

MAARJA JÕELOO

Advances in microarray-based  
copy number variation discovery  
and phenotypic associations





**MAARJA JÕELOO**

Advances in microarray-based copy number  
variation discovery and phenotypic  
associations



UNIVERSITY OF TARTU  
Press

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

This dissertation is accepted for the commencement of the degree of Doctor of Philosophy in Gene Technology on April 29, 2024, by the Council of the Institute of Genomics and the Institute of Molecular and Cell Biology, University of Tartu.

Supervisors: Prof. Reedik Mägi, PhD  
Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia

Prof. Andres Metspalu, MD, PhD  
Estonian Genome Centre, Institute of Genomics, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

Prof. Andres Salumets, PhD  
Department of Obstetrics and Gynaecology, Institute of Clinical Medicine, University of Tartu, Tartu, Estonia

Prof. Andrew P. Morris, PhD  
Division of Musculoskeletal and Dermatological Sciences, The University of Manchester, Manchester, United Kingdom

Reviewer: Assoc. Prof. Age Brauer, PhD  
Chair of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

Opponent: Elliott G. Rees, PhD  
Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, United Kingdom

Commencement: Room No. 105, 23B Riia St., Tartu, on May 31, 2024, at 10:15am.

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

This research was funded by the European Union through the European Regional Development Fund Project No. 2014–2020.4.01.16-0125; the Estonian Research Council grants PRG687 and PRG1911; the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016775; the Archimedes Foundation scholarship in smart specialisation growth areas; Dora Plus PhD student mobility scholarship; and Kristjan Jaak scholarship for short-time visit. Data analyses were carried out in part in the High-Performance Computing Center of the University of Tartu. The thesis was partly written at writing retreats and writing days at the Institute of Genomics.

ISSN 1024-6479 (print)

ISBN 978-9916-27-527-6 (print)

ISSN 2806-2140 (pdf)

ISBN 978-9916-27-528-3 (pdf)

Copyright: Maarja Jõeloo, 2024

University of Tartu Press

www.tyk.ee



European Union  
European Regional  
Development Fund



Investing  
in your future

*“Anything worth doing is worth doing slowly.”*

Mae West



## TABLE OF CONTENTS

|                                                                                                        |    |
|--------------------------------------------------------------------------------------------------------|----|
| LIST OF ORIGINAL PUBLICATIONS .....                                                                    | 9  |
| LIST OF ABBREVIATIONS .....                                                                            | 11 |
| INTRODUCTION.....                                                                                      | 13 |
| 1. REVIEW OF THE LITERATURE.....                                                                       | 14 |
| 1.1. Overview of copy number variations (CNVs).....                                                    | 14 |
| 1.1.1. CNV categorisation and formation .....                                                          | 14 |
| 1.1.2. CNV distribution and abundance in human genome .....                                            | 15 |
| 1.1.3. The mechanisms of CNV effects on phenotype .....                                                | 17 |
| 1.1.4. CNV detection approaches .....                                                                  | 17 |
| 1.2. CNV discovery from SNP microarrays .....                                                          | 19 |
| 1.2.1. SNP microarray signal metrics and algorithms for copy<br>number inference .....                 | 20 |
| 1.2.2. Limitations of SNP microarray-based CNV discovery .....                                         | 22 |
| 1.2.3. Quality control and filtering of microarray-based CNV calls.                                    | 23 |
| 1.3. CNV-phenotype associations .....                                                                  | 24 |
| 1.3.1. Pathogenic CNVs and genomic disorders (GD) .....                                                | 25 |
| 1.3.2. Genome-wide CNV association studies (CNV GWAS) in<br>population biobanks.....                   | 26 |
| 2. AIMS OF THE STUDY.....                                                                              | 29 |
| 3. RESULTS AND DISCUSSION .....                                                                        | 30 |
| 3.1. Omics-informed CNV quality score (Ref. I).....                                                    | 30 |
| 3.1.1. Description of cohort and data preparations .....                                               | 30 |
| 3.1.2. Calculation and comparison of omics-based CNV quality<br>metrics.....                           | 33 |
| 3.1.3. Estimation of false positive and false negative rate of<br>PennCNV .....                        | 34 |
| 3.1.4. Omics-informed CNV quality score (OQS) modelling and<br>validation .....                        | 35 |
| 3.1.5. Incorporating CNV quality in association studies .....                                          | 36 |
| 3.2. Association between 16p11.2 CNV and reproductive traits<br>(Ref. II).....                         | 38 |
| 3.2.1. Description of cohorts and methods .....                                                        | 39 |
| 3.2.2. Association between 16p11.2 BP4-BP5 dosage and<br>pubertal timing.....                          | 41 |
| 3.2.3. Disorders of reproductive health and sexual development of<br>16p11.2 BP4-BP5 CNV carriers..... | 41 |
| 3.2.4. Causal inference between 16p11.2 region gene expression<br>and age at menarche.....             | 42 |

|                                                                                                     |     |
|-----------------------------------------------------------------------------------------------------|-----|
| 3.3. Maps of medically relevant CNV-phenotype associations and driver genes (Ref. III, IV, V) ..... | 43  |
| 3.3.1. Description of cohorts and methods .....                                                     | 44  |
| 3.3.2. Discovery and replication of CNV associations with continuous traits (Ref. III) .....        | 48  |
| 3.3.3. Discovery and replication of CNV associations with common diseases (Ref. IV).....            | 51  |
| 3.3.4. The map of dosage sensitive (DS) genes across human genome (Ref. V) .....                    | 54  |
| CONCLUSIONS .....                                                                                   | 56  |
| SUMMARY IN ESTONIAN .....                                                                           | 57  |
| REFERENCES .....                                                                                    | 60  |
| ACKNOWLEDGMENTS.....                                                                                | 69  |
| PUBLICATIONS .....                                                                                  | 71  |
| CURRICULUM VITAE .....                                                                              | 182 |
| ELULOOKIRJELDUS.....                                                                                | 185 |

## 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. V):

- I **Lepamets M**, Auwerx C, Nõukas M, Claringbould A, Porcu E, Kals M, Jürgenson T, Estonian Biobank Research Team, Morris AP, Võsa U, Bochud M, Stringhini S, Wijmenga C, Franke L, Peterson H, Vilo J, Lepik K, Mägi R, Kutalik Z. 2022. **Omics-informed CNV calls reduce false-positive rates and improve power for CNV-trait associations.** *HGG Adv.* 3(4), 100133.
- II Männik K\*, Arbogast T\*, **Lepamets M\***, Lepik K, Pellaz A, Ademi H, Kupchinsky ZA, Ellegood J, Attanasio C, Messina A, Rotman S, Martin-Brevet S, Dubruc E, Chrast J, Lerch JP, Qiu LR, Laisk T, The 16p11.2 European Consortium, The Simons VIP Consortium, The eQTLGen Consortium, Henkelman RM, Jacquemont S, Herault Y, Lindgren CM, Peterson H, Stehle JC, Katsanis N, Kutalik Z, Nef S, Draganski B, Davis EE, Mägi R, Reymond A. 2019. **Leveraging biobank-scale rare and common variant analyses to identify *ASPHD1* as the main driver of reproductive traits in the 16p11.2 locus.** *bioRxiv*.
- III Auwerx C, **Lepamets M**, Sadler MC, Patxot M, Stojanov M, Baud D, Mägi R, Estonian Biobank Research Team, Porcu E, Reymond A, Kutalik Z. 2022. **The individual and global impact of copy-number variants on complex human traits.** *Am J Hum Genet.* 109(4), 647–668.
- IV Auwerx C, **Jõeloo M**, Sadler MC, Tesio N, Ojavee S, Clark CJ, Mägi R, Estonian Biobank Research Team, Reymond A, Kutalik Z. 2024. **Rare copy-number variants as modulators of common disease susceptibility.** *Genome Med.* 16(1), 5.
- V Collins RL, Glessner JT, Porcu E, **Lepamets M**, Brandon R, Lauricella C, Han L, Morley T, Niestroj L, Ulirsch J, Everett S, Howrigan DP, Boone PM, Fu J, Karczewski KJ, Kellaris G, Lowther C, Lucente D, Mohajeri K, Nõukas M, Nuttle X, Samocha KE, Trinh M, Ullah F, Võsa U, Epi25 Consortium, Estonian Biobank Research Team, Hurles ME, Aradhya S, Davis EE, Finucane H, Gusella JF, Janze A, Katsanis N, Matyakhina L, Neale BM, Sanders D, Warren S, Hodge JC, Lal D, Ruderfer DM, Meck J, Mägi R, Esko T, Reymond A, Kutalik Z, Hakonarson H, Sunyaev S, Brand H, Talkowski ME. 2022. **A cross-disorder dosage sensitivity map of the human genome.** *Cell.* 185(16), 3041–3055.e25.

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

My contributions to the listed publications were as follows:

- Ref. I** Participated in the study design, conducted CNV discovery and all the analyses in Estonian Biobank (EstBB), Lifelines Deep, and SkiPOGH datasets, conducted family-based CNV analysis in UK Biobank (UKBB), drafted and revised the manuscript.
- Ref. II** Participated in the *in silico* study design, conducted CNV discovery and analyses in EstBB and UKBB datasets, performed the initial Mendelian Randomisation analysis, participated in drafting and revising the manuscript.
- Ref. III** Conducted CNV discovery and replication analyses in EstBB, performed the CNV quality assessment, wrote the EstBB description, and revised the manuscript.
- Ref. IV** Conducted CNV discovery and replication analyses in EstBB, revised the manuscript.
- Ref. V** Conducted CNV discovery and data curation in EstBB, revised the manuscript.

## LIST OF ABBREVIATIONS

|                  |                                                                                      |
|------------------|--------------------------------------------------------------------------------------|
| AAM              | Age at menarche                                                                      |
| aCGH             | Array comparative genomic hybridisation                                              |
| AKI              | Acute kidney injury                                                                  |
| ANOVA            | Analysis of variance                                                                 |
| ASD              | Autism spectrum disorder                                                             |
| BAF              | B allele frequency                                                                   |
| BMI              | Body mass index                                                                      |
| bp               | Base pairs                                                                           |
| BP               | Breakpoint                                                                           |
| CGH              | Comparative genomic hybridisation                                                    |
| CKD              | Chronic kidney disease                                                               |
| CMT              | Charcot-Marie-Tooth (disease)                                                        |
| CNP              | Copy number polymorphism                                                             |
| CNV              | Copy number variation                                                                |
| cQS              | Consensus-based (CNV) quality score                                                  |
| DECIPHER         | Database of Chromosomal Imbalance and Phenotype in Humans<br>Using Ensembl Resources |
| DS               | Dosage sensitive                                                                     |
| eQTL             | Expression quantitative trait locus                                                  |
| EstBB            | Estonian Biobank                                                                     |
| EstBB-MO         | Multi-omics set of EstBB                                                             |
| FISH             | Fluorescence In Situ Hybridisation                                                   |
| FN               | False negative                                                                       |
| FP               | False positive                                                                       |
| GE               | Gene expression metric (for CNV quality)                                             |
| GD               | Genomic disorder                                                                     |
| gnomAD           | Genome Aggregation Database                                                          |
| GnRH3            | Gonadotropin-releasing hormone-3                                                     |
| GSA              | Global Screening Array                                                               |
| GTE <sub>x</sub> | Genotype-Tissue Expression (portal)                                                  |
| GWAS             | Genome-wide association study                                                        |
| HI               | Haploinsufficient                                                                    |
| HMM              | Hidden Markov Model                                                                  |
| ICD-10           | International Classification of Disease, 10th Revision                               |
| IED              | Intragenic exonic duplication                                                        |
| LD               | Linkage disequilibrium                                                               |
| LLDeep           | Lifelines Deep                                                                       |
| LOEUF            | Loss-of-function observed/expected upper bound fraction                              |
| LoF              | Loss-of-function                                                                     |
| LRR              | Log R ratio                                                                          |
| mCNV             | Multi-allelic CNV                                                                    |
| MET              | Methylation metric (for CNV quality)                                                 |

|         |                                                |
|---------|------------------------------------------------|
| MR      | Mendelian Randomisation                        |
| MSE     | Mean square error                              |
| MZ      | Mono-zygotic                                   |
| NAHR    | Nonallelic homologous recombination            |
| NHEJ    | Nonhomologous end-joining                      |
| OMIM    | Online Mendelian Inheritance in Man (database) |
| OQS     | Omics-informed (CNV) quality score             |
| OR      | Odds ratio                                     |
| PC      | Principal component                            |
| PFB     | Population frequency of B allele               |
| pHaplo  | Probability of haploinsufficiency (score)      |
| pTriplo | Probability of triplosensitivity (score)       |
| RD      | Read depth                                     |
| RNAseq  | RNA sequencing                                 |
| RP      | Read pair                                      |
| SCr     | Serum creatinine                               |
| SD      | Standard deviation                             |
| SkiPOGH | Swiss Kidney Project on Genes in Hypertension  |
| SNP     | Single nucleotide polymorphism                 |
| SNV     | Single nucleotide variation                    |
| SR      | Split read                                     |
| SV      | Structural variant                             |
| TS      | Triplosensitive                                |
| UKBB    | UK Biobank                                     |
| WES     | Whole-exome sequencing                         |
| WF      | Waviness factor                                |
| WGS     | Whole-genome sequencing                        |
| WHR     | Waist-to-hip ratio                             |

# INTRODUCTION

Copy number variations (CNVs) are deletions and duplications of genomic segments. They were first discovered in the context of developmental delay and neuropsychiatric disorders, where they exhibited severe pathogenic effects. The current knowledge indicates, however, that CNVs are abundant in the genomes of both healthy and diseased individuals. In fact, it is suggested that CNVs affect a substantially larger fraction of a person's genome than single nucleotide variants (SNVs). Nevertheless, the majority of studies aiming to discover the genetic basis of traits predominantly prioritise the investigation of SNVs. This is due to methodological challenges originating from the process of CNV discovery, quality control, and association testing. Still, owing to their propensity to encompass or disrupt genes and regulatory elements, CNVs are believed to have a more significant impact on phenotype compared to SNVs, making CNV studies pivotal for comprehending the complex genetic architecture underlying human traits and diseases.

This thesis aims to advance our understanding of CNVs and their effect on clinically relevant phenotypes. It is divided into two main sections – a literature overview and an experimental part. In the literature part, I introduce the spectrum of CNVs in the human genome. More specifically, I describe how they are categorised, formed, and distributed across the chromosomes, and what are the mechanisms underlying their effects on phenotype. A brief overview is provided for all major CNV detection approaches. However, given the focus of this thesis on CNVs detected from microarray data, an entire subsection is dedicated to the characterisation of this means of CNV discovery. Additionally, the current state of rare CNV-phenotype association studies is described to the best of my knowledge. In the experimental part of the thesis, I present the results from five original publications. First, I present an approach of deriving and employing an omics-informed CNV quality score, which leads to higher statistical power in CNV association studies. Secondly, I zoom in on one well-described pathological CNV and characterise its novel association to pubertal timing. Finally, I elaborate on three collaborative efforts aimed at creating extensive maps for CNV-phenotype associations and dosage dependent genes. Altogether, this thesis strives to improve all three key steps of detecting CNV effects on phenotypes: the pre-analysis quality control, the discovery of associations, and the follow-up studies of potentially causal mechanisms.

# 1. REVIEW OF THE LITERATURE

## 1.1. Overview of copy number variations (CNVs)

### 1.1.1. CNV categorisation and formation

Genetic variants are categorised into two primary classes: ‘small genetic variations’ including single nucleotide variations (SNVs) and short insertions/deletions (i.e., indels), and structural variations (SVs) which refer to alterations in the structure of the genome encompassing larger DNA segments. SV size is typically defined as ranging from 50 base pairs (bp) to entire chromosomes. SVs can overlap genes, regulatory regions, and other functional elements, leading to a spectrum of consequences, from benign polymorphisms to disease susceptibility.

Copy number variations (CNVs) are a specific type of SVs that involve changes in the number of copies of DNA segments. CNVs are divided into two variant classes: deletions (a decrease in the number of copies) and duplications (an increase in the number of copies) (Figure 1). Since CNVs result in net gain or loss of genomic content, they are frequently referred to as ‘unbalanced’ SVs, as opposed to dosage-neutral variants like inversions and translocations which are referred to as ‘balanced’ SVs. Common (>1% frequency) CNVs are sometimes called copy number polymorphisms (CNP), similarly to common SNVs, which are known as single nucleotide polymorphisms (SNP).

CNVs can be categorised as biallelic or multi-allelic. Biallelic CNVs have two primary allelic states, the presence and the absence of a particular deletion or duplication. They can be represented similarly to SNVs and are easier to detect and analyse due to their binary nature. Like SNVs, biallelic CNVs can be heterozygous or homozygous, depending on whether just one or both chromosomal alleles have undergone a deletion or duplication. Multi-allelic CNVs (mCNVs, Figure 1) are variations where there are multiple possible copy number states for a particular genomic segment. For example, two plasma protein coding genes *HPR* and *ORM1* frequently (>20%) show higher copy numbers (up to 8 and 13 copies, respectively) in certain population groups (Handsaker et al., 2015).

Based on their location and frequency CNVs can be classified as either recurrent or sporadic. Recurrent CNVs reappear independently with consistent break-point regions, leading to a higher CNV frequency in the population. Sporadic CNVs, on the other hand, arise spontaneously across the genome and consequently exhibit lower frequency. These two categories also have distinct prevalent formation patterns. The most frequent mechanism for *de novo* formation of recurrent CNVs is nonallelic homologous recombination (NAHR) during meiosis. This means that two genomic regions that are close to identical but not alleles, such as low-copy repeats or segmental duplications, misalign and the subsequent process of DNA replication leads to deletion or duplication of genomic content (Lupski and Stankiewicz, 2005). Analogous sources of misalignment during homologous recombination and CNV formation are variable number tandem repeats (Conrad et al., 2010). Additionally, nonrecurrent or sporadic CNVs

can form during nonhomologous end-joining (NHEJ) which is a reparation mechanism for double-stranded breaks in DNA and which can lead to loss or gain of genetic material due to inaccurate ligation (Lupski and Stankiewicz, 2005). Besides *de novo* formation, CNVs can also appear somatically via similar mechanisms during cell division or DNA damage. This process is particularly common in cancer development (Beroukhi et al., 2010).

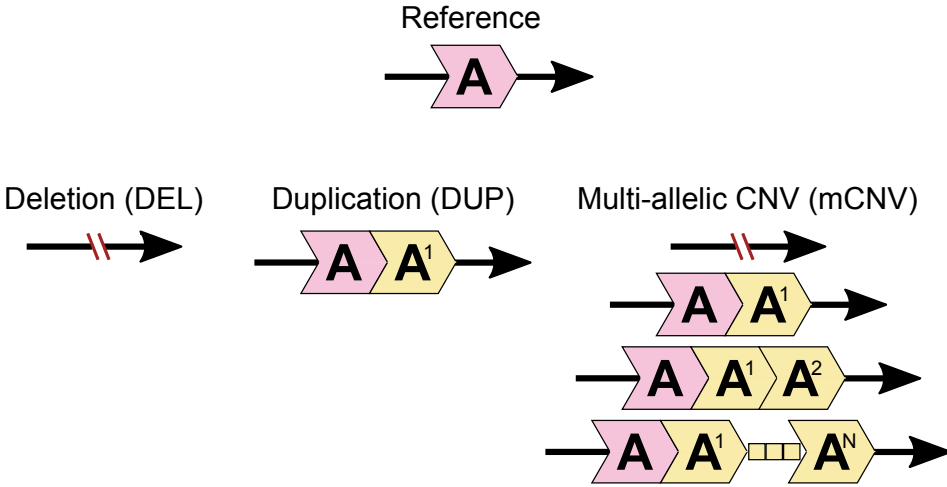


**Figure 1.** Compared to reference, deletions (DEL) decrease and duplications (DUP) increase the copy number of a particular genomic segment (denoted with A). Multi-allelic CNVs (mCNV) exhibit multiple possible copy number states per segment (Inspired by Collins et al. (2020), Figure 1A).

### 1.1.2. CNV distribution and abundance in human genome

Several genome-wide SV and CNV maps have been compiled to quantify the distribution and abundance of CNVs in the human genome (Abel et al., 2020; Almarri et al., 2020; Byrska-Bishop et al., 2022; Collins et al., 2020; Sudmant et al., 2015). Most evaluations so far have been performed in short-read whole-genome sequencing (WGS) studies. Based on these, it has been estimated that 3.7–9.5% of the genome, depending on the variant stringency level and sample diversity, is copy number variable (Conrad et al., 2010; Zarrei et al., 2015). Sudmant et al. (2015) have evaluated that each individual harbours a median of 18.4 Mbp of SVs, out of which a vast majority are CNVs (11.3 Mbp mCNVs, 5.6 Mbp biallelic deletions and 0.5 Mbp biallelic duplications). However, later studies suggest these numbers are actually much higher (Byrska-Bishop et al., 2022). Importantly, CNVs alter a substantially larger number of nucleotides per genome than SNVs (Sudmant et al., 2015).

While exact CNV count and size distribution in studies vary due to differences in CNV definition, methodology, and sample sets, all agree that deletions are on average shorter but detected more frequently compared to duplications. The

median size for deletions detectable from WGS reads is 756 bp, while for duplications it is 4,283 bp (Collins et al., 2020). Both indicate that the majority of CNVs are short (<10 kbp).

Average CNV count per sample in large high-coverage (~30X or more) short-read WGS studies varies in ranges ~1,550–4,781 for deletions, ~490–2,522 for duplications and 425–1,356 for mCNVs, with the highest CNV count per sample found in African ancestry populations (Abel et al., 2020; Byrska-Bishop et al., 2022; Collins et al., 2020; Handsaker et al., 2015), a result aligning with general knowledge about greater genetic diversity within African populations in contrast to non-African populations (Campbell and Tishkoff, 2008). Moreover, leveraging recent long-read WGS technologies for CNV detection have raised the CNV yield per genome up to 25,000, with nearly 70% of the calls not detectable from short reads (Ebert et al., 2021).

A vast majority of CNVs are rare. To date the largest CNV studies show that 91–92% of deletions and 87–95% of duplications have frequency <1% and roughly half of the variants (51–60% of deletions and 47–60% of duplications) detected in recent studies are singletons or found only in a group of close relatives (Abel et al., 2020; Collins et al., 2020). Evidently, CNV size and frequency are negatively correlated (Collins et al., 2020), indicating a more pathogenic effect and negative selection of large CNVs. Still, despite being individually rare, large CNVs are collectively common with over 5% of individuals carrying CNVs >500 kbp and over 1% carrying CNVs >1 Mbp (Itsara et al., 2009).

Common CNVs, defined as CNPs, are often short and exhibit behaviour analogous to SNPs. A significant number of common deletions are in high linkage disequilibrium (LD) with nearby SNPs/indels (median peak  $r^2 > 0.8$ ), while for biallelic duplications the LD is lower with median peak  $r^2 = 0.52$  (Sudmant et al., 2015). This can be due to duplicated alleles being dispersed across genome instead of being in tandem (Conrad et al., 2010).

CNVs have an uneven distribution across the genome. As previously stated, this is influenced by the repetitive sequence context of a particular genomic region. It has also been shown that biallelic deletions and duplication are more frequent near telomeres, while mCNVs are enriched in centromeric regions (Collins et al., 2020). About 3,163 genomic regions have been discovered where CNVs seem to cluster. These include 30 regions that exhibit an excess of CNVs (>4 standard deviations (SD)), indicating loci corresponding to more ‘fragile’ sites (Sudmant et al., 2015). Furthermore, it has been shown that chromosomes 22 and Y show a higher frequency of duplications, while chromosomes 19, 22, and Y display a greater number of deletions when compared to the remainder of the genome (Zarrei et al., 2015). Considering that CNVs often fall into genomic regions that are more challenging to analyse, the actual CNV abundance is likely underestimated in most studies.

### 1.1.3. The mechanisms of CNV effects on phenotype

Varied in their size and location, CNVs possess the capacity to intersect genes, regulatory domains, and other functional elements, leading to a spectrum of outcomes. The effect on human phenotype can arise through diverse molecular mechanisms, including alteration of gene dosage, disruption of coding or regulatory sequence, unmasking recessive alleles, gene fusion and others (Zhang et al., 2009).

Deletions that intersect with genes can result in the loss of entire or partial coding sequence. Similarly, duplications can disrupt the gene and cause frame-shifts. Both are considered as predicted loss-of-function (LoF) mutations. Besides LoF, duplications can cause complete copy gain of genes or a subset of exons (i.e., intragenic exonic duplications or IED). In total, different studies have estimated that SVs alter approximately 162–189 genes per individual, with the majority (~60%) of the variants being LoF mutations (Byrska-Bishop et al., 2022; Collins et al., 2020; Ebert et al., 2021). This suggests that over 25% of all rare protein-truncating mutations are SVs (Collins et al., 2020).

The CNVs that overlap the sequence of dosage sensitive (DS) genes are typically large and rare (Itsara et al., 2009). These variants can lead to haploinsufficiency – a phenomenon where a single operative gene copy is not sufficient for normal function – or triplosensitivity (i.e., duplication intolerance). Currently (2024/01), ClinGen Dosage Sensitivity Map contains 423 and 24 known haploinsufficient (HI) and triplosensitive (TS) genic regions, respectively (HI/TS Score of 3) (ClinGen Dosage Sensitivity, 2024). Interestingly, several genes exhibit bidirectional intolerance to dosage change (Ruderfer et al., 2016). DS genes are typically longer, have well-conserved coding sequence and promoters, and higher expression, especially during early development (Huang et al., 2010). Additionally, a more recent study has shown higher deletion intolerance in cancer-related genes, while duplication-specific constraint suggests strong intolerance to variation altering normal developmental pathways (Aguirre et al., 2019).

