PHG FOUNDATION

making science work for health

Clinical long-read sequencing

DNA sequencing is the process of reading part or all of the DNA, or genome, of an organism. In humans, the genome is over 3 billion DNA base pairs – or letters – in length, and it is not possible to read the whole genome in one piece. To sequence the genome, DNA is first broken into smaller pieces that are each read separately. Bioinformatic techniques compare these short sequences to a reference sequence and piece together this data, like a puzzle, into a continuous sequence.

Comparing the DNA sequence to the reference allows identification of changes to the sequence, known as variants. While many variants are benign, some can play a role in disease. Therefore, sequencing is being used increasingly in clinical genomics to provide a range of tests to inform patient care.

Advances in clinical genomics have been enabled by short-read sequencing (SRS) technologies, where the DNA is broken into pieces up to 300 base pairs long. While these technologies have many advantages, they also have some limitations. There is growing evidence supporting the value of long-read sequencing (LRS) – where DNA is read in longer segments, typically greater than 10,000 base pairs – to overcome some of these challenges. LRS technologies also have several inherent advantages for DNA and RNA sequencing, including epigenetic sequencing and direct RNA sequencing. While there remain barriers to implementation of LRS as part of a clinical service, this technology presents unique opportunities over established genomics technologies.

The essentials

- Long-read sequencing generates continuous sequencing reads from single molecules of nucleic acid (DNA or RNA)
- Long reads are more distinct, compared to shorter reads, reducing ambiguity when assembled together and unlocking previously inaccessible parts of the genome
- The two dominant long-read sequencing technologies are produced by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore)
- LRS has significant potential in clinical genomics, but this utility is yet to be sufficiently proven

WYNG FOUNDATION

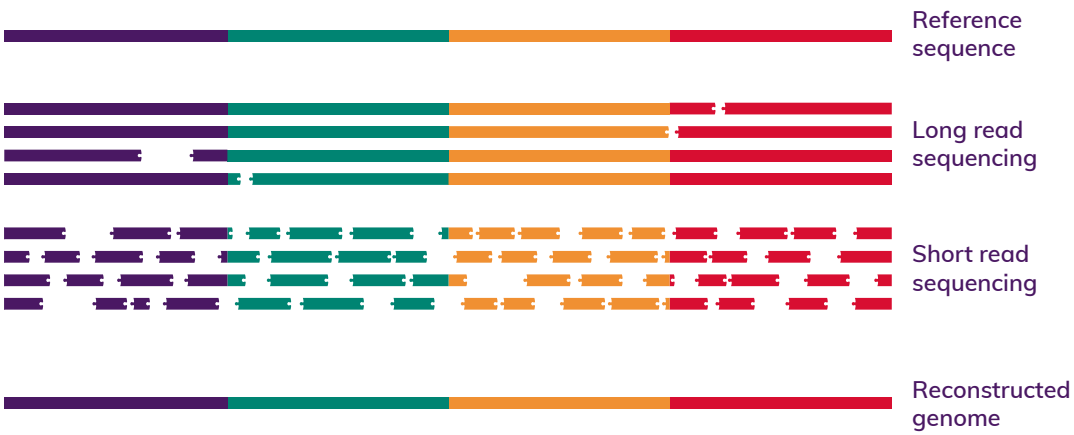
Sequencing in clinical genomics

The rapid development of SRS technologies has allowed ever-increasing volumes of high-quality sequencing data to be generated, in parallel with reduced sequencing costs. SRS technologies can be used to provide a wide range of tests from targeted panels through to whole genome sequencing. However, like a complex puzzle, assembling the genome from shorter pieces, known as reads, can be challenging. Many reads can look similar without additional context, for example, if a section of the genome is particularly repetitive.

Long-read sequencing (LRS) helps by producing data from much longer reads compared to SRS platforms. LRS typically generates read lengths of 10,000 – 100,000 base pairs compared to 75-300 base pair reads for SRS. Under specific conditions, sequence reads in excess of 4 million base pairs have been possible [1].

The process of putting sequence reads in the correct order is known as genome assembly (Figure 1). The advantages of longer reads are that they capture bigger sequences in the genome, are more distinct from each other, and simplify the process of reconstructing the original DNA sequence. For more complex regions of the genome, long reads provide more context, allowing the sequence to be reassembled with less ambiguity and error.

Figure 1. Mapping reads to the genome is a key step before any analysis can be done. Both short-reads and long-read sequencing produces data from fragments of nucleic acid and compares these reads to a reference sequence. Longer reads capture more information in one sequencing improving confidence during genome assembly [2].



Why do we care about long-reads?

Reference genome

The reference genome is the template against which sequencing data is compared to identify variants for analysis. Methods used to construct the reference genome have some limitations resulting in gaps and regions not completely resolved. LRS has been used in recent efforts to generate the first complete human reference genome to resolve these difficult to sequence regions in the genome [3].

Clinical long-read sequencing

Variant detection

Clinical genomics has historically prioritised interpretation – analysis of evidence for the role of a variant in disease – of small variants accurately detected using SRS. Small variants are defined as shorter than 50bp and include single nucleotide variants and small insertions or deletions. LRS is better at identifying larger (>50bp) and more complex types of variants missed or poorly characterised using SRS (Table 1).

Larger genomic variants have a bigger impact on the genome and are more likely to be disruptive, making them a priority in clinical genomics. For variant interpretation, this additional confidence from LRS helps to inform the precise nature of a variant and potential role in disease. Collectively, this will improve diagnostic yield in clinical genomics by expanding the range of variants included in analysis and improving overall detection.

Table 1. Summary of complex variants and how long-reads can improve detection and interpretation.

	What are they?	How do long-reads help?
Structural variants (SV)	Large genomic alterations classified as deletions, duplications, insertions, inversions and translocations resulting in different combinations of DNA gains, losses or rearrangements	LRS generates reads which span breakpoints providing additional context and certainty. This increases the accuracy and reliability of variant calling
Copy number variants (CNV)	A subtype of SV mainly resulting from deletions and duplications resulting in loss or gain of a section of the genome	LRS data produces more even coverage than SRS methods improving accuracy of CNV calling. Additionally, LRS enables identification of SV that cause these CNVs
Complex variants	Large and complex rearrangements arising from a catastrophic genomic event with multiple chromosomal breakpoints, resulting in significant genomic rearrangements	LRS can identify the exact locations of breakpoints and impact on gene function for more complete variant interpretation. This additional information can inform clinical management
Repeat expansion	A type of CNV, repeat expansions are unstable mutations arising from multi-nucleotide DNA sequence repeats that have repeated more times than normal	LRS can read entire sequence repeats to calculate the size of the expansion, including low complexity repeats poorly captured using SRS methods. This introduces bias into short-read sequencing data and will result in missed variants

Clinical long-read sequencing

Some regions of the genome are more difficult to sequence and assemble reliably using SRS, limiting variant detection in genes of known clinical significance (Table 2). Longer reads improve the accuracy of assembly by making the pieces look more distinct overcoming high degrees of variability.

Table 2. Types of regions within the genome where short read sequencing has had more limited utility and examples of diseases where long reads could improve variant detection.

	What are they?	Example
Complex loci	Regions of the genome that are highly repetitive or polymorphic limiting accuracy of variant detection	HLA locus is highly repetitive and polymorphic. LRS improves alignment in this region for more accurate variant interpretation
Low DNA nucleotide diversity	Nucleotide diversity refers to the relative proportion of all nucleotides. High diversity means equal proportions of nucleotides. Regions of low diversity have a higher proportion of either AT or GC resulting in amplification bias when using SRS	Amyotrophic lateral sclerosis (ALS) is caused by a G4C2 repeat expansion in *C9orf72*, which is difficult to sequence because of its size and high GC content. LRS has been used to quantify these expansion overcoming limitations of southern blots which cannot determine the exact nucleotide sequence [4]
Pseudogenes	Pseudogenes, which are imperfect copies of functional genes, have a higher mutation rate than their functional counterpart. There are an estimated 14,000 pseudogenes in the human genome	SRS reads can map ambiguously between these genes. Known disease genes have pseudogenes, for example, *GBA* linked to Parkinson's disease. LRS improves discrimination between these two genomic loci, improving confidence of variant calling

Capturing diversity

Variant detection becomes more challenging the greater the difference between an individual and the reference sequence. The current reference is a composite, made from the genomes of 20 individuals with a single individual contributing most of this sequence. LRS is now being used as part of efforts to develop a pangenome – a more sophisticated and complete reference representative of global genomic diversity. This, pangenome, has the potential to improve genome assembly across populations resulting in improved variant identification.

Clinical long-read sequencing

Genomic resources (i.e. genetic variant databases) are an essential component of genomic analysis to compare variants and inform their role in disease. LRS is now being used to expand the types of variants called in established cohorts (i.e. the 1000 Genomes Project, UK Biobank and deCODE) [5].

Most major genomic research databases are biased towards European populations, and cohorts more representative of global populations are needed. Improvements in these resources will improve diagnostic yields in clinical genomics by improving variant prioritisation and interpretation.

Unlocking 'omics

There is significant interest in the use of 'omics to complement the genomics landscape - that is, fields such as proteomics, metabolomics, metagenomics and transcriptomics. 'Omics have been proposed for a wide range of applications, but so far implementation has been limited. LRS has inherent advantages for 'omics approaches over SRS approaches.

DNA modifications

DNA modifications are changes to the DNA strand that do not affect the sequence, but can change how genes are expressed. Examples of DNA modifications include 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC) and 6-methyladenine (6mA). LRS methods directly sequence nucleic acids, and they can directly detect DNA modifications within raw sequencing data. In comparison, SRS uses amplification to generate sequence data, removing DNA modification as part of this process, and separate sequencing workflows need to be performed to retrieve this data.

LRS incorporates identification of some DNA modifications into base calling algorithms enabling simultaneous detection of methylation abnormalities alongside sequence data. This data could also be used to provide functional evidence when interpreting variants. For example, more than 50 epigenomic signatures have been associated with rare diseases and this could be used as evidence for variant interpretation [6].

Transcriptomics

Transcriptomics is the study of all RNA within a sample made up of individual RNA sequences, known as transcripts. Most research focuses on messenger RNA (mRNA) which is used by the cell for transcription to produce proteins. One gene may code for multiple different mRNAs, known as isoforms. LRS can sequence complete mRNA sequences and this method has been useful in research to expand our knowledge of different isoforms from genes.

Transcriptomics has been implemented in clinical genomics, for example, to detect fusion proteins – chimeric proteins created from the joining of two or more genes that originally coded for separate proteins - sensitive to targeted cancer therapies. However, SRS has limited sensitivity in comparison to LRS for detecting novel transcripts and fusion proteins.

Clinical long-read sequencing

Is LRS ready for use in clinical genomics?

It is only recently that LRS data quality has been considered sufficient for use in clinical genomics. LRS technologies have limitations compared to SRS that need to be addressed, including lower throughput and higher costs compared to SRS. In addition, tools and resources are still being developed to enable LRS data to be analysed more effectively.

In practice, SRS has been used successfully to improve clinical outcomes and LRS is unlikely to replace these tests. Rather, LRS should be considered to complement existing SRS and other genetic testing services.

Conclusions

In clinical genomics, sequencing has become central to many services enabling a wide range of tests. Limitations of short-reads has increased interest in the use of technologies that enable long-read sequencing. Longer-reads have inherent advantages for variant detection and these technologies also have the potential to unlock 'omics. There have been significant improvements in LRS technologies, although limitations remain, suggesting that it is time for clinical genomics to critically consider if these technologies may provide additional value to their services.

References

1. Oxford Nanopore Technologies Ultra-Long DNA Sequencing Kit. 2022; Available from: https://store.nanoporetech.com/ultra-long-dna-sequencing-kit.html.
2. Zhiao C and Xianghuo H. Application of third-generation sequencing in cancer research, 2021. Medical Reviews 1(2): 150-171.
3. Nurk S, Koren S, Rhie R et al. The complete sequence of a human genome, 2022. Science 376 (6588): 44-53.
4. Ebbert MT, Farrugia, SL, Sens JP et al. Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease, 2018. Molecular Neurodegeneration 13(1): 46.
5. Beyter D, Ingimundardottir H, oddson A et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structuRal variants in human diseases and other traits, 2021. Nature Genetics 53(6): 779-786.
6. Levy MA, McKonkey H, Kerkhof J et al. Novel diagnostic DNA methylation epigenetics signatures expand and refine the epigenetic landscape of Mendelian disorders, 2022. Human Genetics and Genomics Advances 3(1): 100075.

Author: Heather Turner

Published: November 2022

UNIVERSITY OF CAMBRIDGE