UNIVERSITY OF TARTU

Faculty of Science and Technology

Institute of Technology

Anastasiia Alekseienko

Systematic interpretation of large-scale GWAS analyses of 5,035 phenotypes

Master's Thesis (30 ECTS)

Curriculum Bioengineering

Supervisors:
Research Fellow of Functional Genomics, PhD Urmo Võsa
Research Fellow of Functional Genomics, PhD Erik Abner

Tartu 2024

Systematic interpretation of large-scale GWAS analyses of 5,035 phenotypes

Abstract

This thesis presents the systematic interpretation of 5,035 genome-wide association studies (GWAS) conducted within the Estonian Biobank, aiming to elucidate the genetic determinants influencing a diverse array of phenotypic traits. Through a review of existing literature and the application of advanced bioinformatic tools, the work done in this thesis outlined the results of main post-GWAS methods, such as the identification of novel variants, SNP heritability estimation, fine-mapping of causal variants, and prioritization of genes associated with complex traits. Around 26% of the variants identified as lead ones in this work are novel and had not been implicated by GWASs before. Fine mapping prioritized single genetic variants for 10.3% of investigated loci, providing hypotheses for further functional studies, and the gene prioritization approach identified the 2,402 lead variants to be related to 804 genes, several of those biologically interpretable. Heritability was reliably inferred for 25% of studied phenotypes, the most heritable traits being “Other specified hypothyroidism” (ICD-10 code E03.8), “Obesity due to excess calories” (E66.0), “Obesity” (E66), “Myopia” (H52.1), and “Hypertension” (I10). These findings are a good starting point for a more in-depth interpretation of loci associated with complex diseases in the Estonian Biobank.

Keywords

Genome-wide association studies; Post-GWAS analysis; Large-scale analysis

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics; B220 Genetics, cytogenetics

Suuremahuliste GWAS analüüside süstemaatiline tõlgendamine 5,035 fenotüübi põhjal

Lühikokkuvõte

Selle magistritöö raames tehti Eesti geenivaramu andmete põhjal läbi viidud 5,035 fenotüüpi hõlmavate ülegenoomsete assotsiatsiooniuuringute (GWAS) süstemaatiline tõlgendamine, eesmärgiga selgitada välja geneetilised variandid, mis mõjutavad erinevate komplekshaigustefenotüübilisi omadusi. Rakendades erinevaid bioinformaatika tööriistu ning kirjanduses varem avaldatud lähenemisi, keskendus käesolev töö uute haigusseoseliste geenivariantide tuvastamisele, SNP-põhise pärilikkuse hindamisele, põhjuslike variantide täppiskaardistamisele ja komplekssete tunnustega seotud geenide prioriseerimisele. Umbes 26% selles uuringus leitud geneetilistes signaalidest polnud varasemalt maailmas avaldatud GWAS-ides kirjeldatud. Geneetilised täppiskaardistamised tuvastasid üksikuid geenivariante 10.3% uuritud lookustest, luues seega hüpoteese jätku-uuringutele. Geenide prioriseerimine geneetilistes lookustes tuvastas 2,402 geenivarianti 804 erinevas geenis ning mitmed neist geenidest olid bioloogiliselt tõlgendatavad. Geneetiline pärilikkus sai kõrge usaldusväärsusega hinnatud 25% uuritud fenotüüpidele ning kõige päritavamad tunnused olid "Muu täpsustatud hüpotüreoos"(RHK-10 kood E03.8), "Ülekaalulisus" (E66 ja E66.0), "Lühinägelikkus" (H52.1) ning "Kõrgvererõhutõbi" (I10). Need tulemused on hea alguspunkt edasisteks komplekshaigustega seotud lookuste uurimiseks Eesti geenivaramu andmete põhjal.

Võtmesõnad

Ülegenoomsed assotsiatsiooniuuringud; GWAS-järgne analüüs; Laiaulatuslik analüüs

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetria; B220 Geneetika, tsütogeneetika.

TABLE OF CONTENTS

TERMS, ABBREVIATIONS AND NOTATIONS

ABF	–	Approximate Bayes Factor
CLPP	–	Colocalization Posterior Probability
CS	–	Credible Set
DEPICT	–	Data-driven Expression Prioritized Integration for Complex Traits
eQTL	–	Expression Quantitative Trait Loci
EstBB	–	Estonian Biobank
FUMA	–	Functional Mapping and Annotation of GWAS
GWAS	–	Genome-wide Association Study
ICD	–	International Classification of Disease
LD	–	Linkage Disequilibrium
LDSC/LDSR	–	Linkage Disequilibrium Score Regression
MAF	–	Minor Allele Frequency
MR	–	Mendelian Randomisation
N_{eff}	–	Effective Sample Size
N	–	Sample Size
NMR	–	Nuclear Magnetic Resonance
PCA	–	Principal Component Analysis
PIP	–	Posterior Inclusion Probability
SNP	–	Single-nucleotide Polymorphism
TSS	–	Transcription Start Site
UKBB	–	UK Biobank
WES	–	Whole Exome Sequencing
WGS	–	Whole Genome Sequencing

INTRODUCTION

Genome-wide association studies (GWASs) have revolutionized the field of genetics by enabling researchers to identify associations between genetic variants and complex traits or diseases across the entire genome. By scanning markers across the genomes of many individuals, GWASs can pinpoint specific genetic variations that correlate with particular phenotypic outcomes. This approach has significantly advanced our understanding of the genetic underpinnings of numerous traits and diseases, highlighting the genetic complexity and the polygenic nature of many conditions [1].

The importance of GWAS lies in its ability to uncover genetic variants that contribute to phenotypic diversity and disease susceptibility. However, such studies alone provide merely statistical data, which needs to be transformed for further understanding and interpretation. While GWAS can identify loci associated with traits, it does not inherently reveal the causal genes, the functional impact of variants, or the biological pathways involved. This limitation necessitates post-GWAS analysis techniques for translating GWAS findings into meaningful biological knowledge.

Post-GWAS analyses are crucial for several reasons. Firstly, they help to estimate the heritability of traits by quantifying the proportion of phenotypic variance attributable to genetic variation, often through methods like linkage disequilibrium score regression. Second, fine-mapping techniques refine GWAS signals and identify the most likely causal variants within associated loci. Then, integrating functional genomics data through quantitative trait loci (QTL) mapping and colocalization analyses can link genetic variants to gene expression changes, thereby prioritizing causal genes. Lastly, pathway and enrichment analyses elucidate the biological processes and pathways affected by the associated variants.

In this thesis, we apply several post-GWAS techniques to analyze 5,035 phenotypes in the Estonian Biobank. The Estonian Biobank provides a rich resource of genetic and phenotypic data, enabling a comprehensive investigation into the genetic architecture of complex traits. Our approach includes identifying lead variants, estimating SNP heritability, fine-mapping associated loci, and prioritizing causal genes, all done in a high-throughput manner. This research aims to extend beyond the initial GWAS findings and take the first steps in uncovering the functional implications of genetic variants by employing a combination of computational tools and bioinformatic methodologies.

This analysis aims to enhance our understanding of the genetic determinants of human traits and diseases. The results from this study will contribute to a more complete understanding of the genetic determinants of human traits and diseases and, in the future, could help support the development of precision medicine strategies.

1 LITERATURE REVIEW

1.1 GENOME-WIDE ASSOCIATION STUDIES: METHODS, CHALLENGES, APPLICATIONS

Genome-wide association study is an approach that aims to identify genomic coding and non-coding variants that demonstrate a statistical correlation with a particular phenotype, e.g., risk for a disease or a trait [2]. Individuals with the same ancestry but with phenotypical differences are enrolled, and the statistical association between each genetic variant and phenotype is tested. Different classes of genetic variants can be tested, with single-nucleotide polymorphisms (SNPs) being the most common.

GWAS commonly identifies groups of related SNPs that demonstrate a statistically significant association with specific traits, referred to as genomic risk loci [1]. Typically, such studies do not directly allow the identification of which variants among the significant SNPs are biologically causal. Many SNPs are associated since they correlate with the actual causal variants. This process is facilitated by linkage disequilibrium – non-random association of the alleles at different variants [3]. This phenomenon occurs when there is a restricted recombination between genetic loci, leading to the non-random inheritance of allele combinations. Often, these functional variants influence the regulation of the expression of nearby genes rather than altering the protein-coding sequences of the genes themselves. As a result, GWASs are primarily effective in pinpointing genetic loci linked to particular traits rather than directly discovering the functional variants or genes. Subsequent bioinformatic analysis and functional characterization are required to delineate these functional variants and genes further.

The typical workflow for GWAS comprises several steps [1]. According to Uffelmann et al. (2021), the study usually takes the following steps:

1. Data Acquisition

Research data can be sourced from study groups or genetic and phenotypic data from biobanks or other databases. It is crucial to account for potential confounders and ensure that recruitment methods do not lead to biases [4]. One should also consider correct phenotype definitions and the selection of the appropriate controls.

2. Collection of Genotypic Data

Microarrays to capture common genetic variations. In contrast, next-generation sequencing techniques are used for comprehensive genome or exome sequencing (WGS/WES).

3. Quality Control Measures

Quality assurance processes are implemented at both the experimental (e.g., genotype determination and DNA quality checks) and computational stages (e.g., removal of problematic single-nucleotide polymorphisms (SNPs) and samples, identification of sample population structures, and principal component analysis).

4. Genotype Phasing and Imputation

Genetic data can be divided into haplotypes (phased). It is needed since most datasets use genotyping arrays, and those usually contain only a small subset of variants in the human genome – the genotyped ones. Missing genotypes can be inferred (imputed) using specific bioinformatic methods, e.g., Beagle for phasing and Eagle for imputation [5, 6], while using

data from corresponding genetic reference panels such as the 1000 Genomes Project or TopMed [7, 8]. Such reference panels can be used to impute common variants in most populations. In the case of rarer variants that are more population-specific, the corresponding more population-specific reference panels are used [9 – mitt et al]. The imputation involves estimating unmeasured genotypes of specific SNPs based on the observed genotypes of other SNPs.

5. Genetic Association Analysis

Tests are conducted to find associations between each genetic variant and trait, employing models that might be additive, dominant, or recessive [9]. Additive models assume the trait effect increases linearly with the number of risk alleles. Dominant models consider the trait effect when at least one risk allele is present, and recessive models assess the trait effect only when two copies of the risk allele are present. Most GWASs, including the ones done for this project, use additive models.

These analyses adjust for confounders, such as population structure, and account for the need to manage multiple hypothesis testing. This is done by applying a P-value threshold, typically 5×10^{-8}. It corresponds to a Bonferroni correction for approximately one million independent tests, a rough estimate of the number of independent common genetic variants in the human genome typically tested in a GWAS. By setting the threshold at 5×10^{-8}, researchers aim to control the genome-wide false positive rate at 0.05, ensuring the findings are robust against the multiple comparisons problem inherent in such large-scale genetic analyses [1].

In GWASs, researchers typically employ linear or logistic regression models to examine associations, choosing the model based on the nature of the phenotype – linear for continuous traits like height, blood pressure, or body mass index and logistic for binary traits such as disease presence or absence [1]. Covariates like age, sex, and ancestry are incorporated to mitigate stratification and counteract confounding demographic factors, though this inclusion may lessen the statistical power for binary traits in specific samples [11]. Ancestry is usually obtained using principal component analysis (PCA) since it can identify and account for genetic population structure by summarizing the genetic variation in the data into principal components, which often correlate with ancestral backgrounds [12]. Including PCs as covariates helps control these differences, reducing the risk of spurious associations.

To enhance genomic discovery's statistical power and improve stratification control, an additional random effect term particular to each individual is used in mixed models, which requires more computational effort [13]. This computational demand can be offset by efficient tools like fastGWA [14]. One of the most widely used tools is REGENIE, designed explicitly for large-scale studies [15]. It employs a two-step approach to efficiently handle the computational burden of analyzing high-dimensional genetic data. An important aspect is that it allows parallel analysis of many phenotypes, saving time and resources. REGENIE fits a whole-genome regression model in the first step to account for polygenic effects and reduce data dimensionality. In the second step, association testing is performed using the residuals from the first step, enabling rapid and accurate identification of genetic variants associated with numerous complex traits in one run [15]. REGENIE's efficient algorithm and scalability make it suitable for analyzing large biobank datasets, where traditional methods may be computationally prohibitive. Regarding population control in genetic association analysis, it is

essential to state that current mixed-effects models better account for population structure by incorporating a relatedness matrix in their analyses.

6. Meta-analysis:

This is an optional step to increase the overall sample size and enhance the power to detect genetic associations that might be too weak to identify in smaller, individual studies. In a meta-analysis, findings from several smaller datasets are integrated using uniform statistical methods to enhance the robustness of results [16]. Each dataset undergoes a separate GWAS analysis, and the resulting statistical summary statistics, such as beta coefficients, odds ratios, and standard errors, are combined using meta-analysis methods. Examples of such methods are fixed-effects models that assume that one actual effect size is standard to all studies and that any observed variation between studies is due to sampling error and random-effects models that allow variations in the effect size between different studies [17].

7. Replication of Results

Reproducibility of findings is sought through either internal replication within the study or external replication in a genetically similar but independent cohort that does not overlap with the initial study group. The replication dataset is independent of the discovery dataset, even if these come from the same study/biobank. This is an integral part of the study since it helps to ensure that the results were obtained not due to random chance, biases, or specific characteristics of a single cohort. By replicating the findings in different cohorts, one can confirm the reliability and validity of their results. Checking reproducibility strengthens the scientific evidence and increases confidence that the genetic associations identified are genuinely related to the traits or diseases studied rather than being false positives. Additionally, reproducible findings are more likely to generalize to broader populations [18].

8. *In Silico* Post-GWAS Analysis and Experimental Validation

Post-genome-wide association studies involve computational analyses leveraging bioinformatic methods and external databases for further genetic insights, including SNP and gene function mapping, pathway and correlation analyses, and risk prediction [19]. Following GWAS, experimental techniques like CRISPR or massively parallel reporter assays and functional validation in human traits or disease models are used to test functional hypotheses [20, 21].

All these steps are represented in Figure 1.

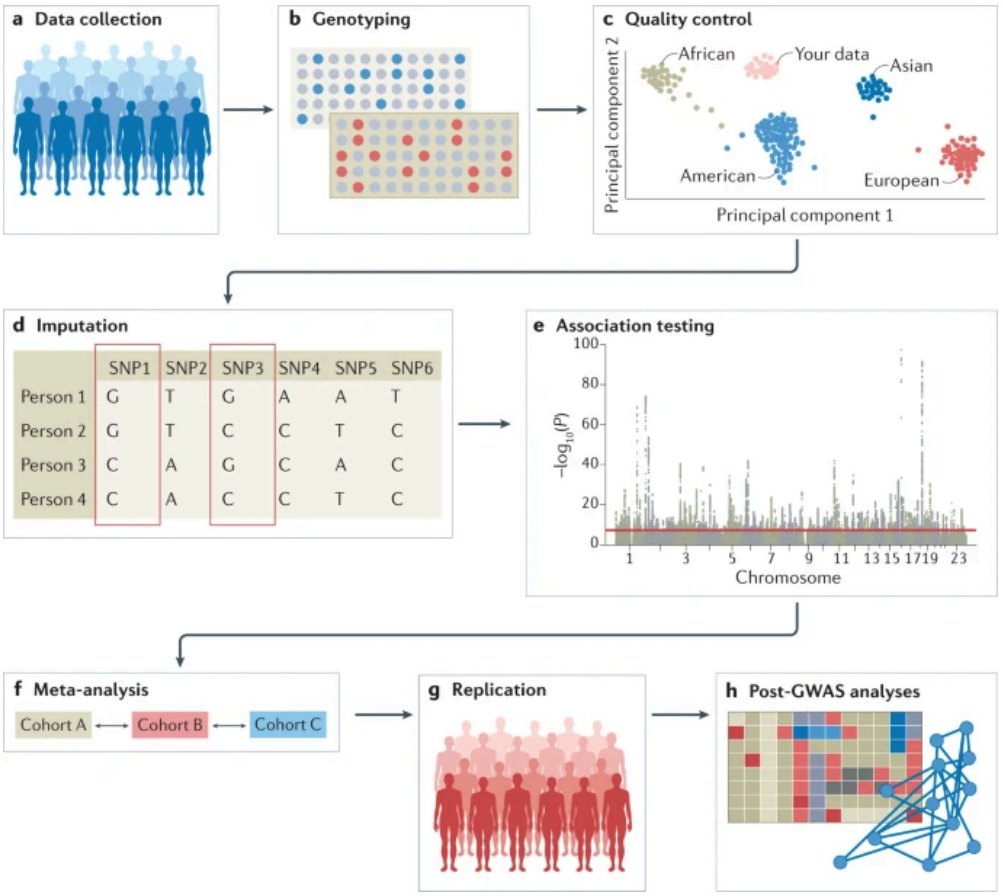

Figure 1. Overview of steps for conducting GWAS [1].

