Published in final edited form as:

Future Microbiol. 2012 November ; 7(11): 1283–1296. doi:10.2217/fmb.12.108.

Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences

Josephine Bryant^{#1}, Claire Chewapreecha^{#1}, and Stephen D Bentley^{*,1,2}

¹Pathogen Genomics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK

²Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, CB2 0QQ, UK

[#] These authors contributed equally to this work.

Abstract

Evolution of bacterial pathogen populations has been detected in a variety of ways including phenotypic tests, such as metabolic activity, reaction to antisera and drug resistance and genotypic tests that measure variation in chromosome structure, repetitive loci and individual gene sequences. While informative, these methods only capture a small subset of the total variation and, therefore, have limited resolution. Advances in sequencing technologies have made it feasible to capture whole-genome sequence variation for each sample under study, providing the potential to detect all changes at all positions in the genome from single nucleotide changes to large-scale insertions and deletions. In this review, we focus on recent work that has applied this powerful new approach and summarize some of the advances that this has brought in our understanding of the details of how bacterial pathogens evolve.

Keywords

bacteria; evolution; genome sequencing; horizontal gene transfer; mutation; recombination; selection

Since 1995 when the first two bacterial pathogens were fully sequenced, *Haemophilus influenzae* [1] and *Mycoplasma genitalium* [2], pathogen biology has benefited from the wealth of information and insight provided by bacterial genomes. We are now in the age of next-generation sequencing, which allows deep investigations into genome variation within individual species or clones. Here we focus on the most recent evolutionary insights provided by large-scale sequencing projects, but cannot ignore the extensive body of work it

© 2012 Wellcome Trust Sanger Institute (Genome Research Limited)

*Author for correspondence: sdb@sanger.ac.uk.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

builds upon. For this reason, we have included both next-generation sequencing studies and the insights gained through comparisons of high-quality reference sequences.

How bacteria evolve: point mutations

Point mutations are often thought of as the raw material of evolution; these small changes include the substitution of one nucleotide with another (often referred to as a single-nucleotide polymorphism [SNP]) or the insertion or deletion of a single nucleotide. Single-nucleotide substitution in a DNA sequence that encodes a protein may produce either a synonymous (silent) codon mutation, and thus no change in the encoded amino acid, or a nonsynonymous mutation, resulting in an amino acid change that may have an impact on the function of the encoded protein. Single nucleotide insertion or deletion in protein coding sequences will result in a frameshift such that the downstream codons, including stop codons, will be translated from a different reading frame, leading to significant alteration of the encoded protein. Point mutations in non-protein-coding DNA sequences can also have functional consequences, particularly if they affect a regulatory element.

Regardless of the consequence of a point mutation, its maintenance in the resulting genome may imply a selective advantage and can be used in comparison with the genomes of related individuals to construct phylogenetic relationships. Studying the function, rate, type and distribution of point mutations can therefore give us insights into the evolutionary pressures on bacterial pathogens.

Function

Identification of point mutations responsible for particular phenotypic changes has often been carried out through ‘bottom-up’ approaches, such as complementation tests [3]. However, identifying the specific mutation responsible for a phenotype is often time consuming and requires previous knowledge of genetic markers for linkage analysis [4]. Whole-genome sequencing (WGS) provides us with a ‘top-down’ approach to associate genotype with phenotype and has recently sped up learning about the basis of virulence and antibiotic-resistance acquisition in pathogens. For example, Renzoni *et al.* used WGS to identify SNPs in two isogenic strains of *Staphylococcus aureus*: a parental strain and a derivative strain that had undergone stepwise *in vitro* selection for resistance to teicoplanin (a glycopeptide antibiotic). Only three SNP differences were identified (*stpI*, *yjbH* and *vraS*) and experimental work confirmed that all three were required to confer the highest level of glycopeptide resistance. Resistance and fitness conferred by a double mutant (*vraS*[G45R], *stpI*[Q12stop]) suggested a synergistic interaction between the two, and the third (*yjbH*[K23stop]) had a more subtle contribution to resistance [5]. Similarly, Comas *et al.* recently reported the discovery of SNPs that compensate for the fitness costs associated with rifampicin resistance in *Mycobacterium tuberculosis* [6]. Previous experimental work had shown that laboratory and clinical strains, both resistant to rifampicin through the same causal point substitution in *rpoB*, had different fitness phenotypes [7]. The clinical strains were found to have a higher level of fitness than the *in vitro*-selected laboratory strains, suggesting that other genomic variations may have evolved to compensate for the reduction in fitness often associated with drug resistance. WGS was carried out on the laboratory-evolved strains in addition to ten pairs of rifampicin-resistant clinical isolates and their

corresponding susceptible counterpart collected from the same individual at an earlier time point. Mapping of the sequence data to a reference genome allowed the identification of 11 nonsynonymous SNPs in eight of the pairs. Most notably, *rpoB* and *rpoC* harbored multiple independent nonsynonymous SNPs in both the laboratory evolved and clinical strains. Mapping of the resultant amino acid changes onto a model of the structure of the *Escherichia coli* RNA polymerase showed that they occurred on the interface between the α - and β -subunits and the compensatory nature of these SNPs was confirmed experimentally. Subsequent analysis of a broader set of genomes showed that the same mutations had occurred independently in different *M. tuberculosis* lineages, further demonstrating the power of WGS to identify subtle genomic variations and to put them into a functional context [6].

Rate

Determining mutation rate can be crucial to understanding the biology and evolution of an organism. Mutation rate is distinct from fixation rate, in that it describes the raw influx of mutations in a genome without the influence of selection or genetic drift. The rate of accumulation of synonymous and intergenic substitutions should, in theory, expose the rate of mutation in a genome in the absence of selection and comparison of genome sequences provides a powerful route to deriving such rates [8].

Previously, estimates of mutation rate in bacteria had been determined experimentally, for example, by measuring the rate of reversion to wild-type for *lacZ* mutant alleles through selection for growth on lactose-minimal media [9]. However, large discrepancies are observed between these experimentally derived rates and those determined directly from genome sequence comparisons. The rate of point mutations in *E. coli* and *Salmonella enterica* estimated from *lacZ* reversion rate is approximately 5×10^{-10} per base pair per generation, while the observed rate of change at apparently neutral positions from the comparison of *E. coli* and *S. enterica* genomes is approximately 4.5×10^{-9} per site per year [8,10]. This discrepancy could be partly due to the non-neutral nature of synonymous point mutations. It has been shown that these sites in the genome are under selective pressures for translational efficiency, where codon bias is observed in more highly translated genes than others [11]. In addition, the controlled experimental conditions of the *lacZ* assay are probably leading to unrealistic estimates as they ignore the environmental stresses and bottlenecks that take place in natural populations of bacteria.

WGS allows an estimation of mutation rates directly from clinical isolates, which are more likely to represent 'wild' rates. WGS of 63 clinical isolates from a single clone of methicillin-resistant *S. aureus* (MRSA) revealed a surprisingly high mutation rate of 3.3×10^{-6} SNPs per site per year; approximately 1000-times faster than that estimated from the comparison of *E. coli* and *S. enterica* genomes. The authors speculate that this elevated rate is due to the higher relatedness of the samples, implying a shorter time since their most recent common ancestor and thereby a shorter time for the removal, by selection, of slightly deleterious mutations [12]. This time dependency of mutation rate has also been demonstrated in eukaryotic datasets [13].

