— REVIEW —

The Genetics of Childhood Disease and Development: A Series of Review Articles

This review by Dr. Chan is the sixth in this series. In his article, Dr. Chan describes the unique features of bacterial genomes and the many mechanisms leading to bacterial diversification. Knowledge of these processes is essential to diagnosis, treatment and prvention of bacterial infections and to understanding host-bacterial interaction in the pathogenesis of disease.

Alvin Zipursky
Editor-in-Chief

Bacterial Genomes and Infectious Diseases

VOON LOONG CHAN

Department of Medical Genetics and Microbiology, and Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada

ABSTRACT

The genome sequencing approach has proved to be highly effective and invaluable for gaining an insight on structure of bacteria genomes and the biology and evolution of bacteria. The diversity of bacteria genomes is beyond expectation. Gaining a full understanding of the biology and pathogenic mechanisms of these pathogens will be a major task because on an average only approximately 69% of the encoded proteins in each genome have known functions. Genome sequence analyses have identified novel putative virulence genes, vaccine candidates, targets for antibacterial drugs, and specific diagnostic probes. Microarray technology that makes use of the genomic sequences of human and bacterial pathogens will be a major tool for gaining full understanding of the complexity of host–pathogen interactions and mechanisms of pathogenesis. (***Pediatr Res*** **54: 1–7, 2003**)

Abbreviations:
G+C, guanosine plus cytidine
HTG, horizontally transferred gene
ORF, open reading frame
PAI, pathogenicity island

The first bacterial genome, *Hemophilus influenzae*, completely sequenced, annotated, and published was in 1995 by Fleischmann *et al.* (1) at The Institute for Genomic Research. Today, 73 prokaryotic (archaeal and bacterial) genomes have been completed and at least 120 others are in various stages of completion (http://www.tigr.org/tdb/mdb/mdbcomplete.html; http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html; http://www.sanger.ac.uk/Projects/Microbes/). Among the completed genomes, 35 are human bacterial pathogens (2–35) (see Table 1). Table 1 presents five key features of these genomes: size, topology, guanosine plus cytidine (G+C) content, total number of predicted open reading frames (ORFs), percentage of the ORFs with unknown function, and percentage of the ORFs that are unique to the species (or strain). The major impetus in obtaining the complete genome sequence of an organism is to gain a better understanding of the biology and evolution of the microbes and, for pathogens, to identify new vaccine candidates, putative virulent genes, and targets for antibiotics. In this review, some of the major findings about

Table 1. *Genomes of pathogenic bacteria*

Genome	Size (Mbp)	Topology	G + C (%)	ORF (Total no.)	Unknown ORF (%)[d]	Unique ORF (%)[e]	Ref.
Borrelia burgdorferi B31	1.52 (1 chr. 21 pla)	1 linear chr, 12 linear & 9 circular Pla	28.6 (chr)	853	12	29	2,3
Brucella melitensis 16M (Biotype 1)	3.29 Chr-I (2.1), Chr-II (1.18)	Both circular	57	3197	22	7	4
Campylobacter jejuni NCTC11168	1.64	Circular	30.6	1654	13.5	8.7	5
Chlamydia pneumoniae CWL029	1.23	Circular	40.6	1073	23	17	6
Chlamydia pneumonia AR39	1.23	Circular	40.6	1052[a]			7
Chlamydia pneumonia J138	1.22	Circular	40.7	1049[b]			8
Chlamydia trachomatis serovar D (D/UW-3/Cx)	1.04	Circular	41.3	894	32.4	28	9
Chlamydia trachomatis MoPn	1.07	Circular	40.3	924[c]			7
Clostridium perfringens 13	3.03 (chr) 0.054 (pla)	Both circular	28.6	2660(chr)	18.9	25.2	10
Escherichia coli K-12 strain MG1655	4.60	Circular	50.79	4288	38.1		11
Escherichia coli O157:H7 (EDL933)	5.5	Circular	50.4	5416	35.8	26[f]	12
Escherichia coli O157:H7 (Sakai)	5.59	Circular (5.5), 1 pla (92.7 kb)	50.54	5447	37.3	30[g]	13
Haemophilus influenzae Rd (KW20)	1.83	Circular	38	1743	42.2	22.3	1
Helicobacter pylori 26695	1.67	Circular	39	1590	31.4		14
Helicobacter pylori J99	1.64	Circular	39	1495	18.4	23.1	15
Listeria monocytogenes EGD-e (serovar1/2a)	2.94	Circular	39	2853	35.3	9.5	16
Mycobacterium leprae TN	3.26	Circular	57.8	1604	32.8	10.3	17
Mycobacterium tuberculosis H37RV	4.41	Circular	65.6	3924	44	16	18
Mycoplasma genitalium G-37	0.58	Circular	32	470	20.4		19
Mycoplasma pneumoniae M129	0.81	Circular	40.0	677	26.7	16.1	20
Neisseria meningitidis MC58	2.27	Circular	51.5	2158	16.0	24.7	21
Neisseria meningitides serogroup A (Z2491)	2.18	Circular	51.8	2121	33.1	26.9	22
Psedomonas aeruginosa PA01	6.3	Circular	66.6	5570	45.8	32	23
Rickettsia prowazekii Madrid E	1.10	Circular	29.1	834	12.5	24.8	24
Salmonella typhi CT18	4.81 (chr) 0.22 (pla1) 0.11 (pla2)	Circular	52.09 (chr)	4599		13.1[h]	25
Salmonella typhimurium LT2	4.86 (chr) 0.094 (pla)	Circular	53 (chr)	4489	29.2	11[i]	26
Staphylococcus aureus N315	2.81 (chr) 0.25 (pla)	Circular	32.8 (chr)	2595	35.0		27
Staphylococcus aureus Mu50	2.88 (chr) 0.25 (pla)	Circular	32.9 (chr)	2697			27
Streptococcus Pneumoniae TIGR4	2.16	Circular	39.7	2236	36.0	20	28
Streptococcus Pneumoniae R6	2.04	Circular	40.0	2043			29
Streptococcus pyogenes serotype M1, SF370	1.85	Circular	38.5	1752	32	26.6	30
Streptococcus pyogenes serotype M18, MGAS8232	1.89	Circular	38.6	1889			31
Treponema pallidum Nichols	1.14	Circular	52.8	1041	45.0	18.9	32
Ureaplasma, Urealyticum Serovar 3	0.75	Circular	25.5	613	47.0	28.0	33
Vibrio cholerae Serotype O1, Biotype E1 Tor N16961	2.96 (chr-1) 1.07 (chr-2)	Both circular	47.7 (chr-1) 46.9 (chr-2)	2770 1115	42.0 59.0		34
Yersinia pestis CO-92 Biovar Orientalis	4.65 (chr) 0.01 (pla1) 0.07 (pla2) 0.096 (pla3)	All circular	47.6 (chr)	4012	28.9	22.2	35

