APPLICATIONS OF NEXT-GENERATION SEQUENCING

Transforming clinical microbiology with bacterial genome sequencing

Xavier Didelot[1], Rory Bowden[1,2,3], Daniel J. Wilson[2,4], Tim E. A. Peto[3,4] and Derrick W. Crook[4,3]

Abstract | Whole-genome sequencing of bacteria has recently emerged as a cost-effective and convenient approach for addressing many microbiological questions. Here, we review the current status of clinical microbiology and how it has already begun to be transformed by using next-generation sequencing. We focus on three essential tasks: identifying the species of an isolate, testing its properties, such as resistance to antibiotics and virulence, and monitoring the emergence and spread of bacterial pathogens. We predict that the application of next-generation sequencing will soon be sufficiently fast, accurate and cheap to be used in routine clinical microbiology practice, where it could replace many complex current techniques with a single, more efficient workflow.

Escherichia coli
A common inhabitant of the guts of many animals, but some strains can cause serious food poisoning, as reminded by the 2011 outbreak in Germany.

Clinical microbiology is a discipline that focuses on rapidly characterizing pathogen samples to direct the management of individual infected patients (diagnostic microbiology) and to monitor the epidemiology of infectious disease (public health microbiology). Applications in epidemiology include detecting outbreaks, monitoring trends in infection and identifying the emergence of new threats. Ongoing developments in DNA-sequencing technologies are likely to affect the diagnosis and monitoring of all pathogens, including viruses, bacteria, fungi and parasites, but for this Review we focus on bacterial pathogens to demonstrate the likely changes that arise from the adoption of routine whole-genome sequencing.

Bacterial pathogens account for much of the worldwide burden of infection. For patients with bacterial infections, the crucial steps are to grow an isolate from a specimen, to identify its species, to determine its pathogenic potential and to test its susceptibility to antimicrobial drugs. Together, this information facilitates the specific and rational treatment of patients. For public health purposes, knowledge also needs to be gained about the relatedness of the pathogen to other strains of the same species to investigate transmission routes and to allow the recognition of outbreaks[1]. Each of the steps in this process of characterizing the pathogen depends on many specialized, species-specific methodologies that have been developed over decades. These require the extensive knowledge base of clinical microbiologists who apply labour-intensive, complex and often slow techniques to yield the relevant information. This multiple-step process takes from days (for the isolation by culture, species identification and susceptibility testing for rapidly growing bacteria, such as *Escherichia coli*) to months (for slow-growing bacteria, such as *Mycobacterium tuberculosis*, or to produce full typing for any pathogen) (FIG. 1).

Ideally, all of the information that is necessary for both individual treatment and public health protection would be gained in a single step. In principle, the genome sequence of an isolate contains all, or nearly all, of the information required to direct treatment and to inform public health measures. Indeed, it is becoming clear that rapid, inexpensive genome sequencing (BOX 1) holds the potential to replace many complex multifaceted procedures that are used to characterize a pathogen after it has been isolated by culture[2,3]. However, there are substantial challenges to be overcome, and success will depend on the development of the genomic knowledge and analytical methods required to extract and interpret this information correctly. Indeed, the application of new sequencing technologies will be highly disruptive, and we predict that it will take many years to transform clinical microbiology laboratories fully. Ultimately, deployment will crucially require substantial validation of genotypic prediction of the phenotype, particularly for antimicrobial resistance; this work is yet to be done. In this Review, we provide a brief overview of current practice, and then we outline the potential of sequencing technology to deliver the following key diagnostic information in the clinical laboratory after culture of an isolate: identification of

[1]*Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK.*
[2]*Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.*
[3]*NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford OX3 9DU, UK.*
[4]*Nuffield Department of Clinical Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK.*
Correspondence to D.W.C.
e-mail: derrick.crook@ndcls.ox.ac.uk

doi:10.1038/nrg3226
Published online
7 August 2012

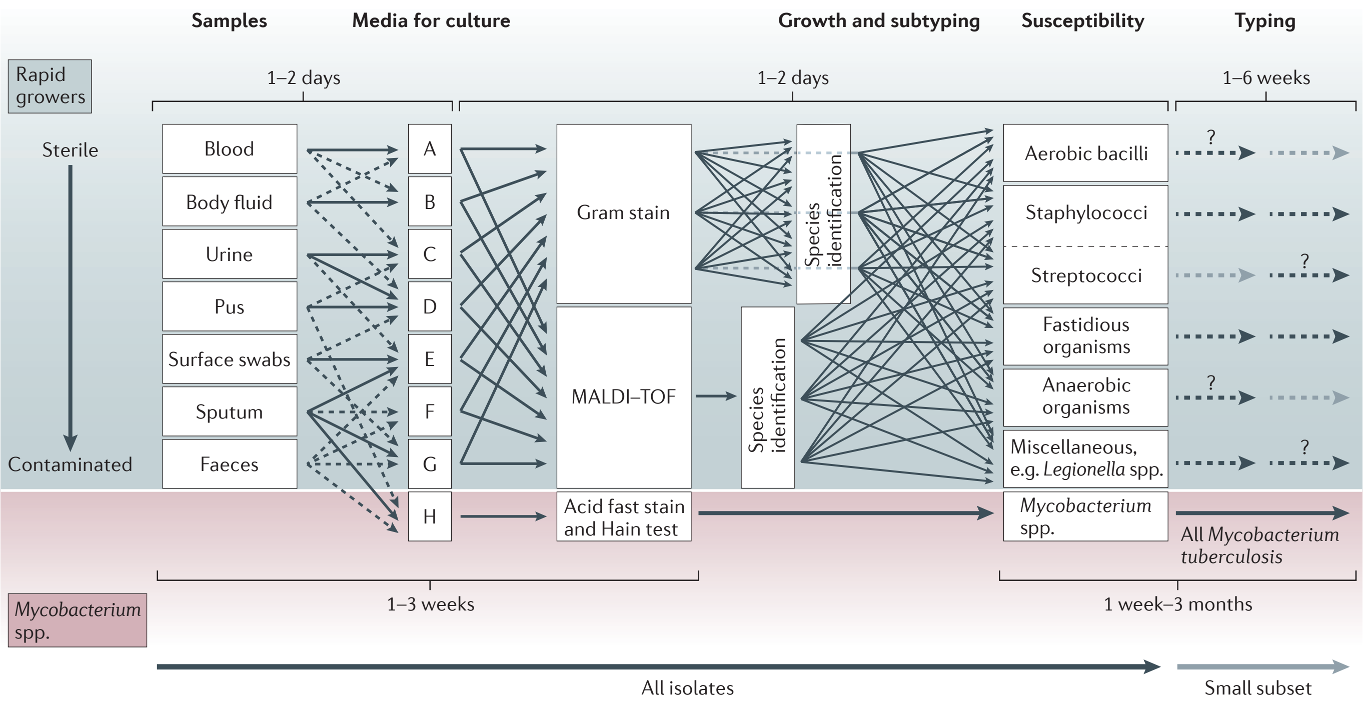


Figure 1 | **Principles of current processing of bacterial pathogens.** A schematic representation of the current workflow for processing samples for bacterial pathogens is presented, showing high complexity and a typical timescale of a few weeks to a few months. The schematic is an approximation that highlights the principal steps in the workflow; it is not intended to be a comprehensive or precise description. Samples that are likely to be normally sterile are often cultured on a rich medium that will support the growth of any culturable organism. Samples that are contaminated with colonizing flora present a challenge for growing the infecting pathogen. Many types of culture media (referred to as selective media) are used to favour the growth of the suspected pathogen; this approach is particularly important for culturing pathogens from faeces. Boxes A to H arbitrarily represent the many different media for culture. The medium H represents a medium designed for growing mycobacteria that have specific growth requirements. When an organism is growing, the morphological appearance and density of growth are properties that need specialist knowledge for deciding whether it is likely to be pathogenic. The likely pathogens are then processed through a complex pathway that has many contingencies to determine species and antimicrobial susceptibility. Broadly, there are two approaches. One approach uses matrix-assisted laser desorption/ionization–time of flight (MALDI–TOF) mass spectrometry for species identification before setting up susceptibility testing. The other uses Gram staining followed by biochemical testing to determine species; susceptibility testing is often set up simultaneously with doing biochemical tests. Categorization of pathogens into groups of species is needed to choose the appropriate susceptibility-testing panel. Finally, depending on the species and perceived likelihood of an outbreak, a small subset of isolates may be chosen for further investigation using a wide range of typing tests that are often only provided by reference laboratories. The dashed lines and question marks are positioned arbitrarily to indicate that the further investigation is varied and happens in only a small number of cases.

species, antimicrobial resistance, presence of virulence determinants, and strain typing to detect outbreaks and support surveillance.