1.2 POST-GWAS ANALYSIS: AN OVERVIEW OF APPROACHES

Traditionally, GWAS outputs only statistical associations; it is not straightforward to interpret how they relate to biological mechanisms [1]. As a result, several post-GWAS methodologies have been developed to address these gaps. In the genomics field, researchers are interested in questions that are not entirely understandable from GWAS results directly:

1. GWAS results alone do not estimate the heritability of traits. Summary-statistics-based methods like Linkage Disequilibrium Score Regression estimate the proportion of phenotypic variance explained by all common variants [22]. The other, less widely used methods include LDAK and HDL [23, 24].

2. While GWAS can identify lead variants associated with traits, it does not pinpoint the causal variants. Due to linkage disequilibrium and allelic heterogeneity – a phenomenon when different mutations in one gene cause the same disease or condition, the variant identified as the lead one might not be causal [25]. Fine-mapping techniques, such as CAVIAR or FINEMAP, help narrow down the list of candidate variants [26, 27]. Variant annotation tools like ANNOVAR and RegulomeDB provide additional functional context to prioritize these variants [28, 29].

3. Identifying the genes affected by significant variants is another challenge. The closest gene approach is often used, but more sophisticated methods like expression Quantitative Trait Loci (eQTL) mapping can link variants to gene expression changes, thus highlighting potential causal genes [30].
4. Understanding the broader biological context of GWAS findings requires enrichment analyses. Tools like FUMA (Functional Mapping and Annotation of GWAS) provide pathway and tissue enrichment analyses [31]. Similarly, DEPICT (Data-driven Expression Prioritized Integration for Complex Traits) offers insights into the biological processes and pathways the associated variants affect [32].
We will take a closer look at the methods used to emphasize these questions further.

1.2.1 Lead SNP Identification Based on Clumping in pGWAS

In the context of post-GWAS analysis, identifying lead SNPs is considered one of the most critical steps that precede further analyses. By definition, lead SNPs are independent SNPs that have reached a minimum P-value threshold, meaning they are independent of each other at the LD threshold [33]. Identifying these SNPs often involves a technique known as "clumping" [34]. It reduces the complexity of genetic data by focusing on the most informative variants.
Clumping typically involves several steps:
1. Selection of a primary SNP:
The SNP with the lowest P-value within a genomic region is initially selected as the primary or lead SNP [33].
2. LD Thresholding:
Surrounding SNPs in high LD (as measured by R^2) with the lead SNP are identified; R^2 quantifies the extent to which the presence of an allele at one SNP predicts the presence of an allele at a second SNP. A standard threshold for R^2 is set, beyond which SNPs are considered to be in significant LD with the lead SNP. The most common R^2 thresholds are 0.1 and 0.2. They are typically set low to remove the variants in partial LD [35].
To calculate R^2, one must compare the observed frequency of haplotypes (combinations of alleles at the two SNPs) in the population with the frequencies expected if the alleles were independently assorted. Mathematically, R^2 is calculated as

$$R^2 = D^2/(p\_A * p\_a * p\_B * p\_b)$$

D^2 is the difference between the observed and expected haplotype frequencies, and p_A, p_a, p_B, and p_b are the allele frequencies of the respective SNPs [36]. An R^2 value of 1 indicates perfect LD, meaning the alleles at the two SNPs are always inherited together. In contrast, an R^2 value of 0 indicates no LD. It is important to note that ideally, the LD is inferred from the same genotype data as GWAS was done [35]. However, ancestry-matched external reference panels are also often used [37].
3. Exclusion of SNPs in High LD:
SNPs that exceed the predefined R^2 threshold relative to the lead SNP are "clumped" together and excluded from further consideration as independent signals. This process assumes that their association signals are not distinct but reflect the lead SNP's effect.

This process is repeated across the genome to evaluate all genomic regions and identify the most significant independent SNPs. For each lead variant, a proxy SNP is identified. A proxy SNP is a genetic variant that is in linkage disequilibrium with another SNP, meaning it is closely associated and can serve as an indirect marker for the genetic region of interest.
Several computational tools are employed to perform the clumping in post-GWAS analyses. The most popular is PLINK – a tool that offers an "ld-clump" function designed for this purpose [38]. It allows researchers to set up parameters such as P-value thresholds, R^2 values for LD, and physical distance limits to identify lead SNPs effectively. Another tool that may be used for clumping is FUMA – an online platform with a double-clumping method. The first clumping round is based on a genome-wide significant P-value threshold (5×10^{-8}) and initial R^2 threshold of 0.6, followed by a second round to identify the most independent SNPs based on a stricter R^2 threshold of 0.1 [31]. This double-clumping process outputs a set of lead SNPs statistically significant and largely independent in linkage disequilibrium. This approach ensures that the identified SNPs represent distinct genetic signals, reducing redundancy and increasing the clarity and interpretability of the genetic associations. However, because it is a web-based platform, FUMA is unsuitable for large-scale analysis, and PLINK was used in our project [38].
Clumping SNPs is essential for several reasons:

1. It simplifies the analysis by reducing the number of SNPs to consider in further interpretation.
2. It helps to better interpret and summarise the results by removing redundant signals from the results and keeping only one variant per signal.
3. Clumped variants can be used for further genetic investigations.
4. The clumped variants can also be used as reference points in defining associated genetic loci. These loci are typically used in downstream analyses such as colocalization or visualizations with regional plots [34].

1.2.2 Fine-mapping in post-GWAS analysis

Fine-mapping is the class of techniques aiming to pinpoint the most likely causal variants within the trait-associated genomic regions identified by GWAS [33]. This process addresses the challenge of linkage disequilibrium (LD), where multiple SNPs are associated due to their proximity and correlation rather than a direct causal relationship. Fine mapping can be done directly on the primary genotype and phenotype data; however, several methods have been developed to use GWAS summary statistics for such analysis [39]. Most fine-mapping methods on summary statistics use some form of Bayesian statistics via different algorithms. The simplest method, ABF (Approximate Bayesian Factor), assumes a single causal variant in the locus [40]. More recently developed methods such as SUSIE and FINEMAP offer the advantage of not assuming a single causal variant in a locus [41, 28]. However, these methods require an LD matrix, which ideally should be derived from the same sample used in the analysis to ensure accuracy. Bayesian approaches are compelling in fine mapping as they calculate each variant's posterior probabilities of causality. These methods produce credible

sets (CS) of variants, which collectively have a high probability of containing causal variants, typically above a 95% or 99% threshold [41].
Credible sets are critical for narrowing the list of potential causal variants from a broad set identified through genome-wide association studies. The crucial difference between a 95% and a 99% credible set lies in their levels of certainty and inclusiveness. A 95% credible set includes variants with a 95% probability of containing the causal variant. In contrast, a 99% credible set has a higher probability threshold, containing variants representing a 99% likelihood of encompassing the causal variant. Consequently, 99% credible sets are typically larger than 95% credible sets because they aim to increase the probability that they contain the causal variant, thus encompassing more variants to ensure a higher certainty. This difference reflects a trade-off between precision and confidence: 99% credible sets offer more assurance than 95%, but they may also include more non-causal variants, potentially complicating further analysis [42].

1.2.3 Prioritization of causal genes

The prioritization of the causal genes is one of the crucial steps in studying the human complex traits genetics. Knowing which gene is causal helps precision medicine target them to study the trait and develop the mechanisms to cure or adjust it. The simplest way of prioritizing potentially causal genes in the locus is based on the measurements of the distance between the GWAS hit and the closest gene. There are two main approaches to measuring this distance – to the closest transcription start site (TSS) or the gene body. This prioritizing method assumes that the gene closest to the GWAS hit is the most likely to be affected by the variant [43]. The gene with the closest TSS is likely relevant because regulatory sequences, such as transcription factor target sites, are often located near the TSS, influencing gene expression. Similarly, SNPs mapping into the gene body could affect gene function by altering coding sequences, splicing, or regulatory regions within introns or exons [44]. It is also possible that the variant is inside the gene; in this case, it is checked, and the measurements are performed accordingly.
Despite its simplicity, the distance-based method for gene prioritization has been shown to perform relatively well. In their study, Tambets et al. (2023) state that the precision and recall heavily depend on the truth sets – chosen gene-variant pairs that assign each GWAS variant to the closest gene [45]. In the same work, the authors declare that the simplistic distance-based approach outperformed three widely used Bayesian methods for colocalization: coloc.susie, coloc.abf, and colocalization posterior probability (CLPP) [46, 47, 48]. Additionally, it has been shown that distance to the gene body (when the GWAS variant fits between the start and end positions of the gene) instead of the closest transcription start site increased both precision and recall [45]. However, TSS mapping is still a simplistic method. More sophisticated approaches involve using QTL (Quantitative Trait Loci) data to prioritize causal genes. The assumption is that if the same variant affects the trait and gene/protein expression levels, that gene is likely causal [49]. Additionally, colocalization methods, which analyze the association profile of all SNPs in a locus, can determine if the genetic signals for the trait and gene expression colocalize, providing more substantial evidence for a causal relationship [46].

Usually, before the prioritization of the genes via distance-based methods, the signals are fine-mapped to identify the causal variants. This method is often performed together with Mendelian Randomisation to enhance the precision of gene prioritization. According to Zuber et al., 2022, mendelian randomization and colocalization are two statistical approaches that summarize data from genome-wide association studies to elucidate relationships between traits and diseases [50]. Despite their similarities, their objectives, implementation, and interpretation differ, reflecting their development for distinct scientific communities. Mendelian randomization evaluates whether genetic predictors of an exposure are associated with an outcome, interpreting such associations as evidence of a causal effect of the exposure on the outcome. In contrast, colocalization examines whether the same or distinct causal variants influence two traits. Each method has its own false positive rate, so one can intersect them to get more precise results while combining these methods. Summary statistics-based Mendelian Randomisation (MR) is conducted between GWAS summary statistics and (e)QTL summary statistics to determine if a gene affects a phenotype, using variants that affect gene or protein expression as instruments [51]. Unlike distance-based and colocalization methods, MR has an advantage as it also assesses the directionality of the effect, establishing whether the gene or protein influences the phenotype rather than the phenotype influencing gene or protein expression [52].

1.2.4 Estimating heritability with Linkage Disequilibrium Score Regression (LDSR)

Narrow-sense heritability h^2 is a genetic measure that tells us how much of a trait's variation can be explained by additive genetic differences among individuals. From it, SNP-based heritability can be distinguished to obtain more precise results. Yang J. et al., 2010 define SNP-based heritability as the proportion of phenotypic variance explained by SNPs variations [53]. The authors showed that common SNPs explain 45% of the variance in height using a mixed linear model in a GWAS dataset of unrelated individuals. For comparison, general narrow-sense heritability for this trait was previously estimated to explain 5% of the variance; this was explained by a smaller dataset [54].

As the estimated variance explained by genome-wide significant SNPs is often much more minor than the heritability estimated from family or twin studies, it is essential to state that family and twin studies can give inflated estimates for complex traits [55]. Siblings usually also share the environment, and if the environment affects traits, we get a higher h^2 estimate than genetic data alone. The findings suggest that for complex traits such as height, brought as an example in some studies, there may be numerous common genetic variants with effects too subtle to meet the strict significance threshold ($P < 5\times10^{-8}$) in genome-wide association studies, despite using sample sizes considered sufficient at the time (n = 1,000 to 10,000 samples before 2010) [56]. This explanation aligns with the idea of polygenic inheritance.

Nevertheless, SNP-based heritability is widely used in post-GWAS analyses. One of the most popular methods is Linkage Disequilibrium Score Regression (LDSR), sometimes called LDSC [57]. It is a regression-based technique that estimates SNP heritability. The LD score quantifies the extent to which each SNP tags neighboring SNPs, with a higher LD score

indicating a greater likelihood of tagging causal variants and thus exhibiting a stronger average association [58].
Each SNP within a population reference panel is assigned an LD score that reflects the genetic variation it accounts for. This helps to differentiate population stratification biases from genuine polygenic influences [59]. By regressing summary statistics from multiple SNPs against the LD scores of each SNP, LDSC assesses the genetic variance each SNP explains. The regression's intercept quantifies potential biases, while the slope estimates the overall variance in the phenotype that all SNPs account for, essentially measuring the heritability derived from the SNPs used in calculating LD scores [58].
LDSC operates under several assumptions. Firstly, it assumes that the variance explained by each SNP is uncorrelated with its LD score, stating that rare SNPs might exhibit larger effect sizes than common ones [58]. However, this may not consistently hold, especially in traits where LD scores correlate with minor allele frequency (MAF). Secondly, the accuracy of LDSC depends on the alignment between the LD reference panel and the target sample. A mismatch due to genetic heterogeneity between the reference and the sample can lead to less accurate estimates. Additionally, LDSR cannot distinguish effects due to genetic drift, as these do not correlate with LD and thus remain indistinguishable by this method [59].

1.3 LARGE-SCALE GENOMIC STUDIES

As genomics has expanded, many large-scale genomic projects have been undertaken to enhance the characterization of genetic diversity across various populations. These initiatives have significantly augmented our knowledge about human genomics, including substantial contributions from European, American, and Asian ancestries. In past years, projects such as the International HapMap and the 1000 Genomes Project have extensively cataloged genetic variation, reducing reliance on extensive sequencing, facilitating flexible definitions of haplotypes, and improving the resolution and refinement of association signals [60, 61].

Despite the advancements, early disease-focused case-control studies often needed more comprehensive data from non-European ancestries. One reason is that disease risk prediction via polygenic risk scores only works well when the GWAS is conducted in one population and the prediction is applied to another. Genetic variants associated with diseases can differ significantly across populations due to varying allele frequencies and environmental interactions. Consequently, polygenic risk scores derived from predominantly European-ancestry GWAS may not capture the complete genetic architecture of diseases in other populations, leading to less accurate risk predictions [62]. However, with the advent of the biobank era, there has been a noticeable increase in the representation of non-European donors in major biobanks like BioBank Japan and the All of Us Research Program [63, 64]. However, European ancestries continue predominating in several major biobanks, including the UK Biobank (UKBB), FinnGen, deCODE, and the Estonian Biobank [65, 66, 67, 68].

Meta-analysis studies are also instrumental for investigating the genetics of complex traits; they involve aggregating and analyzing genetic and health data from multiple biobanks. These

studies identify genetic variants associated with diseases, traits, and health outcomes across cohorts. By combining data from various biobanks, researchers can increase the sample size and statistical power, enhancing the ability to detect genetic associations that may be missed in smaller studies. One example of such extensive studies is GWAS done by Yengo et al., 2022 [69]. Their meta-analysis combined previously done GWAS of the 281 studies reaching a total sample size of around 5.4 million [69]. Additionally, biobank meta-analyses help validate findings across different populations and improve the generalizability of results, ultimately contributing to a better understanding of the genetic basis of diseases and informing personalized medicine approaches.

