Archaeal Genomes

Ronald M Atlas, *University of Louisville, Louisville, Kentucky, USA*
Daniel Drell, *US Department of Energy, Washington DC, USA*
Claire Fraser-Liggett, *The Institute of Genome Sciences, University of Maryland Medical School, Baltimore, MD, USA*

Advanced article

Article Contents

Online posting date: 30th April 2008

The genomes of 46 archaeal species have been fully sequenced and published. As of this writing (September, 2007), 60 more are in various stages of progress. Analyses of these genomes are providing many useful insights into the evolution and functioning of diverse archaea, helping to understand the diverse physiological properties of archaea and their relationships to bacteria and eukarya.

Introduction

The determination of nucleotide sequences of complete genomes – based upon construction, sequencing and assembly of gene libraries – has been or is being accomplished for 106 archaea. This genomic information is providing many useful insights into the evolution and functioning of living organisms. These genomic analyses support the view that the archaea represent an evolutionary lineage distinct from the bacteria and eukarya confirming the view advanced by Woese and Fox 30 years ago (Woese and Fox, 1977). **See also**: Archaea; Archaeal Cells; DNA Sequencing; Genome Sequence Analysis

Many (though not all) of the archaea are extremophiles and their genomes hold great potential for biotechnological applications (see Cavicchioli, 2007). The genome analyses are revealing the gene structure/function relationships that allow the archaea to survive and to function in diverse harsh habitats. It should be noted, however, that our knowledge of gene functions is evolving rapidly and that the annotations for each of the genomes that has been sequenced represents only the current state of knowledge concerning gene functions. For this reason, the numbers of genes and assignments into functional categories have been and continue to evolve from the first version of this Encyclopedia of the Life Sciences entry. **See also**: Extreme Thermophiles

Additionally, not all archaea are extremophiles. Recent work (Samuel and Gordon, 2006) shows that one of the dominant members of the human gut flora is an archaeon – *Methanobrevibacter smithii* – and all sorts of relatives are found in other mammalian species. The archaea are not as 'extreme' (in terms of where they live) as earlier ideas might have suggested.

Many archaea are marine (DeLong, 2003). Archaea of both Crenarchaeota and Euryarchaeota divisions have been reported, some of which inhabit cold environmental niches (e.g. winter surface waters off Antarctica where planktonic Crenarchaeota can account for up to 20% of total microbial rRNA and even a higher representation has been seen in the North Pacific Ocean Gyre at depths from 200 to 5000 m; see DeLong (2003)). The smallest archaeal genome (so far) appears to belong to *Nanoarchaeum equitans*, an obligate symbiont with the chrenarchaeon *Ignicoccus*. *N. equitans* has a genome of only 490 885 bp and encodes proteins that carry out information processing and repair but little else. Ninety-five per cent of its DNA is coding. From its obligate symbiotic (e.g. parasitic) relationship with *Ignicoccus*, it is speculated that the *N. equitans* genome may represent a basal archael lineage (Waters *et al.*, 2003).

Genome sequencing of microbes and other organisms has been (and is being) carried out at high throughput sequencing centres, among them the US Department of Energy's Joint Genome Institute (http://www.jgi.doe.gov), the J. Craig Venter Institute (http://www.venterinstitute.org/jtc/, which recently incorporated The Institute for Genomic Research (TIGR, http://www.tigr.org)), the Wellcome Trust's Sanger Center (http://www.sanger.ac.uk/), the Harvard-MIT Broad Institute (http://www.broad.mit.edu/), the Washington University (St. Louis) Genome Sequencing Center (http://www.genome.wustl.edu) as well as others, and both the number of sequencing centres around the world and their overall throughput are increasing. The current pace of genome sequencing is revealing new genes (Venter *et al.*, 2004) at a rate much faster than annotation efforts can match.