Current clinical microbiology

The principles behind diagnostic bacteriology have changed little over the past 50 years. Most of the output from a microbiological laboratory is dependent on isolating a viable organism. More than a century of experimentation has led to the development of a wide repertoire of methods for isolating culturable bacterial pathogens. After culture, diagnostic characterization depends on a wide range of testing pathways (FIG. 1), many aspects of which are species-specific[4–6]. Complexity and a lack of automation prevent the rapid return of the complete diagnostic information about a bacterial isolate.

The cardinal steps in processing a sample are isolating a pathogen, determining the species, testing antimicrobial susceptibility and virulence and, in specific settings, intra-species typing. The first three steps are crucial for the optimal management of an infected patient, and the last step is valuable for identifying outbreaks and surveillance.

Culture of pathogen. The aim of culture is to investigate the microbial composition of a sample, to identify colonies that deserve further attention and to produce sufficient mass of pure organisms for subsequent use. Although most bacterial diseases are caused by ~20 species (TABLE 1), up to 1,000 other species may sometimes cause disease[6]. Most of these pathogens can be grown in appropriate culture media (using various methods), but a minority (<10%) of infecting bacterial pathogens are

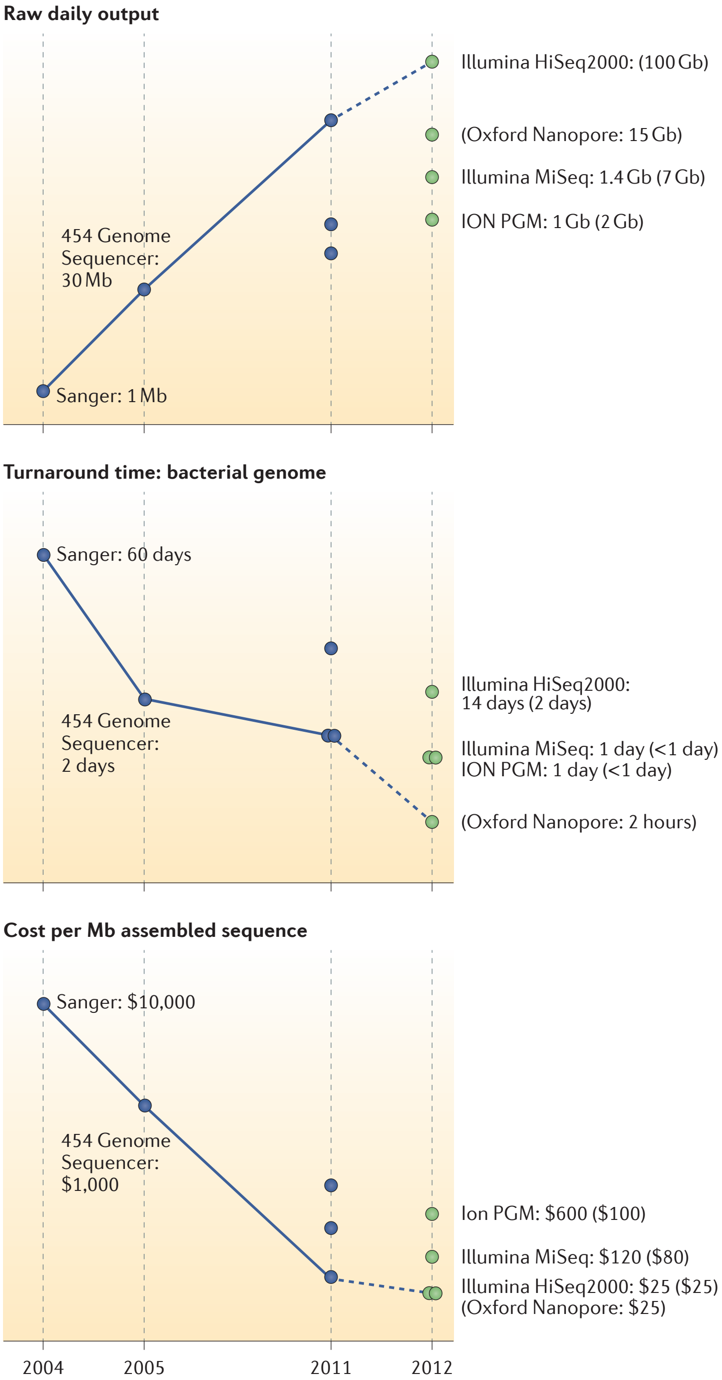
Mycobacterium tuberculosis
The causative agent of tuberculosis, it infects approximately one-third of the human population and claims over one million lives per year, making it the most deadly bacterial pathogen of humans.

Box 1 | Sequencing platforms for clinical microbiology

Released in 2005 with reads of ~110 b, the first next-generation sequencers, from Roche-454, could sequence bacterial genomes in a single run[69]. Initial applications were focused on diversity discovery. Later versions of the 454 platform have increased read length (~500 b) to approach that of Sanger sequencing but at a much lower cost and so have retained a role in producing high-contiguity assemblies of bacterial genomes.

Initially launched in 2006 with short (36 b) reads, Illumina Genome Analysers have captured the bulk of the sequencing market for both microbiology and larger organisms. With incrementally increasing capacity and read length, the current standard configuration (at the end of 2011) delivers ~300 Gb of raw data per eight-lane flow cell in the form of 100 b paired reads. Tagging each sample with its own 6–8 b index sequence allows at least 96 samples to be sequenced simultaneously in each lane. This approach makes the Illumina HiSeq platform useful and cost-effective for large bacterial sample collections.

It is clear that for most uses in microbiology, fast, compact bench-top machines will be preferred to the large, high-capacity machines designed for human sequencing. Two such platforms, the Ion PGM and the Illumina MiSeq, both of which use established chemistries that involve library preparation and amplification as the first steps in sequencing, are becoming popular among microbiologists[70]. In a new platform from Oxford Nanopore Technologies, which is slated for commercial release in 2012 (REF. 71), the sequence of a single DNA molecule passing through a protein nanopore under the control of a processive enzyme is measured as fluctuations in electrical current across a lipid membrane. According to the company, data are collected in real time at around 200–400 bases per second, and they expect up to 1,000 bases per second in the future. These data are translated to sequence information in real time using on-board electronics. The company have said that chips are configured to read 2,000 or 8,000 pores simultaneously and that reads can be up to tens of kilobases in length. Because it reads native DNA, the Oxford Nanopore technology is anticipated to work with fairly crude samples and low DNA concentrations. The company plans two machines: the scalable GridION, in which multiple sequencing units (each with a projected output of ~2 Gb of data an hour) can be combined in parallel; and the single-use, USB-connected MinION, which has a projected hourly capacity of ~150 Mb. If per-base accuracy can be improved to current next-generation standards, the long reads will enable complete genomes to be generated in minutes with either machine. This new technology is the first in a new breed of similarly designed platforms that are likely to produce dramatic improvement is sequencing technology. The figure shows the development of sequencing technologies that are relevant to microbiology, highlighting the continuing increases in throughput and speed and reductions in costs. Values in brackets and represented by green dots are projections for 2012. Cost estimates include machine depreciation over 3 years and service costs, library and consumables costs.



believed to be non-culturable or difficult to grow; for these species, diagnosis currently depends on serological, antigen and nucleic acid amplification tests.

Culture is complex and contingent on the origin of the sample. Samples from usually sterile sites (such as cerebrospinal fluid) and bacterially contaminated samples (such as faeces) represent opposite extremes. For sterile sites, a full report of all organisms present is possible, although not all organisms may be clinically relevant. For highly contaminated samples, isolation of pathogens requires selective media assisted by, for example, inspection of colony morphology and Gram staining. Educated guesses about likely pathogens alter the choice of protocol, and the growth time before further analysis can vary from hours to weeks. A full description of culture methodology is beyond the scope of this Review and is available from extensive literature: for example, a clinical microbiology textbook[5].