Hybrid or fusion genes are formed when two previously independent genes are merged into a chimeric structure. Most frequently, they are created by translocations, inversions and interstitial deletions with breakpoints in different genes (Panagopoulos and Heim, 2021). Fusion genes play a crucial role in cancer development by creating abnormal proteins (Yoshihara et al., 2015; Zhang et al., 2009). Among other fields, the importance of fusion genes is further characterised in the studies of colour-blindness (Deeb, 2005) and pharmacogenetics (Gaedigk et al., 2010).

### 1.1.4. CNV detection approaches

In the early days of genomic aberration detection, cytogenetic karyotyping was the visual examination approach to identify gains and losses of genomic content of 5 Mbp or larger (Dowjat and Włodarska, 1981; Pepler et al., 1968). This was further advanced by the appearance of Fluorescence In Situ Hybridisation (FISH),

which enabled a more focused visualisation of CNVs in specific chromosomal regions (Feuk et al., 2005; Ravnán et al., 2006). A significant breakthrough was achieved through Comparative Genomic Hybridisation (CGH), a method in which test and reference (e.g., tumour and normal) samples are labelled with different fluorescent dyes (green and red) and hybridised to metaphase chromosomes. The relative deficiency or excess of genomic material is detected based on the differences of fluorescent signal intensities between these samples (Kallioniemi et al., 1992). Deviations from the equal amounts of green and red fluorescence indicate chromosomal imbalances in the test sample compared to the reference sample. This was the earliest approach to enable genome-wide screening of CNVs. Still, its detection resolution resembled that of cytogenetic methods.

A notable upgrade to CGH was array CGH (aCGH) (Lucito et al., 2003). Similarly, CNVs are called by contrasting signal intensity between a test and a reference sample. However, the full metaphase chromosomes of CGH are replaced with up to millions of short (>25 bp) immobilised DNA probes as hybridisation targets (Theisen, 2008). The main advantages of aCGH over preceding methods were proportionally higher resolution and the ability to simultaneously perform high-throughput genome-wide CNV detection in larger sample sets. A study has shown that aCGH has an over 10% detection rate of clinically relevant chromosomal abnormalities and it was the first approach for the identification of previously undescribed CNVs as incidental findings with unclear clinical relevance (Shaffer et al., 2007).

In more recent developments, SNP microarrays have gained prominence over aCGH in the field of CNV detection from hybridisation intensities. As CNVs discovered from SNP arrays is the main topic of this thesis, the technology with its possibilities and limitations is thoroughly described in section 1.2.

Sequencing-based methods have distinct advantages over array-based approaches for detecting small CNVs and achieving higher resolution for determining CNV breakpoints. Unlike microarrays, sequencing also enables the simultaneous discovery of other SV types. There are four main SV detection concepts, all of which require the mapping of sequencing reads or contigs to the reference genome (Alkan et al., 2011; Kosugi et al., 2019). First, tools that focus solely on copy number estimation mainly use read depth (RD) methods (Abyzov et al., 2011; Klambauer et al., 2012). These algorithms assume that deletions and duplications show reduced and heightened read depth, respectively, when compared to diploid regions. Secondly, read pair (RP) approaches search for regions where there are inconsistencies between the reference genome and the span and/or orientation of the mapped read pairs. For example, read pairs that map too far apart indicate the presence of a deletion. Additionally, split read (SR) approaches focus on reads that span SV breakpoints and thus do not map continuously to the reference genome. Instead, in SV regions a part of the read aligns to one location, and another part to a different location. Finally, (local) *de novo* assembly approaches can be used to detect SVs by assembling the sequencing reads to contigs and subsequently aligning the contigs to the reference genome. Most of the widely-

used sequencing-based detection algorithms for short-read WGS use a combination of these concepts (Chen et al., 2016; Handsaker et al., 2011; Layer et al., 2014; Rausch et al., 2012). Still, different algorithms are optimised for different SV types and sizes. In order to increase the detection accuracy and sensitivity, it has been advised to use several tools for CNV/SV discovery and combine the results (Byrska-Bishop et al., 2022; Collins et al., 2020; Sudmant et al., 2015; Vialle et al., 2022). This approach has been employed by many large SV studies (Byrska-Bishop et al., 2022; Collins et al., 2020; Sudmant et al., 2015; Vialle et al., 2022).

Despite being an improvement on microarray-based methods, CNV/SV detection from short-read WGS has several constraints. RP, SR, and assembly approaches have limited accuracy in repetitive regions, while RD methods suffer from poor breakpoint precision (Alkan et al., 2011). To overcome these shortcomings, long-read WGS data can be leveraged for CNV/SV discovery. Typically, either *de novo* assembly (Wenger et al., 2019) or alignment-based approaches are used (Heller and Vingron, 2019; Jiang et al., 2020; Smolka et al., 2024). As long reads often span full SV regions, the latter can utilise both inter- and intra-read signatures. However, the high cost of long-read WGS remains the limiting factor for its wide-spread use.

Another group of sequencing-based algorithms is designed specifically for CNV detection from coding regions using whole-exome sequencing (WES) data. Those algorithms rely mostly on RD methods. However, the non-uniformity of read coverage and non-contiguous nature of sequenced regions, demands extra steps for denoising or removal of systematic biases (Babadi et al., 2023; Fromer et al., 2012; Klambauer et al., 2012; Krumm et al., 2012).

Besides common array- and sequencing-based approaches, CNV detection methods have been developed for a few unconventional data layers. Recently, a few alignment-free algorithms for CNV estimation from sequencing reads have been developed (Pajuste and Remm, 2023; Shen and Kidd, 2020). These methods utilise the abundance of short DNA fragments with fixed length (i.e.,  $k$ -mers) calculated directly from raw reads. Additionally, CNV detection approaches have been developed for methylation array (Feber et al., 2014; Marzouka et al., 2016) and RNA sequencing (RNAseq) data (Talevich and Hunter Shain, 2018).

## 1.2. CNV discovery from SNP microarrays

At the time of conducting most of the work for this thesis, microarrays were the most numerous datasets for CNV detection. As being the prime source of SNV information used in genetic association analyses, the already existing microarray intensity data were simply repurposed for CNV discovery in a cost- and time-efficient manner. Besides large-scale data availability, the improved genome coverage of SNP microarrays presented another great advancement over aCGH techniques by enabling CNV detection with  $\sim 1$  kilobase resolution. Although microarray-based methods have more recently given way to WGS-based

approaches due to the increasing number of WGS samples available, they are still frequently and efficiently used for rare CNV detection in large cohorts.

### 1.2.1. SNP microarray signal metrics and algorithms for copy number inference

Similarly to aCGH, SNP microarray-based CNV detection relies on the measurements of hybridisation intensities. There are two key aspects that differentiate between aCGH and SNP microarray data in the context of CNV detection (Alkan et al., 2011). First, only one (test) sample is used in the hybridisation process per microarray. Instead of using a reference sample, the signal is normalised against the expected intensities calculated based on a population of samples. For commercially available microarrays, the reference population is usually chosen by the manufacturer. The normalised signal, referred to as ‘log R ratio’ (LRR), is subsequently employed for CNV detection as in aCGH approach. Second, for each SNP included on the microarray, a pair of probes are used to capture allele-specific fluorescent signal. The relative ratio between the intensities of two alleles per SNP, referred to as ‘B allele frequency’ (BAF), can be incorporated to increase sensitivity of CNV discovery. Both metrics, as well as the signal changes in case a CNV is present, are illustrated in Figure 2.

Based on the underlying statistical framework that infers CNV regions from LRR and BAF values, most of the CNV discovery approaches are classified as either segmentation algorithms (often originally built for aCGH) or Hidden Markov models (HMM). Furthermore, some methods have been developed with certain genotyping platforms (such as Illumina or Affimetrix) in mind while others are more universal (Balagué-Dobón et al., 2022).

The most widely used tools for CNV detection from SNP microarray data use an HMM-based algorithm. This approach assumes that the observed (‘emitted’) intensity per SNP corresponds to a ‘hidden’ copy number state with an integer value (usually between 0 and 4). A transition from one state to another imposes a penalty leading to a sequence of states with longer stretches of consecutive probes with similar values. QuantiSNP (Colella et al., 2007) is one of the first developed software tools of this kind. It uses an Objective Bayes framework for setting prior state probabilities corresponding to a fixed false positive (type I error) rate and incorporates both LRR and BAF values for copy number state inference. PennCNV (Wang et al., 2007) improves the HMM algorithm by combining LRR/BAF values with population frequencies of B alleles (PFB) and SNP locations. Longer distance between neighbouring SNPs results in higher probability of having the copy number state change between them. Furthermore, PennCNV can integrate family information to update the state transition locations (CNV boundaries) in offspring. Both QuantiSNP and PennCNV tools are designed for Illumina and Affimetrix genotyping platforms.

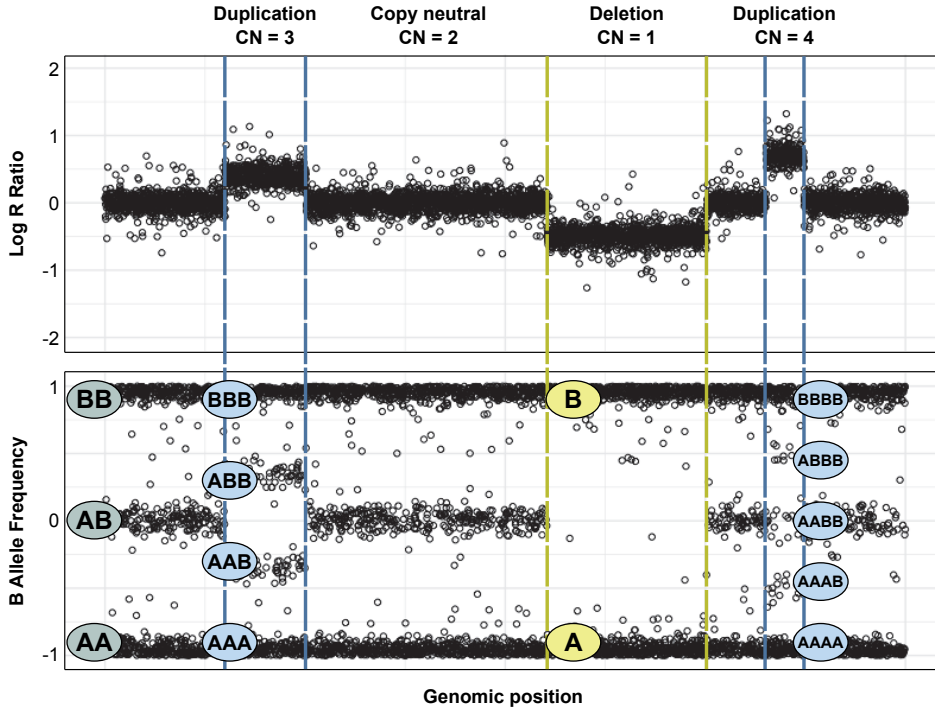


**Figure 2.** Log R ratio (LRR; upper plot) and B allele frequency (BAF; lower plot) signals in copy neutral, deletion (yellow vertical lines), and duplication (blue vertical lines) regions. For each copy variable region, all possible genotypes comprising A and B alleles, are presented. Abbreviations: CN, copy number.

Birdsuite (Korn et al., 2008) is a CNV detection approach optimised for the Affymetrix platform. It has separate modules for genotyping of known common CNV regions and detection of rare or *de novo* CNVs. Unlike in previous software, Birdsuite seeks to overcome the issues of heterogeneity of probe performance caused by different probe sets showing different intrinsic measurement variance and allele frequency spectrum across samples. First, those properties are modelled on the single SNP level and then included in an HMM algorithm together with LRR/BAF and PFB values to search for consistent evidence for CNVs across multiple neighbouring SNP sites. Birdsuite is also one of the few algorithms that combines CNV and SNP information to accurately genotype SNPs within CNV regions resulting in genotypes such as A (in deletions) or AAC (in duplications).

Recently, a haplotype-informed CNV detection algorithm HI-CNV has been developed for biobank-scale datasets (Hujoel et al., 2022). Its aim is to increase the power to detect CNVs carried by multiple individuals that have co-inherited the same CNV-containing haplotype from a common ancestor. First, HI-CNV identifies identity-by-descent DNA segments based on SNV data. Then, for each haplotype a set of “haplotype neighbours” are defined and CNVs are detected by

an HMM algorithm using the intensity data of all haplotype neighbours simultaneously. To our knowledge, this is the first approach to leverage the intensity data across multiple samples to aid in CNV detection.

Additionally, a few commercial CNV discovery software tools exist such as CNVPartition as part of Illumina’s GenomeStudio platform (Illumina GenomeStudio 2.0 Plug-ins, 2017), an analogous software for Affymetrix implemented in Affymetrix Array Power Tools (ThermoFisher Scientific Axiom CNV Summary Tools Software, 2015) and others. Moreover, as part of the iPSYCH Initiative (iPSYCH, 2022), a tool is developed for the CNV detection from amplified DNA of dried blood spots (iPsychCNV, 2017).

At the time of writing, PennCNV stands out as the most extensively used CNV detection algorithm for SNP microarray data (PubMed citations: PennCNV 1,011, QuantiSNP 334, Birdsuite 485; 12.02.2024). This popularity can be attributed to its user-friendliness, optimisation for different genotyping platforms, and commendable performance in terms of specificity, reliability, reproducibility, and mitigation of bias (Balagué-Dobón et al., 2022).

### 1.2.2. Limitations of SNP microarray-based CNV discovery

Although being an improvement over aCGH, CNV discovery from SNP microarray intensities still has several shortcomings. These mainly arise from the SNP distribution on the array or high variance and suboptimal signal-to-noise ratio of probe intensities.

First, as CNV detection requires consistent evidence of altered intensities across neighbouring SNPs, the resolution of the CNV calls depends on the number of good-quality SNPs overlapping the CNV region. It is not feasible to detect precise CNV breakpoints, and short CNVs are often overlooked due to inadequate SNP density of these regions on the microarray. For example, by default, PennCNV requires the intensities of at least three consecutive SNPs to confirm the presence of a CNV event (Wang et al., 2007). However, regions with frequent CNVs often lack sufficient coverage on SNP arrays (i.e., termed “unSNPable genome”) as a result of difficult assay design and implementation (Winchester et al., 2009).

Second, even for regions with good SNP microarray coverage, the signal intensities are normalised against a reference population which is implicitly assumed to be copy neutral (i.e., diploid). This assumption is evidently no longer applicable when considering regions that overlap with common CNV events (Winchester et al., 2009). Moreover, consistent discrepancies between the reference population and the sample of interest increase the signal variability estimates, and consequently impact on the reliability of CNV calls (Pinto et al., 2011). It has been stated that aCGH where only one reference sample is used for signal normalisation tends to have better signal-to-noise ratio compared to SNP arrays (Liu et al., 2019). To alleviate the problem of common CNVs and unSNPable regions, and provide a more uniform coverage of the genome, different microarray manufacturers have started to include specific probes on their SNP microarrays for the detection of CNVs with >5% frequency (Illumina Technical Notes, 2007).

Furthermore, microarray-based methods can usually only distinguish between copy number states ranging from 0 to 4. As the number of copies increases, the differences in signal intensities become subtler and the precise copy number detection becomes more challenging. Moreover, deletions are easier to detect compared to duplications as the former represents a 50% signal intensity reduction, while the latter only a 33% intensity increase (Balagué-Dobón et al., 2022).

All the previously mentioned aspects contribute to the relatively large false negative (FN) rate of SNP microarray-based CNV detection. Although, the exact number of detectable CNVs and FN rate depends on the array and dataset, the general yield appears to be an order of magnitude lower compared to WGS-based approaches (Winchester et al., 2009). To make matters worse, due to high variance and noise in probe-specific intensities, array-based CNV discovery also suffers from a high false positive (FP) rate. A study reveals that, depending on the array, the fraction of CNVs that can be validated using WGS reads range between 0% and 86%, with arrays containing the largest number of probes and extensive exonic coverage being least suitable for CNV discovery (Haraksingh et al., 2017). Another study incorporating three microarray-based detection algorithms – PennCNV, QuantiSNP and CNVPartition – shows that these tools only agree in approximately 20% of the cases (Macé et al., 2016). These findings indicate that strict quality control steps must be taken prior to further analyses.

### 1.2.3. Quality control and filtering of microarray-based CNV calls

While the measures to improve FN rates of CNV detection are limited, FP rates can be better quantified and controlled by quality control and filtering. This is especially important in large-scale CNV genome-wide association studies (CNV GWASs) where manual curation of the CNV data is not feasible. Various filtering approaches have been suggested. These are mainly based on arbitrarily set cut-off values of different CNV or sample-specific parameters, e.g., CNV length, number of calls per sample, metrics calculated based on intensity signals, etc. (Aguirre et al., 2019; Chettier et al., 2014; Kendall et al., 2016; Marshall et al., 2017; Palta et al., 2015; Pinto et al., 2011; Wang et al., 2007).

More specifically, summary statistics (mean and several measures of variation) calculated based on LRR and BAF signals are used to evaluate DNA sample quality. For example, waviness factor (WF), which characterises the amount of dispersion in hybridisation intensity, is a metric often considered. While the causes of “genomic waves” in microarray data are not fully understood, it has been reported that the magnitude of waves correlates with DNA quantity and GC content (Diskin et al., 2008). Higher GC content leads to stronger hybridisation while lower GC content can cause lack of hybridisation (Koltai and Weingarten-Baror, 2008), causing variation in intensity signals. Therefore, to reduce the waviness the intensities are subjected to adjustment based on GC content, which is a feature provided by PennCNV. Another option is to exclude samples based on their WF (e.g., absolute WF > 0.03).

Additionally, most studies filter out individuals based on their overall CNV burden, as a very high number of CNV calls per sample (e.g., >30) is likely to indicate poor quality intensity signals. Furthermore, SNP call-rate is used to exclude poor quality samples (e.g., <96%). In CNV GWASs, individual variants are excluded if they fail a set length or probe composition criteria (e.g., <10 overlapping probes). Moreover, singletons and ultra-rare CNVs are omitted from analyses due to lack of statistical power. On the other hand, in some instances, CNVs with higher frequency (e.g., >5%) are also excluded (Hujoel et al., 2022) as microarrays are not optimised for the detection of common CNVs and rare CNVs are more likely deleterious.

As an alternative to the conventional cut-off-based filtering, a continuous CNV quality score has been developed (Macé et al., 2016). It predicts the probability that a CNV region is consensually called between PennCNV, QuantiSNP and CNVpartition tools. In this thesis, we refer to it as the consensus-based quality score (cQS). Based on simulation studies, it has been estimated that the cQS improves the statistical power of CNV association detection by up to 20% as conventional filtering strategies tend to be too strict and remove many high quality consensus CNV calls (Macé et al., 2016). The main limitation in cQS development, however, is its reliance on only one input (i.e., SNP microarray data) without any independent validation data layer such as aCGH, WGS or other.

### **1.3. CNV-phenotype associations**

It has long been understood that CNVs increase the risk for severe developmental delay and neuropsychiatric disorders (Coe et al., 2014; Rees and Kirov, 2021) – conditions often diagnosed in early childhood. However, when it was found that healthy adults also carry CNVs, their pathogenicity became under question. The discovery revealed the incomplete penetrance of CNVs (Kirov et al., 2014), meaning that individuals carrying a specific (pathogenic) CNV do not consistently express the associated phenotype. Indeed, several studies have shown that 5–10% of the general healthy population carry a potentially pathogenic large or intermediate-size rare CNV (Itsara et al., 2009; Männik et al., 2015). While originally the CNV studies were carried out almost exclusively in clinical cohorts, the establishment of large biobanks, such as UK Biobank (UKBB) (Bycroft et al., 2018), Estonian Biobank (EstBB) (Leitsalu et al., 2015) and others, have enabled to thoroughly study the phenotypic variability of rare CNVs in cohorts with no clinical preselection.

Many CNV-phenotype association analyses have been conducted using CNVs discovered from sequencing data (Billingsley et al., 2023; Chen et al., 2021; Fitzgerald and Birney, 2022; Khan et al., 2018; Werling et al., 2018; Wheeler et al., 2022). However, in this thesis we concentrate on analyses that incorporate microarray-based CNV calls, with special emphasis on methodological aspects in biobank cohorts. Although these methods miss a substantial fraction of CNVs, it is generally believed that the majority of large CNVs with deleterious effects

are captured. This, and easy access to extensive datasets, provides sufficient justification to use those CNVs in association analyses.

The constraints of CNV analyses emerge from incomplete penetrance (i.e., small effect sizes in association studies) and rarity of CNVs, especially in unselected populations. To overcome these issues, collapsing or burden testing methods are often used where, instead of focusing on each CNV individually, the collective burden of genome-wide CNVs or CNVs overlapping a gene or genomic region of interest is analysed. Variant-level CNV GWASs, however, present a set of completely unique methodological challenges that arise from intersecting CNVs having variation in breakpoints, size, dosage and details of gene disruption (Tzeng et al., 2015). As there are no best practice pipelines or software designed specifically for array-based CNV associations, different studies use various methods to overcome these issues. The main approaches are described in section 1.3.2. of this thesis.

### 1.3.1. Pathogenic CNVs and genomic disorders (GD)

Cytogenetic analyses were the first approaches to associate large genomic rearrangements to neurodevelopmental conditions (Vorstman et al., 2006). With the emergence of array-based methods, the detection resolution for CNVs rose significantly, and arrays became the main clinical diagnostic test for individuals with developmental disabilities or congenital anomalies (Miller et al., 2010). However, as the majority of CNVs are rare and/or nonrecurrent, determining the clinical significance of each variant remains challenging. To facilitate interpretation, pathogenic CNVs together with phenotype records, are compiled into the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) (Firth et al., 2009). Currently (2024/03), DECIPHER contains 48,585 CNVs from 47,980 patients.

Pathogenic CNVs associated with severe health conditions are often termed as genomic disorders (GD) (Harel and Lupski, 2018). GDs are characterised by a wide variety of phenotypes and health outcomes, most often linked to developmental delay and neuropsychiatric disorders. Examples of GDs include Charcot-Marie-Tooth (CMT) disease most frequently caused by a duplication in the 17p11.2 region encompassing the *PMP22* gene (Online Mendelian Inheritance in Man (OMIM) (OMIM, 2024) database ID: 118220), DiGeorge or 22q11.2 deletion syndrome (OMIM: 188400), Miller-Dieker syndrome caused by a deletion in chromosome 17p13.3 (OMIM: 247200), Prader-Willi syndrome caused by the loss of function or deletion of a specific set of genes on paternal chromosome 15 (OMIM: 176270), proximal 16p11.2 deletion (OMIM: 611913) and duplication (OMIM: 614671) syndromes, and others. The effect of GDs on phenotype is recognised for being highly pleiotropic. Specifically, the 16p11.2 CNVs have been associated with developmental delay (Shinawi et al., 2010), schizophrenia (McCarthy et al., 2009), autism spectrum disorders (ASD) (Weiss et al., 2008), body mass index (BMI) (Jacquemont et al., 2011; Walters et al., 2010), head size

(Qureshi et al., 2014) and other phenotypes (Crawford et al., 2019; Giannuzzi et al., 2022; Zufferey et al., 2012).

There are several widely-used approaches for linking potentially pathogenic CNVs to phenotypes (Malhotra and Sebat, 2012). First, family cohorts of unaffected parents and affected offspring have been studied to detect (recurrent) *de novo* CNVs associated with the disease, such as autism (Sanders et al., 2015; Sebat et al., 2007), schizophrenia (Kirov et al., 2012), and bipolar disorder (Georgieva et al., 2014; Malhotra et al., 2011). Additionally, case-control sample sets have been compiled for CNV burden and variant-level association detection. For example, greater genome-wide burden of rare CNVs has been associated with schizophrenia (Buizer-Voskamp et al., 2011; Szatkiewicz et al., 2014; Walsh et al., 2008) and autism (Girirajan et al., 2013). Large studies have been conducted to discover and characterise associated pathogenic CNV regions in schizophrenia patients (Marshall et al., 2017; Rees, Walters, Chambert, et al., 2014; Rees, Walters, Georgieva, et al., 2014) and in children with intellectual disability, developmental delay, ASD and congenital malformations (Coe et al., 2014; Cooper et al., 2011).

Conventionally, the penetrance and pleiotropy of pathogenic CNVs have been characterised in clinically preselected cohorts (Curry et al., 2013; Fetit et al., 2020; Michaelovsky et al., 2012; Shriki-Tal et al., 2017; Skre, 1974; Taylor et al., 2023). However, analyses of pathogenic CNVs in biobanks and cohorts of healthy adults, have shed even more light on the effects of those CNVs in the general population. Although seemingly healthy, several studies have indicated that adults carrying known pathogenic CNVs suffer from various conditions detrimental to their health or quality of life. Findings reveal that CNVs linked to schizophrenia and other neurodevelopmental disorders cause impaired performance on cognitive tests and lower educational qualifications (Kendall et al., 2016), and that overall burden of large and rare ( $\leq 0.05\%$ ,  $\geq 250\text{kb}$ ) deletions and duplications is negatively associated with educational attainment in healthy adults (Männik et al., 2015). Additionally, the burden of rare deletions have been linked to lower IQ (Huguet et al., 2018). Furthermore, pathogenic CNVs have been found associated with numerous physical traits such as BMI, hand grip strength, pulse rate, blood pressure, fat percentage and others (Jacquemont et al., 2011; Owen et al., 2018). It is noteworthy that in most cases the effects are in the direction of reduced performance and increased mortality (Owen et al., 2018). Additionally, carriers of pathogenic CNVs suffer from various common health outcomes such as diabetes, hypertension, obesity and renal failure (Crawford et al., 2019).

