MICROBIAL GENOME ANALYSIS: INSIGHTS INTO VIRULENCE, HOST ADAPTATION AND EVOLUTION

Brendan W. Wren

Genome analysis of microbial pathogens has provided unique insights into their virulence, host adaptation and evolution. Common themes have emerged, including lateral gene transfer among enteric pathogens, genome decay among obligate intracellular pathogens and antigenic variation among mucosal pathogens. The advent of post-genomic approaches and the sequencing of the human genome will enable scientists to investigate the complex and dynamic interplay between host and pathogen. This wealth of information will catalyse the development of new intervention strategies to reduce the burden of microbial-related disease.

PATHOGEN
An organism, generally a microorganism, that can cause disease in animals and plants.

VIRULENCE FACTOR
In the strict sense, a determinant that causes damage to the host cell (for example, an exotoxin). In the broader sense, a determinant required for the survival of the pathogen in the host (for example, the ability to acquire iron).

IMMUNOGEN
An antigen that produces a significant immunological response.

Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK.
e-mail: brendan.wren@lshtm.ac.uk

Despite advances in the treatment of infectious disease, pathogenic microorganisms are the single most important threat to health worldwide. Many infectious disease agents have never been controlled, or have re-emerged as global PATHOGENS, whereas others pose a new threat. Media and scientific attention has focused on a range of problems, including the alarming spread of antibiotic resistance, microbial contamination of the food chain, the global resurgence of malaria and tuberculosis, and other emerging and re-emerging infections triggered by lifestyle, political and ecological changes. It is also becoming increasingly clear that microorganisms have a significant causative role in diseases such as cancer (for example, *Helicobacter pylori* and gastric cancer[1]) and heart disease (for example, *Chlamydia pneumoniae*[2]). The need to gain an integrated and comprehensive understanding of the workings of our old adversaries is as great as ever.

Research in microbial pathogenicity has changed fundamentally in the past few years, from a piecemeal approach of characterizing individual determinants to a more global analysis of host–pathogen interactions. This has been fuelled by developments in molecular biology, cell biology and immunology and, more recently, by the ability to determine the complete genome sequence of microorganisms. In just five years since the publication[3] of the first genome sequence of a free-living organism, *Haemophilus influenzae* Rd, over 50 microbial genomes have been completely or partially sequenced (TABLE 1 and TIGR web site). Genome sequence data will soon be available for most microorganisms of medical or economic significance, including plant, fungal and protozoan pathogens. Already genome sequences of several strains within the same species are available (for example, *Helicobacter pylori*[4,5], *Neisseria meningitidis*[6,7] and *Chlamydia trachomatis*[8,9]). Furthermore, the genome sequencing of a pathogen and a closely related non-pathogen (for example, *Listeria monocytogenes* and *L. innocua*) is near completion, which should allow the identification of determinants responsible for host specificity and virulence[10].

The first stage of a typical genome project involves sequencing a six- to tenfold redundant random shotgun library[3]. After sequence assembly, gaps are closed to give the full sequence and this is followed by interpretation and computer annotation of all predicted coding sequences[11,12]. The complete DNA sequence provides a directory of every potential VIRULENCE FACTOR and IMMUNOGEN, and the speed and ease of gene discovery and identification is vastly greater than that achieved by traditional molecular genetic approaches. Many new toxins, adhesins, invasins, polysaccharide surface structures and

Table 1 | Sequenced human bacterial pathogens

Pathogen strain	Disease	Genome size (Mb)	Reference
Borrelia burgdorferi B31	Lyme disease	1.44	21
Campylobacter jejuni NCTC11168	gastroenteritis	1.64	14
Chlamydia pneumoniae CWL029	acute respiratory disease	1.23	79
Chlamydia pneumoniae AR39	atherosclerosis	1.23	9
Chlamydia trachomatis D/UW-3/Cx	trachoma, blindness	1.05	8
Chlamydia trachomatis MoPn	trachoma, blindness	1.07	9
Haemophilus influenzae Rd	meningitis, otitis media	1.83	3
Helicobacter pylori 26695	peptic ulcer, gastric cancer	1.66	4
Helicobacter pylori J99	peptic ulcer, gastric cancer	1.64	5
Listeria monocytogenes EGD-e	listeriosis, abortion	2.95	10
Mycobacterium leprae	leprosy	3.10	32
Mycobacterium tuberculosis H37Rv	tuberculosis	4.40	13
Mycoplasma genitalium G-37	urethritis	0.58	36
Mycoplasma pneumoniae M129	respiratory disease	0.81	80
Neisseria meningitidis A Z2491	meningitis (developing world)	2.18	7
Neisseria meningitidis B MC58	meningitis	2.27	6
Rickettsia prowazekii Madrid E	typhus	1.10	30
Streptococcus pyogenes M1	toxic shock syndrome, scarlet fever, rheumatic fever, necrotizing fascitis	1.85	29
Treponema pallidum Nichols	syphilis	1.14	33
Vibrio cholerae N16961	cholera	4.03	20

See links.

other determinants characteristic of pathogens have been identified by such analyses[3,4,6,13–19]. However, with few exceptions[15,16], full characterization of these determinants awaits further investigation.

The availability of complete annotated genome sequences, coupled with bioinformatics, has spawned the new scientific discipline of comparative genomics. This allows the comparison of whole genome sequence data between strains, species, genera and even kingdoms. Comparative genomics is a powerful approach for studying differences in phenotype, host range and virulence, and is providing important new insights into the molecular evolution of virulence.

This review discusses emerging themes from the comparative analysis of genome sequences from human pathogens, and how this information is helping to formulate hypotheses on virulence, host adaptation and evolution. The concurrent development of post-genomic approaches to determine gene function is described, as well as selected examples of how this technology and genome sequence data can be used to identify new antimicrobial and vaccine targets.

Genome diversity

Comparison of available genome sequences has revealed wide variation in genome organization among bacterial pathogens. The traditional belief that bacteria contain a single circular chromosome with the occasional circular plasmid has been challenged by a number of completed genome sequences. For example, *Vibrio cholerae* has two circular chromosomes of 2.69 megabases (Mb) and 1.07 Mb[20], and *Borrelia burgdorferi* B31 contains a linear chromosome (0.91 Mb) and at least 17 linear and circular plasmids (totalling 0.53 Mb)[21]. The G+C content of sequenced pathogens also varies, from 29% in *B. burgdorferi*[21] to 65% in *Mycobacterium tuberculosis* H37Rv (REF. 13). In *M. tuberculosis* the high G+C content is reflected in the biased amino-acid content of the proteins, which may be important in its unique pathogenic life style[13].

The diversity of microbial gene content is exemplified by the observation that about half of the predicted coding sequences from bacterial genome sequences are of unknown biological function, and around half of these seem to be unique to the individual microorganism. For example, over 40% of the genes from *Escherichia coli*, the best studied of all microorganisms, have no known function[22]. These statistics emphasize our limited knowledge of bacteria and microbial pathogens in general. Elucidating the function of these 'ORFan' (open reading frames with no similarity to known genes) or 'FUN' (function unknown) genes is one of the biggest challenges of the post-genomic era. The extent of differences among pathogens indicates that it may be inappropriate to consider any single bacterial pathogen as a model organism for comparative biological analysis.