Table 1 | **Examples of bacterial pathogens reported by a microbiology laboratory**

Top 25 categories	Classification	Number	Percentage	Cumulative percentage
Escherichia coli	Species	222,094	29.57	29.57
Staphylococcus aureus	Species	128,158	17.06	46.63
Coagulase-negative *Staphylococcus*	Genus	58,429	7.78	54.41
Enterococcus faecalis	Species	48,121	6.41	60.81
Coliforms	Genus	40,844	5.44	66.25
Group B *Streptococcus*	Species	36,934	4.92	71.17
Anaerobes sensitive to metronidazole	Other	35,411	4.71	75.88
Pseudomonas spp.	Genus	20,588	2.74	78.62
Pseudomonas aeruginosa	Species	14,261	1.9	80.52
Proteus mirabilis	Species	13,694	1.82	82.35
Campylobacter jejuni	Species	12,775	1.7	84.05
Group A *Streptococcus*	Species	12,423	1.65	85.7
Haemophilus influenzae	Species	11,967	1.59	87.29
Group G *Streptococcus*	Species	11,188	1.49	88.78
Klebsiella spp.	Genus	9,734	1.3	90.08
Streptococcus viridans	Genus	8,990	1.2	91.28
Streptococcus pneumoniae	Species	6,564	0.87	92.15
Streptococcus constellatus	Species	4,348	0.58	92.73
Proteus spp.	Genus	4,121	0.55	93.28
Staphylococcus saprophyticus	Species	4,055	0.54	93.82
Diphtheroids	Other	4,035	0.54	94.36
Salmonella enterica	Species	3,621	0.48	94.84
Group C *Streptococcus*	Species	3,453	0.46	95.3
Propionibacterium spp.	Genus	2,805	0.37	95.67
Moraxella catarrhalis	Species	2,461	0.33	96
Others (355 categories)		30,060	3.99	100

The top 25 categories from the 15-year output of isolates by the Oxford University Hospitals Trust, UK, microbiology laboratory are shown as an example of the frequency of pathogens isolated by a large service with comprehensive diagnostic throughput. Of 751,134 isolates cultured: 557,581 (74%) were categorized into 301 species using routine phenotypic methods; 157,150 (21%) were characterized to genus or other grouping (71 categories; for example, *Pseudomonas* spp. or coagulase-negative staphylococci, respectively); 36,403 (5%) were isolated but not characterized beyond the Gram stain (not shown). On a global scale, the proportions of species may differ by country. For example *Mycobacterium tuberculosis* will be a major component of laboratory activity in communities with a high prevalence, whereas Oxford has a low incidence of tuberculosis.

Staphylococcus aureus
Found as a harmless colonizer of the skin in ~20% of the human population; it can cause life-threatening symptoms and be resistant to some antibiotics (for example, methicillin-resistant *S. aureus* (MRSA)).

Staphylococcus epidermidis
A normal part of the human skin flora; it can become pathogenic if introduced into deeper tissues following surgery.

16S ribosomal RNA gene
The 16S ribosomal RNA genes are transcribed into the 16S ribosomal RNA molecule, which is a major component of the bacterial small ribosomal subunit. The strong sequence conservation of this molecule makes it ideal for detecting large evolutionary distances between two organisms.

Species identification. Knowing the species of an isolate is often vital to make effective clinical decisions. Determining species is directly informative about pathogenic potential and allows differentiation of infecting pathogens from non-infecting 'contaminating pathogens'. A typical example is that *Staphylococcus aureus* isolated from a blood culture (rather than a skin swab sample) has a high probability of being an infecting pathogen, whereas *Staphylococcus epidermidis* is likely to be a contaminating isolate. Currently, species identification is first based on Gram staining, colony growth and morphology, rapid biochemical reactions and ancillary tests. These take up to 24 hours for organisms that require extensive biochemical panels (for example, using the Vitek system commercialized by BioMerieux or the BD Phoenix system from Beckton Dickinson). 16S ribosomal RNA gene sequencing is increasingly being used in ambiguous cases, but this has a number of drawbacks[7], including the fact that it usually takes a further 2 days[8].

Recently, matrix-assisted laser desorption/ionization–time of flight (MALDI–TOF) mass spectrometry has yielded rapid species identification by analysing the biomolecules that are present in pure suspensions of any isolate and comparing the results with known profiles[9,10]. The cost of a MALDI–TOF mass spectrometer is high (several hundred thousand US dollars), but the running cost is low[11] (around one dollar per isolate), and results can be obtained in minutes[9], so this approach is rapidly being adopted for routine use. Therefore, MALDI–TOF represents an attractive alternative to traditional methods by increasing the speed of identification of species[12], and it highlights the potential of new technology combined with sophisticated software and databases to simplify and improve an important aspect of diagnostic microbiology. However, questions remain about the level of resolution it can achieve[9,13]: for example, in distinguishing between the closely related species *Mycobacterium bovis* and *M. tuberculosis*[14]. In addition, no further information about the isolate, such as antibiotic susceptibility or virulence, is yielded by MALDI–TOF.

Testing for antibiotic resistance. Determining the antimicrobial resistance properties of an isolate is possibly the single most important procedure for managing bacterial infectious disease at the individual-patient level. This is largely because falsely recording an organism as being susceptible to an antibiotic represents a serious risk to the infected patient if they are treated with an ineffective antibiotic. Current knowledge of susceptibility testing is vast and complex and is embodied in guidelines and various textbooks[15]. The phenotypic methods for susceptibility testing are almost exclusively based on the inhibition of growth of the bacteria when exposed to the test antimicrobial.

Practice in infectious diseases is crucially dependent on confidence in this system of testing. However, the sensitivity and specificity of a particular method of testing is based on a comparison with *in vitro* 'gold-standard' susceptibility-testing systems (such as the micro-dilution method[15]), which are regarded as surrogates for clinical outcome[15]. Consequently, even with phenotypic testing, the *in vivo* (that is, clinical) susceptibility of an isolate is not known with complete certainty[15]. Clinicians have come to accept this uncertainty in clinical decision making, and indeed they are familiar with treating some pathogens without testing, as for some organisms there are no accurate tests available.

Advantageously, phenotypic testing yields information not only on those agents to which an organism is resistant but also on those agents to which it is susceptible; this is of direct clinical value. However, no single pathway for resistance and/or susceptibility testing exists. Tests are grouped by species, adding to the complexity and time taken for thorough testing. The tests are subject to many assumptions about the degree of susceptibility based on the minimum inhibitory concentration (MIC), and they require the selection of a 'breakpoint' for each antibiotic: that is, an MIC level above which the isolate is deemed to be resistant to therapy. These breakpoints are chosen on the basis of diverse but imperfect factors such as the distribution of MICs, chemical concentration, mutual interactions between the host and drug, animal models and clinical treatment experience. Consequently, there is considerable debate on how to set the breakpoints, and these are not always agreed across countries and organizations. The effect of susceptibility testing on the clinical response to infection is difficult to study, given the multiple factors that influence patient outcome, so that the sensitivity and specificity for determining resistance or susceptibility of phenotypic tests are often poorly measured. In addition, phenotypic testing has proved to be unreliable in some well-described situations. For example, the emergence of quinolone resistance in *Salmonella enterica* subsp. *enterica* serovar Typhi was not detected by routine phenotypic testing, and these isolates were falsely found to be susceptible[16]. This failure has since been found to be caused by the emergence of a new resistance mechanism, and new testing recommendations have been formulated[17]. Furthermore, complete testing can take days for rapid growers, such as *E. coli* and *S. aureus*, and even months for slow growers, such as *M. tuberculosis*.

Currently, the presence or absence of resistance genes is used in a few situations to direct early treatment of patients. For example, detection of the *mecA* gene determines whether an isolate of *Staphylococcus aureus* is methicillin-susceptible or methicillin-resistant[18], which in turn is associated with increased mortality[19]. Another example is the Hain test, which uses DNA hybridization of primers that are unique to a limited number of common resistance determinants to predict the resistance of *M. tuberculosis* isolates to a few key anti-mycobacterial drugs[20]. This has gained credibility and wide use and is a good proof-of-principle example for the future of using DNA sequence to predict resistance.

Detecting virulence determinants. Identifying virulence determinants is rarely a priority in treating individual patients. There are, however, a few examples in which this is crucial. For example, in *Corynebacterium diphtheriae* infections, detecting the presence of toxin is crucial for administering an antitoxin to the patient[4,5]. Similarly, determining whether a strain of *Clostridium difficile* is toxin producing is important in diagnosing whether *C. difficile* is pathogenic and in determining what treatment is required. Historically, most virulence determinants have been detected using bioassays (for example, the detection of the botulinum toxin) or serotyping (for example, presence of pneumococcal capsule). Increasingly, the detection of virulence factors is based on detecting the bacterial sequences that encode virulence factors using PCR (for example, for factors such as the *C. difficile* toxin B)[21]. These tests are rarely included in the repertoire of routine laboratories and are usually carried out by reference laboratories. For public health purposes, virulence determinants such as capsule type are important, particularly for species in which capsule-based vaccines are in wide use[4,5]: for example, *Haemophilus influenzae* type b, *Streptococcus pneumoniae* and *Neisseria meningitidis*.