1.3.1 Biobanks

Biobanks arose as establishments along with the development of the genomics field and the growing implementation of large-scale studies [70]. The sample types, population, and recruitment method can characterize them. All these characteristics are primarily associated with the type of research planned for the samples. Two prominent examples of biobanks may be population- and disease-focused [71]. The first type, population-based biobanks, achieve high recruitment rates and are particularly effective for exploring the prevalence of genetic disorders and identifying genetic markers associated with widespread diseases. Disease-specific biobanks, on the other hand, might experience lower recruitment rates [71]. They serve as valuable repositories for research focused on the genetic responses to therapeutic interventions, delineating genetic and environmental factors linked to specific conditions and discovering biomarkers that enhance disease classification and monitor its progression.

Numerous large-scale population-based biobanks are available to the scientific community, offering datasets from thousands of extensively genotyped and deeply phenotyped individuals. These participants are typically characterized through detailed questionnaires, comprehensive laboratory measurements, and linkage to electronic health records without being selected for specific disease traits. The attributes of these biobanks, such as the diversity of the samples, the demographic characteristics of the populations sampled, and the participant recruitment methodologies, critically influence the nature and quality of the research outcomes that can be derived from these resources [72].

For instance, population-based biobanks that achieve high recruitment rates are particularly valuable for conducting studies to determine the prevalence of genetic disorders and identify genetic markers associated with common diseases. Conversely, disease-specific biobanks or those confined to a single location, which may face lower recruitment rates, are instrumental in supporting research focused on the genetic determinants of response to therapeutic interventions, elucidating the interconnections between genetic and environmental factors in specific diseases and identifying biomarkers that improve disease classification and monitor progression [73].

Although biobanks were historically focused primarily on European populations, ethnic diversity is becoming more comprehensive. Examples of biobanks dedicated to non-European populations are China Kadoorie Biobank, BioBank Japan, the Mexican Biobank Project, and

the African BioGenome Project [74, 63, 75, 76]. Nevertheless, diversity-related issues remain a bottleneck of such studies. This lack of diversity limits the generalizability and applicability of findings across different ethnic groups, potentially overlooking important genetic variants in non-European populations [77]. Therefore, continuing to include diverse populations in GWAS is essential to uncover a more comprehensive spectrum of genetic variants and to ensure equitable healthcare advancements.

1.3.2 Estonian Biobank

The Estonian Biobank is a population-based biobank that contains about 20% of Estonia's adult population. All ~215,000 participants are genotyped with genome-wide genotyping arrays, while different data layers have been generated for the various subsets of participants (Table 1).

Table 1. Overview of the data layers available in Estonian Biobank (adapted from the Estonian Biobank website (Estonian Biobank, n.d. [78])

OMICS profiling	**Sample size**
Whole genome sequencing	3 000
Whole exome sequencing	2 500
Genome-wide genotyping arrays	200 000
Microbiome (Metagenomics)	2 509
Genome-wide methylation arrays	821
Genome-wide expression arrays	1 100
mRNA sequencing	600
Total RNA sequencing	50
Metabolomics (NMR)	200 000

Overall, the biobank was designed to approximate the adult Estonian population and to reflect the age, sex, and demographical distribution (83% of Estonians, 14% of russians, and 3% of other nationalities) [68]. However, due to recruitment limitations, the data's representativeness is incomplete. The main differences concern the sex distribution since, in the biobank, the percentage of females is approximately 66%, which distorts the actual distribution in the population (Figure 2). Estonian Biobank also has a higher representation of younger people and fewer non-Estonians. The differences between the current state of the Estonian population and the data distribution in the Estonian Biobank are represented in Figure 2.

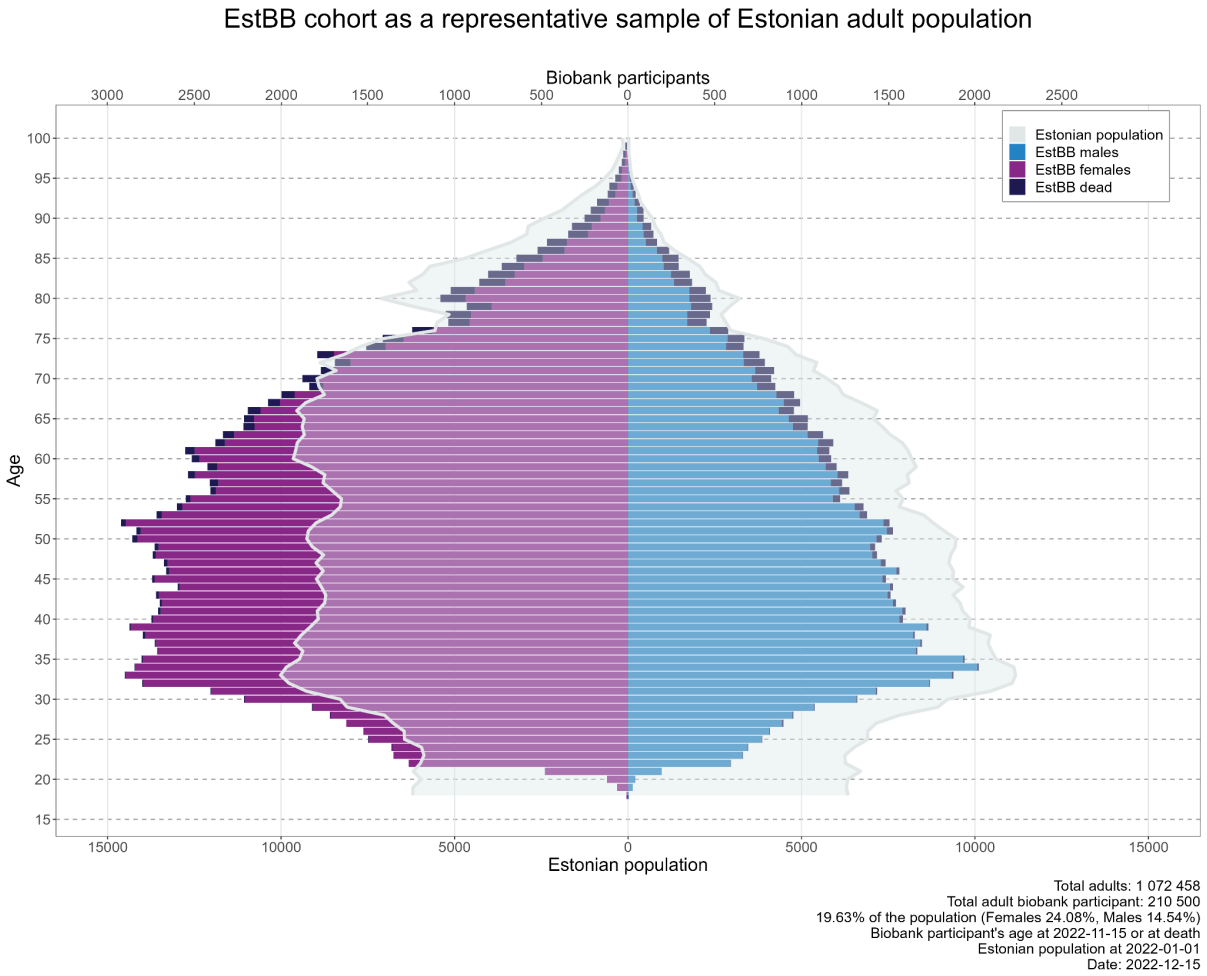

Figure 2. Differences between the Estonian population and its representation in the Estonian Biobank (Image made by Diana Sokurova).

The participants were recruited in two waves in different years: the first ~50,000 joined between 2003 and 2008, and the additional ~150,000 joined in 2018 and 2019. While collecting the data, questionnaires were collected from the participants, which included personal and genealogical data and information about one's physical and dietary habits, education and professional life, etc. The Estonian Biobank is periodically linked with national registries, hospital databases, and the national health insurance fund database, making it an example of a biobank linked with electronic health records [79]. All the diagnoses are recorded in ICD-10 codes, which will be discussed later. The genetic material was genotyped using an Illumina GSA microchip that contains more than 700,000 SNP markers [68]. The data was imputed with the utilization of a population-specific high-coverage reference panel that is comprised of 2244 Estonian individuals. This panel contains fewer haplotypes and variants; however, according to Mitt et al., 2017, the imputation confidence and accuracy of imputed low-frequency and rare variants were higher compared to the available international reference panels [9].

1.4 ICD CODING SYSTEM

Introducing disease classifications is essential not only for genomic studies but also for the development of medicine in general. These classifications help to structure the traits and make them comparable between different medical institutions in various countries. One example of such classification is the International Classification of Disease (ICD), created by the World Health Organization to accurately track traits within a population. The ultimate function of such classification is to enhance the international compatibility and usefulness of mortality statistics [79]. There were different variations of this system; the most current is ICD-11; however, ICD-10 and its variation for the United States, ICD-10 CM, are still widely used. It contains 22 chapters, each related to a particular type of trait [80]. One hundred seventeen countries worldwide, including Estonia, use the unchanged international version of ICD-10; however, in some countries, the system was adapted to the local specificity.

In ICD, the traits and conditions reported by clinicians are translated into the codes using a comprehensive set of rules. The code typically consists of up to 7 digits: characters 1-3 stand for one of 22 diagnosis categories, while characters 4-6 indicate clinical details, such as etiology or severity. The seventh character is an extension of whether it is needed [81]. While reported in sets, a single trait is selected as the underlying cause of death, while others are reported as non-underlying causes [79].

2 THE AIMS OF THE THESIS

The main goal of this study is to perform a large-scale interpretation of 5,035 GWASs in the Estonian Biobank. More specifically, we aim to:

- Identify lead variants and analyze them in terms of novelty;
- Estimate SNP heritability of the traits to determine what percentage of variation of a trait of the trait is explained by common genetic variants;
- Perform Approximate Bayesian fine-mapping to prioritize the causal variants;
- Prioritize the causal genes for the lead variants;
- Summarise and visualize obtained results to set the groundwork for further more focussed interpretation.

Our study will contribute to a more nuanced understanding of genetics and disease susceptibility.

3 EXPERIMENTAL PART

3.1 MATERIALS AND METHODS

3.1.1 Biobank data, genotyping and imputation

For this study, the data from the Estonian Biobank was used. Overall, 5,506 phenotypes coded in the ICD10 system were selected for the project; the total number of samples was 211,658. This selection of the phenotypes of all the available data in the biobank was explained by the fact that only the traits with more than 100 cases were used. The phenotype list used in the study was extracted from the internal database GEVA and prepared by Erik Abner.

The research activities involving biobank participant data were carried out under the approval nr. 1.1-12/624 by the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs), using data according to release application number 3-10/GI/31689 from the Estonian Biobank.

All participants were genotyped at the Core Genotyping Lab of the Institute of Genomics, University of Tartu, using the Illumina Global Screening Array v3.0_EST. This part of the study was done previously, and already prepared data was used for this thesis. Genotyping and PLINK format file creation were performed with Illumina GenomeStudio v2.0.4. If the call rate of the individuals was below 95%, if they were outliers in terms of heterozygosity (more than three standard deviations from the mean), or if their sex based on X chromosome heterozygosity did not match the phenotype data, these were excluded [9]. Variants were filtered before imputation based on a call rate below 95%, Hardy-Weinberg equilibrium P-value less than 1e-4 (for autosomal variants), and a minor allele frequency less than 1%. Initially, in build 37, genotyped variant positions were converted to build 38 using Picard. Phasing and imputation were conducted using Beagle v5.4 with default settings in batches of 5,000 [82]. A population-specific reference panel of 2,695 whole-genome sequenced samples and standard Beagle hg38 recombination maps were used for imputation. Samples not of European ancestry were removed based on principal component analysis. Duplicate and monozygotic twin detection was performed using KING 2.2.7, and one sample from each pair of duplicates was excluded [83].

3.1.2 GWAS

GWAS was performed using the NextFlow pipeline based on the regenie v3.2.1 – C++ program for whole genome regression modeling of extensive genome-wide association studies. The pipeline had previously been developed by the work group member Klavs Jermakovs, and most of the phenotypes interpreted in this thesis had already been GWAS'd previously. The analysis was performed in two steps: first, it fits a null model using leave-one-cation-out (LOCO) cross-validation, and the second step performs association testing using Firth's logistic regression for rare variant analysis to handle potential biases [84]. The covariates used in the GWASs include age, age 2, and 10 principle components derived from principal component analysis of the genotype data.

However, approximately 50 phenotypes were GWASd by the work's author. Before the GWAS, all the phenotypes were divided into non-sex-specific, male and female sex-specific groups, which were GWASd separately. Only the non-sex-specific ones were used in this

study, and all the following analyses were performed on this part of the data. Association analysis in the Estonian Biobank was conducted for all variants with an INFO score greater than 0.4, using the additive model implemented in REGENIE v3.0.3 with standard binary trait settings [85]. Logistic regression was adjusted for current age, age², sex, and the first ten principal components, analyzing only variants with a minimum minor allele count of 2. The general number of the genetic variants was obtained using filtering: for the variants with minor allele frequency > 0.01, the INFO score was set to be > 0.4. For the minor allele frequency < 0.01, the INFO score was set to >0.8.

3.1.3 Quality Control

Quality control was performed using a set of custom R scripts. The initial script was executed in the High-Performance Computing Center to pre-process the data for further analysis. This step was crucial for converting GWAS output results into summary statistics values that will be used for further study. The script also identified lead variants, ensuring the significant associations met the established P-value threshold of 5×10^{-8}. Key statistical metrics, such as the lambda inflation factor, were calculated to detect any potential biases or anomalies in the data and passed down to further analysis. The second script checked the integrity of summary statistic files by calculating the number of variants and chromosomes in each file and filtering variants to a minor allele frequency greater than 0.01. Incomplete GWASs were flagged as failed runs and were re-ran. Associated loci and lead variants were identified through distance-based clumping, using a P-value threshold of $P = 5\times10^{-8}$ and iteratively removing variants within a 1Mb window of each lead variant. For this assessment, the visualizations were done as well.

3.1.4 Identification of novel loci

The novel loci were identified using a custom Nextflow pipeline co-developed by the author and Urmo Võsa. This pipeline integrated previously known genetic associations from the GWAS Catalogue (version v1.0) and OpenTargets Genetics Portal (v2d version) [86, 87].
The pipeline was structured into four key modules, each responsible for specific tasks: parsing summary statistics files, clumping, and identifying proxies, and it overlapped with both datasets. For the linkage disequilibrium clumping, PLINK v1.9 was employed [38]. As an LD reference panel, an Estonian-specific subset of ~2500 whole-genome sequenced (WGS) samples from the Estonian Biobank was used [9]. This step aimed to refine the list of SNPs by removing those in high LD with lead SNPs within a specified genomic window in 1 Mb. Using an R^2 threshold of 0.2 and a window size of 250 kb, the analysis ensured that only independent SNPs were retained. Next, for each lead variant, the module identified proxy SNPs based on their physical proximity and LD to the lead SNPs, enhancing the resolution of the genetic mapping and ensuring that potential signals were thoroughly investigated.
The modules that overlap with the GWAS Catalogue and OpenTargets Database use R scripts to compare the identified SNPs and their proxies against the databases. This comparison was needed to distinguish novel signals from previously reported associations.

The pipeline is available at https://github.com/urmovosa/ldoverlapnf.
The output files were merged and cleaned with Python v3.11.5 library pandas v1.4.0 to maintain unique signals for further analysis.

3.1.5 ICD10 to ontology mapping

After all the loci were analyzed in terms of novelty and overlapping with the GWAS Catalogue and OpenTargets, the next step was to match the trait ICD10 codes used in EstBB GWASs to the ontology codes used by in GWAS Catalogue and OpenTargets databases to characterize the traits. The process was done semi-automatically, and the ontologies were sourced from two databases: the EFO-UKB-mappings repository and the OntoBee website, which provide a set of matched ICD10 codes to different ontologies [88, 89].
The ICD10 codes used in this project matched the corresponding ontology codes from those mapping files.
Based on this, all the significant loci were classified into four categories:
1. Novel signals were not previously found in the two databases with which we overlapped our results.
2. Known signals: These loci were associated with phenotypes that we could not map with the ontology code. These remained classified as known to acknowledge their previous association with at least one phenotype.
3. Novel associations known signals: They referred to newly discovered links identified for the first time through our analysis: the ICD10 code was matched to an ontology term.
4. Known associations: In cases where the mapped ontology term overlapped perfectly with the previously described ontology term, this was considered a known association.
This classification was adapted from Verma et al., 2023 [90]. The final step was visualizing these classifications with a custom Python v3.11.5 script with libraries matplotlib v3.7.2 and seaborn v0.12.2.

3.1.6 Fine mapping

For this step, a Nextflow pipeline was developed by Urmo Võsa and implemented by the author. The methodology involved using the function finemap.abf from R v4.3.0 package coloc (v3.1). After the summary statistics files were loaded into the pipeline, SNPs were filtered based on a minor allele frequency threshold of 0.01 and an imputation INFO score of 0.8. Additionally, a P-value threshold of 5×10^{-8} was applied to define significant loci. Lead SNPs were iteratively identified by selecting the most significant SNP and removing all other SNPs within the specified window, ensuring the independence of the lead SNPs. A window size of +/-1 Mb was used to determine the locus boundaries around each lead SNP, ensuring that the identified lead SNPs were independent and that the same window was set to define the locus boundaries. For each identified locus, fine mapping was performed using Approximate Bayes Factor (ABF) analysis [40]. The default prior probability of 1×10^{-4} was used for fine mapping. This analysis generated Posterior Inclusion Probabilities (PIPs) for each SNP, indicating the likelihood of each variant being causal. Credible sets were defined at

95% and 99% confidence intervals to prioritize variants with the highest probability of causality. Assuming that causal variants are in the results, the 95% or 99% probability sets containing the causal variant were defined for further analysis. These results were analyzed using the Python v3.11.5 library pandas v1.4.0.
The pipeline is available at https://github.com/urmovosa/FinemapAbf/tree/main.

3.1.7 Prioritization of causal genes

Prioritizing causal genes involves a systematic approach to mapping lead variants to relevant genetic regions. The main goal of this method is to prioritize one most likely causal gene for each locus. The set of human genes from the Ensembl database, release 111, was used as a reference [91]. Initially, the positions of all lead variants were assessed to determine if they fell within gene bodies. For variants that did not map within gene bodies, their positions were evaluated around the transcription start sites (TSS) of genes, with a distance threshold of less than 1 Mb. This analysis was done with R v4.3.0 package rtracklayer v1.60.1.
It is essential to consider that some variants may not be assigned to any gene. The results of this analysis were visualized using the Python v3.11.5 library matplotlib.pyplot v3.7.2, facilitating the interpretation and presentation of the spatial relationships between SNPs and genes.

3.1.8 Heritability Estimation

To estimate SNP Heritability, the custom Nextflow pipeline was co-developed by the author and Urmo Võsa.
The pipeline comprises five detailed modules that first parse summary statistics files and prepare them for the primary analysis; the heritability was estimated using LD Score Regression software (LDSC) v1.0.1. LD score reference for European LD Scores from 1000 Genomes were used, and all the variants were filtered to keep only a subset of HapMap 3 variants for the analysis [92, 60]. The LDSC tool is used twice per phenotype: once to compute liability-scale heritability adjusting for sample prevalence and once to calculate observed-scale heritability. The output log files from the LDSC computations to extract key results: heritability estimates, standard errors, and chi-square statistics were collected and summarised into a liability-scale heritability results table and an observed-scale heritability results table.
Overall, the pipeline can easily be reused or rescaled for further research. The Python script for processing the summary statistics and LDSC was adapted from Bulik-Sullivan et al., 2015 [93].
The pipeline is available at https://github.com/aalksnk/HeritabilitySampoNF.
As the study's next step, the effective sample size (N_{eff}) value was calculated to see how the h^2 value related to sample size [94]. It is necessary since some of the phenotypes represented in the study are relatively rare and have unreliable h^2 estimates. N_{eff} is calculated using the following formula:

$$Neff = 4 \div ((1 \div Ncases) + (1 \div Ncontrols))$$

After the calculations, the data was categorized based on the N_{eff} values in the following way: the value is less than 4,500, between 4,500 and 20,000, between 20,000 and 40,000, and 40,000 or higher. Corresponding to these conditions, labels "None," "Low," "Medium," and "High" are assigned. These labels show the confidence of h^2 values obtained from the analysis. If a row in the dataset does not meet the specified conditions, it is labeled as "Unknown." This classification was adapted from UKBB analysis from Neal's lab [94]. Finally, the results were visualized using the Python v3.11.5 library seaborn v0.12.2.
Regarding statistical significance, the P-value for this analysis was calculated with
$P = 0.05/number\ of\ the\ phenotypes$

3.2 RESULTS

3.2.1 GWAS

5,075 phenotypes in ICD10 codes were run through the GWAS pipeline. Most of them were carried out before this thesis research; however, some were run as part of it. Before the GWASs, all the phenotypes were divided into non-sex-specific, male, and female sex-specific. Only non-sex-specific data was used in this study.

3.2.2 Quality control

The quality control of 5,075 phenotypes was implemented to ensure the reliability and accuracy of the phenotype-genotype associations assessed. Each quality control step addresses a specific potential weakness in the dataset, ensuring that subsequent analyses are based on the most complete and accurate data possible. The quality control metrics and criteria were designed to filter out phenotypes with insufficient genetic variants, incomplete chromosomal data, and missing statistical values, which indicate the technical errors in GWAS analysis. A total of 40 phenotypes were identified to have fewer than 18,977,761 genetic variants. This value was checked to ensure that the analyses did not face technical issues that may prevent the correct detection of meaningful associations. Phenotypes not meeting this criterion might suffer from reduced statistical power; hence, they were removed from further analysis.

Two phenotypes were found to have less than 23 chromosomes. This may be explained by the errors in data collection or GWAS running. These phenotypes were also removed from the following analyses and will be re-analyzed in the future.

In terms of completeness of data, there were 1,083 phenotypes identified with at least one missing value in statistical fields, namely the chi-square and the logarithm of the P-value. It is important to mention that for some of the phenotypes checked (these examples were picked randomly, D22, E06.3, I10), the number of absent values per phenotype varies from 1 to 3. That makes these phenotypes sufficient for further analysis. Missing values in these fields can likely be attributed to specific alignments within the dataset. For instance, it may occur that, for a particular genetic variant, the division of cases and controls into genotype groups is such that it adversely impacts the calculation of certain statistics. Comparison with a few previously published GWASs from the same phenotypes as tested in our data indicated that the rest of the associations were similar to previously reported results, suggesting that these few absent values did not affect results substantially; hence, the part of these phenotypes had fewer variants than expected as well and were removed from further analysis.

After all quality control steps, 5,035 phenotypes were enrolled for further analyses. The analysis determined that 990 out of the 5,035 phenotypes possess at least one significant locus, totaling 2,848 significant loci across these phenotypes. This metric is pivotal for discerning phenotypes that demonstrate clear genetic influences; by focusing on phenotypes with significant loci, the study can prioritize those most likely to yield actionable genetic insights.

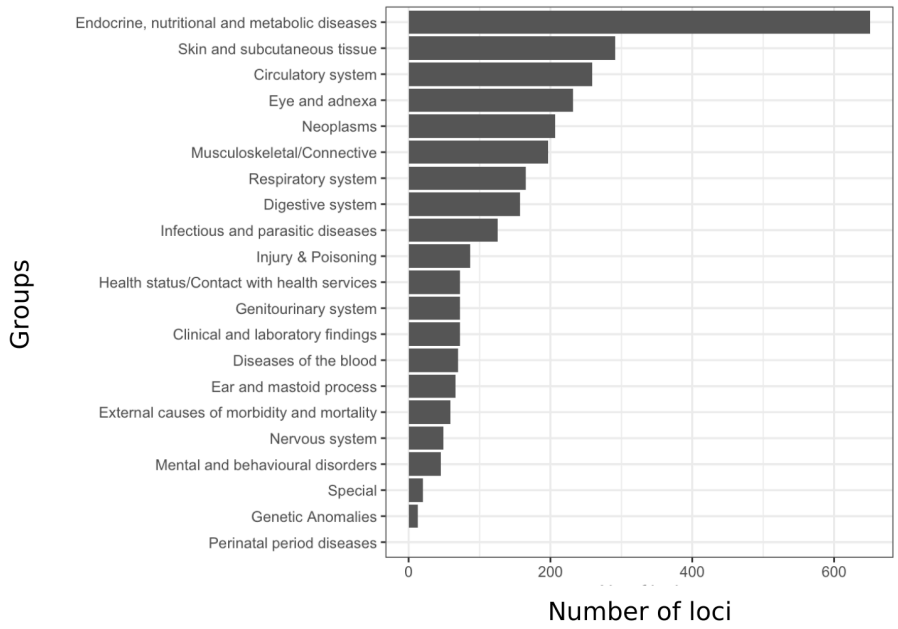

Figure 3. Overview of the number of loci by trait group.

The most notable observation is that phenotypes related to "Endocrine, nutritional, and metabolic diseases" show the highest number of significant loci, closely followed by conditions affecting the "Skin and subcutaneous tissue" and "Circulatory system" (Figure 3). This might indicate the high heritability of some of the phenotypes in these groups, which makes them areas of interest for further genetic research. The abundance of significant loci across different categories reflects the genetic diversity of these conditions and underscores the trait classes where the genetic factors are potentially influential. It is also possible that the phenotypes from these groups have higher numbers of cases; hence, we have the power to identify a substantial number of signals. As the next step, the relation between the number of loci and number of cases was explored.

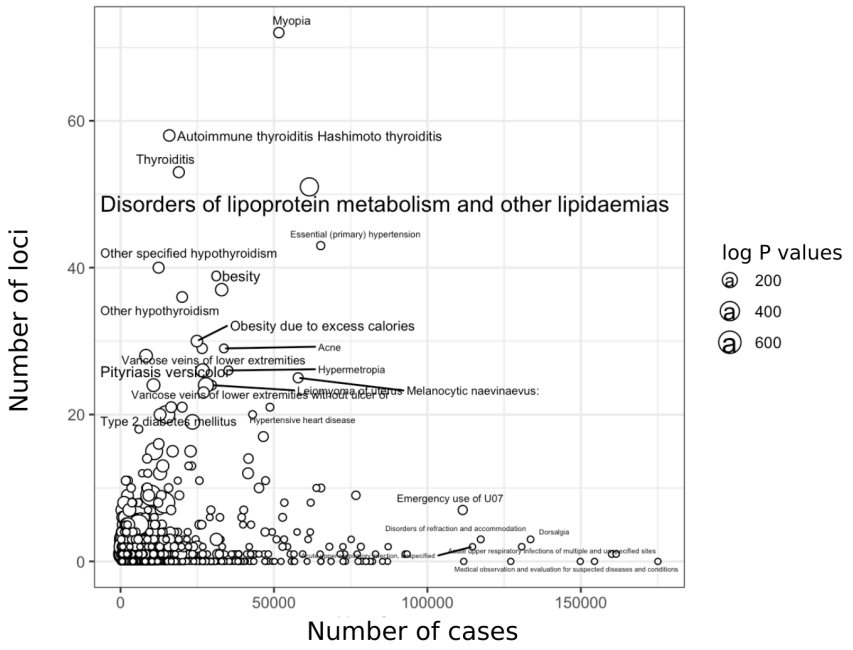

Figure 4. Relation between number of loci and number of cases. Larger circles represent a higher significance level in the association studies regarding log P values.

Exploring the relation between the number of loci and the number of cases shows some expected values for common conditions and provides insights specific to the dataset (Figure 4). For example, the traits from the "Respiratory infections" group have high numbers of cases; however, the numbers of significant loci for such traits are relatively low (e.g., Acute upper respiratory infections of multiple and unspecified sites (J06): 160082 cases, 1 lead variant; Acute pharyngitis (J02): 87204 cases; 2 lead variants.). Such diagnoses have heterogeneous causes, probably causing the limited number of associated loci. On the other hand, common conditions like "Obesity" and "Type 2 diabetes mellitus" also display a significant number of loci, underscoring the complex interplay of genetics in widespread health issues. The phenotypes such as "Myopia" and "Hashimoto thyroiditis" showing a high number of significant loci suggest a strong genetic predisposition. Myopia (H52.1) is the phenotype with the highest number of associated loci in our analysis with 72 lead variants. This phenomenon may be associated with the fact that there are 51,585 cases in EstBB, allowing the high power and also this phenotype seems to be under strong genetic control. It is visible specifically in comparison with other phenotypes in the same trait group, e.g. Conjunctivitis (H10) has more cases (72,459) but fewer associated loci (2). Later the heritability estimation for Myopia will be discussed in terms of population prevalence.

3.2.3. Identification of novel loci

Following the comprehensive analysis and subsequent filtering procedures, the study yielded a dataset comprising 2,745 genetic signals associated with 888 distinct phenotypes. Additionally, a successful matching between ICD10 codes and ontology terms was achieved

for 307 of these phenotypes, facilitating the comparison with previous GWASs. These matched phenotypes have 1,141 signals that were studied further.

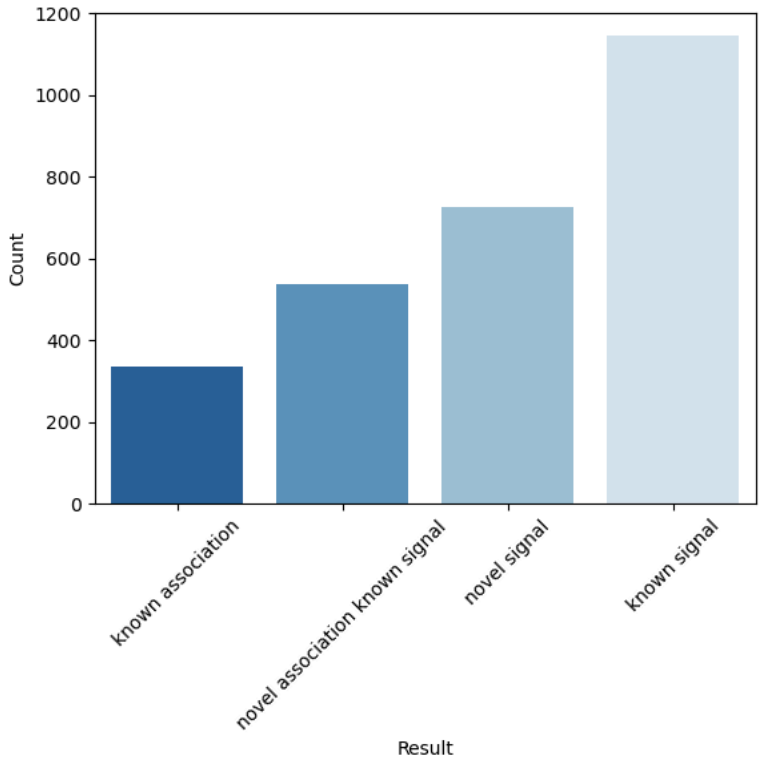

Figure 5. Distribution of the results in terms of novelty.

Out of 2,745 signals from our analyses, 2019 (26.45%) had never previously been associated with any phenotype, and 726 (73.55%) had been associated with at least one phenotype, prioritizing those as potentially interesting loci to follow up. (Figure 5). The results contain 1,145 known signals, confirming the presence of well-established genetic markers.
In order to evaluate the novelty of the results in more detail, the next analysis focussed on the subset of 1141 loci (303 phenotypes) where the phenotype had matched the EFO code, enabling the direct comparison with previous GWASs done for the same phenotypes.

