Bacterial Genomes

Secondary article

Ronald M Atlas, *University of Louisville, Louisville, Kentucky, USA*

Daniel Drell, *US Department of Energy, Washington DC, USA*

Claire Fraser, *The Institute for Genomic Research, Rockville, Maryland, USA*

The genomes of a number of bacterial species have been fully sequenced. Analyses of these genomes are providing useful insights into the evolution and functioning of diverse bacteria and bacterial pathogens.

Introduction

The determination of nucleotide sequences of complete genomes, based upon construction, sequencing and assembly of gene libraries, has been accomplished for a number of bacteria (**Table 1**). This genomic information is providing useful insights into the evolution and functioning of living organisms. Most sequenced bacterial genomes appear to possess about 1000 open reading frames (ORFs) per megabase of DNA – plus or minus 15%.

Genomes of Evolutionarily Diverse Bacteria

Bacterial genome analyses are providing new evolutionary perspectives and enhancing our understanding of the ecology of bacteria. From the evolutionary perspective, short regions of conservation of gene order of clusters of ribosomal proteins appear to have been maintained during the evolutionary divergence of the bacteria, supporting the credibility of using ribosomal RNA (rRNA) analyses for phylogenetic relationships and the exploration of vertical evolutionary change. However, there also appear among the bacteria significant instances of horizontal gene transfers during evolution. Regardless of ecological niche or phylogenetic position, each bacterial genome has a significant proportion (typically greater than 30% and frequently approaching 50%) of unique genes or genes of unknown functions. Bacteria in varied environments exhibit a greater diversity of genes. It should be noted, however, that the status of the annotation is rapidly evolving – the annotations for each of the genomes that has been sequenced represents only the current state of knowledge concerning gene functions.

Aquifex aeolicus

The 1 551 335-bp genome of *Aquifex aeolicus* encodes the metabolic pathways needed for this bacterium to grow as a chemoautotroph on hydrogen, oxygen, carbon dioxide and mineral salts (**Figure 1**) (Deckert *et al.*, 1998). Metabolic flexibility seems to be reduced as a result of the limited genome size. The use of oxygen (albeit at very low concentrations) as an electron acceptor is allowed by the presence of a complex respiratory apparatus. Although *Aquifex aeolicus* is an extreme thermophile, few specific indications of thermophily have been identified within the genome of this bacterium, which was among the first lines of evolutionary divergence from the archaea.

Thermotoga maritima

The 1 860 725-bp genome of *Thermotoga maritima* MSB8 contains 1877 predicted coding regions, 1014 (54%) of which have functional assignments and 863 (46%) of which are of unknown function (**Figure 2**) (Nelson *et al.*, 1999). Genome analysis reveals numerous pathways involved in degradation of sugars and plant polysaccharides, and 108 genes that have orthologues only in the genomes of other thermophilic bacteria and archaea. Of the bacteria sequenced to date, *T. maritima* has the highest percentage (24%) of genes that are most similar to archaeal genes. Eighty-one archaeal-like genes are clustered in 15 regions of the *T. maritima* genome that range in size from 4 to 20 kb. Conservation of gene order between *T. maritima* and archaea in many of the clustered regions suggests that lateral gene transfer may have occurred between thermophilic bacteria and archaea.

Synechocystis sp.

The nucleotide sequence of the 3 573 470-bp genome of the cyanobacterium *Synechocystis* sp. strain PCC6803 was assembled from the sequences of the physical map-based contigs of cosmid clones and of lambda clones and using long PCR products for gap-filling (**Figure 3**) (Kaneko *et al.*, 1996). A total of 3168 potential protein genes have been assigned on the genome, in which 145 (4.6%) are identical to reported genes and 1426 (45.0%) have no apparent similarity to any previously reported genes. Among the potential protein genes assigned, 128 were related to the genes participating in photosynthetic reactions. A notable feature on the gene organization of the genome is that 99

Table 1 Completed and in progress bacterial genome sequencing

Genome	Size (Mb)	Institution	Publication	Relevance
Actinobacillus actinomycetemcomitans HK1651	2.2	University of Oklahoma		Human pathogen
Agrobacterium tumefaciens C58	5.3	University of Washington/Dupont		Plant pathogen causing crown gall, high gene transfer capability
Aquifex aeolicus VF5	1.50	Diversa	Deckert *et al*., 1998	Extreme thermophile, potential for identifying high-temperature enzymes
Bacillus anthracis Ames	4.5	TIGR		Animal and human pathogen, potential biological threat agent
Bacillus halodurans C-125	4.2	Japan Marine Science and Technology Center		Salt-tolerant endospore-forming Gram-positive bacterium
Bacillus stearothermophilus 10		University of Oklahoma		Thermophilic endospore-forming Gram-positive bacterium
Bacillus subtilis 168	4.20	International Consortium	Kunst *et.al*., 1997	Very widely studied Gram-positive bacterium
Bartonella henselae Houston 1	2.00	University of Uppsala		Human pathogen
Bordetella bronchiseptica RB50	4.9	Sanger Centre		Human pathogen
Bordetella parapertussis	3.9	Sanger Centre		Human pathogen
Bordetella pertussis Tohama I	3.88	Sanger Centre		Human pathogen that causes whooping cough
Borrelia burgdorferi B31	1.44	TIGR	Fraser *et al*., 1997; Casjens *et al*., 2000	Human pathogen that causes Lyme disease
Buchnera sp. APS	0.64	University of Tokyo/RIKEN		Endosymbiont
Burkholderia pseudomallei K96243	6.0	Sanger Centre/DERA/Public Health Laboratory Beowulf Genomics		Animal and human pathogen, potential biothreat agent
Campylobacter jejuni NCTC 11168	1.64	Sanger Centre	Parkhill *et al*., 2000	Human pathogen, frequent cause of gastroenteritis
Caulobacter crescentus	3.80	TIGR		Potential for heavy-metal remediation in waste-treatment plant wastewater
Chlamydia pneumoniae	1.23	Genset		Human pathogen, intracellular parasite
Chlamydia pneumoniae AR39	1.23	TIGR	Read *et al*., 2000	Human pathogen, intracellular parasite
Chlamydia pneumoniae CWL029	1.23	UC Berkeley & Stanford	Kalman *et al*., 1999	Human pathogen, intracellular parasite
Chlamydia psittaci GPIC	1.2	TIGR		Human pathogen, intracellular parasite
Chlamydia trachomatis L2	1.038	Genset		Human pathogen, intracellular parasite
Chlamydia trachomatis MoPn	1.07	TIGR	Read *et al*., 2000	Human pathogen, intracellular parasite
Chlamydia trachomatis serovar D(D/UW-3/Cx)	1.05	UC Berkeley & Stanford	Stephens *et al*., 1998	Human pathogen, intracellular parasite

