University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Kwong, JC;Mccallum, N;Sintchenko, V;Howden, BP

Title:
Whole genome sequencing in clinical and public health microbiology

Date:
2015-04-01

Citation:
Kwong, J. C., Mccallum, N., Sintchenko, V. & Howden, B. P. (2015). Whole genome sequencing in clinical and public health microbiology. Pathology, 47 (3), pp.199-210. https://doi.org/10.1097/PAT.0000000000000235.

Persistent Link:
https://hdl.handle.net/11343/262576

License:
CC BY-NC-ND

Pathology (April 2015) **47(3)**, pp. 199–210

OPEN

MOLECULAR DIAGNOSTICS IN MICROBIOLOGY

Whole genome sequencing in clinical and public health microbiology

J. C. KWONG[1,2], N. MCCALLUM[3,4], V. SINTCHENKO[3,4] AND B. P. HOWDEN[1,2,5]

[1]Microbiological Diagnostic Unit Public Health Laboratory, Department of Microbiology and Immunology, the University of Melbourne, at the Doherty Institute for Infection and Immunity, Melbourne, Vic, [2]Department of Infectious Diseases, Austin Health, Heidelberg, Vic, [3]Marie Bashir Institute for Infectious Diseases and Biosecurity and Sydney Medical School, The University of Sydney, Sydney, NSW, [4]Centre for Infectious Diseases and Microbiology-Public Health, Institute of Clinical Pathology and Medical Research, Westmead Hospital, Sydney, NSW, and [5]Department of Microbiology, Monash University, Clayton, Vic, Australia

Summary

Genomics and whole genome sequencing (WGS) have the capacity to greatly enhance knowledge and understanding of infectious diseases and clinical microbiology. The growth and availability of bench-top WGS analysers has facilitated the feasibility of genomics in clinical and public health microbiology. Given current resource and infrastructure limitations, WGS is most applicable to use in public health laboratories, reference laboratories, and hospital infection control-affiliated laboratories. As WGS represents the pinnacle for strain characterisation and epidemiological analyses, it is likely to replace traditional typing methods, resistance gene detection and other sequence-based investigations (e.g., 16S rDNA PCR) in the near future. Although genomic technologies are rapidly evolving, widespread implementation in clinical and public health microbiology laboratories is limited by the need for effective semi-automated pipelines, standardised quality control and data interpretation, bioinformatics expertise, and infrastructure.

Key words: Clinical microbiology, genomics, public health microbiology, sequencing, WGS, whole genome sequencing.

Received 11 December 2014, revised 5 January, accepted 22 January 2015

BACKGROUND

Advances in technology, including the rapidly growing field of genomics, are transforming clinical medicine. The term 'genomics' was first coined in 1986 by Dr Thomas Roderick, a geneticist in Bar Harbour, Maine, and was initially intended as a term to encompass the study and comparison of genomes of various species, including their evolution and relationships.[1] Essentially, genomics involves the application of DNA sequencing and the subsequent analyses using *in vitro* experiments and bioinformatic approaches to study the structure and function of genes, both human and pathogen.

In recent decades, genomics has been used extensively in a research capacity to study infectious agents, with the development of high throughput 'next-generation' sequencing technologies allowing detailed large scale analyses of entire pathogen genomes. However, despite the perceived benefits of sequencing technology to support traditional methods in diagnostic microbiology, there has been limited application in clinical and public health laboratories in Australasia to date.

This review aims to examine applications of current technologies in diagnostic microbiology and to outline the added value and current limitations of genomics, and in particular, bacterial whole genome sequencing (WGS), in order to support microbiologists in future implementation and use of these new technologies in clinical and public health practice.

WGS: METHODS, SEQUENCING TECHNOLOGY AND DATA ANALYSIS

The evolution of sequencing technology

The Human Genome Project instigated a revolution in sequencing technologies resulting in the establishment of high-throughput WGS as an important tool for the study of organisms, both human and microbial. Initial technological advances focussed on enhancing the chain termination sequencing method published by Sanger *et al*. in 1977.[2] These modifications included fluorescent labelling of molecules, development and utilisation of capillary-based instruments, and automation of these processes to allow analysis of multiple samples in parallel.[3]

As Sanger sequencing was limited to <1000 bases, the search for more efficient methods for sequencing long, complex pieces of DNA such as entire chromosomes, led to other approaches. Initially described in 1979, 'shotgun sequencing', where longer segments of DNA were randomly fragmented into smaller segments for Sanger sequencing, was an early step towards facilitating genome sequencing, but was slow and labour-intensive for an entire genome, requiring a map to assemble the sequenced fragments.[4] With the parallel advancements in computation technology and software, this strategy evolved into 'whole-genome shotgun sequencing', which bypassed the need for a genetic map by using bacterial clones to produce a large amount of redundant sequence read data across the genome and utilising newer computation technology to assemble the sequence reads. This method resulted in the landmark sequencing of the *Haemophilus influenzae* genome,[5] the first genome from a free-living organism to be sequenced, and was the most popular and advanced sequencing method until the late 2000s.[6]

Next-generation sequencing

More recently, the invention of high-throughput 'next-generation' sequencing technology, with relatively simple benchtop technology and efficient library preparation protocols, has

Print ISSN 0031-3025/Online ISSN 1465-3931 © 2015 Royal College of Pathologists of Australasia. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License, where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially.
DOI: 10.1097/PAT.0000000000000235

significantly improved the capacity to perform low-cost, efficient WGS, and has made it a feasible tool to enhance clinical diagnostic investigations in near real-time. Next-generation processes generally involve parallel sequencing, producing vast quantities of data that require modern computation methods to assemble the sequence reads.

Figure 1 shows the typical workflow and application of next-generation sequencing that could be applied to clinical microbiology.

There are a number of commercialised next-generation sequencing methods in use and novel technologies emerging onto the market, each with advantages and disadvantages, which have been reviewed in detail previously,[6–12] although several are now outdated with the rapid growth in technology. While this review is not exhaustive, a summary of the current most common sequencing technology is shown in Tables 1–3.

Sequencing options for clinical microbiology: what needs to be considered?

There are a number of important considerations in comparing sequencing platforms for clinical microbiology, and deciding whether to perform in-house sequencing or to out-source to an experienced sequencing service provider.

Cost

The cost of implementation including equipment set up, routine sequencing costs for reagents and consumables as well as post-processing bioinformatics costs is an obvious, but significant factor. These expenses can be measured in cost per sequencing run, cost per organism genome sequenced, or cost per megabase of output data. To be a financially viable option for clinical microbiology laboratories WGS must be able to replace current technologies (e.g., methods for molecular characterisation of pathogens such as pulsed field gel electrophoresis), or provide additional benefits in patient outcomes and clinical or laboratory efficiency.

In-house versus outsourced

In-house sequencing may improve turnaround times for data generation and analyses, however this requires significant investment in technology and data analysis expertise. Although outsourcing may result in longer turnaround times, it may improve overall time and cost efficiency of sequencing by pooling isolates from smaller laboratories with insufficient sample numbers to fill a standard sequencing run. However, clear communication between referrer and provider is paramount to ensure that the clinical questions to be answered with WGS are clear, and that the subsequent analysis is understood and verified by both parties.

Sequencing capacity

Some available technologies allow sequencing a handful of bacterial genomes in a few hours, while others have capacity to sequence 50–100 bacterial genomes in a single run that may take between 1 and 3 days. Flexibility in sequencing throughput, without significant financial implications of cost per sample, should also be considered. A reference microbiology laboratory needs to be able to sequence a large collection of 50–100 samples for epidemiological purposes, but also have the ability to sequence a small number of strains of pathogens of public health concern urgently for a similar cost per sample.

Adaptability

Adaptability of the sequencing platform to upgrades and changing sequencing practices is another factor, with sequencing technology rapidly evolving. The capability of the sequencer to be used for human genome sequencing and for research groups may also allow sharing of resources in smaller centres with lower demand for microbial WGS.

Data quality

The quality of a sequence result can be reported using a score to indicate the quality and accuracy of each nucleotide base call.