Euryarchaeal Genomes

The euryarchaeota comprise a phylum of the archaea containing extremely halophilic organisms, methanogens and some extremely thermophilic aerobes and anaerobes. The

ELS subject area: Microbiology

How to cite:
Atlas, Ronald M; Drell, Daniel; and, Fraser-Liggett, Claire (April 2008) Archaeal Genomes. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0003653.pub2

Methanocaldococcus jannaschii DSM2661

	Genes	Per cent	Colour
Unknown function	31	2.421875	
Conserved hypothetical	534	41.71875	
Protein fate	38	2.96875	
Cell envelope	25	1.953125	
Transcription	21	1.640625	
DNA metabolism	54	4.21875	
Unclassified	0	0	
Energy metabolism	161	12.57813	
Protein synthesis	117	9.140625	
Cellular processes	26	2.03125	
Regulatory functions	19	1.484375	
Amino acid biosynthesis	65	5.078125	
Transport and binding proteins	57	4.453125	
Central intermediary metabolism	20	1.5625	
Fatty acid and phospholipid metabolism	9	0.703125	
Purines, pyrimidines, nucleosides and nucleotides	38	2.96875	
Biosynthesis of cofactors, prosthetic groups and carriers	53	4.140625	
Mobile and extrachromosomal element functions	12	0.9375	
Signal transduction	0	0	
	1280	100	

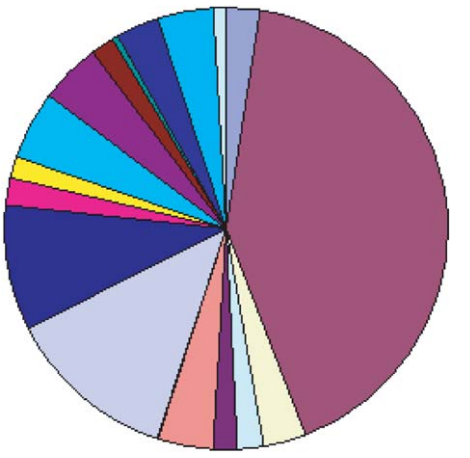

Figure 1 Functional distribution of genes for the archaeon *Methanocaldicoccus jannaschii* DSM2661.

euryarchaeota can be distinguished from the crenarchaeota (see later) on the basis of ribosomal ribonucleic acid (rRNA) sequences but there are other differences as well, among them cell division machinery, deoxyribonucleic acid (DNA) polymerases and histones. The genomes of methane-producing and sulfur-metabolizing euryarchaea that have been sequenced indicate substantial conservation of genes within this kingdom. Genes involved in metabolism are most similar to bacterial genes and those involved in gene expression and replication are closest to those of eukaryotes. The genome sequences also reveal evolutionary divergences that account for the specialized metabolic capabilities of the methanogens and sulfur-metabolizing euryarchaea. To date, 74 sequencing projects are in various stages of completion, with 32 being completed. (http://www.genomesonline.org/gold.cgi) Representative examples are given below. **See also**: Archaeal Chromosome; Euryarchaeota; Methanogenesis: Ecology

Methanocaldococcus jannaschii DSM2661 genome

The complete 1.66-mbp genome sequence of the autotrophic archaeon *Methanocaldococcus jannaschii* DSM2661 and its 58- and 16-kbp extrachromosomal elements have been determined by whole-genome random sequencing (**Figure 1**; Bult *et al*., 1996). A total of 1280 predicted protein-coding genes have been identified; however, only 56% have been assigned a putative cellular role. Although the majority of genes related to energy production, cell division and metabolism in *Methanocaldococcus jannaschii* are most similar to those found in bacteria, most of the genes involved in transcription, translation and replication in *Methanocaldococcus jannaschii* are closer to those found in eukaryotes. **See also**: Genome Mapping

