Bacterial pathogenomics

Mark J. Pallen[1] & Brendan W. Wren[2]

Genomes from all of the crucial bacterial pathogens of humans, plants and animals have now been sequenced, as have genomes from many of the important commensal, symbiotic and environmental microorganisms. Analysis of these sequences has revealed the forces that shape pathogen evolution and has brought to light unexpected aspects of pathogen biology. The finding that horizontal gene transfer and genome decay have key roles in the evolution of bacterial pathogens was particularly surprising. It has also become evident that even the definitions for 'pathogen' and 'virulence factor' need to be re-evaluated.

The sequencing of bacterial genomes (see Glossary) has occurred against the backdrop of an established programme of research on bacterial pathogenesis. Nonetheless, it has uncovered aspects of pathogen biology that were unexpected before the genomic revolution. Here, we examine the 'creative clash' between genomic research and bacterial pathogenesis research, an encounter that has spawned new technologies and new avenues for applied research. In addition, we discuss the forces that have shaped the evolution of bacterial pathogens, and we reappraise human–pathogen interactions in the light of bacterial ecology and evolution.

Genome dynamics

At the start of the genomic era, each project to sequence a bacterial genome was viewed as equivalent in difficulty to a Moon landing. However, the cutting edge soon shifted to determining the genome sequences of multiple strains in each species[1–4], heralding a transformation in our view of bacterial diversity. Comparisons between the genomes of related strains and species of bacterial pathogens, across the whole range of taxonomic variation, have made it clear that a 'one size fits all' approach cannot be applied to the evolutionary dynamics of bacterial virulence[5–7]. Instead, different evolutionary processes predominate in different taxonomic groups.

Three main forces have been found to shape genome evolution: gene gain, gene loss and gene change (that is, any changes that affect the sequences or order of the existing genes) (Fig. 1). In the genome of some bacterial pathogens (for example, *Yersinia pestis*[8]), all three are evident. In addition, differences in the scale and the timing of these changes in different lineages of bacterial pathogens have resulted in at least three main patterns of genome dynamics. First, some genetically uniform lineages, which are also usually reproductively isolated, have emerged recently in evolutionary terms (for example, *Bacillus anthracis* and *Mycobacterium leprae*). Second, recombination can occur between closely related sequences in closely related strains; this is common in naturally competent mucosal pathogens (for example, *Neisseria meningitidis*, *Haemophilus influenzae* and *Streptococcus pneumoniae*). Third, widespread horizontal gene transfer, bringing in new sequences, predominates in certain pathogens (for example, many enterobacteria, and some staphylococci and streptococci).

The smallest-scale variation in bacterial genomes occurs at the level of single-nucleotide polymorphisms (SNPs). SNP detection has been applied extensively to recently emerged genetically uniform pathogens, such as *M. leprae*, *Mycobacterium tuberculosis*, *Y. pestis and B. anthracis* (the last driven by forensic considerations after the anthrax attacks of 2001 in the United States)[9–12]. More recently, whole-genome sequencing has been used to detect SNPs in more variable species, such as *Escherichia coli* and *Francisella tularensis*[13–15]. This approach to SNP detection enabled *E. coli* strains that had diverged for as few as 200 generations to be differentiated[16] and revealed genomic changes in *Burkholderia mallei* after accidental human infection[17]. These studies indicate that the use of whole-genome sequencing could soon become a routine epidemiological tool in bacteriology, as it already is in virology (Box 1).

Genome sequencing has also confirmed that phase variation is a widespread source of intraspecific genotypic and phenotypic variation[18,19]. Several mutational mechanisms are exploited by bacteria to switch gene and/or protein expression on or off. For example, in *Campylobacter jejuni*, the presence of several tens of homopolymeric nucleotide repeat sequences can lead to slippage during DNA replication, resulting in a varied repertoire of structures exposed on the bacterial-cell surface[20]. By contrast, *Bacteroides fragilis* uses DNA inversion to modulate more than 20 genetic loci, which contain genes that encode bacterial surface proteins, polysaccharides and components of regulatory systems[21]. The combinatorial mathematics of phase variation mean that a bacterium with just 20 phase-variable loci can exist in 2^{20} (that is, more than a million) different states.

Horizontal gene transfer

The greatest surprise resulting from the application of genomics to bacteriology was the extent of genomic variability within many bacterial species. Two *E. coli* strains can differ by as much as a quarter

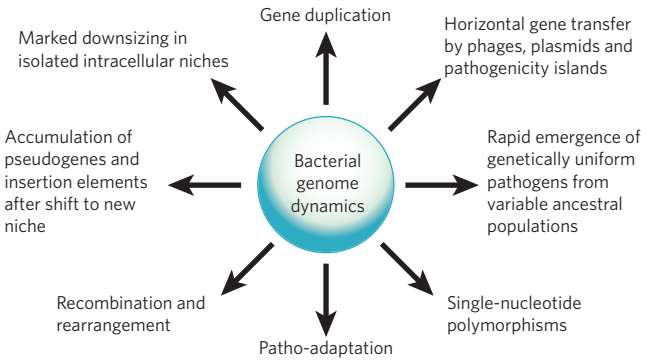

Figure 1 | Bacterial genome dynamics. There are three main forces that shape bacterial genomes: gene gain, gene loss and gene change. All three of these can take place in a single bacterium. Some of the changes that result from the interplay of these forces are shown.

[1]Centre for Systems Biology, University of Birmingham, Birmingham B15 2TT, UK. [2]Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

Box 1 | Technologies and applications in pathogenomics

In 1995, J. Craig Venter and colleagues showed that genomes could be sequenced efficiently by using a whole-genome shotgun approach, facilitated by computational sequence assembly and laboratory-based finishing and closure approaches[85]. This is now the conventional approach in almost all genome-sequencing projects. Arguments continue about whether it is more efficient and informative to obtain complete genome sequences for a small number of strains or to obtain partial sequences for a large collection of strains[86]. Highly efficient and cost-effective, 'next-generation', sequencing technologies hold promise for sequencing many more bacterial genomes at a cost of only a few hundred dollars each[61].

In the field of bacteriology, the genomic revolution catalysed the development of bioinformatics approaches (for example, for comparing genomes and for detecting sequence-based evidence of selection and horizontal gene transfer) and high-throughput experimental technologies (for example, microarray-based transcriptomics, mass mutagenesis, and the use of simpler surrogate hosts such as the microscopic nematode *Caenorhabditis elegans* and the budding yeast *Saccharomyces cerevisiae*[87,88]). The advent of genomics also led to new experimental approaches for assessing genomic variability, including multi-locus sequence typing[89], variable-number tandem-repeat typing[90] and the use of microarrays for genomic comparisons[91]. These approaches have provided insight into the population structure of bacterial pathogens, facilitated the classification of strains by source or pathogenicity and helped to identify pathogen-specific genes in pathogenic lineages[92].

