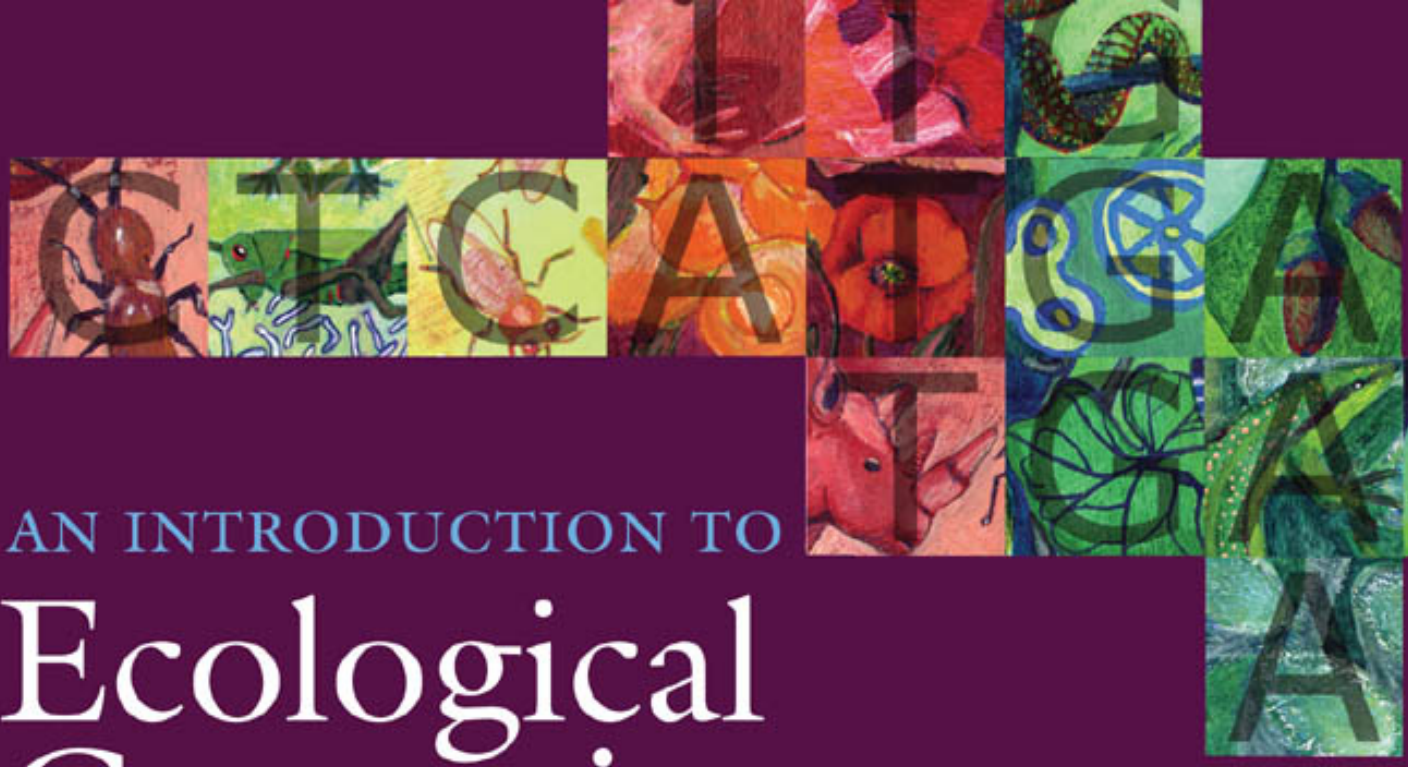
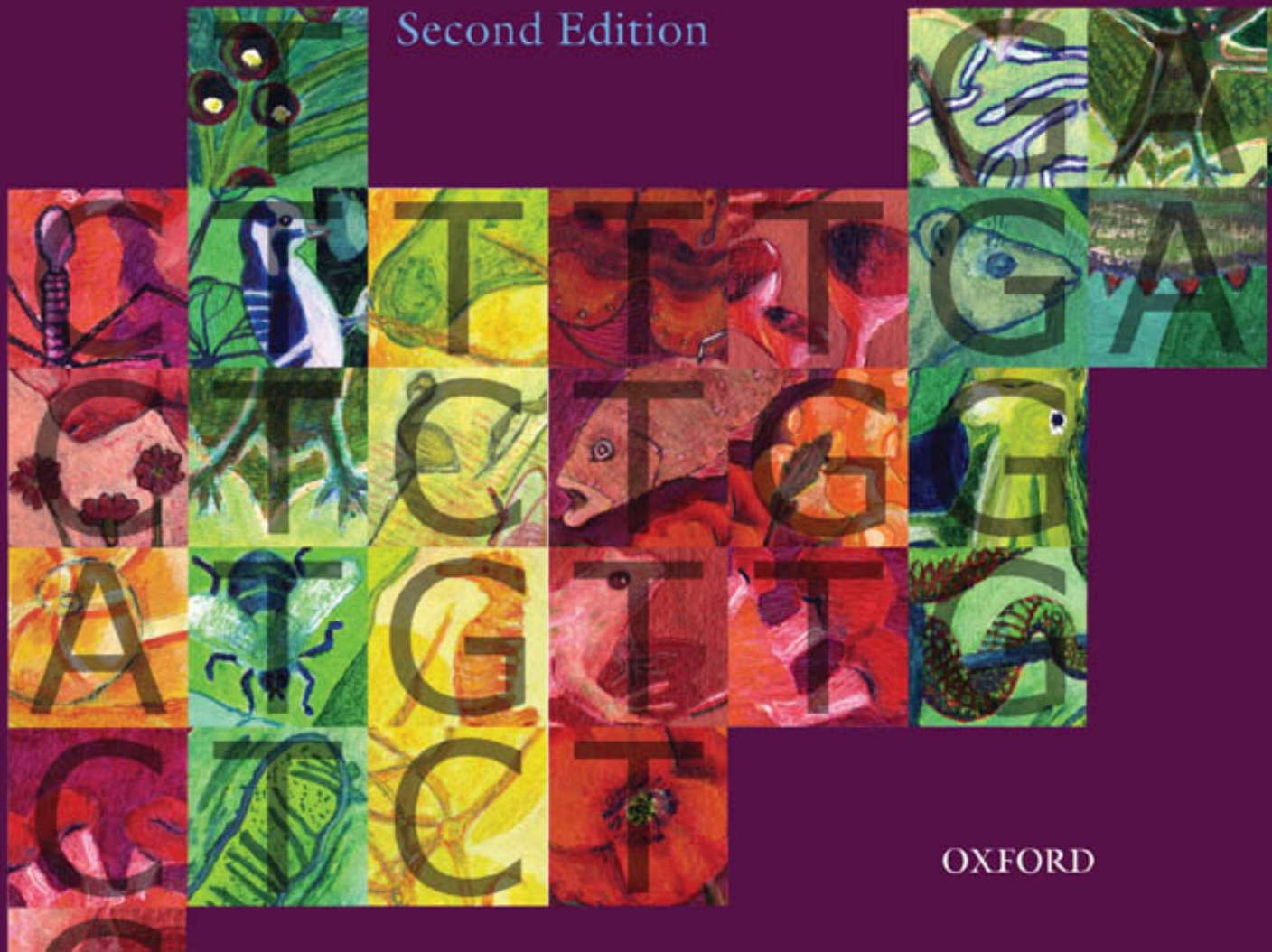


AN INTRODUCTION TO Ecological Genomics

NICO M. VAN STRAALEN
DICK ROELOFS

Second Edition



OXFORD

An Introduction to Ecological Genomics

This page intentionally left blank

An Introduction to Ecological Genomics

Second Edition

Nico M. van Straalen and Dick Roelofs
VU University Amsterdam

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

First edition 2006
Second edition published 2012

British Library Cataloguing in Publication Data
Data available
Library of Congress Cataloging in Publication Data
Library of Congress Control Number: 2011933690
Cover design by Janine Mariën
Typeset by SPI Publisher Services, Pondicherry, India
Printed in Great Britain
on acid-free paper by
CPI Group (UK) Ltd, Croydon, CR0 4YY
ISBN 978–0–19–959468–9 (Hbk)
ISBN 978–0–19–959469–6 (Pbk)
10 9 8 7 6 5 4 3 2

Preface to the second edition

How fast ecological genomics has moved forward since the first edition of this book! The few years that have elapsed since then (2006–2011) have not only seen the rise of extremely fast, so-called 'next-generation' DNA sequencing technology, but also a host of excellent new studies in which genomics technology has been applied to address ecological questions. We are particularly impressed by the massive progress in comparative genomics, phylogenomics, population genomics, and metagenomics. We tried to pay particular attention to the new frontiers created by these fields as we prepared the second edition of this book.

In comparison to the previous edition, we skipped the 'Genome analysis' chapter, as the technology has moved forward to such an extent that a great deal of that chapter was considered outdated. Instead, we added a separate section on methodology and data analysis to Chapter 1, in which we also treat the new-generation sequencing technologies.

Furthermore, we have included a completely new chapter on variation and adaptation, in which we treat the various aspects of genome variability, and pay a good deal of attention to population genomics, a topic of increasing popularity among population-oriented ecological genomicists. In this chapter the reader may also find extensive reference to the issue of neutrality in molecular evolution. In addition we discuss the various aspects of genome architecture in relation to gene expression and epigenetic regulation of genes. Altogether we believe that this new chapter has expanded the scope of the book to include a wider variety of topics of interest to evolutionary ecologists.

We have retained our 'problem-orientated' approach to introduce each chapter, plus an appraisal section at the end. With this design we want to emphasize that, in the end, ecological genomics is just another branch of ecology and that it addresses questions that all ecologists ask, only with different technology. On the other hand, it is our opinion that ecological genomics also brings new questions to the discourse that were not raised before or could not be answered by more traditional ecological approaches.

What was marginally doable in 2005 has now become an impossibility: to summarize all publications of relevance to ecological genomics in a single book. We hope that the reader may find guidance and inspiration in this book to further study more specialized areas of their interest that we could not cover.

We hope this book will support graduate programmes of Ecology, Evolutionary Biology, or similar programmes, and stimulate students to proceed with their career in the exciting field of ecological genomics, while it is—still—relatively new.

Nico M. van Straalen and Dick Roelofs
Amsterdam, February 2011

This page intentionally left blank

Preface to the first edition

This book is an introduction to the exciting new field of ecological genomics, for use in MSc courses and by those beginning their PhD studies.

When we became involved in a national research programme on ecological genomics, or ecogenomics as it became known, we realized that information on this newly emerging subject needed to be brought together. In order to start up a research programme in such a new discipline, not only the students, but also we as teachers, had to get to grips with the subject. Furthermore, although obtaining a PhD implies mastering a specialized field, the PhD student must be able to place this field in a broader context if he or she is to become a mature scientist. This approach may be called the T-model of education; the horizontal bar of the T representing a broad understanding, and the vertical bar an investigation in depth, going down to the root of the problem. Our book uses this approach.

We assume a basic level of knowledge in the biological sciences to BSc level: ecology, evolutionary biology, microbiology, plant physiology, animal physiology, genetics, and molecular biology. We have tried to link up with the content of the most common textbooks in these fields, at the same time realizing that students of ecological genomics have a variety of backgrounds. However, our main targets are students with subjects closely related to ecology and evolutionary biology, which is why we place the emphasis on aspects that we judge to be particularly new to them.

Evolutionary genomics and bioinformatics are companion disciplines to ecological genomics. In the last 10 years interest in both disciplines has grown enormously. Several textbooks on bioinformatics have already been published and subjects encompassed by evolutionary genomics, such as comparative genomics, phylogenetic analysis, and molecular evolution, can now be considered as fields in their own right. They are certainly too large to be covered in an introductory book on ecological genomics; indeed, evolutionary genomics deserves a textbook of its own.

We have organized this book around three issues important in modern ecology, choosing questions for which the links to genomics are best developed. At the outset, we perhaps use rather ambitious phrasing to announce the genomics approach to these ecological questions. Maybe our questions cannot be answered at this stage. However, we decided not to suppress unanswered, and thus open, issues. Instead we hope to stimulate discussion as well as provide factual information. We have included an appraisal section at the end of each chapter to emphasize this question-orientated approach. Combined with information given in the introductory section, this allows the reader to grasp the main points of each chapter, even if the detailed treatment of molecular principles and case studies are left aside.

Case studies are taken from literature published since the year 2000. Nevertheless, a book on genomics runs the risk of becoming outdated very quickly: the rate at which knowledge is being accrued and insight developed is unprecedented. However, we hope that our question-orientated set-up will be useful for some years to come, even when new and better case studies are available.

Before this book was written, journal articles comprised the only literature on ecological genomics. These, although very inspiring, were scattered widely. Today, most textbooks on genetics and evo-

lution have a chapter on genomics. Gibson and Muse published a primer on genome science in 2002, but this did not cover ecological questions. So, for us, writing this book was ploughing unknown ground. We have attempted to add structure to the field, and hopefully have put ecological genomics on the map. However, we welcome constructive criticism and suggestions from our readers.

We thank the colleagues who reviewed parts of the book, suggested issues that had escaped us, or helped with correcting the English: Martin Feder, Claire Hengeveld, Jan Kammenga, René Klein Lankhorst, Bas Kooijman, Jan Kooter, Wilfred Röling, and Martijn Timmermans. We thank Desirée Hoonhout and Karin Uyldert for checking the reference list, and Nico Schaefers, for preparation of the figures. Ian Sherman at Oxford University Press provided us with stimulating discussion. We thank members of the Animal Ecology Department at the Vrije Universiteit for your friendship and encouragement. N.M.vS. also thanks the Faculty of Earth and Life Sciences of the Vrije Universiteit for granting the sabbatical leave during which most of this book was written.

Nico M. van Straalen and Dick Roelofs,
Amsterdam, July 2005

Contents

CHAPTER 1

Ecological genomics and genome analysis

We define ecological genomics as:

a scientific discipline that studies the structure and functioning of a genome with the aim of understanding the relationship between the organism and its biotic and abiotic environments.

With this book we hope to contribute to this new discipline by summarizing the developments over the last ten years and explaining the general principles of genomics technology and its application to ecology. Using examples drawn from the scattered literature, we indicate where ecological questions can be analysed, reformulated, or solved by means of genomics approaches. This first chapter introduces the main purpose of ecological genomics. We describe its characteristics, its interactions with other disciplines, and its fascination with model species. Then we briefly introduce some of the most important technologies and the associated data analysis approaches.

1.1 The genomics revolution invading ecology

The twentieth century has been called the 'century of the gene' (Fox Keller 2000). It began with the rediscovery in 1900 of the laws of inheritance by DeVries, Correns, and Von Tschermak, laws that had been formulated about 40 years earlier by Gregor Mendel. With the appearance of the Royal Horticultural Society's English translation of Mendel's papers, William Bateson suggested in a letter in 1902 that this new area of biology be called genetics. The word gene followed, coined by Wilhelm Ludvig Johannsen in 1909, and then in 1920 the German botanist Hans Winkler proposed the word genome. The term genomics did not appear until the mid-1980s and was introduced in 1987 as the name of a new journal (McKusick and Ruddle 1987). The century ended with the genomics revolution, culminating in the announcement of the completion of a draft version of the human genome in the year 2000.

Realizing the importance of Mendel's papers, William Bateson announced that genetics was to become the most promising research area of the life sciences. One hundred years later one cannot avoid the conclusion that the progress in understanding the role of genes in living systems has indeed been astonishing. The genomics revolution has now expanded beyond genetics, its impact being felt in many other areas of the life sciences, including ecology. In the ecological arena, the interaction between genomics and ecology has led to a new field of research, *evolutionary and ecological functional genomics*. Feder and Mitchell-Olds (2003) indicated that this new multidiscipline 'focuses on the genes that affect evolutionary fitness in natural environments and populations'.

Our definition of ecological genomics given above seems at first sight to include the basic aim of ecology, viewing genomics as a new tool for analysing fundamental ecological questions. However, the merging of genomics with ecology includes more than the incorporation of a toolbox, because with the new technology new scientific questions emerge and existing questions can be answered in a way that was not considered before. We expect

therefore that ecological genomics will develop into a truly new discipline, and will forge a mechanistic basis for ecology that is often felt to be missing. This could also strengthen the relationship between ecology and the other life sciences, because to a certain extent ecological genomicists speak the same language and read the same papers as molecular biologists.

Figure 1.1 illustrates the various fields from which ecological genomics draws and upon which it is still growing. First of all, as indicated by Feder and Mitchell-Olds (2003), ecological genomics is closely linked to evolutionary biology and the associated disciplines of population genetics and evolutionary ecology. Another major area supporting ecological genomics is plant and animal physiology, which has its base in biochemistry and cell biology. A special position is held by microbial ecology, the meeting place of microbiology and ecology, where the use of genomics approaches has proceeded further than in any other subdiscipline of ecology. We consider genomics itself as a mainly technological advance, supporting ecological genomics in the same way as it supports other areas of the life sciences, such as medicine, neurobiology, and agriculture.

The genomics revolution is not only due to advances in molecular biology. Three major technological developments that took place in the 1990s also made it possible: microtechnology, computing, and communication.

Microtechnology. The possibility of working with molecules on the scale of a few micrometres, given by advances in laser technology, has been very important for one of genomics' most conspicuous achievements, the development of the gene chip.

Computing technology. To assemble a genome from a series of sequences requires tremendous computational power. Extensive calculations are also necessary for the analysis of expression matrices and protein databases. Without the advent of high-speed computers and data-storage systems of vast capacity all this would have been impossible.

Communication technology. Consulting genome databases all over the world has become such normal practice that the scientific progress of any genomics laboratory has become completely dependent on communication with the rest of the World Wide Web. The Internet has become an indispensable part of genomics.

The essence of genomics is that it is the study of the genome and its products *as a unitary whole*. In biology, the suffix -ome signifies the collectivity of units (Lederberg and McCray 2001), as for example in coelome, the system of body cavities, and biome, the entire community of plants and animals in a climatic region. In aiming to investigate many genes at the same time genomics differs from ecology,

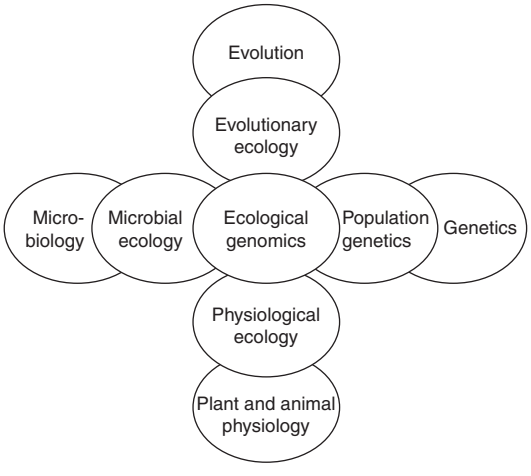


Figure 1.1 The position of ecological genomics in the middle of the other life-science disciplines with which it interacts most intensively.

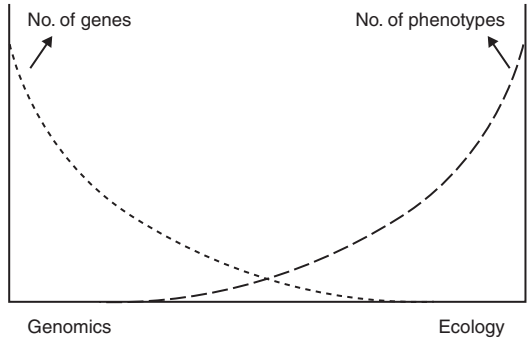


Figure 1.2 The playing field of ecological genomics, in between genomics, with its focus on the single genome of a model organism, studying all the genes that it contains, and ecology, studying a few genes in many species.

which although investigating many phenotypes, usually deals with only a few genes at a time (Fig. 1.2). Ecological genomics borrows from these two extremes, investigating phenotypic biodiversity as well as diversity in the genome. With this new discipline, ecology is enriched by genomics technology and genomics is enriched by ecological questioning and evolutionary views.

Because genomics analyses the genome in its entirety, it transcends classical genetics, which studies genes one by one, relating DNA sequences to proteins and ultimately to heritable traits. Genomics is based on the observation that the impact of one gene on the phenotype can only be understood in the context of the expression of several other genes or, in fact, of all other genes in the genome, plus their products, metabolites, cell structures, and all the interactions between them. This is not to say that every study in genomics deals with everything all the time, but that the mind is set and tools are deployed to maximize awareness of any effects elsewhere in the genome, outside the system under study. Consequently genomics is invariably associated with unexpected findings. The discovery aspect of genomics is expressed aptly in a public-education project of Genome Canada entitled *The GEEE! in Genome* (www.genomecanada.ca).

The work of Spellman and Rubin (2002) and their discovery of *transcriptional territories* in the genome of the fruit fly, *Drosophila melanogaster*, is an example of how the genomics approach can fundamentally alter our way of thinking about the relationship between genes and the environment (see also Weitzman 2002). The authors carried out transcription profiling with DNA microarrays (see Section 1.4) to investigate the expression of almost all of the genes in the fruit fly's genome under 88 different environmental conditions. Their work was in fact a meta-analysis of transcription profiles collected earlier in six separate investigations. Because the complete genome sequence of *Drosophila* is known, it was possible to trace every differentially expressed gene back to its chromosomal position. They concluded that genes physically adjacent in the genome often had similar expression when comparing different environmental challenges. The window of correlated expression appeared to extend to 10 or more adjacent genes and they estimated that 20% of the genome was organized in such 'expression clusters'. Most astonishingly, genes in one cluster proved to be no more similar in structure or function than could be expected from a random arrangement. Spellman and Rubin (2002) suggested that local changes in chromatin structure trigger the expression of large groups of genes together. Thus a gene may be expressed not because there is a particular need for its product, but because its neighbour is expressed for a reason completely unrelated to the function of the first gene. At the moment it is not known whether such mechanisms lead to unexpected correlations between phenotypic traits, but surely the discovery of transcriptional territories could never have been made on a gene-by-gene basis, and this is due to the genomics approach.

1.2 Yeast, fly, worm, and weed

A striking feature of genomics is its focus on a limited number of *model species* with fully sequenced genomes and large research networks organized around them. The genomes of these model species have been sequenced completely and the information is shared on the Internet, allowing scientists to take maximal advantage of progress made by others. This explains the extreme speed with which the field is developing. Ecology does not have a strong tradition in standardized experimentation with one species. Thus the genomics approach is all the more striking to an ecologist, who is often more fascinated by the diversity of life than by a single organism, and engaged in a very wide variety of topics, systems, and approaches. In this section we examine the arguments for introducing model species in ecological genomics.

The best-known completely sequenced genomes, in addition to those of mouse and human, are those of the yeast *Saccharomyces cerevisiae*, the 'fly' *Drosophila melanogaster*, the 'worm' *Caenorhabditis elegans*, and the 'weed' *Arabidopsis thaliana*. Investigations into the genomes of these model organisms are supported by extensive databases on the Internet that provide a wealth of information about genome maps, genomic sequences, annotated genes, allelic variants, cDNAs, and expressed

sequence tags (ESTs), as well as news, upcoming events, and publications. These four model genomes and their relationships with evolutionary related species will be discussed in more detail in Chapter 2. The genomics of the mouse and human are not discussed at length in this book because the model status of these two species has mainly a medical relevance.

The first genome to be sequenced completely was that of *Haemophilus influenzae* (Fleischmann *et al*. 1995). This bacterium is associated with influenza outbreaks, but is not the cause of the disease, which is a virus. Although several years earlier the 'genome' of bacteriophage ΦX174 had been sequenced (Sanger 1977a), 1995 is considered by many as the true beginning of genomics as a science, not in the least because the *H. influenzae* project demonstrated the usefulness of a new strategy of sequencing and assembly (whole-genome shotgun sequencing; see Section 1.4). With 1.8 Mbp the genome of *H. influenzae* was about 10 times larger than that of any virus sequenced before, but still two to four orders of magnitude smaller than the genome of most eukaryotes. Genome sequences of many other prokaryotes soon followed, including that of *Methanococcus jannaschii*, an archaeon living at a depth of 2600 m near a hydrothermal vent on the floor of the Pacific Ocean (Bult *et al*. 1996). The genome of this *extremophile* was interesting because of the many genes that were completely unknown before. In 1989, a large network of scientists embarked on a project for sequencing the yeast genome, which was completed in 1996 and was the first eukaryotic genome to be elucidated (Goffeau *et al*. 1996). Thus, by 1996, the first genomic comparisons were possible between the three domains of life: Bacteria, Archaea, and Eukarya.

The international *Human Genome Project* initiated by the US National Institutes of Health and the US Department of Energy, was launched in 1990 with completion due in 2005. However, in the meantime a private enterprise, Celera Genomics, embarked on a project with the same aim but a different approach and actually overtook the Human Genome Project. The competition was settled with the historic press conference on 26 June 2000, when US President Bill Clinton, J. Craig Venter of Celera Genomics, and Francis Collins of the National Institutes of Health jointly announced that a working draft of the human genome had been completed (Fig. 1.3). Many commentators have qualified this announcement as more a matter of public communication than scientific achievement. At that time the accepted criterion for completion of a genome sequence, namely that only a few gaps or gaps of known size remained to be sequenced and that the error rate was below 1 in 10 000 bp, had not been closely met. The euchromatin part of the genome was not completed until mid-2004, although that milestone was again considered by some to be only the end of the beginning (Stein 2004). Nevertheless, the Human Genome Project can be regarded as one of the most successful scientific endeavours in history and the assembly of the 3.12 billion bp of DNA, requiring some 500 million trillion sequence comparisons, was the most extensive computation that had ever been undertaken in biology.

New ultra high-throughput sequencing techniques, also called next-generation sequencing (the technologies will be explained in Section 1.4), have caused a second revolution in genome sequencing. The number of organisms whose genome has been sequenced completely and published has exceeded 1300 (Liolios *et al.* 2009). By 2010, no fewer than 188 Archaea, 4800 bacterial organisms, and 1524 eukaryotes were the subject of ongoing genome sequencing projects. Bacteria dominate the list, as the small size of their genomes makes these organisms well-suited for whole-genome sequencing.

The list of species with completed genome sequences does not represent a random choice from the Earth's biodiversity. From an ecologist's point of view, the near absence of reptiles, amphibians, molluscs, and annelids is striking, as also is the scarcity of birds and arthropods other than the insects. How does a species come to be a model in genomics? We review the various arguments below, asking whether they would also apply when selecting model species for ecological studies.

Previously established reputation. This holds for yeast, *C. elegans*, *Drosophila*, mouse, and rat. These species had already proved their usefulness as models before the genomics revolution and were adopted by genomicists because so much was

Figure 1.3 From left to right: J. Craig Venter (Celera Genomics), President Clinton, and Francis Collins (National Institutes of Health) on the historic announcement of 26 June 2000 of the completion of a working draft of the human genome. © Win McNamee/Reuters.

known about their genetics and biochemistry and, perhaps just as importantly, because a large research community was interested, could support the work, and use the results.

Genome size. One of the first questions that is asked when a species is considered for whole-genome sequencing is, what is the size of its genome? At least in the beginning, a relatively small genome was a major advantage for a sequencing project. The *genome size* of living organisms ranges across nine orders of magnitude, from 10^3 bp (0.001 Mbp) in RNA viruses to nearly 10^{12} bp (1000 000 Mbp) in some protists, ferns, and amphibians (cf. Fig. 2.1). The puffer fish, *Takifugu rubripes*, was chosen because of its relatively small genome (one-eighth of the human genome). The issue of genome size has become less important over the years, due to the rise of faster and faster sequencing technologies.

Possibility for genetic manipulation. The possibility of genetic manipulation was an important reason why *Arabidopsis*, *Drosophila*, and mouse became such popular genomic models. The ultimate answer about the function of a gene comes from studies in which the genome segment is knocked out, down-regulated, or overexpressed against a genetic background that is the same as that of the wild type. Also, the introduction of constructs in the genome that can report activity of certain genes by means of signal molecules is very important. This can only be done if the species is accessible using recombinant-DNA techniques. Foreign DNA can be introduced using *transposons*; for example, modified P-elements that can 'jump' into the DNA of *Drosophila*, or bacteria such as *Agrobacterium tumefaciens* that can transfer a piece of DNA to a host plant. DNA can also be introduced by physical means, especially in cell cultures, using electroporation, microinjection, or bombardment with gold particles. Another popular approach is post-transcriptional gene silencing using *RNA interference* (RNAi), also called inhibitory RNA expression. The question can be asked,

should the possibility for genetic manipulation be an argument for selecting model species in ecological genomics? We think that it should, knowing that the capacity to generate mutants and transgenes of ecologically relevant species is crucial for confirming the function of genes. Ecologists should also use the natural variation in ecologically relevant traits to guide their explorations of the genome (Koornneef 2004; Tonsor *et al*. 2005, see also Chapter 6). A basic resource for genome investigation can be obtained by using natural varieties of the study species, and developing genetically defined culture stocks.

Medical or agricultural significance. Many bacteria and parasitic protists were chosen because of their pathogenicity to humans. Other bacteria and fungi were taken as genomic models because of their potential to cause plant diseases (phytopathogenicity). Obviously, the sequencing of rice was motivated by the huge importance of this species as a staple food for the world population (Adam 2000). Some agriculturally important species have great relevance for ecological questions; for example, the bacterium *Sinorhizobium meliloti*, a symbiont of leguminous plants, is known for its nitrogen-fixing capacities, but it also makes an excellent model system for the analysis of ecological interactions in nutrient cycling, together with its host *Medicago truncatula*.

Biotechnological significance. Many bacteria and fungi are important as producers of valuable products, for example antibiotics, medicines, vitamins, soy sauce, cheese, yoghurt, and other foods made from milk. There is considerable interest in analysing the genomes of these microorganisms because such knowledge is expected to benefit production processes (Pühler and Selbitschka 2003). Other bacteria are valuable genomic models because of their capacity to degrade environmental pollutants; for example, the marine bacterium *Alcanivorax borkumensis* is a genomic model because it produces surfactants and is associated with the biodegradation of hydrocarbons in oil spills (Röling *et al*. 2004; Head *et al.* 2006).

Evolutionary position. Whole-genome analysis of organisms at crucial or disputed positions in the tree of life can be expected to contribute significantly to our knowledge of evolution. The sea squirt, *Ciona intestinalis*, was chosen as a model because it belongs to a group, the Urochordata, with properties similar to the ancestors of vertebrates. The study of this species should provide valuable information about the early evolution of the phylum to which we belong ourselves. *Methanococcus jannaschii* was chosen for more or less the same reason, because it was the first sequenced representative from the domain of the Archaea. Many other organisms, although not on the list for a genome project to date, have a strong case for being declared as model species for evolutionary arguments. These include the velvet worm, *Peripatus*, traditionally seen as a missing link between the arthropods and annelids, but now classified as a separate phylum in the Panarthropoda lineage (Nielsen 1995), and the springtail, *Folsomia candida*, regarded as ancestral to all insects (Timmermans *et al.* 2008).

Comparative purposes. Over the last few years, genomicists have realized that assigning functions to genes and recognizing promoter sequences in a model genome can greatly benefit from comparison with a set of carefully chosen reference organisms at defined phylogenetic distances. Comparative genomics is developing an increasing array of bioinformatics techniques, such as *synteny analysis*, *phylogenetic footprinting*, and *phylogenetic shadowing* (see Chapter 2), by which it is possible to understand aspects of a model genome from other genomes. One of the main reasons for sequencing the chimpanzee's genome was to illuminate the human genome, and a variety of fungi were sequenced to illuminate the genome of *S. cerevisiae*.

Ecological significance. Since the completion of the human genome ecological arguments have played an increasing role in the selection of species for whole-genome sequencing, and we expect them to become more important in the future. Jackson *et al*. (2002) have formulated arguments for the selection of ecological model species, and we present them in a slightly adapted form.

Diversity of physiologies. The new range of models should embrace diverse phylogenetic lineages, varying in their physiology and life-history strategy. For example, the model plants *Arabidopsis* and rice both employ the C3 photosynthetic pathway. To complement our understanding of primary produc-

tion, genomic analysis of plants utilizing C4 photosynthesis (e.g. sorghum) or crassulacean acid metabolism (CAM) will be highly informative. Considering the diversity of life histories, species differing in their mode of reproduction and dispersal capacity should be chosen; for example, hermaphroditism versus gonochorism, parthenogenesis versus bisexual reproduction, and so on.

Ecological interactions. Species that take part in critical ecological interactions (mutualisms, antagonisms) are obvious candidates for genomic analysis. One may think of mycorrhizal fungi, nitrogen-fixing symbionts, pollinators, natural enemies of pests, parasites, and so on. The most obvious strategy for analysing such interactions is to sequence the genomes of the players involved and to try and understand interactions between them from mutually influential gene expressions. An exciting model is presented by the pea aphid, *Acyrthosiphon pisum*, and its obligate bacterial symbiont, *Buchnera aphidicola*. The symbiotic bacterium has lost many genes, of which a few have been transferred to the genome of its host. Analysis of the two genomes indicates the presence of extensive exchange of metabolites between symbiont and host, for instance a shared amino acid biosynthesis (Perez-Brocal *et al.* 2006; Richards *et al.* 2010).

Suitability for field studies. The wealth of knowledge from experienced field ecologists should play a role in deciding about new 'ecogenomic' models. Not all species lend themselves to studies of behaviour, foraging strategy, habitat choice, population structure, dispersal, or migration in the field, simply because they are too rare, not easily spotted, difficult to sample quantitatively, impossible to mark and recapture, not easy to distinguish from related species, or inaccessible to invasive techniques. Thus suitability for field research is another important criterion. The threespine stickleback, *Gasterosteus aculeatus*, and water flea, *Daphnia pulex*, are considered to be highly suitable for ecological field studies and have a long standing history of ecological investigation. Recently, genome sequences for both of these ecological models have become available.

Feder and Mitchell-Olds (2003) developed a similar series of criteria for an ideal model species in evolutionary and ecological functional genomics (Fig. 1.4). These authors point out that there is currently a discrepancy between classical model species and many ecologically interesting species. Models such as *Drosophila* and *Arabidopsis* are not very suitable for ecological studies, whereas many popular ecological models have a poorly characterized genome and lack a large community of investigators. In some cases a large ecological community is available, but functional genomic studies are difficult for reasons of quite another nature. For example, many ecologists favour wild birds as a study object, but there are ethical objections to genetic manipulation of such species and laboratory experiments are restricted by law.

Still, we foresee that all the major ecological models will also become genomic models. Using next-generation sequencing technologies extremely large amounts of sequence data can by generated in a very cost-effective way. The saturation point could very well be due to the limited number of molecular ecologists in the worldwide scientific community. This is not to say, however, that all questions in ecological genomics require the full-length DNA sequence of a species before they can be answered. Some issues may prove to be solvable with the use of less extensive genomic investigations, for example a gene hunt followed by high-throughput quantitative PCR, rather than transcription profiling of the complete genome (see Section 1.4). In addition, microarray studies with part of the expressed genome are possible even in species lacking a complete DNA sequence. Microarrays can be manufactured at costs that are affordable for small research groups if they are limited to genes associated with a specific function or response pathway (Held *et al.* 2004).

Not all ecological models will enjoy the type of in-depth investigation now dedicated to yeast, fly, worm, and weed. Murray (2000) points out that the development of genome-based tools has a strong element of positive feedback; the rich—that is, widely studied organisms—get richer and the poor get poorer. This development has already been felt in the fields of animal and plant physiology, where many of the species traditionally investigated in comparative physiology and biochemistry have

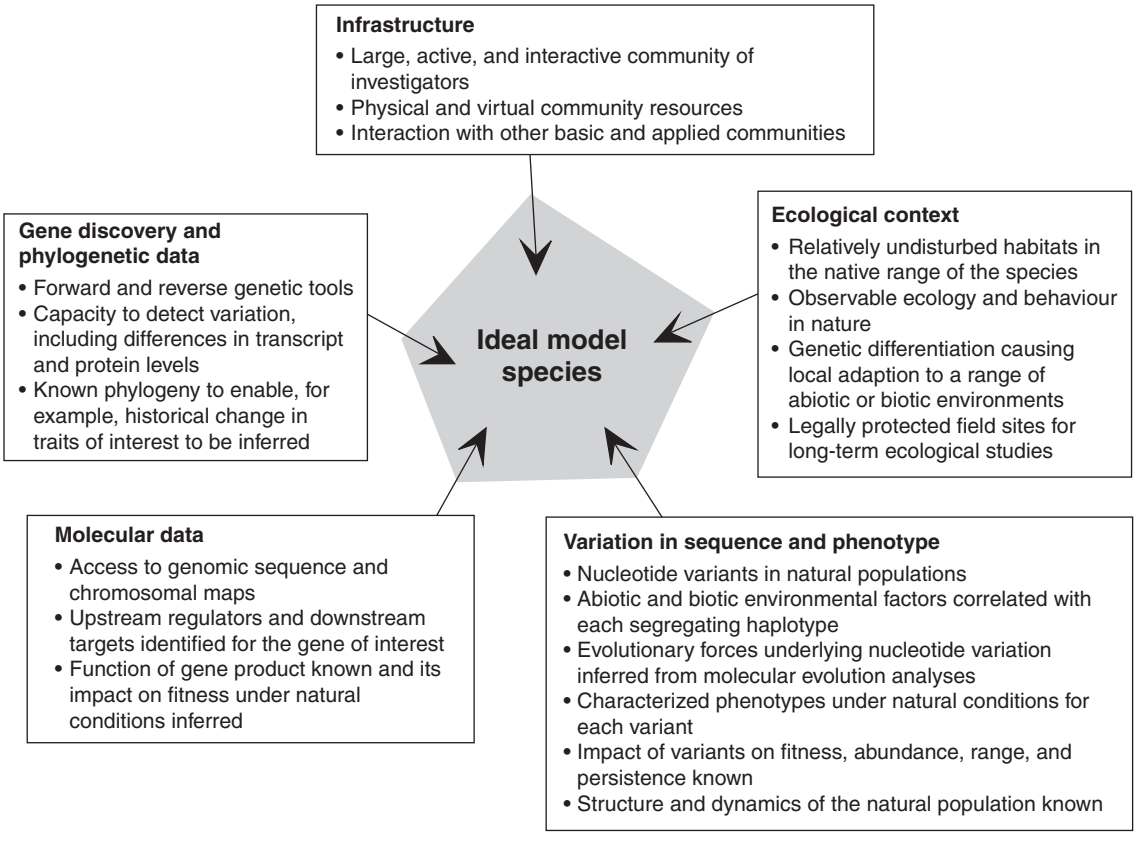


Figure 1.4 Criteria in evolutionary and ecological functional genomics for a model species, according to Feder and Mitchell-Olds (2003). At present few species satisfy all criteria. Reproduced by permission of Nature Publishing Group.

been abandoned in favour of models that can be genetically manipulated to study the function of genes. Murray (2000) predicted that 'the larger its genome and the fewer its students, the more likely work on an organism is to die'. Crawford (2001) has argued, however, that functional genomics should resist this tendency and instead choose species best suited to addressing specific physiological or biochemical processes. For example, the Nobel Prize for Medicine was given to H. A. Krebs for his research on the citric acid cycle, which was conducted on common doves. By modern standards the dove is a non-model species, but it was chosen because its breast muscle is very rich in mitochondria. In animal physiology, *Krogh's principle* assumes that for every physiological problem there is a species uniquely suited for its analysis (Gracey and Cossins 2003). According to this principle, genomic standard species are likely to be suboptimal for at least some problems of physiology, because no model is uniquely suited to answering all questions.

DNA microarrays, with their associated massive generation of data on expression profiles (see Section 1.4), are one of the most tangible features of modern genomics and are often seen as holding the greatest promise for solving problems in ecology. However, not all ecologists are convinced that microarray-based transcription profiling is the best way to advance the genomics revolution into ecology. Some authors suggest that microarrays have already been overtaken by next-generation sequencing methods, which allow gene expression profiles to be developed from brute force sequencing, rather

than from hybridization (see Section 1.4). Thomas and Klaper (2004) saw a drawback in the fact that genome-wide microarrays are available only for genomic model species, whereas the interest of ecologists is with species that are important in the environment and amenable to ecological studies; these two interests do not necessarily coincide.

Some authors have solved these issues by using microarrays of model species to profile the transcriptome of non-models. In these *cross-species hybridizations* it is assumed that there is sufficient homology between the non-model and the model to allow differential expressions to be assessed reliably. For example, rainbow trout can be a reference for other salmonid fish (Von Schalburg *et al*. 2005), and *Arabidopsis thaliana* may function as a model for other species of the family Brassicaceae. As an example of a successful cross-species hybridization study, Van de Mortel *et al.* (2006) studied gene expression in roots of *A. thaliana* and *Thlaspi caerulescens* plants grown under deficient and excess zinc supply. They applied the Agilent Arabidopsis 3 60-mer microarray containing all 27 000 annotated *Arabidopsis* genes complemented by 10 000 nonannotated transcripts. Over 2000 genes showed significant differential expression between *A. thaliana* and *T. caerulescens* at each zinc exposure. Many of these genes appeared to function in metal homeostasis, abiotic stress response, and lignin biosynthesis. Obviously, the success of such experiments is dependent on the sequence divergence between model and non-model, although this is not always decisive (Bar-Or *et al.* 2007). Cross-species hybridization seems to work best when using long probes (cDNAs) and is risky in the case of short oligos. It is always advisable to do a comparative genome hybridization (CGH) to check for probes on the array that have insufficient homology across species (Machado *et al.* 2009).

The use of microarrays in ecology to better understand genetic mechanisms underlying species interactions, adaptations, and evolutionary processes has increased rapidly (Kammenga *et al.* 2007). With the help of next-generation sequencing technology (see Section 1.4) it is feasible to establish a whole transcriptome microarray within 12 months of commencing a project; neither the estimated expense nor the availability of technology need be major obstacles for progress. Given the fact that the number of completely sequenced organisms is increasing month by month, we can expect that within a few years the genomes of almost all ecologically relevant species will be available to be probed to address a multitude of ecological questions.

1.3 -Omics speak

Because of the immediately attractive upswing created by the genomics revolution, and the large financial resources made available in many industrialized countries, adjacent fields of science have adopted terms echoing genomics, leading to a great proliferation of designations such as transcriptomics, proteomics, and metabolomics, such that some biologists have complained that what was molecular biology before is now named after one of the '-omics' but in fact is still molecular biology. Zhou *et al*. (2004) proposed a classification of genomics according to three main categories: approach (structural or functional), scientific discipline (evolutionary genomics, ecological genomics, etc.), and object of study (plant genomics, microbial genomics, etc.). An Internet page maintained by Mary Chitty (Cambridge Healthtech Institute) provides a glossary of the various terms that have arisen with the emergence of '-omics' technologies (www.genomicglossaries.com). There are obvious terms such as pharmacogenomics and cardiogenomics, and awkward ones such as saccharomics (the study of all the carbohydrates in the cell) and vaccinomics (the use of genomics for vaccine development). The three most common extensions of genomics are transcriptomics, proteomics, and metabolomics, and these are introduced briefly here, with reference to Fig. 1.5.

Transcriptomics is the study of all the transcripts that are present at any time in the cell. In principle the transcriptome includes messenger RNAs (mRNAs) in addition to ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and small nuclear RNAs (snRNAs), but transcriptomics is usually limited to mRNA, the template for translation into protein. The main activity in transcriptomics is to obtain a

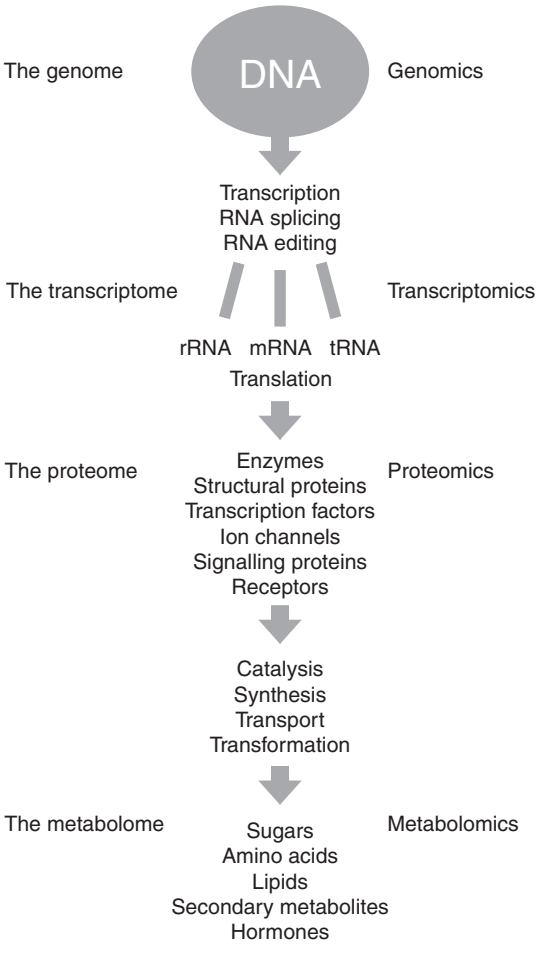


Figure 1.5 The relationship between genomics, transcriptomics, proteomics, and metabolomics.

profile of global gene expression in relation to some condition of interest. Which genes are turned 'on' and 'off' during certain phases of the cell cycle? Which genes are upregulated by certain physiological conditions? Which genes change their expression in response to adaptation to the environment? The study of transcriptomes is a core activity of *functional genomics*, because it does not look at the DNA as such but at its functions.

In general, it is expected that there are more transcripts than there are protein-encoding genes in the genome, even when considering only those genes that are actually transcribed. This is due to the mechanism of *alternative splicing*: the generation of different mRNAs from the same pre-mRNA during the removal of introns. *RNA editing* (post-transcriptional insertion or deletion of nucleotides, or conversion of one base for another) is another reason for incongruence between the genome and the transcriptome.

There are more reasons why a functional analysis of the genome can provide a different picture than an inventory of genes. Obviously, all cells of an organism have the same genome, but not the same transcriptome. Even when looking at cells of the same type, the transcriptome depends on environmental conditions, physiological state, developmental state, and so on. So the transcriptome allows a glimpse of the living cell much more than the genome itself. The argument also holds when making comparisons across species. Classical molecular phylogenetics (see Graur and Li 2000) is based on variation of homologous DNA sequences across species. However, the same structural DNA can be regulated in different ways in different species. We illustrate this argument with an example from Enard *et al*. (2002), who did one of the first studies in *comparative transcriptomics*.

Enard *et al*. (2002) analysed the expression of 18 000 genes in liver, blood leucocytes, and brain tissue of humans, chimpanzee (*Pan troglodytes*), and rhesus monkey (*Macaca mulatta*). The expression patterns in human blood and liver turned out to be more similar to chimpanzees than to rhesus monkeys, which is in accordance with the phylogenetic distances between the three primate species; however, the expression profiles in the brain were more similar between chimpanzee and rhesus monkey than between either of the two primates and human (Fig. 1.6). So, although chimpanzees share 98.7% of their DNA with humans, the human species expresses that DNA in a different manner, especially in the brain. Gene expression in the brain has undergone accelerated evolution compared to gene expression elsewhere in the body, and evolution has resulted in a divergence of humans from chimpanzees, mostly due to regulatory change rather than structural reorganization of the DNA. This suggestion by Enard *et al.* (2002) was confirmed in a later study by Gilad *et al.* (2006) who showed that the

genes differentially expressed between humans and chimpanzee are enriched in transcription factors, the regulatory proteins that affect the expression of other genes.

Proteomics is the study of all the proteins in the cell. As with genomics, proteomics arose thanks to technological innovation, which in this case is *tandem mass spectrometry* (MS/MS) and liquid chromatography coupled to tandem mass spectrometry (LC/MS/MS). The idea is to separate a mixture of soluble proteins by means of chromatography and then to estimate masses, first of the larger peptide and, after a second ionization, of fragments of the same peptide. The fragment patterns provide a fingerprint characteristic of the protein. Interpretation of proteomics data is usually supported by genomic sequence information, in such a way that an observed peptide fragment pattern may be compared to a database of proteins predicted from the genome. Mass spectrometry may also be used to determine the amino acid sequence of a protein. For this application, the protein is cleaved with a protease, for example trypsin, which generates a collection of fragments characteristic of the protein. These fragments may be compared to an *in silico* (computer-simulated) digestion derived from the genome and the known cleavage sites of the protease.

The proteome provides a different picture of a cell's activities to the transcriptome. Several authors have indeed wondered about the lack of correlation between mRNA and protein abundances. One of the reasons for this is the existence of control mechanisms at the ribosomes, where mRNA is translated to peptides. *Translational control* allows the cell to select only certain mRNAs for translation and block others. The selection is often dependent on environmental conditions, so this mechanism allows for physiological adaptation on the level of the proteome, even though the transcriptome remains the same. Another issue is *post-translational modification* or *protein processing*, processes that can greatly affect the function of a protein, for example by acetylation or ubiquitination of the N-terminal residue, hydroxylation of prolines, or cleavage of the molecule into smaller units. The proteome and the genome are linked by many feedback mechanisms, because some proteins are transcription factors necessary for gene activation, others are enzymes involved in transcription or translation, and still others are structural components of chromosomes. So, in a molecular biology context, the living cell can only be understood fully by considering genome, transcriptome, and proteome together.

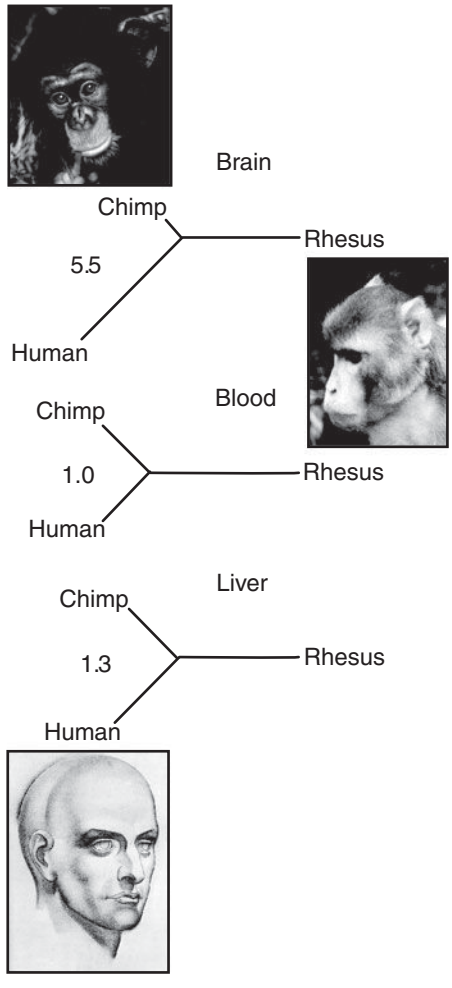


Figure 1.6 Distance trees showing the similarity of gene-expression profiles in brain, blood leucocytes, and liver of human, chimpanzee, and rhesus monkey. Numbers refer to the ratio between the rate of evolution in the human and the chimpanzee lineages, taking the rhesus monkey as an outgroup. Reprinted with permission from Enard *et al*. (2002). Copyright 2002 AAAS.

As an example of a study applying proteomics in an environmental context, consider the work of Martyniuk *et al.* (2010). These authors studied the

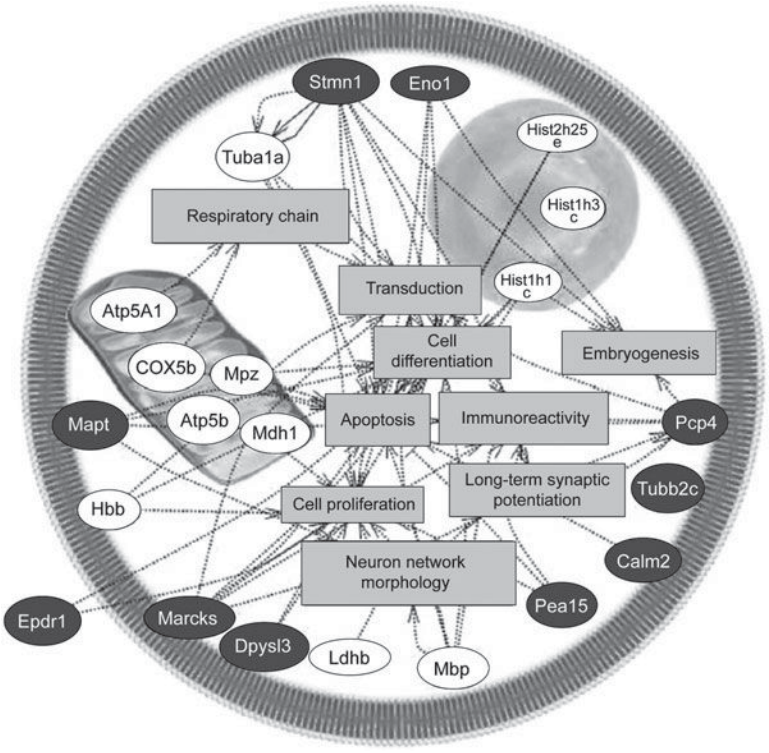


Figure 1.7 Pathway analysis of significantly altered proteins in neuroendocrine tissue of fathead minnow after exposure to 17α-ethinylestradiol (EE_2). Dark ovals indicate protein induction, white ovals indicate protein reduction. Abbreviations are protein functions associated with the depicted cell processes (rectangles). After Martyniuk *et al*. (2010) with permission from Elsevier.

effects of estrogen at the proteomic level on neuronal processes in male fathead minnows. The substance 17α-ethinyloestradiol (EE_2) is a derivative of the sex hormone oestradiol and the most commonly used bioactive estrogen in contraceptive pills. It is released into the environment as xenoestrogen from urine and faeces of medicated individuals and has been associated with feminization of male fish, a process called endocrine disruption. Martyniuk identified 14 downregulated and 12 upregulated proteins. Functional analysis of these proteins revealed that cell differentiation and proliferation, neuron network morphology, and long-term synaptic potentiation were affected by environmentally relevant EE_2 concentrations. The functional relationships among these biological processes are summarized in Fig. 1.7. Interestingly, transcriptional profiles of the genes encoding the affected proteins did not correlate with the direction and level of protein abundance, suggesting complex feedback mechanisms between RNA and protein level regulation. The functional genomics of endocrine disruption will be discussed in more detail in Chapter 5.

Metabolomics is the study of all low-molecular-weight cellular constituents. Usually only metabolites belonging to a limited category are included, for example all soluble carbohydrates, or all metabolites that can be measured by a certain analytical technique such as pyrolysis gas chromatography or infrared spectrometry. No single method can measure the thousands of different chemical compounds that may be present at any time in a cell because of the greatly diverging chemical properties (hydrophilic versus hydrophobic compounds, acids versus bases, reactive versus inert compounds, etc.).

The metabolome requires a diversity of analytical approaches to obtain a complete picture.

An increasing number of proteomic and metabolomic studies are addressing ecological questions. For instance, metabolomic analysis has provided new insights into the mechanisms underlying plant resistance to herbivores (Macel *et al.* 2010) and abiotic stress in earthworms (Bundy *et al.* 2009). A compelling environmental metabolomics study was published by Hines *et al.* (2007). These authors asked the question whether metabolomics can identify hypoxia-induced phenotypes in mussels in a highly variable environment or whether these animals should be stabilized in a controlled laboratory environment prior to sampling. Metabolic fingerprints were generated from adductor muscle tissue of *Mytilus galloprovincialis*. Animals were taken directly from the field, with and without hypoxia, and compared to animals (with and without hypoxia) that were profiled after a 60 hour stabilization period in the laboratory. Surprisingly, the lab acclimation for 60 hours completely blurred the hypoxia signal that was clearly visible in the field-sampled animals. Apparently, direct field sampling minimizes metabolic variability and enables stress-induced phenotyping.

It is essential for an organism to integrate mRNA levels with protein and metabolite fluxes. Thus, a relation between the three levels of organization might be expected. For instance, Cardozo *et al.* (2003) found a very high correlation between mRNA abundance and protein expression in 60 genes of mouse islet cells. In ecological genomics, gene expression is often directly related to fitness without considering protein activities or metabolite fluxes. A provoking review by Feder and Walser (2005) showed that mRNA levels can sometimes predict protein levels but often cannot. In the study on oestrogen disruption in fathead minnows, opposite directions of regulation between proteomic and transcriptomic data were shown. It seems that the relationship between quantities of mRNAs and proteins is often unclear and, if present, is insufficient to explain the interdependencies between them.

The complementation of transcriptomics data with metabolomics seems more promising. In a 'systems toxicology' approach Bundy *et al.* (2008) paid particular attention to metabolic categories supported by high throughput approaches. In a joint analysis of transcriptomes and metabolomes during copper exposure in the earthworm *Lumbricus rubellus*, the authors were able to filter out noise inherent to gene expression and to select those gene expression patterns that were consistent with the metabolome. Such a systems biology analysis may avoid the issue raised by Feder and Walser (2005), that too much confidence is put on transcriptomics to explain phenotypic changes.

Despite the promise of metabolomics, comparative genomics and transcriptomics still dominate ecological genomics studies, and that is why these two -omics techniques do not play a major role in this book. In Chapter 7 we will address some aspects of metabolomics when discussing metabolic networks. Finally, Table 1.1 describes some other terms used in connection with genomics.

With the further development of ecological genomics, applications will also come within reach. One can envisage a multitude of issues where a better knowledge of genomes in the environment can support measures to improve ecosystem health, risk assessment of pollution, conservation of endangered species, and so on (Greer *et al*. 2001). Such applications fall outside the scope of this book; however, we mention two examples below, to sketch the range of possibilities.

Purohit *et al*. (2003) suggested that multilocus DNA fingerprints prepared from environmental samples could act as an *indicator DNA signature* (IDS); for example, fingerprints of microbial soil communities could be indicative of soil pollution. Their suggestion can be extended to involve transcription profiles that are characteristic of certain environmental conditions or physiological states. Figure 1.8 illustrates this principle. When an organism is exposed to polluted soil, this will be accompanied by gene expression that has both a general aspect due to the generality of the stress response and a specific aspect which characterizes the challenge (see also Chapter 5). When the expression profile observed for a suspect soil is compared with a database of reference profiles, the type of pollution and its biological effects may be indicated (Fig. 1.8). This may help to support

decisions about the urgency of remediation measures.

As a second example of a possible application consider the case of soil-borne pathogens. Many pathogens attacking economically important crops are difficult to control by conventional strategies such as the use of host resistance and synthetic pesticides. However, some soils have an inherent capacity to suppress diseases and such soils need lower rates of pesticide application to combat them. *Disease-suppressive capacity* is due to the presence of genes involved in antibiotic production by antagonistic microorganisms (Van Elsas *et al*. 2002; Weller *et al*. 2002; Garbeva *et al*. 2004). In several cases,

Table 1.1 List of some of the more common -omics designations in addition to those discussed in the text

Term	Object of study
Pathogenomics	Genomes of human pathogens: analysis of genes involved in disease generation
Pharmacogenomics	Genomic responses to drugs, analysis of expression profiles that indicate similarity of action across compounds, analysis of genetic polymorphisms that determine a person's disposition to drug action
Toxicogenomics	Mode of action of toxic compounds, development of expression profiles that indicate similarity of toxic action across compounds
Ecotoxicogenomics	Genomic responses of organisms exposed to environmental pollution
Ionomics	All mineral nutrients and trace elements in an organism, for example using inductively coupled plasma mass spectrometry (ICP-MS)

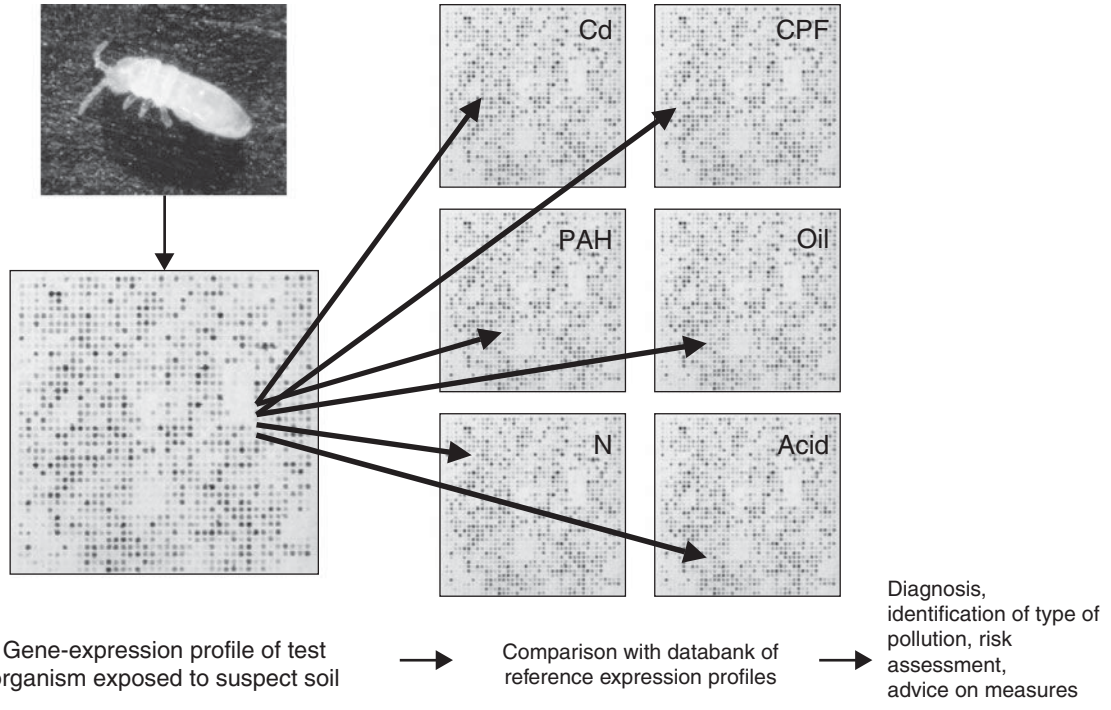


Figure 1.8 Risk assessment of soil pollution can be supported by matching the gene-expression profile of an indicator organism, generated after exposure to a suspect soil, with profiles established as a reference and known to be associated with certain types of pollution. Examples are given of soils polluted by specific substances: CPF, chlorpyrifos; PAH, polycyclic aromatic hydrocarbons.

specific microbial populations have been identified that contribute to disease suppressiveness; however, for most soils, we have little understanding of the consortium of microorganisms and the corresponding genes that are responsible for this critical function. Natural disease-suppressive soils can be regarded as a largely untapped resource for the discovery of new antagonistic microorganisms and antibiotics. We will see several examples of this in Chapter 3. Management strategies can be developed that involve selective stimulation and support of populations of antagonistic microorganisms in the rhizosphere. Genomic methods of soil diagnosis could be used as feedback on agricultural management decisions.

1.4 Genome analysis

Genomics is highly technology driven. The enormous impact of genomics research on medical and agro-technological sciences has inspired commercial life-science companies to develop innovative genomic tools at a tremendously high speed. We abstain from discussing in this book all methods relevant to ecological genomics, as this would be an impossible task. However, we foresee that two important types of technologies will stand out in the future: next-generation sequencing and microarray hybridization. The principles underlying these technologies will be discussed in the remainder of this introductory chapter, including some of the challenges in data analysis and statistics associated with them.

1.4.1 Next-generation sequencing

The field of DNA analysis was dominated for nearly three decades by the sequencing technology developed by Fred Sanger (Sanger *et al.* 1977b). '*Sanger sequencing*', as it is now called, was of crucial importance in sequencing endeavours such as the human genome project, and it earned him a second Nobel Prize in Chemistry in 1980. Until about four years ago, all genome-sequencing projects employed his sequencing principle. Sanger sequencing makes use of the fact that in a PCR the extension by DNA polymerase is terminated if a dideoxynucleotide (ddNTP) is incorporated in the sequence, rather than a normal deoxynucleotide (dNTP). The trick is to make a reaction mix including normal nucleotides and chain-terminator nucleotides in such a way that amplicons are generated that are terminated randomly at all positions of the sequence. The result of the reaction is a collection of DNAs, each with the same sequence starting at the 5′-end of the template, but each differing in length so that every position in the entire sequence is represented by a fragment that terminates at that nucleotide. Four similar reactions are conducted, each having one of four radioactively labelled ddNTPs. Separation of the amplified fragments by electrophoresis and reading the labelled bases in four different lanes allows reconstruction of the sequence of the template. A further breakthrough came from Leroy E. Hood in 1986 with the use of fluorescent labels, allowing a single sequencing reaction containing all four ddNTPs, and detection of DNA fragments using a laser. Machines were developed that could perform DNA sequence analysis completely automatically and send the read-outs to a computer (Smith *et al*. 1986).

The Sanger method is considered a 'first-generation' technology. An increasing demand for technologies delivering fast and accurate massive genome information has catalysed the development of *next-generation sequencing technologies (NGS).* Under this heading are included various alternative strategies for ultra high-throughput DNA sequencing which may be grouped into four categories: (i) microchip-based electrophoretic sequencing, (ii) sequencing by hybridization, (iii) cyclic-array sequencing, and (iv) real-time observation of single molecules (Shendure and Ji 2008).

Typically, these next-generation sequencing approaches deliver an enormous number of reads, from which the sequence of a genome has to be assembled. The *assembly phase* makes use of computer programs that in principle compare every sequence read with every other read to identify overlaps. From these overlaps contigs and scaffolds can be constructed. Since most genomic positions will be sequenced more than once, there will always be matches between reads that provide a basis for defining overlaps. The *coverage* or *sequencing depth* is

the average number of times that a base is sequenced. Traditional sequencing methods ('whole-genome shotgun sequencing') used a sequencing depth of 5 to 10 times, but NGS methods can often reach a coverage of 60 or more. A *contig* is a DNA sequence derived from contiguous reads that overlap sufficiently to align them. A *scaffold* is a larger piece of assembled DNA, consisting of contigs aligned to each other, forming a single joint sequence. Large scaffolds of more than 100 kbp can be mapped onto a chromosomal position. This is done by matching the scaffold sequence with previously cloned genes whose position is known from linkage maps, or by developing probes that can indicate a scaffold's position by fluorescent *in situ* hybridization (FISH) in chromosome preparations. Still, considerable effort has to be devoted to verification and accuracy checks, including additional end sequencing of large insert libraries to close the gaps.

Microchip-based electrophoretic sequencing is based on a conventional Sanger method carried out in a microfabricated device, thereby reducing processing time and reagent consumption. The method remains the best option in terms of read length and accuracy of base calling. The concept of *sequencing by hybridization* is the application of differential hybridization of labelled nucleic acids to an oligonucleotide array to precisely locate and identify variant positions. This method is often called *resequencing*, because a full genome sequence is necessary to generate the microarray. Indeed, it is unclear if this method can be used for *de novo* sequencing, although it is a very cost-effective technology. Another disadvantage is that repetitive sequences throughout the genome are subject to cross-hybridization and will be very difficult to interrogate with probes.

The most common next-generation platforms are based on *cycle-array sequencing*. This principle has been realized in several commercial products. Table 1.2 summarizes the specifications of the most commonly used sequencers. Although these platforms are quite diverse in the design of the array and the biochemistry applied, they share conceptually similar workflows. Random fragmentation of template DNA is taken as a starting point, followed by the ligation of common adaptor sequences. Subsequently, clonally clustered PCR amplicons are generated to achieve millions of sequencing features. In this way PCR amplicons derived from any given single molecule end up spatially clustered. Finally, the sequencing process itself consists of alternating cycles of biochemistry by an enzyme (DNA polymerase or restriction enzymes) and imaging-based base calling.

The main advantages of the cyclic-array strategy over Sanger sequencing are depicted in Fig. 1.9. They include (i) the *in vitro* construction of a library without the need for cloning fragments into plasmids, transformation into *E. coli*, and subsequent colony picking; (ii) an extremely high degree of parallelism that cannot even be achieved by micro-chip based Sanger sequencing; and (iii) immobilization of the array-based features on a solid surface, allowing enzymatic manipulation in a single reaction volume.

Still, cycle-array sequencing confers two major disadvantages: short read lengths and low raw accuracy. On average, base-calls are ten times less accurate than in the Sanger method. The limitation of short sequence read length creates major challenges for algorithms that assemble the sequence information into contigs.

Two commercially available next-generation sequencing platforms are most often used in ecological genomics, namely 454 sequencing from Roche, and the Illumina Genome analyser IIx (also called Solexa sequencing).

The 454 system was the first commercially available next-generation sequencing platform (Margulies *et al.* 2005). A library of short DNA fragments is created, and adaptors bearing universal priming sites are ligated to the fragment ends. After ligation the DNA is denatured into single strands and captured on beads under conditions that favour the binding of a single DNA molecule to a single bead. Each fragment on a bead is amplified by means of emulsion PCR, such that a single bead represents a clonal sequencing feature consisting of several millions of identical bound copies (Fig. 1.10a).

Subsequently, millions of amplified and enriched emPCR beads are ready for loading on a picotiter plate. A picotiter plate consists of wells with a diam-

Table 1.2 Specifications of the most commonly used next-generation sequencing platforms

Platform	Template preparation	Chemistry	Read length (bp)*	Run time (hrs)	Gbp per run	Application
Roche 454 GS FLX Titanium	Fragmentation MP emPCR	Pyrosequencing	400	10	0.5	*De novo* sequencing, exon capture barcode sequencing
Illumina Genome Analyser IIx	Fragmentation MP emPCR	Reversible terminator	100–150	96	18–35	Variant discovery, resequencing, *de novo* sequencing, gene discovery, RNA-seq
Helicos BioSciences Heliscope	Fragmentation MP single molecule	Reversible terminator	35	192	37	Resequencing, RNA-seq
Applied Biosystems SOLiD	Fragmentation MP emPCR	Cleavable probes SBL	50	168	30–50	Resequencing, variant discovery

* Average read-length after Metzker (2010). MP, mate-pair; emPCR, emulsion PCR; SBL, sequencing by ligation

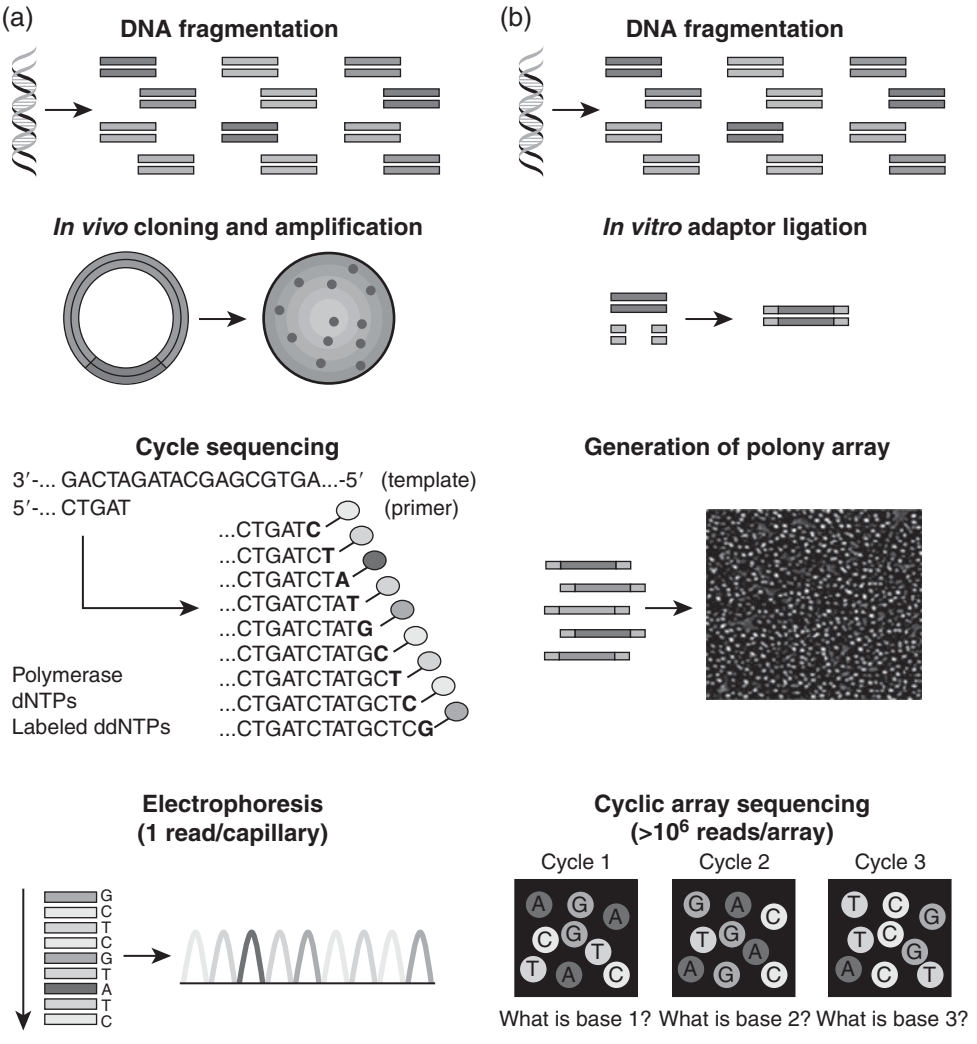


Figure 1.9 Work flow comparison of conventional versus cycle-array sequencing. a) During Sanger sequencing nucleic acids are being fragmented and cloned into plasmids. For each sequence a colony is being picked and this will generate a ladder of ddNTP-terminated dye-labelled products. Subsequently, base-pair calling is achieved by separation of fragments and detection of each of four colored dyes. b) Sequencing by cycle-array sequencing also starts with fragmentation of nucleic acids. However, cloning is omitted and instead common adaptors are ligated upon which fragments are subjected to an amplification protocol that results in an array of millions of spatially immobilized PCR colonies also called *polonies*. Since all polonies are immobilized on a solid surface, a single microlitre-scale reagent volume can manipulate millions of polonies in parallel. Consequently, imaging-based detection of a fluorescent label incorporated with every extension will acquire sequence information from all polonies in parallel. The iterative process of enzymatic incorporation of fluorescently labelled nucleotides followed by imaging are applied to build up a contiguous sequencing read per polony. From Shendure and Ji (2008), reproduced by permission of Nature Publishing Group.

eter that can only hold a single emPCR bead. Such a plate can be occupied by as many as one million individual beads and so a single 454 run can yield over 500 million base-pairs of sequence when taking an average read length of 500 bp. The actual sequencing goes via the pyrosequencing method, a real-time DNA sequencing technique described already in the 1990s by Ronaghi and coworkers (Ronaghi *et al*. 1996). The method relies on the detection of pyrophosphate (PPi), a product of DNA

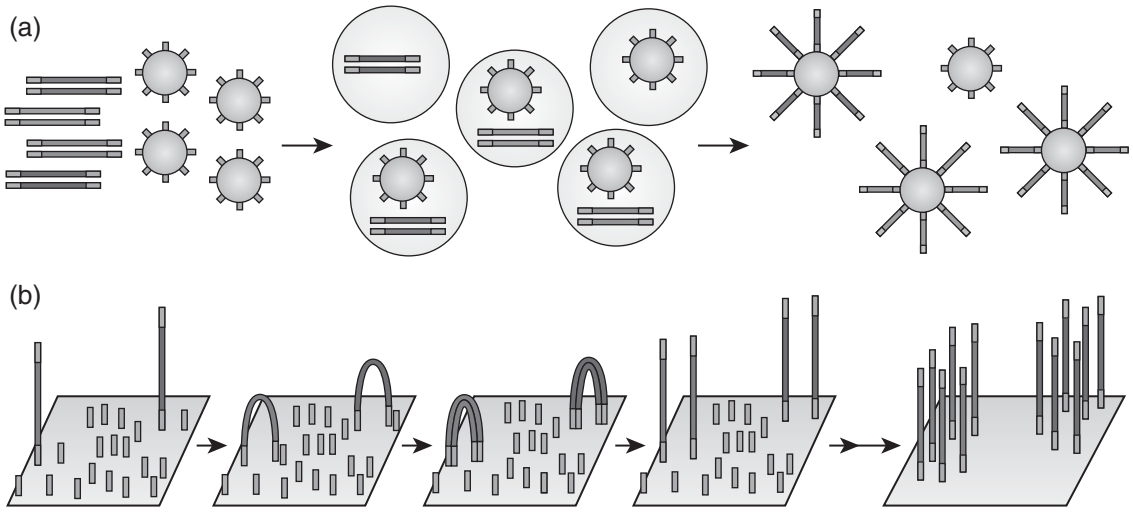


Figure 1.10 Clonal amplification of sequences. a) Emulsion PCR in 454 sequencing. b) Bridge PCR applied in Solexa sequencing. Reproduced from Shendure and Ji (2008), by permission of Nature Publishing Group.

polymerase activity. The Ppi which is formed in the polymerase reaction is converted to ATP by ATP sulfurylase. The ATP is then used by luciferase to catalyse the production of oxyluciferin and light from ATP and luciferin (Fig. 1.11c). Finally, a charge-coupled device (CCD) measures the ATP-driven burst of light that is proportional to the rate of nucleotide incorporation. In this way the sequencing synthesis is monitored in real time (Fig. 1.11d). After removing the sequencing reagents a new cycle is started with another nucleotide. Across multiple cycles (A-C-G-T-A-C-G-T…etc.) the pattern of incorporation of nucleotides detected per well reveals the sequence of the immobilized template.

At a certain point the sequencing may run asynchronously across the wells due to the presence of homopolymer nucleotide stretches such as AAAA or CCCC. The absence of a termination moiety on the displayed nucleotides will cause multiple consecutive incorporations of a base-pair, in such a way that the length of the homopolymer stretch needs to be deduced from the intensity of the light signal. This is the major limitation of 454 sequencing, because deducing homopolymer length from light output intensity is prone to a high error rate. Consequently, the dominant error type is insertion-deletion calling. However, 454 sequencing is the method of choice regarding *de novo* sequence analysis, due to the generation of relatively long read-length, up to 600 bp. In recent years 454 sequencing has become very popular in ecological genomics, since it a good choice when working with organisms about which little prior genomic information is available.

Solexa sequencing (using the Illumina Genome analyser IIx) deviates from 454 sequencing in several aspects. First, library construction is being performed with smaller fragments (150 bp) than in the 454 approach. Second, after adaptor ligation sequencing features are not amplified by emulsion PCR (as in the 454 procedure), but via bridge PCR on a solid phase (Fedurco *et al.* 2006). In this method high-density forward and reverse primers complementary to adaptor sequences in the library are covalently bound to a glass slide. The ratio of the primers to template concentration defines the surface density of the amplified clusters (Fig. 1.10b). Such an amplification procedure can yield 100–200 million sequencing features consisting of around 1000 identical amplicons. Needless to say, the parallel sequencing capacity of this platform is almost beyond imagination. With a current sequence read

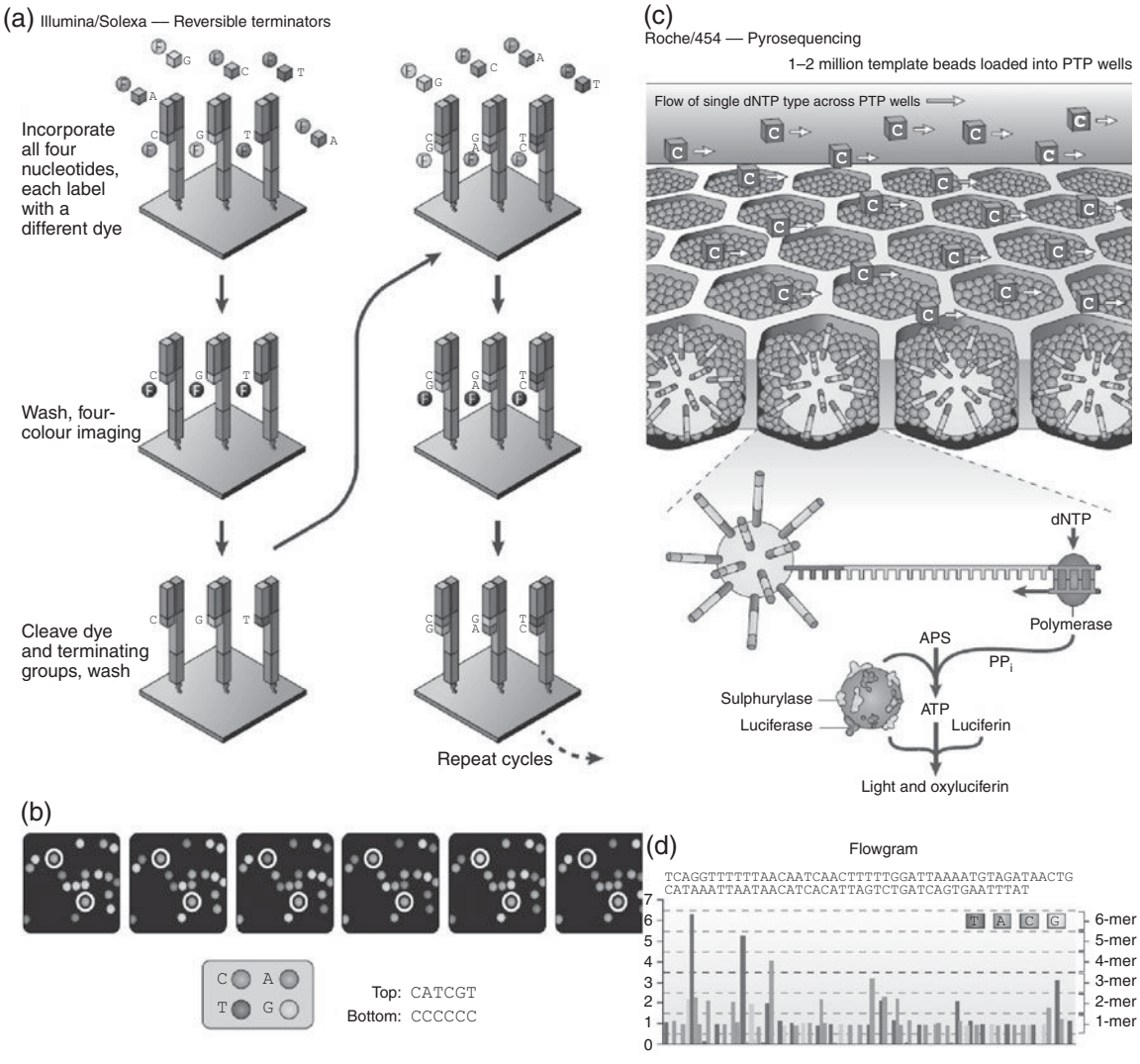


Figure 1.11 Principle of two cycle-array sequencing techniques. a) four-colour cyclic reversible termination (CRT) with 3′-O-azidomethyl reversible chemistry (Illumina/Solexa), b) four-colour imaging of two clones, c) pyrosequencing (454 sequencing), d) light generated by the cascade of enzymatic steps is recorded as a series of peaks, called a flowgram, from which the sequence is inferred. Reproduced from Metzker (2010), by permission of Nature Publishing Group.

length of around 120 bp per sequencing feature, this platform generates 10–30 Gbp of sequencing data in a single run (Metzker 2010).

The actual sequencing chemistry of Solexa sequencing is fundamentally different from 454 sequencing. It makes use of the principle of *cyclic reversible termination* (CRT, Fig. 1.11a), whereby reversible terminators are applied in a cyclic method of nucleotide incorporation. A sequencing cycle includes the addition of four modified deoxynucleotide molecules. Each nucleotide bears one of four fluorescent labels together with a reversible termination moiety at the 3′ hydroxy position. DNA polymerase catalyses the simultaneous incorpora-

tion of the four nucleotides on the primed template of each sequencing feature. The termination moiety prevents incorporation of multiple nucleotides. After washing away any surplus nucleotides, incorporated nucleotides are recording by taking an image (Fig. 1.11b). Finally, the termination moiety is cleaved off together with the fluorescent dye, followed by an additional washing step before a new CRT round is initiated.

Currently, the major drawback of CRT is the relatively high proportion of erroneous incorporation of nucleotides resulting in apparent substitutions in the sequence, in particular when the previous incorporation was a G (Dohm *et al.* 2008). This is due to the fact that the applied DNA polymerases do not efficiently incorporate nucleotides with 3′-blocked terminators. This is also the reason why sequence read lengths with the Solexa sequencing method are short (70–120 bp per sequence feature). However, it is envisaged that better DNA polymerases (with a more efficient incorporation of 3′ modified nucleotides) will be designed in the near future.

Presently, the Illumina/Solexa Genome analyser dominates the NGS market, because of its capability of ultra-deep sequencing; *de novo* Solexa sequencing of a transcriptome usually results in 900 times coverage of each transcript. This means that each unique sequence is sequenced about 900 times. This also implies that Solexa sequencing is an extremely powerful approach for SNP detection (see Section 6.2).

A fundamental limitation in DNA sequencing is that none of the technologies allows uninterrupted reading of the whole genome; the read lengths are also smaller than most transcripts. To cover a whole genome by short sequence reads, sequencing is started from a large number of random positions in the genome, leading to a collection of many reads from which the genome sequence has to be assembled. This strategy is designated as *de novo whole-genome shotgun* (*WGS*) sequencing. The purpose of assembly software is to reconstruct the target genome by aligning overlapping sequence reads. Assembly is relatively straightforward when sequence stretches are unique. This is mostly the case for transcripts representing a transcriptome. However, most genomes contain large regions of repetitive DNA, represented by, for instance, microsatellites (2–10 bp tandem repeat units), minisatellites (more than 10 bp tandem repeat units), and multiple insertions of mobile elements such as transposons and retroviruses. Also, transcriptomes consisting of gene families present difficulties because such families consist of duplicated genes with a high degree of sequence similarity due to their evolutionary origin from one single parental gene. Highly sophisticated assembly algorithms have been developed to address these issues.

A second challenge is computational power. Typically for WGS data, every sequence read needs to be compared to every other read to identify overlap. *De novo* assembly of NGS data presents a special problem due to the short sequence read lengths and the tremendous volume of data. With the read length of around 800 bp generated by the Sanger sequencing method, all previous assembly programs were based on direct overlap of reads and resolving this overlap into long colinear DNA stretches. This method is therefore called *overlap-layout-consensus* (*OLC*) assembly. At the time, Gene Myers developed the Celera assembler based on OLC, which was first successfully applied in WGS assembly of the *Drosophila* genome (Myers *et al.* 2000). It also turned out to be successful later in assembling the human genome. However, such a 'read-centric' method is computationally unfeasible for NGS due to the short sequence lengths and extremely high coverage generated by these methods. Instead, new methods have been developed that rely on transformation of the sequence reads into graphs of very small, fixed-length subsequences called *k-mers*. In this representation, nodes are located for sequences with *k* lengths; an edge is set between two *k*-mers if they represent an adjacent sequence (Fig. 1.12). The graphical representation is also called a 'De Bruijn graph' after the Dutch mathematician, Nicolaas Govert de Bruijn, who described this graph theory in 1946.

The De Bruijn graph approach with *k*-mers is particularly suitable for short read lengths as generated by Solexa sequencing because it does not store individual sequences and compresses highly redundant sequence information. Still, assembling NGS data is very challenging. Miller *et al.* (2010) formulated

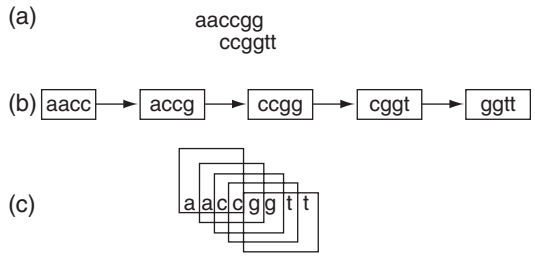


Figure 1.12 Two sequence reads represented by a *k*-mer graph. a) Two reads have an error-free overlap, b) a single *k*-mer graph (k = 4) represents both reads; the alignment is a by-product of graph construction, c) the simple path through the graph implies a colinear contig from which a consensus sequence can easily deduced. After Miller *et al*. (2010), reproduced by permission of Elsevier.

three factors that complicate the use of De Bruijn graph approach in *de novo* assembly: i) since DNA is double-stranded *k*-mer graphs need to implement nodes and edges for both strands; ii) repeats greater than the *k*-mer length may collapse inside the graph and prevent a colinear solution into a single contig sequence; and iii) DNA sequences may contain *palindromes*, a sequence bearing its own reverse complement, that will induce the graphical path to fold back on its previous stretch. This is, however, solved in the assembly algorithm VELVET by allowing only odd-sized *k*-mers which cannot match a reverse complement stretch (Zerbino and Birney 2008). The bioinformatics community is developing an ever increasing number of packages for NGS assembly based on the principles described above. A comprehensive overview is given by Miller *et al*. (2010). However, future NGS platforms will almost certainly produce even larger data volumes against lower costs. Therefore, developers of assembly algorithms will remain to be challenged to deal with still increasing datasets.

1.4.2 Applications of next-generation sequencing in ecological genomics

There are essentially five different areas of application of next-generation sequencing: (i) *de novo* assembly of genomes, (ii) resequencing of genomes, (iii) sequencing environmental DNA (metagenomes), (iv) gene expression profiling (sequencing transcriptomes), and (v) detection of genomic polymorphisms, especially SNPs. In this section we discuss some of these applications.

One of the first papers reporting on 454 sequencing in an ecological model species was that published by Toth *et al.* (2007). In communities of the primitive eusocial wasp, *Polistes metricus*, workers take care of their sibs instead of reproducing themselves. In an attempt to gain more information on the molecular mechanism of this social behaviour Toth *et al*. (2007) sequenced the transcriptome of *Polistes* brain tissue from distinct behavioural groups. Most of the 454 sequence reads were mapped against the honey bee, another hymenopteran insect. The authors showed that gene expression in the brain of workers was similar to gene expression in the brain of foundresses (females that exhibit both reproductive and maternal care). However, striking differences between foundresses and workers were observed in insulin signalling, suggesting that eusociality not only involves reproductive pathways but also nutritional pathways.

Another remarkable example of 454 sequencing in ecological genomics is the study of Vera *et al.* (2008). These authors successfully performed a *de novo* assembly (without the help of a closely related reference sequence) of a good part of the transcriptome of the Glanville fritillary butterfly (*Melitaea cinxiae*), an ecological model species with complex metapopulation dynamics. The transcriptomic information for this butterfly is bound to help in understanding the dynamics of local extinction and recolonization of endangered species in fragmented landscapes.

The giant panda genome (*Ailuropoda melanoleuca*) was the first complex genome to be solely sequenced with the Illumina platform (Li *et al.* 2010). Interest in sequencing the panda genome was triggered by the fact that the ecological traits of this species, such as a low fecundity and a restricted diet primarily made up of bamboo, make it very vulnerable to human population expansion and habitat destruction. At present, only 2500–3000 individuals live in small mountainous habitats of Western China. Furthermore, the giant panda occupies a unique phylogenetic position since it is the living species most close

to the common ancestor of the bear subfamily Ursinae, diverging about 17.9–22.1 million years ago (Krause *et al.* 2008).

Li and coworkers set out to sequence the genome of a 3-year old female panda using only the Illumina Genome analyser platform. First, 37 sequence libraries were generated with varying fragment length ranging from 150 bp to 10 kbp. Subjection of these libraries to the Illumina platform yielded a total of 176 Gbp of high-quality sequence with an average read length of 52 bases. The SOAPdenovo algorithm was used for *de novo* assembly, which applies the De Bruijn graph method. *De novo* assembly used an all-against-all comparison of the generated *k*-mers, which required the algorithm to run on a supercomputer with no less than 521 Giga bytes of random access memory (RAM). The assembly was performed in a stepwise manner, starting with the reads originating from the short sequence-size libraries (150–500 bp sheared fragments). This assembly yielded contigs with an average length of 1.5 kbp. By applying a paired-end approach during sequencing the authors were able to identify pairs of 50 bp flanking sequences that were physically linked in a single DNA fragment. This information turned out to be essential in assembling the relatively small contigs into larger scaffolds. After four rounds of scaffold construction the authors were able to construct a sequence of 2.25 Gbp with 56 times coverage. Thus, they were able to retrieve 94% of the panda's genome, given that the estimated size of the giant panda genome is 2.4 Gb, organized in 20 pairs of autosomes and one pair of sex chromosomes (2n = 42).

Gene prediction analysis was performed by aligning the panda sequence with 20 000 open reading frames (ORFs) from the human and dog genome. In this way, over 19 000 gene loci could be identified in the panda genome. A survey for genes potentially involved in food selection and digestion typically yielded genes associated with a carnivorous digestion system (protease, lactase, invertase, maltase). Genes involved in cellulose breakdown (e.g. endoglucanase, exoglucanase) could not be identified, suggesting that the vegetarian lifestyle, a derived property of the panda unique within the Ursinae, is not dictated by the panda's own genome, but is probably more dependent on the gut microbiome. In summary, the paper of Li *et al*. (2010) clearly showed that NGS sequencing can be used to accurately assemble *de novo* a large complex genome in a cost-effective way.

Another very important development due to NGS is that it allows the resequencing of genomes of established genomic models. Two large projects are of particular interest. The 1000 Genomes is an international consortium established in 2008 to create the most detailed and medically useful information on genetic variation in the human genome through the resequencing of 1000 human genomes. It is supported by the three largest genome centres in the world (Welcome Trust Sanger Institute, UK; Beijing Genomics Institute Shenzhen, China; National Human Genome Research Institute, USA). The 1001 Genomes project was established by ten genome centres in Europe and the US in 2009 to catalogue genetic variation in *Arabidopsis thaliana* through NGS analysis of 1001 different *Arabidopsis* strains. Undoubtedly, such projects will greatly accelerate the discovery of variants that affect quantitative traits (see Chapter 6).

Resequencing genomes using NGS is now also penetrating molecular ecological research. This is nicely illustrated by a study conducted by Turner *et al*. (2010). They investigated local adaptation of *Arabidopsis lyrata* to serpentine soils, which are characterized by a high heavy metal content and low calcium-to-magnesium ratio. *Arabidopsis lyrata* is a perennial representative from the family Brassicaceae and closely related to the genomic model *A. thaliana*. In contrast to *A. thaliana*, however, *A. lyrata* is a self-incompatible outbreeding species. It thrives on various soil substrates including serpentine soils. The question was raised which properties of the genome allow these plants to survive under soil conditions that are toxic to most other plant species.

Two populations from serpentine soils were resequenced by NGS and compared to resequenced pools of *A. lyrata* originating from two neighbouring granite soils. Over eight million SNPs were surveyed and analysed for any association with

soil type; 96 SNPs showed allele frequency differences of greater than 80% between the contrasting soil types. The gene sets associated with differential SNPs were associated with functions such as metal detoxification, calcium metabolism, and magnesium transport. Additional sequencing of a Scottish serpentine *A. lyrata* showed similar associations between these functional loci and serpentine soil, suggesting parallel adaptive evolution of metal tolerance. In conclusion, NGS technology provided an excellent opportunity to study the genomic basis of ecological adaptation and to identify variants associated with functional properties.

NGS can also be used in sequencing genomes of microorganisms in the environment. We will see in Chapter 3 that microbial communities in the environment are still largely unknown, and many microbes are characterized only by bits of their DNA sequence without being brought into culture. A recent study by Nolte *et al*. (2010) described the successful application of 454 sequencing to identify the turnover dynamics and seasonal abundance of protist communities residing in an Austrian lake. The authors showed that the standing diversity of protists is relatively stable and the total sampled taxon richness was about 4000 genotypes. A rarefaction analysis of the data showed that by using NGS the identification of taxa was saturated, which is a major achievement. The authors also analysed relative abundance of three *Spumella* groups (Chrysophyceae) based on their habitat preferences. The first group comprised soil-derived species preferring a temperate moderate warm climate, the second group thrived at cold-temperate sites, and the third group preferred freshwater isolates from warm habitats. Surprisingly, taxa associated with warmer habitats prevailed in the summer protist community, whereas taxa adapted to cooler conditions were abundant in colder months.

These examples are only scraping the surface of the enormous array of possibilities open to ecologists using NGS to address their ecological questions. More examples will be discussed in the course of this book and many more are expected to be published in the near future.

1.4.3 Microarray-based transcription profiling

The well established microarray technology is very relevant for ecological genomics and we feel it will play an important role in the field in the near future due to the fact that analysis tools have been refined to a high qualitative level (Kammenga *et al.* 2007). The aim of transcription profiling is to develop a complete overview of all the genes in a genome that are upregulated or downregulated in response to some factor of interest, in comparison with a designated reference expression.

To develop a genome-wide image of gene expression, all the RNAs present in the cell or tissue at a certain time point need to be assessed, as well as their relative abundance. Obviously, the transcriptome in a cell includes not only mRNAs but also rRNAs, tRNAs, and non-coding iRNAs, however, gene expression is usually focused on mRNA; this RNA is also called *poly(A)-RNA*, because mRNA of eukaryotes is often isolated by taking advantage of the characteristic poly(A) tail on each messenger. The RNA pool is transcribed to complementary DNA (cDNA) with the enzyme *reverse transcriptase* and, depending on the platform, sometimes again to RNA (which is then called *cRNA*) before hybridization.

The most common way to get an overview of all transcripts is to generate two labelled samples, one from the challenged organism or cell and one from the reference, and hybridize these to a large number of DNA sequences, immobilized on a coated glass plate in an ordered array. The hybridization is usually done after mixing the pools, so homologous RNAs from either sample compete for binding to the same probe. Other designs, in which the RNAs do not compete with each other, are also possible as detailed below. The principle of using probes fixed to a glass plate to interrogate the transcriptome was first described by Schena *et al*. (1995). Such a device, denoted a *microarray*, quickly became the cornerstone of transcription profiling in genomics. The term microarray is contrasted with *macroarray*, which uses essentially the same technology but is conducted on a nylon membrane with fewer spots and using radioactive labels rather than fluorescence. The spots on a microarray, which are packed

very close together, are called *probes* or sometimes *features*. The sample of transcripts that is interrogated by the array is called the *target*. Microarray hybridization is sometimes called *reverse hybridization* because the probe represents the immobile phase. This contrasts with traditional Southern blotting, in which the target is immobilized (usually on a membrane) and the (labelled) probe is mobile.

The hybridizations on a microarray are visualized by labelling the transcripts—for example the reference with a green fluorescent label and the test sample with a red label—and the array is scanned using a laser. Each spot on the array that holds a DNA sequence of sufficient homology to one of the sequences in the collection of mRNAs will be labelled red or green if the abundance of that messenger was greater or less in the tested sample compared with the reference. In this way, genes that are upregulated relative to the reference will be labelled with one colour and genes that are downregulated will be labelled with the other. For all spots that correspond to messengers whose abundance in the sample is similar to the reference, equal amounts of label will bind and the spot will be perceived as yellow. Spots for which there is no corresponding messenger will not be labelled at all.

The use of a microarray for transcription profiling can be illustrated by a classical paper by DeRisi *et al*. (1997). These investigators were interested in the changes in physiology that occur in yeast (*S. cerevisiae*) when cells switch from anaerobic growth (fermentation), using glucose as a carbon source and producing ethanol, to aerobic growth, which occurs when glucose is depleted after fast growth and the cells turn to respiring ethanol. This *diauxic shift* is accompanied by a fundamental reorganization of the cell's physiology, in which the expression of many genes changes. Since at the time the microarray had just been introduced (Schena *et al*. 1995) and the genome sequence of yeast completed (Goffeau *et al*. 1996), the authors could apply one of the first genomics approaches to transcription profiling. DeRisi *et al*. (1997) amplified ORFs from the yeast genome using PCRs with primers specific for each gene. These DNA fragments, approximately 6400, corresponding to nearly all the genes in the yeast genome, were printed onto glass slides using a robotic device. In the experiment, cells were harvested at different stages in the growth phase and mRNA was isolated. These mRNAs were reverse-transcribed to cDNAs and labelled with a red carbocyanin (Cy5) label, and this was mixed with a green (Cy3) labelled sample prepared from cells harvested directly after the start of the experiment. After hybridization and washing the array was scanned using a fluorescent confocal microscope and images generated like that shown in Fig. 1.13.

Microarray scans are often represented in colour, where a red spot indicates upregulation, a green spot indicates downregulation, and a yellow spot constant expression; however, the fluorescence intensities of the two carbocyanine labels are not actually 'seen' in colour by the detector, which just records digital values for intensities at different

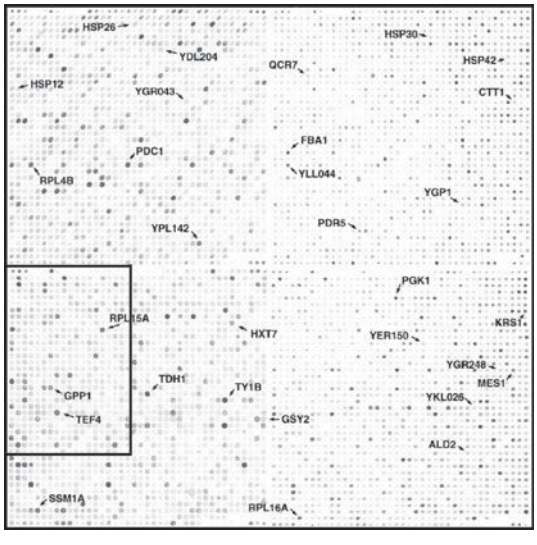


Figure 1.13 Scan of a microarray as used by DeRisi *et al*. (1997). Each spot represents one of 6500 DNA sequences from the yeast genome, printed on a glass plate of 18 × 18 mm. The array was used to detect gene-expression changes occurring when the cells had depleted glucose from the medium after 9.5 hours and shifted to aerobic metabolism. In this negative greyscale picture, hybridization intensities are indicated by different shades of grey. In full-colour images upregulation and downregulation are indicated in red and green, *resp.* Some of the spots are indicated by codes designating the genes. The rectangle on the left indicates a section of the microarray singled out for further analysis (not shown). Reprinted with permission from DeRisi *et al*. (1997). Copyright 1997 AAAS.

wavelengths. Therefore a colour representation of a microarray image is also referred to as having *false colours*. In Fig. 1.13 and elsewhere in this book we represent microarray scans in grey tones, although the difference between upregulation and downregulation is not always discernible in this way. The reader should consult the original publications to obtain complete views of microarray images.

DeRisi *et al*. (1997) discussed their data in terms of the biochemical pathways for carbon and energy metabolism (pentose phosphate pathway, glycolysis, Krebs cycle, glycoxylate cycle). They were able to show that the changes in mRNA abundance during diauxic shift could be mapped onto this biochemical framework and indicated that a significant redirection of metabolites was taking place. It was also apparent that groups of genes responded in a coordinated fashion and seemed to be regulated by a common factor. This observation triggered further studies into the mechanisms by which transcription factors activate sets of genes collectively.

Following the pioneering work of DeRisi *et al*. (1997) early microarrays were almost exclusively manufactured using pen tip deposition of cDNA probes. These arrays were also called spotted arrays. However, this technology posed several quality problems, such as non-uniform feature size, irregular feature shape, and variation in deposited DNA quantity across slides. Therefore, during the course of the 1990s, new techniques for the fabrication of microarrays were introduced which led to so-called high-density oligonucleotide arrays or *gene chips* (Shiu and Borevitz 2008). In these systems, short sequences of 20–25 nucleotides are synthesized directly (*in situ*) on the substrate using photolithography or chemical-based deprotection. In such arrays, all the probes are synthetic and genomic sequence information alone is sufficient to construct the array. The probes can be made in such a way that the most unique part of a transcript is represented; this does not need to be the coding part of the gene; in fact, often the 3′- or 5′-UTR of the mRNA is used, since this is more specific for a transcript than the coding region. It is also common to use multiple probes designed to hybridize to different regions of the same transcript. In addition, each probe is supplemented by a control sequence that has one mismatched base in the middle of the sequence. Because the target cDNA should only bind to the perfect probe and not to the mismatch probe, an accurate measure of transcript abundance can be obtained by subtracting the match signal from the mismatch signal. Hybridization with these arrays is non-competitive; only a single sample of target is applied to each array and comparison with a reference is made across chips. Figure 1.14 provides an overview of the two approaches to transcription profiling (Schulze and Downward 2001).

Oligonucleotide microarrays allow greater coverage of the genome and may be more repeatable across laboratories. The extremely large number of synthetic oligonucleotides that may be packed on an array (some technologies allow for 6.25 million features on a single glass slide) allow each gene to be represented by multiple probes (Nuwaysir *et al*. 2002). This is useful when aiming to detect polymorphisms and different splice variants of the same gene, and it also allows for within-array replication of gene expression.

Recently, arrays have been designed to cover not only the transcriptome, but the full genome sequence of a species. Such arrays are called *tiling arrays*, because the probes are arranged in a colinear contiguous way and are often partially overlapping like the tiles of a roof (Mockler and Ecker 2005, Fig. 1.15). Tiling arrays have become more popular since next-generation sequencing technology has accelerated the number of completed genome sequences. They achieve a truly whole-genome view on the data and are extremely useful in novel gene discovery, analysis of splicing variants, mapping of regulatory DNA motifs, whole-genome methylation analysis, polymorphism analysis, and genome resequencing. Today, the costs of commercially available microarrays are no longer prohibitive to their use in academic research laboratories. In fact, due to their superior quality and associated bioinformatic analysis tools, commercial arrays have almost replaced spotted arrays and have also become the main platform of choice for ecological genomics.

A puzzling discovery associated with the use of tiling arrays in gene expression studies is that transcription can be detected in intergenic regions of the genome. Intergenic regions are areas that have

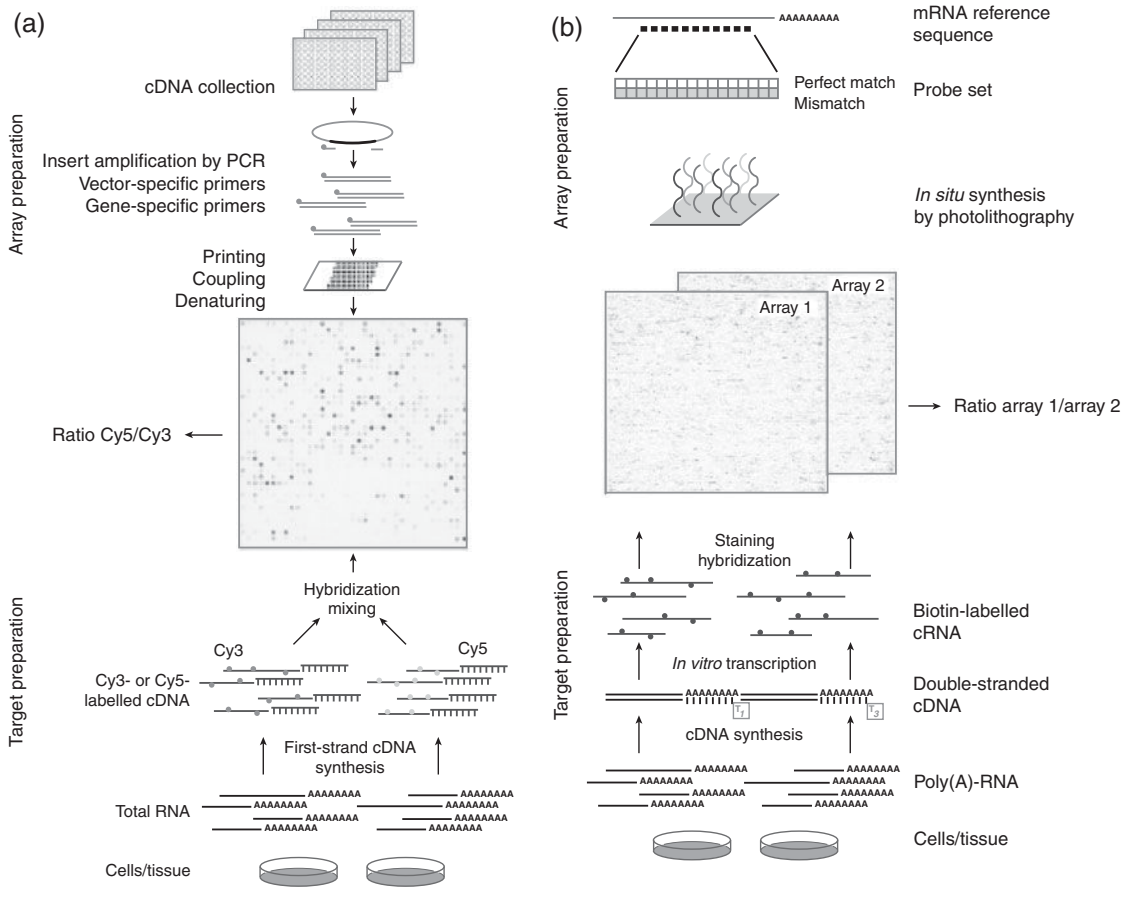


Figure 1.14 Overview of transcription-profiling approaches applied for (a) spotted cDNA microarrays and (b) high-density oligonucleotide microarrays. In the cDNA approach (a), the microarray is printed from a cDNA library (often developed from ESTs), and the probes are isolated by PCR amplification using primers specific to the gene or the vector. PCR products are printed using a high-precision robotic device. The target sample is obtained from RNA isolated from two groups of cells or tissues. This is used for reverse transcription in the presence of nucleotides with fluorescent labels (e.g. Cy3 and Cy5). The two samples are mixed in a hybridization buffer and brought into contact with the array under competitive conditions, such that for each probe the most abundant transcript (with either a Cy3 or a Cy5 label) binds most to the array. Scanning of the array with wavelengths corresponding to the excitation spectra of the two dyes will provide a picture of the transcript abundance profile in the sample, relative to the reference. In the high-density oligonucleotide microarray approach (b), sequence information of the transcriptome of a genomic model species is used to develop probes of 25 nucleotides which provide a perfect match with a unique part of each transcript. In addition, control probes are developed with a single base mismatch. Each transcript sequence is represented by 16–20 different probes. The probes are synthesized *in situ* and fixed directly to the array. The target sample is prepared from poly(A)-RNA isolated from the cells or tissues of interest, which is reverse-transcribed to generate double-stranded cDNA, using a poly(dT) with a transcriptional start site for T7 RNA polymerase; this polymerase is then used to synthesize cRNA using biotinylated nucleotides. The two pools of amplified RNAs are hybridized with two different arrays and target binding is detected by staining with a fluorescent dye coupled to streptavidin, which recognizes the biotin label. Signal intensities of probe sets of the two different arrays are used to calculate relative transcript abundance. After Schulze and Downward (2001), reproduced by permission of Nature Publishing Group.

never been identified as a coding sequence and do not end up as translated peptide. These transcriptional active regions raise many questions. Are these features truly transcribed? Is the synthesis of such RNA regulated? What functional relevance do these transcripts exert and what is their mode of action? A study by Stolc *et al.* (2004) emphasized the fact that transcriptional activity is much more abundant

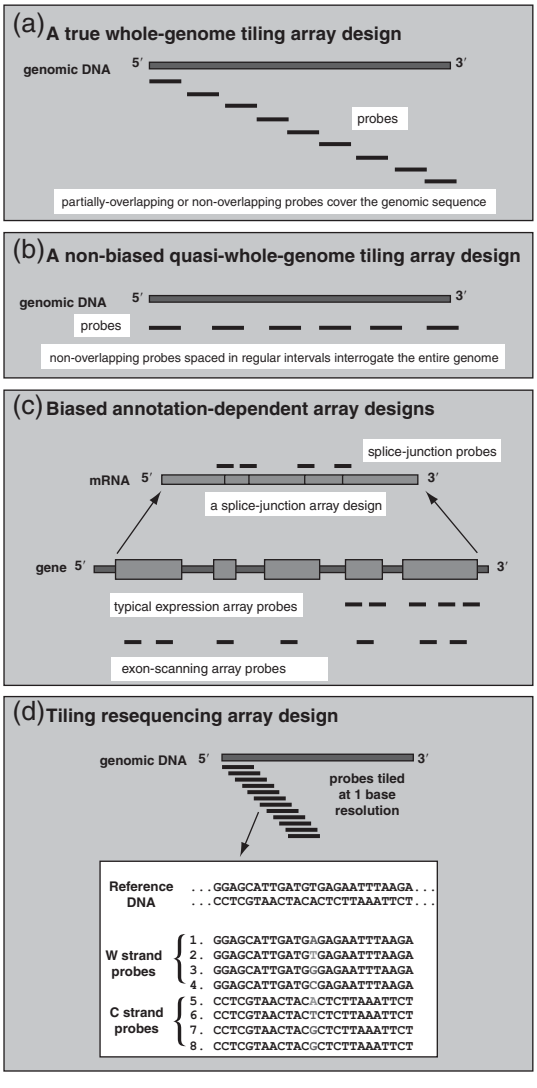


Figure 1.15 Comparison of different whole-genome array designs. (a), (b) unbiased whole-genome tiling array either with or without partial overlap of the adjacent probes; (c) biased whole genome arrays such as typical expression arrays, exon-scanning arrays, or splice-junction arrays contain only probes from known and predicted features of the genome; (d) tiling resequencing arrays where each nucleotide is represented by a set of eight oligonucleotide probes. Reproduced from Mockler and Ecker (2005), by permission of Elsevier.

in the genome than we previously estimated from cDNA/EST sequencing and computational open reading frame (ORF) prediction. These authors applied a tiling array to study gene expression during different life stages of *Drosophila melanogaster*. The transcriptional activity of 97% of all previously predicted and annotated open reading frames was confirmed. Astonishingly, over 40% of so-called *non-exon probes* (*NEP*, situated in intergenic or intron regions) showed significant transcriptional activity. About 15% of these transcriptionally active NEPs showed significant differential expression at some developmental stage, indicating that these regions are developmentally regulated. The authors suggested that the NEPs represent genes that failed to be annotated, or maybe non-discovered exons of annotated genes. With the exact transcriptional position mapped to the genome, they re-evaluated gene prediction with the Genescan gene identification software and were able to corroborate over 50% of NEPs either as unknown genes or as part of existing open reading frames. The functional significance of such transcripts needs to be systematically characterized; RNA interference techniques will probably be of great value.

1.4.4 Processing of microarray data and statistical analysis

The raw data from microarray experiments often come in the form of 16-bit TIFF images. The data-processing phase is conducted by software supplied with microarray scanners, and is not discussed extensively in this book. We just mention three important steps that have to be performed.

Image analysis. The image-processing software must recognize the spots, determine their boundaries, measure the signal coming from each spot, compare it with a local background, and assign the result to the correct probe.

Estimating expression ratios. In competitive hybridization two signals, corresponding to two different fluorescent dyes, are read from each spot. When using gene chips, the reference and query samples are hybridized to different arrays, but in this case spots corresponding to the same probe are also compared. The quotient of the two signals is defined as the expression ratio T_i for each gene i:

$$T_i = \frac{Q_i}{R_i}$$

where Q_i is the signal for gene *i* in the query sample and R_i is the signal for the same gene in the reference sample. In spotted cDNA microarrays, *Q* derives from the Cy5 signal (red) and *R* from the Cy3 signal (green), or vice versa. The expression ratios are always logarithmically transformed and generally the logarithm to the base 2 is applied. This transformation results in a quantity known as *fold change* or *fold regulation* (FR). As a result of the transformation, *FR* takes a value of 0 if there is no change, a value of 1 if there is a twofold increase, and a value of –1 if there is a twofold decrease in expression. A fourfold increase results in FR = 2 and a fourfold decrease in FR = -2. So:

$$FR_1 = {}^2\log T_1 = \frac{{}^{10}\log Q_1 - {}^{10}\log R_1}{{}^{10}\log 2}$$

It should be noted that the use of expression ratios has a disadvantage, namely that information about the actual signal intensity is lost. So genes that are expressed weakly in both the *Q* and *R* samples are treated similarly to genes with an overall strong expression, if the relative up- or downregulation is the same.

Normalization. For various reasons, the FR values obtained cannot be directly compared across replicate measurements or different experiments. The most common source of variation is the use of different amounts of RNA as the starting material from which the target sample was prepared. One popular approach is to consider a regression of log Q_i against log R_i. If the initial amount of RNA is exactly the same for the *Q* and *R* samples, and if the labelling and detection efficiencies are also identical, such a plot would show a cluster of points around a straight line through the origin with slope 1 (of course, individual genes will lie apart from the line due to up- or downregulation). However, often the data do not fall exactly on such a straight line and show a curving trend (Fig. 1.16, left-hand panel). Since the interest lies in deviations from the diagonal, insight may be increased by rotating the plot by 45° and rescaling the axes. This can be done by plotting $M = {}^2\log Q - {}^2\log R$ over $A = {}^2\log Q + {}^2\log R$, and in such a plot *M* should be independent of *A* (Fig. 1.16, right-hand panel). If this is not the case, one can correct the data by subtracting a quantity *c*, which depends on *A* and is defined as the difference between the local deviation of the data from a horizontal line. This correction term is estimated for each value of *A* by means of *local weighted regression* (loess) (Smyth *et al*. 2003).

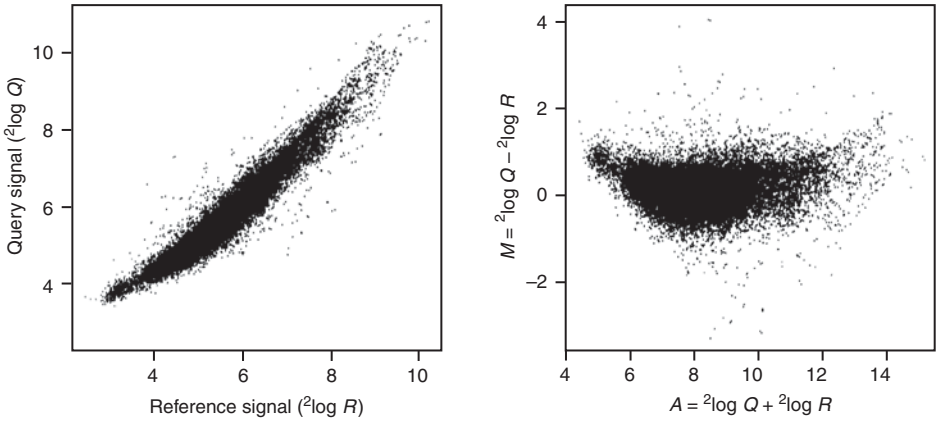


Figure 1.16 (a) Scatterplot of two signal intensities, log Q_i (from the query sample, e.g. fluorescence from Cy5) and log R_i (from the reference sample, e.g. Cy3) over all genes *i* in a microarray expression analysis. Ideally one expects that the average of the data falls along a straight line with slope 1 through the origin. The slightly curved shape indicates intensity-dependent bias. (b) When plotted as $M = {}^2\log Q$—$^2\log R$ over $A = {}^2\log Q + {}^2\log R$, the bias is visualized more clearly. The data can be corrected by subtracting a term which depends on *A* and is estimated by local weighted regression (loess). From Smyth *et al*. (2003) by permission of Humana Press.

After data processing, a file results that can be viewed as a matrix with one very long column in which the FR values of all genes are noted. Usually one experiment involves several samples and these are taken together in one gene-expression matrix with a number of columns, for example different points in time, different physiological states of the organism, or different environments from which the RNA was isolated.

Because such gene-expression matrices are valuable resources for statistical analysis, and a single investigator is often not able to exploit all possible data-analysis techniques, the data are often published on the Internet. This allows other researchers to compare expression profiles across studies, in much the same way as a genomic database is consulted by different people. Brazma *et al*. (2001) considered the requirements that such data matrices should have in order to be valuable to the research community. They developed a standard known as minimum information about a microarray experiment (*MIAME*). This standard stipulates that publication of gene-expression matrices should be accompanied by details about:

- experimental design (conditions, doses, replication, quality-control measures, etc.);
- array design (type of array, source of probes, clone identifications, etc.);
- hybridization conditions (buffer, washing procedure, hybridization time, temperature, etc.);
- measurements (quantification, normalization, data filtering, etc.).

As part of MIAME it is required that the raw image file from the scanner, with relevant scanning parameters, is supplied with the publication.

There are various considerations of experiment design when working with microarrays. Yang and Speed (2002) asked the question of what dangers could occur when not paying adequate attention to design issues? They envisaged two extreme situations. A carelessly designed experiment might be entirely satisfactory but very inefficient in material use and therefore not cost-efficient. At the other extreme, too much emphasis on cost-efficiency might limit or even compromise the interpretation of results. For instance, the researcher must decide whether samples should be hybridized independently from each other on separate arrays (*single channel detection*, e.g. biotin labelled cRNA on an Affymetrix platform) or whether two samples labelled with two different dyes should be competitively hybridized on a single array (*dual channel detection*, e.g. cyanin 3 and cyanin 5 labelled samples on an Agilent platform). Dual channel detection seems much more cost effective, but raises additional technical challenges due to differential dye intensities originating from the differential oxidation kinetics of Cy3 and Cy5 labels. Because microarray slides and dye labels are expensive, the number of replications and the way in which experimental groups are compared to each other are usually carefully optimized. The following issues are important to take into account:

Technical replication. The noise stemming from technical imperfections is taken on board by the use of more than one feature for each gene on a single array, and hybridization of the same RNA sample to different arrays. Also the *dye-swap*, applied in competitive hybridization, is an aspect of technical replication.

Biological replication. The greatest source of noise is usually due to variation across samples. It is essential for any microarray experiment that two or more samples from different replicated RNA isolations are included in the design.

Hybridization scheme. Three common designs are applied: (i) each sample may be tested against the same reference, for example RNA samples from plants grown under different conditions are compared to a large RNA pool from plants grown under standard conditions; (ii) all samples are tested against all others—this is called a *saturated design*; and (iii) each sample is tested against several, but not all, other samples—this is called a *loop design*. The design can be portrayed as a circle of pairwise connected samples, arranged in such a way that the most important contrasts are included. Designs (i) and (ii) are intuitively attractive, but often require a too large a number of arrays. In many cases a loop design is very effective.

The statistical analysis of microarray data is intimately linked to the design of the experiment. Suppose we consider a single factor experiment,

where we would like to compare differential gene expression in three different tissue samples (A, B, and C), so that all pairwise comparisons are of equal importance. Figure 1.17 shows three design choices where R is a common reference. If mRNA yield is very low, design 1 may be the only option. However, if more replicates are available designs II and III become feasible. Note that indirect comparison (design II) requires twice as many arrays. In contrast, direct comparison (design III, loop design) requires fewer arrays and provides more precise comparisons (lowest average variance) among the samples. Therefore loop designs have become very popular. However, estimates of variance may become imprecise with larger sample sizes, because some paths connecting successive treatments become very long. More relevant approaches should consider which comparisons are of most interest and seek a design that gives the highest precision to most relevant comparisons.

Specialized statistical software has been developed that is completely targetted to the analysis of microarray data. Due to the rapid developments within this field a detailed overview of such packages falls beyond the scope of this book. We discuss a few general issues that hold irrespective of the design.

One approach to analysing microarray data is to apply an *analysis of variance* (ANOVA) to each gene separately (Jin *et al*. 2001). For example, if two factors are considered (such as sex and age in a study of gene expression in *Drosophila*), the analysis takes the form of a two-way analysis of variance applied to each gene, for which the model is (Sokal and Rohlf 1995):

$$FR_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where FR_{ijk} is the FR value of this gene expected for replicate k of sex i at age j, μ is the overall mean expression, α_i is the effect of factor A (e.g. sex) at level i, β_j is the effect of factor B (e.g. age) at level j, $(\alpha\beta)_{ij}$ is the interactive effect of A and B together, and ε_{ijk} is the error term. There must be at least two replicate measurements for each combination of factors for such an analysis to be useful. The ANOVAs will lead to a few thousand F tests, one for each gene, and these tests will indicate the genes that are dif-

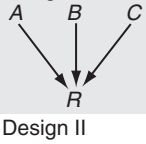
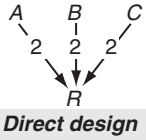
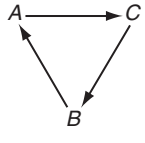
Design choices	Number of slides	Units of material (number of samples)	Average variance
Indirect designs			
Design I 	3	$A = B = C = 1$	2.00
Design II 	6	$A = B = C = 2$	1.00
Direct design			
Design III 	3	$A = B = C = 2$	0.67

Figure 1.17 Designs for single-factor experiments. A, B, and C different treatments; R common reference. Reproduced from Yang and Speed (2002), by permission of Nature Publishing Group.

ferential between sex, that change with age, and that change with age in a sex-dependent manner.

One should be careful though in attaching an absolute value to the outcome of a significance test. Because tests are applied to the same larger database, the *P* values for the *F* tests may not represent the true type I errors. To avoid taking too many false-positive results on board, a significance level of 10^{-4} or even smaller is chosen. Specialized methods are being developed for controlling the false-positive discovery rate (Cheng *et al*. 2004). In addition, the significance of the effect must be balanced against the magnitude of the effect. In judging magnitudes of effects in transcription profiling, up- or downregulation by a factor of 2 is usually chosen as a threshold (FR value of greater than 1 or smaller than –1). Not all effects exceeding this threshold will be significant and not all significant effects will exceed the criterion of twofold change. To illustrate this, Jin *et al*. (2001) developed a presentation known as a *volcano plot*, in which the apparent *P* values are plotted as a function of the FR value, both on a logarithmic scale (Fig. 1.18). Such a plot shows that some genes with a highly significant effect do not fulfil the criterion of twofold change, while on the other hand there are also genes that do fulfil the criterion of twofold change but are not significant. The plot may help to identify these different groups of genes, and focus further research on the most promising among them.

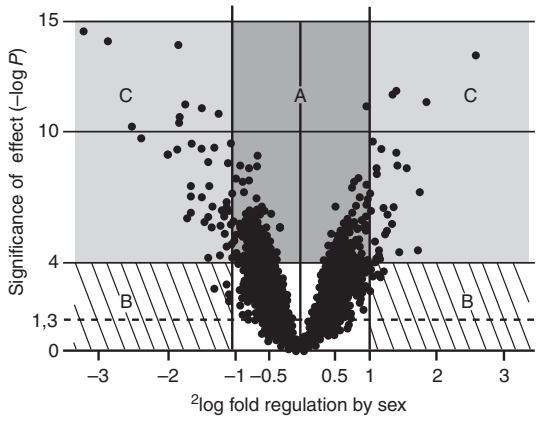


Figure 1.18 Volcano plot showing the relationship between the *P* values of *F* tests applied to gene expressions of *D. melanogaster* as a function of sex, and the magnitude of the effect. Each point is a separate gene. The horizontal axis gives the fold regulation value of the gene with respect to sex (genes to the left are downregulated in males compared to females, genes to the right are upregulated). The ^{2}log value of the gene-expression change is plotted, so a value of 1 implies a twofold upregulation and a value of –1 a twofold downregulation. The vertical axis gives the ^{10}log-transformed reciprocal *P* value, so the line at 1.3 corresponds to $P = 0.05$ and the line at 4 to $P = 10^{-4}$. The latter value was taken by Jin *et al*. (2001) as the 'preset experiment-wise false positive acceptance level'. There are several genes that are significant but have a fold regulation lower than a factor of 2 (region A) and there are also genes that are more than twofold up- or downregulated, but are not significant (regions B). Only genes in regions C are beyond doubt. After Jin *et al*. (2001), by permission of Nature Publishing Group.

In addition to analysis of variance, various non-parametric methods have been proposed, some of the most popular being *SAM*, Statistical Analysis of Microarrays (see www-stat.stanford.edu/~tibs/SAM; Tusher *et al*. 2001) and *RDAM*, Rank Difference Analysis of Microarrays (Martin *et al*. 2004). In non-parametric methods the raw signals from a microarray scan are replaced by ranks and variation among replicates by rank differences.

Gene-by-gene analysis of microarray data, be it parametric (ANOVA) or non-parametric (rank-based), has the disadvantage that it does not consider the correlation structure among the expressions of different genes, otherwise than by controlling the false-discovery rate. Actually, the gene-expression matrix is of a multivariate nature. Each gene may considered a *case* (also called an object) and the samples represent a number of measurements made on that case. Various multivariate statistical techniques can be applied to data organized in this way and one of the most common is some form of *hierarchical clustering*. The logic of clustering is evident from the fact that groups of genes will have similar expression patterns over samples, because they are induced by the same environmental conditions or regulated by the same transcription factors. The most common clustering algorithm to apply to microarray data comes from Eisen *et al*. (1998).

Clustering starts by developing a matrix of pairwise distances between the genes. There are different ways to calculate distances, one of the most straightforward being *Euclidean distance*. Suppose we are considering genes *A* and *B*, and we have observations on gene expression of a_i for gene *A* and

b_i for gene B in sample i, then the Euclidean distance D_{Eucl} between the genes is:

$$D_{Eucl}(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_j)^2}$$

where n is the number of samples. This distance measure is calculated for each pair of genes resulting in a *distance matrix*, which is input to the clustering algorithm. The Euclidean distance is not the only way of defining distances between genes. Other measures are the Minkowski distance, Manhattan distance, and Hamming distance. In addition, the clustering may be based on a similarity measure, such as Pearson correlation, rather than distance. The reader is referred to Causton *et al*. (2003) and textbooks of multivariate statistical analysis for more information.

The object of clustering analysis is to develop a dendrogram that groups together genes with similar expression patterns. There are several principles that can be applied to achieve clustering. In an influential paper on gene-expression data analysis, Eisen *et al*. (1998) applied the so-called *average linkage method*. In this method a computer algorithm screens the matrix of pairwise distances for the smallest value. Then a node is defined between these genes and gene-expression values are calculated for the node by averaging over the two genes involved. The distance matrix is then updated and a new smallest distance is identified. The procedure is repeated until g–1 nodes have been made, where g is the number of genes. Software packages such as that developed by Eisen *et al*. (1998) not only provide a computational procedure but also a pictorial presentation of the clustered gene-expression pattern; each gene is qualified by a colour code, where red is used for upregulated expression and green for downregulated expression. This representation is aptly called a '*heat map*'; we will meet many of them in this book.

It turns out that for gene-expression matrices only a few principal components can actually explain the data. This is due to the fact that many gene expressions are correlated with each other, but also that the number of samples is usually much smaller than the number of genes. So the amount of information in a gene-expression matrix is not as large as it would seem from the long list of genes. By plotting the data in a two-dimensional graph a lot of the information is captured already. Holter *et al*. (2000) express the situation by analogy with spectral analysis of music: 'the complex "music of the genes" is orchestrated through a few underlying patterns, and the genes in a micro-array comprise a set of identically tuned strings'. Kim *et al*. (2001) and Kim and Tidor (2003) showed that dimensionality reduction leads to a limited number of building blocks that can be scaled and added together in various combinations to best reconstruct the data. The authors developed a system in which the genome of *C. elegans* is dissected into 43 *expression mountains*, where each mountain represents a functional group of 5–1818 genes with a high internal correlation of expression.

Another popular approach to analysing gene-expression data is the technique of *self-organizing maps*. A self-organizing map starts with a set of nodes with a simple topology, for example in a two-dimensional grid, and continues by applying an algorithm in which the nodes are mapped onto the highly dimensional space of the data points in such a way that the distance from data points to nodes is as small as possible (Tamayo *et al*. 1999). The number of nodes needs to be specified beforehand, and so the procedure is equivalent to placing a fixed number of flags in a landscape in such a way that the scatter of data points is organized around the flags in an optimal way. The algorithm can dissect a complex dataset in a limited number of gene clusters where each cluster has a characteristic response over treatments.

As an example of a full statistical analysis of a microarray study we discuss the analysis of genome-wide gene-expression changes during the cell cycle of *A. thaliana* (Menges *et al*. 2002, 2003). An oligonucleotide microarray of the genome of *Arabidopsis* (Affymetrix ATH1) was used to look at gene-expression changes during a synchronized cell culture (Fig. 1.19). Plant cells were grown in suspension under continuous agitation and were synchronized by release from a chemical inhibitor. Gene expressions were first analysed by PCA, and this demonstrated that the first principal axis already explained

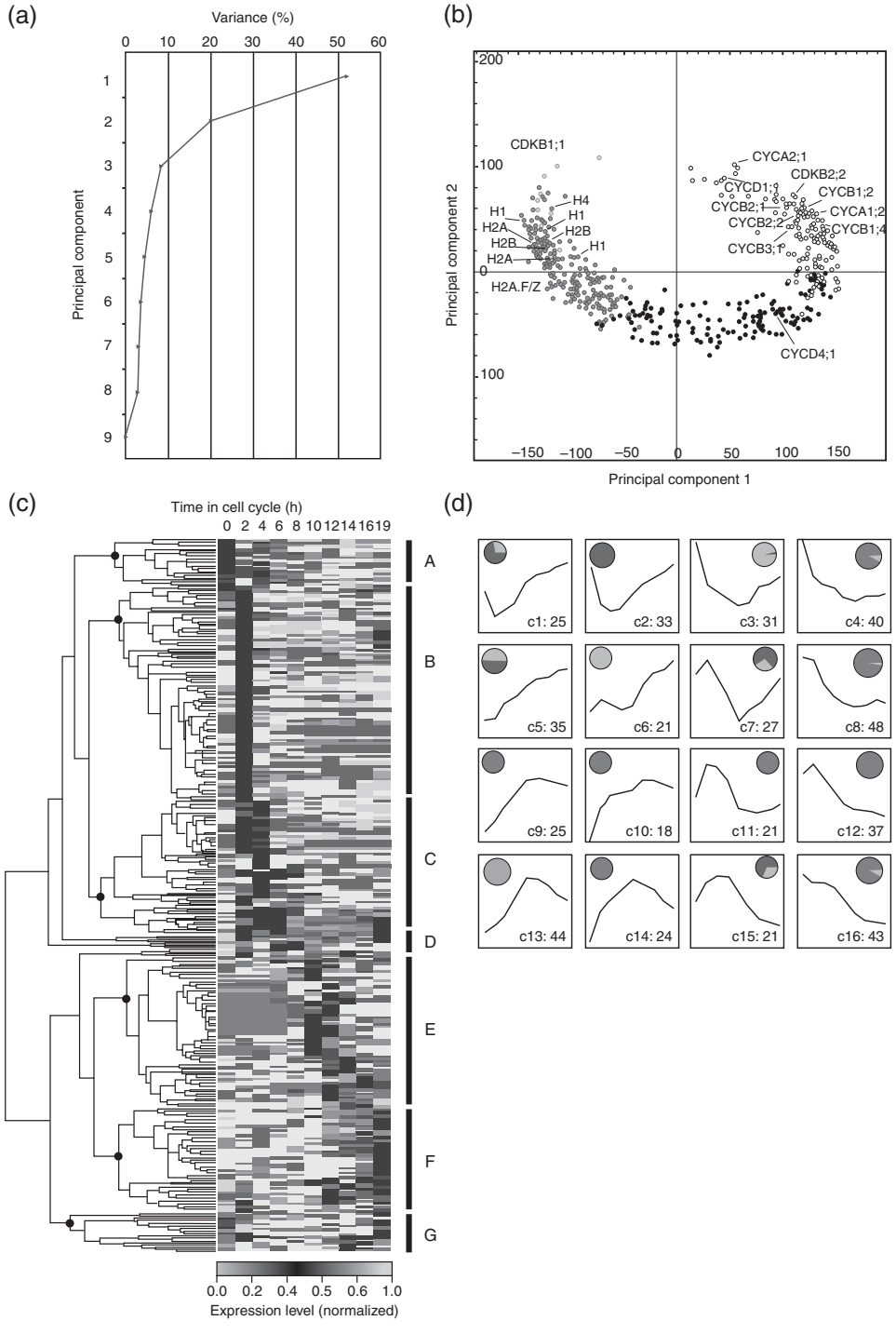


Figure 1.19 Results of statistical analysis of gene expression changes during the cell cycle of a synchronized cell suspension culture of *Arabidopsis*. (a) PCA showing the percentage of explained variance by successive axes. (b) PCA biplot of the data in a state space reduced to two dimensions. (c) Result of hierarchical clustering, showing gene expressions as a 'heat diagram', where different shades of grey indicate upregulation and downregulation. Note the oblique pattern of dark fields indicating peak expressions at different times in the cell cycle. (d) Sixteen patterns of gene expression change over the cell cycle, identified by self-organizing map analysis. The number of genes involved in each pattern (numbered c1–c16) is indicated. The small pie charts show the composition of each cluster with respect to four gene classes characteristic for certain cell-cycle phases (S, M, G_1, G_2). After Menges *et al*. (2002), with permission from Springer.

more than 50% of the variation, while the second added another 20%. In dimensionality-reduced representation two main clusters appeared (Fig. 1.19b). Hierarchical cluster analysis (Fig. 1.19c) confirmed the existence of two main clusters, each of which could be subdivided further. In total, 16 different clusters were recognized, each representing a particular pattern of expression during the cell cycle (Fig. 1.19d). For example, cluster 5 (35 genes) represents genes that increase continuously in expression during the experiment, whereas cluster 16 (43 genes) represents genes with decreasing expression over time. Using different ways of synchronizing the cell culture, Menges *et al*. (2003) finally developed a refined list of 1082 genes (out of some 14 000 genes that could be detected under the conditions chosen) that were cell-cycle regulated, of which 371 have no known function at the moment.

Statistical analyses will usually generate a list of genes ranked or classified according to differential expression over certain phenotypes or treatments. This is, however, not sufficient when considering the underlying biology. In addition to statistical analysis we need a biological analysis which takes the identity and physiological function of the genes into account. The most successful initiative of systematic biological description of biological processes, molecular functions, and cellular components is the *Gene Ontology (GO)* Project (http://www.geneontology.org). Gene ontology essentially contains two components: defined biological terms plus structural relationships between them (*GO ontology*) and the association between genes and ontology terms (*GO annotation*).

Conceptually, GO ontologies are cell-biological knowledge domains that are organized in a *directed acyclic graph (DAG)* so that the GO terms are the nodes and their relationships are represented by edges between nodes. An important feature of DAG is that parent–offspring relationships are defined. The parent terms are more general biological entities when compared to their child terms (Fig. 1.20). The GO annotation is an essential link of a gene to a set of GO terms and is generated by a curator or automatically via bioinformatic predictive methods. When a gene is annotated to a specific term it will automatically inherit its parent terms. Therefore, it is extremely important that any path from term (e.g. vesicle fusion) to roots (e.g. biological process) is biologically accurate. Revisions are needed if this turns out to be incorrect. If the curation is correct the annotation process becomes flexible and very powerful: a gene annotated to vesicle fusion can be directly retrieved with this function, but also through its parental terms.

A GO term always includes an *evidence code* to indicate how the biological information was

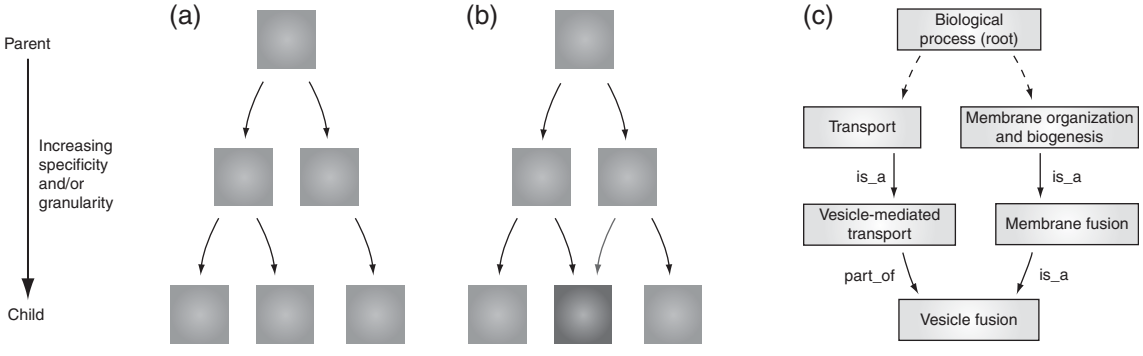


Figure 1.20 Simple trees as compared to acyclic graphs. Boxes represent nodes and arrows edges. a) Simple tree representation: each child has only one parent. b) Directed acyclic graph representation (DAG), where a child can have one or more parents (dark colored node). Example of a child node for vesicle fusion exerting multiple parents (dark box). Reproduced from Rhee *et al.* (2008), by permission of Nature Publishing Group.

Table 1.3 Analysis of gene expression change in a soil-dwelling invertebrate, *Folsomia candida*, in response to desiccation for 8-174 h. The table provides Gene Ontology (GO) terms that were over-represented in lists of differentially expressed transcripts and their *P*-values derived from the gene enrichment method described by Alexa *et al*. (2006). GO ID, unique identifier for Gene Ontology term; GO term, biological process associated with the GO ID; # in GO term, the number of genes present on the array with identical GO ID; # Significant, number of gene probes that show a significant transcriptional response. Reproduced from Timmermans *et al*. (2009), by permission of the Royal Entomological Society.

Exposure period (h)	GO ID	GO Term	P-value	# in GO Term	# Significant
8	GO:0006144	purine base metabolic process	0.008	2	2
	GO:0006189	'*de novo*' IMP biosynthetic process	0.008	2	2
	GO:0006635	fatty acid beta-oxidation	0.024	3	2
	GO:0006716	juvenile hormone metabolic process	0.008	2	2
	GO:0007417	central nervous system development	0.008	21	5
	GO:0008643	carbohydrate transport	0.045	6	3
27	GO:0006078	pentose-phosphate shunt	0.006	3	3
53	GO:0006007	glucose catabolic process	0.0035	9	8
	GO:0007279	pole cell formation	0.043	4	3
	GO:0007613	memory	0.043	6	4
	GO:0008356	asymmetric cell division	0.005	7	5
	GO:0035151	regulation of tube size, open tracheal system	0.043	4	3
174	GO:0006030	chitin metabolic process	0.03	16	10
	GO:0006119	oxidative phosphorylation	0.0005	35	45
	GO:0006959	humoral immune response	0.042	11	8
	GO:0008356	asymmetric cell division	0.042	7	7

retrieved. The codes range from evidence obtained from direct assays ('inferred from direct assay', IDA) to evidence solely based on bioinformatics ('inferred from electronic annotation', IEA). For instance in *Saccharomyces cerevisiae* over 85% of all genes have GO terms inferred from direct assays. In the human genome only 28% of the genes have GO terms assigned through direct assays (Rhee *et al*. 2008). Published annotation analyses should therefore cite the source of the data and be careful about the choosing a certain analysis method.

We expect that GO will become increasingly important in data analysis and functional prediction. Recently, methods have been developed to identify relevant biological processes or functions from gene expression data by assessing the statistical significance of predefined functional gene groups such as GO. Gene set enrichment methods analyse the position of genes in a specific functional group among an ordered list of genes. It is believed that a biological process is more relevant when its gene set members are among the top ranked genes derived from the initial analysis, such as statistical significance of differential expression.

All GO assessments rely on calculating a score for over-representation; for example a specific treatment may result in the induction of a certain set of mRNAs all belonging to a specific biochemical pathway. This will lead to over-representation of GO-terms associated with that pathway. The over-representation score is based on a statistical significance value. For instance, Alexa *et al*. (2006) applied the Fisher exact test to verify whether a GO term was enriched among a gene list. The total number of genes with a single GO term was compared to the number of genes with that particular GO term occurring in the list of significantly regulated genes. A Bonferroni adjustment is then applied to the resulting p-values to correct for multiple testing.

As a practical example we consider the study of Timmermans *et al*. (2009). They used an oligonucleotide microarray platform to elucidate the molecular mechanism of drought tolerance in a species of soil-dwelling invertebrate, *Folsomia candida*. Transcriptional changes were studied in animals exposed to desiccation conditions after various exposure periods. For each exposure period they were able to generate an ordered list of significantly regulated genes. The number of genes in these lists increased from 505 after 8 hours to 2116 after 172 hours. Subsequently, the lists were subjected to the GO term enrichment method developed by Alexa *et al*. (2006). Based on this analysis the authors concluded that GO terms associated with carbohydrate transport, sugar catabolism, and cuticle maintenance were important biological processes involved in combating desiccation stress (Table 1.3). Interestingly, Bayley and Holmstrup (1999) had shown that *F. candida* becomes hyperosmotic during desiccation by accumulating glucose and myoinositol, allowing water vapour to be extracted from the environment even under desiccating conditions. Thus, the transcriptomic data supported previous physiological observations and provided additional mechanistic insight into the remarkable drought tolerance of this soil invertebrate.

CHAPTER 2

Comparing genomes

In this chapter we will deal with the first step that usually follows completion of a genome sequence, which is to inspect the genome for its general properties, such as number of genes, size of gene families, mobile elements distribution, and so on, and compare it with other species, related or unrelated. Genetic model species will be the starting point for our comparisons, but we will explore, wherever possible, links with species from the same clade and highlight the ecological significance of models and their wild relatives.

2.1 Properties of genomes

Once the genome of a species is elucidated, partly or completely, the researcher is able to analyse the properties that characterize the genome as a whole. This is usually not done in isolation, but in comparison with other species, and therefore this part of genome science is called *comparative genomics*. The term *structural genomics*—the study of genome sequences, genetic and physical maps, and so on—is also appropriate here, as a contrast with functional genomics; however, some scientists use this term for the analysis of structural, three-dimensional, properties of proteins, using x-ray diffraction and nuclear magnetic resonance. Comparative genomics draws heavily on bioinformatics and uses computer programs to find patterns in genome sequences across species, to estimate similarities that can support the assignment of gene functions, and to develop phylogenetic trees by which the evolutionary relationships among genomes are visualized (Hedges 2002). Arguments from comparative genomics have become crucial in selecting new species for whole-genome sequencing. Sequencing efforts can be optimized to discover what information the new species can provide about functionalities in already-sequenced genomes and to determine at what evolutionary distances such species should be placed to maximize that information (Eddy 2005). Comparative genomics is now a recognized subdiscipline of genome science with dedicated handbooks (Saccone and Pesole 2003). The field is closely related to molecular phylogenetics, for which we refer the reader to Hughes (1999) and Graur and Li (2000).

Because by far the greatest number of completely sequenced genomes are from prokaryotes, comparative genomics proceeded at a tremendous pace, especially in microbiology. Comparison of genomes has shed new light on the microbial species concept and on the genes associated with specific phenotypes such as physiological functions or specific resistances (Konstantinidis and Tiedje 2005; Achtman and Wagner 2008). In addition, comparative genomics algorithms have been applied to environmental DNA sequences, comparing them to fully sequenced genomes to identify the functional capacities in the environment (Von Mering *et al.* 2007). Before discussing the genomes of microorganisms, plants, and animals in more detail, in this section we provide an overview of the properties of genomes in general.

2.1.1 Genome size

Usually the size of a genome is known with some accuracy before its complete sequence is elucidated, because genome size is essential knowledge for optimal construction of a genomic library and the design of a genome-sequencing project. Genome

size may be estimated using biochemical methods or flow cytometry and is usually expressed in picograms of the haploid genome per cell. This is easily converted to nucleotides by the general formula $G = 0.987 \times 10^9 C$, where G is genome size in base-pairs and C is genome size in pg (roughly, 1 pg is equivalent to 1000 Mbp). Genome size in pg is also known as the *C value*. Estimates of genome size are now available for nearly 5000 species of animals (www.genomesize.com; Gregory 2005) and more than 7000 species of plants (data.kew.org/cvalues). The size of a genome varies dramatically across species. Table 2.1 provides a few examples of fully sequenced genomes that illustrate this diversity.

There is an obvious increase of genome size going from viruses to prokaryotes, and further to unicellular eukaryotes and multicellular eukaryotes. This increase can be related to the increasing complexity of the cells and tissues involved. Viruses do not need their genomes to encode all the proteins that they require for their maintenance and propagation, because they exploit the molecular machinery of the host. The minimal number of genes required for an autonomous self-replicating entity was estimated by Graur and Li (2000) as 256 (based on a comparison of prokaryotic genomes) or 254 (based on a review of knockout studies). This is about half the number of genes in *Mycoplasma genitalium*, which has 468 genes, the lowest number of any independently living organism. Unicellular eukaryotes have a genome size one order of magnitude greater than the average prokaryote; however, the extremes meet each other; the yeast genome is only twice as large as the largest genome of a prokaryote sequenced so far (*Pseudomonas aeruginosa*) and even smaller than that of some Cyanobacteria. Another order of magnitude lies between unicellular and multicellular eukaryotes, but the variability among the latter group is enormous.

Table 2.1 Genome size and number of genes for organisms with completely sequenced genomes

Species	Total size of the genome (kbp)	Estimated no. of protein-encoding genes
Bacteriophage φX174	5.4	10
Mycoplasma genitalium	580	468
Methanococcus jannaschii	1665	1738
Haemophilus influenzae	1830	1743
Escherichia coli	4639	4288
Agrobacterium tumefaciens	5670	5419
Pseudomonas aeruginosa	6264	5570
Saccharomyces cerevisiae	12 610	6128
Caenorhabditis elegans	95 500	18 424
Drosophila melanogaster	123 000	13 601
Arabidopsis thaliana	125 000	25 498
Oryza sativa	466 000	50 820

A second trend in genome sizes across species is the reduction of the genome (*genome miniaturization*) in endosymbiotic organisms and parasites. In the ultimate endosymbiotic entity, the mitochondrion, many genes were lost compared to its alphaproteobacterial ancestor, partly due to deleting functions that were not necessary in the symbiotic lifestyle, and partly due to migration of genes to the nuclear genome of the host. The same holds for chloroplasts; however, chloroplasts (stemming from Cyanobacteria) have retained a significantly larger genome than mitochondria. Genome size reduction is also seen in parasites, and this may explain the very small genome of *M. genitalium*. However, it must be pointed out that there is also an opposite tendency: parasites need to have specialized proteins for adhesion to tissues of their host and to thwart the host's immune response, so they sometimes have larger genomes than expected.

The two trends noted above are about the only ones that can be observed when comparing genome sizes across species. In fact, there is a remarkable lack of correspondence between genome size and organism complexity, especially among eukaryotes. For example, the marbled lungfish, *Protopterus aethiopicus*, has more than 40 times the amount of DNA per cell than humans! Figure 2.1 provides an overview of the ranges for the various taxonomic groups.

The lack of correspondence between genome size and organism complexity has become known as the *C value paradox*. It is obvious that the enormous increase of genome size in some lineages is not accompanied by a proportional increase of the number of genes. This is already evident from Table 2.1, which shows that the number of genes tends to

reach a plateau of some 20 000–50 000 with increasing genome size. In fact, the non-genic fraction of the DNA is the main factor responsible for the *C* value paradox, such that in eukaryotes anything between 30 and 99% of the genome can consist of non-coding DNA (repetitive sequences, mobile elements, introns, intergenic spacers, etc.). Ohno (1972) introduced the term *junk DNA* to stress the fact that, according to him, non-coding DNA is a useless but mostly harmless part of the genome. This view has been opposed by many authors who have pointed out that large parts of the non-coding DNA may have a regulatory function.

How can a genome grow in size? An obvious mechanism, causing a single-step increase, is *polyploidization*. This can concern a duplication of the entire genome (global polyploidization), or parts of the genome (regional polyploidization). Many global polyploidy mutations are highly deleterious because they interfere with cell division and meiosis. In mammals, polyploidy destroys the mechanism of dosage compensation, which normally inactivates one X chromosome in the female to compensate for the lack of complementary genes on the Y chromosome in the male. Still, polyploidies that concern an even number of chromosomes, such as *tetraploidy* (doubling of the diploid genome), sometimes do not have adverse effects on the phenotype and may be a mechanism for evolutionary innovation. A famous example is the tobacco species, *Nicotiana digluta* ($2n = 72$), which is assumed to originate from genome doubling in a sterile hybrid ($2n = 36$) between *Nicotiana tabacum* ($2n = 48$) and *Nicotiana glutinosa* ($2n = 24$).

Polyploidy is especially common in angiosperms and pteridophytes, but rare in gymnosperms and

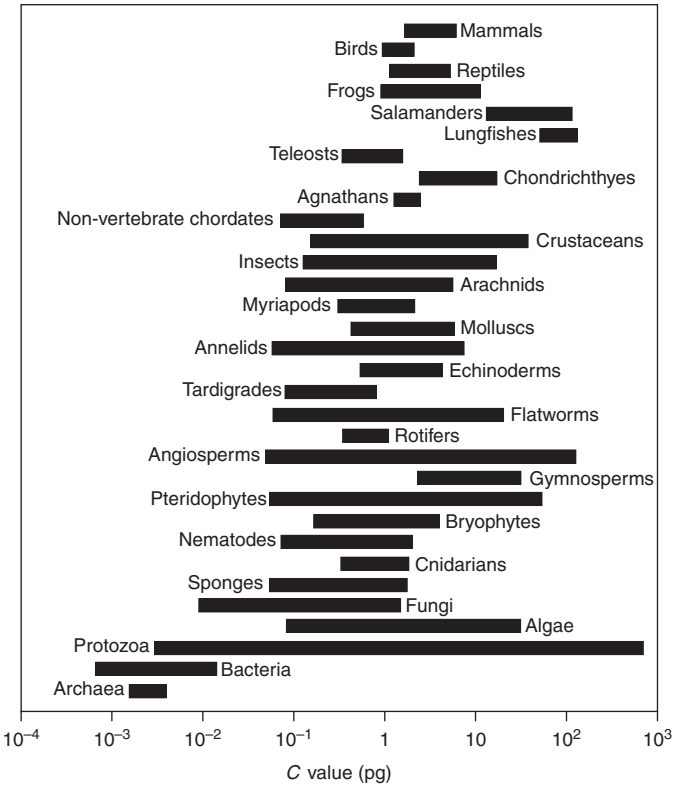


Figure 2.1 Ranges of reported genome sizes (*C* value of the haploid genome, in pg per cell) for different organism groups. From Gregory (2005), with permission from Elsevier.

bryophytes. There are also large differences in ploid levels between plant families. Polyploidy is relatively rare in animals; however, some families of fish, most notably in the Cypriniformes (carp, minnows) and several species of the amphibian order Anura (frogs and toads) are known for their relatively high occurrence of polyploid species. The only known cases of polyploidy in mammals are two species of octodontid rodent, *Tympanoctomys barrerae* and *Pipanacoctomys aureus*, which inhabit the salt deserts of Central Argentina (Gallardo *et al*. 1999, 2004; Fig. 2.2). *T. barrerae* has a *C* value of 16.8 pg, which is twice the amount of its closest relatives (8.2–7.6 pg) and out of the normal range of mammals (see Fig. 2.1).

Triploidy (arising from hybridization of a haploid and a diploid gamete) always leads to sterility, although it does not prevent growth and asexual reproduction in many plants, as shown by the triploid banana plant (*Musa acuminata*). This condition may actually be exploited by plant breeders to produce commercially attractive seedless fruits, for example of watermelon (*Citrullus lanatus*). The polyploid nature of many commercially interesting crops—for example, corn (*Zea mays*), with a genome size of 2500 Mbp, and wheat (*Triticum aestivum*), with a genome of 16 000 Mbp—is a serious challenge for the assembly of complete genome sequences of these species.

In polyploid organisms the term genome size may become ambiguous. Greilhuber *et al*. (2005) have proposed that in addition to the *C* value, which denotes the haploid genome with chromosome number n, another term, *Cx value*, should be used to denote the chromosome base number (x) in a polyploid organism. For a tetraploid, $2n = 4x$, and for a triploid, $2n = 3x$. We should also note that when a tetraploid organism comes into existence, the *C* value doubles but the *Cx* value is not changed, because the number of different chromosomes remains the same. However, when the organism evolves further, paralogous genes may differentiate in function, the two chromosomal complements undergo rearrangements and eventually the tetraploid situation will no longer be distinguishable from a normal diploid. The transition state is called *cryptopolyploidy*.

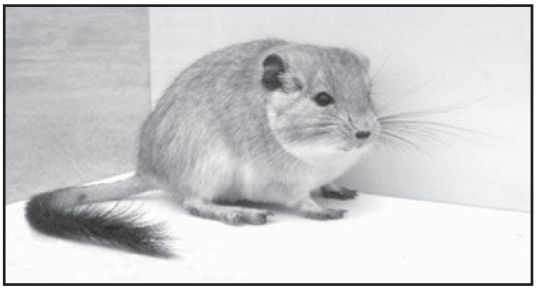


Figure 2.2 The red viscacha rat, *Tympanoctomys barrerae* (Octodontidae, Rodentia), representing a rare case of tetraploidy in mammals. The animal has 100 autosomes plus two sex chromosomes (X and Y). Courtesy of M.T. Gallardo, Universidad Austral de Chile.

According to several authors (e.g. Spring 1997) the vertebrates as a whole may be considered a cryptopolyploid lineage, originating from two successive rounds of duplication, one prior to the divergence of jawless fish (Cyclostomata) and one thereafter (Gnathostomata and higher vertebrates). This hypothesis, also known as the *2R hypothesis*, seems to be supported by a variety of evidence, including the occurrence of several genes in fourfold copies in vertebrate compared to invertebrate genomes. However, Hughes (1999) pointed out that under the 2R hypothesis one expects the phylogeny of duplicated genes in vertebrates to show four clusters arising from three splits, one dividing the tree in two branches, the other two splitting each branch. Hughes (1999) tested this expectation using several gene families, one of which, the *Notch* genes, is reproduced in Fig. 2.3. *Notch* genes encode transmembrane signalling proteins, and were first discovered in *Drosophila* where mutations in this gene are recessive and embryonic lethal due to hypertrophy of neural tissue at the expense of epidermal tissue. In the heterozygous condition these mutations cause notches at the edges of the wings, hence the name of the gene. *Drosophila* has only one *Notch* gene, but vertebrates have four. Despite the fact that the one/four comparison would support the 2R hypothesis, the phylogeny clearly does not (Fig. 2.3). In the tree, the *Notch*4 variant even clusters outside the insect *Notch* and outside vertebrate *Notch*1, 2, and 3. One gene duplication seems to have taken place prior to the split between vertebrates and

insects, whereas two successive duplications followed in the vertebrate lineage. A similar situation holds for seven other gene families. This phylogenetic evidence is in conflict with the idea of two rounds of whole-genome duplication, and the 2R hypothesis remains unsettled.

The second mechanism of genome enlargement, regional genome duplication, will lead to repetition of blocks of genes or single genes. This may occur during meiotic cell division through *unequal crossing-over*, in which case whole genes or even substantial segments of a chromosome are duplicated in one of the two gametes, while the other gamete is left without the corresponding segment.

A third mechanism by which genomes may be enlarged is *duplicative transposition*. Transposable elements are sequences that have an intrinsic capacity to change their genomic location, either by physically moving from one position to another, or by making a copy of themself that 'jumps' to another place in the genome. The mobile element carries its own genes, necessary for catalysing the transposition (including an enzyme called transposase), but in addition to these it may also carry genes or non-coding DNA that have nothing to do with the transposition itself and are thus relocated in the genome. There is an enormous variety of such mobile elements in many animal genomes. In *Drosophila*, one of the transposable elements is the well-known *P element*. It is assumed that a P element was introduced into *D. melanogaster* from another *Drosophila* species by a parasitic mite feeding on the eggs (Hoy 1994). Subsequently, P elements have been exploited by fruit fly geneticists to develop engineered P-element vectors that can introduce exogenous DNA into the germline of the fruit fly.

How could the increase of genome size in eukaryotes, and especially the dramatic proliferation of

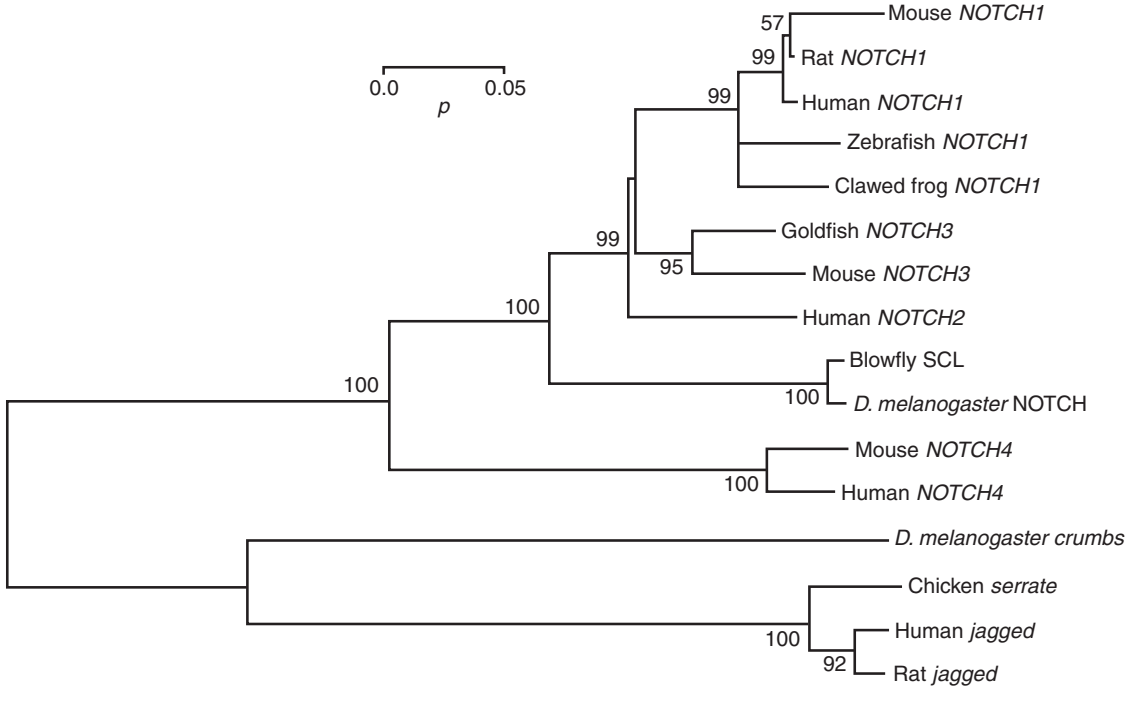


Figure 2.3 Phylogeny of the *Notch* gene family in some vertebrates and insects, constructed on the basis of proportional amino acid differences (*p*). *Serrate*, *Jagged*, and *Crumbs* were used as an outgroup to root the tree. The values along the branches are percentages of support in 1000 bootstrap samples. After Hughes (1999) by permission from Oxford University Press.

non-coding DNA, be maintained during evolution? Lynch and Conery (2003) have argued that genome complexity evolved mainly by neutral mechanisms, which apply to small populations more than large populations. The theory is supported by an obvious negative correlation between genome size and effective population size (Fig. 2.4). The *effective population size*, N_e, is defined in population genetics as the number of individuals of a species in a theoretical ideal population that would have the same magnitude of random genetic drift as an actual population has (Hartl and Clark 1997). N_e is smaller than the real population size for two reasons. First, the actual population fluctuates over time; the effective size is the harmonic mean of abundance over time and so low numbers have a disproportionally large effect on N_e. Second, not all members of the population may actually participate in reproduction and transfer their genes to the next generation. If the sex ratio of the breeding pool departs from unity, for example due to the fact that a few males sire all females and many males have no offspring, this will decrease N_e far below the census population. Population size has a pervasive influence on the genetic structure of a population because small populations are conducive to sampling errors arising from the fact that the genetic material of the offspring generation is a sample from the genetic composition of the parental generation. In small populations, random genetic drift is the most important factor for changes in gene frequencies. Population genetic theory shows that the effect of drift depends not only on N_e, but also on the mutation rate per nucleotide (μ); this is why Lynch and Conery (2003) calculated $N_e\mu$ for every species and correlated this with genome size (Fig. 2.4).

According to Lynch and Conery (2003) there is nothing adaptive in the increase of genome complexity; it can be considered mainly an effect of genetic drift governed by the laws of neutral population genetics. Although this conclusion has been challenged and examples to the contrary have been pointed out (Vinogradov 2004), convincing adaptive explanations for the increase in genome complexity have not been given. For many aspects of genome architecture a neutralist view seems to be a good starting point (Lynch 2007a).

2.1.2 Gene families

In any genome, a large number of genes are similar to each other due to their common descent from a duplication event (like the *Notch* genes in Fig. 2.3). Such genes are said to be *paralogous* to each other. Genes in different species sharing a common ancestor are called *orthologous*. These two different types of homology are illustrated in Fig. 2.5. All the paralogous genes in a genome constitute a *gene family*. Examples of gene families are the globulins, homeobox proteins, esterases, trypsin-like peptidases, G-protein-coupled receptors, cytochrome P450s, and proteins involved in the immune response.

After a duplication event, there are four possible scenarios for the fate of duplicates:

1. Both copies remain active with the same function and together produce twice the amount of mRNA that the ancestral gene did before; this situation may be maintained because of the selective advantage of a large amount of gene product. We call this *superfunctionalization*.
2. One of the copies is silenced by degenerative mutation and the other continues its function. This is called *nonfunctionalization*.
3. One copy acquires a new advantageous function and is preserved by natural selection while the other retains the original function (*neofunctionalization*).
4. Both copies continue to be active but are compromised by deleterious mutations in their regulatory regions to the effect that each gene has a lower expression than the ancestral gene and both duplicates are required to produce the full complement of functions (*subfunctionalization*). This situation is also described as *duplication–degeneration–complementation* (the DDC model).

Analysis of the genomic databases for several eukaryotic species has suggested that the vast majority of gene duplications eventually lead to nonfunctionalization; neutral evolution seems to dominate the fate of duplicated genes (Lynch and Conery 2000). Still, many duplicated genes are preserved and the most important mechanism for their survival seems to be subfunctionalization; the *Hoxb1* genes in zebrafish closely approach this model (Prince and Pickett 2002). Superfunctionalization is

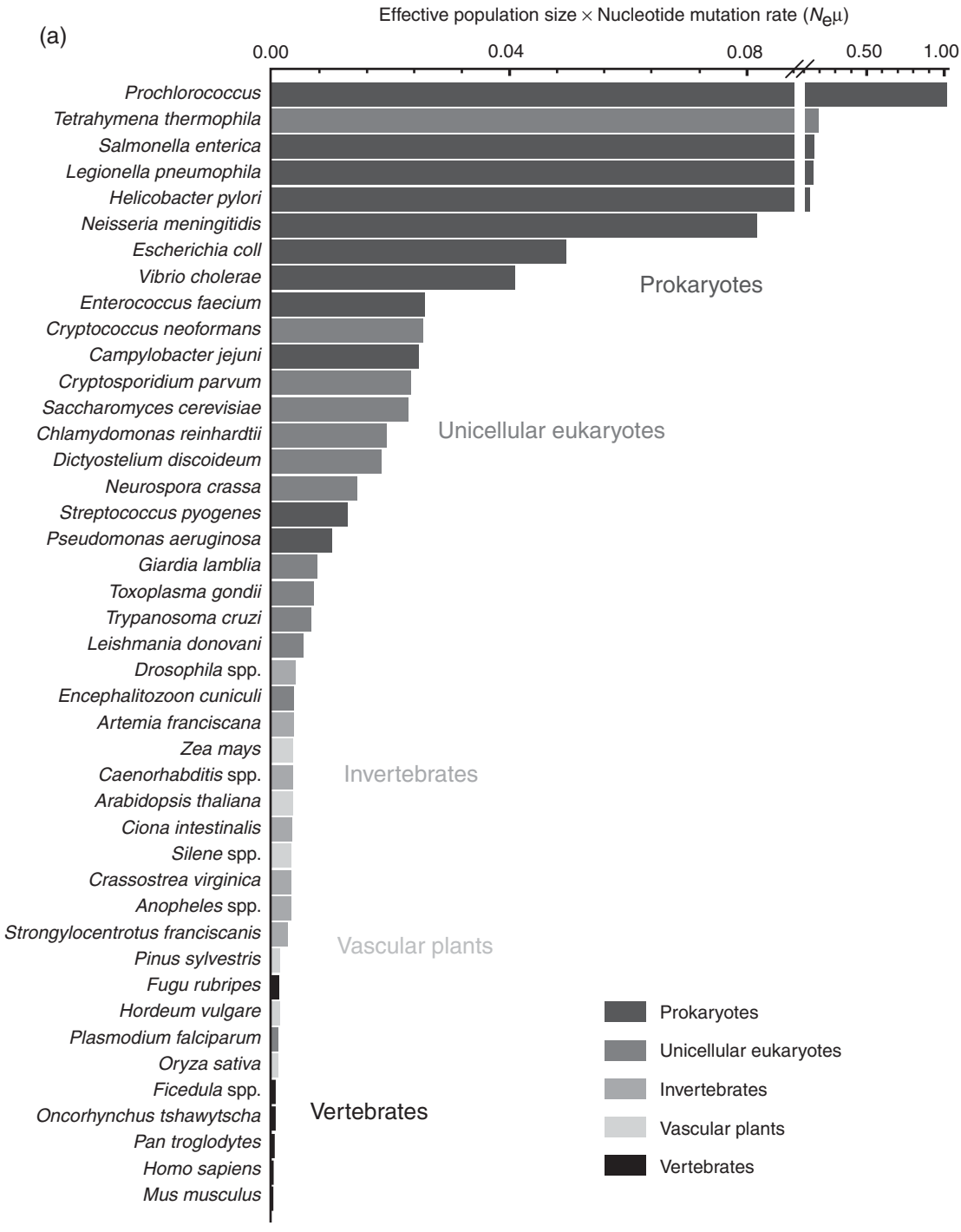


Figure 2.4 (a) Effective population size times mutation rate per nucleotide ($N_e\mu$), plotted for various species. Note that despite the general decreasing trend from prokaryotes to vertebrates, there is some mixing of species across taxonomic categories (e.g. *Tetrahymena thermophila* is a ciliate but has a value of $N_e\mu$ comparable to bacteria). (b) $N_e\mu$ and gene number as a function of genome size. Note the saturation curve for gene number, which is due to the preponderance of non-genic DNA in large genomes and the very good correlation between $N_e\mu$ and genome size. Reprinted with permission from Lynch and Conery (2003). Copyright 2003 AAAS.

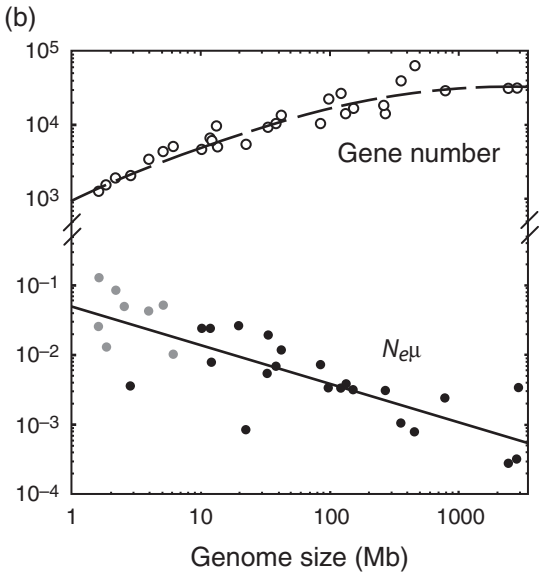


Figure 2.4 (*Continued*)

well known in the case of pesticide resistance in insects where tolerant strains often have duplicated or quadruplicated copies of genes encoding detoxification enzymes (Devonshire and Field 1991). It is assumed that such adaptive gene amplifications contribute significantly to the maintenance of paralogues in the early phase after a duplication (Kondrashov *et al*. 2002). Neofunctionalization is often considered to be a rare evolutionary pathway, but comparative genomics of yeast indicates that it may be more common than thought previously. Kellis *et al*. (2004) analysed the genome of baker's yeast, *Saccharomyces cerevisiae*, by comparison with a related fungus, *Kluyveromyces waltii*, and obtained proof for a whole-genome duplication in the *Saccharomyces* lineage. Moreover, the authors could track the various gene pairs and were able to show that in the majority of cases one of the paralogues retained the ancestral function, while the other underwent accelerated evolution. Interestingly, deletion mutations of the ancestral paralogue were often lethal, whereas mutations in the derived copy never were. These data support the neofunctionalization model and suggest that duplication may indeed be a mechanism for creating evolutionary novelty.