continued

Table 1 – *continued*

Genome	Size (Mb)	Institution	Publication	Relevance
Lactobacillus acidophilus ATCC 700396	1.9	Environmental Biotechnology Institute Dairy Management, Inc./California Research Foundation/ Environmental Biotechnology Institute		Milk digestion inoculum
Lactococcus lactis IL1403	2.35	GENOSCOPE		Production of fermented dairy products
Legionella pneumophila Philadelphia-1	4.0	Columbia Genome Center		Human pathogen
Leptospira interrogans serovar icterohaemorrhagiae Lai	4.8	Chinese National Human Genome Center at Shanghai CNCBD/ Science and Technology Commission of Shanghai		Human pathogen
Listeria innocua Clip11262, rhamnose-negative	3.2	GMP		Human pathogen
Listeria monocytogenes EGD-e	2.94	EC Consortium		Human pathogen
Methylobacterium extorquens		University of Washington		Utilizes C-1 compounds
Methylococcus capsulatus	4.60	TIGR/University of Bergen, Norway		Uses methane as single carbon and energy source; generates pollutant-oxidizing enzymes; used commercially to produce biomass and other proteins
Mycobacterium avium 104	4.70	TIGR		Emerging human pathogen, causes tuberculosis in immunocompromised individuals
Mycobacterium bovis AF2122/97	4.4	Sanger Centre/Institut Pasteur		Animal and human pathogen, causes tuberculosis
Mycobacterium leprae	2.80	Sanger Centre		Human pathogen, causes leprosy
Mycobacterium paratuberculosis K-10	5.00	University of Minnesota		Human pathogen
Mycobacterium tuberculosis CSU#93 (clinical isolate)	4.40	TIGR		Human pathogen, causes tuberculosis
Mycobacterium tuberculosis H37Rv (lab strain)	4.40	Sanger Centre	Cole *et al.*, 1998	Human pathogen, causes tuberculosis
Mycoplasma genitalium G-37	0.58	TIGR	Fraser *et al.*, 1995; Hutchison *et al.*, 1999	Human pathogen; serves as model for minimal set of genes sufficient for free living
Mycoplasma hyopneumoniae 232	0.89	University of Washington		Human pathogen
Mycoplasma mycoides subsp. *mycoides* SC PG1	1.28	The Royal Institute of Technology, Stockholm & The National Veterinary Institute, Uppsala		Human pathogen

continued

Table 1 – *continued*

Genome	Size (Mb)	Institution	Publication	Relevance
Mycoplasma pneumoniae M129	0.81	University of Heidelberg	Himmelreich *et al*., 1996	Human pathogen
Mycoplasma pulmonis	0.95	GENOSCOPE		Human pathogen
Neisseria gonorrhoeae	2.20	University of Oklahoma		Human pathogen, causes gonorrhoea
Neisseria meningitidis MC58	2.27	TIGR	Tettelin *et al*., 2000	Human pathogen
Neisseria meningitidis serogroup A strain Z2491	2.18	Sanger Centre	Parkhill *et al*., 2000	Human pathogen
Neisseria meningitidis Serogroup C strain FAM18	2.2	Sanger Centre		Human pathogen
Nitrosomonas europaea	2.2	JGI		Important in soil nitrogen cycling and ammonia oxidation; degrades chlorinated hydrocarbons; aids incorporation of carbon dioxide into biomass
Nostoc punctiforme ATCC 29133	10	JGI		Fixes carbon dioxide and nitrogen; produces hydrogen; survives acidic, anaerobic, and low-temperature conditions
Pasteurella haemolytica	2.4	LION Bioscience		Animal and human pathogen
Pasteurella multocida Pm70	2.4	University of Minnesota		Animal and human pathogen
Photorhabdus luminescens TT01	5.5	GMP		Luminescent endosymbiont
Porphyromonas gingivalis W83	2.20	TIGR/ Forsyth Dental Center		Human pathogen in oral cavity
Prochlorococcus marinus MED4	2.00	JGI		Abundant in temperate and tropical oceans; important in ocean carbon cycling; absorb blue light efficiently
Pseudomonas aeruginosa PAO1	5.90	University of Washington PathoGenesis		Gram-negative bacterium with diverse metabolism, human pathogen
Pseudomonas putida	6.1	TIGR/German Consortium		High potential for bioremediation by reducing metal and pollutants
Ralstonia solanacearum		GENOSCOPE		Plant pathogen
Rhodobacter capsulatus SB1003	3.70	University of Chicago/Institute of Molecular Genetics		Photosynthetic energy metabolism
Rhodobacter sphaeroides 2.4.1	4.34	University of Texas – Houston Health Science Center		Photosynthetic energy metabolism
Rhodopseudomonas palustris	4.5	JGI		Fixes carbon dioxide; produces hydrogen; biodegrades under both aerobic and anaerobic conditions