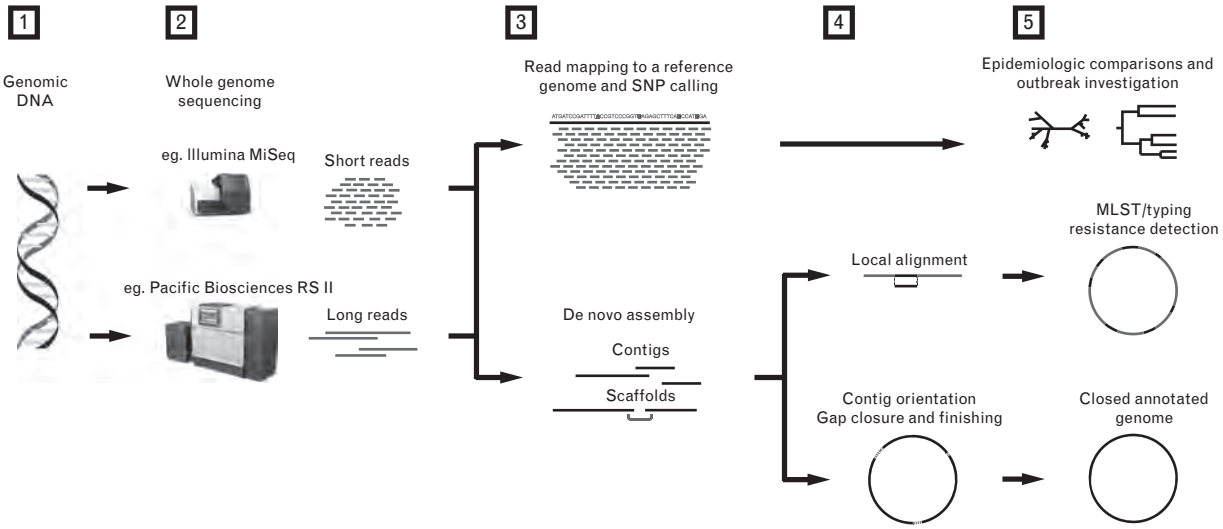


Fig. 1 Whole genome sequencing workflow. (1) DNA extraction from homogeneous microbial samples, e.g., single bacterial colony from a pure culture. (2) Whole genome sequencing using next-generation sequencers. Most high-throughput sequencers produce short reads (e.g., Illumina MiSeq), although long reads from Pacific Biosciences RS II or Illumina TruSeq technology may facilitate *de novo* assembly more readily. (3) SNPs called from read mapping to a reference genome can be used for phylogenetic comparisons to assist in epidemiological and outbreak analyses. Reads can also be assembled de novo into longer contiguous sequences (contigs), and orientated and aligned to form scaffolds. (4) The resulting *de novo* assemblies can be used for further analyses such as typing and resistance detection based on local alignment tools (e.g., BLAST), or can be further finished into a completed or closed genome. This finishing stage usually requires gap closure through extensive 'wet-lab' techniques such as primer walking, and so is generally performed for research purposes, although WGS long reads are increasingly being used to produce more complete *de novo* assemblies and minimise the amount of laboratory work required. (5) Data analysis for outbreak investigation, typing, or resistance detection. Closed annotated genomes can be used as reference genomes for comparison, or can be analysed in further detail.

Table 1 Popular sequencing technology

Traditional sequencing	
Sanger sequencing	• Still widely used for sequencing short segments of DNA (up to 1000 bp) due to the ease and accuracy of sequencing • Labour, time and cost intensive for sequencing entire genomes on a regular basis
Shotgun sequencing	• Involved fragmentation of long strands of DNA into numerous smaller segments for Sanger sequencing • Facilitated initial whole genome sequencing efforts • Shotgun approach still utilised by 'next-generation' sequencing methods
Next-generation sequencing technologies	
Pyrosequencing (Roche 454)	• Detects pyrophosphate release on addition of a complementary nucleotide to determine the template sequence • Lower throughput and subsequently higher sequencing cost per base • One of the earlier next-generation technologies, but now being phased out with Roche intending to cease production in 2016
SOLiD sequencing (Life Technologies)	• Sequencing by Oligonucleotide Ligation and Detection (SOLiD) uses a ligation-based approach • Less popular than Life Technologies' other platform, the Ion Torrent, and likely to be superseded by newer technologies
Ion semiconductor sequencing (Life Technologies Ion Torrent)	• Uses a sequencing-by-synthesis method, detecting changes in pH due to hydrogen ion release with synthesis of complementary DNA • Popular due to lower sequencer cost and speed of sequencing • Requires separate emulsion PCR library amplification prior to sequencing (slow and complicated), though automation can be performed using the separate Ion Chef system • Higher error rates, particularly homopolymers, than other platforms and poor coverage of extremely AT-rich or GC-rich regions • Ion Torrent Personal Genome Machine (PGM) and newer, higher throughput Ion Proton available
Illumina sequencing	• Uses a sequencing-by-synthesis method, detecting release of fluorescent labels from incorporated nucleotides to determine sequence • Current market leader with high sequence throughput, with low error rate and low sequencing cost per base • Limitations of short read sequences and a longer sequencing run time • Several platforms with moderate (MiSeq), moderate-high (NextSeq) and high (HiSeq) throughput • TruSeq long read technology recently introduced to produce synthetic reads of 10 kb in length (currently only HiSeq 2000/2500)
Single molecule real-time sequencing (Pacific Biosciences)	• Novel method – observes natural synthesis of unmodified DNA by DNA polymerase, with reads up to 40 kb in length, using nucleotides with fluorescent labels attached to the terminal phosphate (rather than the base) • Higher raw error rates, but errors are randomly distributed (vs. ends of reads or homopolymers), and overlapping reads can produce a consensus sequence with high accuracy • Has significantly improved *de novo* assembly and bacterial genome completion without needing traditional PCR-based gap closure • High setup cost and low throughput have limited implementation, though outsourcing options are available
Emerging technologies	
Nanopore sequencing (Oxford Nanopore)	• Probably the leader of the pack of benchtop sequencing technologies in development • Detects characteristic disruptions in a current applied across a protein channel or 'nanopore' as each nucleotide of a strand of DNA is passed through the nanopore • Method still being refined, but has the capability of generating long-sequence reads • Two portable/affordable benchtop sequencers available – the MinION (disposable USB stick), and the GridION (rack-mountable)

For example, Illumina uses a Phred-score (see Appendix 1: Glossary, http://links.lww.com/PAT/A30), with a score of 20 (Q20) equating to 1 error every 100 bases, or a 1% error rate, while a score of 30 (Q30) indicates an error rate of 1 every 1000 bases (0.1%). Modern WGS methods aim to achieve a quality score of 30 across the genome, although sequencing for different purposes may have different targets.

Despite the differences between sequencing platforms, in experienced hands, the output from several of the established next-generation sequencers (Tables 2 and 3) appears to be sufficient for most current clinical applications.[13] The potential advantages of long sequence reads for clinical microbiology are still being investigated,[14] although in a research environment, long reads (>5000 bp) have helped overcome many of the limitations of short read data.[15,16] Examples include resolution of tandem repeat units and insertion sequences, identification of smaller circularised sequences such as plasmids, and bridging contiguous sequence gaps that litter *de novo* assemblies from short reads to assist with genome closure. Although this may soon become the standard in bacterial genome sequencing, the advantages are offset by the lower throughput, and higher implementation and sequencing costs, with other sequencers producing output data of sufficient quality and resolution for clinical purposes.

Bioinformatic analysis of sequencing data

With the technological advances in generating large amounts of high quality sequencing data, the bottleneck in implementing whole genome sequencing for clinical purposes has shifted to the post-sequencing data analysis. The term 'bioinformatics' encompasses the handling and analysis of sequencing data, usually with the assistance of computer-based algorithms.

