UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Molecular and Cell Biology

Farid Naghiyev

5S rDNA copy number in WGS data

Bachelor's Thesis (12 ECTS)

Curriculum Science and Technology

Supervisor:
MSc Tarmo Puurand

Tartu 2022

5S rDNA copy number in WGS data

Abstract:

In this thesis, 4 k-mers with two mutations were chosen to describe the copy number of human 5S rDNA in the whole genome data of different individuals. The study found that the number of copies varies between individuals and populations.

Keywords:

5S RNA, DNA, RNA, rDNA, rRNA, k-mer

CERCS: B110

5S rDNA geenikoopia arv täisgenoomi sekveneerimisandmetes

Lühikokkuvõte:

Antud töös valiti suunatult kahe mutatsiooniga 4 k-meeri, mis oma informatiivsusega kirjeldavad ära inimese 5S rDNA koopiaarvu eri indiviidide täisgenoomi andmetes. Töös leiti, et indiviidide ja populatsioonide vahel on koopiaarv varieeruv.

Võtmesõnad:

5S RNA, DNA, RNA, rDNA, rRNA, k-meer

CERCS: B110

TABLE OF CONTENTS

TERMS, ABBREVIATIONS AND NOTATIONS

mRNA – Messenger RNA

NGS – Next Generation Sequencing

NOR – Nucleolus Organizer Regions

rDNA – Ribosomal DNA

rRNA – Ribosomal RNA

SNP – Single-nucleotide Polymorphism

WGS – Whole Genome Sequence

INTRODUCTION

As the next-generation sequencing techniques improve, the difficulty to sequence a full genome has decreased significantly. For the most part, the analysis of genomic data requires either de novo assembly of the genome, mapping to a reference genome, or homology searches from raw reads. All these processes are either time-consuming or error prone.

Due to this, a relatively new oligomer frequency-based method of genome analysis caught a lot of attention as it does not require genome assembly. This method usually uses k-mers – oligomers of length k - to conduct the analysis (Kaplinski et al., 2015). In this thesis, we wanted to find the variability in 5S rDNA copy numbers within a group of people. For that, we used the genomic data of over two thousand individuals that were obtained using whole-genome sequencing. We combined this with k-mer genome analysis to obtain a k-mer list using GenomeTester4.

With this information, we were able to perform a pairwise alignment of 5S rDNA from two complete human genome sequences – CHM13 and GRCh38 - using bioinformatics software, such as MegaX and NCBI Blast. The resulting dot plot from the pairwise alignment of these two sequences depicted the difference between them. Next, we performed a multiple alignment of CHM13 5S rDNA cluster units to see how similar they are to one another.

Lastly, we performed a 5S RNA variability and distribution in human populations by k-mers to identify the difference in variability and distribution within individuals and populations.

1 LITERATURE REVIEW

1.1 Cells – architecture overview

Cells consist of organelles, which are discrete bodies that have different functions within the cell (**Figure 1**). To proliferate cells must divide. The process of division is one of the processes of the cell cycle. This process ensures that both chromosomal DNA molecules and mitochondrial DNA molecules are successfully replicated and divided between two daughter cells. Depending on various features, cells can be either prokaryotes or eukaryotes. Prokaryotes consist of bacteria and archaea. Eukaryotic cells are more complex than prokaryotic cells. Eukaryotes have a selectively permeable cell membrane, which regulates the transport of ions and smaller molecules into the cell and vice versa. Eukaryotic cells have a nucleus that contains most of the cell's DNA (Strachan et al., 2011).

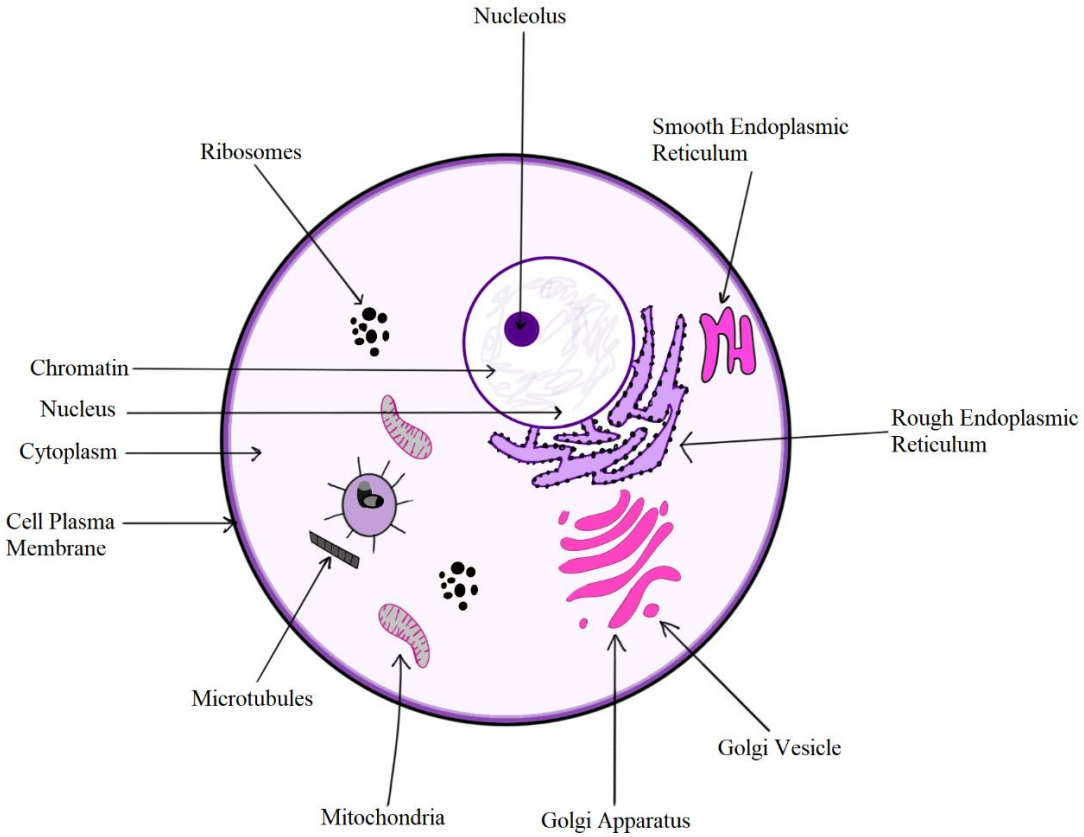

Figure 1. Eukaryotic Cell Structure and Components.

Eukaryotic cells also have various cell membrane-bound and membrane unbound organelles including the cytoskeleton, the centrosome, and the ribosome. Interactions between the cell and its environment are mediated by the plasma membrane. The membrane consists of the lipid bilayer which acts as a barrier between two aqueous compartments.

Another important compartment of the cell is the cytoskeleton. It controls various cell functions like cell shape, cell movement, cell structure support, and muscle contraction. The main parts of the cytoskeleton are the three filaments: actin microfilaments, intermediate microfilaments, and microtubules.

Most of the genetic material is located in the cell nucleus. DNA is isolated from the cytoplasm with the help of the nuclear membrane (Haddad, 2020).

1.2 Ribosomes – human rDNA-s and proteins

Ribosomes play a key role in the production of proteins. They regulate the growth and development of various organisms including eukaryotic and mammalian cells. An average eukaryotic cell consists of 4 RNA molecules and around 80 ribosomal proteins. This structure is held together with the help of around 300 RNA and protein cofactors. The binding process involves almost all 80 ribosomal proteins (Gupta & Warner, 2014).