Practical applications of sequencing bacterial genomes include metabolic reconstruction (allowing the design of improved culture media[93]), glyco-engineering (allowing the biotechnological manipulation of sugar residues on macromolecules)[94] and reverse vaccinology (facilitating the discovery of new vaccine targets[95]). Genomic studies have also focused on the host genome, allowing the identification of host genes that are expressed after bacterial invasion[96] or are associated with susceptibility to infection[97].

of their genomes: for example, the laboratory strain *E. coli* K-12 is missing 1.4 megabases of DNA present in *E. coli* O157 (ref. 3). For many important pathogens, the genes common to all strains within a species (known as the core genome) are a minority component of the entire gene pool for that species (the pan-genome). Furthermore, a distinction can be made between closed pan-genomes and open pan-genomes. For closed pan-genomes, completing the genome sequencing of additional bacterial strains is unlikely to yield new genes. By contrast, for open pan-genomes, each new genome sequence reveals new members of the gene pool for that species[22].

The genomes of some bacterial pathogens have gained genes through gene duplication, resulting in increased numbers of key gene clusters or the expansion of important protein families: for example, in *M. tuberculosis*, the gene families encoding acidic glycine-rich proteins and the gene clusters encoding the secreted protein ESAT6 (early secretory antigenic target 6) and its homologues have undergone extensive rounds of gene duplication[23]. Nonetheless, gene gain as a result of horizontal gene transfer remains the most potent source of 'innovation' and variation. However, unlike viruses, bacteria seldom acquire 'eukaryotic-like' genes from their hosts (although there seem to be some exceptions, for example, *Legionella pneumophila*[24]). Instead, horizontal gene transfer generally occurs between different strains and species of bacteria.

Horizontal gene transfer is mediated by diverse mobile genetic elements, including plasmids, bacteriophages and pathogenicity islands (Table 1). These elements often carry genes that encode factors involved in infection (often termed virulence factors) (Box 2). For example, numerous virulence factors and systems are encoded on plasmids. These virulence-associated plasmids can be large (for example, the plant pathogen *Ralstonia solanacearum* carries such a plasmid of more than 2 megabases[25]). They can also be promiscuous: that is, they can move freely between cells with markedly different chromosomal backgrounds. In the extreme case of enterotoxigenic *E. coli*, the association of promiscuous plasmids with diverse chromosomal lineages is all that defines the pathotype of these bacteria[26].

Bacteriophages (that is, bacterial viruses) can also mediate horizontal gene transfer. Some classic virulence factors, such as diphtheria toxin, are encoded in the genomes of bacteriophages that have integrated into the bacterial chromosome (which are known as prophages)[27]. Genomic analyses show that prophages have a widespread role in driving the diversification of bacterial pathogens as distinct as *E. coli*, *Streptococcus*

Table 1 | Examples of mobile genetic elements that encode virulence factors and are present in human pathogens

Type of mobile element	Pathogen	Virulence factor
Plasmid	*Bacillus anthracis*	Anthrax toxin
	Clostridium tetani	Tetanus toxin
	Enterotoxigenic *Escherichia coli*	Heat-stabile toxin, heat-labile toxin and fimbriae
	Mycobacterium ulcerans	Polyketide toxin
	Salmonella enterica serovar Typhimurium	SpvR, SpvA, SpvB, SpvC and SpvD proteins*
	Shigella spp.	Type III secretion system
	Staphylococcus aureus	Exfoliatin B
	Pathogenic *Yersinia* spp.	Type III secretion system
Prophage	*Corynebacterium diphtheriae*	Diphtheria toxin
	Enterohaemorrhagic *E. coli*	Shiga toxin and type III secretion effectors
	S. aureus	Staphylococcal enterotoxin A, exfoliatin A and Panton–Valentine leukocidin
	Streptococcus pyogenes	Streptococcal pyrogenic exotoxins, DNases and streptococcal phospholipase A_2 (Sla)
	Vibrio cholerae	Cholera toxin
Pathogenicity island	*Clostridium difficile*	Clostridial enterotoxin and clostridial cytotoxin
	Enteropathogenic and enterohaemorrhagic *E. coli*	Type III secretion system
	Uropathogenic *E. coli*	Fimbriae, iron-uptake systems, the capsular polysaccharide and α-haemolysin
	Helicobacter pylori	Cag antigen
	S. enterica	Type III secretion systems
	S. aureus	Toxic-shock toxin, staphylococcal enterotoxin B, enterotoxin C, enterotoxin K and enterotoxin L

*Involved in intracellular survival.

pyogenes and *Staphylococcus aureus*[28–30]. Prophages, particularly those derived from tailed bacteriophages, often carry genes that are superfluous for bacteriophage replication, and these genes are present within distinct 'passenger compartments' at one end of the prophage genome. These compartments are sometimes called morons to reflect that the associated prophage genomes encode more DNA than is necessary for bacteriophage replication alone[31]. The genes in these compartments are often implicated in virulence and can show a bias in base composition that sets them apart from the rest of the prophage and from the genome of the bacterial host. For example, in *E. coli* O157, the passenger compartments of lambdoid prophages contain genes with a low G+C composition that encode effector proteins capable of translocation into host cells by a type III secretion mechanism[29].

Pathogenicity islands are another class of mobile element involved in horizontal gene transfer. The term 'pathogenicity island' originated from the study of uropathogenic *E. coli* but has subsequently been widely applied to bacterial pathogens[32]. Pathogenicity islands are usually defined by five characteristics. First, they are clusters of contiguous genes that are present in some related strains or species but not in others. Second, they are presumed to have been acquired by horizontal gene transfer. Third, they are generally associated with transfer RNA gene loci. Fourth, they typically have a G+C content that differs from that of the host bacterial genome. Fifth, they confer on the host bacterium a complex and distinctive virulence phenotype in a single step. Although some pathogenicity islands carry genes encoding integrases (enzymes that integrate the pathogenicity island into the host DNA), the mechanisms underlying the transfer of pathogenicity islands from one genome to another are unclear in many cases, as is the identity of the donor microorganisms.

Despite their mobility, pathogenicity islands are remarkably well integrated into the global regulatory network of bacterial cells. For example, numerous external factors affect the expression of genes on the locus of enterocyte effacement (LEE) pathogenicity island of *E. coli*, making it part of the 'genomic continent'[33–35]. It is also important to recognize that some 'pathogenicity islands' are deletions in one lineage rather than insertions in another[36]. Therefore, instead of considering the evolution of pathogens as a series of acquisitions of pathogenicity islands, a more sophisticated outlook is that genomes are 'molecular palimpsests': that is, the variable compartment of the genome bears the scars of repeated rounds of gene acquisition and erosion.