Although both 'open source' and commercially available bioinformatic programs/tools have been specifically developed for use in a clinical setting by clinicians with limited bioinformatics knowledge,[17–20] many of these lack the ability to batch

Table 2 Comparison of popular next-generation sequencers: benchtop sequencing platforms for low-moderate throughput

	Illumina MiSeq	Ion Torrent PGM (Life Technologies)	Ion Proton (Life Technologies)	Roche 454 GS Junior
Configuration	Nextera Reagent Kit v3	Ion 318™ Chip v2	Proton I chip	GS Junior Plus
Dimensions	69 × 57 × 52 cm	61 × 51 × 53 cm	54 × 78 × 47 cm	40 × 60 × 40 cm
Weight	54.5 kg	30 kg	59 kg	25 kg
Preparation time	8 hours	8 hours	8 hours	8 hours
Sequencing time	60 hours	4–7 hours	2–4 hours	18 hours
Data output (Gb per run)	13–16 Gb/run	600 Mb – 2 Gb/run	10 Gb/run	50–70 Mb/run
Sequence read length	2 × 300 bp	200 / 400 bp	200 bp	700 bp
Number of S. aureus (~2.9 Mb genome) per run at 30x coverage	75	15	60	1 (at 15x coverage)
Error rate*	Overall 0.1% Indel error rate 0.001 per 100 bp	Overall 0.5–2.5% Indel error rate 1.5 per 100 bp	Not reported	Overall 0.2–1.0% Indel error rate 0.4 per 100 bp
Accuracy	Mostly Q30	Mean Q20 (Q10-Q30)	Not reported	Q20-Q30
Cost of platform (approximate)†	$150,000	$100,000	$150,000	$100,000
Advantages	• Higher accuracy and data output • Low cost per output • Library amplification incorporated	• Low platform cost • Short run time	• Low cost per output • Rapid run time	• Smaller instrument size • Longer read length (up to 800 bp with GS Junior+)
Disadvantages	• Longer run time • Higher platform cost • Shorter read length	• Requires separately amplified sequence libraries by emPCR‡ • Higher indel error rate, particularly with homopolymers • Quality of sequence deteriorates at ends of reads, though can be improved with post-sequencing read clipping • Poor coverage of AT-rich regions • Can be more difficult to assemble	• Requires separately amplified sequence libraries by emPCR‡ • Higher indel error rate, particularly with homopolymers • Quality of sequence deteriorates at ends of reads, though can be improved with post-sequencing read clipping • Poor coverage of AT-rich regions • Can be more difficult to assemble	• More 'hands-on' time – requires manually amplified sequence libraries by emPCR • Higher indel error rate, particularly with homopolymers • Higher cost per output • Requires manually amplified sequence libraries • Roche closing sequencing operations and ceasing production

* Based on Loman *et al.*[7] and Jünemann *et al.*[9]
† Costs are only approximate at time of writing, and may vary substantially – intended only as a rough guide.
‡ emPCR = emulsion PCR. Slow and complicated process; automated amplification systems are available for Ion Torrent/Ion Proton (Ion Chef).

analytical processes on large datasets and customise automation of data analysis pipelines, as a trade-off for the ease of use via a graphical user interface (GUI). The majority of available bioinformatic software requires some knowledge of the text-based command-line of the UNIX or Linux operating systems, allowing custom programming scripts and pipelines to automate data manipulation and analysis in a single step. Table 4 shows examples of bioinformatics tools commonly employed for analysis of bacterial genomes.

In assessing bioinformatics software for analysis of WGS data for clinical microbiology, there are several considerations and criteria to take into account.

Useability

Although Linux-based tools will continue to predominate due to the ability and ease in customising analyses, tools that can be operated through a GUI may be preferred by those unfamiliar with bioinformatics.

Automation

Another key advantage of Linux-based tools, although often requiring some initial work to establish, is a 'pipeline' for specific types of analyses. These pipelines enable 'batching' or sequential running of a number of processes on multiple genomes with a single command, compared with running each component individually, before manually entering the next command.

Speed

In a clinical setting, the ability to obtain a result quickly is often a priority over correcting minor inaccuracies in single nucleotide polymorphism (SNP) calls that do not change the overall result. Bioinformatic tools that are able to analyse multiple samples together and utilise the processing power and resources of modern computers to split large complex processes into smaller processes running in parallel exemplify the 'many hands make light work' proverb, a feature known as multithreading or hyperthreading.

Accuracy and detail

It naturally follows that the accuracy of the analysis is important for clinical microbiology, particularly for organism identification, typing, and resistance detection. However, while research pursuits require accurate and detailed analyses, the additional resolution from this level of detail is not always required for clinical decisions. For example, in inferring

Table 3 Comparison of popular next-generation sequencers: high-end sequencing platforms for high throughput/long reads

	Illumina HiSeq 2500	Illumina NextSeq 500	Roche 454 GS FLX+	Pacific Biosciences RS II
Configuration	Rapid-run mode Dual flow-cell	High output flow cell	Titanium XL+	RS II
Dimensions	119 × 76 × 94 cm	59 × 53 × 64 cm	Upper 74 × 70 × 36 cm Lower 75 × 91 × 93 cm	200 × 77 × 158 cm
Weight	221 kg	83 kg	242 kg	1091 kg
Preparation time	8 hours	8 hours	8 hours	8 hours
Sequencing time	60 hours	30 hours	24 hours	4 hours
Data output (Gb per run)	250–300 Gb/run	100–120 Gb/run	0.7 Gb/run	0.5–1 Gb/run
Sequence read length	2 × 250 bp[*]	2 × 150 bp	700 bp	1,000–40,000 bp[†]
Number of S. aureus (~2.9 Mb genome) per run at 30x coverage	1200[‡]	480[‡]	5	1
Error rate	0.1%	0.1%	0.2–1.0%	14%[§]
Accuracy	Mostly Q30	Mostly Q30	Q20-Q30	Mostly Q30
Cost of platform[‖]	$650,000	$250,000	$500,000	$750,000
Advantages	• Massive throughput (though better suited to human genome sequencing) • Low cost per output • High output and rapid run modes	• High throughput suitable for microbial genomes • Lower instrument cost • Low cost per output • Dimensions suitable for 'benchtop' • Potential for expansion/upgrades	• Read length up to 1000 bp facilitates *de novo* assembly	• Long reads facilitate *de novo* assembly and resolution of repetitive genomic regions • Able to sequence regions of high GC content (results in more uniform coverage of the genome) • Detects modified DNA bases, eg, DNA methylation patterns
Disadvantages	• Longer run time • Short reads limit *de novo* assembly[*] • Higher instrument cost	• Short reads limit *de novo* assembly	• More 'hands-on' time – requires manually amplified sequence libraries by emPCR • Higher cost per output • Roche closing sequencing operations and ceasing production	• Lower output • Higher error rate in individual reads[§] • Higher instrument cost and cost per output • Large instrument size

[*] TruSeq Long Read technology allows sequencing reads of 10,000 bp in length.
[†] N50 = 14,000 bp; i.e., half of the sequence data is contained in reads >14,000 bp.
[‡] Theoretical number for comparison only – requires custom-synthesised indices. Current Illumina Index Kits (Nextera XT) allow up to 384 samples per flow cell.
[§] Error rate is based on raw read error rate. However, as the error model for SMRT sequencing is stochastic, combining reads can produce high quality consensus sequence across all bases. Our experience is that in comparison with sequencing on the Illumina MiSeq, the RS II produces high quality consensus sequences with an error rate approximately 1 per 1000 bases (predominantly homopolymers).
[‖] Costs are only approximate at time of writing, and may vary substantially – intended only as a rough guide

phylogenetic relationships between organisms, Bayesian methods have become popular in estimating a phylogenetic tree. However, while substantially faster neighbour-joining methods may not produce as accurate an evolutionary tree, the resolution is likely to be sufficient and rapid enough for analysing a public health outbreak in real time, where the organisms involved are highly clonal.

Cost

Although there is a large amount of free publicly available software for bioinformatic analysis, these tend to be command-line based with low adaptability across different sequencing platforms. GUI-based software that can be used with relatively little experience is available, though often at a cost, both financial as well as speed and occasionally detail. For example, Applied Maths Bionumerics suite offers a wide range of tools for analysis at a cost of approximately AU$17 per isolate. Galaxy is a free, open source web-based platform for bioinformatics, but requires data uploading and sharing of public servers, which limit the speed of analyses.

Documentation and support

An advantage of commercial software is the availability of user manuals and professional support for troubleshooting. In contrast, while there is usually some documentation for use and limited support available from open-source software developers, many issues require local computing expertise for implementation and troubleshooting.

Public genome data

A number of public repositories of sequencing data are available, with published sequences available for download for comparative genomic analysis. The National Center for Biotechnology Information's (NCBI) GenBank database currently lists just over 3000 annotated complete genome assemblies in addition to 25,000 draft genome assemblies. The NCBI genome data are exchanged with the European Molecular Biology Laboratory's (EMBL) European Nucleotide Archive and the DNA Data Bank of Japan (DDBJ), which together form the International Nucleotide Sequence Database Collaboration. The Global Microbial Identifier initiative (http://www.globalmicrobialidentifier.org/) is another independent collaboration that aims to coordinate a microbial sequencing data collection and collate the collective genomic and metadata to facilitate subsequent analysis on a global scale, although the data are only available to collaborators.