[a] The *C. pneumoniae* AR39 and CWL029 are essentially identical, differed by only 21 single base frame shift mutations and 296 single nucleotide polymorphisms. AR39 has two small insertions (25 and 85 bp), and CWL029 has five insertions (5, 5, 7, 89, and 305 bp).

[b] Three insertions (ranging in size from 27 to 84 nt) are unique to the J138 genome, and five DNA segments (ranging in size from 89 to 1649 nt) are unique to CWL029. The rest of the two genomes are almost identical (identity, 99.9%).

[c] The main difference between the genome of *C. trachomatis* serovar D and MoPn strains are localized to a single plasticity zone, which is 22.9 kb in serovar D and 50.6 kb in MoPn. Strain MoPn also has three novel genes and few rare gene duplications.

[d] Unknown ORF (%) was either given in the cited reference or calculated on the basis of the number of ORFs that are not in clusters of orthologous groups of proteins (COGs) and the numbers that are classified in COGs with unknown function (http://www.ncbi.nlm.nih.gov/COG/).

[e] Unique ORF (%) was either given in the cited reference or calculated on the basis of the numbers of ORFs that are not classified in the COGs.

[f] O157 (EDL933)-specific genes relative to the K-12 strain.

[g] O157 (Sakai)-specific genes relative to the K-12 strain.

[h] Unique to *S. typhi* compared with *S. typhimurium*.

[i] Unique to *S. typhimurium* compared with *S. typhi*.

bacterial genomes and their impact on strategy and approach for investigating mechanisms of pathogenesis, prevention, and treatment of infectious diseases are highlighted.

DIVERSITY OF BACTERIAL GENOMES

Bacteria generally have a double-stranded circular DNA genome. However, a number of species have been shown to have a linear chromosome (36). The best known example is *Borrelia burgdorferi*, which also has the distinction of having the largest number of extrachromosomal elements, 12 linear and nine circular plasmids (2, 3). Unlike other bacteria, *Vibrio cholerae* (34) and *Deinococcus radiodurans* (37) have two circular chromosomes per cell.

The size of a bacterial genome can range from 0.58 Mbp to 10.0 Mbp, and the G+C content can vary from 25% to 75% (36). Bacteria of the same species and closely related species have very similar G+C content and generally are of similar size. However, for some species, the size of the genome can vary markedly. For example, *Escherichia coli* K-12 has a genome size of 4.6 Mbp, whereas *E. coli* O157:H7 is 5.6 Mbp (1 Mbp can encode 1000 extra genes). In contrast, different isolates of *Chlamydia pneumoniae* have similar genomic size, overall organization, gene order of orthologous genes, and predicted proteomes (see Table 1) (8).