Understanding mutation rate has important consequences for our understanding of pathogen evolution and one important example concerns *M. tuberculosis*. Latent infection with *M. tuberculosis* is generally assumed to represent a dormant state for the pathogen and, since only 10% of patients with latent infection suffer reactivation, it has long been assumed that the mutation rate is low or completely absent during this phase [14]. Understanding the dynamics of this process could provide clues to the triggers for reactivation of disease and is vital to assess the potential for acquisition of drug resistance. WGS was used to investigate mutation rate during infection in a nonhuman primate model. Genome analysis of 33 isolates showed the mutation rate during latency to be similar to that during active infection [15]. This has serious implications for therapeutic practices where isoniazid monotherapy is common during latent infection and the risk for selection of resistance may have been underestimated. It has also been shown that mutation rate can vary over the course of an infection. Sequencing of *Pseudomonas aeruginosa* isolates from early and chronic infection of a cystic fibrosis patient revealed that 'chronic' isolates were characterized by a mutator phenotype and a high proportion of nonsynonymous SNPs. Many of the SNPs led to the loss of function of genes associated with bacterial virulence, suggesting that genes that promote acute infection may be selected against during chronic infection. The study also implies that hypermutation in long term infection may serve to promote genetic adaptation [16].

Mutation rates inferred from sampling isolates in a time-stratified manner can also allow the dating of evolutionary events. By plotting the SNP distance from the phylogenetic root of the tree against time, Harris *et al.* were able to estimate the emergence of the ST239 clone of MRSA to the mid-1960s, consistent with the increased use of antibiotics and the first detection of MRSA [12]. This method was similarly applied to whole-genome sequences of a clone *Streptococcus pneumoniae* and showed that the generation of vaccine-escape variants had occurred within the population prior to the introduction of the antipneumococcal vaccine [17].

Type

An advantage of the increasing volume of sequence data available for bacteria is that it allows the testing of evolutionary hypotheses across the entire domain. Recent work by Hershberg and Petrov used bacterial WGS data to investigate the spectrum of mutation types. Surprisingly, this revealed that the spectrum varies very little with most biased towards C/G→T/A transitions. This was found universally, even for clades with particularly high GC contents, which suggests that in the absence of selection, bacterial genomes would approach an equilibrium GC content of 20–30%. Given the large variety of GC contents observed for bacteria (~20–80%), this observation suggests that mutational pressures alone are not responsible for the nucleotide composition of the majority of genomes and that selection probably plays an important role [18]. Although the nature of this selective pressure is yet to be determined, one study observed that the GC content of bacterial communities is dependent on the habitat they were isolated from, which suggests that the environment may be an important contributor [19]. Understanding how this dynamic flux of mutation and selection results in the observed nucleotide composition will give us a greater resolution of understanding of the evolution of pathogen genomes.

Distribution

It has been shown that mutations can follow a nonrandom genomic positional distribution, even in the absence of selection. For example, single cell WGS of *E. coli* subjected to chemical mutagenesis revealed a distinctly nonrandom distribution of mutations along the genome. The most striking of these were long stretches of the genome where only C/G→T/A transitions are observed. Most of the observations could be explained by known biological processes, such as semi-conservative replication and sister-strand exchange [20]. The ability to study genomes at the single cell level provides us with the opportunity to study micro-evolutionary events, such as those that may be lost when sequencing a population of cells where only the consensus variants are analyzed. Genome sequences have also revealed mechanisms evolved to make mutation more likely in some parts of the genome than others. Sequencing of two *Salmonella typhimurium* mutants, without their major DNA repair mechanisms and under strongly reduced selection, found that highly expressed genes have a higher mutation rate than the rest of the genome. This suggests transcription can also influence mutation rate [21].

How bacteria evolve: large sequence variants

Detecting structural variants from WGS data

Comparison of whole genomes allows the identification of large genomic variations, including insertions, deletions, inversions, translocations and duplications, which can all contribute to the unique genotypic composition of each isolate. Such structural variations can be identified through both read mapping and sequence assembly. Mapping of sequence reads to a reference genome can allow the identification of discordant signatures beyond simple point mutations. Disproportionate read coverage can be used to detect deletions (manifested as an absence of reads mapping to that region of the genome) and duplications of the genome (manifested as a doubling of reads mapping to that region of the genome). The span and orientation of paired-end reads can indicate a deletion (if read pairs map at a distance longer than expected from the shotgun library insert size) or insertion event (if read pairs map at a distance that is closer than expected). Failure of one of the paired reads to map could also be consistent with an insertion where the read relates to the inserted sequence that is absent from the reference. Inconsistencies in read pairing and orientation can also give clues to inversions and translocations. Although computationally more demanding, *de novo* assembly, followed by comparison with a high-quality reference, can be used to identify all of the structural variants mentioned above with sometimes greater clarity than could be achieved by mapping alone [22,23].

Large-scale insertions

Acquisition of virulence & antimicrobial resistance genes—WGS comparison has been used to identify insertion events in bacterial genomes, which have been proposed to have played a significant role in the emergence of epidemic clones or even the pathogen species itself. Examples include: insertion of prophages, the latent form of bacteriophage in which genes are incorporated into the bacterial chromosome [24]; or integrative conjugative elements (ICEs), the conjugative self-transmissible elements that can integrate into and excise from the chromosomes. The cargo of these elements includes genes for virulence

factors or drug-resistance determinants. WGS analysis of 154 isolates of *Vibrio cholerae* revealed that the recent waves of pandemic of cholera were associated with the acquisition of an SXT/R391 family ICE encoding resistance to several antibiotics [25]. Similarly, the emerging zoonotic pathogen, *Streptococcus suis*, whose major outbreak was reported in China, was shown to carry additional ICEs and transposons coding sequences associated with drug resistance [26]. Using WGS of 240 isolates from the pandemic clone PMEN1, Croucher *et al.* showed the ubiquitous presence of transposon Tn916 carrying inserted genes for resistance to tetracycline and chloramphenicol, which were probably influential in the success of the clone [17]. The analysis also showed that, as the clone spread globally and encountered new selective pressures, there were multiple independent acquisitions of macrolide resistance genes, again inserted into Tn916. Given the high level of sequence conservation found within the species, detection of insertional events in strains of a recent outbreak or of increased virulence provides powerful evidence for the genetic basis of their success.

Widely detected insertion elements—Genome sequencing has highlighted some promiscuous insertion sequences, transposons or ICEs that are present in many bacterial genera. There is experimental evidence that the Tn916 transposon and its derivatives, with resistance to tetracycline, erythromycin and kanamycin, are able to transfer between distant taxa, including transfer between Gram-positive and Gram-negative bacteria *in vitro* [27,28]. However, recent sequence-based analyses have provided a fuller understanding of the extent of Tn916 mobilization within and between species. Based on PCR amplification and sequencing, the phylogeny of *tetM* on Tn916 from separate studies in enterococci [29,30] and staphylococci [31] indicated that *tetM* could be transferred across bacterial genera by conjugative transposons. Recent WGS-based studies have shown that Tn916, and similar transposons, are very common in many species including *S. aureus* [31], *Clostridium difficile* [32] and taxons of streptococci [33-38] and enterococci [39], all with common features in their insertion target sites. Genome analysis will continue to improve our understanding of the evolution and spread of these type of genetic elements, which are clearly important players in the spread of drug resistance and can also play a role in the evolution of virulence.

Insertional mutagenesis—As well as acquisition of potentially beneficial genes, insertion events can also disrupt the integrity of genes, leading to loss of function. In a comparison of genomes from three *Bordetella* species, Parkhill *et al.* demonstrated that insertion and proliferation of insertion sequence elements (ISEs) has led to widespread gene inactivation, especially in *Bordetella pertussis* where massive expansion of one family of ISEs was observed. One particular event caused disruption of a flagellar operon in *Bordetella parapertussis* and *B. pertussis*, resulting in loss of mobility in these species [40]. In general, the extent of disruption caused by the ISEs correlates with a narrowing of niche. *Bordetella bronchiseptica* can infect multiple mammalian species, while *B. pertussis* is restricted to humans [41]. A similar picture was seen for *Burkholderia mallei*, where a stepwise accumulation of insertion sequences appears to have occurred during adaptation to its obligate lifestyle within horses [42]. More recent WGS studies have detected similar patterns in a range of species of pathogenic bacteria, including *Citrobacter* [43] and

Brucella [44], suggesting that the process is common in bacterial evolution, especially where there has been recent adaptation to a new niche.

The principle of gene disruption via transposition can also be applied as an investigative tool. Commonly referred to as Tn-seq, transposon-induced mutagenesis combined with WGS has recently emerged as a powerful method for identifying genes that are essential for bacterial survival under different environmental conditions. The random nature of transposon insertion is exploited to generate large pools of bacterial mutants. An insertion into essential genes would be detrimental under selective conditions and thus would not be detected from the mutant pools. By sequencing and mapping transposon insertion sites from pools of mutants before and after passage through an environmental challenge, genes essential for survival in that condition can be precisely defined. Using this approach, essential genes were identified in *Salmonella enterica* serovar *typhi*, which included genes that contribute towards bile tolerance, a trait required for carriage of *S. typhi* in the gall bladder [45]. The same approach was used to identify genes in *H. influenzae*, required to delay clearance in a murine pulmonary model [46] and genes required for resistance to the amino-glycoside antibiotic tobramycin in *P. aeruginosa* [47]. The method has also been applied to the eukaryotic pathogen, *Candida albicans*, providing a comprehensive fitness profile [48]. The development of high-throughput sequencing has significantly enhanced the investigative power of random transposon mutagenesis, a technique that has been a cornerstone of molecular micro-biology research for decades. Early studies have reaped rich insights that can only be advanced when the technique is applied to more species under more conditions.

Deletions

Although gene loss can be mediated by insertion of transposable elements, gene loss via deletion, sometimes subsequent to gene disruption, is another major force that drives the evolution of pathogenic bacteria.

Tuning virulence—Deletions have been shown to play a role in shaping pathogen genomes and in many cases, differences in deletion sites have been observed between avirulent and virulent strains. Fookes *et al.* compared whole-genome sequences of a group of two *Salmonella* species that share common ancestry, the less virulent *S. bongori*, which rarely infect warm-blooded animals, and the more virulent *S. enterica*, which causes severe infections in warm-blooded hosts [49]. Multiple regions were shown to be missing in *Salmonella bongori*, which might limit its ability to cause disease compared with *S. enterica*. This included a locus termed *Salmonella* pathogenicity island 2 (*SPI-2*), which encodes the type III secretion systems required for optimal replication within macrophages. Also in *Salmonella*, comparison of the genome of *S. enterica* serovar *typhimurium* LT2 strain with that of its more virulent progenitor identified a number of deletions, some within prophage, which give clues to its attenuated virulence phenotype [50].

Deletion events can also have the effect of increasing virulence in bacterial pathogens. The bacterial flagellum is an organelle that provides mobility for the organism and is an important virulence factor in *E. coli*. Genes encoding the flagella are known to display phase

variation, which is thought to allow evasion from the host immune response [51]. Using WGS of *E. coli* H17, Liu *et al.* described a novel mechanism of phase variation where a region encoding flagella *flnA* was deleted [52]. Additional deletion and complementation tests showed that an upstream integrase mediated the excision within the *flnA* region through site-specific recombination.

Genome reduction—The sequences of bacterial genomes have revealed that many species have a mutational bias towards deletions events [53,54], which results in pseudogenization and gene loss [55]. An important deletion mechanism is through recombination between identical ISEs. Losada *et al.* showed that the mechanism was vital for the evolution of *B. mallei* in which the variable gene sets were frequently flanked by ISEs [56]. This genomic arrangement provides a mechanism where gene sets under reduced selection in the mammalian host, may be excised and subsequently removed. Genome reduction due to gene loss is a general pattern seen in the evolutionary transition from facultative to obligate pathogen, demonstrated by WGS studies for several species. Through WGS of *Shigella*, Feng *et al.* identified common genes that had been decayed through deletion and pseudogenization in five *Shigella* lineages. A large number of genes, including those coding for transporter, virulence, carbon utilization, cell motility and membrane proteins, were lost, possibly due to a lack of selection pressure during evolutionary conversion from free-living to intracellular lifestyle [57]. The same pattern is seen in *Listeria* where virulence associated genes have been recurrently lost during a switch from a facultative pathogen to an obligate saprotroph [58]. Deletions of virulent factors – either *prfA* clusters or internalin genes or both – were shown to occur during speciation to *Listeria seeligeri*, *welshimeri*, *innocua* and *marthii*, corresponding to a change to obligate saprotrophs where virulent factors may no longer be necessary. Genome-wide comparison between related free-living and obligate organisms often enables the process of genome decay to be observed where deletions and transposition gene inactivation act together to facilitate this evolutionary process.

Short repetitive sequences

Repeats can be an important source of variation in bacterial genomes, however, studying them on a genome-wide scale can be challenging, due to the problems in differentiating one repeat copy from another either by assembly or read mapping [59,60]. The ability to do this is also dependent on other sequencing parameters, including read length and fragment size. Despite these difficulties, repeats should not be ignored, as they can be instrumental in the evolution and biology of the bacterium.

With the recent development of software designed for detecting miss-assemblies [61] and the application of mate-pair sequencing, which provides information on the pair-wise constraints on the placement of reads [62], repeats can be detected with greater confidence. Moreover, the availability of an increasing number of whole-genome sequences has allowed new repeat elements to be discovered and illustrates the role of repeats in bacterial genome evolution.

Short palindromic repeats as a barrier to horizontal gene transfer—Introduction of foreign DNA via horizontal gene transfer into bacteria can disrupt genome stability and

most species have evolved defensive mechanisms against an invasion of foreign material. Recent interests have focused on a class of repeat arrays called clustered regularly interspaced short palindromic repeats (CRISPR), which act in an interference pathway to limit phage infection and plasmid conjugation. CRISPR were first noticed in *E. coli* [63] and have since been detected in the genomes of many bacterial taxa. CRISPR are present in approximately 40% of bacteria genomes and display a high degree of genetic variability, which likely corresponds with the large diversity of mobile genetic elements to which bacteria are exposed [64].

Genome and transcriptome sequencing have recently enhanced our understanding of the repetitive elements as they allow novel repeat families to be discovered, as well as documenting their transcription profiles. The anatomy of the element was revealed by both capillary and next-generation sequencing, demonstrating that the conserved CRISPR sequences are separated by short variable sequences (spacers), which match bacteriophage or plasmid sequence and thereby specify the targets of interference pathway. Comparison of CRISPR sequences across different bacterial species has identified a set of conserved genes associated with the loci termed CRISPR-associated (*cas*) genes, whose role in mediating RNA silencing was confirmed by functional studies [65-68] and genome-wide transcriptome profiling was used to confirm an upregulation of small CRISPR RNAs during bacteriophage infection [69]. At the population scale, WGS comparative studies have highlighted diversity of spacer sequences, even in a clonal populations [70], implying a role for diversifying selection on this defensive mechanism against foreign DNA [71].