Table 4 Common software for bioinformatic analysis

De novo assembly
- *De novo* assembly involves using computer algorithms to align overlapping WGS reads to form longer contiguous sequences known as contigs, and order the contigs into a framework of the sequenced genome (scaffolds). Velvet (https://www.Ebi.Ac.Uk/~zerbino/velvet/)[72] and SPAdes (http://bioinf.Spbau.Ru/spades)[73] are two of the more popular assemblers for Illumina short-reads, while Ion Torrent reads are better assembled using MIRA (http://www.Chevreux.Org/projects_mira.Html). Other commonly used assemblers include Newbler (http://swes.Cals.Arizona.Edu/maier_lab/kartchner/documentation/index.Php/home/docs/newbler) for 454 pyrosequencing reads, and the commercial CLC Genomics suite. Assemblers used for PacBio long reads include SPAdes, HGAP[74] and the Celera-MHAP assembler.[75]
- Contigs can be visualised in the Java-based program Mauve (http://gel.Ahabs.Wisc.Edu/mauve/), which can also order and orientate contigs to a reference genome. Alternatively, command-line tools such as MUMmer (http://mummer.Sourceforge.Net/) can be used to automate and batch this process as part of an assembly pipeline.

Annotation
- Genome annotation includes identification of DNA segments of known and probable open reading frames (ORF) that contain gene coding DNA, and matching the identified segments to a database of known gene sequences. Tools include the web-based RAST (http://www.nmpdr.org/FIG/wiki/view.cgi/FIG/Rapid AnnotationServer), NCBI's Prokaryotic Genome Annotation Pipeline (http://www.ncbi.nlm.nih.gov/genome/annotation_prok/) or the command-line tool Prokka (http://www.vicbioinformatics.com/software.prokka.shtml) for automated genome annotation.

Genome visualisation and comparison
- Once assembled and annotated, genomes can be viewed using a genome browser to display the structure and embedded genetic elements of a genome in a graphical format, and manipulate the genome sequence if required. The Wellcome Trust Sanger Institute's Artemis (http://www.sanger.ac.uk/resources/software/artemis/), and the commercially available Geneious Pro suite (http://www.geneious.com/) are examples of genome browsers.
- Visual comparisons of multiple genomes can also be made using the above utilities.

Alignment and read mapping
- Read mapping is the process of aligning reads to a reference, using a combination of local and global alignment. Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) and BWA (http://bio-bwa.sourceforge.net/) are two of the more popular short read alignment algorithms.[76]
- BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi), the most widely used utility for searching a sequence database, uses local alignment of sequence segments. BLAST can be run either as a web-based tool, or incorporated using a command line.
- Whole genome alignment is a computationally intensive process, but can be performed using Mauve or Mugsy/MUMmer (http://mugsy.sourceforge.net/).

SNP/variant calling
- Single nucleotide differences identified from aligning comparator sequences to a reference can be used to describe genetic relationships between isolates. Multiple tools are available,[77] and are frequently incorporated into more automated software packages.
- We use the Nesoni suite of tools (http://www.Vicbioinformatics.Com/software.Nesoni.Shtml) as well as SAMtools (http://samtools.Sourceforge.Net/), Freebayes (https://github.Com/ekg/freebayes) and Nucmer (part of MUMmer).

Phylogenetic analysis
- Phylogenetic trees can be used to analyse and visualise the SNP differences between isolates, although the true phylogeny of a group of isolates is never known. Popular methods include the simpler but rapid neighbour-joining method (most phylogenetic software), and the more complex maximum likelihood approach (RAxML https://github.com/stamatak/standard-RAxML, and PhyML http://atgc.lirmm.fr/phyml/). More recently, Bayesian approaches to estimating phylogenetic relationships have become popular as computation technology has improved. Examples include BEAST (http://beast2.org/), MrBayes (http://mrbayes.sourceforge.net/), and BAPS (http://www.helsinki.fi/bsg/software/).
- SplitsTree and FigTree are examples of phylogenetic software that can calculate neighbour-joining or display trees produced by other software.

Utilities for clinical microbiology
- Species identification can be performed on WGS data by either 16 S characterisation, or by identifying short strings of DNA used in genome assembly (k-mer identification). Both options can be performed on the Danish Center for Genomic Epidemiology Java-based website http://www.genomicepidemiology.org/
- A number of other clinically useful tools are available on this site, including ResFinder for the detection of antimicrobial resistance, and Multi-Locus Sequence Typing. Command-line based tools such as BLAST using *de novo* assemblies, or SRST2 (https://github.com/katholt/srst2)[78] which uses read-mapping on sequencing reads, are better suited to automation, batching of multiple sequence analyses, and incorporation into analysis pipelines.

Databases
- NCBI GenBank (http://www.ncbi.nlm.nih.gov/genbank) and Genome Bank (http://www.ncbi.nlm.nih.gov/genome)
- European Nucleotide Archive (http://www.ebi.ac.uk/ena)
- DNA Databank of Japan (http://www.ddbj.nig.ac.jp/)

Typing databases
- MLST database (http://www.mlst.net/databases/

Antibiotic resistance gene databases
- ARG-ANNOT (http://en.mediterranee-infection.com/article.php?laref=283&titre=arg-annot-)
- ResFinder (https://cge.cbs.dtu.dk/services/data.php)

Multifunction bioinformatic suites
- Geneious Pro (http://www.geneious.com/)
- CLC Genomics (http://www.clcbio.com/products/clc-genomics-workbench/)
- Bionumerics (http://www.applied-maths.com/bionumerics)
- Nesoni (http://www.vicbioinformatics.com/software.nesoni.shtml)
- Harvest (https://github.com/marbl/harvest)
- Galaxy (http://galaxyproject.org/)

A more extensive list of software can be found at http://seqanswers.com/wiki/Software/list.

PERSPECTIVES ON GENOMICS IN CLINICAL MICROBIOLOGY

We recently conducted a qualitative online survey of infectious diseases physicians, microbiologists and other professionals involved in the management of infectious diseases on attitudes towards bacterial whole genome sequencing in Australia and New Zealand. Of 102 respondents, 74% were either clinical microbiologists or infectious diseases physicians, with the remaining 26% either infectious diseases/microbiology trainees

(23%) or research-based professionals (3%). Respondents were predominantly based in Victoria (34%) or New South Wales (25%), although a number represented Western Australia (9%) and New Zealand (8%).

Of respondents, 32% had some prior involvement with WGS, although only 24% reported local capacity to perform WGS. The Illumina MiSeq and the Ion Torrent PGM were the only sequencers used. Although subject to survey bias, all respondents indicated that they thought WGS would be useful in clinical microbiology in the next 5–10 years, primarily for epidemiological surveillance typing, clonality testing for outbreak investigation, and for detection of antimicrobial resistance. Due to concerns about the cost of implementing and conducting WGS, and the current lack of expertise in WGS and bioinformatics, most respondents thought WGS would be most likely used in reference laboratories, tertiary hospital laboratories and research laboratories. However, 83% thought that WGS would be used at least once per month in their laboratory over the next 5–10 years.

Although a qualitative study, this survey indicates a perceived utility of WGS in clinical and public health microbiology, with realistic anticipation that in the current economic climate, this will only be feasible in reference laboratories and large tertiary hospitals. As others have alluded to,[21,22] this model, with a few peripheral nodes and a centralised hub for WGS, would help to facilitate national/international collaboration and standardisation.

WGS IN CLINICAL AND PUBLIC HEALTH MICROBIOLOGY: HOW CAN IT HELP?

Structural and functional genomics

One of the primary investigation tools in microbial research is the use of genomics to characterise an organism, including identification of the genetic elements that may result in pathogenicity, survival, or antimicrobial resistance. As with human genetics, microbial genomics has the capacity to interrogate organisms for key genetic markers that may influence treatment and prognosis of infections. Currently, there are four main potential applications of WGS for bacterial pathogen characterisation in the diagnostic microbiology laboratory: identification, typing, resistance detection, and virulence gene detection.

Identification

Previous studies have illustrated proof-of-concept applications using next-generation sequencing for bacterial identification.[23–28] Given the current costs of sequencing, this is unlikely to surpass current methods such as matrix-assisted laser desorption ionisation-time of flight (MALDI-TOF) for routine bacterial detection for standard isolates. However, WGS may play a key role with organisms that are unable to be identified using routine methods. This includes organisms that often undergo methods such as 16S rDNA sequencing or specific nucleic acid probes to confirm identification, such as *Nocardia* and non-tuberculous mycobacteria, and organisms that are not usually or unable to be readily cultured. A recent report of neuroleptospirosis diagnosed through next-generation sequencing where conventional tests were non-diagnostic highlighted this potential role in diagnostic microbiology.[23]

Typing

Typing of bacterial pathogens for epidemiological surveillance, infection control and outbreak investigation is a more obvious and immediate application of WGS. There are numerous traditional typing methods for several key organisms that are generally performed in centralised reference laboratories, although occasionally will be performed for a specific purpose in routine diagnostic laboratories. For example, surveillance typing of *Listeria monocytogenes* for outbreak monitoring has been previously performed by a number of methods, including serotyping, binary typing, ribotyping, multilocus variable number tandem repeat analysis (MLVA), pulse-field gel electrophoresis (PFGE), and multi-locus sequence typing (MLST). Such diversity of methods with different resolution power is difficult to maintain and may hinder rather than enhance strain comparisons. Furthermore, typing is organism specific and requires constant validation. In contrast, WGS has the capacity to supersede traditional typing methods, through either *in silico* typing, or superior discriminatory capacity.[29,30] For instance, MLST, which is traditionally performed by sequencing of a set of housekeeping genes, can be simulated by mapping WGS reads to the reference sequences of those genes,[31] or using the Basic Local Alignment Search Tool (BLAST) to identify the alleles of the housekeeping genes.[32] The role of WGS as a superior method to typing for epidemiological surveillance and outbreak investigation is described in the 'Comparative genomics' section below.

Resistance detection

There are also potential applications for WGS to assist with antimicrobial resistance detection. A few studies have attempted to validate the accuracy of WGS for predicting antimicrobial resistance, with reasonable concordance.[20,33–35] Current analyses using WGS data can readily detect acquired resistance such as beta-lactamases and aminoglycoside modifying enzymes, although characteristic mutations in critical genes such as *rpoB* can also be detected with prediction of resistance phenotypes. However, these methods are currently unable to reliably predict some resistance mechanisms, for example, resistance resulting from derepression of *ampD*, *ampR* and other regulatory genes of *AmpC* hyperproduction, or vancomycin heteroresistance conferred by mutations in the complex regulatory system that includes *graRS*, *vraSR*, *walKR*, *agr* and *rpoB*. Although current susceptibility methods from organism culture are likely to be more rapid and reliable for routine testing, as with organism identification, WGS methods may be useful for slow-growing organisms, organisms that are unable to be cultured, or where phenotypic susceptibility testing is unreliable, e.g., clarithromycin susceptibility testing for *Mycobacterium abscessus*. For example, WGS was used to rapidly diagnose a case of extensively drug-resistant (XDR) *Mycobacterium tuberculosis*, reducing time to diagnosis from weeks to days, subsequently reducing exposure to ineffective drugs and minimising risk of *de novo* resistance.[36]

Virulence profiling

The other main potential use of WGS data for organism characterisation is detection of genetic markers of virulence, such as Panton-Valentine leukocidin (PVL) in *Staphylococcus aureus*, or Shiga toxin in *Escherichia coli*, although this still remains investigational due to the uncertainty in gene expression and significance of gene presence.[37]

Comparative genomics

The emergence of WGS as a universal replacement for traditional bacterial typing has unveiled its potential as a

powerful tool for epidemiological surveillance of bacterial pathogens, one of the cornerstones of infection control. Although largely performed in research environments, several studies have illustrated the capabilities of WGS to describe the evolution and epidemiology of important infections.[38–46] In an era of increasing antimicrobial resistance, mapping the epidemiology of such multidrug resistant infections to direct public health responses and antimicrobial prescribing practices is vital. In addition to tracking resistant organisms, WGS allows tracking of specific resistance mechanisms, including motifs on mobile genetic elements such as plasmids and elucidation of mechanisms of gene transfer.[47,48] For example, Wright *et al.* demonstrated that patients can be colonised with multiple strains of *Acinetobacter baumanii* capable of interacting within the patient, and that movement of patients and staff between healthcare facilities contributes to strain mixing and diversification.[49]

There have been numerous studies reporting the use of WGS to inform hospital infection control responses to suspected pathogen transmission.[50–58] In 2012, investigators from the US National Institutes of Health (NIH) used WGS to track a suspected outbreak of carbapenem-resistant *Klebsiella pneumoniae*, identifying a single patient as the source for three independent transmission events.[58] Another example was the paradigm-shifting evidence from WGS that multidrug-resistant *Mycobacterium abscessus* subspecies *massiliense* was frequently transmitted between patients with cystic fibrosis, prompting reconsideration of infection control measures.[50] The high discriminatory capacity of WGS has promoted it as the new gold-standard method for strain comparison, surpassing more traditional typing methods for inferring disease transmission and providing one of the strongest arguments for the use of WGS in clinical and public health microbiology laboratories.

Comparative genomic studies have also attempted to clarify transmission events and outbreak propagation. These methods relied upon established 'molecular clocks' to estimate the time to the most common recent ancestor and dates of presumed transmission events, using phylogenomic models.[59] Some defined thresholds for the number of SNPs between independent isolates that are required to infer whether they are epidemiologically linked,[43,60] although mutation and recombination rates vary between species and lineages,[50] and the rates of microevolution of endemic clones may need to be defined in each context.

Culture-independent identification and metagenomics

As alluded to above, WGS has been demonstrated to be a useful tool as a culture-independent method of bacterial identification, predominantly through metagenomic analyses. Although it is yet to be implemented in routine diagnostics, metagenomics involves sequencing all DNA content in a clinical sample, before using bioinformatic analyses to filter out human and non-pathogenic organism DNA to identify the causative agent. Due to the extensive depth of sequencing required for species identification, metagenomic investigations performed on low diversity sterile site samples are likely to produce a greater yield of results, in comparison to high diversity samples such as faeces. High quality samples with sufficient concentrations of genomic nucleic acid, such as tissue or fluid aspirates, are paramount for this application of WGS.

Previous methods including broad-range 16S rRNA PCR and sequencing have been used for diagnosis of culture-negative bacterial infections.[61] However, these methods frequently had low sensitivity if insufficient pathogen DNA was present, and were affected by the presence of contaminating DNA from other bacterial species. Metagenomic analysis of WGS data from a clinical sample has the capacity to overcome these limitations by filtering out unwanted DNA in the post-sequencing analysis. Sensitivity is also potentially greater, as organisms can be identified from a number of different segments of DNA, rather than a specific target segment which may have been altered or fragmented in the pre-testing process.

Aside from research studies on the human microbiome, the other potential application of metagenomics is in novel pathogen discovery. Although it has been successful for identification of some pathogens,[62,63] further testing is required to confirm the validity of novel genomes discovered by next-generation sequencing.[64]

Overall, metagenomic investigations remain experimental as sequencing technology and bioinformatic software to process and analyse metagenomic data is only just emerging.

CONSIDERATIONS FOR IMPLEMENTATION

Although WGS appears promising as an addition to the armoury of tests that are currently used in clinical and public health microbiology, it is yet to be widely implemented. With the significant improvements in cost and ease of sequencing, it is likely that WGS will supersede other molecular technologies including PFGE, MLST, DNA microarray and 16S rDNA sequencing in the near future, if not already. However, until costs become negligible, it is unlikely to be adopted for routine bacterial investigation over standard microscopy and culture, MALDI-TOF identification, and phenotypic antimicrobial susceptibility testing. Currently, the most immediately feasible applications are typing, epidemiological surveillance, and outbreak investigation to inform infection control procedures, most applicable to public health and tertiary hospital laboratories, although this would need to be matched by appropriate bioinformatic expertise. This situation is dynamic, and likely to change as sequencing technology evolves further and knowledge of bioinformatics develops.

Limitations of WGS

There are limitations of WGS that should be recognised. At present, the majority of analyses are based upon single nucleotide variants or SNPs identified from comparisons to a reference genome sequence. Consequently, analyses are dependent on the quality of sequencing and genome assembly, as well as the quality and selection of the reference genome. As current comparative analyses based on SNPs selectively exclude a significant proportion of phylogenetic data, some bioinformaticians have suggested conducting phylogenetic analyses based on all loci in a genome, rather than limiting the analysis to SNPs.[65] However, the significant requirements in computation resources and time would render such analyses unusable in a clinical environment. In contrast to research applications, comparative genomic methods for clinical purposes should aim to utilise more time efficient estimations of phylogenetic relationships that may not be the most accurate approximation, but are of sufficient resolution and accuracy to inform clinical and public health decisions.

Although WGS data can be used to provide detailed genomic information, this does not necessarily translate into knowledge of gene expression and transcription. For example, the presence of *lukSF-PV* does not necessarily equate to PVL production

and/or clinically aggressive *Staphylococcus aureus* infection. Detection of post-transcription RNA with next-generation sequencing can be performed and may help detect gene expression or enzyme hyperproduction (e.g., *BlaZ* or *AmpC* beta-lactamases), but requires a separate RNA sequencing run and is unlikely to replace current phenotypic screening methods.