Diversity of codon usage has also been observed among bacteria. Significant deviation from standard "universal" code has been observed in bacteria with small genomes and/or extreme G+C content. UGA, a stop codon in the standard genetic code, encodes tryptophan in *Mycoplasma* (*e.g. M. genitalium* and *Ureaplasma urealyticum*). The CGG codon for arginine is an unassigned (of no function) codon in *M. capricolon* (25% G+C), whereas in *Micrococcus luteus* (74% G+C), AGA (arginine) and AUA (isoleucine) are unassigned (38).

BACTERIAL EVOLUTION

Deletions, pseudogenes, and reductive evolution. Bacterial evolution is dependent on natural selection operating on mixed populations of parental forms and genetic variants. Genetic variations are generated through spontaneous mutations, intrachromosomal recombinations, and acquisition of DNA from other organisms. Spontaneous mutations produce variant genes, including pseudogenes (nonfunctional genes, usually with one or more nonsense or frame-shift mutations). Intrachromosomal recombinations and spontaneous mutations can generate major deletions, duplications, and translocations, leading to chromosomal and gene reorganization. With the exception of a few gene clusters (*e.g.* ribosomal protein genes), organization of genes is not conserved between distantly related bacteria (39). Even between closely related species, such as *M. genitalium* and *U. urealyticum*, there is little conservation of gene order across large fractions of the genome (19, 33).

Obligate intracellular parasitic bacterial genomes seem to have undergone reductive evolution (40). Evidence of this was observed in the genome of *Rickettsia prowazekii* (24, 41) and *Mycobacterium leprae* (17). In the *M. leprae* (17) and *R. prowazekii* (24) genomes, the protein encoding regions represent only 76.5% and 75.4%, respectively. Thus, large regions of the genome contain noncoding sequences, which may represent regulatory sequences or residual genes mutated beyond recognition. The average coding content of free-living bacteria is approximately 90%. The *Campylobacter jejuni* genome is the most dense with 94.3% (5). Among all of the genomes sequenced, the *M. leprae* genome has the highest level of pseudogenes, representing 23.5% of the genome (17). Many pseudogenes eventually will be lost from the genome through deletions.

Mollicutes (commonly named mycoplasmas) as a group have the smallest genome. The genome of *M. genitalium* is only 0.58 Mbp, which is the smallest of the bacterial genomes sequenced. It is predicted to have 470 ORFs with an average size of 1040 bp and composing 88% of the genome, a value similar to most free-living bacteria (19). The reduction in genome size is not due to an increase in gene density or a decrease in gene size. It seems to have evolved from a much larger genome through successive deletion of nonessential genes and the acquisition of a small number of new genes that are largely involved in transport. There is an almost complete lack of genes involved in amino acid biosynthesis, *de novo* nucleotide biosynthesis, and fatty acid biosynthesis. The genomes of two other Mollicutes, *M. pneumoniae* (20) and *U. urealyticum* (33), have also been sequenced (see Table 1). Comparative analysis of these three genomes and that of other bacteria provide information for defining essential functions of a minimal self-replicating bacterial cell.

Recent comparative sequence analyses and genetics studies of commensal *E. coli*, enteroinvasive *E. coli*, and *Shigella* species showed that deletion of certain large genomic fragments or genes that inhibit virulence can lead to increase in virulence. These large genomic sequences that are absent from pathogenic strains but are present in nonpathogenic isolates are known as "black holes" (42). Formation of black holes is a complementary but inverse pathway to the acquisition of pathogenicity islands (see below) in the evolution of pathogenic lineages.

Horizontally transferred genes. Phylogenetics and comparative sequence analysis of orthologous genes (homologous genes with the same function) from distantly and closely related species provide strong indication that in many bacterial genomes some genes are acquired through a horizontal (lateral) transfer process. A phylogenetic tree generated from orthologous protein or gene sequences from various species should reflect the evolutionary relationship and should be congruent with trees that are generated basing on 16S rRNA sequences. Noncongruence of the trees generated or contradiction to established evolutionary relationship would suggest horizontally transferred genes (HTGs). Horizontal transfer of genes can be mediated by DNA transformation, phage-mediated transduction, or plasmid-mediated conjugation. The G+C content and codon usage of HTGs should resemble that of the genome of the donors allowing easy identification if the donor DNA has significantly different G+C content from that of the recipient. HTGs often exist in clusters, ranging in size from a few kb to 200 kb. These clusters are generally known as genomic islands. Islands that contain genes involved in pathogenesis are called

pathogenicity islands (PAIs) (43). PAIs are often absent from nonvirulent strains, flanked by small direct repeat sequences, associated with a t-RNA gene, and often carry insertion and mobile elements that encode integrases and transposases. Koonin *et al.* (44) estimated the levels of HTG in various sequenced genomes. The estimates vary from a modest level of 1.6% of the genes for *M. genitalium* to 32.6% in *T. pallidum. M. pneumoniae* is also low with a level of 2.1%, and *Pseudomonas aeruginosa*, a proteobacterium found in diverse environment, has an intermediate value of 13.1%. *E. coli* O157:H7 strain EDL933 (enterohemorrhagic) has 177 strain-specific genomic islands scattered over the whole genome that are not found in the *E. coli* K-12 strain MG1655 (nonpathogenic) and MG1655 has 234 genomic islands that are absent in EDL933 strain (12).

Slipped-strand mispairing. Slipped-strand mispairing at homopolymeric (primarily G and C) tracts during replication can generate a high level of mutations at these sites and form hypervariable sequences. Hypervariable homopolymeric sequences were identified in the genome of *C. jejuni* (5) and *Helicobacter pylori* (14, 15). In *C. jejuni*, most of the hypervariable sequences are coincident with the clusters of genes involved in lipooligosaccharide and surface polysaccharide biosynthesis and flagella modification (5). Tract length variation results in translational frameshifting and has been shown to be responsible for phase variation of lipooligosaccharide structure of *C. jejuni* (45, 46) and surface structure of other bacteria, and such variations have been proposed to play a role in adaptive evolution (47).

Bacterial speciation and intraspecies variation. Complete genomic sequences of multiple strains of the same species have been achieved for *Chlamydia pneumoniae*, *C. trachomatis*, *E. coli*, *H. pylori*, *Neisseria meningitides*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Streptococcus pyogenes*. Comparative genomic analyses between strains of the same species have generated valuable information on intraspecies variation and mechanisms of speciation. Intraspecies variation is an inherent feature of bacteria. The magnitude of variation among isolates of the same species can vary significantly for different species and seems to be largely determined by the lifestyle and niches that are occupied by the organism. *C. pneumoniae*, an obligate parasite of human cells, seems to have a very low level of strain variation, presumably because it occupies a stable environment and has limited opportunity to acquire DNA from other bacterial species (6–8). In contrast, *E. coli* strains, which occupy diverse environments and often reside in the presence of numerous and diverse populations of bacteria, have genomes that can differ by as much as 20% in size. Lan and Reeves (48) suggested that there is a need for a species genome concept. They proposed that genes that are found in 95% or more isolates form the core set of genes for the species and genes found in 1% to 95% of isolates are considered auxiliary genes.

VIRULENCE GENES

Comparative sequence analyses enable identification of known virulence proteins with conserved sequences or motifs

and also novel putative virulence proteins and PAIs. The value of genomic sequence is best illustrated by pathogens that are difficult to grow *in vitro* and have poorly characterized genetic systems. Chlamydiae, obligate intracellular eubacteria, are prime examples of such organisms. Complete genome sequence analyses identified species' unique genes and putative virulent genes.

Comparative genomic analyses identified type III protein secretion system genes, which are conserved among bacteria. The type III system transposes effectors and toxins directly into the cytosol of the host cells or into the extracellular milieu. Putative genes encoding type III effector proteins and a *Chlamydia*-specific protein, that may have a role in virulence, were also identified in the *C. trachomatis* MoPn genome (7). The role of a type III secretion system and type III effector proteins in pathogenesis is now well established in pathogenic *E. coli*, *Salmonella*, *Yersinia*, and *Shigella*. In the *M. leprae* genome, a single gene encoding laminin-binding protein, that may be an important virulence factor, was identified (17).

NEW VACCINE CANDIDATES, ANTIBIOTICS TARGETS, AND DNA PROBES

Pathogenic bacteria are becoming resistant to commonly used antibiotics at an alarming rate. Accordingly, there is an urgent need for the development of new antibiotics and vaccines. With the aim of identifying new vaccine candidate genes of *N. meningitides* and *S. pneumoniae*, whole-genome sequence of these two pathogens were scanned to identify ORFs encoding proteins with secretion motifs or similarity to predicted virulence factors. A total of 130 ORFs of *S. pneumoniae* were identified and then cloned into an expression vector. Products were purified and tested for immunogenic activity in a mouse model for induction of protective antibodies against pneumococcal challenge. Six novel antigens encoded by five separate genes of *S. pneumoniae* conferred protection against disseminated infection in the mouse model. These proteins, shown to be widely conserved among different isolates and immunogenic in human infection, are currently being evaluated as a vaccine for the prevention of mucosal infection and invasive disease caused by pneumococci (49). Similar whole-genome scanning identified 570 ORFs of *N. meningitides*, which were cloned into expression vectors, and purified recombinant proteins were then used to immunize mice to generate specific antisera. Using the antisera, seven proteins were localized on the cell surface of *N. meningitides*, induced bactericidal antibodies, and are conserved among different isolates, characteristics of an effective vaccine (50).

Another major benefit of the bacterial genome sequences is in antibacterial drug development (51). Comparative analyses of the encoded proteins of the completed genomes have shown that a significant fraction of the ORFs are unique to the species (see Table 1). Among the sequenced bacterial pathogens, the percentage of genes that are unique to the species can range from 7% to 32%. Some of the proteins encoded by species-specific genes are essential for growth or survival in the infected host and should serve well as novel targets for the development of highly species-specific antibacterial drugs.

