APPLICATIONS OF NEXT-GENERATION SEQUENCING — OPINION

Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions

W. Florian Fricke and David A. Rasko

Abstract | The potential of bacterial whole-genome sequencing (WGS) to complement existing diagnostic infrastructures in clinical microbiology has been shown in proof-of-principle examples and extensively discussed. However, less attention has been drawn to bioinformatic challenges that are associated with the clinical adoption of WGS-based molecular diagnostics. This Perspective article discusses questions that are related to standard operating procedures, computational resource management, and data storage and integration in the context of recent developments in the sequencing and bioinformatics service markets.

The transformative potential of bacterial whole-genome sequencing (WGS) for clinical diagnostics has been widely recognized in the scientific literature[1–9]. Molecular diagnostics (MDx) adopts tools from molecular biology for use in clinical diagnostics; it includes both the identification and the characterization of microorganisms on the basis of the detection and the characterization of nucleic acids. Beyond the research sector, the MDx market is a fast growing segment in the *in vitro* diagnostic space and is expected to grow at a rate of >9% in the next five years[10]. Infectious disease testing in clinical microbiology is estimated to represent 70% of the global MDx market[11]. Bacterial WGS applications have been widely discussed theoretically as a future application of clinical MDx, but such applications have only been shown in practice in few proof-of-concept studies, most of which were from the same established genome sequencing centres. BOX 1 provides an overview of the range of MDx methods that are currently in use in clinical bacterial diagnostics.

The objective of this Perspective article is to provide an overview of the specific bioinformatic challenges that need to be addressed in the transition from proof of concept to widespread clinical implementation of bacterial WGS-based MDx tests in microbial diagnostics. We briefly describe the changes in the sequencing market that are defining the landscape for clinical adoption of WGS. We focus on bioinformatic problems that are associated with creating a technically and economically sound MDx product for clinical implementation — specifically, development of bioinformatic workflows and definition of standard operating procedures, management of computational resources and selection of computational support models, and data integration and storage. Here, we focus on clinical bacteriology, as several comprehensive recent reviews have highlighted the potential for immediate application of WGS-based MDx tools in this field of research[1,8,9]. Bacterial diagnostics are an attractive application for early adoption of WGS owing to their modest sequencing and bioinformatic analysis requirements, the existing clinical microbiology infrastructure and a well-established set of validated, clinically useful diagnostic parameters.

We describe how recent developments in the markets for both sequence data generation and bioinformatic support and, most importantly, how the introduction of both benchtop sequencing and remote cloud computing have laid the basis for widespread, decentralized adoption of WGS-based MDx tools in the clinic.

Genome sequencing in diagnostics

Although the potential of WGS to complement existing clinical microbiology practice for the typing and the characterization of bacterial isolates has long been emphasized, only recently have the first studies been published to clearly demonstrate examples of clinical use, most of which are for high-resolution epidemiological investigations of bacterial pathogens (reviewed in REFS 1,8,9). TABLE 1 provides an overview of seminal publications that highlight cases of the potential use of WGS in bacterial diagnostics. In the near future, applications of WGS in the clinic are predicted to include the drawing of more accurate epidemiological outbreak maps[2,6,7,12], the deciphering of both the evolutionary history and the genetic make-up of particular outbreak isolates[2,3], and forensic assignments of biological samples in the context of biodefense or criminal investigation[4]. Recently, the use of metagenomic sequencing has been proposed for infectious disease detection[13,14].

In clinical practice, an exemplary diagnostic application of bacterial WGS would be the culture-independent *in silico* antimicrobial susceptibility testing. This technique is based on bioinformatic analyses of the genomes of either bacterial pathogens — such as *Salmonella enterica* or other enteric pathogens under routine surveillance by the US National Antimicrobial Resistance Monitoring System[15] — or bacterial isolates that are responsible for drug-resistant tuberculosis infections, such as extensively drug-resistant (XDR) *Mycobacterium tuberculosis* strains[16]. In one study, methicillin-resistant *Staphylococcus aureus* (MRSA)[12] isolates from patients and staff members at a hospital in the United Kingdom were sequenced, and single-nucleotide polymorphisms (SNPs) that were indistinguishable by traditional sequence typing between isolates were compared. This study reported a suspected MRSA transmission within the

Box 1 | **Overview of bacterial MDx in the clinic**

Molecular diagnostic (MDx) assays, or so-called '*in vitro* diagnostic devices', that are approved for clinical use can be summarized as follows, using the list of diagnostic tests that are cleared by the US Food and Drug Administration as a guideline (reviewed in REF. 41).