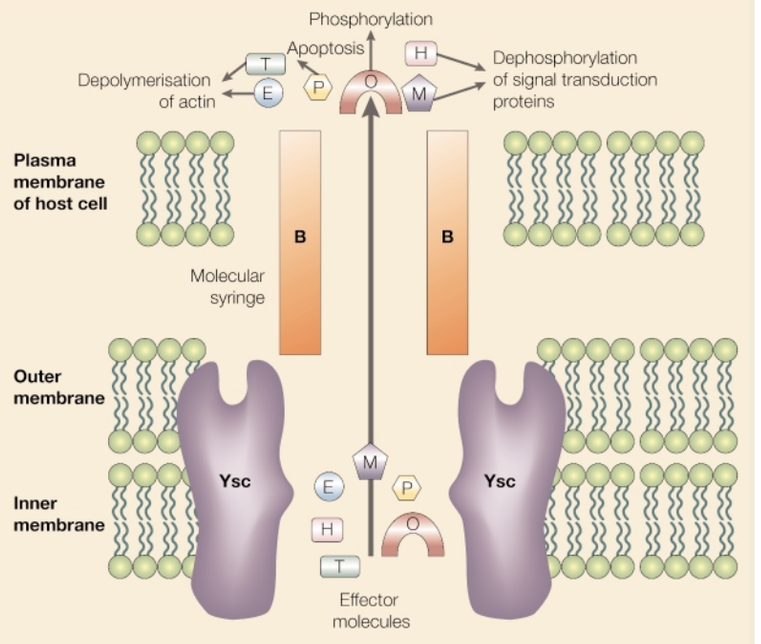
Lateral gene transfer

LATERAL GENE TRANSFER of DNA between different species is well documented, particularly among ENTERIC bacterial pathogens, and is important in the diversification of virulence[23]. DNA can be transferred by several mechanisms, including the uptake of exogenous DNA by

LATERAL GENE TRANSFER
The transfer of DNA, frequently cassettes of genes, between organisms.

ENTERIC PATHOGEN
A pathogen that resides in the gastrointestinal tract.

Box 1 | Type III and type IV bacterial secretion systems



Type III and type IV secretion systems are specialized organelles present in some pathogenic Gram-negative bacteria, which are used to export effector molecules across the bacterial membrane to the extracellular space, or directly into the host cell to modulate host cellular functions. Type III systems are evolutionarily related to each other and to bacterial flagellar organelle structures, and probably originate through lateral gene transfer. Specialized type III secretion systems are found in animal pathogens such as enteropathogenic and enterohaemorrhagic *E. coli*, *Salmonella enterica*, *Bordetella bronchiseptica*, *Pseudomonas aeruginosa*, *Chlamydia*, *Shigella* and *Yersinia* spp., as well as in plant pathogens including *Pseudomonas syringae*, *Ralstonia solanacearum*, *Erwinia* and *Xanthomonas* spp.[25,26].

Type IV systems are a second type of secretion apparatus that have a similar structure to the type III prototype. Conserved type IV gene cassettes have been found in *H. pylori*, *E. coli*, *B. pertussis*, *Legionella pneumophila*, *R. prowazekii* and *Agrobacterium tumefaciens* and have diverse functions ranging from the secretion of toxins to the transfer of DNA to the nucleus of a plant cell[27]. Although gene members within each of the two secretion apparatus pathways are highly conserved and probably originate through lateral gene transfer, the effector molecules they deliver are usually unique for each bacterial species.

Diagram shows the *Yersinia* type III secretion system (Ysc). Ysc consists of YopD, YopF, YopJ, YopL, YopN, YopQ, YopR, YopS, YopU, LcrD and YscC (not shown), connected to YopB (B) allowing the export of effector proteins YopE (E), YopH (H), YopM (M), YopO (O), YopP (P) and YopT (T). The genes encoding proteins for the Ysc are conserved among several animal and plant pathogens and in *Yersinia* are clustered on the 70 kb virulence plasmid.

transformation, conjugation on transferable elements (often transposons), transduction by bacteriophages and genomic recombination[23]. Transfer often involves cassettes of genes ranging in size from 5 to 100 kilobases (kb) and, if they contribute to virulence, such cassettes have been termed pathogenicity islands[24]. On incorporation into a recipient bacterium these DNA regions can convert a benign organism into a pathogen. Examples of the products of pathogenicity islands include type III and type IV secretion systems. Both systems encode specialized organelles that act as molecular syringes to export effector molecules (generally toxins) across the bacterial membrane of Gram-negative bacteria into the host cell to modulate host cellular functions[25–27] (BOX 1). Although the genes encoding the two independent secretion systems are conserved and seem to be acquired by lateral gene transfer, the effector molecules they deliver are often unique for each bacterial species[26,27].

One characteristic of loci acquired by lateral gene transfer is an atypical G+C content relative to the rest of the genome[3,6]. The availability of a complete genome sequence allows genome-wide screening for 'spikes' of G+C variation, offering the opportunity to measure and compare the cumulative effect of lateral gene transfer among pathogens. Comparison of sequenced genomes confirms that bacteria have undergone frequent gene transfer events, many of which act as markers for possible virulence determinants. The gastric pathogen *Helicobacter pylori* 26695 (formerly *Campylobacter pylori*) has over 30 G+C spikes and a number of these contain tell-tale remnants of portable transposon or insertion sequences[4,5,28]. This contrasts with the closely related pathogen *Campylobacter jejuni* NCTC11168, in which there is little evidence of lateral gene transfer[14]. The similarity between these pathogens is mainly restricted to housekeeping genes, and genes required for most functions related to survival, transmission and pathogenesis are remarkably dissimilar. So lateral gene transfer is largely responsible for the genotypic and phenotypic differences between these closely related pathogens. It seems that selective pressures have driven profound evolutionary changes to create two very different pathogens from a relatively recent common ancestor. The genome sequence of a second *H. pylori strain* (J99) confirmed significant lateral gene transfer in this naturally transformable species, with up to 7% of the genes being specific to each strain[5]. Half of these strain-specific genes are clustered in a single hypervariable region, termed a plasticity zone[5].

The genome sequence of the naturally COMPETENT bacterium *Neisseria meningitidis serotype* B MC58 (2.23 Mb) contains 1,910 DNA uptake signal sequences that are known to mediate the integration of exogenous DNA into the chromosome[6]. Analysis of G+C content revealed at least three large regions of foreign DNA, two of which contain genes encoding proteins involved in pathogenicity[6]. Moreover, 22 intact and 29 remnant insertion sequences were detected, indicating further instances of foreign DNA acquisition[6]. The group A streptococci, which are responsible for a diverse array of diseases including scarlet fever and necrotizing fasciitis, have at least four bacteriophage genomes present in their genome sequence, all of which encode one or more EXOTOXINS[29]. Typically, these genes are found adjacent to the phage DNA insertion site on the chromosome and provide clear evidence of lateral gene transfer relating to virulence acquisition[29]. In *V. cholerae*, it has been suggested that the smaller of the two chromosomes can be considered as a mega-plasmid that was captured by a *Vibrio* species in its ancestral past, and

COMPETENCE
The ability of bacteria to take up exogenous DNA molecules.

EXOTOXINS
Generally secreted proteins that cause damage to the host cell.

which provides it with a protective advantage[20]. This selective advantage may relate to the marine part of the *V. cholerae* life cycle (the probable natural habitat of *V. cholerae*) rather than to adaptation to the human host. Overall, the association between lateral gene transfer and the acquisition of virulence or pathogenicity determinants is providing valuable information about potential virulence determinants that may be targets for therapeutic interventions.

