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
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.

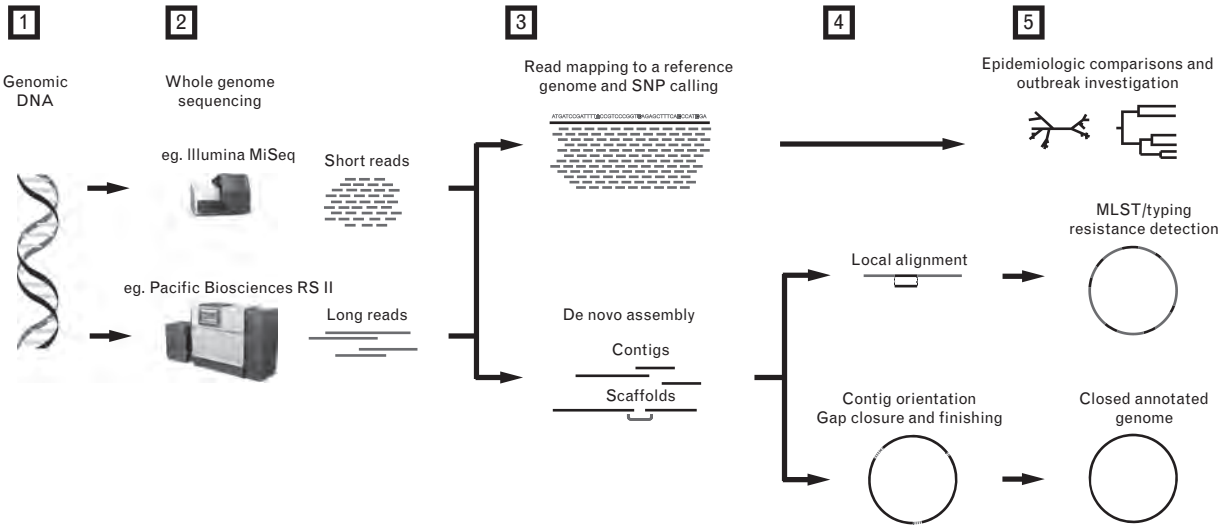

Fig. 1 Whole genome sequencing workflow. (1) DNA extraction from homogeneous microbial samples, e.g., single bacterial colony from a pure culture. (2) Whole genome sequencing using next-generation sequencers. Most high-throughput sequencers produce short reads (e.g., Illumina MiSeq), although long reads 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.