Repetitive sequences that promote gene transfer—As well as being a barrier to horizontal transfer, some classes of repetitive sequences can actually promote the uptake of foreign DNA. A well-known example is the DNA uptake sequence (DUS) in the *Neisseria* genus, which provides recognition to mediate DNA uptake from the environment. WGS was used to identify nearly 2000 copies of 12-bp uptake sequences with different arrangement patterns in each genome, and demonstrated that plasmids containing *Neisseria* DUS were preferentially transformed [72]. A later study, by the same group, showed that DUS are not evenly distributed throughout the genome. Rather, the elements are over-represented in the core genome and under-represented in regions under high diversification [73]. Enrichment of DUS in genes involved in DNA repair, recombination, restriction-modification and replication suggests that *Neisseria* may use transformation to balance out deleterious effects of genome instability in the core genome [74]. Uptake signal sequences, which promote cell competence, were also detected in the genus *Pasteurellaceae*; examples of this genus include the human pathogen *H. influenzae* [75]. WGS has allowed searches for repetitive sequences in other species and an assessment of the potential functionality. Croucher *et al.* identified variation and transcription of repeat sequences in *S. pneumoniae*, another naturally transformable bacteria, including the identification of a previously unrecognized repeat family [76]. The pneumococcal repeats are likely to be simply parasitic elements, but nevertheless, these studies highlight the application of genome and transcriptome sequencing to understanding repetitive sequence in bacterial genomes.

How bacteria evolve: homologous recombination

Homologous recombination is a mechanism allowing the maintenance of genetic diversity in bacterial populations, while counteracting the accumulation of harmful DNA changes in a Muller's ratchet fashion [77,78]. Recombination also provides a mechanism for bacteria to make large evolutionary leaps, such as the acquisition of drug resistance. Classical homologous recombination involves interaction between two sequences with a high nucleotide identity. However, recombination can be responsible for horizontal gene transfer, where more distantly related sequences are exchanged or inserted into bacterial genomes. Additionally, site-specific recombination can mediate the integration of phage genomes or conjugative elements, which involves short stretches of homology between specific sequences of foreign and bacterial DNA. Identification of recombination events from WGS thus provides insights into this important evolutionary force.

Informative investigations into recombination have been possible using sequence data for a handful of genes from large sample databases [79-81] and genome-wide SNP-typing data [82], but the larger genomic datasets present significant computational challenges. Didelot *et al.* searched for recombination in *S. enterica* sub-species *enterica* using a subset of the core genome and showed that recombination has occurred predominantly between members of the same lineage [83].

WGS studies of *S. pneumoniae* have given insights into the selective advantages of recombination. Based on a consideration of both SNP density and phylogeny in the genomes of 240 isolates of the lineage PMEN1 clone [17], greater than 700 recombination events were found throughout the genome with a nonrandom distribution. One of the hotspots for recombination was the locus responsible for the production of the cell's polysaccharide capsule, a surface structure with many different types and the focus of current anti-pneumococcal vaccines (see 'Rate' section above). The authors demonstrated how recombination was responsible for a capsule switch to vaccine-escape serotype 19A, which emerged in the USA following an introduction of a seven-valent conjugate polysaccharide vaccine. Another hotspot was found within Tn916, revealing multiple independent acquisitions of macrolide antibiotic resistance determinants through recombination (see 'Acquisition of virulence & antimicrobial resistance genes' section above). Using the same dataset, Marttinen *et al.* developed a novel Bayesian approach, which allowed the identification of recombinant fragments; this algorithm was able to make use of the large dataset to distinguish the internationally distributed subpopulation from the ancestral European population [84].

Apart from providing selective advantages for existing species, homologous recombination has been shown to drive the process of speciation itself. This principal was illustrated in a WGS study of two populations of *Vibrio cyclitrophicus* that live in different marine habitats [85]. This ecological patterning was explained by a few discrete genomic regions harboring potential habitat-specific genes. These regions were introduced into the ancestral genome of one population by homologous recombination. Moreover, the authors also showed that there was a trend towards genetic exchange within rather than between habitats, which may be accelerating the process of speciation.

Why bacteria evolve: selection

The single base-resolution information gained from WGS allows us to look for evidence of selection. This allows us to investigate the differing selection pressures acting on different species, lineages, gene alleles and even codon and nucleotide positions. Additionally, using *in vitro* evolution of populations of bacteria in combination with WGS allows us to observe evolution in ‘real-time’.

Detecting selection: parallel evolution

Convergent or parallel evolution has long been appreciated as a hallmark of selection. It is indicated by traits reaching a high frequency or fixation in independent lineages. Independent fixation of traits is indicated by the presence of homoplasies, which are variants identified in more than one taxa that are not present in their most recent common ancestor. WGS can be used to observe these patterns of convergent evolution, which allows us to separate adaptive nucleotide changes from neutral ones. This was observed recently in a *Burkholderia dolosa* outbreak in cystic fibrosis patients. This outbreak provided a unique opportunity to observe the course of evolution within 14 patients through the WGS of 112 isolates collected over 16 years. Seventeen genes were found to acquire nonsynonymous mutations independently in separate individuals. Eleven of these belong to functional categories related to pathogenicity, indicating a selection pressure on pathogenic traits during the course of infection. However, interestingly, six of the genes, including the most mutated, have not previously been implicated in the pathogenesis of lung infections. Three of these can be linked through homology to oxygen-dependent gene regulation pathways. It is speculated that this indicates adaptation to the oxygen-depleted mucus environment of the cystic fibrosis lung [86]. As mentioned previously, an accumulation of nonsynonymous SNPs in virulence genes were also observed during chronic *P. aeruginosa* infection of cystic fibrosis patients. The parallels between these two studies suggest that different bacteria can have highly similar ways of adapting during chronic infection [16]. Convergent evolution can also be observed at individual amino acid or nucleotide positions in the genome. One study, comparing 14 *E. coli* genomes, found evidence of positive selection in core genes through the independent generation of variants at the same amino acid position. Interestingly, they found a higher frequency of these ‘hotspot’ mutations in pathogenic species when compared with commensal ones [87]. Homoplasies were also identified through WGS of 63 MRSA strains. Of the 38 homoplastic SNPs identified, 18 were non-synonymous, including ten that correspond to mutations known to confer antibiotic resistance. This indicates that there was frequent independent evolution of resistance. The fact that these traits arise independently in *S. aureus* suggests that resistance occurs via evolution, in addition to the clonal spread of already resistant strains. Understanding how easily antibiotic resistance evolves and subsequently spreads has important implications for our understanding of antibiotic resistance in pathogens, including how it can be tackled and how often we expect it to occur [12].