A major limitation of WGS in clinical laboratories is the lack of validation and utility comparisons in clinical studies. Although proof-of-concept studies are frequently published, sequencing methods and data analyses for clinical projects have been customised around selected pathogens. The development and standardised evaluation of WGS pipelines for clinical and public health laboratories would guide further widespread implementation and add much needed evidence to an emerging field.

Quality control and standardisation

As with all tests that are incorporated into diagnostic and public health laboratory workflows, there is a need for a rigorous quality control process and standardisation of testing. Such measures have not yet been established, and benchmarks for quality control are yet to be determined. What should be the standard acceptable run quality? How should the quality of a genome assembly be assessed? Based on Illumina sequencing, we would recommend a minimum quality score of >30 across the genome with a minimum depth of 30–50-fold (i.e., an average of 30–50 overlapping reads at any particular locus) to produce adequate quality sequence for clinical comparative genomics. Although these have not been validated as the optimal target sequencing metrics, they are consistent with the recommendations recently proposed by the Australian Public Health Laboratory Network.[66] We currently interrogate our read quality and genome assemblies manually, though this is an arduous process for a large number of sequences that might arise from an outbreak, and a standardised automated quality control process is required. Figure 2 summarises some of the key considerations when assessing quality of WGS analyses.

National and international standardisation is also required, particularly with comparative analyses. Decisions about which reference genome is selected, or whether a composite reference genome is used, or which typing method to correlate results with each organism, need to be made at an overarching level and accepted across sequencing sites nationally. However, inter-site comparative analyses based on core genome SNPs may also vary depending on the selection of test isolates. To address this, the Global Microbial Identifier (GMI) initiative requires individual sequencing sites to upload data to a centralised server, which updates real-time phylogenetic analyses with the addition of new strains. An alternative solution has been to use a globally standardised set of core genome housekeeping genes to perform MLST, or whole-genome MLST (wgMLST). This analysis involves typing based on approximately 2000 core genes for each species. At present, both wgMLST and GMI are internationally-driven subscription based services, and are yet to be globally accepted.

However, these methods can create ambiguity when inferring direct pathogen transmission. Although wgMLST will likely be useful as a high resolution typing tool, it excludes a large amount of potentially informative genomic information, such as phages, insertion sequences, and other mobile genetic elements that may indicate direct transmission. A centralised server that includes all strains and a constant reference genome in comparative analyses can generate a phylogenetic signal, but will frequently lack the resolution to infer transmission. As there is frequently substantial genetic diversity within a single species group, the number of core genes that are common to all strains within that group diminishes as the number of strains

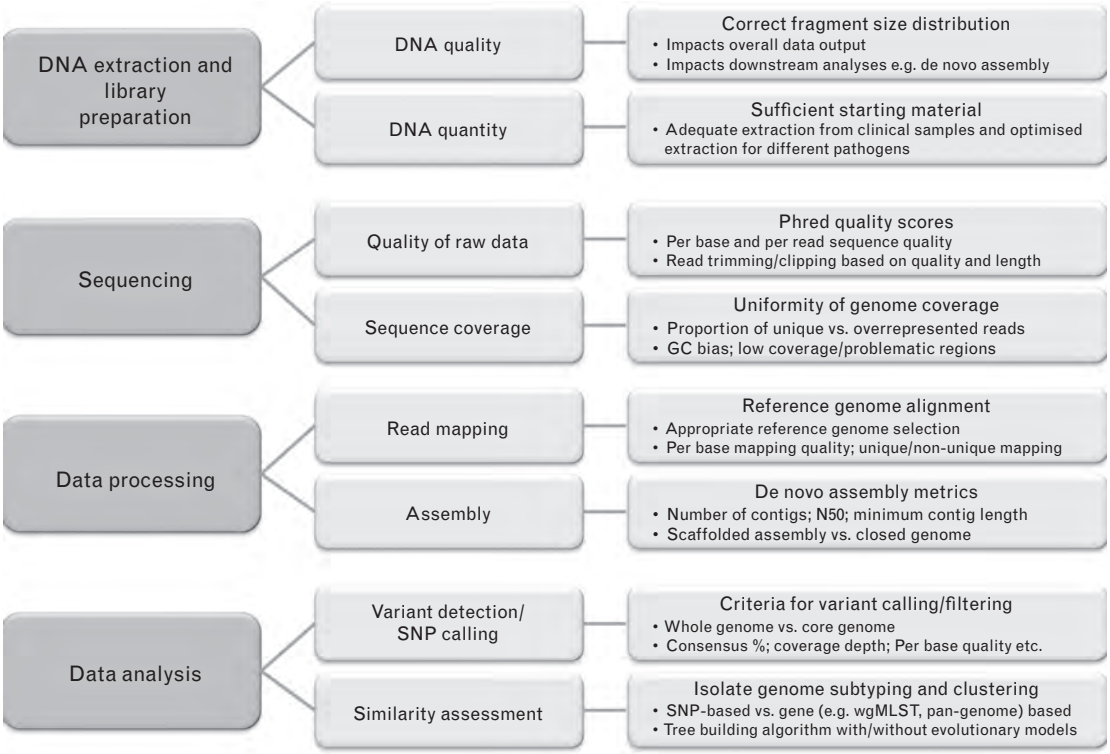


Fig. 2 Key considerations in quality assessment of whole genome sequencing analyses. Contigs, contiguous sequences; GC, genome coverage; SNP, single nucleotide polymorphism; wgMLST, whole genome multi-locus sequence typing.

increases. Conversely, core genome analysis of a clonal subpopulation of that species group can involve many more common genes, resulting in amplification of the number of loci compared, and frequently the number of SNPs identified as well. Thus, for identification of outbreak transmission, a reference genome should ideally be as similar to the outbreak strains as possible, if not part of the outbreak, and comparative genomic analyses should exclude outlying taxa that are not suspected to be part of the outbreak.

A further consideration is at what level of similarity are two isolates considered to be 'identical' or 'related' for inferring disease transmission. In theory, immediate sampling of an isolate that has been passed directly from Host A to Host B should result in identical whole genome sequences with no SNPs identifiable when the strains are directly compared to one another. However, this rarely occurs, due to the background mutation rate of organisms over time and the possible influences of sampling, storage and sequencing. Hence, it is difficult to pre-specify a threshold that defines an outbreak strain from one that is not. Within-host microbial genomic diversity has also been demonstrated for a number of pathogens including *Klebsiella pneumoniae*,[58] *Pseudomonas aeruginosa*,[67] *Burkholderia dolosa*,[68] and *Staphylococcus aureus*. In particular, studies of *S. aureus* have demonstrated genetic diversity of up to 40 SNPs between nasal isolates of the same spa-type and MLST group from a single patient,[69] as well as nasal carriage of multiple strain types.[70] Consequently, although appealing, establishing a fixed threshold of genetic divergence for an outbreak definition is difficult, and while comparative analyses of genomic data can support epidemiological investigations, they are not definitive.

Standardisation of data reporting is adopted to a certain degree for other microbiological tests. However, with the detail and complexity of genomic information, there are likely to be a limited number of clinicians who are able to fully comprehend and interpret all the data from a detailed analysis. Analogous to continuous data variables, results may not be able to be conveyed as a simple dichotomous 'yes' or 'no' report. For example, in an outbreak analysis, two isolates that differed by 100 SNPs might not be considered related if other clustered isolates were differing by 10 SNPs. However, if 90 of the 100 SNPs were co-located within a small genomic segment suggestive of a recombination event, this may still represent evidence of transmission. Lengthy technical descriptions of possible interpretations are unhelpful, and there is a need to develop standardised plain language reports, however there are no studies to guide reporting methods.

Resource and infrastructure requirements

The relative financial costs of WGS are discussed above. Additional costs for other equipment such as a high-end fluorometer for assessing DNA quantity, and a bioanalyser to assess DNA quality should be factored in. Fully automated systems from DNA extraction, sample and library preparation and sequencing have been marketed, although are rare. These systems were developed to maximise efficiency and costs of sequencing, but are geared more towards genomics reference centres.

The post-sequencing bioinformatic costs can also be significant. The emergence and accessibility of next-generation sequencing has resulted in an exponential increase in the amount of data generated from sequencing. The Illumina MiSeq is able to generate up to 15 gigabytes of raw data every 3 days before post-sequencing processing, while the HiSeq can generate up to one terabyte of data every week in high output mode. Although some clinical microbiology laboratories have previously invested in capacity for organism storage, physical sequencing data storage with backup needs to be considered with the implementation of WGS. Cloud-based storage options have been proposed for both workflow and data storage, however this may be impaired by bandwidth and data transfer capabilities. Handling of data confidentiality, security and integrity for these options also needs to be verified, though storage of sequencing data would seem easier than frozen organism storage.

