NEW TECHNOLOGIES: METHODS AND APPLICATIONS — TIMELINE

Twenty years of bacterial genome sequencing

Nicholas J. Loman and Mark J. Pallen

Abstract | Twenty years ago, the publication of the first bacterial genome sequence, from *Haemophilus influenzae*, shook the world of bacteriology. In this Timeline, we review the first two decades of bacterial genome sequencing, which have been marked by three revolutions: whole-genome shotgun sequencing, high-throughput sequencing and single-molecule long-read sequencing. We summarize the social history of sequencing and its impact on our understanding of the biology, diversity and evolution of bacteria, while also highlighting spin-offs and translational impact in the clinic. We look forward to a 'sequencing singularity', where sequencing becomes the method of choice for as-yet unthinkable applications in bacteriology and beyond.

Bacterial genome sequencing is now 20 years old. During this period, the powerful combination of genome sequencing and bioinformatics-driven analysis of sequence data has transformed our understanding of how bacteria function, evolve and interact with each other, with their hosts, and with their surroundings, while also providing numerous avenues for translational impact. Sequence-based analyses have delivered unexpected insights into microbial diversity — from strains to super-phyla — and have allowed us to explore microbial communities. Such approaches have also allowed us to track the spread of infection and helped us devise new drugs and vaccines. We now face the imminent transition of genome sequencing and bioinformatics into the clinic, and the arrival of real-time monitoring of infectious disease outbreaks. The process of sequencing has seen remarkable innovation, so that sequencing projects that used to take years and cost hundreds of thousands of dollars can now be completed in a few days for less than the price of a meal out for two. However, with sequencing no longer a bottleneck, it can take much longer to analyse than to generate sequence data, which brings the problems of big data to bacterial genomics.

In this Timeline article, we present a brief history of the major events that have shaped the sequencing and analysis of bacterial genomes in the past two decades (FIG. 1). We look back to the 1990s and forward to the next decade, and present a chronology that encompasses three technological revolutions: whole-genome shotgun sequencing, high-throughput sequencing and single-molecule long-read sequencing (FIG. 2). Additionally, we highlight scientific and cultural milestones for each phase. We invite the reader to join us on this roller-coaster ride of discovery.

The first revolution

Whole-genome shotgun sequencing. The bacterial genome-sequencing revolution was initiated in the early 1990s, with the launch of consortium-led projects to sequence the genomes of model organisms, such as *Escherichia coli* and *Bacillus subtilis*[1,2] (BOX 1). However, the 'big bang' came in 1995 when Craig Venter, Hamilton Smith and their associates performed the first shotgun sequencing of entire bacterial genomes[3] (FIG. 2). Ironically, the first bacterium to be genome-sequenced was a non-pathogenic strain of *Haemophilus influenzae*, which Smith happened to have to hand because he had used it to obtain the restriction enzyme HindIII in the work that won him the Nobel Prize in Physiology or Medicine in 1978 (with Werner Arber and Daniel Nathans). The first genome paper largely contained a technical description of the method, with few references to the organism's biology[3]. However, it jump-started a race to sequence genomes from pathogens, model organisms and extremophiles. In pursuit of completed genomes, there was an exhortation to "bang out every base and close every gap" (REF. 4), and there was work enough for multiple sequencers, bioinformaticians and annotators on both sides of the Atlantic (BOX 2).

Over the years that followed, we caught a first glimpse of the inner workings of our most fearful microbial adversaries, from the cause of the 'white plague', *Mycobacterium tuberculosis*[5], to the agent of the Black Death, *Yersinia pestis*[6]. Even for model organisms like *E. coli* K-12 (REF. 7) and *B. subtilis*[8], the first genome sequences delivered thousands of new genes, and genome sequencing provided an exciting route to the reconstruction of organismal biology for organisms that were hard or impossible to study *in vitro*, including pathogens like *Treponema pallidum*[9], *Mycobacterium leprae*[10] or *Tropheryma whipplei*[11], or extremophiles like *Deinococcus radiodurans*[12]. For *T. whipplei*, metabolic reconstructions based on this novel genomic information even allowed the design of an axenic growth medium for the organism, which was previously unculturable[13].

Comparative genomics. Analyses of diverse new bacterial genomes revealed important differences in genomic composition and organization from the *E. coli* paradigm. For example, the *Campylobacter jejuni* genome was found to contain several dozen hypervariable homopolymeric repeats (tandem repeats of the same base), concentrated in genes that were responsible for the biosynthesis or modification of surface structures[14]. Similarly, the *Bacteroides fragilis* genome was found to house multiple inverted DNA repeats that mediated antigenic variation in polysaccharides[15]. The *Y. pestis* and *Bordetella pertussis* genomes contained large-scale genomic rearrangements despite having conserved species-specific gene sets[6].

As we entered a new millennium, we started to gain multiple genomes from the same genus or species[16–19]. Given the large-scale conservation of gene order in genetic maps of *E. coli* and *Salmonella enterica*[20], one might have expected one *E. coli* genome to be almost identical to any other *E. coli* genome; instead, the first three *E. coli* genomes revealed an unexpected role for horizontal gene transfer (HGT) in generating strain-to-strain diversity in this species[7,21,22]. Similarly, the genome sequence of the thermophilic bacterium *Thermotoga maritima* provided evidence for extensive HGT between Archaea and Bacteria[23].

Following these findings, it soon became clear that no single conceptual framework could be applied to the genome dynamics of all bacterial lineages. Some lineages were 'celibate', refraining from sexual exchange of DNA and thus showing limited genetic diversity and little or no evidence of recombination or HGT. Several important human pathogens