Outbreak detection and surveillance. Pathogen surveillance and outbreak detection is mostly informal and reactive. The isolates chosen for investigating relatedness to identify outbreaks have been dependent on extemporary choices and are often based on loosely defined epidemiological criteria at the level of the routine clinical laboratory or infection control team. Consequently, many outbreaks are likely to be missed. The typing used to identify epidemic transmission can take months to complete, as most typing schemes are species-specific and depend on many variables, and only a small number of laboratories in the whole world carry out routine

Minimum inhibitory concentration
(MIC). The minimum concentration of an antibiotic that is sufficient to inhibit growth of a bacterial culture.

Clostridium difficile
A leading cause of diarrhoea and more severe conditions, especially in the elderly following disruption of the normal gut flora through the use of antibiotics.

Serotyping
In this context, the classification of bacteria on the basis of their surface antigens.

Haemophilus influenzae
Responsible for a wide range of clinical diseases (but not influenza, as originally thought and as the name might still suggest) especially in young children, it was the first free living organism to have its genome completely sequenced.

Streptococcus pneumoniae
A major cause of pneumonia, it can also cause various other severe conditions and has recently developed resistance to some antibiotics. It causes ~1 million deaths per year, mostly in children.

Neisseria meningitidis
A commensal inhabitant of the nasopharynx in up to one-quarter of the human population; it occasionally gets into the blood, resulting in >100,000 deaths per year through meningitis and septicaemia.

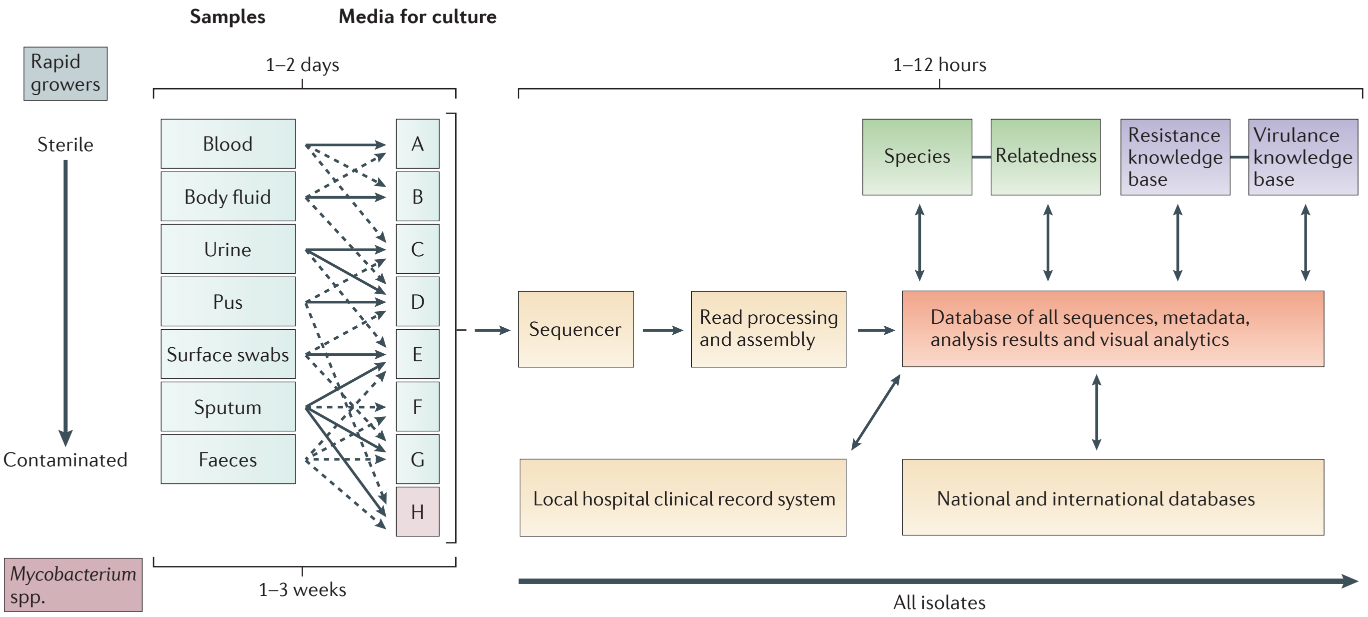


Figure 2 | **Hypothetical workflow based on whole-genome sequencing.** A schematic representation of the workflow anticipated after adoption of whole-genome sequencing is shown, with an expected timescale that could fit within a single day. The culture steps would be the same as those that are currently used in a routine microbiology laboratory. Some types of sample might be directly sequenced (see 'Future directions', not shown here). When a sample or likely pathogen is ready for sequencing, DNA will be extracted. This procedure is becoming simpler, as the input required for successful sequencing is reducing; it is now possible to use as little as 5 ng and to purify this in <30 minutes. For current bench-top machines, it can take as little as 2 hours to prepare the DNA for sequencing, and new platforms (BOX 1) could enable sequencing without preparation. Therefore, bacterial genome sequencing in hours and possibly even minutes is a realistic prospect. After sequencing, the main processes for yielding information will be computational. The development of software and databases is a major challenge to overcome before pathogen sequencing can be deployed in clinical microbiology. Automated sequence assembly algorithms will be necessary to process the raw sequence data (BOX 1). This assembled sequence would then be analysed by modular software to determine species, relationship to other isolates of the same species, antimicrobial resistance profile and virulence gene content. Results of this analysis will be reported through hospital information systems. All of the results will also be used for outbreak detection and infectious diseases surveillance. These developments will require a new large database and other informatics technology and will take time to develop. In particular, it will need 'intelligent systems', which will incorporate elements of machine learning to allow automatic updating of key knowledge bases for species identification, antimicrobial resistance determination and virulence detection. Formal evaluation of such a solution will also need robust testing to ensure that it performs at least as well as current methods.

typing. Methods that are commonly used now include PCR (for example, multiple-locus variable number of tandem repeats (MLV–VNTR) analysis)[22], restriction fragment length polymorphisms (for example, pulsed field gel electrophoresis)[23] or fractional sequencing (for example, multi-locus sequence typing (MLST))[24]. Through substantial investment in monitoring and reference facilities, turnaround time for these methods can be reduced to a few days. However, because most locations do not benefit from such facilities, typing typically contributes little to the immediate control of an outbreak.

Potential of genome sequencing

The major advantage of whole-genome sequencing is to yield all of the available DNA information content on isolates in a single rapid step following culture (sequencing without culture is discussed in the 'Future directions' section). In principle, the result contains all of the data that are currently used for diagnostic and typing needs, even though it is not always yet known how to interpret these data. However, the genome also includes vast amounts of additional data that are currently unavailable from routine processing, thus opening the prospect for large-scale research into pathogen genotype–phenotype associations from routinely collected data. The hurdles to implementing whole-genome sequencing in clinical and public health laboratories are substantial, as widespread adoption would require incorporating the knowledge from more than a century of characterizing pathogens — currently delivered by a skilled workforce — into an entirely new framework of mainly computer-driven genome processing (FIG. 2). This would require a radical shift towards a new operational paradigm for routine laboratories. In addition, a new understanding of genotype-to-phenotype relationships needs to be established, evaluated and deployed in parallel with current routine methods, which will require a major effort leading to gradual replacement of present-day methodologies over many years.

Box 2 | Phylogenetic analysis

Pathogen	Mutations per site per year	Mutations per genome per year	Refs
Staphylococcus aureus	3.0×10^{-6}	8.4	48,57
Clostridium difficile	5.3×10^{-7}	2.3	*
Mycobacterium tuberculosis	1.1×10^{-7}	0.5	87
Streptococcus pneumoniae	1.6×10^{-6}	3.5	82
Helicobacter pylori	1.9×10^{-5}	30.4	88
Vibrio cholerae	8.3×10^{-7}	3.3	37
Escherichia coli	2.26×10^{-7}	1.1	89

*X.D., unpublished observations.

To identify the species of an isolate or to investigate whether this isolate is a part of an outbreak, it is often useful to construct a phylogeny showing how the isolate is related to other strains. To do so, Bayesian phylogenetics is an attractive alternative method to classical non-statistical phylogenetic techniques[72]. The most popular software for Bayesian phylogenetic inference is BEAST[73]. A key advantage of the Bayesian method is that assumptions are made explicitly and can be relaxed or tested. Many such extensions are implemented in BEAST to account for, for example, differences in sampling dates[74], non-constant population sizes[75], non-constant molecular clocks[76] and geographical origins of the individuals[77]. Bayesian phylogenetic methods can be slow for genome-scale data, but parallel computing approaches can help with this issue[78].

