Microbiology in the post-genomic era

Duccio Medini¶, Davide Serruto*¶, Julian Parkhill‡, David A. Relman§, Claudio Donati*, Richard Moxon||, Stanley Falkow§ and Rino Rappuoli**

Abstract | Genomics has revolutionized every aspect of microbiology. Now, 13 years after the first bacterial genome was sequenced, it is important to pause and consider what has changed in microbiology research as a consequence of genomics. In this article, we review the evolving field of bacterial typing and the genomic technologies that enable comparative analysis of multiple genomes and the metagenomes of complex microbial environments, and address the implications of the genomic era for the future of microbiology.

Genome
The entire hereditary information of an organism that is encoded by its DNA (or RNA for some viruses).

Bacterial typing
A procedure for identifying types and strains of bacteria.

Metagenome
The global genetic repertoire of an environmental niche that is constituted by diverse organisms such as free-living microorganisms in the wild or the commensals of a particular niche in a mammalian host (from the Greek '*meta*', meaning beyond, and genome).

**Novartis Vaccines and Diagnostics, 53100 Siena, Italy.*
‡*The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.*
§*Department of Microbiology and Immunology, Stanford University, Stanford, California 94305, USA.*
||*University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK.*
¶*These authors contributed equally to this work.*
Correspondence to R.R.
e-mail: rino.rappuoli@novartis.com
doi:10.1038/nrmicro1901
Published online 13 May 2008

The amount of novel microbial genomic information that is being generated on a daily basis is so vast that multidisciplinary approaches that integrate bioinformatics, statistics and mathematical methods are required to assess it effectively. This information is inspiring a new understanding of microorganisms that appreciates the scale of microbial diversity and acknowledges that the microbial gene pool is considerably larger than expected. Indeed, the availability of only one complete genome sequence for a given taxon, which was a dream only two decades ago, is now considered inadequate for describing the complexity of species and genera and their interrelationships. Advances in genomics are also beginning to drive the discovery of novel diagnostics, drug targets and vaccines. In this article, we review various aspects of the impact of genomics on microbiology, including in the evolving field of bacterial typing, and the genomic technologies that enable the comparative analysis of multiple genomes or the metagenomes of complex microbial environments. We also address the implications of the genomic era for the future of microbiology.

Classification and the challenges of genomics

Historically, bacteria have been classified using convenient traits such as cell structure, cellular metabolism or differences in cellular components. Diversity within a particular species could also be addressed using opportune markers, such as capsular and protein serotypes or the ability to agglutinate or lyse red blood cells. In the 1970s, DNA–DNA hybridization was introduced to differentiate bacterial species. Isolates that showed >70% DNA–DNA homology under standard hybridization conditions were considered to belong to the same species. Later, advances in sequencing techniques allowed the introduction of other markers, such as 16S ribosomal RNA (rRNA), a molecule that is ubiquitous in bacterial and archaeal genomes. 16S rRNA sequence similarities in bacteria and archaea were found to be highly correlated with DNA hybridization — roughly, the 70% cut-off level in DNA–DNA reassociation corresponds to 98% 16S rRNA sequence identity[1,2]. High-throughput sequencing of rRNA molecules from environmental samples showed that most of these sequences fall into clusters of 99% identity[3,4], which suggests a basis for a coherent working definition of a bacterial species and indicates that 16S rRNA sequence conservation of approximately 99% marks the borders of naturally occurring phenotypic clusters or species. This definition holds for most cases even today, with only a few exceptions that include the *Bacillus* genus, for which the 16S rRNA sequences from phenotypically distinct species, such as *Bacillus cereus*, *Bacillus thuringiensis* and *Bacillus anthracis*, differ at only a few bases[5]. Another limitation of 16S rRNA typing is that organisms with polymorphisms in the regions that are used for primer design can be poorly detected with 'universal' primers or completely lost, as in the Nanoarchaeota phylum[6].

Additionally, 16S rRNA analysis is a poor method for resolving sub-populations within species. For this purpose, multilocus enzyme electrophoresis (MLEE)[7], a method that classifies bacteria on the basis of the isoforms of a combination of ~15 metabolic enzymes[8], became the method of choice for many epidemiological studies. MLEE was not widely used, however, because it is low throughput and intensive laboratory work is required. A natural evolution of MLEE that came with the genomic era was multilocus sequence typing (MLST), a typing method that is based on the partial sequences of 7 housekeeping genes of ~450 bp each[9]. MLST is high throughput, allows direct comparison of organisms that are being studied in different laboratories in different parts of the world and has led to a rapidly enlarging database (MLST Public Repository; see Further information), in which almost 30 species are represented.

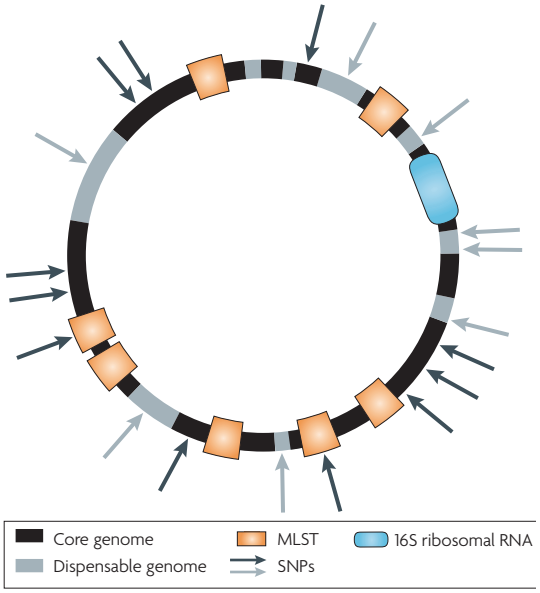

Figure 1 | **Genomic coverage of genetic typing methods.** Shows the core genome, which includes genes that encode proteins which are involved in essential functions, such as replication, transcription and translation, and the dispensable genome, which includes genes that encode proteins that facilitate organismal adaptation. Coverage by 16S ribosomal RNA (rRNA), multilocus sequence typing (MLST) and single-nucleotide polymorphisms (SNPs) is also depicted. For *Neisseria meningitidis*, which has a 2.2 Mb genome[91], the average length of a 16S rRNA gene is ~1.5 Kb[92] and the average length of the MLST loci is ~4 Kb[9]; typing based on 16S rRNA and MLST therefore covers 0.07% and ~0.2% of the *N. meningitidis* genome, respectively. The genome of *Salmonella enterica* serovar Typhi (*S. typhi*) is ~4.8 Mb[93] and SNP gene fragments are present in ~89 Kb[23]; SNP-based typing therefore provides coverage of 2% of the *S. typhi* genome.

One of the limitations of MLST is that for some species there is too little sequence variation in the housekeeping genes that are analysed to provide sufficient discrimination, whereas in other species (the so-called monomorphic species, such as *B. anthracis* and *Salmonella enterica* serovar Typhi (*S. typhi*)), the housekeeping genes are so uniform that all isolates appear to be the same. The other, more general limitation of both 16S rRNA typing and MLST is connected to the limited genome coverage of these methods (FIG. 1).

It is now apparent that bacterial genomes comprise core sequences — including genes that encode proteins which are involved in essential functions, such as replication, transcription and translation, the evolution of which probably correlates with neutral markers — and dispensable sequences that encode proteins which facilitate organismal adaptation. All bacteria face changes in their environment, especially commensal and pathogenic bacteria, which encounter extensive and dynamic variations in their co-evolving hosts. Dispensable sequences are characterized by a variable pattern of presence or absence in different bacterial isolates. They are also often associated with high rates of nucleotide sequence variability and contribute to phenotypic diversity within bacterial populations. Of the variants that are generated, the fittest are retained by natural selection. Examples of dispensable sequences include the hypermutable contingency loci, which have the capacity to generate phase variation and thus provide a powerful combinatorial mechanism of adaptation[10], and pathogenicity islands, which confer fitness advantages that are often associated with pathogenicity and resistance to antimicrobials[11,12]. Many of these sequences include genes which encode proteins that are themselves targets for drugs or antibodies. Understanding the population-wide variation in dispensable sequences in natural populations of bacterial pathogens is therefore of substantial public health importance.

As MLST is based solely on housekeeping genes, however, those population structures that evolve under the pressure of non-neutral evolutionary forces might be missed. In fact, in many pathogens, the evolution of virulence-associated genes, which is mainly driven by interactions with the host and the host immune system, is not directly linked to housekeeping functions and therefore might not correlate with MLST[13] (FIG. 2). The extent of the diversity within bacterial species was emphasized by a comparison of the genomes of seven isolates of *Streptococcus agalactiae* (group B *Streptococcus*; GBS), which suggested that the genome of a bacterial species — the pan-genome — could be many times larger than the genome of a single bacterium. This analysis also showed that in this group the conventional and convenient markers that are widely used to classify pathogenic bacteria, such as those used to classify serotypes, do not correlate with the genomic content of the bacteria (as originally indicated by MLEE studies on *Escherichia coli*[14]), but instead correlate only with the presence or absence of a single gene or gene island that is actively exchanged between bacteria[15].