continued

Table 1 – *continued*

Genome	Size (Mb)	Institution	Publication	Relevance
Rickettsia conorii	1.2	GENOSCOPE		Human pathgoen
Rickettsia prowazekii Madrid E	1.10	University of Uppsala	Andersson *et al*., 1998	Human pathogen
Salmonella enteritidis LK5	4.5	University of Illinois		Human pathogen
Salmonella paratyphi A ATCC 9150	4.60	Washington University Consortium		Human pathogen
Salmonella typhi	4.5	Sanger Centre		Human pathogen
Salmonella typhimurium SGSC1412	4.80	Washington University Consortium		Human pathogen
Salmonella typhimurium TR7095	4.50	Washington University Consortium		Human pathogen
Shewanella putrefaciens MR-1	4.50	TIGR		May degrade toxic organic wastes and sequester toxic metals
Shigella flexneri 2a 301	4.7	Microbial Genome Center		Human pathogen
Sinorhizobium meliloti 1021	6.6	European & Canadian Consortium/ Stanford University		Nitrogen fixation, symbiosis
Staphylococcus aureus COL	2.80	TIGR		Human pathogen
Staphylococcus aureus 8325	2.80	University of Oklahoma		Human pathogen
Staphylococcus aureus MRSA	2.8	Sanger Centre/Trinity College		Human pathogen, resistant to methicillin
Staphylococcus aureus MSSA	2.8	Sanger Centre/Trinity College WTCEID		Human pathogen
Staphylococcus epidermidis ATCC 12228	2.4	Chinese National Human Genome Center at Shanghai/ Shanghai Medical University		Gram-positive commensal
Streptococcus agalactiae ATCC 12403	2.00	GMP		Human pathogen
Streptococcus mutans UAB159	2.20	University of Oklahoma		Tooth decay
Streptococcus pneumoniae type 4	2.20	TIGR		Human pathogen
Streptococcus pneumoniae R6	2.04	Eli Lilly		Human pathogen
Streptococcus pyogenes M1	1.85	University of Oklahoma		Human pathogen
Streptococcus pyogenes Manfredo	1.98	Sanger Centre/ University of Newcastle		Human pathogen
Streptomyces coelicolor A3(2)	8.0	Sanger Centre/ John Innes Centre		Antibiotic production
Synechococcus spp.		JGI		Photosynthetic; important to ocean carbon fixation; genetically tractable
Synechocystis sp. PCC 6803	3.57	Kazusa DNA Research Institute	Kaneko *et al*., 1996	Photosynthetic

continued

Table 1 – *continued*

Genome	Size (Mb)	Institution	Publication	Relevance
Thermotoga maritima MSB8	1.80	TIGR	Nelson *et al.*, 1999	Potential for identifying high-temperature, high-pressure enzymes
Thermus thermophilus HB27	1.82	Goettingen Genomics Laboratory		Thermophile
Thiobacillus ferrooxidans ATCC 23270	2.90	TIGR		Used in mining industry to sequester iron and sulfide
Treponema denticola	3.00	TIGR/ University of Texas		Oral cavity human pathogen
Treponema pallidum Nichols	1.14	TIGR/ University of Texas	Fraser *et al.*, 1998	Human pathogen, causes syphilis
Ureaplasma urealyticum serovar 3	0.75	University of Alabama / PE-ABI		Human pathogen
Vibrio cholerae serotype O1, Biotype El Tor, strain N16961	4.0	TIGR	Heidelberg *et al.*, 2000	Human pathogen, causes cholera
Xanthomonas citri	5.3	Brazilian Consortium		Plant pathogen
Xylella fastidiosa 9a5c	2.68	ONSA Consortium	Simpson *et al.*, 2000	Plant pathogen
Yersinia pestis CO-92 Biovar Orientalis	4.38	Sanger Centre		Human pathogen, causes plague, potential biothreat agent

ORFs with similarity to transposase genes are spread all over the genome, implying that rearrangement of the genome has occurred frequently during or after establishment of this species.

Deinococcus radiodurans

The 3 284 156 bp genome of the radiation-resistant bacterium *Deinococcus radiodurans* R1 is composed of two chromosomes (2 648 638 and 412 348 bp), a megaplasmid (177 466 bp), and a small plasmid (45 704 bp), encoding a total of 3187 ORFs (White *et al.*, 1999). Multiple components distributed on the chromosomes and megaplasmid that contribute to the ability of *D. radiodurans* to survive under conditions of starvation, oxidative stress and high amounts of DNA damage were identified. *D. radiodurans* represents an organism in which all known systems for DNA repair, DNA damage export, desiccation and starvation recovery, and genetic redundancy are present in one cell. Given the 53% of the genome that codes for proteins of unknown functions, additional systems may await discovery and characterization.

Bacillus subtilis

Bacillus subtilis has a genome of 4 214 810 bp with 4100 protein-coding genes. Of these protein-coding genes, 53% are represented once, while a quarter of the genome corresponds to several gene families that have been greatly expanded by gene duplication, the largest family containing 77 putative ATP-binding transport proteins (Kunst *et al.*, 1997). In addition, a large proportion of the genetic capacity is devoted to the utilization of a variety of carbon sources, including many plant-derived molecules. The identification of five signal peptidase genes, as well as several genes for components of the secretion apparatus, is important given the capacity of *Bacillus* strains to secrete large amounts of industrially important enzymes. Many of the genes are involved in the synthesis of secondary metabolites, including antibiotics. The genome contains at least 10 prophages or remnants of prophages, indicating that bacteriophage infection has played an important evolutionary role in horizontal gene transfer.