Recombination in bacteria may be frequent, may occur at rates that vary among lineages and may have effects on sequence diversification (these effects may often be larger than those of mutation)[72,79]. Ignoring the effect of recombination can therefore impair phylogenetic reconstruction[80]. Furthermore, understanding the recombination process itself is often informative about ecological[31] or pathological[81] properties of bacterial lineages. An intuitive approach is to detect recombinant fragments and to account for them during phylogenetic reconstruction[82]. It is possible to do this formally by expanding Bayesian phylogenetic methods to include a model of recombination, as implemented in, for example, ClonalFrame[83] and ClonalOrigin[84].

A phylogenetic tree is not a direct reflection of transmission events[85], but it can still be informative about the way they occurred[86]. In this context, an important first step is to estimate the molecular clock (the rate of molecular substitution) to re-scale the tree in units of time. Such a clock rate can be estimated from longitudinal samples from a single infected individual[87,88]; it can be jointly estimated with the phylogeny in BEAST[57], or it can be estimated from the reconstructed tree by exploiting the correlation between tree root-to-tip distances and year of isolation[37,82]. Such estimates are only reliable if the range of sampling dates is substantial (typically at least 10 years) compared to the time to the most recent common ancestor. The table contains estimates of molecular clock rates for various bacterial pathogens. Although these vary substantially, they are all of the order of one mutation per year per genome. After a molecular clock has been estimated, the common ancestors on the phylogenetic tree can be dated, so that epidemiological interpretations of microevolution become possible, and these are in turn informative about patterns of transmission at a larger scale.

Campylobacter jejuni
A natural colonizer of the digestive tracts of many birds and cattle; it is typically transmitted to humans by ingestion of contaminated food and results in severe diarrhoeal diseases.

Crucially, the translation of sequence technology into new practices in clinical microbiology is facilitated by genetic features of bacteria. Compared with eukaryotic genomes, bacterial genomes are much smaller (2–6 Mb), and bacteria usually possess a single haploid chromosome (although a few possess two haploid chromosomes). However, they are much more diverse than eukaryotic species, partly because ~10–40% of the genome may consist of dispensable sequences that are not shared in all members of the same species[25]. Many of these dispensable elements are also mobile: for example, episomal structures such as plasmids. The plasmids and other mobile elements often encode antibiotic resistance and even virulence determinants, and as such they are highly relevant to clinical microbiology.

Species identification. As highlighted above, the identification of species is a crucial initial step in managing infectious diseases and tracking pathogens. Currently, taxonomic approaches are based on keeping a type strain collection as a gold standard (with the exception of MALDI–TOF, which can use a set of references for each species). Using whole-genome sequencing, this could be replaced by a 'type sequence'. That is, species would be taxonomically defined by their sequence, and the 'type sequence' would constitute a reference point against which to compare sequence data from other isolates. The relationship of the species to all previously sequenced organisms can be determined using phylogenetic analysis (BOX 2).

A ribosomal MLST (rMLST) scheme has recently been proposed[26] that relies on the sequences of 53 genes encoding ribosomal proteins, which are present in all bacteria. Acquiring the sequences of such a large number of genes is best done by first sequencing the whole genome and then extracting individual genes using, for example, BLAST[27]. The BIGSdb database system is an integrated platform that enables users to find many genes in many genomes using BLAST and to record the results for future use[28]. More than 1,900 bacterial genomes from 452 bacterial genera have been analysed using the rMLST scheme[26]. Any newly sequenced genome can easily be added to the database, can have its ribosomal genes extracted and can have its phylogenetic relationships with other genomes assessed. In a separate effort, a new method has recently been developed that allows the automatic *in silico* application to any genome sequence of the MLST schemes of 66 distinct species based on hundreds of genes, thus potentially revealing both the species to which the genome belongs and its sequence type within the relevant MLST scheme[29].

With further development, these comparative approaches could reach the level required to replicate current species identification procedures with high precision. As this is progressively being achieved, our definitions of bacterial species will probably need refining to reflect new accumulated knowledge based on sequence comparison. Indeed, it has already been shown that sequence data, even at the level of fractional sequencing (for example, MLST), is robust at differentiating *Streptococcus pneumoniae* or *Campylobacter jejuni* from closely related species[30,31]. However, it has also revealed that some named species do not represent monophyletic units of diversity: for example, in the case of *Bacillus cereus* and *Bacillus thuringiensis*[32]. Although the increased statistical power of having the whole genomic sequence data considerably improves the precision of such analysis for differentiating all species, it will probably also reveal more ambiguity at the boundaries of currently defined species than has already been recognized from fractional sequencing[31,33]. Such findings are likely to give impetus to a reconsideration of the notion of bacterial species, eventually leading

Box 3 | Assembly and alignment techniques

High-throughput sequencing techniques produce many short (30 to 100 bp) overlapping reads from the target genome. The first task of any analysis is therefore to assemble these reads into larger parts of the genome[90]. A first approach to do so is called 'reference based assembly' and consists of comparing the reads to a previously sequenced 'reference genome' to determine where they fit. Maq[91] and STAMPY[92] are two popular software packages that can carry out such assembly. After reads have been mapped to the reference genome, positions that differ can be found using, for example, SAMtool[93]. A first obvious drawback of this approach is that any element that is absent from the reference genome will not be assembled. A second difficulty is that the ability to map reads accurately to the reference genome decreases with the genetic distance between the target and reference genomes. A closely related reference genome is therefore needed to assemble the target genome accurately. Furthermore, when several genomes are assembled using the same reference, the genomes that are more closely related to the reference will be better assembled, which can introduce substantial biases in downstream analysis.

For these reasons, there is growing interest in assembling genomes in a reference-free manner, a task that is often called '*de novo* assembly' and carried out by, for example, the programs Newbler[94] or Velvet[95]. With new high-throughput-sequencing techniques that can produce longer reads (for example, the Pacific Biosciences platform and, in the future, the Oxford Nanopore platform), individual reads contain larger overlapping regions so that it is easier to see how they fit with each other along the target genome. *De novo* assemblies do not have the two difficulties described above for reference-based assembly: the whole of the genome is assembled, and the quality of the assembly does not depend on the choice of reference. *De novo* assembly, however, suffers from the fact that it results in tens or hundreds of contigs that represent different segments of the genome. Further assembly of these contigs into a complete genome is typically made impossible by the presence of repetitive elements, for which reads from separate elements can have high levels of homology.

When genomes are assembled *de novo*, they need to be aligned before they can be compared. Alignment of bacterial *de novo* genome assemblies is complicated by rearrangements that have destroyed the colinearity of the genomes[96]. The computer package Mauve has been designed to align whole genomes, accounting for rearrangements[97,98]. However, it is limited in the number of genomes that it can align simultaneously (in our experience, up to 20–40 genomes, depending on their diversity). A solution is to align the genomes in a pairwise fashion to a reference, but this raises the same difficulties as described above for reference-based assembly.

An alternative that is useful for most practical purposes is to take a gene-by-gene approach. For example, genes can be retrieved from genomes using BLAST[27]. This gene-querying approach is useful when genes of interest are known in advance: for example, when carrying out species identification using 16S ribosomal DNA[99] or when assessing the presence of known genetic markers of resistance[100] or virulence[101]. A full description of genetic content of the genomes may, however, require automatic annotation. This can be performed by Glimmer[102] or one of the several pipelines based on this program, such as xBASE[103] or DIYA[104].

Vibrio cholerae
The agent of cholera is transmitted via contaminated waters and can cause death through dehydration. It caused millions of deaths in Europe in the ninteenth century but has since mostly disappeared from industrialized countries. It still claims > 100,000 lives per year in developing countries.

Salmonella enterica subsp. *enterica* serovar Typhi
All *Salmonella* cause disease, but the Typhi lineage is the main causative agent of typhoid fever, which claims hundreds of thousands of lives per annum.

to great simplification and clarity to the early steps in diagnostic clinical microbiology. For example, a genomic criterion for species definition has been proposed whereby two isolates belong to the same species if their average nucleotide identity is at least 95%[34], and this was shown closely to replicate current definitions based on DNA–DNA hybridization tests[35].

Several challenges remain to be overcome before routine species identification by whole-genome sequencing can become a reality for most pathogens. This includes achieving a turnaround time approaching hours for sequencing and analysing the isolate data. This will depend on new rapid sequencing (BOX 1), new assembly techniques (BOX 3), new phylogenetic techniques (BOX 2) and developing software and databases that are able to store large numbers of genomes (FIG. 2). Software packages will need to be user-friendly and will need to yield clinically meaningful results. Quality-control procedures will need to be developed as well as criteria for run success, software validation and proficiency testing for laboratories. Before its deployment as a diagnostic system, a detailed clinical evaluation will be needed, including a comparison with currently used methods.