This new information revealed the shortcomings of many of the commonly used classification criteria and raised the question of whether there is a single method or combination of methods that can take into account the similarities and differences both between and within species. One powerful approach that is increasingly being used is the investigation of single-nucleotide polymorphisms (SNPs). Originally developed for use in humans, and then applied to bacteria for the analysis of single genes[16–18], SNPs have recently been used to differentiate *B. anthracis* clinical samples that were collected from a disease outbreak[19], to resolve the population structure of *Mycobacterium tuberculosis*[20,21] and to propose an *M. tuberculosis* typing scheme[22]. More recently, the complex evolutionary history of *S. typhi* was reconstructed by analysing 88 biallelic polymorphisms, including 82 SNPs[23]. This history could be explained by the superimposition of neutral evolution that is associated with an asymptomatic carrier state in the human host and the adaptive evolution that is driven by the rapid transmission of phenotypic changes during acute infection. However, although SNPs can be extremely powerful owing to their provision of greater genomic coverage compared with other classification methods (FIG. 1), their use is still limited and their potential for more general use in bacterial population genetics is still unproven.

16S ribosomal RNA
The 16S ribosomal RNA gene is a component of the small bacterial and archaeal ribosomal subunit. The gene includes hypervariable regions that contain species-specific signature sequences which are useful for bacterial and archaeal identification at the species level.

MLEE
The characterization of bacterial species by the relative electrophoretic mobility of approximately 15 cellular metabolic enzymes.

MLST
An unambiguous procedure for characterizing isolates of bacterial species using the sequences of internal fragments of (usually) seven housekeeping genes. Approximately 450–500 bp internal fragments of each gene are used, as these can be accurately sequenced on both strands using an automated DNA sequencer.

Pan-genome
The global gene repertoire of a bacterial species that comprises the sum of the core and the dispensable genome (from the Greek '*pan*', meaning whole, and genome).

SNP
DNA sequence variation that occurs when a single nucleotide in the genome differs between members of a species.

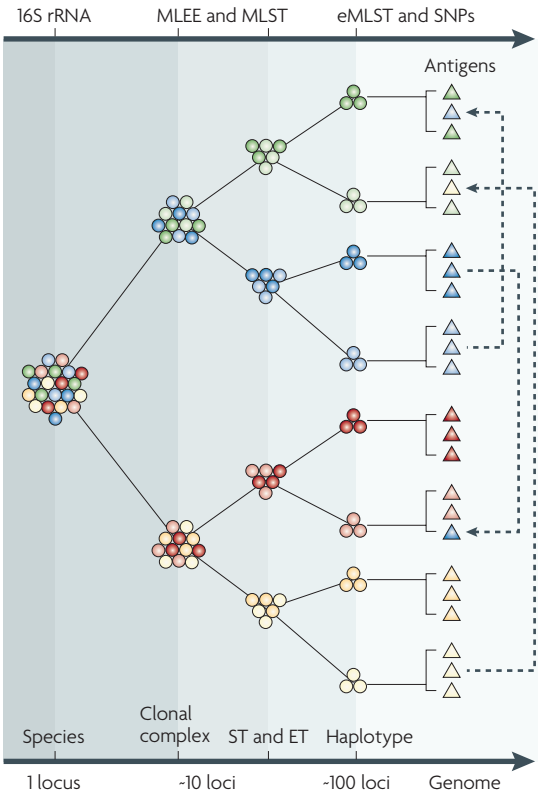

Figure 2 | **Genetic markers and deviations from population structure.** Schematic representation of different resolution levels within a typical population structure as identified by various typing schemes. Ribosomal RNA (rRNA) typing is the gold standard to differentiate species from other members of the same genus, class or even kingdom but, being based on a single locus, frequently lacks intra-species resolution. Multilocus typing schemes that are based on ~10 loci, either via enzyme electrophoresis (MLEE) or housekeeping-gene sequencing (MLST), provide fine intra-species resolution by defining electrophoretic and sequence types (ETs and STs, respectively) and clusters of types that group into clonal complexes. By measuring single-nucleotide polymorphisms (SNPs) at ~100 loci or applying an extended MLST (eMLST) schema that includes dispensable gene sequences, it is possible to further increase the typing resolution and define species-specific haplotypes. However, various genes that encode protein antigens have allelic distributions that do not correlate with MLST classification and, in principle, only complete genome coverage will be able to detect all the non-clonal genetic variations that shape the fine structure of a bacterial population.

In fact, mutations (any change in the DNA sequence) provide a means by which an organism can alter its fitness and evolve through natural selection. The evolutionary fate of any organism can be reduced to a simple paradigm — whether it survives or not. Over time, bacteria face environmental challenges that test their fitness. In encounters with a hostile, uncertain and fluctuating world, the genetic variation of bacterial populations is a major determinant of the equilibrium between the fidelity of genome replication and the genetic variation that is selected by organismal evolution. These sequence variations reflect the life history of the bacteria and their interactions with the environment. Unfortunately, the obvious solution of using high-throughput genomic sequencing to sample all the similarities and differences between bacterial populations might not be as close as is often stated.

Advances in high-throughput genomics

Several billion bases have been entered into GenBank (85,759,586,764 bases in February, 2008) during the 13 years since the 1.8 million bp genome of *Haemophilus influenzae* was completely sequenced. Virtually all of the sequences that have been deposited in GenBank and other sequence databases have been obtained by the Sanger chain-termination method, which was developed in the late 1970s[24] and led to a second Nobel Prize for Fred Sanger. This method was improved considerably in the 1980s[25–27] to increase the average length of the sequence reads from 450 to ~850 bp, which established the gold standard for DNA sequencing.

Perhaps the main limitation of Sanger sequencing technology is the cloning step that is required to make the bacterial libraries and the consequent need for directed sequencing of non-clonable regions. Additionally, large-scale Sanger-sequencing projects require a substantial infrastructure, and today such projects are mostly concentrated in a few large facilities that can generate reliable genomic data (including reliable annotation) at reasonable costs.

In the past few years, unprecedented efforts have therefore been made to develop and deploy new sequencing strategies[28,29]. Three new methods are currently being commercialized that are based on amplification strategies as alternatives to the standard cloning system and use different methods of sequence detection: high-throughput pyrosequencing on beads[30], sequencing by ligation, also on beads[31] and sequencing by synthesis on DNA that is amplified directly on a glass substrate[32,33] (FIG. 3). Most of these new methods use PCR to amplify individual DNA molecules that are immobilized on solid surfaces, either beads or a glass surface, such that all the identical molecules present can be sequenced in parallel using various sequencing approaches. Pyrosequencing uses the light that is emitted by the release of pyrophosphates that are attached to the incorporated bases; the sequencing-by-synthesis method uses fluorescent reversible dye terminators (which is conceptually similar to Sanger sequencing, but with a single base extension); and the sequencing-by-ligation method uses the ligation of a pool of partially random oligonucleotides that are labelled according to the discriminating base or bases. Although extremely fast, these methods still only give short sequence reads, making subsequent sequence assembly problematic. For example, pyrosequencing currently provides read lengths of up to 250 bp and the sequencing-by-synthesis and sequencing-by-ligation methods currently generate read lengths of up to 50 bp. It is highly likely, however, that the read lengths and sequence quality that can be obtained using these methods will improve considerably in the future.