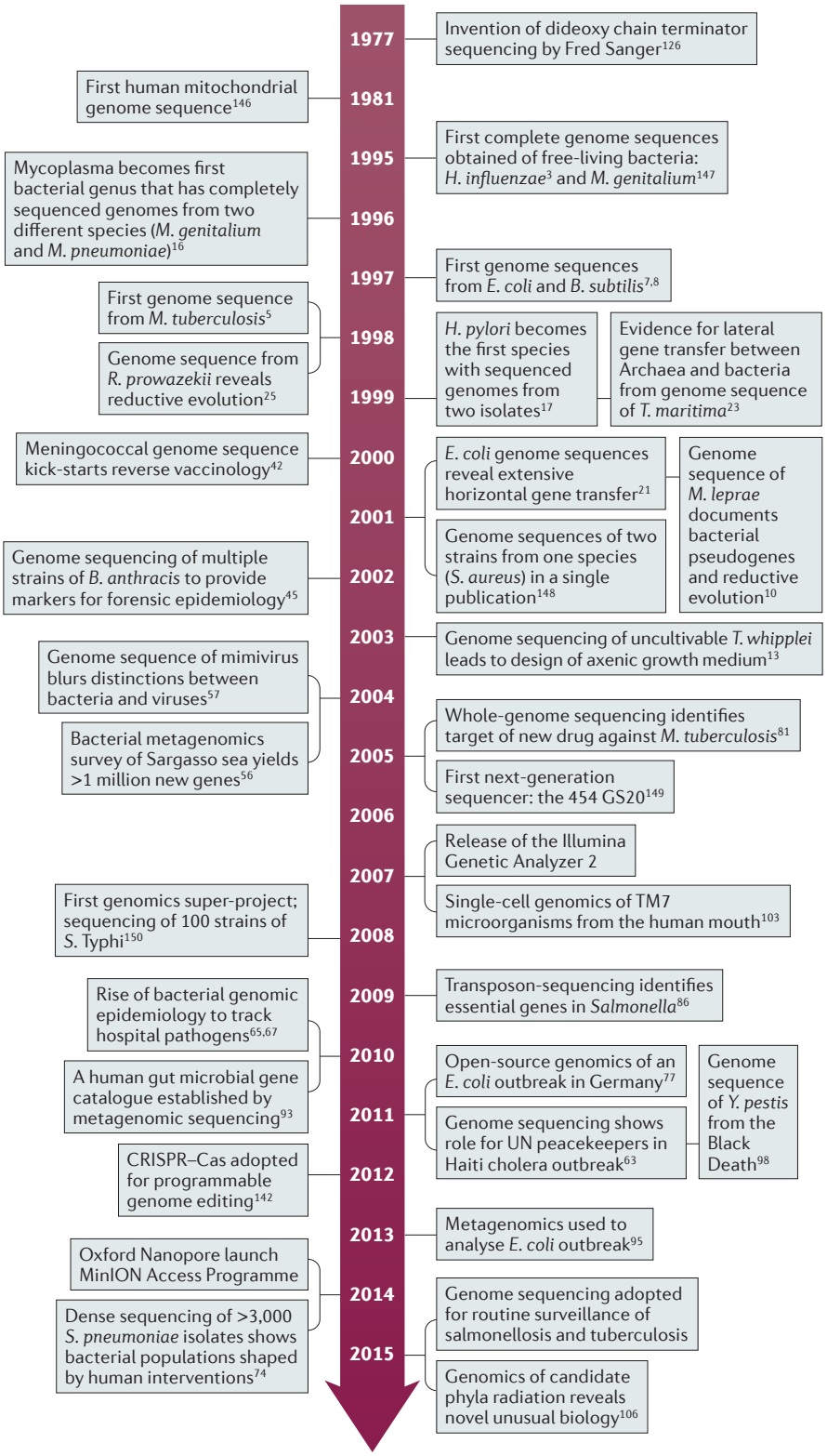

Figure 1 | **Milestones in bacterial genome sequencing.** *B. anthracis*, *Bacillus anthracis*; *B. subtilis*, *Bacillus subtilis*; Cas, CRISPR-associated; CRISPR, clustered regularly interspaced short palindrome repeats; *E. coli*, *Escherichia coli*; *H. influenzae*, *Haemophilus influenzae*; *H. pylori*, *Helicobacter pylori*; *M. genitalium*, *Mycoplasma genitalium*; *M. leprae*, *Mycobacterium leprae*; *M. pneumoniae*, *Mycoplasma pneumoniae*; *M. tuberculosis*, *Mycobacterium tuberculosis*; *R. prowazekii*, *Rickettsia prowazekii*; *S. aureus*, *Staphylococcus aureus*; *S. pneumoniae*, *Streptococcus pneumoniae*; S. Typhi; *Salmonella enterica* subsp. *enterica* serovar Typhi; *T. maritima*, *Thermotoga maritima*; *T. whipplei*, *Tropheryma whipplei*; *Y. pestis*, *Yersinia pestis*.

showed this tidy, tree-like, monomorphic pattern of genome divergence, characterized by single-nucleotide polymorphisms (SNPs) and deletions; examples include *Y. pestis, Bacillus anthracis, Salmonella enterica* subsp. *enterica* serovar Typhi and *M. tuberculosis*[24]. Genome sequences from the intracellular parasite *Rickettsia prowazekii*[25] and from the unculturable leprosy bacillus *M. leprae*[10] provided a glimpse of the process of reductive genome evolution that occurs when sexually isolated lineages adapt to a restricted niche; this process is characterized by the creation of non-functional pseudogenes, followed by complete loss of sequences that are no longer needed for bacterial survival in the intracellular niche[26]. Similar genome erosion has been reported in a cyanobacterial endosymbiont of a fern[27].

Comparing the genomes of pathogens specifically adapted to a particular disease lifestyle with those of close relatives revealed a similar loss of genes that hinder within-host survival[28]. For example, studying the genome of the intracellular pathogen *Shigella* (a set of lineages that belong firmly within the species *E. coli*, but which have retained a separate genus designation to avoid confusion in clinical microbiology) revealed a loss of genes that are responsible for flagellar motility and for the production of the diamine cadaverine, factors that hinder virulence in the new intracellular niche[28]. Comparative genomics also revealed the presence of degenerate gene clusters even in the paradigmic *E. coli* K-12 genome, which has implications for gene annotation and for understanding evolution in this important model organism[29,30].

The view from the genomic high ground also provided new insights into key virulence strategies used by pathogenic bacteria, often overturning assumptions based on studies on a limited set of organisms. Notably, these studies elucidated the evolution and function of bacterial protein secretion systems and protein-targeting mechanisms, such as sortases, which enable the attachment of substrate proteins such as enzymes, pilins and adhesins to the bacterial cell surface. Examples include the discovery that Esx secretion, which mediates the secretion of important antigens in *M. tuberculosis*, was not limited to mycobacteria, but occurred in a wide range of bacteria, or that in the genomes of bacteria such as *Corynebacterium diphtheriae* or *Streptococcus pneumoniae* there were multiple sortase genes, each clustered with genes for sortase substrates[31,32]. Genomic mining of plant pathogens such as *Pseudomonas syringae* revealed many new type III secretion system effectors[33].

