Genomics applications of nanopore long-read sequencing for small to large sized genomes

Liem, M.

— Chapter 6

Summary and discussion

Summary introduction

In this thesis I highlight the applications of Oxford Nanopore Technologies (ONT) sequencing. This technique is a relatively new approach in the sequencing field, where nanopores are embedded in a membrane, DNA molecules are pulled through nanopores and an electrical current serving as the sequencing signal. This technique yields reads-lengths of >10Kbp and has no theoretical upper limit towards read-length. The positive impact on data quality due to improved chemistry is underlined, improved chemistry leads to less sequencing errors and a more homogeneous coverage over complex genomic architectures. Benefits for increased read-lengths are assessed for resolving fragmented genome assemblies that were previously based solely on short-read sequencing data. Furthermore, the assembly of a large genome using ONT data is described, indicating ONT is a suitable candidate for resolving extremely large genomes using sophisticated assembly approaches. And finally, the potential for on-site sequencing is evaluated. Exploiting simplicity, mobility and accuracy provided by this new technique.

The central hypothesis of this thesis is that Oxford Nanopore Technologies long-read sequencing can be valuable for established genomics applications, such as whole genome sequencing (chapters 2–4) and metagenomic characterization of microbial communities (chapter 5). Here, I reconsider this general proposition in light of the results of the preceding chapters. In addition, I discuss the prospects for emerging and future genomics applications based on the possibilities presented by ONT data.

The quality of long- read sequencing and assemblies

ONT sequencing differs from traditional sequencing methods in the way that nucleotides are directly measured using electrical signals as opposed to synthetic copies or surrogate markers such as fluorescent labels. Multiple nucleotides (5-mers) occupy the pore shaft at the same time, hence it is the set of nucleotides that cause the electrical interference to determine the final profile. This profile signal needs to be, algorithmically, untangled to identify a single sequenced base. Therefore, basecalling algorithms yield the interpretable sequencing reads, and by improving basecalling algorithms sequencing data quality is subsequently increased even for previously sequenced projects[1]. ONT initially allowed 30 nucleotides per second to pass through the nanopore. The number of bases processed by a single pore was limited since basecalling algorithms struggled to differentiate nucleotides that passed through the pore too quickly, resulting in extremely low sequencing quality. Restricting the sequencing speed to 30 bases per second yielded ~70% accuracy. Currently ONT can process ~450 bases per second yielding reads >10Kbp with accuracy between ~90-99%. Chapter 2 highlights the effect of improved sequencing speed, basecalling and chemistry for a highly heterozygous yeast strain.

However, to generate accurate haplotypes for this genome additional sequence accuracy is required. From BUSCO analysis we observed genes that remained absent from our best assembly results. The difference in alignment hit, due to error introduction, is highlighted by the comparison of identified genes before and after error correction.

Where more genes are identified when sequence accuracy is increased. Compared to other studies, which use coverages ranging from 70x up to 1000x, our dataset has relatively low coverage. Hence slight coverage increase could aid in resolving any remaining assembly difficulties as well as increase sequence accuracy due to increased evidence for the error-correction procedure[2–4].

Since assembly algorithms struggle to define the ends of circular DNA, circularization for complex genomes remains a challenging task. In this study we have not investigated the architecture of mitochondrial DNA or circularization of plasmids. Hence it would be beneficial to subject the final assembly results to circularization software designed specifically for closing circular contigs from assemblers using long read data[2,5].

In chapter 2 and 3 we have evaluated a multitude of assembly, consensus calling and correction tools, which have performed anywhere from mediocre to promising. Most assembly strategies are comparable and result in relatively small differences. Highlighting the origin of those small discrepancies and deciding upon the final assembly is tedious and a labor-intensive matter. The currently available tools leave room for user-friendly workflows, including base level and genome wide visualizations. These workflows should report progress at alignment, assembly, consensus, and correction level to facilitate decision making for downstream analysis. Off-the-shelf assembly workflows would increase the speed at which genome analysis is performed, and reliefs investigations for large sequencing datasets. The current gold-standard is performing multiple assembly strategies and continue in a result-based fashion. Analysis tools for small to medium-size genomes show comparable yet still slightly different assembly results causing comparison between analyses to be extremely difficult[7,9].

De novo assembly results based on long-reads for small genomes show promising reconstructions. Assemblies for medium-large genome sizes of comparable quality, such as investigated in chapter 4, are increasingly publicly available. However, separated haplotypes of such organisms have yet to be uncovered since base-level quality has only recently become of sufficient quality to accurately phase chromosomal copies[10].

Despite increased capability towards evenly spread coverage, increased read-length and improvement towards low-complexity regions, for ultra large genomes additional development is required[6]. Sequencing ultra large genomes at routine level requires an additional developmental update particularly towards sequencing speed and cost. For instance, sequencing the genome of *Paris japonica*, a plant species with a genome size of unprecedented scale, estimated genome size ~150 Gbp for a single genome copy[8]. Sequencing a genome of this magnitude requires just under one hour on a fully loaded PromethION (that is 48 flow cells, each ~$2.000 and utilizing ~ 2.500 pores at 450 bases per second) for a single genome copy. Hence, although feasible, sequencing at the required sequencing depth for such genomes still takes days and is very resource intensive. Evaluating the distinct ONT improvements towards read-length, read-accuracy and throughput, ONT is a pioneer entering the truly large-genome research area[6].

The cost of genome sequencing

Evaluating the cost of genome sequencing using Moore's Law has made it clear that incredible amounts of sequencing data are going to be generated. These data volumes indicate the necessity of efficient downstream analysis software. Currently, sequencing data has become more affordable as opposed to costs for analyzing large datasets using computer clusters. The benefit of decreased cost and increased sequencing speed and throughput is lost when data analysis requires thousands of CPU hours on an expensive dedicated cluster. We therefore need to provide the scientific community with more sophisticated tools for processing large datasets, that are less computationally intensive, require less memory, are faster and more user friendly.

Sequencing anything anywhere

Standard lab technicians are not experienced with command line tools and do not possess the skills to adequately adapt to alternative results. This clearly present gap could be bridged by using standardized metrics and formats, easily accessible free yet sophisticated software that is backed-up with logical visual representations.

In line with the skewed relation between generating and analyzing data is the size of sequencing machines, currently the smallest sequencing device is just the size of a large USB stick and provides mobility to allow infield sequencing, discussed in chapter 5. However, infield generated data needs to be processed by computer clusters or at least a high-end laptop with sufficient energy supply. Fully exploiting this mobility characteristic requires downscaled processing power and memory consumption.

From amplicon to *in situ* metagenome sequencing and assembly

In chapter 5 we used metagenomics to identify the microbial diversity using ONT, which is a first step in understanding the oceans biocomplexity and ecology. However, to know which species thrive at which locations is only the start of understanding the ecology behind microbial diversity. To functionally assess microbial capabilities full genome assemblies are required, this would for example lead to increased understanding towards resistance mechanisms used by microbial communities to survive the harsh oceanic conditions or reveal the mechanistic property to interchange genomic content through plasmids. Additionally, it would highlight the diversification of species in a time and space fashion, enabling to monitor the health of oceans, seas and rivers that are the foundation of life on land.

To fully allow in-field monitoring of seawater, DNA isolation and library preparation methods need to be performed at location. In chapter 5 we have isolated the DNA under laboratory conditions. Although this procedure follows a very simple guideline, collecting high molecular weight DNA from marine organisms is particularly challenging due to excessive metabolites secretion that co-precipitates with DNA[11]. Hence optimization for high molecular weight DNA isolation regarding on site sequencing needs additional development towards speed and ease-of-use.

Additionally, isolated high molecular weight DNA requires library preparation to allow the sequencing device to bring molecules in proximity of sequencing pores and to read-out bases using an electrical current. Equipment for those preparations should meet desired requirements to be able for *in situ* use. Voltrax library preparation provides a potential solution and is able to prep isolated DNA in a matter of minutes, however, due to the lack of purification steps isolated high molecular weight DNA could be rather contaminated. Hence even with small and easy-to-use devices such as Voltrax, *in situ* DNA isolation and purification remains challenging[11].

Moreover, chemistry required for sequencing requires specific storage limitations; both flow cells and chemistry are temperature-sensitive and refrigerator capacity for in-field expeditions are usually inconsistent due to the lack of adequate power supply[12].

Finally, additional analysis is required to position identified species phylogenetically. Onecodex (used in chapter 5) is beneficial to place organisms quickly and easily into context of existing databases, easing time constraints and labor complexity. However, it lacks branch unity and cannot indicate the genetic distance between species and position them relative to each other. Furthermore, it only offers enhanced functionality using a paid license which adds to resource pressure and making it difficult for researchers to compare results. Previous studies show successful phylogenetic placement under remote conditions using JModelTest, hence this could be a potential candidate for downstream analysis on metagenome samples from seawater[13].

The future of Oxford Nanopore Technologies sequencing and its applications

With the use of the current best flow cells and chemistry sequence accuracies of Q20 are achieved, translating to >99% read accuracy after basecalling. These methods allow molecules attached to the nanopore to be unzipped and both single strand copies are pulled through a pore reading-out the base sequence. Although sequencing both separated single strands was already introduced by Oxford Nanopore Technologies in ~2015, it was later replaced by single strand sequencing chemistry. However, chemistries to sequence both separated single strands have recently been released again by Oxford Nanopore Technologies. Here the information of both single strands is utilized to reduce basecalling errors by combining the sequencing signals. As the double stranded molecule found its way to the pore, one of the two strands is pulled through the pore, called the template strand. Subsequently unzipping the double stranded DNA leaves the 5' end of the complementary strand in proximity of the pore using a tether molecule attached to the membrane. As the sequencing reaches the end of the molecule, with some likelihood, the complement strand immediately follows the template strand through the same pore. From the output signal reads that transition one after the other with similar sequence lengths and complementary base composition are detected as pairs, referred to as a duplex pair.

Earlier basecalling methods either uses single strand signals or join signals from both template and complementary strands, called 'paired decoding'. On the one hand simplex basecalling (processing the signal of a single strand individually) is very fast however yields higher error rates. On the other hand, feeding both strands to a neural network basecalling algorithm decoding base pairs yield high accuracy sequences at the expense of resources and time. Decoding base pairs is computationally intensive, up to five times slower compared to simplex basecalling and therefore lacks scalability[14].

The novel quality increase for 'stereo duplex basecalling' finds its origin by feeding base information, quality scores and the sequencing signal for both template and complement strand to a 'stereo' basecaller. This basecalling method is simple, fast, and robust allowing for better scalability towards generating large amounts of data over a reasonable time frame while yielding Q30 sequencing reads. With read-quality approaching gold-standard sequencing platforms Oxford Nanopore Technologies appears a promising technique for analyses that require high accuracy on base pair level, such as SNP detection and haplotype identification, particularly for polyploid genomes.

Even though we observe an outpacing of Moore's Law (Figure 9 - introduction) regarding sequencing cost, long read sequencing remains relatively expensive. Under more cost efficient conditions long-read sequencing is also a well-suited candidate for functional genomics analysis. The ability to prepare samples libraries without amplification circumvents the introduction of sequence-specific biases, where some molecules are underrepresented, and others excessively amplified allowing precise quantification. Long-read sequences can span full-length transcripts in a single read, hence avoiding complicated transcript assemblies, allowing simplified identification, and requiring fewer sequencing reads to identify the same number of genes compared to short read methods[15]. Furthermore, since full-length transcripts are recorded using single reads, they are exceptionally valuable for the characterization of structural variation such as isoforms. Isoforms can exhibit different functional properties and expression levels, and they are extremely difficult to determine using short reads. Additionally, structural variation is used across a broad spectrum of research areas, where it has shown significant importance to understand cancers in clinical settings all the way up to encoding target traits in agricultural studies. Structural variation spans, in many cases, Mbp stretches in the genome and are impossible to capture in a single read using gold-standard sequencing techniques. Hence those regions are sequenced in a fragmented fashion and reassembled to uncover the full structural variation using gold-standard techniques. This yields misassemblies and the absence of regions that are prone to amplification biases using other sequencing methods. Furthermore, since long reads provide increased alignment specificity the number of ambiguous alignments is significantly reduced, rescuing alignment regions that are lost using short read methods.

And finally, the sensitivity of sequencing signals and developments in artificial intelligence allows nanopore sequencing to detect modified bases. The epigenome is a complicated framework existing of a multitude of chemical compounds dictating DNA functionality. The higher order structure orchestrating the genome function comprises, among others, CpG methylation, nucleosome occupancy, chromatin accessibility, histone modifications and protein binding events that aid in proper segregation of chromosomes[16,17]. The most well-known epigenetics component is CpG methylation and is associated with suppressing gene transcription under hyper methylated promotor conditions or transcription activation for hypo- and hypermethylation of the promotor region and gene body, respectively. A gold-standard method to detect methylation is whole genome bisulfite sequencing, where unmethylated cytosines are replaced, at first using uracil and later by thymine nucleotides, revealing the methylation fingerprint. However, this method requires complicated bisulfite conversion steps, amplification and yields short reads. Hence this strategy is particularly difficult to apply for low complexity regions such as GC-islands. Oxford Nanopore Technologies methylation

identification has been shown to perform with similar accuracy compared to gold-standard methods. In addition, it offers the benefit of increased read-length and the absence of amplification, allowing better alignments for low complexity regions, avoiding complicated laboratory procedures, and only utilizing the sensitive sequencing signal and adjusted basecalling algorithm[18]. Applications for functional genomics and epigenetics have proven their worth for specific scientific bottlenecks and have bridged knowledge gaps of areas left untouched by traditional technologies. The current cost perspective makes Oxford Nanopore Technologies specifically attractive for specialized cases, whether that is to identify genes surrounded by repetitive content, quantify splice variants with repetitive content, generating methylation fingerprints over long range epigenetic elements or to close assembly gaps for large and complex genomes. When Oxford Nanopore Technologies reaches a cost-effective ratio comparable to gold-stand methods it will find its true potential and will open a new era for standardized sequencing and data processing allowing the analysis of anything, by anyone, everywhere.

To boldly go where no man has ever gone before

Suggested by the rapid read-length improvements it becomes more realistic to hypothesize that future sequencing will transform from a read-out of fragments method into a telomer-to-telomere sequencing fashion. Currently, the maximum read-lengths reported are >4 Mbp, compared to >10 Kbp during 2010 indicating it will not be long before end-to-end telomere sequencing is the gold standard. Sequencing entire chromosomes would bring significant benefits compared to current sequencing technologies, as it circumvents assembly for whole genome sequencing altogether. Downscaling computational load will relieve the scientific community of computationally intensive downstream analysis and will free scientist from dedicated computer clusters and command line tools.

Furthermore, sequencing speed is based on the number of nucleotides passing through the nanopore; to protect accuracy speeds are currently limited to 450 nucleotides per second. This sweet spot allows modern deep learning algorithms to determine the base sequence with up to Q30 accuracy. Increasing sequencing speed using those basecalling models would cause accuracy reduction as sequencing signals become too difficult to untangle. However, deep learning improvements resulting in more sophisticated neural network basecallers could increase sequencing speed up to a theoretically derived maximum of $>10^6$ nucleotides per second[19]. Exploiting the maximum sequencing speed could sequence a single human genome copy in just under two hours using a single pore. Such reduced computational pressure and increased sequencing speed will allow analysis of DNA content of any organism on a mediocre laptop in a matter of minutes, as opposed to a matter of days using dedicated and expensive computer clusters.

Additionally, standardized analysis workbenches should aid to reduce the time constraints even further, enabling scientists to navigate through the genomic content quickly and easily in a comprehensive, user-friendly, and visually appealing manner. Although read-lengths approach chromosome lengths, additional progress for chemistries must be made to, among others, avoid the entanglement of such long molecules during the isolation and unzipping of the double stranded DNA molecule.

Another potential application for future Oxford Nanopore Technologies that circumvents cell lysis to obtain high molecular weight DNA is the ability to sequence DNA/ RNA directly from the cell. Bringing the nucleus in proximity of the outer membrane and strategically incorporating a nanopore on both the nuclear envelop as well as on the outer membrane the nuclear interior could be connected to the sequencing pore. Using the intrinsic machinery that regulates proliferation to control entanglement and folding, DNA molecules can exit the nuclear envelope through the outer membrane into the sequencing pore. This would in turn bypass complicated entanglement of very large molecules and at the same time evade DNA molecule breakage that frequently occurs due to invasive laboratory procedures such as pipetting or mechanical lysis.

With one large leap of faith, in line with single cell nucleus sequencing, it might be possible to return the sequenced DNA or RNA through an additional feedback-pore. The unwinding of the DNA strands is then facilitated by proteins to collect and reposition proteins that are attached to the DNA strand. This allows the read-out of a single cell's entire genomic content without the need to sacrifice the sample. And would enable researchers to generate paired datasets that are statistically of incredible value avoiding biological variation on a cellular level.

References

1. Lee J Kerkhof, (2021), Is Oxford Nanopore sequencing ready for analyzing complex microbiomes?, FEMS Microbiology Ecology, Volume 97, Issue 3, fiab001, https://doi.org/10.1093/femsec/fiab001

2. Giselle C. Martín-Hernández, et. al., (2021), Chromosome-level genome assembly and transcriptome-based annotation of the oleaginous yeast Rhodotorula toruloides CBS 14, Genomics, Volume 113, Issue 6, Pages 4022-4027, ISSN 0888-7543, https://doi.org/10.1016/j.ygeno.2021.10.006.

3. Min-Seung Jeon et al., (2023), Life Science Alliance, 6 (4) e202201744; DOI: 10.26508/lsa.202201744

4. Yury A Barbitoff et al., (2021), Chromosome-level genome assembly and structural variant analysis of two laboratory yeast strains from the Peterhof Genetic Collection lineage, G3 Genes|Genomes|Genetics, Volume 11, Issue 4, jkab029, https://doi.org/10.1093/g3journal/jkab029

5. Hunt, M. et al, (2015), Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol 16, 294. https://doi.org/10.1186/s13059-015-0849-0

6. Kathryn Dumschott et al., (2020), Oxford Nanopore sequencing: new opportunities for plant genomics?, Journal of Experimental Botany, Volume 71, Issue 18, Pages 5313–5322, https://doi.org/10.1093/jxb/eraa263

7. De Maio N et al., (2019), Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genom. 5(9):e000294. doi: 10.1099/mgen.0.000294.

8. Jaume Pellicer et al., (2010), The largest eukaryotic genome of them all?, Botanical Journal of the Linnean Society, Volume 164, Issue 1, Pages 10–15, https://doi.org/10.1111/j.1095-8339.2010.01072.x

9. Segerman B (2020) The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. Front. Cell. Infect. Microbiol. 10:527102. doi: 10.3389/fcimb.2020.527102

10. Duan, H., Jones, A.W., Hewitt, T. et al., (2022), Physical separation of haplotypes in dikaryons allows benchmarking of phasing accuracy in Nanopore and HiFi assemblies with Hi-C data. Genome Biol 23, 84. https://doi.org/10.1186/s13059-022-02658-2

11. Sonia Boughattas et al., (2021), Whole genome sequencing of marine organisms by Oxford Nanopore Technologies: Assessment and optimization of HMW-DNA extraction protocols, Ecology and Evolution, https://doi.org/10.1002/ece3.8447

12. Aaron Pomerantz et al., (2018), Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building, GigaScience, Volume 7, Issue 4, giy033, https://doi.org/10.1093/gigascience/giy033

13. Darriba, D., Taboada, G., Doallo, R. et al. , (2012), jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9, 772 https://doi.org/10.1038/nmeth.2109

14. Silvestre-Ryan, J., Holmes, I., (2021), Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing. Genome Biol 22, 38. https://doi.org/10.1186/s13059-020-02255-1

15. Bayega, A., Oikonomopoulos, S., Gregoriou, ME. et al., (2021), Nanopore long-read RNA-seq and absolute quantification delineate transcription dynamics in early embryo development of an insect pest. Sci Rep 11, 7878. https://doi.org/10.1038/s41598-021-86753-7

16. Nicolas Altemose et al., (2022),Complete genomic and epigenetic maps of human centromeres. Science 376, eabl4178. DOI:10.1126/science.abl4178

17. Lee, I., Razaghi, R., Gilpatrick, T. et al., (2020), Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat Methods 17, 1191–1199. https://doi.org/10.1038/s41592-020-01000-7

18. Mitchell R. Vollger et al., (2022), Segmental duplications and their variation in a complete human genome. Science376,eabj6965.DOI:10.1126/science.abj6965

19. Wang Y, Yang Q and Wang Z (2015) The evolution of nanopore sequencing. Front. Genet. 5:449. doi: 10.3389/fgene.2014.00449