- Matrix-assisted laser desorption ionization–time-of-flight (MALDI–TOF) mass spectrometry has become a widely used diagnostic method that is associated with high initial investments but low operating costs. It is used in routine diagnostic practice to identify cultured bacterial isolates[42]. Other applications of this method, such as the detection of antibiotic resistance mechanisms, have also been discussed[43].
- Various types of DNA amplification are used to identify and type infectious agents of sexually transmitted diseases (such as *Chlamydia trachomatis* and *Neisseria gonorrhoeae*) and of health care-associated infections (such as *Clostridium difficile*, *Staphylococcus aureus* and methicillin-resistant *S. aureus* (MRSA), and vancomycin-resistant *Enterococcus* spp. (VRE)) or other pathogens (such as *Mycobacterium tuberculosis*, *Listeria monocytogenes*, group A and group B *Streptococcus* spp. and sepsis-causing bacteria). These tests apply singleplex or multiplex, or qualitative or quantitative (that is, real-time) PCR, as well as other types of DNA amplification (for example, transcription-mediated, strand displacement or isothermal methods).
- Nucleic acid hybridization-based MDx assays that use peptide nucleic acid fluorescence *in situ* hybridization (PNA-FISH) are used for quick identification of pathogens (such as *Escherichia coli*, *Pseudomonas aeruginosa*, *Staphylococcus* spp. and *Streptococcus* spp.), including those from material that is less suitable for PCR (for example, from positive blood cultures)[44].

Compared with most MDx assays in clinical use, bacterial WGS is currently associated with higher cost and personnel effort. It is thus unlikely to replace existing tests but might complement them in the future. Bacterial WGS provides optimal resolution for pathogen typing, including predictions of antibiotic resistance or other phenotypes, and for epidemiological analyses, all of which can be carried out *in silico* on the bacterial genome sequence as bioinformatic multiplex assays. In addition, as currently available MDx assays are still dependent on at least short bacterial cultivation, direct metagenomic sequencing has the potential to reduce clinical response times in the future.

hospital, which confirmed predictions that were previously made on the basis of conventional epidemiological analyses.

A recent contract was awarded by the US Food and Drug Administration to Illumina to place and use the MiSeq sequencing platform — a benchtop sequencer introduced by Illumina to make genome sequencing affordable even for smaller laboratories — in US state and federal laboratories to source-track foodborne enteric pathogens (that is, *S. enterica* and Shiga toxin-producing *Escherichia coli* (STEC)), which attests to the receptiveness of both national health services and public health laboratories to adopt WGS-based MDx tools[17]. This might prove as a starter for their clinical adoption. However, similar cases have not yet been seen in other countries. Julian Parkhill, a co-author of several recent benchmark studies on the use of WGS-based epidemiological tracking of hospital pathogens[2,12,16,18–21], pointed out that "the average clinician in a hospital is not going to be able to do this [WGS analysis], so it has to be automated" (REF. 22).

The economic feasibility of bacterial WGS-based MDx tests in the clinic will mostly depend on the cost and the ease of both sequence generation and bioinformatic sequence analyses. Extensive benchmarking will be required to determine expenses and personnel efforts that are required for sample processing, data analysis and data storage. To this end, recent studies have attempted to compare sequencing platforms[23] and to calculate costs of bioinformatic analyses using 'canned' analysis pipelines in combination with commercial cloud computing services[24,25]. Bioinformatic challenges are possible reasons that WGS has not yet penetrated deeply into the clinical bacterial diagnostic market. These challenges result from the lack of both bioinformatic standards and infrastructure to adequately meet demands on data storage, as well as on the analyses of rapidly changing and increasing amounts of sequence data from multiple next-generation sequencing platforms[26]. BOX 2 gives an overview of the specific bioinformatic challenges and the solutions that characterize the transition of bacterial WGS-based MDx from proof of concept to clinical implementation.

The changing sequencing market

Next-generation sequencing technologies continue to change the genomic field, which results in several options for sequence data generation. These have been reviewed in the context of bacterial WGS-based clinical application[27,28]. Available sequencing platforms vary with respect to required financial investments for initial installation (that is, infrastructure for sample processing, sequencing and sequence analyses) and subsequent usage (that is, the cost per sequencing run or per generated base); the time and effort that are associated with sample preparation and sequencing itself; the degree of automation and personnel training that is required for sample processing; and the specific characteristics of the generated data, such as sequence read length and accuracy, error profile, number of generated reads per run and multiplexing capacity. As a consequence, available sequencing platforms can be more or less suitable for specific genomic applications or for desired production scales, and thus for clinical application in MDx[27,28]. Most recent studies on the use of WGS for bacterial diagnostics applied high-throughput short-read sequencing (TABLE 1), as the accuracy that is afforded by high genome coverage proved to be more important for the applied tests, including the detection of SNPs or of gene presence or absence, than read length. With the clinical market in mind, both Illumina and Life Technologies — the two leading developers of sequencing platforms in this field — have made great advances to simplify the workflow of sequence library preparation and to reduce sequencing runtimes.

Generally, the sequencing market has been following two major trends over the past few years. First, sequencing platforms continue to be optimized for ever-increasing output and lower cost, which are measured in total numbers of base-pair output per sequencing run and in the associated cost per base pair, respectively. As a downside, installation of these increased output platforms often requires substantial financial investments and offers limited flexibility to scale sequencing outputs down to the smaller read numbers that are required for bacterial genomes compared with that of human genomes. High-throughput platforms can thus require more samples to be sequenced in parallel in order to provide economic advantage over smaller sequencing platforms. For example, a full run of the Illumina HiSeq platform can generate more than 200-fold sequencing coverage for more than 250 *E. coli* genomes. This sets a sample number threshold that may be difficult to amass on a daily or weekly basis by a single clinical laboratory that depends on the timely delivery of diagnostic results, thus decreasing the use of these platforms for local installation.