### **1.3.2. Genome-wide CNV association studies (CNV GWAS) in population biobanks**

The focus of CNV association analyses has largely remained on neurodevelopmental conditions. These analyses have assumed that both CNVs and associated disorders are rare and that CNVs have large pathogenic effects. The few exceptions

include a CNV association study on endometriosis (Chettier et al., 2014) and another on cardiovascular disease (Glessner et al., 2020). However, the so-called ‘Common Disease, Rare Variant’ hypothesis argues that rare variants with relatively high penetrance are in fact the major genetic contributors to susceptibility of common disease (Schork et al., 2009). In this light, studying the effect of rare CNVs on a much broader set of phenotypic traits is commendable.

The emergence of large datasets and biobanks have facilitated association testing for rare CNVs detected from SNP microarray intensities. Previously described burden testing is a suitable method to further overcome the issues with low CNV frequencies. Alternatively, instead of collapsing variants to increase statistical power, one may opt for consolidating phenotypes. For association testing between rare CNVs and common human diseases, Li et al. (2020) have combined individuals with disease phenotypes into four major case categories as well as control samples. The authors suggest that due to a significant pleiotropic effect of CNVs, this approach is an acceptable substitute where phenome-wide analyses are not feasible.

A meta-analysis study of associations between rare CNVs and four anthropometric traits was carried out on data from UKBB and 25 smaller cohorts (Macé et al., 2017). The authors used PennCNV for CNV detection and incorporated the cQS quality measure (described in section 1.2.3.) to produce a more precise probabilistic CNV call (Macé et al., 2016). Instead of using full CNV regions, probabilistic copy number dosages were retrieved for each microarray probe used for CNV detection. The dosages were subsequently input into a linear regression model and analysed using custom scripts. Altogether, seven CNV regions were discovered to be associated to one or several of the tested traits.

The largest phenome-wide CNV analysis to date was conducted by Aguirre et al. (2019) analysing 3,157 continuous and disease phenotypes in UKBB. To surmount most of the challenges presented by CNV analyses, this study only incorporated recurrent PennCNV calls with fixed breakpoints and frequency  $>0.0005\%$ . The CNV dosages were coded as SNV effect allele dosages and analysed with PLINK v2 software (Chang et al., 2015), originally designed for SNV analyses. Variant-level analysis and gene-level burden testing were carried out and resulted in 4 and 32 associations per phenotype on average, respectively.

The most recent CNV association study conducted on 56 quantitative traits of UKBB participants, achieved a 6-fold increase in CNV detection sensitivity compared to previous analyses with 269 independent associations involving 97 loci and 252 likely-causal CNVs (Hujoel et al., 2022). This increase can be attributed to a more sensitive HI-CNV pipeline (described in section 1.2.1.) and the inclusion of related individuals in the analysis. The associations were tested with linear mixed models implemented in BOLT-LMM (Loh et al., 2015) and three analysis types were used: probe-level, CNV-level, and gene-level tests. Approximately half of the detected loci revealed putative target genes.

Recent CNV GWASs underscore the importance of CNVs in shaping phenotypic diversity and human disease. It has been discovered that there is an excess of rare CNVs overlapping OMIM morbid genes as well as significant SNV GWAS

loci (Li et al., 2020), and that the majority of associated rare CNVs are the likely-causal variants in the locus linked to phenotype (Hujoel et al., 2022). This further emphasises the need for extensive association and subsequent follow-up analyses. However, CNV studies are few in numbers due to challenges in variant discovery, quality control and analysis methods.

## 2. AIMS OF THE STUDY

The aim of this thesis is to improve the quality assessment of SNP microarray-based CNV calls and leverage the CNV data to detect novel CNV-phenotype associations, either at a genome-wide scale or by focusing on a single well-established pathogenic CNV region.

The specific objectives of the thesis are as follows:

- to utilise multi-omics data layers to evaluate CNV quality and develop an omics-informed quality score for PennCNV calls to improve statistical power in subsequent CNV association analyses;
- to explore a novel association between the 16p11.2 BP4-BP5 syndromic CNV and pubertal timing, focusing on fine-mapping the CNV region to the most likely causal gene using *in vivo* and *in silico* approaches;
- to compile a comprehensive map of rare CNV associations with clinically relevant continuous measurements and disease phenotypes;
- to assess dosage sensitivity of genes encompassed within disease-causing CNVs and predict haploinsufficiency and triplosensitivity of all genes across genome.

### 3. RESULTS AND DISCUSSION

#### 3.1. Omics-informed CNV quality score (Ref. I)

The emergence of biobank datasets has resulted in increased popularity of large-scale CNV GWASs that until recently, due to technical limitations and low variant frequency, had remained predominantly an unexplored field. The abundance of SNP genotyping microarray datasets has made them the main source for inference of CNV information, as compared to considerably lower sample sizes of available WGS data. Currently, PennCNV (Wang et al., 2007) is the most widely-used CNV detection software from SNP microarray signals. However, this and other similar algorithms are prone to FP (and FN) calls. Therefore, intensive quality control is a prerequisite for any CNV association study.

This part of the thesis concentrates on improving the CNV quality prediction methodology for SNP microarray-based detection. In order to discriminate between high and low quality CNV calls, we have incorporated three independent multi-omics data layers – WGS, gene expression from RNAseq, and methylation array intensity data. Our aim was to build an omics-informed CNV quality score (OQS; Figure 3A) model and compare it to a previously published consensus-based quality score (cQS, see section 1.2.3.) (Macé et al., 2016) as well as conventional CNV quality control and filtering approaches relying on cut-off values of CNV and sample-specific parameters (e.g., CNV length, number of calls per sample, mean and standard deviation of intensity signal parameters, etc.) (Chettier et al., 2014; Palta et al., 2015; Pinto et al., 2011; Wang et al., 2007). The model validations were conducted on CNVs shared between close family members and known CNV-phenotype associations (Macé et al., 2017). Furthermore, our data enabled us to evaluate the amount and distribution of FP and FN calls of PennCNV, which are important statistics to note when conducting any CNV GWAS. Although our work is based on PennCNV, the methodological concepts introduced can be applied to any CNV detection algorithm used on sample sets with at least one additional omics layer available.

##### 3.1.1. Description of cohort and data preparations

Samples included in this study originated from four independent population-based cohorts genotyped on various platforms: EstBB (Illumina Infinium Omni-Express-24 and Illumina Global Screening Array (GSA)), Lifelines Deep (LLDeep; Human CytoSNP-12) (Tigchelaar et al., 2015), Swiss Kidney Project on Genes in Hypertension (SkiPOGH; Illumina 2.5) (Alwan et al., 2014) and UKBB (Affymetrix UK Biobank Axiom and Affymetrix UK BiLEVE Axiom arrays). All analyses were carried out under approval of relevant institutional ethics committees and all participants signed a broad informed consent. The sample sizes for datasets used in different steps of the study are summarised in Table 1.

For each cohort, CNV detection was performed with PennCNV software (Wang et al., 2007). Adjacent CNV calls were merged and samples with >200 CNVs or CNVs larger than 10 Mbp were removed. For each sample, we obtained the putative CNV regions together with the values of three CNV-specific (length, number of overlapping SNPs, PennCNV confidence score) and seven sample-specific parameters (number of CNV calls per sample, mean and SD of BAF, BAF drift, mean and SD of LRR, and absolute signal WF; see section 1.2.). Additional microarray size-dependent parameters were considered, such as CNV length and number of calls corrected for SNP density.

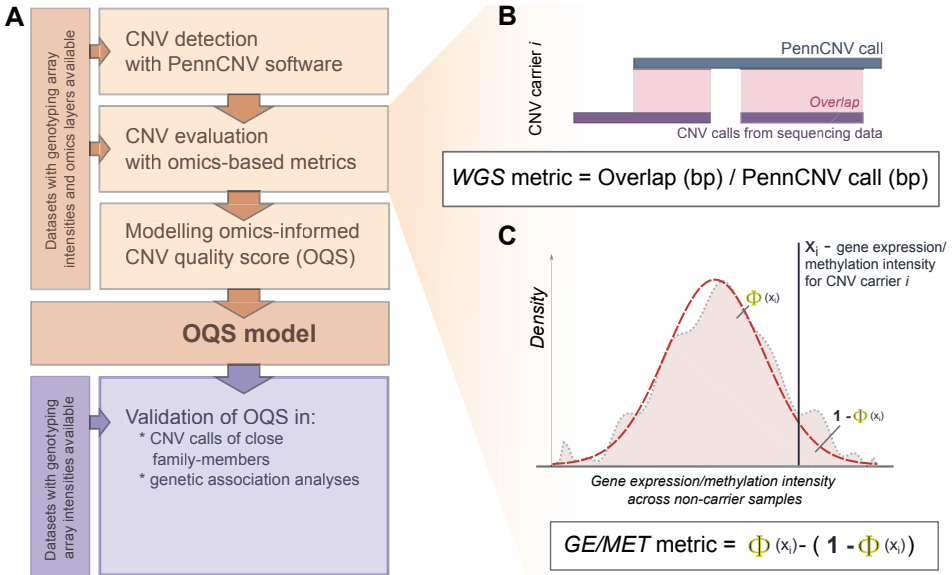


**Figure 3.** (A) CNV quality evaluation and modelling pipeline for PennCNV calls. The quality metrics are estimated based on (B) whole genome sequencing (WGS) and (C) gene expression (GE) and/or methylation intensity (MET) data. (B) WGS metric is calculated as the fraction of a PennCNV call overlapped by WGS CNVs of the same individual. (C) To calculate GE/MET metrics, the reference distribution of expression/intensity based on non-carriers (pink area) is approximated to a normal distribution (red dashed line). The expression/intensity value of each potential CNV carrier ( $x_i$ ) is compared to the reference. The metric is a difference between the fraction of non-carriers with the corresponding value  $\leq x_i$  and those with the corresponding value  $> x_i$  and captures how extreme  $x_i$  is compared to the reference distribution of non-carriers. In case a CNV overlaps with several genes/CpG sites, the metric values are averaged over them. (Adapted from Ref. I, Figure 1)

**Table 1.** Overview of datasets and sample sizes in the analysis steps. N/A, not applicable. (Adapted from Ref. I, Table 1)

| Sample counts                                                      |                  |     |                   |        | Analysis steps                   |                                |   |                  |
|--------------------------------------------------------------------|------------------|-----|-------------------|--------|----------------------------------|--------------------------------|---|------------------|
| Dataset                                                            | n                | WGS | Methylation array | RNAseq | Omics-based metrics calculations | Model selection and validation |   |                  |
|                                                                    |                  |     |                   |        |                                  | Model building                 |   | CNV associations |
| Estonian OmniExpress sample set (N = 7,750)                        |                  |     |                   |        |                                  |                                |   |                  |
| EstBB-MO                                                           | 1,066            | 983 | 295               | 382    | +                                | +                              | − | −                |
| First-degree relatives                                             | 504 <sup>a</sup> | N/A | N/A               | N/A    | −                                | −                              | + | −                |
| Lifelines Deep (LLDeep; N = ~1,500)                                |                  |     |                   |        |                                  |                                |   |                  |
| LLDeep                                                             | 1,383            | N/A | 768               | 1,098  | +                                | +                              | − | −                |
| Swiss Kidney Project on Genes in Hypertension (SkiPOGH; N = 1,128) |                  |     |                   |        |                                  |                                |   |                  |
| SkiPOGH                                                            | 466              | N/A | 148               | 405    | +                                | −                              | − | −                |
| Parent-child pairs                                                 | 319              | N/A | N/A               | N/A    | −                                | −                              | + | −                |
| Estonian Biobank GSA sample set (EstBB-GSA; N = ~200,000)          |                  |     |                   |        |                                  |                                |   |                  |
| EstBB-GSA (unrelated)                                              | 89,516           | N/A | N/A               | N/A    | −                                | −                              | − | +                |
| MZ twins                                                           | 312              | N/A | N/A               | N/A    | −                                | −                              | + | −                |
| First-degree relatives                                             | 79,903           | N/A | N/A               | N/A    | −                                | −                              | + | −                |
| UK Biobank (UKBB; N = ~500,000)                                    |                  |     |                   |        |                                  |                                |   |                  |
| UKBB                                                               | 331,522          | N/A | N/A               | N/A    | −                                | −                              | − | +                |
| MZ twins                                                           | 302              | N/A | N/A               | N/A    | −                                | −                              | + | −                |
| First-degree relatives                                             | 42,032           | N/A | N/A               | N/A    | −                                | −                              | + | −                |

<sup>a</sup>Estonian OmniExpress first-degree relatives do not overlap with EstBB-MO samples.

Additional multi-omics layers (WGS reads, gene expression, and methylation array data) used for CNV quality metric calculations and OQS modelling were available for a subset of EstBB (defined as multi-omics set of EstBB or EstBB-MO), LLDeep and SkiPOGH samples. For all three sample sets, gene expression levels were estimated from RNAseq reads and only genes showing dosage-dependence (Pearson  $R > 0.1$ ) in an independent study (Talevich and Hunter Shain, 2018) were kept. Furthermore, expression values of the remaining genes were corrected for independent study-specific expression quantitative trait loci (eQTLs,  $P < 0.05$ ) in genes' 500 kbp flanking region. Methylation intensity was captured using Illumina Infinium Human Methylation 450k Beadchip array type II CpG sites. WGS CNVs were detected only in EstBB-MO dataset using GenomeSTRiP pipeline (Handsaker et al., 2015).

For quality score model validations, we used shared CNVs of close family members from EstBB, SkiPOGH and UKBB. Pairs of mono-zygotic (MZ) twins and first-degree relatives from EstBB and UKBB were detected using KING-robust kinship estimator (Manichaikul et al., 2010) coefficient ranges [0.354–0.500) and [0.177–0.354), respectively. SkiPOGH is a family-based cohort and parent-child pairs were used there. For final validations in CNV-phenotype association setting, quality-controlled subsets of 89,516 and 331,522 unrelated European ancestry EstBB and UKBB samples, respectively, were included (see Ref. I for details).

### 3.1.2. Calculation and comparison of omics-based CNV quality metrics

For each putative CNV region, we calculated up to three independent omics-based quality metrics using EstBB-MO, LLDeep and SkiPOGH datasets. The WGS metric was estimated as the fraction of a PennCNV call overlapping with WGS CNVs in the same sample (Figure 3B). Methylation (MET) and gene expression (GE) metrics were calculated for CNV regions that overlap with CpG sites and genes, respectively. The calculations were based on the assumption that compared to the general population a true CNV carrier has either excess (i.e., duplication) or lack of (i.e., deletion) genetic material, which results in higher or lower captured (methylation or gene expression) intensity in the observed genomic region, respectively. For example, true deletions lead to lower expression of dosage-dependent genes and decrease overall intensity (sum of methylated and un-methylated intensities) captured from CpG sites, while false deletions do not. Duplications act in reverse. We exploited this phenomenon to validate detected CNV regions.

First, for each included CpG site/gene  $i$ , we constructed the approximately Gaussian intensity/expression distribution (with cumulative distribution function  $\Phi$ ) based on samples assumed to be copy-neutral in this genomic region. Secondly, we compared this reference to the intensity/expression value ( $x_i$ ) of a potential CNV carrier with a measure equivalent to a signed 2-sided tail-probability  $p_i \in [0; 1]$ , which indicates how extreme  $x_i$  is compared to the general population (Figure 3C):

$$p_i = \Phi(x_i) - (1 - \Phi(x_i)).$$

For each CNV, we calculated mean  $p_i$  values across  $n$  CpG sites or genes of which >80% overlap with the CNV, respectively:

$$\text{MET or GE} = \frac{\sum_{i=1}^n p_i}{n}.$$

We expected MET/GE to be  $< 0$  (i.e., decreased intensity) for deletions and  $> 0$  (i.e., increased intensity) for duplications. If this was not the case, the corresponding values were set to 0. Finally, MET and GE were transformed to their absolute values.

All three calculated metrics had bimodal distribution with modes near 1 and 0, indicating their ability to clearly differentiate between true and false calls. Furthermore, within each sample set, the metrics were highly concordant. Moderate to strong correlation was observed between the MET and GE metrics depending on the dataset (Pearson  $R \in [0.59; 0.8]$  for deletions,  $R \in [0.33; 0.57]$  for duplications), while the WGS metric was strongly correlated ( $R \geq 0.7$ ) to both MET and GE metrics. Interestingly, the correlation between the WGS metric and the previously published cQS was considerably lower for comparison ( $R = 0.17$  for deletions,  $R = 0.43$  for duplication). This illustrates the benefit of incorporating omics layers as additional CNV quality validators and suggests that simply using multiple software for array-based CNV detection (as was done for cQS) might not be as informative. Our results also indicate that gene expression and methylation data are a good alternative to WGS for CNV validation in samples with no WGS data available.

As a final step, to keep just one quality value per CNV we defined a combined omics-based metric as ‘the most extreme’ (i.e., furthest from 0.5) metric of the three.

### 3.1.3. Estimation of false positive and false negative rate of PennCNV

Previously, it has been suggested that only approximately 20% of PennCNV calls are high confidence true positives (Macé et al., 2016). By (a) incorporating the previously calculated combined metric and (b) comparing PennCNV output to WGS CNVs, we have evaluated both FP as well as FN PennCNV calls.

We found that the percentages of FP (i.e., absolute combined omics-informed metric  $< 0.1$ ) duplications were 47.5% in EstBB-MO, 28.3% in LLDeep and 59.4% in SkiPOGH. The percentages of FP deletions were equal or higher, with 47.3% in EstBB-MO, 50.5% in LLDeep and 70.9% in SkiPOGH. These results clearly illustrate the necessity of quality control prior further CNV analyses.

Furthermore, we evaluated the genome-wide distribution of FP and FN calls based on comparison with WGS CNVs in 979 EstBB samples. The comparison was carried out in a probe-by-probe manner using Illumina Infinium Omni-Express-24 autosomal probe locations ( $n = 709,358$ ). Overall, 17.0% ( $n = 120,395$ ) of probes had FP rate  $> 0$ . We observed a negative correlation (Pearson  $R = -0.12$ ) between FP rate and CNV frequency indicating that FP calls mostly appear as randomly distributed rare CNV events. Still, we discovered that 0.12% ( $n = 816$ ) of probes contributed to FP “hot spots” that generated FP CNVs with  $> 1\%$  frequency. We suggest excluding these “hot spot” regions from association analyses.

Additionally, we estimated that the fraction of FN CNVs based on overlap with WGS CNVs is as high as 97.3%. This is partly due to the requirement for microarray-based CNVs to overlap with fixed probe locations in order to be detectable. For PennCNV, specifically, a CNV needs to overlap with  $\geq 3$  probes. When considering only WGS CNVs that meet this requirement, the fraction of FN dropped to 75.7% ( $n = 69,889$ ) and continued dropping with increasing number

of overlapping probes (48.4% for  $\geq 10$  probes; 8.8% for  $\geq 30$  probes). Another array-based detection limitation is that all signal normalisation is based on the assumption that the reference population is copy neutral. This hinders the detection of high-frequency CNVs (Winchester et al., 2009). To support that claim, we observed a positive correlation between probe-specific FN rate and CNV frequency (Pearson  $R = 0.19$ ), and CNVs with  $>50\%$  frequency contributed to 31.8% ( $n = 22,196$ ) of all FN CNVs overlapping  $\geq 3$  probes.

#### 3.1.4. Omics-informed CNV quality score (OQS) modelling and validation

The main aim of this project was to be able to extrapolate the omics-based CNV quality estimations to datasets for which only SNP microarray data are available. In order to do that, we built an omics-informed quality score (OQS) model that predicts the previously described combined metrics based on CNV and sample-specific PennCNV output parameters. The full model building workflow is described in detail in Ref. I. Briefly, we fitted a generalised linear regression model with quasi-binomial link function, as the response variable was bimodal instead of being strictly binary (typical for logistic regression). Parameter selection was implemented in a forward stepwise algorithm where in each round a parameter (or interaction) that minimised the 10-fold cross-validation mean square error (MSE) was added to the model. If including additional parameters did not improve the average MSE, the existing model was returned and the final OQS values could be calculated as

$$OQS = \frac{1}{1 + \exp(-(\beta_0 + \sum_{k=1}^n X_k \beta_k))},$$

where  $n$  is the number of selected PennCNV parameters,  $X_k$  is the  $k$ th selected parameter (explanatory variable),  $\beta_0$  is the model intercept term and  $\beta_k$  is the effect estimate of explanatory variable  $X_k$ . OQS can obtain values between 0 and 1, with higher values corresponding to higher confidence CNV calls. We fitted the models separately for deletions and duplications, and for EstBB-MO and LLDeep datasets (SkiPOGH was omitted from model fitting due to the high fraction of FP CNV calls).

To choose the model that is the most generalisable across different genotyping platforms and datasets, we incorporated CNVs of MZ twins and parent-child pairs from SkiPOGH, UKBB and EstBB (not overlapping with EstBB-MO samples). The CNVs were divided into two sets: “familial CNVs”, i.e., CNVs shared between close family members (Jaccard index based on overlapping bp count of at least 0.9) and “non-familial CNVs”. We reasoned that the former contains a much larger subset of true positive CNVs compared to the latter. Therefore, the best models (one for deletions and one for duplications) were chosen as the ones that maximised the mean OQS values between the two CNV groups, averaged over three datasets. The terms and effect sizes of the final models are presented in Ref. I Tables S10 and S11.

The final models were further validated on independent sets of familial and non-familial CNVs of first-degree relatives from UKBB and EstBB (Figure 4). On average, OQS difference between the two CNV groups ranged between 0.22 and 0.35 depending on the dataset. Importantly, OQS showed significantly higher values (0.67–0.82 for deletions and 0.48–0.70 for duplications) for familial CNVs compared to the previously published cQS (0.27–0.32 in deletions and 0.42–0.53 in duplications) with paired Wilcoxon test  $P < 1.4 \times 10^{-21}$  once again illustrating the benefit of omics-based CNV validations.

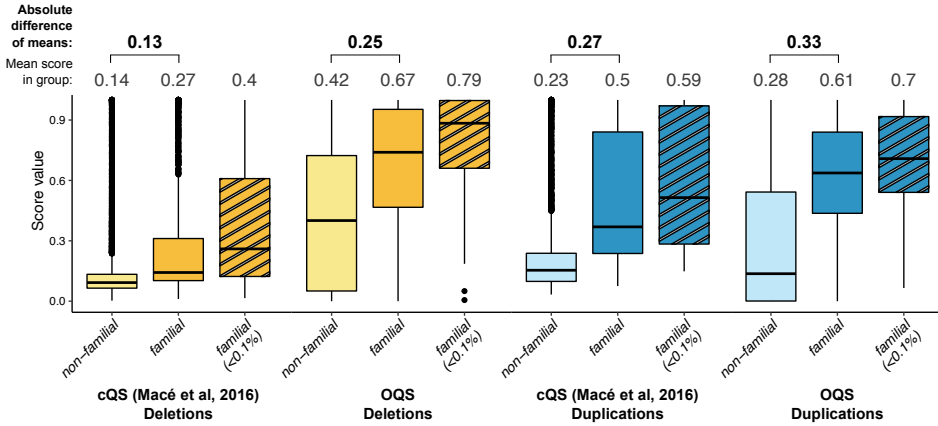


**Figure 4.** Comparison of cQS and OQS on PennCNV deletion (yellow) and duplication (blue) calls of EstBB first-degree relatives. Familial CNVs are likely true positives, while non-familial group contains both true and false positives. We included rare (frequency <0.1%, striped background) familial CNVs as a subset of CNVs less likely to validate in a relative by chance. The mean score of each group and their pairwise difference are shown on top of the figure. All differences are significant with  $p < 1.0 \times 10^{-16}$ . (Ref. I, Figure 3)

### 3.1.5. Incorporating CNV quality in association studies

Predominantly, OQS models were developed with CNV association studies in mind. Therefore, as the principal target, we tested them in an association setting and compared to raw PennCNV output, four sets of previously published quality filtering parameter cut-offs (Chettier et al., 2014; Palta et al., 2015; Pinto et al., 2011; Wang et al., 2007) and cQS (Macé et al., 2016). Our motivation was to evaluate the improvement in statistical power in contrast to other methods.

We selected 21 previously established associations between CNVs and four human anthropometric traits (BMI, height, weight and waist-hip ratio (WHR)) (Macé et al., 2017). We concentrated on mirror type association models, i.e., deletions and duplications affect phenotypes with similar magnitudes but opposite directions. For that, OQS (and cQS) values for deletions were negated prior to incorporating the score into the association models (analogously to SNV dosages). The analysis is then executed in a probe-by-probe manner across probes overlapping with >5 CNVs in the regions of interest.

We replicated 10 associations in EstBB and 18 in UKBB dataset (Bonferroni-corrected  $P < 0.05/21$ ). Using these associations, we aimed to estimate the average relative change in explained variance when using OQS as opposed to other methods. This was done for each phenotype and dataset separately. First, for each significant CNV region, we clumped the probes with  $R^2 > 0.5$  based on randomised probe priority orders 20 times. Second, we calculated the change in variance explained based on the clumped probes retaining 20 values depicting that change. Finally, the average over these repetitions was reported.

We observed that OQS systematically increased the explained variance in most of the comparisons (Figure 5). In EstBB, the increase was 2%–34% and 23%–55% when compared to raw PennCNV and cQS, respectively. In UKBB, OQS showed improvement (up to 10%) in three out of four phenotypes compared to raw PennCNV but compared to cQS the explained variance decreased. Partly, this was to be expected as the associations we replicated were originally detected in the UKBB using the cQS and, thus, the results could be biased towards favouring cQS. Additionally, cQS could simply be more appropriate to use in the UKBB as different scores can suit to different datasets. Importantly, though, our results clearly indicate that using cut-off-based filtering approaches in CNV analyses is too stringent and results in major loss of statistical power. In EstBB, the increase was up to 56% but the change was even more notable in the UKBB where for several filtering sets the improvement of OQS was  $>100\%$  (in some cases  $>400\%$ ).

