生命科学专论

生态基因组学

导论

AN INTRODUCTION TO Ecological Genomics

葛颂 审校
陈明生 选译
[荷]D. 罗洛夫斯
[荷]N. M. 斯特拉伦 编著

科学出版社
www.sciencep.com

An Introduction to Ecological Genomics

生态基因组学导论

〔荷〕N. M. 斯特拉伦
D. 罗洛夫斯 编著

陈明生 选译
葛 颂 审校

科学出版社
北京

图字:01-2007-4863 号

内 容 简 介

本书介绍了一个非常前沿的领域——生态基因组学。这门学科将基因组学的研究手段和方法引入生态学领域,从基因组学的角度考察了三个生态学的基本问题:生态系统中群落的结构和功能、不同的生活史类型和变异以及生态位的界定。

虽然其内容主要来源于零散的第一手研究论文,但在作者的精心组织下,本书的前沿性和系统性得到了统一,适合作为本科生和研究生的参考用书。对生态学研究感兴趣的读者可以学习如何应用基因组学的技术深化生态学的研究;对其他学科(如分子生物学)有兴趣的读者可以学习生态学的基本概念和基础知识,以及如何将基因组学和生态学相结合,从而形成这门新的前沿交叉学科。

图书在版编目(CIP)数据

生态基因组学导论/(荷)斯特拉伦(Straalen, N. M.)等编著;陈明生选译;葛颂审校. —北京:科学出版社,2008

ISBN 978-7-03-020506-3

Ⅰ. 生… Ⅱ. ①斯…②陈…③葛… Ⅲ. 生态学-基因组-研究 Ⅳ. Q14 Q343.1

中国版本图书馆 CIP 数据核字(2007)第 178090 号

责任编辑:王 静 李 晓/责任校对:宋玲玲

责任印制:钱玉芬/封面设计:福瑞来书装

科学出版社出版

北京东黄城根北街 16 号

邮政编码: 100717

http://www.sciencep.com

新蕾印刷厂印刷

科学出版社发行 各地新华书店经销

*

2008 年 1 月第 一 版 开本:787×1092 1/16

2008 年 1 月第一次印刷 印张:20 1/4

印数:1—2 500 字数:461 000

定价:58.00 元

(如有印装质量问题,我社负责调换〈环伟〉)

译者的话

记得上大学时在《植物生态学》课上，教授讲池塘群落的生态、讲食物链、讲营养循环等，当时觉得生态学的研究颇有趣味。之后的我一直从事基因组学的研究，期间偶尔也有与生态学交叉的时刻，但是对于生态学研究什么样的问题，以及如何进行生态学的研究一直有一种朦胧的感觉。令我印象深刻的是在一次国际会议上，听一位国际知名学者讲到Metagenomics(环境基因组学或宏基因组学)。大意是说我们对地球上生命的认识是非常有限的，如果我们从任何地方取一块土壤，其中可能含有成千上万种生命是我们尚未了解的。我们目前认识到的微生物都是能够进行人工培养的，而地球上更多的微生物是无法进行人工培养的。随着生态基因组学的发展，我们可以直接对土壤、海水等地球环境中的微生物进行大规模测序，从而能够逐步揭示这些未知生命的奥秘。

《生态基因组学导论》这本书就很好地介绍了基因组学的研究方法与生态学概念的结合。它涵盖了近年来生态基因组学领域的一些典型实例，省去了我们查找浩如烟海的文献之麻烦，综合了从事生态基因组学研究的基本思路，是帮助我们学习这门新兴交叉学科的很好的参考用书。

生态学研究什么问题？什么是 Metagenomics？如何利用基因组学方法进行生态学研究？带着这些问题，从事生命科学相关学科的青年学者有必要读一下这本书，以开阔视野，为未来的发展增加知识的储备。

陈明生

2007 年 11 月

前　言

生态基因组学是一个崭新的、令人兴奋的研究领域。本书对该领域进行了介绍，适合作为硕士研究生和入门阶段的博士生教程。

当着手制定一个生态基因组学的国家研究计划时，我们意识到有必要把这门新兴学科的各方面信息综合起来。若要在这样一个新学科中建立研究计划，无论是学生还是教师，都需要先掌握这门新学科。尽管获得博士学位意味着精通一门专业领域，可想要成为一名成熟的科学工作者，博士生们必须在精通专业知识的同时有宽广的知识面。这种教学方式可以称为T型教育："T"的横线代表知识的宽广度，竖线代表研究的深度，要一直深到问题的本源。本书就采用了这种方式。

阅读本书需要具备大学本科生物学的基本知识：生态学、进化生物学、微生物学、植物生理学、动物生理学、遗传学以及分子生物学。在编写过程中我们尽量和这些课程的通用教科书相联系，同时也适当考虑生态基因组学的学生背景不一。然而，本书的主要对象还是那些生态学和进化生物学专业的学生，这也是为什么本书将重点放在了对这些学生来说是比较新的研究内容。

进化基因组学和生物信息学是生态基因组学的相伴学科。在过去的十年间，上述两门学科都得到了巨大的发展。多本有关生物信息学的教材已经问世，进化基因组学所包括的学科，如比较基因组学、系统发育分析以及分子进化，已经成为独立的学科。当然这些学科范围太广而不可能全部涵盖在一本入门的生态基因组学教材中，但显然，进化基因组学值得作为一本独立的教材。

本书的组织围绕着现代生态学的三个重大问题，特别是那些与基因组学密切相关的科学问题。一开始，我们使用了挑战性的语言来描述解决生态学问题的基因组学方法，也许我们目前还不能回答这些问题，但我们决不回避这些尚且无法回答的、可以自由探索的问题。我们希望能借此来激发对问题的讨论，同时提供来自实际的证据。我们在每一章的最后增加了一节"初步评估"，用于强调以问题为导向的研究方法。结合在第一章的相关信息，读者能够很快掌握每一章节的主旨，哪怕是先把分子原理的详细论证和例证放在一边。

本书研究实例多选自于2000年以后出版的文献资料。尽管如此，一本基因组学的书总是面临很快过时的风险：基因组学知识积累和认识的速度是前所未有的。但是，我们希望我们采用的以问题为导向的方法在未来若干年内都是有用的，即使有新的和更好的例证出现。

在本书成书之前，学术论文是生态基因组学领域唯一的文献来源。这些文献资料虽然令人鼓舞，但同时也很分散。目前很多有关遗传和进化的书籍都有一个章节论述基因组学。Gibson和Muse在2002年出版了一本基因组学的导读，但该书没有覆盖生态学问题。因此，对于我们来说，写作这本书也是在开拓一个新的领地。我们试图使该领域纹理清晰，希望使生态基因组学成为显学。我们非常欢迎读者提出建设性的批评意见和建议。

我们衷心感谢以下同事审阅了本书部分文稿、提出补漏意见或者帮助修改文法，他们

是 Martin Feder，Claire Hengeveld，Jan Kammenga，René Klein Lankhorst，Bas Kooijman，Jan Kooter，Wilfred Röling 和 Martijn Timmermans。我们感谢 Desiree Hoonhout 和 KarinUyldert 检查文献目录，以及 Nico Schaefers 准备插图。牛津大学出版社的 Ian Sherman 提供了很有启发的讨论。我们感谢阿姆斯特丹自由大学动物生态学系的全体同事的友谊和鼓励。我们作者之一(N. M. van Straalen)同时感谢阿姆斯特丹自由大学地球和生命科学系提供公休的机会，正是在公休假期本书的大部分得以成稿。

Nico M. van Straalen 和 Dick Rolofs

阿姆斯特丹

2005 年 7 月

目　录

CHAPTER 1

什么是生态基因组学？

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 5 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. We also touch on some of its possible applications.

1.1 渗透到生态学领域的基因组学革命

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 humane 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 indeed has 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.

Fig. 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 have their 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.

Microtechnogy. 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, 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

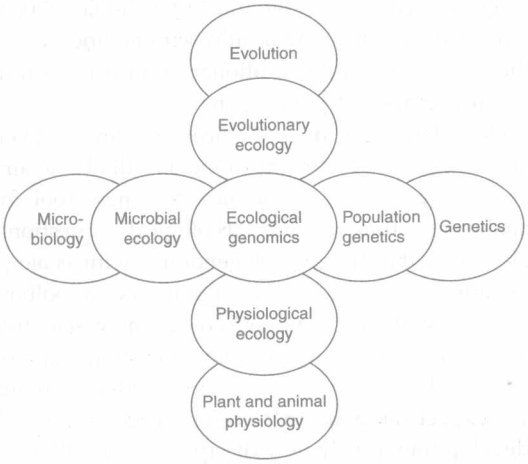

Figure 1.1 The position of ecological genomics in the middle of the other life-science disciplines with which it interacts most intensively.

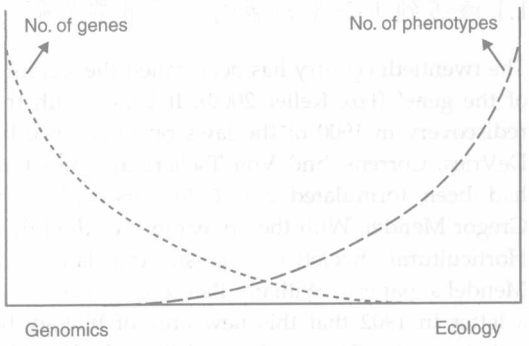

Figure 1.2 The playing field of ecological genomics, in between genomics, with its focus on the single genome of a model organism, studying all the genes that it contains, and ecology, studying a few genes in many species.

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 2.3) 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.

The interactions between the genes within the genome and the dynamic character of the genome on an evolutionary scale have been sketched vividly by Dover (1999) as an *internal tangled bank*. This idea goes back to Darwin (1859) who, after investigating the banks of hollow roads in the English countryside, was intrigued by the great variety of organisms tangled together:

> 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...

Darwin considered the way in which all organisms depended on each other as the template for evolution. Inspired by Darwin, Dover (1999) made a distinction between the 'external tangled bank' (the ecology) and the 'internal tangled bank' (the genome), attributing to them complementary roles in the evolutionary process (Fig. 1.3). 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. This agrees with François Jacob's famous description of 'evolution through tinkering' (Jacob 1977). It should not surprise us that genetic turbulence leaves many traces in the genome that do not have

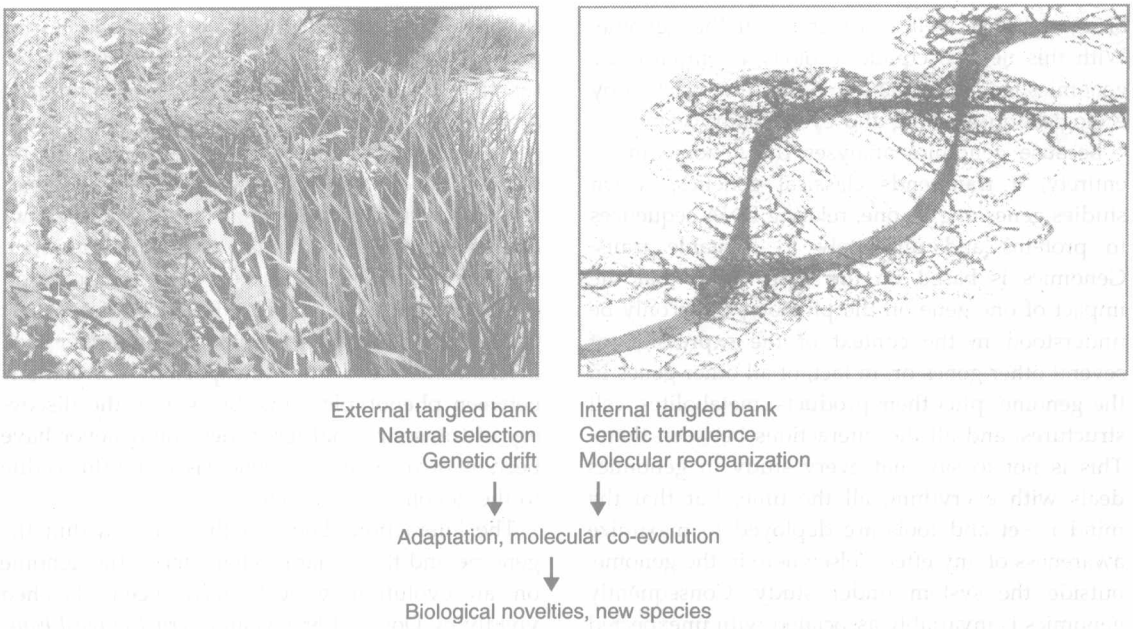

Figure 1.3 Evolution viewed as an interplay between the two 'tanged banks' of genetic turbulence and natural selection. Modified after Dover (1999), by permission of Oxford University Press.

direct negative phenotypic consequences; these traces from the past provide a valuable historical record for genome investigators to discover.

1.2 酵母、果蝇、线虫和拟南芥

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 3. 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 Chapter 2). 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 Eucarya.

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. 1.4). Many commentators have

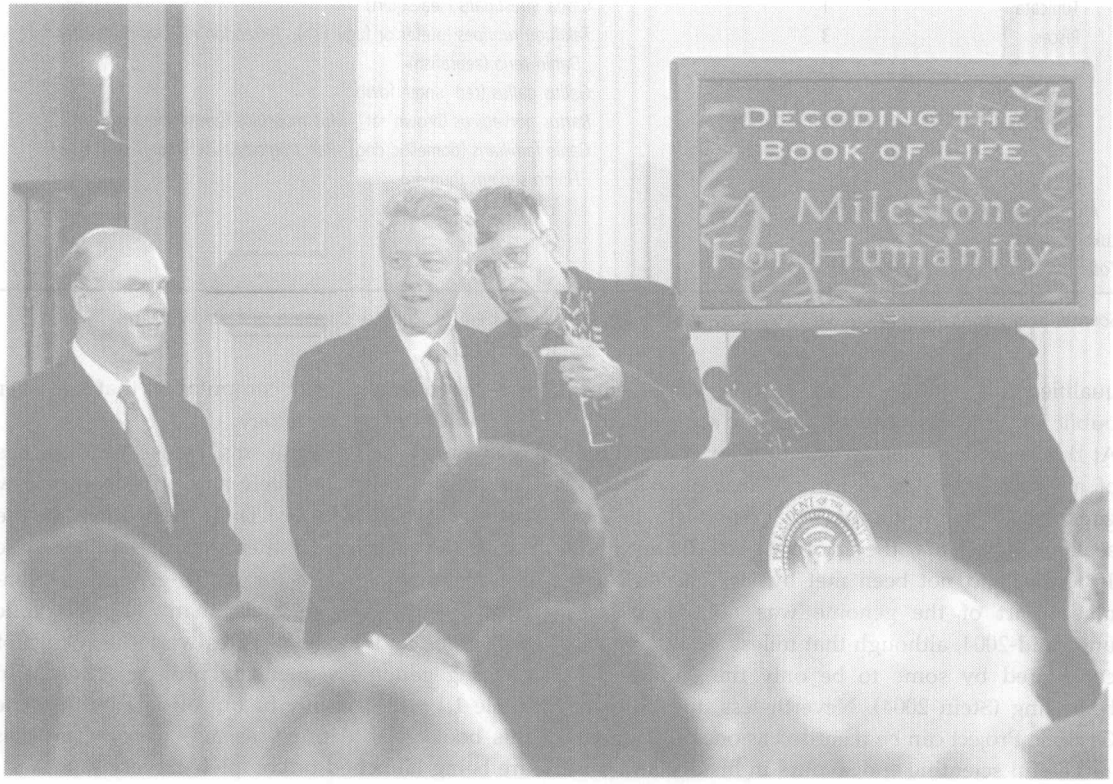

Figure 1.4 From left to right: J. Craig Venter (Celera Genomics), President Clinton, and Francis Collins (National Institutes of Health) on the historic announcement of 26 June 2000 of the completion of a working draft of the human genome. © Win McNamee/Reuters.

Table 1.1 List of complete and published genomes (not including viruses) by June 2005

Taxonomic group	No. of genomes	Remarks on species
Bacteria *total*	211	Many common laboratory models and pathogens
Archaea *total*	21	Several methanogens and extremophiles
Eukarya*		
Myxomycota	1	*Dictyostelium discoideum* (slime mould)
Entamoeba	1	*Entamoeba histolytica* (amoeba causing dysentery)
Apicomplexa	6	Four *Plasmodium* and two *Microsporidium* species
Kinetoplastida	2	*Trypanosoma brucei, Leishmania tropica* (parasites)
Cryptomonadina	1	*Guillardia theta* (flagellated unicellular alga)
Bacillariophyta	1	*Thalassiosira pseudonana* (marine diatom)
Rhodophyta	1	*Cyanidioschyzon merolae* (small unicellular red alga)
Plants	4	*Chlamydomonas reinhardtii* (green alga), *Populus trichocarpa* (black cottonwood), *Arabidopsis thaliana* (thale cress), *Oryza sativa* var. *japponica*, var. *indica* (rice)
Fungi	14	Including *Saccharomyces cerevisiae* (baker's yeast)
Animals		
Nematoda	2	*Caenorhabditis elegans* (free-living roundworm), *Caenorhabditis briggsae*
Insecta	4	*Bombyx mori* (silk worm), *Drosophila melanogaster* (fruit fly), *Anopheles gambiae* (mosquito, malaria vector), *Apis mellifera* (honey bee)
Tunicata	1	*Ciona intestinalis* (sea squirt)
Pisces	3	*Takifugu rubripes* (puffer or fugu fish), *Tetraodon nigroviridis* (puffer fish), *Danio rerio* (zebrafish)
Aves	1	*Gallus gallus* (red jungle fowl)
Mammalia	5	*Rattus norvegicus* (brown rat), *Mus musculus* (house mouse), *Canis familiaris* (domestic dog), *Pan troglodytes* (chimpanzee), *Homo sapiens* (human)
Animals: *total*	16	
Eukarya: *total*	47	
Total	279	

Sources: from www.genomesonline.org, genomenewsnetwork.org, GenBank Nucleotide Sequence Database, and sundry sources.

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 met by far. 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.

The number of organisms whose genome has been sequenced completely and published is now approaching 300 (Table 1.1). Bacteria dominate the list, as the small size of their genomes makes these organisms well-suited for whole-genome sequencing. By June 2005, no fewer than 730 prokaryotic organisms and 496 eukaryotes were the subject of ongoing genome sequencing projects. The list in Table 1.1 will certainly be out of date by the time this book goes to press, as new genome projects are being launched or completed every month.

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 absence of reptiles, amphibians, molluscs, and annelids is striking, as also is the scarcity of birds and arthropods other than the insects. How did 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 proven their usefulness as models before the genomics revolution and were adopted by genomicists because so much was known about their genetics and biochemistry, and, perhaps just as important, 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 (1 000 000 Mbp) in some protists, ferns, and amphibians. The puffer fish, *Takifugu rubripes*, was indeed chosen because of its relatively small genome (one-eighth of the human genome).

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, downregulated, 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). 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 (see the many parasites in Table 1.1). 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).

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, *Ci. 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. *Me. 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*, formerly regarded as a primitive insect, but now suggested to have developed the hexapod bodyplan before the insects separated from the crustaceans (Nardi *et al.* 2003).

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 3), 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. cerevisae*.

Ecological significance. It will be clear that ecological arguments have only played a minor role in the selection of species for whole-genome sequencing, but 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 slightly adapted form.

Biodiversity. 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 genomic knowledge of primary production, new models should be chosen among plants utilizing C4 photosynthesis or crassulacean acid metabolism (CAM). Considering the diversity of life histories, species differing in their mode of reproduction and dispersal capacity should be chosen; for example, hermaphoditism versus gonochorism, parthenogenesis versus bisexual reproduction, etc.

Ecological interactions. Species that take part in critical ecological interactions (mutualisms, antagonisms) are obvious candidates for genomic analysis. One may think of mycorrhizae, nitrogen-fixing symbionts, pollinators, natural enemies of pests, parasites, etc. The most obvious strategy for analysing such interactions would be to sequence the genomes of the players involved and to try and understand interactions between them from mutualisms or antagonisms in gene expression.

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 size, age 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