Gene loss

Bacterial genomes remain about the same size despite the pervasive effects of horizontal gene transfer, so gene gain must be balanced by gene loss[37]. Indeed, it is expected that any gene that is not maintained by natural selection is lost: bacterial genomes are subject to the 'use it or lose it' maxim. Genome sequencing has now provided a series of 'snapshots' that show directly the dynamic processes of gene loss and genome decay (that is, the progressive purging from the genome of unnecessary genes). For example, in many *E. coli* lineages, the Flag-2 and ETT2 gene clusters — which, when intact, span tens of kilobases — have been reduced to small scars occupying only a few hundred base pairs, presumably because they no longer provide any selective advantage to the organism[36,38].

The most surprising snapshots of genome decay have come from recently emerged pathogens that have changed lifestyle, usually to live in a simpler host-associated niche. For example, the genomes of *M. leprae*[39], *Y. pestis*[40] and *Salmonella enterica* serovar Typhi[41] contain hundreds or even thousands of pseudogenes; in the *M. leprae* genome, there are nearly as many pseudogenes as functional genes[39]. These examples contradict the view that every gene in a bacterial genome must have a function and that bacterial genomes never contain 'junk' DNA. Instead, every genome should be viewed as a work in progress, burdened with some non-functional 'baggage of history'.

Another common feature of recently emerged genetically uniform pathogens is the 'proliferation' of transposable elements, particularly insertion sequences, in the genome[42]. This abundance of insertion sequences facilitates homologous recombination within the genome, a process that can result in large-scale chromosomal rearrangements that disrupt the ancestral gene order. In the case of *Y. pestis*, recombination between insertion sequences results in marked anomalies in GC skew (usually a marker of the direction of replication for a given region of chromosome) and in reversible chromosomal rearrangements during *in vitro* growth of the organism[40]. It is unclear whether such large-scale genomic rearrangements have functional relevance.

The most extreme form of genome decay is seen in host-associated bacteria, particularly endosymbionts that have been isolated for long periods in a static and 'undemanding' intracellular niche[43]. Pioneering studies by Siv Andersson and colleagues established that certain intracellular bacteria, such as *Rickettsia prowazekii*, have undergone considerable genomic downsizing, shedding many (or even most) of their ancestral genes[44]. *Buchnera aphidicola*, an aphid endosymbiont, is a pertinent example in that it is a close relative of *E. coli* but has fewer than one-tenth of the genes present in the latter[45]. Extreme genome decay is often accompanied by a shift towards a low G+C content: the largest known shift is in the 160-kilobase genome of the psyllid (jumping plant lice) symbiont *Carsonella ruddii*, which has a G+C content of only 16.5% (ref. 46). But perhaps the most extreme example of bacterial

Box 2 | Defining virulence

In the late nineteenth century, Robert Koch laid the groundwork for establishing a link between pathogens and disease, by putting forward what are now known as Koch's postulates. These postulates are four criteria for determining that a particular organism is the causative agent of a particular disease. First, the organism should be detected in all individuals suffering from the disease but not in their healthy counterparts. Second, it must be possible to isolate the organism from a diseased individual. Third, it must be possible to grow the organism in pure culture. Fourth, the cultured organism must cause disease when introduced into a healthy individual and must be able to be re-isolated from the new host.

Subsequently, it became clear that this is an oversimplified view of host–pathogen interactions, in that most pathogens cause disease across a spectrum, from subclinical infection to severe disease, depending on host factors (for example, the function of the immune system) and bacterial factors (for example, strain-to-strain variation in colonization and virulence factors). In addition, some pathogens cannot be grown in the laboratory, and some cause disease only in partnership with other organisms.

A molecular version of Koch's postulates has been devised by Stanley Falkow, in an attempt to provide a definition of the term 'virulence factor'[98]. This new version has three criteria. First, the potential virulence factor should be found in all pathogenic strains of a species but be absent from their non-pathogenic relatives. Second, specific inactivation of the relevant gene(s) should attenuate virulence in an appropriate animal model. Third, subsequent reintroduction of the gene should restore virulence in the animal model.

Similar to the original Koch's postulates, however, there are problems if these 'molecular Koch's postulates' are applied uncritically. These new postulates rest on the assumption that there is an essential distinction between pathogens and non-pathogens, but bacteria often have different roles in different circumstances. For example, uropathogenic *Escherichia coli* function as commensal microorganisms in the human gut but as pathogens in the human bladder, and enterohaemorrhagic *E. coli* function as commensal microorganisms in the bovine gut but are pathogens in the human gut. Similarly, *Yersinia pestis* is a pathogen of mice and fleas, but the virulence factors are likely to differ in each host.

A key contribution of genomics to this debate is to highlight the tension between the first of Falkow's postulates (virulence factors defined by using comparative genomics) and the rest of the postulates (virulence factors defined by using genetic techniques and models of infection). If the first postulate is enforced — that is, any factors that are common to pathogens and non-pathogens cannot be virulence factors — then some pathogens do not have any virulence factors. If the first postulate is ignored, then many 'virulence factors' turn up in non-pathogens (Table 2).