Genome decay and host adaptation

There is increasing evidence that in some pathogenic microbial species loss of gene function, or genome decay, increases with adaptation to the host. For example, many pseudogenes, often ignored as sequencing artefacts, may be remnants of functional genes from a host-adapted pathogen in the process of downsizing its genome content. This is illustrated by analysis of the genome sequence of the OBLIGATE INTRACELLULAR PATHOGEN *Rickettsia prowazekii*[30]. The *R. prowazekii* (1.11 Mb) genome is packed with pseudogenes, but also has the highest proportion of non-coding DNA in a prokaryote (>24% of total DNA is intergenic). In other rickettsial species, genome downsizing seems to be an ongoing process with the emergence of multiple mutations in each of several genes[31]. On the basis of these observations, the high proportion of intergenic DNA may be the scattered remnants of genes lost in a stepwise process. As the organism acquires an obligate intracellular lifestyle such genes are no longer required — a process of reductive convergent evolution caused by prolonged intracellular life. The *R. prowazekii* genome sequence also shows surprising similarity to mitochondrial genes, indicating a very ancient ancestral link between rickettsia and mitochondria[30,31].

Analysis of the genome of the leprosy bacillus *Mycobacterium leprae* has revealed numerous pseudogenes and extensive genetic downsizing not found in the related organism *M. tuberculosis*[32]. It has been proposed that *M. leprae* has evolved to have the natural minimal gene-set for *Mycobacteria* and is a pathogen on the brink of survival[32]. In contrast to *M. tuberculosis*, *M. leprae* has a very limited metabolic repertoire and limited host range. Genome downsizing is also connected with the observation that intracellular pathogens make extensive use of host cellular processes. For example, the obligate intracellular pathogen *Chlamydia trachomatis* (1.05 Mb) lacks many biosynthetic capabilities, but retains functions for the interconversion of metabolites obtained from the mammalian host cells[8]. Similarly, in the FASTIDIOUS INTRACELLULAR PATHOGEN *Treponema pallidum* most amino acid or cofactor biosynthesis genes are absent[33]. For *C. trachomatis*, analysis of nearest matches for individual genes reveals a wide variety of different organisms (including eukaryotes), implying that complex genetic exchanges occurred in adaptation to obligate intracellular parasitic status[8].

Yersinia pestis, the causative agent of plague, may be a pathogen whose genome is at an intermediate stage of genetic flux. There is evidence of both selective genome expansion by lateral gene transfer and the initial stages of genome downsizing by insertional and point mutations leading to non-functional genes. The acquisition of new sequence by lateral gene transfer may be counterbalanced by genome decay. The organism has acquired three plasmids, one encoding a type III secretion system that is present in other pathogenic *Yersinia* species, and two plasmids specific to *Y. pestis* containing determinants responsible for survival of the organism outside the human host. The genome of *Y. pestis* is full of insertion sequences disrupting up to 100 genes, virtually all of which are uninterrupted and functional in the closely related enteric pathogen *Yersinia pseudotuberculosis*[34]. On the basis of sequence analysis of multiple housekeeping genes, it has been proposed that current *Y. pestis* strains form a homogeneous clone whose last common ancestor existed 1,500 to 20,000 years ago[35]. No sequence diversity was found in any *Y. pestis* gene studied and all were identical or nearly identical to *Y. pseudotuberculosis* orthologues[35]. Compared with *Y. pseudotuberculosis*, *Y. pestis* has recently gained two plasmids to expand its host range and change its pathogenic potential, but through selective pressure has lost dozens of genes, many of which would be required for survival in the human gastrointestinal tract. This may explain why *Y. pestis* infection in humans occurs by subcutaneous injection or the pneumonic route of infection rather than through the gastrointestinal tract.

In general, obligate intracellular parasitic pathogens have reduced genomes compared to free-living organisms, indicating a continual selective pressure for a minimal genome. Apart from genome downsizing, the intracellular habitat of this group of organisms is likely to shield the organism from potential donors of exogenous bacterial DNA. As far as free-living organisms are concerned, *Mycoplasma genitalium* has the smallest known genome (0.58 Mb)[36]. It has unusual physiology and metabolic capacity, possibly reflecting its minimal genome content. *M. genitalium* is an important model organism for determining the minimal number of genes required for host-independent existence[37].

Phase and antigenic variation

Phase variation, the reversible high-frequency gain or loss of a phenotype resulting from changes of expression of single or multiple genes, is a common survival strategy used by bacterial pathogens[38]. Variation of surface structure by pathogens, frequently referred to as antigenic variation, is often used to avoid detection or to outpace a host's immune system[38]. In some pathogens, such variation can occur by the slipped-strand mispairing of repeat sequences during replication. Alteration of the length of these tracts within or immediately upstream of coding sequences causes the translation of the respective protein to move in and out of the correct frame, affecting the synthesis of the protein. Therefore, the phenotype of an individual bacterium within a clonally growing population can change, leading to a stable sub-population with altered properties. Genes with variation in simple repeat sequences have been termed contingency genes and the repeating unit

OBLIGATE INTRACELLULAR PATHOGEN
A pathogen that lives exclusively within the host, and depends on the host for survival.

FASTIDIOUS INTRACELLULAR PATHOGEN
A pathogen that lives within the host and has stringent growth requirements.

can vary from single nucleotides (homopolymeric tracts) to pentanucleotides[39]. Scrutiny of whole-genome sequences is required for the comprehensive identification of such repeats. Such analysis simplifies the identification and investigation of potential contingency genes, which are frequently involved in host adaptation and pathogenesis. Before sequencing of the *H. influenzae* genome, only two examples of contingency genes were known in the organism. The search for simple nucleotide repeat sequences in the *H. influenzae* genome sequence identified a further dozen potential contingency genes, four of which are involved in lipopolysaccharide biosynthesis and four more which are involved in iron uptake[3,40]. Analysis of the *H. pylori* genome sequence indicated the presence of 27 putative phase variable genes based upon the presence of simple repeats[4,5,28]. Two of these repeats were found in independent α-3-fucosyltransferase genes, which have been shown to be responsible for the variable expression of the LEWIS X AND LEWIS Y ANTIGENS on the surface of *H. pylori*[41–43]. *N. meningitidis* serotypes A and B have an unprecedented number of potentially phase variable genes[6,7]. *N. meningitidis* serotype B strain MC58 has 65 putative contingency genes, most of which encode products that seem to have recognized, host-interactive functions (for example, outer membrane proteins, pili, lipopolysaccharides and capsules)[6,44]. The genome of *N. meningitidis* serotype A strain Z2491 contains hundreds of repeat elements ranging from short homopolymeric tracts to complete gene duplications[7]. This repetitive structure probably allows for extreme genome fluidity, which could be important in antigenic variation of this human-specific pathogen.

Strikingly, the *C. jejuni* genome sequence apparently has a high level of genetic variation affecting translation of over 25 contingency genes[14]. During sequencing of the eightfold redundant shotgun library of clones, regions were identified where the sequences of otherwise identical reads varied at a single point[14]. These were mainly homopolymeric tracts of G and C that varied in length by one or more base pairs. Most of the 25 hypervariable regions group into three clusters in the genome, and these are coincident with the clusters of genes responsible for lipo-oligosaccharide (LOS) biosynthesis, capsule biosynthesis and flagellar modification[14]. One of these genes encodes a β-1,3-galactosyltransferase responsible for expression of a GM1 ganglioside (a human nerve tissue structure) mimic on *C. jejuni* LOS. An assay to detect the presence of GM1 is its ability to bind cholera toxin[17]. This assay has been used to demonstrate the on/off reversible switching of this determinant on the surface of *C. jejuni* cells[16] (FIG. 1). This study shows the power of genome sequence data to rapidly identify genes likely to be involved in host–pathogen interactions and to develop hypotheses for 'wet laboratory' research. It also adds to the growing list of contingency genes that encode proteins involved in mimicry of human host molecules (*H. influenzae* lipopolysaccharide[15], *N. meningitidis* type B capsule[6] and *H. pylori* Lewis antigens[41–43]). Why so many proteins involved in generating host-like structures are phase variable is unclear, but the rapid on/off switching of these proteins may contribute to triggering the host immune response during post-infection immune complications associat-

LEWIS ANTIGENS
Fucosylated carbohydrate antigens usually found on the surface of eukaryotic cells. They are structurally related to human ABH blood group systems.