Detecting selection: synonymous & nonsynonymous nucleotide changes

Measuring the proportion of nonsynonymous nucleotide changes (dN) to synonymous nucleotide changes (dS) is a commonly used approach for detecting selection in DNA

sequence data. In general, a $dN/dS > 1$ indicates positive selection, < 1 indicates purifying selection and a ratio at or close to 1 neutral selection (or a balance of positive and purifying selection). Although there are disadvantages to summarizing a complicated spectrum of selection across a whole genome, it can be used as a rough measure of selection to give us insights into species evolution. Comparison of whole-genome sequences has revealed some surprises. Rocha *et al.* compared the genomes of 31 bacterial isolates representing six genera. They found that the more closely related the genomes were, the higher the dN/dS . This indicates that dN/dS , and thus selection, are time dependent. In the species studied, purifying selection effectively acts as a sieve, removing slightly deleterious mutations over time. However, this lag is different between different taxa, probably due to differences in the effective population size, which can have an effect on the efficiency of selection [88]. This pattern has also been observed within a species through WGS of 21 *C. difficile* isolates. The rate of reduction of dN/dS over genetic distance was found to be less than for *E. coli*, indicating a longer delay in the purging of slightly deleterious SNPs, which could possibly reflect a smaller effective population size due to a more restricted biological niche [89].

dN/dS can also reveal differences in selection acting on individual genes. A metagenomic dataset was used to sample a population of *Synechococcus*, a genus of cyanobacteria, from a coastal environment with reads aligned to two available *Synechococcus* reference genomes. Most genes (98%) were found to be under purifying selection, but some genes were found to have a $dN/dS > 1$. Of these, 93% were annotated as hypothetical [90]. This demonstrates the scarcity of functional knowledge we have for a vast proportion of bacterial genes and how they may be important for our understanding of bacterial evolution.

A $dN/dS > 1$ indicates positive or diversifying selection, so it makes sense that often these genes are involved in pathogenic interactions with the host. Twelve complete *Streptococcus* genomes were used to identify genes under positive selection, of which 29% were associated with virulence [91]. Identifying proteins that are involved in the evolutionary arms race between pathogens and their host are essential for vaccine design, as they are probably unsuitable [92]. However, dN/dS is sometimes unsuitable for short-term measures of selection due to the small amount of variation generated.

Observing selection as it happens

Richard Lenski's group at Michigan State University (MI, USA) have been running a long-term experiment (since 1988) tracking genetic changes in 12 initially identical populations of *E. coli* through over 50,000 cell cycle generations, providing an excellent opportunity to observe bacterial evolution [93]. Recent results utilized WGS to investigate the concept of 'evolvability' during a competition experiment. WGS and subsequent mapping to a reference genome enabled the identification of SNPs, in addition to small and large insertions/deletions, to explore their involvement in the success of the clones. By sequencing, the 'eventual winners' they were able to show that they had a greater propensity for adaptability than the 'eventual losers'. They were able to 'replay' evolution by reviving frozen stocks taken at 500 generations into the experiment, to show that the 'eventual winners' did not win by pure chance. Although initially they had a fitness deficit, the 'eventual winners' always overcame this deficit to have a 2.1% greater fitness on average

than the ‘eventual losers’. The authors speculate that this is due to negative epistatic interactions between early mutations found in the ‘eventual loser’, which may have prevented further adaptive changes [94]. Experiments with *E. coli* can also be used to investigate evolutionary paths to a particular trait. Recent work by Toprak *et al.* used a device (morbidosat) that monitors bacterial growth and dynamically regulates antibiotic concentrations in the growth medium. They found that parallel populations followed similar phenotypic trajectories. Comparing different drugs, they found that trimethoprim resistance increased in a stepwise fashion, whereas resistance to chloramphenicol and doxycycline proceeded at a smooth rate. The stepwise trajectory was confirmed by WGS to be due to the rarity of the resistance-conferring mutations and their confinement to one gene (*DHFR*), whereas resistance to chloramphenicol and doxycycline was caused by a diverse combination of SNPs found in many gene classes. This means the waiting time for a trimethoprim resistance mutation is longer than for the other drugs but also that rising resistance levels occur in jumps [95].

The ‘red queen’ hypothesis proposes that continuous adaption is required for a species to maintain relative fitness to the organisms coevolving with it [96]; this is often used to describe host-parasite evolutionary relationships. Recently, this hypothesis was elegantly demonstrated through coevolution experiments with a bacterium, *Pseudomonas fluorescens*, and its viral parasite. The bacteriophage and bacteria can be separated while transferring the populations to a fresh media, which allows the replacement of one of the populations. Evolution experiments were carried out under two conditions: where both bacteria and bacteriophage coevolved and where the bacterial population was kept at a constant genotype. WGS demonstrated that the rate of molecular evolution was higher in the bacteriophage when it coevolved with the bacteria, than when it was evolving with bacteria held at a constant genotype. This antagonistic coevolution demonstrates how the speed of selection is very much affected by the fitness of the organisms it interacts with [97]. These experiments demonstrate how evolutionary experiments can be used in combination with WGS to produce a powerful tool to test evolutionary hypotheses. This is not only important for understanding bacterial evolution, but can also serve as useful models to test hypotheses of natural selection itself.

Future perspective

We stand at a pivotal moment in biological science where the acquisition of a complete genome sequence, something that a decade ago took millions of dollars and several years to achieve, is becoming trivial. Technologies are burgeoning that will allow the generation of complete bacterial genomes within hours for a few tens of dollars and the machinery could be as convenient as a USB drive plugged into your computer. Trends are towards higher accuracy and longer read lengths, thus analyses should become simpler, quicker and better able to resolve complex rearrangements; however, data storage will continue to be a challenge and wider use of distributed systems seem inevitable [98].

Given a robust ability to analyze data, the falling costs of genome sequencing will see its routine introduction in many areas, with perhaps the most exciting in terms of pathogen research, being public health clinical microbiology. Here a plethora of current, often species-

specific techniques could be superseded by the exhaustive and unifying sampling of whole genomes from which detailed genotypes can be acquired and phenotypes derived. These data will reveal the details of pathogen evolution in response to clinical and environmental selective forces and will help to identify emerging new clones.

Advances in sequencing technology are also gradually bringing metagenomics within reach. The current state-of-the-art technology generates heavily fragmented genomes limiting the power of interpretation. Longer reads hold the promise of easily assembled whole genomes, which will help to reveal the true extent of horizontal gene flow between clones and species in the natural niche, be it a human body site or a host-independent environment, such as soil. This will enable a better sampling of the sources of gene pools available to pathogen genomes and the study of microbial populations as a whole. The plethora of new data available will drive the discovery and testing of more evolutionary hypotheses and provide us with an increased understanding of how pathogens evolve.