Methanothermobacter thermautotrophicus Delta H genome

The 1 751 377-bp genome of the thermophilic archaeon *Methanothermobacter thermautotrophicus* Delta H has 2164 open reading frames (ORFs), 1696 (78%) of which have been assigned putative functions based on their similarities to database sequences with assigned functions (**Figure 2**; Smith *et al*., 1997). A total of 377 (17%) of the ORFs are related to sequences with unknown functions, 330 (15%) have little or no homology to identified sequences ('unclassified') and some 1013 of the putative gene

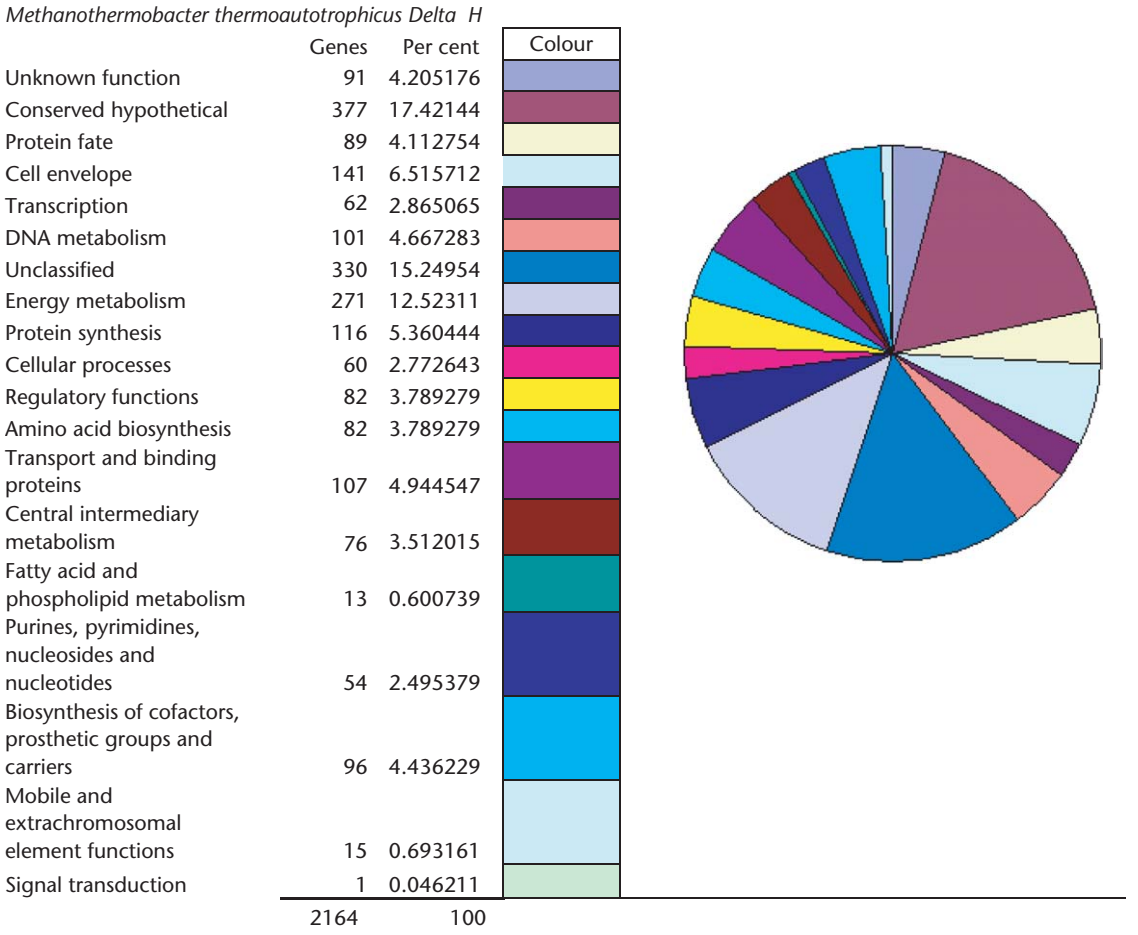

Figure 2 Functional distribution of genes for the archaeon *Methanothermobacter thermautotrophicus* Delta H.