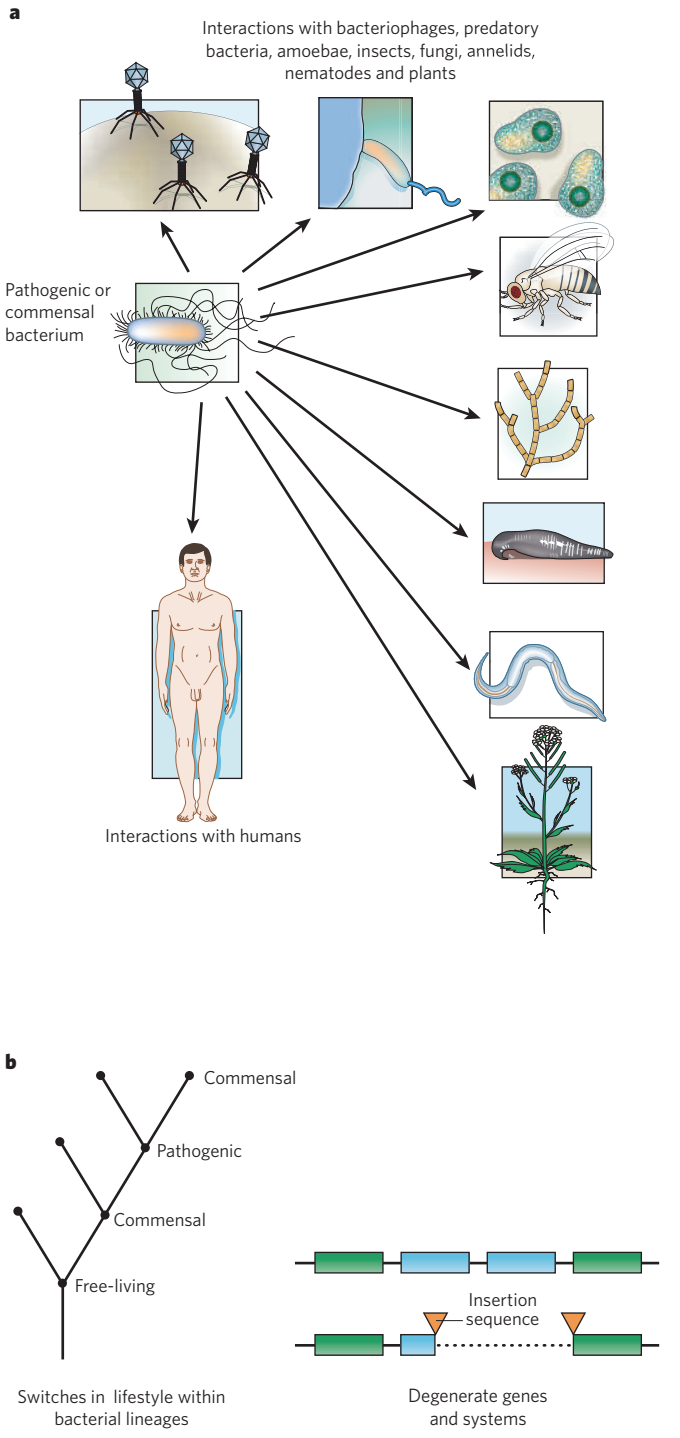

Figure 2 | The eco-evo view of bacterial pathogenomics. a, Pathogenic bacteria and commensal bacteria often share their habitats with bacteriophages, other bacteria, amoebae, insects, nematodes, annelids (such as leeches), fungi, plants and mammals (such as humans). This mixed ecology is a considerable driving force in the evolution of these microorganisms. In this context, it is not surprising that genes encoding 'virulence factors' are found in both human pathogens and non-pathogens. **b**, In addition, consideration of the evolutionary history of a pathogen might be needed to explain some of the features of its genome. Within bacterial genomes, it is common to find remnants of genes or gene clusters that presumably provided an adaptive advantage in the past but are now non-functional (indicated in blue). Also, it should be considered that a microorganism that is pathogenic now might at one time have been a commensal microorganism, and vice versa (indicated by the phylogenetic tree).

genome decay is that of human mitochondria, which belong to the α-proteobacterial lineage and retain a tiny, 17-kilobase, genome (arguably the first bacterial genome to be sequenced[47]).

Less common than genome decay, but more marked in its consequences, is positive selection for gene loss. This occurs as a newly emerged pathogen adapts to its niche and forms part of a process known as pathoadaptation. Pathoadaptation can involve any changes that refine newly formed virulence mechanisms. One example is glucosylation of the surface molecule lipopolysaccharide, which optimizes the exposure of the type III secretion apparatus of *Shigella flexneri*[48]. Pathoadaptation also encompasses gene loss, although it might seem counter-intuitive that losing genes can specifically improve the fitness of a bacterium *in vivo* and make it more pathogenic. The best-known example occurs among the shigellae: loss of the gene *cadA* (which encodes the enzyme lysine decarboxylase) provides a selective advantage in the intracellular niche, because the product of lysine-decarboxylase activity, cadaverine, inhibits the plasmid-encoded virulence factors of these bacteria[49]. Genome sequencing has shown that the genetic mechanisms underlying loss of *cadA* vary between *Shigella* lineages, thus providing an example of convergent evolution in bacterial genomes.

Intriguingly, several recently emerged pathogens (including *Bordetella pertussis*, *B. mallei*, *Y. pestis*, all *Shigella* lineages and some *E. coli* O157 lineages) have independently lost flagellar motility during the transition to a new virulence-associated lifestyle. This suggests that these bacteria are subject to a common pathoadaptive selective pressure, but it is unclear whether the driving force is loss of a target (the protein flagellin) recognized by both the innate immune response and the adaptive immune response in mammals or changes in bacterial metabolism that occur concurrently[50].

An 'eco-evo' perspective on host-pathogen interactions

A glance at the post-genomic landscape shows that our previous knowledge of the ecology and evolution of bacterial pathogenesis was limited. New findings mean that previous assumptions need to be questioned and terms need to be redefined. Among genetically variable bacterial species, it is now clear that a single strain rarely typifies an entire species, particularly because genomics has provided compelling evidence that commonly used laboratory strains (for example, *E. coli* K-12, *S. enterica* serovar Typhimurium LT2, *Pseudomonas aeruginosa* PAO1 and *S. aureus* COL) have undergone marked genotypic and phenotypic changes during their descent from the ancestral free-living isolate[51,52].

Similarly, the readily available bacterial genome-sequence data have challenged the simplistic views that a bacterial pathogen can be understood solely by identifying its virulence factors and that pathogens always evolve from non-pathogens by acquiring virulence genes on plasmids, bacteriophages or pathogenicity islands. Instead, genomics has helped to blur the distinction between pathogens and non-pathogens and between virulence factors and colonization factors. And it has catalysed a copernican shift in how host–pathogen interactions are viewed, a shift away from an anthropocentric focus towards a broader perspective that places interactions between eukaryophilic bacteria and eukaryotes in a wider ecological and evolutionary context (Fig. 2). Inherent in this 'eco–evo' perspective is the need to identify the selective advantages of virulence factors in the broader lifestyle of the pathogen. In addition, 'evolutionary narratives' that interweave genomic changes with ecological shifts can now be constructed. For example, genomic comparisons allow a reconstruction of how the plague bacillus, *Y. pestis* (a rodent and flea pathogen that is occasionally transmitted to humans), evolved from a gastrointestinal pathogen (*Yersinia pseudotuberculosis*) in an evolutionary blink of an eye (about 10,000 years), through the processes of gene gain, loss and rearrangement[8,53,54].