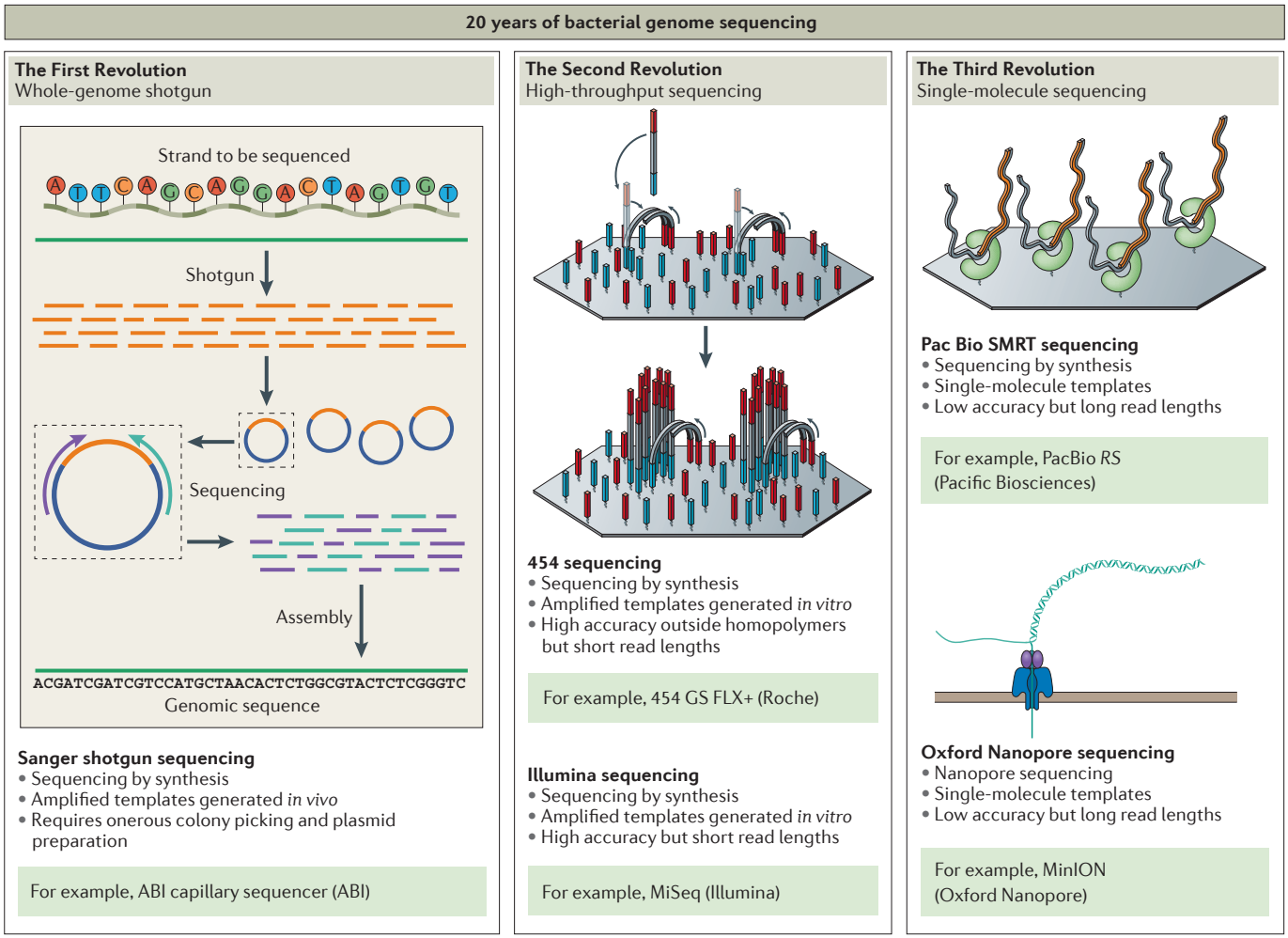

Figure 2 | **Bacterial genomics: the first two decades.** The three revolutions in sequencing technology that have transformed the landscape of bacterial genome sequencing are as follows: whole-genome shotgun sequencing, high-throughput sequencing, and single-molecule long-read sequencing. SMRT, single-molecule real-time.

In other bacteria, particularly naturally competent organisms, such as *Neisseria* spp. or the streptococci, it became clear that recombination blurred the evidence of evolutionary branching and even rendered tree-like thinking inappropriate in understanding genome evolution[34]. The pervasive role of HGT across much of the bacterial world led to a consideration of the key differences between core genomes (the genes present in all strains within a taxonomic group), accessory genomes (the genes present in a single strain, or in some but not all strains) and pan-genomes (the entire gene set, including the core and accessory genomes)[35,36]. Furthermore, this led to the recognition of important roles of genomic islands and mobile genetic elements (MGEs), particularly bacteriophages, in shaping genome evolution and pathogen biology through HGT and genetic rearrangements[37,38].

The discovery of what seemed to be virulence factors encoded in the genomes of non-pathogens led to a new 'eco–evo perspective', in which genomic analyses of pathogens and commensals were embedded in a rich ecological and evolutionary context that takes into account lifestyle shifts (for example, from commensal to pathogen or vice versa) and recognizes that many bacterial virulence factors have been shaped by evolutionary forces outside the context of human–pathogen interactions[39,40]. In this context, a pioneering comparison of isolates from a laboratory-acquired infection with *Burkholderia mallei* provided the first glimpse of short-term within-host genome evolution[41].

Exploiting genomics. Hard on the heels of the first sequencing projects, efforts focused on the information contained in the newfound genomes. In a pioneering approach termed 'reverse vaccinology', Rino Rappuoli and his colleagues sieved through the meningococcal genome for novel vaccine targets. This effort culminated in the recently licensed Bexsero vaccine against menB meningitis[42,43]. Sequencing was also applied to other problems. For instance, in the wake of the 2001 anthrax attacks in the United States, now known as the Amerithrax incident, in which *B. anthracis* was deliberately released into the US postal service, scientists at The Institute for Genomic Research (TIGR) genome-sequenced multiple strains of *B. anthracis*, and a subsequent tour-de-force genomic analysis of samples from the incident led to closure of the case by linking the profile of mutations within the released material to a flask in a government laboratory at Fort Detrick in Maryland[44,45].

Bacterial genome sequencing also spawned a range of high-throughput approaches that fall under the umbrella term

Box 1 | **Sequencing and bioinformatics technologies**

In the 1970s, British biochemist Fred Sanger invented a chain-termination approach to DNA sequencing that revolutionized biology[126], while Roger Staden showed how computer programs could be used to assemble sequences[127]. By the 1990s, steady improvements in sequencing technologies raised the possibility of bacterial whole-genome sequencing. However, the first bacterial genome-sequencing efforts used an onerous hierarchical top-down sequencing approach, in which it was necessary to create and map a library of large-insert clones, then create small-insert libraries from each of these clones, which would finally enable sequencing of these inserts. This top-down approach was largely side-lined by the arrival of bacterial whole-genome shotgun sequencing in 1995 (REF. 3). In this approach, the bacterial genome is broken up randomly into numerous small segments, which are sequenced *en masse* and then assembled into much larger sequences (contigs) using powerful computer programs (FIG. 1). Shotgun sequencing of bacterial genomes in turn spawned novel bioinformatics tools for assembly, gene calling and annotation such as Phred, Phrap, Glimmer and Artemis[128–130].

Sanger sequencing, combined with shotgun cloning in *Escherichia coli*, survived uncontested during the first decade of bacterial genome sequencing, despite the drawbacks of this approach: it remained onerous and expensive, and it could not be used on genes that were toxic to the cloning host, which therefore dropped out of sequencing libraries. The inability to clone such genes was overcome by the high-throughput sequencing revolution, which switched from biology to chemistry for template generation, while also delivering a massive increase in throughput[60]. However, this increase in throughput came at the expense of read length, which ranged from a few dozen to a few hundred base pairs per read. This meant that the new short-read technologies could not deliver finished bacterial genomes, because they were unable to generate accurate assembly across long repeats (derived from insertion sequences, prophages or ribosomal RNA clusters). They also failed to detect large-scale structural variation in genomes (for example, large chromosomal inversions, insertions or duplications).

