Preliminaries & Literature Review

For most of the history of DNA sequencing, expense and difficulty rendered whole-genome sequencing and assembly attainable only for large, highly funded organisations and projects. However, in the last 15 to 20 years, massive advances in automation and speed, and reductions in cost of DNA sequencing have sparked a burst of mutually reinforcing growth in the accessibility of whole-genome sequencing and assembly, application domains, and bioinformatics tools for assembly and downstream analyses [1, 2, 3]. Whole-genome assembly (WGA), in particular *de novo* WGA i.e. assembly of an individual genome without consulting previously resolved sequence, is a maturing and active field of research with well-established paradigms and a variety of tools addressing a range of needs. Nevertheless, there remains ample room at the time of writing for new approaches to continuing challenges, one of which shall be the subject of this thesis.

To set the stage for a more technical description of the specific issues to be addressed and statement of the approach for doing so, we shall start with a brief history of genome sequencing from the perspective of its usage in *de novo* WGA.

A brief history of genome assembly

Modern genome sequencing can be said to have begun with the introduction in 1977, by Frederick Sanger and colleagues, of a method for DNA sequencing with chain-terminating inhibitors [4], now commonly known as Sanger sequencing. This method became the most commonly used DNA sequencing technology for several years, spurring increasing automation and forming the basis of the first commercial DNA sequencing machines [5, 6].

During the "first generation" of genome sequencing represented by Sanger's method, the shotgun process, still dominant today, was first developed and used to assemble long contiguous genomic sequences from shorter reads. The process randomly breaks a target molecule; the resulting fragments are sampled and sequenced to obtain reads [7, 8]. In the case of whole-genome shotgun (WGS) sequencing, the target molecules consist of the chromosomes making up a genome. When a target is oversampled, the resulting reads overlap; they can then be computationally ordered and assembled [9]. The broad applicability of WGS was demonstrated by the 1995 completion of the 1.8-Mbp (million base pair) *Haemophilus influenzae* genome by WGS sequencing [10] and a number of subsequent projects [11]. Milestones were reached when almost all of the 120-Mbp euchromatic portion of the *Drosophila melanogaster* genome [12], and (in the Human Genome Project or HGP) a 2.91-Gbp consensus sequence of the euchromatic portion of the human genome [11], were determined using WGS sequencing with support by other techniques.

Concurrent with the spread and progress of first-generation sequencing technology, a luminescent method for measuring pyrophosphate synthesis was introduced [13]; its application to DNA sequencing via a technique called pyrosequencing was pioneered over the next decade [14, 15] and licensed to the company 454 Life Sciences, where it was deployed in the first major commercially successful "next-generation sequencing" (NGS) machines [6]. These machines coupled pyrosequencing with massively parallelised sequencing reactions, producing up to a million reads ~200-500 bps long in each run, which represented an orders of magnitude increase in sequencing yield [].

The massively parallel output of relatively short shotgun-generated reads is a shared, defining characteristic of several sequencing techniques that emerged over the next decade, which have come to be seen as second-generation sequencing. These techniques are also known by various other names, i.e. the aforementioned next-generation sequencing (NGS), massively parallel sequencing (MPS), and high-throughput sequencing (HTS) [14, Reuter et al. 2015, 3]. Since the release of Illumina's Genome Analyzer II sequencer in 2006, competition between NGS technology manufacturers drove tremendous gains in output and reductions in cost, with raw per-base cost plummeting by four orders of magnitude between 2007 and 2012 [Wetterstrand, 17]. As a result, NGS had almost completely supplanted Sanger sequencing a decade after the HGP was completed in 2001, and Illumina has achieved a near-monopoly on the NGS market. At present, NGS reads are typically a few hundred bps long, and sequencers produce an output of up to 1Tbps, or billions of reads, per run [Reuter et al. 2015, 3, 17].

Alongside the rise to ubiquity of NGS, yet another class of sequencing technologies, sometimes called the third generation, has been advancing rapidly [Heather and Chain 2016, Shendure et al. 2019]. Characterised by long-range single-molecule resolution, these technologies do not require amplification in library preparation, and now routinely produce reads of average length around 10kb, in a range that can exceed 1Mb [Sedlazeck et al. 2018]. They consist of two main approaches, single-molecule real-time (SMRT) sequencing [Roberts et al. 2013] and nanopore-based sequencing [Jain et al. 2016]. Distinct from these true long-read platforms, synthetic long reads are created through library preparation protocols that associate NGS short reads derived from a single larger molecule with the same barcode; these are currently available on platforms from two vendors, Illumina and 10X Genomics [Goodwin et al. 2016]. Another NGS-compatible advance has been the creation of very long-range mate pair-like data from Hi-C and related chromatin crosslinking protocols [Dudchenko et al. 2017, Sedlazeck et al. 2018]. Lastly, new optical mapping technology from BioNano Genomics uses nicking enzymes to create high-resolution sequence motif physical maps that can be *de novo* assembled into scaffolds to complement assembled genomic sequence [Jiao et al. 2017, Sedlazeck et al. 2018].

In principle, if assembled using effectively tailored approaches, single-molecule data offer alternatives or solutions to the drawbacks of NGS technology for *de novo* WGA, including amplification artefacts from library preparation, and difficulty in characterising long repeats and large structural variation [Sedlazeck et al. 2018]. For example, high-quality long-read-only assemblies were created of microbial [Chin et al. 2013], maize [Jiao et al. 2017], and human [Shi et al. 2016, Pendleton et al. 2015] genomes. However, in practice at the time of writing, the relatively high error rates and per-base cost of single-molecule reads make them impractical for many purposes unless used in conjunction with a short-read base assembly [Wick et al. 2017]: they require costly error-correction procedures [Koren et al. 2012], can be very computationally expensive [Zimin et al. 2017, Koren et al. 2017, Sedlazeck et al. 2018], or require prohibitively high coverage—the aforementioned example assemblies used 65x to 103x SMRT read depth or, in one case, 22x and 24x SMRT with 80x optical mapping coverage. Moreover, short reads, particularly Illumina reads, are widely used and likely to remain so for some time due to their high accuracy and low cost [Wick et al. 2017]. Not least, a NGS workflow can be leveraged to include linked reads and produce a high-quality assembly at modest extra cost [Sedlazeck et al. 2018, Jiao and Schneeberger 2017]. Thus, NGS short-read *de novo* WGA is likely to remain highly relevant for some time.

Putting it together: an overview of *de novo* whole-genome shotgun sequence assembly

Of course, raw genomic sequence reads do us no good without tools, e.g. assemblers, that turn them into useful information. For much of the history of DNA sequencing, before NGS drove cost down, genome sequencing and assembly was the primary purpose of DNA sequencing [Shendure 2019]. Thus, genome sequencing, in particular *de novo* whole-genome shotgun (WGS) sequencing, has had a long development,

Context
- Preceded by OLC; reasons doesn't work well for NGS
- de Bruijn graph (DBG) sequence assembly: Rationale & brief history

Technical introduction to DBG WGS assembly
- DBG: definition
- Overview of canonical assembly process; definitions as needed

- importance of PE data (source???)

References

1. Nagarajan N, Pop M. Sequence assembly demystified. *Nature Reviews Genetics.* 2013;14(3),157–167.
2. Ekblom R, Wolf JBW. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*. 2014;7(9):1026–1042.
3. Zhao EY, Jones M, Jones SJM. Whole-genome sequencing in cancer. *Cold Spring Harb Perspect Med.* 2019;9:a034579.
4. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA.* 1977;74(12):5463–5467.
5. Hunkapiller T, Kaiser R, Koop B, Hood L. Large-scale and automated DNA sequence determination. *Science.* 1991;254(5028):59-67.
6. Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. *Genomics.* 2016;107(1):1-8.
7. Sanger F, Coulson AR, Barrell BG, Smith AJ, Roe BA. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J Mol Biol* 1980;143(2):161-178.
8. Setubal J, Meidanis J. Introduction to computational molecular biology. Boston, MA: *PWS Publishing Company*; 1997:19.
9. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics.* 2010;95(6):315-327.
10. Fleischmann RD, Adams MD, White O et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science.* 1995;269(5223):496-512.
11. Venter JC, Adams MD, Myers EW et al. The sequence of the human genome. *Science.* 2001;291(5507):1304-1351.
12. Adams MD, Celniker SE, Holt RA, et al. The genome sequence of Drosophila melanogaster. *Science.* 2000;287(5461):2185-2195.
13. Nyrén PI, Lundin A. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal Biochem* 1985;151(2):504-509.
14. Nyrén PI. Enzymatic method for continuous monitoring of DNA polymerase activity. *Anal Biochem* 1987;167(2):235-238.
15. Ronaghi M, Uhlen M, Nyrén PI. A sequencing method based on real-time pyrophosphate. *Science.* 1998;281(5375):363-365.

16. Margulies M, Egholm M, Altman W, Attiya S. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 2005;437(), pp. 376-380.
17. J. Shendure, H. Ji. Next-generation DNA sequencing. *Nat Biotechnol*, 26 (2008), pp. 1135-1145.
18. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. *Mol Cell*. 2015;58(4):586–597.
19. Wetterstrand, K. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). 2017.
20. Shendure J, Balasubramanian S, Church GM et al. DNA sequencing at 40: past, present and future. *Nature.* 2019;550(7676),345-353.
21. Sedlazeck FJ, Lee H, Darby CA, Schatz, MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nature Reviews Genetics. 2018;19(6),329-346.
22. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. *Genome Biology*. 2013;14(6).
23. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 2016;17, 239.
24. Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17(6), 333–351. doi:10.1038/nrg.2016.49
25. Dudchenko, O. et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
26. Jiao Y, Peluso P, Shi J et al. Improved maize reference genome with single molecule technologies. Nature. 2017;546,524-527.
27. Chin C-S, Alexander DH, Marks P et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*. 2013;10(6),563–569.
28. Shi L, Guo Y, Dong C, et al. Long-read sequencing and *de novo* assembly of a Chinese genome. Nature Communications. 2016;7(1),12065.
29. Pendleton M, Sebra R, Pang AWC et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*. 2015;12(8),780-786.
30. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*. 2017;13(6),e1005595.
31. Koren S, Schatz MC, Walenz BP et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*. 2012;30(7),693-700.
32. Zimin AV, Puiu D, Hall R et al. The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *Gigascience.* 2017;6(11),1–7.
33. Koren S, Walenz BP, Berlin K et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 2017;27(5),722–736.
34. Jiao W-B, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*. 2017;36,64–70.