With many genes present as members of a family, phylogenetic analysis becomes more complicated. Should we include all the genes of a large gene family when comparing them with a set of homologous genes in another species? One possible solution is to limit comparisons to the *core proteome*: the number of distinct families in each organism, counting a cluster of paralogues as one. Rubin *et al*. (2000) made a genome-wide comparison of gene families among the three model eukaryotic species, yeast, nematode, and fruit fly, and compared them with the bacterium, *H. influenzae* (Table 2.2). It is remarkable that *Drosophila* has a core proteome only twice that of yeast. One would not expect that just doubling the number of genes could make up for the increase in complexity associated with multicellularity, the development of a spatially differentiated body plan, the need for communication networks between organs, and the processing of sensory

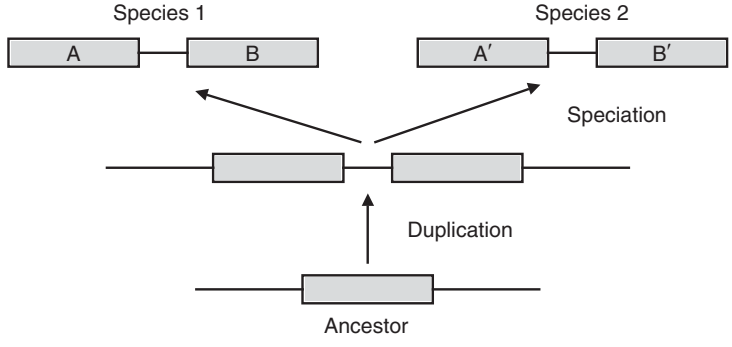


Figure 2.5 Illustrating the difference between paralogous and orthologous homologies. Genes A and B in species 1 are paralogous to each other, as are A′ and B′ in species 2; however, gene A in species 1 and gene A′ in species 2 are orthologues, as are genes B and B′.

Table 2.2 Estimated fraction of genes arising through duplication and estimates of the core proteome in three eukaryotes and one prokaryotic organism

Species	Total no. of predicted genes	No. of genes duplicated	Percentage of paralogues	Total no. of distinct families
Haemophilus influenzae	1709	284	17	1425
Saccharomyces cerevisiae	6241	1858	30	4383
Caenorhabditis elegans	18 424	8971	49	9453
Drosophila melanogaster	13 601	5536	41	8065

Reprinted with permission from Rubin *et al*. (2000). Copyright 2000 AAAS.

information and locomotion! It is also interesting to note that the core proteome of the nematode is the same size as that of the fruit fly, which again is astonishing given the large differences in development and body plan that exist between the two species. The lesson from this comparison must be that complexity in metazoans is not achieved by sheer number of genes. Instead, it is the regulation of these genes, in place and time and in response to the environment, added to the way they are organized in networks, that determines the differences between the species.

2.1.3 Skew, GC content, and codon usage

Turning our attention from genes to nucleotides, we will now explore patterns of the genomic occurrence of the four bases, adenine (A), guanine (G), thymidine (T), and cytosine (C). Obviously, A, G, T, and C are not distributed randomly in a DNA molecule. For the moment, disregarding the sequence itself, we will discuss two quantities in which deviations from random occurrence are expressed, GC skew and GC content.

If there is no mutation bias between the two strands of a DNA molecule, the orientation of any base-pair with respect to the leading and lagging strand is arbitrary and one expects that the number of As on a strand will be equal to the number of Ts; the same is expected for G and C. However, in practice deviations are seen, especially in bacterial genomes. The bias is expressed in a quantity called *GC skew* (S_{GC}), which is defined as:

$$S_{GC} = \frac{f_G - f_C}{f_G + f_C}$$

where f_G is the frequency of G and f_C the frequency of C in a certain segment (window) of DNA. S_{AT} is defined in an analogous way. The expected value of S_{GC} when the frequencies of G and C are the same is zero. Lobry (1996) found that GC skew showed a consistent switch across the origin of replication in three bacterial genomes, from a negative value (bias towards C) leftward of the origin of replication, to a positive value (bias towards G) rightward of the origin. This was confirmed by Blattner *et al*. (1997) when the genome of *E. coli* was completely sequenced (Fig. 2.6). The most likely explanation for skew is mutational bias associated with DNA replication; that is, in some way or another, the way in which the leading and lagging strands are replicated causes the leading strand to become more rich in G. This interpretation is supported by the fact that GC skew is most prominent in intergenic positions and third codon positions (Fig. 2.6), where selective pressure is relaxed compared to the first and second codon positions. That the leading and lagging strands of prokaryotic DNA can differ in mutation rate was confirmed by Tillier and Collins (2000), who showed that sequences of the same gene can vary depending on their orientation in the genome; that is, whether they are encoded in the leading or the lagging strand.

Another phenomenon that characterizes many genomes is bias in GC base-pairs over AT base-pairs. This bias is valid for different segments of a chromosome, between coding and non-coding DNA, and across species. In the majority of cases, the GC content in coding regions of the DNA is higher than in flanking regions of a gene. The 5° flanking region tends to be more GC-rich than the 3° flanking region, which is due to the fact that

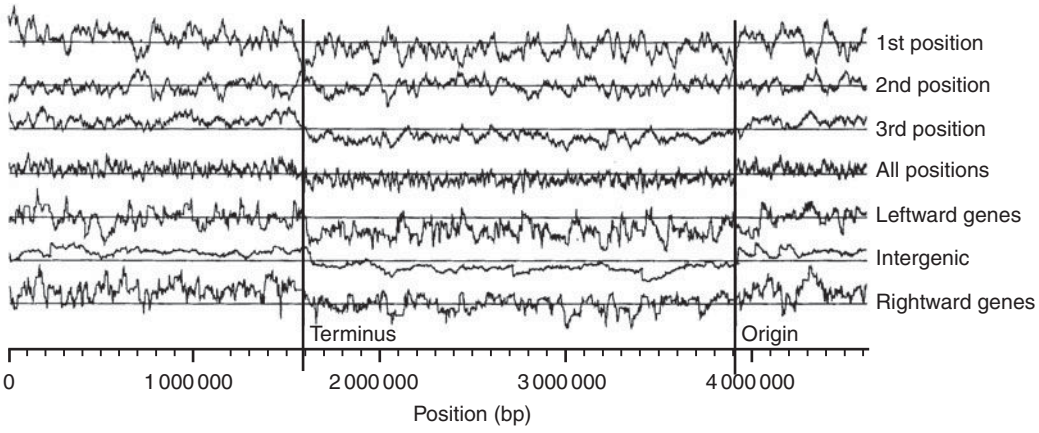


Figure 2.6 GC skew over the entire genome of *E. coli*, expressed as $(f_G - f_C)/(f_G + f_C)$ in 10 kbp windows of the genome, drawn in linear representation (in fact, the molecule is circular). GC skew is plotted separately for the three codon positions, for leftward genes, rightward genes, and non-protein coding regions. A negative value (below the line) indicates bias towards C, and a positive value means bias towards G. The two vertical lines indicate the origin of replication (right) and terminus of replication (left). Reprinted with permission from Blattner *et al*. (1997). Copyright 1997 AAAS.

promoter sequences tend to be relatively rich in GC. The GC content is also biased over longer stretches of DNA (around 300 kbp). Pieces of DNA with similar GC content are called *isochores*. This term refers to fractions of equal density obtained when DNA is subjected to CsCl gradient ultra-centrifugation. When mammalian DNA is sheared mechanically it tends to fall apart into fragments belonging to five discrete families, called L1, L2, H1, H2, and H3, each with a different GC content (P_{GC}; Bernardi 2000). The GC-poor families L1 and L2 ($P_{GC} < 44\%$) are poor in genes, whereas the GC-rich families H1 ($44 \le P_{GC} < 47$), H2 ($47 \le P_{GC} < 52$), and H3 ($P_{GC} \ge 52\%$) are increasingly rich in genes. The human genome consists of a mosaic of such isochores (Fig. 2.7) and the same holds for other mammals. The distribution of isochores is correlated with the banding pattern of metaphase chromosomes seen in microscopic preparations using fluorescent dyes. The short isochores may indicate so-called *CpG islands*, GC-rich stretches of DNA of at least 200 bp with an overrepresentation of GC dinucleotide repeats (the p in CpG indicates that it is a dinucleotide, not a base-pair). Such CpG islands are indicative of genes that are switched on frequently.

In invertebrates and plants there is less heterogeneity of GC content in the genome; very long isochores are observed in the centre of chromosome I in *Arabidopsis* and in the centre of chromosome III of *C. elegans* (Fig. 2.7). There is no correlation between gene density and GC content in the *C. elegans* genome; however, in *Drosophila* there is such a correlation, which is consistent with the greater variability of GC content over the genome (Oliver *et al*. 2001).

In addition to variation along the chromosome, there are also conspicuous differences between species in GC content. The GC content of Bacteria and Archaea varies between 25 and 75%, depending on the species, whereas the GC content of eukaryotes varies within a much smaller range, between 40 and 55%. GC contents in coding regions of the DNA are now known for more than 20 000 species; a database, fed by GenBank sequences for full-length proteins, is maintained by Nakamura *et al*. (2000). Table 2.3 lists some examples of GC content taken from this database.

There is no good explanation for why the GC content of a genome should vary between species. One of the forces acting upon the GC content is asymmetry in substitution rates. If the rate of substitution from G/C to A/T is denoted as u and the rate of substitution from A/T to G/C as v, then the equilibrium content of GC is expected to take the value P_{GC}, where (Graur and Li 2000)

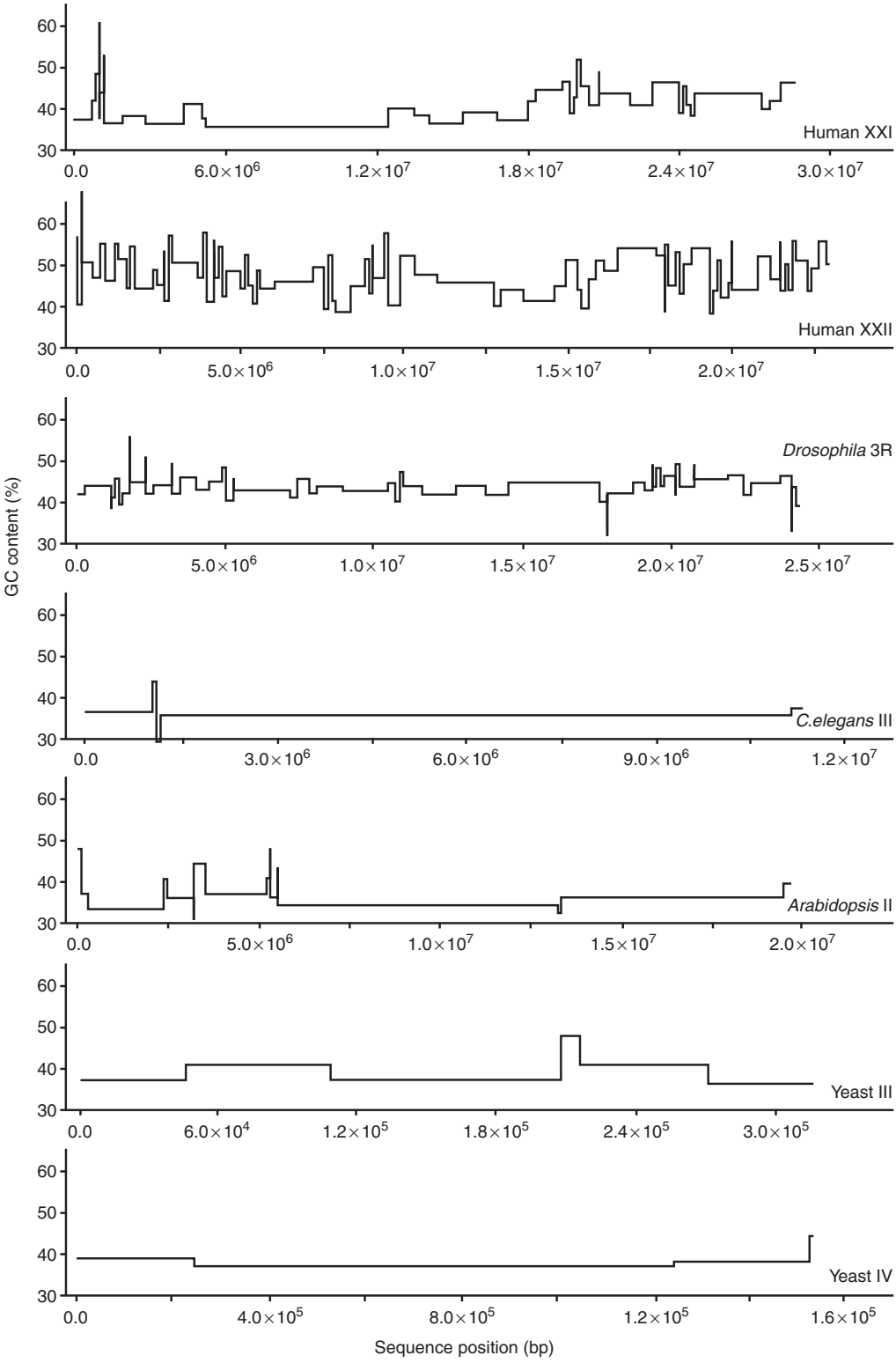


Figure 2.7 Isochore maps of some eukaryotic chromosomes, calculated from the genomic databases with an improved 'compositional segmentation' algorithm. For human chromosomes XXI and XXII the largest available contigs were taken. Note that there is much more structure in the human and fruit fly genomes, indicating variable gene density along the chromosome, than in the *C. elegans*, *Arabidopsis*, and yeast genomes, indicating a more even spread of genes. After Oliver *et al*. (2001), with permission from Elsevier.

Table 2.3 GC contents of coding regions of genomes of some selected species from the three domains of life

Species	GC (%)			
	Overall	First codon position	Second codon position	Third codon position
Archaea				
Methanococcus maripaludis	34.08	44.70	32.29	25.26
Halobacterium salinarum	64.85	68.07	43.04	83.45
Thermoplasma acidophilum	47.37	49.55	37.55	55.02
Sulfolobus solfataricus	36.49	43.22	33.65	32.59
Pyrococcus furiosus	31.09	49.46	34.56	39.24
Bacteria				
Sinorhizobium meliloti	63.14	64.88	45.23	79.30
Escherichia coli	51.39	58.27	40.84	55.06
Leptospira interrogans	36.42	44.70	34.62	29.94
Actinomyces naeslundi	67.51	63.74	49.50	89.27
Thermotoga maritima	46.40	52.24	34.40	52.55
Eukarya				
Saccharomyces cerevisiae (budding yeast)	39.76	44.60	36.61	38.09
Caenorhabditis elegans (roundworm)	42.93	49.95	38.93	39.90
Arabidopsis thaliana (thale cress)	44.60	50.90	40.52	42.37
Drosophila melanogaster (fruit fly)	53.97	55.87	41.51	64.52
Fundulus heteroclitus (mummichog fish)	53.93	55.10	40.89	65.81
Danio rerio (zebrafish)	50.94	54.50	40.82	57.52
Xenopus laevis (clawed frog)	46.97	52.75	39.94	48.22
Vipera aspis (aspis viper)	52.20	44.05	43.17	69.38
Anas platyrhynchos (mallard duck)	51.27	53.89	41.16	58.78
Sus scrofa (wild pig)	54.44	56.26	41.81	65.25

Source: Data from www.kazusa.or.jp/codon (Nakamura *et al.* 2000).

$$P_{GC} = \frac{\qquad}{u+}$$

If the two rates are equal, P_{GC} is expected to be 1/2, or 50%. A high GC ratio is evidence for v being larger than u. The asymmetry of substitutions is called *GC mutational pressure*. However, GC content is not only shaped by mutation, but to some extent also by selection. One of the selective constraints arises from codon usage. We know that several amino acids are encoded in the DNA by more than one triplet, but this does not imply that all possible triplets are used in proportion. In most amino acids there is a bias towards the use of certain codons because the tRNAs of these codons are more abundant (*codon usage bias*). Depending on the number of Gs in the preferred codon, selection can constrain the GC content. Another source of selection is due to the superior stability of the G–C bond, which uses three hydrogen bridges, whereas the A–T bond uses two. It has been argued by several researchers that high ambient temperatures would favour high GC contents and that endothermic ('warm blooded') animals (birds and mammals) would have higher GC contents than ectothermic ('cold-blooded') animals (most reptiles, amphibians, fish, and invertebrates).

If selection is a major factor influencing the GC content and if GC mutational pressure drives the GC content upwards, one would expect that P_{GC} is higher in the third codon positions of a protein-encoding gene, compared with the first and second

positions. There is some evidence that this is indeed the case both for prokaryotes and eukaryotes (Table 2.3); however, whether this is indeed indicative of selection is questionable; bias introduced by mobile elements with a high GC content is usually ignored (Duret and Hurst 2001). The possible role of temperature stability in the evolution of GC content was critically examined by Hughes (1999) and he concluded that the hypothesis is not supported strongly by the data. Among thermophilic prokaryotes there are species with low and high GC contents and within the vertebrates the picture is not consistent either (Table 2.3). The strongest effect in the GC content seems to stem from phylogenetic constraints: taxonomically related species tend to have similar GC contents. Although purifying selection may play a modest role, the dominant factors acting upon GC content seem to be neutral processes, GC mutational pressure, and random drift.

2.1.4 Gene order

In genetics, two loci are called *syntenic* if they are located on the same chromosome (Russell 2002). In genomics, however, the term *synteny* is used to indicate a situation where a series of genes is arranged in the same order on different genomes (Gibson and Muse 2002). Passarge *et al*. (1999) have rightly pointed out that this new usage is incorrect and etymologically awkward. Another term that may be more appropriate is *colinearity*; however, in this book we will comply with the most common usage and use synteny in the genomics understanding. The presence of synteny between two genomes is somewhat dependent on the scale of analysis. On the level of the chromosomes, synteny may be demonstrated by techniques such as *chromosome painting* (using interspecies fluorescent *in situ* hybridization); however, this does not exclude the presence of extensive rearrangements on the level of individual genes. The term *microsynteny* is used to indicate detailed sequence comparisons between individual genes within a chromosomal segment.

In any genome, genes are found to be organized in clusters and these clusters sometimes maintain the same order across species, even across groups as far apart as mammals and fish. Well-known examples of synteny are histone genes, *Hox* gene clusters, and the genes of the major histocompatibility complex (MHC). There are also large synteny blocks, covering hundreds of kilobase pairs, between the genomes of rice and *Arabidopsis* (see Section 2.2).

The *Hox* genes are a famous example of long-range synteny. Indeed, the same order of genes can be found in the *Hox* clusters of nematodes, insects, and mammals. *Hox* genes encode transcription factors that regulate developmental patterns across the anterior–posterior axis of bilaterian animals (Carroll *et al*. 2005). Macroevolutionary relationships in the animal kingdom can be partly understood as duplications followed by neo- and nonfunctionalization of essentially the same pattern of *Hox* genes (Amores *et al*. 1998; Carroll 2000).

Synteny analysis is an important tool in comparative genomics. The relative order of genes in one species can provide clues about the presence or even the function of genes in another species. Similarly, by looking at the order of genes in a cluster, one can discover genes by homology to another species that were missed by automatic gene-finding algorithms. Synteny analysis can also be a tool to reveal duplications across species, for example by searching two regions in one genome that have the same gene order as one region in another genome (*doubly conserved synteny*; Kellis *et al*. 2004).

How could such blocks of gene order be maintained while other regions of the genome are reshuffled extensively by recombination? How can it be that some genes are free to move through the genome while others are tied, for millions of years, to the same neighbours? One of the reasons could be selective pressure acting upon the cluster as an integrated whole. This is certainly the case if genes are organized in operons, as in all prokaryotic and some eukaryotic genomes (see Sections 2.2 and 2.3). Another functional constraint is illustrated by the *Hox* genes. In most animals, the order of these genes along the chromosome is the same as the order of their expression domains along the anterior–posterior body axis. In addition, the genes at the front end of the complex are expressed earlier in development than the ones at the back end. These observations suggest that it is the requirement for coherent

temporal expression that is maintaining colinearity of the *Hox* cluster (Patel 2004). The genetic developmental system that governs the basic positional information of tissues in all animals was called the *zootype* (Slack *et al*. 1993).

Another reason for conservation of gene order, proposed more recently, is *interdigitization* of regulatory elements. We know that the expression of genes is controlled by regulatory elements, usually in the 5′ region of the gene. It turns out that some regulatory elements may be physically linked to genes close by, or even be located in the introns of other genes. The fixation of regulatory elements inside the territory of neighbouring genes thus forges a physical bond resulting in close linkage between the genes. Another way in which regulatory elements may link genes together occurs when the expression of a group of genes is controlled by a single *locus-control region*. The principle of the locus-control region was first discovered in the β-globulin cluster of the human genome, but similar regions, presumably participating in dynamic chromatin alteration, have now been found in other gene clusters. In all these cases genes cannot move independently from each other without gaining a severe selective disadvantage. The shuffling and reorganization of the genome during evolution, as highlighted in Chapter 6 and indicated by the term genetic turbulence, is inevitably limited to some extent by such processes.

An example of synteny analysis is provided by a comparison between two distantly related species of nematode, sharing a common ancestor 300–500 million years ago (Guiliano *et al*. 2003). A nematode parasite of vertebrates, *Brugia malayi* (order Ascarida), was compared with the well-known model species *C. elegans* (order Rhabditida). Whereas the genome of the latter species is completely known and many genes are annotated, the genome of *Brugia* was only known incompletely at the time; the comparison was undertaken partly to reveal more of the function of genes in *Brugia* from knowledge of *C. elegans*. Figure 2.8 shows an alignment of two overlapping contigs of *B. malayi* with a homologous portion of the genome of *C. elegans*. Eleven putative *Brugia* genes in this region can be homologized to genes in the *C. elegans* genome and, except for one case, the order of the genes is the same in the two species; in two cases the direction of transcription was reversed. Although earlier comparisons between two related species, *Caenorhabditis briggsae* and *C. elegans*, had

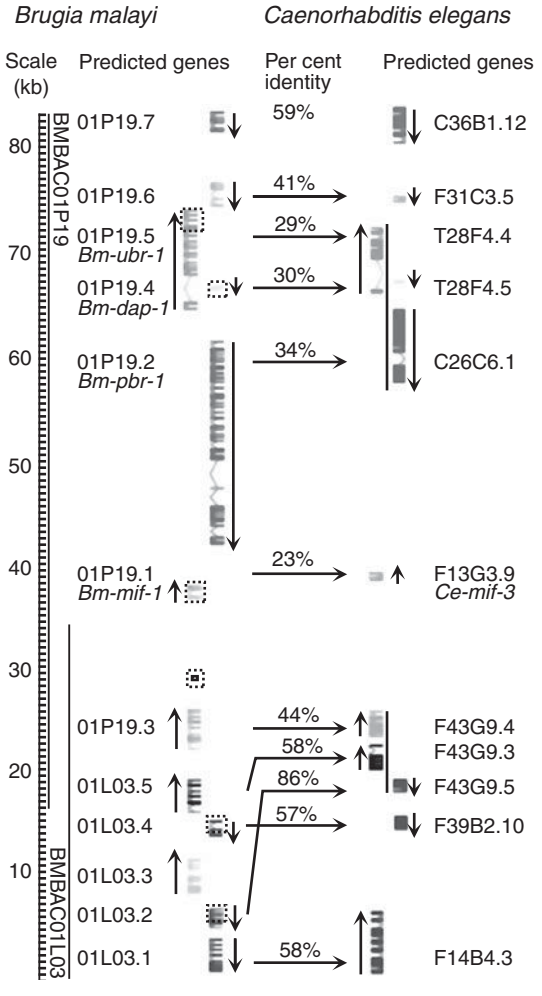


Figure 2.8 Synteny between two distantly related species of nematode, *Brugia malayi* and *Caenorhabditis elegans*. An alignment is shown of two overlapping contigs from a genomic library of *Brugia* (BMBAC01P19 and BMBAC01L03) with the *C. elegans* genome. Exons of genes are indicated by boxes, with brackets between them to indicate introns. Genes are designated by code names from top to bottom (01P19.7, etc.), but have different codes in the different species. The direction of transcription is indicated for each gene by a vertical arrow. The horizontal arrows indicate matches of putative *B. malayi* genes in the *C. elegans* genome. From Guiliano *et al*. (2003), by permission of BioMed Central.

shown that nematodes have a high rate of intrachromosomal rearrangement, the present observations demonstrate that there may still be local clusters of synteny across relatively large phylogenetic distances. To explain this, Guiliano *et al*. (2003) could not identify a common function for the gene cluster but suggested that promoters or *cis*-acting regulatory elements could be embedded within other cluster members, so synteny could be maintained by interdigitization of regulatory sequences.

2.2 Prokaryotic genomes

The genomes of prokaryotes typically consist of one circular molecule, representing a haploid genome, characterized by an *origin of replication* and two semicircles of sequence that are read leftward and rightward, with the terminus of replication positioned on the opposite side of the molecule. The origin of replication is a stretch of 200–300 bp that has characteristic sequences allowing DNA unwinding and binding of the initiator protein DnaA. The genes in prokaryotic genomes are organized in *operons*, clusters of genes, the expressions of which are regulated jointly by interactions between operator and repressor proteins (cf. Fig. 2.9). The structural genes in an operon are transcribed into a single *polygenic mRNA*. The situation that an mRNA molecule contains protein-coding sequences from more than one gene (cistron) is called *polycistronic*. It was long thought that polycistronic transcription was limited to prokaryotes and that all eukaryotic mRNAs were monocistronic; however, it turned out in 1994 that the nematode *C. elegan*s has approximately 25% of its genes organized in polycistronic units of two or more members (Hodgkin *et al*. 1995), so polygenic transcription is not exclusive to prokaryotes.

The genes in a microbial genome are usually subdivided into two main classes: *informational genes* and *operational genes*. In the first category (also called *essential genes*) are genes related to information processing, such as transcription, translation, replication, and so on. These genes define the 'essence' of the species. They are usually large, complex systems, with many interactions, located on the main chromosome of the cell and less prone to lateral gene transfer (see below). The second category includes genes that function in basic cell maintenance processes and metabolism, such as protein, energy metabolism, and phospholipid acid biosynthesis. These genes are usually members of small assemblies of a few gene products and are more often found on plasmids.

2.2.1 Chromosomes and plasmids

As an example of the architecture of a prokaryote genome, we reproduce in Fig. 2.9 the sequence of *H. influenzae*, published by Fleischmann *et al*. (1995). The figure summarizes some general aspects of the sequence and provides a great deal of information on restriction sites, arrangement of gene clusters, GC content, tandem repeats, and so on. The relatively small prokaryotic genomes are the only ones for which this information can be reproduced in a still more-or-less readable figure, using 16 different colours (shades of grey in Fig. 2.9) to indicate various functional details (colour-blind people are at a real disadvantage in genomics!). An analysis of the functional categories of these genes allows some first insight into the functions that can and cannot be performed by the organism. For example, the genome of *H. influenzae* has a complete metabolic machinery required for survival and reproduction as a free-living organism, but it lacks genes such as the fimbrial gene cluster that encodes proteins allowing the bacterium to attach to host cells. This is explained by the fact that the sequenced strain is not pathogenic. The sequence also includes a cryptic μ-like prophage, which is a segment resembling the DNA of bacteriophage μ. We will see below that many bacterial and archaeal genomes contain genetic material from bacteriophage origin. A bacteriophage that is 'sitting' in its host genome, being replicated with the chromosome of the host and waiting to become active, is called a *prophage*. In some cases these prophage sequences become debilitated and are no longer functional but are still recognizable in the genome.

Although the single circular chromosome is the most common genome organization among prokaryotes, an increasing number of exceptions to this are being identified. In addition to circular

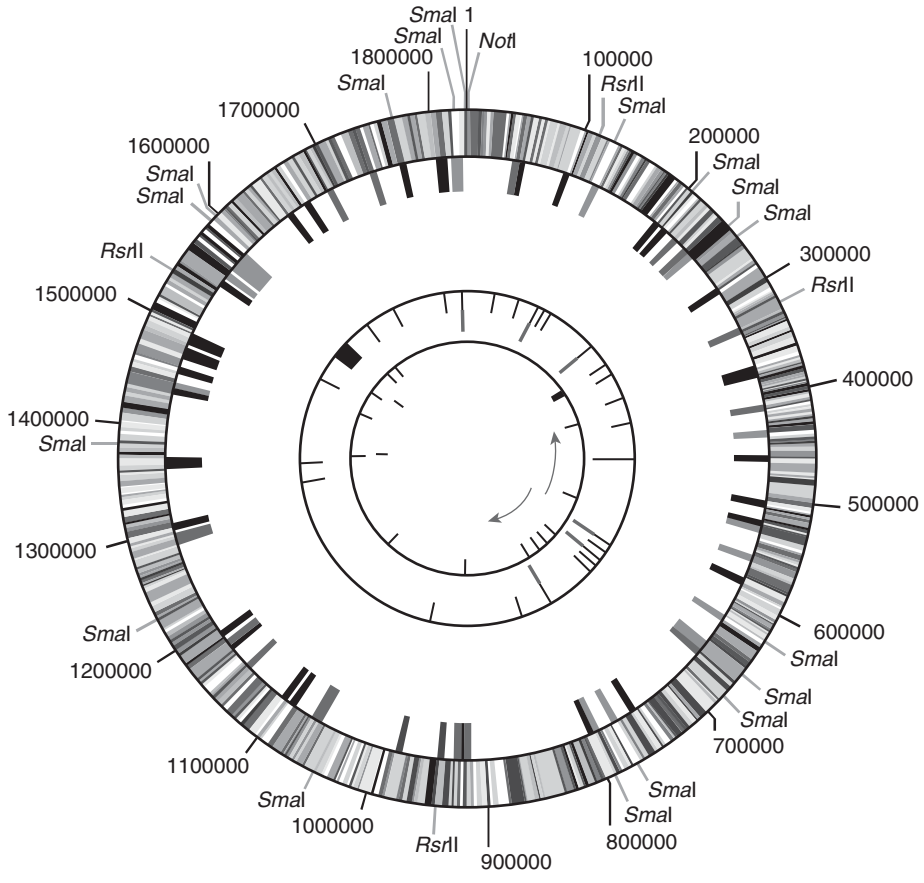


Figure 2.9 Circular representation of the annotated genome of *H. influenzae*, as published by Fleischmann *et al*. (1995). Its characteristics are indicated by four concentric circles and rings. Outer perimeter, restriction sites. The *Not*I restriction site occurs only once and was arbitrarily chosen as the site at which to begin the clockwise numbering of base pairs (top of the figure). Outer ring, locations of genes that could be assigned a role. In the original paper genes were classified according to 16 different functions, which in this figure are represented by different grey tones. Second ring, regions of high GC content (grey) and high AT content (black). Third concentric circle, locations of six ribosomal operons (extended markers), tRNAs (short markers), and a cryptic m-like prophage (black box). Fourth (inner) circle, locations of simple tandem repeats. The origin of replication is indicated by two outward-pointing arrows near base pair no. 603 000. Two potential termination sites are indicated near the opposite midpoint of the inner circle. Reprinted with permission from Fleischmann *et al*. (1995). Copyright 1995 AAAS.

chromosomes, many prokaryotes have smaller circular DNA molecules, called plasmids. The classical example of a bacterial plasmid is the *F-plasmid* of *E. coli*, which is transferred from a donor (F^+ cell) to a recipient (F^- cell) during *conjugation*. The F-plasmid carries genes associated with the formation of *F-pili*, hair-like surface structures which allow the physical union between the F^+ and F^- mating types. Modified prokaryotic plasmids have been used extensively as cloning vectors to develop DNA libraries. For example, bacterial artificial chromosomes (BACs) were modelled after F-plasmids and use the F-plasmid's origin of replication.

Like the F-plasmid, many bacterial plasmids contain genes dedicated towards specific, so-called non-essential, functions. These functions may relate to antibiotic resistance, pathogenicity, or specialized metabolic pathways. In *Agrobacterium tumefaciens* (Alphaproteobacteria), one of the plasmids carries a discrete set of genes (*transforming DNA*, or

T-DNA) that can be transferred to a plant host and, when expressed in the host, promote the synthesis of plant growth hormones that redirect the local development of plant tissue to form a gall. In the gall, the bacterium also directs the plant to produce specific amino acid derivatives (*opines*) that can be used as the sole carbon and nitrogen source for *Agrobacterium*. The transfer of the T-DNA and its integration in the chromosome of the host is supported by proteins encoded in virulence (*vir*) genes, which are also on the plasmid. The plasmid is thus specialized for gall-formation and is therefore called a *tumour-inducing plasmid* or Ti plasmid. The capacity of the Ti plasmid of *A. tumefaciens* to introduce foreign DNA into a host has been widely exploited by plant molecular biologists to artificially transform plants.

A similar specialized function located on a plasmid is present in the symbiotic nitrogen-fixing soil bacterium *Sinorhizobium meliloti*. This bacterium belongs to a group of three closely related Alphaproteobacteria of the family Rhizobiaceae that now have been sequenced completely (*S. meloti*, *Mesorhizobium loti*, and *A. tumefaciens*; Galibert *et al*. 2001; Goodner *et al*. 2001; Wood *et al*. 2001). *S. meloti* is a symbiont of alfalfa (*Medicago sativa*) and because of the profound ecological importance of symbiotic nitrogen fixation in the global nitrogen cycle, genomic studies of these types of bacteria are particularly relevant for ecological genomics. Figure 2.10 shows the position of the three Rhizobiaceae in relation to some other bacteria from the protobacterial lineages with fully-sequenced genomes.

The plasmid of *S. meliloti* which carries most of the symbiotic functions is called SymA or *symbiotic plasmid*. An extensive annotation of gene functions of this large plasmid (1354 kbp, comparable in size to an entire prokaryotic genome!) was published by Barnett *et al*. (2001). Specific nodulation (*nod*) genes on the plasmid encode biosynthetic enzymes for the so-called *Nod factors*, molecules which are active in very low concentrations (down to 10^{-12} M) in the signalling between plants and rhizobia in the early stages of root-nodule formation. The expression of Nod genes is triggered, via a signal transduction pathway, by phenolic compounds excreted by plant roots. The plasmid also carries several genes encoding enzymes necessary for nitrogen fixation,

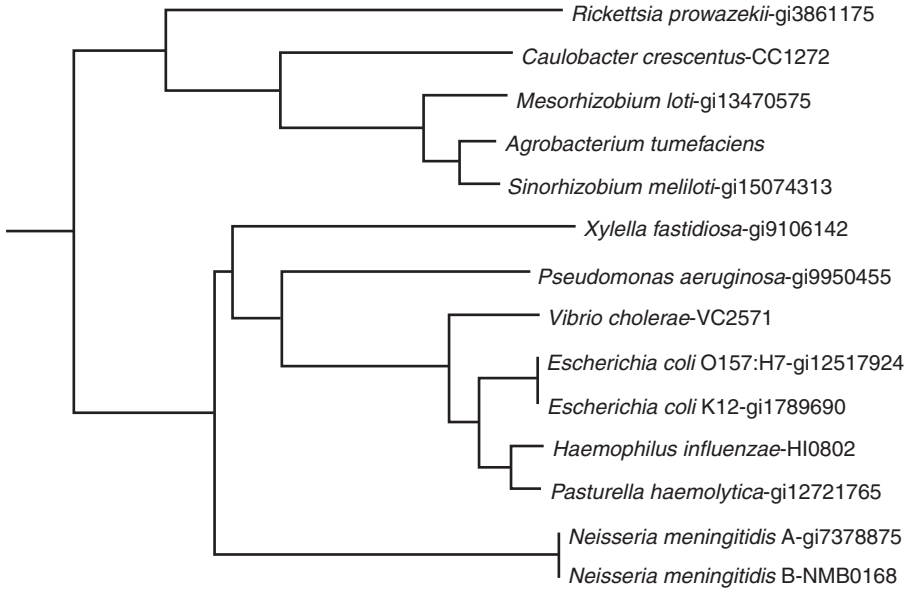


Figure 2.10 Phylogeny, based on the RpoA (RNA polymerase α) gene, for 14 bacteria with fully-sequenced genomes from the Alpha-, Beta-, and Gammaproteobacteria phyla. Reprinted with permission from Wood *et al*. (2001). Copyright 2001 AAAS.

denitrification, and opine metabolism, and many genes involved with transport and osmotic stress. It is interesting to note that no informational genes, for example relating to DNA replication or cell division, were found on pSymA, demonstrating that, despite its size, it is a true plasmid.

In addition to circular chromosomes and plasmids, some bacteria also have *linear chromosomes*. A particularly complicated situation is present in the genome of *A. tumefaciens*, which was published by Wood *et al*. (2001) and Goodner *et al*. (2001). It consists of one circular chromosome (2.8 Mbp), one linear chromosome (2.1 Mbp), and two plasmids, which in this case are called *megaplasmids* because they are relatively large (543 and 214 kbp). The smaller of the two plasmids is the Ti plasmid referred to above. The linear chromosome appeared to carry 35% of the genome's genes, including those encoding ribosomal and DNA-replication proteins, as well as 21 complete metabolic pathways. The presence of these genes confirmed the chromosome-like nature of the linear element; however, other genes on the same chromosome, especially those involved in conjugation, are reminiscent of a plasmid. Most interestingly, the sequence revealed that there was a *repABC* operon (a gene known to be associated with replication of circular chromosomes) near the centre of the chromosome, plus an inversion in the GC skew (see Section 2.1.3). This seems to indicate that the linear chromosome has a bidirectional, plasmid-like mode of replication. Why *A. tumefaciens* maintains such a complicated arrangement of its genome and the advantages of having multiple chromosomes remain a mystery. Multiple chromosomal elements are especially prominent among proteobacterial phyla and spirochaetes, and this in itself would suggest phylogenetic determination of the tendency to form extra-chromosomal elements.

2.2.2 Lateral gene transfer

For many years it has been known that microorganisms can absorb DNA directly from the environment. The relative ease by which antibiotic resistance can be donated from one bacterium to another constitutes further proof that genetic information is not only transferred during cell division (vertical transmission), but also from one intact cell to another (*lateral* or *horizontal transmission*). Lateral gene transfer can occur via three mechanisms (Zhou and Thompson 2004), as follows.

Transformation. This is the uptake of DNA directly from the environment. If prokaryotic cells can do this they are called *competent*. Very few bacteria are competent during their whole life cycle, but some are during certain physiological stages.

Transduction. Bacteriophages can transfer DNA between species if two host species share similar bacteriophage receptors. Transduction may concern random pieces of the host DNA, packaged during phage assembly in the lytic cycle, or it may be limited to the sequences flanking the insertion site.

Conjugation. Lateral gene transfer can occur via specialized plasmids during physical contact between F^+ and F^- cells, as discussed above.

Lateral gene transfer is not limited to bacteria of the same species, it can also occur among species of widely different origin and even between Bacteria and Archaea. The latter phenomenon was discovered by Nelson *et al*. (1999), when they sequenced the genome of the thermophilic bacterium *Thermotoga maritima* (Fig. 2.11). *Thermotoga* derives its name from the sheath-like envelope that surrounds the cell (the 'toga') and the fact that it has a temperature optimum for growth of 80 °C. It is usually placed in a separate lineage of the Bacteria, called Thermotogales, because of some unique characteristics, including rRNA sequences that are unusual among the Bacteria and a set of fatty acids that is only found in this group. Phylogenetic analysis, aimed at enlightening the position of *T. maritima* within the Bacteria, resulted in a great lack of congruence when different genes were used as a basis for the comparison. Further analysis of the ORFs in the *T. maritima* genome showed that no less than 24% of the predicted genes were most similar to proteins in archaeal species, rather than to Bacteria. The Archaea-like genes were found to lie in clusters (islands) in the genome of *T. maritima*; in several of these islands even the archaeal gene order was conserved. These observations and those by Aravind *et al*. (1998) on another thermophilic bacterium,

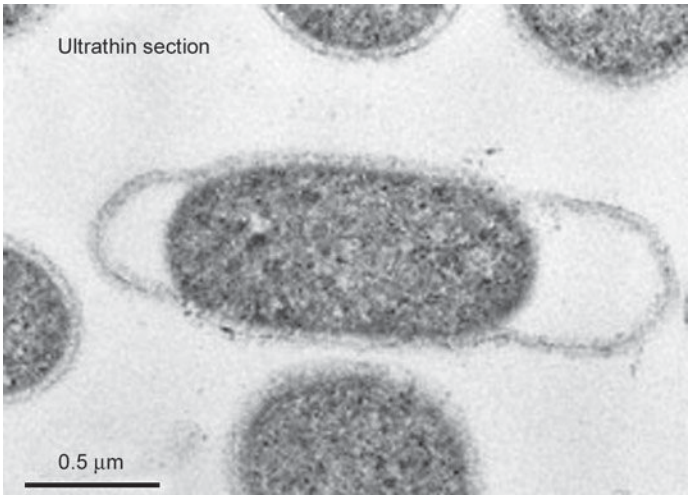


Figure 2.11 Electron micrograph of *Thermotoga maritima*, a thermophilic bacterium belonging to the group Thermotogales, which was isolated originally from a geothermal heated marine sediment at Vulcano, Italy. Courtesy of K.O. Stetter, University of Regensburg.

Aquifex aeolicus, provided great support to the theory that lateral gene transfer was not to be considered an oddity, but a very significant process for many microorganisms.

Koonin *et al*. (2001) performed a quantitative analysis of the frequency of lateral gene transfer by analysing the genomes of 8 archaeal and 22 bacterial species. They estimated that the percentage of new genes acquired from another domain (Bacteria, Archaea, or Eukarya) was on average 0.9% for the bacterial genomes and 3.4% for the archaeal genomes. When looking at interspecies transfers within the Bacteria, the percentages of acquisition of new genes varied considerably, from 0.4 to 19.8%, depending on the group. A particularly high frequency of foreign DNA is found in the genomes of the Spirochaetales. In general, it turned out that bacteria living at high temperatures had more archaeal genes in their genomes than mesophilic bacteria, and bacteria with a parasitic lifestyle more often had eukaryotic genes in their genome than non-parasitic bacteria. It therefore seems that lateral gene transfer is especially common among organisms that live in close proximity to each other.

Several authors have pointed out that not all evidence for lateral gene transfer is equally reliable. If lateral gene transfer is inferred only from the genome sequence, showing that certain ORFs have a BLAST match outside the group considered, this is not sufficient evidence, because alternative explanations may be given, such as the loss of genes in some lineages or widely diverging rates of evolution across the groups compared (Eisen 2000; Nesbø *et al*. 2001). To prove that lateral gene transfer has occurred, one needs to conduct a gene-by-gene phylogenetic analysis. As an example we discuss the work of Deppenheimer *et al*. (2002).

Deppenheimer *et al*. (2002) sequenced the genome of *Methanosarcina mazei* (Euryarchaeota), an archaeon of great ecological importance since it derives its energy from fermenting simple organic substrates to methane. Methane production in underwater sediments and inundated land (notably rice paddies) is an important link in the global carbon cycle. The genome of *M. mazei* (4.1 Mbp) was more than twice as large as other methanogenic Archaea that had been sequenced completely. Of the 3371 identified ORFs, no less than 1043 (31%) had their closest homologue not in an archaeal but in a bacterial species. Unlike the situation in *Thermotoga*, the bacterial genes in *Methanosarcina* did not cluster together in islands, but were found scattered in the genome. Gene phylogenies in which the laterally transferred genes were compared with orthologues in other Bacteria and Archaea showed that the foreign genes clustered with bacterial homologues, rather than with Archaea (Fig. 2.12).

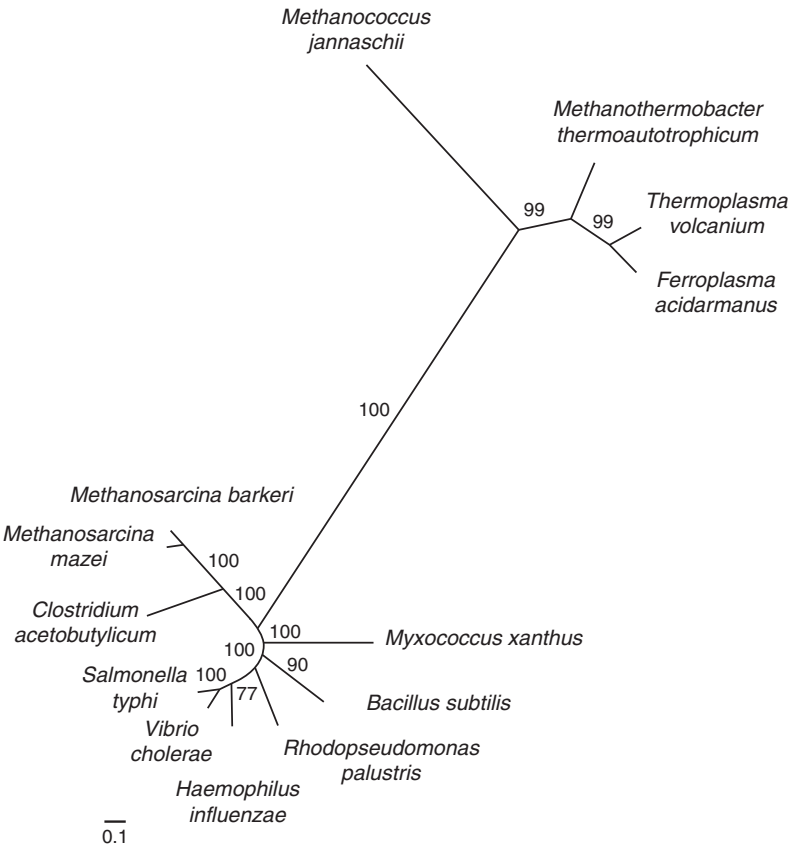


Figure 2.12 Unrooted phylogenetic tree, constructed using the neighbour-joining principle, for ATP-dependent Lon protease, a gene suspected of lateral transfer. The tree shows two clusters of archaeal (top right) and bacterial (bottom left) genes; however, the gene from the archaeon *M. mazei* clusters with the bacteria, rather than with the Archaea. This greatly supports the hypothesis that ATP Lon protease was acquired by *M. mazei* from a bacterial donor. Scale bar indicates the proportion of amino acid difference. After Deppenheimer *et al*. (2002), by permission of Horizon Press.

Detailed analysis of the metabolic role of the bacterial genes in *Methanosarca* showed that the imported genes had considerably enlarged the metabolic spectrum of the archaeon. The suggestion from the data was that the metabolism of *M. mazei* has evolved from a simple methanogenic pathway (using hydrogen and carbon dioxide to produce methane), to a much more versatile substrate spectrum, allowing the use of acetate, methanol, and methylamine. Interestingly, most of the laterally transferred genes seemed to have been obtained from obligate and facultative anaerobic bacteria; that is, from organisms that live in the same microenvironment as the methanogenic archaeon (sediments, inundated land). This is in line with observations on thermophilic communities showing that the greatest transfer is between organisms living close together.

Over the last few years, it has become obvious that lateral gene transfer is not limited to prokaryotes but is equally important, although maybe less frequent, between prokaryotes and multicellular eukaryotes (Andersson 2005; Doolittle *et al*. 2008; Keeling and Palmer 2008). Also in this case the principle of close proximity prevails: lateral gene transfer is most often seen between symbionts and their hosts, between parasites and their hosts, and between animals and their food. The first proof of a

prokaryote–eukaryote gene transfer came from plant-parasitic nematodes, who appeared to have β-1,4-endoglucanase activity of their own, necessary to degrade cellulose in plant cell walls. While cellulase activity is never found to originate in animals, this was due to lateral gene transfer from a prokaryote, mostly likely a *Rhizobium* (Smant *et al.* 1998). Screening of plant-parasitic nematode genomes has up to now revealed the presence of 13 genes of bacterial origin (Scholl *et al.* 2003).

One of the more spectacular examples of lateral gene transfer into an animal genome is from the sea slug *Elysia chlorotica*. This animal uses photosynthetic plastids from its food, the heterokont alga *Vaucheria litorea*, to perform its own photosynthesis in specialized cells of the digestive tract. This in itself is already a wonderful example of functional flexibility from the snail's side, since a chloroplast genome usually encodes no more than 10% of the proteins needed in photosynthesis. Obviously, the snail must have photosynthesis-supporting genes in its own genome; one of these genes, *psbO*, a manganese-stabilizing protein from photosystem II, has been identified as such (Rumpho *et al*. 2008). The transfer of this gene from the alga to the snail's genome is suggested to contribute to the stability of the host–plastid interaction.

Similar cases of lateral gene transfer are found in symbiontic relationships between bacteria and insects. Many plant-feeding insects culture bacteria in specialized cells of the digestive tract ('bacteriocytes') and use these obligate symbionts as a source of dietary items that are rare in plant sap, such as amino acids. These bacteria are usually specialized lineages of the Gammaproteobacteria, such as the genera *Buchnera* (in aphids), *Wigglesworthia* (in tsetse flies), *Blochmania* (in carpenter ants), and *Baumannia* (in 'sharpshooter' homopterans). In addition to these bacteriome-associated obligate symbionts insects carry several facultative symbionts, mutualists, and parasites such as *Wolbachia* and *Rickettsia* (Moran *et al.* 2008).

Several genes are now being identified in insect genomes that originate from endosymbiontic bacteria. One of the best studied cases is the aphid genome, which has received genes from the bacterial symbiont, *Buchnera aphidicola* (Nikoh and Nakabachi 2009). One of the genes, called rare lipoprotein A (*rlpA*) is found only in prokaryotes, never in any animal except aphids. Phylogenetic analysis suggests that the lateral gene transfer had already taken place early in the evolution of the aphids, since *rlpA* genes in three different aphids form a monophyletic group mirroring the evolution of the aphids themselves (Fig. 2.13).

Later gene transfer may even include the exchange of genes in tripartite coevolutionary relationships, such as in the case of the homopteran xylem-feeding insect *Homalodisca coagulata* ('sharpshooter') which carries two specialized bacteria, *Baumannia cicadellinicola* (Gammaproteobacteria) and *Sulcia muelleri* (Bacteroidetes). The two symbionts conduct complementary functions: *Baumannia* is specialized in the production of vitamins and co-factors for the host, while *Sulcia* does most of the synthesis of amino acids. Both bacteria have lost many metabolic pathways otherwise essential for bacterial life and so depend on each other, as well as on their host. Like in the photosynthetic snail, lateral gene transfer in such systems will contribute greatly to stability of the coevolutionary relationship.

As indicated by the examples above, lateral gene transfer is an important mechanism for recruiting new ecological functions. By lateral gene transfer, microorganisms and eukaryotes may be able to exploit new ecological niches that were inaccessible prior to the event. However, not all lateral gene transfers are necessarily to be viewed within a purely adaptive framework. The presence of foreign genes in a genome might well be a consequence, rather than a cause, of adaptation (Nesbø *et al*. 2001). Mira *et al*. (2001) viewed the bacterial genome as resulting from an evolutionary balance between ongoing recruitment and removal processes. Lateral gene transfer is the most important recruitment process. If there is no pressure from natural selection to maintain a newly recruited gene, it will quickly be removed or inactivated. Assuming that, at least in bacterial genomes, there is a bias towards a higher rate of deletions over insertions, this explains the small and streamlined genomes of many bacteria. According to G. Gottschalk (personal communication) the large

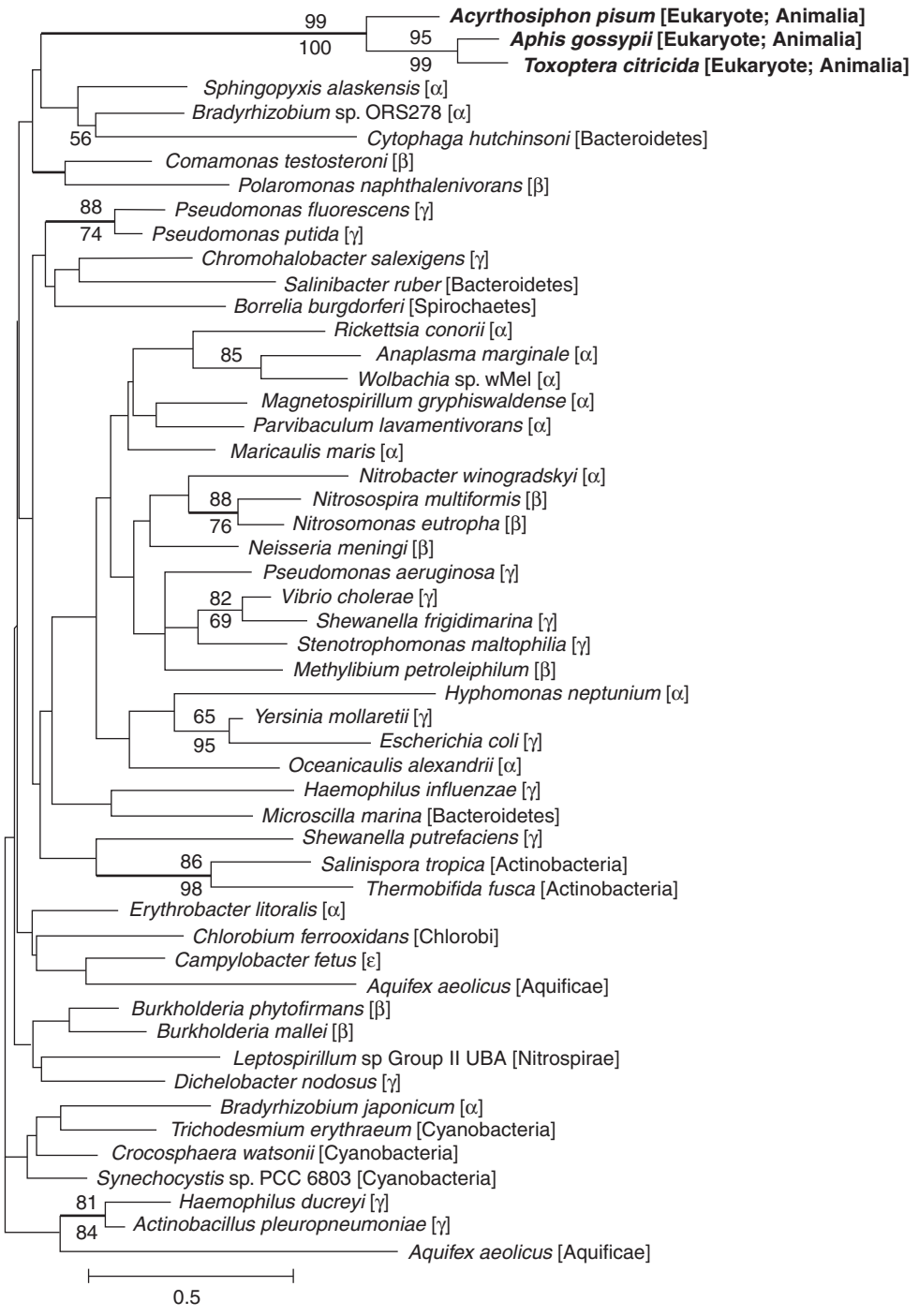


Figure 2.13 Phylogenetic tree of rare lipoprotein A (*rlpA*), based on 76 aligned amino acid sites. Thickened lines indicate Bayesian posterior probabilities greater than 95%. Bootstrap values are indicated for the (topologically identical) neighbour-joining tree (above) and maximum likelihood tree (below). All taxa are Bacteria, except *Acyrthosiphon*, *Aphis*, and *Toxoptera*, which are aphids. α, β, γ: protobacterial phyla. Afer Nikoh and Nakabachi (2009), with permission from Biomed Central.

genomes of methanogens may just be a consequence of the fact that removal of laterally transferred genes is a slow process in these organisms. This argument *a fortiori* holds for animals. The function of many laterally transferred genes in animals is not at all clear (e.g. in bdelloid rotifers, Gladyshev *et al.* 2008) and this may be seen as supporting the neutral point of view.

Are all genes equally subjected to lateral gene transfer? Jain *et al*. (1999) postulated the *complexity hypothesis*, which states that genes that have few interactions with other genes will integrate more easily into a new genomic background and are therefore more likely to be successful in lateral gene transfer, compared with genes that are part of a complicated network and dependent on many other genes. This hypothesis explains the greater tendency of operational genes to be transferred, compared to informational genes; however, there are many exceptions to the hypothesis (Zhou and Thompson 2004).

Although the actual rate of lateral gene transfer is considered to be low compared to the life cycle of microorganisms, lateral gene transfer may leave a permanent trace in the genome; the presence of DNA of 'foreign' origin dating from a transfer event millions of years ago is still visible to the present-day genome investigator. At the same time, lateral gene transfer complicates the construction of phylogenetic trees, because the phylogenetic reconstruction of one gene may be different from the reconstruction of another gene if one of them underwent lateral gene transfer. Even the classical gene used for prokaryote phylogeny, the 16 S rRNA gene, which is assumed to be less prone to lateral gene transfer than operational genes, has caused some problems. Phylogenies based on 16 S rRNA are not always consistent with those derived from another essential gene, RNA polymerase. Consequently, the phylogenetic history of a gene is not always a correct indicator of the phylogenetic history of the organisms themselves. In a much-discussed paper, Doolittle (1999) proposed a radical way out of this dilemma, which is to abolish the whole idea of a universal phylogenetic tree of life. Instead, Doolittle (1999) argued that the accepted taxonomic categories, for example Bacteria and Archaea, may be used as convenient descriptors of shared genes, but not as diagnostic indicators of common ancestry. Doolittle (1999) presented a sketch of early evolution in which the base of life is seen as a highly reticulate structure, a network of promiscuous gene exchange, from which the three main domains of life finally rise (Fig. 2.14).

Despite the fact that the importance of lateral gene transfer is now well established, the argument that it is an impediment to classifying prokaryotes is not necessarily true. Snel *et al*. (1999) developed a genome-based phylogenetic approach that goes one step further than just comparing the sequences of genes. These authors considered the fraction of genes shared between genomes as a measure of distance between two species, as follows:

$$d_{AB} = 1 - \frac{n_{AB}}{n_A}$$

where d_{AB} is the distance between genomes A and B, n_{AB} is the number of genes shared (using an arbitrary threshold level for orthology), and n_A is the number of genes in the smallest genome of the two. So in this method the phylogenetic distance between two species is characterized by a single parameter, not by as many parameters as there are shared genes. The phylogeny based on a matrix of

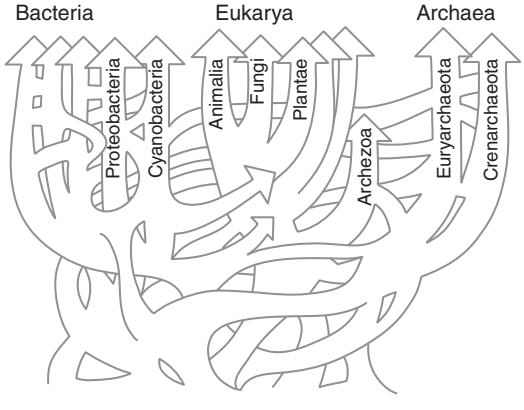


Figure 2.14 The network phylogeny of life which, according to Doolittle (1999), must replace the traditional phylogenetic tree to account for lateral gene transfer. Reprinted with permission from Doolittle (1999). Copyright 1999 AAAS.

pairwise distances calculated in this way was called a *genome phylogeny*. Snel *et al*. (1999) applied this approach to 12 fully sequenced prokaryotes and showed that the resulting tree was actually quite similar to the tree generated by the 16 S rRNA gene. The authors concluded that, despite lateral gene transfer, there is still a very strong phylogenetic signature in the gene content of prokaryotic genomes. However, we should realize that a genome phylogeny captures the central trend of evolutionary history, but does not provide the complete picture (Wolf *et al*. 2002).

2.2.3 From bacteria to organelles

The haploid, circular structure of prokaryotic genomes extends to the genomes of mitochondria and chloroplasts, which have a similar arrangement but are smaller due to loss of many genes. Gray *et al*. (1999) indicated that the most probable ancestor of the mitochondrion is to be found in the order Rickettsiales of the Alphaproteobacteria. Members of this group include various obligate intracellular parasites such as *Rickettsia* and *Wolbachia*. The species *Rickettsia prowazekii*, the causative agent of a form of typhus transmitted by lice, was long considered to have the most mitochondrion-like genome. However, when more members of the group were sequenced, such as *Wolbachia pipientis*, an obligate intracellular parasite of Diptera (Wu *et al*. 2004), doubt was cast on the grouping of mitochondria within the Rickettsiales. Still, evolutionary analysis supports the hypothesis that mitochondria share a common ancestor with the Alphaproteobacteria. Also, the common view remains that the collective mitochondrial genomes are monophyletic; that is, they all originate from the same ancestor.

The most bacterium-like mitochondrial genome is found in the flagellate *Reclinomonas americana* (Sarcomastigophora, Histionida). When the mitochondrial DNA (mtDNA) of this species was sequenced (Lang *et al*. 1997) it was considered by some as a missing link between bacteria and mitochondria, because of its unusually large number of genes (97, including all the proteins found in other sequenced mtDNAs). The organelle genome database (GOBASE), coordinated by the University of Montreal in Canada (gobase.bcm.umontreal.ca), has now collected 177 000 mitochondrial sequences and 41 000 chloroplasts (O'Brien *et al.* 2009). The mitochondrial genomes of protists probably hold the key to the evolution of the group, because they comprise most of the phylogenetic diversity within the eukaryotes (Gray *et al*. 1999).

Mitochondria obey the rule that endosymbiosis is accompanied by genome miniaturization (see Section 2.1). In the course of evolution, most of the genetic information for mitochondrial biogenesis and function has moved to the nuclear genome; the proteins needed by the mitochondrion are synthesized in the cytoplasm and then transported across the mitochondrial membrane. Still, the mitochondrial genome encodes several RNAs and proteins essential for mitochondrial function, mostly respiratory complexes of the electron transport chain such as NADH ubiquinone oxidoreductase, succinate ubiquinone oxidoreductase, ubiquinol cytochrome *c* oxidoreductase, and cytochrome *c* oxidase.

Comparisons among mitochondrial genomes are troubled by the fact that loss of genes has occurred many times independently. For example, genome miniaturization in the bacterial rickettsias has taken place independently of miniaturization in the mitochondrial lineages of protists. To complicate things further, some plant mitochondria contain genes that originate from chloroplasts; this holds for two tRNA genes in the mtDNA of *Arabidopsis*. The result is that the size and the gene content of mitochondria are remarkably divergent between the eukaryote lineages.

Gray *et al*. (1999) pointed out that mtDNAs come in two basic types, designated as ancestral and derived. *Ancestral mitochondrial genomes* (for example, the one from *Reclinomonas americana*) have retained clear vestiges of their eubacterial ancestry, with many non-animal genes, tightly packed in a genome with no or few introns. *Derived mitochondrial genomes* are characterized by substantial reduction in genome size, marked divergence of rRNA genes, and adoption of biased codon-usage patterns in protein genes. All metazoan and most fungal mtDNAs fall into this category.

The size of mitochondrial genomes varies between less than 6 kbp in *Plasmodium falciparum*

(the human malaria parasite, belonging to the phylum Apicomplexa) to more than 200 kbp in land plants. Gene content is similarly variable across species. In angiosperms, the mitochondrial genome has evolved to become recombinationally active, which has led to extensive rearrangements of genes, breaking up bacterial gene clusters, and loss of tRNA genes. The mitochondrial genome of *Arabidopsis* is among the largest of the eukaryotes, but it does not encode many more genes than some of the protist mtDNAs. An overview of the diversity of mitochondrial genome size and content is given in Fig. 2.15.

Interestingly, mitochondrial genomes may fragment into different molecules. This is thought to be due to recombination between repeat segments in different parts of the genome. An example is found

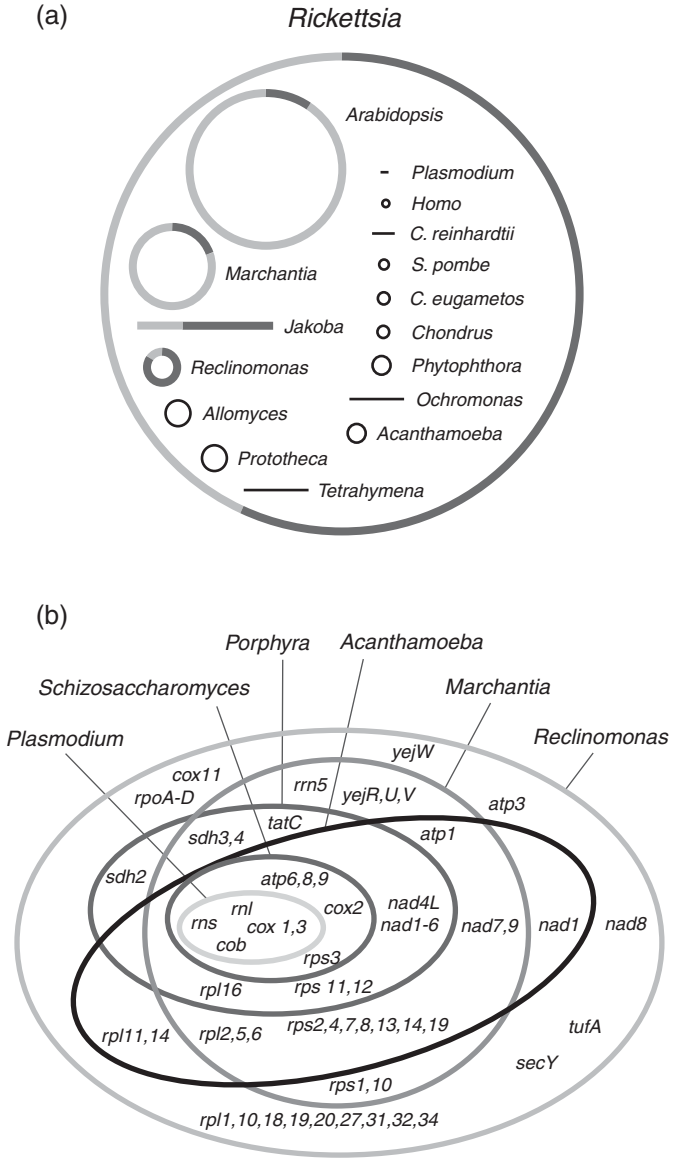


Figure 2.15 Genome size and gene content of mitochondrial DNAs across a wide range of species. (a) Circles and lines represent circular and linear chromosomes, with the ORFs of known function shown as dark lines. The major groups to which the species mentioned belong are as follows: *Rickettsia* (Alphaproteobacteria), *Arabidopsis* (angiosperm plant), *Marchantia* (liverwort), *Jakoba*, *Reclinomonas* (flagellates), *Allomyces* (fungus), *Prototheca* (green alga), *Tetrahymena* (ciliate), *Acanthamoeba* (amoeba), *Ochromonas* (golden alga), *Phytophthora* (oomycete), *Chondrus* (red alga), *Chlamydomonas eugamatos*, *Chlamydomonas reinhardtii* (green algae), *Schizosaccharomyces pombe* (yeast), *Homo* (human), and *Plasmodium* (malaria parasite). (b) Gene complement of mitochondrial genomes, showing the overlap between species. Each ellipse corresponds to one organism and includes all the mitochondrial genes of that organism. For example, the tiny mitochondrial genome of *Plasmodium* has four genes, which are also found in all other mitochondria, while all known mitochondrial genes are found in the mitochondrial genome of *Reclinomonas*. The genes are designated by code names. Reprinted with permission from Gray *et al*. (1999). Copyright 1999 AAAS.

in the potato cyst nematode, *Globodera pallida* (Tylenchida, Heteroderidae), which has a mitochondrial genome consisting of six different circular small chromosomes each 6.3–9.5 kbp. The twelve mitochondrial genes are scattered over the six units, but the ribosomal genes are all concentrated on one of them (Armstrong *et al*. 2000). Even more surprising is that the frequencies of these mitochondrial components differ between populations of the nematode. Such small mitochondrial genomes are called *subgenomic-sized mtDNAs*; they are also found in some green algae and higher plants. Another peculiar situation is due to the presence of plasmids inside mitochondria. Mitochondrial plasmids are especially ubiquitous among filamentous fungi and some of them cause a syndrome of growth loss and early senescence when inserted into ribosomal genes of the mitochondrial genome (Maas *et al*. 2005).

In addition to mitochondria, the other main eukaryotic organelle of bacterial origin is the chloroplast. Comparative studies on bacterial and chloroplast genomes have now demonstrated convincingly that chloroplasts are derived from a cyanobacterium related closely to the present species *Nostoc punctiforme*. It is also evident that chloroplast genomes jointly are one monophyletic group; that is, they all descend from the same ancestor. This is not to imply that the symbiotic event that gave rise to the chloroplast occurred only once. In fact, it is assumed that the initial inclusion of a cyanobacterium (which led to red algae, green algae, and higher plants) was followed by a second endosymbiotic event, probably involving a red alga, which produced cells in which the chloroplast was surrounded by four, rather than two, membranes. From this type of cell three different evolutionary lineages are assumed to have originated (Tudge 2000; Raven and Allen 2003; Falkowksi *et al*. 2004):

- Brown algae (Phaeophyta), diatoms (Bacillariophyta), golden algae (Chrysophyta), yellow green algae (Xanthophyta), and water moulds (Oomycota), with the chloroplast secondarily lost in the Oomycota.
- Dinoflagellates (Dinoflagellata), Apicomplexa (*Plasmodium* and other parasites), and ciliates (Ciliata), with a strong reduction of the chloroplast in apicomplexans and a complete loss of the chloroplast in ciliates.
- Kinetoplastids (parasites like *Trypanosoma*) and Euglenozoa, with secondary loss of the chloroplast in kinetoplastids and some euglenoids.

So, the present-day scattered distribution of photosynthetic capacity across the eukaryotes is explained not only by gains but also by losses of chloroplasts. Losses are especially prominent in the lines that obtained the chloroplast through double symbiosis. Different degrees of chloroplast reduction can be observed in the phylum Apicomplexa. These unicellular organisms are characterized by an organelle called an *apicoplast*, an assumed relict of a chloroplast, with a greatly reduced genome; however, not all species have this organelle. A genomic survey of the human parasite, *Cryptosporidium parvum*, which lacks an apicoplast, demonstrated the presence of 31 genes of likely cyanobacterial and other prokaryote origin in the genome, confirming the theory that apicomplexans evolved from a plastid-containing lineage (Huang *et al*. 2004).

The genome of a chloroplast typically encodes 60–200 proteins, which is more than an order of magnitude less than the genome of a cyanobacterium, which encodes at least 1500 proteins. Genes in the preplastid were either lost or transferred to the nucleus. Studies on *Arabidopsis* nuclear and chloroplast genomes (Martin *et al*. 2002) have shown that thousands of genes have been transferred. Some 4500 genes in the nuclear genome of *Arabidopsis*, or 18% of the genes, appear to have a bacterial (chloroplast) origin. This is not to say that the products of these genes are all functional in the chloroplast; actually, more than half of the originally chloroplastic genes are now targeted to other cell compartments. To complicate things further, the protein products of many nuclear genes that were not acquired from the plastid ancestor are now targeted to the plastid! In this complicated interplay between genomes, many issues remain unresolved, including the fundamental question of why chloroplasts have a genome at all, if genes can be transferred so easily to the nucleus (Raven and Allen 2003). There must be some crucial selective advantage in retaining some genes in chloroplasts, but not others.

2.3 Eukaryotic genomes

2.3.1 Protist genomes

Microbial eukaryotes, that is protists, can be considered to comprise all other eukaryotic lineages, including animals, plants, and fungi. The modern DNA-based phylogenetic trees of the Eukarya support a division of five main lineages (Simpson and Patterson 2008; Lane and Archibald 2008; Keeling *et al.* 2005):

- Unikonta, including slime moulds, amoebozoans, fungi, and animals;
- Rhizaria, including radiolarians, foraminiferans and cercozoans;
- Excavata, including euglenids and several parasitic lineages: kinetoplastids, diplomonads, and parabasalids;
- Archaeplastida, including glaucophytes, red algae, green algae, charophytes and higher plants;
- Chromalveolata, including haptophytes, oomycetes, brown algae, diatoms, cilates, dinoflagellates, and apicomplexans.

Sometimes not five but eight different supergroups are discerned and the precise evolutionary relationships among these 'highways' of eukaryotic evolution are unclear at the moment. Still, it is obvious that diversification of eukaryotes occurred very early, directly after the first eukaryotic cell with a mitochondrion appeared, and that endosymbiotic events including chloroplasts were a very important aspect of the evolutionary events. According to the most widespread view, four of the five highways (all except Unikonta) were once green, but many lost their chloroplasts later.

Given the fact that protists have diversified already for more than a billion years, tremendous differences in genome architecture between the lineages may be expected. This is indeed the case. Genome sequencing has elucidated many unusual and interesting genomic features in every group. An overview of these highlights is given in Table 2.4.

The recent sequencing of the genome of *Ectocarpus silliculosis*, a brown alga related to kelp (Cock *et al.* 2010), has shed some interesting light on the evolu-

Table 2.4 Some unusual features of protist genomes (after Pain *et al.* 2005; Katz and Bhattacharya 2008; Martens *et al.* 2008; Bowler *et al.* 2008; Moustafa *et al.* 2009; Haas *et al.* 2009; Cock *et al.* 2010)

Clade	Species analysed	Peculiar genomic features
Kinetoplastida	*Trypanosoma brucei, T. cruzi*, *Leishmania major*	Absolute strand polarity with long arrays of cotranscribed genes; conserved synteny between species
Lobosea	*Entamoeba histolytica*	High levels of repetitive sequences (rRNAs, tRNAs, transposons), no evidence of telomere repeats
Microsporidia	*Encephalitozoon cuniculi*, *Antonospora locustae*	Dramatic genome reduction, conserved arrangement of chromosomes and conservation of gene order
Apicomplexa	*Plasmodium falciparum, P. chabaudi, P. yoelli, P. berghei, P. vivax, Cryptosporidium parvum, C. hominis*	Greatly reduced genome, loss of several metabolic pathways, core function genes located in chromosome centres, species-specific genes at chromosome ends often associated with antigenic variation, many genes undergoing immune selection
Oomycetes	*Phytophthora infestans, P. sojae, P. ramorum*	Large interspecies variation of genome size, blocks of conserved gene order with high gene density separated by variable regions of high repeat content, rich population of transposons in *P. infestans*, many new gene families and expanded disease effector gene families
Diatoms	*Thalassiosira pseudonana*, *Phaeodactylum tricornutum*	Rich in metabolic genes, intraspecific genomic variation, genomic footprints of several cryptic endosymbiotic events, involving green algae and red algae
Brown algae	*Ectocarpus silliculosis*	Many introns per gene, complex photosynthetic apparatus, receptor kinases (involved in evolution of multicellularity) different from plants and animals
Slime moulds	*Dictyostelium discoideum*	High gene-density, rich in simple sequence repeats and transposable elements

tion of multicellularity. Most evolutionary biologists agree that multicellularity has arisen several times in the evolution of protists: in fungi, animals, red algae, green algae, higher plants, and brown algae. A partial form of multicellularity is seen in slime moulds. Multicellularity has also been lost several times after it had been gained, most frequently within the fungi, leading to unicellular fungal forms (yeasts). A key group of genes associated with multicellularity are the membrane-spanning receptor kinases. These proteins have been shown to play key roles in developmental processes such as cell differentiation and cellular patterning. The animal receptor kinases are tyrosine kinases, and they form a monophyletic lineage separate from the plant receptor kinases, which are serine/threonine kinases. Interestingly, the *Ectocarpus* receptor kinases also form a monophyletic group, again separate from the animal and plant receptor kinases (Cock *et al.* 2010). This is consistent with the idea of multiple evolution of multicellularity from different unicellular ancestors.

Comparative genomics of microbial eukaryotes is only just beginning. It is hampered by the fact that so little is known about their biology. In addition, some protists have extremely large genomes, while others, especially the parasitic forms, have extremely small genomes. This, added to the frequent and ancient lateral gene transfers associated with organelle symbioses makes the elucidation of protist genome evolution a daunting task (Katz and Bhattacharya 2008). Genome sequencing and analysis is highly biased towards the many protist parasites with small genomes (*Plasmodium*, *Toxoplasma*, Microsporidia). Still, microbial eukaryotes include many representatives of great ecological importance, with major influences on the global nutrient cycles, such as coccolithophorans (Haptophyta), dinoflagellates, diatoms, and brown algae. The greatest promise for ecological genomics is most likely to be found in the large and diverse group of chromalveolates.

2.3.2 Yeast and other fungi

Fungi have received a great deal of attention from genome researchers. Hundreds of species are being sequenced and for many species a complete genome assembly is available. Fungi are producers of many important biologically active products, including enzymes and antibiotics, and play key roles in many bio-catalysed processes such as the production of bread, beer, wine, and cheese. Therefore many fungi have been sequenced to reveal the cell factory responsible for the production of these compounds and to possibly enhance their use in biotechnology.

Among the fungal genomes, the ascomycete order Saccharomycetales is best covered, because the exploration of fungal genomics began in that group with the model species *Saccharomyces cerevisiae*. The unicellular mode of life, ease of culturability, and the relatively small genome were important factors in the success of investigating baker's yeast. Probably more is known about yeast genomics, biochemistry, and physiological responses to environmental conditions than about any other species of eukaryote.

Although yeasts can be isolated from many natural habitats (fruits, leaf surfaces, decaying plants, soil), they are not common objects of study among ecologists. Spencer and Spencer (1997a, 1997b) provide a general overview of the biology of yeasts, with particular emphasis on their taxonomy and the type of natural habitat colonized by them. The designation yeast is not a proper taxon within the fungal kingdom, it just indicates that we are dealing with a unicellular fungus that can grow in colonies by budding or fission, like bacteria. Yeasts are contrasted with filamentous, mycelium-forming fungi. Although most yeasts are classified in the Ascomycota, they are also found in the Basidiomycota and Deuteromycota. *S. cerevisiae* is a yeast associated with bread preparation. The closely related *Saccharomyces* used in beer and wine production are considered to be separate species (Spencer and Spencer 1997a).

Table 2.5 lists some species of sequenced fungi of interest to ecologists. Fungi play an important role in ecosystems as degraders of recalcitrant organic matter. An example of this category is *Phanerochaete chrysosporium*, or white rot fungus, which is found commonly on dead trees and wood fragments on the forest floor (Martinez *et al*. 2004). White rot owes its name to the fact that the fungus 'bleaches' wood

by degrading the (brown-coloured) lignin, rendering the white cellulose visible. Lignin forms protective sheaths around cellulose fibrils in plant cell walls. The biodegradation of lignin by white rot is supported by unique extracellular oxidative enzymes (peroxidases and oxidases) that act non-specifically via the generation of free radicals attacking the lignin molecule. There are many other ecologically important fungi commonly found on dead organic material that contribute to processing of organic matter and cycling of nutrients in natural ecosystems. The mushroom fungus *Schizophyllum commune* and the ascomycete *Trichoderma reesei* are two other organic matter degraders of which the genome is now sequenced.

Of particular interest to community ecologists are fungi engaging in symbiotic relationships with plant roots. These symbioses come in two different types: *ectomycorrhizas* and *arbuscular mycorrhizas*. The function for both associations is based on the exchange of sugars from plant to fungus and the exchange of phosphorus and nitrogen from fungus to plant. Despite the fact that the ecological benefits in these associations are similar, the structural and molecular organization is very different between the ectomycorrhizas and the arbuscular mycorrhizas.

Ectomycorrhizas are associated with the roots of trees and are mostly found among basidiomycetes although some ectomycorrhizal fungi, such as black truffle, are actually ascomycetes. Ectomycorrhizal symbiosis is a crucial component of forest ecosystems and receives extensive study in the framework of sustainable forestry. The genome of *Laccaria bicolor* provided various new insights into the genes and gene families likely associated with symbiosis (Martin *et al.* 2008; Martin and Nehls 2009). Research is aimed towards the definition of a '*symbiosis toolkit*', the collective genome characteristics that are crucial to mycorrhizal formation and function. In the *Laccaria* genome a battery of genes was identified encoding effector-type small secretory proteins (SSPs), several of which are only expressed in symbiotic tissues. These mycorrhiza-induced SSP

Table 2.5 List of fungal species of ecological importance with completely sequenced genomes

Species	Genome size (Mbp)	Growth form, habitat, ecological importance
Ascomycota, Saccharomycetales		
Saccharomyces cerevisiae	12	Baker's yeast, important physiological model organism
Ascomycota, Sordaliales		
Magnaporthe grisea	40	Filamentous, causes blast disease in rice and other domesticated grasses
Neurospora crassa	43	Filamentous mould, important model for genetics and molecular biology
Trichoderma reesei	34	Mycelium abundant in forest soil, general degrader of biomass
Ascomycota, Pezizales		
Tuber melanosporum	125	Ectomycorrhizal symbiont with oak and hazel trees
Glomeromycota, Glomerales		
Glomus intraradices	16	Model for arbuscular mycorrhizae of many forbs and grasses, evolution of mutualisms
Basidiomycota, Ustilaginales		
Ustilago maydis	21	Pathogen of maize, model for smut fungi
Basidiomycota, Aphyllophorales		
Phanerochaete chrysosporium	30	White rot, in fallen trees, on forest floor, degrades lignin
Basidiomycota, Agaricales		
Schizophyllum commune	39	White rot, degradation of branches and timber of deciduous trees
Lacccaria bicolor	65	Ectomycorrhizal symbiosis with conifers and deciduous trees, crucial role in tree growth

Sources: www.genomenewsnetwork.org, www.ebi.ac.uk/genomes/index.html, www.genomesonline.org, Kellis *et al.* (2003, 2004); Kämper *et al.* (2006); Martinez *et al.* (2008); Martin *et al.* (2008, 2010); Ohm *et al.* (2010)

genes may be associated with the control of metabolic pathways in the host plant, or the suppression of plant defence mechanisms. They may play a decisive role in the establishment of the symbiotic relationship between fungus and plant root, although further comparative analysis must confirm whether they are just specific to *Laccaria* or to all ecotomycorrhizal associations. Other gene families that have undergone expansion in *Laccaria* are GTPases, proteases and transporters. These families likewise contain candidate genes for the 'symbiosis toolkit'.

Interestingly, *L. bicolor* lacks the enzymes necessary to degrade plant cell walls and so obviously cannot grow as a decomposer of dead plant remains; it is highly dependent on its host for carbohydrate provision. The nutrient provision it delivers in return is mainly focused on nitrogen, including nitrogen of animal origin. *L. bicolor* has genes to import a wide spectrum of inorganic and organic nitrogen sources found in forest soils.

Further comparative analysis of ectomycorrizal fungi must show whether there are commonalities across species in the fungal gene repertoire underlying symbiosis. Inspection of the genome of the Périgord black truffle fungus, *Tuber melanosporum*, suggests that symbiogenesis has evolved in a quite different way in this ascomycete (Martin *et al.* 2010). First of all, the *T. melanosporum* genome contains fewer genes than *L. bicolor* (only 7500, versus 20 614 in *L. bicolor*), despite the fact that it is tremendously expanded and has by far the largest genome size of any fungus (125 Mbp), which is due to proliferation of multicopy transposable elements. Furthermore, *T. melanosporum* does induce carbohydrate-cleaving enzymes in symbiotic tissues, which suggests that it degrades plant cell walls as part of the colonization of plant roots. These divergencies at the molecular level again illustrate that similar ecological outcomes can be brought about by very different genomic mechanisms.

The completion of the genome of the arbuscular mycorrhizal fungus *Glomus intraradices* has been characterized as a 'long hard road' (Glomus Genome Consortium 2008). This is due to the fact that the concept of an 'individual' is unclear in this fungus. The mycelium of *Glomus* is multinucleate and individual nuclei can move within the mycelium and also between hyphae of different individuals upon *anastomosis* (non-sexual fusion of branches of different mycelia). Consequently, different nuclei from the same mycelium may not have the same genotype, a condition that seriously disturbs interpretation of genome sequencing. A high level of polymorphism is present in the *Glomus* genome, such that some regions of the genome appear as different haplotypes in the assembled scaffolds. To complicate things further, the genome is rich in short repeated sequences that increase the chance of misassemblies and are easily mistaken for different alleles.

When finally available to the ecological community, genomic resources such as genotyping methods and gene expression arrays for *Glomus* will be of tremendous importance. It is estimated that 200 000 species, two-thirds of all terrestrial plants, rely to a greater or lesser extent on arbuscular mycorrhizal (AM) symbiosis. AM fungi therefore make a very large contribution to global nutrient cycles for carbon and phosphorus and they are considered the 'mother of plant root endosymbioses' (Parniske 2008). In addition, AM fungi, being accessible to experiments under controlled conditions in greenhouses, are an important model for the establishment of mutualistic relationships in general (Kiers *et al.* 2010). Mutualistic stability can be established only if cheating by one of the partners can be detected by the other, and punished. Such mechanisms indeed are operating during AMF-plant symbioses (Kiers and Van der Heijden 2006).

While the genome of *S. cerevisiae* is only slightly larger than the average prokaryote genome, the genomes of other fungi are generally in between those of prokaryotes and multicellular eukaryotes, that is ten times larger than the average bacterial genome (Table 2.5). The question can be asked: why does baker's yeast have such a small genome? Braun *et al*. (2000) suggest that this is due to a process of 'streamlining' by loss of genes. Detailed comparisons of *S. cerevisiae* with the filamentous fungus *Ashbya gossypii* (Dietrich *et al*. 2004) and the unicellular *Kluyveromyces waltii* (Kellis *et al*. 2004) have confirmed that the evolution of *S. cerevisiae* has included a whole-genome duplication, followed by

extensive rearrangements and loss of genes. These conclusions are in accordance with the idea that unicellularity in fungi is an apomorphic condition and that yeasts evolved independently several times from multicellular ancestors, not the other way around.

Neurospora crassa (orange bread mould), a mycelium-forming fungus, was sequenced shortly after *S. cerevisiae* (Galagan *et al*. 2003). *N. crassa* has been a famous model organism for genetic and biochemical studies since the classical experiments of Beadle and Tatum in 1942, who proved that for every enzyme there is one gene (the *one gene/one enzyme hypothesis*). Although *Neurospora* is an ascomycete like *S. cerevisiae*, it is more similar to animals than to yeast in several ways. For example, unlike yeast but like mammals, it has a clearly discernable circadian rhythm, it methylates DNA to control gene expression, and it has complex I in the respiratory chain. With all this biochemical research, *Neurospora* is the best-characterized of the filamentous fungi, but its ecology remains relatively unexplored. Being moderately thermophilic, *Neurospora* was thought to occur mainly in moist tropical and subtropical regions, but recent surveys have also found *Neurospora* colonizing trees and shrubs killed by wildfires in temperate regions (Jacobson *et al*. 2004; Fig. 2.16). Mycological inventories showed that in North America isolates were comprised predominantly of a single species, *Neurospora discreta*, but in southern Europe species collected included *N. crassa*, *N. discreta*, *N. sitophila*, and *N. tetrasperma*.

The life cycle of *S. cerevisiae* is of the *diplobiontic* type; that is, it cycles through two distinct phases, one diploid and one haploid (Fig. 2.17). The diploid stage grows vegetatively by budding off new cells and forming colonies. Under certain conditions, such as deprivation of carbon or nitrogen, it can form stress-resistant ascospores. Exactly how *sporulation* is triggered is currently under investigation. Ascospores are of two types, called *a* and α. They sit together in the ascus and upon germination produce two so-called *mating types*. These haploid cells can grow vegetatively by budding, a property that provides unique opportunities to geneticists, because the expression of traits in this stage is not confounded by dominance: the phenotype is a direct result of the genotype. The two haploid mating types may interact with each other by means of hormones, which induce a characteristic change in shape, leading to pear-shaped cells, called *shmoos* after the lovable creatures in Li'l Abner's comic strip from 1948. The process of *sexual conjugation* can occur only between opposite mating types. It involves a complicated series of cell-surface changes to facilitate fusion and is mediated by the hormones in a manner that is mating-type specific. The life cycle of *N. crassa* is similar, except that the diploid vegetative stage is suppressed and the zygote proceeds directly to form an ascus.

SGD™ is the Saccharomyces Genome Database (www.yeastgenome.org) where information about the molecular biology and genetics of baker's yeast is filed and presented to the world. The database includes a variety of search options that allow one to consult the genome sequence; analysis tools such as BLAST, programs for homology searches, and information about protein structure, as well as contact details for more than 1000 people in the yeast research community. The database also provides a list of recently published papers on all aspects of yeast molecular biology and links to databases of the other fungi. The organizational principles of the database were discussed by Dwight *et al*. (2004). Over the years the SGD has seen a dramatic increase in its usage, and has served as a template for other databases. The success of SGD, as measured by the numbers of pages viewed, user responses, and number of downloads, is due in large part to the network philosophy that has guided its mission and organization since it was established in 1993. The yeast genome was the first for which microarrays were developed. Expression arrays allowing one to address all genes of both *S. cerevisiae* and *Schizosaccharomyces pombe* at the same time are now commercially available.

A series of in-depth comparative genomic studies has been conducted in which the genome of *S. cerevisiae* was compared with other fungi in and outside of the order Saccharomycetales (Brachat *et al*. 2003; Cliften *et al*. 2003; Kellis *et al*. 2003, 2004; Dietrich *et al*. 2004; Dujon *et al*. 2004). An important argument for sequencing species that have a known

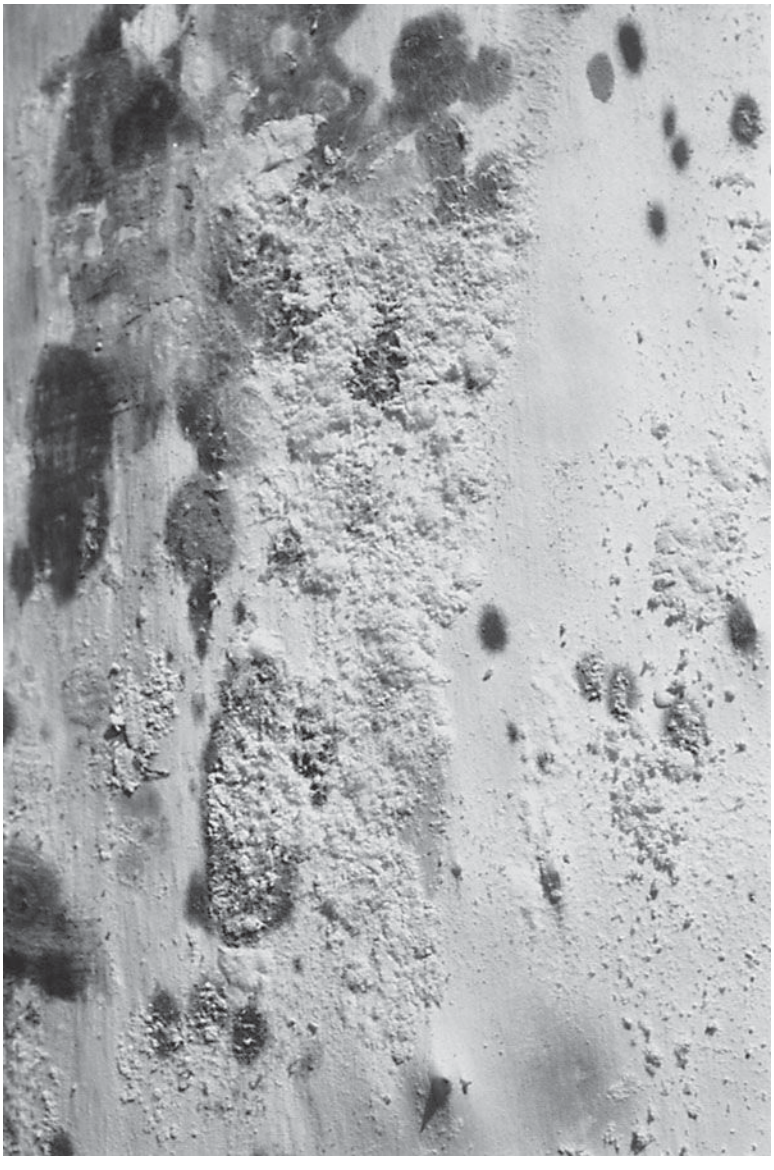


Figure 2.16 *Neurospora* growing under tree bark. Courtesy of D.J. Jacobson, Stanford University.

phylogenetic relationship with *S. cerevisiae* is that it could help in the identification of genes and regulatory elements in the genome of *S. cerevisiae*. Kellis *et al*. (2003) analysed the relationship among orthologous genes using a *reading frame conservation test*. This test classifies each ORF in *S. cerevisiae* as biologically meaningful or meaningless, depending on the proportion of the sequence over which conservation with other species is observed. Each of the other species was considered a 'voter', 'approving' or 'rejecting' the sequence in *S. cerevisiae*. Obviously, the procedure carries a risk that true genes under strong selective pressure in one of the species are rejected as biologically meaningless, but this was

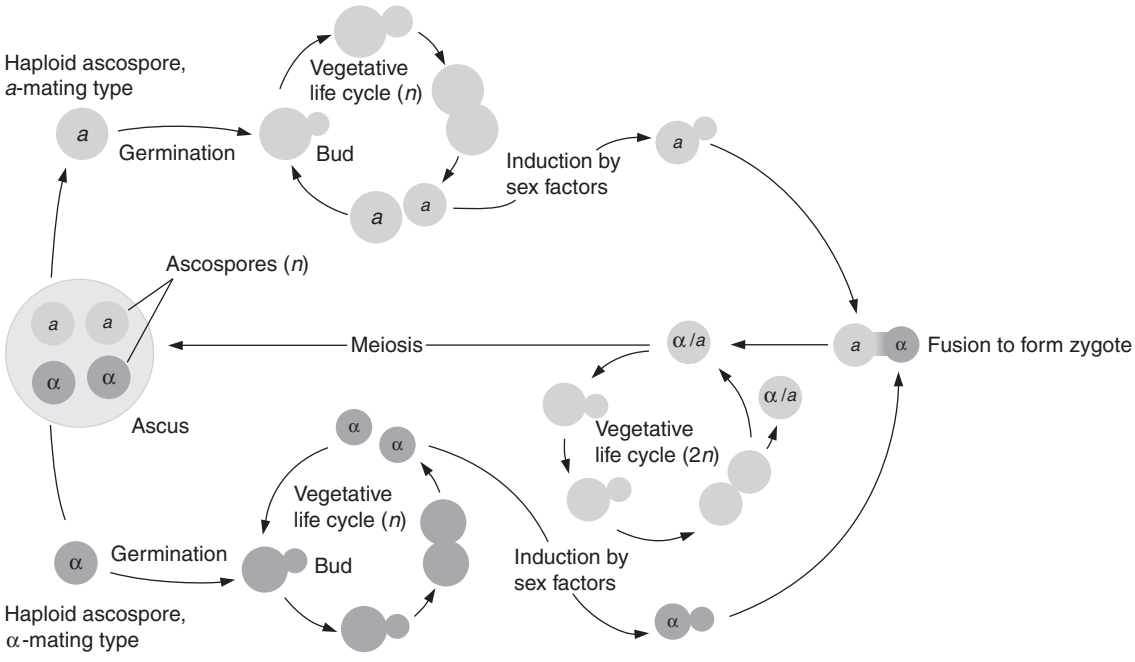


Figure 2.17 Life cycle of the yeast *S. cerevisiae*. After Russell (2002), with permission from Pearson Education Inc.

prevented by looking in detail at each rejection. Confidence in the method was increased when it appeared that only a few already annotated ORFs were rejected as genes. Inspection showed that in all of these possibly false-negative cases the annotated ORFs were indeed likely to be spurious. The analysis of Kellis *et al*. (2003) pruned the yeast gene catalogue of 503 genes, leaving only 20 unresolved ORFs in the database and decreasing the number of protein-encoding genes with more than 100 amino acids to 5538.

Further insight into the *S. cerevisae* genome was obtained from comparisons with more distantly related species. Dujon *et al*. (2004) sequenced four species from the hemiascomycete group, *Candida glabrata*, *Kluyveromyces lactis*, *Debaryomyces hansenii*, and *Yarrowia lipolytica*, and compared their genomes with that of *S. cerevisiae*. A total of approximately 24 200 novel genes was identified, and their translation products were classified into about 4700 families. Pairwise comparisons were made between the species to establish the degree of sequence divergence between orthologous genes. It appeared that the five yeast species together spanned a genetic diversity comparable to the entire phylum Chordata. For example, the average sequence identity between orthologous genes (translated into proteins) between *S. cerevisiae* and *C. glabrata* was 65%, between *S. cerevisiae* and *K. lactis* 60%, and between *S. cerevisiae* and *Y. lipolytica* 49%. This is less than the average sequence identity of proteins between mouse and fugu fish (70%) and comparable to that found between the urochordate sea squirt, *Ciona intestinalis*, and the mammals! The lesson of this large-scale comparative genomics study was that the evolutionary distance between yeasts, despite their very similar morphology, is extremely large.

2.3.3 Nematodes

Although molecular biologists have developed the habit of calling the nematode *C. elegans* a 'worm', the animal has nothing to do with the true worms, the Annelida, since it is classified in a completely separate phylum, Nematoda. Phylogenomic analysis has demonstrated that this phylum is related to

the arthropods and belongs to the so-called moulting animals, the Ecdysozoa (Dopazo and Dopazo 2005); the annelids are classified with the molluscs in another superphylum, Lophotrochozoa.

C. elegans is one of the rhabditid nematodes, a group of tiny, free-living, bacteria-feeding animals, living in soils, dead organic material, or wherever there are bacteria. On the basis of rRNA gene sequences, 23 species are classified in the genus *Caenorhabditis*, including *C. briggsae*, the other nematode whose genome has been sequenced completely. Classification of nematodes is complicated by the fact that the external structure of the animals is not very diversified. The morphology of the most important diagnostic characters, the mouthparts, and other aspects of external morphology do not always fit with the molecular data and therefore the names assigned to higher-order categories in the classical taxonomy are sometimes illogical when arranging the species according to a molecular phylogeny. For example, the order Rhabditida does not indicate a monophyletic group, but appears to fall into at least two different phylogenetic lineages (see http://nematol.unh.edu/phylogeny.php, and Blaxter *et al*. 1998).

Despite their morphological uniformity, the phylum Nematoda is extremely diverse from a genetic point of view. Analysing a large collection of ESTs (>250 000) from 30 different nematode species, Parkinson *et al*. (2004) found that 30–50% of the transcriptome of each species was unique to that species. Consequently, a single nematode like *C. elegans* can reveal only a small fraction of the genomic diversity of even its own phylum. A phylogeny of 53 species of nematodes, based on small-subunit rRNA sequences, is given in Fig. 2.18 (Blaxter *et al*. 1998). The figure also provides information on feeding habits, which diverge widely within the nematodes as a whole; one can find bacteriovores (like *C. elegans*), fungivores, predators, omnivores, plant parasites, and animal parasites. Figure 2.18 shows that there is no phylogenetic conservation of feeding habits; feeding modes are scattered throughout the tree. The great biodiversity of feeding habits, life-history patterns, and colonizing capacity makes the Nematoda a very suitable group for community bioindication. When each species is given a score on a scale of colonizers to persisters, the weighted sum of these scores for a given community (the *maturity index*) can be used as an indicator of habitat quality (Bongers and Ferris 1999).

The phylum Nematoda includes many parasites, such as the well-known intestinal roundworm of pigs, *Ascaris lumbricoides*, the small human pinworm *Enterobius vermicularis*, which infects 30–80% of schoolchildren in western countries, and various species causing serious diseases in tropical countries, such as *Onchocerca volvulus*, the causative agent of river blindness (onchocerciasis), which is spread by the bite of an infected blackfly. Experiments have shown that the inflammatory response in the human eye causing blindness, which is triggered by the presence of dying nematode microfilariae, is not only due to the worm itself, but also to toxins excreted by an endosymbiotic *Wolbachia* bacterium (Saint André *et al*. 2002). So the genomic studies on *C. elegans* have important ramifications for parasite research (Blaxter *et al*. 1998) and scientific networks are currently in development that address the field of nematode parasitomics: for example, the Filarial Genome Network (http://xyala.cap.ed.ac.uk/research/nematodes/fgn/filgen1.html), named after one of the parasitic species in the onchocercid group, *Filaria martis*. Although the interest in parasite genomics is exclusively medical at the moment, parasites are important agents in the population dynamics of many wild species and progress in the medical sector could well have a future spin-off to ecology. There are also several nematodes that form cysts in plant roots or otherwise damage below-ground plant tissues. These plant-parasitic nemotodes are also subject to intense genome investigations. The genome of the root-knot nematode *Meloidogyne incognita*, a parasite of tomato, cotton, and coffee, is already available (Abad *et al.* 2008), while final genome assemblies for the potato cyst nematodes *Globodera pallida* and *G. rostochiensis* are expected to be released soon.

C. elegans is a 1 mm-long, transparent animal with *sequential hermaphoditism* and *self-fertilization*. Sperm cells are made first and stored in the spermathecae. Then the animal switches to the production of oocytes, which are fertilized by sperm from the

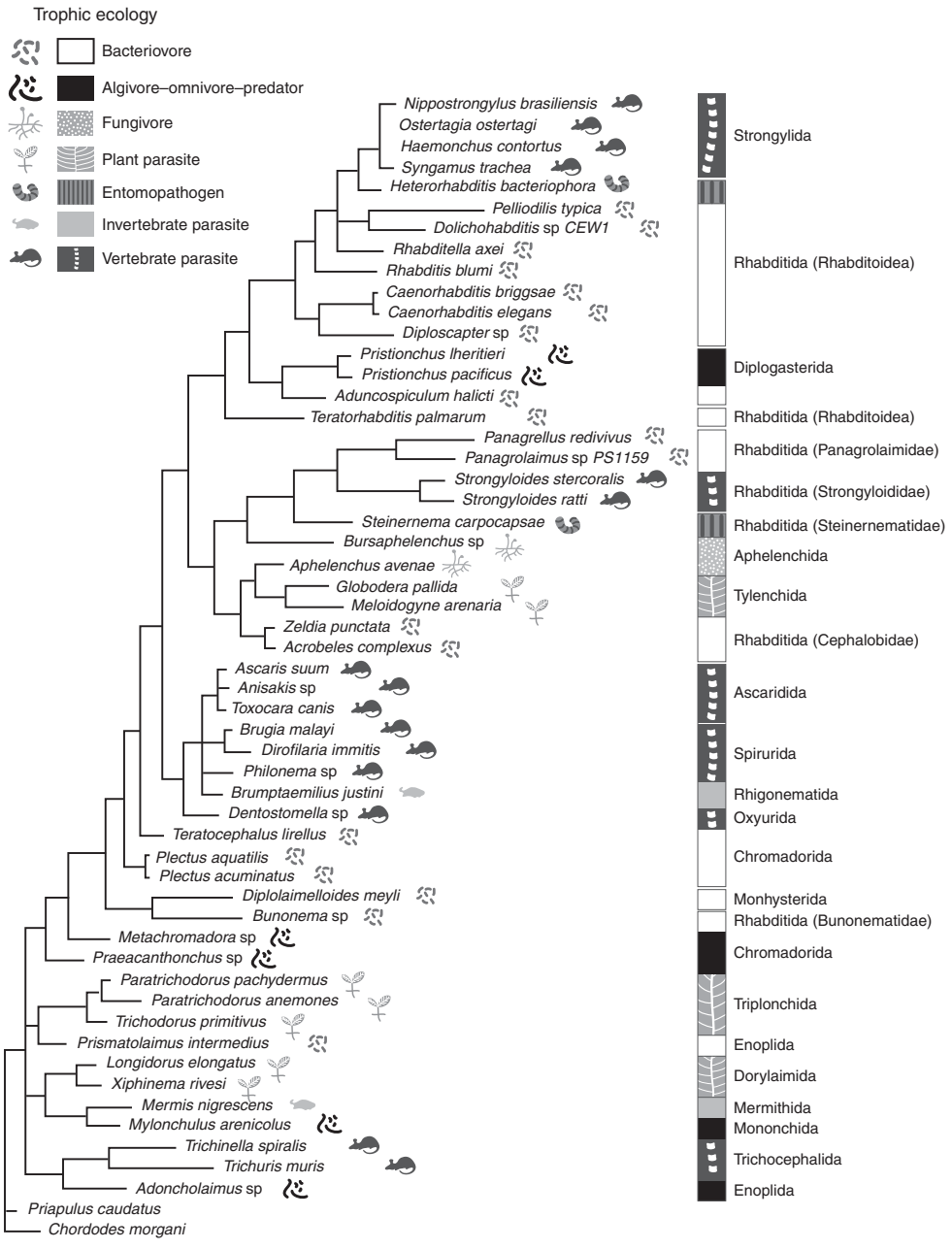


Figure 2.18 Phylogeny of 53 nematodes, based on sequences of the small-subunit rRNA gene. With each species an indication of the feeding habit is given (see key). The right-hand side indicates the orders of classical nematode taxonomy. After Blaxter *et al*. (1998), by permission of Nature Publishing Group.

same individual, mature partly while still in the body, and develop to the first larval stage, which emerges after the eggs have been laid. The time from egg to egg-laying adult is 5.5 days at 15 ºC and 3.5 days at 20 ºC. Around 10 eggs are in the body at any time, but the animal can produce more than 300 progeny during its adult lifetime, which may take another 14 days. This phenomenal reproductive capacity, within hours to days, was an important consideration when the species was chosen as a model. In addition to hermaphrodites, males sometimes occur. These males fertilize the hermaphrodites, as there are no gonochoristic females. The sex-determining system is chromosomal and the males lack one sex chromosome (hermaphrodites are XX, males are X0). The possibilities offered by this type of reproductive cycle are very convenient for genetic work, because clones can be made from hermaphroditic lines with no signs of inbreeding depression, and males can be used for cross-fertilization.

The life cycle includes four larval stages, each separated by a moult (Fig. 2.19). The development of gonads and the production of sperm are already taking place during the larval stage. In addition to the four normal larval stages there is a resting stage, called the *dauer larva* (after the German word for endure), which is actually an arrested third larval stage, in which the animal goes into a state of torpor and does not eat, although it can move slightly and may live for several months. The dauer larval stage is induced by adverse conditions such as crowding and food scarcity. When terminated, the dauer stage proceeds to the fourth larval stage. It is assumed that the dauer larva is the nematode's dispersal stage. Several features point towards an increased propensity to be transported, either by wind or by other animals. The dauer dries itself out, secretes an extra cucticle and develops a behaviour known as nictation (winking); it tends to crawl up objects that protrude from the surface, stands on its tail, and waves its head back and forth (Riddle 1988). An important aspect of the dauer is its extreme longevity, which may last several months rather than the normal 20 days. The dauer stage of *C. elegans* is an important model for investigating the genomic basis of longevity (see Chapter 4).

A very peculiar property of *C. elegans* is that it has a completely determinate developmental pattern, which is fixed for all 959 cells of the body. This was the reason why the animal was initially chosen as a model for developmental studies by Sydney Brenner at the beginning of the 1960s (Brown 2004). Brenner was inspired by earlier German work on the nervous system of the intestinal parasitic nematode *Ascaris suum*, which had shown that the fate and location of each cell could be traced, and was the same in all individuals. The original *C. elegans* strain, on which the research in Cambridge was started by Brenner, came from the laboratory of Ellsworth Dougherty in Berkeley, who had cultured *C. elegans* for several years. It is assumed that the culture actually originated from mushroom compost collected near Bristol, UK (Fitch and Thomas 1997).

C. elegans is a cosmopolitan species. More than 20 different strains have been isolated from soils of North America, Europe, and Australia (Fitch and Thomas 1997). Despite this broad distribution, the species is not a popular object of study among field ecologists, because it is very difficult to distinguish from other species in the same group and its distribution seems to be restricted to synanthropic habitats such as compost heaps and manure. The natural genetic variation in outdoor *C. elegans* populations remains unexplored to date, although it might be a valuable source of new hypotheses on the regulation of genetic pathways (Kammenga *et al.* 2008).

The complete genome sequence of *C. elegans* was the first to be published for a multicellular organism (the *C. elegans* Sequencing Consortium 1998). The WormBase consortium has continued to edit the sequence, brought the estimated error rate down to 1 in 100 000, and closed the last gap in November 2002. This makes the *C. elegans* sequence the first and so far only metazoan genome database that has reached sequencing closure for all of the chromosomes. The interface on the World Wide Web (www.wormbase.org), described by Harris *et al*. (2004), offers a rich source of information, not only on the complete genome sequence but also on mutant phenotypes, genetic markers, developmental lineages of the worm, and bibliographic resources, including paper abstracts and author contact information. The

genome sequence of the related species, *C. briggsae*, is now completely integrated into WormBase, which allows comparative analysis of orthologues and synteny. WormBase also contains extensive information from large-scale genome analyses, microarray expression studies, and the assignment of gene ontology terms to gene products. New data releases are published regularly and from time to time a 'freeze' of the software and the database is deposited, which can be downloaded.

The *C. elegans* sequence was announced in 1998 as a 'platform for investigating biology'. The consortium realized that the importance of the genome sequence of *C. elegans* extended beyond nematodes proper, and in fact could be considered the basic formula for constructing a multicellular animal, in the same way that the sequence of the *S. cerevisiae* genome contains all the information for making and maintaining a unicellular eukaryote. In addition, because nematodes branched off early in the evolutionary tree of life, the *C. elegans* sequence was considered an outgroup for all other bilaterian animals, from Annelida to Chordata. (Hodgkin *et al*. 1995). However, that view proved to be wrong when representatives from the annelid worms were sequenced, which have retained the ancestral bilaterian features to a greater extent than nematodes (Miller and Ball 2009). An overview of the specific features of the *C. elegans* genome is given in Table 2.6.

At the time of publication of the *C. elegans* genome sequence, information was available on only a few

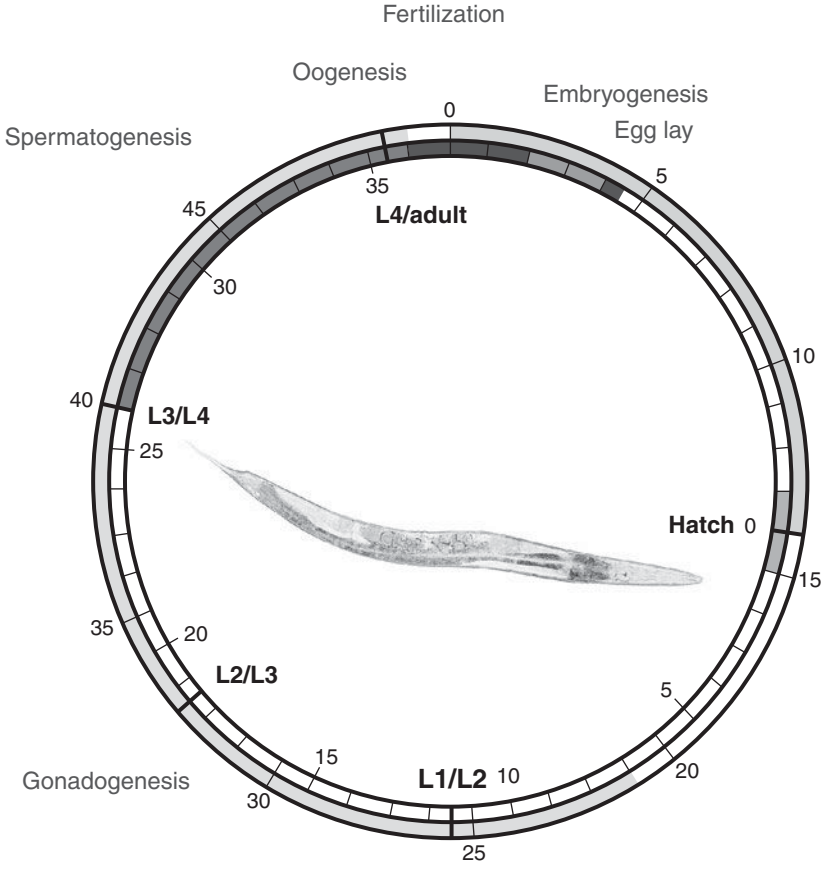


Figure 2.19 Life cycle (from egg to adult) of *C. elegans*, when cultured in the laboratory with abundant food (*E. coli*) at 25 ºC. The outer scale is marked in hours since fertilization, the inner scale in hours since hatching. L1, L2, L3, and L4 are the first to fourth larval stages. The adult can live for several days more. After Wood *et al*. (1980), with permission from Elsevier.

Table 2.6 General features of the genome of *C. elegans* (*C. elegans* Sequencing Consortium 1998)

Category	Features
Protein-encoding genes	Many genes (25%) organized in cistronic units (Section 2.2); three times more genes than in yeast (see Table 2.1); more genes than was estimated from genetic studies
RNA genes	Many tRNAs on the X chromosome; several RNA genes in introns of protein-encoding genes; rRNA genes in long tandem arrays
Gene density	Uniform GC content (see Fig. 2.7); fairly constant gene density across chromosomes
Repetitive DNA	Tandem repeats account for 2.7% of the genome; inverted repeats account for 3.6% of the genome; repeat sequences overrepresented in introns; 38 different families of dispersed repetitive sequences associated with transposition; dispersed repetitive sequences abundant on the arms of the chromosomes

other eukaryote genomes. *C. elegans* could be compared with yeast and bacteria such as *E. coli*, as well as to the then-available gene content of the human genome. It turned out that 36% of the predicted *C. elegans* genes had a human homologue and that no less than 74% of the human genes had a homologue in *C. elegans* (Fig. 2.20). The similarity of *C. elegans* to *Homo sapiens* was greater than that to yeast or bacteria. This comparison demonstrated for the first time the striking unity that underlies the genomes of organisms as different as nematode and human. This tendency was reconfirmed many times when more eukaryotic genome sequences became available.

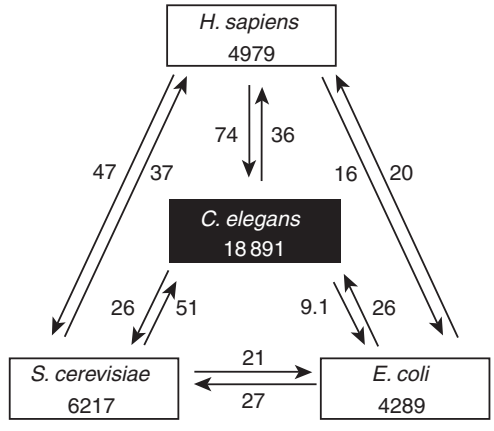


Figure 2.20 Pairwise comparison of predicted proteins in four species. The numbers adjacent to the arrows indicate the percentage of proteins in an organism that has a match in the organism indicated by the arrow. The numbers in the boxes indicate the actual number of proteins compared. Reprinted with permission from The *C. elegans* Sequencing Consortium (1998). Copyright 1998 AAAS.

2.3.4 *Drosophila* and other arthropods

In terms of numbers of species, the arthropods as a group surpass any other phylum in the animal kingdom. Of some 1.25 million known animal species, about four-fifths of these are arthropods; the class Insecta by itself represents almost three-quarters of all described animal species. Considering the small size of most members, it is probable that as many species again remain undescribed. However, the vast biodiversity of arthropods is not matched by proportional investment in genome sequencing (Heckel 2003). Only a limited number of species of arthropod have been sequenced completely. Pionerees were the fruit fly *D. melanogaster* (Adams *et al*. 2000), the malaria mosquito *An. gambiae* (Holt *et al*. 2002), the silk worm *Bombyx mori* (Biology Analysis Group 2004), and the honey bee, *Apis mellifera* (Honeybee Genome Sequencing Consortium 2006). Several other species now contribute to this list of sequenced arthropods (Table 2.7).

In the beginning, many arthropods were considered for sequencing because of their medical (disease transmission) or agronomical (pest) importance, however, this situation has changed considerably since two very important models of evolutionary ecology were added to the list. These are the three *Nasonia* species (*Nasonia* Genome Working Group 2010) and the water flea *Daphnia*.

Nasonia is a parasitoid wasp in which the females lay their eggs inside a host; the larvae developing from these eggs consume the host internally and

Table 2.7 List of arthropod species, other than *D. melanogaster*, *An. gambiae*, *B. mori*, and *A. mellifera* for which whole-genome sequencing projects are completed

Taxonomic group	Arguments, relevance
Insecta, Diptera	
Drosophila pseudobscura and ten other *Drosophila* species	Comparison with *D. melanogaster*, mechanisms of radiative speciation, phylogenetic shadowing
Aedes aegypti (yellow fever mosquito)	Disease transmission, comparative basis for haematophagy, together with *An. gambiae*, *C. pipiens*, and other mosquitoes
Culex quinquefasciatus (southern house mosquito)	Vector for Western Nile Virus, St Louis encaphalitis virus and filarial nematodes, model for blood-feeding mosquitoes
Glossina morsitans (tsetse fly)	Transmitter of *Trypanosoma*, cause of sleeping sickness
Insecta, Hymenoptera	
Nasonia vitripennis, *N. giraulti*, *N. longicornis* (parasitoid wasp)	Genetics of parasitoid behaviour, biological control of agricultural pests, reproductive allocation, life-history evolution
Insecta, Coleoptera	
Tribolium castaneum (red flour beetle)	Model organism for genetics and developmental biology
Insecta, Lepidoptera	
Heliothis virescens (tobacco budworm)	Larva attacks various crops such as alfalfa, clover, cotton, soybean, and tobacco
Insecta, Hemiptera	
Acyrthosiphon pisum (pea aphid)	Important agricultural pest, model for insect-plant relations, coevolution with obligate bacterial symbionts
Insecta, Phthiraptera	
Pediculus humanus (human louse)	Vector of typhus, skin irritation
Acari, Ixodidae	
Ixodes scapularis (black-legged tick)	Parasite of wildlife, vector for transmission of Lyme disease and other vectorial diseases
Crustacea, Cladocera	
Daphnia pulex (water flea)	Important model species in aquatic ecology, widely accepted test organism in water-quality assessment

eventually kill it. Parasitoid wasps are very important as beneficial insects in the control of agricultural pests. *Nasonia* species parasitize on several cyclorraphan flies including houseflies, which are themselves easy to culture, and this contributes to their popularity among experimental ecologists. *Nasonia* also has a short life-cycle (around 2 weeks) and produces many offspring. The genomics resources developed for *Nasonia* will also benefit work on other parasitoid wasps such as *Asobara* and *Cotesia*.

The genome of *Nasonia* appeared to be rich in transposable elements and other repetitive DNA (*Nasonia* Genome Working Group 2010), much more than in the other hymenopteran, *Apis mellifera*. The abundance of repetitive DNA is one of the reasons why *Nasonia* has a high density of microsatellite loci. One of the main discoveries in the *Nasonia* genome was the presence of all three DNA methyltransferases (three *Dnmt1* genes, one *Dnmt2*, and one *Dnmt3*). This situation is like the one in *Apis mellifera*, but unlike *Drosophila* which has only *Dnmt2*. In *Apis mellifera*, DNA methylation is important for caste development, and in *Nasonia* it may be related to sex determination. We will discuss in Chapter 6 the role of DNA methylation in the regulation of gene expression. Other key findings of the *Nasonia* genome analysis include 79 genes encoding venom proteins, and an ancient lateral gene transfer involving Pox viruses, *Wolbachia* and *Nasonia*. Genetic tools and gene expression resources are becoming available and will ensure that *Nasonia*

will develop into a main model of evolutionary ecology.

Another set of species that raises great expectations among ecologists are the water fleas *Daphnia pulex* and *Daphnia magna*. Water fleas (Crustacea, Cladocera) are extremely important organisms in aquatic food chains. On the one hand, they have a top-down effect by grazing and controlling algal populations, and on the other hand they have a bottom-up effect as food for fish. Numerous limnologists have studied population dynamics, life history, energy metabolism, and vertical migration of daphnids for nearly three centuries. The reproductive biology of *Daphnia* is very suitable for genomic studies because it can reproduce by apomictic parthenogenesis under favourable conditions but also by sexual mechanisms under unfavourable conditions. The sexually produced resting stage is a saddle-shaped structure called an *ephippium*, which contains diapausing eggs; indeed, centuries-old *Daphnia* can be resurrected from ephippia recovered from natural sediments. Due to the combination of sexual reproduction and clonal propagation in localized populations, water fleas are differentiated in many ecotypes, showing diverging rates of genetic variation, depending on the temporal stability of their habitat and the rate of outbreeding (De Meester 1996). *D. magna* is an internationally accepted standard test organism in ecotoxicology and water-quality assessment in general.

The Daphnia Genomics Consortium (DGC) coordinates the release of the genome sequence (http://daphnia.cgb.indiana.edu) and a great variety of other genomic tools, such as nearly 2000 microsatellite markers (many linked to gene loci), a fine-scale genetic map, and no less than 50 000 cDNAs, expressed in a great variety of environments and developmental stages. A database, wFleaBase (http://wfleabase.org, Colbourne *et al*. 2004), provides the infrastructure to share genomics data and protocols via the World Wide Web. Techniques for cell culture and genetic transformation are in development. The genome sequencing effort was initially focused on *D. pulex*, but now also covers *D. magna*. It is expected that, among animal models, *Daphnia* will be one of the most promising species for realizing a true blend of ecology and genomics.

It is worthwhile probing the base of the phylum Arthropoda when evolutionary arguments are being advanced for future sequencing projects. How do insects relate to myriapods, arachnids, and crustaceans? This debate has occupied entomologists for a very long time (Gillot 1980); a final answer is now coming from comparative genomics studies. Timmermans *et al.* (2008) made an alignment of 48 ribosomal proteins sequenced as part of ETS projects in a great variety of arthropods and developed a phylogenetic tree reproduced here as Fig. 2.21. The phylogeny clearly shows that Hexapoda (including Insecta and the apterygote groups) is monophyletic. This contradicts earlier claims (Nardi *et al.* 2003) that the hexapod body plan had evolved twice. On the other hand, Crustacea is obviously paraphyletic. In fact there are two groups of crustaceans, one involving waterfleas and brine shrimp, the other including crabs, crayfish, and so on. The former group is actually more related to the hexapods than to the other crustaceans! Paraphyly of Crustacea is also supported by other phylogenomic analyses (e.g. Regier *et al.* 2010).

Within the Diptera (the two-winged flies), *D. melanogaster* belongs to the family Drosophilidae, or fruit flies, a large cyclorraphan family with almost 3000 species in 60 genera described worldwide. *Drosophila melanogaster* actually falls in the subgenus *Sophophora* and a proposal has been made to rename it officially as *Sophophora melanogaster*, but this (luckily) has had little follow-up. Members of the drosophilid family are found in association with fermenting substances, most notably fruit, but also in specialized habitats such as the sap of bleeding tree wounds, slime fluxes, rotting cacti, and flower heads as well as in general habitats such as on fungi and decaying leaves on forest floors. The adults (3–5 mm) are attracted by the alcoholic odours emanating from the activities of bacteria and yeasts colonizing a substrate rich in sugars. Eggs are driven into the substrate by means of the female's ovipositor and the maggot-like larvae emerging from the eggs develop inside the substrate through three stages followed by a pupa; at 25 ºC new adults emerge within 10 days of hatching. *D. melanogaster* is believed to have evolved in Africa, but is now found worldwide in many

synanthropic habitats, such as fruit markets and garbage cans. The classical booklet by Demerec and Kaufmann (1950) is still a valuable source of information about the biology of *Drosophila*.

The publication of the sequence of the *Drosophila melanogaster* genome by Adams *et al*. (2000) was a landmark achievement because it marked nine decades of *Drosophila* research, starting with T.H. Morgan's discovery in 1910 of a mutant white-eyed fly and leading to major conceptual or technical breakthroughs in our understanding of animal genetics over the course of the century (Rubin and Lewis 2000). The *Drosophila* genome was the second and largest animal genome sequenced at the time and the short period (less than 1 year) in which the work was completed, as a combined academic and industry effort, was impressive. Interestingly, the superstar status that the fruit fly already enjoyed before the sequencing began was actually something of an impediment to starting the project (Rubin and Lewis 2000). Over 1300 individual genes, nearly 10% of all the genes in *Drosophila* had already been cloned and studied, and with such success a whole-genome sequence was considered unnecessary. Nevertheless, it was realized that the majority of genes eluded study by traditional methods because in *Drosophila*, like in *C. elegans* and *Arabidopsis*, many genes do not show obvious phenotypes when mutated.

The *Drosophila* genome-sequencing project was the first that had to deal with a substantial amount of heterochromatin, one-third (60 Mbp) of the genome. Attention was concentrated on the 120 Mbp of euchromatin, because the heterochromatic regions were intractable to the sequencing method. Heterochromatin was present in the centromeres of the two autosomes, one half of the X chromosome, the complete Y chromosome, and the very small fourth chromosome. The heterochromatic, gene-poor, regions consisted primarily of simple sequence satellites and transposons. Interestingly, the transition zones between heterochromatin and euchromatin regions appeared to contain many previously unknown genes.

The genome of *Drosophila* was found to contain around 13 600 genes; taking alternative splicing into account, 14 113 transcripts were postulated (Adams *et al*. 2000). Rubin *et al*. (2000) compared the fruit fly genes with those of *C. elegans* and *S. cerevisiae*, the only two other eukaryote genomes sequenced at the time. A summary of these comparisons is shown in Fig. 2.22. Around 35% of the protein-encoding genes had a match in the nematode genome and 16.5% in the yeast genome. These relatively low values, which by themselves are in accordance with the phylogenetic relationships between the three species, showed that the genome of *Drosophila* was only remotely related to the other eukaryotes. In fact, further comparisons including a mammalian EST database demonstrated that half of the fly proteins showed similarity with mammalian proteins, which suggested that the *Drosophila* proteome is more similar to mammalian proteomes than are those of nematodes or yeast.

The basis for comparative genomics of fruit flies was strengthened considerably when, in 2007, the number of sequenced *Drosophila* genomes was extended to 12 (*Drosophila* 12 Genomes Consortium 2007). The amount of coding DNA in the 12 species appeared to be comparable, however, the total size of the genomes varied by a factor of 3 (130 Mbp in *D. mojavensis* and 364 Mbp in *D. virilis*). This is mainly due to the amount of DNA attributed to transposons. Gregory and Johnston (2008) examined the variation in genome size across 67 species of drosophilids, and noted a correlation with larval developmental time (from egg to adult): the species with the largest genomes had the longest developmental times. The adaptive explanation was that a large genome slows down the rate of cell division.

The comparative analysis of the 12 Drosophila genomes also revealed many blocks of synteny between the species, although individual genes have often changed places. The latter was especially evident from the *Hox* genes, for which seven different cluster arrangements were identified. None of the species has retained the ancestral condition. That the *Hox* genes can switch places so easily is contrary to the idea that their physical arrangement is constrained for functional reasons. The 12 species had more or less the same gene content, but there were contractions or expansions of gene families in almost all cases in one or more of the species. Gene families particularly subjected to species-specific

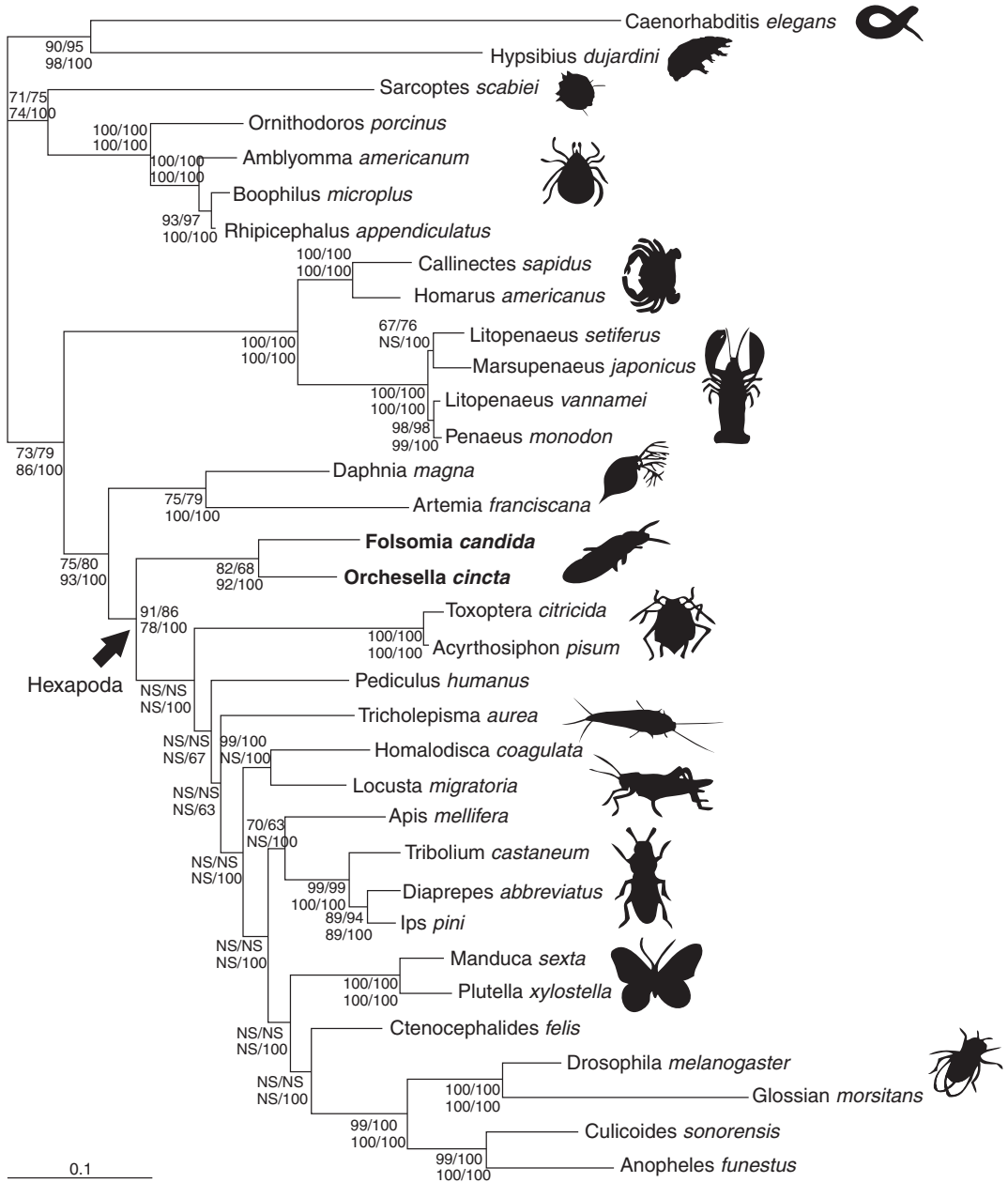


Figure 2.21 Phylogenetic tree of Pancrustacea, developed from an alignment of 48 ribosomal protein sequences. The tree shows that both Pancrustacea (Crustacea plus Hexapoda) and Hexapoda are monophyletic lineages, but Crustacea itself is paraphyletic. Two species of Collembola cluster as a sister group of the insects; they are indicated in bold because of their crucial position in the debate about monophyly of Hexapoda (cf. Nardi *et al*. 2003). Reproduced from Timmermans *et al.* (2008), by permission of BioMed Central.

evolutionary dynamics were proteins of the immune system, chemoreception, and defence against toxicants. This was related to the feeding habits of each species: dietary generalists tend to maintain the highest diversity of chemoreception genes. All in all, comparative study of the 12 *Drosophila* genomes has allowed many interesting insights into the evolutionary dynamics of genes and genomes.

Shortly after the genome of *Drosophila melanogaster*, the genome of the malaria mosquito *An. gambiae* (Diptera, Nematocera, Culicidae) was published (Holt *et al*. 2002). This sequencing project posed another challenge that had not been encountered in any sequencing exercise so far, namely the high degree of genetic polymorphism in the clone library used. *An. gambiae* populations are highly structured into several genetic types, with different habitat preferences but indistinguishable morphologies. The heterogeneous genetic background caused difficulties in genome assembly, because haplotype variation was difficult to distinguish from repeat sequences. The starting stock for the sequencing was a long-term culture maintained at the Institut Pasteur in Paris, which was established from a cross between a strain originating in Nigeria with a strain from Kenya, followed by three rounds of outbreeding. It was believed that the strain mostly represented the *Savanna* form of *An. gambiae*, presently found in western Kenya. Variability was not distributed uniformly throughout the genome; it was much higher in the automosomes than in the X chromosome. It was assumed that the genome of the strain used had resulted from a complex introgression between two different chromosomal forms (cytotypes); lack of introgression in the X chromosome (as a consequence of the hemizygous condition of the male) could explain the lower degree of polymorphism in this chromosome.

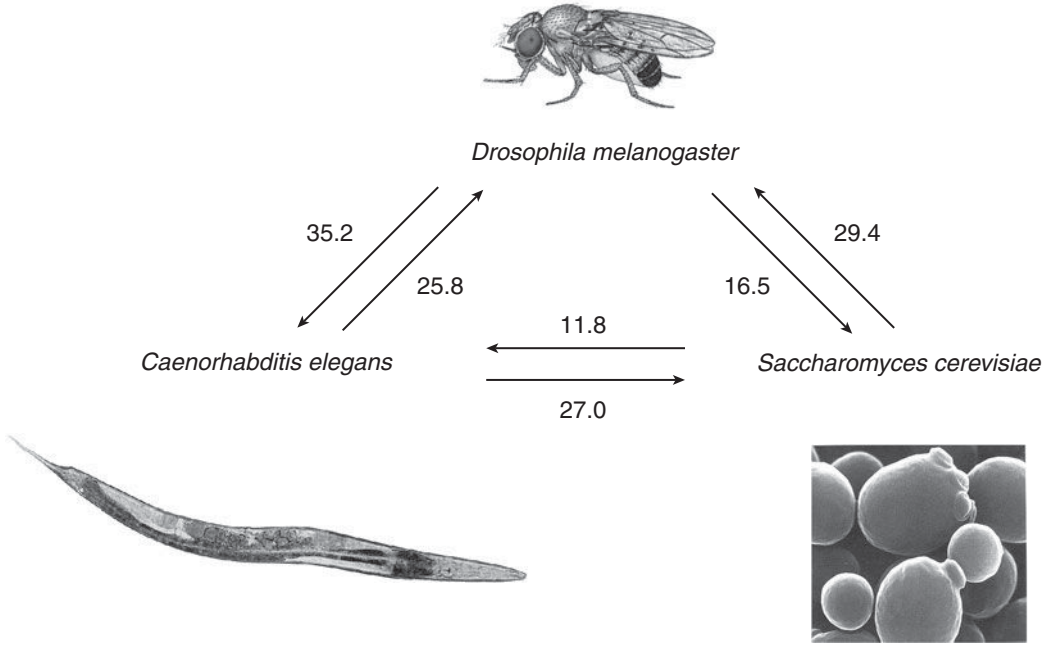


Figure 2.22 Pairwise similarities of predicted proteins in the genomes of fruit fly, nematode, and yeast. The numbers adjacent to the arrows indicate the percentage of proteins in an organism that has a match in the organism indicated by the arrow (at a BLAST *E* value of 10^{-10} or smaller), relative to the number of proteins in the first species. Each set of pairs was analysed without consideration of the third proteome. Data from Rubin *et al*. (2000).

Compared with *Drosophila*, the genome of *An. gambiae* demonstrates several features that relate to its haematophagous mode of nutrition. Several gene families were found to be greatly expanded; that is, they appeared to consist of many more, and more diversified, members than the orthologous *Drosophila* family (Zdobnov *et al*. 2002). A prime example of gene-family expansion is in genes containing a fibrinogen domain (FBN genes; Christophides *et al*. 2002). The FBN genes of invertebrates are involved in the *innate immune response* (see Section 5.4.2). FBN genes combat intruders by binding to foreign surfaces. The expansion of these genes in *Anopheles* might reflect selection pressure from growth of bacteria in the gut after a blood meal or from defence against *Plasmodium* and other parasites. Other genes of particular importance to *An. gambiae* are those related to antioxidant defence, which is a challenge to animals feeding on vertebrate blood, because the blood-meal-derived haem group is a catalyst of free oxygen radicals. In summary, the genome of *Anopheles* is twice as large as the *Drosophila* genome, but this is not only due to expansion of specific gene families, but also to loss of non-coding DNA in *Drosophila*. The fraction of the genome which consists of introns and intergenic spacers is considerably larger in *Anopheles* (Table 2.8). The comparative analysis of Zdobnov *et al*. (2002) also showed that several genes known from yeasts, nematodes, and plants are absent from the two insect genomes. Examples are nine genes related to sterol metabolism, which is in accordance with the fact, noted by entomologists a long time ago, that insects cannot synthesize their own sterols. Table 2.8 summarizes some characteristic features of the genomes of *Anopheles* and *Drosophila*.

The genome sequences of drosophilids are curated by the FlyBase consortium (http://flybase.org). Along with other responsibilities, FlyBase is committed to maintaining up-to-date annotations of the genome (FlyBase Consortium 2003). A special challenge is to connect the genome annotations to the vast amount of information on phenotypic characters available for *Drosophila*. Since the original release, reannotations have changed the contents of the genetic database considerably, due to division of genes into two genes, merging previously separate genes, and addition or deletion of exons. The web pages of FlyBase provide an enormous amount of information on genetic maps, cytological maps, genes, alleles, gene products, protein function, protein location, gene expression, transposons, transgene constructs, fruit fly stocks, collections, fly anatomy, literature, references, and fruit fly investigators. For transcription profiling the GeneChip® Drosophila Genome 2.0 Array from Affymetrix is often used. This microarray provides comprehensive coverage of the transcribed *Drosophila* genome using 18 880 probe sets, analysing over 18 500 transcripts.

Table 2.8 General features of the genomes of two insects, *D. melanogaster* (fruit fly) and *An. gambiae* (malaria mosquito)

	Anopheles	Drosophila
Total genome size	278 Mbp	123 Mbp
Total coding DNA	19.3 Mbp (7%)	23.8 Mbp (19%)
Total intron DNA	43.0 Mbp (15%)	27.6 Mbp (22%)
Total intergenic DNA	216.0 Mbp (78%)	71.3 (58%)
Number of genes	13 683	13 472
Number of exons	50 609	54 537
Genomic GC content	35.2%	41.1%

Sources: From Holt *et al*. (2002) and Zdobnov *et al*. (2002).

AnoBase is the *An. gambiae* genomic and biological database (www.anobase.org). The site also contains reference material and links to other mosquito resources, as well as current news and conference information. Anobase works in collaboration with Genoscope, the French Government's sequencing centre, who along with Celera Genomics were involved in the initial sequencing of the mosquito genome, and several organizations that specialize in research on tropical diseases, such as The Malaria Research and Reference Reagent Resource Center (MR4). Genome information on a range of arthropod disease vectors such *Aedes aegypti*, *Ixodes*, *Culex*, *Pediculus*, and *Anopheles* can be obtained from the VectorBase website: www.vectorbase.org.

2.3.5 Plant genomes

The plant kingdom is a monophyletic evolutionary lineage, including green algae, mosses, ferns, and seed plants. Genome sequencing up to now has

been completed for two representatives from the green algae, *Chlamydomonas reinhardtii* and *Ostreococcus tauri*, and several higher plants including *A. thaliana* (thale cress), *Oryza sativa* (rice), and *Populus trichocarpa* (black cottonwood). Many sequencing projects are being planned and genomic databases, for example collections of ESTs, are being developed worldwide. Researchers often organize themselves around a group of related plant species; for example the Legume Genomics network focuses on *Medicago truncatula* as a model, the Solanaceae Genomics Network concerns tomato and potato, collaborators in TreeGenes are interested in forest genetics and the genome sequencing of forest trees, the Multinational Brassica Genome Project addresses the various *Brassica* species, SoyBase focuses on the soybean *Glycine max*, and BeanGenes addresses *Phaseolus* and *Vigna* species, among others.

The very large, often polyploid, genomes of some agriculturally important plants are a serious obstacle for full-genome sequencing (Table 2.9). This is especially valid for species from the family Poaceae (grasses), which includes the cereals (Triticaceae). Still, it is expected that even the genomes of these plants—despite their very large size—will be sequenced eventually. For some plant species commercial microarrays for gene-expression analysis are already available, even though a complete genome assembly and annotation has not yet been conducted. In these cases the microarrays were developed from publicly accessible databases complemented by EST sequences submitted by consortium members. Examples are the Affymetrix GeneChip® Soybean Genome Array, which can be used to study gene expression of over 37 500 soybean transcripts, and the Affymetrix GeneChip® Barley Genome Array that was designed in collaboration with the international barley community.

Sequencing the genome of the green alga, *C. reinhardtii* (Chlorophyta, Volvocales), was inspired by the fact that this small unicellular organism served as a classical model for biochemical research into photosynthesis and cell motility (Harris 2001; Grossman *et al*. 2003; Gutman and Niyogi 2004). A key feature of its success was the availability of a mutation in a photosynthetic regulatory mechanism called the state transition. The state transition involves allocation of light-harvesting proteins from photosystem II to photosystem I in response to the wavelength of incident light. Because this shift is larger and more easily measured in *Chlamydomonas* than in higher plants the alga was a preferred organism for eukaryotic photosynthesis research. *Chlamydomonas* is also a very suitable organism for the study of experimental evolution. It can reproduce both sexually and asexually, and this property has been exploited by Kaltz and Bell (2002) to demonstrate that sexual, genetically diverse, populations were better able to adapt to a new hostile environment than equivalent asexual lines. From a comparative genomics point of view the genome of *C. reinhardtii* is interesting because together with *Arabidopsis* it forms a pair bracketing the entire lineage of green plants. *Chlamydomonas* species are also studied in freshwater algology and are commonly observed in plankton samples under the microscope as tiny, rapidly swimming flagellates. Thus a connection between the growing genomics insights and ecological studies seems very possible, although a practical difficulty is the very large diversity in the genus *Chlamydomonas*, which has no less than 600 species worldwide. The Chlamy Center, found at www.chlamy.org, provides entry to a genome

Table 2.9 Genome sizes of some agriculturally important crops, in comparison with *Arabidopsis*

Species	Scientific name	Genome size (Mbp)
Cabbages		
Thale cress	*Arabidopsis thaliana*	125
Oilseed rape	*Brassica napus*	1200
Cereals		
Rice	*Oryza sativa*	420
Barley	*Hordeum vulgare*	4800
Wheat	*Triticum aestivum*	16 000
Corn	*Zea mays*	2500
Legumes		
Garden pea	*Pisum sativum*	4100
Soybean	*Glycine max*	1100
Nightshades		
Potato	*Solanum tuberosum*	1800
Tomato	*Lycopersicon esculentum*	1000

Source: From Adam (2000). Reproduced by permission of Nature Publishing Group.

browser, an EST database, and various microarray projects, as well as a system for ordering specific strains. Another green alga, *Ostreococcus tauri* (Chlorophyta, Mamiellales) has also been sequenced (Derelle *et al.* 2006). This species has extremely small cells and a very densely packed genome of only 11.5 Mbp; it provides insight into the minimal number of genes necessary for a photosynthetic eukaryote.

Only one species of moss is presently completely sequenced, *Physcomitrella patens* (Funariales) (Rensing *et al.* 2008). This plant is a genetic model very suitable for reverse genetics (targeted gene knock-down). It takes a position in between aquatic algae such as *Chlamydomonas* and higher plants, such as *Arabidopsis*, and so is also interesting from an evolutionary point of view. From the comparative genomics study reported in Rensing *et al.* (2008) it may be concluded that the last common ancestor of all land plants lost several genes associated with movement in aquatic habitats (components of the flagellum), it lost dynein-mediated transport of materials through the cytoplasm, it gained however several genes related to abiotic stress resistance, such as the abscisic acid signalling pathway, and it showed an overall increase in genome complexity. These comparisons, when added to more genome information on bryophytes, aids reconstruction of the evolutionary events that marked the transition in the plant lineage to full terrestrial life.

Among the sequenced species of higher plants, the legume *M. truncatula* takes a special position because of its high ecological relevance. *Medicago* is seen as a model for legumes in general, a family of plants that includes well-known genera such as *Lathyrus*, *Phaseolus*, *Glycine*, *Trifolium*, *Medicago*, *Pisum*, *Vicia*, and *Lotus*, among which are many agriculturally important species that provide a major source of protein for the human population. Under nitrogen-limiting conditions, leguminous plants are able to establish a symbiotic relationship with bacteria from the family Rhizobiaceae, including *S. meliloti*, in itself a genomic model species (see Section 2.2). Symbiotic nitrogen fixation is a very important link in the global nitrogen cycle and so genomic studies of the legume–rhizobium relationship have great added value in comparison to *Arabidopsis*, which does not offer this possibility. In addition, *Medicago* can also be used to study arbuscular mycorrhizal symbiosis. *M. truncatula* is a close relative of alfalfa (*M. sativa*), but unlike the latter species, it has an annual life cycle and only half the genome size of the alfalfa genome. It exhibits simple genetics and a genome highly conserved with alfalfa and pea (*Phaseolus vulgaris*) and moderately conserved with soybean. The *Medicago* sequencing is supported by a dense physical map of nearly 200 000 ESTs, and a growing array of functional genomic tools, see www.medicago.org and Town (2006).

Populus trichocarpa (family Salicaceae) was the first tree that was considered for whole-genome sequencing. Forest biologists have developed strong justifications for why trees should be viewed as model systems in plant genomics (Bradshaw *et al*. 2000; Wullschleger *et al*. 2002). The physiology of trees includes a number of processes that cannot be understood from the model herbaceous plants, such as perennial growth, large size and complex crown structure, extensive secondary xylem, and bud dormancy (Taylor 2002; Brunner *et al*. 2004). Knowledge of the poplar genome greatly contributes to the growth of a research area known as *forest genomics*, which focuses on questions related to mechanisms of wood formation, stress resistance, pathogen resistance, genetic diversity of trees, conservation, and tree breeding. Poplar is an important model species for forest genomics, along with spruce and to a lesser extent pine. With its genome consisting of 480 Mbp, more than four times larger than *Arabidopsis*, but still some 40 times smaller than pine, poplar has a relatively small genome among trees. The sequencing was done by the International Populus Genome Consortium (IPGC) in which US, Canadian, and Swedish scientists collaborated (www.ornl.gov/sci/ipgc). The draft genome sequence was published in 2006 (Tuskan *et al.* 2006).

The importance of poplar as a genomic model extends beyond the aims of forestry; poplar is also a very suitable species for fundamental ecological genomics, as illustrated by the following example. Cottonwoods (*Populus trichocarpa*) are central to the structure and ecosystem functioning of riparian forests of North American rivers, whereas black

poplar (*Populus nigra*) fulfils the same role in European riparian systems. Ecological interactions in such ecosystems involve engineering by beavers (North American *Castor canadensis* and European *Castor fiber*), herbivory by the arthropod community on the leaves (caterpillars, beetles, aphids, etc.), and decomposition of fallen leaves by a diverse community of microorganisms, invertebrate shredders, and detritivores. Interestingly, the principal steering force in these interactions appears to emanate from the genetics of the tree. A number of hybrids from the genus *Populus* occur naturally and hybridization is associated with marked variations in leaf form and chemical composition of the leaves. Wimp *et al*. (2004) showed that the genetic variation of a cottonwood stand, estimated by AFLP fingerprints, was strongly correlated with the Shannon–Weaver index of the leaf arthropod community: genetically diverse stands had the highest species richness. The effect is most probably mediated by the chemical composition of the leaves, particularly the concentration of tannins. In the same system, leaf tannin was negatively correlated with decomposition and nitrogen mineralization of poplar litter (Schweitzer *et al*. 2004). The fact that genetic diversity of a dominant species in an ecosystem pervades into the herbivorous and decomposing communities has important implications for conservation strategies. In addition, the system offers exciting possibilities for establishing a direct relationship between ecosystem functions and gene-expression profiles in the tannin-synthesis pathway.

In addition to *Medicago* and *Populus*, two species with a great ecological relevance, several agriculturally important plant species have been subjected to genome sequencing over the last years. These include grapevine (*Vitis vinifera*, Grapevine Consortium 2007), sorghum (*Sorghum bicolor*, Paterson *et al.* 2009), cucumber (*Cucumis sativus*, Huang *et al.* 2009), soybean (*Glycine max*, Schmutz *et al*. 2010), and castor bean (*Ricinus communis*, Chan *et al*. 2010). These species are not the main interest of ecologists, but they may provide a basis for comparison and so support interpretation of responses seen in wild relatives.

Despite the promises of *Medicago* and *Populus*, at the moment *A. thaliana* is the best-characterized genomic model plant by far. Thale cress is a species of the Brassicaceae family, with a wide distribution in the morthern hemisphere (Fig. 2.23). The species is native to western Eurasia but is now found in the wild throughout Europe, the Mediterranean, the East African highlands, and eastern and central Asia (Hoffmann 2002). It has also been introduced into America and Australia (Fig. 3.23). Johannes Thal (hence, *thaliana*) first described *Arabidopsis* in the sixteenth century in the German Harz Mountains, although he called it *Pilosella siliquosa* at the time. The name underwent several changes before *A. thaliana* was settled upon in 1842.

A. thaliana is a small annual plant, 5–30 cm high, growing in open areas with sandy soil, along paths, and in agricultural fields. The life cycle is that of a winter annual, which germinates and grows in autumn, survives winter as a rosette, and flowers in early spring. The developmental switch from vegetative growth to reproduction, involving erection of the flower stem (*bolting*) is an important issue of research in ecological genomics (see Chapter 4). In fact the life history is more flexible than a typical winter annual, because there are also variants that germinate in spring and flower in July, and others that have more or less lost their phenological tuning. *A. thaliana* has a high level of self-pollination. It does not cross-hybridize with its relatives, because the number of chromosomes is reduced to five, whereas all its closest relatives have a haploid chromosome number of eight. AFLP fingerprinting has shown that the population structure of *A. thaliana* over its native geographical range is shaped by postglacial colonization from the Iberian peninsula and the near east, leading to a suture zone in central Europe (Sharbel *et al*. 2000).

A. thaliana is never a dominant species in wild vegetation. Its suitability as a model for ecological field work is limited by its sparse occurrence and its narrow phenological window. Molecular ecologists have been looking for related species that lend themselves better to ecological studies (Mitchell-Olds 2001). There are 10 species of *Arabidopsis* among the approximately 3000 species of Brassicaceae. *Arabidopsis halleri* and *Arabidopsis lyrata* are closely related species with a perennial life cycle. Genera within the same clade are *Cardamine*,

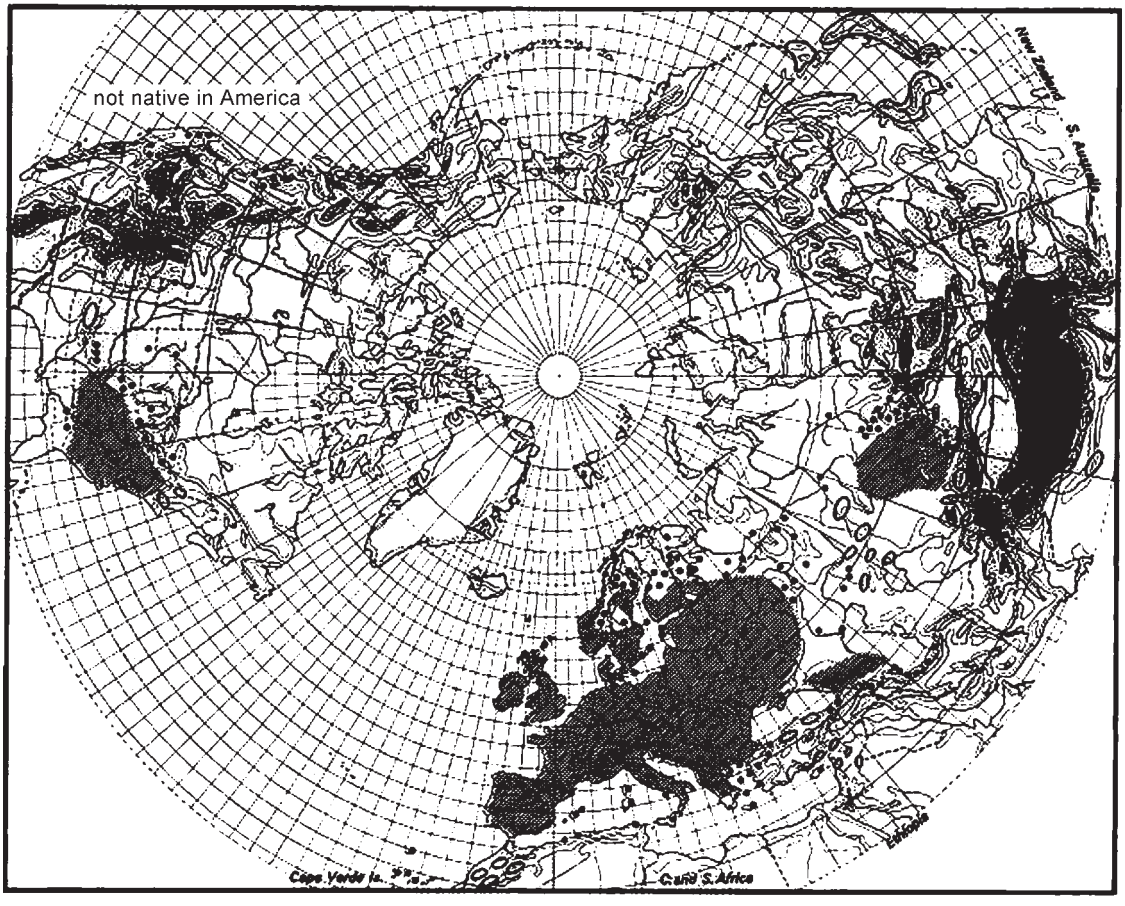


Figure 2.23 Global geographical distribution of *A. thaliana*. Courtesy of Koeltz Scientific Books.

Rorippa, *Barbarea*, *Arabis*, and *Thlaspi*. Taken together these cruciferan species comprise a wide array of life cycles and ecological niches. Several of these species seem to be suitable as 'wild' counterparts of *A. thaliana*, and four of them are pictured in Fig. 2.24.

Scientific research on *A. thaliana* started in the beginning of the twentieth century with microscopic studies on the chromosomes, but it was not until the 1970s that molecular geneticists discovered its suitability as a model. A genetic map was developed by Maarten Koornneef and coworkers at the beginning of the 1980s and physical maps of the genome, based on RFLP and AFLP fingerprints, were developed thereafter. The possibilities of transforming *Arabidopsis*, first using *Agrobacterium* and later more advanced genetic-engineering techniques, were also developed; extensive mutant collections were built concurrently. The Multinational Coordinated Arabidopsis thaliana Genome Research Project was launched in 1990 and the Arabidopsis Genome Initiative (AGI) started sequencing the genome in 1996 on a chromosome-by-chromosome basis. Chromosomes 2 and 4 were completed first and published in 1999. With the completion of chromosomes 1, 3, and 5 in 2000 the genome sequence was essentially complete and this was considered a hallmark event for plant biology (Walbot 2000).

The first detailed analysis of the *Arabidopsis* genome content provided many surprises

(Arabidopsis Genome Initiative 2000). From the total genome of 125 Mbp, 115.4 Mbp had been fully sequenced. This sequence appeared to contain 25 498 predicted genes, significantly more than *C. elegans* and *Drosophila* (see Table 2.2). When the protein-coding genes were compared with the genomes of other eukaryotes and prokaryotes, many matches were found (Fig. 2.25). It even turned out that in the *Arabidopsis* genome genes may be found that share clear homology with human disease genes! However, the percentage of genes matching those of other species depended greatly on the functional category. Among the genes related to transcription only a small percentage (8–23%) had a match in another species, whereas among the genes related to protein synthesis up to 60% corresponded to a gene in another species. Overall, the similarity with prokaryote genomes was significantly less than with eukaryotes, but in the functional category of energy metabolism more than 30% of the plant genes were similar to a bacterial gene. This is obviously a consequence of the trans-

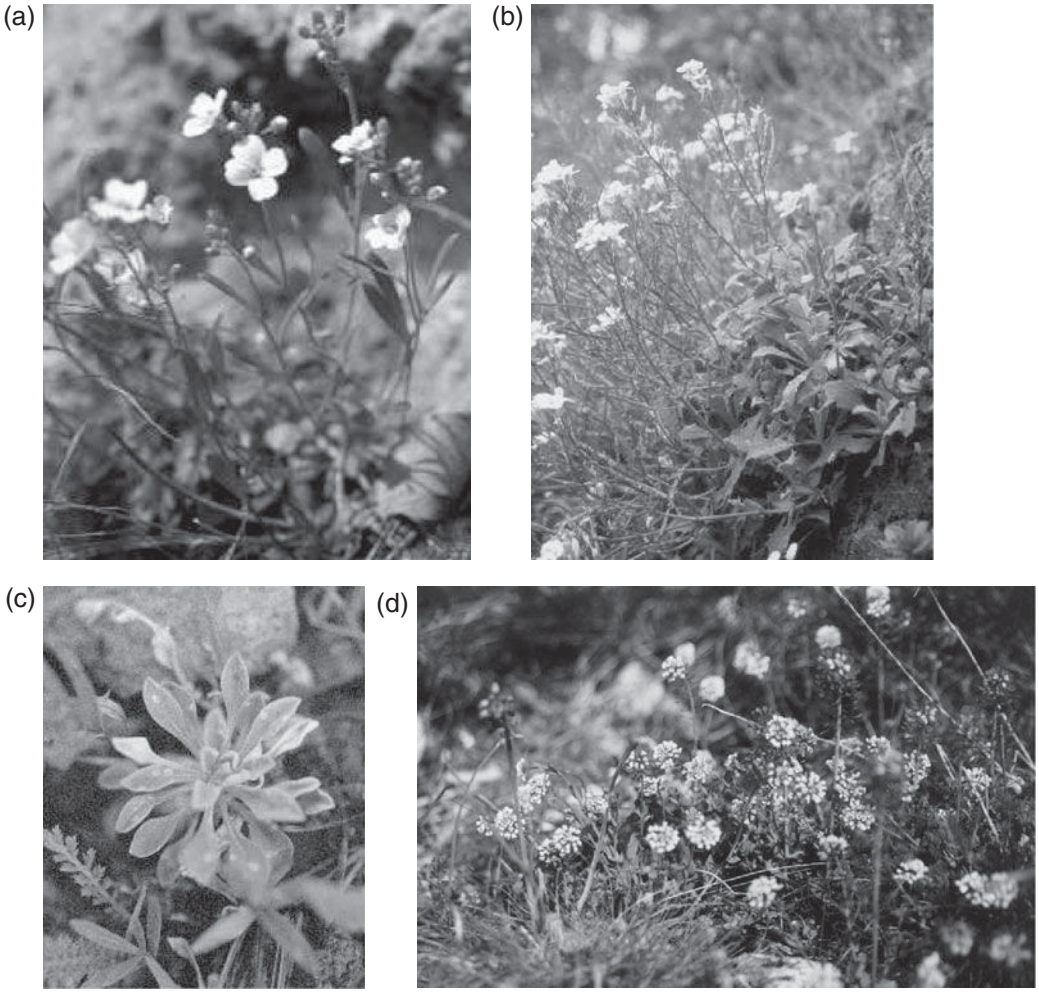


Figure 2.24 Wild relatives of *A. thaliana*: (a) *Arabidopsis petraea*, (b) *Arabis alpina*, (c) *Boechera holboelii*, and (d) *Thlaspi caerulescens*. Courtesy of T. Mitchell-Olds, Max Planck Institute, Jena and C. Lefèbvre, Free University of Brussels.

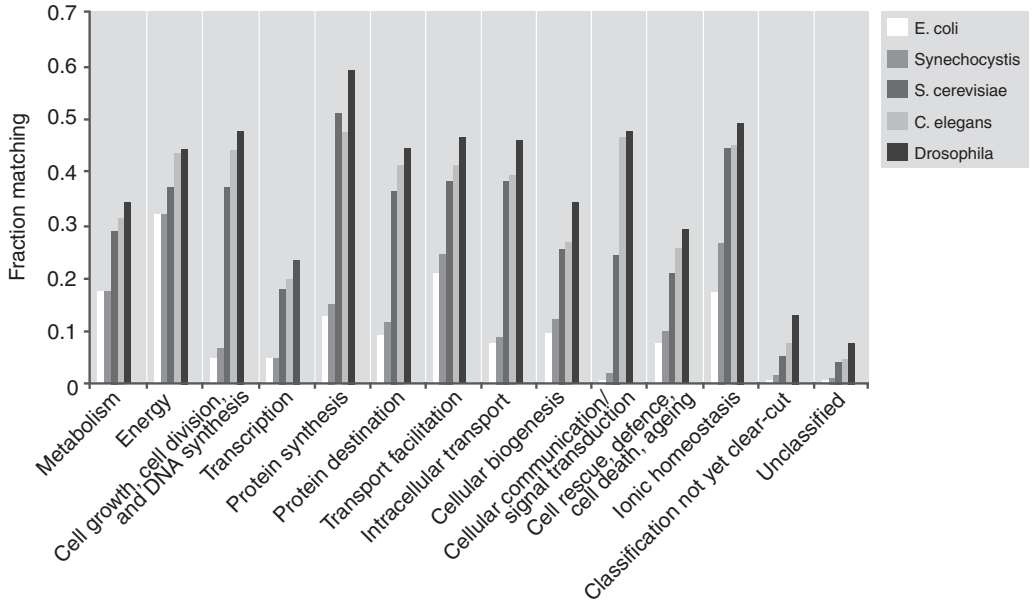


Figure 2.25 Functional analysis of the *Arabidopsis* genes predicted from the genome sequence, showing the similarities between *Arabidopsis* functional gene categories and bacterial genomes (*E. coli* and *Synechocystis*, a cyanobacterium) and those of yeast, nematode, and fruit fly. The *y* axis indicates the fraction of *Arabidopsis* genes in a functional category showing a BLAST match with the respective reference genome. The right to use this figure provided courtesy of members of the Arabidopsis Genome Initiative and Nature magazine. This figure first appeared in *Nature* 408, 796–815 (2000).

fer of chloroplast genes to the nuclear genome. Maybe less surprisingly, in the category of cellular communication and signal transduction hardly any match was found between *Arabidopsis* genes and those of bacteria, but the correspondence with the (unicellular) yeast genome was also relatively low (Fig. 2.25).

Why has the *Arabidopsis* genome 87% more genes than *Drosophila melanogaster*? Two explanations have been given (Arabidopsis Genome Initiative 2000). First, individual genes have been subjected to wide-scale amplification events, generating large arrays of tandems and dispersed gene families; unequal crossing-over may be the predominant mechanism involved. Second, the genome of *A. thaliana* has undergone a whole-genome duplication after it diverged from most other dicotyledons (Bowers *et al*. 2003), classifying *A. thaliana* as a cryptotetraploid species (see Section 2.1). These two genome-enlargement mechanisms have led to a considerable degree of *genetic redundancy* in the genome; that is, more than one gene has the same function. This is consistent with observations from genetic engineering studies which show that many genes can be knocked-out in *Arabidopsis* without any phenotypic consequences. The Arabidopsis Genome Initiative speculated that such large-scale duplication events may be needed to generate new functions, and that creating new functions by duplication is more common in plants than in animals, where novelties are more often generated by rearrangements of promoters and alternative splicing.

The possibility of ancient polyploidy in model plants was analysed in more detail by Blanc and Wolfe (2004), using whole-genome data and EST sequences for 14 different species. The authors estimated the sequence divergence between the two genes of a paralogous pair by looking at the average number of substitutions without amino acid alteration (number of synonymous substitutions per synonymous site, K_s). The frequency distribution of K_s values over all the genes is a cue to the timing of the

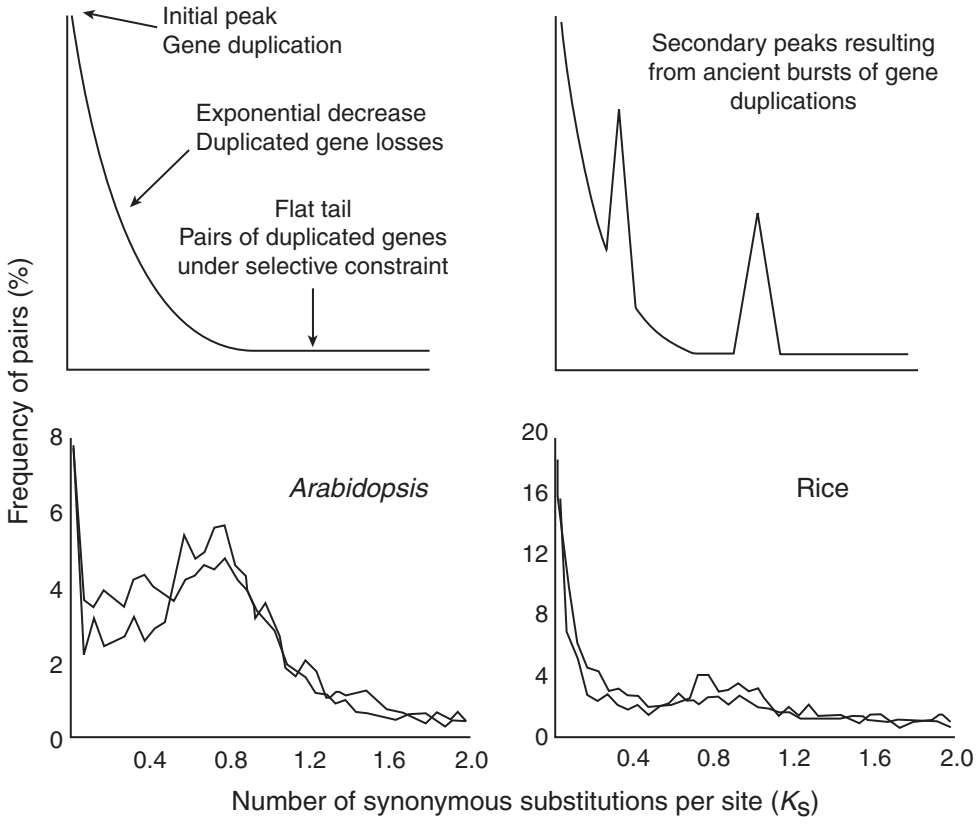


Figure 2.26 Top: theoretical age distributions of pairs of duplicated genes in a genome. The general decrease of the curve indicates that fewer and fewer genes remain as recognizable duplicate pairs with increasing time since duplication (measured by the number of synonymous substitutions per synonymous site, K_s). Peaks in the curve are indicative of 'cohorts' of synchronous duplications. Bottom: distribution of K_s values of paralogous gene pairs in *A. thaliana* (left) and *O. sativa* (right). Distributions are shown for genomic gene sequences and for partial, sequenced cDNAs (ESTs). These two approaches result in practically the same pattern. The peak in the *Arabidopsis* curve around $K_s = 0.7–0.8$ is indicative of an ancient polyploidy event. In the rice genome the distribution conforms mostly to the theoretical prediction of the top-left panel. After Blanc and Wolfe (2004). Copyright American Society of Plant Biologists.

duplication process (Fig. 2.26). *Arabidopsis* obviously has a peak in the frequency distribution around a K_s value of 0.8, which is indicative of synchronized duplication of many genes together. The most likely explanation for synchrony is a polyploidization of the whole genome, dated around 25–26.7 million years ago. This was followed by extensive rearrangements and an accelerated loss of genes, with the consequence that the *Arabidopsis* genome is now relatively small among plant genomes (Table 2.9) and constitutes a complicated mosaic of duplicated genes. In rice, the distribution of K_s values is much more similar to the theoretical expectation following from a continuous process of individual duplications; however, there is a small elevation in the distribution, which according to Blanc and Wolfe (2004) is indicative of a partial chromosomal duplication dated at 70 million years (Fig. 2.26).

Rice was the second higher plant species with a completely sequenced genome. In fact, two different projects were conducted, one by Syngenta focusing on the *japonica* subspecies (Goff *et al*. 2002), and one by the Beijing Genomics Institute, focusing on the most widely cultivated subspecies in China, *O. sativa indica* (Yu *et al*. 2002). The *indica* genome was 466 Mbp in size, with the number of genes

estimated to be between 46 022 and 55 615; the *japonica* data were similar. Again these counts show that the number of genes in plants can be much higher than in animals. Rice and *Arabidopsis* belong to two different lineages of angiosperm plants, the monocotyledons and dicotyledons, which diverged around 200 million years ago; however, despite this ancient evolutionary divergence, there appears to be a considerable degree of homology between individual genes. Goff *et al*. (2002) estimated that 85% of the *Arabidopsis* predicted proteins had a homologue in the rice genome and that 31% of the proteins shared between *Arabidopsis* and rice were not found in fruit fly, nematode, or yeast. Almost all genes related to disease resistance in *Arabidopsis* are also found in rice. These data show that the defence against pathogens is a very basic element of plant biology and is highly conserved between dicotyledons and monocotyledons.

Despite the large number of orthologues shared between *Arabidopsis* and rice, the degree of synteny between these two species is very limited. There is, however, a great deal of genome synteny (colinearity) between the species of the tribus Triticaceae, which in addition to rice includes wheat, barley, rye, and some wild plants of the genus *Aegilops* (goatgrass). Analysing the genetic maps of the Triticaceae, Devos and Gale (2000) showed that only two chromosomal rearrangements need to be assumed to achieve colinearity between the genome of *Aegilops tauschii* (Tausch's goatgrass) and *Hordeum vulgare* (barley), whereas seven rearrangements can explain the relationship between *Ae. tauschii* and rye (*Secale cereale*). Similar syntenic relationships hold for the family Poaceae in general. So the sequence of the relatively small rice genome allows identification of chromosomal segments in other species. However, on a smaller scale (microsynteny), numerous discontinuities in gene order between wheat and rice were identified by Sorrells *et al*. (2003), so the use of rice as a model for cross-species gene isolation in other Triticaceae could prove to be limited.

The website for the Arabidopsis Information Resource (TAIR; http://arabidopsis.org) allows researchers to search for genes, proteins, alleles, markers, and so on, and provides various analysis tools, such as sequence viewers, map viewers, BLAST protocols, and microarray analysis. There are also a great number of links, for example to the Arabidopsis Biological Resource Center, which has thousands of stocks in the form of clones or seeds, which are shipped around the world. The website includes a search engine for publications on *Arabidopsis* genomics in the widest sense. A frequently used platform for transcription profiling in *Arabidopsis* is the Affymetrix Arabidopsis Genome Array ATH1, which has probes for 24 000 genes. The most comprehensive Arabidopsis microarray is the GeneChip® Arabidopsis Tiling 1.0R Array, which has 3.2 million probe pairs tiled across the complete non-repetitive genome. With this array novel transcripts can be identified and advanced methods such as whole-genome chromatin immunoprecipitation experiments can be applied.

2.3.6 The deuterostome lineage

Within the animal kingdom, the protostomes exhibit by far the greatest diversity of body plans. We have seen earlier in this chapter that the greatest attention in invertebrate genome sequencing is paid to nematodes and arthropods, two phyla that belong to the Ecdysozoa lineage of the Protostomia. The other main protostomian lineage, Lophotrochozoa, has received relatively little interest from genome researchers. Still, this clade contains large phyla such as Mollusca and Annelida, and the question may be raised as to what these animals can tell us about the ancestral body plan of the first bilatarian animals, the so-called *Urbilateria*.

