POLICY BRIEFING

phg foundation
making science work for health

What is long read sequencing?

October 2018

Emma Johnson
emma.johnson@phgfoundation.org

Sobia Raza
sobia.raza@phgfoundation.org

DNA sequencing – the process of reading part or all of the DNA of an organism – is helping to improve clinical care across different areas of medicine, from rare diseases and cancers, to the management of infectious diseases.

Progress has been accelerated by the advancement of high-throughput next-generation sequencing (NGS) technologies, which are capable of reading the code of millions of small fragments of DNA in parallel. These have enabled faster sequencing with increased throughput, at falling costs. In recent years, new technologies that are capable of sequencing longer strands of DNA by reading single DNA molecules, have advanced and become more prominent. This briefing explains what long-read sequencing (LRS) is, and how it differs from established short-read sequencing (SRS). The second, accompanying briefing, *Long-Read Sequencing: Ready for the Clinic?* describes the potential of these technologies for diagnostic sequencing in a clinical setting, and in this context the challenges with implementing the technology.

The essentials

- Single molecule, 'true' long-read sequencers enables the production of reads that are considerably longer than those resulting from SRS. This has several inherent advantages
- LRS can sequence parts of the genome that cannot easily be sequenced by short-read sequencing. Longer reads are more likely to look distinct compared to shorter reads, allowing them to be assembled together with less ambiguity
- The two dominant producers of 'true' long-read sequencing technologies are Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore)

UNIVERSITY OF CAMBRIDGE

POLICY BRIEFING

phg foundation
making science work for health

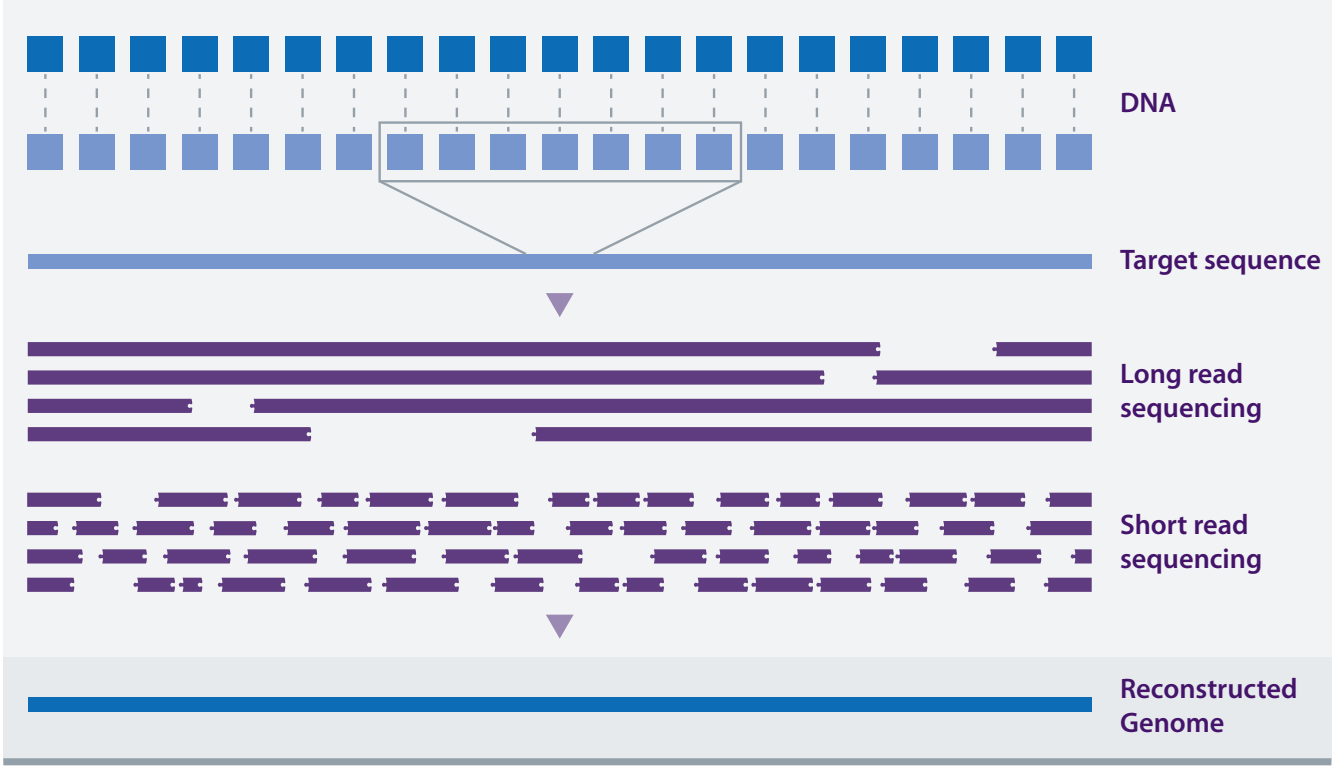
What is long-read sequencing?

The genome of most organisms (including humans) is too long to be sequenced as one continuous string. Using next-generation 'short-read' sequencing, DNA is broken into short fragments that are amplified (copied) and then sequenced to produce 'reads'. Bioinformatic techniques are then used to piece together the reads like a jigsaw, into a continuous genomic sequence.

LRS allows for the retrieval of much longer (>10,000bp) sequencing reads than widely-used SRS systems (75-300bp). Some long-read sequencing (LRS) platforms have produced sequence reads of 882,000bp [1], with some user groups reporting reads of over 2,000,000bp (2MB) [2]; however, read lengths of 10,000-100,000bp are more common.

True LRS technologies – sometimes referred to as third generation sequencers – directly sequence single molecules of DNA in real time, often without the need for amplification. This direct sequencing approach enables the production of reads that are considerably longer than those resulting from SRS. Other, 'synthetic' long-read sequencing approaches utilise modified sample processing and conventional SRS to computationally reconstruct long reads from shorter sequencing reads. True LRS represents the greatest departure from widely used short-read systems.

Currently, the two dominant producers of 'true' long-read sequencing technologies are Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore). Both have developed platforms for 'real-time' sequencing of nucleic acids (DNA and RNA) that is faster than current short-read technologies.



The benefits

There are several inherent benefits in using longer reads for the examination of genomic data; these can have advantages for clinical genome analysis.

- **Genome assembly:** The human genome is over 3 billion DNA base pairs in length and contains many repetitive stretches of genetic code. Like a complex jigsaw, reassembling the genome from short reads can be challenging, as many fragments look highly similar without additional context. Long-read data can make this task simpler as the reads are more likely to look distinct, allowing them to be assembled together with less ambiguity and error. Improvements in genome assembly are helping to close gaps in our knowledge of the genome and allow for a better understanding of the genetic causes of disease.

- **Variant detection:** Some features of individual genomes are particularly difficult to detect and quantify with SRS technologies, for example: large and complex rearrangements, large insertions or deletions of DNA, repetitive regions, highly polymorphic regions, or regions with low DNA nucleotide diversity. Long reads can span across larger parts of these regions, so are able to detect more of these variants, which may be clinically relevant. LRS may also enhance the 'genome-wide' detection of certain variants [3].

- **Haplotype phasing:** In areas such as reproductive medicine it can be useful to know whether genetic variants exist on the same copy of the chromosome. This can be determined using a process known as haplotype phasing. Long reads are able to provide the long-range information for resolving haplotypes without additional statistical inference, maternal/paternal sequencing, or sample preparation, as is required for an approximation of phasing using SRS.

Beyond producing long reads, true LRS technologies have other features that present new opportunities. Amongst these are:

- **Portability:** In contrast to other sequencing platforms, Nanopore's devices rely on detecting electronic rather than optical signals. This allows them to design devices as small as a memory (USB) stick, making them highly portable. Many other sequencers, including the vast majority of SRS systems, are large desktop or free-standing machines. Nanopore's MinION device has been used to sequence samples in the field during the Ebola and Zika virus outbreaks and has even been used in space.

- **Real-time sequencing and speed:** Compared to the fixed run times of SRS systems, both PacBio and Oxford Nanopore offer faster sequencing runs. PacBio provides options for rapid sequencing that can be completed in <24hours, from sample preparation to analysis. Nanopore technologies permit real-time analyses and allow experimental run time to be determined by the user, giving the user the ability to track data collection and begin analyses as desired. This provides additional flexibility and speed, and removes the need for batch sequencing of multiple samples which is currently required for cost-effective SRS. It is particularly useful when examining small genomes (such as those of many pathogens) or specific genomic regions.

- **Other 'omics:** Long-read technologies have been used to directly sequence RNA. They may also allow simultaneous detection of epigenetic modifications (chemical modifications to DNA/RNA that affect how genes are expressed), although additional bioinformatic interpretation is required. Separate sequencing runs need to be performed to retrieve this information using current SRS systems.

Conclusion

The inherent benefits of utilising longer reads for genome reconstruction and analysis, alongside the additional potential advantages true LRS systems present for genome analysis, could be beneficial for the diagnosis of several diseases and disorders. However, LRS systems also present their own challenges, and come with some limitations; this and their potential for use in clinical sequencing is discussed in the accompanying briefing.

Acknowledgements

We are grateful to Dr Shehla Mohammed reviewing this briefing, and Dr Sarah James for researching this topic.

Conflict of interest statement

PHG Foundation provides occasional analytical services to Oxford Nanopore Technologies (ONT). However, this briefing is the result of PHG Foundation's independent analysis and views, and is not linked to any third party. No external funding has been received to support the development of this briefing nor has ONT had any involvement in its preparation.

References

1. Jain M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol, 2018.
2. Payne A. et al. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. BioRxiv, 2018. **https://doi.org/10.1101/312256**
3. Stancun MC. et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Comms, 2017.