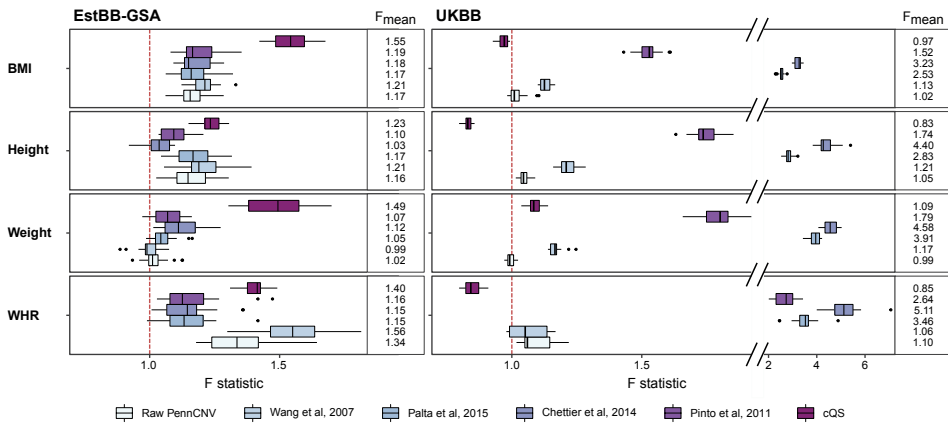


**Figure 5.** Change of variance explained when using OQS over raw PennCNV, four published quality filtering approaches, and cQS in the EstBB-GSA (left) and UKBB (right), depicted as distribution of F statistics. Explained variance is increased when  $F > 1$  and decreased when  $F < 1$ . Larger F values indicate greater improvement in statistical power when using OQS over the given reference approach. (Ref. I, Figure 4A)

Still, for most comparisons, modest improvements were anticipated as false CNV calls only introduce a small amount of noise per region (FP “hot spots” excluded). Additionally, the replicated associations mostly incorporated easily-detectable well-studied CNV events. However, we expect to gain more statistical power in intermediate-quality CNV regions for which the phenotype associations have previously not been detected. Furthermore, whenever there are additional omics data layers available, we encourage the scientific community to develop sample set-specific quality models that are tailored to work best in given datasets of interest. Therefore, the true extent of improvement offered by OQS and omics-based CNV validations is yet to be determined in future CNV association studies.

### **3.2. Association between 16p11.2 CNV and reproductive traits (Ref. II)**

Among the most studied genomic disorders in clinical and population cohorts are the proximal 16p11.2 deletion (OMIM: 611913) and reciprocal duplication (OMIM: 614671) syndromes. The 16p11.2 locus encompasses five clusters of segmental duplications (BP1-BP5) that cause multiple rare recurrent rearrangements, among them a ~600 kbp CNV (denoted by its breakpoints as BP4-BP5). The CNV harbours 28 genes and has been associated with developmental delay, schizophrenia, and ASD (McCarthy et al., 2009; Shinawi et al., 2010; Weiss et al., 2008). Deletion carriers exhibit severe obesity, while duplication carriers show significantly reduced weight (Jacquemont et al., 2011; Walters et al., 2010). The deletions and duplications are associated with a head size increase (macrocephaly) and decrease (microcephaly), respectively (Qureshi et al., 2014). Furthermore, 16p11.2 BP4-BP5 deletions have been linked to higher risk of diabetes, osteoarthritis, hypertension, asthma, anaemia and renal problems (Crawford et al., 2019). A recent study has associated the 16p11.2 CNV to white blood cell counts (Giannuzzi et al., 2022). Large SNV GWAS meta-analyses have also linked this region to age at menarche (Day et al., 2017; Perry et al., 2014).

In this part of the thesis, we expanded on the existing knowledge about the association between variants in the 16p11.2 region and reproductive traits by linking the CNV directly to pubertal timing and reproductive tract disorders in both clinical and biobank cohorts. Additionally, we conducted follow-up analyses to characterise and fine-map the region down to the most likely causal gene affecting pubertal timing. In this thesis, we mainly concentrate on *in silico* CNV analyses. However, we briefly touch the *in vivo* follow-up, which have shed further light on possible genic interaction in this region.

### 3.2.1. Description of cohorts and methods

Computational analyses of this study were carried out on two clinically unselected population cohorts: UKBB (analysed in two release cohorts, i.e., discovery and replication) and EstBB. Additionally, 16p11.2 BP4-BP5 CNV carrier probands and their close relatives were included from the Pan-European 16p11.2 cohort and The Simons Variation in Individuals (Simons VIP) Collection of 16p11.2 Families clinical cohort. The project was approved by the Ethics Review Committee on Human Research of the University of Tartu, Estonia and The Committee of Human Research Ethics of the Canton of Vaud, Switzerland. All participants signed an informed consent.

CNV detection was conducted by using the PennCNV (v1.0.4) software. All participants with >200 CNV calls or SNP genotyping call-rate <98% were excluded. A participant was labelled as a 16p11.2 BP4-BP5 CNV carrier if they had >70% overlap with the corresponding CNV region as defined in DECIPHER database (chr16:29,606,852-30,199,855, human genome build 19) (DECIPHER CNV Syndromes, 2024; Firth et al., 2009). All UKBB and EstBB 16p11.2 CNVs were further confirmed with the independent iPsychCNV detection software (iPsychCNV, 2017). Additionally, the carrier status of a selection of EstBB participants had previously been validated with quantitative polymerase chain reaction (Männik et al., 2011). No CNV quality score (such as cQS or OQS) was employed in this study, given that large well-established recurrent CNV regions, such as 16p11.2 BP4-BP5, are typically straightforward to detect.

Finally, if more than one close family member was a CNV carrier, only one randomly selected member was retained. Also, all CNV carriers from UKBB replication cohort that had a family member among carriers from UKBB discovery cohort were excluded. The cohort-specific numbers of CNV carriers and copy-neutral controls included in the analyses are presented in Table 2.

For all participants, self-reported onset of puberty and BMI was obtained from baseline recruitment questionnaires. We used age at menarche (AAM) as a proxy for female pubertal timing. Prior to conducting the analyses, we excluded the women with extreme outlier AAM (<5 or >25). For male participants, we only had indicative data from UKBB questionnaires where men assessed based on their subjective observations whether they had first facial hair and voice break at a “younger than average”, “average” or “older than average” age. Furthermore, we studied the abundance of diagnoses related to disorders of reproductive tract and fertility in 16p11.2 BP4-BP5 carrier groups. In UKBB, diagnoses in the form of ICD-10 (International Classification of Disease, 10th Revision) codes were extracted from initial assessment visit health reports (2006–2010). In EstBB, we queried the ICD-10 codes from the Estonian Health Insurance Fund database (Tervisekassa, 2024). Additionally, we obtained the relevant disorders from the medical history of Simons VIP participants.

**Table 2.** Sample counts for controls (CTRL), 16p11.2 deletion (DEL) and duplication (DUP) carriers as well as association statistics with female pubertal timing proxied by age at menarche (AAM). Abbreviations: Eur./non-Eur., Europeans/non-Europeans.

|                              |                  | Cohorts                     |                               |                        |                      |                          |                             |
|------------------------------|------------------|-----------------------------|-------------------------------|------------------------|----------------------|--------------------------|-----------------------------|
|                              |                  | UKBB<br>discovery<br>(Eur.) | UKBB<br>replication<br>(Eur.) | UKBB<br>(non-<br>Eur.) | EstBB                | Simons<br>VIP<br>16p11.2 | Pan-<br>European<br>16p11.2 |
| <b>Sample counts</b>         |                  |                             |                               |                        |                      |                          |                             |
| Males                        | CTRL             | 56,047                      | 112,890                       | 35,870                 | 11,955               |                          |                             |
|                              | DEL              | 18                          | 36                            | 13                     | 4                    | 27                       |                             |
|                              | DUP              | 11                          | 36                            | 8                      | 4                    | 24                       |                             |
| Females                      | CTRL             | 62,217                      | 133,615                       | 43,747                 | 22,354               |                          |                             |
|                              | DEL              | 11                          | 23                            | 18                     | 8                    | 22                       | 14                          |
|                              | DUP              | 23                          | 28                            | 4                      | 10                   | 20                       | 11                          |
| <b>Age at menarche (AAM)</b> |                  |                             |                               |                        |                      |                          |                             |
| Mean<br>[years]              | CTRL             | 12.9                        | 13.0                          | 13.0                   | 13.5                 |                          |                             |
|                              | DEL              | 11.4                        | 11.9                          | 11.9                   | 12.3                 | 11.0                     | 11.6                        |
|                              | DUP              | 14.4                        | 14.0                          | 14.7                   | 15.6                 | 13.3                     | 13.5                        |
| ANOVA<br>(P-val.)            | DEL vs<br>CTRL   | $1.3 \times 10^{-2}$        | $4.5 \times 10^{-2}$          | 0.19                   | 0.19                 |                          |                             |
|                              | DUP vs<br>CTRL   | $7.8 \times 10^{-5}$        | $1.1 \times 10^{-3}$          | 0.13                   | $9.8 \times 10^{-6}$ |                          |                             |
|                              | Mirror<br>effect | $1.9 \times 10^{-5}$        | $6.5 \times 10^{-4}$          | 0.13                   | $2.4 \times 10^{-5}$ | $7.7 \times 10^{-5}$     | $3 \times 10^{-3}$          |

To detect the differences of mean BMI and AAM in 16p11.2 dosage groups, we used the analysis of variance (ANOVA) method. As due to small sample sizes the normality assumption was weak for AAM, the results were also validated using a non-parametric Wilcoxon rank-sum test. For binary and categorical outcomes, Fisher’s exact test was used. Furthermore, two-sample univariate (Zhu et al., 2016) and multivariate (McDaid et al., 2017) Mendelian Randomisation (MR) was conducted to test whether the expression of 16p11.2 genes can be causally linked to AAM. For this analysis, the most recent AAM GWAS (Day et al., 2017) and whole-blood eQTL data obtained from the eQTLGen consortium (Võsa et al., 2021) were used. Altogether, 12 (43% out of 28) 16p11.2 BP4-BP5 region genes with strong genetic instruments ( $F > 20$ ) were included in MR. We set the significance threshold to  $\alpha_{SMR} = 0.05/12$  and required  $P > 0.0009$  in HEIDI (Heterogeneity in Independent Instruments) test (Giambartolomei et al., 2014) to discard cases where instead of a colocalising causal variant there are two distinct but correlated genetic variants affecting AAM and gene expression. Finally, for all 16p11.2 genes, mRNA over-expression and CRISPR/Cas9 genome editing was done in *Tg(gnrh3:egfp)* transgenic zebrafish (Abraham et al., 2009) to investigate duplication and deletion effects, respectively, via automated live

imaging (see Ref. II Supplemental Methods). *Tg(gnrh3:egfp)* is a model designed for studying gonadotropin-releasing hormone-3 (GnRH3) system, which is known to be crucial for reproduction in all vertebrates.

### 3.2.2. Association between 16p11.2 BP4-BP5 dosage and pubertal timing

First, to assess the validity of the 16p11.2 BP4-BP5 CNV carriers, we confirmed a previously detected highly significant and replicable association between the CNV dosage and BMI in UKBB discovery cohort ( $P = 5.3 \times 10^{-25}$ ), UKBB replication cohort ( $P = 3.8 \times 10^{-44}$ ) and EstBB ( $P = 6.4 \times 10^{-8}$ ) with deletions increasing (UKBB males: difference of BMI between carriers and non-carriers ( $\Delta$ ) = +6.8, UKBB females:  $\Delta$  = +8.5) and duplications decreasing (UKBB males:  $\Delta$  = -1.2, UKBB females:  $\Delta$  = -1.9) the BMI. Next, we observed a previously unexplored association between 16p11.2 BP4-BP5 and AAM in UKBB discovery cohort female participants of European origin ( $P = 1.9 \times 10^{-5}$ , ANOVA, corrected for BMI, age, and four principal components (PCs)). Compared to copy-neutral controls (mean AAM in group ( $\mu_{AAM}$ ) = 12.9 years), the deletion carriers had AAM 1.5 years earlier ( $\mu_{AAM}$  = 11.4 years, Wilcoxon  $P = 0.00098$ ) and the duplication carriers 1.5 years later ( $\mu_{AAM}$  = 14.4 years, Wilcoxon  $P = 0.002$ ) (Figure 6A). These results were further validated in Europeans from UKBB replication cohort ( $P = 6.5 \times 10^{-4}$ ) and EstBB ( $P = 2.4 \times 10^{-5}$ ) as well as in Simons VIP ( $P = 7.7 \times 10^{-5}$ ) and Pan-European ( $P = 3.4 \times 10^{-3}$ ) clinical cohorts (Table 2). We separately analysed the full sample set of UKBB non-European women and the results, despite not being statistically significant, were consistent with other cohorts (controls:  $\mu_{AAM}$  = 13.0; deletions:  $\mu_{AAM}$  = 11.9,  $P = 0.0056$ ; duplications:  $\mu_{AAM}$  = 14.7,  $P = 0.093$ ).

For male participants of UKBB discovery cohort we observed a non-significant trend for deletion carriers towards earlier voice break ( $OR = 2.5$ , Fisher's exact test  $P = 0.21$ ) and first appearance of facial hair ( $OR = 2.7$ ,  $P = 0.08$ ). The duplication carriers trended towards later voice break ( $OR = 1.4$ ,  $P = 0.52$ ) as well as later facial hair ( $OR = 2.8$ ,  $P = 0.09$ ) (Figure 6B). Similar tendencies were present in UKBB replication cohort (see Ref. II Table S4). These results suggest that the effect of 16p11.2 BP4-BP5 CNV on pubertal timing, though more easily detectable in women due to data availability, is not sex-specific.

### 3.2.3. Disorders of reproductive health and sexual development of 16p11.2 BP4-BP5 CNV carriers

To further characterise the effect of 16p11.2 BP4-BP5 CNV on sexual health, we tested for associations between CNV carrier status and diagnoses of female reproductive tract disorders. With deletions and duplications combined, we detected nominally significant associations in EstBB for “absent, scanty and rare menstruation” (ICD-10: N91;  $OR = 4.4$ ,  $P = 0.013$ ), “non-inflammatory disorders

of ovary, fallopian tube and broad ligament” (N83;  $OR = 5.2$ ,  $P = 0.003$ ) and “inflammatory diseases of female pelvic organs” (N70-N77;  $OR = 4.1$ ,  $P = 0.012$ ) (Figure 6C). Altogether, 50% and 64% of EstBB deletion and duplication carriers, respectively, had been diagnosed with irregular/absent menstruation or various hormonal and ovarian disorders.

Unfortunately, these analyses resulted in no significant associations in UKBB, possibly due to UKBB containing mostly post-menopausal women (40–69 years of age at recruitment in 2006–2010) with little verified diagnosis information from before 1997 (1981 for Scottish participants). We did, however, observe that 16p11.2 deletion carriers had a significantly lower number of live births (population average  $\mu = 1.81$ ) in both UKBB discovery ( $\mu_{del} = 0.91$ ,  $P = 0.013$ ) and replication cohorts ( $\mu_{del} = 0.78$ ,  $P = 2.9 \times 10^{-5}$ ). Additionally, female duplication carriers had an elevated risk of miscarriage in discovery cohort ( $OR = 2.5$ ,  $P = 0.037$ ). This finding did not replicate in the second cohort.

Although we were unable to replicate many of our findings, the over-representation of reproductive disorders in 16p11.2 BP4-BP5 deletion and duplication carriers indicates a clear pattern which can be further studied in cohorts with larger numbers of affected samples.

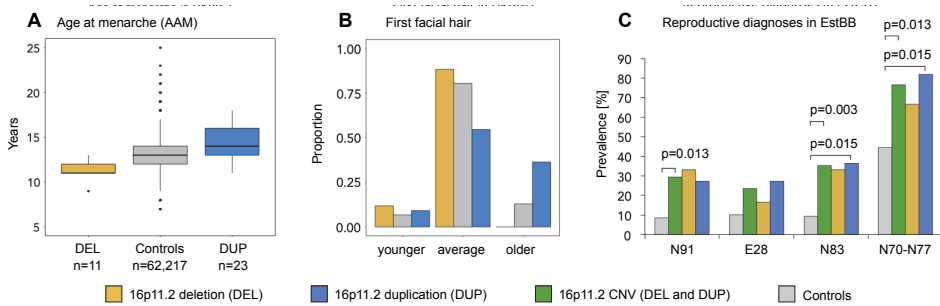


**Figure 6.** Distribution of age at menarche (AAM) (A) and relative age at first facial hair (“younger than average”, “average”, and “older than average”) (B) among 16p11.2 BP4-BP5 deletion (DEL) and duplication (DUP) carriers and copy neutral controls in UKBB discovery cohort. (C) Prevalence of reproductive health phenotypes among 16p11.2 CNV carriers and controls in EstBB. N91 – absent, scanty, and rare menstruation; E28 – ovarian dysfunction; N83 – non-inflammatory disorders of ovary, fallopian tube, and broad ligament; N70-N77 – inflammatory diseases of female pelvic organs.

### 3.2.4. Causal inference between 16p11.2 region gene expression and age at menarche

Simple association analysis can link a genetic variant to a phenotype but is unable to suggest the underlying mechanism resulting in the phenotypic change. Since the 16p11.2 BP4-BP5 region has high gene-density with 28 overlapping genes, we hypothesised that the CNV effect on AAM could be mediated via decreased or elevated expression levels resulting from deletion or duplication of one or

multiple of these genes. Therefore, as the final step of this study, we decided to investigate if any of the genes falling into the 16p11.2 CNV region show considerable evidence for being causally linked to AAM. We conducted univariate MR which revealed three such genes: *INO80E* (standardised causal  $\beta = 0.098$ ,  $P = 1.3 \times 10^{-10}$ ), *KCTD13* ( $\beta = -0.15$ ,  $P = 4.5 \times 10^{-6}$ ) and *MAPK3* ( $\beta = 0.032$ ,  $P = 9.8 \times 10^{-10}$ ). Two of those genes, *INO80E* and *KCTD13*, also remained significant in multivariate MR. The latter has previously been identified as the key actor in 16p11.2 dosage-dependent micro- and macrocephaly (Golzio et al., 2012) and neurodevelopmental disorders (Coe et al., 2014; Marshall et al., 2017).

Still, the close proximity and overlapping eQTLs of 16p11.2 region genes hinders drawing any final conclusions from MR results alone. Furthermore, only 43% (12/28) of the genes in the region could be tested as the rest lacked strong genetic instruments in the eQTLGen dataset. To overcome these limitations, we individually modified the expression of all 16p11.2 genes in *Tg(gnrh3:egfp)* transgenic zebrafish model and monitored GnRH3 patterning *in vivo*. We found that both under- and over-expression of *ASPHD1*, a gene with unknown function not included in MR, resulted in significantly decreased GnRH3 neuron area (deletions decrease 13% compared to controls,  $P = 0.0031$ ; duplications decrease 19%,  $P < 0.0001$ ). When over-expressing *ASPHD1* and *KCTD13* simultaneously, the decrease was even more pronounced (24%,  $P < 0.0001$ ), suggesting potential interaction. According to the GTEx tissue-specific gene expression dataset (GTEx Consortium et al., 2017), *ASPHD1* is mainly expressed in brain and pituitary gland, which further supports our hypothesis.

Overall, this study presents an elegant example of an in-depth investigation of a previously unreported CNV-phenotype association. We used interdisciplinary methods to suggest *ASPHD1* (modified by *KCTD13*) as the causal gene for pubertal timing. Additional *in silico* and *in vivo* analyses can help to further clarify the exact underlying mechanism.

### 3.3. Maps of medically relevant CNV-phenotype associations and driver genes (Ref. III, IV, V)

Large association analyses have revealed the broad contribution of rare CNVs on neurodevelopmental conditions (Coe et al., 2014; Marshall et al., 2017) as well as common complex traits (Aguirre et al., 2019; Hujoel et al., 2022; Macé et al., 2017). As a result of the emergence of biobank datasets, large-scale hypothesis-free association testing in CNV GWASs has become more feasible even for low-frequency copy number alterations. This in turn has enabled the assessment of dosage sensitivity of genes overlapping rare CNVs.

This part of the thesis summarises the findings of three large multi-phenotype CNV GWASs, each analysing the effects of rare CNVs from SNP microarray-based detection methods on over 50 medically relevant phenotypes. In Ref. III

and IV, we focus on association detection with continuous traits and common diseases, respectively. We bring examples of associations to further illustrate the general patterns of CNV effects and pleiotropy. As the author of this thesis has conducted analyses on EstBB, the examples are mainly selected from the results that replicated within this dataset. Ref. V presents a CNV GWAS with disease phenotypes using the largest sample size to date, comprising nearly 1 million individuals from 17 clinical or population cohorts. The results were further leveraged to estimate probabilities of haploinsufficiency (pHaplo) and triplosensitivity (pTriplo) for all autosomal genes.

### 3.3.1. Description of cohorts and methods

In general, there are two approaches used in large GWASs to assure the validity of the results: (1) the associations are detected from the initial discovery cohort and later replicated in a second independent cohort, or (2) the associations are independently detected from multiple cohorts and the results are meta-analysed into one large study. We used the former approach in Ref. III and IV where UKBB and EstBB datasets were leveraged as discovery and replication cohorts, respectively. On the other hand, in Ref. V the CNV and phenotype counts from EstBB were combined with 16 other cohorts into a CNV association meta-analysis. As currently there are no best practice pipelines for CNV GWASs, we use diverse methodological approaches in our analyses, chosen according to the primary objectives of the study and the respective dataset. For example, in Ref. III and Ref. IV, we use the previously described approach of utilising the dosages obtained for SNP microarray probes as CNV proxies (Macé et al., 2017). The method is suitable for genome-wide analyses that incorporate intersecting CNVs with various sizes and frequencies. In Ref. V we use the sliding window and gene-based approaches, which are well suited for studying large and rare CNVs with mostly recurrent breakpoints. The full methodological overview of the three CNV GWASs as well as the specific quality control steps taken in EstBB cohort for each study are depicted in Table 3.

Additionally, for comprehensive CNV quality evaluations in Ref. III, we used 966 unrelated EstBB samples that had both genotyping array and WGS CNVs detected with GenomeSTRiP pipeline (Handsaker et al., 2015) available (a subset described in section 3.1.). For each probe, Pearson's correlation was calculated between the two CNV profiles. Due to inconsistency in CNV breakpoints and potential CNV fragmentation, proxy probes ( $r \geq 0.5$ ) from 250 kbp flanking regions were considered in case the correlation was  $<0.75$  for the original index probe.