products (54%) are most similar to polypeptide sequences described previously for other organisms in the domain Archaea. Comparisons with the *Methanocaldococcus jannaschii* genome show extensive divergence between these two methanogens; approximately 352 (19%) of *Methanothermobacter thermautotrophicus* ORFs encode sequences with >50% homology to *Methanocaldococcus jannaschii* polypeptides, and there is little conservation in the relative locations of orthologous genes. When the *Methanothermobacter thermautotrophicus* ORFs are compared with sequences from only the eukaryal and bacterial domains, about 42% are more similar to bacterial sequences and approximately 13% are more similar to eukaryal sequences. The bacterial domain-like gene products include the majority of those predicted to be involved in cofactor and small molecule biosyntheses, intermediary metabolism, transport, nitrogen fixation, regulatory functions and interactions with the environment. Most proteins predicted to be involved in DNA metabolism, transcription and translation are more similar to eukaryal sequences.

See also: Bacterial Genomes

Archaeoglobus fulgidus DSM4304 genome

Ar. fulgidus DSM4304 was the first sulfur-metabolizing organism to have its genome sequence determined. Its genome of 2 178 400 bp contains 1809 ORFs (**Figure 3**; Klenk *et al.*, 1997). The information-processing systems and the biosynthetic pathways for essential components (nucleotides, amino acids and cofactors) have extensive correlation with their counterparts in the archaeon *Methanocaldococcus jannaschii*, but in contrast to *Methanocaldococcus jannaschii*, *Ar. fulgidus* has fewer restriction-modification systems, and none of its genes appears to contain inteins (genetic elements coded by sequences internal to other expressed genes, but which excise themselves during posttranslational protein processing; inteins often have endonuclease activity and seemingly behave as parasitic genetic elements. For a review, see Gogarten *et al.*, (2002)). A third (658 ORFs) of the *Ar. fulgidus* genome encodes functionally uncharacterized (yet conserved) proteins, two-thirds of which are shared with *Methanocaldococcus jannaschii* (428 ORFs). Another

Archaeoglobus fulgidus DSM4304

	Genes	Per cent	Colour
Unknown function	50	2.76	
Conserved hypothetical	658	36.37	
Protein fate	53	2.93	
Cell envelope	36	1.99	
Transcription	27	1.49	
DNA metabolism	42	2.32	
Unclassified	0	0.00	
Energy metabolism	242	13.38	
Protein synthesis	114	6.30	
Cellular processes	48	2.65	
Regulatory functions	77	4.26	
Amino acid biosynthesis	76	4.20	
Transport and binding proteins	129	7.13	
Central intermediary metabolism	28	1.55	
Fatty acid and phospholipid metabolism	79	4.37	
Purines, pyrimidines, nucleosides and nucleotides	40	2.21	
Biosynthesis of cofactors, prosthetic groups and carriers	64	3.54	
Mobile and extrachromosomal element functions	46	2.54	
Signal transduction	0	0.00	
	1809	100	

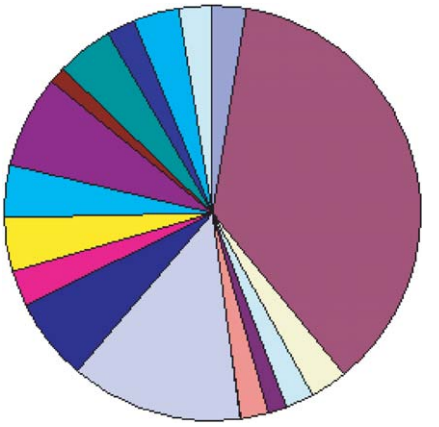

Figure 3 Functional distribution of genes for the archaeon *Archaeoglobus fulgidus* DSM4304.

quarter of the *Ar. fulgidus* genome encodes apparently unique proteins indicating substantial archaeal gene diversity.