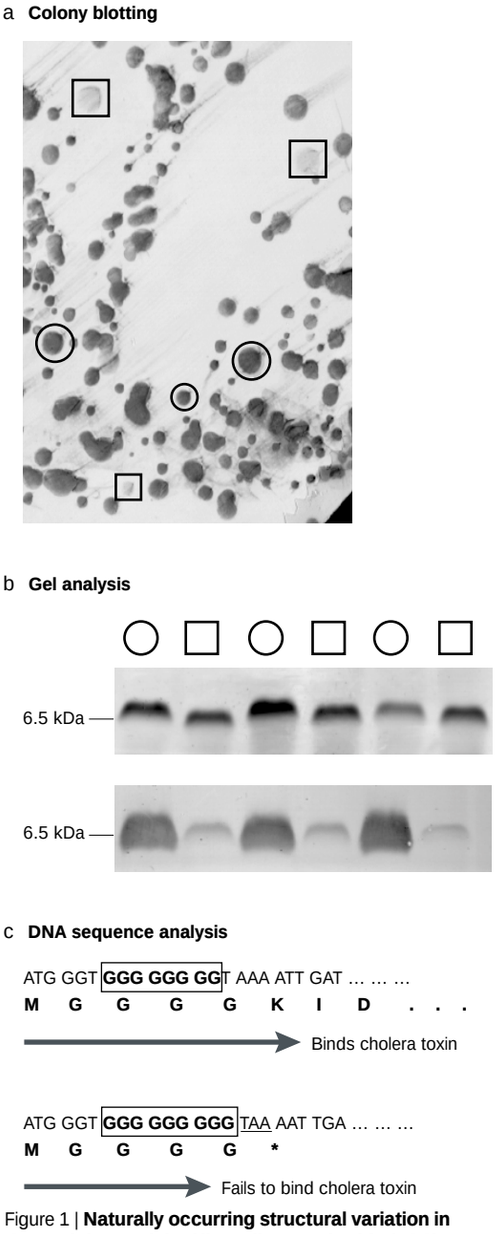


Figure 1 | **Naturally occurring structural variation in *Campylobacter jejuni* lipo-oligosaccharide (LOS) mediated by the intragenic homopolymeric G tract of the *wlaN* contingency gene[16]. a** | Colony blotting of a wild-type population of *C. jejuni* NCTC11168 with cholera toxin. **b** | Gel analysis shows increased mobility of LOS on tricine gels from colonies that fail to bind cholera toxin (upper panel squares) and reduced crossreactivity with cholera toxin of LOS from colonies that fail to bind cholera toxin (lower panel squares). (Note residual binding probably due to revertants in natural population.) **c** | DNA sequence of *wlaN* from selected colonies. Colonies that bind cholera toxin (circles in **a**), have 8 G bases; whereas colonies that fail to bind cholera toxin (squares in **a**), have 9 G bases resulting in a frameshift and termination of translation.

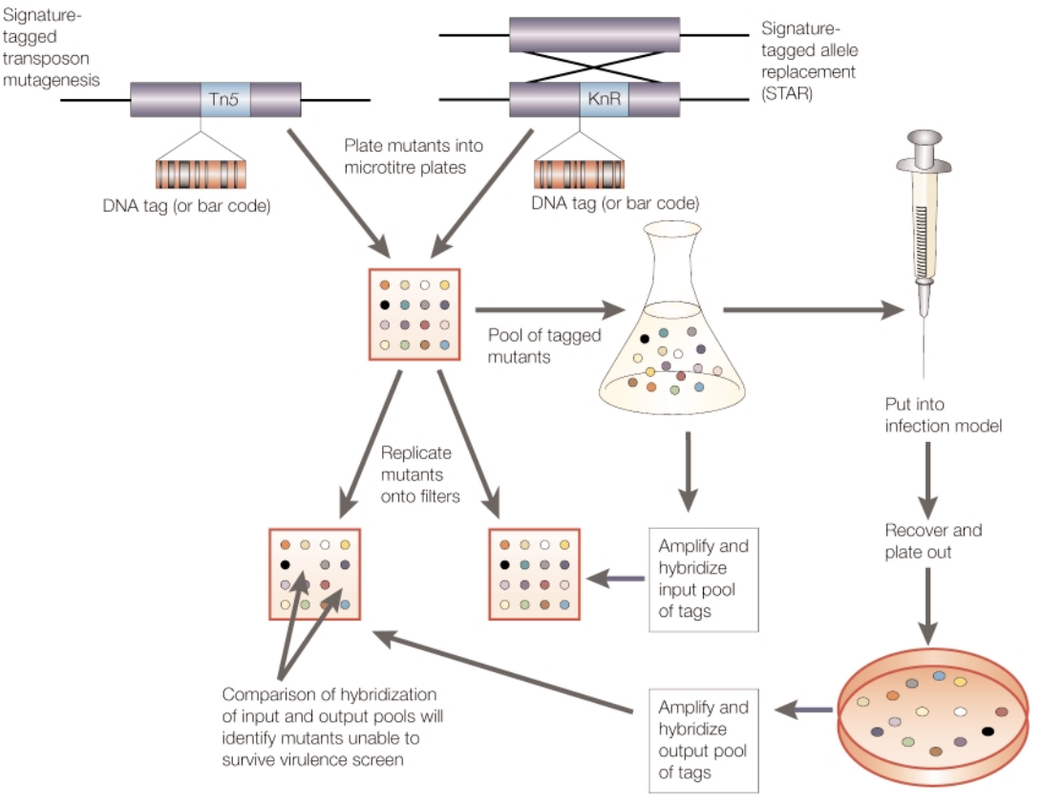


Figure 2 | **Signature-tagged mutagenesis — high-throughput analysis of *in vivo*-expressed genes.** Mutants generated by transposons or allele replacement using DNA tagged transposon 5 (Tn*5*) or kanamycin antibiotic resistance cassette (KnR), respectively. Input and output pools of mutants are generated and compared to identify mutants unable to survive *in vivo*[49].

ed with some of these pathogens.

In addition to genes associated with the synthesis of cell surface structures, contingency genes are found in genes that encode DNA restriction/modification systems (for example, DNA methyltransferase[45]), indicating that the uptake and restriction of DNA may phase vary in *H. influenzae*[3,40], *N. meningitidis*[6], *H. pylori*[4,28] and *C. jejuni*[14]. Contingency genes seem to be almost exclusively associated with MUCOSAL PATHOGENS with relatively small genome sequences (*H. influenzae*, *N. meningitidis*, *N. gonorrhea*, *H. pylori*, *C. jejuni* and *Campylobacter coli*[46]). So slipped-strand mis-pairing may be a simple or even primordial mechanism for making changes in a small genetic repertoire among selected groups of pathogens.

Gene duplication is another mechanism used by microbial pathogens for generating variation in surface structure. The genome sequence of *H. pylori* encodes a family of over 30 paralogous outer membrane proteins that may play a role in generating antigenic variation[4]. These are absent in the related *C. jejuni* genome sequence. Similarly, genome analysis of *M. tuberculosis* revealed two unusual families of glycine-rich proteins with repetitive structure constituting 10% of the genome[13]. These gene families may represent a source of antigenic variation for *M. tuberculosis*[13]. The biological significance of the multiple plasmid-encoded genes found in *B. burgdorferi* B31 is not clear, but it has been postulated that they may be involved in antigenic variation or immune evasion[21].