**Table 3.** Overview of CNV GWAS methodology and quality control in EstBB dataset.

|                               | Anwerx et al., 2022 (Ref. III)                                                                                                                                                                                            | Anwerx et al., 2024 (Ref. IV)                                                                                                                                                                                   | Collins et al., 2022 (Ref. V)                                                                                                                                                                                                                                      |
|-------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| General study characteristics |                                                                                                                                                                                                                           |                                                                                                                                                                                                                 |                                                                                                                                                                                                                                                                    |
| Datasets                      | Discovery analysis was conducted in UKBB (N = ~500,000; age at recruitment 40-69; ~54% female). Replication was carried out in EstBB (N = ~200,000; age at recruitment ≥18; 66% female).                                  | Same as Ref. III                                                                                                                                                                                                | The study incorporated 17 independent datasets (including EstBB) from national biobanks to genetic testing laboratories.                                                                                                                                           |
| Cohort harmonisation          | All CNV detection and quality control steps were harmonised between UKBB and EstBB (see below).                                                                                                                           | Same as Ref. III                                                                                                                                                                                                | The CNV detection approaches and initial sample and CNV quality filtering were cohort-specific. EstBB methodology is described below. Later all CNV sets were harmonised based on CNV length and frequency.                                                        |
| Study sample size             | UKBB: 331,522; EstBB: 89,516                                                                                                                                                                                              | UKBB: 331,522; EstBB: 90,211                                                                                                                                                                                    | 950,278 (458,326 cases and 491,952 controls), out of which 78,842 were EstBB participants                                                                                                                                                                          |
| Phenotypes                    | The study included <b>57 medically relevant continuous traits</b> (46 in EstBB), including anthropometric traits, hematological measurements, blood biomarkers, and sex-specific traits.                                  | The study included <b>60 common disease phenotypes and disease burden</b> defined as the number of diagnoses (out of 60) an individual has received. In UKBB dataset, age at first diagnosis was also examined. | The study included <b>54 disease phenotypes</b> , including 24 neurological and 28 non-neurological traits, mapped to Human Phenotype Ontology (HPO) (Köhler et al., 2019) terms.                                                                                  |
| Phenotype preparation         | Prior analyses, the phenotypes were inverse normal transformed and corrected for sex (except for sex-specific traits), age, age <sup>2</sup> , genotyping batch and up to 40 first PCs calculated based on SNV genotypes. | Cases were defined based on ICD-10 codes. Controls were identified as individuals lacking the case ICD-10 codes, along with an expanded range of conditions related to the assessed disorder.                   | The final set of HPO terms were reduced to those with ≥3000 cases and ≥300 cases in at least three datasets. Out of redundant HPO terms with Jaccard index >80%, only one was retained. Controls were designated as individuals without any of the case HPO terms. |

Table 3. (continues)

|                                   | Auwerx et al., 2022 (Ref. III)                                                                                                                                                                                                                                                                                                                                                             | Auwerx et al., 2024 (Ref. IV)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | Collins et al., 2022 (Ref. V)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|-----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| CNV size                          | <10 Mbp                                                                                                                                                                                                                                                                                                                                                                                    | <10 Mbp                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | ≥100 kbp and ≤20 Mbp                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| CNV frequency                     | >0.005%                                                                                                                                                                                                                                                                                                                                                                                    | >0.01%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | <1% in all study cohorts and four published CNV catalogues (Abel et al., 2020; Almarri et al., 2020; Byrskabishop et al., 2022; Collins et al., 2020)                                                                                                                                                                                                                                                                                                                                                                                      |
| Chromosomes                       | In discovery analysis autosomes, pseudoautosomal regions and X chromosome were included. In replication, only autosomes were considered.                                                                                                                                                                                                                                                   | Same as Ref. III                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Only autosomes were included.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| Analysis models                   | deletion-only, duplication-only and mirror (deletions and duplications both affect the phenotype but in different directions) models                                                                                                                                                                                                                                                       | deletion-only, duplication-only, mirror and U-shape (deletions and duplications both affect the phenotype in the same direction) models                                                                                                                                                                                                                                                                                                                                                                                                                                                  | deletion-only and duplication-only models                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| CNV association analysis overview | CNVs were converted to probe x sample matrix, with entries corresponding to the presence (1 for duplications, -1 for deletions) or absence (0) of a CNV overlapping a particular probe in a given sample. All unique probes were imported to PLINK (v1.9) software (Chang et al., 2015), where CNV dosage were coded and analysed similarly to SNP allele dosages using linear regression. | CNVs were converted to probe x sample matrix analogously to Ref. III. All unique probes were imported to PLINK (v1.9). For each phenotype, Fisher test was used to assess dependence between disease and copy number statuses and probes with $p \leq 0.001$ and a minimum of two disease cases among CNV carriers were retained for the main analysis. Covariates were chosen from age, sex, genotyping batch and up to 40 first PCs calculated based on SNV genotypes. For each phenotype, only nominally significant ( $p \leq 0.05$ ) covariates were included in the main analysis. | Using custom R scripts, two approaches were used for CNV analyses: (a) full genome was divided into 200 kbp sliding windows with 10 kbp step size. For each window, CNVs with >50% overlap were included and one-sided Fisher's exact test was used to compare the ratios of CNV carriers between cases and controls per phenotype per cohort. The results were combined using fixed-effects meta-analysis; (b) analogously to previous analysis, exome-wide analysis was performed using gene-based approach instead of a sliding window. |

Table 3. (continues)

|                                                            | Auwerx et al., 2022 (Ref. III)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Auwerx et al., 2024 (Ref. IV)                                                                                                                                | Collins et al., 2022 (Ref. V)                                                                                                                                                                                                                                                                                                                                                     |
|------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Genome-wide significance threshold                         | $\alpha = 4.2 \times 10^{-6}$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | $\alpha = 7.5 \times 10^{-6}$ for genome-wide significance and $\alpha = 1.2 \times 10^{-7}$ for experiment-wide significance (correcting for 60 phenotypes) | $\alpha = 3.7 \times 10^{-6}$ for sliding window and $\alpha = 2.9 \times 10^{-6}$ for gene-based analysis (with $p < 0.05$ in $\geq 2$ cohorts), Benjamini-Hochberg false discovery rate $< 1\%$                                                                                                                                                                                 |
| Methodological steps and quality control for EstBB dataset |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                   |
| Sample filtering                                           | Samples with genetic and self-reported gender mismatch, and high SNV genotype missingness ( $> 2\%$ ) were excluded. To reduce the amount of somatic CNVs in the dataset, all blood malignancy cases were further eliminated. As part of CNV quality control, we excluded all samples from outlier plates (two full batches with inflated genotyping intensity and plates with $> 3$ samples with $> 200$ CNV calls) and individual samples with $> 200$ CNV calls. Finally, only unrelated samples of European ancestry were kept. |                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                   |
| CNV detection                                              | CNVs were detected using PennCNV (v1.0.5) software (Wang et al., 2007).                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                   |
| CNV defragmentation                                        | Adjacent CNVs with $\leq 20\%$ gap length were merged.                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | Same as Ref. III                                                                                                                                             | Adjacent CNVs were merged if after extending the CNVs by 25% of their original length to either side, the extended CNVs overlapped.                                                                                                                                                                                                                                               |
| CNV quality assessment overview                            | CNV quality was evaluated using a predefined quality model (defined as consensus-based CNV quality score or cQS in section 3.1. of this thesis) (Macé et al., 2016). True CNVs were defined using a stringent cut-off of $ cQS  \geq 0.5$ and lower quality calls were removed.                                                                                                                                                                                                                                                     | Same as Ref. III                                                                                                                                             | For CNV quality filtering a conventional approach where exclusion criteria were set based on PennCNV output parameters was used. Namely, only samples with $\leq 50$ raw CNV calls, as well as CNVs with a PennCNV confidence score $\geq 15$ and overlap with $\geq 10$ probes were retained. We further required the CNV length to be $\geq 25$ kbp prior defragmentation step. |

In Ref. V we specifically aimed to use the CNV-phenotype association results to assess the dosage dependence of genes within disease-causing CNVs and on a genome-wide scale. Briefly, as large CNVs encompass multiple adjacent genes, all of which are not drivers, we repurposed a Bayesian GWAS fine-mapping algorithm to detect the 95% credible set of most likely causal genes for each associated locus (Kichaev et al., 2014). In this analysis, we prioritised the genes based on the association statistics from gene-based analysis and 145 gene-level features extracted from various sources (including GTEx v7 (GTEx Consortium et al., 2017), Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al., 2015) and Swissprot (Bairoch and Apweiler, 2000)). We labelled genes as “highly confident” or “confident” if their posterior inclusion probability was  $\geq 0.8$  or 0.2–0.8, respectively. We further leveraged the gene-based association analysis statistics to compute the likelihood that the gene belonged to manually curated sets of HI or TS genes. We then modelled these likelihoods based on the 145 gene-level annotations, which enabled us to obtain pHaplo and pTriplo scores for 18,641 autosomal protein-coding genes.

### 3.3.2. Discovery and replication of CNV associations with continuous traits (Ref. III)

When studying the landscape of detected CNV regions in UKBB dataset, we observed that 82% of probes had an overlapping CNV in at least one individual, while 81% of those probes had CNV frequency  $\leq 0.005\%$ . On the other hand, 39% of all samples had at least one high-confidence CNV and 10% had more than one. These results further illustrate the paradigm that, although rare individually, CNVs collectively are quite common, making them even more important subject for closer studying.

All unique probes passing the CNV frequency cut-off  $> 0.005\%$  were included in the CNV GWAS across 57 medically relevant continuous traits in UKBB dataset. We detected 86, 68 and 50 significantly associated CNV-phenotype pairs for mirror, deletion-only and duplication-only models, respectively, spanning 47 phenotypes (Figure 7). These results could be combined into 131 independent associations as findings from the mirror model often (in 73% cases) replicated with either the deletion-only or the duplication-only model type. Out of these, 61 (47%) had data available for replication analysis in EstBB cohort. We detected 6 signals that replicated with Bonferroni corrected  $P < 0.05/61 = 8.2 \times 10^{-4}$  (Table 4) and additional 22 nominally significant signals, which was 7.2 times more than expected by chance (two-sided binomial test:  $P = 7.8 \times 10^{-14}$ ). The effect sizes were concordant between the two datasets indicating the robustness of the associations (Figure 8).

The quality assessment for the trait-associated CNV regions based on the comparison to WGS CNVs showed high sensitivity and specificity. We found high concordance ( $r \geq 0.75$ ) for the calling profiles of 85% of the probes. Out of the false positives (only detected by PennCNV), 92% correlated ( $r \geq 0.5$ ) with a

highly concordant probe nearby, indicating potential CNV fragmentation in WGS data. Only 7% of the probes were labelled as false negatives (only detected in WGS).

We observed highly pleiotropic effects in UKBB for some of the well-established CNV regions including the 16p11.2 BP4-BP5 CNV region that we found associated to 26 traits. For example, the 16p11.2 deletion was linked to levels of insulin-like growth factor 1 (IGF-1) and insulin resistance, both previously connected to short stature and increased adipose mass (Kargi and Merriam, 2013), which are known characteristics of 16p11.2 deletion carriers. Four of the 16p11.2 BP4-BP5 associations (BMI, WHR, height, and weight) replicated in EstBB, while an additional two (age at menarche and levels of alanine aminotransferase) replicated with nominal significance.

Several of the trait-associated CNVs overlapped with known Mendelian disorder loci. For example, the 17q12 deletion (OMIM: 614527) is known to cause highly penetrant autosomal dominant renal cysts and diabetes syndrome, characterised by renal defects. In UKBB dataset, we found that the 17q12 duplication (OMIM: 614526) is associated with kidney damage biomarkers (cystatin C:  $\beta = 0.15 \text{ mg/L}$ ,  $P = 4.2 \times 10^{-17}$ ; serum creatinine (SCr):  $\beta = 13.0 \text{ } \mu\text{mol/L}$ ,  $P = 2.7 \times 10^{-16}$ ; serum urea:  $\beta = 0.93 \text{ mmol/L}$ ,  $P = 9.1 \times 10^{-10}$ ). Association with SCr reached nominal significance in EstBB ( $P = 0.0015$ ). While deletion carriers were excluded from the analyses due to low variant frequency ( $n = 6$  in UKBB), they did exhibit elevated levels of all three kidney biomarkers, suggesting that both deletions and duplications cause renal abnormalities.

Another example of new insights for a well-established CNV involves the 17p12 duplication, a variant linked to CMT disease type 1A (OMIM: 118220) that causes nerve damage and decreased muscle mass. We observed an association to hand grip strength ( $\beta = -9.8 \text{ kg}$ ,  $P = 4.1 \times 10^{-39}$ ) and lower levels of SCr ( $\beta = -13.8 \text{ } \mu\text{mol/L}$ ,  $P = 6.5 \times 10^{-18}$ ) in UKBB. The latter also replicated in EstBB ( $\beta = -94.1 \text{ } \mu\text{mol/L}$ ,  $P = 5.2 \times 10^{-4}$ ), while hand grip strength was not available. From previous studies it is known that SCr levels can be used as a proxy for muscle mass (Thongprayoon et al., 2016). We observed that, despite the clear aetiology, only 36% of duplication carriers had previously been diagnosed with CMT or relevant neuropathies. This subgroup also presented a more extreme decrease of grip strength compared to the rest of the duplication carriers (one-sided T-test:  $P = 0.005$ ). Therefore, our results suggest an incomplete penetrance for the 17p12 duplication in the general population.

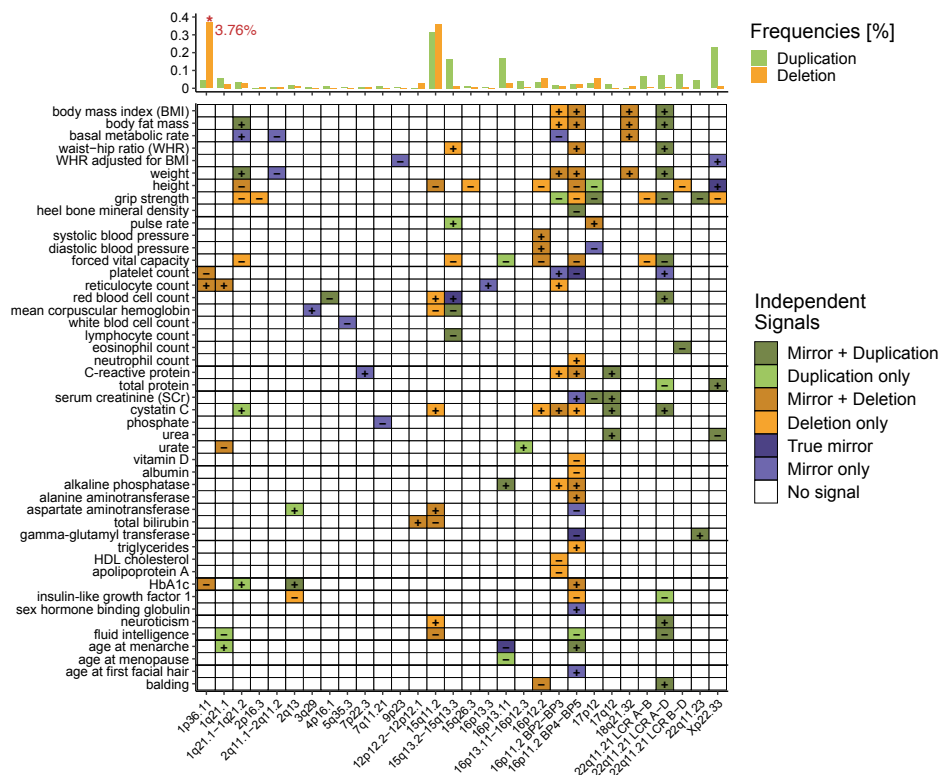


**Figure 7.** Independent genome-wide significant associations between CNV (x axis) and traits (y axis). (+) and (-) signs mark the direction of the association. (Ref. III, Figure 2D)

**Table 4.** Association statistics ( $\beta$  and P-values) for CNV-trait associations that replicated in EstBB with Bonferroni corrected significance threshold ( $P < 8.2 \times 10^{-4}$ ). Row numbers correspond to labels in Figure 8. (Adapted from Ref. III, Figure 3B)

|   | CNV region<br>(model type)  | Trait  | UKBB    |                       | EstBB   |                      |
|---|-----------------------------|--------|---------|-----------------------|---------|----------------------|
|   |                             |        | $\beta$ | P                     | $\beta$ | P                    |
| 1 | 1q21.1-1q21.2<br>(mirror)   | height | 4.3     | $7.8 \times 10^{-22}$ | 4.6     | $3.3 \times 10^{-5}$ |
| 2 |                             | BMI    | 6.1     | $3.6 \times 10^{-29}$ | 7.0     | $3.4 \times 10^{-7}$ |
| 3 | 16p11.2 BP4-BP5             | WHR    | 0.1     | $1.5 \times 10^{-27}$ | 0.1     | $7.6 \times 10^{-6}$ |
| 4 | (deletion-only)             | height | -8.2    | $3.3 \times 10^{-27}$ | -9.0    | $2.5 \times 10^{-6}$ |
| 5 |                             | weight | 11.1    | $5.0 \times 10^{-12}$ | 15.2    | $5.2 \times 10^{-4}$ |
| 6 | 17p12<br>(duplication-only) | SCr    | -13.8   | $6.5 \times 10^{-18}$ | -94.1   | $5.2 \times 10^{-4}$ |

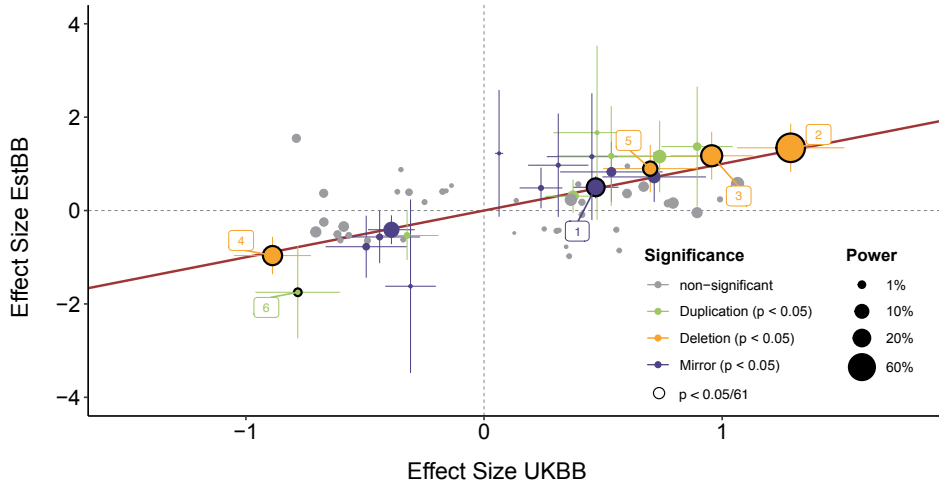


**Figure 8.** EstBB versus UKBB standardised effect sizes with 95% confidence intervals. Point size indicates power at  $\alpha = 0.05/61$ . Labels correspond to row numbers in Table 4. (Adapted from Ref. III, Figure 3A)

An additional example involves a 16p13.11 CNV region and female reproductive timing. In UKBB we detected an association between the deletion and increased AAM ( $\beta = 1.1$  years,  $P = 3.6 \times 10^{-7}$ ), while the duplication was linked to decreased AAM ( $\beta = -0.6$  years,  $P = 2.0 \times 10^{-10}$ ) and age at menopause ( $\beta = -1.8$  years,  $P = 1.7 \times 10^{-6}$ ). The associations with duplication were replicated in EstBB with nominal significance (menarche:  $\beta = -0.6$  years,  $P = 1.8 \times 10^{-2}$ ; menopause:  $\beta = -2.6$  years,  $P = 4.5 \times 10^{-2}$ ). Previously, a SNP association to AAM has been observed in this locus (Day et al., 2017). A potential causal gene, however, remained unclear. In our study, we pinpointed *MARF1*, a gene overlapping the CNV but not nearest to the SNP GWAS top hit. Based on the Genome Aggregation Database (gnomAD) (gnomAD, 2024) the gene is under evolutionary constraint and literature also supports its female-specific effect on reproduction across different species including humans (Katari et al., 2018; Kawaguchi et al., 2020; Su, Sugiura, et al., 2012; Su, Sun, et al., 2012).

### 3.3.3. Discovery and replication of CNV associations with common diseases (Ref. IV)

As a logical extension of CNV associations with continuous traits, we proceeded with association testing for a curated set of 60 common diseases using the same CNV datasets. In the UKBB discovery set we detected 40, 41, 21, and 38 independent genome-wide significant signals for mirror, U-shape (similar effect in both deletions and duplications), duplication-only and deletion-only model types, respectively, which we combined into 73 unique disease risk-increasing or

burden associations (32 when using the stringent experiment-wide significance threshold) (Figure 9). Altogether, these associations mapped to 45 unique CNV regions. We found that the genes within these disease-causing CNVs had stronger intolerance towards LoF mutations compared to “background genes” affected by CNVs not associated with any disease (probability of LoF intolerance: Wilcoxon  $P = 1.3 \times 10^{-4}$ ; loss of function observed/expected upper bound fraction (LOEUF) (Karczewski et al., 2020):  $P = 1.9 \times 10^{-7}$ ). Additionally, the genes overlapping disease-causing CNVs were more likely to be triplosensitive (Wilcoxon  $P = 9.0 \times 10^{-19}$  for genes in duplication regions,  $P = 1.0 \times 10^{-23}$  for genes in deletion regions).

We sought to replicate our findings in EstBB cohort. We managed to test 49 out of 73 associations. We detected 3 signals that replicated with Bonferroni corrected  $P \leq 0.05/49 = 1.0 \times 10^{-3}$  significance and an additional 4 nominally significant signals ( $P \leq 0.05$ ) (Table 5). Due to over three times smaller sample size in EstBB compared to UKBB, the replication power was strongly limited. Still, the results corresponded to 2.9-fold (two-sided binomial test:  $P = 0.011$ ) and 16.3-fold ( $P = 1.1 \times 10^{-4}$ ) enrichment for replication at  $P \leq 0.05$  and  $P \leq 5.0 \times 10^{-3}$ , respectively, compared to what would be expected by chance. Out of the replicated findings, five had additional support from previously published SNV or CNV association studies (including Ref. III) for related phenotypes and two overlapped with relevant morbid OMIM genes.

Out of 45 associated CNV regions, nine mapped to known GD (using a set compiled by Crawford et al. (2019)). For seven of these CNVs, we found associations with multiple phenotypes in UKBB. Similarly to Ref. III, the highest pleiotropy was observed for the 16p11.2 BP4-BP5 deletion, which we could link to higher disease burden as well as increased risk of 12 individual diseases across multiple organ systems. In EstBB, the associations could be replicated for three disease phenotypes – acute kidney injury (AKI), hypertension, and chronic kidney disease (CKD) (Table 5). As 16p11.2 BP4-BP5 has been linked with severe obesity (Walters et al., 2010), which is a risk factor in itself, we further tested if the CNV-disease association could be driven by BMI. The CNV effect on the risk for five diseases remained significant after adjusting the model for BMI.

Previously in Ref. III, we demonstrated that 17q12 deletions and duplications were associated with kidney damage markers. To further emphasise the link between the 17q12 CNV region and renal problems, we showed in UKBB, and replicated in EstBB, that both deletions and duplications increase the risk for CKD (Table 5). We observed that the prevalence of CKD was 33.3% and 16.9% in deletion and duplication carriers, respectively, compared to 4.4% in copy neutral individuals. Moreover, the presence of the CNV lowered the age at CKD onset (Cox proportional hazards model: hazard ratio = 4.6,  $P = 1.3 \times 10^{-7}$ ). A recent clinical study agrees that the 17q12 CNV is the most common GD aetiology among CKD patients (Verbitsky et al., 2023).

We studied the CNV regions associated to seemingly unrelated phenotypes to dissect their pattern of pleiotropy. For example, neuropsychiatric and developmental conditions have been linked to two deletions in the 15q13 region (OMIM: 612001). A longer deletion, defined as BP4-BP5, fully encompasses the shorter D-CHRNA7-BP5 deletion. Also, 15q13 duplication carriers have been shown to exhibit psychiatric disease (van Bon et al., 2009). Based on the overlap with the short deletion, *CHRNA7* is proposed as the causal gene (Gillentine and Schaaf, 2015). In our CNV GWAS we detected associations in UKBB with 15q13 BP4-BP5 duplication and three non-neurological disorders: AKI (Table 5), haemorrhagic stroke (odds ratio ( $OR$ ) = 7.5,  $P = 4.3 \times 10^{-6}$ ), and anaemia ( $OR = 4.9$ ,  $P = 3.2 \times 10^{-6}$ ). However, no associations were found for D-CHRNA7-BP5 duplication. The results suggest that despite mapping to the same CNV locus, the causes for neurological and non-neurological conditions lie in different genes.

**Figure 9.** CNV-disease association map. Colour indicates the best model type. Point size and black contours indicate the level of EstBB replication significance. (Adapted from Ref. IV, Figure 2)

**Table 5.** Association statistics (OR and P-values) for CNV-trait associations that replicated in EstBB ( $P < 0.05$ ). Abbreviations: AKI, acute kidney injury; CKD, chronic kidney disease; HTN, hypertension; PD, Parkinson’s disease. (Adapted from Ref. IV, Figure 3B)

|                      | CNV region<br>(model type)            | Trait  | UKBB |                       | EstBB |                      |
|----------------------|---------------------------------------|--------|------|-----------------------|-------|----------------------|
|                      |                                       |        | OR   | P                     | OR    | P                    |
| Bonferroni           | 16p11.2 BP4-BP5<br>(deletion-only)    | AKI    | 16   | $5.6 \times 10^{-20}$ | 42    | $2.1 \times 10^{-4}$ |
|                      | 15q13.2-15q13.3<br>(duplication-only) | AKI    | 4.6  | $7.1 \times 10^{-7}$  | 12    | $2.7 \times 10^{-4}$ |
|                      | 17q12<br>(U-shape)                    | CKD    | 6.5  | $5.9 \times 10^{-9}$  | 18    | $8.6 \times 10^{-4}$ |
| Nominal significance | 16p11.2 BP4-BP5<br>(deletion-only)    | HTN    | 4.6  | $3.4 \times 10^{-7}$  | 5.0   | $4.0 \times 10^{-3}$ |
|                      | 15q13.2-15q13.3<br>(mirror)           | asthma | 0.17 | $1.2 \times 10^{-6}$  | 0.29  | $6.2 \times 10^{-3}$ |
|                      | 16p11.2 BP4-BP5<br>(U-shape)          | CKD    | 4.4  | $2.7 \times 10^{-8}$  | 8.4   | 0.033                |
|                      | 11p15.4<br>(mirror)                   | PD     | 0.05 | $1.9 \times 10^{-7}$  | 0.05  | 0.047                |

Altogether, the series of examples here and in section 3.3.2 illustrate several interesting phenomena observable from CNV-phenotype associations. First, it further emphasises the high pleiotropy and variable expressivity of GD loci. We demonstrated that in the general population pathogenic CNVs, such as CKD-associated deletion and duplication of 17q12, affect relevant biomarkers without necessarily causing clinically diagnosable disorders. This, and higher statistical power due to the continuous nature of biomarker measurements, make them excellent proxies to study various underlying pathologies in any GWAS analysis. Secondly, CNV-phenotype associations exhibit substantial overlap with previously detected SNV GWAS loci, suggesting that both mutational classes should be studied in combination in order to detect the exact genetic mechanisms and causal genes. We observed that genes within phenotype-associated CNV regions tend to have higher evolutionary constraint and dosage sensitivity compared to background genes. This phenomenon is further studied in the next section of the thesis.

### 3.3.4. The map of dosage sensitive (DS) genes across human genome (Ref. V)

As discussed in section 1.1.3., the main mechanisms of CNV effect on phenotype include the loss or complete copy gain of DS genes caused by deletions or duplications, respectively. In this study, we aimed to annotate all protein-coding genes with probabilities of haploinsufficiency and triplosensitivity and provide a map of DS genes for the wider scientific community.

As the initial step, we tested for associations with rare (<1% frequency) CNVs and 54 disease phenotypes in a meta-analysis cohort of 950,278 individuals. First, we used the approach of a 200kb sliding window in 10kb steps and performed the association analysis per phenotype for CNVs overlapping each window (see Table 3). The significant results were compiled to 795 unique CNV-phenotype associations that mapped to 163 distinct rare CNVs (69 deletions and 94 duplications). We compared our results to an established set of 95 GDs and found that 55.8% (53/95) of them were present among our disease-causing CNVs. However, when testing the 95 GDs selected *a priori* and using the Bonferroni-corrected significance cut-off  $P \leq 0.05/95 = 5.26 \times 10^{-4}$ , we retained another 15 GDs associated with our phenotypes. Those were included in our set of disease-causing CNVs, totalling 178 regions carried forward to subsequent analyses. We observed that these regions were significantly more likely to overlap with phenotype-matched OMIM genes (one-tailed permutation test:  $P = 3.8 \times 10^{-4}$ ) and were enriched for genes under strong mutational constraint ( $P = 0.006$ ), indicating a presence of at least one DS driver gene. This was further supported by the excess of damaging *de novo* SNVs (Fu et al., 2022; Kaplanis et al., 2020) within associated CNV regions, which often concentrated almost entirely in just one or two genes.