Escherichia coli

The genome of *Escherichia coli* K-12, a nonpathogen, has a 4 639 221-bp sequence of circular duplex DNA (Blattner *et al.*, 1997). There are 4288 protein-coding annotated genes, 38% of which have no attributed function. Protein-coding genes account for 87.8% of the genome, 0.8% encodes stable RNAs, and 0.7% consists of noncoding repeats, leaving approximately 11% for regulatory and other functions. The proportion of uncharacterized ORFs (38%) is similar to the proportion of unassigned ORFs in

Gene role	Genes	Per cent	Colour
Unknown	37	3.15	⬛
Hypothetical	412	35.10	⬜
Protein fate	43	3.67	🟪
Cell envelope	53	4.52	🟩
Transcription	18	1.53	⬜
DNA metabolism	46	3.92	🟨
Other categories	2	0.17	⬜
Energy metabolism	135	11.50	🟩
Protein synthesis	103	8.79	🟪
Cellular processes	63	5.38	🟥
Regulatory functions	32	2.73	🟦
Amino acid biosynthesis	65	5.55	🟧
Transport and binding proteins	37	3.15	🟦
Central intermediary metabolism	12	1.02	🟫
Fatty acid and phospholipid metabolism	16	1.36	🟪
Purines, pyrimidines, nucleosides, and nucleotides	39	3.33	🟧
Biosynthesis of cofactors, prosthetic groups, and carriers	58	4.95	⬜
Number of genes	1171	100.00	

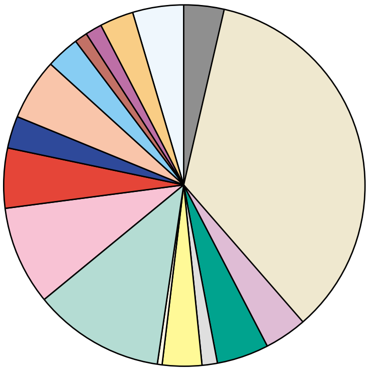

Figure 1 Functional distribution of genes for the bacterium *Aquifex aeolicus* VF5.

Gene role	Genes	Per cent	Colour
Unknown	83	5.50	⬛
Hypothetical	404	26.80	⬜
Protein fate	49	3.25	🟪
Cell envelope	73	4.84	🟩
Transcription	16	1.06	⬜
DNA metabolism	54	3.58	🟨
Other categories	13	0.86	⬜
Energy metabolism	195	12.90	🟩
Protein synthesis	105	6.96	🟪
Cellular processes	49	3.25	🟥
Regulatory functions	70	4.64	🟦
Amino acid biosynthesis	72	4.77	🟧
Transport and binding proteins	189	12.50	🟦
Central intermediary metabolism	44	2.91	🟫
Fatty acid and phospholipid metabolism	15	0.99	🟪
Purines, pyrimidines, nucleosides, and nucleotides	45	2.98	🟧
Biosynthesis of cofactors, prosthetic groups, and carriers	31	2.05	⬜
Number of genes	1507	100.00	

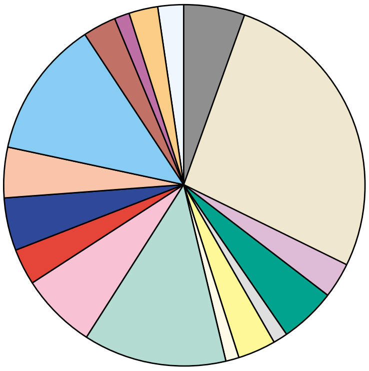

Figure 2 Functional distribution of genes for the bacterium *Thermotoga maritima* MSB8.

Gene role	Genes	Per cent	Colour
Unknown	38	2.05	⬛
Hypothetical	729	39.30	⬜
Protein fate	63	3.40	🟪
Cell envelope	67	3.61	🟩
Transcription	25	1.34	⬜
DNA metabolism	46	2.48	🟨
Other categories	72	3.88	⬜
Energy metabolism	221	11.90	🟩
Protein synthesis	96	5.18	🟪
Cellular processes	52	2.80	🟥
Regulatory functions	76	4.10	🟦
Amino acid biosynthesis	70	3.77	🟧
Transport and binding proteins	107	5.77	🟦
Central intermediary metabolism	50	2.69	🟫
Fatty acid and phospholipid metabolism	26	1.40	🟪
Purines, pyrimidines, nucleosides, and nucleotides	42	2.26	🟧
Biosynthesis of cofactors, prosthetic groups, and carriers	72	3.88	⬜
Number of genes	1852	100.00	

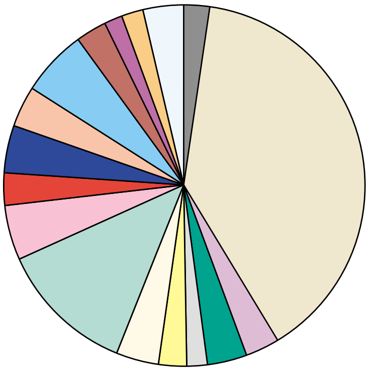

Figure 3 Functional distribution of genes for the bacterium *Synechocystis* sp. PCC 6803.

other sequenced bacterial genomes. The largest well-defined functional group consists of 281 transport and binding proteins, and there are an additional 146 putative transport and binding proteins. Of 2192 predicted operons, 73% have only one gene, 16.6% have two genes, 4.6% have three genes, and 6% have four or more genes. All of them have at least one promoter, either known or predicted. Of 2405 operon regions with predicted promoters, 68% contain one promoter, 20% contain two promoters, and 12% contain three or more promoters. There are a number of transposable elements that are implicated in the generation of many spontaneous mutations by a variety of mechanisms. The genome contains insertion sequence (IS) elements, phage remnants and many other patches of unusual composition indicating genome plasticity through horizontal transfer.

Genomes of Pathogenic Bacteria

The genomes of a number of pathogens have been sequenced with the aim of understanding the basis for disease and finding cures and means of disease prevention. With regard to pathogenicity there appears to have been significant gene reduction among many pathogenic bacterial species, reflecting their specialized ecological niches. There also appears to be extensive horizontal gene transfers among the pathogenic bacteria, where islands of related genes have moved horizontally, contributing to the dispersion of pathogenicity.

Escherichia coli O157:H7

The genome of *Escherichia coli* O157:H7, which causes haemorrhagic colitis and haemolytic uraemic syndrome, contains 1387 genes not found in the nonpathogenic *E. coli* K-12 (Nicole *et al.*, 2001). These new genes are encoded in strain-specific clusters of diverse sizes. Most differences in overall gene content between *E. coli* O157:H7 and *E. coli* K-12 are attributable to horizontal transfer, and offer a wealth of candidate genes that may be involved in pathogenesis. Base substitution has introduced variation into most gene products – even among conserved regions of the two strains. The new genes in *E. coli* O157:H7 appear to have arisen as multiple events over the 4.5 million years of evolutionary divergence from a common ancestor with *E. coli* K-12. These genes encode virulence factors, alternative metabolic capacities, several prophages and other new functions. The magnitude and the distribution of the divergent genes suggest that the evolution of pathogenicity is complex.

Haemophilus influenzae

The first complete nucleotide sequence of a nonviral, free-living life form, the bacterium *Haemophilus influenzae*, 1 830 137 bp encoding 1743 open reading frames, was accomplished in 1995 using shotgun cloning and alignment of contigs (**Figure 4**) (Fleischmann *et al*., 1995). Genes encoding the complete glycolytic pathway and the production of fermentative end products were identified. Also identified were genes encoding functional anaerobic electron transport systems that depend on inorganic electron acceptors such as nitrates, nitrites, and dimethyl sulfoxide. Genes encoding three enzymes of the classic tricarboxylic acid (TCA) cycle appear to be absent from the genome – citrate synthase, isocitrate dehydrogenase and aconitase were not identified. Six rRNA operons were identified, each containing three subunits and a variable spacer region in the order. Over a third of the open reading frames could not be definitively assigned a functional role (389 had no functional role and 347 matched hypothetical functions), indicating much remains to be learned about bacterial genomics.

Mycoplasma genitalium

The complete nucleotide sequence of the sexually transmissible pathogen *Mycoplasma genitalium*, which contains 580 070 bp and encodes 480 ORFs, is the smallest known genome of any free-living organism (**Figure 5**) (Fraser *et al*., 1995; Hutchison *et al*., 1999). Thirty per cent of the genome (140 genes) is devoted to the structure and function of the cytoplasmic membrane. The portion of the *Mycoplasma* genome dedicated to coding lipoproteins is relatively large and suggests that this class of membrane proteins is important to the cell. Translation requires nearly 90 different proteins whereas DNA replication only requires about 30 proteins. A surprising 4.5% of the genome is used for systems that evade mammalian host cell responses. *M. genitalium* has only five regulatory functional genes, less than 10% of the number found in *H. influenzae*. It appears from transposon mutational analyses that 265–350 of the 480 protein-coding genes of *M. genitalium* are essential for growth conditions, including about 100 genes of unknown function. Relatively few *M. genitalium* genes have functions related to biosynthesis and metabolism and the limited metabolic capacity is compensated for by a relatively high proportion of transport genes. Genes for some metabolic pathways appear to be essential, including the eight genes encoding ATP–proton-motive force interconversion activities and the genes involved in glycolysis, which appears to be the major source of ATP.

Mycoplasma pneumoniae

The genome of the bacterium *Mycoplasma pneumoniae* M129, which causes pneumonia, has 816 394 bp, encoding 677 ORFs and 39 genes coding for various RNA species (Himmelreich *et al*., 1996). Of the predicted ORFs, 75.9% show significant similarity to genes/proteins of other organisms while only 9.9% do not reveal such similarities.

Gene role	Genes	Per cent	Colour
Unknown	30	1.99	⬛
Hypothetical	353	23.50	⬜
Protein fate	69	4.59	🟪
Cell envelope	102	6.79	🟩
Transcription	28	1.86	⬜
DNA metabolism	91	6.06	🟨
Other categories	19	1.26	⬜
Energy metabolism	143	9.52	🟩
Protein synthesis	117	7.79	🟪
Cellular processes	67	4.46	🟥
Regulatory functions	64	4.26	🟦
Amino acid biosynthesis	71	4.73	🟧
Transport and binding proteins	165	10.90	🟦
Central intermediary metabolism	42	2.79	🟫
Fatty acid and phospholipid metabolism	34	2.26	🟪
Purines, pyrimidines, nucleosides, and nucleotides	50	3.33	🟧
Biosynthesis of cofactors, prosthetic groups, and carriers	56	3.73	⬜
Number of genes	1501	100.00	

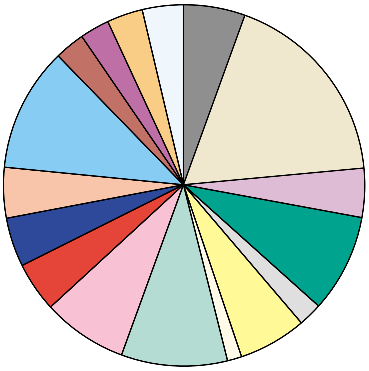

Figure 4 Functional distribution of genes for the bacterium *Haemophilus influenzae* KW20.