That lophotrochozoan invertebrates indeed represent an ancestral form of animal body plan is indicated by the remarkable vertebrate-like features of the marine annelid worms *Platynereis dumerilii* and *Pomatoceros lamarckii* (Miller and Ball 2009). Firstly, the genes of *Platynereis* are rich in introns, a situation they share with vertebrates, not with *C. elegans* or *Drosophila* (Raible *et al.* 2005). Secondly, the immune-related transcriptome of *Platynereis* shows closer phylogenetic relationships with deuterostomes than with ecdysozoans (Altinicek and Vilcinskas 2007). Thirdly, *Platynereis* has been shown to have a ParaHox gene cluster with a relatively

complex expression pattern like that in higher animals (Hui *et al.* 2009). This evidence suggests that to some extent the lophotrochozoan genome organization represents the Urbilateria better than the derived genomes of nematodes and fruit flies. In fact, the genome of the sea anemone, *Nematostella vectensis*, also shows some remarkable architectural similarities with vertebrates, more than with *C. elegans* and fruit flies (Putnam *et al.* 2007). The basal features of animal gene repertoire and genome organization were already present in the first eumetazoans, with subsequent losses and specializations in many later lineages. There is still a world to win for comparative genomics of invertebrates outside ecdysozoans.

Genomes of higher animals are discussed jointly here with reference to the subkingdom Deuterostomia, which includes the phyla Pterobranchia, Echinodermata, Hemichordata, and Chordata. The California purple sea urchin, *Strongylocentrotus purpuratus*, is the best investigated model of the invertebrate deuterostomes. Its phylogenetic position within the Bilateria, relative to other genomic models, is given in Fig. 2.27.

Analysis of the sea urchin genome produced many surprises (Sea Urchin Genome Sequencing Consortium 2006). Almost all gene families found in vertebrates are also found in this simple animal. Surprisingly this includes genes that in vertebrates are associated with vision, balance, hearing, and chemoreception. Does the sea urchin possess hitherto unexplored sensory capacities? It is hard to imagine. Several gene families have undergone expansion in the sea urchin, independent of expansion in the vertebrates. A remarkable case concerns the innate immune system, which is extremely extensive. Around 4 to 5% of the sea urchin's genome is involved with immune functions. This includes a vast family of Toll-like receptors, a large family of NLR (NACHT and leucine-rich repeat proteins) genes, and a similarly large family of SRCR (scavenger receptor cysteine-rich domain proteins) genes. Why sea urchins would need such a sophisticated immune system is not known. It might have something to do with the long lifetime (up to 100 years) that this animal can attain (Rast *et al.* 2006), but it might also be a purely neutral (non-adaptive) feature (Lynch 2007a).

Strongylocentrotus shows the greatest number or reciprocal matches between genes with the human and mouse genomes (around 25% of the genes have a genuine orthologue in the other genome), see Fig. 2.28. The number of reciprocal pairs between sea urchin and mouse is 1.5 times the number of matches with fruit flies, and the number of matches with nematodes is even lower. These comparisons show how strongly the echinoderms are positioned in the deuterostome lineage, together with the chordates. The initial idea of a vertebrate–echinoderm link, proposed in 1908, was based solely on

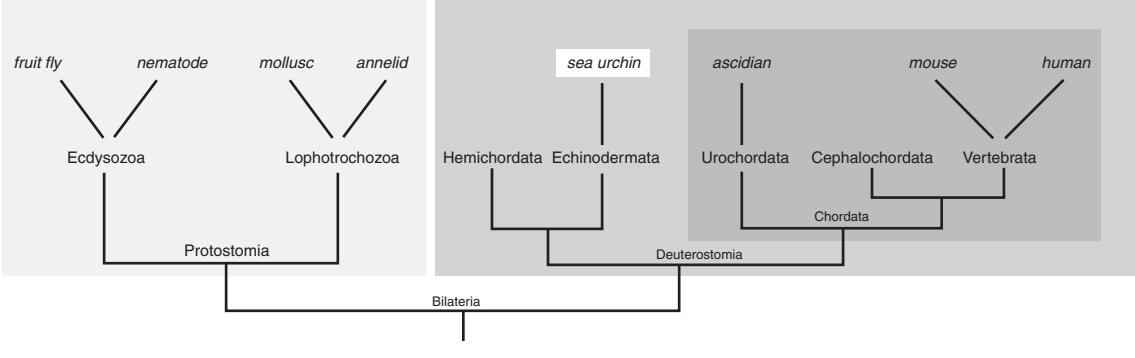


Figure 2.27 Phylogenetic position of various animal models in the overall evolutionary scheme of the Bilateria, highlighting the sea urchin. Please note that the early branching of the chordates is under discussion: there is evidence to show that urochordates, not cephalochordates, are a sister group of the vertebrates (Bourlat *et al.* 2006). Reproduced from Sea Urchin Genome Sequencing Consortium (2006). Copyright 2006 AAAS.

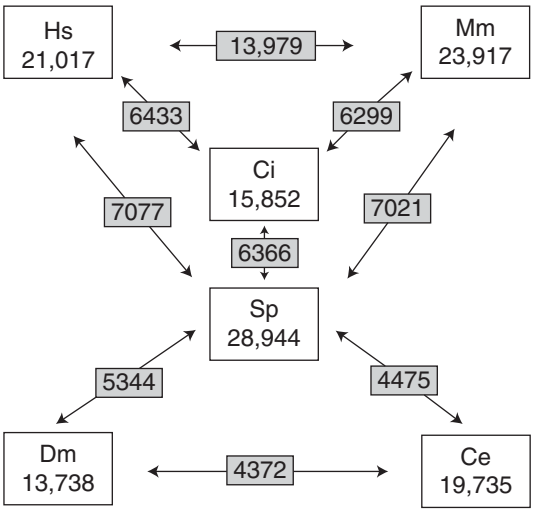


Figure 2.28 Reciprocal matches (genuine orthologues) between six genomic model species, illustrating the strong ties of sea urchins and sea squirts within the deuterostome lineage. The number of orthologues is indicated along the arrows, and the total number of International Protein Index Database gene sequences is given in the boxes. Hs *Homo sapiens*, Mm *Mus musculus*, Ci *Ciona intestinalis*, Sp *Strongylocentrotus purpuratus*, Dm *Drosophila melanogaster*, Ce *Caenorhabditis elegans*. Reproduced from Sea Urchin Genome Sequencing Consortium (2006). Copyright 2006 AAAS.

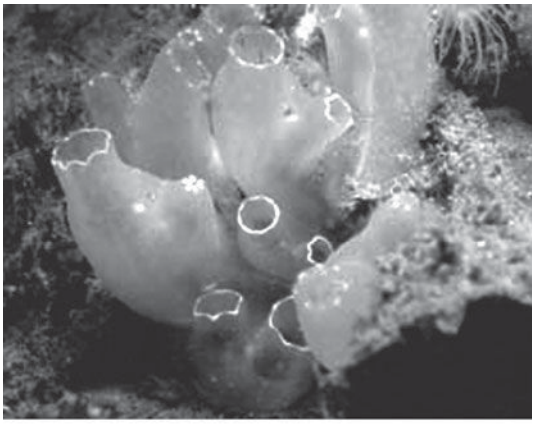
morphological similarity of the embryos. Modern comparative genomics now places this proposal on a very firm footing. While *Strongylocentrotus* itself is a model mainly for developmental biology, the availability of its genome will undoubtedly support genomics research in the marine coastal environment, where sea urchins make up an often conspicuous element of the ecology.

An interesting view of the origin of vertebrates is obtained from the genomes of the two invertebrate chordates sequenced so far, the sea squirt, *Ci. intestinalis* (Dehal *et al*. 2002; Cañestro *et al*. 2003), and the lancelet (amphioxus), *Branchiostoma floridae* (Putnam *et al.* 2008). Sea squirts belong to the chordate subphylum Urochordata, also called Tunicata, after the tunic, a tough fibrous cover excreted from the skin in which the animal is contained. Lancelets belong to another chordate subphylum, Cephalochordata. For a long time the lancelets were considered as holding the key to the vertebrate body plan, mostly because of the pharyngeal gill slits, a feature of the chordate zootype. So strongly was the vertebrate body plan exemplified by amphioxus in zoology classes, that students attending a summer school at the Marine Biological Laboratory at Woods Hole on Cape Cod composed a song on it, following the World War I tune 'It's a Long Way to Tipperary': 'It's a Long Way from Amphioxus'. This song today would be neglecting the precise evolutionary relationships within the chordates, because nowadays it is agreed that Urochordates, not Cephalochordates, are a sister group to the chordates (Bourlat *et al.* 2006), cf. Fig. 2.27.

As an adult, *Ci. intestinalis* is sessile and attached to an underwater substrate where it filters food particles by pumping water through its elaborate pharynx, a basket-like structure, which fills most of the tunic. The name squirt is due to the regular pulses of water driven out of the exhalent siphon (Fig. 2.29a). The larva of a sea squirt looks very much like a jawless fish, and is equipped with a chorda and a dorsal nerve cord, externally resembling a tadpole (Fig. 2.29b). Unlike *Ciona*, amphioxus does not have an elaborate metamorphosis and so its relationship with the chordate body plan is more obvious from its external morphology.

Ci. intestinalis is a solitary, small, and relatively short-lived marine animal that colonizes solid substrates in the sublittoral zone, such as protected rocky shores, ship wrecks, and buoys. Due to its rapid colonizing capacity, it is sometimes a conspicuous and abundant representative of the 'fouling' community. With their large filtration capacity, the animals act as filters and so contribute to purification of coastal waters, although by the same mechanism they accumulate chemicals and are used for biomonitoring of coastal sea pollution. Ecological work on *Ciona* and other tunicates aims at answering questions about settlement in relation to density and intraspecific competition. Local populations seem to be highly dynamic and are characterized by cyclic retreat and recolonization events. Because of this type of population dynamics, ecologists are interested in geographical population genetic structure; microsatellite markers have been developed to support such analyses (Procaccini *et al*. 2000). The genomic information on *Ciona* has, however, not yet penetrated into ecological studies.

(a)



(b)

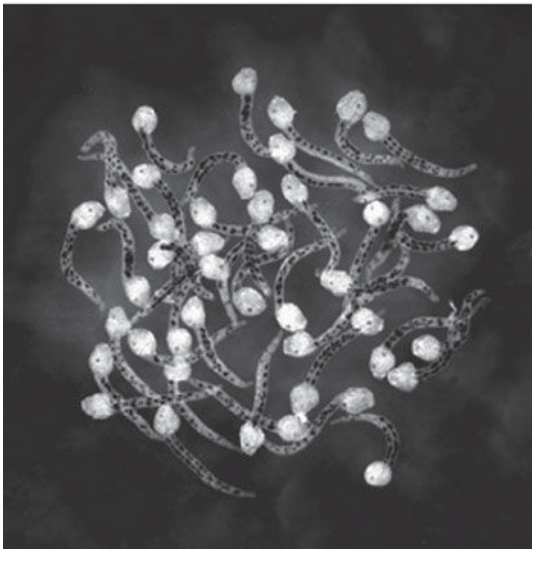


Figure 2.29 (a) Adult sea squirts. (b) A group of larvae. David Keys (photo) and Leila Hornick (artistic rendering), courtesy of the U.S. Department of Energy Joint Genome Institute. © 2005 The Regents of the University of California.

The genomes of tunicates are considerably smaller than those of vertebrates, and *Ciona*'s genome measures about 160 Mbp (Dehal *et al*. 2002). The gene content represents an interesting blend between ancient protostome signatures and chordate innovations, with some tunicate autapomorphisms added. Dehal *et al*. (2002) found a total of 15 852 protein-encoding genes and these were compared with the gene complements of *Drosophila*, *C. elegans*, puffer fish, and mammals. It turned out that 60% of the genes shared homology with fruit flies and nematodes, so these represent the core physiological and developmental machinery that is common to all bilaterian animals. A few hundred of these genes have a stronger similarity to fruit fly or nematode than to any vertebrate, and so these genes represent functions that were present in the invertebrates, but were lost in the vertebrate lineage. Examples are chitin synthase (there is no chitin exoskeleton in chordates), phytochelatin synthase (the role of the zinc-binding molecule phytochelatin was taken over by metallothionein), and haemocyanin (the copper-containing blood pigment of arthropods and bivalves, absent from vertebrates). Another 16% of the genes lacked a homologue in the protostome groups, but had a clear vertebrate counterpart. These genes apparently have arisen on the deuterostome branch before the split between tunicates and vertebrates. Then another 21% of the genes had no clear homologue in fruit fly, nematode, fish, or mammal and represent tunicate-specific genes.

Interestingly, *Ciona*'s genome has genes related to the synthesis and degradation of cellulose (cellulose synthase and several endoglucanases), genes that are only found in bacteria and plants, never in animals, except nematodes, which gained them by lateral gene transfer, see section 2.2.2. The presence of these genes in *Ciona* is related to the composition of the tunic, which is built largely of a cellulose-like carbohydrate called tunicin. How *Ciona* obtained these genes (a dramatic example of lateral gene transfer?) remains a mystery, but obviously it has been a very significant event in the evolution of this group (Matthysse *et al*. 2004). *Ciona*'s genome has all the genes related to the innate immune system, as in *Anopheles* and *Drosophila*, but genes implicated in adaptive immunity could not be found. This suggests that the adaptive immune system is an apomorphy of the vertebrates, not of the chordates as a whole. Ascidians are also known for their extremely high body concentration of vanadium, several orders of magnitude higher than any other animal. Vanadium is accumulated in specialized blood cells, vanadocytes, where it is localized in

intracellular vacuoles, together with a similarly high concentration of sulphate. Three vanadium-binding proteins, *vanabins*, have been characterized in *Ascidia sydneiensis samea* (Ueki *et al*. 2003) and five vanabins are encoded in the genome of *Ci. intestinalis* (Trivedi *et al*. 2003). However, a genome-wide analysis of the peculiar vanadium metabolism of ascidians has not yet been conducted.

Turning our attention from urochordates to vertebrates, we note that several species of fish presently serve as genomic models: *Takifugu rubripes*, *Tetraodon nigroviridis* (both puffer fish, family Tetraodontidae), *Danio rerio* (zebrafish, family Cyprinidae), *Oryzias latipes* (medaka, family Adrianichthyidae), *Gasterosteus aculeatus* (three-spined stickleback, family Gasterosteidae), and so on. To a certain extent, genomic information on zebrafish and fugu can be extrapolated to other fish species as long as genes are conserved. This cannot be stretched too far, probably not beyond the family Cyprinidae for the zebrafish. Ecologists working on other species will need to avail themselves of genomic sequences for their own model, but the community of fish biologists is rather fragmented (Clark *et al*. 2003). EST databases and microarrays are being developed for a considerable number of species including atlantic salmon (*Salmo salar*), carp (*Cyprinus carpio*), largemouth bass (*Micropterus salmoides*), fathead minnow (*Pimephales promelas*), killifish (*Fundulus heteroclitus*), and long-jawed mudsucker (*Gillichtys mirabilis*). We will see many examples, especially in Chapters 5 and 6, in which fish were used to exemplify aspects of ecological and evolutionary genomics.

Takifugu rubripes (also known as *Fugu rubripes*) was proposed in 1993 by Sydney Brenner as a genomic model because with its small genome (470 Mbp) it would allow a cost-effective way of illuminating the human genome. In the far east the fish is not only known for its small genome but also for containing the extremely toxic compound tetrodotoxin, which, with an oral LD_{50} to mammals of 15 µg per kg of body weight, is one of the most potent toxins known. Japanese men practice the habit of eating 'fugu' fish in restaurants that have obtained a special licence allowing the cook to separate the flesh from the hypertoxic liver and ovaria. The International Fugu Genome Consortium was formed in the year 2000, coordinated by the Institute of Molecular and Cell Biology in Singapore, in collaboration with groups in the UK and the USA (www.fugu-sg.org). The sequence was released less than two years later (Aparicio *et al*. 2002).

Because the *Takifugu* genome assembly remained highly fragmented, another team, coordinated by the French sequencing centre Genoscope, started on a related puffer fish, *Te. nigroviridis*. This species has an even smaller genome and it offered the additional advantage of being a popular aquarium fish, easily maintained in tap water. The name puffer is derived from the fish's habit of inflating itself when it is threathened. The analysis of the genome, published by Jaillon *et al*. (2004), revealed several interesting trends about gene duplications in the actinopterygian fish lineage (ray-finned fish, as opposed to lobe-finned fish, the Sarcopterygii, such as lung fish and coelocanths). The genome of *Tetraodon* measured 342 Mbp and had 28 918 putative protein-encoding genes, 1.8 times more than in *Ciona* but somewhat less than in *Takifugu* (31 059). The slightly smaller genome size was ascribed to the absence of transposable elements, which are rather abundant in fugu fish. Careful analysis of the content of each of the 21 *Tetraodon* chromosomes allowed reconstruction of a duplication event in the actinopterygian lineage (Fig. 2.30). Assuming that the original number of chromosomes of the ancestral gnathostome (jawed fish) was 12, a duplication event, followed by 10 different chromosomal rearrangements (fusions and translocations), can explain the present organization of the 21 chromosomes. The duplication is assumed to have taken place later in the evolution of the Actinopterygii, close to the origin of the Teleostei (modern bony fish), because some early-branching actinopterygian fish (bichirs, Polypteriformes) do not have the duplication. Similar conclusions were reached by Christoffels *et al*. (2004) in an analysis of the fugu genome.

The model of Jaillon *et al*. (2004) is consistent with an earlier analysis of the vertebrate *Hox* genes by Amores *et al*. (1998). These authors had sequenced all 50 *Hox* genes of zebrafish and analysed paralogous and orthologous homologies across zebrafish, fugu fish, and mouse. The pattern of *Hox*-gene

clustering could be explained by assuming that the ancestor of the Gnathostomata lineage had four clusters of *Hox* genes, each cluster consisting in principle of 13 genes. This system was continued in the Tetrapoda (amphibians, reptiles, birds, and mammals), but several losses led to a total of no more than 40 genes in the mouse, still arranged in four clusters. In the Actinopterygii lineage a duplication, identical to the one discussed by Jaillon *et al*. (2004), was assumed to have taken place, leading to eight clusters, followed by the loss of one cluster and several individual genes, leading to 50 *Hox* genes in zebrafish, arranged in seven clusters. How the ancestral gnathostome acquired its complement of four clusters, through two rounds of whole genome duplication (the 2R hypothesis), or through two local duplications, is not yet resolved (see the discussion in Section 2.1.). The duplications of the *Hox* genes, both in the early evolution of the chordates and in the actinopterygian lineage, may have spurred innovation of the body plan and subsequent radiation of the highly successful vertebrate groups (Venkatesh 2003).

Although the two puffer fish had a head start as genomic models because of their unusually small genomes, the ultimate fish model is the zebrafish, *D. rerio*. This species has many experimental advantages, including ease of culture, a transparent embryo, and ample possibilities for manipulations such as cell labelling, transplantation, microinjection, and mutagenesis. Genome analysis of zebrafish started with the production of extensive genetic maps, to accelerate the molecular localization of mutations, and to allow comparisons of genome location with other vertebrates (Woods *et al*. 2000). The Zebrafish Information Network (ZFIN) now

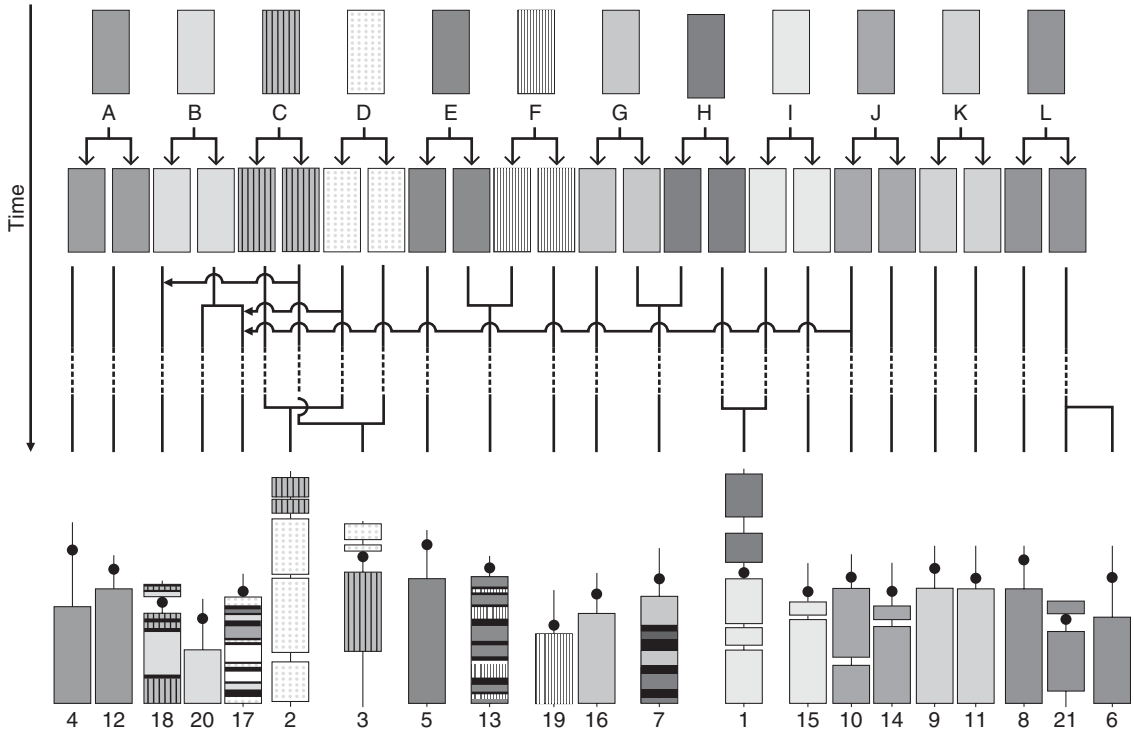


Figure 2.30 Model, inspired by detailed genomic analysis of the puffer fish *Te. nigroviridis*, showing how the present 21 chromosomes of teleost fish can be derived from an ancestral gnathostome karyotype with 12 chromosomes, by assuming a whole-genome duplication event followed by 10 different rearrangements of chromosomal segments (fusions and translocations). After Jaillon *et al*. (2004), by permission of Nature Publishing Group.

serves as the zebrafish model organism database. The design of the network (http://zfin.org) is described by Sprague *et al*. (2001); it aims to maintain the definitive reference datasets of zebrafish research information, and to facilitate the use of zebrafish as a model for human biology.

The genomes of amphibians and reptiles are hardly explored at the moment, despite the fact that these animals are among the most popular ecological study objects. The first genome sequence for an amphibian, the Western clawed frog, *Xenopus tropicalis*, was published only in 2010 (Hellsten *et al*. 2010). This sequence will also aid investigators that study its more popular cousin, the African clawed frog, *Xenopus laevis*. A community of *Xenopus* investigators has developed an information resource, Xenbase (www.xenbase.org), where data on the sequence are offered along with genomic tools, as well as an archive of basic biological information about clawed frogs including animations showing anatomical features and developmental patterns. Sequencing *Xenopus* was mostly inspired by the eminent possibilities that the animal offers for studies into early embryonic development and cell biology; however, it may also have some relevance to ecological studies in herpetology.

Regarding birds, two species stand out, the red jungle fowl (*Gallus gallus*) (International Chicken Genome Sequencing Consortium 2004) and the zebra finch, *Taeniopygia guttata* (Warren *et al.* 2010). The red jungle fowl, native to Southeast Asia, is the ancestor of the various domestic breeds of chicken. The chicken genome will be an important resource for poultry science and applied avian studies. The zebra finch is an important model for neurosciences because of its communication through learned vocalizations. Thus, neither of these two is an ecological model in itself; however, the genomes of chicken and zebra finch may provide important templates for assembly and annotation of the genomes of other birds.

A tremendously rich system for investigating questions of population structure, life history, and behaviour is offered by wild bird species such as the great tit (*Parus major*); sequencing such species would be a real breakthrough for ecological genomics. Luckily, genomics resources such as genetic maps and genome-wide SNP markers are being developed for this species (Van Bers *et al.* 2010). Interestingly, assembly of the short reads obtained in this study with Illumina 1G sequencing methodology was greatly supported by alignments to the zebra finch and the chicken genome. This is due to the fact that genome size and architecture have remained more or less constant during the evolution of birds, and so syntenic relationships have been maintained over relatively wide phylogenetic distances. Birds as a whole have small genomes, a feature which was originally ascribed to the metabolic efficiency associated with flight, but is now known to root deep in the saurisschian origin of birds, even before the animals could fly (Bonneaud *et al.* 2008).

Many mammalian genomes are being sequenced, some because of their relevance to husbandry (cattle, pig, horse, dog), others because of their medical importance (mouse, rat, rhesus macaque), their phylogenetic closeness to man (chimpanzee, gorilla, orangutan), their conservational value (elephant), or their crucial position in the tree of life (platypus). Although each of these genomes has its own exciting features we do not discuss them in this book because none of the species is an important ecological model in itself. Still, these genomes are relevant to ecological genomics as a reference; they may provide support in assembly and gene annotation for ecological models. For example, population genomics using microsatellite markers in the wolf was greatly aided by the availability of the dog genome (Hagenblad *et al.* 2009). Likewise, studies on voles, monkeys, and deer will be aided by the genomic information of mouse, macaque, and cattle, respectively.

This completes our overview of prokaryotic and eukaryotic model genomes and the promises that they hold for ecology. The field of comparative genomics is rapidly growing and we believe that many discoveries are still in store. Comparative genomics provides an indispensable evolutionary foundation for the still teneral state of ecological genomics.

CHAPTER 3

Structure and function in communities

In this chapter we will address one of the most fundamental issues of ecology: the relationship between ecosystem processes and species richness in communities, or as Lawton (1994) put it, 'What do species do in ecosystems?'. Ecological genomics opens new avenues to explore this question. We will review scientific evidence concerning genome diversity in the environment and the function of genomes in nutrient cycles. Because microorganisms are in a key position at many crucial links of nutrient cycles, most of this chapter will deal with the ecological genomics of microorganisms.

3.1 The biodiversity and ecosystem functioning synthetic framework

A summary of the ecological framework that forms the background to this chapter is given by Naeem *et al.* (2002). At the beginning of the 1990s ecologists reformulated a question that had already existed for a long time in ecology; namely, what is the relation between structure and function in ecosystems? Structure includes all quantities that can still be observed in a snapshot of the system at a particular moment in time. This includes things like species richness, biomass, dominance structure, and feeding groups. The functional aspects include the processes that cannot be observed in a snapshot but need to be monitored in time, such as primary production, respiration, degradation of organic matter, and nitrification. With the staggering loss of biodiversity that we observe today, the question may be asked, how will ecological functions respond? Conversely, if we are interested in protecting functions, is it possible to achieve this aim through protecting the structure? An answer to these questions requires a scientific underpinning of the ecological importance of biodiversity.

Since 1993, when a group of scientists congregated in Bayreuth and the seed for the 'biodiversity and ecosystem functioning synthetic framework' was planted (Schulze and Mooney 1993), the role of biodiversity in maintaining ecological functions has been subject to intense theoretical and experimental analysis. These developments were also spurred by the Rio Convention on Biological Diversity held in 1992, followed by the spreading realization that global biodiversity is under serious threat. The issues were addressed by theoretical models, food-web analysis, microcosm experiments, and field-plot investigations.

In general it is assumed that there is an asymmetrical relationship between structure and function; that is, protection of functions does not require protection of all structures, whereas on the other hand protection of structures will always guarantee protection of functions. However, it is still a matter of debate what kind of form this asymmetrical relationship should take. Several alternative hypotheses have been formulated that differ from each other in the extent to which a decrease in the number of species endangers an important function of the system. The hypotheses are discussed in terms of graphs in which some ecosystem process is plotted as a function of the number of species in the ecosystem (Fig. 3.1). The argument is, what happens to ecosystem function (plotted on the vertical axis) when biodiversity (plotted on the horizontal axis) decreases or increases? Although Fig. 3.1 pictures 6 different relationships, other authors have distinguished no

less than 50 different hypotheses, which can, however, be categorized in three main classes (Lawton 1994; Naeem *et al.* 2002), as follows.

The redundant species hypothesis. With a decrease in biodiversity, ecosystem functions are unaffected up to the point where only a small number of key species remains; if one of these species is removed, the system collapses. The idea is that many species in the ecosystem are redundant in the sense that they are at least partly substitutable while their contribution to the ecosystem process can be taken over by other, functionally similar, species.

The rivet hypothesis. With a decrease in biodiversity, ecosystem functions decrease proportionally (linearly or in steps). The idea is that every species makes a (smaller or larger) contribution to the process, so if it is removed, that contribution is subtracted from the process.

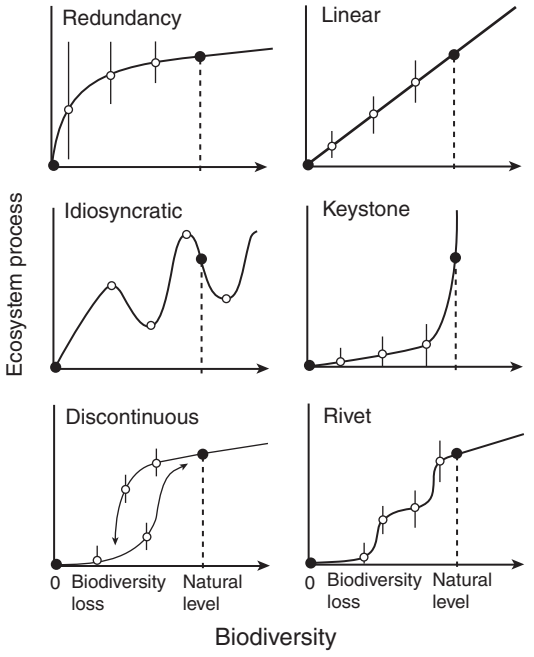


Figure 3.1 Graphical representation of six different hypotheses about the relationship between biodiversity and ecological processes. The central idea is that, commencing with the natural level of biodiversity and moving in the direction of a decrease (to the left), there are different ways in which the function of a system can change; in every case it ends at a zero level when no biodiversity is left. From Naeem *et al.* (2002), with permission from Oxford University Press.

The idiosyncratic hypothesis. There is no universal relationship between structure and function; rather the relationship is ecosystem-specific. In one case there may be a strong reduction of function with a loss of biodiversity, in another case there may hardly be an effect or maybe even an increase.

The general feeling among ecologists is that *functional redundancy* indeed plays an important role in many ecosystems. The argument is supported by observations on systems in which members of the community are suppressed by toxicants. In soil ecology it is well known that respiration is considerably less sensitive to the effects of toxic substances than nitrification, which is attributed to the fact that all heterotrophic organisms contribute to respiration, while only a few bacterial genera are responsible for nitrification (Domsch 1984). In studies of heavy-metal contamination in forest ecosystems, it has been demonstrated that in a gradient of pollution around a metal-smelting works a considerable loss of species of fungi can occur, whereas soil respiration is hardly affected and decreases only at the very high levels of pollution close to the source (Nordgren *et al.* 1983).

Still, after more than a decade of ecological research the central question of the synthetic framework cannot be answered in a simple way. It is still very difficult to refute any one of the six hypotheses of Fig. 3.1. In very general terms, one may conclude that a minimal number of species is necessary to allow a system to function, and a larger number of species is necessary to guarantee stability of the processes in a changing environment (Loreau *et al.* 2001). In addition, two points have emerged showing that analysis of the problem may benefit from narrowing down the scope of the question.

First, it appears that the way in which biodiversity influences ecological processes depends on the way in which these processes are limited. This issue is of evident importance in aquatic ecosystems, in which primary production can either be limited by the substrate (nutrient loading, such as phosphorus or carbon), or by the capacity of the organisms to process that substrate (the biomass and the number of producer species). In *capacity-limited systems* the substrate is supplied rapidly enough so that every functional unit is saturated and the rate of through-

put is insensitive to changes in the rate of substrate supply. An example is nitrogen fixation in cyanobacteria, which is usually not limited by the availability of nitrogen gas but by the biomass of fixating organisms, which in turn are limited by other factors (e.g. zooplankton grazing, phosphate). A capacity-limited process is sensitive to a loss of biodiversity, because any reduction in the number of functional units will decrease the process rate. In a *substrate-limited system*, the capacities of the functional units are not fully deployed and if such a system loses biodiversity the overall throughput may remain unchanged because each functional unit can easily increase its share in the process (Levine 1989; Van Straalen 2002). Substrate limitation may be less important in terrestrial ecosystems and even less in below-ground systems, because of the abundance of dead organic matter as a food source for the decomposer community. However, even soil communities may be limited, for example by nutrient imbalance (Pokarzhevskii *et al.* 2003) and by microhabitat heterogeneity causing spatial dislocation between the food and the hungry.

Second, there is an increasing awareness that biodiversity as such is not as important as biodiversity in relation to the properties of the species. That is, to evaluate the effects of diminishing species richness on ecosystem processes we must look at the biodiversity of ecological traits in a community, not only at species numbers (Berg and Ellers 2010). As an example, consider the work by Walker *et al.* (1999), who investigated vegetation structure in Australian savannahs. The authors noted that dominant plant species in the same community tend to be positioned apart from each other when classified according to species-specific attributes, such as height, biomass, specific leaf area, longevity, and leaf-litter quality. Rare species may contribute to resilience of the vegetation because they often have attributes similar to the dominant species and may act as a functional substitute. Another way to phrase the issue is the *principle of complementarity*: the stability of the system benefits if species complement each other in their function. The argument extends to communities of decomposer invertebrates. In microcosm experiments with earthworms, isopods, and millipedes Heemsbergen *et al.* (2004) demonstrated that the effect of detritivore invertebrates on soil respiration and litter breakdown depended not on the species composition per se, but on the *functional dissimilarity* within that community. The suggestion was that positive interactions in the community cause a functionally dissimilar assembly to have a larger effect on soil processes than a functionally similar assembly, independent of the number of species.

The biodiversity and ecosystem functioning synthetic framework has not yet been probed using the genomics approach. Yet there is a lot of mechanistic knowledge, especially in microbiology, about the actors behind biogeochemical cycles. This knowledge has increased tremendously with the large-scale sequencing and transcription profiling of microbial genomes. In the sections below we explore a possible link between the two fields of investigation; community ecology and microbial functional genomics.

3.2 Measurement of microbial biodiversity

To estimate the number of species that are present in a specific habitat is more difficult than it may seem. The situation is aptly described by the following phrase from the classical book by Charles Elton (Elton 1927), the founder of animal ecology:

> Two boys of rather good powers of observation were sent into a wood in summer to discover as many animals as they could, returned after half an hour and reported that they had seen two birds, several spiders, and some flies—that was all. When asked how many species of all kinds of animals they thought there might be in the wood one replied after a little hesitation 'a hundred', while the other said 'twenty'. Actually there were probably over ten thousand.

Now it is obvious that if Elton's boys had been asked to include microorganisms, their estimate would have been even more inaccurate. For obvious reasons, estimating the biodiversity of microorganisms is more difficult than estimating species richness of plants or animals. In addition, microbiologists struggle with an even more fundamental question; that is, it is often not clear what constitutes a microbial species.

According to classical bacteriological taxonomy, an isolate is recognized as a proper species if its morphology is described plus some key aspects of its metabolism (trophic system, substrate use, etc.). Two isolates are considered to belong to the same species if their DNAs are similar by more than 70% or if there is less than a 5 °C difference in the melting temperature of a DNA–DNA hybridization (Wayne *et al.* 1987). Obviously, species that cannot be put into pure culture cannot be characterized in this way. It is estimated that anything between 50 and 99% of microorganisms may belong to this group of *unculturables* and these remain undescribed as species, although fragments of their genome may be sequenced from the environment. Why so many organisms cannot be cultured in the laboratory is unclear and probably there are many reasons, including specific growth conditions, unknown nutrient requirements, very slow growth, and special surfaces to which cells must attach. Still, microbiologists have discovered that some 'uncultivable' bacteria can be brought into culture when placed in close proximity to other species, from which they are separated only by a membrane; apparently, chemical signals from other members of the community are sometimes crucial to induce growth (Kaeberlein *et al.* 2002). We will see later in this chapter that genomics approaches provide another solution to the problem: the DNA of species in the environment can be assembled and its functions characterized without even attempting to put them into a culture tube.

The problem of what constitutes a microbial species can also be approached from a comparative point of view (Achtman and Wagner 2008). Konstantinidis and Tiedje (2005) and Konstantinidis *et al.* (2006) compared a number of completely sequenced prokaryotic genomes in an attempt to discover natural dividing lines between species. The often used criterion of 70% similarity in DNA–DNA hybridization appeared to correspond to an *average nucleotide identity* (*ANI*), on the level of the genome, of 95%. Still, some groups of prokaryotes were found to be much more variable than others, and this often relates to their ecology. A clear clustering of strains, with sharp divides between the species, was more obvious in organisms with narrow ecological niches, for example pathogens in a host or microbes living under extreme conditions. In lineages with a broad ecological niche, with strains living under many different ecological conditions, there was more continuity in the genomes and species borders were less well defined. So it seems that to resolve the species problem in prokaryotes, the ecology of the species must play a role in addition to evolution and phylogenetics.

Microorganisms have been given little attention in ecological studies until recently. The last decade has produced a new awareness of microbial diversity and the suitability of microorganisms to address questions of fundamental ecological importance (Øvreås 2000; Horner-Devine *et al.* 2004; Kassen and Rainey 2004; Jessup *et al.* 2004). Microorganisms have been reported from extreme habitats in which they are the only type of organism surviving, such as hot springs, deep ocean vents, volcanic crater lakes, and sediments under permanent ice cover. Such extreme habitats hold many surprises in store. For example, a completely new phylum of Archaea, the Nanoarchaeota, was discovered in a hot submarine vent north of Iceland and a new division of Euryarchaeota was found in a hypersaline anoxic basin in the Mediterranean Sea (Huber *et al.* 2002; Van der Wielen *et al.* 2005). The development of universal phylogenetic trees on the basis of genes that are common to all life forms has demonstrated that the biodiversity of the Bacteria and Archaea is at least as large as that of the whole domain of the Eukarya (Fig. 3.2).

3.2.1 Diversity of small-subunit rRNA genes

In the so-called *polyphasic taxonomy* of current microbiology a species is differentiated on both genetic and phenotypic grounds. The genetic characterization is derived from the sequence of the small-subunit rRNA gene. From basic biochemistry we know that the size of ribosomes may be characterized by Svedberg units (S), a measure of sedimentation velocity during ultracentrifugation (1 S corresponds to 10^{-13} s). The prokaryotic ribosome measures 70 S and is made up of a *small subunit* (SSU) of 30 S, consisting of 21 proteins and an RNA molecule of 16 S, and a *large subunit* (LSU), measuring 50 S, consist-

ing of 34 proteins and two RNA molecules, one 23 S and the other 5 S. In the prokaryotic genome, the genes encoding these RNAs are organized in an *rRNA transcription unit* (*rrn* region), with the 16, 23, and 5 S rRNA genes lying behind each other, separated by spacers and being transcribed as one unit. The 16S rRNA gene (also called the *SSU rRNA gene*) has been chosen as the basic diagnostic instrument of prokaryote phylogeny and classification. The gene is assumed to fall into the category of essential genes, which are not, or at least infrequently, subjected to lateral transfer (see Section 2.2.2).

The size of ribosomal components is slightly different in eukaryotes (Table 3.1). The SSU rRNA

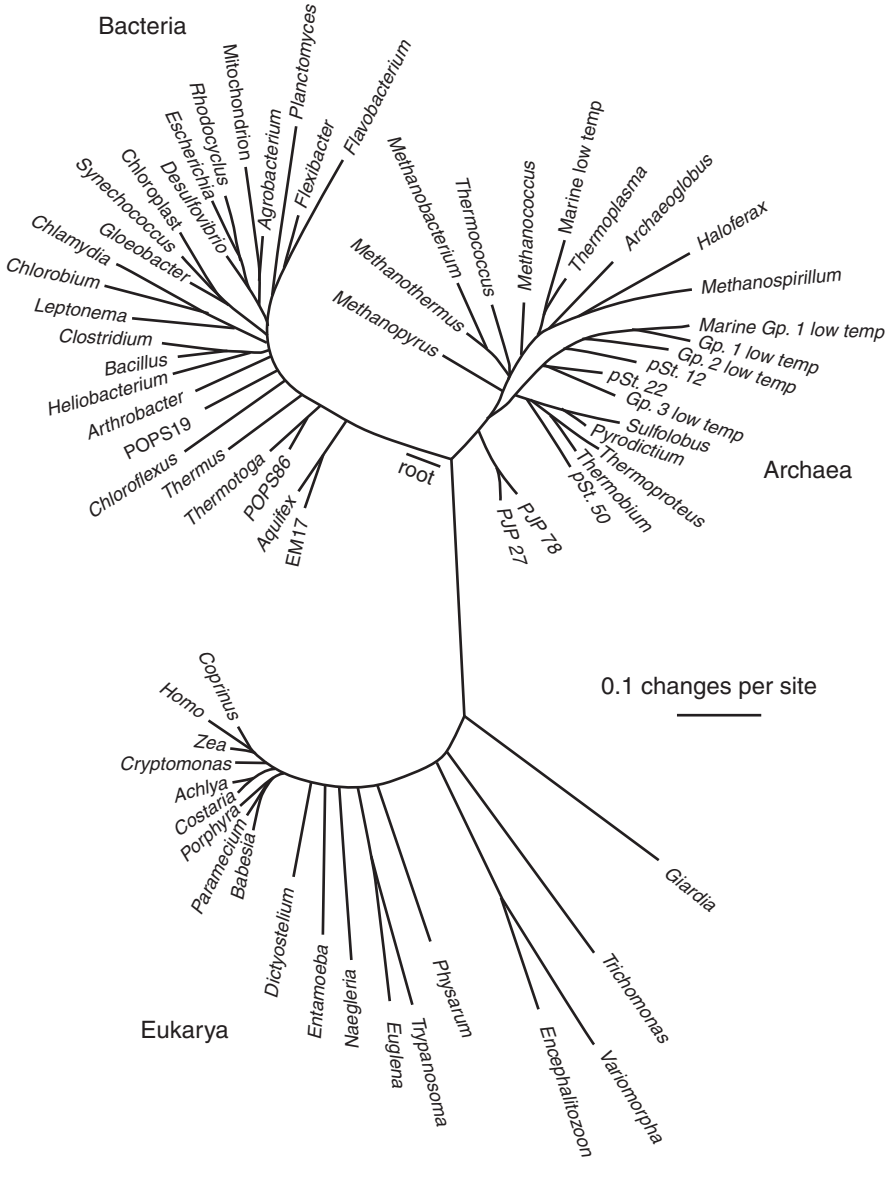


Figure 3.2 The universal phylogenetic tree of life, based on small-subunit rRNA gene sequences. Reprinted with permission from Pace (1997). Copyright 1997 AAAS.

Table 3.1 Composition of ribosomes of prokaryotes and eukaryotes

	Prokaryotes	Eukaryotes
Overall size	70 S	80 S
Size of SSU	30 S	40 S
Proteins in SSU	~21	~30
RNA in SSU	16 S, 1500 bp	18 S, 2300 bp
Size of LSU	50 S	60 S
Proteins in LSU	~34	~50
RNAs in LSU	23 S, 2900 bp	28 S, 4200 bp
	5 S, 120 bp	5.8 S, 160 bp
		5 S, 120 bp

LSU, large subunit; SSU, small subunit. *Source*: From Madigan *et al.* (2002).

is 17–18 S in eukaryotes (18 S in vertebrates) and the LSU rRNA molecule, which is 23 S in prokaryotes, is enlarged to 28 S. In addition, eukaryotes have the prokaryote rRNA genes in their mitochondria and chloroplasts. The fact that all life forms have the same basic organization of rRNA genes allows comparison across domains and the development of phylogenies such as that shown in Fig. 3.2.

The reason why the 16S rRNA is particularly suitable as an anchor for prokaryote classification is that it shows a mosaic of conserved and variable regions. The molecule is shown in Fig. 3.3. A characteristic feature is that the RNA molecule folds into a determinate structure with many short duplex regions as well as hairpin loops. The secondary structure of the molecule is crucial to its function in translation of mRNA to peptides, and so the nucleotide sequences of the duplex regions are highly conserved. Other parts of the molecule can undergo substitutions without change of function and so these parts provide an apomorphic signature, characterizing a certain prokaryotic lineage. There are nine variable regions in the molecule, numbered V1–V9 (Fig. 3.3).

Microbiologists have agreed that a 3% difference in the overall 16S rRNA sequence is to be considered as the species boundary. Justification for this is obtained from the fact that if the genomic DNA of two bacteria hybridizes by more than 70% (see above), the 16S rRNA always differs by less than 3% (Madigan *et al.* 2002). Since the genomic hybridization threshold is considered a valid species boundary, two organisms differing in their 16S rRNA by more than 3% can always be considered different species. However, the converse is not true; there are many valid species differing by less than 3% in their 16S rRNA. For example, in the genus *Bacillus* there is a high degree of similarity among species and some species (*Bacillus cereus* and *Bacillus anthracis*) even have completely identical 16S rRNA sequences.

Because of the diagnostic value of the 16S rRNA gene, an enormous number of sequences have accumulated in the international nucleotide sequence databases (GenBank, EMBL). There is an internationally accepted system for numbering the base-pair positions, modelled on *E. coli*. The Ribosomal Database Project (RDP-II; http://rdp.cme.msu.edu) provides alignment and annotation of a large number of rRNA sequences to serve microbial taxonomy and biodiversity research. In release 10 of RDP-II, an alignment is provided of nearly 1.5 million SSU rRNA sequences, prepared using a program that incorporates constraints from rRNA secondary structure (Cole *et al.* 2006). An annotated collection of more than 2.5 thousand oligonucleotide probe sequences targeting rRNAs is available from probe-Base (www.microbial-ecology.de/probebase; Loy *et al.* 2003). The use of databases for optimal probe design has become a crucial element in the design of oligonucleotide microarrays used for detection of species of functional genes (DeSantis *et al.* 2003; Wagner *et al.* 2007).

The 16S rRNA gene is the basis for a popular method of community profiling called *denaturating-gradient gel electrophoresis* (DGGE). The principle of this method was borrowed from medical research, where it was applied for the detection of mutations. Muyzer *et al.* (1993) suggested that the same method could be applied to PCR-amplified 16S rRNA genes. A PCR was designed that amplifies a segment of the gene, using primers targeting sites of the molecule in which the sequence is conserved across all bacteria (*universal bacterial 16S primers*). The amplicon, however, is chosen to span a variable region, so that when the PCR is applied to an environmental sample containing a community of microorganisms, a mixture of fragments is produced with identical

lengths but diverging sequences. These fragments are separated on the basis of differential sensitivity to denaturation in a gradient of urea (DGGE) or temperature (*temperature-gradient gel electrophoresis*, TGGE). The sensitivity to denaturation is determined mainly by the GC content of the sequence. Sequences with higher GC content will denature later and run further on the gel. To prevent complete separation of the two DNA strands, a so-called *GC clamp* is included in the amplicon by extension of the 5′ primer with 40 bp of a GC-rich sequence; this ensures that the sequence is halted in the gel as a stable, partially melted molecule. Preferably the profile is calibrated by marker samples that contain a mixture of known sequences isolated from a clone library from the same environment.

For many years, the 16S rRNA-targeted DGGE approach was one of the most popular methods with which to profile microbial communities in the environment. An example of such a study is the work by Röling *et al.* (2001). These authors were interested in microbial communities of groundwater and their potential to degrade aromatic compounds (benzene, toluene, xylene), leaching from a landfill. Groundwater samples were taken from bore holes at different distances and microbial communities were profiled using DGGE. Profiles from samples taken in the plume were clearly different from those outside the plume (Fig. 3.4). Sequencing of cloned 16S rRNA gene fragments confirmed the trend. In the uncontaminated area upstream of the landfill the community was dominated by Betaproteobacteria, but directly beneath the landfill Gram-positive bacteria dominated. At the end of the gradient the community was dominated by Deltaproteobacteria whereas Betaproteobacteria reappeared. The main determinant of community structure was the redox potential; under iron-reducing conditions in the plume the community was dominated by *Geobacter* species, which are known to be able to degrade organic compounds.

In addition to DGGE, variation in 16S rRNA genes can also be visualized by restriction polymorphisms. *Amplified ribosomal DNA restriction analysis* (ARDRA) applies a restriction to rRNA genes amplified by PCR using universal bacterial primers. Endonuclease digestion of the PCR products will lead to a collection of differently sized DNA fragments and when these fragments are separated by electrophoresis a profile is seen that is characteristic of a community, due to differences between species in the restriction sites. A further development in this direction is *fluorescently tagged ARDRA* (FT-ARDRA), in which the terminal fragment of the restriction is marked by labelling one of the two PCR primers, thereby producing only one band per species. Following separation of the fluorescently labelled fragments in an automatic sequencer with laser detection, a species-composition profile of the community is obtained. FT-ARDRA is a form of *terminal restriction fragment length polymorphism* (T-RFLP); however, this more general term is often used interchangeably with FT-ARDRA.

3.2.2 Microarray-based community surveys

Because gel-based approaches such as DGGE have the disadvantage that only a limited number of species can be resolved, more sophisticated methods using genomics technology have been developed for profiling complex communities. 'A ray of hope' is expected to come from microarray-based hybridization approaches (Polz *et al.* 2003; Zhou 2003; Sessitsch *et al.* 2006; Wagner *et al.* 2007). In studies on plants and animals, microarrays are most commonly used as tools for transcription profiling, however, in microbial ecology microarrays serve to detect or identify organisms and diagnose communities. The principle is that DNAs from a microbial community, or PCR-amplified fragments of environmental DNA, are hybridized to a microarray which contains probes from a series of known species. Those species in the community whose probes are on the microarray will show their presence by hybridization. Unlike in the case of transcription profiling, hybridizations are usually not conducted with two pools of DNA in competition, unless two communities are being compared directly.

The concept of a diagnostic microarray was first suggested by Cho and Tiedje (2001). These authors used four species of *Pseudomonas* as reference for the preparation of probes. The genome of each species was fragmented mechanically and fragments of

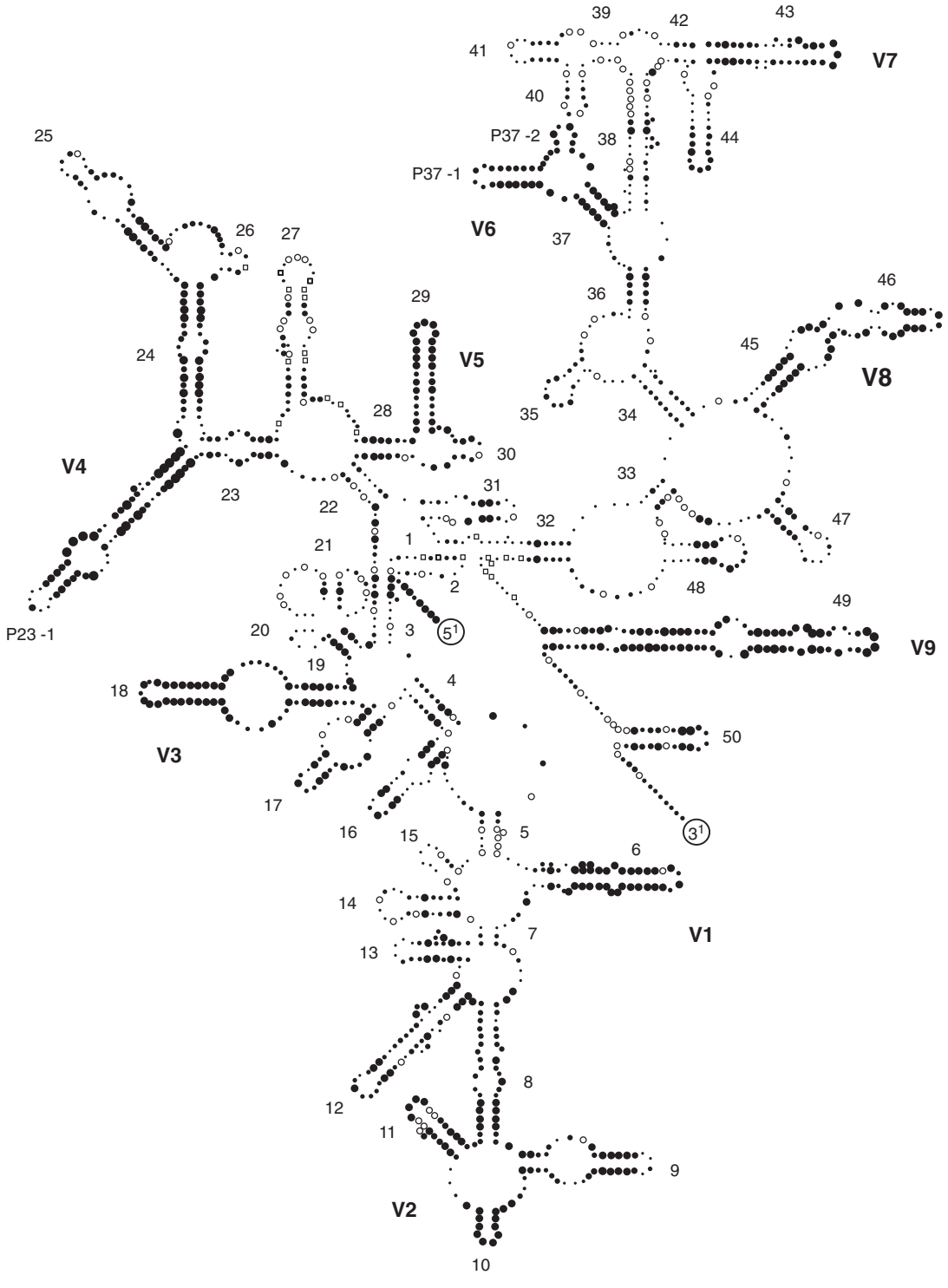


Figure 3.3 Model of a 16S rRNA molecule, showing nine variable regions numbered V1–V9. Each dot represents a nucleotide and its variability across species is indicated by the size of the dot (in five classes). From Neefs *et al.* (1993) with permission from Oxford University Press.

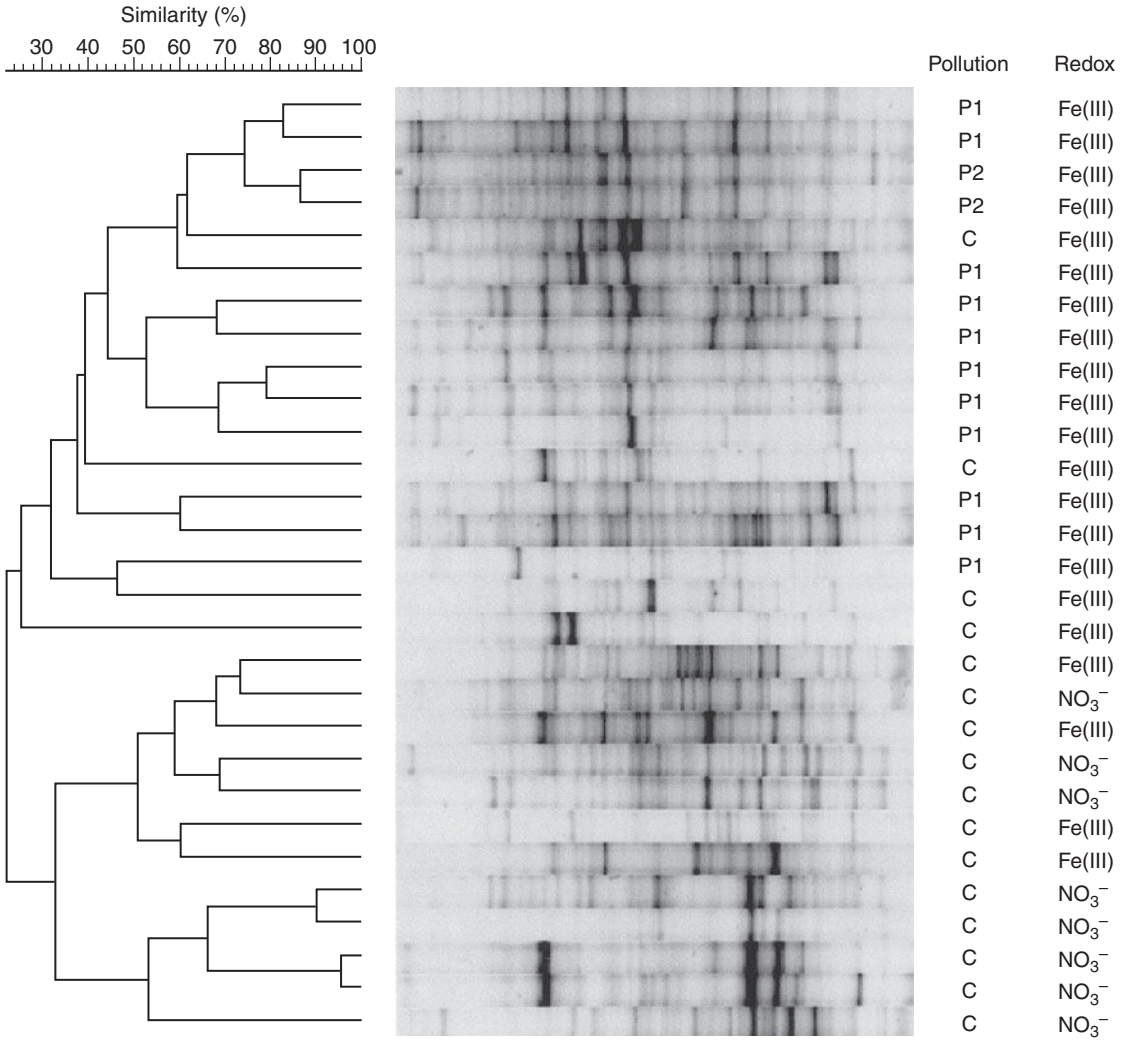


Figure 3.4 DGGE profiles of bacterial communities in an aquifer polluted by landfill leachate. Profiles are shown for samples at different distances and depths; for each lane the level of pollution is indicated (P, in the plume; C, outside the plume, below, above or remote). The right-hand column identifies the dominant redox process of the sample: Fe(III), iron reduction;, denitrification. The profiles were clustered on the basis of similarity measured by Pearson's correlation coefficient, using the unweighted pair-group method using an arithmetic average (UPGMA). There is a clear separation between clean samples and polluted samples, although the latter group comprises three clean samples (which were actually suspected of being influenced by the plume as well). Degradation of contaminants is taking place under iron-reducing conditions. From Röling *et al.* (2001) by permission of the American Society for Microbiology.

1–2 kbp (60–90 per species) were spotted on a coated slide, then hybridized with bacterial cultures. Comparing the hybridization patterns of related species and strains, a distance tree could be developed that was consistent with the phylogenetic tree obtained from 16S rRNA sequences. This provided a proof of principle on the basis of which many new technologies have been developed. Five different types of microarray can be distinguished nowadays, depending on the nature and the source of the

probes used (Zhou 2003; Zhou and Thompson 2004; Gentry *et al.* 2006; see Table 3.2).

One approach is to use the entire genome of a number of target species whose presence is expected (*community genome arrays*). This approach, based on the principle developed by Voordouw *et al.* (1991), but then called *reverse sample genome probing*, does not require prior knowledge of genome sequences, it just spots the whole bacterial genome on the array. This can be done with microorganisms that can be isolated in pure culture; the array then detects the species in a sample from the environment. Another, similar approach is to focus on the open reading frames in single or multiple genomes (*whole genome arrays*).

A third category, *metagenomic arrays* relies on sequence information from environmental DNA. Probes are developed using bioinformatic analysis of a large dataset obtained by means of high-throughput sequencing of environmental samples (see, e.g. Chariton *et al.* 2010). The probes can be designed without prior knowledge of the species, they only derive from the presence of sequences in the environment. This approach has gained popularity recently because sequencing environmental DNA has become much more powerful and cheaper using next generation sequencing technology.

A fourth type of microarray is based on 16S rRNA sequences in databases such as RDP. Because the rRNA gene is diagnostic for prokaryote classification, this type of microarray is called a *phylogenetic microarray* (perhaps taxonomic microarray would be a better qualification, because the information retrieved is essentially taxonomic, not evolutionary; it is the design of the probes that relies on phylogenetic information). This type of microarray quickly gained popularity among microbial ecologists, especially since the introduction of a standardized version of such an array, called '*PhyloChip*' by Gary Anderson and Todd DeSantis (DeSantis *et al.* 2003, 2005, 2007).

Finally, arrays have been developed on the basis of functional genes, often genes related to nutrient cycles such as nitrification or sulphate reduction. These arrays are called *functional gene arrays.* Depending on which parts of the gene are used as probes, functional gene arrays may detect the same gene in a wide variety of organisms, so the information retrieved is not of a taxonomic nature, unless species-specific parts of the gene are targeted. The most popular version of a functional gene array is the '*GeoChip*' which was introduced by He *et al.* (2007).

Using microarrays for detection, complex communities can be very rapidly profiled without the

Table 3.2 Overview of five types of microarray approach used in microbial ecology to detect microorganisms in the environment

	Phylogenetic arrays (phylochips)	**Functional gene arrays (geochips)**	**Metagenomic arrays**	**Community genome arrays**	**Whole genome arrays**
Source of probes	Selected regions of 16S rRNA genes	Genes with specific function in an ecological process	Environmental DNA	Entire genomic DNA of species	ORFs in whole genome
Type of information provided	Taxonomy, species identification	Potential functions	Potential environmental functions	Community composition	Phylogenetic and functional
Specificity	Species, strains, mutants	All species within 80–90% sequence homology	Species	Species to strains	Strains, SNPs
Sensitivity (ng of DNA detectable)	500	1–8	Not determined	0.2	0.1

Source: Adapted from Zhou (2003), Gentry *et al.* (2006), and Sessitsch *et al.* (2006)

need for culturing. In principle, the technique can even be implemented in a portable system, allowing species identification in the field within 50 min (Bavykin *et al.* 2001). However, we need to realize that, unlike gel-based methods such as DGGE, microarray-based profiling can only be conducted with species for which genomic information is already available, because this information is required to develop the probes on the array. This implies that the only species that are detected are those that we already have knowledge of—although sometimes that knowledge is limited to the probe sequence alone.

A strong argument in favour of the use of microarrays in microbial detection is that it can, at least in principle, replace PCR amplification of target sequences. PCRs applied to a heterogeneous collection of DNAs from the environment often suffer from preferential or aspecific amplification of certain sequences over others (*PCR bias*), especially when the PCR is addressed to a specific, limited group of organisms. This is due to the initially small number of template copies that must be found in a large background of DNA to be ignored. Biodiversity screens in which a PCR step is included are vulnerable to the criticism that the diversity in the amplified assemblage may be different from the diversity of the environment. With microarrays the PCR step can (in principle) be skipped altogether, as demonstrated by some authors (Small *et al.* 2001; Urakawa *et al.* 2002; El Fantroussi *et al.* 2003). In this strategy RNA extracted from environmental samples is fragmented, labelled, and hybridized directly with the microarray. A multiple melting-curve analysis is then performed to detect aspecific hybridizations. Although this strategy seems to be particularly applicable in an environmental context, most authors still prefer to start with a PCR to obtain sufficient target DNA for hybridization (e.g. Loy *et al.* 2002; Peplies *et al.* 2003; Taroncher-Oldenburg *et al.* 2003). Such a PCR uses universal bacterial primers (targeting a phylogenetic marker gene in all bacteria) to minimize possible bias, after which specific groups are picked out by microarray hybridization. For functional gene arrays, prior gene-specific PCR amplification is unpractical because of the large variety of genes addressed by the array. Sometimes *rolling circle amplification* is applied, a unidirectional rapid amplification of circular genomes. This technique is especially useful when direct hybridization of environmental DNA to a microarray is ineffective due to microbial communities with low biomass.

DeSantis *et al.* (2005) provided proof of principle for the use of microarrays for diagnosis of environmental DNA. The authors used a background of environmental DNA originating from outdoor air samples and spiked those samples with known quantities of rRNA genes from five different species. There was a strong linear relationship between the intensity signal from the microarray and the known amount of DNA spiked, over a wide range of concentrations (Fig. 3.5). This work showed that it is in principle possible to use microarrays for quantitative detection of microorganisms in the environment.

Despite the general enthusiasm for the use of microarrays in microbial ecology, several challenges remain. These have been listed by Wagner *et al.* (2007) and can be summarized as follows:

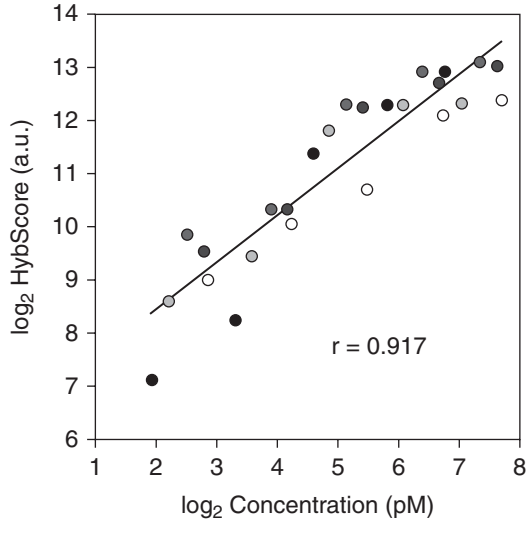


Figure 3.5 Double logarithmic plot of PhyloChip-based hybridization intensities versus spiked-in rRNA gene copies, on a background of environmental DNA. The differently filled circles refer to different microbial species. Reproduced from DeSantis *et al.* (2005), by permission of the Federation of European Microbial Societies.

- *Specificity*. Ideally, every probe on the array will hybridize only to a fully matched sequence in the environment, however, this has been difficult to achieve.
- *Sensitivity*. Current microarrays can detect the 20 to 2000 most abundant target populations in the environment, however, this may not be sufficient if crucial functions are conducted by species with low abundance.
- *Quantification*. As shown in Fig. 3.5, some authors have obtained very good linear relationships between microarray hybridization and the amount of DNA, however, other steps in the procedure (e.g. DNA extraction, PCR) may limit the potential for quantification.
- *Identification*. Microarrays obviously can only identify the organisms about which DNA information was already available. By expanding the databases used for probe design this problem will gradually diminish, but given the enormous breadth of microbial diversity, it may remain a problem for a long time.
- *Reliability*. More replication and confirmation with other techniques is necessary to enhance the reliability of microarray results.

How is a microarray designed from sequence information? Different taxa in the hierarchical classification system of microorganisms are addressed. For example, in a microarray system targeting *E. coli* some probes will be universal for all bacteria, others are specific for the phylum Proteobacteria, and still others for the γ subdivision of Proteobacteria; then some probes target only the group of enteric bacteria and the most specific probes address specific genera of Enterobacteria, such as *Escherichia* or *Salmonella*. So DNA from *E. coli* should hybridize with all these probes, and DNA from *Nitrosomonas*, a betaproteobacterium, should hybridize only with the first two. From the hybridization pattern of a complex community an indication can be obtained of the taxonomic composition; however, not all groups of bacteria can be resolved in this way. In most cases the discriminatory power is limited to genera; in several cases, when genera are very similar in their 16S rRNA gene, only the higher taxonomic levels are resolved. The taxa (be it genera, families, or divisions) that can be detected in this way are designated as *operational taxonomic units* (OTUs).

The strategy of hierarchical arrangement of SSU RNA probes was applied by Wilson *et al.* (2002) to develop a photolithography gene chip with 31 179 oligonucleotide probes. As in the case of gene chips used for transcription profiling, each 20-mer probe was paired with a control probe in which the eleventh nucleotide was replaced by a mismatch position; these control probes should not hybridize with the target and so serve to control for nonspecific effects. The number of probe pairs targeting a specific taxon ranged between 1 and 70, depending on the completeness of sequence information of the rRNA region. In total there were 1945 prokaryotic and 431 eukaryotic sequences represented on the chip. By way of a test, the chip was used to detect bacteria and fungi recovered from filtered air samples. The results showed that the airborne microbial assembly was dominated by bacteria from the Gram-positive *Bacillus-Lactobacillus-Streptococcus* subdivision and the Gram-positive high-G+C group; in addition significant amounts of ascomycete and basidiomycete DNA (presumably spores) were found in the 'aeroplankton'. There was a good correspondence between chip-based detection and PCR amplification of 16S RNA genes followed by cloning and sequencing; however, cloning revealed 28 novel sequences that did not correspond well to any phylogenetic group in the RDP database and so were not detected by the microarray.

The principle of microbial detection in air was further developed with a gene chip of 500 000 probes allowing detection of around 9000 OTUs (Andersen *et al.* 2004; Brodie *et al.* 2007). A sampling campaign was conducted with this array to survey the air of two cities in the USA, Austin and San Antonio, over several months. The use of microarrays to gauge air biota had a mainly medical relevance, and was specifically motivated in the USA by the need for bioterrorism defence strategies against possible mischievous release of pathogenic agents. In the monitoring campaign, significant seasonal trends were observed; the abundance of bacteria was correlated negatively with air temperature and positively with the dew point. An interesting observation was that, although the time trends were

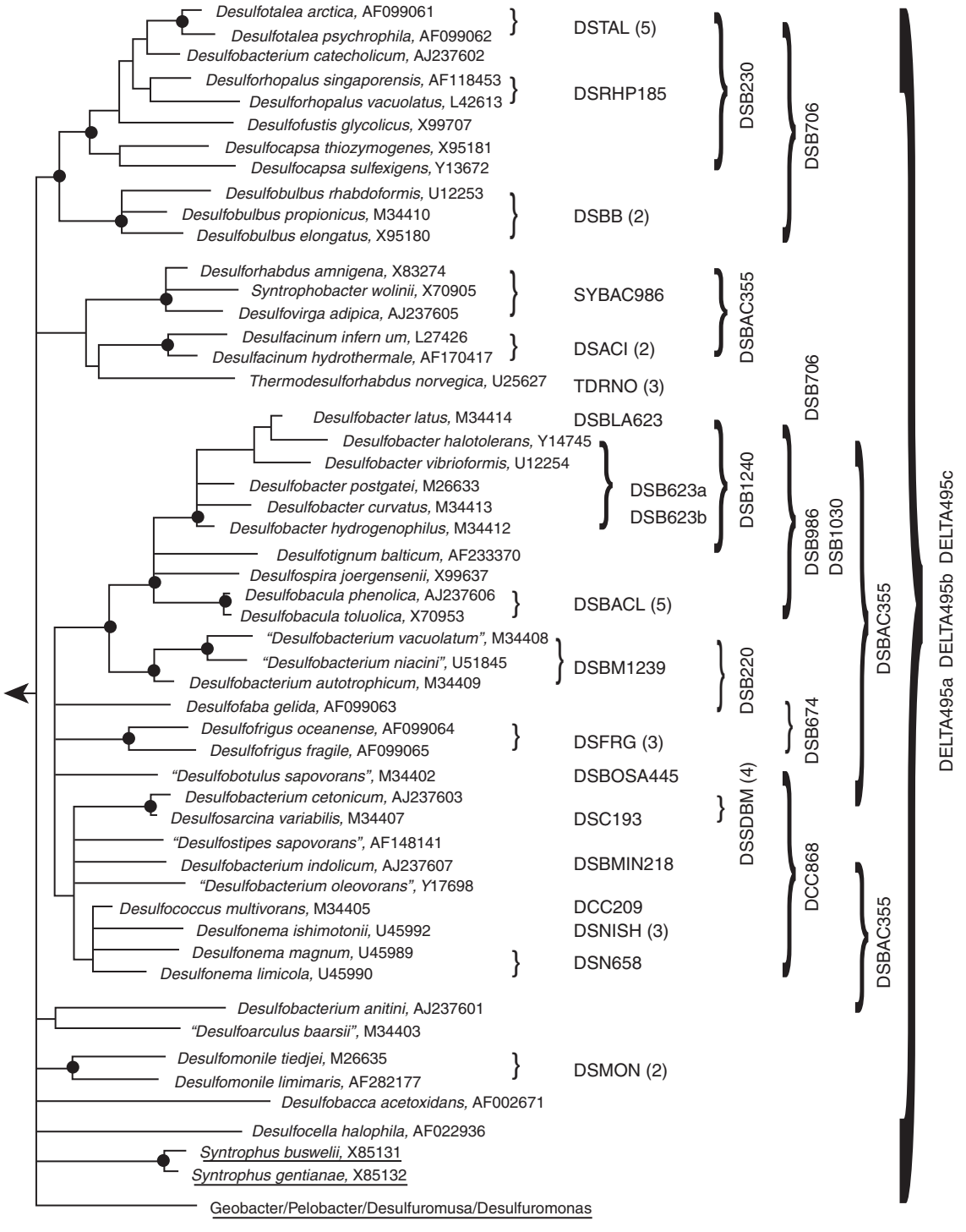
similar for the two cities, the compositions of the microbial assemblies were different, indicating that in addition to climatic determinants there are local sources of airborne microorganisms. Low levels of *Bacillus anthracis* (causing anthrax) and *Clostridium botulinum* (causing botulism) were detected, because these bacteria are normally found associated with livestock and occur naturally in sediments. Inevitably, some of these bacteria are aerosolized and become airborne with the wind. It is therefore necessary to know the background levels of these pathogens so as to single out situations hazardous to human health.

Similar detection strategies applied in ecological settings, for example using DNA from soils, sediments, or water meets more difficulties than air sampling, because the chemical matrix in which the environmental DNA is contained is much more complex. In particular, humic derivatives are notorious for causing disruption to DNA samples. Nevertheless, several successful attempts have been made. Loy *et al.* (2002) developed an oligonucleotide microarray called '*SRP-PhyloChip*' for cultivation-independent detection of sulphate-reducing prokaryotes (SRPs). Sulphate reducers form a highly heterogeneous, polyphyletic, group of bacteria. The capacity to gain energy from sulphate reduction has evolved independently several times in different lineages. In some cases species that can and cannot perform sulphate reduction are taxonomically closely related. This situation makes it impossible to target sulphate reducers on the basis of a PCR with a single 16S rRNA sequence; instead, several different PCRs would be needed. For microarrays, which can deal with many probes in parallel, the situation presents no problem. Sulphate reducers (134 recognized species in total) are present in seven different lineages of prokaryotes: Deltaproteobacteria, Nitrospirae, Clostridia, Thermodesulfobiaceae, Thermodesulfobacteria (all Bacteria), Euryarchaeota, and Crenachaeota (Archaea) (Muyzer & Stams 2008). Figure 3.6 provides an example of the affiliations in two orders of the Deltaproteobacteria.

The alignments of 16S RNA sequences led to the development of 138 18-mer probes, which in conjunction are diagnostic for the community of sulphate reducers. For example, the genus *Desulfotalea* was specifically detected by five probes and was also targeted by three probes with broader specificities (Fig. 3.6). In some cases, however, the taxonomic resolution did not go further than a group of three genera (see SYBAC986 in Fig. 3.6). If the array is hybridized with a sample of PCR-amplified 16S rRNA genes, a read-out of hybridizations will produce a list of taxa present in the sample. The SRP-PhyloChip was used to investigate the sulphate-reducing community of a hypersaline cyanobacterial mat from Solar Lake, Egypt. Bacteria from the genera *Desulfonema* and *Desulfomonile* were identified and this was confirmed by cloning and sequencing.

Another genomics survey of bacterial community composition was applied by Valinsky *et al.* (2002). These authors were interested in the *disease-suppressiveness* of agricultural soils. Agricultural experience shows that soils can become suppressive to soil-borne pathogens if they are managed in certain ways. One of the best-known examples is suppression of the fungus *Gaeumannomyces graminis* var. *tritici* (Ascomycota), which causes a root disease of wheat worldwide (the syndrome is described as 'take-all'). When wheat is cropped in continued monoculture, one or more severe outbreaks of the disease usually follow, but thereafter the soil spontaneously develops a suppressiveness and the fungus is unable to grow any more in such soils (take-all decline). This is correlated with an increase in the

Figure 3.6 Showing a phylogeny of sulphate-reducing bacteria in the orders Desulfobacterales and Syntrophobacterales of the Deltaproteobacteria. The tree was developed on the basis of 16S rRNA sequences using maximum parsimony and other methods. Non-sulphate reducers are underlined. The probes used on the array are indicated by short names such as DSTAL, DSRHP185, etc. Braces indicate the taxonomic span of the probes, e.g. DELTA495a, DELTA495b, and DELTA495c target all Deltaproteobacteria, DSB706 targets the upper clade of 11 species plus *T. norvegica*, DSB230 targets a group of eight species within the upper clade, and DSRHP185 targets two species of the genus *Desulforhopalus*. Numbers in parentheses next to a probe name indicate that more than one probe was used to target different parts of the rRNA molecule, e.g. five different DSTAL probes target the genus *Desulfotalea*. The codes after the species represent GenBank accession numbers of their 16S rRNA sequences. Adapted from Loy *et al.* (2002), by permission of the American Society for Microbiology.


Desulfotalea arctica, AF099061
Desulfotalea psychrophila, AF099062
Desulfobacterium catecholicum, AJ237602
DSTAL (5)
Desulforhopalus singaporensis, AF118453
Desulforhopalus vacuolatus, L42613
DSRHP185
Desulfofustis glycolicus, X99707
Desulfocapsa thiozymogenes, X95181
Desulfocapsa sulfexigens, Y13672
Desulfobulbus rhabdoformis, U12253
Desulfobulbus propionicus, M34410
Desulfobulbus elongatus, X95180
DSBB (2)
DSB230
DSB706
Desulforhabdus amnigena, X83274
Syntrophobacter wolinii, X70905
Desulfovirga adipica, AJ237605
SYBAC986
Desulfacinum infern um, L27426
Desulfacinum hydrothermale, AF170417
DSACI (2)
DSBAC355
Thermodesulforhabdus norvegica, U25627
TDRNO (3)
Desulfobacter latus, M34414
DSBLA623
Desulfobacter halotolerans, Y14745
Desulfobacter vibrioformis, U12254
Desulfobacter postgatei, M26633
Desulfobacter curvatus, M34413
Desulfobacter hydrogenophilus, M34412
DSB623a
DSB623b
DSB1240
DSB706
Desulfotignum balticum, AF233370
Desulfospira joergensenii, X99637
Desulfobacula phenolica, AJ237606
Desulfobacula toluolica, X70953
DSBACL (5)
DSB986
DSB1030
"Desulfobacterium vacuolatum", M34408
"Desulfobacterium niacini", U51845
Desulfobacterium autotrophicum, M34409
DSBM1239
DSB220
Desulfofaba gelida, AF099063
DSB674
DSBAC355
Desulfofrigus oceanense, AF099064
Desulfofrigus fragile, AF099065
DSFRG (3)
"Desulfobotulus sapovorans", M34402
DSBOSA445
Desulfobacterium cetonicum, AJ237603
Desulfosarcina variabilis, M34407
DSC193
DSSDBM (4)
"Desulfostipes sapovorans", AF148141
Desulfobacterium indolicum, AJ237607
DSBMIN218
"Desulfobacterium oleovorans", Y17698
DCC868
Desulfococcus multivorans, M34405
DCC209
Desulfonema ishimotonii, U45992
DSNISH (3)
Desulfonema magnum, U45989
Desulfonema limicola, U45990
DSN658
DSBAC355
Desulfobacterium anitini, AJ237601
"Desulfoarculus baarsii", M34403
Desulfomonile tiedjei, M26635
Desulfomonile limimaris, AF282177
DSMON (2)
Desulfobacca acetoxidans, AF002671
Desulfocella halophila, AF022936
Syntrophus buswelii, X85131
Syntrophus gentianae, X85132
Geobacter/Pelobacter/Desulfuromusa/Desulfuromonas
DELTA495a DELTA495b DELTA495c

abundance of bacteria of the genus *Pseudomonas*, some of which are known to produce an antifungal compound, 2,4-diacetylphloroglucinol. This well-known example of soils developing suppressiveness is a model for research at the interface between plant pathology and soil microbiology (Weller *et al.* 2002; De Souza *et al.* 2003; Garbeva *et al.* 2004).

Valinsky *et al.* (2002) investigated a case of soil suppressiveness towards the plant-parasitic nematode *Heterodera schachtii* (sugar beet cyst nematode). Two adjacent agricultural fields, one suppressive and the other not, were sampled and microbial communities were screened using an array-based method. The authors used a membrane to which the soil-extracted and PCR-amplified 16S rRNA genes were fixed, while different radioactively labelled oligonucleotide probes were added to detect a 16S rRNA sequence of a specific species or group of species. A great number of bacterial clusters were identified, which were grouped into five major taxa (Table 3.3). There were obvious differences between the soils; the disease-suppressive soil had fewer *Bacillus* species, more Alphaproteobacteria, and fewer Enterobacteria. Interestingly, DGGE analysis of the same soils revealed only 13 bands and did not detect any bands differential between the two soils. This illustrates the enormous increase in resolution that can be obtained by genome-wide analyses. Further research is necessary to reveal the mechanistic relationships between community composition and suppressiveness to cyst nematodes in these soils.

Table 3.3 Taxonomic composition of bacterial rRNA clones obtained from two agricultural soils using an array-based screening method

	Number of clones Reference soil	Soil suppressive to sugar beet cyst nematode
Bacillus	405	35
Cytophaga–Flexibacter–Bacteroides group	5	25
Actinobacteria	130	185
Alphaproteobacteria	10	142
Beta- and gammaproteobacteria	162	87
Enterobacteria	127	8

Source: From Valinsky *et al.* (2002).

3.2.3 Statistical approaches to prokaryote diversity

The recurrent question of how many species of prokaryote there are has also been addressed using statistical approaches applied to genome data (Hughes *et al.* 2001; Curtis *et al.* 2002; Ward 2002; Ward *et al.* 2008). When a community is sampled and the number of species is counted in successive samples, the total number of species retrieved will show a curve of diminishing returns. A plot of the cumulative number of observed species versus sample size produces a *species-accumulation curve*, also called collector's curve, which tends to level off to a plateau beyond a certain sample size (Krebs 1999). Assuming a suitable model (e.g. a hyperbola or a negative exponential) the total number of species in the community may be estimated by extrapolation, even though a fraction of species remains unnoticed. Such methods are commonly applied to field inventories of plants and insects, but they are rarely found in microbial ecology.

Hughes *et al.* (2001) reviewed various microbial datasets and showed how the ecological methodology can be applied to estimate species richness in a microbial setting. One of the datasets analysed was from McCaig *et al.* (1999). These authors had sequenced 16S rRNA genes from clones isolated from soils of two upland pastures in Scotland, one natural but grazed by sheep in summer, the other reseeded, fertilized, and grazed during the whole season ('improved'). Applying the 97% similarity criterion as a species boundary, there were no differences in observed species richness between the two sites (113 in the fertilized habitat and 114 in the natural grassland); however, there were differences in species composition. Hughes *et al.* (2001) reanalysed the data and applied Chao's estimator for total species richness, S_T, which is

$$S_T = S_{obs} + \frac{n_1^2}{2n_2}$$

where S_{obs} is the observed number of species, n_1 is the number of singletons (species captured once), and n_2 is the number of doubletons (species captured twice). The theory also accounts for a standard error of this estimate. Applying this equation to the Scottish grassland data produced the species-accumulation curve shown in Fig. 3.7. Total species richness S_T was estimated as 590 OTUs in the natural grassland and 467 OTUs in the improved grassland. Due to the significant number of rare species (singletons relative to doubletons), this estimated total number of species in the community was considerably higher than the actual number observed. There seemed to be more species in the natural grassland soil than in the fertilized soil, although the difference was not statistically significant.

Another approach to estimation of microbial diversity was applied by Curtis *et al.* (2002). These authors used the distribution of species over-abundance classes as a basis for extrapolation. We know from basic ecology that in any community there are a few species with very high abundance, whereas most species have intermediate abundance. The distribution of species over abundances can be characterized by a function *f*, where

$$S_T = \int_{N_{min}}^{N_{max}} f(N)\,dN$$

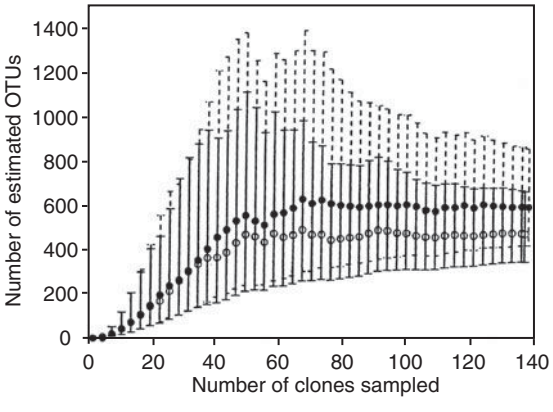


Figure 3.7 Estimates of species richness according to Chao applied to 16S rRNA sequences differing more than 3% from each other (OTUs) in two upland pastures in Scotland, one natural (•), the other fertilized (o), with 95% confidence intervals (dashed lines for the natural grassland, solid lines for the fertilized habitat). From Hughes *et al.* (2001), by permission from the American Society for Microbiology.

in which S_T is the total number of species in the community, N_{min} the abundance of the least abundant species, and N_{max} the abundance of the most abundant species. Integrating *f* over all abundances is equivalent to cumulating species over all abundance classes, from rare to dominant. Community ecologists have argued that *f* often takes a bell-shaped symmetrical form if abundance is grouped into geometric-scale units, for example in classes whose widths increase by a factor of 2 with every successive group (octave scale; Krebs 1999). The function *f* then approaches a normal curve on a logarithmic scale. *Lognormal species-abundance curves* are assumed to hold particularly well for microorganisms, which exhibit highly dynamic and random growth, influenced by many independent factors. The lognormal distribution is defined by only two parameters, allowing S_T to be estimated from the number of species in one of the classes (e.g. the largest class) plus the spread of the distribution (Krebs 1999). However, these two quantities are still difficult to measure in the case of microorganisms, therefore Curtis *et al.* (2002) developed an expression allowing S_T to be estimated from two quantities easily accessible to measurement: the total number of individuals in the community (N_T) and the abundance of the most abundant member (N_{max}). Applying the formula to data for microbial clone abundance reported in the literature, the authors were able to estimate prokaryote species diversity for several ecosystems. In addition, they also estimated the global diversity of Bacteria and Archaea in the sea (Table 3.4). A similar approach, but using model-free approximation of the species-abundance, was applied by Gans *et al.* (2005). These authors estimated the number of bacterial species per 10 g of soil as 8×10^6, exceeding previous estimates by one or two orders of magnitude.

The analyses of Curtis *et al.* (2002) and Gans *et al.* (2005) depend crucially on abundance data, which, even for the most abundant organisms—prokaryotes—are not very reliable. Nevertheless, these estimates show that prokaryote species richness may be orders of magnitude greater than can ever be achieved with clone libraries; even the most extensive clone libraries do not reach beyond a few hundred species. When better quantitative data become

available in the future the estimates can be refined; however, to develop a reliable empirical species-abundance curve for a microbial community will be a serious experimental challenge. For the moment the theoretical predictions seem to stand on firmer ground. There are also some intriguing questions inherent in the estimates in Table 3.4; for example, why the global species richness of the Archaea is so much lower than that of the Bacteria and why the biodiversity of soil ecosystems increases less with increasing scale than the biodiversity in the sea.

Estimates of species richness can also be derived from the sequence diversity observed in environmental DNA. This approach was applied in a large-scale sequencing campaign examining marine biodiversity (Venter *et al.* 2004). The Sorcerer II expedition is an undertaking of the Venter Institute aimed at analysing the vast untapped and unseen world of oceanic biota around the world (www.sorcerer2expedition.org). In the first phase of the project, samples were collected from the Sargasso Sea, off the coast of Bermuda, and that DNA was cloned into plasmid vectors and sequenced to capture as many organisms as possible (see also Section 3.4 below). Obviously, when sampling the community at random, the most abundant species will be sequenced many times and the rare species only occasionally. Using an adapted assembly algorithm with enormous computational power, scaffolds were constructed that could be assigned to species. A 'genomic species' was defined as a clustering of assemblies or unassembled reads with more than 94% sequence similarity.

Table 3.4 Estimates of local and global biodiversity of prokaryotes, derived from abundance data assuming a lognormal distribution of species over abundance classes

Environmental compartment	Estimated number of species
Bacteria in1 ml of ocean water	163
Bacteria in the entire sea (pelagic)	2 000 000
Archaea in the entire sea (pelagic)	20 000
Bacteria in an average lake	8000
Bacteria in 1 g of soil	6300–38 000
Bacteria in 100 g of soil	100 000–1 000 000
Bacteria in 1000 kg of soil	4 000 000
All soil bacteria in the world	4 500 000
Bacteria in the global atmosphere	4 000 000
Bacteria in 1 ml of sewage sludge	70
All bacteria in a sewage works	500

Source: From Curtis *et al.* (2002).

The method for estimating species-richness from sequence data is based on the number of times that sequences are observed from the same genome. Obviously, the most abundant genome will be sequenced many times and the rare genomes only occasionally. By analysing how the sequencing coverage decreases with decreasing abundance, an estimate of the number of genomes can be obtained. Let $P(r)$ be the probability that any particular base position in a collection of DNAs is sequenced r times. $P(r)$ can be modelled as a Poisson distribution, for which the mean equals the average depth of coverage. Similarly, for a mixture of rare and abundant genomes, different species have different depths of coverage and $P(r)$ is a composite Poisson distribution. By fitting the theoretical distribution to the data, the number of genomes at each depth of coverage can be estimated (Table 3.5), and the sum of these numbers is an estimate of the number of species in the community. For a 200 l sample from the Sargasso Sea, species richness was estimated as at least 1800, which is consistent with the extrapolations from the model of Curtis *et al.* (2002) discussed above.

Inherent in these extrapolation methods is that the total estimated number of species (genomes) is above, sometimes far above, the actual total number of phylotypes observed. By reversing the argument one may therefore also estimate by what amount the sampling effort has to increase to cover the larger part of the community. This was done in a statistical analysis of metagenomic sequencing data by Quince *et al.* (2008). They came to the astonishing conclusion that in species-rich habitats such as soils and deep-ocean hydrothermal vents, hundreds of times the current number of samples will be required to obtain just 90% of the taxonomic diversity based on 16S rRNA genes. For a complete survey of the whole metagenome tens of thousands times the current sequencing effort is necessary. Models like this can form the basis for a rational planning of biodiversity surveys.

Our *tour d'horizon* of prokaryote biodiversity screening has demonstrated an inherent weakness of microbial ecology compared to plant and animal ecology: the difficulty of obtaining reliable estimates for abundance and species richness. It is expected that genomics approaches will improve on this situation in the near future. Still, sequences cloned from environmental samples are rarely identical to sequences of cultured bacteria in gene databases. Several investigations report on completely new bacterial lineages. For example, only a few years ago it was realized that the phylum Acidobacteria, previously known from only three described species, is as diverse as the whole superphylum of Proteobacteria. DNA sequences of this group are found everywhere in clone libraries; often no less than 30% or even 50% of the 16S rRNA gene inserts in clone libraries from soil sample belong to Acidobacteria, which is indicative of a very important (but unknown) ecological role (Quaiser *et al.* 2003). Obviously, a levelling-off in the collector's curve of global microorganisms is not even in sight.

Table 3.5 Illustrating a model to estimate genome (species) diversity in microplankton communities analysed by sequencing

Sequencing coverage depth	Fraction of assembly consensus organisms	Expected fraction of genome sequenced	Genomes
25	0.0055	1.0	2.5
21	0.0050	1.0	2.3
13	0.0035	1.0	1.6
9	0.0040	1.0	1.8
7	0.0080	0.999	3.6
6	0.0047	0.998	2.1
4	0.0100	0.982	4.6
2.4	0.0258	0.909	12.8
2	0.0700	0.865	36.4
0.25	0.8635	0.221	1756.7
Total	1.0		1824.4

Notes: Assembly consensus sequences are classified by the sequencing depth coverage. For example, 0.55% of the organisms have been sequenced 25 times and such sequencing covered the whole genome. This results in an estimated number of 2.5 genomes falling into this class, assuming a mixture of Poisson distributions. The total number of species estimated for a 200-l sample from the Sargasso Sea is 1824. Reprinted with permission from Venter *et al*. (2004). Copyright 2004 AAAS.

3.3 Microbial genomics of biogeochemical cycles

The crucial role of microorganisms in biogeochemical cycles has been known for a long time, but only recently have we gained insight into the architecture and diversity of the genetic systems responsible for these ecological functions. In addition, new and unexpected links between element cycles have been discovered that were not taken into account before. The new evidence is coming from two sources: detection of functional genes using microarrays and cloning of environmental genomes. The second topic will be discussed in Section 3.4, and this section is devoted to an overview of the genes and genomic background associated with conversions in the major element cycles, as well as examples of studies using microbial functional genomics in the field.

All elements in the Earth's crust and the atmosphere interact with the biosphere and for some elements, especially those that are used as essential building blocks or catalysts in cells, the interaction is particularly strong. Many elements undergo transitions from one chemical compound to another, from one redox state to another, or from one environmental compartment to another. Many of these transitions are catalysed by enzymes in specific biological entities. To understand and predict biogeochemical cycles it is necessary to know the mechanisms by which organisms direct the transitions in these cycles. In this section we will therefore focus on the genomic determinants of such transitions and conversions. Knowledge of the responsible genes and the organisms in which these are expressed can be used to develop screening instruments by which biogeochemical functions in environmental samples may be assessed. Table 3.6 is a guide for the discussion. We focus on key links in the carbon, nitrogen, sulphur, phosphorus, iron, calcium, and silicium cycles. We do not discuss all the biochemical aspects of nutrient cycles; for a full treatment of this the reader is referred to a microbiology textbook such as Madigan *et al.* (2002).

3.3.1 Key genes in the carbon cycle

Starting with carbon, we note that photosynthesis is the most important link in the cycling of nutrients through the biosphere, because it is the process upon which almost all other life depends. Most organisms that can use sunlight for the generation of metabolic energy (phototrophs) divert that energy to the synthesis of organic molecules from CO_2 and so are *photoautrophic*. There are also phototrophs that rely on organic carbon for their growth, although they use light as a source of energy; these are called *photoheterotrophs*. Another distinction relates to the source of reducing power for the conversion of CO_2 to organic carbon. This can come from water, producing O_2 as a byproduct, from reduced sulphur compounds in the environment such as H_2S, or from molecular hydrogen, H_2. The first type of photosynthesis, which is practised by cyanobacteria, protists, green algae, and land plants is called *oxygenic* (producing oxygen), whereas the second type, practised by purple bacteria and other prokaryotes, is called *anoxygenic*. The use of H_2S or H_2 as a source of reduction equivalents is not always driven by light; it can also be coupled to chemoautotrophic energy generation, although the use of water and the production of O_2 is always driven by light.

Interception of light by photoautotrophs relies on chlorophyll, a porphyrin molecule with a magnesium atom in the centre of the ring structure. Several chlorophylls are known, which differ from each other in substituents on the porphyrin ring and side chains attached to it. These substituents influence the spectroscopic properties of the Mg-porphyrin, and so the different chlorophylls have different absorption maxima. Cyanobacteria, green algae, and land plants have chlorophyll *a* and *b*, several algal groups (Phaeophyta, Chrysophyta, Dinoflagellata) have chlorophyll *a* and *c*, and red algae have a type of chlorophyll *a* called bacteriochlorophyll *a*. In addition, most phototrophic organisms have various accessory pigments that support the capture of light (phycobilins) or protect against damage from radicals (carotenoids).

The main genomic models for investigating bacterial photosynthesis are purple bacteria of the genera *Rhodobacter* (Alphaproteobacteria, Rhodobacterales) and *Rhodopseudomonas* (Alphaproteobacteria, Rhizobiales). These bacteria are anoxygenic phototrophs, which generate sulphur from H_2S rather than oxygen from water. *Rhodobacter sphaeroides* is readily amenable to genetic manipulation, allowing the development of mutants that are deficient in some aspect of the process, and this has greatly helped in reconstructing the genomic organization of the photosynthetic apparatus. In *Rhodobacter*, the genes involved in photosynthesis are clustered in operons in a 50 kb region of the chromosome called the *photosynthetic cluster*. The operons are transcribed in a highly coordinated manner to allow the correct assembly of new photosynthetic complexes. The cluster involves *bch* genes, which encode proteins involved in the synthesis of bacteriochlorophyll; *crt* genes, which encode proteins involved with carotenoid synthesis; *puh* and *puf* genes, which encode pigment-binding peptides of the reaction centre; and *puc* genes, encoding assembly factors. These genes may be used to develop probes for functional gene microarrays (Table 3.6).

An important model for oxygenic photosynthesis in bacteria is *Synechocystis*, a cyanobacterium from the order Chroococcales, the fourth microorganism and the first photosynthetically active species to have its genome sequenced completely (Kaneko *et al.* 1996). We know from plant physiology that cyanobacteria, green algae, and plants have a more complicated system for the flow of electrons than the anoxygenic phototrophs; namely, they use two interconnected photosystems. In the first step, conducted by *photosystem II*, water is split into oxygen and hydrogen, whereas the second step, conducted by *photosystem I*, produces NADPH. The double system is also called *noncyclic photophosphorylation* because the electrons liberated from chlorophyll by light excitation are transferred to $NADP^+$, rather than back to chlorophyll as in anoxygenic or cyclical photophosphorylation. The evolution of photosystem II in Cyanobacteria around 3.5 billion years ago changed the face of the Earth because it allowed the use of water as an electron donor and the production of oxygen, which came to accumulate in the atmosphere. The subunits of the two photosystems are encoded by genes of the *psb* cluster (for photosystem I) and *psa* (for photosystem II). Other genes

Table 3.6 Overview of genes associated with key links in biogeochemical cycles, classified by element and process

Element	General process	Specific process	Gene products, enzymes	Indicative genes or gene clusters
Carbon	Anoxygenic photosynthesis	Photopigment biosynthesis	Enzymes of bacteriochlorophyll and carotenoid biosynthesis	*bch, crt*
		Assembly of photosynthetic complex	Light-harvesting and reaction-centre complex subunits and assembly factors	*pufLM, puhA, pucBA*
	Oxygenic photosynthesis	Photopigment biosynthesis	Enzymes of chlorophyll, phycobilin, and carotenoid pathways	*apc, cpc*
		Assembly of photosynthetic complex	Photosystem I and II subunits	*psa, psb*
	Carbon fixation	Calvin cycle	Ribulose bisphosphate carboxylase complex (Rubisco) subunits, phosphoribulokinase	*rbc, cbb, prk*
		Reverse citric acid cycle	ATP citrate lyase	*acl*
	Non-chlorophyll phototrophy	Light-driven energy generation	Proteorhodopsin	*gpr*
	Decomposition	Polysaccharide catabolism	Amylase, cellulase, pectinase, chitinase	*amy, cel, pem pel, chi,*
		Lipid catabolism	Phospholipase, acyl-CoA dehydrogenase	*phlHGFACBDE*
		Lignin degradation	Peroxidase	*mnp*
	Methanotrophy	Methane oxidation	Methane monooxygenase	*pmoCAB*
	Methanogenesis		Methylreductase, heterodisulphide reductase	*mrc, hdr*
Nitrogen	Nitrogen fixation	Nitrogenase	Dinitrogenase α and β, dinitrogenase reductase, electron-transport flavoproteins	*nifKDH, fixABCX*
	Denitrification		Nitrate reductase, nitrite reductase, NO reductase, N_2O reductase	*napA, nrfA, narG, nirK, nirS, norB, nosZ*
	Anammox	Anaerobic ammonia oxidation	Hydrazine synthase	*hzsA, hzsB, hzsC*
	Assimilatory nitrate reduction		Nitrate reductase, nitrite reductase	*nasA, nirA, nirB*
	Nitrification	Aerobic ammonia oxidation	Ammonia monooxygenase, hydroxylamine oxidoreductase	*amoA, hao*
		Nitrite oxidation	Nitrite oxidoreductase	*norB*
	Ammonification		Glutamate dehydrogenase, urease	*gdh, ureC*
Sulphur	Sulphate reduction		Adenosine phosphosulphate reductase, sulphite reductase	*apsA, dsrAB, aprBA*
		Dimethyl sulphide generation	Cleavage of dimethylsulphoniopropionate	*dddD*
	Sulphur oxidation		Sulphite oxidase, sulphur dehydrogenase	*soxB*
Phosphorus	Polyphosphate hydrolysis		Exopolyphosphatase, polyphosphate kinase	*ppx, ppk*
Iron	Iron reduction	Periplasmic electron shuttling, iron uptake	c-Type cytochromes, iron sensor proteins	*omcAB, mtrB, cctA, cymA, fecIR*
	Iron oxidation	Electron shuttling	Rusticyanin, c-type cytochromes	*rus, cyc1, cyc2, coxABCD*
Calcium	Formation of calcareous skeletons	Calcium carbonate precipitation	Vacuolar-type ATPase proton pump	V-ATPase
Silicium	Formation of frustules, spicules	Silica uptake, silica precipitation	Silaffin, frustulin, silicatein	*Sil*

of the photosynthetic complex are the enzymes of the pathways for chlorophyll synthesis and accessory photopigments (Table 3.6).

Several mechanisms are known for the fixation of CO_2 in organic molecules, but the *Calvin cycle* is the most widespread. Both photoautotrophs and chemoautotrophs use the generation of ATP and NADPH to fuel CO_2 fixation. The Calvin cycle uses two key enzymes, ribulose bisphosphate carboxylase (Rubisco) and phosphoribulokinase. Rubisco is the most abundant enzyme of the biosphere and it shows a remarkable constancy across all autotrophic organisms, from bacteria to flowering plants. The complex consists of several subunits, encoded by *rbc* and *cbb* genes. In plants some of the subunits are encoded in the chloroplast, others in the nuclear genome. The second enzyme that is unique to the Calvin cycle is phosphoribulokinase, which is encoded by genes from the *prk* cluster. In green sulphur bacteria (*Chlorobium* is the best-investigated genus), and in some chemoautotrophic bacteria and Archaea, an alternative CO_2-fixation pathway is active, the so-called *reverse citric acid cycle*. Most of the enzymes in this cycle are the same as in the normal Krebs cycle, except for the enzyme ATP citrate lyase, which is specific to this pathway and is used as a genetic marker for reverse citric acid carbon fixation. In *Chlorobium* the enzyme is encoded by two adjacent ORFs, *acIB* and *acIA* (Kanao *et al.* 2001).

Only recently, thanks to a genomics approach, has another mechanism of phototrophy been implicated in the carbon cycle. In this pathway energy is generated from light, not through chlorophyll but through a retinal-dependent light-activated proton pump, *proteorhodopsin*. This mechanism was known previously only from some Archaea (in which the protein was called bacteriorhodopsin) until Béjà *et al.* (2000a) discovered that rhodopsin-like genes were associated with uncultured Gammaproteobacteria of the SAR86 clade in the sea. Further research has shown that this protein is very common in marine bacterioplankton and is present well outside of the phylum Proteobacteria (Béjà *et al.* 2001; De la Torre *et al.* 2003; Venter *et al.* 2004). Proteorhodopsin most probably supports a photoheterotrophic lifestyle in which light is used to generate energy, relieving respiratory costs in bacteria that do not fix CO_2. However, proteorhodopsin may be coupled to an as-yet-unknown photoautotrophic pathway.

Turning to the catabolic part of the carbon cycle we note that a great variety of enzymes are used by heterotrophs to degrade polymers from photosynthetically fixed carbon. These are all part of the *decomposition* subsystem, which returns CO_2 or CH_4 to the atmosphere. Decomposition is a crucial process occurring in soils and sediments. It actually involves more than just breakdown of organic matter, it includes mineralization of nutrients and synthesis of humus, leading to humus–clay complexes that stabilize the soil. All organic matter eventually ends in the decomposer network, to which an enormous variety of heterotrophic organisms contribute. The number of genes that can be used as indicators of the decomposition process is potentially very large; we have mentioned in Table 3.6 only a few enzymes associated with carbohydrate and lipid catabolism. Many of these enzymes are present in bacteria, fungi, and animals; however, cellulase (more correctly called β-1–4-glucanase) has a limited distribution. Most fungi can degrade cellulose, but only few groups of bacteria; no animal, except the tunicates (see Section 2.3), has cellulase. Consequently all animals depend on microorganisms for the degradation of this abundant polysaccharide. Many animals have recruited a specialized microflora in one of their gut compartments to do the job of cellulose degradation for them. The capacity for lignin degradation is even more limited in nature. Certain basidiomycetes called wood-rotted fungi are responsible for this important link in the carbon cycle and they do this by means of aspecific peroxidases. We have seen in Section 2.3 that the white rot fungus *Phanerochaete chrysosporium* was selected as a genomic model for lignin degradation.

In addition to the polymeric substrates attacked by decomposition reactions, a great variety of simpler organic compounds may be used by heterotrophic microorganisms as a source of carbon and to generate energy (e.g. disaccharides, organic acids, phospholipids). A special case is C_1 metabolism, which is conducted by some Gamma- and Alphaproteobacteria (*Methylomonas*, *Methylosinus*) that are able to grow on

compounds with only one carbon atom and thus have to synthesize all carbon–carbon bonds themselves. Methane is the most common one-carbon substrate; the microorganisms that use methane as their sole source of carbon are called *methanotrophs*. These organisms are found wherever a stable source of methane is combined with availability of oxygen; for example in the transition zone between oxic and anoxic strata of lakes and soils. Methane oxidation is an important link in the carbon cycle, because it converts methane back into cell material and CO_2. A key enzyme in the methane oxidation pathway is methane monooxygenase, a haem protein of the same family as ammonium monooxygenase (see below), which is encoded by the *pmo* gene cluster (Table 3.6).

The last link in the carbon cycle to be discussed is the process of methanogenesis. This is an extremely important process in the climate-change issue because methane is a very effective greenhouse gas. Methane is generated in marshes, swamps, lake sediments, paddy fields, and landfill sites under anoxic conditions by a group of strictly anaerobic Archaea, called methanogens. These organisms are found in five orders of the phylum Euryarchaeota and include the genomic model species *Methanococcus jannaschiii*, a hyperthermophile from the ocean floor which was the first fully sequenced archaeon, as well as many mesophilic species such as *Methanosarcina mazei*, whose genome was discussed in Section 2.2. The most important substrate used in methanogenesis is CO_2, which is reduced to CH_4 using H_2 as an electron donor in the overall reaction:

$$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O$$

In addition to CO_2, many methanogens can use simple alcohols and organic acids as well—for example methanol, methylamines, formate, and acetate—but not larger organic molecules such as sugars. This implies that methanogens must rely on decomposing (CO_2 producing) and fermenting (formate- and acetate-producing) microorganisms in their immediate surroundings for the supply of necessary substrates. This trophic interdependence within microbial communities is called *syntrophy*. The reactions leading to the reduction of CO_2 to CH_4 are quite complex and involve some unique enzymes and coenzymes. The terminal step is conducted by the methyl reductase complex, which is encoded by the *mrc* gene cluster. The reaction catalysed by methyl reductase produces not only free methane but also leads to the formation of a complex of two coenzymes (coenzyme B and coenzyme M), linked via a disulphide bridge. This disulphide is then reduced to produce the free coenzymes by the enzyme heterodisulphide reductase, another unique enzyme found only in methanogens; its three subunits are encoded by genes of the *hdr* cluster (Table 3.6).

3.3.2 Key genes in the nitrogen cycle

Figure 3.8 shows an overview of the nitrogen cycle. Most of the inorganic nitrogen on Earth is present as inert dinitrogen gas, N_2, which surrounds all plants and animals, but is unavailable to them. Only certain prokaryotes can utilize dinitrogen gas and reduce it to ammonia, which is then taken up into the biosphere. *Nitrogen fixation* is therefore a key link in the nitrogen cycle (process 1 in Fig. 3.8), although on a global scale it is now surpassed by industrial nitrogen fixation on behalf of the production of mineral fertilizers (the Haber process). Among the nitrogen-fixing organisms are free-living aerobic bacteria such as Cyanobacteria; free-living anaerobic bacteria, such as *Clostridium*; symbiotic bacteria of the family Rhizobiaceae, living in association with leguminous plants; and symbiotic bacteria of the actinomycete genus *Frankia*, living in association with nonleguminous plants such as alder trees (*Alnus*) and buckthorn (*Hippophae*). The ability to fix nitrogen is not associated with a monophyletic group of prokaryotes, but is found scattered among several bacterial and archaeal lineages, phototrophs, chemoorganotrophs, and chemolithotrophs. Nitrogen fixers are also called *diazotrophs* (deriving nutrition from dinitrogen).

The reduction of N_2 to NH_3 is a highly energy-demanding process, catalysed by the enzyme complex *nitrogenase*, which consists of two units, dinitrogenase and dinitrogenase reductase. Dinitrogenase contains iron and molybdenum in its active centre, whereas

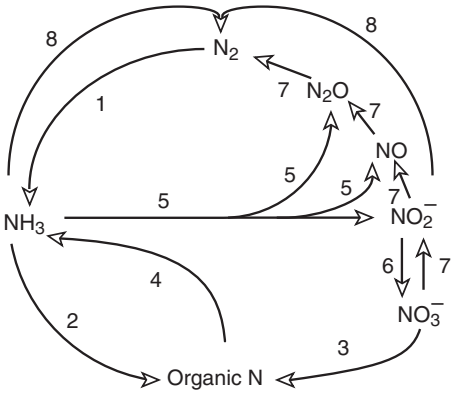


Figure 3.8 Overview of the nitrogen cycle. 1, Nitrogen fixation; 2, ammonium assimilation (uptake by microorganisms and plants); 3, reductive nitrate assimilation (uptake); 4, ammonification (deamination, mineralization, decomposition); 5, aerobic ammonia oxidation; 6, nitrite oxidation; 5+6, nitrification; 7, denitrification; 8, anaerobic ammonia oxidation. Modified after Kowalchuk and Stephen (2001) and Jetten (2008), reproduced with permission from Annual Reviews.

dinitrogenase reductase contains only iron. Because the fixation of N_2 requires such a large amount of energy, the activity of these enzyme complexes is highly regulated. The genetic architecture of nitrogen fixation has been studied in detail in the model species *Klebsiella pneumoniae* (Gammaproteobacteria, Enterobacteria), a species normally living in soil or water but occasionally causing pneumonia in humans, hence its name. The structural and regulatory genes associated with nitrogen fixation are organized in a complex network of operons, the *nif* regulon, encoding no less than 20 different proteins all dealing with regulation, maturation, and assembly of the nitrogenase complex (Fig. 3.9). The various *nif* genes are indicated by letters; for example, *nifK* encodes the *β*-subunit of denitrogenase, *nifD* encodes a subunit of the same enzyme, and *nifH* encodes the enzyme dinitrogenase reductase. The regulon also includes several proteins that support the processing of Mo, its insertion in the apo-enzyme, and expression regulators and inhibitors, and so on. Similar *nif* regulons are present in other nitrogen fixers such as cyanobacteria. Some of the *nif* genes have been used succesfully to develop probes for assessing diazotroph microbial communities in the field (see below). In addition to the *nif* cluster, symbiotic nitrogen fixers have a *fixABCX* gene cluster, which encodes *electron-transport flavoproteins* (ETFs), which presumably support the electron transport to nitrogenase. This *fix* gene cluster is under transcriptional control by one of the regulatory proteins from *nif*, NifA.

Nitrogen can assume six different oxidation states, with the valency of the nitrogen atom varying from +5 in nitrate to –3 in organic N (amino groups and heterocyclic compounds). Nitrogen fixation is a reduction process because it changes N from oxidation state 0 to –3. Another reduction takes place in the process of *denitrification*, where N changes from +5 to 0. Denitrification (process 7 in Fig. 3.8) is the biological conversion of nitrate to nitrogen gas, a process that is detrimental in agriculture, because it diminishes the effect of fertilization, but favourable in other situations, such as sewage treatment, because it removes eutrophicating nitrate from the effluent. Most denitrifying prokaryotes are members of the Proteobacteria and they represent a metabolically versatile group, growing both aerobically and anaerobically. Some denitrifying bacteria will also use other electron acceptors such as ferric ion (Fe^{3+}). Four consecutive enzyme systems are involved with denitrification, converting nitrate subsequently to nitrite, NO, N_2O, and finally N_2. The enzymes are referred to as nitrate reductase, nitrite reductase, NO reductase, and N_2O reductase, respectively. Some bacteria, such as *E. coli*, can only carry out the first two steps and therefore generate NO, not N_2. Natural denitrification, although considered beneficial in the fight against eutrophication, has negative side effects because it is always accompanied by the emission of NO and N_2O, which in themselves are greenhouse gases and by reaction with water in the atmosphere contribute to the acidity of rain. The reduction of nitrate in the environment is also called *dissimilative nitrate reduction*, to distinguish it from nitrate reduction following uptake in plants, which is called *assimilative nitrate reduction* (process 3 in Fig. 3.8).

The biochemistry of denitrification has been studied in detail in *Paracoccus denitrificans*, which can conduct all four denitrification steps. This organism can express both a periplasmic nitrate reductase (between cell membrane and outer membrane) and a membrane-bound nitrate reductase, where the

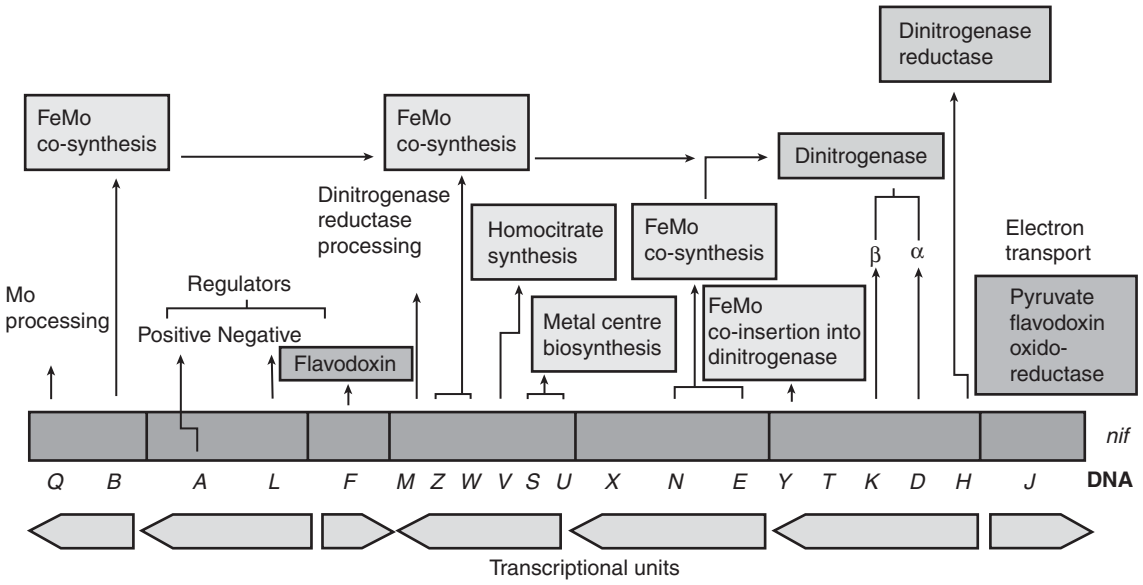


Figure 3.9 Schematic representation of the *nif* regulon in *Klebsiella pneumoniae.* The 20 *nif* genes are indicated by letters (*Q*, *B*, *A*, etc.). They are transcribed as seven polycistronic messengers indicated below the genes; the direction of transcription differs between operons. The functions of the various proteins are indicated in the boxes and by arrows. From Madigan *et al.* (2002), with permission from Pearson Education.

latter is formed only under anaerobic conditions and also functions exclusively under these conditions. A third nitrate reductase is active in the cytoplasm and associated with assimilatory nitrate reduction. The membrane-bound reductase is encoded by a *nar* gene cluster and the periplasmatic nitrate reductase by the *nap* operon. As in the case of nitrogen fixation, these clusters include both structural genes, encoding subunits of the enzymes, and regulatory proteins. Another gene cluster, *nir*, encodes nitrite reductase, whereas *nor* genes encode NO reductase, and *nos* genes encode N_2O reductase. Because of the toxicity of nitrite and nitric oxide, the expression of these genes must be highly coordinated; a comprehensive set of regulators (including nitrite sensors) is involved, as well as transporters that may prevent intracellular accumulation of nitrite. Van Spanning *et al.* (2005) provided a review of the distribution of denitrification-associated genes in Bacteria and Archaea.

The third link in the nitrogen cycle that we discuss here is the process of *nitrification*, of which the first and rate-limiting step is aerobic ammonia oxidation (process 5 in Fig. 3.8). Until recently it was assumed that, unlike nitrogen fixation and denitrification, the capacity for ammonia oxidation had a limited distribution within the bacterial kingdom. However, metagenomic screening has suggested that ammonia oxidation may be more widespread (see Section 3.4). Several terrestrial ammonia-oxidizing bacteria are present in the Betaproteobacteria, involving the well-known genera *Nitrosomonas* and *Nitrosospira*. Another group of ammonia oxidizers is found in the Gammaproteobacteria (*Nitrosococcus*), but these organisms seem to have a limited distribution, being mainly marine. Nitrite oxidizers are found in the Alphaproteobacteria (Nitrobacter), the Deltaproteobacteria (*Nitrospina*), and the phylum Nitrospirae (Kowalchuk and Stephen 2001). The β-ammonia-oxidizing bacteria *Nitrosomonas*/*Nitrosospira* group forms a monophyletic lineage and can be probed with a single set of 16S rRNA primers. Most ammonia-oxidizing bacteria are autotrophic; in addition to the system for nitrate oxidation they have the Calvin cycle for CO_2 fixation. The ATP and reducing power requirements for this process, added to the relatively limited amount of ATP generated by nitrate oxidation,

may explain the slow growth of these bacteria in laboratory cultures.

The key enzyme in aerobic ammonia oxidation is *ammonia monooxygenase*, which oxidizes ammonia to hydroxylamine (NH_2OH), which is then oxidized further to nitrite by hydroxylamine oxidoreductase. Ammonia monooxygenase splits O_2, and incorporates one of the oxygen atoms in NH_3, while the other reacts with H^+ to form water. Similar reaction schemes using monooxygenase enzymes apply to the oxidation of methane and organic compounds such as benzene. The genome of *Nitrosomonas europaea* has been completely sequenced and serves as the main genetic model for nitrification research (Chain *et al.* 2003). Two genes in the *amo* cluster of *Nitrosomonas*, *amoA* and *amoB*, encode the two subunits of monooxygenase, whereas a third gene, *amoC*, encodes a supporting membrane protein. *AmoA* is one of the genes that is used in microarray-based profiling of environmental samples (see below). The enzyme nitrite oxidoreductase, which is used by nitrite oxidizers to produce nitrate from nitrite, is encoded in the *nor* operon, which is essentially the same gene cluster as used by denitrifiers to reduce NO to N_2O.

In addition to being oxidized under aerobic conditions, ammonia can also be oxidized under anoxic conditions by means of a process called anammox. This involves the joint use of ammonia and nitrite in the overall reaction:

$$NH_4^+ + NO_2^- \rightarrow N_2 + 2H_2O$$

indicated by process 8 in Fig. 3.8. The first organism that was identified as responsible for anoxic ammonia oxidation, *Brocadia anammoxidans*, is a member of the bacterial phylum Planctomycetes, a somewhat unusual group of prokaryotes because they have membrane-enclosed compartments inside the cell. One such compartment, the anammoxosome, is geared specifically towards the anammox process. *Br. anammoxidans* is an autotroph and uses nitrite as an electron donor to fix CO_2. This is the same principle used by nitrite oxidizers (e.g. *Nitrobacter*) and it was long assumed that the same enzymes were used. However, the real key enzymes in the anammox reaction were identified after the complete genome of another anammox bacterium, *Kuenenia stuttgartiensis*, was sequenced (Strous *et al.* 2006). They are hydrazine synthase (*hzsABC*) and hydrazine oxidoreductase (*hzoAB*). The first enzyme catalyses the synthesis of hydrazine, N_2H_4, from NO and NH_4^+; subsequently, hydrazine is reduced by hydrazine oxidoreductase to generate N_2. The *hzsABC* gene cluster is best used as a marker gene for screening the presence of anammox potential in the environment (Table 3.6).

Finally, the mineralization of organic nitrogen to ammonia (*ammonification*) takes place during the decomposition of organic material and involves *deaminase* reactions. Many heterotrophic organisms can do this and there are many different deaminases, so no key enzyme for ammonification can be indicated. Under neutral to acid pH conditions ammonia from decomposition reacts with water and is present in the environment as an NH_4^+ ion, adsorbed to the sediment or soil-exchange complex. At low pH very little NH_4^+ dissociates to form NH_3 and because NH_3, not the ammonium ion, is required for ammonia oxidation, low pH was long thought to be a limiting factor for nitrification; however, even at pH 3 in nitrogen-saturated coniferous forest soil a substantial rate of nitrification by acid-tolerant *Nitrosopira* has been measured (Laverman *et al.* 2001).

This overview of the nitrogen cycle shows that it is accompanied by a great variety of redox reactions and conducted by many different microorganisms. For most of the transitions the key enzymes have been well characterized. It is evident that in some cases similar enzymes conduct different reactions in different organisms and apparently their action depends on the context of the organism in which they are expressed. Several genes are indicative of key steps in the nitrogen cycle and these can be used in environmental genome-profiling studies.

3.3.3 Other nutrient cycles

Sulphur, like nitrogen, can take different oxidation states and the transitions between these are exploited by many microorganisms. One of the best known reactions is *dissimilative sulphate reduction*, the conversion of sulphate to sulphide, which can

be liberated as smelly H_2S gas or precipitate with cations. Sulphate-reducing bacteria are found in seven different lineages of Bacteria and Archaea, as we have seen above (Fig. 3.6). The species *Desulfovibrio vulgaris* (Deltaproteobacteria), whose genome has been sequenced (Heidelberg *et al.* 2004), is a model species for the study of sulphate reduction. Most sulphate-reduction processes take place under anoxic conditions and use electron donors from organic compounds. Intertidal sediments are prime examples of sulphate-reducing environments, because of the combination of high sulphate availability in seawater with abundant organic material from salt-marsh vegetation and anoxic conditions due to regular flooding. The black colour of intertidal sediments is due to the precipitation of sulphide with iron to form FeS. In the sea itself, sulphate reduction is limited by the availability of carbon sources. In addition to sulphate reduction in the environment, sulphate is also reduced after uptake by microorganisms, plants, and animals (*assimilative sulphate reduction*), because the major form of sulphur in organic molecules is the reduced form of the thiol (SH) group in proteins.

The reduction of sulphate proceeds through a series of steps. Sulphate is first activated by binding to ATP to form adenosine phosphosulphate (APS) before it can be reduced to sulphite by the enzyme APS reductase. In the next step sulphite is reduced to sulphide by sulphite reductase. The two reactions require electrons from a suitable donor, either hydrogen or acetate, which is oxidized to water or CO_2. The *aps* locus of sulphate reducers encodes the key enzyme APS reductase, and *dsrA* and *dsrB* encode the α and β subunits of sulphite reductase. The three genes are highly conserved in several deep-branching phyla of the Bacteria and Archaea and seem to have been subject to lateral gene transfer. Sulphate reducers often occur in close association with methanogenic Archaea and with methanotrophic (methane oxidizing) Archaea and they use each other's products (another case of syntrophy). Taken together, the only oxidants that this small community requires are CO_2 and sulphate, compounds which existed before the Earth became oxygenated by photosynthesis. Therefore, communities consisting of sulphate reducers, methanogens, and methanotrophs, found in salt marsh sediment and in association with hydrothermal vents on the ocean floor, could represent the modern descendants of extremely ancient communities dating back to the beginning of the Proterozoic era (Teske *et al.* 2003).

In addition to H_2S, a major source of sulphur emission to the atmosphere is due to the volatile compound dimethyl sulphide (DMS; CH_3–S–CH_3). Microbial mats of salt marshes are well-known emission sources for DMS. This compound is a degradation product of β-dimethylsulphoniopropionate (DMSP), which is produced in large amounts by many marine algae, cyanobacteria, and salt-marsh plants, probably to protect them against osmotic shock (Yoch 2002). DMS is assumed to have an anti-greenhouse effect, because sulphate aerosols generated by DMS act as cloud-condensing nuclei, reducing the amount of sunlight reaching the Earth's surface. However, by generating sulphate, DMS may contribute significantly to acid rain in places where there is no air pollution and the concentration of aerosols is low. When the DMSP-containing organisms die off or are grazed by zooplankton, DMSP becomes available in the environment and is converted to DMS by means of an enzyme from the family of type III acyl coenzyme A transferases. The enzyme catalyses the binding of acyl coenzyme A to DMSP, after which it is cleaved to form DMS. The gene responsible for this reaction (*dddD*) was recently cloned from the marine bacterium *Marinomonas* (Todd *et al.* 2007). Other bacteria may have other ways of making DMS, because *dddD* homologues could not be found in DMS emitting strains of *Sulfobacter* and *Roseovarius*. Bacteria from the genus *Roseobacter* are also known for DMS emission; some live in close association with dinoflagellates, which are major producers of DMSP, and it is assumed that the bacteria may benefit from the DMSP produced by the dinoflagellates (Miller and Belas 2004).

Reduced sulphur compounds such as elemental sulphur and sulphide can also be used as electron donors in the process of *sulphur oxidation*. Sulphur oxidizers often live attached to a surface, because elemental sulphur does not dissolve in water. Oxidation, from either S^0 or S^{2-}, leads to the

production of sulphite, sulphate, and H^+ and may lower the ambient pH considerably. Sulphur oxidizers are autotrophs that fix CO_2 by means of the Calvin cycle. The sulphur-oxidizing system was investigated in detail in the Gram-negative lithoautotrophic species *Paracoccus pantotrophus* (Friedrich *et al.* 2001, 2005). The *sox* gene cluster of *P. pantotrophus* comprises at least 15 different protein-encoding genes. Seven of these genes encode four proteins with sulphite oxidase and sulphur dehydrogenase activities. The cluster also contains transcriptional regulators and transport regulators. Evidence has emerged that similar proteins are present in at least 19 other bacteria (Friedrich *et al.* 2005). Homologies can be drawn between the genes of the various species (Fig. 3.10), and a common mechanism for all sulphur-oxidizing bacteria is emerging (Friedrich *et al.* 2001). The sulfur oxidation pathways of Archaea seem to be quite different from those in bacteria, as Archaea do not seem to have *sox* genes. The best investigated species in this respect is *Acidianus ambivalens*, a member of the Sulfobolales. A key enzyme of sulphur oxidation was characterized in this species, a cytoplasmic sulphur oxygenase reductase (SOR) (Friedrich *et al.* 2005).

Regarding the *phosphorus cycle*, we note that it is dominated by physical processes (erosion, precipitation) more than by biological influences. Phosphorus undergoes hardly any microbe-mediated redox reactions in the environment; an exception is the process of phosphite oxidation conducted by lithoautotrophic sulphate reducers (Schink and Friedrich 2000), in which the redox state of phosphorus changes from +3 to +5. All organisms need phosphate, which they take up using membrane-bound transport systems, of which there is a great variety. Phosphorus may be present in the environment as polyphosphates (long chain-polymers of phosphate, contrasted with orthophosphate, PO_4^{3-}) and these need to be hydrolysed before take-up. Microorganisms use exopolyphosphatase to hydrolyse polyphosphates. The *ppx* operon of *E. coli* K12 encodes this enzyme, and presumably many microorganisms have similar loci. In addition, all cells, prokaryotic and eukaryotic, accumulate polyphosphate internally, where it functions as a phosphate reserve and as a sequestration mechanism for otherwise toxic cations, for example Ni and Pb (Kulaev and Kulakoskaya 2000). Another major source of phosphate in the environment is organic phosphorus in dead organic matter, from which phosphate is released during decomposition by phosphatase activity of heterotrophic microorganisms. The fact that biota do not directly alter the chemical appearance of phosphorus in the environment to a great extent does not diminish the important growth-limiting effect that phosphate has in most aquatic systems and to an extreme degree in the sea.

Iron has two oxidation states, ferric (III) and ferrous (II), and undergoes redox reactions in the environment catalysed by microorganisms. Since the valency of the iron atom is correlated with differences in solubility and stability of its compounds, iron redox reactions have important consequences for the biogeochemical cycles of other elements, especially sulphur and phosphorus. For example, Fe^{2+} may precipitate with S^{2-} to form insoluble FeS, preventing the release of gaseous H_2S, and Fe^{3+} oxides may bind phosphate, preventing the dissolution of phosphate in water. The latter phenomenon is well-known among limnologists who are familiar with the fact that release of phosphate from sediment into the overlying water is much greater under anaerobic conditions than from the same sediment under aerobic conditions.

The most important bacteria conducting *iron reduction* in the environment are from the genus *Geobacter* (Deltaproteobacteria). Other iron reducers are *Shewanella* (mainly marine), *Geothrix*, and *Anaeromyxobacter* (North *et al.* 2004). Most iron reducers can also use other metal ions as electron acceptors, such as uranium (VI) and manganese (IV). Various organic molecules (e.g. acetate, benzene, and toluene) can act as electron donors in Fe^{3+}-dependent anaerobic growth. By oxidizing organic molecules in conjunction with iron reduction, iron-reducing bacteria may contribute significantly to the natural purifying capacity of subsoils, for example when ferric-rich aquifers are polluted by leachate from landfills or oil-storage tanks (Lovley 2003; see also Fig. 3.4). The genomes of *Shewanella oneidensis* and *Geobacter sulfurreducens* have been sequenced completely (Heidelberg *et al.*

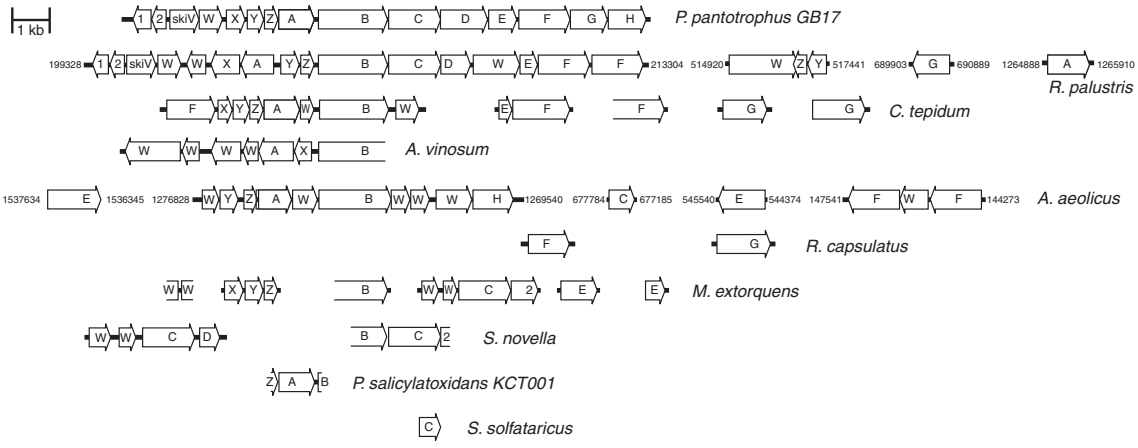


Figure 3.10 Map of the *sox* gene cluster of *Paracoccus pantotrophus* and other sulphur-oxidizing bacteria. Homologies are indicated by common letters. Explanation of gene codes: 1, ArsR-type transcriptional regulator; 2, periplasmic thioredoxin; skiV, membrane-bound transporter involved with cytochrome *c* biogenesis; W, periplasmic thioredoxin; X, Y, Z, A, B, C, and D, structural genes encoding four proteins—SoxXA, SoxYZ, SoxB, and SoxCD; E–H, proteins of unknown function. *P. pantotrophus*, *Paracoccus pantotrophus*; *R. palustris*, *Rhodopseudomonas palustris*; *C. tepidum*, *Chlorobium tepidum*; *A. vinosum*, *Allochromatium vinosum*; *A. aeolicus*, *Aquifex aeolicus*; *R. capsulatus*, *Rhodobacter capsulatus*; *M. extorquens*, *Methylobacterium extorquens*; *S. novella*, *Starkeya novella*; *P. salicylatoxidans*, *Pseudaminobacter salicylatoxidans*; *S. solfataricus*, *Sulfolobus solfataricus.* After Friedrich *et al.* (2001), by permission of the American Society for Microbiology.

2002; Methé *et al*. 2003). The genomes of these species contain an unprecedented number of c-type cytochromes, some with two or more haem groups. The abundance of cytochromes highlights the importance of electron transport to these organisms. Three different strategies are followed by Fe^{3+} reducing bacteria (Lovley *et al*. 2004): they can attach to an iron-rich surface and transfer electrons directly to the mineral, they can release iron chelators to bring iron to the cell where it can be reduced, or they can use extracellular *electron shuttles* to transfer electrons back and forth between the mineral and the cell. A model for the latter mechanism, iron reduction at a distance, was proposed by Croal *et al.* (2004) and Lovley *et al.* (2004) (see Fig. 3.11a).

Different iron-reducing bacteria seem to have different sets of cytochromes, although there are also some genes that are common among the three species *G. sulfurreducens*, *Sh. oneidensis*, and *Desulfovibrio vulgaris.* These genes are potential markers for iron reduction activity in the environment. Recently, Chin *et al.* (2004) showed that expression of *omcB*, a gene encoding an outer membrane *c*-type cytochrome (Fig. 3.11a), was very well correlated with Fe-reduction rates in *G. sulfurreducens*. It is not yet certain that this gene can be used as a universal indicator of Fe-reducing capacity in the environment; the relatively small overlap between the derived gene complements of different iron-reducing bacteria suggests that their metal-reducing capabilities are not simply related to their sharing an exclusive set of genes.

In addition to metal reductase activity associated with anaerobic respiration, iron is also reduced during uptake into the cell (assimilative iron reduction). Fe is often scarce in the environment because of the insolubility of $Fe(OH)_3$. Therefore many organisms have developed mechanisms of scavenging Fe^{3+} by excreting compounds in the medium with an extremely high iron affinity. Citrate is often used for this purpose, but also molecules completely geared to Fe chelation, *siderophores*, are used. These compounds may either donate Fe to a membrane-bound transporter or be taken up entirely by pinocytosis after binding to a membrane receptor. Genes encoding ferric citrate transporters and iron siderophore receptor proteins are found in many bacteria. Expression of siderophore receptor proteins is regulated by membrane-bound sensor systems encoded by *fecIR genes*.

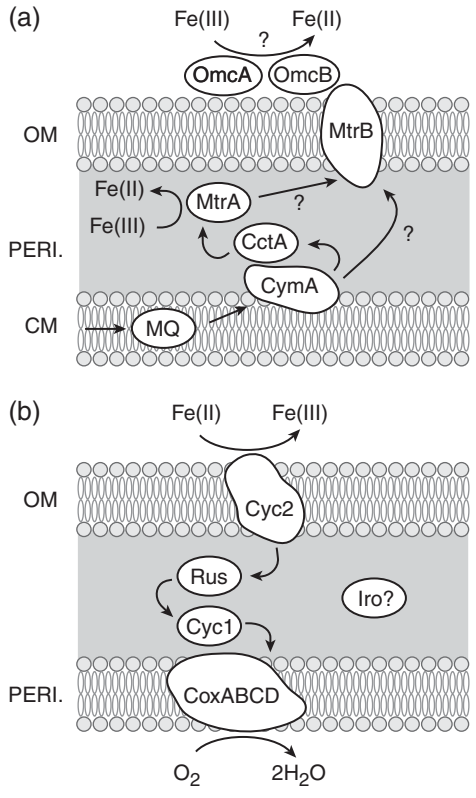


Figure 3.11 Models for electron-shuttling processes involved in (a) Fe^{2+} reduction by *Shewanella oneidensis* MR-1 and (b) Fe^{3+} oxidation by *Acidithiobacillus ferrooxidans*. OM, outer membrane; PERI., periplasmatic space; CM, cell membrane; OmcA and OmcB, outer-membrane cytochromes A and B; MtrA and MtrB, metal-reduction proteins A and B; CctA, small tetrahaem *c*-type cytochrome A; CymA, inner-membrane cytochrome A; MQ, menaquinone; Cyc1 and Cyc2, c-type cytochromes 1 and 2; Rus, rusticyanin; Iro, ferredoxin; CoxABCD, cytochrome oxidase complex. After Croal *et al.* (2004) with permission from Annual Reviews.

Few bacteria use *iron oxidation* as an energy-yielding reaction; large amounts of iron need to be oxidized in order to deliver sufficient energy for growth. Iron oxidation can take place under anaerobic conditions by nitrate reducers and photoautotrophs (Straub *et al.* 2004), but most iron-oxidizing bacteria are aerobic acidophiles. At neutral pH hardly any energy can be gained from iron oxidation under aerobic conditions because ferrous iron then oxidizes spontaneously to ferric iron. Iron oxidation leads to the formation of $Fe(OH)_3$ precipitates; this process is recognizable in many places with a boreal climate, where iron-rich groundwater from acidic marshes comes into contact with aerobic sandy soils. Another environment where iron oxidation is important is acid-mine drainage; a very acid leachate develops under the influence of iron oxidizers and sulphur oxidizers when iron sulphide comes into contact with the air in coal mines (see Section 4.4).

Model species for iron oxidation are *Acidithiobacillus* (*Thiobacillus*) *ferrooxidans* (Gammaproteobacteria) and *Leptospirillum ferrooxidans* (Nitrospirae). A crucial role in iron oxidation by *Ac. ferrooxidans* is played by a periplasmic copper-containing protein, *rusticyanin* (Fig. 3.11b). This protein oxidizes ferrous iron in the periplasm and then reduces a cytochrome in the cell membrane. After a very short electron-transport chain O_2 is reduced to water under consumption of H^+, allowing the generation of ATP. The *rus* operon of *Ac. ferrooxidans* encodes the rusticyanin protein as well as two c-type cytochromes (*cyc1* and *cyc2*) and four subunits of a cytochrome *c* oxidase (*coxABCD*). Recent work by Yarzábal *et al.* (2004) has shown that expression of this operon is induced by ferrous irons and so *rus* seems an excellent candidate for indicating the capacity of iron oxidation in the environment (Table 3.6).

Regarding the *calcium cycle*, a significant influence, at least in oceanic systems, is due to coccolithophorid algae (Coccolithophorida, Haptophyta). *Emiliania huxleyi* is a well-known representative from this group, recognized by its beautiful platelets of calcite, *coccoliths*, which form a skeleton structure around the cell (Fig. 3.12). *Em. huxleyi*, although very small, is extremely common in the ocean, such that extensive fields of *Em. huxleyi* blooms can be observed even from space. Coccolithophorids are favourite study objects among palaeontologists because they are abundant in marine sediments extending back to the Cambrian era, and may be used as indicators for climate reconstruction. The ecological relevance of these organisms is due to the fact that they fix carbon not only for synthesis of organic molecules but also for coccolith formation. By sedimentation of cells to the bottom of the ocean significant amounts of carbon and calcium are withdrawn

from the biological sphere of influence. The coccoliths are formed inside the cell in specialized calcifying vesicles of the Golgi apparatus, and later pushed outwards. The formation of calcifying vesicles is an interesting model for *biomineralization*. It is assumed to involve a vacuolar-type ATPase (V-ATPase), a proton pump that is known from many other eukaryotes as well. Ca^{2+} and ions are assumed to precipitate as $CaCO_3$ on a macromolecular template inside the vesicle. Coccolithophores have very large genomes, exceeding 200 Mbp. Wahlund *et al.* (2004) developed a cDNA library and sequenced 3000 ESTs. The library was found to contain multiple copies of genes for calnexin and calreticulin, two chaperones that play a key role in calcium homeostasis, but the authors did not find a V-ATPase. The complete genome of *Em. huxleyi* is presently being sequenced by the US Joint Genome Institute. Further analysis of the genome sequence is needed before a definitive marker gene for calcification can be identified.

The *silicium cycle* in the ocean is greatly influenced by diatoms (Bacillariophyta) who incorporate silicium in their silica cell wall or *frustule*. Like coccolithophorids, diatoms withdraw carbon from the atmosphere by sinking to the bottom of the sea, and in doing so also remove silicium from the sea water. The delicate structure of diatom frustules and their bewildering array of forms (Fig. 3.13) have fascinated biologists ever since the invention of the microscope and their reproducible fine structure presently also raises interest among nanotechnologists (Bradbury 2004). The genomes of *Thalassiosira pseudonana*, a diatom with a centric cell shape, resembling a Petri dish, and of *Phaeodactylum tricornutum*, a species with elongate, pennate cells, have recently been sequenced (Armbrust *et al.* 2004; Bowler *et al.* 2008). Not surprisingly, the diatom genomes contain many genes with a function in silicium metabolism. Armbrust *et al.* (2004) identified three genes that encode transporters for active uptake of silicic acid. Silica precipitates in a silica-deposition vesicle, in a process controlled by a family of phosphoproteins, called silaffins, that are embedded in the frustule while it is being formed. Five such silaffins could be identified in the genome of *Th. pseudonana*, but surprisingly these did not match proteins of similar function in other diatoms and sponges. Another group of proteins associated with frustule formation are the frustulins (casing glycoproteins), which protect the outside of the frustule.

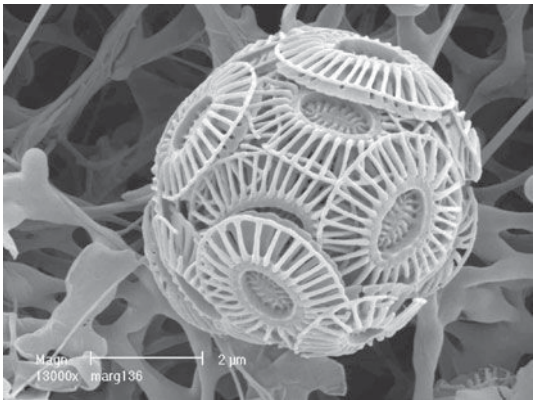


Figure 3.12 Scanning electron micrograph of *Emiliania huxleyi*, a representative of the Coccolithophorida (Chrysophyta, golden algae), well-known for its beautiful external platelets, coccoliths, made of calcium carbonate. Coccolithophorids are a key link in the oceanic carbon and calcium cycles. Courtesy of Jeremy Young and Markus Geisen, Palaeontology Dept., The Natural History Museum, London.

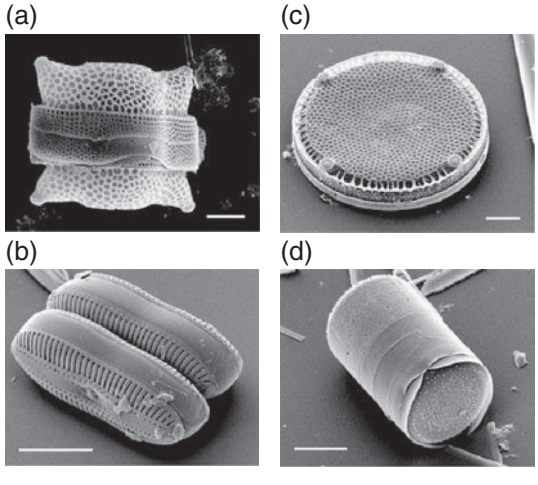


Figure 3.13 Scanning electron micrographs of centric and pennate species of diatom. (a) *Biddulphia reticulata*, (b) *Diploneis* sp., (c) *Eupodiscus radiatus* (single valve), (d) *Melosira varians*. Scale bars, 10 mm. Courtesy of Mary Ann Tiffany, San Diego State University.

3.3.4 Microarray-based detection of functional genes in the environment

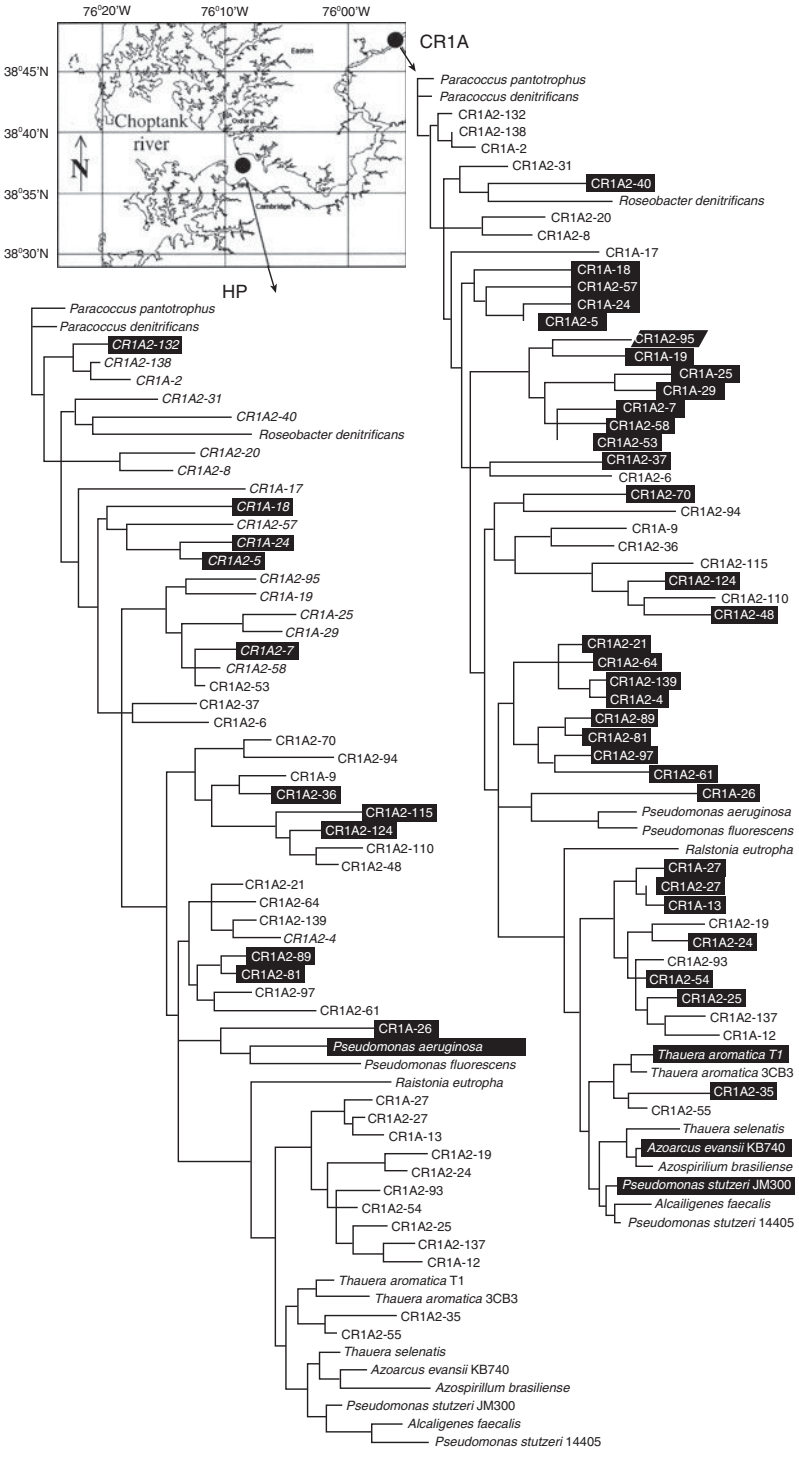
Not all the genes discussed above and mentioned in Table 3.6 have been exploited for environmental screening with genomics technology. Until now, the nitrogen cycle has attracted most attention, primarily because it offers such a rich variety of enzymatic systems, many of which have well-described biochemistries (Ye and Thomas 2001). We saw in Table 3.2 that *functional gene microarrays* are designed to detect the presence of genes with specific functions in the environment, irrespective of the species. Depending on hybridization conditions, all genes showing a sequence similarity of more than 80% with the probes will be detected. Obviously, microarray screening of functional genes provides an indication of the capacity to conduct the function, it does not demonstrate the function itself. Assessment of the function itself requires isolation of mRNA, followed by competitive microarray hybridization, as in the case of transcription profiling in model organisms. Such studies are more difficult in environmental microbial communities than in model eukaryotes, because of the instability of microbial mRNA, which is characterized by long transcription products without poly(A) tails covering several genes, and the ensuing difficulty of obtaining reliable RNA samples other than rRNA from environmental matrices (Ye *et al.* 2001).

One of the first attempts to apply DNA array technology for assessing functional gene diversity was made by Wu *et al.* (2001). These authors developed a prototype microarray with 104 probes representing PCR-amplified genes from bacteria involved in nitrogen cycling: the nitrite reductase genes *nirS* and *nirK*, and the ammonium monooxygenase gene *amoA*. Strong hybridization signals were obtained with DNA extracted from a marine sediment and from a surface soil. The diversity of *nirS* and *nirK* genes was similar in the two environments, indicating comparable capacities for denitrification. Similar results were obtained by Tiquia *et al.* (2004), who used an oligonucleotide microarray with several hundred 50-mer probes developed from genes in the cycles of nitrogen, carbon, and sulphur (*nirS*, *nirK*, *amoA*, *nifH*, *pmoA*, and *dsrAB*). Although the potential for detection was clearly demonstrated, sensitivity was still a critical issue in these hybridization experiments, because they used DNA extracted directly from the environment without prior PCR amplification. It is estimated that only genes from populations contributing more than 5% to the environmental genomes can be detected in this way (Cho and Tiedje 2002); however, this figure can be brought down to 1% by using modified 70-mer probes printed on special substrates (Denef *et al.* 2003).

Bodrossy *et al.* (2003) developed a functional gene microarray for methanotrophs. A survey of published *pmoA* and *amoA* gene sequences allowed the construction of a nested set of 68 probes that together could diagnose almost the entire known diversity of methanotrophs. A microarray with these probes was fabricated and used to survey methanotroph diversity in soils and microcosms. The study confirmed that all cells representing more than 5% of the targeted community could be detected.

As an example of an early study using a functional gene microarray in an ecological context we discuss the work by Taroncher-Oldenburg *et al.* (2003). These authors developed a microarray with 70-mer oligonucleotides targeting genes encoding the functions of nitrification (*amoA*), nitrogen fixation (*nifH*), and denitrification (*nirK*, *nirS*). Samples were taken from sediments in the Choptank river (in Maryland, USA) and extracted DNA was amplified with gene-specific primers. Data for *nirS* diversity are shown in Fig. 3.14. Among 64 *nirS* probes, 29 were detected in a sample from the upstream area, and 11 in the downstream area. Hybridizations were distributed over a great variety of sequences; that is, they were not limited to a distinct cluster in

Figure 3.14 Comparison of *nirS* (cytochrome cd_1-containing nitrite reductase) gene diversity in two samples taken from sediments of the Choptank river, Maryland, USA. A total of 64 *nirS* probes (each 70 bp) were spotted on a microarray and the similarity between these sequences is shown as a dendrogram. Probes were developed from environmental *nirS* sequences isolated earlier (indicated by codes) and from pure cultures of bacteria (indicated by species names). Positive hybridizations are shown in white type on a black background. The diversity of *nirS* sequences is greater in the upstream location (CR1A) than in the downstream location (HP). After Taroncher-Oldenburg *et al.* (2003) by permission of the American Society for Microbiology.


76°20'W
76°10'W
76°00'W
38°45'N
38°40'N
38°35'N
38°30'N
Choptank
river
N
Easton
Oxford
Cambridge
CR1A
HP
Paracoccus pantotrophus
Paracoccus denitrificans
CR1A2-132
CR1A2-138
CR1A-2
CR1A2-31
CR1A2-40
Roseobacter denitrificans
CR1A2-20
CR1A2-8
CR1A-17
CR1A-18
CR1A2-57
CR1A-24
CR1A2-5
CR1A2-95
CR1A-19
CR1A-25
CR1A-29
CR1A2-7
CR1A2-58
CR1A2-53
CR1A2-37
CR1A2-6
CR1A2-70
CR1A2-94
CR1A-9
CR1A2-36
CR1A2-115
CR1A2-124
CR1A2-110
CR1A2-48
CR1A2-21
CR1A2-64
CR1A2-139
CR1A2-4
CR1A2-89
CR1A2-81
CR1A2-97
CR1A2-61
CR1A-26
Pseudomonas aeruginosa
Pseudomonas fluorescens
Raistonia eutropha
CR1A-27
CR1A2-27
CR1A-13
CR1A2-19
CR1A2-24
CR1A2-93
CR1A2-54
CR1A2-25
CR1A2-137
CR1A-12
Thauera aromatica T1
Thauera aromatica 3CB3
CR1A2-35
CR1A2-55
Thauera selenatis
Azoarcus evansii KB740
Azospirillum brasiliense
Pseudomonas stutzeri JM300
Alcaligenes faecalis
Pseudomonas stutzeri 14405

the dendrogram of probe sequences. The data suggest that the composition of denitrification genes is quite different between the two sampling sites. It is also evident that the diversity of denitrification genes is lower in the downstream location. Denitrification rates measured in sediment cores from the downstream river station were also lower than in the upstream location. So, this study suggests a positive correlation between community biodiversity and ecological function, in accordance with the 'rivet hypothesis' mentioned at the beginning of this chapter. The differences in the composition of the denitrifier community were possibly related to an environmental gradient of salinity and dissolved organic carbon in the river.

Following on from these pioneering studies, the most commonly used microarray for detection of functional genes became the GeoChip developed by He *et al.* (2007), and a later version of it, called GeoChip 3.0 (He *et al.* 2010). GeoChip 3.0 addresses genes from 292 gene families with a total of 27 812 oligonucleotide (50-mer) probes. These genes were selected to represent specific microbe-mediated reactions in the cycles of carbon, nitrogen, phosphorus, and sulphur, and in energy metabolism, antibiotic resistance, metal resistance, and contaminant degradation. Using the array in a field experiment, He *et al*. (2010) were able to show that plots in which a monoculture of plants was growing had fewer functional genes in the soil bacterial community compared to plots with four, nine, or sixteen species of plants.

The functional gene microarray GeoChip 2.0 has been applied in several ecological studies. Yergeau *et al.* (2007) used it to analyse the diversity of carbon and nitrogen cycling genes in microbial communities across a latitudinal gradient in Antarctica. The diversity and abundance of several genes could be correlated with the environmental gradient, for example cellulose degradation genes and denitrification genes were correlated with temperature, and nitrogen fixation genes were correlated with soil covered by lichens. In a follow-up study, the abundance of functional genes was linked to the diversity of taxonomic groups as assessed with the PhyloChip (Yergeau *et al.* 2009). This study was the first to combine GeoChip data with PhyloChip data. A multivariate statistical analysis (principal coordinates analysis, PcoA) of GeoChip and PhyloChip data is summarized in Fig. 3.15. This display shows that some functional genes are highly correlated with certain taxonomic groups. For example, chitinase and mannase genes were correlated with Bacteroidetes, methane oxidation genes with Alphaproteobacteria, and cellulase genes with Actinobacteria. Although these correlations do not demonstrate that the identified functional genes are actually present in the genomes of the identified taxa, they provide a first insight into how functional diversity and species diversity link together. Over all there was a good consistency between the similarities of communities as assessed by the PhyloChip and the GeoChip approaches. This suggests that functional genes detected in the environment are strongly linked to the composition of the local community as determined by phylogenetic marker genes.

For further mechanistic insight in the processes conducted by microbial communities, diagnostic microarrays should be combined with other methods. An interesting new development is the combination of *stable-isotope probing* (SIP) with microarray analysis (Dumont and Murrell 2005; Wagner *et al.* 2007). In SIP, a substrate (e.g. methanol), enriched with a stable isotope (e.g. ^{13}C) is added to an environmental sample and after a designated incubation time DNA is extracted. DNA from microorganisms that have taken up the substrate can be separated from other DNA by using CsCl density-gradient centrifugation. The diversity of this DNA can then be screened with a microarray, usually targeting the 16S rRNA gene. By focusing only on DNA in which ^{13}C is incorporated, a profile is obtained of a single functional guild, defined by the capacity to metabolize the substrate (DeLong 2001; Gray and Head 2001; Radajewski *et al.* 2003; Wellington *et al.* 2003).

Another promising, newly developed, technique goes under the name of isotope array (Adamcyk *et al.* 2003). As in SIP, a labelled substrate is added to a community, but the label is a radioisotope such as ^{14}C. rRNA is then extracted, labelled with a fluorescent label, and screened for 16S rRNA diversity using a phylochip. By scanning for both fluorescence and radioactivity one can single out species from a

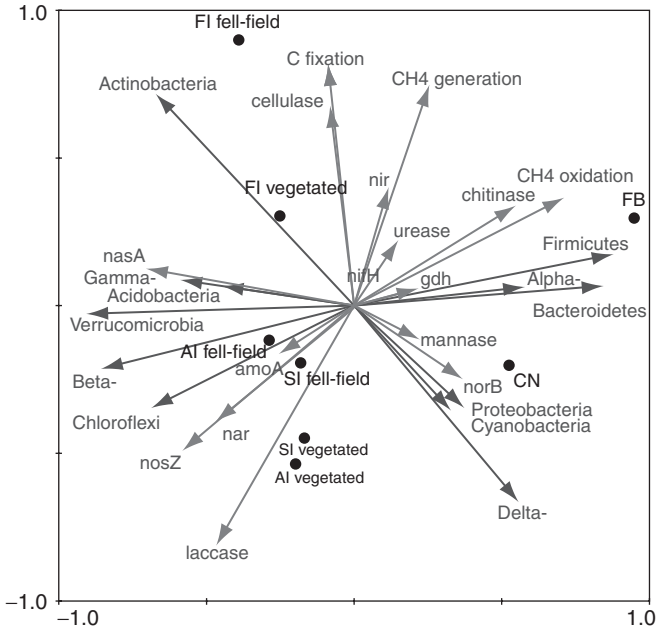


Figure 3.15 Multivariate analysis of microbial communities in a transect of Antarctic soil communities, assessed using the PhyloChip and the GeoChip microarray approaches in combination. The graph shows associations between specific taxonomic groups (Actinobacteria, Firmicutes, Alphaproteobacteria, etc.), and the abundance of functional genes, e.g. celullase, urease, chitinase, etc. Reproduced from Yergeau *et al.* (2009) by permission of the International Society for Microbial Ecology.

designated functional guild that have proved to be active because they have taken up the substrate. Adamcyk *et al.* (2003) used this approach to profile nitrifying communities in activated sludge from two different wastewater-treatment plants. Diversity of ammonia-oxidizing bacteria in the sludge was assessed from the incorporation of added bicarbonate (leading to radioactively labelled rRNA) combined with hybridization to ammonia-oxidizing bacterium-specific 16S rRNA gene probes (which is possible since ammonia-oxidizing bacteria constitute a monophyletic lineage). Such a combination of different molecular tools seems to be particularly promising.

3.4 Reconstruction of functions from environmental genomes

An exciting new development in microbial genomics is the exploration of communities by analysis of metagenomic libraries or by direct sequencing metagenomes. Fragments of DNA are isolated from the environment and cloned into vectors such as BACs or fosmids, followed by probing, sequencing, or screening for functions. If one is only interested in community surveys, not in demonstrating functionalities, the cloning step can be skipped and the environmental DNA sequenced directly using next-generation sequencing technology. This approach, designated *community genomics* or *metagenomics*, allows insight into microbial diversity, and it may lead to the discovery of new genes and functions, novel metabolic pathways, and previously unknown properties of microorganisms that cannot be cultured (Ball and Trevors 2002; Handelsman 2004; Riesenfeld *et al.* 2004a; Tiedje and Zhou 2004; Allen and Banfield 2005; DeLong 2005; Schleper *et al.* 2005). For ecological genomics it is important to realize that metagenomic analysis may lead to reconstruction of functions from genome sequences of organisms that have never been cultured. Comparative analysis of different environments has

demonstrated that metagenomes contain habitat-specific signatures that can be used for environmental diagnosis (Tringe *et al.* 2005). Sometimes genes found in metagenomes may be 'brought to life' in the laboratory by expressing the DNA segment in a suitable host.

A characteristic property of the metagenomics approach is that genes and functions are studied without consideration of the species from which the DNA derives. The metagenome of a habitat thus consists of the collective genomes of all organisms together. Of course, such an approach has its limitations, because a specific cell environment and the joint expression of several genes together in a delimited volume are often crucial to the function. In addition, the genomes of different species will differ in dynamics and responses to environmental change, and so the composition of a metagenome could be highly variable in time (DeLong 2001). Nevertheless, several important discoveries have been made by probing communities in this way, as we will see in the examples discussed below.

Two different approaches may be discerned for screening environmental genome libraries: *function-driven screening* and *sequence-driven screening* (Schloss and Handelsman 2003). In the first approach, the aim is to identify clones in a library that express a certain function, often one that has potential applications in medicine, agriculture, or ecology. For example, one may be interested in genes from biosynthetic pathways for antibiotics or genes associated with crucial links in biogeochemical cycles. Usually the frequency of active clones is quite low, so one needs a simple assay by which large numbers of clones in a library can be tested quickly. A clever, high-throughput, method was developed for this purpose by Uchiyama *et al.* (2005). The authors made use of the fact that many genes are induced by the substrate that they catabolize. A vector containing green fluorescent protein, suitable for shotgun cloning, was used with the effect that host cells with an insert carrying the promoter of the target gene expressed green fluorescent protein in the presence of the target substrate; these cells could then be sorted by an automated cell-sorting system.

In the second type of screening, sequence-driven screening, one uses hybridization probes to detect clones containing a desired known sequence. The probe can be a phylogenetic anchor, such as a 16S rRNA sequence, or a specific functional gene. Using next-generation sequencing, the tedious development of a clone library is not even necessary and the environmental DNA can be analysed directly. A variant to sequence-driven screening was proposed recently by Sebat *et al.* (2003), who screened a metagenomic library with a microarray. In their study, the microarray consisted of probes from a stable reference community, cloned in a cosmid library. Hybridizations with the metagenomic library were evidence of the presence of certain species. A great advantage of microarray-based screening of libraries is that many probes are used in parallel.

Library screening is usually followed by partially or completely sequencing the clones of interest. In addition, the library may also be sequenced from the start, without prior screening, if one is interested in all sequences. In this case the library is not prepared with large-insert cloning vectors such as BACs, but with small-insert plasmid vectors that are suitable for direct sequencing. By picking out clones at random, assembly of multiple complete genomes is then attempted. We will discuss examples of all these approaches below.

3.4.1 Marine community genomics

One of the most appealing results of community genomics is the discovery by Béjà *et al.* (2000a) of proteorhodopsin in the ocean. As we have seen above, proteorhodopsin is a retinal-dependent light-driven proton pump, which may support a photoheterotrophic lifestyle of marine bacteria. The widespread presence of this pathway in the carbon cycle was an unexpected outcome of genomics exploration. Crucial to the approach applied by Béjà *et al.* (2000a) was the construction of a large-insert BAC library after preparation of DNA using a special type of electrophoresis, *pulsed-field gel electrophoresis* (Béjà *et al.* 2000b). With this technique, applied to environmental DNA digests, it was possible to isolate high-molecular-mass DNA fragments up to several hundred kbp. The library was screened with 16S rRNA probes to survey the taxonomic diversity. Béjà

et al. (2000a) decided to sequence a 130 kbp genomic fragment from a clone in which the 16S probe had detected an rRNA sequence of an uncultivated member of marine Gammaproteobacteria, the SAR86 group. Sequencing the rest of the clone revealed an ORF for a rhodopsin-like protein called proteorhodopsin, which showed similarity with rhodopsin genes from extreme halophilic Archaea and the fungus *Neurospora crassa* (Fig. 3.16).

Rhodopsins act as transmembrane channels that can bind the chromophore retinal (a derivative of vitamin A) to become sensitive to light. Absorption of light energy by the protein–retinal complex leads to a series of conformational shifts, promoting the transport of ions across the cell membrane (Fig. 3.17). In the case of proton transporters, the outside surface of the cell membrane will become charged with protons and the resulting electrochemical membrane potential creates a motive force for another membrane-bound molecule, H^+-ATPase, to drive ATP synthesis. Three groups of rhodopsins are present in Archaea, one group acting as chloride pumps (halorhodopsins), another as proton pumps (bacteriorhodopsins), and the third as photosensory receptors (sensory rhodopsins). The last group of molecules is related to the opsin proteins found in eyes throughout the animal kingdom. In *N. crassa* a related opsin protein acts in the maintenance of circadian rhythmicity.

That the sequence found in the marine BAC clone represented a functional protein and not a pseudogene of some sort was proven by recombinant expression. *E. coli* cells, transfected with the rhodopsin sequence, expressed the protein and it was shown that a combination of retinal and yellow light triggered cross-membrane proton transport in *E. coli* cell suspensions (Fig. 3.16). Subsequent research using membrane preparations collected directly from seawater exposed to laser-flash photolysis demonstrated that similar photoactive molecules were very common in the environment (Béjà *et al.* 2001).

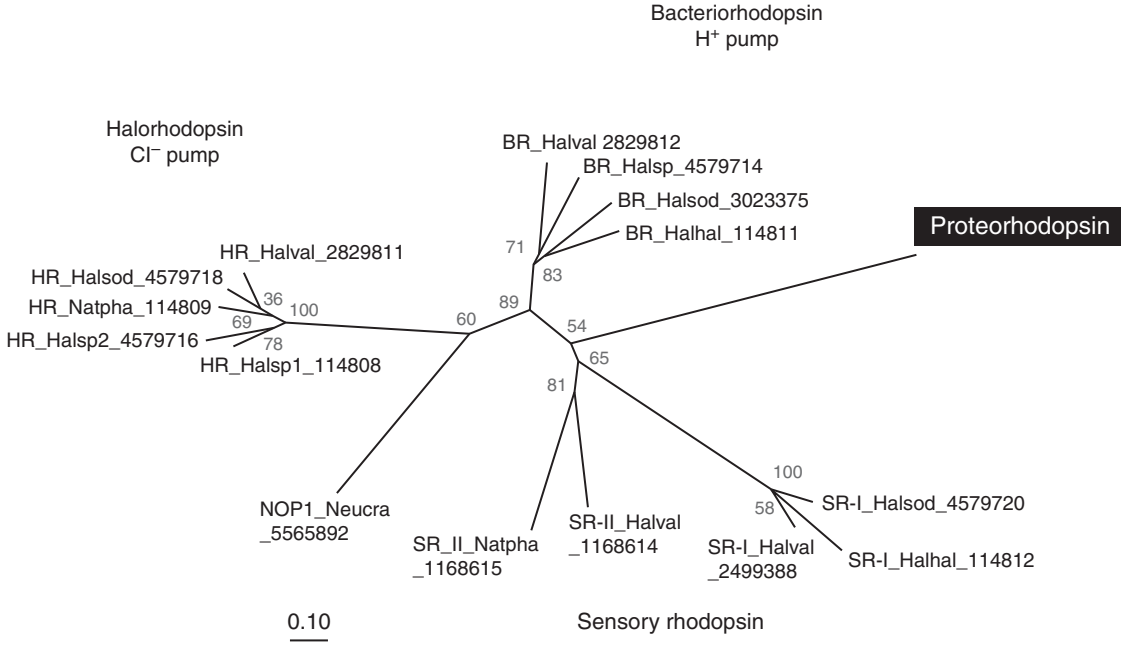


Figure 3.16 Unrooted phylogenetic tree of the proteorhodopsin sequence of an uncultured marine gammaproteobacterium found by Béjà *et al.* (2000a), aligned with rhodopsins in Archaea and the fungus *Neurospora crassa*. HR, halothodopsin, light-driven chloride pumps; BR, bacteriorhodopsin, light-driven proton pumps; SR, sensory rhodopsin; Halsod, *Halorubrum sodomense*; Halhal, *Halobacterium salinarum*; Halval, *Haloarcula vallismortis*; Natpha, *Natronomonas pharaonis*; Halsp, *Halobacterium* sp. (all Archaea); Neucra, *N. crassa* (Ascomycota). The scale bar indicates the proportion of amino acid difference. Reprinted with permission from Béjà *et al.* (2000a). Copyright 2000 AAAS.

The widespread occurrence of proteorhodopsins in the sea was confirmed by the '*Global Ocean Sampling*' (*GOS*) expedition (Venter *et al.* 2004; Rusch *et al.* 2007; Yooseph *et al.* 2007). The first phase of this project was focused on the Sargasso Sea off the coast of Bermuda; in the second phase the expedition explored Northwest Atlantic and Eastern Tropical Pacific ecosystems. From the Sargasso Sea DNA, a total of 1.36 Gbp of microbial DNA sequence was generated, from at least 1800 genomic species, including 148 previously unknown bacterial phylotypes and 1.2 million previously unknown genes. Some distinct groups of sequence scaffolds could be distinguished, one clearly related to a *Burkholderia* species, others to *Shewanella*, *Prochlorococcus*, and a SAR86 gammaproteobacterium. The presence of *Burkholderia*, a nutritionally versatile genus of Betaproteobacteria, was unexpected, because this genus was considered typical for terrestrial environments. Similarly, *Shewanella* is an abundant genus in aquatic, nutrient-rich environments. The presence of these organisms in the open ocean shows that they have a wider ecological amplitude than thought previously, or that there are nutrient-rich microhabitats (possibly associated with marine animals or anthropogenic waste) in which they may survive.

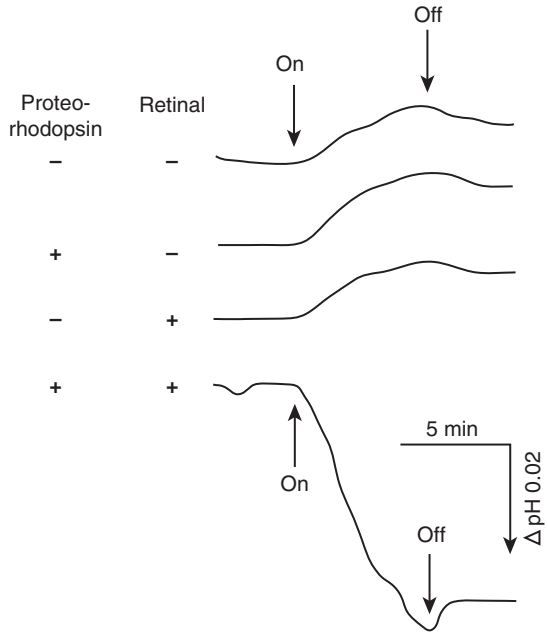


Figure 3.17 Diagram showing the pH change in a medium with cell suspensions of *E. coli* expressing a proteorhodopsin. In the presence of both the protein and retinal, an outward transport of protons occurs when the cells are exposed to yellow light (> 485 nm, indicated by On/Off), leading to a decrease in the pH of the medium. Reprinted with permission from Béjà *et al.* (2000a). Copyright 2000 AAAS.

In the metagenome of the Sargasso Sea Venter *et al.* (2004) found 782 different rhodopsin-like genes, which were classified into 13 distinct subfamilies. Four of these families consisted of the archaeal, fungal, and sensory rhodopsins mentioned above, but nine families were related to sequences from uncultured species, including seven only known from the Sargasso Sea samples. Analysis of scaffolds containing both a taxonomic marker (σ subunit of RNA polymerase) and a rhodopsin gene demonstrated that rhodopsins are not limited to the Gammaproteobacteria in which Béjà *et al.* (2000a) had first discovered them. For example, in one scaffold a rhodopsin was found together with a σ subunit RNA polymerase from the phylum Flavobacteria.