Table 1 | **Recent seminal publications on bacterial diagnostic applications of WGS**

Organism	Study	Methodology	Major findings	Refs
Clostridium difficile	A global collection of *C. difficile* isolates, primarily from patients in hospitals	WGS (Illumina GAIIx and HiSeq), read mapping (BWA) and SNP identification (SAMtools), and phylogenetic tree prediction (SplitsTree4, PhyML and BEAST)	Two lineages of *C. difficile* were found, as opposed to a single outbreak lineage that was previously expected	19
Escherichia coli	An outbreak of Shiga toxin-positive, enteroaggregative *E. coli* in Germany and some parts of Europe	WGS (454 GS FLX, Illumina HiSeq, Ion Torrent PGM and Pacific Biosciences RS II), *de novo* assembly and annotation	Novel virulence factor assemblages were identified	3, 45–47
	Outbreak of enterohemorrhagic *E. coli* (EHEC) on a farm in the United Kingdom	WGS (454 GS FLX and Illumina GAIIx), *de novo* assembly (Velvet and Newbler), read mapping to best assembly (Bowtie) and SNP identification (SAMtools)	Distinct EHEC lineages that infected both animals and humans were characterized, each of which potentially had a role in the dissemination of the pathogen	21
Klebsiella pneumoniae	Outbreak of carbapenem-resistant *K. pneumoniae* in a hospital in the United States	WGS (454 GS FLX), *de novo* assembly (Newbler) and annotation (NCBI PGAAP), alignment and SNP identification (Mauve)	A rapid hospital spread of antibiotic-resistant *K. pneumoniae* was detected by integrating clinical and WGS data for transmission studies	5
Legionella pneumophila	A collection of clinical and environmental *Legionella* spp. isolates in the United Kingdom	WGS (Illumina MiSeq), read mapping (SMALT) and SNP identification (SSAHA), and phylogenetic tree prediction (RAxML)	*L. pneumophila* outbreak isolates were distinguished from non-outbreak isolates	20
Methicillin-resistant *Staphylococcus aureus* (MRSA)	An outbreak in a neonatal intensive care unit in the United Kingdom	WGS (Illumina MiSeq), read mapping to reference (SMALT) and SNP identification (SSAHA), and phylogenetic tree prediction (RAxML)	MRSA outbreak isolates were identified in the neonatal care unit, which could be linked to isolates that were identified in the community	2
	Two independent human cases that were linked to livestock reservoirs on farms in Denmark	WGS (Illumina HiSeq), read mapping (SMALT) and SNP identification (SSAHA), phylogenetic tree prediction (RAxML), *de novo* assembly (Velvet) and alignment (Mauve and Artemis comparison tool)	Zoonotic transmission of MRSA isolates that had novel resistance markers from animals to humans was detected	18
Mycobacterium tuberculosis	An outbreak in British Columbia, Canada	WGS (Illumina GAII), read mapping to reference and SNP identification (SSAHA), and phylogenetic tree prediction (ClustalX and GARLI)	Using a combination of WGS and social networking, an outbreak of drug-resistant *M. tuberculosis* was tracked in underserved populations	6
	A single patient with extensively drug-resistant (XDR) tuberculosis	Direct DNA isolation from a three-day culture, WGS (Illumina MiSeq), read mapping (SMALT) and SNP identification (BCFtools and SAMtools), and phylogenetic tree prediction (RAxML)	A mixed infection that was undetected by standard genotyping was identified; antibiotic resistance prediction was consistent with phenotypic analysis	16
Bacterial communities	A mummified body (>200 years) of a patient with tuberculosis	Direct metagenomic DNA extraction from lung tissue, metagenomic sequencing (Illumina MiSeq), *de novo* assembly (CLC workbench) and read mapping to reference (BWA)	High coverage (>30 fold) of *M. tuberculosis* genome was generated; a mixed infection was identified	14
	Fecal specimens from patients with diarrhoea in Germany	Metagenomic DNA isolation and sequencing (Illumina MiSeq and HiSeq), *de novo* assembly using random sequence subsets (Ray Meta), read mapping against assembly (BWA) and phylogenetic profiling (Bowtie and Metaphlan)	A draft genome of a Shiga toxin-positive *E. coli* outbreak strain was assembled from fecal metagenomic data in a sample subset; suspected diarrhoeal disease agents in other samples (*C. difficile*, *Campylobacter jejuni* and *Salmonella enterica*) were detected	13

BCF, binary call format; BEAST, Bayesian evolutionary analysis by sampling trees; BWA, Burrows–Wheeler aligner; GAIIx, Genome Analyzer IIx; GARLI, genetic algorithm for rapid likelihood inference; GS FLX, genome sequencer FLX; NCBI, National Center for Biotechnology Information; PGAAP, prokaryotic genomes automatic annotation pipeline; PGM, personal genome machine; PhyML, phylogenetic estimation using maximum likelihood; RAxML, randomized axelerated maximum likelihood; SAM, sequence alignment/map; SNP, single-nucleotide polymorphism; SSAHA, sequence search and alignment by hashing algorithm; WGS, whole-genome sequencing.