Gene role	Genes	Per cent	Colour
Unknown	12	2.51	⬛
Hypothetical	168	35.20	⬜
Protein fate	21	4.40	🟪
Cell envelope	29	6.07	🟩
Transcription	13	2.72	⬜
DNA metabolism	29	6.07	🟨
Other categories	0	0.00	⬜
Energy metabolism	33	6.91	🟩
Protein synthesis	90	18.80	🟥
Cellular processes	6	1.25	🟥
Regulatory functions	5	1.04	🟦
Amino acid biosynthesis	0	0.00	🟧
Transport and binding proteins	33	6.91	🟦
Central intermediary metabolism	7	1.46	🟫
Fatty acid and phospholipid metabolism	8	1.67	🟪
Purines, pyrimidines, nucleosides, and nucleotides	19	3.98	🟧
Biosynthesis of cofactors, prosthetic groups, and carriers	4	0.83	⬜
Number of genes	477	100.00	

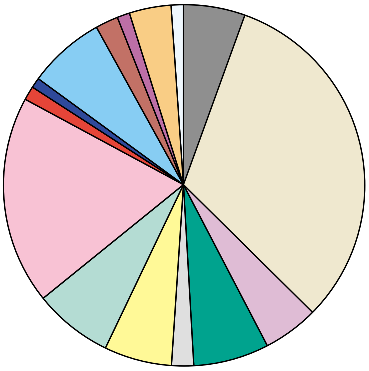

Figure 5 Functional distribution of genes for the bacterium *Mycoplasma genitalium* G-37KW20.

M. pneumoniae has a small genome resulting from the evolutionary loss of complete biosynthetic pathways (e.g. no pathway for complete amino acid synthesis). The limited genomic capability of *M. pneumoniae* forces it to act as an obligate parasite to obtain required nutrients. Surprisingly, the *M. genitalium* genome is almost entirely included within the *M. pneumoniae* genome, raising interesting questions about the evolutionary relationship of these two closely related organisms.

Ureaplasma urealyticum

The 751 719-bp genome of the urinary tract pathogen *Ureaplasma urealyticum* is a circular chromosome with 1362 ORFs. About half of the genome has been assigned known or hypothetical functions. Of this portion, 23% of the genes are involved in protein synthesis.

Borrelia burgdorferi

The genome of the bacterium *Borrelia burgdorferi* B31, the aetiologic agent of Lyme disease, contains a linear chromosome of 910 725 bp and at least 17 linear and circular plasmids with a combined size of more than 533 000 bp (Fraser *et al.*, 1997; Casjens *et al.*, 2000). The chromosome contains 853 genes encoding a basic set of proteins for DNA replication, transcription, translation, solute transport and energy metabolism, but no genes for cellular biosynthetic reactions. Of 430 genes on 11 plasmids, most have no known biological function; 39% of plasmid genes are paralogues that form 47 gene families. The biological significance of the multiple plasmid-encoded genes is not clear, although they may be involved in antigenic variation or immune evasion. Genetic rearrangements appear to have contributed to a surprisingly large number of apparently nonfunctional pseudogenes, an unusual feature for a bacterial genome.

Treponema pallidum

The genome sequence of *Treponema pallidum*, which causes syphilis, contains 1 138 006 bp encoding 1041 ORFs (**Figure 6**) (Fraser *et al.*, 1998). Systems for DNA replication, transcription, translation and repair are intact, but catabolic and biosynthetic activities are minimized. No phosphoenolpyruvate:phosphotransferase carbohydrate transporters have been identified and the overall number of identifiable transporters is small. Potential virulence factors include a family of 12 potential membrane proteins and several putative haemolysins. Comparison of the *T. pallidum* and *Borrelia burgdorferi* genomes indicates considerable diversity among pathogenic spirochaetes.

Campylobacter jejuni

Campylobacter jejuni, which causes gastroenteritis, has a circular chromosome of 1 641 481 bp that is predicted to encode 1654 proteins and 54 stable RNA species (Parkhill *et al*., 2000). The genome is unusual in that there are virtually no insertion sequences or phage-associated sequences and very few repeat sequences. One of the most striking findings in the genome is the presence of hypervariable sequences (short homopolymeric nucleotide sequences) that occur frequently in genes encoding the biosynthesis or modification of surface structures, and in closely linked genes of unknown function. The apparently high rate of variation of these variable sequences may be important in the survival strategy of *C. jejuni*.

Helicobacter pylori

Helicobacter pylori strain 26695, which causes peptic ulcers, has a circular genome of 1 667 867 bp and 1590 predicted coding sequences (**Figure 7**) (Tomb *et al*., 1997; Alm *et al*., 1999). Sequence analysis indicates well-developed systems for motility, scavenging iron and DNA restriction and modification. A high number of genes for adhesins, lipoproteins and other outer membrane proteins probably contribute to pathogenicity and the ability to attach to mucosa. Consistent with its restricted niche, *H. pylori* has few regulatory networks, and limited metabolic repertoire and biosynthetic capacity. Its survival in acid conditions depends, in part, on its ability to establish a positive inside-membrane potential in low pH. Comparison of the complete genomic sequences of two unrelated *H. pylori* isolates reveals similar overall genomic organization, gene order and predicted proteomes; only 6–7% of the genes are specific to each strain, with almost half of these genes being clustered in a single hypervariable region.

Chlamydia trachomatis

Analysis of the 1 042 519-bp sexually transmissible *Chlamydia trachomatis* genome, encoding 847 ORFs, reveals a mosaic of genes, including a large number of genes with phylogenetic origins from eukaryotes, which implies a complex evolution for adaptation to obligate intracellular parasitism (Read *et al*., 2000; Stephens *et al*., 1998). *C. trachomatis* lacks many biosynthetic capabilities, but retains functions for performing key steps and interconversions of metabolites obtained from mammalian host cells. The genome reveals numerous potential virulence-associated proteins.