All these technologies have their particular limitations in terms of read length or accuracy profiles. Integration of Sanger capillary sequencing and one or more of the

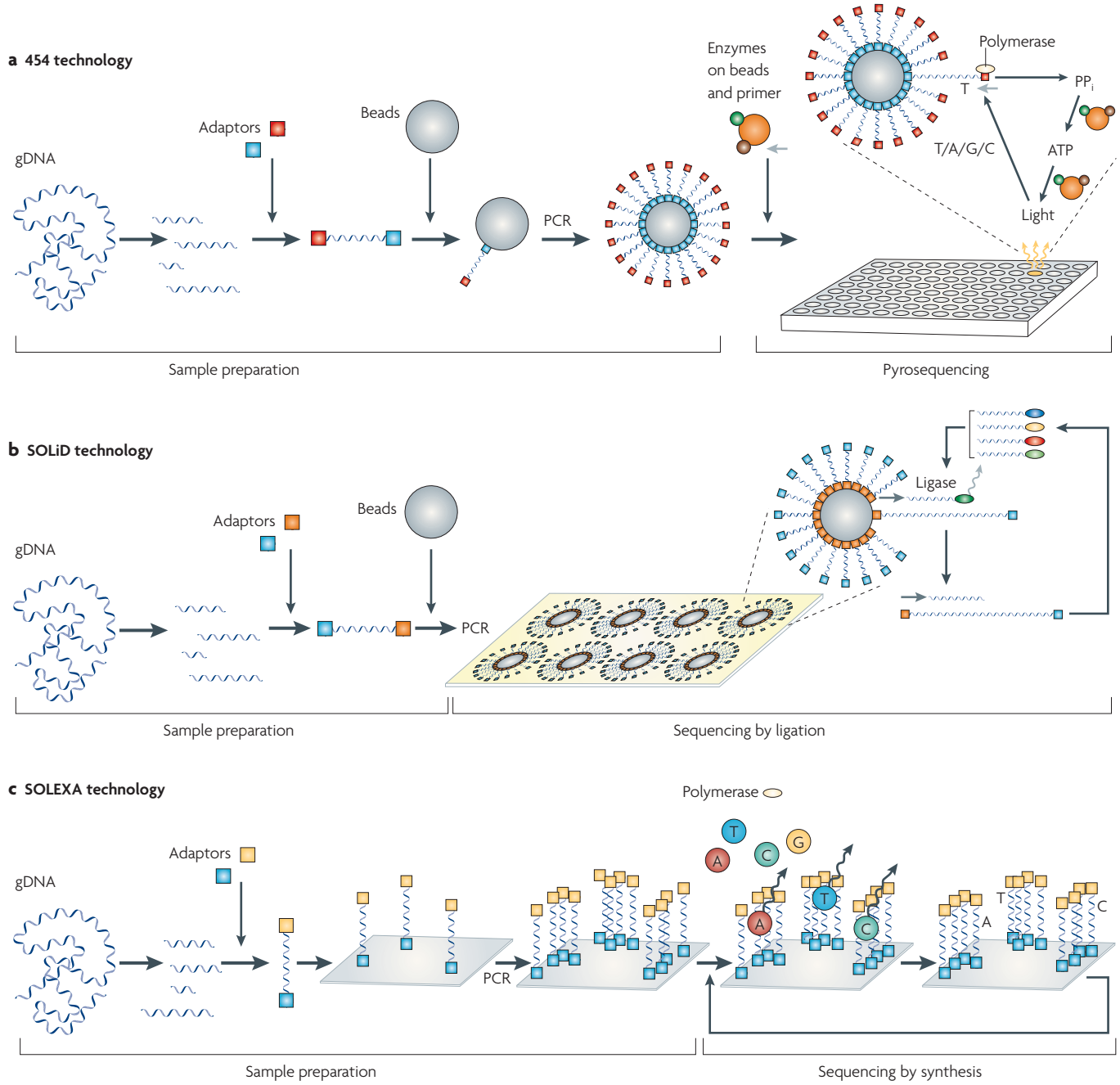

Figure 3 | **Post-Sanger sequencing technologies. a** | The 454 sequencing method is a highly parallel, two-step approach. First, the DNA is sheared and oligonucleotide adaptors are attached. Each fragment is attached to a bead and the beads are PCR amplified within droplets of an oil–water emulsion. This generates multiple copies of the same DNA sequence on each bead. Second, the beads are captured in picolitre-sized wells in a fabricated substrate and pyrosequencing (pyrophosphate-based sequencing) is performed in parallel on each DNA fragment as shown (the DNA fragment has been artificially elongated in the figure). Nucleotide incorporation is detected by the release of inorganic pyrophosphate (PPi), which leads to the enzymatic generation of photons: PPi is released and converted to ATP and luciferase uses the ATP to generate light. The cycle is iteratively repeated for each of the four bases. The average read length has already increased from 110 bp to approximately 250 bp, and future developments will probably increase it further to over 400 bp.. **b** | SOLiD technology has an amplification procedure that is conceptually similar to that of 454, but the sequencing strategy is radically different. Beads are deposited onto glass slides and the sequence is determined by sequential hybridization and ligation of partially random oligonucleotides with a central determined base (or pair of bases) that is identified by a specific fluorophore. After the colour is recorded, the ligated oligonucleotide is cleaved and removed and the process is then repeated. The reads that are generated are currently ~25 bp, but will probably increase to more than 50 bp in the future. **c** | The first step of SOLEXA sequencing is based on the amplification of DNA on a solid surface using fold-back PCR and anchored primers. Multiple cycles of the solid-phase amplification followed by denaturation create clusters of ~1,000 copies of single-stranded DNA molecules. Sequencing is performed sequentially using primers, DNA polymerase and four fluorophore-labelled, reversibly terminating nucleotides. After the incorporation of a nucleotide, the image is captured and the identity of the first base is recorded. The terminators and fluorophores are then removed and the incorporation, detection and identification steps are repeated. The average read length is currently ~40 bp, but this will also probably increase in the future.

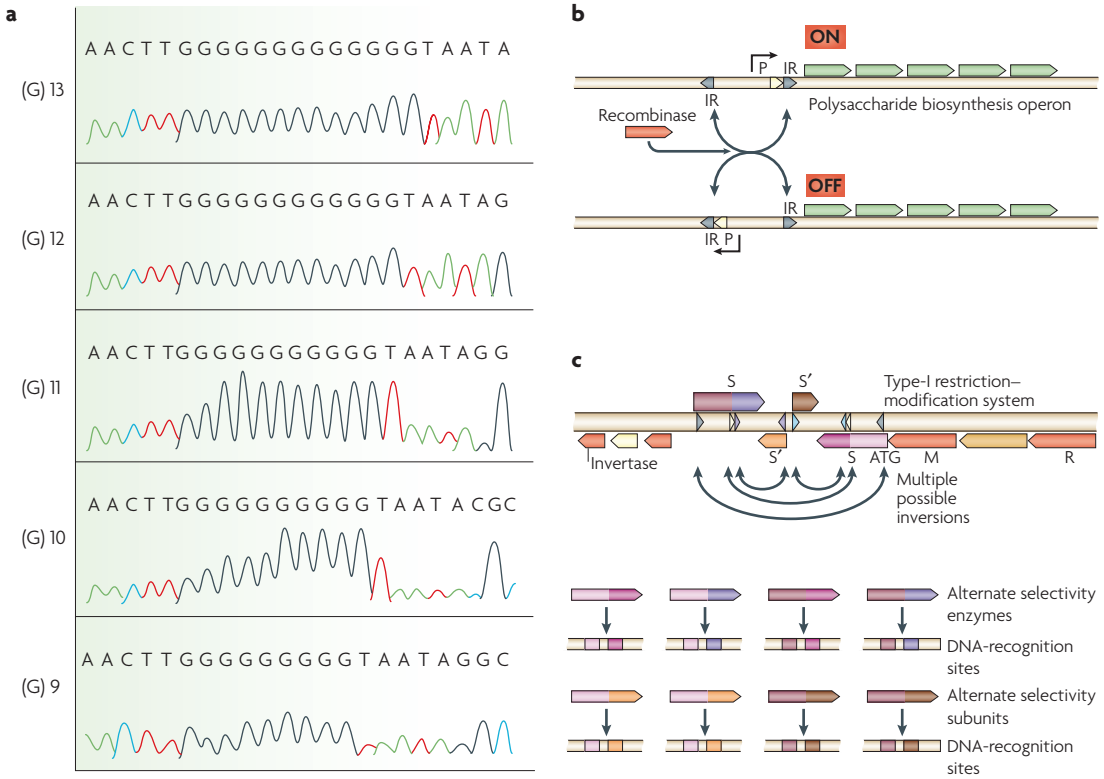

Figure 4 | **Rapid mechanisms that generate diversity within a clonal culture.** Specific mechanisms in many microorganisms introduce diversity into a clonal culture that results in heterogeneous DNA reads in shotgun sequencing. **a** | Schematic depiction of the hypervariable repeats in *Campylobacter jejuni*[35], showing regions of the chromosome where different reads showed a different number of bases in a particular run of the same base (guanosine; G). **b** | Invertible promoters of *Bacteroides fragilis*[38,73]: rapid inversion of a small section of DNA, which is mediated by a site-specific recombinase, is used to control the expression of entire operons that encode surface-polysaccharide production, which can be switched on and off as single entities with a single promoter. **c** | Random expression of different alternative proteins from a single locus in *Bacteroides fragilis*: rapid DNA inversion is also used to exchange parts of the coding sequences of expressed genes with sequences from silent cassettes, thereby altering the encoded protein sequence. IR, inverted repeat; M, S and R, methylation, specificity and restriction subunits of the restriction endonuclease system; P, promoter.

next-generation methods has been shown to be, at present, the best solution for *de novo* bacterial genome sequencing in terms of sequence quality and cost-effectiveness[34]. However, increases in read length and throughput will probably substantially broaden the work that can be accomplished with these technologies; indeed, in 2004, the National Institutes of Health set the scientific community the '$1,000 human genome' challenge to be achieved by 2015, which looks more and more achievable. The great enthusiasm that was generated by these advances has led at least some in the microbial genomics community to consider the alternative goal of a '$1,000 genome in a day' to be closer for bacteria than for the larger and more complex genomes of eukaryotic species. However, bacterial genomes are extremely variable in terms of the presence or absence of genes and often have large regions that need to be sequenced *de novo*, which potentially limits the usefulness of these new short-read technologies. Moreover, even the highly clonal populations of bacteria that are used for sequencing contain considerable sequence variability, such that the genome sequence that is obtained represents a population average rather than the sequence of any of the individual bacteria. This variability is not a sequencing artefact that can be addressed by increasing the signal to noise ratio with higher coverage or various technical improvements, but is in fact *bona fide* variation that is generated by specific bacterial mechanisms (discussed below)[35–39] (FIG. 4).