References

Papers of special note have been highlighted as:

■ of interest

■ ■ of considerable interest

1. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995; 269(5223):496–512. [PubMed: 7542800]
2. Fraser CM, Gocayne JD, White O, et al. The minimal gene complement of *Mycoplasma genitalium*. *Science*. 1995; 270(5235):397–403. [PubMed: 7569993]
3. Wassenaar TM, Gaastra W. Bacterial virulence: can we draw the line? *FEMS Microbiol. Lett.* 2001; 201(1):1–7. [PubMed: 11445159]
4. Nowrousian M, Teichert I, Masloff S, Kück U. Whole-genome sequencing of *Sordaria macrospora* mutants identifies developmental genes. *G3*. 2012; 2(2):261–270. [PubMed: 22384404]
5. Renzoni A, Andrey DO, Jouselin A, et al. Whole genome sequencing and complete genetic analysis reveals novel pathways to glycopeptide resistance in *Staphylococcus aureus*. *PLoS One*. 2011; 6(6):e21577. [PubMed: 21738716]
6. Comas I, Borrell S, Roetzer A, et al. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* 2011; 44(1):106–110. [PubMed: 22179134]
7. Gagneux S, Long CD, Small PM, Van T, Schoolnik GK, Bohannon BJ. The competitive cost of antibiotic resistance in *Mycobacterium tuberculosis*. *Science*. 2006; 312(5782):1944–1946. [PubMed: 16809538]
8. Ochman H. Neutral mutations and neutral substitutions in bacterial genomes. *Mol. Biol. Evol.* 2003; 20(12):2091–2096. [PubMed: 12949125]
9. Cupples CG, Miller JH. A set of *lacZ* mutations in *Escherichia coli* that allow rapid detection of each of the six base substitutions. *Proc. Natl Acad. Sci. USA*. 1989; 86(14):5345–5349. [PubMed: 2501784]
10. Ochman H, Elwyn S, Moran NA. Calibrating bacterial evolution. *Proc. Natl Acad. Sci. USA*. 1999; 96(22):12638–12643. [PubMed: 10535975]
11. Sharp PM, Li WH. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* 1987; 4(3):222–230. [PubMed: 3328816]
12. Harris SR, Feil EJ, Holden MT, et al. Evolution of MRSA during hospital transmission and intercontinental spread. *Science*. 2010; 327(5964):469–474. [PubMed: 20093474] [■ ■ Landmark]

paper that demonstrates the power of next-generation sequencing to detect transmission of pathogens on both local and international scales.]

13. Ho SY, Shapiro B, Phillips MJ, Cooper A, Drummond AJ. Evidence for time dependency of molecular rate estimates. *Syst. Biol.* 2007; 56(3):515–522. [PubMed: 17562475]
14. Dye C, Williams BG. The population dynamics and control of tuberculosis. *Science.* 2010; 328(5980):856–861. [PubMed: 20466923]
15. Ford CB, Lin PL, Chase MR, et al. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat. Genet.* 2011; 43(5):482–486. [PubMed: 21516081] [■ Used whole-genome sequencing to challenge the idea that *Mycobacterium tuberculosis* does not evolve during latent infection.]
16. Smith EE, Buckley DG, Wu Z, et al. Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc. Natl Acad. Sci. USA.* 2006; 103(22):8487–8492. [PubMed: 16687478]
17. Croucher NJ, Harris SR, Fraser C, et al. Rapid pneumococcal evolution in response to clinical interventions. *Science.* 2011; 331(6016):430–434. [PubMed: 21273480]
18. Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 2010; 6(9):e1001115. [PubMed: 20838599]
19. Foerster KU, Von Mering C, Hooper SD, Bork P. Environments shape the nucleotide composition of genomes. *EMBO Rep.* 2005; 6(12):1208–1213. [PubMed: 16200051]
20. Parkhomchuk D, Amstislavskiy V, Soldatov A, Ogryzko V. Use of high throughput sequencing to observe genome dynamics at a single cell level. *Proc. Natl Acad. Sci. USA.* 2009; 106(49):20830–20835. [PubMed: 19934054]
21. Lind PA, Andersson DI. Whole-genome mutational biases in bacteria. *Proc. Natl Acad. Sci. USA.* 2008; 105(46):17878–17883. [PubMed: 19001264]
22. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 2011; 12(5):363–376. [PubMed: 21358748]
23. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods.* 2009; 6(11 Suppl.):S13–S20. [PubMed: 19844226]
24. Novick RP, Christie GE, Penades JR. The phage-related chromosomal islands of Gram-positive bacteria. *Nat. Rev. Microbiol.* 2010; 8(8):541–551. [PubMed: 20634809]
25. Mutreja A, Kim DW, Thomson NR, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature.* 2011; 477(7365):462–465. [PubMed: 21866102]
26. Holden MT, Hauser H, Sanders M, et al. Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*. *PLoS One.* 2009; 4(7):e6072. [PubMed: 19603075]
27. Roberts AP, Cheah G, Ready D, Pratten J, Wilson M, Mullany P. Transfer of Tn916-like elements in microcosm dental plaques. *Antimicrob. Agents Chemother.* 2001; 45(10):2943–2946. [PubMed: 11557498]
28. Bertram J, Stratz M, Durre P. Natural transfer of conjugative transposon Tn916 between Gram-positive and Gram-negative bacteria. *J. Bacteriol.* 1991; 173(2):443–448. [PubMed: 1846142]
29. Rizzotti L, La Gioia F, Dellaglio F, Torriani S. Molecular diversity and transferability of the tetracycline resistance gene *tet(M)*, carried on Tn916-1545 family transposons, in enterococci from a total food chain. *Antonie Van Leeuwenhoek.* 2009; 96(1):43–52. [PubMed: 19333776]
30. Agerso Y, Pedersen AG, Aarestrup FM. Identification of Tn5397-like and Tn916-like transposons and diversity of the tetracycline resistance gene *tet(M)* in enterococci from humans, pigs and poultry. *J. Antimicrob. Chemother.* 2006; 57(5):832–839. [PubMed: 16565159]
31. De Vries LE, Christensen H, Skov RL, Aarestrup FM, Agerso Y. Diversity of the tetracycline resistance gene *tet(M)* and identification of Tn916- and Tn5801-like (Tn6014) transposons in *Staphylococcus aureus* from humans and animals. *J. Antimicrob. Chemother.* 2009; 64(3):490–500. [PubMed: 19531603]
32. Mullany P, Williams R, Langridge GC, et al. Behaviour and target site selection of the conjugative transposon Tn916 in two different strains of toxigenic *Clostridium difficile*. *Appl. Environ. Microbiol.* 2012; 78(7):2147–2153. [PubMed: 22267673]

33. Del Grosso M, Scotto d'Abusco A, Iannelli F, Pozzi G, Pantosti A. Tn2009, a Tn916-like element containing *mef(E)* in *Streptococcus pneumoniae*. Antimicrob. Agents Chemother. 2004; 48(6): 2037–2042. [PubMed: 15155196]
34. Quintero B, Araque M, Van Der Gaast-De Jongh C, Hermans PW. Genetic diversity of Tn916-related transposons among drug-resistant *Streptococcus pneumoniae* isolates colonizing healthy children in Venezuela. Antimicrob. Agents Chemother. 2011; 55(10):4930–4932. [PubMed: 21788464]
35. Mingoia M, Tili E, Manso E, Varaldo PE, Montanari MP. Heterogeneity of Tn5253-like composite elements in clinical *Streptococcus pneumoniae* isolates. Antimicrob. Agents Chemother. 2011; 55(4):1453–1459. [PubMed: 21263055]
36. Ciric L, Mullany P, Roberts AP. Antibiotic and antiseptic resistance genes are linked on a novel mobile genetic element: Tn6087. J. Antimicrob. Chemother. 2011; 66(10):2235–2239. [PubMed: 21816764]
37. Hraoui M, Boutiba-Ben Boubaker I, Doloy A, Ben Redjeb S, Bouvet A. Molecular mechanisms of tetracycline and macrolide resistance and emm characterization of *Streptococcus pyogenes* isolates in Tunisia. Microb. Drug Resist. 2011; 17(3):377–382. [PubMed: 21612508]
38. Haenni M, Saras E, Bertin S, Leblond P, Madec JY, Payot S. Diversity and mobility of integrative and conjugative elements in bovine isolates of *Streptococcus agalactiae*, *S. dysgalactiae* subsp. *dysgalactiae*, and *S. uberis*. Appl. Environ. Microbiol. 2010; 76(24):7957–7965. [PubMed: 20952646]
39. Sun J, Sundsfjord A, Song X. *Enterococcus faecalis* from patients with chronic periodontitis: virulence and antimicrobial resistance traits and determinants. Eur. J. Clin. Microbiol. Infect. Dis. 2011; 31(3):267–272. [PubMed: 21660501]
40. Parkhill J, Sebaihia M, Preston A, et al. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat. Genet. 2003; 35(1):32–40. [PubMed: 12910271]
41. Bemis, DA. *Bordetella*. 3rd ed.. Blackwell Publishing; UK: 2004. p. 259–272.
42. Song H, Hwang J, Yi H, et al. The early stage of bacterial genome-reductive evolution in the host. PLoS Pathog. 2010; 6(5):e1000922. [PubMed: 20523904]
43. Petty NK, Feltwell T, Pickard D, et al. *Citrobacter rodentium* is an unstable pathogen showing evidence of significant genomic flux. PLoS Pathog. 2011; 7(4):e1002018. [PubMed: 21490962]
44. Audic S, Lescot M, Claverie JM, Cloeckert A, Zygmunt MS. The genome sequence of *Brucella pinnipedialis* B2/94 sheds light on the evolutionary history of the genus *Brucella*. BMC Evol. Biol. 2011; 11:200. [PubMed: 21745361]
45. Langridge GC, Phan MD, Turner DJ, et al. Simultaneous assay of every *Salmonella typhi* gene using one million transposon mutants. Genome Res. 2009; 19(12):2308–2316. [PubMed: 19826075]
46. Gawronski JD, Wong SM, Giannoukos G, Ward DV, Akerley BJ. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. Proc. Natl Acad. Sci. USA. 2009; 106(38):16422–16427. [PubMed: 19805314]
47. Gallagher LA, Shendure J, Manoel C. Genome-scale identification of resistance functions in *Pseudomonas aeruginosa* using Tn-seq. MBio. 2011; 2(1):e00315–10. [PubMed: 21253457]
48. Oh J, Fung E, Price MN, et al. A universal TagModule collection for parallel genetic analysis of microorganisms. Nucleic Acids Res. 2010; 38(14):e146. [PubMed: 20494978]
49. Fookes M, Schroeder GN, Langridge GC, et al. *Salmonella bongori* provides insights into the evolution of the Salmonellae. PLoS Pathog. 2011; 7(8):e1002191. [PubMed: 21876672]
50. Jarvik T, Smillie C, Groisman EA, Ochman H. Short-term signatures of evolutionary change in the *Salmonella enterica* serovar *typhimurium* 14028 genome. J. Bacteriol. 2010; 192(2):560–567. [PubMed: 19897643]
51. Bonifield HR, Hughes KT. Flagellar phase variation in *Salmonella enterica* is mediated by a posttranscriptional control mechanism. J. Bacteriol. 2003; 185(12):3567–3574. [PubMed: 12775694]

52. Liu B, Hu B, Zhou Z, et al. A novel nonhomologous recombination-mediated mechanism for *Escherichia coli* unilateral flagellar phase variation. *Nucleic Acids Res.* 2012; 40(10):4530–4538. [PubMed: 22287625]
53. Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 2001; 17(10):589–596. [PubMed: 11585665]
54. Moran NA. Microbial minimalism: genome reduction in bacterial pathogens. *Cell.* 2002; 108(5): 583–586. [PubMed: 11893328]
55. Toft C, Andersson SG. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat. Rev. Genet.* 2010; 11(7):465–475. [PubMed: 20517341]
56. Losada L, Ronning CM, Deshazer D, et al. Continuing evolution of *Burkholderia mallei* through genome reduction and large-scale rearrangements. *Genome Biol. Evol.* 2010; 2:102–116. [PubMed: 20333227]
57. Feng Y, Chen Z, Liu SL. Gene decay in *Shigella* as an incipient stage of host-adaptation. *PLoS One.* 2011; 6(11):e27754. [PubMed: 22110755]
58. Den Bakker HC, Cummings CA, Ferreira V, et al. Comparative genomics of the bacterial genus *Listeria*. Genome evolution is characterized by limited gene acquisition and limited gene loss. *BMC Genomics.* 2010; 11:688. [PubMed: 21126366]
59. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 2011; 13(1):36–46. [PubMed: 22124482]
60. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nat. Methods.* 2011; 8(1):61–65. [PubMed: 21102452]
61. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.* 2008; 9(3):R55. [PubMed: 18341692]
62. Wetzel J, Kingsford C, Pop M. Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. *BMC Bioinformatics.* 2011; 12:95. [PubMed: 21486487]
63. Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J. Bacteriol.* 1987; 169(12):5429–5433. [PubMed: 3316184]
64. Grissa I, Vergnaud G, Poursel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics.* 2007; 8:172. [PubMed: 17521438]
65. Delihias N. Impact of small repeat sequences on bacterial genome evolution. *Genome Biol. Evol.* 2011; 3:959–973. [PubMed: 21803768]
66. Marraffini LA, Sontheimer EJ. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* 2010; 11(3):181–190. [PubMed: 20125085]
67. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. *Science.* 2010; 327(5962):167–170. [PubMed: 20056882]
68. Karginov FV, Hannon GJ. The CRISPR system: small RNA-guided defense in bacteria and archaea. *Mol. Cell.* 2010; 37(1):7–19. [PubMed: 20129051]
69. Agari Y, Sakamoto K, Tamakoshi M, Oshima T, Kuramitsu S, Shinkai A. Transcription profile of *Thermus thermophilus* CRISPR systems after phage infection. *J. Mol. Biol.* 2010; 395(2):270–281. [PubMed: 19891975]
70. Diez-Villasenor C, Almendros C, Garcia-Martinez J, Mojica FJ. Diversity of CRISPR loci in *Escherichia coli*. *Microbiology.* 2010; 156(Pt 5):1351–1361. [PubMed: 20133361]
71. Haerter JO, Trusina A, Sneppen K. Targeted bacterial immunity buffers phage diversity. *J. Virol.* 2011; 85(20):10554–10560. [PubMed: 21813617]
72. Ambur OH, Frye SA, Tonjum T. New functional identity for the DNA uptake sequence in transformation and its presence in transcriptional terminators. *J. Bacteriol.* 2007; 189(5):2077–2085. [PubMed: 17194793]
73. Treangen TJ, Ambur OH, Tonjum T, Rocha EP. The impact of the neisserial DNA uptake sequences on genome evolution and stability. *Genome Biol.* 2008; 9(3):R60. [PubMed: 18366792]
74. Schoen C, Tettelin H, Parkhill J, Frosch M. Genome flexibility in *Neisseria meningitidis*. *Vaccine.* 2009; 27(Suppl. 2):B103–B111. [PubMed: 19477564]