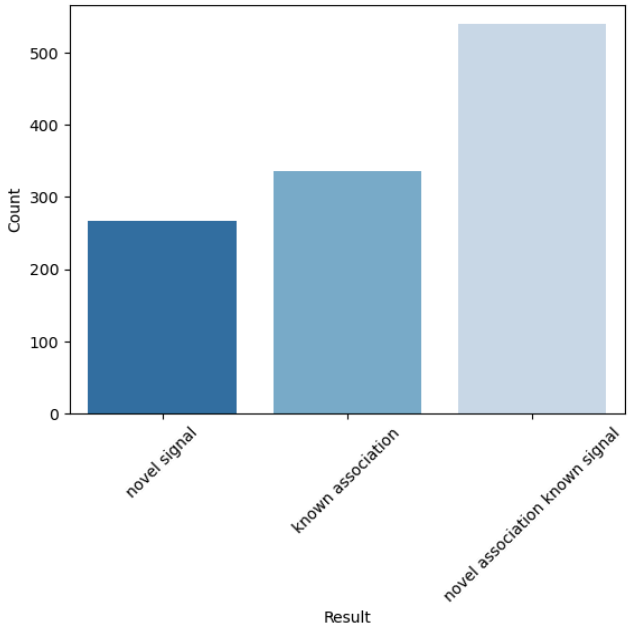

Figure 6. Distribution of mapped phenotypes in terms of novelty.

Out of those matched to the ontologies signals, 335 (29.36%) were found previously for the same phenotype, 267 (23%) were novel, and 539 (47.64.%) were interesting cases where the locus was previously associated with a different phenotype than the one analyzed in the current study (Figure 6).

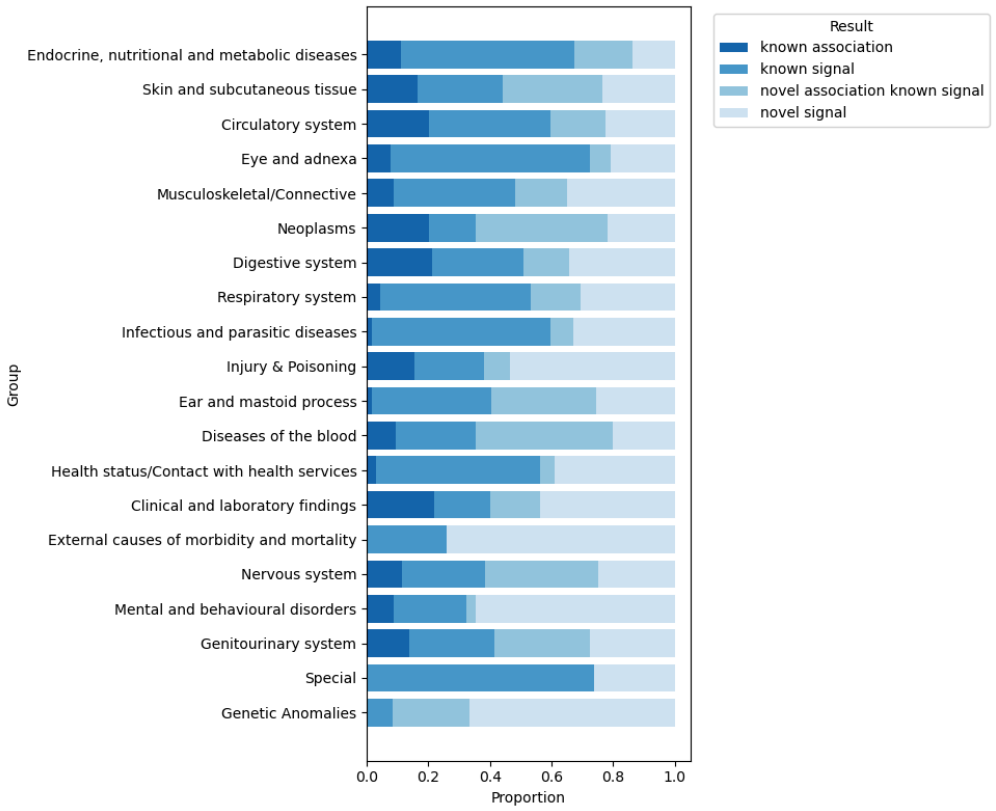

Figure 7. Proportional distribution of signals between phenotype groups.

In the full scope of diseases, some groups, like endocrine, nutritional, and metabolic disorders, have a higher proportion of known signals and associations, while others, like genetic anomalies, have a notable prevalence of novel signals, indicating areas of possible further genetic research (Figure 7). Some other groups, e.g., "Mental and behavioral disorders" and "Injury and poisoning", also have high proportions of novel signals. In contrast, the "Eye and adnexa" group and "Infectious and parasitic disorders" have an extensive percentage of known signals. One reason for these variations might be the specifics in the phenotypes and phenotype classes; for some phenotypes, Estonian Biobank data might have more power to detect associations than in previous studies, and for others, our study may have less power.

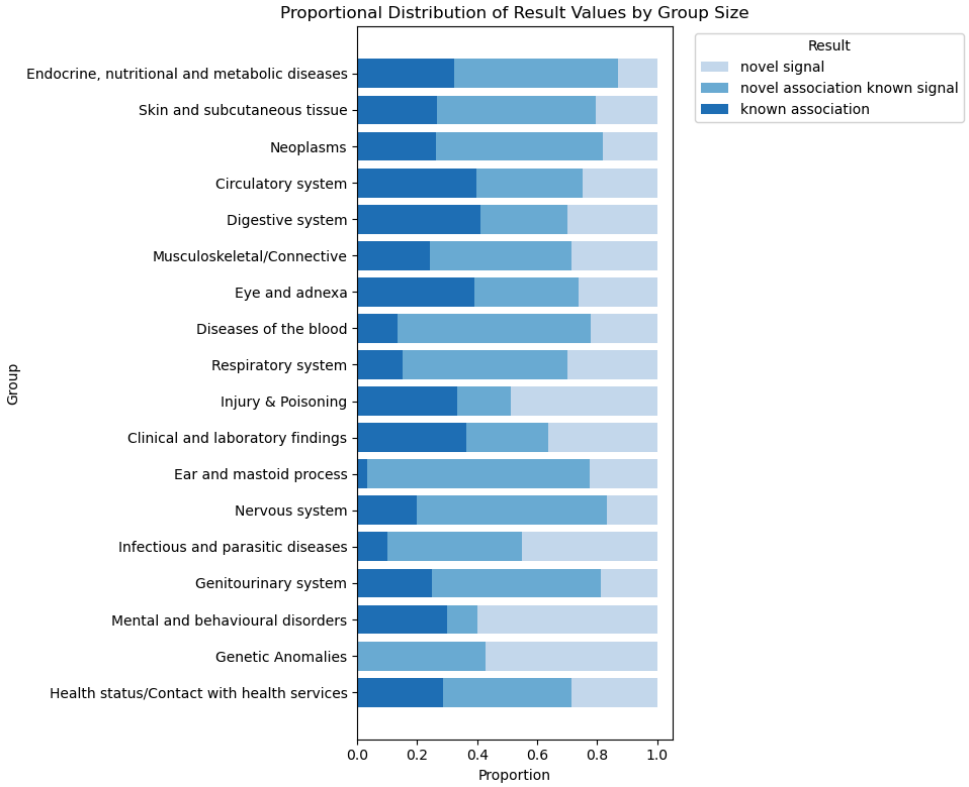

Figure 8. Proportional distribution of signals in matched to ontologies phenotypes between phenotype groups.

Taking a look only at the mapped phenotypes, the distribution of the signals varies (Figure 8). For example, mapped phenotypes in a group of genetic anomalies do not contain any known associations, while almost half of the associations are known for the phenotypes related to the digestive system. As said before, while analyzing these results one must keep in mind the differences in the number of cases for each group of diseases. The imperfect mapping between different phenotypes and differences in phenotype definitions between the ICD10 system and other systems in the GWAS Catalog and OpenTargets database may also explain some of these variations.

For a closer look, we can analyze the novelty of the signals on a scale of one phenotype. From a subset of the phenotypes containing one novel signal, Migraine (G43) was outlined as an example, as a relatively well-studied phenotype mapped to the ontology terms (Figure 9). In our analysis, there were 22,172 migraine cases, whereas the largest migraine GWAS meta-analysis consisted of 102,084 migraine cases [95].

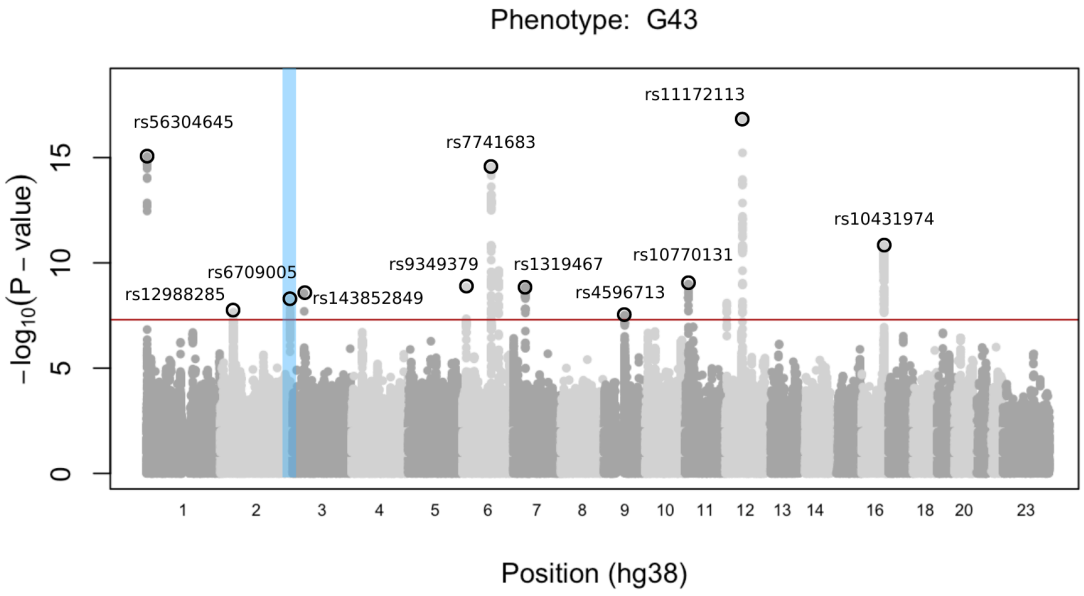

Figure 9. Manhattan plot for Migraine (G43). The novel signal is highlighted.

The analysis detected 11 signals, 1 novel – rs6709005 on chromosome 2. The automatic classification of the other signals is shown in Table 2.

Table 2. The automatic classification of the lead variants of Migraine (G43).

SNP	**Chromosome**	**Type of the signal**	**Previously associated trait**	**Previously done GWAS study ID**
rs56304645	1	Known association	Neurological disease	PMID: 29397368 [96]
rs1319467	7	Novel association known signal	Pulse pressure measurement	PMID: 34594039 [97]
rs10431974	16	Novel association known signal	Chronic obstructive pulmonary	PMID:33106845 [98]

			disease	
rs11172113	12	Known association	Migraine without aura	PMID:33959723 [99]
rs10770131	11	Novel association known signal	Brain region volumes	PMID:33959723 [99]
rs4596713	9	Known association	Headache	NEALE2_6159_1 [100]
rs7741683	6	Known association	Migraine	PMID:31015401 [101]
rs9349379	6	Novel association known signal	Heart disease	PMID: 22745674 [102]
rs143852849	3	Known association	Migraine without aura	PMID:33959723 [99]
rs12988285	2	Novel association known signal	Nose morphology measurement	PMID:33959723 [99]
rs6709005	2	Novel signal	None	None

Overall, there were 5 known associations and 5 novel association to the known signals detected (Table 2).

3.2.4 Fine mapping

2,834 loci were included in the analysis. As described previously, in genetic fine-mapping, a credible set is a group of variants containing the causal variant with a specified probability. Although 99% of credible sets are more precise, 95% contain fewer variants. This difference makes the 95% ones easier to interpret, specifically on a large scale, which is why these sets have been chosen for further analysis. There are 294 credible sets (10.3%) in the current dataset that contained only one variant, suggesting a high probability that these are indeed the causal variants.

In the results of genetic fine-mapping, the statistical analysis of the 95% credible sets revealed a notable discrepancy between the central tendency and the range of values. The mean size 95% credible set size was calculated to be approximately 342.23, substantially higher than the

median value of 16. This significant difference between the mean and median suggests a highly skewed distribution with extreme values. Indeed, the minimum value observed in the dataset is 1, while the maximum value is dramatically higher at 19,720. This wide range suggests that while most credible set sizes are relatively small, there are exceptional cases with much larger sizes, which could be due to the presence of loci with extensive linkage disequilibrium structures. It is essential to consider that in this work, a simplistic method of fine-mapping was used; hence the more complex methods may provide more precise results that will differ from the obtained ones.

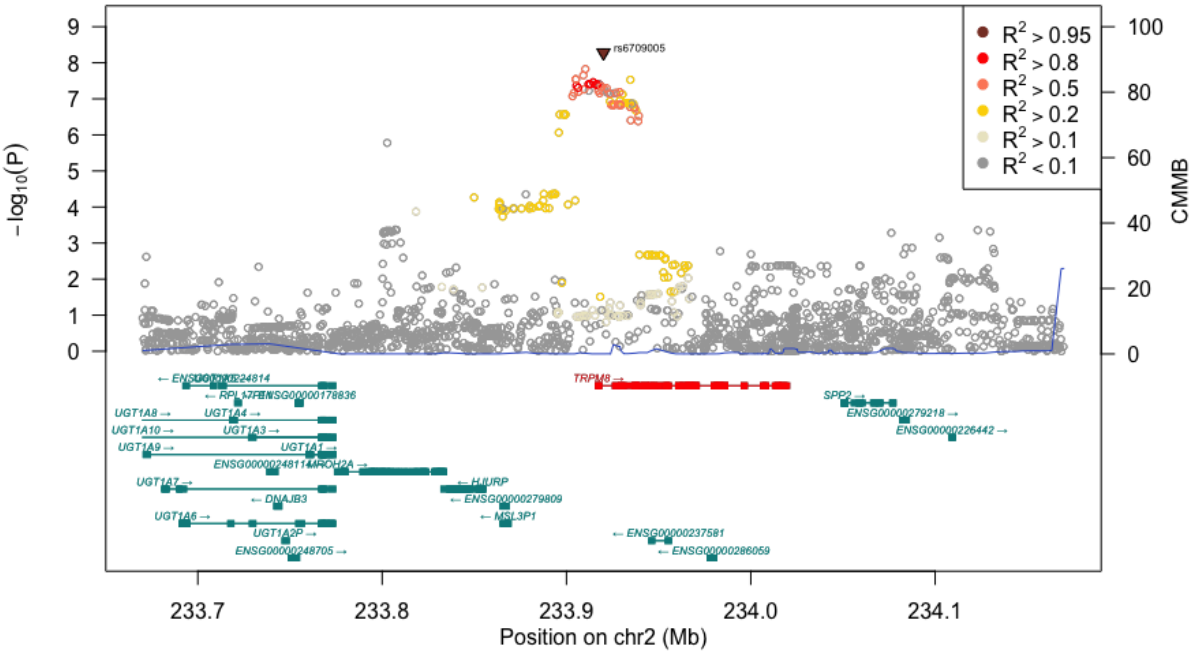

Figure 10. Regional plot for rs6709005 in Migraine (G43). The gradient of the colors represents the variations in R^2. Lead variant rs6709005 is the sole member of a 95% credible set and is outlined.

For the previously detected novel signal in Migraine (G43), the regional forest plot showed that according to the fine-mapping results, it is likely to be the causal variant (Figure 10). This preliminary overview gives us an insight into its causality; however, it is essential to consider more stringent fine-mapping techniques for further research. It is also shown that this signal is assigned to the gene TRPM8, which will be discussed further.

3.2.5 Prioritization of potentially causal genes

For this step, the lead variants identified with LD clumping were used. Altogether, all 2,402 lead SNPs were assigned to 804 genes. 338 variants (14.1%) are considered to be in the intergenic regions since they were not assigned to any gene through the analysis. Most of the assigned genes are protein-coding long noncoding RNA and other smaller subcategories, such

as different types of pseudogenes and missense RNA. The overall distribution can be seen in Figure 11.

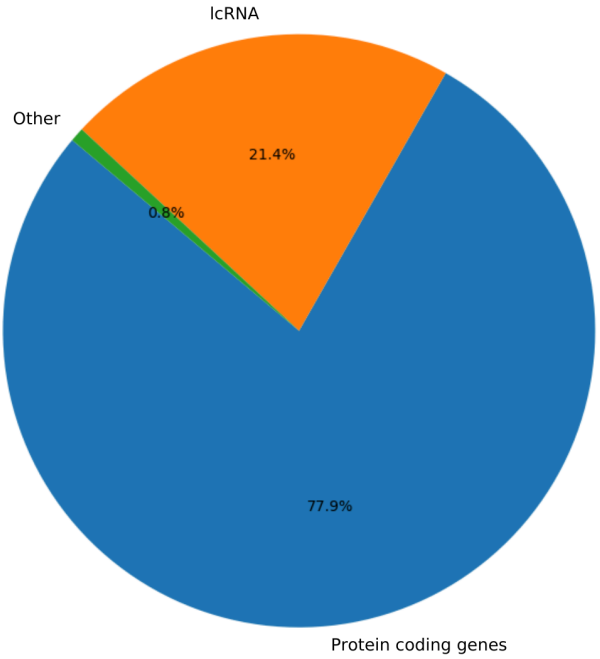

Figure 11. Distribution of the gene types assigned to lead SNPs.

In terms of the genes to which the largest number of various phenotypes were assigned, we have the results shown in Table 3.

Table 3. The genes that have the highest amount of associations.

Gene name	**Number of lead variants assigned**
FTO	26
ABO	21
F5	19
LOXL1	19
CDKN2B-AS1	17

These genes are associated with significant biological functions known to be related to the phenotypes that correspond to our findings in terms of gene prioritization. FTO (Fat Mass and Obesity Associated Gene) is widely studied for its strong association with obesity and body mass index [103]. Variations in the FTO gene are some of the most well-known genetic factors contributing to obesity across populations. The gene's role is complex and appears to impact energy balance and metabolic regulation. It was shown that the SNPs assigned to the gene are related to the phenotypes from our analysis such as “type 2 diabetes mellitus” and subtypes E11, E11.7, E11.8, E11.9; “obesity” and subtypes E66, E66.0, E66.8, E66.9;

“disorders of lipoprotein metabolism and other lipidaemias” E78; “diabetic retinopathy” H36.0; “hypertension and hypertensive heart disease” I10, I11; “cholelithiasis” K80; “elevated blood glucose level” R73; “dietary counseling and surveillance” Z71.3.

The ABO (ABO Blood Group System) gene is responsible for determining the ABO blood group system (ABO locus has three primary allelic forms: A, B, and O, that define the blood type), which is crucial not only in blood transfusion compatibility but also has implications in various diseases. Research has linked variations in this gene with risks of cardiovascular diseases, infectious diseases, and even cancer [104]. In our data, the following phenotypes are associated with this gene: “iron deficiency anemia” D50; “disorders of lipoprotein metabolism and other lipidaemias” E78, E78.0, E78.2; “pulmonary embolism” I26; “venous embolism and thrombosis”, “varicose veins and other disorders of veins” I82, I83, I87; “hemorrhage from respiratory passages” R04 and “COVID-19” U07.1, U07.2.

F5 is the gene coding for Coagulation Factor V, which plays a critical role in the blood clotting process. Mutations in F5, such as the well-known Factor V Leiden mutation, increase the risk of developing abnormal blood clotting, leading to deep vein thrombosis and pulmonary embolism [105]. The assigned SNPs are related to “benign lipomatous neoplasm” D17; “iron deficiency anemia” D50; “other coagulation defects” D68, D68.2, D68.8; “pulmonary embolism” I26; “venous embolism and thrombosis”, “varicose veins and other disorders of veins” I82, I83, I87; and “hemorrhage from respiratory passages” R04.

LOXL1 (Lysyl Oxidase-Like 1) has been prominently associated with pseudoexfoliation syndrome, a systemic condition that significantly increases the risk of glaucoma. Variants in this gene are linked to changes in extracellular matrix metabolism, influencing ocular structures and potentially leading to increased intraocular pressure and optic nerve damage [106]. This gene was assigned to SNPs related to “cataracts” H25, H26; “glaucoma” H40; “mechanical complication of intraocular lens” T85.2; “ophthalmic devices associated with adverse incidents” Y77; and “presence of intraocular lens” Z96.1.

CDKN2B-AS1 (Cyclin-Dependent Kinase Inhibitor 2B Antisense RNA 1) is also known as ANRIL. This gene regulates the cell cycle, cellular aging, and apoptosis. It is particularly noted for its association with various cancers and cardiovascular diseases. Variants in CDKN2B-AS1 are associated with an increased risk of coronary artery disease, highlighting its importance in vascular cell proliferation and inflammation processes [107]. In our data, “glaucoma” H40; “angina” I20; “myocardial infarction” I21; “chronic ischaemic heart disease” I25; “heart failure” I50; and “atherosclerosis” I70 relate to this gene.

These genes represent a diverse array of physiological functions and pathophysiological implications. They are subjects of significant interest in further genetic and epidemiological research on understanding and managing human health and diseases.

Another way to look at the results of the causal gene prioritization is to check the biological sense in terms of one phenotype. For example, migraine G43 has 11 lead variants, 8 of which were matched to the genes. They include SUGCT (Succinyl-CoA: Glutarate-CoA Transferase), PHACTR1 (Phosphatase And Actin Regulator 1), APOER (apolipoprotein E receptor) – the gene that codes a protein forming a receptor found in the plasma membrane of cells involved in receptor-mediated endocytosis and CFDP1 – a neuroblastoma susceptibility gene that regulates transcription factors of the noradrenergic cell identity [108, 109, 110, 111].

All these genes code the proteins that take part in signal transduction and thus can be biologically related to the migraine.
Interestingly, this phenotype has 1 novel signal, rs6709005, assigned to TRPM8 (Transient Receptor Potential Cation Channel Subfamily M Member 8) [112]. This gene encodes a protein that forms a calcium channel; the release of calcium ions plays an important role in signal transduction between the cells which leads to the feeling of pain [113]. This finding explains the potential biological function of this novel variant and makes it an interesting target for further studies.

3.2.6 Heritability estimation

The mean h^2 over all 5,035 phenotypes was 0.058 and 6% showed a significant difference from 0 (Bonferroni-corrected P<0.05 P= 9.93×10^{-6}). 1,396 phenotypes showed a negative h^2 value (Figure 12). These unrealistic estimates are known limitations of LDSC and are caused by low-prevalence phenotypes where the low number of cases does not allow precise estimation of h^2 [114]. This pattern is also visible in the current results (Figure 13).

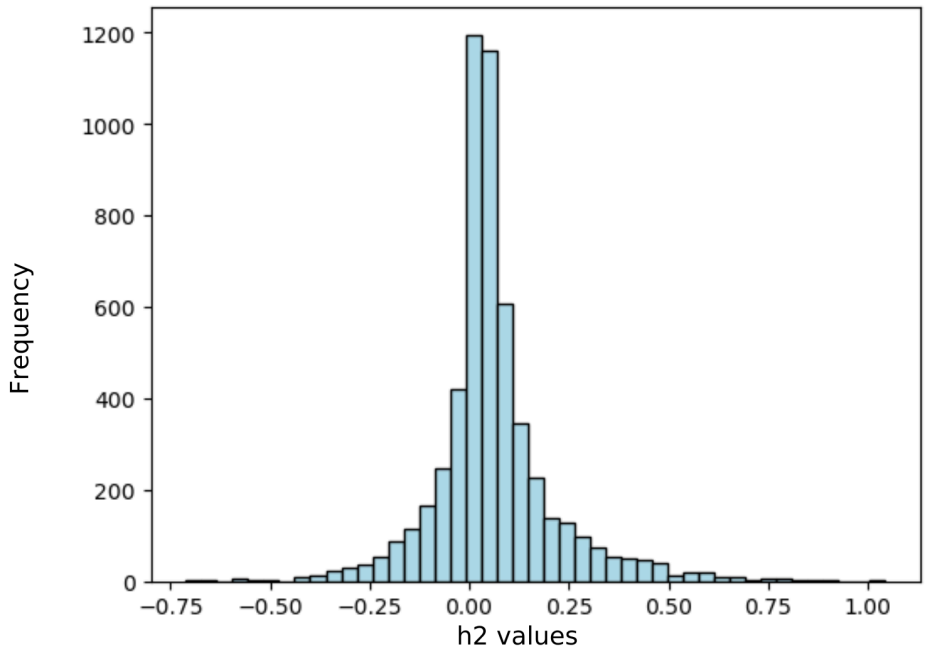

Figure 12. Distribution of h^2 values

In the overall distribution of h^2, the values range from approximately -0.75 to 1.00. The data distribution shows a robust central tendency near 0, most of the values are observed between -0.25 and 0.25. The shape of the distribution is roughly bell-shaped, suggesting a normal distribution but with slight negative skewness towards the left, as seen by the tail on the left side.

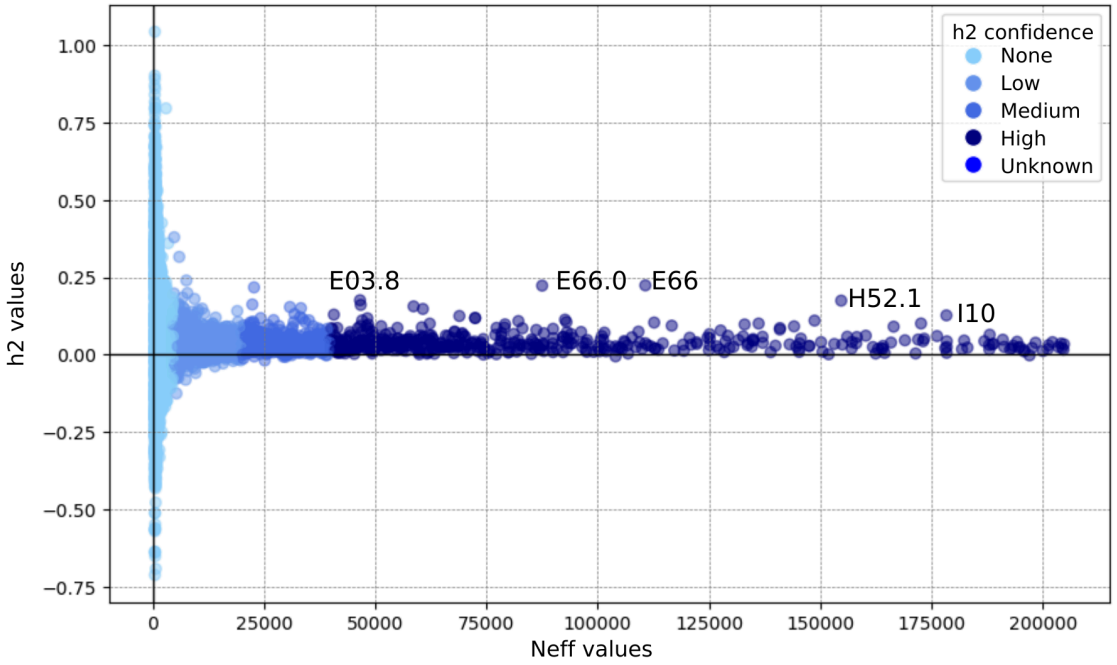

Figure 13. The relation between h^2 values and N_{eff} values. Different colors represent different levels of N_{eff} confidence, from lighter shades for none h^2 confidence to darker shades representing higher levels of h^2 confidence.

The relation between N_{eff} values and h^2 values suggests a low h^2 across varying N_{eff} values since most data points cluster around the zero mark on the Y-axis (Figure 13). The density of points near zero and their confidence levels predominantly in the lower range indicate that higher N_{eff} values do not necessarily correspond with higher h^2 values. We observe less negative h^2 estimates for more frequent phenotypes, confirming that LDSC gives unreliable signals for small sample sizes (N<5000). Also, the relationship between heritability estimates and sample sizes is similar to the one observed in UKBB, which supports the robustness of our results [118]. Some points with high confidence are outstanding in terms of h^2 values: E03.8 (Other specified hypothyroidism), E66.0 (Obesity due to excess calories), E66 (Obesity), H52.1 (Myopia), and I10 (Hypertension). These and some other examples of previously well-studied phenotypes were outlined. These results are visualized in Figure 14.

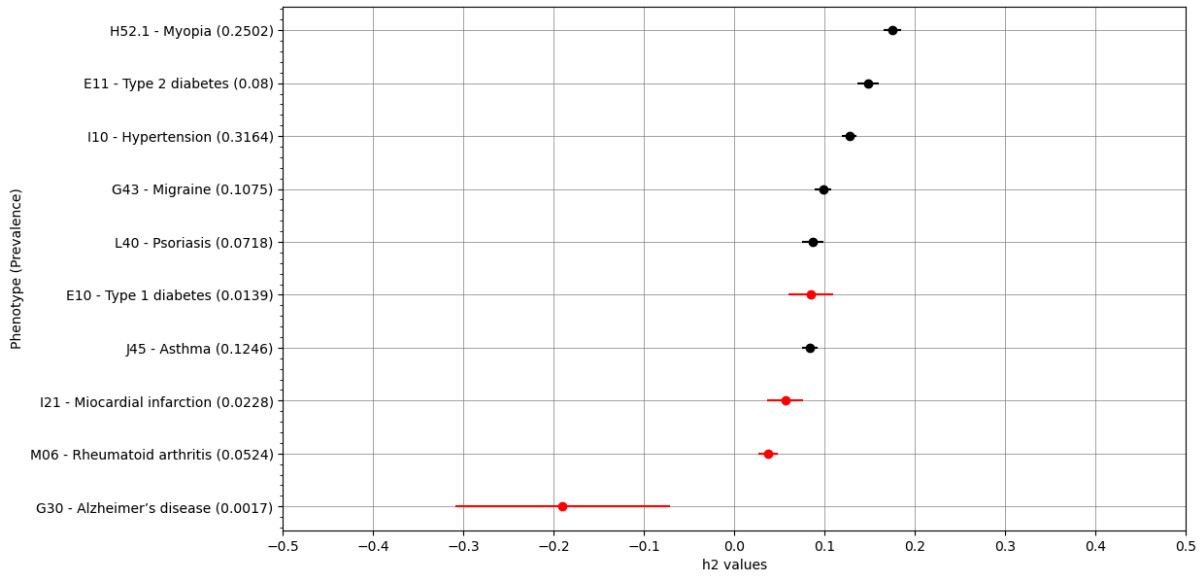

Figure 14. Estimated h^2 values for a set of well-known phenotypes. The plot illustrates the heritability estimates of various phenotypes, with horizontal error bars representing the standard errors of these estimates. The length of the error bars indicates the confidence level of the heritability estimates, with wider bars signifying greater uncertainty. The color coding in the plot differentiates the statistical significance of the heritability values: black bars indicate significant values (Bonferroni-corrected $P < 0.05$), while red bars denote insignificant values. The number in parentheses next to each phenotype name represents the prevalence of that condition within the studied population.

From the set of the phenotypes chosen for representation, myopia (H52.1) shows a relatively high h^2 0.1749, combining it with high h^2 confidence, suggesting significant genetic influence, whereas Alzheimer's disease (G30) shows an unrealistic negative h^2 value -0.1905, probably due to small effective sample size (Figure 14). These results were compared to those previously studied in Neale's lab [115]. For myopia, the h^2 value in UKBB is 0.0967, with a prevalence of 0.0813. For Alzheimer's disease, the estimated h^2 in UKBB is -0.399, with a population prevalence of 0.000329. The differences in h^2 between biobanks may be explained by the differences in the prevalences in the first place and by differences in phenotype definitions.

3.3 DISCUSSION

In this thesis, the bioinformatic interpretation of 5,035 GWASs in the Estonian Biobank was done. Among the 5,035 GWAS analyses conducted in the Estonian Biobank, the initial analysis concentrated on LD clumping and identifying lead variants. This step was needed to define a set of genetic variants that impact the analyzed traits most.

We observed a significant fraction of novel loci, 26.45%. Such a high percentage of novel loci may be explained by the phenotype definitions that differ from the ones in other countries and may provide different insights. It is also possible that some of these variants are population-specific for the Estonian population, which needs further investigation.