One solution to this problem could be provided by the recently devised single-molecule sequencing techniques, either by mimicking the natural enzymatic process of DNA synthesis by DNA polymerase[40] or by reading single bases as DNA molecules pass through nanopores[41–43]. Although these techniques are at an early stage of development, they represent a radical change in sequencing methodology and avoid the need for analysis of a mixture of multiple templates for the same DNA molecule. Although this will provide new opportunities for sequencing well-defined individual organisms, which have genetic information that is clearly distinguished from the genetic information of other individuals, this will not change the fact that, in the bacterial world, interaction with environmental niches or specific host niches often occurs at the population level. Consequently, both the averaging process that is required to generate a consensus sequence from a population and direct sequencing from a single bacterial cell will hide

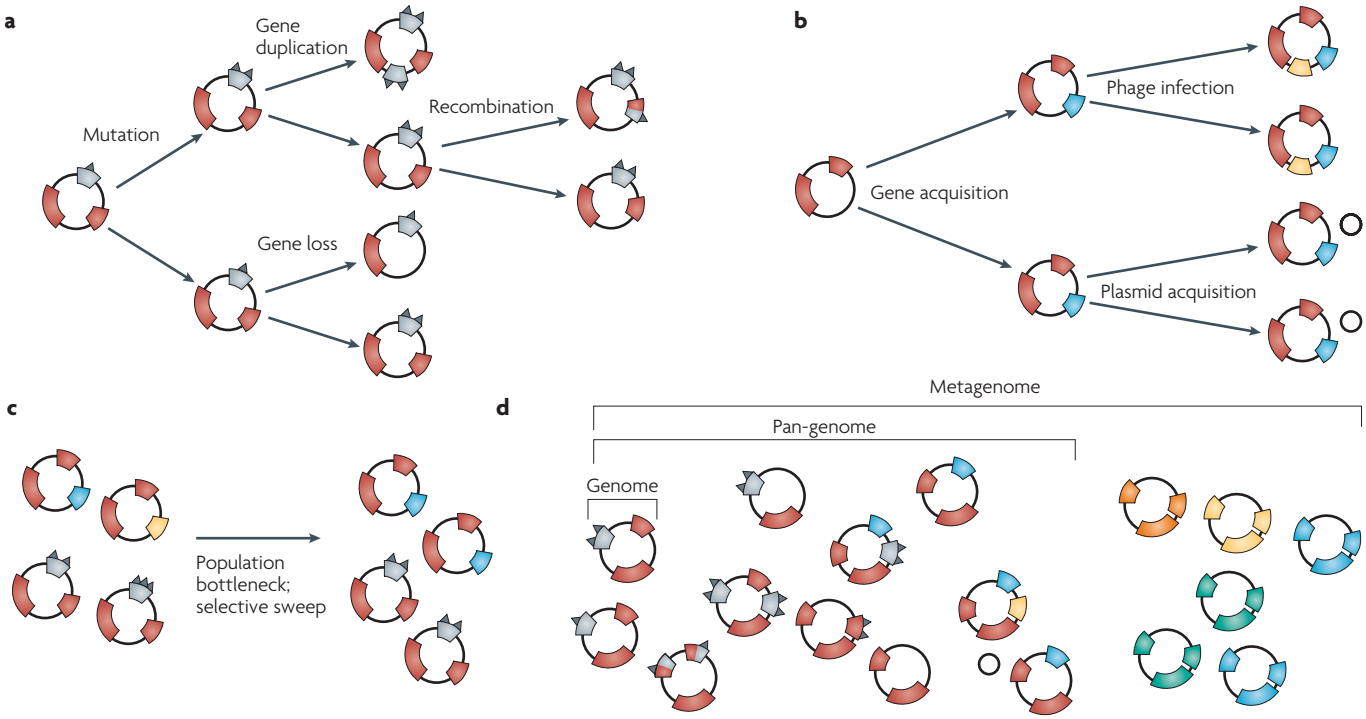

Figure 5 | **Molecular evolutionary mechanisms that shape bacterial species diversity: one genome, pan-genome and metagenome.** Intra-species (**a**), inter-species (**b**) and population dynamic (**c**) mechanisms manipulate the genomic diversity of bacterial species. For this reason, one genome sequence is inadequate for describing the complexity of species, genera and their inter-relationships. Multiple genome sequences are needed to describe the pan-genome, which represents, with the best approximation, the genetic information of a bacterial species. Metagenomics embraces the community as the unit of study and, in a specific environmental niche, defines the metagenome of the whole microbial population (**d**).

crucial information on the adaptive strategies that are encoded in intrinsically variable bacterial genomes. Direct PCR amplification and detailed study of particular loci could therefore remain crucial steps in bacterial genome sequencing, even if complete closure of the genome is not strictly required. The new sequencing technologies have greatly reduced the amount of time and money that is required to determine a complete bacterial genome sequence, but, as a bacterial genome represents the average of a population, the process still requires human interpretation of raw data and the integration of diverse sequencing methods. Paradoxically, therefore, although the target of a $1,000 genome in a day is much closer than could have been predicted a decade ago, it is clear that much more remains to be learned from 'simpler' targets, such as single bacterial genomes.

From genome to pan-genome and metagenome

When the first complete bacterial genome sequence was published, it was commonly thought that a few dozen more genomes, chosen representatively from the bacterial and archaeal domains, would be sufficient to describe the gene pool of the entire microbial world. Today, as the number of fully sequenced microbial genomes approaches 700, it is clear that microbial diversity has been vastly underestimated and we are just 'scratching the surface'. Variability, which as discussed above is already present within a single clonal culture that is established from a single cell, increases greatly within a single bacterial species (the pan-genome) and goes beyond sensible estimation when we consider the gene pool of microorganisms in the environment (the metagenome; FIG. 5).

The size of the genomes of known, free-living bacteria ranges from 0.16 Mb for *Candidatus* Carsonella ruddii to 10 Mb for *Solibacter usitatus*. Evolutionary forces seem to drive the size of the genome. On the one hand, many parasitic or symbiotic microorganisms do not need all of the genetic material that would be necessary to support life if they were independent of their hosts and genes that are redundant therefore tend to be eliminated, leading to a small genome. On the other hand, bacteria that face continuously variable environments need a large gene pool to address different needs; these bacteria have the largest genomes, which, in some cases, can reach twice the size of the smallest eukaryotic genomes[44].

In many species, there is extensive genomic plasticity. For example, completion of the genome sequence of *E. coli* O157:H7 revealed that this strain possesses >1,300 strain-specific genes compared with *E. coli* K12; these genes encode proteins that are involved in virulence and metabolic capabilities[45]. Moreover, when the genomes of three *E. coli* strains (K12, O157:H7 and the uropathogenic strain CFT073) were compared, only 39.2% of genes could be found in all three strains[46]. Other reports have also revealed an extensive amount of genomic diversity among strains of a single species[15,47,48].

From these studies, it is evident that it is not possible to characterize a species from a single genome sequence. But

how many genome sequences are necessary? The answer might vary from species to species. The study by Tettelin and colleagues[15] discussed above that examined the diversity of 8 isolates of *S. agalactiae* revealed that each new genome had an average of 30 genes that were not present in any of the previously sequenced genomes, which suggested that the number of genes associated with this species could, theoretically, be unlimited. Therefore, the best approximation to describe a species could be made by using the concept of the pan-genome. The pan-genome can be divided into three elements: a core genome that is shared by all strains; a set of dispensable genes that are shared by some but not all isolates; and a set of strain-specific genes that are unique to each isolate.

For *S. agalactiae*, the core genome encodes the basic aspects of *S. agalactiae* biology and the dispensable and strain-specific genes, which are largely composed of hypothetical, phage-related and transposon-related genes[49], contribute to its genetic diversity. This contrasts with *B. anthracis*, as the pan-genome for this species can be adequately described by just four genome sequences. This difference in the nature of the pan-genome reflects several factors, including: the different lifestyles of the two organisms (exposure of *S. agalactiae* to diverse environments versus the occupation of a more isolated biological niche by *B. anthracis*); the ability of each species to acquire and stably incorporate foreign DNA, an advantage in niche adaptation from the acquisition of laterally transferred DNA; and the recent evolutionary history of each species. Recently, a refined model for the *H. influenzae* pan-genome was proposed, which predicted that even if the overall number of genes that pertain to this species is finite, the number of genes in the pan-genome will still be four to five times higher than the number of genes in the *H. influenzae* core genome[50]. A similar analysis of 17 *Streptococcus pneumoniae* genomes revealed a core genome of 1,454 genes and a pan-genome (named the supra-genome) of approximately 5,000 genes[51]. Based on the authors' assumptions, 142 genomes would need to be sequenced to obtain the complete *S. pneumoniae* genome. In conclusion, there is agreement that the size of the pan-genome is much larger than the size of the genome of a single isolate of a particular species. However, the theoretical interpretation of these data with regard to the ultimate size of the pan-genome differs slightly depending on the assumptions that are used for the mathematical models. Tettelin *et al.*[15] assume that the pan-genome can be large and theoretically unlimited, whereas Hogg and Hiller[50] make more conservative estimates and predict a large, but finite, pan-genome. Determining which of these hypotheses is correct will require the accumulation of more data to facilitate the construction of more accurate mathematical models.

Whatever its ultimate size, the pan-genome reflects the selective pressure on several species to generate new adaptive combinations by recombining and constantly restructuring gene variants (alleles) in the population and by lateral gene transfer between species. Several natural processes — transport by viruses (transduction), bacterial 'mating' (conjugation) and the direct uptake of DNA from the environment (transformation) — carry genetic information from one species to another[52]. These processes, which are regulated and evolutionarily conserved, are turned on when they are most likely to result in gene transfer, and genes that must function together are often transferred together as genomic islands (for example, pathogenicity islands)[11].

The concept of the pan-genome is not just a theoretical exercise; it also has fundamental practical applications in vaccine research. Recently, it was shown that the design of a universal protein-based vaccine against GBS was only possible using dispensable genes[53]. In addition, the sequencing of multiple genomes was instrumental in discovering the presence of pili in GBS, group A *Streptococcus* and *S. pneumoniae*, an essential virulence factor that had been missed by conventional technologies for a century[54].

Although the size of the gene pool for a species can be estimated by mathematical modelling, the size of the gene pool for the microbial biosphere is beyond any credible model. The projects that have been completed to date have focused on species that can be grown in culture, but it has been estimated that >99% of the bacteria in the environment cannot be cultured in the laboratory[55]. In nature, bacteria occupy diverse environmental niches as complex communities in which they interact with each other and with the surrounding environment, and acquire and discard genes by lateral gene transfer events (FIG. 5). This emphasizes the importance of the local environment in shaping the genomic evolution of individual community members. In the past few years, a new approach known as metagenomics (also called environmental genomics or community genomics) has emerged that involves the use of genome sequencing and other genomic technologies to study microorganisms directly in their natural habitats[56]. Metagenomics embraces the community as the unit of study, and includes sequence-based and product- or function-based methods for analysing environmental samples directly, without the need for isolation of discrete organisms. This process has been pursued in recent years for a few microbial communities and habitats: an acid-mine drainage site, the Sargasso Sea, agricultural soil, a deep-sea whale skeleton and the human distal gut[57–61]. These studies have identified hundreds of unknown bacteria and viruses and millions of new genes, revealing an unexpected degree of diversity[62] (FIG. 5). The use of metagenomic approaches has also provided a new perspective on host–microorganism mutualistic interactions (BOXES 1,2).

Complex microbial communities are found in a range of human habitats, including the female reproductive tract, the skin, the oral cavity and the gut. These communities have co-evolved with their human host and play an important part in human health and disease[63]. For example, the human intestinal microbiota is composed of >1,000 species and the concept of humans as 'super-organisms' is highlighted by estimates that the human microbiome contains roughly 100 times more genes than the human genome[64–66]. The microbiome can be viewed as a human 'accessory' genome that complements the functions which are provided by the human germline and provides the host with flexibility, diversification and adaptability in the face of a rapidly changing environment.

Core genome
The pool of genes that is shared by all the strains of the same bacterial species.

Lateral gene transfer
The mechanism by which an individual of one species transfers genetic material (that is, DNA) to an individual of a different species.

Metagenomics
The study of the genomic repertoire of all the organisms that live in a particular environment and their activities as a collective. The genomic analysis is applied to entire communities of microorganisms, which bypasses the need to isolate and culture individual microbial species.

Box 1 | **Metagenomics of mammalian hosts**

Although metagenomics can be applied to the cellular and genomic variants that arise during the growth of a single cell, it is more frequently applied to complex environmental samples that contain multiple strains or species, which emphasizes the community as the unit of study[56]. Currently, DNA sequencing is the most common form of analysis, but other community components have been targeted, including RNA (for example, using high-density hybridization platforms or pyrosequencing), proteins (for example, using mass spectroscopy) and other metabolites. Libraries of molecules can be screened for function, thereby enhancing the information that can be extracted from these complex environments. Because it remains difficult or impossible to obtain pure cultures of most members of naturally occurring microbial communities, traditional approaches for the analysis of individual community members impose major constraints and limitations.

Although this field is still in its infancy, metagenomics has already revealed previously unappreciated features of microbial genomics and microbial biology, and has hinted at the kinds of insights into the complex microbial communities upon which we, humans, depend that will be provided by this approach. These features include: information on metabolic capabilities, evidence of host co-evolution and, perhaps, community-wide selection; the basis of cell–cell signalling; and emergent features of complex communities that would not otherwise be apparent from the study of their individual components. For example, the metagenomic sequence data of faecal specimens from healthy, lean individuals emphasized the importance of isoprenoid and essential-vitamin biosynthesis, amino acid, xenobiotic and dietary polysaccharide metabolism, and methanogenesis by demonstrating a relative enrichment of genes that are involved in pathways associated with these processes[60]. Similar analyses of gut specimens from obese mice revealed enrichment of genes that are associated with the harvesting of energy from dietary plant polysaccharides, which presumably reflects metabolic activities within the gut microbiome that complement the digestive capabilities of the host[61].

A number of large international efforts have now been initiated to explore and gain a better understanding of the human microbiome[89,90]. Metagenomics will play a prominent part in these efforts. There are several potential benefits from a metagenomic-enabled characterization of the human microbiome: early recognition and prediction of community disturbances that are associated with disease; identification and validation of novel end points for treatment (restoration of health) that are based on patterns of microbial community structure; identification of novel targets for therapy (for example, microbial pathways and inter-cellular communication signals); discovery of novel therapeutics from the products and components of human microbial communities; identification of novel strategies for preventing pathogen invasion; and new insights into the mechanisms of disease.

Before these potential benefits of human-based metagenomics can be realized, however, several challenges must be addressed. Much of the bacterial diversity in the human body is found at the level of species and strains. The prominence of strain-level diversification has major implications for shotgun sequencing of the human microbiota: error rates must be reduced, more sophisticated assembly algorithms will be required and the association between genotype and phenotype for closely related strains must be more closely examined. In addition, we have not yet determined the relevant spatial scale at which clinical sampling of the human body should be undertaken. An equally important and still under-appreciated challenge is the need for robust, standardized, clinical metadata on human-derived specimens, without which effective analysis of metagenomic data is severely impaired. Finally, the dramatic unevenness of the human microbiota (for which there is large variation in the relative abundance of different members) and the potential crucial importance of rare members (for example, as keystone species) demands improved methods for community analysis, including methods for normalization, selective enrichment and cultivation of previously ignored phylotypes, and single-cell analysis. Single-cell genomic analyses are becoming increasingly feasible[85]. As technologies for the precise manipulation of single microbial cells improve, the targeted analysis of individual members (especially rare members) from complex communities will complement the shotgun approach of metagenomics and enhance the interpretation of community-wide data, as well as facilitate the analysis of rare members that have crucial roles in the stability and net functions of the community as a whole.

To date, more than 100 metagenomic projects are ongoing, the goals of which vary from characterizing a particular species to understanding the dynamics of an entire microbial community. The most obvious advantage of sequencing DNA from natural samples is the capacity to access a broad range of genome sequences. From a broader perspective, metagenomics has the ability to capture the genomic diversity within a natural population, thereby offering the promise of assessing biodiversity in a new way that is independent of the bacterial species concept[67]. One possible outcome of these projects is the characterization of bacterial populations as a continuum of genomic possibilities[68].