75. Redfield RJ, Findlay WA, Bosse J, Kroll JS, Cameron AD, Nash JH. Evolution of competence and DNA uptake specificity in the Pasteurellaceae. *BMC Evol. Biol.* 2006; 6:82. [PubMed: 17038178]
76. Croucher NJ, Vernikos GS, Parkhill J, Bentley SD. Identification, variation and transcription of pneumococcal repeat sequences. *BMC Genomics.* 2011; 12:120. [PubMed: 21333003]
77. Muller HJ. The need for recombination to prevent genetic deterioration. *Genetics.* 1964; 48:903–903.
78. Moran NA. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA.* 1996; 93(7):2873–2878. [PubMed: 8610134]
79. Marttinen P, Baldwin A, Hanage WP, Dowson C, Mahenthiralingam E, Corander J. Bayesian modeling of recombination events in bacterial populations. *BMC Bioinformatics.* 2008; 9:421. [PubMed: 18840286]
80. Hanage WP, Fraser C, Tang J, Connor TR, Corander J. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. *Science.* 2009; 324(5933):1454–1457. [PubMed: 19520963]
81. Corander J, Connor TR, O'Dwyer CA, Kroll JS, Hanage WP. Population structure in the *Neisseria*, and the biological significance of fuzzy species. *J. R. Soc. Interface.* 2011; 9(71):1208–1215. [PubMed: 22072450]
82. Pearson T, Giffard P, Beckstrom-Sternberg S, et al. Phylogeographic reconstruction of a bacterial species with high levels of lateral gene transfer. *BMC Biol.* 2009; 7:78. [PubMed: 19922616]
83. Didelot X, Bowden R, Street T, et al. Recombination and population structure in *Salmonella enterica*. *PLoS Genet.* 2011; 7(7):e1002191. [PubMed: 21829375]
84. Marttinen P, Hanage WP, Croucher NJ, et al. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* 2012; 40(1):e6. [PubMed: 22064866]
85. Shapiro BJ, Friedman J, Cordero OX, et al. Population genomics of early events in the ecological differentiation of bacteria. *Science.* 2012; 336(6077):48–51. [PubMed: 22491847] [■ Elegantly demonstrated that homologous recombination can drive early ecological differentiation of bacteria.]
86. Lieberman TD, Michel JB, Aingaran M, et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* 2011; 43(12):1275–1280. [PubMed: 22081229] [■ Whole-genome sequencing was used to detect parallel evolution of a pathogen between cystic fibrosis patients.]
87. Chattopadhyay S, Weissman SJ, Minin VN, Russo TA, Dykhuizen DE, Sokurenko EV. High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *Proc. Natl Acad. Sci. USA.* 2009; 106(30):12412–12417. [PubMed: 19617543]
88. Rocha EP, Smith JM, Hurst LD, et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* 2006; 239(2):226–235. [PubMed: 16239014]
89. He M, Sebaihia M, Lawley TD, et al. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc. Natl Acad. Sci. USA.* 2010; 107(16):7527–7532. [PubMed: 20368420]
90. Tai V, Poon AF, Paulsen IT, Palenik B. Selection in coastal *Synechococcus* (cyanobacteria) populations evaluated from environmental metagenomes. *PLoS One.* 2011; 6(9):e24249. [PubMed: 21931665]
91. Anisimova M, Bielawski J, Dunn K, Yang Z. Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. *BMC Evol. Biol.* 2007; 7:154. [PubMed: 17760998]
92. Aguilera G, Refregier G, Yockteng R, Fournier E, Giraud T. Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infect. Genet. Evol.* 2009; 9(4):656–670. [PubMed: 19442589]
93. Lenski RE. Evolution in action: a 50,000-generation salute to Charles Darwin. *Microbe.* 2011; 6:30–33. [■ Describes an important ongoing long-term *Escherichia coli* experiment that has and will provide insights into natural selection.]
94. Woods RJ, Barrick JE, Cooper TF, Shrestha U, Kauth MR, Lenski RE. Second-order selection for evolvability in a large *Escherichia coli* population. *Science.* 2011; 331(6023):1433–1436. [PubMed: 21415350]

95. Toprak E, Veres A, Michel JB, Chait R, Hartl DL, Kishony R. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat. Genet.* 2011; 44(1):101–105. [PubMed: 22179135]
96. Van Valen L. Molecular evolution as predicted by natural selection. *J. Mol. Evol.* 1974; 3(2):89–101. [PubMed: 4407466]
97. Paterson S, Vogwill T, Buckling A, et al. Antagonistic coevolution accelerates molecular evolution. *Nature.* 2010; 464(7286):275–278. [PubMed: 20182425]
98. Baker M. Next-generation sequencing: adjusting to data overload. *Nat. Methods.* 2010; 7(7):495–499.
99. Seth-Smith HM, Harris SR, Persson K, et al. Co-evolution of genomes and plasmids within *Chlamydia trachomatis* and the emergence in Sweden of a new variant strain. *BMC Genomics.* 2009; 10:239. [PubMed: 19460133]

Executive summary

How bacteria evolve: point mutations

- Precise derivation of bacterial mutation rates will enable better identification of pathogen transmissions and outbreaks, as well as prediction of adaptation during pathogenesis and application of clinical interventions.
- The distribution and type of point mutations that occur in bacteria have been found to be nonrandom.

How bacteria evolve: large sequence variants

- Widespread sampling of mobile genetic elements will reveal how they are acquired and spread among pathogen populations.
- Gene loss and genome reduction have frequently been found to be associated with the evolution of virulence and niche adaptation in pathogens.

How bacteria evolve: homologous recombination

- Understanding homologous recombination has important implications for the spread of drug resistance.

Why bacteria evolve: selection

- Sequences reveal genes under selection that give clues to host and environmental interactions and provide useful targets for new clinical interventions.

Figure 1. Mapping and assembly with sequencing reads

Both mapping to a high-quality reference and *de novo* assembly of sequencing reads can provide us with evolutionary information. However, the resolution of this information and their possible applications vary. Mapping enables the rapid identification of high-quality SNPs, which can be used to build phylogenies for both evolutionary and epidemiological inference. However, *de novo* assembly is required to study larger variants, such as insertion/deletions, mobile elements and rearrangements.

SNP: Single-nucleotide polymorphism.

Figure 2. Detecting single nucleotide variation and large-scale insertions and deletions from read mapping

An example looking at *Chlamydia trachomatis* chromosome and plasmid sequences. The pale blue bar highlights a deletion site observed from read data which has significant implications for the effective typing of this pathogen (see [99] for further details).

(A) Depth of coverage for reads from a test genome mapped against a reference genome sequence represented by the orange bar (chromosome) and brown bar (plasmid). The depth of coverage steps up in the plasmid region due to its higher copy number per cell relative to the chromosome. Towards the right of the plasmid region the plot drops to zero, indicating a deletion in the test sample relative to the reference. (B) Stack view of the mapped reads showing those that are exactly equivalent in sequence and length to each other (green), those that are unique in sequence and length (blue) and positions where the reads disagree with the reference (red). Sporadic red marks likely represent errors, while true variants are indicated by vertical red lines where all reads mapped to that position indicate the variant. As with (A), the deletion region shows no mapped reads. (C) Insertion/deletion size can be estimated by plotting read pair inferred size on a log scale. The (log) mapped insert size is plotted on the y-axis to show that, where there is an insertion/deletion, the mapped insert size is increased relative to the mean insert size, which is the true fragment length. In essence, by mapping to a reference that does not have the deletion, we artificially introduce sequence length between the reads, thus increasing the mapped insert size. This reveals the presence of the deletion by the increase in inferred size and drop in coverage. The inferred insert size calculated from this subset of reads is far bigger than the normal size range of read pairs in this region. In addition, the absence of lines linking paired reads within the normal size range across this region is indicative of deletion. Image provided courtesy of S Harris and T Carver (Wellcome Trust Sanger Institute, UK).