Antimicrobial and vaccine targets. Whole-genome sequences will provide unprecedented opportunities for vaccine design, as the complete inventory of genes encoding every virulence factor and potential immunogen will be available for selection. DNA itself is an attractive reagent for immunization strategies as it is stable and easy to produce. An approach to identifying candidates for DNA vaccines is to immunize animals with pools of test DNA fragments and then re-examine the positive pools to identify the DNA molecule that elicited the most potent immune response. This method, called expression library screening or genomic vaccination, has been tested for the mouse pathogen *Mycoplasma pulmonis*, and through the availability of whole genome sequences offers a systematic unbiased approach to the discovery of vaccine candidates[47].

Another systematic approach has been applied to identify seven new *N. meningitidis* vaccine targets directly from the genome sequence through a combination of bioinformatics, high-throughput expression and immunological screening[48]. From the 2,158 predicted coding sequences of the *N. meningitidis* type B MC58 genome, 570 encoding potential surface antigens were selected for expression in *E. coli*. These included genes encoding predicted membrane pro-

MUCOSAL PATHOGEN
A pathogen that frequents the mucosal surface (for example, nose, lungs and gastrointestinal tract) of the host.

Box 2 | Functional genomics of microbial pathogens

Potential applications of microarrays

- Differential gene expression (DGE) by hybridizing mRNA extracted under varying environmental conditions.
- DGE comparing mutant to the wild-type strains, particularly to decipher regulatory networks.
- Genomotyping — hybridization of the DNA content of a test strain (for example, clinical or environmental isolates) against the genome of a sequenced strain arrayed on the solid support. An example is the whole-genome comparison of *M. tuberculosis* with Bacille Calmette-Guérin (BCG)[78].
- Testing genome plasticity of an individual strain by DNA hybridization.

Potential applications of proteomics

- Characterization of genetic regulatory pathways modulated by a myriad of environmental signals.
- Study of post-translational modifications.
- Study of protein complexes.
- Identification of immunogenic proteins by immunoblotting — useful for identifying vaccine targets.
- Determination of mechanisms of drug action and identifying drug targets.

teins, proteins with signal sequences and lipoproteins. Genes likely to be involved in antigenic variation, such as contingency genes and genes with homology to human DNA sequences, were excluded. Of the 570 genes targeted, 350 were successfully expressed as fusion proteins in *E. coli* and were selected to immunize mice. On the basis of strong antibody responses, 85 fusion proteins were tested for their ability to induce a bactericidal antibody response *in vitro*, a property known to correlate with vaccine efficacy in humans. Further criteria, such as conservation among the most prevalent *N. meningitidis* serogroups, narrowed down the list to seven surface-exposed antigens. These are now being evaluated as vaccine candidates against type B meningitis. This brute-force approach was initiated as the genome sequencing of *N. meningitidis* type B MC58 was in progress and demonstrates the value of bioinformatics in rational experimental design[6,48]. Finally, genome sequences can be used to generate attenuated microbial pathogens, which are vaccine candidates in their own right, and can also be used as vectors for vaccine delivery. Attenuation can be approached at the genome level by identifying and deleting genes or sets of genes required for virulence to produce optimally attenuated, safe and inexpensive vaccines.

Until very recently, antimicrobial drug discovery remained a largely random process. The identification of virulence factors from genome sequences reveals drug targets that might disable a pathogen. Comparative analysis of sequenced microbial pathogen genomes can identify genes that are conserved in genomes of phylogenetically diverse pathogens and so are likely to be essential. Such genes could provide targets for new broad-spectrum antimicrobial agents. Genomics also allows screening for genes that are conserved in all or most pathogens, but which are absent in eukaryotes. Gene products from such genes are attractive targets because drugs targeted against them are less likely to be toxic to humans.

The birth of functional genomics

The acquisition and analysis of genome sequence data is not an end in itself, but a starting point for generating testable hypotheses on pathogenesis. Homology provides clues, but does not establish gene function. Furthermore, a large percentage of putative genes from microbial pathogens have no matches to known genes. The avalanche of genome sequence data has coincided with important technological progress in three 'wet laboratory' research areas: gene mutagenesis, nucleic acid hybridization technology and protein chemistry. Spectacular advances in computation and bioinformatics have also occurred, which are vital to the success of functional genomics approaches. These advances will redirect experimental strategies from the piecemeal study of individual genes or operons towards comprehensive analyses of the entire gene and protein complement of the microbial cell — a complete approach to the functional characterization of microbial pathogens at the mutational, transcriptional and protein expression levels.

Mutational analysis. The construction of defined mutants by transposon mutagenesis or allelic replacement has proved to be a powerful method for determining gene function and for dissecting out virulence determinants in numerous microbial pathogens. An extension of this approach is to label each mutant with a unique DNA signature tag, which permits simultaneous analysis of several hundred mutants for phenotypic features[49]. By negative selection, mutants that fail to be recovered from the host following inoculation of a mixed pool of mutants can be identified and so the host is used as a high-throughput screen for the *in vivo* fitness of mutants (FIG. 2). This approach greatly reduces the number of animal experiments required for the assessment of microbial virulence. The use of DNA signature tags was validated using *Salmonella* and a mouse model of typhoid fever, a strategy that identified a second type III secretion system as well as other potential virulence determinants[49,50]. Signature-tagged mutagenesis has since been successfully used in the identification of virulence-associated factors in several bacterial pathogens including *S. aureus*, *V. cholerae*, *N. meningitidis*, *Streptococcus pneumoniae*, *Legionella pneumophila*, *Yersinia enterocolitica*, *Proteus mirabilis*, *M. tuberculosis* and *Brucella suis*[51–58]. Signature-tagged mutagenesis is particularly useful for pathogens whose genomes have been sequenced, as this aids determination of the location of the transposon insertion site.

Because transposons often fail to integrate, or integrate non-randomly into the chromosome of many bacteria, transposon mutagenesis is not universally applicable. Allelic replacement is a useful alternative for the construction of defined deletion mutants. Furthermore, the era of complete genomic sequences means that the large-scale systematic construction of defined mutants over the entire genome is now possible. This approach is being attempted for *Bacillus subtilis*[59] and *Saccharomyces cerevisiae*[60] and is planned for several bacterial