Second, benchtop sequencers have recently been introduced to the market. Compared with larger sequencing platforms, these sequencers require smaller capital investments for both installation and infrastructure, and they promise fast and simple sequence generation in a standard laboratory environment. These platforms could be economically feasible for the small- to mid-size health care setting, in spite of higher operating costs per individual sequenced genome. Correspondingly, benchtop sequencing has been proposed for *in situ* implementation in the clinical setting, and the use of these platforms for WGS-based bacterial MDx applications has been successfully examined and validated[7,27].

It is important to note that the next-generation-sequencing market could change markedly, as the field is expecting the introduction of nanopore technologies for fast and affordable long-read sequencing from low-input samples[29]. WGS-based diagnostic protocols will most probably have to be adapted, as the bioinformatic field is evolving to accommodate changing sequence

data types. However, the required modifications to bioinformatic protocols for bacterial WGS-based MDx might be less extreme than expected, considering that several of the first bioinformatic sequence analysis tools, such as the basic local alignment search tool (BLAST)[30], continue to be widely used today, more than 20 years after their publication. Moreover, in some aspects, the field is moving back to its roots, as bacterial genome sequencing started with the generation of fairly long (>900 bp) Sanger sequence reads.

In the academic setting, the bacterial genomics field is already experiencing increased decentralization owing to the introduction of benchtop sequencers, with even smaller laboratories successfully deploying next-generation platforms[31]. These decentralized sequencing operations provide a paradigm for future widespread application of bacterial WGS in the clinic. Today, small to mid-size hospitals often use central commercial services for routine bacterial diagnostic tests, whereas larger hospitals maintain microbiology laboratories on site. With this infrastructure in place, both the logistics and the financial requirements would be modest for commercial diagnostic services or for hospital-integrated microbiology laboratories to venture into the bacterial WGS-based MDx service space (FIG. 1). For the integration of bacterial WGS efforts into existing clinical microbiology practice, different models are conceivable, depending on the requirements for sample throughput, the control over both data and analysis parameters, and the integration with additional research activities (BOX 3). Decentralized diagnostic networks that are in close association with health care providers offer general advantages, as they can stimulate research collaborations, shorten reaction times to outbreak scenarios and might ultimately lead to better quality of care. The specific bioinformatic challenges to developing and supporting market-ready bacterial WGS-based MDx products are outlined below.

Box 2 | **From proof of concept to clinical implementation of WGS-based MDx**

Requirements for the implementation of whole-genome sequencing (WGS)-based molecular diagnostics (MDx) in the clinic are detailed below.

Standard operating procedures

- Standardization: select and assemble workflow from a selection of available bioinformatic methods, programs and configurable options
- Optimization: increase efficiency for specific application (or applications) but not for overall functionality
- Validation: prove robustness and reproducibility of diagnostic results using real-life data and a 'locked-down' analysis workflow in independent blinded validation
- Automation: reduce configurable options and combine analysis steps into a single automated pipeline to increase both user-friendliness and potential for high-throughput application, as well as to reduce training requirements

Computational support model

- Software: replace the bioinformatic workbench model that is optimized for functional flexibility with a pipeline model of pre-selected, automated workflow that has reduced options for user configuration
- Hardware: reduce local hardware requirements and take advantage of on-demand online cloud resources for computationally demanding processes and workloads
- Computation: afford decentralized diagnostic operations with seamless support of online cloud resources for computation and reference data management

Data management

- Reference data: create and maintain a centralized, expandable and up-to-date reference database for all diagnostic tests
- Patient information: secure sensitive patient data during data storage, transfer, processing and analyses, for example, by complying with regulations of the US Health Insurance Portability and Accountability Act
- Sequence data: provide open access to de-identified genome sequence data output from MDx tests

Commercial product

- Product type: offer diagnostic product as bioinformatic service with or without hardware and/or software support
- Certification: obtain required certifications depending on product type, for example, US Clinical Laboratory Improvement Amendment (CLIA) or 'post-CLIA bioinformatics services'

Bioinformatic challenges

Standard operating procedures. In the academic research setting, the analysis of sequence data often involves an iterative process of testing, evaluating and optimizing specific steps in the analysis, which can include the application of multiple bioinformatic methods, tools and parameters. This optimization is less driven by economic factors — such as simplicity, reproducibility and efficiency of the analysis — all of which are key concerns for clinical implementation, but it is instead driven by the completeness and the accuracy of results, as well as by the conformity of the analysis with community-accepted standards. However, clinical MDx applications require definition of a robust bioinformatic sequence analysis workflow in order to ensure standardization, validation and automation, which helps to reduce both costs and times of such analyses.

