DISSERTATIONES
MATHEMATICAE
UNIVERSITATIS
TARTUENSIS
125

MART KALS

Computational and statistical methods for DNA sequencing data analysis and applications in the Estonian Biobank cohort

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
125

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
125

MART KALS

Computational and statistical methods for DNA sequencing data analysis and applications in the Estonian Biobank cohort

UNIVERSITY OF TARTU
Press

Institute of Mathematics and Statistics, Faculty of Science and Technology, University of Tartu, Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in mathematical statistics on November 21, 2018, by the Council of the Institute of Mathematics and Statistics, Faculty of Science and Technology, University of Tartu.

Supervisor: Krista Fischer, PhD
Professor, Institute of Mathematics and Statistics, Faculty of Science and Technology, University of Tartu, Tartu, Estonia
Senior Research Fellow, Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

Opponents: Academy Research Fellow Taru Tuulia Tukiainen, PhD
Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

Associate Professor Tanel Kaart, PhD
Chair of Animal Breeding and Biotechnology, Institute of Veterinary Medicine and Animal Sciences, Estonian University of Life Sciences, Tartu, Estonia

Commencement: J. Liivi 2-403, Tartu, on December 28, 2018, at 12:15 pm

The publication of this dissertation is granted by the Institute of Mathematics and Statistics, University of Tartu.

This research was funded by EU FP7 grant 313010, EU H2020 grants 692145, 676550, the European Regional Development Fund, the Centre of Excellence in Genomics (EXCEGEN) and GENTRANSMED (Project No. 2014-2020.4.01. 15-0012), the Development Fund of the University of Tartu (SP1GVARENG), the Estonian Doctoral School of Mathematics and Statistics (NMTMM09577) (NLTMS16154), TerVE programme grant PerMed I, NIASC grant No. 62721, the Estonian Research Council grants IUT20-60, IUT24-6, ETF9353, PUT1665P, and the Archimedes Foundation.

ISSN 1024-4212
ISBN 978-9949-77-895-9 (print)
ISBN 978-9949-77-896-6 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications, referred to in the text by Roman numerals (Ref. I to Ref. VI):

I Nikopensius, T., Annilo, T., Jagomägi, T., Gilissen, C., **Kals, M.**, Krjutškov, K., Mägi, R., Eelmets, M., Gerst-Talas, U., Remm, M., Saag, M., Hoischen, A., Metspalu, A. (2013). Non-syndromic tooth agenesis associated with a nonsense mutation in ectodysplasin-A (*EDA*). *Journal of Dental Research, 92*(6), 507–511.

II Nikopensius, T., Saag, M., Jagomägi, T., Annilo, T., **Kals, M.**, Kivistik, P.A., Milani, L., Metspalu, A. (2013). A missense mutation in *DUSP6* is associated with class III malocclusion. *Journal of Dental Research, 92*(10), 893–898.

III Vaher, U., Nõukas, M., Nikopensius, T., **Kals, M.**, Annilo, T., Nelis, M., Õunap, K., Reimand, T., Talvik, I., Ilves, P., Piirsoo, A., Seppet, E., Metspalu, A., Talvik, T. (2014). *De novo SCN8A* mutation identified by whole-exome sequencing in a boy with neonatal epileptic encephalopathy, multiple congenital anomalies, and movement disorders. *Journal of Child Neurology, 29*(12), NP202–206.

IV Guo, M.H., Nandakumar, S.K., Ulirsch, J.C., Zekavat, S.M., Buenrostro, J.D., Natarajan, P., Salem, R.M., Chiarle, R., Mitt, M., **Kals, M.**, Pärn, K., Fischer, K., Milani, L., Mägi, R., Palta, P., Gabriel, S.B., Metspalu, A., Lander, E.S., Kathiresan, S., Hirschhorn, J.N., Esko, T., Sankaran, V.G. (2017). Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(3), E327–E336.

V Mitt, M.*, **Kals, M.***, Pärn, K.*, Gabriel, S.B., Lander, E.S., Palotie, A., Ripatti, S., Morris, A.P., Metspalu, A., Esko, T., Mägi, R., Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics*, *25*(7), 869–876.

VI **Kals, M.**, Nikopensius, T., Läll, K., Sikka, T.T., Suvisaari, J., Salomaa, V., Ripatti, S., Palotie, A., Metspalu, A., Palta, P., Mägi, R. (2018). Advantages of genotype imputation with ethnically-matched reference panel for rare variant association analyses. *Manuscript*.

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

My contributions to the listed publications were as follows:

Refs. I–III Analysed and interpreted the sequencing data, participated in the writing of the manuscript.

Ref. IV Participated in sequencing data QC analysis, performed CNV calling and analysis, and revised the manuscript.

Ref. V Participated in study design, constructed Estonian-specific imputation reference panel, prepared the figures, contributed to the data analysis, and participated in writing of the first draft of the manuscript.

Ref. VI Participated in study design, constructed ethnically-matched imputation reference panel, prepared the figures, conducted the data analysis, and drafted the manuscript.

***** These authors contributed equally to this work.

LIST OF ABBREVIATIONS

1000G	1000 Genomes Project
bp	base pair
CNV	copy number variation
DNA	deoxyribonucleic acid
EGCUT	Estonian Genome Center, University of Tartu
EMR	electronic medical record
EstFin	Estonian-Finnish
GATK	Genome Analysis Toolkit
GWAS	genome-wide association study
HGMD	Human Gene Mutation Database
HMM	hidden Markov model
HRC	Haplotype Reference Consortium
indel	small insertion or deletion
IRP	imputation reference panel
kb	kilobase (1,000 base pairs)
LD	linkage disequilibrium
LoF	loss-of-function
M	million
MAF	minor allele frequency
mRNA	messenger RNA
NGS	next-generation sequencing
NHLBI-ESP	NHLBI Exome Sequencing Project
NR	non-reference
RNA	ribonucleic acid
SER	switch error rate
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SV	structural variation
tag SNP	tagging SNP
VQSR	variant quality score recalibration
WES	whole exome sequencing
WGS	whole genome sequencing

INTRODUCTION

Molecular genetics started to progress after the discovery of the double helix structure of DNA in 1953. After the development of Sanger sequencing in the 1970s to identify the sequence of bases in the DNA, the need for new technologies has grown tremendously. As a sign of a great effort, the first human genome was sequenced in 2000. Since then, the development of the second generation sequencing or so-called 'next-generation sequencing' (NGS) has rapidly evolved. Advances in NGS technology have led us to the situation where we can, in a cost-effective way, identify the full spectrum of genetic variation across the genome.

In connection with the development of ultra-high-throughput sequencing, there is a demand for better computational methods and resources. It all stems from the bioinformatics challenge that millions to billions of sequenced short DNA reads, 35–300 base pairs long, have to be put back together.

Many nations have launched a large-scale genomic analysis of their population, including Estonia. As a pioneer, Estonian Biobank has collected, analysed, and integrated genomic data for a long time. Almost 10 years ago, we started to generate and process NGS data and I have played a role in this process from the first day. There are many useful applications for the NGS data and I present some of them in my thesis. In the first part, I give an overview of exome sequencing and its use in medical genetics. In particular, I give an overview on how to detect rare and *de novo* mutations underlying Mendelian diseases. Whole genome sequencing is introduced in the next part, where I show its importance in covering the full spectrum of genome at a scale not previously attainable. Also, to perform large-scale association analysis more efficiently, I introduce an imputation reference panel specific to Estonians and show its advantages in imputing genetic variants, especially rare ones.

1. REVIEW OF THE LITERATURE

1.1. Introduction to genetics

Genetics is a branch of biology that studies the heritability and variation in organisms. DNA (deoxyribonucleic acid) is the hereditary material present in the nucleus of almost every cell of an organism. The information in DNA is stored as a code based on only four chemical bases (nucleotides): adenine (A), cytosine (C), guanine (G), and thymine (T). The full human DNA sequence contains more than 3 billion nucleotides present as base pairs (bps), and more than 99% of those bases are the same in all individuals. The sequence of these bases determines phenotypic characteristics (e.g. gender and blood type) and retains the information necessary for building and maintaining an organism. Human DNA is bundled into 46 chromosomes – 22 pairs of autosomes and one pair of allosomes (sex chromosomes). The autosome pairs are numbered (1–22), while the allosome pair usually consists of two X chromosomes in women or one X and one Y chromosome in men. Having two copies of each chromosome makes humans diploid organisms.

Only a small fraction of the human DNA (~1–2%) contains protein-coding regions. These regions are called genes, where coding parts (exons) are interrupted by noncoding parts (introns). It is estimated that there are ~20,000 human protein-coding genes (Ezkurdia et al. 2014). During transcription, the entire gene is copied into a pre-mRNA (messenger ribonucleic acid), and during the process of RNA splicing, introns are removed and the resulting mRNA is translated into amino acids to form proteins. Except for the genes, the exact functional role of the large majority of the DNA remains unclear (Rands et al. 2014) (however, it has been found that great deal of genetic variation is hidden in the form that do not produce obvious phenotypic differences).

Each gene resides at a specific locus (location on a chromosome) in two copies, one copy of the gene inherited from each parent. At each genomic position, one copy of the gene is named as allele and two copies together are referred as genotype. A given gene may have multiple different alleles, though only two alleles are present at the locus of any individual. The alleles may differ from each other (heterozygous genotype) or be the same (homozygous genotype). The relative frequency of an allele at a locus in a particular population is the allele frequency. Allele frequencies may vary between human populations (e.g. between geographical or ethnic groups).

There can be differences in both the composition and the structure of the DNA between individuals. Single nucleotide variants (SNVs) occur when a single base (A, C, G, or T) is altered in the DNA sequence and are the most common type of genetic variation. A subclass of SNVs are called as single-nucleotide polymorphisms (SNPs), corresponding to the SNVs where at least a certain proportion (e.g. > 1%) of individuals carry a different nucleotide than the majority of the population at a specific locus. SNVs may fall within coding sequences of genes, non-coding regions of genes, or in the regions between

genes (intergenic regions). SNVs within the coding sequence may change the amino acid sequence of the produced protein, but do not do it necessarily, due to the degeneracy of the genetic code. SNVs in the coding region that do not affect the protein sequence are called synonymous, in contrast to nonsynonymous variants that change the amino acid sequence of a protein. SNVs that are not in protein-coding regions may still affect gene splicing, transcription factor binding, mRNA degradation, or the sequence of noncoding RNA.

Small insertions or deletions (indels) occur when a stretch of DNA ranging from a single nucleotide to 1 kilobase (kb) in the genome is either present or absent. Small indels are the second most frequent type of genetic variation in human genomes. In terms of base pair variation, indels cause similar level of variation as SNVs. Structural variations (SVs) are generally defined as a regions of DNA approximately 1 kb and larger in size (Freeman et al. 2006) and can include inversions, balanced translocations or genomic imbalances (indels), commonly referred to as copy number variations (CNVs), because they effectively change the DNA copy number (the number of copies of a particular gene in the genotype of an individual). Despite the fact that small indels are highly abundant, they have received far less attention compared to SNVs and SVs, because they are more challenging to detect and validate (Mullaney et al. 2010).

Meiosis is a specialized type of cell division that reduces the chromosome number by half. During meiosis, one maternal and one paternal chromosome form a pair of homologous chromosomes and exchange parts of their DNA (genetic recombination), resulting in the recombination of the two original chromosomes. Due to the genetic recombination, the offspring have a different set of alleles and genes compared to their parents. A particular set of alleles at linked loci that are present on one of the two homologous chromosomes are called haplotype blocks. The non-random association between loci within a haplotype block is called linkage disequilibrium (LD).

For a set of N heterozygous SNVs, there are 2^N possible haplotypes that could underlie the genotypes (Figure 1). However, in the presence of LD within this set, the number of actual haplotypes that are present in the population can be considerably smaller. Knowledge of the LD structure within the population makes it possible to identify a set of tagging SNPs (tag SNPs) – once the genotypes of these SNPs are known, the entire haplotype block is uniquely determined (or at least known with a sufficiently high probability).

Figure 1. SNPs, haplotypes and tag SNPs. SNPs and haplotypes from the same genomic regions in four individuals are illustrated. (A) Three SNPs with two possible alleles are highlighted in different colours, rest of the loci are identical (black). (B) Haplotypes are combinations of alleles that are inherited together and are located at nearby SNPs on the same chromosome. In total, 23 SNPs are presented, and three are from the panel (A). (C) Three tag SNPs out of the 23 SNPs are sufficient to identify these four haplotypes uniquely.

1.2. Next-generation sequencing

Since the discovery of the double helix structure of DNA (Watson & Crick 1953) and the development of Sanger sequencing (Sanger & Coulson 1975) to detect the sequence of DNA bases, the field of DNA sequencing has rapidly progressed. Sanger sequencing method provides ultimate resolution for genome analysis, but is limited by the low number (96) of parallel reactions, which makes it time consuming and expensive. Despite that, it was the most widely used sequencing method for approximately 40 years and remains in wide use for smaller-scale projects.