Gene role	Genes	Per cent	Colour
Unknown	36	4.75	⬛
Hypothetical	176	23.20	⬜
Protein fate	47	6.20	🟪
Cell envelope	53	7.00	🟩
Transcription	25	3.30	⬜
DNA metabolism	51	6.73	🟨
Other categories	0	0.00	⬜
Energy metabolism	54	7.13	🟩
Protein synthesis	97	12.80	🟪
Cellular processes	77	10.10	🟥
Regulatory functions	22	2.90	🟦
Amino acid biosynthesis	7	0.92	🟧
Transport and binding proteins	59	7.79	🟦
Central intermediary metabolism	6	0.79	🟫
Fatty acid and phospholipid metabolism	11	1.45	🟪
Purines, pyrimidines, nucleosides, and nucleotides	21	2.77	🟧
Biosynthesis of cofactors, prosthetic groups, and carriers	15	1.98	⬜
Number of genes	757	100.00	

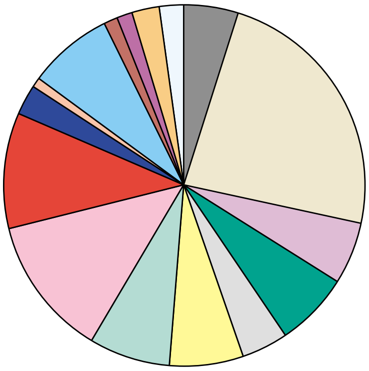

Figure 6 Functional distribution of genes for the bacterium *Treponema pallidum* Nichols.

Gene role	Genes	Per cent	Colour
Unknown	24	2.19	⬛
Hypothetical	187	17.10	⬜
Protein fate	41	3.75	🟪
Cell envelope	100	9.14	🟩
Transcription	10	0.91	⬜
DNA metabolism	90	8.23	🟨
Other categories	17	1.55	⬜
Energy metabolism	100	9.14	🟩
Protein synthesis	99	9.05	🟪
Cellular processes	126	11.50	🟥
Regulatory functions	25	2.28	🟦
Amino acid biosynthesis	42	3.84	🟧
Transport and binding proteins	88	8.05	🟦
Central intermediary metabolism	24	2.19	🟫
Fatty acid and phospholipid metabolism	25	2.28	🟪
Purines, pyrimidines, nucleosides, and nucleotides	38	3.47	🟧
Biosynthesis of cofactors, prosthetic groups, and carriers	57	5.21	⬜
Number of genes	1093	100.00	

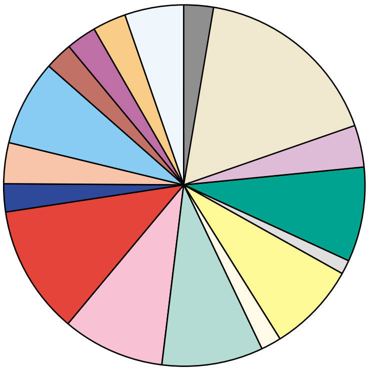

Figure 7 Functional distribution of genes for the bacterium *Helicobacter pylori* 26695.

Chlamydia pneumoniae

Analysis of the 1 230 230-bp *Chlamydia pneumoniae* genome, encoding 547 ORFs, reveals 214 protein-coding sequences not found in *C. trachomatis*, most without homologues to other known sequences (Read *et al.*, 2000; Kalman *et al.*, 1999). There have been multiple large inversion events since the species divergence of *C. trachomatis* and *C. pneumoniae*, apparently oriented around the axis of the origin of replication and the termination region. The striking synteny of the *Chlamydia* genomes and the prevalence of tandemly duplicated genes are evidence of minimal chromosome rearrangement and foreign gene uptake, presumably owing to the ecological isolation of the obligate intracellular parasites.

Rickettsia prowazekii

The 1 111 523-bp genome of *Rickettsia prowazekii*, an obligate intracellular parasite that causes epidemic typhus, contains 834 protein-coding genes (Andersson *et al.*, 1998). Phylogenetic analyses indicate that *R. prowazekii* is more closely related to mitochondria than is any other microbe studied so far. *R. prowazekii* genes show similarities to those of mitochondrial genes. No genes for anaerobic glycolysis are found in either *R. prowazekii* or mitochondrial genomes. A complete set of genes for the tricarboxylic acid cycle and the respiratory-chain complex is found in *R. prowazekii* so that ATP production in *Rickettsia* is the same as that in mitochondria. The *R. prowazekii* genome contains the highest proportion of noncoding DNA (24%) detected so far in a bacterial genome. As gene elimination appears to be characteristic of pathogen evolution, such noncoding sequences may be destined to be eliminated from the genome.

Mycobacterium tuberculosis

The genome of *Mycobacterium tuberculosis* H37Rv, which causes human tuberculosis, comprises 4 411 529 bp that code for approximately 4000 genes (**Figure 8**) (Cole *et al.*, 1998). Of the 3924 open reading frames, precise functions have been assigned to 40%, 44% exhibit high homology to known and hypothetical proteins, and 16% resemble no known proteins and may account for specific mycobacterial functions. *M. tuberculosis* differs radically from other bacteria in that a very large portion of its coding capacity is devoted to the production of enzymes involved in lipogenesis and lipolysis, and to two families of glycine-rich proteins with a repetitive structure that may represent a source of antigenic variation. *M. tuberculosis* contains examples of every known lipid and polyketide biosynthetic system, including enzymes usually found in mammals and plants as well as the common bacterial systems. There are approximately 250 different enzymes involved in fatty acid metabolism in *M. tuberculosis* compared with only 50 in *E. coli*.

Gene role	Genes	Per cent	Colour
Unknown	335	11.60	⬛
Hypothetical	725	25.10	⬜
Protein fate	79	2.74	🟪
Cell envelope	114	3.96	🟩
Transcription	37	1.28	⬜
DNA metabolism	64	2.22	🟨
Other categories	70	2.43	⬜
Energy metabolism	227	7.88	🟩
Protein synthesis	114	3.96	🟪
Cellular processes	101	3.50	🟥
Regulatory functions	164	5.69	🟦
Amino acid biosynthesis	84	2.91	🟧
Transport and binding proteins	162	5.62	🟦
Central intermediary metabolism	274	9.52	🟫
Fatty acid and phospholipid metabolism	168	5.83	🟪
Purines, pyrimidines, nucleosides, and nucleotides	56	1.94	🟧
Biosynthesis of cofactors, prosthetic groups, and carriers	104	3.61	⬜
Number of genes	2878	100.00	

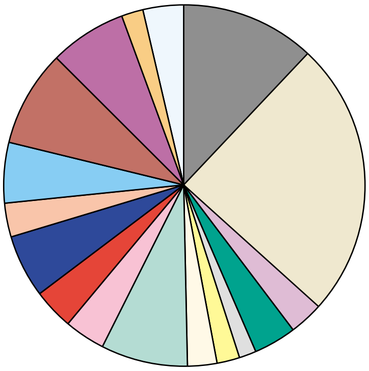

Figure 8 Functional distribution of genes for the bacterium *Mycobacterium tuberculosis* CSU 93.