Standardization of entire analysis protocols using a defined set of bioinformatic tools and analysis parameters guarantees reproducible diagnostic results. This reproducibility allows validation of diagnostic results as part of the developmental process of the clinical MDx product. Such validation should include large blinded clinical cohort studies. Parameters will need to be defined and validated for bacterial WGS-based MDx applications to associate specific genetic features with phenotypes. For example, standardization requires the definition of set thresholds to identify the presence of a genetic feature in the WGS data set; that is, how many individual sequence reads need to match a reference gene or locus, what should be the minimal required sequence identity between these reads and the reference, and how much of the reference locus needs to be covered by the matching sequence reads. The US Institute of Medicine formulated guidelines for sound scientific practice for the validation of so-called 'omics-based' tests, which included independent validation of the robustness of the test using a 'locked down', or frozen, computational model that cannot be changed during the validation process[32].

The automation of defined complex bioinformatic workflows that involve multiple individual steps facilitates analysis optimization, affords high-throughput data processing and reduces user training requirements. Most bioinformatic support systems that are available for bacterial sequence analyses rely

on 'workbench' models that allow users to choose from a variety of analysis procedures and tools, which provides maximum support for customized analyses. However, such flexibility is unnecessary in the clinical setting, in which the implementation of a standardized MDx test that runs with minimal configurable options will be more desirable to support their widespread use by clinical personnel without bioinformatic training. Ideally, such an MDx product would directly link to the raw bacterial WGS data that come from the sequencer and, in a fully automated way, generate a clear, concise human-readable diagnostic report, as well as an electronic output in a format that can be integrated with existing hospital informatics systems.

Data processing and computational resource management. Reasonably fast and computationally inexpensive bioinformatic workflow protocols have been effectively used in recent bacterial WGS-based MDx studies to identify specific genetic elements of interest (such as antibiotic resistance or virulence genes) or to determine the evolutionary position of a genome in relation to a set of references (that is, to draw phylogenetic trees). Short-read mapping tools that use the Burrows–Wheeler transform can align an entire bacterial WGS data set to a reference genome in less than an hour even with modest hardware support; for example, ~2.5 million *E. coli* reads can be processed on the credit card-sized single-board Raspberry Pi computer[33]. Mapping results can be parsed to infer information about the presence or absence of genes and to identify SNPs. The bioinformatic challenge for this type of bacterial WGS-based analysis includes the ability to provide scalable and elastic computational support to simultaneously process hundreds or more genomes at any time.

In the academic field, computational support for bioinformatic sequence analysis is typically provided using either local hardware or online computational resources. Local support can be provided through individual desktop computers or through local server networks. For example, two commercial providers of workbench programs for bioinformatic sequence analyses — CLC bio and Geneious — offer software for local installation either on a local desktop or on a server for additional computational support. Apart from the advantages that automated analysis pipelines provide over workbench models that force users to select and configure analysis parameters, the use of locally installed systems for bacterial genome sequence analyses requires substantial hardware support to allow parallel processing of typical clinical sample loads.

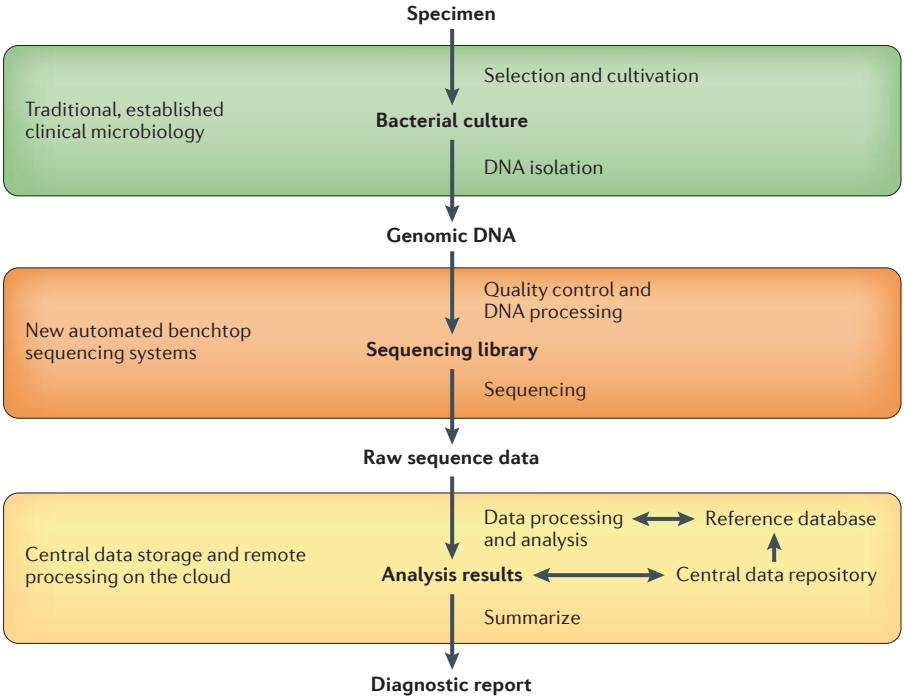

Figure 1 | **Workflow for bacterial WGS-based MDx.** A summary of the proposed workflow is shown for bacterial whole-genome sequencing (WGS)-based molecular diagnostics (MDx), which uses benchtop sequencing for decentralized sequence generation, as well as the cloud for both central data storage and remote data processing. Double-headed arrows indicate that the central data repository and reference database are constantly updated on the basis of new analysis results.

