UNIVERSITY OF TARTU
FACULTY OF SCIENCE AND TECHNOLOGY
INSTITUTE OF MOLECULAR AND CELL BIOLOGY
CHAIR OF BIOINFORMATICS

Daria Kholodniuk

Utilizing the Knowledge of the Completely Assembled Human Chromosomes X and 8

Bachelor's Thesis (12 ECTS)

Curriculum Science and Technology

Supervisor:
Ph.D. Reidar Andreson

TARTU 2021

ABSTRACT

Utilizing the knowledge of the completely assembled human chromosomes X and 8

DNA sequencing is the process of determining the sequence of nucleotides in a DNA chain. It is of great importance in various fields of science, as it provides information "encoded" in our genome. From 1990 to 2003, scientists within the Human Genome Project worked on sequencing the human genome. The assembly was complete and brought many discoveries, but it was not perfect and had many gaps. In 2020, a team of scientists from the Telomere-to-Telomere consortium announced a new version of the human genome, where the gaps were filled with the help of modern sequencing technologies. They also succeeded in completely assembling human chromosomes X and 8. One of the aims of my thesis is to theoretically study CHM13hTERT cell line and the manipulations done on it in order to understand what contribution they made to genome sequencing. Moreover, the plan of how one can adapt information about completely sequenced chromosomes X and 8 is suggested in this work.

Keywords: NGS sequencing, complete human chromosomes, telomeres, repeats

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Täielikult järjestatud inimese kromosoomide X ja 8 analüüsimisvõimalused

DNA sekveneerimiseks nimetatakse protsessi, mille käigus määratakse DNA ahelas olevate nukleotiidide järjekord. Sekveneerimine võimaldab tänu genoomi “kodeeritud” informatsiooni tuvastamisele tulemusi rakendada erinevates teadusharudes. Ajavahemikus 1990-2003 suutsid Inimese Genoomi Projektis (*Human Genome Project*) osalevad teadlased ära järjestada inimese genoomse DNA järjestuse. Tulemusena saadud genoomi järjestus rajas teed mitmetele bioloogilistele avastustele. Siiski ei olnud tegemist täieliku järjestusega, kuna see sisaldas mitmeid kaardistamata regioone. 2020. aastal avalikustasid Telomere-to-Telomere konsortsiumis osalevad teadlased uue inimese genoomi versiooni, mille puhul olid paljud puuduvad DNA regioonid uute sekveneerimise meetoditega kaardistatud. Neil õnnestus täielikult järjestada inimese kromosoomide X ja 8 järjestused. Käesoleva uurimistöö üheks eesmärgiks oli uurida CHM13hTERT rakuliini ja sellel tehtud muudatusi mõistmaks, kuidas

seda kasutada genoomide järjestamisel. Lisaks arutlen, kuidas võiks edasi analüüsida täielikult sekveneeritud kromosoomide X ja 8 andmeid ning mis võiks olla selle tulemuseks.

Võtmesõnad: NGS sekveneerimine, täielikud inimese kromosoomid, telomeerid, kordused

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

TABLE OF CONTENTS

ABBREVIATIONS

ABI SOLiD	Sequencing by Oligonucleotide Ligation and Detection
bp	Base Pair
CLR	Continuous Long Reads
dNTP	Deoxyribonucleotide
ddNTP	Dideoxyribonucleotide
DLS	Direct Label and Stain
FGS	First Generation Sequencing
GEM	Gel Bead-In-Emulsions
HiFi	High-Fidelity
HMW DNA	High-Molecular-Weight DNA
HOR	Higher Order Repeat
HGP	Human Genome Project
mtDNA	Mitochondrial DNA
NGS	Next Generation Sequencing
PacBio	Pacific Biosciences
PITC	Phenylisothiocyanate
Poll	Klenow fragment of DNA Polymerase I
PCR	Polymerase Chain Reaction
PGM	Personal Genome Machine
SGS	Second Generation Sequencing
SMS	Single Molecule Sequencing
STR	Short Tandem Repeat
TRF analysis	Terminal Restriction Fragment Analysis
VNTR	Variable Number Tandem Repeat

INTRODUCTION

Genome sequencing is the process of determining the DNA nucleotide order of an organism's genome. Knowledge of DNA sequences has become essential for biological research and crucial for applied fields such as biotechnology, medical diagnosis and genetic counseling. One of the largest collaborative biological projects, the Human Genome Project, has been working on determination of the base pairs that make up human DNA for 13 years. The resulting human reference genome (GRCh38) is the most accurate and complete vertebrate genome ever obtained (Miga et al., 2020). However, not a single chromosome was completed and there were many gaps left.

Last year the Telomere-to-Telomere consortium announced that they managed to surpass the previously obtained genome. Moreover, scientists have assembled the first complete set of human chromosomes: Chromosome X and Chromosome 8. The discovery of chromosome sequences is an undeniably important event for science. Thanks to this discovery, various genomic mutations and genetic diseases can be better studied. For such a discovery, Consortium has used a combination of different technologies: Oxford Nanopore, PacBio CLR, PacBio HiFi, 10X Genomics, as well as BioNano DLS and Arima Genomics HiC. What is particularly important for this study is that T2T[1] used a CHM13hTERT cell line.

Haploid CHM13hTERT cell line (derived from a complete hydatidiform mole CHM13) is suitable for effective sequencing due to its haploidy. Genomes come from a single sperm that has gone through post-meiotic chromosomal duplication. Hence, genomes are homozygous for one set of alleles. This cell line was previously used to fill gaps in the human reference genome and other studies. Karyotyping also confirmed a karyotype 46,XX without abnormalities (Miga et al., 2020). This allows scientists to avoid the laborious assembling of both haplotypes of a diploid genome. The scientists avoided sequencing two non-identical chromosomes from a normal human cell. They needed a certain type of cell that would have a set of two identical chromosomes. Such cells provide a larger set of DNA for sequencing than, for example, a male cell with only one X chromosome. It also avoids sequence differences that occur when analyzing two X chromosomes of a typical female cell.

The task is to theoretically study these cells and the manipulations done on them in order to understand what contribution they made to genome sequencing. Moreover, the plan of how one can adapt information about completely sequenced chromosomes X and 8 is suggested in this work.

[1] Telomere-to-Telomere consortium

LITERATURE REVIEW

1.1 DNA Sequencing

Knowing the nucleic acid sequence is crucial because those sequences contain important information about biochemical properties and inheritance features of various living organisms. Thus, the method one can use for obtaining such sequences is essential for research works and understanding life in general (Griffiths, 2012). DNA sequencing is the process of determining the sequence of nucleotides in DNA chains. Before the actual approaches for DNA sequencing were introduced, there were only technologies for determining protein chains' sequences (order of amino acids determination): Edman degradation (Edman, 1950), (Doucet and Overall, 2011) and later mass spectrometry (Yates, 2011).

Edman degradation is an ordered amino acid identification from N to C terminus. The mechanism of that method is that Edman reagent (PITC) reacts with an amino group at the N terminus of a peptide at basic pH and the modified amino acid is seceded as the thiazolinone derivative at acidic pH (Smith, 2001). At the end thiazolinone product is identified with help of chromatography (Castagnola et al., 1988). One of the disadvantages of that method is that only short purified peptide chains can be assessed - meaning it cannot be used for huge complex protein compounds which are present in most biological organisms.

Mass spectrometry is a method for compound identification as well. In short, molecules are being 'weighted'. Components are being ionized and the mass-ion charge ratio is measured (Glish and Vachet, 2003). The overall process of tandem mass spectrometry involves degradation of proteins to a mixture of peptides which are fractionated by high-performance liquid chromatography (Coon et al., 2018).

Those protein sequencing technologies were not suitable for DNA sequencing, because DNA components, nucleotides, are smaller than amino acids and they are very similar to each other. At the same time, the DNA chain itself is longer (Hutchison, 2007). First works have been done on the RNAs because they are shorter than DNA, it is easier to produce in culture and they are not complicated with complementarity (Kukurba and Montgomery, 2015). Analytical chemistry methods were able only to measure nucleotide composition but not the order of them (Holley et al., 1961). But those methods were mixed with selective ribonuclease treatments (Gerashchenko and Gladyshev, 2016) and it became possible to obtain the first nucleic acid sequence alanine tRNA from *Saccharomyces cerevisiae* (Holley, 1965).

Contemporaneously Fred Sanger with his colleagues came up with resembling technology based on the detection of radiolabelled partial-digestion fragments after two-dimensional fractionation (Sanger, 1965).

1.2 First Generation Sequencing

Two-dimensional fractionation was changed to single separation by polynucleotide length via electrophoresis through polyacrylamide gels (Maniatis, 1975). That method was mainly used in the 'plus and minus' technique which was created by Alan Coulson and Fred Sanger (Sanger, 1975) and in Walter Gilbert's chemical cleavage technique (Maxam and Gilbert, 1977). In 1977 Sanger offered a new method which was called chain-termination or dideoxy technique (Sanger, 1977).

1.2.1 'Plus and Minus' Sanger's Technique

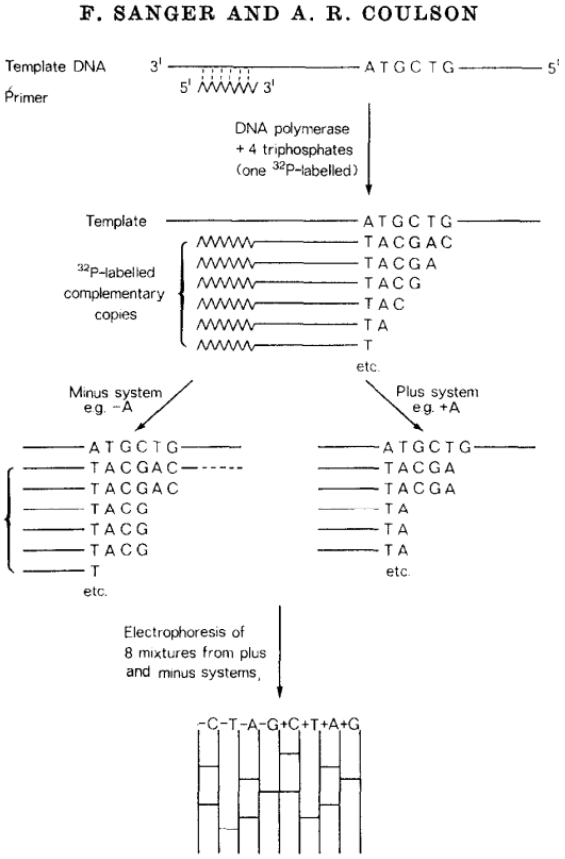

FIG. 1. The principle of the method.

Figure 1. *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase*. Single DNA helix fragment is used as a base in polymerase copying reaction. For primers synthetic oligonucleotides or natural subfragments obtained by hydrolysis with restriction endonucleases are used. Poll from E.coli is used as an enzyme. There are two stages in the method.

First, under limited conditions polymerase reaction is carried out in the presence of all four types of dNTPs (one of them is labeled at the alpha position of phosphate), resulting in a set of products of incomplete copying of the template fragment. The mixture is purified from unbound deoxynucleoside triphosphates and divided into eight pairs. In the 'plus' system four reactions are carried out in the presence of each four types of nucleotides, in the 'minus' system in the absence of them. As a result, in the 'minus' system termination takes place before the dNTP of a certain type, in the 'plus' system - after it. The eight obtained samples are separated by electrophoresis, the signal is 'read' and the DNA sequence is determined. Image adapted from: Sanger and Coulson (1975).

'Plus and minus' method involves using a primer with radiolabeled nucleotides which undergoes an interaction with DNA polymerase for synthesis. That includes two reactions: polymerization ('plus') reaction in which each nucleotide type is present and 'minus' reaction in which each nucleotide type is missing. In the result, in 'minus' system termination occurs before the dNTP of a given nucleotide type, whereas in 'plus' system termination occurs after it. Eight resulting products are then put into polyacrylamide gel and results are analyzed. This way the first DNA genome was sequenced *bacteriophage ϕX174 (PHix)* (Sanger et al., 1977) (Figure 1).

1.2.2 Maxam-Gilbert Method

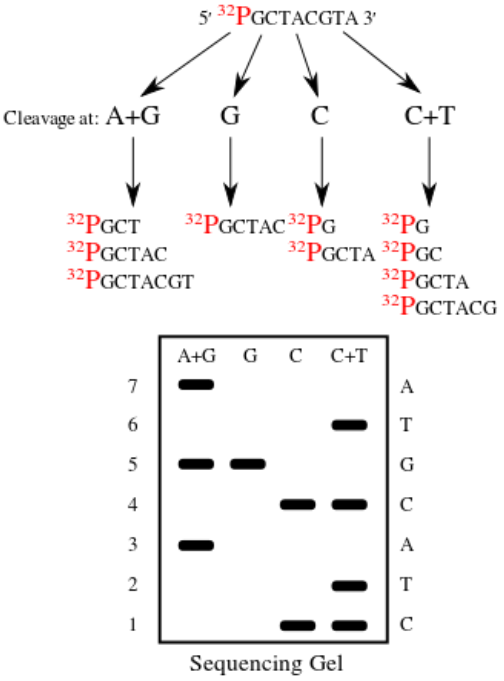

Figure 2. *An example Maxam–Gilbert sequencing reaction*. It requires radioactive labeling at 5'-end of the DNA fragment to be sequenced (by kinase reaction using gamma-32 P ATP) and it requires DNA purification. Chemical treatment breaks one or two of the four nucleotide bases in each of the

four reactions (G, A+G, C, C+T). Purines are depurinated with formic acid, guanines (as well as adenines to some extent) are methylated with dimethyl sulfate and pyrimidines are hydrolyzed using hydrazine. Sodium chloride is added to hydrazine reaction to inhibit the reaction of thymine to reaction with C only. The modified DNAs are cleaved with hot piperidine in the position of the modified base. Obtained fragments undergo acrylamide gel electrophoresis. Image adapted from: https://en.wikipedia.org/wiki/Maxam%E2%80%93Gilbert_sequencing

Radiolabeled DNA is treated with chemicals which break the chain in different places and then it is put into the polyacrylamide gel. DNA fragments result in different lengths and that is how they are being distinguished. This technique can be considered as first generation sequencing (Figure 2).

1.2.3 Sanger Sequencing

In this method chemical analogues dNTP are used, monomers of the DNA. ddNTPs are lacking 3'-hydroxyl groups which are needed for DNA chains, and they cannot create a bond with 5'-phosphate of the next dNTP (Chidgeavadze, 1984). This method can be performed manually or with the sequencing machine. The process can be divided into three main parts: PCR with fluorescent chain-terminating ddNTPs, size separation by capillary gel electrophoresis, laser excitation and detection by sequence machine or manually assessed results (Figure 3). Polyacrylamide gel is used instead of agarose due to its high resolving power, meaning that it can separate DNA fragments that differ only in 1 bp (Stellwagen, 2009). In the sequencing machine fluorescence is used. Laser excites the fluorescence tags in every band and the computer analyzes emitted light. Output is called a chromatogram which shows the fluorescent peak of each nucleotide (Pareek, 2011).

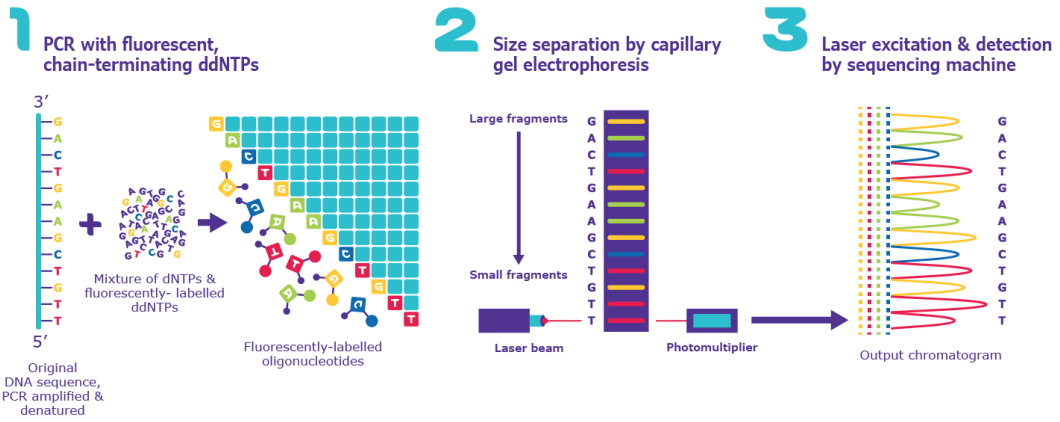

Figure 3. *Three Basic Steps of Automated Sanger Sequencing.* Mixture of dNTPs and fluorescently-labeled ddNTPs is added to the original DNA sequence which is PCR amplified and

denatured. It all results in fluorescently-labeled oligonucleotides, which are then put into acrylamide gel and electrophoresis is done. Fragments are separated by their lengths, a laser beam goes through the sample, and the result is detected by the machine. Image adapted from: https://www.sigmaaldrich.com/technical-documents/articles/biology/sanger-sequencing.html

1.3 Second Generation Sequencing

With the development of PCR and recombinant DNA technologies there was a huge contribution into overall genomics development (Elrich, 2013). It was thanks to the better abilities to obtain high concentrations of clean DNAs needed for sequencing. Eventually, new dideoxy sequencers, ABI PRISM (genetic analyzer), have made it possible to sequence many samples simultaneously, which made a huge contribution to the Human Genome project.

Second generation sequencing can be divided into two major types: sequencing by hybridization and sequencing by synthesis (Slatko et al., 2018).

1.3.1 Sequencing by Hybridization

The process of hybridization can be defined as the process when one DNA strand binds to its complementary strand in the DNA double-helix. In the sequencing by hybridization sets of oligonucleotides are hybridized under certain conditions that allow detection of complementary sequences in the target nucleic acid (Figure 4) (Drmanac et al., 2002).

Figure 4. *Probe overlaps and assembly branches.* **A.** Here the sequence sample containing known primer sequences PPPPPPP at each end is illustrated and two 7-mer repeat sequences are shown. **B.** The assembly using overlapped probes is started from the 5'-end. First, the positive probe that

corresponds to the primer sequence is selected. Sequence extension continues in the 3'-direction until the first branching point with two possible options for further extension occurs (GGTCCCTc and GGTCC-CTa). **C.** Branching process leads to many alternative sequences. There is a requirement for the sequence to end with 3'-primer and for sequence subfragments between branching points to occur only once - that leads to assembly of two possible sequences (**D**), which differ in placement of CAA and AAT sub-sequences. Image adapted from: Drmanac et al. (2002).

1.3.2 Sequencing by Synthesis

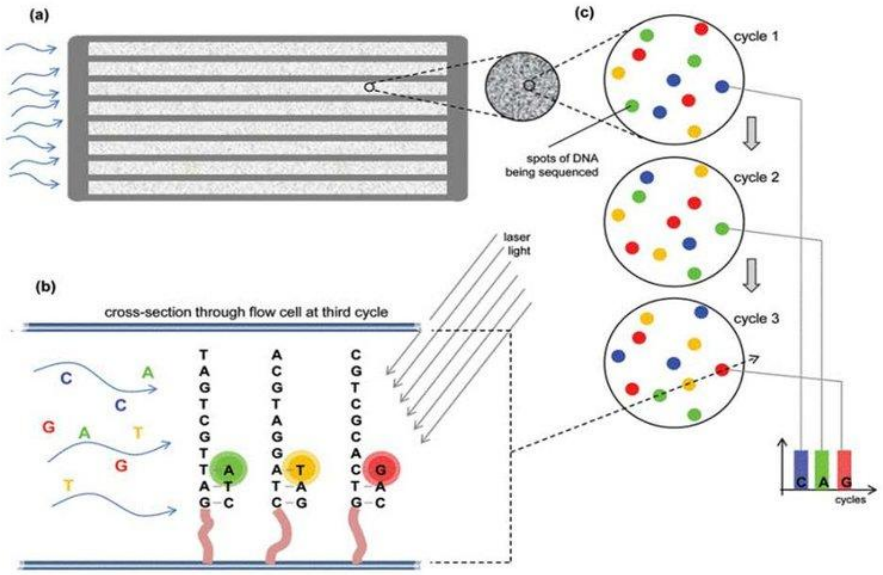

Figure 5. *Illustration of Illumina sequencing by synthesis method.* **(a)** Flow cell overview. **(b)** Incorporation of nucleotides results in fluorescence release. **(c)** Zoomed in the flow cell - different nucleotides with their specific fluorescence color. Image adapted from: Bucher (2019).

Method of sequencing by synthesis allows the sequence of one DNA chain by synthesis of complementary chains while each nucleotide junction is registered. During the reaction DNA matrix is immobilized, nucleotides solutions are added and washed sequentially after each sequencing cycle. That technology was originally developed by Shankar Balasubramanian and David Klenerman at the University of Cambridge (Bentley et al., 2008). They created the company Solexa in 1998 to commercialize their sequencing method. Later, Illumina purchased Solexa and improved the original technology. Sequencing by synthesis is similar to Sanger sequencing but this method involves modified dNTPs that contain a terminator

which is able to block further polymerization (so only a single base can be added using a polymerase enzyme to each growing DNA copy strand) (Figure 5).

Illumina sequencing method is considered to be already the next generation sequencing, although sequencing by synthesis itself is referred to as second generation sequencing.

1.4 Next Generation Sequencing

Next generation sequencing (NGS) combines different modern sequencing technologies, the main work principle of which involves simultaneous sequencing of many samples. NGS technologies include: Illumina (Solexa) sequencing, Roche 454 sequencing, Ion Torrent: Proton/ PGM sequencing. In NGS DNA is broken into short pieces, amplified and sequenced. NGS is famous for speed superiority compared to all previous sequencing technologies. NGS techniques have been used for whole viral genome sequencing (e.g. new influenza viruses), detection of viral genome variability and monitoring of low-abundance antiviral drug-resistance mutations (Barzon et al., 2011). NGS methods have different biochemical backgrounds and various sequencing protocols (Roche 454, Illumina GA, Ion Torrent PGM, ligation sequencing SOLiD) and they differ in lengths. Therefore, the SOLiD system is the most suitable for applications which need high throughput of sequences, but not long reads, for instance, RNA-sequencing projects or genome re-sequencing. While 454 and Illumina give data for de novo assembly (Figure 6) (Liu et al., 2012).

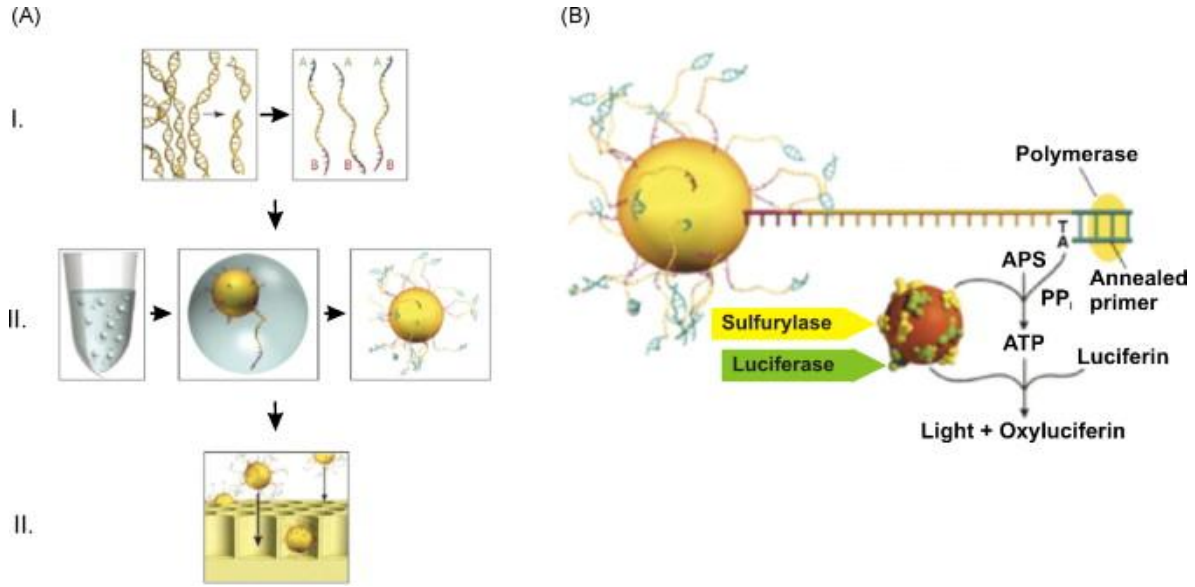

Figure 6. *Example of NGS sequencing technology.* **(A)** Outline of the GS 454 DNA sequencer workflow. (I) Library construction ligates 454-specific adapters to DNA fragments (indicated as A and B) and couples amplification beads with DNA in an emulsion PCR to amplify fragments before

sequencing (II). The beads are loaded into the picotiter plate (III). (B) Schematic illustration of the pyrosequencing reaction which occurs on nucleotide incorporation to report sequencing-by-synthesis. Image adapted from: Mardis (2007)

1.5 Third Generation Sequencing

There are still debates on how to distinguish third generation sequencing (TGS) from second, but in this work third generation sequencing is referred to the technologies that are able to synthesize separate molecules without DNA amplification in advance. The first single molecule sequencing (SMS) technology was invented by Stephen Quake lab (Braslavsky et al., 2003). Its working principle was close to the one of Illumina but without bridge amplification. DNA templates are attached to flat surface and then propriety fluorescent reversible terminator dNTPs (Bowers et al., 2009) are washing each base in turn and imaging before cleavage and cycling the next base over (Orenstein, 2009). In comparison with NGS, where DNA is broken into short pieces, amplified and sequenced, TGS directly synthesises single DNA molecules. The advantage of long reads is that they contain more information than short reads. Such long reads are beneficial for genome assemblies.

1.5.1 Pacific Biosciences (PacBio)

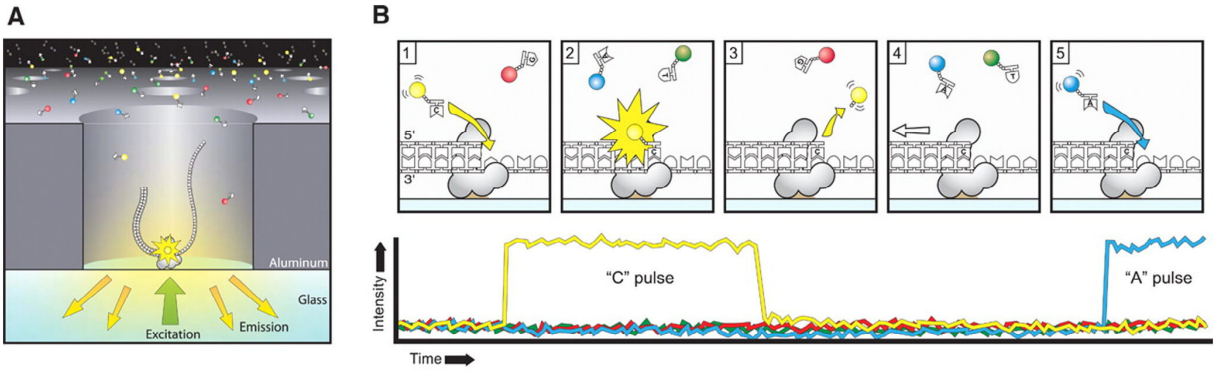

Figure 7. *Sequencing via light pulses.* Four nucleotide types are labelled by different fluorescent colors (red, yellow, green and blue for G, C, T and A correspondingly), they have different emission spectrums. Nucleotide is held in the detection volume by the polymerase - a light pulse which identifies the base is produced. **1.** Fluorescnetly labelled nucleotide connects with DNA template in the active centre of polymerase. **2.** Light emission corresponding to the base (e.g. yellow for C). **3.** The dye linker-pyrophosphate product is separated from the nucleotide and diffuses out of the ZMW to end the fluorescent pulse. **4.** Polymerase is moved onto the next position. 5. The next nucleotide connects with the template in the active centre of polymerase and initiates the next fluorescent impulse (e.g. “A” pulse). Image adapted from: Eid et al. (2009)

Single Molecule, Real-Time (SMRT) Sequencing is the core technology of PacBio (www.pacb.com) long-read sequencing platforms. It does not require PCR amplification. Zero-mode waveguides (ZMWs), subwavelength optical nanostructures fabricated in a thin metallic film, are powerful analytical tools that are capable of confining an excitation volume to the range of attoliters, which allows individual molecules to be isolated for optical analysis at physiologically relevant concentrations of fluorescently labelled biomolecules. Arrays of such nanostructures can be transformed into systems for analyzing the mass of single-molecule reactions in real time, which is the basic work principle of PacBio SMRT. (Plénat et al., 2017), (CD Genomics, n.d.)

PacBio SMRT Sequencing (www.cd-genomics.com) uses a ZMW innovation to identify the ideal fluorescent signal from the strong fluorescent background, which is caused by freely floating nucleotides. The connection of DNA polymerase to template DNA is attached to the lower glass ZMW surface. There is a laser which comes through that lower glass ZMW surface but does not go inside it completely due to the ZMW size (it is less than lightwave length). Therefore, that allows a selective excitation and light identification, which is excited by nucleotides that are used for base elongation (Figure 7). (Heather and Chain, 2016)

TGS has been widely used in genome research since PacBio released their PacBio RS II in 2013. In 2015 they launched their new sequencer - PacBio Sequel. Table 1 provides the comparison between PacBio RS II and PacBio Sequel.

	RS II	**Sequel**
Average read length	10-15 kb	8-12 kb
ZMWs	150,000	1,000,000
Data size	500Mb - 1 Gb	5-10 Gb
SMRT cell No./Run	1-16	1-16
Run time	0.5-6 hours	0.5-6 hours
Multiplex Amplicons	384	1536

Table 1. *RS II and Sequel comparison*. Data is taken from: www.pacb.com

1.5.1.1 PacBio CLR and HiFi

HiFi reads are highly accurate long reads, which provide accuracy up to >99.9% like short reads and Sanger sequencing. HiFi reads are a type of data obtained by using the circular consensus sequencing (CCS) mode on one of the PacBio Sequel Systems (www.pacb.com)

(Figure 8). HiFi reads are used for whole genome sequencing and RNA sequencing. HiFi reads are generated by calling consensus from subreads generated by multiple enzyme passes around a circularized template. As a result, the HiFi read is both long and accurate. (Wenger et al., 2019)

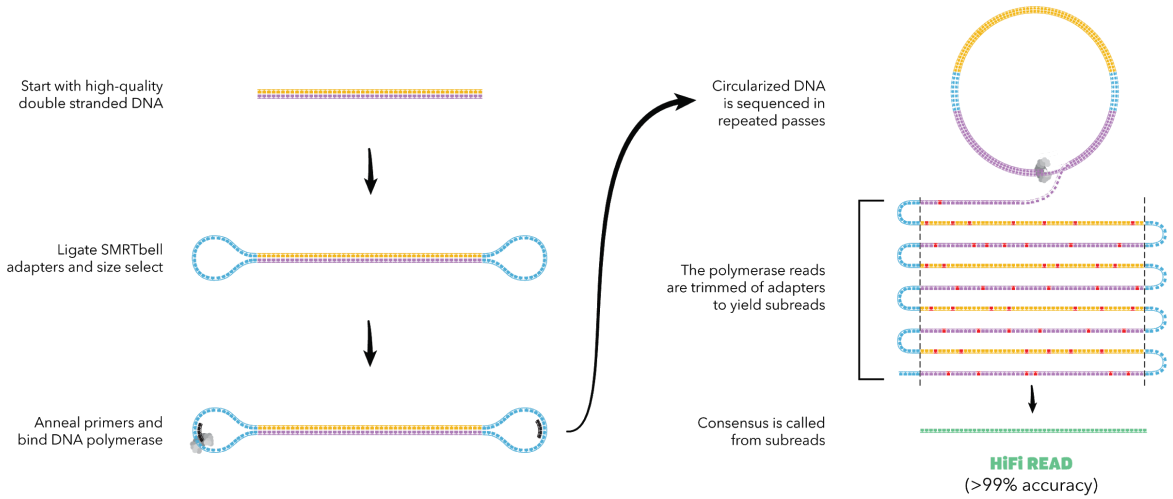

Figure 8. *HiFi sequencing illustration.* Image adapted from: www.pacb.com

The replication processes in each ZMW are recorded like a 'light impulses movie' and impulses from each ZMW can be interpreted as a sequence of bases, which is called continuous long read (CLR). PacBio RS II usually creates movies with a length of 0.5-4 hours. SMRTbell comprises a closed loop structure where polymerase replicates one strand of the target dsDNA and can continue incorporating bases of the adapter and then the other strand. If the lifetime of the polymerase is continuous enough then it can be used for sequencing several times in a single CLR. CLR can be divided into subreads by recognizing and cutting out the adaptor sequences. The consensus sequence of multiple subreads in a single ZMW produces a more accurate circular consensus sequence (CCS) read. A CCS read cannot be created if the target DNA is too long to be sequenced several times in a CLR, and only a single subread is output instead. (Rhoads and Au, 2015)

1.5.2 Oxford Nanopore

The main principle of nanopore sequencing method is analyzing DNA molecules while it is being pulled through a very little pore suspended in a membrane. Changes in electric currents occur which enable the detection of the base sequence (Ansorge et al., 2017). Nanopore sequencing method was invented by Oxford Nanopore and it allows to sequence DNA/RNA molecules in real time without the need of PCR amplification or labeling by chemicals.

Electrophoresis is usually used in nanopore sequencing to transport an unknown sample through a hole 10^{-9} m in diameter. An electrolyte solution is always present in the nanopore system, and an electric current emerges in the system under the influence of the constant electric field. The magnitude of the electric current density on the surface of a nanopore is determined by the size of the nanopores as well as the composition of the DNA or RNA that resides within the nanopore. Because the samples are close enough to the nanopores, they cause characteristic changes in the electric current density on the surfaces of the nanopores, enabling the sequencing process. (Bharagava et al., 2019)

Nanopore sequencing has been used successfully to characterize plasmids encoding carbapenemase in enterobacteriaceae isolated from wastewater treatment plants (Ludden et al., 2017) and to determine the correlation between resistome phenotype and E. coli genotype in urban wastewater (Xia et al., 2017).

1.5.3 10x Genomics

10x Genomics (www.10xgenomics.com) is a biotech company based in the United States that develops and manufactures gene sequencing technologies for use in scientific research. Serge Saxonov, Ben Hindson, and Kevin Ness founded it in 2012. In the upgraded human genome sequencing process T2T consortium used Linked-Reads by 10x Genomics (Logsdon et al., 2021), (Miga et al., 2020).

10x Genomics' Linked-Reads sequencing technology uses microfluidics to separate and barcode HMW DNA to create a data type that provides contextual information about the genome from short reads. This technology enables combining multiple analyses into a single powerful workflow with minimal input. Simplest terms, linked reads give a wealth of information based on short read sequencing data, allowing to discover information that one was previously unaware existed. (10x Genomics, n.d.)

10x Genomics' Linked-Read strategy employs emulsion technology in conjunction with a microfluidic instrument to separate long DNA fragments into micelles known as "GEM" (gel beads in emulsions). Within each GEM, portions of separated long DNA fragments are amplified, and each of these amplification products is labeled with a barcode unique to that GEM. The barcodes are used to classify sequences that are close in proximity in the genome after Illumina sequencing. Though each DNA fragment is often completely sequenced, data from several overlapping fragments originating from the same genomic region may be combined to create a read cloud. As a result, long stretches of the genome can be reconstructed more precisely using connected reads than using conventional whole genome

shotgun sequencing. De novo genome assembly, haplotype phasing, and structural variant detection are all supported by the details preserved from long fragments. (Ott et al., 2018)

1.5.4 BioNano DLS and Arima Genomics HiC

Direct Label and Stain (DLS) technology from Bionano is a new non-destructive chemistry for labeling genomic DNA sequence motifs that improves all aspects of Bionano genome mapping (www.bionanogenomics.com).

Before the DLS, DNA was labeled with nicking endonucleases to generate sequence-specific breaks that were then fluorescently labeled and fixed. This method is very accurate and precise, but where nodes on opposite strands are close together, it causes systemic double-strand splits, reducing the contiguity of Bionano maps. The DLS chemistry preserves sample DNA and prevents molecule splits in a systematic manner. A single enzymatic marking reaction is accompanied by cleaning and staining in the DLS protocol. There is no need to patch nicks, resulting in a more effective process. The Bionano de novo genome assemblies have the best contiguity regions that have ever been seen. (Bionano Genomics, n.d.)

The contiguity of maps produced by DLS improves the construct quality of the Bionano genome even further (Formenti et al., 2018). Furthermore, the DLE-1 enzyme (Direct Label Enzyme) is compatible with a wide variety of organisms, allowing for rapid and high-quality genome assembly in almost all research areas. De novo assemblies based on DLS-labeled genomes are among the most contiguous conceivable, typically providing full scaffolds of chromosomal arms. DLS-generated assemblies set the benchmark for accuracy and precision in the system, often outperforming recent reference genome publications. (Bionano Genomics, n.d.)

Hi-C and HiChIP are proximity ligation strategies for capturing three-dimensional (3D) organizational structure and long-range, gene regulatory chromatin interactions, in which genomic sequences that are closer together in three-dimensional space can be spaced linearly apart. As a consequence, prior to sequencing, sequence and structure detail is kept over long distances. Results, once sequenced, can be used for different genomic applications, such as identifying promoter-enhancer interactions for gene regulation studies and Hi-C contig scaffolds for genome assemblies for de novo chromosome recognition. (Mumbach, 2016), (Ando-Kuri, 2018)

In T2T research Arima Genomics (www.arimagenomics.com) used four restriction enzymes to build Hi-C libraries in duplicate. Following the modified chromatin digestion, the digested

ends were numbered, proximally ligated, and then the proximally ligated DNA was purified (Miga et al., 2020).

2.1 hTERT Immortalized Cell Lines

hTERT-immortalized primary cells more precisely resemble cell physiology in vivo. These cells are formed from differentiated cells and display tissue-specific characteristics, express differentiation-specific proteins, and shape structures similar to those seen in vivo. In vitro, primary cells immortalized with hTERT have improved proliferative ability. They exhibit the characteristics of a continuous cell line in terms of progression. Primary cells immortalized with hTERT, unlike primary cells, do not mature after many passages and proceed to proliferate and demonstrate primary cell morphological features. The karyotype and genotype of hTERT-immortalized primary cells are stable, and there are no transformation-related modifications, such as tumorigenicity (Lee et al., 2004). Transfection of hTERT into primary human cells results in chromosome telomere elongation and retention (Bikkul et al., 2018). Forced expression of only hTERT in certain cases helps cells to inhibit replicative senescence and resolve growth crises, effectively contributing to immortality. More than one immortalizing agent may be needed in certain cases to effectively immortalize a specific cell type (ATCC, n.d.).

2.2 Haploid CHM13hTERT Cell Line

T2T scientists have selected the effectively haploid cell line CHM13hTERT for sequencing to avoid the difficulty of assembling all haplotypes of the diploid genome. This cell line was generated by a complete hydatidiform mole and has the karyotype 46, XX (Miga et al., 2020). These uterine moles' genomes are derived from a single sperm that has undergone postmeiotic chromosomal replication and is hence consistently homozygous with one group of alleles. CHM13 has previously been used to correct mistakes and fill in the gaps in human reference records, to test genome assemblers and cause diploid variants, and to look at human segmental duplications. The CHM13 line's karyotyping revealed that it has a healthy 46, XX karyotype with no chromosomal abnormalities. In the early 2000s, CHM13 cells were cultured from one case of complete hydatidiform at Magee-Womens Hospital (Pittsburgh, PA) as part of research. CHM13 cells were cultured, karyotyped with Q-banding, and immortalized with human telomerase reverse transcriptase (hTERT). (Miga et al., 2020)

3.1 Chromosome X - Complete Assembly

The X chromosome contains several characteristics that are peculiar to the human genome. Males inherit a single maternal X chromosome, while females inherit the X chromosome from both parents. The mechanism of X chromosome inactivation (XCI) suppresses gene expression on one of the female X chromosomes in the early stages of development, and this chromosome then remains inactive in somatic tissues. The dormant chromosome is reactivated in the female germ line and undergoes meiotic recombination with the second X chromosome. During meiosis, the male X chromosome cannot recombine its whole length; alternatively, recombination is restricted to small parts at the ends of the X chromosome arms, which recombine with corresponding fragments on the Y chromosome (Ross et al., 2005). For years, the biological effects of sex chromosome evolution have sparked strong interest in the human X chromosome.

Manual assembly has been assigned to Chromosome X by T2T consortium. This is due to its high initial assembly continuity, disproportional involvement in Mendelian diseases (the X chromosome is associated with a relatively large number of Mendelian diseases, and despite the fact that chromosome X comprises just 4% of all human genes, it is related to 10% of Mendelian diseases (Ross et al., 2005)), and unique behavior during its development (Migeon and Kennedy, 1975). De novo assembly was disrupted in three locations: the centromere and two almost identical segmental duplications. Those duplications were manually fixed using ultra long reads that completely covered the repeats and were tightly attached on both sides. Droplet digital PCR (ddPCR) verified that the latest assembly correctly represents CHM13 genome tandem repeats, including the 3.1-Mb centromeric DXZ1 array. For each technology, polishing was done in two rounds of iterative polishing, in such order: Oxford Nanopore, PacBio, and finally Illumina linked reads. (Miga et al., 2020)

3.2 Chromosome X and Its Potential Connection to Telomere Length

Telomeres are chromosomal end regions that are not able to bind to other chromosomes and are responsible for protective function. Telomeres are composed of specialized linear chromosomal DNA with short tandem repeats. Telomeric repeats are conservative - all vertebrate repeats are TTAGGG. Telomeres wear off in somatic cells over time due to the failure of DNA polymerase to synthesise DNA copy from the very end. It can only bind nucleotides to pre-existing 3'-hydroxyl groups. Due to this, DNA polymerase needs a primer to which it can bind the first nucleotide. This is known as the end-replication problem, and it

is thought to be one of the causes of ageing (Rizvi et al., 2014), (Shammas, 2011). Telomere length is determined by a variety of factors, including cellular replication history or telomerase, which consists of the RNA component hTR and the catalytic protein component hTERT. Telomeric repeats may be added to the ends of chromosomes through telomerase (Wong and Collins, 2003). There have been studies that suggest that telomere length can be inherited (Jeanclos et al., 2000), (Slagboom et al., 1994), but the mechanism of inheritance has not been determined.

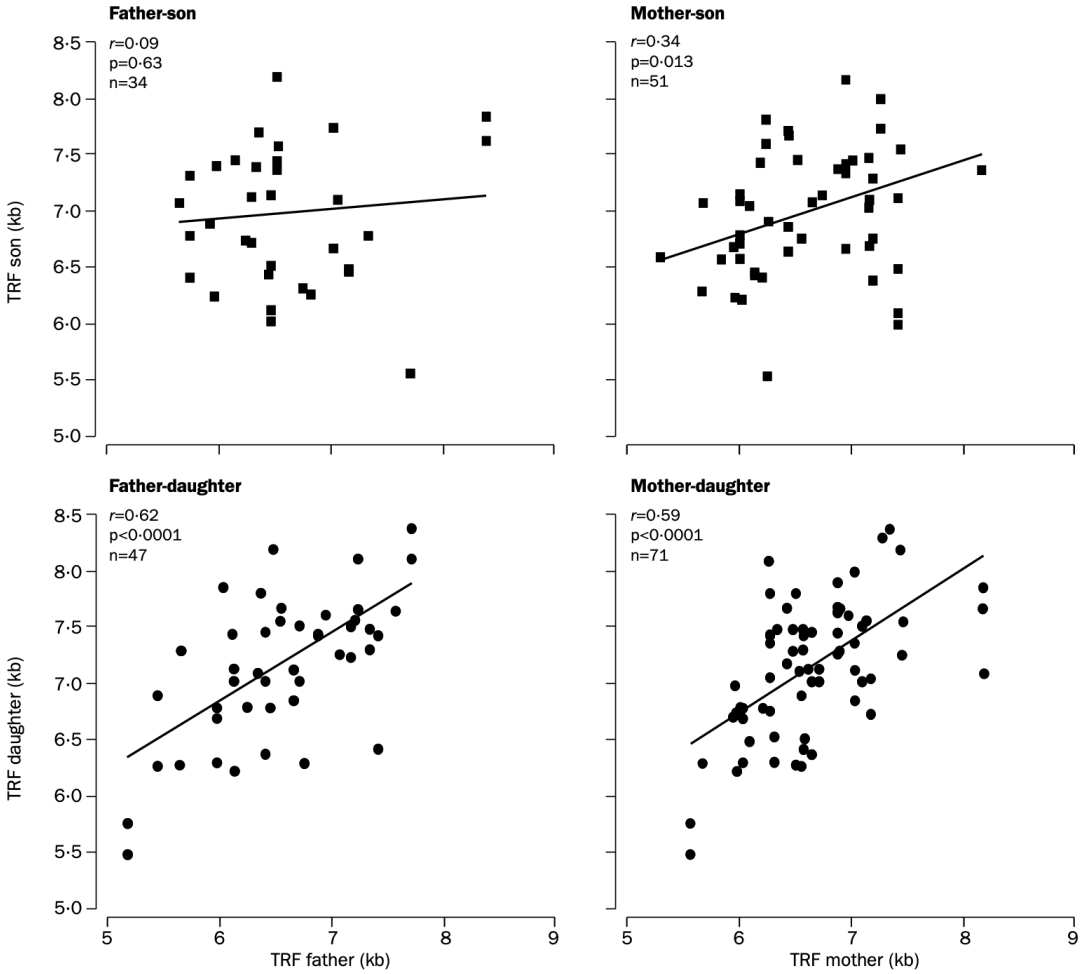

Figure 9. *Intrafamilial relations for TRF length.* Image adapted from: Nawrot, T. S., Staessen, J. A., Gardner, J. P., & Aviv, A. (2004). Telomere length and possible link to X chromosome. *The Lancet*, *363*(9408), 507–510. https://doi.org/10.1016/s0140-6736(04)15535-9

The Nawrot et al. (2004) study, published in The Lancet, found a clear concordance of TRF (terminal restriction fragment) length between mothers and offspring, as well as between fathers and daughters, but not between fathers and sons (Figure 9). This demonstrates the X-linked mechanism of inheritance (Nawrot et al., 2004).

DKC1, a gene on the X chromosome, encodes a protein called dyskerin. This protein is needed for the stable accumulation of the telomerase component hTR (MacNeil et al., 2019), (Mitchell et al., 1999). Patients with inherent X-linked dyskeratosis congenita triggered by a missense mutation in the DKC1 gene have lower hTR concentrations, lower telomerase activity, and shorter TRF length. This suggests that polymorphisms in the DKC1 gene, as well as other unknown genes on the X chromosome, will influence TRF length in the population.

Since the T2T team has completed X chromosome assembly, it is now possible to investigate the relationship between biological ageing and telomere length, as well as genes that regulate telomere length and telomerase activity.

Telomere length can be measured in a variety of ways, with terminal restriction fragmentation serving as the gold standard. Quantitative PCR is more advantageous because it needs less DNA, which is ideal for epidemiological studies involving a large number of participants. There is also a technique that employs fluorescent probes to determine not only mean telomere lengths but also chromosome-specific telomere lengths. However, such a method has a limitation: it can only be used on mitotically active cells. (Montpetit et al., 2014)

3.3 Terminal Restriction Fragments Analysis

The analysis of Terminal Restriction Fragments (TRF) is recognised as a standard approach for determining telomere length. TRF research employs DNA hybridisation technology, in which genome DNA fragments are digested by restriction enzymes and telomere DNA repeats and some subtelomeric DNA are left behind. These are isolated using agarose gel electrophoresis, moved to a filter membrane, and hybridized with oligonucleotide probes labelled with chemiluminescence or radioactivity to visualize telomere restriction fragments. This technique takes a large amount of DNA, but it can measure the telomere length distribution of a population of cells and can be presented in absolute kilo bases. There is an updated TRF method in which genomic DNA is digested by restriction enzymes (that do not cut telomeres) and then electrophoresis in agarose gel is performed. The gel is then dried, and the DNA is denaturated before being hybridized in situ with a radioactively labeled oligonucleotide probe. This in-situ hybridisation prevents telomeric DNA loss and increases signal intensity. The gels are then visualized using phosphor screens, and the length of the telomeres is estimated using a graphic program (Jenkins et al., 2017). The laboratories of Drs. Wooding Wright and Jerry Shay at the University of Texas Southwestern offered such an approach (Herbert et al., 2003), (Mender and Shay, 2015).

3.3.1 TeloTool

Gohring et al. (2013) published a paper with their invention called "TeloTool" - a new software that allows detailed statistical analysis of TRF results. A variety of methodical biases from previously mentioned methods are eliminated using this new technique, including conclusions based on probe intensity corrections. This software offers a consistent method for quickly and reliably extracting quantitative data from TRF autoradiogram. Its widespread use would enable accurate comparison of datasets produced in different laboratories. TeloTool was written in Matlab (Mathworks) and operates on a 64-bit Windows platform, requiring the installation of the MATLAB Compiler Runtime (version 7.17(R2012a)) (Gohring et al., 2013). This program includes following steps: image processing, lane detection, ladder manipulation and fit, statistical analysis, probe correction and background correction.

3.4 Quantitative PCR for Telomere Length Measurement

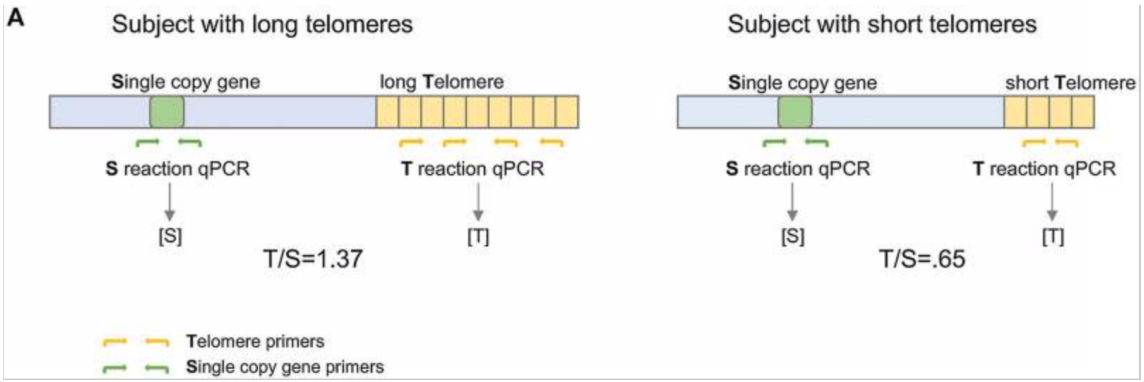

Figure 10. *Schematic diagram of qPCR telomere length measurement as described in Cawthon, 2002.* There are examples of two subjects with high and low T/S ratios, indicating long and short average telomere length. These are arbitrary numbers, but they fall within the usual range of values for normal, healthy adults where a pooled human genomic DNA is used as the reference standard. Image adapted from: Lin et al. (2018)

One more method exists to measure telomere length. Richard Cawthon explained qPCR telomere length measurement in his 2002 paper (Cawthon, 2002). The telomere signal abundance per genome determined by qPCR reflects the average telomere length in a given DNA sample, according to the method's general concept. The sum of input genomic DNA is quantified by calculating the qPCR product of a single copy gene, which is then used to normalize the telomere reaction signal. The T/S ratio that results reflects the average telomere length per chromosome (Figure 10). Cawthon identified a monochrome multiplex qPCR

(mmQPCR) method in a subsequent paper that simultaneously tests the signals of telomeres and a single copy gene in a one-tube reaction (Cawthon, 2009).

4.1 Chromosome 8 - Complete Assembly

Chromosome 8 is one of the 23 pairs of chromosomes in humans. Chromosome 8 covers about 145 million base pairs (the building material of DNA) and accounts for between 4.5 and 5.0 percent of total DNA in cells (Tabarés-Seisdedos & Rubenstein, 2009, p. 56). A large area of 15 megabases on distal 8p seems to have a significantly higher mutation rate, which appears to have increased in homo sapiens compared to other sequenced mammals. This rapidly developing area includes a variety of genes linked to innate immunity and the nervous system, as well as loci that seem to be under positive selection, such as the large defensin (DEF) gene cluster and MCPH, a gene that may have led to the development of larger brain size in great apes. The information from chromosome 8 should help researchers better understand normal and disease genetics, as well as genome evolution (Nusbaum et al., 2006).

Unlike the human X chromosome assembly, T2T used both ultra-long Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) HiFi data to eliminate breaks in human chromosome 8. They began by generating 20-fold sequence coverage of ultra-long ONT data and 32.4-fold sequence coverage of PacBio HiFi data from a complete hydatidiform mole (CHM13). Then, from CHM13 PacBio HiFi results, complex regions in chromosome 8 were constructed by generating a library of singly specific nucleotide k-mers (SUNKs) (Sudmant et al., 2010), or sequences of length k that occur approximately once per haploid genome. The SUNKs were also validated using Illumina data from the same genome and used to barcode ultra-long ONT reads. The researchers assembled ultra-long ONT reads with extremely identical barcodes into an initial sequence scaffold that traverses each chromosome 8 gap. T2T researchers enhanced the sequence scaffolds' base-pair accuracy by replacing the raw ONT sequence with concordant PacBio HiFi contigs and inserting them into a previously created linear assembly of human chromosome 8. (Logsdon et al., 2021)

According to Logsdon et al. (2021) the total telomere-to-telomere sequence of the human chromosome 8 is 146,259,671 bases long and contains 3,334,256 bases that were absent from the existing reference genome (GRCh38). The majority of the adds are located on different chromosome regions: a 644-kb copy number polymorphic β-defensin gene cluster that maps to chromosome 8p23, the entire centromere equivalent to 2.08 Mb of α-satellite HORs (higher order repeats), an 863-kb 8q21.2 variable number tandem repeat (VNTR), and all telomeric

regions that end with the typical TTAGGG sequence. The T2T consortium validated the assembly using optical maps (Bionano Genomics), single-cell DNA template strand sequencing (Falconer and Lansdorp, 2013), and references to completed BAC sequences as well as Illumina whole-genome sequencing data obtained from the same source genome. Unlike the human reference genome, which has two segmental duplications, CHM13 (newly sequenced genome) has three: a 544-kb segmental duplication on the distal end and two 693- and 644-kb segmental duplications on the proximal end. Each segmental duplication region contains at least five β-defensin genes, and as a result, researchers have found five additional β-defensin genes that are nearly identical to the reference at the amino acid level. (Logsdon et al., 2021)

4.2 Actions in Perspective After Chromosome 8 Assembly

Chromosome 8 has some distinguishing characteristics. Despite covering just around 5% of the overall genome, it contains a vast array of genes and pseudogenes that are responsible for brain function and the formation of cancer tumors. Crohn's disease is particularly linked to chromosome 8, since this chromosome contains the beta-defensin gene cluster, and incomplete expression of this gene causes Crohn's disease (Fellermann et al., 2006).

The T2T consortium selected chromosome 8 because it contains a small centromere as well as a polymorphic neocentromere (Logsdon et al., 2021). However, the current finished assembly allows researchers to not only investigate centromere-related diseases, but also to scan the entire sequence for the occurrence of Alu elements. Repeated sequences account for more than 55% of the human genome (Tang et al., 2018), (Hubley et al., 2016). These repetitive sequences may be classified as tandem repeats or interspersed repeat elements (segmental duplications and transposable elements). The Alu element is the most common transposable element of the human genome. A standard Alu element is a 300 bp long transposable nucleotide chain (Schmid and Jelinek, 1982). Alu repeats can play a role in cancer growth. The mutational theory of tumorigenesis outlines the causes of the transformation of a normal cell into a cancerous cell by the occurrence of somatic mutations, which cancer cells pass on to all future generations (Moolgavkar and Knudson, 1981). The function of Alu repeats as one of the most hypervariable loci of the human genome becomes particularly clear in the mechanism of somatic mutations during carcinogenesis (Mitchell et al., 1991), (Muratani et al., 1991). This is also supported by the fact that the absence of allelic heterozygosity in tumors is often observed by Alu-PCR. As a result, it is difficult to deny that Alu repeats have taken on the role of genetic markers of the mutation mechanism involved in the formation of

cancer. The mutator hypothesis is based on the idea that there is a specific gene or genes that are extremely vulnerable to repair system failures. Genes containing unstable microsatellite DNA may be the first candidates for this position. The state of their relative stability is controlled by the main protein of hMSH2 reparation. A mutated and normal allele of this gene coexist in heterozygous cells. Simultaneously, the cells maintain a normal phenotype, and this process persists until the normal allele is no longer normal as a result of a newly emerging mutation. Tumor cells, in reality, only carry mutant alleles. The disappearance of heterozygosity at the locus encoding the hMSH2 protein increases the mutability of not only oncosuppressor genes, but also other genes, and this is manifested most prominently in mutations of defective genetic structures, such as DNA repeats. (Orth et al., 1994), (Lukash et al., 1996).

4.3 Alignment-free Method for Detecting Alu Elements

A group of Tartu University scientists developed a great quick alignment-free approach for studying polymorphic insertions of Alu elements in the human genome. It consists of two pipelines for detecting novel polymorphic insertions in raw sequencing reads. One pipeline looks for Alu elements that are present in a given organism but not in the reference genome (REF– elements), while the other looks for possible polymorphic Alu elements that are present in the reference genome but not in all individuals (REF+ elements). The REF– discovery pipeline takes about 2 hours per person to complete, while the REF+ discovery pipeline takes about 20 minutes. Depending on the hardware setup, genotyping possible polymorphic insertions takes between 0.4 and 4 hours per person. Furthermore, they suggest an alignment-free genotyping approach that calls the genotype directly from raw sequencing reads based on the frequency of insertion or deletion-specific 32-mer pairs. (Puurand et al., 2019)

5.1 VNTR and HOR

Talking about repetitions in the genome, it is worth mentioning: Variable Number Tandem Repeats (VNTR) and Higher Order Repeats (HOR). Those are also associated with various health conditions, just as Alu elements. For VNTR a “classic” example would be association with type I diabetes. Insulin gene *INS* contains a VNTR which is 14bp long and varies from 26 to 200 repeats. Variation in this VNTR has been associated with expression of the *INS* gene and risk for type 1 diabetes (Durinovic-Belló et al., 2010). HOR structures are tandem arrays of larger repeat units which consist of multiple basic repeat units. Due to the fact that

those regions include large numbers or repetitions, they are susceptible to mutations, epigenetic changes and modifications in gene expression patterns with a potential link to human disease (Rich et al., 2014). Thus, long tandem repeats study will be useful for working on a treatment of satellite-related diseases, such as FSHD , for example. In this thesis it will be suggested to use adVNTR-NN tool for VNTR calling and genotyping and Alpha-CENTAURI tool for estimating HOR structure, which is created to deliver feedback on the organisation of local repeats for each read and allows to monitor repeat units rearrangements and the position of the array's transition to non-satellite DNA (Sevim et al., 2016).

THE AIMS OF THE THESIS

- To emphasize the importance of the newly sequenced human genome and give an overview of the process of sequencing in general.
- To give an overview of CHM13hTERT cell line features, used in the T2T study and discuss possible outcomes and contribution of T2T study for future science.
- To suggest the possibility of studying newly sequenced human chromosomes X and 8.
- To overview methods that scientists used to improve the previously sequenced genome.

EXPERIMENTAL PART

As an experimental part, a plan for using different bioinformatic tools is proposed that can be used on the data obtained after the complete assembly of chromosomes X and 8.

1. Data Sources

	GenBank	*File format*	*Description*	*File size*	*Citation*
Chromosome X	CP068255.2	FASTA	Homo sapiens isolate CHM13 chromosome X 154259625 bp	156.5 MB	Miga et al. (2020)
Chromosome 8	CP068270.1	FASTA	Homo sapiens isolate CHM13 chromosome 8 146259322 bp	148.3 MB	Logsdon et al. (2021)

Table 2. *Representation of data sources.* Data is adapted from www.ncbi.nlm.nih.gov

2. Tools for Analysing Chromosome Sequences

There are many different bioinformatics tools that can be used to carry out various sequence studies *in silico*. This chapter presents only a few of them that can be used with the participation of the obtained data on the sequenced chromosomal sequences. Here the emphasis is on which discoveries were presented from the T2T consortium. The first is the complete assembly of telomeres and the TeloTool program, which, although it requires prior TRF analysis, is able to determine the telomere length from the pictures obtained from PhosphoImager. Then, Alpha-CENTAURI, a tool that helps to evaluate the structure of higher order repeats. This is possible using long reads data in conjunction with the archive of already known satellite monomers. This is a useful program for working with chromosome data, where centromeres are also completely reconstructed. adVNTR-NN is a program for detecting VNTRs, which can detect mutations as well. Eventually, AluMine - alignment-free method to discover Alu elements.

2.1 TeloTool for Telomere Length

TRF analysis should be conducted prior to using TeloTool.

TRF analysis usually begins with DNA isolation, followed by DNA digestion (2.5 g DNA for 4 h to overnight at 37 °C with a mixture of restriction enzymes (HhaI, HinF1, MspI, HaeIII, RsaI, AluI) and Gel migration, in which a radiolabeled TRF ladder and an unlabeled

molecular weight marker are loaded on one or both sides of the samples. The DNA is then hybridized, after hybridization with a telomere sequence-specific probe, radiolabeled TRF markers can be visualized. Gel Red can be used to image unlabeled, digested plasmid DNA, but not the telomere sequence specific probe. Then a hot probe is prepared, the gel is washed, and the PhosphoImager is used to scan it. These steps are described in accordance with Mender and Shay's (2015) protocol. As stated earlier, TRF analysis results can be processed using TeloTool.

TeloTool is only available for use on 64-bit Windows platforms. The 64bit MATLAB Compiler Runtime (MCR) (version 7.17) should be downloaded and installed (R2012a). The software itself is available for free from mathworks.com. (Gohring et al., 2013)

2.2 Alpha-CENTAURI

Alpha-CENTAURI (ALPHA satellite CENTromeric AUtomated Repeat Identification) is a bioinformatic tool for assessing local HOR structure within long read whole genome sequencing datasets. Alpha-CENTAURI can forecast HOR patterns by clusters of discrete repeat units. This is possible if providing long-read data and a training dataset of repeat units from a satellite family of interest. As a result, the user is given the HOR structure classification: defined either as *regular* (the one that contains the same ordering of near-identical monomers) or *irregular* (the one that contains a rearrangement, inversion or discontinuous spacing in monomers when compared to other HORs on the same read). Alpha-CENTAURI accepts two input files: a FASTA file with long reads and an Hidden Markov Models (HMM) archive constructed with proven alpha-satellite monomers. (Sevim et al., 2016)

Documentation and source code for Alpha-CENTAURI are available at github.com.

2.3 adVNTR-NN for VNTR Detection

adVNTR-NN is a program for detecting Variable Number Tandem Repeats (VNTR) in sequence data. It detects diploid repeating counts for VNTRs and recognizes potential mutations in the VNTR sequences using both NGS short reads (Illumina HiSeq) and SMRT reads (PacBio). Shallow neural networks are used for quick read recruiting, followed by sensitive HMM for genotyping. VNTRs are known as "hidden polymorphisms." Despite high mutation rates and reported examples of function variations, VNTR genotyping is not used in Mendelian or GWAS pipelines. This is largely due to technological difficulties. To reliably

identify VNTR genotypes, the software employs a combination of quick filtering followed and HMM-based genotyping. (Bakhtiari et al., 2021)
adVNTR-NN is also available at github.com.

2.4 Alu Elements with AluMine

These discovery methods can be implemented to any whole-genome sequencing data to look for novel Alu insertions. The AluMine method enables the detection of new Alu elements from tumor sequences, enabling researchers to investigate the somatic insertion of Alu elements in cancer cells and their function in carcinogenesis. The pre-compilation of 32-mer pairs enables the genotyping of known Alu element insertions in high-coverage sequencing results. It is suggested to use the software to find Alu repeats in completely sequenced chromosomes since it operates with sequencing data. It is worth noting that the X chromosome was historically recognized as the location where the Alu family of short interspersed nuclear elements (SINEs) is below average, which is consistent with the chromosome's gene-poor composition (Ross et al., 2009). Alu components from the X chromosome, on the other hand, may be used as genetic markers in human population genetics (Athanasiadis et al., 2007). It is therefore fair to propose using a tool for detecting Alu repeats primarily in the chromosome 8 sequence. Since this chromosome is closely linked to the development of different forms of cancer, and the function of Alu elements is also linked to tumor formation, as previously highlighted. It has already been mentioned that AluMine can detect repeats that exist in the reference genome but not in the individual's genome, and vice versa. The complete chromosome assembly also enables them to be used as a new reference genome.

Below is a description of the use of scripts (the steps are taken in accordance with Puurand et al. (2019) protocol available on github.com).

2.4.1 REF-plus Discovery Scripts

Before you start working, you must first download the local server by entering "git clone https://github.com/bioinfo-ut/AluMine". The first step in REF-plus discovery is to use a script to look for all possible Alu insertions. This script looks for Alu repeat signatures which are 10bp long and Target Site Duplication sequences (270-350bp) in the reference genome. Then, using BLAST, determine if the detected candidate elements are homologous to known existing Alu elements. Following that, using 25-mer lists, search within the chimp genome. This process gets rid of older elements that are probably fixed in both organisms. To execute

the scripts manually, open pipeline_ref_plus.sh in a text editor and specify the paths to the reference and chimp genome in FASTA format.

2.4.2 REF-minus Discovery Scripts

The following files are essential components of the REF-minus discovery procedure: find_ref_minus_candidates_bam.pl, gtester, and ref_minus_post_gtester.pl.

The scripts find_ref_minus_candidates_bam.pl and find_ref_minus_candidates_fastq.pl look for 10bp long Alu sequences in BAM or FASTQ files and output all available variants with a 25bp flanking sequence. The 25bp flanking sequences are then found in the reference genome using 'gtester.' The ref_minus_post_gtester.pl eliminates set Alu repeats (present within the human reference genome) and gives a pair of 32-mers for those which are not present in the reference genome. Prior to launching the REF-minus discovery pipeline, it is necessary to open the file pipeline_ref_minus.sh via text editing program and specify the pathways to sample files and human chromosome files. Formats for input could be BAM, single FASTQ, multiple FASTQ and gnuzipped FASTQ.

DISCUSSION

The primary goal of the thesis is to concentrate on a breakthrough that fundamentally alters all already learned about the human genome. The study's aim is also to demonstrate technology evolution and the advantages it brings to mankind. Since the introduction to this thesis started with a general understanding of DNA sequencing, it is worth noting that this technique has many advantages for people. In personalized medicine, for example, the idea of genome sequencing comes first. And it doesn't stop there. After all, the evidence collected from sequencing allows for the development of more reliable and modern therapeutic approaches, as well as the prescription of drugs that would undoubtedly be effective. The simplest example is the introduction of a new treatment for breast cancer (Sheikh-Hosseini et al., 2021). In short, it includes the ability to "work" specifically with the gene or genes that are the

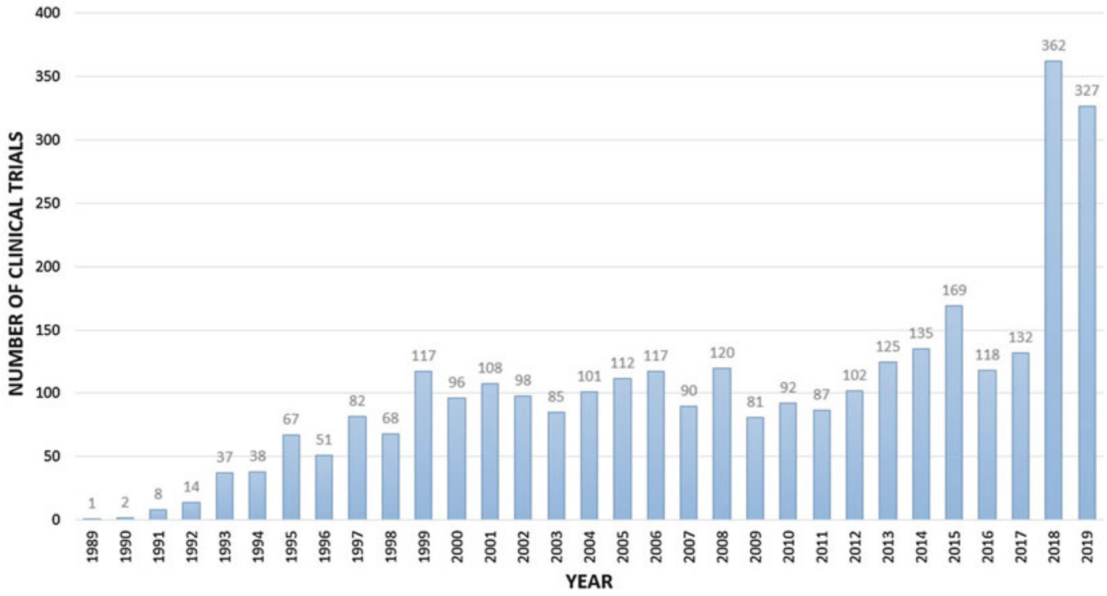

Figure 11. *From 1989 to 2019, the number of gene therapy clinical trials per year increased*. Image adapted from: Arjman et al. (2019)

source of the mutation leading to breast cancer. There are many strategies for such cases: gene replacement, antisense technique, RNA interference therapy, growth receptors method, suicide gene technique etc. We will not go into detail about these approaches here, but it is important to understand that, unlike so-called conservative therapies such as chemotherapy or radiation, the genetic approach is preferable in that it operates exclusively on the root of the issue and does not expose the rest of the body to harm. Despite the fact that gene therapy is

not able to cure absolutely all diseases today, the number of gene therapy clinical trials per year has grown significantly from 1989 to 2019 (Figure 11).

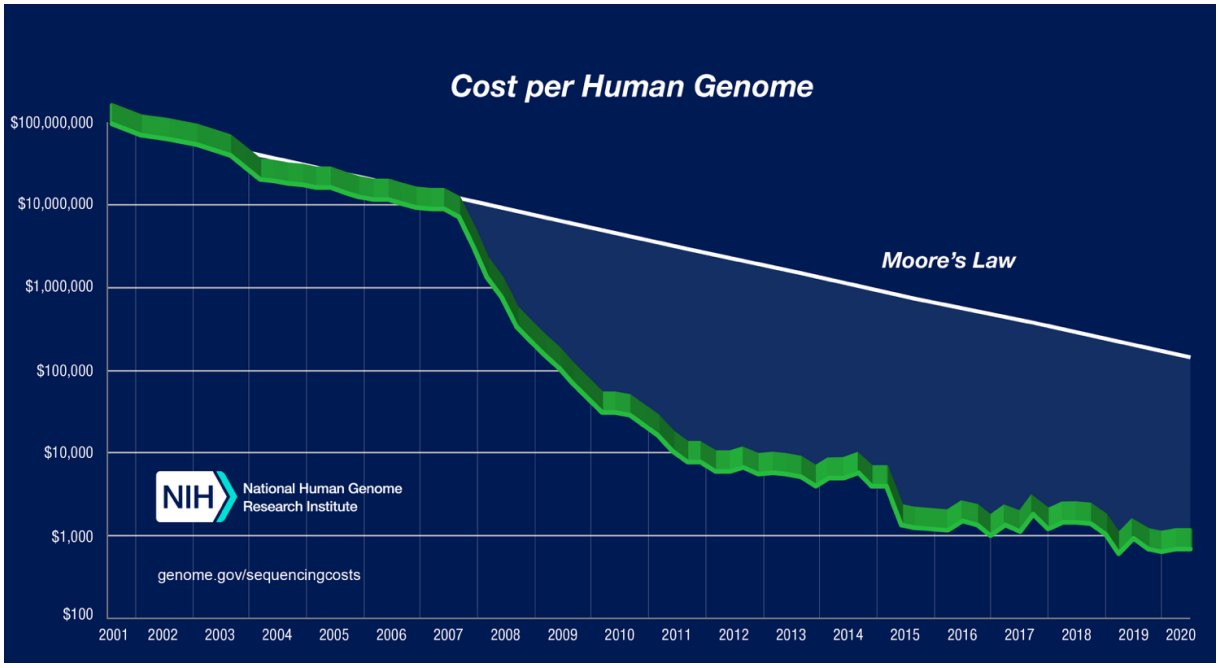

Figure 12. *The cost per Human Genome*. The graph includes hypothetical data based on Moore's Law, which describes a long-term trend in the computer hardware industry involving the doubling of 'compute power' every two years. Image adapted from: www.genome.gov

Of course, it is also worth noting that the procedures associated with genome sequencing are quite financially costly (Schwarze et al., 2020), and there is also an ethical side to the issue. For instance, individual identifiability, adequacy of consent, stereotyping and stigmatization (Foster and Sharp, 2006). However, these problems are not insoluble; on the contrary, they stimulate further work with DNA sequencing. For example, talking about costs, the evolution from first generation sequencing to third generation sequencing has made the sequencing process faster and cheaper (Figure 12).

DNA sequencing methods have made their way into the field of forensic science. To detect length difference in short tandem repeat (STR) markers, the majority of forensic DNA studies currently use PCR and capillary electrophoresis (CE)-based fragment analysis methods. Scientists around the world are investigating the utility of NGS technology for forensic studies, since there are drawbacks to CE-based research, such as the inability to analyze multiple genetic polymorphisms in a single reaction using a single workflow, low-resolution genotyping of current markers, loss of valuable genomic knowledge from damaged DNA samples, and low-resolution mtDNA and mixture analysis. Of course, DNA sequencing in

forensic science also faces some difficulties during research, because often the DNA sample is insufficient or the sample is contaminated (Yang et al., 2014).

The Human Genome Project (HGP) has made significant contributions to the development of sequencing technology. The project, for example, was crucial in the development and eventual widespread use of second generation sequencing technologies, which started in the mid-2000s (Kulski, 2016). The need for the creation and advancement of new technologies gradually arose during the first genome sequencing project, demonstrating the significance and positive impact of the concept of sequencing the entire human genome. HGP has made a huge contribution and has had a significant impact on various fields of science and life. The results of the project helped to create and classify the ‘parts list’ of most human genes (Venter et al., 2001). In turn, the emergence of such a list contributed to the development of a science called "systems biology" (Ideker et al., 2001). HGP played an important role in the development of proteomics (the science of proteins and their functions). It also aided mass spectrometry work by supplying reference sequences. The project also shed light on evolution, making it possible to assess how much and whether different organisms are generally related to each other. And an ability to see the proof of the existence of a common ancestor is now possible (Theobald, 2010). One of the problems with genome sequencing is the interpretation of the data obtained, which was much more difficult during HGP than it is today. Accordingly, another contribution of the project was to unite computer scientists, mathematicians, engineers and theoretical physicists with biologists for the development of advanced statistical and mathematical data methods. And, as stated earlier, DNA sequencing is of great importance in medicine, and thanks to HGP, a reference model of the human genome has emerged and the possibility and use of advanced sequencing technologies. It is also very important that thanks to such a great discovery, we now know that we do not have any genes that determine our race. We are all equal, and this is excellent evidence against racial discrimination (Foster and Sharp, 2004; Royal and Dunston, 2004).

However, HGP in its results did not have a "perfect" sequence, there were still gaps, 5% in the human genome were not sequenced due to technical problems (Hood and Rowen, 2013). The largest gaps were located at centromeres and telomeres due to the fact that those regions have a big number of repeating sequences. And here, accordingly, the T2T consortium comes into play. The most important goal of their work was to finally close the gaps of the previous assembly of the human genome. In this they were helped by the combination of several sequencing technologies, which have already been mentioned in the literature review.

Particular contributions were made by the developments of PacBio, Oxford Nanopore and PacBio HiFi. PacBio showed that they can be used to sequence the human genome based on long-reads (using a cell line from the hydatidiform mole, which later became the reference genome). Then Oxford Nanopore was also confirmed for use in genome sequencing. Finally, PacBio HiFi has emerged which provides 99.8% accuracy when sequencing the HG002 human genome reference standard at an average read length of 13.5 kb (Wrighton, 2021). Special attention should be paid to the fact that the use of the CHM13hTERT cell line greatly simplified the work due to its homozygosity between chromosomes. And so, the T2T consortium has assembled a v1.0 assembly of a complete human genome, and what exactly is important for this thesis - fully sequenced chromosomes 8 and X.

In the previous version of the genome, as we already know, there were many gaps due to repetitive sequences. Since they were not in the database, these repeat-rich sequences were often omitted from genetics and genomics research, limiting the spectrum of interaction and functional analysis. This means that with the updated version of the genome, the doors are opening for us to study, for example, diseases associated with centromeres, diseases associated with telomeres, diseases associated with different repeating regions. It is now possible to better understand the function of chromosomes in general, to study genomic variation and finally to improve the technologies in biomedicine that use short-read mapping to a reference genome. If we talk about chromosomes X and 8 specifically, then the complete assembly of each of them also makes an undeniable contribution. So, for example, based on this thesis, early studies argued that on the X chromosome there are various unknown genes that affect the TRF length in the population. Accordingly, having a complete sequence of the X chromosome, it is possible to discover new genes and correlation between TRF length and those genes can be estimated. Moreover, the presented X-linked inheritance of telomere length and the new complete version of the chromosome also allow for general statistical studies, revealing certain patterns. TeloTool is present in this thesis as the only one bioinformatic tool which requires prior “wet” lab experiments, but it is very beneficial in correctly estimating telomere length. As we now know the complete sequence of telomeres in the reference genome, by measuring telomere lengths in populations, it is possible to conduct comparative studies.

If we consider the contribution of the assembly of chromosome 8, then it is worth noting that during this assembly, five additional beta-defensin genes were discovered that could contribute to a better understanding of Crohn's disease. Since it was already mentioned that on chromosome 8 there is a beta-defensin gene cluster, which is directly related to the

development of Crohn's disease. Chromosome 8 itself is in general closely related to the development of cancerous tumors, which makes the proposal for the study of Alu repeats even more important. Indeed, according to the information given, these repeating elements also play a role in the development of cancer. Alu repeats are inherently highly polymorphic (Epstein et al., 1990), so it makes them “good candidates” for susceptibility to mutational variability. AluMine tool could look for Alu insertions in the new reference genome and identify which ones are novel, unknown til this moment. And then, vice versa, new data can become a new reference for testing different individuals.

Further use of the obtained data is based on the completed centromeres. Centromeres in many complex organisms include long near identical arrays of tandem repeats. Those can be either organized into multi-monomeric repeating units or HOR. HORs are very variable in terms of repeat number or repeat structure. Having a large number of repeat units, they are susceptible for mutations as well. Alpha-CENTAURI is helpful in analyzing long sequencing reads for de novo repeat discovery. Then the tool could use FASTA sequences from the table as a reference.

VNTR is a region of DNA sequence where a short sequence of nucleotides is present as a tandem repeat, its length differs between different people. They are often found in the coding regions. There is a strong correlation between VNTR and Mendelian diseases. Diseases happen due to point mutation in the repeat unit or change in the number of repeat units. For instance, bipolar disorder is caused when a 30 bp repeat unit in *MAOA* gene (chromosome X) is repeated 4 times instead of 2-5 (Byrd et al., 2018). New data from completed chromosomes can contribute into studying other disease-linked VNTRs by revealing previously unknown variants. The obtained data about chromosomes X and 8 could be used as a new reference.

SUMMARY

DNA sequencing is the determination of the sequence of nucleotides in a DNA chain. There has been a long way from the earliest attempts at sequencing to the already powerful enough techniques in use today. DNA sequencing brought together scientists from various scientific fields in the first complete assembly of the human genome, in a peculiar way "obliging" them to develop new discoveries. Largely because of this, the Telomere-to-Telomere consortium were able to fill in the gaps of the previous human genome. This was done using a combination of different technologies: Oxford Nanopore, PacBio CLR, PacBio HiFi, 10x Genomics, BioNano DLS and Arima Genomics HiC. Also contributing was the CHM13hTERT cell line, obtained from a hydatidiform mole and immortalized. It is unique due to its haploid nature and, thus, greatly facilitated assembly. The new version of the genome assembled by T2T opens the door for us to even greater discoveries, such as understanding diseases and establishing evolutionary processes. It is especially important that the consortium also completely completed the assembly of chromosomes X and 8. The gaps were in places where there are a lot of repeats. And, as you know, such places are especially closely associated with the development of various diseases due to higher mutation rates. There are a number of bioinformatics devices for various types of analysis, and here it was proposed to use TeloTool, Alpha-CENTAURI, adVNTR-NN, and AluMine to study telomere length, assess the HOR structure, determine VNTR and find new Alu repeats, respectively.

REFERENCES

Ando-Kuri, M., Rivera, I. S. M., Rowley, M. J., & Corces, V. G. (2018). Analysis of Chromatin Interactions Mediated by Specific Architectural Proteins in Drosophila Cells. *Methods in Molecular Biology*, 239–256. https://doi.org/10.1007/978-1-4939-7768-0_14

Ansorge, W. J., Katsila, T., & Patrinos, G. P. (2016). Perspectives for Future DNA Sequencing Techniques and Applications [E-book]. In G. P. Patrinos (Ed.), *Molecular Diagnostics* (3rd ed., pp. 141–153). Academic Press. https://doi.org/10.1016/B978-0-12-802971-8.00008-0

Arjmand, B., Larijani, B., Sheikh Hosseini, M., Payab, M., Gilany, K., Goodarzi, P., Parhizkar Roudsari, P., Amanollahi Baharvand, M., & Hoseini Mohammadi, N. S. (2019). The Horizon of Gene Therapy in Modern Medicine: Advances and Challenges. *Advances in Experimental Medicine and Biology*, 33–64. https://doi.org/10.1007/5584_2019_463

Arjmand, B., Larijani, B., Sheikh Hosseini, M., Payab, M., Gilany, K., Goodarzi, P., Roudsari, P. P., Baharvand, M. A., & Mohammadi, N. H. (2019). From 1989 to 2019, the number of gene therapy clinical trials per year increased [Graph]. In *The Horizon of Gene Therapy in Modern Medicine: Advances and Challenges*.

Athanasiadis, G., Esteban, E., Via, M., Dugoujon, J. M., Moschonas, N., Chaabani, H., & Moral, P. (2007). The X chromosome Alu insertions as a tool for human population genetics: data from European and African human groups. *European Journal of Human Genetics*, *15*(5), 578–583. https://doi.org/10.1038/sj.ejhg.5201797

Barzon, L., Lavezzo, E., Militello, V., Toppo, S., & Palù, G. (2011). Applications of Next-Generation Sequencing Technologies to Diagnostic Virology. *International Journal of Molecular Sciences*, *12*(11), 7861–7862. https://doi.org/10.3390/ijms12117861

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., . . . Smith, A. J. (2008). Accurate whole human genome sequencing

using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Bharagava, R. N., Purchase, D., Saxena, G., & Mulla, S. I. (2019). Applications of Metagenomics in Microbial Bioremediation of Pollutants [E-book]. In S. Das & H. R. Dash (Eds.), *Microbial Diversity in the Genomic Era* (1st ed., pp. 469–470). Academic Press. https://doi.org/10.1016/B978-0-12-814849-5.00026-5

Bowers, J., Mitchell, J., Beer, E., Buzby, P. R., Causey, M., Efcavitch, J. W., Jarosz, M., Krzymanska-Olejnik, E., Kung, L., Lipson, D., Lowman, G. M., Marappan, S., McInerney, P., Platt, A., Roy, A., Siddiqi, S. M., Steinmann, K., & Thompson, J. F. (2009). Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods*, *6*(8), 593–595. https://doi.org/10.1038/nmeth.1354

Braslavsky, I., Hebert, B., Kartalov, E., & Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences*, *100*(7), 3960–3964. https://doi.org/10.1073/pnas.0230489100

Bucher, P. (2019a). Illustration of Illumina sequencing by synthesis method [Illustration]. In *Metagenomics Profiling of Tumours Using 16S-rRNA Amplicon Based Next Generation Sequencing* (p. 8).

Bucher, P. (2019b, March). *Metagenomics Profiling of Tumours Using 16S-rRNA Amplicon Based Next Generation Sequencing*. https://www.researchgate.net/publication/332752369_Metagenomics_Profiling_of_Tumours_Using_16S-rRNA_Amplicon_Based_Next_Generation_Sequencing

Byrd, A. L., Manuck, S. B., Hawes, S. W., Vebares, T. J., Nimgaonkar, V., Chowdari, K. V., Hipwell, A. E., Keenan, K., & Stepp, S. D. (2018). The interaction between monoamine oxidase A (MAOA) and childhood maltreatment as a predictor of personality pathology in females: Emotional reactivity as a potential mediating mechanism. *Development and Psychopathology*, *31*(1), 361–377. https://doi.org/10.1017/s0954579417001900

Castagnola, M., Cassiano, L., De Cristofaro, R., Landolfi, R., Rossetti, D. V., & Bettolo, G. B. M. (1988). High-performance liquid chromatography in protein sequence determinations. *Journal of Chromatography A*, *440*, 231–251. https://doi.org/10.1016/s0021-9673(00)94527-8

Cawthon, R. M. (2002). Telomere measurement by quantitative PCR. *Nucleic Acids Research*, *30*(10), 47e–447. https://doi.org/10.1093/nar/30.10.e47

Cawthon, R. M. (2009). Telomere length measurement by a novel monochrome multiplex quantitative PCR method. *Nucleic Acids Research*, *37*(3), e21. https://doi.org/10.1093/nar/gkn1027

Chidgeavadze, Z., Beabealashvilli, R., Atrazhev, A., Kukhanova, M., Azhayev, A., & Krayevsky, A. (1984). 2′, 3′-Dideoxy-3′ aminonucleoside 5′- triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. *Nucleic Acids Research*, *12*(3), 1671–1686. https://doi.org/10.1093/nar/12.3.1671

Coon, J. J., Syka, J. E., Shabanowitz, J., & Hunt, D. F. (2005). Tandem Mass Spectrometry for Peptide and Protein Sequence Analysis. *BioTechniques*, *38*(4), 519–523. https://doi.org/10.2144/05384te01

Dewey, F. E., Pan, S., Wheeler, M. T., Quake, S. R., & Ashley, E. A. (2012). DNA Sequencing. *Circulation*, *125*(7), 931–944. https://doi.org/10.1161/circulationaha.110.972828

Doucet, A., & Overall, C. M. (2011). Amino-Terminal Oriented Mass Spectrometry of Substrates (ATOMS). *Methods in Enzymology*, *501*, 275–293. https://doi.org/10.1016/b978-0-12-385950-1.00013-4

Drmanac, R. (2002). Probe overlaps and assembly branches [Illustration]. In *Advances in Biochemical Engineering/Biotechnology* (77th ed., pp. 75–101).

Drmanac, R., Drmanac, S., Chui, G., Diaz, R., Hou, A., Jin, H., Jin, P., Kwon, S., Lacy, S., Moeur, B., Shafto, J., Swanson, D., Ukrainczyk, T., Xu, C., & Little, D. (2002). Sequencing by Hybridization (SBH): Advantages, Achievements, and Opportunities. *Advances in Biochemical Engineering/Biotechnology*, 75–101. https://doi.org/10.1007/3-540-45713-5_5

Durinovic-Belló, I., Wu, R. P., Gersuk, V. H., Sanda, S., Shilling, H. G., & Nepom, G. T. (2010). Insulin gene VNTR genotype associates with frequency and phenotype of the autoimmune response to proinsulin. *Genes & Immunity*, *11*(2), 188–193. https://doi.org/10.1038/gene.2009.108

Edman, P., Högfeldt, E., Sillén, L. G., & Kinell, P. O. (1950). Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chemica Scandinavica*, *4*, 283–293. https://doi.org/10.3891/acta.chem.scand.04-0283

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., & Rank, D. (2009, January). Sequencing via light pulses [Illustration]. In *Science* (p. 134).

Epstein, N., Nahor, O., & Silver, J. (1990). The 3’ ends of alu repeats are highly polymorphic. *Nucleic Acids Research*, *18*(15), 4634. https://doi.org/10.1093/nar/18.15.4634

Erlich, H. A. (2013). Development and Evolution of PCR. *Genetic Engineering & Biotechnology News*, *33*(7), 32–33, 45. https://doi.org/10.1089/gen.33.7.14

Falconer, E., & Lansdorp, P. M. (2013). Strand-seq: A unifying tool for studies of chromosome segregation. *Seminars in Cell & Developmental Biology*, *24*(8–9), 643–652. https://doi.org/10.1016/j.semcdb.2013.04.005

Fellermann, K., Stange, D. E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C. L., Reinisch, W., Teml, A., Schwab, M., Lichter, P., Radlwimmer, B., & Stange, E. F. (2006). A Chromosome 8 Gene-Cluster Polymorphism with Low Human Beta-Defensin 2 Gene Copy Number Predisposes to Crohn Disease of the Colon. *The American Journal of Human Genetics*, *79*(3), 439–448. https://doi.org/10.1086/505915

Formenti, G., Chiara, M., Poveda, L., Francoijs, K. J., Bonisoli-Alquati, A., Canova, L., Gianfranceschi, L., Horner, D. S., & Saino, N. (2018). SMRT long reads and Direct Label and Stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow (Hirundo rustica rustica). *GigaScience*, *8*(1). https://doi.org/10.1093/gigascience/giy142

Foster, M. W. (2006). Ethical issues in medical-sequencing research: implications of genotype-phenotype studies for individuals and populations. *Human Molecular Genetics*, *15*(90001), R45–R49. https://doi.org/10.1093/hmg/ddl049

Foster, M. W., & Sharp, R. R. (2004). Beyond race: towards a whole-genome perspective on human populations and genetic variation. *Nature Reviews Genetics*, *5*(10), 790–796. https://doi.org/10.1038/nrg1452

Gerashchenko, M. V., & Gladyshev, V. N. (2016). Ribonuclease selection for ribosome profiling. *Nucleic Acids Research*, *45*(2), e6. https://doi.org/10.1093/nar/gkw822

Glish, G. L., & Vachet, R. W. (2003). The basics of mass spectrometry in the twenty-first century. *Nature Reviews Drug Discovery*, *2*(2), 140–150. https://doi.org/10.1038/nrd1011

Göhring, J., Fulcher, N., Jacak, J., & Riha, K. (2013). TeloTool: a new tool for telomere length measurement from terminal restriction fragment analysis with improved probe intensity correction. *Nucleic Acids Research*, *42*(3), e21. https://doi.org/10.1093/nar/gkt1315

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Herbert, B., Shay, J. W., & Wright, W. E. (2003). Analysis of Telomeres and Telomerase. *Current Protocols in Cell Biology*, *20*(1). https://doi.org/10.1002/0471143030.cb1806s20

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., & Zamir, A. (1965). Structure of a Ribonucleic Acid. *Science*, *147*(3664), 1462–1465. https://doi.org/10.1126/science.147.3664.1462

Holley, R. W., Apgar, J., Merrill, S. H., & Zubkoff, P. L. (1961). NUCLEOTIDE AND OLIGONUCLEOTIDE COMPOSITIONS OF THE ALANINE-, VALINE-, AND TYROSINE-ACCEPTOR "SOLUBLE" RIBONUCLEIC ACIDS OF YEAST. *Journal of the American Chemical Society*, *83*(23), 4861–4862. https://doi.org/10.1021/ja01484a040

Hood, L., & Rowen, L. (2013). The human genome project: big science transforms biology and medicine. *Genome Medicine*, *5*(9), 79. https://doi.org/10.1186/gm483

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F., & Wheeler, T. J. (2015). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, *44*(D1), D81–D89. https://doi.org/10.1093/nar/gkv1272

Hutchison, C. A. (2007). DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*, *35*(18), 6227–6237. https://doi.org/10.1093/nar/gkm688

Ideker, T., Galitski, T., & Hood, L. (2001). A New Approach to Decoding Life: Systems Biology. *Annual Review of Genomics and Human Genetics*, *2*(1), 343–372. https://doi.org/10.1146/annurev.genom.2.1.343

Jeanclos, E., Schork, N. J., Kyvik, K. O., Kimura, M., Skurnick, J. H., & Aviv, A. (2000). Telomere Length Inversely Correlates With Pulse Pressure and Is Highly Familial. *Hypertension*, *36*(2), 195–200. https://doi.org/10.1161/01.hyp.36.2.195

Jenkins, F. J., Kerr, C. M., Fouquerel, E., Bovbjerg, D. H., & Opresko, P. L. (2017). Modified Terminal Restriction Fragment Analysis for Quantifying Telomere Length Using In-gel Hybridization. *Journal of Visualized Experiments*, *125*. https://doi.org/10.3791/56001

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11). https://doi.org/10.1101/pdb.top084970

Kulski, J. K. (2016). Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. *Next Generation Sequencing - Advances, Applications and Challenges*. Published. https://doi.org/10.5772/61964

Lee, K. M., Choi, K. H., & Ouellette, M. M. (2004). Use of exogenous hTERT to immortalize primary human cells. *Cytotechnology*, *45*(1–2), 33–38. https://doi.org/10.1007/10.1007/s10616-004-5123-3

Lin, J., Smith, D. L., Esteves, K., & Drury, S. (2019a). Telomere length measurement by qPCR – Summary of critical factors and recommendations for assay design. *Psychoneuroendocrinology*, *99*, 271–278. https://doi.org/10.1016/j.psyneuen.2018.10.005

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, *2012*, 1–11. https://doi.org/10.1155/2012/251364

Lizardi, P. M. (2008). Next-generation sequencing-by-hybridization. *Nature Biotechnology*, *26*(6), 649–650. https://doi.org/10.1038/nbt0608-649

Logsdon, G. A., Vollger, M. R., Hsieh, P., Mao, Y., Liskovykh, M. A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P. C., Rhie, A., de Lima, L. G., Dvorkina, T., Porubsky, D., Harvey, W. T., Mikheenko, A., Bzikadze, A. V., Kremitzki, M., Graves-Lindsay, T. A., Jain, C., . . . Eichler, E. E. (2021). The structure, function and evolution of a complete human chromosome 8. *Nature*, *593*(7857), 101–107. https://doi.org/10.1038/s41586-021-03420-7

Ludden, C., Reuter, S., Judge, K., Gouliouris, T., Blane, B., Coll, F., Naydenova, P., Hunt, M., Tracey, A., Hopkins, K. L., Brown, N. M., Woodford, N., Parkhill, J., & Peacock, S. J. (2017). Sharing of carbapenemase-encoding plasmids between Enterobacteriaceae in UK sewage uncovered by MinION sequencing. *Microbial Genomics*, *3*(7). https://doi.org/10.1099/mgen.0.000114

Lukash, L. L., Shvachko, L. P., & Kostetskaya, K. V. (1996). Mobile genetic elements in mutagenic process, recombination and malignization of human cells. *Biopolymers and Cell*, *12*(2), 7–19. https://doi.org/10.7124/bc.00041a

MacNeil, D. E., Lambert-Lanteigne, P., & Autexier, C. (2019). N-terminal residues of human dyskerin are required for interactions with telomerase RNA that prevent RNA degradation. *Nucleic Acids Research*, *47*(10), 5368–5380. https://doi.org/10.1093/nar/gkz233

Maniatis, T., Jeffrey, A., & Van deSande, H. (1975). Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. *Biochemistry*, *14*(17), 3787–3794. https://doi.org/10.1021/bi00688a010

Mardis, E. R. (2007). Example of NGS sequencing technology [Illustration]. In *Trends in Genetics* (p. 135).

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, *74*(2), 560–564. https://doi.org/10.1073/pnas.74.2.560

Mender, I., & Shay, J. (2015). Telomere Restriction Fragment (TRF) Analysis. *BIO-PROTOCOL*, *5*(22). https://doi.org/10.21769/bioprotoc.1658

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., . . . Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, *585*(7823), 79–84. https://doi.org/10.1038/s41586-020-2547-7

Migeon, B. R., & Kennedy, J. F. (1975). Evidence for the Inactivation of an X Chromosome Early in the Development of the Human Female. *Obstetrical & Gynecological Survey*, *30*(10), 693–694. https://doi.org/10.1097/00006254-197510000-00018

Mitchell, G. A., Labuda, D., Fontaine, G., Saudubray, J. M., Bonnefont, J. P., Lyonnet, S., Brody, L. C., Steel, G., Obie, C., & Valle, D. (1991). Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. *Proceedings of the National Academy of Sciences*, *88*(3), 815–819. https://doi.org/10.1073/pnas.88.3.815

Mitchell, J. R., Wood, E., & Collins, K. (1999). A telomerase component is defective in the human disease dyskeratosis congenita. *Nature*, *402*(6761), 551–555. https://doi.org/10.1038/990141

Montpetit, A. J., Alhareeri, A. A., Montpetit, M., Starkweather, A. R., Elmore, L. W., Filler, K., Mohanraj, L., Burton, C. W., Menzies, V. S., Lyon, D. E., & Jackson-Cook, C. K. (2014). Telomere Length: A review of methods for measurement. *Nursing Research*, *63*(4), 289–299. https://doi.org/10.1097/nnr.0000000000000037

Moolgavkar, S. H., & Knudson, A. G. (1981). Mutation and Cancer: A Model for Human Carcinogenesis. *JNCI: Journal of the National Cancer Institute*, *66*(6), 1037–1052. https://doi.org/10.1093/jnci/66.6.1037

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., & Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods*, *13*(11), 919–922. https://doi.org/10.1038/nmeth.3999

Muratani, K., Hada, T., Yamamoto, Y., Kaneko, T., Shigeto, Y., Ohue, T., Furuyama, J., & Higashino, K. (1991). Inactivation of the cholinesterase gene by Alu insertion: possible mechanism for human gene transposition. *Proceedings of the National Academy of Sciences*, *88*(24), 11315–11319. https://doi.org/10.1073/pnas.88.24.11315

Nawrot, T. S., Staessen, J. A., Gardner, J. P., & Aviv, A. (2004a). Telomere length and possible link to X chromosome. *The Lancet*, *363*(9408), 507–510. https://doi.org/10.1016/s0140-6736(04)15535-9

Nawrot, T. S., Staessen, J. A., Gardner, J. P., & Aviv, A. (2004b). Intrafamilial relations for TRF length [Graph]. In *The Lancet* (pp. 507–510).

Nusbaum, C., Mikkelsen, T. S., Zody, M. C., Asakawa, S., Taudien, S., Garber, M., Kodira, C. D., Schueler, M. G., Shimizu, A., Whittaker, C. A., Chang, J. L., Cuomo, C. A., Dewar, K., FitzGerald, M. G., Yang, X., Allen, N. R., Anderson, S., Asakawa, T., Blechschmidt, K., . . . Lander, E. S. (2006). DNA sequence and analysis of human chromosome 8. *Nature*, *439*(7074), 331–335. https://doi.org/10.1038/nature04406

Orth, K., Hung, J., Gazdar, A., Bowcock, A., Mathis, J. M., & Sambrook, J. (1994). Genetic instability in human ovarian cancer cell lines. *Proceedings of the National Academy of Sciences*, *91*(20), 9495–9499. https://doi.org/10.1073/pnas.91.20.9495

Ott, A., Schnable, J. C., Yeh, C. T., Wu, L., Liu, C., Hu, H. C., Dalgard, C. L., Sarkar, S., & Schnable, P. S. (2018). Linked read technology for assembling large complex and polyploid genomes. *BMC Genomics*, *19*(1). https://doi.org/10.1186/s12864-018-5040-z

Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, *52*(4), 413–435. https://doi.org/10.1007/s13353-011-0057-x

Plénat, T., Yoshizawa, S., & Fourmy, D. (2017). DNA-Guided Delivery of Single Molecules into Zero-Mode Waveguides. *ACS Applied Materials & Interfaces*, *9*(36), 30561–30562. https://doi.org/10.1021/acsami.7b11953

Puurand, T., Kukuškina, V., Pajuste, F. D., & Remm, M. (2019a). AluMine: alignment-free method for the discovery of polymorphic Alu element insertions. *Mobile DNA*, *10*(1). https://doi.org/10.1186/s13100-019-0174-3

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Rich, J., Ogryzko, V. V., & Pirozhkova, I. V. (2014). Satellite DNA and related diseases. *Biopolymers and Cell*, *30*(4), 249–259. https://doi.org/10.7124/bc.00089e

Rizvi, S., Raza, S. T., & Mahdi, F. (2015). Telomere Length Variations in Aging and Age-Related Diseases. *Current Aging Science*, *7*(3), 161–167. https://doi.org/10.2174/1874609808666150122153151

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G. R., Burrows, C., Bird, C. P., Frankish, A., Lovell, F. L., Howe, K. L., Ashurst, J. L., Fulton, R. S., Sudbrak, R., Wen, G., Jones, M. C., Hurles, M. E., . . . Bentley, D. R. (2005). The DNA sequence of the human X chromosome. *Nature*, *434*(7031), 325–337. https://doi.org/10.1038/nature03440

Royal, C. D. M., & Dunston, G. M. (2004). Changing the paradigm from "race" to human genome variation. *Nature Genetics*, *36*(S11), S5–S7. https://doi.org/10.1038/ng1454

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, *265*(5596), 687–695. https://doi.org/10.1038/265687a0

Sanger, F., Brownlee, G., & Barrell, B. (1965). A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of Molecular Biology*, *13*(2), 373–398. https://doi.org/10.1016/s0022-2836(65)80104-8

Sanger, F., & Coulson, A. (1975a). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441–448. https://doi.org/10.1016/0022-2836(75)90213-2

Sanger, F., & Coulson, A. R. (1975b). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase [Illustration]. In *Journal of Molecular Biology* (p. 442).

Schmid, C., & Jelinek, W. (1982). The Alu family of dispersed repetitive sequences. *Science*, *216*(4550), 1065–1070. https://doi.org/10.1126/science.6281889

Schwarze, K., Buchanan, J., Fermont, J. M., Dreau, H., Tilley, M. W., Taylor, J. M., Antoniou, P., Knight, S. J. L., Camps, C., Pentony, M. M., Kvikstad, E. M., Harris, S., Popitsch, N., Pagnamenta, A. T., Schuh, A., Taylor, J. C., & Wordsworth, S. (2019). The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genetics in Medicine*, *22*(1), 85–94. https://doi.org/10.1038/s41436-019-0618-7

Sevim, V., Bashir, A., Chin, C. S., & Miga, K. H. (2016). Alpha-CENTAURI: assessing novel centromeric repeat sequence variation with long read sequencing. *Bioinformatics*, *32*(13), 1921–1924. https://doi.org/10.1093/bioinformatics/btw101

Shammas, M. A. (2011). Telomeres, lifestyle, cancer, and aging. *Current Opinion in Clinical Nutrition and Metabolic Care*, *14*(1), 28–34. https://doi.org/10.1097/mco.0b013e32834121b1

Sheikh-Hosseini, M., Larijani, B., Gholipoor Kakroodi, Z., Shokoohi, M., Moarefzadeh, M., Sayahpour, F. A., Goodarzi, P., & Arjmand, B. (2021). Gene Therapy as an Emerging Therapeutic Approach to Breast Cancer: New Developments and Challenges. *Human Gene Therapy*. Published. https://doi.org/10.1089/hum.2020.199

Shvachko, L. P. (1998). The mutation processes resulting to cancer: the role of ALU repeats in genetic instability. *Biopolymers and Cell*, *14*(5), 389–395. https://doi.org/10.7124/bc.0004e3

Slagboom, P. E., Droog, S., & Boomsma, D. I. (1994). Genetic determination of telomere size in humans: a twin study of three age groups. *American journal of human genetics*, *55*(5), 876–882

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next‐Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1), 1–6. https://doi.org/10.1002/cpmb.59

Smith, J. B. (2001). Peptide Sequencing by Edman Degradation. *Encyclopedia of Life Sciences*. Published. https://doi.org/10.1038/npg.els.0002688

Stellwagen, N. C. (2009). Electrophoresis of DNA in agarose gels, polyacrylamide gels and in free solution. *Electrophoresis*, *30*(S1), S188–S195. https://doi.org/10.1002/elps.200900052

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., & Eichler, E. E. (2010). Diversity of Human Copy Number Variation and Multicopy Genes. *Science*, *330*(6004), 641–646. https://doi.org/10.1126/science.1197005

Tabarés-Seisdedos, R., & Rubenstein, J. L. R. (2009). Chromosome 8p as a potential hub for developmental neuropsychiatric disorders: implications for schizophrenia, autism and cancer. *Molecular Psychiatry*, *14*(6), 564. https://doi.org/10.1038/mp.2009.2

Tang, W., Mun, S., Joshi, A., Han, K., & Liang, P. (2018). Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp

sequence increase. *DNA Research*, *25*(5), 521–533. https://doi.org/10.1093/dnares/dsy022

Theobald, D. L. (2010). A formal test of the theory of universal common ancestry. *Nature*, *465*(7295), 219–222. https://doi.org/10.1038/nature09014

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., . . . Zhu, X. (2001). The Sequence of the Human Genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., . . . Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wong, J. M., & Collins, K. (2003). Telomere maintenance and disease. *The Lancet*, *362*(9388), 983–988. https://doi.org/10.1016/s0140-6736(03)14369-3

Wrighton, K. (2021, February 10). *Filling in the gaps telomere to telomere*. Nature. https://www.nature.com/articles/d42859-020-00117-1?error=cookies_not_supported&code=878f7ecd-6521-4a8d-982c-9b06184e09e2

Xia, Y., Li, A. D., Deng, Y., Jiang, X. T., Li, L. G., & Zhang, T. (2017). MinION Nanopore Sequencing Enables Correlation between Resistome Phenotype and Genotype of Coliform Bacteria in Municipal Sewage. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.02105

Yang, Y., Xie, B., & Yan, J. (2014). Application of Next-generation Sequencing Technology in Forensic Science. *Genomics, Proteomics & Bioinformatics*, *12*(5), 190–197. https://doi.org/10.1016/j.gpb.2014.09.001

Yates, J. R. (2011, July 28). *A century of mass spectrometry: from atoms to proteomes*. Nature Methods. https://www.nature.com/articles/nmeth.1659?error=cookies_not_supported&code=6ea2e6ee-67b5-47a8-8031-21e94bd383bb

INTERNET SOURCES

10x Genomics. (n.d.). *Linked-Reads Genomics*. Retrieved April 2021, from https://www.10xgenomics.com/products/linked-reads

An example Maxam–Gilbert sequencing reaction. (n.d.). [Illustration]. Wikipedia. https://en.wikipedia.org/wiki/File:Maxam-Gilbert_sequencing_en.svg

ATCC. (n.d.). *hTERT Immortalized Primary Cells*. Retrieved April 3, 2021, from https://www.lgcstandards-atcc.org/Products/Cells_and_Microorganisms/hTERT_Immortalized_Cell_Lines.aspx?geo_country=ee

Bionano Genomics. (n.d.). *DLS Technology*. Retrieved April 1, 2021, from https://bionanogenomics.com/technology/dls-technology/

CD Genomics. (n.d.). *Overview of PacBio SMRT sequencing: principles, workflow, and applications – CD Genomics*. Retrieved March 2021, from https://www.cd-genomics.com/pacbio-smrt-system-single-molecule-real-time-sequencing.html

Göhring, J., & Fulcher, N. (2013, December). *jagoehring/TeloTool*. GitHub. https://github.com/jagoehring/TeloTool

Griffiths, A. J. F. (n.d.). *DNA sequencing | genetics*. Encyclopedia Britannica. Retrieved February 2021, from https://www.britannica.com/science/DNA-sequencing

Lin, J., Smith, D. L., Esteves, K., & Drury, S. (2019b). *Schematic diagram of qPCR telomere length measurement as described in Cawthon, 2002* [Diagram]. ScienceDirect. https://ars.els-cdn.com/content/image/1-s2.0-S0306453018307789-gr1a_lrg.jpg

National Human Genome Research Institute. (n.d.). *The cost per Human Genome* [Graph]. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Orenstein, D. (2009, August). *Professor sequences his entire genome at low cost, with small team*. Stanford News. https://news.stanford.edu/news/2009/august10/genome-081009.html

Puurand, T., Kukuškina, V., Pajuste, F.-D., & Remm, M. (2019b). *bioinfo-ut/AluMine*. GitHub. https://github.com/bioinfo-ut/AluMine#readme

Sanger DNA Sequencing Steps and Method. (n.d.). Merck. Retrieved February 2021, from https://www.sigmaaldrich.com/technical-documents/articles/biology/sanger-sequencing.html

Sevim, V., Chin, J., Bashir, A., & Miga, K. (n.d.). *volkansevim/alpha-CENTAURI*. GitHub. Retrieved May 2021, from https://github.com/volkansevim/alpha-CENTAURI

Three Basic Steps of Automated Sanger Sequencing. (n.d.). [Illustration]. Merck. https://www.sigmaaldrich.com/content/dam/sigma-aldrich/articles/biology/marketing-assets/sanger-sequencing_steps_process_diagram.png

Appendix

Glossary

DNA ligation The term refers to an enzymatic reaction that creates a covalent bond between two biomolecules. Ligation is used to repair single and double strand DNA breaks as part of natural cellular processes such as DNA replication	DNA amplification Production of multiple copies of a sequence of DNA
Q-banding The fluorescent staining procedure, which employs quinacrine, is used to recognize individual chromosomes and structural abnormalities based on the banding pattern produced	DNA hybridization Technique that measures the degree of genetic similarity between pools of DNA sequences, usually used to determine the genetic distance between two organisms

NON-EXCLUSIVE LICENCE TO REPRODUCE THESIS AND MAKE THESIS PUBLIC

I, Daria Kholodniuk

(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

Utilizing the Knowledge of the Completely Assembled Human Chromosomes X and 8,

(*title of thesis*)

supervised by Ph.D. Reidar Andreson

(*supervisor's name*)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Daria Kholodniuk

20/05/2021