A more fundamental consequence of the eco–evo view is that it is now expected that what, at first, seem to be virulence factors are encoded in the genomes of 'non-pathogens' (Table 2). There are several reasons for this. First, it is now clear, both from genomic and pathogenesis studies, that pathogens, commensal microorganisms and symbionts rely

Table 2 | Examples of 'virulence systems' encoded in the genomes of both pathogenic and non-pathogenic bacteria

Virulence factor	Role in virulence	Homologues in non-pathogens	Potential explanation for presence in non-pathogens	References
Type III secretion system	Role in infection with many human pathogens, including chlamydiae, salmonellae, shigellae and yersiniae	Remnants of type III secretion systems and effectors in commensal strains of *Escherichia coli*, including the laboratory strain *E. coli* K-12	Had a role in a former niche (a degenerate system)	29, 36
		Type III secretion systems in environmental bacteria: for example, *Myxococcus xanthus*, *Verrucomicrobium spinosum*, *Desulfovibrio vulgaris* and non-pathogenic *Yersinia* spp.	Mediate uncharacterized interactions with nematodes, and amoebae and other microscopic eukaryotes in terrestrial and aquatic environments	67
		Type III secretion systems in symbiotic bacteria: for example, *Photorhabdus luminescens*, *'Hamiltonella defensa'*, *Aeromonas veronii*, *Sodalis glossinidius* and *Protochlamydia amoebophila*	Mediate symbiosis with plants, nematodes, leeches, insects and amoebae	67–73
Type VI secretion system	Role in infection with *Vibrio cholerae* or *Pseudomonas aeruginosa*	Type VI secretion systems in environmental bacteria: for example, *Rhodopirellula baltica*, *Hahella chejuensis* and *Oceanobacter* sp.	Mediate uncharacterized interactions with nematodes, and amoebae and other microscopic eukaryotes in aquatic environments	74, 75
ESAT6 and associated Esx secretion system	Key virulence determinant of *Mycobacterium tuberculosis* and *Staphylococcus aureus*, and major attenuating factor in the bacillus Calmette–Guérin (BCG) vaccine against tuberculosis	Esx gene clusters in *Bacillus subtilis*, *Bacillus licheniformis*, *Bacillus halodurans*, *Clostridium acetobutylicum*, *Listeria innocua* and *Streptomyces coelicolor*	Mediate uncharacterized interactions with nematodes, and amoebae and other microscopic eukaryotes in terrestrial and aquatic environments, or involved in conjugative transfer of plasmids	76–81
Specific invasion genes (for example, *yijP*, *ibeB* and *ompA*)	Contribute to invasion of *E. coli* in animal models of meningitis	Invasion genes in commensal strains of *E. coli*, including the laboratory strain *E. coli* K-12	Are a short-sighted local adaptation that does not contribute to transmission (a dead-end trait)	82–84

on similar strategies and molecular systems in their interactions with eukaryotic hosts (for example, phase variation)[21,55]. Second, it is also understood that bacteria sometimes produce virulence factors that provided an advantage only in a previous, now non-existent, niche. Last, it has also become evident that many bacterial pathogens infect humans only incidentally and often produce virulence factors that are active against non-mammalian adversaries as diverse as plants, insects, protozoans, nematodes, predatory bacteria and bacteriophages. Inherent in this view is the realization that many bacterial virulence factors have been shaped by evolutionary forces outside the context of human–pathogen interactions, and only by studying these forces can the emergence of human infections be understood.

Enterohaemorrhagic *E. coli* strains, particularly *E. coli* O157, provide a compelling test case for the eco–evo view. *E. coli* O157 is a rare but devastating pathogen of humans, but it is also a common commensal microorganism of the bovine gut. Genomic comparisons have helped to explain how this pathogen has evolved from a non-pathogenic ancestor by acquiring virulence factors encoded on various mobile elements (for example, Shiga toxin, which is encoded on a bacteriophage, and the type III secretion system encoded on the LEE pathogenicity island)[56]. However, recent studies have shown that a pilus-adherence factor that is crucial to the virulence of *E. coli* O157 in humans is also carried by commensal strains of *E. coli*[57]. Also, similarly to many commensal strains of *E. coli*, *E. coli* O157 carries remnants of a gene cluster (ETT2) encoding a virulence-associated secretion system that is now thought to be inactive[36]. It is only through an eco–evo view that the evolution and transmission of *E. coli* O157, and why it produces such lethal virulence factors, might be understood. One potential explanation is that the 'virulence factors' of *E. coli* O157, such as Shiga toxin, help it to colonize the bovine gut. However, the evidence for this hypothesis is equivocal at best[58]. Instead, a recent study shows that the Shiga-toxin-encoding bacteriophage increases bacterial survival in the presence of a grazing ciliate, *Tetrahymena pyriformis*, indicating that interactions with non-mammalian adversaries might have driven the evolution of this virulence factor[59]. Similar points can be made about many other pathogens: for example, it has long been known that the pathogenesis of legionellosis in humans relies on mechanisms that legionellae use to subvert amoebae[60]. Clearly, in the post-genomic era, even for important human pathogens, humans can no longer be considered to be the centre of the bacterial universe.

Future challenges

The clearest challenge for future studies of bacterial pathogenomics is coping with the flood of new data unleashed by the arrival of affordable and quick, 'next-generation', sequencing technologies[61]. Now that the cost of sequencing bacterial genomes fits comfortably within the budget of a standard research project grant, it is set to become an integral and routine part of research programmes. Therefore, within the next decade, tens of thousands of bacterial genomes will be sequenced[62]. And the focus will shift from the mechanics of generating sequence data to the problems of analysing it, creating an urgent need for better ways to compare and visualize genomic data. Also, there are likely to be many more incompletely sequenced genomes, as the efficiency of the finishing stage of a genome project lags behind the rapid pace of next-generation whole-genome shotgun sequencing.

For genetically uniform species, particularly those with potential as bioterrorism agents, the resequencing of hundreds of isolates (for example, by using tiling arrays) will drive forward forensic genomics. For more variable species, such as *E. coli*, *S. enterica* and *S. aureus*, the focus will be on defining the extent of the pan-genome and on developing improved approaches to understanding epidemiology, particularly for those that cause hospital-acquired infections. The most important challenge will still be to add functional relevance to genome sequences, a challenge that will continue to drive the application of high-throughput 'omics' approaches to the study of virulence.

Furthermore, sequencing the genomes of environmental organisms and carrying out metagenomic surveys of diverse environments will provide not only an improved understanding of microbial biodiversity but also insight into the evolution of bacterial factors that are involved in human disease[63,64]. Metagenomic surveys of eukaryote-associated bacterial communities will strengthen our understanding of the ecology of bacterial infections (for example, the micro-ecological changes that accompany antibiotic-associated diarrhoea) and help to shed light on the pathogenesis of polymicrobial infections, such as those that cause periodontal disease and bacterial vaginosis[65]. Similarly, studying the metagenomics of bacteriophage populations will help to unravel the connections between these mobile elements and the evolution of virulence.

In addition to the genomic technologies discussed here, evolutionary game theory will need to be applied so that the complex interactions between bacteriophages, the virulence factors that they encode, the bacteria that they infect and the eukaryotic targets of their virulence

Glossary

Bacteriophage A virus that infects bacteria. A bacteriophage can either lyse a cell or integrate into its genome.