The trade-offs inherent in this new reliance on short-read sequencing led to changes in emphasis: from whole-genome sequencing of new species to large-scale resequencing of closely related genomes from the same species; from finished genomes to draft genomes; and from *de novo* assembly to mapping against a reference genome. To some extent, this revolution changed what it meant to say one "had sequenced a genome", because genomes were now usually left incomplete, with most of the effort going into mapping differences between related genomes, rather than identifying and annotating new genes. This fuelled efforts in bioinformatics to devise new tools for analysing short-read data, such as Newbler, SOAPdenovo and Velvet[131].

The shortcomings of short-read sequencing provided the impetus for the third revolution: the arrival of long-read, single-molecule sequencing. Single-molecule real-time (SMRT) technology relies, like short-read sequencing, on a sequencing-by-synthesis approach[110]. However, the sequencing reactions occur in such small volumes that base calling becomes possible from unamplified, single-molecule targets, which facilitates long-read sequencing. Nanopore sequencing provides an innovative alternative to sequencing-by-synthesis, in which strands of DNA pass through the nanopore, and successive bases trigger changes in current that can be used to generate a sequence[132]. Both third-generation technologies have driven the development of bioinformatics tools aimed at getting the most out of long-read data, such as HGAP (hierarchical genome assembly process) and Nanopolish[112,133].

The steady accumulation of genomic and metagenomic data has brought the problems of 'big data' to bacteriology. Cloud computing provides an attractive solution, which provides easier access to data and facilitates sharing of tools and resources[134]. As a result, cloud computing has been adopted by sequencing companies (such as Illumina's Base Space), commercial providers (such as the Amazon Cloud) and academic consortia (such as the UK's Medical Research Council (MRC)-funded Cloud Infrastructure for Microbial Bioinformatics (CLIMB) project).

'functional genomics'. Early on, genome sequences were used to design microarrays that could be used to compare genome contents and interrogate patterns of global gene expression[46–48]. Similarly, the availability of complete genome sequences primed efforts in structural genomics and proteomics[49–51]. When combined with novel mutagenesis approaches that facilitated high-throughput screening of gene function[52], these functional genomics efforts delivered unparalleled insights into the biology of pathogens and model organisms, together with some unexpected spin-offs (BOX 3).

Sequencing the biosphere. From the start, bacterial genome sequencing made significant inroads into environmental microbiology, documenting the lifestyles of extremophiles while also providing insights into the diversity and evolution of life. For example, the genomes of *Shewanella oneidensis* and *Geobacter sulfurreducens* provided new insights into the process of metal ion reduction in the environment and opened up new possibilities for bioremediation[53,54], whereas genome sequencing revealed unsuspected diversity among marine aerobic anoxygenic phototrophs[55]. In 2004, Venter and colleagues applied shotgun metagenomic sequencing to the microbial contents of the Sargasso Sea[56], delivering over a million new predicted protein sequences into the databases. Similarly, sequencing the 1.2 Mb genome of mimivirus, which was originally isolated from amoebae growing in the water of a cooling tower of a hospital, blurred the distinction between bacteria and viruses[57]. Efforts such as the Genomic Encyclopedia of Bacteria and Archaea (GEBA) set about applying genome sequencing to as many organisms as possible[58].

The second revolution

High-throughput sequencing. High-throughput or next-generation sequencing reached bacteriology in the second half of the 2000s[59,60] (FIG. 2). This was clearly an idea whose time had come, as multiple platforms hit the marketplace in quick succession (reviewed in REF. 60), accompanied by new bioinformatics approaches. By 2012, a fresh round of innovation led to the emergence of benchtop sequencing platforms[61]. These laser-printer-sized instruments came with modest set-up and running costs, and turn-around times measured in days; this meant that, for the first time, bacterial genome sequencing could move out of sequencing centres and into universities and public health laboratories.

Translational clinical bacterial genomics. The arrival of high-throughput sequencing coincided with and energized the development of SNP-based phylogenetic analyses of bacterial pathogens. Pioneering efforts by Mark Achtman and colleagues at the Wellcome Trust Sanger Institute showed how these approaches could be used to capture global population genomics and the links between genomic diversity and geography for important pathogens such as *S.* Typhi[62]. Newsworthy applications of this approach included uncovering a politically charged link between a cholera outbreak in Haiti and Nepalese peacekeepers[63], and showing that humans transmitted leprosy to armadillos, who then transmitted it back to those who handled or ate these animals[64].

SNP-based analyses were also applied in a small-scale, high-impact fashion to the genomic epidemiology of outbreaks, with pioneering applications to the hospital

outbreaks of *S. aureus* and *Acinetobacter baumannii*[65–67]. These efforts were then followed by substantially extensive analyses of a range of pathogens that can be found in hospitals (for example, *S. aureus*, *Clostridium difficile* and carbapenemase-producing Enterobacteriaceae)[68–70] and in the community (for example, tuberculosis and drug-resistant gonorrhoea)[71,72]. The exquisite resolution of these approaches has allowed the reconstruction of transmission chains while also documenting within-patient pathogen diversity, showing that 'clonal' does not mean 'identical', that pathogens evolve within the host, and that multiple genotypes of a pathogen can coexist at a given site[73]. Whole-genome sequencing has also documented bacterial adaptation to therapeutic interventions in patients, such as the use of antibiotics and vaccines[74–76].

The increasing tractability of high-throughput sequencing has seen this approach move ever closer to routine clinical and public health microbiology. For example, rapid benchtop sequencing, open data release and social media catalysed the analysis of genomes during an outbreak of Shiga-toxin-producing *E. coli*[77]. Furthermore, in a landmark study, every significant bacterial pathogen isolated during a single day in a clinical microbiology laboratory was genome-sequenced, illustrating the feasibility and utility of this approach in clinical practice[78]. As a result of these technological advances, there is now considerable interest in determining how reliably one can deduce phenotype from genotype when considering resistance or virulence markers in bacteria[79,80].

Fresh applications, fresh challenges. High-throughput sequencing has found additional new applications in drug discovery and in functional genomics. For example, SNP-based comparisons between the genomes of a sensitive parent strain and a resistant daughter strain can be used to identify the targets of new drugs[81–83]. Notably, functional genomics has been re-energized, especially owing to the emergence of transposon sequencing (Tn-Seq), an approach discovered independently by four different research groups[82,84–86]. Tn-Seq exploits massively parallel screening of transposon libraries to identify genes and pathways that contribute to fitness in different environments. Additionally, various new approaches combine macromolecular crosslinking with high-throughput sequencing. These include: CHIP–Seq (chromatin immunoprecipitation followed by sequencing)[87], which is providing detailed global maps of the interactions between proteins and genomes; chromatin confirmation capture (3C), which is turning one-dimensional genome sequences into three-dimensional maps[88]; and Hi-C, an approach for elucidating the cellular colocalization of DNA sequences that holds great promise in metagenomics[89]. The genome-wide association study (GWAS) approach commonly used in human genetics has also been applied to bacterial genomics and shown early success[90], although it is unclear at present how widely it will be used.

In parallel with the relentless rise of the microbiome across the scientific agenda and in the public eye[91], high-throughput sequencing has been harnessed for culture-independent approaches to microbial ecology and even for diagnosis. David Relman and colleagues showed how molecular bar-coding approaches could be combined with high-throughput sequencing to achieve unprecedented depths of coverage in microbial community profiling[92]. Similarly, shotgun metagenomics took on a new lease of life, with the first in-depth studies of the gut microbiomes of humans and other animals[93,94]. In recent years, there has been a growing interest in using metagenomics to deliver a new culture-independent paradigm in diagnostic microbiology, as demonstrated by the recent proof-of-principle studies on an outbreak of Shiga-toxin-producing *E. coli* O104:H4, a suspected outbreak of severe pneumonia and a case of neuroleptospirosis[95–97].

High-throughput sequencing has been applied to other areas, from the study of ancient pathogens to the analysis of single cells. For example, ancient DNA research has delivered pathogen genomes from the past, including genomes from Black Death, from a medieval *Brucella* strain and from eighteenth-century tuberculosis[98–100]. In addition, multiple displacement amplification, which is an isothermal amplification approach that relies on random hexamers and a high-fidelity polymerase for whole-genome amplification, has delivered bacterial genome sequences from low-biomass samples, including single cells[101]. This approach has provided reference genomes for numerous candidate phyla, known

Box 2 | The sociology of sequencing: from sequencing centre to benchtop

Two large sequencing centres were established in 1992, along with associated sequencing programmes, one on each side of the Atlantic: Craig Venter set up The Institute for Genomic Research (TIGR) in Rockville, Maryland, while the Wellcome Trust established the Sanger Centre near Cambridge, in England. The primary focus of the two centres was the human genome, but both cranked out many bacterial genome sequences over the years that followed. They were subsequently joined by several other major sequencing centres, including the French National Sequencing Centre, Genoscope, in Évry, near Paris; the US Department of Energy's Joint Genome Institute (JGI), in Walnut Creek, California; the Whitehead Institute, and then the Broad Institute, both located in Cambridge, Massachusetts; and the Human Genome Sequencing Center situated at Baylor College of Medicine in Houston, Texas.

In the early years, there was some friendly rivalry between centres. TIGR beat Sanger to the first bacterial genome sequence[3], but the Sanger beat TIGR to the first *Mycobacterium tuberculosis* genome[5]. The Sanger beat Genoscope to the first genome of *Tropheryma whipplei*[11], but French investigators skilfully used the genome sequence to design a new culture medium[13]. In the first decade of bacterial genome sequencing, publications regularly appeared in high-impact journals, and sequencing centres engaged widely with the scientific community — both through individual projects and through microbial genome meetings, at which the latest exciting breakthroughs were announced to a riveted audience[135].

Relationships between research communities and sequencing projects were complex. For each community of researchers focused on a given microorganism, the arrival of a genome sequence re-drew the research landscape and forced them to adapt, sometimes reluctantly, to new post-genomic circumstances and opportunities. Many were grateful for what the sequencing centres delivered, but some felt frustration at their inability to control the tempo and agenda of the genome-sequencing projects, which sometimes dragged on for years.

The UK's first bacterial genome-sequencing project outside of a major sequencing centre was completed in 2007 (REF. 136). Within a couple of years, the disruptive effect of high-throughput sequencing, particularly benchtop sequencing, brought bacterial genome sequencing into the average university set-up[137] while also driving population-biology projects, with thousands of bacterial genomes per project becoming routine[74,138].

As whole-genome sequencing becomes the default approach for a range of research and clinical applications, it remains unclear how far institutions should try to centralize, de-centralize or outsource sequencing capacity. No longer a grand voyage of discovery, but an undemanding technical exercise, bacterial genome sequencing now often falls to Ph.D. or even project students. How times have changed!

Box 3 | **Spin-offs from bacterial genomics**

As cogent proof that translational research cannot be scripted, bacterial genomics has delivered several unplanned but important spin-offs. One early example was the unexpected discovery of novel glycosylation systems encoded in the *Campylobacter jejuni* genome[14]. This has led to a vibrant programme of glycoengineering, which promises to deliver new highly immunogenic glycoconjugate vaccines[139,140].

Comparative genome analyses, by Eugene Koonin and his collaborators, led to the recognition that the CRISPRs (clustered regularly interspaced short palindrome repeats) and variable arrays of the CRISPR-associated (Cas) genes seen in many bacterial genomes represented a prokaryotic immune system that targeted specific sequences from bacteriophages[141]. This primed the exploitation of these systems by Jennifer Doudna and others via genome engineering of humans, plants and animals — an advance that arguably counts as one of the greatest scientific breakthroughs of this millennium[142,143].

The availability of large genomic and metagenomic data sets has fuelled bioprospecting and provided many of the components of the toolkits used by synthetic biology[144]. It has also underpinned chemical synthesis of the genome of a free-living organism[145].

previously only from molecular barcodes, thereby filling in gaps in the genomic tree of life[102]. In 2007, the first genomes were obtained for the evasive TM7 phylum from single cells taken from the human mouth and from soil[103,104]. Five years later, a genome from another candidate phylum, TM6, was recovered from a hospital sink drain using a highly automated single-cell genomics platform[105]. Improvements in laboratory and bioinformatics pipelines for metagenomics mean that this approach can also provide assembled genome sequences from uncultured organisms. A recent example includes the recovery of multiple genomes from a 'candidate phyla radiation' super-phylum, which represents novel and unusual biology across a large part of the bacterial domain[106].

The advances in high-throughput sequencing have also had an impact on bacterial taxonomy. Genome and metagenome sequencing have become easier to use, which has increased their practical utility compared to bacterial taxonomy — a discipline that remains conservative and still insists on using mid-twentieth century approaches[107]. A number of studies have compared traditional and genome-sequence-based taxonomies[108,109], and now that most bacterial taxa are known only from culture-independent approaches, it is perhaps time to reconsider use of traditional approaches, particularly as genome sequences provide reliable, reproducible digital taxonomic data.