Although many WGS analyses can be theoretically performed on standard desktop computers, the computational power required to process and analyse more than 50 genomes in a clinically actionable timeframe is considerable. The optimal specifications of this technology are beyond the scope of this review, and require local expertise.

Comparisons with human genomics

As with bacteria, human DNA has also been sequenced in research and clinical settings. While some parallels can be drawn between human and microbial genomics, there are a number of differences. The human genome is over 3 billion base pairs in size, a thousand times the size of the average bacterial genome. Somatic human cells are diploid with 23 pairs of chromosomes, while the majority of bacterial genomes, if not haploid in genome content, behave in a haploid manner. Several bacteria are polyploid organisms, e.g., *Neisseria gonorrhoeae*, although the significance of this has not be determined.

Clinical human genomics has focussed on identifying defined functional genetic mutations that result in disease. With the size of the human genome, high throughput next-generation sequencing has been used to perform targeted capture sequencing of exomes (the collective protein-coding regions of the genome) on large numbers of samples,[71] replacing previous DNA microarrays. Rapidly declining costs of sequencing and improvement in sequencing technology has resulted in greater utilisation of whole genome sequencing, providing more comprehensive human genome data.

With the significant number of short reads required to assemble the human genome, human WGS is more resource intensive. For example, on Illumina's NextSeq 500 in high output mode, a single run can theoretically sequence the entire genomes of up to 500 bacteria with an average genome size of 3 Mbp with 30× coverage. Alternatively, a single human genome or up to 10 exomes can be sequenced with similar metrics.

There are also differences in post-sequencing bioinformatics. Human genomics largely involves searching for defined mutations in specific segments of DNA, although other analyses such as genome wide association studies (GWAS) have gained popularity to search for genetic markers of disease. Assuming the most significant mutations occur in coding regions, attention can be focussed on exome analysis. Adequate coverage over these regions to ensure high sequence accuracy is critical. In contrast, SNPs in non-coding regions can still be informative to comparative bacterial phylogenomics. Detection of small indels is essential to identify mutations in the human genome, but is not always required for bacterial outbreak investigation. While detection of recombination in bacterial genomes is important to identify horizontal genetic exchange

and improve phylogenetic signal, such analyses are more of interest in the evolution of the human genome, rather than informing clinical human genetics.

Given these important differences, it follows that different expertise is required for human and bacterial genomics, and although there are similarities that may allow some sharing of certain resources such as a sequencing platform and reagents, each requires a different sequencing and bioinformatics configuration.

FUTURE DIRECTIONS

Whole genome sequencing has undisputed applications in research to enhance our understanding in numerous facets of infectious diseases and microbiology. Research into these aspects, including pathogen evolution, epidemiology and virulence determinants, and development and spread of antimicrobial resistance mechanisms, indirectly influences microbiology and clinical infectious diseases practices, and has the ultimate goal of improving patient care. Genomics is also increasingly being used in identifying potential drug and vaccine targets, and the increasing use of metagenomic analyses are starting to build our understanding of microbial ecosystems including the human microbiome.

There are still limitations that hinder widespread implementation of WGS in clinical and public health microbiology as a test performed in real-time to directly inform clinical practices. Even with rapidly improving sequencing efficiency, WGS is unlikely to surpass current methods for routine bacterial identification and antimicrobial susceptibility testing in the near future. It would seemingly have an expanding role in public health, reference and infection control laboratories for detailed isolate characterisation, outbreak investigation, and detection of disease transmission. As sequencing becomes more widely available and utilised, more routine use in diagnostic laboratories for pathogen identification in culture-negative samples, and metagenomic investigation of polymicrobial samples for pathogen and 'resistome' identification may be adopted.

It is clear that user-friendly bioinformatic pipelines are key to facilitating more widespread use of WGS, with more widespread bioinformatics expertise. Until this bottleneck is overcome, the most immediate implementation strategy is for centralised state reference laboratories to perform WGS and data analysis, with peripheral centres outsourcing to the reference laboratories as required. As uptake of WGS improves and costs decline, a more powerful epidemiological surveillance system could be established with sequencing performed at peripheral nodes, with bioinformatic analysis and oversight of the sequencing at the centralised reference centre. In an era of increasing drug resistance globally, ease of international travel, and little investment into antimicrobial drug development, utilising the few but powerful tools we have available to monitor and curb the spread of infectious diseases is paramount.

Acknowledgements: JCK is supported by a postgraduate scholarship from the National Health and Medical Research Council (NHMRC), Australia (APP1074824). BPH is supported by a NHMRC fellowship (APP1023526).

Conflicts of interest and sources of funding: The authors state that there are no conflicts of interest to disclose.

Address for correspondence: Professor Benjamin Howden, Microbiological Diagnostic Unit Public Health Laboratory, Department of Microbiology and Immunology, the University of Melbourne, The Doherty Institute for Infection amd Immunity, 792 Elizabeth Street, Melbourne, Vic 3000, Australia. E-mail: bhowden@unimelb.edu.au

References

1. Kuska B. Beer, Bethesda, and biology: how ''genomics'' came into being. *J Natl Cancer Inst* 1998; 90: 93.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977; 74: 5463–7.
3. Green ED. Strategies for the systematic sequencing of complex genomes. *Nat Rev Genet* 2001; 2: 573–83.
4. Staden R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* 1979; 6: 2601–10.
5. Fleischmann RD, Adams MD, White O, *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995; 269: 496–512.
6. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010; 11: 31–46.
7. Loman NJ, Misra RV, Dallman TJ, *et al*. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012; 30: 434–9.
8. Loman NJ, Constantinidou C, Chan JZ, *et al*. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* 2012; 10: 599–606.
9. Junemann S, Sedlazeck FJ, Prior K, *et al*. Updating benchtop sequencing performance comparison. *Nat Biotechnol* 2013; 31: 294–6.
10. Miyamoto M, Motooka D, Gotoh K, *et al*. Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics* 2014; 15: 699.
11. Quail MA, Smith M, Coupland P, *et al*. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012; 13: 341.
12. Liu L, Li Y, Li S, *et al*. Comparison of next-generation sequencing systems. *J Biomed Biotechnol* 2012; 2012: 251364.
13. Harris SR, Torok ME, Cartwright EJ, *et al*. Read and assembly metrics inconsequential for clinical utility of whole-genome sequencing in mapping outbreaks. *Nat Biotechnol* 2013; 31: 592–4.
14. Chin CS, Sorenson J, Harris JB, *et al*. The origin of the Haitian cholera outbreak strain. *N Engl J Med* 2011; 364: 33–42.
15. McCoy RC, Taylor RW, Blauwkamp TA, *et al*. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* 2014; 9: e106689.
16. Conlan S, Thomas PJ, Deming C, *et al*. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing *Enterobacteriaceae*. *Sci Transl Med* 2014; 6: 254ra126.
17. Blankenberg D, Coraor N, Von Kuster G, *et al*. Integrating diverse databases into an unified analysis framework: a Galaxy approach. *Database* 2011; 2011: . bar011.
18. Lazarus R, Kaspi A, Ziemann M, *et al*. Creating reusable tools from scripts: the Galaxy Tool Factory. *Bioinformatics* 2012; 28: 3139–40.
19. Gupta SK, Padmanabhan BR, Diene SM, *et al*. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother* 2014; 58: 212–20.
20. Zankari E, Hasman H, Cosentino S, *et al*. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012; 67: 2640–4.
21. Köser CU, Ellington MJ, Cartwright EJP, *et al*. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog* 2012; 8: e1002824.
22. Fricke WF, Rasko DA. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat Rev Genet* 2014; 15: 49–55.
23. Wilson MR, Naccache SN, Samayoa E, *et al*. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med* 2014; 370: 2408–17.
24. Hasman H, Saputra D, Sicheritz-Ponten T, *et al*. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J Clin Microbiol* 2014; 52: 139–46.
25. Loman NJ, Constantinidou C, Christner M, *et al*. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *J Am Med Assoc* 2013; 309: 1502–10.
26. Larsen MV, Cosentino S, Lukjancenko O, *et al*. Benchmarking of methods for genomic taxonomy. *J Clin Microbiol* 2014; 52: 1529–39.
27. Naccache SN, Federman S, Veeraraghavan N, *et al*. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res* 2014; 24: 1180–92.