Metagenomics of single organisms

The genomic analysis of mixed populations is increasingly being applied to studies of environmental samples that contain large numbers of different species and genera. However, it is not often appreciated that even a standard shotgun sequence of a single organism can be considered to be a metagenomic sample of a population. In most cases, shotgun-sequence libraries are made from large amounts of DNA that has been isolated from clonal cultures. However, to generate a contiguous sequence, and ensure accuracy, these libraries are over-sampled, usually by eightfold to tenfold, with multiple individual sequence reads contributing to each base in the final sequence. The mechanics of library construction mean that, effectively, each individual read has probably come from a different cell, and therefore the shotgun sequence is an eightfold-to-tenfold-deep sample of the population at each base position. As the sample is usually clonal, this often has no real consequence. However, if there is variation within the population, then the redundant sampling will contain that variation, which can be detected in the shotgun sequences.

Box 2 | **Bacterial pathogenesis in the post-genomic era**

Pathogenicity, like other microbial traits, is a reflection of a particular specialization, which could be as simple as living in the ocean or in soil. There are similar genetic processes at play and common themes of survival. The application of genetic and molecular methods to the study of microorganisms that cause infection and disease has been propelled by the fact that we now possess at least one complete genomic sequence of virtually all bacterial species that cause human infectious disease. Yet, one cannot simply examine the complete genome of a bacterium on a computer screen and deduce from its sequence whether or not it is a pathogen. There is no 'core' set of genes that defines pathogenicity for a particular bacterial species, and it is still necessary to do experiments to understand bacterial pathogenesis, even in the post-genomic era. Moreover, the distinction between what is a pathogen, as compared with a commensal species, is blurred. Many of the recognized pathogens, such as the pneumococcus and the meningococcus, are far more likely to be carried asymptomatically than to cause clinical disease. What we have considered to be virulence factors are, in reality, colonization factors in these bacteria. Bacterial pathogenesis is not necessarily reflected in death and disease (except, perhaps, when writing a grant application).

To understand what constitutes a pathogen and pathogenicity, one must take into account that the evolution of pathogenicity paralleled the evolution of the host. The initial adaptation of bacteria to humans took place at a time when human life expectancy was much shorter than it is today. The extension of human life, paradoxically through the control of microorganisms and infectious diseases, now places an extraordinary selective pressure on microorganisms to evolve their specialization to survive in humans. This entails more than antibiotic resistance. We have observed, for example, the evolution of staphylococcal pathogenicity and transmissibility in the past few decades in response to the selective pressure of antibiotics, changes in human demographics and the appearance of new microbial opportunities as a result of non-genetic host changes. The host–pathogen dynamic in humans is still a 'work in progress', as reflected by the deluge of emerging infectious diseases.

Fortunately, it is the availability of the genome of both hosts and the microorganisms that live upon and within these hosts that provides us with the newest and, in our opinion, the best experimental approaches to examine the contribution of specific genes to the host–pathogen dynamic. Our ability to monitor the response of a host to a wild-type bacterial strain, as compared with a single defined mutant of this strain, tells us a great deal about bacterial pathogenesis and host-defence mechanisms. Similarly, our increasing ability to examine the response to microbial challenge in hosts that differ by a defined genetic alteration supplies us with complementary information about both the microorganism and its host. Thus, the genomes of both hosts and their pathogens have become the new foundation for research on the biological implications of pathogenicity and pathogenesis.

This becomes of more than esoteric interest when dealing with bacteria that are pathogens or commensals of larger organisms. These bacteria have a problem as, barring rare mutations, growth by binary fission generates a clonal population of identical cells, but evading or avoiding an immune system or colonizing a highly variable environment requires diversity. Bacteria can escape this constraint in two ways: by promoting the exchange of DNA with other related organisms or rapidly generating diversity within an otherwise clonal growth. The population sampling effect of shotgun sequencing allows us to identify and understand these mechanisms directly from raw genome-sequence data.

Diversity-generating mechanisms usually involve the random, but heritable, on or off switching of surface-exposed structures that leads to mixed populations of cells. This phenomenon is known as phase variation and has been extensively studied for many years[10]. The simplest mechanism of this switching is the random change in length (during DNA replication) of short repetitive tracts of bases. The genome sequencing of *Campylobacter jejuni*, a bird commensal and human pathogen, identified regions of the chromosome where different reads showed a different number of bases in a particular run of the same base[35] (FIG. 4a). Investigations ruled out experimental errors as the source of this variation and showed that the effect was due to the shotgun library sampling of individual cells, which, in fact, had different sequences at these specific loci. The context of these variants was investigated, and most were found to be within protein-coding sequences and had the effect of switching the reading-frame of the gene such that it could, or could not, be translated. Thus, the expected clonal population was in fact a mixture of genomes with variant sequences that expressed different subsets of surface proteins. The metagenomic, population-sampling effect of the genome-sequence libraries allowed the immediate identification of these variant sequences and the proteins they affected, which provided an overview of the variable gene set of the organism directly from the assembled sequence. A similar effect was observed in the related organism *Helicobacter pylori*[69], but not in *Neisseria meningitidis*[70,71], which is another organism that is known to use this mechanism of phase variation. This suggests that variation occurs at different rates in the two groups and indicates that the single-organism metagenomics approach only works for rapid-variation mechanisms.

A second common mechanism for randomly varying the expression of surface structures is DNA inversion. Here, a small section of DNA is inverted by a cleavage-and-ligation reaction that is mediated by a site-specific recombinase. At its simplest, the effect of this inversion can be to alter the direction of transcription from a promoter within the inverted segment, thereby switching on or off the expression of genes that are downstream of the promoter. Again, this mechanism has been known and studied for some time, most notably in the control of phase variation of the flagellar antigen of *Salmonella enterica* serovar Typhimurium[72], but population sampling through single-organism metagenomic sequencing allows its presence and extent to be immediately identified within whole genomes directly from the sequence data.

One good example of this is *Bacteroides fragilis*, a human commensal and opportunistic pathogen[38,73]. Genome sequencing revealed that this mechanism was active, rapid and used to control the expression not just of surface proteins but of entire operons that encode surface-polysaccharide production, which can be switched on or off as single entities with a single promoter (FIG. 4b). The genome sequencing of *B. fragilis* also directly showed that this mechanism is used not only to alter the expression of genes, but also to exchange parts of the coding sequences of expressed genes with sequences from silent cassettes, thereby altering the encoded protein sequence (FIG. 4c). This allows the random expression of different alternative proteins from a single locus, a process that has also been described (and elucidated from genomic shotgun sequences) in *Mycoplasma pulmonis*[74].

A fundamental level of diversity between bacterial genomes is at the level of the single base. It is clear that bacteria can generate diversity at this level by point mutation and DNA recombination. Mechanisms have also been uncovered that use, for example, reverse transcription to generate point mutations in targeted regions; this was initially described in a bacteriophage of *Bordetella bronchiseptica*[75]. Metagenomic analysis of shotgun data from *Tropheryma whipplei*, a human pathogen, has also provided evidence of single-base diversity generation on a genomic level[39]. In this case, the rate of diversity generation by the organism in its natural habitat was less clear-cut, as this fastidious organism had to be grown in human cell culture for 17 months to provide enough DNA for sequencing. However, it was evident that consistent variation between shotgun sequences could only be seen in specific locations in the genome that corresponded to the coding sequences of surface-expressed proteins. It was proposed that the mechanism for generating variation in *T. whipplei* is the transfer of base-pair variants from non-coding repeats elsewhere in the genome, possibly by a gene-conversion-like mechanism.

It is clear therefore that it is possible to perform metagenomics on single organisms, through careful analysis of the variation within shotgun sequences and an understanding that this variation represents diversity within an otherwise clonal population. In turn, this allows us to identify and examine the biological consequences of rapid diversity-generating mechanisms in numerous bacterial pathogens and commensals.

Implications for microbiology and conclusions

Over the past decade, genomic technologies have revolutionized microbiology and will probably continue to do so during the next decade. The information that is being added to sequence databases is increasing exponentially and every day we are in a better position to describe microorganisms by their genome, bacterial species by their pan-genome and even complex microbial environments by their metagenome. The pan-genome concept, which was predicted by a mathematical model before it was demonstrated biologically[15,76], is an example of how mathematics is becoming increasingly important in microbiology, not only to manage new information, but also to drive new discoveries by challenging conventional biological assumptions. We look forward to developments in the new discipline of systems microbiology, a research area that aims to accurately describe the dynamics of the microbial world with the assistance of mathematical models.

In the meantime, as our knowledge increases, we realize that we are just scratching the surface of the microbial world. We still have much to learn about non-cultivable microorganisms, human and animal metagenomes and the metagenomes of soil, oceans and extreme environments. The deposition into GenBank (see Further information) of 1.8 million genes from a single environmental investigation[77] — twice the size of the entire pre-existing gene database — is an example of our underestimation of the diversity of the microbial gene repertoire. In this study, the analysis of the extraordinary amount of fragmented sequence data required the development of *ad hoc* strategies for sequence reconstruction and clustering into protein families to reduce the dataset into a tractable form[78]. Unsupervised methods of protein-family generation from complete genome data have also been proposed, to enable high-throughput protein classification to be achieved without human intervention[79]. The question for contemporary microbiologists is how can we productively apply the avalanche of new information without getting lost in the details?

Some applications of the genomic era can already be seen. One of the first of these, reverse vaccinology[80,81], led to the identification of hundreds of genes that encode antigens which mediate protection by validated mechanisms, and made it possible to develop vaccines against bacteria for which only a limited number of antigens was previously known. However, in general, we have little information on the biological role of the novel antigens that are identified by reverse vaccinology, and further laboratory work is therefore needed to understand the biology of these new antigens and make more informed decisions on the use of vaccines that are based on these targets.

Similarly, new antigens usually segregate in the bacterial population in a manner that is independent of conventional markers, such as those used to define serotypes, and genetic markers, such as those used in MLST, and so we must identify new ways to type bacteria. New typing systems need to incorporate whole genome sequences, including non-core genes, instead of just a few core genetic loci, as has been the case so far. Whether new typing systems will be based on SNPs or a combination of revised MLST and SNPs, or will use novel genetic markers that will be selected from the analysis of complete genomes, it is still too early to know. However, it is likely that they will follow the example of MLST in being open, internet-based collaborative models that allow researchers from hundreds of different laboratories and countries to contribute their data to centralized and publicly accessible databases[82].

Assuming this open-access paradigm also prevails in the forthcoming genome-wide bacterial population studies, this will produce an invaluable database of genotypic and phenotypic characteristics that relate to each single isolate, thereby allowing us to perform association studies that are aimed at identifying genotypic traits, such as

Reverse vaccinology
A genomic approach to vaccine development that searches the entire genetic repertoire of a pathogen for protective antigens.

specific polymorphisms, virulence factors and pathogenicity islands, as determinants of pathogenicity, carriage and enhanced or reduced transmission.

Constructing new typing systems will require the collection of epidemiological data that allow us to reconstruct the global population biology of each species. New mathematical models that are based on a comparison of multiple whole-genome sequences have recently been proposed[83] to identify homologous recombination events that disrupt a clonal pattern of inheritance. Correct inference of ancestry, particularly from the perspective of using the whole genome as a marker, is fundamental to our reconstruction of a coherent and potentially complete picture of bacterial population structures. This kind of approach was used to analyse the relationship between different *S. enterica* serovars: a recent convergent evolution between *S. typhi* and *S. paratyphi* A was suggested that was confined to one quarter of the genome owing to a highly non-random pattern of homologous recombination that, possibly, was connected with the adaptation of both lineages to the interaction with the human host[84].

Finally, metagenomics will probably become the driving force in microbiology in the future by shedding light on the 'dark matter' of the microbial world[85]. The single-cell genetic analysis of rare and non-cultivable microorganisms and the analysis of the metagenomes of the complex environments they inhabit will disclose information on unknown microorganisms, genomes and microbial communities that will almost certainly change the way we view microbiology[86]. Metagenomic studies performed in specific niches of the human host could produce a complete picture of the functional repertoire of bacterial communities as shaped by host–pathogen interactions and would complement information on the population structure[64,87]. Together, these data should facilitate our understanding of the molecular bases of specific interplays with the host[88] and define the boundaries between pathogenic and commensal sub-populations of the same species.

In conclusion, although genomic information is challenging our existing knowledge of bacterial species, typing systems and population biology, further clarification of these concepts is a challenge that is at the intersection of diverse disciplines. In the long term, mathematics might provide an accurate description of the microbial world, but, in the meantime, we need to go back to the laboratory to try to understand the biological relevance of the information that has been generated by genomics.

1. Dighe, A. S. *et al.* Comparison of 16S rRNA gene sequences of genus *Methanobrevibacter*. *BMC Microbiol.* **4**, 20 (2004).
2. Keswani, J. & Whitman, W. B. Relationship of 16S rRNA sequence similarity to DNA hybridization in prokaryotes. *Int. J. Syst. Evol. Microbiol.* **51**, 667–678 (2001).
3. Acinas, S. G. *et al.* Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**, 551–554 (2004).
4. Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. Lond. B* **361**, 1929–1940 (2006).
5. Sacchi, C. T. *et al.* Sequencing of 16S rRNA gene: a rapid tool for identification of *Bacillus anthracis*. *Emerg. Infect. Dis.* **8**, 1117–1123 (2002).
6. Huber, H. *et al.* A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67 (2002).
7. Selander, R. K. *et al.* Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl. Environ. Microbiol.* **51**, 873–884 (1986).
8. Caugant, D. A., Kristiansen, B. E., Froholm, L. O., Bovre, K. & Selander, R. K. Clonal diversity of *Neisseria meningitidis* from a population of asymptomatic carriers. *Infect. Immun.* **56**, 2060–2068 (1988).
9. Maiden, M. C. *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA* **95**, 3140–3145 (1998).
The first description of the MLST approach and its validation for *N. meningitidis*.
10. Moxon, R., Bayliss, C. & Hood, D. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.* **40**, 307–333 (2006).
11. Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. Genomic islands in pathogenic and environmental microorganisms. *Nature Rev. Microbiol.* **2**, 414–424 (2004).
Comprehensive review that presented results on lateral gene transfer for pathogenicity islands in pathogenic bacteria. Also showed that lateral gene transfer is a universal mechanism of evolution.
12. Falkow, S. in Escherichia coli *and* Salmonella typhimurium*: Cellular and Molecular Biology* (eds Neidhardt, F. C. *et al*.) 2723–2729 (ASM, Washington DC, 1996).
13. Urwin, R. *et al.* Distribution of surface protein variants among hyperinvasive meningococci: implications for vaccine design. *Infect. Immun.* **72**, 5955–5962 (2004).
14. Caugant, D. A. *et al.* Genetic diversity in relation to serotype in *Escherichia coli*. *Infect. Immun.* **49**, 407–413 (1985).
15. 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).
First quantification of the diversity of a single bacterial species on the basis of genomic sequences of multiple strains. The authors introduced and defined the pan-genome concept.
16. Moorhead, S. M., Dykes, G. A. & Cursons, R. T. An SNP-based PCR assay to differentiate between *Listeria monocytogenes* lineages derived from phylogenetic analysis of the *sigB* gene. *J. Microbiol. Methods* **55**, 425–432 (2003).
17. Robertson, G. A. *et al.* Identification and interrogation of highly informative single nucleotide polymorphism sets defined by bacterial multilocus sequence typing databases. *J. Med. Microbiol.* **53**, 35–45 (2004).
18. Weissman, S. J., Moseley, S. L., Dykhuizen, D. E. & Sokurenko, E. V. Enterobacterial adhesins and the case for studying SNPs in bacteria. *Trends Microbiol.* **11**, 115–117 (2003).
19. Read, T. D. *et al.* Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* **296**, 2028–2033 (2002).
20. Alland, D. *et al.* Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. *J. Bacteriol.* **185**, 3392–3399 (2003).
21. Gutacker, M. M. *et al.* Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* **162**, 1533–1543 (2002).
22. Filliol, I. *et al.* Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J. Bacteriol.* **188**, 759–772 (2006).
23. Roumagnac, P. *et al.* Evolutionary history of *Salmonella typhi*. *Science* **314**, 1301–1304 (2006).
24. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
25. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).
26. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336–341 (1987).
27. Madabhushi, R. S. Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions. *Electrophoresis* **19**, 224–230 (1998).
28. Hall, N. Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* **210**, 1518–1525 (2007).
This comprehensive review describes and discusses the recent advances in DNA-sequencing technologies.
29. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nature Methods* **5**, 16–18 (2008).
30. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
31. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
32. Bennett, S. Solexa Ltd. *Pharmacogenomics* **5**, 433–438 (2004).
33. Bennett, S. T., Barnes, C., Cox, A., Davies, L. & Brown, C. Toward the 1,000 dollars human genome. *Pharmacogenomics* **6**, 373–382 (2005).
34. Goldberg, S. M. *et al.* A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl Acad. Sci. USA* **103**, 11240–11245 (2006).
35. Parkhill, J. *et al.* The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–668 (2000).
36. Gundogdu, O. *et al.* Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. *BMC Genomics* **8**, 162 (2007).
37. Baker, S. & Dougan, G. The genome of *Salmonella enterica* serovar Typhi. *Clin. Infect. Dis.* **45** (Suppl. 1), 29–33 (2007).
38. Cerdeno-Tarraga, A. M. *et al.* Extensive DNA inversions in the *B. fragilis* genome control variable