The third revolution

Single-molecule, long-read sequencing. The first long-read technology to achieve widespread use was single-molecule real-time (SMRT) sequencing from Pacific Biosciences[110] (FIG. 2). Recent publications have shown that this approach, on its own or combined with short-read sequencing, can deliver high-quality assemblies[111,112]. This, in turn, is taking us back to the era of complete, reference-quality genome sequences. This is exemplified by an ongoing collaboration between the Wellcome Trust Sanger Institute and Public Health England to use SMRT sequencing to deliver reference genomes for 3,000 bacterial strains from the UK's National Collection of Type Cultures[113], which will not only add value to this well-curated collection but also deliver new insights into genomic and metabolic diversity. SMRT sequencing has also proven useful in unravelling plasmid diversity in multidrug-resistant hospital pathogens, such as *Enterobacter cloacae*[114], and has the potential to go beyond four-base sequencing to reveal genome-wide patterns of methylation and other chemical modifications that control the biology of bacteria or the virulence of pathogens[115].

Despite its benefits, with a US$700,000 price tag and large instrument size, SMRT sequencing is largely restricted to major sequencing centres, although a cheaper instrument is promised in 2016. An alternative approach, nanopore sequencing, promises single-molecule long-read sequencing for the masses, with Oxford Nanopore's MinION instrument under evaluation by numerous eager early-adopters. Similar to SMRT sequencing, nanopore sequencing can generate reads that are long enough to span large-scale repeats, and early proof-of-principle studies suggest that it can deliver genome-scale assemblies for bacteria[116–118]. Furthermore, nanopore sequencing has already been applied to the analysis of a *Salmonella* outbreak and to the detection of resistance genes in Gram-negative isolates and in *S. aureus*[119,120]. However, unlike the SMRT platform, the MinION is small and portable, which enables near-patient or in-the-field sequencing, as evidenced by its use during the 2014–2015 Ebola outbreak in West Africa[121]. As with SMRT sequencing, nanopore sequencing is far less accurate than established short-read technologies, although recent improvements have been documented[122].

Future prospects. So, what can we expect from the third decade of bacterial genome sequencing? Despite suggestions to the contrary[123], we expect the gold rush to continue and to see the $1,000 human genome matched by the $1 bacterial genome. Perhaps rather fancifully, we have predicted a 'sequencing singularity', whereby sequencing becomes the method of choice for as-yet unthinkable applications. The recent report of encoding and then sequencing Shakespeare's sonnets in a DNA format illustrates the point[124]. But who knows what will happen when it becomes as easy to sequence a bacterial genome as it is to perform a pregnancy test?

Clearly, whole-genome sequencing will soon overtake phenotypic methods for the identification and characterization of bacterial isolates, whether in clinical practice or in taxonomy. The arrival of accurate, high-throughput long-read sequencing will transform genomic epidemiology, forcing us to think beyond single colonies and SNPs. How far metagenomics can replace culture methods as a diagnostic approach remains to be seen, but for some samples (such as faeces or urine) one could imagine bacterial metagenomics integrated into a microfluidics-driven nanopore-based comprehensive macromolecular monitoring approach that could capture sequences from pathogen and host, DNA, RNA and proteins, to assay and investigate infection, inflammation and neoplasia all in one workflow[125].

Liberated from the laboratory by field-compatible sequencing devices, environmental microbiologists will steadily sequence more of the microbial biosphere. No one can know what 'unknown unknowns' await us in terms of microbial diversity, but, as with mimivirus, we may yet again have to rewrite the textbooks. For all microbiologists — clinical or environmental, basic or applied — a brave new world awaits us.

Nicholas J. Loman is at the Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, UK.

Mark J. Pallen is in the Microbiology and Infection Unit, Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK.