Commensal microorganism A microorganism that benefits from living in close contact with a human or animal but has no direct beneficial or detrimental effects on its host.

Core genome The set of genes found in all members of a single species.

Eco–evo perspective A perspective in which organisms are evaluated broadly in the light of evolution and ecology, rather than narrowly by the constraints of their behaviour in the laboratory or in human infection.

Eukaryophilic A term applied to any bacterium that interacts with eukaryotes in its natural environmental niche. It does not specify whether the interaction is pathogenic or symbiotic.

Finishing stage The final phase in a genome-sequencing project, in which all gaps between contigs are closed and all ambiguities are resolved. This is much more labour-intensive than the shotgun phase, so the sequencing of many genomes is now left unfinished.

GC skew A measure that reflects bias for guanine bases on the leading strand of DNA and cytosine bases on the lagging strand.

Genome The complete set of genetic information in an organism. In bacteria, this includes the chromosome(s) and plasmids.

Horizontal gene transfer Any process in which an organism transfers genetic material to another cell that is not its offspring. This process is in contrast to vertical gene transfer, which is much more common and occurs when genetic material is passed from parent to offspring or, more generally, from ancestor to descendent.

Insertion sequence The simplest type of transposable element in bacteria. It contains only the genes required for its own transposition.

Metagenomics The high-throughput study of sequences from multiple genomes recovered from environmental samples that contain mixed populations.

Next-generation sequencing A set of novel approaches to DNA sequencing that dispenses with the need to create libraries of cloned sequences in bacteria and holds the promise of providing faster and cheaper sequencing.

Pan-genome The set of all genes found in members of a single species.

Pathoadaptation The genetic changes that occur after transition to a new pathogenic lifestyle and ensure that the bacterium becomes fitter in its new host niche. These can include changes in the sequences of genes, alterations in gene expression and loss of genes.

Pathogen An organism that can cause disease in another organism.

Pathogenicity island A cluster of genes acquired by horizontal gene transfer that encodes products that contribute to virulence.

Phase variation A spontaneous genetically defined switch between expression of alternative, usually surface-associated, proteins that occurs at a high frequency.

Plasmid An extrachromosomal DNA molecule that can replicate autonomously within a bacterial cell.

Resequencing The application of genome sequencing to a close relative of a strain that has already been sequenced. The availability of a template genome greatly facilitates the finishing stage of sequencing. In contrast to resequencing, *de novo* sequencing, in which a large proportion of the genome is novel sequence, is much more challenging.

Single-nucleotide polymorphism A variation between two genomes that involves a single base pair.

Symbiont An organism that lives in close contact with another organism (for example, a bacterium and a eukaryote) in a relationship in which both partners benefit. Endosymbionts are symbionts that live within the body or cells of another organism.

Type III secretion One of several processes by which bacteria export proteins to the external environment. Type III secretion is required for the biosynthesis of the main organelle of motility in bacteria, the flagellum. It is also responsible for the translocation of bacterial effector proteins from pathogenic or symbiotic bacteria to the cytoplasm of their eukaryotic partners, where these proteins subvert eukaryotic-cell functions to the advantage of the bacterium.

Virulence factor A factor that is produced by a pathogen and required for it to cause disease. It should be noted that defining this term precisely is difficult.

Whole-genome shotgun sequencing An approach to determining the sequence of a genome in which the genome is broken into numerous small fragments. These fragments are then sequenced en masse. The individual sequences are assembled into larger sequences (known as contigs) that correspond to substantial portions of the genome. Typically, more than 99% of the genome can be sequenced using this approach, before the finishing stage.

factors can be understood. Similar approaches will also be needed to solve the conundrum of invasive disease: for example, to explain why meningococci cause meningitis, despite the fact that the disease has no role in the transmission of the bacteria[66]. Evolutionary systems-biology approaches will also be useful for understanding the evolution and regulation of complex virulence systems, the interactions between pathogens and their host, and the co-evolution of their genomes.

The first decade of bacterial genomics has afforded unprecedented insights into the evolution of virulence. The next decade holds the promise of being even more rewarding as the new eco–evo view of host–pathogen interactions draws on ever more genome and metagenome sequences. ■

1. Read, T. D. *et al.* Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* **296,** 2028–2033 (2002).
2. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. *Proc. Natl Acad. Sci. USA* **102,** 13950–13955 (2005).
3. Hayashi, T. *et al.* Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8,** 11–22 (2001); erratum **8,** 96 (2001).
4. Kuroda, M. *et al.* Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet* **357,** 1225–1240 (2001).
5. Fraser-Liggett, C. M. Insights on biology and evolution from microbial genome sequencing. *Genome Res.* **15,** 1603–1610 (2005).
6. Lawrence, J. G. Horizontal and vertical gene transfer: the life history of pathogens. *Contrib. Microbiol.* **12,** 255–271 (2005).
7. Raskin, D. M., Seshadri, R., Pukatzki, S. U. & Mekalanos, J. J. Bacterial genomics and pathogen evolution. *Cell* **124,** 703–714 (2006).
8. Wren, B. W. The yersiniae — a model genus to study the rapid evolution of bacterial pathogens. *Nature Rev. Microbiol.* **1,** 55–64 (2003).
9. Pearson, T. *et al.* Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc. Natl Acad. Sci. USA* **101,** 13536–13541 (2004).
10. Touchman, J. W. *et al.* A North American *Yersinia pestis* draft genome aequence: SNPs and phylogenetic analysis. *PLoS ONE* **2,** e220 (2007).
11. Gutacker, M. M. *et al.* Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* **162,** 1533–1543 (2002).
12. Monot, M. *et al.* On the origin of leprosy. *Science* **308,** 1040–1042 (2005).

13. Hayashi, K. *et al.* Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol. Syst. Biol.* **2,** doi:10.1038/msb4100049 (2006).
14. Zhang, W. *et al.* Probing genomic diversity and evolution of *Escherichia coli* O157 by single nucleotide polymorphisms. *Genome Res.* **16,** 757–767 (2006).
15. Chaudhuri, R. R. *et al.* Genome sequencing shows that European isolates of *Francisella tularensis* subspecies *tularensis* are almost identical to US laboratory strain Schu S4. *PLoS ONE* **2,** e352 (2007).
16. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309,** 1728–1732 (2005).
17. Romero, C. M. *et al.* Genome sequence alterations detected upon passage of *Burkholderia mallei* ATCC 23344 in culture and in mammalian hosts. *BMC Genomics* **7,** 228 (2006).
18. Moxon, R., Bayliss, C. & Hood, D. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.* **40,** 307–333 (2006).
19. van der Woude, M. W. & Baumler, A. J. Phase and antigenic variation in bacteria. *Clin. Microbiol. Rev.* **17,** 581–611 (2004).
20. Parkhill, J. *et al.* The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403,** 665–668 (2000).
21. Cerdeno-Tarraga, A. M. *et al.* Extensive DNA inversions in the *B. fragilis* genome control variable gene expression. *Science* **307,** 1463–1465 (2005).
22. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15,** 589–594 (2005).
23. Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393,** 537–544 (1998).
24. Bruggemann, H., Cazalet, C. & Buchrieser, C. Adaptation of *Legionella pneumophila* to the host environment: role of protein secretion, effectors and eukaryotic-like proteins. *Curr. Opin. Microbiol.* **9,** 86–94 (2006).
25. Salanoubat, M. *et al.* Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* **415,** 497–502 (2002).
26. Turner, S. M. *et al.* Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages. *J. Clin. Microbiol.* **44,** 4528–4536 (2006).
27. Freeman, V. J. Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *J. Bacteriol.* **61,** 675–688 (1951).
28. Brussow, H., Canchaya, C. & Hardt, W. D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **68,** 560–602 (2004).
29. Tobe, T. *et al.* An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc. Natl Acad. Sci. USA* **103,** 14941–14946 (2006).
30. Ohnishi, M., Kurokawa, K. & Hayashi, T. Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? *Trends Microbiol.* **9,** 481–485 (2001).
31. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. The origins and ongoing evolution of viruses. *Trends Microbiol.* **8,** 504–508 (2000).
32. Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. Genomic islands in pathogenic and environmental microorganisms. *Nature Rev. Microbiol.* **2,** 414–424 (2004).
33. Zhang, L. *et al.* Regulators encoded in the *Escherichia coli* type III secretion system 2 gene cluster influence expression of genes within the locus for enterocyte effacement in enterohemorrhagic *E. coli* O157:H7. *Infect. Immun.* **72,** 7282–7293 (2004).
34. Nakanishi, N. *et al.* ppGpp with *DksA* controls gene expression in the locus of enterocyte effacement (LEE) pathogenicity island of enterohaemorrhagic *Escherichia coli* through activation of two virulence regulatory genes. *Mol. Microbiol.* **61,** 194–205 (2006).
35. Laaberki, M. H., Janabi, N., Oswald, E. & Repoila, F. Concert of regulators to switch on LEE expression in enterohemorrhagic *Escherichia coli* O157:H7: interplay between Ler, GrlA, HNS and RpoS. *Int. J. Med. Microbiol.* **296,** 197–210 (2006).
36. Ren, C. P. *et al.* The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli*, is present in the majority of strains but has undergone widespread mutational attrition. *J. Bacteriol.* **186,** 3547–3560 (2004).
37. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17,** 589–596 (2001).
38. Ren, C. P., Beatson, S. A., Parkhill, J. & Pallen, M. J. The *Flag-2* locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from *Escherichia coli*. *J. Bacteriol.* **187,** 1430–1440 (2005).
39. Cole, S. T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409,** 1007–1011 (2001).
40. Parkhill, J. *et al.* Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413,** 523–527 (2001).
41. Parkhill, J. *et al.* Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413,** 848–852 (2001).
42. Siguier, P., Filee, J. & Chandler, M. Insertion sequences in prokaryotic genomes. *Curr. Opin. Microbiol.* **9,** 526–531 (2006).
43. Wernegreen, J. J. For better or worse: genomic consequences of intracellular mutualism and parasitism. *Curr. Opin. Genet. Dev.* **15,** 572–583 (2005).
44. Andersson, S. G. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396,** 133–140 (1998).
45. Perez-Brocal, V. *et al.* A small microbial genome: the end of a long symbiotic relationship? *Science* **314,** 312–313 (2006).
46. Nakabachi, A. *et al.* The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* **314,** 267 (2006).
47. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290,** 457–465 (1981).
48. West, N. P. *et al.* Optimization of virulence functions through glucosylation of *Shigella* LPS. *Science* **307,** 1313–1317 (2005).
49. Maurelli, A. T. Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. *FEMS Microbiol. Lett.* **267,** 1–8 (2007).
50. Leatham, M. P. *et al.* Mouse intestine selects nonmotile *flhDC* mutants of *Escherichia coli* MG1655 with increased colonizing ability and better utilization of carbon sources. *Infect. Immun.* **73,** 8039–8049 (2005).
51. Fux, C. A., Shirtliff, M., Stoodley, P. & Costerton, J. W. Can laboratory reference strains mirror 'real-world' pathogenesis? *Trends Microbiol.* **13,** 58–63 (2005).
52. Hobman, J. L., Penn, C. W. & Pallen, M. J. Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully? *Mol. Microbiol.* **64,** 881–885 (2007).
53. Achtman, M. *et al.* Microevolution and history of the plague bacillus, *Yersinia pestis*. *Proc. Natl Acad. Sci. USA* **101,** 17837–17842 (2004).
54. 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).
55. van der Woude, M. W. Re-examining the role and random nature of phase variation. *FEMS Microbiol. Lett.* **254,** 190–197 (2006).
56. Wick, L. M., Qi, W., Lacher, D. W. & Whittam, T. S. Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7. *J. Bacteriol.* **187,** 1783–1791 (2005).
57. Rendón, M. A. *et al.* Commensal and pathogenic *Escherichia coli* use a common pilus adherence factor for epithelial cell colonization. *Proc. Natl Acad. Sci. USA* **104,** 10637–10642 (2007).
58. Sheng, H., Lim, J. Y., Knecht, H. J., Li, J. & Hovde, C. J. Role of *Escherichia coli* O157:H7 virulence factors in colonization at the bovine terminal rectal mucosa. *Infect. Immun.* **74,** 4685–4693 (2006).
59. Meltz Steinberg, K. & Levin, B. R. Grazing protozoa and the evolution of the *Escherichia coli* O157:H7 Shiga toxin-encoding prophage. *Proc. R. Soc. B* **274,** 1921–1929 (2007).
60. Albert-Weissenberger, C., Cazalet, C. & Buchrieser, C. *Legionella pneumophila* — a human pathogen that co-evolved with fresh water protozoa. *Cell. Mol. Life Sci.* **64,** 432–448 (2007).
61. Hall, N. Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* **210,** 1518–1525 (2007).
62. Field, D., Wilson, G. & van der Gast, C. How do we compare hundreds of bacterial genomes? *Curr. Opin. Microbiol.* **9,** 499–504 (2006).
63. Rusch, D. B. *et al.* The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.* **5,** e77 (2007).
64. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
65. Brogden, K. A., Guthmiller, J. M. & Taylor, C. E. Human polymicrobial infections. *Lancet* **365,** 253–255 (2005).
66. Meyers, L. A., Levin, B. R., Richardson, A. R. & Stojiljkovic, I. Epidemiology, hypermutation, within-host evolution and the virulence of *Neisseria meningitidis*. *Proc. R. Soc. B* **270,** 1667–1677 (2003).
67. Pallen, M. J., Beatson, S. A. & Bailey, C. M. Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: a Darwinian perspective. *FEMS Microbiol. Rev.* **29,** 201–229 (2005).
68. Ffrench-Constant, R. H. *et al.* A genomic sample sequence of the entomopathogenic bacterium *Photorhabdus luminescens* W14: potential implications for virulence. *Appl. Environ. Microbiol.* **66,** 3310–3329 (2000).
69. Moran, N. A., Degnan, P. H., Santos, S. R., Dunbar, H. E. & Ochman, H. The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. *Proc. Natl Acad. Sci. USA* **102,** 16919–16926 (2005).
70. Silver, A. C. *et al.* Interaction between innate immune cells and a bacterial type III secretion system in mutualistic and pathogenic associations. *Proc. Natl Acad. Sci. USA* **104,** 9481–9486 (2007).
71. Skorpil, P. *et al.* NopP, a phosphorylated effector of *Rhizobium* sp. strain NGR234, is a major determinant of nodulation of the tropical legumes *Flemingia congesta* and *Tephrosia vogelii*. *Mol. Microbiol.* **57,** 1304–1317 (2005).
72. Horn, M. *et al.* Illuminating the evolutionary history of chlamydiae. *Science* **304,** 728–730 (2004).
73. Dale, C., Young, S. A., Haydon, D. T. & Welburn, S. C. The insect endosymbiont *Sodalis glossinidius* utilizes a type III secretion system for cell invasion. *Proc. Natl Acad. Sci. USA* **98,** 1883–1888 (2001).
74. Pukatzki, S. *et al.* Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system. *Proc. Natl Acad. Sci. USA* **103,** 1528–1533 (2006).
75. Mougous, J. D. *et al.* A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus. *Science* **312,** 1526–1530 (2006).
76. Pym, A. S., Brodin, P., Brosch, R., Huerre, M. & Cole, S. T. Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and *Mycobacterium microti*. *Mol. Microbiol.* **46,** 709–717 (2002).
77. Lewis, K. N. *et al.* Deletion of RD1 from *Mycobacterium tuberculosis* mimics bacille Calmette–Guerin attenuation. *J. Infect. Dis.* **187,** 117–123 (2003).
78. Brodin, P. *et al.* Dissection of ESAT-6 system 1 of *Mycobacterium tuberculosis* and impact on immunogenicity and virulence. *Infect. Immun.* **74,** 88–98 (2006).
79. Pallen, M. J. The ESAT-6/WXG100 superfamily — and a new Gram-positive secretion system? *Trends Microbiol.* **10,** 209–212 (2002).
80. Desvaux, M. *et al.* Genomic analysis of the protein secretion systems in *Clostridium acetobutylicum* ATCC 824. *Biochim. Biophys. Acta* **1745,** 223–253 (2005).
81. Burts, M. L., Williams, W. A., DeBord, K. & Missiakas, D. M. EsxA and EsxB are secreted by an ESAT-6-like system that is required for the pathogenesis of *Staphylococcus aureus* infections. *Proc. Natl Acad. Sci. USA* **102,** 1169–1174 (2005).
82. Huang, S. H. *et al.* Identification and characterization of an *Escherichia coli* invasion gene locus, *ibeB*, required for penetration of brain microvascular endothelial cells. *Infect. Immun.* **67,** 2103–2109 (1999).
83. Huang, S. H., Stins, M. F. & Kim, K. S. Bacterial penetration across the blood–brain barrier during the development of neonatal meningitis. *Microbes Infect.* **2,** 1237–1244 (2000).
84. Huang, S. H., Wan, Z. S., Chen, Y. H., Jong, A. Y. & Kim, K. S. Further characterization of *Escherichia coli* brain microvascular endothelial cell invasion gene *ibeA* by deletion, complementation, and protein expression. *J. Infect. Dis.* **183,** 1071–1078 (2001).
85. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).
86. Parkhill, J. The importance of complete genome sequences. *Trends Microbiol.* **10,** 219–220 (2002).
87. Dorer, M. S. & Isberg, R. R. Non-vertebrate hosts in the analysis of host–pathogen interactions. *Microbes Infect.* **8,** 1637–1646 (2006).

88. Hilbi, H., Weber, S. S., Ragaz, C., Nyfeler, Y. & Urwyler, S. Environmental predators as models for bacterial pathogenesis. *Environ. Microbiol.* **9,** 563–575 (2007).
89. Maiden, M. C. Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* **60,** 561–588 (2006).
90. Lindstedt, B. A. Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. *Electrophoresis* **26,** 2567–2582 (2005).
91. Dorrell, N., Hinchliffe, S. J. & Wren, B. W. Comparative phylogenomics of pathogenic bacteria by microarray analysis. *Curr. Opin. Microbiol.* **8,** 620–626 (2005).
92. Champion, O. L. *et al.* Comparative phylogenomics of the food-borne pathogen *Campylobacter jejuni* reveals genetic markers predictive of infection source. *Proc. Natl Acad. Sci. USA* **102,** 16043–16048 (2005).
93. Renesto, P. *et al.* Genome-based design of a cell-free culture medium for *Tropheryma whipplei*. *Lancet* **362,** 447–449 (2003).
94. Wacker, M. *et al.* N-linked glycosylation in *Campylobacter jejuni* and its functional transfer into *E. coli*. *Science* **298,** 1790–1793 (2002).
95. Mora, M., Donati, C., Medini, D., Covacci, A. & Rappuoli, R. Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach. *Curr. Opin. Microbiol.* **9,** 532–536 (2006).
96. Jenner, R. G. & Young, R. A. Insights into host responses against pathogens from transcriptional profiling. *Nature Rev. Microbiol.* **3,** 281–294 (2005).
97. Hill, A. V. Aspects of genetic susceptibility to human infectious diseases. *Annu. Rev. Genet.* **40,** 469–486 (2006).
98. Falkow, S. Molecular Koch's postulates applied to microbial pathogenicity. *Rev. Infect. Dis.* **10** (suppl. 2), S274–S276 (1988).

Acknowledgements We thank L. Snyder, J. Kelly, D. Baker, L. Bingle and S. Andersson for critical reading of the manuscript. We acknowledge the Biotechnology and Biological Sciences Research Council for funding numerous genomic research projects in our laboratories, and the Wellcome Trust (particularly the Wellcome Trust Sanger Institute) for facilitating bacterial genome sequencing in the United Kingdom. This article is dedicated to the memory of C. A. Hart.

Author Information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to the authors (m.pallen@bham.ac.uk; brendan.wren@lshtm.ac.uk).