The rDNA (ribosomal DNA) gene repeats are well-researched parts of the chromosome. Their repetitive structure allows studies on cellular processes such as DNA replication, recombination, and transcription. In eukaryotic cells, the rRNA (ribosomal RNA) gene repeats are located on the chromosome, in a bundle of rDNA repeats (Kobayashi, 2014). Usually, these repeats are located at the NORs (Nucleolar organizer regions). The number of rDNA copies may significantly vary among different species, as the genes are mostly generated via crossing-over and/or sister chromatid exchange (Gregory, 2005). The main product of rDNA is ribosomal RNA (rRNA). Ribosomes are composed of rRNA and ribosomal proteins. The main function of a ribosome is to translate mRNA to protein (Kobayashi, 2014).

The human ribosome consists of two major components: rRNA and roughly 80 ribosomal proteins. Four types of rRNA are present in humans: 18, 5.8, 28, and 5S rRNA. The ribosomes consist of two subunits – the 40S small subunit which contains 18S rRNA and proteins, and the 60S large subunit which contains 28S, 5.8S, 5S rRNA, and proteins (**Figure 2**) (Nieto et al., 2020).

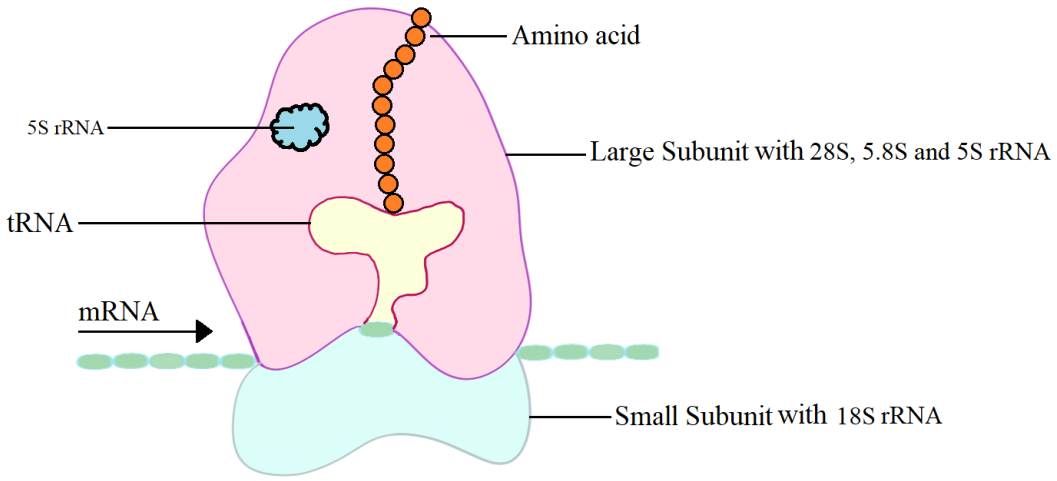

Figure 2. The structure of a ribosome.

1.3 rDNA genes in detail

The human genome contains approximately 400 copies of rDNA tandemly arrayed in NORs on the five acrocentric chromosomes (13, 14, 15, 21, and 22 chromosomes) (**Figure 3**). Each item includes 13.3 kb (kilobase) of encoding the 28, 5.8, 18S, and 45S rRNAs, and a non-coding IGS (intergenic spacer) (**Figure 4.** Schematic representation of an rDNA repeat (Agrawal & Ganley, 2018; Symonová, 2019).). These rRNAs, combined with 5S rRNA allow for the formation of the ribosome's nucleic acid backbone (Malinovskaya et al., 2018).

rDNA chromatin occurs in at least three different states. The first state is transcriptionally active rDNA copies. In this state, rDNA is euchromatic, hypomethylated at CpG sites, and marked with histone modifications that are related to transcriptionally active nucleoplasmic genes (Hamperl et al., 2013; Malinovskaya et al., 2018). The transcriptionally active rDNA copies stand for nearly 50% of the total rDNA copies in the genome.

In the second state rDNA copies are either inactive non-methylated or insignificantly methylated. They are usually found in the same structure of the nucleolus as are transcriptionally active copies and are considered a normal part of the operation of the nucleolus.

The last state (third) consists of inactive hypermethylated rDNA copies, which account for the heterochromatin around the nucleolus (Malinovskaya et al., 2018).

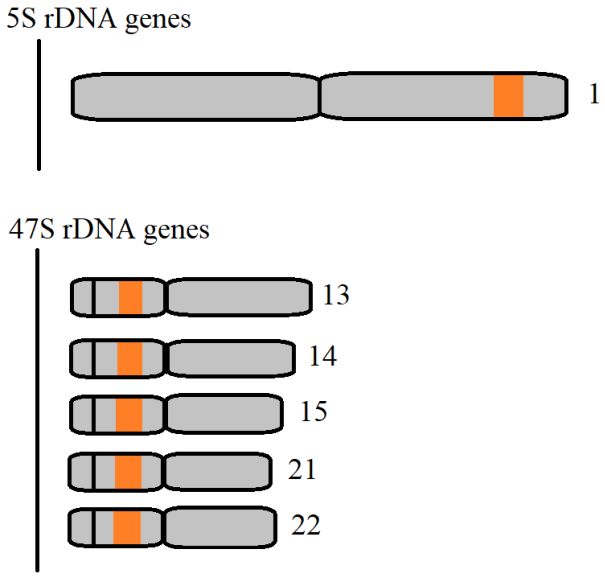

Figure 3. rDNA genes.

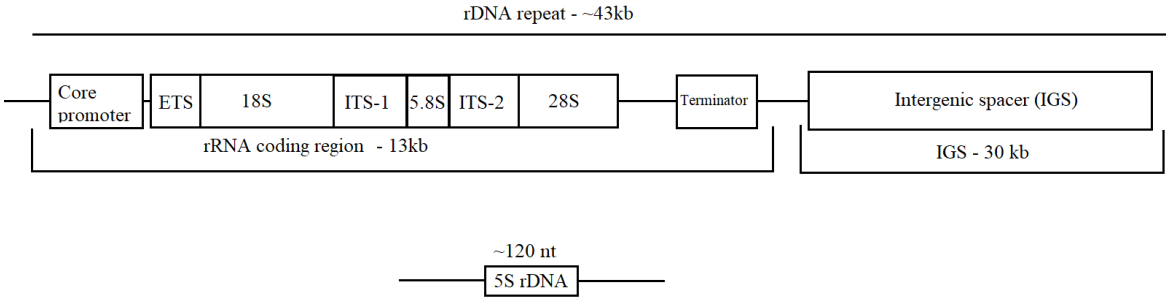

Figure 4. Schematic representation of an rDNA repeat (Agrawal & Ganley, 2018; Symonová, 2019).

1.4 Pseudogenes

Pseudogenes are repetitive and non-coding genes that have lost the ability to produce functional proteins and due to this are considered "junk DNA" (Pink et al., 2011). The term "pseudogene" was thought up in 1977 when C. Jacq, et al discovered a version of the 5S rRNA coding gene that was truncated but retained homology with the normal gene in Xenopus laevis (Jacq et al., 1977).

Following this, pseudogenes were rarely discovered (Mighell et al., 2000).

Pseudogenes can be transcriptionally silent or active. There are numerous ways for pseudogenes to form (**Figure 5.** The structure of the three types of pseudogenes (Pink et al., 2011).). Unitary pseudogenes originate because of spontaneous mutations that prevent transcription and translation of the gene. Duplicated pseudogenes form because of tandem

replications or uneven crossing-over, which in turn disable their protein-coding ability due to mutations. Another type of pseudogenes is retrotransposed pseudogenes. They originated when an mRNA gene is reverse-transcribed and inserted into a different position on the genome, which means that they do not contain introns. The insertion of mRNA into the genome is controlled by long interspersed nuclear element 1 (L1). The number of pseudogenes is surprisingly similar to the number of coding genes ranging from ten thousand to twenty thousand pseudogenes in humans, where the majority of pseudogenes are retrotransposed pseudogenes.

Although most pseudogenes lost their ability to be transcribed due to mutations in the promoter or integration into silent regions of the genome, numerous pseudogenes can be transcribed, such as tumor suppressor PTEN (Fujii et al., n.d.; Pink et al., 2011).

Some pseudogenes have promoters that control their transcriptional activity, whereas pseudogenes that do not have personal promoters rely on the promoters of nearby genes. Housekeeping genes with high expression tend to produce more PPs, along with other highly transcribed shorter RNAs. This fact can be proved by the small number of ribosomal encoding genes which account for ~20% of PPs (Pink et al., 2011).

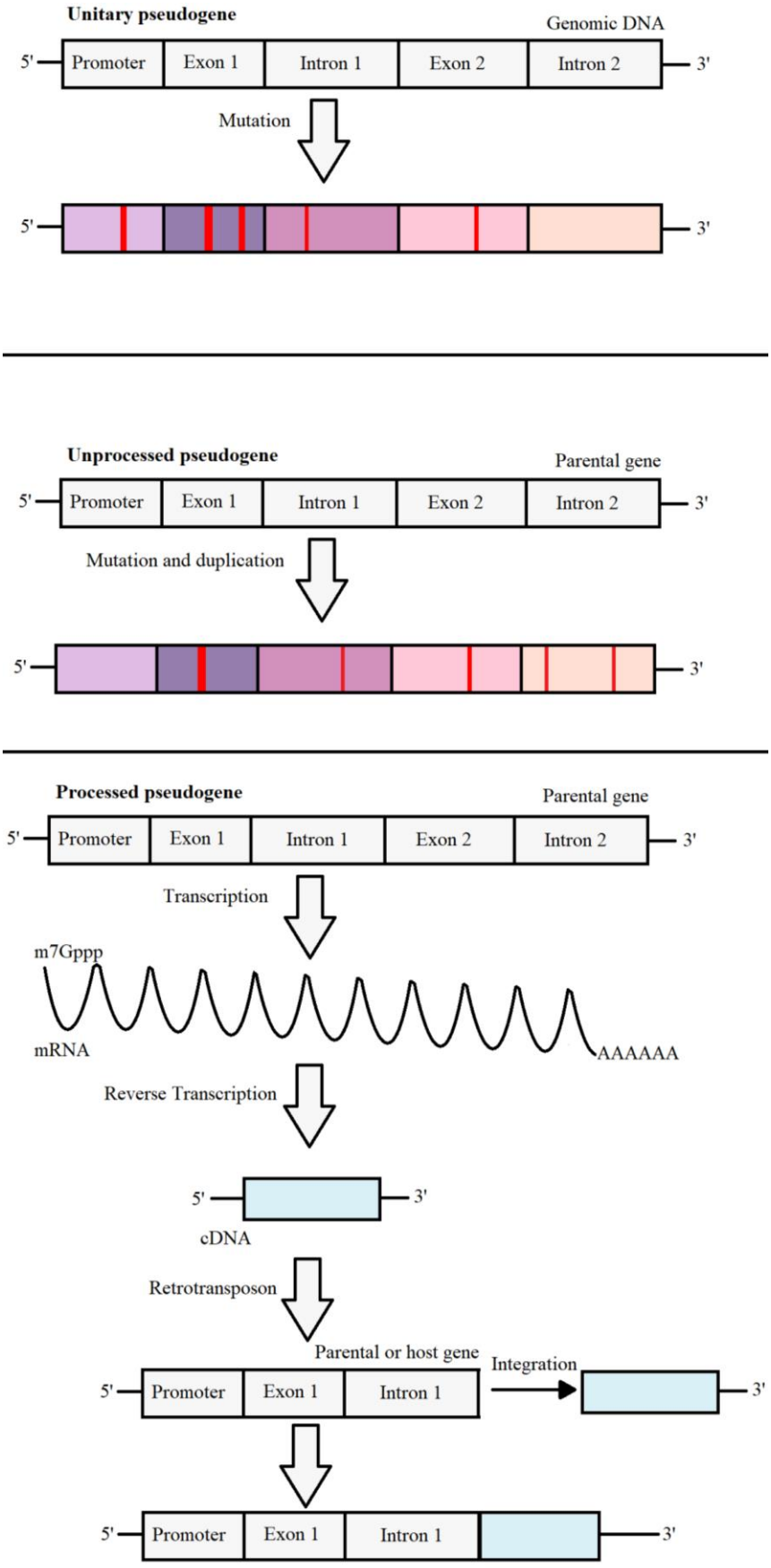

Figure 5. The structure of the three types of pseudogenes (Pink et al., 2011).

1.5 GRCh38 and CHM13 – complete human genome

The human genome consists of 22 diploid chromosomes and mitochondrial DNA. The human reference genome is essential for much sequencing-based research. One of these reference genomes is GRCh38, which nowadays is considered the most accurate human genome sequence. The sequencing was performed using Sanger sequencing. Compared to its predecessor GRCh37, GRCh38 altered 8000 nucleotides, corrected gaps and misassembled regions, and overall improved the diversity of reference by including alternate loci across the regions.

Despite all the advantages of GRCh38, it is not perfect. One of its main unresolved problems is the highly repetitive centromeres. Modeled centromeres were used in GRCh38 to fill the gaps. The models identify alpha-satellite centromere sequences from reads. Then they represent the approximate repeats for every alpha-satellite sequence. Even though centromeres represented in this way are just an approximation, it is a solid improvement compared to GRCh37 in which centromeres were represented as gaps.

Additionally, the reference genome represents only one allele per genomic site even though in diploid regions there are two alleles per individual (Guo et al., 2017).

Another human genome sequence is a new human genome sequence CHM13, an improved version of the GRCh38 genome, which includes gapless assemblies for all chromosomes except the Y chromosome, corrects many errors, and introduces almost 200 million base pairs of sequence containing almost 2000 gene predictions. The GRCh38 assembly contains 151 Mbp (mega base pairs) of unknown sequence, which includes duplications, gene arrays, and rDNA arrays, which are considered necessary for fundamental cellular processes. One of the large reference gaps includes human satellite repeat arrays and short arms of all acrocentric chromosomes. Additionally, GRCh38 shows a genome-wide deletion bias which is suggestive of incomplete assembly.

The fundament of the T2T (telomere to telomere) - CHM13 assembly is a high-resolution assembly string graph constructed directly from HiFi reads (Nurk et al., 2022).

1.6 Sequencing methods (Sanger, NGS-Illumina, Pacbio, Oxford Nanopore)

There are many ways to determine the nucleotide sequence of DNA. I will review some of the most popular methods available for wide use.

Sanger sequencing is the method of selective insertion of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication. Since both alleles of the autosomal locus are sequenced simultaneously, the Sanger sequencing method can miss significant portions of low-level mutations and at higher allele fractions it can misidentify mutations (Jamuar et al., 2016).

Even though the Sanger sequencing method is a fundamental breakthrough in DNA sequencing, it is considered relatively slow by current NGS standards (Next-generation sequencing). Nevertheless, it is still a useful choice for experiments where very high throughput is not required.

The need for cheaper but more effective sequencing methods gave birth to a new sequencing standard “NGS”.

One of the leading nextgen (next generation) sequencing technologies is Illumina. It is considered a second-generation sequencing method and based on a bridge amplification technique in which DNA molecules with specific adapters ligated on the corresponding ends are used as substrates for repeated amplification synthesis reactions on a glass slide. The glass slide contains oligonucleotide sequences that are connected to the ligated adapters. The spacing of oligonucleotide is done in a certain way, which after the number of amplifications, creates cloned cluster which contains approximately 1000 copies of each oligonucleotide fragment. Illumina sequencing supports many protocols such as DNA sequencing, RNA sequencing, and ChIP-sequencing.

Along with the second generation NGS, another sequencing generation has been created to sequence long DNA and RNA. This sequencing generation is called the third generation. One of the leaders in this area is PacBio (Pacific Biosciences). This technology allows scientists to sequence large data (50 kb or even longer). The method relies on bonding a specifically engineered DNA polymerase, with bound DNA of choice to the bottom of a ZMW (zero-mode waveguide) well. A ZMW is a small compartment that moves light energy into a very small area. Due to the specific design of ZMW, imaging occurs on the bottom of the ZMW, where the DNA polymerase, with bound DNA to it, inserts every base to a corresponding growing chain. Each nucleotide is labeled with a phosphor-lin. Then the growing chain will be able to differentiate between bases as the correct fluorescently labeled nucleotide is bound. This enables imaging at a very fast (up to millisecond) time scale.

The template preparation process involves the production of a circular double-stranded DNA molecule with the adapter sequence complementary to the primers that were used to initiate

DNA synthesis on the template. As a result, the polymerase can read a large template many times until the polymerase stops.

Despite all the features, PacBio has, it is prone to a high error rate. However, this can be overcome as the errors are stochastic and not systematic, meaning that sequencing the template multiple times shall result in a high accuracy sequence.

An even larger amount of DNA data (hundreds of kb up to hundreds of Gb) can be sequenced with the newest sequence generation "fourth-generation sequencing". One of the fourth-generation methods is Nanopore-based DNA sequencing. The company behind this technology is Oxford Nanopore Technologies. They offer various sequencing devices, from "portable" models (MinION) to high throughput models (PromethION). The technology is based on protein nanopores which are in electrically resistant polymer membranes through which current changes occur as nucleotides pass through the detector. Long dsDNA (double-stranded DNA) molecules are first bound to the processive enzyme. When this complex reaches a nanopore, one of the strands will enter the nanopore where the translocation rate through the pore is controlled by DNA polymerase. The processive enzyme will allow the DNA to be condensed through it. By moving through the pore, the nucleotide disrupts the current that has been applied to the nanopore. Each nucleotide provides an electronic signal that is recorded in real-time as a current disruption event (Slatko et al., 2018).

Table 1. Comparison of NGS methods (Kwong et al., 2015; Minervini et al., 2020).

Sequencing method	Read length	Error Rate
Illumina	Up to 600 bp	<0.1%
PacBio	Up to 20 kb	14% (can decrease up to ~0.1% for SNV and 4% for Indel)
Nanopore Sequencing	Up to 2 Mbps	Up to 5% for SNV and up to 10% for Indel

1.7 Sequencing coverage

Even though sequencing techniques have dramatically improved over the years, the sequencing cost remains substantial and varies from experiment to experiment thus higher coverage of sequencing will require higher costs. The coverage can be expected or actual. The expected coverage is the average number of times every nucleotide is expected to be sequenced within a certain amount of reads of a given length. On the other hand, actual coverage depicts the actual number of times that a nucleotide in the reference is covered by an aligned read (Sims et al., 2014).

Redundancy of coverage can be depicted as the number of reads that contribute to each consensus base. To calculate the redundancy of coverage it is required to divide the lengths of all of the sequencing reads that contribute to the sequence by the number of bases in the sequence (Bouck et al., 1998). Redundancy of coverage can also be called the depth of coverage, or shortly the depth.

The percentage of target bases that are sequenced in the given amount of time is called the breadth of coverage. In a best-case scenario, the sequencing method would successfully read all nucleotides sequentially from one end to the other of each chromosome, which would ensure that all polymorphic alleles could be identified, and all long and near-identical repetitive regions could be placed in a genome assembly. However, in a real case scenario, read lengths are relatively short – about 250 nucleotides – and can potentially have sequence errors. Increasing the number of sequencing reads is the key to fixing this problem. Nevertheless, increasing the depth of coverage is not a universal fix as it cannot fill sequence gaps that are caused by repetitive regions with lengths that are equal to or higher than those of reads.

In de novo genome sequencing the major factors that determine the required depth are the error rate of the selected sequencing method, the assembly algorithms, the repeat complexity of the genome, and the read length.

Nowadays, high-quality assemblies are produced using hybrid methods, in which the advantages of high-depth and short-read sequencing are combined with the advantages of lower-depth and longer-read sequencing. A modern approach to sequencing genomes with a huge number of repeats is to barcode and sequence with approximately 20x depth all read that are derived from clusters of short DNA fragments (Bouck et al., 1998).

1.8 K-mers

As high-throughput sequencing technologies progress, larger amounts of data can be processed. Due to this, many tools now require counts of substrings of length k (k-mers) in genomic sequencing reads. These k-mers can be used in various bioinformatics applications, such as genome assembly and multiple sequence alignment.

K-mer counting involves counting the number of substrings of a length k in a set of strings where k is a positive integer. K-mer counting is useful in de novo genome projects, where genomic characteristics such as genome size are estimated by analyzing the k-mer distribution frequency. Quantitative features of repetitive DNA can be determined by the distribution of frequencies of long k-mers.

Progress in NGS technologies provides us with longer reads. Due to this, longer Illumina reads experience lower accuracy. K-mers with large values can improve the accuracy of these reads. There are various approaches to counting k-mers that can be categorized based on the approach they use.

A popular method is to calculate k-mers using the sorting approach, where k-mers will be extracted from each reading. Then, the obtained k-mer frequencies can be counted and repeating k-mers put into the correct position. Many applications can perform this calculation (Manekar & Sathe, 2018). One of them is GenomeTester4 (GListMaker), which uses the sorting approach to collect all k-mers from the input file and sort them to produce the final k-mer count. One of the advantages of GenomeTester4 is multithreading which improves the speed of k-mer counting (Kaplinski et al., 2015).

If the genome of interest is perfectly amplified, then all the pieces of the reads should be evenly distributed over all regions, and the k-mer histogram which essentially represents normal distribution will form (**Figure 6.** This figure represents the k-mer frequency distribution histogram. The black line represents the number of k-mers with a certain frequency. Point P represents the main peak point.) (Sohn & Nam, 2016).

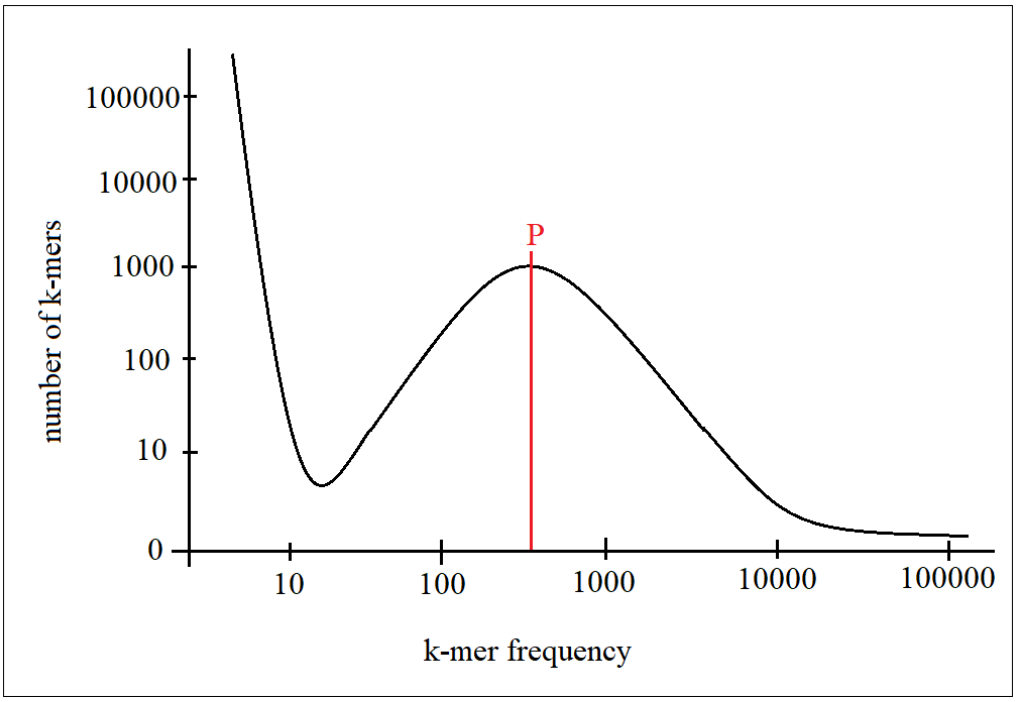

Figure 6. This figure represents the k-mer frequency distribution histogram. The black line represents the number of k-mers with a certain frequency. Point P represents the main peak point.

2 THE AIMS OF THE THESIS

1. Dot plot of 5S RNA between GRCh38 and CHM13 assembly.
2. Multiple alignment of CHM13 5S rDNA cluster units.
3. 5S RNA variability and distribution in human populations by k-mers.

3 EXPERIMENTAL PART

3.1 MATERIALS AND METHODS

3.1.1 1000 Genome project

2504 individuals from 26 populations including Africa, East Asia, Europe, South Asia, and the Americas were sampled. All individuals were sequenced using whole-genome sequencing with a mean depth of 7.4x, and targeted exome sequencing with a mean depth of 65.7x. People and their first-degree relatives were genotyped using high-density SNP microarrays. multi-allelic SNPs, indels, and various groups of structural variants. Variant discovery used 24 sequence analysis tools and machine-learning algorithms to distinguish between high-quality variants and false positives (The 1000 Genomes Project Consortium et al., 2015). Reference genome sequences GRCh38 and CHM13 were used for multiple pair-wise alignments. The sequences were converted to FASTA format.

3.1.2 K-mer list

To create and analyze our k-mer list, we used a software package named GenomeTester4. The GenomeTester4 consists of 3 tools: GListMaker, GListCompare, and GListQuery. GlistMaker is used to generate the k-mer count list from the nucleotide sequences. It uses FASTA or FASTQ file format to obtain nucleotide sequences. Then it uses arrays to store k-mers from the input file, which are later sorted and adjacent instances of the same k-mer are counted during the collation phase.

GListCompare performs algebraic operations with two lists of choice, such as union, intersection, and difference. All operations are performed simultaneously over both lists and written into a new combined list. The union of two lists contains entries that were present in either list, while the intersection contains only entries that were present in both lists. The complement provides the entries that were present only in the first list.

Ultimately, GListQuery was able to search for user-inputted sequences in the generated list. The input can be either single k-mer sequences or k-mer lists (Kaplinski et al., 2015).

3.1.3 Sequence alignment

To perform pair-wise alignment we used the NCBI blastn tool. We performed a megablast with the word size of 28. The query sequence was 5S RNA from GRCh38 and the subject sequence was 5S RNA from CHM13 human sequence. The resulting output file is a dot plot

(dot matrix) that shows regions of similarity, based on the blast results. For the multiple alignments, we used MEGA11 software and the BioEdit software.

3.1.4 K-mer selection

We selected 4 different k-mers discovered by visual searching in whole genome alignments in multiple individuals. With minimal numbers of k-mers, we chose k-mer with 2 mismatches (**Table 2**).

Table 2. k-mers have 2 mismatches and in sum, they are all possible variants included in the human genome. K-mers are not inside the 5S RNA gene but in flanking region.

K-mer	K-mer
1	GATGGATGGAGAGATAGAAACCGAG
2	GATGGAT**C**GAGAGATAGA**G**ACCGAG
3	GATGGATGGAGAGATAGA**G**ACCGAG
4	GATGGAT**C**GAGAGATAGAAACCGAG

3.2 RESULTS

3.2.1 Dot plot of 5S RNA between GRCh38 and CHM13 assembly

A dot plot between 5S RNA GRCh38 and CHM13 using NCBI blastn resulted in a dot plot, where the x-axis represents the query sequence, which is GRCh38, and the numbers represent the bases or residues of the query sequence. Y-axis represents the subject sequence, which is CHM13, and the numbers represent the bases or residues of the subject sequence. Vertical bars represent duplications and in total, we got 19 bars. This means that in positions where these bars exist, the subject sequence coincides with the query sequence. Numbers and red lines represent the positions of duplications on the Y-axis (**Figure 7**).

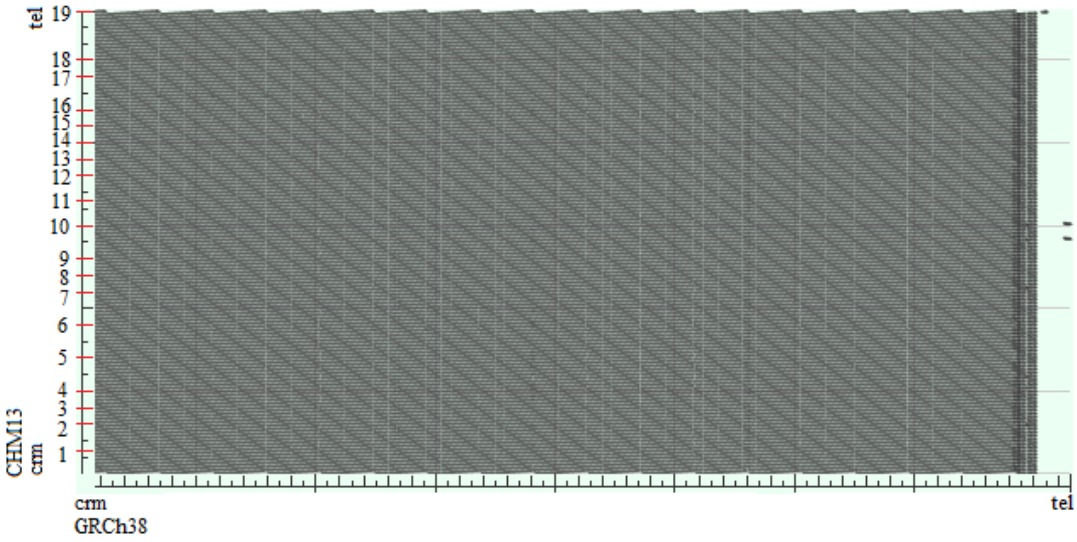

Figure 7. This dot plot represents the pairwise alignment of 5S RNA between GRCh38 (x-axis) and CHM13 (y-axis). „Crm“ stands for „centromere“, while „tel“ stands for „telomere“.

3.2.2 Multiple alignment of CHM13 5S rDNA cluster units

Multiple alignment of CHM13 5S rDNA showed us that the 5S rDNA sequences in CHM13 are very similar to one another, with only occasional changes. For example, Guanine is substituted by Thymine in position 62. (**Figure 8**).

Figure 8. Partial multiple alignment of CHM13 5S rDNA. The first column contains the position number of the tandemly repeated block start we have used in the chm13 assembly. Dots mean the same nucleotide as it is in the first line. 5S rDNA copies are in high similarity.

3.2.3 5S RNA variability and distribution in human populations by k-mers

Used k-mers showed variability in distribution within individuals and populations. K-mer frequencies normalized with sequencing coverage give us an indicative number of 5S RNA gene numbers, even the k-mers' numbers (Appendix Table). Our samples have 55-233 copies of 5S rDNA genes, every individual has copies of 1st k-mer, 55 individuals have 2nd k-mer and 4 individuals with African ancestry have 3rd k-mer.

Table 3. rDNA k-mer presence in different populations. The first k-mer is presented in every individual, therefore it is ancestral or there is pressure to change sequences, because the second common k-mer contains 2 mismatches.

Population	k-mer 1	k-mer 2	k-mer 3	k-mer 4
African	3	3	2	
Bengali	12	4		
British	6	1	1	
Chinese	4	1		
Colombian	10	1		
Finnish	15	8	1	
Gambian	1	1	1	
Gujarati	11	1		
Han	18	2	1	
Iberian	5	1	2	
Indian	12	7	1	
Japanese	15	1		
Kinh	4	1		
Mexican	1	1		
Peruvian	9	6		
Puerto	20	9	3	
Punjabi	11	4	1	
Sri	12	5	1	
Sri Lanka	1			
Toscani	4	2		
Utah	2	2		
Total	176	61	14	

DISCUSSION

The current work's main object was to find the minimal number and most informative k-mers for survey 5S gene distribution in different populations around the world. The found k-mers provided us with this information. Most surprising finding is that the second common k-mer contains 2 mismatches. This genetic phenomenon is often seen in single copy regions. But in genes with copy number over 50, it takes time to be fixed.

We have not used computational method for selecting k-mers, so we did it manually in directed search for it, find most informative ones. Manual work was successful after several WGS samples by just looking mismatches in alignments, which were occurring repeatedly after several kilobases of sequence.

Multiple alignment between the different 5S rDNA repeated blocks showed high similarities, so we may choose multiple other k-mers. Although, those k-mers may be good for copy number estimation, they may not be good for identifying very old splitting event. Every block contains other mutations but surveying them was not this work topic.

Pairwise alignment visualizing tool dot plot showed us similarity between CHM13 and GRCh38 assembly but the copy number difference between them was around 100. This fact alone proves copy number variability presence. K-mer method is not precise in copy number estimation, but probably there is no other in silico method to describe indicative amount of gene copies. 5S rDNA copy number varies between 55-233 in individuals.

First and second k-mer are presented around the world, but 3rd k-mer is only in African ancestry population. Denisovan human have most frequent 3rd k-mer and small amount of 1st k-mer, Neanderthal have only 1st k-mer with high copy number. For better understanding distribution in other populations, more individuals are required to be studied.

SUMMARY

In this thesis, our goal was to identify the difference in copy number of human 5S rDNA using the whole genome data of different populations and individuals, with the help of 100 Genome project and 4 different k-mers discovered by visual searching.

The dot plot showed us that 5S rDNA in GRCh38 and CHM13 have many regions of close similarity. This proved that our reference genomes were suitable to use.

By performing the multiple alignment of CHM13 5S rDNA cluster units, we ensured that these units were similar, despite the minuscule differences in nucleotides in some units.

Our k-mers showed a substantial variability between individuals and populations. And, although every individual had a set of 1^{sts} k-mer, other k-mers were found in different individuals, and in total our samples had 55-233 copies of 5S rDNA genes. We found that every individual had copies of 1st k-mer, 55 individuals had 2nd k-mer and 4 individuals with African ancestry had 3rd k-mer.

The most surprising fact was that the second most popular k-mer was the k-mer with 2 mismatches. We suspect that happened because there was a pressure to change the sequences.

REFERENCES

Agrawal, S., & Ganley, A. R. D. (2018). The conservation landscape of the human ribosomal RNA gene repeats. *PLOS ONE*, *13*(12), e0207531. https://doi.org/10.1371/journal.pone.0207531

Bouck, J., Miller, W., Gorrell, J. H., Muzny, D., & Gibbs, R. A. (1998). Analysis of the Quality and Utility of Random Shotgun Sequencing at Low Redundancies. *Genome Research*, *8*(10), 1074–1084. https://doi.org/10.1101/gr.8.10.1074

Fujii, G. H., Morimoto, A. M., Berson, A. E., & Bolen, J. B. (n.d.). *Transcriptional analysis of the PTEN/MMAC1 pseudogene, CPTEN*. 5.

Gregory, T. R. (2005). Genome Size Evolution in Animals. In *The Evolution of the Genome* (pp. 3–87). Elsevier. https://doi.org/10.1016/B978-012301463-4/50003-6

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., & Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, *109*(2), 83–90. https://doi.org/10.1016/j.ygeno.2017.01.005

Gupta, V., & Warner, J. R. (2014). Ribosome-omics of the human ribosome. *RNA*, *20*(7), 1004–1013. https://doi.org/10.1261/rna.043653.113

Haddad, L. A. (2020). Cellular structure and molecular cell biology. In *Clinical Molecular Medicine* (pp. 17–45). Elsevier. https://doi.org/10.1016/B978-0-12-809356-6.00002-2

Hamperl, S., Wittner, M., Babl, V., Perez-Fernandez, J., Tschochner, H., & Griesenbeck, J. (2013). Chromatin states at ribosomal DNA loci. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1829*(3–4), 405–417. https://doi.org/10.1016/j.bbagrm.2012.12.007

Jacq, C., Miller, J. R., & Brownlee, G. G. (1977). A pseudogene structure in 5S DNA of Xenopus laevis. *Cell*, *12*(1), 109–120. https://doi.org/10.1016/0092-8674(77)90189-1

Jamuar, S. S., D'Gama, A. M., & Walsh, C. A. (2016). Somatic Mosaicism and Neurological Diseases. In *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry* (pp. 179–199). Elsevier. https://doi.org/10.1016/B978-0-12-800105-9.00012-3

Kaplinski, L., Lepamets, M., & Remm, M. (2015). GenomeTester4: A toolkit for performing basic set operations - union, intersection and complement on k-mer lists. *GigaScience*, *4*(1), 58. https://doi.org/10.1186/s13742-015-0097-y

Kobayashi, T. (2014). Ribosomal RNA gene repeats, their stability and cellular senescence. *Proceedings of the Japan Academy, Series B*, *90*(4), 119–129. https://doi.org/10.2183/pjab.90.119

Kwong, J. C., Mccallum, N., Sintchenko, V., & Howden, B. P. (2015). Whole genome sequencing in clinical and public health microbiology. *Pathology*, *47*(3), 199–210. https://doi.org/10.1097/PAT.0000000000000235

Malinovskaya, E. M., Ershova, E. S., Golimbet, V. E., Porokhovnik, L. N., Lyapunova, N. A., Kutsev, S. I., Veiko, N. N., & Kostyuk, S. V. (2018). Copy Number of Human Ribosomal Genes With Aging: Unchanged Mean, but Narrowed Range and Decreased Variance in Elderly Group. *Frontiers in Genetics*, *9*, 306. https://doi.org/10.3389/fgene.2018.00306

Manekar, S. C., & Sathe, S. R. (2018). A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience*. https://doi.org/10.1093/gigascience/giy125

Mighell, A. J., Smith, N. R., Robinson, P. A., & Markham, A. F. (2000). Vertebrate pseudogenes. *FEBS Letters*, *468*(2–3), 109–114. https://doi.org/10.1016/S0014-5793(00)01199-6

Minervini, C. F., Cumbo, C., Orsini, P., Anelli, L., Zagaria, A., Specchia, G., & Albano, F. (2020). Nanopore Sequencing in Blood Diseases: A Wide Range of Opportunities. *Frontiers in Genetics*, *11*, 76. https://doi.org/10.3389/fgene.2020.00076

Nieto, B., Gaspar, S. G., Moriggi, G., Pestov, D. G., Bustelo, X. R., & Dosil, M. (2020). Identification of distinct maturation steps involved in human 40S ribosomal subunit biosynthesis. *Nature Communications*, *11*(1), 156. https://doi.org/10.1038/s41467-019-13990-w

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). *The complete sequence of a human genome*. 11.

Pink, R. C., Wicks, K., Caley, D. P., Punch, E. K., Jacobs, L., & Francisco Carter, D. R. (2011). Pseudogenes: Pseudo-functional or key regulators in health and disease? *RNA*, *17*(5), 792–798. https://doi.org/10.1261/rna.2658311

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121–132. https://doi.org/10.1038/nrg3642

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

Sohn, J., & Nam, J.-W. (2016). The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics*, bbw096. https://doi.org/10.1093/bib/bbw096

Strachan, T., Read, A. P., & Strachan, T. (2011). *Human molecular genetics* (4th ed). Garland Science.

Symonová, R. (2019). Integrative rDNAomics—Importance of the Oldest Repetitive Fraction of the Eukaryote Genome. *Genes*, *10*(5), 345. https://doi.org/10.3390/genes10050345

The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G. R., Steering committee, Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Appendix

K-mer frequencies in 1000G individuals. 176 individuals from different populations, sequencing coverage, 4 k-mer frequencies, and appropriate 5S gene cluster copy number. One k-mer is always 0 as expected.

Id	Population	Cov	k-mer 1	k-mer 2	k-mer 3	k-mer 4	copy nr 1	copy nr 2	copy nr 3	copy nr 4
HG00101	British	31	3465	0	0	0	112	0	0	0
HG00117	British	35	5256	0	0	0	150	0	0	0
HG00136	British	30	3498	0	0	0	117	0	0	0
HG00140	British	31	982	2397	0	0	32	77	0	0
HG00160	British	29	4793	0	0	0	165	0	0	0
HG00182	Finnish	31	3555	0	0	0	115	0	0	0
HG00187	Finnish	30	3972	0	0	0	132	0	0	0
HG00189	Finnish	26	3418	0	0	0	131	0	0	0
HG00252	British	30	4985	0	2	0	166	0	0	0
HG00271	Finnish	30	1276	1914	0	0	43	64	0	0
HG00278	Finnish	33	62	4041	0	0	2	122	0	0
HG00290	Finnish	40	6459	0	4	0	161	0	0	0
HG00308	Finnish	27	2442	0	0	0	90	0	0	0
HG00311	Finnish	30	2036	1107	0	0	68	37	0	0
HG00325	Finnish	27	1545	1734	0	0	57	64	0	0
HG00329	Finnish	32	3761	0	0	0	118	0	0	0
HG00358	Finnish	32	94	4582	0	0	3	143	0	0
HG00360	Finnish	29	2753	1337	0	0	95	46	0	0
HG00371	Finnish	33	2666	2529	0	0	81	77	0	0
HG00372	Finnish	29	1527	2645	0	0	53	91	0	0
HG00375	Finnish	31	3118	0	0	0	101	0	0	0
HG00403	Han	29	3966	0	0	0	137	0	0	0
HG00421	Han	31	1459	2946	0	0	47	95	0	0
HG00445	Han	33	1281	1965	0	0	39	60	0	0
HG00457	Han	30	3290	0	0	0	110	0	0	0
HG00459	Han	30	4179	0	0	0	139	0	0	0
HG00475	Han	30	2751	0	0	0	92	0	0	0
HG00524	Han	29	3146	0	0	0	108	0	0	0
HG00536	Han	27	3600	0	0	0	133	0	0	0
HG00556	Han	26	2456	0	0	0	94	0	0	0
HG00580	Han	31	3150	0	0	0	102	0	0	0
HG00628	Han	31	3512	0	0	2	113	0	0	0
HG00653	Han	46	5497	0	0	0	120	0	0	0
HG00707	Han	29	2859	0	0	0	99	0	0	0
HG00881	Chinese	32	3589	0	0	0	112	0	0	0
HG01051	Puerto	27	1571	1665	0	0	58	62	0	0
HG01054	Puerto	27	2067	828	0	0	77	31	0	0
HG01069	Puerto	45	62	2465	0	0	1	55	0	0
HG01072	Puerto	26	3366	0	0	0	129	0	0	0
HG01088	Puerto	31	2415	0	0	0	78	0	0	0

HG01097	Puerto	29	1115	1561	0	0	38	54	0	0
HG01101	Puerto	31	75	3942	0	0	2	127	0	0
HG01104	Puerto	28	4023	0	0	0	144	0	0	0
HG01110	Puerto	32	4254	0	0	0	133	0	0	0
HG01112	Colombian	30	2782	0	0	0	93	0	0	0
HG01124	Colombian	30	2894	0	0	0	96	0	0	0
HG01130	Colombian	29	3904	0	0	0	135	0	0	0
HG01133	Colombian	31	3812	0	0	0	123	0	0	0
HG01139	Colombian	31	3363	0	0	0	108	0	0	0
HG01142	Colombian	32	2001	867	0	0	63	27	0	0
HG01161	Puerto	30	3680	0	0	0	123	0	0	0
HG01164	Puerto	29	3859	0	0	0	133	0	0	0
HG01167	Puerto	42	2560	4532	0	0	61	108	0	0
HG01187	Puerto	29	2997	0	21	0	103	0	1	0
HG01190	Puerto	30	4475	0	0	0	149	0	0	0
HG01200	Puerto	30	3551	0	0	0	118	0	0	0
HG01253	Colombian	31	3591	0	0	0	116	0	0	0
HG01305	Puerto	29	1290	58	2677	0	44	2	92	0
HG01311	Puerto	31	3020	0	0	0	97	0	0	0
HG01325	Puerto	30	2514	0	0	0	84	0	0	0
HG01344	Colombian	31	1715	0	0	0	55	0	0	0
HG01402	Puerto	30	2932	15	942	0	98	1	31	0
HG01412	Puerto	29	888	837	0	0	31	29	0	0
HG01431	Colombian	30	3230	0	0	0	108	0	0	0
HG01494	Colombian	30	2658	0	0	0	89	0	0	0
HG01509	Iberian	27	2981	0	0	0	110	0	0	0
HG01512	Iberian	28	4328	0	0	0	155	0	0	0
HG01527	Iberian	30	3864	0	2	0	129	0	0	0
HG01530	Iberian	29	2454	1242	0	0	85	43	0	0
HG01565	Peruvian	36	5792	0	0	0	161	0	0	0
HG01571	Peruvian	31	1249	1670	0	0	40	54	0	0
HG01583	Punjabi	32	4193	0	0	0	131	0	0	0
HG01890	African	30	2255	72	1611	0	75	2	54	0
HG01892	Peruvian	29	554	1493	0	0	19	51	0	0
HG01920	Peruvian	29	1893	1570	0	0	65	54	0	0
HG01938	Peruvian	33	3789	0	0	0	115	0	0	0
HG01961	Peruvian	27	1551	1654	0	0	57	61	0	0
HG01974	Peruvian	27	1190	1645	0	0	44	61	0	0
HG02104	Peruvian	32	1566	2090	0	0	49	65	0	0
HG02134	Kinh	30	1251	1257	0	0	42	42	0	0
HG02141	Kinh	30	2811	0	0	0	94	0	0	0
HG02224	Iberian	30	3472	0	2	0	116	0	0	0
HG02253	Peruvian	30	2627	0	0	0	88	0	0	0
HG02373	Chinese	44	2736	3815	0	0	62	87	0	0
HG02409	Chinese	30	2601	0	0	0	87	0	0	0
HG02410	Chinese	30	3433	0	0	0	114	0	0	0
HG02470	African	30	61	1734	2453	0	2	58	82	0
HG02512	Kinh	32	3132	0	0	0	98	0	0	0

HG02521	Kinh	30	2978	0	0	0	99	0	0	0
HG02536	African	32	1664	2710	0	0	52	85	0	0
HG02681	Punjabi	29	1087	2302	0	0	37	79	0	0
HG02684	Punjabi	30	2815	3	0	0	94	0	0	0
HG02783	Punjabi	30	3249	0	0	0	108	0	0	0
HG02786	Punjabi	31	1293	1788	0	0	42	58	0	0
HG02789	Punjabi	26	3522	0	0	0	135	0	0	0
HG02982	Gambian	28	4513	19	15	0	161	1	1	0
HG03009	Bengali	28	2291	1593	0	0	82	57	0	0
HG03015	Punjabi	30	2628	0	2	0	88	0	0	0
HG03594	Bengali	28	2585	0	0	0	92	0	0	0
HG03644	Sri	27	3268	0	0	0	121	0	0	0
HG03660	Punjabi	33	3068	0	0	0	93	0	0	0
HG03680	Sri	27	1579	1163	0	0	58	43	0	0
HG03691	Sri Lanka	32	3493	0	0	0	109	0	0	0
HG03693	Sri	30	3497	0	0	0	117	0	0	0
HG03697	Sri	30	1743	1478	0	0	58	49	0	0
HG03702	Punjabi	32	3995	0	0	0	125	0	0	0
HG03708	Punjabi	37	3443	0	0	0	93	0	0	0
HG03716	Indian	28	1964	1651	0	0	70	59	0	0
HG03718	Indian	27	47	2850	0	0	2	106	0	0
HG03727	Indian	32	1360	1600	0	0	43	50	0	0
HG03767	Punjabi	27	1350	3064	0	0	50	113	0	0
HG03792	Indian	30	1031	2393	0	0	34	80	0	0
HG03812	Bengali	28	1227	847	0	0	44	30	0	0
HG03821	Bengali	29	5397	0	0	0	186	0	0	0
HG03824	Bengali	27	2368	0	0	0	88	0	0	0
HG03837	Sri	30	1041	2139	0	0	35	71	0	0
HG03846	Sri	29	3144	0	0	0	108	0	0	0
HG03848	Sri	29	1392	1442	0	0	48	50	0	0
HG03870	Indian	26	3214	0	0	0	124	0	0	0
HG03875	Indian	35	3792	2	3	0	108	0	0	0
HG03885	Sri	29	1344	1851	0	0	46	64	0	0
HG03900	Sri	29	2791	0	0	0	96	0	0	0
HG03917	Bengali	33	4800	3	0	3	145	0	0	0
HG03920	Bengali	27	1871	0	0	0	69	0	0	0
HG03941	Bengali	29	2060	0	0	0	71	0	0	0
HG03953	Sri	24	3743	0	2	0	156	0	0	0
HG03965	Indian	33	2813	4	0	0	85	0	0	0
HG03971	Indian	28	2214	0	0	0	79	0	0	0
HG03990	Sri	33	3622	0	0	0	110	0	0	0
HG04033	Sri	33	3696	0	0	0	112	0	0	0
HG04080	Indian	34	3105	0	0	0	91	0	0	0
HG04093	Indian	29	2938	0	0	0	101	0	0	0
HG04140	Bengali	33	3312	0	0	0	100	0	0	0
HG04158	Bengali	29	2364	0	0	0	82	0	0	0
HG04173	Bengali	31	3022	0	0	0	97	0	0	0
HG04182	Bengali	29	1339	3144	0	0	46	108	0	0

HG04225	Indian	36	2733	1691	0	0	76	47	0	0
HG04238	Indian	29	1720	0	0	0	59	0	0	0
NA11932	Utah	29	2599	1461	0	0	90	50	0	0
NA12282	Utah	31	464	1773	0	0	15	57	0	0
NA18534	Han	28	2927	0	0	0	105	0	0	0
NA18536	Han	27	2512	0	2	0	93	0	0	0
NA18612	Han	41	2797	0	0	0	68	0	0	0
NA18620	Han	32	3770	0	0	0	118	0	0	0
NA18749	Han	28	3939	0	0	0	141	0	0	0
NA18940	Japanese	30	2845	0	0	0	95	0	0	0
NA18943	Japanese	29	880	1241	0	0	30	43	0	0
NA18944	Japanese	28	3047	0	0	0	109	0	0	0
NA18948	Japanese	29	2702	0	0	0	93	0	0	0
NA18952	Japanese	31	2754	0	0	0	89	0	0	0
NA18953	Japanese	29	2680	0	0	0	92	0	0	0
NA18960	Japanese	31	2636	0	0	0	85	0	0	0
NA18961	Japanese	29	6750	0	0	0	233	0	0	0
NA18966	Japanese	32	2665	0	0	0	83	0	0	0
NA18967	Japanese	31	2954	0	0	0	95	0	0	0
NA18970	Japanese	36	4006	0	0	0	111	0	0	0
NA18971	Japanese	30	1732	0	0	0	58	0	0	0
NA18983	Japanese	27	2120	0	0	0	79	0	0	0
NA19079	Japanese	43	2997	0	0	0	70	0	0	0
NA19091	Japanese	32	3296	0	0	0	103	0	0	0
NA19655	Mexican	31	1080	2532	0	0	35	82	0	0
NA20520	Toscani	36	1658	1303	0	0	46	36	0	0
NA20527	Toscani	27	868	1061	0	0	32	39	0	0
NA20543	Toscani	30	3513	0	0	0	117	0	0	0
NA20758	Toscani	28	4941	0	0	0	176	0	0	0
NA20845	Gujarati	30	1777	2917	0	0	59	97	0	0
NA20911	Gujarati	28	4959	0	0	0	177	0	0	0
NA21091	Gujarati	30	3142	0	0	0	105	0	0	0
NA21094	Gujarati	41	4151	0	0	0	101	0	0	0
NA21100	Gujarati	27	2815	0	0	0	104	0	0	0
NA21117	Gujarati	27	2483	0	0	0	92	0	0	0
NA21118	Gujarati	29	3051	0	0	0	105	0	0	0
NA21123	Gujarati	31	2384	0	0	0	77	0	0	0
NA21124	Gujarati	29	2537	0	0	0	87	0	0	0
NA21127	Gujarati	35	2370	0	0	0	68	0	0	0
NA21133	Gujarati	38	3714	0	0	0	98	0	0	0

NON-EXCLUSIVE LICENCE TO REPRODUCE THESIS AND MAKE THESIS PUBLIC

I, Farid Naghiyev,

(*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,

5S rDNA copy number in WGS data,

(*title of thesis*)

supervised by Tarmo Puurand.

(*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 4.0, 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.

Farid Naghiyev

27/05/2022