Horizon Scanning Series
The Future of Precision Medicine in Australia

Sequencing

This input paper was prepared by Professor Dave Burt, Dr Yuanyuan Cheng and Dr Ken McGrath

Suggested Citation

Burt, D, Cheng, Y, McGrath, K (2017). The Future of Precision Medicine in Australia: Sequencing. Input paper for the Horizon Scanning Project "The Future of Precision Medicine in Australia" on behalf of the Australian Council of Learned Academies, www.acola.org.au.

Sequencing

This paper was prepared by Professor Dave Burt, Dr Yuanyuan Cheng and Dr Ken McGrath

1. Abstract

For the last decade, short-read sequencing (e.g. Illumina) has dominated genomics research, providing rapid and economical ways to generate draft genomes and investigate variation in the genomes of individuals and populations. The information provided by short-reads is incomplete, but the technology is changing. Long-read sequencing (e.g. Pacific Biosciences) overcomes these limitations and produces complete coverage and assembly of genomes and transcriptomes, while tolerating genomic complexities such as high GC-content, repeat regions, phasing, and polyploidy. Advances in bioinformatics combine both types of data, and are further aided by long range physical mapping of DNA molecules (e.g. optical mapping e.g. \Bionanogenomics and Hi-c interaction maps, e.g. Dovetail Genomics), which compliment traditional genetic information (cytogenetics and genetic linkage). The greatest potential for long-read sequencing stems from advances in nanopore technology (e.g. Oxford nanopore), providing direct detection and real-time sequencing of single DNA/RNA molecules. The information it provides will simplify genome assembly, enable direct RNA sequencing and targeted assays. As this technology advances, it reveals that our limitation of understanding will likely be linked to the integrity of the DNA and RNA molecules themselves. The trajectory of nanopore technology is likely to deliver a simple, rapid, inexpensive system capable of sequencing intact chromosome-length DNA molecules, in real-time. This will enable direct *de novo* sequencing to define the genomes of individuals and populations.

2. Introduction

Since the successful assembly of the human genome there has been a rapid increase in our knowledge and understanding of the function and role of genetic variation of genomes in health and disease. This was made possible by parallel advances in both sequencing technologies and data analysis, but further progress is still constrained by limited throughput, data processing and high costs of sequencing on a population scale (rather than the original limits on individuals.

Since the release of high throughput sequencing platforms early in 21st century (next generation sequencing or NGS) there have been continued developments. These advances have been possible by incorporating novel technologies increasing read lengths, genome coverage up to entire chromosomes, at reduced cost. These advances were such that it now possible to use genome sequencing as a clinical tool in our fight against disease, both inherited and somatic.

However, moving from the research environment and into the clinic is not a simple task, and presents a whole set of new problems in addition to the limitations of current sequencing technologies (quality, costs, etc). Shorty read technologies are still cheaper, faster but are constrained by high error rates and short read lengths, making them still the method of choice for population screens rather than using more expensive, lower throughput long read technologies. In addition, there is also a reluctance to use genome sequencing rather than traditional cheaper and more specific technologies (Sanger, arrays, gene panels, etc.) which are often cheaper, faster and less computationally intensive. But of course, these old methods are very limited and can only diagnose 10-20% vs 40-60% of cases using NGS methods. The reluctance to change is also influenced by the initial high investment costs on some of these technologies. Goodwin et al (2016) provides a thorough review and comparison of the current technologies.

3. Short-Read Sequencing

For the last decade, short-read sequencing (also known as next-generation sequencing or NGS) has dominated genomics research, providing rapid and economical ways to generate draft genomes and investigate variation in the genomes of individuals and populations. Advances in NGS technologies have led

to a substantial drop in the cost of genome sequencing, bringing the price for generating a whole genome down to around AUD1000 (according to Australian Genome Research Facility, AGRF) by 2017, but we must also add in the bioinformatics cost to process and assemble the genome sequence. This has allowed the exponential growth in the number of genomes sequenced in the past decade, giving rise to 1000s of released animal and plant genomes, and 100,000s of human genomes (expected to increase to millions in next few years) providing unprecedented large data for investigating the genetic basis of complex biological traits and processes across diverse species. Also, remarkably, over 100K genome assemblies have been generated for a wide range of bacterial, viral, and fungal strains, including numerous pathogens of clinical and veterinary importance. These efforts have not only greatly improved our understanding of the microbial community inhabiting our body and the surrounding environments (through microbiome research), but also enabled the development of novel diagnostic solutions for infectious diseases that allow sensitive and rapid detection of variants and mutations (e.g. HIV, Fisher et al., 2015; TB, Pankhurst et al., 2016).