Correspondence to M.J.P.
e-mail: m.pallen@warwick.ac.uk

doi:10.1038/nrmicro3565
Published online 9 November 2015

1. Burland, V., Plunkett, G., Daniels, D. L. & Blattner, F. R. DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: organizational symmetry around the origin of replication. *Genomics* **16**, 551–561 (1993).
2. Glaser, P. *et al. Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325 degrees to 333 degrees. *Mol. Microbiol.* **10**, 371–384 (1993).
3. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
4. Parkhill, J. In defense of complete genomes. *Nat. Biotechnol.* **18**, 493–494 (2000).
5. Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
6. Parkhill, J. *et al.* Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**, 523–527 (2001).
7. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
8. Kunst, F. *et al.* The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256 (1997).
9. Fraser, C. M. *et al.* Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375–388 (1998).
10. Eiglmeier, K. *et al.* The decaying genome of *Mycobacterium leprae*. *Lepr Rev.* **72**, 387–398 (2001).
11. Bentley, S. D. *et al.* Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*. *Lancet* **361**, 637–644 (2003).
12. White, O. *et al.* Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286**, 1571–1577 (1999).
13. Renesto, P. *et al.* Genome-based design of a cell-free culture medium for *Tropheryma whipplei*. *Lancet* **362**, 447–449 (2003).
14. Parkhill, J. *et al.* The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–668 (2000).
15. Patrick, S. *et al.* Multiple inverted DNA repeats of *Bacteroides fragilis* that control polysaccharide antigenic variation are similar to the hin region inverted repeats of *Salmonella* typhimurium. *Microbiology* **149**, 915–924 (2003).
16. Himmelreich, R. *et al.* Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**, 4420–4449 (1996).
17. Alm, R. A. *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176–180 (1999).
18. Read, T. D. *et al.* Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* **28**, 1397–1406 (2000).
19. Fleischmann, R. D. *et al.* Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.* **184**, 5479–5490 (2002).
20. Liu, S. L., Hessel, A. & Sanderson, K. E. Genomic mapping with I-Ceu I, an intron-encoded endonuclease specific for genes for ribosomal RNA, in *Salmonella* spp., *Escherichia coli*, and other bacteria. *Proc. Natl Acad. Sci. USA* **90**, 6874–6878 (1993).
21. 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).
22. Welch, R. A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **99**, 17020–17024 (2002).
23. Nelson, K. E. *et al.* Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329 (1999).
24. Achtman, M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu. Rev. Microbiol.* **62**, 53–70 (2008).
25. Andersson, S. G. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140 (1998).
26. Darby, A. C., Cho, N. H., Fuxelius, H. H., Westberg, J. & Andersson, S. G. Intracellular pathogens go extreme: genome evolution in the Rickettsiales. *Trends Genet.* **23**, 511–520 (2007).
27. Vigil-Stenman, T., Larsson, J., Nylander, J. A. & Bergman, B. Local hopping mobile DNA implicated in pseudogene formation and reductive evolution in an obligate cyanobacteria-plant symbiosis. *BMC Genomics* **16**, 193 (2015).
28. Maurelli, A. T. Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. *FEMS Microbiol. Lett.* **267**, 1–8 (2007).
29. 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).
30. 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).
31. Pallen, M. J., Lam, A. C., Antonio, M. & Dunbar, K. An embarrassment of sortases — a richness of substrates? *Trends Microbiol.* **9**, 97–102 (2001).
32. Pallen, M. J. The ESAT-6/WXG100 superfamily — and a new Gram-positive secretion system? *Trends Microbiol.* **10**, 209–212 (2002).
33. Collmer, A., Lindeberg, M., Petnicki-Ocwieja, T., Schneider, D. J. & Alfano, J. R. Genomic mining type III secretion system effectors in *Pseudomonas syringae* yields new picks for all TTSS prospectors. *Trends Microbiol.* **10**, 462–469 (2002).
34. Spratt, B. G. & Maiden, M. C. Bacterial population genetics, evolution and epidemiology. *Phil. Trans. R. Soc. Lond. B* **354**, 701–710 (1999).
35. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594 (2005).
36. 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).
37. 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).
38. 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).
39. Holden, M., Crossman, L., Cerdeno-Tarraga, A. & Parkhill, J. Pathogenomics of non-pathogens. *Nat. Rev. Microbiol.* **2**, 91 (2004).
40. Pallen, M. J. & Wren, B. W. Bacterial pathogenomics. *Nature* **449**, 835–842 (2007).
41. 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).
42. Pizza, M. *et al.* Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287**, 1816–1820 (2000).
43. Vernikos, G. & Medini, D. Bexsero® chronicle. *Pathog. Glob. Health* **108**, 305–316 (2014).
44. Rasko, D. A. *et al. Bacillus anthracis* comparative genome analysis in support of the Amerithrax investigation. *Proc. Natl Acad. Sci. USA* **108**, 5027–5032 (2011).
45. Read, T. D. *et al.* Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* **296**, 2028–2033 (2002).
46. Cummings, C. A. & Relman, D. A. Using DNA microarrays to study host–microbe interactions. *Emerg. Infect. Dis.* **6**, 513–525 (2000).
47. Harrington, C. A., Rosenow, C. & Retief, J. Monitoring gene expression using DNA microarrays. *Curr. Opin. Microbiol.* **3**, 285–291 (2000).
48. Dorrell, N. *et al.* Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res.* **11**, 1706–1715 (2001).
49. Schmid, M. B. Structural proteomics: the potential of high-throughput structure determination. *Trends Microbiol.* **10**, S27–S31 (2002).
50. Matte, A., Jia, Z., Sunita, S., Sivaraman, J. & Cygler, M. Insights into the biology of *Escherichia coli* through structural proteomics. *J. Struct. Funct. Genomics* **8**, 45–55 (2007).
51. Lipton, M. S. *et al.* Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc. Natl Acad. Sci. USA* **99**, 11049–11054 (2002).
52. Brown, J. S. *et al.* Signature-tagged and directed mutagenesis identify PABA synthetase as essential for *Aspergillus fumigatus* pathogenicity. *Mol. Microbiol.* **36**, 1371–1380 (2000).
53. Heidelberg, J. F. *et al.* Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nat. Biotechnol.* **20**, 1118–1123 (2002).
54. Methe, B. A. *et al.* Genome of *Geobacter sulfurreducens*: metal reduction in subsurface environments. *Science* **302**, 1967–1969 (2003).
55. Beja, O. *et al.* Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415**, 630–633 (2002).
56. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
57. Raoult, D. *et al.* The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).
58. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056–1060 (2009).
59. Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res.* **15**, 1767–1776 (2005).
60. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
61. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439 (2012).
62. Holt, K. E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat. Genet.* **40**, 987–993 (2008).
63. Chin, C. S. *et al.* The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364**, 33–42 (2011).
64. Truman, R. W. *et al.* Probable zoonotic leprosy in the southern United States. *N. Engl. J. Med.* **364**, 1626–1633 (2011).
65. Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010).
66. Azarian, T. *et al.* Whole-genome sequencing for outbreak investigations of methicillin-resistant *Staphylococcus aureus* in the neonatal intensive care unit: time for routine practice? *Infect. Control Hosp. Epidemiol.* **36**, 777–785 (2015).
67. Lewis, T. *et al.* High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. *J. Hosp. Infect.* **75**, 37–41 (2010).
68. Koser, C. U. *et al.* Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N. Engl. J. Med.* **366**, 2267–2275 (2012).
69. Eyre, D. W. *et al.* Diverse sources of *C. difficile* infection identified on whole-genome sequencing. *N. Engl. J. Med.* **369**, 1195–1205 (2013).
70. Snitkin, E. S. *et al.* Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci. Transl Med.* **4**, 148ra116 (2012).
71. Gardy, J. L. *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* **364**, 730–739 (2011).
72. Grad, Y. H. *et al.* Genomic epidemiology of *Neisseria gonorrhoeae* with reduced susceptibility to cefixime in the USA: a retrospective observational study. *Lancet Infect. Dis.* **14**, 220–226 (2014).
73. Paterson, G. K. *et al.* Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. *Nat. Commun.* **6**, 6560 (2015).
74. Chewapreecha, C. *et al.* Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46**, 305–309 (2014).
75. Loman, N. J. *et al.* Clonal expansion within pneumococcal serotype 6C after use of seven-valent vaccine. *PLoS ONE* **8**, e64731 (2013).
76. Hornsey, M. *et al.* Whole-genome comparison of two *Acinetobacter baumannii* isolates from a single patient, where resistance developed during tigecycline therapy. *J. Antimicrob. Chemother.* **66**, 1499–1503 (2011).
77. Rohde, H. *et al.* Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N. Engl. J. Med.* **365**, 718–724 (2011).
78. Long, S. W. *et al.* A genomic day in the life of a clinical microbiology laboratory. *J. Clin. Microbiol.* **51**, 1272–1277 (2013).
79. Parkhill, J. What has high-throughput sequencing ever done for us? *Nat. Rev. Microbiol.* **11**, 664–665 (2013).
80. Gordon, N. C. *et al.* Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. *J. Clin. Microbiol.* **52**, 1182–1191 (2014).
81. Andries, K. *et al.* A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science* **307**, 223–227 (2005).

82. Abrahams, K. A. *et al.* Identification of novel imidazo[1,2-a]pyridine inhibitors targeting *M. tuberculosis* QcrB. *PLoS ONE* **7**, e52951 (2012).
83. Remuinan, M. J. *et al.* Tetrahydropyrazolo[1,5-a] pyrimidine-3-carboxamide and *N*-benzyl-6′,7′-dihydros piro[piperidine-4,4′-thieno[3,2-c]pyran] analogues with bactericidal efficacy against *Mycobacterium tuberculosis* targeting MmpL3. *PLoS ONE* **8**, e60933 (2013).
84. Goodman, A. L. *et al.* Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6**, 279–289 (2009).
85. van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods* **6**, 767–772 (2009).
86. Langridge, G. C. *et al.* Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. *Genome Res.* **19**, 2308–2316 (2009).
87. Galagan, J., Lyubetskaya, A. & Gomes, A. ChIP-Seq and the complexity of bacterial transcriptional regulation. *Curr. Top. Microbiol. Immunol.* **363**, 43–68 (2013).
88. Umbarger, M. A. *et al.* The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol. Cell* **44**, 252–264 (2011).
89. Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).
90. Sheppard, S. K. *et al.* Genome-wide association study identifies vitamin B_5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc. Natl Acad. Sci. USA* **110**, 11923–11927 (2013).
91. Hanage, W. P. Microbiology: microbiome science needs a healthy dose of scepticism. *Nature* **512**, 247–248 (2014).
92. Dethlefsen, L., Huse, S., Sogin, M. L. & Relman, D. A. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.* **6**, e280 (2008).
93. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
94. Sergeant, M. J. *et al.* Extensive microbial and functional diversity within the chicken cecal microbiome. *PLoS ONE* **9**, e91941 (2014).
95. Loman, N. J. *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA* **309**, 1502–1510 (2013).
96. Fischer, N. *et al.* Rapid metagenomic diagnostics for suspected outbreak of severe pneumonia. *Emerg. Infect. Dis.* **20**, 1072–1075 (2014).
97. Wilson, M. R. *et al.* Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N. Engl. J. Med.* **370**, 2408–2417 (2014).
98. Bos, K. I. *et al.* A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* **478**, 506–510 (2011).
99. Kay, G. L. *et al.* Recovery of a medieval *Brucella melitensis* genome using shotgun metagenomics. *MBio* **5**, e01337–e01314 (2014).
100. Kay, G. L. *et al.* Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat. Commun.* **6**, 6717 (2015).
101. Walker, A. & Parkhill, J. Single-cell genomics. *Nat. Rev. Microbiol.* **6**, 176–177 (2008).
102. Lasken, R. S. & McLean, J. S. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat. Rev. Genet.* **15**, 577–584 (2014).
103. Marcy, Y. *et al.* Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl Acad. Sci. USA* **104**, 11889–11894 (2007).
104. Podar, M. *et al.* Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl. Environ. Microbiol.* **73**, 3205–3214 (2007).
105. McLean, J. S. *et al.* Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc. Natl Acad. Sci. USA* **110**, E2390–E2399 (2013).
106. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
107. Ramasamy, D. *et al.* A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int. J. Syst. Evol. Microbiol.* **64**, 384–391 (2014).
108. Chan, J. Z., Halachev, M. R., Loman, N. J., Constantinidou, C. & Pallen, M. J. Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. *BMC Microbiol.* **12**, 302 (2012).
109. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl Acad. Sci. USA* **102**, 2567–2572 (2005).
110. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
111. Bashir, A. *et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nat. Biotechnol.* **30**, 701–707 (2012).
112. Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
113. Wellcome Trust Sanger Institute. *Public Health England reference collections. Wellcome Trust Sanger Institute* [online], https://www.sanger.ac.uk/resources/downloads/bacteria/nctc/ (2015).
114. Stoesser, N. *et al.* Dynamics of MDR *Enterobacter cloacae* outbreaks in a neonatal unit in Nepal: insights using wider sampling frames and next-generation sequencing. *J. Antimicrob. Chemother.* **70**, 1008–1015 (2015).
115. Korlach, J. & Turner, S. W. Going beyond five bases in DNA sequencing. *Curr. Opin. Struct. Biol.* **22**, 251–261 (2012).
116. Madoui, M. A. *et al.* Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* **16**, 327 (2015).
117. Karlsson, E., Larkeryd, A., Sjodin, A., Forsman, M. & Stenberg, P. Scaffolding of a bacterial genome using MinION nanopore sequencing. *Sci. Rep.* **5**, 11996 (2015).
118. Quick, J., Quinlan, A. R. & Loman, N. J. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *Gigascience* **3**, 22 (2014).
119. Judge, K., Harris, S. R., Reuter, S., Parkhill, J. & Peacock, S. J. Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. *J. Antimicrob. Chemother.* **70**, 2775–2778 (2015).
120. Quick, J. *et al.* Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol.* **16**, 114 (2015).
121. Check Hayden, E. Pint-sized DNA sequencer impresses first users. *Nature* **521**, 15–16 (2015).
122. Loman, N. J. & Watson, M. Successful test launch for nanopore sequencing. *Nat. Methods* **12**, 303–304 (2015).
123. Hall, N. After the gold rush. *Genome Biol.* **14**, 115 (2013).
124. Goldman, N. *et al.* Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80 (2013).
125. Lo, Y. M. & Chiu, R. W. Plasma nucleic acid analysis by massively parallel sequencing: pathological insights and diagnostic implications. *J. Pathol.* **225**, 318–323 (2011).
126. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
127. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6**, 2601–2610 (1979).
128. Rieder, M. J., Taylor, S. L., Tobe, V. O. & Nickerson, D. A. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res.* **26**, 967–973 (1998).
129. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).
130. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945 (2000).
131. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315–327 (2010).
132. Eisenstein, M. Oxford Nanopore announcement sets sequencing sector abuzz. *Nat. Biotechnol.* **30**, 295–296 (2012).
133. Loman, N. J., Quick, J & Simpson, J. T. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
134. Drake, N. How to catch a cloud. *Nature* **522**, 115–116 (2015).
135. Pallen, M. J. Microbial genomes. *Mol. Microbiol.* **32**, 907–912 (1999).
136. 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).
137. Loman, N. J. *et al.* High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.* **10**, 599–606 (2012).
138. Nasser, W. *et al.* Evolutionary pathway to increased virulence and epidemic group A *Streptococcus* disease derived from 3,615 genome sequences. *Proc. Natl Acad. Sci. USA* **111**, E1768–E1776 (2014).
139. Wacker, M. *et al.* N-linked glycosylation in *Campylobacter jejuni* and its functional transfer into *E. coli*. *Science* **298**, 1790–1793 (2002).
140. Cuccui, J. *et al.* Exploitation of bacterial N-linked glycosylation to develop a novel recombinant glycoconjugate vaccine against *Francisella tularensis*. *Open Biol.* **3**, 130002 (2013).
141. Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* **1**, 7 (2006).
142. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
143. Pennisi, E. The CRISPR craze. *Science* **341**, 833–836 (2013).
144. Cameron, D. E., Bashor, C. J. & Collins, J. J. A brief history of synthetic biology. *Nat. Rev. Microbiol.* **12**, 381–390 (2014).
145. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).
146. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).
147. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
148. Kuroda, M. *et al.* Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet* **357**, 1225–1240 (2001).
149. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
150. Baker, S. *et al.* High-throughput genotyping of *Salmonella enterica* serovar Typhi allowing geographical assignment of haplotypes and pathotypes within an urban District of Jakarta, Indonesia. *J. Clin. Microbiol.* **46**, 1741–1746 (2008).

Acknowledgements
N.J.L. and M.J.P. are supported by the Medical Research Council (MRC)-funded Cloud Infrastructure for Microbial Bioinformatics (CLIMB) project (reference number MR/L015080/1).

Competing interests statement
The authors declare competing interests: see Web version for details.

FURTHER INFORMATION
Genomic Encyclopedia of Bacteria and Archaea (GEBA): http://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF