Assignment 1 – Bioinformatics Theory and Practice

Bioinformatic challenges of long-read nucleic acid sequencing technologies

Raman Chahal, 160067554
Word Count:1916

1 Introduction

In the case for most organisms, whose genome is very long, genome sequencing as one continuous string is not feasible. To overcome this problem, "next-generation" short-read sequencing (SRS) breaks the DNA into small fragments, which are subsequently amplified and sequenced to generate reads (75-300bp). Bioinformatic analysis then computationally pieces the reads together into a continuous genomic sequence.

However, long-read sequencing (LRS), a third-generation approach, allows for obtaining sequence reads >10Kbp, in some cases producing sequence reads >880Kbp (Jain M, *et al*., 2018) and >2MB (Payne A, *et al*., 2018). These long sequence reads are enabled by the real-time direct-sequencing of single molecules of DNA, without the need for amplification. This third-generation approach differs from synthetic long-read sequencing, which employs modified sample processing and conventional SRS to computationally construct long reads from short reads.

Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore) are currently the most established developers of true long-read sequencing technologies. Both companies have produced devices capable of real-time nucleic acid sequencing (DNA and RNA) quicker than current short-read technologies.

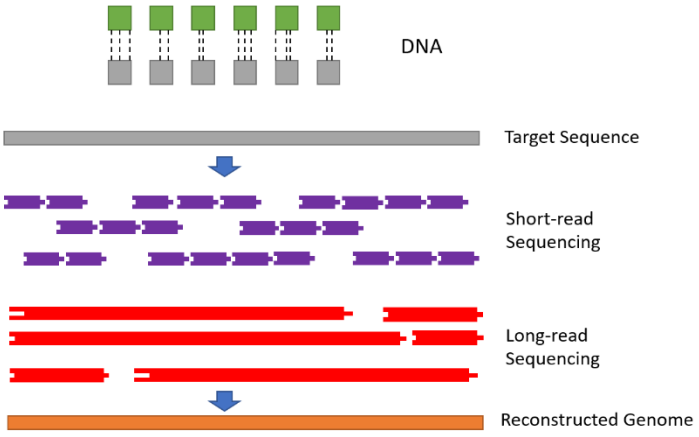


Figure 1. Visual representation of long reads vs short reads during genome sequencing. The main difference between LRS and SRS being the substantial increase in read length.

Long-read sequencing has some inherent benefits than short-read sequencing when examining genomic data. Since the human genome is very large (>3 billion DNA base pairs) and contains many repeating stretches of DNA, genome assembly using short-reads represents a huge challenge and can lead to complications, as there are high similarities between fragments without further context. Using long-read sequencing, this task is made easier as the longer reads are typically more distinct, allowing their subsequent assembly to be less ambiguous and carried out with fewer errors. These advances in genome assembly technologies help better our understanding of the genome and can beneficial in determining genetic causes of disease.

Sequencing Platform	General facts	Major applications	Challenges to Bioinformatics
Oxford Nanopore Technologies (MinION)	Each long read molecule has an average length ~10kb - ~1Mb; currently more expensive than SRS when factoring in data storage costs etc	De novo genome assembly, epigenetic modification detection, structrual variant detection and gene isoform resolution	Raw reads have high error rates, predominantly due to flase insertions. This is corrected by new alignment and error correction algorithms
PacBio Sequencing (SMRT)	Each long read molecule has an average length ~10kb - ~100kb; Much more expensive than conventional short-read sequencing	De novo genome assembly, epigenetic modification detection, structrual variant detection and gene isoform resolution	Raw reads have high error rates, predominantly due to flase deletions and homopolymer errors. This is corrected by new alignment and error correction algorithms

Table 1. Differences between sequencing platforms offered by Oxford Nanopore Technologies and PacBio are shown, and the bioinformatic challenges faced by the respective companies are outlined.

The SMRT (single molecule real time) sequencing approach used by PacBio utilises specialised flow cells, containing thousands of picolitre transparent wells called zero-mode waveguides (ZMW) (Levene, M. J. *et al*., 2003). DNA polymerase is fixed at the bottom of the well allowing the DNA strand to pass through the ZMW. A laser-camera records the colour and duration of emitted light caused by the incorporation of a dNTP at the bottom of the ZMW, each colour change corresponds to a specific base. The cleavage of the fluorophone bound to the dNTP by polymerase allowing it to diffuse out of the ZMW before a new labelled dNTP is incorporated. This method of sequencing also allows generation of a Circular Consensus Sequence (CCS), as a circular template allows multiple sequencing from the polymerase repeatedly traversing through the circular molecule.

Nanopore sequencing, as offered by Oxford Nanopore, does not monitor nucleotide incorporations guided

by a DNA temple. This method can directly detect the DNA arrangement of the native ssDNA strand, without interpreting light, or colour signals. For sequencing, DNA is translocated through a protein nanopore while a current is passed (through the pore) (Clarke, J, *et al.*, 2009). During translocation, a sensor measures changes in ionic current with constant sample frequency. Using machine learning, the changes in current data is used to obtain a consensus sequence (Jain, *et al.*, 2016).

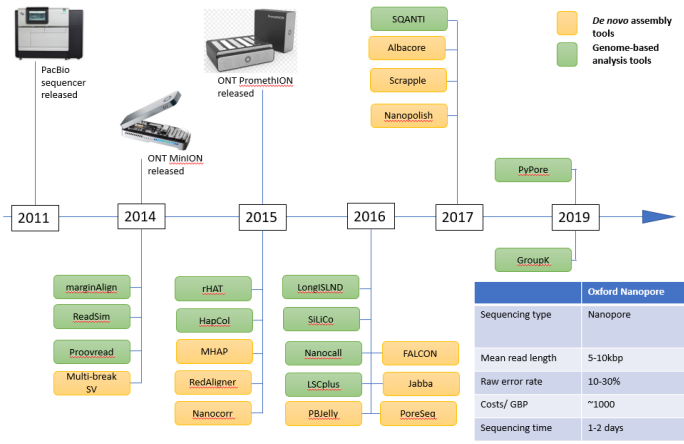
2 Bioinformatics tools for LRS and Bioinformatic challenges they face when using long reads

The data obtained using Oxford Nanopore's device MinION is integrated into Metrichor (EPI2ME), a cloud-based analytics company. By using cloud-based data storage, data interpretation workflows are automated, allowing real-time tracking, predicting and analysis of biological data. Oxford Nanopore also distribute additional analysis tools, such as MinKNOW and Guppy. Using data generated from these instruments could be useful in detection of DNA abnormalities in certain diseases, including complex structural variants, repetitive regions, regions that are polymorphic and vast deletions or insertions of DNA. By using long reads, which span a much larger area of these regions in contrast to SRS, variants such as these are easier to detect once the genome has been assembled.

There are two main types of genome assembly: *de novo* and reference-based genome assembly. *De novo* assembly is performed when there is no prior knowledge of the genome and is typically used for transcript detection with new introns or changed splice sites. Reference-based assembly requires prior knowledge, i.e. a reference sequence that is close to the sequenced data and is used to identify small mutations (indels) and single-nucleotide polymorphisms.

LRS genome assembly (*de novo* assembly) involves multiple steps: mapping of the initial reads, read error correction, correct read assembly and assembly polishing. Overlap-based algorithms, such as OLC (overlap-layout-consensus), are used to assemble the reads (Lannoy C, *et al.*, 2017). Alignments between the long reads are generated, before a consensus sequence of the contigs is made from the calculated overlap graph. The errors made in generating the consensus sequence are corrected by one of two ways. The first method aligns the long reads against themselves, whereas the second method has the long reads corrected by short reads (Lannoy C, *et al.*, 2017).

Bioinformatic tools developed for LRS allow for the quick and easy annotation, arrangement and navigation of large variant data sets from numerous platforms. Figure 2 illustrates the development of third-generation tools over the past 6 years.



	Oxford Nanopore	PacBio
Sequencing type	Nanopore	Single-molecule real-time (SMRT)
Mean read length	5-10kbp	10-15kb
Raw error rate	10-30%	10-15%
Costs/ GBP	~1000	~500+ (outsource)
Sequencing time	1-2 days	2-3 hours

Figure 2. **(A)** Timeline of released bioinformatic tools used for long-reads. The difference between reference-based assembly tools and *de novo* assembly tools are highlighted by green and yellow boxes respectively. **(B)** Table outlining differences between LRS offered by Oxford Nanopore and PacBio

For Oxford Nanopores PromethION device, a high-throughput instrument with higher capacity than their previous model MinION, the alignment tool RefAligner aligns and maps each molecule referentially using a dynamic programming algorithm (*de novo assembly*). This is done by determining the regions in a genomic sequence which are match the best (Mak AC, *et al.*, 2016). The matches are scored using the regional *in silico* nicking sites on the reference sequence and how the fluorescent labels are distributed.

Genome-based (reference-based) sequence analysis utilises different tools in comparison to *de novo* assembly. Figure 3. shows different tools used by Oxford Nanopore and PacBio (SMRT) for reference-based genome analysis.

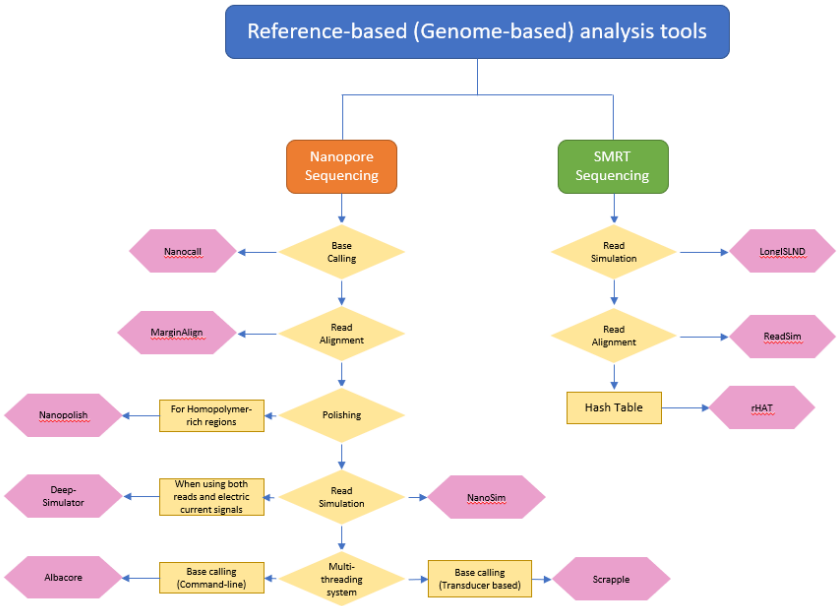


Figure 3. A decision tree for choosing a suitable sequencing analysis tool when using Nanopore or PacBio (SMRT) sequencing. When performing reference-based sequence analysis, decisions (as demonstrated by the diamond box) need to be made depending on sequencing type. For the reads used for read alignment in the SMRT platform, the read lengths need to be identified and use of a hash table needs to be determined. When using nanopore reads a decision must be made to determine if the reads should be performed for read alignment or base calling.

One of the biggest challenges when using long reads is accuracy. In comparison with SRS, the sequence accuracy of LRS depends on the quality of testing sample: >99% for individual base calls with SRS (Illumina Inc, 2014) compared to ~85%-90% with LRS systems (Mahmoud M. *et al*., 2017). However, the error rates can be reduced when performing multiple sequencing, or polishing the assemblies, to enable high consensus accuracy.

Data obtained using PacBio (SMRT) has lower coverage and greater sequence errors than current short read sequencing technologies, such as Illumina. During genome assembly, efficient detection of true overlaps requires specially designed algorithms. Presently, the need for improving small overlap detection, or detecting overlaps with errors in both reads, is high. Once these issues have been addressed, metagenomic assembly using LRS will be improved.

Existing tools for LRS genome assembly typically rely on two graph models: de Bruijn and overlap graphs. In cases of low error rates, using de Bruijn graph is theoretically more advantageous as the size of the graph does not increase substantially with the sequence coverage, which for Illumina datasets is typically high. As LRS has an inherently high error rate and low coverage, using the overlap graph is best suited to assemble the genome (Koren S, *et al.*, 2012). To create the overlap graph, read pairs which share overlaps need to be identified – indicating the sequencing of the reads occurs at the same loci in the genome. Despite the plethora of sequence alignment programs available for aligning the overlaps (Altschul SF, *et al.,* 1990) (Schwartz S, *et al*., 2003), the computational cost for high-throughput sequencing is too expensive. Short-read overlap detection software, utilising hash tables or Burrows-Wheeler transform, cannot directly be applied to LRS as the error rates are too high (Simpson JT, *et al.,* 2010) (Gonnella G, *et al*., 2012).

In a 2019 study by Nan Du et al, an overlap detection program (GroupK) was produced for LRS reads based on grouped short *k*-mer hits. By using short *k*-mer hits, oppose to longer hits used by other programs, the sensitivity for small overlap detection is increased. The research group are the first to use group hits to detect long read overlaps and have shown GroupK as an effective bioinformatic tool for sensitive overlap detection, particularly for datasets with low sequencing coverage. (Nan Du, et al., 2019)

Recent advances in Oxford Nanopore sequencing has increased the throughput of device such as PromethION to 10-20 Gb, allowing millions of reads to be generated. A research group led by Semeraro have developed a python toolbox called PyPore. For each completed read by Nanopore, all sequencing data is stored in FAST5 format. Currently, there is limited available software that can deal with FAST5 and facilitate downstream analysis from this file format (Legett, *et al.*, 2015). Pypore uses the raw FAST5 files to check quality of alignment to a reference genome and allows exploration of the sequence genome through generated HTML files.

The advent of bioinformatics tools such as Pypore will yield many opportunities for genome sequencing. The widespread use of ONT devices (MinION), predominantly due to low price (although the costs of LRS systems vary: Oxfords device MinION costs around $1000 for the starter pack, whereas PacBio Sequel costs $350,000.), simplistic sample preparation and inherent benefits to using long reads also depend on portability of nanopore sequencing devices. In contrast to the majority to SRS systems, which are typically large free-standing machines or desktops, MinION is designed as a small USB portable device. This was proven to be particularly useful during the Ebola and Zika outbreaks in 2016 and 2017 respectively. (Hoenen T, *et al*., 2016) (Quick J, *et al*., 2017)

These benefits have enabled smaller bioinformatics laboratories to sequence and analyse genomes by themselves. Although some features of Pypore are shared with other programs, several features such as G-C content estimation, interactive results sum-ups and plotting alignment modules make the program unique.
See supplementary figures for an overview of all the main tools currently used to complete each step in genome sequencing.

3 Biomedical applications of LRS systems

LRS has enabled researchers to identify a myriad of diseases, which result from genetic alterations that occur within genes, as well as within in non-coding regions. (Ashley EA, *et al*., 2016) Therefore, LRS has been quickly adopted by clinicians to tailor patient's treatment depending on their unique mutation profile (Ng *SB et* al., 2009) (Ashley EA, *et* al., 2010). Furthermore, long-read sequencing is used in haplotype phasing in reproductive medicine to determine whether genetic variants occur on the copy of the chromosome. Haplotypes can be resolved using long-range information provided by the long reads, without the need for further statistical analysis or sequencing of the maternal/paternal genome, which would be required if using short reads.

In a 2017 study by Stancun MC. et al, LRS, using MinION and NanoSV, was shown to enhance "genome-wide" detection of structural variants such as chromothripsis (Stancun MC, *et al.,* 2017). By using long-

reads, efficient genetic variation phasing enabled the researchers to determine chromothripsis breakpoint origins, helping resolve the complex rearrangement structure. Short-reads were shown to be unable to identify a large proportion of novel variants of inherited SVs, which were later shown as retrotransposon insertions.

Currently, thousands of patients are screened, by karyotyping or copy number profiling, for pathogenic structural variations. Despite these methods being cost-effective and robust, they lack the capability for small or copy-balanced SVs to be mapped, as well as lacking base-pair resolution accuracy. They also cannot resolve complex SVs (Alkan C, *et al*., 2011). Hence, MinION sequencing could be implemented by clinicians as a screening tool for patients expressing congenital diseases.

The computational strategy proposed by the researchers enabled accurate SV phasing straight from the nanopore reads, without the need for statistical intervention or suggesting phase based on sequenced family members (Hehir-Kwa JY, *et al*., 2016). Phasing SVs using reads is especially advantageous when there is a small population frequency for SV classes, and for *de novo* variations.

Supplementary figures

Nanopore base calling tool	Consensus quality score	Read quality score
Albacore	21.9	9.2
Scrappie	22.4	9.3
DeepNano	N/A	N/A
Guppy	23.0	9.7
Metrichor	N/A	N/A
Nanocall	N/A	N/A

Supplementary Figure 1. Base calling programs for nanopore sequencing are shown, where base calling is the process of transforming raw electric current signal data into a nucleotide sequence. The quality scores are based from a 2019 study by Wick RR (Wick RR, *et al*., 2019)

Aligning tools for long reads	Algorithm
BWA	Burrows-Wheeler Aligner's Smith–Waterman Alignment
Minimap(2)	Hash table
GraphMap	Gapped spaced seeds
LAST	Adaptive seeds
NGMLR	Smith-Waterman alignment and k-mer search
Kart	Divide and Conquer

Supplementary Figure 2. Long read nanopore sequencing aligners are shown with their respective algorithm.

Sequence assembly tools	Features
A Brujin	Used in *de novo* assembly for long reads
Nanopolish	Oxford-nanopore software package used for consensus sequence calculation
HINGE	Assembles long-reads using 'hinge' methodology
NanoPipe	Uses LAST alignment to calculate consensus sequence
PBJelly	Fills in gaps between aligned long sequence reads and high-confidence draft assembled sequences
SMART *de novo*	Assembles reads without error correction

Supplementary Figure 3. List of long sequence assembly tools and a brief description. (Makałowski, W *et al.,* 2019)

Variant detection tools	Features
Clair	Uses a neural-network for variant calling
Nanopolish	Oxford-nanopore software package used for SNP and indel mutation detection. Uses raw signal data.
NanoPipe	Uses LAST alignment to calculate consensus sequence
PBHoney	Uses long-reads to identify variants
Sniffles	Used to detect structural variations from long-reads

Supplementary Figure 4. List of variant detection tools and a brief description. All the tools apart from Nanopolish detect variants by comparison with a consensus sequence after base calling. (Makałowski, W *et al.,* 2019)

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. Levene, M. J. et al. Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299, 682–686 (2003).
4. Clarke, J. et al. Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol. 4, 265–270 (2009).
5. Jain, Miten et al. “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community.” *Genome biology* vol. 17,1 239. 25 Nov. 2016, doi:10.1186/s13059-016-1103-0
6. de Lannoy C, de Ridder D, Risse J. The long reads ahead: de novo genome assembly using the MinION. F1000Res 2017;6:1083.
7. Mak AC, Lai YY, Lam ET, et al. Genome-wide structural variation detection by genome mapping on nanochannel arrays. Genetics 2016; 202:351–62.
8. Understanding Illumina Quality Scores. San Diego, CA: Illumina, Inc, 2014; Pub. No. 770-2012-058
9. Mahmoud M. et al. Efficiency of PacBio long read correction by 2nd generation Illumina sequencing. Genomics, 2017. https://doi.org/10.1016/j.ygeno.2017.12.011
10. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012; 30(7):693–700.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10.
12. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. Human–mouse alignments with BLASTZ. Genome Res. 2003; 13(1):103–7.
13. Simpson JT, Durbin R. Efficient construction of an assembly string graph using the FM-index. Bioinformatics. 2010; 26(12):367–73.
14. Gonnella G, Kurtz S. Readjoiner: a fast and memory efficient string graph-based sequence assembler. BMC Bioinformatics. 2012; 13(1):82.
15. Du, N., Chen, J. & Sun, Y. Improving the sensitivity of long read overlap detection using grouped short k-mer matches. BMC Genomics 20, 190 (2019) doi:10.1186/s12864-019-5475-x
16. Leggett R.M. et al. (2015) NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. Bioinformatics, 32, 142–144.
17. Hoenen T, Groseth A, Rosenke K, et al. Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. *Emerg Infect Dis*. 2016;22(2):331–334. doi:10.3201/eid2202.151796
18. Quick, J., et al. (2017). "Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples." Nature Protocols 12: 1261.
19. Ashley EA. Towards precision medicine. Nat. Rev. Genet. 2016;17:507–522. doi: 10.1038/nrg.2016.86.
20. Ng SB, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461:272–276. doi: 10.1038/nature08250.
21. Ashley EA, et al. Clinical assessment incorporating a personal genome. Lancet. 2010;375:1525–1535. doi: 10.1016/S0140-6736(10)60452-7.
22. Stancun MC. et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Comms, 2017.
23. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 2011;12:363–376. doi: 10.1038/nrg2958.
24. Hehir-Kwa JY, et al. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. Nat. Commun. 2016;7:12989. doi: 10.1038/ncomms12989
25. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford nanopore sequencing. Genome Biology. 2019;20:129
26. Makałowski, W., Shabardina, V. Bioinformatics of nanopore sequencing. *J Hum Genet* (2019) doi:10.1038/s10038-019-0659-4