DISSERTATIONES
BIOLOGICAE
UNIVERSITATIS
TARTUENSIS
453

TARMO PUURAND

Human genome studies with k-mer frequencies

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
453

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
453

TARMO PUURAND

Human genome studies with k-mer frequencies

UNIVERSITY OF TARTU
Press
1632

Institute of Molecular and Cell Biology, University of Tartu, Estonia

This dissertation is accepted for the commencement of the degree of Doctor of Philosophy in Bioinformatics on June 17, 2025, by the Council of the Institute of Molecular and Cell Biology, University of Tartu.

Supervisors: Prof. Maido Remm, PhD
Institute of Molecular and Cell Biology, University of Tartu
Tartu, Estonia

Lauris Kaplinski, PhD
Institute of Molecular and Cell Biology, University of Tartu,
Tartu, Estonia

Reviewer: Prof. Ants Kurg, PhD
Institute of Molecular and Cell Biology, University of Tartu,
Tartu, Estonia

Opponent: Prof. Kateryna Makova, PhD
Elbery College of Science, Pennsylvania State University,
PA 16802, United States of America

Commencement: Room No. 105, 23B Riia St., Tartu, on September 2, 2025, at 14.15.

The publication of this dissertation is granted by the Institute of Molecular and Cell Biology at the University of Tartu, Estonia.

This work was funded by Estonian Research Council Grant PRG2706; Estonian Research Council Grant TEM-TA35, 2021-2027.1.01.24-0627; Estonian Research Council institutional grant IUT34–11 and the EU ERDF grant No. 2014–2020.4.01.15–0012 (Estonian Center of Excellence in Genomics and Translational Medicine). Data analyses were partly carried out at the High Performance Computing Center of the University of Tartu, Estonia.

ISSN 1024-6479 (print)
ISBN 978-9916-27-929-8 (print)
ISSN 2806-2140 (pdf)
ISBN 978-9916-27-930-4 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

I **Puurand T**, Kukuškina V, Pajuste FD and Remm M. (2019). AluMine: alignment-free method for the discovery of polymorphic Alu element insertions. Mobile DNA, 10:31.

II Pajuste FD, Kalpinski L, Möls M, **Puurand T**, Lepamets M, Remm M. (2017). FastGT: an alignment-free method for calling common SNPs directly from raw sequencing reads. Scientific Reports, 7:2537.

III Örd T, **Puurand T**, Örd D, Annilo T, Möls M, Remm M and Örd T. (2020). A human-specific VNTR in the TRIB3 promoter causes gene expression variation between individuals. PLoS Genet., 16(8):e1008981.

IV Kaplinski L, Möls M, **Puurand T**, Pajuste FD, Remm M (2021). KATK: Fast genotyping of rare variants directly from unmapped sequencing reads. Human Mutation 42(6):777–786.

V Kaplinski L, Möls M, **Puurand T**, Remm M (2023). DOCEST-fast and accurate estimator of human NGS sequencing depth and error rate. *Bioinform Adv* 3(1):vbad084.

VI **Puurand T**, Möls M, Kaplinski L, Maal K, Krjutskov K., Salumets A., Kivisild T, Remm M. (2025). Y-mer: A k-mer based method for determining human Y chromosome haplogroups from ultra-low sequencing depth data. *Submitted to Genome Biology.* https://doi.org/10.21203/rs.3.rs-5042960/v2

The publications listed above have been reprinted with the permission of the copyright owners.

My contributions to the listed publications were as follows:

Ref. I Created study design of polymorphic Alu element discovery and detection in WGS data, preparing data and running analyses

Ref. II Selecting samples, dbSNP version and preparing k-mer lists

Ref. III VNTR motif copy number measuring in WGS data, participation in writing the original manuscript

Ref. IV Testing and validating KATK and GATK algorithms including read mapping, variant calling and comparison of the results.

Ref. V Variant calling with GATK

Ref. VI Created study design of using repeats in human chrY haplogroup prediction, preparing k-mer frequencies data and running analysis, writing the original manuscript

LIST OF ABBREVIATIONS

1000G	Thousand genome project
bp	base pairs in nucleotide sequence
cM	centimorgan
CNV	copy number variation
FASTQ	text-based format for storing a biological sequence and its corresponding quality scores
FLA	fragment length analysis
GATK	Genome Analysis Toolkit
HERV	Human endogenous retrovirus
HG	haplogroup
INDEL	insertions and deletions
ISCN	International System for Human Cytogenetic Nomenclature
k-mer	short, fixed length nucleotide sequence for computer-friendly calculations
LINE	Long interspersed nuclear element
MAF	minor allele frequency
MEI	mobile element insertion
NGS	next-generation sequencing
ONT	Oxford Nanopore Technologies PLC
PacBio	Pacific Biosciences Inc.
PAV	presence/absence variation (for mobile elements)
PCR	Polymerase chain reaction
PP	processed pseudogene
QC	quality control
RFLP	restriction fragment length polymorphism
SINE	short interspersed nuclear element
SNP	single-nucleotide variant/polymorphism
SRS	short reads sequencing
SRA	Sequence read archive
SV	sequence variants, i.e. duplications, deletions, inversions, translocations in larger scale
SVA	Sine-VNTR-Alu complex element
STR	short tandem repeat
tagSNP	single nucleotide polymorphism used to define and represent haplotypes in regions of linkage disequilibrium

TGS	third generation sequencing
TSD	target site duplication
VCF	variant calling format
VNTR	variable number of tandem repeats
WGS	whole genome sequencing

INTRODUCTION

Human genomes contain variations that had been suspected for a long time but only became visible to human eyes about 55 years ago with the introduction of chromosome staining using Giemsa dye and the comparative descriptions provided by the ISCN standard. Since the adoption of this staining method (also known as the banding method), progressively newer and better techniques have been developed. Each new method has contributed additional insights to our understanding of the distribution of various types of genetic variation among humans and human populations, as well as differences between healthy individuals and those affected by diseases.

Sequencing the entire human genome, along with its assembly, was once technologically impossible, and partial sequencing was prohibitively expensive. As a result, over the past 20 years, research focus shifted to DNA microarray-based technologies, which allow for the identification of many single nucleotide variations. These microarrays have been used to describe disease-associated loci and human ancestry through statistical analyses of variation combinations detected using these chips. Such calculations have relied on thousands of population-descriptive sequencing datasets generated using the short-read method. Today, sequencing data from over a million samples worldwide are available.

Currently, there is a paradigm shift in whole-genome sequencing from short-read to long- and ultra-long-read technologies. The short-read method cannot reliably detect tandem repeats or gene conversion-related repetitive sequences, which are characteristic of certain mutations. Additionally, aligning short reads presents challenges in poorly described genomic regions, though these can sometimes be analyzed using k-mer frequency methodologies if the regions carry biologically or locationally informative signals. While short-read sequencing generates a composite view of both sister chromosomes, it can still yield meaningful insights when the summed genotype aligns with the observed phenotype.

The advantages of long reads are particularly evident in the analysis of STR (short tandem repeats), VNTR (variable number tandem repeats), and CNV (copy number variation) patterns. Long reads provide both the exact number of repeats and the haplotype information for the chromosomal region, which are often difficult to reconstruct using population-based tagSNP combinatorics typical of short-read alignment. This limitation arises because copy numbers of repeats mutate independently in tandem arrays, and not all structural variations (SVs), such as translocations and mobile element insertions (MEIs), have been fully mapped.

Ultra-long-read technologies have further advanced the ability to determine sequences in the most technically challenging genomic regions, such as satellite sequences in heterochromatin and centromere areas. These regions, characterized by extremely high numbers of tandem repeats, were previously difficult to resolve. Now that the positions of satellite sequences in the genome have been mapped, short reads can become more informative due to fixed and non-fixed mutations

within these sequences. For example, the most common motif in the human genome, heterochromatic (TTCCA)n, is abundant on the Y chromosome and on chromosomes 1, 9, and 16. These sequences contain mutations dating back millions of years and can serve as markers for describing genetic variation, especially in cases where recombination between the heterochromatic regions of the two sister chromosomes has been minimal.

The first part of this thesis provides a historical overview of methods used to detect genetic variations over the past 55 years, tracing their evolution alongside advances in chemical, physical, and information technology. The second part explores the application of k-mer-based methods for detecting variation using short reads, where reads were analyzed without prior alignment to the human reference genome.

1. REVIEW OF THE LITERATURE

1.1. Brief history of DNA-based human variation discovery

During the early 20th century, human chromosomes remained largely unexplored, with estimates of chromosome count varying between 16 and 48 (Ferguson-Smith, 2015). Speculations surrounding the diploid cell chromosome count was definitively resolved in 1956, when Tjio and Levan demonstrated a count of 46 (Tjio and Levan, 1956).

One lingering question pertained to the size and composition of the human genome. The advent of chromosome banding techniques around 1969–70 marked a pivotal innovation in cytogenetics (Ferguson-Smith, 2015). These techniques revealed distinct darker and lighter regions along chromosomes, known as bands. Various banding methods emerged, with Q-banding being the first, relying on quinacrine mustard and quinacrine dihydrochloride to generate a fluorescent banding pattern. G-bands, produced with a Giemsa dye mixture, closely resembled Q-bands, while reverse-stained methods created R-bands. Additionally, C-staining targeted heterochromatin, T-staining focused on telomeres, and NOR-staining highlighted nucleolus organizing regions.

The subsequent query revolved around variability between sister and homologous chromosomes in closely related species. Studies utilizing Q- and C-band localizations in newborns revealed greater chromosomal variation in human populations than previously believed. Notably, chromosomes 3, 4, 13, 14, 15, 21, and 22 exhibited variable band intensities in specific regions. Pericentric inversions in chromosome 9 were also observed (McKenzie and Lubs, 1975). Comparisons of ancestral karyotypes across humans, chimpanzees, gorillas, and orangutans identified pericentric inversions as the most common type of chromosome rearrangement (de Grouchy *et al.*, 1972).

However, these initial studies were not exhaustive, prompting more detailed reviews (Mitchell and Gosden, 1978). Subsequent inquiries focused on phenotype determinations, such as identifying disease-causing variants and selecting variants for evolutionary studies.

Before the Human Genome Project, chromosome banding stood as the sole universal method for comprehensive genome analysis. Over the years, geneticists accumulated more data, enabling them to pose increasingly nuanced questions.

The culmination of these efforts came in the last decade of the 20th century with the Human Genome Project, aimed at creating a physical map of chromosomes with nucleotides precisely positioned in the reference genome with 1 bp accuracy. Chromosome painting offered an accuracy of 1 million bp, while linkage-based accuracy depended on recombination events, marker counts, and heterozygosity rates, with marker distances measured in centimorgans (cM).

Variations in the human genome, stemming from various mechanisms, vary in size and mutation rate. Despite advancements in NGS methods, which can sequence up to 4 million bp with a 15% error rate in a single read, challenges

persist in navigating the most complex tandemly repeated sequences in the human genome, although sophisticated error repair tools for ultra-long reads offer promise (Koren *et al.*, 2017).

1.2. Variations in the human genome

The human genome is dynamic, with variations originating from either germline or somatic cells. The mutation rate per cell division is likely similar for both types. Germline variations are more predictable because they are often shared among many people, and over time, more of these mutations are identified. In contrast, somatic mutations are primarily discovered through the study of disease-causing variations and typically affect only certain parts of an organ rather than the entire body. Variations can occur in coding sequences or in non-coding sequences, sometimes referred to as "junk" DNA. However, we do not yet know if these "junk" sequences have any significant function over a human's lifetime.

Variation may be classified on various bases, such as a) known biological meaning, b) variant length, c) whether bi- or multiallelic, d) the mechanism creating the variation. Variations may cover only 1 nucleotide or as many as millions of nucleotides. The most variable feature in everyone is the size of the genome.

The smallest variations in length include single-nucleotide substitutions, single-nucleotide deletions, and single-nucleotide insertions. The most widely studied variations in genetic research are single-nucleotide polymorphisms (SNPs). On average, the diploid human genome contains 3.53 to 4.65 million differences compared to the reference genome, depending on the population (The 1000 Genomes Project Consortium *et al.*, 2015; Byrska-Bishop *et al.*, 2022). At birth, each child carries approximately 45 new mutations, each with a length of 1 bp. Two-nucleotide-long insertions and deletions are less common but are readily identified and genotyped.

Less-studied variants are short tandem repeats (STR) and variable number tandem repeats (VNTR) because of short-read sequencing technology. In studies of 6,487 WGS samples, 366,013 polymorphic STRs were found (Shi *et al.*, 2023). STRs have a motif length of 2–6 base pairs (bp), and VNTR 7-100...150 bp, and STR analysis is used when they are located in euchromatin. Tandemly repeated sequences in heterochromatin are named satellite sequences, and their length together is impressive. Probably all tandemly repeated sequences in the human genome are multiallelic in their copy numbers, which means that the mutation rate is much higher than in biallelic variants: there should be a reason for this, and the likely reason is repeated sequences and motifs.

Very large mutations like complex chromosomal translocations and deletions are, in most cases, lethal. Inversions and translocations, however, probably affect only fertility in the next generation. Repeated sequences are transposable elements like Alu (Batzer and Deininger, 2002), Line, SVA, and PP or sequences which are not 'jumping' in nature, such as STR, VNTR, and CNV. Transposable elements are biallelic. A new Alu element insertion happens in every 21 live births, a Line

element in every 212 and SVA elements in every 916 (Xing *et al.*, 2009). Processed pseudogene (PP) insertion (Abyzov *et al.*, 2013) frequency has not been calculated, but there are numerous publications describing insertions not featured in reference genomes (also in the Pangenome section). Discovering pipelines found 55 new PPs in the 1KG project (2500 individuals) and 13 in SweGen samples (1000 individuals). Swedish samples found 806 insertions in individuals and only 4 of them (inc. HLA-DRB1, 3 cases) were found in 7 individuals (Ten Berk De Boer, Bilgrav Saether and Eisfeldt, 2023). One class of repeated sequences is satellite sequences, which are mainly located near telomeres or centromeres, both of which also contain repeated motifs with lengths of 5 or 6 bp. Satellite sequences in centromeres are tandemly repeated with motif lengths of approximately 171 bp (Suzuki, Myers and Morishita, 2020). Full-length sequences of centromeres with variable lengths have become available very recently, and the splitting time of old branches has been dated back 1.4 million years (Logsdon *et al.*, 2024). It is clear that every centromere evolves independently, as very few recombination events have been detected. Additionally, a 4.1x higher mutation rate has been found in centromeres compared to regions outside centromeres.

Mutations may occur when DNA polymerase incorporates a wrong nucleotide during synthesis, during double-strand break (DSB) repair, or in recombination events during germline cell divisions. As more and more samples are sequenced on a whole-genome scale, initiatives like the All of Us Research Program have identified over 1 billion sequence variants in 245,388 participants. Of these, 275 million variants were previously unreported, and 3.9 million of these have coding consequences (The All of Us Research Program Genomics Investigators *et al.*, 2024).

Mutation rate frequencies and size plots from 2006 were created before next-generation sequencing (NGS) platforms came into use, so frequencies were calculated from fractions of markers and individuals (Figure 1). The number of *de novo* mutations per birth depends on the ages of the parents, with an average of 35 mutations coming from the father and 10 from the mother (Goldmann *et al.*, 2016).

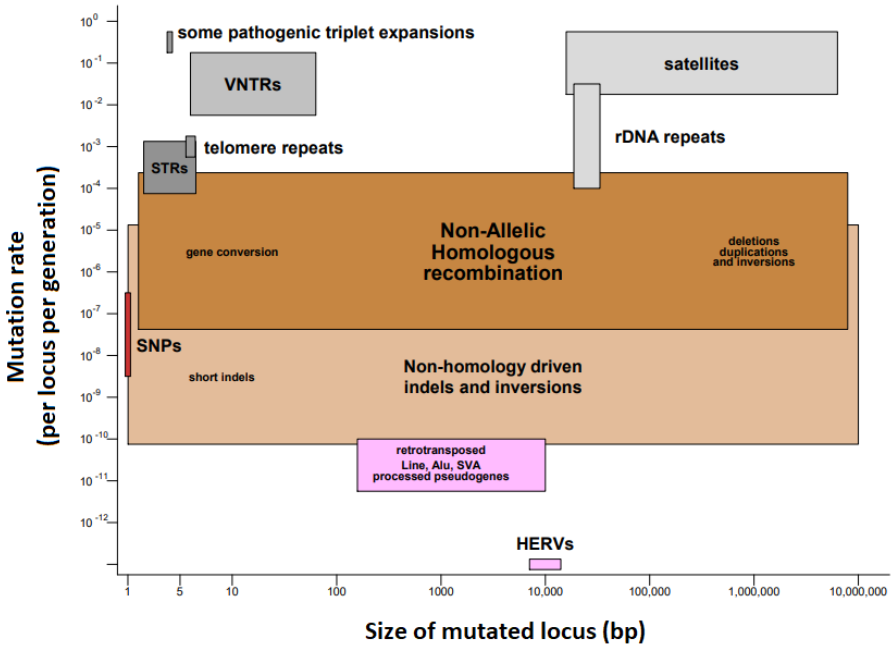


Figure 1. Different types of human variation differ in their length and mutation rate per live birth (modificated, Freeman, 2006). Structural variants (SVs) (Feuk, Carson and Scherer, 2006) are shown in the skin-colored areas. The pink-colored areas represent inserted variations that are located randomly in the human genome and were identified before the NGS era in somatic cancer samples. The red area represents SNPs, which are the markers most commonly used in human genome studies. The gray areas mark tandemly repeated sequences, which are less studied due to the limitations of sequencing technologies. New HERV insertions are extremely rare, and no scale number is available for HERVs.

1.3. Mapping of human chromosomes

A chromosome map serves as the cornerstone for unraveling and characterizing the intricacies of chromosome variability. Its construction draws upon a tapestry of historical knowledge, aimed at addressing specific inquiries by scrutinizing disparities within variable chromosome regions or the distances separating them. Various types of maps employ distinct methodologies: the cytogenetic map employs ideograms depicting G-banding patterns, while the recombination map relies on recombination frequencies among utilized markers (Figure 2). The radiation hybrid map, on the other hand, gauges the frequency of radiation-induced breaks between sequence-tagged sites (STS), and the physical map quantifies distances in nucleotides between studied loci.

These maps are interconnected with physical maps, wherein each marker possesses an exact or approximate position from the chromosome p-arm telomere. Platforms housing human genetic data, such as dbVAR, Ensembl, and UCSC

browsers (see WEB resources), prioritize the inclusion of cytogenetic bands alongside genetic information annotations. These annotations are crafted with the aid of the cytoband file format, an outgrowth of the endeavors following the creation of the human chromosomes physical map (Figure2, BAC Resource Consortium *et al.*, 2001; Wang and LaFramboise, 2019).

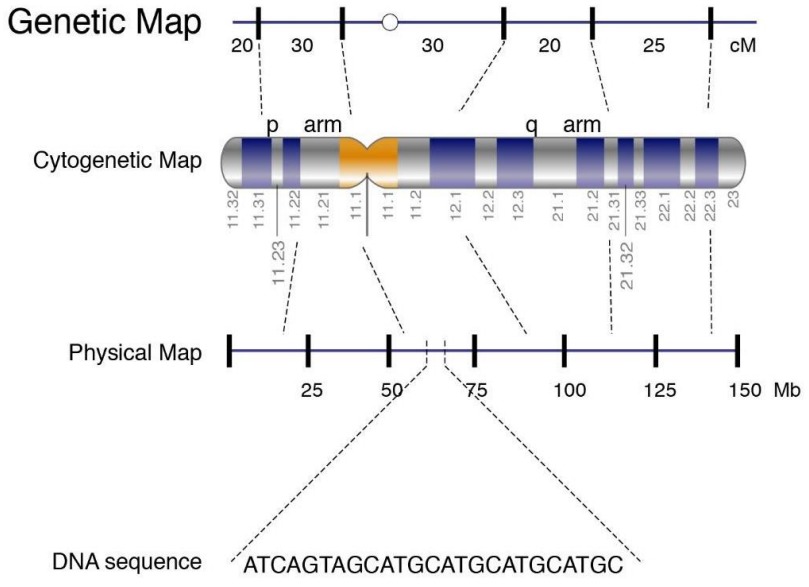


Figure 2. Genome size can be measured in centimorgans, chromosome band-length count or number of base pairs in length for all chromosomes.

1.3.1. Cytogenetic map

The cytogenetic map consists of standard, or reference, ideograms of chromosomes created using stained standard metaphase chromosomes. These chromosomes are visually examined under a microscope, revealing a continuous series of bands with no banded areas. Banding level standards, including 300, 400, 550, 700, and 850, are determined based on the number of bands in the haploid chromosome set of 22 autosomes and the X and Y chromosomes (see Figure 3). High-resolution banding techniques may be employed at different stages of the cell cycle, using methods that induce premature chromosome condensation. Bands ranging from 550 to 850 are typically sufficient for practical purposes, as they allow for the visualization of large variations through comparison.

Regions and bands are numbered from the centromere outward along chromosome arms. The band number should include the chromosome number, the arm symbol (p or q), the region number, and the band number within the region. Centromere numbers are designated as p10 and q10, although they are not presented on the map. Pericentromeric heterochromatin regions of all chromosomes, short arms of all acrocentric chromosomes, and regions such as 1q12, 3q11.2,

9q12, 16q11.2, 19p12, and 19q12 vary in length. Telomeres consist of 5 to 20 kilobases (kb) of tandem hexanucleotide $(CCCTTA)_n$ minisatellite repeats and stain darkly with T-banding (International Standing Committee on Human Cytogenomic Nomenclature *et al.*, 2020).

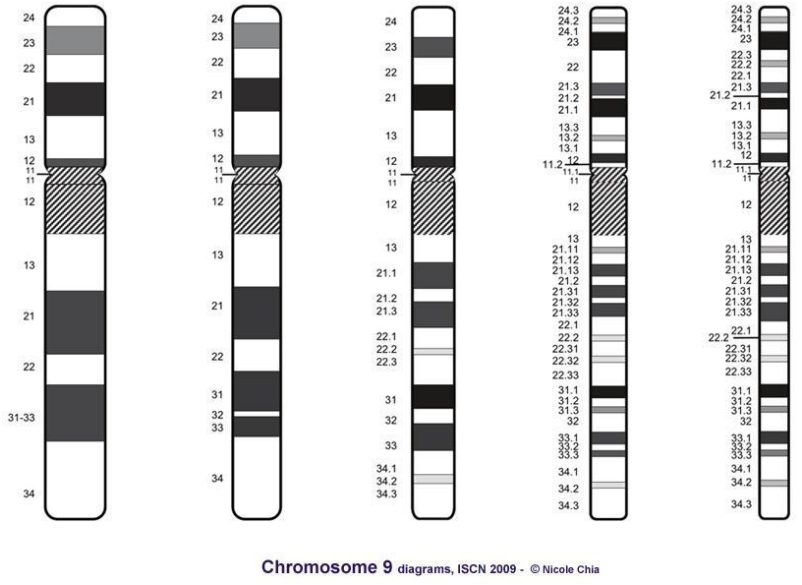


Figure 3. Banding pattern of human chromosomes ISCN 2020 (International Standing Committee on Human Cytogenomic Nomenclature *et al.*, 2020).

1.3.2. Genetic map

The genetic map is established based on the distances between two loci, defined as the anticipated number of crossovers occurring on a single chromosome between these loci, measured in Morgans. One centimorgan (1 cM) signifies one recombination event between two loci in a generation. The recombination rate describes the number of recombination events per one million nucleotides (cM/Mb). Regions with more recombination events typically exhibit higher recombination rates. The sum of all distances contributes to the total genome length (Figure 2).

Historically, the recombination map provided statistical insights into the minimal number of individuals required for mapping, particularly for Mendelian inherited diseases primarily caused by single genes. Constructing a genetic map necessitates a 3-generation family or families with a considerable number of siblings. The quality of the map depends on the count and informativeness of the markers used. In the case of multiallelic markers, informativeness stems from the rate of heterozygosity, indicating the average heterozygous calls per individual, whereas for biallelic markers, it comes from the minor allele frequency (MAF). In both scenarios, a higher value indicates greater informativeness.

The pattern of recombination rate is sex-specific (Kong *et al.*, 2002), population-specific (Spence and Song, 2019), breed-specific in cattle (Shen *et al.*, 2018), and individual-specific (Broman *et al.*, 1998; Dréau *et al.*, 2019). Creating whole-genome coverage of individual-specific recombination maps is costly, and it is even more expensive to repeat the process multiple times to ensure statistically validated experiments. Mice have had individual maps created from 27,710 recombinant molecules using the ReMIX method and Illumina sequencing technology.

The most recent accurate recombination map resolution achieved an average of 682 bp from 4.5 million crossover events, with a sex-averaged genome length of 3391 cM (Halldorsson *et al.*, 2019).

1.3.3. Physical maps

At the core of the human genome lies the chromosome physical map, a sequence of nucleotides extending from the telomere of the p-arm to the last nucleotide of the q-arm, with each base pair assigned coordinates counted from the end of the p-arm (Figure 2). The most prominent physical maps include the human reference genome, constructed during the Human Genome Project (International Human Genome Sequencing Consortium *et al.*, 2001), and the Celera assembly (Venter *et al.*, 2001), developed concurrently. Over the past 20 years, the reference genome created by the Human Genome Project has been refined for greater accuracy and reduced gaps. However, limitations in Sanger sequencing technology mean that sequencing reads with lengths of approximately 1000 base pairs are insufficient for constructing an ideal physical map of the human genome. Gaps resulting from repeated sequences are filled with “N” placeholders and lack biological annotations. Once the primary map is established, various population-specific physical maps emerge, albeit as clones with minor variations.

The Telomere-to-Telomere (T2T) consortium has invested six years in creating new human physical maps without gaps, primarily using sequencing methods such as Pacific Biosciences and Oxford Nanopore (Nurk *et al.*, 2022). The newly built reference genome CHM13 gives a first view of the most difficult satellite regions for sequencing and assembling through technological capture of centromere regions (Altemose, 2022). With established genome references, precise coordinates can be assigned to variations present in individual genomes for a given version.

1.3.4. Pangenomes

A pangenome refers to every genome constructed from multiple individuals, but the latest concept of the pan-genome suggests that when mapping Next-Generation Sequencing (NGS) reads, each read is integrated within graph-based pangenomes. This differs from using a standard reference genome, which functions as a single linear physical map with gaps. Pangenome graph construction starts with finding similarities in input genomes sequenced with long-read sequencing

methods and it is fundamental to add broad variability of indiviuals from different populations in that step. Pangenome graph increases the accuracy of variant calling and genotyping in resequencing projects because of improved mappability (Hickey *et al.*, 2024; Sirén *et al.*, 2024). Several pangenomes exist for humans, with one of the most ambitious projects compiling 350 phased, diploid genomes representing diverse global populations (Liao *et al.*, 2023).

For example, the addition of the first 47 genomes revealed 119 million base pairs (bp) of sequence and identified 1,115 gene duplications compared to the GRCh38 reference (Liao *et al.*, 2023). Similarly, 58 samples from 36 Chinese populations contributed 189 million bp and uncovered 1,367 gene duplications compared to GRCh38. Additionally, this effort identified 5.9 million small variants and 34,223 structural variants which were previously undocumented (Gao *et al.*, 2023).

In another instance, the Arab pangenome, derived from 43 individuals of various ethnicities, revealed an additional 101 million bp and 838 gene duplications not found in GRCh38 (Uddin *et al.*, 2023). Furthermore, the Aboriginal pangenome project adopted a unique approach, focusing on mutations based on variant types. They underscored the challenges of using different reference genomes by emphasizing variant analysis and the complexities involved (Reis *et al.*, 2023).

1.3.5. Gene mapping

Gene mapping results in what is also known as a gene map, but at a personal level. While the recombination map is population-based, the gene map may contain variations discovered from cytogenetic or molecular methods specific to one individual. Two main standards are used: ISCN for personal karyotype (International Standing Committee on Human Cytogenomic Nomenclature *et al.*, 2020) and VCF (Variant Call Format) for variants detected from next-generation sequencing data or microarray (Danecek *et al.*, 2011).

Since most human genomes are similar, the gene map only describes positions that differ from the reference. The ISCN standard describes differences from the normal karyotype or ideogram with broad examples of various mutations and arrangements (International Standing Committee on Human Cytogenomic Nomenclature *et al.*, 2020). For instance, a karyotype with Down syndrome due to trisomy 21 is described as 47, XY, +21, or with Robertsonian Translocation as 46, XX, der(13;21)(q10;q10), +21. Chromosome inversion is denoted as 46, XX, inv(9), (p11q13), and increased length of heterochromatin on the long arm of chromosome 16 as 46, XX, 16qh+. Notably, significant differences in banding are often associated with phenotypes, such as in sub-fertile couples with karyotypes 46, XX, 16qh+ and 46, XY, inv(2)(p11q13),inv(9)(p11q13) (Srebniak *et al.*, 2004).

In contrast, a VCF file is a text file containing meta-information lines, a header line, and data lines, with information on the position of the reference sequence(s) (Danecek *et al.*, 2011). It primarily displays biallelic SNP and short INDEL data,

while other mutations like multiallelic STR or VNTR and SV are more complex for visual inspection. The VCF file is the final step of the genotyping pipeline in WGS or microarray signal intensity processing, combining variations found in millions of positions into a single file. During post-processing, variant positions with low quality are filtered out, and the entire process is automated and standardized.

While the VCF file provides variation information in fine detail, NGS methods, except for Oxford Nanopore, have limitations compared to ISCN, especially in regions with repeated sequences. Even different assemblies yield different sizes, as seen with band 6q21 (Liehr, 2021). The ISCN2020 standard has introduced microarray-based nomenclature, such as arr(X,1-22)x2 for normal females and arr(X,Y)x1,(1-22)x2 for normal males. Loss of a chromosome region is described as arr[GRCh38]4q32.2q35.1(163,146,681_183,022,312)x1, and sequence-based nomenclature for duplication as seq[GRCh38]dup(8)(q24.21q24.21)NC_000008.11:g.128746677-128749160dup (International Standing Committee on Human Cytogenomic Nomenclature *et al.*, 2020).

1.4. Molecular methods for describing variations in the human genome

Each of the main methods for discovering and describing variants in the human genome throughout time has its own scale and resolution (Table 1). Next-Generation Sequencing (NGS) provides comprehensive coverage of the genome, enabling the detection of various types of variants, such as Single Nucleotide Variations (SNPs), short Insertions and Deletions (indels), structural variants, and Copy Number Variations (CNVs). Microarrays are suitable for detecting previously known mutations and population-specific haplotypes linked to them. Painting is particularly useful for identifying large-scale *de novo* mutations that are not present in pangenomes and chromosome-copy number changes. NGS Ultra-long Reads can identify large-scale *de novo* mutations due to their ability to generate long sequencing reads, offering a more comprehensive view of the genome.

Table 1. Main methods for discovering and describing variants with mutation length, resolution and detected variant types.

Method	Scale	Resolution	Variants
Painting	Genome	10Mb	+10Mb SV
Sanger sequencing	1kb	1bp	SNP, STR, VNTR, breakpoints for SV
PCR (+Sanger or gel)	1kb to 25kb	1bp to 1kb	SNP, STR, VNTR
FLA	25kb	Motif copy	STR, VNTR
Microarrays	1bp	1bp	SNP, CNV 0-4x
NGS	Genome	1bp	All variants

There are other methods for chromosome-scale variation detection, such as Bionano Inc.'s optical genome mapping (OGM) and Fluorescence In Situ Hybridization (FISH), but they are like microarray detection on the basis of a reference genome with appropriate painting.

1.4.1. Chromosomes painting

In 1971, Seabright developed a chromosome banding method where, after a brief treatment with trypsin, chromosome preparations are stained with Giemsa stain (Figure 4). The preprocessing of chromosomes involves a 72-hour culture of peripheral blood cells. Groups of 12 to 30 mitotic chromosomes are studied under the microscope and described according to the ISCN (International Standing Committee on Human Cytogenomic Nomenclature et al., 2020) standards This process is routinely used to detect cancer and chromosomal abnormalities. In the Human Pangenome Project (Liao *et al.*, 2023), cell lines are monitored for mosaicism, and relevant data are documented in the Coriell Institute's sample descriptions.

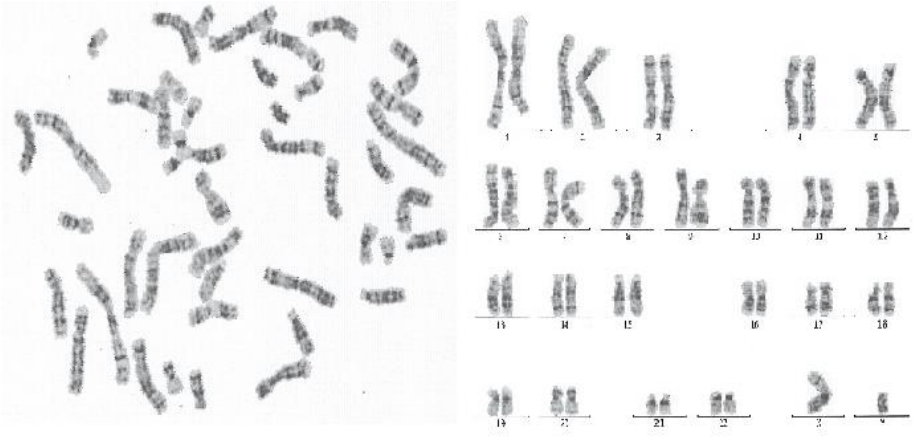
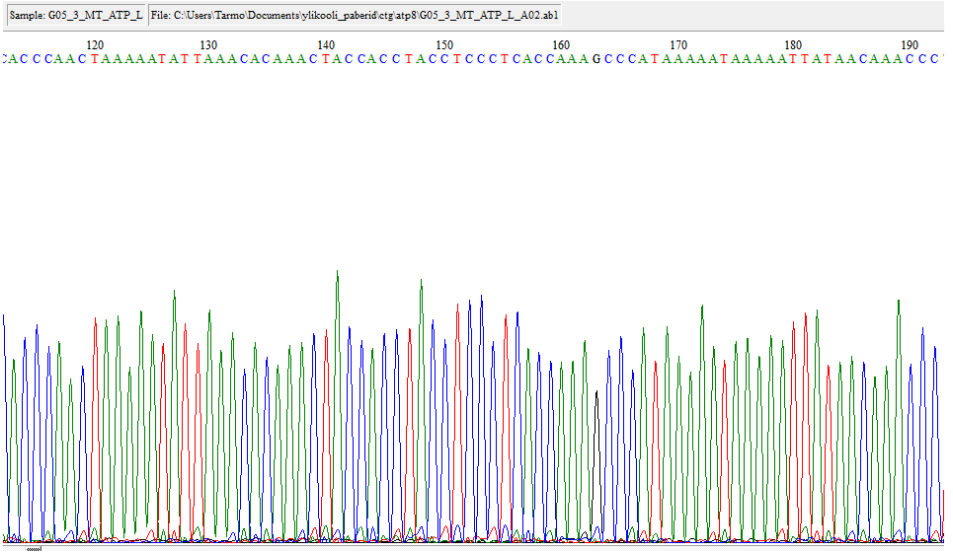


Figure 4. Chromosome painting results are karyotypes based on 20 chromosome clusters, evaluated visually (A, one cluster) and after collecting them in order (B). The author's karyotype is 46,XY in the figure.

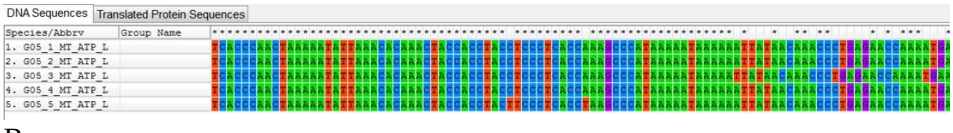
Chromosome banding and karyotyping are time-consuming processes, and automation is increasingly replacing manual efforts for chromosome grouping (Gregory and Maher, 2010). Recent advancements in AI have the potential to streamline this work further, allowing lab workers to focus on verifying the results instead of manually grouping chromosomes, ranging from 12–30 to all chromosome sets (Bokhari *et al.*, 2022).

1.4.2. Sanger sequencing

The initial Sanger sequencing method uses restriction fragments from ϕX174 DNA as a primer and radiolabeled dideoxynucleosides in four parallel channels within an acrylamide gel (Sanger and Coulson, 1975). After its commercialization by Applied Biosystems in 1986, Sanger sequencing became the most widely used sequencing method. It played a crucial role in the Human Genome Project. Even today, Sanger sequencing is necessary for evaluating certain mutations detected by NGS platforms (Figure 5), which are not suitable for PCR. The Sanger sequencing method is based on incorporating fluorescently labeled ddNTPs into a dNTP mix, and detecting them in the order of the template, primarily using capillary electrophoresis.



A.



B.

Figure 5. Sanger sequencing graphical views for ATP8 gene fragment. A. The sequencing output consists of fluorescent dye signal intensity graphs for four channels, as detected by the sequencer. Each channel corresponds to one of the four nucleotides (A, T, C, G), and the intensity of the fluorescent signals represents the presence of each nucleotide in the sequence. B. The BioEdit program provides a visual representation of the sequencing data, highlighting differences between individuals. Each nucleotide is color-coded, making it easy to identify and compare variations in the sequences. Experiment design, analysis, and graph by author, 2016.

1.4.3. PCR with agarose gel

With PCR and gel electrophoresis, the length of PCR products can be detected. For size detection, a DNA ladder is used to estimate the size of the products. Differences in product size can arise from several factors:

- **Restriction Fragment Length Polymorphism (RFLP)**: This method detects SNP alleles that affect the cutting sites of restriction enzymes.
- **Detection of Deletions**: Deletions up to 5 kilobase pairs (kb) can be detected using the same PCR primer pairs (Figure 6).
- **Copy Number Variations**: Differences in the number of Short Tandem Repeats (STR) and Variable Number Tandem Repeats (VNTR) motifs.
- **Primers Near Deletion Breakpoints**: Using PCR primers close to larger deletion breakpoints can identify significant deletions.

PCR is commonly used to validate the results of next-generation sequencing (NGS) projects. It is especially useful for sampling a limited number of polymorphic loci when manual evaluation is feasible.

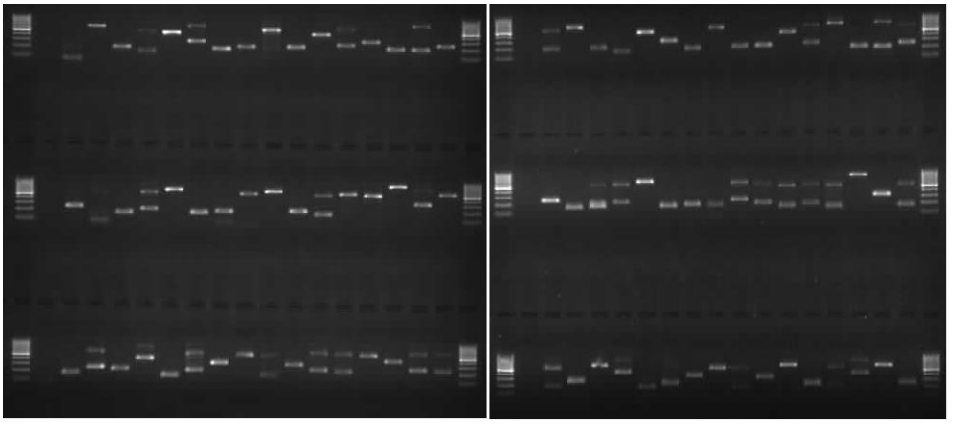


Figure 6. Using PCR, the presence of the Alu element can be detected by comparing the lengths of PCR products in two individuals. Size markers and negative control are included in the analysis. An upper band indicates the presence of the Alu element. A lower band indicates the absence of the Alu element in a homozygous state. The presence of two bands, with a size difference of approximately 300 bp, indicates a heterozygous state. This experiment and the accompanying photographs were conducted by Kadri Maal in 2019.

1.4.4. Fragment length analysis

Fragment length analysis allows the separation of DNA fragments by size, which can then be compared with a standard size marker included in every run (Figure 7). Various instruments are used for this purpose, and multiple samples can be

analyzed simultaneously if their DNA fragment sizes allow for differentiation. This method is particularly effective for typing STR and VNTR copy numbers. It is preferable from gel electrophoresis when the repeated motifs are shorter, and the copy number variations are within a lower range. For instance, the DRD4 and MUC20 VNTRs are located in coding regions, with the copy numbers of their repeated units being 2–11 times 48 bp and 2–6 times 57 bp (Lichter *et al.*, 1993; Higuchi *et al.*, 2004).

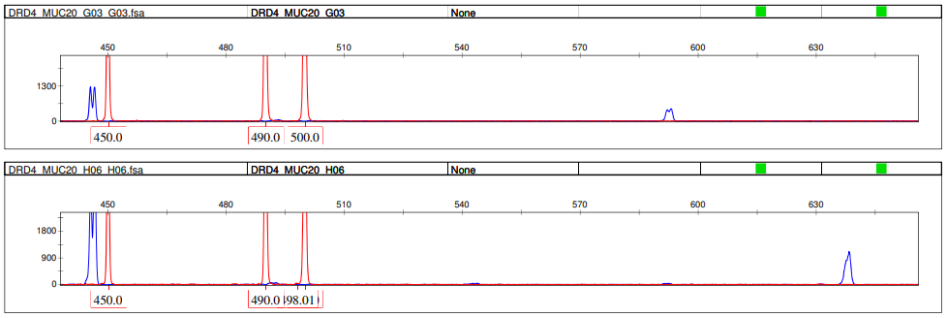


Figure 7. The upper graph illustrates DRD4 alleles with a 3/6 combination (blue peaks), while the lower graph shows a 3/7 alleles combination (blue peaks). Red peaks represent the size markers used for reference. The MUC20 variant detection is not included in this analysis. The experiment design, analysis, and graph were created by Maris Laving in 2018.

1.4.5. Microarrays

Microarrays are primarily used to enhance the power of genome-wide association studies (GWAS), identify disease-causing variants, and conduct comparable analyses in cancer-control hybridization profiles. In this context, we focus on arrays that detect only one nucleotide in a single run (Figure 8). This detection involves identifying the nucleotide adjacent to the probe attached to the solid surface. Over the past 20 years, significant technological advancements have enabled the genotyping of over 4 million scientifically relevant positions in the human genome. A comparison of 28 commonly used microarrays revealed that up to 60% of known CYP2D6 alleles, almost 100% of Class I HLA alleles, but only 50% of Class II HLA alleles can be accurately called after imputation (Verlouw *et al.*, 2021).

In principle, a k-mer functions similarly to an attached probe on an array, where the signal intensity corresponds to the k-mer frequency. The k-mer frequency provides more “bits” of information, especially after signal oversaturation in the scanner, since the light for each spot is consistent during the run. Differences in microarray signal intensities allow for the prediction of chromosome region copy numbers, ranging from single-copy region deletions to duplications of both chromosome regions, i.e., 0-4x (Emanuel and Saitta, 2007).

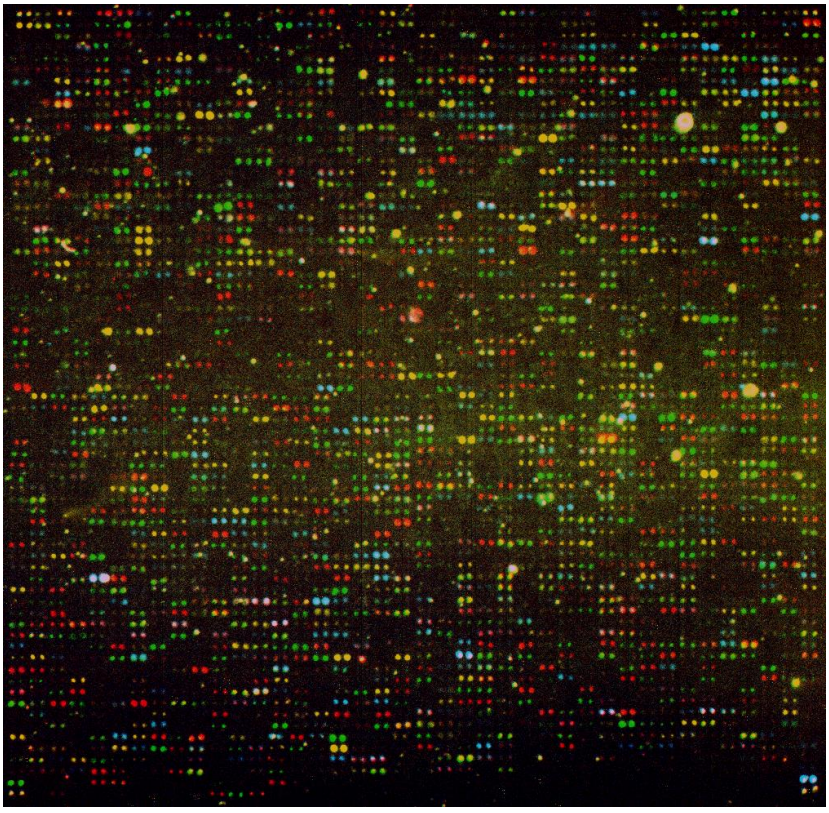


Figure 8. The pseudo-colored representation of chromosome 22 using the Asper Biotech's AS APEX microarray by Genorama employs four channels to indicate nucleotide presence: A (yellow), C (red), G (green), and T (blue). Signals are duplicated, and data from both DNA strands are utilized. Heterozygous signals appear as a mixture of colors, reflecting the presence of different nucleotides at that position. This visual representation combines wet lab PCR work with image processing, as conducted by the author in 15th of September 2000.

1.4.6. Next-generation sequencing

Next-generation sequencing (NGS) refers to advanced sequencing methods that arose following the commercialization of high-throughput capillary Sanger sequencing systems. These methods are categorized into two generations based on read length: second generation, known for producing short reads, and third generation with long-read technologies.

Illumina NGS exemplifies second-generation sequencing and has become the predominant technology in use today. It enables sequencing of DNA fragments up to 250 base pairs in length from both ends of inserts. The workflow is highly efficient and user-friendly, supported by optimized kits tailored for diverse applications. While Illumina platforms generate short reads, paired-end sequencing can facilitate the detection of insertions and deletions within reads, albeit longer

structural variations may be overlooked. The output of sequencing instruments, measured in base pairs per run, varies depending on instrument specifications, and there are discrepancies across different models (Reuter, Spacek and Snyder, 2015).

PacBio represents a third-generation NGS technology that employs single molecule real-time sequencing detection. Historically, until 2019, it utilized Continuous Long Reads (CLR), characterized by a high error rate of around 15%. These errors were primarily due to indels resulting from variations in polytrack lengths. Since 2019, PacBio has transitioned to HiFi (High-Fidelity) technology. This advancement involves sequencing the same DNA insert, averaging 15 kilobase pairs in length, multiple times using special adapters. This approach generates a consensus circular sequencing read with an error rate close to zero, significantly enhancing accuracy compared to previous CLR methods. In addition to its improved accuracy, PacBio HiFi technology also provides methylation information, offering researchers valuable insights into epigenetic modifications alongside sequence data.

Oxford Nanopore Technologies (ONT), another third-generation sequencing technology, distinguishes itself with its remarkable sequencing capabilities that do not rely on DNA polymerases and DNA labeling. Despite having an error rate similar to PacBio CLR reads, ONT can generate reads exceeding 4 million base pairs in length. This is made possible by employing duplex-read technology, where sequencing of one DNA strand is followed by sequencing of the same insert on the complementary strand. This approach reduces the error rate to approximately 0.5%, while achieving a maximum read length of 400 kilobase pairs. ONT's strengths are particularly evident in applications requiring *de novo* assembly of species with large genomes and high repeat consistency at the chromosome level. Additionally, ONT technology holds significant potential for real-time detection of DNA and RNA modifications; it can distinguish between nucleotides with and without modifications, providing insights into epigenetic variations. Moreover, ONT platforms facilitate variant calling during sequencing, enabling rapid and confident diagnosis in clinical settings. This capability enhances the utility of ONT for diverse research and diagnostic applications, leveraging its long read lengths and advanced detection capabilities.

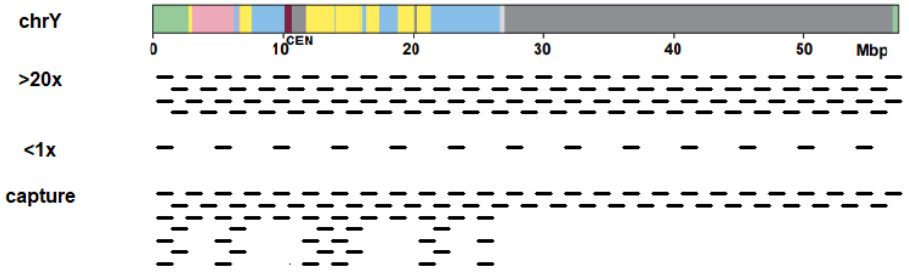


Figure 9. In human NGS sequencing, three main strategies are commonly employed: high coverage sequencing, low coverage sequencing, and capture-based sequencing.

High Coverage Sequencing (>20x): This strategy involves sequencing the genome at a high depth, meaning each base pair is sequenced multiple times. High-coverage sequencing provides comprehensive and accurate data, particularly useful for detecting variants with high confidence. It ensures robust coverage of both unique and repetitive regions of the genome (Figure 9).

Low-Coverage Sequencing (<1x): Conversely, low coverage sequencing involves sequencing the genome at a lower depth, with each base pair being covered fewer times. This approach is more cost-effective and faster than high coverage sequencing but may sacrifice accuracy and sensitivity, especially for detecting rare variants or structural variations (Figure 9).

Capture-Based Sequencing: This method involves enriching specific genomic regions of interest before sequencing. In the context of chrY chromosome capture, for instance, only the targeted regions are sequenced, which allows for precise calculation of sequencing depth in those regions. Sequences of a repetitive nature are typically excluded from capture-based approaches, ensuring a focused analysis on specific genomic loci without interference from repetitive elements (Figure 9).

Long-read methods enable hidden variants to be found when previous standard sequencing platforms do not work because of complex repeated sequences (Fadaie *et al.*, 2021).

1.5. Software for analysis of NGS data

In the analysis of sequence reads, the choice of software often depends on whether researchers are discovering new sequences or already know the origin of the reads. The length of the reads and overall length of the genome are critical factors influencing this decision.

Discovering New Sequences (*De novo* Assembly): *De novo* assembly is preferred when the reads are long enough (typically hundreds to thousands of base pairs). This method constructs the entire genome or transcriptome sequence from scratch without relying on a reference sequence. Long reads simplify the assembly process by bridging repetitive regions and complex genomic structures more effectively.

Mapping to a Reference Genome: When reads are shorter, typically up to 150 base pairs, mapping to a reference genome or transcriptome is commonly employed. This approach aligns the short reads to a known reference sequence to determine where they originated from. Mapping allows researchers to identify genetic variations, such as single nucleotide polymorphisms (SNPs) and small insertions/deletions (indels), relative to the reference sequence.

1.5.1. *de novo* assembly

De novo assembly involves constructing longer nucleotide sequences by piecing together overlapping shorter sequences. Several software programs are designed for this purpose, each with its own approach and capabilities. For short reads, notable assemblers include SPAdes (Bankevich *et al.*, 2012) and Velvet (Zerbino and Birney, 2008). Velvet, initially developed during the early days of NGS, when reads were typically 25–50 base pairs, utilizes de Bruijn graphs. In simulations with prokaryotic data, Velvet achieved an N50 length of up to 50 kb, while on simulated mammalian Bacterial Artificial Chromosomes (BACs), it reached a 3 kb N50 length. As NGS read lengths have increased, newer assemblers have demonstrated improved performance in achieving gapless *de novo* assemblies of mammalian genomes.

A significant challenge in *de novo* assembly arises from tandemly repeated sequences found in diploid eukaryotic genomes. To overcome this hurdle, it is crucial to use reads that are as long as possible and derived from homozygous genomes. An exemplary application of this approach is seen in assembling the human cell lines CHM1 and CHM13 centromeres, where high-coverage (66x PacBio and 98x ONT) long reads were generated (Logsdon *et al.*, 2024). For chromosomes 2, 7, 9, and 20, the hifiasm (see WEB resources) assembler was utilized, achieving comprehensive assembly. For other chromosomes, a strategy involving Singly Unique Nucleotide k-mers (SUNKs) was employed to barcode PacBio reads and link them with ultra-long ONT reads carrying similar barcodes, facilitating their bridging in assembly processes.

1.5.2. Mapping the reads to the reference genome, calling variants, visualizing

After completing a *de novo* assembly to create a reference genome for a species, subsequent sequencing projects often involve mapping new sequencing reads onto this assembly. This reference genome serves as a baseline for comparison and further analysis.

One of the most widely used pipelines for mapping and variant calling is the Genome Analysis Toolkit (GATK). GATK is particularly optimized for processing data from short reads generated by platforms like Illumina. Here's how the GATK pipeline typically operates (Figure 10):

1. **Mapping**: Individual sequencing reads (in FASTQ format) from each sample in the project are aligned or mapped to the reference genome using alignment tools such as BWA or Bowtie.
2. **Variant Calling**: After mapping, variants (e.g., SNPs, indels) are identified by comparing the mapped reads to the reference genome. GATK employs sophisticated algorithms to accurately call variants while considering factors like read quality and mapping ambiguity.

3. **Joint Calling**: Variants from all samples in the project are jointly analyzed and called together. This approach enhances the accuracy of variant calling by leveraging information across multiple samples, improving the detection of variants that may be present in only a subset of samples.
4. **Quality Control (QC)**: Variants called by GATK undergo stringent quality control checks to ensure reliability. Only variants that meet predefined quality criteria are included in the final Variant Call Format (VCF) file.
5. **VCF Output**: The final output of the GATK pipeline is a VCF file containing the list of variants detected across all samples, along with associated quality scores and other relevant information.

Overall, GATK's robust pipeline facilitates comprehensive variant calling and analysis, making it a cornerstone tool in genomics research, particularly for projects using short-read sequencing data from platforms like Illumina.

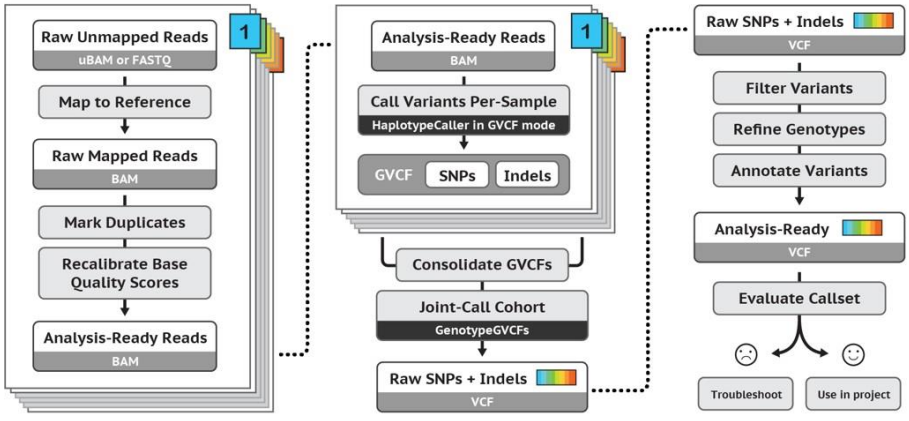


Figure 10. GATK4 Pipeline contains numerous steps, picked with QC-s. Input is FASTQ files and output is VCF file (WEB resources, GATK4, pipeline section).

As Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) are distinct sequencing platforms, their mapping and variant-calling pipelines are tailored to accommodate their unique characteristics. ONT sequencing data analysis tools and pipelines are often consolidated and accessible through platforms like EPI2ME (see WEB resources). These platforms offer integrated solutions for base calling, mapping, and variant calling, streamlining the workflow for ONT users. For PacBio sequencing data, the HiFi-human-WGS-WDL pipeline (see WEB resources) on GitHub provides a comprehensive set of tools and workflows. These pipelines are specifically optimized for PacBio's HiFi sequencing technology, which focuses on generating highly accurate long reads. Both ONT and PacBio sequencing platforms commonly use minimap2 for mapping long reads to a reference genome. Minimap2 is well-suited for handling the unique characteristics of long reads, including higher error rates and structural

variations. Visualization of complex genomic data from long reads is crucial for interpretation. Techniques such as haplotype coloring, as described by De Coster and Rademakers (De Coster and Rademakers, 2023), enable researchers to visualize haplotype differences within populations by using different colors, as do tools that display gene orientations in complex genomic regions, like principal bundle composition (Chin *et al.*, 2023).

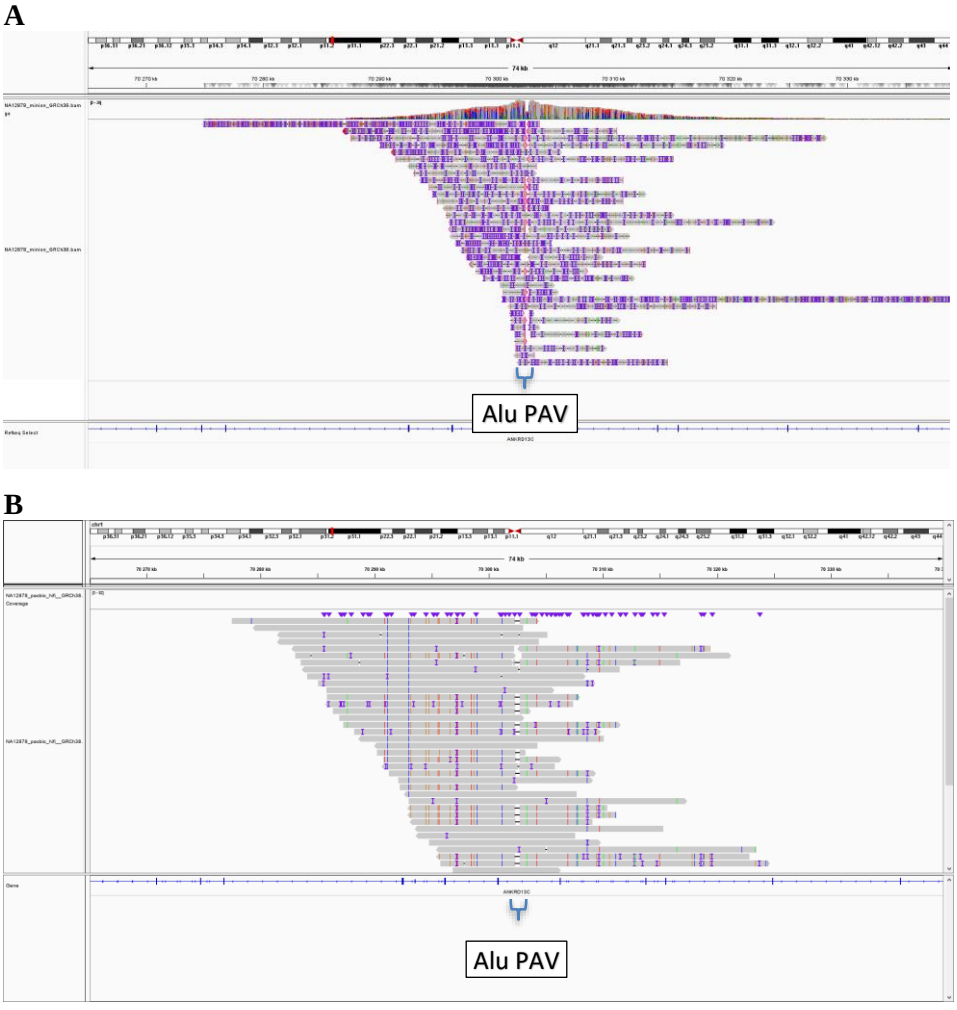


Figure 11. Integrated Genome Viewer (IGV) (Robinson *et al.*, 2011) alignment view for Alu element presence in heterozygous state in sample NA12878 in positions chr1:70302444-70302752 (reference GRCh38). ONT (A) and Pacbio (B) IGV view. Old method from PacBio, Continuous Long Reads (CLS) is equal to ONT, but after improvements in PacBio sequencing method, PacBio has fewer mismatches than ONT when comparing to reference (lilac bars). Note that haplotypes with Alu insertion have fewer SNP alleles not in human reference genome. Figures created by Kadri Maal.

1.5.3. Graph based pangenomes

The pangenome is anchored to a reference genome because the broad types of variants need to have physical locations. There needs to be an evaluation method to determine the minimal set of individuals such that no reads from new individual sequencing data fail to find a location during mapping. The construction of graph-based human pangenomes and associated analysis tools is currently in an intensive development phase. Both high- and low-level visualization programs help in understanding sequences represented in the pangenome, as well as those from mapped individuals not included in the pangenome graph. (Sirén *et al.*, 2024; Secomandi *et al.*, 2025).

A graph based human pangenome "chm13-90c.r518.gfa" file is constructed with the minigraph-cactus program and contains 90-chromosome SV variability, including chm13v2.0 as the reference (Hickey *et al.*, 2024). Each segment has code as a descriptor in Graphical Fragment Assembly (GFA) format (see WEB resources). The individual NA12878, sequenced with ONT technology and mapped to the pangenome chm13-90c.r518.gfa has the same Alu element PAV (Figure 11, Table2) represented in sequence read order ">s6877>s6878>s6879", and without Alu ">s6877>s6879". "s6878" is the Alu element sequence code and ">s6877", ">s6879" are flanking sequence codes (Schloissnig *et al.*, 2024). The pangenome graph built with minigraph for HPRC year-1 samples (Li, 2022) was used to map the ONT reads with minigraph.

Table 2. First 6 columns marks ONT reads for sample NA12878 (WEB resoursces), standard individual in methods comparisons, mapped to chm13-90c.r518.gfa parameters, read ID, read length, mapping coordinates, orientation and pangenome sequences graph codes order. Codes order indicates presence and absence of Alu element PAV (s6878) and orientation.

readID	read length	mapping coordinates		orien-tation	graph codes order
82fb86ee-17bc-48ca-9c53-1d2ec7ebf37e	56598	20092	56585	+	>s6877>s6879
2025e707-6c41-4e22-aa7b-cc480097f1ff	60178	22	60165	+	<s6879<s6877
f4a12c51-dbba-4187-b701-c619e56621be	30125	10	18338	+	<s6879<s6877
494229a4-f178-4fec-a329-a3c92393ec0d	61147	1	31658	+	>s6877>**s6878**>s6879
494229a4-f178-4fec-a329-a3c92393ec0d	61147	31664	61133	+	<s6879<**s6878**<s6877
cbb2a91b-e7dd-4e33-9be7-830d78a96017	7924	825	7895	+	<s6879<**s6878**<s6877
8cc9b583-6dde-4cf6-b584-a780e703cdc7	10695	197	10686	+	>s6877>**s6878**>s6879

1.6. Alignment-free analysis approaches for different variant types

We can categorize genetic approaches based on their use of alignment into three categories:

1. No alignment information is used.
2. Previously aligned information, such as reference sequences, is used to set up alignment-free applications.
3. All analysis is based on aligned sequences

In modern research, developing fully alignment-free approaches is challenging because most technologies are geared towards alignment-based studies, which offer simpler unification. However, fully alignment-free analysis allows for the investigation of consistencies and frequencies in a sequence. Before the NGS era, there was little data available for frequency analysis, so statistical approaches mainly focused on sequences with consistent differences, and the number of sequences studied was limited (Ren *et al.*, 2018). Consequently, most current approaches are only partially alignment-free and use substring frequencies to validate results with statistical tools. Alignment-based methods are not the focus of this thesis; they are only used for preliminary data creation and comparison of results.

1.6.1. k-mers, their length and frequencies

Alignment-free analysis primarily relies on the detection of k-mer presence, with a more advanced approach using the frequencies of these k-mers for comparison in studies.

k-mers are widely used with different applications under different names: BLAST-searching seed is the same string (Altschul *et al.*, 1990). k-mers are used to create *de novo* assemblies of genomes (Shi and Yip, 2020).

Massive Illumina-based parallel sequencing gives the possibility to quantify k-mers, which should be evenly distributed over all regions in the diploid genome. The distribution of k-mers frequencies gives us an indication of genome size (Sun *et al.*, 2018), historic genome duplications (Daccord *et al.*, 2017), hybrid-based heterozygosity (Mixão and Gabaldón, 2020), mixed genomes and even symbiotic or parasitic species living in community (Kumar and Blaxter, 2011). Quantification gives indicative information of the hardest-to-describe sequences in the genomes, high-identity regions with a high copy number, like centromeres (Arora *et al.*, 2021). Therefore, k-mers in sets or independently may have signatures in certain applications.

k-mer counting becomes more important with growing data volumes to be expanded. In the era of limited sequences per individual, k-mer frequencies were not important; it was only necessary to have enough overlapping nucleotides between reads to construct longer contigs (overlapping sequences). Then the human reference genome was published, when k-mer frequencies were used to

count PCR primer binding sites in the genome to avoid unwanted PCR products (Andreson *et al.*, 2006). With short-read sequencing methods, numerous k-mer counting programs soon became available, such as Jellyfish (Marçais and Kingsford, 2011), KMC (Deorowicz *et al.*, 2015), and glistmaker (Kaplinski, Lepamets and Remm, 2015). k-mer counting is a simple thing: no gaps or mismatches are used during counting. Querying the k-mer counts for different applications is fast and allows mismatches. However, bacterial genomes are small and even 100x sequencing depth for a 3 MB genome is still only 10% of a 1x human genome. Different types of data structures are used to overcome difficulties with memory, hard disk speed and data transfer (Wang *et al.*, 2021) and longer k-mers give a chance to be more specific in applications. Kaarme (Díaz-Domínguez, Leinonen and Salmela, 2024), CHTKC (Wang *et al.*, 2021), DSK (Rizk, Lavenier and Chikhi, 2013), and Gebril (Erbert, Rechner and Müller-Hannemann, 2017) programs allow the use of k-mers at least 301 in length (Díaz-Domínguez, Leinonen and Salmela, 2024).

In this context, k-mers refers to substrings of length k created from the reference genome or sequencing data with a step size of 1 (Figure 12). k-mer frequencies indicate how many times a specific k-mer appears in the reference genome or sequencing data (Figure 13). Although k-mer spacing can be greater than one, we can avoid this in variation discovery and genotyping pipelines, due to the potential loss of information density. Sequencing coverage indicates the average number of times each nucleotide from the reference genome is represented in sequencing reads (Sims *et al.*, 2014). Additionally, k-mers may be created *in silico* based on genomic rules.

```
AATTAAGCACCTGGTTTGACAAAGAATTTTCGGCCGGGCG
 ATTAAGCACCTGGTTTGACAAAGAA
  TTAAGCACCTGGTTTGACAAAGAAT
   TAAGCACCTGGTTTGACAAAGAATT
    AAGCACCTGGTTTGACAAAGAATTT
     AGCACCTGGTTTGACAAAGAATTTT
      GCACCTGGTTTGACAAAGAATTTTC
       CACCTGGTTTGACAAAGAATTTTCG
        ACCTGGTTTGACAAAGAATTTTCGG
         CCTGGTTTGACAAAGAATTTTCGGC
          CTGGTTTGACAAAGAATTTTCGGCC
           TGGTTTGACAAAGAATTTTCGGCCG
            GGTTTGACAAAGAATTTTCGGCCGG
             GTTTGACAAAGAATTTTCGGCCGGG
              TTTGACAAAGAATTTTCGGCCGGGC
               TTGACAAAGAATTTTCGGCCGGGCG
```

Figure 12. In this example, k-mers are 25 nucleotides long and created with a step size of 1. These k-mers originate from Alu element insertions, which exhibit polymorphisms in their presence. Polymorphic mobile element insertion (MEI) breakpoint k-mers are highly informative for species identification, as the likelihood of an insertion occurring at the same position at a different time is virtually zero.

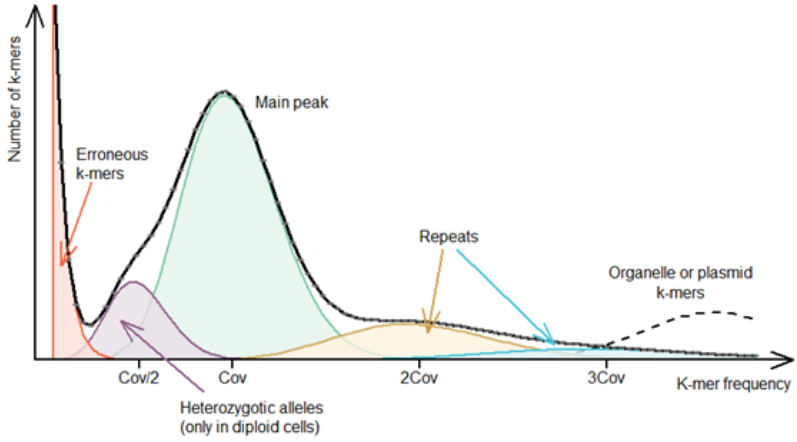


Figure 13. The k-mer frequency distribution in a whole-genome sequencing (WGS) sample. The main peak forms the k-mers from regions with single-copy coverage (Cov) in chromosomes. The Cov/2 peak represents k-mers found in only one sister chromosome or in the non-recombining parts of sex chromosomes in males. K-mers from the repeated regions in the human genome form peaks 2Cov and more Covs (Figure created by Märt Möls, used here with his permission).

In a randomly generated genome with no duplicate events, minimal k should be up-rounded integer $k=\log_4(\text{genome size})$ (Table 3).

Table 3. Minimal k for unique k-mers for reference genomes. As genome size is based on the forward strand, k-mer count for both strands is twice as large if complementary strands are accounted independently.

Organism	Genome size (bp)	K for fw strand	K for fw and rv strand
SARS-CoV	30000	8	8
E.coli	5000000	11	12
Arabidopsis thaliana	119000000	14	14
Diospyros lotus	624000000	15	16
Homo sapiens	3100000000	16	17

Eukaryotic genomes tend to grow and collect repeats and minimal actual k may be lower after counting all presented k-mers in real sequencing data. However, in calculations, k is up to 32 because of computer architecture and to increase speed.

1.6.2. Genome size estimation

Accurate genome size measurement is possible when a *de novo* assembly is 100% accurate, as it allows for the genome size to be calculated by summing the lengths of the assembled chromosomes. In diploid organisms, the genome size is the total length of all sister and sex chromosomes. When sequences are unassembled or unmapped, k-mer-based measurement can be used. Theoretically, the k-mer

based calculation of a diploid genome size involves dividing the cumulative sum of all k-mer frequencies by the average sequencing depth and the k-mer length. This method yields highly accurate results, but genome size is traditionally measured based on the haploid genome, which includes both sex chromosomes. According to the FindGSE publication by Sun et al. (Sun *et al.*, 2018), the lengths of 142 human genomes were measured, ranging from 2,950 Mb to 3,115 Mb, showing a 6% difference between individuals.

1.6.3. Detection of known SNPs

There are numerous programs for identifying the most frequent variants, including short variants, SNPs, and indels. Once variant discovery is performed in multiple individuals or pooled samples, population-specific or sample-set-specific common variants can be detected without the need for resource-intensive variant-calling or mapping steps (Grytten, Dagestad Rand and Sandve, 2022). Most GWAS studies use marker selection for microarrays, which have instrument-specific limitations regarding marker locations and total marker count. Sequence-based detection, however, allows for genotyping of ten or more times the number of marker alleles, depending on the algorithms used. Programs such as Malva (Denti *et al.*, 2019), Bayestyper (The Danish Pan-Genome Consortium *et al.*, 2018), PanGenie (Ebler *et al.*, 2022)and KAGE (Grytten, Dagestad Rand and Sandve, 2022) are all alignment-free genotype callers that use surrounding sequences in k-mers. Among these, only PanGenie and KAGE utilize haplotype information, meaning they consider other nearby existing variants. Additionally, KAGE uses a graph-based pan-genome and allele frequency information. KAGE is highly efficient, capable of genotyping a full sample with 15x coverage in only about 12 minutes using 16 computer cores.

1.6.4. Detection of novel SNPs

Novel SNPs can be classified into two categories: those that have not been described yet due to the limited number of sequenced individuals or insufficient SNP data in databases (e.g., dbSNP, dbVAR), and *de novo* SNPs that appear in children but are absent in their parents. An example of a tool for identifying these SNPs is the program developed by Kimura and Koike (Kimura and Koike, 2015), which compares a dictionary of short NGS reads to a dictionary based on a reference genome. Their test set included simulated data, 5 whole exome sequencing (WES) samples, 3 RNAseq samples, and 3 whole-genome sequencing (WGS) samples. They found that the overall precomputing time for creating a dictionary for WGS data was three times faster than with alignment-based mapping, and SNP calling was twenty times faster, with a sensitivity of 93% and a specificity of 95%. Another tool, Cobasi, is designed for detecting *de novo* mutations in children, using trio samples (Gómez-Romero *et al.*, 2018).

In theory, all biallelic SNPs in a personal genome can be identified in a fully alignment-free manner if heterozygous k-mers (Figure 13) have another k-mer with one mismatch (or more, provided the allele count is 2) within the same heterozygous peak. These k-mers should be included in a set of 50 k-mers, as k-mers spanning the SNP position. Identified SNP alleles may either uniquely align to the reference genome or may not be found in the reference genome due to the complexity of repeated sequences or the absence of the SNP-containing sequence in both alleles in the reference genome (Hurgobin and Edwards, 2017). An allele is considered new if it has not been previously described.

1.6.5. Detection of SVs

Structural variants (SVs) with fixed combinatorics of copy-number changes in alleles contain mutations specific to alleles in a locus. Younger alleles or actively shuffling chromosome regions can only be precisely typed with long and ultra-long reads. However, if precision is not the primary concern, the copy number (CN) of regions can be described using k-mer frequencies. There are numerous applications and ready-to-use programs where the k-mer selection procedure varies. The most precise CN estimators use reference-genome-based gene or gene-specific k-mers, as genes can duplicate locally or jump to another location in the genome via translocation or LINE-mediated transposition through processed pseudogene duplication.

The publication by Pajuste and Remm (Pajuste and Remm, 2023), titled "GeneToCN," uses gene-specific k-mer frequencies to estimate the copy numbers of AMY1, AMY2A, and AMY2B, validating these copy numbers with ddPCR results. The concordance between the results for 38 individuals was 99%. In the same study, the copy numbers of SMN, NPY4R, and the LPA Kringle IV-2 domain genes were also predicted.

Theoretically, any region in the genome can be duplicated, and an already-duplicated region can expand or reduce by the step of the initial duplication. A fully alignment-free method involves selecting k-mers from a personal genome located in peaks above 2n, such as 3n, 4n, etc., and normalizing frequencies with sequencing coverage. These k-mers can locate regions not covered in the reference genome due to the difficulty of assembling regions with a repetitive nature.

1.6.6. Detection of copy number of tandem repeats

Due to the poor mappability of reads, variant calling from variable number tandem repeat (VNTR) regions is challenging. VNtyper addresses this by using the Kastrel algorithm (Audano, Ravishankar and Vannberg, 2018), which generates k-mers from short-read sequencing (SRS) data and the reference genome (Saei *et al.*, 2023). After detecting active regions, haplotypes are reconstructed nucleotide by nucleotide. Sliding window k-mer frequencies are used to detect mutations. VNtyper employs SnaPshot amplification (enrichment) to identify disease-

causing mutations in the test gene, particularly in MUC1-positive (autosomal dominant tubulointerstitial kidney disease) families.

Human chromosome ends, or telomeres, consist of the TTAGGG motif repeated in varying copy numbers among individuals. The TelFinder program determines the length and repeat times of k-mers to identify telomeric motif sequences (Sun *et al.*, 2023). In a study using TelSeq (Ding *et al.*, 2014), the mean length of terminal restriction fragments (mTRFs) was compared with the k-mer-based average telomere length. The study found that the k-mer-based length was 5.63 kb compared to the mTRF length of 6.97 kb, with an annual shortening rate of 34.5 bp/year versus 19.8 bp/year.

2. AIMS OF THE STUDY

This study aims to showcase the effectiveness of k-mers in identifying and characterizing variations within genome sequences. To achieve this, we have three specific patterns of development for computational methods.

1. Development of computational methods for detection of the depth of coverage and describing its relevance for accurate analysis.
2. Development of computational methods for detection of biallelic variants.
3. Development of computational methods for detection of multiallelic variants.

3. RESULTS AND DISCUSSION

Initially, every mutation occurs just once in a single chromosome in a population, giving rise to a variant that is biallelic. If the chromosome carrying the new variant is fortunate, the mutation can gradually spread throughout the population over time. In cases where the mutation event results in sequences with a tandemly repeated nature, the copy number of these tandemly repeated motifs may increase over time. Traditional sequencing methods like short-read NGS and capillary Sanger sequencing struggle to accurately assemble sequences that contain tandem repeats longer than the sequencing read itself.

Table 4. Main variant types covered and reported in publications included in the current thesis. CNV is not covered but is in the list of publications: this work used simplified median-based CNV copy-number detection for CYP2D6 with k-mers.

Variation	Bi/multiallelic (mostly)	Repeatedly the same mutation within another human	Discovery	Detection
SNP	Bi	+	Ref IV	Ref II
STR	Multi	+	Ref IV	–
VNTR	Multi	+	Ref III	Ref III
CNV*	Multi	+–	–	–
MEI	Bi	–	Ref I	Ref I
Satellite/Complex	Highly multi	–	–	Ref VI

The idea of reference genomes is to have one standard as a basis for describing variants in individual genomes. Historically, they have gaps because of limitations in the sequencing methods. The main parts of WGS and WES are focused on discovering and describing short variants with final VCF files, offered by a pipeline calling multiple variants, such as the most commonly used GATK. These methods are mapping-based. Our work with k-mers used mapping-based information only when we visually checked our results or just cut some parts of mapped reads to speed up the throughput of analysis. The information we have obtained comes from thousands of genomes, mainly from the 1000G project and Estonian Genome Center (EGC) full-genome sequencing data. First, we made a k-mer list in preparation for adding information to the sequencing data. Note that the list included all the sequencing information, not only from mappable regions. When the lists were good, then they must be used multiple times to obtain different genomic answers (Genometester4, program glistquery). Instead of occupying an additional 35 GB per individual on the hard disk, we can also use a one-time counting application where the hard disk records only a tiny part compared to a list file (Genometester4, program gmer_counter). All the k-mer lists we used were 25-mer lists, meaning all sequences were cut into 25bp pieces and the time of their presence in sequences was counted (and stored in the list file).

It is worth noting that our k-mer-based methods were planned and developed without the use of artificial intelligence (AI) algorithms.

3.1. Depth of coverage and k-mer length: relevance for accurate analysis (Ref I–VI)

It is crucial to understand the terms "depth" and "coverage" for optimizing DNA sequencing to obtain high-quality and informative data at minimal cost. The human genome's variability, influenced by factors like sex chromosomes, copy-number variations (CNVs), and satellite DNA, can reach as much as 8%. This variability poses challenges in accurately predicting genome size when measuring DNA concentration.

Sequencing depth (or read depth) refers to the number of times a specific nucleotide position in the genome is sequenced. It provides a nucleotide-based measure of how well a particular genomic region is covered by sequencing reads.

Sequencing coverage is a reference-genome-based metric that indicates the number of unique sequencing reads aligning to a specific region in the reference genome. For example, a 20x coverage means that, on average, each position in the reference genome is covered by reads 20 times.

In k-mer analyses, terms such as median and average sequencing coverage are used, but the main peak in k-mer frequency distribution reflects the true sequencing depth. This peak represents the distribution of k-mers from single-copy regions of the genome (Figure 13), offering insights into the actual sequencing depth achieved.

Our research indicated that sequencing coverage below 20x can lead to decreased concordance between non-reference variants and true variants. Factors such as low sequencing coverage or sample contaminants are elaborated in detail in section 3.3.2 of our study.

The choice of minimum k in k-mer length selection depends on the specific applications and available computational resources. Initially, constraints like limited computing power (e.g., 1–2 CPUs, 128 GB RAM, 200 GB HDD) influenced decisions. For instance, developing a 16-mer blacklist using GenomeMasker in early efforts required considerable computational time and storage due to software and hardware limitations (Andreson et al., 2006).

Over time, advancements led to rediscovering k-mers in contexts such as polymorphic Alu element insertions, driven by empirical efforts to enhance PCR primer quality and genotype known variations in the human genome using NGS data (Wang *et al.*, 2006). Today, Illumina-based k-mer lengths typically range optimally from 24 to 27 base pairs, as shorter k-mers tend to have higher frequencies of secondary alignments in the reference genome.

This evolutionary journey highlights how technological and methodological advancements have refined our understanding and utilization of k-mers in genomic research, enabling more precise and efficient analyses despite earlier computational limitations.

3.2. Biallelic variants (Ref I, II, IV)

k-mers representing biallelic variants are in the k-mer distribution graph (Figure 13) in the "cov/2" peak. Because alignment-free sequencing is blind to locations, then biallelic variants allele found after a mapping procedure in repeated areas are filtered out or those results are error-prone. Biallelic variants may present outside reference genome sequences and those variants are outside any genetic studies if they are not described. Nevertheless, biallelic variants are most commonly used in genetic studies, especially SNPs in microarrays. It should be noted that oligonucleotide probes used in microarrays have similar restrictions in use to those of k-mers; they must have a unique location inside the genome. Hence, if only a microarray-based subset of SNPs is under attention, alignment-free variant detection is faster than using alignment-based variant detection.

3.2.1. Genotyping known biallelic SNPs (Ref II)

FastGT is a powerful tool designed for swiftly calling approximately 30 million known SNPs from FASTQ files, accomplishing this task in less than an hour. However, when the reads are mapped and stored in BAM or CRAM formats, an additional 1–3 hours is required for unpacking. This efficiency makes FastGT particularly advantageous when handling numerous samples or when computing resources are limited and not on high-performance computing (HPC) clusters.

The primary use of whole-genome sequencing (WGS) often revolves around creating imputation panels. FastGT facilitates this process by enabling an SNP prescan, allowing regions deemed unnecessary for further calculations to be filtered out based on criteria such as allele count, existing SNP patterns along chromosomes, or allele frequency. This prefiltering step expedites the mapping process by removing unneeded reads or digitally contaminated samples (e.g., demultiplexing errors). Samples from closely related individuals can also be deferred for subsequent analysis.

FastGT excels at identifying SNP allele deletions or duplications due to its extensive marker database, which surpasses the markers available on commonly used arrays. For instance, in Illumina Platinum genomes, FastGT achieves a remarkable 99.96% concordance with 30,328,283 marker genotypes. Key functions like gmer_counter and gmer_caller within FastGT efficiently manage this large dataset.

Our study has determined that a sequencing depth of 20x and a k-mer length of 25 base pairs are optimal for FastGT's operations. Within the 30 million k-mers analyzed, 23,832 markers specific to chromosome Y have been identified, with 5,241 of these markers exhibiting frequencies in male samples from the EGC. Comparing these markers against the Illumina Global Screening Array v.2 Manifest file (GSA-24v3-0_A2.csv) revealed an intersection of 11,188 markers from a total of 654,027 tagSNPs.

There is potential to redesign marker selection criteria using different targeting and quality rules, allowing for the inclusion of k-mers that cover indels and CNV (Copy-Number Variation) describing variants. For instance, the GSAv3 includes 10,118 positions covering indels and 3,846 positions covering CNVs. FastGT's ability to predict copy numbers up to 4 aligns with array-based CNV prediction programs.

In conclusion, FastGT stands out for its rapid and accurate SNP-calling capabilities, robust marker genotype database, and flexibility in handling diverse genetic variations, making it a valuable tool in genetic research and clinical applications.

3.2.2. Genotyping unknown biallelic SNPs (Ref IV)

The main motivation behind developing alignment-free applications for variant calling was the lack of real-time support for variant detection in second-generation sequencing platforms, and the time-consuming nature of post-sequencing alignment-based pipelines such as GATK (Genome Analysis Toolkit). These alignment-based pipelines can take anywhere from 10 to 12 hours on a single computer to call variants in exonic regions, which may not be feasible for users who do not have access to high-performance computing (HPC) resources and require urgent results.

The human genome comprises over 3 billion nucleotides and includes a plethora of simple and complex structural variations (SVs). Tools like KATK (k-mer Alignment Tool Kit) are designed to rapidly detect primarily biallelic SNPs and short indels by employing simplifications that do not cover all regions of the genome comprehensively. These simplifications often involve focusing on regions that are single-copy in the reference genome, where tagging k-mers can uniquely locate variants that are known and documented in databases like dbSNP.

In studies comparing KATK and GATK using simulated data from cell lines CHM1 and CHM13 (which are diploid homozygous cell lines), KATK demonstrated 83.4% concordance with GATK in calling variants within exonic regions (3% of the genome). Both tools missed approximately 2.1% of variants, and an additional 15% of variants were missed by one or the other tools, possibly due to quality control issues in variant calling.

KATK serves as an alternative for users who work with specific subsets of the genome (e.g., certain genes or amplicons), lack experience with complex variant calling pipelines, and need rapid results. It can call variants at 20x coverage for 3% of the genome in just 45 minutes, with calling time linearly correlated with the number of reads.

However, it is important to note that KATK does not currently address capture bias effectively. Capture bias refers to uneven representation of genomic regions in sequencing data due to differences in the efficiency of capturing certain sequences during the experimental process. Until capture bias is resolved or accounted for through equalization of sequencing coverage or other statistical adjustments of read placement, caution is advised in using KATK for applications where accurate representation of all genomic regions is critical.

3.2.3. Detection of Alu-element insertions (Ref I)

The discovery and analysis of Alu element insertions in the human genome have been revolutionized by advances in genomic sequencing technologies. Before the advent of whole-genome sequencing (WGS), identification of these elements was extremely challenging, often relying on serendipitous discoveries during techniques like gel electrophoresis (as noted by Margus Putku in personal communication), where discoveries could be sudden and unexpected.

Alu elements are a type of Short INterspersed Element (SINE) that replicate and insert themselves into the genome, contributing significantly to genomic diversity. They are characterized by conserved features such as target site duplications (TSD), a "start-signal" sequence (GGCCGGGCGC), and a poly-A tail. These elements can be several hundred nucleotides long and are found interspersed throughout the genome.

The tool "AluMine," developed for identifying Alu elements, leverages biological rules derived from these characteristic features. Specifically, AluMine searches for key k-mers that encompass a flanking sequence in the 5' direction (containing the TSD), followed by the "start-signal" sequence or its immediate vicinity in the 3' direction (as illustrated in Figure 14). This approach ensures that potential Alu element insertions are identified based on their unique genomic signatures.

One of the challenges historically associated with Alu-element discovery involves accurately determining the length and orientation of the TSD, as well as describing their genotype presence or absence in one or both alleles. Previous short-read-based discovery programs often faced misunderstandings in these areas, prompting the development of AluMine to provide more accurate and comprehensive detection capabilities. There were missed calls between the 1000 Genomes Phase 1 and Phase 3 studies, as well as with AluMine, even in the commonly used test individual NA12878. Over half of the variant calls overlapped across all three calling algorithms. Since its inception in 2012, AluMine has aimed to identify both known Alu elements present in the human reference genome, as well as new insertions that are absent from the reference. By compiling variant call format (VCF) files for individuals from the Estonian Genome Center (EGC) and other cohorts, AluMine facilitates further research into the impact of these elements on genomic structure and function.

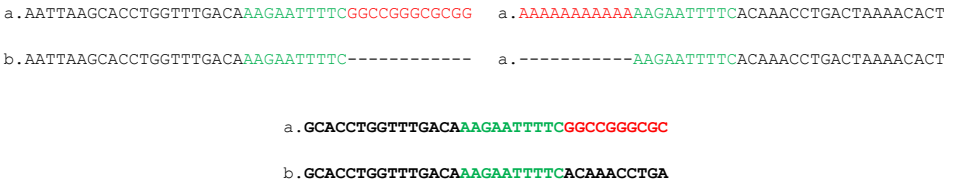


Figure 14. Polymorphic in presence (or PAV) Alu element insertion, see also figure 11. Allele a contains an Alu element and allele b does not. The reference sequence is allele a. Red letters denote the inserted element with poly-A tail, green letters denote target site duplication (TSD) regions, black letters are nucleotides outside of the insertion site TSD. Upper sequences show the ends of the inserted Alu element. The lower box contains breakpoint sequences for querying with BLAST, k-mer applications or script in sequencing data (a. Alu plus and b. Alu minus).

Our findings regarding Alu element insertions in human genomes have shed light on their prevalence and distribution compared to the reference genome and across different populations. The following is a breakdown of the key points discovered using the AluMine software.

1. **Total Alu Element Insertions in Individuals:**
   - On average, each individual has approximately 1,574 Alu element insertions that differ from the reference genome.
   - Among these, 1,045 insertions are not present in the reference genome (PAV – Presence/Absence Variation), indicating novel insertions not previously documented.
   - Additionally, 588 insertions are present in the reference genome but exhibit variation in individual genomes.

2. **Reference Genome (GRCh38) and Cross-Species Comparison:**
   - The human reference genome (GRCh38) contains 15,834 Alu element insertions, which are known to be absent in the chimpanzee genome.
   - Some of these insertions may not be fixed across all humans, suggesting variability in Alu-element presence among different populations.

3. **Denisovan and Neanderthal Genome Comparisons:**
   - AluMine is capable of detecting Alu elements not only in modern human genomes but also in the genomes of Denisovans and Neanderthals, indicating conservation and evolutionary relevance across hominin species.

4. **Computational Efficiency of AluMine:**
   - The process of discovering new Alu-element insertions per individual using AluMine takes approximately 2 hours.
   - Analyzing the states of 15,834 insertions across genomes requires about 20 hours.
   - Genotyping these insertions is completed in less than 4 hours.
   - AluMine is designed to be computationally efficient compared to other variant discovery pipelines, which can take more than 24 hours for similar tasks.

5. **Insights from RNAseq Data:**
   - AluMine's capability extends to analyzing RNAseq data, where it detects Alu-element insertions specifically in the 3' untranslated regions (UTRs) of mRNA transcripts. This finding underscores the utility of AluMine in exploring the functional implications of these insertions in gene regulation and expression.

Overall, AluMine represents a valuable tool in genomic research for comprehensively identifying and characterizing Alu-element insertions across human populations and related species. Its efficiency and alignment-free approach make it particularly suitable for large-scale studies involving genomic variation and evolutionary analysis.

Long-read sequencing methods will improve the accuracy of calling PAVs (Bilgrav Saether and Eisfeldt, 2024). In the discovery phase of novel Alu insertions (REF– in publication Ref I), we used 25 bp of 5' flanking nucleotides to locate them in the reference genome. In the context of long-read data, alignment-free, k-mer-based strategies need to be re-evaluated, as every PAV-containing read can be mapped to the human pangenome, and the use of k-mer combinatorics continues to expand.

3.3. Multiallelic variants (Ref III, VI)

Most multiallelic variations in the human genome are tandemly repeated sequences, such as STRs (Short Tandem Repeats), VNTRs (Variable Number Tandem Repeats), and CNVs (Copy-Number Variations) found in euchromatin, as well as satellite sequences found in heterochromatin. New variants are often discovered under the assumption that at least biallelic presence has been previously described. The TRF program (Benson, 1999) identifies tandemly repeated sequences not only with exact copies but also with mismatches and gaps. The k-mer method works when k-mers have a localization-specific signature, which may be within the motif or span the breakpoints of motif ends for VNTRs. This method also applies to ancient mutations in higher-level motifs, such as the 2400bp (Hsat1B) and 3600bp (HSat3) sequences in chrY heterochromatin.

3.3.1. Estimating copy number of VNTRs (Ref III)

It was known that the TRIB3 promoter region contains 33 bp long-motif VNTR with copy numbers of 2, 3 or 5. We expected that 1x and 4x should also exist. Commonly used k-mer methods do not easily find them because rare alleles 1x and 4x have unique allele counts only in allele combinations 1x/1x, 1x/2x and 4x/5x (Table 4).

Table 4. TRIB3 promoter region VNTR motif copy number in combinations of two haplotypes. Rare 1x and 4x shared haplotypes are detectable only on combinations 1x/1x, 1x/2x and 4x/5x (red). Other 1x and 4x combinations (yellow) are not unique among all other motifs and copy number combinations (white).

Nx/Nx	1	2	3	4	5
1	2	3	4	5	6
2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10

Detection of 1x to 5x tandem repeat variants was also achieved using k-mers. The length of these k-mers included their flanking regions, resulting in lengths of 40, 73, 106, 139, and 140 base pairs. With a sequencing depth of 20x, each allele should ideally appear 10 times in the reads. However, for 140-mer k-mers and 150 bp long reads, only 10 reads would fit this window, and sequencing errors can reduce the count of suitable reads. By searching for k-mers (as detailed in Table 5) within raw sequencing data, we discovered that 1x and 4x VNTR motif copy number variants are present in EGC and 1000G individuals. Following this discovery, we conducted more precise luciferase-based functional experiments using EGC samples, demonstrating that expression levels increase with each additional copy-number step. Global allele frequencies are graphically presented in Reference III. Motif copy numbers greater than 5 certainly exist, but confirming this would require more individuals to be sequenced with long NGS reads. For instance, HG01944 likely has a copy number greater than 5, as detected by substituting the first 4 nucleotides of the 5x VNTR motif with the last 4 nucleotides.

Table 5. k-mer sequences for detecting TRIB3 promoter region VNTR haplotypes in short, 150 bp-length, reads. Copy numbers may be 1–5 and there are no known SNPs in motifs. Used k-mers are 40–140 bp in length and nucleotides coming from repeated motifs are in red.

Nx	k-mer	k-mer sequence with minimum flanking sequence
1x	40-mer	GGCT(GATTAGCTCCGGTTTGCATCACCCGGACCGGGG)$_1$GCC
2x	73-mer	GGCT(GATTAGCTCCGGTTTGCATCACCCGGACCGGGG)$_2$GCC
3x	106-mer	GGCT(GATTAGCTCCGGTTTGCATCACCCGGACCGGGG)$_3$GCC
4x	139-mer	GGCT(GATTAGCTCCGGTTTGCATCACCCGGACCGGGG)$_4$GCC
5x	140-mer	GGCT(GATTAGCTCCGGTTTGCATCACCCGGACCGGGG)$_4$GATT

3.3.2. Detection of chrY haplogroups with k-mer profiles (Ref VI)

Human Y chromosome haplogroups (HGs) are the standard for describing the non-recombining portion of the Y chromosome. These HGs are defined by one or more mutations that distinguish them from each other in a clonal manner. In principle, a single k-mer with a frequency of 1 could predict the HG of the human Y chromosome (chrY). However, the probability of finding such a k-mer in whole-genome sequencing (WGS) data using the GRCh38 assembly is approximately 1 in 30,000,000, which is the length of euchromatin in chrY. There are a different number of SNPs associated with certain HGs, and if the read length is 150 bp, the probability increases significantly by 150 times the number of HG-specific SNPs times the k-mer length. This calculation (150 * 300 * 25) increases the probability by approximately 1.25 million times, resulting in a probability of 1 in 24 for all SNPs. This suggests that a sequencing depth of 0.04x is necessary to accurately detect HG-specific SNPs on chrY. If reads are shorter or fragmented, as is often the case with ancient DNA (aDNA), the minimum coverage required increases. This limitation applies to using SNP data, both with and without capture, in WGS. Instead of relying solely on SNP information, chrY can leverage HG-specific blocks with numerous mutations. These blocks provide additional count information that can be useful in HG identification. For chrY heterochromatin and other tandem repeats, we used a simple hypothesis: in a specific HG, deletions occur in one or more segments of satellite DNA, while other regions of the heterochromatin may undergo multiplication of nearby segment sequences. In another HG, a similar process occurs in different segments. Over generations, these processes create HG-specific patterns of segment-specific k-mer frequencies (see Figure 15).

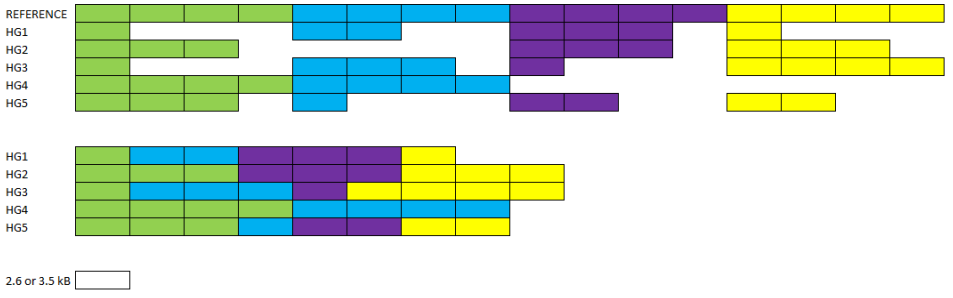


Figure 15. A simplified schema of tandemly repeated 2.6 and 3.5kb blocks with block-specific mutations can be visualized with colors representing different mutations (as shown in the figure). On average, the heterochromatin in chromosome Y (chrY) spans approximately 30Mb by reference genomes (REFERENCE), although this length can vary (HG1-HG5). The composition of these blocks contains an amount of information equivalent to 1,800 times that of mitochondrial DNA (mDNA).

To predict HGs from ultra-low sequencing data, we first built a model using high-coverage samples with known HGs, then tested and applied it (Figure 16). Creating the model involved three steps in k-mer selection and ended with model building:

1. **Selecting Human Chromosome Y Specific k-mers**
   - Generating k-mer lists from assembled Y chromosomes, or
   - Using male WGS-based k-mer lists. Sequencing errors are removed using a cutoff function during list generation with the GenomeTester4 package (Kaplinski, Lepamets and Remm, 2015).
   - Excluding k-mers presented in the "female" lists.

2. **Selecting the Most Informative k-mers Based on HGs of Interest**
   - Chromosome Y-specific k-mer frequencies from individual k-mer lists are normalized using average values from a subset of "NIPT" chrY k-mers (~36,000 evenly distributed across the euchromatic region of chrY between the PAR regions) (Sauk *et al.*, 2018). These normalized values approximate the copy number of each k-mer on chromosome Y, as they originate from the 1x region in the GRCh38 assembly.
   - For each HG selected for model training, Mann-Whitney U tests are used to compare the average counts of each k-mer between the target HG and all other HGs.
   - From each HG, the 10,000 most specific and 10,000 least specific k-mers are selected as the most informative.

3. **Preparing Input Data for the Model**
   - Informative k-mers from the training set are compared against additional k-mers used for sequencing depth estimation, and intersecting k-mers are removed.
   - An input table is created with raw k-mer frequencies for each individual selected for the model.

4. **Model Building**
   - A model is built using distance-based clustering across all informative k-mers.
   - During HG calling, the HG with the smallest distance to the sample is assigned as the best match.

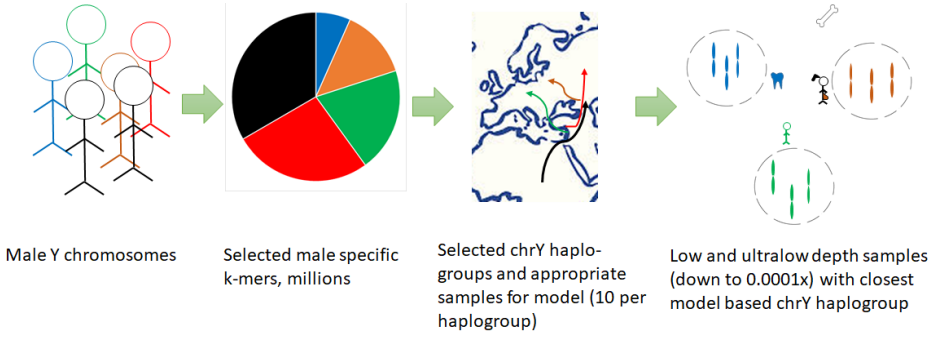


Figure 16. Workflow of building Y-mer model, testing and using. Y-mer method relies on Y chromosome-specific k-mers, 25 base pair sequences absent in the female genome, whose frequencies within the Y chromosome are similar within HGs but distinct between them. Millions of such k-mers exist in high-coverage WGS samples that are required to build the HG prediction models. Only a small fraction of the k-mers used in the model would be required to be present in ultra-low sequencing depth samples (aDNA, NIPT, forensics) for their successful HG prediction. The bone in the figure belongs to a female sample because it is too far from the HG groups.

In simulated dilutions, we may say that the HG prediction threshold for samples in-model is 0.0005x and out-of-sample samples 0.001x (Figure 17), i.e., at a minimum, 100–500 times less DNA is needed. Deeper analysis shows that 60% of k-mers come from repeated sequences.

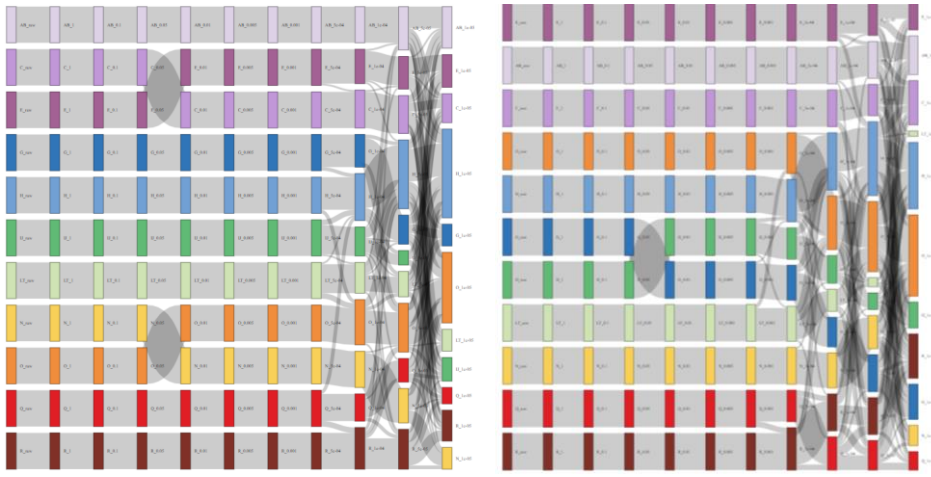


Figure 17. Testing simulated dilutions (columns raw, 1x, 0.1x, 0.05x, 0.01x, 0.005x, 0.001x, 0.0005x, 0.0001x, 0.00005x, 0.00001x) in samples used in (left) and not used (right) in model building. The sensitivity difference is 5 times, but still it is 0.005x in samples not used in the model. Grey color indicates HG prediction accuracy dynamics: other colors represent HGs (AB, C, E, G, H, IJ, LT, N, O, Q, R).

A massive number of Illumina-based next-generation sequencing (NGS) projects use capture techniques for gene panels, exomes, or targeted amplification (Figure 9). In these cases, sequencing depth is uneven across the chromosome. However, this does not affect HG prediction sensitivity, as long as the capture region does not specifically target chrY repeats. PCR enrichment for bacteria, followed by WGS (Lavania *et al.*, 2018), retains human DNA fragments, allowing for HG prediction even in these samples (unpublished results). Due to the variability of whole-exome sequencing (WES) protocols, we cannot recommend their use for HG prediction until improvements have been made.

CONCLUSIONS

In whole-genome sequencing (WGS), sequencing reads include all chromosomal sequences, although they may not be evenly covered when mapped to the reference genome. Mapping-based approaches filter out chromosomal regions not listed in the reference genome. Until the development of an ideal reference genome, sequencing technology, and variant-calling algorithms, k-mer-based methods can be used to partially fill these gaps and enhance genomic analysis applications. This study examined variation data in whole-genome samples from 2,500 Estonian Biobank participants and 300 from the 1000 Genomes Project.

In this thesis, we demonstrated the following:

1. **k-mer frequencies are quantifiable and informative**, similar to signal intensities in microarrays. While copy number variations detected via microarrays are defined on a scale of 0–4, k-mer methods have no such limits, enabling the detection of variations not identifiable through microarray or mapping-based methods. k-mer frequencies are normalized using either the mean/median coverage of the sequencing dataset, the mean/median coverage of the neighboring region of interest, or another suitable normalization variable. It is important to note that mapping-based and k-mer-based coverage values do not match, as mapping-based values depend on the reference genome length, which can differ between individuals by up to 10% of the total genome length.
2. **k-mer frequencies allow the genotyping of both known and unknown biallelic mutations**, such as SNPs and PAVs. k-mers describing biallelic markers move one step across the variation site, forming two k-mer sets: one representing each allele. The frequency ratio of these k-mers indicates the presence of one, the other, or both alleles in the sample. Care should be taken to ensure the k-mer is unique to a single location in the genome and within one mismatch distance to avoid overlap with unmapped mutations. This means many variations detectable by mapping-based methods are missed with 32-mers, as mapping-based methods use longer reads (150–300 base pairs) for unique placement.
3. **Normalized k-mer frequencies can accurately predict copy numbers of multiallelic tandem repeats**, such as VNTRs. The human genome contains over 260,000 tandemly repeated sequences with motifs of 6–100 base pairs. When the product of the motif length and copy number exceeds the read length (over 150 base pairs in this study), mapping-based methods fail to reliably determine the variant, even when the reference sequence contains the longest repeat. With mapping-free methods like ours, the combined copy number of the two alleles is determined, enabling correlation analysis with other traits of the same sample. Notably, unreported normalized k-mer results showed 80% concordance with gel electrophoresis results. We believe the

20% discrepancy could be addressed by identifying somatic mutations using long-read sequencing.

4. **Human chrY HG-specific k-mers can be identified and used to predict HGs from WGS data**, even at very low sequencing depths. Half of the human Y chromosome consists of heterochromatin with the primary motif TTCCA, which is the most common 25-mer (in five-copy repetitions) in the human genome. Over millions of years, mutations have expanded this motif to 2.5–3.5 kb motifs. In chrY HG determination, repetitive sequences account for 70% of the significance. These repeat-rich chromosomal regions enable the classification of HGs in previously unclassified samples at coverage levels as low as 0.005–0.0001x, depending on the HG. Tandem repeat mutations, feared to occur rapidly, are either not as fast or exhibit distinct repeat patterns for different HGs. Extremely low-coverage samples originate from ancient burial sites, forensic cases, cell-free DNA, or purposefully low-coverage sequencing experiments. This method is also suitable for analyzing enriched samples because enriched genome segments and repeats do not overlap.

SUMMARY IN ESTONIAN

Inimese genoomi uuringud k-mer sagedustega

Genoomid on dünaamilised, muteerumine genoomides on permanentne protsess. Imetajate, sh. inimese genoom on väga suur, 3,2 miljardit aluspaari samas, kui tavalise bakteri genoom on 5 miljonit aluspaari ning viirusel kümned tuhanded. Kui inimese genoomi järjestust lähemalt vaadata, siis üle poole genoomist moodustavad kordused, mis paiknevad sõltuvalt tekkimise ja funktsiooni tõttu kas hajusalt üle genoomi või tandeemselt. Sekveneerimismetoodikad määravad ära genoomi osad, mida saab varieeruvuse tuvastamiseks uurida. Viimase 25 aasta jooksul on inimese genoomi uuringutes toimunud väga suur edasiminek.

Variatsioonide tuvastamisel inimese WGS andmetest kasutatakse valdavalt joonduspõhist meetodit, mis põhineb lugemite paigutamisel referentsjärjestusele. Esimene inimese referentsjärjestuse versioon sai teadlastele kättesaadavaks 2001. aasta veebruaris, WGS andmed aga peale 2007. aastat. Vahepealsel ajal, 2001–2007, olid teadlastele kättesaadavad SNP alleelivariandid, mis tuvastati mikrokiipidega. DNA mikrokiipide kasutus SNP variatsioonide määramisel on tänapäevalgi väga lai, Eesti Geenivaramu 200000 ja 23andMe Holding Co. üle 15 miljoni proovi on genotüpiseeritud Illumina Inc. kiibiplatvormi kasutades.

Käesolevas töös on genoomi uurimise andmeühikuks k-merid (25 aluspaari pikad), mis on valitud kas referentsjärjestusest, WGS andmetest või teoreetiliselt eksisteerivate seast ning lisaks on sekveneerimisandmetest leitud k-meride sagedused. Edasine variatsiooni kirjeldamine on joondusvaba, mistõttu joondusele kuluva aja võrra on variatsiooni tuvastus kiirem.

Kui SNP põhisel kiibil on andmepunkte miljonites, siis WGS andmetes on k-mer niipalju, kui inimese genoomis erinevaid k pikkusi järjestusi on, üle 2,5 miljardi. Siin töös me näitasime, et dbSNP andmebaasis olevad ühenukleotiidsed variatsioonid on tuvastatavad nii nagu on tuvastatavad need, mis dbSNP andmebaasis veel kirjas ei ole. Siin juures tuleb rõhutada, et iga SNP ei ole kiibikõlbulik oma järjestuse spetsiifikast tulenevalt ning k-mer põhisel tuvastamisel on täpselt samasugused variatsiooni asukoha arvust tingutud piirangud.

SNP-de tuvastamisel, mis põhineb joondusel, leitakse nii juba teadaolevad kui ka uued SNPd. Kirjeldamine õnnestub hästi geene sisaldavates kromosoomi piirkondades, kus geeni ja tandeemse korduse koopiaarv ei ole väga suur. DNA järjestuse korduvates osades on aga raske muutusi tuvastada, kuna joondamisel lühikesed lugemid ei suuda neid piirkondi ühtlaselt katta. Ka joondusvabad meetodid, mis põhinevad k-meride analüüsil, on piiratud täpsusega seni, kuni populatsioonis eksisteerivaid alleele ei ole täispikkuses järjestatud. Lahenduseks on pakutud graafipõhise pangenoomi kasutamist, mis võimaldab täpsemalt määrata kõik variatsioonid. Kuni täiuslikku pangenoomi pole olemas, saab k-meride sageduste põhjal hinnata ligikaudset geeni- ja tandeemse korduse koopia arvu.

Käesolevas uuringus näidatakse, et k-mer põhine koopiaarvu hinnang on bioloogiliselt tähenduslik, kuna see seostub hästi transkriptsioonifaktorite seostumiskohtade arvu ja allavoolu oleva geeni avaldumise tasemega VNTR-järjestustes.

Käeoleva töö kõige põnevam tehniline ja teaduslik osa on Y-kromosoomi haplogruppide määramine väga vähese DNA koguse (0.005x–0.0001x e. 0.5–0.01% genoomist) põhjal, kasutades miljoneid k-meride kombinatsioone. Tavaliselt vajatakse usaldusväärseks SNP alleelide määramiseks vähemalt 20x katvust. Y-kromosoomist on pool tandeemselt korduv ning sellest põhjustatuna oli nimetatud piirkond jäetud varasemates uuringutes kõrvale. Meie lähenemine on eriline, kuna kasutame just seda kordust kui DNA loomuliku „võimendusest", mis oleks kui laialdaselt kasutatav PCR tehnoloogia. Aja jooksul on meie poolt kasutatavad korduvad piirkonnad kogunud unikaalseid muutusi – nii DNA lõigu väljalõikamist kui ka juurde kasvamist. Kuna need piirkonnad ei rekombineeru ei X ega teistsuguse Y kromosoomiga, tekib Y kromosoomi haplogrupile iseloomulik muster. Meie k-meride kombinatoorikal põhinev meetod koos kaugus-põhise mudeliga suudabki määrata Y haplogrupi ka sellistest proovist, mille puhul varasemad meetodid ebaõnnestusid.

REFERENCES

Abyzov, A. *et al.* (2013) 'Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division', *Genome Research*, 23(12), pp. 2042–2052. Available at: https://doi.org/10.1101/gr.154625.113.

Altemose, N. (2022) 'A classical revival: Human satellite DNAs enter the genomics era', *Seminars in Cell & Developmental Biology*, p. S1084952122001379. Available at: https://doi.org/10.1016/j.semcdb.2022.04.012.

Altschul, S.F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*, 215(3), pp. 403–410. Available at: https://doi.org/10.1016/S0022-2836(05)80360-2.

Andreson, R. *et al.* (2006) 'GENOMEMASKER package for designing unique genomic PCR primers', *BMC Bioinformatics*, 7(1), p. 172. Available at: https://doi.org/10.1186/1471-2105-7-172.

Arora, U.P. *et al.* (2021) 'Population and subspecies diversity at mouse centromere satellites', *BMC Genomics*, 22(1), p. 279. Available at: https://doi.org/10.1186/s12864-021-07591-5.

Audano, P.A., Ravishankar, S. and Vannberg, F.O. (2018) 'Mapping-free variant calling using haplotype reconstruction from k-mer frequencies', *Bioinformatics*. Edited by B. Berger, 34(10), pp. 1659–1665. Available at: https://doi.org/10.1093/bioinformatics/btx753.

BAC Resource Consortium, T. *et al.* (2001) 'Integration of cytogenetic landmarks into the draft sequence of the human genome', *Nature*, 409(6822), pp. 953–958. Available at: https://doi.org/10.1038/35057192.

Bailey, J.A. *et al.* (2002) 'Recent Segmental Duplications in the Human Genome', *Science*, 297(5583), pp. 1003–1007. Available at: https://doi.org/10.1126/science.1072047.

Bankevich, A. *et al.* (2012) 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing', *Journal of Computational Biology*, 19(5), pp. 455–477. Available at: https://doi.org/10.1089/cmb.2012.0021.

Batzer, M.A. and Deininger, P.L. (2002) 'Alu repeats and human genomic diversity', *Nature Reviews Genetics*, 3(5), pp. 370–379. Available at: https://doi.org/10.1038/nrg798.

Benson, G. (1999) 'Tandem repeats finder: a program to analyze DNA sequences', *Nucleic Acids Research*, 27(2), pp. 573–580. Available at: https://doi.org/10.1093/nar/27.2.573.

Bilgrav Saether, K. and Eisfeldt, J. (2024) 'Detecting transposable elements in long-read genomes using sTELLeR', *Bioinformatics*. Edited by P. Robinson, 40(11), p. btae686. Available at: https://doi.org/10.1093/bioinformatics/btae686.

Bokhari, Y. *et al.* (2022) 'ChromoEnhancer: An Artificial-Intelligence-Based Tool to Enhance Neoplastic Karyograms as an Aid for Effective Analysis', *Cells*, 11(14), p. 2244. Available at: https://doi.org/10.3390/cells11142244.

Broman, K.W. *et al.* (1998) 'Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination', *The American Journal of Human Genetics*, 63(3), pp. 861–869. Available at: https://doi.org/10.1086/302011.

Byrska-Bishop, M. *et al.* (2022) 'High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios', *Cell*, 185(18), pp. 3426–3440.e19. Available at: https://doi.org/10.1016/j.cell.2022.08.004.

Chin, C.-S. *et al.* (2023) 'Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes', *Nature Methods*, 20(8), pp. 1213–1221. Available at: https://doi.org/10.1038/s41592-023-01914-y.

Daccord, N. *et al.* (2017) 'High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development', *Nature Genetics*, 49(7), pp. 1099–1106. Available at: https://doi.org/10.1038/ng.3886.

Danecek, P. *et al.* (2011) 'The variant call format and VCFtools', *Bioinformatics*, 27(15), pp. 2156–2158. Available at: https://doi.org/10.1093/bioinformatics/btr330.

De Coster, W. and Rademakers, R. (2023) 'NanoPack2: population-scale evaluation of long-read sequencing data', *Bioinformatics*. Edited by C. Alkan, 39(5), p. btad311. Available at: https://doi.org/10.1093/bioinformatics/btad311.

Denti, L. *et al.* (2019) 'MALVA: Genotyping by Mapping-free ALlele Detection of Known VAriants', *iScience*, 18, pp. 20–27. Available at: https://doi.org/10.1016/j.isci.2019.07.011.

Deorowicz, S. *et al.* (2015) 'KMC 2: fast and resource-frugal k-mer counting', *Bioinformatics (Oxford, England)*, 31(10), pp. 1569–1576. Available at: https://doi.org/10.1093/bioinformatics/btv022.

Díaz-Domínguez, D., Leinonen, M. and Salmela, L. (2024) 'Space-efficient computation of k-mer dictionaries for large values of k', *Algorithms for molecular biology: AMB*, 19(1), p. 14. Available at: https://doi.org/10.1186/s13015-024-00259-1.

Ding, Z. *et al.* (2014) 'Estimating telomere length from whole genome sequence data', *Nucleic Acids Research*, 42(9), p. e75. Available at: https://doi.org/10.1093/nar/gku181.

Dréau, A. *et al.* (2019) 'Genome-wide recombination map construction from single individuals using linked-read sequencing', *Nature Communications*, 10(1), p. 4309. Available at: https://doi.org/10.1038/s41467-019-12210-9.

Ebler, J. *et al.* (2022) 'Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes', *Nature Genetics*, 54(4), pp. 518–525. Available at: https://doi.org/10.1038/s41588-022-01043-w.

Emanuel, B.S. and Saitta, S.C. (2007) 'From microscopes to microarrays: dissecting recurrent chromosomal rearrangements', *Nature Reviews Genetics*, 8(11), pp. 869–883. Available at: https://doi.org/10.1038/nrg2136.

Erbert, M., Rechner, S. and Müller-Hannemann, M. (2017) 'Gerbil: a fast and memory-efficient k-mer counter with GPU-support', *Algorithms for molecular biology: AMB*, 12, p. 9. Available at: https://doi.org/10.1186/s13015-017-0097-9.

Fadaie, Z. *et al.* (2021) 'Long-read technologies identify a hidden inverted duplication in a family with choroideremia', *Human Genetics and Genomics Advances*, 2(4), p. 100046. Available at: https://doi.org/10.1016/j.xhgg.2021.100046.

Ferguson-Smith, M.A. (2015) 'History and evolution of cytogenetics', *Molecular Cytogenetics*, 8(1), p. 19. Available at: https://doi.org/10.1186/s13039-015-0125-8.

Feuk, L., Carson, A.R. and Scherer, S.W. (2006) 'Structural variation in the human genome', *Nature Reviews Genetics*, 7(2), pp. 85–97. Available at: https://doi.org/10.1038/nrg1767.

Freeman, J.L. (2006) 'Copy number variation: New insights in genome diversity', *Genome Research*, 16(8), pp. 949–961. Available at: https://doi.org/10.1101/gr.3677206.

Gao, Y. *et al.* (2023) 'A pangenome reference of 36 Chinese populations', *Nature*, 619(7968), pp. 112–121. Available at: https://doi.org/10.1038/s41586-023-06173-7.

Goldmann, J.M. *et al.* (2016) 'Parent-of-origin-specific signatures of de novo mutations', *Nature Genetics*, 48(8), pp. 935–939. Available at: https://doi.org/10.1038/ng.3597.

Gómez-Romero, L. *et al.* (2018) 'Precise detection of de novo single nucleotide variants in human genomes', *Proceedings of the National Academy of Sciences of the United States of America*, 115(21), pp. 5516–5521. Available at: https://doi.org/10.1073/pnas.1802244115.

Gregory, C. and Maher, E. (2010) 'Automating Giemsa banding of chromosomes: protocol for and evaluation of the use of a programmable, high-throughput automatic stainer', *Biotechnic & Histochemistry*, 84(6), pp. 337–345. Available at: https://doi.org/10.3109/10520290902879250.

de Grouchy, J. *et al.* (1972) 'Karyotypic evolution in man and chimpanzees. A comparative study of band topographies after controlled denaturation', *Annales De Genetique*, 15(2), pp. 79–84.

Grytten, I., Dagestad Rand, K. and Sandve, G.K. (2022) 'KAGE: fast alignment-free graph-based genotyping of SNPs and short indels', *Genome Biology*, 23(1), p. 209. Available at: https://doi.org/10.1186/s13059-022-02771-2.

Halldorsson, B.V. *et al.* (2019) 'Characterizing mutagenic effects of recombination through a sequence-level genetic map', *Science*, 363(6425), p. eaau1043. Available at: https://doi.org/10.1126/science.aau1043.

Hickey, G. *et al.* (2024) 'Pangenome graph construction from genome alignments with Minigraph-Cactus', *Nature Biotechnology*, 42(4), pp. 663–673. Available at: https://doi.org/10.1038/s41587-023-01793-w.

Higuchi, T. *et al.* (2004) 'Molecular Cloning, Genomic Structure, and Expression Analysis of MUC20, a Novel Mucin Protein, Up-regulated in Injured Kidney', *Journal of Biological Chemistry*, 279(3), pp. 1968–1979. Available at: https://doi.org/10.1074/jbc.M304558200.

Hurgobin, B. and Edwards, D. (2017) 'SNP Discovery Using a Pangenome: Has the Single Reference Approach Become Obsolete?', *Biology*, 6(1), p. 21. Available at: https://doi.org/10.3390/biology6010021.

International Human Genome Sequencing Consortium *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921. Available at: https://doi.org/10.1038/35057062.

International Standing Committee on Human Cytogenomic Nomenclature *et al.* (eds) (2020) *ISCN 2020: an international system for human cytogenomic nomenclature (2020)*. Basel ; Hartford: Karger.

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

Kimura, K. and Koike, A. (2015) 'Ultrafast SNP analysis using the Burrows-Wheeler transform of short-read data', *Bioinformatics (Oxford, England)*, 31(10), pp. 1577–1583. Available at: https://doi.org/10.1093/bioinformatics/btv024.

Kong, A. *et al.* (2002) 'A high-resolution recombination map of the human genome', *Nature Genetics*, 31(3), pp. 241–247. Available at: https://doi.org/10.1038/ng917.

Koren, S. *et al.* (2017) 'Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation', *Genome Research*, 27(5), pp. 722–736. Available at: https://doi.org/10.1101/gr.215087.116.

Kumar, S. and Blaxter, M.L. (2011) 'Simultaneous genome sequencing of symbionts and their hosts', *Symbiosis*, 55(3), pp. 119–126. Available at: https://doi.org/10.1007/s13199-012-0154-6.

Lavania, M. *et al.* (2018) 'Enriched whole genome sequencing identified compensatory mutations in the RNA polymerase gene of rifampicin-resistant Mycobacterium leprae strains', *Infection and Drug Resistance*, 11, pp. 169–175. Available at: https://doi.org/10.2147/IDR.S152082.

Li, H. (2022) 'Minigraph pangenome graphs for HPRC year-1 samples'. Zenodo. Available at: https://doi.org/10.5281/ZENODO.6983934.

Liao, W.-W. *et al.* (2023) 'A draft human pangenome reference', *Nature*, 617(7960), pp. 312–324. Available at: https://doi.org/10.1038/s41586-023-05896-x.

Lichter, J.B. *et al.* (1993) 'A hypervariable segment in the human dopamine receptor D $_4$ (*DRD4*) gene', *Human Molecular Genetics*, 2(6), pp. 767–773. Available at: https://doi.org/10.1093/hmg/2.6.767.

Liehr, T. (2021) 'About classical molecular genetics, cytogenetic and molecular cytogenetic data not considered by Genome Reference Consortium and thus not included in genome browsers like UCSC, Ensembl or NCBI', *Molecular Cytogenetics*, 14(1), pp. 20, s13039-021-00540–7. Available at: https://doi.org/10.1186/s13039-021-00540-7.

Logsdon, G.A. *et al.* (2024) 'The variation and evolution of complete human centromeres', *Nature*, 629(8010), pp. 136–145. Available at: https://doi.org/10.1038/s41586-024-07278-3.

Marçais, G. and Kingsford, C. (2011) 'A fast, lock-free approach for efficient parallel counting of occurrences of k-mers', *Bioinformatics (Oxford, England)*, 27(6), pp. 764–770. Available at: https://doi.org/10.1093/bioinformatics/btr011.

McKenzie, W.H. and Lubs, H.A. (1975) 'Human Q and C chromosomal variations: distribution and incidence', *Cytogenetic and Genome Research*, 14(2), pp. 97–115. Available at: https://doi.org/10.1159/000130330.

Mitchell, A.R. and Gosden, J.R. (1978) 'Evolutionary relationships between man and the great apes', *Science Progress*, 65(259), pp. 273–293.

Mixão, V. and Gabaldón, T. (2020) 'Genomic evidence for a hybrid origin of the yeast opportunistic pathogen Candida albicans', *BMC Biology*, 18(1), p. 48. Available at: https://doi.org/10.1186/s12915-020-00776-6.

Nurk, S. *et al.* (2022) 'The complete sequence of a human genome', *Science*, 376(6588), pp. 44–53. Available at: https://doi.org/10.1126/science.abj6987.

Pajuste, F.-D. and Remm, M. (2023) 'GeneToCN: an alignment-free method for gene copy number estimation directly from next-generation sequencing reads', *Scientific Reports*, 13(1), p. 17765. Available at: https://doi.org/10.1038/s41598-023-44636-z.

Reis, A.L.M. *et al.* (2023) 'The landscape of genomic structural variation in Indigenous Australians', *Nature*, 624(7992), pp. 602–610. Available at: https://doi.org/10.1038/s41586-023-06842-7.

Ren, J. *et al.* (2018) 'Alignment-Free Sequence Analysis and Applications', *Annual Review of Biomedical Data Science*, 1(1), pp. 93–114. Available at: https://doi.org/10.1146/annurev-biodatasci-080917-013431.

Reuter, J.A., Spacek, D.V. and Snyder, M.P. (2015) 'High-Throughput Sequencing Technologies', *Molecular Cell*, 58(4), pp. 586–597. Available at: https://doi.org/10.1016/j.molcel.2015.05.004.

Rizk, G., Lavenier, D. and Chikhi, R. (2013) 'DSK: k-mer counting with very low memory usage', *Bioinformatics (Oxford, England)*, 29(5), pp. 652–653. Available at: https://doi.org/10.1093/bioinformatics/btt020.

Robinson, J.T. *et al.* (2011) 'Integrative genomics viewer', *Nature Biotechnology*, 29(1), pp. 24–26. Available at: https://doi.org/10.1038/nbt.1754.

Saei, H. *et al.* (2023) 'VNtyper enables accurate alignment-free genotyping of MUC1 coding VNTR using short-read sequencing data in autosomal dominant tubulo-interstitial kidney disease', *iScience*, 26(7), p. 107171. Available at: https://doi.org/10.1016/j.isci.2023.107171.

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

Sauk, M. *et al.* (2018) 'NIPTmer: rapid k-mer-based software package for detection of fetal aneuploidies', *Scientific Reports*, 8(1), p. 5616. Available at: https://doi.org/10.1038/s41598-018-23589-8.

Schloissnig, S. *et al.* (2024) 'Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project'. Available at: https://doi.org/10.1101/2024.04.18.590093.

Secomandi, S. *et al.* (2025) 'Pangenome graphs and their applications in biodiversity genomics', *Nature Genetics*, 57(1), pp. 13–26. Available at: https://doi.org/10.1038/s41588-024-02029-6.

Shen, B. *et al.* (2018) 'Characterization of recombination features and the genetic basis in multiple cattle breeds', *BMC Genomics*, 19(1), p. 304. Available at: https://doi.org/10.1186/s12864-018-4705-y.

Shi, C.H. and Yip, K.Y. (2020) 'A general near-exact k-mer counting method with low memory consumption enables *de novo* assembly of 106× human sequence data in 2.7 hours', *Bioinformatics*, 36(Supplement_2), pp. i625–i633. Available at: https://doi.org/10.1093/bioinformatics/btaa890.

Shi, Y. *et al.* (2023) 'Characterization of genome-wide STR variation in 6487 human genomes', *Nature Communications*, 14(1), p. 2092. Available at: https://doi.org/10.1038/s41467-023-37690-8.

Sims, D. *et al.* (2014) 'Sequencing depth and coverage: key considerations in genomic analyses', *Nature Reviews Genetics*, 15(2), pp. 121–132. Available at: https://doi.org/10.1038/nrg3642.

Sirén, J. *et al.* (2024) 'Personalized pangenome references', *Nature Methods*, 21(11), pp. 2017–2023. Available at: https://doi.org/10.1038/s41592-024-02407-2.

Spence, J.P. and Song, Y.S. (2019) 'Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations', *Science Advances*, 5(10), p. eaaw9206. Available at: https://doi.org/10.1126/sciadv.aaw9206.

Srebniak, M. *et al.* (2004) 'Subfertile couple with inv(2),inv(9) and 16qh+', *Journal of Applied Genetics*, 45(4), pp. 477–479.

Sun, H. *et al.* (2018) 'findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies', *Bioinformatics*. Edited by I. Birol, 34(4), pp. 550–557. Available at: https://doi.org/10.1093/bioinformatics/btx637.

Sun, Q. *et al.* (2023) 'Large-Scale Detection of Telomeric Motif Sequences in Genomic Data Using TelFinder', *Microbiology Spectrum*, 11(2), p. e0392822. Available at: https://doi.org/10.1128/spectrum.03928-22.

Suzuki, Y., Myers, E.W. and Morishita, S. (2020) 'Rapid and ongoing evolution of repetitive sequence structures in human centromeres', *Science Advances*, 6(50). Available at: https://doi.org/10.1126/sciadv.abd9230.

Ten Berk De Boer, E., Bilgrav Saether, K. and Eisfeldt, J. (2023) 'Discovery of non-reference processed pseudogenes in the Swedish population', *Frontiers in Genetics*, 14, p. 1176626. Available at: https://doi.org/10.3389/fgene.2023.1176626.

The 1000 Genomes Project Consortium *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. Available at: https://doi.org/10.1038/nature15393.

The All of Us Research Program Genomics Investigators *et al.* (2024) 'Genomic data in the All of Us Research Program', *Nature*, 627(8003), pp. 340–346. Available at: https://doi.org/10.1038/s41586-023-06957-x.

The Danish Pan-Genome Consortium *et al.* (2018) 'Accurate genotyping across variant classes and lengths using variant graphs', *Nature Genetics*, 50(7), pp. 1054–1059. Available at: https://doi.org/10.1038/s41588-018-0145-5.

Tjio, J.H. and Levan, A. (1956) 'The chromosome number of man', *Hereditas*, 42(1–2), pp. 1–6. Available at: https://doi.org/10.1111/j.1601-5223.1956.tb03010.x.

Uddin, M. *et al.* (2023) 'A draft Arab pangenome reference'. Available at: https://doi.org/10.21203/rs.3.rs-3490341/v1.

Venter, J.C. *et al.* (2001) 'The Sequence of the Human Genome', *Science*, 291(5507), pp. 1304–1351. Available at: https://doi.org/10.1126/science.1058040.

Verlouw, J.A.M. *et al.* (2021) 'A comparison of genotyping arrays', *European Journal of Human Genetics* [Preprint]. Available at: https://doi.org/10.1038/s41431-021-00917-7.

Wang, J. *et al.* (2006) 'dbRIP: A highly integrated database of retrotransposon insertion polymorphisms in humans', *Human Mutation*, 27(4), pp. 323–329. Available at: https://doi.org/10.1002/humu.20307.

Wang, J. *et al.* (2021) 'CHTKC: a robust and efficient k-mer counting algorithm based on a lock-free chaining hash table', *Briefings in Bioinformatics*, 22(3), p. bbaa063. Available at: https://doi.org/10.1093/bib/bbaa063.

Wang, J. and LaFramboise, T. (2019) 'CytoConverter: a web-based tool to convert karyotypes to genomic coordinates', *BMC Bioinformatics*, 20(1), p. 467. Available at: https://doi.org/10.1186/s12859-019-3062-4.

Xing, J. *et al.* (2009) 'Mobile elements create structural variation: analysis of a complete human genome', *Genome Research*, 19(9), pp. 1516–1526. Available at: https://doi.org/10.1101/gr.091827.109.

Yunis, J.J. and Prakash, O. (1982) 'The Origin of Man: A Chromosomal Pictorial Legacy', *Science*, 215(4539), pp. 1525–1530. Available at: https://doi.org/10.1126/science.7063861.