Over the last two decades, technologies across multiple fields were brought together in the development of so-called ‘next-generation’ or ‘massively parallel shotgun’ sequencing instrument for large-scale, routine sequencing. This phenomenon was especially true at the beginning of the current century, largely because of the efforts of Human Genome Project to sequence the human genome (Lander et al. 2001; Venter et al. 2001; Human Genome Sequencing Consortium 2004). Since then, the number of companies involved in NGS and developed technologies has increased rapidly, along with the corresponding field of bioinformatics. Today, NGS is a commonly used term describing ultra-high-throughput sequencing methods that allow the sequencing of millions to billions of DNA fragments in a single instrument run. The cost of DNA sequencing has fallen from a billion dollars (the first human genome) to a few thousands of dollars (nowadays) per human genome. In addition to rapid turnaround and reasonable price compared to the previous methods, NGS allows

studying all kinds of genomic variation (e.g. SNVs, indels, SVs) in a single experiment.

Although sequencing instruments differ in many aspects, they rely on a few conceptually similar core technologies (Metzker 2010). Several distinct NGS platforms are commercially available, the most well-known manufacturers currently are Illumina (Bentley et al. 2008), Pacific Biosciences (Eid et al. 2009), Ion Torrent (Rothberg et al. 2011), and Oxford Nanopore (Quick et al. 2014). Detailed overviews of existing sequencing platforms and their properties are given by Reuter et al. (2015), Levy and Myers (2016).

Currently, Illumina's platforms dominate the sequencer market share. The HiSeq 2500 System features two run modes: a high-throughput mode, which outputs up to one terabase (Tb) of data (up to four billion reads) in six days, and a quick but less cost-effective rapid mode, which produces a 30× coverage (the number of sequenced bases at given position) human genome in 27 hours. The HiSeq X Ten System is specialized for WGS. It consists of a set of 10 HiSeq X instruments, able to deliver 18,000 human genomes at 30× coverage per year.

The performance of different NGS instruments can be evaluated by a variety of metrics like throughput, read length, cost per base, and error rate. Most mainstream NGS systems have short read lengths (35–300 bp), which have a limited use for *de novo* assembly (Treangen & Salzberg 2012), and for the detection of structural variations in high resolution. NGS technologies also have a noticeably higher error rates in base calls than Sanger sequencing, which affects the reliability of detecting genomic variation. Still, it is estimated that long-read technologies (e.g. Pacific Biosciences and Oxford Nanopore), producing thousands of bases per read, have even higher sequencing error rates (Besser et al. 2018; Fox et al. 2014).

Several large-scale sequencing projects have been finished, but many more are ongoing or in the planning stages. The 1000 Genomes Project (1000G) (Gibbs et al. 2015) was the first initiative to sequence the genomes of a large number of individuals. It was conducted in 2008–2015 and contained data for 2,504 individuals from 26 populations. Other examples of large-scale publicly funded projects are the Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Network 2012) that has generated maps of the key genomic changes in many types of cancer, and the related Encyclopedia of DNA Elements (ENCODE) project (The ENCODE Project Consortium 2012) to build a comprehensive list of functional elements in the human genome. In addition, several populations have announced their own sequencing efforts, such as the ongoing Genomics England's '100,000 Genomes Project' (Marx 2015), aiming to sequence 100,000 cancer or rare disease patients. Also, there are an increasing number of ongoing consortium-based projects. For example, the data consortium Genome Aggregation Database (gnomAD) (Lek et al. 2016) has the goal of aggregating and harmonizing sequencing data from a wide variety of projects. In general, most of the data from large-scale NGS projects has been made available for the larger scientific community and it is now implausible to

conduct any type of genomic analysis without using publicly available data (e.g. reference genome, variant frequencies).

On many occasions, it is necessary to select genomic regions of interest before sequencing. For instance, one can sequence all protein-coding regions of the genome. Compared to WGS, it is more cost-effective, focusing only on the ~1–2% of the genome that encodes proteins. These approaches are referred as 'whole exome sequencing' (WES) and based on DNA enrichment to target only the regions of interest (Mamanova et al. 2010). Several technologies for targeted sequence capture have been developed, with the most well-known platforms from Illumina, Agilent Technologies, and Roche NimbleGen.

Turner and colleagues (2009) point out that in the analysis of WES data, at least eight relevant performance metrics should be considered: (1) capture specificity (the ability to capture only the target regions), (2) uniformity (the equal capture over the targets), (3) completeness (the ability to capture all targets), (4) allelic bias (the equal capture of two alleles of a heterozygous variant), (5) multiplexity (the ability to capture all target regions in a single experiment), (6) input requirements (the amount and quality of input DNA), (7) scalability (the flexibility to handle large sample sizes), and (8) cost (if two methods perform equally or sample sizes are large).

2. AIMS OF THE STUDY

The aim of this thesis was to introduce computational and statistical methods for the analysis of next-generation DNA sequencing data, and its possible applications in the Estonian Biobank cohort.

The specific objectives of the thesis were as follows:

1. To use exome sequencing analysis in gene discovery for Mendelian diseases (Refs. I–III).
2. To illustrate how population-based whole genome sequencing provides insight into hematopoietic regulatory mechanisms (Ref. IV).
3. To quantify how much a population-specific imputation reference panel improves imputation accuracy of low-frequency and rare variants as compared to the reference panels based on diverse populations (Ref. V).
4. To quantify the advantages of genotype imputation with an ethically-matched reference panel for rare variant association analysis (Ref. VI).

3. BIOINFORMATICS AND STATISTICAL METODOLOGY

3.1. Bioinformatics processing

To turn sequenced DNA fragments into biologically meaningful information, several bioinformatics steps have to be taken. The bioinformatics processing of NGS data is quite similar across the different platforms. Any data processing from raw sequencing reads to 'analysis-ready' genomic variations includes the following steps: (1) pre-processing, (2) genotype calling, and (3) post-processing (Figure 2). However, each step introduces bias which has to be measured and taken into account.

3.1.1. Base-calling

Most of the NGS technologies rely on the detection of illumination signals from billions of clusters of DNA templates. In other words, base-calling algorithms infer the actual nucleotide information from a multitude of high-resolution images. The signal intensities are also used for the calculation of per-base quality scores. Although base-calling errors may be specific for sequencing platforms, they can all be transformed into the standard Phred quality score (Ewing et al. 1998), given the following formula:

$$Q_{Phred} = -10 \log_{10} P\,(error). \quad (1)$$

For example, an error rate of 1% corresponds to a Phred score of 20. Distributions of Phred quality scores vary between platforms. For instance, Illumina platforms are more error-prone in later cycles, and there are noticeable differences between error rates for SNVs and indels across platforms. The ability to reduce the error rate of base calls has important consequences for the downstream analysis.

3.1.2. Pre-processing

The method of assembly relying on a reference sequence is called mapping, and the method not using reference is referred as *de novo* assembly. Assembly is a complex computational challenge, in which billions of reads have to be placed at their correct genomic origin. As all platforms generate reads of much shorter lengths than the length of human genome, the target genome is over-sampled with short reads from random positions. Because shorter reads have a higher probability of being mapped to several identical locations, it increases the computational complexity and uncertainty, and makes them not suitable for *de novo* assembly. In addition, alignment algorithms have to take into account sequencing errors and deviations from the reference sequence due to SNVs and indels. Therefore, alignment is complicated for regions with large differences between the reference and the sequenced genome. *De novo* assembly techni-

ques, which mostly rely on graph-based representation (Sundquist et al. 2007; Zerbino & Birney 2008), may provide potential solutions. Most mapping algorithms for short reads (like BWA (Li & Durbin 2009) and Bowtie (Langmead et al. 2009)) make use of Burrows-Wheeler transform (Burrows et al. 1994) or are hash-based.

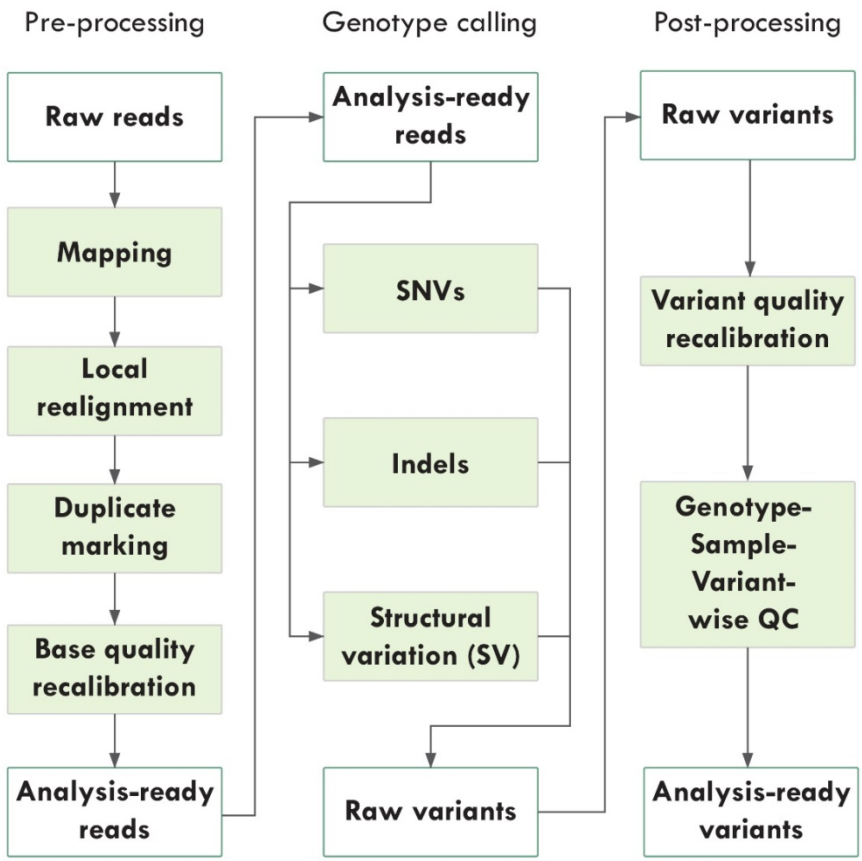

Figure 2. The GATK best practice framework from raw next-generation sequencing data to high-quality genotypes. Pre-processing converts raw-reads into a set of aligned reads with associated quality scores (per-base quality scores for each base and mapping quality scores for each read). Genotype calling is applied using a Bayesian approach, resulting in genotype calls and associated quality scores. Typically, multiple samples are called simultaneously. Post-processing separates true variation from artefacts.

A best practice guideline of the Genome Analysis Toolkit (GATK) (DePristo et al. 2011; Van der Auwera et al. 2013) is nearly ubiquitously used and accepted as a gold standard pipeline for the NGS data processing. This guidelines states that initial alignments in sequence alignment map (SAM) file have to be sorted

and converted to a binary alignment map (BAM) file to make the analysis faster. Duplicate reads have to masked, because they are sequenced from the same DNA molecule and should not counted in genotype calling. To eliminate mapping artefacts, initial alignments are refined by a local realignment to identify the most consistent placement of the reads in respect to indels. As genotype calling algorithms depend on the base quality scores provided by the sequencing machines, these scores need to be recalibrated, because initial estimates may be inaccurate.

3.1.3. Genotype calling

Early genotype detection methods counted the number of times each allele is observed and applied fixed thresholds, but this leads to a loss of information regarding individual read qualities (Nielsen et al. 2011). Most current algorithms utilize a Bayes' theorem for calculating conditional likelihoods of genotypes given the read data for a particular individual at a particular site.

There are several approaches for assigning priors: (1) it can be equal for all genotypes, (2) rely on external information (e.g. public databases), or (3) it can be improved by jointly analysing multiple individuals. The genotype with the highest posterior probability is generally chosen. The genotype likelihood can be also calculated using the per-base quality scores. One can take into account the pattern of LD at nearby sites, e.g. GATK HaplotypeCaller performs local *de novo* assembly of haplotypes in the region of interest (Van der Auwera et al. 2013).

Nowadays, the calling algorithms detect SNVs and small indels simultaneously, but the detection of indels is still more problematic due to sequencing errors and alignment artefacts.

3.1.4. Post-processing

After calling, each genotype, sample and variant must be evaluated and filtered based on quality estimates. The GATK best practice guideline suggests a two-step variant quality score recalibration (VQSR). In the first step, variant quality score is calculated using machine learning methods to assign a well-calibrated probability to each variant call in a raw call set. In the second step, calculated score can be used for separating true positive and false positive calls. The end product of the quality control is a variant call format (VCF) file containing high-quality variant calls that can be used in downstream analyses (Van der Auwera et al. 2013).

Finally, variants are usually annotated to assess their molecular and clinical significance. A variety of variant annotators are available, the most widely used are Variant Effect Predictor (McLaren et al. 2016), Annovar (Wang et al. 2010), and snpEFF (Cingolani et al. 2012). Although annotators aggregate plenty of

data from different sources, depending on the study design and research question, information from additional databases may be required. In general, variant-level annotation may include the following information:

1) variant description according to the internationally accepted standard (e.g. HGVS nomenclature (Den Dunnen et al. 2016));
2) allele frequencies in population databases (e.g. NHLBI Exome Sequencing Project (NHLBI-ESP) (Auer et al. 2016), 1000G (Gibbs et al. 2015), and ExAC/gnomAD (Lek et al. 2016)) and, if available, in-house databases;
3) *in silico* pathogenicity predictions (e.g. PolyPhen-2 (Adzhubei et al. 2010), SIFT (Kumar et al. 2009), and CADD (Kircher et al. 2014));
4) evolutionary conservation scores (e.g. PhyloP (Pollard et al. 2010));
5) additional annotations (e.g. pathogenic variants databases like Human Gene Mutation Database (HGMD) (Stenson et al. 2009) or known gene-disease databases like Online Mendelian Inheritance in Man (OMIM) (McKusick 2007)) can be added.

3.2. Identification of disease genes by exome sequencing

WES and WGS allow to explore rare variants explaining the heritability of complex traits as well to identify genes underlying rare disorders (known as Mendelian diseases). The number of rare Mendelian diseases is estimated to be ~7000, and for more than two-third of these underlying genes have been discovered (McKusick 2007). For Mendelian diseases, the rapid growth of the number of identified causal variants started after the wide deployment of WES (Kaiser 2010). Identification of variants that underlie both complex and Mendelian traits provides important knowledge about disease mechanisms and biological pathways that should lead to improved diagnostics, prevention strategies and potential therapeutic targets (Dietz 2010). WES is favourable for disease gene discovery, because most variants that are known to underlie rare Mendelian diseases are protein-altering (Stenson et al. 2009), and majority of rare, protein-coding variants are predicted to be deleterious (Kryukov et al. 2007). Therefore, the exome represents a highly enriched subset of the genome in which to search for variants with large effect sizes (Ng et al. 2010).

In general, WES strategies depend on the mode of inheritance of a disease, the pedigree or population structure, whether a phenotype arises due to *de novo* or inherited variants, and the extent of locus heterogeneity for a disease (Bamshad et al. 2011). As tens of thousands of genomic variants can be identified by WES in each individual, a key challenge is how to efficiently distinguish disease-related variants from non-pathogenic variants and sequencing artefacts. The most successful strategy of the identification of a novel disease gene relies on discrete filtering. First, to prioritize rare or novel candidate variants, sequenced data is filtered against population frequency datasets (e.g. dbSNP (Sherry et al. 2001), NHLBI-ESP, 1000G, in-house databases) to exclude variants that are present at frequencies higher than the expected carrier

frequency. It is important to note that described prioritization may discard the pathogenic variant, because the underlying assumption that the filtering datasets do not contain individuals with the studied disease may be not hold. Or, the disease causing variant is present in the population at low frequency in a heterozygous state. This risk is especially relevant for recessive disorders, in which variant causes disease if present in a homozygous or compound heterozygous state. Next, candidate variants can be prioritized based on their functional class (e.g. nonsynonymous, loss-of-function (LoF) variants), by evolutionary conservation scores (e.g. PhyloP) and *in silico* pathogenicity predictions (e.g. PolyPhen-2, SIFT, CADD), because pathogenic variants tend to have a markedly higher conservation than benign variants (Cooper et al. 2010).

Another important factor that can assist unraveling of candidate variants is the use of pedigree information. Not necessarily all individuals in a large pedigree have to be sequenced, but the choice depends on the relationships in the pedigree and the frequency of a disease causing variant. Sequencing the two or three most distantly related individuals with the phenotype of interest can restrict the genomic search space most effectively. For the discovery of *de novo* mutations, the most favourable is parent-child trio analysis aiming to filtering out all inherited variants. A detailed overview of common WES strategies is given by Gilissen et al. (2012). Finally, all detected findings have to be validated by Sanger sequencing.

Although, WES has detected thousands of clinically relevant candidate variants, there exist constant need for improved sequencing methodology, statistical and bioinformatics methods for better detection, prioritization and interpretation. There are several reasons influencing the success of WES such as sequencing not covering target regions entirely, bioinformatics artefacts, the disease-causing variant is in the non-coding region, multiple candidate variants of unknown significance left after filtering or possible misinterpretation of clinical significance of identified variants.

3.3. Genotype imputation

Genotype imputation is a method for statistically inferring untyped genotypes in a set of partially genotyped individuals, lending information from a densely genotyped reference panel of phased haplotypes. Haplotype phasing refers to the statistical estimation of haplotypes from the genotype data. Imputation methods attempt to identify haplotype sharing between individuals in the target set and in an imputation reference panel (IRP) (Marchini et al. 2007; Li et al. 2009). Intuitively, any two individuals can share short stretches of chromosomal segments from a distant common ancestor. The true haplotypes underlying the observed genotype data are assumed to be imperfect mosaics of the reference haplotypes. Points where the reference haplotype changes from one to another represent the historical recombination. The observed alleles may differ from the

alleles on the underlying reference haplotypes because of mutations, genotype errors, or erroneously assigned matches.

The main advantage of genotype imputation is that it allows to study variants that have not been directly genotyped and thereby to increase the resolution of genome-wide association studies (GWAS) (Figure 3). Also, imputation is useful for combining association results across studies that used different genotyping arrays and facilitates fine-mapping to localise association signals by increasing genetic variant density in candidate genomic regions (Liu et al. 2010).

3.3.1. Genotype imputation methods

Fundamentally, imputation is very similar to phasing, so most imputation algorithms are based on population genetic models that were originally used in phasing methods. The most important distinction between phasing and imputation datasets is that the latter include large proportions of systematically missing genotypes (Howie et al. 2009). Most methods for haplotype phase inference can also be used to perform imputation, but there are imputation methods that are independent of haplotype phase inference (Browning 2008).

As the number of haplotypes increases, it becomes increasingly difficult to efficiently apply the classical recursive computation algorithms. If the number of individuals being phased is N, then the complexity of the algorithm is quadratic. The ability to limit the number of states is essential for datasets with larger numbers of individuals (e.g. GWAS-sized datasets). Different imputation methods have been developed, summarized in Table 1 by Das et al. (2018), but the majority of them based on the statistical model for patterns of LD among multiple markers introduced by Li and Stephens (2003). In this framework, a subset of haplotypes is selected as a reference set, and each reference haplotype represents a hidden state of the hidden Markov models (HMMs) at each marker. For instance, this framework is implemented in MaCH (Li et al. 2010), IMPUTE (Marchini et al. 2007; Howie et al. 2009), minimac (Howie et al. 2012; Fuchsberger et al. 2015; Das et al. 2016), and the most recent Beagle (Browning & Browning 2016) algorithms. Although all of them employ the HMM, they differ from each other in how they define the state space and the parameters of the HMM.

The HMM framework is the most widely used method for inference of haplotype phase and missing genotypes. In an HMM, there are a set of observations that can be used to generate underlying hidden states (Rabiner 1989). In case of missing genotype inference, the observed unphased genotypes represent the observed data of the HMM, whereas an underlying and unobserved set of phased genotypes represent the hidden states. A Markov model is applied to the hidden states along the chromosome.

Figure 3. An overview of genotype imputation. Genotype imputation uses a densely genotyped reference panel of phased haplotypes to infer untyped genotypes in partially genotyped individuals. (A) An individual is partially genotyped, with a large number of missing genotypes (question marks). (B) Imputation methods require that haplotypes are estimated from typed genotype data. (C) These haplotypes are compared to the densely genotyped reference panel of haplotypes. Haplotypes of unrelated individuals over short stretches of DNA may be related to each other by being identical by descent, therefore their haplotypes can be modeled as a mosaic of haplotypes of other individuals. Untyped alleles are inferred based on these matched haplotypes.

The Li and Stephens model state space is represented as a two-dimensional grid of HMM states (Figure 3C), where rows represent reference haplotypes and columns reference panel markers. Each allele (on each reference haplotype) represents an HMM state. Each observed haplotype (Figure 3B) proceeds from left to right through the all reference markers (from the first to the last). When the path switches between reference haplotypes (rows), a new segment in the mosaic of reference haplotypes starts. This is determined by the HMM transition probabilities and closely related to the population recombination rate. The HMM emission probabilities determine the difference between the observed allele and the reference allele. Given an observed haplotype with missing allele, the probabi-

lity of each possible path through the HMM states can be calculated with the HMM forward-backward algorithm (Rabiner 1989). Imputed allele probabilities at a marker are obtained from the state probabilities. The probability that the target haplotype carries a particular allele (probability of imputed allele) is the sum of all the state probabilities corresponding to reference haplotypes that carry the allele.

Specialized algorithms are used to compute HMMs. For example, the Viterbi algorithm (Viterbi 1967) to find the most likely sequence of hidden states, and the Baum forward–backward algorithm (Baum & Eagon 1967) to compute posterior probabilities of hidden states. Then, one can use the Baum–Welch algorithm (or equivalently the EM (expectation–maximization) algorithm (Dempster et al. 1977)) to fit model parameters by maximizing the likelihood. Alternatively, Bayesian models typically use Markov chain Monte Carlo (MCMC) sampling, attempting to explore the entire model space. Details of the HMM algorithms are given by Rabiner (1989).

As an example, let us consider the algorithm implemented in IMPUTE (Marchini et al. 2007) to impute untyped genotypes. Assume that we have data at L diallelic autosomal variants with two alleles coded as 0 and 1. Let denote $H = \{H_1, \dots, H_N\}$ a set of N known haplotypes at L markers, where $H_i = \{H_{i1}, \dots, H_{iL}\}$ and $H_{ij} \in \{0,1\}$. Let $G = \{G_1, \dots, G_K\}$ denote the genotype data on the K individuals with $G_i = \{G_{i1}, \dots, G_{iL}\}$ and $G_{ik} \in \{0,1,2, missing\}$. To impute the missing genotypes, partition G into two disjoint sets: a set T that is typed in both the target individuals and the reference panel, and a set U that is untyped in the target sample but typed in the reference panel, $G = \{G_T, G_U\}$. The joint distribution of typed and untyped genotype data is assumed, and that each individual's genotype vector can be considered independently of the others. Then

$$\mathrm{P}(G_U|G_T, H) \propto \mathrm{P}(G_U, G_T|H) = \mathrm{P}(G|H) = \prod_{i=1}^{K} \mathrm{P}(G_i|H). \qquad (2)$$

Missing genotypes are inferred through each individual's genotype vector G_i, conditional on H and a set of parameters. Corresponding HMM can be written as

$$\mathrm{P}(G_i|H, \theta, \rho) = \sum_{Z_i^{(1)}, Z_i^{(2)}} \mathrm{P}\left(G_i \middle| Z_i^{(1)}, Z_i^{(2)}, \theta\right) \mathrm{P}\left(Z_i^{(1)}, Z_i^{(2)} \middle| H, \rho\right), \qquad (3)$$

where $Z_i^{(1)} = \left\{Z_{i1}^{(1)}, \dots, Z_{iL}^{(1)}\right\}$ and $Z_i^{(2)} = \left\{Z_{i1}^{(2)}, \dots, Z_{iL}^{(2)}\right\}$ are two sequences of hidden states at the L sites with $Z_{il}^{(j)} \in \{1, \dots, N\}$. These hidden states can be thought of as the pair of haplotypes in the set H that are being copied to form the genotype vector G_i. The term $\mathrm{P}\left(Z_i^{(1)}, Z_i^{(2)} \middle| H, \rho\right)$ models how the pair of

copied haplotypes changes along the sequence and is defined by a Markov chain in which switching between states depends on an estimate of the fine-scale recombination map (ρ) across the genome. The initial state of the Markov chain is assumed to follow the Uniform distribution:

$$\mathrm{P}\left(Z_{i1}^{(1)}, Z_{i1}^{(2)} \middle| H, \rho\right) = \frac{1}{N^2}. \quad (4)$$

The term $\mathrm{P}\left(G_i \middle| Z_i^{(1)}, Z_i^{(2)}, \theta\right)$ allows each observed genotype vector to differ through mutation from the genotypes determined by the pair of copied haplotypes and is controlled with the mutation parameter θ. The model allows for recurrent mutation at each site, but assumes a uniform mutation rate across the genome, $\theta = \left(\sum_{i=1}^{N-1} \frac{1}{i}\right)^{-1}$. Exact marginal probabilities for the missing genotypes that are conditional on the observed genotype data in the vector G_i are obtained using the forward-backward algorithm for HMMs. The transition and mutation probabilities with precise forms of the HMM terms are given by Marchini et al. (2007).

One can separate imputation into two steps: first, estimate the haplotypes for each individual (pre-phasing) and then impute missing genotypes into these estimated haplotypes (Howie et al. 2012). This approach reduces the complexity of the imputation step from quadratic to linear in the number of reference haplotypes, because it is much faster to match a phased haplotype to one reference haplotype than to match two unphased genotypes to a pair of reference haplotypes.

3.3.2. Imputation reference panels

As discussed by Das et al. (2018), several factors can affect imputation accuracy such as haplotype phasing in reference and study samples, density of genotyping array, size of reference panel, similarities of LD patterns and allele frequencies between study samples and the reference panel.

Publicly available imputation reference panels (IRPs) like the International HapMap Project (Frazer et al. 2007; International HapMap 3 Consortium 2010), 1000G (Gibbs et al. 2015) and Haplotype Reference Consortium (HRC) (McCarthy et al. 2016) have been commonly used for imputation. The first large-scale imputation studies mostly used the HapMap (Phase II) IRP, which consists of microarray-based genotypes from 270 individuals at 3.1 million (M) variants (Morris et al. 2012; Speliotes et al. 2010). The 1000G project was the first large-scale IRP based on WGS, containing eventually 2,504 individuals from 26 populations across the world and up to 81.7 M variants (Artigas et al. 2015; Leeuwen et al. 2016; Gormley et al. 2016). HRC contains 32,488 reference individuals, mostly with European ancestry and up to 39.2 M variants.

Although above-mentioned ethnically heterogeneous IRPs allow robust imputation of common variants (minor allele frequency (MAF) ≥ 5%) and low-frequency variants (0.5 ≤ MAF < 5%), they have only limited imputation accuracy for rare (MAF < 0.5%) variants (Pasaniuc et al. 2012; Zheng et al. 2012). To date, the largest publicly accessible IRP is the Trans-Omics for Precision Medicine (TOPMed) WGS program, containing 62,784 individuals from diverse populations and about 463 M variants.

During the last few years, the results indicate that an IRP specific to the particular population (referred to as population-specific IRP) can improve the imputation of rarer variants due to more similar allele frequencies and greater relatedness between the imputed individuals and the IRP. Population-specific IRPs further advance the imputation accuracy of common and low-frequency variants in the relevant population (Pistis et al. 2015; Gudbjartsson et al. 2015; Deelen et al. 2014). They achieve a higher imputation accuracy compared to the 1000G panel even in the case of smaller panel sizes (Zhou et al. 2017; Lin et al. 2018). For example, using imputed data of Sardinian WGS-based IRP, Sidore et al. (2015) detected several variants associated with circulating lipid levels in Sardinians. In the UK10K project, where the British population-specific IRP was combined with 1000G reference panel, several novel genetic variants associated with medically relevant phenotypes were discovered (Walter et al. 2015; Huang et al. 2015). In addition, several studies have demonstrated the benefit of population-specific IRPs to discover rare variants associated with diseases (Holm et al. 2011; Jonsson et al. 2013; Helgason et al. 2013; Steinthorsdottir et al. 2014).

3.3.3. Phasing and imputation accuracy measures

A standard measure to assess phasing accuracy is the switch error rate (SER) (Stephens & Donnelly 2003). A switch error occurs when a heterozygous site has phase switched in respect to the previous heterozygous site. The SER is the proportion of pairs of heterozygous sites where a switch error has occurred out of the total number of possible pairs and can only be assessed when the true haplotypes are known (e.g. in simulated data, or when nuclear family data is available). SER is zero if all the heterozygotes are phased correctly.

Several measures have been proposed to assess the accuracy of the imputed dose of an allele without the knowledge of the true allele dose:

1) The squared correlation r^2 between the imputed and true dose of an allele (Das et al. 2018; Howie et al. 2012). Let $X = 1$ if a chromosome carries the allele of interest and let $X = 0$ otherwise. Let Z be the estimated posterior allele probability that $X = 1$. The posterior allele probabilities are correctly calibrated if $E(X|Z) = Z$. If the posterior allele probabilities are correctly calibrated, we can use the law of total expectation and the fact that $X^2 = X$ to obtain

$$E(X^2) = E(X) = E\big(E(X|Z)\big) = E(Z) \tag{5}$$

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = E(Z) - E(Z)^2 \tag{6}$$

$$\begin{aligned}\mathrm{Cov}(X,Z) &= E(XZ) - E(X)E(Z) = E\big(E(XZ|Z)\big) - E\big(E(X|Z)\big)E(Z) \\ &= E(Z^2) - E(Z)E(Z) = \mathrm{Var}(Z).\end{aligned} \tag{7}$$

Thus,

$$r^2 = \frac{\big(\mathrm{Cov}(X,Z)\big)^2}{\mathrm{Var}(X)\mathrm{Var}(Z)} = \frac{\mathrm{Var}(Z)}{\mathrm{Var}(X)} = \frac{E(Z^2) - E(Z)^2}{E(Z) - E(Z)^2}. \tag{8}$$

2) The allelic R^2 estimates the correlation between the true allele dose and the most probable (i.e. best guess) allele dose (Browning & Browning 2008). When the most probable target allele is the same for all samples, allelic R^2 cannot be computed.
3) IMPUTE 'INFO' measure is not directly correlation based, but is the ratio of the observed and complete information (Marchini & Howie 2010).

All measures are highly correlated and designed so that 0 indicates a complete uncertainty in the imputed alleles and 1 refers to no uncertainty (Marchini & Howie 2010). All of these measures can be interpreted as the approximate reduction in sample size when testing imputed alleles instead of the true alleles (Pritchard & Przeworski 2001). In general, poorly imputed markers are removed from downstream analysis based on some threshold (e.g. threshold of 0.3 or larger for meta-analysis and 0.7 or larger in single cohort GWAS) (Zeggini et al. 2008).

When true genotypes are known (e.g. NGS data is available for the imputed individuals), one can estimate imputation accuracy using non-reference (NR) sensitivity and non-reference discordancy rate (DePristo et al. 2011). Let X_i be the number of NR alleles for genotype call i in call set X and $X_{nr} = \{i \in X: X_i > 0\}$. Then

$$\mathrm{NR}_{\text{sensitivity}}(E,C) = \frac{|E_{nr} \cap C_{nr}|}{|C_{nr}|} \tag{9}$$

and

$$\mathrm{NR}_{\text{discordancy rate}}(E,C) = \frac{|i \in E_{nr} \cup C_{nr}: E_i \neq C_i|}{|E_{nr} \cup C_{nr}|}, \tag{10}$$

where E and C represent evaluated and compared to (inferred with an independent methodology) call sets, respectively. For multiple samples, (9) and (10) are averaged over samples and generally over MAF categories.

4. RESULTS AND DISCUSSION

4.1. Cohort description

The Estonian Biobank is a population-based biobank of the Estonian Genome Center, University of Tartu (EGCUT), containing almost 52,000 individuals of the Estonian population (aged ≥18 years), which closely reflects the age, sex and geographical distribution of the Estonian adult population (Leitsalu et al. 2015). All biobank participants have signed a broad informed consent form, which allows linking to national registries, electronic health record databases and hospital information systems. The majority of biobank participants have been analysed using genotyping arrays, and many of the samples have undergone extensive genomic characterization such as exome sequencing (∼2,700 individuals) and high-coverage PCR-free genome sequencing (∼3,000 individuals). In the following studies, we have taken advantage of the valuable data resource afforded by the Estonian Biobank to conduct several types of analysis.

4.2. Exome sequencing analysis (Refs. I–III)

The first part of the thesis deals with clinical exome sequencing analysis aiming to detect rare disease-related variants. The main challenges of the analysis are: (1) to convert sequenced reads to the genotype calls, and (2) to identify disease-related alleles among multiple non-pathogenic alleles and discard sequencing and alignment artefacts. The first part was conducted in a similar manner for all three studies: the raw sequencing data were aligned against the GRCh37/hg19 human genome reference using BWA. The GATK best practice guideline was applied for further BAM processing, and the GATK VQSR was used for variant quality control. All variants were annotated with Variant Effect Predictor, and a custom script was used for additional annotations. In particular, we added allele frequencies from population databases such as 1000G and NHLBI-ESP, pathogenic variant database HGMD, and *in silico* pathogenicity predictions SIFT and PolyPhen-2. In addition, allele counts from our increasing in-house database of variants detected among all NGS analyses (WGS and WES) performed in the Estonian Biobank were included to the annotation.

For the second step, to find causal disease-related variants, we determined the mode of inheritance and the extent of locus heterogeneity. To find rare or novel variants in the same gene shared among affected individuals, we used the variants databases – presence in the dbSNP or MAF > 1% in the 1000G or NHLBI-ESP datasets shortened our candidate variants list. Another exclusion criterion was if the variants were present in the in-house set of Estonian NGS samples. Finally, we were focusing on nonsynonymous and LoF variants, *in silico* analysis was performed to verify the variant's pathogenic status, and all findings were confirmed by Sanger sequencing.

4.2.1. Analysis of non-syndromic tooth agenesis (Ref. I)

Tooth agenesis, the congenital absence of one or more permanent teeth, is the most common abnormality of human dentition with a prevalence 2.2% – 10.1% (Polder et al. 2004). Affected family members often demonstrate a significant variability with regard to the position, morphology, symmetry, and number of teeth involved. Tooth agenesis occurs either in association with genetic syndromes based on the presence of other inherited abnormalities, as a non-syndromic familial trait or as a sporadic finding (Gorlin et al. 2001). Familial tooth agenesis has been reported to have either an autosomal-dominant, autosomal-recessive, or X-linked mode of inheritance. It is reported that more than 300 genes are involved in tooth morphogenesis (Thesleff 2006), indicating that there are many genes underlying regulatory mechanisms of tooth agenesis.

We conducted exome sequencing analysis in an Estonian family with variable degrees of tooth agenesis (Figure 1, Ref. I). Particularly, we sequenced one unaffected and four affected individuals. After bioinformatics data processing and variant prioritization, we detected 235 novel variants that were shared by affected female patients. Among these variants, we discovered a novel nonsense mutation c.874G>T (p.Glu292X) in the TNF homology domain of *EDA* – a previously known tooth agenesis candidate gene encoding for ectodysplasin-A. Sanger sequencing confirmed that all affected female patients were heterozygous carriers, while both the unaffected father and a half-brother did not carry this mutation (Figure 2A, Ref. I). Parental testing demonstrated that this variant arose *de novo*, and the risk-associated allele was transmitted to affected offspring from their mother. The Glu292 position is highly conserved in the other known *EDA* proteins (Figure 2C, Ref. I), suggesting that it has an important function in the protein. We confirmed that *EDA* mutations are involved in the underlying regulatory pathways of the development of teeth, but further in-depth molecular studies are required to clarify their role.

4.2.2. Analysis of class III malocclusion (Ref. II)

Class III malocclusion is a heterogeneous dentofacial phenotype that is skeletally characterized by the overgrowth of the mandible, the undergrowth of the maxilla, or a combination of both with a prevalence of 4–23% (Singh 1999). The inheritance pattern of class III malocclusion is controversial, and it is presumed to occur as a multifactorial trait for a majority of affected individuals with a variety of phenotypic subtypes. Numerous genome-wide linkage scans have identified chromosomal regions that might harbour susceptibility genes for class III malocclusion in several populations (Yamaguchi et al. 2005; Frazier-Bowers et al. 2009; Li et al. 2011; Jang et al. 2010; Cruz et al. 2011).

We performed exome sequencing in four affected and one unaffected siblings from an Estonian family consisting of 21 members from four generations (Figure 1, Ref. II). Following bioinformatics data processing and variant

filtering detected 14 rare non-synonymous SNVs shared among the four affected male siblings and a carrier female (Table 2, Ref. II), including a heterozygous missense mutation c.545C>T (p.Ser182Phe) in exon 2 of the *DUSP6* gene (encoding the dual-specificity phosphatase 6) shared by all five siblings. This missense mutation affects a highly conserved amino acid, and *in silico* analysis predicted this variant to be probably pathogenic. None of the remaining rare or novel variants were considered plausible candidates. Sanger sequencing was used for confirmation of the findings (Figure 2A, Ref. II). The Ser182 position is highly conserved in *DUSP6* proteins of other species (Figure 2B, Ref. II), suggesting that this residue is important for the function of the *DUSP6* protein.

This study demonstrates that class III malocclusion familial distribution may be explained by the presence of a dominant major gene under the influence of other modifier genes and environmental factors. The increased knowledge of genetic risk factors for class III malocclusion will be necessary for an understanding of how the molecular mechanisms underlying this phenotype may influence the response to dentofacial and orthodontic treatment and allows clinicians to develop more effective targeted intervention strategies to prevent the development of class III malocclusion.

4.2.3. Analysis of epileptic encephalopathy with neonatal beginning (Ref. III)

Epileptic encephalopathies refer to a severe condition where epileptic activity itself can contribute to progressive cognitive, behavioral, and motor dysfunction. However, the encephalopathic effect of seizures can occur in association with any form of epilepsy (Berg et al. 2010). Children with severe early-onset epilepsies are thought to be at more risk and typically have a poor prognosis (Cross & Guerrini 2013). Several genes have been associated with early infantile epileptic encephalopathy, but determining the underlying cause can be challenging because of genetic and phenotypic heterogeneity.

Exome sequencing was performed in affected boy and his healthy parents. After bioinformatics data processing and variant prioritization, we identified a novel heterozygous missense mutation c.3979A>G in exon 22 of *SCN8A*, predicting a p.Ile1327Val substitution. *In silico* analysis suggested that the mutation has a deleterious effect on the protein function. Also, the affected amino acid is located at an extremely conserved position (Figures 2A and 2B, Ref. III). The variant c.3979A>G was confirmed as arising *de novo* in the proband with Sanger sequencing (Figure 2C, Ref. III).

This study implicated *SCN8A* in the pathogenesis of epileptic encephalopathy with neonatal beginning and demonstrates the value of WES data in clinical settings. Further investigations will be worthwhile to determine the prevalence and significance of *SCN8A* mutations in epileptic encephalopathies.

4.3. Population-based genome sequencing analysis with blood cell measurements (Ref. IV)

Hematopoiesis is a process by which blood cells are formed. Although hematopoiesis is perturbed in a variety of human blood disorders and shows considerable interindividual variation, the underlying basis of the disease etiology and variation remains incompletely understood (Sankaran & Orkin 2013). Several studies have shown that genetic variants can influence blood cell measurements, and result in rare blood disorders (Van der Harst et al. 2012; Ulirsch et al. 2016; Orrù et al. 2013).