Moreover, the gene-based association analysis was conducted. It resulted in 5,680 significant associations across 739 unique genes. None of the associated loci were novel compared to the sliding window approach. However, 32 additional CNV-phenotype associations were discovered. Leveraging the gene-based association statistics and 145 gene-level features, we were able to fine-map 90 (55.2%) out of 163 rare associated CNV loci from the original analysis to a set of candidate driver genes.

Finally, we developed pHaplo and pTriplo scores to predict haploinsufficiency and triplosensitivity of 18,641 autosomal genes based on 145 gene-level features (see section 3.3.1.). We estimated the score cut-offs such that the CNV effect sizes would be comparable with LoF effects in genes that are well-established as constrained for protein-truncating variants (Karczewski et al., 2020). Based on these cut-offs (pHaplo  $\geq 0.86$  and pTriplo  $\geq 0.94$ ) we defined 2,987 HI and 1,559 TS genes. We observed that bidirectionally DS (both HI and TS) genes are mainly defined by their evolutionary constraint. In general, HI genes appear to be larger, further from other genes and with a high abundance of *cis* regulatory elements, whereas TS genes were smaller, GC-rich, and located in gene-dense regions.

In summary, the study presents the largest CNV GWAS to date and a preliminary insight into the dosage sensitivity patterns of the genes on a genome-wide scale. Both resources are valuable additions to CNV-phenotype association maps described in Ref. III and IV, all of which are poised to be invaluable in advancing research and practice in human disease and clinical genetics.

## CONCLUSIONS

In order to dissect the genetic mechanisms underlying different phenotypes, it is imperative to investigate variant types beyond solely SNVs and small indels. Copy number alterations detected from microarray intensities provide an excellent resource to study effects of rare CNVs on human traits and common diseases. Nevertheless, CNV studies remain challenging and methodological improvements for reliable CNV detection, quality control, association testing, and follow-up analyses remain a work in progress.

The main conclusions drawn from this thesis are as follows:

- Multi-omics data can be leveraged to assess CNV quality. The quality metrics based on three independent data layers – WGS, gene expression levels, and methylation array intensities – all exhibited high concordance and ability to distinguish between true and false calls from popular detection software PennCNV. Furthermore, the metrics can be modelled based on PennCNV output parameters. The resulting omics-informed quality score (OQS) improved statistical power in CNV-phenotype association studies, showing up to 34% and 55% increase in variance explained compared to raw PennCNV calls and a previously published CNV quality score, respectively.
- A well-established pathogenic 16p11.2 BP4-BP5 CNV region has a mirroring effect on pubertal timing. In females, deletion and duplication carriers had 1.5 years earlier and 1.5 years later age at menarche, respectively, compared to copy neutral controls. Similar trends were observed for male pubertal timing, indicating that the effect is not sex-specific. *In silico* and *in vivo* analyses suggested potential causal role of *ASPHD1* and *KCTD13* genes in the 16p11.2 region.
- Rare CNVs have significant impact on medically relevant continuous measurements and common diseases. In the largest multi-phenotype CNV GWAS to date, we observed phenotype associations with 163 distinct CNVs. The results aid in dissecting novel associations with highly pleiotropic genomic disorders and variable expressivity of CNVs overlapping loci linked to monogenic conditions. The associated CNVs encompassing known SNV GWAS loci help to detect the exact genetic mechanisms and pinpoint the potential causal genes.
- Disease-causing CNVs are enriched with DS genes. Furthermore, by modelling the likelihood of haploinsufficiency and triplosensitivity based on features of these genes, it is possible to assess the dosage sensitivity of all genes on a genome-wide scale. In total, we defined 2,987 and 1,559 likely HI and TS autosomal genes, respectively, which is a valuable resource for human disease research and clinical interpretation.

## SUMMARY IN ESTONIAN

### Arengud mikrokiibi-põhiste koopiaarvu variatsioonide tuvastamises ja fenotüüpidega seostamises

Koopiaarvu variatsioonid (ingl. k. *copy number variation* ehk CNV) on piirkonnad DNA-s, mille koopiaarv erineb tavapärasele diploidsele genoomile omasest kahest koopiast. CNV-de tekkimise tulemusel on genoomipiirkonnad kadunud (nimetatakse deletsiooniks) või mitmekordistunud (nimetatakse duplikatsiooniks). Kuigi algselt seostati CNV-sid eelkõige raskekujuliste arenguhäiretega, siis nüüdseks on selge, et CNV piirkondi leidub kõigi inimeste DNA-s ning nad mõjutavad oluliselt suuremat osa genoomist ning omavad ka suuremat efekti inimese tervisele ja muudele tunnustele kui palju-uuritud ühenukleotiidsed polümorfismid (ingl. k. *single nucleotide polymorphism* ehk SNP). Sellest hoolimata on CNV-sid ja nende seoseid inimese fenotüübiga vähe uuritud, põhjuseks CNV-de määramise ja analüüsimise meetodika keerukus. Seetõttu on oluline nii CNV-de tuvastamise meetodika arendamine kui ka uute CNV-de ja fenotüüpide vaheliste seoste leidmine.

Toetudes teaduskirjandusele, annab väitekirja esimene osa ülevaate CNV-de jaotusest inimese genoomis, CNV-de määramisest ja mõjust inimese fenotüübile. Eraldi peatükk on pühendatud CNV-de määramisele SNP mikrokiibi intensiivsuse andmetelt. Suurte olemasolevate andmemahtude tõttu on mikrokiipide kasutamine üheks enamlevinud CNV-de tuvastamise meetodikaks. Populaarsed koopiaarvu määramise programmid, näiteks PennCNV, kasutavad eelkõige kahte mõõdetavat suurust iga kiibi markeri kohta: oodatava ja empiirilise signaali-intensiivsuse logaritmitud suhet (ingl. k. *Log R Ratio* ehk LRR) ja ühe SNP alleeli intensiivsuse osakaalu teise alleeli suhtes (ingl. k. *B Allele Frequency* ehk BAF). Muutused LRR ja BAF signaalides viitavad CNV olemasolule vaadeldavas genoomipiirkonnas. Samas on kiibipõhisel CNV-de määramisel mitmeid olulisi puudujääke. Näiteks ei saa kiibiga määrata lühikesi CNV-sid, millel pole kiibi markeritega piisavalt ülekattet. Samuti ei võimalda signaaliintensiivsuste analüüs täpselt määrata suurte (>4) koopiaarvudega CNV-sid. Need puudused suurendavad valenegatiivsete CNV-de hulka, kuid kiibipõhisel tuvastamisel on märkimisväärselt probleemiks ka valepositiivsete CNV-de osakaal. Seetõttu vajavad CNV andmed enne edasise analüüsi põhjalikku kvaliteedikontrolli. Enamasti kasutatakse madala kvaliteediga proovide või CNV-de välja filtreerimiseks (küllaltki juhuslikult määratud) piirväärtuseid LRR ja BAF signaalide koondstatistikutele (nt keskmist ja standardhälvet) ja muudele parameetritele, näiteks CNV pikkus, CNV-de arv proovi kohta jne.

Kõigist puudustest hoolimata on aga mikrokiibi intensiivsustest võimalik edukalt määrata eelkõige pikki harvaesinevaid CNV-sid, millel on inimese tervisele suur (enamasti negatiivne) mõju. Teaduskirjanduse ülevaate viimane peatükk teeb kokkuvõtte seni läbiviidud seoseuuringutest harvade CNV-dega, mille tulemuste põhjal saab seda mõju täpsemalt kirjeldada. Valdav enamus sellistest uuringutest keskendub CNV-de seostele erinevate neuroarenguliste ja -psühhiaatriliste

haigustega, näiteks autism, skisofreenia ja teised. Käesolevas väitekirjas on eriline tähelepanu pühendatud aga ülegenoomsetele CNV seoseuuringutele (ingl. k. *genome-wide CNV association study* ehk CNV GWAS) suurtel biopankade andmetel, mida on seni läbi viidud vaid loetud hulk. Nende uuringute tulemused rõhutavad CNV-de olulisust fenotüübilise varieeruvuse kujunemisel ja haiguste põhjustamisel.

Väitekirja teine osa annab ülevaate viiest doktoritöö osaks olevast teadusartiklist. Esimene artikkel keskendub CNV kvaliteedihinnangute parendamisele, et tõsta edasistes analüüsides kasutatavate CNV-de usaldusväärsust ja seeläbi statistilist võimsust fenotüübiliste seoste leidmiseks. Kasutasime kiibipõhiste CNV-de kvaliteedi hindamiseks kolme sõltumatut nõ „oomika“ andmekihti: täisgenoomide järjestustest tuvastatud CNV-sid, geeniekspressiooni ja metülatsioonikiibi intensiivsuseid. Analüüs baseerus eeldusel, et tõesed mikrokiibilt tuvastatud CNV-d on määratavad ka sama indiviidi täisgenoomist ning et nad mõjutavad geeniekspressiooni ja metülatsioon-intensiivsuseid oodatavas suunas, st deletsioonid madalamaks ja duplikatsioonid kõrgemaks. Tulemuseks saime iga CNV ja oomika andmekihi kohta ühe numbrilise suuruse, mis peegeldas vastava CNV usaldusväärsust. Nägime, et kõik kasutatud andmestikud sobisid tõeste ja valepositiivsete CNV-de eraldamiseks ning olid üksteisega tugevas korrelatsioonis (sõltuvalt andmestikust kuni  $R = 0.8$ ). Paljudel biopankadel ja andmekogudel aga ülalmainitud lisa-andmekihid puuduvad ja olemas on vaid mikrokiibi andmed. Et ka sellisel juhul saaks CNV-de usaldusväärsust kirjeldada, lõime oma andmeid kasutades CNV kvaliteedimudeli. Selleks hindasime oomika andmetelt arvatud CNV kvaliteediväärtuseid, kasutades sisendina kiibipõhiste CNV-de parameetreid (CNV pikkus, ülekate kiibi markeritega, LRR/BAF signaali statistikud jne). Saadud mudeli testimisel sõltumatul Eesti Biopanga andmestikul nägime, et meie kvaliteedimudeli kasutamine suurendas seoseanalüüsi võimsust kuni 55% võrreldes varasemalt avaldatud CNV kvaliteedimudeliga. Meie tulemused andsid kinnitust, et eri andmekihtide informatsiooni kaasamine CNV-de tuvastamise protsessi aitab suurendada määratud CNV-de usaldusväärsust, mis omakorda võimaldab leida uusi CNV-de ja fenotüüpide vahelisi seoseid.

Teises artiklis uurisime ühte konkreetset teadaolevat haigusseoselist CNV-d, mis paikneb 16p11.2 genoomiregioonis, ning selle varem avastamata seost puberteedi alguse vanusega. Leidsime, et 16p11.2 deletsioon põhjustab tüdrukutel keskmiselt 1,5 aastat varasemat ja duplikatsioon 1,5 aastat hilisemat menstruatsiooni algust. Sarnast trendi nägime ka poiste puhul, kus deletsioonikandjad raporteerisid võrreldes eakaaslastega varasemat ja duplikatsioonikandjad vastavalt hilisemat häälemurret ja karvakasvu algust näopiirkonnas. Lisaks oli meie eesmärgiks tuvastada 16p11.2 piirkonnast puberteediiga põhjuslikult mõjutav geen. Selleks viisime läbi arvutusliku Mendeli randomiseerimise analüüsi ning *in vivo* analüüsi sebrakalas. Nende kahe analüüsimeetodika tulemusel järeldasime, et tõenäoliselt on fenotüüp põhjustatud kahe geeni, *ASPHD1* ja *KCTD13* interaktsioonist. Artikkel on elegantne näide CNV seose- ja sellele järgnevatest

jätkuanalüüsides ning loob aluse edasisteks uuringuteks inimese reproduktiivtervise ja 16p11.2 genoomipiirkonna sõltuvustest.

Viimased kolm artiklit kirjeldavad CNV GWAS analüüse, mille tulemusena tuvastasime kokku mitusada seost harvaesinevate CNV-de ning erinevate haiguste ja meditsiiniliste mõõtmiste (näiteks antropomeetrilised tunnused või vereproovist määratavad kontsentratsioonid) vahel. Meie tulemused näitasid, et CNV efektid on tihti pleiotroopsed, st seotud mitmete väga erinevate fenotüüpidega. Heaks näiteks on juba varem mainitud 16p11.2 CNV piirkond, mida me seostasime vastavalt 12 haigus- ja 26 pideva mõõdetava tunnusega. Lisaks avastasime haigusseoselisi CNV-sid genoomipiirkondadest, mida on varasemalt analüüsitud SNP GWAS-is või mida on seostatud monogeensete haigustega. CNV-de uurimine sellistes regioonides aitab tuvastada haiguste tekke aluseks olevat geneetilist arhitektuuri ja mehhanisme. CNV GWAS tulemused võimaldasid meil ka hinnata geenide doosituvust, et tuvastada olukordi, kus deletsiooni tulemusel vähenenud või duplikatsiooni tulemusel suurenenud geenikoopiate arv ei ole optimaalne organismi edukaks toimimiseks. Kokku leidsime me 2 987 deletsiooni ja 1 559 duplikatsiooni mitte-taluvat valku kodeerivat geeni, mis on suur edasimineku seni teadaolevate doositundlike geenide arvust (ClinGen andmebaasi alusel on vastavad geenide arvud 423 ja 24). Nii CNV-de ja fenotüüpide vaheliste seoste kui ka doositundlike geenide kaardistamine on oluline samm edasisteks teadusuuringuteks ja CNV efektide kliiniliseks interpreteerimiseks.

Kokkuvõttes puudutab antud väitekiri kolme olulist CNV-uuringute aspekti: analüüsile eelnevat CNV-de usaldusväarsuse hindamist, CNV-de seoseanalüüse ning viimaste tulemustel põhinevaid jätkuanalüüse põhjuslike geenide tuvastamiseks. Käesolevast tööst, nagu ka mitmetest teistest teaduspublikatsioonidest, johtub, et vaid SNP-del baseeruvad seoseanalüüsid ei suuda erinevate haiguste ja muude tunnuste geneetilist tagapõhja täielikult määrata ning muude variatsioonitüüpide, sh CNV informatsiooni kaasamine on tänapäeva geneetilises epidemiooloogias äärmiselt oluline samm.

## REFERENCES

- Abel HJ, Larson DE, Regier AA, et al. (2020) Mapping and characterization of structural variation in 17,795 human genomes. *Nature* 583(7814): 83–89.
- Abraham E, Palevitch O, Gothliff Y, et al. (2009) The zebrafish as a model system for forebrain GnRH neuronal development. *General and comparative endocrinology* 164(2–3): 151–160.
- Abyzov A, Urban AE, Snyder M, et al. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research* 21(6): 974–984.
- Aguirre M, Rivas MA and Priest J (2019) Phenome-wide Burden of Copy-Number Variation in the UK Biobank. *American journal of human genetics* 105(2): 373–383.
- Alkan C, Coe BP and Eichler EE (2011) Genome structural variation discovery and genotyping. *Nature reviews. Genetics* 12(5): 363–376.
- Almarri MA, Bergström A, Prado-Martinez J, et al. (2020) Population Structure, Stratification, and Introgression of Human Structural Variation. *Cell* 182(1): 189–199.e15.
- Alwan H, Pruijm M, Ponte B, et al. (2014) Epidemiology of masked and white-coat hypertension: the family-based SKIPOGH study. *PloS one* 9(3): e92522.
- Babadi M, Fu JM, Lee SK, et al. (2023) GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nature genetics* 55(9): 1589–1597.
- Bairoch A and Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* 28(1): 45–48.
- Balagué-Dobón L, Cáceres A and González JR (2022) Fully exploiting SNP arrays: a systematic review on the tools to extract underlying genomic structure. *Briefings in bioinformatics* 23(2).
- Beroukhim R, Mermel CH, Porter D, et al. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature* 463(7283): 899–905.
- Billingsley KJ, Ding J, Jerez PA, et al. (2023) Genome-Wide Analysis of Structural Variants in Parkinson Disease. *Annals of neurology* 93(5): 1012–1022.
- Buizer-Voskamp JE, Muntjewerff J-W, Genetic Risk and Outcome in Psychosis (GROUP) Consortium Members, et al. (2011) Genome-wide analysis shows increased frequency of copy number variation deletions in Dutch schizophrenia patients. *Biological psychiatry* 70(7): 655–662.
- Bycroft C, Freeman C, Petkova D, et al. (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562(7726): 203–209.
- Byrska-Bishop M, Evani US, Zhao X, et al. (2022) High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* 185(18): 3426–3440.e19.
- Campbell MC and Tishkoff SA (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annual review of genomics and human genetics* 9: 403–433.
- Chang CC, Chow CC, Tellier LC, et al. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4: 7.
- Chen L, Abel HJ, Das I, et al. (2021) Association of structural variation with cardiometabolic traits in Finns. *American journal of human genetics* 108(4): 583–596.
- Chen X, Schulz-Trieglaff O, Shaw R, et al. (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. Available at: <http://dx.doi.org/10.1093/bioinformatics/btv710>.

- Chettier R, Ward K and Albertsen HM (2014) Endometriosis is associated with rare copy number variants. *PLoS one* 9(8): e103968.
- ClinGen Dosage Sensitivity (2024). Available at: <https://dosage.clinicalgenome.org> (accessed January 2024).
- Coe BP, Witherspoon K, Rosenfeld JA, et al. (2014) Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nature genetics* 46(10): 1063–1071.
- Colella S, Yau C, Taylor JM, et al. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic acids research* 35(6): 2013–2025.
- Collins RL, Brand H, Karczewski KJ, et al. (2020) A structural variation reference for medical and population genetics. *Nature* 581(7809): 444–451.
- Conrad DF, Pinto D, Redon R, et al. (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464(7289): 704–712.
- Cooper GM, Coe BP, Girirajan S, et al. (2011) A copy number variation morbidity map of developmental delay. *Nature genetics* 43(9): 838–846.
- Coutelier M, Holtgrewe M, Jäger M, et al. (2022) Combining callers improves the detection of copy number variants from whole-genome sequencing. *European journal of human genetics: EJHG* 30(2): 178–186.
- Crawford K, Bracher-Smith M, Owen D, et al. (2019) Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *Journal of medical genetics* 56(3): 131–138.
- Curry CJ, Rosenfeld JA, Grant E, et al. (2013) The duplication 17p13.3 phenotype: analysis of 21 families delineates developmental, behavioral and brain abnormalities, and rare variant phenotypes. *American journal of medical genetics. Part A* 161A(8): 1833–1852.
- Day FR, Thompson DJ, Helgason H, et al. (2017) Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nature genetics*. forthcoming April 24, 2017. DOI: 10.1038/ng.3841.
- DECIPHER CNV Syndromes (2024). Available at: <https://www.deciphergenomics.org/disorders/syndromes/list> (accessed 2017).
- Deeb SS (2005) The molecular basis of variation in human color vision. *Clinical genetics* 67(5): 369–377.
- Diskin SJ, Li M, Hou C, et al. (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic acids research* 36(19): e126.
- Dowjat K and Włodarska I (1981) G-banding patterns in mouse lymphoblastic leukemia L1210. *Journal of the National Cancer Institute* 66(1): 177–182.
- Ebert P, Audano PA, Zhu Q, et al. (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372(6537).
- Escamilla CO, Filonova I, Walker AK, et al. (2017) Kctd13 deletion reduces synaptic transmission via increased RhoA. *Nature* 551(7679): 227–231.
- Feber A, Guilhamon P, Lechner M, et al. (2014) Using high-density DNA methylation arrays to profile copy number alterations. *Genome biology* 15(2): R30.
- Fetit R, Price DJ, Lawrie SM, et al. (2020) Understanding the clinical manifestations of 16p11.2 deletion syndrome: a series of developmental case reports in children. *Psychiatric genetics* 30(5): 136–140.
- Feuk L, MacDonald JR, Tang T, et al. (2005) Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS genetics* 1(4): e56.

- Firth HV, Richards SM, Bevan AP, et al. (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American journal of human genetics* 84(4): 524–533.
- Fitzgerald T and Birney E (2022) CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank. *Cell genomics* 2(8): 100167.
- Fromer M, Moran JL, Chambert K, et al. (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *American journal of human genetics* 91(4): 597–607.
- Fu JM, Satterstrom FK, Peng M, et al. (2022) Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nature genetics* 54(9): 1320–1331.
- Gaedigk A, Fuhr U, Johnson C, et al. (2010) CYP2D7–2D6 hybrid tandems: identification of novel CYP2D6 duplication arrangements and implications for phenotype prediction. *Pharmacogenomics* 11(1). Future Medicine: 43–53.
- Georgieva L, Rees E, Moran JL, et al. (2014) De novo CNVs in bipolar affective disorder and schizophrenia. *Human molecular genetics* 23(24): 6677–6683.
- Giambartolomei C, Vukcevic D, Schadt EE, et al. (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics* 10(5): e1004383.
- Giannuzzi G, Chatron N, Mannik K, et al. (2022) Possible association of 16p11.2 copy number variation with altered lymphocyte and neutrophil counts. *NPJ genomic medicine* 7(1): 38.
- Gillentine MA and Schaaf CP (2015) The human clinical phenotypes of altered CHRNA7 copy number. *Biochemical pharmacology* 97(4): 352–362.
- Girirajan S, Johnson RL, Tassone F, et al. (2013) Global increases in both common and rare copy number load associated with autism. *Human molecular genetics* 22(14): 2870–2880.
- Glessner JT, Li J, Desai A, et al. (2020) CNV Association of Diverse Clinical Phenotypes from eMERGE reveals novel disease biology underlying cardiovascular disease. *International journal of cardiology* 298: 107–113.
- gnomAD (2024). Available at: <https://gnomad.broadinstitute.org/> (accessed 2020).
- Golzio C, Willer J, Talkowski ME, et al. (2012) KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* 485(7398): 363–367.
- GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC) – Analysis Working Group, Statistical Methods groups – Analysis Working Group, et al. (2017) Genetic effects on gene expression across human tissues. *Nature* 550(7675): 204–213.
- Handsaker RE, Korn JM, Nemesh J, et al. (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature genetics* 43(3): 269–276.
- Handsaker RE, Van Doren V, Berman JR, et al. (2015) Large multiallelic copy number variations in humans. *Nature genetics* 47(3): 296–303.
- Haraksingh RR, Abyzov A and Urban AE (2017) Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC genomics* 18(1): 321.
- Harel T and Lupski JR (2018) Genomic disorders 20 years on-mechanisms for clinical manifestations. *Clinical genetics* 93(3): 439–449.

- Heller D and Vingron M (2019) SVIM: structural variant identification using mapped long reads. *Bioinformatics* 35(17): 2907–2915.
- Huang N, Lee I, Marcotte EM, et al. (2010) Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics* 6(10): e1001154.
- Huguet G, Schramm C, Douard E, et al. (2018) Measuring and Estimating the Effect Sizes of Copy Number Variants on General Intelligence in Community-Based Samples. *JAMA psychiatry* 75(5): 447–457.
- Hujoel MLA, Sherman MA, Barton AR, et al. (2022) Influences of rare copy-number variation on human complex traits. *Cell* 185(22). biorxiv.org: 4233-4248.e27.
- Illumina GenomeStudio 2.0 Plug-ins (2017). Available at: <https://support.illumina.com/downloads/genomestudio-2-0-plugin-ins.html> (accessed 2024).
- Illumina Technical Notes (2007). Available at: [https://www.illumina.com/Documents/products/technotes/technote\\_cnv\\_decode.pdf](https://www.illumina.com/Documents/products/technotes/technote_cnv_decode.pdf).
- iPSYCH (2022). Available at: <https://ipsych.dk/en/> (accessed 2024).
- iPsychCNV (2017). Available at: <https://biopsyk.dk/ipsychcnv/> (accessed 2024).
- Itsara A, Cooper GM, Baker C, et al. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *American journal of human genetics* 84(2): 148–161.
- Jacquemont S, Reymond A, Zufferey F, et al. (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* 478(7367): 97–102.
- Jiang T, Liu Yongzhuang, Jiang Y, et al. (2020) Long-read-based human genomic structural variation detection with cuteSV. *Genome biology* 21(1): 189.
- Kallioniemi A, Kallioniemi O-P, Sudar D, et al. (1992) Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors. *Science* 258(5083): 818–821.
- Kaplanis J, Samocha KE, Wiel L, et al. (2020) Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* 586(7831): 757–762.
- Karczewski KJ, Francioli LC, Tiao G, et al. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581(7809): 434–443.
- Kargi AY and Merriam GR (2013) Diagnosis and treatment of growth hormone deficiency in adults. *Nature reviews. Endocrinology* 9(6): 335–345.
- Katari S, Aarabi M, Kintigh A, et al. (2018) Chromosomal instability in women with primary ovarian insufficiency. *Human reproduction* 33(3): 531–538.
- Kawaguchi S, Ueki M and Kai T (2020) Drosophila MARF1 ensures proper oocyte maturation by regulating nanos expression. *PloS one* 15(4): e0231114.
- Kendall KM, Rees E, Escott-Price V, et al. (2016) Cognitive Performance Among Carriers of Pathogenic Copy Number Variants: Analysis of 152,000 UK Biobank Subjects. *Biological psychiatry*. forthcoming August 18, 2016. DOI: 10.1016/j.biopsych.2016.08.014.
- Khan FF, Melton PE, McCarthy NS, et al. (2018) Whole genome sequencing of 91 multiplex schizophrenia families reveals increased burden of rare, exonic copy number variation in schizophrenia probands and genetic heterogeneity. *Schizophrenia research* 197: 337–345.
- Kichaev G, Yang W-Y, Lindstrom S, et al. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics* 10(10): e1004722.
- Kirov G, Pocklington AJ, Holmans P, et al. (2012) De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular psychiatry* 17(2): 142–153.

- Kirov G, Rees E, Walters JTR, et al. (2014) The penetrance of copy number variations for schizophrenia and developmental delay. *Biological psychiatry* 75(5): 378–385.
- Klambauer G, Schwarzbauer K, Mayr A, et al. (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic acids research* 40(9): e69.
- Köhler S, Carmody L, Vasilevsky N, et al. (2019) Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic acids research* 47(D1): D1018–D1027.
- Koltai H and Weingarten-Baror C (2008) Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. *Nucleic acids research* 36(7): 2395–2405.
- Korn JM, Kuruvilla FG, McCarroll SA, et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature genetics* 40(10): 1253–1260.
- Kosugi S, Momozawa Y, Liu X, et al. (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome biology* 20(1): 117.
- Krumm N, Sudmant PH, Ko A, et al. (2012) Copy number variation detection and genotyping from exome sequence data. *Genome research* 22(8): 1525–1532.
- Layer RM, Chiang C, Quinlan AR, et al. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome biology* 15(6): R84.
- Leitsalu L, Haller T, Esko T, et al. (2015) Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International journal of epidemiology* 44(4): 1137–1147.
- Li YR, Glessner JT, Coe BP, et al. (2020) Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. *Nature communications* 11(1): 255.
- Liu M, Fang L, Liu S, et al. (2019) Array CGH-based detection of CNV regions and their potential association with reproduction and other economic traits in Holsteins. *BMC genomics* 20(1): 181.
- Loh P-R, Tucker G, Bulik-Sullivan BK, et al. (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics* 47(3): 284–290.
- Lucito R, Healy J, Alexander J, et al. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome research* 13(10): 2291–2305.
- Lupski JR and Stankiewicz P (2005) Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS genetics* 1(6): e49.
- Macé A, Tuke MA, Beckmann JS, et al. (2016) New quality measure for SNP array based CNV detection. *Bioinformatics* 32(21): 3298–3305.
- Macé A, Tuke MA, Deelen P, et al. (2017) CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. *Nature communications* 8(1): 744.
- Malhotra D and Sebat J (2012) CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* 148(6): 1223–1241.
- Malhotra D, McCarthy S, Michaelson JJ, et al. (2011) High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* 72(6): 951–963.
- Manichaikul A, Mychaleckyj JC, Rich SS, et al. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22): 2867–2873.

- Männik K, Parkel S, Palta P, et al. (2011) A parallel SNP array study of genomic aberrations associated with mental retardation in patients and general population in Estonia. *European journal of medical genetics* 54(2): 136–143.
- Männik K, Mägi R, Macé A, et al. (2015) Copy number variations and cognitive phenotypes in unselected populations. *JAMA: the journal of the American Medical Association* 313(20): 2044–2054.
- Marshall CR, Howrigan DP, Merico D, et al. (2017) Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature genetics* 49(1): 27–35.
- Marzouka N-A-D, Nordlund J, Bäcklin CL, et al. (2016) CopyNumber450kCancer: baseline correction for accurate copy number calling from the 450k methylation array. *Bioinformatics*. Available at: <http://dx.doi.org/10.1093/bioinformatics/btv652>.
- McCarthy SE, Makarov V, Kirov G, et al. (2009) Microduplications of 16p11.2 are associated with schizophrenia. *Nature genetics* 41(11). Nature Publishing Group: 1223–1227.
- McDaid AF, Joshi PK, Porcu E, et al. (2017) Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nature communications* 8: 15842.
- Michaelovsky E, Frisch A, Carmel M, et al. (2012) Genotype-phenotype correlation in 22q11.2 deletion syndrome. *BMC medical genetics* 13: 122.
- Miller DT, Adam MP, Aradhya S, et al. (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *American journal of human genetics* 86(5). Elsevier BV: 749–764.
- OMIM (2024). Available at: <https://www.omim.org> (accessed 2024).
- Owen D, Bracher-Smith M, Kendall KM, et al. (2018) Effects of pathogenic CNVs on physical traits in participants of the UK Biobank. *BMC genomics* 19(1): 867.
- Pajuste F-D and Remm M (2023) GeneToCN: an alignment-free method for gene copy number estimation directly from next-generation sequencing reads. *Scientific reports* 13(1): 17765.
- Palta P, Kaplinski L, Nagirnaja L, et al. (2015) Haplotype phasing and inheritance of copy number variants in nuclear families. *PloS one* 10(4): e0122713.
- Panagopoulos I and Heim S (2021) Interstitial Deletions Generating Fusion Genes. *Cancer genomics & proteomics* 18(3): 167–196.
- Pepler WJ, Smith M and van Niekerk WA (1968) An unusual karyotype in a patient with signs suggestive of Down's syndrome. *Journal of medical genetics* 5(1): 68–71.
- Perry JRB, Day F, Elks CE, et al. (2014) Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* 514(7520): 92–97.
- Pinto D, Darvishi K, Shi X, et al. (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature biotechnology* 29(6): 512–520.
- Qureshi AY, Mueller S, Snyder AZ, et al. (2014) Opposing brain differences in 16p11.2 deletion and duplication carriers. *The Journal of neuroscience: the official journal of the Society for Neuroscience* 34(34): 11199–11211.
- Rausch T, Zichner T, Schlattl A, et al. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28(18): i333–i339.

- Ravnan JB, Tepperberg JH, Papenhausen P, et al. (2006) Subtelomere FISH analysis of 11 688 cases: an evaluation of the frequency and pattern of subtelomere rearrangements in individuals with developmental disabilities. *Journal of medical genetics* 43(6): 478–489.
- Rees E and Kirov G (2021) Copy number variation and neuropsychiatric illness. *Current opinion in genetics & development* 68: 57–63.
- Rees E, Walters JTR, Georgieva L, et al. (2014) Analysis of copy number variations at 15 schizophrenia-associated loci. *The British journal of psychiatry: the journal of mental science* 204(2): 108–114.
- Rees E, Walters JTR, Chambert KD, et al. (2014) CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1. *Human molecular genetics* 23(6): 1669–1676.
- Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539): 317–330.
- Ruderfer DM, Hamamsy T, Lek M, et al. (2016) Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nature genetics* 48(10): 1107–1111.
- Sanders SJ, He X, Willsey AJ, et al. (2015) Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* 87(6): 1215–1233.
- Schork NJ, Murray SS, Frazer KA, et al. (2009) Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development* 19(3): 212–219.
- Sebat J, Lakshmi B, Malhotra D, et al. (2007) Strong association of de novo copy number mutations with autism. *Science* 316(5823): 445–449.
- Shaffer LG, Bejjani BA, Torchia B, et al. (2007) The identification of microdeletion syndromes and other chromosome abnormalities: cytogenetic methods of the past, new technologies for the future. *American journal of medical genetics. Part C, Seminars in medical genetics* 145C(4): 335–345.
- Shen F and Kidd JM (2020) Rapid, Paralog-Sensitive CNV Analysis of 2457 Human Genomes Using QuickK-mer2. *Genes* 11(2).
- Shinawi M, Liu P, Kang S-HL, et al. (2010) Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *Journal of medical genetics* 47(5): 332–341.
- Shriki-Tal L, Avrahamy H, Pollak Y, et al. (2017) Psychiatric disorders in a cohort of individuals with Prader–Willi syndrome. *European psychiatry: the journal of the Association of European Psychiatrists* 44. Cambridge University Press: 47–52.
- Skre H (1974) Genetic and clinical aspects of Charcot-Marie-Tooth's disease. *Clinical genetics* 6(2). Wiley: 98–118.
- Smolka M, Paulin LF, Grochowski CM, et al. (2024) Detection of mosaic and population-level structural variants with Sniffles2. *Nature biotechnology*. forthcoming January 2, 2024. DOI: 10.1038/s41587-023-02024-y.
- Su Y-Q, Sugiura K, Sun F, et al. (2012) MARF1 regulates essential oogenic processes in mice. *Science* 335(6075): 1496–1499.
- Su Y-Q, Sun F, Handel MA, et al. (2012) Meiosis arrest female 1 (MARF1) has nuage-like function in mammalian oocytes. *Proceedings of the National Academy of Sciences of the United States of America* 109(46): 18653–18660.
- Sudmant PH, Rausch T, Gardner EJ, et al. (2015) An integrated map of structural variation in 2,504 human genomes. *Nature* 526(7571): 75–81.
- Szatkiewicz JP, O'Dushlaine C, Chen G, et al. (2014) Copy number variation in schizophrenia in Sweden. *Molecular psychiatry* 19(7): 762–773.

- Talevich E and Hunter Shain A (2018) CNVkit-RNA: Copy number inference from RNA-Sequencing data. *bioRxiv*. Available at: <https://www.biorxiv.org/content/10.1101/408534v1.abstract> (accessed 8 September 2021).
- Taylor CM, Finucane BM, Moreno-De-Luca A, et al. (2023) Phenotypic shift in copy number variants: Evidence in 16p11.2 duplication syndrome. *Genetics in medicine: official journal of the American College of Medical Genetics* 25(1): 151–154.
- Tervisekassa (2024). Available at: <https://www.tervisekassa.ee/en> (accessed July 2016).
- Theisen A (2008) Microarray-based comparative genomic hybridization (aCGH). *Nature Education*. forthcoming 2008.
- ThermoFisher Scientific Axiom CNV Summary Tools Software (2015). Available at: <https://www.thermofisher.com/ee/en/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/axiom-cnv-summary-tools-software.html> (accessed 2024).
- Thongprayoon C, Cheungpasitporn W and Kashani K (2016) Serum creatinine level, a surrogate of muscle mass, predicts mortality in critically ill patients. *Journal of thoracic disease* 8(5): E305–11.
- Tigchelaar EF, Zhernakova A, Dekens JAM, et al. (2015) Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ open* 5(8): e006772.
- Tzeng J-Y, Magnusson PKE, Sullivan PF, et al. (2015) A New Method for Detecting Associations with Rare Copy-Number Variants. *PLoS genetics* 11(10): e1005403.
- van Bon BWM, Mefford HC, Menten B, et al. (2009) Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. *Journal of medical genetics* 46(8): 511–523.
- Verbitsky M, Krishnamurthy S, Krithivasan P, et al. (2023) Genomic Disorders in CKD across the Lifespan. *Journal of the American Society of Nephrology: JASN* 34(4): 607–618.
- Vialle RA, de Paiva Lopes K, Bennett DA, et al. (2022) Integrating whole-genome sequencing with multi-omic data reveals the impact of structural variants on gene regulation in the human brain. *Nature neuroscience* 25(4): 504–514.
- Vorstman JAS, Staal WG, van Daalen E, et al. (2006) Identification of novel autism candidate regions through analysis of reported cytogenetic abnormalities associated with autism. *Molecular psychiatry* 11(1): 1, 18–28.
- Võsa U, Claringbould A, Westra H-J, et al. (2021) Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature genetics* 53(9): 1300–1310.
- Walsh T, McClellan JM, McCarthy SE, et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320(5875): 539–543.
- Walters RG, Jacquemont S, Valsesia A, et al. (2010) A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* 463(7281): 671–675.
- Wang K, Li M, Hadley D, et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research* 17(11): 1665–1674.
- Weiss LA, Shen Y, Korn JM, et al. (2008) Association between microdeletion and microduplication at 16p11.2 and autism. *The New England journal of medicine* 358(7): 667–675.

- Wenger AM, Peluso P, Rowell WJ, et al. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology* 37(10): 1155–1162.
- Werling DM, Brand H, An J-Y, et al. (2018) An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nature genetics* 50(5): 727–736.
- Wheeler MM, Stilp AM, Rao S, et al. (2022) Whole genome sequencing identifies structural variants contributing to hematologic traits in the NHLBI TOPMed program. *Nature communications* 13(1): 7592.
- Winchester L, Yau C and Ragoussis J (2009) Comparing CNV detection methods for SNP arrays. *Briefings in functional genomics & proteomics* 8(5): 353–366.
- Yoshihara K, Wang Q, Torres-Garcia W, et al. (2015) The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* 34(37): 4845–4854.
- Zarrei M, MacDonald JR, Merico D, et al. (2015) A copy number variation map of the human genome. *Nature reviews. Genetics* 16(3): 172–183.
- Zhang F, Gu W, Hurler ME, et al. (2009) Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics* 10: 451–481.
- Zhu Z, Zhang F, Hu H, et al. (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics* 48(5): 481–487.
- Zufferey F, Sherr EH, Beckmann ND, et al. (2012) A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders. *Journal of medical genetics* 49(10). BMJ Publishing Group Ltd: 660–668.

## ACKNOWLEDGEMENTS

“They stumble that run fast,” said William Shakespeare. Yet even I did not expect to take almost eight years to finish my PhD studies. Nor did my supervisors, Reedik Mägi, Andres Metspalu, Andres Salumets and Andrew Morris, which gives me all the more reason to thank them for their lasting support, guidance, and feedback. I especially wish to thank Reedik, who first led me to CNV studies (probably regretting it later) and have since put up with stubborn me. Thank you for being easily approachable, consistently responsive to all my random questions, and for your caring and good-humoured companionship on this journey. I thank Andres for being a true inspirational leader. You have always made me feel that, by working at Estonian Genome Centre, I am part of something grand. I thank my reviewer Age for critical reading of my work under tight time constraints. I also wish to thank my supervisors back from my Bachelor’s and Master’s studies – Maido, Märt, and Lauris, who first paved the way for my interest in science and genomics.

I am especially grateful to Triin. Despite not formally being my supervisor, you have been a friend and mentor to guide my way in academia as a young female scientist. Thank you for your willingness to hear me out whenever I feel like ranting about my work or other aspects of life. You have ever been my main source of work-related self-confidence. Also, thank you for reading and commenting on several parts of this thesis.

Over the years, many people have contributed to my work both from Estonian Genome Centre and abroad. A big thank you to all my colleagues and co-authors! Thank you, Urmo, Margit, Mart, Maris, Priit, and others for scientific discussions and valuable input to my research. Moreover, thank you, Lili, for introducing me to pharmacogenetics and helping me ease back into science after my (short) maternity leave. I admire your passion for research, which has truly infected me. Also, I am lucky to have been able to gain additional experience from multiple research groups across Europe during my PhD studies. For that I thank Prof. Zoltán Kutalik and Prof. Alexandre Reymond from Lausanne, Prof. Lude Franke from Groningen, Prof. Cecilia Lindgren from Oxford, and all their respective group members. I especially want to thank Chiara for fruitful collaboration and in general for being a nice person to work with. I thank Kaido for making several of those trips with me, for guiding me in statistics, for being a great hiking partner during off hours, and for inspiring me towards better work ethics and self-improvement.

My greatest gratitude goes out to all my fellow PhD students. Dear Fanny, without our everlasting conversations and coffee breaks, both of us would have defended our PhDs long ago, and yet... what is a PhD compared to a true friend! Thank you for all your support as we both try to balance our work and life as best we can. Dear Kreete, I truly admire your resilience and strength. I feel like just talking to you makes all my problems seem less overwhelming. Thank you for sharing in my joys and struggles of PhD and life. Now that the weather is nice

again, we should go hiking more! Dear Kristi K., you are one of the sweetest and most compassionate persons I have ever met. Thank you for always showing your support with kind words and actions. Dear Natalia, Jessica, Anette, Nele, Marili, Kertu, Oliver, Silva, Kristi L., Merli, Tuuli, and everyone else, being on this journey together has made us aware of each other's struggles in ways no-one else could fully understand. You all have a special place in my heart. I could not have done it without you!

Nor could I have done it without my dear family. My dear Mother and Father, dear Andreas, thank you for all your help and support throughout the years to get me where I am today. Dear Kābi, thank you for forcing me outside on days when I did not want to get out of bed. My dear little Hansel, I think it is safe to say that you have taught me more than anyone else in my adult life, even about my work. Who would have thought that PhD could be done in ten-minute chunks trapped under a napping baby! Also, since you joined us, work feels like vacation from home and home feels like vacation from work, so now my life is an everlasting vacation, right? And finally, my dearest Peeter, no words can say how lucky I am to have you. Thank you for being the most loving and supportive husband, and for providing the emotionally secure environment for me to pursue my dreams (and PhD) in! Ma armastan sind!

## **PUBLICATIONS**

## CURRICULUM VITAE

**Name:** Maarja Jõeloo (née Lepamets)  
**Date of birth:** 23.12.1990  
**Contact:** Estonian Genome Centre, Institute of Genomics, University of Tartu, Riia 23b, 51010, Tartu, Estonia  
**E-mail:** maarja.joeloo@ut.ee

### Education:

2021–... PhD in Gene Technology, Institute of Genomics, Faculty of Science and Technology, University of Tartu, Estonia (transferred due to a change in structural unit)  
2016–2021 PhD in Gene Technology, Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu, Estonia  
2013–2015 MSc in Computer Science, Institute of Computer Science, Faculty of Mathematics and Computer Science, University of Tartu, Estonia  
2010–2014 BSc in Gene Technology, Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu, Estonia  
2010–2013 BSc in Mathematical Statistics, Institute of Mathematical Statistics, Faculty of Mathematics and Computer Science, University of Tartu, Estonia

### Professional employment:

2016–... Estonian Genome Centre, Institute of Genomics, University of Tartu, specialist  
2016 Competence Centre on Health Technologies, bioinformatician  
2012–2015 Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu, programmer

### Publications:

Kaplinski, L; **Lepamets, M**; Remm, M (2015). GenomeTester4: a toolkit for performing basic set operations – union, intersection and complement on k-mer lists. *GigaScience*, 4 (58), 1–8.  
Pajuste, F; Kaplinski, L; Möls, M; Puurand, T; **Lepamets, M**; Remm, M (2017). FastGT: an alignment-free method for calling common SNVs directly from raw sequencing reads. *Scientific Reports*, 7 (1), 2537–2537.  
Roosaare, M; Vaher, M; Kaplinski, L; Möls, M; Andreson, R; **Lepamets, M**; Kõressaar, T; Naaber, P; Kõljalg, S; Remm, M (2017). StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. *PeerJ*, 5, e3353–e3353.

- Kõressaar, T; **Lepamets, M**; Kaplinski, L; Raime, K; Andreson, R; Remm, M (2018). Primer3\_masker: integrating masking of template sequence with primer design software. *Bioinformatics*, 34 (11), 1937–1938.
- Reisberg, S; Krebs, K; **Lepamets, M**; Kals, M; Mägi, R; Metsalu, K; Lauschke, VM.; Vilo, J; Milani, L (2019). Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. *Genetics in Medicine* 21, 1345–1354.
- Läll, K; **Lepamets, M**; Palover, M; Esko, T; Metspalu, A; Tõnisson, N; Padrik, P; Mägi, R; Fischer, K (2019). Polygenic prediction of breast cancer: comparison of genetic predictors and implications for risk stratification. *BMC Cancer*, 19 (1), 557.
- Lee, AS.; Rusch, J; Lima, AC.; Usmani, A; Huang, N; **Lepamets, M**; Vigh-Conrad, KA.; Worthington, RE.; Magi, R; Wu, X; Aston, KI; Atkinson, JP.; Carrell, DT.; Hess, RA.; O'Bryan, MK.; Conrad, DF (2019). Rare mutations in the complement regulatory gene CSMD1 are associated with male and female infertility. *Nat Commun*, 10, 4626.
- Männik, K\*; Arbogast, T\*; **Lepamets, M\***; Lepik, K; Pellaz, A; Ademi, H; Kupchinsky, ZA; Ellegood, J; Attanasio, C; Messina, A; Rotman, S; Martin-Brevet, S; Dubruc, E; Chrast, J; Lerch, JP; Qiu, LR; Laisk, T; ...; Davis, EE; Mägi, R; Reymond, A (2019). Leveraging biobank-scale rare and common variant analyses to identify ASPHD1 as the main driver of reproductive traits in the 16p11.2 locus. *bioRxiv*.
- Laisk, T; **Lepamets, M**; Koel, M; Abner, E; Metspalu, A; Mägi, R (2021). Genome-wide association study identifies five risk loci for pernicious anemia. *Nat Commun*, 12 (1), 3761.
- Auwerx, C; **Lepamets, M**; Sadler, MC; Patxot, M; Stojanov, M; Baud, D; Mägi, R; Estonian Biobank Research Team; Porcu, E; Reymond, A; Kutalik, Z (2022). The individual and global impact of copy-number variants on complex human traits. *Am J Hum Genet*, 109 (4), 647–668.
- Collins, RL; Glessner, JT; Porcu, E; **Lepamets, M**; Brandon, R; Lauricella, C; Han, L; Morley, T; Niestroj, LM; Ulirsch, J; Everett, S; Howrigan, DP; Boone, PM; Fu, J; Karczewski, KJ; Kellaris, G; Lowther, C; ...; Brand, H; Talkowski, ME (2022). A cross-disorder dosage sensitivity map of the human genome. *Cell*, 185 (16), 3041–3055.
- Lepamets, M**; Auwerx, C; Nõukas, M; Claringbould, A; Porcu, E; Kals, M; Jürgenson, T; Morris, AP; Estonian Biobank Research Team; Võsa, U; Bochud, M; Stringhini, S; Wijmenga, C; Franke, L; Peterson, H; Vilo, J; Lepik, K; Mägi, R; Kutalik, Z (2022). Omics-informed CNV calls reduce false-positive rates and improve power for CNV-trait associations. *HGG Advances*, 3 (4), 100133.
- Pujol-Gualdo, N; Läll, K; **Lepamets, M**; Estonian Biobank Research Team; Rossi, HR; Arffman, RK; Piltonen, TT; Mägi, R; Laisk, T (2022). Advancing our understanding of genetic risk factors and potential personalized strategies for pelvic organ prolapse. *Nat Commun*, 13 (1), 3584.

- Rahmioglu, N; Mortlock, S; Ghiasi, M; Møller, PL; Stefansdottir, L; Galarneau, G; Turman, C; Danning, R; Law, RH; Sapkota, Y; Christofidou, P; Skarp, S; Giri, A; Banasik, K; Krassowski, M; **Lepamets, M**; ...; Morris, AP; Zondervan, KT (2023). The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions. *Nat Genet*, 55 (3), 423–436.
- Koel, M; Vösa, U; **Jõeloo, M**; Läll, K; Pujol-Gualdo, N; Laivuori, H; Lemmelä, S; Estonian Biobank Research Team; FinnGen; Daly, M; Palta, P; Mägi, R; Laisk, T (2023). GWAS meta-analyses clarify the genetics of cervical phenotypes and inform risk stratification for cervical cancer. *Hum Mol Genet*, 32 (12), 2103–2116.
- Hanssen, R; Auwerx, C; **Jõeloo, M**; Sadler, MC; Estonian Biobank Research Team; Henning, E; Keogh, J; Bounds, R; Smith, M; Firth, HV; Kutalik, Z; Farooqi, IS; Reymond, A; Lawler, K (2023). Chromosomal deletions on 16p11.2 encompassing SH2B1 are associated with accelerated metabolic disease. *Cell Rep Med*, 4 (8), 101155.
- Auwerx, C; **Jõeloo, M**; Sadler, MC; Tesio, N; Ojavee, S; Clark, CJ; Mägi, R; Estonian Biobank Research Team; Reymond, A; Kutalik, Z (2024). Rare copy-number variants as modulators of common disease susceptibility. *Genome Med*, 16 (1), 5.
- Kalnapenkis, A\*; **Jõeloo, M\***; Lepik, K; Kukuškina, V; Kals, M; Alasoo, K; Estonian Biobank Research Team; Mägi, R; Esko, T; Vösa, U (2024). Genetic determinants of plasma protein levels in the Estonian population. *Sci Rep*, 14 (1), 7694.

### Supervised dissertations:

- Tuuli Jürgenson Bachelor's Degree, Institute of Mathematics and Statistics, University of Tartu, Estonia, "The effect of copy number variations on adverse drug reactions" (2019)

### Awards and stipends:

- 2013 Scholarship to Vienna BioCenter Summer School, a 9-week research program for undergrads in Life Sciences
- 2013 Skype and IT Academy scholarship to Master's student of Computer Science
- 2015 Erasmus scholarship for 12-week traineeship at Prof. Richard Mott's lab, Wellcome Trust Centre for Human Genetics, Oxford University, UK
- 2016 Scholarship for PhD student in smart specialisation growth areas from Archimedes Foundation
- 2017 Dora Plus PhD student mobility scholarship to work 5 months as a visiting PhD student at Prof. Alexandre Reymond's lab and Prof. Zoltán Kutalik's lab, University of Lausanne, Switzerland
- 2017 Kristjan Jaak scholarship for short-time visit at Prof. Lude Franke's lab, Department of Genetics, University Medical Center Groningen, the Netherlands

## ELULOOKIRJELDUS

**Nimi:** Maarja Jõeloo (neiupõlvenimi: Lepamets)  
**Sünniaeg:** 23.12.1990  
**Address:** Eesti Geenivaramu, genoomika instituut, Tartu Ülikool, Riia 23b, 51010, Tartu, Eesti  
**E-post:** maarja.joeloo@ut.ee

### Haridus:

2021–... Doktoriõpe geenitehnoloogia erialal, genoomika instituut, loodus- ja täppisteaduste valdkond, Tartu Ülikool, Eesti (viidud üle seoses struktuuriüksuse muutusega)  
2016–2021 Doktoriõpe geenitehnoloogia erialal, molekulaar- ja rakubioloogia instituut, loodus- ja täppisteaduste valdkond, Tartu Ülikool, Eesti  
2013–2015 Magistriõpe arvutiteaduse erialal, arvutiteaduse instituut, matemaatika-informaatikateaduskond, Tartu Ülikool, Eesti  
2010–2014 Bakalaureuseõpe geenitehnoloogia erialal, molekulaar- ja rakubioloogia instituut, loodus- ja tehnoloogiateaduskond, Tartu Ülikool, Eesti  
2010–2013 Bakalaureuseõpe matemaatilise statistika erialal, matemaatilise statistika instituut, matemaatika-informaatikateaduskond, Tartu Ülikool, Eesti

### Teenistuskäik:

2016–... Eesti Geenivaramu, genoomika instituut, Tartu Ülikool, spetsialist  
2016 Tervisetehnoloogiate Arenduskeskus AS, bioinformaatik  
2012–2015 Molekulaar- ja rakubioloogia instituut, Tartu Ülikool, programmeerija

### Teaduspublikatsioonid:

Kaplinski, L; **Lepamets, M**; Remm, M (2015). GenomeTester4: a toolkit for performing basic set operations - union, intersection and complement on k-mer lists. *GigaScience*, 4 (58), 1–8.  
Pajuste, F; Kaplinski, L; Möls, M; Puurand, T; **Lepamets, M**; Remm, M (2017). FastGT: an alignment-free method for calling common SNVs directly from raw sequencing reads. *Scientific Reports*, 7 (1), 2537–2537.  
Roosaare, M; Vaher, M; Kaplinski, L; Möls, M; Andreson, R; **Lepamets, M**; Kõressaar, T; Naaber, P; Kõljalg, S; Remm, M (2017). StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. *PeerJ*, 5, e3353–e3353.