Table 3.7 provides an overview of some functional aspects of the Sargasso Sea sequence data. As an example, consider the presence of an ammonia monooxygenase gene sequence associated with an archaeal taxonomic marker, which indicates scope for archaeal nitrification in this environment. Previously, marine biologists had argued that nitrification in the ocean was hardly possible due to the sensitivity of chemoautotrophic bacteria to high levels of UV irradiation. Nitrification by Archaea would not be inhibited by UV light and this activity would be in accordance with the relatively high nitrite concentrations that are seen along with nitrate at certain times of the year in the Sargasso Sea. Many microbiologists have wondered why archaeal ammonia oxidation had not been noted before; that it was found by environmental DNA sequencing is often cited as an example of the 'discovery' aspect of ecological genomics. Soon after publication of the Sargasso Sea study, Könneke *et al.* (2005) isolated a crenarchaeote strain from a marine nitrifying enrichment culture that was shown to be capable of autotrophic ammonia oxidation. It was named *Nitrosopumilis maritimus*. Further analysis of

archaeal *amoA* sequences indicated that ammonia oxidizing archaea (AOA) were found in many different marine habitats, both water columns and sediments, and, in contrast to ammonia oxidizing bacteria (AOB), covered a great phylogenetic diversity (Francis *et al.* 2007).

In a commentary to the paper by Venter *et al.* (2004), Falkowksi and De Vargas (2004) remarked that the massive sequencing approach is reaching its limits when applied to community genomes. For example, despite the huge sequencing effort, only two nearly complete genomes (those of *Burkholderia* and *Shewanella*) could be reconstructed, and this could only be achieved by using already existing databases as a reference to support the assembly. The major part of the community is represented by rare organisms, and to obtain 95% coverage of these more than an order of magnitude of sequencing depth would be needed. However, the statistical analysis of Quince *et al.* (2008) indicates that for the data of the Global Ocean Sampling a fivefold increase in sampling effort would be sufficient to reach 90% coverage of the metagenome, that is, all genes from all taxa. With the extremely powerful next-generation sequencing technology, a complete enumeration of the microbial metagenome of surface oceanic samples seems to come within reach. However, this still excludes the larger cells of eukaryotes, which pose special problems due to the extremely large genomes of some representatives, such as coccolithophores (see Section 3.3.3).

The second phase of the GOS expedition produced nearly five times more information compared to the Sargasso Sea data: 6.3 billion base-pairs of DNA sequence, twice the size of the human genome (Gross 2007; Rusch *et al.* 2007; Yooseph *et al.* 2007; Kannan *et al.* 2007; Williamson *et al.* 2008). The value of this dataset is twofold: first, the sequence data are spatially structured, that is, samples were taken from different oceanic locations. This allows questions to be answered about the geographic factors that shape the diversity of oceanic microbial communities. Second, the data are so extremely numerous that they called for new bioinformatics methodology. Most reads failed to assemble in contigs, due to the immense diversity of sequence. Although a few microbial clades dominated the community, 85% of the assembled sequences and 57% of unassembled sequence were unique. To deal with such diversity, Rusch *et al.* (2007) applied a new assembly method, termed '*fragment recruitment*'. This method capitalizes on the completion of several hundred full-genome sequences of prokaryotes (around 600 at the time of the study). All reads were matched against

Table 3.7 List of some remarkable functional insights reconstructed from WGS sequencing of Sargasso Sea microbial DNA by Venter *et al.* (2004)

Genomic property	Functional relevance
782 different rhodopsin genes belonging to 13 protein families	Rhodopsin-mediated phototrophy is very common in oceanic bacterial plankton
Rhodopsin gene in scaffold bearing a taxonomic anchor from the Flavobacteria/Cytophaga group	Rhodopsin-mediated phototrophy distributed well outside the Proteobacteria
Ammonia monooxygenase in archaeal-associated scaffold	Oceanic nitrification not limited to the Bacteria; there are nitrifiers among the Archaea, and ammonia oxidation is not inhibited by UV light
Genes encoding phosphonate and high-affinity phosphate transporters; many genes responsible for utilization of pyrophosphates and polyphosphates	Versatile use of phosphorus compounds in oceanic environment to deal with severe phosphorus limitation
Gene homologous to umuCD DNA damage-induced DNA polymerase of E. coli found on plasmid	Resistance against UV damage by allowing DNA replication even when damaged by UV irradiation
Genes for arsenate, mercury, copper, and cadmium resistance found on plasmids	Possible role of oceanic microorganisms in trace-metal cycling in an oligotrophic environment
At least 50 bacteriophage gene groupings in scaffolds and 150 in singletons	High diversity of phages in oceanic bacterial community; significant fraction of bacteria infected

these reference genomes, which produced perfect alignments in 30% of the cases and less optimal alignments, indicating a distant evolutionary relationship, in 70% of the cases. Recruited reads are likely to come from microbes identical or closely related to the reference organism. Every genome tested recruited some reads, but five bacterial genera recruited the majority: *Prochlorococcus*, *Synechococcus*, *Pelagibacter*, *Shewanella*, and *Burkholderia*. The first three of these jointly represented about 50% of the recruited reads and 15% of all reads in the GOS dataset. Interestingly again a relatively high abundance of *Burkholderia* was found, a genus that was not expected to be so important in oceanic ecosystems before J.C. Venter published the Sargasso Sea data.

Another analysis of the GOS data aimed at identifying open reading frames. These ORFs were subjected to cluster analysis and extensive pairwise comparisons to generate around six million predicted proteins, at that time 1.8 times the number of protein clusters already in public databases (Yooseph *et al.* 2007). This enormous extension of protein sequence information led to some astonishing conclusions. For example, some protein families considered to be unique to eukaryotes were found to be prevalent in prokaryotes as well. To illustrate this point, Kannan *et al.* (2007) analysed the diversity of protein kinases (the microbial '*kinome*'). Prokaryotic protein kinases are histidine-aspartate protein kinases, which are distinct from the eukaryotic protein kinases families. However, analysis of the GOS data showed that proteins with a eukaryotic protein kinase-like domain are also prevalent in bacteria and most likely play an important role in these organisms (Kannan *et al.* 2007).

Expansion of the universe of protein families is also illustrated by the discovery of a new clade of *RuBisCo* large subunit. Ribulose 1–5 biphosphate carboxylase oxygenase can be said to be the most important enzyme on Earth because it is crucial for the fixation of carbon dioxide and the generation of all biomass. There are four forms of RuBisCo. The most common is Form I, an octameric structure, which is found in plants and algae but also in Proteobacteria and deep-sea bacteria. Form II is a dimeric structure and is found in several other bacteria, such as *Rhodospirilium*. Form III occurs in oligomeric structures and is found in Archaea, while form IV, also called RuBisCo-like protein (RLP), is found in bacteria that do not fix carbon. In addition to these clusters already known, analysis of the GOS data identified a new group of 65 RuBisCo sequences that did not cluster with any known RuBisCo and formed a distinct clade, a sister group of Form II in a supergroup also containing Form II and RLPs (Fig. 3.18). In the new proteins, some of the amino acids in the active site of RuBisCo were found to be mutated, which suggests that these proteins have evolved to adopt new catalytic functions, like the RLPs, and there might be more than one type of RuBisCo protein in some microorganisms. The fact that in this dataset a completely new clade could be discovered in one of the best investigated protein families, is another illustration of the fact that we have not yet fully explored the 'universe' of protein families.

The marine metagenomics studies have also shed new light on the diversity of viral communities (Breitbart *et al*. 2002; Williamson *et al.* 2008). Viruses represent a very important factor in biogeochemical

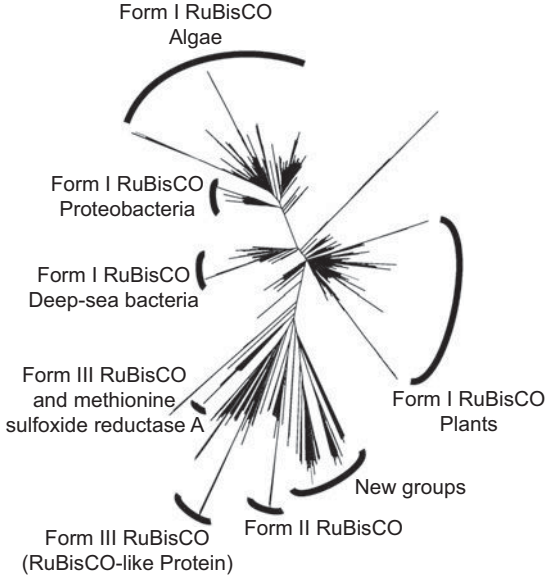


Figure 3.18 Phylogeny of RuBisCo large subunit, illustrating the diversity of clades after addition of the GOS data to sequences that were known earlier. Four clusters of RuBisCo are recognized and a completely new clade was found in the GOS data, related to Form II. Reproduced from Yooseph *et al.* (2007).

cycles and microbial biodiversity; by means of transduction they interfere with the genomes of their hosts, which for the marine ecosystem are mostly bacteria and algae. Viral activities are also considered important drivers of microbial community diversity by 'killing the winner' and promoting growth conditions of species with low abundance (Weinbauer and Rassoulzadegan 2004). Obtaining an overview of the biodiversity of viruses is difficult, because these organisms do not possess universal taxonomic anchors like the 16S rRNA gene in prokaryotes. The DNA polymerase gene *pol* can be used as a taxonomic marker for a subset of viruses (Short and Suttle 2002). New taxonomic systems are being developed that use all the genes in a viral genome to determine distances between species. This is elaborated in the Phage Proteomic Tree, a database and taxonomic algorithm for classifying bacteriophages (Edwards and Rohwer 2005). However, not many viral genomes have actually been sequenced completely (Paul *et al.* 2002). Another issue is that viruses cannot be cultured outside their hosts, which in the case of marine bacteria are themselves mostly uncultured. So, a metagenomic approach to surveying viral biodiversity seems very appropriate.

In the study of Breitbart *et al.* (2002) free-living viruses were collected by differential filtration and density-gradient centrifugation from surface sea water at two sites—Scripps Pier and Mission Bay—along the coast of California, USA. Special precautions must be taken when cloning viral DNA in a bacterial host, due to the presence of modified nucleotides and genes that could lyse the host. A total of 1934 sequences were obtained in a WGS approach, of which 70% showed no significant hits on sequences reported previously in GenBank. Among the remaining sequences no more than 34% were annotated in GenBank as viral sequences, and the rest were sequences of Archaea, Bacteria, and Eukarya, as well as mobile elements and repeat sequences. It appears that viral genomes carry a significant amount of DNA that originates from their hosts. About 83% of the viral sequences were related to bacteriophages, and these were classified further over the major groups of phages (Fig. 3.19). The viral community seemed to differ between the two sampling sites. Viral genomes at Scripps Pier were more 'bacterial' in origin, whereas viral genomes of Mission Bay had a more eukaryotic signature. Among the phage types, the Siphoviridae (λ-type phages) were more dominant at Mission Bay than at Scripps Pier.

Why the viral community should differ between two sampling stations and whether there is any ecological relevance in such differences remains uncertain. A possibility could be that viral community composition is a reflection of eukaryotic versus prokaryotic dominance of the plankton, for example due to algal blooms of variable composition. Analysis of the GOS data (Willamson *et al.* 2008) similarly yielded a high abundance of viral sequences. The prevalence of many physiological functions in viral DNA, and the fact that microorganisms can exchange DNA via viral-mediated lateral gene transfer, suggest that the total collection of viral DNA in the oceans is an important genetic reservoir generating microbial diversity.

3.4.2 The soil metagenome

Soil organisms have been most valuable sources of all kinds of natural products ever since the Scottish bacteriologist Alexander Fleming discovered in 1928 that the soil fungus *Penicillium* produced a substance that killed *Staphylococcus* bacteria. Many other products derived from microbial secondary metabolites have been used to develop antibiotics, anticancer drugs, fungicides, immunosuppressive agents, enzyme inhibitors, antiparasitic agents, herbicides, insecticides, and growth promoters. Over the years, most of the microorganisms that can be cultured in the laboratory have been examined thoroughly for the production of compounds with biological activity, and biotechnological investigators have gained the impression that the limits of what these organisms can yield in terms of valuable products has been reached. However, as we have seen above, any environment, and certainly the soil, holds a great diversity of uncultured microorganisms that remain to be investigated. With the advent of metagenomic recombinant DNA technology (Handelsman *et al.* 2002) it became technically feasible to screen soil microorganisms for new

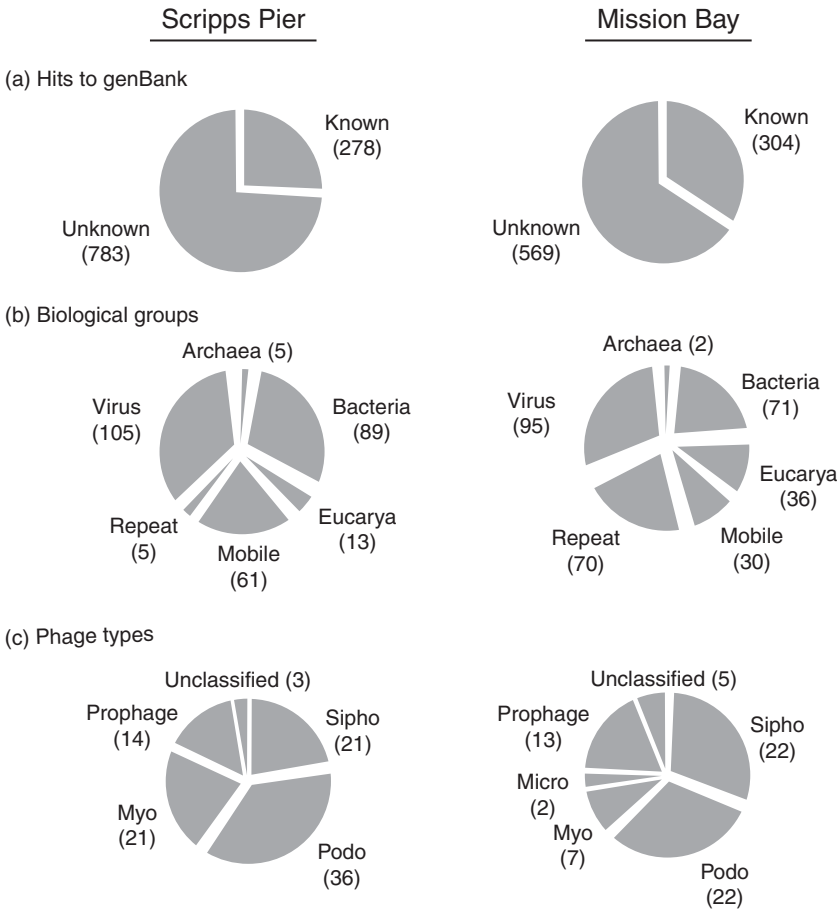


Figure 3.19 Overview of the content of viral genomes recovered from two marine coastal sampling stations, Scripps Pier and Mission Bay in California, USA. (a) A total of 1934 sequences were BLASTed to GenBank and 582 sequences produced a significant hit. (b) These sequences were classified according to sequence annotation, and 200 sequences were truly viral. (c) Among the 200 viral sequences 166 were from bacteriophages and these were classified according to the main phage families: Sipho, Siphoviridae (l-like); Podo, Podoviridae (T7-like); Myo, Myoviridae (T4-like); Micro, Microviridae (fX174-like). From Breitbart *et al.* (2002), by permission of the National Academy of Sciences of the United States of America.

functionalities without culturing them. This opportunity has raised great expectations and renewed interest in gene mining. The soil has been likened to Lady Bountiful (Rondon *et al*. 1999) and is considered a rich source for the discovery of novel natural products (Lorenz and Schleper 2002; Cowan *et al*. 2004; Daniel 2004, 2005).

How high is the probability of finding a novel product by functional screening of a metagenomic library? Gabor *et al*. (2004) explored this question in a theoretical way by analysing existing genome sequences of microorganisms. Assuming a random approach to expression cloning, it was argued that the probability of isolating an expressed gene in a metagenomic library depends on the mechanism by which that gene is expressed. The minimal requirements for gene expression in a host include the presence of a promoter for transcription and a ribosome-binding site for initiation of translation. Both of these sites must be recognized by the expression machinery of the host. If expression involves *trans*-acting factors from the host—for example special transcription factors, inducers, and so on—or if modifying enzymes are necessary for the gene

product to become functional, the situation becomes much more complicated. Calculations were made for three modes of expression to estimate the number of clones that would have to be screened before a target gene was recovered with a probability of greater than 90%. For the most simple case, independent expression, the expected number of clones was found to depend on the size of the insert and decreased to around 3000 with an insert size increasing to 100 kbp. It was also estimated that 40% of the genes can be found in this way. So, if a metagenomic screening effort covers a library of several thousand clones, each with an insert of 100 kbp, there is a fair chance that a designated gene will be found.

A pioneering study in soil metagenomics was the work of Rondon *et al.* (2000). These authors developed BAC libraries with DNA isolated from agricultural soil in Wisconsin, USA. The largest library had 24 576 clones with an average insert size of 44.5 kbp, whereas 10% of the clones had an insert size of between 70 and 80 kb. It was estimated that this library contained 1000 Mbp of DNA; given an average density in microbial genomes of one gene per 1000 bp, about 1 million genes were expected to be present in the library. Screening of the library for biological activities employed a variety of strategies. For example, to find clones expressing the enzyme cellulase, plates with the host cells were overlaid with agar containing brilliant red hydroxyethyl cellulose, and a yellow halo around the colony was taken as an indicator of cellulase activity. In this way, a great variety of enzymatic activities were screened, including β-lactamase, keratinase, chitinase, and amylase.

A remarkable discovery coming from the metagenomic library screening conducted by Rondon *et al.* (2000) concerned a clone that had antibacterial activity against *Bacillus subtilis* and *Staphylococcus aureus*, but not against *E. coli*. The clone in which this activity was found was sequenced completely and it appeared to contain 29 ORFs, including a cluster of 8 genes associated with phosphate transport. This showed that it is possible for BAC clones to contain complete, intact, operons. Fourteen of the twenty-nine ORFs could not be assigned a function. Using transposon mutagenesis, the genes were mutated to see which one was responsible for the antibacterial properties of the clone. Finally a single candidate gene was identified, of which the predicted amino acid sequence had several repeat units (Fig. 3.20). The authors also considered the *hydrophobicity profile* of this molecule. This is a plot of scores along the sequence, in which each amino acid is given a number indicating the degree of hydrophobicity (see Lesk 2002). The profile of the unknown protein showed seven peaks, which is indicative of a membrane pump (amino acids anchored in the membrane have a high score for hydrophobicity if they are to be embedded stably in the lipophilic environment of the membrane). Recombinant expression of the protein confirmed that it conferred antibacterial activity to the host; however, the protein itself was not active. The sequence shows significant homology with a gene in the genome of *Bdellovibrio bacteriovorus*, a deltaproteobacterium whose genomic sequence was presented by Rendulic *et al*. (2004). *Bdellovibrio* is a predator of other bacteria and its genome is considered a valuable reservoir of antimicrobial substances. The nature of the possibly novel mechanism of antibacterial activity in the soil metagenomic library remains unknown to date (M. Rondon, personal communication).

Functional metagenomic screening exercises similar to Rondon *et al.* (2000) have been applied by many other authors (Henne *et al.* 2000; Gillespie *et al.* 2002; Knietsch *et al.* 2003; Voget *et al.* 2003; Piel *et al.* 2004; Riesenfeld *et al.* 2004b). These studies were all motivated by the search for products with a medical or biotechnological relevance. Metagenomic profiling for ecological functions (e.g. crucial links in biogeochemical cycles, organic matter decomposition, and allelochemical production) has not yet been attempted.

Important insights into soil microbial communities have also come from sequence-driven screening of soil metagenomic libraries. Quaiser *et al.* (2002) were interested in the Crenarchaeota, a group of mesophilic Archaea that are frequently detected in soil communities; some of them are especially common in the rhizosphere. The Crenarchaeota are only known from their 16S rRNA sequences and have never been brought into pure culture. Quaiser *et al.*

(a)

MSFMKRFFCSCLTVAVILTACFSAAAQSE**GTLD**VSFNTTGVRYEDFGGAD
DKAMAVAV*QLDGKIVSVG*SSEVSGSGIDEAVVRYNSD**GTLD**SSFGTGGKV
TTAIGPGTSSDIAYSVVI*QSDGKIVVAG*SAAGISGTETDFAIVRYNAN**GT**
LDTSFGGTGKVTTPFGVATSADVANSVAL*QADGNIVPAG*YADDGSGADFA
LARYNTN**GSLD**ASFDTFGKTTTAIGAGTLGDFAQAVAI*QSDGKIVAAG*WT
EAASGLSIDFALARYNTN**GSLD**ATFDGDGKVITTVGSSTTFDLANAVLV*Q*
*ADGKIVAGG*FSDSLSSGADEALVRYNTN**GSLD**TSFDTDGIVITAIGPGTY
FDIAKAIVL*QPDGKIIAAG*YTDDLLVGFPSTDLALARYNVD**GSLD**TSFNA
DGKATIDLGGTEIINGAAIYAGNRIVVAGSSASNFLTARIWIATLVTAAP
VTVSGRITDERGRALKGVSVTLTDQDGVSSVASTNGFGYYRFTRVESGGT
HFLHATDRGYTFAPPVRIVDTKSDVSDADFVGTKQKGKPNSTR

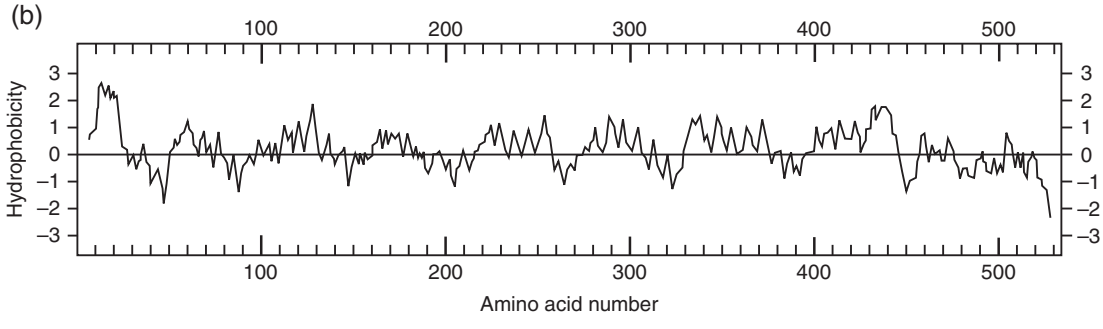


Figure 3.20 (a) Predicted amino acid sequence of a protein from an unknown soil organism, responsible for conferring antibacterial activity when expressed in *E. coli*. Amino acids are indicated by their standard single-letter codes. The ORF was discovered after screening a metagenomic BAC library made from DNA extracted from an agricultural soil. Repeated units are shown underlined, in bold, or in italic. (b) Hydrophobicity profile of the ORF, showing seven clusters of hydrophobic amino acids (peaks above the line), which is indicative of a membrane protein. The horizontal axis shows amino acids numbered from N- to C-terminus and the vertical axis is a dimensionless score for the hydrophobicity of each amino acid. For details see the text. After Rondon *et al.* (2000), by permission of the American Society for Microbiology.

(2002) screened a fosmid metagenomic library developed from DNA extracted from a calcareous grassland near Darmstadt, Germany, using a 16S rRNA probe. A clone of 33 925 bp which contained a crenarchaeotic 16S rRNA gene was sequenced entirely. It contained 17 ORFs, of which 5 could not be assigned a function. An overview of the organization of the clone is given in Fig. 3.21.

Among several others, an interesting cluster of genes was encoded in the crenarchaeote clone, which showed high similarity to the *fixABCX* gene cluster of nitrogen-fixing bacteria. In a phylogenetic analysis, the *fixA* gene of this operon clustered with *fixA* genes from two other Archaea, both hyperthermophilic crenarchaeotes. Together they seem to form a separate group which goes back to

the ancestor of the Crenarchaeota, rather than to other Archaea or Bacteria via lateral gene transfer. As mentioned in Section 3.3.1, *fix* genes encode electron-transport flavoproteins and are coexpressed with the nitrogenase (*nif*) gene cluster in nitrogen-fixing bacteria. Presumably electron-transport flavoproteins provide the electron transport to the nitrogenase complex. In the symbiotic plasmid of *Rhizobium etli* the *fix* genes are tightly associated with the *nif* cluster (González *et al.* 2003). The *fixABCX* operon is highly conserved in diazotrophs as well as in a wide variety of other bacterial and archaeal species; however, its function remains unknown in non-nitrogen fixing species. In *E. coli* *fix* genes are related to the carnitine pathway, a transport system of long-chain fatty acids into the mitochondrion, prior to fatty acid oxidation. None of the obligately aerobic Archaea that contain the *fix* genes are capable of nitrogen fixation; nitrogen fixation in Archaea is limited to strictly anaerobic euryarchaeotes such as *Methanosarcina* and *Methanococcus*. The intriguing presence of the *fix* gene cluster in Crenarchaeota cannot yet be evaluated from a functional point of view. Several other genes have been recovered from the same metagenomic libraries, some of which are unique to mesophilic Crenarchaeota; however, specific physiological traits have not yet been identified (Treusch *et al*. 2004).

Another group of microorganisms that has enjoyed interest from soil metagenomics research are the Acidobacteria, a relatively unknown phylum of the bacterial kingdom until the work of Quaiser *et al.* (2003). A soil metagenomic library appeared to contain 15 clones from this group. Six of these clones were sequenced and several genes were identified. Phylogenies on the basis of the purine-biosynthesis gene *purF* were consistent with 16S rRNA-based phylogenies and showed that the Acidobacteria should be considered a separate phylum of the Bacteria, sharing a common ancestor with the Proteobacteria (Fig. 3.22). The analysis of the metagenomic clones also revealed many deduced protein sequences, but only few gave hints of specific metabolic traits in Acidobacteria. After this study, many authors have confirmed that Acidobacteria are among the most abundant microorganisms in soil ecosystems (Liles *et al.* 2003; Janssen 2006; Fullthorpe *et al.* 2008; Lauber *et al.* 2009; Kielak *et al.* 2009). Acidobacteria make up an average of 20%, sometimes more than 40%, of soil bacterial communities and they comprise at least eight different phylogenetic lineages (Janssen 2006). Recently, the complete genomes of three acidobacterial strains were sequenced and this provided some insights into the ecology of this abundant phylum (Ward *et al.* 2009). The suggestion from the genome analysis, combined with available culture experience, is that within the bacterial kingdom Acidobacteria represent typically K-selected organisms: they are long-lived, divide slowly, exhibit a low metabolic rate, and are able to grow under low-nutrient conditions. This may partly explain their abundance in so many different soil ecosystems.

Another remarkable finding of soil metagenomics studies was the discovery of archaeal nitrification. We have seen in Section 3.4.1 that archaeal *amoA*

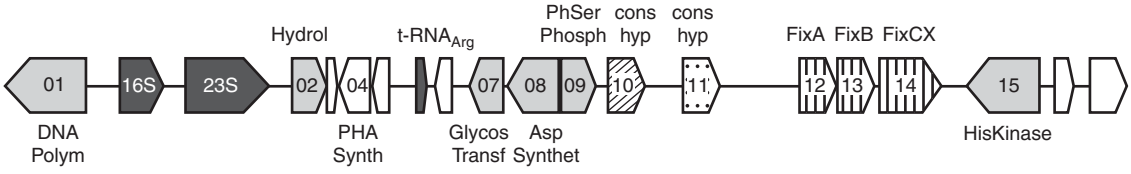


Figure 3.21 Schematic representation of a 34 kbp archaeal fosmid clone recovered from a calcareous grassland near Darmstadt, Germany. Different shadings indicate the most significant phylogenetic affinity of putative protein-coding genes to the Archaea (diagonal stripes), Bacteria (dots), Archaea and Bacteria (vertical stripes), or Archaea, Bacteria, and Eukarya (light grey). Genes without homology assignment are white. 01, Family B DNA polymerase; 02, α/β hydrolase; 04, polyhydroxyalkanoate synthase; 07, glycosyl transferase group 1; 08, asparagine synthetase; 9, phosphoserine phosphatase; 10, conserved hypothetical protein; 11, transmembrane protein; 12–14, *fixABCX*; 15, sensory transduction histidine kinase. From Quaiser *et al.* (2002), by permission of Blackwell Science.

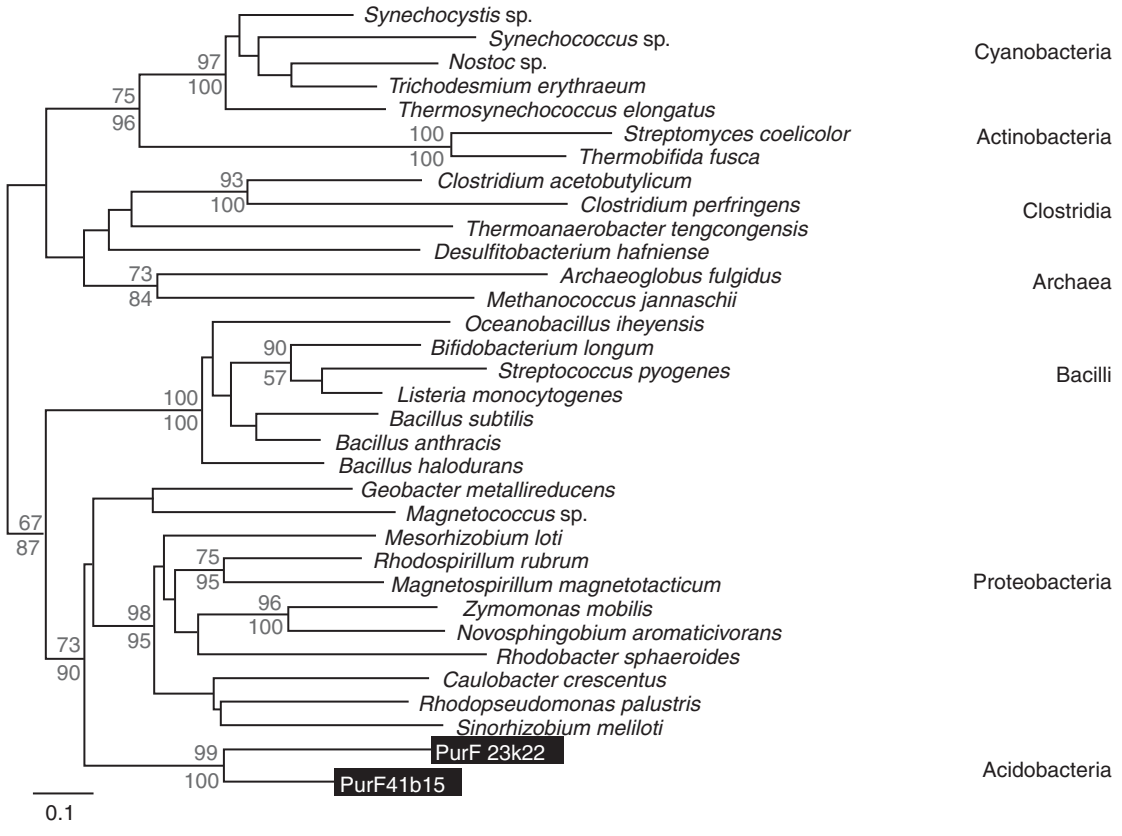


Figure 3.22 Phylogenetic analysis using the neighbour-joining algorithm of *purF* (amidophosphoribosyl transferase, a gene in the purine biosynthesis) sequences of several prokaryotes. Numbers on the nodes are bootstrap values. The position of two sequences from Acidobacteria, found in a metagenomics library of grassland soil DNA, demonstrates that this group should be considered a separate lineage within the bacterial kingdom. The scale bar indicates the proportion of amino acid difference. From Quaiser *et al.* (2003), by permission of Blackwell Science.

genes were first discovered in marine ecosystems (Venter *et al.* 2004; Könneke *et al.* 2005; Francis *et al.* 2007). However, soon thereafter crenarchaeote ammonia monooxygenase genes were also found in soil ecosystems. Treusch *et al.* (2006) identified a nitrite reductase and fragments of *amoA* and methane monooxygenase genes in a metagenomic library of a sandy grassland soil. Leininger *et al.* (2006) subsequently screened twelve different soils for the presence of bacterial and crenarchaeote *amoA* genes. The astonishing conclusion was that ammonia oxidizing Archaea (AOA) are more dominant than ammonia oxidizing bacteria (AOB) (1.5 to 232 more crenarchaeote *amoA* copies compared to bacteria). A chemical compound used as a marker for Crenarchaeota, crenarchaeol, was highly correlated with the number of crenarchaeote *amoA* copies. It is now generally agreed that in many soils AOA are more abundant than AOB. Whether these groups occupy different ecological niches in soil or live in competition with each other remains unknown to date. It is also unclear how the differential abundance of AOB and AOA translates to their share in the overall nitrification activity of a soil. The *amoA* genes of Archaea are clearly distinct from those of Bacteria and this might indicate a difference in their biochemical functionality (Cavicchioli *et al.* 2006).

Metagenomic studies of soil ecosystems have reached a new phase with the application of next generation sequencing methods. Projects are launched to sequence the complete metagenome of a reference soil (Vogel *et al.* 2009). A special chal-

lenge will be how to best interrogate DNA libraries in order to find the needles in the metagenomic haystack (Kowalchuk *et al.* 2007).

3.4.3 Genomes in extreme environments

Extreme environments usually harbour only a few species that have developed special characteristics to survive conditions that are lethal to most other organisms. The small and specialized communities found in extreme environments are interesting subjects for ecological research. Since such communities involve only a few species, interactions between them and with the environment remain tractable, and each species fills a complete niche. Some extreme environments are considered models for extraterrestrial conditions and the study of such environments is motivated by the fact that microorganisms may leave biosignatures in rocks when they become fossilized (Walker *et al*. 2005). Extreme environments are also a rewarding object of study in genomics. Among the habitats investigated by ecological genomicists are polar deserts, deep-sea hydrothermal vents, and acid-mine drainage.

One of the most exciting studies in metagenomics of extreme environments is that of Tyson *et al.* (2004), who investigated the metabolism of a small community of prokaryotes in an acid-mine drainage biofilm by means of the near-complete reconstruction of their genomes. *Acid-mine drainage* is a common environmental problem occurring in many coal and iron mines when pyrite (FeS_2) comes into contact with the air. This initially leads to the slow chemical oxidation of sulphide to sulphate and lowering of the pH. As we have seen above (Section 3.3) under low pH conditions it becomes profitable for iron-oxidizing microorganisms such as *Ac. ferrooxidans* to gain energy from the oxidation of Fe^{2+} to Fe^{3+} Since this reaction allows only a little energy gain large amounts of Fe^{2+} have to be oxidized. The resulting Fe^{3+} ions may react back with pyrite to oxidize the metal sulphide bond, leading to more sulphate and a further lowering of the pH. Once the cycle of pyrite oxidation has started and acidophilic microorganisms have established themselves, the reaction tends to propagate: Fe^{3+} is regenerated and large quantities of acid are produced. The overall reaction is:

$$FeS_2 + 14Fe^{3+} + 8H_2O \rightarrow 15Fe^{2+} + 2SO_4^{2-} + 16H^+$$

Acid water and reduced iron will leach from the system and when this comes into contact with neutral surface water Fe^{2+} oxidizes spontaneously to Fe^{3+} and a precipitate of $Fe(OH)_3$ is formed. Microorganisms play a central role in the generation of acid-mine drainage and they may develop a pink biofilm growing under very acid conditions (down to pH 0.5).

Tyson *et al*. (2004) developed a small-insert plasmid library from an acid-mine drainage biofilm in the Richmond Mine, California, USA, and sequenced the library to a depth of 10. The reads could be assembled into near-complete genomes of two major members of the community, a bacterium from the *Leptospirillum* group II, related to L. *ferrooxidans* (Nitrospirae), and a previously unknown, uncultured archaeon designated *Ferroplasma* type II (Euryarchaeota, Thermoplasmatales). Partial genome sequences were obtained for another *Leptospirillum* species (*Leptospirillum* group III), a species of *Sulfobacillus*, and a *Ferroplasma* type I. The sequences of *Leptospirillum* type II, although arguably from different individuals, were nearly all the same; the average rate of nucleotide polymorphism was no greater than 0.08%. In the Ferroplasma type II genome, there was however a fair degree of polymorphism (2.2%), and these polymorphisms had a peculiar distribution in which 'hot spots' were interspersed with longer homogeneous regions, suggesting that the *Ferroplasma* type II genome should be considered a mosaic of at least three different strains, combined with each other through recombination. Although this detailed reconstruction of population-genetic structure from an environmental genome was already unique, Tyson *et al*. (2004) were also able to identify a significant number of functional genes in the two dominant species and link these to the geochemical processes in the biofilm (Fig. 3.23).

In the genome of *Leptospirillum* group II 63% of the ORFs could be assigned a function, while in *Ferroplasma* type II the corresponding figure was 58%. Both species are iron oxidizers but seem to use different systems to gain energy from this reaction. *Leptospirillum* has a system comparable to the one discussed in Section 3.3.3 for *Ac. ferrooxidans*, in

which a periplasmic cytochrome oxidizes Fe^{2+} and transfers the electron to a membrane-bound electron-transport chain (Fig. 3.23). However, the genome of *Leptospirillum* does not encode rusticyanin, but rather a 'red cytochrome' with the same function. In *Ferroplasma* (which lacks a cell wall), a putative membrane-bound 'blue copper protein' conducts the initial oxidation of Fe^{2+}. This copper-containing protein shared sequence characteristics with both rusticyanin of *Ac. ferrooxidans* and sulphocyanin of the sulphur-oxidizing archaeon *Su. solfataricus* (see Fig. 3.10). The precise relationships between these electron-transport proteins remain to be elucidated, but from the data at hand it can already be concluded that different species may use different proteins for similar functions, and that there are homologies between the proteins used for iron and sulphur oxidation. The situation seems to be comparable to the electron-shuttling proteins used in iron reduction discussed above: a comparison between *Geobacter*, *Shewanella*, and *Desulfovibrio* showed that an evolutionarily diverse set of proteins may be used for similar functions in different species (see Section 3.3).

The genome of *Leptospirillum* encoded all the genes of the Calvin cycle and so *Leptospirillum* is probably a chemoautotroph. The lack of genes from the Calvin cycle in *Ferroplasma* and the presence of a large number of ATP-binding cassette (ABC)-type sugar transporters suggests that this species is heterotrophic. Interestingly, neither of the two species had genes for nitrogen fixation, but these were found in a third, incompletely reconstructed species, *Leptospirillum* group III. It is remarkable that the latter species, which has a much lower abundance than *Leptospirillum* group II, plays such a key role in the community. Eukaryotic heterotrophs have also been identified in the acid-mine drainage communities, including two distinct groups of ascomycete fungi (Baker *et al*. 2004). Thus the community as a whole is a self-sustaining unit, independent of organic input from the surface; it can fix nitrogen and carbon from the air and uses iron oxidation as a prime mechanism for energy generation. The composition of the community with a few dominant genome types is consistent with the extreme character of the environment, providing only a few niches.

Obviously, microorganisms living in the extremely acid conditions of acid-mine drainage must be very tolerant to low pH. One aspect of tolerance is a high expression of proteins involved in combating oxidative stress and protein denaturation. We will see in Chapter 5 how such defence mechanisms provide protection against cellular stress. A proteomics study of the Richmond Mine acid-mine drainage biofilm confirmed that stress-defence proteins were a very dominant feature of the community proteome (Ram *et al.* 2005). Another key feature of acid tolerance is the nature of the cell membrane. A cell membrane with very low proton permeability is found in the acidothermophilic archaeon *Picrophilus torridus*, whose genome sequence was presented by Fütterer *et al.* (2004). The genome of *P. torridus* encodes an extraordinary large number of secondary transport systems and this suggests that the steep proton gradient across the cell membrane is used extensively for transport of metabolites. Acidophilic microorganisms must also be tolerant to extremely high levels of dissolved metal and metalloid ions, which reach concentrations in the upper millimolar range. Dopson *et al.* (2003) reviewed the various mechanisms by which microorganisms can achieve tolerance to As, Cu, Zn, Cd, Ni, Hg, and other metals. Most of the tolerance mechanisms in acidophilic microorganisms involve P-type ATPase efflux pumps, sometimes combined with reductases (in the case of Hg and As). The locus conferring As resistance has been particularly well studied. The *ars* operon encodes a reductase, an efflux pump, and a regulatory protein. A genome-wide inventory of heavy metal-tolerance genes was made for the non-acidophilic species *Pseudomonas putida*, the genome of which was sequenced in 2003 (Cánovas *et al*. 2003). No fewer than 61 ORFs were found that were likely to be involved in metal(loid) tolerance or homeostasis and some systems appeared to be duplicated. Loci were found for metal-specific uptake pumps, metal-reduction enzymes, chelating proteins, and metal-efflux pumps as well as several transcriptional regulators. Extrapolation of this machinery to acidophilic species may not be warranted, however, because the extreme acid environment poses additional requirements beyond 'normal' metal tolerance.

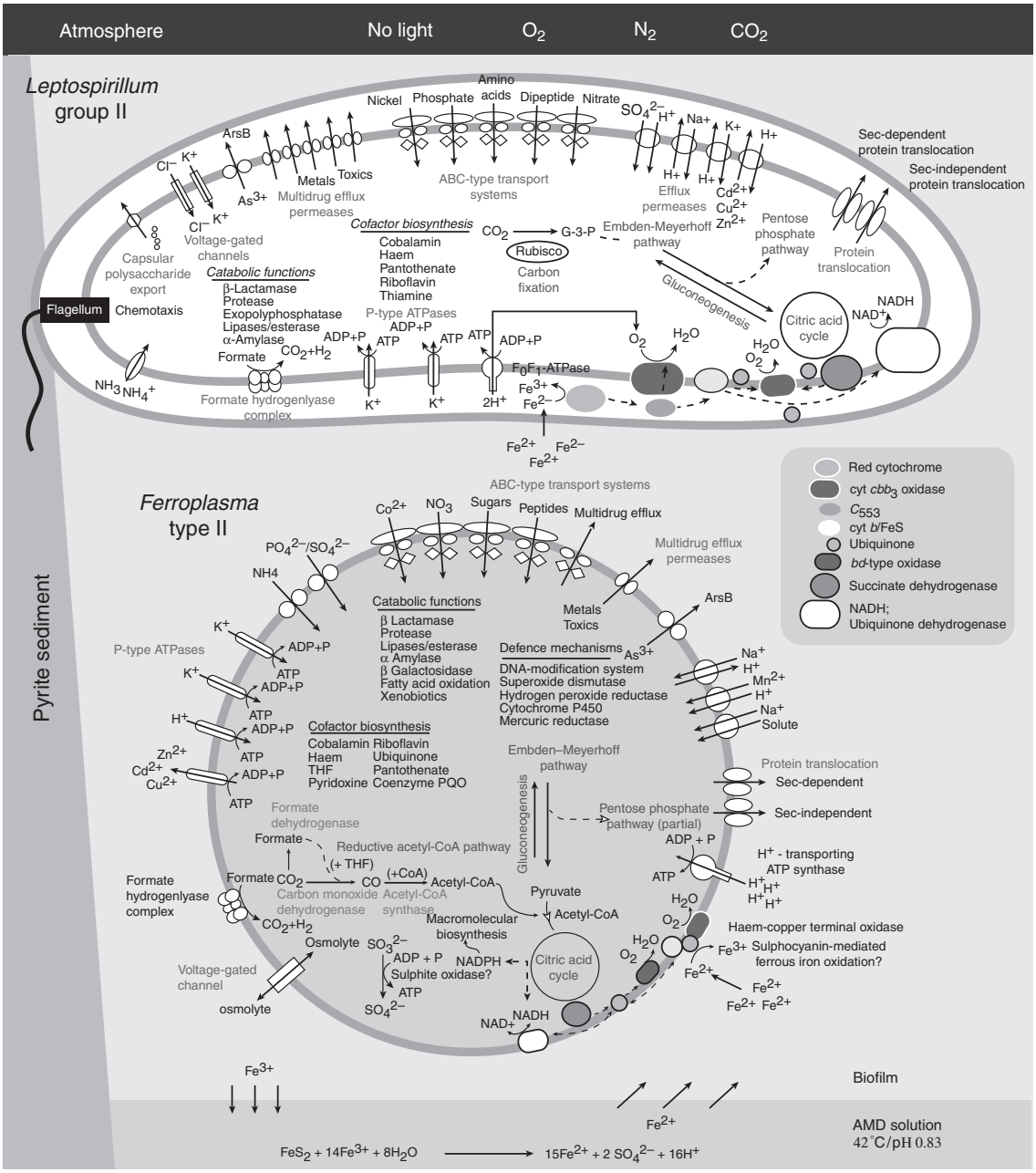


Figure 3.23 Cell metabolism of the genomes of *Leptospirillum* group II and *Ferroplasma* type II in an acid-mine drainage (AMD) biofilm with pyrite sediment and an acid-mine drainage solution of pH 0.83. Note the inner and outer membrane of the Gram-negative bacterium *Leptospirillum* and the presence of iron-oxidizing red cytochrome in the periplasm, coupled to a short electron-transport chain donating electrons to O_2. In the archaeon *Ferroplasma* type II, which lacks a cell wall, a membrane-bound sulphocyanin-mediated iron-oxidation system is assumed to be present. Both species have a complete citric acid cycle and several catabolic functions. *Leptospirillum* also has a complete Calvin cycle for CO_2 fixation, while *Ferroplasma* has plenty of ABC-type transporters for sugar uptake. The overall reaction combining microbial Fe^{2+} oxidation with reduction of Fe^{3+} by reaction with pyrite leading to acid generation is shown below. THF, tetrahydrofolate. After Tyson *et al*. (2004) by permission of Nature Publishing Group.

Another extreme habitat investigated by genome researchers are *hydrothermal vent systems* on the ocean floors. The sequencing of the first archaeon in 1996, *Me. jannaschii*, illustrates the interest in this environment. The interest is especially strong because of the antiquity and evolutionary significance of these systems, allowing inferences about the very first microbial pathways. Hydrothermal vents on the ocean floor are places where the Earth's crust drifts apart, allowing seawater to mix with hot minerals, leading to spurts of fluid at rates of 1–2 m/s with temperatures of 270–380 °C. A range of specialized, thermophilic Archaea and Bacteria have adapted to the high temperatures in these systems with growth optima between 60 (thermophiles) and 105 °C (hyperthermophiles). The energy generated by these microbial communities is derived from inorganic and geothermal sources (sulphide, H_2, reduced metals, CO_2, CH_4; Reysenbach and Shock 2002). The main metabolic strategies are methanogenesis, sulphur reduction (see Section 3.3), and the so-called knallgas reaction ($O_2 + 2H_2 \rightarrow 2H_2O$). Culture-independent screening using 16S rRNA probes has shown the presence of several Archaea, including a new phylum, the Korarchaeota, only discovered in 1999 in a hot spring at Yellowstone National Park, USA, the Obsidian Pool. Another new archaeal phylum, the Nanoarchaeota, was reported in 2002 (Huber *et al.* 2002). This phylum is represented by tiny cells that live attached to a sulphur-reducing archaeon of the genus *Ignicoccus*. They appeared to possess a 16S rRNA gene that has many base-pair substitutions in regions of the gene considered to be universal; for this reason the group was not found previously when using regular 16S PCR primers or fluorescent probes.

Figure 3.24 Picture of *Alvinella pompejana*, a polychaete worm occurring near hydrothermal vent ecosystems on the ocean floor, carrying a dense layer of symbiontic bacteria on its dorsal surface. The Pompeii worm is the most thermotolerant animal known. Courtesy of the University of Delaware.

Surprisingly, thermal-vent ecosystems are also colonized by animals. A peculiar example is the polychaete worm, *Alvinella pompejana* (Pompeii worm), that is extremely tolerant to high temperature (Fig. 3.24). With a body temperature that can reach 80 °C, this species is the most thermotolerant animal known. Population-genetic studies have shown that, despite the island-like structure of their habitat, hydrothermal polychaete worms are able to disperse via larval stages, and genetic differentiation is determined by hydrography and topography of the ocean (Hurtado *et al.* 2004). The dorsal surface of *Al. pompejana* is covered with bacteria of the group Epsilonproteobacteria, which are considered symbionts, providing the worm with organic material. Campbell *et al.* (2003) developed a fosmid library of the symbiontic bacteria and showed that it contained two enzymes from the reversed citric acid cycle, a system used for CO_2 fixation in green sulphur bacteria in lieu of the Calvin cycle (see Table 3.6). Therefore it is likely that these symbionts are chemoautotrophs, fixing CO_2 in the same way as green sulphur bacteria.

Similar symbiotic relationships occur in the Vestimentifera, a peculiar group of tube-living marine animals, formerly classified as a separate animal phylum (Pogonophora), but now considered an order of the Annelida on the basis of molecular data. Vestimentiferans lack a normal intestinal tract but have developed a specialized tissue, the *trophosome*, in which symbiontic chemolithotroph (sulphur-oxidizing) bacteria live, nourishing the animal. Molecular investigations of these symbionts

living in different hosts and different geographical sites have shown that the transmission from one generation to another is mainly horizontal; that is, the larvae acquire the symbionts from their environment, prior to attachment (Di Meo *et al*. 2000).

Community genomic studies in extreme environments are likely to deliver more surprises when complete metagenomic libraries have been sequenced. We will then reach a position in which complete metagenomes of different ecosystems

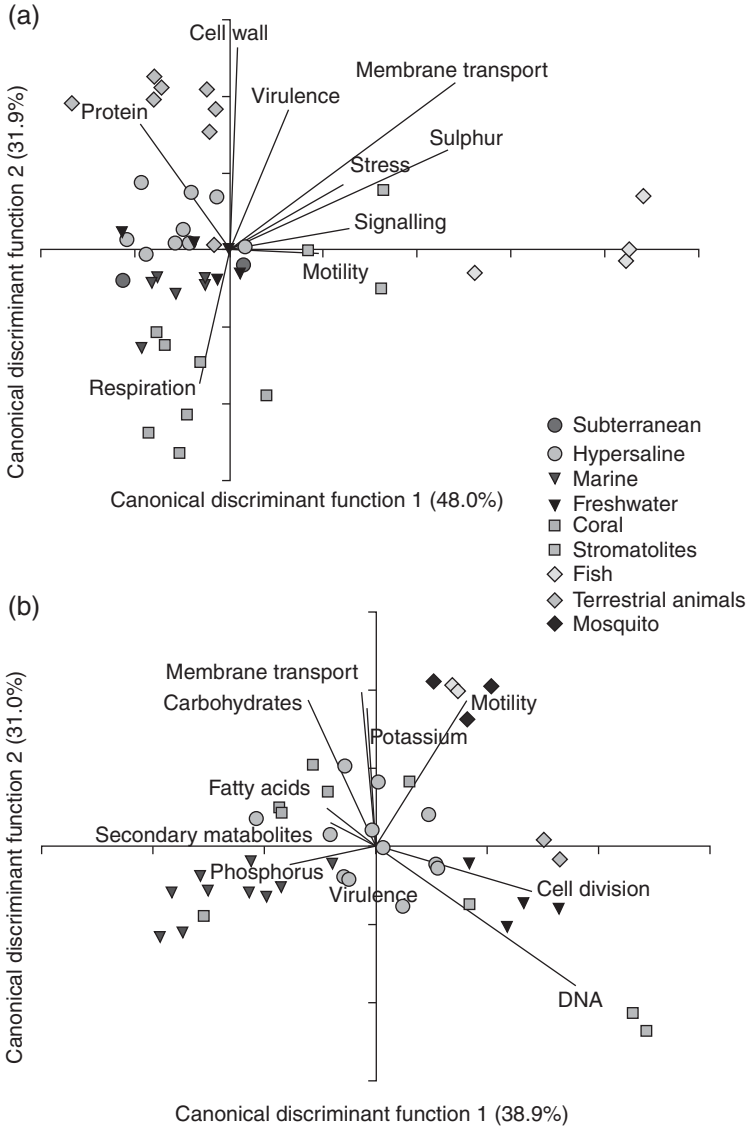


Figure 3.25 Biplot of a canonical discriminant analysis applied to (a) microbial, and (b) viral metagenomes from nine different habitats, indicated with different symbols. The multivariate analysis explains 80% of the variation in microbial communities and 70% in the viral communities. The length of each vector indicates the influential strength that a specific metabolic category of genes has on the separation of the nine metagenomes. Vectors pointing in opposite direction have opposing effects on gene abundances. Reproduced from Dinsdale *et al.* (2008), with permission from Nature Publishing Group.

can be compared. A glimpse of the insights that can be achieved by such a comparative meta-metagenomics study is provided by Dinsdale *et al.* (2008). These authors compared 15 million pyrosequence reads obtained from DNA samples of nine different biomes: a subterranean mine, a hypersaline pond, a marine environment, a freshwater environment, a coral-associated microbial community, an aquaculture fish-associated community, a terrestrial animal-associated community, and mosquito-associated microbes. The metagenomes were compared for the presence of genes representative of specific metabolic pathways and biochemical functions. A summary of the analysis is given in Fig. 3.25 in the form of a biplot obtained by canonical discriminant analysis. In this plot the gene abundances classified by various functional categories are plotted as vectors that discriminate the metagenomes from each other. Each habitat can be characterized by a specific metabolic profile. For example, the genome of the coral-associated microbial community has a relatively high percentage of genes falling into the metabolic category 'respiration' (20%), while in the terrestrial animal-associated community this is only 3%. The analysis was done for microbial communities as well as viral communities. Like the marine viral metagenomic studies discussed above, there was a striking metabolic diversity among the viral sequences, which act as a reservoir and store of potential evolutionary change in their microbial hosts.

3.5 Genomic approaches to biodiversity and ecosystem function: an appraisal

Microbial ecology has developed a wide variety of methods that can be used to study functions in ecosystems and it can rely on a rich background of biochemistry and genetics of the organisms involved. Genes associated with crucial links in all major biogeochemical cycles have been identified. The presence and diversity of these genes can be studied without consideration of the species in which they are present, as indicators of an ecosystem's capacity to perform specific functions. Most of these genes are known from microbial biochemistry and microbial genetics but the possibilities offered by this strong background have not yet been exploited fully in an ecological context. Ecological studies have focused on genes from the nitrogen cycle, in particular denitrification genes. The few genomics studies that have been conducted with the aim of answering questions about biodiversity and ecological function seem to indicate a positive correlation between overall function and species biodiversity, supporting the rivet hypothesis (see Section 3.1). Community genomics studies also confirm that species carrying out indispensable key functions are not always among the dominant members of the community.

Ideally, assessment of functions in the environment would use transcription profiling, starting with extraction of functional mRNAs, as is done in model eukaryotes. This is more difficult in prokaryotes, and in microbial studies it is mostly limited to profiling pure cultures, rather than communities in the field. However, there are some promising new technical developments in which whole community RNA amplification is applied to obtain sufficient amounts of microbial mRNA from environmental samples (Gao *et al.* 2007). Usually, however, the potential for microbial functions is assessed from the presence, not the expression, of functional genes. With this limitation, the various diagnostic microarray approaches hold good promise for large-scale profiling and monitoring. However, technological improvement is still necessary especially with respect to specificity, to allow phylochips and geochips to be used as reliable diagnostic instruments. Brute-force sequencing, using ultra high-throughput technology, is developing as an alternative strategy.

Microbial ecology faces a tough problem, in that many organisms in the environment cannot be cultured in the laboratory and so remain undescribed as proper species. With every survey of environmental samples, new sequences are detected and a levelling-off of the collector's curve is not yet in sight, not even for common habitats such as lake sediments and forest soils. The most common system to classify microorganisms is the SSU rRNA gene, although some groups remain unnoticed in this way, as demonstrated by the recent discovery of a completely new archaeal phylum, the Nanoarchaeota. Theoretical arguments, based on

lognormal distributions of species over abundance classes, predict that prokaryote biodiversity may be an order of magnitude greater than even the maximum indicated by current genomics surveys.

Recent studies have also shown that spatial aspects and geographic barriers are more important in microbial community composition than thought before (Whitaker *et al*. 2003; Fullthorpe *et al.* 2008). We have seen evidence for this in the various soil genomics studies, but also in the aeroplankton surveys discussed in Section 3.2, where two different sampling locations revealed differing community compositions, suggesting that there are local sources of airborne microbes. Studies of the relationship between structure and function in microbial communities should include an assessment of the sources of biodiversity, as in community studies of island biogeography (Curtis and Sloan 2004).

Community genomics is a recent field where exciting new discoveries are being made. This has been made possible by new techniques of library construction and massive sequencing capacity using next-generation technology. The two studies published in 2004 on Sargasso Sea microplankton and on acid-mine drainage biofilms are particularly impressive and announced the beginning of a major new direction of research. The question can be asked, are there limits to what brute-force sequencing of the environment may reveal? There are theoretical arguments to support the point that in species-rich ecosystems even the most massive of next-generation sequencing methods will not be sufficient to reach a reasonable coverage of the metagenome. However, a full reconstruction of the microbial metagenome (> 90% coverage of the gene complement) is likely to be within reach for marine ecosystems and ecosystems in extreme environments. A particular interesting aspect of extreme environments is the mutual interdependency of genomes in such communities, varying from syntrophy to symbiosis.

The community genomics or metagenomics approach has also brought renewed interest in bioprospecting; that is, exploring novel functions by screening metagenomic libraries in which unknown genes from the environment are expressed in a heterologous host. This strategy has until now mostly been focused on functions of biotechnological relevance (new enzymes, antibiotics pathways); however, we see no reason why the same approach could not be applied to solving ecological questions.

Has ecological genomics solved the basic ecological question about biodiversity and ecological functions? Not yet. Only preliminary answers can be given at the moment, but a major breakthrough is likely to come soon.

CHAPTER 4

Life-history patterns

The great diversity of life cycles in nature, from short-lived annual plants to long-lived trees, from prolific reproducers such as oysters to economical types such as the albatross, has fascinated many biologists. An important scientific question has arisen from this fascination: can the life cycle of a species be understood in the context of the environment in which it lives? Attempts to answer this question have given rise to a large body of ecological literature, including observational data (population census, age distributions), experimental data (life-history manipulation by food or temperature), genetic data (heritable variation in life-history traits), and theoretical models (demography and quantitative genetics). Life histories of model organisms have also been analysed using molecular and genomic approaches. In this chapter we will visit this exciting blend of life-history ecology and genomics.

4.1 The core of life-history theory

When Charles Darwin travelled with the H.M.S. Beagle to Tierra del Fuego and the Falkland Islands in 1834, he was surprised to find, on counting the eggs of a large white *Doris* (a sea slug), how extraordinarily numerous they were. The slugs produce their eggs as a long ribbon, rolled up in a cone, and adhered to the rock. The inquisitive naturalist, extrapolating from a part of the structure, estimated the total number of eggs in one spire to be at least 600 000. Yet the animal itself was certainly not very common. Although he often searched under the stones, he could find only seven individuals. He then added in a footnote (Darwin 1845):

> No fallacy is more common with naturalists than that the numbers of an individual species depend on its powers of propagation.

If the abundance of a species does not depend on the number of offspring, why are there such large differences between species in reproductive styles? That such differences exist is obvious. The mathematical biologist A.J. Lotka distinguished between the 'lavish type' and the 'economical type' (Lotka 1924). Many marine animals, including Darwin's slug, obviously belong to the first type, whereas humans, with their low birth rate and long lifespan, are an example of the second type. Also within relatively homogeneous groups of organisms large differences between species may exist. For example, in birds clutch size varies from one egg in petrels and condors up to 20 eggs in pheasants and partridges. Likewise, among lizards the number of eggs laid in a season varies per species from 2 to 20. An even more extreme variation may be observed among plant species. The extent of reproductive output across species is usually positively correlated with juvenile mortality rate and negatively with longevity, so each species with stationary population size strikes a balance between mortality and natality; however, the weights on each side of the balance vary enormously between species.

An answer to the question of why the power of propagation differs so much between species comes from life-history theory. Characteristic for this area of population ecology is the recognition that the various *life-history traits*, also called *vital rates*, such as juvenile mortality, age at maturity, adult body size, clutch size, and longevity, cannot be isolated from each other. The theory considers all these traits

jointly, including the interrelationships among them. An important concept is the presence of *trade-offs*. A trade-off is a negative correlation between two life-history traits in such a way that an increase of one trait (e.g. clutch size) imposes a cost to another (e.g. chick mortality). However, the term trade-off is also used in a broader sense to indicate any negative correlation between two traits, whether they are causally related or not. Given the trade-off structure among life-history traits, the theory attempts to explain patterns across species by assuming that the life history as a whole is subjected to natural selection and is optimized with respect to the environmental conditions to which the organism is exposed, subject to lineage-specific constraints. Life-history theory has developed into a major field of population ecology and the subject is summarized in several textbooks, such as those by Stearns (1992) and Roff (2002).

Life histories can be described using the formalism of *demography*, the science of age-structured populations and their dynamics. The aim of demographic analysis is to develop models in which life-history traits, such as mortality and fertility, are linked to population size and age structure. In human demography such models are used to forecast future population growth and composition, given fertility and mortality schedules. In ecology, the same models are used to estimate the optimal life history, given a trade-off structure among the vital rates and a criterion for optimality, which is usually taken as the population growth rate. In other words, it is assumed that the survival and fertility of a species are optimized in such a way that, accounting for constraints from trade-offs, population growth rate is at a maximum. In addition to trade-offs, vital rates are also constrained by *lineage-specific effects*, which arise from the body plan, the physiological capacities, and the network of gene interactions posed by the phylogenetic history of the group to which the species belongs, which, for example, prevents a bird from producing as many eggs as a mussel.

As an example of how arguments are developed in life-history theory, we consider a relatively simple model outlined by Roff (2002), who considers a hypothetical animal in a constant environment with a simplified life history about which the following two assumptions are made. First, mortality rate, indicated by θ, is constant. This implies that throughout life a constant fraction of a birth cohort is removed and the survival of the cohort is described by a negative exponential. Such a survival curve is often seen in species for which mortality mainly comes from external sources, such as from predators. Second, fertility is constant within one life cycle, but there are different options, depending on the age at which reproduction starts. If reproduction starts later, fertility is higher; a linear relationship is assumed between fertility and the age at maturity. Such a relationship is often seen in organisms in which reproductive capacity increases with body size. By postponing reproduction, the animal can reach a larger body size and consequently achieve greater fertility.

These assumptions are displayed graphically in Fig. 4.1. Three possible fertility scenarios are plotted: one that starts at age 5 and maintains a reproductive output of 15 per time unit (the total number of eggs laid by a female surviving to age 20 would be 225), one that starts at age 10 and produces 40 eggs per time unit (total number of eggs produced from age 10 to 20 would be 400), and one that starts at age 15 and produces 65 eggs per time unit (total output 325 eggs). If all animals survive to age 20, the best of these three scenarios would be the second; however, a still-better scenario is to start at age 11, since this would provide a total reproductive output of 405 eggs (= (20–11) × 45). However, reproductive output must be corrected for the fewer and fewer animals remaining to produce eggs and so the actual optimum will be lower than 11. How much? A quantitative valuation can be given by considering a common criterion for optimality, net reproductive rate, R_0, which is given by

$$R_0 = \int^{\infty} l(x)m(x)\mathrm{d}x$$

where $m(x)$ is fertility (number of offspring produced per time unit by individuals aged x), $l(x)$ is survival from birth to age x (as a fraction), and α is the age at maturity (i.e. first reproduction). Net reproductive rate is the average number of individuals in the next generation by which an individual

of the present generation is replaced. Although Roff (2002) uses R_0 as a criterion for optimality, most authors prefer a parameter called intrinsic population growth rate, which can be calculated from $l(x)$ and $m(x)$ in a similar but more complicated way. R_0 ignores variation in generation time but is preferred in field studies with more or less stationary populations (Kozlowski 1993). The use of both net reproductive rate and intrinsic growth rate is supported by the *central theorem of demography*, which states that any population with time-invariant fertility and mortality schedules will, after some generations, attain a stable age distribution and then grow exponentially. A strong implication of the central theorem is that the growth factor per generation, R_0, can be estimated beforehand, directly from the life history; we need not wait until exponential growth is actually realized. In this way, R_0 can be considered an indicator measuring the fitness of a life history.

The equation above shows that to estimate R_0 one needs information about fertility and mortality rates for each age class; in the case considered by Roff (2002), $l(x) = e^{-\theta x}$ and $m(\alpha) = a + b\alpha$, where θ is the rate of mortality (set at 0.2 in the example), and a and b are constants (set at –10 and 5, respectively). This parameterization leads to

$$R_0(\quad) = \frac{1}{}(a + b\quad)e^{-}$$

The optimal age at maturity, $\hat{}$, is that which maximizes R_0, and is found by evaluating

$$\frac{\partial R_0}{\partial} = \left(\frac{b}{} - a - \quad b\right)e^{-}\quad = 0$$

which produces

$$\hat{} = \frac{1}{} - \frac{a}{b}$$

So conditions promoting postponement of reproduction (large $\hat{}$) include a low rate of mortality (small θ) and a slow increase of fertility with age (small b). If two related species live under conditions of unequal mortality, everything else being

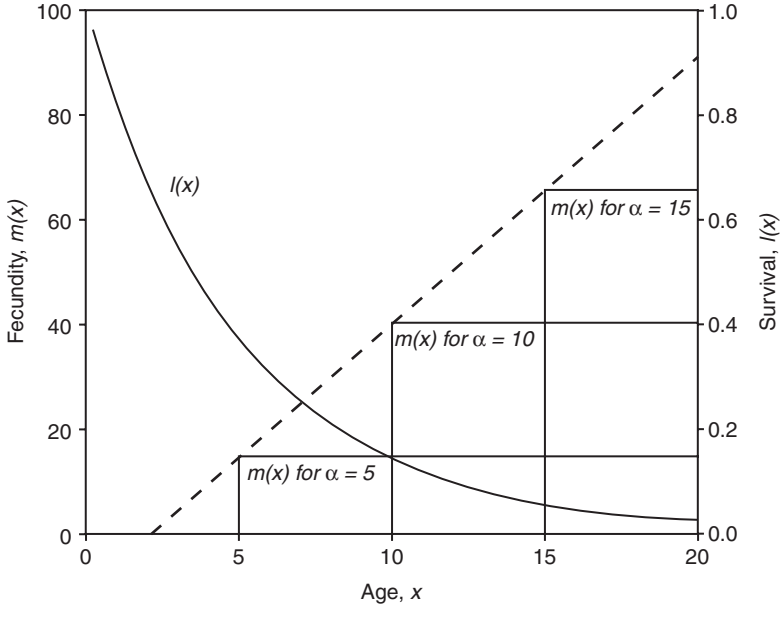


Figure 4.1 A hypothetical life history in which survival is a single exponential given by $l(x)=e^{-\theta x}$, where θ is the rate of mortality, and fertility, $m(x)$, is constant with age, but dependent on age at maturity, α, as $m=a+b\alpha$. Parameters are chosen arbitrarily as follows: $\theta = 0.4$, $a = -10$, and $b = 5$. Three different fertility scenarios are shown, for $\alpha = 5$, 10, and 15. Using the net reproductive rate R_0 as a criterion (see text) it can be shown that the optimal value for α is 7. After Roff (2002), reproduced with permission from Sinauer Associates.

equal, the one living under the highest mortality regime should have the shortest juvenile period. In the numerical example of Fig. 4.1, $\hat{} = 7$ and the net reproductive rate of the optimal life history is 30.8.

This simple example shows how mathematical models can help to draw inferences on the optimality of life-history traits if some basic attributes of the species and its environment are given. Whereas age at maturity, body size, and clutch size are important traits for animals, for plants traits such as germination fraction, shoot–root allocation, flowering time, and number of seeds are considered. Much more complicated elaborations of the basic demographic principles are discussed in Roff (2002), including fluctuating and predictably changing environments. A crucial role in many models is played by trade-offs; in the present case it was assumed that the organism could increase its fertility by postponing maturity and growing larger first, and that there was a linear relationship between the two. Such a mechanism is often considered a consequence of *energy allocation*: what is spent on one side cannot be spent on the other side. The idea of trade-offs due to energy allocation is very old and can be traced back to the '*loi de balancement*' proposed by Geoffroy Saint Hilaire in 1818 (Leroi 2001). Darwin (1859) noted that artificial selection of domestic animals and plants showed many examples of correlated responses or 'compensation of growth'. He referred to both Geoffroy and Goethe in stating 'if nourishment flows to one part or organ in excess, it rarely flows, at least in excess, to another part; thus it is difficult to get a cow to give much milk and to fatten readily'.

The principle of energy acquisition and allocation was developed by Kooijman (2000) into a systematic physiology-based framework of growth, reproduction, and aging, the *dynamic energy budget model* (DEB model). In this model, food uptake is assumed to be proportional to body surface and assimilated energy is converted into reserves. The reserve pool is in dynamic equilibrium with a mobile pool available to all organs of the body. A fixed proportion (κ) of the circulating pool is spent on growth plus maintenance, and the remaining portion, $1-\kappa$ on development plus reproduction. This aspect of the model is designated as the κ *rule* for allocation. Energy taken up by somatic tissues is first used for maintenance, and the remainder is used for growth. In this way growth competes directly with maintenance, but reproduction competes with growth only through the κ rule. This model can explain why many animals continue to grow after the onset of reproduction. Their growth slows down due to the increasing maintenance costs of a larger body, not directly through competition with reproduction. In the model, the onset of reproduction, which is due to the $1-\kappa$ flux being redirected from development to reproduction, does not create the discontinuities and inconsistencies that are present in other allocation models. The great variety of examples discussed in Kooijman's (2000) book illustrates that the DEB model is a powerful instrument for analysing energetic relationships since it argues from first principles rooted in thermodynamics and emphasizes the similarities across organisms as different as yeast, waterfleas, parasites, fish, and birds.

Despite the importance of allocations and trade-offs in life-history theory, reliable empirical measurements are difficult. This is especially annoying because often the outcome of an optimization procedure depends critically on the shape of a trade-off function, for example a convex relationship between reproduction and survival predicts *iteroparity* (repeated ongoing reproduction), whereas a concave relationship predicts *semelparity* (a single, large reproductive output followed by death). Empirical studies are hardly able to distinguish between these two forms of trade-off. In addition, trade-offs may be masked by fluctuation in the resource that is the subject of allocation. For example, if the total energy available for growth and reproduction increases due to increasing food intake, the allocation between them becomes invisible, and the correlation between growth and reproduction at the phenotypic level may turn from negative to positive (Van Noordwijk and De Jong 1986). There has been a tendency to measure trade-offs in terms of negative genetic correlations between life-history traits, either by pedigree analysis or by selection, using the formalism of quantitative genetics; however, such estimates have not produced satisfactory results because very large sample sizes are needed to resolve the presence of genetic correlations (Roff 2002).

In addition to energy allocation, two other classes of mechanism can cause negative associations between life-history traits: *negative* (*antagonistic*) *pleiotropy* and *co-regulation by signalling pathways* (Zera and Harshman 2001; Leroi *et al*. 2005). Geneticists define pleiotropy as the phenomenon that one gene affects two or more phenotypic traits. Negative pleiotropy arises when expression of a single gene affects one trait in a positive way and another trait in a negative way. This can also be true for signalling pathways: the same biochemical cascade or hormone signal may affect one process in a positive direction, and another in a negative direction. Negative pleiotropy may be a more common mechanism for negative association between life-history traits than energy allocation. Leroi *et al*. (2005) examined several classes of genes involved with the regulation of longevity and concluded that many of them have antagonistic effects on life-history traits. Additional evidence for the importance of negative pleiotropy comes from the literature on tolerance to pesticides (Van Straalen and Hoffmann 2000). Many pesticide tolerances are associated with apparent 'costs' to vitality or reproductive capacity, but these costs are more often due to metabolic side effects of a gene mutation that confers tolerance, than to an energy drain towards detoxification of the pesticide.

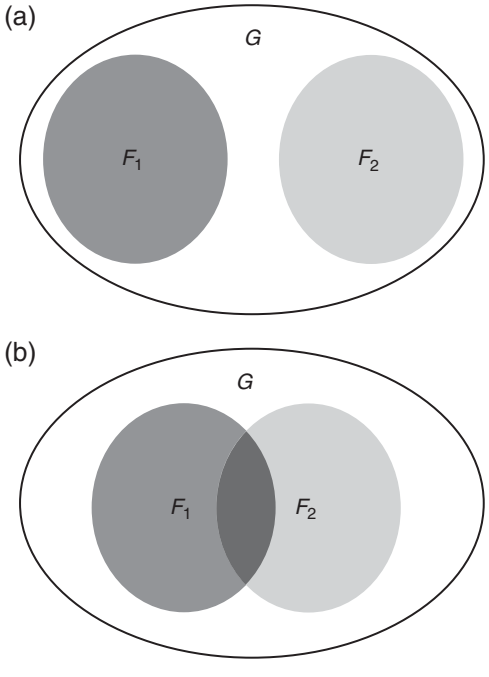


Figure 4.2 Theoretical illustration of how trade-offs between life-history traits might be explained by conflicts over gene expression. G is the set of all expressed genes in an organism, F_1 is the set whose expression changes in response to function 1, and F_2 to function 2. If there is overlap between F_1 and F_2, as in (b), a trade-off will arise if genes need to be upregulated in response to 1 and downregulated in response to 2. After Stearns and Magwene (2003) by permission of the University of Chicago Press.

A new framework for addressing questions about life-history patterns may come from genome-wide gene-expression studies. Genomics may serve as a biomolecular underpinning of the mathematical framework, providing mechanistic explanations for the various trade-offs assumed in life-history theory (Roff 2007). Stearns and Magwene (2003) suggested that trade-offs can be seen as *conflicts over gene expression*. This argument is illustrated in Fig. 4.2. If two functions in an organism—for example, reproduction and longevity—are regulated by two sets of genes, a trade-off between the functions may arise if the two sets overlap and if, for example, one function calls for upregulation and the other for downregulation. Such a trade-off can also arise from conflicting signals downstream of gene expression, as we will see below.

Finally, we want to point out that life-history traits always result from interaction between genotypic determinants and the environment. Even those traits that are under strong genetic control, for example clutch size in birds, may vary depending on environmental conditions. The way in which a life-history trait is shaped by an environmental factor can itself be considered an aspect of the life-history pattern. For example, some birds will tune their clutch size to environmental food supply, taking advantage of years with abundant food availability and minimizing reproduction in bad years; other birds just lay the same number of eggs irrespective of the environment. This aspect of a life history is denoted as *phenotypic plasticity*, and the function describing a life-history trait as a function of an environmental factor is called a *norm of reaction*. A reaction norm is a property of a genotype;

different genotypes may have different slopes of the reaction norm, allowing natural selection to act on plasticity. Until now plasticity as an adaptive trait has been analysed mainly by statistical methods (multivariate quantitative genetics), but genomic technology is opening up new perspectives for providing a genetic basis to phenotypic plasticity (Pigliucci 1996; Gibson 2008).

In this chapter we will explore the genomic basis of life-history patterns, emphasizing proximate causation in the life history. There is a lot of literature on the evolution of longevity and aging that is left undiscussed here; the reader is referred to Kirkwood and Austad (2000), Partridge (2001), and Kenyon (2010) for reviews of theories and experimental data. The molecular work on aging, which is mainly on *C. elegans*, *Drosophila*, and mouse, is often inspired by its possible extrapolation to human gerontology and is not very well integrated in ecological studies. We nevertheless believe that a molecular connection can provide a mechanistic basis for life-history theory and so increase its power to explain the variety of life-history patterns in nature.

4.2 Longevity and aging

The nematode *C. elegans* has developed into a classical model for the study of aging. Most of this research is motivated by a possible connection with aging in humans, but the new insights accumulated over the last few years are of equal relevance to ecology and life-history theory. It also appears that the properties of the molecular machinery regulating aging are shared with *Drosophila* and are even conserved across the whole animal kingdom. The main reason why *C. elegans* became a model for aging was due to the discovery of mutants that showed extended longevity. It turned out that the genes mutated to extend longevity were the same as those associated with the formation of dauer larvae. The dauer larva is a developmentally arrested stage, the entry of which is triggered by adverse conditions such as food shortage or crowding (see Section 2.3). Here we will consider the signalling pathways associated with dauer formation as well as extended longevity.

4.2.1 The insulin signalling pathway

The first report of single-gene mutations that affected lifespan and reproduction in *C. elegans* dates back to 1988. The gene involved was called *age-1*, and mutations in this locus increased longevity and decreased hermaphrodite fertility (Friedman and Johnson 1988). Later, Kenyon *et al*. (1993) discovered that two other single-gene mutations could affect lifespan. One of these longevity-regulating genes was known as *daf-2* (daf from dauer formation); animals mutated at the *daf-2* locus lived twice as long as the wild-type worms (Fig. 4.3). It was also shown that a second gene, *daf-16*, is required for the lifespan-extending effect of *daf-2*, because the double mutant, *daf-2*; *daf-16*, had a normal lifespan (Fig. 4.3). Interestingly, the reproductive output of the *daf-2* mutant was hardly decreased (the brood size was 212 ± 36 eggs in the long-lived mutant versus 278 ± 35 in the wild type) and it was also shown that ablation of the germ-line precursor cells, effectively sterilizing the adult, did not increase lifespan. Thus the increased lifespan of the *daf-2* mutant was not a consequence of lower reproduction. Later several other genes were identified that can extend longevity when altered in *C. elegans*, and in total approximately 70 are known today. These genes include regulators of metabolism, genes involved in sensory

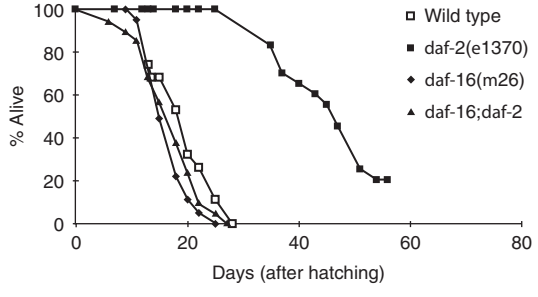


Figure 4.3 Survival curves for *C. elegans* mutated in the *daf-2* gene (*daf-2*(*e1370*)), compared with a control group (wild type). Survival curves are also shown for nematodes mutated in the *daf-16* gene (*daf-16*(*m26*)) and in both genes (*daf-16*;*daf-2*). Median survival times are 17 days for *m26*, 17 days for *m26*;*e1370*, 19 days for the wild type, and 46 days for *e1370*. The comparison of mutants demonstrates that *daf-2* dysfunction increases longevity by more than a factor of 2, and this effect requires a functional copy of *daf-16*. From Kenyon *et al*. (1993) by permission of Nature Publishing Group.

perception, and reproduction genes. In addition, several genes regulating longevity were found to be associated with defence against oxidative stress. One specific group of genes receiving much attention appeared to encode proteins of the *insulin signalling pathway*.

Insulin is a peptide hormone that in vertebrates is secreted from groups of cells associated with diverticula of the gut, in mammals taking the form of the islets of Langerhans in the pancreas. Insulin-like peptides and insulin-like growth factors (IGFs) are also present in invertebrates, usually in their highest concentrations in the gut. Insulin in mammals has a central role in carbohydrate and lipid metabolism, the best-known effect being increased uptake of glucose from the blood by muscles and adipose tissue and the formation of glycogen by the liver. Insulin and IGF do not enter their target cells but instead react with a receptor protein in the cell membrane, the *insulin/IGF receptor*. This protein has an extracellular domain, binding insulin or IGF, and a cytosolic domain, which acts as a tyrosine kinase: when activated it catalyses the phosphorylation of tyrosine residues in cytosolic proteins. These proteins in turn activate others in a complicated cascade, finally leading to phosphorylation of a cytosolic protein known as DAF-16. DAF-16 is a transcription factor of the so-called *forkhead* family. When active, this transcription factor switches on a series of genes that form a programme of dauer-larva formation. As long as DAF-16 is phosphorylated by DAF-2 signalling it cannot enter the nucleus and so is effectively inactivated as a transcription factor (Lin *et al*. 2001). The secretion of insulin-like peptides is under control of the nervous system, which receives sensory input from the mouth region. So the whole system seems to be targeted towards translating information about food availability in the environment into either normal development (DAF-16 inhibited by activated DAF-2) or a dauer programme (DAF-16 activated by relaxation of DAF-2; Braeckman *et al*. 2001; Olsen *et al*. 2003; Fig. 4.4). The pathway is referred to as insulin/IGF-1 signalling.

The fact that mutations in the DAF-2/DAF-16 pathway regulate longevity as well as dauer formation strongly suggests that the effect of *daf-2* knockout on longevity is basically a dauer programme 'mis-expressed' in the adult. Indeed, Dillin *et al*. (2002), using RNAi to suppress *daf-2* and *daf-16* at different times in the life cycle, found that the same pathway regulates aging in the adult and dauer formation in the larvae. Jones *et al*. (2001), analysing gene expression in dauer and non-dauer populations of *C. elegans*, showed that the dauer-specific transcriptome was greatly enriched in genes that previously were known to regulate longevity. The study also showed that dauer formation is not to be considered a true resting stage from a genetic point of view, because no fewer than 18% of the genes detected were specifically upregulated in the dauer larva, compared to a mixed-stage population. Similarly, Wang and Kim (2003) in a microarray study classified 1984 genes as dauer-regulated, which is 11% of the *C. elegans* genome. The dauer-enriched genes include several enzymes characteristic for anaerobic metabolism (Holt and Riddle 2003).

The long-lived non-dauer *daf-2* mutant illustrates that it is possible to uncouple part of the dauer programme (the part that confers extended longevity) from the main programme leading to quiescence. The mutations in *daf-2* and other genes of the insulin/IGF-1 signalling pathway that affect longevity are weak mutations. Strong mutations in the same genes cause the *C. elegans* larvae to go into dauer dormancy regardless of environmental cues. This indicates that there may be thresholds in the levels of endocrine signalling such that a mild decrease of signalling is already sufficient to start the anti-aging programme, whereas a further decrease is necessary to enter the dauer stage. These thresholds could also be dependent on temperature; some *age-1* mutants that develop into long-lived adults at normal culture temperature will go into dauer diapause at high temperature (27 °C).

Increased longevity similar to decreased DAF-2 signalling is also seen when the nematodes are cultured under *dietary restriction*, also called *caloric restriction*. This refers to a diet in which food intake is limited to 30–40% of the intake shown by animals fed *ad libitum*. In *C. elegans* this can be achieved by diluting the bacteria in the medium or by applying an axenic medium with heat-killed *E. coli* cells.

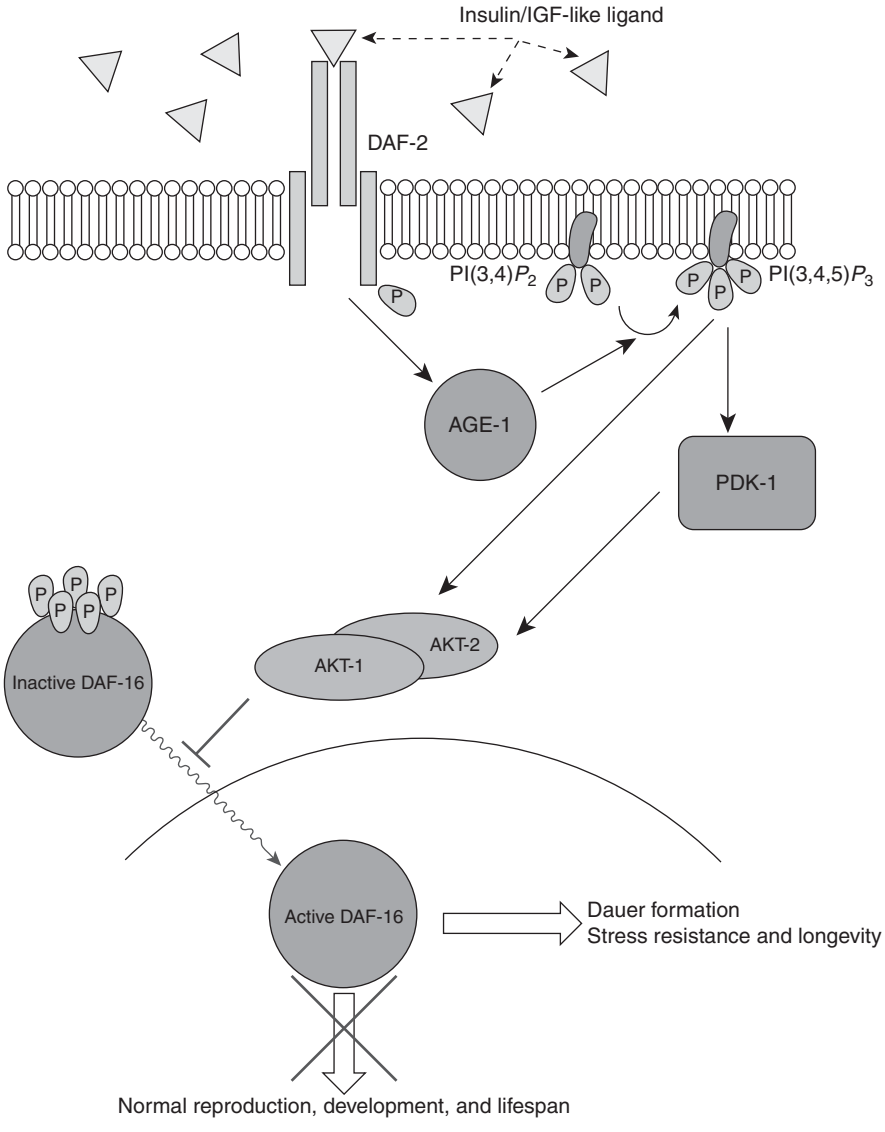


Figure 4.4 Scheme of the insulin/IGF-1 signalling pathway, showing how extracellular insulin-like peptides may trigger a cascade of events, starting with binding to DAF-2, a receptor protein in the cell membrane, and eventually leading to phosphorylation of DAF-16, a transcription factor of the forkhead family. By decreased insulin/IGF-1 signalling DAF-16 is activated, allowed to enter the nucleus, trigger a programme of dauer formation, stress resistance, and longevity, and suppress normal reproduction. P, phosphate group; PDK-1, phosphoinositide-dependent kinase 1; PI(3,4)P_2, phosphatidylinositol 3,4-bisphosphate; PI(3,4,5)P_3, phosphatidylinositol 3,4,5-trisphosphate. After Olsen *et al*. (2003), with permission from Springer.

Because the insulin signalling pathway is associated with carbohydrate metabolism, and the apparent cues for dauer induction come from food availability, it seems logical to conclude that dietary restriction promotes longevity through the same DAF-2/DAF-16 pathway. However, this appears not to be the case. Houthoofd *et al*. (2003) did experiments with *daf-2* and *daf-16* mutants both exposed to dietary restriction (Table 4.1). Their data show clearly that dietary restriction causes an increase in

the median survival time by a factor of two and a half, both in the wild type and in the *daf-16* mutant. In addition, a restricted diet can boost the median lifespan of the *daf-2* mutant (which is already increased by a factor of 1.7 compared with the wild type), to a record value of 90.9 days (maximum, 136 days). In human terms these animals would correspond to healthy 500 years old! Interestingly, extension of lifespan in this way is not accompanied by decreased metabolic activity; actually, respiration was elevated substantially by dietary restriction as were activities of antioxidant enzymes, such as superoxide dismutase (Houthoofd *et al*. 2002).

Another mutation that extends lifespan in *C. elegans* seems to act mostly independently of the insulin/IGF-1 signalling pathway. Mutations in the so-called *clk-1* gene (the name derives from clock biological timing abnormality) increase longevity in association with a slowing down of the rate of many processes, including cell division, rhythmic behaviour, rate of feeding, and mitochondrial respiration. The correlation between longevity and metabolic rate supports the *rate of living hypothesis* of aging, which states that aging is a consequence of accumulating metabolic damage, in particular from endogenous oxygen radicals (Finkel and Holbrook 2000; Hekimi and Guarante 2003). *Reactive oxygen species* such as superoxide anion (O_2^-), hydroxyl radical ($OH^\bullet$), and hydrogen peroxide (H_2O_2) are amply generated by the electron-transport chain in the mitochondrion. The major source is complex III, in which ubiquinone, alias coenzyme Q, resides. This electron-transport molecule has an unstable intermediate that can easily donate electrons directly to molecular oxygen rather than to complex III. The mutated *clk-1* protein cannot perform the final step in the biosynthesis of ubiquinone and as a consequence the precursor, demethoxyubiquinone, is incorporated in the electron-transport chain. Paradoxically, this alternative component appears to be less prone to the production of reactive oxygen species. Another role for *clk-1* has been suggested by Branicky *et al*. (2000). The protein could have a regulatory role by reporting to the nucleus on the metabolic state of the mitochondria in such a way that the rate of living may be adjusted to the generation of energy. If *clk* is mutated nuclear genes do not receive the right information on respiration and set the rate of living at a default level that is lower than normal. This theory is interesting because it would imply that the effect of respiration on aging acts through a metabolic switch, rather than through a direct impact of damage accumulation (Guarante and Kenyon 2000).

In addition to mutations, surgical operations can be applied to *C. elegans* to increase longevity. An interesting effect is seen after removal of the germ line. As noted in Chapter 2, *C. elegans* has a completely determinate developmental pattern in which the destination of every cell is fixed as soon as it comes into existence; the adult has exactly 959 cells, not including eggs and sperm cells, which descend from the zygote in a fixed lineage. Two cells, Z2 and Z3, give rise to the germ line by continuous division during development. The germ cells differentiate into sperm during the L4 stage or oocytes during adulthood (see Fig. 2.19). Removing the germ-line precursor cells by means of a laser extends the lifespan of *C. elegans* by about 60% and this effect requires the presence of DAF-16. Animals that lack germ cells due to a mutation are also long-lived. In a series of elegant experiments measuring the survival curves of various mutants, Arantes-Oliveira *et al*. (2002) showed that the lifespan-suppressive effect of the germ line is not dependent on the sperm or oocytes themselves, but only on proliferating, active germ-line precursor cells. So, despite the obvious negative correlation between reproduction and longevity in this system, there does not seem to be a simple trade-off in the classical sense that energy allocated to reproduction would detract from maintenance and so increase the rate of aging. Rather, a signal from the germ-line stem cells directs both aging and reproduction, maybe by altering the production of a steroid hormone or by altering the response to such a hormone (Arantes-Oliveira *et al*. 2002).

An important issue in the metabolic network affecting aging in *C. elegans* is the question of which genes act upstream and which downstream in the insulin signalling pathway. Taking DAF-16 as a reference point, *upstream* genes are defined as those that affect expression or activity of DAF-16. To this category belong the neurosecretory signals leading

Table 4.1 Survival times of different populations of *C. elegans* showing that effects of dietary restriction increase longevity independent of the insulin signalling pathway

Population	Medium lifespan (days; mean ± S.E.)	
	On normal medium	Under dietary restriction on axenic medium
N2 (wild type)	14.4 ± 0.1	36.4 ± 0.2
daf-16(*m26*) mutant	12.9 ± 0.1	32.6 ± 0.2
daf-2(*e1370*) mutant	24.3 ± 0.2	90.9 ± 0.4

Source: After Houthoofd *et al*. (2003).

to secretion of insulin-like factors, the genes encoding the insulin-like peptides themselves, the insulin/IGF receptor DAF-2, and the various genes in the signalling cascade leading to phosphorylation of DAF-16, such as *age-1* and *pdk-1*. The *downstream* genes include the ones that are regulated by the transcription factor DAF-16 and whose expression contributes to longevity and stress resistance. The difference between upstream and downstream genes is difficult to make when looking at gene expression as such, but can be unravelled by studying mutants that are knocked out at crucial positions in the pathway. For example, Murakami and Johnson (2001) studied a transmembrane tyrosine kinase gene called *old-1*, which if overexpressed in *C. elegans* increases longevity by a factor of 1.5; using transgenic nematodes in which the *old-1* gene was fused to green fluorescent protein the authors observed that the positive effect of *old-1* on longevity was absent in a mutant in which *daf-16* was knocked out. This makes it likely that *old-1* is regulated by DAF-16; that is, it is downstream of DAF-16.

4.2.2 Genome-wide analysis of lifespan modulation

The various single-gene mutation studies and other manipulations have demonstrated that the effect of insulin/IGF signalling on longevity in *C. elegans* is linked to a large number of other processes, such as reproduction, lipid metabolism, diapause entry, and stress resistance. This has made the situation increasingly complex and it has become difficult to see the bigger picture. However, genome-wide microarray studies (Murphy *et al*. 2003; Golden and Melov 2004; McElwee *et al*. 2004) have made an important breakthrough, presenting an interpretative framework for earlier data and even a new theory of aging.

Murphy *et al*. (2003) aimed to identify all the genes that act downstream of DAF-16 by means of a cDNA microarray survey. These authors focused on genes that showed opposite expression profiles in *daf-2*- versus *daf-16*-knockout mutants. They also treated animals with RNAi of selected genes to confirm that these genes had an effect on longevity. A division into two classes was made: class-1 (lifespan-extending) genes were upregulated in *daf-2* mutants and in *daf-2* (RNAi) animals, but downregulated in animals in which both *daf-16* and *daf-2* were inhibited by RNAi. The second class of genes (lifespan-shortening) was defined by genes that displayed the opposite profile. A relatively succinct list of genes could be identified as belonging to either class and the most prominent ones of each class are listed in Table 4.2.

Inspecting the genes in the two classes, it is obvious that the first class contains many genes of *stress-defence systems* (Table 4.2). Heat-shock proteins are molecular chaperones that support the folding of other proteins; many of them are highly inducible by several stress factors, including a heat shock, in which they were first described. Cytochrome P450 is an enzyme that catalyses the oxidation of aromatic lipophilic compounds, including many xenobiotics, but also endogenic substances such as steroids. Catalase is an enzyme that supports the

catalysis of hydrogen peroxide, a very reactive oxygen species that may cause damage to membranes. It is also striking that class 1 includes several proteins that are known to be involved in antibacterial defence. Stress-defence systems will be discussed in more detail in Chapter 5. The class 2 genes listed in Table 4.2 are less well defined; in addition to a yolk protein, class 2 includes many proteins of unknown function.

Several of the differential expressions reported in Table 4.2 were also found in a microarray study by Golden and Melov (2004), who compared a *daf-2* mutant with the wild type. A recent metabolomics study (Fuchs *et al*. 2010) has identified characteristics of the carbohydrate profile that are indicative of long life: upregulation of the glyoxylate shunt and neoglucogenesis. These longevity-associated signatures were indeed expected from the gene expression studies discussed above. However, the long-life metabolome also showed shifts in amino acid metabolism that cannot be related easily to altered gene expressions of the insulin/IGF pathway, indicating a still unknown additional mechanism of longevity assurance in nematodes.

How DAF-16 upregulates or downregulates all the genes listed in Table 4.2 is not yet clear. Lee *et al*. (2003) identified 17 genes in the genome of *C. elegans* that were orthologous between *C. elegans* and *D. melanogaster* and had a putative binding site for DAF-16 in their promoter (between the start site and 1 kbp upstream). Expression analysis confirmed that six of them were differentially expressed between a *daf-2* and a *daf-2*; *daf-16* mutant, as expected in the case of genes with a

Table 4.2 Overview of genes in *C. elegans* acting downstream of DAF-16, identified by differential expression using a microarray and reduced expression of *daf-2* and *daf-16* by RNAi

Gene	Brief description
Class 1: upregulated by DAF-16 and positive regulators of longevity	
ctl-2	Peroxisomal catalase
dod-1	Member of cytochrome P450 family
hsp-16.1	Heat-shock protein
lys-7	Enzyme associated with response to pathogenic bacteria
dod-2	Thaumatin (sweet protein) associated with plant pathogenesis
hsp12.6	Heat-shock protein, α crystalline
mtl-1	Metallothionein-like, cadmium-binding protein
gei-7	Member of family of malate synthase/isocitrate lyase
dod-3	Protein of unknown function
dod-4	Aquaporin, member of family of transmembrane channels
Class 2: downregulated by DAF-16 and negative regulators of longevity	
dod-17	Protein of unknown function
nuc-1	Endonuclease associated with apoptosis
dod-18	Member of *Maf*-like transcription factors
dod-19	Protein of unknown function
gcy-6	Putative guanylate cyclase, catalysing cGMP second messenger
dod-20	Protein of unknown function
dod-21	Protein of unknown function
vit-5	Vitellogenin, 170 kDa yolk protein
mtl-2	Protein of unknown function
dod-22	Protein of unknown function

Notes: Only the first 10 most prominent genes are shown for each class.

Source: From Murphy *et al*. (2003).

causal relationship to the insulin/IGF-1 signalling pathway. Interestingly, and in accordance with Murphy *et al*. (2003), both the upregulated and the downregulated genes had the consensus binding motif. However, none of the 17 genes identified by Lee *et al*. (2003) is shared with the list of Murphy *et al*. (2003), which indicates that other binding sites and *trans*-acting factors other than DAF-16 may be involved.

Murphy *et al*. (2003) also discovered that a gene encoding an insulin-like peptide, *ins-7*, was present among the class 2 members. This protein is not only downstream of the DAF-16 pathway, but, being insulin-like, also acts as an activator of DAF-2. So, any positive signal on DAF-2 will amplify the pathway, via inhibition of DAF-16, relieving the negative regulation of *ins-7* expression and INS-7 stimulation of DAF-2. This positive-feedback loop in the system was thought to contribute to synchrony across cells in the animal. When dauer larvae sense the presence of favourable food conditions and the DAF-2 pathway is activated in some cells by increasing insulin-like peptides, an amplified signal to other cells will prevent the animals emerging from the dauer stage with a mixture of dauer and non-dauer cells.

Further fine-tuning of the insulin/IGF pathway is achieved by coregulation of different *daf-16* isoforms. Two isoforms of *daf-16* were known to exist in the genome of *C. elegans*, *daf-16a* and *daf-16b*, of which *daf-16a* was assumed to regulate longevity, stress responses, and dauer formation, as described above. However, recently a new isoform, called *daf-16d/f* was discovered (Kwon *et al.* 2010), which appeared to have the greatest effect on longevity, while *daf-16a* alone was insufficient for lifespan regulation. The three isoforms have overlapping but slightly different susceptibilities to the protein kinases of the insulin/IGF pathway, different tissue expression, and different spectra of downstream genes that they activate. *Daf-16d/f* was shown to be the most important regulator of longevity, with smaller effects on lipid metabolism, stress responses, and development, while *daf-16a* influenced longevity and lipid metabolism, and *daf-16b* was mainly involved in regulating development. By means of these differential but overlapping activity profiles the nematode as a whole is able to fine-tune its responses to changing environmental conditions.

An overview of the aging-regulator system of *C. elegans* is given in Fig. 4.5 (Gems and McElwee 2003). The bigger picture integrates many aspects of previous studies. It explains why there are many genes with a small additive effect on aging: they are all regulated by the same transcription factor. The presence of yolk proteins in class 2 is consistent with the link between aging and reproduction, while the presence of antioxidant enzymes in class 1 is in accordance with the well-known relationship between oxygen radicals, cell damage, and aging. Another confirmation of the picture is found in Hsu *et al*. (2003) who showed that the transcription factor HSF-1, which is a regulator of the heat-shock response, also influences aging. Overexpression of *hsf-1* extends lifespan, while reducing it shortens lifespan. So the two transcription factors DAF-16 and HSF-1 partly regulate the same genes with similar effects on aging.

Another genome-wide survey of insulin/IGF-1 signalling-mediated longevity in *C. elegans* is found in the work of McElwee *et al*. (2003, 2004). These authors compared the expression profiles of dauer larvae with those of *daf-2* mutants and argued that genes promoting longevity would have an expression signature similar to the dauer profile. It turned out that in fact around 21% of genes upregulated in dauer larvae are also upregulated in the *daf-2* mutants, and a similar pattern holds for the downregulated genes. Like Murphy *et al*. (2003), McElwee *et al*. (2004) noted a prominent representation of genes associated with stress-defence systems among the positive regulators of longevity. Several of these genes are known to be associated with the so-called *drug metabolism* or *biotransformation system*. This system of interacting enzymes, which is studied extensively in toxicology, involves two phases in which lipophilic, often aromatic, compounds are first activated and then conjugated to form water-soluble complexes that can be excreted. Phase I of the biotransformation pathway is conducted by enzymes of the *cytochrome P450* family. These are haem proteins that can oxidize aromatic compounds to phenols, epoxides, and quinones, which can then be subjected to

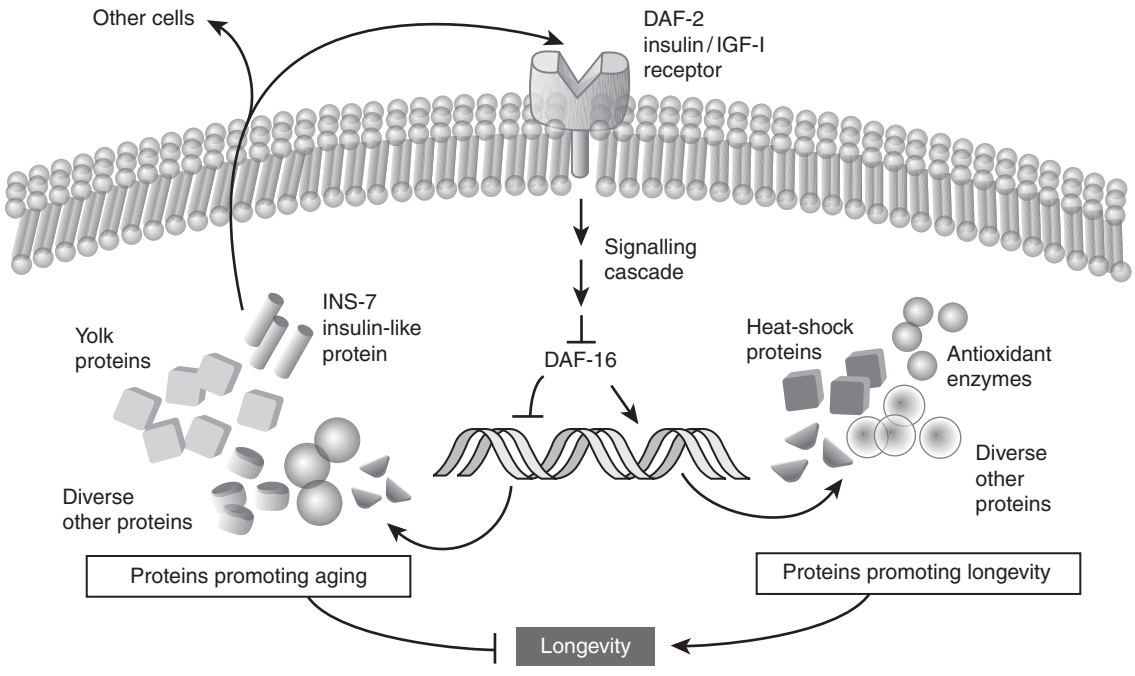


Figure 4.5 The insulin signalling pathway of *C. elegans* converges on repression of the transcription factor DAF-16. DAF-16 itself downregulates expression of proteins promoting aging and upregulates expression of proteins promoting longevity. Among the factors promoting aging is an insulin-like protein, INS-7, that acts upon DAF-2 as a positive-feedback link in the network and a signal to other cells. After Gems and McElwee (2003), by permission of Nature Publishing Group.

conjugation by enzymes of phase II, which attach endogenic compounds such as sulphate, glucose, glucuronic acid, or glutathione to the activated product of phase I. The system acts against an enormous variety of lipophilic compounds including many xenobiotics (drugs, environmental pollutants, and plant secondary metabolites) as well as endogenic aromatics such as steroid hormones. We will learn more about this defence system in Section 5.2.

The upregulation of many enzymes of the drug metabolism system in long-lived *daf-2* mutants suggested to McElwee *et al*. (2004) that this system plays a central role in protection from the metabolic damage that accumulates with age (Fig. 4.6). Various cellular processes as well as xenobiotics produce ample reactive molecules that can cause permanent damage to cellular constituents. The biotransformation system can prevent such damage and upregulation of its enzymes is assumed to promote longevity. This new theory of aging is attractive since it assigns a central biological role to the biotransformation system which obviously did not evolve only to metabolize man-made chemicals such as drugs. On the other hand, the new theory could turn out to be a bit too simplistic, since biotransformation does not only detoxify chemicals, it can also activate chemicals to intermediates that cause more harm than the substrate itself. The detoxifying role of the biotransformation system depends on a delicate balance between phase I and phase II enzyme activity; a general upregulation of all enzymes would be very ineffective and could even be damaging. In addition, many biotransformation enzymes have a great number of isoforms, with different induction profiles, where each substrate needs another isoform to be metabolized effectively.

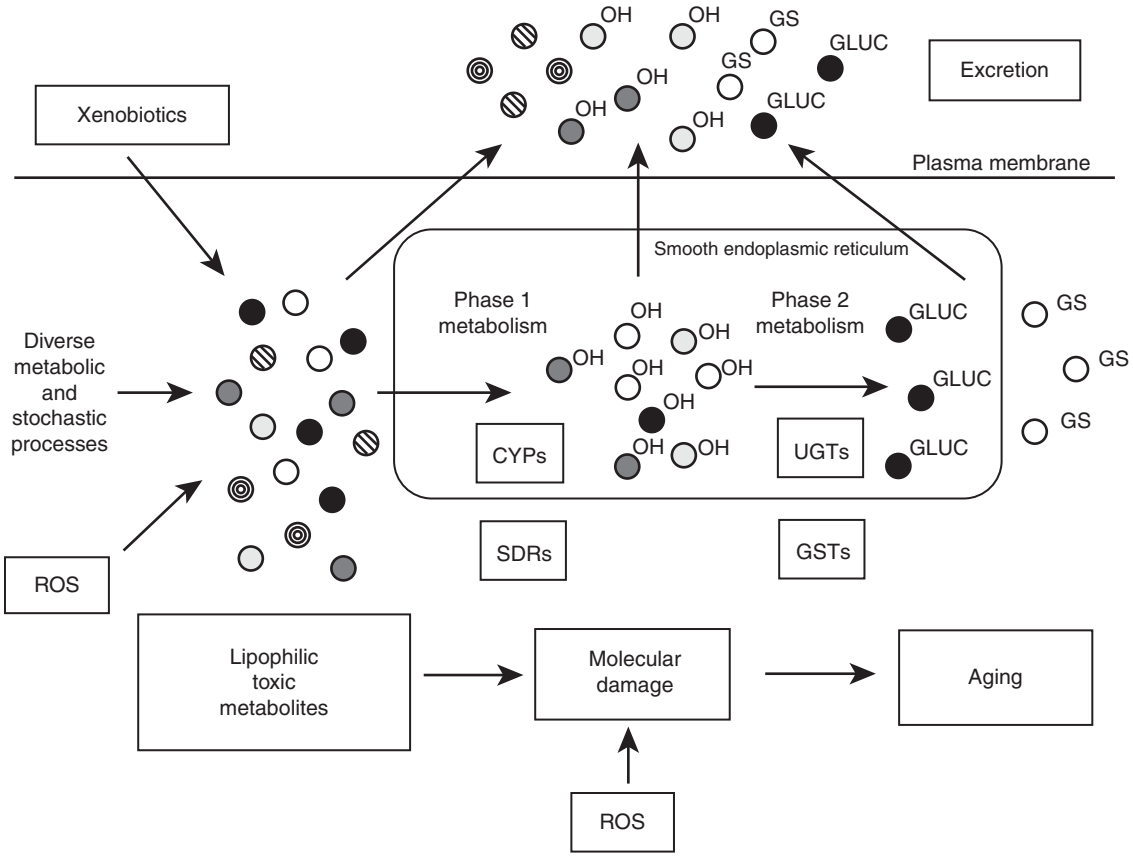


Figure 4.6 Schematic representation of the theory by McElwee *et al*. which holds that a major contribution to the cellular aging process comes from lipophilic toxins generated by cellular activity and xenobiotics, which are biotransformed by phase I and phase II metabolism. ROS, reactive oxygen species; CYPs, cytochrome P450s; SDRs, short-chain dehydrogenase/reductases; UGTs, uridinediphosphate glucosyltransferases and uridinediphosphate glucuronosyltransferases; GSTs, glutathione S-transferases; GLUC, glucosyl or glucuronosyl; GS, glutathionyl. From McElwee *et al*. (2004), by permission of the American Society for Biochemistry and Molecular Biology.

4.2.3 Longevity-regulating systems across species

One might think that the regulation of lifespan seen in *C. elegans* is tied so specifically to the dauer stage, which is absent in insects and vertebrates, that it does not have a validity outside nematodes. However, the converse is true! Fundamental aspects of the longevity programme elaborated in *C. elegans* appear to be conserved across organisms as widely different as yeast, *Drosophila*, and mouse; that is, the aging mechanisms are 'public' rather than 'private' (Gems and Partridge 2001; Partridge and Gems 2002a; Partridge and Pletcher 2003). This has stimulated hope for extrapolation to human longevity assurance, but it is also of relevance to ecology, since we may expect that several aspects of the genomic determination of longevity in model species are also valid for non-models in an ecological context.

In *D. melanogaster*, as in nematodes, the discovery of mutants with elongated lifespans has been the trigger for aging research. A first hint that regulation of aging was conserved between nematodes and fruit flies came from the observation that, as in *C. elegans*, *D. melanogaster* mutants disturbed in the insulin/IGF signalling pathway have a longer

lifespan. The proteins involved in the insulin/IGF-1 signalling pathway of *Drosophila* are INR (insulin/IGF receptor; homologous with DAF-2 in *C. elegans*), an insulin receptor substrate designated CHICO, a phosphoinositide 3-kinase, and a protein kinase B. Most attention has been paid to mutations in the gene encoding CHICO, which confer a dwarf phenotype with reduced fecundity, named after the smallest of the Marx brothers. Clancy *et al*. (2001), studying heterozygous and homozygous *D. melanogaster chico*^{1} mutants, were able to show that female median lifespan was increased from 48 to 57 days in the heterozygote and further to 63 days in the homozygote. The longevity-extending effect could be 'rescued' by introducing a P-element containing a *chico*(+) construct in the heterozygote *chico*^{1}. This reduced the median survival time back to a normal value of 49 days. This type of genetic manipulation, by which not only mutants but also engineered *rescue transgenes* are studied, is necessary by present molecular genetic standards to prove that a gene is causally linked to an observed phenotype.

Interestingly, the effect of *chico*^{1} acted more strongly in females than in males. The male heterozygous *chico*^{1} showed only a small increase in lifespan, whereas the homozygotes lived for even less time than the wild-type flies. This suggests that female reproductive tissues might have something to do with the regulation of longevity in *Drosophila*. Mutations in the other components of the insulin signalling pathway, for example in *InR*, the *Drosophila* homologue of *daf-2*, produced longer lifespans in mutants with a mild reduction of signalling (Tatar *et al*. 2001). In addition, Clancy *et al*. (2001) noted that the effect of *chico*^{1} was not due to reduced fecundity per se. This was obvious when *chico*^{1} was introduced in a strain carrying a mutation that blocks oogenesis and causes complete sterility. *chico*^{1} was still able to extend the lifespan of these sterile flies. Also, *chico*^{1} flies did not have a lower metabolism per unit of body mass (Hulbert *et al*. 2004).

Another commonality between the lifespan-regulatory mechanisms of *C. elegans* and *D. melanogaster* is the involvement of a forkhead transcription factor homologous to the mammalian *Foxo*, which is called *dFoxo* in *Drosophila* and is a homologue of the nematode *daf*-16. Overexpression of *dFoxo* in *Drosophila* fat body increased female lifespan by 20–50% and reduced fecundity by 50% (Giannakou *et al*. 2004). In accordance with Clancy *et al*. (2001) there was no effect in male flies.

Like *C. elegans*, *Drosophila* also shows the effects of dietary restriction. The response of lifespan to food concentration during larval culture shows an optimum curve, in which the optimum falls below normal. Low food density shortens lifespan due to a starvation effect, and high density shortens lifespan due to accelerated aging (Fig. 4.7). Interestingly, in this work the response of *chico*^{1} mutants to food was very similar to the wild-type flies, but was shifted to a higher food density (Clancy *et al*. 2002). The data can be explained by assuming that *chico*^{1} flies are less sensitive to food concentration, but respond similarly otherwise. This suggests that the metabolic changes induced by *chico*^{1} are in some way similar to those of dietary restriction and act along overlapping mechanisms. However, this conclusion is somewhat at variance with work on *C. elegans* (see above) and mice (Bartke *et al*. 2001), which has suggested that dietary restriction and reduced insulin signalling rely on two separate mechanisms. As in the case of reduced insulin/IGF-1 signalling, the lifespan-extending effect of dietary restriction is not a consequence of reduced fertility (Mair *et al*. 2004).

We should note that median survival time is only a summary statistic of the survival curve and it does not capture the difference between two treatments well when the shape of the curve is also affected. In principle, an increase in median lifespan can be brought about by two mechanisms: first, by a reduction in the rate of mortality at all ages, extending the survival curve by making it less steep, and second, by postponing the age at which mortality begins, causing a shift of the curve towards higher ages. The effect of dietary restriction on *Drosophila* is obviously of the first type. There does not seem to be a memory effect of previous food intake: flies that were subjected to dietary restriction after having been fed *ad libitum* for different times all showed the same immediate reduction of mortality rate (Mair *et al*. 2003). This would imply that dietary restriction

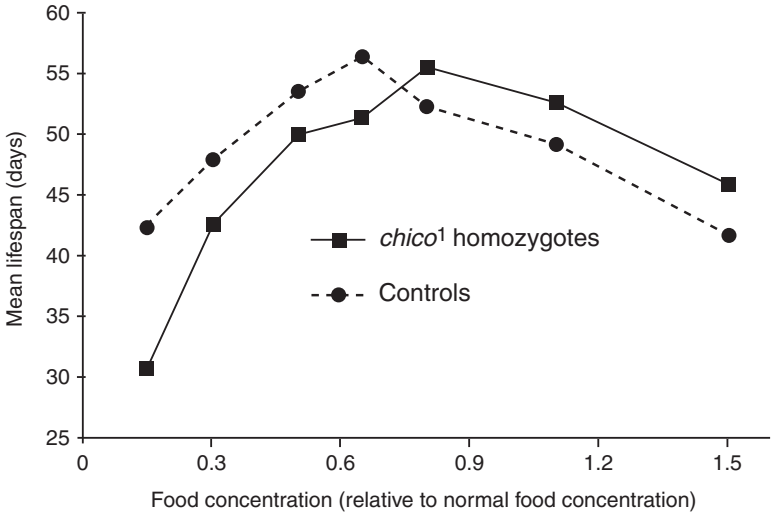


Figure 4.7 Interaction between dietary restriction (dilution of food) and reduced insulin signalling (*chico*[1] mutant, backcrossed into the wild-type strain) on adult lifespan. Food concentration is given as a fraction of normal food concentration. Median lifespan was estimated from a survival curve recorded by monitoring deaths every 2–3 days. Reprinted with permission from Clancy *et al*. (2002). Copyright 2002 AAAS.

does not exert its beneficial effect on lifespan by preventing accumulation of damage. This is in accordance with the situation in mice, where it appears that dietary restriction slows the rate of aging (changing the slope of the survival curve), while mutations in insulin/IGF-1 signalling delay the onset of aging (Bartke *et al*. 2001).

Summarizing, the insulin/IGF-1 signalling pathway is definitely involved in regulation of longevity in *Drosophila*, as it is in nematodes, but the effects in *Drosophila* are slightly more complicated; they seem to depend more on the rate of pathway activity than in *C. elegans*. Also, the effects of dietary restriction on lifespan seem to be less than in *C. elegans* and some aspects of the stress response, for example heat resistance, that are induced by *daf-2* mutation in *C. elegans* are not seen in *chico*[1] in *Drosophila*. Finally, reduced insulin/IGF-1 signalling activity has an effect on lifespan only in female *Drosophila*, but this aspect cannot be compared with *C. elegans* because all experiments in that species have been done with hermaphrodites, not with males.

The evolutionary conservation of longevity regulation is not limited to *C. elegans* and *Drosophila*, but extends to other animals. Holzenberger *et al*. (2003) were inspired by the work on nematodes and fruit flies and investigated the longevity of mice with defective IGF signalling. Whereas *C. elegans* has only one insulin-like receptor, mammals have several; the one that is a homologue of DAF-2 and *InR* is insulin-like growth factor-type 1 receptor, IGF-1R. Holzenberger *et al*. (2003) studied a transgenic mouse in which the essential exon 3 of the *igf-1R* gene was deleted. Controls were included to show that the knock-out allele was transcribed but not translated into a functional product; in the absence of dosage compensation by the wild-type allele the heterozygote had about half the level of IGF-1R protein. This mutant mouse lived 26% longer than the wild type, but the effect was more pronounced in females (33%) compared with males (16%, not significantly different from the wild type). The mutant was also appreciably more resistant to oxidative stress, which was tested by injection of paraquat. Paraquat is a herbicide and a notorious redox cycler, producing large amounts of oxygen radicals, especially superoxide anion, when activated by biotransformation. In addition, Holzenberger *et al*. (2003) noted that the *igf-1R* mutation had hardly any effect on growth and no effect at all on physical activity, food intake, and fertility of the mice. In another study Blüher *et al*.

(2003) studied mice with a fat-specific insulin receptor-knockout and these animals likewise showed an extended lifespan, which in contrast to Holzenberger *et al*. (2003) held for both females and males and was accompanied with reduced adiposity (less fat and a leaner body).

Some aspects of longevity regulation in animals may even be linked to yeast (Guarante and Kenyon 2000; Hekimi and Guarante 2003). A key regulator of aging in yeast is a gene involved in silencing chromatin, called *sir2* (silent information regulator 2). *Gene silencing* is a process by which segments of a chromosome are excluded from transcription. SIR-2 silences specific targets of chromatin by deacetylation of histones (proteins that are packed with the DNA in chromosomes), which introduces a local change in chromatin by which DNA is rendered inaccessible to transcription. In yeast, SIR-2 is involved with the regulation of the sexual cycle, and it mediates the formation of spores. Upregulation of *sir2* leads to a longer lifespan (greater number of cell divisions before the larger (mother) cell shows signs of senescence). A longer lifespan has also been reported for nematodes with a *sir2* duplication (Tissenbaum and Guarante 2001). In addition, it has been shown in yeast that caloric restriction, in the sense of limited glucose supply, increases longevity as well as upregulates *sir2*. Finally, *sir2* is also upregulated in long-lived fruit flies under dietary restriction (Rogina *et al*. 2002).

So, all in all, there is considerable agreement between the three genomic model species of nematode, fruit fly, and mouse, in at least the main aspects of longevity regulation. Some aspects of the system may even extend to yeast. The picture is most complete in *C. elegans*, but without doubt genome-wide surveys of aging-related gene expression will shed more light on similarities across species. The similarities were reviewed by Tatar *et al*. (2003) and Kenyon (2010), and are reproduced here in Fig. 4.8. Upstream of the network there is a sensory mechanism providing input to the central nervous system; this controls the secretion of insulin-like peptides, which suppress the insulin/IGF-1 signalling pathway, starting with *daf-2* in *C. elegans*, *InR* in *Drosophila*, and *igf-1R* in mouse. This then leads to the activation of a forkhead transcription factor, which triggers a great variety of gene expressions and leads to a complicated network, including steroid hormones and signals from the germ line. The final outcome is that growth and reproduction are decreased and anti-aging mechanisms promoted. The division of the body into gonad and soma in this network is crucial (Fig. 4.8).

4.2.4 Trade-off or independent control?

The results discussed above illustrate that the integration of genomics into life-history theory is now in full flight. At the same time, some of the results have come as a kind of shock, because some mutants seem to invalidate a fundamental theorem of life-history theory, the trade-off between life-history attributes. Experimental studies suggest that reproduction, metabolic rate, and longevity can sometimes be uncoupled from each other, which is in flagrant conflict with the idea that there should be a 'cost to reproduction' to prevent Darwin's demon—a hypothetical organism that produces an infinite number of offspring directly after birth and lives forever—from conquering the world (Barnes and Partridge 2003). Is there really such a conflict?

In many of the experimental studies reviewed above there is an obvious negative correlation between longevity and fertility. Figure 4.9 summarizes the data on the different *daf-2* genotypes in *C. elegans* (Leroi 2001). From the viewpoint of life-history theory, it is sufficient to know this relationship and use it to draw inferences about the ultimate effects for the optimal life history. From the viewpoint of ecological genomics, however, it is necessary to understand why this negative correlation is there. It then turns out that decreasing reproduction is not the *cause* of the increased longevity, but that the two processes are regulated by a common mechanism with two different consequences, one usually (but not always) suppressing reproduction, the other increasing stress resistance and (more in females than in males) longevity. Above all, the studies show that the negative relationship in Fig. 4.9 is not due to competition between soma and gonad for a limited resource.

Another new insight into the negative relationship between survival and reproduction has come

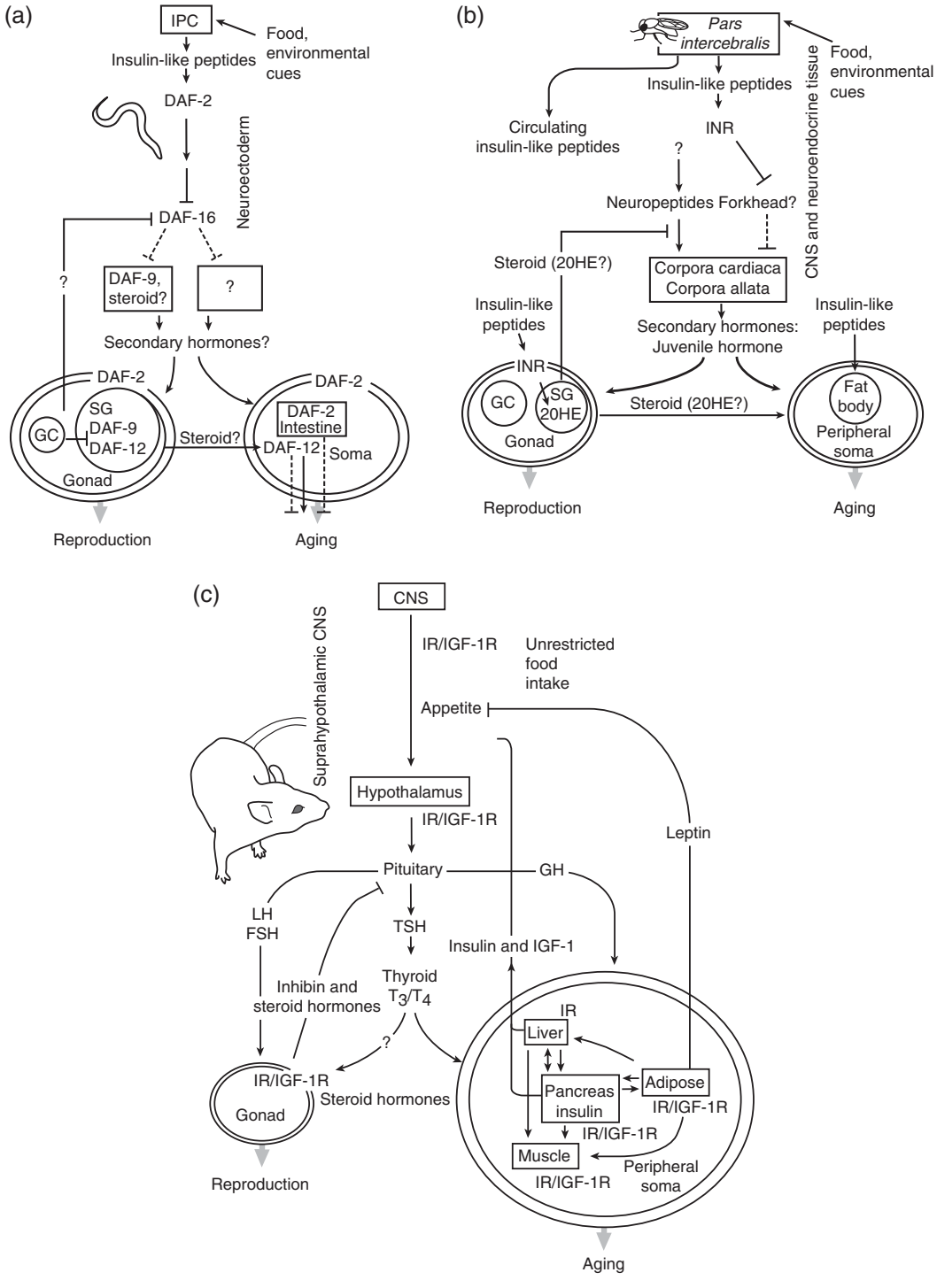


Figure 4.8 Model for endocrine regulation of longevity in (a) *C. elegans*, (b) *D. melanogaster*, and (c) *Mus musculus*, showing how environmental cues are translated into neuroendocrine signals, acting upon an insulin/IGF signalling pathway and triggering a variety of hormonal effects, including steroid responses. The ultimate effect is that different priorities are given to gonad versus soma; that is, reproduction and growth versus aging. CNS, central nervous system; FSH, follicle-stimulating hormone; GC, germ-line cells; GH, growth factor; 20HE, 20-hydroxy-ecdysone; INR, insulin receptor; IPC, insulin-producing cells; IR/IGF-1R, insulin receptor/IGF-1 receptor; LH, luteinizing hormone; SG, somatic gonad tissue; TSH, thyroid-stimulating hormone; T_3, 3,3′,5-tri-iodothyronine; T_4, thyroxine. Reprinted with permission from Tatar *et al*. (2003). Copyright 2003 AAAS.

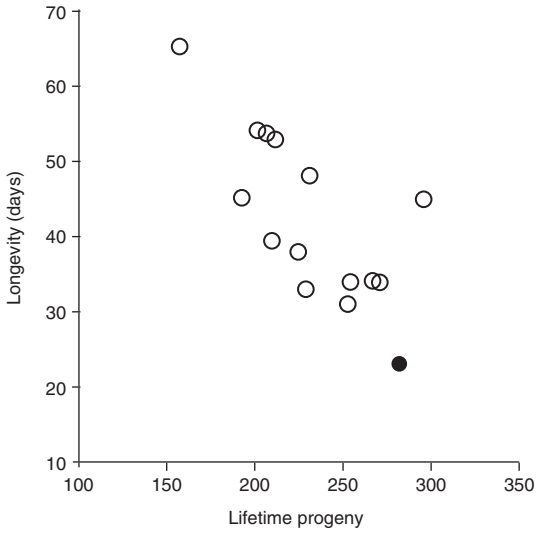


Figure 4.9 Negative correlation between reproduction and median longevity of different genotypes of *daf-2* mutants of *C. elegans*, which have different levels of insulin/IGF signalling. The wild type is indicated by a filled circle. The rather strong negative correlation ($r = -0.75$) between the two processes would be seen as evidence for a trade-off in life-history theory; however, molecular analysis of the mechanism provides no evidence for resource allocation, but rather suggests the action of a hormonal signal with two opposite effects. Reprinted from Leroi (2001), with permission of Elsevier.

from a recent study by Grandison *et al.* (2009). These authors showed that the loss of fecundity in long-lived, dietary restricted *Drosophila* can be restored by adding back amino acids to the diet. The greatest effect was seen for methionine: adding this amino acid to the calorie-restricted diet increased fecundity almost up to the fully fed flies, with no shortening of lifespan. So it seems that longevity and fecundity are both regulated by the diet, but find their optimum at different diet compositions. This again indicates that the simple energy allocation model is insufficient to explain the negative relationship between longevity and fecundity.

Leroi (2001) argued that life-history ecologists too often cannot resist the temptation to postulate some kind of energy allocation wherever they find a negative correlation between two life-history traits. Geoffroy's 'loi de balancement' is assumed to reign whether it applies or not. This conclusion has been challenged, however, by Lessels and Colegrave (2001) and Barnes and Partridge (2003), who argued that the lack of longevity extension seen when the gonad of *C. elegans* is ablated does not negate the presence of an energy-allocation mechanism if resources once mobilized but not exploited for reproduction are lost. In other words, 'removing the bucket will not stop the tap'. Barnes and Partridge (2003) argue that the idea of a cost of reproduction has survived the challenge in this case.

Our feeling is that the issue is mainly semantic. Life-history ecologists tend to use the term trade-off in a general sense, without specifying what kind of mechanism is behind a negative correlation. This is perfectly reasonable because life-history theory seeks explanations in evolutionary terms and looks for ultimate causation, rather than proximate mechanisms. Still, we feel that there should be some understanding between these two worlds and evolutionary explanations should be consistent with the underlying physiology and molecular genetics. In particular, if a trade-off is assumed to imply allocation, it should be made clear what kind of resource is being allocated.

However intricate the genetic or hormonal regulatory pathways, no genomic model organism can escape the energy-conservation laws of thermodynamics. An issue which tends to be neglected in this respect is that in many experiments the actual intake of energy is not measured. This is often due to the difficulty of measuring ingestion and assimilation in animals that take up food from the very substrate in which they live (nematodes, earthworms, fly larvae, flour beetles, etc.). In many allocation arguments the intake of food per unit body mass or surface area is just considered constant, or is assumed to depend only on environmental availability (e.g. prey density). However, food intake is also under control of internal drives. Many animals will eat more when there is a higher internal demand, for example due to the onset of egg production. If this goes unnoticed to the experimenter any energy allocation will be masked.

The question may be asked, why are lifespan mutants not more common in nature, if they live longer and some even maintain normal reproduction? Partridge and Gems (2002b) argued that the life-shortening effect of insulin/IGF-1 signalling could be a negative pleiotropic effect of a fitness

increase at earlier ages, the advantage being that efficient insulin/IGF-1 signalling promotes fertility in *Drosophila* and unarrested development in *C. elegans*. Another part of the answer could lie in the conditions acting upon life-history traits in the wild. Walker *et al*. (2000) demonstrated that long-lived nematodes carrying a mutation in *age-1* quickly disappeared from a mixed culture in which they had to compete with wild-type worms under periodic starvation periods. So there are certainly fitness costs to a long life, but they are not always visible under standard laboratory conditions. Unfortunately little is known about the field ecology of *C. elegans* and this seriously hampers a better understanding of the ecological relevance of life-history mutations (see Section 2.3).

A final remark concerns the ecological relevance of a long life. In almost all organisms in the wild, the lifetime of the average individual falls far below the lifespan seen under protected conditions in captivity or laboratory culture. Everyone who has observed great tit parents flying to and fro all day to feed their nestlings knows what fate awaits the great majority of caterpillars in the surroundings. In an earlier career, studying the demography of field populations of Collembola in forest soils, one of us estimated the fraction of hatchlings that would survive to lay their first clutch as 0.7–6.9% and the life expectancy at hatching as 2.0–3.8 weeks (Van Straalen 1985); the great majority of juveniles disappear into the guts of beetles, mites and spiders. Still, it is not very difficult to keep individuals of the same species alive in the laboratory for more than two years! So do laboratory observations on longevity have any relevance for the field? Kirkwood and Austad (2000) argued that the principal determinant in the evolution of longevity is the level of extrinsic mortality. Animals that live a short life in the wild due to extrinsic sources are expected to live also short (but longer than in the wild) lives if the external sources of mortality are taken away in a cage or laboratory; the pattern of species differences remains the same. So there could indeed be a relationship between aging phenomena observed in the laboratory and the vicissitudes of outdoor life, but it remains quite a challenge for ecological genomics to demonstrate that relationship.

4.3 Gene-expression profiles in the life cycle

The analysis of the Roff (2002) model given at the beginning of this chapter has shown that age at maturation is a very important life-history trait.

It is a well-known fact in life-history theory that when it comes to maximizing intrinsic population growth rate, decreasing the age at first reproduction has the greatest effect, especially when reproductive effort is already high. So it is interesting to explore how expression of genes develops with age and how genomic networks underly the timing of important events in the life history, such as the onset of reproduction. The question is: are there specific expression programmes that mark the various life-history stages?

4.3.1 Developmental stage

Gene-expression profiles tend to change markedly in organisms that develop through distinct stages. This is obvious from work conducted on the model organisms *C. elegans* and *D. melanogaster* (Reinke and White 2002). An exemplary study with nematodes was conducted by Hill *et al*. (2000). These authors isolated oocytes, eggs, juveniles, egg-laying adults, and two-week-old adults (see Fig. 2.19) and profiled the transcriptome of each stage using an oligonucleotide microarray. The statistical technique of 'self-organizing maps' was used to cluster the genes into groups that had similar expression changes throughout the stages. In total, 4221 genes were classified and 36 different clusters were recognized, of which four, A–D, are shown in Fig. 4.10.

It is obvious that each life stage has its characteristic expression profile. In the four groups in Fig. 4.10 cluster A can be considered adult genes, cluster B are larval genes, cluster C are reproduction and egg development genes, and cluster D represents exclusive egg genes. However, none of the clusters, except possibly D, is expressed in one stage only. There are marker genes for each category; for example, *vit-6*, a gene encoding a vitellogenin protein, falls into class A, a cuticular collagen, *dpy-13*, is typical for class B, *mom-2*, required for polarization of the MDS cell (one of the blastomeres in the four-cell

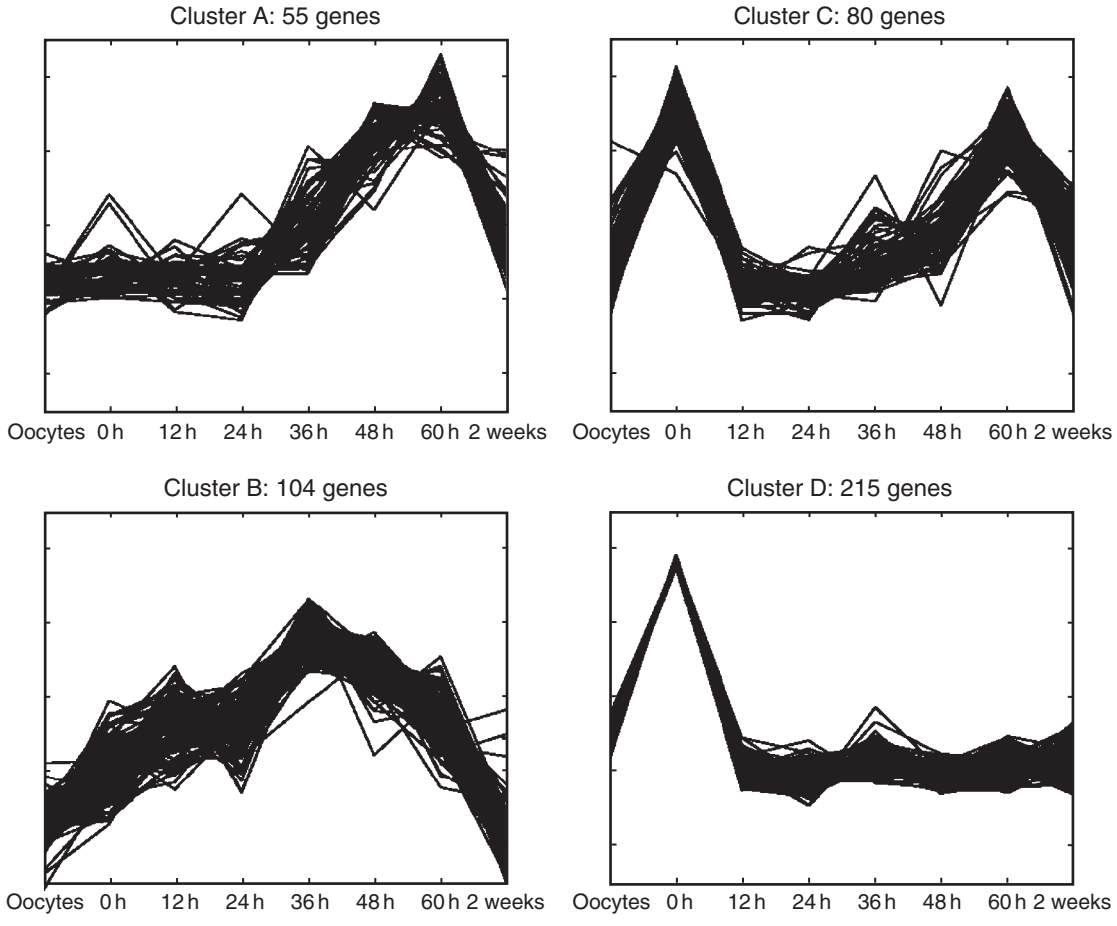


Figure 4.10 Showing changes of gene expression through the life cycle of *C. elegans* for four clusters of genes. The number of genes in each cluster is indicated above the panel. The profiles were all normalized to the same amplitude, so no units are indicated on the *y*-axis. Reprinted with permission from Hill *et al*. (2000). Copyright 2000 AAAS.

stage), is characteristic of class C. Class D appeared to contain several transcription factors and rare messages, but many of these genes had very low expression levels and the data were not considered reliable; nevertheless, class D seemed to contain mostly egg-specific genes (Fig. 4.10). In short, the genes falling into each category reflect the most prominent activity in a life stage.

A similar transcription-profiling study was done by Jiang *et al*. (2001), who used microarrays spotted with PCR fragments from a genomic library rather than oligonucleotide chips. These authors recognized 25 groups of genes with distinct expression patterns across the stages. Particular attention was paid to a group of genes called *cyclins*, which encode proteins involved with cell division. Cyclins are named after their periodic appearance in the cell cycle and they activate cyclin-dependent kinases, which trigger various events in the cell cycle. In *C. elegans*, almost all cell divisions occur at the egg stage, in which the number of cells increases from one to around 700, and in the L4 larva and the adult, in which there is extensive proliferation of cells in the gonads. The *C. elegans* genome encodes seven cyclin genes, which have a similar expression pattern over the stages (Fig. 4.11). They show peak

expressions in the embryo, the L4 stage, and the adult, as expected.

Another perspective on the *C. elegans* data may be obtained by clustering the genes according to their similarity with other genomic models. Hill *et al*. (2000) classified genes as 'core' genes (shared among yeast, nematode, and fruit fly), 'animal' genes (shared between nematode and fruit fly), and 'worm' genes (unique to the nematode). When comparing these gene categories over the life stages the proportion of core genes decreased from egg to adult and the 'worm' genes increased (Fig. 4.12). So, one can say that in the course of development, *C. elegans* becomes more and more a nematode and loses some characteristics of an average animal. The genes expressed in the early stages are most similar to other animals. This pattern is reminiscent of Ernst Haeckel's *biogenetic law*, which states that 'ontogeny recapitulates phylogeny'; reformulated in genomic wording the law may read that genes expressed in early development tend to be more evolutionarily conserved than genes expressed in adult life. It would be interesting to explore whether similar tendencies are valid for other species; with the availability of more and more genome-wide stage-specific expression profiles, a thorough test of the pattern discovered by Hill *et al*. (2000) would be feasible.

Also in *Drosophila*, a considerable part of the gene complement shows stage-specific gene expression. Here comparisons were made between eggs, larvae, pupae, and adults (Arbeitman *et al*. 2002). Interestingly, many genes in *Drosophila* appear to be expressed in two waves during development, with embryonic expressions being recapitulated in pupae, and larval expressions recapitulated in adults. Judging from their expression profiles, pupae are like eggs and adults are like larvae. This is understandable from the processes going on in the various stages. The pupa is a stage of intense reorganization of the body, in which most of the larval tissues degenerate and many adult tissues are newly developed from the imaginal discs, processes which are comparable with the extensive cell differentiation in the egg, and apparently involve partly the same genes. The larval stage is characterized by extensive energy acquisition and growth,

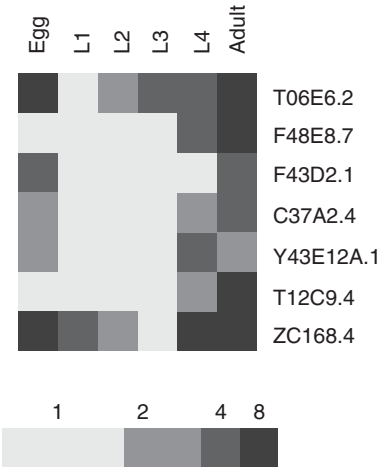


Figure 4.11 Expression of seven cyclin genes through the life stages of *C. elegans*. Expressions were assessed with a microarray spotted with PCR products from a genomic library, and hybridization was relative to cDNA from a mixed-stage population. After Jiang *et al*. (2001), by permission of the National Academy of Sciences of the United States of America.

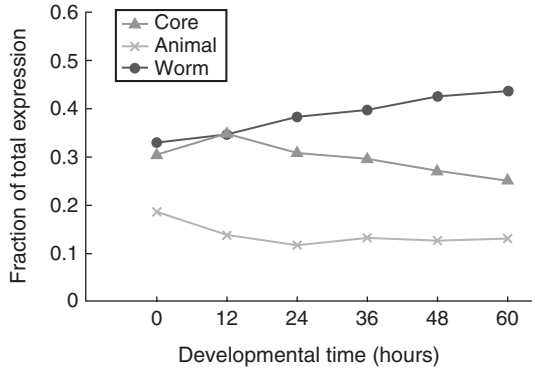


Figure 4.12 Fraction of genes assigned to three categories ('core', shared among yeast, nematode, and fruit fly; 'animal', shared between nematode and fruit fly; 'worm', unique to nematode) among 4221 genes showing differential expression during the life cyle of *C. elegans*, as detected by oligonucleotide microarray transcription profiling. Reprinted with permission from Hill *et al*. (2000). Copyright 2000 AAAS.

which is parallelled in the adult by energy acquisition directed to reproduction.

Arbeitman *et al*. (2002) also showed that in the early egg stage a very large proportion of the transcripts are maternal; that is, they are deposited by

the mother during oogenesis. Most of these maternal transcripts degrade during the first hours of embryogenesis, but some persist to the first larva. Maternally derived messengers indicate a mechanism by which the mother can direct the development of her offspring in a non-genetic way. Such *maternal effects* are very common in insects. Well-known examples can be found in the mechanisms of developmental plasticity in response to seasonal change in temperate climates. The photoperiod experienced by an ovipositing female insect often determines the likelihood that the offspring will go into diapause. This is achieved by adding maternal factors to the egg that will direct its development into a diapausing stage. Such phenomena are a well-known source of annoyance to entomologists doing quantitative genetics experiments, because maternal effects confound any estimation of heritability from parent–offspring comparisons. However, maternal effects are now increasingly recognized as being shaped by natural selection to act as a mechanism by which maternal experience is translated into increased fitness of the offspring (Mousseau and Fox 1998). The study of Arbeitman *et al*. (2002) shows how powerful the mother can be in dominating the transcriptome of the early egg.

The diversity of life-cycle transcript changes is further expanded by expression of different combinations of exons of the same gene in different life-history stages (*exon-specific expression*). Stolc *et al*. (2004) profiled the transcriptome of *D. melanogaster* eggs, larvae, pupae, and adults using an oligonucleotide microarray with no less than 179 972 36-mer probes. In this microarray, exons of the same gene were targeted by different probes, allowing the researchers to profile not only the fluctuations of gene expression of individual genes, but also the changes in expression of exons from the same gene. This is especially relevant if genes undergo *alternative splicing*; that is, the mature mRNA is composed of a subset of exons from the primary transcript, depending on the conditions. Using exon-specific probes on the microarray, Stolc *et al*. (2004) could indicate several genes for which the abundance of exons was not the same in the different stages. For example, for one gene designated as CG8946, exon 1 showed a peak in larvae, whereas exon 4 peaked in pupae, and four other exons were expressed synchronously (Fig. 4.13b). For some genes the expression of different exons was even anticorrelated (Fig. 4.13c). It is likely that such stage-dependent exon-specific expressions are crucially important in the developmental process of each stage. These results imply, quite disquietingly, that a vast amount of variation in gene expression may be missed in microarray studies that use cDNAs or that assess only a subset of exons from each gene.

4.3.2 Diapause

Many organisms have a dormant stage in which they shut down activities such as locomotion and develop tolerance against adverse conditions. The best-known example of life-cycle dormancy is the phenomenon of diapause in insects with seasonal phenologies. Depending on the species, diapause may apply to the egg, larva, pupa, or adult. Are there specific gene-expression patterns characteristic of diapause stages? Surprisingly, this question can only be answered in a rudimentary sense.

Although there is a tremendous amount of literature on environmental regulation of diapause (especially photoperiod), the circadian clock, and the hormonal changes associated with diapause entry and termination, the molecular genetics of diapause have received little attention (Denlinger 2002). This may be due to the fact that *D. melanogaster* is not the best model for studying diapause, since in this species diapause consists of only a weak reproductive arrest. Better models are the silkworm, *Bombyx mori*, which has a facultative egg diapause, and the flesh fly, *Sarcophaga crassipalpis*, which has pupal diapause. Many agricultural pest species have been used for diapause research, for obvious reasons.

Photoperiod is an important driver of diapause entry in many insects, since it is an accurate predictor of seasonal change. It is widely assumed that the circadian clock provides the mechanism by which insects can perceive the photoperiod. Photoperiodic clock genes are known from the *Drosophila* genome and several of them have been sequenced in other insects such as *Bo. mori*, the tobacco hornworm *Manduca sexta*, and *S. crassipalpis* (Goto and Denlinger 2002). The gene *period* (*per*) has been

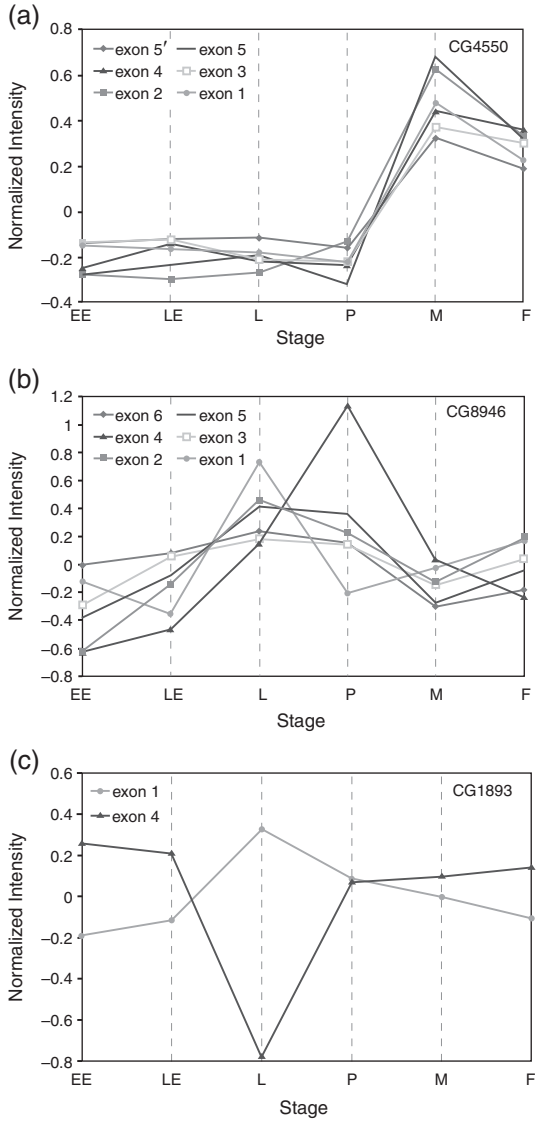


Figure 4.13 Exon-specific expression profiles for three genes over the life stages of *D. melanogaster* (EE, early egg; LE, late egg; L, larva; P, pupa; M, male; F, female), assessed using a microarray in which different oligonucleotide probes targeted different exons of the same gene. Expression was normalized by the mean and the standard deviation. (a) In gene CG4550 (*ninA*, a G-protein-coupled photoreceptor expressed in the eye) all exons are expressed synchronously. (b) In gene CG8946 (*Sly*, a sphingosine-1-phosphate lyase involved with phospholipid metabolism) exon 1 peaks in larvae and exon 4 peaks in pupae. (c) The two exons of gene CG1893 (encoding a product with phospholipid scramblase activity) show anticorrelated expression over the stages. Reprinted with permission from Stolc *et al*. (2004). Copyright 2004 AAAS.

examined in relation to diapause, but how the circadian clock triggers the entry into diapause is not clear at all. It turns out that *Drosophila* carrying a null mutant of *per* can enter diapause just as well as the wild type, although they do not have a circadian rhythm.

Denlinger (2002), drawing together a large number of mostly pre-genomic studies, provides an overview of genes that have their expression regulated by diapause. Table 4.3 gives a summary of this information. Genes are subdivided into six categories, depending on whether they are up- or downregulated early or late in diapause, or not modulated at all. The great majority of genes are downregulated, although some are specifically upregulated. This demonstrates that, like the dauer stage of *C. elegans*, insect diapause does not imply a simple shutting down of the genome but involves an active and specific programme of gene regulation. This is also evident from the fact that in individuals destined for diapause the stage preceding the diapause is longer than normal. This stage includes physiological preparations and sometimes specific behaviours aimed at locating suitable microhabitats for hibernation or aestivation.

Some of the differential expressions in Table 4.3 could be due to master genes regulating the various processes during diapause; other genes may act downstream of regulatory cascades, being up- or downregulated as a consequence of master genes. Still other genes have specific functions in the diapause, such as those related to cold-hardiness, defence against microbial infection, and so on. Some genes are expressed intermittently during diapause, possibly in response to the periodic boosts of respiration triggered by pulses of juvenile hormone.

Table 4.3 shows that the picture of diapause-specific gene expression in insects is still incomplete. Some gene expressions are reminiscent of dormancies in other species. This is especially applicable to the upregulation of heat-shock protein 70, which is consistently reported in insect diapause and is similar to the pattern seen in the *C. elegans* dauer larva (see above) and several other organisms, even fungi. The universal involvement of heat-shock proteins in dormancy and life extension is striking and must relate to a very fundamental role of these proteins in

the cell (see also Section 5.2). In insect diapause, not all heat-shock proteins are upregulated, however: some are not regulated at all, and *Hsp90* is actually downregulated (Table 4.3).

4.3.3 Adult life and sex

After the attainment of adulthood, gene-expression profiles tend to change only little. This is evident both in *C. elegans* (Lund *et al*. 2002) and in *D. mela-*

Table 4.3 Overview of gene products regulated by diapause in insects

Gene or gene product	Function	Species
Not influenced by diapause		
Heat-shock cognate protein 70 (*hsc70*)	Cellular stress response	*S. crassipalpis, C. fumifera*
28 S ribosomal protein	Protein synthesis	*S. crassipalpis*
Cyclins E, p21, and p53	Cell cycle	*S. crassipalpis*
Glutathione S-transferase	Detoxification	*C. fumiferana*
Ubiquitin	Cellular stress response	*C. fumiferana*
Downregulated throughout diapause		
Mitochondrial phosphate transport protein	Respiration	*C. fumiferana*
Midgut enzymes	Food assimilation	*L. dispar*
Heat-shock protein 90 (*hsp90*)	Cellular stress response	*S. crassipalpis*
Proliferating cell nuclear antigen (*pcna*)	Cell cycle	*S. crassipalpis*
cdc2-related Ser/Thr kinase	Cell cycle	*Bo. mori*
Upregulated throughout diapause		
Heat-shock protein 23 (*hsp23*)	Cellular stress response	*S. crassipalpis*
Heat-shock protein 70 (*hsp70*)	Cellular stress response	*S. crassipalpis, R. pomonella, O. nubilalis*
Heat-shock protein 70A (*hsp70A*)	Cellular stress response	*L. decemlineata*
E26 transforming sequence (ETS) protein	Transcription factor	*Bo. mori*
7.9 kDa peptide	Function unknown	*G. atrocyanea*
Alkaline phosphatase	Various functions	*L. dispar*
Upregulated in early diapause, downregulated in late diapause		
pScD41, clone from brain-enrichment library	Function unknown, possibly a retrotransposon	*S. crassipalpis*
Samui	Cold-induced activator of sorbitol dehydrogenase	*Bo. mori*
55 kDa protein in gut	Function unknown	*L. dispar*
Downregulated in early diapause, upregulated in late diapause		
Ultraspiracle (*usp*)	Ecdysone receptor	*S. crassipalpis*
Sorbitol dehydrogenase (*sdh*)	Conversion of sorbitol into glycogen	*Bo. mori*
Defensin	Microbial defence	*C. fumiferana*
45 kDa actin-like protein in brain	Function unknown	*L. dispar*
Genes expressed intermittently during diapause		
60 S ribosomal protein PO	Protein synthesis	*S. crassipalpis*

Notes: Species investigated were *S. crassipalpis* (flesh fly; Diptera, Sarcophagidae), *Choristoneura fumiferana* (spruce budworm; Lepidoptera, Tortricidae), *Lymantria dispar* (gypsy moth; Lepidoptera, Lymantriidae), *Bo. mori* (silkworm; Lepidoptera, Bombycidae), *Rhagoletis pomonella* (apple maggot fly; Diptera, Tephritidae), *Ostrinia nubilalis* (European corn borer; Lepidoptera, Pyralidae), *Leptinotarsa decemlineata* (Colorado potato beetle; Coleoptera, Chrysomelidae), *Gastrophysa atrocyanea* (leaf beetle; Coleoptera, Chrysomelidae).

Source: After Denlinger (2002).

nogaster (Zou *et al*. 2000; Jin *et al*. 2001). Applying rigid statistical criteria, Lund *et al*. (2002) found only 164 genes, representing about 1% of the *C. elegans* genome, to change from the first day of adult life to senescence (19 days). Many of these genes were related to stress resistance. The 27 heat-shock genes on the microarray all showed a common expression profile, rising over the first part of adult life and decreasing later in life.

Another category of differential expressions consisted of insulin-like proteins. Three of the insulin genes increased in expression, but one, *Ins-7*, showed a decreased expression during life. *Ins-7* was identified in the study of Murphy *et al*. (2003) as belonging to a group downregulated by DAF-16 and contributing to a positive-feedback loop in the insulin/IGF-1 signalling signalling pathway (see Section 4.2). Decreased *Ins-7* expression during life, as found by Lund *et al*. (2002), is consistent with upregulation of stress-responsive genes by increased DAF-16 activity. So it seems that normal aging in *C. elegans* is not accompanied by a specific genetic programme. The changing expression profiles seem to indicate a response to accumulating cellular damage from reactive oxygen species and unfolded proteins, rather than being part of the developmental repertoire.

Also in *Drosophila*, the effects of age on gene expression in normal (wild-type) flies tend to be weak. Zou *et al*. (2000) measured genome-wide changes in expression of flies between 3 and 50 days of age, using an EST microarray. A total of 127 genes were found to be regulated by age, and these were classified according to three categories: genes associated with reproduction, metabolic genes, and stress-responsive genes. In all categories there was a trend towards decreasing expression with age, but among the stress-responsive genes some were upregulated. In a similar study, Jin *et al*. (2001) compared 1 and 6 week-old flies of different sexes and from different strains using replicated microarray hybridizations. Expression profiles were mostly affected by sex, less so by strain, and only weakly by age. Analysis of variance showed that sex, in combination with the sex × strain interaction component, accounted for between 60 and 90% of the variation in gene expression. Some of the genes that were age-related in the study of Zou *et al*. (2000) were confirmed as age-related by statistical significance in the study of Jin *et al*. (2001), but several others were not. The large influence of sex in this study is a notable result. Much of the sex bias can probably be attributed to reproductive activities that are naturally different between male and female.

In *C. elegans* most studies are done with hermaphrodites, but Jiang *et al*. (2001) compared expression profiles of hermaphrodites and males. No fewer than 2171 genes (12% of all the genes in the genome) were found to be sex-regulated. About half of the male-specific genes were expressed in the gonad and the sperm, and the other half were expressed in the soma. Male *C. elegans* have an elaborate mating behaviour to find the hermaphrodite's vulva with their tails and to ejaculate sperm into it. In conjunction with this behaviour, males have 115 neurons and neuronal support cells not found in the hermaphrodite, whereas hermaphrodites have only eight sex-specific neurons. Sex-specific expression patterns of the whole animal may partly be due to genes expressed specifically in these cells. One of the groups of sperm-enriched genes is found in the L4 larva, which is the major sperm-producing stage of the hermaphrodite.

The sex bias in gene expression in *Drosophila* and *C. elegans* is corroborated by the fact that on the phenotypic level strains of the same species tend to be more different in one of the sexes than in the other. On a larger scale, the same phenomenon is seen in the differences between closely related species of fruit fly: such differences tend to be larger in males compared to females. A comparison of gene-expression profiles of the sibling species *Drosophila simulans* and *D. melanogaster* showed that 50% of the genes that differed in expression between the species did so in a sex-dependent manner. Almost all the genes that differed by more than a factor of four were male-specific genes (Ranz *et al*. 2003). These data underline the importance of sex-dependent selection in the differentiation of the *Drosophila* species cluster.

The relatively minor age-specific expression changes observed in the studies reviewed above are more or less at variance with a detailed study by Pletcher *et al*. (2002) on transcript profiles in aging *D. melanogaster*. Although these authors analysed only females, so the effects of age cannot be weighed

against the effects of sex, the differential expressions they observed indicated that nearly 23% of fruit fly genes were modulated by age, a figure much higher than reported by Jin *et al*. (2001) and Zou *et al*. (2000), and by other studies on rodents and rhesus monkeys. There is, however, some degree of similarity across studies in the types of genes regulated by age. Pletcher *et al*. (2002) found decreasing expression for genes involved in chorion formation, which is obviously due to reproductive senescence, and upregulation of stress-response genes such as cytochrome P450, as in the other studies. Interestingly, aging was also associated with increased expression of antibacterial peptides. This suggests that microbial infection is an important factor in the life of *Drosophila* in laboratory culture.

Whereas some of the age-specific expression profiles are only valid for certain model organisms, and maybe only under specific culturing conditions, ecological genomics of life histories should aim at finding regulatory signatures of life-history events that are of more general validity. Such signatures can be detected when expression profiles are compared across species. One of the first studies to use this approach was by McCarroll *et al*. (2004). These authors analysed microarray data from *C. elegans* and *D. melanogaster*, in which expressions were compared between two stages, young adults (0 days for *C. elegans*, 3 days for *D. melanogaster*) and mature adults (6 days for *C. elegans*, 23 days for *D. melanogaster*). The measurements were then paired systematically between orthologous genes from the two organisms. The authors were interested in genes with the same expression ratio in the two species; that is, genes which if downregulated with aging in *Drosophila* are also downregulated with aging in *C. elegans*. The methodology is illustrated in Fig. 4.14.

Classifying the genes by Gene Ontology categories, McCarroll *et al*. (2004) identified three categories of shared transcriptional profiles with decreasing expression during adult life:

- genes involved in oxidative metabolism (respiratory chain, citric acid cycle),
- genes involved in catabolism (peptidases) and DNA repair, and
- genes involved in molecular transport functions, such as ion transporters and ABC transporters.

So, the shared transcriptional profile of aging seems to involve a decreased commitment to energy generation and active movement of ions, transmitters, and nutrients. Interestingly, these changes are implemented abruptly early in adult life, and few further changes occur later in life. So it seems that, if there is a developmentally timed transcriptional regulation of aging, it is a programme that is associated with the onset of adulthood. What we see later in life, for example upregulation of stress-response genes in the studies reviewed above, is more a consequence than a cause of aging.

The type of *comparative functional genomics* applied by McCarroll *et al*. (2004) is a promising new strategy for identifying the general genomic basis of life-history events and removing some of the noise and contingency that (not unlike ecological studies) seems to be inherent in many microarray-based transcription profiles.

4.3.4 Flowering time in *Arabidopsis*

The developmental patterns of plants show considerably more flexibility than those of animals. Most of the development process in plants occurs postembryonically through the action of shoot and root apical meristems. Consequently, there is ample opportunity for plants to adapt their morphology to specific environmental conditions, whereas the body plan of animals is more or less fixed and only certain aspects of it may depend on the environment. One of the major developmental transitions in the life of a plant is the switch from vegetative growth to reproductive development. The timing of this switch in relation to environmental conditions is of crucial adaptive value; fitness will be lost if reproduction and seed set fall in an unfavourable season or are not synchronized with other members of a population in out-crossing species. This implies that flowering time, also called *bolting time*, and especially the way in which it is modulated by environmental cues, is an important life-history trait optimized by natural selection.

The major model for the genomics of flowering time is *A. thaliana*. There is some work on other Brassicaceae and on crops such as corn and rice, but relatively little in comparison with *Arabidopsis* and

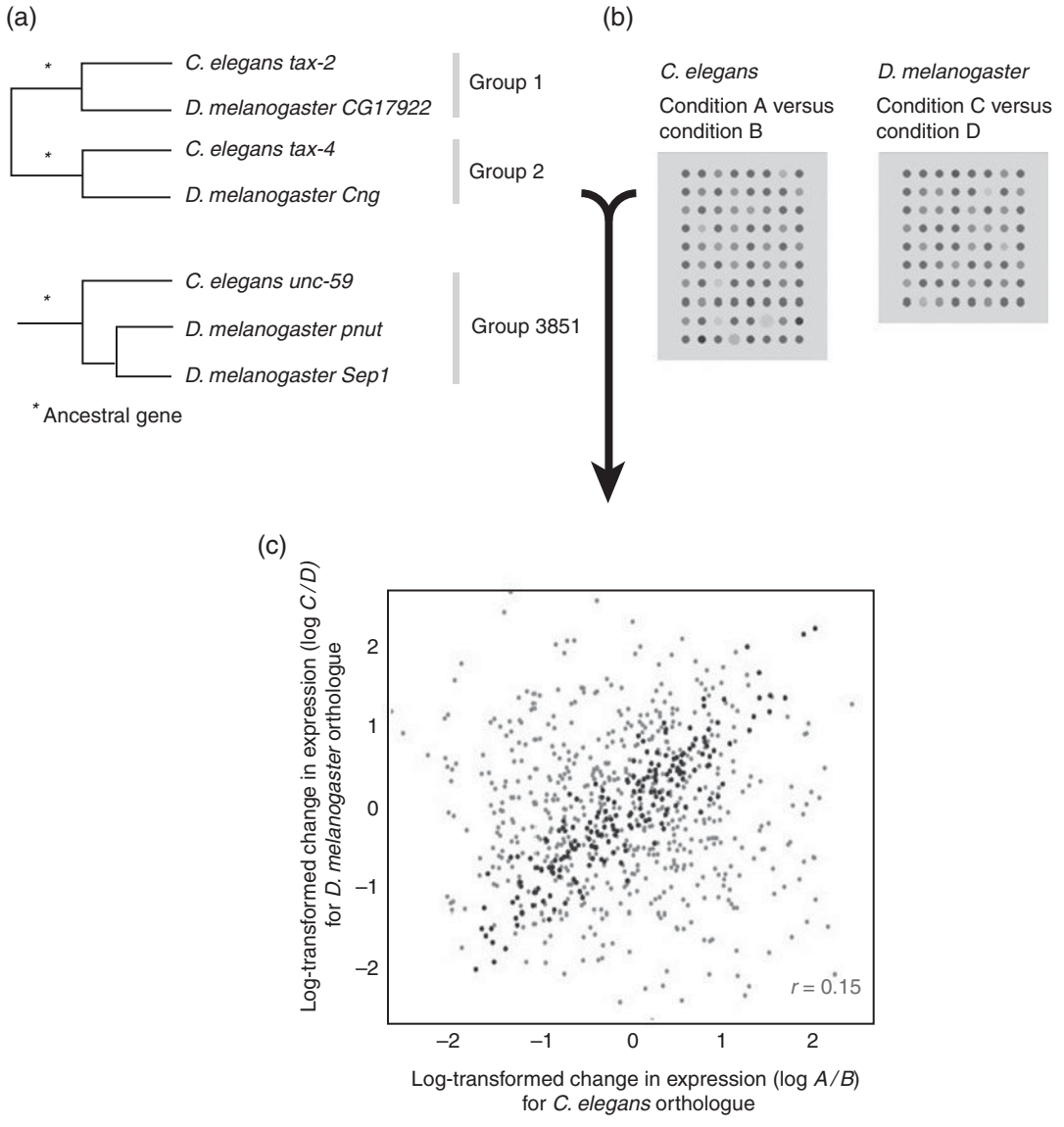


Figure 4.14 Illustrating the principle of comparative functional genomics, as applied by McCarroll *et al*. (2004) to identify transcriptional profiles of aging, shared across *C. elegans* and *D. melanogaster*. (a) Phylogenetic analysis identifies orthologous pairs between the species. If more than one paralogue is present within a species, the most conserved orthologous gene pair is identified. (b) Transcription profiling with microarrays is applied to reveal the relative change in expression when comparing two conditions (e.g. young and old age). (c) Groups of genes are selected according to the Gene Ontology annotation system and the expression ratio for the two species is plotted in a correlation diagram. If there is a significant correlation, the group is said to have a conserved regulation and may be considered an expression signature of the treatment applied in the two species. From McCarroll *et al.* (2004), by permission of Nature Publishing Group.

therefore we will focus initially on this model. Despite its present widespread distribution (Fig. 2.23), *Arabidopsis* has its origin in a northern, seasonal climate and so the timing of the reproductive switch with respect to daylength and temperature is a crucial aspect of its life history. *A. thaliana* is a facultative long-day plant, which implies that bolting is stimulated by the long days of early spring. Bolting also requires *vernalization*: a prolonged period of cold (3–8 weeks at 4 °C or lower), necessary to allow flowering in spring.

Through the analysis of mutants a complicated network of some 50 genes has been identified by which flowering time is regulated (Putterill *et al.* 2004; Flowers *et al.* 2009; Ehrenreich *et al.* 2009). Interestingly, most of the gene products act as regulatory proteins, for example transcription factors, RNA binding proteins, signal transducers, and kinases. The number of genes involved illustrates that flowering time is an extreme example of polygenic control. Genetic research has shown that the flowering-time genes fit into four regulatory networks (Koornneef *et al*. 1998; Mouradov *et al*. 2002; Ratcliffe and Riechmann 2002; Simpson and Dean 2002; Putterill *et al*. 2004), which will be discussed in detail below.

The 'default' life cycle of *Arabidopsis* seems to be a direct floral transition early in life, and this can be achieved by mutations in some of the flowering-time genes. Such 'rapid-cycling' varieties are useful under laboratory conditions, but in the wild type the floral transition is actively repressed. The gene *FLOWERING LOCUS C* (*FLC*) is a crucial element of this repression; it encodes a MADS-box transcription factor, which regulates several floral-identity genes in the apical meristem. Consequently, conditions that stimulate bolting and shorten flowering time must remove the repressive action exerted by the *FLC* locus. This way of regulation, in which an environmental signal removes a repressor, is thought to be more stable and more specific than a system in which the environmental signal acts as a direct positive regulator (Casal *et al*. 2004). We saw above that a similar system of double-negative regulation governed the dauer transition in *C. elegans*.

The MADS-box genes, to which *FLC* belongs, are a large family of very old transcription factors, which have diversified in plants more than in animals (Becker and Theißen 2003). They derive their name from the *MADS box*, a highly conserved 180 bp consensus sequence, which like the homeobox in the *Hox* genes encodes a DNA-binding domain. There are two types of MADS-box gene, type I and II, which are both represented in plants, fungi, and animals, and which are assumed to derive from a very old duplication in the stem of the eukaryotes, more than 1000 million years ago. Little is known about the type I genes in plants; type II genes have diversified considerably and have given rise to genes that control all the developmental processes in plants, such as the ontogeny of roots, flowers, seeds, and fruits. Loss-of-function mutations in MADS-box genes cause homeotic transformations (replacing, for example, sepals with carpels or petals with stamens), indicating that these genes act to determine the identity of a meristem or a primordium. The type II MADS-box genes in plants belong to two families, MIKC*-type genes and MIKC^c-type genes, where MIKC^c is subdivided into four subfamilies (Becker and Theißen 2003). *FLC* belongs to one of the MIKC^c subfamilies.

As indicated above, four interacting pathways control flowering time in *A. thaliana*: the photoperiod and light-quality pathway, the vernalization-response pathway, the autonomous pathway, and the gibberellin signalling pathway. These pathways are interconnected and converge on *FLC* and other flowering regulators such as *FT* and *SOC1* to activate the floral-identity genes *AP1* and *LFY* (Fig. 4.15).

The *photoperiod-response pathway* integrates information from the daylength to promote flowering in response to long days. A crucial gene in this pathway is *CO* (*CONSTANS*), named after the mutant that confers insensitivity to daylength. CO induces early flowering by activating *FT*, which regulates the floral-identity gene *AP1* (Fig. 4.15). The expression of *CO* is influenced by the circadian clock, a system of autoregulated genes with feedback loops such that clock proteins regulate their own expression and appear in an oscillating fashion. The clock is entrained by photoreceptors such as phytochromes (*PHYA* and *PHYB*) and cryptochromes (*CRY1* and *CRY2*). Entrainment ensures that the

period of the circadian rhythm is synchronized with the daily cycle of light and dark. The circadian feedback loop generates a series of oscillating outputs including rhythmic expression of *CO*. The peak of expression falls in the second half of the day; however, to exert its action, the CO protein must be stabilized and this occurs only in the light. CO is stabilized by PHYA in far-red light and by CRY1/CRY2 in blue light (Valverde *et al*. 2004). This recent insight into the post-translational regulation of CO provides an explanation for the *external coincidence model*, which holds that photoperiodic responses are regulated by a signal with circadian expression that must fall in the light to trigger the response. In the case of *CO*, the peak of expression at the end of the day coincides with the light only on long days; CO is then stabilized and may act upon *FT*. In short days, the peak of *CO* expression falls in the dark, the CO protein is then unstable, *FT* expression is not upregulated, and flowering is delayed (Yanovsky and Kay 2003; Hayama and Coupland 2004; Putterill *et al*. 2004; Schepens *et al*. 2004).