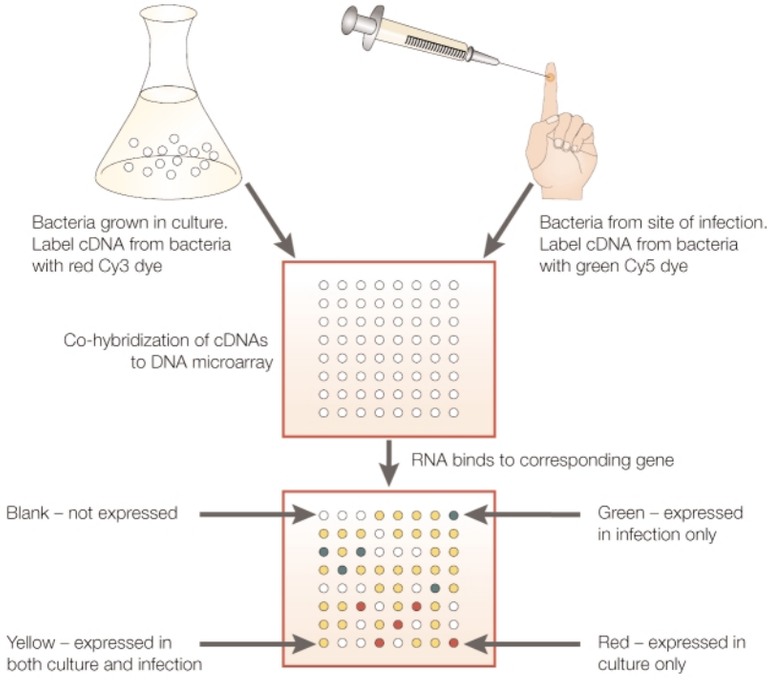


Figure 3 | **Principles of transcriptome analysis using a DNA microarray.** To measure relative differences in gene expression, sample DNA or RNA is labelled (normally by fluorescence) and hybridized to the array. For example, mRNA from cells grown under standard culture conditions is labelled with the red fluorescent dye Cy3 and sample mRNA from a site of infection is labelled with the green fluorescent dye Cy5. After co-hybridization to the microarray, the fluorescence intensity of each spot on the array is read to determine the relative abundance of mRNA from the two test conditions. A red signal indicates gene expression only in cells grown in culture, a green signal indicates gene expression only during infection and a yellow signal indicates genes expressed in both conditions.

pathogens. The coupling of DNA tags with allelic replacement has been referred to as signature-tagged allele replacement (STAR) (FIG. 2). STAR does not require the use of transposons, but enables a systematic unbiased genetic analysis of the genome. A further benefit of this approach is the ability to ascertain which genes are essential for the viability of the organism. Those genes that are refractory to mutagenesis are likely to be important for the viability of the organism, and so their gene products are potential antimicrobial drug targets.

Transcriptome analysis. Transcriptome analysis offers the potential for the simultaneous measurement of expression levels for all transcripts (messenger RNAs) from a genome. This gives a 'snapshot' of the transcriptional activity of all genes in a genome, which can be assayed at a particular time-point in growth or in any environment[61]. Given the inhospitable *in vivo* and the varied *ex vivo* environments encountered by most microbial pathogens, transcriptome analysis holds particular promise for identifying and determining the functions of differentially regulated, virulence-associated genes. The technique involves extracting the mRNA expressed under a range of environmental conditions and hybridizing these sequences to a high-density gridded microarray of the DNA content of an organism. Such high-throughput analysis allows massive parallel gene expression and gene discovery studies to be undertaken[61] (FIG. 3, BOX 2). DNA microarray analysis will complement other technologies such as *in vivo* expression technology and differential fluorescence analysis to identify and investigate which bacterial genes are differentially expressed in the host[62–64].

Proteome analysis. The case for global monitoring of mRNA applies equally to proteins, with the added advantage that post-translational modifications, which frequently play key roles in host–pathogen interactions, can also be studied. Proteomics, the study of the complete set of proteins that is expressed and modified in a cell, is an important and rapidly evolving discipline that is readily applicable to microbial pathogens.

Recent improvements in high-sensitivity biological mass spectrometry have provided a powerful adjunct to traditional 2D SDS–PAGE gel electrophoresis[65,66]. Proteins cut out of a 2D SDS–PAGE gel can be analysed by mass spectrometry. The data can be used to find the best match in a sequence database, making it possible to go from a spot on a 2D SDS–PAGE gel to protein identification in a matter of hours. So the entire complement of soluble proteins expressed by a cell (the proteome) can be defined. This kind of approach has already been used to provide insights into the function of a specific subset of the proteome such as the cell envelope for *Salmonella typhimurium*[67]. Proteome studies are made even more powerful when applied to an organism whose genome has been sequenced. For example, ORFan or FUN gene products can be identified as functional proteins. Several proteome studies are now underway for bacterial pathogens including *M. genitalium*, *S. typhimurium*, *M. tuberculosis*, *H. pylori*, *S. aureus* and *Pseudomonas aeruginosa*[68–73] (BOX 2, LINKS).

Knowledge of the interactions between proteins is also valuable for predicting protein function. A variety of biochemical assays have been developed to measure such interactions. A widely used method is the yeast two-hybrid system — a procedure to identify proteins that bind to a protein of interest, or to define domains or residues critical for an interaction. Protein–protein interaction maps of selected components of microbial pathogens should provide invaluable data on structural components of the cell. This has been applied to some microbial pathogens to study outer membrane proteins and type III secretion systems[74–76]. Genome-wide analysis has now been attempted on *H. pylori*, identifying 1,200 protein–protein interactions linking >800 proteins[77]. Further genome-wide studies are underway for *S. aureus* and *S. pneumoniae*, paving the way for comparative proteome analysis of microbial pathogens (LINKS).

Conclusions

Analysis of whole genome sequences from a range of pathogens shows the diversity and adaptability of this specialized group of organisms. Scrutiny of sequence data has provided much evidence of lateral gene transfer, genome decay and variation of gene expression among different microbial pathogens. This supports

an evolutionary scenario involving vertical diversification by mutagenesis, punctuated by frequent lateral gene transfer resulting in a global mosaic genome structure.

Surprisingly, the function of the vast majority of genes specific to microbial pathogens remains to be determined. A multidimensional analysis, looking at genome sequences, mutants, transcripts and proteins, will greatly advance our understanding of the biology of microbial pathogens. With judicious choice of stimuli and detection assays, the regulation and activity of major sets of genes and proteins, and the interactions between them, can now be rapidly determined, taking an important step towards a complete functional understanding of these problematic microorganisms. This understanding should revolutionize microbial research as did the discovery of the microscope.

A current stumbling block to reaping the benefits of the post-genome era is the insufficiency of bioinformatic tools to analyse genome sequence data and DNA microarray data. In the future, sustained improvements in software, computing speed and information storage will considerably increase the scale of problems we tackle to understand the biology and evolution of microbial pathogens. The development of a database of nucleotide differences among strains should allow the design of a universal microbial pathogen microarray (or biochip), which would have wide applications in studying the epidemiology, population genetics, molecular phylogeny and evolution of microbial pathogens, as well as diagnostic applications. Finally, the availability of the human genome sequence will enable scientists to understand the dynamic and complex interactions between host and pathogen, providing further scope for the rational design of intervention strategies to reduce the burden of microbial-related disease.

Links

GENOME PROJECTS *Borrelia burgdorferi* B31 | *Campylobacter jejuni* NCTC11168 | *Chlamydia pneumoniae* CWL029 | *Chlamydia pneumoniae* AR39 | *Chlamydia trachomatis* D/UW-3/Cx | *Chlamydia trachomatis* MoPn | *Haemophilus influenzae* Rd | *Helicobacter pylori* 26695 | *Helicobacter pylori* J99 | *Listeria monocytogenes* EGD-e | *Mycobacterium leprae* | *Mycobacterium tuberculosis* H37Rv | *Mycoplasma genitalium* G-37 | *Mycoplasma pneumoniae* M129 | *Neisseria meningitidis* A Z2491 | *Neisseria meningitidis* B MC58 | *Rickettsia prowazekii* Madrid E | *Streptococcus pyogenes* M1 | *Treponema pallidum* Nichols | *Vibrio cholerae* N16961 | *Saccharomyces cerevisiae* | *Bacillus subtilis* | *Yersinia pestis* | *Escherichia coli*

FURTHER INFORMATION TIGR | Genomes online | Sanger Centre | Pasteur Institute | Los Alamos sexually transmitted pathogens database | Genomics: a global resource | Kyoto encyclopaedia of genes and genomes | Base-peak mass spectrometry | *E. coli* proteome | Australian proteome analysis facility | The Brown lab microarray guide | Bacterial microarrays at St. George's | Virtual genome centre | MIPS genome annotation | Sanger sequence viewer | Wren laboratory homepage

ENCYCLOPEDIA OF LIFE SCIENCES Immune mechanisms against extracellular pathogens | Antimicrobial resistance: control

Update — added in proof

The genome sequence of the opportunistic pathogen *Pseudomonas aeruginosa* was recently published[81]. At 6.3 Mb, *P. aeruginosa* has the largest bacterial genome sequenced to date. Consistent with its environmental adaptability, it contains an excess of genes for solute transport and regulation.The genome sequence has also allowed essential *P. aeruginosa* genes to be identified using the mariner transposon and genetic footprinting[82].