Among various short-read sequencing technologies, Illumina has been holding the largest market share of NGS instruments and services due to its wide range of platforms offering different levels of throughput to suit the needs of different applications, such as whole-genome sequencing, transcriptome sequencing, targeted gene sequencing, ChIP-seq, etc. One of the main advantages of Illumina systems attributes to the high data output, which enables sequencing of more samples at greater depth in less time; for instance, the HiSeq 4000 System can generate up to 1.5 Tb data per run, allowing the sequencing of up to 12 human genomes or 100 transcriptome samples in fewer than 3.5 days (manufacturer's data) and the latest Novaseq even more at 3 Tb per run and 48 human genomes. The resulting high sequencing coverage and depth coupled with high data quality contribute to a low base-specific error rate (<0.1%) in the consensus sequences (Bentley et al., 2008), which represents the greatest strength of Illumina (as well as many other short-read sequencing platforms) nowadays in comparison to the fast-growing long-read sequencing technologies.

However, short-read sequencing has several key limitations owing to the short-read length (up to a few hundred bases). Genome assemblies generated from short reads are commonly highly fragmented, with the genomic data comprising many gap-containing scaffolds (computationally assembled sequences) that are relatively small compared to the actual chromosomal DNA. These draft quality genomes are missing critical information on complex regions within the genome, such as those containing repetitive sequences or have undergone structural rearrangements (inversions, translocations, copy number variations, etc.), GC-rich regions, and centromeric and telomeric regions. Also importantly, genes with a complex structure or have a high degree of duplication are often under-represented or incorrectly assembled in draft genomes (Korlach et al., 2017). Similar problems extend to the transcriptomes produced with short reads, which can lack robust information on isoforms of gene transcripts. These issues associated with short-read sequencing undermine its suitability for investigating questions that rely on data accuracy, such as association studies aiming to pinpoint genes or genetic variants underlying disease susceptibility. Considering this, genomics research has seen a major shift towards long-read sequencing in the recent few years.

4. Long-Read Sequencing

Long-read sequencing overcomes the limitations of short reads and produces complete coverage and assembly of genomes and transcriptomes, while tolerating genomic complexities such as high GC-content, repeat regions, phasing, and polyploidy.

There are currently two main types of long-read technologies: (a) the synthetic type of approaches (e.g. Illumina TruSeq Synthetic Long-Read 'Moleculo' and 10x Genomics) utilises existing short-read sequencing platforms and employs barcodes to enable grouping and computational assembly of short reads from a large DNA fragment, generating up to 100 kb synthetic sequence length; and (b) single-molecule long-read sequencing approaches (e.g. Pacific Biosciences) fundamentally differ from short-read technologies in that they do not rely on clonal amplification and sequencing of small fragmented molecules, but instead produce a single continuous read spanning the entire length of a large template, preserving valuable information that can be lost through template fragmentation. The single-molecule approach can generate reads up to hundreds of kilobases without the need of computational assembly, allowing confident resolution of large

structural features in genomes and transcriptomes. For this reason, single-molecule sequencing platforms have become the most popular long-read sequencing platforms in recent years.

Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing system is one of the most established and widely used long-read technologies. Essentially, the system converts a double-stranded DNA molecule into a single-stranded circular template by attaching a hairpin adaptor at each end of the DNA; then the circular template is sequenced on a SMRT Cell chip, with each nucleotide passing through a DNA polymerase and emitting light signal multiple times. To ensure sequencing efficiency, template molecules are also subject to a size-selection process (usually around 20 kb), which eliminates molecules that are too small or too large. This allows the current SMRT platform to deliver half of data in reads longer than 20 kb, with a maximum read length of over 60 kb.

Many studies using PacBio SMRT sequencing have demonstrated the advantages of long reads over short-read sequencing in generating high-contiguity genome assembly, revealing complex genomic structural variations, haplotype phasing, and characterising full-length transcripts. SMRT sequencing of a human genome with 40x coverage helped closed 50 out of 164 gaps in the human reference genome, most of which were found to be enriched for repeats and GC-rich content (Chaisson et al., 2015). PacBio based resequencing of hummingbird and zebra finch genomes, which both have previously been sequenced on short-read platforms, improved the contiguity of each reference genome by 200 folds and 150 folds, respectively; the phased assemblies also helped correct formerly misassembled genes and inaccurately resolved phasing between haplotypes (Korlach et al., 2017). A recent study using targeted SMRT sequencing to genotype PKD1 gene in patients with autosomal dominant polycystic kidney disease (ADPKD) successfully overcame complexities of the gene due to its high GC-content and homology with multiple pseudogenes, demonstrating high sensitivity (94.7%) of the strategy for identifying patients carrying ADPKD-causing variants (Borràs et al., 2017). This example nicely illustrates the potential use of long-read sequencing in diagnostics, especially for diseases that are associated with complex genomic regions.

Despite the capability of PacBio sequencing in generating high-contiguity and high-accuracy genomic data, the cost remains a major limitation of this technology, making it less accessible for researcher working on whole genome analysis of large genomes e.g. in population studies. The SMRT sequencing system has a high error rate in single-pass reads (11% median) and requires higher than 30x coverage to achieve >99.999% accuracy (QV50; manufacturer data). For a human genome, this requires about 14 SMRT Cells (on the new PacBio Sequel instrument 7with Gb per run), costing more than AUD6K and the cost is coming down. As a solution to this problem, hybrid genome assembly approaches have been proven a cost-effective way to generate high quality reference genomes.

5. Advances in bioinformatics and Hybrid Approach

Advances in bioinformatics enable the combination of both short-read and long-read types of data, and are further aided by genome-wide physical mapping technologies, such as optical mapping (e.g. Bionano genomics) and Hi-C chromatin interaction mapping (currently main providers are Dovetail Genomics and Phase Genomics), which provide long range information for ordering sequences and assigning them at chromosome level. It has been shown that short-read sequences combined with long range mapping data can generate relatively inexpensive *de novo* human genome assembly with chromosome-scale scaffolding (Burton et al., 2013). Similarly, both optical and Hi-C mapping can massively improve PacBio long-read based genomes by doubling their N50 values (Jiao et al., 2017). The recently released domestic goat genome, reportedly the most continuous *de novo* mammalian assembly produced to date, was constructed using a sophisticated hybrid approach, integrating five data sets including PacBio long-read-based contigs, Illumina reads, BioNano Irys optical maps, Phase Hi-C scaffolding data, and a radiation hybrid map (Bickhart et al., 2017). After evaluating different scaffolding strategies, the authors reported that initial scaffolding of PacBio contigs with optical mapping data followed by Hi-C data yielded the most contiguous assembly with the lowest level of contig mis-orientation in comparison to the radiation hybrid map. The Illumina short-read data (23x coverage) played a critical role in assembly polishing, contributing to the identification and correction of over 1.0 Mb errors (insertions, deletions, and substitutions) within the assembly.

6. Nanopore Technology

The greatest potential for both long-read sequencing and native-detection of DNA modification stems from advances in nanopore technology, providing direct, real-time sequencing of single DNA molecules.

The key benefits of this sequencing are three-fold. Firstly, by allowing sequencing ultra-long (1Mb) regions of DNA and RNA, the bioinformatics overheads required for assembly of the data are greatly reduced, and in some cases, eliminated (http://www.biorxiv.org/content/early/2017/07/31/170373). Secondly, by directly detecting the native DNA and RNA molecules as they existed *in situ*, it enables epigenetic signatures of the DNA to be discriminated, from simple CpG methylation (https://link.springer.com/article/10.1007%2Fs00401-017-1743-5) though to rare or novel epigenetic modifications in both DNA (http://biorxiv.org/content/early/2017/04/13/127100) and RNA (http://biorxiv.org/content/early/2017/04/29/132274). Finally, the nature of direct detection coupled with little or no bioinformatics assembly provides a same-day timeframe from question to answer (https://link.springer.com/article/10.1007%2Fs00401-017-1743-5).