Pyrococcus horikoshii OT3 genome

The 1 738 505-bp genome of the extreme thermophile *P. horikoshii* OT3 exhibits a total of 1401 ORFs, 727 (52%) of which are related to genes with putative function, 640 (46%) to sequences resembling sequences in other organisms but with no known function and 163 (12%) unrelated to any previously sequenced genes (**Figure 4**; Kawarabayasi *et al.*, 1998). A considerable number of ORFs appear to have been generated by sequence duplication.

Halobacterium sp. NRC-1 genome

The 2 571 010-bp genome of the halophile *Halobacterium* sp. NRC-1 reveals 2630 predicted ORFs, 36% of which do not resemble previously reported proteins. NRC-1 is aerobic, mesophilic and exhibits optimal growth at salinities 10 × over seawater. The large number of transposable insertion sequences in the genome is thought to help explain the genomic plasticity observed in halophiles (Lateral Gene Transfer, a topic beyond the scope of this chapter). The genome shows the presence of pathways for uptake and utilization of amino acids, photosensory and signal transduction systems and DNA management systems resembling those in eukaryotic organisms (see Ng *et al.*, 2000; Gan *et al.*, 2006).

Crenarchaeal Genomes

The crenarchaea also constitute a major division within the archaea and include microbial species with some of the highest known growth temperatures yet observed. Interestingly, the crenarchaeota may comprise the majority of marine archaea (Madigan and Martinko, 2005). The genomes of 32 crenarchaea that have been sequenced indicate a high degree of gene duplication during their evolution. The genes encoding the metabolic capabilities of the crenarchaea provide the physiological capabilities for living under conditions of extremely high temperatures in geothermal regions. To date, 34 sequencing projects are in various stages of completion, with six completed and published (http://www.genomesonline.org/gold.cgi). Representative examples are given below.

See also: Crenarchaeota; Extreme Thermophiles; Microorganisms in High-temperature Sulfur Environments

Pyrococcus horikoshii OT3

	Genes	Per cent	Colour
Unknown function	34	2.426838	
Conserved hypothetical	640	45.68166	
Protein fate	40	2.855103	
Cell envelope	25	1.78444	
Transcription	19	1.356174	
DNA metabolism	38	2.712348	
Unclassified	129	9.207709	
Energy metabolism	108	7.708779	
Protein synthesis	90	6.423983	
Cellular processes	47	3.354747	
Regulatory functions	25	1.78444	
Amino acid biosynthesis	42	2.997859	
Transport and binding proteins	61	4.354033	
Central intermediary metabolism	27	1.927195	
Fatty acid and phospholipid metabolism	6	0.428266	
Purines, pyrimidines, nucleosides and nucleotides	39	2.783726	
Biosynthesis of cofactors, prosthetic groups and carriers	31	2.212705	
Mobile and extrachromosomal element functions	0	0	
Signal transduction	0	0	
	1401	100	

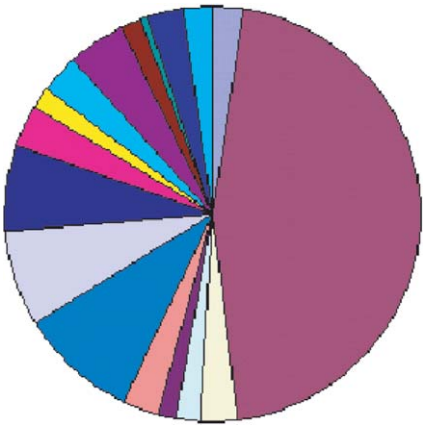

Figure 4 Functional distribution of genes for the archaeon *Pyrococcus horikoshii* (shinkaj) OT3.

Aeropyrum pernix genome

The complete sequence of the 1 669 695-bp genome of the aerobic hyperthermophilic crenarchaeon *Ae. pernix* K1, which optimally grows at 95°C, has been determined by the whole genome shotgun method with some modifications (**Figure 5**; Kawarabayasi *et al*., 1999). As potential protein-coding regions, a total of 1153 ORFs have been assigned, having a somewhat higher gene density than is usually seen in the bacteria. All the genes in the tricarboxylic acid (TCA) cycle except for that of α-ketoglutarate dehydrogenase were included, and instead of the α-ketoglutarate dehydrogenase gene, the genes coding for the two subunits of 2-oxoacid: ferredoxin oxidoreductase have been identified. A set of 213 ORFs (18%) did not show any significant similarity to previously reported ORFs and 448 (39%) were conserved hypotheticals. Sequence comparison among the assigned ORFs suggests that a considerable number have been generated by sequence duplication. **See also**: Citric Acid Cycle; Genome Sequence Analysis

Sulfolobus acidocaldarius genome

The complete sequence of the 2 225 959-bp genome of the aerobic thermoacidophilic crenarchaeon *S. acidocaldarius* strain DSM639, which optimally grows at 80°C and pH 2, has been determined by the whole genome shotgun method with some modifications (Chen, *et al*., 2005). As potential protein-coding regions, a total of 2292 ORFs have been assigned, a number based on comparisons with two related *Sulfolobus* sequences. Many of the smaller genes were identified for the first time on the basis of comparison of three *Sulfolobus* genome sequences. Of the protein-coding genes, 305 are exclusive to *S. acidocaldarius* and 866 are specific to the *Sulfolobus* genus. In addition, *S. acidocaldarius* contains genes for a characteristic restriction-modification system, a UV damage excision repair system, thermopsin and an aromatic ring dioxygenase, all of which are absent from genomes of other *Sulfolobus* species. However, it lacks genes for some of their sugar transporters, consistent with it growing on a more limited range of carbon sources.

In the previous version of this Encyclopedia entry, a table was included listing the archaeal sequences then completed. Since that time, both the number of sequences completed and the rate of appearance of new sequences have grown to the point where it is impractical to list them all in a table that would soon be obsolete. Numerous web sites catalogue microbial genome sequence information, among them:

Genomes Online: http://www.genomesonline.org/
NCBI prokaryotes (including 29 archaea): http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi
DOE Joint Genome Institute Microbes Online: http://www.microbesonline.org/

Aeropyrum pernix K1			
Gene role	Genes	Per cent	Colour
Unknown function	23	1.99	
Conserved hypothetical	448	38.86	
Protein fate	27	2.34	
Cell envelope	10	0.87	
Transcription	18	1.56	
DNA metabolism	25	2.17	
Unclassified	190	16.48	
Energy metabolism	94	8.15	
Protein synthesis	95	8.24	
Cellular processes	19	1.65	
Regulatory functions	5	0.43	
Amino acid biosynthesis	34	2.95	
Transport and binding proteins	75	6.50	
Central intermediary metabolism	19	1.65	
Fatty acid and phospholipid metabolism	18	1.56	
Purines, pyrimidines, nucleosides and nucleotides	22	1.91	
Biosynthesis of cofactors, prosthetic groups and carriers	30	2.60	
Mobile and extrachromosomal element functions	1	0.09	
Signal transduction	0	0	
	1153	100	

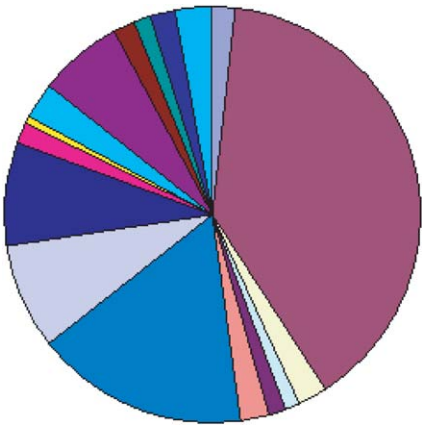

Figure 5 Functional distribution of genes for the archaeon *Aeropyrum pernix* K1.