Online resources for academic use have been provided either as fixed workbenches that are pre-installed on a remote server or as Infrastructure as a Service that requires users to install software on a cloud server before using its resources. Compared with locally installed computational support systems, central resource management provided by cloud services tends to afford economies of scale that result in greater data storage and processing capacities at lower prices. Non-commercial sequence analysis software that is provided as an online service includes the versatile Galaxy platform[34] and the Rapid Annotation using Subsystem Technology (RAST) server for bacterial genome annotation[35]. Users typically access these platforms with a standard web browser and take advantage of both the hardware resources that are available to them and the software that is pre-installed on the server. These systems do not require users to install software locally and provide a lot of flexibility to customize analyses; they are therefore popular in the academic community. However, clinical MDx applications are likely to be better served by combining online computational resources with streamlined and automated 'canned' analysis software systems.

The flexibility of the pay-as-you-go support model of the cloud could be an example for the decentralized implementation of highly scalable bioinformatic support for bacterial MDx. Computational resources on the cloud can accommodate even substantial analysis loads, including multiple tasks that are carried out in parallel. The elasticity of the cloud to theoretically provide these resources on-demand at any given time makes it possible that even bursts of extreme computational demand can be met on short notice.

For cloud-based bioinformatic applications, such as Galaxy Cloudman[36] and the Cloud Virtual Resource (CloVR)[37], to carry out analyses on the cloud, users upload and install software together with the input data. In CloVR, the software is assembled into automated bacterial sequence analysis pipelines that are pre-installed on a virtual computer (that is, a virtual machine)[24,37], which greatly facilitates installation and affords portability across different computer operating systems, including local desktop computers and online cloud servers. As the cloud provides on-demand computational resources and storage space, both of which

Box 3 | **Examples of clinical bioinformatic support models**

The following three cases provide examples for different bioinformatic support models that use local and online cloud-based infrastructures for low- and high-throughput sample processing.

Case 1. Health care settings that have low sample throughput and limited resources can install and use software on local single- or multi-central processing unit (CPU) desktop computers, which provides strict control over software, sequence data and associated health information. For limited data loads, local hardware can adequately handle bioinformatic needs in reasonable timeframes and at moderate costs. Examples of this could include independent physician offices and small hospitals in rural locations.

Case 2. Hospitals that have higher sample throughput can afford a set-up of more extensive local computational infrastructures to handle bioinformatic analyses economically. However, peak demands in computational support — for example, in cases of disease outbreaks — could rapidly outstrip the capacity for local analyses. In addition, apart from initial investments, information technology support for cluster maintenance will be required, which complicates cost calculations. Examples of this would be modest to large health care facilities and centralized commercial laboratory service providers.

Case 3. Hospitals that seek to cover peaks in computational demands or that try to avoid heavy capital investments to set up computational support systems, while trying to increase sample throughput, can take advantage of on-demand online cloud computing services. Data and analysis results are sent over the Internet for processing. Regulations on online data security during transfer and storage have not yet been clearly defined. Additionally, customization of bioinformatic software on the cloud is limited, which leads to reliance on validated analysis pipelines. With this approach, control over data, software and analysis parameters is partially transferred to the bioinformatic service provider, in a model similar to that of the commercial centralized laboratory services that are currently in use at many hospitals. Examples of this could be all of the above, including large private practices and hospitals of all sizes, especially those without the financial means to invest in local bioinformatics infrastructure.

are paid for by the hour, data outputs should be downloaded after completion of the analysis and all remaining data and software removed from the cloud, which CloVR supports in a fully automated, seamless manner.

Data storage and integration. Bacterial WGS-based MDx relies heavily on the use of reference databases, for example, the use of reference strains to carry out isolate typing and phylogenetic analysis or to predict antimicrobial resistance phenotypes on the basis of comparison with known marker genes. In addition, if such MDx system is widely implemented, substantial amount of new bacterial WGS data will be generated on a continuous basis. These data need to be integrated back into the reference databases for subsequent iterations of the MDx application. For example, if a hospital identifies an MRSA isolate as part of its routine surveillance programmes, then the corresponding genotypic information will be important for the inclusion of this isolate in subsequent analyses in order to discover and track potential MRSA transmission events in the same hospital, as well as for long-term national and international surveillance. Beyond their direct use *in situ*, the generated WGS data will also be a valuable academic and/or commercial resource, for example, if genetic features can be associated with a specific outbreak after multiple bacterial isolates have been sequenced and analysed. These genetic loci can then be either diagnostic markers for commercial use, or targets for functional research or vaccine development[38]. Although commercial interests can foster and accelerate both the development and the implementation of new WGS-based MDx products, the immense value of the generated output for both the public health community and the scientific research community should be a strong argument to politically mandate open data sharing, as commercial interests might otherwise result in company-owned, proprietary genome sequence databases.

Central data storage provides economic benefits that are associated with unified data management. New data repositories could become integrated with established non-private databases, such as those maintained at the US National Center for Biotechnology Information (NCBI), which should guarantee open data sharing between commercial MDx test providers and the academic research community. Consistent data types and formats will have to be adapted to handle the enormous amount of data that are generated by routine clinical use of WGS. The current paradigm is to store raw sequence reads, which represents a great data burden and might not provide the best use for integration into future MDx applications. For example, the raw data from a single *E. coli* genome sequenced at ~250-fold coverage that are deposited at the NCBI Short Read Archive amounts to ~500 megabytes, whereas the corresponding GenBank file (accession number: AIFA00000000), which contains all the information of the assembled and annotated genome that is required for phylogenetic or genotypic characterization, has a size of less than two megabytes. This 250-fold reduction in the data 'footprint' substantially decreases both the costs and the efforts that are associated with long-term bacterial WGS data management.

Regulations

As MDx enters the clinical paradigm, the control of information is becoming an important issue at the laboratory, provider and patient levels. Online data transfer, processing and storage will require safety precautions to protect sensitive patient information. Commercial online resource providers, such as the Amazon web services, have recognized this problem and responded by obtaining appropriate security certifications. Additional regulatory requirements need to be defined by the legislative body. Although providers of clinical laboratory services in the United States are obligated to obtain the Clinical Laboratory Improvement Amendment (CLIA) certification, similar standards are lacking for bioinformatic sequence analyses. A California bill sponsored by the consumer genomics firm 23andMe introduced the term 'post-CLIA bioinformatics services' to distinguish laboratory services that are regulated by the CLIA from post-laboratory bioinformatic analysis services[39]. According to this bill, which has been discussed controversially[40], post-CLIA bioinformatics services would not require approval or review of the algorithm by a government regulatory body but instead by a designated individual who is vaguely defined to possess a background in either bioinformatics or biostatistics. Although the main focus of current discussions is on human MDx, the legislative regulation of both the definitions and the validations that are applied to bioinformatic analysis parameters in bacterial MDx will probably also become more relevant in the near future.

Future directions

Clinical application of bacterial WGS-based MDx will be crucial for global efforts to identify, prevent and treat infectious diseases. In order to be truly successful, applications need to become widely available to

clinical laboratories, including remote field settings and resource-poor hospitals in the developing world. For this scenario we envision a model that will require little more than a power supply, an Internet connection and an individual who has minimal laboratory skills to integrate bacterial genome sequencing as a diagnostic tool with essentially limitless applications in any health care setting. The first step towards integration of bacterial WGS-based MDx into the clinic could be its adoption by national health services and public health laboratories. These services operate on a defined set of clinical pathogens and diagnostic parameters, and at a scale that is large enough to allow timely validation and optimization of future MDx tests.

W. Florian Fricke and David A. Rasko are at the Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, Maryland, USA.

Correspondence to W.F.F.
e-mail: wffricke@som.umaryland.edu

doi:10.1038/nrg3624
Published online 26 November 2013

1. Koser, C. U. *et al.* Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog.* **8**, e1002824 (2012).
2. Koser, C. U. *et al.* Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N. Engl. J. Med.* **366**, 2267–2275 (2012).
3. Rasko, D. A. *et al.* Origins of the *E. coli* strain causing an outbreak of haemolytic-uraemic syndrome in Germany. *N. Engl. J. Med.* **365**, 709–717 (2011).
4. Rasko, D. A. *et al. Bacillus anthracis* comparative genome analysis in support of the Amerithrax investigation. *Proc. Natl Acad. Sci. USA* **108**, 5027–5032 (2011).
5. Snitkin, E. S. *et al.* Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci Transl Med.* **4**, 148ra116 (2012).
6. Gardy, J. L. *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* **364**, 730–739 (2011).
7. 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).
8. Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. & Crook, D. W. Transforming clinical microbiology with bacterial genome sequencing. *Nature Rev. Genet.* **13**, 601–612 (2012).
9. Pallen, M. J., Loman, N. J. & Penn, C. W. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Curr. Opin. Microbiol.* **13**, 625–631 (2010).
10. Butkus, B. Report values global MDx market at $11B by 2015; projects qPCR to remain key driver. *PCR Insider* [online], http://www.genomeweb.com/pcrsample-prep/report-values-global-mdx-market-11b-2015-projects-qpcr-remain-key-driver (2012).
11. Rosewell, G. A. *Global Molecular Diagnostic Market: Opportunities and Future Forecast* (Renub Research, 2009).
12. Harris, S. R. *et al.* Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect. Dis.* **13**, 130–136 (2013).
13. Loman, N. J. *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA* **309**, 1502–1510 (2013).
14. Chan, J. Z. *et al.* Metagenomic analysis of tuberculosis in a mummy. *N. Engl. J. Med.* **369**, 289–290 (2013).
15. Crump, J. A. *et al.* Antimicrobial resistance among invasive nontyphoidal *Salmonella enterica* isolates in the United States: National Antimicrobial Resistance Monitoring System, 1996 to 2007. *Antimicrob. Agents Chemother.* **55**, 1148–1154 (2011).
16. Koser, C. U. *et al.* Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. *N. Engl. J. Med.* **369**, 290–292 (2013).
17. Illumina gets 5-year FDA contract to identify food bugs. *Reuters* [online], http://www.reuters.com/article/2012/09/18/illumina-fda-idUSL4E8KI3RN20120918 (2012).
18. Harrison, E. M. *et al.* Whole genome sequencing identifies zoonotic transmission of MRSA isolates with the novel *mecA* homologue *mecC*. *EMBO Mol. Med.* **5**, 509–515 (2013).
19. He, M. *et al.* Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nature Genet.* **45**, 109–113 (2013).
20. Reuter, S. *et al.* A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. *BMJ Open* **3**, e002175 (2013).
21. Underwood, A. P. *et al.* Public health value of next-generation DNA sequencing of enterohemorrhagic *Escherichia coli* isolates from an outbreak. *J. Clin. Microbiol.* **51**, 232–237 (2013).
22. Cossins, D. Real-time outbreak sequencing. *The Scientist* [online], http://www.the-scientist.com/?articles.view/articleNo/33771/title/Real-time-Outbreak-Sequencing/ (2012).
23. Loman, N. J. *et al.* High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Rev. Microbiol.* **10**, 599–606 (2012).
24. Angiuoli, S. V., White, J. R., Matalka, M., White, O. & Fricke, W. F. Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. *PLoS ONE* **6**, e26624 (2011).
25. Dudley, J. T., Pouliot, Y., Chen, R., Morgan, A. A. & Butte, A. J. Translational bioinformatics in the cloud: an affordable alternative. *Genome Med.* **2**, 51 (2010).
26. Kahn, S. D. On the future of genomic data. *Science* **331**, 728–729 (2011).
27. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotech.* **30**, 434–439 (2012).
28. Junemann, S. *et al.* Updating benchtop sequencing performance comparison. *Nature Biotech.* **31**, 294–296 (2013).
29. Maitra, R. D., Kim, J. & Dunbar, W. B. Recent advances in nanopore sequencing. *Electrophoresis* **33**, 3418–3428 (2012).
30. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
31. Hadfield, J. & Loman, N. J. Next generation genomics: world map of high-throughput sequencers [online], http://omicsmaps.com/ (2013).
32. Micheel, C. M., Nass, S. J., & Omenn, G. S. (eds) *Evolution of Translational Omics: Lessons Learned and the Path Forward* (The National Academies Press, 2012).
33. Swan, D. Short-read alignment on the Raspberry Pi. *Eridanusdotnet* [online], http://eridanus.net/?p=10689 (2013).
34. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).
35. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
36. Afgan, E. *et al.* Galaxy CloudMan: delivering cloud compute clusters. *BMC Bioinformatics* **11** (Suppl. 12), S4 (2010).
37. Angiuoli, S. V. *et al.* CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* **12**, 356 (2011).
38. Seib, K. L., Zhao, X. & Rappuoli, R. Developing vaccines in the era of genomics: a decade of reverse vaccinology. *Clin. Microbiol. Infect.* **18** (Suppl. 5), 109–116 (2012).
39. SB-482 Biological data analysis services: regulation. *California legislative information* [online], http://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=200920100SB482 (2009).
40. Ray, T. What's in a name? Experts question bill calling DTC genomics 'post-CLIA bioinformatics services'. *Pharmacogenom. Reporter* [online], http://www.genomeweb.com/dxpgx/whats-name-experts-question-bill-calling-dtc-genomics-post-clia-bioinformatics-s (2009).
41. Emmadi, R. *et al.* Molecular methods and platforms for infectious diseases testing a review of FDA-approved and cleared assays. *J. Mol. Diagn.* **13**, 583–604 (2011).
42. Croxatto, A., Prod'hom, G. & Greub, G. Applications of MALDI–TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiol. Rev.* **36**, 380–407 (2012).
43. Hrabak, J., Chudackova, E. & Walkova, R. Matrix-assisted laser desorption ionization-time of flight (MALDI–TOF) mass spectrometry for detection of antibiotic resistance mechanisms: from research to routine diagnosis. *Clin. Microbiol. Rev.* **26**, 103–114 (2013).
44. Forrest, G. N. PNA FISH: present and future impact on patient management. *Expert Rev. Mol. Diagn.* **7**, 231–236 (2007).
45. Brzuszkiewicz, E. *et al.* Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). *Arch. Microbiol.* **193**, 883–891 (2011).
46. 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).
47. Rohde, H. *et al.* Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N. Engl. J. Med.* **365**, 718–724 (2011).

Acknowledgements
The authors thank S. Angiuoli, A. D. Harris and J. K. Johnson for their feedback and suggestions. Funding for this study was provided by the US National Institute of Allergy and Infectious Diseases, the US National Institutes of Health, under project number 1R21AI100192.

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

FURTHER INFORMATION
Amazon web services: http://aws.amazon.com/security/
CloVR: http://clovr.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF