



**FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA**

**DEPARTMENT OF
CYBERNETICS**

Bachelor's Thesis

Analysis of Tools and Pipeline for Processing Long-Read Sequencing Data

Barbora Soukupová



**FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA**

**DEPARTMENT OF
CYBERNETICS**

Bachelor's Thesis

Analysis of Tools and Pipeline for Processing Long-Read Sequencing Data

Barbora Soukupová

Thesis advisor

Ing. Lucie Houdová, Ph.D.

© 2023 Barbora Soukupová.

All rights reserved. No part of this document may be reproduced or transmitted in any form by any means, electronic or mechanical including photocopying, recording or by any information storage and retrieval system, without permission from the copyright holder(s) in writing.

Citation in the bibliography/reference list:

SOUKUPOVÁ, Barbora. *Analysis of Tools and Pipeline for Processing Long-Read Sequencing Data*. Pilsen, Czech Republic, 2023. Bachelor's Thesis. University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics. Thesis advisor Ing. Lucie Houdová, Ph.D.

ZÁPADOČESKÁ UNIVERZITA V PLZNI

Fakulta aplikovaných věd
Akademický rok: 2022/2023

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení: **Barbora SOUKUPOVÁ**
Osobní číslo: **A21B0440P**
Studijní program: **B0714A150005 Kybernetika a řídicí technika**
Specializace: **Umělá inteligence a automatizace**
Téma práce: **Analýza nástrojů a pipeline pro zpracování long-read sequencing dat**
Zadávající katedra: **Katedra kybernetiky**

Zásady pro vypracování

1. Seznamte s metodami získávání DNA dat (zaměřte se na LRS – Long-Read Sequencing).
2. Prostudujte datové formáty LRS technologií.
3. Proveďte rešerši dostupných přístupů a analytických nástrojů z pohledu specifického využití.
4. Zhodnoťte možnosti a přínosy využití LRS pro budoucí klinické uplatnění.

Rozsah bakalářské práce: **30-40 stránek A4**
Rozsah grafických prací:
Forma zpracování bakalářské práce: **tištěná**
Jazyk zpracování: **Angličtina**

Seznam doporučené literatury:

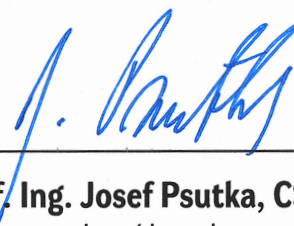
Dle doporučení vedoucí práce.

Vedoucí bakalářské práce: **Ing. Lucie Houdová, Ph.D.**
Katedra kybernetiky

Datum zadání bakalářské práce: **17. října 2022**
Termín odevzdání bakalářské práce: **22. května 2023**



Doc. Ing. Miloš Železný, Ph.D.
děkan



Prof. Ing. Josef Psutka, CSc.
vedoucí katedry

V Plzni dne 17. října 2022

Declaration

I hereby declare that this Bachelor's Thesis is completely my own work and that I used only the cited sources, literature, and other resources. This thesis has not been used to obtain another or the same academic degree.

I acknowledge that my thesis is subject to the rights and obligations arising from Act No. 121/2000 Coll., the Copyright Act as amended, in particular the fact that the University of West Bohemia has the right to conclude a licence agreement for the use of this thesis as a school work pursuant to Section 60(1) of the Copyright Act.

Plzeň, on 22 May 2023

.....

Barbora Soukupová

The names of products, technologies, services, applications, companies, etc. used in the text may be trademarks or registered trademarks of their respective owners.

Abstract

Long-read sequencing is a technology that has revolutionised the field of genomics. Unlike previously used sequencing techniques, which are able to read only short fragments of DNA, long-read sequencing can produce reads of tens of thousands of bases in length, providing a new and more comprehensive view of the genome. This thesis introduces the principles of long-read sequencing and its advantages and disadvantages over other sequencing methods. It provides an analysis of the tools and pipelines that are available for long-read sequencing data analysis as well as the current applications.

Abstrakt

Long-read sekvenování je technologie, která způsobila revoluci v oblasti genomiky. Na rozdíl od dříve používaných sekvenačních technik, které dokážou přečíst pouze krátké úseky DNA, long-read sekvenování dokáže číst až desítky tisíc bází najednou, což poskytuje nový a komplexnější pohled na genom. Tato práce představuje principy long-read sekvenování a jeho výhody a nevýhody oproti jiným metodám sekvenování. Cílem této práce je poskytnout přehled dostupných nástrojů a pipeline, které jsou k dispozici pro analýzu dat z long-read sekvenování, a také shrnout současné aplikace.

Keywords

DNA Sequencing • Long-Read Sequencing • Next-Generation Sequencing • Long-Read Data Analysis • Data Analysis Tools • Data Analysis Pipeline

Acknowledgement

I would like to express my gratitude to my advisor Ing. Lucie Houdová, PhD, who guided me throughout this project. I would also like to thank Mgr. Robin Klieber for his expert consultation and Mgr. Marie Paulusová Tesková for proofreading.

Contents

1	Introduction	3
2	DNA Sequencing Overview	5
2.1	DNA Replication	5
2.2	Sanger Sequencing	5
2.3	Next-Generation Sequencing	7
2.3.1	Illumina Sequencing	7
2.4	Long-Read Sequencing	8
2.4.1	PacBio SMRT Sequencing	8
2.4.2	Nanopore Sequencing	10
3	Comparison of Sequencing Methods	11
3.1	Read Length and Run Time	11
3.2	Sequencing Accuracy	12
3.2.1	Phred Score	12
3.2.2	Accuracy Comparison	12
3.3	Cost	13
4	Data Formats	15
4.1	SAM/BAM Format	15
4.1.1	Header Section	15
4.1.2	Alignment Section	15
4.2	HDF	16
4.3	FASTA/FASTQ	17
4.3.1	FASTA	17
4.3.2	FASTQ	17
4.4	PacBio Data Format	18
4.5	Nanopore Data Formats	18
4.5.1	FAST5	18
4.5.2	POD5	19

5	Long-Read Sequencing Pipeline and Available Tools	21
5.1	Common Steps in Long-Read Data Analysis	21
5.1.1	Basecalling	21
5.1.2	Quality Control	23
5.1.3	Read Alignment	24
5.1.4	Error Correction	25
5.1.5	Other Tools	25
6	Applications of Long-Read Sequencing Technologies	27
6.1	<i>De Novo</i> Assembly	27
6.2	Variant Calling	28
6.3	Epigenetics	29
6.4	Direct RNA Sequencing	31
6.5	Field Laboratory	31
6.6	Applications of Long-Read Sequencing in Cancer Genomics	32
6.6.1	Detecting Genetic Aberrations in Human Cancer	32
6.6.2	HLA Typing	32
7	Conclusion	33
	List of Terms and Abbreviations	35
	List of Figures	37
	List of Tables	39
	Bibliography	41

Introduction

1

DNA sequencing is the process of determining the exact order of nucleotides within a DNA molecule. Since its discovery, more than half a century ago, this technology has revolutionised the field of molecular biology, allowing scientists to study the genetic information contained within an organism's DNA.

The evolution of sequencing technologies can be divided into three generations [1]. The most important technology of the first generation is undoubtedly Sanger sequencing. This technique was responsible for what would arguably become the largest and most important biomedical project of the 20th Century - completion of the full human genome. Thirteen years and approximately 3 billion dollars later, the International Human Genome Sequencing Consortium announced the finalised near-completed sequence of the human genome [2]. Albeit impressive, it also showcased the main limitations of Sanger sequencing - speed and cost.

In the first decade of the 21st century, the sequencing field was swiftly overtaken by next-generation sequencing (NGS) technologies. This is mainly due to the increased speed, throughput and decreased costs. In the past years, next-generation sequencing has evolved drastically, achieving various milestones in the sequencing field as well as lowering the price per human genome to \$1000. Although NGS is nowadays considered a standard for a number of applications, there are still issues to overcome. [3]

In recent years a third generation of sequencing technologies also known as long-read sequencing (LRS) has come into the spotlight. Promising to tackle some of the most challenging obstacles in the sequencing field, it has been chosen as the Method of the year 2022 by Nature Methods [4]. The main advantage it holds over its predecessors is the ability to produce very long reads, which is especially suitable for several applications such as *de novo* assembly or structural variant detection.

This thesis aims to provide a comprehensive review of the current state of LRS technology and its applications. It will focus on the most prominent LRS technologies, their advantages and disadvantages and their comparison with NGS and Sanger sequencing. It will provide a summary of the data formats currently used by LRS technologies as well as the available state-of-the-art tools that are commonly utilised.

1. Introduction

The thesis will also explore the challenges associated with long-read sequencing, such as the complexity of data analysis and the limitations of current sequencing technologies. An overview of contemporary applications with a particular focus on clinical use will be given along with possible future outlooks.

DNA Sequencing Overview

2

In this chapter, first, the process of DNA replication is introduced, which is essential for most of the sequencing techniques within this thesis, and the most widespread sequencing approaches (Sanger sequencing and Next-generation sequencing) are described. The chapter proceeds to cover long-read sequencing, explaining the mechanics behind long-read sequencing made by Pacific Biotechnologies (PacBio) and Oxford Nanopore Technologies (ONT).

2.1 DNA Replication

DNA replication is the process of creating two identical DNA molecules from one original DNA template. A DNA strand is made from a specific sequence of 4 different 'building blocks' called nucleotides, those nucleotides differ only in the nitrogenous base they contain (Adenine, Cytosine, Guanine, or Thymine). When DNA is in its double-helix form, the opposing nucleotides are paired through hydrogen bonds forming a so-called base pair (bp), this pairing, however, has certain rules - Adenine can be only paired with Thymine and Cytosine can be only paired with Guanine. Therefore if the exact base sequence of one strand is known, the sequence of the other strand is also known - they are complementary. The double helix that forms the DNA molecule is unravelled during replication, resulting in two individual strands. Each of those strands then serves as a template for a new DNA molecule. A specific enzyme called DNA Polymerase then starts incorporating appropriate nucleotides to complete the DNA synthesis. This process continues until the second strands are fully synthesized, forming once again a double-helix structure (Figure 2.1). [5]

2.2 Sanger Sequencing

The first person to develop a widely commercially successful, and arguably the most famous, DNA sequencing method was Frederick Sanger in 1977. This discovery

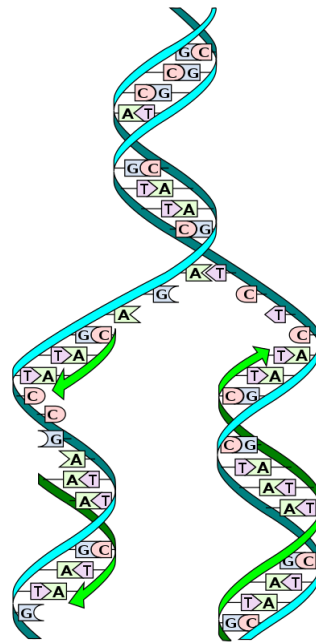


Figure 2.1: DNA replication schema [6]

revolutionised the field of genomics and his technique dominated DNA sequencing for the next 30 years launching the era of first-generation sequencing. This method uses the principle of DNA replication. [7]

To successfully perform Sanger sequencing, first a double-stranded DNA that needs to be sequenced is amplified thousands of times using polymerase chain reaction (PCR). Subsequently, the amplified DNA is denatured forming a single-stranded DNA. Afterwards, all the multiplied strands are put into a test tube along with standard nucleotides and special fluorescent labelled dideoxynucleotides (ddNTPs: ddATP, ddCTP, ddGTP, ddTTP) in a much lower concentration. Using DNA polymerase replication for each of the strands is started. Once in a while, the DNA polymerase incorporates ddNTP instead of a standard nucleotide, which terminates the DNA synthesis due to the chemical structure of ddNTPs. After this process is finished, a number of DNA fragments of varying lengths ending with fluorescently labelled ddNTP, are left. Finally, all those fragments are ordered by length and the final sequence is determined thanks to the differences in fluorescent signals of each ddNTP. [8]

This process was performed manually until 1987 when Applied Biosystems (now ThermoFisher) introduced a machine, that would automate the sequencing method improving both accuracy and speed. Even though other companies have since developed their own automated systems based on Sanger sequencing, Applied Biosystems is the only one that has not been discontinued. [9]

2.3 Next-Generation Sequencing

Next-generation sequencing, also known as massively parallel sequencing, is an umbrella term used for techniques that have vastly improved the speed and cost of DNA sequencing. They have become widespread after the year 2000 and are still in the lead position in the sequencing business. All NGS techniques involve sequencing millions of short DNA fragments which are subsequently stitched together. This presents one of the biggest issues with NGS technologies - the post-sequencing data analysis [10]. There are many NGS technologies, however, the key player that currently dominates the whole sequencing field is Illumina's sequencing by synthesis (SBS) and therefore it will be used as the primary example of NGS in this thesis.

2.3.1 Illumina Sequencing

The whole process of Illumina sequencing consists of four steps - library preparation, cluster generation, sequencing and finally data analysis. First, a DNA library is prepared by dividing the DNA sample into short fragments of approximately 200 - 600 bp and ligating sequencing adapters to both ends of the DNA fragments. These fragments are subsequently denatured to form two individual strands and loaded onto a flow cell where they are anchored using oligonucleotides (short synthetic DNA strands complementary to the ligated adapter).

At this stage, the fluorescent signal would be too weak to be detected, therefore the strands are directly on the flow cell copied to form bigger clusters of around 1000 identical single-stranded DNA fragments. At the end of this process, up to a billion clusters of single-stranded DNA fragments per flow cell are created, resulting in a much stronger sequencing signal.

Illumina sequencing

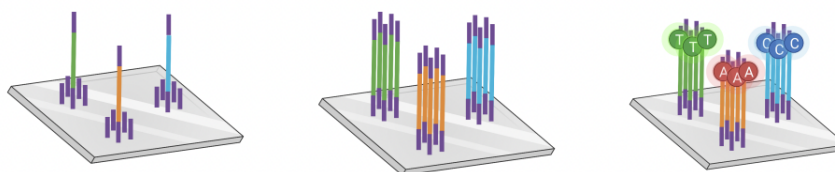


Figure 2.2: Cluster formation during Illumina sequencing [11]

The actual sequencing is based on sequencing by synthesis where a single-stranded DNA is replicated using modified nucleotides labelled with a fluorescent tag with attached reversible terminators. Once the correct base is incorporated, the sequencing process is stopped thanks to the terminators. An image is taken, capturing the fluorescent signal from the tag. The fluorescent tag and the terminator group are

then cleaved, making it possible to include another modified nucleotide. This cycle is repeated for the desired sequencing length (limit 300 bp).

The last step of Illumina sequencing is the data analysis, first, the actual DNA sequence is determined using the previously captured images (Figure 2.3). Each image contains the signal from several clusters and each cluster represents one DNA fragment. Millions of clusters and therefore millions of DNA fragments are being sequenced at the same time. Finally, the resulting short DNA reads are stitched back together to form the final sequence of the original DNA molecule. [9]

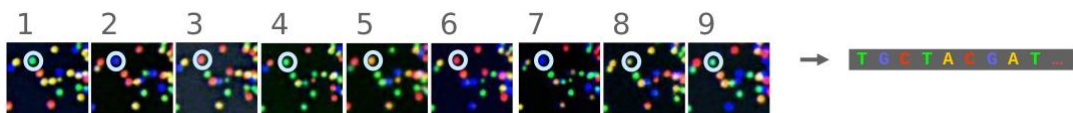


Figure 2.3: Determining the final sequence from images captured during Illumina sequencing [12]

2.4 Long-Read Sequencing

Long-read sequencing, also known as third-generation sequencing (TGS) emerged in the last decade and has since been developing rapidly. The main features that differentiate TGS technologies are real-time sequencing, single-molecule sequencing, and the ability to produce very long reads (more than 10 kilobases [kb] on average) [13, 14]. It also allows DNA sequencing without the need for PCR amplification in some use cases.

Currently, there are two leading players in the field of long-read sequencing. The first is Pacific Biotechnologies, whose 'single-molecule real-time' (SMRT) sequencing was presented in 2009 [15]. The second is Oxford Nanopore Technologies, the author of nanopore sequencing introduced in 2014 [13].

2.4.1 PacBio SMRT Sequencing

To effectively use the SMRT sequencing technique we first need to prepare the DNA library. The library specific to SMRT sequencing is called the SMRTbell library (Figure 2.4). It is created by attaching hairpin adapters to both ends of a double-stranded DNA sample, creating a closed loop. Primer and polymerase are then added to the library.

The prepared samples of the SMRTbell library are loaded onto a chip called an SMRT cell. Each SMRT cell contains up to 25 million wells called zero-mode waveguides (ZMWs). The nanometre structure of a ZMW reduces the volume of

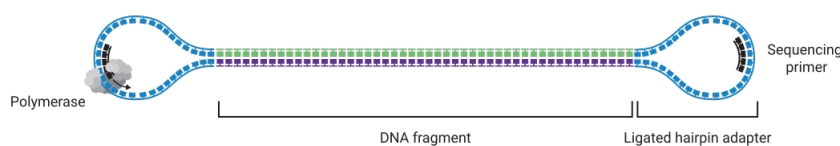


Figure 2.4: SMRTbell library [16]

observation significantly, this allows for the detection of a single fluorophore despite the high concentration of fluorescently labelled nucleotides.

A single DNA template is anchored at the bottom of the ZMW. At this point, nucleotides are introduced into the process. The four nucleotides (Adenine, Thymine, Cytosine, Guanine) are each labelled with a different coloured fluorophore (fluorescent dye) with a characteristic emission spectrum and therefore are easily distinguishable. They diffuse in and out of the ZMWs at a microsecond rate. When the polymerase encounters the correct base (nucleotide), it incorporates it in several milliseconds. During this time the fluorophore is excited, emitting a fluorescent signal that is recorded in real-time by sensitive cameras. DNA sequencing is assembled from these fluorescent signals. After the nucleotide is fully incorporated the fluorophore is severed and flows away from the detection zone of the ZMW. The polymerase is then able to incorporate another base. [15] The sequencing process is shown in Figure 2.5.

In an ideal case, each ZMW is loaded with one SMRTbell and produces one read. Realistically, about one-third to one-half of ZMWs will produce a viable read, resulting in approximately 365 000 reads from one run on the Sequel instrument [9].

PacBio currently offers several long-read sequencing systems - Sequel, Sequel II, Sequel IIe, and their newest addition the Revio [17].

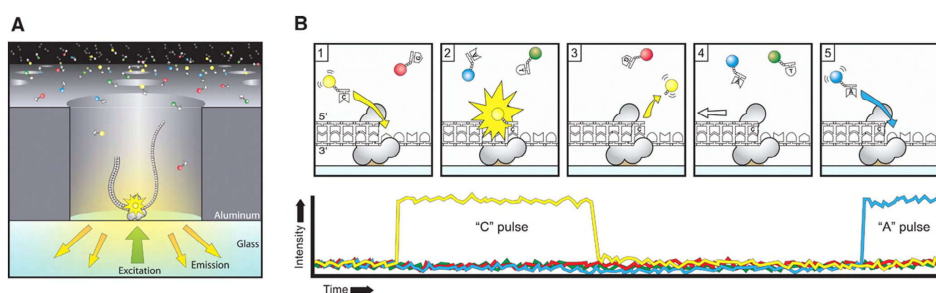


Figure 2.5: SMRT sequencing schema [15]

A: DNA template immobilised at the bottom of the ZMW **B:** Nucleotides labelled with different fluorescent tags are being incorporated, subsequently the fluorescent signal is recorded

2.4.2 Nanopore Sequencing

This technique does not utilize the principles of DNA replication, instead, a single-stranded DNA is passed through a perforation of nanometric size (nanopore). First, a DNA library is prepared, if a double-stranded DNA sample is to be sequenced, two adapters (leader and hairpin) are preloaded with motor proteins [18] and ligated to both its ends, if only a single strand is sequenced, the hairpin adapter is not necessary [19]. The library is then loaded onto a flow cell where 2048 nanopores are housed [20]. The nanopores are embedded in an electrically resistant membrane made out of synthetic polymers. Due to the high electrical resistance of the membrane, when a potential is applied across the membrane, it creates an ionic current that flows through the nanopore. The negatively charged DNA molecule is guided to the proximity of the nanopore by the leader adapter, subsequently, it is unzipped by the motor protein and driven through the nanopore from the side that is negatively charged to the positively charged side. The speed is controlled by the motor protein. The current is measured in real-time and when specific bases pass through the nanopore, it causes a distinct disruption in the current, making it possible to translate the measured current into a sequence of bases using computational algorithms. The above process is shown in Figure 2.6. Once the first (template) strand of DNA has completely passed through the nanopore, the second strand, connected by the hairpin adapter, can also be sequenced. This is called the 2D read. If only one strand of the DNA is sequenced, it is called a 1D read [21].

ONT's sequencing instruments include MinION, GridION, PromethION and Flongle [22].

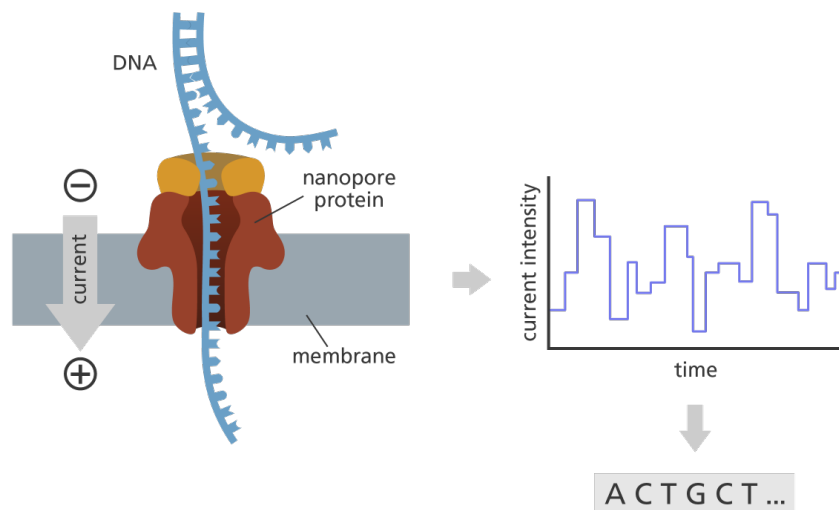


Figure 2.6: Nanopore sequencing schema [23]

Comparison of Sequencing Methods

3

In this chapter, three key criteria for sequencing methods are selected - accuracy, cost and read length. An overview of the performance of long-read sequencing methods in each category is provided and compared with Sanger sequencing and Illumina sequencing.

3.1 Read Length and Run Time

The obvious advantage of long-read sequencing is the drastic length improvement. Nanopore sequencing can routinely produce 150,000 bases long reads [24] and their longest ultra-long reads go as long as 4,000,000 bases [22], which is the current maximum of any sequencing technology. ONT's read length is limited mainly by the length of the input DNA, which indicates, that if provided with a long enough quality DNA template the sequencing lengths could reach new heights. PacBio's SMRT sequencing read length is directly tied to the polymerase used during sequencing. It generally achieves shorter read lengths than ONT, the newest sequencing system from PacBio - Revio claims to be able to standardly achieve read lengths up to 20 kb with the longer reads being over 60 kb [25]. Both Sanger sequencing and Illumina sequencing cannot compete with LRS technologies when it comes to read lengths, Sanger 3730xl reads about 400-900 bases and Illumina's instruments provide a read length of 2 x 150 bp - 2 x 300 bp. [26]

Hand in hand with read length comes the run time of the sequencing instruments. Nanopore sequencing is unique in the way that it has no fixed run-time, it can run as little as minutes up to 72 hours with a maximum speed of 420 bases per second resulting in the theoretical maximal output of 50 Gb per MinION's flow cell. PacBio's instruments have varied run times from approximately 1 hour (for <10 kb reads) [27] to 30 and 24 hours on the Sequel II and Revio respectively. The yield is 360 Gb of high fidelity (HiFi) reads per max run time for Revio and 30 Gb of HiFi reads for Sequel IIe [28]. Sanger 3730xl can produce up to 84kb in 3 hours [26], Illumina's smallest capacity instrument iSeq 100 runs 9.5–19 hrs providing up to 1.2 Gb and

3. Comparison of Sequencing Methods

their production scale NovaSeq X can run up to 48 hrs with the max output of 16 Tb [29].

Generation	Platform	System	Read length	Max output per run	Max run time [h]
Long-read	PacBio	Revio	15-20 kb	360 Gb	24
	PacBio	Sequel II	15-20 kb	30 Gb	30
	ONT	MinION	From short to ultra-long (>4Gb)	50 Gb	72
	ONT	PromethION	From short to ultra-long (>4Gb)	up to 14 Tb	72
NGS	Illumina	iSeq 100	2 x 300 bp	1.2 Gb	19
	Illumina	NovaSeq X	2 x 300 bp	16 Tb	48
Sanger	Applied Biosystems	Sanger 3730xl	400-900 bp	84 kb	3

Table 3.1: Read-length, output and run time of selected sequencing platforms

3.2 Sequencing Accuracy

3.2.1 Phred Score

To assess the accuracy of sequencing techniques a Phred quality score (Q score) is commonly used [30]. It indicates the probability of the base being recognised incorrectly and is defined as

$$Q = -10 \log_{10}(E).$$

Phred Quality Score [Q]	Error [E]	Accuracy (1 - Error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%

Table 3.2: Examples of Q values and corresponding error rates

3.2.2 Accuracy Comparison

The accuracy of DNA sequencing can vary depending on several factors such as the length of the reads, the quality of the input DNA, or the bioinformatics analysis

used to process the data. To this day, Sanger sequencing is still considered the gold standard of sequencing methods and is widely used when a highly accurate DNA sequence is needed. This technique operates with an error rate of less than 1 in 10,000 - Q40 [31]. Generally, Illumina sequencing can produce very accurate results as well with an error rate of approximately 0.1% to 1%. This means that for a typical Illumina sequencing run, 99.9% to 99% of the base calls will be correct [32].

PacBio's SMRT sequencing has in its raw form a relatively high error rate (13%) [13], however, the error rate can be reduced. Thanks to the circular nature of the SMRTbell library, each DNA template can be sequenced repeatedly. The errors during SMRT sequencing are mostly random, therefore when several reads of the same template are compared, those random errors can be eliminated to some degree. This process is called circular consensus sequence (CCS) and the final reads are called HiFi reads. The results they yield are comparable to Illumina sequencing - Q20 when the template is sequenced approximately 4 times and Q30 when 10 times. [33]

A similar thing can be applied to Nanopore sequencing. The error rate is significantly higher (15%) when it comes to raw reads, but data obtained from 2D reads achieve Q30 accuracy [34]. It is also necessary to note that the errors occurring during nanopore sequencing are, unlike those of PacBio's sequencing, biased. The occurrence of transitions is considerably higher than transversions, likely due to the fact that Adenin and Guanin as well as Cythosin and Thymin have similar shapes. Therefore, when they pass through the nanopore during sequencing, the current disruptions are also similar, making them more difficult to differentiate. [35]

Generation	Platform	Error Rate
Long-read	PacBio (SMRT)	Q30 - HiFi
	ONT	Q30 - 2D Q20 - 1D
NGS	Illumina	Q30
Sanger	Applied Biosystems	Q40

Table 3.3: Error rates of different sequencing platforms

3.3 Cost

The cost of DNA sequencing can be assessed based on a variety of criteria, this section will include the cost of instruments as well as the approximate cost of sequencing one billion bases (Gb), this estimation does not include the cost of the material. As it was not always possible to get pricing for the local market, prices are listed either in EUR or USD depending on availability.

3. Comparison of Sequencing Methods

ONT's MinION is the cheapest option with instruments' prices starting as low as \$1,000. The PromethION on the other hand is considerably more expensive with the starter pack valued at \$225,000, unlike MinION, it already includes a computing unit, making it more suitable for conventional uses in the laboratory. The price per Gb is lower with higher throughput putting minION at around €12 and PromethION at €8 [36].

PacBio systems are generally more expensive, the newest Revio sequencing instrument has a list price of \$779,000 [37], its price per Gb is comparable to ONT with €9. The Sequel II has lower throughput raising the price per Gb to €17 while the instrument itself costs €650,000.

The price of Illumina's sequencers varies from \$19,900 for iSeq 100 up to \$985,000 for NovaSeq X. The price per Gb is set around \$485 for the iSeq 100 down to \$2 with the newest NovaSeq series instruments, which is currently the lowest price of any sequencing platform. One of the biggest issues with Sanger sequencing has always been its cost, although the instrument price is similar, the \$5000 [38] price per Gb is incomparable. [39] [26]

Generation	Platform	System	Price per Gb	Instrument Cost
Long-read	PacBio (SMRT)	Revio	€9	\$779,000
		Sequel II	€17	€550,306
	ONT	MinION	€12	\$1,000
		PromethION	€8	\$225,000
NGS	Illumina	iSeq 100	€485	\$19,900
	Illumina	NovaSeq X	\$2	\$985,000
Sanger	Applied Biosciences	Sanger 3730xl	\$5000	\$95,000

Table 3.4: Instrument prices and prices per Gb for different sequencing platforms

Data Formats

4

4.1 SAM/BAM Format

Sequence Alignment/Map (SAM) format is a text format designed for efficient aligned sequenced data storage [40]. BAM (Binary Alignment/Map) format is the binary equivalent of SAM format. It consists of an optional header section and an alignment section.

4.1.1 Header Section

The header section contains metadata such as information about the reference sequence read information or comments. If present, it always has to be in front of the alignment section. Every header line has to start with the symbol '@' and is followed by a one or two-letter header record type that determines the type of metadata included. The header record type is then followed by a data field containing a tab-delimited tag-value pair (excluding the comment record type). Each pair follows the 'TAG:VALUE' format, with 'TAG' being a two-letter string containing the format and information of 'VALUE'. The order of data fields is irrelevant.

4.1.2 Alignment Section

Unlike the header section, lines in this section do not start with the '@' symbol. Each line of the alignment section traditionally represents a linear alignment of a segment and it consists of 11 mandatory tab-separated fields (Figure 4.1). If a value of mandatory field is not available, the field still has to be included, with the value containing either '0' or '*' depending on the field type. After the mandatory fields, a variable number of optional fields can be added.

Col	Field	Type	Regex/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0, 2 ¹⁶ - 1]	bitwise FLAG
3	RNAME	String	* [:rname:^*]=[:rname:]*	Reference sequence NAME ¹¹
4	POS	Int	[0, 2 ³¹ - 1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0, 2 ⁸ - 1]	MAPping Quality
6	CIGAR	String	* ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	* [:rname:^*]=[:rname:]*	Reference name of the mate/next read
8	PNEXT	Int	[0, 2 ³¹ - 1]	Position of the mate/next read
9	TLEN	Int	[-2 ³¹ + 1, 2 ³¹ - 1]	observed Template LENgth
10	SEQ	String	* [A-Za-z=.]+	segment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Figure 4.1: List of mandatory fields in the alignment section

4.2 HDF

The HDF format, or Hierarchical Data Format, is a file format designed for storing and managing large datasets. It is structured as a hierarchy of objects, where each object can contain data or other objects. The most basic object is dataset, which is a multi-dimensional array of data. Datasets can be organized into groups. Attributes can also be attached to datasets and groups to provide additional metadata. An example of HDF format structure is in Figure 4.2.

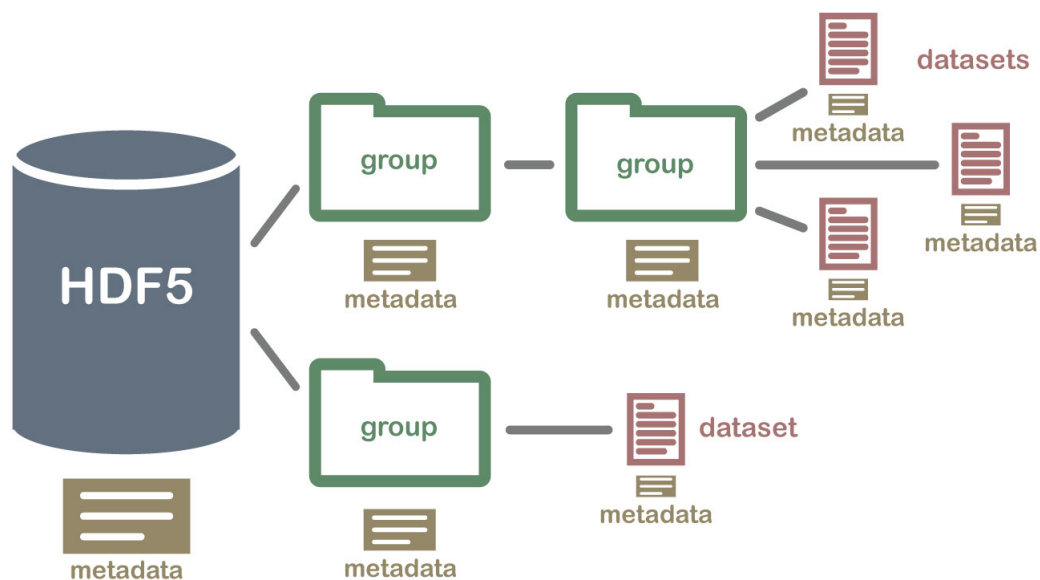


Figure 4.2: HDF5 structure [41]

The format provides a set of libraries and tools that allow for efficient access to the data in the file. Additionally, the HDF Group, the organization responsible

for maintaining the HDF format, provides a number of software tools for working with HDF files, including HDFView, a graphical viewer for HDF files, and h5py, a Python library for working with HDF files.

One of the key advantages of the HDF format is its ability to handle large and complex data sets. HDF files can store data in compressed formats, which can significantly reduce file size and improve performance. Additionally, the hierarchical structure of the format makes it easy to organize and manage datasets.

4.3 FASTA/FASTQ

4.3.1 FASTA

FASTA is a standard format in the sequencing industry. It represents each nucleotide sequence in two lines. The first (header) line starts with the symbol “>” and is followed by the sequence identifier and optionally some other description. The second line contains only the sequence in the form of single letters each representing one nucleic acid. A file can contain more than one nucleotide sequence. It is a text-based format and therefore easily readable by humans. An example of a FASTA file can be seen in Figure 4.3.

```

Header ● >VIT_201s0011g03530.1
Sequence ● AATTAAGCATAAAATACTCACTCTTACCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
● GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
Header ● >VIT_201s0011g03540.1
Sequence ● CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCACAAAACCC
● AGCCTCTGAGACACCACCTCAAACCTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
Header ● >VIT_201s0011g03550.1
Sequence ● CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
● GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA

```

Figure 4.3: An example of FASTA file with multiple sequences [42]

4.3.2 FASTQ

FASTQ is a data format used predominantly by NGS technologies. It is similar to FASTA, however in addition to the actual sequence it also contains the respective Phred scores. Each sequence consists of 4 lines:

- **Sequence identifier** (starts with the symbol '@');
- **The Sequence;**
- **'+' sign;**

- **Quality scores.**

An example of a FASTQ file can be seen in Figure 4.4.

```
Identifier  @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence   TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign   +
Quality scores hhhhhhhhhghghghhhhhfhhhhffhfff'ee['X]b[d[ed'[Y[^Y
Identifier  @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence   GATTTGTATGAAAGTATACAACATAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign   +
Quality scores hhhhghhcgghghggfcffdhfehhhhcehdchhdhahehfffde'bVd
```

Figure 4.4: An example of FASTQ file with multiple sequences [42]

4.4 PacBio Data Format

PacBio is currently using the standard BAM format for both aligned and unaligned data with some minor changes. PacBio includes several extra header lines containing information such as the 'platform model' which includes the instrument series (e.g. Revio, Sequel) or 'Bio Sample Name' containing information about the biological samples. The full extent of PacBio's modifications to the BAM format can be found at <https://pachiofileformats.readthedocs.io/en/12.0/BAM.html>.

4.5 Nanopore Data Formats

4.5.1 FAST5

Sequencers from ONT provide raw data in the FAST5 format based on hierarchical data format version 5 (HDF5) which is suitable to store a large amount of data. A FAST5 file works similarly to a file system, however, unlike the original HDF format it has a defined schema. A typical FAST5 file usually contains the raw signal data as well as some other metadata.

ONT historically defines two types of FAST5 files: single-read and multi-read. The single-read file contains only one read whereas the multi-read file contains multiple reads, however, the single-read files are no longer actively in use and is therefore recommended to convert them into multi-read files. To work with FAST5 files a standard HDF library and tools can be used. ONT developed a Python library for working with FAST5 - `Ont_fast_api`. Within this library, a single-read to multi-read file converter is included.

4.5.2 POD5

ONT has recently introduced a new file format called POD5. The data in POD5 consists of three Apache Arrow tables combined into one container format. The 3 Apache Arrow tables are:

- **Run** - Contains experiment-level information;
- **Reads** - Contains metadata for individual reads;
- **Signal** - Contains the raw signals from reads.

Together with the new data format, ONT has also released a C++ library, C interface, and Python module to work with POD5. As well as a tool to convert FAST5 format to POD5 available at <https://pod5.nanoporetech.com/>.

Long-Read Sequencing Pipeline and Available Tools

5

Long-read sequencing offers a wide range of applications, however, the analysis of LRS data presents unique challenges due to the higher error rates and the increased complexity of the data. To address these challenges, a variety of new bioinformatics approaches, software tools and pipelines have been developed to support the processing and analysis of LRS data. This chapter contains a description of the common steps in long-read data analysis (Figure 5.1) including the available tools.

5.1 Common Steps in Long-Read Data Analysis



Figure 5.1: A pipeline illustrating common steps in LRS data analysis

5.1.1 Basecalling

Basecalling is a process of converting raw data from a sequencing instrument into a sequence of nucleotides. It is the first step in the analysis of long-read sequencing data. It is an ever-evolving field. In the past decades, basecalling tools were largely based on hidden Markov models, this has changed in recent years, with a shift towards neural networks, which means that a sufficient dataset is needed to train the model. Moreover, if there are anomalies such as base modifications present in the sequenced DNA, the training data have to contain them as well in order for the basecaller to identify them properly. [43]

5.1.1.1 PacBio

The basecalling process developed directly by PacBio is called Circular Consensus Sequence. In each ZMW the DNA template is sequenced multiple times, generating multiple 'passes' of the same sequence. During each pass, a 'subread' is recorded. The consensus sequence is then generated by aligning and comparing the subreads to each other to identify errors and overlapping regions (Figure 5.2). The CSS quality is limited by the number of subreads generated, generally, the more subreads the more accurate the final sequence. This is, however, restricted by the longevity of polymerase, a longer DNA template will result in fewer possible subreads. At a minimum, two full subreads are needed for CSS. Each consensus sequence represents DNA template from a single ZMW. The current research suggests that for a 13.5 kb library, CCS achieves Q30 accuracy, although it is important to note that the average computing time required was 3,035 CPU core hours per SMRT cell [44].

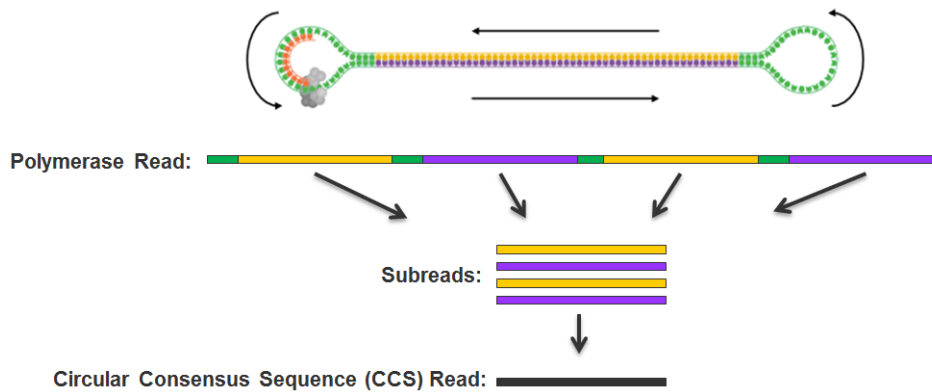


Figure 5.2: Circular consensus sequence [45]

The algorithm for CSS still utilises a hidden Markov model. In 2022 however, PacBio formed a partnership with Google Health and developed DeepConsensus, a deep learning based approach that works on top of CSS and further reduces the error rate in HiFi reads by 41.9% [46].

Even though the raw signal data from SMRT sequencing are available, there are no third-party basecallers designed to work directly from the signal data.

5.1.1.2 ONT

As described in chapter 3, ONTs raw reads have a significant amount of errors that are bias prone, therefore, the basecalling process can be more challenging than PacBio's. There is a high number of third-party software available for basecalling ONT's data and ONT itself has developed an amount of basecalling tools such as Guppy and Bonito.

Guppy

The first to be introduced is the Guppy basecaller. It is integrated into the MinKNOW (primary software on all ONT's sequencing devices). Guppy is based on recurrent neural networks and offers three variants of basecalling models - a Fast model, a High accuracy model, and a Super accurate model. A unique feature of the Fast model is that it was specifically designed to keep up with the sequencing speed (at the cost of lower accuracy). This means the whole process of sequencing and basecalling is done virtually in real-time. The HAC and SUP models are more accurate, however, they are also more computationally intensive (3 and 24 times respectively compared to the Fast model). On a state-of-the-art GPU (NVIDIA Tesla P40 24GB) the HAC model takes approximately 14 hours per flow cell for 11-15 kb library [47].

Bonito

Bonito is the latest PyTorch based basecaller developed by ONT. It is based on convolutional neural networks, specifically the QuartzNet model originally designed for speech recognition [48]. Compared to Guppy, it has increased the basecalling accuracy by 1%. One disadvantage of Bonito is the relatively slow speed making it difficult to use in practice [49].

Third-party basecallers

Of the many existing third-party basecallers, only a handful of them is being regularly updated and it is unclear to which extent they reflect the latest progress in long-read sequencing. The examples listed below are some of the most current basecallers.

Fast-Bonito is a basecaller based on ONT's Bonito, which can be several times faster than Bonito, depending on the hardware used, tackling the main issue of the Bonito basecaller [49].

Another notable basecaller is **DeepNano-Coral** which focuses on real-time basecalling. It is optimised to run using Coral Edge Tensor Processing Unit - a small energy-efficient accelerator attached by USB and surpasses the basecalling quality of Guppy's Fast model [50].

5.1.2 Quality Control

Once the reads have been basecalled, the next step is typically performing quality control. This includes providing an overview of statistical information about the dataset qualities such as read length distribution, accuracy, and GC contents analysis. Quality control tools usually only provide a general summary of the dataset

and are intended to help the user understand the data and find potential issues. It is up to the user to determine how to use this information. Quality control is particularly important in long-read sequencing, where individual reads can be highly variable in quality, or even ‘non-sensical’ - it has been reported that LRS runs can contain a number of reads that have no connection to any other molecule within the sequenced library [51].

There are several pipelines tailored specifically for LRS quality control, a popular example would be **LongQC** (Long Read Quality Control) [51], a pipeline that focuses on the comprehensive evaluation of both ONT’s and PacBio’s data or **MinIONQC** [52] designed to provide a rapid diagnostic of data from MinION. An established quality control tool **FastQC** [53] has also been partially adapted for long-read data.

5.1.3 Read Alignment

In the vast majority of applications, mapping reads to the reference genome is a necessary step. Read alignment (or read mapping) poses an important challenge since the quality of the alignment can severely impact the outcome of any further analyses.

The typical aligning algorithm follows three steps. The first is generating an index of the reference genome. This ensures rapid localization of any subsequences in the reference genome. Followed by determining the potential positions of all the reads. Lastly, each read is thoroughly compared to its potential position to determine the final location of the read as well as any deviations from the reference genome. The second step tends to be difficult for NGS data, because of their small read lengths many possible locations are identified. LRS has an advantage - due to its long reads the position determination is relatively unambiguous. However, the last step may be problematic for LRS, because it involves focusing on smaller details which is challenging considering the higher error rate. [54]

Although the general steps of the aligning algorithms are the same, they need to be tailored to the data they are working with. Some NGS aligners have been adapted to LRS, but their performance is subpar specifically when it comes to computing time, therefore new LRS-specific approaches had to be developed. Currently, the leading state-of-the-art aligner used for both ONT’s and PacBio’s data is **minimap2** [55]. A comparative analysis of aligners showed that when not limited by computational costs, it is desirable to use a combination of alignment tools to get the best results, however, when the number of tools used is limited to one, **minimap2** consistently outperforms other LRS aligners [56]. Other popular LRS aligners include **BLASR** [57] or **NGMLR** [58].

5.1.4 Error Correction

Even though the accuracy of long-read sequencing methods is rising, error correction is still a crucial step in many long-read applications. There are currently two distinct approaches to error correction - hybrid and non-hybrid.

Hybrid Methods

Hybrid methods capitalize on the high accuracy of short-read data in order to perform error correction on longer reads. The important prerequisite for using these methods is the availability of the short-read data [59]. There is a wide variety of available hybrid tools such as **Hercules** [60], **Jabba** [61], and **HALC** [62].

Non-hybrid Methods

Non-hybrid methods of error correction utilise only long-read data eliminating the need for accurate short-read data. Among the most commonly used tools in this category are the **FLAS** [63], **LoRMA** [64] and the error correction mode included in **Canu** (a comprehensive pipeline for long-read data analysis) [65]. All of them are compatible with both PacBio and ONT data.

Although generally, the performance of hybrid methods is currently still superior in both correction performance and computing costs to non-hybrid methods, there are instances where non-hybrid methods are preferable. Simply put, non-hybrid methods lack the additional accuracy of short-read data, therefore, they usually need additional sequencing depth to achieve the same accuracy. A comprehensive study demonstrated that when the error rate is higher, at around 18%, hybrid correction is more effective at low sequencing depths. However, as the sequencing depth increases, non-hybrid correction becomes more efficient. When error rates are lower, at 12% or less, non-hybrid correction outperforms hybrid correction even with low-coverage datasets. Additionally, it is necessary to take into consideration the complexity of the organisms being sequenced, as it has a significant impact on the quality of correction. Specifically, the quality of correction of hybrid methods significantly decreases when dealing with complex and repeat-rich organisms. [66]

5.1.5 Other Tools

It is important to mention that the number of available tools tailored to long-read data analysis is growing extremely fast. It is not within the scope of this thesis to mention every tool available. To help the TGS community effectively navigate the available software, an open-source database that aims to collect and catalogue the available tools has been developed [67]. As of today, the database is available at <https://long-read-tools.org/> and is still being regularly updated. A recently published

5. Long-Read Sequencing Pipeline and Available Tools

paper focused specifically on Nanopore sequencing also provides a summary of the available tools [68].

Applications of Long-Read Sequencing Technologies

6

This chapter will explore some of the key applications of third-generation sequencing technologies as well as provide the reader with a choice of tools available for each application. In the included applications, LRS has either outperformed other sequencing methods or demonstrated significant benefits.

6.1 *De Novo* Assembly

De novo assembly is a process of reconstructing an organism's genome from scratch, without relying on a reference genome. In other words, it involves piecing together DNA fragments obtained during sequencing to create a complete genome sequence. Although NGS technologies have advanced the task of *de novo* assembly greatly, there are still several key aspects that can be improved by using LRS. Assembling repetitive regions is one of those as well as lowering the computational cost - it is especially demanding to assemble data generated through NGS due to the small lengths of sequenced fragments [69].

Long-read sequencing is showing great potential in the field of *de novo* genome assembly, for the first time in history we were able to fill the last missing pieces and fully complete the sequence of the human genome [70]. It has also been demonstrated that in ideal conditions it is possible to assemble a whole human genome in less than 10 minutes which is a significant achievement when considering that the same task required over 100,000 CPU hours just a few years ago [71].

Several studies [72, 73, 74] comparing the performance of assemblers on PacBio's or ONT's data showed that when aiming for accurate assembly and not limited by resources, it is advised to use a combination of some of the more robust assemblers (e.g. **Canu** [65] + **SMARTdenovo** [75]). When using only one assembler, the best-

performing assemblers in terms of accuracy include **Canu**, **SMARTdenovo**, and **Flye** [76]. The main drawback of these tools is their computing time, requiring thousands of CPU hours to assemble a human genome in good quality. One of the both fast and accurate assemblers is **Raven** [72].

Recent studies suggest [77, 78] that especially for human genome *de novo* assembly the best standalone method is PacBio's HiFi reads. Due to their improved accuracy, the assemblers require lower amounts of data thus lowering the computing cost while retaining high accuracy. Assemblers specifically for HiFi reads are **HiCanu** [79] and **hifiasm** [80].

6.2 Variant Calling

Variant calling is the process of identifying changes in a genome. In the human genome, around 99.9% of the DNA sequence is identical for every member of our species. A number of these genomic differences can have a massive impact on a variety of things from physical differences to predisposition to many diseases such as diabetes or cancer [81]. The most commonly observed change in the human genome is a single nucleotide variant (SNV) - a substitute of one single base [82]. Other shorter variants up to 50bp are called short insertion-deletion variations (indels). Changes longer than 50bp are commonly referred to as structural variants (SVs) [83]. There is a much lower number of SVs and indels than SNVs, however, their consequences are generally more prominent thanks to their size [84].

In the past decade, variant calling using NGS technologies helped uncover the effects of a large number SNVs and indels. On the other hand, SVs remained largely understudied, mainly because they are notoriously challenging to detect using NGS technologies for two main reasons. First, SVs are frequently present in regions that are difficult to sequence using NGS (repetitive regions) and second, the read length of NGS is often too short to discover some of the longer, more complex SVs. Due to their long read length and ability to sequence repetitive regions long-read sequencing technologies are being used to detect large, complex SVs with increasing frequency [85]. Both ONT's and PacBio's techniques have proven to be highly effective in detecting SVs, for example, in 2019 a study sequenced 15 human genomes using SMRT sequencing and found almost 100,000 structural variants, of which a large portion was previously unknown [86]. Using SMRT sequencing, an undiscovered 12.4 kb SV was detected in progressive myoclonic epilepsy [87]. Nanopore sequencing helped discover a 7.1 kb SV in Mendelian disease [88].

Some of the popular variant callers used for SVs are **Pbsv** [89], **Sniffles2** [58] and **cuteSV** [90].

LRS technologies have been used to detect SNVs and indels as well, however, the accuracy is very much dependent on the error rate of the sequenced data especially

when the errors are not randomly distributed, which is the case for ONT sequencing. This can be partially resolved by increasing the sequencing coverage, which would also mean increasing the computing costs [19]. A number of variant callers tailored to detect SNVs have emerged, some of them, such as Google Health's **DeepVariant** [91], ONT's **Medaka** [92] or **NanoCaller** [93] are based on deep learning which makes them adaptable to unique error profiles of LRS technologies [44].

In order to efficiently detect both SNVs and SVs some studies propose using hybrid methods for variant calling combining data from both LRS and NGS technologies. A hybrid variant caller **HELLO** was able to reduce the error rate when calling indels by up to 30% when compared with a state-of-the-art variant calling tool that used only short reads [94].

6.3 Epigenetics

Modifications in both DNA and RNA are a crucial part of various biological processes from ageing to serious diseases including cancer or Alzheimer's and Parkinson's disease [95, 96]. There are several causes of DNA modifications, for example, radiation or oxidation damage, however, the most extensively investigated is methylation. The specific modification that occurs the most in both plants and animals is 5-methylcytosine (5mC), sometimes even dubbed the fifth base. It is also the most studied modification, likely because of the availability of accurate sequencing techniques. The most widespread technique is short-read bisulfate sequencing which involves first treating the DNA sample with sodium bisulfate. Using bisulfate, however, leads to the degradation of the DNA as well as more complicated PCR amplification resulting in the need for large quantities of input DNA. Moreover, certain modifications are likely to be found in repetitive regions, such as 5mC in CG regions, which are challenging to read for NGS technologies. [97]

Both PacBio and Nanopore sequencing can, to a certain extent, detect base modifications. Nanopore sequencing determines the nucleotide base based on characteristic changes of current when the base passes through the nanopore. When a modified base goes through the pore, the current change is unique as well, therefore it is able to distinguish them (Figure 6.1). There are two distinct approaches when detecting modified bases from long-read data, direct and non-direct. When using the direct approach, modified bases are called from raw signal data using a basecaller with an extended alphabet. ONT's Guppy is able to detect m5C and N6-methyladenosine (m6A) modifications directly, this, unfortunately, comes hand in hand with decreased overall basecalling accuracy. [33]

The non-direct approach involves first performing a standard basecalling with the four canonical bases and then aligning the basecalled sequence to either the raw signal or reference sequence and using various statistical methods to call the

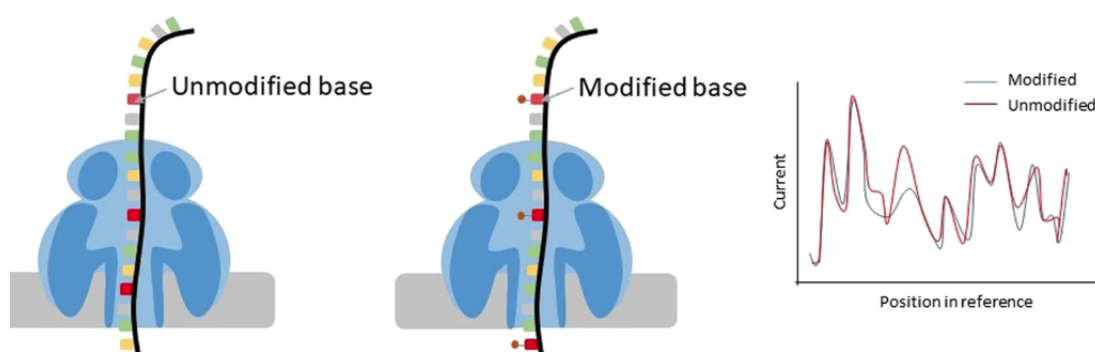


Figure 6.1: Difference in current between modified and unmodified base during ONT sequencing [98]

likelihood of modification. ONT currently employs two tools in this category: **Megalodon** and the more recent **Ramora**. Ramora performs the modification analysis in parallel to the standard basecalling by Guppy and is integrated with the MinKNOW software. To detect modified bases, a variety of third-party tools have been developed such as **NanoMod** [99], **Nanopolish** [100] and **DeepMod** [101].

SMRT sequencing has also been used to detect modified bases. It builds on the idea that base modifications affect the kinetics of the polymerase during sequencing, effectively creating a difference in time between normal and modified bases, when incorporating nucleotides, this is depicted in Figure 6.2 [102]. Some modifications have more prominent kinetics profiles than others. For example, 6-methyladenine and 4-methylcytosine (6mA, 4mC) both have strong kinetic profiles and PacBio recommends 25x coverage per strand. On the other hand, 5mC's kinetic profile is more subtle, thus increasing the recommended coverage tenfold (250x coverage per strand) which leads to a significantly shorter read length (< 2000 bp) [103]. Modification analysis can be performed directly in SMRT Link (a native tool made by PacBio) and with the introduction of HiFi reads PacBio's instruments can simultaneously perform standard basecalling and basecalling with extended alphabet detecting 5mC.

There are similar limitations for both ONT and PacBio. In order to identify a certain base modification the model must be prepared for this particular type by training it with relevant data which are often in short supply. Because the amplification of DNA would result in the loss of base modifications, these methods require a large amount of native, unamplified DNA [105].

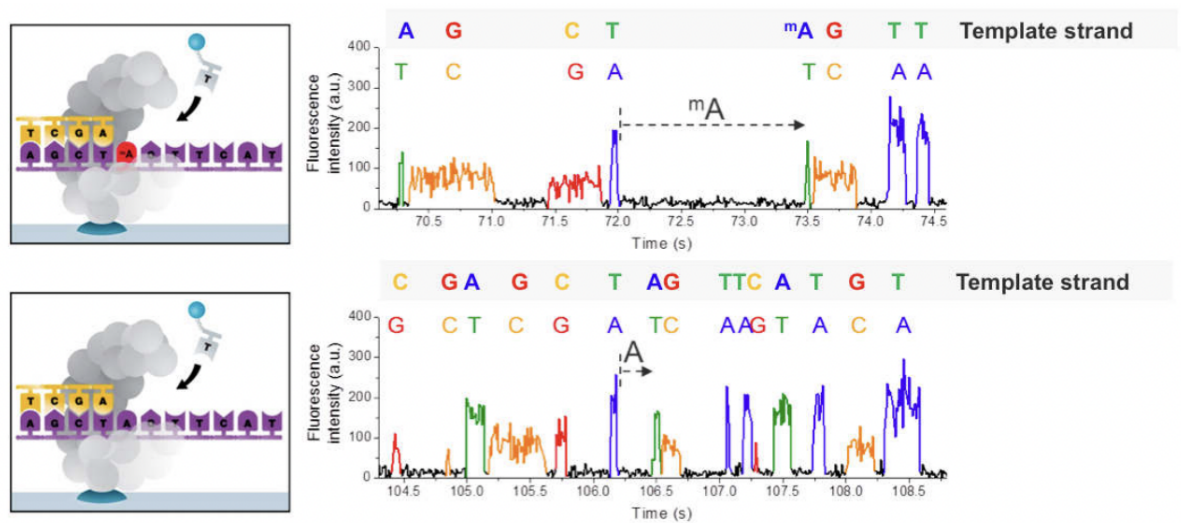


Figure 6.2: Difference in kinetics between modified and unmodified base during PacBio sequencing [104]

6.4 Direct RNA Sequencing

Even though this thesis mainly talks about DNA, it is important not to overlook one long-read sequencing application - the ability to directly sequence RNA. Due to the nature of ONT's sequencing method, it is possible to sequence RNA directly. This has been demonstrated several times, one of them being the direct sequencing of the Influenza A virus genome in 2018 [106]. Although it is possible to use PacBio's instruments to sequence RNA, it cannot be considered direct sequencing because the sequenced RNA molecule needs to be converted to complementary DNA (cDNA) and PCR amplified first [107].

6.5 Field Laboratory

A unique trait of ONT's MinION is its small size. The original MinION device itself weighs 87g and to be functional, it only needs to be connected to a laptop via USB. ONT has also released self-contained MinION versions that require no other computational resources and can perform some tasks in offline mode (basecalling). This option is especially suitable for working in the field [108].

In 2015, MinION instruments were used in Guinea for real-time genomic surveillance of the Ebola epidemic. Yielding results in less than 24 hours after receiving a positive DNA sample, and showing that it is possible to provide real-time genomic monitoring of emerging outbreaks of infectious diseases even with limited resources

[109]. Similarly, MinION was used to monitor the dengue virus in Brazil in 2019 [110].

6.6 Applications of Long-Read Sequencing in Cancer Genomics

Despite the considerable advantages, LRS is not yet that widely used in clinical settings. This section describes two clinical applications of long-read sequencing that use the principles described so far in this chapter.

6.6.1 Detecting Genetic Aberrations in Human Cancer

Human cancer is largely caused by various genome aberrations. Next-generation sequencing is particularly well suited to identify short indels and SNVs, however, longer genomic aberrations and the ones in repetitive regions are difficult to identify for NGS. The ability of LRS to detect large structural variants poses a great advantage in this field. Nanopore sequencing was recently used to identify new SVs in colorectal cancer [111] and to analyse lung cancer genomes [112]. Both PacBio and Nanopore sequencing demonstrated improved performance in identifying SVs in breast cancer genomes [113]. Despite the success, the LRS error rate remains too high for it to be used alone. This shows that ideally NGS should be used in tandem with LRS in order to achieve the best results in clinical cancer care [114].

6.6.2 HLA Typing

HLA (Human Leukocyte Antigen) typing is the process of matching a patient to its potential donor. This is used primarily in hematopoietic stem cells or solid-organ transplantations and can have significant consequences if performed insufficiently. Currently, HLA typing is done using short-read technologies, which can be problematic and lead to ambiguous HLA typing due to their read lengths [115]. Several studies have shown that LRS can detect previously unknown polymorphisms and provide results comparable to or even surpassing those of NGS [116, 117]. Nanopore sequencing is specifically used to develop new, much faster, HLA typing methods [118, 119]. Although the use of LRS for HLA typing is in its beginning stages, it is clear that it can bring significant improvements.

Besides HLA typing, the human genome contains other types of specific genes that can be studied in a similar fashion. One such example is Killer cell immunoglobulin-like receptors [120].

Conclusion

7

First, this thesis introduced the three generations of sequencing techniques, with a particular focus on the two most prominent LRS techniques made by Oxford Nanopore Technologies and Pacific Biosciences. A comparison of the techniques in three categories was provided and showed that LRS are much superior when it comes to read length, especially ONT with their ultra-long reads (>4Mb). Although the error rate is higher, the trends of recent years suggest that it will continue to decrease, especially since PacBio's CCS has achieved an error rate comparable to Illumina sequencing. The price per Gb of sequenced bases remains somewhat higher, but reduction can be expected, primarily for higher throughput instruments such as PacBio's Revio or ONT's PromethION. The cost of the instruments is comparable to NGS with the exception of the MinION, which with its price of \$1,000 opens the door to a new era of low-cost DNA sequencing.

The next chapter included information about the various data formats LRS currently uses. It is important to note that the data formats used by both PacBio and ONT are being updated or changed somewhat regularly, however, the vast majority of tools for subsequent data analysis operate with FASTQ or BAM input data which has been the industry standard since NGS have taken over the sequencing field.

The thesis then moved on to describe the common steps in long-read data analysis including Basecalling, Quality control, Read alignment and Error correction. For all aforementioned steps, a variety of available tools was provided as well as some evaluation of their performance. It is expected that as long-read sequencing becomes more widespread, the number of analytical tools will rise.

Finally, the most promising applications for LRS were described. Starting with *de novo* assembly, where the long reads of LRS have already proven to be useful. Especially PacBio's HiFi reads have the potential to be the best standalone sequencing method for *de novo* assembly. Moving on to variant calling, where LRS demonstrated its ability to detect large structural variants, some of which were previously undetectable using NGS. Another promising application is modified base detection, unlike NGS, long-read sequencing technologies are able to directly detect various base modifications without any initial DNA treatment. Moreover, LRS technologies

7. Conclusion

are great at detecting modifications in repetitive regions. The two last-mentioned applications are direct RNA sequencing and laboratory in the field which are unique to ONT.

Special attention was given to utilization in cancer genomics where, although still understudied, LRS has proven to be promising. Two specific applications were mentioned - HLA Typing and the detection of structural variations in human cancer.

Long-read sequencing has undergone considerable advancement in recent years. The technique has evolved to the point where it can now be considered for routine deployment in clinical practice. The long reads and fast run times will launch the era of personal genomics, specifically tailored care to individual patients. Although it still has its limitations, it is likely that with the development of new bioinformatical approaches, expansion of the available LRS datasets as well as advancements in sequencing chemistry, the limitations will be at least partially overcome.

List of Terms and Abbreviations

A, C, G, T	Adenine, Cytosine, Guanine and Thymine
Allele	One of two or more versions of DNA sequence occurring at a given location (locus)
bp	Base Pair
CCS	Circular Consensus Sequence
ddNTP	Dideoxynucleotides - ddATP, ddCTP, ddGTP, ddTTP
DNA	Deoxyribonucleic Acid
Gb	Giga Base - One Billion Bases
Hairpin adapter	A short nucleotide sequence that binds to one end of both strands of DNA, effectively joining them together and allowing the two strands to be sequenced consecutively
HiFi	High Fidelity
HLA	Human Leukocyte Antigen
indel	Insertion-Deletion Variant
kb	Kilo Base - One Thousand Bases
LRS	Long-Read Sequencing
Mb	Mega Base - One Million Bases
NGS	Next-Generation Sequencing
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCR	Polymerase Chain Reaction

List of Terms and Abbreviations

RNA	Ribonucleic Acid
SBS	Sequencing by Synthesis
SNV	Single Nucleotide Variant
SV	Structural Variant
TGS	Third-Generation Sequencing
Transversion	The interchange of a purine (A, G) base for a pyrimidine (C, T) base, or vice versa
Transitions	The interchange of a purine (A, G) base for another purine base or pyrimidine (C, T) base for another pyrimidine base
ZMW	Zero-Mode Waveguide

List of Figures

2.1	DNA replication schema [6]	6
2.2	Cluster formation during Illumina sequencing [11]	7
2.3	Determining the final sequence from images captured during Illumina sequencing [12]	8
2.4	SMRTbell library [16]	9
2.5	SMRT sequencing schema [15]	9
2.6	Nanopore sequencing schema [23]	10
4.1	List of mandatory fields in the alignment section	16
4.2	HDF5 structure [41]	16
4.3	An example of FASTA file with multiple sequences [42]	17
4.4	An example of FASTQ file with multiple sequences [42]	18
5.1	A pipeline illustrating common steps in LRS data analysis	21
5.2	Circular consensus sequence [45]	22
6.1	Difference in current between modified and unmodified base during ONT sequencing [98]	30
6.2	Difference in kinetics between modified and unmodified base during PacBio sequencing [104]	31

List of Tables

3.1	Read-length, output and run time of selected sequencing platforms . . .	12
3.2	Examples of Q values and corresponding error rates	12
3.3	Error rates of different sequencing platforms	13
3.4	Instrument prices and prices per Gb for different sequencing platforms	14

Bibliography

1. HEATHER, James M; CHAIN, Benjamin. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016, vol. 107, no. 1, pp. 1–8.
2. INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. Finishing the euchromatic sequence of the human genome. *Nature*. 2004, vol. 431, no. 7011, pp. 931–945.
3. GOODWIN, Sara; MCPHERSON, John D; MCCOMBIE, W Richard. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 2016, vol. 17, no. 6, pp. 333–351.
4. Method of the Year 2022: long-read sequencing. *Nat. Methods*. 2023, vol. 20, no. 1, p. 1.
5. MORGAN, David Owen. *The Cell Cycle: Principles of Control*. New Science Press, 2007.
6. MADPRIME. *DNA replication schema*. 2023-04-16. Available also from: https://commons.wikimedia.org/wiki/File:DNA_replication_split.svg.
7. SHENDURE, Jay et al. DNA sequencing at 40: past, present and future. *Nature*. 2017, vol. 550, no. 7676, pp. 345–353.
8. FRANÇA, Lilian T C; CARRILHO, Emanuel; KIST, Tarso B L. A review of DNA sequencing techniques. *Q. Rev. Biophys.* 2002, vol. 35, no. 2, pp. 169–200.
9. SLATKO, Barton E; GARDNER, Andrew F; AUSUBEL, Frederick M. Overview of Next-Generation Sequencing Technologies. *Curr. Protoc. Mol. Biol.* 2018, vol. 122, no. 1, e59.
10. HU, Taishan; CHITNIS, Nilesh; MONOS, Dimitri; DINH, Anh. Next-generation sequencing technologies: An overview. *Hum. Immunol.* 2021, vol. 82, no. 11, pp. 801–811.
11. KROON, Emma. *CRISPR Screen to Identify Genes Regulating Melanoma Cell Invasiveness*. 2023-05-28. Available also from: <https://www.diva-portal.org/smash/get/diva2:1738587/FULLTEXT01.pdf>.

12. UNIVERSITY OF EXETER. *Part 1. Short Read Genomics: Introduction*. 2023-05-14. Available also from: <https://biomedicalhub.github.io/genomics/01-part1-introduction.html>.
13. DIJK, Erwin L van; JASZCZYSZYN, Yan; NAQUIN, Delphine; THERMES, Claude. The Third Revolution in Sequencing Technology. *Trends Genet.* 2018, vol. 34, no. 9, pp. 666–681.
14. MOHAMMADI, Mohammad M; BAVI, Omid. DNA sequencing: an overview of solid-state and biological nanopore-based methods. *Biophys. Rev.* 2022, vol. 14, no. 1, pp. 99–110.
15. EID, John et al. Real-time DNA sequencing from single polymerase molecules. *Science.* 2009, vol. 323, no. 5910, pp. 133–138.
16. PACIFIC BIOSCIENCES. *Sequencing 101: from DNA to discovery — the steps of SMRT sequencing*. 2023-04-16. Available also from: <https://www.pacb.com/blog/steps-of-smrt-sequencing/>.
17. PACIFIC BIOSCIENCES. *Sequencing Systems*. 2023-01-13. Available also from: <https://www.pacb.com/sequencing-systems/>.
18. IP, Camilla L C et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Res.* 2015, vol. 4, p. 1075.
19. MAGI, Alberto; SEMERARO, Roberto; MINGRINO, Alessandra; GIUSTI, Betti; D'AURIZIO, Romina. Nanopore sequencing data analysis: state of the art, applications and challenges. *Brief. Bioinform.* 2018, vol. 19, no. 6, pp. 1256–1272.
20. JAIN, Miten; OLSEN, Hugh E; PATEN, Benedict; AKESON, Mark. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 2016, vol. 17, no. 1, p. 239.
21. WANG, Yunhao; ZHAO, Yue; BOLLAS, Audrey; WANG, Yuru; AU, Kin Fai. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 2021, vol. 39, no. 11, pp. 1348–1365.
22. OXFORD NANOPORE TECHNOLOGIES. *Product specifications*. 2023-04-01. Available also from: [https://nanoporetech.com/products/specifications#comparison\[tab\]=platform](https://nanoporetech.com/products/specifications#comparison[tab]=platform).
23. YOURGENOME. *What is Oxford Nanopore Technology (ONT) sequencing?* 2022-10-01. Available also from: <https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing/>.
24. ROBERTS, Richard J; CARNEIRO, Mauricio O; SCHATZ, Michael C. The advantages of SMRT sequencing. *Genome Biol.* 2013, vol. 14, no. 7, p. 405.

25. PACIFIC BIOSCIENCES. *REVIO SYSTEM*. 2023-04-01. Available also from: <https://www.pacb.com/revio/>.
26. LIU, Lin et al. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012, vol. 2012, p. 251364.
27. PACIFIC BIOSCIENCES. *ALTERNATIVE SIZE SELECTION METHODS FOR SMRTBELL® PREP KIT 3.0*. 2023-04-01. Available also from: <https://www.pacb.com/wp-content/uploads/Technical-note-Alternative-size-selection-methods-for-SMRTbell-prep-kit-3.0.pdf>.
28. PACIFIC BIOSCIENCES. *HIFI WGS AT SCALE ON THE SEQUEL IIe SYSTEM*. 2023-04-01. Available also from: <https://www.pacb.com/wp-content/uploads/HiFi-whole-genome-sequencing-at-scale-on-the-Sequel-IIe-system.pdf>.
29. ILLUMINA. *Illumina sequencing platforms*. 2023-04-16. Available also from: <https://www.illumina.com/systems/sequencing-platforms.html>.
30. EWING, B; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 1998, vol. 8, no. 3, pp. 186–194.
31. WANG, Xin Victoria; BLADES, Natalie; DING, Jie; SULTANA, Razvan; PARMIGIANI, Giovanni. Estimation of sequencing error rates in short reads. *BMC Bioinformatics.* 2012, vol. 13, p. 185.
32. STOLER, Nicholas; NEKRUTENKO, Anton. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform.* 2021, vol. 3, no. 1, lqab019.
33. AMARASINGHE, Shanika L et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 2020, vol. 21, no. 1, p. 30.
34. OXFORD NANOPORE TECHNOLOGIES. *NANOPORE SEQUENCING ACCURACY*. 2023-04-01. Available also from: <https://nanoporetech.com/accuracy>.
35. DELAHAYE, Clara; NICOLAS, Jacques. Sequencing DNA with nanopores: Troubles and biases. *PLoS One.* 2021, vol. 16, no. 10, e0257521.
36. OXFORD NANOPORE TECHNOLOGIES. *Oxford Nanopore Technologies store*. 2023-04-16. Available also from: <https://store.nanoporetech.com/>.
37. PACIFIC BIOSCIENCES. *PacBio Announces Record Orders, Including Orders for 76 Revio Systems Received in the Fourth Quarter of 2022*. 2023-04-01. Available also from: <https://www.pacb.com/press-releases/pacbio-announces-record-orders-including-orders-for-76-revio-systems-received-in-the-fourth-quarter-of-2022/>.

38. KIRCHER, Martin; KELSO, Janet. High-throughput DNA sequencing—concepts and limitations. *Bioessays*. 2010, vol. 32, no. 6, pp. 524–536.
39. TEDERSOO, Leho; ALBERTSEN, Mads; ANSLAN, Sten; CALLAHAN, Benjamin. Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. *Appl. Environ. Microbiol.* 2021, vol. 87, no. 17, e0062621.
40. LI, Heng et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009, vol. 25, no. 16, pp. 2078–2079.
41. NATIONAL ECOLOGICAL OBSERVATORY NETWORK. *Hierarchical Data Formats - What is HDF5?* 2023-05-04. Available also from: <https://www.neonscience.org/resources/learning-hub/tutorials/about-hdf5>.
42. HOSSEINI, Morteza; PRATAS, Diogo; PINHO, Armando J. A Survey on Data Compression Methods for Biological Sequences. *Information*. 2016, vol. 7, no. 4, p. 56.
43. WICK, Ryan R; JUDD, Louise M; HOLT, Kathryn E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* 2019, vol. 20, no. 1, p. 129.
44. WENGER, Aaron M et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 2019, vol. 37, no. 10, pp. 1155–1162.
45. PACIFIC BIOSCIENCES. *Pacific Biosciences Terminology*. 2023-05-04. Available also from: http://files.pacb.com/software/smrtanalysis/2.2.0/doc/smrtportal/help/!SSL!/Webhelp/Portal_PacBio_Glossary.htm.
46. BAID, Gunjan et al. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat. Biotechnol.* 2023, vol. 41, no. 2, pp. 232–238.
47. MURIGNEUX, Valentine et al. MicroPIPE: validating an end-to-end workflow for high-quality complete bacterial genome construction. *BMC Genomics*. 2021, vol. 22, no. 1, p. 474.
48. KRIMAN, Samuel et al. QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions. 2019. Available from arXiv: 1910.10261 [eess.AS].
49. XU, Zhimeng et al. Fast-bonito: A faster deep learning based basecaller for nanopore sequencing. *Artificial Intelligence in the Life Sciences*. 2021, vol. 1, p. 100011.
50. PEREŠINI, Peter; BOŽA, Vladimír; BREJOVÁ, Broňa; VINAR, Tomáš. Nanopore base calling on the edge. *Bioinformatics*. 2021, vol. 37, no. 24, pp. 4661–4667.

51. FUKASAWA, Yoshinori; ERMINI, Luca; WANG, Hai; CARTY, Karen; CHEUNG, Min-Sin. LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. *G3*. 2020, vol. 10, no. 4, pp. 1193–1196.
52. LANFEAR, R; SCHALAMUN, M; KAINER, D; WANG, W; SCHWESSINGER, B. MinIONQC: fast and simple quality control for MinION sequencing data. *Bioinformatics*. 2019, vol. 35, no. 3, pp. 523–525.
53. ANDREWS, Simon. *Illumina sequencing platforms*. 2023-05-01. Available also from: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
54. ALSER, Mohammed et al. Technology dictates algorithms: recent developments in read alignment. *Genome Biol*. 2021, vol. 22, no. 1, p. 249.
55. LI, Heng. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018, vol. 34, no. 18, pp. 3094–3100.
56. LOTEMPIO, Jonathan; DÉLOT, Emmanuèle; VILAIN, Eric. *Benchmarking long-read genome sequence alignment tools for human genomics applications*. 2021.
57. CHAISSON, Mark J; TESLER, Glenn. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*. 2012, vol. 13, p. 238.
58. SEDLAZECK, Fritz J et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*. 2018, vol. 15, no. 6, pp. 461–468.
59. ZHANG, Haowen; JAIN, Chirag; ALURU, Srinivas. A comprehensive evaluation of long read error correction methods. *BMC Genomics*. 2020, vol. 21, no. Suppl 6, p. 889.
60. FIRTINA, Can; BAR-JOSEPH, Ziv; ALKAN, Can; CICEK, A Ercument. Hercules: a profile HMM-based hybrid error correction algorithm for long reads. *Nucleic Acids Res*. 2018, vol. 46, no. 21, e125.
61. MICLOTTE, Giles et al. Jabba: hybrid error correction for long sequencing reads. *Algorithms Mol. Biol*. 2016, vol. 11, p. 10.
62. BAO, Ergude; LAN, Lingxiao. HALC: High throughput algorithm for long read error correction. *BMC Bioinformatics*. 2017, vol. 18, no. 1, p. 204.
63. BAO, Ergude; XIE, Fei; SONG, Changjin; SONG, Dandan. FLAS: fast and high-throughput algorithm for PacBio long-read self-correction. *Bioinformatics*. 2019, vol. 35, no. 20, pp. 3953–3960.

64. SALMELA, Leena; WALVE, Riku; RIVALS, Eric; UKKONEN, Esko. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics*. 2017, vol. 33, no. 6, pp. 799–806.
65. KOREN, Sergey et al. *Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation*. 2017.
66. MORISSE, Pierre; LECROQ, Thierry; LEFEBVRE, Arnaud. *Long-read error correction: a survey and qualitative comparison*. 2021.
67. AMARASINGHE, Shanika L; RITCHIE, Matthew E; GOUIL, Quentin. long-read-tools.org: an interactive catalogue of analysis methods for long-read sequencing data. *Gigascience*. 2021, vol. 10, no. 2.
68. CHEN, Pin et al. Portable nanopore-sequencing technology: Trends in development and applications. *Front. Microbiol.* 2023, vol. 14, p. 1043967.
69. LIAO, Xingyu et al. Current challenges and solutions of de novo assembly. *Quant. Biol.* 2019, vol. 7, no. 2, pp. 90–109.
70. NURK, Sergey et al. *The complete sequence of a human genome*. 2021.
71. KIRSCHKE, Melanie; SCHATZ, Michael C. Democratizing long-read genome assembly. *Cell Syst.* 2021, vol. 12, no. 10, pp. 945–947.
72. VASER, Robert; ŠIKIĆ, Mile. Time- and memory-efficient genome assembly with Raven. *Nature Computational Science*. 2021, vol. 1, no. 5, pp. 332–336.
73. WANG, Jinming et al. Systematic Comparison of the Performances of De Novo Genome Assemblers for Oxford Nanopore Technology Reads From Piroplasm. *Front. Cell. Infect. Microbiol.* 2021, vol. 11, p. 696669.
74. JUNG, Hyungtaek; JEON, Min-Seung; HODGETT, Matthew; WATERHOUSE, Peter; EYUN, Seong-Il. Comparative Evaluation of Genome Assemblers from Long-Read Sequencing for Plants and Crops. *J. Agric. Food Chem.* 2020, vol. 68, no. 29, pp. 7670–7677.
75. LIU, Hailin; WU, Shigang; LI, Alun; RUAN, Jue. SMARTdenovo: a de novo assembler using long noisy reads. *GigaByte*. 2021, vol. 2021, gigabyte15.
76. KOLMOGOROV, Mikhail; YUAN, Jeffrey; LIN, Yu; PEVZNER, Pavel A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 2019, vol. 37, no. 5, pp. 540–546.
77. VOLLGER, Mitchell R et al. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann. Hum. Genet.* 2020, vol. 84, no. 2, pp. 125–140.

78. GAVRIELATOS, Marios; KYRIAKIDIS, Konstantinos; SPANDIDOS, Demetrios A; MICHALOPOULOS, Ioannis. Benchmarking of next and third generation sequencing technologies and their associated algorithms for de novo genome assembly. *Mol. Med. Rep.* 2021, vol. 23, no. 4.
79. NURK, Sergey et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* 2020, vol. 30, no. 9, pp. 1291–1305.
80. CHENG, Haoyu; CONCEPCION, Gregory T; FENG, Xiaowen; ZHANG, Haowen; LI, Heng. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods.* 2021, vol. 18, no. 2, pp. 170–175.
81. BUCHANAN, Janet A; SCHERER, Stephen W. Contemplating effects of genomic structural variation. *Genet. Med.* 2008, vol. 10, no. 9, pp. 639–647.
82. ZVERINOVA, Stepanka; GURYEV, Victor. Variant calling: Considerations, practices, and developments. *Hum. Mutat.* 2022, vol. 43, no. 8, pp. 976–985.
83. SINGH, Ashish Kumar et al. Detecting copy number variation in next generation sequencing data from diagnostic gene panels. *BMC Med. Genomics.* 2021, vol. 14, no. 1, p. 214.
84. LAPPALAINEN, Tuuli; SCOTT, Alexandra J; BRANDT, Margot; HALL, Ira M. Genomic Analysis in the Age of Human Genome Sequencing. *Cell.* 2019, vol. 177, no. 1, pp. 70–84.
85. KOBOLDT, Daniel C. Best practices for variant calling in clinical sequencing. *Genome Med.* 2020, vol. 12, no. 1, p. 91.
86. AUDANO, Peter A et al. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell.* 2019, vol. 176, no. 3, 663–675.e19.
87. MIZUGUCHI, Takeshi et al. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. *J. Hum. Genet.* 2019, vol. 64, no. 5, pp. 359–368.
88. MIAO, Hefan et al. Long-read sequencing identified a causal structural variant in an exome-negative case and enabled preimplantation genetic diagnosis. *Hereditas.* 2018, vol. 155, p. 32.
89. PACIFIC BIOSCIENCES. *pbsv*. 2023-05-08. Available also from: <https://github.com/PacificBiosciences/pbsv>.
90. JIANG, Tao et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* 2020, vol. 21, no. 1, pp. 1–24.
91. POPLIN, Ryan et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* 2018, vol. 36, no. 10, pp. 983–987.

92. OXFORD NANOPORE TECHNOLOGIES. *Medaka*. 2023-05-08. Available also from: <https://github.com/nanoporetech/medaka>.
93. AHSAN, Mian Umair; LIU, Qian; FANG, Li; WANG, Kai. NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks. *Genome Biol.* 2021, vol. 22, no. 1, pp. 1–33.
94. RAMACHANDRAN, Anand; LUMETTA, Steven S; KLEE, Eric; CHEN, Deming. *HELLO: A hybrid variant calling approach*. 2020.
95. ESTELLER, Manel. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.* 2007, vol. 8, no. 4, pp. 286–298.
96. CLARK, Tyson A; SPITTLE, Kristi E; TURNER, Stephen W; KORLACH, Jonas. Direct detection and sequencing of damaged DNA bases. *Genome Integr.* 2011, vol. 2, p. 10.
97. GOUIL, Quentin; KENIRY, Andrew. Latest techniques to study DNA methylation. *Essays Biochem.* 2019, vol. 63, no. 6, pp. 639–648.
98. XU, Liu; SEKI, Masahide. Recent advances in the detection of base modifications using the Nanopore sequencer. *J. Hum. Genet.* 2020, vol. 65, no. 1, pp. 25–33.
99. LIU, Qian; GEORGIEVA, Daniela C; EGLI, Dieter; WANG, Kai. NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data. *BMC Genomics.* 2019, vol. 20, no. Suppl 1, p. 78.
100. SIMPSON, Jared. *Nanopolish*. 2023-05-08. Available also from: <https://github.com/jts/nanopolish>.
101. LIU, Qian et al. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat. Commun.* 2019, vol. 10, no. 1, p. 2449.
102. FANG, Gang et al. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* 2012, vol. 30, no. 12, pp. 1232–1239.
103. BEAULAURIER, John et al. Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nat. Commun.* 2015, vol. 6, p. 7438.
104. PACIFIC BIOSCIENCES. *Detecting DNA Base Modifications Using Single Molecule, Real-Time Sequencing*. 2023-05-16. Available also from: https://www.pacb.com/wp-content/uploads/2015/09/WP_Detecting_DNA_Base_Modifications_Using_SMRT_Sequencing.pdf.

105. POLLARD, Martin O; GURDASANI, Deepti; MENTZER, Alexander J; PORTER, Tarryn; SANDHU, Manjinder S. Long reads: their purpose and place. *Hum. Mol. Genet.* 2018, vol. 27, no. R2, R234–R241.
106. KELLER, Matthew W et al. Direct RNA Sequencing of the Coding Complete Influenza A Virus Genome. *Sci. Rep.* 2018, vol. 8, no. 1, p. 14408.
107. STARK, Rory; GRZELAK, Marta; HADFIELD, James. RNA sequencing: the teenage years. *Nat. Rev. Genet.* 2019, vol. 20, no. 11, pp. 631–656.
108. SIM, Justin; CHAPMAN, Brendan. In-field whole genome sequencing using the MinION nanopore sequencer to detect the presence of high-priced military targets. *Aust. J. Forensic Sci.* 2019, vol. 51, no. sup1, S86–S90.
109. QUICK, Joshua et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature.* 2016, vol. 530, no. 7589, pp. 228–232.
110. JESUS, Jaqueline Goes de et al. Genomic detection of a virus lineage replacement event of dengue virus serotype 2 in Brazil, 2019. *Mem. Inst. Oswaldo Cruz.* 2020, vol. 115, e190423.
111. XU, Luming et al. Long-read sequencing identifies novel structural variations in colorectal cancer. *PLoS Genet.* 2023, vol. 19, no. 2, e1010514.
112. SAKAMOTO, Yoshitaka et al. Long-read sequencing for non-small-cell lung cancer genomes. *Genome Res.* 2020, vol. 30, no. 9, pp. 1243–1257.
113. AGANEZOV, Sergey et al. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res.* 2020, vol. 30, no. 9, pp. 1258–1273.
114. SAKAMOTO, Yoshitaka et al. Phasing analysis of lung cancer genomes using a long read sequencer. *Nat. Commun.* 2022, vol. 13, no. 1, p. 3464.
115. MAYOR, Neema P et al. Recipients Receiving Better HLA-Matched Hematopoietic Cell Transplantation Grafts, Uncovered by a Novel HLA Typing Method, Have Superior Survival: A Retrospective Study. *Biol. Blood Marrow Transplant.* 2019, vol. 25, no. 3, pp. 443–450.
116. JOHANSSON, Tiira; KOSKELA, Satu; YOHANNES, Dawit A; PARTANEN, Jukka; SAAVALAINEN, Päivi. Targeted RNA-Based Oxford Nanopore Sequencing for Typing 12 Classical HLA Genes. *Front. Genet.* 2021, vol. 12, p. 635601.
117. TURNER, Thomas R et al. Widespread non-coding polymorphism in HLA class II genes of International HLA and Immunogenetics Workshop cell lines. *Hladnikia.* 2022, vol. 99, no. 4, pp. 328–356.

118. LIU, Chang; BERRY, Rick. Rapid High-Resolution Typing of Class I HLA Genes by Nanopore Sequencing. *Methods Mol. Biol.* 2020, vol. 2120, pp. 93–99.
119. DE SANTIS, Dianne; TRUONG, Linh; MARTINEZ, Patricia; D’ORSOGNA, Lloyd. Rapid high-resolution HLA genotyping by MinION Oxford nanopore sequencing for deceased donor organ allocation. *Hladnikia.* 2020, vol. 96, no. 2, pp. 141–162.
120. DOWNING, Jonathan; D’ORSOGNA, Lloyd. High-resolution human KIR genotyping. *Immunogenetics.* 2022, vol. 74, no. 4, pp. 369–379.