Zerbino, D.R. and Birney, E. (2008) 'Velvet: Algorithms for de novo short read assembly using de Bruijn graphs', *Genome Research*, 18(5), pp. 821–829. Available at: https://doi.org/10.1101/gr.074492.107.

WEB RESOURCES

http://atlasgeneticsoncology.org/index.html

UCSC t2t

https://genome.ucsc.edu/cgi-bin/hgTracks?db=hub_2395475_t2t-chm13-v1.1&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr22%3A1%2D51324926&hgsid=1256505275_H7AIohPRxRPXXz66KqpOk8zPShCV

dbVAR

https://www.ncbi.nlm.nih.gov/dbvar/,
https://www.ncbi.nlm.nih.gov/dbvar/browse/org/?assm=GCF_000001405.39

ENSEMBL

https://www.ensembl.org/Homo_sapiens/Location/Genome

UCSC

https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr10%3A94762681%2D94855547&hgsid=1256754841_RLvwtxH4OqeIEcyuI7syqU5ZNMTa

Cytoband

http://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/ cytoBand.txt.gz

EPI2ME

https://labs.epi2me.io/

HiFi-human-WGS-WDL

https://github.com/PacificBiosciences/HiFi-human-WGS-WDL

hifiasm

https://github.com/chhylp123/hifiasm

gfatools

https://github.com/lh3/gfatools/tree/master

NCBI Taxonomy Browser

https://www.ncbi.nlm.nih.gov/taxonomy/

GATK4

https://gatk.broadinstitute.org/hc/en-us

NA12878

https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1KG_ONT_VIENNA/gaf

ACKNOWLEDGMENTS

This work started in 2000 or in 2001, but before the first human genome publication, in the company Asper Biotech AS, where Georgi Slavin wrote in FoxPro, only for me, a little program to see on-screen, line-by-line PCR primers and shorter sequences' locations along human chromosome 22 in lighter pixels[*]. I thank him from the bottom of my heart for his kindness in writing the most important computer game for me. The second important event happened in 2010, when I saw Sulev Kõks WGS data mapped to the human reference genome[**]. I thank him for sequencing his own genome. In spring 2022 we had a small discussion with Toomas Kivisild about difficulties in finding paternal relationships between bones in burial places: I thank him for sharing the problem. In 2023, we had a discussion with Andres Metspalu about finding cancer signatures in liquid biopsy: I thank him for the discussion, and the k-mer frequencies era will continue in other applications after long-read sequencing becomes less costly and more popular[***].

I thank my supervisors, Prof. Maido Remm and Dr. Lauris Kaplinski, for giving me the resources I needed to complete this work, especially for calculations, because I had created all the processed data we had used in previous analyses on my own, and the computer running time was enormous while handling knowledge that was unknown or unproven before. It is always necessary to conduct statistical proof, even if you know the result in advance. I thank other members of the group for their patience in listening to my endless stories about genetics, car repairs, house insulation, forestry, the economy, gardening, etc.

I thank Hendrik Pavel for open-minded management in Asper Biotech AS and the Estonian Genome Center for human WGS data preparation, and for sharing and storing raw data.

I would like to thank Raivo Hirmo for maintaining my Unix skills, as well as the system administrator of the National Library of Estonia, two IT employees from the head office of the Estonian Savings Bank, and the system administrator of the Library of the Estonian Academy of Sciences from 1993 to 1999.

I also thank my family for their patience because they have heard all my stories, such as about the friendship between Wolbachia and the woodlouse or pea aphid, many times.

---

[*] key bioinformational observations for the first phase of k-mer studies in humans, k-mers uses in reference genome

[**] key bioinformational observations for the second phase of k-mer studies in humans, k-mer uses in NGS

[***] postulation of the third phase of k-mer studies in humans, k-mer uses in NGS with modifications

PUBLICATIONS

CURRICULUM VITAE

Name: Tarmo Puurand
Date of birth: 15th of July 1969, Tallinn, Estonia
Address: Chair of Bioinformatics, Institute of Molecular and Cell Biology, Riia str. 23, 51010, Tartu, Estonia
Phone: +372 5649 3702
e-mail: tarmo.puurand@gmail.com

Education:

1976–1984	Tallinn 27. 8th grade school
1984–1987	Tallinn 1st secondary school
1987–1993	Diploma University of Tartu, 1st year chemistry department, 2 years in army, Department of Virology
2002–2004	MSc University of Tartu, Department of Biotechnology
2020–2024	PhD student University of Tartu, Department of Bioinformatics

Professional employment:

2000–2004	Asper Biotech Ltd., scientist
2004–2024	University of Tartu, Estonian Biocentre, scientist and scientific programmer

Scientific interests:

During my graduate studies, I participated in research on phenol-degrading bacteria and the proteins EBNA1 of the Epstein-Barr virus (EBV) and E2 of the Human Papillomavirus (HPV). At the beginning of my employment at Asper Biotech Ltd., I developed two ultra-fast workflows for designing effective PCR primers and analyzing big chip data. These workflows allowed me to investigate recombination events in human chromosome 22 using SNP markers during my MSc studies. At Asper Biotech Ltd., I also developed a microarray for differentiating strains of the human hepatitis B virus (HBV). Inspired by comparative chromosomal painting of primate genomes, I have since focused on developing methods (k-mer based since 2010) to describe the variability of human repetitive sequences, including VNTRs, SVs and heterochromatin, which I summarize in this doctoral dissertation.

Supervised dissertations:

Signe Kalamees, bachelor's degree, 2007, supervisors Andres Salumets and Tarmo Puurand, Male infertility and microsatelite variations. University of Tartu, Faculty of Biology and Geography, Institute of Molecular and Cell Biology.

Kairit Kolsar, bachelor's degree, 2011, supervisor Tarmo Puurand, Sequencing angiosperms and whole genome duplications in plant evolution. University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology.

Madis Sarapuu, bachelor's degree, 2016, supervisors Tarmo Puurand and Maris Teder-Laving, Copy number determination of the human *Per3* gene tandem repeat from second-generation sequencing data. University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology.

Sylvia Krupp, bachelor's degree, 2017, supervisor Tarmo Puurand, Determining the size of human genome using k-mer methodology. University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology.

Karl-Sander Erss, bachelor's degree, 2017, supervisor Tarmo Puurand, Estimation of average telomere length from second-generation sequencing data. University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology.

Kadri Maal, Master's degree, 2021, supervisors Lili Milani and Tarmo Puurand, Cytochrome P450 2C19 deletions in the Estonian population. University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology.

Farid Naghiyev, bachelor's degree, 2022, supervisor Tarmo Puurand, 5S rDNA copy number in WGS data. University of Tartu, Faculty of Science and Technology, Institute of Technology, Institute of Molecular and Cell Biology.

Carmen Beljaev, bachelor's degree, 2025, supervisor Tarmo Puurand, Finding correlation between chromosomespecific k-mer frequency and centromere length, University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology.

Katariina Ilmoja, bachelor's degree, 2025, supervisor Tarmo Puurand, Construction of the human chromosome 15 centromere phylogenetic tree based on individuals from Human Pangenome version I, University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology.

Anette Stražev, Master's degree, 2025, supervisors Piret van der Sman and Tarmo Puurand, Application of High-Throughput Sequencing for Virus Detection in Estonian Seed Potato and Identification of Potato Virus Y Strains Using Bioinformatic Methods. University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology.

Publications:

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lõhmussaar E, Zernant J, Tõnisson N, Remm M, Mägi R,

Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I. A first-generation linkage disequilibrium map of human chromosome 22. Nature. 2002 Aug 1; 418(6897):544–8. doi: 10.1038/nature00864. Epub 2002 Jul 10. PMID: 12110843.

Kaplinski L, Andreson R, Puurand T, Remm M. MultiPLX: automatic grouping and evaluation of PCR primers. Bioinformatics. 2005 Apr 15; 21(8):1701–2. doi: 10.1093/bioinformatics/bti219. Epub 2004 Dec 14. PMID: 15598831.

Andreson R, Puurand T, Remm M. SNPmasker: automatic masking of SNPs and repeats across eukaryotic genomes. Nucleic Acids Res. 2006 Jul 1;34 (Web Server issue):W651–5. doi: 10.1093/nar/gkl125. PMID: 16845091; PMCID: PMC1538889.

Karmin M, Saag L, Vicente M, Wilson Sayres MA, Järve M, Talas UG, Rootsi S, Ilumäe AM, Mägi R, Mitt M, Pagani L, Puurand T, Faltyskova Z, Clemente F, Cardona A, Metspalu E, Sahakyan H, Yunusbayev B, Hudjashov G, DeGiorgio M, Loogväli EL, Eichstaedt C, Eelmets M, Chaubey G, Tambets K, Litvinov S, Mormina M, Xue Y, Ayub Q, Zoraqi G, Korneliussen TS, Akhatova F, Lachance J, Tishkoff S, Momynaliev K, Ricaut FX, Kusuma P, Razafindrazaka H, Pierron D, Cox MP, Sultana GN, Willerslev R, Muller C, Westaway M, Lambert D, Skaro V, Kovačevic L, Turdikulova S, Dalimova D, Khusainova R, Trofimova N, Akhmetova V, Khidiyatova I, Lichman DV, Isakova J, Pocheshkhova E, Sabitov Z, Barashkov NA, Nymadawa P, Mihailov E, Seng JW, Evseeva I, Migliano AB, Abdullah S, Andriadze G, Primorac D, Atramentova L, Utevska O, Yepiskoposyan L, Marjanovic D, Kushniarevich A, Behar DM, Gilissen C, Vissers L, Veltman JA, Balanovska E, Derenko M, Malyarchuk B, Metspalu A, Fedorova S, Eriksson A, Manica A, Mendez FL, Karafet TM, Veeramah KR, Bradman N, Hammer MF, Osipova LP, Balanovsky O, Khusnutdinova EK, Johnsen K, Remm M, Thomas MG, Tyler-Smith C, Underhill PA, Willerslev E, Nielsen R, Metspalu M, Villems R, Kivisild T. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. 2015 Apr; 25(4): 459–66. doi: 10.1101/gr.186684.114. Epub 2015 Mar 13. PMID: 25770088; PMCID: PMC4381518.

Pajuste FD, Kaplinski L, Möls M, Puurand T, Lepamets M, Remm M. FastGT: an alignment-free method for calling common SNPs directly from raw sequencing reads. Sci Rep. 2017 May 31;7(1):2537. doi: 10.1038/s41598-017-02487-5. PMID: 28566690; PMCID: PMC5451431.

Tasa T, Krebs K, Kals M, Mägi R, Lauschke VM, Haller T, Puurand T, Remm M, Esko T, Metspalu A, Vilo J, Milani L. Genetic variation in the Estonian population: pharmacogenomics study of adverse drug effects using electronic health records. Eur J Hum Genet. 2019 Mar;27(3):442–454. doi: 10.1038/s41431-018-0300-6. Epub 2018 Nov 12. PMID: 30420678; PMCID: PMC6460570.

Puurand T, Kukuškina V, Pajuste FD, Remm M. AluMine: alignment-free method for the discovery of polymorphic Alu element insertions. Mob DNA. 2019 Jul 18; 10:31. doi: 10.1186/s13100-019-0174-3. PMID: 31360240; PMCID: PMC6639938.

Örd T, Puurand T, Örd D, Annilo T, Möls M, Remm M, Örd T. A human-specific VNTR in the TRIB3 promoter causes gene expression variation between individuals. PLoS Genet. 2020 Aug 3;16(8):e1008981. doi: 10.1371/journal.pgen.1008981. PMID: 32745133; PMCID: PMC7425993.

Kaplinski L, Möls M, Puurand T, Pajuste FD, Remm M. KATK: Fast genotyping of rare variants directly from unmapped sequencing reads. Hum Mutat. 2021 Jun;42(6):777–786. doi: 10.1002/humu.24197. Epub 2021 Apr 1. PMID: 33715282.

Kaplinski L, Möls M, Puurand T, Remm M. DOCEST-fast and accurate estimator of human NGS sequencing depth and error rate. Bioinform Adv. 2023 Jul 18;3(1):vbad084. doi: 10.1093/bioadv/vbad084. PMID: 37641716; PMCID: PMC10460481.

ELULOOKIRJELDUS

Nimi:	Tarmo Puurand
Sünniaeg ja koht:	15. juuli 1969, Tallinn, Eesti
Aadress:	Bioinformaatika Õppetool, Tartu Ülikooli Raku- ja Molekulaarbioloogia Instituut, Riia mnt. 23, 51010, Tartu, Eesti
Telefon	+37256493702
e-mail:	tarmo.puurand@gmail.com

Hariduskäik:

1976–1984	Tallinna 27. 8-klassiline kool, hilisem Rahumäe Põhikool
1984–1987	Tallinna 1. Keskkool, endine ja hilisem Gustav Adolfi Gümnaasium
1987–1993	Tartu Ülikool, 1-ne aasta keemia osakond, ajateenistus 2 aastat, bioloogia diplom, viroloogia eriala
2002–2004	Tartu Ülikool, Bioloogia-Geograafia teaduskond MSc, geenitehnoloogia
2020–2024	Tartu Ülikool, Loodus ja Täppisteaduste valdkond, TÜMRI, bioinformaatika doktorant

Erialane teenistuskäik:

2000–2004	Asper Biotech Ltd., teadur
2004–2024	Tartu Ülikool, Eest Biokeskus, Eesti Geenivaramu, Biodata OÜ. teadur, erakorraline teadur, nooremteadur, teaduslik programmeerija

Teadustegevus:

Diplomioõppe raames osalesin fenooli lagundavate bakterite ning Epstein-Barri viiruse (EBV) valgu EBNA1 ja Inimese Papiloomiviiruse (HPV) valgu E2 uuringute juures. Asper Biotech Ltd. töötamise alguses defineerisin heade PCR'i praimerite tegemise ja APEX kiibi andmete analüüsi kaks oma aja kohta ülikiiret töövoogu suurte andmehulkade jaoks, mis võimaldasid MSc õpingute aja sees uurida inimese 22-s kromosoomis toimuvaid rekombinatsioonijuhte esmakordselt SNP markeritega. Asper Biotech Ltd.-s töötasin välja inimese hepatiit B viiruse (HBV) tüvesid eristava APEX mikrokiibi. Edaspidi olen põhisuunana tegelenud primaatide genoomide kromosoomide värvimise võrdlevast (Yunis and Prakash, 1982) ja segmentaalsete duplikatsioonide (Bailey *et al.*, 2002) töödest innustunult inimese korduvate järjestuse sh. VNTR'ide, SV-de ja heterokromatiini varieeruvuse kirjeldamiseks meetodite väljatöötamisega sh. aastast 2010 k-mer põhiste, millest käesolevas doktoritöös ülevaate teen.

Juhendamised:

Signe Kalamees, bakalaureusekraad, 2007, juhendajad Andres Salumets ja Tarmo Puurand, Mehepoolne viljatus ja mikrosatelliitsed variatsioonid. Tartu Ülikool, Bioloogia-geograafia teaduskond, Molekulaar- ja rakubioloogia instituut.

Kairit Kolsar, bakalaureusekraad, 2011, juhendaja Tarmo Puurand, Katteseemnetaimede sekveneerimisprojektid ja genoomiduplikatsioonid taimede evolutsioonis. Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut.

Madis Sarapuu, bakalaureusekraad, 2016, juhendajad Tarmo Puurand ja Maris Teder-Laving, Inimese *Per3* geeni tandeemse korduse koopiaarvu määramine teise põlvkonna sekveneerimisandmetest. Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut.

Sylvia Krupp, bakalaureusekraad, 2017, juhendaja Tarmo Puurand, Inimese genoomi suuruse määramine k-meer metoodikaga. Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut.

Karl-Sander Erss, bakalaureusekraad, 2017, juhendaja Tarmo Puurand, Telomeeri keskmise pikkuse hindamine teise generatsiooni sekveneerimisandmetest. Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut.

Kadri Maal, magistrikraad, 2021, juhendajad Lili Milani ja Tarmo Puurand, Tsütokroom P450 2C19 deletsioonide esinemine Eesti populatsioonis. Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut.

Farid Naghiyev, bakalaureusekraad, 2022, juhendaja Tarmo Puurand, 5S rDNA copy number in WGS data. Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Tehnoloogiainstituut, Molekulaar- ja rakubioloogia instituut.

Carmen Beljaev, bakalaureusekraad, 2025, juhendaja Tarmo Puurand, Kromosoomipõhise k-meer sageduse ja tsentromeeri pikkuse vahelise korrelatsiooni leidmine, Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut.

Katariina Ilmoja, bakalaureusekraad, 2025, juhendaja Tarmo Puurand, Inimese 15. kromosoomi tsentromeeripuu konstrueerimine inimese pangenoomi versioon I indiviidide põhjal, Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut.

Anette Stražev, magistrikraad, 2025, juhendajad Piret van der Sman ja Tarmo Puurand, Kõrge läbilaskevõimega sekveneerimise rakendamine viiruste tuvastamiseks Eesti seemnekartulis ja kartuli Y-viiruse tüvede identifitseerimine bioinformaatiliste meetoditega. Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut.

Publikatsioonid: Loetletud inglisekeelse CV rubriigis publikatsioonid ('Publications')

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

1. **Toivo Maimets**. Studies of human oncoprotein p53. Tartu, 1991, 96 p.
2. **Enn K. Seppet**. Thyroid state control over energy metabolism, ion transport and contractile functions in rat heart. Tartu, 1991, 135 p.
3. **Kristjan Zobel**. Epifüütsete makrosamblike väärtus õhu saastuse indikaatoritena Hamar-Dobani boreaalsetes mägimetsades. Tartu, 1992, 131 lk.
4. **Andres Mäe**. Conjugal mobilization of catabolic plasmids by transposable elements in helper plasmids. Tartu, 1992, 91 p.
5. **Maia Kivisaar**. Studies on phenol degradation genes of *Pseudomonas* sp. strain EST 1001. Tartu, 1992, 61 p.
6. **Allan Nurk**. Nucleotide sequences of phenol degradative genes from *Pseudomonas sp.* strain EST 1001 and their transcriptional activation in *Pseudomonas putida.* Tartu, 1992, 72 p.
7. **Ülo Tamm**. The genus *Populus* L. in Estonia: variation of the species biology and introduction. Tartu, 1993, 91 p.
8. **Jaanus Remme**. Studies on the peptidyltransferase centre of the *E.coli* ribosome. Tartu, 1993, 68 p.
9. **Ülo Langel**. Galanin and galanin antagonists. Tartu, 1993, 97 p.
10. **Arvo Käärd**. The development of an automatic online dynamic fluorescense-based pH-dependent fiber optic penicillin flowthrought biosensor for the control of the benzylpenicillin hydrolysis. Tartu, 1993, 117 p.
11. **Lilian Järvekülg**. Antigenic analysis and development of sensitive immunoassay for potato viruses. Tartu, 1993, 147 p.
12. **Jaak Palumets**. Analysis of phytomass partition in Norway spruce. Tartu, 1993, 47 p.
13. **Arne Sellin**. Variation in hydraulic architecture of *Picea abies* (L.) Karst. trees grown under different enviromental conditions. Tartu, 1994, 119 p.
13. **Mati Reeben**. Regulation of light neurofilament gene expression. Tartu, 1994, 108 p.
14. **Urmas Tartes**. Respiration rhytms in insects. Tartu, 1995, 109 p.
15. **Ülo Puurand**. The complete nucleotide sequence and infections *in vitro* transcripts from cloned cDNA of a potato A potyvirus. Tartu, 1995, 96 p.
16. **Peeter Hõrak**. Pathways of selection in avian reproduction: a functional framework and its application in the population study of the great tit (*Parus major*). Tartu, 1995, 118 p.
17. **Erkki Truve**. Studies on specific and broad spectrum virus resistance in transgenic plants. Tartu, 1996, 158 p.
18. **Illar Pata**. Cloning and characterization of human and mouse ribosomal protein S6-encoding genes. Tartu, 1996, 60 p.
19. **Ülo Niinemets**. Importance of structural features of leaves and canopy in determining species shade-tolerance in temperature deciduous woody taxa. Tartu, 1996, 150 p.

20. **Ants Kurg**. Bovine leukemia virus: molecular studies on the packaging region and DNA diagnostics in cattle. Tartu, 1996, 104 p.
21. **Ene Ustav**. E2 as the modulator of the BPV1 DNA replication. Tartu, 1996, 100 p.
22. **Aksel Soosaar**. Role of helix-loop-helix and nuclear hormone receptor transcription factors in neurogenesis. Tartu, 1996, 109 p.
23. **Maido Remm**. Human papillomavirus type 18: replication, transformation and gene expression. Tartu, 1997, 117 p.
24. **Tiiu Kull**. Population dynamics in *Cypripedium calceolus* L. Tartu, 1997, 124 p.
25. **Kalle Olli**. Evolutionary life-strategies of autotrophic planktonic microorganisms in the Baltic Sea. Tartu, 1997, 180 p.
26. **Meelis Pärtel**. Species diversity and community dynamics in calcareous grassland communities in Western Estonia. Tartu, 1997, 124 p.
27. **Malle Leht**. The Genus *Potentilla* L. in Estonia, Latvia and Lithuania: distribution, morphology and taxonomy. Tartu, 1997, 186 p.
28. **Tanel Tenson**. Ribosomes, peptides and antibiotic resistance. Tartu, 1997, 80 p.
29. **Arvo Tuvikene**. Assessment of inland water pollution using biomarker responses in fish *in vivo* and *in vitro.* Tartu, 1997, 160 p.
30. **Urmas Saarma**. Tuning ribosomal elongation cycle by mutagenesis of 23S rRNA. Tartu, 1997, 134 p.
31. **Henn Ojaveer**. Composition and dynamics of fish stocks in the gulf of Riga ecosystem. Tartu, 1997, 138 p.
32. **Lembi Lõugas**. Post-glacial development of vertebrate fauna in Estonian water bodies. Tartu, 1997, 138 p.
33. **Margus Pooga**. Cell penetrating peptide, transportan, and its predecessors, galanin-based chimeric peptides. Tartu, 1998, 110 p.
34. **Andres Saag**. Evolutionary relationships in some cetrarioid genera (Lichenized Ascomycota). Tartu, 1998, 196 p.
35. **Aivar Liiv**. Ribosomal large subunit assembly *in vivo*. Tartu, 1998, 158 p.
36. **Tatjana Oja**. Isoenzyme diversity and phylogenetic affinities among the eurasian annual bromes (*Bromus* L., Poaceae). Tartu, 1998, 92 p.
37. **Mari Moora**. The influence of arbuscular mycorrhizal (AM) symbiosis on the competition and coexistence of calcareous grassland plant species. Tartu, 1998, 78 p.
38. **Olavi Kurina**. Fungus gnats in Estonia (*Diptera: Bolitophilidae, Keroplatidae, Macroceridae, Ditomyiidae, Diadocidiidae, Mycetophilidae*). Tartu, 1998, 200 p.
39. **Andrus Tasa**. Biological leaching of shales: black shale and oil shale. Tartu, 1998, 98 p.
40. **Arnold Kristjuhan**. Studies on transcriptional activator properties of tumor suppressor protein p53. Tartu, 1998, 86 p.
41. **Sulev Ingerpuu**. Characterization of some human myeloid cell surface and nuclear differentiation antigens. Tartu, 1998, 163 p.

42. **Veljo Kisand**. Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa**. Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa**. Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik**. Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo**. Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo**. Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots**. Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero**. Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees**. Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks**. Cholecystokinin (CCK) – induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and serotonin. Tartu, 1999, 123 p.
52. **Ebe Sild**. Impact of increasing concentrations of O_3 and CO_2 on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva**. Electron microscopical analysis of the synaptonemal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna**. Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro**. Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane**. Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm**. Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg**. Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild**. The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu**. Studies of the TOL plasmid transcription factor XylS. Tartu, 2000, 88 p.
61. **Dina Lepik**. Modulation of viral DNA replication by tumor suppressor protein p53. Tartu, 2000, 106 p.
62. **Kai Vellak**. Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu, 2000, 122 p.

63. **Jonne Kotta**. Impact of eutrophication and biological invasionas on the structure and functions of benthic macrofauna. Tartu, 2000, 160 p.
64. **Georg Martin**. Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000, 139 p.
65. **Silvia Sepp**. Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira**. On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000, 96 p.
67. **Priit Zingel**. The role of planktonic ciliates in lake ecosystems. Tartu, 2001, 111 p.
68. **Tiit Teder**. Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu, 2001, 122 p.
69. **Hannes Kollist**. Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu, 2001, 80 p.
70. **Reet Marits**. Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu, 2001, 112 p.
71. **Vallo Tilgar**. Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major,* breeding in Nothern temperate forests. Tartu, 2002, 126 p.
72. **Rita Hõrak**. Regulation of transposition of transposon Tn*4652* in *Pseudomonas putida*. Tartu, 2002, 108 p.
73. **Liina Eek-Piirsoo**. The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002, 74 p.
74. **Krõõt Aasamaa**. Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002, 110 p.
75. **Nele Ingerpuu**. Bryophyte diversity and vascular plants. Tartu, 2002, 112 p.
76. **Neeme Tõnisson**. Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002, 124 p.
77. **Margus Pensa**. Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003, 110 p.
78. **Asko Lõhmus**. Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003, 168 p.
79. **Viljar Jaks**. p53 – a switch in cellular circuit. Tartu, 2003, 160 p.
80. **Jaana Männik**. Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003, 140 p.
81. **Marek Sammul**. Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003, 159 p
82. **Ivar Ilves**. Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003, 89 p.
83. **Andres Männik**. Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003, 109 p.

84. **Ivika Ostonen**. Fine root structure, dynamics and proportion in net primary production of Norway spruce forest ecosystem in relation to site conditions. Tartu, 2003, 158 p.
85. **Gudrun Veldre**. Somatic status of 12–15-year-old Tartu schoolchildren. Tartu, 2003, 199 p.
86. **Ülo Väli.** The greater spotted eagle *Aquila clanga* and the lesser spotted eagle *A. pomarina*: taxonomy, phylogeography and ecology. Tartu, 2004, 159 p.
87. **Aare Abroi**. The determinants for the native activities of the bovine papillomavirus type 1 E2 protein are separable. Tartu, 2004, 135 p.
88. **Tiina Kahre**. Cystic fibrosis in Estonia. Tartu, 2004, 116 p.
89. **Helen Orav-Kotta**. Habitat choice and feeding activity of benthic suspension feeders and mesograzers in the northern Baltic Sea. Tartu, 2004, 117 p.
90. **Maarja Öpik**. Diversity of arbuscular mycorrhizal fungi in the roots of perennial plants and their effect on plant performance. Tartu, 2004, 175 p.
91. **Kadri Tali**. Species structure of *Neotinea ustulata*. Tartu, 2004, 109 p.
92. **Kristiina Tambets**. Towards the understanding of post-glacial spread of human mitochondrial DNA haplogroups in Europe and beyond: a phylogeographic approach. Tartu, 2004, 163 p.
93. **Arvi Jõers**. Regulation of p53-dependent transcription. Tartu, 2004, 103 p.
94. **Lilian Kadaja**. Studies on modulation of the activity of tumor suppressor protein p53. Tartu, 2004, 103 p.
95. **Jaak Truu**. Oil shale industry wastewater: impact on river microbial community and possibilities for bioremediation. Tartu, 2004, 128 p.
96. **Maire Peters**. Natural horizontal transfer of the *pheBA* operon. Tartu, 2004, 105 p.
97. **Ülo Maiväli**. Studies on the structure-function relationship of the bacterial ribosome. Tartu, 2004, 130 p.
98. **Merit Otsus**. Plant community regeneration and species diversity in dry calcareous grasslands. Tartu, 2004, 103 p.
99. **Mikk Heidemaa**. Systematic studies on sawflies of the genera *Dolerus, Empria,* and *Caliroa* (Hymenoptera: Tenthredinidae). Tartu, 2004, 167 p.
100. **Ilmar Tõnno**. The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N_2 fixation in some Estonian lakes. Tartu, 2004, 111 p.
101. **Lauri Saks**. Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004, 144 p.
102. **Siiri Rootsi**. Human Y-chromosomal variation in European populations. Tartu, 2004, 142 p.
103. **Eve Vedler**. Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.
104. **Andres Tover**. Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 126 p.
105. **Helen Udras**. Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005, 100 p.

106. **Ave Suija**. Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005, 162 p.
107. **Piret Lõhmus**. Forest lichens and their substrata in Estonia. Tartu, 2005, 162 p.
108. **Inga Lips**. Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005, 156 p.
109. **Krista Kaasik**. Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005, 121 p.
110. **Juhan Javoiš**. The effects of experience on host acceptance in ovipositing moths. Tartu, 2005, 112 p.
111. **Tiina Sedman**. Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005, 103 p.
112. **Ruth Aguraiuja**. Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005, 112 p.
113. **Riho Teras**. Regulation of transcription from the fusion promoters generated by transposition of Tn*4652* into the upstream region of *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 106 p.
114. **Mait Metspalu**. Through the course of prehistory in India: tracing the mtDNA trail. Tartu, 2005, 138 p.
115. **Elin Lõhmussaar**. The comparative patterns of linkage disequilibrium in European populations and its implication for genetic association studies. Tartu, 2006, 124 p.
116. **Priit Kupper**. Hydraulic and environmental limitations to leaf water relations in trees with respect to canopy position. Tartu, 2006, 126 p.
117. **Heili Ilves**. Stress-induced transposition of Tn*4652* in *Pseudomonas Putida.* Tartu, 2006, 120 p.
118. **Silja Kuusk**. Biochemical properties of Hmi1p, a DNA helicase from *Saccharomyces cerevisiae* mitochondria. Tartu, 2006, 126 p.
119. **Kersti Püssa**. Forest edges on medium resolution landsat thematic mapper satellite images. Tartu, 2006, 90 p.
120. **Lea Tummeleht**. Physiological condition and immune function in great tits (*Parus major* l.): Sources of variation and trade-offs in relation to growth. Tartu, 2006, 94 p.
121. **Toomas Esperk**. Larval instar as a key element of insect growth schedules. Tartu, 2006, 186 p.
122. **Harri Valdmann**. Lynx (*Lynx lynx*) and wolf (*Canis lupus*) in the Baltic region: Diets, helminth parasites and genetic variation. Tartu, 2006. 102 p.
123. **Priit Jõers**. Studies of the mitochondrial helicase Hmi1p in *Candida albicans* and *Saccharomyces cerevisia*. Tartu, 2006. 113 p.
124. **Kersti Lilleväli**. Gata3 and Gata2 in inner ear development. Tartu, 2007, 123 p.
125. **Kai Rünk**. Comparative ecology of three fern species: *Dryopteris carthusiana* (Vill.) H.P. Fuchs, *D. expansa* (C. Presl) Fraser-Jenkins & Jermy and *D. dilatata* (Hoffm.) A. Gray (Dryopteridaceae). Tartu, 2007, 143 p.

126. **Aveliina Helm**. Formation and persistence of dry grassland diversity: role of human history and landscape structure. Tartu, 2007, 89 p.
127. **Leho Tedersoo**. Ectomycorrhizal fungi: diversity and community structure in Estonia, Seychelles and Australia. Tartu, 2007, 233 p.
128. **Marko Mägi**. The habitat-related variation of reproductive performance of great tits in a deciduous-coniferous forest mosaic: looking for causes and consequences. Tartu, 2007, 135 p.
129. **Valeria Lulla**. Replication strategies and applications of Semliki Forest virus. Tartu, 2007, 109 p.
130. **Ülle Reier**. Estonian threatened vascular plant species: causes of rarity and conservation. Tartu, 2007, 79 p.
131. **Inga Jüriado**. Diversity of lichen species in Estonia: influence of regional and local factors. Tartu, 2007, 171 p.
132. **Tatjana Krama**. Mobbing behaviour in birds: costs and reciprocity based cooperation. Tartu, 2007, 112 p.
133. **Signe Saumaa**. The role of DNA mismatch repair and oxidative DNA damage defense systems in avoidance of stationary phase mutations in *Pseudomonas putida.* Tartu, 2007, 172 p.
134. **Reedik Mägi**. The linkage disequilibrium and the selection of genetic markers for association studies in european populations. Tartu, 2007, 96 p.
135. **Priit Kilgas**. Blood parameters as indicators of physiological condition and skeletal development in great tits (*Parus major*): natural variation and application in the reproductive ecology of birds. Tartu, 2007, 129 p.
136. **Anu Albert**. The role of water salinity in structuring eastern Baltic coastal fish communities. Tartu, 2007, 95 p.
137. **Kärt Padari**. Protein transduction mechanisms of transportans. Tartu, 2008, 128 p.
138. **Siiri-Lii Sandre**. Selective forces on larval colouration in a moth. Tartu, 2008, 125 p.
139. **Ülle Jõgar**. Conservation and restoration of semi-natural floodplain meadows and their rare plant species. Tartu, 2008, 99 p.
140. **Lauri Laanisto**. Macroecological approach in vegetation science: generality of ecological relationships at the global scale. Tartu, 2008, 133 p.
141. **Reidar Andreson**. Methods and software for predicting PCR failure rate in large genomes. Tartu, 2008, 105 p.
142. **Birgot Paavel**. Bio-optical properties of turbid lakes. Tartu, 2008, 175 p.
143. **Kaire Torn**. Distribution and ecology of charophytes in the Baltic Sea. Tartu, 2008, 98 p.
144. **Vladimir Vimberg**. Peptide mediated macrolide resistance. Tartu, 2008, 190 p.
145. **Daima Örd**. Studies on the stress-inducible pseudokinase TRB3, a novel inhibitor of transcription factor ATF4. Tartu, 2008, 108 p.
146. **Lauri Saag**. Taxonomic and ecologic problems in the genus *Lepraria* (*Stereocaulaceae*, lichenised *Ascomycota*). Tartu, 2008, 175 p.

147. **Ulvi Karu**. Antioxidant protection, carotenoids and coccidians in greenfinches – assessment of the costs of immune activation and mechanisms of parasite resistance in a passerine with carotenoid-based ornaments. Tartu, 2008, 124 p.
148. **Jaanus Remm**. Tree-cavities in forests: density, characteristics and occupancy by animals. Tartu, 2008, 128 p.
149. **Epp Moks**. Tapeworm parasites *Echinococcus multilocularis* and *E. granulosus* in Estonia: phylogenetic relationships and occurrence in wild carnivores and ungulates. Tartu, 2008, 82 p.
150. **Eve Eensalu**. Acclimation of stomatal structure and function in tree canopy: effect of light and CO_2 concentration. Tartu, 2008, 108 p.
151. **Janne Pullat**. Design, functionlization and application of an *in situ* synthesized oligonucleotide microarray. Tartu, 2008, 108 p.
152. **Marta Putrinš**. Responses of *Pseudomonas putida* to phenol-induced metabolic and stress signals. Tartu, 2008, 142 p.
153. **Marina Semtšenko**. Plant root behaviour: responses to neighbours and physical obstructions. Tartu, 2008, 106 p.
154. **Marge Starast**. Influence of cultivation techniques on productivity and fruit quality of some *Vaccinium* and *Rubus* taxa. Tartu, 2008, 154 p.
155. **Age Tats**. Sequence motifs influencing the efficiency of translation. Tartu, 2009, 104 p.
156. **Radi Tegova**. The role of specialized DNA polymerases in mutagenesis in *Pseudomonas putida.* Tartu, 2009, 124 p.
157. **Tsipe Aavik**. Plant species richness, composition and functional trait pattern in agricultural landscapes – the role of land use intensity and landscape structure. Tartu, 2009, 112 p.
158. **Kaja Kiiver**. Semliki forest virus based vectors and cell lines for studying the replication and interactions of alphaviruses and hepaciviruses. Tartu, 2009, 104 p.
159. **Meelis Kadaja**. Papillomavirus Replication Machinery Induces Genomic Instability in its Host Cell. Tartu, 2009, 126 p.
160. **Pille Hallast**. Human and chimpanzee Luteinizing hormone/Chorionic Gonadotropin beta (*LHB/CGB*) gene clusters: diversity and divergence of young duplicated genes. Tartu, 2009, 168 p.
161. **Ain Vellak**. Spatial and temporal aspects of plant species conservation. Tartu, 2009, 86 p.
162. **Triinu Remmel**. Body size evolution in insects with different colouration strategies: the role of predation risk. Tartu, 2009, 168 p.
163. **Jaana Salujõe**. Zooplankton as the indicator of ecological quality and fish predation in lake ecosystems. Tartu, 2009, 129 p.
164. **Ele Vahtmäe**. Mapping benthic habitat with remote sensing in optically complex coastal environments. Tartu, 2009, 109 p.
165. **Liisa Metsamaa**. Model-based assessment to improve the use of remote sensing in recognition and quantitative mapping of cyanobacteria. Tartu, 2009, 114 p.

166. **Pille Säälik**. The role of endocytosis in the protein transduction by cell-penetrating peptides. Tartu, 2009, 155 p.
167. **Lauri Peil**. Ribosome assembly factors in *Escherichia coli.* Tartu, 2009, 147 p.
168. **Lea Hallik**. Generality and specificity in light harvesting, carbon gain capacity and shade tolerance among plant functional groups. Tartu, 2009, 99 p.
169. **Mariliis Tark**. Mutagenic potential of DNA damage repair and tolerance mechanisms under starvation stress. Tartu, 2009, 191 p.
170. **Riinu Rannap**. Impacts of habitat loss and restoration on amphibian populations. Tartu, 2009, 117 p.
171. **Maarja Adojaan**. Molecular variation of HIV-1 and the use of this knowledge in vaccine development. Tartu, 2009, 95 p.
172. **Signe Altmäe**. Genomics and transcriptomics of human induced ovarian folliculogenesis. Tartu, 2010, 179 p.
173. **Triin Suvi**. Mycorrhizal fungi of native and introduced trees in the Seychelles Islands. Tartu, 2010, 107 p.
174. **Velda Lauringson**. Role of suspension feeding in a brackish-water coastal sea. Tartu, 2010, 123 p.
175. **Eero Talts**. Photosynthetic cyclic electron transport – measurement and variably proton-coupled mechanism. Tartu, 2010, 121 p.
176. **Mari Nelis**. Genetic structure of the Estonian population and genetic distance from other populations of European descent. Tartu, 2010, 97 p.
177. **Kaarel Krjutškov**. Arrayed Primer Extension-2 as a multiplex PCR-based method for nucleic acid variation analysis: method and applications. Tartu, 2010, 129 p.
178. **Egle Köster**. Morphological and genetical variation within species complexes: *Anthyllis vulneraria* s. l. and *Alchemilla vulgaris* (coll.). Tartu, 2010, 101 p.
179. **Erki Õunap**. Systematic studies on the subfamily Sterrhinae (Lepidoptera: Geometridae). Tartu, 2010, 111 p.
180. **Merike Jõesaar**. Diversity of key catabolic genes at degradation of phenol and *p*-cresol in pseudomonads. Tartu, 2010, 125 p.
181. **Kristjan Herkül**. Effects of physical disturbance and habitat-modifying species on sediment properties and benthic communities in the northern Baltic Sea. Tartu, 2010, 123 p.
182. **Arto Pulk**. Studies on bacterial ribosomes by chemical modification approaches. Tartu, 2010, 161 p.
183. **Maria Põllupüü**. Ecological relations of cladocerans in a brackish-water ecosystem. Tartu, 2010, 126 p.
184. **Toomas Silla**. Study of the segregation mechanism of the Bovine Papillomavirus Type 1. Tartu, 2010, 188 p.
185. **Gyaneshwer Chaubey**. The demographic history of India: A perspective based on genetic evidence. Tartu, 2010, 184 p.

186. **Katrin Kepp**. Genes involved in cardiovascular traits: detection of genetic variation in Estonian and Czech populations. Tartu, 2010, 164 p.
187. **Virve Sõber**. The role of biotic interactions in plant reproductive performance. Tartu, 2010, 92 p.
188. **Kersti Kangro**. The response of phytoplankton community to the changes in nutrient loading. Tartu, 2010, 144 p.
189. **Joachim M. Gerhold**. Replication and Recombination of mitochondrial DNA in Yeast. Tartu, 2010, 120 p.
190. **Helen Tammert**. Ecological role of physiological and phylogenetic diversity in aquatic bacterial communities. Tartu, 2010, 140 p.
191. **Elle Rajandu**. Factors determining plant and lichen species diversity and composition in Estonian *Calamagrostis* and *Hepatica* site type forests. Tartu, 2010, 123 p.
192. **Paula Ann Kivistik**. ColR-ColS signalling system and transposition of Tn*4652* in the adaptation of *Pseudomonas putida.* Tartu, 2010, 118 p.
193. **Siim Sõber**. Blood pressure genetics: from candidate genes to genome-wide association studies. Tartu, 2011, 120 p.
194. **Kalle Kipper**. Studies on the role of helix 69 of 23S rRNA in the factor-dependent stages of translation initiation, elongation, and termination. Tartu, 2011, 178 p.
195. **Triinu Siibak**. Effect of antibiotics on ribosome assembly is indirect. Tartu, 2011, 134 p.
196. **Tambet Tõnissoo**. Identification and molecular analysis of the role of guanine nucleotide exchange factor RIC-8 in mouse development and neural function. Tartu, 2011, 110 p.
197. **Helin Räägel**. Multiple faces of cell-penetrating peptides – their intracellular trafficking, stability and endosomal escape during protein transduction. Tartu, 2011, 161 p.
198. **Andres Jaanus**. Phytoplankton in Estonian coastal waters – variability, trends and response to environmental pressures. Tartu, 2011, 157 p.
199. **Tiit Nikopensius**. Genetic predisposition to nonsyndromic orofacial clefts. Tartu, 2011, 152 p.
200. **Signe Värv**. Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.
201. **Kristjan Välk**. Gene expression profiling and genome-wide association studies of non-small cell lung cancer. Tartu, 2011, 98 p.
202. **Arno Põllumäe**. Spatio-temporal patterns of native and invasive zooplankton species under changing climate and eutrophication conditions. Tartu, 2011, 153 p.
203. **Egle Tammeleht**. Brown bear (*Ursus arctos*) population structure, demographic processes and variations in diet in northern Eurasia. Tartu, 2011, 143 p.
205. **Teele Jairus**. Species composition and host preference among ectomycorrhizal fungi in Australian and African ecosystems. Tartu, 2011, 106 p.

206. **Kessy Abarenkov**. PlutoF – cloud database and computing services supporting biological research. Tartu, 2011, 125 p.
207. **Marina Grigorova**. Fine-scale genetic variation of follicle-stimulating hormone beta-subunit coding gene (*FSHB*) and its association with reproductive health. Tartu, 2011, 184 p.
208. **Anu Tiitsaar**. The effects of predation risk and habitat history on butterfly communities. Tartu, 2011, 97 p.
209. **Elin Sild**. Oxidative defences in immunoecological context: validation and application of assays for nitric oxide production and oxidative burst in a wild passerine. Tartu, 2011, 105 p.
210. **Irja Saar**. The taxonomy and phylogeny of the genera *Cystoderma* and *Cystodermella* (Agaricales, Fungi). Tartu, 2012, 167 p.
211. **Pauli Saag**. Natural variation in plumage bacterial assemblages in two wild breeding passerines. Tartu, 2012, 113 p.
212. **Aleksei Lulla**. Alphaviral nonstructural protease and its polyprotein substrate: arrangements for the perfect marriage. Tartu, 2012, 143 p.
213. **Mari Järve**. Different genetic perspectives on human history in Europe and the Caucasus: the stories told by uniparental and autosomal markers. Tartu, 2012, 119 p.
214. **Ott Scheler**. The application of tmRNA as a marker molecule in bacterial diagnostics using microarray and biosensor technology. Tartu, 2012, 93 p.
215. **Anna Balikova**. Studies on the functions of tumor-associated mucin-like leukosialin (CD43) in human cancer cells. Tartu, 2012, 129 p.
216. **Triinu Kõressaar**. Improvement of PCR primer design for detection of prokaryotic species. Tartu, 2012, 83 p.
217. **Tuul Sepp**. Hematological health state indices of greenfinches: sources of individual variation and responses to immune system manipulation. Tartu, 2012, 117 p.
218. **Rya Ero**. Modifier view of the bacterial ribosome. Tartu, 2012, 146 p.
219. **Mohammad Bahram**. Biogeography of ectomycorrhizal fungi across different spatial scales. Tartu, 2012, 165 p.
220. **Annely Lorents**. Overcoming the plasma membrane barrier: uptake of amphipathic cell-penetrating peptides induces influx of calcium ions and downstream responses. Tartu, 2012, 113 p.
221. **Katrin Männik**. Exploring the genomics of cognitive impairment: whole-genome SNP genotyping experience in Estonian patients and general population. Tartu, 2012, 171 p.
222. **Marko Prous**. Taxonomy and phylogeny of the sawfly genus *Empria* (Hymenoptera, Tenthredinidae). Tartu, 2012, 192 p.
223. **Triinu Visnapuu**. Levansucrases encoded in the genome of *Pseudomonas syringae* pv. tomato DC3000: heterologous expression, biochemical characterization, mutational analysis and spectrum of polymerization products. Tartu, 2012, 160 p.
224. **Nele Tamberg**. Studies on Semliki Forest virus replication and pathogenesis. Tartu, 2012, 109 p.

225. **Tõnu Esko**. Novel applications of SNP array data in the analysis of the genetic structure of Europeans and in genetic association studies. Tartu, 2012, 149 p.
226. **Timo Arula**. Ecology of early life-history stages of herring *Clupea harengus membras* in the northeastern Baltic Sea. Tartu, 2012, 143 p.
227. **Inga Hiiesalu**. Belowground plant diversity and coexistence patterns in grassland ecosystems. Tartu, 2012, 130 p.
228. **Kadri Koorem**. The influence of abiotic and biotic factors on small-scale plant community patterns and regeneration in boreonemoral forest. Tartu, 2012, 114 p.
229. **Liis Andresen**. Regulation of virulence in plant-pathogenic pectobacteria. Tartu, 2012, 122 p.
230. **Kaupo Kohv**. The direct and indirect effects of management on boreal forest structure and field layer vegetation. Tartu, 2012, 124 p.
231. **Mart Jüssi**. Living on an edge: landlocked seals in changing climate. Tartu, 2012, 114 p.
232. **Riina Klais**. Phytoplankton trends in the Baltic Sea. Tartu, 2012, 136 p.
233. **Rauno Veeroja**. Effects of winter weather, population density and timing of reproduction on life-history traits and population dynamics of moose (*Alces alces*) in Estonia. Tartu, 2012, 92 p.
234. **Marju Keis**. Brown bear (*Ursus arctos*) phylogeography in northern Eurasia. Tartu, 2013, 142 p.
235. **Sergei Põlme**. Biogeography and ecology of *alnus*- associated ectomycorrhizal fungi – from regional to global scale. Tartu, 2013, 90 p.
236. **Liis Uusküla**. Placental gene expression in normal and complicated pregnancy. Tartu, 2013, 173 p.
237. **Marko Lõoke**. Studies on DNA replication initiation in *Saccharomyces cerevisiae.* Tartu, 2013, 112 p.
238. **Anne Aan**. Light- and nitrogen-use and biomass allocation along productivity gradients in multilayer plant communities. Tartu, 2013, 127 p.
239. **Heidi Tamm**. Comprehending phylogenetic diversity – case studies in three groups of ascomycetes. Tartu, 2013, 136 p.
240. **Liina Kangur**. High-Pressure Spectroscopy Study of Chromophore-Binding Hydrogen Bonds in Light-Harvesting Complexes of Photosynthetic Bacteria. Tartu, 2013, 150 p.
241. **Margus Leppik**. Substrate specificity of the multisite specific pseudouridine synthase RluD. Tartu, 2013, 111 p.
242. **Lauris Kaplinski**. The application of oligonucleotide hybridization model for PCR and microarray optimization. Tartu, 2013, 103 p.
243. **Merli Pärnoja**. Patterns of macrophyte distribution and productivity in coastal ecosystems: effect of abiotic and biotic forcing. Tartu, 2013, 155 p.
244. **Tõnu Margus**. Distribution and phylogeny of the bacterial translational GTPases and the Mqsr/YgiT regulatory system. Tartu, 2013, 126 p.
245. **Pille Mänd**. Light use capacity and carbon and nitrogen budget of plants: remote assessment and physiological determinants. Tartu, 2013, 128 p.

246. **Mario Plaas**. Animal model of Wolfram Syndrome in mice: behavioural, biochemical and psychopharmacological characterization. Tartu, 2013, 144 p.
247. **Georgi Hudjašov**. Maps of mitochondrial DNA, Y-chromosome and tyrosinase variation in Eurasian and Oceanian populations. Tartu, 2013, 115 p.
248. **Mari Lepik**. Plasticity to light in herbaceous plants and its importance for community structure and diversity. Tartu, 2013, 102 p.
249. **Ede Leppik**. Diversity of lichens in semi-natural habitats of Estonia. Tartu, 2013, 151 p.
250. **Ülle Saks**. Arbuscular mycorrhizal fungal diversity patterns in boreonemoral forest ecosystems. Tartu, 2013, 151 p.
251. **Eneli Oitmaa**. Development of arrayed primer extension microarray assays for molecular diagnostic applications. Tartu, 2013, 147 p.
252. **Jekaterina Jutkina**. The horizontal gene pool for aromatics degradation: bacterial catabolic plasmids of the Baltic Sea aquatic system. Tartu, 2013, 121 p.
253. **Helen Vellau**. Reaction norms for size and age at maturity in insects: rules and exceptions. Tartu, 2014, 132 p.
254. **Randel Kreitsberg**. Using biomarkers in assessment of environmental contamination in fish – new perspectives. Tartu, 2014, 107 p.
255. **Krista Takkis**. Changes in plant species richness and population performance in response to habitat loss and fragmentation.Tartu, 2014, 141 p.
256. **Liina Nagirnaja**. Global and fine-scale genetic determinants of recurrent pregnancy loss. Tartu, 2014, 211 p.
257. **Triin Triisberg**. Factors influencing the re-vegetation of abandoned extracted peatlands in Estonia. Tartu, 2014, 133 p.
258. **Villu Soon**. A phylogenetic revision of the *Chrysis ignita* species group (Hymenoptera: Chrysididae) with emphasis on the northern European fauna. Tartu, 2014, 211 p.
259. **Andrei Nikonov**. RNA-Dependent RNA Polymerase Activity as a Basis for the Detection of Positive-Strand RNA Viruses by Vertebrate Host Cells. Tartu, 2014, 207 p.
260. **Eele Õunapuu-Pikas**. Spatio-temporal variability of leaf hydraulic conductance in woody plants: ecophysiological consequences. Tartu, 2014, 135 p.
261. **Marju Männiste**. Physiological ecology of greenfinches: information content of feathers in relation to immune function and behavior. Tartu, 2014, 121 p.
262. **Katre Kets**. Effects of elevated concentrations of CO_2 and O_3 on leaf photosynthetic parameters in *Populus tremuloides*: diurnal, seasonal and interannual patterns. Tartu, 2014, 115 p.
263. **Külli Lokko**. Seasonal and spatial variability of zoopsammon communities in relation to environmental parameters. Tartu, 2014, 129 p.
264. **Olga Žilina**. Chromosomal microarray analysis as diagnostic tool: Estonian experience. Tartu, 2014, 152 p.

265. **Kertu Lõhmus**. Colonisation ecology of forest-dwelling vascular plants and the conservation value of rural manor parks. Tartu, 2014, 111 p.
266. **Anu Aun**. Mitochondria as integral modulators of cellular signaling. Tartu, 2014, 167 p.
267. **Chandana Basu Mallick**. Genetics of adaptive traits and gender-specific demographic processes in South Asian populations. Tartu, 2014, 160 p.
268. **Riin Tamme**. The relationship between small-scale environmental heterogeneity and plant species diversity. Tartu, 2014, 130 p.
269. **Liina Remm**. Impacts of forest drainage on biodiversity and habitat quality: implications for sustainable management and conservation. Tartu, 2015, 126 p.
270. **Tiina Talve**. Genetic diversity and taxonomy within the genus *Rhinanthus.* Tartu, 2015, 106 p.
271. **Mehis Rohtla**. Otolith sclerochronological studies on migrations, spawning habitat preferences and age of freshwater fishes inhabiting the Baltic Sea. Tartu, 2015, 137 p.
272. **Alexey Reshchikov**. The world fauna of the genus *Lathrolestes* (Hymenoptera, Ichneumonidae). Tartu, 2015, 247 p.
273. **Martin Pook**. Studies on artificial and extracellular matrix protein-rich surfaces as regulators of cell growth and differentiation. Tartu, 2015, 142 p.
274. **Mai Kukumägi**. Factors affecting soil respiration and its components in silver birch and Norway spruce stands. Tartu, 2015, 155 p.
275. **Helen Karu**. Development of ecosystems under human activity in the North-East Estonian industrial region: forests on post-mining sites and bogs. Tartu, 2015, 152 p.
276. **Hedi Peterson**. Exploiting high-throughput data for establishing relationships between genes. Tartu, 2015, 186 p.
277. **Priit Adler**. Analysis and visualisation of large scale microarray data, Tartu, 2015, 126 p.
278. **Aigar Niglas**. Effects of environmental factors on gas exchange in deciduous trees: focus on photosynthetic water-use efficiency. Tartu, 2015, 152 p.
279. **Silja Laht**. Classification and identification of conopeptides using profile hidden Markov models and position-specific scoring matrices. Tartu, 2015, 100 p.
280. **Martin Kesler**. Biological characteristics and restoration of Atlantic salmon *Salmo salar* populations in the Rivers of Northern Estonia. Tartu, 2015, 97 p.
281. **Pratyush Kumar Das**. Biochemical perspective on alphaviral nonstructural protein 2: a tale from multiple domains to enzymatic profiling. Tartu, 2015, 205 p
282. **Priit Palta**. Computational methods for DNA copy number detection. Tartu, 2015, 130 p.
283. **Julia Sidorenko**. Combating DNA damage and maintenance of genome integrity in pseudomonads. Tartu, 2015, 174 p.

284. **Anastasiia Kovtun-Kante**. Charophytes of Estonian inland and coastal waters: distribution and environmental preferences. Tartu, 2015, 97 p.
285. **Ly Lindman**. The ecology of protected butterfly species in Estonia. Tartu, 2015, 171 p.
286. **Jaanis Lodjak**. Association of Insulin-like Growth Factor I and Corticosterone with Nestling Growth and Fledging Success in Wild Passerines. Tartu, 2016, 113 p.
287. **Ann Kraut**. Conservation of Wood-Inhabiting Biodiversity – Semi-Natural Forests as an Opportunity. Tartu, 2016, 141 p.
288. **Tiit Örd**. Functions and regulation of the mammalian pseudokinase TRIB3. Tartu, 2016, 182. p.
289. **Kairi Käiro**. Biological Quality According to Macroinvertebrates in Streams of Estonia (Baltic Ecoregion of Europe): Effects of Human-induced Hydromorphological Changes. Tartu, 2016, 126 p.
290. **Leidi Laurimaa**. *Echinococcus multilocularis* and other zoonotic parasites in Estonian canids. Tartu, 2016, 144 p.
291. **Helerin Margus**. Characterization of cell-penetrating peptide/nucleic acid nanocomplexes and their cell-entry mechanisms. Tartu, 2016, 173 p.
292. **Kadri Runnel**. Fungal targets and tools for forest conservation. Tartu, 2016, 157 p.
293. **Urmo Võsa**. MicroRNAs in disease and health: aberrant regulation in lung cancer and association with genomic variation. Tartu, 2016, 163 p.
294. **Kristina Mäemets-Allas**. Studies on cell growth promoting AKT signaling pathway – a promising anti-cancer drug target. Tartu, 2016, 146 p.
295. **Janeli Viil**. Studies on cellular and molecular mechanisms that drive normal and regenerative processes in the liver and pathological processes in Dupuytren's contracture. Tartu, 2016, 175 p.
296. **Ene Kook**. Genetic diversity and evolution of *Pulmonaria angustifolia* L. and *Myosotis laxa sensu lato* (Boraginaceae). Tartu, 2016, 106 p.
297. **Kadri Peil**. RNA polymerase II-dependent transcription elongation in *Saccharomyces cerevisiae*. Tartu, 2016, 113 p.
298. **Katrin Ruisu**. The role of RIC8A in mouse development and its function in cell-matrix adhesion and actin cytoskeletal organisation. Tartu, 2016, 129 p.
299. **Janely Pae**. Translocation of cell-penetrating peptides across biological membranes and interactions with plasma membrane constituents. Tartu, 2016, 126 p.
300. **Argo Ronk**. Plant diversity patterns across Europe: observed and dark diversity. Tartu, 2016, 153 p.
301. **Kristiina Mark**. Diversification and species delimitation of lichenized fungi in selected groups of the family Parmeliaceae (Ascomycota). Tartu, 2016, 181 p.
302. **Jaak-Albert Metsoja**. Vegetation dynamics in floodplain meadows: influence of mowing and sediment application. Tartu, 2016, 140 p.

303. **Hedvig Tamman**. The GraTA toxin-antitoxin system of *Pseudomonas putida*: regulation and role in stress tolerance. Tartu, 2016, 154 p.
304. **Kadri Pärtel**. Application of ultrastructural and molecular data in the taxonomy of helotialean fungi. Tartu, 2016, 183 p.
305. **Maris Hindrikson**. Grey wolf (*Canis lupus*) populations in Estonia and Europe: genetic diversity, population structure and -processes, and hybridization between wolves and dogs. Tartu, 2016, 121 p.
306. **Polina Degtjarenko**. Impacts of alkaline dust pollution on biodiversity of plants and lichens: from communities to genetic diversity. Tartu, 2016, 126 p.
307. **Liina Pajusalu**. The effect of CO_2 enrichment on net photosynthesis of macrophytes in a brackish water environment. Tartu, 2016, 126 p.
308. **Stoyan Tankov**. Random walks in the stringent response. Tartu, 2016, 94 p.
309. **Liis Leitsalu**. Communicating genomic research results to population-based biobank participants. Tartu, 2016, 158 p.
310. **Richard Meitern**. Redox physiology of wild birds: validation and application of techniques for detecting oxidative stress. Tartu, 2016, 134 p.
311. **Kaie Lokk**. Comparative genome-wide DNA methylation studies of healthy human tissues and non-small cell lung cancer tissue. Tartu, 2016, 127 p.
312. **Mihhail Kurašin**. Processivity of cellulases and chitinases. Tartu, 2017, 132 p.
313. **Carmen Tali**. Scavenger receptors as a target for nucleic acid delivery with peptide vectors. Tartu, 2017, 155 p.
314. **Katarina Oganjan**. Distribution, feeding and habitat of benthic suspension feeders in a shallow coastal sea. Tartu, 2017, 132 p.
315. **Taavi Paal**. Immigration limitation of forest plants into wooded landscape corridors. Tartu, 2017, 145 p.
316. **Kadri Õunap**. The Williams-Beuren syndrome chromosome region protein WBSCR22 is a ribosome biogenesis factor. Tartu, 2017, 135 p.
317. **Riin Tamm**. In-depth analysis of factors affecting variability in thiopurine methyltransferase activity. Tartu, 2017, 170 p.
318. **Keiu Kask**. The role of RIC8A in the development and regulation of mouse nervous system. Tartu, 2017, 184 p.
319. **Tiia Möller**. Mapping and modelling of the spatial distribution of benthic macrovegetation in the NE Baltic Sea with a special focus on the eelgrass *Zostera marina* Linnaeus, 1753. Tartu, 2017, 162 p.
320. **Silva Kasela**. Genetic regulation of gene expression: detection of tissue- and cell type-specific effects. Tartu, 2017, 150 p.
321. **Karmen Süld**. Food habits, parasites and space use of the raccoon dog *Nyctereutes procyonoides*: the role of an alien species as a predator and vector of zoonotic diseases in Estonia. Tartu, 2017, p.
322. **Ragne Oja**. Consequences of supplementary feeding of wild boar – concern for ground-nesting birds and endoparasite infection. Tartu, 2017, 141 p.
323. **Riin Kont**. The acquisition of cellulose chain by a processive cellobiohydrolase. Tartu, 2017, 117 p.

324. **Liis Kasari**. Plant diversity of semi-natural grasslands: drivers, current status and conservation challenges. Tartu, 2017, 141 p.
325. **Sirgi Saar**. Belowground interactions: the roles of plant genetic relatedness, root exudation and soil legacies. Tartu, 2017, 113 p.
326. **Sten Anslan**. Molecular identification of Collembola and their fungal associates. Tartu, 2017, 125 p.
327. **Imre Taal**. Causes of variation in littoral fish communities of the Eastern Baltic Sea: from community structure to individual life histories. Tartu, 2017, 118 p.
328. **Jürgen Jalak**. Dissecting the Mechanism of Enzymatic Degradation of Cellulose Using Low Molecular Weight Model Substrates. Tartu, 2017, 137 p.
329. **Kairi Kiik**. Reproduction and behaviour of the endangered European mink (*Mustela lutreola*) in captivity. Tartu, 2018, 112 p.
330. **Ivan Kuprijanov**. Habitat use and trophic interactions of native and invasive predatory macroinvertebrates in the northern Baltic Sea. Tartu, 2018, 117 p.
331. **Hendrik Meister**. Evolutionary ecology of insect growth: from geographic patterns to biochemical trade-offs. Tartu, 2018, 147 p.
332. **Ilja Gaidutšik**. Irc3 is a mitochondrial branch migration enzyme in *Saccharomyces cerevisiae.* Tartu, 2018, 161 p.
333. **Lena Neuenkamp**. The dynamics of plant and arbuscular mycorrhizal fungal communities in grasslands under changing land use. Tartu, 2018, 241 p.
334. **Laura Kasak**. Genome structural variation modulating the placenta and pregnancy maintenance. Tartu, 2018, 181 p.
335. **Kersti Riibak**. Importance of dispersal limitation in determining dark diversity of plants across spatial scales. Tartu, 2018, 133 p.
336. **Liina Saar**. Dynamics of grassland plant diversity in changing landscapes. Tartu, 2018, 206 p.
337. **Hanna Ainelo**. Fis regulates *Pseudomonas putida* biofilm formation by controlling the expression of *lapA*. Tartu, 2018, 143 p.
338. **Natalia Pervjakova**. Genomic imprinting in complex traits. Tartu, 2018, 176 p.
339. **Andrio Lahesaare**. The role of global regulator Fis in regulating the expression of *lapF* and the hydrophobicity of soil bacterium *Pseudomonas putida*. Tartu, 2018, 124 p.
340. **Märt Roosaare**. *K*-mer based methods for the identification of bacteria and plasmids. Tartu, 2018, 117 p.
341. **Maria Abakumova**. The relationship between competitive behaviour and the frequency and identity of neighbours in temperate grassland plants. Tartu, 2018, 104 p.
342. **Margus Vilbas**. Biotic interactions affecting habitat use of myrmecophilous butterflies in Northern Europe. Tartu, 2018, 142 p.

343. **Liina Kinkar**. Global patterns of genetic diversity and phylogeography of *Echinococcus granulosus* sensu stricto – a tapeworm species of significant public health concern. Tartu, 2018, 147 p.
344. **Teivi Laurimäe**. Taxonomy and genetic diversity of zoonotic tapeworms in the species complex of *Echinococcus granulosus* sensu lato. Tartu, 2018, 143 p.
345. **Tatjana Jatsenko**. Role of translesion DNA polymerases in mutagenesis and DNA damage tolerance in Pseudomonads. Tartu, 2018, 216 p.
346. **Katrin Viigand**. Utilization of α-glucosidic sugars by *Ogataea* (*Hansenula*) *polymorpha*. Tartu, 2018, 148 p.
347. **Andres Ainelo**. Physiological effects of the *Pseudomonas putida* toxin grat. Tartu, 2018, 146 p.
348. **Killu Timm**. Effects of two genes (DRD4 and SERT) on great tit (*Parus major*) behaviour and reproductive traits. Tartu, 2018, 117 p.
349. **Petr Kohout**. Ecology of ericoid mycorrhizal fungi. Tartu, 2018, 184 p.
350. **Gristin Rohula-Okunev**. Effects of endogenous and environmental factors on night-time water flux in deciduous woody tree species. Tartu, 2018, 184 p.
351. **Jane Oja**. Temporal and spatial patterns of orchid mycorrhizal fungi in forest and grassland ecosystems. Tartu, 2018, 102 p.
352. **Janek Urvik**. Multidimensionality of aging in a long-lived seabird. Tartu, 2018, 135 p.
353. **Lisanna Schmidt**. Phenotypic and genetic differentiation in the hybridizing species pair *Carex flava* and *C. viridula* in geographically different regions. Tartu, 2018, 133 p.
354. **Monika Karmin**. Perspectives from human Y chromosome – phylogeny, population dynamics and founder events. Tartu, 2018, 168 p.
355. **Maris Alver**. Value of genomics for atherosclerotic cardiovascular disease risk prediction. Tartu, 2019, 148 p.
356. **Lehti Saag**. The prehistory of Estonia from a genetic perspective: new insights from ancient DNA. Tartu, 2019, 171 p.
357. **Mari-Liis Viljur**. Local and landscape effects on butterfly assemblages in managed forests. Tartu, 2019, 115 p.
358. **Ivan Kisly**. The pleiotropic functions of ribosomal proteins eL19 and eL24 in the budding yeast ribosome. Tartu, 2019, 170 p.
359. **Mikk Puustusmaa**. On the origin of papillomavirus proteins. Tartu, 2019, 152 p.
360. **Anneliis Peterson**. Benthic biodiversity in the north-eastern Baltic Sea: mapping methods, spatial patterns, and relations to environmental gradients. Tartu, 2019, 159 p.
361. **Erwan Pennarun**. Meandering along the mtDNA phylogeny; causerie and digression about what it can tell us about human migrations. Tartu, 2019, 162 p.

362. **Karin Ernits**. Levansucrase Lsc3 and endo-levanase BT1760: characterization and application for the synthesis of novel prebiotics. Tartu, 2019, 217 p.
363. **Sille Holm**. Comparative ecology of geometrid moths: in search of contrasts between a temperate and a tropical forest. Tartu, 2019, 135 p.
364. **Anne-Mai Ilumäe**. Genetic history of the Uralic-speaking peoples as seen through the paternal haplogroup N and autosomal variation of northern Eurasians. Tartu, 2019, 172 p.
365. **Anu Lepik**. Plant competitive behaviour: relationships with functional traits and soil processes. Tartu, 2019, 152 p.
366. **Kunter Tätte**. Towards an integrated view of escape decisions in birds under variable levels of predation risk. Tartu, 2020, 172 p.
367. **Kaarin Parts**. The impact of climate change on fine roots and root-associated microbial communities in birch and spruce forests. Tartu, 2020, 143 p.
368. **Viktorija Kukuškina**. Understanding the mechanisms of endometrial receptivity through integration of 'omics' data layers. Tartu, 2020, 169 p.
369. **Martti Vasar**. Developing a bioinformatics pipeline gDAT to analyse arbuscular mycorrhizal fungal communities using sequence data from different marker regions. Tartu, 2020, 193 p.
370. **Ott Kangur**. Nocturnal water relations and predawn water potential disequilibrium in temperate deciduous tree species. Tartu, 2020, 126 p.
371. **Helen Post**. Overview of the phylogeny and phylogeography of the Y-chromosomal haplogroup N in northern Eurasia and case studies of two linguistically exceptional populations of Europe – Hungarians and Kalmyks. Tartu, 2020, 143 p.
372. **Kristi Krebs**. Exploring the genetics of adverse events in pharmacotherapy using Biobanks and Electronic Health Records. Tartu, 2020, 151 p.
373. **Kärt Ukkivi**. Mutagenic effect of transcription and transcription-coupled repair factors in *Pseudomonas putida.* Tartu, 2020, 154 p.
374. **Elin Soomets**. Focal species in wetland restoration. Tartu, 2020, 137 p.
375. **Kadi Tilk**. Signals and responses of ColRS two-component system in *Pseudomonas putida.* Tartu, 2020, 133 p.
376. **Indrek Teino**. Studies on aryl hydrocarbon receptor in the mouse granulosa cell model. Tartu, 2020, 139 p.
377. **Maarja Vaikre**. The impact of forest drainage on macroinvertebrates and amphibians in small waterbodies and opportunities for cost-effective mitigation. Tartu, 2020, 132 p.
378. **Siim-Kaarel Sepp**. Soil eukaryotic community responses to land use and host identity. Tartu, 2020, 222 p.
379. **Eveli Otsing**. Tree species effects on fungal richness and community structure. Tartu, 2020, 152 p.
380. **Mari Pent**. Bacterial communities associated with fungal fruitbodies. Tartu, 2020, 144 p.

381. **Einar Kärgenberg**. Movement patterns of lithophilous migratory fish in free-flowing and fragmented rivers. Tartu, 2020, 167 p.
382. **Antti Matvere**. The studies on aryl hydrocarbon receptor in murine granulosa cells and human embryonic stem cells. Tartu, 2021, 163 p.
383. **Jhonny Capichoni Massante**. Phylogenetic structure of plant communities along environmental gradients: a macroecological and evolutionary approach. Tartu, 2021, 144 p.
384. **Ajai Kumar Pathak**. Delineating genetic ancestries of people of the Indus Valley, Parsis, Indian Jews and Tharu tribe. Tartu, 2021, 197 p.
385. **Tanel Vahter**. Arbuscular mycorrhizal fungal biodiversity for sustainable agroecosystems. Tartu, 2021, 191 p.
386. **Burak Yelmen**. Characterization of ancient Eurasian influences within modern human genomes. Tartu, 2021, 134 p.
387. **Linda Ongaro**. A genomic portrait of American populations. Tartu, 2021, 182 p.
388. **Kairi Raime**. The identification of plant DNA in metagenomic samples. Tartu, 2021, 108 p.
389. **Heli Einberg**. Non-linear and non-stationary relationships in the pelagic ecosystem of the Gulf of Riga (Baltic Sea). Tartu, 2021, 119 p.
390. **Mickaël Mathieu Pihain**. The evolutionary effect of phylogenetic neighbourhoods of trees on their resistance to herbivores and climatic stress. Tartu, 2022, 145 p.
391. **Annika Joy Meitern**. Impact of potassium ion content of xylem sap and of light conditions on the hydraulic properties of trees. Tartu, 2022, 132 p.
392. **Elise Joonas**. Evaluation of metal contaminant hazard on microalgae with environmentally relevant testing strategies. Tartu, 2022, 118 p.
393. **Kreete Lüll**. Investigating the relationships between human microbiome, host factors and female health. Tartu, 2022, 141 p.
394. **Triin Kaasiku**. A wader perspective to Boreal Baltic coastal grasslands: from habitat availability to breeding site selection and nest survival. Tartu, 2022, 141 p.
395. **Meeli Alber**. Impact of elevated atmospheric humidity on the structure of the water transport pathway in deciduous trees. Tartu, 2022, 170 p.
396. **Ludovica Molinaro**. Ancestry deconvolution of Estonian, European and Worldwide genomic layers: a human population genomics excavation. Tartu, 2022, 138 p.
397. **Tina Saupe**. The genetic history of the Mediterranean before the common era: a focus on the Italian Peninsula. Tartu, 2022, 165 p.
398. **Mari-Ann Lind**. Internal constraints on energy processing and their consequences: an integrative study of behaviour, ornaments and digestive health in greenfinches. Tartu, 2022, 137 p.
399. **Markus Valge**. Testing the predictions of life history theory on anthropometric data. Tartu, 2022, 171 p.
400. **Ants Tull**. Domesticated and wild mammals as reservoirs for zoonotic helminth parasites in Estonia. Tartu, 2022, 152 p.

401. **Saleh Rahimlouye Barabi**. Investigation of diazotrophic bacteria association with plants. Tartu, 2022, 137 p.
402. **Farzad Aslani**. Towards revealing the biogeography of belowground diversity. Tartu, 2022, 124 p.
403. **Nele Taba**. Diet, blood metabolites, and health. Tartu, 2022, 163 p.
404. **Katri Pärna**. Improving the personalized prediction of complex traits and diseases: application to type 2 diabetes. Tartu, 2022, 190 p.
405. **Silva Lilleorg**. Bacterial ribosome heterogeneity on the example of bL31 paralogs in *Escherichia coli.* Tartu, 2022, 189 p.
406. **Oliver Aasmets**. The importance of microbiome in human health. Tartu, 2022, 123 p.
407. **Henel Jürgens**. Exploring post-translational modifications of histones in RNA polymerase II-dependent transcription. Tartu, 2022, 147 p.
408. **Mari Tagel**. Finding novel factors affecting the mutation frequency: a case study of tRNA modification enzymes TruA and RluA. Tartu, 2022, 176 p.
409. **Marili Sell**. The impact of environmental change on ecophysiology of hemiboreal tree species – acclimation mechanisms in belowground. Tartu, 2022, 163 p.
410. **Kaarin Hein**. The hissing behaviour of Great Tit (*Parus major*) females reflects behavioural phenotype and breeding success in a wild population. Tartu, 2022, 96 p.
411. **Maret Gerz**. The distribution and role of mycorrhizal symbiosis in plant communities. Tartu, 2022, 206 p.
412. **Kristiina Nõomaa**. Role of invasive species in brackish benthic community structure and biomass changes. Tartu, 2023, 151 p.
413. **Anton Savchenko**. Taxonomic studies in Dacrymycetes: *Cerinomyces* and allied taxa. Tartu, 2023, 181 p.
414. **Ahto Agan**. Interactions between invasive pathogens and resident mycobiome in the foliage of trees. Tartu, 2023, 155 p.
415. **Diego Pires Ferraz Trindade**. Dark diversity dynamics linked to global change: taxonomic and functional perspective. Tartu, 2023, 134 p.
416. **Madli Jõks**. Biodiversity drivers in oceanic archipelagos and habitat fragments, explored by agent-based simulation models. Tartu, 2023, 116 p.
417. **Ciara Baines**. Adaptation to oncogenic pollution and natural cancer defences in the aquatic environment. Tartu, 2023, 164 p.
418. **Rain Inno**. Placental transcriptome and miRNome in normal and complicated pregnancies. Tartu, 2023, 145 p.
419. **Daniyal Gohar**. Diversity, genomics, and potential functions of fungus-inhabiting bacteria. Tartu, 2023, 138 p.
420. **Sirli Rosendahl**. Fitness effects of chromosomal toxin-antitoxin systems in *Pseudomonas putida.* Tartu, 2023, 154 p.
421. **Mathilde Frédérique E. André**. New Guinea, a hotspot for Human evolution: settlement history and adaptation in northern Sahul. Tartu, 2023, 202 p.

422. **Vlad-Julian Piljukov**. Biochemical characterization of Irc3 helicase. Tartu, 2023, 137 p.
423. **Gerli Albert**. Carbon use strategies of macrophyte communities in the northeastern Baltic Sea: implications for a high CO_2 environment. Tartu, 2023, 128 p.
424. **Mariann Koel**. The molecular interactions between trophoblast and endometrial cells in embryo implantation. Tartu, 2023, 171 p.
425. **Robin Gielen**. Diversity and ecological role of pathogenic fungi in insect populations. Tartu, 2023, 139 p.
426. **Kaspar Reier**. Quantity, stability and disparity of ribosomal components in *Escherichia coli* stationary phase. Tartu, 2023, 151 p.
427. **Linda Rusalepp**. The impact of environmental drivers and competition on phenolic metabolite profiles in hybrid aspen and silver birch. Tartu, 2023, 153 p.
428. **Eliisa Pass**. The effect of managed forest-wetland landscapes on forest grouse and nest predation. Tartu, 2023, 115 p.
429. **Sanni Färkkilä**. Methods for studying plant-fungal interactions – reflecting on the old, the new and the upcoming. Tartu, 2024, 147 p.
430. **Maarja Jõeloo**. Advances in microarray-based copy number variation discovery and phenotypic associations. Tartu, 2024, 209 p.
431. **Natàlia Pujol Gualdo**. Decoding genetic associations of female reproductive health traits. Tartu, 2024, 205 p.
432. **Sirelin Sillamaa**. The role of helicases Hmi1 and Irc3 in yeast mitochondrial DNA maintenance. Tartu, 2024, 189 p.
433. **Iris Reinula**. Genetic variation of grassland plants in changing landscapes. Tartu, 2024, 201 p.
434. **Vi Ngan Tran**. The cellular dynamics and epithelial morphogenesis in *Drosophila* wing development. Tartu, 2024, 158 p.
435. **Slendy Julieth Rodríguez Alarcón**. Intraspecific trait diversity in plants: characterizing effects of trait variation on community assembly and ecosystem functioning. Tartu, 2024, 129 p.
436. **Arun Kumar Devarajan**. Microbes and climate change: insights from plant-microbe interactions in rice phyllosphere and soil microbiomes in subarctic grasslands. Tartu, 2024, 224 p.
437. **Leonard Owuraku Opare**. Rearing density effects on a commercially important insect species. Tartu, 2024, 145 p.
438. **Siqiao Liu**. The effect of anthropogenic disturbance on soil fungal communities. Tartu, 2024, 172 p.
439. **Kertu Liis Krigul**. The gut microbiome at the interface of human health and disease. Tartu, 2024, 158 p.
440. **Danat Yermakovich**. The evolutionary history of complex traits: implications of archaic admixture. Tartu, 2024, 153 p.
441. **Yiming Meng**. Plant mycorrhizal type and status in the global flora. Tartu, 2024, 200 p.

442. **Iryna Yatsiuk**. Evolution, species delimitation and diversity in myxomycetes: *Arcyria* and allied genera. Tartu, 2024, 193 p.
443. **Daniela León Velandia**. Mycorrhizal trait distribution and composition in plant communities under natural gradients. Tartu, 2024, 121 p.
444. **Bruno Paganeli**. Dark diversity methods for prioritization of areas and species in nature conservation. Tartu, 2024, 155 p.
445. **Mario Reiman**. Placental transcriptome in normal and complicated pregnancies. Tartu, 2025, 167 p.
446. **Maarja Kõrkjas**. Dynamics of tree-related microhabitats in live forest trees and its links with biodiversity. Tartu, 2025, 134 p.
447. **Eleonora Beccari**. Mapping and exploring trait spaces across the tree of life. Tartu, 2025, 190 p.
448. **Jack R. Hall**. Dissolved organic carbon dynamics of Baltic Sea macroalgae: production, bioavailability and ecosystem effects. Tartu, 2025, 135 p.
449. **Artjom Stepanjuk**. Function of adhesion molecules and signalling pathways in human endometrial and embryonic models. Tartu, 2025, 247 p.
450. **Marianne Kivastik**. Heterostylous plants in an era of global change: the role of local, landscape and climatic actors. Tartu, 2025, 167 p.
451. **Yehor Yatsiuk**. Large tree-cavities as key structures for forest biodiversity. Tartu, 2025, 215 p.
452. **Ovidiu Copoț**. Relevance of eDNA, citizen science, and species distribution modelling for fungal conservation. Tartu, 2025, 198 p.