gene expression. *Science* **307**, 1463–1465 (2005).
39. Bentley, S. D. *et al.* Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*. *Lancet* **361**, 637–644 (2003).
40. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proc. Natl Acad. Sci. USA* **100**, 3960–3964 (2003).
41. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl Acad. Sci. USA* **93**, 13770–13773 (1996).
42. Storm, A. J. *et al.* Fast DNA translocation through a solid-state nanopore. *Nano Lett.* **5**, 1193–1197 (2005).
43. Storm, A. J., Chen, J. H., Zandbergen, H. W. & Dekker, C. Translocation of double-strand DNA through a silicon oxide nanopore. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **71**, 051903 (2005).
44. Fraser-Liggett, C. M. Insights on biology and evolution from microbial genome sequencing. *Genome Res.* **15**, 1603–1610 (2005).
45. Perna, N. T. *et al.* Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–533 (2001).
46. 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).
47. Brochet, M. *et al.* Genomic diversity and evolution within the species *Streptococcus agalactiae*. *Microbes Infect.* **8**, 1227–1243 (2006).
48. Brzuszkiewicz, E. *et al.* How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains. *Proc. Natl Acad. Sci. USA* **103**, 12879–12884 (2006).
49. Tettelin, H., Medini, D., Donati, C. & Masignani, V. Towards a universal group B *Streptococcus* vaccine using multistrain genome analysis. *Expert Rev. Vaccines* **5**, 687–694 (2006).
50. Hogg, J. S. *et al.* Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol.* **8**, R103 (2007).
51. Hiller, N. L. *et al.* Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J. Bacteriol.* **189**, 8186–8195 (2007).
52. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
53. Maione, D. *et al.* Identification of a universal group B *Streptococcus* vaccine by multiple genome screen. *Science* **309**, 148–150 (2005).
The use of different GBS strains for genome analysis and screening enabled the development of a universal vaccine.
54. Lauer, P. *et al.* Genome analysis reveals pili in group B *Streptococcus*. *Science* **309**, 105 (2005).
55. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68**, 669–685 (2004).
56. National Research Council. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet* (The National Academies, Washington DC, 2007).
This book describes the applications and challenges of metagenomics.
57. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).
58. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
Reported the first assembled genome to emerge from a metagenomic analysis of environmental samples.
59. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
The first work in which the whole-genome shotgun-sequencing technique was applied to an environmental sample. Greatly improved our understanding of the complexity and variability of the world of uncultured bacteria.
60. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355–1359 (2006).
Described a metagenomic approach that helped define the gene content and encoded functional attributes of the gut microbiome in healthy humans through whole-genome shotgun sequencing.
61. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–1031 (2006).
Showed that the microbiome of obese mice has an increased capacity to harvest energy from the diet through metagenomic and biochemical analysis. Also identified the gut microbiome as being a contributing factor to the pathophysiology of obesity.
62. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. *Nature Rev. Genet.* **6**, 805–814 (2005).
This review presented the methodological advances that have allowed natural populations to be sequenced and gave examples of metagenomic studies that have used these techniques.
63. Dethlefsen, L., McFall-Ngai, M. & Relman, D. A. An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature* **449**, 811–818 (2007).
Discussed how humans have co-evolved with their microbial partners and examined how evolutionary and ecological principles are relevant to human–microorganism relationships.
64. Versalovic, J. & Relman, D. How bacterial communities expand functional repertoires. *PLoS Biol.* **4**, e430 (2006).
65. Ley, R. E., Peterson, D. A. & Gordon, J. I. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**, 837–848 (2006).
66. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635–1638 (2005).
67. Doolittle, W. F. & Papke, R. T. Genomics and the bacterial species problem. *Genome Biol.* **7**, 116 (2006).
68. Goldenfeld, N. & Woese, C. Biology's next revolution. *Nature* **445**, 369 (2007).
69. Alm, R. A. *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176–180 (1999).
70. Parkhill, J. *et al.* Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502–506 (2000).
71. Tettelin, H. *et al.* Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287**, 1809–1815 (2000).
72. Silverman, M., Zieg, J., Hilmen, M. & Simon, M. Phase variation in *Salmonella*: genetic analysis of a recombinational switch. *Proc. Natl Acad. Sci. USA* **76**, 391–395 (1979).
73. Kuwahara, T. *et al.* Genomic analysis of *Bacteroides fragilis* reveals extensive DNA inversions regulating cell surface adaptation. *Proc. Natl Acad. Sci. USA* **101**, 14919–14924 (2004).
74. Chambaud, I. *et al.* The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res.* **29**, 2145–2153 (2001).
75. Liu, M. *et al.* Reverse transcriptase-mediated tropism switching in *Bordetella* bacteriophage. *Science* **295**, 2091–2094 (2002).
76. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594 (2005).
77. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5**, e77 (2007).
78. Yooseph, S. *et al.* The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5**, e16 (2007).
79. Medini, D., Covacci, A. & Donati, C. Protein homology network families reveal step-wise diversification of Type III and Type IV secretion systems. *PLoS Comput. Biol.* **2**, e173 (2006).
80. Rappuoli, R. Reverse vaccinology. *Curr. Opin. Microbiol.* **3**, 445–450 (2000).
Definition of the reverse-vaccinology approach and its applications.
81. Pizza, M. *et al.* Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287**, 1816–1820 (2000).
The first application of reverse vaccinology: identified new vaccine candidates against *N. meningitidis* B.
82. Jolley, K. A., Chan, M. S. & Maiden, M. C. mlstdbNet — distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics* **5**, 86 (2004).
83. Didelot, X. & Falush, D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266 (2007).
84. Didelot, X., Achtman, M., Parkhill, J., Thomson, N. R. & Falush, D. A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? *Genome Res.* **17**, 61–68 (2007).
85. 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).
86. Dethlefsen, L. & Relman, D. A. The importance of individuals and scale: moving towards single cell microbiology. *Environ. Microbiol.* **9**, 8–10 (2007).
87. Dethlefsen, L., Eckburg, P. B., Bik, E. M. & Relman, D. A. Assembly of the human intestinal microbiota. *Trends Ecol. Evol.* **21**, 517–523 (2006).
88. Relman, D. A. Genome-wide responses of a pathogenic bacterium to its host. *J. Clin. Invest.* **110**, 1071–1073 (2002).
89. Relman, D. A. & Falkow, S. The meaning and impact of the human genome sequence for microbiology. *Trends Microbiol.* **9**, 206–208 (2001).
90. Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449**, 804–810 (2007).
Together with Reference 89, discusses the Human Microbiome Project, with the focus on its potential outcomes and conceptual and experimental challenges.
91. Bentley, S. D. *et al.* Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *PLoS Genet.* **3**, e23 (2007).
92. Sacchi, C. T., Whitney, A. M., Reeves, M. W., Mayer, L. W. & Popovic, T. Sequence diversity of *Neisseria meningitidis* 16S rRNA genes and use of 16S rRNA gene sequencing as a molecular subtyping tool. *J. Clin. Microbiol.* **40**, 4520–4527 (2002).
93. Deng, W. *et al.* Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. *J. Bacteriol.* **185**, 2330–2337 (2003).

Acknowledgements
This paper was inspired by a meeting entitled 'Microbial population genomics: sequence, function and diversity', which was organized by D.M., D.S. and R.R. in Siena, Italy, 17–19 January 2007. The authors thank M. Achtman, S. Bentley, C. Buchreiser, D. Caugant, L.L. Cavalli Sforza, A. Covacci, J. Dunning Hotopp, D. Falush, P. Glaser, J. Hacker, W. Hanage, D. Hood, M. Maiden, J. Telford and H. Tettelin for their contribution to the meeting and G. Corsi for artwork.

The authors declare competing financial interests: see web version for details.

DATABASES

Entrez Genome Project: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
Bacillus anthracis | *Bacillus cereus* | *Bacillus thuringiensis* | *Bacteroides fragilis* | *Bordetella bronchiseptica* | *Campylobacter jejuni* | *Candidatus* Carsonella ruddii | *Escherichia coli* CFT073 | *Escherichia coli* K12 | *Escherichia coli* O157:H7 | *Haemophilus influenzae* | *Helicobacter pylori* | *Mycobacterium tuberculosis* | *Mycoplasma pulmonis* | *Neisseria meningitidis* | *Solibacter usitatus* | *S. typhi* | *Streptococcus agalactiae* | *Streptococcus pneumoniae* | *Tropheryma whipplei*

FURTHER INFORMATION

GenBank database: http://www.ncbi.nlm.nih.gov/Genbank/index.html
Genomes OnLine Database (GOLD): http://www.genomesonline.org/
***Nature* Collection on Metagenomics:** http://www.nature.com/nature/supplements/collections/metagenomics
MLST Public Repository: http://www.mlst.net/
***Nature Reviews Microbiology* Focus on Metagenomics:** http://www.nature.com/nrmicro/focus/metagenomics/index.html
Ribosomal Database Project II: http://rdp.cme.msu.edu/
ClonalFrame homepage: http://bacteria.stats.ox.ac.uk/
454 technology: http://www.454.com
SOLiD technology: http://www.appliedbiosystems.com
SOLEXA technology: http://www.illumina.com

ALL LINKS ARE ACTIVE IN THE ONLINE PDF