- Kõressaar, T; **Lepamets, M**; Kaplinski, L; Raime, K; Andreson, R; Remm, M (2018). Primer3\_masker: integrating masking of template sequence with primer design software. *Bioinformatics*, 34 (11), 1937–1938.
- Reisberg, S; Krebs, K; **Lepamets, M**; Kals, M; Mägi, R; Metsalu, K; Lauschke, VM.; Vilo, J; Milani, L (2019). Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. *Genetics in Medicine* 21, 1345–1354.
- Läll, K; **Lepamets, M**; Palover, M; Esko, T; Metspalu, A; Tõnisson, N; Padrik, P; Mägi, R; Fischer, K (2019). Polygenic prediction of breast cancer: comparison of genetic predictors and implications for risk stratification. *BMC Cancer*, 19 (1), 557.
- Lee, AS.; Rusch, J; Lima, AC.; Usmani, A; Huang, N; **Lepamets, M**; Vigh-Conrad, KA.; Worthington, RE.; Magi, R; Wu, X; Aston, KI; Atkinson, JP.; Carrell, DT.; Hess, RA.; O'Bryan, MK.; Conrad, DF (2019). Rare mutations in the complement regulatory gene CSMD1 are associated with male and female infertility. *Nat Commun*, 10, 4626.
- Männik, K\*; Arbogast, T\*; **Lepamets, M\***; Lepik, K; Pellaz, A; Ademi, H; Kupchinsky, ZA; Ellegood, J; Attanasio, C; Messina, A; Rotman, S; Martin-Brevet, S; Dubruc, E; Chrast, J; Lerch, JP; Qiu, LR; Laisk, T; ...; Davis, EE; Mägi, R; Reymond, A (2019). Leveraging biobank-scale rare and common variant analyses to identify ASPHD1 as the main driver of reproductive traits in the 16p11.2 locus. *bioRxiv*.
- Laisk, T; **Lepamets, M**; Koel, M; Abner, E; Metspalu, A; Mägi, R (2021). Genome-wide association study identifies five risk loci for pernicious anemia. *Nat Commun*, 12 (1), 3761.
- Auwerx, C; **Lepamets, M**; Sadler, MC; Patxot, M; Stojanov, M; Baud, D; Mägi, R; Estonian Biobank Research Team; Porcu, E; Reymond, A; Kutalik, Z (2022). The individual and global impact of copy-number variants on complex human traits. *Am J Hum Genet*, 109 (4), 647–668.
- Collins, RL; Glessner, JT; Porcu, E; **Lepamets, M**; Brandon, R; Lauricella, C; Han, L; Morley, T; Niestroj, LM; Ulirsch, J; Everett, S; Howrigan, DP; Boone, PM; Fu, J; Karczewski, KJ; Kellaris, G; Lowther, C; ...; Brand, H; Talkowski, ME (2022). A cross-disorder dosage sensitivity map of the human genome. *Cell*, 185 (16), 3041–3055.
- Lepamets, M**; Auwerx, C; Nõukas, M; Claringbould, A; Porcu, E; Kals, M; Jürgenson, T; Morris, AP; Estonian Biobank Research Team; Võsa, U; Bochud, M; Stringhini, S; Wijmenga, C; Franke, L; Peterson, H; Vilo, J; Lepik, K; Mägi, R; Kutalik, Z (2022). Omics-informed CNV calls reduce false-positive rates and improve power for CNV-trait associations. *HGG Advances*, 3 (4), 100133.
- Pujol-Gualdo, N; Läll, K; **Lepamets, M**; Estonian Biobank Research Team; Rossi, HR; Arffman, RK; Piltonen, TT; Mägi, R; Laisk, T (2022). Advancing our understanding of genetic risk factors and potential personalized strategies for pelvic organ prolapse. *Nat Commun*, 13 (1), 3584.

- Rahmioglu, N; Mortlock, S; Ghiasi, M; Møller, PL; Stefansdottir, L; Galarneau, G; Turman, C; Danning, R; Law, RH; Sapkota, Y; Christofidou, P; Skarp, S; Giri, A; Banasik, K; Krassowski, M; **Lepamets, M**; ...; Morris, AP; Zondervan, KT (2023). The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions. *Nat Genet*, 55 (3), 423–436.
- Koel, M; Vösa, U; **Jõeloo, M**; Läll, K; Pujol-Gualdo, N; Laivuori, H; Lemmelä, S; Estonian Biobank Research Team; FinnGen; Daly, M; Palta, P; Mägi, R; Laisk, T (2023). GWAS meta-analyses clarify the genetics of cervical phenotypes and inform risk stratification for cervical cancer. *Hum Mol Genet*, 32 (12), 2103–2116.
- Hanssen, R; Auwerx, C; **Jõeloo, M**; Sadler, MC; Estonian Biobank Research Team; Henning, E; Keogh, J; Bounds, R; Smith, M; Firth, HV; Kutalik, Z; Farooqi, IS; Reymond, A; Lawler, K (2023). Chromosomal deletions on 16p11.2 encompassing SH2B1 are associated with accelerated metabolic disease. *Cell Rep Med*, 4 (8), 101155.
- Auwerx, C; **Jõeloo, M**; Sadler, MC; Tesio, N; Ojavee, S; Clark, CJ; Mägi, R; Estonian Biobank Research Team; Reymond, A; Kutalik, Z (2024). Rare copy-number variants as modulators of common disease susceptibility. *Genome Med*, 16 (1), 5.
- Kalnapenkis, A\*; **Jõeloo, M\***; Lepik, K; Kukuškina, V; Kals, M; Alasoo, K; Estonian Biobank Research Team; Mägi, R; Esko, T; Vösa, U (2024). Genetic determinants of plasma protein levels in the Estonian population. *Sci Rep*, 14 (1), 7694.

#### Juhendatud väitekirjad:

- Tuuli Jürgenson Bakalaureusetöö, matemaatika ja statistika instituut, Tartu Ülikool, Eesti, „Koopiaarvu variatsioonide mõju ravimi kõrvaltoimete tekkimisele” (2019)

#### Stipendiumid:

- 2013 Stipendium Viini Biokeskuse suvekooli (9-nädalane teadusprogramm bakalaureuse- ja magistriõppuritele eluteaduste valdkonnas)
- 2013 Skype ja IT Akadeemia magistristipendium arvutiteadustes
- 2015 Erasmuse stipendium praktika läbiviimiseks välisriigis (prof. Richard Motti grupp, Wellcome Trusti Inimgeneetika keskus, Oxfordi Ülikool, Suurbritannia)
- 2016 Sihtasutus Archimedes nutika spetsialiseerumise kasvuvaldkondade doktorandistipendium
- 2017 Dora Plus pikaajalise õpirände doktorandistipendium (prof. Alexandre Reymond' ja prof. Zoltán Kutaliku grupp, Lausanne-i Ülikool, Šveits)
- 2017 Kristjan Jaagu lühiajalise õpirände stipendium (prof. Lude Franke grupp, Groningeni Ülikool, Holland)

## DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

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

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

42. **Veljo Kisand.** Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa.** Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa.** Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik.** Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo.** Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo.** Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots.** Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero.** Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees.** Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks.** Cholecystokinin (CCK) – induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and serotonin. Tartu, 1999, 123 p.
52. **Ebe Sild.** Impact of increasing concentrations of O<sub>3</sub> and CO<sub>2</sub> on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva.** Electron microscopical analysis of the synaptone-mal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna.** Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro.** Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane.** Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm.** Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg.** Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild.** The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu.** Studies of the TOL plasmid transcription factor XylS. Tartu, 2000, 88 p.
61. **Dina Lepik.** Modulation of viral DNA replication by tumor suppressor protein p53. Tartu, 2000, 106 p.
62. **Kai Vellak.** Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu, 2000, 122 p.

63. **Jonne Kotta.** Impact of eutrophication and biological invasions on the structure and functions of benthic macrofauna. Tartu, 2000, 160 p.
64. **Georg Martin.** Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000, 139 p.
65. **Silvia Sepp.** Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira.** On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000, 96 p.
67. **Priit Zingel.** The role of planktonic ciliates in lake ecosystems. Tartu, 2001, 111 p.
68. **Tiit Teder.** Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu, 2001, 122 p.
69. **Hannes Kollist.** Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu, 2001, 80 p.
70. **Reet Marits.** Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu, 2001, 112 p.
71. **Vallo Tilgar.** Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major*, breeding in Northern temperate forests. Tartu, 2002, 126 p.
72. **Rita Hõrak.** Regulation of transposition of transposon Tn4652 in *Pseudomonas putida*. Tartu, 2002, 108 p.
73. **Liina Eek-Piirsoo.** The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002, 74 p.
74. **Krõõt Aasamaa.** Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002, 110 p.
75. **Nele Ingerpuu.** Bryophyte diversity and vascular plants. Tartu, 2002, 112 p.
76. **Neeme Tõnisson.** Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002, 124 p.
77. **Margus Pensa.** Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003, 110 p.
78. **Asko Lõhmus.** Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003, 168 p.
79. **Viljar Jaks.** p53 – a switch in cellular circuit. Tartu, 2003, 160 p.
80. **Jaana Männik.** Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003, 140 p.
81. **Marek Sammul.** Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003, 159 p.
82. **Ivar Ilves.** Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003, 89 p.
83. **Andres Männik.** Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003, 109 p.

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

106. **Ave Suija.** Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005, 162 p.
107. **Piret Lõhmus.** Forest lichens and their substrata in Estonia. Tartu, 2005, 162 p.
108. **Inga Lips.** Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005, 156 p.
109. **Krista Kaasik.** Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005, 121 p.
110. **Juhan Javoiš.** The effects of experience on host acceptance in ovipositing moths. Tartu, 2005, 112 p.
111. **Tiina Sedman.** Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005, 103 p.
112. **Ruth Aguraiuja.** Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005, 112 p.
113. **Riho Teras.** Regulation of transcription from the fusion promoters generated by transposition of Tn4652 into the upstream region of *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 106 p.
114. **Mait Metspalu.** Through the course of prehistory in India: tracing the mtDNA trail. Tartu, 2005, 138 p.
115. **Elin Lõhmussaar.** The comparative patterns of linkage disequilibrium in European populations and its implication for genetic association studies. Tartu, 2006, 124 p.
116. **Priit Kupper.** Hydraulic and environmental limitations to leaf water relations in trees with respect to canopy position. Tartu, 2006, 126 p.
117. **Heili Ilves.** Stress-induced transposition of Tn4652 in *Pseudomonas Putida*. Tartu, 2006, 120 p.
118. **Silja Kuusk.** Biochemical properties of Hmi1p, a DNA helicase from *Saccharomyces cerevisiae* mitochondria. Tartu, 2006, 126 p.
119. **Kersti Püssa.** Forest edges on medium resolution landsat thematic mapper satellite images. Tartu, 2006, 90 p.
120. **Lea Tummeleht.** Physiological condition and immune function in great tits (*Parus major* L.): Sources of variation and trade-offs in relation to growth. Tartu, 2006, 94 p.
121. **Toomas Esperk.** Larval instar as a key element of insect growth schedules. Tartu, 2006, 186 p.
122. **Harri Valdmann.** Lynx (*Lynx lynx*) and wolf (*Canis lupus*) in the Baltic region: Diets, helminth parasites and genetic variation. Tartu, 2006. 102 p.
123. **Priit Jõers.** Studies of the mitochondrial helicase Hmi1p in *Candida albicans* and *Saccharomyces cerevisiae*. Tartu, 2006. 113 p.
124. **Kersti Lilleväli.** Gata3 and Gata2 in inner ear development. Tartu, 2007, 123 p.
125. **Kai Rünk.** Comparative ecology of three fern species: *Dryopteris carthusiana* (Vill.) H.P. Fuchs, *D. expansa* (C. Presl) Fraser-Jenkins & Jermy and *D. dilatata* (Hoffm.) A. Gray (Dryopteridaceae). Tartu, 2007, 143 p.

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

147. **Ulvi Karu.** Antioxidant protection, carotenoids and coccidians in green-finches – assessment of the costs of immune activation and mechanisms of parasite resistance in a passerine with carotenoid-based ornaments. Tartu, 2008, 124 p.
148. **Jaanus Remm.** Tree-cavities in forests: density, characteristics and occupancy by animals. Tartu, 2008, 128 p.
149. **Epp Moks.** Tapeworm parasites *Echinococcus multilocularis* and *E. granulosus* in Estonia: phylogenetic relationships and occurrence in wild carnivores and ungulates. Tartu, 2008, 82 p.
150. **Eve Eensalu.** Acclimation of stomatal structure and function in tree canopy: effect of light and CO<sub>2</sub> concentration. Tartu, 2008, 108 p.
151. **Janne Pullat.** Design, functionlization and application of an *in situ* synthesized oligonucleotide microarray. Tartu, 2008, 108 p.
152. **Marta Putrinš.** Responses of *Pseudomonas putida* to phenol-induced metabolic and stress signals. Tartu, 2008, 142 p.
153. **Marina Semtsenko.** Plant root behaviour: responses to neighbours and physical obstructions. Tartu, 2008, 106 p.
154. **Marge Starast.** Influence of cultivation techniques on productivity and fruit quality of some *Vaccinium* and *Rubus* taxa. Tartu, 2008, 154 p.
155. **Age Tats.** Sequence motifs influencing the efficiency of translation. Tartu, 2009, 104 p.
156. **Radi Tegova.** The role of specialized DNA polymerases in mutagenesis in *Pseudomonas putida*. Tartu, 2009, 124 p.
157. **Tsipe Aavik.** Plant species richness, composition and functional trait pattern in agricultural landscapes – the role of land use intensity and landscape structure. Tartu, 2009, 112 p.
158. **Kaja Kiiver.** Semliki forest virus based vectors and cell lines for studying the replication and interactions of alphaviruses and hepaciviruses. Tartu, 2009, 104 p.
159. **Meelis Kadaja.** Papillomavirus Replication Machinery Induces Genomic Instability in its Host Cell. Tartu, 2009, 126 p.
160. **Pille Hallast.** Human and chimpanzee Luteinizing hormone/Chorionic Gonadotropin beta (*LHB/CGB*) gene clusters: diversity and divergence of young duplicated genes. Tartu, 2009, 168 p.
161. **Ain Vellak.** Spatial and temporal aspects of plant species conservation. Tartu, 2009, 86 p.
162. **Triinu Remmel.** Body size evolution in insects with different colouration strategies: the role of predation risk. Tartu, 2009, 168 p.
163. **Jaana Salujõe.** Zooplankton as the indicator of ecological quality and fish predation in lake ecosystems. Tartu, 2009, 129 p.
164. **Ele Vahtmäe.** Mapping benthic habitat with remote sensing in optically complex coastal environments. Tartu, 2009, 109 p.
165. **Liisa Metsamaa.** Model-based assessment to improve the use of remote sensing in recognition and quantitative mapping of cyanobacteria. Tartu, 2009, 114 p.

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

186. **Katrin Kepp.** Genes involved in cardiovascular traits: detection of genetic variation in Estonian and Czech populations. Tartu, 2010, 164 p.
187. **Virve Sõber.** The role of biotic interactions in plant reproductive performance. Tartu, 2010, 92 p.
188. **Kersti Kangro.** The response of phytoplankton community to the changes in nutrient loading. Tartu, 2010, 144 p.
189. **Joachim M. Gerhold.** Replication and Recombination of mitochondrial DNA in Yeast. Tartu, 2010, 120 p.
190. **Helen Tammert.** Ecological role of physiological and phylogenetic diversity in aquatic bacterial communities. Tartu, 2010, 140 p.
191. **Elle Rajandu.** Factors determining plant and lichen species diversity and composition in Estonian *Calamagrostis* and *Hepatica* site type forests. Tartu, 2010, 123 p.
192. **Paula Ann Kivistik.** ColR-ColS signalling system and transposition of Tn4652 in the adaptation of *Pseudomonas putida*. Tartu, 2010, 118 p.
193. **Siim Sõber.** Blood pressure genetics: from candidate genes to genome-wide association studies. Tartu, 2011, 120 p.
194. **Kalle Kipper.** Studies on the role of helix 69 of 23S rRNA in the factor-dependent stages of translation initiation, elongation, and termination. Tartu, 2011, 178 p.
195. **Triinu Siibak.** Effect of antibiotics on ribosome assembly is indirect. Tartu, 2011, 134 p.
196. **Tambet Tõnissoo.** Identification and molecular analysis of the role of guanine nucleotide exchange factor RIC-8 in mouse development and neural function. Tartu, 2011, 110 p.
197. **Helin Räägel.** Multiple faces of cell-penetrating peptides – their intracellular trafficking, stability and endosomal escape during protein transduction. Tartu, 2011, 161 p.
198. **Andres Jaanus.** Phytoplankton in Estonian coastal waters – variability, trends and response to environmental pressures. Tartu, 2011, 157 p.
199. **Tiit Nikopensius.** Genetic predisposition to nonsyndromic orofacial clefts. Tartu, 2011, 152 p.
200. **Signe Värvi.** Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.
201. **Kristjan Vääk.** Gene expression profiling and genome-wide association studies of non-small cell lung cancer. Tartu, 2011, 98 p.
202. **Arno Põllumäe.** Spatio-temporal patterns of native and invasive zooplankton species under changing climate and eutrophication conditions. Tartu, 2011, 153 p.
203. **Egle Tammelaht.** Brown bear (*Ursus arctos*) population structure, demographic processes and variations in diet in northern Eurasia. Tartu, 2011, 143 p.
205. **Teele Jairus.** Species composition and host preference among ectomycorrhizal fungi in Australian and African ecosystems. Tartu, 2011, 106 p.

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

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

246. **Mario Plaas.** Animal model of Wolfram Syndrome in mice: behavioural, biochemical and psychopharmacological characterization. Tartu, 2013, 144 p.
247. **Georgi Hudjašov.** Maps of mitochondrial DNA, Y-chromosome and tyrosinase variation in Eurasian and Oceanian populations. Tartu, 2013, 115 p.
248. **Mari Lepik.** Plasticity to light in herbaceous plants and its importance for community structure and diversity. Tartu, 2013, 102 p.
249. **Ede Leppik.** Diversity of lichens in semi-natural habitats of Estonia. Tartu, 2013, 151 p.
250. **Ülle Saks.** Arbuscular mycorrhizal fungal diversity patterns in boreo-nemoral forest ecosystems. Tartu, 2013, 151 p.
251. **Eneli Oitmaa.** Development of arrayed primer extension microarray assays for molecular diagnostic applications. Tartu, 2013, 147 p.
252. **Jekaterina Jutkina.** The horizontal gene pool for aromatics degradation: bacterial catabolic plasmids of the Baltic Sea aquatic system. Tartu, 2013, 121 p.
253. **Helen Vellau.** Reaction norms for size and age at maturity in insects: rules and exceptions. Tartu, 2014, 132 p.
254. **Randel Kreitsberg.** Using biomarkers in assessment of environmental contamination in fish – new perspectives. Tartu, 2014, 107 p.
255. **Krista Takkis.** Changes in plant species richness and population performance in response to habitat loss and fragmentation. Tartu, 2014, 141 p.
256. **Liina Nagirnaia.** Global and fine-scale genetic determinants of recurrent pregnancy loss. Tartu, 2014, 211 p.
257. **Triin Triisberg.** Factors influencing the re-vegetation of abandoned extracted peatlands in Estonia. Tartu, 2014, 133 p.
258. **Villu Soon.** A phylogenetic revision of the *Chrysis ignita* species group (Hymenoptera: Chrysididae) with emphasis on the northern European fauna. Tartu, 2014, 211 p.
259. **Andrei Nikonov.** RNA-Dependent RNA Polymerase Activity as a Basis for the Detection of Positive-Strand RNA Viruses by Vertebrate Host Cells. Tartu, 2014, 207 p.
260. **Eele Õunapuu-Pikas.** Spatio-temporal variability of leaf hydraulic conductance in woody plants: ecophysiological consequences. Tartu, 2014, 135 p.
261. **Marju Männiste.** Physiological ecology of greenfinches: information content of feathers in relation to immune function and behavior. Tartu, 2014, 121 p.
262. **Katre Kets.** Effects of elevated concentrations of CO<sub>2</sub> and O<sub>3</sub> on leaf photosynthetic parameters in *Populus tremuloides*: diurnal, seasonal and inter-annual patterns. Tartu, 2014, 115 p.
263. **Küllli Lokko.** Seasonal and spatial variability of zoopsammon communities in relation to environmental parameters. Tartu, 2014, 129 p.
264. **Olga Žilina.** Chromosomal microarray analysis as diagnostic tool: Estonian experience. Tartu, 2014, 152 p.

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

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

303. **Hedvig Tamman.** The GraTA toxin-antitoxin system of *Pseudomonas putida*: regulation and role in stress tolerance. Tartu, 2016, 154 p.
304. **Kadri Pärtel.** Application of ultrastructural and molecular data in the taxonomy of helotialean fungi. Tartu, 2016, 183 p.
305. **Maris Hindrikson.** Grey wolf (*Canis lupus*) populations in Estonia and Europe: genetic diversity, population structure and -processes, and hybridization between wolves and dogs. Tartu, 2016, 121 p.
306. **Polina Degtjarenko.** Impacts of alkaline dust pollution on biodiversity of plants and lichens: from communities to genetic diversity. Tartu, 2016, 126 p.
307. **Liina Pajusalu.** The effect of CO<sub>2</sub> enrichment on net photosynthesis of macrophytes in a brackish water environment. Tartu, 2016, 126 p.
308. **Stoyan Tankov.** Random walks in the stringent response. Tartu, 2016, 94 p.
309. **Liis Leitsalu.** Communicating genomic research results to population-based biobank participants. Tartu, 2016, 158 p.
310. **Richard Meitern.** Redox physiology of wild birds: validation and application of techniques for detecting oxidative stress. Tartu, 2016, 134 p.
311. **Kaie Lokk.** Comparative genome-wide DNA methylation studies of healthy human tissues and non-small cell lung cancer tissue. Tartu, 2016, 127 p.
312. **Mihhail Kurašin.** Processivity of cellulases and chitinases. Tartu, 2017, 132 p.
313. **Carmen Tali.** Scavenger receptors as a target for nucleic acid delivery with peptide vectors. Tartu, 2017, 155 p.
314. **Katarina Oganjan.** Distribution, feeding and habitat of benthic suspension feeders in a shallow coastal sea. Tartu, 2017, 132 p.
315. **Taavi Paal.** Immigration limitation of forest plants into wooded landscape corridors. Tartu, 2017, 145 p.
316. **Kadri Õunap.** The Williams-Beuren syndrome chromosome region protein WBSR22 is a ribosome biogenesis factor. Tartu, 2017, 135 p.
317. **Riin Tamm.** In-depth analysis of factors affecting variability in thiopurine methyltransferase activity. Tartu, 2017, 170 p.
318. **Keiu Kask.** The role of RIC8A in the development and regulation of mouse nervous system. Tartu, 2017, 184 p.
319. **Tiia Möller.** Mapping and modelling of the spatial distribution of benthic macrovegetation in the NE Baltic Sea with a special focus on the eelgrass *Zostera marina* Linnaeus, 1753. Tartu, 2017, 162 p.
320. **Silva Kasela.** Genetic regulation of gene expression: detection of tissue- and cell type-specific effects. Tartu, 2017, 150 p.
321. **Karmen Süld.** Food habits, parasites and space use of the raccoon dog *Nyctereutes procyonoides*: the role of an alien species as a predator and vector of zoonotic diseases in Estonia. Tartu, 2017, p.
322. **Ragne Oja.** Consequences of supplementary feeding of wild boar – concern for ground-nesting birds and endoparasite infection. Tartu, 2017, 141 p.
323. **Riin Kont.** The acquisition of cellulose chain by a processive cellobiohydrolase. Tartu, 2017, 117 p.

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

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

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

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

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

- 422. **Vlad-Julian Piljukov.** Biochemical characterization of Irc3 helicase. Tartu, 2023, 137 p.
- 423. **Gerli Albert.** Carbon use strategies of macrophyte communities in the northeastern Baltic Sea: implications for a high CO<sub>2</sub> environment. Tartu, 2023, 128 p.
- 424. **Mariann Koel.** The molecular interactions between trophoblast and endometrial cells in embryo implantation. Tartu, 2023, 171 p.
- 425. **Robin Gielen.** Diversity and ecological role of pathogenic fungi in insect populations. Tartu, 2023, 139 p.
- 426. **Kaspar Reier.** Quantity, stability and disparity of ribosomal components in *Escherichia coli* stationary phase. Tartu, 2023, 151 p.
- 427. **Linda Rusalepp.** The impact of environmental drivers and competition on phenolic metabolite profiles in hybrid aspen and silver birch. Tartu, 2023, 153 p.
- 428. **Eliisa Pass.** The effect of managed forest-wetland landscapes on forest grouse and nest predation. Tartu, 2023, 115 p.
- 429. **Sanni Färkkilä.** Methods for studying plant-fungal interactions – reflecting on the old, the new and the upcoming. Tartu, 2024, 147 p.