28. Byrd AL, Perez-Rogers JF, Manimaran S, *et al*. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. *BMC Bioinformatics* 2014; 15: 262.
29. Bartels MD, Petersen A, Worning P, *et al*. Comparing whole-genome sequencing with Sanger sequencing for spa typing of methicillin-resistant *Staphylococcus aureus*. *J Clin Microbiol* 2014; 52: 4305–8.
30. Athey TB, Teatero S, Li A, *et al*. Deriving group A Streptococcus typing information from short-read whole-genome sequencing data. *J Clin Microbiol* 2014; 52: 1871–6.
31. Inouye M, Conway TC, Zobel J, *et al*. Short read sequence typing (SRST): multi-locus sequence types from short reads. *BMC Genomics* 2012; 13: 338.
32. Larsen MV, Cosentino S, Rasmussen S, *et al*. Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol* 2012; 50: 1355–61.
33. Stoesser N, Batty EM, Eyre DW, *et al*. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. *J Antimicrob Chemother* 2013; 68: 2234–44.
34. Gordon NC, Price JR, Cole K, *et al*. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. *J Clin Microbiol* 2014; 52: 1182–91.
35. Zankari E. Comparison of the web tools ARG-ANNOT and ResFinder for detection of resistance genes in bacteria. *Antimicrob Agents Chemother* 2014; 58: 4986.
36. Koser CU, Bryant JM, Becq J, *et al*. Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. *N Engl J Med* 2013; 369: 290–2.
37. Knobloch JK, Niemann S, Kohl TA, *et al*. Whole-genome sequencing for risk assessment of long-term Shiga toxin-producing *Escherichia coli*. *Emerg Infect Dis* 2014; 20: 732–3.
38. Leopold SR, Goering RV, Witten A, *et al*. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. *J Clin Microbiol* 2014; 52: 2365–70.
39. Howden BP, Holt KE, Lam MM, *et al*. Genomic insights to control the emergence of vancomycin-resistant enterococci. *mBio* 2013; 4: e00412–3.
40. Petty NK, Ben Zakour NL, Stanton-Cook M, *et al*. Global dissemination of a multidrug resistant *Escherichia coli* clone. *Proc Natl Acad Sci USA* 2014; 111: 5694–9.
41. Grad YH, Kirkcaldy RD, Trees D, *et al*. Genomic epidemiology of *Neisseria gonorrhoeae* with reduced susceptibility to cefixime in the USA: a retrospective observational study. *Lancet Infect Dis* 2014; 14: 220–6.
42. Chua KYL, Seemann T, Harrison PF, *et al*. The dominant Australian community-acquired methicillin-resistant *Staphylococcus aureus* clone ST93-IV [2B] is highly virulent and genetically distinct. *PLoS One* 2011; 6: e25887.
43. Eyre DW, Cule ML, Wilson DJ, *et al*. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. *N Engl J Med* 2013; 369: 1195–205.
44. Gire SK, Goba A, Andersen KG, *et al*. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 2014; 345: 1369–72.
45. Johnson JR, Tchesnokova V, Johnston B, *et al*. Abrupt emergence of a single dominant multidrug-resistant strain of *Escherichia coli*. *J Infect Dis* 2013; 207: 919–28.
46. Stinear TP, Holt KE, Chua K, *et al*. Adaptive change inferred from genomic population analysis of the ST93 epidemic clone of community-associated methicillin-resistant *Staphylococcus aureus*. *Genome Biol Evol* 2014; 6: 366–78.
47. Sivertsen A, Billstrom H, Melefors O, *et al*. A multicentre hospital outbreak in Sweden caused by introduction of a vanB2 transposon into a stably maintained pRUM-plasmid in an *Enterococcus faecium* ST192 clone. *PLoS One* 2014; 9: e103274.
48. Stoesser N, Giess A, Batty EM, *et al*. Genome sequencing of an extended series of NDM-*Klebsiella pneumoniae* neonatal infections in a Nepali hospital characterizes the extent of community versus hospital-associated transmission in an endemic setting. *Antimicrob Agents Chemother* 2014.
49. Wright MS, Haft DH, Harkins DM, *et al*. New insights into dissemination and variation of the health care-associated pathogen *Acinetobacter baumannii* from genomic analysis. *mBio* 2014; 5: e00963–1013.
50. Bryant JM, Grogono DM, Greaves D, *et al*. Whole-genome sequencing to identify transmission of *Mycobacterium* abscessus between patients with cystic fibrosis: a retrospective cohort study. *Lancet* 2013; 381: 1551–60.
51. Harris SR, Cartwright EJ, Torok ME, *et al*. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect Dis* 2013; 13: 130–6.
52. Long SW, Beres SB, Olsen RJ, *et al*. Absence of patient-to-patient intrahospital transmission of *Staphylococcus aureus* as determined by whole-genome sequencing. *mBio* 2014; 5: e01692–714.
53. Price JR, Golubchik T, Cole K, *et al*. Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* in an intensive care unit. *Clin Infect Dis* 2014; 58: 609–18.
54. Epson EE, Pisney LM, Wendt JM, *et al*. Carbapenem-resistant *Klebsiella pneumoniae* producing New Delhi metallo-β-lactamase at an acute care hospital, Colorado, 2012. *Infect Control Hosp Epidemiol* 2014; 35: 390–7.
55. Wendt JM, Kaul D, Limbago BM, *et al*. Transmission of methicillin-resistant *Staphylococcus aureus* infection through solid organ transplantation: confirmation via whole genome sequencing. *Am J Transplant* 2014.
56. Sherry NL, Porter JL, Seemann T, *et al*. Outbreak investigation using high-throughput genome sequencing within a diagnostic microbiology laboratory. *J Clin Microbiol* 2013; 51: 1396–401.
57. Reuter S, Harrison TG, Koser CU, *et al*. A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. *BMJ Open* 2013; 3: e002175.
58. Snitkin ES, Zelazny AM, Thomas PJ, *et al*. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci Transl Med* 2012; 4: 148ra16.
59. Lindsay JA. Evolution of *Staphylococcus aureus* and MRSA during outbreaks. *Infect Genet Evol* 2014; 21: 548–53.
60. Walker TM, Lalor MK, Broda A, *et al*. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med* 2014; 2: 285–92.
61. Rampini SK, Bloemberg GV, Keller PM, *et al*. Broad-range 16S rRNA gene polymerase chain reaction for diagnosis of culture-negative bacterial infections. *Clin Infect Dis* 2011; 53: 1245–51.
62. Palacios G, Druce J, Du L, *et al*. A new arenavirus in a cluster of fatal transplant-associated diseases. *N Engl J Med* 2008; 358: 991–8.
63. Bhatt AS, Freeman SS, Herrera AF, *et al*. Sequence-based discovery of *Bradyrhizobium enterica* in cord colitis syndrome. *N Engl J Med* 2013; 369: 517–28.
64. Naccache SN, Greninger AL, Lee D, *et al*. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol* 2013; 87: 11966–77.
65. Bertels F, Silander OK, Pachkov M, *et al*. Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Mol Biol Evol* 2014; 31: 1077–88.
66. Public Health Laboratory Network (PHLN). Ensuring national capacity in genomics-guided public health laboratory surveillance. 8 Jan 2015; cited 20 Jan 2015. http://www.health.gov.au/internet/main/publishing.nsf/Content/ohp-phln-pubs-genome-sequencing-report.htm
67. Smith EE, Buckley DG, Wu Z, *et al*. Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc Natl Acad Sci USA* 2006; 103: 8487–92.
68. Lieberman TD, Michel JB, Aingaran M, *et al*. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat Genet* 2011; 43: 1275–80.
69. Golubchik T, Batty EM, Miller RR, *et al*. Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. *PLoS One* 2013; 8: e61319.
70. Mongkolrattanothai K, Gray BM, Mankin P, *et al*. Simultaneous carriage of multiple genotypes of *Staphylococcus aureus* in children. *J Med Microbiol* 2011; 60: 317–22.
71. Ng SB, Turner EH, Robertson PD, *et al*. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009; 461: 272–6.
72. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008; 18: 821–9.
73. Bankevich A, Nurk S, Antipov D, *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012; 19: 455–77.
74. Chin CS, Alexander DH, Marks P, *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013; 10: 563–9.
75. Koren S, Harhay GP, Smith TP, *et al*. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol* 2013; 14: R101.
76. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* 2010; 11: 473–83.
77. Pabinger S, Dander A, Fischer M, *et al*. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* 2014; 15: 256–78.
78. Inouye M, Dashnow H, Raven L, *et al*. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* 2014; 6: 90.