Species-specific antibiotics have the potential to reduce the imminent problem of interspecies transfer of drug-resistant genes and reduce nonspecific toxicity to the beneficial commensal microflora in the gut.

An important and immediate benefit of having sequenced the complete genome of bacteria is the potential for developing rapid and highly species-specific DNA probes and immuno-probes for the identification of pathogens. Using PCR technology, it should be possible to develop multitarget probes based on conserved species-specific genes and known virulence genes. Rapid and reliable specific tests will improve treatment of infectious diseases and reduce the levels of misuse of antibiotics. Genomic sequences have also had an impact on the development of typing schemes of infectious agents. A typing scheme based on multilocus sequences has been developed for various pathogens and may well become the gold standard for typing bacterial pathogens (52).

BACTERIA–HOST INTERACTIONS

The availability of a complete genomic sequence makes it possible to examine the global transcription profile of a cell. Both bacterial and mammalian (mouse, human) genome sequences can be used in microarray technology to define the expression profile of pathogens and the host cells. The global transcription effects on host cells by various bacterial pathogens, including *Listeria monocytogenes*, *Salmonella*, *Pseudomonas aeruginosa*, and *Bordetella pertussis* have been analyzed by using microarray technology (53). Rosenberger *et al.* (54) identified novel macrophage genes whose level of expression are altered in *S. typhimurium* infection or when treated with lipopolysaccharide. Similarly, Cohen *et al.* (55) identified 74 up-regulated RNAs and 23 down-regulated host RNAs in *L. monocytogenes*–infected human promyelocytic THP1 cells. Infection of human bronchial epithelial cells (BEAS-2B) by *B. pertussis* results in an increase in transcriptional levels of 33 genes and decrease in transcriptional levels of 65 genes (56). Many of the up-regulated genes encode proinflammatory cytokines (*e.g.* IL-8, IL-6, and growth-related oncogene-1), and many of the down-regulated genes encode transcriptional factors and cellular adhesion molecules. Understanding the molecular basis of the host response to bacterial infections is critical for preventing disease and tissue damage resulting from the host response. Furthermore, an understanding of host transcriptional changes induced by the microbes can be used to identify specific protein targets for drug development.

POSTGENOME RESEARCH

Analysis of the 35 completed bacterial pathogen genomes clearly signals how little we comprehend the biology of these pathogens as on the average approximately 31% of the predicted ORFs have unknown functions (Table 1). Understanding the functions and how these genes and their products are regulated are some of the major tasks confronting us. New and better algorithms and programs for structure prediction and identification of new conserved motifs in proteins are needed. Putative virulence genes identified from the sequenced genomes by bioinformatics must be verified experimentally by construction of isogenic mutants and testing them using appropriate animal models. There is an urgent demand for good animal models for some pathogen-induced diseases (*e.g.* campylobacteriosis). Inexpensive animal models with a large repertoire of knock-out mutants defective in innate immune response or signal transduction will also be in great demand. Gene expression profiling technology has recently been expanded to examine the expression profile of a specific cell population (gastric parietal cells) in mice with or without infection by *H. pylori* (57). Similar studies can be extended to other pathogens and animal model organisms. Such studies will provide a clearer picture of the molecular events that occur in human infection. This knowledge is critical for gaining a better understanding of the mechanisms of pathogenesis.

THE NEW "OMICS" ERA

Completion of the large number of microbial genomes and the human genome provide enormous impetus to develop and implement new techniques to manage and exploit this sequence information leading to creation of a new generation of "omics" enterprises, which emphasize on comparative and functional aspects of genomics, transcriptomics, proteomics, metabolomics, infectomics, pharmacogenomics, immunoproteomics, and many others (58–61). An essential component of the "omics" era is the development of new computational methods (bioinformatics) that aim to solve biologic problems (62). New advances in bioinformatics are major driving forces in many areas of biologic research.

REFERENCES

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu LI, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Craig VJ 1995 Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512
2. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb JF, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, Hanson M, van Vugt R, Palmer N, Adams MD, Gocayne J, Weidman J, Utterback T, Watthey L, McDonald L, Artiach P, Bowman C, Garland S, Fujii C, Cotton MD, Horst K, Roberts K, Hatch B, Smith HO, Venter JC 1997 Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature 390:580–586
3. Casjens S, Palmer N, van Vugt R, Huang WM, Stevenson B, Rosa P, Lathigra R, Sutton G, Peterson J, Dodson RJ, Haft D, Hickey E, Gwinn M, White O, Fraser CM 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*. Mol Microbiol 35:490–516
4. DelVecchio VG, Kapatral V, Redkar RJ, Patra G, Mujer C, Los T, Ivanova N, Anderson I, Bhattacharyya A, Lykidis A, Reznik G, Jablonski L, Larsen N, D'Souza M, Bernal A, Mazur M, Goltsman E, Selkov E, Elzer PH, Hagius S, O'Callaghan D, Letesson JJ, Haselkorn R, Kyrpides N, Overbeek R 2002 The genome sequence of the facultative intracellular pathogen *Brucella melitensis*. Proc Natl Acad Sci USA 99:443–448
5. Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, Jagels K, Karlyshev AV, Moule S, Pallen MJ, Penn CW, Quail MA, Rajandream MA, Rutherford KM, van Vliet AH, Whitehead S, Barrell BG 2000 The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 403:665–668
6. Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW, Stephens RS 1999 Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. Nat Genet 21:385–389
7. Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T, Berry K, Bass S, Linher K, Weidman J, Khouri H, Craven B, Bowman C, Dodson R, Gwinn M, Nelson W, DeBoy R, Kolonay J, McClarty G, Salzberg SL, Eisen J, Fraser CM 2000 Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res 28:1397–1406

8. Shirai M, Hirakawa H, Kimoto M, Tabuchi M, Kishi F, Ouchi K, Shiba T, Ishii K, Hattori M, Kuhara S, Nakazawa T 2000 Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. Nucleic Acids Res 28:2311–2314

9. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, Koonin EV, Davis RW 1998 Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282:754–759

10. Shimizu T, Ohtani K, Hirakawa H, Ohshima K, Yamashita A, Shiba T, Ogasawara N, Hattori M, Kuhara S, Hayashi H 2002 Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. Proc Natl Acad Sci USA 99:996–1001

11. Blattner FR, Plunkett 3rd G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y 1997 The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1474

12. Perna NT, Plunkett 3rd G, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR 2001 Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409:529–533

13. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H 2001 Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 8:11–22

14. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzegerald LM, Lee N, Adams MD, Weidman JM, Fujii C, Bowman C, Watthey L, Wallin E, Hayes WS, Borodovsky M, Karp PD, Smith HO, Fraser CM, Venter JC 1997 The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388:539–547

15. Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL, Carmel G, Tummino PJ, Caruso A, Uria-Nickelsen M, Mills DM, Ives C, Gibson R, Merberg D, Mills SD, Jiang Q, Taylor DE, Vovis GF, Trust TJ 1999 Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397:176–180

16. Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, Baquero F, Berche P, Bloecker H, Brandt P, Chakraborty T, Charbit A, Chetouani F, Couve E, de Daruvar A, Dehoux P, Domann E, Dominguez-Bernal G, Duchaud E, Durant L, Dussurget O, Entian KD, Fsihi H, Portillo FG, Garrido P, Gautier L, Goebel W, Gomez-Lopez N, Hain T, Hauf J, Jackson D, Jones LM, Kaerst U, Kreft J, Kuhn M, Kunst F, Kurapkat G, Madueno E, Maitournam A, Vicente JM, Ng E, Nedjari H, Nordsiek G, Novella S, de Pablos B, Perez-Diaz JC, Purcell R, Remmel B, Rose M, Schlueter T, Simoes N, Tierrez A, Vazquez-Boland JA, Voss H, Wehland J, Cossart P 2001 Comparative genomics of *Listeria* species. Science 294:849–852

17. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, Maclean J, Moule S, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutherford KM, Rutter S, Seeger K, Simon S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Taylor K, Whitehead S, Woodward JR, Barrell BG 2001 Massive gene decay in the leprosy bacillus. Nature 409:1007–1011

18. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry 3rd CE, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG 1998 Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537–544

19. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman JL, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison 3rd CA, Venter JC 1995 The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403

20. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R 1996 Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 24:4420–4449

21. Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ, Nelson WC, Gwinn ML, DeBoy R, Peterson JD, Hickey EK, Haft DH, Salzberg SL, White O, Fleischmann RD, Dougherty BA, Mason T, Ciecko A, Parksey DS, Blair E, Cittone H, Clark EB, Cotton MD, Utterback TR, Khouri H, Qin H, Vamathevan J, Gill J, Scarlato V, Masignani V, Pizza M, Grandi G, Sun L, Smith HO, Fraser CM, Moxon ER, Rappuoli R, Venter JC 2000 Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 287:1809–1815

22. Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T, Davies RM, Davis P, Devlin K, Feltwell T, Hamlin N, Holroyd S, Jagels K, Leather S, Moule S, Mungall K, Quail MA, Rajandream MA, Rutherford KM, Simmonds M, Skelton J, Whitehead S, Spratt BG, Barrell BG 2000 Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature 404:502–506

23. Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E, Westbrock-Wadman S, Yuan Y, Brody LL, Coulter SN, Folger KR, Kas A, Larbig K, Lim R, Smith K, Spencer D, Wong GK, Wu Z, Paulsen IT 2000 Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature 406:959–964

24. Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG 1998 The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133–140

25. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N, Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, O'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG 2001 Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature 413:848–852

26. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S, Layman D, Leonard S, Nguyen C, Scott K, Holmes A, Grewal N, Mulvaney E, Ryan E, Sun H, Florea L, Miller W, Stoneking T, Nhan M, Waterston R, Wilson RK 2001 Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature 413:852–856

27. Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki K, Nagai Y, Lian J, Ito T, Kanamori M, Matsumaru H, Maruyama A, Murakami H, Hosoyama A, Mizutani-Ui Y, Takahashi NK, Sawano T, Inoue R, Kaito C, Sekimizu K, Hirakawa H, Kuhara S, Goto S, Yabuzaki J, Kanehisa M, Yamashita A, Oshima K, Furuya K, Yoshino C, Shiba T, Hattori M, Ogasawara N, Hayashi H, Hiramatsu K 2001 Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet 357:1225–1240

28. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ, Durkin AS, Gwinn M, Kolonay JF, Nelson WC, Peterson JD, Umayam LA, White O, Salzberg SL, Lewis MR, Radune D, Holtzapple E, Khouri H, Wolf AM, Utterback TR, Hansen CL, McDonald LA, Feldblyum TV, Angiuoli S, Dickinson T, Hickey EK, Holt IE, Loftus BJ, Yang F, Smith HO, Venter JC, Dougherty BA, Morrison DA, Hollingshead SK, Fraser CM 2001 Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 293:498–506

29. Hoskins J, Alborn Jr WE, Arnold J, Blaszczak LC, Burgett S, DeHoff BS, Estrem ST, Fritz L, Fu DJ, Fuller W, Geringer C, Gilmour R, Glass JS, Khoja H, Kraft AR, Lagace RE, LeBlanc DJ, Lee LN, Lefkowitz EJ, Lu J, Matsushima P, McAhren SM, McHenney M, McLeaster K, Mundy CW, Nicas TI, Norris FH, O'Gara M, Peery RB, Robertson GT, Rockey P, Sun PM, Winkler ME, Yang Y, Young-Bellido M, Zhao G, Zook CA, Baltz RH, Jaskunas SR, Rosteck Jr PR, Skatrud PL, Glass JI 2001 Genome of the bacterium *Streptococcus pneumoniae* strain R6. J Bacteriol 183:5709–5717

30. Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, Primeaux C, Sezate S, Suvorov AN, Kenton S, Lai HS, Lin SP, Qian Y, Jia HG, Najar FZ, Ren Q, Zhu H, Song L, White J, Yuan X, Clifton SW, Roe BA, McLaughlin R 2001 Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc Natl Acad Sci USA 98:4658–4663

31. Smoot JC, Barbian KD, Van Gompel JJ, Smoot LM, Chaussee MS, Sylva GL, Sturdevant DE, Ricklefs SM, Porcella SF, Parkins LD, Beres SB, Campbell DS, Smith TM, Zhang Q, Kapur V, Daly JA, Veasy LG, Musser JM 2002 Genome sequence and comparative microarray analysis of associated with acute rheumatic fever outbreaks. Proc Natl Acad Sci USA 99:4668–4673

32. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum K, Sodergren E, Hardham JM, McLeod MP, Salzberg S, Peterson J, Khalak H, Richardson D, Howell JK, Chidambaram M, Utterback T, McDonald L, Artiach P, Bowman C, Cotton MD, Fujii C, Garland S, Hatch B, Horst K, Roberts K, Sandusky M, Weidman J, Smith HO, Venter JC 1998 Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281:375–388

33. Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH 2000 The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. Nature 407:757–762

34. Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam L, Gill SR, Nelson KE, Read TD, Tettelin H, Richardson D, Ermolaeva MD, Vamathevan J, Bass S, Qin H, Dragoi I, Sellers P, McDonald L, Utterback T, Fleischmann RD, Nierman WC, White O 2000 DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406:477–483

35. Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, Baker S, Basham D, Bentley SD, Brooks K, Cerdeno-Tarraga AM, Chillingworth T, Cronin A, Davies RM, Davis P, Dougan G, Feltwell T, Hamlin N, Holroyd S, Jagels K, Karlyshev AV, Leather S, Moule S, Oyston PC, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG 2001 Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 413:523–527

36. Casjens S 1998 The diverse and dynamic structure of bacterial genomes. Annu Rev Genet 32:339–377

37. White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, Dodson RJ, Haft DH, Gwinn ML, Nelson WC, Richardson DL, Moffat KS, Qin H, Jiang L, Pamphile W, Crosby M, Shen M, Vamathevan JJ, Lam P, McDonald L, Utterback T, Zalewski C, Makarova KS, Aravind L, Daly MJ, Minton KW, Fleischmann RD, Ketchum KA, Nelson KE, Salzberg S, Smith HO, Venter JC, Fraser CM 1999 Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 286:1571–1577

38. Santos MAS, Ueda T, Watanabe K, Tuite MF 1997 The non-standard genetic code of *Candida* spp.: an evolving genetic code of a novel mechanism for adaptation? Mol Microbiol 26:423–431