To perform a GWAS analysis with 14 blood cell measurements, we used high-coverage WGS data of 2,284 individuals and chip-based SNVs of 14,904 individuals from the Estonian Biobank (Figure S1, Ref. IV). For a small subset of the samples (∼2000), blood cell measurements were directly assayed in a laboratory, for the rest, the measurements were accessed from their electronic medical records (EMRs). In general, the blood cell measurements were strongly correlated (Figure S2, Ref. IV), also laboratory-based and EMR-based values showed high concordance (Figure S3, Ref. IV).

The single variant analysis detected 17 genome-wide significant associations across the various blood cell measurements (Table 1, Ref. IV). All but one of these associations have been previously reported and highlighted important biological mechanisms. However, we detected a previously undiscovered association with basophil counts near *CEBPA* gene (rs78744187; $P = 6.19 \times 10^{-38}$) (Figure 1, Ref. IV). Following fine-mapping in 17 detected regions provided insight into the molecular regulatory mechanisms. Gene-based burden testing of rare variants (MAF < 5%) did not detect significant associations.

We demonstrated that high-coverage WGS data can be used to discover novel common variants associated with human traits and diseases compared to chip-based or imputed data. We also showed that in a population-based biobank study, one can link genetic data with EMRs to greatly increase sample sizes. Although for rare variant analysis, only WGS-based datasets may be likely unpowered.

4.4. Genotype imputation using population-specific reference panel (Refs. V-VI)

A GWAS is a widely-used instrument for detecting associations between genetic variants and phenotypic traits, which mostly captures small to modest effect sizes. However, even in aggregate, these explain only a small fraction of the heritability of studied traits. Although GWASs have successfully identified thousands of common (MAF > 5%) trait-related variants, they are unpowered to detect associations with rare variants. To increase the resolution of GWASs, genotype imputation is routinely implemented to incorporate variants that are not directly genotyped.

Although several factors influence imputation accuracy, the genetic similarity between individuals in the imputation reference panel and in the genotyped individuals seems to be very dominant, especially for the imputation of rare variants (MAF < 1%). In the last part of the thesis, we introduce a WGS-based population-specific imputation reference panel and apply it in Estonians. We systematically study its impact on imputation accuracy of rare variants (Ref. V) and downstream effects of genome-wide association analysis (Ref. VI).

4.4.1. Evaluation of imputation accuracy of rare variants using population-specific imputation reference panel (Ref. V)

Imputation accuracy in a specific population depends largely on the size of IRP, and genetic similarities between IRP and study samples. However, can smaller population-specific IRPs outperform a large number of reference haplotypes from diverse populations?

To address this question, we used high-coverage WGS data of 2,244 individuals from the EGCUT, and created an imputation reference panel specific to Estonians. For comparison, we used two ethnically heterogeneous IRPs (1000G and HRC), and two combinations of these panels (EGCUT + 1000G and 1000G + EGCUT) to impute SNVs into 6,394 Estonians (Table 2, Ref. V). IRPs and chip-based genotype data was pre-phased using SHAPEIT2 (Delaneau et al. 2013), followed by IMPUTE2 (Howie et al. 2009) imputation. IMPUTE2 allows improving imputation accuracy by using two reference panels simultaneously by pooling haplotype information across both IRPs.

We compared phasing speed and accuracy with three programs – SHAPEIT2, SHAPEIT2-RA (for read-aware) and Eagle2 (Loh et al. 2016). We observed that switch error rate was slightly smaller using SHAPEIT2 or SHAPEIT2-RA, but it was achieved by more time-consuming computation compared to Eagle2 (Table 1, Ref. V). SHAPEIT2-RA did not outperform SHAPIT2 in phasing accuracy.

For each IRP, we studied the number of SNVs as a function of the imputation confidence estimate (INFO-value) and performed separate analysis for 'well-imputed' (INFO > 0.4) and 'confidently imputed' (INFO > 0.8) SNVs. Although the number of total variants and well-imputed variants obtained with the larger diverse panels (1000G and HCR) exceeded the corresponding numbers for the population-specific panel, the situation was reversed for confidently imputed SNVs (Figure 1b, Ref. V). Looking the same by MAF categories of the imputed SNVs (Figure 2, Ref. V), the number of imputed common (MAF ≥ 5%) variants was very similar across IRPs, but we detected large differences for rare variants (MAF < 0.5%), where 3.48 M, 2.54 M and 1.86 M SNVs were imputed confidently with EGCUT, HRC and 1000G panels, respectively. Population-specific panel outperformed similarly in the analysis of confidently imputed LoF and missense variants, providing almost twice as many rare LoF variants compared to both diverse panels (Figure 3, Ref. V).

For further validation, we had exome sequencing data available for 505 imputed EGCUT individuals. Treating these WES-based genotype calls as 'gold standard', we calculated sensitivity and discordancy rates for each imputed datasets. For well-imputed common SNVs, all of the IRPs gave similarly high sensitivities (88.5–92.4%) (Figure 4a, Ref. V) and low discordancy rates (1.9–3.4%) (Figure 4b, Ref. V). For low-frequency ($0.5 \leq MAF < 5\%$) and rare SNVs, the three panels that included data from the population-specific panel (EGCUT, EGCUT+1000G, and 1000G+EGCUT) yielded a higher sensitivity and lower discordancy rate compared to more diverse panels (Table 3, Ref. V). The differences were greater for rare SNVs than for low-frequency variants. Notably, one-quarter (24.7%) of rare SNVs imputed from the 1000G IRP had incorrect genotype calls, whereas the proportion was substantially lower with the EGCUT IRP alone (14.1%). We observed similar results for confidently imputed variants (Figure S4 and Table S4, Ref. V). In the analysis of finer MAF categories, the accuracy of genotype imputation of well-imputed variants decreased in the lower MAF bins for all compared IRPs (Figures S5–S9, Ref. V). But in case of a population-specific IRP (Figure S7, Ref. V), imputation accuracy was significantly better for rare variants, achieving relatively confident imputation of variants down to MAF of 0.2%.

We did not detect any major differences when imputing common variants. But imputation of rare variants in Estonians reveals that a population-specific IRP (or used in combination with publicly available references such as the 1000G IRP) outperforms larger IRPs from diverse populations. The majority of rare variants imputed with 1000G or HRC IRPs have low confidence. It is important to note that high imputation confidence estimates (like INFO-value) do not guarantee that the corresponding genotypes are inferred correctly. For example, diverse IRPs contain divergent haplotypes, which are not present in the target samples and may result in SNVs that are not actually polymorphic in the study population. Also, we saw that publically available IRPs are limited in imputing population-specific SNVs.

4.4.2. Advantages of genotype imputation with ethnically-matched reference panel for rare variant association analyses (Ref. VI)

Population-specific IRPs are implemented in several populations (e.g. in Finns, British, Dutch, Sardinians and Icelandic populations), but their genome-wide downstream consequences are not comprehensively explored. We determined and quantified these differences by performing several comparative evaluations of the biologically motivated analysis scenarios.

We developed a WGS-based IRP containing ethnically closely related 2,279 Estonians and 1,856 Finns, which is referred to as the Estonian-Finnish (EstFin) ethnically matched IRP. We imputed 36,716 unrelated Estonian Biobank samples with the EstFin, and ethnically mixed 1000G IRPs (Figure S1 and Table S1,

Ref. VI) and conducted comparative GWAS and gene-wise association testing of rare variants (MAF < 1%) with body mass index and seven complex traits (Table S2, Ref. VI). Only confidently imputed (INFO > 0.8) variants (SNVs and indels) were considered and all association analysis performed separately in both imputed datasets (Figure 1, Ref. VI).

Genome-wide association analysis, followed by fine-mapping and MAF-enriched analysis, did not detect any major differences between the imputed datasets. Single variant analysis replicated previously reported common variant associations – 12 and 13 significant loci based on EstFin and 1000G IRPs, respectively (Figure S3 and Table 1, Ref. VI). In the gene-based analysis of rare (MAF < 1%) coding (LoF and missense) variants, we observed 10-fold differences in the number of tested genes and significant gene-trait associations between reference panels. In the EstFin-based imputed data we detected 48 gene-trait associations and only four in the 1000G-based data (Figures 3 and S4, Table 2, Ref. VI).

We demonstrated empirically that imputed data based on ethnically-matched panel is very promising for rare variant analysis – it captures more population-specific variants and makes it possible to efficiently identify novel findings compared to ethnically-mixed panels.

CONCLUSIONS

Next-generation sequencing technology enables large-scale, routine sequencing in large cohorts. In the current thesis, we demonstrated that the analysis of NGS data has a huge potential in several fields, but also requires a massive computational power. Also, with the increase of data volumes, there is an incessant need for the development of computational and statistical methods.

Exome sequencing has been implemented successfully in clinical practice. Covering the whole spectrum of protein-coding regions in a cost-effective way, it opens new opportunities for quick and exact large-scale screenings not attainable with alternative methods like Sanger sequencing or the use of imputed datasets. Particularly, exome sequencing is efficient for detecting very rare and *de novo* mutations. The first part of the thesis demonstrated that this approach is suitable for the Estonian clinical data as well. We analysed three families with Mendelian diseases and detected potentially causative gene variants for each case. These projects highlighted that a tight collaboration between data scientists and medical geneticists can lead to findings with considerable impact in the research of rare genetic disorders and have the potential to lead to successful therapies in the future.

Also, we experienced that rich genomic data is not always sufficient for a successful study. Population-based biobanks (like Estonian Biobank) provide numerous opportunities for expanding phenotypic datasets by additional measurements or taking advantages from national databases. We used additional blood cell measurements from the electronic medical records and our genome-wide scan detected previously undiscovered association with basophil counts near *CEBPA* gene, and highlighted their role in the autoimmune regulation. This example opens new dimensions for scanning underlying genetic basis for a variety of traits and diseases.

To increase the resolution of genome-wide association analysis, genotype imputation is routinely implemented to incorporate variants that are not directly genotyped. Imputation is performed based on reference haplotypes. We had an opportunity to construct an imputation reference panel to Estonians based on high-coverage genome sequencing data. We showed that the utilization of a population-specific reference panel provided significantly higher imputation confidence for rare variants compared to larger, multi-ethnic panels. Also, the population-specific panel yielded a higher sensitivity and lower discordancy rate than the more diverse panels. In the downstream association analysis, we did not experience differences in respect to the common genetic variants, but observed a huge gain in gene-based rare variant analysis. As one of the main results of this thesis, the Estonian-specific imputation reference panel is created, tested and ready to serve for a long time. This includes genotyping and imputing the data in the framework of ongoing personalised medicine initiative to invite 100,000 Estonians to join the Estonian Biobank cohort, with the purpose to develop more precise and improved disease prevention and treatment guides.

SUMMARY IN ESTONIAN

Arvutuslikud ja statistilised meetodid DNA sekveneerimisandmete analüüsimiseks ja rakendused TÜ Eesti Geenivaramu andmetel

Sekveneerimismeetodite areng viimastel aastakümnetel on olnud tormiline. Selle tulemusena määrati inimese genoomi DNA järjestus käesoleva sajandi alguses. Tänapäeval võimaldavad nn. teise põlvkonna sekveneerimisel (*next-generation sequencing*, NGS) põhinevad meetodid kulu-efektiivselt määrata inimese genoomi järjestuse vähem kui ööpäevaga. Seejuures toodetakse väga suuri andmemahtusid, mis omakorda tekitavad mitmeid väljakutseid nii informaatika kui statistika valdkonnas. Näiteks on tekkinud vajadus suurte arvutusklastrite järele, samuti peab järjepidevalt arendama andmetele sobilikke meetodeid, statistilisi mudeleid ja analüüsitarkvara.

Enne igat analüüsi vajavad sekveneerimisandmed põhjalikku eel-protsessimist ja kvaliteedikontrolli, kus võetakse arvesse võimalikke vigu. Kui andmed on analüüsimiseks valmis, on nende kasutusala väga lai, kuna nende abil on võimalik mõõta väga täpselt geneetilist variatsiooni kogu genoomis võrreldes varasemate meetoditega.

Paljud riigid on alustanud suuremahulisi geeniuuringuid, kaasaarvatud Eesti. TÜ Eesti Geenivaramu on juba aastatel 2002–2011 kogunud enam kui 50 000 inimese geeniproovi ja käesoleval aastal lisandub sellele veel 100 000. Praeguseks hetkeks on üle 5 500 geenidoonori DNA-d analüüsitud erinevate NGS meetoditega ja just neile andmetele keskendubki käesolev doktoritöö. Välja on pakutud üldine raamistik NGS andmete töötluseks ning lisaks on uuritud, kuidas võimalikult hästi arvestada Eesti päritolu isikute võimalikku geneetilist eripära.

Üheks levinud NGS meetodiks on eksoomi ehk kõigi valku kodeerivate geenipiirkondade sekveneerimine. See meetod on laialdaselt rakendust leidnud meditsiinigeneetikas ühe geeni poolt määratud ehk mendeliaarsete haiguste geenimutatsioonide tuvastamisel, kuna võimaldab efektiivselt leida harvu ja *de novo* geenivariante. Doktoritöö esimene osa demonstreerib, et selline lähenemine töötas ka Eesti andmetel, kus me analüüsisime kolme perekonna andmeid. Kõigil kolmel juhul tegime eksoomi sekveneerimisega kindlaks patogeense mutatsiooni, mis lubab tulevikus välja töötada paremaid ravimeetodeid. Siinkohal peab lisama, et niisugused projektid eeldavad head koostööd statistikute ja meditsiinigeneetikute vahel.

Samuti viisime läbi kogu genoomi sekveneerimisandmete analüüsi kliinilise vere näitajatega. See analüüs tõi välja populatsioonipõhise biopanga eelised, mis lisaks rikkalikele genoomiandmetele sisaldab ka väärtuslikku informatsiooni erinevate haiguste ja tunnuste kohta ning vajadusel võimalust neid täiendada linkimisega üleriigilistest andmebaasidest. Selles uuringus me küsisime puuduolevaid kliinilise vere andmeid E-tervise andmebaasist. Sel viisil teostatud uuring aitas meil tuvastada olulisi seoseid *CEBPA* geenivariantide ja basofiilide

arvu vahel, kusjuures viimasel on roll mitmete autoimmuunhaiguste sümptomaatikas. Seega näitasime, et populatsioonipõhise biopanga andmetega on võimalik ka edaspidi edukalt uurida geenide seoseid huvipakkuvate tunnuste ja haigustega.

Ülegenoomsetes assotsiatsiooniuuringutes analüüsitakse eelkõige genotüübiandmeid, mis on saadud nn genotüpiseerimiskiipide abil, mis võimaldavad määrata kuni miljon erinevat geneetilist varianti. Selliste uuringute võimsuse suurendamiseks kasutatakse puuduvate geenivariantide ennustamist ehk imputeerimist, mis põhineb haplotüüpide referentsandmestikul. Muutmaks just Eesti päritolu isikute andmeanalüüsi tõhusamaks, kasutasime TÜ Eesti Geenivaramu genoomide sekveneerimisandmeid eestlaste haplotüüpide referentsandmestiku loomiseks. Seejärel imputeerisime puuduvaid geenivariante kolmel viisil – kasutades nii eestlaste-spetsiifilist kui ka kahte multi-etnilist referentspaneeli. Võrdlustulemused näitasid, et eestlaste-spetsiifilise referentspaneeli kasutamisel õnnestus määrata rohkem parema kvaliteediga geenivariante, ning loodud paneeli eelis tuleb eriti esile harvaesinevate variantide puhul.

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249.

Artigas, M. S., Wain, L. V., Miller, S., Kheirallah, A. K., Huffman, J. E., Ntalla, I., … Tobin, M. D. (2015). Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nature Communications*, *6*(1), 8658.

Auer, P. L., Reiner, A. P., Wang, G., Kang, H. M., Abecasis, G. R., Altshuler, D., … Leal, S. M. (2016). Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. *The American Journal of Human Genetics*, *99*(4), 791–801.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, *12*, 745–755.

Baum, L. E., & Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. *Bulletin of the American Meteorological Society*, *73*, 360–363.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59.

Berg, A. T., Berkovic, S. F., Brodie, M. J., Buchhalter, J., Cross, J. H., van Emde Boas, W., … Scheffer, I. E. (2010). Revised terminology and concepts for organization of seizures and epilepsies: Report of the ILAE Commission on Classification and Terminology, 2005-2009. *Epilepsia*, *51*(4), 676–685.

Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L., & Trees, E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical Microbiology and Infection*, *24*(4), 335–341.

Browning, B. L., & Browning, S. R. (2008). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, *84*(2), 210–223.

Browning, B. L., & Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *The American Journal of Human Genetics*, *98*(1), 116–126.

Browning, S. R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. *Human Genetics*, *124*(5), 439–450.

Burrows, M., Burrows, M., & Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm. *Technical report 124*. Digital Equipment Corporation, Palo Alto, CA, USA.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92.

Cooper, G. M., Goode, D. L., Ng, S. B., Sidow, A., Bamshad, M. J., Shendure, J., & Nickerson, D. A. (2010). Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nature Methods*, *7*(4), 250–251.

Cross, J. H., & Guerrini, R. (2013). The epileptic encephalopathies. *Handbook of Clinical Neurology*, *111*, 619–626.

Cruz, R. M., Hartsfield, J. K., Falcão-Alencar, G., Koller, D. L., Pereira, R. W., Mah, J., … Oliveira, S. F. (2011). Exclusion of Class III Malocclusion Candidate Loci in Brazilian Families. *Journal of Dental Research*, *90*(10), 1202–1205.

Das, S., Abecasis, G. R., & Browning, B. L. (2018). Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics*, *19*(1), 73–96.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287.

Deelen, P., Menelaou, A., Van Leeuwen, E. M., Kanterakis, A., Van Dijk, F., Medina-Gomez, C., … Swertz, M. A. (2014). Improved imputation quality of low-frequency and rare variants in European samples using the "Genome of the Netherlands." *European Journal of Human Genetics*, *22*(11), 1321–1326.

Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm on JSTOR. *Journal of the Royal Statistical Society*, *39*(1), 1–38.

Den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., Mcgowan-Jordan, J., … Taschner, P. E. M. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human Mutation*, *37*, 564–569.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498.

Dietz, H. C. (2010). New Therapeutic Approaches to Mendelian Disorders. *New England Journal of Medicine*, *363*(9), 852–863.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, *323*(5910), 133–138.

Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, *8*(3), 175–185.

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., … Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, *23*(22), 5866–5878.

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., & Loeb, L. A. (2014). Accuracy of Next Generation Sequencing Platforms. *Next Generation, Sequencing & Applications*, *1*: 1000106.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., … Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861.

Frazier-Bowers, S., Rincon-Rodriguez, R., Zhou, J., Alexander, K., & Lange, E. (2009). Evidence of Linkage in a Hispanic Cohort with a Class III Dentofacial Phenotype. *Journal of Dental Research*, *88*(1), 56–60.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., … Lee, C. (2006). Copy number variation: new insights in genome diversity. *Genome Research*, *16*(8), 949–961.

Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). minimac2: faster genotype imputation. *Bioinformatics*, *31*(5), 782–784.

Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Gilissen, C., Hoischen, A., Brunner, H. G., & Veltman, J. A. (2012). Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics*, *20*(5), 490–497.

Gorlin, R., Cohen, M., & Hennekam, R. (2001). *Syndromes of the head and neck*. New York: NY: Oxford University Press.

Gormley, P., Anttila, V., Winsvold, B. S., Palta, P., Esko, T., Pers, T. H., … Palotie, A. (2016). Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nature Genetics*, *48*(8), 856–866.

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., … Stefansson, K. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, *47*(5), 435–444.

Helgason, H., Sulem, P., Duvvari, M. R., Luo, H., Thorleifsson, G., Stefansson, H., … Stefansson, K. (2013). A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. *Nature Genetics, 45*, 1371–1374.

Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadottir, H. T., Zanon, C., … Stefansson, K. (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nature Genetics*, *43*(4), 316–323.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959.

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, *5*(6), e1000529.

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., … Zhang, W. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications*, *6,* 8111.

Human Genome Sequencing Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945.

International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*(7311), 52–58.

Jang, J. Y., Park, E. K., Ryoo, H. M., Shin, H. I., Kim, T. H., Jang, J. S., … Kwon, T. G. (2010). Polymorphisms in the *Matrilin-1* Gene and Risk of Mandibular Prognathism in Koreans. *Journal of Dental Research*, *89*(11), 1203–1207.

Jonsson, T., Stefansson, H., Steinberg, S., Jonsdottir, I., Jonsson, P. V, Snaedal, J., … Stefansson, K. (2013). Variant of *TREM2*Associated with the Risk of Alzheimer's Disease. *New England Journal of Medicine*, *368*(2), 107–116.

Kaiser, J. (2010). Affordable "exomes" fill gaps in a catalog of rare diseases. *Science (New York, N.Y.)*, *330*(6006), 903.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315.

Kryukov, G. V., Pennacchio, L. A., & Sunyaev, S. R. (2007). Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies. *The American Journal of Human Genetics*, *80*(4), 727–739.

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073–1081.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25.

Leeuwen, E. M. van, Sabo, A., Bis, J. C., Huffman, J. E., Manichaikul, A., Smith, A. V, … Duijn, C. M. van. (2016). Meta-analysis of 49 549 individuals imputed with the 1000 Genomes Project reveals an exonic damaging variant in ANGPTL4 determining fasting TG levels. *Journal of Medical Genetics*, *53*(7), 441–449.

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M. L., Alavere, H., Snieder, H., … Metspalu, A. (2015). Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. *International Journal of Epidemiology*, *44*(4), 1137–1147.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … Consortium, E. A. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291.

Levy, S. E., & Myers, R. M. (2016). Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics*, *17*(1), 95–115.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

Li, N., & Stephens, M. (2003). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, *165*(4), 4 2213-2233.

Li, Q., Li, X., Zhang, F., & Chen, F. (2011). The Identification of a Novel Locus for Mandibular Prognathism in the Han Chinese Population. *Journal of Dental Research*, *90*(1), 53–57.

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*(8), 816–834.

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype Imputation. *Annual Review of Genomics and Human Genetics*, *10*(1), 387–406.

Lin, Y., Liu, L., Yang, S., Li, Y., Lin, D., Zhang, X., & Yin, X. (2018). Genotype imputation for Han Chinese population using Haplotype Reference Consortium as reference. *Human Genetics*, *137*(6), 431–436.

Liu, J. Z., Tozzi, F., Waterworth, D. M., Pillai, S. G., Muglia, P., Middleton, L., … Marchini, J. (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nature Genetics*, *42*(5), 436–440.

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., … Price, A. L. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, *48*(11), 1443–1448.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., … Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods*, *7*(2), 111–118.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, *39*(7), 906–913.

Marx, V. (2015). The DNA of a nation. *Nature*, *524*(7566), 503–505.
McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., … Marchini, J. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283.
McKusick, V. A. (2007). Mendelian Inheritance in Man and its online version, OMIM. *American Journal of Human Genetics*, *80*(4), 588–604.
McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122.
Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46.
Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V, Steinthorsdottir, V., … McCarthy, M. I. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, *44*(9), 981–990.
Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, *19*(R2), R131–R136.
Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., … Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, *42*(1), 30–35.
Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443–451.
Orrù, V., Steri, M., Sole, G., Sidore, C., Virdis, F., Dei, M., … Cucca, F. (2013). Genetic Variants Regulating Immune Cell Levels in Health and Disease. *Cell*, *155*(1), 242–256.
Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., … Price, A. L. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*, *44*(6), 631–635.
Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., … Sanna, S. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: Implications for cost-effective study designs. *European Journal of Human Genetics*, *23*(7), 975–983.
Polder, B. J., Van't Hof, M. A., Van der Linden, F. P. G. M., & Kuijpers-Jagtman, A. M. (2004). A meta-analysis of the prevalence of dental agenesis of permanent teeth. *Community Dentistry and Oral Epidemiology*, *32*(3), 217–226.
Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–121.
Pritchard, J. K., & Przeworski, M. (2001). Linkage Disequilibrium in Humans: Models and Data. *The American Journal of Human Genetics*, *69*(1), 1–14.
Quick, J., Quinlan, A. R., & Loman, N. J. (2014). A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GigaScience*, *3*(1), 22.
Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rands, C. M., Meader, S., Ponting, C. P., & Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genetics*, *10*(7), e1004525.

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, *58*(4), 586–597.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., … Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348–352.

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441–448.

Sankaran, V. G., & Orkin, S. H. (2013). Genome-wide association studies of hematologic phenotypes: a window into human hematopoiesis. *Current Opinion in Genetics & Development*, *23*(3), 339–344.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311.

Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., … Abecasis, G. R. (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature Genetics*, *47*(11), 1272–1281.

Singh, G. D. (1999). Morphologic determinants in the etiology of class III malocclusions: A review. *Clinical Anatomy*, *12*(5), 382–405.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., … Loos, R. J. F. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*(11), 937–948.

Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., … Stefansson, K. (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nature Genetics*, *46*(3), 294–298.

Stenson, P. D., Ball, E. V, Howells, K., Phillips, A. D., Mort, M., & Cooper, D. N. (2009). The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalised genomics. *Human Genomics*, *4*(2), 69.

Stephens, M., & Donnelly, P. (2003). A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. *The American Journal of Human Genetics*, *73*(5), 1162–1169.

Sundquist, A., Ronaghi, M., Tang, H., Pevzner, P., & Batzoglou, S. (2007). Whole-Genome Sequencing and Assembly with High-Throughput, Short-Read Technologies. *PLoS ONE*, *2*(5), e484.

The Cancer Genome Atlas Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330–337.

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74.

Thesleff, I. (2006). The genetic basis of tooth development and dental defects. *American Journal of Medical Genetics Part A*, *140A*(23), 2530–2535.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46.

Turner, E. H., Ng, S. B., Nickerson, D. A., & Shendure, J. (2009). Methods for Genomic Partitioning. *Annual Review of Genomics and Human Genetics*, *10*(1), 263–284.

Ulirsch, J. C., Nandakumar, S. K., Wang, L., Giani, F. C., Zhang, X., Rogov, P., … Sankaran, V. G. (2016). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell*, *165*(6), 1530–1545.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics* (Vol. 43, p. 11.10.1-11.10.33). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., … Chambers, J. C. (2012). Seventy-five genetic loci influencing the human red blood cell. *Nature*, *492*(7429), 369–375.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, *291*(5507), 1304–1351.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269.

Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., … Zhang, W. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*(7571), 82–89.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164.

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of Nucleic Acids. *Nature*, *171*, 737–738.

Yamaguchi, T., Park, S. B., Narita, A., Maki, K., & Inoue, I. (2005). Genome-wide Linkage Analysis of Mandibular Prognathism in Korean and Japanese Patients. *Journal of Dental Research*, *84*(3), 255–259.

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., … Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, *40*(5), 638–645.

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829.

Zheng, H.-F., Ladouceur, M., Greenwood, C. M. T., & Richards, J. B. (2012). Effect of Genome-Wide Genotyping and Reference Panels on Rare Variants Imputation. *Journal of Genetics and Genomics*, *39*(10), 545–550.

Zhou, W., Fritsche, L. G., Das, S., Zhang, H., Nielsen, J. B., Holmen, O. L., … Willer, C. J. (2017). Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. *Genetic Epidemiology*, *41*(8), 744–755.

ACKNOWLEDGMENTS

It has been a marathon. A large amount of details have to be perfect for an excellent performance. Starting with coaches, I would like to emphasize the importance of my supervisor, Professor Krista Fischer, who has improved my techniques in statistics through the years. In addition, unofficial supervisors Priit Palta and Reedik Mägi have introduced me a beauty of bioinformatics. During the years, Professor Andres Metspalu has believed in me and supported my career. Also, my special gratitude goes to all co-authors I have worked and published together.

I have been surrounded by a special crowd at the Estonian Genome Center. It has been a great pleasure to work and produce an excellent science. My appreciation goes to Krista Liiv for all kind of paperwork to make things much easier for me. I have been blessed with roommates Tom Haller and Tõnis Tasa, who have balanced my working days with shared lunch hours and fruitful conversations.

Bioinformatics without a playground is impossible. During the years I have used tens of thousands computing hours at the High Performance Computing Center of the University of Tartu – Ivar Koppel and your team, thank you for providing such a great environment and massive support.

I would like to thank the Institute of Mathematics and Statistics for useful lectures. I am grateful to the Doctoral School of Mathematics and Statistics, especially Rainis Haller and Eve Oja, for fellowship nominations.

I wish to thank Paula Ann Kivistik, Tiit Nikopensius, and Märt Roosaare for critical reviewing of my thesis and helpful comments. I express my gratitude to the core facility and gene donors in the Estonian Genome Center for their enthusiasm.

Without family everything is worthless. Mom and Dad, thank you for sharing the genes. Tiina, thank you for your endless love and support. Johanna and Pipi-Linda, you are awesome.

PUBLICATIONS

CURRICULUM VITAE

Name: Mart Kals
Date of birth: March 10, 1981
Citizenship: Estonian
Address: Estonian Genome Center, Institute of Genomics,
University of Tartu
Riia 23B, 51010, Tartu, Estonia
Phone: (+372) 737 4041
E-mail: mart.kals@ut.ee

Education:

2010–2018	University of Tartu, doctoral studies in mathematical statistics
2004–2005	University Limburg, Belgium, master's studies in biostatistics
1999–2003	University of Tartu, bachelor's studies in mathematical statistics
1996–1999	Hugo Treffner Gymnasium

Employment:

2010–	Estonian Genome Center, Institute of Genomics, University of Tartu, specialist
2007–2009	Astra Export & Trading AB, Estonian Branch Office, business analyst
2005–2007	Resta Ltd., consultant

Supervised dissertations:

2016 Co-supervision of the master's thesis of Madli Tamm "Detecting cancer related mutations in cell-free DNA of lung cancer patients" (MSc in gene technology)

2011 Co-supervision of the bachelor's thesis of Silva Kasela "Genome-wide association study and its practical conducting on data from the Estonian Genome Center of the University of Tartu" (BSc in mathematical statistics)

LIST OF PUBLICATIONS

Tasa, T., Krebs, K., **Kals, M.**, Mägi, R., Lauschke, V., Haller, T., … Milani, L. (2018). Genetic variation in the Estonian population: pharmacogenomics study of adverse drug effects using electronic health records. *European Journal of Human Genetics*, (published online).

Alver, M., Palover, M., Saar, A., Läll, K., Zekavat, S. M., Tõnisson, N., …, **Kals, M.**, … Esko, T. (2018). Recall by genotype and cascade screening for familial hypercholesterolemia in a population-based biobank from Estonia. *Genetics in Medicine*, (published online).

Reisberg, S., Krebs, K., Lepamets, M., **Kals, M.**, Mägi, R., Metsalu, K., … Milani, L. (2018). Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. *Genetics in Medicine*, (published online).

Teeäär, T., Serg, M., Paapstel, K., Kals, J., **Kals, M.**, Zilmer, M., … Kampus, P. (2018). Heart rate reduction decreases central blood pressure in sick sinus syndrome patients with a permanent cardiac pacemaker. *Journal of Human Hypertension*, *32*(5), 377–384.

Lieberg, J., Pruks, L.-L., **Kals, M.**, Paapstel, K., Aavik, A., & Kals, J. (2018). Mortality After Elective and Ruptured Abdominal Aortic Aneurysm Surgical Repair: 12-Year Single-Center Experience of Estonia. *Scandinavian Journal of Surgery*, *107*(2), 152–157.

Pappa, L., **Kals, M.**, Kivistik, P. A., Metspalu, A., Paal, A., & Nikopensius, T. (2017). Exome analysis in an Estonian multiplex family with neural tube defects – a case report. *Child's Nervous System*, *33*(9), 1575–1581.

Mitt, M. *, **Kals, M.** *, Pärn, K. *, Gabriel, S. B., Lander, E. S., Palotie, A., … Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics*, *25*(7), 869–876.

Guo, M. H., Nandakumar, S. K., Ulirsch, J. C., Zekavat, S. M., Buenrostro, J. D., Natarajan, P., … **Kals, M.**, …, Sankaran, V. G. (2017). Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(3), E327–E336.

Vals, M.-A., Pajusalu, S., **Kals, M.**, Mägi, R., & Õunap, K. (2017). The Prevalence of PMM2-CDG in Estonia Based on Population Carrier Frequencies and Diagnosed Patients (pp. 13–17). Springer, Berlin, Heidelberg.

Mihailov, E., Nikopensius, T., Reigo, A., Nikkolo, C., **Kals, M.**, Aruaas, K., … Metspalu, A. (2017). Whole-exome sequencing identifies a potential TTN mutation in a multiplex family with inguinal hernia. *Hernia*, *21*(1), 95–100.

Ivanov, M., **Kals, M.**, Lauschke, V., Barragan, I., Ewels, P., Käller, M., … Ingelman-Sundberg, M. (2016). Single base resolution analysis of 5-hydroxymethylcytosine in 188 human genes: implications for hepatic gene expression. *Nucleic Acids Research*, *44*(14), 6756–6769.

Polfus, L. M., Khajuria, R. K., Schick, U. M., Pankratz, N., Pazoki, R., Brody, J. A., …, **Kals, M.**, … Sankaran, V. G. (2016). Whole-Exome Sequencing Identifies Loci Associated with Blood Cell Traits and Reveals a Role for Alternative GFI1B Splice Variants in Human Hematopoiesis. *The American Journal of Human Genetics*, *99*(2), 481–488.

Ganna, A., Genovese, G., Howrigan, D. P., Byrnes, A., Kurki, M. I., Zekavat, S. M., …, **Kals, M.**, … Neale, B. M. (2016). Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nature Neuroscience*, *19*(12), 1563–1565.

Gormley, P., Anttila, V., Winsvold, B. S., Palta, P., Esko, T., Pers, T. H., …, **Kals, M.**, … Palotie, A. (2016). Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nature Genetics*, *48*(8), 856–866.

Tšuiko, O., Nõukas, M., Žilina, O., Hensen, K., Tapanainen, J. S., Mägi, R., …, **Kals, M.**, … Kurg, A. (2016). Copy number variation analysis detects novel candidate genes involved in follicular growth and oocyte maturation in a cohort of premature ovarian failure cases. *Human Reproduction*, *31*(8), 1913–1925.

Pervjakova, N., Kasela, S., Morris, A. P., **Kals, M.**, Metspalu, A., Lindgren, C. M., … Mägi, R. (2016). Imprinted genes and imprinting control regions show predominant intermediate methylation in adult somatic tissues. *Epigenomics*, *8*(6), 789–799.

Maasalu, K., Nikopensius, T., Kõks, S., Nõukas, M., **Kals, M.**, Prans, E., … Märtson, A. (2015). Whole-exome sequencing identifies de novo mutation in the COL1A1 gene to underlie the severe osteogenesis imperfecta. *Human Genomics*, *9*(1), 6.

Hagg, S., Fall, T., Ploner, A., Magi, R., Fischer, K., Draisma, H. H., …, **Kals, M.**, … Ingelsson, E. (2015). Adiposity as a cause of cardiovascular disease: a Mendelian randomization study. *International Journal of Epidemiology*, *44*(2), 578–586.

Fall, T., Hägg, S., Ploner, A., Mägi, R., Fischer, K., Draisma, H. H. M., …, **Kals, M.**, … ENGAGE Consortium. (2015). Age- and sex-specific causal effects of adiposity on cardiovascular risk factors. *Diabetes*, *64*(5), 1841–1852.

Putku, M., **Kals, M.**, Inno, R., Kasela, S., Org, E., Kožich, V., … Laan, M. (2015). CDH13 promoter SNPs with pleiotropic effect on cardiometabolic parameters represent methylation QTLs. *Human Genetics*, *134*(3), 291–303.

Haller, T., Kals, M., Esko, T., Magi, R., & Fischer, K. (2015). RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Briefings in Bioinformatics*, *16*(1), 39–44.

Bonder, M., Kasela, S., **Kals, M.**, Tamm, R., Lokk, K., Barragan, I., … Milani, L. (2014). Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genomics*, *15*(1), 860.

Vals, M.-A., Õiglane-Shlik, E., Nõukas, M., Shor, R., Peet, A., **Kals, M.**, … Õunap, K. (2014). Coffin–Siris Syndrome with obesity, macrocephaly, hepatomegaly and hyperinsulinism caused by a mutation in the ARID1B gene. *European Journal of Human Genetics*, *22*(11), 1327–1329.

Vaher, U., Nõukas, M., Nikopensius, T., **Kals, M.**, Annilo, T., Nelis, M., … Talvik, T. (2014). *De Novo SCN8A* Mutation Identified by Whole-Exome Sequencing in a Boy with Neonatal Epileptic Encephalopathy, Multiple Congenital Anomalies, and Movement Disorders. *Journal of Child Neurology*, *29*(12), NP202–NP206.

Vahi, P.-S., **Kals, M.**, Kõiv, L., & Braschinsky, M. (2014). Preoperative corticosteroid injections are associated with worse long-term outcome of surgical carpal tunnel release. *Acta Orthopaedica*, *85*(1), 102–106.

Nikopensius, T., Saag, M., Jagomägi, T., Annilo, T., **Kals, M.**, Kivistik, P. A., … Metspalu, A. (2013). A Missense Mutation in *DUSP6* is Associated with Class III Malocclusion. *Journal of Dental Research*, *92*(10), 893–898.

Ivanov, M., **Kals, M.**, Kacevska, M., Barragan, I., Kasuga, K., Rane, A., … Ingelman-Sundberg, M. (2013). Ontogeny, distribution and potential roles of 5-hydroxymethylcytosine in human liver function. *Genome Biology*, *14*(8), R83.

Fall, T., Hägg, S., Mägi, R., Ploner, A., Fischer, K., Horikoshi, M., …, **Kals, M.**, … consortium, for the E. N. for G. and G. E. (ENGAGE). (2013). The Role of Adiposity in Cardiometabolic Traits: A Mendelian Randomization Analysis. *PLoS Medicine*, *10*(6), e1001474.

Nikopensius, T., Annilo, T., Jagomägi, T., Gilissen, C., **Kals, M.**, Krjutškov, K., … Metspalu, A. (2013). Non-syndromic Tooth Agenesis Associated with a Nonsense Mutation in Ectodysplasin-A *(EDA)*. *Journal of Dental Research*, *92*(6), 507–511.

Ivanov, M., **Kals, M.**, Kacevska, M., Metspalu, A., Ingelman-Sundberg, M., & Milani, L. (2013). In-solution hybrid capture of bisulfite-converted DNA for targeted bisulfite sequencing of 174 ADME genes. *Nucleic Acids Research*, *41*(6), e72–e72.

McQuillan, R., Eklund, N., Pirastu, N., Kuningas, M., McEvoy, B. P., Esko, T., …, **Kals, M.**, … Consortium, on behalf of the Roh. (2012). Evidence of Inbreeding Depression on Human Height. *PLoS Genetics*, *8*(7), e1002655.

Uusküla, A., McMahon, J. M., **Kals, M.**, Talu, A., Abel-Ollo, K., Rüütel, K., & Jarlais, D. C. Des. (2013). Risk for Heterosexual HIV Transmission Among Non-Injecting Female Partners of Injection Drug Users in Estonia. *AIDS and Behavior*, *17*(3), 879–888.

Uuskula, A., **Kals, M.**, & McNutt, L.-A. (2011). Assessing non-response to a mailed health survey including self-collection of biological material. *The European Journal of Public Health*, *21*(4), 538–542.

Uusküla, A., Des Jarlais, D. C., **Kals, M.**, Rüütel, K., Abel-Ollo, K., Talu, A., & Sobolev, I. (2011). Expanded syringe exchange programs and reduced HIV infection among new injection drug users in Tallinn, Estonia. *BMC Public Health*, *11*(1), 517.