Testing for antibiotic resistance. In principle, it should be possible to predict resistance phenotypes by identifying genetic determinants of antimicrobial resistance and thus to permit rapid antibiotic treatment decision making. Currently, there are a few examples (including from *S. aureus*[36], *Vibrio cholerae*[37] and *Burkholderia dolosa*[38]) in which genetic determinants of antimicrobial resistance identified from whole-genome data are consistent with recorded variation in phenotype. These early data suggest that a sequence-based approach holds substantial promise. Indeed, a few methods for predicting antibiotic resistance from genetic rather than phenotypic data are already widely used: for example, the detection by PCR of mecA, which confers methicillin resistance in *S. aureus*[39], and sequences that are known to encode resistance to isoniazid, rifampicin, ethambutol, aminoglycosides, capreomycin and fluoroqinolones in *M. tuberculosis* (known as the Genotype MTBDR assay[40]). In principle, whole-genome data could improve these tests, as computational querying of the sequence may be more sensitive than using PCR primers, and it would be easier to search for more determinants.

Several challenges need to be overcome to achieve clinical adoption of whole-genome sequencing in resistance prediction. First, a comprehensive set of genetic determinants of antimicrobial resistance would need to be identified for each species. Such genetic determinants include: the presence of genes that confer resistance (such as TEM β-lactamase[41]); point mutations in essential genes (such as in *rpoB*, which confers rifampicin resistance[42]); and changes in the expression of genes (for example, reversion in the mutant operator sequences of *E. coli ampC*, leading to an increase in β-lactamase expression[43]). Importantly, even where resistance determinants are well characterized, others may be revealed by further research[44]. Furthermore, new mechanisms of antimicrobial resistance arise all too frequently: recent examples include quinolone resistance in *Salmonella typhi*[16], New Delhi metallo-β-lactamase-1 in Enterobacteriaciae[45] and multi-resistance in *Neisseria gonorrhoeae*[46]. Therefore, compiling a list of genetic determinants of resistance would be an ongoing task.

The sequence details of these determinants would need to be incorporated into a database that is kept up-to-date (to include novel resistance determinants) and that allows international data exchange via, for example, the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, USA, and the European Centre for Disease Prevention and Control (ECDC) in Stockholm, Sweden. Such a database would also facilitate identifying and reporting trends in resistance and new acquisition of resistance genes from other species. Predictions about the resistance and susceptibility from sequence data need to be accurate: falsely inferring susceptibility where the isolate is resistant represents a substantial risk to the patient. Therefore, performance needs to be established to high degrees of confidence in robust and well-powered clinical studies before deployment in a regulated environment. For example, in the United Kingdom, this would require Clinical Pathology Accreditation, and in the United States this would require approval from the Federal Food and Drug Administration.

Therefore, although sequence data have the potential to support fast and cheap identification of resistance, we envisage a two-pronged approach that combines ongoing comparison of clinical outcome data with genetic data and phenotypic resistance screening. For example, ongoing phenotypic testing will be needed to identify new resistance and to keep the proposed database up-to-date.

Detecting virulence determinants. The genetic basis of many recognized virulence phenotypes is known, and yet our understanding of virulence factors is incomplete. The genome sequence of an isolate could yield information on all of the known virulence factors in one step and could create the opportunity for the discovery of new virulence factors through association studies that link the isolate genomic data with patient disease manifestation and outcome data. One early example of a finding from such an association study was the discovery of a prophage associated with whether *Neisseria meningitidis* causes meningitis[47]. Another example is the finding that non-synonymous mutations in specific genes in *S. aureus* occurred just before the development of invasive disease[48].

More recently, whole-genome sequencing of isolates from major outbreaks has demonstrated the potential for identifying recognized virulence genes and pathogenicity gene clusters and for providing new understanding of virulence factors. For example, the recent analysis of whole-genome data from *E. coli* O104 (REFS 49,50) showed the speed and precision of whole-genome sequencing. Draft sequencing took three days using the IonTorrent PGM[51], and the first assembly was released two days later[52,53]. Within a week of data becoming available, the strain was shown to be a novel *E. coli* O104:H4 variant that had acquired a prophage encoding Shiga toxin 2 and additional virulence and antibiotic resistance determinants[50]. Similarly, sequencing of isolates from the 2010 Haitian *Vibrio cholerae* outbreak was claimed to be achievable in less than a day using the PacBio system[54], and sequence analysis allowed the detection and characterization of a toxin encoded by the CTX phage[55].

Similarly to the situation for antimicrobial resistance, identifying virulence determinants from analysis of whole genomic sequences is at an early stage, and substantial challenges need to be overcome before implementing this approach in a routine service environment. In particular, it requires the development of a database that includes all known virulence determinants and can incorporate new determinants. New software is needed to analyse genome sequences for the presence and absence of known virulence determinants as well as conducting ongoing association studies as described for antimicrobial resistance. The requirement for high sensitivity is generally lower for identifying virulence factors than for antimicrobial resistance, as identifying virulence has major clinical consequences in only a few cases.

New Delhi metallo-β-lactamase-1
An enzyme that confers extensive antibiotic resistance; first characterized in 2008.

Chlamydia trachomatis
The cause of > 100 million sexually transmitted infections annually, as well as trachoma, which is an infection of the eye that can result in blindness.

Outbreak detection and surveillance. Genome sequences potentially provide a high-resolution, accurate and reproducible means for relating organisms. For example, sequencing the genomes of a diverse collection of *Chlamydia trachomatis* isolates has demonstrated the limitations of current clinical typing techniques for identifying phylogenetic relationships[56]. Compelling examples of the effectiveness of whole-genome analyses for unravelling the origins and dispersal of pathogens at regional and global scales have recently been published. This approach was used to investigate the emergence and global dispersal of ST239 isolates of methicillin-resistant *S. aureus*[57]. In another example, the emergence of serotype 19A pneumococcal capsular variants, following the introduction in the United States of a pneumococcal vaccine, was documented and its spread tracked across the country[58]. A comparative study of 154 whole genomes of *Vibrio cholerae* enabled the history of pandemic cholera over the past 50 years to be compiled, revealing that the seventh and current cholera pandemic has comprised three successive, partially overlapping waves with strong geographical and temporal structure[37]. In *Mycobacterium leprae*, genome sequencing of isolates from 50 patients and 33 wild armadillos showed that these animals represent a major source of zoonotic transmission of leprosy in the southern United States[59]. In a previous study, the spread of *M. leprae* was shown to follow human migration and historical trade routes[60]. Finally, a comparison of 17 whole genomes and SNP typing in 286 globally representative isolates established strong geographical clustering in *Yersinia pestis* that is compatible with a Chinese origin for the Black Death pandemic[61].

Early reports also strongly suggest that using sequencing to detect outbreaks that include person-to-person transmission within communities and hospitals is a major benefit to health care; this has been recently shown for *S. aureus* and *C. difficile* using rapid bench-top sequencing[62,63]. A report on using whole-genome sequencing to study a tuberculosis outbreak on Vancouver Island[64] suggested that genealogical analysis of whole genomic sequences could be a major advance for tuberculosis contact tracing compared with the current cumbersome approaches. The current approaches depend heavily on identifying transmission networks through interviews, supplemented by *M. tuberculosis*-specific MIRU–VNTR typing[65], which is less discriminatory than whole-genome sequencing. Similar observations have been reported for a subset of MRSA isolates cultured from a hospital in Thailand, suggesting that phylogenetic analysis could be used to infer local hospital transmission[57]. The previously discussed studies of *V. cholerae*[55] and shigatoxin-producing *E. coli* O104 (REFS 49,50) indicate that sequencing can also rapidly provide a clear understanding of the origins of a local outbreak.

Whole-genome sequencing is becoming the method of choice in research settings for monitoring pathogens over long time courses and wide geographical scales, as well as for identifying outbreaks. Sequence data gathered for diagnostic purposes can be accumulated for pathogen surveillance, outbreak detection and evolutionary studies. In principle, detection of an outbreak could occur as early as the first secondary case. Consequently, deployment of sequencing technology for diagnostic purposes in local laboratories would also meet the needs for surveillance, as long as the genome sequences can be linked with the epidemiological information. To be fully useful, data would have to be shared locally, nationally and internationally: new integrated approaches to store epidemiological and genomic data jointly are under development[66]. It can be expected that national reference laboratories will adopt whole-genome sequencing as a single technology for typing all pathogens — replacing many species-specific typing methods — even if this is not done in the near future in routine diagnostic laboratories. A number of agencies, including the Health Protection Agency in England, UK (which will be called Public Health England from 2013), are exploring the adoption of whole-genome sequencing, initially to supplement current methods for typing high-value pathogens with the intention of implementing this approach more widely as the preferred typing method for outbreak investigation and pathogen surveillance.