39. Siefert JL, Martin KA, Abdi F, Widger WR, Fox GE 1997 Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. J Mol Evol 45:467–472

40. Moran NA 2002 Microbial minimalism Genome reduction in bacterial pathogens. Cell 108:583–586

41. Andersson JO, Andersson SG 1999 Genome degradation is an ongoing process in *Rickettsia*. Mol Biol Evol 16:1178–1191

42. Maurelli AT, Fernandez RE, Bloch CA, Rode CK, Fasano A 1998 "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive *Escherichia coli*. Proc Natl Acad Sci USA 95:3943–3948

43. Hacker J, Kaper JB 2000 Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol 54:641–679

44. Koonin EV, Makarova KS, Aravind L 2001 Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709–742

45. Linton D, Gilbert M, Hitchen PG, Dell A, Morris HR, Wakarchuk W, Gregson NA, Wren BW 2000 Phase variation of a β-1,3 galactosyl-transferase involved in generation of the ganglioside GM_1-like lipo-oligosaccharide of *Campylobacter jejuni*. Mol Microbiol 37:501–514

46. Guerry P, Szymanski CM, Prendergast MM, Hickey TE, Ewing CP, Pattarini DL, Moran AP 2002 Phase variation of *Campylobacter jejuni* 81–176 lipooligosaccharide affects ganglioside mimicry and invasiveness in vitro. Infect Immun 70:787–793

47. Moxon ER, Rainey PB, Nowak MA, Lenski RE 1994 Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr Biol 4:24–33

48. Lan R, Reeves PR 2000 Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends Microbiol 8:396–401

49. Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, Choi GH, Barash SC, Rosen CA, Masure HR, Tuomanen E, Gayle A, Brewah YA, Walsh W, Barren P, Lathigra R, Hanson M, Langermann S, Johnson S, Koenig S 2001 Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. Infect Immun 69:1593–1598

50. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, Galeotti CL, Luzzi E, Manetti R, Marchetti E, Mora M, Nuti S, Ratti G, Santini L, Savino S, Scarselli M, Storni E, Zuo P, Broeker M, Hundt E, Knapp B, Blair E, Mason T, Tettelin H, Hood DW, Jeffries AC, Saunders NJ, Granoff DM, Venter JC, Moxon ER, Grandi G, Rappuoli R 2000 Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 287:1816–1820

51. Tang CM, Moxon ER 2001 The impact of microbial genomics on antibacterial drug development. Annu Rev Genomics Hum Genet 2:259–269

52. Enright MC, Spratt BG 1999 Multilocus sequence typing. Trends Microbiol 7:482–487

53. Rappuoli R 2000 Pushing the limits of cellular microbiology: microarrays to study bacteria-host cell intimate contacts. Proc Natl Acad Sci USA 97:13467–13469

54. Rosenberger CM, Scott MG, Gold MR, Hancock RE, Finlay BB 2000 *Salmonella typhimurium* infection and lipopolysaccharide stimulation induce similar changes in macrophage gene expression. J Immunol 164:5894–5904

55. Cohen P, Bouaboula M, Bellis M, Baron V, Jbilo O, Poinot-Chazel C, Galiegue S, Hadibi EH, Casellas P 2000 190 Monitoring cellular responses to *Listeria monocytogenes* with oligonucleotide arrays. J Biol Chem 275:11181–11190

56. Belcher CE, Drenkow J, Kehoe B, Gingeras TR, McNamara N, Lemjabbar H, Basbaum C, Relman DA 2000 From the cover: the transcriptional responses of respiratory epithelial cells to *Bordetella pertussis* reveal host defensive and pathogen counter-defensive strategies. Proc Natl Acad Sci USA 97:13847–13852

57. Mills JC, Syder AJ, Hong CV, Guruge JL, Raaii F, Gordon JI 2000 A molecular profile of the mouse gastric parietal cell with and without exposure to *Helicobacter pylori*. Proc Natl Acad Sci USA 98:13687–13692

58. Vihinen M 2001 Bioinformatics in proteomics. Biomol Eng 18:241–248

59. Wasinger VC, Corthals GL 2002 Proteomic tools for biomedicine. J Chromatogr B Analyt Technol Biomed Life Sci 771:33–48

60. Huang SH, Triche T, Jong AY 2002 Infectomics: genomics and proteomics of microbial infections. Funct Integr Genomics 1:331–344

61. Haas G, Karaali G, Ebermayer K, Metzger WG, Lamer S, Zimny-Arndt U, Diescher S, Goebel UB, Vogt K, Roznowski AB, Wiedenmann BJ, Meyer TF, Aebischer T, Jungblut PR 2002 Immunoproteomics of Helicobacter pylori infection and relation to gastric disease. Proteomics 2:313–324

62. Goodman N 2002 Biological data becomes computer literate: new advances in bioinformatics. Curr Opin Biotechnol 13:68–71