DOE Joint Genome Institute Integrated Microbial Genomes: http://img.jgi.doe.gov/cgi-bin/pub/main.cgi
TIGR Comprehensive Microbial Resource: http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi
The Wellcome Trust Sanger Institute: http://www.sanger.ac.uk/Info/Statistics/
Broad Institute Microbial Sequencing Center: http://www.broad.mit.edu/seq/msc/
Venter Institute Microbial Sequencing: https://research.venterinstitute.org/moore/
UCSD CAMERA Cyberinfrastructure Database: http://camera.calit2.net
Washington University St. Louis Genome Sequencing Center, Microbes Group: http://www.genome.wustl.edu/sub_genome_group_index.cgi? GROUP = 3

This is far from a comprehensive list and includes very little private-sector sequencing data, much of which is proprietary.

References

Bult C, White O, Olsen G *et al*. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**: 1058–1073.

Cavicchioli R (ed.) (2007) *Archaea: Molecular and Cellular Biology*. Washington DC: ASM Press.

Chen L, Brugger K, Skovgaard M *et al*. (2005) The genome of *Sulfolobus acidocalderius*, a model organism of the Crenarchaeota. *Journal of Bacteriology* **187**: 4992–4997.

DeLong E (2003) Oceans of archaea. *ASM News* **69**: 503–511.

Gan RR, Yi EC, Chiu Y *et al*. (2006) Proteome analysis of *Halobacterium* sp. NRC-1 facilitated by the biomodule analysis tool BMSorter. *Molecular and Cellular Proteomics* **5**: 987–997.

Gogarten JP, Senejani AG, Zhaxybayeva O, Olendzenski L and Hilario E (2002) Inteins: structure, function, and evolution. *Annual Reviews of Microbiology* **56**: 263–287.

Kawarabayasi Y, Hino Y, Horikawa H *et al*. (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Research* **6**: 145–152.

Kawarabayasi Y, Sawada M, Horikawa H *et al*. (1998) Complete sequence and gene organization of the genome of a hyperthermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Research* **5**: 55–76.

Klenk HP, Clayton RA, Tomb JF *et al*. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**: 364–370.

Madigan M and Martinko J (eds) (2005) *Brock Biology of Microorganisms*, 11th edn. Upper Saddle River, NJ: Prentice Hall.

Ng WV, Kennedy SP, Mahairas GG *et al*. (2000) Genome sequence of *Halobacterium* species NRC-1. *Proceedings of the National Academy of Sciences of the USA* **97**: 12176–12181.

Samuel BS and Gordon JL (2006) A human gnotobiotic mouse model of host-archaeal-bacterial mutualism. *Proceedings of the National Academy of Sciences of the USA* **103**: 10011–10016.

Smith DR, Doucette-Stamm LA, Deloughery C *et al*. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* delta H: functional analysis and comparative genomics. *Journal of Bacteriology* **179**: 7135–7155.

Venter JC, Remington K, Heidelberg JF *et al*., (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Waters E, Hohn MJ, Ahel I *et al*. (2003) The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. *Proceedings of the National Academy of Sciences of the USA* **100**: 12984–12988.

Woese CR and Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the USA* **74**: 5088–5090.

Further Reading

Dixon B (1994) *Power Unseen: How Microbes Rule the World*. Oxford, England: W.H. Freeman & Co.

Fraser CM, Read T and Nelson K (eds) (2004) *Microbial Genomes* Totowa, NJ: Humana Press.

Fraser-Ligett CM (2005) Insights on biology and evolution from microbial genome sequencing. *Genome Research* **15**(12): 1603–1610.

Handelsman J and Tiedje J (co-chairs, Committee on Metagenomics) (2007) *The New Science of Metagenomics: Revealing the Secrets of our Microbial Planet*. Washigton, DC: National Academies Press.

Tringe SG, von Mering C, Kobayashi A *et al*. (2005) Comparative Metagenomics of microbial communities. *Science* **308**(5721): 554–557.