Future directions

Clinical microbiology is on the threshold of incorporating genome sequencing into routine practice. Although this Review focuses on the promise of this technology for bacterial pathogens, there is also rapid progress towards its adoption for viral, fungal and parasitic pathogen diagnostics and surveillance. The potential advantages of sequencing as a primary technology, and the requirement for robust evaluation, have been set out in this Review.

It is likely that commercial developments based on sequencing technologies will focus on steps in current processing of cultured isolates that are discrete, high-cost and high-value. An example in which adoption may occur soon is in the analysis of mycobacterial cultures. Whole-genome sequencing is likely soon to provide, at a lower cost, all of the information that is currently provided by the MTBDR assay[40] and also more details about species identification and resistance determinants. Similarly, sequencing could yield, at little additional cost, more definitive typing information than MIRU–VNTR testing. As discussed above, another setting in which adoption of whole-genome sequencing has already started is the investigation of putative outbreaks of major pathogens.

In this Review, we have focused on cases in which the pathogen has been cultured, but there is also potential for sequencing without culturing: that is, to sequence the entire DNA in a sample (for example, pus, cerebrospinal fluid or sputum). Such a metagenomics approach has been used to define the microbiomes of diverse samples and environments[67,68]. Approaches such as bioinformatically masking the human sequences then assembling pathogen genomes *de novo* or mapping reads to a reference genome from the hypothesized pathogen are likely to be useful, subject to the availability of sufficient data to overcome the low proportion of pathogen DNA in a clinical sample. In samples in which pathogen cell counts are low (such as *M. tuberculosis* that is present among many other organisms in sputum or the blood of a bacteraemic patient with 1–100 bacterial-colony-forming units per millitre), recovering complete bacterial genome sequences may depend on very cheap, fast sequencing or enhanced methods to deplete background material. New, very fast single-molecule long-read sequencing approaches (BOX 1) should make it possible to sequence at great depth and low cost.

Adopting whole-pathogen sequencing would require major changes in the organization, skill mix and infrastructure of diagnostic laboratories and would therefore be disruptive, even if the main use of sequencing were after culture of the pathogen. Areas for focus will be strengthening competence in bioinformatics and software development. Advances are required in databases, efficient software and algorithms for analysis, software that automatically updates knowledge bases and sophisticated links between pathogen genomics databases and patient clinical record systems. To ensure that the benefits are accessible to the wider community, especially where a number of providers (commercial or otherwise) are developing systems, information needs to be shared in line with agreed standards. The opportunities for global surveillance of infectious diseases are vast, but political resolve is required to enable the sharing of sequence and meta-data on a global scale.

1. Burlage, R. S. *Principles of Public Health Microbiology* (Jones & Bartlett Learning, 2012).
2. Relman, D. A. Microbial genomics and infectious diseases. *N. Engl. J. Med.* **365**, 347–357 (2011).
3. Parkhill, J. & Wren, B. W. Bacterial epidemiology and biology — lessons from genome sequencing. *Genome Biol.* **12**, 230 (2011).
4. Mandell, G. L., Bennett, J. E. & Dolin, R. *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases* (Churchill Livingstone/Elsevier, 2010).
5. Murray, P. R., Rosenthal, K. S. & Pfaller, M. A. *Medical Microbiology* (Mosby/Elsevier, 2009).
6. Warrell, D. A., Cox, T. M. & Firth, J. D. *Oxford Textbook of Medicine* (Oxford Univ. Press, 2010).
7. Janda, J. M. & Abbott, S. L. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.* **45**, 2761–2764 (2007).
8. Clarridge, J. E. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* **17**, 840–862, (2004).
9. Seng, P. *et al.* Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Clin. Infect. Dis.* **49**, 543–551 (2009).
10. van Veen, S. Q., Claas, E. C. & Kuijper, E. J. High-throughput identification of bacteria and yeast by matrix-assisted laser desorption ionization-time of flight mass spectrometry in conventional medical microbiology laboratories. *J. Clin. Microbiol.* **48**, 900–907 (2010).
11. Cherkaoui, A. *et al.* Comparison of two matrix-assisted laser desorption ionization-time of flight mass spectrometry methods with conventional phenotypic identification for routine identification of bacteria to the species level. *J. Clin. Microbiol.* **48**, 1169–1175 (2010).
12. Gaillot, O. *et al.* Cost-effectiveness of switch to matrix-assisted laser desorption ionization-time of flight mass spectrometry for routine bacterial identification. *J. Clin. Microbiol.* **49**, 4412 (2011).
13. Stevenson, L. G., Drake, S. K. & Murray, P. R. Rapid identification of bacteria in positive blood culture broths by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J. Clin. Microbiol.* **48**, 444–447 (2010).
14. Shitikov, E. *et al.* Mass spectrometry based methods for the discrimination and typing of mycobacteria. *Infect. Genet. Evol.* **12**, 838–845 (2012).
15. Lorian, V. *Antibiotics in Laboratory Medicine* (Lippincott Williams & Wilkins, 2005).
16. Wain, J. *et al.* Quinolone-resistant *Salmonella typhi* in Viet Nam: molecular basis of resistance and clinical response to treatment. *Clin. Infect. Dis.* **25**, 1404–1410 (1997).

17. Cavaco, L. M. Hasman, H., Xia, S. & Aarestrup, F. M. *qnrD*, a novel gene conferring transferable quinolone resistance in *Salmonella enterica* serovar Kentucky and *Bovismorbificans* strains of human origin. *Antimicrob. Agents Chemother.* **53**, 603–608 (2009).
18. Bode, L. G., van Wunnik, P., Vaessen, N., Savelkoul, P. H. & Smeets, L. C. Rapid detection of methicillin-resistant *Staphylococcus aureus* in screening samples by relative quantification between the *mecA* gene and the *SA442* gene. *J. Microbiol. Methods* **89**, 129–132 (2012).
19. Cosgrove, S. E. *et al.* Comparison of mortality associated with methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* bacteremia: a meta-analysis. *Clin. Infect. Dis.* **36**, 53–59 (2003).
20. Barnard, M., Albert, H., Coetzee, G., O'Brien, R. & Bosman, M. E. Rapid molecular screening for multidrug-resistant tuberculosis in a high-volume public health laboratory in South Africa. *Am. J. Respir. Crit. Care Med.* **177**, 787–792 (2008).
21. Knetsch, C. W. *et al.* Comparison of real-time PCR techniques to cytotoxigenic culture methods for diagnosing *Clostridium difficile* infection. *J. Clin. Microbiol.* **49**, 227–231 (2011).
22. Lindstedt, B. A. Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. *Electrophoresis* **26**, 2567–2582 (2005).
23. Goering, R. V. Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease. *Infect. Genet. Evol.* **10**, 866–875 (2010).
24. Maiden, M. C. Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* **60**, 561–588 (2006).
25. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594 (2005).
26. Jolley, K. A. *et al.* Ribosomal multi-locus sequence typing: universal characterisation of bacteria from domain to strain. *Microbiology* **158**, 1005–1015 (2012).
This is a database system for whole genomes that provides a smooth transition for users from working with MLST to working with genomes.
27. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
28. Jolley, K. A. & Maiden, M. C. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595 (2010).
29. Larsen, M. V. *et al.* Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.* **50**, 1355–1361 (2012).
30. Hanage, W. P. *et al.* Using multilocus sequence data to define the pneumococcus. *J. Bacteriol.* **187**, 6223–6230 (2005).
31. Sheppard, S. K., McCarthy, N. D., Falush, D. & Maiden, M. C. Convergence of *Campylobacter* species: implications for bacterial evolution. *Science* **320**, 237–239 (2008).
32. Priest, F. G., Barker, M., Baillie, L. W., Holmes, E. C. & Maiden, M. C. Population structure and evolution of the *Bacillus cereus* group. *J. Bacteriol.* **186**, 7959–7970 (2004).
33. Hanage, W. P., Fraser, C. & Spratt, B. G. Fuzzy species among recombinogenic bacteria. *BMC Biology* **3**, 6 (2005).
34. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl Acad. Sci. USA* **102**, 2567–2572 (2005).
This was the first description of computational criteria to define bacterial species on the basis of whole-genome sequencing.
35. Goris, J. *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* **57**, 81–91 (2007).
36. McAdam, P. R., Holmes, A., Templeton, K. E. & Fitzgerald, J. R. Adaptive evolution of *Staphylococcus aureus* during chronic endobronchial infection of a cystic fibrosis patient. *PLoS ONE* **6**, e24301 (2011).
37. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–465 (2011).
38. Lieberman, T. D. *et al.* Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature Genet.* **43**, 1275–1280 (2011).
39. Wolk, D. M. *et al.* Multicenter evaluation of the Cepheid Xpert methicillin-resistant *Staphylococcus aureus* (MRSA) test as a rapid screening method for detection of MRSA in nares. *J. Clin. Microbiol.* **47**, 758–764 (2009).
40. Hilleman, D. *et al.* Use of the genotype MTBDR assay for rapid detection of rifampin and isoniazid resistance in *Mycobacterium tuberculosis* complex isolates. *J. Clin. Microbiol.* **43**, 3699–3703 (2005).
41. Livermore, D. M. β-lactamases in laboratory and clinical resistance. *Clin. Microbiol. Rev.* **8**, 557–584 (1995).
42. Boehme, C. C. *et al.* Rapid molecular detection of tuberculosis and rifampin resistance. *N. Engl. J. Med.* **363**, 1005–1015 (2010).
43. Caroff, N., Espaze, E., Gautreau, D., Richet, H. & Reynaud, A. Analysis of the effects of -42 and -32 *ampC* promoter mutations in clinical isolates of *Escherichia coli* hyperproducing ampC. *J. Antimicrob. Chemother.* **45**, 783–788 (2000).
44. Devasia, R. *et al.* High proportion of fluoroquinolone-resistant *Mycobacterium tuberculosis* isolates with novel gyrase polymorphisms and a gyrA region associated with fluoroquinolone susceptibility. *J. Clin. Microbiol.* **50**, 1390–1396 (2012).
45. Walsh, T. R., Weeks, J., Livermore, D. M. & Toleman, M. A. Dissemination of NDM-1 positive bacteria in the New Delhi environment and its implications for human health: an environmental point prevalence study. *Lancet Infect. Dis.* **11**, 355–362 (2011).
46. Bolan, G. A., Sparling, P. F. & Wasserheit, J. N. The emerging threat of untreatable gonococcal infection. *N. Engl. J. Med.* **366**, 485–487 (2012).
47. Bille, E. *et al.* A chromosomally integrated bacteriophage in invasive meningococci. *J. Exp. Med.* **201**, 1905–1913 (2005).
This was the first example of an association-mapping study to determine virulence factors in *N. meningitidis*.
48. Young, B. C. *et al.* Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc. Natl Acad. Sci. USA* **109**, 4550–4555 (2012).
This was a detailed investigation of *S. aureus* within-host genomic diversification over a period of time that revealed probable evolution towards increased virulence.
49. Rohde, H. *et al.* Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *N. Engl. J. Med.* **365**, 718–724 (2011).
50. Rasko, D. A. *et al.* Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N. Engl. J. Med.* **365**, 709–717 (2011).
This was an epidemiological investigation based on whole-genome sequencing for the 2011 outbreak of *E. coli* in Germany.
51. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
52. Suerbaum, S. No tech gaps in *E. coli* outbreak. *Nature* **476**, 33 (2011).
53. Mellmann, A. *et al.* Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE* **6**, e22751 (2011).
54. Korlach, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol.* **472**, 431–455 (2010).
55. Chin, C. S. *et al.* The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364**, 33–42 (2011).
This is a study of the origin of the ongoing Haitian outbreak of *Vibrio cholerae* based on whole-genome comparison with other strains.
56. Harris, S. R. *et al.* Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. *Nature Genet.* **44**, 413–419 (2012).
This is an example of how current typing techniques can be misleading compared to whole-genome sequencing.
57. Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010).
This was one of the first studies to demonstrate the great potential of whole-genome sequencing to reconstruct person-to-person transmission pathways within a hospital.
58. Golubchik, T. *et al.* Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nature Genet.* **44**, 352–355 (2012).
This is an example of the great evolutionary potential that highly recombinogenic bacteria have in order to escape epidemiological interventions.
59. Truman, R. W. *et al.* Probable zoonotic leprosy in the southern United States. *N. Engl. J. Med.* **364**, 1626–1633 (2011).
60. Monot, M. *et al.* Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nature Genet.* **41**, 1282–1289 (2009).
61. Morelli, G. *et al.* *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genet.* **42**, 1140–1143 (2010).
62. Koser, C. U. *et al.* Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N. Engl. J. Med.* **366**, 2267–2275 (2012).
63. Eyre, D. W. *et al.* A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. *BMJ Open* **2**, e001124 (2012).
In this paper, a demonstration is provided of the usefulness of bench-top sequencing to answer epidemiological questions in near real-time.
64. Gardy, J. L. *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* **364**, 730–739 (2011).
65. Cardoso Oelemann, M. *et al.* The forest behind the tree: phylogenetic exploration of a dominant *Mycobacterium tuberculosis* strain lineage from a high tuberculosis burden country. *PLoS ONE* **6**, e18256 (2011).
66. Aanensen, D. M., Huntley, D. M., Feil, E. J., al-Own, F. & Spratt, B. G. EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection. *PLoS ONE* **4**, e6968 (2009).
67. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nature Rev. Genet.* **13**, 260–270 (2012).
68. Kuczynski, J. *et al.* Experimental and analytical tools for studying the human microbiome. *Nature Rev. Genet.* **13**, 47–58 (2012).
69. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
70. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotech.* **30**, 562 (2012).
71. Check Hayden, E. Nanopore genome sequencer makes its debut. *Nature* 17 Feb 2012 (doi:10.1038/nature.2012.10051).
72. Didelot, X. in *Bacterial Population Genetics in Infectious Disease* 37–60 (John Wiley & Sons, 2010).
73. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
74. Rodrigo, A. G. *et al.* Coalescent estimates of HIV-1 generation time *in vivo*. *Proc. Natl Acad. Sci. USA* **96**, 2187–2191 (1999).
75. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005).
76. Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
77. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, e1000520 (2009).
78. Suchard, M. A. & Rambaut, A. Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25**, 1370–1376 (2009).
79. Vos, M. & Didelot, X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* **3**, 199–208 (2009).
80. Schierup, M. H. & Hein, J. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879–891 (2000).
81. 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).
82. Croucher, N. J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).

83. Didelot, X. & Falush, D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266 (2007).
84. Didelot, X., Lawson, D., Darling, A. & Falush, D. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* **186**, 1435–1449 (2010).
85. Pybus, O. G. & Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Rev. Genet.* **10**, 540–550 (2009).
86. Cottam, E. M. *et al.* Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007. *PLoS Pathog.* **4**, e1000050 (2008).
87. Ford, C. B. *et al.* Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nature Genet.* **43**, 482–486 (2011).
88. Kennemann, L. *et al. Helicobacter pylori* genome evolution during human infection. *Proc. Natl Acad. Sci. USA* **108**, 5033–5038 (2011).
89. Reeves, P. R. *et al.* Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. *PLoS ONE* **6**, e26907 (2011).
90. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Rev. Genet.* **12**, 443–451 (2011).
91. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
92. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
93. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
94. Chaisson, M. J. & Pevzner, P. A. Short read fragment assembly of bacterial genomes. *Genome Res.* **18**, 324–330 (2008).
95. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
96. Darling, A. E., Miklos, I. & Ragan, M. A. Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.* **4**, e1000128 (2008).
97. Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
98. Darling, A. E. Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
99. Yarza, P. *et al.* The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst. Appl. Microbiol.* **31**, 241–250 (2008).
100. Liu, B. & Pop, M. ARDB—Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **37**, D443–D447 (2009).
101. Wu, H.-J., Wang, A. H. J. & Jennings, M. P. Discovery of virulence factors of pathogenic bacteria. *Curr. Opin. Chem. Biol.* **12**, 93–101 (2008).
102. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
103. Chaudhuri, R. R. *et al.* xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res.* **36**, D543–546 (2008).
104. Stewart, A. C., Osborne, B. & Read, T. D. DIYA: a bacterial annotation pipeline for any genomics lab. *Bioinformatics* **25**, 962–963 (2009).

Acknowledgements

D.W.C. and T.A.E.P. are part-funded by the UK National Institute for Health Research (NIHR) Oxford Biomedical Research Centre and are NIHR Senior Investigators. X.D., D.J.W. and R.B. are funded by the UK Clinical Reasearch Collaboration. We thank J. Paul for helpful advice and suggestions.

Competing interests statement

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

FURTHER INFORMATION

Modernising Medical Microbiology project: http://www.modmedmicro.ac.uk

***Nature Reviews Genetics* Series on Applications of next-generation sequencing:** http://www.nature.com/nrg/series/nextgeneration/index.html

***Nature Reviews Genetics* Series on Translational genetics:** http://www.nature.com/nrg/series/translational/index.html

ALL LINKS ARE ACTIVE IN THE ONLINE PDF