A surprisingly large part of the genome of *A. thaliana* undergoes day–night oscillations. A genome-wide transcription-profiling study by Harmer *et al*. (2000) classified 6% of the probes on an oligonucleotide gene chip as cycling with a period of between 20 and 28 h. Four categories were recognized: (i) light-harvesting centres, photosynthesis genes, phytochromes, and cryptochromes; (ii) genes involved with photoprotective pigment pathways, such as flavonoids and anthocyanins; (iii) enzymes involved in resistance to chill and drought, for example catalysing lipid modification; and (iv) enzymes of the carbon, nitrogen, and sulphur pathways. The authors were also able to identify a conserved 9 bp

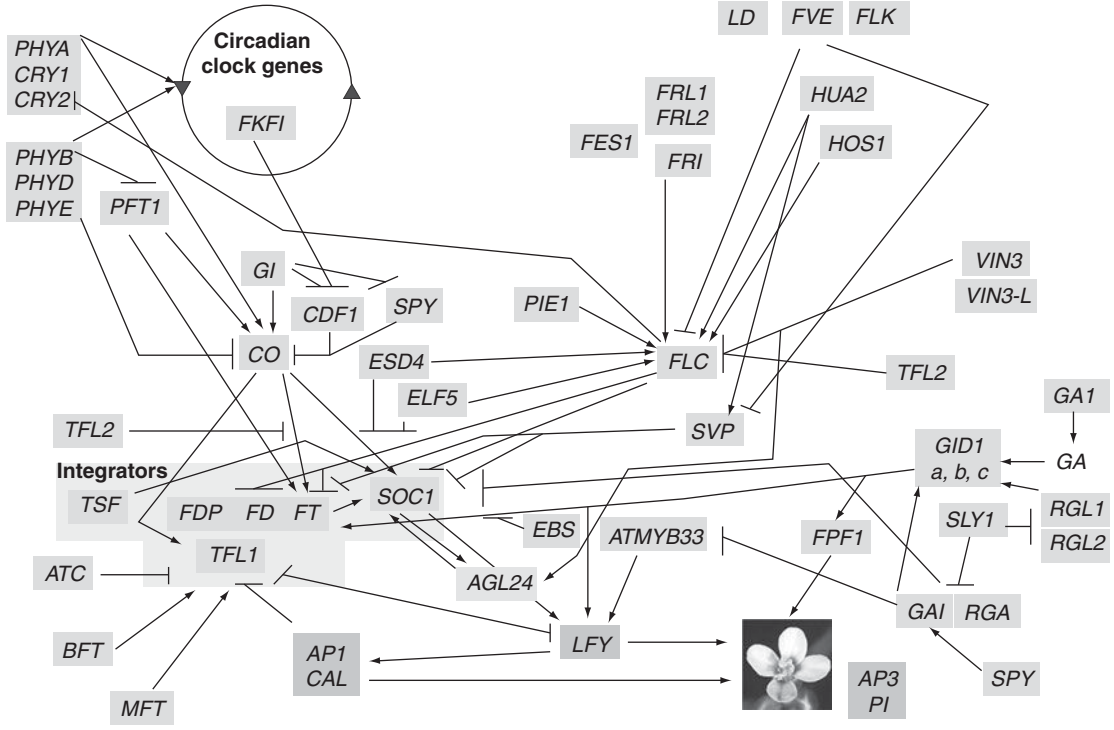


Figure 4.15 The genetic network for the regulation of flowering time in *A. thaliana*. The network can be seen as consisting of four interacting pathways: the photoperiod pathway in which *CO*, *CONSTANS* plays a key role, the vernalization and autonomous pathways converging on *FLC*, *FLOWERING LOCUS C*, and the GA signalling pathway involving *SOC1*, *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* and *FPF1*, *FLORAL PROMOTING FACTOR 1*. Reproduced from Ehrenreich *et al.* (2009), by permission of the Genetics Society of America.

motif in the promoters of 35 cycling genes, which suggests that these genes are regulated by the same transcription factor. A similar study was reported by Schaffer *et al*. (2001), who used a microarray with 11 521 *Arabidopsis* ESTs showing that 11% of the genes were expressed diurnally. Obviously, such widespread rhythmicity in the genome reflects the pervasive influence of light on the metabolism of phototrophic organisms; at the same time, it reminds experimenters how important it is to standardize the time of the day when harvesting RNA in studies looking at gene expression in plants.

The *vernalization pathway* for the control of flowering time integrates information from the past temperature regime. A crucial gene in this pathway is *FRI* (*FRIGIDA*). This gene promotes the transcription of *FLC* and so represses the floral transition. Mutations in *FRI* remove this positive regulation and cause loss of the vernalization requirement. The *FRI* locus is particularly relevant in the ecology of *Arabidopsis*, because it shows natural variation associated with flowering time (Flowers *et al.* 2009). When a comparison is made of flowering times of *Arabidopsis* accessions of different geographic origin, plants from low latitude tend to flower earlier in a common garden than plants from high latitude. However, this latitudinal cline is only found in ecotypes that have a functional *FRI* gene (Stinchcombe *et al*. 2004), suggesting that *FRI* mediates the transduction of latitudinal information into regulation of flowering time.

An important issue in understanding the response to vernalization has been the question of how vernalized plants can remember a cold treatment and flower several weeks later, even when temperatures are higher for some time after the cold signal. It turns out that cold treatment induces an altered state in the shoot apex which can be passed on through mitotic cell divisions, even in the absence of cold. Recent research has shown that gene silencing due to changes in chromatin structure is the basis of this cellular memory (Bastow *et al*. 2004; Sung and Amasino 2004).

We know from basic biochemistry that in eukaryotic chromosomes the DNA double helix is wound around groups of small globular proteins, histones, forming nucleosomes. These histones have tails protruding outward from a nucleosome, in which the lysine residues are normally acetylated. Because histones are highly positively charged proteins, acetylation of the tails is necessary to prevent the formation of aggregates of nucleosomes and to maintain a loose chromatin structure in which DNA is accessible to transcription.

Sung and Amasino (2004) identified a regulatory gene in *Arabidopsis* called *VERNALIZATION INSENSITIVE 3* (*VIN3*), which, in conjunction with two other vernalization genes, *VNR1* and *VNR2*, inactivates *FLC* by local deacetylation of histones. *Histone deacetylation* causes condensation of chromatin and shuts off DNA from transcription. This process is catalysed by histone deacetylase complex (HDAC), a cluster of molecules involving a DNA-binding protein and an acetyltransferase enzyme. The condensed state of chromatin is transferred to the daughter cells when cells divide and so provides an epigenetic mechanism of inheritance.

VIN3 is the most upstream component of the vernalization pathway identified so far, but it is still not known how the protein senses cold. Sung and Amasino (2004) suggest that VIN3 might be a receptor for phosphoinositides (a group of phospholipids) in the nucleus, and could perceive changes in the composition of these compounds that occur during cold exposure.

The third pathway for regulating flowering time in *Arabidopsis*, the *autonomous pathway*, integrates information from the developmental stage of the plant. In the default state it represses *FLC* like the vernalization pathway (Fig. 4.15). *Arabidopsis* mutants of the autonomous pathway are early-flowering but retain the photoperiodic response, and so the flowering signal is independent of environmental cues (hence the use of the term autonomous). However, work by Blázquez *et al*. (2003) has shown that expression of genes in the autonomous pathway depends on temperature; this would represent a system of thermosensory control of flowering time that acts in parallel to vernalization. The autonomous pathway involves a group of six different genes, which act upon *FLC* in two different ways. One of the mechanisms involves inactivation of *FLC* by histone deacetylation, like in the vernalization pathway (He *et al*. 2003). This is mediated by a locus *FLD*, which encodes a protein that forms

part of an HDAC. Another gene product of the autonomous pathway, FVE, is probably part of the same HDAC (Amasino 2004; Ausín *et al*. 2004; Kim *et al*. 2004). In a series of elegant experiments He *et al*. (2003) were able to show that a specific region in the first intron of *FLC* acted as a binding site for the HDAC (Fig. 4.16). Mutations in *FLD*, as well as deletion of a 294 bp region from intron 1, prevented binding of HDAC to the *FLC* gene, causing continued transcription of *FLC* and late flowering.

The second regulatory input on *FLC* from the autonomous pathway comes from the floral-promotion genes *FCA* and *FY*. The way in which these genes regulate *FLC* expression represents another complicated but beautiful example of gene regulation underlying life-history traits (Eckardt 2002; Macknight *et al*. 2002; Amasino 2003; MacDonald and McMahon 2003; Quesada *et al*. 2003; Simpson *et al*. 2003). The *FCA* gene includes 20 introns, which are spliced out during mRNA assembly; however, introns 3 and 13 are spliced alternatively, leading to four different transcripts, designated as a, b, g, and d. Only the g transcript functions in the control of flowering time; it encodes a protein that, together

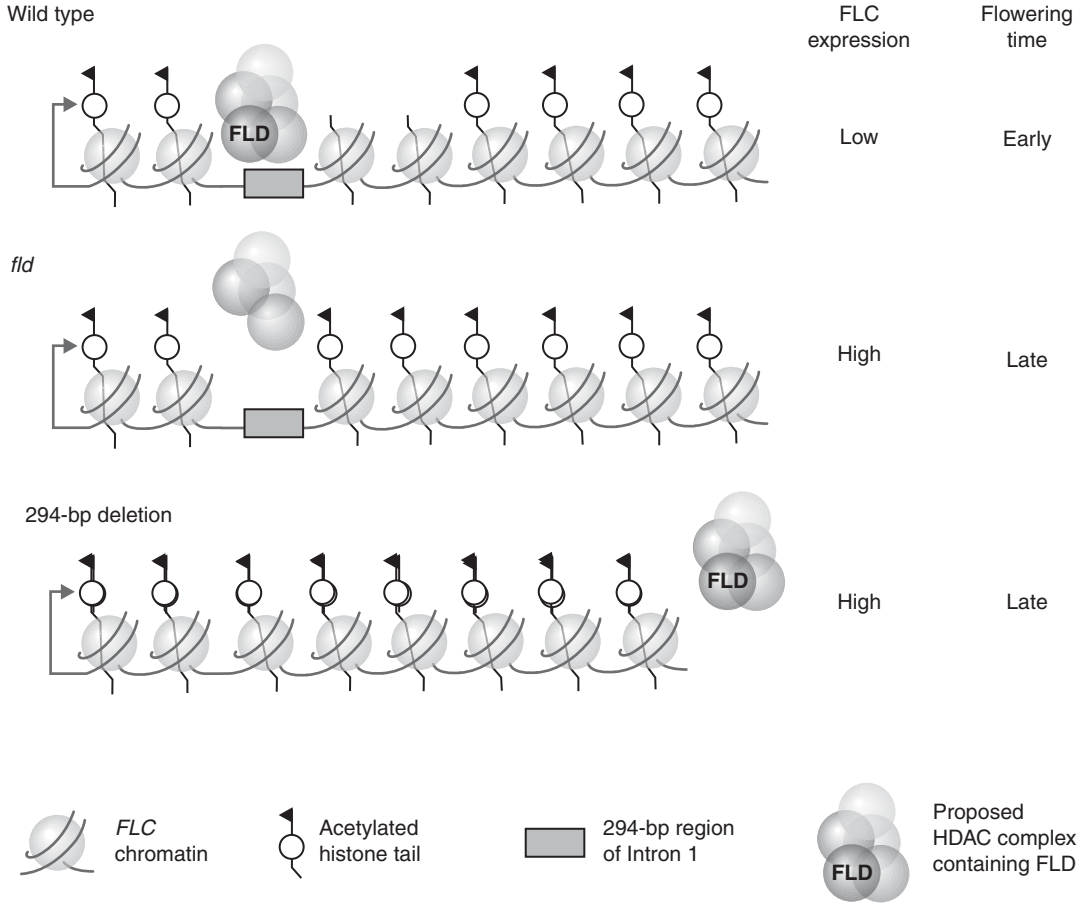


Figure 4.16 Regulation of *FLOWERING LOCUS C* (*FLC*), a floral repressor of *Arabidopsis*, by *FLD* (*FLOWERING LOCUS D*). In the wild type, *FLC* expression is suppressed by deacetylation of histones in the vicinity of FLC, conducted by an HDAC, of which FLD is a part. If *FLD* is mutated or if a 294 bp region in the first intron of *FLC* is deleted, HDAC is no longer able to deacetylate the *FLC* histones, alleviating *FLC* from transcription suppression and causing late flowering. Reprinted with permission from Bastow and Dean (2003). Copyright 2003 AAAS.

with the FY protein, suppresses *FLC* mRNA and thus stimulates flowering. In addition, FY and FCA proteins feed back upon the activity of the *FCA* gene because they promote the formation of the inactive b transcript, by polyadenylation in intron 3. This negative-feedback loop is essential to maintain a low concentration of active transcripts of *FCA*. When the autoregulation breaks down, active FCA is formed, causing inhibition of FLC mRNA and induction of flowering (Fig. 4.17). The *alternative splicing* of *FCA* transcripts provides a mechanism by which *FLC* expression can be fine-tuned during development and limited to certain tissues.

It is not known what endogenous signal is directing the autonomous pathway or triggers expression of *FLD*, *FVE*, *FCA*, and *FY*. Plants must pass through a juvenile phase before they can flower, but it is unclear how this is sensed internally. It might be that the pathway is not regulated dynamically, but functions constitutively to maintain high levels of *FLC* expression throughout early development (Simpson and Dean 2002). In addition, parts of the pathway may act to transfer information about environmental temperatures, as indicated above. The inhibitory effect of *FLC* on flowering is due its repressive effect on *FT* (*FLOWERING LOCUS T*; Fig. 4.15), one of the regulator genes in the network. Another gene suppressing *FT* has been identified, designated *EBS* (*EARLY BOLTING IN SHORT DAYS*; Gómez-Mena *et al*. 2001; Piñeiro *et al*. 2003). The sequence of this gene suggests that it is part of a chromatin-remodelling complex, and this suggests another case of epigenesis in flowering-time regulation.

The fourth pathway acting upon flowering time is the *gibberellin signalling pathway*. Gibberellic acid (GA) is a plant hormone with important effects on growth regulation, promoting germination, stem elongation, and fruit growth. Mutations in GA biosynthesis and GA receptors disrupt the floral transition, and have many other effects on plant development. One way in which gibberellins promote flowering is by increasing transcriptional activity of the floral meristem-identity gene *LFY* (*LEAFY*), through upregulation of a MADS transcription factor gene, *SOC1* (*SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1*). It is not known which endogeneous or exogeneous signals are integrated by the GA input in the flowering-time network.

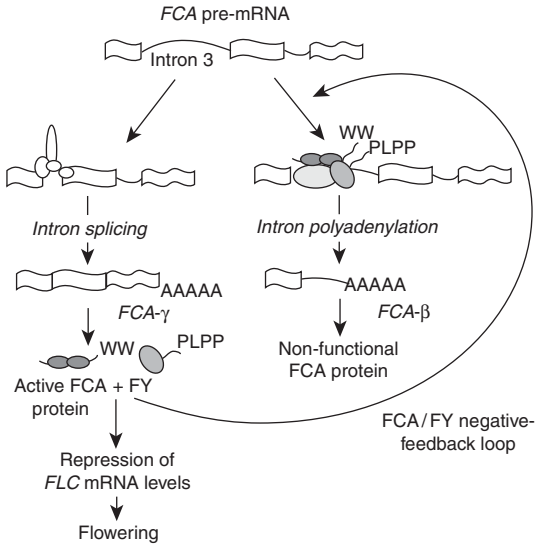


Figure 4.17 Scheme of the input from *FCA* and *FY* in the regulation of flowering time in *Arabidopsis*. Intron 3 of the *FCA* pre-mRNA transcript can be spliced out in two different ways, one of which is promoted by FY and FCA proteins, leading to intron polyadenylation and a non-functional b transcript. The other pathway leads to a fully functional g transcript, which in conjunction with FY represses *FLC* mRNA. Since FLC is a flowering repressor, this stimulates flowering. WW, a 35–40-amino acid protein domain engaging in protein–protein interactions; PLPP, amino acid motif of the FY protein, interacting with the WW domain of FCA. After Putterill *et al*. (2004), by permission of John Wiley & Sons.

Most of the insights discussed above have been obtained through analysis of mutants and naturally occurring varieties. Genome-wide expression-profiling studies have confirmed the regulatory networks derived from single-gene mutants, but at the same time have suggested the involvement of even more genes. Schmid *et al*. (2003) identified two previously unknown genes repressed by transition to long days, named *SCHLAFMÜTZE* (*SMZ*; meaning night cap) and *SCHNARCHZAPFEN* (*SNZ*; meaning snoring uvula). Knockout mutants of *SNZ* and *SMZ* flowered normally, showing that the two genes have a redundant function, and could not have been detected by genetic screens using mutants. They both encode a so-called AP2 domain, which makes them potential binding proteins for

micro-RNAs, which are short, single-stranded, RNA molecules that can base-pair with complementary sequences in mRNA, blocking their translation. Regulation of mRNA by microRNA has been implicated as an important mechanism in plant morphogenesis. The study of Schmid *et al*. (2003) suggests that microRNAs can also play a role in the regulation of flowering responses to daylength.

4.3.5 Regulation of flowering time in other plants

Is the genetic network regulating flowering time in *Arabidopsis* applicable to other plant species? Similarities are found in the Brassicaceae, the plant family to which *Arabidopsis* belongs (Kole *et al*. 2001). In *Brassica rapa* several flowering-time QTLs were identified and high-resolution mapping showed that one of them, *VFR2*, was homologous to *Arabidopsis FLC*. It seemed to have the same phenotypic effect: late flowering, downregulated by vernalization in the biennual genotype of *B. rapa*. Kim *et al*. (2003) identified genes sharing strong homology with *AGAMOUS-LIKE 20*, a MADS-box transcription factor downstream of *FLC* in *B. rapa* as well as in two other Brassicaceae, *Cardamine flexuosa* and *Draba nemorosa*. The expression pattern of *AGL20* in *C. flexuosa* during the floral transition was similar to *Arabidopsis*, which together with the data from Kole *et al*. (2001) suggested that at least some aspects of the regulation of flowering time are conserved between *Arabidopsis* and other Brassicaeae.

It is likely that most of the flowering-time genes identified in *Arabidopsis* are also present in plants outside the Brassicaceae, but they do not necessarily have the same function. For example, Petersen *et al*. (2004) recovered MADS-box genes of perennial ryegrass, *Lolium perenne*, which were differentially expressed during transition from vegetative to reproductive growth induced by vernalization. They used a differential-display technique with one PCR primer targeting a monocotyledon-specific conservative region in the MADS-box gene. After sequencing nine ryegrass MADS-box genes and studying their expression with quantitative real-time PCR the authors noted both similarities and differences with the *Arabidopsis* system. Genes with sequence homology to the *Arabidopsis AP1* subfamily appeared to have a different expression pattern and possibly a different function in vernalization, compared to *Arabidopsis*.

A more detailed comparison of flowering-time regulation is possible between *Arabidopsis* and rice (Hayama *et al*. 2003; Griffiths *et al*. 2003; Hayama and Coupland 2004; Putterill *et al*. 2004). Like many economically important crops *Oryza sativa* is a short-day plant; it promotes flowering when daylength falls below a critical threshold. Would this suggest that the mechanisms for regulation of flowering are entirely different from the long-day plant *Arabidopsis*? Recent genetic studies have elucidated part of the photoperiodic control of flowering time in rice. Several rice genes have been shown to be orthologues of *Arabidopsis* genes: a gene named *Hd1* is homologous to *CO*, and *Hd3a* is an orthologue of *FT*. However, *CO* promotes expression of *FT* in *Arabidopsis* whereas *Hd1* suppresses expression of *Hd3a* in rice (Hayama *et al*. 2003). So the photoperiodic response is reversed by using the same set of regulatory genes, regulated differently. Figure 4.18 provides an overview of the similarities in the regulation of flowering time between the two plant species.

Obviously, the usefulness of *A. thaliana* as a model for life-history investigation is limited by the fact that it represents only one type—a winter annual—out of the large variety of plant life histories. Biennuals, perennials, shrubs, and trees may have completely different ways of dealing with the problem of optimal timing of reproduction. Examination of the gene content of the *Populus trichocarpa* genome shows that most of the *Arabidopsis* flowering-time genes have counterparts in the poplar genome; however, a significant exception is the central floral repressor *FLC* which seems to lack an orthologue in poplar (Brunner and Nilsson 2004). The *FLC* subgroup of MADS-box genes seems to be specific to the Brassicaceae lineage. This raises questions as to the function of the other flowering-time genes in poplar. Despite the fact that a large part of the flowering-time network may be conserved across species the functions of the genes are not necessarily the same in species with different life histories.

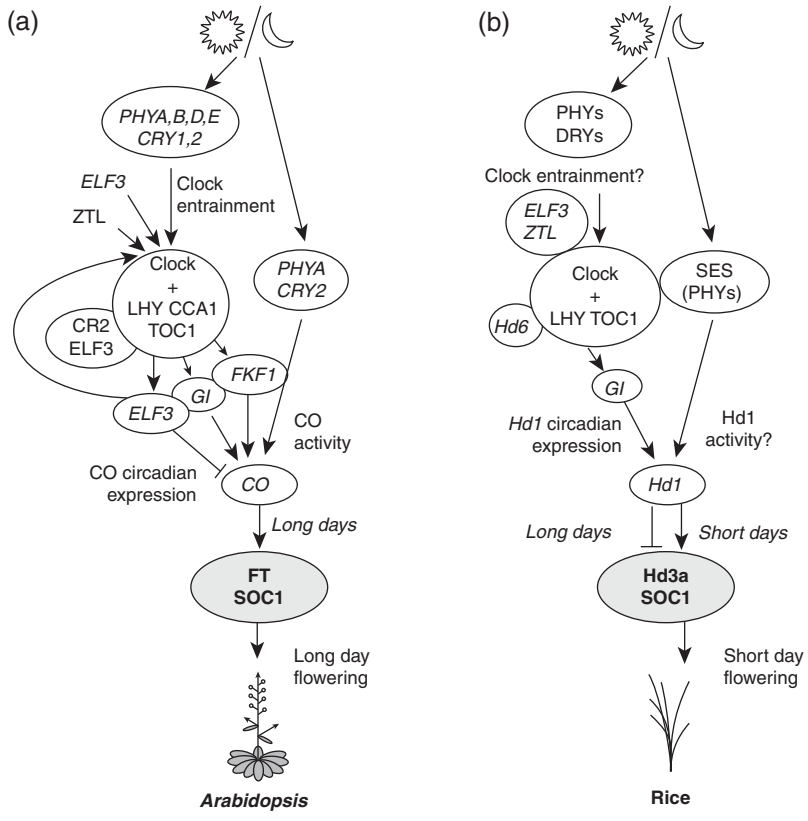


Figure 4.18 Similarities and differences between the photoperiodic control of flowering in (a) *Arabidopsis* and (b) rice. *Arabidopsis* flowering is promoted by long days while rice flowering is promoted by short days but repressed by long days. Despite the different phenotypic responses many elements of the regulatory network are the same. After Putterill *et al*. (2004) by permission of John Wiley & Sons.

4.4 Phenotypic plasticity of life-history traits

Many life-history attributes do not attain fixed values but respond to environmental conditions in an adaptive fashion. Part of the variation in life histories over species stems from the fact that different species have adopted different ways of responding to environmental factors. The functional relationship between the phenotype and the environment is called a *norm of reaction* (see Section 4.1). Life-history theory assumes that such a norm itself can be subject to natural selection and is optimized with respect to certain environmental conditions. We have discussed plant flowering time in the previous section as a life-history trait; however, its response to photoperiod and temperature can be seen as a reaction norm. In this section we will discuss more examples of phenotypic plasticity and its genomic underpinning.

There is considerable confusion as to how the primarily ecological concept of a reaction norm should be interpreted on the genetic level. One view holds that reaction norms are no more than a set of life-history traits measured in different environments, and that one has to analyse each trait in each environment as a separate variable, taking into account cross-environment correlations; this approach is also known as the *character-state view of phenotypic plasticity* (Roff 2002). Another approach emphasizes the continuity of the reaction norm across environments. This is especially appropriate if

environments form a natural gradient, for example in the case of temperature. Plasticity is then defined by the response in the average environment, plus the slope of the reaction norm if it is linear. For non-linear reaction norms there is no obvious indicator of plasticity, although the first derivative of the reaction-norm value in the average environment may be used as a local measure of plasticity when that environment changes (De Jong 1995). This approach is known as the *reaction-norm view of phenotypic plasticity* (Roff 2002). It has been suggested that in addition to genes determining the average response, there should exist 'genes for plasticity'; that is, genes which determine the slope of the reaction norm. Natural selection emanating from variable environments could promote such plasticity genes. In particular, plasticity is favoured (Schlichting and Smith 2002):

- if environmental change is frequent,
- if environmental cues are reliable,
- if environmental variation is fine-grained in space or time,
- in the case of coarse-grained temporal variation, if change is predictable, or
- in the case of fine-grained temporal variation, if there is a predictable sequence.

The reaction-norm perspective is very much dominated by the terminology of quantitative genetics and by statistical analysis of phenotypic data across environments. Molecular data are changing this perspective (Pigliucci 1996; Schlichting and Smith 2002; Gibson 2008). The example of flowering time discussed above, as well as the cases discussed below, illustrate that it may be difficult to reconcile the quantitative-genetics views of plasticity with modern genomic insights. The concept of a plasticity gene does not have a clear interpretation on the molecular level (Gibson 2008). Rather, plastic life-history traits are determined by networks of gene expressions, which integrate multiple cues from the environment with signals from the internal metabolism in such a way that the difference between genes for plasticity and genes for an average response cannot be seen. In this section we discuss three cases—polyphenism and body size in insects, and shade avoidance in plants—to illustrate this argument. As we will see, the genomic underpinning of life-history plasticity is less well developed than some of the other issues discussed in this chapter, so the section is phrased in terms of examples rather than a general theory.

4.4.1 Polyphenetic development

Many insects will develop not only one phenotype but several at the same time or different phenotypes in a temporal sequence, phenomena known as *polyphenisms*. Obviously, when the same genotype develops different phenotypes, there must be developmental switches that are sensitive to environmental conditions. Therefore, polyphenetic development can be considered a prime case of phenotypic plasticity, also called *discrete phenotypic plasticity* to differentiate it from the *continuous phenotypic plasticity* described by reaction norms. Functional genomic analysis should be able to reveal the expression programmes underlying development of alternative phenotypes.

Spectacular examples of polyphenisms are found in insects that can develop different morphotypes in different seasons or in response to changes in the diet. Several examples are discussed by Nijhout (2003a), including spring and summer forms of nymphalid butterflies, gregarious and non-gregarious phases of locusts, soldier and worker castes in ants, and hornless and horned mating types in dung beetles. In all cases investigated, polyphenetic switching of the developmental pathway is triggered by a hormonal signal. Development is canalized into two or more alternative pathways, through an altered hormone titre, an altered threshold sensitivity to the hormone, an altered timing of hormone secretion, or an altered timing of the hormone-sensitive period. Often the environmental signal triggering an alternative condition falls considerably before the actual developmental switch itself; when a decision on a specific pathway has been made, development usually becomes irreversible after some time and the animal is destined to become a specific morph.

Seasonal polyphenisms in *Bicyclus* butterflies (family Satyridae) have been investigated in detail. There are about 70 species in the genus *Bicyclus*,

which inhabit a variety of habitats in Central Africa. Most species have a wet-season morph with brown-coloured wings, and conspicous eye-spots and bands, as well as a dry-season morph with dull colours and cryptic wing patterns. This seasonal diphenism can be understood as an adaptation to seasonal changes in the habitat, expressing camouflage against a background of dead vegetation in the dry season, when the butterflies are less active, and predator deflection using eyespots in the wet season, when butterflies are more active. The two morphs can be produced in laboratory cultures maintained at different temperatures (Fig. 4.19). However, it is not certain whether temperature alone is the most important environmental cue determining the morphs, since temperature can be negatively correlated with humidity in one place and positively in another (Roskam and Brakefield 1999).

As Fig. 4.19 illustrates, a major difference between seasonal morphs lies in the eyespots on the wings. Butterfly eyespots are pattern elements composed of concentric rings of coloured scales. These rings develop around an organizing centre, a so-called *developmental focus*, which during metamorphosis induces the surrounding scale-building cells to produce a designated pigment (Brunetti *et al*. 2001; Beldade and Brakefield 2002; Carroll *et al*. 2005). Four stages can be recognized in the development of eyespots on butterfly wings. The first step already takes place in the imaginal disc of the last larval instar, when subdivisions (fields) of the wing are defined. In the second stage, foci are established within specific fields. The establishment of foci takes place in each field separately, which explains why mutants of butterflies with multiple eyespots on the wing can have one of the spots missing without any consequence to other spots (Monteiro *et al*. 2003). Specification of the eyespot focus is indicated by expression of a homeobox transcription factor, *Distal-less* (*Dll*), a *Hox* gene that has been cloned in *Drosophila* and many other species. Due to conservation of *Hox* genes throughout the animal kingdom, antibodies against the *Drosophila* Hox proteins also provide reactivity to the orthologous Hox proteins of other insects, which allows the expression of these developmental proteins to be localized in developing butterfly wings (Brunetti *et al*. 2001; Beldade *et al*. 2002). In the third stage, which takes place in the early pupa, a signal from the focus induces the surrounding cells to adopt a certain colour. The type of pigment that these cells adopt seems to depend on the sensitivity thresholds of the responding cells, whereas the strength of the signal determines the size of the eyespot. The fate determination around the focus uses the *hedgehog* signalling pathway, one of the major so-called toolkit genes of developmental genetics (Carroll *et al*. 2005). This pathway, like the insulin signalling pathway discussed in Section 4.2, translates the binding of an extracellular ligand into regulation of a transcription factor, which in this case is *cubitus interruptus*. Finally, the fourth phase of eyespot formation is the pigmentation itself, which takes place in the late pupal stage.

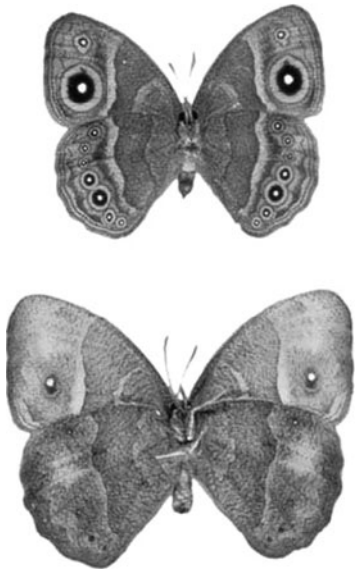


Figure 4.19 Seasonal polyphenism in *Bicyclus anynana* (Lepidoptera, Satyridae) includes a wet-season morph (top) and a dry-season morph (bottom). Courtesy of P.M. Brakefield, University of Leiden.

Which developmental switches are made to produce the alternative seasonal morphs of butterflies like *Bicyclus* is not known precisely. Given the knowledge about eyespot formation summarized above, it is likely that the switch will involve

regulation of developmental genes such as *Distal-less*, or one of the genes in the *hedgehog* signalling pathway. The involvement of ecdysteroid hormones in seasonal polyphenisms has been demonstrated for several butterfly species, so it seems reasonable to assume that the upstream part of the regulation consists of an endocrine signal, which in some way or another is sensitive to an environmental cue. Seasonal polyphenism represents a fascinating area of research where genomics, through developmental genetics, can meet ecology and evolution in a fruitful manner.

Another well-known case of polyphenism is represented by the castes of many social insects (termites, ants, wasps, and bees). The best-investigated caste system is that of the worker/queen polyphenism in the honey bee, *Apis mellifera*. Experiments have shown convincingly that the development of a larva into a worker or a queen is completely under environmental control, rather than reflecting a genetic predisposition of some larvae to follow a distinct developmental pathway. Larvae are induced to develop into queens when fed a rich mixture of food throughout development, including royal jelly, a secretion from the mandibular gland of nursing workers which contains a higher concentration of sugar than worker jelly. As a consequence, queen larvae develop faster, grow larger, and have larger corpora allata, the source of juvenile hormone, which controls the differentiation of oocyctes and the production of vitellogenic proteins by the fat body. Levels of juvenile hormone are considerably higher in queen larvae than in worker larvae, especially in the fifth instar, the stage just before pupation. Up to the fourth instar (2.5–3.5 days old) the queen developmental pathway is reversible, but after entry into the fifth stage a change of feeding regime has no further influence on the phenotype appearing after pupation. Obviously, the alternative developmental pathways are regulated by larval nutrition, mediated by endocrine signals, which in some way are translated into gene expressions.

Are the castes of honey bees characterized by specific expression signatures? In an early study, Evans and Wheeler (1999) looked at differential gene expression between queen and worker larvae of *Apis mellifera* using an SSH protocol; this was followed later by macroarray expression profiling using 144 cDNAs from enriched libraries (Evans and Wheeler 2000, 2001). Gene expressions were compared between early fourth instar (bipotential) larvae and fifth instar larvae destined to workers or queens. Several loci were confirmed to be differentially expressed and, interestingly, many of these were downregulated in queen larvae and the data showed that workers resemble the bipotential young larvae more than queens. So the suggestion is that the worker programme is the default pathway, which must be altered actively to produce a queen programme, and this latter programme includes switching off many genes and turning on a limited set of queen genes.

Figure 4.20 provides a summary of differential gene expression in honey bee castes. Young larvae (Y in Fig 4.20) overexpressed two heat-shock proteins and several proteins related to RNA processing. Among the genes specifically upregulated in worker larvae (W) is hexamerin 2, a member of a group of hexameric storage proteins, which serve as a source of energy during metamorphosis and adult life. Such hexameres also accumulate in the haemolymph of insects preparing for diapause (Denlinger 2002). A cytochrome P450 was also upregulated consistently in workers and this was also found in another bee species, *Melipona quadrifasciata* (Judice *et al*. 2004). The expression profile of queen larvae (Q) shows overexpression of ATP synthase and cytochrome oxidase I. This is consistent with earlier work by Corona *et al*. (1999) who, using differential display, identified three mitochondrial proteins, mitochondrial translation-initiation factor, cytochrome oxidase subunit 1, and cytochrome *c*, that were upregulated consistently in queen larvae compared with worker larvae. So, these limited data suggest that the worker larvae expression profile is characterized by allocation to storage and the queen larvae profile by genes reflecting higher mitochondrial respiration.

Developmental genomics of caste determination in honey bees is an active field of research. The early studies were limited by the relatively small number of genes analysed. The recently completed genome sequence of *A. mellifera* will allow a significant acceleration of expression profiling. Several EST

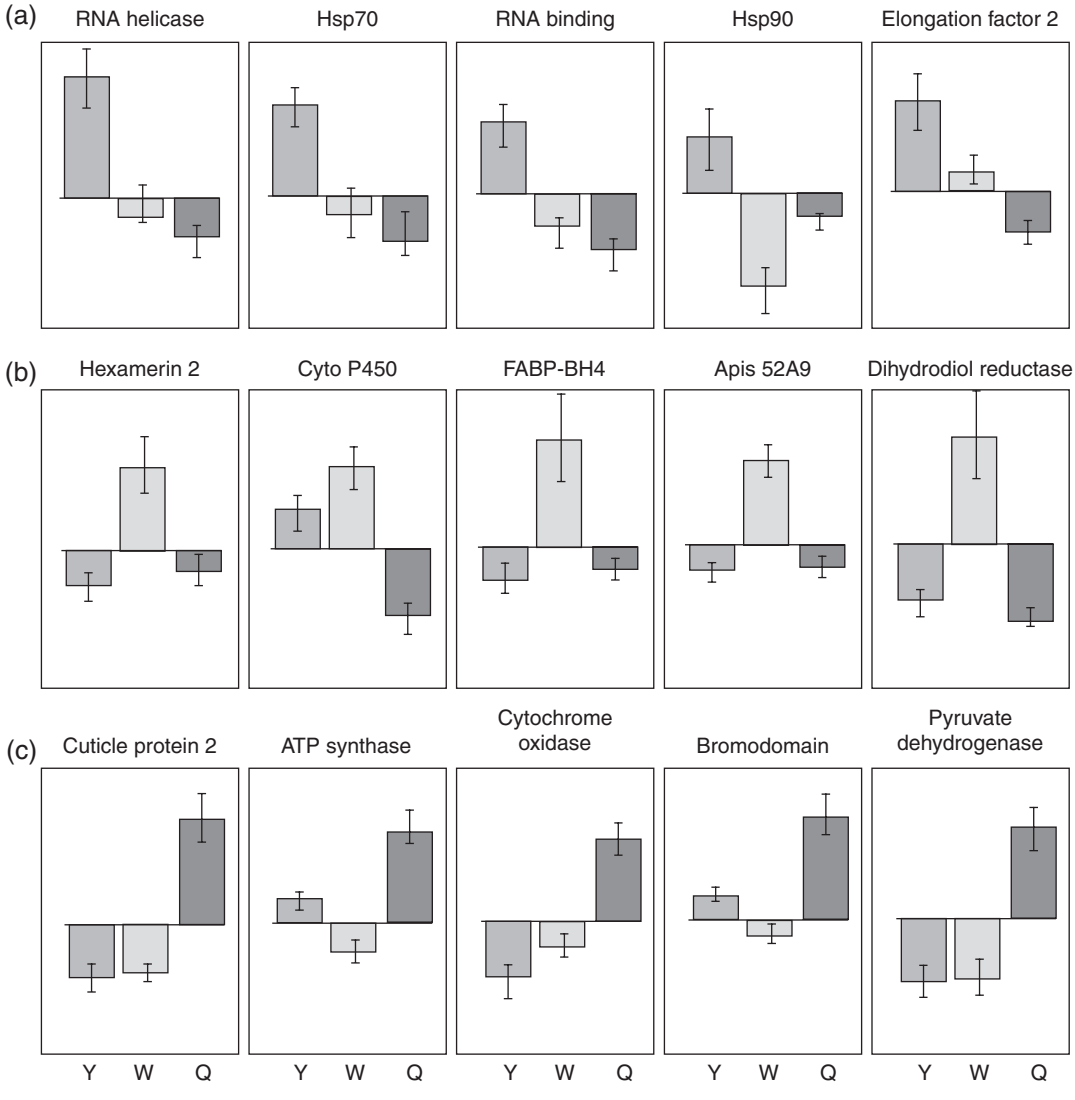


Figure 4.20 Overview of 15 genes of honey bee, *A. mellifera*, differentially expressed between fourth instar larvae (Y; young, bipotential), and fifth instar larvae destined to become workers (W) or queens (Q). Normalized expression levels are shown, with standard errors, estimated from non-competitive hybridization of cDNA samples with a macroarray of 144 probes, derived from SSH libraries. Shown are genes with high expression in (a) young larvae, (b) worker larvae, and (c) queen larvae. After Evans and Wheeler (2000), by permission of BioMed Central.

libraries have been developed allowing a more detailed analysis of the honey bee transcriptome (Whitfield *et al*. 2002; Nunes *et al*. 2004). Most likely the final answer is to be found in epigenetics. A recent study showed that DNA methylation by DNA methyltransferase 3 suppresses the developmental pathway towards a queen (Kocharski *et al*. 2008). The *A. mellifera* genome contains a lot of RNAi genes, some of them not known in other insects, and these genes are assumed to play a role in caste differentiation (Honey Bee Genome Sequencing Consortium 2006). Genome-wide analysis of honey bee castes will not only shed more light on the developmental and metabolic aspects

of caste differentiation, but also on the biological basis of differential behaviours in general.

The approaches taken in the honey bee would also be applicable to other eusocial insects, including ants, a favourite object of study in ecology. Many ant species have a more complicated system of caste differentiation and task division than honey bees; castes may include workers as well as soldiers, males, and alate (winged) females. The intriguing variety of ant social behaviours would be a very rich source of genomic discovery.

4.4.2 Body size

Body size is one of the most important life-history traits. Even though body size does not appear explicitly in demographic tables of fertility and mortality from which fitness is estimated (see Section 4.1), it is implicated in changes of these vital rates with age, for example because increasing size allows higher fertility and often entails a lower mortality (see Fig. 4.1). The adult body size of a species is also one of the few variables that predicts successfully a large number of ecological attributes, such as metabolic rate and energy intake and assimilation. We saw in Chapter 2 that body size was also associated with population size, which is one the main determinants of genome size according to the neutral theory of population genetics (see Fig. 2.4). Body size is also tightly linked to questions of shape and form, because physical forces have essentially different impacts on small objects compared to large objects, including cells and organisms. In his marvelous book, *On Growth and Form*, D'Arcy Thompson (1917) expressed it succinctly: 'Size of body is no mere accident'.

Despite the almost universal importance of body size, its determination by physiology and genes continues to be a formidably vexing problem (Nijhout 2003b). The mechanisms that determine body size are obviously contingent on the underlying framework for regulating cell size, cell number, and the size of organs; however, if an organism is to attain a characteristic body size (which many animals do) there must be additional mechanisms, responding to information generated all over the body. One way of looking at body size is to see it as a consequence of halting growth. The question then becomes not so much how to attain a fixed size but when to stop growing. This point of view is particularly applicable to holometabolous insects, which only grow in the larval stage and whose adult body size depends entirely on the mass reached just before pupation. Entomological research has shown that one of the major factors determining body size of insects is a threshold reached in the larval stage, the so-called *critical weight*. This is defined as the minimal weight above which further feeding and growth are not strictly required for successful pupation (Davidowitz *et al*. 2003). The strength of this concept is that it is causally linked to a series of endocrine events inducing the onset of pupation in the last larval instar. At the critical weight, the corpora allata stop producing juvenile hormone, which removes this hormone's suppression of prothoracicotropic hormone secretion, allowing the prothoracic glands to secrete ecdysone, which then sets into motion a cascade of events leading to metamorphosis. Larval growth stops when this sequence of endocrine events culminates in a peak of ecdysteroids, at which point the larva has grown to a size above the critical threshold, depending on the time delay for induction of ecdysteroid secretion and the photoperiod.

Detailed research on the tobacco hornworm, *Manduca sexta* (Lepidoptera, Sphingidae), has shown that additive genetic variation exists for critical weight, and this has allowed selection for adult body size in laboratory cultures. In addition, two other determinants contribute to genetic variation of body size in *M. sexta*, prothoracicotropic hormone delay time and growth rate (D'Amico *et al*. 2001). From research on *Drosophila* we know that growth itself is influenced by signals from the insulin/IGF-1 signalling pathway (Brogiolo *et al*. 2001; see Section 4.2). Inspired by mammalian research, another signalling pathway has shown to be involved in the regulation of growth rate of *Drosophila*, which centres around a protein called *target of rapamycin* (TOR). This protein was first discovered in yeast through mutants resistant to the cytotoxic effects of the fungicide, rapamycin. TOR is part of a signalling cascade that interacts partly with insulin/IGF-1 signalling, is sensitive to amino acids in the haemolymph, and influences

metabolism and growth (Britton *et al*. 2002; Oldham and Hafen 2003). Figure 4.21 provides a schematic overview of the various determinants of body size in insects.

The physiological evidence reviewed makes it understandable that body size is both genetically determined and plastic. Indeed, numerous experiments with insects have shown that adult body size is affected by environmental conditions, the best-studied response being the effect of temperature. Rearing temperature has a consistent effect on body size: insects, in fact almost all animals, grow larger at low temperatures and smaller at high temperatures. This trend is also present in the field: populations at higher latitude attain a larger body size than tropical populations of the same species. In *D. melanogaster*, the latitudinal cline of body size has a high heritable component and is accompanied by clines of allele frequencies, starvation resistance, and differential gene expressions (De Jong and Bochdanovits 2003; Bochdanovits *et al.* 2003; Bochdanovits and De Jong 2004).

One of the few genomic approaches to the problem of body-size determination is found in the work by Li *et al.* (2006), Gutteling *et al.* (2007), and Kammenga *et al.* (2007). These authors analysed the genetic mechanisms of the thermal responses of body size in *C. elegans*. In the reference strain N2, originating from Bristol, UK, body size decreases with environmental temperature in accordance with the temperature-body size rule discussed above. However, in CB, a strain from Hawaii, the effect of temperature on body size is much smaller (in fact not significant). In addition, within each strain there is substantial genetic variation for the slope of the body size-temperature response (Fig. 4.22a). A panel of recombinant inbred lines, obtained from a cross between N2 and CB was genotyped with SNPs to locate a quantitative trait locus (QTL) responsible for thermal plasticity of body size.

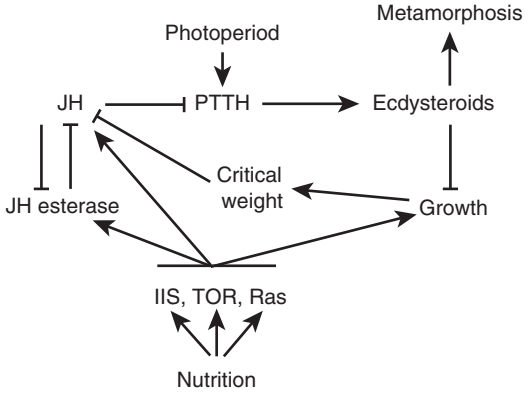


Figure 4.21 Tentative schematic overview of relationships involved in the determination of adult body size in insects, inspired by research on tobacco hornworm, *M. sexta*. Adult size is determined by cessation of larval growth before pupation, which is triggered by ecdysteroids that appear when the titre of juvenile hormone falls upon reaching the critical weight. JH, juvenile hormone; PTTH, prothoracicotropic hormone; IIS, insulin/IGF signalling pathway; TOR, target of rapamycin signalling pathway; Ras, Ras signalling pathway. Based on Nijhout (2003b) and other sources.

The QTL analysis identified genomic regions associated with body size at 12 ºC, with body size at 24 ºC and loci associated with the slope of the thermal reaction norm. Interestingly, more loci were found for body size at 12 ºC than for body size at 24 ºC, while the 24 ºC locus on chromosome IV was distinct from the 12 ºC loci, which were all located on chromosome III (Fig. 4.22b). Several QTLs were found for the slope of the thermal reaction norm, most of them in the same regions as the QTLs for either temperature. Apparently, the genes influencing body size at a specific temperature also influence the responsiveness of body size to temperature. However, a region designated as TRB (thermal reaction norm for body size) was found to be distinct from the body size loci (Fig. 4.22b). This TRB locus affected only the response of body size to temperature, not body size at any temperature as such. It is therefore a true candidate 'locus for plasticity'.

Capitalizing on the complete genome sequence of *C. elegans*, Kammenga *et al.* (2007) located a gene inside the TRB region, *tra-3*, as a candidate controlling the thermal slope of body size. *Tra-3* encodes a protein with high homology to mammalian calpain regulatory proteases, which are involved in calcium signalling and the regulation of cell size. *Tra-3* of *C. elegans* was first reported to be associated with sex determination, transforming XX females into males if mutated (hence its name, *transformer 3*). Its role in temperature-dependent body-size regulation was only discovered by Kammenga *et al.* (2007).

Cold environments are known to cause an increase of intracellular calcium and so variation in *tra-3* could cause variation in the way in which cell size, and ultimately body-size, responds to temperature. The CB strain from Hawaii carries a single nucleotide mutation in *tra-3*, which was shown to decrease the calcium-binding activity of the TRA-3 protein. This work is a unique example illustrating how ecological genomics, coupled to classical QTL analysis, can lead to a better molecular understanding of the polymorphisms in temperature plasticity of body size.

The analyses of Bochdanovits and De Jong (2004) in *Drosophila* and Kammenga *et al*. (2007) in

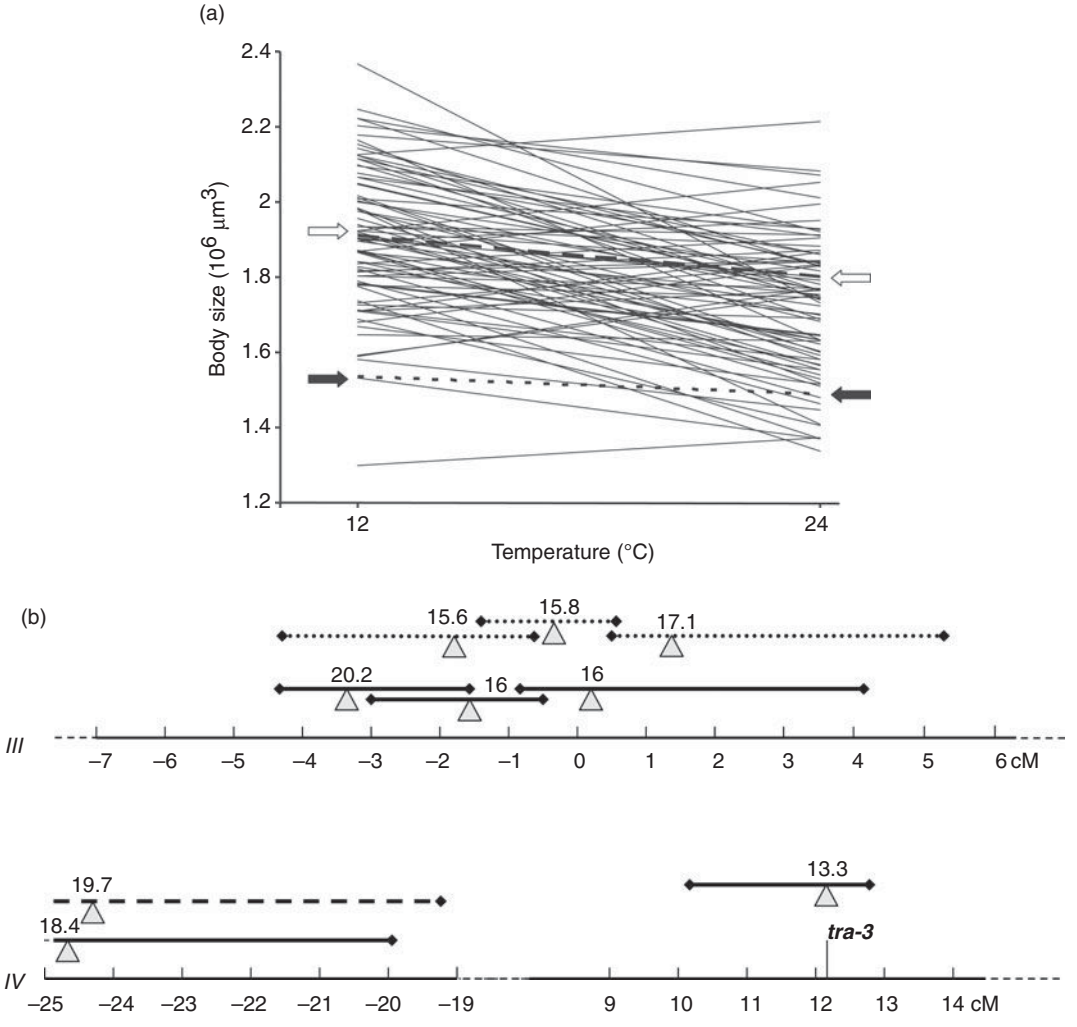


Figure 4.22 (a) Thermal reaction norms for body size in 80 recombinant inbred lines of *C. elegans* and their parental strains N2 and CB. The response of the parent strain N2 (from the UK), indicated by open arrows on the sides, is much stronger than the response of the CB strain from Hawaii, indicated by filled arrows. (b) Position of body-size QTLs on chromosomes III and IV of *C. elegans*. Only the chromosomal portions with body-size QTLs are shown. Triangles designate the peaks of the QTLs and horizontal lines their confidence intervals. Numbers above the confidence intervals indicate the likelihood ratio values. Dotted lines: QTLs for body size at 12 °C, dashed line: QTL for body size at 24 °C, solid lines: QTLs for thermal reaction norms for body-size (TRB). *tra-3* (*transformer 3*): candidate gene for body-size plasticity. From Kammenga *et al.* (2007).

C. elegans illustrate that trade-offs between life-history traits may indeed be due to genes with pleiotropic effects, but how these genes may exert such effects on the phenotype is far from clear. From a biochemical point of view, genes causing trade-offs in life histories are expected to be part of the pathways involved in nutrient allocation, body growth, and energy metabolism. These pathways are known as IIS, TOR, and Ras (cf. Fig. 4.21). In accordance with this expectation, Bochdanovits and De Jong (2004) found a Ras protein (CG9611) to be among the candidate genes associated with high body weight but low survival of *D. melanogaster*. Ras proteins are part of the Ras/mitogen-activated protein kinase (MAPK) signalling pathway, which like the insulin/IGF-1 signalling and TOR pathways transduces signals from extracellular growth factors (in this case the so-called epidermal growth factor, EGF) into metabolic action in the cell (see Fig. 5.3). The Ras/MAPK pathway is known for its role in regulating cell growth, cell proliferation, and survival. The repeated occurrence of genes from signalling pathways involved in transducing nutritional and growth factor information into cellular metabolism is probably indicative of a crucial role of these pathways in the regulation of life-history traits and possible trade-offs among them.

4.4.3 Shade avoidance

The last plastic life-history trait to be discussed concerns the remarkable flexibility seen in plant growth when exposed to different light intensities or light spectra. Part of this flexibility can be understood as a mechanism to avoid shade and therefore this type of phenotypic plasticity is known as the *shade-avoidance syndrome*. Shade avoidance is observed when plants are exposed to low ratios of red to far-red light, such as would prevail during growth under a closed canopy. The response involves suppression of branching, emphasis on vertical elongation, and early flowering. This is assumed to contribute to plant fitness, since it allows completion of the life cycle and seed set before being overgrown by competitors.

Plants have three groups of light-sensitive molecules that translate aspects of the light regime into metabolic action: phototropins, cryptochromes, and phytochromes (Chen *et al*. 2004). *Phototropins* are sensitive to UV-A and blue light; they are involved in the well-known phototropic responses of plants (roots grow away from light, stems bend towards light), and also in regulating chloroplast movements and stomatal opening. *Cryptochromes* are also sensitive to UV-A and blue light and are involved in *de-etiolation* (the transition of dark-grown seedlings to phototrophically competent plants), photoperiod-dependent induction of flowering, and entrainment of the circadian oscillator. *Phytochromes* are sensitive to red and far-red light and are involved in seed germination and shade avoidance. Many of the light-dependent responses of plants are modulated by networks involving both cryptochromes and phytochromes (see the discussion of flowering time in Section 4.3), but the shade-avoidance response is regulated only by phytochromes and this is the reason for discussing them in a little more detail here.

There are five different phytochrome genes in *Arabidopsis*, designated *PHYA* to *PHYE*. Similar families of phytochromes are known in species throughout the plant kingdom, from algae to angiosperms. The proteins encoded by these genes differ in their spectral properties and the rates of conversion between active and inactive states; consequently they have different physiological roles. Mutants of *Arabidopsis* defective in one of the phytochrome genes have confirmed this differentiation of function. PHYB plays a predominant role in the shade-avoidance response, with redundant action of PHYD and PHYE (Schlichting and Smith 2002).

How can phytochromes translate light signals into gene expression? A large amount of detailed knowledge is available on the first steps in this process (Chen *et al*. 2004; Schepens *et al*. 2004). The light-interception component of the system is not the phytochrome itself, but a *chromophore* called phytochromobilin. The role of this chromophore is similar to the role of retinal bound to proteorhodopsin, as discussed in Section 3.4: it triggers a state transition of the main molecule. The phytochrome protein may undergo a conversion between two relatively stable states: a red-light-absorbing Pr form and a far-red-absorbing Pfr form. The Pfr form is assumed to be the active configuration.

Phytochromes are present as dimeric molecules in the cytoplasm. When at least one of the units is activated to the Pfr form, the molecule as a whole can be imported into the nucleus (Fig. 4.23). In the nucleus, phytochromes are localized in so-called *nuclear bodies*. In general, nuclear bodies, also called speckles, are considered subcompartments of the nucleus that carry out specific functions, such as mRNA splicing. The nature of the nuclear bodies recruiting activated phytochromes is unknown. They seem to represent structures by which phytochromes can transfer a signal to transcription activators. Both subunits of the phytochrome must be in their active form before localization in these bodies can occur.

Expression of a large number of genes is triggered by phytochromes. Transcription profiling using microarrays applied to *Arabidopsis* mutants defective in one of the phytochrome genes have revealed the genome-wide nature of the phytochrome signalling pathway (Ma *et al*. 2001; Quail 2002; Wang *et al*. 2002; Devlin *et al*. 2003). The first targets seem to be transcription factors, since these genes respond within 1 h of the light signal. Presumably, many of the other expression changes are downstream targets of transcription factors. No less than 26 different cellular pathways were found to be regulated by phytochrome signalling—some suppressed, others promoted—in a coordinated fashion (Ma *et al*. 2001).

Devlin *et al*. (2003) proposed seven functional categories for the genes differentially regulated by shade in *Arabidopsis*. Genes related to photosynthesis, fatty acid metabolism, and redox metabolism were mostly downregulated. Genes acting upon the cell wall, including pectinesterases and pectate lyases, were mostly upregulated; these genes support loosening of the cell wall. Genes regulated by the plant growth hormone auxin made up a large proportion of the upregulated genes. These switches

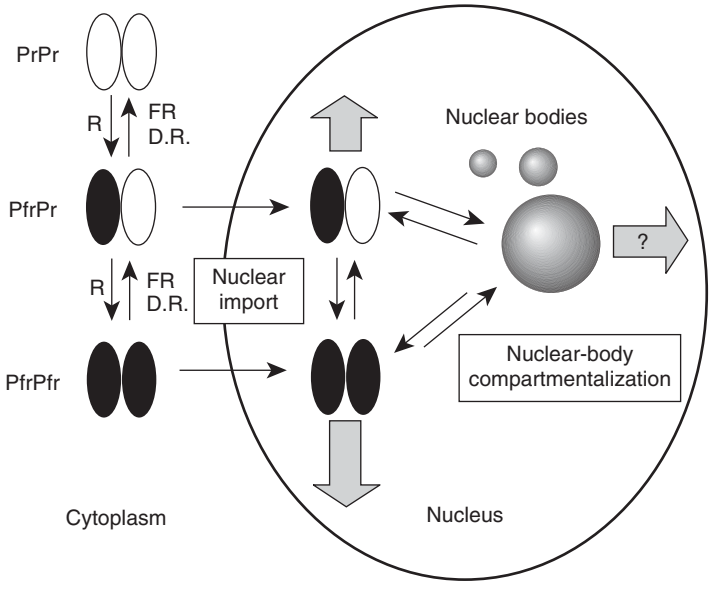


Figure 4.23 Relocalization of phytochromes in the cell, using PHYB as a model. Interception of red light by a chromophore attached to PHYB leads to activation of one or two of the subunits of the dimeric molecule (transition from Pr to Pfr). The activated molecule may be imported into the nucleus where it is compartmentalized into nuclear bodies. Localization into nuclear bodies is necessary for triggering gene expression. The PfrPfr form is incorporated preferentially into nuclear bodies over the PfrPr form. DR, dark reversion (slow spontaneous decay of Pfr to Pr in the dark); R, red light; FR, far-red light. After Chen *et al*. (2004), reproduced with permission from Annual Reviews.

in the transcriptome are all understandable in light of the pronounced elongated growth and accelerated flowering shown by the shade-avoidance phenotype.

Several studies have demonstrated the importance of plant hormones (auxins and gibberellins) in regulating the shade-avoidance growth response. A role of the volatile plant hormone ethylene is also suggested. In a study on tobacco, Pierik *et al*. (2004) showed that ethylene-insensitive plants had a reduced response to shade. The effect of ethylene requires GA because plants with inhibited GA production showed hardly any ethylene-dependent response to shading.

The study of phytochrome-mediated shade avoidance, although far from complete, demonstrates that it is possible, in principle, to explain the process of phenotypic plasticity in terms of gene expression regulated by environmental signals (Schlichting and Smith 2002). With sufficient sophisticated knowledge of the signalling pathways and the expression programmes they trigger, it would be possible to account for the widely different plant phenotypes that may develop from the same genotype. On the other hand, in considering the large number of genes involved and the way their expression is organized into networks, rather than in linear causal chains, this could turn out to be an extremely difficult task. With this open-ended challenge we close our discussion of phenotypic plasticity.

4.5 Genomic approaches to life-history patterns: an appraisal

Genomic analyses of life histories, as reviewed in this chapter, have revealed a number of molecular principles underlying regulation and determination of key life-history events. These can be summarized as follows.

Environmental information (food, crowding, light, daylength, temperature) often modulates life-history traits and this is the basis of phenotypic plasticity. In animals, such information is typically processed by the nervous system, then translated into a hormonal signal, which acts upon signalling pathways to steer gene expressions in target cells. Signalling pathways consist of a membrane-bound extracellular receptor, connected to an intracellular system of kinases, able to trigger a cascade of biochemical events, finally leading to activation of a transcription factor, which then triggers gene expression. The involvement of the insulin signalling pathway in regulating growth, reproduction, and longevity is a prime example of this principle, but also the induction of insect diapause and the developmental switches in polyphenisms act similarly. In plants, phytochromes play a major role in the translation of light signals.

Second, we may note that if an organism has the capacity to follow more than one developmental option, for example a dormancy stage in addition to an active stage, or flowering in addition to vegetative growth, the alternative pathway is often constitutively present throughout life but repressed until an environmental cue triggers its appearance. The execution of the alternative programme then becomes a question of simply removing a repression, rather than staging the whole programme anew. So-called resting stages (e.g. diapause) are not quiescent at all from a molecular point of view. Despite the fact that many genes are downregulated during diapause, the entry, maintenance, and exit of diapause involves active upregulation of several other genes. We have seen indications of such 'hidden' developmental programmes in nematodes (the dauer larva programme) and in *Arabidopsis* (the floral transition). Interestingly, mutations in *C. elegans* show that it is possible to uncouple part of this programme (life extension) from the main pathway leading to dauer formation, and thus to confer increased longevity to the non-dauer adult.

Third, we have seen that expression of life-history traits often comes with intricate mechanisms of regulation that go further than simple transcriptional regulation. We have seen four examples of this: alternative splicing, gene silencing, RNAi, and post-translational regulation of protein stability. Alternative splicing was seen in a *Drosophila* study showing that different splice variants of the same gene are expressed in different life-history stages, and in *Arabidopsis*, where one of the loci for flowering induction is suppressed by a feedback loop promoting an inactive splice variant of the

gene. Alternative splicing seems to be a mechanism by which gene expression can be fine-tuned in time (developmental stage) or space (specific tissues). The second type of non-transcriptional regulation, gene silencing, was observed in the vernalization response of *Arabidopsis*, where histone deacetylation is employed to suppress floral suppression and allow bolting. Gene silencing was also involved in mutations conferring lifespan extension in yeast, nematodes, and fruit flies. Such epigenetic regulation seems to act as a memory of environmental events, which can trigger an adequate response in cells that do not witness the event themselves but are imprinted by their mother cell. The third type of non-transcriptional regulation is RNAi. Both microRNAs (implicated in the photoperiodic pathway of the *Arabidopsis* floral transition) and RNA-binding proteins (in the autonomous pathway) have been discussed above as means to selectively inhibit mRNAs. Finally, post-translational regulation was noted in the light-dependent stabilization of CONSTANS protein by phytochromes and crytochromes in the photoperiodic pathway regulating flowering time in *Arabidopsis*.

The fourth molecular principle illustrated in this chapter is that life-history traits are underpinned by a complex network of gene expression and feedback loops. Networks are characterized by upstream and downstream functional units. In the upstream part, a simple trigger may act upon a signalling pathway. If such upstream genes are mutated, this often has very large effects on the phenotype and may introduce a syndrome of correlated altered life-history traits (see the *daf-2* mutants of *C. elegans*). Downstream of the network we often see a host of gene expressions, which are triggered by a limited number of transcription factors, acting upon all genes sharing a certain motif in their promoter. An obvious example of a downstream gene-expression cascade was the microarray analysis of longevity in *C. elegans*, in which a large number of genes upregulated as well as downregulated by the transcription factor DAF-16 were identified. Another characteristic of a network is the principle of *convergence*, which holds that different pathways are integrated by acting (some positively, some negatively) on a single integrator gene. The floral repressor *FLC*, which integrates the autonomous and the vernalization pathway in *A. thaliana*, is a clear example of this. A third property of networks is *redundancy* or parallelism. This is especially obvious in organisms with duplicated genomes, where more than one copy of the same gene exists and mutants do not have recognizable phenotypes due to another copy of the gene performing a similar function. The discovery of *SCHLAFMÜTZE* and *SCHNARCHZAPFEN* in the photoperiodic pathway of *Arabidopsis* floral transition is a clear example of this.

A final molecular lesson from the examples given in this chapter concerns the extrapolation of genomic information across species. As discussed in Chapter 2, comparative genomics has developed a whole gamut of instruments by which fine- and coarse-scaled comparisons of genomic sequences are made. In this chapter we have seen one example, the photoperiodic response of *Arabidopsis* and rice, where the genes themselves were conserved between species, but regulated in a different way, with the consequence that *Arabidopsis* flowering is stimulated by long days and rice flowering is stimulated by short days. If this principle of similar structure but different function in related species is common, homology of coding sequences is insufficient for extrapolating across species; rather, the basis for extrapolation should lie in functional comparative genomics. A similar disparity between the transcriptome and the genome was presented as an example in Chapter 1, when gene expression in different organs of *H. sapiens* and *Pan troglodytes* was compared (Fig. 1.6). In the present chapter we have seen that the first methodologies to analyse transcriptional profiles shared across species are now being developed.

Is ecological genomics about to make a major contribution to life-history theory? Obviously the genomics revolution has already intruded into the analysis of life histories, but we see four major limitations. In the first place, the outcomes of microarray studies published so far sometimes differ widely between similar studies. This is especially obvious from the appallingly small overlap of genes reported in different studies to be regulated by age in *Drosophila*. Maybe what is lacking here is a good definition of the conditions under which test

animals are cultured or exposed. Animal physiologists know that physiological responses of animals are influenced by many environmental factors, some of them difficult to standardize, such as food quality, air humidity, microbial infection, pheromones from conspecifics, and so on. Consequently, a genome-wide gene-expression profile may partly reflect such non-standardized aspects of the test organism. Only replication across studies and robust statistical analysis can reveal the universal responses among those that are just contingent on specific experimental conditions.

A second limitation is that most of the studies conducted to date have focused on a few genetic model species. Although some generalities have been indicated in the sections above, the regulation of aging though the insulin/IGF-1 signalling pathway being a case in point, most of the work is concentrated on *C. elegans*, *Drosophila*, and *Arabidopsis*, species that represent only a tiny fraction of the array of life histories in nature. Broadening the comparative basis will surely benefit a further integration of life-history theory and ecological genomics.

A third limitation is the focus on laboratory observation. Because any life history, especially mortality, has an important component external to the organism (predation, microbial infection, competition, spatial heterogeneity), which is difficult to mimic in the laboratory, the relevance of gene-expression profiles and mutants only observed in the laboratory can be doubted. An example illustrating this argument is a study by Weinig *et al*. (2002), who demonstrated several QTLs for flowering time in *Arabidopsis* that are only expressed under certain field conditions and thus not found in laboratory mutants. A good strategy for ecological genomics may be to exploit the natural variation of genomic programmes in wild organisms and then to work out mutations and expression profiles relating to these 'eco-variable' loci (Kammenga *et al.* 2008).

Finally, we note that many gene-expression studies have not yet left the descriptive stage. It is one thing to draw up an inventory of transcriptional profiles associated with some life-history event, but quite another to explain the causal relationship between these profiles and the life history. Only in the cases of longevity in *C. elegans* and flowering time in *Arabidopsis* is such an understanding coming within reach.

Despite these limitations we see a glorious future for a further merger between ecological genomics and life-history theory. This will involve establishing a link between crucial phenotypic phenomena, such as trade-off and plasticity on the one hand and gene expression, pleiotropy, and molecular signalling on the other (Roff 2007). We expect that such bifaceted discussion about fundamental concepts of life-history theory will contribute to a smooth transition between evolutionary explanation, the underlying physiology, and molecular genetics.

CHAPTER 5

Stress responses

Stress is a fundamental aspect of life and a major aspect of natural selection in the wild. Ecologists have studied the responses of plants and animals to environmental stress factors since the 1960s. Previously known as physiological ecology or ecophysiology, these studies are now often called stress ecology. The study of stress responses on the genomic level has produced new insights into the mechanisms that enable plants and animals to survive in harsh environments and that limit the distribution of species. Biochemical studies have shown that, on the cellular level, there is surprising degree of uniformity in the stress responses of different species, even to widely different environmental stress factors. Genomic studies have reinforced this idea while at the same time providing new insights into the coherence of the cellular stress response. Stress is evoked in an organism at the edges of its ecological niche. The extent to which the organism is able to deal with such stresses determines the limits of its ecological amplitude. There is therefore a logical link between genomic analysis of the stress response and the ecology of the species. In this chapter we aim to introduce the reader to the ecological genomics of stress analysis.

5.1 Stress and the ecological niche

Even the most casual observer of natural systems will note that many species tend to occupy a characteristic place in nature. The idea is demonstrated most vividly by gradient studies, for example the distribution of plants on a salt marsh, where each species tends to be limited to a certain zone by a combination of soil texture, redox potential, salt, and lime. Ecologists have invented the concept of an ecological niche to organize their thoughts about the ways in which organisms fit into their environment. The inception of this concept in the ecological literature is attributed to the American ornithologist Joseph Grinnell with his now classical paper on the California thrasher published in 1917, but the most widely used and influential elaboration of the niche concept is that of Hutchinson (1957). Hutchinson defined the niche in terms of any number of conditions and resources that limit the distribution of a species. The niche was pictured as an *n*-dimensional hypervolume that envelops those values of continuously varying environmental factors that allow long-term survival of the species. As an illustration of the Hutchinsonian niche concept we reproduce a two-dimensional picture of fitness in the collembolan *Folsomia candida* as a function of zinc exposure and food density (Fig. 5.1; Noël *et al*. 2006).

Hutchinson (1957) realized that a distinction should be made between the *fundamental niche*, which comprises all the conditions under which a species potentially may occur, and the *realized niche*, which is usually more narrow than the fundamental niche due to competition in the field. The fundamental niche is observed in laboratory experiments and in the field when competitors are absent. At the edges of its fundamental niche a species is less well equipped to face competition with others and so, when competition is important, it will give way at the edges and occupy a smaller section of the gradient, the realized niche. The formal definitions of the ecological niche by Hutchinson (1957) have spurred a great variety of studies aiming to explain community structure in the wild from the traits of individual species and their responses to environmental factors.

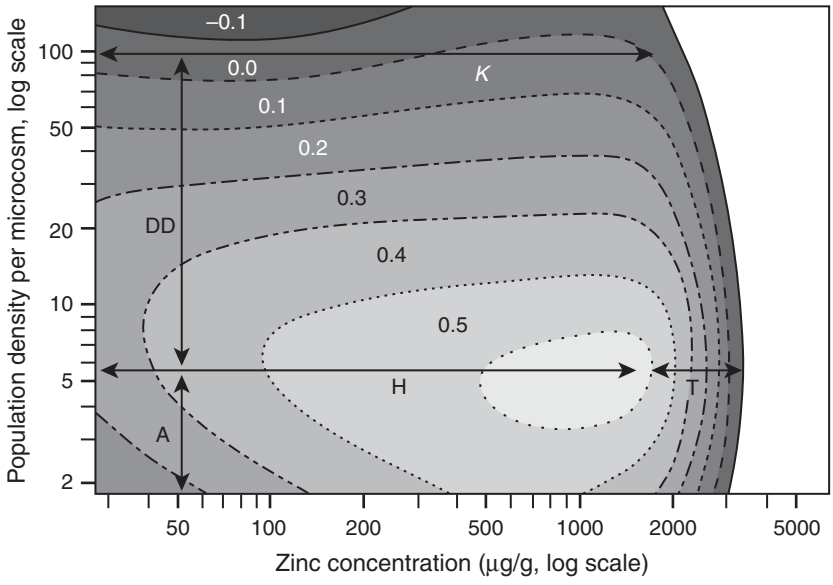


Figure 5.1 Graph illustrating the niche concept in two dimensions. Contours of population growth rate of the soil-dwelling collembolan *Folsomia candida* are plotted as a function of dietary zinc concentration and initial population density per microcosm (reflecting food density). Four different ranges of effect are indicated. At high exposure levels zinc becomes toxic and suppresses fitness (T); with increasing population size fitness decreases due to crowding (DD, inverse density dependence; *K*, carrying capacity); at low exposure levels zinc supports fitness by a stimulatory effect on growth, a phenomenon known as hormesis (H); at low food density fitness decreases due to absence of inter-individual communication stimulating consumption (Allee effect; A). Courtesy of H. Noël and R.M. Sibly, University of Reading.

The ecological niche concept touches the very heart of ecology—the relationship between species and their environment—but this has not prevented the proliferation of a great deal of confusion in the literature. Several reviews have pointed out the historical context of this confusion, which is partly due to independent elaboration by Joseph Grinnell and Charles Elton (Chase and Leibold 2003). On the one hand the concept emphasizes the requirements of species, but on the other hand it includes the species' role or impact on its environment. This Janus-faced property of the niche caused such confusion that some ecologists in the 1970s suggested avoiding the term niche altogether. However, Chase and Leibold (2003) revisited the concept, setting the stage for a new synthetic approach, framed by the recent developments of ecology. Their new, synthetic, definition of ecological niche runs as follows.

> The joint description of the environmental conditions that allow a species to satisfy its minimum requirements so that the birth rate of a local population is equal to or greater than its death rate, along with the set of per capita effects of that species on these environmental conditions.

This definition joins the two components of the niche mentioned above, and it recognizes that whenever resources and environmental conditions are altered by the organisms themselves, as in the case of predators depleting their prey or ecological engineers altering the physical structure of their environment, this aspect should be included in the niche concept. The definition by Chase and Leibold (2003) also recognizes that the difference between mortality and natality, in other words fitness, is the ultimate measure in which the requirements of the organism are to be expressed.

The niche concept has been highly instrumental in community ecology, as it allows explanations of community structure from the point of view of the placing of species along a gradient of resources or conditions according to their ecological niche.

Various competition models have been derived that predict how niche overlap between species is minimized by competition and the maximum overlap between the niches of two adjacent species that will allow coexistence of both. However, less attention has been paid to the question of how the niche itself is shaped by underlying determinants founded in the physiology of the species. The science of environmental physiology, also called ecophysiology or physiological ecology, addresses these aspects. Several physiology textbooks have been written that reach out to ecology, including that by Larcher (2003) on plants, and that by Schmidt-Nielsen (1997) on animals. In this chapter we likewise aim to explore the reductionist path, from niche to genomics.

Environmental physiology has a strong focus on studies of stress. The reason for this bias is that the regulatory mechanisms allowing homeostasis of the *milieu interne* are best seen when these mechanisms are put to the test by pushing the organism to the borders of its ecological niche. This is how we will approach the issue in this chapter; by studying responses at the edges of the ecological niche we aim to reveal the regulatory mechanisms that promote fitness both within and outside the niche.

Like the niche, the concept of stress has a long and confusing semantic history in ecology. Some authors have argued that ecological stress should be defined by analogy to the physical concept of stress, which would imply that it is an external constraining or impelling force applied to an ecological system. Most biologists, however, consider stress as an internal state, brought about by a hostile environment or negative social interaction. Nowadays, there seems to be agreement on the fact that a distinction should be made between a *stressor* (an external factor), the *stress* (an internal state brought about by a stressor), and the *stress response* (a cascade of internal changes triggered by stress). Although the concept of stress can be defined at various levels of ecological integration, stress is most commonly studied in the context of individual organisms, whereas stress responses are studied on the cellular, biochemical, and genomic levels.

We need to realize that the concept of stress is not absolute; it can only be defined with reference to the normal range of ecological function; that is, with reference to the ecological amplitude or ecological niche of the species. What is an extremely stressful condition for one organism (e.g. the absence of free air) is quite normal for another organism (a fish). A definition of ecological stress that incorporates this idea runs as follows (Van Straalen 2003).

> Ecological stress is a condition evoked in an organism by one or more environmental factors that bring the organism near to or over the edges of its fundamental ecological niche.

This definition complies with the common physiological usage of the term, which is that stress is an internal condition, not an external factor. In addition, stress has the following properties: (i) it is usually transient, (ii) it involves a syndrome of specific physiological responses, and (iii) it is accompanied by the induction of mechanisms that counteract its consequences. Our niche-based definition of stress is illustrated schematically in Fig. 5.2.

The stress response can take different forms, depending on the timescale. Calow (1989) distinguished two main types, *proximate* and *ultimate* responses. The proximate response implies induction of physiological, biochemical, and genomic mechanisms (*physiological adaptation*) that allow

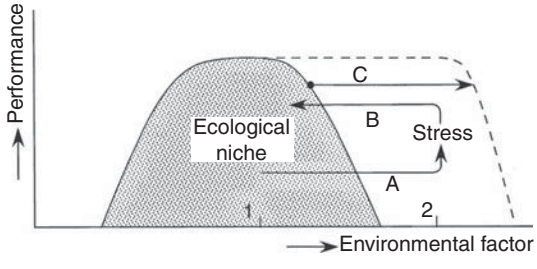


Figure 5.2 Graph illustrating a definition of stress based upon the ecological niche of a species. Ecological stress arises when the intensity of an environmental factor increases from 1 to 2 in such a way that in position 2 the organism is placed outside the niche (A). This will evoke stress and stress-response reactions, which fade away when the environmental factor relaxes and the organism returns to its niche (B). Another type of response is to move the border of the niche (C) by genetic adaptation in such a way that position 2 is not experienced as stress anymore. Reproduced from Van Straalen (2003), with permission from the American Chemical Society.

survival while the stress prevails. Such mechanisms cannot be maintained forever without consequences for normal cell function and so a return to the niche is necessary for long-term maintenance of fitness. The ultimate response implies that genotypes with a greater than average innate capacity to resist the stress are favoured and replace the ones with lower resistance in the next generation. Then, after some generations, the whole population consists of resistant genotypes (*genetic adaptation*). The boundaries of the niche have been shifted to include the organism's new position, and what was stress before is not stress anymore (Fig. 5.2).

The existence of genetic adaptation makes us realize that the ecological niche is not a property of a species as a whole, but may show variation between populations of a single species. In that case, a species with wide ecological amplitude (a *euryoecious* species) may consist of several local populations, each with narrow amplitude (*stenoecious* populations). Consequently, what is experienced as stress for one population is normal for another population of the same species. Such genetically determined polymorphisms in response to stress have been investigated often in evolutionary ecology and provide some fascinating examples of microevolution in real time, such as pesticide resistance in insects and metal tolerance in plants. In addition, stress may lower the threshold for expression of traits and so increase phenotypic variation and accelerate evolution (Hoffmann and Hercus 2000).

In our discussion of stress responses we will emphasize *conditions* more than *resources*. Resources are environmental factors that can be consumed and belong to the impact component of the ecological niche. Conditions are factors in the habitat, such as temperature, humidity, osmotic value, and oxygen tension, that are not consumed and can be altered only slightly by the organism. They can be plotted along an axis, as in Fig. 5.2, and the ecological amplitude of the species can be marked by values that depend on the species' physiology. We must add that our niche-based definition of stress does not encompass all stress phenomena. For example, stress may arise in animals upon sight of a predator, or when being chased away by a group member, or when witnessing an overwhelming natural event. These kinds of stress are difficult to relate to the concept of ecological niche, but many aspects of the internal state evoked by social and mental factors are similar to those imposed by a harsh environment.

A considerable number of functional genomic studies have been conducted with model organisms under stress. Due to the early availability of a full genomic microarray a lot of work has been done on yeast, which has developed into a classical model for the study of stress responses. There are also a fair number of genomic studies on stress responses in *Drosophila* and *Arabidopsis*. As in previous chapters, we will discuss the studies on models even when taken out of the context of their ecology, to illustrate the principles. From there we will try to draw conclusions about the relationship between stress and the ecological niche, which may hold equally for other species.

5.2 The main defence mechanisms against cellular stress

All organisms must deploy stress-defence systems with more or less specific tasks to cope with disturbances and restore normal physiological conditions after disturbance. These mechanisms are found in any cell and several systems are conserved throughout life, from prokaryotes to eukaryotes. The widespread occurrence of stress-defence mechanisms indicates that from very early in the evolution of life the defence against disturbances was a crucial problem to solve. Stress defence is thus closely linked to the idea of *homeostasis*, the tendency to regulate the internal state at a level independent from the changeable environment. Some systems that evolved in the early days of life have maintained their tasks mostly unchanged throughout evolution. We have already glimpsed some stress-defence systems in Chapter 4, when discussing the issue of longevity. It was noted then that upregulated stress defence in the wide sense is one of the most obvious signatures of a long life. In this section we will discuss the various stress-defence systems in more detail.

Korsloot *et al*. (2004) reviewed the cellular stress-defence responses with an emphasis on arthropods. In the book, the authors distinguished five different systems: (i) basal signal transduction systems, (ii) stress proteins, (iii) the oxidative stress response, (iv) metallothionein and associated systems, and (v) mixed-function oxygenase. It was also noted that there are many crosslinks between the different stress-defence systems. These cross-links help to coordinate the cellular response, which is needed to maintain integrity. In addition, many genes of the stress-defence system have promoters responding to more than one challenge. For example, the metallothionein promoter has metal-responsive elements enabling induction by metal stress, but it also has antioxidant-responsive elements and steroid hormone receptor-binding sites. Korsloot *et al*. (2004) even went one step further and argued that the different systems cooperate as a single, *integrated, cellular stress-defence system*. In the course of this chapter, we will meet genomic evidence that suggests this may indeed be the case. However, before presenting the genome-wide profiling studies we will briefly discuss the five best-investigated systems separately. Later sections will show that these five are by no means the only stress-responsive systems in the cell, but they serve to illustrate the most important principles of how stress-induced gene expression is brought about.

A theme common to all stress-induced gene expression is that stress signals converge on the activation of transcription factors, which bind to specific DNA sequences in the promoters of stress-induced genes. More generally these factors are called *trans*-acting factors and the DNA sequences to which they bind are *cis*-regulatory elements. We will discuss the architecture of gene expression regulation more extensively in Section 6.3. Here we note that screening of the 5´ region of a gene and identification of potential binding sites for transcription factors can help considerably to understand the biochemical context and the function of a gene. Conversely, when groups of genes are up- or downregulated in concert, and the same transcription factor-binding site appears in their promoters, this may indicate that they are regulated by the same transcription factor. One can never be sure, however, that a certain sequence, even if it conforms to a *cis*-regulatory element consensus sequence, is acting as a transcription factor-binding site *in vivo*, because transcriptional regulation is an extremely complicated process and is very much context-dependent (Wray *et al*. 2003). TRANSFAC® is a database on eukaryotic transcription factors, their genomic binding sites, and their DNA-binding profiles (www.gene-regulation.com/pub/databases.html). An overview of consensus sequences of transcription factor-binding sites appearing in this chapter is given in Table 5.1.

We will see in this chapter that there are many similarities across species in the way cells respond to stress. On the level of the cell, a universal mechanism seems to exist by which defence and homeostasis are organized. In addition, many aspects of cellular defence are not specific to the stressor, but are seen under a wide variety of conditions. Several authors have therefore argued that it should be possible to define a '*universal minimal stress proteome*', that is, a minimal set of proteins that are involved in the cellular stress response in all species and under all conditions. Kültz (2005) discussed the properties of such a universal stress response and identified 44 proteins as members of the minimal stress proteome. These proteins can be subdivided into six different functional groups:

- redox regulation (18 genes),
- DNA damage sensing and repair (4 genes),
- molecular chaperones (5 genes),
- protein degradation (6 genes),
- fatty acid and lipid metabolism (3 genes),
- energy metabolism (5 genes),
- other functions (3 genes).

Kültz (2005) also proposed making a distinction between the *cellular stress response* (*CSR*) and the *cellular homeostasis response* (*CHR*). While the cellular stress response can be viewed as a rescue operation aimed at restoring the integrity of macromolecules and redox potential after a disturbance, the cellular homeostasis response deals with environmental change and aims at maintaining homeostasis under variable conditions. The CHR is not triggered by macromolecular damage or oxidative burst but by stressor-specific sensors that monitor the conditions

outside the cell and trigger intracellular adaptive responses. For example, gene expressions in fish gills induced by changing osmolarity of the medium are part of the CHR, not necessarily the CSR. However, the two cascades, CHR and CSR, interact in numerous ways.

The scientific literature on cellular stress and homeostasis covers an extensive territory of biochemistry. It would be a hopeless task to try and cover all this ground here; instead we aim to present those aspects of stress-defence mechanisms that we think are necessary to understand the genomic studies on stress responses in an ecological context. We pay special attention to the pathways by which stress is translated into gene expression.

5.2.1 Stress-activated protein kinase signalling pathways

The principle of a signal transduction pathway was introduced in Chapter 4. An extensively discussed system in that chapter was the insulin/IGF receptor, which proved to be associated with the regulation of longevity in various animals. The insulin-signalling pathway relays a hormonal signal into a cellular response, via a cascade of molecular interactions, eventually leading to inactivation of a transcription factor. We saw the same principle in the TOR and Ras pathways acting upon body size in insects and in the phytochrome signalling pathway regulating shade avoidance in plants. The principle of signal transduction is also employed in

Table 5.1 Overview of stress-related transcription factors and consensus sequences of their DNA binding sites (http://www.gene-regulation.com/pub/databases.html)

Transcription factor	DNA-binding site	Consensus sequence (5′ → 3′)	Context
Activator protein 1 (AP-1)	AP-1-binding element	TGA(C/G)T(A/C)A	General stress response
Heat-shock factor (HSF)	Heat-shock-responsive element (HSE)	NGAANNGAANNG AAN	Heat shock, general stress response
Nuclear factor erythroid 2-related factor 2 (Nrf2)	Antioxidant-responsive element (ARE), electrophile- responsive element (EpRE)	TGACNNNGC	Induction of antioxidant enzymes and ROS-scavenging systems
Metal-responsive-element-binding transcription factor (MTF-1)	Metal-responsive element (MRE)	TGC(A/G)CNC	Induction of metallothionein
Aryl hydrocarbon receptor (AhR)	Xenobiotic-responsive element (XRE), dioxin-responsive element (DRE), aryl hydrocarbon-responsive element (AhRE)	TGCGTGAGAAGA (human, mouse, rat, guinea pig)	Induction phase I and II biotransformation enzymes
Centromere-binding factor 1 (CBF1)	Centromere DNA element 1 (CDE1)	(A/G)TCAC(A/G)T G (yeast)	Regulation of sulphur amino acid biosynthesis pathway, induced by exposure to heavy metals
Msn2p, Msn4p	Stress-response element (STRE)	GATGACGTGT (Msn4) (yeast)	General stress defence and carbohydrate homeostasis in yeast
DREB family transcription factors	Dehydration-responsive element (DRE)	CCGAC	Induction of defence against cold and drought in plants
Abscisic acid-responsive element binding factor (ABF)	Abscisic acid-responsive element (ABRE)	(C/T)ACGTGGC	Water stress signalling in plants
Hypoxia-inducible factor (HIF-1)	Hypoxia responsive element (HRE)	TACGTGCT (human)	Metabolic switches related to hypoxia
Oestrogen receptor (ER-α)	Oestrogen-responsive element (ERE)	AGGTCANNNTGA CCT (human, mouse, cattle, clawed frog, chicken)	Activation of processes related to oogenesis and other female functions

Notes: In some cases more than one name is given to the same factor or binding site. Some consensus sequences are specific to certain groups of organisms (where indicated). N, any nucleotide; ROS, reactive oxygen species.

one of the most elaborate stress-responsive systems, the *mitogen-activated protein kinase pathway* (MAPK pathway; Chang and Karin 2001).

The name MAPK derives from the action of mitogens, chemical substances that stimulate cell division and often have a tumour-promoting action. Several secondary metabolites from plants are notorious mitogens, the best known being phorbol esters from the family Euphorbiaceae (spurges), such as the tropical plant *Croton tiglium*. The mitogenic action of phorbol esters is due to their activation of the MAPK pathway; some of them are used as model compounds in studies of cell biology. It became clear later that MAPK is also involved in transducing stress signals.

The classical MAPK pathway is activated by the binding of an extracellular signal molecule such as a mitogen to a receptor protein on the cell membrane (Fig. 5.3). In addition to mitogens, MAPK is sensitive to cytokines, hormones, and a variety of other molecules. These substances activate a *receptor tyrosine kinase* (RTK), a molecule with an extra-

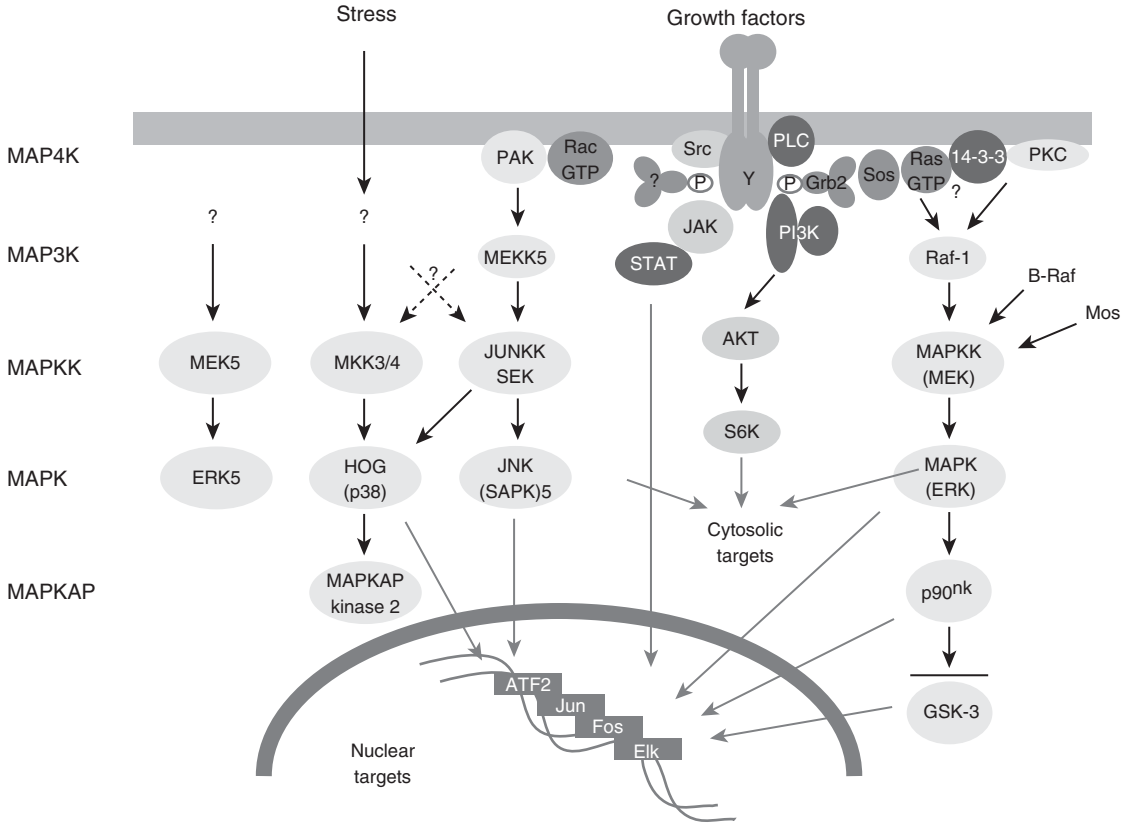


Figure 5.3 Scheme of the mammalian MAPK signalling pathway, illustrating the great complexity of reactions. The figure shows a portion of a cell with the plasma and nuclear membranes. The classical pathway (on the right) starts with binding of a growth factor to an extracellular receptor (Y, receptor tyrosine kinase), leading to three successive waves of kinase activity (MAP3K, MAPKK, MAPK) and activation of MAPKAPs. When activated, several of the downstream kinases can translocate to the nucleus and trigger transcriptional regulation of a variety of genes. Kinases can also influence cytoplasmic targets and contribute to translational control. Parts of the MAPK pathway are activated by stress signals (on the left), but the cascade of events is less well known than that triggered by cell growth factors. ATF, activating transcription factor; GSK, glycogen synthase kinase; JAK, Janus kinase; PAK, p21-activated kinase; PI3K, phosphoinositide 3-kinase; PKC, protein kinase C; PLC, phospholipase C; S6K, ribosomal protein S6 kinase; SEK, SAPK kinase; STAT, signal transducer and activator of transcription.

cellular part to which the signalling molecule can bind, and an intracellular domain with kinase activity. When the extracellular receptor is occupied, a conformational change is imposed upon the intracellular domain, which then becomes an attractive binding site for adapter proteins, and their binding triggers further molecular reactions. The cascade of reactions following on RTK activation is extremely complex. More than 100 different proteins have been described to participate in the MAPK network; many of them are *kinases*, enzymes that affect other proteins by phosphorylating amino acids critical for the three-dimensional structure, resulting in activation or inactivation of the target. The MAPK cascade involves three successive tiers of kinase activity, which—from downstream to upstream—are designated MAPK, MAPK kinase (MAPKK, MKK, or MEK), and MAPKK kinase (MAPKKK, MAP3K, MEK kinase, or MEKK). The upstream proteins associated with binding to RTK and activation of MAP3K are sometimes called MAP4K (Fig. 5.3). Downstream of the three-tiered cascade, MAP kinases activate *effector kinases*, also called MAPK-activated proteins (MAPKAPs).

One way in which the MAPK proteins, and the effector kinases activated by them, exert their action is by translocation to the nucleus. An example is a kinase called *extracellular signal-regulated kinase* (ERK), which when activated can pass through the nuclear envelope and there activate transcription factors. Important targets for ERK are c-Jun and c-Fos, which are components of the dimeric transcription factor *activator protein 1* (AP-1), a protein which turns up as a downstream effector in various stress responses. The combination of two different proteins into a single active unity is called *heterodimerization*. Many of the transcription factors activated by MAPK are dimeric proteins, which only become active by combining the two components into one functional protein. Although transcription factors are important MAPK targets, MAPKs may also influence post-transcriptional processes in the cytoplasm, for example by contributing to mRNA stabilization.

MAPK is also activated by stress (Fig. 5.3). The part of the pathway that is specifically associated with the stress response is called the *stress-activated protein kinase* (SAPK) *signalling pathway*. However, the chain of events in SAPK is less well known than in the case of the classical pathway activated by mitogens. The stress signal may be transduced along a distinct pathway, interacting with the classical pathway, but it may also affect enzymes of the classical pathway directly. Two important kinases which are activated by stress and can translocate to the nucleus to activate transcription factors are *HOG* (product of high osmolarity gene) and *c-Jun N-terminal kinase* (JNK). Some of the MAPK proteins are known under different names in different organisms; for example HOG was first described in yeast but its orthologue in higher organisms is known as p38, whereas JNK is also known as SAPK5.

As noted above, the action of protein kinases involves phosphorylation of certain amino acids in other proteins. The two stress-related effector kinases, HOG and JNK, can only transfer to the nucleus when phosphorylated. However, they may also be dephosphorylated by the action of protein phosphatases. Phosphatases specific for MAPK are called *MAPK phosphatases* (MKPs; Tamura *et al*. 2002). The action of these phosphatases is very important for relaxation of the MAPK signal. There are four different MKP families; some of them act on specific effector kinases (e.g. only on JNK) and others have a broader action spectrum (e.g. they dephosphorylate three effector kinases—HOG, JNK, and ERK). Interestingly, MKP genes are among the many genes activated by MAPK signalling. Their activity thus constitutes a negative feedback, which attenuates the MAPK signal after being triggered by stress or a growth factor.

One of the possible ways in which stress may activate SAPK signalling is by inhibition of MKPs. Oxidative stress in particular may lead to the oxidation of sensitive thiol groups in protein phosphatases and so allow SAPK signalling by removing the negative feedback (Korsloot *et al*. 2004). Direct activation of kinases is another possible mechanism (Fig. 5.3).

The fact that different stimuli all activate MAPK signalling, yet can activate different genes, suggests the presence of a regulatory or coordinative system within the cascade. The existence of distinct but not mutually exclusive mechanisms, and the involve-

ment of a large number of kinases, many of them acting upon each other, could contribute to the achievement of specificity in the responses to different stimuli. Another aspect contributing to specificity is the time-dependency of kinase activity. Depending on the strength of feedback loops between the proteins activated by a certain stress factor, the downstream signal may come as a single pulse or as a sustained push, which could trigger different effector proteins.

5.2.2 Heat-shock proteins

Heat-shock proteins are the best-investigated components of the cellular stress response. The traditional way of studying them is by applying a heat shock: exposure of the organism or a cell culture for a period of 30 min to a few hours to supraoptimal temperature, such as 39 ºC. Later research showed that heat-shock proteins are not only induced by heat but also by many other environmental stress factors, including cold, food depletion, osmotic stress, and toxicants, and so the term *stress proteins* is actually more appropriate. However, the term heat-shock protein is now so widely used and accepted that we retain it here.

Heat-shock proteins have been found in all organisms in which they were sought (Feder and Hofmann 1999). In addition, essential features of their molecular structure are conserved over the entire tree of life, from Archaea to mammals. This extremely wide occurrence and conservation suggests that heat-shock proteins must play a very fundamental role in the cellular defence against stress. The consensus view is that this role is due to support in protein folding and unfolding, which involves four aspects: (i) stabilizing essential structural proteins, especially around the nucleus and of organelles such as ribosomes and spliceosomes, (ii) assisting transfer of proteins across membranes, by unfolding and refolding, (iii) supporting refolding of proteins denaturated by stress, and (iv) supporting the degradation of aberrant proteins. On the basis of their ability to align with other proteins and facilitate changes in their three-dimensional structure, heat-shock proteins are often called *molecular chaperones*. Possibly the conservation of heat-shock proteins throughout all living beings stems from the fact that a very specific three-dimensional structure is required for conducting this chaperone function. The ability of heat-shock proteins to protect cells against the adverse effects arising from accumulation of denatured proteins can be seen as a logical extension of their normal function as molecular chaperones.

The diversity of stress proteins varies per taxonomic group. As an example we discuss here the stress proteins of *Drosophila* (Table 5.2; Korsloot *et al*. 2004). Four families may be discerned; Sp90, Sp70, small heat-shock proteins, and ubiquitin. The proteins themselves are named after their apparent molecular mass on an electrophoresis gel, so Hsp70 has a molecular mass of about 70 kDa. Another type of classification is between *inducible* and *constitutive* proteins. Proteins of the latter category are called *heat-shock cognate proteins* (e.g. Hsc70). They are always present in the cell during normal physiological function and are not usually induced by stress factors, although under certain circumstances they may be induced a little. Proteins of the former category, especially Hsp70, are known for their very strong inducibility. Induction of heat-shock proteins was long used as a classical model of gene regulation, because the heat shock is so easily brought about and induction is very strong. Even nowadays, the promoters of heat-shock genes are often deployed in gene-expression studies, using genetic constructs in which the coding region of a gene of interest is fused with a promoter from an inducible heat-shock gene.

Organisms may have more than one representative of the heat-shock proteins indicated in Table 5.2. For instance, *D. melanogaster* has two types of Hsp70, encoded by genes at different chromosomal loci. At each locus, two or more repeats of the same gene are present, depending on the fly stock. The organization of heat-shock genes varies both within and between species and this variation is possibly relevant for the ecological response to environmental extremes (Feder and Hofmann 1999).

Each of the stress proteins mentioned in Table 5.2 has specific functions during normal cell metabolism. Members of the Sp90 family, including *Drosophila* Hsp83, are constitutively present, but

Table 5.2 Overview of stress proteins in *D. melanogaster*, following Korsloot *et al*. (2004)

Family	Name	Location in unstressed cell	Location under cellular stress	Function
Sp90	Hsp83	Cytoplasm	Cytoplasm	Stabilizes specific receptor proteins and chaperones peptides to membranes
Sp70	Hsp70	Nucleus (cytoplasm)	Nucleus, nucleolus, cytoplasm	Highly inducible, rescue of aberrant proteins during stress response
	Hsp68	Nucleus, cytoplasm	Nucleus, cytoplasm	Inducible; function unknown
	Hsc70	Cytoplasm	Cytoplasm, nucleus	Protein binding in cytoplasm
	Hsc70b	Cytoplasm	Cytoplasm (nucleus)	Protein binding in cytoplasm
	Hsc71	Mitochondrion	Mitochondrion	Protein binding in mitochondrion
	Hsc72	Endoplasmic reticulum	Endoplasmic reticulum	Protein binding in endoplasmic reticulum
Small Hsps	Hsp22, Hsp23, Hsp26, Hsp27	Cytoplasm	Nucleus (cytoplasm)	Participation in development and stress response
Ubiquitin	Ubiquitin	Nucleus, nucleolus, cytoplasm	Cytoplasm (nucleus)	Histone binding, tagging proteins for degradation

heat, anoxic conditions, and glucose deprivation increase their level. They are sometimes referred to as *glucose-regulated stress proteins*. A specific role for Hsp83 lies in stabilizing steroid hormone receptors. As long as a steroid does not occupy the receptor, Hsp83 binding ensures a receptive configuration. Upon binding of the hormone, Hsp83 detaches from the receptor, which then takes another configuration; the receptor may then enter the nucleus to act as a transcription factor. This is a mechanism by which cells sensitive to steroid hormones translate a hormonal signal into gene expression. A similar function of Hsp83 lies in stabilizing the heat-shock factor, HSF, and the aryl hydrocarbon receptor (see below). Expression of Hsp83 varies with life stage and tissue; Hsp83 is developmentally induced in gonad tissue during oogenesis, which is understandable from the steroid hormone-receptor-binding function.

The Sp70 family includes the inducible heat-shock proteins Hsp70 and Hsp68 (Table 5.2). These proteins are abundantly expressed within minutes after commencement of a heat treatment. After induction Hsp70 is mostly found in the nucleus and on cell membranes, where it ensures protection of essential proteins against denaturation. In the recovery phase after termination of a heat treatment, Hsp70 is translocated to the cytoplasm and participates in the degradation of damaged proteins. It is likely that Hsp68 fulfils a similar role, but its precise function is less well described. When the cell is not in stress, Hsp70 is hardly detectable, so the induction has the characteristics of a rescue operation, with Hsp70 turning out only in case of emergency. However, we know from Chapter 4 that a modest continuous upregulation of Hsp70 can have beneficial effects: it increases lifespan. A large continued increase in Hsp70 levels has detrimental effects, especially in tissues with a high rate of cell division and growth. The reason is, as we will see below in more detail, that induction of Hsp70 is accompanied by a redirection of protein synthesis in which priority is given to heat-shock protein synthesis and degradation of aberrant proteins, while normal protein synthesis is blocked.

The heat-shock cognate proteins Hsc70 and Hsc70b have roles similar to the inducible stress proteins, but they are constitutively present at high levels to assist in the folding of peptides. In particular, polypeptide chains that leave the ribosome after translation need assistance to assume the correct three-dimensional structure. It is assumed that negatively charged clusters in the Hsc bind to positively charged amino acid residues of the peptide; this binding changes the local conformation of the complex and causes a fold in the chain by forcing any neutral residues adjacent to the charged residues into the relatively hydrophobic environment inside

the fold. The other two heat-shock cognate proteins, Hsc71 and Hsc72, are active in the mitochondrion and the endoplasmic reticulum, respectively. Both proteins are encoded in the nuclear genome, so Hsc71 must be transported across the mitochondrial membrane after translation in the cytoplasm.

Low-molecular-mass heat-shock proteins vary considerably in molecular mass and number across species and they form a rather heterogeneous collection with limited sequence similarity to each other. Four different small heat-shock proteins have been identified in *Drosophila* (Table 5.2). The regulation and function of small heat-shock proteins are complicated. Some small heat-shock proteins are activated by phosphorylation in response to stimuli that also activate the MAPK signalling pathway (see above). These are constitutively present heat-shock proteins that play a role in protecting and chaperoning special proteins and enzymes in normal cell metabolism. However, other small heat-shock proteins are induced in large quantities without phosphorylation and these inducible heat-shock proteins seem to be directed mainly towards nuclear structures and RNA.

The fourth family of heat-shock protein consists of the small globular protein *ubiquitin*, which like the other heat-shock proteins is extremely well conserved; only three amino acids differ between yeast and *H. sapiens*. Ubiquitin can be conjugated to other proteins by a ubiquitin protein lyase and when proteins are 'tagged' in this way they are destined for ATP-dependent cytoplasmic protein degradation. The cytoplasmic proteolytic system promotes the turnover of proteins and increases availability of free amino acids for protein synthesis. Another regulatory role for ubiquitin is due to its binding to histones. In Chapter 5 we saw that active transcription of DNA requires histones to be acetylated; if not, histones will interact with each other and cause condensation of chromatin. A similar effect is due to ubiquitin: ubiquitination of histones prevents chromatin condensation. Under stressful conditions an enzyme, ubiquitin hydrolase, is induced, which removes ubiquitin from the histones (*de-ubiquitination*) and thereby causes a general decrease of DNA processing. The liberated ubiquitin moves to the cytoplasm and participates in tagging proteins for proteolytic degradation. The system can thus be understood as a shift of priorities under stress from DNA transcription to protein degradation.

How is the activity of inducible heat-shock proteins regulated? One way in which the cell can increase the amount of heat-shock proteins is by transcriptional activation of *Hsp* genes. Inducible heat-shock protein genes all share certain conserved sequences in their promoter, so-called *heat-shock elements* (HSEs), see Table 5.1. These are the binding sites for a very important transcriptional activator, *heat-shock factor* (HSF). In yeast and *Drosophila* only one HSF has been identified, but in mammals there appear to be two, HSF1 and 2, where HSF1 is homologous to *Drosophila* HSF. Biochemical research has shown that HSF needs to undergo *homo-trimerization* to become active as a transcription factor (Fig. 5.4). This means that three HSF molecules must bind together; the trimeric HSF can then bind to the HSEs in the promoter of heat-shock genes.

In unshocked *Drosophila* cells, HSF is present in the nucleus in the monomeric form. Binding of heat-shock proteins, in particular Hsp83 and Hsp70, to HSF monomers prevents trimerization (Morimoto *et al*. 1994; Santoro 2000). When heat shock leads to the appearance of aberrant proteins in the cytoplasm, heat-shock proteins are drawn to these proteins with an affinity greater than their binding to HSF. Consequently, HSF is no longer prevented from trimerization and the activated HSF may bind to the heat-shock elements in the promoter of *Hsp* genes. The binding itself is however not yet sufficient for maximum transcriptional activation; HSF must also be phosphorylated. This is brought about by a protein kinase, which by itself is activated by an external stimulus, transduced, for example, through the SAPK pathway. The activation of a transcription factor by phosphorylation while already bound to DNA is an example of *transactivation*. The complete model of HSF activation is shown in Fig. 5.4.

The scheme illustrated in Fig. 5.4 illustrates how environmental stress regulates the induction of heat-shock protein at three levels: (i) translocation of protein kinases to the nucleus, (ii) removal of the inactivation conferred by heat-shock proteins on HSF, and (iii) transactivation of HSF by protein

kinase. Even this rather complicated scheme is still a simplification. For example, trimerization of HSF is not only dependent on the appearance of aberrant proteins and heat-shock-protein detachment; it may also be directly triggered by changes in the cellular redox state and by a peak in Ca^{2+} ions. In addition, the details are likely to differ from one species to another. Our scheme is mostly inspired by the situation in *Drosophila*, which is similar to the situation in mammals except that some heat-shock proteins have different molecular masses; for example, Hsp83 of *Drosophila* is homologous to human Hsp90. The heat-shock factor of yeast differs considerably from animal HSF. It lacks one of the domains (a leucine zipper) that allows stabilization of the monomeric configuration. Consequently, it is not stored in monomeric form in the nucleus like the animal HSF, but is already bound to the DNA in its trimeric form, requiring only phosphorylation to be activated.

The regulatory mechanisms depicted in Fig. 5.4 all relate to transcriptional processes; however, the stress response also includes regulatory mechanisms at the level of RNA processing, translation, and mRNA degradation. One of the processes that is downregulated during severe heat shock is RNA splicing: the extrusion of intron sequences from the primary transcript, occurring at specialized bodies in the nucleus, the spliceosomes. The consequence is that all pre-mRNAs that have introns in their sequence cannot be processed and are not translated into functional proteins. Interestingly, the

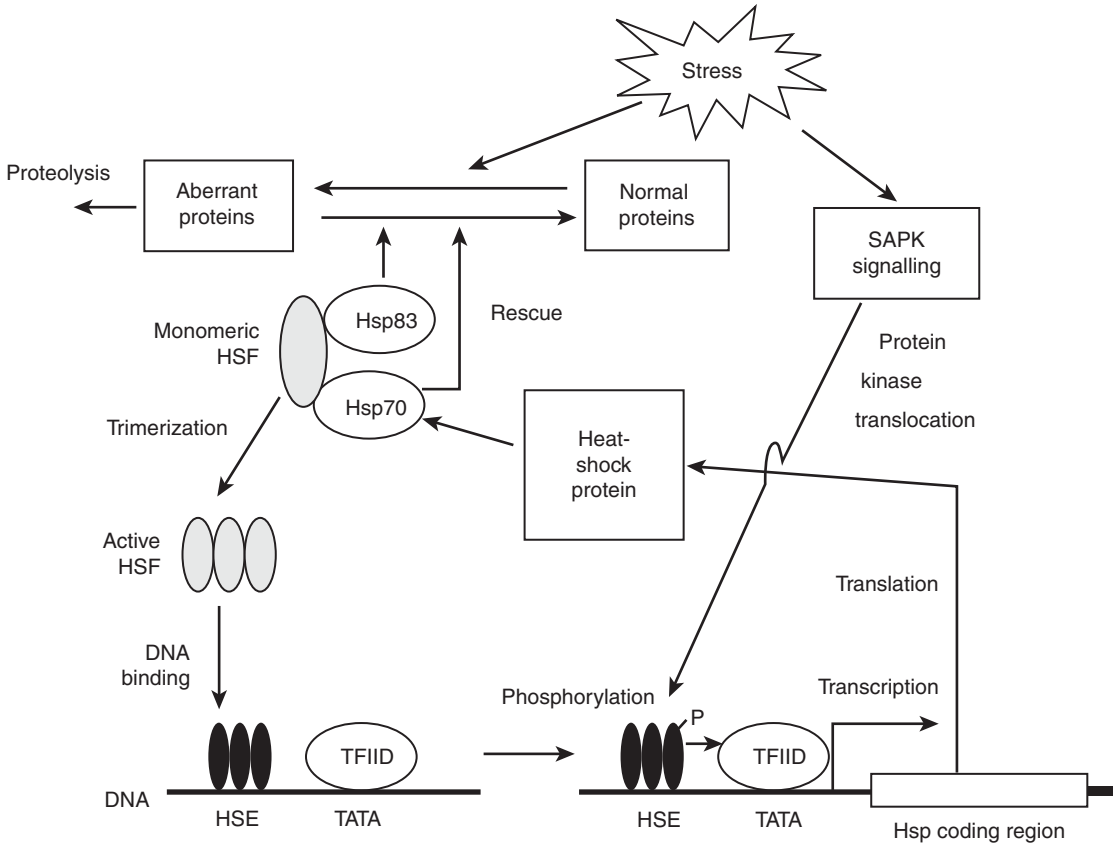


Figure 5.4 Schematic representation of the chain of events leading to induction of heat-shock proteins as a consequence of stress. TATA, TATA box; TFIID, transcription factor IID, a TATA-box-binding protein facilitating the formation of a transcription-initiation complex.

inducible heat-shock genes themselves lack any introns and so can circumvent the splicing block. Translation itself is not inhibited: it can proceed and produce a massive amount of heat-shock protein. In addition, as we saw above, stress may cause de-ubiquitination of histones, because ubiquitin moves to the cytoplasm where it is needed for tagging damaged proteins. The loss of ubiquitin from histones causes chromatin condensation, which is another mechanism to suppress protein synthesis. Obviously, the inducible heat-shock genes must be exempt from inactivation by chromatin condensation, which is brought about by a basal promoter-binding element (known as a GAGA factor) being permanently bound to the *Hsp* promoter; this protein displaces the histone-containing nucleosomes from the gene and so prevents chromatin condensation around the *Hsp* gene.

The heat-shock response continues to act as a model *par excellence* of homeostatic mechanisms of the living cell. It illustrates a great variety of biochemical regulatory mechanisms, of which we have discussed only the most salient features. Later in this chapter we will see how important heat-shock proteins are in the genome-wide responses to stress.

5.2.3 The oxidative stress-response system

Under the term oxidative stress come a variety of phenomena that are an unavoidable consequence of aerobic metabolism. By definition all aerobic organisms need oxygen, but at the same time they must avoid the inherently cytotoxic effects of oxygen. Toxicity of oxygen is not due to O_2 itself but to reactive oxygen derivatives, generated by cellular processes. These derivatives are jointly referred to as *reactive oxygen species* (ROS). Several of these ROS are *free radicals*; that is, molecules or elements with one or more unpaired electrons in the outer orbital. The best-known ROS are the free radicals superoxide radical ($O_2^{\bullet -}$), hydroxyl radical ($OH^\bullet$), and nitric oxide radical ($NO^\bullet$; the lack of a paired electron is indicated by $^\bullet$ in the chemical formula). Non-radical ROS are hydrogen peroxide (H_2O_2) and singlet oxygen ($^1O_2$). Molecular oxygen itself has two unpaired electrons, so strictly speaking it is also a radical, but the reactivity of O_2 is limited due to the fact that the two electrons have equal spin (the so-called spin restriction). In ROS the spin restriction is lifted, which is why these forms are called activated oxygen.

ROS are produced at many places in the cell. Superoxide anion is produced abundantly in the respiratory chain of the mitochondria, the light-harvesting reactions of the choroplast, the reduction–oxidation reactions catalysed by cytochromes of the smooth endoplasmic reticulum, and the xanthine dehydrogenase pathway, which is involved with the degradation of purines to urate. Singlet oxygen is produced by so-called photo-sensibilization reactions, in which light energy is absorbed by molecules such as riboflavin, chlorophyll, and retinol, and transferred to molecular oxygen. Not all ROS are equally reactive. The most reactive species are hydroxyl radical and singlet oxygen, which react immediately with a suitable molecule and so inflict injury mainly on local cellular structures. H_2O_2 and superoxide anion are more stable and can move through the cell by diffusion; H_2O_2 can even pass cell membranes. The damaging effect of these ROS is mostly due to their ability to generate hydroxyl radicals. These reactions are catalysed by transition metals, such as iron and copper. For example, $OH^\bullet$ is generated from H_2O_2 in the so-called *Fenton reaction*:

$$Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + OH^\bullet + OH^-$$

Although the concentration of free iron in the cell is very low (most iron is bound in porphyrins and ferritin and not available for the Fenton reaction), some iron is bound to low-molecular-mass chelators such as citrate, and this can catalyse the generation of $OH^\bullet$ from H_2O_2.

In Chapter 4 we saw that ROS are implicated in aging and senescence according to the free radical theory of aging. The cellular basis for this theory is that ROS may cause damage to many macromolecules in the cell, including proteins, DNA, and lipids. Protein damage may be caused by oxidation of free thiol (SH) groups, leading to loss of function. DNA damage may be due to thymidin dimers and strand breaks. Lipid damage is due to a process known as *lipid peroxidation*. This is a chain reaction

in which an oxygen radical abstracts a hydrogen atom from an unsaturated bond in a fatty acid chain, which is followed by molecular rearrangement and uptake of oxygen, to form a lipid peroxyradical ($LOO^{\bullet}$), which can abstract $H^{\bullet}$ from another lipid. Polyunsaturated fatty acid chains of membrane lipids are particularly sensitive to lipid peroxidation. Peroxidation causes loss of membrane flexibility, loss of activity of membrane-bound enzymatic processes, and, in the most severe form, lipid destruction followed by the appearance of volatile alkanes, alkenes, and aldehydes.

Obviously, there are great advantages in the use of oxygen gas as an electron acceptor in cellular respiration; however, damage induced by ROS is an unavoidable side effect. Protection against ROS damage is an evolutionary necessity that must compensate for the disadvantage. A great variety of antioxidant systems are deployed, using two strategies: *scavenging* (neutralization of ROS by reaction with a reductant) and *enzymatic transformation* (dismutation or reduction) to a non-reactive form. Table 5.3 provides an overview of the major antioxidant systems. Scavenging and enzymatic systems to protect cells from oxidative stress are found in almost all organisms; however, anaerobic and microaerophilic bacteria and Archaea have an oxidoreductase not found anywhere else (Lumppio *et al*. 2001). Several antioxidant enzymes are localized in specific organelles of the cell (Table 5.3).

Among the different protective systems listed in Table 5.3 we briefly highlight the role of *glutathione*, which acts both as a scavenger on its own and as a substrate in enzymatic reactions. Glutathione is a tripeptide, γ-glutamyl-cysteinyl-glycine, which in reduced form (written GSH) has a free thiol group on the cysteine residue. In the transition to the oxidized form (GSSG) two thiol groups react with each other, losing two electrons, as in the reaction catalysed by glutathione peroxidase:

$$H_2O_2 + 2GSH \rightarrow GSSG + 2H_2O$$

Table 5.3 Major antioxidant systems protecting the cell against injury from radical oxygen species

Antioxidant system	Primary localization	Actions
Copper- and zinc-containing superoxide dismutase (Cu/Zn SOD)	Cytosol, nucleus	Catalyses dismutation of O_2^{i-} to H_2O_2
Manganese-containing superoxide dismutase (Mn SOD)	Mitochondrion	Catalyses dismutation of O_2^{i-} to H_2O_2
Rubredoxin oxidoreductase	Only found in anaerobic bacteria and Archaea	Catalyses reduction of O_2^{i-} to H_2O_2
Catalase (CAT)	Peroxisomes	Catalyses reduction of H_2O_2 to H_2O
Glutathione peroxidase (GSH-Px)	Cytosol, mitochondrion	Catalyses reduction of H_2O_2 to H_2O
Glutathione peroxidase (GSH-Px,m)	Lipid membranes	Catalyses reduction of lipid hydroperoxides
Glutathione reductase	Cytosol, mitochondrion	Catalyses reduction of low-molecular-mass disulphides
Ascorbate peroxidase	Chloroplast (only in plants)	Catalyses reduction of H_2O_2 to H_2O
Thioredoxin (TRX)	Cytosol, mitochondrion, nucleus	Restores redox state by reducing oxidized thiols, scavenges H_2O_2 and $OH^{\bullet}$
α-Tocopherol (vitamin E)	Lipid membranes	Scavenges H_2O_2, $OH^{\bullet}$, and $LOO^{\bullet}$; interrupts lipid peroxidation
β-Carotenoid (vitamin A)	Lipid membranes	Scavenges O_2^{i-} and peroxyl radicals
Ascorbate (vitamin C)	Throughout the cell	Scavenges O_2^{i-} and $OH^{\bullet}$; contributes to regeneration of vitamin E
Glutathione	Throughout the cell	Scavenges O_2^{i-} and organic free radicals; substrate in enzymatic reduction reactions

Source: From various sources.

Oxidized glutathione is then again reduced at the expense of reduction equivalents from NADPH by means of glutathione reductase:

$$GSSG + NADPH + H^+ \rightarrow 2GSH + NADP^+$$

Since glutathione is responsible for the majority of ROS-protection reactions and it is also involved in conjugating lipophilic substances in xenobiotic biotransformation reactions (see below), the maintenance of a large pool of reduced glutathione is very important for the vitality of the cell. Cellular stress due to ROS, altered redox state, and xenobiotic metabolism may cause *glutathione depletion*.

To be effective as protective mechanisms, the antioxidant systems listed in Table 5.3 must be upregulated when the cell perceives oxidative stress. This is indeed the case. Recent research has shown that genes encoding antioxidant protective enzymes all have a characteristic sequence in their promoter, designated the *antioxidant-responsive element* (ARE). This element is also called the *electrophile-responsive element* (EpRE), after researchers discovered that not only oxidative stress but also electrophile chemicals could activate the element. As in the case of heat-shock proteins, the sequence serves to bind a transcription factor which is activated by stress. The factor binding to AREs has been identified as *Nrf2*, also called NF-E2 (an abbreviation of *nuclear factor erythroid 2-related factor 2*; Nguyen *et al*. 2003, 2004; Jaiswal 2004; Kobayashi and Yamamoto 2005). Under normal physiological conditions, Nrf2 is continuously degraded under the influence of a protein known as *Keap1* (*Kelch-like ECH-associating protein 1*). This protein supports the tagging of Nrf2 with ubiquitin, after which it is destined for cytoplasmic proteolysis. Degradation of Nrf2 is compensated by continuous synthesis, but under normal conditions the pool of Nrf2 in the cytoplasm is very small. Under conditions of oxidative stress Keap1 is inactivated and Nrf2 is stabilized (Fig. 5.5). Presumably, the oxidative-stress signal is transduced through the SAPK pathway leading to activation of protein kinase C (PKC) and other cytosolic factors. Phosphorylation of Nrf2 by PKC results in the release of Nrf2 from its repressor. Nrf2 may then translocate to the nucleus, and will bind to an ARE. However, before becoming fully active it needs to undergo *heterodimerization*. This is comparable to the formation of activator protein AP-1 by heterodimerization of c-Jun and c-Fos (see Section 5.2.1 on SAPK signalling). In the case of Nrf2, the partner protein is suggested to be a member of the family of Maf proteins. Maf proteins are nuclear factors that, like Nrf2, may bind to ARE. In the absence of Nrf2, they exert negative control over ARE-mediated gene expression. How this negative control is lifted by Nrf2 is not known precisely. Anyway, expression of the antioxidant system seems to be regulated by both negative and positive agents and this could allow fine-tuning to the physiological needs of the cell.

Our present knowledge of the regulation of antioxidant systems is almost completely limited to human cells and *Drosophila*. One of the reasons for this is that the issue raises a good deal of interest in medical research: tumour cells are characterized by an upregulated oxidative-stress response. How the knowledge generated in the medical sector translates to an ecological context of animals and plants in the wild is difficult to evaluate at the moment; however, it may be expected that, like the stress-activated signalling pathways and the stress proteins, most of components of the oxidative-stress response are evolutionarily conserved. In the genome-wide studies discussed later in this chapter we will discuss some evidence supporting this statement.

5.2.4 Metallothionein and associated systems

Metallothionein is a peculiar protein with an extremely high affinity for free metal ions such as Zn^{2+}, Cd^{2+}, and Cu^{+}. In the presence of sufficient metallothionein the free concentrations of these metals in solution are reduced to the picomolar range. Interestingly, iron is not bound by metallothionein but instead has its own pathways of uptake and intracellular transport, involving ferrotransferrin and ferritin. As we will see below, changes in iron trafficking are part of the general stress response, but the regulation of these changes does not appear to interact with metallothionein

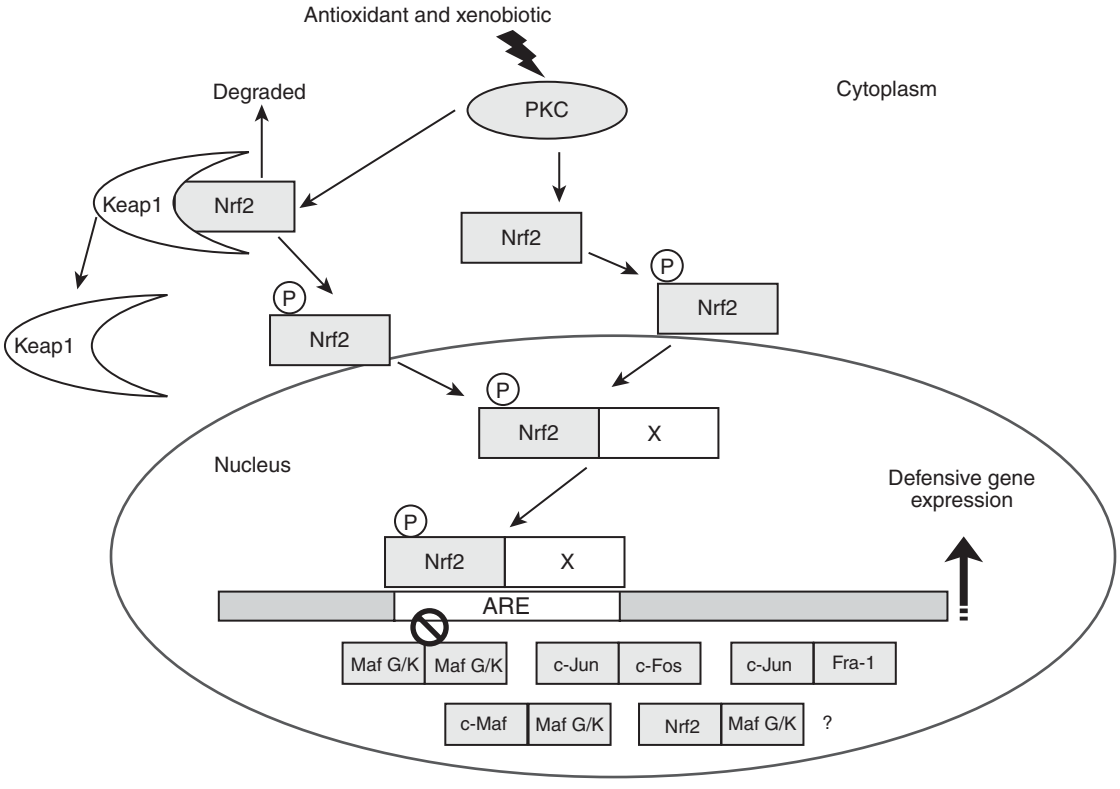


Figure 5.5 Scheme of Nrf2 activation and induction of antioxidant genes by ARE binding. Nrf2 in the cytoplasm is bound to a protein, Keap1, which promotes its degradation by ubiquitination. An oxidative-stress signal, transduced via SAPK, activates protein kinases that may phosphorylate Nrf2 and liberate it from inactivation of Keap1. In addition, Keap1 may be destabilized by reactive chemicals, for example through alkylation of cysteine residues critical for its activity. Either mechanism allows Nrf2 to translocate to the nucleus. There it undergoes dimerization with an as-yet unknown partner (X) and triggers expression of genes with AREs in their promoter. A variety of nuclear factors (Maf G/K and others) normally inhibit ARE-mediated gene expression, but are removed by the Nrf2/X heterodimer. From Jaiswal (2004), with permission from Elsevier.

and therefore we leave iron out of consideration here.

As the name indicates, metallothionein contains not only a high amount of metal but also sulphur. This is due to an extraordinarily high percentage of cysteine residues (around 30%). These cysteines participate in the formation of *metal-thiolate clusters*, in which the thiol groups of several cysteines coordinate with a group of metal atoms. In the metallothionein of mammals there are two clusters, one binding four metal ions using eleven cysteines, the other binding three metal ions with nine cysteines. These two clusters appear as two separate protein domains, a C-terminal α-domain (four-metal cluster) and an N-terminal β-domain (three-metal cluster), which are separated by a short linker sequence (Fig. 5.6). There are no aromatic amino acids in metallothionein; the whole protein is very hydrophilic.

The two-cluster structure of metallothionein has also been found in other vertebrates, invertebrates, and plants; however, in some species the α and β clusters are in reversed configuration with respect to their N- and C-terminal positions. Other species have two three-metal clusters. The metallothionein genes (*Mt*s) of *Drosophila* are an oddity among animals. *Drosophila* has four *Mt* genes, each encoding a small metallothionein with one metal-binding domain (Valls *et al*. 2000; Egli *et al*. 2003). Such single-cluster *Mt*s have not been found in other

animals up to now, but they are present in fungi. In plants the *Mt* genes encode proteins in which the two metal-thiolate clusters are connected by a very long link of non-cysteine amino acids (Cobbett and Goldsbrough 2002). So it appears that the evolution of metallothionein, quite unlike the stress proteins and the components of signalling pathways discussed above, has come with a considerable reshuffling of the molecule.

When metallothionein was first described, the function attributed to it was to regulate the cellular concentrations of essential metals. This was assumed to involve donating metals to specific metal-requiring ligands (enzymes, zinc fingers, structural proteins), while preventing aspecific binding to macromolecules by keeping the free concentrations of metals very low. In addition some metallothioneins turned out to be highly inducible by non-essential heavy metals (e.g. Cd) and this suggested a detoxification role. This classical, dual role of metallothionein has come under fire recently. The following issues describe the more complicated situation today.

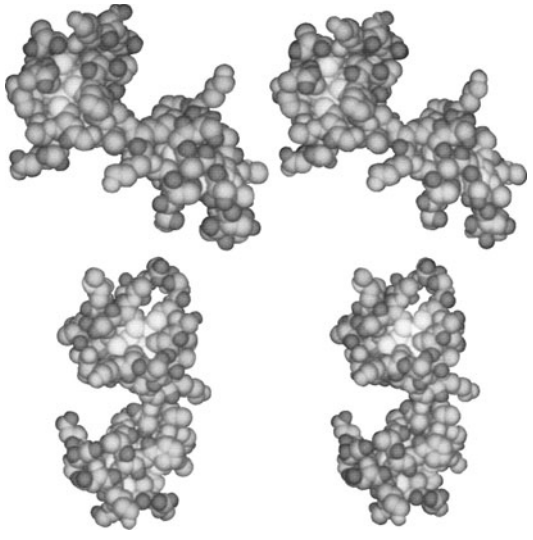


Figure 5.6 Model, based on crystallography data, of rat liver metallothionein with five cadmium and two zinc atoms, bound in two metal-thiolate clusters. The molecule can be viewed in three dimensions by stereoscopy. From Robbins *et al*. (1991), with permission from Elsevier.

First, it turned out that metallothionein is induced not only by metal ions, but also by a variety of other stresses, including changes in redox state, oxidative stress, and stress hormone signals. This suggests that the protein might be a member of the integrated stress response and has other roles as well. A function as a scavenger of free radicals is often suggested. Second, metallothionein should not be considered a single protein. Many organisms have more than one *Mt* gene; the human genome has no less than 16 of them. Most invertebrates investigated so far have two genes, one strongly inducible by cadmium and encoding a cadmium-binding protein, the other not inducible and encoding a copper-binding protein. The presence of a specific zinc-binding metallothionein in invertebrates is doubtful; the copper- and cadmium-binding metallothioneins of *Drosophila*, nematodes, earthworms, and snails are not inducible by zinc. Third, other metal-chelating substances have been found; this happened initially in plants, hence the name *phytochelatins*. They are the main zinc-binding ligands in plant cytoplasm and could play a similar role in invertebrates. Phytochelatins are peptides of variable length with the general formula $(\gamma\text{-Glu-Cys})_n$-Gly, where *n* varies from 2 to 5. The peptides are synthesized from glutathione by the enzyme *phytochelatin synthase*. The gene encoding this enzyme is not only found in plants but also in nematodes, earthworms, and chironomids (Cobbett and Goldsbrough 2002). Phytochelatin synthase is also present in the genome of the tunicate *Ci. intestinalis*, as mentioned in Section 2.3.6, but it is absent from vertebrates.

Induction of metallothionein by exposure to heavy metals has been studied extensively, but the mechanism is not yet clear. As in the case of antioxidant enzymes, inducibility is due to the presence of specific sequences in the promoter of the gene, which bind a transcription factor activated by metals. In the case of metallothionein these sequences are called *metal-responsive elements* (Table 5.1). Such elements are not only present in the promoters of *Mt* genes, but also in those of genes encoding membrane-bound zinc transporters and enzymes associated with glutathione synthesis (Andrews 2001). The best-characterized transcription factor binding

to these sequences is *metal-responsive element-binding transcription factor* (MTF). Induction by metals takes place by activation of MTF in the cytoplasm, followed by translocation to the nucleus; however, how MTF is activated by metals is not clear. One model suggests that in uninduced circumstances MTF is inhibited by a factor called metallothionein transcription inhibitor (MTI). This MTI has possible binding sites for zinc, which if occupied would result in the release of MTF (Palmiter 1994; Roesijadi 1996; Haq *et al*. 2003). Under this model, activation of MTF by cadmium is explained by an increase in the free zinc concentration in the cell, brought about by cadmium displacing zinc from cellular binding sites (Fig. 5.7).

The validity of the model for MTF activation by free zinc may be limited to mammals and cannot be extrapolated without modification to fish or invertebrates. In rainbow trout it was shown that silver could activate metallothionein expression without mediation by zinc (Mayer *et al*. 2003). In *Drosophila*, zinc itself does not activate MTF, although cadmium does. In earthworms and nematodes induction must involve an entirely different mechanism because metal-responsive elements seem to be completely absent; yet induction of the *Mt* gene by cadmium is possible (Stürzenbaum *et al*. 2001). Another, non-MTF/metal-responsive element-like induction mechanism is found in yeast. Here metallothionein (*CUP1*) is induced primarily by copper and the system is more direct than the one described above. A transcription factor has been isolated which binds to upstream activation sequences only when it contains copper; this, probably in combination with other transcription factors, results in enhanced transcription (Thiele 1992).

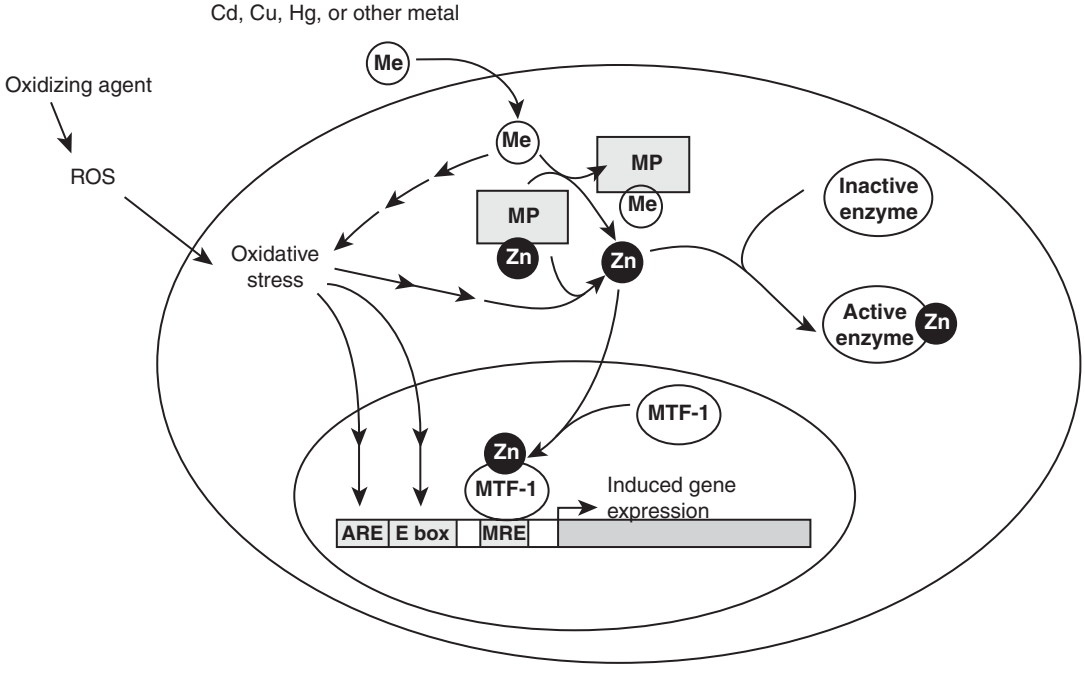


Figure 5.7 Model for induction of metallothionein in mammals. Heavy metals such as cadmium, copper, and mercury can displace zinc from metal-containing ligands (MP), thereby increasing the free zinc concentration in the cell. Excess free zinc then activates metal-responsive element-binding transcription factor-1, MTF-1, which binds to metal-responsive elements (MREs) in the promoter of the metallothionein gene. Additional transcriptional activation is recruited from oxidative stress signals (activating AREs), and adenomajor late transcription factor (MLTF), binding to the E box. After Haq *et al*. (2003), with permission from Elsevier.

There is a great deal of tissue specificity in metallothionein induction. Dallinger *et al*. (1997) isolated two metallothioneins from the snail *Helix pomatia*. One of these binds only copper and is present mainly in the mantle while the other binds only cadmium and is present in the midgut gland. In zebrafish, *Mt2* is induced by methylmercury but only in the liver—not in muscle or brain—despite the fact that methylmercury accumulates mainly in the brain (Gonzalez *et al*. 2005). The metallothioneins of mammals are also tissue-specific. Among the four groups of iso-enzymes (MTI–IV), MTI and MTII are inducible and expressed in nearly every cell; these proteins seem to qualify best as members of the stress-response. MTIII and MTIV are expressed constitutively and only in specific tissues: MTIII is expressed only in the brain and MTIV is found only in squamous epithelium cells.

Metallothionein has many links with the stress-activated systems discussed above, especially with SAPK signalling and the antioxidant stress response. One of the interactions is due to the fact that after induction by metals MTF must still be activated by kinase activity (Zhang *et al*. 2001). Yu *et al*. (1997) showed that in mammalian cell lines metallothionein induction by metals can be suppressed by protein kinase C inhibitors, suggesting that MTF must be phosphorylated by stress signalling. In addition, sequencing of metallothionein promoters has revealed that they contain not only metal-responsive elements, but also anti-oxidant responsive elements. This explains why metallothionein is also induced by oxidative stress. A model of metallothionein induction in mammals, involving both MTF and antioxidant signalling, is represented in Fig. 5.7.

Finally, we note that the metal-scavenging function of metallothionein should be supplemented by other metal-handling systems if it is to make any sense in the metabolism of the cell. Exactly how excess metal is removed after being bound by metallothionein is not completely known. One of the mechanisms suggested is that metallothionein donates excess metals to vesicles of the lysosomal system. Especially in invertebrates, tissues with an intestinal and hepatic function are full of these vesicles, which in electron microscope images become visible as electron-dense granules (Hopkin 1989). These granules may be removed from the cell by exocytosis or apoptosis of the cell. This scheme is supported by Liao *et al*. (2002), who identified a cadmium-responsive gene (*Cdr-1*) in the genome of *C. elegans*, encoding a putative metal pump associated with the lysosomal membrane. A remarkable feature of this gene was that it was cadmium-specific and not induced by oxidative stress.

5.2.5 The mixed-function oxygenase system

Aromatic endogenous compounds such as steroid hormones, signalling molecules, vitamins, feeding deterrents, and anti-herbivore toxins are all metabolized in the cell by enzyme-mediated reactions covered under the general term *biotransformation*. Also, compounds coming from the environment, such as plant toxins, environmental pollutants, and drugs are subject to biotransformation. One of the most important enzyme families participating in these reactions is known by the name of the *cytochrome P450s*. The genes encoding these enzymes, which are designated with the prefix *Cyp*, can be found in all genomes, including those of most prokaryotes. About 200 different families of *Cyp* genes have been described; they represent probably the most diverse superfamily of enzyme systems known. Their evolutionary success story is due to frequent gene duplication, as well as gene conversion, genome duplication, gene loss, and lateral transfer (Werck-Reichhart and Feyereisen 2000; Werck-Reichhart *et al*. 2002). Comparative genomic analysis has shown that cytochrome P450 diversification has occurred independently in different lines (Ranson *et al*. 2002). Many species have several tens of *Cyp* genes (*D. melanogaster* has 84 and *C. elegans* has 74), but their number has exploded into the hundreds in plants (*A. thaliana* has 249 and *O. sativa* has 323). The extreme diversity in plants is thought to result from the increased need for versatile defence mechanisms in sessile organisms, which cannot avoid stress factors by moving away. A complicated nomenclature is used for distinguishing the various cytochrome P450 families and higher groupings,

designated as clans (http://drnelson.uthsc.edu/cytochromeP450.html).

In eukaryotes cytochrome P450 is anchored in the membrane of the smooth endoplasmic reticulum, which can be isolated by differential centrifugation as a fraction containing microsomal vesicles; that is why P450 activity is also called *microsomal monooxygenase*. Other designations are aryl hydrocarbon hydroxylase, aromatase, and mixed-function oxidase. The latter term obviously relates to the huge diversity of substrates that can be attacked by cytochrome P450 enzymes. Most of the biotransformation reactions are oxidations introducing a hydroxyl group on to an aromatic ring, according to the following overall reaction:

$$RH + O_2 + NADPH + H^+ \rightarrow ROH + H_2O + NADP^+$$

where R is any substrate. This scheme shows that molecular oxygen is split into two oxygen atoms, one of which is introduced into the substrate and the other of which reacts with two hydrogen atoms to form water. The enzyme's reactive centre, which is responsible for the binding of oxygen, contains an iron atom in a porphyrin ring structure. The introduction of a hydroxyl group on to an aromatic ring makes the substrate more polar and therefore more water-soluble, which is important if it has to be excreted.

Cytochrome P450 usually conducts its biotransformation reactions in conjunction with conjugation enzymes such as glucuronyl transferase, sulphotransferase, and glutathione S-transferase. These enzymes transfer a polar endogenous compound (glucuronic acid, sulphate, or glutathione, respectively) to the product of the oxidation catalysed by cytochrome P450. Because of the sequence of events, the initial oxidation by cytochrome P450 is called *phase I biotransformation*, and the conjugations thereafter are called *phase II biotransformations*. Interestingly, the phase II enzymes seem to have diverged to nearly the same degree as cytochrome P450; for example the *Drosophila* genome has 27 annotated genes encoding a glutathione S-transferase.

The importance of biotransformation reactions is well recognized in ecological biochemistry (see Harborne 1997). In plants no less than 15–25% of protein-encoding genes may be involved with biotransformation reactions of secondary metabolism. The evolution of novel secondary compounds presents a fascinating illustration of how metabolic diversity may be generated by gene duplication, repeated evolution, and convergence (Pichersky and Gang 2000; Wittstock and Gershenzon 2002). In animals, biotransformation is often deployed to detoxify plant toxins and excrete them in water-soluble form. Some specialist herbivores sequester plant-derived toxins in their bodies to support their own defence against predators; others metabolize them to reproductive pheromones (Nishida 2002). It is often assumed that the great diversity of plant secondary compounds to which herbivores are exposed was an important selective force in the evolution of biotransformation mechanisms. In pharmacology and toxicology biotransformation is studied intensively because it determines the half-life of drugs and environmental pollutants in the body.

In the context of the stress response we must pay special attention to biotransformation reactions directed towards plant toxins and xenobiotics. Such reactions are usually considered as detoxifications, which increase the water solubility of the compound and assist its elimination. However, depending on the chemical, phase I reactions may give rise to intermediate compounds that are more reactive than the parent compound. Such compounds may react with macromolecules (DNA, proteins) before phase II metabolism can detoxify them. This process is called *bioactivation*. A well-known example is the activation of certain polycyclic aromatic hydrocarbons (a group of chemicals occurring in crude oil, diesel exhaust, and tar) to very reactive intermediates that are highly mutagenic and carcinogenic. The production of such very reactive compounds with obvious negative metabolic effects can be seen as an unavoidable evolutionary trade-off against the capacity to metabolize toxins and xenobiotics. In addition, with some substrates cytochrome P450 produces large amounts of ROS as a by-product. These two negative side effects of biotransformation may explain why upregulated monooxygenase is often accompanied by upregulation of antioxidant enzymes and heat-shock proteins. Figure 5.8 provides an overview of the different possibilities

for the fate of a foreign compound that undergoes biotransformation.

Many biotransformation enzymes are greatly inducible; that is, their activity can increase by several orders of magnitude when they are exposed to certain chemical compounds. However, not all isozymes of cytochrome P450 are inducible to the same degree and not all chemical compounds induce the same set of enzymes. Traditionally, toxicologists have made a discrimination between two types of inducer: *phenobarbital-type inducers* (PB-type inducers) and *3-methylcholanthrene-type inducers* (3MC-type inducers). The two compounds, phenobarbital and 3-methylcholanthrene, are used as model substrates. The distinction is important because only the 3MC-type inducers act according to a mechanism of receptor-mediated transcriptional regulation. The common property of 3MC-type inducers is a planar molecular structure, such as is present in certain dioxins, certain polychlorinated biphenyls, and polycyclic aromatic hydrocarbons such as benzo(*a*)pyrene and 3-methylcholanthrene itself. 3MC-type more than PB-type induction is associated with cytotoxic effects. The most potent inducer in this class, and at the same time the most toxic anthropogenic chemical, is 2,3,7,8-tetrachlorodibenzo(para)dioxin (TCDD), a compound which arises as a byproduct of the manufacture of chlorinated pesticides and other organochlorines.

Induction of cytochrome P450, especially the genes known as *Cyp1a1* and *Cyp1a2*, is initiated by binding of the inducer with a cytosolic receptor, the *aryl hydrocarbon receptor* (Ah receptor). In the uninduced state this receptor is stabilized by heat-shock protein Hsp83 (Hsp90 in mammals). This role of Hsp83 is very similar to its stabilization of the heat-shock factor HSF and the steroid hormone receptor (see Fig. 5.4 and Table 5.2). If a 3MC-type inducer

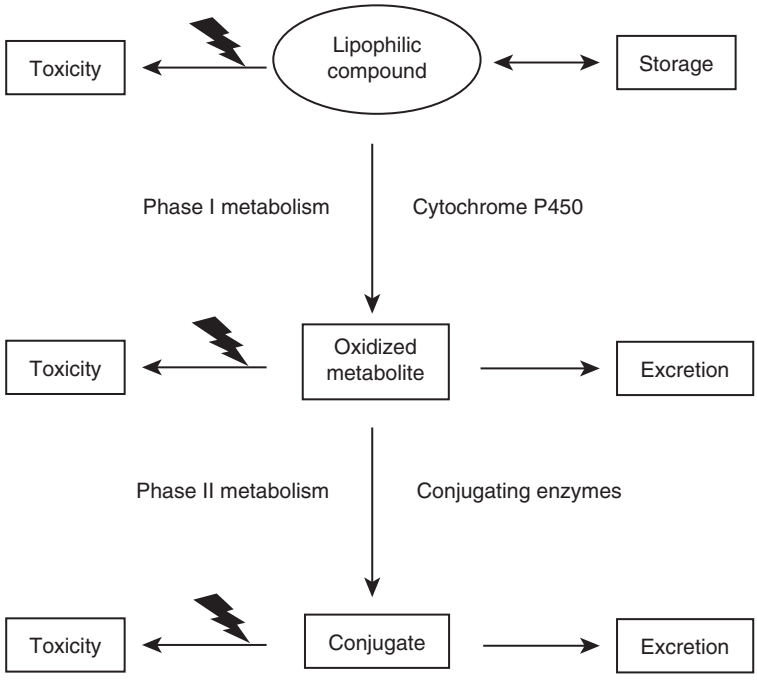


Figure 5.8 Overview of the fate of lipophilic compounds subjected to two phases of biotransformation. Phase I metabolism is conducted by cytochrome P450; phase II metabolism consists of conjugation enzymes such as sulphotransferase (transferring a sulphate moiety to the activated product of phase I metabolism), uridinediphosphate glucuronyl transferase (transferring glucuronic acid), and glutathione S-transferase (transferring glutathione). In several cases the product of phase I metabolism is very reactive and more toxic than the original compound. This also happens sometimes with phase II reaction products.

binds to the Ah receptor, the protein is activated and can translocate to the nucleus, where it is phosphorylated by protein kinase C and forms a transcriptional activator complex with another protein, *Ah receptor nuclear translocator* (also known as *ARNT*). The complex then binds to sequences known as *xenobiotic-responsive elements*. The elements are also called *dioxin-responsive elements* because of the use of dioxin as a model compound; the term Ah receptor elements is also used. Such sequences are present in the promoters of both phase I and phase II genes. Genes activated by the Ah receptor are jointly referred to as the *Ah battery* (Nebert *et al*. 2000). The group involves at least two P450 genes (*Cyp1a1* and *Cyp1a2*) and four genes involved with phase II biotransformation and the antioxidant stress response. Interestingly, the promoters of phase II biotransformation enzyme genes in the Ah battery contain not only xenobiotic-responsive elements but also AREs.

There is a close link between xenobiotic biotransformation and oxidative stress (Nebert *et al*. 2000; Kong *et al*. 2001; Fig. 5.9). Some xenobiotics such as dioxins and polychlorinated biphenyls are very potent inducers of *Cyp1a1*, but are themselves hardly metabolized by the cytochrome P450 enzymes. Instead, upregulated enzyme activity generates a lot of ROS and induces prolonged oxidative stress. In addition, some metabolites generated by P450 activity are very *electrophilic*, which means that they react easily with other compounds to compensate for their shortage of electrons. Electrophiles and

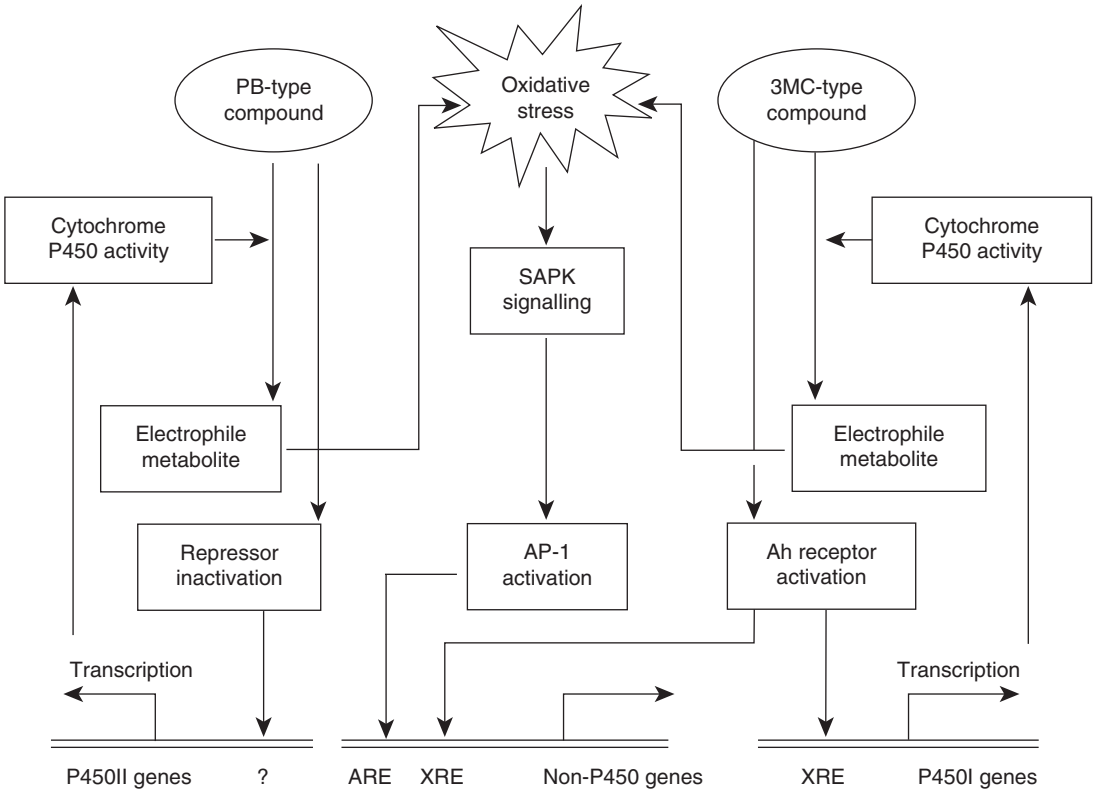


Figure 5.9 Summary of the various pathways leading to induction of biotransformation enzymes. Three inducing agents can trigger activity, PB-type inducers, oxidative stress, and 3MC-type inducers. The first type of inducer presumably activates transcription of cytochrome P450 II genes by binding to a repressor. The latter type of substance induces cytochrome P450 I genes via activation of the Ah receptor, which binds to a xenobiotic-responsive element (XRE) in the promoters of P450 I genes. In addition, P450 activity generates electrophilic metabolites and ROS and this may, through SAPK signalling, induce antioxidant genes and non-P450 genes from the Ah battery.

oxygen radicals induce antioxidant enzymes by the mechanisms discussed above. The presence of AREs in the promoters of phase II biotransformation enzymes ensures that these genes are also induced. The chronic toxicity of compounds such as dioxin is ascribed to a situation of sustained oxidative stress in the whole organism.

Compounds of the PB-type predominantly induce cytochrome P450s of the IIB group and to a certain extent also members of the III family, which are normally induced by steroid hormones. PB-type induction is not as specific as 3MC-induction and it does not depend on the Ah receptor. The precise mechanism is not known. One possibility is that PB-type inducers activate cytochrome P450 by binding to a cytosolic repressor, causing derepression of *Cyp* genes. However, the great structural variety of PB-type compounds makes it unlikely that this mechanism holds for all inducers. Another possibility is that BP-type inducers introduce a change in redox state, upon which SAPK signalling is triggered, leading indirectly to transcriptional upregulation of *Cyp* genes. A comprehensive summary of the various pathways that may lead to induction of biotransformation activities, based on Nebert *et al*. (2000) and Korsloot *et al*. (2004), is given in Fig. 5.9.

5.3 Heat, cold, drought, salt, and hypoxia

Most of the genome-wide stress analyses concern model organisms exposed to stress factors in the laboratory. The aim of such studies is usually to reveal biochemical regulatory mechanisms in the cell or to identify genes that act as targets of signalling cascades. The most extensively studied organism in this respect is baker's yeast, *S. cerevisiae*. There is also a lot of work on stress responses in mammalian cell lines conducted in the context of tumour cell biology. Hardly ever are cellular stress studies aimed at explaining the performance of organisms in the wild. It must be noted, however, that genomic studies in ecologically relevant systems are only just beginning. It is to be expected that essential insights obtained from the work on model organisms can be translated to ecologically relevant contexts, since, as we have seen above, several aspects of the stress response are conserved over large parts of the tree of life.

In this section we address a number of physical factors of the environment that elicit stress responses when they attain extreme values. Among these factors, temperature stands out as a major determinant of the niche, because the great majority of species—prokaryotes, ectothermic animals, and plants—cannot regulate the temperature of their internal environment. In these organisms the rate of all metabolic processes is ultimately determined by the ambient temperature. The importance of temperature is easily demonstrated by the fact that numerous species exhibit a distribution range that is bounded by some aspect of temperature, for example a winter frost isotherm. Other major conditions that determine niche boundaries are humidity, salinity, oxygen tension, and redox potential. Responses to drought have received a lot of attention in plant studies, because water deficit is often the most severe limiting factor for crop productivity. In this section we will explore what kind of stress responses are triggered by abiotic factors at the niche edge.

5.3.1 Responses to abiotic stress factors in yeast

S. cerevisiae has been called the vanguard of a truly integrative biology, because with its limited number of protein-encoding genes (around 6000), the availability of mutants deleted for each of these genes, and the early development of tools such as microarrays, it seemed possible to capture all interactions, including the transcriptome, proteome, and metabolome, into an integrative approach of the living cell. Stress-response studies are an important part of yeast integrative biology, because by removing the cell from its normal operating range the various compensatory and regulatory mechanisms are forced to reveal themselves.

Two pioneering studies of whole-genome responses to stress in yeast were published at around the same time: Gasch *et al*. (2000) and Causton *et al*. (2001). These authors studied the transcriptome of yeast under a variety of stress factors: temperature shocks, chemicals generating ROS, osmotic shock, and nutrient depletion. The

profiles were studied in time and for some agents as a function of the dose. Interestingly, no two expression programmes were precisely the same in terms of the genes affected, the magnitude of expression alteration, and the changes in time. Gasch *et al*. (2000) introduced the term *choreography of expression* to describe the sequence of events occurring after a specific stimulus. Each stress factor seemed to trigger its own choreography; the uniqueness of each programme highlights the precision by which yeast responds to environmental change.

Despite the specific choreographies, the studies also showed that a large fraction of the yeast genome responded to stress in a stereotypical manner. Gasch *et al*. (2000) identified two clusters, one upregulated and one downregulated, of around 900 genes in total—more than 14% of the yeast genome—which demonstrated similar responses across the various stress factors. These genes together were designated as the *environmental stress response*. In a similar manner, Causton *et al*. (2001) identified 499 genes, corresponding to around 10% of the yeast genome, which were common to most of the transcriptional changes observed when yeast cells were exposed to a number of different environmental changes. They called these genes the *common environmental response* (CER). Table 5.4 lists the functional categories to which the genes of the environmental stress response belong. The concept of CER is similar to the CSR concept proposed by Kültz (2005) (see Section 5.2), however, the genes identified by Gasch *et al.* (2000) and Causton *et al.* (2001) are specific to yeast and only a few of them overlap with the more universal set of Kültz's CSR genes.

The genes of the environmental stress response fit into a syndrome of changed priority from protein synthesis to protective mechanisms. Almost all repressed genes have something to do with translation at the ribosomes, whereas many upregulated genes relate to stress-defence mechanisms, such as scavenging of ROS, anti-oxidant defence, and repair of aberrant proteins. A major fraction of the upregulated genes is due to heat-shock proteins and other stress proteins discussed in Section 5.2. Another group of upregulated genes with a less obvious stress-defence function involve enzymes of sugar metabolism, especially in the pathways of trehalose and glycogen. The genome-wide regulation of carbohydrate metabolism under stress seems to be specific to yeast and may reflect the pervasive importance of carbohydrates in the natural environment of yeast. Interestingly, both synthetic and catabolic enzymes of carbohydrate metabolism were upregulated under stress. Gasch *et al*. (2000) suggest that these apparently conflicting functions may reflect the need for the cell to increase its capacity for regulated flux of carbohydrates so as to rapidly buffer energy reserves and manage osmotic instability.

The coherent induction and repression of genes belonging to the environmental stress response would suggest that they are all regulated by a single master process. Earlier research had suggested a key role for the yeast transcription factors Msn2p and Msn4p. These important transcription factors are associated with changes of nutritional state and diauxic shift; they regulate many genes related to carbohydrate metabolism, but are also involved in the stress response. Translocation of Msn2p to the nucleus under the influence of SAPK and other signalling pathways is a characteristic feature of the yeast stress response (Görner *et al.* 1998). Msn2p and Msn4p exert transcriptional control of stress-responsive genes by binding to a specific *stress-response element* in the promoters of these genes. However, the genomic work by Gasch *et al*. (2000) showed that not all stress responses were dependent on these transcription factors. For example, genes of the thioredoxin cluster were induced in *msn2;msn4* mutants to the same degree as in the wild type, suggesting that there must be alternative regulators of environmental-stress-response gene expression. In fact, it seems to be more a rule than an exception that genes are regulated by more than one transcription factor, depending on specific environmental conditions. A possible additional group of stress-responsive gene regulators are the *yeast activator protein* (Yap) factors. There are eight *Yap* genes in the yeast genome and for five of them a function in some aspect of the stress response has been established (Rodrigues-Pusada *et al*. 2004).

Numerous other studies have been published on genome-wide responses to stress in yeast (see Gasch and Werner-Washburne 2002 and Hohmann and Mager 2003); however, since the focus of these

Table 5.4 List of functional categories containing genes regulated in the environmental stress response of yeast

Functional categories of genes repressed	Functional categories of genes induced
Growth-related processes	Carbohydrate metabolism
RNA processing, RNA splicing	Cellular redox reactions and antioxidant defence
Translation initiation and elongation	Protein folding
Nucleotide synthesis	Protein degradation and vacuolar functions
Secretion	DNA-damage repair
Ribosomal proteins	Intracellular signalling

Source: From Gasch *et al.* (2000).

studies is mainly biochemical, discussing them would lead us too far away from the ecological focus of this chapter. From the sample studies discussed above we may conclude that yeast cells respond to environmental stress factors by means of a number of essentially independent pathways that are integrated in an overall genomic expression programme. When cells are exposed to two or more stress factors simultaneously the resulting expression programme largely approximates the sum of each individual stress response. Coherence is brought about by plenty of cross-talk between the various pathways and the presence of different transcription factor-binding sites in the promoters of stress-responsive genes. It is also obvious from the yeast studies that stress responses elicited by sudden changes of abiotic conditions are essentially transient. The remodelling of the transcriptome reflects an adaptation phase in which the cell adjusts its metabolic machinery to the new conditions. Transcript levels turn back to normal even when the stress factor persists. In accordance with this model, the time over which the genome shows altered transcription is correlated with the seriousness of the disturbance. Finally, the discovery of a common set of genes (the environmental stress response) involved in defence against a variety of physical and chemical stressors is an important lesson that is equally applicable to other organisms.

5.3.2 Plant responses to drought, cold, and salt

Plants have a remarkable ability to cope with environmental stress factors, including extremes of temperature, humidity, and salinity. The genomic responses to such stresses are of potential importance to agriculture, because a better understanding of abiotic stress tolerance may improve the basis for breeding of crop plants. In some parts of the world cultivation of stress-resistant crops under marginal conditions is the only way to increase food production. Whereas people have selected plant species for 10 000 years to grow under a variety of climatic conditions, breeding for stress tolerance has proved difficult because the traits involved are determined by multiple genes. Using a genome-wide approach it might be possible to identify factors upstream in a stress-signalling cascade and so increase the likelihood that determinants of genome-wide stress tolerance can be identified, manipulated, and possibly introduced into crop plants. It may be assumed that the fundamental aspects of stress tolerance are present in all plants, but what distinguishes species, it seems, is how fast and how persistently the stress-tolerance machinery is engaged (Bohnert *et al*. 2001).

Extremes of temperature, humidity, and salinity are most often studied in plant stress-response studies, but a shortage of nutrients (nitrate, sulphate, etc.) is an equally important factor that may limit the ecological niche of a species. Responses to nutritional stress are, however, more specific than responses to physical stress and the two hardly interact with one another. Therefore we leave nutritional stress out of consideration here. A recent thorough overview of the various molecular aspects of plant responses to drought and salt stress is given by Bartels and Sunkar (2005).

Genomic studies of *Arabidopsis* have shown that there is a great deal of commonality in the responses to drought, cold, and salt. Seki *et al*. (2001, 2002) developed a cDNA microarray of *Arabidopsis* genes and monitored the expression of 7000 genes when plants were desiccated, exposed to 4 ºC, or grown in hydroponic solution with 250 mM NaCl. In total, 277 genes were upregulated more than fivefold by drought, 53 were cold-inducible, and 194 were induced by high salinity (Fig. 5.10). Among these genes, 22 responded to all three stress factors. A large number of genes overlapped between drought and salinity, and fewer between either of these factors and cold. The fact that the greatest number of genes is induced by drought suggests that tolerance to drought requires the largest transcriptional alteration in plant cells and that water deficit may be considered the most severe limiting factor of plant growth. However, as Seki *et al*. (2002) admit, this result may also be due to the intensity of the stress factors applied; dose-dependence of gene expression was not investigated in this study.

The various *Arabidopsis* genes induced by drought, cold, and high salinity may be classified into two functional groups. The first group includes proteins that play a direct role in combating stress. These genes include heat-shock proteins, osmoprotectants, water-channel proteins, sugar transporters, and potassium transporters. Each of these proteins is targeted to solve a specific aspect of the stress condition. For example, KIN proteins induced by cold have a unique ability to prevent freezing of fluids by neutralizing ice nucleators. Aquaporins (members of a larger family of major intrinsic proteins, MIPs) regulate the flux of water across the membrane. A group of proteins called late embryogenesis-abundant (LEA) proteins have a role similar to heat-shock proteins and protect macromolecules from denaturation. Table 5.5 provides an overview of the various upregulated genes involved directly in stress tolerance. The second group of drought-, cold-, and high-salinity regulated genes contains mainly regulatory proteins. These are transcription factors, protein kinases, protein phosphatases, and genes associated with plant hormones and signalling molecules. No fewer than 40 genes, which is 11% of all stress-regulated genes in the study of Seki *et al*. (2002), encoded transcription factors. In a similar study Chen *et al*. (2002) found 57 transcription factors in the *Arabidopsis* genome to be regulated by one or more stress factors (cold, salt, wounding, pathogens). Both studies illustrate the importance of transcriptional control in the tolerance to stress.

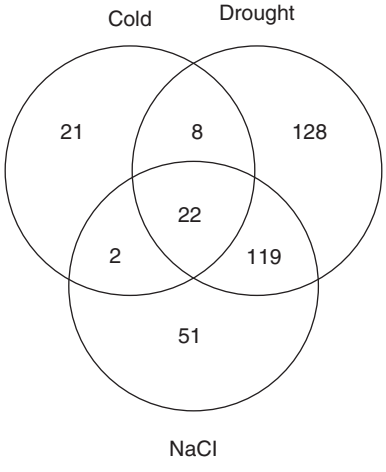


Figure 5.10 Venn diagram of gene expression in *A. thaliana* Columbia exposed to three stress factors: cold, drought, and high salinity (NaCl). In each intersection the number of genes is specified whose expression ratio showed more than fivefold upregulation compared to unstressed plants. From Seki *et al*. (2002), by permission of Blackwell Science.

One specific group of stress-inducible transcription factors in plants is the *DREB* family, which belongs to the larger group of AP-2/ERF-type transcription factors. DREB proteins bind to the so-called *dehydration-responsive element* (DRE), a 9-bp conserved DNA sequence which is found in promoters of several stress-responsive plant genes (see Table 5.1). The element is assumed to exert an important *cis*-acting influence on stress-responsive gene expression. DREB1s are involved in cold-responsive gene expression, whereas DREB2s are associated mainly with drought. Altering the expression of such master genes can be a simple lever for increased stress tolerance. For example, overexpression of DREB1A in transgenic *Arabidopsis* results in enhanced tolerance to drought, cold, and salt (Kasuga *et al*. 1999).

Another control mechanism regulating the stress response in plants goes via the plant hormone *abscisic acid* (ABA). ABA is an important stress-responsive plant hormone that triggers a signalling pathway ultimately converging on a *cis*-acting DNA sequence known as *abscisic-acid-responsive element* (ABRE). Transcription factors binding to ABREs belong to the large group of *basic leucine zipper* (bZIP) proteins. The seemingly awkward name derives from a DNA-binding domain rich in basic residues adjacent to a leucine zipper domain which supports dimerization of the protein (required for DNA binding). Among the 22 genes that were upregulated by all three stress factors in the study of Seki *et al*. (2002), 16 contained a DRE in their promoter and 15 contained an ABRE, illustrating the importance of both abscisic acid-dependent and -independent gene regulation. The presence of different transcription factor-binding sites in promoters of the same gene may explain the partial overlap between transcriptional profiles of different stress factors.

In Section 5.2 it was mentioned that the expression of several stress-related genes (e.g. metallothionein) shows a high degree of tissue-specificity. This is also applicable to the transcription profiles of plants exposed to abiotic stress factors. In particular, roots and leaves may show diverging transcriptional profiles. Kreps *et al*. (2002) monitored expression of 8100 *Arabidopsis* genes using an oligonucleotide gene chip and identified 2409 genes with a greater than twofold change over controls when plants were exposed to salt, osmotic, and cold stress. However, expression of many genes was specific for either roots or leaves, especially in the response to cold. Less than 14% of the cold-specific changes were shared between roots and leaves. Some transcripts had different tissue-specific temporal dynamics; for example, a gene was expressed initially in both roots and leaves but the transcript in

Table 5.5 List of gene categories associated with genomic responses to drought, cold, or high salinity in *A. thaliana*

Group of genes	Functional significance
Proteins involved directly with stress tolerance	
Late embryogenesis-abundant (LEA) proteins and heat-shock proteins	Protect macromolecules from denaturation
Cold-inducible (KIN) proteins	Inhibit ice-crystal growth and neutralize ice nucleators
Osmoprotectant biosynthesis-related proteins	Production of sugar and proline as osmolytes protecting cells from dehydration
Carbohydrate metabolism-related proteins and sugar transporters	Transport of sugars through plasma membrane and tonoplast to adjust osmotic pressure
Water-channel proteins (aquaporins)	Regulate water flux over plasma membrane and tonoplast in relation to osmotic homeostasis
Potassium transporters	Control potassium and sodium uptake in relation to salinity tolerance
Detoxification enzymes	Protection against ROS and electrophiles
Proteases, protease inhibitors, and senescence-related genes	Increase protein turnover and availability of amino acids; accelerate leaf scenescence
Ferritin	Protects cells from iron-catalysed oxidative damage through the Fenton reaction
Lipid-transfer proteins	Repair of stress-induced membrane damage, alteration of lipid composition of membranes
Regulatory proteins	
Transcription factors	40 different genes encoding DNA-binding proteins regulating stress-inducible genes
Protein kinases and protein phosphatases	Transducing stress signals and regulating stress-inducible genes
Plant hormone-related genes	Biosynthesis, regulation, and action of ethylene, jasmonic acid, and auxin

Notes: Two main categories are shown; one group that is involved in direct combating of stress, another that is involved in signal transduction and transcriptional regulation.
Source: After Seki *et al*. (2002).

the root disappeared quickly and only in the leaves was it observed as a sustained and consistent change. The importance of time-specific responses was also underlined in a study by Kawasaki *et al*. (2001) on salt stress in rice: genes for general stress defence were induced within 15 min, and most genes reached a peak after 1 h, but some only subsided after 7 days.

In the real life of plants, abiotic stress factors often occur simultaneously. This is especially valid for drought and heat in semi-arid or desert environments. The question is, can the common environmental stress response provide protection against two such factors at the same time? Interestingly, this does not seem to be the case, at least not for heat and drought in plants. Rizhsky *et al*. (2004) showed that to combat both heat and drought, plants deploy a partial combination of two multigene defence pathways, plus an additional 454 genes that are expressed specifically during a combination of drought and heat. As shown in Fig. 5.11, there is actually very little similarity between the responses of *Arabidopsis* to drought and heat. Only 29 genes were found to overlap. The largest overlap was between the responses to heat and a combination of heat and drought. This suggests that large portions of the defence programme against heat are also turned out in defence against drought, but in addition new genes are activated to deal with the combination. The combined response is characterized by enhanced respiration, suppressed photosynthesis, and accumulation of sucrose and other sugars. It was particularly striking that the amino acid proline, which is a common osmoprotectant accumulating under drought stress, was not accumulated when plants were exposed to both drought and heat. This suggests that the combination of drought and heat imposes a different kind of stress to plant cells compared to drought alone. Perhaps proline is avoided and sucrose favoured because heat ameliorates the toxicity of drought-induced proline.

The study of Rizhsky *et al*. (2004) suggests that there is an element of 'collision' in the defence pathways to different stress factors. The presence of antagonism in gene-expression programmes was also suggested by Tamaoki *et al*. (2003). These authors investigated the role of three plant

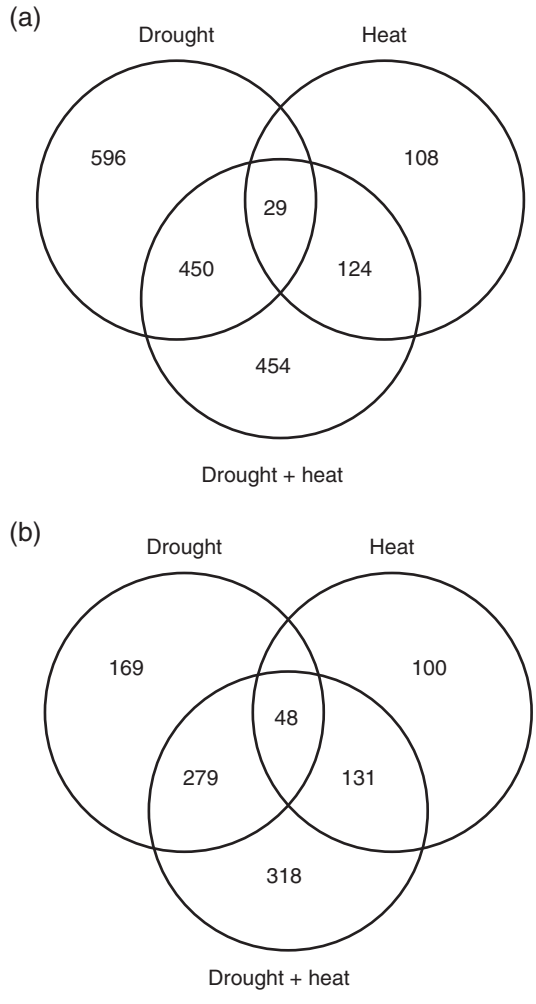


Figure 5.11 Venn diagrams showing genes regulated by drought (decrease over 6 days of plant water content to 70–75%), heat (38 °C for 6 h), and a combination of the two treatments in *A. thaliana* Columbia analysed using an oligonucleotide gene chip. (a) Upregulated genes; (b) downregulated genes. After Rizhsky *et al*. (2004). Copyright American Society of Plant Biologists.

hormones—ethylene, jasmonic acid, and salicylic acid—in the regulation of gene expression of *Arabidopsis* exposed to ozone. By studying the stress responses of mutants disturbed in each of the hormone signalling pathways, it became obvious that ethylene regulated the ozone response of 73 genes, jasmonic acid regulated 62 genes, and salicylic acid regulated 24 genes; however, there was a considerable overlap between these genes. Many defence

genes induced by ethylene and jasmonic acid signals were suppressed by salicylic acid signalling, suggesting that the salicylic acid pathway acts as an antagonist to the other two pathways. Such interactions in gene-expression programmes of plants are in contrast to the additive nature of stress responses assumed for yeast (Gasch and Werner-Washburne 2002; see above).