1. Blaser, M. J. Linking *Helicobacter pylori* to gastric cancer. *Nature Med.* **6**, 376–377 (2000).
2. Rosenfeld, M. E. *et al.* Chlamydia, inflammation, and atherogenesis. *J. Infect. Dis.* **181**, S492–S497 (2000).
3. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
First sequencing of a free-living organism and first description of use of the shotgun strategy for whole genome sequencing.
4. Tomb, J. F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547 (1997).
5. Alm, R. A. *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176–180 (1999).
Sequence determination of a second strain within a species. Identification of a genome plasticity zone.
6. Tettelin, H. *et al.* Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287**, 1809–1815 (2000).
7. Parkhill, J. *et al.* Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502–506 (2000).
8. Stephens, R. S. *et al.* Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**, 754–759 (1998).
9. Read, T. D. *et al.* Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* **28**, 1397–1406 (2000).
10. Glaser, P. & Cossart, P. The *Listeria monocytogenes* genome project. *Genomes 2000: International conference on microbial and model genomes* **20** (2000).
11. Fraser, C. M. & Fleischmann, R. D. Strategies for whole microbial genome sequencing and analysis. *Electrophoresis* **18**, 1207–1216 (1997).
12. Frangeul, L. *et al.* Cloning and assembly strategies in microbial genome projects. *Microbiology* **145**, 2625–2634 (1999).
13. Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
14. Parkhill, J. *et al.* The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–668 (2000).
Sequencing *Campylobacter jejuni* allowed the direct identification of contingency genes from the shotgun sequence.
15. Hood, D. W. *et al*. Use of the complete genome sequence information of Haemophilus influenzae strain Rd to investigate lipopolysaccharide biosynthesis. *Mol. Microbiol*. **22**, 951–965 (1996).
First example of the exploitation of genome sequence data to investigate the biology of a pathogenic organism.
16. Linton, D. *et al.* Phase variation of a β-1,3 galactosyltransferase involved in generation of the ganglioside GM1-like lipo-oligosaccharide of *Campylobacter jejuni*. *Mol. Microbiol.* **37**, 501–515 (2000).
17. Linton, D. *et al.* Multiple *N*-acetyl neuraminic acid synthetase (*neuB*) genes in *Campylobacter jejuni*: identification and characterization of the gene involved in sialylation of lipo-oligosaccharide. *Mol. Microbiol.* **35**, 1120–1134 (2000).
18. Karlyshev, A. V., Linton, D., Gregson, N. A., Lastovica, A. J. & Wren, B. W. Genetic and biochemical evidence of a *Campylobacter jejuni* capsular polysaccharide that accounts for Penner serotype specificity. *Mol. Microbiol.* **35**, 529–541 (2000).
19. Wren, B. W. *et al.* Characterization of a haemolysin from *Mycobacterium tuberculosis* with homology to a virulence factor of *Serpulina hyodysenteriae*. *Microbiology* **144**, 1205–1211 (1998).
20. Heidelberg, J. F. *et al.* DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*, *Nature* **406**, 477–483 (2000).
21. Fraser, C. M. *et al.* Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580–586 (1997).
22. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474 (1997).
23. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
24. Hacker, J., Blum-Oehler, G., Muhldorfer, I. & Tschape, H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.* **23**, 1089–1097 (1997).
25. Cheng, L. W. & Schneewind, O. Type III machines of Gram-negative bacteria: delivering the goods. *Trends Microbiol.* **8**, 214–220 (2000).
26. Galan, J. E. & Collmer, A. Type III secretion machines: bacterial devices for protein delivery into host cells. *Science* **284**, 1322–1328 (1999).
27. Covacci, A., Telford, J. L., Del Giudice, G., Parsonnet, J. & Rappuoli, R. *Helicobacter pylori* virulence and genetic geography. *Science* **284**, 1328–1333 (1999).
28. Saunders, N. J., Peden, J. F., Hood, D. W. & Moxon, E. R. Simple sequence repeats in the *Helicobacter pylori* genome. *Mol. Microbiol.* **27**, 1091–1098 (1998).
29. Ferretti, J. J. *et al.* The complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Genomes 2000: International conference on microbial and model genomes* **21** (2000).
30. Andersson, S. G. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140 (1998).
Analysis of the genome sequence of the obligate intracellular pathogen *Rickettsia prowazekii* revealed extensive genome downsizing and the possible origins of mitochondria.
31. Andersson, J. O. & Andersson, S. G. E. A century of typhus, lice and *Rickettsia*. *Res. Microbiol.* **151**, 143–150 (2000).
32. Brosch, R., Gordon, S. V., Eiglmeier, K., Garnier, T. & Cole, S. T. Comparative genomics of the leprosy and

tubercle bacilli. *Res. Microbiol.* **151**, 135–142 (2000).
33. Fraser, C. M. *et al.* Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375–388 (1998).
34. Buchrieser, C. *et al.* The 102-kilobase pgm locus of *Yersinia pestis:* sequence analysis and comparison of selected regions among different *Yersinia pestis* and *Yersinia pseudotuberculosis* strains. *Infect. Immunol.* **67**, 4851–4861 (1999).
35. Achtman, M. *et al. Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc. Natl Acad. Sci. USA* **96**, 14043–14048 (1999).
36. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
37. Hutchison, C. A. *et al.* Global transposon mutagenesis and a minimal *Mycoplasma genome*. *Science* **286**, 2165–2169 (1999).
38. Henderson, I. R., Owen, P. & Nataro, J. P. Molecular switches—the ON and OFF of bacterial phase variation. *Mol. Microbiol.* **33**, 919–932 (1999).
39. Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4**, 24–33 (1994).
40. Hood, D. W. *et al.* DNA repeats identify novel virulence genes in *Haemophilus influenzae*. *Proc. Natl Acad. Sci. USA* **93**, 11121–11125 (1996).
Demonstration of the rapid identification of repeat sequences for the convenient identification of new virulence determinants.
41. Wang, G., Rasko, D. A., Sherburne, R. & Taylor, D. E. Molecular genetic basis for the variable expression of Lewis Y antigen. in *Helicobacter pylori*: analysis of the alpha (1,2) fucosyltransferase gene. *Mol. Microbiol.* **31**, 1265–1274 (1999).
42. Appelmelk, B. J. *et al.* Phase variation in *Helicobacter pylori* lipopolysaccharide due to changes in the lengths of poly(C) tracts in alpha3-fucosyltransferase genes. *Infect. Immunol.* **67**, 5361–5366 (1999).
43. Wang, G., Ge, Z., Rasko, D. A. & Taylor, D. E. Lewis antigens in *Helicobacter pylori:* biosynthesis and phase variation. *Mol. Microbiol.* **36**, 1187–1196 (2000).
44. Saunders, N. J. *et al.* Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitis* strain MC58. *Mol. Microbiol.* **37**, 207–215 (2000).
45. De Bolle, X. *et al.* The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. *Mol. Microbiol.* **35**, 211–222 (2000).
46. Park, S. F., Purdy, D. & Leach, S. Localized reversible frameshift mutation in the flhA gene confers phase variability to flagellin gene expression in *Campylobacter coli*. *J. Bacteriol.* **182**, 207–210 (2000).
47. Barry, M. A., Lai, W. C. & Johnston, S. A. Protection against mycoplasma infection using expression-library immunization. *Nature* **377**, 632–635 (1995).
48. Pizza, M. *et al.* Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287**, 1816–1820 (2000).
Remarkable demonstration of the use of whole-genome sequence data from an important pathogen *Neisseria meningitidis* to identify seven conserved surface antigens as vaccine candidates.
49. Hensel, M. *et al.* Simultaneous identification of bacterial virulence genes by negative selection. *Science* **269**, 400–403 (1995).
First description of signature-tagged mutagenesis which allowed the identification of a second type III secretion system (spill) in *Salmonella typhimurium*.
50. Shea, J. E., Hensel, M., Gleeson, C. & Holden, D. W. Identification of a virulence locus encoding a second type III secretion system in *Salmonella typhimurium*. *Proc. Natl Acad. Sci. USA* **93**, 2593–2597 (1996).
51. Mei, J. M., Nourbakhsh, F., Ford, C. W. & Holden, D. W. Identification of *Staphylococcus aureus* virulence genes in a murine model of bacteraemia using signature-tagged mutagenesis. *Mol. Microbiol.* **26**, 399–407 (1997).
52. Chiang, S. L. & Mekalanos, J. J. Use of signature-tagged transposon mutagenesis to identify *Vibrio cholerae* genes critical for colonization. *Mol. Microbiol.* **27**, 797–805 (1998).
53. Claus, H., Frosch, M. & Vogel, U. Identification of a hotspot for transformation of *Neisseria meningitidis* by shuttle mutagenesis using signature-tagged transposons. *Mol. Gen. Genet.* **259**, 363–371 (1998).
54. Polissi, A. *et al.* Large-scale identification of virulence genes from *Streptococcus pneumoniae*. *Infect. Immunol.* **66**, 5620–5629 (1998).
55. Edelstein, P. H., Edelstein, M. A., Higa, F. & Falkow, S. Discovery of virulence genes of *Legionella pneumophila* by using signature tagged mutagenesis in a guinea pig pneumonia model. *Proc. Natl Acad. Sci. USA* **96**, 8190–8195 (1999).
56. Darwin, A. J. & Miller, V. L. Identification of *Yersinia enterocolitica* genes affecting survival in an animal host using signature-tagged transposon mutagenesis. *Mol. Microbiol.* **32**, 51–62 (1999).
57. Camacho, L. R., Ensergueix, D., Perez, E., Gicquel, B. & Guilhot, C. Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol. Microbiol.* **34**, 257–267 (1999).
58. Zhao, H., Li, X., Johnson, D. E. & Mobley, H. L. Identification of protease and rpoN-associated genes of uropathogenic *Proteus mirabilis* by negative selection in a mouse model of ascending urinary tract infection. *Microbiology* **145**, 185–195 (1999).
59. Ogasawara, N. Systematic functional mutagenesis of *Bacillus subtilis* genes. *Res. Microbiol.* **152**, 129–134 (2000).
60. Winzeler, E. A. *et al.* Functional characterization of the *Saccharomyces cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
61. Brown, P. O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21**, 33–37 (1999).
62. Mahan, M.J., Slauch, J. M. & Mekalanous, J. J. Selection of bacterial virulence genes that are specifically induced in host tissues. *Science* **259**, 686–688 (1993).
63. Valdivia, R. H. & Falkow, S. Fluorescence-based isolation of bacterial genes expressed within host cells. *Science* **277**, 2007–2011 (1993).
64. Strauss, E. J., Falkow, S. Microbial pathogenesis: genomics and beyond. *Science* **276**, 707–712 (1997).
65. Pappin, D. J. Peptide mass fingerprinting using MALDI-TOF mass spectrometry. *Methods Mol. Biol.* **64**, 165–173 (1997).
66. Fernandez, J., Gharahdaghi, F. & Mische, S. M. Routine identification of proteins from sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) gels or polyvinyl difluoride membranes using matrix assisted laser desorption/ionization-time of flight-mass spectrometry (MALDI-TOF-MS). *Electrophoresis* **19**, 1036–1045 (1998).
67. Qi, S. Y., Moir, A. & O'Connor, C. D. Proteome of *Salmonella typhimurium* SL1344: identification of novel abundant cell envelope proteins and assignment to a two-dimensional reference map. *J. Bacteriol.* **178**, 5032–5038 (1996).
68. Wasinger, V. C., Pollack, J. D. & Humphery-Smith, I. The proteome of *Mycoplasma genitalium* Chaps-soluble component. *Eur. J. Biochem.* **267**, 1571–1582 (2000).
69. O'Connor, C. D., Farris, M., Fowler, R. & Qi, S. Y. The proteome of *Salmonella enterica* serovar *typhimurium*: current progress on its determination and some applications. *Electrophoresis* **18**, 1483–1490 (1997).
70. Sonnenberg, M. G. & Belisle, J. T. Definition of Mycobacterium tuberculosis culture filtrate proteins by two-dimensional polyacrylamide gel electrophoresis, N-terminal amino acid sequencing, and electrospray mass spectrometry. *Infect. Immunol.* **65**, 4515–4524 (1997).
71. Jungblut, P. R. *et al.* Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. *Mol. Microbiol.* **33**, 1103–1117 (1999).
72. Tekaia, F. *et al.* Analysis of the proteome of *Mycobacterium tuberculosis in silico*. *Tuber. Lung Dis.* **79**, 329–342 (1999).
73. Jungblut, P. R. *et al.* Comparative proteome analysis of *Helicobacter pylori*. *Mol. Microbiol.* **36**, 710–725 (2000).
74. Williams, J. M., Chen, G. C., Zhu, L. & Rest, R. F. Using the yeast two-hybrid system to identify human epithelial cell proteins that bind gonococcal Opa proteins: intracellular gonococci bind pyruvate kinase via their Opa proteins and require host pyruvate for growth. *Mol. Microbiol.* **27**, 171–186 (1998).
75. Hartland, E. L. *et al.* Binding of intimin from enteropathogenic *Escherichia coli* to Tir and to host cells. *Mol. Microbiol.* **32**, 151–158 (1999).
76. Day, J. B. & Plano, G. V. A complex composed of SycN and YscB functions as a specific chaperone for YopN in *Yersinia pestis*. *Mol. Microbiol.* **30**, 777–788 (1998).
77. Rain, J. C. & Legrain, P. Functional proteomics on microbial pathogens. *Genomes 2000: International conference on microbial and model genomes* **26** (2000).
78. Behr, M. A. *et al.* Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**, 1520–1523 (1999).
79. Kalman, S. *et al.* Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature Genet.* **21**, 385–389 (1999).
80. Himmelreich, R. *et al.* Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**, 4420–4449 (1996).
81. Stover, C. K. *et al.* Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406**, 959–964 (2000).
82. Wong, S. M. & Mekalanos, J. J. Genetic footprinting with mariner-based transposition in *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA* **97**, 10191–10196 (2000).

Acknowledgements

Work in the author's laboratory is supported by the BBSRC, the Wellcome Trust and Beowulf Genomics. I acknowledge Dennis Linton and Elaine Allan for helpful comments and a critical review of the manuscript.