Neisseria meningitidis

The genome of the serogroup A strain of *Neisseria meningitidis* Z2491, which causes bacterial meningitis, consists of 2 184 406 bp with 2121 predicted coding sequences (Tettelin *et al.*, 2000). The most notable feature of the genome is the presence of many hundreds of repetitive elements, ranging from short repeats, positioned either singly or in large multiple arrays, to insertion sequences and gene duplications of one kilobase or more. Many of these repeats appear to be involved in genome fluidity and antigenic variation in this important human pathogen. The analysis of the genome indicates three mechanisms of repeat-mediated antigenic variation within the *N. meningitidis* genome: on/off switching and transcriptional modulation of gene expression by slipped-strand mispairing of short tandem repeats; intragenomic recombination of localized repeats leading to the use of different C-termini for surface-exposed proteins; and intergenomic gene conversion of specific surface-associated genes associated with large arrays of global repeats, mediated by the internalization of related DNA through the highly repetitive DNA uptake sequence.

Vibrio cholerae

The genomic sequence of the Gram-negative bacterium *Vibrio cholerae* El Tor N16961, which causes cholera, consists of 4 033 460 bp with two circular chromosomes of 2 961 146 bp and 1 072 314 bp that together encode 3885 open reading frames (Heidelberg *et al.*, 2000). The vast majority of recognizable genes for essential cell functions (e.g. DNA replication, transcription, translation and cell wall biosynthesis) and pathogenicity (e.g. toxins, surface antigens and adhesins) are located on the large chromosome. In contrast, the small chromosome contains a larger fraction (59%) of hypothetical genes compared with the large chromosome (42%), and carries a gene capture system (the integron island) and host 'addiction' genes that are typically found on plasmids. There are 105 duplications with at least one of each ORF on each chromosome indicating there have been recent crossovers between chromosomes. The extensive duplication of genes involved in scavenging behaviour (chemotaxis and solute transport) suggests the importance of these gene products in *V. cholerae* biology, notably its ability to inhabit diverse environments. *V. cholerae* has numerous transport proteins with broad substrate specificity and the corresponding catabolic pathways to enable it to survive and to grow in varied ecosystems.

Xylella fastidiosa

The complete genome sequence of *Xylella fastidiosa* clone 9a5c, which causes citrus variegated chlorosis of orange trees, comprises a 2 679 305-bp circular chromosome and two plasmids of 51 158 bp and 1285 bp. Putative functions can be assigned to 47% of the 2904 predicted coding regions (Simpson *et al*., 2000). Efficient metabolic functions are predicted, with sugars as the principal energy and carbon source, supporting existence in the nutrient-poor xylem sap. Orthologues of some of these proteins have only been identified in animal and human pathogens; their presence in *X. fastidiosa* indicates that the molecular basis for bacterial pathogenicity is both conserved and independent of host. At least 83 genes are bacteriophage-derived and include virulence-associated genes from other bacteria, providing direct evidence of phage-mediated horizontal gene transfer.

References

Alm RA, Ling LS, Moir DT *et al*. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**: 176–180.

Andersson SG, Zomorodipour A, Andersson JO *et al*. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**: 133–140.

Blattner FR, Plunkett G III, Bloch CA *et al*. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**: 1453–1474.

Casjens S, Palmer N, van Vugt R *et al*. (2000) A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Molecular Microbiology* **35**: 490–516.

Cole ST, Brosch R, Parkhill J *et al*. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**: 537–544.

Deckert G, Warren PV, Gaasterland T *et al*. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**: 353–358.

Fleischmann RD, Adams MD, White O *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496–512.

Fraser CM, Gocayne JD, White O *et al*. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**: 397–403.

Fraser CM, Casjens S, Huang WM *et al*. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**: 580–586.

Fraser CM, Norris SJ, Weinstock GM *et al*. (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**: 375–388.

Heidelberg JF, Eisen JA, Nelson WC *et al*. (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**: 477–483.

Himmelreich R, Hilbert H, Plagens H *et al*. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Research* **24**: 4420–4449.

Hutchison CA III, Peterson SN, Gill SR *et al*. (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**: 2165–2169.

Kalman S, Mitchell W, Marathe R *et al*. (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature Genetics* **21**: 385–389.

Kaneko T, Sato S, Kotani H *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**: 109–136.

Kunst F, Ogasawara N, Moszer I *et al*. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**: 249–256.

Nelson KE, Clayton RA, Gill SR *et al*. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**: 323–329.

Nicole TP, Plunkett G, Burland V *et al*. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**: 529–533.

Parkhill J, Wren BW, Mungall K *et al*. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**: 665–668.

Read TD, Brunham RC, Shen C *et al*. (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Research* **28**: 1397–1406.

Simpson AJG, Reinach FC, Arruda P *et al*. (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature* **406**: 151–157.

Stephens RS, Kalman S, Lammel C *et al*. (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**: 754–759.

Tettelin H, Saunders NJ, Heidelberg J *et al*. (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287**: 1809–1815.

Tomb JF, White O, Kerlavage AR *et al*. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**: 539–547.

White O, Eisen JA, Heidelberg JF *et al*. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286**: 1571–1577.