Although most of the mechanistic knowledge on stress tolerance comes on account of *A. thaliana*, plant biologists are also very interested, for obvious reasons, in the stress responses of crop species such as rice and barley (Kawasaki *et al*. 2001; Ozturk *et al*. 2002). In addition, the study of species growing naturally under extreme conditions might add insights that cannot be attained from mesophilic plants. Models for the study of halotolerance are the ice plant, *Mesembryanthemum crystallinum*, and the green alga *Dunaliella salina* (Cushman and Bohnert 2000; Bohnert *et al*. 2001). The resurrection plant, *Craterostigma plantagineum*, which shows a remarkable ability to restrict cell damage during desiccation and rehydration of its tissues, is a promising model for xerotolerance. Gene-discovery programmes in such naturally tolerant models have focused on EST sequencing of stressed and unstressed libraries. Such studies have demonstrated that ESTs related to stress are underrepresented in the current genomic databases, which suggests that there may still be unknown mechanisms involved in plant stress tolerance. Comparative genomics of the type discussed in the context of aging in Section 4.2 has an important role to play here. Comparisons among stress-tolerant species from different evolutionary lineages may help to identify the universal gene complement underlying stress tolerance in plants.

5.3.3 Abiotic stress responses in animals

In addition to yeast, fruit flies have been widely used as a model for studying genome-wide responses to elevated temperature. As expected, supraoptimal temperature induces heat-shock proteins, but it also causes a great variety of other changes in the transcriptome. As an example, consider the study of Leemans *et al*. (2000). These authors found that 74 genes were affected significantly in *D. melanogaster* embryos exposed to a mild heat shock (36 ºC for 25 min). Among these 74 genes, 36 had increased and 38 had decreased expression levels (Fig. 5.12). A very strong induction was seen for the small heat-shock proteins *Hsp22*, *Hsp26*, *Hsp27*, and *Hsp23*; this induction was many times greater than the fold regulation found for other genes. No induction was seen for the heat-shock cognate proteins *Hsc70-1*, *Hsc70-4*, and *Hsc70-5*; however, a small degree of upregulation was noted for *Hsc70-3*. Two signal transduction genes were also upregulated, *Shark* and *Rabgap1. Shark* encodes a protein kinase involved in the JNK cascade and *Rabgap1* is a Ras GTPase activator; both proteins are part of signalling transduction pathways regulating cell growth (see Fig. 5.3). Finally, many changes were seen in genes encoding proteins involved with transcriptional regulation and metabolism, and these changes included both up- and downregulation (Fig. 5.12).

The study of Leemans *et al*. (2000) illustrates that the transcriptional change induced by heat shock may be more complicated than suggested by the biochemical work discussed in Section 5.2. Although the strong upregulation of heat-shock proteins is in accordance with the earlier theory, the many changes seen in other functional categories suggest that there is no simple downregulation of overall protein synthesis, but a more complicated adjustment to the metabolic needs of the cell. Some of the responses documented by Leemans *et al*. (2000) may be specific to the life stage (embryos); for example, the abundance of small heat-shock proteins, which are known to be developmentally expressed. Also we must remember that the heat-shock response involves post-transcriptional regulatory mechanisms, so not all transcriptional changes depicted in Fig. 5.12 need to be expressed at the protein level.

Although the heat-shock response has become the most widely used model for studies on stress-induced transcriptional change, other stress responses may sometimes be more closely linked to an environmental context. An example illustrating this point is the response to *hypoxia* (low oxygen levels) in fish. Dissolved oxygen is a limiting factor for many fast-swimming, active fish species. Fish

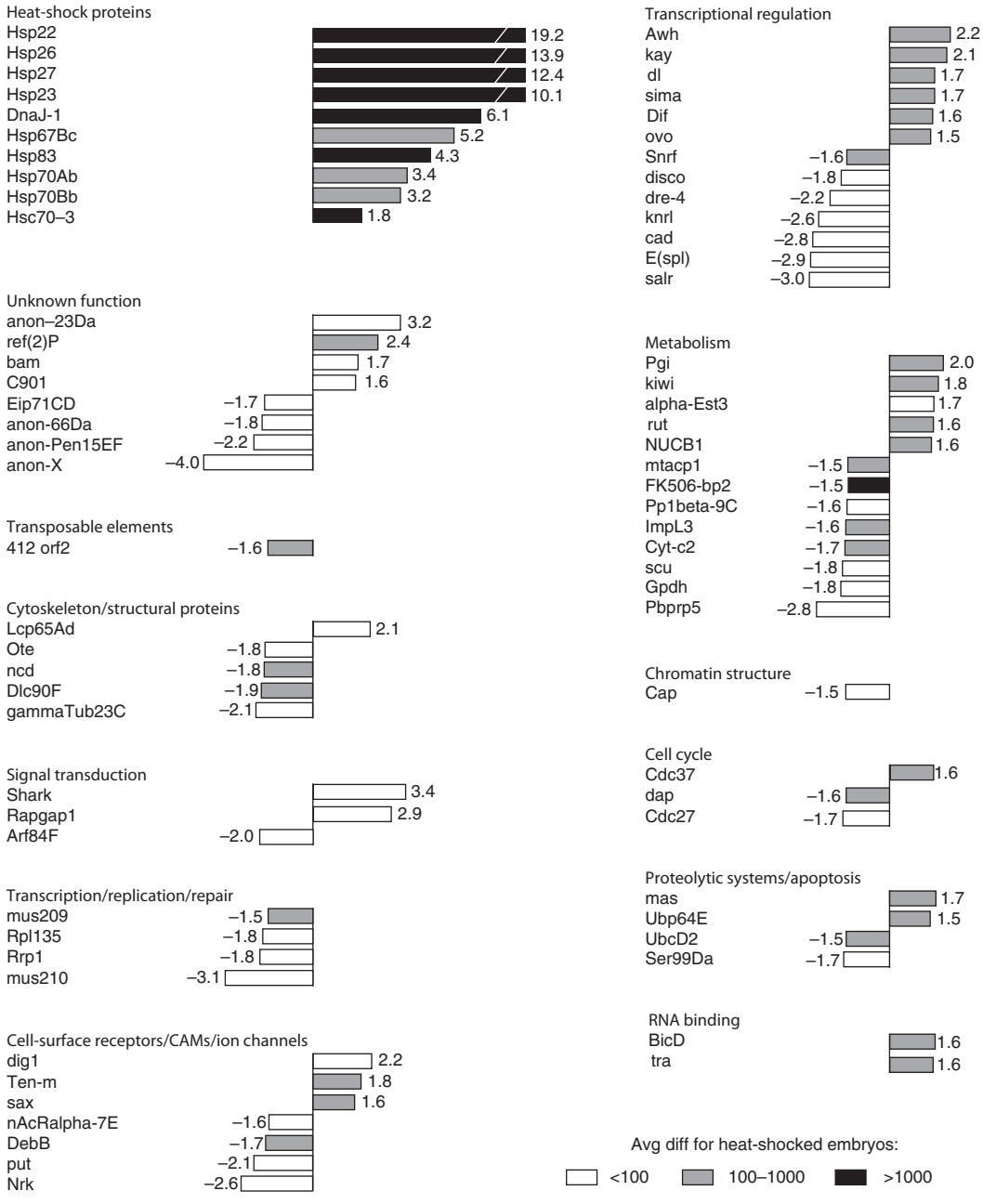


Figure 5.12 Overview of differentially expressed genes in *D. melanogaster* embryos exposed to heat shock, in comparison with non-heat-shocked embryos. The genes are grouped according to functional class; bars represent fold regulation. CAM, cell-adhesion molecules. In the key, the darkness of the bar indicates the extent to which expression in heat-shocked embryos differs (on average) from the standard condition. Avg Diff is the ratio of the two expressions. After Leemans *et al*. (2000) by permission of the National Academy of Sciences of the United States of America.

species inhabiting sediment burrows in estuarine ecosystems are particularly tolerant of hypoxic conditions and the study of such naturally tolerant species is expected to shed light on cellular responses to hypoxia in general.

From mammalian research it is known that gene expression induced by hypoxia is controlled by *hypoxia-inducible factor 1* (HIF-1) in a manner which is very comparable to the action of Nrf2 in the oxidative stress response (see Section 5.2). HIF-1 comes in two subunits, HIF-1α and HIF-1β, which are both expressed constitutively at high levels (Wenger 2002; Schulte 2004). HIF-1β is identical to the Ah receptor nuclear translocator (ARNT), a protein that we met in Section 5.2 as a dimerization partner of the Ah receptor. Bound to HIF-1α, it promotes translocation of HIF-1 to the nucleus. However, under normal physiological conditions HIF-1α is degraded rapidly through ubiquitin-mediated cytoplasmic proteolysis. Degradation is initiated by enzymatic hydroxylation of two proline residues, a process that is sensitive to the oxygen concentration in the cell. Under low oxygen conditions prolyl hydroxylase is inhibited, with the consequence that HIF-1α is no longer degraded and the heterodimer can translocate to the nucleus. A second hypoxia-dependent regulatory mechanism is located in the nucleus. To act as a transcription factor, HIF-1α must bind to nuclear factor p300, which is prevented by another enzymatic hydroxylation, in this case of an asparagine. Under hypoxic conditions this hydroxylation is also impeded, allowing HIF-mediated transcription to occur. Interestingly, HIF-1 is also required for heat acclimation and contributes even to metal resistance (Katschinski and Glueck 2003; Treinin *et al*. 2003). Such cross-tolerance phenomena present another illustration of the interconnectedness between the various stress-response pathways.

In a pioneering genomics study Gracey *et al*. (2001) documented transcription profiles of the long-jawed mudsucker, *Gillichthys mirabilis* (Perciformes, Gobiidae), exposed to hypoxic conditions (0.8 mg/l at 15 ºC). SSH was used to generate cDNA libraries enriched with hypoxia-regulated genes. More than 5000 PCR-amplified cDNA clones were printed on an array, which was then used to monitor transcriptional change in liver and muscle tissue of fish under hypoxic conditions. Clones that were differentially expressed by two-and-a half-fold or greater were sequenced and their putative function was established by homology to database sequences. A total of 126 distinct hypoxia-regulated cDNAs were found, of which 75 could be identified by homology.

Gracey *et al*. (2001) were able to interpret many of the changes in gene expression in terms of an ecophysiological strategy employed by the fish to allow its survival under hypoxic conditions (Table 5.6). The transcription profile suggested a metabolic switch in which, very rapidly after the onset of hypoxia, the major energy-requiring processes, like protein synthesis and locomotion, were repressed. Then, after about 24 h, the metabolic machinery was directed towards anaerobic ATP production and synthesis of glucose from non-carbohydrate sources (*gluconeogenesis*). The changes in genes for amino acid metabolism indicate that amino acids are the main source for gluconeogenesis. At the same time, cell growth and proliferation is repressed by means of binding circulating IGFs and by attenuation of MAPK signalling.

Among the genes with unclear function was an inducible pseudogene encoding an antisense mRNA matching the 5´ end of retinoblastoma-binding protein 2 (RBP2). An *antisense RNA* is a sequence complementary to a certain mRNA, the translation of which is suppressed by binding to the 5´ end. The production of antisense RNA in response to stress is a means to antagonize the expression of other genes, which may contribute to the fine-regulation of metabolic processes. That this type of mechanism was involved in the response to hypoxia was not known before.

The long-jawed mudsucker study is remarkable because it illustrates several of the points raised in Chapter 1 as dilemmas of ecological genomics: (i) it is possible to explore the transcriptional profile of an organism about which no sequence data existed before, (ii) cDNA microarrays can be developed from SSH libraries without prior knowledge of the genome, (iii) microarrays are useful as exploratory instruments, (iv) even with a limited sequencing effort, a great deal of insight into the ecophysiology

Table 5.6 Summary of transcriptional change observed in long-jawed mudsucker, *G. mirabilis*, exposed to hypoxic conditions

Functional category	Tissue	Examples of genes	Possible functional significance
Energy metabolism	Liver	Lactate dehydrogenase (+), enolase (+), trisephosphate isomerase (+)	Maintenance of glucose homeostasis by gluconeogenesis
Locomotion and contraction	Muscle	α-Tropomyosin (–), myosin heavy chain (–), myosin regulatory light chain 2A (–)	Decreased locomotory activity
Translation, protein synthesis	Muscle	Elongation factor 2 (–), several ribosomal proteins (–)	Reduced protein synthesis
Iron metabolism	Liver	Haem oxygenase-1 (+), ferritin (+), transferrin (+)	Increased production of erythrocytes
Cell growth and proliferation	Liver	IGF binding protein 1 (+), MAPK phosphatase (+)	Suppression of cell growth by attenuating MAPK signalling
Amino acid metabolism	Liver	S-Adenosylmethionine synthase (+), tyrosine aminotransferase (+)	Catabolism of amino acids used for gluconeogenesis

Notes: The direction of change is indicated with + or –.

Source: After Gracey *et al*. (2001).

of a species can be reconstructed, and (v) the study of species adapted to extreme conditions can reveal new insights into homeostatic mechanisms that may be relevant for many other species (Gracey 2007).

5.4 Herbivory and microbial infection

Not only abiotic conditions, but also biotic factors may limit the ecological niche and can elicit specific stress responses in plants and animals. Such biotic factors can be other organisms that decrease the fitness of plants or animals by consumptive action. Ecologists distinguish herbivores (animals consuming plants or parts of them), predators (animals consuming other animals after catching and killing them), parasites (organisms living inside or on the surface of plants or animals, diverting resources from the host to themselves), and parasitoids (animals living inside other animals but killing the host to complete the life cycle). The initial contact between the two players in such interactions is invariably accompanied by stress and specific defence responses, especially in the victim being preyed upon, being consumed, or acting as a host. In this section we will review studies dealing with stress responses associated with attack by herbivores and pathogenic microorganisms. The importance of both processes in regulating the abundance of species in the wild stands beyond doubt; however, herbivory has received much more attention from ecologists than parasitism and disease. Genomic responses to microbial infection of plants is a major topic in plant pathology, but this is mostly studied in an agricultural context and therefore not discussed here. Predation, although a popular subject among ecologists, has not been studied at the genomic level and so is also not considered.

5.4.1 Plant defence against insect herbivory

Plants respond to herbivore attack with a wide array of defence mechanisms. One of the strategies employed is to synthesize *secondary compounds* such as alkaloids and terpenoids that are toxic to herbivores and pathogens. Such compounds are called secondary because they do not belong to the metabolism of primary cell constituents (carbohydrates, proteins, and lipids). Secondary compounds can be synthesized after initial damage by the herbivore (*damage-induced defence*), or they may be synthesized at all times (*constitutive defence*). The first strategy has the advantage that the costs associated with biosynthesis only burden the plant when actually under attack; however, the disadvantage is that the defence may not be rapid enough, or the initial damage may be too severe (Wittstock and Gershenzon 2002). The second strategy is more effective, but obviously costs of biosynthesis are

incurred for as long as the plant grows. The costs of anti-herbivore defences can indeed be considerable, especially under competitive growing conditions, and so inducibility is assumed to have evolved as a cost-saving mechanism (Zavala *et al*. 2004). We will focus on damage defence induced by insects in this section.

Induced defence against herbivores can be direct or indirect. Direct responses aim at preventing further feeding of the herbivore by some kind of toxic action such as sensory irritation, gut convulsion, or paralysis. Indirect responses involve the use of volatile alarm chemicals to attract predators and parasitoids of the herbivore. There is a great deal of specificity in these responses, involving chemical communication between plants, herbivores, and natural enemies, which has led to the concept of *tritrophic interaction* (Price *et al*. 1980). A better understanding of the genomics of herbivore defence by plants may thus have important ramifications outside the plant proper and benefit community ecology (Dicke *et al*. 2004).

Over the last decade, significant progress has been made in identifying the nature of volatile chemicals that are produced by plants when attacked by insects (Kessler and Baldwin 2002; Dicke *et al*. 2003). Three groups of chemicals can be distinguished: (i) so-called *green leaf volatiles*, C_6 alcohols and aldehydes which are synthesized quickly after damage and are not very specific to the plant species or the type of leaf damage, (ii) terpenoids from the octadecanoid pathway, commonly called *oxylipins*, which are emitted slowly, typically 24 h after the damage and are more specific, and (iii) derivatives of the aromatic amino acid precursor shikimate, such as *methyl salicylate*, which are emitted after herbivore damage but not after mechanical wounding. Volatile chemicals are not only released locally from damaged leaves but also from undamaged parts of the plants in distinct temporal patterns (a *systemic response*). Some constituents of the volatile releases are just a passive consequence of damage to cell compartments, notably vacuoles and trichomes; however, many are due to *de novo* biosynthesis under control of three plant hormones: ethylene, jasmonic acid, and salicylic acid. The question is, how is the signal emitted from herbivore-induced damage transduced into regulation of these biosynthetic pathways?

Studies on *Arabidopsis* provided the first cues on the nature of transcriptional changes induced by herbivory in plants. Reymond *et al*. (2000) prepared an array with 150 EST probes of genes that were known to be involved in stress defence. Temporal change of gene expression was analysed in response to mechanical wounding and feeding by cabbage white caterpillars, *Pieris rapae* (Lepidoptera, Pieridae). Mechanical wounding of the leaf induced a clearly recognizable stress response in *Arabidopsis*. Upregulation was seen for many genes from the stress-responsive pathways discussed in Section 5.2, including several MAPKs, a metallothionein, two glutathione S-transferases, and a cytochrome P450. A cluster of 17 genes was regulated in a coherent fashion; this cluster included general stress genes as well as genes implicated in the synthesis of the plant hormone *jasmonic acid*, and known to be induced by this hormone (Fig. 5.13). Earlier research had shown that three key enzymes in the jasmonate biosynthetic pathway (lipoxygenase (LOX), hydroperoxidaselyase (HPL), and allene oxide synthase (AOS)) can be used as indicators for wounding and these three enyzmes were also responsive in the microarray study (Fig. 5.13). There was a very good correlation between expression of the inducible gene cluster and leaf concentrations of jasmonate, while the metabolic precursors of jasmonate, OPDA and dnOPDA, rose more slowly, peaking around 6 h after wounding. The observations suggest strongly that the transcriptional response to wounding is triggered by a burst of jasmonate within 1 h, followed by jasmonate synthesis.

The availability of *Arabidopsis* mutants insensitive to jasmonate allowed the role of this trigger to be assessed in more detail. About one half of the genes regulated by wounding were no longer induced or repressed in a jasmonate-insensitive mutant, demonstrating that for a significant number of genes the plant response to wounding depends strictly on the action of jasmonate. The other half of the genes responded independently of jasmonate and this group included many genes that were also activated by water stress. So it seems that the transcriptional profile of wounding includes signatures

triggered by jasmonate signalling, which are specific to wounding, as well as water stress signalling, which, as we have seen above, is regulated by abscisic acid.

Further experiments by Reymond *et al*. (2000) have demonstrated that the transcriptional profile of wounding is not the same as the profile induced by herbivory. Feeding by cabbage white caterpillars induced many transcripts that were also induced by mechanical wounding, but additional genes were induced that mainly responded to feeding. Interestingly, there was an under-representation of water stress-induced genes in the herbivore-induced transcriptional profile. These observations suggest that insects in some way or another can minimize the dehydration stress experienced by the plant while damaging the leaf. Behavioural mechanisms can possibly explain these observations. Many herbivorous insects remove tissues only from the edge of the leaf and make semicircular holes without cutting the midvein. In doing so they will cause less damage than most mechanical-wounding operations.

It must be noted that transcriptomic changes following wounding and herbivory are not restricted to those triggered by jasmonate and abscisic acid; they involve many other effects that indicate a major metabolic switch, as in the case of abiotic stress responses. Transcripts related to photosynthesis and ribosomal processes are found to be downregulated by wounding, whereas expression of genes associated with protein turnover, carbohydrate metabolism, cell-wall modification, and antimicrobial defence was increased (Moran *et al*. 2002; Hui *et al*. 2003; Zhu-Salzman *et al*. 2004). The genome-wide nature of these changes suggests that in addition to jasmonic acid signalling, which is specific to wounding, other signalling pathways are also activated, for example those triggered by salicylic acid, ethylene, and abscisic acid. The changes can be summarized as a metabolic switch from growth priority to stress-defence priority.

The fact that some aspects of the transcriptional changes to herbivory are specific to insect feeding and do not occur after mechanical wounding suggests that certain chemical cues emanating from the insect are recognized by the plant. Two classes of

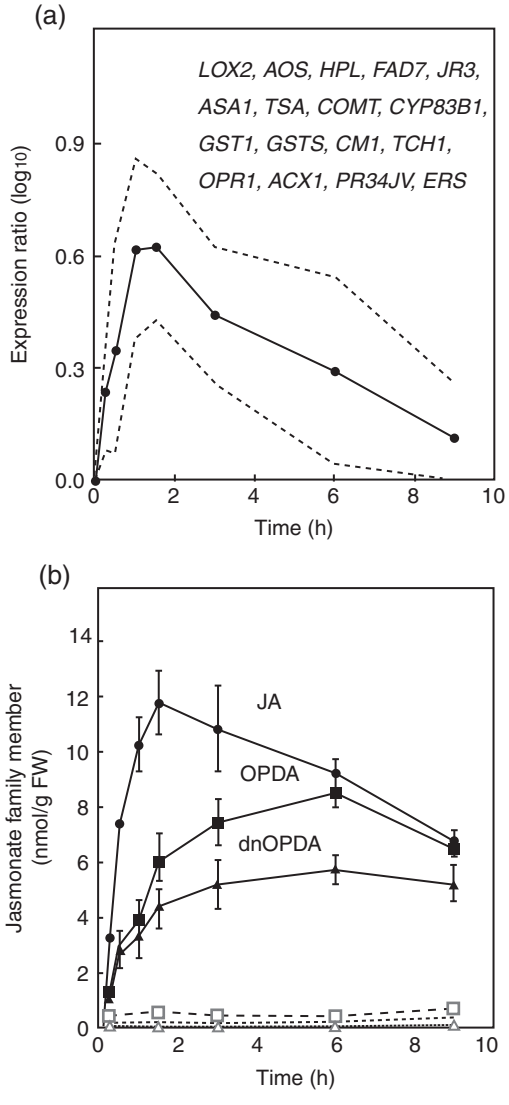


Figure 5.13 Temporal changes of gene expression in *Arabidopsis* after mechanical wounding of leaves. (a) Average expression of 17 genes with similar temporal change; dashed lines indicate the standard deviation. LOX2, lipoxygenase; AOS, allene oxide synthase; HPL, hydroperoxide lyase; FAD7, fatty acid desaturase; JR3, aminohydrolase; ASA1, anthranilate synthase α subunit; TSA, tryptophan synthase α subunit; COMT, O-methyltransferase; CYP83B1, cytochrome P450; GST1 and GST5, glutathione S-transferases; CM1, chorismate mutase; TCH1, calmodulin; OPR1, OPDA reductase; ACX1, acyl-CoA oxidase; PR3AIV, chitinase; ER5, late embryonesis-abundant (LEA)-like protein. (b) Temporal change of leaf concentrations of jasmonate (JA), 12-oxo-phytodienoic acid (OPDA), and dinor OPDA (dnOPDA). The open symbols represent control measurements without wounding. After Reymond *et al*. (2000). Copyright American Society of Plant Biologists.

these so-called *stress elicitors* have been identified (Kessler and Baldwin 2002; Korth 2003). The first class includes digestive enzymes present in the saliva of the herbivore, such as β-glucosidase, glucose oxidase, and alkaline phosphatase. The second class comprises amino acid conjugates of fatty acids, fatty-acid-amino acid conjugates (FACs), which are present in the digestive tract, frass, and regurgitate of insects (Schittko *et al*. 2001; Roda *et al*. 2004). The fatty acid moiety in these compounds derives from the plant itself; FACs probably represent products of phase II biotransformation in the insect, aimed at solubilizing and excreting lipophilic compounds (see Section 5.2). Recognition of chemical cues that are specifically associated with insect feeding allows plants to fine-tune their defence to certain herbivores and to attract natural enemies of the herbivore by releasing specific volatiles.

An interesting complication is that some herbivores are able to express detoxification enzymes of the cytochrome P450 family even before the plant has been able to produce defensive toxins. Herbivores do this by responding to the jasmonate and salicylate burst of the plant. This would provide protection in the critical window between accumulation of plant defence compounds and production of enzymes that can metabolize them. This *eavesdropping* on plant defence signals is especially advantageous in polyphagous herbivores such as the noctuid *Helicoverpa zea*, for which this mechanism was first described (Li *et al*. 2002).

The interaction between herbivore-induced damage and chemical signalling has been investigated in great detail in a system consisting of wild tobacco, *Nicotiana attenuata* (Solanaceae), and its specialist herbivore, tobacco hornworm, *M. sexta* (Lepidoptera, Sphingidae; Fig. 5.14). Wild tobacco is an interesting model because it is diploid, exhibits a large degree of phenotypic plasticity, and evolved in a habitat that can be considered primordial to the agricultural niche, the environment created by wildfires in woodlands (Baldwin 2001). *N. attenuata* uses two different defence mechanisms against herbivores; the production of nicotine and the production of terpenoid volatiles. The level of *nicotine*, a potent neurotoxin, is greatly induced by mechanical wounding under the influence of jasmonate signalling. Nicotine production seems to act mainly as a direct defence against browsing mammalian herbivores. Specialist herbivores such as *M. sexta* can store nicotine in their tissues without any toxicity and may even use it to defend themselves against avian or mammalian predators. Consequently, the plant must rely on other chemical defences, such as oxylipins, to combat such specialist herbivores. Oxylipins are also assumed to support indirect defence by attracting natural enemies of the herbivore. The ecological importance of jasmonate-induced oxylipins in the defence against insect herbivores was demonstrated elegantly by Kessler *et al*. (2004). These authors used transformed lines of *N. attenuata* which were planted outdoors in an experimental field. Mutants in which key enzymes of the jasmonate pathway (LOX, HPL, and AOS) had been silenced were not only more vulnerable to damage by sphingid caterpillars but also attracted new herbivores such as leaf hoppers and a chrysomelid beetle.

Genomic investigations into the tobacco– herbivore interaction have used a microarray which was developed from cDNAs identified by pre-genomic

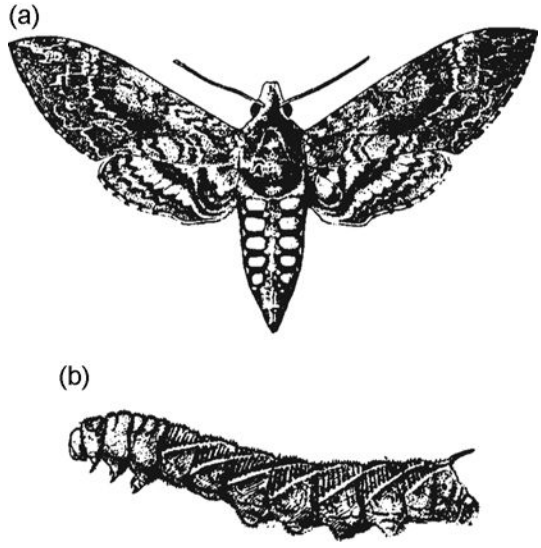


Figure 5.14 Tobacco hornworm, *M. sexta* (Lepidoptera, Sphingidae), a specialist herbivore of wild tobacco and used as a model for genomic studies of plant responses to insect herbivory. (a) Adult; (b) larva. From Gillot (1980), with permission from Springer.

differential screening techniques such as differential display, cDNA-AFLP, and subtractive hybridization (Halitschke *et al*. 2004; Hui *et al*. 2003). Despite the small number of genes on the array (initially 241), which precludes a truly genome-wide inventory of expression change, the use of a targeted approach proved to be quite successful. Development of small *boutique arrays*, encompassing a focused selection of cDNAs from genes relevant in a certain context, is a good strategy for ecological laboratories working with incompletely sequenced organisms (Held *et al*. 2004). Later the herbivore stress array was extended to 789 probes, represented by 50-mer oligonucleotides (Heidel and Baldwin 2004).

The studies on the tobacco herbivore community have revealed interesting patterns of herbivore-specific gene expressions (Heidel and Baldwin 2004; Voelckel *et al*. 2004; Voelckel and Baldwin 2004). Comparisons were made between transcription profiles induced by different species of chewing lepidopteran: *M. sexta* (Sphingidae), a specialist herbivore, and two generalists, *Heliothis virescens* and *Spodoptera exigua* (both Noctuidae). Chemical analysis had shown that the composition of stress elicitors (FACs) in the regurgitate of these species is varied. The *M. sexta* regurgitate is dominated by a FAC named *N*-linolenoyl-l-glutamate. This is absent in the other two species, which both have a compound known as volicitin, or *N*-(17-hydroxylinolenoyl)-l-glutamine. The reason for these species-specific FAC profiles is unknown; maybe they relate to different substrate specificities of biotransformation enzymes of the insects' xenobiotic metabolism. The FAC profiles were correlated with herbivore-induced gene expression in the plant; there was a large overlap between the transcriptional profiles of the two noctuids, whereas the overlap between either of the noctuids and *M. sexta* was much smaller (Voelckel and Baldwin 2004).

Other differences between herbivore-elicited transcriptional profiles were observed in a comparison of diverging herbivore-feeding guilds. Chewing herbivores such as sphingid and noctuid caterpillars consume pieces of leaf tissue completely; however, mirid bugs (Heteroptera, Miridae) puncture holes in the tissue and feed on the cell contents, while aphids insert their stylet between the cells and suck only the phloem. Comparing *M. sexta*, *Tupiocoris notatus* (Heteroptera, Miridae), and *Myzus nicotianae* (Homoptera, Aphididae) Voelckel *et al*. (2004) observed that the aphids elicited only weak responses, both in a qualitative sense (fewer genes affected) and in a quantitative sense (fold regulations were lower). An overview of the differences in terms of numbers of genes is given in Fig. 5.15. Thus the herbivore-induced transcriptomic changes

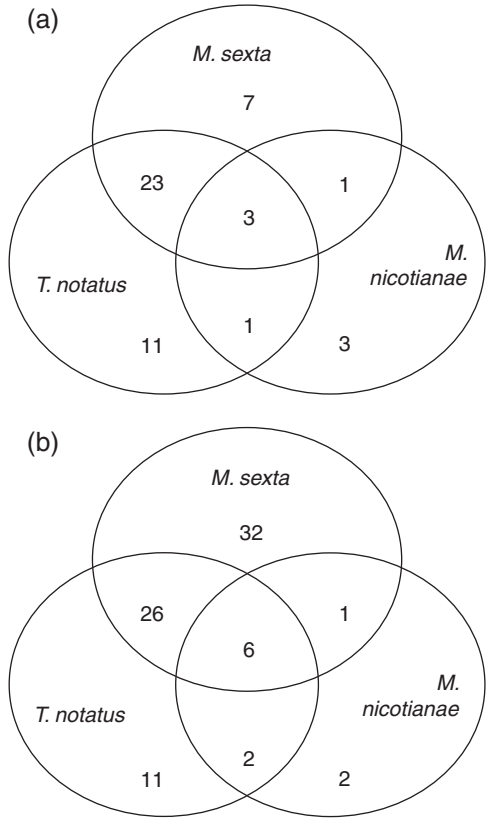


Figure 5.15 Venn diagram of the number of cDNAs showing differential expression in *N. attenuata* in response to three herbivorous insects, *M. sexta* (Lepidoptera, Sphingidae), *Tupiocoris notatus* (Heteroptera, Miridae), and *Myzus nicotianae* (Homoptera, Aphididae). The different feeding modes employed by these herbivores induce distinct but overlapping gene expression, with the largest overlap between *M. sexta* and *T. notatus*. (a) Upregulated genes; (b) downregulated genes. After Voelckel *et al*. (2004), by permission of Blackwell Science.

could be proportional to the degree of damage caused by the herbivore. It is also relevant to note that FACs or other elicitors have never been isolated from aphids (Moran *et al*. 2002). Interestingly, Voelckel *et al*. (2004) found a plant gene encoding the enzyme glutamate synthase to be induced by aphid feeding, but not by the other two herbivores. This could indicate an alteration of the plant amino acid metabolism imposed by the aphid. Amino acids are a limiting resource for aphid growth and any upregulation of their concentration in the phloem would greatly benefit the aphid. These data added to an earlier study by Moran *et al*. (2002) suggest that plant responses to aphids are fundamentally different from responses to chewing herbivores. The response to aphids has less of a wounding signature and includes facilitation of the host plant by the herbivore.

Our short account of the defence mechanisms associated with plant–herbivore interactions demonstrates that there are still important gaps in our knowledge. On the mechanistic side, the source of stress elicitors in the plant-feeding insect is not known and also the plant receptor upon which these elicitors act remains to be elucidated. On the ecological side, the tritrophic aspect of herbivore defence needs more attention, as well as the way in which the different stress responses interact with different feeding groups of herbivores. These questions indicate that genomic analysis of plant–herbivore interaction represents a highly promising research area, where genomics meets ecology, with pay-offs to both sides.

5.4.2 Genomics of the immune response in *Drosophila*

In Chapter 2 we saw that the genomes of invertebrates, including the tunicate *Ci. intestinalis*, do not encode proteins of the *adaptive immune system*, which only evolved in the vertebrate lineage, to supplement the older *innate immune system*. The adaptive immune system is to be considered a major evolutionary innovation: it is specific to particular antigens, it shows an extremely large diversity partly inherited and partly acquired during maturation of the system, and it builds up a memory of previous antigen encounters. The latter property implies that when an antigen reacts with a clone of cells with specificity for that antigen, these clones expand greatly and adapt to give the highest possible specificity for the antigen. All these properties are lacking in the innate immune system, which constitutes a general first-line defence with low specificity.

The organization of the innate immune response of invertebrates shows many similarities with the vertebrate innate response, suggesting that they have a common origin, and that defence against microorganisms was already a priority in the first metazoans (Hoffmann and Reichhart 2002). Research on the innate immune response has greatly benefitted from the use of *Drosophila* as a model and some important molecules involved, such as the antifungal compound *drosomycin*, were first isolated from *Drosophila*. Studies in *Drosophila* can reveal aspects of the human innate immune response that may otherwise be obscured by the adaptive response (Govind and Nehm 2004). In addition, the evolutionary conservation of the innate immune system implies that the principles discovered in *Drosophila* have a general validity for all animals.

The link between between immunology and ecology is only weakly developed at the moment. Without doubt pathogens are a very important aspect of ecological functioning in both plants and animals. Diseases can limit the distribution of species or prevent their establishment in newly colonized habitats. Lee and Klasing (2004) therefore called for a role for immunology in invasion biology. Disease and parasitism are also potent evolutionary driving forces. Continued adaptation to parasites was implicated in Van Valen's (1973) *Red Queen hypothesis*, which holds that evolution is very much in line with the remark made by the Red Queen, whom Alice met in Lewis Carroll's famous book *Through the Looking Glass*, saying 'Here you see, it takes all the running you can do to keep in the same place'. Still, the number of studies that bridge the two fields, ecology and immunology, is limited.

An interesting target of investigation in ecological immunology is the *major histocompatibility complex* (MHC), a large cluster of genes encoding proteins involved in the adaptive immune response

of vertebrates. MHC proteins bind to specific recognition sites of antigens and present them on the surface of leucocytes. A part of the molecule, a cup-like structure responsible for the recognition of a specific topographical structure of the antigenic molecule (the *epitope*), shows an extremely high degree of sequence polymorphism. The MHC genes are therefore interesting markers for resolving population structure and dispersal (Beebee and Rowe 2004). Beebee and Rowe (2004) also discuss the possibility that the striking polymorphism of the MHC complex is maintained partly by selective mate choice. Obviously, genetic variation of MHC genes may be the cause of differential susceptibility to disease, although not many studies have actually demonstrated this in an ecological context. A case in point is a study on Atlantic salmon, *Salmo salar* (Salmoniformes, Salmonidae), in which an association has been found between allele frequencies at the MHC IIB locus and susceptibility to bacterial infection (Langefors *et al*. 2001). One of the alleles had a particularly high frequency among fish resistant to the enteric pathogen *Aeromonas salmonicida* (Deltaproteobacteria).

The innate immune system, although less specific than the adaptive system, nevertheless shows considerable transcriptional change upon infection. This has become obvious from studies in *Drosophila*, which we discuss here to illustrate the genome-wide nature of immune responses and the signalling pathways involved (De Gregorio *et al*. 2001; Irving *et al*. 2001; Dionne and Schneider 2002; Govind and Nehm 2004). Microbial infection in *Drosophila* activates a number of processes. One cascade leads to blood coagulation at the site of infection followed by the production of melanin, which is toxic to microorganisms. This process of *melanization* is accompanied by encapsulation of the pathogen. Another reaction is a massive synthesis of *antimicrobial peptides* by the fat body; these peptides bind to specific surface molecules of bacteria and other invaders. Third, blood cells analogous to mammalian *macrophages* become active as microbe engulfers.

The responses of the innate immune system are controlled by two signalling pathways, *Toll* and *immune deficiency* (Imd; Fig. 5.16). The Toll protein is a membrane-bound system with an extracellular receptor that recognizes a cytokine called Spätzle. However, the Spätzle protein must first be cleaved to become active, and this is achieved by means of a proteolytic cascade originating from the infection. Activation of Toll triggers a cytoplasmic signalling pathway converging on two transcription factors of the nuclear factor-κB family called DIF and DORSAL, which translocate to the nucleus and promote transcriptional activation of a number of antimicrobial peptides. The Toll cascade is mainly directed towards pathogenic fungi and Gram-positive bacteria and it includes the antifungal protein drosomycin. Compared to Toll, the Imd pathway is less well described and the extracellular cascades triggering the receptor remain undefined at the moment. Imd signalling may activate caspases leading to apoptosis of the cell, but it may also act upon a protein complex called IKK signalosome, where a transcription factor called RELISH is activated. RELISH promotes transcription of genes encoding peptides directed towards Gram-negative bacteria (Fig. 5.16).

Infection of *Drosophila* with the Gram-negative bacterium *E. coli*, the Gram-positive *Micrococcus luteus*, and the entomopathogenic fungus *Beauveria bassiana* (Hyphomycetes) leads to a transcriptional response in genes from many different functional classes, including actin-associated proteins, calcium-binding proteins, cell-adhesion proteins, heat-shock proteins, and many others. The Toll and Imd genes were also induced. In total 543 genes were found to be differentially expressed (Irving *et al*. 2001), which nevertheless seems a modest number in comparison to some of the abiotic stress responses discussed above. Irving *et al*. (2001) could not find specific signatures for infection by Gram-positive bacteria, Gram-negative bacteria, or fungi; however, bacterial infection seemed to regulate a larger number of genes than fungal infection. De Gregorio *et al*. (2001) selected a set of 400 *Drosophila* genes as immune-regulated, of which 230 were induced and 170 repressed. Among these genes a large number had not previously been associated with the immune response and only 34% could be assigned a designated immunological function.

An overview of functional classification of immune-regulated genes of *Drosophila* is provided in Table 5.7. Several genes had already been identified as immune-regulated in previous studies, but many more were added to the list. A surprising aspect of the genomic inventory was the large number of trypsin-like serine proteases and the many antimicrobial peptides of unknown function. Maybe these induced proteins represent new classes of antimicrobial action. Another unexpected property of the genomic immune response was that it involved sequestration of extracellular iron. What the immune response has to do with iron metabolism is difficult to see, but maybe the role of iron as a catalyst of ROS is relevant in this respect. A change in cellular iron trafficking was also seen in the response of plants to abiotic stress (see Section 5.3). Genes related to iron metabolism, as well as genes of the immune system, seem to be part of a general stress response that is activated by a large number of factors, including environmental contaminants.

The two pioneering genomic studies on the immune response of *Drosophila* leave many questions unanswered (Dionne and Schneider 2002). For example, it is unclear to what extent the responses are due to wounding alone. The plant studies discussed above have demonstrated that damaging a

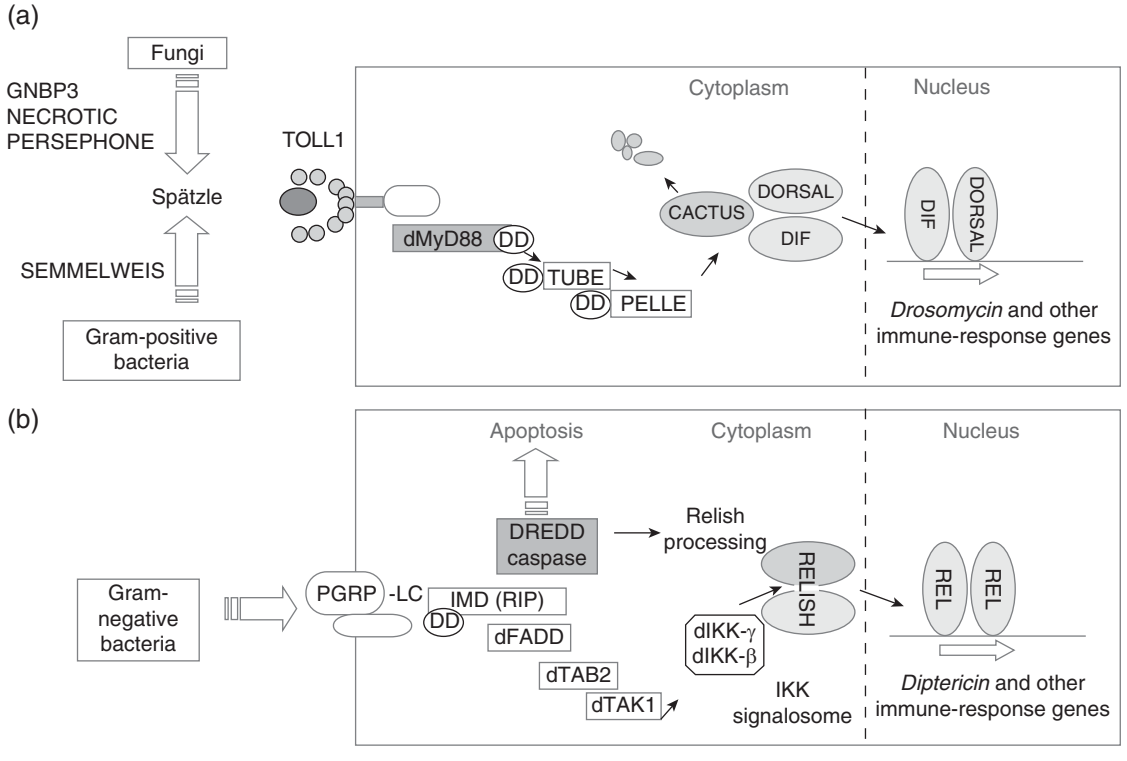


Figure 5.16 Scheme of two signal transduction pathways of the innate immune response in *Drosophila*. (a) The Toll pathway, which is characterized by a proteolytic cascade, cleavage of Spätzle, activation of transcription factors DIF and DORSAL, and production of peptides directed against fungi and Gram-positive bacteria. (b) The Imd pathway, converging on the transcription factor RELISH and the production of peptides directed against Gram-negative bacteria. DD, death domain; this is a heterodimerization domain present in several proteins involved in signal transduction, originally described for proteins involved with cell death. PGRP-LC, peptidoglycan-recognition protein LC, where LC stands for low complexity, a certain class of bacterial surface proteins. d before a protein name indicates *Drosophila*; e.g. dMyD88 is the *Drosophila* homologue of MyD88, a macrophage differentiation marker. For more information on the other protein identifiers, the reader is referred to Flybase (http://flybase.org). Reproduced from Govind and Nehm (2004).

Table 5.7 List of functional categories of genes induced by septic injury (Gram-positive and Gram-negative bacteria) and fungal infection in *D. melanogaster*

Gene categories	Functional significance
Recognition and phagocytosis	
Peptidoglycan-binding proteins	Bind to cell envelope of Gram-positive bacteria
Imaginal-disc growth factor proteins	Stimulate cell growth required for wound healing
Thiolester proteins	Forming a complement by binding to invader surface
Serine protease cascades	
Trypsin-like serine proteases	Extracellular signalling molecules of the Toll pathway
Serpins	Inhibition of trypsin-like serine proteases
Serine protease inhibitors of the Kunitz family	Possibly similar to serpins, but not previously implicated in immune response
Melanization and coagulation	
Pro-phenoloxidase activating enzymes	Proteolytical activation of phenoloxidase from its precursor, conversion of dopamine to melanin
Fibrinogen-like protein	Possible role in blood clotting
Antimicrobial peptides	
Drosomycin	Protein toxic to fungal metabolism
IM-2	Small antimicrobial protein, precise function unknown
Signalling pathways	
Cytokine-like small peptides	Activation of stress signalling pathways
Genes of the Toll pathway	Regulation of antifungal peptide production
Genes of the Imd pathway	Regulation of antibacterial peptide production
Proteins of JNK signalling pathway	General stress response (see Section 5.2)
Iron metabolism	
Transferrins and iron transporters	Sequestration of extracellular iron

tissue can already activate some 50 genes. In addition, the large number of genes with unclear function in the immune response makes interpretation difficult. Further work on mutants may help to resolve these issues. It is also expected that the knowledge on innate immune responses in *Drosophila* will be an important guide to explore the immune system of vectors of human disease, such as the malaria mosquito, *An. gambiae* (Dimopoulos *et al*. 2000; Osta *et al*. 2004). In addition, the conclusion regarding the ecological relevance of immunogenomics must come from field studies assessing genome-wide immune responses in wild animals.

5.5 Toxic substances

The study of ecological effects of toxic substances in the environment is designated as *ecotoxicology* (Walker *et al*. 2001). This multidisciplinary science is a meeting place of environmental chemists, toxicologists, and ecologists. Chemists determine the concentration of substances in the environment and study their distribution over environmental compartments and chemical ligands; toxicologists analyse uptake kinetics, biotransformation, and metabolic effects of toxicants; ecologists study the effects of toxic insults at the population, community, and ecosystem levels. Ecotoxicology traditionally has a strong link with environmental policy. Through this link, scientific support is provided for decisions about issues like standard setting, remediation of contaminated sites, and pesticide registration. In industrialized countries new substances are produced continuously, but their application in society is regulated by safety requirements concerning human and environmental health. Most legislatory systems require that new substances must be tested for their possible adverse effects on ecological receptors before they are admitted to the market.

A well-known principle in toxicology is that toxicity is not an absolute property of a substance, but that the effect of a substance depends on the dose given to the organism. The classical poison is effective in very low amounts, but in principle all substances can be toxic if dosed highly enough. This was already recognized by the Austrian alchemist and physician A.P.T.B. von Hohenheim (1493–1541), better known as Paracelsus, who wrote (Koeman 1996):

> Alle Ding sind Gifft…Allein die Dosis macht daß ein Ding kein Gifft ist (Everything is a poison…it is only the dose that makes it not a poison).

Following the Paracelsus principle, an important activity of toxicologists is the establishment of *dose–effect relationships*, which in ecotoxicology usually take the form of a graph in which some aspect of the performance of a tested organism (e.g. growth or reproduction) is plotted as a function of the concentration of a given toxicant in water, soil, or air. From such a graph two important benchmarks are estimated: the exposure concentration at which a 50% effect is observed (EC_{50}) and the highest exposure at which still no effect is seen (NEC, no effect concentration). Ecotoxicologists are usually concerned with effects that show up after chronic (long-term) exposure and, unlike human toxicologists, study end points that are important for the ecological functions of an organism. In the case of animals, many ecological functions are associated with feeding and behaviour—for example, macroinvertebrates grazing to suppress algal blooms, or earthworms burrowing to improve soil structure—that is why end points in ecotoxicology can be different from those in human toxicology.

The application of genomic technology in toxicology is called *toxicogenomics* (Lovett 2000; Pennie *et al*. 2000; Burczynsky 2003; Waters and Fostel 2004), and its ecological counterpart is *ecotoxicogenomics* (Snape *et al*. 2004). An important aim of toxicogenomics is to characterize the mode of action of toxicants on the basis of expression profiles. When two toxicants induce the same set of genes in a target organ, they are likely to have the same mode of action (Hamadeh *et al*. 2002). New substances, such as drugs, industrial chemicals, or pesticides, can be screened for their transcription profile and when the profile is compared with a database of earlier-investigated chemicals any similarities may provide an indication of the hazardous properties of the compound. The first commercial microarrays, designed to screen induction of enzymes in the human liver, were developed at the end of the twentieth century. It is likely that such tools will also be developed for environmental applications, but standardized assays accepted by regulatory authorities are not yet available.

In this section we will address the question of how genomic technology can improve our insight into ecotoxicity of environmental chemicals. Out of the huge number of chemicals that may cause environmental problems we have selected three classes of toxicant: heavy metals, pesticides, and endocrine disrupters. These three groups have very different environmental effects and serve to illustrate the principles of ecotoxicogenomics.

5.5.1 Heavy metals

Under the term heavy metal fall all elements in the Earth's crust with a density of greater than 5 g/cm^3 in their metallic form. Thus defined, a large proportion of the periodic table of elements belongs to this category; however, many heavy metals are very rare or extremely unavailable and are of no environmental concern. The toxicity of heavy metals is not due to the metal itself, but to ionic forms and other chemical species (e.g. Pb^{2+}, $HgCH_3^+$, and $Cr_2O_7^{2-}$). The active and toxic form of a metal usually constitutes only a small proportion of the total concentration in an environmental compartment, and depends on properties of the environment as well as the metal. One of the most important influences is due to environmental pH: a low pH promotes dissociation of metal complexes and may increase the fraction of metal present in ionic form without changing the total concentration. The dynamic processes occurring at the interface of environmental ligands and biotic surfaces are studied under the heading of *bioavailability*. Unfortunately, bioavailability is usually ignored in laboratory toxicity experiments, and many of the toxicity data reported in

the literature refer to artificial media in which the speciation of metals is biased towards a high fraction of free, ionic metal forms. Total concentrations in such studies cannot be extrapolated easily to the environment. This also holds for most of the toxicogenomic studies using heavy metals.

Studies in yeast were the first to reveal the genome-wide effects of exposure to heavy metals (Gross *et al*. 2000; Momose and Iwahashi 2001; Vido *et al*. 2001; Eide 2001). Many yeast genes are induced by cadmium, including obvious genes such as heat-shock proteins, but also unexpected genes, such as genes related to the synthesis of methionine. In fact, the whole sulphur-salvage pathway, including sulphate uptake, sulphate reduction, methionine synthesis, and glutathione synthesis, was upregulated (Fig. 5.17). This is all understandable since cadmium is a sulphur-seeking metal and its detoxification requires the sulphur-containing amino acid cysteine. In addition, cadmium introduces oxidative stress, which, as we have seen above, is counteracted by means of enzymes such as glutathione S-transferase, which again requires reduced sulphur.

Interestingly, Momose and Iwahashi (2001) as well as Vido *et al*. (2001) did not find induction of metallothionein when yeast cells were exposed to cadmium. Yeast has a copper-binding metallothionein (*CUP1*) that is induced by copper, but not by cadmium. Instead, cadmium in yeast follows a glutathione-dominated pathway, comparable to the phytochelatin pathway of zinc in plants and invertebrates. This is consistent with the finding that among the strongly upregulated genes was a transporter of metal-glutathione, which translocates metals to the vacuole. Metallothionein in yeast is more strongly induced by oxidative stress than by cadmium; however, at higher doses cadmium also causes oxidative stress, so the two effects are difficult to separate in practice.

When analysing the promoter sequences of genes affected by cadmium, Momose and Iwahashi (2001) noted that many of them had a sequence motif known as centromere DNA element I, CDEI. This element binds a transcription factor Cbf1, which is known to regulate the sulphur amino acid-biosynthesis pathway. The results of Momose and Iwahashi (2001) suggest that this pathway is also activated under cadmium stress, maybe through the depletion of glutathione. It is well known among toxicologists that glutathione depletion is a common consequence of oxidative stress and metal stress (see Section 5.2). All in all, the yeast study is a nice example of how the results of a transcription-profiling exercise can be understood from mechanistic knowledge of the underlying processes.

Gene expression profiles induced by metals bear a strong metal-dependent signature. Despite obvious commonalities such as signatures of oxidative stress, sulphur salvage, iron homeostasis, and immune defence, which are found in many metal-induced transcriptomes, each metal also induces a specific set of genes. Conversely, gene expression profiles can be classified using these metal-specific genes. A study illustrating this point is Nota *et al.* (2010). The soil-living invertebrate *Folsomia candida* was exposed to different metal concentrations in soil at levels corresponding to the doses causing 10% (EC10) and 50% (EC50) effect on reproduction after 4 weeks. However, gene expression was assessed after 2 days. Generating expression profiles with a 5122 probe oligonucleotide microarray, a set of 188 classifier genes was defined. Cluster analysis based on these genes showed both similarities and differences between the metals. The transciptomes of lead and zinc were most similar to each other but rather different from cadmium, while chromium clustered completely separately from the other heavy metals (Fig. 5.18). Another striking feature of this study was that the 10% and 50% effect concentration clustered together for each metal, showing that the dose effect, when assessed by short-term gene expression, is smaller than the differences between the metals.

5.5.2 Pesticides

Pesticides, also called crop-protection products, represent a wide range of chemical substances, most of them organic, often with very complex structures, designed for a specific target on the pest organism. Despite the intention of combating only pests, no pesticide is 100% selective against the pest and almost all pesticides have greater or smaller side effects on non-target organisms. Such side effects can be manifested in places far away from the site of application if the pesticide is persistent

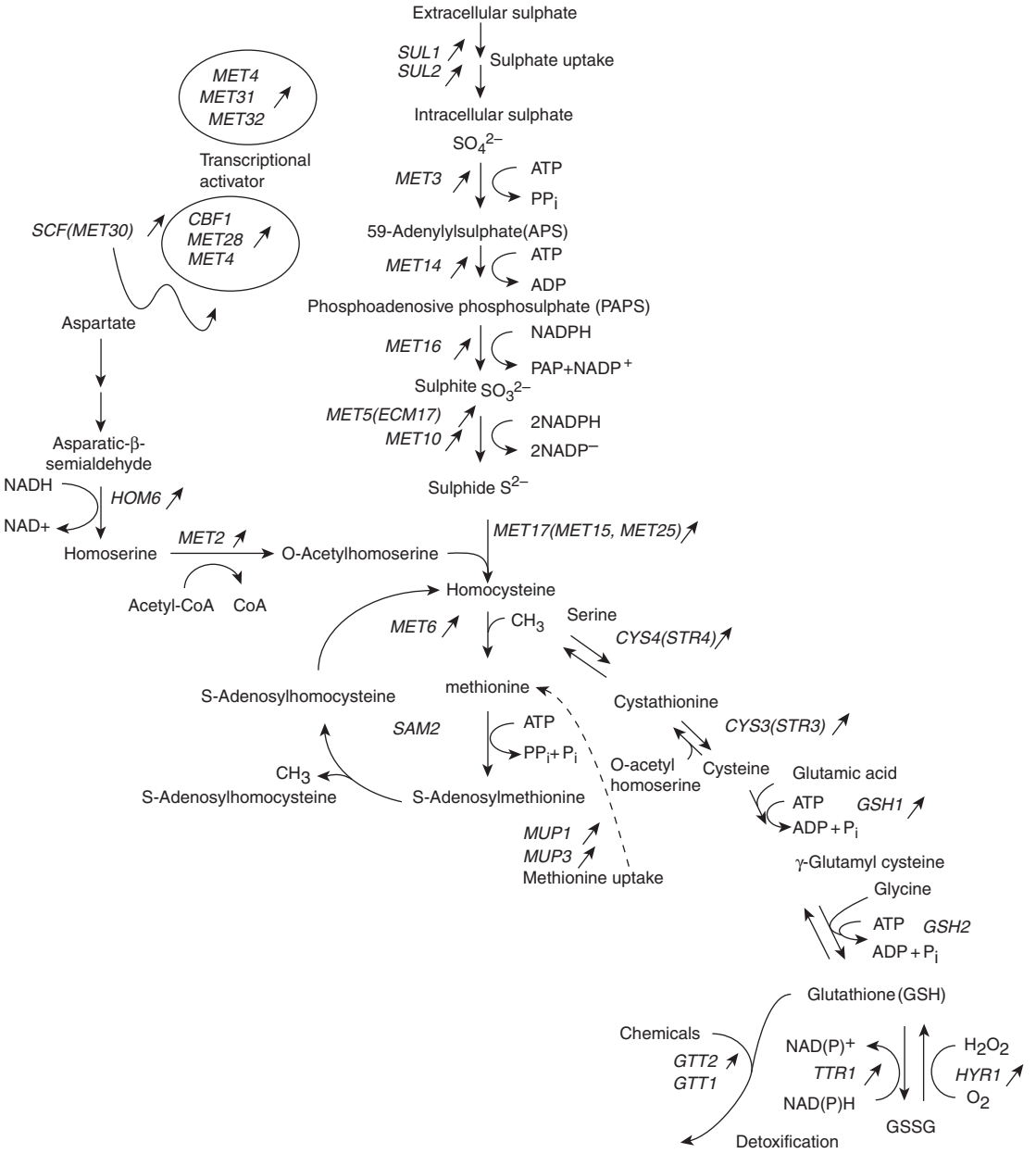


Figure 5.17 Effects of cadmium on the sulphur-salvage pathway of yeast (*S. cerevisiae*). The pathway starts with uptake of sulphate, followed by the formation of phosphoadenosine phosphosulphate (PAPS), sulphate reduction, synthesis of cysteine, methionine, and glutathione (GSSG). Genes found to be upregulated by cadmium are indicated with an oblique arrow next to the name. After Momose and Iwahashi (2001) by permission of the Society of Environmental Toxicology and Chemistry.

and transported easily by air or water. Modern pesticides are short-lived and effective in very low doses. It is expected that genomic analysis will contribute to a more sensitive assessment of possible side effects and to an ever increasing precision in the development of new products. This expectation is reflected by the interest in toxicogenomics shown by the pesticide industry.

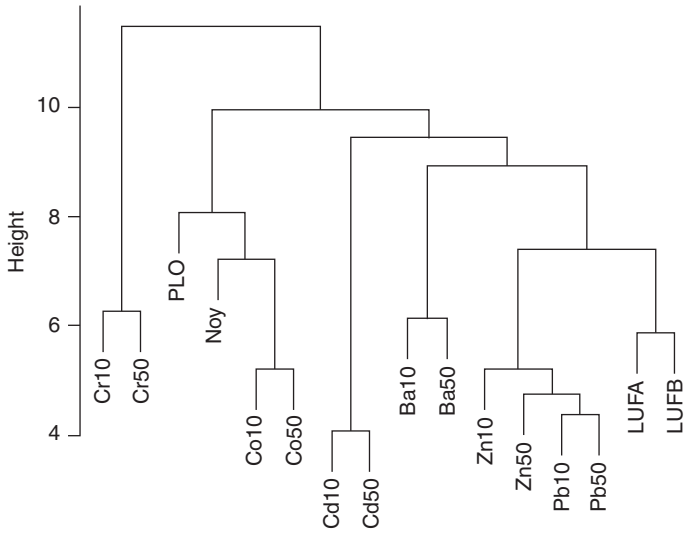


Figure 5.18 Hierarchical cluster analysis of average ^{2}log gene expression values of 188 classifier genes in the soil-living invertebrate *F. candida* (Hexapoda, Collembola), exposed to metal-contaminated soils for 2 days. Each metal was mixed into the soil at levels corresponding to the EC10 (reproduction, 4 weeks) and the EC50. Cr chromium, Co cobalt, Cd cadmium, Ba barium, Zn zinc, Pb lead. LUFA, LUFB, two reference soils. PLO, Noy, field-contaminated soils. Reproduced from Nota *et al*. (2010), by permission of Oxford University Press.

In contrast to heavy metals, pesticides are designed to affect a specific biochemical target, such as a single photosynthesis protein, a specific ion channel in the nervous system, or a key enzyme in a biosynthetic pathway. The biochemical damage caused by reacting with such a specific target is called the *primary lesion*. An example is the reaction of organophosphate insecticides with the enzyme acetylcholinesterase, which when inhibited will cause accumulation of acetylcholine, a neurotransmitter, in the synaptic cleft, leading to uncoordinated behaviour, spasms, and mortality in insects. From this fundamental toxicological principle one would expect that gene-expression profiles induced by pesticides would be limited to a small fraction of the genome. This appears not to be the case. Most pesticides, when administered to organisms outside of the context of agriculture, are less selective than expected. We discuss a few examples to illustrate this point.

Paraquat (1,1´-dimethyl-4,4´-dipyridilium dichloride) is a herbicide used to kill weeds prior to emergence of the crop and to destroy foliage of crops such as potatoes in preparation for harvesting. Under the influence of UV light paraquat enters into a redox-cycling process, producing large amounts of ROS that destroy the surface of leaves in an aspecific manner. Such redox cycles are also triggered easily inside organisms once paraquat is activated by cytochrome P450. Because of this bioactivation, paraquat is rather toxic to humans, with lung damage being the first apparent effect. Due to its ability to generate ROS, paraquat is often used as a model agent to impose oxidative stress on animals in the laboratory.

Girardot *et al*. (2004) exposed *D. melanogaster* to paraquat and two other oxidative stress agents, H_2O_2 and tunicamycin, and assessed gene expression using the Affymetrix Drosophila Genome Array. No fewer than 1111 genes (12% of the probes) were found to be up- or downregulated in flies exposed to the highest dose, illustrating the genome-wide nature of oxidative-stress defence. The genes included many representatives of the general stress response, such as cytochrome P450s, glutathione S-transferases, peptidases, and triacylglycol lipases. These expressions are all indicative of increased effort towards detoxification,

removal of damaged proteins, and repair of membrane lipids. A notable aspect was that iron-binding proteins were specifically induced by paraquat. As noted above, effects on iron metabolism are observed in many different contexts (drought stress in *Arabidopsis*, hypoxia in fish, and microbial infection in *Drosophila*), suggesting that changes in iron metabolism are a part of the general stress response.

A striking feature revealed by the paraquat study was that some of the expression was highly specific to a family of isoenzymes. This was very obvious in the cases of the cytochrome P450s and the glutathione S-transferases. As illustrated in Fig. 5.19, Cyp9b2 was hardly affected by paraquat, Cypb18a1 was repressed, and Cypb4e3 was greatly induced. Similarly, glutathione S-transferase E10 was downregulated, but glutathione S-transferase D5 was upregulated. A similar gene-specific inducibility was found among the cytochrome P450s of *C. elegans* (Menzel *et al*. 2001). Among the 10 *CYP35* genes of the nematode, four were found to be induced strongly by a few specific substrates, four were induced weakly by a wide range of substrates, and two were not inducible at all. This variation reflects the different modes of cytochrome P450 induction, as highlighted in Fig. 5.9.

Monitoring gene expression in organisms in the wild is often proposed as a strategy to evaluate exposure to pollutants in the environment (e.g. Snell *et al*. 2003). The observations of Girardot *et al*. (2004) make it very clear that such measurements need to be specific to a particular isozyme to be of any indicative value. This means that the probes on a microarray must be able to discriminate between the various isoforms within gene families. Such specificity is easier to achieve with quantitative PCR than with cDNA microarrays; in the study of Girardot *et al*. (2004) there was a very good correspondence between Q-RT-PCR measurements of expression and microarray hybridizations (Fig. 5.19).

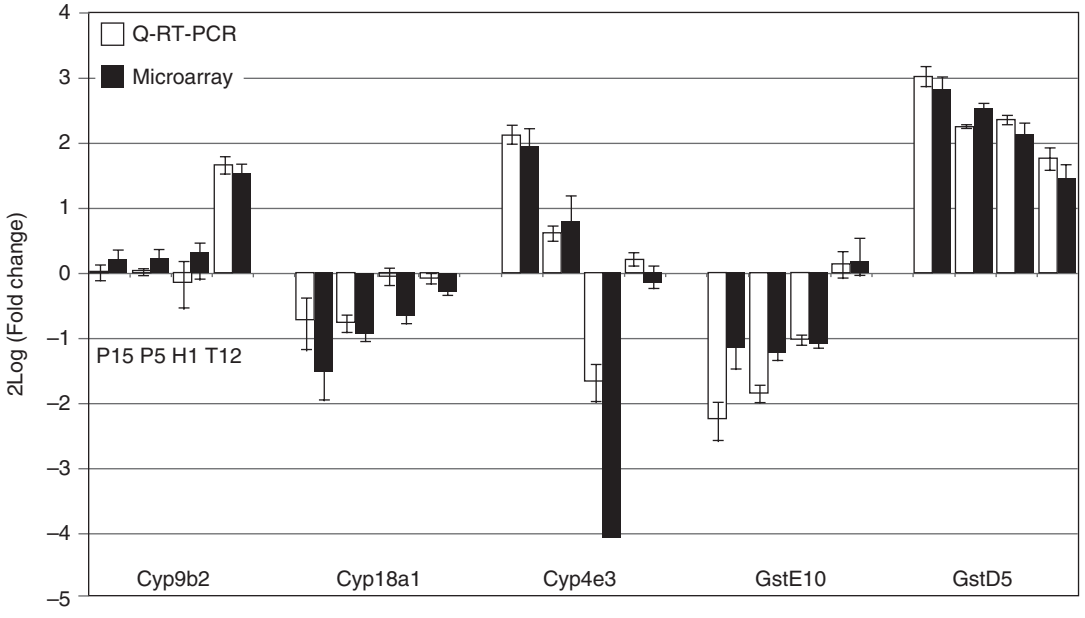


Figure 5.19 Fold regulation of five *D. melanogaster* genes observed when male adults were exposed for 24 h to 15 mM paraquat in the medium (P15), 5 mM paraquat (P5), 1% H_2O_2 (H1), or 12 µM tunicamycin (T12). Data are shown for two methods of differential expression screening: quantitative PCR applied to reverse-transcribed mRNA (Q-RT-PCR) and microarray hybridization. Cyp, cytochrome P450 (three isoenzymes are shown); Gst, glutathione S-transferase (two isoenzymes). After Girardot *et al*. (2004) with permission from BioMed Central.

Whereas the study on paraquat was a typical mechanistic laboratory analysis without direct relevance to environmental exposure, similar transcription-profiling protocols are being developed with the aim of applying them to animals exposed to pesticides or other pollutants in the wild (Sansone *et al*. 2004; Straub *et al*. 2004; Miracle and Ankley 2005; Moens *et al*. 2006; Soetaert *et al*. 2006; Poynton *et al*. 2007; Gong *et al*. 2007; Owen *et al*. 2008). Studies in the coastal marine environment have progressed furthest at the moment, because the application of physiological and molecular markers in monitoring programmes, such as 'mussel watch', was started earlier there than in other environmental compartments. We will discuss one marine study to illustrate the direction that the field is taking.

Pacific oysters, *Crassostrea gigas* (Bivalvia, Ostreidae), live in coastal marine habitats and are exposed to pesticides from surface run-off and river discharge. In particular water-soluble and persistent herbicides, such as atrazine, diuron, and isoproturon (Fig. 5.20), can reach estuarine and coastal environments from agricultural run-off and river discharge. To assess the stress response of oysters exposed to herbicides, Tanguy *et al*. (2005) developed forward and reverse SSH libraries from the gills and digestive glands of animals exposed to a cocktail of the three herbicides shown in Fig. 5.20. Some 137 sequences were retrieved as showing differential expression and the genes could be assigned to six major metabolic functions: (i) xenobiotic detoxification, (ii) nucleic acid and protein regulation, (iii) respiration, (iv) cell communication, (v) cyto-skeleton maintenance, and (vi) energy metabolism. Among the genes consistently upregulated by herbicide exposure was a glutamine synthetase. The enzyme encoded by this gene plays an important role in detoxification of ammonia, synthesis of glutamine, and clearance of glutamate as a neurotransmitter. How these functions relate to pesticide exposure is unclear; however, induction by a wide variety of compounds suggests that glutamine synthetase is part of a general stress response rather than a specific detoxification pathway.

It is also interesting to note from this study that even at relatively low and environmentally relevant

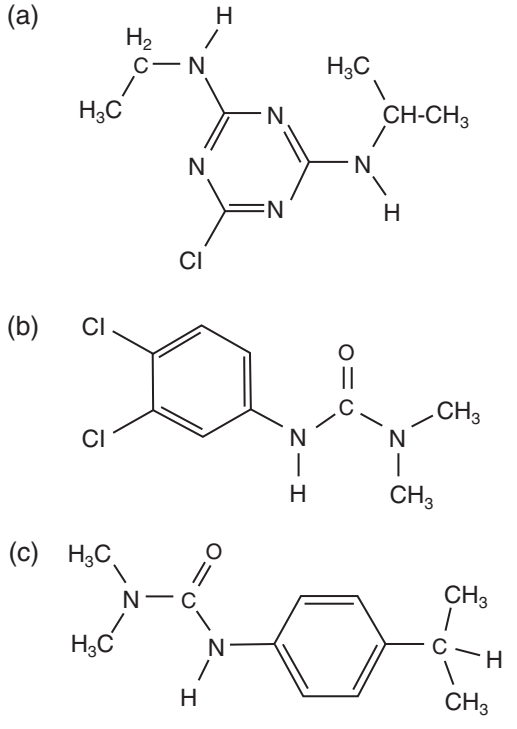


Figure 5.20 Structural formulae of three water-soluble and relatively persistent herbicides: (a) atrazine (6-chloro- *N*-ethyl-*N*-isopropyl-1,3,5-triazine-2,4-diamine), (b) diuron (3-(3,4-dichlorophenyl)-1,1-dimethyl ureum), and (c) isoproturon (3-(4-isopropyl)phenyl-1,1-dimethyl ureum).

exposure concentrations (0.5–2 µg/l) many genes are regulated by herbicides, even though herbicides are not known for their great toxicity to animals. The altered transcriptional profile indicates a broad metabolic effect including a general upregulation of energy production. What the long-term effects of this alteration may be remains uncertain; however, the gene-expression changes seen by Tanguy *et al*. (2005) are consistent with earlier physiological measurements on mussels, in which energy metabolism, often expressed as *scope for growth*, is one of the most sensitive indicators of pollution (Bayne 1989). The data suggest that bivalves in the marine environment possess great flexibility to respond to xenobiotic exposure with adequate transcriptional change, but that this change could alter their filtration capacity and so their ecological functioning in the long term.

5.5.3 Endocrine disrupters

Towards the end of the 1990s environmental scientists began to realize that a wide variety of chemicals in the environment could disrupt endocrine functions of animals. The initial discovery was made by Soto *et al*. (1991), who reported that breast cancer cells sensitive to oestrogen responded to a then-unknown compound leaching in very low amounts from laboratory plasticware made of polystyrene (Colborn *et al*. 1996). The compound was identified as *p*-nonylphenol, one of a family of synthetic chemicals called alkylphenols, which are added to polystyrene and polyvinylchloride to improve stability of the plastic. Alkylphenols are also produced as biodegradation products of alkyl polyethoxylate detergents, and are therefore found in sewage effluent and wastewater from septic tanks. The effect of nonylphenol on oestrogen-sensitive cells appeared to be due to its binding to the *oestrogen receptor*, one of several steroid hormone-binding proteins present in the cytoplasm. When activated by oestrogen the receptor undergoes homodimerization, translocates to the nucleus, and binds to *oestrogen-responsive elements*, regulating transcription of a great variety of genes. Although endogenous steroid hormones such as 17β-oestradiol have the greatest affinity for the oestrogen receptor, numerous compounds, of which some are very abundant in the environment, have been shown to interfere with steroid hormone receptors, either as *agonists* (mimicking the effect of natural hormones) or *antagonists* (suppressing the natural action of a steroid by blocking its receptor). In addition, several environmental chemicals influence hormone metabolism by inhibiting certain forms of cytochrome P450 that metabolize steroids, for example the conversion of testosterone into oestradiol, an activity known as *aromatase*.

Endocrine-disrupting chemicals may be pesticides or metabolites of pesticides, as well as industrial and household chemicals. Oestrogen-active compounds are also found naturally in plants, and are called *phytoestrogens*. A well-known case of phytoestrogen action is the occurrence of isoflavonoids in Australian clover, *Trifolium subterraneum* (Leguminosae), which was identified as the cause of impaired sexual performance in sheep. Phytoestrogens from soybean (*Glycine max*, Leguminosae) are considered as an alternative to oestrogen therapy for menopausal symptoms. Why plants would produce compounds that affect the endocrine system of vertebrates is not clear. Maybe it is just a side-effect, while the main function of these compounds lies elsewhere. For example, isoflavonoids are also implicated in signalling between plants and microorganisms in the rhizosphere. Still, it is often assumed that plants have evolved phytoestrogens as a defence strategy against herbivory. If this is the case, one would expect effects of endocrine disruptors to be more severe in carnivores than in herbivores, because herbivores have had ample opportunity to adapt to plant-derived endocrine disrupters. This prediction, put forward by Wynne-Edwards (2001), still needs to be evaluated.

Endocrine disruptors are associated with a range of adverse effects observed in terrestrial and aquatic wildlife (almost exclusively vertebrates), varying from developmental disorders to decreased fertility. One of the effects which has attracted a lot of attention is the occurrence of intersex fish due to the feminization of males. An indicator of feminization is expression of the gene *Vtg*, which encodes an egg-yolk protein, *vitellogenin*, which is normally only expressed in the female gonad. The levels of vitellogenin in male trout, caged in rivers at several distances from sewage-treatment facilities in the UK, were found to decrease with increasing distance from the sewage outlets (Harries *et al*. 1997). Following the discovery of such effects, many industrialized countries have started programmes to screen all existing chemicals for their possible endocrine-disruptive properties.

That many chemicals can be labelled as potential endocrine disrupters is now beyond doubt. Laboratory studies have demonstrated that anti-conception oestrogens, alkylphenols, phthalates, and some organochlorine pesticides can cause reproductive disorders in fish in the nanogram and lowmicrogram per litre range (Matthiessen 2000; Mills and Chichester 2005; Sumpter 2005). However, evidence that these chemicals actually impair populations of fish in the wild is less convincing. One of the problems is the lack of reliable and sensitive

methods to assess the reproductive status of wild fish. Against this background, endocrine disruption recently became a favourite model for studies in ecotoxicogenomics. Because steroid hormone receptors are regulators of gene expression, any agonistic or antagonistic impact on such receptors should be clearly visible in transcription profiles.

Larkin *et al*. (2002) developed a screening method for profiling 132 genes in largemouth bass, *Micropterus salmoides* (Perciformes, Centrarchidae). Their 'boutique' gene array was used subsequently to assess transcriptional profiles in fish exposed to *p*-nonylphenol and DDE (2,2-bis(*p*-chlorophenyl)-1,1-dichloroethylene), a degradation product of the insecticide DDT (2,2-bis(*p*-chlorophenyl)-1,1,1-trichloroethane) and a suspect hormone disrupter. Oestradiol was used as a positive control. As Fig. 5.21 shows, there was considerable overlap between the expression profiles of nonylphenol and oestradiol. Four vitellogenin genes were induced, plus two choriogenins. These effects were expected because it is known that oestradiol induces the production of yolk proteins as part of the process of oogenesis. Induction of aspartic protease may also be related to this process, since work on zebrafish has suggested that this enzyme plays a role in post-translational processing of vitellogenin in the liver prior to secretion into the bloodstream. The role of the other genes in the steroid hormone response is less clear.

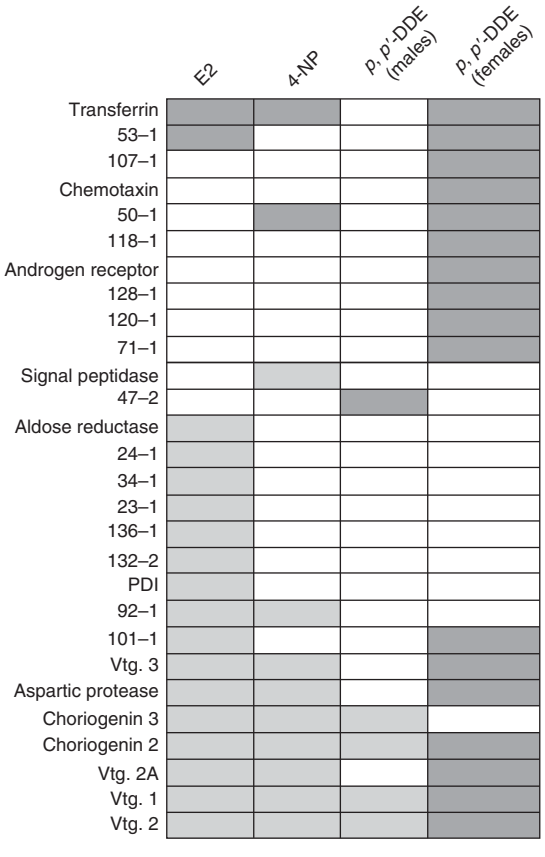


Figure 5.21 Summary of gene-expression changes in largemouth bass (*Micropterus salmoides*) exposed to oestradiol (E_2), *p*-nonylphenol (4-NP), and *p,p´*-DDE, a stable degradation product of DDT. Genes upregulated are shown by dark-grey shading; genes downregulated are shown by light-grey shading. Vtg, vitellogenin; PDI, protein disulphide isomerase. Genes with unknown function are indicated by codes. After Larkin *et al*. (2002) with permission from Elsevier.

Interestingly, some genes were regulated by nonylphenol but not by oestradiol. This suggests that nonylphenol cannot be considered a pure oestrogen mimic but must have additional modes of action that are independent of the oestrogen receptor. The same conclusion was reached in a proteomics study on zebrafish by Shrader *et al*. (2003). The compound DDE downregulated several oestrogen-responsive genes and a number of others (Fig. 5.21). It was striking that DDE had a much stronger effect on females than on males. This compound obviously interferes with fish reproduction but its mode of action requires further investigation.

We must conclude that, as in the case of heavy metals, endocrine disrupting compounds do not represent a homogeneous class from a mode-of-action point of view. This is also confirmed by a gene expression profiling study by Moens *et al.* (2006). These authors analysed genomic responses in *Cyprinus carpio* (carp) exposed to fourteen different compounds including derivatives of oestradiol, testosteron, industrial chemicals such as dibutyl phthalate, and pesticides such as vinclozolin. Despite the fact that all these compounds are classified as reference endocrine disruptors by the OECD, they produced different expression profiles. A minimal subset of genes could be defined that allowed discrimination of the compounds.

A synthetic view of the mode of action of endocrine disrupters is not yet available. It may be

expected that the picture will be quite complicated. There are several different steroid hormone receptors in the cell that can be activated by steroids such as testosterone, oestradiol, cortisol, progesterone, and others. Endocrine disrupters may activate but also inactivate these receptors; in addition, some endocrine disrupters also interact with the receptors for thyroid hormone and retinoids and the cascades triggered by these receptors may interact with steroid hormone transcriptional activation. This suggests that hormone disruption is not to be considered a simple linear process. A genome-wide understanding of endocrine disruption in fish will probably only come from species with completely sequenced genomes, such as zebrafish.

This example of endocrine disruption completes our short overview of ecotoxicogenomics. It is obvious that this subdiscipline of ecological genomics is lagging behind in comparison with other areas of stress ecology, such as abiotic stress in plants. Still, the examples shown illustrate the great potential of ecotoxicogenomics to contribute to the classification and risk assessment of chemicals (Ankley *et al.* 2006; Van Straalen and Roelofs 2008; Van Aggelen *et al.* 2010). Gene expression profiling is advocated by pointing out three advantages over classical bioassays:

- greater specificity and insight in the mode of action,
- greater sensitivity; assessment of lower environmental concentrations, and
- more rapid assessment, in the order of days rather than weeks.

Despite these obvious advantages, data generated by genomic tools such as microarrays are not yet accepted routinely in regulatory ecotoxicology. Bishop *et al*. (2001) analysed the situation and made three recommendations that we endorse: (i) risk assessors must be proactive and involved in identifying research issues that ecotoxicogenomicists should address, (ii) risk assessors must assist genome investigators in finding ways of interpreting gene-expression data that support insights into hazards and dose–effect relationships, and (iii) there must be a continuing and effective dialogue in both directions, so as to maximize information exchange in a way that can inform policy decisions.

5.6 Genomic approaches to ecological stress: an appraisal

We started this chapter by framing stress responses in the context of the ecological niche. Have the genomic insights reviewed in this chapter deepened our insight into what constitutes the boundaries of the ecological niche? It seems fair to admit that, more than in the case of community ecology (Chapter 3) and life-history analysis (Chapter 4), genomics and niche theory still seem to exist on different planets. It may be that our aim for this chapter was a bit too ambitious; however, some general principles have emerged that can help us to connect the two disciplines.

We have seen that stress factors in the environment can impinge on many aspects of the metabolism of cells. Outlines were given for several pathways that translate stress signals into gene expression. Although the details vary per pathway, some common properties emerge: (i) in the case of physical stress, deviation from normal conditions is noted by a stress-specific sensing system; (ii) in the case of chemical stress, there is a specific interaction with a cytosolic receptor protein; (iii) some stress pathways have a system of double-negative control—that is, stress removes the degradation of an activator; (iv) the stress signal is transduced via a more or less complicated network, often involving protein kinases; (v) there is a lot of interaction (cross-talk) between the pathways triggered by different stress signals; (vi) most stress-transduction systems converge on a transcription factor that is translocated to the nucleus; (vii) additional nuclear factors are often needed for activation of the transcription factor; (viii) gene expression is promoted by binding to specific DNA sequences, which are present in a battery of genes; and (ix) many genes have more than one transcription factor-binding site and thus are activated by more than one stress signal.

We have also seen that there is both commonality and divergence in the stress response. Some aspects of the stress response, for example induction of heat-shock proteins, are very general and can be found in nearly every cell under nearly every stress. Other aspects are specific for the stressor and trigger a limited number of genes addressing the stress

factor; for example, the induction of proteins preventing the growth of ice crystals in body fluids.

Despite the great increase in knowledge represented by the genomic studies reviewed in this chapter the biological significance of stress-induced transcription profiles still needs further examination. It is not impossible that environmental stress induces groups of genes that are not functionally related. In Chapter 1 we mentioned a study by Spellman and Rubin (2002), who noted the presence of transcriptional territories in the genome of *Drosophila*: sets of physically adjacent but functionally unrelated genes that are expressed jointly in association with chromatin remodelling. Some of the unusual gene expressions observed in transcription profiles may be due to these effects. The lesson could be that it is always necessary to frame gene expression in terms of a biological scheme, such as the stress-transducing systems and transcription factor cascades discussed in this chapter.

Ecological studies applying genome-wide approaches to assessing physiological stress in animals or plants under natural conditions have not yet been published. The most commonly analysed single-gene indicator of stress is Hsp70 (Feder and Hofmann 1999). Expression of heat-shock proteins is used to obtain an indication of physiological stress and this may explain ecological interactions in the field. For example, Burnaford (2004) showed that a species of chiton, *Katarina tunicata* (Mollusca, Polyplacophora), was experiencing temperature stress during low tide on a rocky shore under unshaded conditions, and this could explain its association with a species of kelp, *Hedophyllum sessile* (Phaeophyta, Laminariales), under which it finds shelter. The author showed that it was abiotic stress, not predation risk or lack of food, that limited the animal's habitat use on exposed rock surface. In this study stress was measured by Hsp70 expression; one can easily imagine how transcription profiling could further deepen the insight into explaining habitat choice and niche dimensions of organisms in the wild.

The question may be asked, does the transcription profile of an organism under stress contain all the information needed to identify the nature of the stress factor? That is, can we read the cause of stress from the transcriptome? This *inverse approach* to transcription profiling seems particularly relevant in an applied context, when genomics is used to assess the quality of the environment. Although it seems reasonable to assume that this question can be answered in the affirmative, up to now a proof of principle is lacking.

Problems facing the inverse approach are twofold. In the first place, antagonism in mixtures may mask gene expressions. If one stress factor suppresses the inducing action of another, a combination of the two will not be noted in the transcription profile. Studies on plant responses to abiotic stress reviewed in this chapter have shown that such antagonistic interactions may be realistic. Secondly, the inverse approach only works if there is a monotonic relationship between the intensity of the stress and the degree of gene expression. At low stress intensities this may indeed be the case, but under severe stress some genes induced by low stress may be repressed due to toxicity or other disturbing factors. In that case a low level of gene expression found in a transcription profile does not have an indicative value because it may have two different causes.

In general, the student of stress responses can learn from toxicology that biological responses are always dose-dependent. Toxicologists are trained to characterize intensity and duration of exposure carefully and to always include several exposure levels in an experiment. A similar attentiveness to the exposure side of stress is often lacking in biochemical stress studies.

In conclusion we note that the large gap between niche theory and genomics is fed by the fact that almost all genomic studies on stress responses are conducted in the safe environment of the laboratory. We may expect that genomic studies will now be expanded quickly to include profiling of organisms under natural conditions, in gradients of environmental stress, in extreme environments, or as a function of their distributional range. This chapter has illustrated that the genomic tools and their biochemical interpretation are ready for ecologists to take and apply.

CHAPTER 6

Variation and adaptation

Polymorphisms in the DNA sequence of an organism are a potential source of phenotypic variation and can contribute to microevolution and speciation. Evolutionary biologists, maybe even more than other biologists, are aware of the fact that the genome of any organism is not a fixed entity but varies from one individual to another and changes over time due to mutation, recombination, and selection. The characterization of such variation has been a long-term objective in population genetics. However, in the early days of genome science, an emphasis was often placed upon 'the' genome of an organism, as a unique entity, with a single sequence that had to be revealed. Heterozygosity and polymorphisms were regarded a nuisance and usually ignored. Now, population genetics and genomics have joined forces. Currently, a complete overview of genetic variation can be obtained by comparing multiple genomes of the same species. The variation in a genome is considered an extremely useful resource for identifying disease genes, geographic patterns, and locally varying selection pressures. Once a reference genome for a species has been characterized, it becomes much easier to identify the variability in that genome. Various projects to catalogue genome-wide genetic variation are now underway, such as the '1000 genomes' project for humans, the '1001 genomes' project for *Arabidopsis*, and the 192 *Drosophila* strains project. In this chapter we will discuss the many different ways in which mutation can introduce variation in a genome and how genome-wide polymorphisms can be analysed to identify factors acting as selective agents on organisms in the wild.

6.1 The internal tangled bank

On the last page of *The Origin of Species* Charles Darwin summarized his idea of evolution in a much-cited passage (Darwin 1859):

> It is interesting to contemplate an entangled bank, clothed with many plants of many kinds, with birds singing on the bushes, with various insects flitting about, and with worms crawling through the damp earth…

We do not know what inspired Darwin to express himself in his way; was it the verge of the sand walk behind his house of was it a hollow road in the English landscape that evoked the image of a tangled bank? Anyway the concept has become so famous that the tangled bank became an emblem of evolutionary theory as a whole.

Darwin's tangled bank emphasizes the environment as the place where evolution is operating and identifies natural selection as the main causative agent for change. Darwin could not know how heritable variation was generated, although he made the correct assumption that it was a blind process. Now that we are in the age of genomics, we witness an enormous growth of our knowledge on the very mechanisms about which Darwin was ignorant. Some authors have therefore applied Darwin's analogy of the tangled bank to the genome. Dover (1999) made a distinction between the 'external tangled bank' (the ecology) and the 'internal tangled bank' (the genome), attributing to them complementary functions in the evolutionary process (Fig. 6.1). The concept of the internal tangled bank emphasizes the role of *genetic turbulence* (gene duplication, genetic sweeps, exon shuffling, transposition, etc.) in the genome and it illustrates that

there is ample scope for 'innovation from within'. These innovations are then checked against the external tangled bank, and this constitutes the process of evolution. Dover's way of looking at the evolutionary process comes close to François Jacob's famous description of 'evolution through tinkering' (Jacob 1977).

Genetic turbulence and tinkering leave many traces in the genome, some with fitness consequences but many with only neutral or weak effects on the phenotype. These traces are there for genome scientists to discover; they provide a valuable historical record to reconstruct evolution.

In the past, ecologists have sometimes overemphasized the adaptive aspect of evolution, assuming that natural selection would inevitably lead to a phenotype that was optimized, in all of its biological traits, against the environment. This 'adaptionist programme' was heavily attacked by Gould and Lewontin (1979) in their paper 'The spandrels of the San Marco and the Panglossian paradigm', one of the most sharply positioned, but also heavily debated, essays of evolution. Gould and Lewontin called for a pluralistic approach, allowing explanations of form, function, and behaviour that do not necessarily include adaptation and selection. In a review of the spandrels paper 20 years later, Pigliucci and Kaplan (2000) concluded that the truth may be somewhere in the middle. Early evolutionists had certainly underestimated the connectedness in genetic networks, causing many constraints to the evolution of any one character. A system of 'survival of the barely tolerable' seems to be more applicable than 'survival of the fittest'.

In the genomics era there is even more reason to be careful with adaptionist explanations. Many genome biologists hold the opinion that nonadaptive processes such as mutation, recombination, and genetic drift are dominating the architecture of genomes. According to the neutral theory (Kimura 1983), now often called the *strictly neutral theory*, DNA sequence variation is dominated by mutations that are either selectively neutral or deleterious. Deleterious mutations are expected to be eliminated rapidly by purifying selection. Beneficial mutations are expected to be extremely rare. Later, the strictly

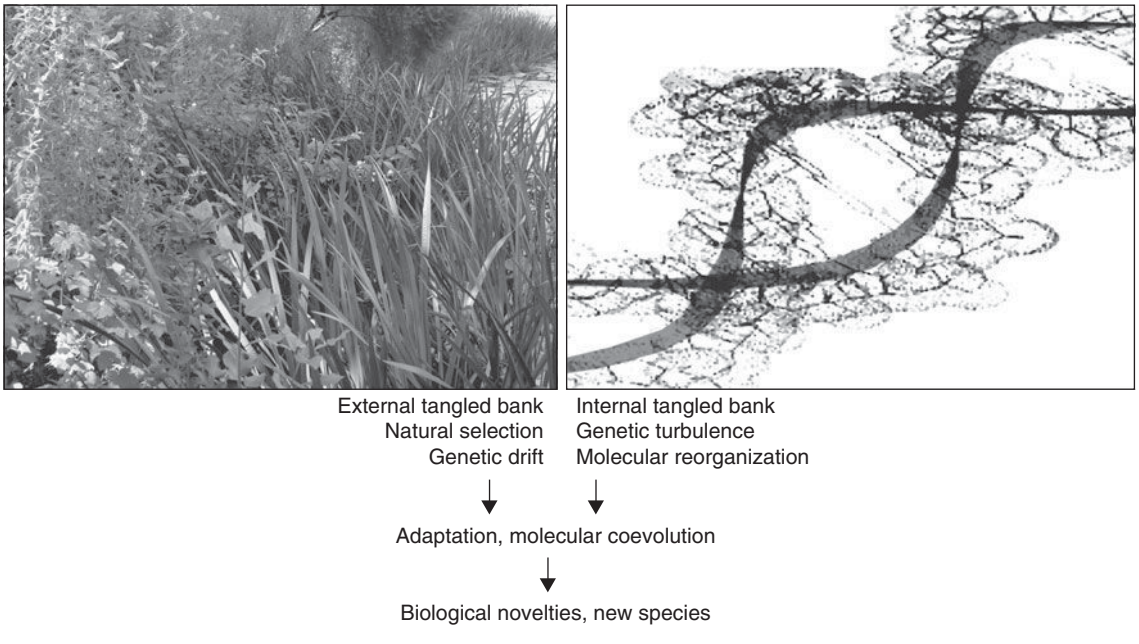


Figure 6.1 Evolution viewed as an interplay between the two 'tangled banks' of genetic turbulence and natural selection. Modified after Dover (1999), by permission of Oxford University Press.

neutral theory was expanded to include slightly deleterious mutations and weak selection. This is known as the *nearly neutral theory* of molecular evolution. Slightly deleterious mutations are not quickly removed by negative selection and can even go to fixation due to random genetic drift. Such mutations, being subject to both selection and drift, would make up a significant fraction of the polymorphisms in a population (Ohta 1992).

The debate about the role of natural selection versus mutation and neutral processes is still going on in recent literature. Nei (2007) argued that a considerable portion of both molecular and phenotypic evolution is caused by neutral or nearly neutral processes. He calls for a new '*mutation theory of phenotypic evolution*', in which evolutionary change is driven mainly by mutations, with natural selection playing a minor role. Likewise, Lynch (2007a) warned against the 'frailty' of adaptive hypotheses for the origins of organismal complexity. To illustrate the point, Lynch (2007a) provided a list of eleven aspects of genome architecture that can only be explained after accounting for nonadaptive evolutionary forces. These aspects include genome-wide A/T composition, isochore structure, preservation of duplicate genes and pseudogenes, differential proliferation of mobile elements, and others. However, some authors have argued that the reach of selection includes the genome. Hurst (2009) pointed at the existence of selection acting upon synonymous mutations and selection favouring clustering of genes with coordinated expression. We will discuss evidence supporting both positions in this chapter.

In the sections to follow, we will explore what variation exists in the genome of a single species, how it can be visualized, and how it can be analysed to infer natural selection and adaptation, with neutral variation taken as a null hypothesis. We will specifically address the upcoming field of population genomics: the analysis of genome-wide variation in populations. Then we will consider the factors causing variation in gene expression, including promoter evolution and developmental change. We expect that a further growth of ecological genomics will contribute to forging mechanistic links between genomic polymorphisms and phenotypic fitness traits.

6.2 Genomic polymorphisms

Spontaneous mutations in the genome are the primary source of all genetic variation. By way of introduction, this section discusses the different types of mutation that can occur in the genome and what type of polymorphisms may result. We show that genomics tools have facilitated new and accurate estimations of mutation rates and their effects on phenotypic change. In addition, we will see that genome-wide analysis of polymorphisms in field populations may provide a signature of selection.

Basically, DNA can mutate in four ways: *substitution*, *deletion*, *insertion*, and *inversion*. Substitution is the change of a nucleotide for another and is generally caused by errors in DNA replication. Substitutions can be divided into two classes: *transitions* and *transversions*. A transition is the change of a purine into another purine (A/G) or of a pyrimidine into another pyrimidine (C/T). A change from a purine to a pyrimidine or vice versa is called a transversion. If a substitution causes a nucleotide site to become polymorphic we call this site a *single nucleotide polymorphism* (SNP, pronounced *snip*).

If substitutions take place in coding regions they can either be *synonymous* or *nonsynonymous*. A synonymous substitution does not change the amino acid sequence of the encoded protein and is therefore also called *silent*. However, such mutations may not always be selectively neutral since there are specific biases in codon use (see Chapter 1). Nonsynonymous substitutions do change the amino acid composition of the encoded protein.

Finally, substitution can result in a stop codon, which is called a *nonsense mutation*. If we look closely at the properties of the genetic code we can conclude that synonymous substitutions occur at the third codon position. Most (but not all) substitutions at the first codon position cause an amino acid replacement and all nucleotide substitutions at the second codon position cause either an amino acid replacement or a nonsense mutation. Substitutions and the resulting SNPs are scattered randomly throughout the genome with a higher abundance in non-coding regions. As we shall see in Section 6.2.2, the distribution of SNPs in a genome is a welcome resource for high-density mapping studies and the

identification of associations with quantitative phenotypic traits (QTLs).

Insertions and deletions (*indels*) involve a variable number of nucleotides ranging from one base to large blocks of DNA. They occur at quite high frequency in non-coding regions of the genome. Indels in a coding region, unless they are a multiple of three bases, will cause a *frameshift mutation*, that is a shift of the open reading frame over one or more bases, often quickly leading to a premature stop codon. The main source of small indels is again errors in DNA replication, whereas large indels seem to be caused by unequal crossing-over and transposition. Transposons or transposable elements cause DNA segments to change chromosomal position thereby generating indels. Unequal crossing-over is believed to be very important in the generation of multigene families such as rRNA genes.

A special case of genomic variation, extremely popular among ecologists, is formed by *microsatellites*, more generally called *simple sequence repeats* (*SSRs*). These represent loci where short sequences (two to six base-pairs) are tandemly repeated in variable numbers. They are usually visualized by a simple PCR targeting conserved sequences flanking the locus, delivering a PCR product of variable length. Microsatellite polymorphisms are generated by slippage during DNA replication. When the nascent strand disassociates from the template strand in a region containing repeats, the nascent strand may easily shift across a number of bases to reanneal slightly out of phase. This causes the nascent strand to be longer or shorter by one or more repeat units.

For the statistical analysis of population surveys of microsatellite frequencies, different mutation models have been proposed, the extremes being the *stepwise mutation model* (*SMM*) in which microsatellite alleles can mutate to another allele with one repeat unit less or more, and the *infinite alleles model* (*IAM*) in which any one allele can mutate to any other allele. The distinction between these models is important when making inferences from allele frequencies in population surveys. Microsatellites may be prone to homoplasy (independent mutation events leading to the same length fragment), which may cause problems when mutation rates are very high and when there are allele size constraints. These and other issues of microsatellite applications in population ecology are discussed by Estoup *et al.* (2002).

When microsatellites involve three base-pairs of a codon, they may lead to polymorphisms in protein composition. In eukaryotes repeats of glutamine, asparagine, and alanine in proteins are fairly common. So microsatellites may not always be considered neutral markers, especially not if they encode triplet repeats.

Finally, *mobile genetic elements* are an important source of polymorphism in a genome. There is a tremendous variety of such elements in the genomes of eukaryotes, including viruses, plasmids, and transposons. We have seen in Chapter 2 that the extent of mobile element proliferation in a genome is an important determinant of genome size differences between species. Mobile elements can be inserted everywhere in the genome, often without functional consequences, however, when inserted in the 5′ region of a gene, they may have an effect on transcription. An illustrative example is provided by DDT resistance in *Drosophila*: insertion of an Accord transposable element into the promoter of a cytochrome P450 gene (*Cyp6g1*) causes overexpression of this gene and production of a large amount of biotransformation enzyme that can degrade DDT. The *Cyp6g1* locus has undergone an 'adaptive walk' with several mobile element insertions and duplications, where each step provided a selective advantage (Daborn *et al.* 2002; Schmidt *et al.* 2010).

Some genes, due to their constitutively uncondensed chromatin structure, are particularly rich in mobile elements. This is the case with the heat-shock genes in *Drosophila*, which have an extraordinary number of P-elements in their promoters (Walser *et al.* 2006). The effect of P-elements on heat-shock gene expression causes a large amount of natural variation for selection to act upon.

Despite the obvious possibilities, mobile genetic elements are not often used for genotyping in natural populations, however, many of the polymorphisms used in anonymous restriction-based fingerprinting methods such as AFLP might

actually rely on variation in the insertion of mobile elements.

6.2.1 Patterns of synonymous and nonsynonymous substitutions

For a full understanding of genomic polymorphisms, an appreciation of the mutation process is needed. However, measuring mutation rates is extremely difficult, because the frequency of mutation within each generation is very low. The traditional way to study genomic mutation rate is to perform *mutation accumulation* (MA) experiments. The standard way to design such an experiment is described very well by Keightley and Charlesworth (2005) and may be outlined as follows. An isogenic line from the organism of interest is taken as the inbred progenitor, and is subdivided into several MA lines. Going through several generations of inbreeding (selfing or full-sib mating) these lines accumulate mutations at random, causing divergence at the phenotypic as well as the DNA level. One or a few individuals from each generation are isolated to form the next generation so that the potential role of selection is diminished. The inbred progenitor population is considered to be mutation-free and should be preserved in that state. The mean fitness of all MA lines is estimated at generation time t. Also, the mean fitness of the mutation-free population can be inferred. The first parameter of interest is the change in mean fitness per generation that is due to MA (ΔM). Secondly, we can estimate the increase in genetic variance in fitness among inbred lines per generation (V_m). If we assume that the average deleterious effects of mutation (s) are equal, then $\Delta M = U_d s$, where U_d is the genomic rate for mutations affecting fitness per generation. Furthermore, Keightley and Eyre-Walker (1999) showed that $V_m = U_d s^2$, so that U_d can be estimated from the formula: $U_d = \Delta M^2 / V_m$.

A lot of the work on MA has been performed using *D. melanogaster* as a model organism, but data are also available for *C. elegans*, *E. coli*, and humans. Keightley and Eyre Walker (1999) showed that U_d estimates vary greatly, ranging from 0.00017 in *E. coli* to 0.47 in humans. In *D. melanogaster* several U_d values have been estimated in different studies and these data vary over an order of magnitude (0.02–0.47), whereas the effect of deleterious mutations was estimated at 3%. We have to conclude that these figures are not very reliable and probably underestimate the actual mutation rate in a genome. The method only detects mutations of moderate effect that change phenotypic traits, whereas mutations with very small effects and mutations caused by transposon activity are not recognized.

The problems associated with mutation-rate estimates based on phenotypic traits were overcome in a study on *C. elegans* by Denver *et al.* (2004). These authors applied a genomics approach that provides a direct estimate of mutation rate in DNA sequences from a set of MA lines. Some 4 Mbp of randomly selected DNA segments was sequenced directly from a set of MA lines after t = 280, 353, and 396 generations. A per-nucleotide mutation rate can be determined from the sequence data, which can be used to calculate a haploid genomic mutation rate U_t per generation. Denver *et al.* (2004) detected 30 mutations, which translates to a mutation rate of 2.1×10^{-8} mutations per site per generation and a U_t value of approx. 2.1 mutations per genome per generation. Surprisingly, the direct estimates of Denver *et al.* (2004) are at least 10 times higher than previous estimates (Drake *et al.* 1998). Furthermore, 17 of the 30 observed mutations were indel mutations, and 13 of these 17 indels were insertions.

How can these high mutation rates and the predominance of insertions be explained? Denver *et al.* (2004) suggested that their study was less biased by genetic selection than previous studies. It seems that pseudogenes, on which earlier studies were based, do not evolve neutrally but may be under selection (Hirotsune *et al.* 2003). Alternatively, Rosenberg and Hastings (2004) propose that phenotypes are masking molecular variation in such a way that mutation rates are higher than predicted. They also propose that accumulation of harmful mutations in MA lines can induce increased mutability. Harmful mutations will cause cellular stress, and in response to that stress mutation rates may increase. However, Keightley and Charlesworth (2005) state that this mutator state is unlikely, because such a genetically unstable state would induce a rapid decline of fitness, and this was not observed.

It remains difficult to make generalizations about mutation rates in organisms. The genomic data on direct mutation-rate estimates generated by Denver *et al.* (2004) raise new questions, for instance on how their data on *C. elegans* relates to earlier estimates on *Drosophila*. As proposed by Keightley and Charlesworth (2005), it may be worthwhile repeating the Denver approach in *Drosophila*.

If DNA sequences of the same gene from two individuals or from two species are aligned, differences are often observed. As discussed above most of these differences do not result in differences in the amino acid sequence of the encoded proteins, but some do. The extent to which differences between two sequences imply differences in amino acid composition is measured by the K_a/K_s *ratio*, defined by Hurst (2002) as the ratio between the number of nonsynonymous substitutions per nonsynonymous site (K_a) to the number of synonymous substitutions per synonymous site (K_s) in a specific segment (window) of DNA sequence. This ratio is also designated the d_n/d_s *ratio*, where d_n is comparable to K_a and d_s comparable to K_s. The rate of synonymous substitutions (K_s) is often equated to neutral mutation rate of the gene, whereas K_a indicates the degree of protein evolution (mostly functional). The K_a/K_s ratio of a certain sequence therefore tells us something about how that sequence has evolved (by neutral processes or under selection).

Usually K_a is much smaller than K_s, because a change in the amino acid is less likely than a silent substitution. This is due to the fact that selection eliminates deleterious mutations to keep the function of a protein intact. So under *purifying selection*, $K_a/K_s < 1$, approaching 0 for complete conservation. However, sometimes K_a is greater than K_s, for instance in genes associated with the immune system that are coevolving with parasites (Nei *et al.* 1997). $K_a/K_s > 1$ indicates selection acting positively on protein change in one sequence compared to the other.

Yang and Bielawski (2000) evaluated two ways for measuring molecular adaptation. The first class is based on intuition; the method of Nei and Gojobori (1986) is best known. Synonymous and nonsynonymous sites and synonymous and nonsynonymous differences are counted between two sequences. Subsequently, the synonymous and nonsynonymous rates are corrected for multiple substitutions at the same site using simplistic assumptions to yield estimated K_a and K_s values. Nei and Gojobori (1986) assume equal rates of transition (T to C, A to G, and vice versa) and transversion (T or C to A or G and vice versa), and uniform codon usage. However, this method yields incorrect estimates of K_a and K_s because the assumptions have been proved to be unrealistic. The method tends to perform especially poorly when codon usage is biased due to differences in translation efficiency (genes that have a high transcription/translation rate are biased in their codon usage as compared to genes that are transcribed/translated at a low rate).

A more reliable class of methods was developed using the principle of maximum likelihood. The codon is considered to be the unit of selection, and a model is developed for substitutions in the codon. Parameters in the model are, for instance, sequence divergence, transition/transversion rate ratio, and K_a/K_s ratio; these parameters are estimated from sequence comparisons by applying the maximum-likelihood principle. Parameters are also corrected for biased codon usage. An important feature of these methods is that a statistical test can be applied to test whether the K_a/K_s ratio is significantly different from unity (neutral evolution). Several software packages have been developed to study molecular adaptation using the K_a/K_s ratio.

As an example, we discuss a study by Talbert *et al.* (2004), who analysed adaptive evolution in genes encoding *centromere-binding protein C* (CENP-C). Centromere-binding proteins play an important role in fixing microtubules to the centromeric region of a chromosome, which is essential in the formation of spindles during mitosis and meiosis. It is remarkable that the centromere has such a conserved function, although its DNA sequence is non-coding and highly variable. The centromere-binding proteins seem to coevolve with the rapidly changing satellite DNA sequences in the centromere. Talbert *et al.* (2004) screened the sequences of *Cenpc* genes in pairs of related species and applied a K_a/K_s analysis to locate regions of selection (Fig. 6.2). Statistically significant selection was detected in the regions encoding DNA-binding domains. Mammals showed stretches of

positive selection, whereas negative selection seemed to be more pronounced in maize and sorghum in the comparable region. Why the centromere and consequently the centromere-binding proteins would evolve so rapidly is not known. Talbert *et al.* (2004) propose a model in which there is competition between centromere variants during female meiosis. The female meiotic spindle has an asymmetric shape, such that the 'stronger' centromere variants may be better captured and included in the meiotic product that becomes the egg nucleus. Such biased inheritance has also been described for other loci, especially

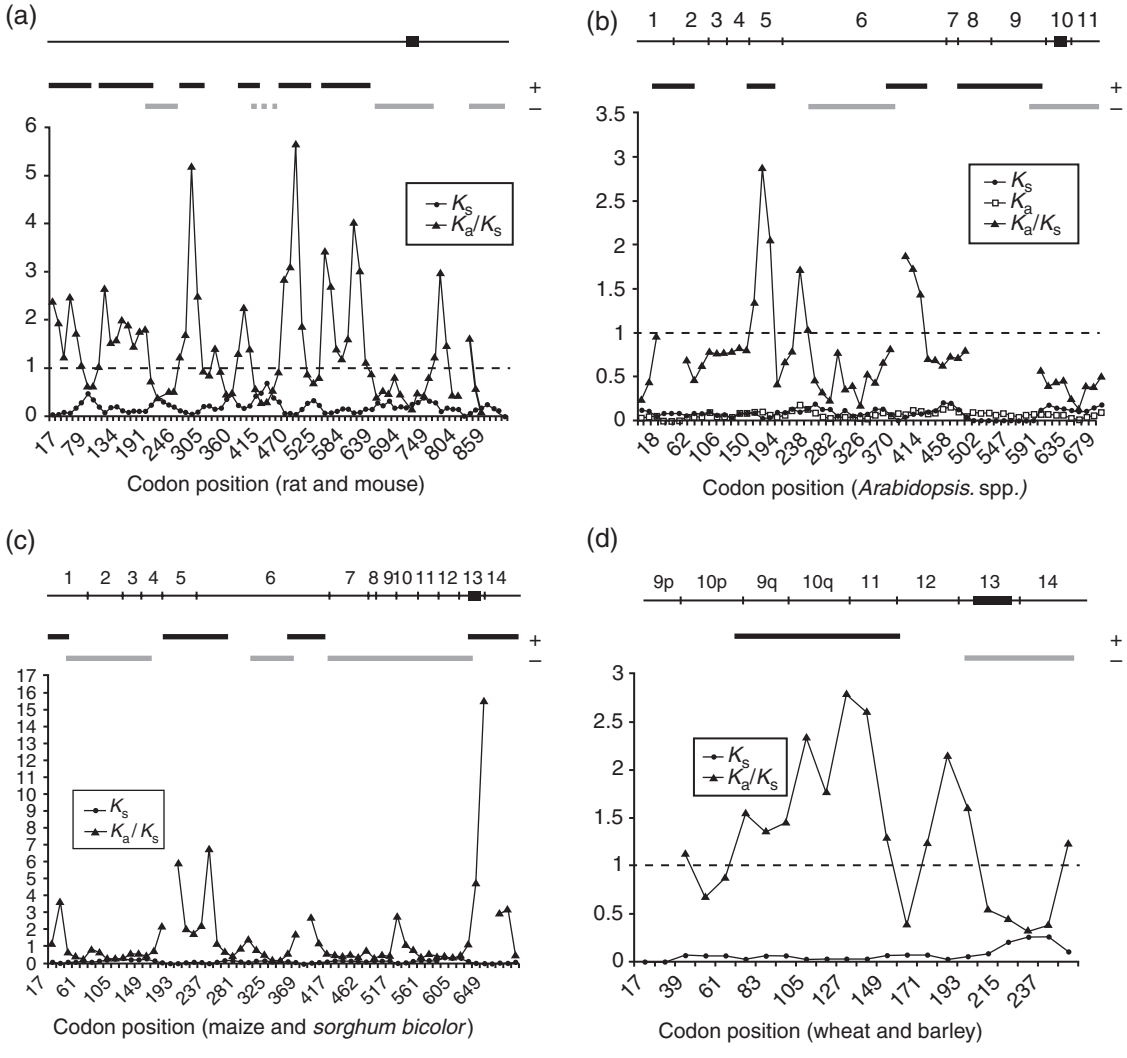


Figure 6.2 Patterns of variation between pairs of related species across the sequence of centromere-binding protein C (*Cenpc*) genes displayed using a sliding window analysis of K_a/K_s ratios. Each point represents the value of K_a/K_s for a 99-nucleotide (33-codon) window plotted against the midpoint of the window. The aligned coding sequence is plotted at the top of each graph. On the sequences, the black rectangles indicate the locations of 24-amino acid CENP-C motifs, the defining structure of this type of protein. Exons are indicated by numbers for the plant sequences. Beneath the sequences regions of positive (black bars) and negative (grey bars) selection are indicated. (a) Rat and mouse, (b) *A. thaliana* and *Arabidopsis arenosa*, (c) maize and *Sorghum bicolor*, and (d) wheat and barley, exons 9p–14. From Talbert *et al.* (2004), by permission of BioMed Central.

in *Drosophila*, and is known among geneticists as *segregation distortion* or *meiotic drive*. According to Talbert *et al.* (2004) meiotic drive can explain the apparent selection on DNA-binding centromere proteins.

Another example of the use of the K_a/K_s ratio is illustrated by Liberles *et al.* (2001), who studied adaptive evolution of amino acids at the genomic level. They calculated K_a/K_s values on nodes of branches within evolutionary lineages. The evolutionary lineages were taken from the Master Catalog, a compilation of sequence alignments and evolutionary trees for all protein modules encoded by genes in Genbank, constructed by Benner *et al.* (2000). They focused on subtrees containing only chordates and Embryophyta (mosses, ferns, and higher plants) and could identify branches with high K_a/K_s values. These branches may be indicative of positive selection, where the mutated protein has a higher fitness than the ancestral form, probably associated with a change in function. The gene families that display high K_a/K_s values were stored in *The Adaptive Evolution Database* (TAED). Currently, TAED 2.1 (www.bioinfo.no/tools/TAED) contains 6657 families that are fully processed. In 10–20% of these families positive selection was determined in at least one branch. High K_a/K_s values on branch points in evolutionary protein trees may be caused by gene amplification followed by functional differentiation of the paralogues. Orthologous genes under different selective regimes in different species may also be found. However, the database cannot distinguish between paralogues and orthologues, so the results should be interpreted cautiously. Besides gene families that were identified previously to be under positive selection, such as the MHC proteins (proteins of the adaptive immune system), quite a number of families were newly identified in TAED to have undergone change in function. The authors conclude that TAED is a useful resource for biologists searching for potential examples of molecular adaptation as a starting point for further experimental study.

We can conclude that the use of K_a/K_s ratios for measuring molecular adaptation in coding sequences of the genome is a very effective approach. We expect that databases like TAED will become an important framework for ecological genomics to study adaptation at the molecular level. The challenge will be to integrate this information with gene-expression profiling and to link molecular variation to phenotypic differences between related species.

6.2.2 Quantitative characters

One way to link genomics to ecologically important traits is through the concept of *Quantitative Trait Locus* (QTL): a polymorphic genomic segment containing one or more genes that affect the variation in a quantitative phenotypic trait, such as body size, clutch size, flowering time, disease resistance, and so on. The position of a QTL in the genome is usually established by linkage to a set of polymorphic marker loci, each segregating in a Mendelian fashion, which have to be available in large numbers with an even spread throughout the genome. QTL analysis is a very powerful approach to elucidate evolutionary and ecological processes because it allows the researcher to focus on phenotypic traits that contribute to fitness or are relevant in terms of a specific ecology (Erickson *et al.* 2004).

Traditionally identification of QTLs has been a major area of research in plant and animal breeding (Tanksley 1993; Jansen and Stam 1994). QTL mapping in this context used controlled crosses, preferably starting with two inbred parent strains that differ in the trait of interest. When the offspring from two such inbred parents are mated, recombination breaks up the linkage between traits in the parental chromosomes. Subsequent sib-mating for several generations produces a set of *recombinant inbred lines* (*RILs*), each of which contains a nearly homozygous segment from one of the parental chromosomes. The segregation of markers in these RILs is then correlated with the phenotypic trait. Data analysis is aimed at producing a graph in which the association between the trait and the marker, expressed as the *LOD* (logarithm of the odds) *score*, is plotted as a function of the position in a genetic map (cf. Fig. 6.3d).

Until the 1980s the lack of a sufficient number of polymorphic markers greatly hampered the identification of QTLs in ecologically important species. Often less than a few hundred markers could be used and only a small fraction of the total genetic

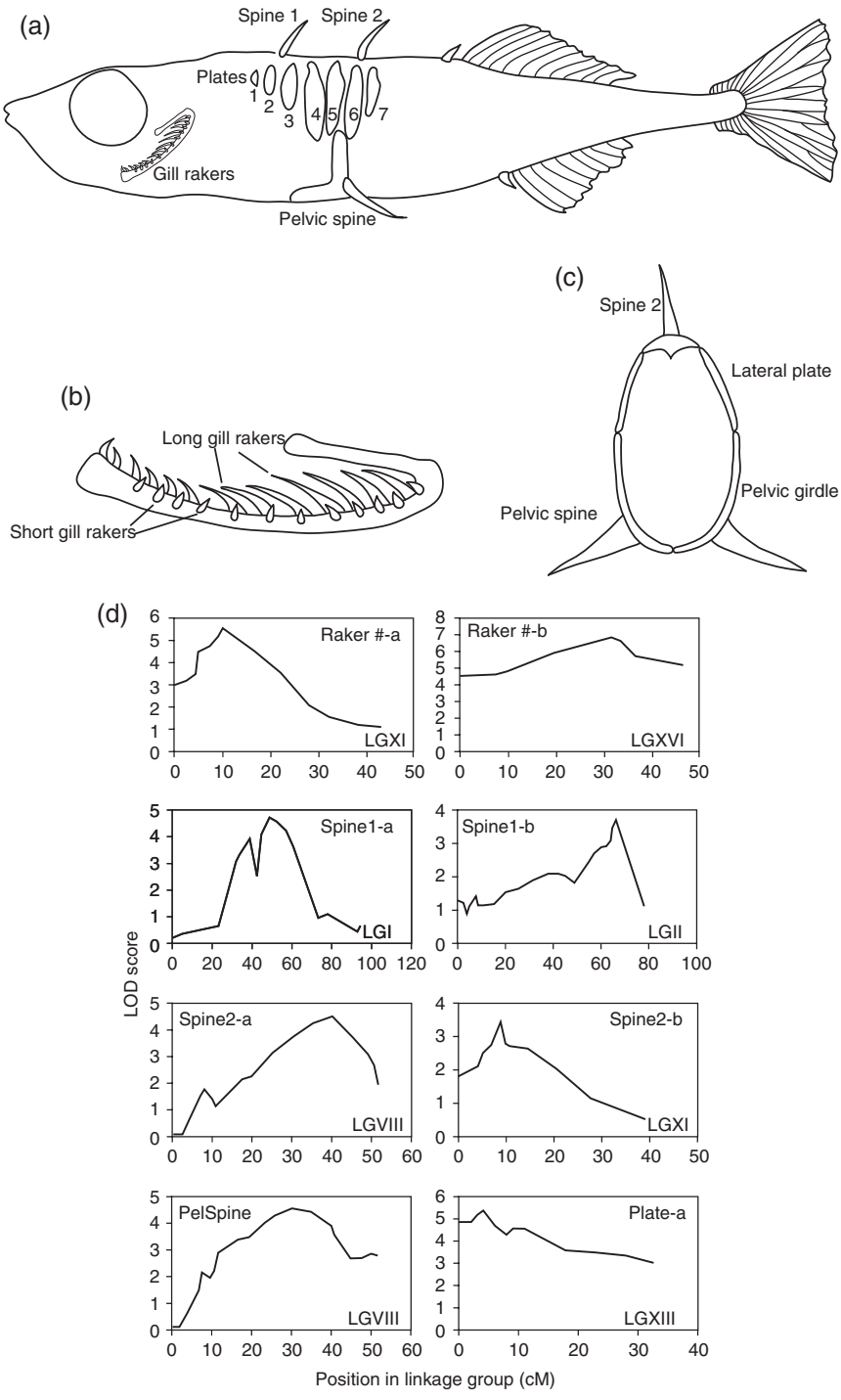


Figure 6.3 Mapping of morphological traits differentiating subspecies of the threespined stickleback (*Gasterosteus aculeatus*). (a, b, c) Schematic drawings of fish indicating the biometric characters investigated. (d) LOD (logarithm-of-the-odds, log-likelihood ratio) scores for various characters as a function of the position (in centiMorgans, cM) in the linkage group. The peaks in the LOD scores show that there are two QTLs for the number of short gill rakers, one in linkage group (LG) XI (Raker #-a) and one in linkage group XVI (Raker #-b), two QTLs for length of spine 1, two for length of spine 2, one for pelvic spine length, and one for lateral plate size. From Peichel *et al.* (2001), by permission of Nature Publishing Group.

variation could be covered. QTLs for ecologically important traits rarely mapped to individual genes; often a region of several thousand to some millions of base-pairs remained for molecular analysis. However, in a few cases it proved to be possible to pinpoint an ecologically relevant QTL to a specific gene. One of these successes was obtained in work on the threespined stickleback, *Gasterosteus aculeatus* (Gasterosteiformes).

The threespined stickleback is an originally marine, anadromous fish species, but populations have permanently colonized a variety of freshwater habitats. Being reproductively isolated from each other, these populations have developed into a wide range of subspecies with different morphologies, habitat preferences, behaviours, and life cycles. For example, in lakes in British Columbia two ecotypes are present, one specializing in benthic habitats, with a larger body-size, reduced spines and armour plates, and fewer gill rakers, the other a pelagic form that is more streamlined and has larger eyes and well-developed spines, gill rakers, and armour plates. Despite reproductive isolation in the field the two forms can be crossed by artificial means in the laboratory. Peichel *et al.* (2001) developed a genetic map from such crosses after genotyping the animals with 438 microsatellite loci. QTLs were identified for biometric characters of spines, armour plates, and rakers, and these loci were mapped into the linkage groups defined by the microsatellite markers (Fig. 6.3).

In further work, the attention was focused on the QTL for lateral plates (Colosimo *et al.* 2004, 2005). New markers were developed that narrowed down the plate morphology QTL to 0.68 cM. Two of the markers that were very closely linked with the QTL were used to screen a BAC library of the stickleback genome. After identification of the BAC containing the locus, further sequencing and the use of fine-mapping identified intron 2 of the stickleback *Ectodysplasin* gene (*Eda*) as the source of plate morphology variation. In mammals, *Eda* encodes a secreted signalling molecule involved in the development of ectodermal derivatives such as teeth, hair, and dermal bones. The *Eda* genes of low-plated and fully plated populations differed in a large number of synonymous substitutions, but also in four nonsynonymous ones. A genetic screen of 25 low-plated and completely plated stickleback populations from the US, Europe, and Japan, showed that *Eda* sequences of the low-plated populations all clustered together, except for one Japanese population. The clustering of *Eda* followed the lateral plate morphology, not the phylogenetic history or the geography of the fish. Therefore the authors could rule out the possibility that the low-plated alleles have spread through migration from a single low-plated source population. Instead it is very likely that substitutions in the *Eda* gene have occurred many times and have given rise to repeated loss of armour plates.

This work, and that of Shapiro *et al.* (2004) and Cresko *et al.* (2004), convincingly shows that rapid evolution of morphological phenotypes, through parallel genetic mechanisms, is very common in sticklebacks. Foster and Baker (2004) speculated that this might be due to specific aspects of the genetic architecture, such as instability of certain genome regions. Anyway, the threespined stickleback proved to be a fascinating system for exploring the genomic basis of adaptive radiation and parallel evolution.

Another example illustrating the identification of a molecular mechanism underlying a quantitative trait is provided by the case of flowering time in *Arabidopsis* (El-Assal *et al.* 2001). *A. thaliana* from the Cape Verde islands flower much earlier than laboratory strains from temperate regions, and are hardly sensitive to day length. Mapping with a variety of molecular markers had located an '*early day length insensitivity*' (EDI) QTL to a 50 kbp region at one end of chromosome 1. The genomic sequence showed that this region contained 15 open reading frames (ORFs); however, the gene *cry2* (cryptochrome-2; encoding a photoreceptor protein) was considered a good candidate as the causal agent of the EDI syndrome. Sequence analysis showed that the Cape Verde mutant of this gene differed from the laboratory strain at twelve nucleotide positions: four in the promoter, one in the 3´-UTR, and seven in the coding region (Fig. 6.4). One of these mutations, leading to the substitution of a valine residue for a methionine in the protein, was proved to be the cause of photoperiod insensitivity.

The CRY-2 protein appears to control a signalling pathway involving genes that promote flowering (see Section 4.3.4); under short-day conditions, the amount of CRY-2 protein is greatly downregulated during the photoperiod and this suppresses early flowering in plants from temperate regions under short day length. The mutated protein is less sensitive to the light-induced downregulation and that is why the Cape Verde plants flower earlier. It is obvious that in the tropical Cape Verde islands there is less need for suppression of flowering in response to short photoperiods, so the plants with mutated CRY-2 would have increased in frequency on the Cape Verde islands and natural selection finally drove the mutation to fixation.

The genomic revolution has opened up new prospects for QTL mapping by using high-density *single nucleotide polymorphisms* (*SNPs*) maps (Borevitz and Nordborg 2003, Brumfield *et al.* 2003). As mentioned above, SNPs are positions in the genome at which at least some individuals of a species have a base-pair different from the most common form. Depending on the species, there is an SNP every 50 (in *Drosophila*) to every 1000 (in humans) base-pairs. Next-generation sequencing methods have provided very powerful instruments for discovering these

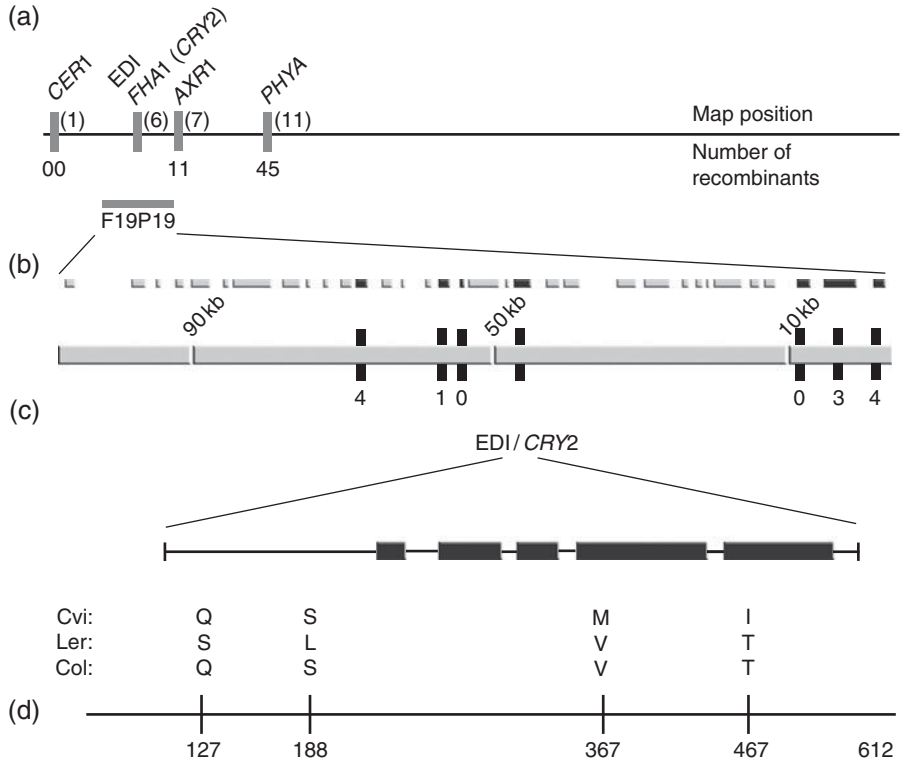


Figure 6.4 Map-based isolation of the EDI QTL in *A. thaliana*. (a) Linkage map of chromosome 1 showing the position of various molecular markers and the EDI QTL. F19P19 is the designation of the clone containing the locus. (b) Physical map of the F19P19 clone. The shaded boxes represent open reading frames according to the *Arabidopsis* genome sequence. The black markers represent seven newly developed molecular markers, used to localize the QTL further. (c) Genomic structure of the *CRY2* gene, including the 5′- and 3-UTRs and five exons (black boxes). (d) Part of the CRY2 protein sequence, showing four variable amino acids. Q, glutamine; S, serine; L, leucine; M, methionine; V, valine; I, isoleucine; T, threonine; Cvi, Cape Verde mutant (with EDI); Ler, laboratory strain (not EDI); Col, Colombia strain (not EDI). After El-Assal *et al.* (2001), reproduced by permission of Nature Publishing Group.

SNPs. The reason is that the new methods use a very large sequencing depth, up to sixtyfold coverage, meaning that a position is sequenced 60 times on average. Advanced bioinformatics software is then used to identify reads with matching sequences but differing in one base. When the technology is applied to a mixture of genotypes, for example from different populations of the species, it is possible to detect millions of SNPs in a genome or transcriptome, especially when a reference genome is already available.

To illustrate these developments, an ongoing sequencing project for the great tit (*Parus major*) may serve as an example. The great tit is a very important model species in population, evolutionary, and behavioural ecology but it still lacks a sequenced reference genome. However, recently Van Bers *et al.* (2010) using Illumina 1G technology, generated 2 Gbp of new sequence information for this species. The aim of the project was to identify a large number of SNPs for the construction of a linkage map and QTL analysis. In general it is difficult to identify SNPs in species with an unsequenced genome, such as the great tit, because there is no reference to compare the reads with. In addition, short reads lack the genomic context around the SNPs which is necessary to design probes for use in genotyping. However, Kerstens *et al.* (2009) indicated a solution for this problem. These authors showed that it is possible to assemble short sequence reads into contigs and thus to build a reference set of contigs. Mapping of the short reads onto these contigs then provides the basis for SNP discovery. This strategy was successfully followed by Van Bers *et al.* (2010) who identified no less than 20 000 novel SNPs for the great tit. They were also able to align a significant number of contigs to the zebra finch genome. This suggests that model species such as zebra finch and chicken might be very helpful in the development of genomics resources for wild bird species, thanks to the relatively uniform genome size and low rate of syntenic re-arrangements in birds as a whole (cf. Edwards 2008; Backström *et al.* 2008).

The concept of Quantitative Trait Locus applies not only to genomic locations encoding polymorphic proteins, but also to loci important in the regulation of other genes. In microarray studies, investigators have found that gene expression profiles show variability between individuals that are partly heritable. Therefore, it is possible to treat each expression level as a quantitative trait and to use QTL methodology to dissect the expression profile into its underlying genetic components. The genetic locus causing variation in the transcription of a certain gene is called an *expression QTL* (*eQTL*). One of the first studies on eQTL mapping, discussed in Chapter 4, was a genomic analysis of phenotypic plasticity in *C. elegans* (Li *et al.* 2006).

In principle, the techniques for identifying eQTLs are the same as for regular QTLs, however, the high-throughput aspect and the sheer number of loci considered simultaneously create special challenges and call for new statistical techniques. An advantage of eQTLs is that the phenotypic character considered is of a molecular nature (expression of a gene), and is therefore directly and mechanistically connected to the genome much more than the complex traits studied in normal QTL analysis, for example body size. Therefore eQTLs take a position somewhat in between sequence variation and the phenotype (Rockman and Kruglyak 2006; Mackay *et al.* 2009).

An eQTL may be positioned just upstream of the gene itself and is then called a *cis-eQTL*. For example, there might be sequence variation in the promoter of a gene that affects its expression. Consequently the promoter is a *cis-eQTL*, since it represents a proximal locus causing heritable variation in expression of the gene. Alternatively, the eQTL might be positioned remotely (distantly) from the gene whose expression it affects, and this is called a *trans-QTL*. For example, there might be a polymorphism in a gene encoding a transcription factor that affects the expression of the target gene. We have to realize, however, that there is some confusion in the use of the words *cis* and *trans* regulatory variation. In eQTL analysis these terms do not refer to the mechanism of action, but to the genomic position, inferred from linkage. There are many indirect ways in which distant loci can influence the expression of a particular gene, some of which could be due to *cis*-acting regulation of a *trans*-acting factor. Rockman and Kruglyak (2006) therefore

prefer the more neutral terms *local* and *distant regulatory variation*. However, the terms *cis* and *trans* regulation have become so common that we continue to use them in this book (cf. Section 6.3.1).

In eQTL microarray studies the positions of eQTLs are often plotted against the position of the gene that they regulate, in a graph like Fig. 6.5. In such a graph, *cis*-eQTLs are all on the diagonal, because their genomic position is about the same as the regulated gene. All off-diagonal eQTLs are *trans*-eQTLs; the most interesting are the ones which affect more than one gene at the same time and these appear as vertical bundles in the graph (*trans*-acting hotspots).

Most of the expression QTL studies to date are restricted to genetic model species such a *C. elegans* and *Arabidopsis*. These studies have revealed an extremely complicated pattern of gene regulation. In an extensive study of eQTL mapping in *Arabidopsis*, West *et al.* (2007) found that the majority of genes in the *Arabidopsis* genome are controlled by an eQTL, that is, for almost any one gene there is a heritable factor elsewhere in the genome controlling its expression. In an analysis of 211 recombinant inbred lines the authors identified no less than 36 871 distinct eQTLs. One-third of these expressions were regulated in *cis*; the remaining *trans*-eQTLs were not homogeneously distributed across the genome but concentrated in hotspots, especially on chromosome 11. Some of these *trans*-eQTLs regulated hundreds of transcripts elsewhere in the genome. This study illustrates how extremely complex the regulation of gene expression is when seen in a genome-wide perspective. The interconnectedness of the genetic factors affecting gene expression of just a single gene is overwhelming.

In a similar *Arabidopsis* study, focused on the genetic network for flowering time, Keurentjes *et al.* (2007) concluded that 46% of the genes were regulated by *cis*-eQTLs. The authors also showed that *cis*-regulation usually has a stronger effect on transcript abundance than *trans*-regulation. Both Keurentjes *et al.* (2007) and West *et al.* (2007) showed that most *trans*-eQTLs involved small phenotypic effects (small heritable expression differences). These two studies also illustrate that, to capture the complexity of genome-wide gene regulation, eQTL analysis must use a massive number of markers that cover the genome completely.

The link between quantitative characters and genomics has triggered the development of a new field, called *genetical genomics*, also called *quantitative genomics* ((Jansen and Nap 2001; Kliebenstein 2009). The aim of genetical genomics is to identify the genetic factors that cause variation in gene expression. This is seen as an important step towards a true understanding of the complicated way in which gene networks influence the expression of quantitative traits in the phenotype.

In a review of the genetics of quantitative traits in the genomics era, Mackay *et al.* (2009) introduced another concept, *Quantitative Trait Nucleotide* (*QTN*). This was defined as a locus with a causal relationship to some endophenotype, that is the network of transcripts, proteins, or metabolites that covaries with allelic variation of the locus. The authors argued that the challenge of QTL analysis is to understand how a change in an endophenotype affects the phenotypic trait of interest. A causal relationship should be sought between QTNs, eQTLs, endophenotypes, and quantitative traits.

As a complement to QTL mapping, another approach, called *genome-wide association mapping* (*GWA*) is upcoming (Bergelson and Roux 2010). GWA mapping, like QTL mapping uses linkage disequilibrium as a tool to link specific phenotypes to genes, but it is not based on genotyping experimental populations, but populations varying naturally. GWA studies take advantage of natural variation and the patterns of linkage and recombination that have evolved over thousands of generations. The GWA approach has been made possible by modern high-throughput genotyping methods, using SNP chips, and is especially applicable to species with a good reference genome sequence. In the year 2010 the first study of GWA mapping in plants (*Arabidopsis*) was published (Atwell *et al.* 2010). It identified the genetic determinants of 107 phenotypic characters such as flowering time, growth, seed dormancy, and pathogen resistance. This new approach is likely to bring a breath of fresh air to the area of gene discovery and is especially relevant for ecology because it focuses on natural variation (Bergelson and Roux 2010; Kammenga *et al.* 2008).

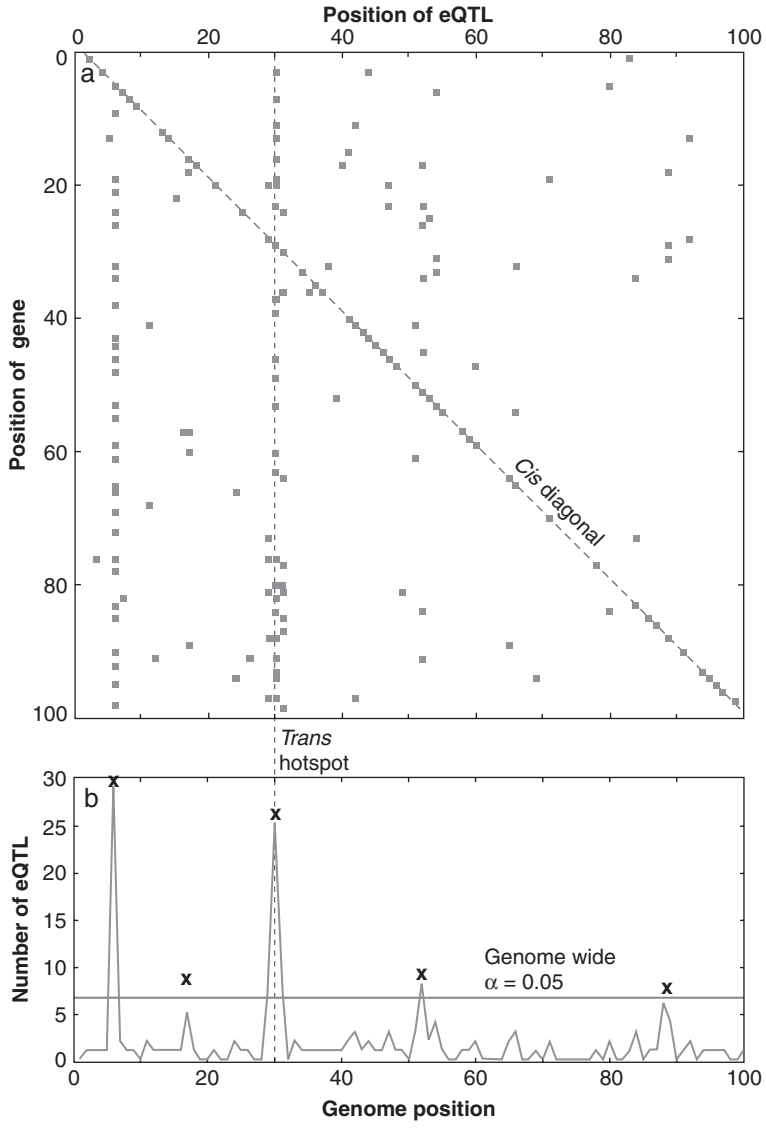


Figure 6.5 Representation of a hypothetical eQTL analysis. In the upper panel (a), the position of a locus influencing the expression of a target gene is plotted as a function of the position of that target gene in the genome. Positions are indicated here on an arbitrary scale (0–100). Expression QTLs just proximal to the target gene (e.g. a promoter) all fall on the diagonal (*cis*-eQTLs), all other spots represent *trans*-eQTLs; *trans* loci that influence the expression of many other genes irrespective of their genomic location appear as vertical bundles. Two of these *trans*-acting hotspots are shown in the lower panel (b). From Kliebenstein (2009), by permission of Annual Reviews.

6.2.3 Marker-based population genomics

If a species occupies a heterogeneous geographical area or lives in contrasting habitats, some populations may experience selective forces differing from those elsewhere. This may cause local population differentiation in polymorphic loci: some alleles confer a fitness advantage in one environment while other alleles are favoured in another environment. This well-known principle, variously called local adaptation, ecotype differentiation, or population differentiation, is expected to take place at all

polymorphic loci with alleles causing differential fitness.

Conversely, if one finds contrasting allele frequencies for some loci in the field, this may be taken as evidence that selection is occurring or has occurred. Allele frequencies of neutral loci will vary in a random manner between subpopulations, due to effects of demography and drift, while alleles under selection will vary in a consistent way, in correlation with an environmental factor. Loci under selection will therefore deviate from the general pattern of differentiation between subpopulations. Using genome-wide genotyping, and screening for loci with deviating allele frequencies, provides, at least in principle, a very strong method to identify genes underlying ecologically relevant traits.

Population genomics combines genome-wide analysis using next-generation sequencing technology with classical concepts from population genetics (Luikart *et al.* 2003). In the last few years the approach has become extremely popular among molecular ecologists (Beaumont and Balding 2004; Storz 2005,; Stinchcombe and Hoekstra 2008). It is also known as *genome scanning*. An important advantage is that it is based on field data, so any pattern found is evidence of natural selection in the wild, not in an artificial laboratory environment.

Variation of allele frequencies across populations is usually described by the F-statistic, introduced by Sewall Wright. If two subpopulations differ in allele frequencies for some locus, there will be a deficiency of heterozygotes for that locus in the population at large. This is known as the *Wahlund effect* in population genetics. A measure of the degree of differentiation can be found by subtracting the heterozygosity within each subpopulation from the heterozygosity that would be observed if the subpopulations jointly were to act as a single random-mating population. The F-statistic, F_{ST} is therefore written as:

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

where H_T is the heterozygosity among individuals when subpopulations are taken together, and H_S is the heterozygosity in each subpopulation (Hartl and Clark 1997). Obviously F_{ST} varies between 0 (no substructure) and 1 (completely separate population structures).

When heterozygosities are taken over many loci, F_{ST} measures the average genetic distance between subpopulations, and so it can also be taken as a measure of coalescence time (Storz 2005); if there is subdivision, the coalescence time derived from loci within subpopulations is smaller than the coalescence time derived from the same loci across the whole population.

For the purpose of population genomics, F_{ST} is not averaged over many loci, but is estimated for each locus separately. In addition, population genomics aims for a large number of loci (e.g. obtained from a SNP genotyping screen) that can be arranged in a physical map of the genome. What results is a plot of F_{ST} as a function of genomic position (hence genome scan), or as a function of expected heterozygosity (Namroud *et al.* 2008, Fig. 6.6). The F_{ST}/heterozygosity graph is most popular because it does not actually require a genome map. In such graphs, the genome-wide background of F_{ST} values reflects the influence of neutral processes such as drift and migration, while the outliers indicate selection.

Outlier loci are defined by having an F_{ST}-value that is significantly above the genomic average. These loci are candidate sites for *diversifying (directional) natural selection*. Values of F_{ST} below the genomic average are taken as evidence for *balancing selection*. We illustrate the outlier identification with an example from Namroud *et al.* (2008). Six natural populations of white spruce (*Picea glauca*), occurring in contrasting ecological habitats in Québec, Canada, were genotyped for 534 SNPs. These SNPs were identified in an EST library, so they represent polymorphisms in close linkage with protein-coding genes. Testing for outliers identified 20 loci with deviant F_{ST} values (Fig. 6.6). The genes associated with the SNPs were classified to five putative biological functional categories: growth, reproduction, abiotic and biotic stress, and wood formation. So the growth of white spruce in Canada involves significant local adaptation to climatic conditions for genes related to growth and stress resistance (Namroud *et al.* 2008).

Although genome scans for differential selection are usually based on F_{ST}, there are other measures of

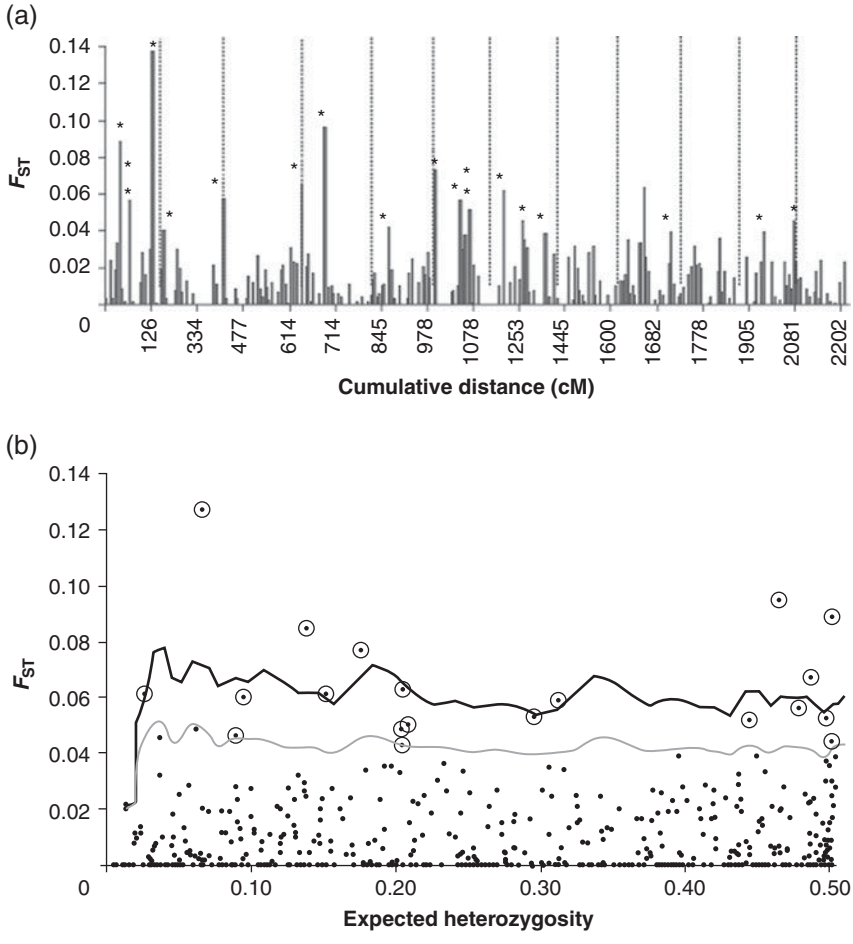


Figure 6.6 Distribution of F_{ST} values derived from 534 SNP polymorphisms over six natural populations of white spruce, *Picea glauca*, in Québec, Canada. The F_{ST}s are plotted as a function of genomic position, where the 12 linkage groups are separated by dotted lines (a), and as a function of expected heterozygosity (b). Statistical tests were conducted to identify 20 outlier SNPs; these are indicated by asterisks in (a) and by circles in (b). In Fig. (b), the grey line indicates the upper 95% confidence interval and the solid line the upper 99% confidence interval of the neutral model. After Namroud *et al.* (2008) by permission of Blackwell Publishing.

population substructure at hand. Holsinger and Weir (2009) discuss four more. One of these statistics, R_{ST}, is sensitive to the mutational distance between alleles. This applies to microsatellites, because under the stepwise mutational model for microsatellite evolution alleles increase or decrease in length by one repeat unit at a time. R_{ST} includes the squared difference between repeat numbers and so genetic variation between alleles with a large mutational distance between them is weighted heavier than variation for closely related alleles (Holsinger and Weir 2009).

A crucial element of the genome scan procedure is the statistical test for outliers. In fact there are different ways to estimate the distribution of F_{ST} under the null hypothesis, depending on assumptions made. In general, the background value of F_{ST} will be determined by a pattern of migration between the subpopulations. If there is little migration, genetic drift will cause the subpopulations to differ from each other and a high, genome-wide, score for F_{ST} will result. Conversely, if migration is very frequent, there will be little differentiation of genetic structure (F_{ST} will approach 0). The most common

approach is to assume a simple island model, in which migration (gene flow) from any subpopulation to another is a constant. Beaumont and Nichols (1996) showed that alternative scenarios in which migration between subpopulations is more complicated (e.g. subpopulations isolated by distance) produced about the same result for the expected distribution of F_{ST}. However, Excoffier *et al.* (2009) argued that a hierarchically structured island model can generate quite different results.

In the model of Excoffier *et al.* (2009) a population is subdivided by islands, but the subpopulations on each island consist of different demes that exchange more migrants with each other than with demes on other islands. The authors also took into account the genetic distances between alleles in a stepwise mutation model for microsatellites (see introductory section of this chapter) and used R_{ST} rather than F_{ST} for microsatellite data. The new model was applied to an earlier study of seven populations of marine and freshwater sticklebacks (*Gasterosteus aculeatus*) by Mäkinen *et al.* (2008). These authors had used 103 microsatellite loci to analyse genetic population subdivision and used F_{ST} to identify loci under directional and balancing selection. They found three loci to be under directional selection and fifteen loci under balancing selection. An important conclusion from the paper was that balancing selection, rather than directional selection, is the predominant mode of selection in the wild.

However, application of the hierarchical island model to Mäkinen's stickleback data led to a quite different conclusion. Excoffier *et al.* (2009) identified two loci as showing evidence of directional selection and only three as showing balancing selection. The reason why the earlier analysis identified so many loci under balancing selection was due to the fact that these loci have very low F_{ST} values, which is due to a relatively high heterozygosity within a subpopulation, caused by the deme structure. The significance of these loci in the genome scan must be considered a false positive result. Identification of significant outliers appears to be rather sensitive to assumptions made for generating the null distribution. This is an important lesson to take into account when doing genome scans.

The loci used in genome scans (AFLPs, STRs, SNPs) will usually be neutral markers, so if they show indications of selection it is not because of the locus itself, but because of their genetic linkage to other, selectable loci. The chance of directly finding functional loci under selection depends on the likelihood that a selected locus is in linkage disequilibrium with one of the markers, which again depends on the density by which the genome is sampled by markers. In addition, the likelihood of finding an association depends on the genomic distance over which linkage disequilibrium is maintained, that is, the recombination rate around the selected locus. If the recombination rate is low, there may be a considerable stretch of DNA on which markers can 'hitch-hike' with a locus under selection.

The extent of linkage disequilibrium in a genome depends on many ecological factors such as the reproductive strategy, rate of outbreeding, and demographic history of the population, so it is difficult to provide any general guidance. In a genome scan for resistance loci in the malaria parasite, *Plasmodium falciparum*, Storz (2005) concluded that a mapping density of one marker per 50 kbp would be sufficient. In the wild yeast, *Saccharomyces paradoxus*, linkage disequilibrium was found to extend over a length of 25 kbp for a European population and 50 kbp for a population from the far east; in baker's yeast, *S. cerevisiae*, recombination between strains is more frequent and linkage disequilibrium tends to fall to a background value within 10 kbp to 30 kbp (Tsai *et al.* 2008; Liti *et al.* 2009; Schacherer *et al.* 2009). However, in some other species, linkage may cover much larger segments of DNA. Figure 6.7 gives the relationship between linkage disequilibrium and genomic distance for a population of Scandinavian wolves (Hagenblad *et al.* 2009). These authors found a very high degree of linkage disequilibrium across 250 microsatellite loci (mean D' = 0.219 for the loci across different chromosomes), while within a chromosome, linkage disequilibrium between loci extended to distances of 50 Mbp, longer than any other outbreeding species. This may relate to a severe founder effect followed by inbreeding; all Scandinavian wolves descend from a single breeding pair that in 1970 migrated into

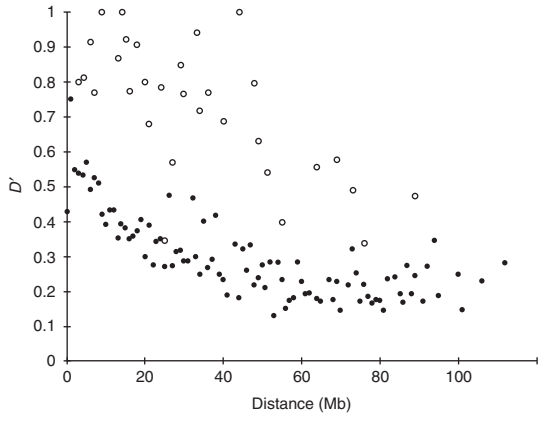


Figure 6.7 Linkage disequilibium (measured by D') as a function of physical distance between loci on the same chromosome. The filled circles are for autosomes, the open circles for the X chromosome. Note that D' is higher for the X-chromosome than for the autosomes due to the fact that the X-chromosome only recombines in females. The background level of linkage disequilibrium (between autosomes) is 0.219, which within a chromosome is reached after about 50 Mbp. From Hagenblad *et al.* (2009), by permission from Blackwell Publishing.

Scandinavia from the Finnish–Russian population, plus a third male founder whose genes appeared in the population in 1991.

One strategy to increase the likelihood that mapping markers for a genome scan will be positioned close to genes under selection is to focus marker development on the transcriptome, rather than on the genome itself. The 3′ and 5′ untranslated regions of expressed sequence tags (ESTs) often contain a significant number of polymorphisms, including microsatellites. If a marker in a 3′ UTR has a significant outlier value for F_{ST}, one can directly identify the gene and attach a functional significance to the selective pressure. This strategy, introduced by Vasemägi *et al.* (2005), was also applied by Oetjen and Reusch (2007) and in the study by Namroud *et al.* (2008) discussed above.

Since 2004, a number of genome scanning studies have been done to reveal loci under selection in wild populations. Table 6.1 provides an overview of these studies. While AFLP markers and SSRs (microsatellites) are still popular markers for genome scans, the future probably lies with SNPs, either EST-associated or genomic, especially since next-generation sequencing and SNP-detection technologies can deal with hundreds of thousands of such loci in a genome.

Table 6.1 shows that current genome scans find between 1 to 10% of the loci bearing a signature of selection. Most likely these estimates are on the high side and may be biased by false positives, especially in the older studies. Somewhat higher percentages are found in studies focusing on EST-associated SNPs. The greatest power is with those cases where the genomic background for F_{ST} is low, that is, in situations where populations are lowly differentiated in general, but still strongly selected locally for some specific trait. Under these conditions outlier F_{ST} values stand out clearly.

It is also obvious from the studies summarized in Table 6.1 that hardly any one of them has, as yet, established a causal link between the loci under selection and some environmental factor acting as a selective force. A lot of more painstaking work, involving a combination of genetics, functional genomics, physiology, and ecology will be needed to find out how the outlier loci identified in a genome scan contribute to differential fitness and local adaptation. What will also help is to correlate the allele frequencies in different subpopulations with environmental factors. For example, Poncet *et al.* (2010) identified 78 loci possibly under selection in a large AFLP screen conducted for more than 200 subpopulations of the alpine plant *Arabis alpina*. By correlating allele frequencies with environmental variables they were able to deduce that mean annual minimum temperature was the factor with the greatest effect on genetic population structure of this plant.

Wood *et al.* (2008) outlined a possible molecular strategy for following up a genome scan. A study in the periwinkle *Littorina saxatilis* had identified a number of AFLP loci contrasting between two ecotypes, one a thick-shelled morph resistant to crab predation, another a thin-shelled morph occupying shores exposed to strong wave action but lacking in crabs. The differential AFLP loci were used to develop probes for screening a genomic BAC library, followed by sequencing the positive BACs. This provided the genomic context of the differential markers, but unfortunately all AFLPs turned out to

Table 6.1 Overview of recent population genomics studies applying genome scans to identify loci under selection in wild populations

Species	Markers	Sign. loci identified	Ecological context	Ref.
Lake whitefish	440 AFLPs	8–14	Benthic (normal) versus limnetic (dwarf) ecotypes	1,2
Atlantic salmon	78 EST-associated SSRs, 17 genomic SSRs	25	Geographic distribution in Northern Europe, salinity	3
Norway spruce	150 AFLPs, genomic SSRs and EST-associated SSRs	9	Geographic differentiation among European populations	4
Eastern oyster	215 AFLPs	3	Geographic cline	5
Common frog	392 AFLPs	8	Altitudinal gradient	6
Eel grass	14 EST-associated SSRs, 11 genomic SSRs	3	Tidal flat versus tidal creek habitats	7
White spruce	543 EST-associated SNPs	47	Geography, temperature, aridity	8
Hawk moth-pollinated violet	369 AFLPs	9	Divergence of floral morphology	9
Honey bee	444 SNPs	5	Expansion from Africa, continental distribution	10
Atlantic cod	318 EST-associated SNPs	26	Geographic cline, temperature	11
Threespined stickleback	103 SSRs	20	Freshwater versus marine environments	12
Killifish	300 AFLPs	24	Adaptation to organic pollution	13
Scandinavian wolf	250 SSRs	21	Comparison with Russian wolves, founder effects	14
Large pine weevil	83 AFLPs	10	Geography and abiotic variables versus host plant	15
Aedes aegypti mosquito	500 Miniature Repeat Transposable Element-derived markers	5	Use of Bt insecticide, development of resistance	16
Aedes rusticus mosquito	155 AFLPs	5	Use of Bt-insecticide, development of resistance	17
African rainforest lizard	191 AFLPs	13	Lowland rainforest to montane forest and ecotone gradient	18
Three-spined stickleback	45.000 SNPs from Restriction-Associated DNA tags	9	Adaptation to freshwater habitats	19

References: 1: Campbell and Bernatchez (2004); 2: Rogers and Bernatchez (2005); 3: Vasemägi *et al.* (2005); 4: Acheré *et al.* (2005); 5: Murray and Hare (2006); 6: Bonin *et al.* (2006); 7: Oetjen and Reusch (2007); 8: Namroud *et al.* (2008); 9: Herrera and Bazaga (2008); 10: Zayed and Whitfield (2008); 11: Moen *et al.* (2008); 12: Mäkinen *et al.* (2008); 13: Williams and Oleksiak (2008); 14: Hagenblad *et al.* (2009); 15: Manel *et al.* (2009); 16: Bonin *et al.* (2009); 17: Paris *et al.* (2010); 18: Freedman *et al.* (2010); 19: Hohenlohe *et al.* (2010).

be located in non-coding regions. A functional interpretation of the differential markers is still difficult, even though there was strong evidence for local adaptation in this study.

In summary we can conclude that population genomics is a potentially strong approach, forging a synthesis between ecological and molecular approaches to evolutionary biology. It allows us to identify loci associated with fitness differences in natural populations (Ellegren and Sheldon 2008). Up to now, however, the studies have been limited by their inability to uncover the functional significance of genome segments bearing a signature of selection (Stinchcombe and Hoekstra 2008). The studies reviewed above teach us that the effectiveness of a genome scan depends on:

1. The *genetic architecture* of the locus, especially the extent of linkage disequilibrium and the rate at which this declines to background with increasing genomic distance; if the rate of recombination is high, the hitch-hiking effect around a locus will be small and loci under selection will be difficult to catch with random markers.
2. The *intensity of selection*. Strong selection, either locally varying or contrasting between phenotypic variants is an obvious requirement for identifying genes under selective pressure.
3. *Population structure*. The genome-wide background for differentiation is determined by the way in which the population is subdivided in local units, each subjected to drift but connected by migration. In the case of unidentified subdivision, for example due to deme structure within an island, it may be difficult to detect outlier F_{ST}s, due to a high threshold value.

6.2.4 Sequence-based population genomics

The population genomics approaches discussed in Section 6.2.3 rely on patterns of allele frequencies within and across populations, but do not usually utilize the sequences themselves because the genes involved are not yet known. Another approach of population genomics is based on analysis of DNA sequences rather than frequencies of markers (see, e.g., Nielsen 2005; Li *et al.* 2008). We designate this approach *sequence-based population genomics*. Li *et al.* (2008) called it *reverse ecology*. While marker-based population genomics starts with allele frequencies observed in the field and attempts to identify the underlying genes, sequence-based population genomics takes the reverse route: it starts with DNA sequences and attempts to draw conclusions about the ecology that has shaped these sequences.

Population genomics of DNA sequences is enhanced greatly by the technical possibilities of next-generation sequencing methods, allowing sequence analysis of not one but several genomes from a single species. The analysis requires advanced bioinformatics and statistics techniques due to the very large datasets involved and the great variety of assumptions that can be made regarding the underlying genetics. A complete treatment of the models and tests involved falls beyond the scope of this book; the reader is referred to Graur and Li (2000) and other specialist books of comparative genome analysis. Here we will only focus on two of the most common procedures that test for deviations from neutrality, the McDonald-Kreitman test and Tajima's D statistic.

Like with the marker-based (forward) approach, an important starting point for analysis is the assumption of neutral evolution. Tests demonstrating deviations from neutrality are then interpreted as a signature of selection. One way to test for non-neutrality in coding DNA is to compare polymorphisms (divergences) that are fixed within a population, but variable across populations, and polymorphisms that are present both within and between populations. Under the neutrality assumption we expect that the ratio between these two categories is the same for synonymous polymorphisms as nonsynonymous polymorphisms. If however the nonsynonomous mutations are over-represented in the fixed category, this indicates that one or more of the populations are influenced by selection.

A popular statistical test to evaluate the assumption of neutrality across populations or species is the *McDonald–Kreitman test*. It uses a simple 2 x 2 table in which counts are made of nonsynonymous and synonymous substitutions classified according to whether they are fixed in one of the populations or polymorphic within populations. Under the null hypothesis, there is independence between rows and columns in such a table (McDonald and Kreitman 1991). From the table one can also derive

a quantitative measure of positive selection, the so-called *A/S ratio*: the ratio between amino acid to synonymous substitutions (Fay *et al.* 2002).

Unfortunately, results of the McDonald–Kreitman test are affected by the presence of polymorphisms due to slightly deleterious mutations. These are mutations upon which purifying selection acts only weakly, and so the polymorphism is not only determined by selection, but also by random genetic drift. This effect is not easily removed by deleting rare variants (Charlesworth and Eyre-Walker 2008). Shapiro *et al.* (2007) noted that adjustments are necessary when the McDonald–Kreitman test is applied to frequency tables of data that are pooled over loci with different degrees of polymorphism.

Another popular method to test for non-neutrality uses a quantity known as *Tajima's D statistic*. Unlike the McDonald–Kreitman test, this method is applicable to coding and non-coding DNA alike, because it does not rely on synonymity of substitutions, but on the distribution of polymorphisms across a series of homologous DNA sequences. The argument goes as follows.

Suppose we have sequenced a homologous segment of DNA in a number of individuals. Due to substitutions there are various alleles for that segment. If the mutation rate is low most of the sites will be monomorphic. Also, if the DNA sequence is sufficiently long, the total number of sites available for mutation is so large that any new mutation appears at a site that was previously monomorphic. The DNA segment is considered as a long series of unlinked nucleotides where each nucleotide has a low probability of mutation and becoming a SNP. This situation approaches the assumption of the *infinite sites model* developed by Kimura (1983). Using this model it is possible to estimate the expected total number of polymorphic sites in the sample. This quantity is called θ_W.

Another model, called the *infinite alleles model*, considers the mutation process in a slightly different way. In this model every new mutation is said to create a new allele. Any DNA segment of a certain length has a very large number of possible alleles, for example a segment of 100 bp has 4^{100} possible alleles. Using this model it is possible to estimate the expected number of pairwise differences: if we draw two alleles at random from the population, we note how often they are different from each other. This quantity is equal to the *average heterozygosity*, because in a panmictic population heterozygotes can be viewed as having two randomly drawn alleles that are different. Average heterozygosity is designated θ_T.

According to Tajima (1989), the two quantities θ_T and θ_W estimate the same thing: $\theta = 4N_e\mu$ (four times the product of effective population size N_e and the mutation rate, μ), when the population is evolving neutrally. So under the assumption of neutrality the two quantities should be equal. This suggests that their difference, called Tajima's D statistic, can be used as a measure of departure from neutrality:

$$D = \hat{\theta}_T - \hat{\theta}_W$$

where the 'hats' indicate that these are estimators of the parameters, rather than the parameters themselves. One way to estimate Tajima's D is to equate θ_T to the average number of different nucleotides between all possible pairs in the sample, and θ_W to the total number of segregating (polymorphic) sites in the sample. Based on computer simulations, Tajima (1989) argued that the distribution of D under the null hypothesis of neutrality approached a beta distribution. If D calculated for a sample falls outside the critical value derived from this distribution, the null hypothesis of neutral evolution is rejected. Later, an improved method for constructing critical values was introduced (Simonsen *et al.* 1995) and this study confirmed that Tajima's D (with the new critical values) is the most powerful test among a number of other tests within this class.

If Tajima's D is significantly negative, two situations may apply, positive or negative selection. A negative D alone cannot discriminate between these two cases, but additional information from another statistic, known as *Fay and Wu's H statistic* can. H measures departures from neutrality that are reflected in the difference between high-frequency and low-frequency alleles. If the polymorphisms are skewed towards an excess of mutations segregating at high frequencies (and a lack of polymorphisms in the low frequency category), then $H < 0$. This implies *positive selection*, that is selection favouring new advantageous mutations. If there is an excess of polymorphisms at low frequencies,

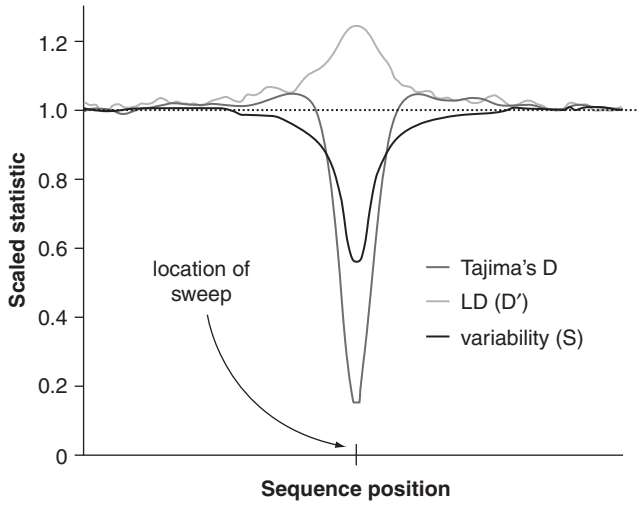


Figure 6.8 Simulation illlustrating the effect of strong positive selection (selective sweep) on the genetic variation of DNA sequences around a selected locus: strongly negative Tajima's D, increased linkage disequilibium (LD, measured by D'), and low overall polymorphism (S). From Nielsen (2005) by permission from Annual Reviews.

then H > 0, and this indicates *negative (purifying) selection*, that is selection against new deleterious mutations. Finally, if Tajima's D is significantly positive, there is an excess of polymorphisms both in the low and the high frequency categories, compared to neutral expectations, this indicates *balancing selection*.

Strong positive selection acting upon a newly derived mutation is also designated as a *selective sweep*. Figure 6.8 indicates the local consequences of a selective sweep in the genome: strongly negative Tajima's D, increased linkage disequilibrium (hitchhiking of neighbouring genomic segments), and low overall polymorphism.

Several studies in *Drosophila* have shown that positive and negative selection are an important aspect of the polymorphism structure (e.g. Andolfatto 2005; Shapiro *et al.* 2007; Haddrill *et al.* 2008). Shapiro *et al.* (2007) studied sequence diversity of 419 genes in 21 lines of *Drosophila melanogaster*. The genes were classified as to their position in the genome: that is whether they were

Table 6.2 Estimates of nucleotide diversity of 419 autosomal loci in 21 *Drosophila melanogaster* lines, separated by recombination rate. From Shapiro *et al.* (2007) by permission of the National Academy of Sciences of the USA

Genomic site	Tajima's D statistic	Fay and Wu's H statistic
Normal recombination, 252 genes		
Nonsynonymous positions	–0.585**	0.320**
Synonymous positions	–0.134	–0.311*
Non-coding positions	–0.147*	–0.158
Low recombination, 167 genes		
Nonsynonymous positions	–0.468**	0.188*
Synonymous positions	–0.288*	–0.364
Non-coding positions	–0.211	–0.120

**$p < 10^{-8}$, *$p < 10^{-2}$

located in chromosomal regions of normal or low recombination. Tajima's D and Fay and Hu's H were calculated for nonsynonymous, synonymous and non-coding positions (Table 6.2).

The data by Shapiro *et al.* (2007) reveal the presence of strong negative (purifying) selection in nonsynonymous coding positions, both in genomic areas of normal recombination and in areas of low recombination ($D < 0$, $H > 0$). The silent and non-coding sites tend to have both negative D and H, and so bear a signature of positive selection. That selection may act upon non-coding regions is confirmed by other *Drosophila* studies, for example Andolfatto (2005) and Haddrill *et al.* (2008), however, the details differ between *D. melanogaster* and *D. simulans*. In *D. simulans*, all non-coding DNA is subject to negative selection, while in *D. melanogaster* it is mainly positive selection that operates on non-coding DNA (Table 6.2). The differences between these closely related species may relate to the fact that *D. melanogaster* has experienced a reduction in effective population size, which obscures the effect of selection.

In summary, the sequence-based population genomic studies of *Drosophila* indicate that selection has played a major role in the recent evolution of this genus. In addition, a substantial fraction of the polymorphisms were driven to fixation in one or more of the sister species. This abundance of comparative genomics evidence supporting selection in fruit flies is somewhat at variance with the situation in yeast, to be discussed later in this chapter (Section 6.3.4). It is also not known at the moment whether the conclusions on selection acting upon the genome of a species are contingent on specific life histories or demographic events in the past, which may differ between the species. More comparative data are needed to shed light on these issues.

6.3 Regulatory and structural change

A fundamental question of molecular evolution is how much phenotypic change is due to changes in the structure of proteins and how much is due to altered expression of these proteins, that is, what is the locus of evolution? There is a divergence of opinions on this issue. Evolutionary developmental biology, also called evo-devo, tends to place a strong emphasis on regulatory change as the prime mechanism for evolutionary changes in body plan, while physiologists tend to emphasize the structural changes in proteins as the prime mechanism of adaptation. We will discuss examples of both in this section.

As an illustration of the importance of structural change we may recall the textbook example of temperature adaptation in the enzyme lactate dehydrogenase (LDH) (Hill *et al.* 2008). Lactate dehydrogenase is a key enzyme of energy metabolism. In muscles, it converts the product of glycolysis, pyruvic acid, to lactate when the citric acid cycle and the electron transport chain are inhibited by low availability of oxygen. In this way the organism can produce ATP from anaerobic glycolysis during energy-demanding events of short duration. Conversely, LDH catalyses the reverse reaction of lactate to pyruvate when oxygen availability increases again. The action of the enzyme depends critically on the affinity to its substrate, which is measured by the reciprocal Michaelis–Menten constant, $1/K_m$. This binding affinity normally increases with decreasing temperature, which would pose a problem for organisms that live in cold environments because molecular flexibility is necessary to maintain an appropriate catalytic rate. Without temperature adaptation, the catalytic rate of the enzyme would be saturated at very low substrate concentrations.

To understand how adaptation has taken place we must know that among vertebrates three different isoforms of LDH exist, of which the A-isoform is mainly expressed in muscle tissue, while two others are expressed in the heart and in the testis. The muscle variant of the enzyme consists of four identical peptides all encoded by a gene called *LDHA*, and the tetrameric active enzyme is designated LDH-A_4.

Measurements have been done of LDH-A_4 affinity to pyruvate in different species of fish, living in different thermal environments, varying from warm-water gobies, living at water temperatures up to 35 ºC, to Antarctic fish, living at water temperatures of a constant –1.9 ºC. Because fish are cold-blooded animals their tissues will also be at these temperatures or only slightly above. For each

species of fish the biophysical laws of molecular kinetics determine that $1/K_m$ increases with decreasing temperature. However, the cold-adapted fish have a much lower substrate binding affinity than the warm-adapted fish when they are measured at the same temperature (as long as they both survive at this temperature). Consequently, when $1/K_m$ of each fish is measured in its own ecological niche, that is, the temperature range of its normal habitat, the substrate binding affinities are more or less the same (Hill *et al.* 2008). This strong temperature compensation effect is brought about by the fact that each fish has evolved a different molecular variant of *LDHA*.

The *LDHA* locus shows several synonymous and nonsynonymous substitutions between different species of fish (Fig. 6.9). Detailed protein analyses have been done to determine which substitutions are critical for the temperature compensation effect shown by cold-adapted fish (Fields and Houseman 2004). The research was focused on Antarctic icefish, Notothenioidei, a suborder of the Perciformes, which contains 122 species endemic to Antarctic waters, all living constantly at a water temperature 1–2 ºC below zero.

Using recombinantly expressed proteins and site-directed mutagenesis, Fields and Houseman (2004) were able to show that a single amino acid substitution is sufficient to alter the substrate affinity of the notothenioid LDH-A_4 molecule to a level similar to the one in warm-adapted fish. Interestingly, this substitution is located rather distantly from the active centre; it is in one of the alpha helices of the protein that move during the docking of pyruvate in the catalytic centre (Fig. 6.9). This substitution, possibly in concert with others, is responsible for the remarkable properties of the LDH-A_4 adaptation in Antarctic fish.