Uusküla, A., **Kals, M.**, Kosenkranius, L., McNutt, L.-A., & DeHovitz J, J. (2010). Population-based type-specific prevalence of high-risk human papillomavirus infection in Estonia. *BMC Infectious Diseases*, *10*(1), 63.

Braschinsky, M., Zopp, I., **Kals, M.**, Haldre, S., & Gross-Paju, K. (2010). Bladder dysfunction in hereditary spastic paraplegia: what to expect? *Journal of Neurology, Neurosurgery, and Psychiatry*, *81*(3), 263–266.

Kaur, S., Zilmer, K., Kairane, C., **Kals, M.**, & Zilmer, M. (2008). Clear differences in adiponectin level and glutathione redox status revealed in obese and normal-weight patients with psoriasis. *British Journal of Dermatology*, *159*(6), 1364–1367.

Uusküla, A., **Kals, M.**, Denks, K., Nurm, Ü., Kasesalu, L., DeHovitz, J., & McNutt, L. A. (2008). The prevalence of chlamydial infection in Estonia: a population-based survey. *International Journal of STD & AIDS*, *19*(7), 455–458.

Uuskula, A., **Kals, M.**, Rajaleid, K., Abel, K., Talu, A., Ruutel, K., … Des Jarlais, D. (2008). High-prevalence and high-estimated incidence of HIV infection among new injecting drug users in Estonia: need for large scale prevention programs. *Journal of Public Health*, *30*(2), 119–125.

Kals, J., Starkopf, J., Zilmer, M., Pruler, T., Pulges, K., Hallaste, M., …, **Kals, M.**, … Soomets, U. (2008). Antioxidant UPF1 attenuates myocardial stunning in isolated rat hearts. *International Journal of Cardiology*, *125*(1), 133–135.

Aavik, A., Lieberg, J., Kals, J., Pulges, A., **Kals, M.**, & Lepner, U. (2008). Ten Years Experience of Treating Aorto-Femoral Bypass Graft Infection with Venous Allografts. *European Journal of Vascular and Endovascular Surgery*, *36*(4), 432–437.

Kals, J., Kampus, P., **Kals, M.**, Pulges, A., Teesalu, R., Zilmer, K., … Zilmer, M. (2008). Inflammation and oxidative stress are associated differently with endothelial function and arterial stiffness in healthy subjects and in patients with atherosclerosis. *Scandinavian Journal of Clinical and Laboratory Investigation*, *68*(7), 594–601.

Kals, J., Kampus, P., **Kals, M.**, Teesalu, R., Zilmer, K., Pulges, A., & Zilmer, M. (2007). Arterial elasticity is associated with endothelial vasodilatory function and asymmetric dimethylarginine level in healthy subjects. *Scandinavian Journal of Clinical and Laboratory Investigation*, *67*(5), 536–544.

Kals, J., Kampus, P., **Kals, M.**, Pulges, A., Teesalu, R., & Zilmer, M. (2006). Effects of stimulation of nitric oxide synthesis on large artery stiffness in patients with peripheral arterial disease. *Atherosclerosis*, *185*(2), 368–374.

Kals, J., Kampus, P., **Kals, M.**, Zilmer, K., Kullisaar, T., Teesalu, R., … Zilmer, M. (2006). Impact of Oxidative Stress on Arterial Elasticity in Patients with Atherosclerosis. *American Journal of Hypertension*, *19*(9), 902–908.

* These authors contributed equally to this work // Antud autorid panustasid võrdselt.

ELULOOKIRJELDUS

Nimi: Mart Kals
Sünniaeg: 10. märts 1981
Kodakondsus: Eesti
Aadress: Eesti geenivaramu teaduskeskus, genoomika instituut, Tartu Ülikool
Riia 23B, 51010, Tartu, Eesti
Telefon: (+372) 737 4041
E-mail: mart.kals@ut.ee

Haridus:

2010–2018	Tartu Ülikool, matemaatilise statistika doktoriõpe
2004–2005	Limburgi Ülikool, Belgia, biostatistika magistriõpe, MSc biostatistika erialal
1999–2003	Tartu Ülikool, matemaatika bakalaureuseõpe, BSc matemaatilise statistika erialal
1996–1999	Hugo Treffneri gümnaasium

Teenistuskäik:

2010–	Eesti geenivaramu teaduskeskus, genoomika instituut, Tartu Ülikool, spetsialist
2007–2009	Astra Export & Trading AB, Eesti filiaal, müügianalüütik
2005–2007	Resta Ltd., konsultant

Juhendatud väitekirjad:

2016 Madli Tamme magistritöö „Kasvajaseoseliste mutatsioonide määramine kopsuvähiga uuritavate rakuvabast DNA-st“ kaasjuhendaja (MSc geenitehnoloogia erialal)

2011 Silva Kasela bakalaureusetöö „Ülegenoomne assotsiatsiooniuuring ja selle praktiline läbiviimine TÜ Eesti Geenivaramu andmete põhjal“ kaasjuhendaja (BSc matemaatilise statistika erialal)

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. **Mati Heinloo.** The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu, 1991, 23 p.
2. **Boris Komrakov.** Primitive actions and the Sophus Lie problem. Tartu, 1991, 14 p.
3. **Jaak Heinloo.** Phenomenological (continuum) theory of turbulence. Tartu, 1992, 47 p.
4. **Ants Tauts.** Infinite formulae in intuitionistic logic of higher order. Tartu, 1992, 15 p.
5. **Tarmo Soomere.** Kinetic theory of Rossby waves. Tartu, 1992, 32 p.
6. **Jüri Majak.** Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu, 1992, 32 p.
7. **Ants Aasma.** Matrix transformations of summability and absolute summability fields of matrix methods. Tartu, 1993, 32 p.
8. **Helle Hein.** Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu, 1993, 28 p.
9. **Toomas Kiho.** Study of optimality of iterated Lavrentiev method and its generalizations. Tartu, 1994, 23 p.
10. **Arne Kokk.** Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu, 1995, 165 p.
11. **Toomas Lepikult.** Automated calculation of dynamically loaded rigid-plastic structures. Tartu, 1995, 93 p, (in Russian).
12. **Sander Hannus.** Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74 p, (in Russian).
13. **Sergei Tupailo.** Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu, 1996, 134 p.
14. **Enno Saks.** Analysis and optimization of elastic-plastic shafts in torsion. Tartu, 1996, 96 p.
15. **Valdis Laan.** Pullbacks and flatness properties of acts. Tartu, 1999, 90 p.
16. **Märt Põldvere.** Subspaces of Banach spaces having Phelps' uniqueness property. Tartu, 1999, 74 p.
17. **Jelena Ausekle.** Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu, 1999, 72 p.
18. **Krista Fischer.** Structural mean models for analyzing the effect of compliance in clinical trials. Tartu, 1999, 124 p.
19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
20. **Jüri Lember.** Consistency of empirical k-centres. Tartu, 1999, 148 p.
21. **Ella Puman.** Optimization of plastic conical shells. Tartu, 2000, 102 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.

23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
25. **Maria Zeltser.** Investigation of double sequence spaces by soft and hard analitical methods. Tartu, 2001, 154 p.
26. **Ernst Tungel.** Optimization of plastic spherical shells. Tartu, 2001, 90 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 p.
28. **Rainis Haller.** *M(r,s)*-inequalities. Tartu, 2002, 78 p.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
30. Töö kaitsmata.
31. **Mart Abel.** Structure of Gelfand-Mazur algebras. Tartu, 2003. 94 p.
32. **Vladimir Kuchmei.** Affine completeness of some ockham algebras. Tartu, 2003. 100 p.
33. **Olga Dunajeva.** Asymptotic matrix methods in statistical inference problems. Tartu 2003. 78 p.
34. **Mare Tarang.** Stability of the spline collocation method for volterra integro-differential equations. Tartu 2004. 90 p.
35. **Tatjana Nahtman.** Permutation invariance and reparameterizations in linear models. Tartu 2004. 91 p.
36. **Märt Möls.** Linear mixed models with equivalent predictors. Tartu 2004. 70 p.
37. **Kristiina Hakk.** Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 137 p.
38. **Meelis Käärik.** Fitting sets to probability distributions. Tartu 2005, 90 p.
39. **Inga Parts.** Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140 p.
40. **Natalia Saealle.** Convergence and summability with speed of functional series. Tartu 2005, 91 p.
41. **Tanel Kaart.** The reliability of linear mixed models in genetic studies. Tartu 2006, 124 p.
42. **Kadre Torn.** Shear and bending response of inelastic structures to dynamic load. Tartu 2006, 142 p.
43. **Kristel Mikkor.** Uniform factorisation for compact subsets of Banach spaces of operators. Tartu 2006, 72 p.
44. **Darja Saveljeva.** Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 117 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
46. **Annely Mürk.** Optimization of inelastic plates with cracks. Tartu 2006. 137 p.
47. **Annemai Raidjõe.** Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97 p.
48. **Olga Panova.** Real Gelfand-Mazur algebras. Tartu 2006, 82 p.

49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
50. **Margus Pihlak.** Approximation of multivariate distribution functions. Tartu 2007, 82 p.
51. **Ene Käärik.** Handling dropouts in repeated measurements using copulas. Tartu 2007, 99 p.
52. **Artur Sepp.** Affine models in mathematical finance: an analytical approach. Tartu 2007, 147 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
54. **Kaja Sõstra.** Restriction estimator for domains. Tartu 2007, 104 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
57. **Evely Leetma.** Solution of smoothing problems with obstacles. Tartu 2009, 81 p.
58. **Ants Kaasik.** Estimating ruin probabilities in the Cramér-Lundberg model with heavy-tailed claims. Tartu 2009, 139 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
60. **Indrek Zolk.** The commuting bounded approximation property of Banach spaces. Tartu 2010, 107 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
63. **Marek Kolk.** Piecewise Polynomial Collocation for Volterra Integral Equations with Singularities. Tartu 2010, 134 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
65. **Larissa Roots.** Free vibrations of stepped cylindrical shells containing cracks. Tartu 2010, 94 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
68. **Olga Liivapuu.** Graded q-differential algebras and algebraic models in noncommutative geometry. Tartu 2011, 112 p.
69. **Aleksei Lissitsin.** Convex approximation properties of Banach spaces. Tartu 2011, 107 p.
70. **Lauri Tart.** Morita equivalence of partially ordered semigroups. Tartu 2011, 101 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.

72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.
74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
75. **Nadežda Bazunova.** Differential calculus $d^3 = 0$ on binary and ternary associative algebras. Tartu 2011, 99 p.
76. **Natalja Lepik.** Estimation of domains under restrictions built upon generalized regression and synthetic estimators. Tartu 2011, 133 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
80. **Marje Johanson.** $M(r, s)$-ideals of compact operators. Tartu 2012, 103 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
82. **Vitali Retšnoi.** Vector fields and Lie group representations. Tartu 2012, 108 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
85. **Erge Ideon**. Rational spline collocation for boundary value problems. Tartu, 2013, 111 p.
86. **Esta Kägo.** Natural vibrations of elastic stepped plates with cracks. Tartu, 2013, 114 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
88. **Boriss Vlassov.** Optimization of stepped plates in the case of smooth yield surfaces. Tartu, 2013, 104 p.
89. **Elina Safiulina.** Parallel and semiparallel space-like submanifolds of low dimension in pseudo-Euclidean space. Tartu, 2013, 85 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
93. **Kerli Orav-Puurand.** Central Part Interpolation Schemes for Weakly Singular Integral Equations. Tartu, 2014, 109 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multiparty computation. Tartu, 2015, 201 p.

95. **Kaido Lätt.** Singular fractional differential equations and cordial Volterra integral operators. Tartu, 2015, 93 p.
96. **Oleg Košik.** Categorical equivalence in algebra. Tartu, 2015, 84 p.
97. **Kati Ain.** Compactness and null sequences defined by ℓ_p spaces. Tartu, 2015, 90 p.
98. **Helle Hallik.** Rational spline histopolation. Tartu, 2015, 100 p.
99. **Johann Langemets.** Geometrical structure in diameter 2 Banach spaces. Tartu, 2015, 132 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
105. **Md Raknuzzaman.** Noncommutative Galois Extension Approach to Ternary Grassmann Algebra and Graded q-Differential Algebra. Tartu, 2016, 110 p.
106. **Alexander Liyvapuu.** Natural vibrations of elastic stepped arches with cracks. Tartu, 2016, 110 p.
107. **Julia Polikarpus.** Elastic plastic analysis and optimization of axisymmetric plates. Tartu, 2016, 114 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.
113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
115. **Tiina Kraav.** Stability of elastic stepped beams with cracks. Tartu, 2017, 126 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.

117. **Silja Veidenberg.** Lifting bounded approximation properties from Banach spaces to their dual spaces. Tartu, 2017, 112 p.
118. **Liivika Tee.** Stochastic Chain-Ladder Methods in Non-Life Insurance. Tartu, 2017, 110 p.
119. **Ülo Reimaa.** Non-unital Morita equivalence in a bicategorical setting. Tartu, 2017, 86 p.
120. **Rauni Lillemets.** Generating Systems of Sets and Sequences. Tartu, 2017, 181 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.
123. **Kaur Lumiste.** Improving accuracy of survey estimators by using auxiliary information in data collection and estimation stages. Tartu, 2018, 112 p.
124. **Paul Tammo.** Closed maximal regular one-sided ideals in topological algebras. Tartu, 2018, 112 p.