While the novelty was studied not only for the lead variants but also for the secondary and tertiary ones, this thesis centers on the initial steps of post-GWAS analyses, and we opted to include only lead variants to simplify the method application. The variants not considered as lead ones may be used in further research. For this and further steps, standard GWAS P-value thresholds were used for all of the phenotypes; however, because of the many phenotypes tested, a more stringent threshold might be considered.

The matching of ICD10 codes to the ontologies was a bottleneck of this study. Since it was done semi-automatically, it was impossible to map all the ICD10 codes used in such a study, and the result distribution was inflated in terms of novelty. It is important to note that other studies suggest the same approach facing the same issue, for example, Verma et al., 2023 [90]. It is also essential to consider that public databases were used for the overlap. There are no stringent rules for defining phenotypes; thus, given the vast number of published studies, the differences in wording arise as a limitation that is hard to overcome. For further research, one should consider searching for different combinations of techniques for ontology mapping.

One of the most simplistic fine-mapping methods, Approximate Bayesian Factor fine mapping, was used to facilitate high-throughput analyses of thousands of loci. This approach allowed us to capture ~10% of loci with a single causal variant; these findings are valuable for further analysis. Other more complex methods, such as SuSiE and FINEMAP, do not assume the detection of a single causal variant in each locus can be used in the future. These methods require input of LD between variants, which is more complex in high-throughput analyses. Benner et al., 2017 state that because such approaches use the LD between variants, reference genotype panels play a crucial role in the application [116]. The size of the reference panel and the simplicity of the method may lead to false positive results: the variants that may be considered lead ones may be suitable in terms of parameters and calculations, but in fact, they may not be the causal ones. Applying more complex methods in a targeted manner should be considered when analyzing this dataset further.

There is a difference in the final number of lead variants analyzed by LD clumping and fine-mapping, which can explain the difference in the methods: in LD clumping, the lead variants were identified based on the distance and LD, and in ABF fine-mapping, the loci were defined based on the distance alone.

The prioritization of the causal gene was done in one of the simplistic ways, basing the assignment on the distance to the gene bodies and the TSS. Overall, we were able to map

around 86% of the lead variants to the closest genes, which is high enough for such a direct approach; however, one must keep in mind that it can also produce some false positive results. From the small subset used as an example that was checked in more detail, all the associations made biological sense regarding genes and their known functions. Considering this, we can assume that the prioritization of the causal genes was done sufficiently, and despite the proportion of the likely incorrect assignments, does not invalidate the overall results.

Regarding SNP heritability estimation, the values reached at most 25% with high-confidence traits. The liability-scaled heritability for all the results highlighted separately does not contradict the previously described values available [115]. These insights suggest variations in prevalence in the populations, and different phenotype definitions may explain the correctness of the method we used and the differences.

In conclusion, this thesis serves as a first step in the post-GWAS analysis. It gives valuable insights, providing the possibility for further research applications using more sophisticated methods in a more targeted manner to gain more specific local insights on the outcomes of GWASs.

SUMMARY

In this thesis, a comprehensive analysis of 5,035 genome-wide association studies conducted within the Estonian Biobank was post-analyzed using an extensive set of methods. Identifying novel lead variants, fine-mapping causal variants, prioritization candidate genes, and estimating SNP heritability collectively contribute to a deeper understanding of the genetic underpinnings of complex traits and diseases.

The identification and analysis of lead variants across these studies revealed 2,745 lead variants, 726 of which are considered novel since they were not described previously in public databases: GWAS Catalogue and OpenTargets Database. Regarding novelty and matching to the ontologies, 335 signals were specified as known associations and 539 as novel associations known signals, outlining that our GWASs identified previously reported associations and not described previously associations. These findings started further investigations into the biological mechanisms underlying traits and potential statistical studies.

The Approximate Bayesian fine-mapping method facilitated the prioritization of 2,834 causal variants associated with the analyzed traits, identifying 45 variants with high posterior probabilities of causality. The sizes of credible sets vary widely, suggesting that while most credible set sizes are relatively small, there are exceptional cases with much larger sizes that can be explained by the presence of loci with extensive linkage disequilibrium structures. These findings pinpoint potential causal variants; however, since it is a simplistic method, the more complicated ones may differ in the outputs.

The gene prioritization analysis conducted in this study revealed that 2,407 lead SNPs of 2,745 were assigned to 804 genes, most comprising the protein-coding ones. The most prevalent genes among the assigned ones were checked in terms of the SNPs related to them to check the quality of the method. All the assignments made sense in terms of biological relations and functionality. This analysis sets the stage for further investigations into the molecular mechanisms of disease susceptibility and trait variability.

The estimation of SNP heritability for the analyzed traits combined with high confidence yielded a heritability estimate of 0.25, indicating that at most ~25% of the phenotypic variance in the studied traits can be explained by common genetic variants. The heritability estimation of the picked examples, such as type 2 diabetes and hypertension, correspond to the values obtained in previously done studies. These findings suggest the estimation's quality and provide insight into the genetic contribution to trait variability.

Overall, the results of this thesis offer a rich set of genetic insights derived from the systematic interpretation of a vast array of GWAS analyses. These findings expand our knowledge of the genetic landscape and pave the way for future genomics and precision medicine research.

ACKNOWLEDGEMENTS

First, I would like to thank my supervisors, Erik Abner and Urmo Võsa, for all their guidance and assistance, patience on each step of my work, and support all on my way through it.
Secondly, I express my gratitude to all the people who made this research possible: all the Biobank participants for their impact, Tartu HPC for allowing such a large-scale work, Genotyping Core Lab for preparing and imputing the genotype data, and Klavs Jermakovs for running most of the GWASs.
I also wish to thank the University of Tartu for allowing me to study here in the first place and for the neverending support of Ukrainian students and Ukrainians in general.
I want to acknowledge my parents and closest friends for always staying by my side and believing in me even when I did not. Having such a support group is a blessing, and I will forever be grateful for it.
Last but not least, I want to thank myself since this is the only person who knows the cost of this journey; nevertheless, she did it anyway.

REFERENCES

1. Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-021-00056-9
2. *Genome-wide association studies(Gwas). (n.d.). Retrieved 22 May 2024, from https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies*
3. Slatkin, M. (2008). Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. *Nature Reviews. Genetics*, *9*(6), 477–485. https://doi.org/10.1038/nrg2361
4. Holmes, M. V., Ala-Korpela, M., & Smith, G. D. (2017). Mendelian randomization in cardiometabolic disease: Challenges in evaluating causality. *Nature Reviews. Cardiology*, *14*(10), 577–590. https://doi.org/10.1038/nrcardio.2017.78
5. BEAGLE. (n.d.). Retrieved May 22, 2024, from https://bio.tools/BEAGLE
6. Loh, P.-R., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics*, *48*(7), 811–816. https://doi.org/10.1038/ng.3571
7. 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393
8. Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y
9. Mitt, M., Kals, M., Pärn, K., Gabriel, S. B., Lander, E. S., Palotie, A., Ripatti, S., Morris, A. P., Metspalu, A., Esko, T., Mägi, R., & Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics: EJHG*, *25*(7), 869–876.
10. Setu, T. J., & Basak, T. (2021). An introduction to basic statistical models in genetics. *Open Journal of Statistics*, *11*(06), 1017–1025. https://doi.org/10.4236/ojs.2021.116060
11. Pirinen, M., Donnelly, P., & Spencer, C. C. A. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics*, *44*(8), 848–851. https://doi.org/10.1038/ng.2346
12. Gaspar, H. A., & Breen, G. (2019). Probabilistic ancestry maps: A method to assess and visualize population substructures in genetics. *BMC Bioinformatics*, *20*(1), 116. https://doi.org/10.1186/s12859-019-2680-1

13. Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W.-Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, *50*(9), 1335–1341. https://doi.org/10.1038/s41588-018-0184-y
14. Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, *51*(12), 1749–1755. https://doi.org/10.1038/s41588-019-0530-8
15. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, *53*(7), 1097–1103. https://doi.org/10.1038/s41588-021-00870-7
16. Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)*, *26*(17), 2190–2191. https://doi.org/10.1093/bioinformatics/btq340
17. Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12
18. Kraft, P., Zeggini, E., & Ioannidis, J. P. A. (2009). Replication in genome-wide association studies. *Statistical Science*, *24*(4). https://doi.org/10.1214/09-STS290
19. He, Z., Chu, B., Yang, J., Gu, J., Chen, Z., Liu, L., Morrison, T., Belloy, M. E., Qi, X., Hejazi, N., Mathur, M., Le Guen, Y., Tang, H., Hastie, T., Ionita-laza, I., Sabatti, C., & Candès, E. (2024). *Beyond guilty by association at scale: Searching for causal variants on the basis of genome-wide summary statistics*. https://doi.org/10.1101/2024.02.28.582621
20. *Crispr*. (n.d.). Retrieved 22 May 2024, from https://www.genome.gov/genetics-glossary/CRISPR
21. Long, E., Yin, J., Funderburk, K. M., Xu, M., Feng, J., Kane, A., Zhang, T., Myers, T., Golden, A., Thakur, R., Kong, H., Jessop, L., Kim, E. Y., Jones, K., Chari, R., Machiela, M. J., Yu, K., Iles, M. M., Landi, M. T., … Choi, J. (2022). Massively parallel reporter assays and variant scoring identified functional variants and target genes for melanoma loci and highlighted cell-type specificity. *The American Journal of Human Genetics*, *109*(12), 2210–2229. https://doi.org/10.1016/j.ajhg.2022.11.006
22. Bulik-Sullivan, B. (2015). *Relationship between ld score and haseman-elston regression*. https://doi.org/10.1101/018283
23. Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, *91*(6), 1011–1021. https://doi.org/10.1016/j.ajhg.2012.10.010
24. Ning, Z., Pawitan, Y., & Shen, X. (2020). High-definition likelihood inference of genetic correlations across human complex traits. *Nature Genetics*, *52*(8), 859–864. https://doi.org/10.1038/s41588-020-0653-y

25. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E., & Halperin, E. (2010). Leveraging genetic variability across populations for the identification of causal variants. *American Journal of Human Genetics*, *86*(1), 23–33. https://doi.org/10.1016/j.ajhg.2009.11.016
26. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., & Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics*, *198*(2), 497–508. https://doi.org/10.1534/genetics.114.167908
27. Benner, C., Spencer, C. C. A., Havulinna, A. S., Salomaa, V., Ripatti, S., & Pirinen, M. (2016). FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics (Oxford, England)*, *32*(10), 1493–1501. https://doi.org/10.1093/bioinformatics/btw018
28. Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164. https://doi.org/10.1093/nar/gkq603
29. Dong, S., Zhao, N., Spragins, E., Kagda, M. S., Li, M., Assis, P., Jolanki, O., Luo, Y., Cherry, J. M., Boyle, A. P., & Hitz, B. C. (2023). Annotating and prioritizing human non-coding variants with RegulomeDB v.2. *Nature Genetics*, *55*(5), 724–726. https://doi.org/10.1038/s41588-023-01365-3
30. *Expression quantitative trait locus—An overview | sciencedirect topics*. (n.d.). Retrieved 22 May 2024, from https://www.sciencedirect.com/topics/neuroscience/expression-quantitative-trait-locus
31. Watanabe, K., Taskesen, E., van Bochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, *8*(1), 1826. https://doi.org/10.1038/s41467-017-01261-5
32. Pers, T. H., Karjalainen, J. M., Chan, Y., Westra, H.-J., Wood, A. R., Yang, J., Lui, J. C., Vedantam, S., Gustafsson, S., Esko, T., Frayling, T., Speliotes, E. K., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Boehnke, M., Raychaudhuri, S., Fehrmann, R. S. N., Hirschhorn, J. N., & Franke, L. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*, *6*, 5890. https://doi.org/10.1038/ncomms6890
33. Adam, Y., Samtal, C., Brandenburg, J., Falola, O., & Adebiyi, E. (2021). Performing post-genome-wide association study analysis: Overview, challenges and recommendations. *F1000Research*, *10*, 1002. https://doi.org/10.12688/f1000research.53962.1
34. Privé, F., Vilhjálmsson, B. J., Aschard, H., & Blum, M. G. B. (2019). Making the most of clumping and thresholding for polygenic scores. *The American Journal of Human Genetics*, *105*(6), 1213–1221. https://doi.org/10.1016/j.ajhg.2019.11.001
35. VanLiere, J. M., & Rosenberg, N. A. (2008). Mathematical properties of the r^2 measure of linkage disequilibrium. *Theoretical Population Biology*, *74*(1), 130–137. https://doi.org/10.1016/j.tpb.2008.05.006
36. *Hartl, d. , & clark, a. G.(2007). Principles of population genetics. Sunderland, ma sinauer associates. - References—Scientific research publishing*. (n.d.). Retrieved 22 May 2024, from https://www.scirp.org/reference/ReferencesPapers?ReferenceID=2033907

37. Shi, H., Burch, K. S., Johnson, R., Freund, M. K., Kichaev, G., Mancuso, N., Manuel, A. M., Dong, N., & Pasaniuc, B. (2020). Localizing components of shared transethnic genetic architecture of complex traits from gwas summary data. *American Journal of Human Genetics*, *106*(6), 805–817. https://doi.org/10.1016/j.ajhg.2020.04.012
38. Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8
39. Hutchinson, A., Asimit, J., & Wallace, C. (2020). Fine-mapping genetic associations. *Human Molecular Genetics*, *29*(R1), R81–R88. https://doi.org/10.1093/hmg/ddaa148
40. Wakefield, J. (2009). Bayes factors for genome‐wide association studies: Comparison with *P* ‐values. *Genetic Epidemiology*, *33*(1), 79–86. https://doi.org/10.1002/gepi.20359
41. Wang, G., Sarkar, A., Carbonetto, P., & Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *82*(5), 1273–1300. https://doi.org/10.1111/rssb.12388
42. Wellcome Trust Case Control Consortium, Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M. M., Auton, A., Myers, S., Morris, A., Pirinen, M., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Hall, A. S., Hattersley, A. T., … Donnelly, P. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, *44*(12), 1294–1301. https://doi.org/10.1038/ng.2435
43. Mountjoy, E., Schmidt, E. M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., Fumis, L., Hayhurst, J., Buniello, A., Karim, M. A., Wright, D., Hercules, A., Papa, E., Fauman, E. B., Barrett, J. C., Todd, J. A., Ochoa, D., Dunham, I., & Ghoussaini, M. (2021). An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nature Genetics*, *53*(11), 1527–1533. https://doi.org/10.1038/s41588-021-00945-5
44. Cano-Gamez, E., & Trynka, G. (2020). From gwas to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00424
45. Tambets, R., Kolde, A., Kolberg, P., Love, M. I., & Alasoo, K. (2023). *Extensive co-regulation of neighbouring genes complicates the use of eQTLs in target gene prioritisation*. bioRxiv. https://doi.org/10.1101/2023.09.29.560109
46. Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genetics*, *17*(9), e1009440. https://doi.org/10.1371/journal.pgen.1009440
47. *Introduction*. (n.d.). [Computer software]. Retrieved 22 May 2024, from https://cran.r-project.org/web/packages/coloc/vignettes/a03_enumeration.html
48. Gloudemans, M. J., Balliu, B., Nachun, D., Schnurr, T. M., Durrant, M. G., Ingelsson, E., Wabitsch, M., Quertermous, T., Montgomery, S. B., Knowles, J. W., & Carcamo-Orive, I. (2022). Integration of genetic colocalizations with physiological and pharmacological perturbations identifies cardiometabolic disease genes. *Genome Medicine*, *14*(1), 31. https://doi.org/10.1186/s13073-022-01036-8

49. Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., GTEx Consortium, Nicolae, D. L., Cox, N. J., & Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), 1091–1098. https://doi.org/10.1038/ng.3367
50. Zuber, V., Grinberg, N. F., Gill, D., Manipur, I., Slob, E. A. W., Patel, A., Wallace, C., & Burgess, S. (2022). Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *American Journal of Human Genetics*, *109*(5), 767–782. https://doi.org/10.1016/j.ajhg.2022.04.001
51. Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., Tan, V. Y., Yarmolinsky, J., Shihab, H. A., Timpson, N. J., Evans, D. M., Relton, C., Martin, R. M., Davey Smith, G., Gaunt, T. R., & Haycock, P. C. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *eLife*, *7*, e34408. https://doi.org/10.7554/eLife.34408
52. Davey Smith, G., & Hemani, G. (2014). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, *23*(R1), R89–R98. https://doi.org/10.1093/hmg/ddu328
53. Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. https://doi.org/10.1038/ng.608
54. Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, *109*(4), 1193–1198. https://doi.org/10.1073/pnas.1119675109
55. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494
56. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., & Visscher, P. M. (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nature Genetics*, *49*(9), 1304–1310. https://doi.org/10.1038/ng.3941
57. Barry, C.-J. S., Walker, V. M., Cheesman, R., Davey Smith, G., Morris, T. T., & Davies, N. M. (2022). How to estimate heritability: A guide for genetic epidemiologists. *International Journal of Epidemiology*, *52*(2), 624–632. https://doi.org/10.1093/ije/dyac224
58. Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M. J., Price, A. L., & Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*(3), 291–295. https://doi.org/10.1038/ng.3211

59. Young, A. I., Frigge, M. L., Gudbjartsson, D. F., Thorleifsson, G., Bjornsdottir, G., Sulem, P., Masson, G., Thorsteinsdottir, U., Stefansson, K., & Kong, A. (2018). Relatedness disequilibrium regression estimates heritability without environmental bias. *Nature Genetics*, *50*(9), 1304–1310. https://doi.org/10.1038/s41588-018-0178-9
60. The International HapMap Consortium. (2003). The international hapmap project. *Nature*, *426*(6968), 789–796. https://doi.org/10.1038/nature02168
61. Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., & Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the *hla* genes in the 1000 genomes project phase i data. *G3 Genes|Genomes|Genetics*, *5*(5), 931–941. https://doi.org/10.1534/g3.114.015784
62. Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., & Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, *10*(1), 3328. https://doi.org/10.1038/s41467-019-11112-0
63. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., Murakami, Y., Yuji, K., Furukawa, Y., Zembutsu, H., Tanaka, T., Ohnishi, Y., Nakamura, Y., BioBank Japan Cooperative Hospital Group, & Kubo, M. (2017). Overview of the biobank japan project: Study design and profile. *Journal of Epidemiology*, *27*(3S), S2–S8. https://doi.org/10.1016/j.je.2016.12.005
64. The All of Us Research Program Investigators. (2019). The "all of us" research program. *New England Journal of Medicine*, *381*(7), 668–676. https://doi.org/10.1056/NEJMsr1809937
65. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779
66. Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K., Reeve, M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Parolo, P. D. B., Lehisto, A., Kanai, M., Mars, N., Rämö, J., … Palotie, A. (2022). *FinnGen: Unique genetic insights from combining isolated population and national health register data*. medRxiv. https://doi.org/10.1101/2022.03.03.22271360
67. Chadwick, R. (1999). The Icelandic database—Do modern times need modern sagas? *BMJ : British Medical Journal*, *319*(7207), 441–444. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1127047/
68. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P. C., Mägi, R., Milani, L., Fischer, K., & Metspalu, A. (2015). Cohort profile: Estonian biobank of the estonian genome center, university of tartu. *International Journal of Epidemiology*, *44*(4), 1137–1147. https://doi.org/10.1093/ije/dyt268
69. Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A. U., Jiang, Y., Raghavan, S., Miao, J., Arias, J. D., Graham, S. E.,

Mukamel, R. E., Spracklen, C. N., Yin, X., Chen, S.-H., Ferreira, T., Highland, H. H., … Hirschhorn, J. N. (2022). A saturated map of common genetic variants associated with human height. *Nature*, *610*(7933), 704–712. https://doi.org/10.1038/s41586-022-05275-y

70. Hewitt, R., & Watson, P. (2013). Defining biobank. *Biopreservation and Biobanking*, *11*(5), 309–315. https://doi.org/10.1089/bio.2013.0042
71. Vaught, J. (2020). Biobanking during the covid-19 pandemic. *Biopreservation and Biobanking*, *18*(3), 153–154. https://doi.org/10.1089/bio.2020.29069.jjv
72. Hewitt, R., & Hainaut, P. (2011). Biobanking in a fast moving world: An international perspective. *Journal of the National Cancer Institute. Monographs*, *2011*(42), 50–51. https://doi.org/10.1093/jncimonographs/lgr005
73. Yuille, M., van Ommen, G.-J., Bréchot, C., Cambon-Thomsen, A., Dagher, G., Landegren, U., Litton, J.-E., Pasterk, M., Peltonen, L., Taussig, M., Wichmann, H.-E., & Zatloukal, K. (2008). Biobanking for europe. *Briefings in Bioinformatics*, *9*(1), 14–24. https://doi.org/10.1093/bib/bbm050
74. Chen, Z., Lee, L., Chen, J., Collins, R., Wu, F., Guo, Y., Linksted, P., & Peto, R. (2005). Cohort profile: The kadoorie study of chronic disease in china(Kscdc). *International Journal of Epidemiology*, *34*(6), 1243–1249. https://doi.org/10.1093/ije/dyi174
75. Silva-Zolezzi, I., Hidalgo-Miranda, A., Estrada-Gil, J., Fernandez-Lopez, J. C., Uribe-Figueroa, L., Contreras, A., Balam-Ortiz, E., del Bosque-Plata, L., Velazquez-Fernandez, D., Lara, C., Goya, R., Hernandez-Lemus, E., Davila, C., Barrientos, E., March, S., & Jimenez-Sanchez, G. (2009). Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(21), 8611–8616. https://doi.org/10.1073/pnas.0903045106
76. Ebenezer, T. E., Muigai, A. W. T., Nouala, S., Badaoui, B., Blaxter, M., Buddie, A. G., Jarvis, E. D., Korlach, J., Kuja, J. O., Lewin, H. A., Majewska, R., Mapholi, N., Maslamoney, S., Mbo'o-Tchouawou, M., Osuji, J. O., Seehausen, O., Shorinola, O., Tiambo, C. K., Mulder, N., … Djikeng, A. (2022). Africa: Sequence 100,000 species to safeguard biodiversity. *Nature*, *603*(7901), 388–392. https://doi.org/10.1038/d41586-022-00712-4
77. Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell*, *177*(1), 26–31. https://doi.org/10.1016/j.cell.2019.02.048
78. *Estonian biobank*. (2021, December 13). https://genomics.ut.ee/en/content/estonian-biobank
79. CDC. (2019). *ICD - ICD-10 - International Classification of Diseases, Tenth Revision*. CDC. https://www.cdc.gov/nchs/icd/icd10.htm
80. *ICD-10 Version:2019*. (n.d.). Retrieved 22 May 2024, from https://icd.who.int/browse10/2019/en
81. *Understanding the icd-10 code structure*. (n.d.). Retrieved 22 May 2024, from https://www.healthnetworksolutions.net/index.php/understanding-the-icd-10-code-structure

82. Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *American Journal of Human Genetics*, *108*(10), 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005
83. Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559
84. Suhas, S., Manjunatha, N., Kumar, C. N., Benegal, V., Rao, G. N., Varghese, M., & Gururaj, G. (2023). Firth's penalized logistic regression: A superior approach for analysis of data from India's National Mental Health Survey, 2016. *Indian Journal of Psychiatry*, *65*(12), 1208–1213. https://doi.org/10.4103/indianjpsychiatry.indianjpsychiatry_827_23
85. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, *53*(7), 1097–1103. https://doi.org/10.1038/s41588-021-00870-7
86. *Gwas catalog*. (n.d.). Retrieved 22 May 2024, from https://www.ebi.ac.uk/gwas/home
87. *Open targets genetics*. (n.d.). Retrieved 22 May 2024, from https://genetics.opentargets.org/
88. *EBISPOT/EFO-UKB-mappings*. (2024, March 21). GitHub. https://github.com/EBISPOT/EFO-UKB-mappings
89. *Ontobee*. (n.d.). Ontobee.org. Retrieved May 22, 2024, from https://ontobee.org/
90. Verma, A., Huffman, J. E., Rodriguez, A., Conery, M., Liu, M., Ho, Y.-L., Kim, Y., Heise, D. A., Guare, L., Panickan, V. A., Garcon, H., Linares, F., Costa, L., Goethert, I., Tipton, R., Honerlaw, J., Davies, L., Whitbourne, S., Cohen, J., … Liao, K. P. (2023). Diversity and scale: Genetic architecture of 2,068 traits in the va million veteran program. *medRxiv*, 2023.06.28.23291975. https://doi.org/10.1101/2023.06.28.23291975
91. *Homo_sapiens—Ensembl genome browser 112*. (n.d.). Retrieved 22 May 2024, from https://useast.ensembl.org/Homo_sapiens/Info/Index
92. *Alkes group*. (n.d.). Retrieved 22 May 2024, from https://alkesgroup.broadinstitute.org/
93. Bulik-Sullivan, B. (2015). *Relationship between ld score and haseman-elston regression*. https://doi.org/10.1101/018283
94. *Heritability of >4,000 traits & disorders in UK Biobank*. (n.d.). Retrieved 22 May 2024, from https://nealelab.github.io/UKBB_ldsc/
95. Hautakangas, H., Winsvold, B. S., Ruotsalainen, S. E., Bjornsdottir, G., Harder, A. V. E., Kogelman, L. J. A., Thomas, L. F., Noordam, R., Benner, C., Gormley, P., Artto, V., Banasik, K., Bjornsdottir, A., Boomsma, D. I., Brumpton, B. M., Burgdorf, K. S., Buring, J. E., Chalmer, M. A., de Boer, I., … Pirinen, M. (2022). Genome-wide analysis of 102,084 migraine cases identifies 123 risk loci and subtype-specific risk alleles. *Nature Genetics*, *54*(2), 152–160. https://doi.org/10.1038/s41588-021-00990-0
96. Meng, W., Adams, M. J., Hebert, H. L., Deary, I. J., McIntosh, A. M., & Smith, B. H.

(2018). A genome-wide association study finds genetic associations with broadly-defined headache in uk biobank(N=223,773). *EBioMedicine*, *28*, 180–186. https://doi.org/10.1016/j.ebiom.2018.01.023

97. Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., Ishigaki, K., Suzuki, A., Suzuki, K., Obara, W., Yamaji, K., Takahashi, K., Asai, S., Takahashi, Y., Suzuki, T., … Okada, Y. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics*, *53*(10), 1415–1424. https://doi.org/10.1038/s41588-021-00931-x
98. Kim, W., Prokopenko, D., Sakornsakolpat, P., Hobbs, B. D., Lutz, S. M., Hokanson, J. E., Wain, L. V., Melbourne, C. A., Shrine, N., Tobin, M. D., Silverman, E. K., Cho, M. H., & Beaty, T. H. (2021). Genome-wide gene-by-smoking interaction study of chronic obstructive pulmonary disease. *American Journal of Epidemiology*, *190*(5), 875–885. https://doi.org/10.1093/aje/kwaa227
99. Dönertaş, H. M., Fabian, D. K., Valenzuela, M. F., Partridge, L., & Thornton, J. M. (2021). Common genetic associations between age-related diseases. *Nature Aging*, *1*(4), 400–412. https://doi.org/10.1038/s43587-021-00051-5
100. *Data-Field 6159*. (n.d.). Retrieved 22 May 2024, from https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=6159
101. Wu, Y., Byrne, E. M., Zheng, Z., Kemper, K. E., Yengo, L., Mallett, A. J., Yang, J., Visscher, P. M., & Wray, N. R. (2019). Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nature Communications*, *10*(1), 1891. https://doi.org/10.1038/s41467-019-09572-5
102. Hager, J., Kamatani, Y., Cazier, J.-B., Youhanna, S., Ghassibe-Sabbagh, M., Platt, D. E., Abchee, A. B., Romanos, J., Khazen, G., Othman, R., Badro, D. A., Haber, M., Salloum, A. K., Douaihy, B., Shasha, N., Kabbani, S., Sbeite, H., Chammas, E., el Bayeh, H., … FGENTCARD Consortium. (2012). Genome-wide association study in a Lebanese cohort confirms PHACTR1 as a major determinant of coronary artery stenosis. *PloS One*, *7*(6), e38663. https://doi.org/10.1371/journal.pone.0038663
103. *FTO FTO alpha-ketoglutarate dependent dioxygenase [Homo sapiens (Human)]—Gene—NCBI*. (n.d.). Retrieved 22 May 2024, from https://www.ncbi.nlm.nih.gov/gene/79068
104. *ABO ABO, alpha 1-3-N-acetylgalactosaminyltransferase and alpha 1-3-galactosyltransferase [Homo sapiens (Human)]—Gene—NCBI*. (n.d.). Retrieved 22 May 2024, from https://www.ncbi.nlm.nih.gov/gene/28
105. *F5 coagulation factor v [homo sapiens (Human)]—Gene—Ncbi*. (n.d.). Retrieved 22 May 2024, from https://www.ncbi.nlm.nih.gov/gene/2153
106. *LOXL1 lysyl oxidase like 1 [Homo sapiens (Human)]—Gene—NCBI*. (n.d.). Retrieved 22 May 2024, from https://www.ncbi.nlm.nih.gov/gene/4016
107. *Cdkn2b-as1 cdkn2b antisense rna 1 [homo sapiens (Human)]—Gene—Ncbi*. (n.d.). Retrieved 22 May 2024, from https://www.ncbi.nlm.nih.gov/gene/100048912
108. *Sugct succinyl-coa:glutarate-coa transferase [homo sapiens (Human)]—Gene—Ncbi*. (n.d.). Retrieved 22 May 2024, from https://www.ncbi.nlm.nih.gov/gene/79783

109. *PHACTR1 phosphatase and actin regulator 1 [Homo sapiens (Human)]—Gene—NCBI*. (n.d.). Retrieved 22 May 2024, from https://www.ncbi.nlm.nih.gov/gene/221692
110. *LRP1 LDL receptor related protein 1 [Homo sapiens (Human)]—Gene—NCBI*. (n.d.). Retrieved 22 May 2024, from https://www.ncbi.nlm.nih.gov/gene/4035
111. *CFDP1 craniofacial development protein 1 [Homo sapiens (Human)]—Gene—NCBI*. (n.d.). Retrieved 22 May 2024, from https://www.ncbi.nlm.nih.gov/gene/10428
112. *TRPM8 transient receptor potential cation channel subfamily M member 8 [Homo sapiens (Human)]—Gene—NCBI*. (n.d.). Retrieved 22 May 2024, from https://www.ncbi.nlm.nih.gov/gene/79054
113. Kowalska, M., Prendecki, M., Piekut, T., Kozubski, W., & Dorszewska, J. (2021). Migraine: Calcium channels and glia. *International Journal of Molecular Sciences*, *22*(5), 2688. https://doi.org/10.3390/ijms22052688
114. Ojavee, S. E., Kutalik, Z., & Robinson, M. R. (2022). Liability-scale heritability estimation for biobank studies of low-prevalence disease. *American Journal of Human Genetics*, *109*(11), 2009–2017. https://doi.org/10.1016/j.ajhg.2022.09.011
115. *Ukb snp-heritability browser*. (n.d.). Retrieved 22 May 2024, from https://nealelab.github.io/UKBB_ldsc/h2_browser.html
116. Benner, C., Havulinna, A. S., Järvelin, M.-R., Salomaa, V., Ripatti, S., & Pirinen, M. (2017). Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *American Journal of Human Genetics*, *101*(4), 539–551. https://doi.org/10.1016/j.ajhg.2017.08.012

Non-exclusive licence to reproduce thesis and make thesis public

I, Anastasiia Alekseienko,

1. I grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis
   Systematic interpretation of large-scale GWAS analyses of 5,035 phenotypes, supervised by Erik Abner and Urmo Võsa.
2. I grant the University of Tartu the permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work from 22/05/2024 until the expiry of the term of copyright,
3. I am aware that the author retains the rights specified in points 1 and 2.
4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Anastasiia Alekseienko
22/05/2024