This classical work on LDH, and many other studies of molecular physiology, suggests that physiological adaptations tend to rely on structural changes in proteins. This is contrasted with adaptations in development that would mainly require regulatory changes. The latter option is inspired by the widely held view among evolutionary zoologists, that the macroevolutionary changes seen in the animal kingdom are due to changes in the regulation of developmentally important genes. For example, Valentine (2004, p. 77) argues that 'it is evident that the evolution of regulatory gene systems, rather than of structural alleles, has been chiefly responsible for the sorts of morphological innovations revealed by the fossil record'. Similarly, Carroll *et al.* (2005) explain that developmental pathways are regulated by a relatively small number of genes, which encode transcription factors and components of signalling pathways; these genes are designated the 'toolkit' of development and are

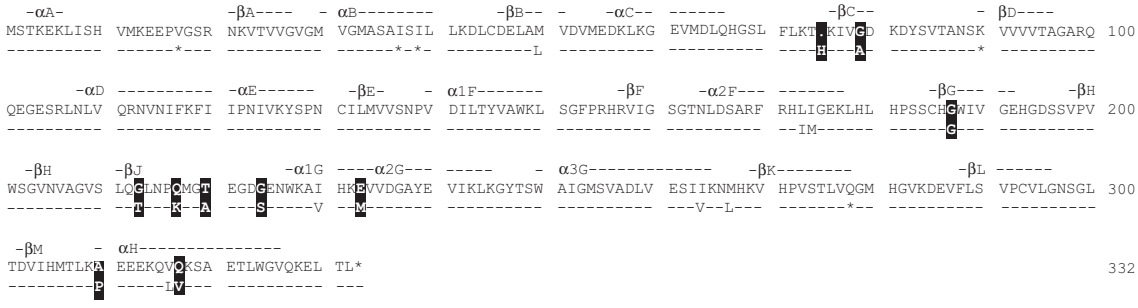


Figure 6.9 Alignment of LDH-A amino acid sequences from cold-adapted Antarctic fish (middle line) with LDH-A from temperate fish (lower line; dashes indicate agreement with Antarctic fish). The top line shows various aspects of protein structure. The sequence of Antarctic fish is a consensus of nine different notothenioid species, the sequence for temperate fish is a consensus of six different teleosts. Asterisks indicate sites where there is agreement between temperate and cold-adapted sequences for some species but not all. Amino acids substitutions that are likely to have functional consequences are highlighted. Research has shown that the substitution at position 233 in the α1G-2G domain (methionein, M, to glutamic acid, E) is crucial in the cold-adaptation of LDH-A_4. From Fields and Houseman (2004), by permission of the Society of Molecular Biology and Evolution.

widely conserved among different animal taxa. Obviously, if the toolkit proteins stay the same, morphological change must come from alterations in their expression, for example tissue specificity or time dependence. Carroll *et al.* (2005, p. 231) therefore state that there is '. . . persuasive evidence that the diversification of regulatory DNA, while generally maintaining coding function, is the most available and most frequently exploited mode of genetic diversification in animal evolution.'

To support the importance of regulatory evolution three arguments are often given:

1. The non-coding nature of regulatory DNA (the absence of open reading frames) allows more freedom for mutations without immediate deleterious effects.
2. Regulatory elements often come in the form of modules which can evolve independently from each other, creating a rich source of variation.
3. Combination of different regulatory elements in a single promoter allows evolution of developmental novelties by new combinations of gene regulation.

The view that developmental adaptation is mainly due to regulatory change was challenged by Hoekstra and Coyne (2007). They argued strongly against the 'theoretical imperative of evo-devo', as formulated above. In this section we will discuss various examples that contribute to the ongoing discussion about the importance of structural versus regulatory variation in the genome.

6.3.1 Evolutionary changes of gene regulation in *cis* and *trans*

An important, although rather crude, classification of gene regulation mechanisms considers the site in the DNA where regulatory factors exert their action, relative to the coding sequence of the gene considered. Regulatory proteins that bind to sites close to the transcription initiation site are said to act as *cis*-regulatory factors. Usually this concerns the so-called proximal promoter in the 5′ region of the gene, up to around 1500 bp upstream of the initiation site. However, there is no general agreement about the length of DNA that can still be called the *cis*-regulatory region of a gene.

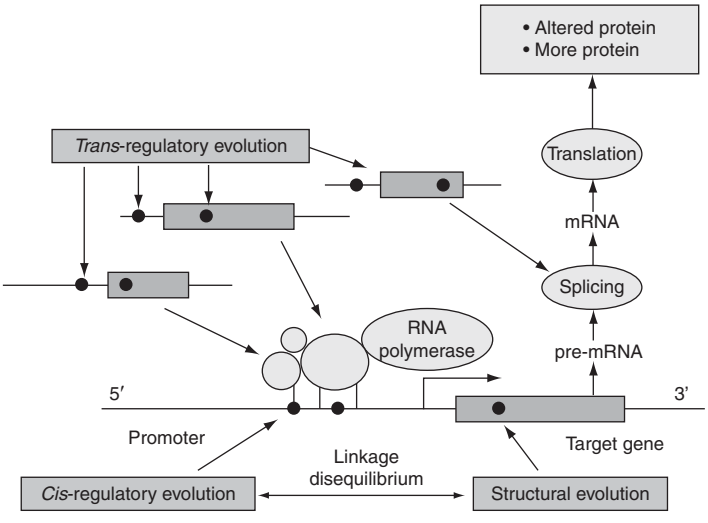


Figure 6.10 Schematic view of the possible evolutionary changes affecting a single target gene. Mutations are indicated by dots in the sequence. They may be located in the coding sequence of the target gene (*structural evolution*) or elsewhere (*regulatory evolution*). Mutations in the proximal promoter may change the action of regulatory factors (*cis-regulatory evolution*). Mutations in genes encoding the regulatory factors themselves, as well as mutations in the promoters of these factors, are considered as *trans*-regulatory change with respect to the target gene.

Evolutionary changes in *cis*-regulatory gene expression concern substitutions in transcription factor binding sites (TFBSs), insertions and deletions altering the spacing of TFBSs, recombinatory events creating new combinations, and so on. Evolutionary changes in *trans*-regulation concern the structure of a regulatory factor or the expression of such factors (Fig. 6.10). So there is an enormous variety of mechanisms by which gene expression can be changed by *trans*-acting factors. Also, whether a factor acts in *cis* or in *trans* should always be viewed relative to a target gene; for example, changes in *cis*-regulation of a gene encoding a DNA-binding regulatory protein are *trans*-acting when considered relative to the target gene.

The question of how much evolutionary change is due to altered *cis*-regulation, and how much to *trans*-regulatory divergence has received quite some attention in the recent literature (Wittkopp *et al.* 2004, 2008a,b; Ranz and Machado 2006; Graze *et al.* 2009; Gagneur *et al.* 2009). Wittkopp *et al.* (2004) introduced a new approach to address the question: allele-specific expression profiling in hybrids. This strategy compares the expression of genes in two different homozygous lines, with the expression of the two alleles of the same gene in a hybrid. If the difference in gene expression between the two homozygotes is only due to *cis*-acting elements, the ratio between the allele-specific expressions in the hybrid should be the same as the ratio between the expressions of the same gene in the two homozygotes. The condition for pure *cis*-regulatory divergence is therefore:

$$\frac{E_A(AB)}{E_B(AB)} = \frac{E_A(AA)}{E_B(BB)}$$

where $E_A(AB)$ and $E_B(AB)$ are the expressions of alleles A and B in the heterozygote (hybrid), $E_A(AA)$ is the expression of allele A in homozygote AA, and $E_B(BB)$ the expression of allele B in homozygote B.

Conversely, if the two alleles are expressed equally in the hybrid, while they are expressed differentially in the homozygotic lines, the difference between the homozygotic lines must be due to genetic factors acting in *trans* upon the gene of interest. Consequently, there will be no correlation between the arguments on either side of the equation above. So the regulatory divergence can be attributed 100% to *trans*-acting factors if:

$$\frac{E_A(AB)}{E_B(AB)} = 1$$

for all values of $E_A(AA)/E_B(BB)$.

In a study by Wittkopp *et al.* (2004), the technology of pyrosequencing provided an opportunity to discriminate between transcripts coming from one or the other allele. A crucial point of the design is that the two alleles differ at least in a single diagnostic nucleotide, which will generally be the case if we study two interbreeding species. A PCR is then designed that amplifies a segment around the divergent site, while pyrosequencing is used to characterize the amplicons. Pyrosequencing produces a signal that is proportional to the quantity of nucleotide added to the extending primer (cf. Chapter 1), so the polymorphic site can be recognized in the pyrogram and the relative amount of either transcript estimated. Next-generation sequencing methods now allow such local analyses to be upscaled to genome-wide screens of allele-specific expression (Pastinen 2010).

Another approach to allele-specific expression analysis, applied by Graze *et al.* (2009) and Gagneur *et al.* (2009) is to use microarrays. In this case the investigator should specify which probes have an exact match to both strains and which probes match to one or the other strain at polymorphic sites. This approach is most profitable when using a tiling array, because in this case maximal use is made of polymorphic positions scattered in the genome.

Wittkopp *et al.* (2004) initially applied the technique to hybrids between closely related species of fruit fly, *Drosophila melanogaster* and *D. simulans*. These species can produce viable offspring (although sterile) and so provide a convenient model to distinguish *cis* and *trans*-regulatory differences between species. One of the main results of their study is reproduced in Fig. 6.11.

There was a good correlation between expression ratios of species and hybrids in almost half of the cases (12 out of 28 genes on the diagonal in

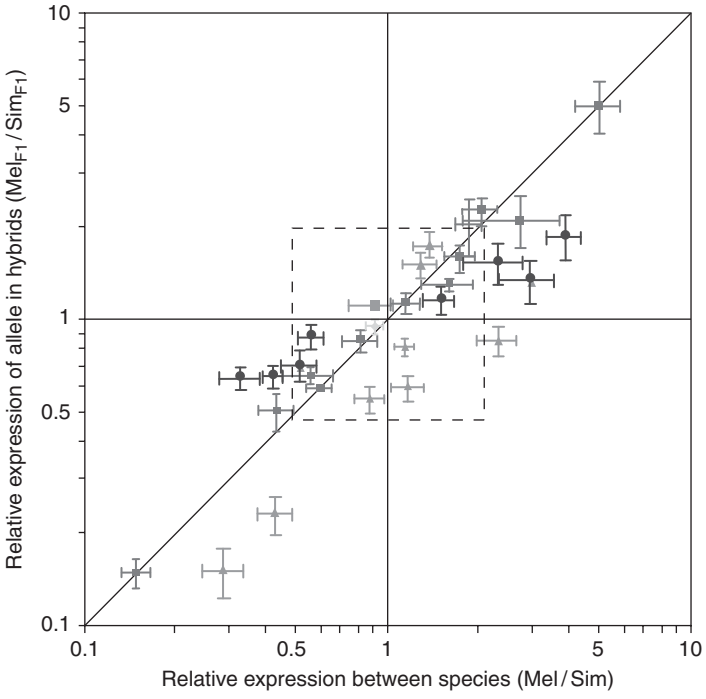


Figure 6.11 Correlation diagram showing differential expression of 28 different genes between two closely related species of *Drosophila* (*D. melanogaster*, abbreviated Mel, and *D. simulans*, Sim) on the horizontal axis, against the expression ratio of the Mel and Sim alleles in a hybrid between the two species. If differential expression is completely due to differences in *cis*-regulation, all the points should be on the diagonal. In the case of 100% *trans*-regulatory divergence, all the genes should be on the middle horizontal line. From Wittkopp *et al.* (2004), by permission of Nature Publishing Group.

Fig. 6.11), which indicates 100% *cis*-regulatory divergence for these genes, while none of the genes showed a complete *trans*-regulatory divergence (none on the middle horizontal line in Fig. 6.11). Of the sixteen genes with both *cis* and *trans*-regulatory divergence eight had a lower and eight a higher divergence than expected. So these results allow the conclusion that *cis*-regulatory divergence of gene expression is quite important in the evolutionary split between species, and it is accompanied by additional changes in *trans*-regulation in about 50% of the cases.

In further work, the same methodology was applied to four strains of *D. melanogaster* and three strains of *D. simulans*, allowing interspecific divergence to be compared with intraspecific divergence. Interestingly, the data for comparisons between strains of the same species were less compliant with the model of *cis*-regulatory divergence than the data for the interspecific comparison (Wittkopp *et al.* 2008a). These conclusions, if they can be generalized, suggest that genetic variation of gene expression between individuals of the same species is mostly due to divergence in *trans*-acting factors, while the evolutionary split between two species tends to amplify differences in *cis*-regulation. This would be partly in line with the evo-devo point of view discussed above.

The type of regulatory divergence may of course depend on the class of gene. In another *Drosophila* study it was shown that *cis*-regulatory effects are often found for genes with a large degree of heritable expression variation (Hughes *et al.* 2006). Evolutionary changes of *cis*-regulation tend to cause large and variable effects on gene expression. This is in line with the *Arabidopsis* studies cited in Section 6.2.2, which showed that loci regulated by *trans*-acting factors tend to concern small and additive phenotypic

effects. In addition, Hughes *et al.* (2006) showed that genetically variable expression is not equally found for genes of all functional categories. Among genes annotated for biological process, underrepresented GO terms were: *development*, *cell communication*, and *regulation of biological process*, and among genes annotated for molecular function categories underrepresented were *signal transduction*, and *transcriptional regulation*, while *catalytic activity* was overrepresented. This suggests that genes encoding catalytic or general metabolic functions are often under diversifying selection, maintaining the genetic variation in their expression. Indeed, frequency-dependent selection could be particularly important for functions dealing with responses to variable environments. On the other hand, genes encoding functions in development are expected to experience strong purifying selection because small deviations in transcript abundance will disrupt the high level of integration in developmental pathways.

6.3.2 Promoter architecture and evolution

A eukaryotic promoter is the stretch of DNA located upstream of a transcription initiation site; it influences the transcription of a downstream coding sequence. There is no general agreement as to how far a promoter extends in the 5′ region. Often a distinction is made between the *proximal promoter*, roughly extending to around 550 bp upstream, and the *distal promoter*, which has no clearly defined ending. In some cases binding sites located 1 Mbp from the initiation site still contribute to transcriptional regulation of the gene. However, the first 550 bp usually possess all activity necessary for basal expression.

The *core promoter* is the region, 200 bp directly upstream of the transcription start site, where the transcription complex is assembled. It is characterized by *basal promoter elements* such as the TATA box, the initiator element (Inr), and the downstream promoter element (DPE). Although earlier studies suggested that the core promoter had highly conserved structural elements, we know now that there is a tremendous variability. Not all core promoter elements are found in all genes and in all species. A lot is known about the TATA box, but this element turns out to be present in only 10–20% of the genes in most species.

The complex of proteins associated with the core promoter region is called *pre-initiation transcription complex* (*PITC*); when binding to a TATA-box it consists of TATA-binding protein (TBP), TBP-associated factors, and RNA polymerase II. However, the PITC on its own can provide only a low level of transcription due to instability of its association with DNA. Without activation only a background level of transcripts is produced.

Transcriptional activity is significantly increased by the action of gene-specific transcription factors that bind further upstream to the proximal promoter. These transcription factors have specific recognition sites, known as transcription factor binding sites (TFBSs), which are characterized by conserved DNA motifs, often called *cis*-regulatory elements. We already met several of these elements in the context of stress responses, and a number of consensus sequences are listed in Table 5.1. It is assumed that transcription factors interact directly with the PITC, however in some cases transcription factors need to be trans-activated by phosphorylation before they can do so. Transcription factors stabilize the binding of the PITC to the DNA and so increase the frequency by which RNA polymerase II starts transcribing.

Finally the level of transcription can be further boosted by *enhancers* that bind upstream, sometimes far away from the transcriptional complex. It is assumed that enhancers bind a coactivator protein, which then recruits a histone-modifying enzyme to create a more favourable chromatin environment for transcription. Alternatively, the coactivator may bind a kinase that activates RNA polymerase by phosphorylation. The complete sequence of events for the assembly of a transcription regulatory complex is given in Fig. 6.12 (Farnham 2009).

The most common bioinformatics technique to characterize promoter sequences is *phylogenetic footprinting*. This approach aims to find conserved regulatory motifs by homology searches in genomic databases. The availability of more and more full genome sequences has been very helpful in this respect and many online search algorithms are now available (see Lemos *et al.* 2008). At the same time,

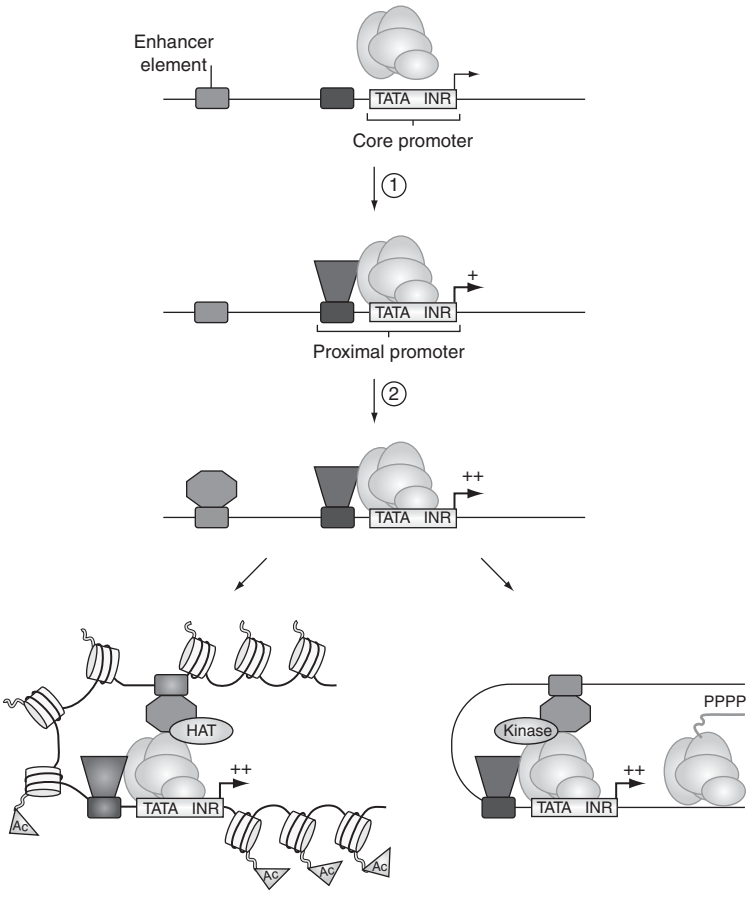


Figure 6.12 Model for eukaryotic transcriptional regulation. Activation of the preinitiation transcription complex requires several steps (1–2). In the third step there are two alternative options: recruitment of a histone-modifying enzyme (HAT), or a kinase. From Farnham (2009) by permission of Nature Publishing Group.

the sequence-based approach to promoter analysis meets severe limitations, especially in plants (for some reason, plant promoters seem to be less well defined than animal promoters). Given the enormous variability of transcriptional regulation, the sequence itself does not provide any clue to its function. It may happen that conserved transcription factor binding sites are found in a promoter, but the transcription factor itself remains elusive. Only functional studies, demonstrating that the binding of a specific protein to a promoter element is necessary for transcription will prove the functionality of a promoter. Popular techniques for demonstrating protein binding to DNA are *electromobility shift assays* (*EMSA*), which are based on the fact that DNAs with proteins bound to them behave differently in electrophoretic gels, and *DNase footprinting*, which is based on the local protection from enzymatic cleavage provided by proteins bound to DNA.

Over the last few years our knowledge of protein–DNA interactions has grown considerably, thanks to the development of techniques such as *chromatin immunoprecipitation* followed by microarray hybridization (*ChIP-Chip*) or by sequencing (*ChIP-Seq*). These technologies (for a short explanation, see Farnham 2009) allow a genome-wide localization of binding sites to be made for any protein of interest. The ENCODE (ENCyclopedia Of DNA Elements) project is a collaborative effort to identify all the functional elements in the human genome

sequence (ENCODE Consortium 2007). Using the new technologies it has been shown that transcription factors bind to an astonishing large number of sites in the genome: thousands of sites may be occupied by transcription factors at any time, in diverse regions of the genome. The binding sites are not limited to genes normally induced by a transcription factor, but include sites far away from the transcription complex, as well as intragenic regions, such as introns, and even exons. The functional significance of this diversity of DNA–protein interactions is not yet clear.

Another strategy to characterize factors important in transcriptional regulation is not to start with the proteins and screen for binding sites in the genome, but to start with a promoter and analyse which factors will bind to that stretch of DNA. For this problem the technique of *yeast one-hybrid* screens is often used. This is a complicated molecular procedure which uses a genetic construct in yeast cells ('bait vector') to screen for interactions between proteins expressed from a cDNA library and a target regulatory DNA element (the bait).

As an example of a successful yeast one-hybrid study, a paper by Assunção *et al.* (2010) is instructive. These authors were interested in the regulation of zinc transporters in plants in zinc-deficient environments. *Arabidopsis thaliana* has four zinc transporter genes, *ZIP1* to 4, where *ZIP4* is highly expressed under zinc deficient conditions. The promoter sequences of these genes did not give any clue as to what proteins might regulate their expression. However, using a yeast one-hybrid assay, Assunção *et al.* (2010) were able to identify two proteins binding to the *ZIP4* promoter, bZIP19 and bZIP23. These are two transcription factors of a family called basic region/leucine zipper motif, which in *Arabidopsis* has 75 members. The bZIP transcription factors generally are involved with pathogen defence and environmental signalling. In Section 5.3.2 we met them in the context of plant abiotic stress responses. Assunção *et al.* (2010) were also able to define a 10-bp DNA sequence to which bZIP19 and bZIP23 can bind, and which is present in genes responding to zinc deficiency. This work illustrates that the technique of yeast one-hybrid can be very successful in identifying regulatory proteins starting from a promoter sequence. However, in non-model organisms the task is much more difficult due to insufficient characterization of the array of possible regulatory proteins and the large number of false positives usually found.

Promoter sequences are assumed to evolve much faster than coding sequences, because they are not limited by an open reading frame. Therefore it is expected that genetic variation in promoters is generally larger than in coding sequences. The question is: will polymorphisms in promoter sequences have functional consequences, that is, will they affect the phenotype and cause differential evolutionary success? Wray *et al.* (2003) listed the following theoretical arguments why this should be so:

1. Changes in the precise regulation of transcription with respect to tissue, age, and time, should have evolutionary consequences, because the timing of gene expression is often crucial (production of a digestive enzyme at the right moment, production of a developmental regulator in the right tissue, etc.).
2. Mutations in the promoters of transcription factor genes can produce highly coordinated pleiotropic effects on the phenotype. For example, increasing the drought resistance of a plant can be done more easily, and with less negative pleiotropy, by altering the expression of a master regulator than by altering many different genes. That is why evolution is expected to act upon polymorphisms in the expression of transcriptional regulators.
3. The fact that developmentally active transcriptional regulators (e.g. *Hox* genes) are so extremely well conserved across the animal kingdom suggests that the expression of these proteins has evolved more than their structure.
4. Promoters are expected to be more evolvable than coding regions; their modular structure allows for a mosaic-type of evolution in which discrete aspects of the overall expression can be changed by new combinations of motifs.

In addition to these theoretical arguments, there is ample evidence now that promoter polymorphisms can have evolutionary consequences. As an example of a detailed study of promoter polymorphisms

in natural populations and their functional significance, we discuss the work of Janssens *et al.* (2007).

Janssens *et al.* (2007) studied the promoter of a gene encoding a metal-binding protein, metallothionein, in a species of soil-living invertebrate, *Orchesella cincta* (Hexapoda: Collembola). Earlier research had shown that there was a large degree of heritable variation ($h^2 = 0.36$) in the inter-individual expression of this protein, and that overexpression contributes to tolerance to cadmium, a toxic heavy metal found in soils of mining sites and around metal-smelting industries. Cloning of the metallothionein promoter (pmt) from animals with different metallothionein expressions revealed a large degree of polymorphism. In total nine different promoter alleles were characterized (Fig. 6.13a). Several putative transcription factor binding sites were identified, their number and relative position differing between the alleles. The similarity between the promoter sequences comes in the form of a mosaic of blocks of a few hundred base-pairs. For example, the alleles *pmtA1* and *pmtA2* share a block of ± 500 bp with *pMtD1* and another block of ± 400 bp with *pmtB*. So A1 and A2 most likely originated by recombination with D1 and B, as indicated in a reticulate phylogenetic network (Fig. 6.13b). Such a network analysis is appropriate when recombination is an important mechanism for generating variation; the phylogenetic reconstruction is different from a normal phylogenetic analysis, because the reticulate network has splits as well as junctions.

Further work, using reporter assays in a *Drosophila* S2 cell line, showed that alleles *pmtD2* and *pmtF* conferred the highest expression when induced by cadmium. The *pmtC* allele, however, was hardly inducible, while the others showed a moderate induction. In field populations six to nine different promoter alleles are commonly found, and so there is significant natural polymorphism. In a survey of 20 sites with varying levels of soil contamination, the frequency of the *pmtD2* allele was significantly (positively) correlated with the cadmium concentration in the soil (Janssens *et al.* 2008). This is in excellent agreement with the fact that this allele confers overexpression of metallothionein, which protects the animal from cadmium toxicity. This case study nicely illustrates the point made by Wray *et al.* (2003), that polymorphisms in promoters can affect differential transcriptional responses and such polymorphisms can entail differential phenotypes with microevolutionary consequences.

Studies of promoter architecture and adaptive variation in natural populations, such as the one discussed above, are still rather rare. Most of our knowledge of the evolutionary aspects of promoter architecture derives from comparative analysis between the human genome and the genomes of mouse and yeast. In yeast the notion has arisen that there are two different classes of promoters, one controlling stable gene expression, the other controlling highly inducible expression. Interestingly, the promoters that were associated with great responsiveness to environmental conditions also diverged the most when different yeast strains were compared. Expression divergence among strains correlated well with gene responsiveness. This suggests that natural selection reinforces the divergence of those genes that had an already existing large variability in gene expression (Tirosh *et al.* 2009).

Since gene responsiveness and expression divergence are ultimately due to promoter architecture, the question arises, what causes this fundamental difference in promoter function? A good candidate for an explanation is the TATA box. In contrast to what had been believed before, not all genes have a TATA box, especially not in mammals. In fact TATA-driven promoters in mammals are the exception, not the rule (Farnham 2009). Still, the TATA box is a common element of core promoters and its function is suggested to be to increase the extent of re-initiation of the transcriptional complex. With the TATA box stabilizing the association of RNA polymerase with DNA, other factors can more easily bind to the complex and so there will be a multiplicative effect on gene expression. In addition, the proximal region of TATA-driven promoters is more often occupied by nucleosomes that may help recruit chromatin regulators which further fine-tune expression.

In accordance with this theory, Tirosh *et al.* (2006) found that the occurrence of TATA-containing promoters was especially high among genes with high expression divergence among yeast strains (Fig. 6.14a). When grouped in functional categories, this

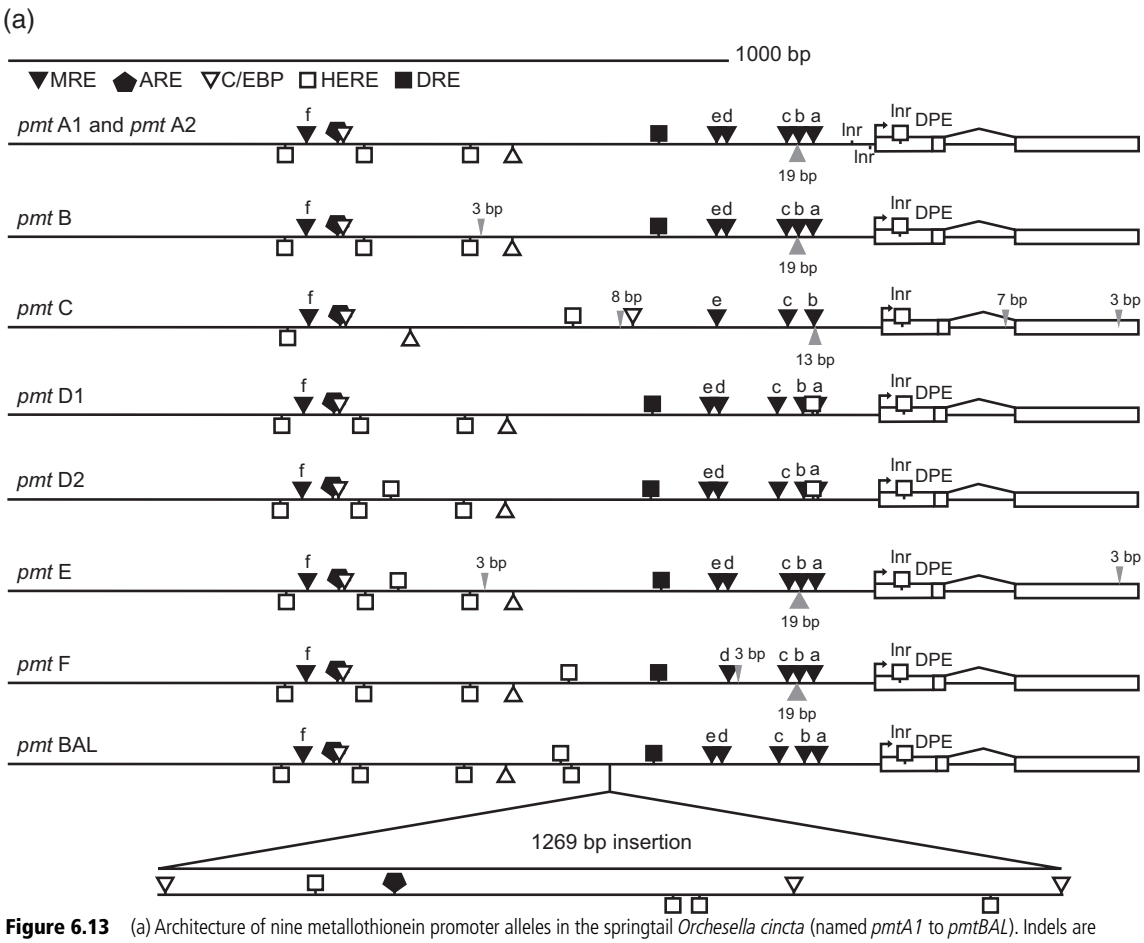


Figure 6.13 (a) Architecture of nine metallothionein promoter alleles in the springtail *Orchesella cincta* (named *pmtA1* to *pmtBAL*). Indels are indicated by arrows. Putative transcription factor binding sites are indicated: MRE, metal responsive element (labelled a, b, c . . .), ARE, antioxidant responsive element, C/EBP, binding site for CCAAT enhancer binding protein, HERE, 20-hydroxy-ecdysone responsive element, Inr, Initiator sequence, DPE, downstream promoter element. The coding sequence (including one intron) is seen on the right. (b) Reticulate network of the same promoter alleles, established using Splitstree v4, NeighborNet, and Reticulate. The network analysis illustrates that recombinatory events have been important in the evolution of these sequences. For example, alleles A1 and A2 most likely originated from recombination with D1 and B. The metallothionein promoter of a related species, *Orchesella villosa*, was used as an outgroup. From Janssens *et al.* (2007), by permission of BioMed Central.

trend became even more obvious: genes associated with plasma membrane proteins or defence against stress showed a high divergence across yeast strains and were also enriched in TATA-containing genes. TATA box promoters might be especially important in stress defence genes. It is known that genes with a TATA box are associated with high levels of intrinsic noise, that is, with large cell-to-cell variability of expression (López-Maury *et al.* 2008). Such a system might be eminently suitable for providing a sudden burst of gene expression under environmental stress and this might be an important function of the TATA box promoter. Consequently, the same genes might be used in evolutionary adaptation and enhanced stress tolerance (Roelofs *et al.* 2010).

We have seen above, in work on *cis*- and *trans*-regulatory change in *Drosophila*, that divergence between species is more often associated with differences in *cis*-regulation of genes, while variation across strains within a species is often due to divergent *trans*-acting influences on gene expression. Tirosh *et al.* (2009) argued that, indeed, expression flexibility is also more dependent on *trans*-acting regulation, especially through signals from the environment (Fig. 6.14b). If TATA box-driven promoters are primarily associated with gene regula-

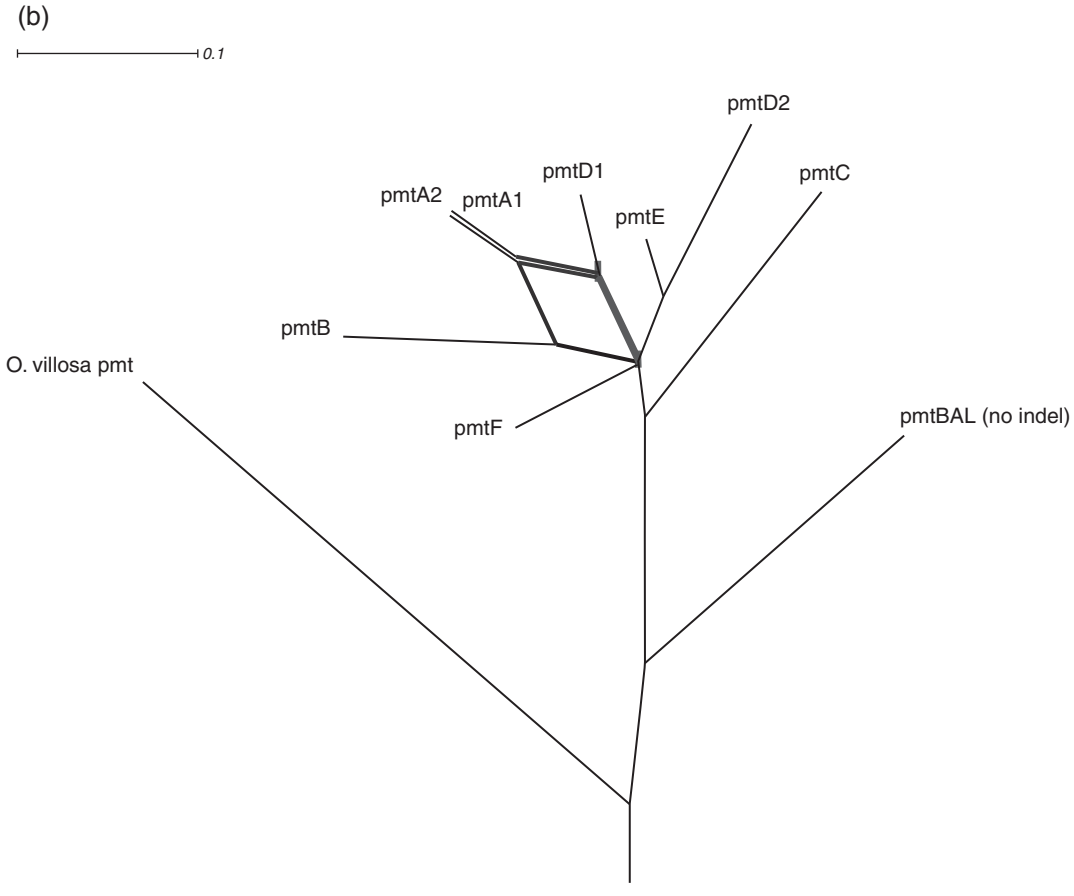


Figure 6.13 (*Continued*)

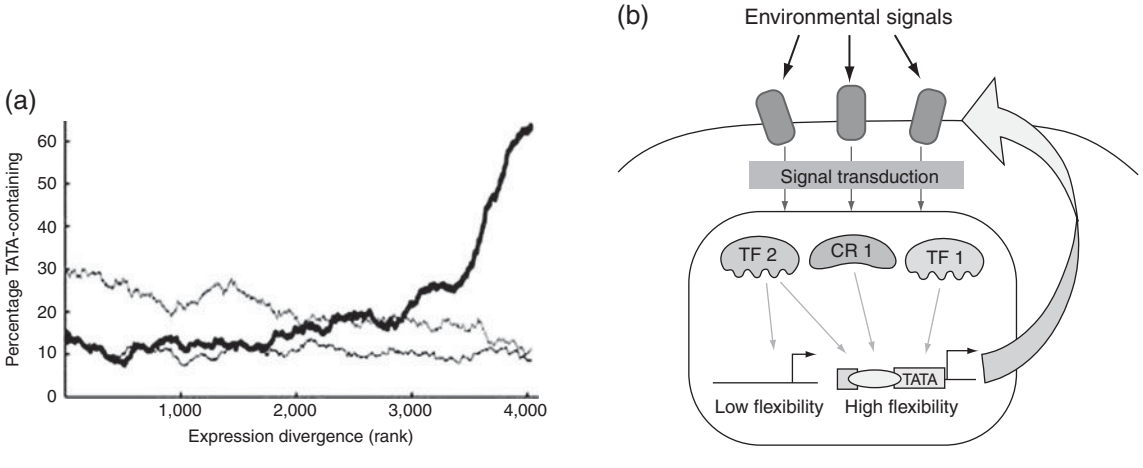


Figure 6.14 (a) Percentage of TATA containing genes in windows of 400 genes across the genome of *Saccharomyces*, as a function of the expression divergence of these genes among yeast strains. The solid line is for all TATA boxes, the thin lines are for nonconservative TATA boxes (upper) and for gene coding sequences (lower; in the latter case the X-axis denotes sequence divergence rather than expression divergence). From Tirosh *et al.* (2006), by permission of Nature Publishing Group. (b) Scheme illustrating how expression flexibility, mediated by the TATA box, may be associated with increased dependence on *trans*-acting factors regulated by environmental signals. From Tirosh *et al.* (2009) by permission of Biomed Central.

tion in *trans*, this would be well in line with the divergence of such genes across strains found by Wittkopp *et al.* (2008b). So promoter architecture may partly explain evolvability of gene expression.

Another contrast in the classification of promoters is due to the presence of CpG islands in the 5′ end of the gene (the p in CpG is usually written to distinguish a CG dinucleotide from a CG base-pair). CpG islands are stretches of DNA where a significant part of the sequence consists of CG dinucleotides. Especially among mammalian promoters, there is a clear distinction between *CpG-poor* and *CpG island promoters*. In CpG-poor promoters, the CpG content is equivalent to the genomic average (1 every 100 bp), while CpG island promoters have CpG dinucleotides about every 10 bp. Similarly, the G+C content of CpG poor promoters is around 42%, while G+C in CpG island promoters reaches values of 65%. In Section 2.1.3 we saw that variation of the GC content across a chromosome determines the isochore structure, where GC rich regions are associated with high densities of genes.

The distribution of CpG islands in the genome is highly suggestive of a role in transcriptional regulation. We will see in Section 6.4.1 that CpG dinucleotides are the sites where DNA methylation of cytosine takes place, however, when CpG dinucleotides are packed in CpG islands they are often devoid of methylation. These regions are sites of high transcriptional activity and open chromatin, most likely due to the binding of proteins that block DNA methyltransferase.

The functional significance of CpG islands in a promoter is not yet clear. In the human genome, all housekeeping genes have a CpG island promoter, while CpG poor promoters are more often associated with tissue-specific genes. However, gene regulation can operate very well without the benefits of CpG islands. In the human genome, functionally similar genes, even genes expressed in the same cell type, for example α-globin and β-globin, differ in the presence of a CpG-island promoter (α-globin has one, β-globin not). Also, some genes which rely on a CpG island promoter for their expression in the human genome are driven by a CpG poor promoter in the mouse genome (Antequerra 2003).

In a genome-wide analysis of mammalian promoter structure, Carnici *et al.* (2006) found that CpG poor promoters usually have a TATA box and a well-defined transcription start site, while CpG island promoters have a broader region of transcription initiation. They also found genes that have a mixed promoter make-up: both CpG islands and a TATA box. From a comparison of the mouse and human promoter architecture, Carnici *et al.* (2006) concluded that CpG island promoters seem to be rapidly evolving, while TATA box containing promoters are more constrained. This conclusion is rather at variance with the analysis of Tirosh *et al.* (2006) in yeast (Fig. 6.14). This could point at a fundamental difference between the mammalian machinery for transcriptional control and the one used by invertebrates, fungi, and plants. The CpG island configuration seems to be especially important in mammals and does not play a similar role in other organisms. There is an enormous *terra incognita* for ecological genomics to better define the relationship between promoter structure, environmental conditions, and expression divergence for organisms in the wild.

6.3.3 Neutral and adaptive variation of gene expression

That gene expression can be considered an individual, heritable trait, and that it can be subject to selection in wild populations, was first demonstrated in fish studies. Two species played an important role in paving the way for functional genomics in relation to adaptation in the field (Fig. 6.15):

1. Killifish, also called mummichog or mud minnow, *Fundulus heteroclitus* (Cyprinodontiformes), a species with a very wide distributional range along the east coast of North America, living in intertidal habitats and coastal marshes but also colonizing inland lakes and streams. It has been an important model in animal physiology for work on osmoregulation, hypoxia, and energy metabolism, and it is also used as a test species in ecotoxicology (Burnett *et al.* 2007).

2. Lake whitefish, *Coregonus clupeaformis* (Salmoniformes), living in deep lakes in the Northern

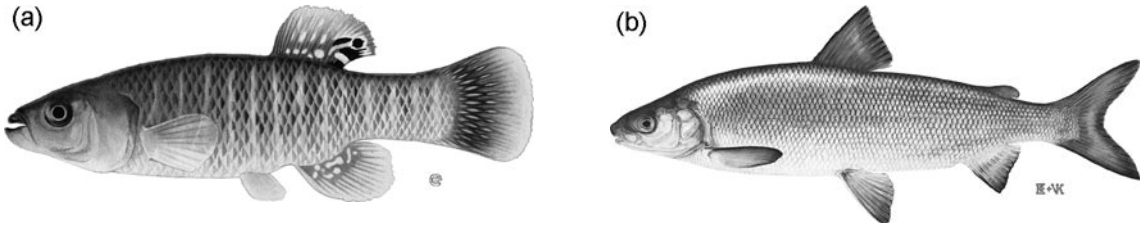


Figure 6.15 Two model fish species of ecological genomics, *Fundulus heteroclitus* (a) and *Coregonus clupeaformis* (b). Images provided courtesy of the New York State Department of Environmental Conservation. All rights reserved.

United States and Canada. This species shows a sympatric divergence between limnetic and epibenthic forms, not unlike the threespined stickleback populations discussed above (Section 6.2.2). Lake whitefish is an important model for the study of adaptive divergence and reproductive isolation (Bernatchez *et al.* 2010).

In both fish species, an appreciable number of genes differ significantly in expression level between individuals, even when the animals are acclimated to the same conditions. In a microarray study with *Fundulus heteroclitus* 18% of 904 genes analysed showed such inter-individual variation (Oleksiak *et al.* 2002). A similar percentage of variable genes is found when comparing different strains of *Drosophila* or yeast. In the *Fundulus* study, fish were sampled from the field, then maintained in laboratory culture for 6 months under identical conditions, and a single tissue (heart ventricle) was sampled. This design is expected to minimize effects due to tissue-specific gene expression and variable physiological conditions. What remains are genetic factors, although maternal effects and carry-over (e.g. imprinting) from early-life environmental exposures cannot be excluded. The magnitude of the variation between individuals typically amounted to a factor of 1.5, but for some genes ranged to a factor of 4. When comparing populations, gene expression differences are often found to lie in the same range. This implies that analysis of between-population variation can only be accounted for by sampling at the level of individual animals.

The differences in gene expression between individual *Fundulus* are indicative of the existence of different 'physiotypes' in the population. Cardiac metabolism was measured using three substrates representing different physiological activities (glucose, fatty acid, and lactate), and was correlated with the expression of genes from three main metabolic pathways for energy metabolism: glycolysis, tricarboxylic acid cycle, and oxidative phosphorylation. The gene expressions associated with these pathways were taken together by multivariate statistics and summarized into principal components. A correlation analysis then showed that gene expression in a pathway did not have the same effect on energy metabolism for each physiotype (Fig. 6.16). For example, in individuals of group 1, glucose metabolism was greatly influenced by gene expression of glycolytic enzymes, but in group 2 glucose metabolism depended more on the expression of oxidative phosphorylation enzymes (Oleksiak *et al.* 2004; Crawford and Oleksiak 2007).

These studies show that gene expressions may be meaningful predictors of physiological performance and that the regulation of energy metabolism differs between individual fish, most likely due to genetic factors. Because cardiac metabolism is an important aspect of fish life in variable environments, these differences most likely are not neutral but present relevant phenotypic variation for selection to act upon.

As stated earlier, to prove that gene expression differences between populations are adaptive, one has to consider neutral variation (genetic drift) as a null hypothesis. This is a crucial aspect of any comparative approach, since populations sampled in different geographic areas will always differ from each other due to isolation and genetic drift, not necessarily proving local adaptation. The question

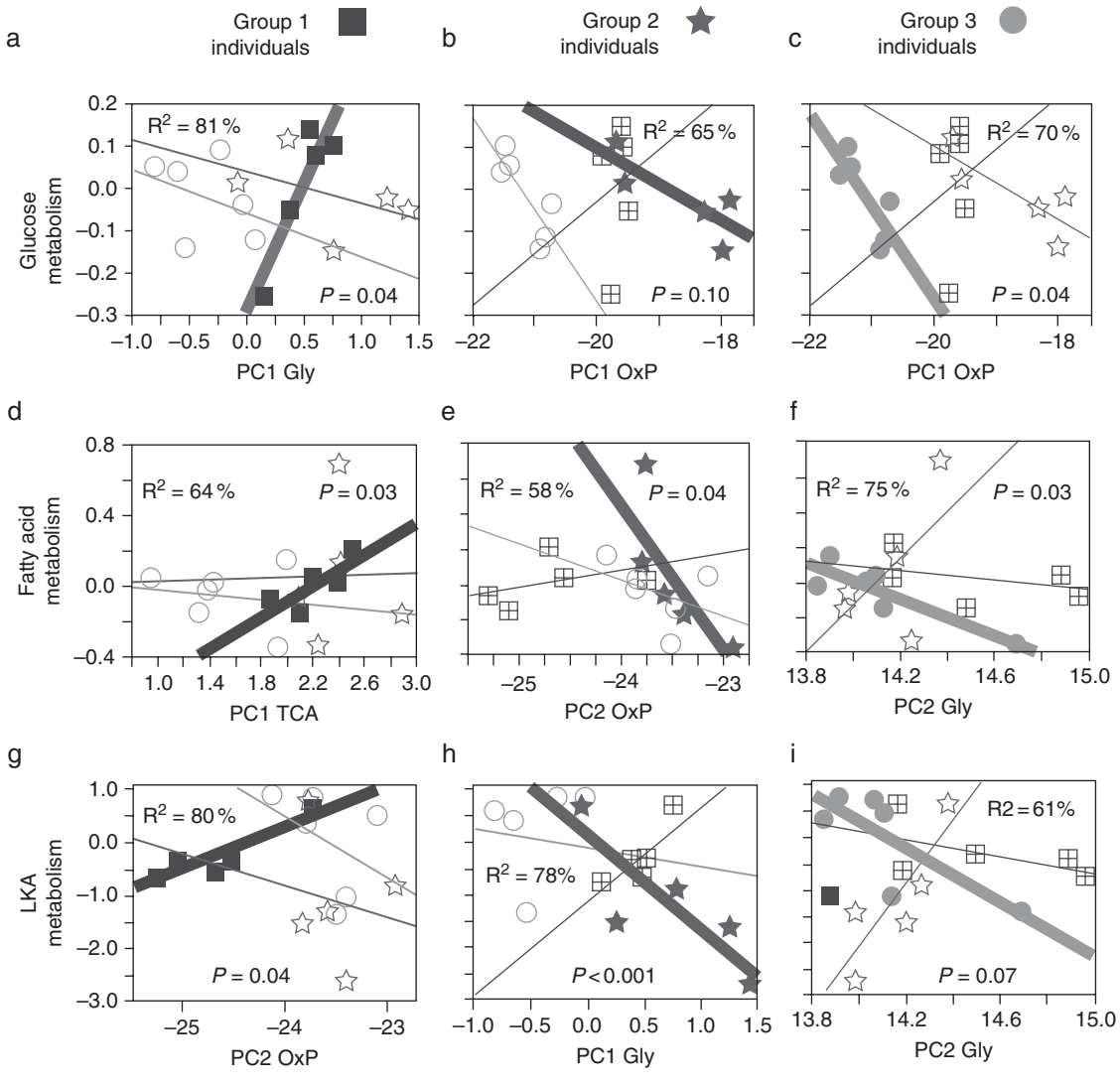


Figure 6.16 Correlation between cardiac metabolism of killifish, measured using different substrates (glucose, fatty acids, and lactate/ketones/alcohol, LKA), as a function of the most informative principal component (PC1 or PC2) of gene expression for genes associated with three metabolic pathways, glycolysis (Gly), oxidative phosphorylation (OxP), and tricarboxylic acid cycle (TCA). The correlations are given for three groups of fish (Group 1, Group 2, and Group 3). Bold lines, R^2 values and p values in each diagram indicate the relationship between metabolism and gene expression of the group involved; the thin lines are for the other two groups. The data indicate that there are different 'physiotypes' among these fish, and each type has its own way of regulating cardiac metabolism by gene expression. From Oleksiak *et al.* (2004), by permission of Nature Publishing Group.

is similar to the test for outliers in population genomics (Section 6.2.3): only gene expression differences exceeding neutral expectations can be considered indicative of selection. Whitehead and Crawford (2006a, b) recognized this problem and developed an approach, using microsatellite markers, to correct for neutral evolution.

The principle of the method applied by Whitehead and Crawford (2006a) is that populations are genotyped by neutral markers in addition to the

functional genomics profiling. The neutral markers are used to draw up a phylogenetic tree from which genetic distances between populations can be read, and gene expression is regresssed on these distances. This will provide a measure of explained variance due to phylogeny, r^2, which is then used as a null hypothesis for adaptive explanations.

In a study of five populations of *Fundulus heteroclitus* along the Atlantic coast of the US, the population analysis suggested a divide off the Hudson river, plus an effect of isolation-by-distance in a northern as well as a southern direction. Gene expressions in liver were regressed on genetic distance and on temperature, and the two different r^2 values were plotted against each other in a correlation diagram (Fig. 6.17). A positive association between these variables is expected, which is due to the fact that temperature is showing the same geographic gradient as genetic distance. For 13 genes, however, the temperature r^2 exceeded the r^2 for genetic distance. Only for those 13 genes is there evidence for directional selection on top of the phylogenetic signal (other genes might also be under selection, but in these genes temperature adaptation is obscured by a colinear effect of genetic distance).

The 13 genes showing a signature of temperature adaptation in their expression, not unexpectedly, were found to be related to energy metabolism and stress responses; most of them were known from temperature acclimation studies in other fish. Interestingly, the gene encoding lactate dehydrogenase B (LDH-B) was not among the 13, despite the fact that earlier research (Schulte *et al.* 2000) had demonstrated a twofold difference in *Ldh-B* expression among northern and southern populations of *F. heteroclitus*. This difference was shown to be due to a substitution in a glucocorticoid responsive element in the promoter of *Ldh-B* and was interpreted as an adaptation to temperature. The fact that *Ldh-B*

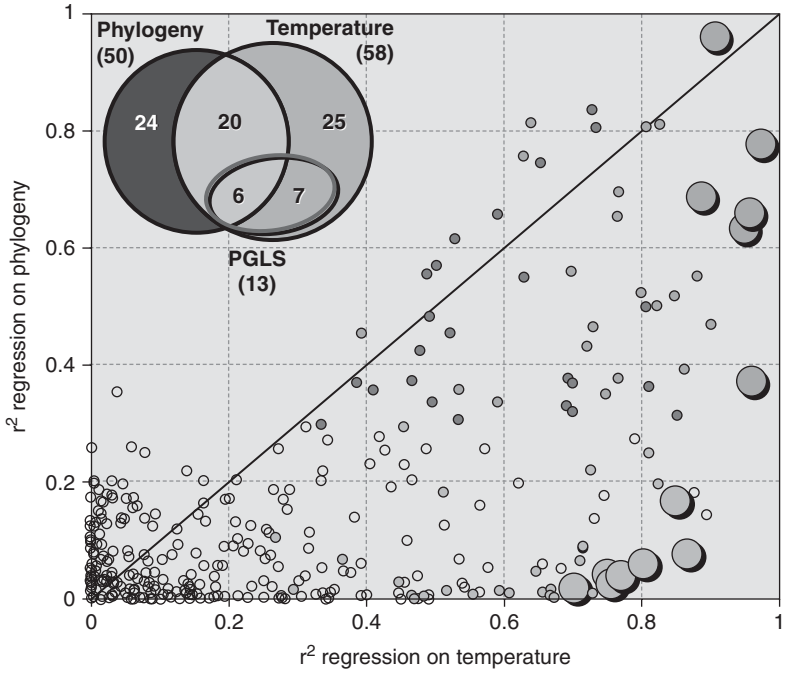


Figure 6.17 Correlation between the fraction of explained variance in liver gene expression (r^2) due to genetic distance (vertical axis), and r^2 due to temperature (horizontal axis) in five populations of killifish, *Fundulus heteroclitus*, in a geographic gradient. Every dot represents a gene. The enlarged spots indicate 13 genes for which there is significant regression with habitat temperature after correction for phylogeny (also designated as PGLS in the Venn diagram, inset). From Whitehead and Crawford (2006a), with permission from the National Academy of Sciences of the USA.

was not identified as a gene under selection in the genomics study of Whitehead and Crawford (2006a) might be due to the very stringent criteria applied. All in all, the *Fundulus* studies illustrate the important point that some form of correction for neutral processes must be applied when studying gene expression in geographically separate populations.

Another important evolutionary principle is illustrated by work on lake whitefish: convergent evolution of body form driven by parallel transcriptional change. American lake whitefish, *Coregonus clupeaformis*, occupies a wide range of habitats with substantial phenotypic variation both among and within populations. In several lakes of North America it has diverged into two morphological phenotypes, a limnetic dwarf form that lives in sympatry with the normal benthic form. Despite gene flow and hybridization the two ecotypes remain reproductively isolated (Bernatchez *et al.* 2010). A similar divergence has happened in the European sister species, *Coregonus lavaretus*. In addition, two other species in the whitefish complex, *C. artedi* (cisco, a North American species) and *C. albula* (vendace, a species from Eurasia) only consist of dwarf forms (Fig. 6.18).

It is assumed that the benthic ecotype is the ancestral form, and that dwarf morphs have evolved several times in the lake whitefish complex. Some species have dwarfed completely (cisco, vendace), while in others the divergence started later and there is presently a polymorphism in the population. The dwarf ecotype of *C. clupeaformis* is found in lakes in which the cisco (*C. artedi*) is not present. The absence of cisco seems to create an ecological opportunity for whitefish to develop a limnetic life-form, while in the presence of cisco such a dwarf ecotype would have to compete with the cisco, which is the better competitor. In *C. clupeaformis*, the evolution of limnetic life-forms has happened in parallel in various lakes, and is not more than 15 000 years old.

The dwarf and normal whitefish utilize different microhabitats (limnetic and benthic environments, respectively), and differ not only in total body size, but also in several other phenotypic characteristics, including the number and position of gill rakers, growth rate, size at maturity, metabolic rate, energy conversion, and also swimming behaviour. Predation is suggested as the main selective force driving these differences; predation pressure, being high in surface habitats, logically calls for earlier maturation, higher metabolism and activity, as well as predator avoidance strategies such as fast swimming.

Population genomics studies, including genome scans for F_{ST} outliers and QTL analysis have pro-

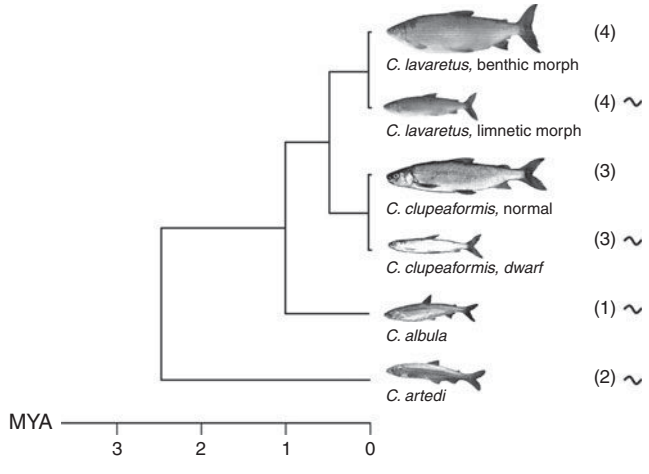


Figure 6.18 Phylogeny based on mtDNA restriction polymorphisms, of four species in the whitefish complex and their divergence into limnetic (indicated by a wave) and benthic ecotypes. MYA: million years before present. From Jeukens *et al.* (2008), by permission of Oxford University Press.

vided strong evidence that these differences are due to directional selection (Campbell and Bernatchez 2004; Rogers and Bernatchez 2005, see Section 6.2.3 and Table 6.1). In the QTL analysis using controlled crossings in the laboratory, 35 AFLP-associated loci were identified that were significantly associated with body growth. Subsequently, 27 of these AFLP loci were used to screen four natural populations, each with a sympatric species pair. The F_{ST} outlier analysis, using the principles explained in Section 6.2.3, resulted in eight loci for which there was evidence of directional selection.

That only 8 out of 27 growth-related loci were identified as being under selection implied that the majority of the QTLs showed neutral levels of divergence between the two ecotypes. Apparently, the differences in body growth are partly due to non-adaptive genetic polymorphisms. This again illustrates the importance of neutral evolutionary processes, even in this case where the phenotypic divergence is so obvious. However, the analysis used a limited number of AFLP markers and therefore the possibility exists that important QTLs were missed. It would be interesting to redo the genetic analysis with the modern high-throughput SNP genotyping approach (see Section 6.2.2).

These population genomics studies were followed-up by transcription profiling studies in muscle, liver, and brain tissues (Derome *et al.* 2006; St-Cyr *et al.* 2008; Whiteley *et al.* 2008). Transcription profiling was done using a cDNA microarray with 16 000 probes from Atlantic salmon (*Salmo salar*). Although such a heterologous hybridization approach has obvious drawbacks with regard to efficiency and bias (see Chapter 1), the comparison between normal and dwarf whitefish is probably not affected.

Comparing gene expressions in the liver of normal and dwarf whitefish, a considerable number of genes were found to differ significantly (St-Cyr *et al.* 2008). This comparison was made in each of two lakes and it turned out that 248 genes differed between the two morphs in both lakes. Of these genes, 92 differed in the same direction across the lakes, that is, if a gene was upregulated in the dwarf morph in one lake, it was also upregulated in the dwarf of the other lake. The other 156 genes did not have this pattern: they were differentially regulated but not consistently across the lakes. Looking at the functions of the parallel genes, the liver transcriptome of the dwarf whitefish was found to be enriched in energy metabolism, lipid metabolism, iron homeostasis, detoxification, and sexual maturation, while the normal whitefish transcriptome was enriched in protein synthesis and cell cycle genes. In muscle tissue, genes involved in the regulation of muscle contraction and energy metabolism were found to discriminate the dwarf and normal ecotypes (Derome *et al.* 2006).

We should realize that the two lakes studied represent two independent evolutionary events in which the whitefish diverged into limnetic and benthic forms. The similarity in body forms is a nice example of *parallel evolution*. The interesting conclusion from the transcriptomic studies is, however, that the same trend seems to hold on the molecular level: similar changes in gene regulation have occurred in the two lakes. This suggests that genetic constraints limit the number of options for the genome to evolve. Wherever there is a similar selection pressure different populations will take the same genetic option, because there are only a few, and arrive at the same phenotype.

The trends found for ecotype differentiation within lake whitefish have also been extrapolated to the two dwarf species in the whitefish complex, cisco (*Coregonus artedi*) and vendace *(Coregonus albula*). The North American cisco and the European vendace represent whitefish that are completely dwarfed, lacking any normal size populations (cf. Fig. 6.18). The question was asked whether the genetic mechanisms underlying dwarphism in *C. clupeaformi*s and *C. lavaretus* also hold for the evolution of cisco and vendace.

Initially, a microarray study seemed to confirm that genes functionally involved with the regulation of swimming in muscle tissue of the dwarf ecotype of *C. clupeaformis* were also upregulated in the cisco (Derome and Bernatchez 2006). This indeed suggested that the same adaptive mechanisms might underly the switch from a benthic to a limnetic lifestyle within and between species. However, a subsequent qPCR study (Jeukens *et al.* 2008) could not

confirm this. Although strong evidence was found for parallelism in gene expression for two liver genes, anionic trypsin and carboxylesterase, two genes expressed in muscle did not show the same pattern. Among two genes expressed in the eye, one did not and the other did confirm parallel evolution (Jeukens *et al.* 2008). An overview of these patterns is given in Table 6.3. These data illustrate that the phenotypic similarity between cisco and vendace on the one hand, and the dwarf ecotypes of normal whitefish on the other, might be due to real convergent evolution. However, the evolution of dwarf forms within *C. clupeaformis* and *C. lavaretus* seems to be a case of parallel evolution at the gene regulation level.

The differentiation of fish in limnetic and benthic life-forms is a fascinating example of parallel evolution in action. It is, however, not limited to whitefish. To date four case studies illustrate a similar pattern:

1. *Gasterosteus aculeatus* (threespined stickleback), discussed in Section 6.2.2 (Peichel *et al.* 2001, Colosimo *et al.* 2005),
2. *Coregonus clupeaformis* and related whitefish species, discussed above (Bernatchez *et al.* 2010),
3. *Telmatochromis temporalis*, a cyclid fish from Lake Tanganyika, which in two areas developed a dwarf morph that finds shelter in empty snail shells (Takahashi *et al.* 2009),
4. *Amphilophus astorqui* and *A. zaliosus*, two closely related cichlids endemic to Lake Apoyo, Nicaragua (Elmer *et al.* 2010). *A. zaliosus* is a recent limnetic species evolved by sympatric speciation.

Obviously, the studies at the moment are only scraping the surface of the matter; more detailed whole-genome gene expression and SNP genotyping studies in this wonderful model system are necessary to delineate the various processes of parallel and convergent evolution at the genetic and the phenotypic level.

6.3.4 Variation at the phenotypic, gene expression, and DNA sequence levels

We have seen earlier in this Chapter that variation of gene expression has a strong heritable component; to be subject to natural selection, however, gene expression must also have phenotypic consequences.

Table 6.3 Overview of gene expressions measured by quantitative real-time PCR in liver, muscle, and eye tissue of lake whitefish from various locations in North America and Europe. Up- or downregulation in the limnetic (dwarf) form, relative to the normal form, is indicated by arrows. Cisco was compared to the mean of the North American normal whitefish, and vendace to the mean of the normal European whitefish. Under the hypothesis of parallel evolution across all lakes and species, arrows in the same row are expected to point in the same direction. Modified after Jeukens *et al.* (2009)

Gene	*C. clupeaformis*			*C. lavaretus*				*Car*	*Cal*
	North America			Europe				NA	Eu
	1	2	3	4	5	6	7	8	6
Anionic trypsin in liver	↑	↑	↑	↑	↑	↑	↑	↓	↓
Carboxylesterase in liver	↑	↑	↑	↑	↑	↑	↑	↓	↓
Parvalbumin in muscle	↑	↓	↑	↑	↓	↑	↑	↑	↑
Lactate dehydrogenase A in muscle	↑	↑	↑	↑	↓	↑	↓	↓	↓
Opsin SW1 in eye	n/a	n/a	n/a	↓	↓	↓	n/a	n/a	↓
Opsin LWS in eye	n/a	n/a	n/a	↑	↓	↓	n/a	n/a	↓

Locations: 1, Cliff Lake, 2, Indian Pond, 3, Laboratory culture, 4, Lake Zürich, 5, Lake Lucerne, 6, Pasvik river, 7, Lake Stuorajávri, 8, Lac des Trente-et-un-Milles and Lac Florent. Car, *Coregonus artedi* (cisco, North American dwarf whitefish), Cal, *Coregonus albula* (vendace, European dwarf whitefish), n/a, no data.

An important question, therefore, is how variation at the levels of DNA sequence, gene expression, and the phenotype are related to one another. Is the sequence variation observed in the genome mostly due to neutral processes with little phenotypic effects, as is argued by the neutral theory of molecular evolution, or is there a correlation between sequence divergence and phenotypic variation?

The yeast, *Saccharomyces cerevisiae*, is a good organism with which to study this question because its genome is very well known and phenotypic variation can be studied easily. For this reason, *Saccharomyces* is presently considered an important model, not only for molecular biology, but also for ecology and evolution (Landry *et al.* 2006a; Replansky *et al.* 2008).

There is substantial phenotypic variation in budding yeast, which has allowed the culturing of a

diverse set of strains used for many centuries in the manufacture of wine, sake, soy sauce, beer, and bread. The various biotechnological phenotypes, such as baker's yeast, brewer's yeast, wine yeast, and so on, all belong to the species *Saccharomyces cerevisiae*. The closely related wild species, *Saccharomyces paradoxus*, has never been found in association with human activities, although it sometimes lives in sympatry with *S. cerevisiae* in the wild. Yeast strains differ greatly in their physiological capacities, and in their tolerance to drugs, alcohol, salt, copper, and so on. There is also morphological variation in colony form and colour between the strains. Some isolates may develop a red-brown or rust colour; others show a pseudo-mycelium or a filigreed growth form.

Phenotyping in yeast can be done using high-throughput technology. Isolates are grown in multi-well arrays, in a large variety of different stressful conditions, including cold, heat, low pH, high ethanol, toxins, and so on, while growth of the colonies is monitored over time. Growth lag time, doubling time, and growth efficiency are measured for each condition. In this way it is possible to characterize strains more or less automatically, using hundreds of quantitative phenotypic traits. The data can be expressed as an average strain-specific sensitivity relative to a reference strain.

DNA sequencing shows that polymorphism in yeast is considerable. In a genome-wide comparison of an Italian and a Californian isolate, using the standard laboratory strain as a reference, Doniger *et al.* (2008) found 88 000 polymorphisms, of which 97% were SNPs and 7% were indels. In a more extensive genome-wide survey, involving more than 70 isolates from soil, rotting fruit, bread, grapes, beer, sake, and various clinical sources, Liti *et al.* (2009) reported 235 127 polymorphisms, with nearly the same distribution over SNPs and indels as reported by Doniger *et al*. (2008). Even more SNPs (1.3 million) were identified in a comparison of 63 isolates using a whole-genome tiling array (Schacherer *et al.* 2009).

Yeast strains also show some variation in chromosome arrangement. Wine yeasts have a reciprocal translocation which has moved a segment of DNA containing a sulphite membrane transporter gene to another chromosome. On the other chromosome the sulphite transporter has come under the influence of a new, highly active promoter and is overexpressed. This adds to the increased sulfite resistance of wine yeast. In addition, some strains of *S. cerevisiae* have obtained a piece of DNA, carrying an allele of the *EHD3* gene from the closely related wild species *S. paradoxus*.

The genome-wide inventories of genetic variation in yeast have also revealed novel sequences in some of the strains that cannot be matched with the reference genome. The unplaced sequences involve 38 new open reading frames, mostly in the subtelomeric regions. Very little divergence between the strains was due to gene copy number variation, except for the rRNA genes. The cosmopolitan occurrence of yeast and its occupation of many different habitats may have contributed to shaping this large amount of genetic variation.

In line with the large degree of genetic polymorphism, gene expression variation in yeast is widespread; it involves genes associated with a variety of basal processes, such as amino acid metabolism, sulphur assimilation, and protein degradation. This seems to indicate that strain variation in yeast is not due to specific operational genes and specialized phenotypes such as resistances and virulence in prokaryotes, but involves many aspects of the basal metabolism. However, the expression differences generally are small—less than a factor of 2 on the average (Townsend *et al.* 2003). Divergence concerns not only gene expression differences as such, but also the responses of the various genotypes to environmental conditions (Landry *et al.* 2006b). Some yeast strains are more plastic than others when grown under different conditions (synthetic wine most, nitrogen-limited medium, etc.). Among these genes, a significant number show genotype-by-environment interaction, that is, their expression in response to the different growth media depends on the strain. Interestingly, this set is biased towards genes with a paralogue elsewhere in the genome; why this should be so is not entirely clear.

The relationship between DNA sequence variation and gene expression divergence may be illustrated by a study by Fay *et al.* (2004). These authors analysed nine strains of *S. cerevisiae* originally

sampled from vineyards, rotting fruit, and oak tree exudates in Italy and the US. To estimate genetic distances between the strains, 31 polymorphic sites in three genes were used. Looking at gene expression, 516 genes showed differences between strains; these represented 8.4% of the number of probes on the microarray used. Pairwise DNA sequence divergence was significantly correlated with expression difference (Fig. 6.19).

The correlation between DNA sequence variation and divergence of gene expression is confirmed by other studies using high-throughput phenotypic screens (Liti *et al.* 2009). These authors reported a good correlation between 200 quantitative phenotypic traits and genome similarity across more than 70 fully sequenced isolates of yeast.

How much of the phenotypic variation in yeast is responsive to selection from the environment is not clear. Doniger *et al.* (2008) could not detect any signature of positive (adaptive) selection. Instead the authors found many deleterious SNPs; about 36% of nonsynonymous substitutions were found to be under negative selection. For 22% of the nonsynonymous SNPs a disruption of protein function was likely. More than 1000 deleterious SNPs were found that disturbed open reading frames conserved across a range of other fungi. In addition, a considerable number of deleterious SNPs were discovered within conserved non-coding regions. Many of these deleterious SNPs probably have phenotypic consequences; some polymorphisms, especially when in *cis*-regulatory regions of genes, may contribute to the variation in gene expression.

The lack of evidence for positive selection in yeast is similar to the situation in *Arabidopsis*. In early plant population genetic studies, claims have been made that 20% to 30% of the loci studied bear a signature of selection. However, Wright and Gaut (2005) argued that many of the older reports are biased because the genes studied are not a random sample from the genome, and appropriate controls for demographic effects (population size and structure) are often

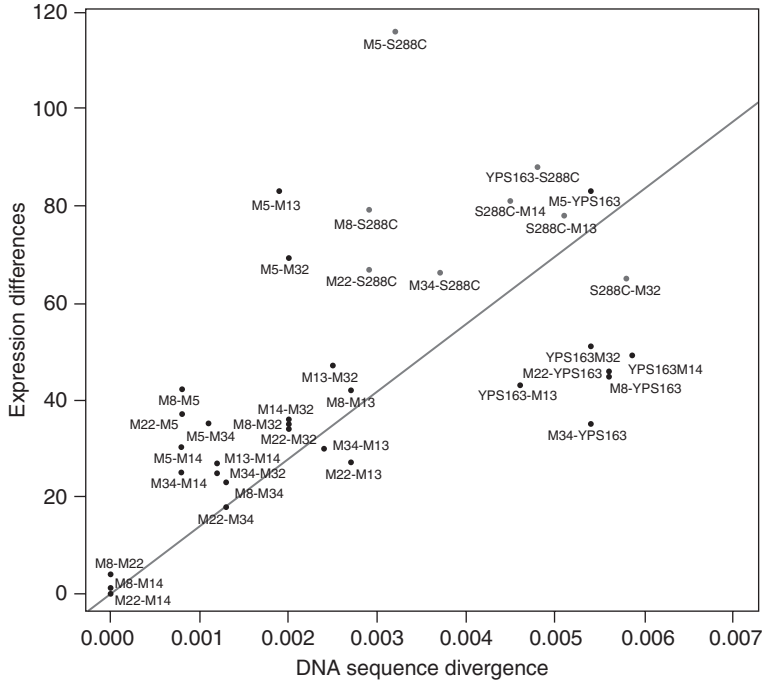


Figure 6.19 Correlation between pairwise differences in average gene expression and DNA sequence divergence across nine different natural isolates of baker's yeast, *Saccharomyces cerevisiae*. From Fay *et al.* (2004), by permission from BioMed Central.

lacking. Also candidate loci were often not checked for their fitness effects or ecologically relevant phenotypes. Actually the picture in *Arabidopsis* is one of purifying selection on deleterious polymorphisms and a near-absence of strong positive selection.

The situation in yeast and *Arabidopsis* rather contrasts with the *Drosophila* studies discussed above. In fruit flies there is quite a bit of evidence for positive selection acting on many loci (e.g. Shapiro *et al.* 2007). Nearly 30% of the amino acid substitutions between *D. melanogaster* and its close relatives are to be considered adaptive. Why the three genomic model species differ so much in this important aspect of their genetic variation is not clear at the moment. Maybe it is due to some crucial aspect of the life cycle, for example outcrossing or population size, which are both higher in fruit flies. In addition, there is a feeling that the statistical tests used to identify selection are often very conservative and unable to detect adaptive evolution even if quite prevalent (Ford 2002; Charlesworth and Eyre-Walker 2008).

In conclusion, the relationship between phenotypic variation, gene expression, and DNA sequence divergence appears to be complicated. Yeast studies show that there is no dearth of variation at the DNA sequence level; gene expression as well as phenotypic variation increases with sequence divergence, depending on strain-by-environment interaction. Although much of this phenotypic variation is expected to be adaptive, a signature of positive selection cannot always be easily detected in genome-wide analyses. Instead, the picture in yeast and *Arabidopsis* is dominated by weak purifying selection. How the situation will turn out for the many species studied by ecologists is difficult to tell at the moment. At any rate, the adaptive phenotypes and their genetic loci that allow adequate responses to specific environmental conditions are hidden within a large body of neutral variation and numerous deleterious mutations.

6.4 Epigenetic variation and developmental change

The use of genome-wide technologies has given a great impetus to the study of *evolutionary developmental biology*, abbreviated '*evo-devo*'. This relatively new branch of evolutionary biology investigates the mechanisms by which evolutionary changes in gene structure and gene regulation have changed the way in which an organism presents itself in three-dimensional space: its shape and form. Evolutionary developmental biology is founded in molecular genetics, it includes embryology and comparative morphology and it attempts to explain the origin and changes of animal design (Carroll *et al.* 2005).

Evo-devo is related to another recent branch of evolutionary biology, *ecological developmental biology*, abbreviated '*eco-devo*'. This is an approach to try and relate the developmental pathways of an organism to its environment. It studies the ways in which organisms integrate signals from the environment in their development, and also how the normal development of an organism is disturbed by stress, drugs, or environmental toxicants (Gilbert and Epel 2009).

In this section we will highlight some of the molecular principles at the interface between evo-devo, eco-devo, and genomics.

6.4.1 Epigenetics and evolution

The term epigenetics was originally introduced by Conrad Waddington to indicate the way in which a given genotype interacts with its environment to produce a phenotype. In fact, the original meaning of 'epigenetics' was very close to the present-day usage of 'development'. Later a much more limited definition of epigenetics was adopted; in the modern view, epigenetics studies the mechanisms by which cellular gene expression is regulated by non-genetic ('on top of genetics') means. Epigenetics tries to explain how gene expression in one cell type differs from gene expression in another cell type, how that signature of cell-type dependent gene expression is transmitted to other cells in the same cell lineage, and how the environment can influence cellular gene expression by leaving a more or less permanent imprint on the genome.

Epigenetic marks in the genome constrain the number of options for the cell to develop and function, and so introduce a memory effect in the developmental trajectory. They are one way to impose

and maintain a certain identity to a cell, and allow for the differentiation between, for example, a kidney cell and a liver cell, even though these cells have the same DNA sequence. The constraints placed by epigenetic marks are one of the main reasons why it is so difficult to clone whole animals from differentiated cells.

Epigenetic marks are often induced by environmental factors. They represent effects that last longer than the immediate response to a changing condition that we studied in Chapter 5 of this book. Often an environmentally-induced epigenetic mark will stay in the genome for the rest of an individual's life. Epigenetics is usually involved when we see long-lasting effects of early development (e.g. in rats, licking and nursing of young by the parents sets an epigenetic mark that is needed for normal parental care behaviour when these young are adults themselves). Because epigenetics is an important vehicle by which organisms respond and adapt to the environment, it is a prime field of investigation for ecologists (Bossdorf *et al.* 2008).

A special case of epigenetics is *genetic imprinting*, a term usually reserved for the situation in which the epigenetic mark is given by one of the parents. The expression of imprinted genes is controlled by specific regions known as *imprinting control regions*. A parental epigenetic mark in this region protects the genes against erasure in the zygote, ensuring that the imprint is maintained throughout development. Around 100 of such imprinted genes have been identified in humans and mice. Imprinting is best known for the transcriptional silencing of one X chromosome in mammalian females, to achieve, for sex-linked genes, an mRNA dosage equal to the males, which lack such a chromosome.

The epigenetic instructions for gene expression do not only have a cell-type or environment-dependent signature, they are also a source of individual variation. Individuals in a population do not all possess the same epigenome due to differences in their DNA sequence, their epigenetic marks obtained from their parents, their early development, and the environmental influences to which they have been exposed. Such variation may be visible to natural selection and thus subject to evolution if there is a heritable component in it. In some publications epigenetics is equated to heritable imprints, but this is not correct and is an unnecessary limitation of the term.

There is no doubt that epigenetic marks are transmitted from parent to daughter cells during mitosis, contributing to the maintenance of cell identity in a cell lineage. Less known is that transfer of epigenetic marks across meiosis to the offspring also occurs, although there is some doubt about the generality of it. Jablonka and Raz (2009) provided a list of over 100 described cases of transgenerational epigenetic inheritance in prokaryotes and eukaryotes. They applied strict criteria to filter any dubious studies, and presented only work in which effects of an inducing agent caused a phenotypic change that reappeared in the offspring for at least two generations.

It seems that epigenetic inheritance is more common in plants and fungi than in animals. This may be due to the different ways in which the germline is separated from the soma in plants and fungi compared to animals. Developmentally induced epigenetic markers in somatic cells may be more easily transferred to germline cells in plants when these somatic cells assume germline functions later in development. In animals the germline is separated from the soma very early in development and so the scope for transfer of epigenetic markers is more limited.

Another explanation for the difference in occurrence of epigenetic inheritance between mammals and plants is the fact that in mammals epigenetic markers are erased and reset two times in the life cycle: during production of sperm cells and during early development of the embryo. Directly after fertilization the whole genome is cleaned of epigenetic marks and then rapidly marked again in specific places. If epigenetic markers are inherited, it is due to incompleteness of the erasure operation. In plants, however, there is no evidence for such widespread resetting of the epigenetic landscape and so plant offspring can receive epigenetic instructions from their parents more easily than mammals.

Among the conditions inducing heritable epigenetic marks, stress plays an important role. This may involve physiological stress such as bad nutrition, heat shock, UV irradiation, and exposure to

toxicants, but also stress associated with domestication and selection for tameness. In plants, a lot of epigenetic reprogramming occurs after hybridization and polyploidization, and some of these changes appear to be heritable. The epigenetic repatterning following chromosome changes is designated *genomic stress* (Rapp and Wendel 2005). Epigenetic changes may be seen as a way to generate phenotypic diversity after a period of stress or a genetic bottleneck. While traditional evolutionary theory predicts a loss of genetic variation with a concomitant loss of phenotypic diversity, the epigenetic view predicts an increase of phenotypic diversity after a bottleneck (Rapp and Wendel 2005, Fig. 6.20).

The evolutionary implications of epigenetic inheritance are potentially great. Jablonka and Raz (2009) argue that, if indeed it turns out that epigenetic inheritance is ubiquitous we are facing a new phase of evolutionary biology in which Darwinian, Lamarckian, and saltational mechanisms are replacing the traditional 'modern synthesis'.

Epigenetic marks transmitted to offspring are less stable than true genetic traits; they usually persist for a number of generations, but tend to dwindle away if the inducing conditions are not repeated. This implies that there must be an interaction between genetic and epigenetic evolutionary mechanisms, for example through the mechanism indicated by West-Eberhard (2005). This author argued that through environmentally induced, flexible epigenetic patterning, many species can develop a range of morphologies or physiotypes from the same genotype. For example, the parallel development of dwarf limnetic forms in several species of fish, as discussed in Section 6.3.3 suggests that the ancestral species availed themselves of a plastic developmental programme that allowed the production of various phenotypes, including dwarf forms living at the surface and normal forms living near the bottom. As long as

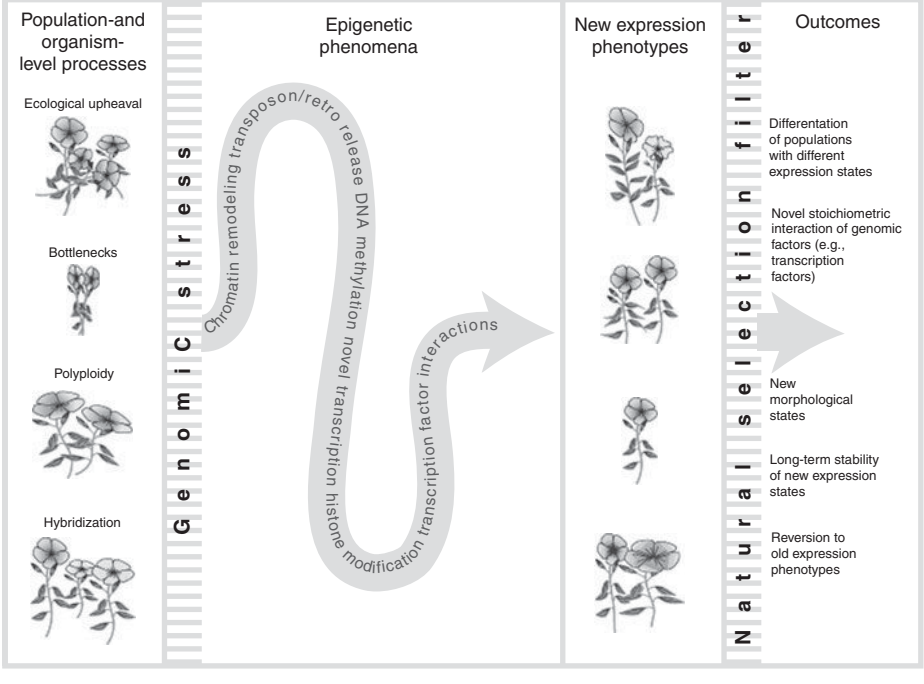


Figure 6.20 Conceptual view of how population level processes (left), following genomic stress and through epigenetic inheritance systems, can lead to new phenotypes, which, after natural selection, can lead to novel evolutionary outcomes (right). From Rapp and Wendel (2005), by permission of New Phytologist.

these phenotypes were produced through partly heritable epigenetic means, mutations that would fix the new developmental pathway would add to the fitness of the new morph. So the phenotypic changes would precede genetic changes and genes would not be initiators but followers. West-Eberhard (2005) called this *genetic accommodation*, a process similar to Waddington's *canalization*.

6.4.2 Epigenetic mechanisms: DNA methylation

The various molecular mechanisms of epigenetic gene regulation are usually classified into three categories: (i) *DNA methylation*, (ii) *histone modification*, and (iii) *small interfering RNA*. These three mechanisms are intimately linked since they influence each other; more than one mechanism may be involved in the transcriptional regulation of a single gene.

DNA methylation is the best described epigenetic mechanism. Normally a certain fraction of the DNA is always methylated, depending on the species, but hyper- or hypomethylation can have strong effects on gene expression. By far the most common target for methylation is the 5-carbon of cytosine. The methylated form of cytosine is often abbreviated *5mC*. A *DNA methyltransferase* (*DNMT*) enzyme transfers a methyl group from S-adenosylmethionine (SAM) to cytosine. All eukaryotes use SAM as a methyl source for this reaction. The methyl group attached to cytosine does not affect the base-pairing, but it protrudes into the major groove of the double helix where it alters the chemical environment around the DNA, influencing binding of proteins.

There are many different methods for the detection of DNA methylation. The genome-wide percentage is measured by complete digestion of the DNA into single nucleotides and characterizing these by a combination of liquid chromatography and mass spectrometry. For sequence-specific analysis, two different experimental strategies are applied, one using sequencing, the other using microarrays. The specificity is based on chemical modification of the DNA using bisulfite, which converts cytosine to uracil but leaves 5mC unchanged.

Another technology applies a combination of methylation-sensitive and methylation-insensitive restriction. Methylation-sensitive restriction is done by means of endonucleases which cleave the DNA only at sites that are not methylated. Restriction is followed by PCR (*amplification of intermethylated sites*, *AIMS*) to display the fragments with methylated cytosines at the restriction site. Still another set of techniques is based on immunoprecipitation with anti-5mC antibody (*methylated DNA immunoprecipitation*, *MeDIP*) or on proteins that bind to methylated CpG sequences (e.g. *methylated-CpG island recovery assay*, *MIRA*). A complete discussion of the large variety of methods, which are all in rapid progress, falls beyond the scope of this book.

In mammals, DNA methylation is targeted towards cytosines in CpG dinucleotides. Such sequences occur everywhere in the genome, and reach a high frequency in so called *CpG islands*. As we saw in Section 6.3.2 these are stretches of DNA of more than 500 bp with a high frequency (more than 55%) of CpG dinucleotides. Such islands are often found in the 5′ region of genes and are usually involved in the binding of a transcription activation complex, especially in so-called *CpG island promoters*. Interestingly, while CpG sequences are the main targets for methylation, CpGs within CpG islands remain free of methylation. In plants, however, DNA methylation is not restricted to CpGs; cytosines are methylated in many sequence contexts; extensive methylation across the whole genome is found in non-CpG DNA.

The percentage of DNA methylation varies enormously between species. In plants more than 30% of the cytosines may be methylated, for mammals and birds this rate is about 5%, fish and amphibians have about 10%, insects and other invertebrates no more than 3%, often less. In most animal genomes, DNA methylation is quite heterogenous: there are large domains of methylated DNA, separated by similar domains of non-methylated DNA, however, in vertebrates methylation is distributed more evenly over the genome (Bird 2002).

In mammals three different DNA methyltransferase genes, *Dnmt1*, *Dnmt2*, and *Dnmt3* are present. DNMT1 is mainly involved with variable, environment-dependent methylation. It also ensures the maintenance of methylation marks across cell divi-

sions. This is aided by the fact that it has a tendency to methylate one strand of DNA if the other strand is already methylated. If there is epigenetic inheritance across meiosis, this also involves DNMT1 activity. DNMT3 is mainly involved with *de novo* methylation during the period of early development. The function of DNMT2 is not clear, although it is highly conserved throughout the animal kingdom. It might be involved in functions other than DNA methylation, for example methylation of tRNA.

Which factors direct the methylation of DNA? There must be some kind of guiding principle, because, depending on the cell type, only specific genes are silenced while others remain active. There is a suggestion that the absence of transcription factors such as Sp1 or enhancers attracts DNA methylation. In that case methylation would be a secondary event and the gene would have to be already inactivated by other means. Another idea is that DNA methylation is targeted by small RNAs (see below).

If not continuously maintained, DNA methylation is lost through two processes known as *passive* and *active demethylation*. Passive demethylation occurs during DNA replication, when the methylation marks are not copied to the newly formed strand. After several cell divisions the methylation mark is lost through dilution. Active demethylation involves enzymatic excision of methylcytosine from the DNA and replacement by unmethylated cytosine, a kind of DNA repair. The enzymes involved are only described in detail for *Arabidopsis* and are not yet well characterized for animals.

In vertebrates, DNA methylation plays a crucial role in development. Suppression of *Dnmt1* and *Dnmt3* in mice causes embryonic lethality. However, many invertebrates do not depend on DNA methylation for their development. Interestingly, the two genetic model species, *Drosophila melanogaster* and *Caenorhabditis elegans*, hardly show any methylation in their genome. How these species can successfully complete development without a system that is so crucial in vertebrates remains enigmatic to date. *C. elegans* lacks all three *Dnmt* genes while fruit flies and mosquitoes have only *Dnmt2*. However, the silk worm, *Bombyx mori* has both *Dnmt1* and *Dnmt2*, while the honeybee and the sea anemone *Nematostella vectensis* have the full set of *Dnmt*s (Cañestro *et al.* 2007). This implies that the ancestor of all eumetazoans had already three *Dnmt* genes and some of these genes were lost multiple times in various protostome lineages. One aspect related to the loss of DNA methylation in fruit flies could be that their development is very short and rather atypical for insects in general. The flour beetle, *Tribolium castaneum* an upcoming model for evo-devo, has a much more characteristic developmental pattern.

Also in the deuterostomes, contraction of the *Dnmt* toolkit has occurred. Comparative genomics has revealed that sea urchins (Echinodermata), *Amphioxus* (Cephalochordata), and *Ciona* (Urochordata, Ascidicaea) all have a complete set of *Dnmt*s, however, the urochordate *Oikopleura* (Larvacea) has only *Dnmt2* (Cañestro *et al.* 2007). This is the more striking since the morphological development of ascidians and larvaceans is very similar. This discrepancy between genetic divergence and conservation of body plans is called the *inverse paradox* of developmental biology.

In plants the epigenetics machinery is similar to animals and partly uses orthologous genes. *Arabidopsis* has three methyltransferases: *MET1* (*METHYL TRANSFERASE 1*) which is a homologue of *Dnmt1* and likewise is involved in maintenance methylation in CpG islands, *DOMAINS REARRANGED METHYLASE 2* (*DRM2*), a homologue of *Dnmt3*, and *CHROMOMETHYLASE 3* (*CMT3*). The latter two proteins mainly target non-CpG sites. *CMT3* is a plant-specific gene, without a known homologue in animal genomes. Similar to the situation in vertebrates, plant DNA methylation is very important for normal development. However, *CMT3* and *MET1* are also especially important for inactivation of transposable elements. Double knock-outs of *CMT3* and *MET1* in *Arabidopsis* release a family of transposable elements, the CACTA family, from suppression. Indeed, one of the theories on the origin of DNA methylation argues that it evolved specifically to inactivate mobile genetic elements. In this theory, the system was recruited only later in some lineages to guide development.

The first comprehensive genome-wide DNA methylation map was developed for *Arabidopsis thaliana* (Zhang *et al.* 2006). These authors detected

well over 25 000 regions of the *Arabidopsis* genome that were significantly methylated, covering nearly 19% of the whole DNA sequence. Extensive methylation was found in each chromosome, especially in and around the centromeres. There was a very good correlation between the presence of repeat elements in a genomic segment and the degree of DNA methylation. On a genome-wide scale, the distribution of DNA methylation reflected the distribution of heterochromatin, which contains a lot of transposons and other repetitive sequences (Fig. 6.21).

Outside the repetitive DNA, methylation was mostly found in genes with no known expression and in pseudogenes. Known pseudogenes had a much higher degree of methylation than expressed genes (Fig. 6.21). This confirms that DNA methylation is often associated with transcriptional silencing, although it may not be always the reason for it. However, the greatest surprise from the Zhang *et al.* (2006) paper was that methylation is also very abundant in the coding region of genes. Around one-third of the *Arabidopsis* genes were found to be methylated in the coding region but not within their promoter. The authors called this *body methylation*. Methylation was biased towards the middle of the gene and declined towards the 3′ and 5′ end of an ORF (Vaughn *et al.* 2007). In addition, introns tend to have a lower degree of methylation than exons (Chodavarapu *et al.* 2010). The reason for these patterns is still elusive, but it might have to do with the position of DNA on a nucleosome. DNA methyltransferases preferentially target DNA bound to nucleosomes and nucleosome-bound DNA is enriched in exons (Chodavarapu *et al.* 2010).

Looking at the expression of genes classified as 'body methylated', 'promoter methylated', and 'unmethylated', the body-methylated genes were found to have the highest expression, followed by the unmethylated genes, while the promoter-methylated genes had the lowest expression. So, in sharp contrast to transposable elements, in which methylation is obviously associated with silencing, methylation of genes is correlated with expression, rather than with silencing. In fact Vaughn *et al.* (2007) suggested that DNA methylation in *Arabidopsis* does not play an active role in regulating gene expression.

Arabidopsis studies have also revealed a high level of epigenetic diversity between populations. Vaughn *et al.* (2007) tested 18 genomic loci in 96 different accessions from two ecotypes. When each gene was scored 'methylated' or 'unmethylated', a more or less random pattern appeared. In most accessions 9 of the 18 loci were methylated, but in two accessions none of them were. Two of the genes were not methylated in any of the accessions, but none of the genes were methylated in all of them. In addition, there was no consistency between the clustering of accessions according to methylation pattern and the phylogenetic tree derived from haplotype sharing.

In an extensive genome-wide study of methylation at CCGG restriction sites, again a significant

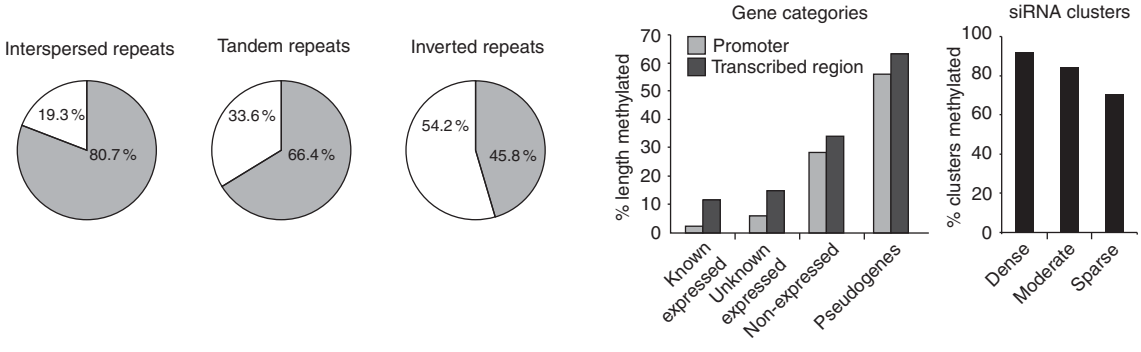


Figure 6.21 Summary of DNA segments found to be methylated in the genome of *Arabidopsis thaliana*. Left: percentage methylation in different categories of repetitive DNA. Right: degree of methylation in four different categories of genes. After Zhang *et al.* (2006), by permission of Elsevier.

level of DNA methylation polymorphism was found, which involved 17 to 19% of the open reading frames and 5 to 13% of the promoters with CCGG sites (Zhang *et al.* 2008, Table 6.4). There was no clear difference in level of methylation when genes and promoters were classified as methylation-constitutive (showing no methylation polymorphism across accessions) and methylation-polymorphic.

The functional significance of these DNA methylation polymorphisms is unclear at the moment. The general picture is that methylation within genes is a more or less random process and the variation across populations has little consequences for gene expression. The level of methylation is generally lower in promoter sequences, and the degree of polymorphism seems to be slightly higher, compared to genes. Extensive promoter methylation could have negative consequences and might be under purifying selection. This may also explain the decrease of genic methylation towards the 5′ end of the ORF.

One of the very first studies reporting on epigenetic variation in wild populations is that of Herrera and Bazaga (2008). These authors studied a long-lived perennial species of violet, *Viola cazorlensis*, which is endemic to the Sierra de Cazorla in Southeast Spain. Using a methylation-sensitive restriction analysis (MS-AFLP), along with normal AFLP, they were able to document the degree of polymorphism within and between 14 populations, not only in regular AFLP markers, but also in methylation-susceptible markers. It turned out that the methylation-susceptible markers showed a very high degree of polymorphism, higher than the regular markers. However, the most striking result of this study was a significant association between ten outlier AFLP loci identified in a previous population genomics study (see Table 6.1) and the differentially methylated loci. Epigenetic variation was strongly linked to adaptive differentiation between populations. The authors suggested that local selection for floral traits (flowering time and flower morphology) acts upon epigenetic markers in correlation with normal genetic variation. The most likely explanation is that epigenetic variation between populations is a direct consequence of DNA-sequence variation (Richards *et al.* 2010).

Ecological epigenetics and population epigenomics are only just beginning. The *Viola cazorlensis* study illustrates that we might find a lot of variation in DNA methylation in natural populations that is related to local adaptation. On the other hand, the genome-wide studies in *Arabidopsis* warn us that the functional significance of variation in DNA methylation is not clear at all. It will not be at all easy to prove that differential methylation of specific loci is functionally related to fitness in specific environmental contexts, however, this is certainly a very promising road to follow.

Table 6.4 Summary of genome-wide constitutive and polymorphic methylation markers at CCGG sites in genes and promoters of wild accessions of *Arabidopsis thaliana*. From Zhang *et al.* (2008)

Number of CCGG sites in:	Constitutive methylation marks	Polymorphic methylation marks
Genes—total	20 609	20 609
Methylated CCGGs in genes	3448 (17%)	432 (13%)
Promoters—total	3246	3246
Methylated CCGGs in promoters	176 (5%)	432 (13%)
Intergenic—total	8276	8276
Methylation CCGGs in intergenic regions	877 (11%)	755 (10%)

6.4.3 Epigenetic mechanisms: chromatin remodelling

Another epigenetic mechanism that is often seen in conjunction with DNA methylation is modification of *histones*. We know that in eukaryotic chromosomes DNA is tightly packed into a complicated structure called *chromatin*, which involves winding of the DNA molecule around histone proteins; 146 bp of DNA are wrapped twice around a protein complex consisting of eight histones, the *nucleosome*. Without the association to proteins and the higher-order winding structure, it would be impossible to pack the very long DNA molecule of a eukaryote in the cell nucleus. Depending on chemical modifications, nucleosomes can be packed closely together, in which case the chromatin is said to be closed and

the DNA region is inaccessible to transcription (*heterochromatin*); alternatively, histone proteins may interact loosely with each other in which case the chromatin is open (*euchromatin*) and the DNA region is active.

In eukaryotic chromosomes, there are five different histone proteins designated H1, H2A, H2B, H3, and H4. They are all small globular molecules with a net positive charge which facilitates their binding to the generally negatively charged DNA. The eight histones in a nucleosome are built of two tetramers, one consisting of two H3 and two H4 proteins, the other consisting of two H2A-H2B dimers. H1 is a linker histone which is associated with the 50 bp of DNA separating each pair of nucleosomes. In addition to histones, other proteins are often found to be associated with chromatin, including the so-called *high mobility group proteins* (*HMG proteins*). These proteins are very different from histones, as they do not have the positive charge and dense structure of histones and are not present in the same amounts in different cell types.

The amino acid sequences of histones are extremely well conserved throughout the eukaryotes. For example, there is only one amino acid difference between H4 of mammals and plants. Within a species, however, different variants of histones may exist. For example the human genome has a family of twelve different H3 encoding genes, which are grouped into three clusters. Histone variants may replace each other as part of a functional change in gene regulation, for example, in active chromatin H3 is often found to be replaced by H3.3, and therefore H3.3 is seen as a marker for gene activity.

In addition to replacement of histones by different variants, the main chromatin regulatory mechanism is enzymatic modification of the histones. The N-terminus of the histones extends like a tail beyond the nucleosome and may undergo a variety of covalent binding processes. Different amino acids in the tail may be subject to different modifications. For example lysine residues may be acetylated, methylated, ribosylated (binding of ADP-ribose), ubiquitinated (binding of ubiquitine), or sumoylated (binding of *small ubiquitine-like modifier* (SUMO) protein); arginines can be methylated and acetylated, while serines and threonines can be phosphorylated. Another layer of complexity is added by the fact that some amino acids can be modified in one, two, or three different places. Although most of the modifications occur on the N-terminal part of the histone, there are some exceptions, for example ubiquitination of the C-terminal tails of H2A and H2B, and acetylation as well as methylation of the globular domain of H3 (Fig. 6.22).

It is obvious that there is an immense complexity of information in these histone modifications, many of them having functional consequences for gene transcription. The total of histone modifications in a piece of chromosome is sometimes called the *histone code*. The histone code is often written in short-hand, for example H3K4me indicates a histone H3 protein of which the lysine residue (K) at the fourth position in the protein is methylated. Figure 6.22 gives an overview of the gamut of possibilities in the histone code (Bhaumik *et al.* 2007).

Addition of negatively charged groups to histones tails, such as acetate to lysine, neutralizes the positive charge of the histone, stabilizes the physical structure of nucleosomes, and contributes to maintenance of euchromatin. Conversely, deacetylation and methylation of aminoacids may lead to condensation of chromatin and formation of heterochromatin. In Section 4.3.4 we saw an example of gene regulation by histone modification: *FLOWERING LOCUS C* (*FLC*) can be inactivated by histone deacetylation; inactivation of *FLC*, required for flowering, is part of the stimulatory effect of vernalization on bolting in *Arabidopsis*.

Histone modifications are conducted by enzymatic activity. There are four main families of enzymes involved, which are named in a straightforward way according to their function: *histone acetyltransferases* (*HATs*), *histone deacetylases* (*HDACs*), *histone methyltransferases* (*HMTs*), and *histone kinases*. These enzymes often operate as part of larger complexes. For example, in the case of *FLC* inactivation by vernalization, a protein encoded by the *FLD* locus is an integral part of the HDAC; suppression of *FLC* therefore requires activity of *FLD* (Section 4.3.4).

It is not yet clear how histone modifications are targeted and maintained across cell divisions. How

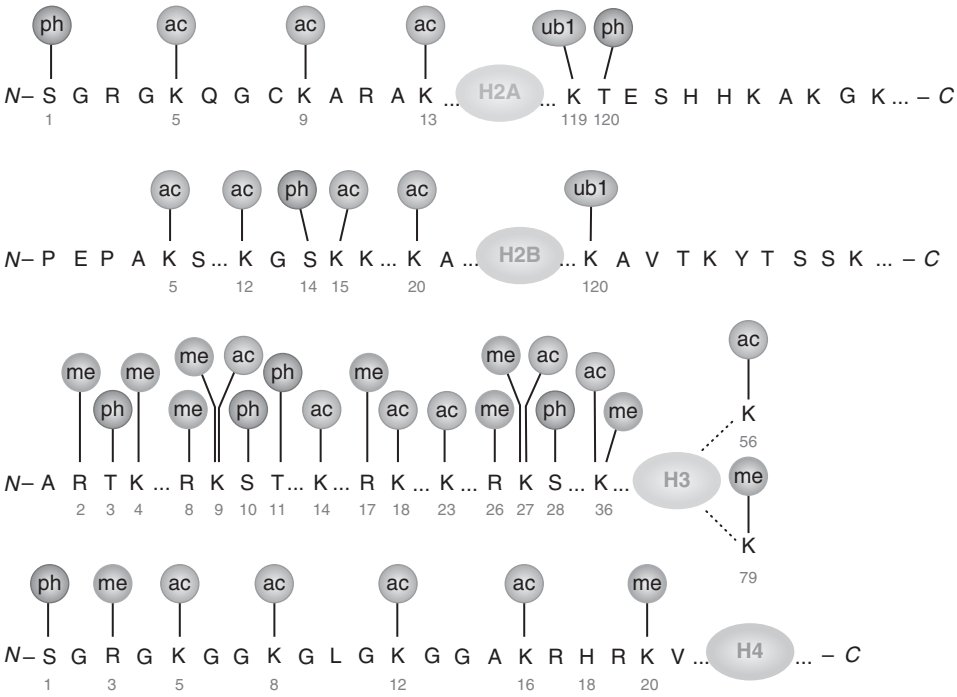


Figure 6.22 Overview of known histone modifications. The N- and C-termini of H2A, H2B, H3, and H4 are shown, with the amino acids indicated by their single-letter codes (e.g. S = serine, K = lysine, etc.), and the various substituents which may covalently bind to these amino acids: ph, phosphate; ac = acetate; ub1, ubiquitine; me, methyl. From Bhaumik *et al.* (2007), reproduced by permission of Nature Publishing Group.

can environmental or developmental factors induce histone modifications at specific regions of the chromosome, causing a desired activation or inactivation of genes? In addition, if histone modifications determine the phenotype of a cell, how is that information transmitted to daughter cells during mitosis? In the process of DNA replication, the nucleosomes disassemble from the DNA and old and new histones are divided randomly across the daughter cells. The fact that DNA is physically separated from the histone proteins during one phase in the cell cycle presents a severe challenge to the idea of chromatin inheritance.

One possible solution is to assume that histone modifications are coupled to histone variants (Henikoff *et al.* 2004). In fission yeast ATP-dependent protein complexes have been discovered that can specifically replace one histone variant (e.g. H2A), by its variant (e.g. H2A.Z, which is inserted near silent chromatin to prevent its spread). It has also become clear that this replacement machinery is replication-independent, that is, it is distinct from the complex that assembles nucleosomes at the replication fork. A model suggests that regions with a high density of specific histone variants, for example H3.3, recruit histone remodelling complexes which remove H3 inserted during replication, and replace them by H3.3, thus maintaining a region of actively transcribed DNA across cell division.

There is also an interaction between DNA methylation and histone modification. Methylated DNA attracts histone deacetylase complexes, and so DNA methylation may lead to chromatin condensation. Conversely, chromosomal regions with modified histones are targets for DNA methylation. The two epigenetic mechanisms, DNA methylation and histone modification, seem to reinforce each other and maybe such a positive feedback loop is a necessary element of the epigenetic gene regulatory mechanism. However, to make it still more complicated,

different histone modifications may also influence each other in a negative way. For example, H3K9me prevents acetylations in H3 and H4 such as H3K14ac and H4K5ac.

Recently a new mechanism of epigenetic modification has been discovered, the action of small, non-coding interfering RNAs (siRNAs). How these RNAs regulate chromatin structure and cause gene silencing has been investigated most extensively in fission yeast, *Schizosaccharomyces pombe* (Grewal and Jia 2007). Unlike budding yeast, fission yeast has large heterochromatic regions which, combined with its small genome, make it a suitable model for the study of chromatin structure.

One of the best studied processes is *RNA-induced transcriptional silencing* (*RITS*). Small regulatory RNA molecules are generated by cleavage of double-stranded RNAs which originate from exogenous sources, such as viruses and mobile genetic elements. The cleavage is carried out by a ribonuclease enzyme called DICER. The siRNAs are then included in a RITS complex, which binds to target DNA matching the sequence of the siRNA. So the sequence of the siRNA provides specificity in the location of RITS complex. Formation of the RITS complex recruits histone modification enzymes and DNA methylation, leading to the condensation of chromatin and gene repression.

Many of the pathways of chromatin regulation are conserved across the eukaryotes, however, this does not imply that chromatin condensation and decondensation have the same functions over all species. Chromatin remodelling is an extremely complicated process with many facets. In fission yeast, heterochromatin formation is targeted to repetitive structures in the DNA that are rendered recombinationally inert. The main purpose therefore seems to lie in protecting the cell from mutagenic transposition events. However, the association of chromatin condensation with repetitive DNA is also used to organize vital chromosomal structures such as telomeres and centromeres, which consist of constitutively condensed chromatin. When comparing species it appears that different species emphasize different aspects of chromatin regulation (Grewal and Jia 2007). This might well reflect different *cooptation* events: the same molecular pathway has been recruited by different species for slightly different purposes, for example the same histone modification might be involved in gene silencing in one species, and in gene activation in another species, depending on the chromosome context.

The various histone modification mechanisms, and the tremendous complexity of interactions between them, are potentially powerful sources of phenotypic plasticity and natural selection. However, to date no studies have been published in which natural variation of histone modification has been documented, let alone placed in an ecological context. An extremely large and unexplored field of research is awaiting discoveries by ecological genomicists.

6.4.4 Morphological variation and developmental adaptation

The various epigenetic mechanisms that we discussed in the previous sections all contribute to the variation of shape and form that is so striking in the natural world. The link between DNA and morphological diversity is studied by the science of *evolutionary developmental biology*, *evo-devo* (Carroll *et al.* 2005; Gilbert and Epel 2009). Genomics has now provided an extensive list of 'toolkit' components that are used in developmental processes to generate a functional phenotype from a zygote.

Two important molecular principles have emerged from evo-devo research. Firstly, the same genetic machinery can be coopted by different species to serve different functions; this explains why species can share the same toolkit elements but still differ greatly in their general morphology. For example, some genes in a wing-expressed sequence tag library for the butterfly *Bicyclus anynana* are not associated with wing development in *Drosophila* and apparently have different developmental functions in different species (Beldade *et al.* 2006).

Secondly, the same morphogenetic process can be regulated by distinct genetic pathways. This is illustrated by a classical developmental process, morphogenesis of the egg-laying structure (vulva) in nematodes. Studies have focused on vulva development of *Pristionchus pacificus*, a nematode whose genome has been recently sequenced and which in

general morphology and life cycle is comparable to *C. elegans*. However, vulva development in *P. pacificus* requires different signalling pathways and other cell determination mechanisms compared to *C. elegans*, despite the fact that the final structure is similar (Sommer 2009).

The evolutionary flexibility of developmental processes is also illustrated by a comparison of *Tribolium castaneum*, the red flour beetle, with *Drosophila melanogaster*. The complete genome of *Tribolium* was published in 2008 (Tribolium Genome Sequencing Consortium 2008), but the beetle was already a model for developmental biology in insects for many years. *Tribolium* develops by a process called a *short germ band embryogenesis*, in which subsequent segments are added from a posterior growth zone. This type of development is similar to basal arthropods, such as millipedes, and also resembles the proliferation of segments in vertebrates. However, it is very different from embryogenesis of *Drosophila* in which all segments form simultaneously during the blastoderm stage. Apparently, the *Drosophila* lineage underwent considerable evolutionary change, maybe related to the need for rapid development in temporary habitats (rotting fruit), leading to many apomorphies and traits less typical for insects and other animals.

In connection with the diverging proliferation of segments, the extra-embryonic membranes of the two insects are also very different. *T. castaneum* has two membranes, the *serosa* surrounding the whole embryo and the *amnion* covering the embryo on the ventral side. However, *Drosophila* has a single membrane, *amnioserosa*, which covers the embryo only partly on the dorsal side. Two genes are involved in the differentiation of the serosa in *Tribolium*, *Tribolium castaneum zerknüllt 1* (*Tc-zen1*), and *Tc-zen2*. These genes stem from a recent duplication which has led to functional differentiation: *zen1* specifies the differentiation of the serosa, while *zen2* determines the fusion between amnion and serosa, which is necessary for dorsal closure. It is assumed that *zen* is derived from a class of *Hox3* genes, which originally specified the head region. In *Drosophila*, the head region is specified by a gene called *bicoid*. There is a great similarity between the phenotypes caused by loss of *bicoid* in *Drosophila* and loss of *zen1* in *Tribolium*, which is essentially due to the fact that both the head region and the serosa are situated at the very front part of the embryo and are specified very early in development. The switch to quick development and reduction of the serosa membrane in higher dipterans has caused a change in the embryonic fate map. Because *zen* expression became less important, this provided a favourable starting point for the evolution of *bicoid* in *Drosophila* (Van der Zee *et al.* 2005).

A number of different species are used in evo-devo research, however the basic concepts are studied in a limited number of them. Sommer (2009) argued that evo-devo models should be chosen in evolutionary vicinity to classical model species. For example, *Tribolium*, *Nasonia*, and *Daphnia* might be evo-devo models matching *Drosophila*; *Antirrhinium* (snapdragon) can be considered an evo-devo model in the vicinity of *Arabidopsis*. By choosing related species that are distinct but still share many homologies in their evo-devo toolkits, the changes in development causing morphological diversity can be much better studied than in a set of completely unrelated species.

The future of evo-devo research is seen in a synthesis with population genetics (Müller 2007; Sommer 2009). It has been commonplace to argue that the theoretical framework of the modern synthesis, with its emphasis on allele frequency change as a basis for evolution, has completely neglected the issue of development. However, this statement is unfair as evo-devo is not in conflict with population genetics theory, it just adds another level of explanation. By joining evo-devo and population genetics the reach of evolutionary theory is expanded in a way completely compatible with the modern synthesis. Conversely, such a joining of forces also implies that the issues of neutrality and adaptativeness of molecular change, as discussed at length above, also apply to developmental processes. Some authors (e.g. Lynch 2007b) do not exclude the idea that the neutrality argument also applies to the origin of organismal complexity. There is no reason why developmental processes underlying the origin of multicellularity, and complexity of shape and form would be immune to nonadaptive evolutionary forces.

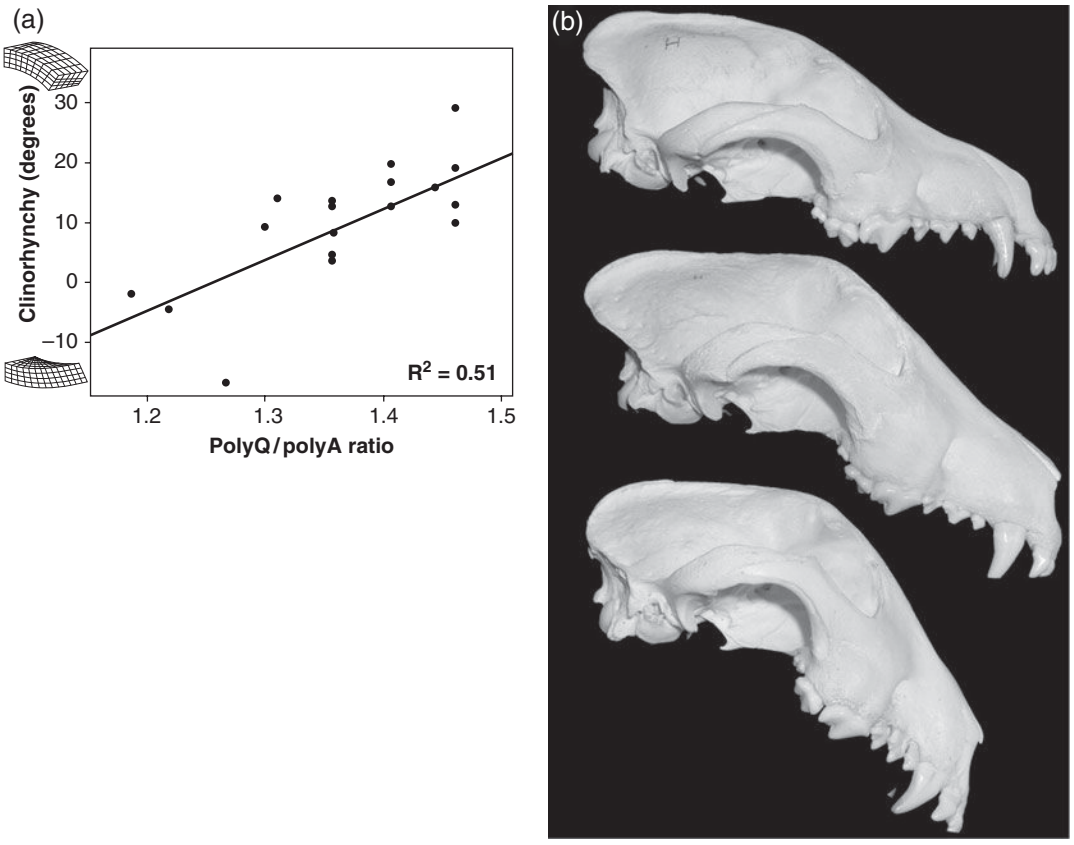


Figure 6.23 (a) Genotyping of a developmental gene, *Runx-2*, revealed that the polyQ/polyA ratio of two SSRs domains in this gene increased from 1.357 in 1931 to 1.462 in 1976. The change of this ratio, being correlated with clinorhynchy of the face across modern dog breeds, illustrates that selection for a downward bending face has acted upon SSR polymorphisms in *Runx2*. (b) Showing pure-bred bull terrier skulls from 1931 (above), 1950 (middle) and 1976 (below). From Fondon and Garner (2004) by permission of The National Academy of Sciences USA.

To study population genetics of developmental complexity we should focus on natural variation in developmental toolkit genes or their regulation. We discuss two recent studies that illustrate how such questions can be approached.

One source of variation in developmental patterns are tandem repeats in transcription factors. We have seen in Section 6.2.3 that polymorphic simple sequence repeats (SSRs) are abundant in many genomes and are often used as markers for population structure. Many SSRs involve non-coding DNA, however, trinucleotide repeats may also be present in protein-coding genes. Such microsatellites may be associated with specific phenotypes and hence may be subject to selection. Recently, Lynch and Wagner (2008) have pointed out that SSRs are particularly abundant in proteins that regulate gene expression. This is rather at variance with the long-held belief that developmental change always relies on changes in gene expression, most notably on *cis*-regulatory change (cf. Hoekstra and Coyne 2007).

Fondon and Garner (2004) made an inventory of SSRs in developmental genes across different dog breeds and humans. In 37 of such genes polymorphisms were found. In addition, there was evidence that in dogs these genes have been subject to selection much more than in humans. This was shown by the extensive polymorphisms among dog breeds and also by the fact that the repeats in the SSRs are

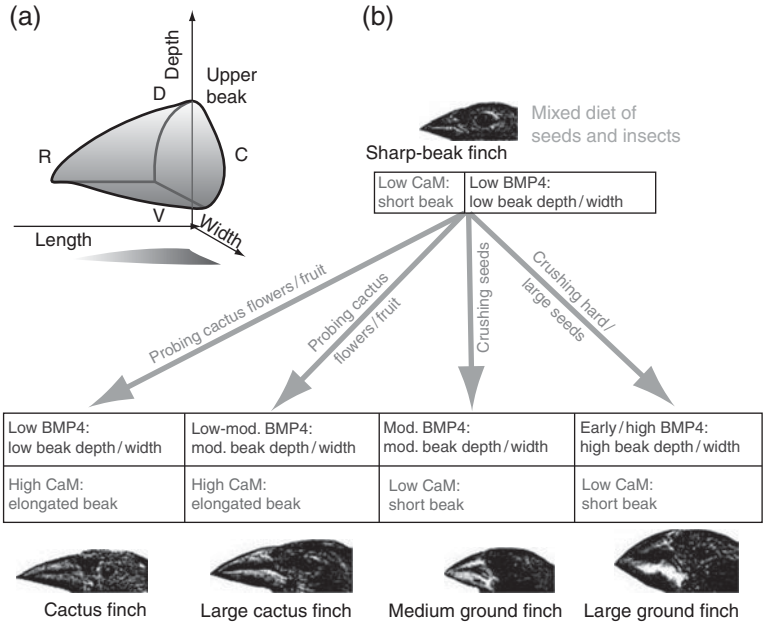


Figure 6.24 Expression differences in *Bmp4* and *CaM* and their combinations, can explain the evolution of four major types of beak morphology in five finches from the Galapagos. *CaM* overexpression is associated with a long beak, while high or early expression of *Bmp4* is associated with large depth and width. From Abzhanov *et al.* (2006), by permission of Nature Publishing Group.

more 'pure', that is contain fewer deviations or interruptions across the tandemly repeated core sequences. Such 'repeat purity' is evidence for selection acting upon length polymorphisms.

In the dog study, two polymorphisms were studied in detail, a polyglutamine and a polyalanine microsatellite in the *Runx2* gene, which encodes a developmentally active transcription factor (*runt-related transcription factor 2*). A historical reconstruction showed that over the years, artificial selection for clinorhynchy (downward bending face) in bull terriers had increased the length of the polyglutamine domain, relative to the polyalanine domain, in *Runx2* (. 6.23).

The second example illustrating an upcoming link between evo-devo and population genetics is due to the Darwin's Galapagos finches, the celebrated example of adaptive speciation and natural selection acting upon beak morphology. Research into the mechanisms of beak development has identified two genes, *bone morphogenetic protein 4* (*Bmp4*) and *calmodulin* (*CaM*) as key factors determining beak evolution. The variation in beak morphology appears to be due to the degree of expression of these genes in the beak primordium of the embryo; *Bmp4* expression affects beak width and depth, while *CaM* expression affects beak length (Abzhanov *et al.* 2004, 2006; Campàs *et al.* 2010).

The role of *Bmp4* was identified in a candidate gene approach, targeting genes known to be involved in craniofacial development. The expression domain of *Bmp4* in beak primordia was correlated with adult beak morphology across five *Geospiza* species, and this suggested the involvement of *Bmp4* in beak width and depth. The suggestion was confirmed by manipulating *Bmp4* expression in chicken embryos which showed similar phenotypes (Abzhanov *et al.* 2004).

However, this could not explain the heavy but elongated beaks typically shown by cactus finches; other candidates among the known craniofacial development genes were not immediately obvious. Therefore a cDNA microarray profiling study was done to screen for differential transcripts associated with long beaks (Abzhanov *et al.* 2006); this identified *calmodulin* (*CaM*) as a candidate. Calmodulin is

a calcium-binding protein which is a key component of the calcium-dependent signal tranduction pathway. It had not been implied in craniofacial development before. Comparison of *CaM* expression in embryos of seven *Geospiza* finches provided a clear suggestion that high expression of CaM in a specific developmental stage promotes the elongated beak morphology as seen in cactus finches. Again, this was confirmed by manipulation experiments in chicken embryos.

Thus, two different developmental proteins can explain beak morphology in Darwin's finches. In principle, beaks can evolve along three axes: length, depth, and width. However, depth and width are closely linked, as they are determined by the same developmental protein, while length can evolve independently of width and depth. The combination of these expression differentials can explain the variation across four main types of ground finch beaks (Fig. 6.24).

The evolution of expression differences underlying avian beak morphology has not yet been traced to specific sequence variation in the DNA. Which genetic changes in the signalling pathways are responsible for differential *Bmp4* and *CaM* expression remains a puzzle to be solved.

6.5 Genomic approaches to variation and adaptation: an appraisal

The genomic perspective on genetic variation, as discussed in this chapter, has revealed tremendous complexity. A distinction should be made between DNA sequence variation and variation in epigenetic markers such as methylation patterns, histone variants, and histone modifications. The epigenetic machinery is of great potential importance to ecologists because it provides a link between the environment and the development of specific phenotypes. However, due to its complexity epigenetics is still little studied by ecologists. Hopefully this situation is going to change soon, since we now have genome-wide methods for screening epigenetic markers. DNA methylation is the obvious mechanism to focus on; our understanding of the histone code, although also of potentially great importance for responding to the environment is not yet complete enough to allow applications in an ecological context.

As to DNA sequence variation, a distinction can be made between substitutions in coding regions and substitutions in non-coding regulatory DNA. The latter can affect gene expression in *cis* or in *trans*. As any one gene can vary its expression by both *cis* and *trans*-acting variation, some of them with large, others with small, effects and many of them interacting with each other, the sources of variation are manifold. As stated in the introduction of this chapter, we will have to look at molecular evolution as a shifting network, rather than as a loose collection of genes changing in frequency. The network perspective is not at variance at all with classical population genetics, however, it places a greater emphasis on interaction effects due to genetic linkage, genomic location, epistasis, and pleiotropy than is usually done in population genetics textbooks.

In this chapter we have seen a few examples in which phenotypic changes could be related to a limited number of genetic factors. However, such cases will be the exception rather than the rule. Adaptation will not come by sweeps alone and the use of statistical methods only aimed at identifying such sweeps may well be missing the point (Pritchard and Di Rienzo 2010). It is much more likely that most adaptive processes, especially those that are important in the environment, proceed by small changes in multiple genes, interacting in a network.

Perhaps the greatest issue in the study of molecular variation is the discussion between the neutralist point of view and the adaptation perspective. Our review of population genomics studies has shown that we have not yet reached a stage in which specific polymorphisms can be pinpointed to alterations of the phenotype that contribute to fitness in a specific environmental context. This is partly due to the fact that many of the polymorphisms revealed reflect geographic isolation and neutral substitutions, rather than adaptive phenomena. The (relatively rare) adaptive changes have to be searched like needles in a haystack of neutral variation. They can only be found with extremely strong screens,

such as SNP genotyping based on next-generation sequencing methodology.

The study of molecular variation relies heavily on statistical analysis of sequence data. We have discussed various statistical tests in this chapter to identify directional and disruptive selection. The testing of alternative evolutionary hypotheses on sequence data is a crucial element in any study of variation and adaptation. However, it should not distract us from functional analysis of the DNA polymorphisms. The final proof for an adaptive change comes from a demonstration of its contribution to fitness and to show this we need to involve biochemistry, physiology, and development, in addition to statistical analysis of DNA sequences.

Will the evolutionary paradigm, as formulated by Darwin and the modern synthesis, change under the influence of the genomics revolution? We are convinced that a shift is indeed occurring, although we argue that this may be a question of emphasis, rather than a complete overturn. The following issues are indicative of such a shift: (i) the network perspective of molecular evolution, placing a much greater emphasis on interactions and negating the simplistic idea that there is a gene for every phenotype; (ii) the realization that neutral processes play an important role in the origins of genome architecture, and most likely also in the origin of new body plans; (iii) the possibility that some aspects of inheritance may involve epigenetic rather than genetic mechanisms; and (iv) the identification of new evolutionary mechanisms such as cooptation and genetic accommodation of alternate developmental pathways.

CHAPTER 7

Integrative ecological genomics

As indicated in Chapter 1, a large part of genomics can be qualified as a science of discovery. An impressive amount of new and often unexpected information comes from sequencing genomes and analysing transcriptomes. In this book we have seen many examples of surprises that have arisen from discovery. There is, however, another side to pure discovery science: sequencing of genomes and cataloguing gene expressions are basically descriptive processes and not hypothesis-driven. We do not argue that descriptive activities are always unscientific. Every science includes an inductive phase; unprejudiced exploration is a wealthy source of new hypotheses. In fact, a large part of ecology can be characterized as mainly descriptive, especially the painstaking work conducted by field biologists who document the whereabouts of species in the wild. The point is, in good ecology this type of research is only one phase in the scientific cycle, and is followed up by hypothesis-driven research, either in the laboratory or in the field again. In the same spirit, molecular biologists are convinced that the discovery science of genomics must be followed up by integrative, hypothesis-driven, new activities. In this chapter we discuss some of the post-genomic, integrative approaches and evaluate their relevance to ecology.

7.1 The need for integration: systems biology

Systems analysis has been a part of ecology for a long time. It originates in the work of Eugene P. Odum, who in his classical book *Fundamentals of Ecology* in 1953 made a plea for considering the ecosystem as a unit in its own right, with characteristics that refer to the whole and that are amenable to investigation in the same way as a physiologist investigates a single organism. Odum (1953) also recognized that an ecosystem had properties that were not measurable at the levels of populations or communities but were characteristic of the system as a unitary whole. Such properties were called *emergent properties*. From the beginning, systems ecology relied strongly on mathematical analysis of energy fluxes and nutrient cycles, using differential equations and computer simulation. Systems ecology remained an integrative but limited part of ecology, next to the much larger fields of community and population ecology, but in the 1990s it saw a renaissance, spurred by research programmes on acid rain and climate change, which called for large-scale approaches (Schindler 1987).

It is interesting to see how around the year 2000 the developments in ecology were mirrored by similar developments in molecular biology, leading to the birth of a new field, *systems biology* (Ideker *et al*. 2000; Kitano 2002). In molecular biology, the immediate cause was a feeling that the trend towards more and more reductionism was reaching its limits and in some respects was even impeding progress (Van Regenmortel 2004; Strange 2005). According to the reductionist paradigm, understanding is to be gained from studying processes underlying, and thus determining, the phenomenon of interest. Therefore biological processes should be deconstructed into their component parts and the physicochemical properties studied to achieve understanding. Reductionism argues that any biological process ultimately finds its foundation in the laws of chemistry and physics. What is missing in this description is that many biological phenomena

do not only depend on the properties of the component parts, but also on other biological phenomena. Gene expression takes place in a certain cellular context, which is moulded by expressions of other genes. Even the expression of a single gene may require gene products from at least 10 other genes; for example, components of the polymerase complex, transcription factors, enhancers, and modifiers. Recognizing the complexity of gene expression, molecular biologists started to analyse the network of interactions between the expression of different genes.

Despite the explosive use of the term systems biology in the biochemical literature since 2004, according to Cornish-Bowden and Cárdenas (2005) not all activities grouped under this heading can be regarded as true systems biology. Systems biology is not to be considered a way of integrating information from diverse components into a model of the system as a whole, it must be seen as *a view on the whole to understand the parts*. So systems biology argues from the whole to the components, not the other way around. Seen in this way, systems biology and reductionism are not necessarily in conflict with each other.

Westerhoff and Palsson (2004) gave an historical perspective on the origin of systems biology. These authors recognized two independent lines of development: one originating in the discovery of DNA, followed by recombinant technology and large-scale sequencing, the other originating in non-equilibrium thermodynamics, followed by Jacob and Monod's work on the *lac* operon and molecular kinetics (Fig. 7.1). The second timeline was dominated by modelling of molecular processes but this was often seen as rather theoretical and not based in 'real' biology due to the lack of adequate data. The quantum leap of data acquisition in the first timeline, brought about by the genomics revolution, caused a sudden convergence of the two developments and marked the birth of systems biology.

The structure of systems biology can also be viewed as an activity in which the two lines of inquiry delineated in Fig. 7.1 are fused into a circular

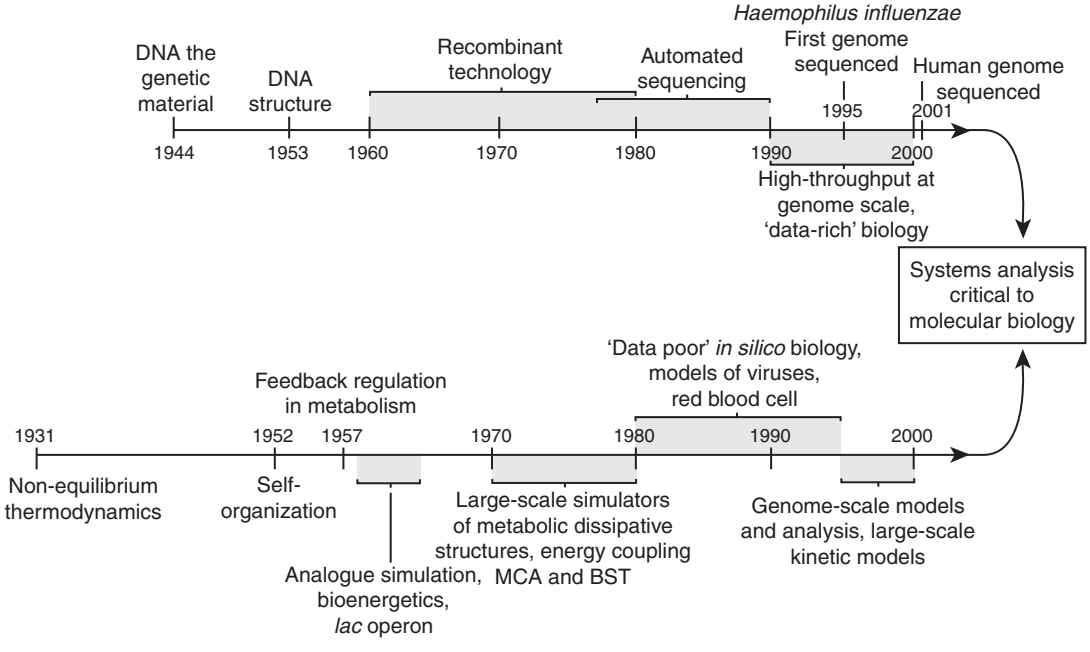


Figure 7.1 Scheme showing how systems biology emerged at the end of the twentieth century from the convergence of two lines of enquiry that had remained separate for several decades. MCA, metabolic control analysis; BST, biological systems theory. From Westerhoff and Palsson (2004), by permission of Nature Publishing Group.

process (Fig. 7.2; Kitano 2002). In this cyclic view, hypothesis-driven modelling, computer simulation, and predictions follow up genomic experiments. This part of the cycle is also called *in silico* biology (Palsson 2000). A tremendously important role is played by systematic perturbation of the genome using high-throughput genetic manipulation. For instance, a collection of deletion mutants is now available for essentially all genes in the genome of *S. cerevisiae* and tools exist for systematic manipulation of numerous genes via different constructs in many strains simultaneously. Analysing the phenotypic consequences of such genetic manipulations provides a vast and rich experimental basis on which hypotheses from computer simulation can be tested.

One way in which the complexity of genomic interactions can be analysed is by considering gene expression as a network. Indeed, a great deal of systems biology is concerned with *network analysis*. A first logical step in such an analysis is to group genes, proteins, and metabolites into functional units. According to the concept of *modularity*, the functions of a cell can be partitioned into a number of modules where membership of a module is defined by a specific task, which is separable from those of other modules. Four different types of functional module may be discerned: (i) physically delineated molecular machines such as the ribosome or the flagellum; (ii) signalling cascades, in which membership of the module is defined by initial binding of a signal molecule such as in insulin/IGF-1 signalling (Section 4.2) or MAPK (Section 5.2); (iii) collections of genes that are all regulated by the same transcription factor, such as the Ah battery (Section 5.2; these are called *transcription modules*); and (iv) networks defined by the processing of a substrate or a group of metabolites, for example glycolysis, Krebs cycle, or nutrient salvage (see Fig. 5.17). A systematic collection of biochemical pathways representing our current state of knowledge of molecular interactions in the living cell is provided by the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (www.genome.jp/kegg).

The greatest advances in metabolic network analysis have been made in *E. coli* and *S. cerevisiae*. One of the lines of investigation focuses on classification of genes according to regulatory modules defined by common responses to environmental conditions. Segal *et al*. (2003) analysed a large number of microarray experiments and proposed a classification of the yeast genome into 50 distinct modules, where each module showed a specific response to a set of conditions, while all the genes of a module were regulated in concert. For example, a 'respiratory module' was defined as consisting of 55 genes, of which 39 encoded respiratory proteins and 6 encoded enzymes of glucose metabolism. The module was primarily regulated by the transcription factor Hap4, and 29 of the 55 genes had a known Hap4-binding site in their promoter. In addition, 32 genes had an STRE element, which is recognized by the stress-responsive transcription factor Msn4 (see Section 5.3.1). This type of classification is an important step towards an integrative understanding of responses to environmental conditions in yeast and eventually also in models of ecological relevance.

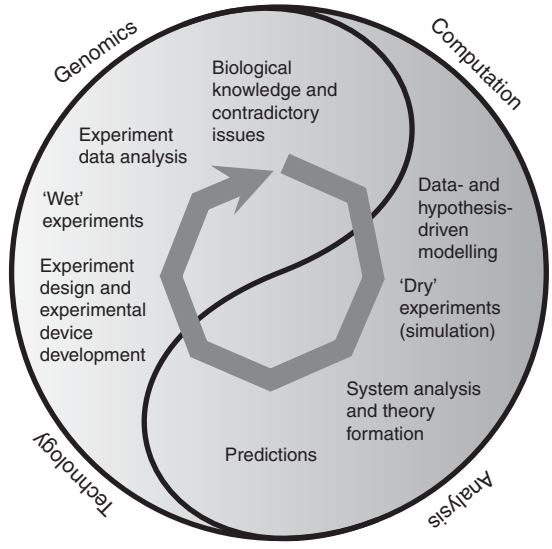


Figure 7.2 Illustrating the structure of systems biology. The cycle of research is characterized by construction of models based on genomics data, followed by computer simulation (*in silico* or dry experimentation), which leads to prediction about the function of specific parts of a genomic network. These predictions are then tested using 'wet' experiments, for example genetic manipulations in the network of interest, which leads to renewed biological knowledge and refinement of models. Reprinted with permission from Kitano (2002). Copyright 2002 AAAS.

In addition to classifying genes according to their transcriptional regulation, systems biologists also map entire sets of genes in a genome (Featherstone and Broadie 2002; Tong *et al*. 2004; Zhang *et al*. 2005). To decide whether two genes interact with each other a technique known as *synthetic genetic array* (SGA) analysis is used. The availability of deletion mutants for many genes in the genome is a crucial element in this technique and this condition is satisfied in yeast. An SGA screen starts with a viable mutant that has a query gene deleted. That mutant is crossed into an array of viable deletion mutants for many other genes to generate double mutants which are isolated automatically and scored for loss of fitness or lethality. Double mutants with reduced fitness are indicative of interaction between the two genes, because if each of the single mutants is viable while the double mutant is not, this suggests that the gene products buffer each other in the wild type. Interactions thus defined are often found between genes belonging to the same gene ontology category, which confirms the validity of the interpretation.

Analysis of genetic networks in yeast has revealed that most of them have a structure characterized as *scale-free topology* (Featherstone and Broadie 2002; Tong *et al*. 2004). This structure implies that the distribution of links over the nodes of a network is not random but biased towards a few highly connected nodes ('hubs'), which participate in a large number of metabolic reactions, while the majority of nodes are hooked on only via a single link. Technically speaking, in a scale-free network the probability $P(k)$ that a node interacts with k other nodes (the *degree distribution* of the network) follows a power law, where $P(k) \sim k^{-\gamma}$. The parameter γ approximates a value of 2.2 for all organisms. When this relationship is plotted on a graph (Fig. 7.3), the horizontal axis represents the degree of interaction that a gene has in the genome (highly interactive genes are on the right), whereas the vertical axis represents the number of genes that is found for each degree. In an

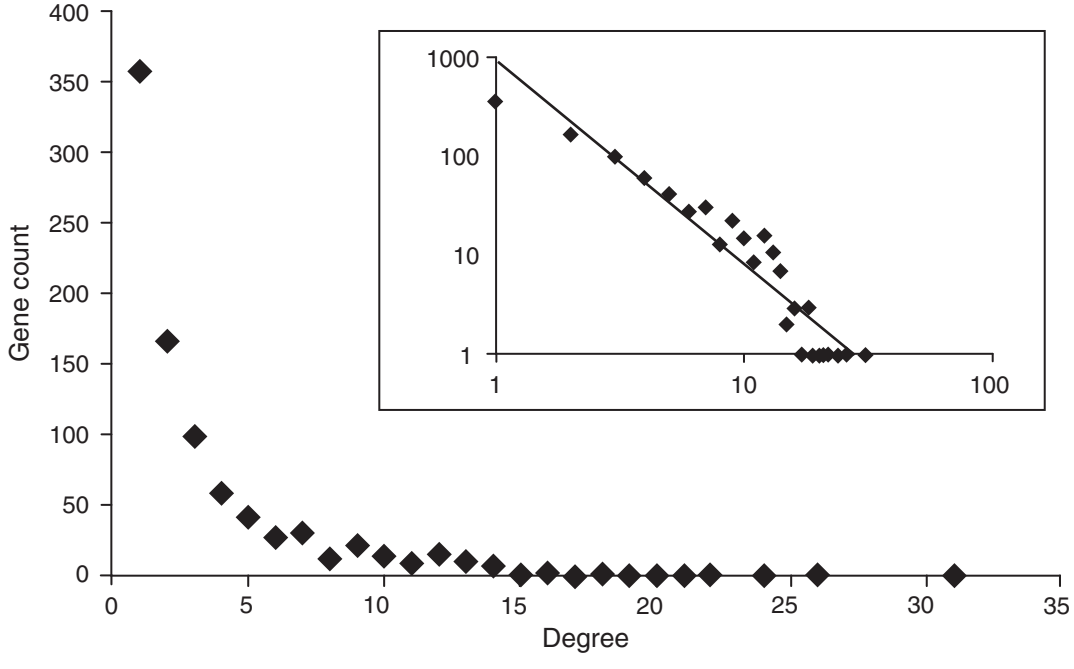


Figure 7.3 Degree distribution of interactions between yeast genes established through SGA analysis. A frequency distribution is shown of the number of interactions (degree) that yeast genes have with other genes. The shape of the distribution is suggestive of a power law, which is confirmed by the linear relationship obtained when the two axes are expressed logarithmically (inset). Networks with a power law distribution of degrees are called scale-free. An example of the topology is shown in Fig. 7.4a. Reprinted with permission from Tong *et al*. (2004). Copyright 2004 AAAS.

SGA screen conducted by Tong *et al*. (2004) 132 different yeast genes were queried and each of these genes was tested for interactions with 4700 other genes. As expected, genes with only one interaction were most abundant but some had interactions with 32 other genes (Fig. 7.3).

The scale-free property of genetic-interaction networks is in conflict with the idea of modularity referred to above. The existence of modules implies some degree of organization and a more or less uniform distribution of links over nodes. These two seemingly opposing tendencies are reconciled in the concept of *hierarchical modularity*, introduced by Ravasz *et al*. (2002). In a hierarchical modular network there are many small, highly connected modules that combine into larger units according to a power law. The hierarchical network has a scale-free topology with embedded modularity (Fig. 7.4c). Ravasz *et al*. (2002) conducted an extensive analysis of the metabolic networks of 43 different organisms and found that in all species investigated there was significant evidence for hierarchical modularity.

Network analysis in yeast is becoming more and more comprehensive. Drawing together information from SGA screens, protein–DNA interaction, and protein–protein interaction, Zhang *et al*. (2005) developed a network involving no less than 5831 nodes (genes, proteins) and 154 659 links between them. Such extremely complicated networks can be analysed for the occurrence of *network motifs*: patterns of interconnectedness between three or more genes that occur more frequently than expected from a random combination of links. Overlying a collection of network motifs is a *network theme*: an interconnected cluster of motifs reflecting a common organizational principle. Network themes can often be linked to a biological process and there are only a few of them in the cell. At a still higher organizational level is defined a *thematic map*, which captures the dynamic relationships between themes.

We have dwelt on network theory to illustrate the approaches that systems biologists are applying in order to understand the metabolism of the living cell. The question arises, should ecological genomics follow a similar path to understand ecological systems and link genomics to ecology? Network analysis is not uncommon in community ecology. Food webs can be considered as networks and ecologists are interested in what properties of food webs contribute to their stability. One such property, discussed by Neutel *et al*. (2002), is the pattern of *interaction strengths* across the web. Interaction strength of a trophic couple is a measure of the influence of one species on the population increase of another species when both are near equilibrium. The negative effect of a predator on its prey is usually larger than the positive effect of a prey on its predator; therefore, interaction strengths are greater for top-down relationships than for bottom-up relationships. Interaction strengths are particularly important when there are loops in the web; that is, paths that return on the same species without visiting other species more than once. The mean interaction strength of all the pairs in a loop is called the *loop weight*. Examination of 104 published food webs revealed that the longer the loop, the lower its loop weight. Mathematical analysis showed that this property contributes greatly to the stability of the network. So, loop weight in a food web can be considered a network motif of a significance comparable to the ones identified by Zhang *et al*. (2005) for metabolic networks in the cell.

We believe that approaches similar in spirit to systems biology should ultimately be adopted to enable genomic answers to ecological questions. That is, a systems approach is needed to link genomics to ecosystem function, to life-history pattern, and to the ecological niche. Chapters 3, 4, 5, and 6 of this book have shown that such links are currently far from complete. When ecological processes are governed by a limited number of signal transduction pathways, as in some of the stress responses discussed in Chapter 5, a network analysis of interactions seems to be a suitable option (see the figures in Chapter 5 showing induction of gene expression by stress). Also, in the cases of nutrient cycles catalysed by well-characterized genetic complexes in microorganisms, a systems approach seems to be feasible. However, in general it is difficult to forecast which type of systems biology is required as an integrative framework for ecological genomics. In addition, we see three main issues that are lacking

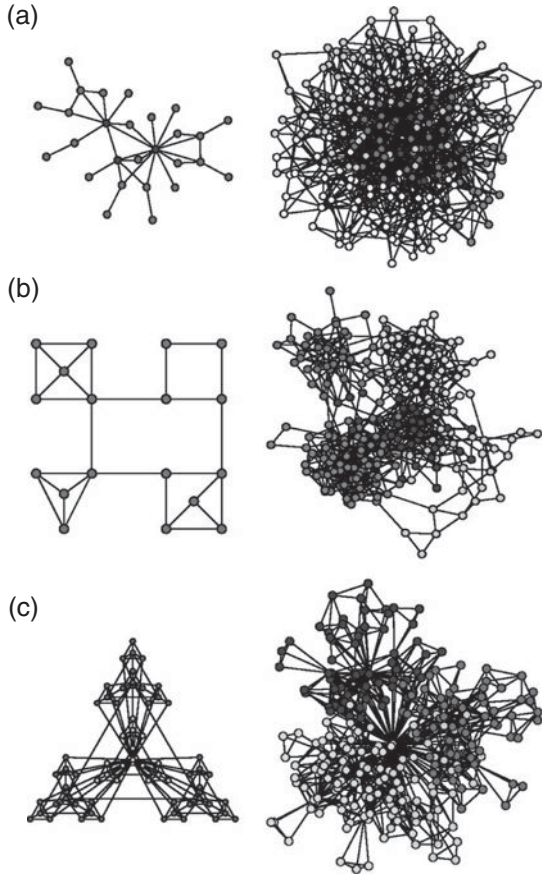


Figure 7.4 Three types of model for complex networks. (a) A scale-free model characterized by a few, highly connected nodes; (b) a modular network consisting of four highly interlinked units, connected to each other by a few links; and (c) a hierarchical network, which combines a scale-free topology with hierarchical clustering. Reprinted with permission from Ravasz *et al*. (2002). Copyright 2002 AAAS.

in the present systems-biology approaches but which are nevertheless crucial for ecology.

Spatial considerations. Space is a very important aspect of ecological analysis. Many processes in communities require some degree of proximity between different organisms (e.g. in the case of syntrophy). Reproduction in animals usually requires physical contact between the sexes; in flowering plants pollen has to travel from anther to pistil and the distance between these organs matters. Stratification of the environment, or other types of heterogeneity, are often crucial for species to coexist with each other. Colonization events can alter the functioning of ecosystems dramatically if the colonizer is an invader and outcompetes local species. All these issues, so obvious for an ecologist, are completely absent from present-day systems biology. It is unclear how genomics and systems biology can contribute to spatial ecology.

Temporal considerations. Like space, time is another important dimension of ecological analysis. Any ecosystem bears all kinds of traces from its prior development; the way in which an ecosystem functions is determined partly by what went before. Historical issues are, for example, reflected in the build-up of the soil profile and the presence of peat accumulated from previous plant growth. If birds for some reason in the past decided to use one piece of land rather than another, this situation may continue for a long time because habitat use is often culturally transmitted or imprinted in the offspring. So, ecologists are accustomed to the fact that some phenomena in nature can only be understood by referring to past events, and that there is a more or less predictable succession of events during the development of an ecosystem. It is unclear how such temporal phenomena can be reconciled with a systems-biology perspective.

Bi-directional interaction with the environment. In the systems-biology treatment of the living cell, the environment of the cell is considered as given; expressed in ecological terms, it is a set of external conditions that are not altered by activities of the cell. In real ecological systems, the environment is altered significantly by organisms, for example by gas exchange, by consuming resources, or by altering physical structure (e.g. burrowing by earthworms). Including such interactions in a systems-biology model would imply that resources and nutrients are considered dynamic variables that can be depleted or otherwise altered by the organism. Such bi-directional interactions are not yet part of systems-biology analyses.

7.2 Ecological control analysis

In addition to network theory, which emphasizes topology and structure, systems biology also deploys a set of quantitative tools, aiming to capture the

dynamics of transcriptome, proteome, and metabolome in mathematical terms. One such approach developed in the 1970s goes under the name of *metabolic control analysis (MCA)*. The foundations of MCA theory were laid by Henrik Kacser in 1973 (Fell 1992; Kacser and Burns 1995). The theory focuses on the flux of metabolites through a biochemical pathway, where a certain substrate undergoes successive modifications catalysed by enzymes. An example is the well-known glycolytic pathway, which involves 10 different catalytic steps starting with phosphorylation of glucose and ending in the formation of pyruvate. Under normal physiological conditions, the concentration of each intermediary metabolite will be constant, while there is a constant flux of glucose down to pyruvate. It is the flux of substrate that drives the metabolism of the cell, not the concentration. The question asked in MCA is, which enzyme exerts the greatest control over the flux?

In the formularium of MCA, the flux of substrate through the pathway is designated by J. The control of an enzyme i over the flux is expressed as the increase of J that would be brought about by a small increase in enzyme concentration, e_i. If an enzyme exerts a large control, the flux will increase strongly with a small increase of e_i; that is, the derivative of J with respect to e_i will be large. The *flux-control coefficient* C_i is therefore defined as the partial derivative of J with respect to e_i, relative to J and e_i themselves, under steady-state conditions:

$$C_i = \frac{\partial J e_i}{\partial e_i J} = \frac{\partial \ln J}{\partial \ln e_i}$$

Flux-control coefficients are dimensionless constants that characterize the action of an enzyme in the pathway. However, they should not be considered properties of an enzyme as such, because their value depends on the state of the whole pathway. Control coefficients can only be measured while the whole pathway is intact; this can be done by measuring the flux as accumulation of the end product in cells with experimentally manipulated enzyme concentrations, for example using specific inhibitors. Another approach is to use genetic means, for example by comparing homozygotes with heterozygotes, by mutating a gene to downregulate its activity, by introducing a plasmid carrying an extra copy of the gene, or by replacing the promoter of the gene by an artificial promoter that can be modulated by an external factor. The rise of DNA technology has greatly expanded the possibilities for manipulating enzyme activities and has allowed measurements of flux control that were difficult to achieve with earlier biochemical methods (Jensen *et al*. 1995).

An interesting property of flux control coefficients in a metabolic pathway is that, leaving aside some special cases, the sum of their values over the whole pathway equals unity:

$$\sum_{i=1}^{n} C_i = 1$$

where n is the number of enzymes in the pathway. This so-called *summation theorem* of MCA holds independent of the structure of the pathway; it is valid for linear pathways such as glycolysis, but also for non-linear pathways, including those with feedback, coupling, and branching.

The summation theorem, which was proved by Kacser and Burns (1995), implies that control is not necessarily attributable to a single enzyme. The idea that there is always one enzyme in a metabolic pathway that is rate-limiting is not supported by theory or biochemical experiments. Instead MCA leads to the concept of *distributed control*: all enzymes in a metabolic pathway exert some degree of control over the flux. It is possible that all enzymes take an equal share in the control, but the most common situation is that several enzymes in a pathway contribute only little to flux control. This is supported by evidence from heterozygotes that carry a defective mutant allele; in such genotypes the activity of the enzyme involved may be reduced to 50% in the case of full recessiveness, but still the flux of metabolites through the pathway is often hardly affected.

A second important concept in MCA concerns the way in which enzyme activities depend on the substrate concentration. This is expressed in the *elasticity coefficient* ε, which is defined for any enzyme i as the partial derivative of its reaction rate v_i with respect to the substrate, S, again normalized to the substrate concentration and the rate:

$$\varepsilon_i = \frac{\partial \upsilon_i}{\partial S} \frac{S}{\upsilon_i} = \frac{\partial \ln \upsilon_i}{\partial \ln S}$$

Elasticity coefficients are determined by the kinetic properties of the enzymes. In contrast to the flux-control coefficient, they are local properties and can be studied using the enzyme when it is isolated from the pathway. Flux-control coefficients lose their meaning when the enzyme is isolated, but elasticity can still be studied *in vitro*.

MCA theory further learns that there is a relationship between elasticity coefficients and flux-control coefficients. This is a consequence of the fact that under steady-state conditions the flux neither accumulates nor oscillates in the pathway. If we consider two adjacent steps, 1 and 2, in a metabolic pathway, we have two elasticities, ε_1 and ε_2, and two flux-control coefficients, C_1 and C_2. The relationship between these quantities is:

$$C_1\varepsilon_1 + C_2\varepsilon_2 = 0$$

This relationship is known as the *connectivity theorem*. For a formal proof the reader is referred to Kacser and Burns (1995). The relationship can be generalized from two adjacent steps to the whole metabolic pathway, in which case the connectivity theorem takes the form of a matrix equation.

The connectivity theorem is one of the strongest results of MCA, since it implies a formal link between enzyme kinetics and flux control. If the elasticity coefficients of all enzymes in a pathway are known, their flux control coefficients can be derived by applying the connectivity theorem. Derivation of flux control from elasticities is called *forward control analysis*. Equally interesting is *reverse control analysis*, which argues from control coefficients to elasticities. Westerhoff *et al*. (1994) showed that through a single matrix inversion step the *in-situ* elasticities of the enzymes in a pathway can be obtained from properties of the pathway as a whole, the flux-control coefficients.

MCA is interesting because it offers a powerful framework with potential application to ecology. A link between MCA and ecology is developed in *trophic control analysis* (Getz *et al*. 2003). In this approach the MCA methodology is applied to food chains rather than biochemical pathways. Trophic chains are different from biochemical pathways in one crucial aspect: the flux process is not conservative. Due to growth, respiration, excretion, and mortality in food chains, biomass is lost from the flux of matter; control analysis has to take this into account. Getz *et al*. (2003) developed an MCA-like food web model by which it is possible to investigate which trophic groups exert the largest control over the biomass flux. The aim of the study was to test the trophic cascade hypothesis, which holds that predators exert control over their prey at the top of the food chain in such a way that the top-down control trickles down to lower trophic levels and so also limits populations at the base of the food web. Trophic control analysis demonstrated, however, that control is distributed over several different levels of the trophic chain, rather than residing in one particular place.

Another type of application of MCA to ecology goes under the name *ecological control analysis*. This type of analysis stays close to MCA by focusing on the flux of some material through a community in the absence of significant biomass production from the flux. This may hold as an approximation for microbial communities under anaerobic conditions in which growth is extremely slow and almost all of the carbon is dissimilated. As a consequence enzyme concentrations and reaction stoichiometries can be assumed to be in steady state and are conserved. Röling *et al*. (2007) considered a simple syntrophic community consisting of three functional groups of microorganisms, degrading a specific organic substrate by fermentation (Fig. 7.5). The fermenting microorganisms produce acetate and hydrogen but they are inhibited by these products and cannot grow when acetate and hydrogen accumulate in their environment. In a syntrophic community, fermentation products are utilized by terminal electron-accepting microorganisms such as iron reducers, sulphate reducers, and methanogens (see Section 3.3). It is assumed in the model that acetate and hydrogen are used by different microorganisms (Fig. 7.5).

Röling *et al*. (2007) estimated numerical values for the elasticity coefficients in this system and showed that in contrast to current views, degradation fluxes are not always limited by a single functional group, but can be controlled by several groups simultaneously. The model was used to explore under what

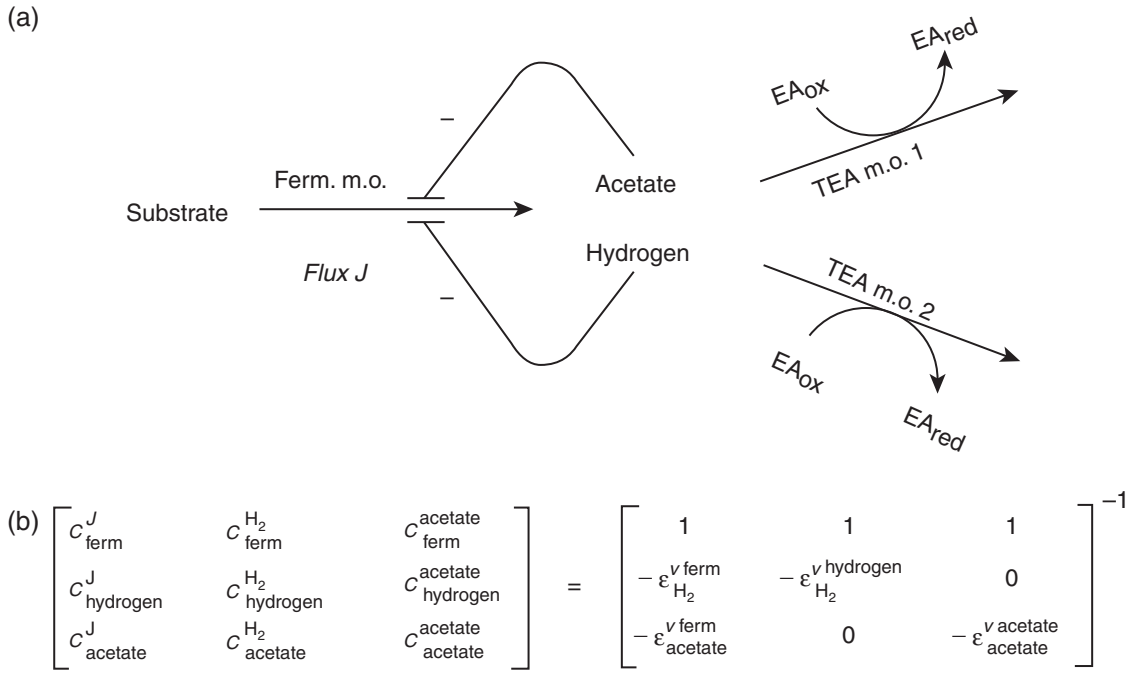


Figure 7.5 (a) Model of a syntrophic microbial community consisting of three members, one fermenting an organic substrate and producing acetate and hydrogen, the other two consuming the intermediates. Arrows indicate material transfer; the curved lines ending with bars indicate negative control of the products. Ferm. m.o., fermenting microorganisms; TEA m.o., terminal electron-accepting microorganisms; EA_{ox}, EA_{red}, oxidized and reduced electron acceptors, respectively; *J*, flux of substrate. (b) Control matrix (left) obtained as the inverse of the elasticity matrix (right) that corresponds to the scheme shown in (a). *C*, flux-control coefficient; the superscript indicates which flux (*J*, hydrogen, acetate) is controlled, the subscript indicates the controlling functional group. ε, Elasticity coefficient; the superscript indicates the functional group, the subscript indicates the substrate. From Röling *et al*. (2007), with permission from John Wiley and Sons.

conditions degradation of a substrate is controlled by which group of microorganisms. It turned out that under denitrifying and iron-reducing conditions flux control by terminal-electron accepting microorganisms is generally negligible and control is exerted mainly by the fermentors; however, under less favourable redox conditions flux control by terminal electron-accepting microorganisms can strongly increase. The redox potential is therefore a dominant driver that determines where flux control is concentrated. The results have very important practical implications, for example in the remediation and management of polluted groundwater systems.

For the moment, ecological control analysis seems to fit best with problems of food web ecology and biogeochemical cycles. However, we see no reason why application should be limited to these fields. We have seen several examples in this book that can be phrased in terms of fluxes (contaminants, oxygen radicals, resource allocation) and that could be analysed with some kind of control analysis. There is also a link between metabolic flux and population genetics, as explored in an early paper by Dykhuizen *et al*. (1987). These authors analysed enzymes of the *lac* operon of *E. coli* and showed that the β-galactosidase enzyme had only a small control coefficient with regard to fitness. Therefore, genetic variants of this enzyme in which enzyme activity is changed mildly in comparison to the wild type will mostly be selectively neutral. This was in contrast with the β-galactosidase permease locus, which encodes an enzyme with strong control over fitness. Thus it was predicted that mutations in the latter

locus, even if they change enzyme activity by only a little, could be under directional selection. The evolutionary fate of an allele is thus determined by the flux-control coefficient of the enzyme it encodes. There are many examples where alternative variants of enzymes exert a different degree of flux control. Ecologists can utilize the natural variation in such enzymes to test the evolutionary implications of MCA.

7.3 Outlook

For the final section of this book we will discuss some emerging issues that we think will become important for ecological genomics in the near future. While genomics develops at staggering speed, so it is difficult to predict how much attention each of these issues will get; those we have selected follow logically from the topics addressed in this book.

7.3.1 Organization of model species communities

One of the most frequently experienced obstacles of ecological genomics at the moment is the availability of just a few fully-sequenced genomes of ecologically relevant species. However, we believe that this situation has changed radically with the availability of next-generation sequencing technology, and that many ecologically relevant species will soon have a sufficiently large genomic database to allow genome-wide analyses of ecological questions. Transcription profiling using microarrays is expected to be a major activity for ecological genomics in the near future, but direct sequencing of the transcriptome is expected to replace some of the microarray applications.

The real obstacle for ecological genomics is not the sequencing itself but the assembly, annotation, and maintenance of the extremely large databases that come with next-generation sequencing. In addition, the creation of sufficiently large collaborative networks around model species is important. As we have noted above, ecologists are fascinated more by the biodiversity of species than by one representative species, and have difficulty in accepting the idea of model species. Still, it is unavoidable that ecological laboratories intending to specialize in genomics will collaborate and agree on a single model. Benefit from genome information can also be obtained by working on a wild, close relative of a genomic model, as we have seen in the case of Brassicaceae related to *Arabidopsis* (Chapter 2).

7.3.2 Large-scale sequencing of the environment

In Chapter 3 we saw that complete genomes can be reconstructed from DNA extracted from the environment, allowing analysis of metabolic potentials of organisms that have never been cultured in the laboratory. This strategy seems to be especially fruitful when applied to simplified communities in extreme environments and in not-too-complicated ecosystems such as the ocean. Such large-scale sequencing projects produce a hardly imaginable amount of new information and without doubt new discoveries are in store. It seems that an order-of-magnitude increase in sequencing effort is still needed to apply this approach effectively to mesophilic complex communities including eukaryotes, such as soils. If technology and computing power allows for such an increase in the near future, large-scale sequencing of the environment could make an enormously important contribution to mapping the Earth's biodiversity and archiving sequences of species that risk extinction in the near future.

In addition to sequencing, functional screening of metagenomic libraries is also expected to provide a lot of new information. As we have discussed in Chapter 3, metagenomic studies at the moment are focused on discovering new genes or biosynthetic pathways for products that have potential application in medicine, biotechnology, or agriculture. We expect that in the near future metagenomics will also benefit ecology. Many functions important for nutrient cycling, mutualism, quorum sensing, plant–microbe signalling, and so on are still hidden in the genomes of partly unknown microorganisms. New technological developments that combine genomics with localization and the use of tracers seem to be particularly promising.

7.3.3 Wild transcription profiling

The great majority of genomic studies reviewed in this book were conducted with organisms in a laboratory environment. The microbial studies reviewed in Chapter 3 are an exception and that is why we have argued that the process of merging ecology with genomics has shown the greatest progress in microbiology. However, microbial ecological genomics has to deal with challenges that plague botanists and zoologists to a lesser degree, the *terra incognita* of biodiversity.

As the genomes of more and more ecologically relevant model species are sequenced, transcription profiling of organisms collected directly from their natural environment will come within reach. Such studies will be crucial to understanding the dynamics and ecological relevance of transcription profiles. It could very well be that the profiles that have been observed until now in laboratory-cultured organisms will differ greatly from those collected in the wild. This may especially hold for species that face conditions in the field that are very different from their laboratory environment. For example, most ecologists are convinced that disease and parasitism are important limiting factors in the field. Do transcription profiles of plants and animals in the wild contain signatures of frequently upregulated immune responses? We don't know.

7.3.4 The mechanistic framework

Sequencing and transcription profiling run the risk of only scraping the surface of what ecological genomics has to offer. Indeed, collecting endless lists of genes and gene-expression profiles is not an aim in itself. In this chapter we discussed the necessity of linking genomics with hypothesis-driven research. In our opinion, gene-expression profiling makes sense only if the genes sooner or later can be positioned in an analytical framework that is grounded in physiological or biochemical knowledge. Even ecologists, if they take ecological genomics seriously, will have to get to grips with the mechanisms of the processes they study. We have seen that the most successful stories of ecological genomics discussed in Chapter 4, on longevity in *C. elegans* and flowering time in *Arabidopsis*, came from previous painstaking work in genetics and biochemistry. We argued in Chapter 1 that microarray-based transcription profiling should perhaps be viewed as just an exploratory instrument, or as only a stop on the way to some basic question, not as a goal in itself. We should avoid the impression, sometimes given, that genomics is a very advanced form of stamp collecting.

7.3.5 New methods of data analysis

The field of bioinformatics is developing a tremendous number of new analytical tools to match the high-throughput pipelines of comparative and functional genomics. Most of these methods are aimed at dimension reduction using various types of multivariate statistics, cluster analysis, pattern recognition, and so on. Another avenue is the application of univariate models such as analysis of variance in a gene-by-gene manner, which raises statistical issues about false-positive results in large datasets. Ecologists are traditionally acquainted with statistics as a necessary tool for the analysis of noisy data; however, the sheer size of genomics datasets adds a new dimension to statistical analysis in ecological genomics.

We argued in Chapter 1 that the statistical approaches presently developed for microarray-derived transcription profiles are not necessarily optimal for ecological genomics. In particular, we argued that clustering of conditions may better fit with ecological problems than clustering of genes.

Another challenge for bioinformatics in ecology is how to define the normal operating range of species; that is, the set of expression profiles that reflects routine physiological variation and the absence of stress. Finally, the development of integrative approaches of the type visited earlier in this chapter, linking genomics and transcriptomics to questions of community structure, life-history, and ecological niche, may be the greatest challenge of all.

7.3.6 Comparative genomics

In the last few years the field of comparative genomics has expanded its scope tremendously and is

responsible for many exciting discoveries. Using techniques such as phylogenetic footprinting and phylogenetic shadowing information is obtained about the content of a genome in such a way that maximal use is made of homologies elsewhere in the tree of life. Using comparative genomics, organismal groups with unclear phylogenetic affiliation will soon be positioned solidly. Especially in the domain of eukaryotic microorganisms (protists) the tree of life is under thorough revision due to comparative genomics. Interestingly, new species are being sequenced not for the sake of these species but for the sake of a more or less distant model (usually humans). This development can actually benefit ecology. An example is the availability of a full genome sequence for the sea squirt, *Ci. intestinalis*, which was sequenced for comparative reasons, but is an equally good model for ecological studies in the coastal marine environment. We expect that the science of molecular evolution, supported by comparative genomics, is looking towards a glorious future.

7.3.7 Focus on natural variation

Natural variation in gene expression among and within field populations can be an extremely powerful tool for revealing evolutionary mechanisms, as illustrated by the various population genomics studies discussed in Chapter 6. These studies suggest that inter-individual variation of transcriptional regulation is very common in natural populations. Such variation, which is ultimately due to polymorphisms in promoter sequences or variation in *trans*-acting factors or signal transduction pathways, can be an important template for natural selection. The recent literature seems to suggest that, at least in animals, evolution acting upon transcriptional regulation could be at least as important as evolution of coding sequences.

The use of natural variation is also important because it allows ecological genomics to get away from the present focus on laboratory mutants (see the work presented in Chapter 4). Knocking-out genes to analyse their function is a crucial tool in mechanistic studies, but the relevance of the results for understanding functions in the environment is limited. The point is, we need to know the fitness consequences of gene mutations under ecologically relevant conditions (Kammenga *et al.* 2008). Where nature is already providing this variation, it is best to make optimal use of it.

The by now classical approaches of quantitative genetics and QTL analysis are badly integrated with ecological genomics at the moment. We foresee a renewed interest in quantitative character analysis, using QTL analysis as well as genome-wide association mapping, when ecological genomics develops further. There is an increasing realization that even the expression of a single gene behaves as a quantitative character because several other gene products (transcription factors, enhancers, repressors, ribosomal factors, and splice factors) are involved with gene expression. Expression of single genes, when measured with quantitative methods such as real-time PCR, can be subjected to the same statistical analysis as traditional quantitative characters like body size and development time. The link between quantitative characters and genomics has developed further into a new field, called *genetical genomics* (see Chapter 6).

7.3.8 Epigenetics

We know now that regulatory information that is not defined in the DNA can still be transmitted between cells, and from one generation to another, using DNA methylation, histone acetylation, and so on. By means of epigenetic processes cells may canalize their potential expression repertoire and adopt specific functions or states. In addition, an epigenetically regulated state may be transmitted in cell lineages. Such epigenetic processes seem to be especially important when transmitting information that must be remembered in the daughter cells in order to respond in an adequate way to some environmental factor. We have seen several examples of epigenetic processes in this book, for example the regulation of flowering time in *Arabidopsis* by temperature (Chapter 4). Still, epigenetics has had little influence on ecology (Jablonka and Lamb 2002). We call for more attention on this important phenomenon, for example by studying the frequency of epigenetic variants in natural populations.

One of the mechanisms of epigenetic regulation is due to RNAi, a phenomenon by which double-stranded RNAs specifically suppress a target protein by degradation of its mRNA or by transcriptional gene silencing (Novina and Sharp 2004). Following the discovery of RNAi in nematodes and plants in the 1990s, realization has come that it is a widespread natural phenomenon and an evolutionarily ancient mechanism of genome defence. It may be expected that RNAi is also involved in regulating ecologically important processes, but this has not yet been demonstrated.

7.3.9 The unification of biology

The enormous scientific success of biology in the course of the last century is evident from the fact that many other disciplines have adopted biological epithets. This is not limited to biochemistry and biophysics, but extends to disciplines such as biogeology, bioarchaeology, and biopsychology. As a consequence of this spreading movement, biology has become such an extensive scientific field that it is hardly possible to view the whole. What was just zoology, botany, and microbiology before is now scattered over a wide variety of subdisciplines. The same diversifying trend is present in ecology itself, which has to cover the widest span of all: from molecule to ecosystem. How sustainable is this situation? Should we fear for the future of ecology as a homogeneous science? Not denying the tendency to diversify, which is typically seen in any growing field, we also expect an opposite trend, the unification of biology. Genomics could very well turn out to act as a new point of crystallization, bringing biologists of various colours to the same core. As demonstrated by the recent establishment of a Gordon Research Conference series on Evolutionary and Ecological Functional Genomics (Feder and Mitchell-Olds 2003), there is an increasing interest among molecular biologists in issues of evolution and ecology. Likewise, more and more ecologists are becoming interested in the molecular biology of ecological phenomena. We hope that this book has contributed at least a little to this admittedly ambitious perspective.

References

Abad, P., Gouzy, J., Aury, J.-M., Castagnone-Sereno, P., Danchin, E.G.J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V.C., *et al.* (2008) Genome sequence of the metazoan plant-parasitic nematode *Meloidyne incognita*. *Nature Biotechnology* **26**: 909–15.

Abzhanov, A., Protas, M., Grant, B.R., Grant, P.R., and Tabin, C.J. (2004) *Bmp4* and morphological variation of beaks in Darwin's finches. *Science* **305**: 1462–565.

Abzhanov, A., Kuo, W.P., Hartmann, C., Grant, B.R., Grant, P.R., and Tabin, C.J. (2006) The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature* **442**: 563–7.

Acheré, V., Favre, M., Besnard, G., and Jeandroz, S. (2005) Genomic organization of molecular differentiation in Norway spruce (*Picea abies*). *Molecular Ecology* **14**: 3191–201.

Adam, D. (2000) Now for the hard ones. *Nature* **408**: 792–3.

Achtman, M. and Wagner, M. (2008) Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology* **6**: 431–40.

Adamcyk, J., Hesselsoe, M., Iversen, N., Horn, M., Lehner, A., Nielsen, P.H., Schloter, M., Roslev, P., and Wagner, M. (2003) The isotope array, a new tool that employs substrate-mediated labeling of rRNA for determination of microbial community structure and function. *Applied and Environmental Microbiology* **69**: 6875–87.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F. *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185–95.

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**: 1600–1607.

Allen, E.A. and Banfield, J.F. (2005) Community genomics in microbial ecology and evolution. *Nature Reviews Microbiology* **3**: 489–98.

Altinicek, B. and Vilcinskas, A. (2007) Analysis of the immune-related transcriptome of a lophotrochozoan model, the marine annelid *Platynereis dumerilii*. *Frontiers in Zoology* **4**: 18.

Amasino, R.M. (2003) Flowering time: a pathway that begins at the 3´end. *Current Biology* **13**: R670–2.

Amasino, R. (2004) Take a cold flower. *Nature Genetics* **36**: 111–12.

Amores, A., Force, A., Yan, Y.-L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.-L. *et al*. (1998) Zebrafish *hox* clusters and vertebrate genome evolution. *Science* **282**: 1711–14.

Andersen, G.L., DeSantis, T.Z., Murray, S.R., and Moberg, J.P. (2004) *Phylogenetic chip for microbial detection*. 10th International Symposium on Microbial Ecology, Cancun, Mexico. International Society of Microbial Ecology, Geneva: 110.

Andersson, J.O. (2005) Lateral gene transfer in eukaryotes. *Cellular and Molecular Life Sciences* **62**, 1182–97.

Andolfatto, P. (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**: 1149–52.

Andrews, G.K. (2001) Cellular zinc sensors: MTF-1 regulation of gene expression. *BioMetals* **14**: 223–37.

Ankley, G.T., Daston, G.P., Degitz, S.J., Denslow, N.D., Hoke, R.A., Kennedy, S.W., Miracle, A.L., Perkins, E.J., Snape, J., Tillit, D.E., *et al.* (2006) Toxicogenomics in regulatory ecotoxicology. *Environmental Science and Technology* **40**: 4055–65.

Antequerra, F. (2003) Structure, function and evolution of CpG island promoters. *Cellular* and *Molecular Life Sciences* **60**: 1647–58.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A. *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**: 1301–10.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

Arantes-Oliveira, N., Apfeld, J., Dillin, A., and Kenyon, C. (2002) Regulation of life-span by germ-line stem cells in *Caenorhabditis elegans*. *Science* **295**: 502–5.

Aravind, L., Tatusov, R.L., Wolf, Y.I., Walker, D.R., and Koonin, E.V. (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends in Genetics* **14**: 442–4.

Arbeitman, M.N., Furlong, E.E.M., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W., and White, K.P. (2002) Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297**: 2270–5.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M. *et al*. (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**: 79–86.

Armstrong, M.R., Blok, V.C., and Phillips, M.S. (2000) A multipartite mitochondrial genome in the potato cyst nematode *Globodera pallida*. *Genetics* **154**: 181–92.

Assunção, A.G.L., Herrero, E., Lin, Y.-F., Huettel, B., Talukdar, S., Smaczniak, C., Immink, R.G.H., Van Eldik, M., Fiers, M., Schat, H., *et al.* (2010) *Arabidopsis thaliana* transcription factors bZIP19 and bZIP23 regulate the adaptation to zinc deficiency. *Proceedings of the National Academy of Sciences USA* **107**: 10296–301.

Ausín, I., Alonso-Blanco, C., Jarillo, J.A., Ruiz-Garcia, L., and Martínez-Zapater, J.M. (2004) Regulation of flowering time by FVE, a retinoblastome-associated protein. *Nature Genetics* **36**: 162–6.

Backström, N., Fagerberg, S., and Ellegren, H. (2008) Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. *Molecular Ecology* **17**: 964–80.

Baker, B.J., Lutz, M.A., Dawson, S.C., Bond, P.L., and Banfield, J.F. (2004) Metabolically active eukaryotic communities in extremely acidic mine drainage. *Applied and Environmental Microbiology* **70**: 6264–71.

Baldwin, I.T. (2001) An ecologically motivated analysis of plant-herbivore interactions in native tobacco. *Plant Physiology* **127**: 1449–58.

Ball, K.D. and Trevors, J.T. (2002) Bacterial genomics: the use of DNA microarrays and bacterial artificial chromosomes. *Journal of Microbiological Methods* **49**: 275–84.

Bar-Or, C., Czosnek, H., and Koltai, H. (2007) Cross-species microarray hybridizations: a developing tool for studying species diversity. *Trends in Genetics* **23**: 200–7.

Barnes, A.I. and Partridge, L. (2003) Costing reproduction. *Animal Behaviour* **66**: 199–204.

Barnett, M.J., Fisher, R.F., Jones, T., Komp, C., Abola, A.P., Barloy-Hubler, F., Bowser, L., Capela, D., Galibert, F., Gouzy, J. *et al*. (2001) Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* megaplasmid. *Proceedings of the National Academy of Sciences USA* **98**: 9883–8.

Bartels, D. and Sunkar, R. (2005) Drought and salt tolerance in plants. *Critical Reviews in Plant Sciences* **24**: 23–58.

Bartke, A., Wright, J.C., Mattison, J.A., Ingram, D.K., Miller, S.R., and Roth, G.S. (2001) Extending the lifespan of long-lived mice. *Nature* **414**: 412.

Bastow, R., Mylne, J.S., Lister, C., Lippman, Z., Martienssen, R.A., and Dean, C. (2004) Vernalization requires epigenetic silencing of *FLC* by histone methylation. *Nature* **427**: 164–7.

Bavykin, S.G., Akowksi, J.P., Zakhariev, V.M., Barsky, V.E., Perov, A.N., and Mirzabekov, A.D. (2001) Portable system for microbial sample preparation and oligonucleotide microarray analysis. *Applied and Environmental Microbiology* **67**: 922–8.

Bayley, M. and Holmstrup, M. (1999) Water vapor absorption in arthropods by accumulation of myoinositol and glucose. *Science* **285**: 1909–11.

Bayne, B.L. (1989) Measuring the biological effects of pollution: the mussel watch approach. *Water Science and Technology* **21**: 1089–100.

Beaumont, M.A. and Balding, D.J. (2004) Identifying adaptive divergence among populations from genome scans. *Molecular Ecology* **13**: 969–80.

Beaumont, M.A. and Nichols, R.A. (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London, Series B* **263**: 1619–25.

Becker, A. and Theißen, G. (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Molecular Phylogenetics and Evolution* **29**: 464–89.

Beebee, T. and Rowe, G. (2004) *Introduction to Molecular Ecology*. Oxford University Press, Oxford.

Béjà, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P., Jovanovich, S.B., Gates, C.M., Feldman, R.A., Spudich, J.L. *et al*. (2000a) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**: 1902–6.

Béjà, O., Suzuki, M.T., Koonin, E.V., Aravind, L., Hadd, A., Nguyen, L.P., Villacorta, R., Amjadi, M., Garrigues, C., Jovanovich, S.B. *et al*. (2000b) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environmental Microbiology* **2**: 516–29.

Béjà, O., Spudich, E.N., Spudich, J.L., Leclerc, M., and DeLong, E.F. (2001) Proteorhodopsin phototrophy in the ocean. *Nature* **411**: 786–9.

Beldade, P. and Brakefield, P.M. (2002) The genetics and evo-devo of butterfly wing patterns. *Nature Reviews Genetics* **3**: 442–52.

Beldade, P., Brakefield, P.M., and Long, A.D. (2002) Contribution of *Distal-less* to quantitative variation in butterfly eyespots. *Nature* **415**: 315–18.

Beldade, P., Rudd, S., Gruber, J.D., and Long, A.D. (2006) A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics* 7: 130.

Benner, S.A., Liberles, D.A., Chamberlin, S.G., Govindarajan, S., and Knecht, L. (2000) Functional inferences from reconstructed evolutionary biology involving rectified databases—An evolutionarily grounded approach to functional genomics. *Research in Microbiology* **151**: 97–106.

Berg, M.P. and Ellers, J. (2010) Trait plasticity in species interactions: a driving force of community dynamics. *Evolutionary Ecology* **24**: 617–29.

Bergelson, J. and Roux, F. (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. *Nature Reviews Genetics* **11**: 867–79.

Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* **241**: 3–17.

Bernatchez, L., Renaut, S., Whiteley, A.R., Derome, N., Jeukens, J., Landry, L., Lu, G., Nolte, A.W., Østbye, K., Rogers, S.M., *et al.* (2010) On the origin of species: insights from the ecological genomics of lake whitefish. *Philosophical Transactions of the Royal Society of London. B. Biological Sciences* **365**: 1783–800.

Bhaumik, S.R., Smith, E., and Shilatifard, A. (2007) Covalent modifications of histones during development and disease pathogenesis. *Nature Structural* and *Molecular Biology* **14**: 1008–16.

Biology Analysis Group and Genome Analysis Group (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* **306**: 1937–40.

Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes* and *Development* **16**: 6–21.

Bishop, W.E., Clarke, D.P., and Travis, C.C. (2001) The genomic revolution: what does it mean for risk assessment? *Risk Analysis* **21**: 983–7.

Blanc, G. and Wolfe, K.H. (2004) Widespread paleopolyploidy in model species inferred from age distributions of duplicate genes. *The Plant Cell* **16**: 1667–78.

Blattner, F.R., Plunket, III G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K12. *Science* **277**: 1453–62.

Blaxter, M.L., De Ley, P., Garey, J.R., Liu, L.X., Scheldeman, P., Vierstraete, A., Vanfleteren, J.R., Mackey, L.Y., Dorris, M., Frisse, L.M. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**: 71–5.

Blázquez, M.A., Ahn, J.H., and Weigel, D. (2003) A thermosensory pathway controlling flowering time in *Arabidopsis thaliana*. *Nature Genetics* **33**: 168–71.

Blüher, M., Kahn, B.B., and Kahn, C.R. (2003) Extended longevity in mice lacking the insulin receptor in adipose tissue. *Science* **299**: 572–4.

Bochdanovits, Z. and De Jong, G. (2004) Antagonistic pleiotropy for life-history traits at the gene expression level. *Proceedings of the Royal Society of London* (supplement) **271**: S75–8.

Bochdanovits, Z., Van der Klis, H., and De Jong, G. (2003) Covariation of larval gene expression and adult body size in natural populations of *Drosophila melanogaster*. *Molecular Biology and Evolution* **20**: 1760–6.

Bodrossy, L., Stralis-Pavese, N., Murrell, J.C., Radajewski, S., Weilharter, A., and Sessitsch, A. (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. *Environmental Microbiology* **5**: 566–82.

Bohnert, H.J., Ayoubi, P., Borchert, C., Bressan, R.A., Burnap, R.L., Cushman, J.C., Cushman, M.A., Deyholos, M., Fischer, R., Galbraith, D.W. *et al.* (2001) A genomics approach towards salt stress tolerance. *Plant Physiology and Biochemistry* **39**: 295–311.

Bongers, T. and Ferris, H. (1999) Nematode community structure as a bioindicator in environmental monitoring. *Trends in Ecology and Evolution* **14**: 224–8.

Bonin, A., Taberlet, P., Miaud, C., and Pompanon, F. (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Molecular Biology* and *Evolution* **23**: 773–83.

Bonin, A., Paris, M., Tetreau, G., David, J.-P., and Després, L. (2009) Candidate genes revealed by a genome scan for mosquito resistance to a bacterial insecticide: sequence and gene expression variations. *BMC Genomics* **10**: 551.

Bonneaud, C., Burnside, J., and Edwards, S.V. (2008) High-speed developments in avian genomics. *BioScience* **58**: 587–95.

Bossdorf, O., Richards, C.L., and Pigliucci, M. (2008) Epigenetics for ecologists. *Ecology Letters* **11**: 106–15.

Bourlat, S.J., Juliusdottir, T., Lowe, C.J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E.S., Thorndyke, M., Nakano, H., Kohn, A.B., *et al.* (2006) Deuterostome

phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* **444**: 85–8.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433–8.

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., *et al.* (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239–44.

Brachat, S., Dietrich, F.S., Voegeli, S., Zhang, Z., Stuart, L., Lerch, A., Gates, K., Gaffney, T.D., and Philippsen, P. (2003) Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biology* **4**: R45.

Bradbury, J. (2004) Nature's nanotechnologists: unveiling the secrets of diatoms. *PLoS Biology* **2**: e347–e306.

Bradshaw, Jr, H.D., Ceulemans, R., Davis, J., and Stettler, R. (2000) Emerging model systems in plant biology: poplar (*Populus*) as a model forest tree. *Journal of Plant Growth Regulation* **19**: 306–13.

Braeckman, B.P., Houthoofd, K., and Vanfleteren, J.R. (2001) Insulin-like signaling, metabolism, stress resistance and aging in *Caenorhabditis elegans*. *Mechanisms of Ageing and Development* **122**: 673–93.

Branicky, R., Bénard, C., and Hekimi, S. (2000) *clk-1*, mitochondria and physiological rates. *BioEssays* **22**: 48–56.

Braun, E.L., Halpern, A.L., Nelson, M.A., and Natvig, D.O. (2000) Large-scale comparison of fungal sequence information: mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae*. *Genome Research* **10**: 416–30.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C. *et al*. (2001) Minimum information about a microarray experiment (MIAME)—towards standards for microarray data. *Nature Genetics* **29**: 365–71.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., and Rohwer, F. (2002) Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences USA* **99**: 14250–5.

Britton, J.S., Lockwood, W.K., Li, L., Cohen, S.M., and Edgar, B.A. (2002) *Drosophila*'s insulin/PI3-kinase pathway coordinates cellular metabolism with nutritional conditions. *Developmental Cell* **2**: 239–49.

Brody, E.L., DeSantis, T.Z., Moberg Parker, J.P., Zubietta, I.X., Piceno, Y.M., and Andersen, G.L. (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 299–304.

Brogiolo, W., Stocker, H., Ikeya, T., Rintelen, F., Fernandez, R., and Hafen, E. (2001) An evolutionary conserved function of the *Drosophila* insulin receptor and insulin-like peptides in growth control. *Current Biology* **11**: 213–21.

Brown, A. (2004) *In the Beginning Was the Worm. Finding the Secrets of Life in a Tiny Hermaphrodite*. Pocket Books, London.

Brumfield, R.T., Beerli, P., Nickerson, D.A., and Edwards, S.V. (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution* **18**: 249–56.

Brunetti, C.R., Selegue, J.E., Monteiro, A., French, V., Brakefield, P.M., and Carroll, S.B. (2001) The generation and diversification of butterfly eyespot color patterns. *Current Biology* **11**: 1578–85.

Brunner, A.M. and Nilsson, O. (2004) Revisiting tree maturation and floral initiation in the poplar functional genomics area. *New Phytologist* **164**: 43–51.

Brunner, A.M., Busov, V.B., and Strauss, S.H. (2004) Poplar genome sequence: functional genomics in an ecologically dominant plant species. *Trends in Plant Science* **9**: 49–56.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., Fitzgerald, L.M., Clayton, R.A., Gocayne, J.D. *et al*. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**: 1058–73.

Bundy, J.G., Sidhu, J.K., Rana, F., Spurgeon, D.J., Svendsen, C., Wren, J.F., Stürzenbaum, S.R., Morgan, A.J., and Kille, P. (2008) 'Systems toxicology' approach identifies coordinated metabolic responses to copper in a terrestrial non-model invertebrate, the earthworm *Lumbricus rubellus*. *BMC Biology* **6**: 25.

Bundy, J.G., Davey, M.P., and Viant, M.R. (2009) Environmental metabolomics: a critical review and future perspectives. *Metabolomics* **5**: 3–21.

Burnett, K.G., Bain, L.J., Baldwin, W.S., Callard, G.V., Cohen, S., Di Gulio, R.T., Evans, D.H., Gómez-Chiarri, M., Hahn, M.E., Hoover, C.A., *et al.* (2007) *Fundulus* as the premier teleost model in environmental biology: Opportunities for new insights using genomics. *Comparative Biochemistry and Physiology, Part D* **2**: 257–86.

Burnaford, J.L. (2004) Habitat modification and refuge from sublethal stress drive a marine plant-herbivore association. *Ecology* **85**: 2837–47.

Calow, P. (1989) Proximate and ultimate responses to stress in biological systems. *Biological Journal of the Linnean Society* **37**: 173–81.

Campàs, O., Mallarino, R., Herrel, A., Abzhanov, A., and Brenner, M.P. (2010) Scaling and shear transformations capture beak scale variation in Darwin's finches. *Proceedings of the National Academy of Sciences USA* **107**: 3356–60.

Campbell, B.J., Stein, J.L., and Cary, S.C. (2003) Evidence of chemolithoautotrophy in the bacterial community associated with *Alvinella pompejana*, a hydrothermal vent polychaete. *Applied and Environmental Microbiology* **69**: 5070–8.

Campbell, D. and Bernatchez, L. (2004) Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. *Molecular Biology* and *Evolution* **21**: 945–56.

Cañestro, C., Bassham, S., and Postlewaith, J.H. (2003) Seeing chordate evolution through the *Ciona* genome sequence. *Genome Biology* **4**: 208.

Cañestro, C., Yokai, H., and Postlewaith, J.H. (2007) Evolutionary developmental biology and genomics. *Nature Reviews Genetics* **8**: 932–42.

Cánovas, D., Cases, I., and De Lorenzo, V. (2003) Heavy metal tolerance and metal homeostasis in *Pseudomonas putida* as revealed by complete genome analysis. *Environmental Microbiology* **5**: 1242–56.

Cardozo, A.K., Berthou, L., Kruhøffer, M., Ørntoft, T., Nicolls, M.R., and Eizirik, D.L. (2003) Gene microarray study corroborates proteomic findings in rodent islet cells. *Journal of Proteome Research* **2**: 553–5.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics* **38**: 626–35.

Carroll, S.B., Grenier, J.K., and Weatherbee, S.D. (2005) *From DNA to Diversity*. Blackwell Publishing, Malden.

Casal, J.J., Fankhauser, C., Coupland, G., and Blázquez, M.A. (2004) Signalling for developmental plasticity. *Trends in Plant Science* **9**: 309–14.

Causton, H.C., Quackenbush, J., and Brazma, A. (2003) *Microarray Gene Expression Data Analysis. A Beginner's Guide*. Blackwell Science, Malden.

Causton, H.C., Ren, B., Koh, S.S., Harbison, C.T., Kanin, E., Jennings, E.G., Lee, T.I., True, H.L., Lander, E.S., and Young, R.A. (2001) Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the Cell* **12**: 323–37.

Cavicchioli, R., DeMaere, M.Z., and Thomas, T. (2006) Metagenomic studies reveal the critical and wide-ranging ecological importance of uncultivated archaea: the role of ammonia oxidizers. *BioEssays* **29**: 11–14.

*C. elegans*Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**: 2012–18.

Chain, P., Lamerdin, J., Larimer, F., Regala, W., Lao, V., Land, M., Hauser, L., Hooper, A., Klotz, M., Norton, J. *et al.* (2003) Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea*. *Journal of Bacteriology* **185**: 2759–73.

Chan, A.P., Crabtree, J., Zhao, Q., Lorenzi, H., Orvis, J., Puiu, D., Melake-Berhan, A., Jones, K.M., Redman, J., Chen, G., *et al.* (2010) Draft genome sequence of the oilseed species *Ricinus communis*. *Nature Biotechnology* **28**: 951–6.

Chang, L. and Karin, M. (2001) Mammalian MAP kinase signalling cascades. *Nature* **410**: 37–40.

Chariton, A.A., Court, L.N., Hartley, D.M., Colloff, M.J., and Hardy, C.M. (2010) Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment* **8**: 233–8.

Charlesworth, J. and Eyre-Walker, A. (2008) The McDonald-Kreitman test and slightly deleterious mutations. *Molecular Biology and Evolution* **25**: 1007–15.

Chase, J.M. and Leibold, M.A. (2003) *Ecological Niches. Linking Classical and Contemporary Approaches*. The University of Chicago Press, Chicago.

Chen, M., Chory, J., and Fankhauser, C. (2004) Light signal transduction in higher plants. *Annual Review of Genetics* **38**: 87–117.

Chen, W., Provart, N.J., Glazebrook, J., Katagiri, F., Chang, H.-S., Eulgem, T., Mauch, F., Luan, S., Zou, G., Whitham, S.A. *et al.* (2002) Expression profile matrix of Arabidopsis transcription factor genes suggests their putative functions in response to environmental stresses. *The Plant Cell* **14**: 559–74.

Cheng, C., Pounds, S.B., Boyett, J.M., Pei, D., Kuo, M.-L., and Roussel, M.F. (2004) Statistical significance threshold criteria for analysis of microarray gene expression data. *Statistical Applications in Genetics and Molecular Biology* **3**: 36.

Chin, K.-J., Esteve-Núñez, A., Leang, C., and Lovley, D.R. (2004) Direct correlation between rates of anaerobic respiration and levels of mRNA for key respiratory genes in *Geobacter sulfurreducens*. *Applied and Environmental Microbiology* **70**: 5183–9.

Cho, J.-C. and Tiedje, J.M. (2001) Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. *Applied and Environmental Microbiology* **67**: 3677–82.

Cho, J.-C. and Tiedje, J.M. (2002) Quantitative detection of microbial genes by using microarrays. *Applied and Environmental Microbiology* **68**: 1425–30.

Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J., *et al.* (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* **466**: 388–92.

Christoffels, A., Koh, E.G.L., Chia, J.-M., Brenner, S., Aparicio, S., and Venkatesh, B. (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fish. *Molecular Biology and Evolution* **21**: 1146–51.

Christophides, G.K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P.T., Collins, F.H., Danielli, A., Dimopoulos, G. *et al*. (2002) Immunity-related genes and gene families in *Anopheles gambiae*. *Science* **298**: 159–65.

Clancy, D.J., Gems, D., Harshman, L.G., Oldham, S., Stocker, H., Hafen, E., Leevers, S.J., and Partridge, L. (2001) Extension of life-span by loss of CHICO, a *Drosophila* insulin receptor substrate protein. *Science* **292**: 104–6.

Clancy, D.J., Gems, D., Hafen, E., Leevers, S.J., and Partridge, L. (2002) Dietary restriction in long-lived dwarf flies. *Science* **296**: 319.

Clark, M.S., Crawford, D.L., and Cossins, A. (2003) Meeting review: worldwide genomic resources for non-model fish species. *Comparative and Functional Genomics* **4**: 502–8.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**: 71–6.

Cobbett, C. and Goldsbrough, P. (2002) Phytochelatins and metallothioneins: roles in heavy metal detoxification and homeostasis. *Annual Review of Plant Biology* **53**: 159–82.

Cock, J.M., Sterck, L., Rouzé, P., Scornet, D., Allen, A.E., Amoutzias, G., Anthouard, V., Artiguenave, F., Aury, J.-M., Badger, J.H., *et al.* (2010) The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**: 617–21.

Colborn, T., Myers, J.P., and Dumanoski, D. (1996) *Our Stolen Future*. Little, Brown and Company, Boston.

Colbourne, J.K., Singan, V.R., and Gilbert, D.G. (2004) wFleaBase: the *Daphnia* genome database. *BMC Bioinformatics* **6**: 45.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Bandela, A.M., Cardenas, E., Garrity, G.M., and Tiedje, J.M. (2006) The ribosomal database project (RDP-II): introducing *myRDP* space and quality controlled public data. *Nucleic Acids Research* **35**: D169–72.

Colosimo, P.F., Peichel, C.L., Nereng, K.S., Blackman, B.K., Shapiro, M.D., Schluter, D., and Kingsley, D.M. (2004) The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biology* **2**: 0635–41.

Colosimo, P.F., Hosemann, K.E., Balabhadra, S., Villreal Jr, G., Dickson, R.M., Grimwood, J., Schmutz, J., Myers, R.M., Schluter, D., and Kingsley, D.M. (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**: 1928–33.

Cornish-Bowden, A. and Cárdenas, M.L. (2005) Systems biology may work when we learn to understand the parts in terms of the whole. *Biochemical Society Transactions* **33**: 516–19.

Corona, M., Estrada, E., and Zurita, M. (1999) Differential expression of mitochondrial genes between queens and workers during caste determination in the honeybee *Apis mellifera*. *Journal of Experimental Biology* **202**: 929–38.

Cowan, D.A., Arslanoglu, A., Burton, S.G., Baker, G.C., Cameron, R.A., Smith, J.J., and Meyer, Q. (2004) Metagenomics, gene discovery and the ideal biocatalyst. *Biochemical Society Transactions* **32**: 298–302.

Crawford, D.L. (2001) Functional genomics does not have to be limited to a few select organisms. *Genome Biology* **2**: Interactions/1001.1.

Crawford, D.L. and Oleksiak, M. (2007) The biological importance of measuring individual variation. *The Journal of Experimental Biology* **210**: 1613–21.

Cresko, W.A., Amores, A., Wilson, C., Murphy, J., Currey, J., Phillips, P., Bell, M.A., Kimmel, C.B., and Postlewaith, J.H. (2004) Parallel genetic basis for repeated evolution of armor plate loss in Alaskan threespine stickleback populations. *Proceedings of the National Academy of Sciences USA* **101**: 6050–5.

Croal, L.R., Gralnick, J.A., Malasarn, D., and Newman, D.K. (2004) The genetics of geochemistry. *Annual Review of Genetics* **38**: 175–202.

Curtis, T.P. and Sloan, W.T. (2004) Prokaryote diversity and its limits: microbial community structure in nature and implications for microbial ecology. *Current Opinion in Microbiology* **7**: 221–6.

Curtis, T.P., Sloan, W.T., and Scannell, J.W. (2002) Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences USA* **99**: 10494–9.

Cushman, J.C. and Bohnert, H.J. (2000) Genomic approaches to plant stress tolerance. *Current Opinion in Plant Biology* **3**: 117–24.

Daborn, P.J., Yen, J.L., Bogwitz, M.R., Le Goff, G., Feil, E., Jeffers, S., Tijet, N., Perry, T., Heckel, D., Batterham, P., *et al.* (2002) A single P450 allele associated with insecticide resistance in *Drosophila*. *Nature* **297**: 2253–6.

Dallinger, R., Berger, B., Hunziker, P., and Kägi, J.H.R. (1997) Metallothionein in snail Cd and Cu metabolism. *Nature* **388**: 237–8.

D'Amico, L.J., Davidowitz, G., and Nijhout, H.F. (2001) The developmental and physiological basis of body size evolution in an insect. *Proceedings of the Royal Society of London Series B* **268**: 1589–93.

Daniel, R. (2004) The soil metagenome—a rich resource for the discovery of novel natural products. *Current Opinion in Biotechnology* **15**: 199–204.

Daniel, R. (2005) The metagenomics of soil. *Nature Reviews Microbiology* **3**: 470–8.

Darwin, C. (1845) *The Voyage of The 'Beagle'*. J.M. Dent & Sons, London.

Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. Penguin Books, Harmondsworth.

Davidowitz, G., D'Amico, L.J., and Nijhout, H.F. (2003) Critical weight in the development of insect body size. *Evolution & Development* **5**: 188–97.

De Gregorio, E., Spellman, P.T., Rubin, G.M., and Lemaitre, B. (2001) Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. *Proceedings of the National Academy of Sciences USA* **98**: 12590–5.

De Jong, G. (1995) Phenotypic plasticity as a product of selection in a variable environment. *American Naturalist* **145**: 493–512.

De Jong, G. and Bochdanovits, Z. (2003) Latitudinal clines in *Drosophila melanogaster*: body size, allozyme frequencies, inversion frequencies, and the insulin-signalling pathway. *Journal of Genetics* **82**: 207–23.

De la Torre, J.R., Christianson, L.M., Béjà, O., Suzuki, M.T., Karl, D.M., Heidelberg, J., and DeLong, E.F. (2003) Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proceedings of the National Academy of Sciences USA* **100**: 12830–5.

De Meester, L. (1996) Local genetic differentiation and adaptation in freshwater zooplankton populations: Patterns and processes. *Ecoscience* **3**: 385–99.

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D.M. *et al.* (2002) The draft genome of *Ciona intestinalis*: insights into the chordate and vertebrate origins. *Science* **298**: 2157–67.

DeLong, E.F. (2001) Microbial seascapes revisited. *Current Opinion in Microbiology* **4**: 290–5.

DeLong, E.F. (2005) Microbial community genomics in the ocean. *Nature Reviews Microbiology* **3**: 459–69.

Demerec, M. and Kaufmann, B.P. (1950) *Drosophila Guide. A Guide to Introductory Studies of the Genetics and Cytology of Drosophila melanogaster. With an Appendix Containing a Series of Experiments to be Conducted by the Beginning Student*. The Lord Baltimore Press, Baltimore.

Denef, V.J., Park, J., Rodrigues, J.L.M., Hashsham, S.A., and Tiedje, J.M. (2003) Validation of a more sensitive method for using spotted oligonucleotide DNA microarrays for functional genomics studies on bacterial communities. *Environmental Microbiology* **5**: 933–43.

Denlinger, D.L. (2002) Regulation of diapause. *Annual Review of Entomology* **47**: 93–122.

Denver, D.R., Morris, K., Lynch, M., and Thomas, W.K. (2004) High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* **430**: 679–82.

Deppenheimer, U., Johann, A., Hartsch, T., Merkl, R., Schmitz, R.A., Martinez-Arias, R., Henne, A., Wiezer, A., Bäumer, S., Jacobi, C. *et al.* (2002) The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between Bacteria and Archaea. *Journal of Molecular Microbiology and Biotechnology* **4**: 453–61.

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A.Z., Robbens, S., Partensky, F., Degroeve, S., Echeynié, S., Cooke, R., *et al.* (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proceedings of the National Academy of Sciences USA* **103**: 11647–11652.

DeRisi, J.L., Iyer, V.R., and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**: 680–6.

Derome, N. and Bernatchez, L. (2006) The transcriptomics of ecological convergence between 2 limnetic coregonine fishes (Salmonidae). *Molecular Biology and Evolution* **23**: 2370–8.

Derome, N., Duchesne, P., and Bernatchez, L. (2006) Parallelism in gene transcription among sympatric lake whitefish (*Coregonus clupeaformis* Mitchell) ecotypes. *Molecular Ecology* **15**: 1239–49.

DeSantis, T.Z., Dubosarskiy, I., Murray, S.R., and Andersen, G.L. (2003) Comprehensive aligned sequence construction for automated design of effective probes (SASCADE-P) using 16S rDNA. *Bioinformatics* **19**: 1461–8.

DeSantis, T.Z., Stone, C.E., Murray, S.R., Moberg, J.P., and Andersen, G.L. (2005) Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. *FEMS Microbiology Letters* **245**: 271–8.

DeSantis, T.Z., Brodie, E.L., Moberg, J.P., Zubieta, I.X., Piceno, Y.M., and Andersen, G.L. (2007) High-density universal 16S rRNA microarray analysis reveals broader

diversity than typical clone library when sampling the environment. *Microbial Ecology* **53**: 371–83.

De Souza, J.T., Weller, D.M., and Raaijmakers, J.M. (2003) Frequency, diversity, and activity of 2,4-diacetylphloroglucinol-producing fluorescent *Pseudomonas* spp. in Dutch take-all decline soils. *Phytopathology* **93**: 54–63.

Devlin, P.F., Yanovsky, M.J., and Kay, S.A. (2003) A genomic analysis of the shade avoidance response in Arabidopsis. *Plant Physiology* **133**: 1617–29.

Devos, K.M. and Gale, M.D. (2000) Genome relationships: the grass model in current research. *The Plant Cell* **12**: 637–46.

Di Meo, C.A., Wilbur, A.E., Holben, W.E., Feldman, R.A., Vrijenhoek, R.C., and Cary, S.C. (2000) Genetic variation among endosymbionts of widely distributed vestimentiferan tubeworms. *Applied and Environmental Microbiology* **66**: 651–8.

Dicke, M., Van Poecke, R.M.P., and De Boer, J.G. (2003) Inducible indirect defence of plants: from mechanisms to ecological functions. *Basic and Applied Ecology* **4**: 27–42.

Dicke, M., Van Loon, J.J.A., and De Jong, P.W. (2004) Ecogenomics benefits community ecology. *Science* **305**: 618–19.

Dietrich, F.S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pöhlmann, R., Luedi, P., Choi, S. *et al*. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304**: 304–7.

Dillin, A., Crawford, D.L., and Kenyon, C. (2002) Timing requirements for insulin/IGF-1 signaling in *C. elegans*. *Science* **298**: 830–4.

Dimopoulos, G., Casavant, T.L., Chang, S., Scheetz, T., Roberts, C., Donohue, M., Schultz, J., Benes, V., Bork, P., Ansorge, W. *et al*. (2000) *Anopheles gambiae* pilot gene discovery project: Identification of mosquito innate immunity genes from expressed sequence tags generated from immune-competent cell lines. *Proceedings of the National Academy of Sciences USA* **97**: 6619–24.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., Furlan, M., Desnues, C., Haynes, M., Linlin, L., *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature* **452**: 629–32, Corrigendum Vol. 455, p. 830.

Dionne, M.S. and Schneider, D.S. (2002) Screening the fruitfly immune system. *Genome Biology* **3**: reviews 1010.1–10.2.

Dohm, J.C., Lottaz, C., Borodina, T., Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* **36**: e105.

Domsch, K.H. (1984) Effects of pesticides and heavy metals on biological processes in soil. *Plant and Soil* **76**: 367–78.

Doniger, S.W., Kim, H.S., Swain, D., Corcuera, D., Williams, M., Yang, S.-P., and Fay, J.C. (2008) A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genetics* **4**: e1000183.

Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science* **284**: 2124–8.

Doolittle, W.F., Nesbø, C.L., Bapteste, E., and Zhaxybayeva, O. (2008) Lateral gene transfer. In *Evolutionary Genomics and Proteomics*, M. Pagel, and A. Pomiankowski (eds.), Sinauer Associates, Inc., Publishers, Sunderland: 45–79.

Dopazo, H. and Dopazo, J. (2005) Genome-scale evidence of the nematode-arthropod clade. *Genome Biology* **6**: R41.

Dopson, M., Baker-Austin, C., Koppineedi, P.R., and Bond, P.L. (2003) Growth in sulfidic environments: metal resistance mechanisms in acidophilic microorganisms. *Microbiology* **149**: 1959–70.

Dover, G. (1999) Human evolution: our turbulent genes and why we are not chimps. In *The Human Inheritance. Genes, Language, and Evolution*, B. Sykes (ed.). Oxford University Press, Oxford: 75–92.

Drake, J.W., Charlesworth, B., Charlesworth, D., and Crow, J.F. (1998) Rates of spontaneous mutation. *Genetics* **148**: 1667–86.

Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203–18.

Dujon, B. (1996) The yeast genome project: what did we learn? *Trends in Genetics* **12**: 263–70.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E. *et al*. (2004) Genome evolution in yeasts. *Nature* **430**: 35–44.

Dumont, M.G. and Murrell, J.C. (2005) Stable isotope probing—linking microbial identity to function. *Nature Reviews Microbiology* **3**: 499–504.

Duret, L. and Hurst, L.D. (2001) The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. *Molecular Biology and Evolution* **18**: 757–62.

Dwight, S.S., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dolinski, K., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J., Hong, E.L. *et al*. (2004) *Saccharomyces* genome database: underlying principles and organisation. *Briefings in Bioinformatics* **5**: 9–22.

Dykhuizen, D.E., Dean, A.M., and Hartl, D.L. (1987) Metabolic flux and fitness. *Genetics* **115**: 25–31.

Eckardt, N.A. (2002) Alternative splicing and the control of flowering time. *The Plant Cell* **14**: 743–7.

Eddy, S.R. (2005) A model of the statistical power of comparative genome sequence analysis. *PLoS Biology* **3**: e10.

Edwards, R.A. and Rohwer, F. (2005) Viral metagenomics. *Nature Reviews Microbiology* **3**: 504–10.

Edwards, S.V. (2008) A *smörgåsbord* of markers for avian ecology and evolution. *Molecular Ecology* **17**: 945–6.

Egli, D., Selvaraj, A., Yepiskoposyan, H., Zhang, B., Hafen, E., Georgiev, O., and Schaffner, W. (2003) Knockout of 'metal-responsive transcription factor' MTF-1 in *Drosophila* by homologous recombination reveals its central role in heavy metal homeostasis. *EMBO Journal* **22**: 100–8.

Ehrenreich, I.M., Hanzawa, Y., Chou, L., Roe, J.L., Kover, P.X., and Purugganan, M.D. (2009) Candidate gene association mapping of Arabidopsis flowering time. *Genetics* **183**: 325–35.

Eide, D.J. (2001) Functional genomics and metal metabolism. *Genome Biology* **2**: 1028.1–1028.3.

Eisen, J.A. (2000) Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Current Opinion in Genetics & Development* **10**: 606–11.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA* **95**: 14863–8.

El Fantroussi, S., Urakawa, H., Bernhard, A.E., Kelly, J.J., Noble, P.A., Smidt, H., Yershov, G.M., and Stahl, D.A. (2003) Direct profiling of environmental microbial populations by thermal dissociation analysis of native rRNAs hybridized to oligonucleotide arrays. *Applied and Environmental Microbiology* **69**: 2377–82.

El-Assal, S.E.-D., Alonso-Blanco, C., Peeters, A.J.M., Raz, V., and Koornneef, M. (2001) A QTL for flowering time in *Arabidopsis* reveals a novel allele of *CRY2*. *Nature Genetics* **29**: 435–40.

Ellegren, H. and Sheldon, B.C. (2008) Genetic basis of fitness differences in natural populations. *Nature* **452**: 169–75.

Elmer, K.R., Fan, S., Gunter, H.M., Jones, J.C., Boekhoff, S., Karuku, S., and Meyer, A. (2010) Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. *Molecular Ecology* **19** (Suppl. 1): 197–211.

Elton, C. (1927) *Animal Ecology*. Methuen & Co., London.

ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R. *et al*. (2002) Intra- and interspecific variation in primate gene expression patterns. *Science* **296**: 340–3.

Erickson, D.L., Fenster, C.B., Stenøien, H.K., and Price, D. (2004) Quantitative trait locus analyses and the study of evolutionary process. *Molecular Ecology* **13**: 2505–22.

Estoup, A., Jarne, P., and Cornuet, J.-M. (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* **11**: 1591–604.

Evans, J.D. and Wheeler, D.E. (1999) Differential gene expression between developing queens and workers in the honey bee, *Apis mellifera*. *Proceedings of the National Academy of Sciences USA* **96**: 5575–80.

Evans, J.D. and Wheeler, D.E. (2000) Expression profiles during honeybee caste determination. *Genome Biology* **2**: research 0001.1–0001.6.

Evans, J.D. and Wheeler, D.E. (2001) Gene expression and the evolution of insect polyphenisms. *BioEssays* **23**: 62–8.

Excoffier, L., Hofer, T., and Foll, M. (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**: 285–98.

Falkowksi, P.G. and De Vargas, C. (2004) Shotgun sequencing in the sea: a blast from the past? *Science* **304**: 58–60.

Falkowksi, P.G., Katz, M.E., Knoll, A.H., Quigg, A., Raven, J.A., Schofield, O., and Taylor, F.J.R. (2004) The evolution of modern eukaryotic phytoplankton. *Science* **305**: 354–60.

Farnham, P.J. (2009) Insights from genomic profiling of transcription factors. *Nature Reviews Genetics* **10**: 605–16.

Fay, J.C., Wyckoff, G.J., and Wu, C.-I. (2002) Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**: 1024–6.

Fay, J.C., McCullough, H.L., Sniegowksi, P.D., and Eisen, M.B. (2004) Population genetic variation in gene expression is associated with phenotypic variation in *Saccharomyces cerevisiae*. *Genome Biology* **5**: R26.

Featherstone, D.E. and Broadie, K. (2002) Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *BioEssays* **24**: 267–74.

Feder, M.E. and Mitchell-Olds, T. (2003) Evolutionary and ecological functional genomics. *Nature Reviews Genetics* **4**: 649–55.

Feder, M.E. and Walser, J.-C. (2005) The biological limitations of transcriptomics in elucidating stress and stress responses. *Journal of Evolutionary Biology* **18**: 901–10.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., Turcatti, G. (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research* **34**: e22.

Fell, D.A. (1992) Metabolic Control Analysis: a survey of its theoretical and experimental development. *Biochemical Journal* **286**: 313–30.

Fields, P.A. and Houseman, D.E. (2004) Decreases in activation energy and substrate affinity on cold-adapted A_4-lactate dehydrogenase: evidence from the Antarctic notothenioid fish *Chaenocephalus aceratus*. *Molecular Biology* and *Evolution* **21**: 2246–55.

Finkel, T. and Holbrook, N.J. (2000) Oxidants, oxidative stress and the biology of ageing. *Nature* **408**: 239–47.

Fitch, D.H.A. and Thomas, W.K. (1997) Evolution. In *C. Elegans II*, D.L. Riddle, T. Blumenthal, B.J. Meyer, and J.R. Priess (eds), Cold Spring Harbor Monograph Series. Cold Spring Harbor Press, Cold Spring Harbor: 815–50.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M. *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496–512.

Flowers, J.M., Hanzawa, Y., Hall, M.C., Moore, R.C., and Purugganan, M. (2009) Population genomics of the *Arabidopsis thaliana* flowering time gene network. *Molecular Biology and Evolution* **26**: 2475–86.

FlyBase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Research* **31**: 172–5.

Fondon III J.W. and Garner, H.R. (2004) Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences USA* **101**: 18058–63.

Ford, M.J. (2002) Applications of selective neutrality tests to molecular ecology. *Molecular Ecology* **11**: 1245–62.

Foster, S.A. and Baker, J.A. (2004) Evolution in parallel: new insights from a classic system. *Trends in Ecology and Evolution* **19**: 456–9.

Fox Keller, E. (2000) *The Century of the Gene*. Harvard University Press, Cambridge, MA.

Francis, C.A., Beman, J.M., and Kuypers, M.M.M. (2007) New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *The ISME Journal* **1**: 19–27.

Freedman, A.H., Thomassen, H.A., Buermann, W., and Smith, T.B. (2010) Genomic signals of diversification along ecological gradients in a tropical lizard. *Molecular Ecology* **19**: 3773–88.

Friedman, D.B. and Johnson, T.E. (1988) A mutation in the *age-1* gene in *Caenorhabditis elegans* lengthens life and reduces hermaphrodite fertility. *Genetics* **118**: 75–86.

Friedrich, C.G., Rother, D., Bardischewsky, F., Quentmeier, A., and Fischer, J. (2001) Oxidation of reduced inorganic sulfur compounds by bacteria: emergence of a common mechanism? *Applied and Environmental Microbiology* **67**: 2873–82.

Friedrich, C.G., Bardischewsky, F., Rother, D., Quentmeier, A., and Fischer, J. (2005) Prokaryotic sulfur oxidation. *Current Opinion in Microbiology* **8**: 253–9.

Fuchs, S., Bundy, J.G., Davies, S.K., Viney, J.M., Swire, J.S., and Leroi, A.M. (2010) A metabolic signature of long life in *Caenorhabditis elegans*. *BMC Biology* **8**: 14.

Fullthorpe, R.R., Roesch, L.F.W., Riva, A., and Triplett, E.W. (2008) Distantly sampled soils carry few species in common. *The ISME Journal* **2**: 901–10.

Fütterer, O., Angelov, A., Liesegang, H., Gottschalk, G., Schleper, C., Schepers, B., Dock, C., Antranikian, G., and Liebl, W. (2004) Genome sequence of *Picrophilus torridus* and its implications for life around pH 0. *Proceedings of the National Academy of Sciences USA* **101**: 9091–6.

Gabor, E.M., Alkema, W.B.L., and Janssen, D.B. (2004) Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environmental Microbiology* **6**: 948–58.

Gagneur, J., Sinha, H., Perocchi, F., Bourgon, R., Huber, W., and Steinmetz, L.M. (2009) Genome-wide allele- and strand-specific expression profiling. *Molecular Systems Biology* **5**: 274.

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.-J., Smirnov, S., Purcell, S. *et al*. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**: 859–68.

Galibert, F., Finan, T.M., Long, S.R., Pühler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M.J., Becker, A., Boistard, P. *et al*. (2001) The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293**: 668–72.

Gallardo, M.H., Bickham, J.W., Honeycutt, R.L., Ojeda, R.A., and Köhler, N. (1999) Discovery of tetraploidy in a mammal. *Nature* **401**: 341.

Gallardo, M.H., Kausel, G., Jiménez, A., Bacquet, C., González, C., Figuerora, J., Köhler, N., and Ojeda, R. (2004) Whole-genome duplications in South American desert rodents (Octodontidae). *Biological Journal of the Linnean Society* **82**: 443–51.

Gans, J., Wolinsky, M., and Dunbar, J. (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**: 1387–90.

Gao, H., Yang, Z.K., Gentry, T.J., Wu, L., Schadt, C.W., and Zhou, J. (2007) Microarray-based analysis of microbial community RNAs by whole-community RNA amplification. *Applied and Environmental Microbiology* **73**: 563–71.

Garbeva, P., Van Veen, J.A., and Van Elsas, J.D. (2004) Microbial diversity in soil: selection of microbial populations by plant and soil type and implications for disease suppressiveness. *Annual Review of Phytopathology* **42**: 243–70.

Gasch, A.P. and Werner-Washburne, M. (2002) The genomics of yeast responses to environmental stress. *Functional and Integrative Genomics* **2**: 181–92.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* **11**: 4241–57.

Gems, D. and Partridge, L. (2001) Insulin/IGF signalling and ageing: seeing the bigger picture. *Current Opinion in Genetics & Development* **11**: 287–92.

Gems, D. and McElwee, J.J. (2003) Microarraying mortality. *Nature* **424**: 259–61.

Gentry, T.J., Wickham, G.S., Schadt, C.W., He, Z., and Zhou, J. (2006) Microarray applications in microbial research. *Microbial Ecology* **52**: 159–75.

Getz, W.M., Westerhoff, H.V., Hofmeyr, J.-H.S., and Snoep, J.L. (2003) Control analysis of trophic chains. *Ecological Modelling* **168**: 153–71.

Giannakou, M.E., Goss, M., Jünger, M.A., Hafen, E., Leevers, S.J., and Partridge, L. (2004) Long-lived *Drosophila* with over-expressed dFOXO in adult fat body. *Science* **305**: 361.

Gibson, G. (2008) The environmental contribution to gene expression profiles. *Nature Reviews Genetics* **9**: 575–81.

Gibson, G. and Muse, S.V. (2002) *A Primer of Genome Science*. Sinauer Associates, Sunderland, MA.

Gilad, Y., Oshlack, A., Smyth, G.K., Speed, T.P., and White, K.P. (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440**: 242–5.

Gilbert, S.F. and Epel, D. (2009) *Ecological Developmental Biology. Integrating Epigenetics, Medicine and Evolution*. Sinauer Associates, Inc., Sunderland.

Gillespie, D.E., Brady, S.F., Bettermann, A.D., Cianciotto, N.P., Liles, M.R., Rondon, M.R., Clardy, J., Goodman, R.M., and Handelsman, J. (2002) Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Applied and Environmental Microbiology* **68**: 4301–6.

Gillott, C. (1980) *Entomology*. Plenum Press, New York.

Girardot, F., Monnier, V., and Tricoire, H. (2004) Genome wide analysis of common and specific stress responses in adult drosophila melanogaster. *BMC Genomics* **5**: 1471–2164/5/74.

Gladyshev, E.A., Meselson, M., and Arhipova, I.R. (2008) Massive horizontal gene transfer in bdelloid rotifers. *Science* **320**: 1210–3.

Glomus Genome Consortium (2008) The long hard road to a completed *Glomus intraradices* genome. *New Phytologist* **180**: 747–50.

Goff, S.A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92–100.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. *et al*. (1996) Life with 6000 genes. *Science* **274**: 546–67.

Golden, T.R. and Melov, S. (2004) Microarray analysis of gene expression with age in individual nematodes. *Aging Cell* **3**: 111–24.

Gómez-Mena, C., Piñeiro, M., Franco-Zorilla, J.M., Salinas, J., Coupland, G., and Martínez-Zapater, J.M (2001) *early bolting in short days*: An Arabidopsis mutation that causes early flowering and partially suppresses the floral phenotype of leafy. *The Plant Cell* **13**: 1011–24.

Gong, P., Guan, X., Inouye, L.S., Pirooznia, M., Indest, K.J., Athow, R.S., Deng, Y., and Perkins, E.J. (2007) Toxicogenomic analysis provides new insights into molecular mechanisms of 2,4,6-trinitrotoluene in *Eisenia fetida*. *Environmental Science and Technology* **41**: 8195–202.

Gonzalez, P., Dominique, Y., Massabuau, J.C., Boudou, A., and Bourdineaud, J.P. (2005) Comparative effects of dietary methylmercury on gene expression in liver, skeletal muscle, and brain of the zebrafish (*Danio rerio*). *Environmental Science and Technology* **39**: 3972–80.

Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B.S., Cao, Y., Askenazi, M., Halling, C. *et al*. (2001) Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* **294**: 2323–8.

Görner, W., Durchschlag, E., Martinez-Pastor, M.T., Estruch, F., Ammerer, G., Hamilton, B., Ruis, H., and Schüller, C. (1998) Nuclear localization of the C_2H_2 zinc finger protein Mns2p is regulated by stress and protein kinase A activity. *Genes & Development* **12**: 586–97.

Goto, S.G. and Denlinger, D.L. (2002) Short-day and long-day expression patterns of genes involved in the flesh fly clock mechanism: *period*, *timeless*, *cycle* and *cryptochrome*. *Journal of Insect Physiology* **48**: 803–16.

Govind, S. and Nehm, R.H. (2004) Innate immunity in fruit flies: a textbook example of genomic recycling. *PLoS Biology* **2**: 1065–8.

Gould, S.J. and Lewontin, R.C. (1979) The spandrels of San Marco and the panglossian paradigm. *Proceedings of the Royal Society of London, Series B* **205**: 581–98.

Gracey, A.Y. (2007) Interpreting physiological responses to environmental change through gene expression profiling. *The Journal of Experimental Biology* **209**: 1584–92.

Gracey, A.Y. and Cossins, A.R. (2003) Application of microarray technology in environmental and comparative physiology. *Annual Review of Physiology* **65**: 231–59.

Gracey, A.Y., Troll, J.V., and Somero, G.N. (2001) Hypoxia-induced gene expression profiling in the euryoxic fish *Gillichthys mirabilis*. *Proceedings of the National Academy of Sciences USA* **98**: 1993–8.

Grandison, R.C., Piper, M.D.W., and Partridge, L. (2009) Amino-acid imbalance explains extension of lifespan by dietary restriction in *Drosophila*. *Nature* **462**: 1061–4.

Grapevine Consortium: The French-Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests hexaploidization in major angiosperm phyla. *Nature* **499**: 463–7.

Graur, D. and Li, W.-H. (2000) *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Gray, M.W., Burger, G., and Lang, B.F. (1999) Mitochondrial evolution. *Science* **283**: 1476–81.

Gray, N.D. and Head, I.M. (2001) Linking genetic identity and function in communities of uncultured bacteria. *Environmental Microbiology* **3**: 481–92.

Graze, R.M., McIntyre, L.M., Main, B.J., Wayne, M.L., and Nuzhdin, S.V. (2009) Regulatory divergence in *Drosophila melanogaster* and *D. simulans*, a genomewide analysis of allele-specific expression. *Genetics* **183**: 547–61.

Greer, C.W., Whyte, L.G., Lawrence, J.R., Masson, L., and Brousseau, R. (2001) Genomic technologies for environmental science. *Environmental Science and Technology* **35**: 360A–6A.

Gregory, T.R. (ed.) (2005) *The Evolution of the Genome*. Elsevier Academic Press, Amsterdam.

Gregory, T.R. and Johnston, J.S. (2008) Genome size diversity in the family Drosophilidae. *Heredity* **101**: 228–38.

Greilhuber, J., Dolezel, J., Lysák, M.A., and Bennett, M.D. (2005) The origin, evolution and proposed stabilization of the terms 'Genome size' and 'C-value' to describe nuclear DNA contents. *Annals of Botany* **95**: 255–60.

Grewal, S.I.S. and Jia, S. (2007) Heterochromatin revisited. *Nature Reviews Genetics* **8**: 35–46.

Griffiths, S., Dunford, R.P., Coupland, G., and Laurie, D.A. (2003) The evolution of *CONSTANS*-like gene families in barley, rice, and *Arabidopsis*. *Plant Physiology* **131**: 1855–67.

Gross, C., Kelleher, M., Iyer, V.R., Brown, P.O., and Winge, D.R. (2000) Identification of the copper regulon in *Saccharomyces cerevisae* by DNA microarrays. *Journal of Biological Chemistry* **275**: 32310–6.

Gross, L. (2007) Untapped bounty: sampling the seas to survey microbial biodiversity. *PLoS Biology* **5**: e85.

Grossman, A.R., Harris, E.E., Hauser, C., Lefebvre, P.A., Martinez, D., Rokhsar, D., Shrager, J., Silflow, C.D., Stern, D., Vallon, O., and Zhang, Z. (2003) *Chlamydomonas reinhardtii* at the crossroads of genomics. *Eukaryotic Cell* **2**: 1137–50.

Guarante, L. and Kenyon, C. (2000) Genetic pathways that regulate ageing in model organisms. *Nature* **408**: 255–61.

Guiliano, D.B., Hall, N., Jones, S.J.M., Clark, L.N., Corton, C.H., Barrell, B.G., and Blaxter, M.L. (2003) Conservation of long-range synteny and microsynteny between the genomes of two distantly related nematodes. *Genome Biology* **3**: research/0057/I–14.

Gutman, B.L. and Niyogi, K.K. (2004) Chlamydomonas and Arabidopsis. A dynamic duo. *Plant Physiology* **135**: 607–10.

Gutteling, E.W., Riksen, J.A.G., and Kammenga, J.E. (2007) Mapping phenotypic plasticity and genotype-environment interaction affecting life-history traits in *Caenorhabditis elegans*. *Heredity* **98**: 28–37.

Haas, B.J., Kamoun, S., Zody, M.C., Jiang, R.H.Y., Handsaker, R.E., Cano, L.M., Grabherr, M., Kodira, C.D., Raffaele, S., Torto-Alalibo, T., *et al.* (2009) Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* **461**: 393–8.

Haddrill, P.R., Bachtrog, D., and Andolfatto, P. (2008) Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Molecular Biology* and *Evolution* **25**: 1825–34.

Hagenblad, J., Olsson, M., Parker, H.G., Ostrander, E.A., and Ellegren, H. (2009) Population genomics of the inbred Scandinavian wolf. *Molecular Ecology* **18**: 1341–51.

Halitschke, R., Ziegler, J., Keina ¨nen, M., and Baldwin, I.T. (2004) Silencing of hydroperoxide lyase and allene oxide synthase reveals substrate and defense signaling crosstalk in *Nicotiana attenuata*. *The Plant Journal* **40**: 35–46.

Hamadeh, H.K., Bushel, P.R., Jayadev, S., Martin, K., DiSorbo, O., Sieber, S., Bennett, L., Tennant, R., Stoll, R., Barrett, J.C. *et al.* (2002) Gene expression analysis reveals chemical-specific profiles. *Toxicological Sciences* **67**: 219–31.

Handelsman, J. (2004) Metagenomics: applications of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* **68**: 669–85.

Handelsman, J., Liles, M., Mann, D., Riesenfeld, C., and Goodman, R.M. (2002) Cloning the metagenome: culture-independent access to the diversity and functions of the uncultivated microbial world. *Methods in Microbiology* **33**: 241–55.

Haq, F., Mahoney, M., and Koropatnick, J. (2003) Signaling events for metallothionein induction. *Mutation Research* **533**: 211–26.

Harborne, J.B. (1997) *Introduction to Ecological Biochemistry*. Academic Press, London.

Harmer, S.L., Hogenesch, J.B., Straume, M., Chang, H.-S., Han, B., Zhu, T., Wang, X., Kreps, J.A., and Kay, S.A. (2000) Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. *Science* **290**: 2110–3.

Harries, J.E., Sheahan, D.A., Jobling, S., Matthiessen, P., Neall, P., Sumpter, J.P., Tylor, T., and Zaman, N. (1997) Estrogenic activity in five United Kingdom rivers detected by measurement of vitellogenesis in caged male trout. *Environmental Toxicology and Chemistry* **16**: 534–42.

Harris, E.H. (2001) *Chlamydomonas* as a model organism. *Annual Review of Plant Physiology and Plant Molecular Biology* **52**: 363–406.

Harris, T.W., Chen, N., Cunningham, F., Tello-Ruiz, M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Chan, J. *et al*. (2004) WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Research* **32**: D411–7.

Hartl, D.L. and Clark, A.G. (1997) *Principles of Population Genetics*, 3rd edn. Sinauer Associates, Sunderland, MA.

Hayama, R. and Coupland, G. (2004) The molecular basis of diversity in the photoperiodic flowering responses of Arabidopsis and rice. *Plant Physiology* **135**: 677–84.

Hayama, R., Yokoi, S., Tamaki, S., Yano, M., and Shimamoto, K. (2003) Adaptation of photoperiodic control pathways produces short-day flowering in rice. *Nature* **422**: 719–22.

He, Y., Michaels, S.D., and Amasino, R.M. (2003) Regulation of flowering time by histone acetylation in *Arabidopsis*. *Science* **302**: 1751–4.

He, Z., Gentry, T.J., Schadt, C.W., Wu, L., Liebich, J., Chong, S.C., Huang, Z., Wu, W., Gu, B., Jardine, P., *et al.* (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *The ISME Journal* **1**: 67–77.

He, Z., Deng, Y., Van Nostrand, J.D., Tu, Q., Xu, M., Hemme, C.L., Li, X., Wu, L., Gentry, T.J., Yin, Y., *et al.* (2010) GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. *The ISME Journal* **4**: 1167–79.

Head, I.M., Jones, D.M., and Röling, W.F.M. (2006) Marine microorganisms make a meal of oil. *Nature Reviews Microbiology* **4**: 173–82.

Heckel, D.G. (2003) Genomics in pure and applied entomology. *Annual Review of Entomology* **48**: 235–60.

Hedges, S.B. (2002) The origin and evolution of model organisms. *Nature Reviews Genetics* **3**: 838–49.

Heemsbergen, D.A., Berg, M.P., Loreau, M., Van Hal, J.R., Faber, J.H., and Verhoef, H.A. (2004) Biodiversity effects on soil processes explained by interspecific functional dissimilarity. *Science* **306**: 1019–20.

Heidel, A.J. and Baldwin, I.T. (2004) Microarray analysis of salicylic acid- and jasmonic acid-signalling in responses of *Nicotiana attenuata* to attack by insects from multiple feeding guilds. *Plant, Cell and Environment* **27**: 1362–73.

Heidelberg, J.F., Paulsen, I.T., Nelson, K.E., Gaidos, E.J., Nelson, W.C., Read, T.D., Eisen, J.A., Seshadri, R., Ward, N., Methe, B. *et al*. (2002) Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nature Biotechnology* **20**: 1118–23.

Heidelberg, J.F., Seshadri, R., Haveman, S.A., Hemme, C.L., Paulsen, I.T., Kolonay, J.F., Eisen, J.A., Ward, N., Methe, B., Brinkac, L.M. *et al*. (2004) The genome sequence of the anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. *Nature Biotechnology* **22**: 554–9.

Hekimi, S. and Guarante, L. (2003) Genetics and the specificity of the aging process. *Science* **299**: 1351–4.

Held, M., Gase, K., and Baldwin, I.T. (2004) Microarrays in ecological research: a case study of a cDNA microarray for plant-herbivore interactions. *BMC Ecology* **4**: 13.

Hellsten, U., Harland, R.M., Gilchrist, M.J., Hendrix, D., Jurka, J., Kapitonov, V.V., Ovcharenko, I., Putnam, N.H., Shu, S., Taher, L., *et al.* (2010) The genome of the western clawed frog *Xenopus tropicalis*. *Science* **328**: 633–6.

Henikoff, S., Furuyama, T., and Ahmad, K. (2004) Histone variants, nucleosome assembly and epigenetic inheritance. *Trends in Genetics* **20**: 320–6.

Henne, A., Schmitz, R.A., Bömeke, M., Gottschalk, G., and Daniel, R. (2000) Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. *Applied and Environmental Microbiology* **66**: 3113–16.

Herrera, C.M. and Bazaga, P. (2008) Population-genomic approach reveals adaptive floral divergence in discrete populations of a hawk moth-pollinated violet. *Molecular Ecology* **17**: 5378–90.

Hill, A.A., Hunter, C.P., Tsung, B.T., Tucker-Kellogg, G., and Brown, E.L. (2000) Genomic analysis of gene expression in *C. elegans*. *Science* **290**: 809–12.

Hill, R.W., Wyse, G.A., and Anderson, M. (2008) *Animal Physiology. Second Edition*. Sinauer Associates, Sunderland.

Hines, A., Oladiran, G.S., Bignell, J.P., Stentiford, G.D., and Viant, M.R. (2007) Direct sampling of organisms from the field and knowledge of their phenotype: Key recommendations for environmental metabolomics. *Environmental Science and Technology* **41**: 3375–81.

Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Wynshaw-Boris, A., Sugiyama, F., Takahashi, S., Yagami, K.-I., and Yoshiki, A. (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **423**: 91–6.

Hodgkin, J., Plasterk, R.H.A., and Waterston, R.H. (1995) The nematode *Caenorhabditis elegans* and its genome. *Science* **270**: 410–14.

Hoekstra, H.E. and Coyne, J.A. (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**: 995–1016.

Hoffmann, A.A., and Hercus, M.J. (2000) Environmental stress as an evolutionary force. *BioScience* **50**: 217–26.

Hoffmann, J.A. and Reichhart, J.-M. (2002) *Drosophila* innate immunity: an evolutionary perspective. *Nature Immunology* **3**: 121–6.

Hoffmann, M.H. (2002) Biogeography of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae). *Journal of Biogeography* **29**: 125–34.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., and Cresko, W.A. (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics* **6**: e1000862.

Hohmann, S. and Mager, W.H., eds. (2003) *Yeast Stress Responses. Topics in Curent Genetics, Vol. 1*, Springer-Verlag, Berlin.

Holsinger, K.E. and Weir, B.S. (2009) Genetics in geographically structure populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics* **10**: 639–50.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M.C., Wides, R. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**: 129–49.

Holt, S.J. and Riddle, D.L. (2003) SAGE surveys *C. elegans* carbohydrate metabolism: evidence for an anaerobic shift in the long-lived dauer larva. *Mechanisms of Ageing and Development* **124**: 770–800.

Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R., and Federoff, N.V. (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proceedings of the National Academy of Sciences USA* **97**: 8409–14.

Holzenberger, M., Dupont, J., Ducos, B., Leneuve, P., Géloe ¨n, A., Even, P.C., Cervera, P., and Le Bouc, Y. (2003) IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. *Nature* **421**: 182–7.

Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**: 931–49.

Hopkin, S.P. (1989) *Ecophysiology of Metals in Terrestrial Invertebrates*. Elsevier Applied Science, London.

Horner-Devine, M., Carney, K.M., and Bohannan, B.J.M. (2004) An ecological perspective on bacterial biodiversity. *Proceedings of the Royal Society of London, Series B* **271**: 113–22.

Houthoofd, K., Braeckman, B.P., Lenaerts, I., Brys, K., De Vreese, A., Van Eygen, S., and Vanfleteren, J.R. (2002) Axenic growth upregulates mass-specific metabolic rate, stress resistance, and extends life span in *Caenorhabditis elegans*. *Experimental Gerontology* **37**: 1369–76.

Houthoofd, K., Braeckman, B.P., Johnson, T.E., and Vanfleteren, J.R. (2003) Life extension via dietary restriction is independent of the Ins/IGF-1 signalling pathway in *Caenorhabditis elegans*. *Experimental Gerontology* **38**: 947–54.

Hoy, M.A. (1994) *Insect Molecular Genetics*. Academic Press, San Diego.

Hsu, A.-L., Murphy, C.T., and Kenyon, C. (2003) Regulation of aging and age-related disease by DAF-16 and heat-shock factor. *Science* **300**: 1142–5.

Huang, J., Mullapudi, N., Lancto, C.A., Scott, M., Abrahamsen, M.S., and Kissinger, J.C. (2004) Phylogenomic evidence supports past endosymbiosis, intracelullar and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biology* **5**: R88.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., Ni, P., *et al.* (2009) The genome of the cucumber, *Cucumis sativus* L. *Nature Genetics* **41**: 1275–81.

Huber, H., Hohn, M.J., Rachel, R., Fuchs, T., Wimmer, V.C., and Stetter, K.O. (2002) A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**: 63–7.

Hughes, A.L. (1999) *Adaptive Evolution of Genes and Genomes*. Oxford University Press, New York.

Hughes, J.B., Hellmann, J.J., Ricketts, T.H., and Bohannan, B.J. (2001) Counting the uncountable: statistical

approaches to estimating diversity. *Applied and Environmental Microbiology* **67**: 4399–406.

Hughes, K.A., Ayroles, J.F., Reedy, M.M., Drnevich, J.M., Rowe, K.C., Ruedi, E.A., Cáceres, C.E., and Paige, K.N. (2006) Segregating variation in the transcriptome: *cis* regulation and additivity of effects. *Genetics* **173**: 1374–1355.

Hui, D., Iqbal, J., Lehman, K., Gase, K., Saluz, H.P., and Baldwin, I.T. (2003) Molecular interactions between the specialist herbivore *Manduca sexta* (Lepidoptera, Sphingidae) and its natual host *Nicotiana attenuata*: V. Microarray analysis and further characterization of large-scale changes in herbivore-induced mRNAs. *Plant Physiology* **131**: 1877–93.

Hui, J.H.L., Raible, F., Korchagina, N., Dray, N., Samain, S., Magdelenat, G., Jubin, C., Segurens, B., Balavoine, G., Arendt, D., *et al.* (2009) Features of the ancestral bilaterian inferred from *Platynereis dumerilii* ParaHox genes. *BMC Biology* **7**: 43.

Hulbert, A.J., Clancy, D.J., Mair, W., Braeckman, B.P., Gems, D., and Partridge, L. (2004) Metabolic rate is not reduced by dietary-restriction or by lowered insulin/IGF-1 signalling and is not correlated with individual lifespan in *Drosophila melanogaster*. *Experimental Gerontology* **39**: 1137–43.

Hurst, L.D. (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics* **18**: 486–7.

Hurst, L.D. (2009) Evolutionary genomics and the reach of selection. *Journal of Biology* **8**: 12.

Hurtado, L.A., Lutz, R.A., and Vrijenhoek, R.C. (2004) Distinct patterns of genetic differentiation among annelids of eastern Pacific hydrothermal vents. *Molecular Ecology* **13**: 2603–15.

Hutchinson, G.E. (1957) Concluding remarks. In *Cold Spring Harbor Symposia on Quantitative Biology. Volume XXII Population Studies: Animal Ecology and Demography*. Cold Spring Harbor Press, New York: 415–27.

Ideker, T., Galitski, T., and Hood, L. (2000) A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics* **2**: 343–72.

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695–716.

Irving, P., Troxler, L., Heuer, T.S., Belvin, M., Kopczynski, C., Reichhart, J.-M., Hoffmann, J.A., and Hetru, C. (2001) A genome-wide analysis of immune responses in *Drosophila*. *Proceedings of the National Academy of Sciences USA* **98**: 15119–24.

Jablonka, E. and Lamb, M.J. (2002) The changing concept of epigenetics. *Annals of the New York Academy of Sciences* **981**: 82–96.

Jablonka, E. and Raz, G. (2009) Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *The Quarterly Review of Biology* **84**: 131–76.

Jackson, R.B., Linder, C.R., Lynch, M., Purugganan, M.D., Somerville, S., and Thayer, S.S. (2002) Linking molecular insight and ecological research. *Trends in Ecology and Evolution* **19**: 409–14.

Jacob, F. (1977) Evolution and tinkering. *Science* **196**: 1161–6.

Jacobson, D.J., Powell, A.J., Dettman, J.R., Saenz, G.S., Barton, M., Hiltz, M.D., Dvorachek Jr, W.H., Glass, N.L., Taylor, J.W., and Natvig, D.O. (2004) *Neurospora* in temperate forests of western North America. *Mycologia* **96**: 66–74.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Maucell, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A. *et al*. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate protokaryotype. *Nature* **431**: 946–57.

Jain, R., Rivera, M.C., and Lake, J.A. (1999) Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences USA* **96**: 3801–6.

Jaiswal, A.K. (2004) Nrf2 signaling in coordinated activation of antioxidant gene expression. *Free Radical Biology & Medicine* **36**: 1199–207.

Jansen, R.C. and Stam, P. (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**: 1447–55.

Jansen, R.C. and Nap, J.-P. (2001) Genetical genomics: the added value from segregation. *Trends in Genetics* **17**: 388–91.

Janssen, P.H. (2006) Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Applied and Environmental Microbiology* **72**: 1719–28.

Janssens, T.K.S., Mariën, J., Cenijn, P., Legler, J., Van Straalen, N.M., and Roelofs, D. (2007) Recombinational micro-evolution of functionally different metallothionein promoter alleles from *Orchesella cincta*. *BMC Evolutionary Biology* **7**: 88.

Janssens, T.K.S., Del Rio Lopez, R., Mariën, J., Timmermans, M.J.T.N., Montagne-Wajer, M., Van Straalen, N.M., and Roelofs, D. (2008) Comparative population analysis of metallothionein promoter alleles suggests stress-induced microevolution in the field. *Environmental Science* and *Technology* **42**: 3873–8.

Jensen, P.R., Snoep, J.L., Molenaar, D., Van Heeswijk, W.C., Khodolenko, B.N., Van der Gugten, A.A., and Westerhoff, H.V. (1995) Molecular biology for flux control. *Biochemical Society Transactions* **23**: 367–70.

Jessup, C.M., Kassen, R., Forde, S.E., Kerr, B., Buckling, A., Rainey, P.B., and Bohannan, B.J.M. (2004) Big questions, small worlds: microbial model systems in ecology. *Trends in Ecology and Evolution* **19**: 189–97.

Jetten, M.S.M. (2008) The microbial nitrogen cycle. *Environmental Microbiology* **10**: 2903–9.

Jeukens, J., Bittner, D., Knudsen, R., and Bernatchez, L. (2008) Candidate genes and adaptive radiation: insight from transcriptional adaptation to the limnetic niche among coregonine fishes (*Coregonus* sp., Salmonidae). *Molecular Biology* and *Evolution* **26**: 155–66.

Jiang, M., Ryu, J., Kiraly, M., Duke, K., Reinke, V., and Kim, S.K. (2001) Genome-wide analysis of developmental and sex-regulated expression profiles in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences USA* **98**: 218–23.

Jin, W., Riley, R.M., Wolfinger, R.D., White, K.P., Passador-Gurgel, G., and Gibson, G. (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics* **29**: 389–95.

Jones, S.J.M., Riddle, D.L., Pouzyrev, A.T., Velculescu, V.E., Hillier, L., Eddy, S.R., Stricklin, S.L., Baillie, D.L., Waterston, R., and Marra, M.A. (2001) Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans*. *Genome Research* **11**: 1346–52.

Judice, C., Hartfelder, K., and Pereira, G.A.G. (2004) Caste-specific gene expression in the stingless bee *Melipone quadrifasciata*—are there common patterns in highly social bees? *Insectes Sociaux* **51**: 352–8.

Kacser, H. and Burns, J.A. (1995) The control of flux (with additional comments by H. Kacser and D.A. Fell). *Biochemical Society Transactions* **23**: 341–66.

Kaeberlein, T., Lewis, K., and Epstein, S.S. (2002) Isolating 'uncultivable' microorganisms in pure culture in a simulated natural environment. *Science* **296**: 1127–9.

Kaltz, O. and Bell, G. (2002) The ecology and genetics of fitness in *Chlamydomonas*. XII. Repeated sexual episodes increase rates of adaptation to novel environment. *Evolution* **56**: 1743–53.

Kammenga, J.E., Doroszuk, A., Riksen, J.A.G., Hazendonk, E., Spiridon, L., Petrescu, A.-J., Tijsterman, M., Plasterk, R.H.A., and Bakker, J. (2007) A *Caenorhabditis elegans* wild type defies the temperature-size rule owing to a single nucleotide polymorphism in *tra-3*. *PLoS Genetics* **3**: e34.

Kammenga, J.E., Herman, M.A., Ouborg, N.J., Johnson, L., and Breitling, R. (2007) Microarray challenges in ecology. *Trends in Ecology and Evolution* **22**: 273–9.

Kammenga, J.E., Phillips, P.C., De Bono, M., and Doroszuk, A. (2008) Beyond induced mutants: using worms to study natural variation in genetic pathways. *Trends in Genetics* **24**: 178–85.

Kämper, J., Kahmann, R., Bölker, M., Ma, L.-J., Brefort, T., Saville, B.J., Banuett, F., Kronstadt, J.W., Gold, S.E., Müller, O., *et al.* (2006) Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* **444**: 97–101.

Kanao, T., Fukui, T., Atomi, H., and Imanaka, T. (2001) ATP-citrate lyase from the green sulfur bacterium *Chlorobium limicola* is a heteromeric enzyme composed of two distinct gene products. *European Journal of Biochemistry* **268**: 1670–8.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S. *et al*. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Research* **3**: 109–36.

Kannan, N., Taylor, S.S., Zhai, Y., Venter, J.C., and Manning, G. (2007) Structural and functional diversity of the microbial kinome. *PLoS Biology* **5**: e17.

Kassen, R. and Rainey, P.B. (2004) The ecology and genetics of microbial diversity. *Annual Review of Microbiology* **58**: 207–31.

Kasuga, M., Liu, G., Miura, S., Yamaguchi-Shinozaki, K., and Shinozaki, K. (1999) Improving plant drought, salt, and freezing tolerance by gene transfer of a single stress-inducible transcription factor. *Nature Biotechnology* **17**: 287–91.

Katschinski, D.M. and Glueck, S.B. (2003) Hot worms can handle heavy metal. Focus on 'HIF-1 is required for heat acclimation in the nematode *Caenorhabditis elegans*'. *Physiological Genomics* **14**: 1–2.

Katz, L.A. and Bhattacharya, D. (eds.) (2008) *Genomics and Evolution of Microbial Eukaryotes*, Oxford University Press, Oxford.

Kawasaki, S., Borchert, C., Deyholos, M., Wang, H., Brazille, S., Kawai, K., Galbraith, D., and Bohnert, H.J. (2001) Gene expression profiles during the initial phase of salt stress in rice. *The Plant Cell* **13**: 889–905.

Keeling, P.J., Burger, G., Durnford, D.G., Lang, B.F., Lee, R.W., Pearlman, R.E., Roger, A.J., and Gray, M.W. (2005) The tree of eukaryotes. *Trends in Ecology and Evolution* **20**: 670–6.

Keeling, P.J. and Palmer, J.D. (2008) Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* **9**: 605–18.

Keightley, P.D. and Eyre-Walker, A. (1999) Terumi Mukai and the riddle of deleterious mutation rates. *Genetics* **153**: 515–23.

Keightley, P.D. and Charlesworth, B. (2005) Genetic instability of *C. elegans* comes naturally. *Trends in Genetics* **21**: 67–70.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241–54.

Kellis, M., Birren, B.W., and Lander, E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**: 617–24.

Kenyon, C. (2010) The genetics of ageing. *Nature* **464**: 504–12, Erratum Vol. **467**: 622.

Kenyon, C., Chang, J., Gensch, E., Rudner, A., and Tabtiang, R. (1993) A *C. elegans* mutant that lives twice as long as wild type. *Nature* **366**: 461–4.

Kerstens, H.H.D., Crooijmans, R.P.M.A., Veenendaal, A., Dibbits, B.W., Chin-A-Woeng, T.F.C., Den Dunnen, J.T., and Groenen, M.A.M. (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics* **10**: 479.

Kessler, A. and Baldwin, I.T. (2002) Plant responses to insect herbivory: the emerging molecular analysis. *Annual Review of Plant Biology* **53**: 299–328.

Kessler, A., Halitschke, R., and Baldwin, I.T. (2004) Silencing the jasmonate cascade: induced plant defenses and insect populations. *Science* **305**: 665–8.

Keurentjes, J.J.B., Fu, J., Terpstra, I.R., Garcia, J.M., Van den Ackerveken, G., Snoek, L.B., Peeters, A.J.M., Vreugdenhil, D., Koornneef, M., and Jansen, R.C. (2007) Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 1708–13.

Kielak, A., Pijl, A.S., Van Veen, J.A., and Kowalchuk, G.A. (2009) Phylogenetic diversity of *Acidobacteria* in a former agricultural soil. *The ISME Journal* **3**: 378–82.

Kiers, E.T., Palmer, T.M., Ives, A.R., Bruno, J.F., and Bronstein, J.L. (2010) Mutualisms in a changing world: an evolutionary perspective. *Ecology Letters* **13**: 1459–74.

Kiers, E.T. and Van der Heijden, M.G.A. (2006) Mutualistic stability in the arbuscular mycorrhizal symbiosis: exploring hypotheses of evolutionary cooperation. *Ecology* **87**: 1627–36.

Kim, H.-J., Hyun, Y., Park, J.-Y., Park, M.-J., Park, M.-K., Kim, M.D., Kim, H.-J., Lee, M.H., Moon, J., Lee, I., and Kim, J. (2004) A genetic link between cold responses and flowering time through *FVE* in *Arabidopsis thaliana*. *Nature Genetics* **36**: 167–71.

Kim, K.W., Shin, J.-H., Moon, J., Kim, M., Lee, J., Park, M.-C., and Lee, I. (2003) The function of flowering time gene *AGL20* is conserved in crucifers. *Molecules and Cells* **16**: 136–41.

Kim, P.M. and Tidor, B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Research* **13**: 1706–18.

Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N., and Davidson, G.S. (2001) A gene expression map for *Caenorhabditis elegans*. *Science* **293**: 2087–92.

Kimura, M. (1983) *The Neutral Theory of Evolution.* Cambridge University Press, Cambridge.

Kirkwood, T.B.L. and Austad, S.N. (2000) Why do we age? *Nature* **408**: 233–8.

Kitano, H. (2002) Systems biology: a brief overview. *Science* **295**: 1662–4.

Kliebenstein, D. (2009) Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annual Review of Plant Biology* **60**: 93–114.

Knietsch, A., Waschkowitz, T., Bowien, S., Henne, A., and Daniel, R. (2003) Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on *Escherichia coli*. *Applied and Environmental Microbiology* **69**: 1408–16.

Kobayashi, M. and Yamamoto, M. (2005) Molecular mechanisms activating the Nrf2-Keap1 pathway of antioxidant gene regulation. *Antioxidants & Redox Signaling* **7**: 385–94.

Koeman, J.H. (1996) Toxicology, history and scope of the field. In *Toxicology. Principles and Applications*, R.J.M. Niesink, J. De Vries, and M.A. Hollinger (eds). CRC Press, Boca Raton, FL: 3–14.

Kole, C., Quijada, P., Michaels, S.D., Amasino, R.M., and Osborn, T.C. (2001) Evidence for homology of flowering-time genes *VFR2* from *Brassica rapa* and *FLC* from *Arabidopsis thaliana*. *Theoretical and Applied Genetics* **102**: 425–30.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. (2002) Selection in the evolution of gene duplications. *Genome Biology* **3**: research 0008.1–0008.9.

Kong, A.-N.T., Owuor, E., Yu, R., Hebbar, V., Chen, C., Hu, R., and Mandlekar, S. (2001) Induction of xenobiotic enzymes by the MAP kinase pathway and the antioxidant or electrophile response element (ARE/EpRE). *Drug Metabolism Reviews* **33**: 255–71.

Könneke, M., Bernhard, A.E., De la Torre, J.R., Walker, C.B., Waterbury, J.B., and Stahl, D.A. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**: 543–809.

Konstantinidis, K.T. and Tiedje, J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences USA* **102**: 2567–72.

Konstantinidis, K., Ramette, A., and Tiedje, J.M. (2006) The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society of London. B. Biological Sciences* **361**: 1929–40.

Kooijman, S.A.L.M. (2000) *Dynamic Energy and Mass Budgets in Biological Systems*. Cambridge University Press, Cambridge.

Koonin, E.V., Makarova, K.S., and Aravind, L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology* **55**: 709–42.

Koornneef, M. (2004) Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annual Review of Plant Biology* **55**: 141–72.

Koornneef, M., Alonso-Blanco, C., Peeters, A.J.M., and Soppe, W. (1998) Genetic control of flowering time in Arabidopsis. *Annual Review of Plant Physiology and Plant Molecular Biology* **49**: 345–70.

Korsloot, A., Van Gestel, C.A.M., and Van Straalen, N.M. (2004) *Environmental Stress and Cellular Response in Arthropods*. CRC Press, Boca Raton, FL.

Korth, K.L. (2003) Profiling the response of plants to herbivorous insects. *Genome Biology* **4**: 221.

Kowalchuk, G.A., Speksnijder, A.G.C.L., Zhang, K., Goodman, R.M., and Van Veen, J.A. (2007) Finding the needles in the metagenomic haystack. *Microbial Ecology* **53**: 475–85.

Kowalchuk, G.A. and Stephen, J.R. (2001) Ammonia-oxidizing bacteria: a model for molecular microbial ecology. *Annual Review of Microbiology* **55**: 485–529.

Kozlowski, J. (1993) Measuring fitness in life-history studies. *Trends in Ecology and Evolution* **8**: 84–5.

Krause, J., Unger, T., Nocon, A., Malaspinas, A.-S., Kolokotronis, S.-O., Stiller, M., Soibelzon, L., Spriggs, H., Dear, P.H., Briggs, A.W., *et al.* (2008) Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evolutionary Biology* **8**: 220.

Krebs, C.J. (1999) *Ecological Methodology*. Addison Wesley Longman, Menlo Park, CA.

Kreps, J.A., Wu, Y., Chang, H.-S., Zhu, T., Wang, X., and Harper, J.F. (2002) Transcriptome changes for *Arabidopsis* in response to salt, osmotic en cold stress. *Plant Physiology* **130**: 2129–41.

Kucharski, R., Maleszka, J., Foret, S., and Maleszka, R. (2008) Nutritional control of reproductive status in honeybees via DNA methylation. *Science* **319**: 1827–30.

Kulaev, I. and Kulakoskaya, T. (2000) Polyphosphate and phosphate pump. *Annual Review of Microbiology* **54**: 709–34.

Kültz, D. (2005) Molecular and evolutionary basis of the cellular stress response. *Annual Review of Physiology* **67**: 225–57.

Kwon, E.-S., Narasimhan, S.D., Yen, K., and Tissenbaum, H.A. (2010) A new DAF-16 isoform regulates longevity. *Nature* **466**: 498–502.

Landry, C.R., Townsend, J.P., Hartl, D.L., and Cavalieri, D. (2006a) Ecological and evolutionary genomics of *Saccharomyces cerevisiae*. *Molecular Ecology* **15**: 575–91.

Landry, C.R., Oh, J., Hartl, D.L., and Cavalieri, D. (2006b) Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* **366**: 343–51.

Lane, C.E. and Archibald, J.M. (2008) The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends in Ecology and Evolution* **23**: 268–75.

Lang, B.F., Burger, G., O'Kelly, C.J., Cedergren, R., Golding, G.B., Lemieux, C., Sankoff, D., Turmel, M., and Gray, M.W. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387**: 493–7.

Langefors, Å., Lohm, J., Grahn, M., Andersen, Ø., and Von Schantz, T. (2001) Association between major histocompatibility complex class IIB alleles and resistance to *Aeromonas salmonicida* in Atlantic salmon. *Proceedings of the Royal Society of London, Series B* **268**: 479–85.

Larcher, W. (2003) *Physiological Plant Ecology*, 4th edn. *Ecophysiology and Stress Physiology of Functional Groups*. Springer, Berlin.

Larkin, P., Folmar, L.C., Hemmer, M.J., Poston, A.J., Lee, H.S., and Denslow, N.D. (2002) Array technology as a tool to monitor exposure of fish to xenoestrogens. *Marine Environmental Research* **54**: 395–9.

Lauber, C.L., Hamady, M., Knight, R., and Fierer, N. (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology* **75**: 5111–20.

Laverman, A.M., Speksnijder, A.G.C.L., Braster, M., Kowalchuk, G.A., and Verhoef, H.A. (2001) Spatio-temporal stability of an ammonia-oxidizing community in a nitrogen-saturated forest soil. *Microbial Ecology* **42**: 35–45.

Lawton, J.H. (1994) What do species do in ecosystems? *Oikos* **71**: 367–74.

Lederberg, J. and McCray, A.T. (2001) 'Ome sweet' omics—A genealogical treasure of words. *The Scientist* **15**: 8.

Lee, K.A. and Klasing, K.C. (2004) A role of immunology in invasion biology. *Trends in Ecology and Evolution* **19**: 523–9.

Lee, S.S., Kennedy, S., Tolonen, A.C., and Ruvkun, G. (2003) DAF-16 target genes that control *C. elegans* lifespan and metabolism. *Science* **300**: 644–7.

Leemans, R., Egger, B., Loop, T., Kammermeier, L., He, H., Hartman, B., Certa, U., Hirth, F., and Reichert, H. (2000) Quantitative transcript imaging in normal and heat-shocked *Drosophila* embryos by using high-density oligonucleotide arrays. *Proceedings of the National Academy of Sciences USA* **97**: 12138–43.

Leininger, S., Urich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G.W., Prosser, J.I., Schuster, S.C., and Schleper, C. (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **442**: 806–9.

Lemos, B., Landry, C.R., Fontanillas, P., Renn, S.C.P., Kulathinal, R., Brown, K.L., and Hartl, D.L. (2008) Evolution of genomic expression. In *Evolutionary Genomics* and *Proteomics* (eds Pagel, M. and Pomiankowski, A.), pp. 81–118. Sinauer Associates, Inc. Publishers, Sunderland.

Leroi, A.M. (2001) Molecular signals versus the Loi de Balancement. *Trends in Ecology and Evolution* **16**: 24–9.

Leroi, A.M., Bartke, A., De Benedictis, G., Franceschi, C., Gartner, A., Gonos, E., Feder, M.E., Kisivild, T., Lee, S., Kartal-Özer, N. *et al*. (2005) What evidence is there for the existence of individual genes with antagonistic pleiotropic effects? *Mechanisms of Ageing and Development* **126**: 421–9.

Lesk, A.M. (2002) *Introduction to Bioinformatics*. Oxford University Press, Oxford.

Lessels, K. and Colegrave, N. (2001) Molecular signals or the *Loi de Balancement*? *Trends in Ecology and Evolution* **16**: 284–5.

Levine, S.N. (1989) Theoretical and methodological reasons for variability in the responses of aquatic ecosystem processes to chemical stress. In *Ecotoxicology: Problems and Approaches*, S.A. Levin, M.A. Harwell, J.R. Kelly, and K.D. Kimball (eds). Springer Verlag, New York: 145–79.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., *et al.* (2010) The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**: 311–17.

Li, X., Schuler, M.A., and Berenbaum, M.R. (2002) Jasmonate and salicylate induce expression of herbivore cytochrome P450 genes. *Nature* **419**: 712–15.

Li, Y., Álvarez, O.A., Gutteling, E.W., Tijsterman, M., Fu, J., Riksen, J.A.G., Hazendonk, E., Prins, P., Plasterk, R.H.A., Jansen, R.C., *et al.* (2006) Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genetics* **2**: e222.

Li, Y.F., Costello, J.C., Holloway, A.K., and Hahn, M.W. (2008) 'Reverse ecology' and the power of population genomics. *Evolution* **62**: 2984–94.

Liao, V.H.-C., Dong, J., and Freedman, J.H. (2002) Molecular characterization of a novel, cadmium-inducible gene from the nematode *Caenorabditis elegans*. *Journal of Biological Chemistry* **277**: 42049–59.

Liberles, D.A., Schreiber, D.R., Govindarajan, S., Chamberlin, S.G., and Benner, S.A. (2001) The Adaptive Evolution Database (TAED). *Genome Biology* **2**: research 0028.1–0028.6.

Liles, M.R., Manske, B.F., Bintrim, S.B., Handelsman, J., and Goodman, R.M. (2003) A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Applied and Environmental Microbiology* **69**: 2684–91.

Lin, K., Hsin, H., Libina, N., and Kenyon, C. (2001) Regulation of the Caenorhabditis elegans longevity protein DAF-16 by insulin/IGF-1 and germline signaling. *Nature Genetics* **28**: 139–45.

Liolios K, Chen IMA, Mavromatis K, *et al.* (2009) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* **38**: D346–54.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V., *et al.* (2009) Population genomics of domestic and wild yeasts. *Nature* **458**: 337–41.

Lobry, J.R. (1996) Asymmetric substitution pattern in the two DNA strands of bacteria. *Molecular Biology and Evolution* **13**: 660–5.

López-Maury, L., Marguerat, S., and Bähler, J. (2008) Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics* **9**: 583–93.

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J.P., Hector, A., Hooper, D.U., Huston, M.A., Raffaelli, D., Schmid, B. *et al.* (2001) Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* **294**: 804–8.

Lorenz, P. and Schleper, C. (2002) Metagenome—a challenging source of enzyme discovery. *Journal of Molecular Catalysis B: Enzymatic* **19–20**: 13–9.

Lotka, A.J. (1924) *Elements of Physical Biology*. Dover Publications, New York.

Lovett, R.A. (2000) Toxicologists brace for the genomics revolution. *Science* **289**: 536–7.

Lovley, D.R. (2003) Cleaning up with genomics: applying molecular biology to bioremediation. *Nature Reviews Microbiology* **1**: 35–44.

Lovley, D.R., Holmes, D.E., and Nevin, K.P. (2004) Dissimilatory Fe(III) and Mn(IV) reduction. *Advances in Microbial Physiology* **49**: 219–86.

Loy, A., Lehner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J., Schleifer, K.-H., and Wagner, M. (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all

recognized lineages of sulfate-reducing prokaryotes in the environment. *Applied and Environmental Microbiology* **68**: 5064–81.

Loy, A., Horn, M., and Wagner, M. (2003) probeBase: an online resource for rRNA-targeted oligonucleotide probes. *Nucleic Acids Research* **31**: 514–16.

Luikart, G., England, P.R., Tallmon, D., Jordan, S., and Taberlet, P. (2003) The power of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**: 981–94.

Lumppio, H.L., Shenvi, N.V., Summers, A.O., Voordouw, G., and Kurtz, Jr, D.M. (2001) Rubrerythrin and rubredoxin oxidoreductase in *Desulfovibrio vulgaris*: a novel oxidative stress protection system. *Journal of Bacteriology* **183**: 101–8.

Lund, J., Tedesco, P., Duke, K., Wang, J., Kim, S.K., and Johnson, T.E. (2002) Transcriptional profile of aging in *C. elegans*. *Current Biology* **12**: 1566–73.

Lynch, M. (2007a) *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA.

Lynch, M. (2007b) The frailty of adaptive hypothesis for the origins of organismal complexity. *Proceedings of the National Academy of Sciences USA* **104,** Suppl. 1: 8597–604.

Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–5.

Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* **302**: 1401–4.

Lynch, V.J. and Wagner, G.P. (2008) Resurrecting the role of transcription factor change in developmental evolution. *Evolution* **62**: 2131–54.

Ma, L., Li, J., Qu, L., Hager, J., Chen, Z., Zhao, H., and Deng, X.W. (2001) Light control of Arabidopsis development entails coordinated regulation of genome expression and cellular pathways. *The Plant Cell* **13**: 2589–607.

Maas, M.F.P.M., Van Mourik, A., Hoekstra, R.F., and Debets, A.J.M. (2005) Polymorphism for pKALILO based senescence in Hawaiian populations of *Neurospora intermedia* and *Neurospora tetrasperma*. *Fungal Genetics and Biology* **42**: 224–32.

MacDonald, C.C. and McMahon, K.W. (2003) The flowers that bloom in the spring: RNA processing and seasonal flowering. *Cell* **113**: 671–2.

Macel, M., Van Dam, N.M., and Keurentjes, J.J.B. (2010) Metabolomics: the chemistry between ecology and genetics. *Molecular Ecology Resources* **10**: 583–93.

Machado, H.E., Pollen, A.A., Hofman, H.A., and Renn, S.C.P. (2009) Interspecific profiling of gene expression informed by comparative genomic hybridization: A review and a novel approach in African cichlid fishes. *Integrative and Comparative Biology* **49**: 644–58.

Mackay, T.F.C., Stone, E.A., and Ayroles, J.F. (2009) The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**: 565–77.

Macknight, R., Duroux, M., Laurie, R., Dijkwel, P., Simpson, G., and Dean, C. (2002) Functional significance of the alternative transcript processing of the Arabidopsis floral promoter *FCA*. *The Plant Cell* **14**: 877–88.

Madigan, M.T., Martinko, J.M., and Parker, J. (2002) *Brock Biology of Microorganisms*. Prentice Hall, Pearson Education, Upper Saddle River, NJ.

Mair, W., Goymer, P., Pletcher, S.D., and Partridge, L. (2003) Demography of dietary restriction and death in *Drosophila*. *Science* **301**: 1731–3.

Mair, W., Sgrò, C.M., Johnson, A.P., Chapman, T., and Partridge, L. (2004) Lifespan extension by dietary restriction in female *Drosophila melanogaster* is not caused by a reduction in vitellogenesis or ovarian activity. *Experimental Gerontology* **39**: 1011–19.

Mäkinen, H.S., Cano, J.M., and Merilä, J. (2008) Identifying footprints of directional and balancing selection in marine and freshwater three-spined stickleback (*Gasterosteus aculeatus*) populations. *Molecular Ecology* **17**: 3565–82.

Manel, S., Conord, C., and Després, L. (2009) Genome scan to assess the respective role of host-plant environmental constraints on the adaptation of a widespread insect. *BMC Evolutionary Biology* **8**: 288.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–80.

Martens, C., Vandepoele, K., and Van de Peer, Y. (2008) Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 3427–32.

Martin, D.E., Demougin, P., Hall, M.N., and Bellis, M. (2004) Rank difference analysis of microarrays (RDAM), a novel approach to statistical analysis of microarray expression profiling data. *BMC Bioinformatics* **5**: 148.

Martin, F., Aerts, A., Ahren, D., Brun, A., Danchin, E.G.J., Duchaussoy, F., Gibon, J., Kohler, A., LIndquist, E., Pereda, V., *et al.* (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbosis. *Nature* **452**: 88–92.

Martin, F. and Nehls, U. (2009) Harnessing ectomycorrizal genomics for ecological insights. *Current Opinion in Plant Biology* **12**: 508–15.

Martin, F., Kohler, A., Murat, C., Balestrini, R., Coutinho, P.M., Jaillon, O., Montanini, B., Morin, E., Noel, B., Percudani, R.,

et al. (2010) Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature* **464**: 1033–8.

Martinez, D., Larrondo, L.F., Putnam, N., Sollewijn Gelpke, M.D., Huang, K., Chapman, J., Helfenbein, K.G., Ramaiya, P., Detter, J.C., Larimer, F. *et al*. (2004) Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nature Biotechnology* **22**: 695–700; Erratum, *Nature Biotechnology* **22**: 899.

Martinez, D., Berka, R.M., Henrissat, B., Saloheimo, M., Arvas, M., Baker, S.E., Chapman, J., Chertkov, O., Coutinho, P.M., Cullen, D., *et al*. (2008) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nature Biotechnology* **26**: 553–60.

Martyniuk, C.J., Kroll, K.J., Doperalski, N.J., Barber, D.S., and Denslow, N.D. (2010) Environmentally relevant exposure to 17 alpha-ethinylestradiol affects the telencephalic proteome of male fathead minnows. *Aquatic Toxicology* **98**: 344–53.

Matthiessen, P. (2000) Is endocrine disruption a significant ecological issue? *Ecotoxicology* **9**: 21–4.

Matthysse, A.G., Deschet, K., Williams, M., Marry, M., White, A.R., and Smith, W.C. (2004) A functional cellulose synthase from ascidian epidermis. *Proceedings of the National Academy of Sciences USA* **101**: 986–91.

Mayer, G.D., Leach, A., Kling, P., Olsson, P.-E., and Hogstrand, C. (2003) Activation of the rainbow trout metallothionein-A promoter by silver and zinc. *Comparative Biochemistry and Physiology Part B* **134**: 181–8.

McCaig, A.E., Glover, L.A., and Prosser, J.I. (1999) Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures. *Applied and Environmental Microbiology* **65**: 1721–30.

McCarroll, S.A., Murphy, C.T., Zou, S., Pletcher, S.D., Chin, C.-S., Jan, Y.N., Kenyon, C., Bargmann, C.I., and Li, H. (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature Genetics* **36**: 197–204.

McDonald, J.H. and Kreitman, M. (1991) Adaptive evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652–4.

McElwee, J., Bubb, K., and Thomas, J.H. (2003) Transcriptional outputs of the *Caenorhabditis elegans* forkhead protein DAF-16. *Aging Cell* **2**: 111–21.

McElwee, J.J., Schuster, E., Blanc, E., Thomas, J.H., and Gems, D. (2004) Shared transcriptional signature in *Caenorhabditis elegans* dauer larvae and long-lived *daf-2* mutants implicates detoxification system in longevity assurance. *Journal of Biological Chemistry* **279**: 44533–43.

McKusick, V.A. and Ruddle, F.H. (1987) A new discipline, a new name, a new journal. *Genomics* **1**: 1–2.

Menges, M., Hennig, L., Gruissem, W., and Murray, J.A.H. (2002) Cell cycle-regulated gene expression in *Arabidopsis*. *Journal of Biological Chemistry* **277**: 41987–2002.

Menges, M., Hennig, L., Gruissem, W., and Murray, J.A.H. (2003) Genome-wide gene expression in an *Arabidopsis* cell suspension. *Plant Molecular Biology* **53**: 423–42.

Metzker, M.L. (2010) Sequencing technologies – the next generation. *Nature Reviews Genetics* **11**: 31–46.

Menzel, R., Bogaert, T., and Achazi, R. (2001) A systematic gene expression screen of *Caenorhabditis elegans* cytochrome P450 gene reveals CYP35 as strongly xenobiotic inducible. *Archives of Biochemistry and Biophysics* **395**: 158–68.

Methé, B.A., Nelson, K.E., Eisen, J.A., Paulsen, I.T., Nelson, W., Heidelberg, J.F., Wu, D., Wu, M., Ward, N., Beanan, M.J. *et al*. (2003) Genome of *Geobacter sulfurreducens*: metal reduction in subsurface environments. *Science* **302**: 1967–9.

Miller, D.J. and Ball, E.E. (2009) The gene complement of the ancestran bilaterian – Urbilateria a monster? *Journal of Biology* **8**: 89.

Miller, J.R., Koren, S., and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics* **95**: 315–27.

Miller, T.R. and Belas, R. (2004) Dimethylsulfoniopropionate metabolism by *Pfiesteria*-associated *Roseobacter* spp. *Applied and Environmental Microbiology* **70**: 3383–91.

Mills, L. and Chichester, C. (2005) Review of evidence: are endocrine-disrupting chemicals in the aquatic environment impacting fish populations? *Science of the Total Environment* **343**: 1–34.

Mira, A., Ochman, H., and Moran, N.A. (2001) Deletion bias and the evolution of bacterial genomes. *Trends in Genetics* **17**: 589–96.

Miracle, A.L. and Ankley, G.T. (2005) Ecotoxicogenomics: linkages between exposure and effects in assessing risks of aquatic contaminants to fish. *Reproductive Toxicology* **19**: 321–6.

Mitchell-Olds, T. (2001) *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. *Trends in Ecology and Evolution* **16**: 693–700.

Mockler, T.C., and Ecker, J.R. (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85**: 1–15.

Moen, T., Hayes, B., Nilsen, F., Delghandi, M., Fjalestad, K.T., Fevolden, S.E., Berg, P.R., and Lien, S. (2008) Identification and characterization of novel SNP

markers in Atlantic cod: evidence for directional selection. *BMC Genetics* **9**: 18.

Moens, L.N., Van der Ven, K., Van Remortel, P., Del-Favero, J., and De Coen, W. (2006) Expression profiling of endocrine-disrupting compounds using a customized *Cyprinus carpio* microarray. *Toxicological Sciences* **93**: 298–310.

Momose, Y. and Iwahashi, H. (2001) Bioassay of cadmium using a DNA microarray: genome-wide expression patterns of *Saccharomyces cerevisiae* response to cadmium. *Environmental Toxicology and Chemistry* **20**: 2353–60.

Monteiro, A., Prijs, J., Bax, M., Kahhaart, T., and Brakefield, P.M. (2003) Mutants highlight the modular control of butterfly eyespot patterns. *Evolution & Development* **5**: 160–7.

Moran, P.J., Cheng, Y., Cassell, J.L., and Thompson, G.A. (2002) Gene expression profiling of *Arabidopsis thaliana* in compatible plant-aphid interactions. *Archives of Insect Biochemistry and Physiology* **51**: 182–203.

Moran, N.A., McCutcheon, J.P., and Nakabachi, A. (2008) Genomics and evolution of heritable bacterial symbionts. *Annual Review of Genetics* **42**: 165–90.

Morimoto, R.I., Tissières, A., and Georgopoulos, C. (eds) (1994) *The Biology of Heat Shock Proteins and Molecular Chaperones*. Cold Spring Harbor Press, New York.

Mouradov, A., Cremer, F., and Coupland, G. (2002) Control of flowering time: interacting pathways as a basis for diversity. *The Plant Cell* **14** (supplement) 2002: S111–30.

Mousseau, T.A. and Fox, C.H. (1998) The adaptive significance of maternal effects. *Trends in Ecology and Evolution* **13**: 403–7.

Moustafa, A., Beszteri, B., Maier, U.G., Bowler, C., Valentin, K., and Bhattacharya, D. (2009) Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**: 1724–6.

Müller, G.B. (2007) Evo-devo: extending the evolutionary synthesis. *Nature Reviews Genetics* **8**: 943–9.

Murakami, S. and Johnson, T.E. (2001) The OLD-1 positive regulator of longevity and stress resistance is under DAF-16 regulation in *Caenorhabditis elegans*. *Current Biology* **11**: 1517–23.

Murphy, C.T., McCarrol, S.A., Bargmann, C.I., Fraser, A., Kamath, R.S., Ahringer, J., Li, H., and Kenyon, C. (2003) Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature* **424**: 277–84.

Murray, A.W. (2000) Whither genomics? *Genome Biology* **1**: comment 003.1–003.6.

Murray, M.C. and Hare, M.P. (2006) A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, *Crassostrea virginica*. *Molecular Ecology* **15**: 4229–42.

Muyzer, G., De Waal, E.C., and Uitterlinden, A.G. (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology* **59**: 695–700.

Muyzer, G. and Stams, A.J.M. (2008) The ecology and biotechnology of sulphate-reducing bacteria. *Nature Reviews Microbiology* **6**: 441–54.

Myers, E.W., Sutton, G.G., Delcher, A.L., *et al.* (2000) A whole-genome assembly of *Drosophila*. *Science* **287**: 2196–204.

Naeem, S., Loreau, M., and Inchausti, P. (2002) Biodiversity and ecosystem functioning: the emergence of a synthetic ecological framework. In *Biodiversity and Ecosystem Functioning*, M. Loreau, S. Naeem, and P. Inchausti (eds). Oxford University Press, Oxford: 3–11.

Nakamura, Y., Gojobori, T., and Ikemura, T. (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research* **28**: 292.

Namroud, M.-C., Beaulieu, J., Juge, N., Laroche, J., and Bousquet, J. (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology* **17**: 3599–613.

Nardi, F., Spinsanti, G., Boore, J.L., Carapelli, A., Dallai, R., and Frati, F. (2003) Hexapod origins: monophyletic or paraphyletic? *Science* **299**: 1887–9.

Nasonia Genome Working Group (2010) Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* **327**: 343–8.

Nebert, D.W., Roe, A.L., Dieter, M.Z., Solis, W.A., Yang, Y., and Dalton, T.P. (2000) Role of the aromatic hydrocarbon receptor and [Ah] gene battery in the oxidative stress response, cell cycle control, and apoptosis. *Biochemical Pharmacology* **59**: 65–85.

Neefs, J.-M., Van der Peer, Y., De Rijk, P., Chapelle, S., and De Wachter, R. (1993) Compilation of small ribosomal subunit RNA structures. *Nucleic Acids Research* **21**: 3025–49.

Nei, M. (2007) The new mutation theory of phenotypic evolution. *Proceedings of the National Academy of Sciences USA* **104**: 12235–42.

Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**: 418–26.

Nei, M., Gu, X., and Sitnikova, T. (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences USA* **94**: 7799–806.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A. *et al*. (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**: 323–9.

Nesbø, C.L., L'Haridon, S., Stetter, K.O., and Doolittle, W.F. (2001) Phylogenetic analyses of two 'archaeal' genes in *Thermotoga maritima* reveal multiple transfers between Archaea and Bacteria. *Molecular Biology and Evolution* **18**: 362–75.

Neutel, A.-M., Heesterbeek, J.A.P., and De Ruiter, P.C. (2002) Stability in real food webs: weak links in long loops. *Science* **296**: 1120–3.

Nguyen, T., Sherrat, P.J., and Pickett, C.B. (2003) Regulatory mechanisms controlling gene expression mediated by the antioxidant response element. *Annual Review of Pharmacology and Toxicology* **43**: 233–60.

Nguyen, T., Yang, C.S., and Pickett, C.B. (2004) The pathways and molecular mechanisms regulating Nrf2 activation in response to chemical stress. *Free Radical Biology & Medicine* **37**: 433–41.

Nielsen, C. (1995) *Animal Evolution. Interrelationships of the Living Phyla*. Oxford University Press, Oxford.

Nielsen, R. (2005) Molecular signatures of natural selection. *Annual Review of Genetics* **39**: 197–218.

Nijhout, H.F. (2003a) Development and evolution of adaptive polyphenisms. *Evolution & Development* **5**: 9–18.

Nijhout, H.F. (2003b) The control of body size in insects. *Developmental Biology* **261**: 1–9.

Nikoh, N. and Nakabachi, A. (2009) Aphids acquired symbiotic genes via lateral gene transfer. *BMC Biology* **7**: 12.

Nishida, R. (2002) Sequestration of defensive substances from plants by Lepidoptera. *Annual Review of Entomology* **47**: 57–92.

Noël, H.L., Hopkin, S.P., Hutchinson, T.H., Williams, T.D., and Sibly, R.M. (2006) Towards a population ecology of stressed environments: the effects of zinc on the springtail *Folsomia candida*. *Journal of Applied Ecology* **43**: 325–32.

Nolte, V., Pandey, R.V., Jost, S., Medinger, R., Ottenwälder, B., Boenigk, J., and Schlötterer, C. (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Molecular Ecology* **19**: 2908–15.

Nordgren, A., Bååth, E., and Söderström, B. (1983) Microfungi and microbial activity along a heavy metal gradient. *Applied and Environmental Microbiology* **45**: 1829–37.

North, N.N., Dollhopf, S.L., Petrie, L., Istok, J.D., Balkwill, D.L., and Kostka, J.E. (2004) Change in bacterial community structure during *in situ* biostimulation of subsurface sediment cocontaminated with uranium and nitrate. *Applied and Environmental Microbiology* **70**: 4911–20.

Nota, B., Verweij, R.A., Molenaar, D., Ylstra, B., Van Straalen, N.M., and Roelofs, D. (2010) Gene expression analysis reveals a gene set discriminatory to different metals in soil. *Toxicological Sciences* **115**: 34–40.

Novina, C.D. and Sharp, P.A. (2004) The RNAi revolution. *Nature* **430**: 161–4.

Nunes, F.M.F., Valente, V., Sousa, J.F., Cunha, M.A.V., Pinheiro, D.G., Maia, R.M., Araujo, D.D., Costa, M.C.R., Martins, W.K., Carvalho, A.F. *et al*. (2004) The use of open reading frame ESTs (ORESTES) for analysis of the honey bee transcriptome. *BMC Genomics* **5**: 84.

Nuwaysir, E.F., Huang, W., Albert, T.J., Singh, J., Nuwaysir, K., Pitas, A., Richmond, T., Gorksi, T., Berg, J.P., Ballin, J. *et al*. (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Research* **12**: 1749–55.

Odum, E.P. (1953) *Fundamentals of Ecology*. W.B. Saunders Co., Philadelphia.

Oetjen, K. and Reusch, T.B.H. (2007) Genome scans detect consistent divergent selection among subtidal vs. intertidal populations of the marine angiosperm *Zostera marina*. *Molecular Ecology* **16**: 5156–67.

Ohm, R.A., De Jong, J.F., Lugones, L.G., Aerts, A., Kothe, J.E., De Vries, R.P., Record, E., Levasseur, A., Baker, S.E., Bartholomew, K.A., *et al.* (2010) Genome sequence of the model mushroom *Schizophyllum commune*. *Nature Biotechnology* **28**: 957–63.

Ohno, S. (1972) So much 'junk' DNA in our genome. In *Evolution of Genetic Systems*, H.H. Smith, H.J. Price, A.H. Sparrow, F.W. Studier, and J.D. Yourno (eds). Gordon and Breach, New York: 366–70.

Ohta, T. (1992) The nearly neutral theory of molecular evolution. *Annual Review of Ecology* and *Systematics* **23**: 263–86.

Oldham, S. and Hafen, E. (2003) Insulin/IGF and target of rapamycin signaling: a TOR de force in growth control. *Trends in Cell Biology* **13**: 79–85.

Oleksiak, M., Churchill, G.A., and Crawford, D.L. (2002) Variation in gene expression within and among natural populations. *Nature Genetics* **32**: 261–6.

Oleksiak, M., Roach, J.L., and Crawford, D.L. (2004) Natural variation in cardiac metabolism and gene expression in *Fundulus heteroclitus*. *Nature Genetics* **37**: 67–72.

Oliver, J.L., Bernaola-Galván, P., Carpena, P., and Román-Roldán, R. (2001) Isochore maps of eukaryotic genomes. *Gene* **276**: 47–56.

Olsen, A., Sampayo, J.N., and Lithgow, G.L. (2003) Aging in *C. elegans*. In *Aging of Organisms*, H.D. Osiewacz (ed.). Kluwer Academic Publishers, Dordrecht: 163–99.

Osta, M.A., Christophides, G.K., and Kafatos, F.C. (2004) Effect of mosquito genes on *Plasmodium* development. *Science* **303**: 2030–2.

Øvreås, L. (2000) Population and community level approaches for analysing microbial diversity in natural environments. *Ecology Letters* **3**: 236–51.

Owen, J., Hedley, B.A., Svendsen, C., Wren, J.F., Jonker, M.J., Hankard, P.K., Lister, L.J., Stürzenbaum, S.R., Morgan, A.J., Spurgeon, D.J., *et al.* (2008) Transcriptome profiling of developmental and xenobiotic responses in a keystone animal, the oligochaete annelid *Lumbricus rubellus*. *BMC Genomics* **9**: 266.

Ozturk, Z.N., Talamé, V., Deyolos, M., Michalowski, C.B., Galbraith, D.W., Gozokirmizi, N., Tuberosa, R., and Bohnert, H.J. (2002) Monitoring large-scale changes in transcript abundance in drought- and salt-stressed barley. *Plant Molecular Biology* **48**: 551–73.

Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science* **276**: 734–40.

Pain, A., Crossman, L., and Parkhill, J. (2005) Comparative apicomplexan genomics. *Nature Reviews Microbiology* **3**: 454–5.

Palmiter, R.D. (1994) Regulation of metallothionein genes by heavy metals appears to be mediated by a zinc-sensitive inhibitor that interacts with constitutively active transcription factor, MTF-1. *Proceedings of the National Academy of Sciences USA* **91**: 1219–23.

Palsson, B. (2000) The challenges of in silico biology. *Nature Biotechnology* **18**: 1147–50.

Paris, M., S., B., Bonin, A., Collado, A., David, J.-P., and Despres, L. (2010) Genome scan in the mosquito *Aedes rusticus*: population structure and detection of positive selection after insecticide treatment. *Molecular Ecology* **19**: 325–37.

Parkinson, J., Mitreva, M., Whitton, C., Thomson, M., Daub, J., Martin, J., Schmid, R., Hall, N., Barrell, B., Waterston, R.H. *et al*. (2004) A transcriptomic analysis of the phylum Nematoda. *Nature Genetics* **36**: 1259–67.

Parniske, M. (2008) Arbuscular mycorrhiza: the mother of plant root endosymbioses. *Nature Reviews Microbiology* **6**: 763–75.

Partridge, L. (2001) Evolutionary theories of ageing applied to long-lived organisms. *Experimental Gerontology* **36**: 641–50.

Partridge, L. and Gems, D. (2002a) The evolution of longevity. *Current Biology* **12**: R544–6.

Partridge, L. and Gems, D. (2002b) Mechanisms of ageing: public or private? *Nature Reviews Genetics* **3**: 165–75.

Partridge, L. and Pletcher, S.D. (2003) Genetics of aging in *Drosophila*. In: *Aging of Organisms* H.D. Osiewacz (ed.). Kluwer Academic Publishers, Dordrecht: 125–61.

Passarge, E., Horsthemke, B., and Farber, R.A. (1999) Incorrect use of the term synteny. *Nature Genetics* **23**: 387.

Pastinen, T. (2010) Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics* **11**: 533–8.

Patel, N.H. (2004) Time, space and genomes. *Nature* **431**: 28–9.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., *et al.* (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551–6.

Paul, J.H., Sullivan, M.B., Segall, A.M., and Rohwer, F. (2002) Marine phage genomics. *Comparative Biochemistry and Physiology Part B* **133**: 463–76.

Peichel, C.L., Nereng, K.S., Ohgi, K.A., Cole, B.L.E., Colosimo, P.F., Buerkle, C.A., Schluter, D., and Kingsley, D.M. (2001) The genetic architecture of divergence between threespine stickleback species. *Nature* **414**: 901–5.

Pennie, W.D., Tugwood, J.D., Oliver, G.J.A., and Kimber, I. (2000) The principles and practice of toxicogenomics: applications and opportunities. *Toxicological Sciences* **54**: 277–83.

Peplies, J., Glöckner, F.O., and Amann, R. (2003) Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. *Applied and Environmental Microbiology* **69**: 1397–1407.

Perez-Brocal V., Gil, R., Ramos, S., *et al.* (2006) A small microbial genome: The end of a long symbiotic relationship? *Science* **314**, 312–13.

Petersen, K., Didion, T., Anderson, C.H., and Nielsen, K.K. (2004) MADS-box genes from perennial ryegrass differentially expressed during transition from vegetative to reproductive growth. *Journal of Plant Physiology* **161**: 439–47.

Pichersky, E. and Gang, D.R. (2000) Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends in Plant Science* **5**: 439–45.

Piel, J., Hui, D., Fusetani, N., and Matsunaga, S. (2004) Targeting modular polyketide synthases with iteratively acting acyltransferases from metagenomes of uncultured bacterial consortia. *Environmental Microbiology* **6**: 921–7.

Pierik, R., Cuppens, M.L.C., Voesenek, L.A.C.J., and Visser, E.J.W. (2004) Interactions between ethylene and

gibberellins in phytochrome-mediated shade avoidance responses in tobacco. *Plant Physiology* **136**: 2928–36.

Pigliucci, M. (1996) How organisms respond to environmental changes: from phenotypes to molecules (and vice versa). *Trends in Ecology and Evolution* **11**: 168–73.

Pigliucci, M. and Kaplan, J. (2000) The fall and rise of Dr Pangloss: adaptionism and the *Spandrels* paper 20 years later. *Trends in Ecology* and *Evolution* **15**: 66–70.

Piñeiro, M., Gómez-Mena, C., Schaffer, R., Martínez-Zapater, J.M., and Coupland, G. (2003) EARLY BOLTING IN SHORT DAYS is related to chromatin remodeling factors and regulates flowering in Arabidopsis by repressing *FT*. *The Plant Cell* **15**: 1552–62.

Pletcher, S.D., Macdonald, S.J., Marguerie, R., Certa, U., Stearns, S.C., Goldstein, D.B., and Partridge, L. (2002) Genome-wide transcript profiles in aging and calorically restricted *Drosophila melanogaster*. *Current Biology* **12**: 712–23.

Pokarzhevskii, A.D., Van Straalen, N.M., Zaboev, D.P., and Zaitsev, A.S. (2003) Microbial links and element flows in nested detrital food-webs. *Pedobiologia* **47**: 213–24.

Polz, M.F., Bertilsson, S., Acinas, S.G., and Hunt, D (2003) A(r)ray of hope in analysis of the function and diversity of microbial communities. *Biological Bulletin* **204**: 196–9.

Poncet, B.N., Herrmann, D., Gugerli, F., Taberlet, P., Holderegger, R., Gielly, L., Rioux, D., Thuiller, W., Aubert, S., and Manel, S. (2010) Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Molecular Ecology* **19**: 2896–907.

Poynton, H.C., Varshavsky, J.R., Chang, B., Cavigiolio, G., Chan, S., Holman, P.S., Loguinov, A.V., Bauer, D.J., Komachi, K., Theil, E.C., *et al.* (2007) *Daphnia magna* ecotoxicogenomics provides mechanistic insights into metal toxicity. *Environmental Science and Technology* **41**: 1044–50.

Price, P.W., Bouton, C.E., Gross, P., McPheron, B.A., Thompson, J.N., and Weis, A.E. (1980) Interactions among three trophic levels: Influence of plants on interactions between insect herbivores and natural enemies. *Annual Review of Ecology and Systematics* **11**: 41–65.

Prince, V.E. and Pickett, F.B. (2002) Splitting pairs: the diverging fates of duplicated genes. *Nature Reviews Genetics* **3**: 827–37.

Pritchard, J.K. and Di Rienzo, A. (2010) Adaptation – not by sweeps alone. *Nature Reviews Genetics* **11**: 665–7.

Procaccini, G., Pischetola, M., and Di Lauro, R. (2000) Isolation and characterization of microsatellite loci in the ascidian *Ciona intestinalis* (L.). *Molecular Ecology* **9**: 1924–6.

Pühler, A. and Selbitschka, W. (2003) Genome research on bacteria relevant for agriculture, environment and biotechnology. *Journal of Biotechnology* **106**: 119–20.

Purohit, H.J., Raje, D.V., Kapley, A., Padmanabhan, P., and Singh, R.N. (2003) Genomic tools in environmental impact assessment. *Environmental Science and Technology* **37**: 337A–68A.

Putnam, N.H., Srivastata, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A.A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V., *et al.* (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organisation. *Science* **317**: 86–94.

Putnam, N.H., Butts, T., Ferrier, D.E.K., Furlong, R.F., Hellsten, U., Kawashima, T., Rubinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.-K., *et al.* (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**: 1064–71.

Putterrill, J., Laurie, R., and Macknight, R. (2004) It's time to flower: the genetic control of flowering time. *BioEssays* **26**: 363–73.

Quail, P.H. (2002) Phytochrome photosensory signalling networks. *Nature Reviews Molecular Cell Biology* **3**: 85–93.

Quaiser, A., Ochsenreiter, T., Klenk, H.-P., Kleftzin, A., Treusch, A.H., Meurer, G., Eck, J., Sensen, C.W., and Schleper, C. (2002) First insight into the genome of an uncultivated crenarchaeote from soil. *Environmental Microbiology* **4**: 603–11.

Quaiser, A., Ochsenreiter, T., Lanz, C., Schuster, S.C., Treusch, A.H., Eck, J., and Schleper, C. (2003) Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. *Molecular Microbiology* **50**: 563–75.

Quesada, V., Macknight, R., Dean, C., and Simpson, G.G. (2003) Autoregulation of *FCA* pre-mRNA processing controls *Arabidopsis* flowering time. *EMBO Journal* **22**: 3142–52.

Quince, C., Curtis, T.P., and Sloan, W.T. (2008) The rational exploration of microbial diversity. *The ISME Journal* **2**: 997–1006.

Radajewski, S., McDonald, I.R., and Murrell, J.C. (2003) Stable-isotope probing of nucleic acids: a window to the function of uncultured microorganisms. *Current Opinion in Biotechnology* **14**: 296–302.

Raible, F., Tessmar-Raible, K., Osoegawa, K., Wincker, P., Jubin, C., Balavoine, G., Ferrier, D., Benes, V., De Jong, P., Weissenbach, J., *et al.* (2005) Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science* **310**: 1325–6.

Ram, R.J., VerBerkmoes, N.C., Thelen, M.P., Tyson, G.W., Baker, B.J., Blake, II, R.C., Shah, M., Hettich, R.L., and

Banfield, J.F. (2005) Community proteomics of a natural microbial biofilm. *Science* **308**: 1915–20.

Ranson, H., Claudianos, C., Ortelli, F., Abgrall, C., Hemingway, J., Sharakhova, M.V., Unger, M.F., Collins, F.H., and Feyereisen, R. (2002) Evolution of supergene families associated with insecticide resistance. *Science* **298**: 179–81.

Ranz, J.M., Castillo-Davis, C.I., Meiklejohn, C.D., and Hartl, D.L. (2003) Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. *Science* **300**: 1742–5.

Ranz, J.M. and Machado, C.A. (2006) Uncovering evolutionary patterns of gene expression using microarrays. *Trends in Ecology* and *Evolution* **21**: 29–37.

Rapp, R.A. and Wendel, J.F. (2005) Epigenetics and evolution. *New Phytologist* **168**: 81–91.

Rast, J.P., Courtney Smith, L., Loza-Coll, M., Hibino, T., and Litman, G.W. (2006) Genomic insights into the immune system of the sea urchin. *Science* **314**: 952–6.

Ratcliffe, O.J. and Riechmann, J.L. (2002) Arabidopsis transcription factors and the regulation of flowering time: a genomic perspective. *Current Issues in Molecular Biology* **4**: 77–91.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabási, A.-L. (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297**: 1551–5.

Raven, J.A. and Allen, J.F. (2003) Genomics and chloroplast evolution: what did cyanobacteria do for plants? *Genome Biology* **4**: 209.

Regier, J.C., Shultz, J.W., Zwick, A., Ball, B., Wetzer, R., Martin, J.W., and Cunningham, C.W. (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-encoding sequences. *Nature* **463**: 1079–83.

Reinke, V. and White, K.P. (2002) Developmental genomic approaches in model organisms. *Annual Review of Genomics and Human Genetics* **3**: 153–78.

Rendulic, S., Jagtap, P., Rosinus, A., Eppinger, M., Baar, C., Lanz, C., Keller, H., Lambert, C., Evans, K.J., Goesmann, A. *et al*. (2004) A predator unmasked: life cycle of *Bdellovibrio bacteriovorus* from a genomic perspective. *Science* **303**: 689–92.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E.A., Kamisugi, Y., *et al.* (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**: 64–9.

Replansky, T., Koufopanou, V., Greig, D., and Bell, G. (2008) *Saccharomyces sensu stricto* as a model system for evolution and ecology. *Trends in Ecology and Evolution* **23**: 494–501.

Reymond, P., Weber, H., Diamond, M., and Farmer, E.E. (2000) Differential gene expression in response to mechanical wounding and insect feeding in *Arabidopsis*. *The Plant Cell* **12**: 707–19.

Reysenbach, A.-L. and Shock, E. (2002) Merging genomes with geochemistry in hydrothermal ecosystems. *Science* **296**: 1077–82.

Rhee, S.Y., Wood, V., Dolinksi, K., and Draghici, S. (2008) Use and misuse of the gene ontology annotations. *Nature Reviews Genetics* **9**: 509–15.

Richards, C.L., Bossdorf, O., and Verhoeven, K.J.F. (2010) Understanding natural epigenetic variation. *New Phytologist* **187**: 562–4.

Riddle, D.L. (1988) The dauer larva. In *The Nematode Caenorhabditis elegans*, W.B. Wood and the Community of *C. elegans* Researchers (eds). Cold Spring Harbor Press, Cold Spring Harbor: 393–412.

Riesenfeld, C.S., Schloss, P.D., and Handelsman, J. (2004a) Metagenomics: genome analysis of microbial communities. *Annual Review of Genetics* **38**: 525–52.

Riesenfeld, C.S., Goodman, R.M., and Handelsman, J. (2004b) Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental Microbiology* **6**: 981–9.

Rizhsky, L., Liang, H., Shuman, J., Shulaev, V., Davletova, S., and Mittler, R. (2004) When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress. *Plant Physiology* **134**: 1683–96.

Robbins, A.H., McRee, D.E., Williamson, M., Collet, S.A., Xuong, N.H., Furey, W.F., Wang, B.C., and Stout, C.D. (1991) Refined crystal structure of Cd, Zn metallothionein at 2.0 Å resolution. *Journal of Molecular Biology* **221**: 1269–93.

Rockman, M.V. and Kruglyak, L. (2006) Genetics of global gene expression. *Nature Reviews Genetics* **7**: 862–72.

Roda, A., Halitschke, R., Steppuhn, A., and Baldwin, I.T. (2004) Individual variability in herbivore-specific elicitors from the plant's perspective. *Molecular Ecology* **13**: 2421–33.

Rodrigues-Pousada, C.A., Nevitt, T., Menezes, R., Azevedo, D., Pereira, J., and Amaral, C. (2004) Yeast activator proteins and stress response: an overview. *FEBS Letters* **567**: 80–5.

Roelofs, D., Morgan, J., and Stürzenbaum, S. (2010) The significance of genome-wide transcriptional regulation in the evolution of stress tolerance. *Evolutionary Ecology* **24**: 527–39.

Roesijadi, G. (1996) Metallothionein and its role in toxic metal regulation. *Comparative Biochemistry and Physiology* **113C**: 117–23.

Roff, D.A. (2002) *Life History Evolution*. Sinauer Associates, Sunderland, MA.

Roff, D.A. (2007) Contributions of genomics to life-history theory. *Nature Reviews Genetics* **8**: 116–25.

Rogers, S.M. and Bernatchez, L. (2005) Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). *Molecular Ecology* **14**: 351–61.

Rogina, B., Helfand, S.L., and Frankel, S. (2002) Longevity regulation by *Drosophila* Rpd3 deacetylase and caloric restriction. *Science* **298**: 1745.

Röling, W.F.M., Van Breukelen, B., Braster, M., Lin, B., and Van Verseveld, H.W. (2001) Relationships between microbial community structure and hydrochemistry in a landfill leachate-polluted aquifer. *Applied and Environmental Microbiology* **67**: 4619–29.

Röling, W.F.M., Van Breukelen, B.M., Bruggeman, F.J., and Westerhoff, H.V. (2007) Ecological control analysis: being(s) in control of mass flux and metabolite concentrations in anaerobic degradation processes. *Environmental Microbiology* **9**: 500–11.

Röling, W.F.M., Milner, M.G., Jones, D.M., Fratepietro, F., Swannell, R.P.J., Daniel, F., and Head, I.M. (2004) Bacterial community dynamics and hydrocarbon degradation during a field-scale evaluation of bioremediation on a mudflat beach contaminated with buried oil. *Applied and Environmental Microbiology* **70**: 2603–13.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M., and Nyren, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* **242**: 84–9.

Rondon, M.R., Goodman, R.M., and Handelsman, J. (1999) The Earth's bounty: assessing and accessing soil microbial diversity. *Trends in Biotechnology* **17**: 403–9.

Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R., Loiacono, K.A., Lynch, B.A., MacNeil, I.A., Minor, C. *et al*. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and Environmental Microbiology* **66**: 2541–7.

Rosenberg, S.M. and Hastings, P.J. (2004) Worming into genetic instability. *Nature* **430**: 625–6.

Roskam, J.C. and Brakefield, P.M. (1999) Seasonal polyphenism in *Bicyclus* (Lepidoptera: Satyridae) butterflies: different climates need different cues. *Biological Journal of the Linnean Society* **66**: 345–56.

Rubin, G.M. and Lewis, E.B. (2000) A brief history of Drosophila's contributions to genome research. *Science* **287**: 2216–18.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W. *et al*. (2000) Comparative genomics of the eukaryotes. *Science* **287**: 2204–15.

Rumpho, M.E., Worful, J.M., Lee, J., Kannan, K., Tyler, M.S., Bhattacharya, D., Moustafa, A., and Manhart, J.R. (2008) Horizontal gene transfer of the algal nuclear gene *psbO* to the photosynthetic sea slug *Elysia chlorotica*. *Proceedings of the National Academy of Sciences USA* **105**: 17867–71.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffmann, J.M., Remington, K., *et al*. (2007) The *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through Eastern tropical Pacific. *PLoS Biology* **5**: e77.

Russell, P.J. (2002) *iGenetics*. Pearson Education/Benjamin Cummings, San Fransisco.

Saccone, C. and Pesole, G. (2003) *Handbook of Comparative Genomics*. John Wiley & Sons, Hoboken, NJ.

Saint André, A.V., Blackwell, N.M., Hall, L.R., Hoerauf, A., Brattig, N.W., Volkmann, L., Taylor, M.J., Ford, L., Hise, A.G., Lass, J.H. *et al*. (2002) The role of endosymbiotic Wolbachia bacteria in the pathogenesis of river blindness. *Science* **295**: 1892–5.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison III, C.A., Slocombe, P.M., and Smith, M. (1977a) Nucleotide sequence of bacteriophage ΦX174 DNA. *Nature* **265**: 687–95.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977b) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA* **74**: 5463–7.

Sansone, S.A., Morrisson, N., Rocca-Serra, P., and Fostel, J.M. (2004) Standardization initiatives in the (eco)toxicogenomics domain: a review. *Comparative and Functional Genomics* **5**: 633–41.

Santoro, M.G. (2000) Heat shock factors and the control of the stress response. *Biochemical Pharmacology* **59**: 55–63.

Schacherer, J., Shapiro, J.A., Ruderfer, D.M., and Kruglyak, L. (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458**: 342–5.

Schaffer, R., Landgraf, J., Accerbi, M., Simon, V., Larson, M., and Wisman, E. (2001) Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis. *The Plant Cell* **13**: 113–23.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467–70.

Shendure, J., and Ji, H. (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135–45.

Schepens, I., Duek, P., and Fankhauser, C. (2004) Phytochrome-mediated light signalling in Arabidopsis. *Current Opinion in Plant Biology* **7**: 564–9.

Schindler, D.W. (1987) Detecting ecosystem responses to anthropogenic stress. *Canadian Journal of Fisheries and Aquatic Sciences* **44**: 6–25.

Schink, B. and Friedrich, M. (2000) Phosphite oxidation by sulphate reduction. *Nature* **406**: 37.

Schittko, U., Hermsmeier, D., and Baldwin, I.T. (2001) Molecular interactions between the specialist herbivore *Manduca sexta* (Lepidoptera, Sphingidae) and its natural host *Nicotiana attenuata*. II. Accumulation of plant mRNAs in response to insect-derived cues. *Plant Physiology* **125**: 701–10.

Schleper, C., Jurgens, G., and Jonuscheit, M. (2005) Genomic studies of uncultivated Archaea. *Nature Reviews Microbiology* **3**: 479–88.

Schlichting, C.D. and Smith, H. (2002) Phenotypic plasticity: linking molecular mechanisms with evolutionary outcomes. *Evolutionary Ecology* **16**: 189–211.

Schloss, P.D. and Handelsman, J. (2003) Biotechnological prospects from metagenomics. *Current Opinion in Biotechnology* **14**: 303–10.

Schmid, M., Uhlenhaut, N.H., Godard, F., Demar, M., Bressan, R., Weigel, D., and Lohmann, J.U. (2003) Dissection of floral induction pathways using global expression analysis. *Development* **130**: 6001–12.

Schmidt, J.M., Good, R.T., Appleton, B., Sherrard, J., Raymant, G.C., Bogwitz, M.R., Martin, J.H., Daborn, P.J., Goddard, M.E., Batterham, P., *et al.* (2010) Copy number variation and transposable elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genetics* **6**: e10000998.

Schmidt-Nielsen, K. (1997) *Animal Physiology. Adaptation and Environment*, 5th edn. Cambridge University Press, Cambridge.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J.-L., Mitros, T., Nelson, W.C., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature* **463**: 178–83.

Scholl, E.H., Thorne, J.L., McCarter, J.P., and Bird, D.M. (2003) Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. *Genome Biology* **4**: R39.

Schulte, P.M. (2004) Changes in gene expression as biochemical adaptations to environmental change: a tribute to Peter Hochachka. *Comparative Biochemistry and Physiology Part B* **139**: 519–29.

Schulte, P.M., Glémet, H.C., Fiebig, A.A., and Powers, D.A. (2000) Adaptive variation in lactate dehydrogenase-B gene expression: Role of a stress-responsive regulatory element. *Proceedings of the National Academy of Sciences USA* **97**: 6597–602.

Schulze, A. and Downward, J. (2001) Navigating gene expression using microarrays—a technology review. *Nature Cell Biology* **3**: E190–5.

Schulze, E.-D. and Mooney, H.A. (eds) (1993) *Biodiversity and Ecosystem Function*. Springer-Verlag, Berlin.

Schweitzer, J.A., Balley, J.K., Rehill, B.J., Martinsen, G.D., Hart, S.C., Lindroth, R.L., Keim, P., and Whitham, T.G. (2004) Genetically based trait in a dominant tree affects ecosystem processes. *Ecology Letters* **7**: 127–34.

Sea Urchin Genome Sequencing Consortium (2006) The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**: 941–52.

Sebat, J.L., Colwell, F.S., and Crawford, R.L. (2003) Metagenomic profiling: microarray analysis of an environmental genomic library. *Applied and Environmental Microbiology* **69**: 4927–34.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* **34**: 166–76.

Seki, M., Narusaka, M., Abe, H., Kasuga, M., Yamaguchi-Shinozaki, K., Carninci, P., Hayashizaki, Y., and Shonozaki, K. (2001) Monitoring expression pattern of 1300 *Arabidopsis* genes under drought and cold stresses by using a full-length cDNA microarray. *The Plant Cell* **13**: 61–72.

Seki, M., Narusaka, M., Ishida, J., Nanjo, T., Fujita, M., Ouno, Y., Kamiya, A., Nakajima, M., Enju, A., Sakurai, T. *et al.* (2002) Monitoring the expression profiles of 7000 *Arabidopsis* genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *The Plant Journal* **31**: 279–92.

Sessitsch, A., Haeckl, E., Wenzl, P., Kilian, A., Kostic, T., Stralis-Pavese, N., Sandjong, B.T., and Bodrossy, L. (2006) Diagnostic microbial microarrays in soil ecology. *New Phytologist* **171**: 719–36.

Shapiro, J.A., Huang, W., Zhang, C., Hubisz, M.J., Lu, J., Turissini, D.A., Fang, S., Wang, H.-Y., Hudson, R.R., Nielsen, R., *et al.* (2007) Adaptive genic evolution in the *Drosophila* genomes. *Proceedings of the National Academy of Sciences USA* **104**: 2271–6.

Shapiro, M.D., Marks, M.E., Peichel, C.L., Blackman, B.K., Nereng, K.S., Jónsson, B., Schluter, D., and Kingsley, D.M. (2004) Genetic and developmental basis of evolutionary pelvis reduction in threespine sticklebacks. *Nature* **428**: 717–23.

Sharbel, T.F., Huabold, B., and Mitchell-Olds, T. (2000) Genetic isolation by distance in *Arabidopsis* biogeogra-

phy and postglacial colonization of Europe. *Molecular Ecology* **9**: 2109–18.

Shiu, S.H., and Borevitz, J.O. (2008) The next generation of microarray research: Applications in evolutionary and ecological genomics. *Heredity* **100**: 141–9.

Short, S.M. and Suttle, C.A. (2002) Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. *Applied and Environmental Microbiology* **68**: 1290–6.

Shrader, E.A., Henry, T.R., Greeley, Jr, M.S., and Bradley, B.P. (2003) Proteomics in zebrafish exposed to endocrine disrupting chemicals. *Ecotoxicology* **12**: 485–8.

Simonsen, K.L., Churchill, G.A., and Aquadro, C. (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**: 413–29.

Simpson, A.G.B. and Patterson, D.J. (2008) Current perspectives on high-level groupings of protists. In *Genomics and Evolution of Microbial Eukaryotes*, L.A. Katz, and D. Bhattacharya (eds.), Oxford University Press, Oxford: 7–30.

Simpson, G.C. and Dean, C. (2002) Arabidopsis, the Rosetta stone of flowering time? *Science* **296**: 285–9.

Simpson, G.G., Dijkwel, P.P., Quesada, V., Henderson, I., and Dean, C. (2003) FY is an RNA 3′ end-processing factor that interacts with FCA to control the *Arabidopsis* floral transition. *Cell* **113**: 777–87.

Slack, J.M.W., Holland, P.W.H., and Graham, C.F. (1993) The zootype and the phylotypic stage. *Science* **361**: 490–2.

Small, J., Call, D.R., Brockman, F.J., Straub, T.M., and Chandler, D.P. (2001) Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. *Applied and Environmental Microbiology* **67**: 4708–16.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B.H., and Hood, L.E. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* **321**: 674–9.

Smant, G., Stokkermans, J.P.W.G., Yan, Y., De Boer, J.M., Baum, T.J., Wang, X., Hussey, R.S., Gommers, F.J., Henrissat, B., Davis, E.L., *et al.* (1998) Endogenous cellulases in animals: isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proceedings of the National Academy of Sciences USA* **95**: 4906–11.

Smyth, G.K., Yang, Y.H., and Speed, T. (2003) Statistical issues in cDNA microarray analysis. In *Functional Genomics: Methods and Protocols*, M.J. Brownstein and A.B. Khodurksy (eds). Humana Press, Totowa, NJ: 111–36.

Snape, J.R., Maund, S.J., Pickford, D.B., and Hutchinson, T.H. (2004) Ecotoxicogenomics: the challenge of integrating genomics into aquatic and terrestrial ecotoxicology. *Aquatic Toxicology* **67**: 143–54.

Snel, B., Bork, P., and Huynen, M. (1999) Genome phylogeny based on gene content. *Nature Genetics* **21**: 108–10.

Snell, T.W., Brogdon, S.E., and Morgan, M.B. (2003) Gene expression profiling in ecotoxicology. *Ecotoxicology* **12**: 475–83.

Soetaert, A., Moens, L.N., Van der Ven, K., Van Leemput, K., Naudts, B., Blust, R., and De Coen, W.M. (2006) Molecular impact of propiconazole on *Daphnia magna* using a reproduction-related cDNA array. *Comparative Biochemistry and Physiology Part C* **142**: 66–76.

Sokal, R.R. and Rohlf, F.J. (1995) *Biometry. The Principles and Practice of Statistics in Biological Research*, 3rd edn. W.H. Freeman and Company, San Fransisco.

Sommer, R.J. (2009) The future of evo-devo: model systems and evolutionary theory. *Nature Reviews Genetics* **10**: 416–22.

Sorrells, M.E., La Rota, M., Bermudez-Kandianis, C.E., Greene, R.A., Kantety, R., Munkvold, J.D., Miftahudin, J., Mahmoud, A., Ma, X., Gustafson, P.J. *et al*. (2003) Comparative DNA sequence analysis of wheat and rice genomes. *Genome Research* **13**: 1818–27.

Soto, A.M., Justicia, H., Wray, J.W., and Sonnenschein, C. (1991) p-Nonyl-phenol: an estrogenic xenobiotic released from 'modified' polystyrene. *Environmental Health Perspectives* **92**: 167–73.

Spellman, P.T. and Rubin, G.M. (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. *Journal of Biology* **1**: 5.

Spencer, J.F.T. and Spencer, D.M. (1997a) Ecology: where yeasts live. In *Yeasts in Natural and Artificial Habitats*, J.F.T. Spencer and D.M. Spencer (eds). Springer-Verlag, Berlin: 33–58.

Spencer, J.F.T. and Spencer, D.M. (1997b) Taxonomy: the names of the yeasts. In *Yeasts in Natural and Artificial Habitats*, J.F.T. Spencer and D.M. Spencer (eds). Springer-Verlag, Berlin: 11–32.

Sprague, J., Doerry, E., Douglas, S., Westerfield, M., and the ZFIN Group (2001) The Zebrafish Information Network (ZFIN): a resource for genetic, genomic and developmental research. *Nucleic Acids Research* **29**: 87–90.

Spring, J. (1997) Vertebrate evolution by interspecific hybridisation—are we polyploid? *FEBS Letters* **400**: 2–8.

St-Cyr, J., Derome, N., and Bernatchez, L. (2008) The transcriptomics of life-history trade-offs in whitefish species pairs (*Coregonus* sp.). *Molecular Ecology* **17**: 1850–70.

Stearns, S.C. (1992) *The Evolution of Life Histories*. Oxford University Press, Oxford.

Stearns, S.C. and Magwene, P. (2003) The naturalist in a world of genomics. *American Naturalist* **161**: 171–80.

Stein, L.D. (2004) End of the beginning. *Nature* **431**: 915–16.

Stinchcombe, J.R. and Hoekstra, H.E. (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* **100**: 158–70.

Stinchcombe, J.R., Weinig, C., Ungerer, M., Olsen, K.M., Mays, C., Halldorsdottir, S.S., Purugganan, M.D., and Schmitt, J. (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene *FRIGIDA*. *Proceedings of the National Academy of Sciences USA* **101**: 4712–17.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., Van Batenburg, M.F., Rifkin, S.A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P.E. *et al*. (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306**: 655–60.

Storz, J.F. (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* **14**: 671–88.

Strange, K. (2005) The end of 'naive reductionism': rise of systems biology or renaissance of physiology? *American Journal of Physiology Cell Physiology* **288**: 968–74.

Straub, P.F., Higham, M.L., Tanguy, A., Landau, B.J., Phoel, W.C., Hales, Jr, L.S., and Thwing, T.K.M. (2004) Suppression subtractive hybridization cDNA libraries to identify differentially expressed genes from contrasting fish habitats. *Marine Biotechnology* **6**: 386–99.

Strous, M., Pelletier, E., Mangenot, S., Rattei, T., Lehner, A., Taylor, M.W., Horn, M., Daims, H., Bartol-Mavel, D., Wincker, P., *et al*. (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440**: 790–4.

Stürzenbaum, S.R., Winters, C., Galay, M., Morgan, A.J., and Kille, P. (2001) Metal ion trafficking in earthworms. Identification of a cadmium-specific metallothionein. *Journal of Biological Chemistry* **276**: 34013–18.

Sumpter, J.P. (2005) Endocrine disrupters in the aquatic environment: an overview. *Acta Hydrochimica et Hydrobiologica* **33**: 9–16.

Sung, S. and Amasino, R.M. (2004) Vernalization in *Arabidopsis thaliana* is mediated by the PHD finger protein VIN3. *Nature* **427**: 159–64.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–95.

Takahashi, S., Watanabe, H., Munehara, H., Rüber, L., and Hori, M. (2009) Evidence for divergent natural selection of a Lake Tangangyika cichlid inferred from repeated radiations in body size. *Molecular Ecology* **18**: 3110–19.

Talbert, P.B., Bryson, T.D., and Henikoff, S. (2004) Adaptive evolution of centromere proteins in plants and animals. *Journal of Biology* **3**: 18.

Tamaoki, M., Nakajima, N., Kubo, A., Aono, M., Matsuyama, T., and Saji, H. (2003) Transcriptome analysis of O_3-exposed *Arabidopsis* reveals that multiple signal pathways act mutually antagonistically to induce gene expression. *Plant Molecular Biology* **53**: 443–56.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences USA* **96**: 2907–12.

Tamura, S., Hanada, M., Ohnishi, M., Katsura, K., Sasaki, M., and Kobayashi, T. (2002) Regulation of stress-activated protein kinase signaling pathways by protein phosphatases. *European Journal of Biochemistry* **269**: 1060–6.

Tanguy, A., Boutet, I., Laroche, J., and Moraga, D. (2005) Molecular identification and expression study of differentially regulated genes in the Pacific oyster *Crassostrea gigas* in response to pesticide exposure. *FEBS Journal* **272**: 390–403.

Tanksley, S.D. (1993) Mapping polygenes. *Annual Review of Genetics* **27**: 205–33.

Taroncher-Oldenburg, G., Griner, E.M., Francis, C.A., and Ward, B.B. (2003) Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. *Applied and Environmental Microbiology* **69**: 1159–71.

Tatar, M., Bartke, A., and Antebi, A. (2003) The endocrine regulation of aging by insulin-like signals. *Science* **299**: 1346–51.

Tatar, M., Kopelman, A., Epstein, D., Tu, M.-P., Yin, C.-M., and Garofalo, R.S. (2001) A mutant *Drosophila* insulin receptor homolog that extends life-span and impairs neuroendocrine function. *Science* **292**: 107–1110.

Taylor, G. (2002) *Populus*: Arabidopsis for forestry. Do we need a model tree? *Annals of Botany* **90**: 681–9.

Teske, A., Dhillon, A., and Sogin, M.L. (2003) Genomic markers of ancient anaerobic microbial pathways: sulfate reduction, methanogenesis, and methane oxidation. *Biological Bulletin* **204**: 186–91.

Thiele, D.J. (1992) Metal-regulated transcription in eukaryotes. *Nucleic Acids Research* **20**: 1183–91.

Thomas, M.A. and Klaper, R. (2004) Genomics for the ecological toolbox. *Trends in Ecology and Evolution* **19**: 439–45.

Thompson, D.A.W. (1917) *On Growth and Form*. Cambridge University Press, Cambridge.

Tiedje, J.M. and Zhou, J. (2004) Future perspectives: genomics beyond single cells. In *Microbial Functional*

Genomics, J. Zhou, D.K. Thompson, Y. Xu, and J.M. Tiedje (eds). John Wiley & Sons, Hoboken, NJ: 477–86.

Tillier, E.R.M. and Collins, R.A. (2000) The contribution of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *Journal of Molecular Evolution* **50**: 249–57.

Timmermans, M.J.T.N., Roelofs, D., Mariën, J., and Van Straalen, N.M. (2008) Revealing pancrustacean relationships: Phylogenetic analysis of ribosomal protein genes places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear DNA markers. *BMC Evolutionary Biology* **8**: 83.

Timmermans, M.J.T.N., Roelofs, D., Nota, B., Ylstra, B., and Holmstrup, M. (2009) Sugar sweet springtails: on the transcriptional response of *Folsomia candida* (Collembola) to desiccation stress. *Insect Molecular Biology* **18**: 737–46.

Tiquia, S.M., Wu, L., Chong, S.C., Passovets, S., Xu, D., Xu, Y., and Zhou, J. (2004) Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. *BioTechniques* **36**: 664–75.

Tirosh, I., Barkai, N., and Verstrepen, K.J. (2009) Promoter architecture and the evolvability of gene expression. *Journal of Biology* **8**: 95.

Tirosh, I., Weinberger, A., Carmi, M., and Barkai, N. (2006) A genetic signature of interspecies variations in gene expression. *Nature Genetics* **38**: 830–4.

Tissenbaum, H.A. and Guarante, L. (2001) Increased dosage of a *sir-2* gene extends lifespan in *Caenorhabditis elegans*. *Nature* **410**: 227–30.

Todd, J.D., Rogers, R., Li, Y.G., Wexler, M., Bond, P.L., Sun, L., Curson, A.R.J., Malin, G., Steinke, M., and Johnston, A.W.B. (2007) Structural and regulatory genes required to make the gas dimethyl sulfide in bacteria. *Science* **315**: 666–9.

Tong, A.H.Y., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M. *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science* **303**: 808–13.

Tonsor, S.J., Alonso-Blanco, C., and Koornneef, M. (2005) Gene function beyond the single trait: natural variation, gene effects and evolutionary ecology in *Arabidopsis thaliana*. *Plant, Cell and Environment* **28**: 2–20.

Toth, A.L., Varala, K., Newman, T.C., Miquez, F.E., Hutchison, S.K., Willoughby, D.A., Simons, J.F., Egholm, M., Hunt, J.H., Hudson, M.E., and Robinson, G.E. (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science* **318**: 441–4.

Town, C.D. (2006) Annotating the genome of *Medicago truncatula*. *Current Opinion in Plant Biology* **9**: 122–7.

Townsend, J.P., Cavalieri, D., and Hartl, D.L. (2003) Population genetic variation in genome-wide gene expression. *Molecular Biology* and *Evolution* **20**: 955–63.

Treinin, M., Shliar, J., Jiang, H., Powell-Coffman, J.A., Bromberg, Z., and Horowitz, M. (2003) HIF-1 is required for heat acclimation in the nematode *Caenorhabditis elegans*. *Physiological Genomics* **14**: 17–24.

Treusch, A.H., Kletzin, A., Raddatz, G., Ochsenreiter, T., Quaiser, A., Meurer, G., Schuster, S.C., and Schleper, C. (2004) Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. *Environmental Microbiology* **6**: 970–80.

Treusch, A.H., Leininger, S., Kletzin, A., Schuster, S.C., Klenk, H.-P., and Schleper, C. (2006) Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environmental Microbiology* **7**: 1985–95.

Tribolium Genome Sequencing Consortium (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452**: 949–55.

Tringe, S.G., Von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., Podar, M., Short, J.M., Mathur, E.J., Detter, J.C. *et al*. (2005) Comparative metagenomics of microbial communities. *Science* **308**: 554–7.

Trivedi, S., Ueki, T., Yamaguchi, N., and Michibata, H. (2003) Novel vanadium-binding proteins (vanabins) identified in cDNA libraries and the genome of the ascidian *Ciona intestinalis*. *Biochimica et Biophysica Acta* **1630**: 64–70.

Tsai, I., Bensasson, D., Burt, A., and Koufopanou, V. (2008) Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proceedings of the National Academy of Sciences USA* **105**: 4957–62.

Tudge, C. (2000) *The Variety of Life*. Oxford University Press, Oxford.

Turner, T.L., Bourne, E.C., Von Wettberg, E.J, Hu, T.T. and Nuzhdin, S.V. (2010) Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics* **42**: 260–3.

Tusher, V.G., Tibshrinai, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA* **98**: 5116–21.

Tuskan, G.A., DiFazio, S., Jansson, S., Bohlma, J., Grigoriev, I., Helsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa*. *Science* **313**: 1596–1604.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.A., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., and Banfield, J.F. (2004) Community

structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Uchiyama, T., Abe, T., Ikemura, T., and Watanabe, K. (2005) Substrate-induced gene-expression screening of environmental metagenomic libraries for isolation of catabolic genes. *Nature Biotechnology* **23**: 88–93.

Ueki, T., Adachi, T., Kawano, S., Aoshima, M., Yamaguchi, N., Kanamori, K., and Michibata, H. (2003) Vanadium-binding proteins (vanabins) from a vanadium-rich ascidian *Ascidia sydneiensis samea*. *Biochimica et Biophysica Acta* **1626**: 43–50.

Urakawa, H., Noble, P.A., El Fantroussi, S., Kelly, J.J., and Stahl, D.A. (2002) Single-base-pair discrimination of terminal mismatches by using oligonucleotide microarrays and neural network analyses. *Applied and Environmental Microbiology* **68**: 235–44.

Valentine, J.W. (2004) *On the Origin of Phyla*. The University of Chicago Press, London.

Valinsky, L., Della Vedova, G., Scupham, A., Alvey, S., Figueroa, A., Yin, B., Hartin, R.J., Chrobak, M., Crowley, D.E., Jiang, T., and Borneman, J. (2002) Analysis of bacterial community composition by oligonucleotide fingerprinting of rRNA genes. *Applied and Environmental Microbiology* **68**: 3243–50.

Valls, M., Bofill, R., Romero-Isart, N., Gonzàlez-Duarte, R., Abián, J., Carrascal, M., Gonzàlez-Duarte, P., Capdevila, M., and Atrian, S. (2000) *Drosophila* MTN: a metazoan copper-thionein related to fungal forms. *FEBS Letters* **467**: 189–94.

Valverde, F., Mouradov, A., Soppe, W., Ravenscroft, D., Samach, A., and Coupland, G. (2004) Photoreceptor regulation of CONSTANS protein in photoperiodic flowering. *Science* **303**: 1003–6.

Van Aggelen, G., Ankley, G.T., Baldwin, W.S., Bearden, D.W., Benson, W.H., Chipman, J.K., Collette, T.W., Craft, J.A., Denslow, N.D., Embry, M.R., *et al.* (2010) Integrating omic technologies into aquatic ecological risk assessment and environmental monitoring: hurdles, achievements, and future outlook. *Environmental Health Perspectives* **118**: 1–5.

Van Bers, N.E.M., Van Oers, K., Kerstens, H.H.D., Dibbits, B.W., Crooijmans, R.P.M.A., Visser, M.E., and Groenen, M.A.M. (2010) Genome-wide SNP detection in the great tit *Parus major* using high throughput sequencing. *Molecular Ecology* **19**: 89–99.

Van de Mortel, J.E., Villanueva, L.A., Schat, H., Kwekkeboom, J., Coughlan, S., Moerland, P.D., Van Themaat, E.V.L., Koornneef, M., Aarts, M.G.M. (2006) Large expression differences in genes for iron and zinc homeostasis, stress response, and lignin biosynthesis distinguish roots of *Arabidopsis thaliana* and the related metal hyperaccumulator *Thlaspi caerulescens*. *Plant Physiology* **142**, 1127–47.

Van der Wielen, P.W.J.J., Bolhuis, H., Borin, S., Daffonchio, D., Corselli, C., Giuliano, L., D'Auria, G., De Lange, G.J., Huebner, A., Varnavas, S.P. *et al.*, and the BioDeep Scientific Party (2005) The enigma of prokaryotic life in deep hypersaline anoxic basins. *Science* **307**: 121–3.

Van der Zee, M., Berns, N., and Roth, S. (2005) Distinct functions of the *Tribolium zerknüllt* genes in serosa specification and dorsal closure. *Current Biology* **15**: 624–36.

Van Elsas, J.D., Garbeva, P., and Salles, J. (2002) Effects of agronomical measures on the microbial diversity of soils as related to the suppression of soil-borne plant pathogens. *Biodegradation* **13**: 29–40.

Van Noordwijk, A.J. and De Jong, G. (1986) Acquisition and allocation of resources: their influence on variation in life history tactics. *American Naturalist* **128**: 137–42.

Van Regenmortel, M.H.V. (2004) Reductionism and complexity in molecular biology. *EMBO Reports* **5**: 1016–20.

Van Spanning, R.J.M., Delgado, M.J., and Richardson, D.J. (2005) The nitrogen cycle: denitrification and relationship to N_2 fixation. In Nitrogen Fixation in Agriculture, *Forestry, Ecology and the Environment*, D. Werner and E. Newton (eds). Kluwer Academic Publishers, Dordrecht: 277–342.

Van Straalen, N.M. (1985) Comparative demography of forest floor Collembola populations. *Oikos* **45**: 253–65.

Van Straalen, N.M. (2002) Assessment of soil contamination—a functional perspective. *Biodegradation* **13**: 41–52.

Van Straalen, N.M. (2003) Ecotoxicology becomes stress ecology. *Environmental Science and Technology* **37**: 324A–30A.

Van Straalen, N.M. and Hoffmann, A.A. (2000) Review of experimental evidence for physiological costs of tolerance to toxicants. In *Demography in Ecotoxicology*, J.E. Kammenga and R. Laskowski (eds). John Wiley and Sons, Chichester: 147–61.

Van Straalen, N.M. and Roelofs, D. (2008) Genomics technology for assessing soil pollution. *Journal of Biology* **7**: 19.

Van Valen, L. (1973) A new evolutionary law. *Evolutionary Theory* **1**: 1–30.

Vasemägi, A., Nilsson, J., and Primmer, C.R. (2005) Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). *Molecular Biology* and *Evolution* **22**: 1067–76.

Vaughn, M.W., Tanurdzic, M., Lippman, Z., Jiang, H., Carrasquillo, R., Rabinowicz, P.D., Dedhia, N., McCombie, W.R., Agier, N., Bulski, A., *et al.* (2007) Epigenetic natural variation in *Arabidopsis thaliana*. *PLoS Biology* **5**: e174.

Venkatesh, B. (2003) Evolution and diversity of fish genomes. *Current Opinion in Genetics & Development* **13**: 588–92.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W. *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Vera, J.C., Wheat, C.W., Fescemyer, H.W., *et al*. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**: 1636–47.

Vido, K., Spector, D., Lagniel, G., Lopez, S., Toledano, M.B., and Labarre, J. (2001) A proteome analysis of the cadmium response in *Saccharomyces cerevisae*. *Journal of Biological Chemistry* **276**: 8469–74.

Vinogradov, A.E. (2004) Testing genome complexity. *Science* **304**: 389–90.

Voelckel, C. and Baldwin, I.T. (2004) Generalist and specialist lepidopteran larvae elicit different transcriptional responses in *Nicotiana attenuata*, which correlate with larval FAC profiles. *Ecology Letters* **7**: 770–5.

Voelckel, C., Weisser, W., and Baldwin, I.T. (2004) An analysis of plant-aphid interactions by different microarray hybridization techniques. *Molecular Ecology* **13**: 3187–95.

Vogel, T.M., Simonet, P., Hirsch, P.R., Jansson, J.K., Tiedje, J.M., Van Elsas, J.D., Bailey, M.J., Nalin, R., and Philippot, L. (2009) TerraGenome: a consortium for the sequencing of a soil metagenome. *Nature Reviews Microbiology* **7**: 252.

Voget, S., Leggewie, C., Uesbeck, A., Raasch, C., Jaeger, K.-E., and Streit, W.R. (2003) Prospecting for novel biocatalysts in a soil metagenome. *Applied and Environmental Microbiology* **69**: 6235–42.

Von Mering, C., Hugenholtz, P., Raes, J., Tringe, S.G., Doerks, T., L.J., J., Ward, N., and Bork, P. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**: 1126–30.

Von Schalburg, K.R., Rise, M.L., Cooper, G.A., Brown, G.D., Ross Gibbs, A., Nelson, C.C., Davidson, W.S., and Koop, B.F. (2005) Fish and chips: Various methodologies demonstrate utility of a 16,006-gene salmonid microarray. *BMC Genomics* **6**: 126.

Voordouw, G., Voordouw, J.K., Karkhoff-Schweizer, R.R., Fedorak, P.M., and Westlake, D.W.S. (1991) Reverse sample genome probing, a new technique for identification of bacteria in environmental samples by DNA hybridization, and its application to the identification of sulfate-reducing bacteria in oil field samples. *Applied and Environmental Microbiology* **57**: 3070–8.

Wagner, M., Smidt, H., Loy, A., and Zhou, J. (2007) Unravelling microbial communities with DNA-microarrays: challenges and future directions. *Microbial Ecology* **53**: 489–506.

Wahlund, T.M., Hadaegh, A.R., Clark, R., Nguyen, B., Fanelli, M., and Read, B. (2004) Analysis of expressed sequence tags from calcifying cells of marine coccolithophorid (*Emiliania huxleyi*). *Marine Biotechnology* **6**: 278–90.

Walbot, V. (2000) A green chapter in the book of life. *Nature* **408**: 794–5.

Walker, B., Kinzig, A., and Langridge, J. (1999) Plant attribute diversity, resilience, and ecosystem function: the nature and significance of dominant and minor species. *Ecosystems* **2**: 95–113.

Walker, D.W., McGoll, G., Jenkins, N.L., Harris, E.E., and Lithgow, G.L. (2000) Evolution of lifespan in *C. elegans*. *Nature* **405**: 296–7.

Walker, C.H., Hopkin, S.P., Sibly, R.M., and Peakall, D. (2001) *Principles of Ecotoxicology*, 2nd edn. Taylor & Francis, London.

Walker, J.J., Spear, J.R., and Pace, N.R. (2005) Geobiology of a microbial endolithic community in the Yellowstone geothermal environment. *Nature* **434**: 1011–14.

Walser, J.-C., Chen, B., and Feder, M.E. (2006) Heat-shock promoters: targets for evolution by P-transposable elements in *Drosophila*. *PLoS Genetics* **2**: e165.

Wang, J. and Kim, S.K. (2003) Global analysis of dauer gene expression in *Caenorhabditis elegans*. *Development* **130**: 1621–34.

Wang, W., Cherry, M., Botstein, D., and Li, H. (2002) A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences USA* **99**: 16893–8.

Ward, B.B. (2002) How many species of prokaryotes are there? *Proceedings of the National Academy of Sciences USA* **99**: 10234–6.

Ward, D.M., Cohan, F.M., Bhaya, D., Heidelberg, J.F., Kühl, M., and Grossman, A.R. (2008) Genomics, environmental genomics and the issue of microbial species. *Heredity* **100**: 207–19.

Ward, N.L., Challacombe, J.F., Janssen, P.H., Henrissat, B., Coutinho, P.M., Wu, M., Xie, G., Haft, D.H., Sait, M., Badger, J., *et al.* (2009) Three genomes from the phylum *Acidobacteria* provide insight into the lifestyles of these

microorganisms in soils. *Applied and Environmental Microbiology* **75**: 2046–56.

Warren, W.C., Clayton, D.F., Ellegren, H., Arnold, A.P., Hillier, L.W., Künstner, A., Searle, S., White, S., Vilella, A.J., Fairley, S., *et al.* (2010) The genome of a songbird. *Nature* **464**: 757–62.

Waters, M.D. and Fostel, J.M. (2004) Toxicogenomics and systems toxicology: aims and prospects. *Nature Reviews Genetics* **5**: 936–48.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E. *et al*. (1987) Report of the *ad hoc* committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic Bacteriology* **37**: 463–4.

Weinbauer, M.G. and Rassoulzadegan, F. (2004) Are viruses driving microbial diversification and diversity? *Environmental Microbiology* **6**: 1–11.

Weinig, C., Ungerer, M.C., Dorn, L.A., Kane, N.C., Toyonaga, Y., Halldorsdottir, S.S., Mackay, T.F.C., Purugganan, M.D., and Schmitt, J. (2002) Novel loci control variation in reproductive timing in *Arabidopsis thaliana* in natural environments. *Genetics* **162**: 1875–84.

Weitzman, J.B. (2002) Transcriptional territories in the genome. *Journal of Biology* **1**: 2.

Weller, D.M., Raaijmakers, J.M., McSpadden Gardener, B.B., and Thomashow, L.S. (2002) Microbial populations responsible for specific soil suppressiveness to plant pathogens. *Annual Review of Phytopathology* **40**: 309–48.

Wellington, E.M.H., Berry, A., and Krsek, M. (2003) Resolving functional diversity in relation to microbial community structure in soil: exploiting genomics and stable isotope probing. *Current Opinion in Microbiology* **6**: 295–301.

Wenger, R.H. (2002) Cellular adaptation to hypoxia O_2 sensing protein hydroxylases, hypoxia-inducible O_2^- sensing protein hydroxylases, hypoxia-inducible transcription factors and O_2-regulated gene expression. *FASEB Journal* **16**: 1151–62.

Werck-Reichhart, D. and Feyereisen, R. (2000) Cytochromes P450: a success story. *Genome Biology* **1**: reviews 3003.1–3003.9.

Werck-Reichhart, D., Bak, S., and Paquette, S. (2002) Cytochromes P450. In *The Arabidopsis Book*, C.R. Somerville and E.M. Meyerowitz (eds). American Society of Plant Biologists, Rockville, IL: 10.119/tab.0028.

West, M.A.L., Kim, K., Kliebenstein, D.J., Van Leeuwen, H., Michelmore, R.W., Doerge, R.W., and St Clair, D.A. (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* **175**: 1441–50.

West-Eberhard, M.J. (2005) Developmental plasticity and the origin of species differences. *Proceedings of the National Academy of Sciences USA* **102**: 6543–9.

Westerhoff, H.V. and Palsson, B.O. (2004) The evolution of molecular biology into systems biology. *Nature Biotechnology* **22**: 1249–52.

Westerhoff, H.V., Hofmeyr, J.-H.S., and Khodolenko, B.N. (1994) Getting into the inside of cells using metabolic control analysis. *Biophysical Chemistry* **50**: 273–83.

Whitaker, R.J., Grogan, D.W., and Taylor, J.W. (2003) Geographic barriers isolate endemic populations of hyperthermophilic Archaea. *Science* **301**: 976–8.

Whitfield, C.W., Band, M.R., Bonaldo, M.F., Kumar, C.G., Liu, L., Pardinas, J.R., Robertson, H.M., Soares, M.B., and Robinson, G.E. (2002) Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. *Genome Research* **12**: 555–66.

Whitaker, R.J., Grogan, D.W. and Taylor, J.W. (2003) Geographic barriers isolate endemic populations of hyperthermophilic Archaea. *Science* 301: 976–8.

Whitehead, A. and Crawford, D.L. (2006a) Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Sciences USA* **103**: 5425–30.

Whitehead, A. and Crawford, D.L. (2006b) Variation within and among species in gene expression: raw material for evolution. *Molecular Ecology* **15**: 1197–211.

Whiteley, A.R., Derome, N., Rogers, S.M., St-Cyr, J., Laroche, J., Labbe, A., Nolte, A., Renaut, S., Jeukens, J., and Bernatchez, L. (2008) The phenomics and expression quantitative trait locus mapping of brain transcriptomes regulating adaptive divergence in lake whitefish species pairs (*Coregonus* sp.). *Genetics* **180**: 147–64.

Williams, L.M. and Oleksiak, M.F. (2008) Signatures of selection in natural populations adapted to chronic pollution. *BMC Evolutionary Biology* **8**: 282.

Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C.S., Sutton, G., *et al.* (2008) The Sorcerer II Global Ocean Sampling Expedition: Metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* **3**: e1456.

Wilson, K.H., Wilson, W.J., Radosevich, J.L., DeSantis, T.Z., Viswanathan, V.S., Kuczmarski, T.A., and Andersen, G.L. (2002) High-density microarray of small-subunit ribosomal DNA probes. *Applied and Environmental Microbiology* **68**: 2535–41.

Wimp, G.M., Young, W.P., Woolbright, S.A., Keim, P., and Whitham, T.G. (2004) Conserving plant genetic diversity for dependent animal communities. *Ecology Letters* **7**: 776–80.

Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2004) Evolutionary changes in *cis* and *trans* gene regulation. *Nature* **430**: 85–8.

Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2008a) Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genetics* **40**: 346–50.

Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2008b) Independent effects of *cis*- and *trans*-regulatory variation on gene expression in *Drosophila melanogaster*. *Genetics* **178**: 1831–5.

Wittstock, U. and Gershenzon, J. (2002) Constitutive plant toxins and their role in defense against herbivores and pathogens. *Current Opinion in Plant Biology* **5**: 300–7.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V., and Koonin, E.V. (2002) Genome trees and the tree of life. *Trends in Genetics* **18**: 472–9.

Wood, H.M., Grahame, J.W., Humphray, S., Rogers, J., and Butlin, R.K. (2008) Sequence differentiation in regions identified by a genome scan for local adaptation. *Molecular Ecology* **17**: 3123–35.

Wood, W.B., Hecht, R., Carr, S., Vanderslice, R., Wolf, N., and Hirsh, D. (1980) Parental effects and phenotypic characterization of mutations that affect early development in *Caenorhabditis elegans*. *Developmental Biology* **74**: 446–69.

Wood, D.W., Setubal, J.C., Kaul, R., Monks, D.E., Kitajima, J.P., Okura, V.K., Zhou, Y., Chen, L., Wood, G.E., Almeida, Jr, N.F. *et al*. (2001) The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* **294**: 2317–23.

Woods, I.G., Kelly, P.D., Chu, F., Ngo–Hazelett, P., Yan, Y.-L., Huang, H., Postlewaith, J.H., and Talbot, W.S. (2000) A comparative map of the zebrafish genome. *Genome Research* **10**: 1903–14.

Wray, G.A., Hahan, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., and Romano, L.A. (2003) The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution* **20**: 1377–419.

Wright, S.I. and Gaut, B.S. (2005) Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology* and *Evolution* **22**: 506–19.

Wu, L., Thompson, D.K., Li, G., Hurt, R.A., Tiedje, J.M., and Zhou, J. (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Applied and Environmental Microbiology* **67**: 5780–90.

Wu, M., Sun, L.V., Vamathevan, J., Riegler, M., Deboy, R., Brownlie, J.C., McGraw, E.A., Martin, W., Esser, C., Ahmadinejad, N. *et al*. (2004) Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile elements. *PloS Biology* **2**: 0327–41.

Wullschleger, S.D., Jansson, S., and Taylor, G. (2002) Genomics and forest biology: *Populus* emerges as the perennial favourite. *The Plant Cell* **14**: 2651–5.

Wynne-Edwards, K.E. (2001) Evolutionary biology of plant defenses against herbivory and their predictive implications for endocrine disruptor susceptibility in vertebrates. *Environmental Health Perspectives* **109**: 443–8.

Yang, Y.H., and Speed, T.P. (2002) Design issues for cDNA microarray experiments. *Nature Reviews Genetics* **3**: 579–88.

Yang, Z. and Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution* **15**: 496–503.

Yanovsky, M.J. and Kay, S.A. (2003) Living by the calendar: how plants know when to flower. *Nature Reviews Molecular Cell Biology* **4**: 265–75.

Yarzábal, A., Appia-Ayme, C., Ratouchniak, J., and Bonnefoy, V. (2004) Regulation of the expression of the *Acidithiobacillus ferrooxidans rus* operon encoding two cytochromes c, a cytochrome oxidase and rusticyanin. *Microbiology* **150**: 2113–23.

Ye, R.W. and Thomas, S.M. (2001) Microbial nitrogen cycles: physiology, genomics and applications. *Current Opinion in Microbiology* **4**: 307–12.

Ye, R.W., Wang, T., Bedzyk, L., and Croker, K.M. (2001) Applications of DNA microarrays in microbial systems. *Journal of Microbiological Methods* **47**: 257–72.

Yergeau, E., Kang, S., He, Z., Zhou, J., and Kowalchuk, G.A. (2007) Functional microarray analysis of nitrogen and carbon cycling genes across an Antarctic latitudinal gradient. *The ISME Journal* **1**: 163–79.

Yergeau, E., Schoondermark-Stolk, S.A., Brodie, E.L., Déjean, S., DeSantis, T.Z., Gonçalves, O., Piceno, Y.M., Andersen, G.L., and Kowalchuk, G.A. (2009) Environmental microarray analysis of Antarctic soil microbial communities. *The ISME Journal* **3**: 340–51.

Yoch, D.C. (2002) Dimethylsulfionopropionate: its sources, role in the marine food web, and biological degradation to dimethylsulfide. *Applied and Environmental Microbiology* **68**: 5804–15.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W., *et al.* (2007) The *Sorcerer II* Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biology* **5**: e16.

Yu, C.-W., Chen, J.-H., and Lin, L.-Y. (1997) Metal-induced metallothionein gene expression can be inactivated by protein kinase C inhibitor. *FEBS Letters* **420**: 69–73.

Yu, J., Hu, S., Wang, J., Wong, G.K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**: 79–92.

Zayed, A. and Whitfield, C.W. (2008) A genome-wide signature of positive selection in ancient and recent invasive expansions of the honey bee *Apis mellifera*. *Proceedings of the National Academy of Sciences USA* **105**: 3421–6.

Zavala, J.A., Patankar, A.G., Gase, K., and Baldwin, I.T. (2004) Constitutive and inducible trypsin proteinase inhibitor production incurs large fitness costs in *Nicotiana attenuata*. *Proceedings of the National Academy of Sciences USA* **101**: 1607–12.

Zdobnov, E.M., Von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R.R., Christophides, G.K., Thomasova, D., Holt, R.A., Subramanian, G.M. *et al*. (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**: 149–59.

Zera, A.J. and Harshman, L.G. (2001) The physiology of life history trade-offs in animals. *Annual Review of Ecology and Systematics* **32**: 95–126.

Zerbino, D.R., and Birney, E. (2008) Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* **18**: 821–9.

Zhang, B., Egli, D., Georgiev, O., and Schaffner, W. (2001) The *Drosophila* homolog of mammalian zinc finger factor MTF-1 activates transcription in response to heavy metals. *Molecular and Cellular Biology* **21**: 4505–14.

Zhang, L.V., King, O.D., Wong, S.L., Goldberg, D.S., Tong, A.H.Y., Lesage, G., Andrews, B., Bussey, H., Boone, C., and Roth, F.P. (2005) Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *Journal of Biology* **4**: 6.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.-L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E., *et al.* (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* **126**: 1189–201.

Zhang, X., Shiu, S.-H., Cal, A., and Borevitz, J.O. (2008) Global analysis of genetic, epigenetic and transcriptional polymorphisms in *Arabidopsis thaliana* using whole genome tiling arrays. *PLoS Genetics* **4**: e1000032.

Zhou, J. (2003) Microarrays for bacterial detection and microbial community analysis. *Current Opinion in Microbiology* **6**: 288–94.

Zhou, J. and Thompson, D.K. (2004) Application of microarray-based genomic technology to mutation analysis and microbial detection. In *Microbial Functional Genomics*, J. Zhou, D.K. Thompson, Y. Xu, and J.M. Tiedje (eds). John Wiley & Sons, Hoboken, NJ: 451–76.

Zhou, J., Thompson, D.K., and Tiedje, J.M. (2004) Genomics: toward a genome-level understanding of the structure, functions, and evolution of biological systems. In *Microbial Functional Genomics*, J. Zhou, D.K. Thompson, Y. Xu, and J.M. Tiedje (eds). John Wiley & Sons, Hoboken, NJ: 1–19.

Zhu-Salzman, K., Salzman, R.A., Ahn, J.-E., and Koiwa, H. (2004) Transcriptional regulation of sorghum defense determinants against a phloem-feeding aphid. *Plant Physiology* **134**: 420–31.

Zou, S., Meadows, S., Sharp, L., Jan, L.Y., and Jan, Y.N. (2000) Genome-wide study of aging and oxidative stress response in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences USA* **97**: 13726–31.

Index