Nanopore structures can determine the sequence of DNA by measuring changes in electrical resistance of the pore. This can be in response to either the DNA moving through pore (e.g. Oxford Nanopore Technology), or coupled to probes/markers that move through the pore in an order determined by the original DNA sequence (e.g. Genia, Stratos Genomics and NABsys).

The current leaders of Nanopore sequencing technology are Oxford Nanopore Technologies (ONT), who launched their pocket-sized "MinION" sequencer in 2014. ONT uses a biological protein to create a nanopore through a synthetic membrane, across which electrical current flows. The electrical resistance of the DNA as it passes through the pore effects changes in the measured amperage across the membrane, which is in turn interpreted by a neural network to predict the most likely base that was introduced into the pore (http://www.nature.com/nbt/journal/v26/n10/full/nbt.1495.html). Such an approach has a high error rate compared to other platforms, although the per-base error has dropped significantly from around 40% at launch) to below 5%; however, with enough coverage of the target gene/genome, error rates fall to below 0.01% (http://www.biorxiv.org/content/early/2017/05/18/139840).

The compact size of the platform has led to its use in remote and extreme environments, including in Antarctica https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5362188/, zero-gravity (https://www.ncbi.nlm.nih.gov/pubmed/28725742) and even the International Space Station (http://www.biorxiv.org/content/biorxiv/early/2016/09/27/077651.full.pdf). This portability has the potential for a nanopore device to be used in point-of-care diagnostics, where rapid identification of a genetic element is required such as pathogen outbreaks (http://jcm.asm.org/content/early/2017/03/02/JCM.02483-16.abstract) , and time-critical oncology https://link.springer.com/article/10.1007%2Fs00401-017-1743-5) .

Several other companies are also developing their own versions of Nanopore sequencers e.g. Genia (Kumar et al., 2012), Stratos Genomics (https://www.stratosgenomics.com/), and NABsys (http://www.biorxiv.org/content/early/2017/05/18/139840). Each of these addresses the problem of high error rates on direct Nanopore sequencing by coupling the DNA sequence with a high signal-to-noise marker, that is then detected by the Nanopore rather than the DNA itself. These sequencing-by-proxy methods should dramatically increase base-accuracy of sequencing; however, it uncouples the sequence detection from the original DNA strand, and in the process removes the ability to detect epigenetic modifications.

Despite the exciting potential of Nanopore sequencing, there remain several barriers to adoption. The primary limitation to Nanopore sequencing is the high error rate (currently ~5%), in comparison to other platforms (e.g. Illumina, at ~0.1%). However, this is being address by the providers, and each iteration of the technology has seen error rates decrease.

Another limitation of Nanopore sequencing is the high "cost per base". This metric is used to compare the economy of sequencing genomes across various platforms. The $/bp on the MinION has dropped (primarily due to a large increase in date yields) and is currently ~$50 per Gb (https://nanoporetech.com/products#comparison), which is similar to the MiSeq platform, although the larger Illumina platforms can achieve costs of <$10/Gb. However, this metric fails to appreciate that the inherent value of a base as part of a long read is greater than the value of a base that is part of a short read.

As Nanopore technology improves, researchers are rapidly discovering that the upper limit on read-length is no longer the sequencing platform, but rather how the DNA is handled during extraction and preparation for sequencing. Even pipetting long DNA strands - that after extraction are no longer protected by their histone proteins - will fragment it to <100kb size ranges, reducing the potential for ultra-long reads (http://www.biorxiv.org/content/early/2017/04/20/128835). Advancements in both DNA extraction and library preparation techniques will be required to achieve the sequencing "holy-grail" of an end-to-end read of an entire intact chromosome.

Nanopore sequencing will continue to improve in quality and reduce in price. As the technology advances, systems will likely move away from biological, protein-based nanopores to solid-state nanopores such as graphene (http://www.nature.com/nature/journal/v467/n7312/abs/nature09379.html). This will enable industrial scale manufacturing at low cost, providing affordable, robust portable devices for rapid sequencing of DNA from any source. Not only will this integrate with POC diagnostic devices, but it will make its way into user electronics, providing mass-produced tools of citizen science.

The information Nanopore sequencing provides will simplify genome assembly, enable direct RNA sequencing and targeted assays. As this technology advances, it reveals that our limitation of understanding will likely be linked to the integrity of the DNA and RNA molecules themselves.

7. Gaps in Australia (and how this may compare internationally)

If the future of Precision Medicine is a clinical sequencing facility in each major hospital, key gaps are in funding and core accredited facilities around Australia. However, this is not a problem specific to Australia, it is world-wide but is expected to rapidly change over the next few years. Whole genome sequencing is a skill and resource mostly found in academic centres and not in the clinical setting, this needs to change.

8. Where the field is heading and what opportunities and challenges the next 10 years may bring for Australia

The trajectory of nanopore technology is likely to deliver a simple, rapid, inexpensive system capable of sequencing intact chromosome-length DNA molecules, in real-time. The speed of sequencing may require DNA molecules to be less than 1 Mb, to be assembled rapidly. Such long molecules can allow each paternal and maternal chromosome to be sequenced and thus provide haplotype data. This will enable direct *de novo* sequencing to define the genomes of individuals and populations. Direct sequencing of RNA is also likely to be more common in Precision Medicine, providing a direct read out of gene expression in tissue biopsies and blood samples allowing the quantitation of each gene transcript. Rapid sequencing devices are needed in a clinical setting, probably based on combining microfluidic devices to extract and deliver HMW DNA molecules to a nanopore sequencing unit. So, there are opportunities for new technologies, such as those being developed at the Australian Institute of Nanobiotechnology, QLD (AINB). Such systems will need to be linked and integrated with software to process and deliver sequencing data for DNA or RNA molecules.

To realize these ambitions, significant resourcing will need to be directed into coordinated training and infrastructure, allowing this new technology to run in parallel with existing procedures and then allowing existing procedures to be phased out where genomic analysis is superior. Training is required both for bioinformaticians to establish, validate and extend the pipelines for precision medicine, but also for clinical staff to ensure that they are fully informed to act appropriately upon the results. The harmonization of

platforms for data generation, data analysis, data access and secure data storage will facilitate monitoring at the national level.

9. Conclusion

Precision medicine has the capacity to make a major impact on the diagnosis and treatment of simple and complex genetic diseases, as well cancer and infectious diseases in Australia. To achieve this goal funding is required for resources, upskilling, technology development and increased integration between Academic Centres of Research Excellence and State Hospital and Health Services.

References

Bentley, D. R. and Balasubramanian, S. and Swerdlow, H. P. and Smith, G. P., et al. (2008) 'Accurate whole human genome sequencing using reversible terminator chemistry', Nature, 456(7218), pp. 53-59.

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., et al. (2017) 'Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome', Nat Genet, 49(4), pp. 643-650.

Borràs, D. M., Vossen, R. H. A. M., Liem, M., Buermans, H. P. J., et al. (2017) 'Detecting PKD1 variants in polycystic kidney disease patients by single-molecule long-read sequencing', Human Mutation, 38(7), pp. 870-879.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., et al. (2013) 'Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions', Nat Biotech, 31(12), pp. 1119-1125.

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., et al. (2015) 'Resolving the complexity of the human genome using single-molecule sequencing', Nature, 517(7536), pp. 608-611.

Fisher, R. G., Smith, D. M., Murrell, B., Slabbert, R., et al. (2015) 'Next generation sequencing improves detection of drug resistance mutations in infants after PMTCT failure', Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology, 62, pp. 48-53.

Goodwin, S., John D. McPherson, J. M., Richard McCombie, W. R. (2016) 'Coming of age: ten years of nextgeneration sequencing technologies' Nat Rev Genetics 17, 333-351.

Jiao, W.-B., Garcia Accinelli, G., Hartwig, B., Kiefer, C., et al. (2017) 'Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data', Genome Research.

Korlach, J., Gedman, G., King, S., Chin, J., et al. (2017) 'De Novo PacBio long-read and phased avian genome assemblies correct and add to genes important in neuroscience research', bioRxiv.

Kumar, S., Tao, C., Chien, M., Hellner, B., et al. (2012) 'PEG-Labeled Nucleotides and Nanopore Detection for Single Molecule DNA Sequencing by Synthesis' Sci Report, 2, 684.

Pankhurst, L. J., del Ojo Elias, C., Votintseva, A. A., Walker, T. M., et al. (2016) 'Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study', The Lancet Respiratory Medicine, 4(1), pp. 49-58.