Norwegian University
of Life Sciences

Master's Thesis 2017 30 ECTS
Faculty of Chemistry, Biotechnology and Food Science

Evaluation of long-read Nanopore sequencing in genome studies

Thomas Dailey Strand
MSc Chemistry and Biotechnology

Acknowledgements

This master's thesis was performed at the Faculty of Chemistry, Biotechnology and Food Science at the Norwegian University of Life Sciences (NMBU), with Professor Knut Rudi as main supervisor and Postdoctoral Ekaterina Avershina as co-supervisor.

First of all, I would like to thank my main supervisor, Knut Rudi, for giving me the opportunity to write this master's thesis, and acquainting me to his research team, the MiDiv group. Your positive attitude, insight and excellent sense of humor made writing this thesis a memorable experience. Then I would like to give a big thanks to Ekaterina Avershina for being a great co-supervisor, assisting me tremendously with the writing and structuring of my thesis. Your knowledge, willingness to help and ability to teach, amazes me. Thank you, Inga Leena Angell, for helping me during my practical work in the laboratory. And thanks to the rest of the MiDiv group for answering any questions I have had and making me feel welcome.

Thanks to Misti Dawn Finton and Davide Porcellato and for assisting me during the Nanopore sequencing run. Thanks to my friend, Nathan Lau, for guiding me in Linux while using the Canu assembly tool for processing the Nanopore sequencing data. Thanks to my girlfriend, Ane Brækken, and my family for supporting me. And finally, thanks to my fellow master student, Åsmund Andersen, for introducing me to Knut Rudi, and for being an excellent workmate.

Ås, May 2017

Thomas Dailey Strand

Abstract

The development of the latest generation of sequencing technologies, known as third-generation sequencing, has revolutionized sequence assembly due to long reads compared to the short reads of second-generation sequencing technologies. However, these long reads are highly error-prone (~8 – 40 % error rate). In this thesis, we performed an evaluation of using the Nanopore MinION platform, developed by Oxford Nanopore Technologies, for *de novo* assembly of the *Bifidobacterium longum* genome. In addition, we tested whether Amplified Fragment Length Polymorphism (AFLP) fragments, sequenced on the Illumina MiSeq platform, have the potential to correct highly error-prone Nanopore reads. The MinION sequencing run generated highly erroneous reads, resulting in long stretches of A's and T's in the output data, revealing the issue of G/C-bias related to Nanopore sequencing. However, an *in silico* experiment demonstrated the suitability of short AFLP fragments to correct long reads that had up to 40 % of erroneous bases. This demonstrates their high potential for error-correction.

Keywords: Nanopore, *de novo* assembly, AFLP

Sammendrag

Utviklingen av den nyeste generasjonen innen sekvenseringsteknologi, kjent som tredje-generasjons sekvensering, har revolusjonert hvordan man rekonstruerer sekvenser sammenliknet med andre-generasjons sekvensering på grunn av evnen til å sekvensere lange sekvenser. En ulempe med å sekvensere lange sekvenser, er derimot høy sekvenseringsfeil (~8 – 40 %). I denne masteroppgaven ble det gjort en vurdering av bruksevnen til sekvenseringsplatformen Nanopore MinION, utviklet av Oxford Nanopore Technologies, til å generere data for å rekonstruere genomet til *Bifidobacterium longum*, uten å sammenlikne med referanser (*de novo*). I tillegg ble det testet om 'Amplified Fragment Length Polymorphism' (AFLP) fragmenter, sekvensert med Illumina MiSeq, kunne brukes til å rette opp sekvenseringsfeil i lange Nanopore sekvenser. Sekvenseringsforsøket med MinION gav sekvenser med veldig høy feilrate som ikke kunne brukes i videre analyse, og resulterte i generering av lange sekvenser bestående av basene A og T. Dette avslørte sekvenseringsproblemet tilknyttet høyt innhold av basene G og C under Nanopore sekvensering. Det ble derimot gjort et simulert forsøk som demonstrerte evnen AFLP fragmenter har til å rette opp sekvenseringsfeil i lange sekvenser, som bestod av opptil 40 % feilaktige baser. Dette demonsterer potensialet disse fragmentene har til å rette opp sekvenseringsfeil.

Nøkkelord: Nanopore, *de novo* rekonstruksjon, AFLP

Abbreviations

(M)bp(s)	- (Mega) base pair (s)
cDNA	- Complementary DNA
DNA	- Deoxyribonucleic acid
dNTP	- Deoxynucleoside triphosphate
dsDNA	- Double-stranded DNA
gDNA	- Genomic DNA
HGP	- Human Genome Project
MiDiv	- Microbial Diversity
Min	- Minutes
NGS	- Next-generation sequencing
ONT	- Oxford Nanopore Technologies
PacBio	- Pacific Biosciences
PCR	- Polymerase chain reaction
RNA	- Ribonucleic acid
SD	- Standard deviation
Sec	- Seconds
SGS	- Second generation sequencing
ssDNA	- Single-stranded DNA
TGS	- Third generation sequencing
v	- Version
WGS	- Whole-genome sequencing

Table of contents

1 Introduction

1.1 First-generation sequencing

In 1977, Frederick Sanger was the first person to sequence a genome containing single-stranded DNA (ssDNA) from a bacteriophage (Sanger et al. 1977). The method he used for sequencing is known today as Sanger sequencing. This opened the door to one of the fastest growing fields in modern biology; genome sequencing (Schuster 2008).

Sanger sequencing, or the chain termination method, is based on the separation and detection of labeled dideoxynucleotides (ddNTPs) through capillary electrophoresis. As labeled ddNTPs are incorporated into to growing ssDNA, synthesis is terminated, creating different lengths of ssDNA. The termination process occurs because of the lack of the OH-group on the ddNTP, which is responsible for crosslinking dNTPs during synthesis. When the ssDNA fragments travel through the capillary, a light emitted by one of the four fluorochromes on terminal ddNTP is captured. Since ssDNA fragments are sorted by size, the light signal can be translated into DNA sequences (Shendure & Ji 2008).

Sanger sequencing was used when first sequencing the human genome, which was achieved through the Human Genome Project (HGP, 1990-2003) (McCarthy 2010), and is still used today having undergone major improvements regarding fragment read-lengths and base calling accuracy (Shendure & Ji 2008). However, this method was expensive and had a very low throughput. Therefore, to be able to establish and build solid databases with information regarding human health and prokaryotic/eukaryotic evolution (Bahassi el & Stambrook 2014; Mardis 2011), the sequencing technology needed to evolve and expand in speed, accuracy and volume of data acquisition (Margulies et al. 2005).

1.2 Second-generation sequencing

Second-generation sequencing (SGS) or next-generation sequencing (NGS) is a term used for all sequencing methods and platforms that are based on high-throughput sequencing. High-throughput is the term used when describing vast amounts of data being generated in a sequencing run, and is achieved through ‘massively parallel sequencing’ which is a reference

to the amount of sequences being processed simultaneously during a sequencing run (Behjati & Tarpey 2013; Lu et al. 2016; Mikheyev & Tin 2014).

SGS was introduced in 2005 and has dominated the market the last decade (Illumina 2016b; Quail et al. 2012). Due to SGS, genome sequencing became orders of magnitude cheaper, faster and generated much more sequencing data as compared to Sanger sequencing (Illumina 2016c). This has resulted in the generation of enormous amounts of data (Deschamps et al. 2016).

The main strengths of SGS are very accurate reads (>99%) (Quail et al. 2012; Salipante et al. 2014) and high throughput when sequencing. SGS is mostly based on sequencing small fragments of DNA/RNA (short reads), spanning up to a few hundred base pairs (bps). The most common SGS technologies are based on sequencing by synthesis (SBS) (Illumina, Roche/454 Pyrosequencing), sequencing by ligation (Applied Biosystems SOLiD) and sequence determination based on change in pH (Ion Torrent PGM) (Berglund et al. 2011; Mardis 2013).

Sequencing by synthesis

SBS is a technology that is based on the detection of deoxynucleoside triphosphates (dNTPs) as they are added into a growing DNA strand (Illumina 2010). SBS is currently one of the most widely adopted and successful NGS technologies worldwide (Reuter et al. 2015; Schirmer et al. 2015), and is the foundation for all Illumina sequencing platforms (Illumina 2016b; Quail et al. 2012), Ion Torrent PGM, as well as Roche/454 Pyrosequencing (Berglund et al. 2011).

1.2.1 Illumina sequencing

During a sequencing cycle with Illumina technology, one labeled dNTP is added to a growing nucleic acid chain (Mardis 2011) and the emitted fluorescence is captured to determine the dNTP added during synthesis. The sequencing process is initiated by ligating adapter sequences onto each end of the template DNA fragments generated from extracted DNA. The fragments are then added to a glass flow cell with complementary adapter sequences that bind and immobilize these fragments. Clusters of template DNA are then generated by solid-phase bridge amplification which generates up to 1000 copies in close proximity of each

template DNA. The bridge amplification is initiated by adding unlabeled nucleotides and enzyme. Double-stranded bridges are then formed on the solid-phase substrate from nucleotides being incorporated by the enzyme. This results in several million of single-molecule clusters which is possible because there is no involvement of photolithography, mechanical spotting, or positioning of beads into wells (Illumina 2010). After the cluster generation, four labeled reversible terminators, primers and DNA polymerase are added to begin sequencing. The emitted fluorescence from each cluster is captured as an image as each labeled dNTP is added to the growing DNA strand. This process is repeated until the template DNA fragment has been fully sequenced.

During paired-end sequencing, each DNA fragment is read from both sides (5'-end and 3'-end) for an overlap consensus in the sequencing. The limitation with this technique is that it only reads up to a maximum of 300 bp from each side (Schirmer et al. 2015), meaning fragments longer than 600 bp will not have any overlap. This challenge is overcome by producing fragments with a length of ~500 bp, making sure there is sufficient overlap when sequencing the fragment.

Illumina has several platforms that differ in produced output and imaging technique. The benchtop sequencing platforms include MiSeq (4-channel imaging), MiniSeq (2-channel imaging) and NextSeq (2-channel imaging). The production-scale platforms include HiSeq (4-channel imaging) and NovaSeq (2-channel imaging). The 2-channel imaging uses a combination of dyes to label each of the four nucleotides instead of the separate dye for each nucleotide as used in 4-channel imaging (Illumina 2016b).

1.3 Third-generation sequencing

Despite the success of SGS and their highly accurate reads, the short read length became a challenge when performing *de novo* genome assembly, as well as the extensive time needed to perform a sequencing run (Berlin et al. 2015; Lu et al. 2016; Mikheyev & Tin 2014; Nagarajan & Pop 2009). Sequence assembly or reconstruction refers to the process of aligning and merging DNA fragments into the original sequence, and *de novo* assembly is when the original sequence is reconstructed without any reference (Mardis 2013). Complex genomes often contain long repetitive genomic regions, which can span across several hundred bps or even

thousands. The short reads of SGS are unable to surpass these regions, thus making it a challenge or even impossible to assemble them (Ashton et al. 2015; Lee et al. 2016; Rhoads & Au 2015).

Based on this information, third generation sequencing (TGS) was introduced in an effort to complement the current methods of SGS in hope of diminishing the challenges they have (Istace et al. 2017; Munroe & Harris 2010). TGS refers to long-read technology able to sequence a single molecule at a time, like the Oxford Nanopore Technologies (ONT) MinION or Pacific Biosciences (PacBio) Single-Molecule Real-Time (SMRT) sequencing. The aim for TGS was to generate long reads (tens of kilobases) to exceed the repetitive genomic regions, making it much less of a challenge to assemble these areas, while also reducing the sequencing time compared to SGS (from days to hours), and diminishing of sequencing biases due to not needing PCR amplification (Lu et al. 2016; Meldrum et al. 2011; Schadt et al. 2010). Strengths and weaknesses of the SGS and TGS technologies are mostly complementary, which inspired the hybrid sequencing strategy (Rhoads & Au 2015), combining technologies to achieve accurate assemblies of complex genomes.

1.3.1 Nanopore technology

The idea behind using nanopores for sequencing was first introduced during the 1990s (Feng et al. 2015) and had its breakthrough when a functional Nanopore sequencing platform was developed after academic laboratories achieved several discoveries within the Nanopore research field (Jain et al. 2016).

The MinION is a TGS platform that operates by ‘pulling’ or translocating DNA or RNA through a nanopore (Jain et al. 2016). The platform was developed by ONT and released in 2014, and is a small highly portable device that can sequence a single molecule of DNA or RNA, generating data in real-time. Nanopores are currently the only sequencing technology that are able to sequence RNA directly, avoiding the use of complementary DNA (cDNA) and PCR amplification (ONT 2017a). However, this thesis will not focus any further on the use of RNA for sequencing.

The MinION is connected to a computer through the USB port and has user-friendly software (MinKNOW) that is easily navigated and displays critical information during a sequencing run

(ONT 2015), like run parameters and tracking the activity of available nanopores. In addition to storing sequencing data locally, data is also uploaded and stored online through an online tool called Metrichor. Metrichor is a cloud based analysis service developed by Metrichor (an Oxford Nanopore company), which helps connect and share valuable information on all living organisms across the world. They call this 'The Internet of Living Things' and their goal is to provide real-time biological data analysis (Metrichor 2015). During a sequencing run, Metrichor will also provide base calling for all reads performed by the MinION (Istace et al. 2017).

The MinION utilizes flow cells containing up to 2048 individual protein nanopores set in an electrically resistant polymer membrane. When performing a sequencing run with the MinION, a voltage is applied across this membrane, which creates an ionic current that passes through the nanopore, translocating the DNA through the nanopore (Deschamps et al. 2016). Genomic DNA fragments are generated using transposase, which also ligates adapters to the free ends of the DNA fragments. These adapters facilitate strand capture several thousand-folds, as well as concentrating the DNA substrates at the membrane surface in close proximity to the nanopores (Jain et al. 2016). As DNA is translocated through the nanopore by the current, there is a characteristic disruption for each molecule, called events. These events can be measured which makes it possible to identify the order of nucleic acids (Laver et al. 2015; Lu et al. 2016).

This device has achieved reads of tens of kilobases in short time (Lu et al. 2016), and should theoretically only be limited to the length of the strand itself (unless limitations are introduced during sample preparation) (Ip et al. 2015). The MinION also generates data in real-time, which means the read results becomes known as soon as each base passes through the nanopore. This has huge clinical impact as one would be able to identify problems or abnormalities, and if needed, apply treatment consecutively as the sequencing takes place (Dunne et al. 2012; Lu et al. 2016). Real-time data generation, combined with MinION's size, cost, simple library preparation (no amplification step needed) and, portability, creates a general advantage over other TGS platforms (Istace et al. 2017). Not needing any amplification step during library preparation is a critical factor for diminishing sequencing biases during sequencing (Madoui et al. 2015).

There are two different ways the MinION can perform sequencing; 1D and 2D. The difference between the two methods is whether single-stranded DNA (ssDNA) or double stranded DNA (dsDNA) is sequenced (Jain et al. 2016). The library preparation for 2D sequencing is more extensive than the 1D preparation, involving the ligation of a hairpin adapter to connect the template and complementary strand, linking them into one strand after denaturing the dsDNA (Mikheyev & Tin 2014). This allows for sequencing of both the template and complementary strands, yielding a slightly higher read accuracy than 1D sequencing, but at the cost of lower throughput and shorter read length (Jain et al. 2016; Koren et al. 2017).

1.4 Genome analysis

1.4.1 Whole-genome sequencing

When researching an organism, the most comprehensive way of obtaining information is by using whole-genome sequencing (WGS) i.e. determining the full DNA sequence in a single sequencing run (Ekblom & Wolf 2014; Illumina 2016a). The constant decline in sequencing cost and higher sequencing quality since the introduction of NGS technologies have increased the potential of WGS by allowing it to be used more frequently (Bahassi el & Stambrook 2014).

The general procedure when performing WGS is first randomly shearing genomic DNA into fragments. Then adapters are ligated onto each end of these fragments before being amplified through PCR and then sequenced. There are several variants of this procedure, and Illumina's method, Nextera, will be detailed as an example:

The first step is fragmenting and tagging (tagmentation) of genomic DNA by the Nextera transposome. Then the tagmented DNA is cleaned up before being amplified through PCR, generating a DNA library. After amplification, the fragments of the required size are selected by AMPure XP beads before being sequenced (Illumina 2016d).

1.4.2 Amplified Fragment Length Polymorphism

DNA fingerprinting or profiling is a technique used to identify species based on characteristics of their DNA, and is performed by looking at specific DNA fragments in a DNA sample. Amplified Fragment Length Polymorphism (AFLP) is a DNA fingerprinting technique used to

identify differences and similarity-patterns in DNA between closely related species of plants, fungus, bacteria and animals by using selective PCR amplification to detect genomic restriction fragments (Gibson et al. 1998). This is a technique developed and published during the 90s (Vos et al. 1995), and is closely related to restriction fragment length polymorphism (RFLP). However, instead of looking at the differences in length of restriction fragments, the AFLP technique displays either the presence or absence of restriction fragments. A core strength of the AFLP technique is its applicability to any organism without any prior sequence knowledge (Maughan et al. 1996; Vuylsteke et al. 2007).

Genomic DNA is digested by a pair of restriction enzymes, usually EcoRI and MseI. EcoRI is a rare cutter (cutting every ~4096 bp) and MseI is a frequent cutter (cutting every ~150 bp) (Vos et al. 1995). The MseI is responsible for generating the fragments, while EcoRI limits the number of effective amplicon fragments that will be amplified. After digestion, adapters are then ligated onto each end of the fragments. These adapters consist of a core sequence and an enzyme-specific sequence. The enzyme-specific sequence is what allows for correct ligation onto the restriction fragments. Three subsets of fragments are formed after ligation (EcoRI-EcoRI, MseI-MseI and EcoRI-MseI). Fragments located between the different enzymes (EcoRI-MseI) are then selected and amplified by PCR by using selection primers that are complimentary to the adapter sequences (Ajmone-Marsan et al. 1997), restriction area, and a few nucleotides within the restriction fragments. The amplified fragments are then separated on a denaturing gel by gel electrophoresis and visualized by exposure of the gel to UV-light (Masny & Plucienniczak 2001). The resulting banding patterns can then be analyzed to determine the DNA fingerprints associated with a given sample (Vos et al. 1995; Vuylsteke et al. 2007). The amplified fragments can also be identified through sequencing.

1.5 Data analysis

1.5.1 Canu assembly tool

After a sequencing run with the MinION, error correction of the raw data can be performed before generating contigs by assembling the reads. There are several different software packages developed specifically for processing Nanopore or TGS data as previous software algorithms do not work well with the long error-prone TGS reads (Simpson 2015).

Canu (Phillippy et al. 2015) is a successor of the Celera assembler (the assembler previously used for PacBio and Nanopore data), and is a TGS assembler tailored for processing and assembling long noisy reads - like Nanopore MinION reads (Koren et al. 2017). Canu can assemble both 1D and 2D Nanopore reads, and utilizes the MinHash Alignment Process (MHAP), which is a new algorithm introduced by the creators of Canu, and is designed to overcome the computational bottleneck of overlapping noisy, single-molecule reads. The algorithm is based on a hierarchical method which processes the sequencing data by using multiple rounds of read overlapping (Berlin et al. 2015), and does not require any complementary data to supplement the long reads (Chin et al. 2013).

Canu was chosen for this thesis to process the MinION data based on it being one of the tools most recently updated, analysis feedback (Istace et al. 2017; Judge et al. 2016), and having specific noise correction for long, noisy reads (Koren et al. 2017).

1.5.2 Geneious analysis tool

As sequencing technologies develop, providing both larger amounts and more challenging data due to the increase in read length, it is crucial having access to the appropriate analysis tools to achieve success. Geneious (Biomatters 2005) is a framework providing several analysis methods, complementary data sources and visualization tools designed to more easily overcome the challenges of conducting research-focused bioinformatics (Kearse et al. 2012).

1.6 Thesis objectives

Reads generated through Nanopore sequencing gives an advantage for *de novo* assembly due to the long reads. However, these long reads contain high sequencing error, making it difficult to accurately assemble (Lu et al. 2016; Zimin et al. 2017). Therefore, highly accurate SGS reads can be helpful to increase the accuracy of assembly.

The assembly of longer reads also show a better contiguity than the assembly of shorter reads during *de novo* sequencing (Salzberg et al. 2012), making shorter reads less valuable when it comes to *de novo* genome assembly of large and complex genomes with long repetitive regions (Miller et al. 2010). Hence, many genomes that have been sequenced by SGS methods,

and are available through various databases, are not complete (Koren et al. 2013; Phillippy et al. 2008; Risse et al. 2015).

The new sequencing methods and technologies introduced over the last few years (TGS) are designed for diminishing or overcoming the challenges of using SGS when assembling complex genomes (Bahassi el & Stambrook 2014). These methods have much longer read lengths than SGS as previously mentioned, and to date mostly used in a complementary effort with SGS to fill each other's gaps.

Because of the issues related to SGS with *de* novo genome assembly, the new approach by combining methods (hybrid sequencing/assembly) to achieve accurate, unfragmented results will be tested and evaluated with the established TGS method; Nanopore sequencing.

Aim

The main aim of this thesis was to evaluate the use of long error prone sequencing reads in combination with short accurate reads in order to obtain accurate *de novo* genome assemblies.

Sub goals:

- Use the MinION for sequencing the *Bifidobacterium longum* and attempt to assemble the data into contigs.
- Asses the error correction potential of AFLP fragments, sequenced by Illumina MiSeq
- Determine the error rate threshold for use of AFLP fragments by *in silico* generated data.

2 Materials and methods

A list of all commercial agents and instruments used are listed in Appendix 1.

A schematic representation of the work flow of the thesis is shown in Figure 1.

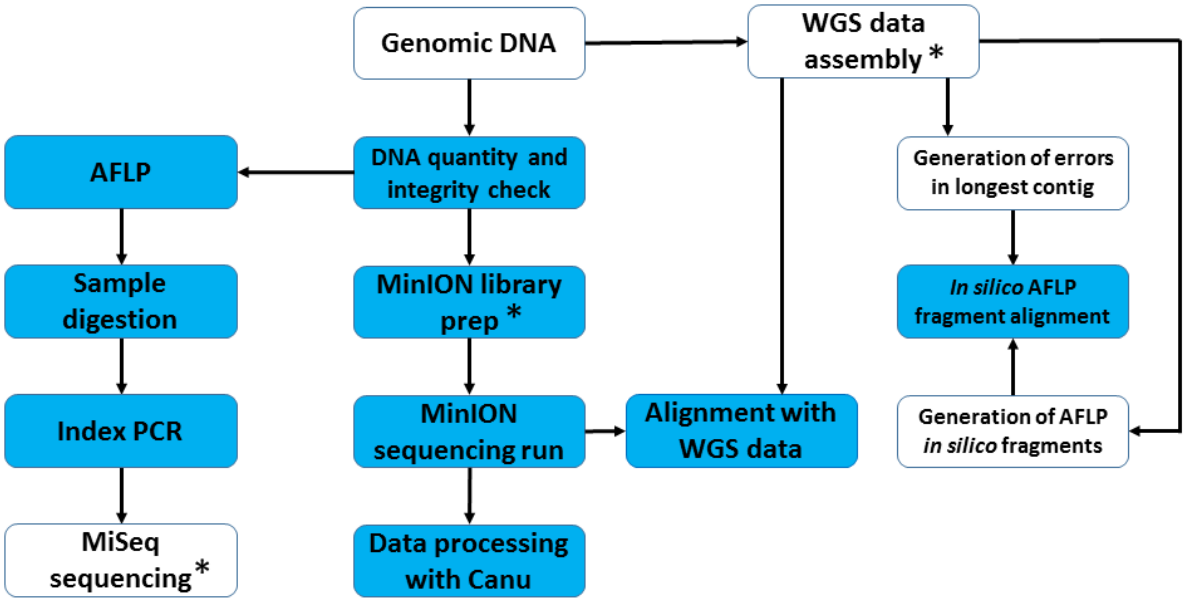


Figure 1. A flow-chart representation of the workflow of this thesis. Light blue nodes indicate what was performed by me for this thesis, and the uncolored nodes indicate work that was done by other people from the MiDiv group. Nodes marked with a star (*) indicates that only *B. longum* was used for this step.

2.1 Samples description

Genomic DNA from was extracted using phenol chloroform extraction (Sambrook & Russell 2006) prior to the start of this thesis and stored at -20 °C until analysis. All experiments were performed in duplicates. *B. longum*, *B. adolescentis* and *B. infantis* were used.

2.2 DNA integrity and concentration

Quant-iT Assay

Genomic DNA was measured using the Invitrogen Quant-iT dsDNA High-Sensitivity (HS) kit (Invitrogen Inc.) on the Qubit fluorometer according to manufacturer's recommendations.

Gel electrophoresis

Ten µL of genomic DNA was tested on a 0,8 % agarose gel to check integrity of the DNA. The DNA was treated with 1 µL RNAse for 30 min in room temperature prior to the run.

2.3 Library preparation of AFLP amplicons for Illumina MiSeq sequencing

This AFLP method follows an in-house protocol developed by the MiDiv lab, and is not publicly available. All steps of the AFLP protocol were performed in room temperature (~20 °C) unless otherwise specified.

2.3.1 AFLP protocol

Restriction cutting

A restriction cutting mix consisted of EcoRI enzyme (8U), MseI enzyme (4U), gDNA template (10 ng) and Cut Smart Buffer (1x). Nuclease-free water was added up to a total volume of 20 µL for each sample. The mixes were then incubated for 1 hour at 37 °C.

Adapter ligation

A ligation mix was made consisting of EcoRI adapter mix (0,5 µM), MseI adapter mix (5 µM), T4 Reaction Buffer (1x) and T4 DNA Ligase (1 µL). The EcoRI and MseI adapters used in the mix are shown in Table 1.

Table 1. Overview of forward (fwd) and reverse (rev) adapter sequences used for making the EcoRI and MseI adapter mixes in the AFLP process.

Mix	Adapter name	Adapter sequence
EcoRI adapter mix	EcoRI_adapter_Fwd	5'-CTCGTAGACTGCGTACC-3'
	EcoRI_adapter_Rev	5'-AATTGGTACGCAGTCTAC-3'
MseI adapter mix	MseI_adapter_Fwd	5'-GACGATGAGTCCTGAG-3'
	MseI_adapter_Rev	5'-TACTCAGGACTCAT-3'

After the 1 hour incubation of the restriction cutting mix, 5 µL of the adapter ligation mix was added, resulting in a total volume of 25 µL. Incubation at 37 °C was then continued for another 3 hours.

PCR amplification

Each reaction contained HOTFIREPol RTL (1x), EcoRI primer (0,2 µM), MseI primer (0,2 µM) and template DNA (5 µL, from the 25 µL volume mix made in the previous step). Nuclease-free water was then added up to a total volume of 25 µL in each sample. The EcoRI and MseI primers used are shown in Table 2.

Table 2. Overview of the EcoRI and MseI primers used for PCR amplification of the restriction cutting and adapter ligation mixes made in the AFLP process.

Primer name	Primer sequence
EcoRI_primer	GACTGCGTACCAATTC
MseI_primer	GATGAGTCCTGAGTAA

The following thermo-cycling protocol was used:

Activation at 95 °C for 15 min, 25 cycles of; denaturation at 95 °C for 30 sec, annealing at 56 °C for 1 min and elongation at 72 °C for 1 min, with a final elongation step at 72 °C for 7 min. PCR products were then tested on a 1 % agarose gel.

PCR product clean up with AMpure XP beads

Equal volumes of the PCR product and AMPure XP beads (1x beads) were mixed by pipetting up and down 10 times in a 1,5 mL Eppendorf tube for each sample.

Tubes were then incubated at room temperature for 5 min. After 5 min, the tubes were placed on a magnetic stand for at least 2 min until all supernatant had cleared from the beads. The following steps were all conducted with tubes on a magnetic stand:

The supernatant was carefully removed and discarded. The beads were then washed with freshly prepared 80 % ethanol as follows:

120 µL 80 % ethanol was added to each tube carefully in order to avoid resuspending the beads. The tubes were then incubated in room temperature for 30 sec, then the supernatant was carefully removed and discarded afterwards. This washing step was repeated and excess ethanol was removed at the end of the washing step. After the removal of all the ethanol, beads were then air dried for 15 min.

Tubes were then removed from the magnetic stand and 25 µL PCR water was added to each tube and gently mixed to resuspend the beads by pipetting up and down 10 times. The tubes were then incubated at room temperature for 2 min before being placed back onto the magnetic stand. Once the supernatant had cleared, 20 µL of the eluted DNA was transferred to a new 1,5 mL Eppendorf tube to be used for index PCR.

Index PCR of cleaned PCR products

For this step, Illumina Forward Index Primers 7 and 8 were used (one for each sample), as well as Illumina Reverse Index Primer 8 for both samples. The index primers used are shown in Table 3.

Table 3 Overview of the Illumina forward and reverse primers used when performing Index PCR of cleaned PCR products in the AFLP process.

Primer name	Primer sequence
Illumina Forward Primer 7	5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC ACGACGCTCTTCCGATCTGTGAAAGACTGCGTACCAATTC-3'
Illumina Forward Primer 8	5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA CGACGCTCTTCCGATCTGTGGCCGACTGCGTACCAATTC-3'
Illumina Reverse Primer 8	5'CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCTGATGAGTCCTGAGTAA-3'

A master mix containing FIREPol RTL (1x), Forward Index Primer (0,2 µM), Reverse Index Primer (0,2 µM) and Template DNA (2 µL) was prepared with unique forward and reverse index primer pairs. Nuclease-free water was then added to each sample resulting in a total volume of 25 µL.

The following thermo-cycling protocol was used for the index PCR:

Activation at 95 °C for 5 min, 10 cycles of; denaturation at 95 °C for 30 sec, annealing at 56 °C for 1 min and elongation at 72 °C for 1 min, with a final elongation step at 72 °C for 7 min.

2.4 Nanopore MinION sequencing

Prior to a sequencing run with the MinION, a quality control (QC) of the flow cell being utilized is mandatory to reveal any faulty manufacturing or low quality of the flow cell in terms of the amount of active nanopores.

For the sequencing run, flow cell v. FLO-MIN106 R9 and MinION software MinKNOW v. 1.4.2 was used together with Metrichor.

2.4.1 Nanopore sequencing protocol

The Nanopore sequencing run was performed using the 'Rapid Sequencing Kit SQK-RAD002 following the 'Rapid Sequencing (SQK-RAD002) Protocol' for 1D MinION sequencing. All reads obtained from the Nanopore MinION during the sequencing run was template reads (1D).

Library preparation

Template DNA was prepared by mixing 200 ng genomic DNA and nuclease-free water in an 1,5 mL Eppendorf tube resulting in a total volume of 7,5 mL. The content was mixed thoroughly by inversion and spun down in a microfuge.

A tagmentation mix was made consisting of the template DNA and FRM with volumes being 7,5 µL and 2,5 µL respectively, resulting in a total volume of 10 µL. The mix was incubated in a thermal cycler at 30 °C for 1 min and then at 75 °C for 1 min.

An adapter ligation was performed by adding 1 µL of RAD to the tagmentation mix, mixed by inversion and spun down. 0,2 µL of Blunt/TA Ligase Master Mix was then added to the mix, mixed by inversion, spun down and incubated for 5 min at room temperature.

Priming the SpotON Flow Cell

The flow cell was inspected and made sure no air bubbles were present, and that the buffer was continuous throughout the flow cell.

A priming mix was made by mixing 480 µL RBF and 520 µL nuclease-free water.

The flow cell priming port was then loaded with 800 µL priming mix. After 5 min, the remaining primer mix was then loaded.

Library loading

This step was completed during the 5 min wait time in the above step and the resulting library loading mix was loaded immediately after the final primer mix was loaded.

A library loading mix was made consisting of RBF (25,5 µL), DNA library (11,0 µL), LLB (26,6 µL) and nuclease-free water (12 µL) resulting in a total volume of 75 µL.

The mix was then mixed by inversion and spun down before all of it being loaded dropwise into the SpotON sample port.

The sequencing run was then initiated through the MinKNOW software and ran uninterrupted for 48 hours.

2.5 Illumina MiSeq sequencing

2.5.1 AFLP Illumina library preparation

AFLP Index PCR products were processed and sequenced by Illumina MiSeq sequencing following Illumina's recommendations. Briefly explained, the Index PCR products were cleaned, normalized and pooled before being submitted for sequencing using the MiSeq Reagent Kit v3 (600 cycles).

2.6 Data analysis

2.6.1 Sequence data

After the Nanopore MinION sequencing run, the raw .fast5 files were converted to .fastq files and joined through the software package poRe (Watson et al. 2015) for R Studio v. 1.0.44.

The Canu assembly tool v. 1.4 was used for error-correcting and assembling the Nanopore MinION reads. Assembly was done on a virtual computer, Linux Ubuntu v. 16.04.2, through the virtual machine software Oracle VM VirtualBox v. 5.1.14, following Canu recommendations. No polishing step of the Nanopore data was used in this thesis.

2.6.2 WGS data assembly

WGS data was assembled *de novo* using Geneious v. 8.0.5 with default settings. Geneious performs *de novo* assembly using a greedy algorithm, similar to the one used for multiple sequence alignment. A blast-like algorithm is used to identify closely matching sequences and merging these into contigs. This process is repeated until all matching sequences have been matched into contigs. *De novo* assembly through Geneious only works well if there is high coverage in the data being used. Geneious will also always produce the same result due to the deterministic method used for the assembly.

2.6.3 Generation of errors in longest contig

Matlab v. R2016b was used to generate random mutations into the longest contig using the formula below:

$$Contig\ length * error\ rate = number\ of\ mutated\ nucleotides$$

Five different contigs with mutations assigned to random mutations were generated, ranging from 10 - 50 % error rate. Nucleotides located in these randomly assigned positions to-be-mutated were identified and wrong nucleotides were inserted following the pattern G → C, A → T, T → A and C → G. If the position had an ambiguous nucleotide, one of the four nucleotides were randomly inserted.

2.6.4 Generation of AFLP in silico fragments

Generating *in silico* AFLP fragments based on the longest contig obtained from *de novo* genomic DNA assembly. The Geneious analysis tool v. 8.0.5 was used for creating the restriction map. The enzymes EcoRI and MseI were selected to perform the restriction cutting, generating a table for theoretical positions of where they would cut. From the table of restriction cutting positions, AFLP fragments located between EcoRI and MseI within the length of 100-500 bp were selected and extracted using Matlab v. R2016b.

2.6.5 Use of in silico AFLP fragments to error correct longest contig

Generated AFLP fragments were mapped to contigs with varying error rate, using Geneious v. 10.1.3. The algorithm used for this is a mapper based on the seed and expand style. The reference is first indexed into locations containing words (series of bases of a specific length). Each fragment (read) is then processed separately, identifying if reads have any correlating words with the reference. These correlations are seed points, where the matching range is later expanded outwards, covering the whole read. A score is then assigned to each mapping based on the number of mismatches or gaps introduced during mapping.

3 Results

3.1 Overview of sequence data

The integrity of the genomic DNA is visualized in Figure 2. DNA concentration was measured to be 146 ng/µL using a Qubit fluorometer.

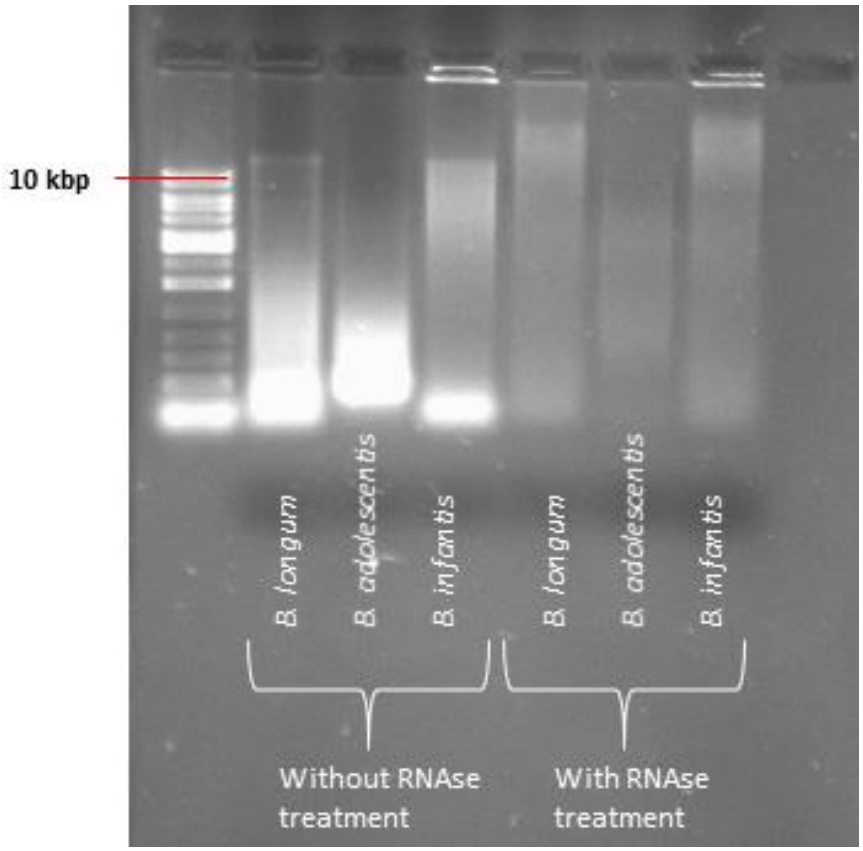


Figure 2. Gel electrophoresis of genomic DNA at 2x stock dilution on a 0,8 % agarose gel. A 1 kbp DNA Ladder was used.

3.2 Nanopore sequencing run

The QC performed on the MinION flow cell revealed the amount of active nanopores to be >1100, which is within the recommended range.

3.2.1 Choosing assembly software

Several tools for processing the Nanopore MinION data was considered for this thesis. A list of these tools is shown in Table 4.

Table 4. List of Nanopore data processing tools

Software name	Applications	Recently updated	Specific noise correction for long, noisy reads
Canu	A single molecule sequence assembler for high-noise data	Yes	Yes
Miniasm	Ultrafast *de novo* assembly for long noisy reads	No	Yes
NanoOK	Analysis tool for Nanopore sequencing data	Yes	No
Nanopolish	Software package for signal-level analysis of ONT sequencing data	No	No
NaS	Hybrid assembler generating synthetic long reads from Nanopore MinION and Illumina data	No	No
Poreseq	Error correction and variant calling algorithm for Nanopore sequencing	No	No

When considering the different tools, Canu was one of the most recently updated, including specific noise correction for long, noisy reads. Links to the different software are listed in Appendix 5.

3.2.2 Sequence assembly

Out of the 19867 raw reads obtained from the MinION sequencing run, only 1735 were accepted by Metrichor after base calling.

All Metrichor accepted reads were fed as an input to Canu for processing, but only 114 of the 1735 reads had an acceptable length for assembly, and of those 114, only 44 reads were used for the final assembly (Figure 3). The number of base pairs before error correction was 821973 bp, while after correction the number was 811875 bp. The 1621 reads not used for assembly had a total of 249781 bp, while the 114 reads accepted by Canu had a total of 835263 bp. An overview of Nanopore MinION sequencing read distribution and a graphical representation of Canu read error-correction can be viewed in Appendix 2 and 3 respectively.

According to the Canu assembly logs, both error correction and trimming were successful, but the final assembly failed to complete due to an unknown error. An example of a read used for assembly can be seen in Appendix 4.

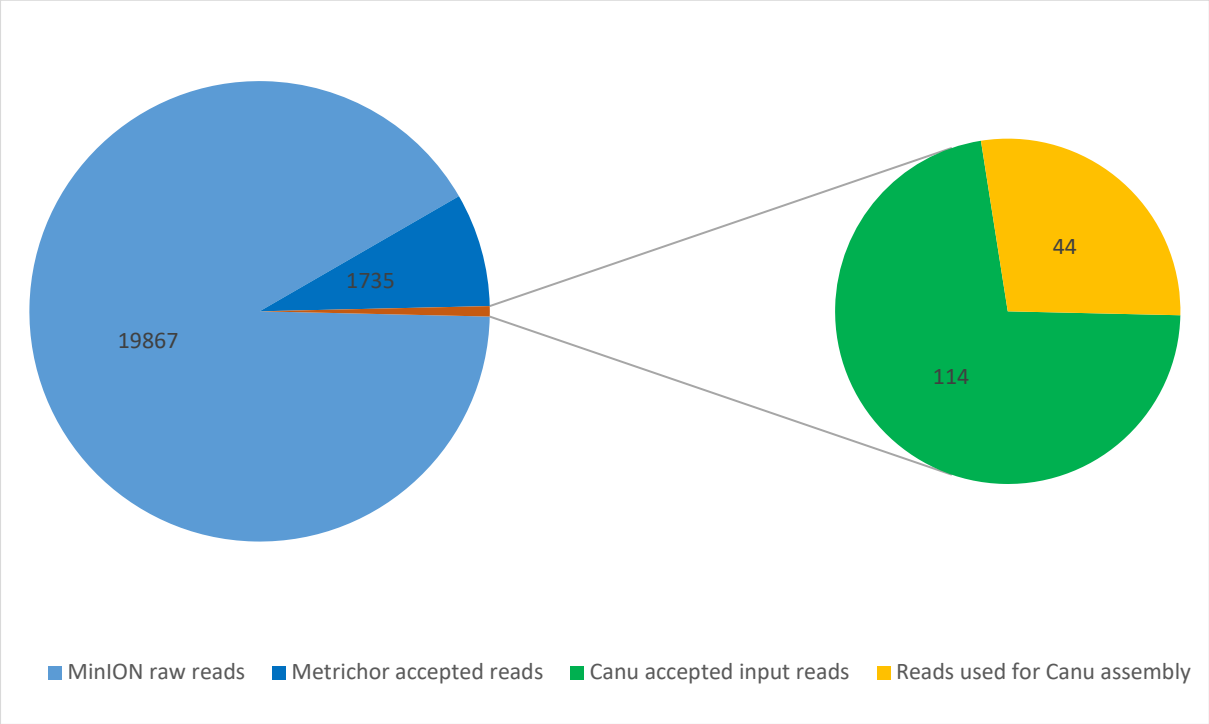


Figure 3. Graphical representation of the MinION data output and the number of reads used by Canu for genome assembly.

3.3 WGS data assembly

The *de novo* assembly of the WGS data obtained prior to the start of this thesis resulted in the generation of 332 contigs (Table 5), with ~1x genome coverage. The longest contig was ~290 kbp long and assembled from 30093 reads.

Table 5. Output from the *de novo* assembly of WGS data performed with Geneious.

Statistics	Unused reads	All contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	2678	332	332	71
Min length (bp)	35	106	106	1011
Median length (bp)	-	338	338	8101
Mean length (bp)	232	7415	7415	33455
Max length (bp)	301	289607	289607	289607
N50 length (bp)	-	93556	93556	93556
Number of contigs >= N50	-	9	9	9
Length sum (bp)	623013	2461829	2461829	2375336

3.4 Nanopore MinION data mapped towards WGS contigs

Nanopore sequencing reads of three different error thresholds (no error, 5 % and 20 %), were mapped towards WGS contigs. An overview of the mapping is displayed in Table 6, and examples of contig mapping are visualized in Figure 4.

Table 6. Overview of Nanopore MinION data alignment towards WGS contigs.

Error rate (%)	20	5	0
Number of contigs	25	9	24
Number of reads (mean)	39,52	1,56	53,42
Number of reads (SD)	177,26	1,07	238,45
Identical sites (average %)	65,42	98,74	55,86

The percentage of identical sites with 5 % error threshold was high (98,7 %), however, these sites represented only a minor fraction of the Nanopore reads (Figure 5).

Figure 4. Picture of the Nanopore data mapping towards a WGS contig in Geneious. The right picture is a part of the interface of Geneious, with the black lines representing all the nanopore reads that were mapped. The short black line inside the area marked by the green circle is the actual WGS contig. The top left picture (indicated by the green arrow) is a zoomed in version of the contig. The bottom left picture is a zoomed in version of the area inside the red circle.

Figure 5. Picture of the Nanopore data mapping towards a WGS contig in Geneious. The top picture of the figure is showing the whole contig, with the pink line representing the trimmed part of the Nanopore read. The middle picture is a more zoomed in version of the top picture, more clearly showing the small part of the read not trimmed. The bottom picture is showing the small part of the Nanopore read mapping towards the WGS contig.

3.5 AFLP

3.5.1 Gel electrophoresis of PCR products

The gel electrophoresis image of Index PCR products can be viewed in Figure 6.

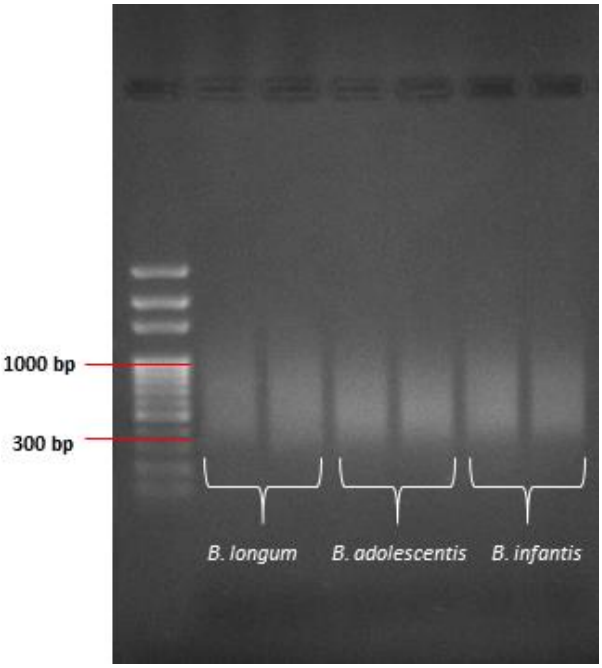


Figure 6. Gel electrophoresis of the Index PCR products generated during the AFLP procedure. A 100 bp DNA Ladder was used.

3.5.2 Generation of *in silico* AFLP fragments and mapping towards longest WGS contig

There were 484 *in silico* AFLP fragments generated from the longest contig. From those, 101 were fragments with EcoRI on one side and MseI on the other side, and only 7 of those were between 100-500 bp long. These 7 were used for alignment towards the longest contig with different error rates, ranging from 10-50 %, including the original contig with no error. Results from the alignment are displayed in Table 7.

Table 7. Overview of the *in silico* generated AFLP fragments and their alignment towards the longest contig from WGS *de novo* assembly containing different rates of error.

Error rate (%)	Number of aligned AFLP fragments	Identical sites (%)	Pairwise identity (%)	Minimum sequence length (bp)
0	7	100	100	216
10	7	90,9	90,9	216
20	7	80,0	80,0	216
30	7	69,6	69,6	216
40	7	60,5	60,5	216
50	3	44,1	44,3	313

3.5.3 Lab-generated AFLP fragments

A total of 131525 sequence pairs were generated during the sequencing run (Table 8).

Table 8. Results from the lab-generated AFLP fragments sequenced by Illumina MiSeq.

			Duplicate 1	Duplicate 2
Number of read pairs			79129	52396
Length	Mean ± SD	FWD	246,2 ± 57,2	231,1 ± 65,9
		REV	257,9 ± 57,1	245,3 ± 67,0
	Min	FWD	21	24
		REV	27	30
	Max	FWD	279	279
		REV	285	285

Due to time constraint, these data were not analysed further.

4 Discussion

4.1 Evaluation of Nanopore MinION sequencing results

Since the announcement of the MinION in 2012, and its later beta-release in 2014, several studies and evaluations of the platform has been performed (Ip et al. 2015). Both the sequencing chemistry and associated software of the MinION has undergone major changes and overhauls to improve the performance of the platform (Jain et al. 2017; Lu et al. 2016). A study by Laver et al. (2015) estimated the sequencing error rate of the MinION to be 38,2 % after base calling, which is regarded as highly inaccurate. Another study by Jain et al. (2016) claimed the error rate to be <8 %, which demonstrates the positive influence of sequencing chemistry and software improvements, while also strengthening MinION's potential.

Despite the rapid development and improvements of sequencing technologies and associated software, incorrect reads during sequencing are still an issue for TGS technologies, specifically Nanopore reads (Ma et al. 2017). However, the constant improvement of accuracy in data obtained through sequencing, and algorithms to process these data, have made it possible to sequence a human genome and obtain a highly complete and contiguous *de novo* assembly, as shown by Jain et al. (2017). High sequencing accuracy is critical when performing *de novo* sequence assembly since fragments with errors might not be recognized as they should be, thus they could either be wrongly placed during the assembly or be so erroneous that fragments become unrecognizable. This could confound assembly, resulting in generation of inaccurate contigs (Utturkar et al. 2014). However, TGS technologies have been demonstrated to generate highly accurate assemblies when combined with complementary data (hybrid assembly) (Koren et al. 2017).

The genome size of *B. longum* is ~2 Mbp (Schell et al. 2002), while the total number of Metrichor accepted reads generated by the MinION consisted of ~1 Mbp, meaning the coverage of the sequencing run was below 1x. To achieve high-quality *de novo* assemblies with Nanopore data alone, it is recommended to have above 20x coverage (Koren et al. 2017).

The single sequencing run performed for this thesis generated almost exclusively reads consisting of long stretches of A's and T's, resulting in low quality data. More runs could have been performed, but time was a limiting factor. The highly erroneous sequences generated

this sequencing run could be a result of mainly two reasons: First, the recommended length of genomic DNA fragments when performing a sequencing run with the MinION is >30 kbp by ONT's own recommendation (ONT 2017b), and the DNA fragments used for this sequencing run was below that. Second, sequencing performed with the MinION seem to have issues regarding G/C-rich genomes (Laver et al. 2015). Regions which are G/C-rich tend to have a lower base sequencing accuracy and it is common to observe sequencing difficulties in these regions, with difficulties being a larger variation and less representation in sequencing coverage (Lu et al. 2016). However, G/C-bias has also been reported as a result from base calling error bias, favouring an uncertain G/C read into A/T (Deschamps et al. 2016). The MinION generated reads for this thesis are thus considered to be artefact sequences as *B. longum* has an established genomic G/C-value of ~60 %. Artefact sequences are sequences that are generated through for example sequencing preparations or during the run itself due to various reasons, that does not represent the original sequence (Do & Dobrovic 2012). Artefact sequences present after a sequencing run with the MinION is not unfamiliar (Goodwin et al. 2015), and reads longer than the actual DNA fragment being sequenced have also been observed, likely caused by the MinION instrument, although not verified (Wang et al. 2015).

Another challenge regarding accurate sequence assembly is computational bottlenecks, more precisely computing algorithms and treatment of sequence data (Nagarajan & Pop 2009). Longer reads introduced by TGS tend to have a higher computational requirement than the shorter SGS reads (Jansen et al. 2017). According to Jansen et al. (2017), assembly of the human genome from PacBio SMRT reads would take thousands of CPU hours, which should encourage the development of how to more efficiently process long-read data. The process of assembly for TGS is very different than SGS. Exact algorithms had to be used for SGS data processing, and all the work and development of Illumina data processing does not work well with longs reads (Simpson 2015). New tools tailored for TGS data processing have been developed and are being improved to further advance the quality of data assembly from TGS data alone, but also methods used in a hybrid approach, combining sequencing data from both SGS and TGS. ONT has not yet developed an own pipeline for their Nanopore data, however, the single-molecule assembler, Canu, that is able to process Nanopore MinION sequencing data, has been developed (Koren et al. 2017).

4.1.1 Data processing with Canu

For this thesis, Canu was chosen as the preferred software for processing the Nanopore MinION data. The decision was based on other reviews of the software, how it processes the Nanopore data, that it was one of the most recently updated software packages available, and it had specific noise correction for long, noisy reads (Koren et al. 2017). However, the process of finding a suitable tool for processing the Nanopore data was slightly challenging. ONT do not supply one themselves, nor do they provide any specific recommendations of their own. Several tools for processing data was mentioned through other papers in regards to TGS technologies (Jain et al. 2016; Lu et al. 2016), and the decision of using Canu was made after conducting research on each of these tools.

Canu was able to error correct the MinION reads, and the original read length compared with the corrected read length yielded almost identical results. The correlation between the data are close to linear apart from a few evident outliers in the bottom left corner and one in the top right corner (Appendix 2). In general, most reads were only corrected with a few bps except for six reads which were highly corrected. For the assembly, only 44 reads were used, resulting in a mere 0,22 % of the raw reads obtained from the MinION sequencing run. All reads not used were discarded either because of little or no overlap between the reads or due to insufficient read length. The low number of reads used for assembly, combined with the low genome coverage from the sequencing run, is not sufficient for accurate *de novo* assembly with Nanopore data as previously mentioned.

4.2 Nanopore data mapping towards WGS data

For this thesis, Illumina MiSeq data was used as a reference because of the high sequencing accuracy (Quail et al. 2012; Salipante et al. 2014). The results from the Nanopore data mapping towards the WGS reference contigs shows the low Nanopore data quality. Even though numerous reads mapped, the mapping was still bad or could not be trusted as it was only a short part of the read that mapped, and most of the reads did not map towards the contig, but to each other. Furthermore, these self-mapping do not carry any significance either because of the high density of the nucleotides A and T. The number of contigs that the Nanopore data mapped towards were low considering the total number of available contigs. The average number of reads that correlated to each contig was very low, additionally, the SD

of the reads used were substantial, meaning the distribution of reads between contigs were highly uneven. At both 0 and 20 % error thresholds, the number of nucleotides mapping correctly (identical sites) were low. This indicates different nucleotides at the same position, further demonstrating the low quality of the Nanopore data.

4.3 Assessment of AFLP error correction

AFLP data was generated for this thesis, but was not used due to time constrain.

The *in silico* experiment of mapping generated AFLP fragments towards contigs with different error rates demonstrates the potential for these fragments to improve genome assembly. Up to an error rate of 40 % in the contig, all AFLP fragments mapped without error. At 50 % error, only three of the fragments mapped, with one also being mapped at the wrong position, leading to an increase in contig length due to the fragment extending the contig. The error rate of Nanopore reads range from ~8 % (Jain et al. 2016) to ~40 % (Laver et al. 2015), meaning the suitability of the fragments should not be compromised as they in general should improve accuracy of Nanopore data and their error prone reads due to their ability to correctly map, even at 40 % error rate.

4.4 Prospects

While the *in silico* experiment of using generated AFLP fragments to correct long error prone reads was successful, more time was needed to use real AFLP data to determine their suitability.

For proper evaluation of the MinION, more sequencing runs need to be performed since data from one run is not enough to determine the potential of the platform.

From the MinION sequencing data obtained for this thesis, G/C bias was evident, thus improvement to the sequencing technology and optimization of existing protocols are necessary. However, the Nanopore technology is rapidly developing, and potentially these problems could be solved in the near future.

5 Conclusions

The MinION sequencing run performed for this thesis generated highly erroneous sequencing results, which prevented the downstream analysis. For this reason, it is difficult to say anything conclusive about the performance of the MinION. Although we did not succeed in generating long Nanopore reads, we evaluated to what extent AFLP fragments can be used through an *in silico* experiment. AFLP fragments were accurately mapped on long contigs with up to a 40 % error rate, suggesting their high potential for error-correction. Better protocols or improvements to the sequencing technology is needed due to the current G/C-bias.

References

Ajmone-Marsan, P., Valentini, A., Cassandro, M., Vecchiotti-Antaldi, G., Bertoni, G. & Kuiper, M. (1997). AFLP markers for DNA fingerprinting in cattle. *Anim Genet*, 28 (6): 418-26.

Ashton, P. M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J. & O'Grady, J. (2015). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol*, 33 (3): 296-300.

Bahassi el, M. & Stambrook, P. J. (2014). Next-generation sequencing technologies: breaking the sound barrier of human genetics. *Mutagenesis*, 29 (5): 303-10.

Behjati, S. & Tarpey, P. S. (2013). What is next generation sequencing? *Arch Dis Child Educ Pract Ed*, 98: 236-238.

Berglund, E. C., Kiialainen, A. & Syvanen, A. C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Investig Genet*, 2: 23.

Berlin, K., Koren, S., Chin, C. S., Drake, J. P., Landolin, J. M. & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33: 623-630.

Biomatters. (2005). *Geneious - A powerful and comprehensive suite of molecular biology and NGS analysis tools*. Available at: http://www.geneious.com/ (accessed: 14 Apil 2017).

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddelston, J., Eichler, E. E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10: 563-569.

Deschamps, S., Mudge, J., Cameron, C., Ramaraj, T., Anand, A., Fengler, K., Hayes, K., Llaca, V., Jones, T. J. & May, G. (2016). Characterization, correction and de novo assembly of an Oxford Nanopore genomic dataset from *Agrobacterium tumefaciens*. *Sci Rep*, 6: 28625.

Do, H. & Dobrovic, A. (2012). Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase. *Oncotarget*, 3 (5): 546-58.

Dunne, W. M., Jr., Westblade, L. F. & Ford, B. (2012). Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. *Eur J Clin Microbiol Infect Dis*, 31 (8): 1719-26.

Ekblom, R. & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications* (9): 7: 1026-1042.

Feng, Y., Zhang, Y., Ying, C., Wang, D. & Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genomics Proteomics Bioinformatics*, 13 (1): 4-16.

Gibson, J. R., Slater, E., Xerry, J., Tompkins, D. S. & Owen, R. J. (1998). Use of an amplified-fragment length polymorphism technique to fingerprint and differentiate isolates of Helicobacter pylori. *Journal of Clinical Microbiology*, 36 (9): 2580-2585.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C. & McCombie, W. R. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res*, 25 (11): 1750-6.

Illumina. (2010). Illumina Sequencing Technology. Available at: https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf (accessed: 31 March 2017).

Illumina. (2016a). *A high-resolution view of the entire genome*. Available at: https://www.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing.html (accessed: 27 December 2016).

Illumina. (2016b). Illumina Two-Channel SBS sequencing technology. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/techspotlights/techspotlight_two-channel_sbs.pdf (accessed: 10 April 2017).

Illumina. (2016c). An Introduction to Next-Generation Sequencing Technology. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf (accessed: 10 April 2017).

Illumina. (2016d). Nextera DNA Library Prep Reference Guide. Available at: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nexteradna/nextera-dna-library-prep-reference-guide-15027987-01.pdf (accessed: 30 April 2017).

Ip, C. L., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., Leggett, R. M., Eccles, D. A., Zalunin, V., Urban, J. M., et al. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Res*, 4: 1075.

Istace, B., Friedrich, A., d'Agata, L., Faye, S., Payen, E., Beluche, O., Caradec, C., Davidas, S., Cruaud, C., Liti, G., et al. (2017). de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience*, 6 (2): 1-13.

Jain, M., Olsen, H. E., Paten, B. & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*, 17: 239.

Jain, M., Koren, S., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., et al. (2017). Nanopore sequencing and assembly of a human genome with ultra-long reads. *BioRxiv*.

Jansen, H. J., Liem, M., Jong-Raadsen, S. A., Dufour, S., Weltzien, F. A., Swinkels, W., Koelewijn, A., Palstra, A. P., Pelster, B., Spaink, H. P., et al. (2017). Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *BioRxiv*.

Judge, K., Hunt, M., Reuter, S., Tracey, A., Quail, M. A., Parkhill, J. & Peacock, S. J. (2016). Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology. *Microbial Genomics*.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28 (12): 1647-9.

Koren, S., Harhay, G. P., Smith, T. P., Bono, J. L., Harhay, D. M., McVey, S. D., Radune, D., Bergman, N. H. & Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol*, 14 (9): R101.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*, 27 (5): 722-736.

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K. & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif*, 3: 1-8.

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., McCombie, W. R. & Schatz, M. C. (2016). Third-generation sequencing and the future of genomics. *BioRxiv*.

Lu, H., Giordano, F. & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics*, 14 (5): 265-279.

Ma, X., Stachler, E. & Bibby, K. (2017). Evaluation of Oxford Nanopore MinION Sequencing for 16S rRNA Microbiome Characterization. *BioRxiv*.

Madoui, M. A., Engelen, S., Cruaud, C., Belser, C., Bertrand, L., Alberti, A., Lemainque, A., Wincker, P. & Aury, J. M. (2015). Genome assembly using Nanopore-guided long and error-free DNA reads. *Bmc Genomics*, 16.

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470 (7333): 198-203.

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)*.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437 (7057): 376-80.

Masny, A. & Pluciennicza k, A. (2001). Fingerprinting of bacterial genomes by amplification of DNA fragments surrounding rare restriction sites. *Biotechniques*, 31 (4): 930-936.

Maughan, P. J., Saghai Maroof, M. A., Buss, G. R. & Huestis, G. M. (1996). Amplified fragment length polymorphism (AFLP) in soybean: species diversity, inheritance, and near-isogenic line analysis. *Theor Appl Genet*, 93 (3): 392-401.

McCarthy, A. (2010). Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chem Biol*, 17 (7): 675-6.

Meldrum, C., Doyle, M. A. & Tothill, R. W. (2011). Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin Biochem Rev*, 32 (4): 177-95.

Metrichor. (2015). *Internet of Living Things*. Available at: https://vimeo.com/125796581 (accessed: 12 January 2017).

Mikheyev, A. S. & Tin, M. M. Y. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources*, 14 (6): 1097-1102.

Miller, J. R., Koren, S. & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95 (6): 315-27.

Munroe, D. J. & Harris, T. J. R. (2010). Third-generation sequencing fireworks at Marco Island: advances in sequencing platforms promise to make this technology more accessible. *Nat Biotechnol*, 28 (5): 426+ (accessed: 6 March 2017).

Nagarajan, N. & Pop, M. (2009). Parametric complexity of sequence assembly: theory and applications to next generation sequencing. *J Comput Biol*, 16 (7): 897-908.

ONT. (2015). *MinKNOW - Primary Data Analysis*. Available at: https://nanoporetech.com/analyse (accessed: 4 April 2017).

ONT. (2017a). Nanopores allow direct sequencing of full-length RNA strands and modified RNA nucleotides. Available at: https://nanoporetech.com/publications/nanopores-allow-direct-sequencing-full-length-rna-strands-and-modified-rna-nucleotides (accessed: 2 May 2017).

ONT. (2017b). Rapid Sequencing protocol (SQK-RAD002). Available at: https://community.nanoporetech.com/protocols/rapid-sequencing-sqk-rad002/v/rse_9018_v2_revk_21nov2016/overview-of-the-rapid-sequ.

Phillippy, A. M., Schatz, M. C. & Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol*, 9 (3): R55.

Phillippy, A. M., Koren, S. & Walenz, B. P. (2015). *Canu assembly tool*. Available at: https://github.com/marbl/canu.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13: 341.

Reuter, J. A., Spacek, D. V. & Snyder, M. P. (2015). High-throughput sequencing technologies. *Mol Cell*, 58 (4): 586-597.

Rhoads, A. & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics*, 13 (5): 278-89.

Risse, J., Thomson, M., Patrick, S., Blakely, G., Koutsovoulos, G., Blaxter, M. & Watson, M. (2015). A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. *Gigascience*, 4: 60.

Salipante, S. J., Kawashima, T., Rosenthal, C., Hoogestraat, D. R., Cummings, L. A., Sengupta, D. J., Harkins, T. T., Cookson, B. T. & Hoffman, N. G. (2014). Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Appl Environ Microbiol*, 80 (24): 7583-91.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*, 22 (3): 557-67.

Sambrook, J. & Russell, D. W. (2006). Purification of nucleic acids by extraction with phenol:chloroform. *CSH Protoc*, 2006 (1).

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74 (12): 5463-7.

Schadt, E. E., Turner, S. & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, 19 (2): R227-R240.

Schell, M. A., Karmirantzou, M., Snel, B., Vilanova, D., Berger, B., Pessi, G., Zwahlen, M. C., Desiere, F., Bork, P., Delley, M., et al. (2002). The genome sequence of Bifidobacterium longum reflects its adaptation to the human gastrointestinal tract. *Proc Natl Acad Sci U S A*, 99 (22): 14422-7.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T. & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43 (6).

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat Methods*, 5 (1): 16-8.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol*, 26 (10).

Simpson, J. T. (2015). Error correction, assembly and consensus algorithms for MinION data. Available at: https://londoncallingconf.co.uk/lc/2015-plenary#128666688 (accessed: 23 January 2017).

Utturkar, S. M., Klingeman, D. M., Land, M. L., Schadt, C. W., Doktycz, M. J., Pelletier, D. A. & Brown, S. D. (2014). Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*, 30 (19): 2709-16.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., et al. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res*, 23 (21): 4407-14.

Vuylsteke, M., Peleman, J. D. & van Eijk, M. J. (2007). AFLP technology for DNA fingerprinting. *Nat Protoc*, 2 (6): 1387-98.

Wang, J., Moore, N. E., Deng, Y. M., Eccles, D. A. & Hall, R. J. (2015). MinION nanopore sequencing of an influenza genome. *Front Microbiol*, 6: 766.

Watson, M., Thomson, M., Risse, J., Talbot, R., Santoyo-Lopez, J., Gharbi, K. & Blaxter, M. (2015). poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics*, 31 (1): 114-5.

Zimin, A. V., Puiu, D., Luo, M. C., Zhu, T., Koren, S., Marcais, G., Yorke, J. A., Dvorak, J. & Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res*, 27 (5): 787-792.

Appendix 1: Instruments and reagents used for this thesis

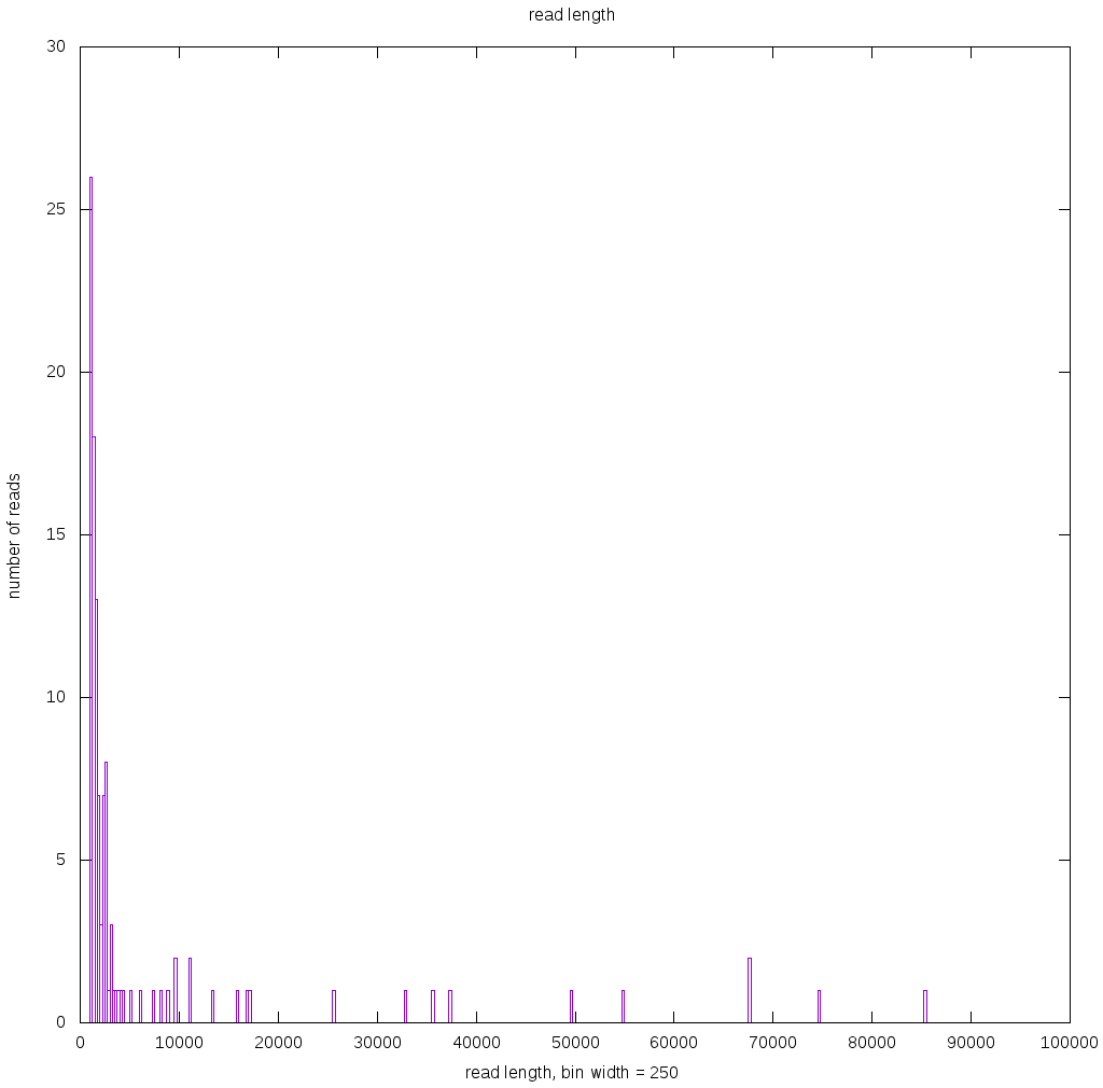
Supplementary Table 1. List of commercial reagents and kits used for this thesis

Name	Manufacturer
Rapid Sequencing Kit SQK-RAD002	Oxford Nanopore Technologies Ltd
EcoRI enzyme	New England Biolabs Inc.
MseI enzyme	New England Biolabs Inc.
Cut Smart buffer	New England Biolabs Inc.
Nuclease-free water	Amresco Inc.
EcoRI adapter mix	New England Biolabs Inc.
MseI adapter mix	New England Biolabs Inc.
T4 DNA ligase	New England Biolabs Inc.
T4 reaction buffer	New England Biolabs Inc.
5x HOTFIREPol	Solis BioDyne Inc.
EcoRI primer	Sigma-Aldrich Co. LLC.
MseI primer	Sigma-Aldrich Co. LLC.
5x FIREPol RTL	Solis BioDyne Inc.
Illumina Forward Index Primer 7	Illumina Inc.
Illumina Forward Index Primer 8	Illumina Inc.
Illumina Reverse Index Primer 8	Illumina Inc.
AMPure XP beads (1x)	Life Sciences
Ethanol	Kemetyl AS
PeqGreen RNA/DNA Dye	PeqLab Inc.
1 kbp DNA Ladder	Solis BioDyne Inc.
100 bp DNA Ladder	Solis BioDyne Inc.
Quant-iT Buffer	Invitrogen Inc.
Quant-iT Reagent	Invitrogen Inc.
Qubit dsDNA HS Standard 1	Invitrogen Inc.
Qubit dsDNA HS Standard 2	Invitrogen Inc.

Supplementary Table 2. List of commercial equipment used for this thesis

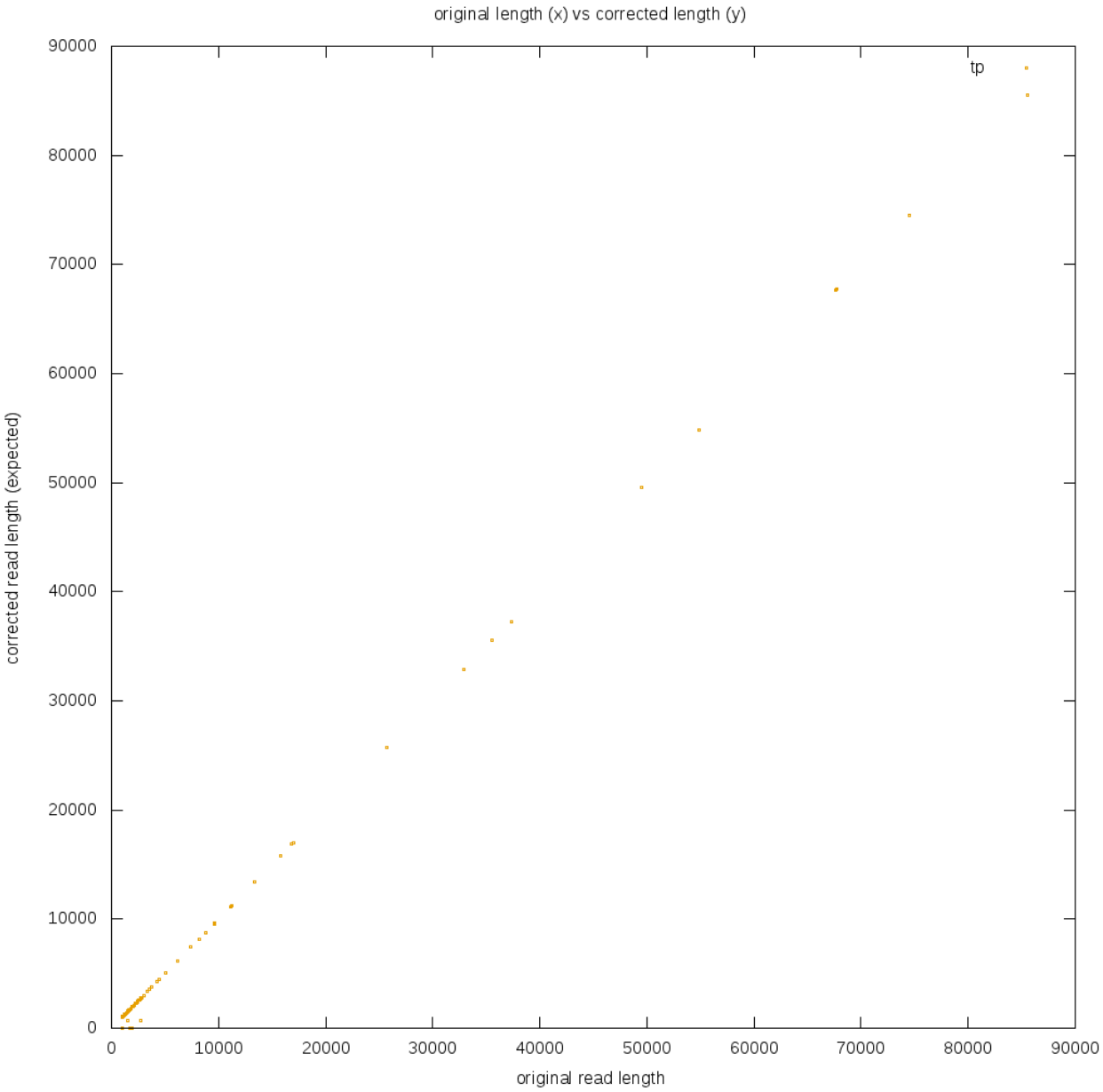
Name	Manufacturer
Nanopore MinION	Oxford Nanopore Technologies Ltd
MiSeq System	Illumina Inc.
Qubit fluorometer	Invitrogen Inc.
2720 Thermal Cycler	Applied Biosystems Inc.
Gel system: Mini-Sub Cell GT	Bio Rad

Appendix 2: Distribution of MinION reads



Supplementary Figure 1. Canu output graph. Distribution of Nanopore MinION reads from one genomic DNA sequencing run. The x-axis shows how many bases are associated with each read and the y-axis shows the number of reads connected to each length.

Appendix 3: Nanopore data correction graph



Supplementary Figure 2. Canu output graph. The original Nanopore MinION read lengths data compared with read lengths after being corrected by Canu.

Appendix 4: Example read used for assembly by Canu

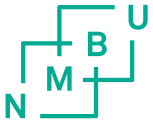
```
>read1_0
CCTTATAGATATTGTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
```

```
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATT
ATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTGTTATTATTGTTATTGTT
ATTGTTATTGTTGTTATTGTTGTTATTGTTATTGTTGTA
ATTATTATTATTATTATTATTATTATTATTGTTATTTTA
```

Appendix 5: Links to the different Nanopore data processing tools

Appendix Table 3. Links to the different Nanopore data processing tools considered for this thesis. Table 4 includes more information on each of these tools.

Software name	Link
Canu	https://github.com/marbl/canu
Miniasm	https://github.com/lh3/miniasm
NanoOk	https://github.com/TGAC/NanoOK
Nanopolish	https://github.com/jts/nanopolish
NaS	https://github.com/institut-de-genomique/NaS
Poreseq	https://github.com/tszalay/poreseq



Norges miljø- og biovitenskapelig universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway