UNIVERSITY OF HELSINKI

Functional insights into FinnGen disease associations via Finnish transcriptome sequencing

Genetics and Molecular Biosciences,
Molecular and Analytical Health Biosciences
Master's thesis

Author:
Elna-Maria Satokangas

Supervisor(s):
Taru Tukiainen, Docent

19.5.2025
Helsinki

Faculty: Biological and Environmental Sciences
Degree programme: Genetics and Molecular Biosciences
Study track: Molecular and Analytical Health Biosciences
Author: Elna-Maria Satokangas
Title: Functional insights into FinnGen disease associations via Finnish transcriptome sequencing
Level: Master's thesis
Month and year: May 2025
Number of pages: 43
Keywords: eQTL, GWAS, FinnGen, GTEx, colocalization analysis
Supervisor or supervisors: Taru Tukiainen, Docent
Where deposited: HELDA – Digital Repository of the University of Helsinki
Additional information:
Abstract:
Genome-wide association studies (GWAS) identify genetic variants associated with traits, most located in non-coding regions, suggesting they regulate gene expression rather than altering protein composition. Expression quantitative trait loci (eQTL) mapping complements this by identifying genetic variants that influence gene expression. Colocalization studies integrate these approaches to determine whether the same genetic variant influences a disease trait and gene expression level, providing insight into disease mechanisms. This thesis investigates how genetic factors regulate gene expression in the liver tissue using eQTL data from 131 Finnish individuals (FinnLiver) and analyses colocalization results from FinnLiver transcriptomics with disease and trait associations from FinnGen to identify GWAS loci that share causal variants with eQTL data. Finally, the study compares these findings with liver-derived colocalization results from U.S. individuals in the GTEx study to assess whether differences in colocalization signals are due to differences in data origin.

eQTL analysis identified 1,836 unique genes with significant eQTLs, showing a relationship with higher gene expression and greater tolerance to loss-of-function mutations compared to genes without significant eQTLs. Colocalization analysis revealed 180 significant colocalizations between FinnLiver eQTLs and FinnGen GWAS data, concentrated in circulatory system and metabolic disease endpoints. Despite a smaller sample size, FinnLiver had almost twice as many significant colocalizations with FinnGen when compared to GTEx. Furthermore, when examining liver-related disease endpoints, only seven colocalizations were found in both FinnLiver and GTEx datasets, while 99 were unique to FinnLiver and 29 to GTEx. However, there was no clear evidence that the unique colocalizations were driven by population-specific data. Technical differences and the population heterogeneity within GTEx were likely more relevant factors explaining the observed differences.

Tiedekunta: Bio- ja ympäristötieteellinen tiedekunta
Koulutusohjelma: Genetiikan ja molekulaaristen biotieteiden maisteriohjelma
Opintosuunta: Molekulaariset ja analyyttiset terveyden biotieteet
Tekijä: Elna-Maria Satokangas
Työn nimi: Toiminnallisia näkökulmia FinnGenin tautiyhteyksiin suomalaisen transkriptomisekvensoinnin avulla
Työn laji: Maisterin tutkielma
Kuukausi ja vuosi: Toukokuu 2025
Sivumäärä: 43
Avainsanat: eQTL, GWAS, FinnGen, GTEx, kolokalisaatioanalyysi
Ohjaaja(t): Taru Tukiainen, Dosentti
Säilytyspaikka: HELDA – Helsingin yliopiston digitaalinen arkisto
Lisätiedot:
Abstrakti:
Genominlaajuiset assosiaatiotutkimukset (GWAS) auttavat tunnistamaan geneettisiä variantteja, jotka liittyvät erilaisiin sairauksiin ja ominaisuuksiin. Suurin osa näistä varianteista sijaitsee geenien ulkopuolisilla, ei-koodaavilla alueilla, mikä viittaa siihen, että proteiinirakenteen sijaan ne säätelevät geenien ilmentymistä. Geenien ekspression kvantitatiivisten lokusten (eQTL) kartoitus täydentää GWAS-lähestymistapaa tunnistamalla variantteja, jotka säätelevät geenien ilmentymistä. Kolokalisaatioanalyysi yhdistää nämä kaksi lähestymistapaa ja selvittää, onko jokin yksittäinen geneettinen variantti yhteydessä sekä geenin ilmentymiseen että tiettyyn fenotyyppiin tarjoten syvempää ymmärrystä tautimekanismeista. Tässä tutkielmassa tutkitaan geenien ilmentymisen säätelyä maksakudoksessa analysoimalla 131 suomalaiselta yksilöltä kerättyä eQTL-aineistoa (FinnLiver), sekä analysoimalla transkriptomidatan ja FinnGenin tautiyhteyksien kolokalisaatioita, tavoitteena tunnistaa yhteisiä kausaalisia variantteja aineistojen välillä. Lisäksi näitä löydöksiä verrataan yhdysvaltalaisilta yksilöiltä kerättyyn GTEx-projektin maksan transkriptomidataan perustuviin kolokalisaatiotuloksiin, jotta voidaan arvioida, kuinka paljon kolokalisaatiosignaalien erot johtuvat käytetyn aineiston alkuperästä.

eQTL-analyysissä tunnistettiin yhteensä 1 836 geeniä, joilla oli tilastollisesti merkitseviä eQTL-assosiaatioita. Näillä geeneillä havaittiin olevan korkeampi geeniekpressio sekä suurempi sietokyky loss-of-function-mutaatiota kohtaan verrattuna geeneihin, joilla ei havaittu merkitseviä eQTL:itä. Kolokalisaatioanalyysi paljasti 180 merkitsevää kolokalisaatiota FinnLiverin eQTL-aineiston ja FinnGenin GWAS-signaalien välillä, pääosin verenkiertoelimistön ja aineenvaihdunnan sairauksissa. Pienemmästä näytemäärästä huolimatta, FinnLiver-aineistolla havaittiin lähes kaksinkertainen määrä merkitseviä kolokalisaatioita verrattuna GTEx-aineistoon. Tarkasteltaessa maksan kannalta oleellisia tautiryhmiä, ainoastaan 7 kolokalisaatiota oli yhteisiä FinnLiverin ja GTExin välillä, kun taas 99 oli uniikkeja FinnLiverille ja 29 GTExille. Tulosten perusteella ei kuitenkaan voitu osoittaa, että erot johtuivat populaatiokohtaisista tekijöistä, vaan ne selittynevät todennäköisemmin teknisillä eroilla tai populaatiorakenteella.

Table of contents

1 Abbreviations

CLPA – Causal Posterior Agreement

eQTL – Expression Quantitative Trait Loci

GTEx – Genotype-Tissue Expression Project

GWAS – Genome-Wide Association Studies

LoF – Loss-of-Function

LOEUF – Loss-of-Function Observed/Expected Upper Fraction

PP.H4 – Posterior Probability for Hypothesis 4

sQTL – Splicing Quantitative Trait Loci

SNP – Single Nucleotide Polymorphisms

SuSiE – Sum of Single Effects

TSS – Transcription Start Site

2 Introduction

2.1 Genetic basis of complex traits

Human diversity arises from genetic variation, with traits shaped by alleles distributed across the 23 pairs of chromosomes. These alleles are inherited from different parents, but the process of meiosis introduces additional combinations of variants. During meiosis, independent assortment and crossover events create unique combinations of alleles in gametes, ensuring that even siblings, apart from identical twins, have a unique genome.

Recent genomic studies have provided more profound insights into human genetic variation. On average, individuals differ from the reference genome at approximately 4.1 to 5 million genomic positions, with over 99.9% of these differences being single nucleotide polymorphisms (SNPs) and small insertions and deletions (1000 Genomes Project Consortium, 2015). Although structural variants are less frequent, they impact a larger portion of the genome—approximately 20 million bases. The distribution of genetic variants varies significantly across populations, with African ancestry populations exhibiting the highest genetic diversity. While the vast majority of genetic variants in the human genome are rare, most variants found within an individual genome are common. In fact, only 1–4% of variants in a typical genome have a frequency below 0.5%.

Unlike Mendelian traits, which are typically governed by a single gene, complex traits and diseases are influenced by multiple genetic loci, each contributing a small effect (Gallagher & Chen-Plotkin 2018). Furthermore, many genes involved in polygenic traits exhibit pleiotropy, meaning they influence multiple, typically related traits, although effects on seemingly unrelated traits have also been observed (Watanabe et al., 2019). On top of that, the environment has its own contribution to traits (Argentieri et al., 2025). Lifestyle choices, exposure to toxins, and socioeconomic factors lead to variations among individuals with similar genetic backgrounds and can affect the severity, onset, or progression of traits over time. This intricate interplay between genetics and environment presents a major challenge in understanding the genetic basis of complex traits.

Genes critical for maintaining proper cellular and organismal function rarely tolerate any change (Karczewski et al., 2020). This intolerance, particularly to loss-of-function (LoF) mutations, is referred to as gene constraint. The Loss-of-Function Observed/Expected Upper Fraction (LOEUF) score is commonly used to quantify gene constraint. A lower LOEUF score indicates higher constraint, meaning that LoF mutations are rarely observed in the population due to strong selective pressure against them. Conversely, genes with higher LOEUF scores are less constrained and more tolerant of variation.

To understand how genetic variants influence traits, researchers often examine the minor allele frequency (MAF), which measures how common the less frequent allele of a genetic variant is within a population (Jackson et al., 2018). MAF provides valuable insights into the distribution of genetic variation and helps reveal the evolutionary dynamics that shape the prevalence of different alleles. Rare variants with low MAFs often have larger effects on traits but are harder to detect, particularly in smaller studies (Huang et al., 2024). Mutations that cause high disease risk tend to reduce fitness, making them less likely to be transmitted widely within a population (Umans et al., 2021). As a result, common variants typically have smaller individual effects, but collectively, they play a significant role in shaping polygenic traits, such as height or disease risk.

2.2 Genetic regulation of gene expression

Gene expression refers to the process of transcribing genetic information from DNA into RNA (Pai et al., 2015). This RNA may then be translated into proteins or, in some cases, function directly as an RNA molecule. Variations in genetic sequences can affect this regulation, leading to differences in gene expression levels among individuals. Understanding how genetic variation influences gene expression is crucial for uncovering the molecular mechanisms behind phenotypic diversity and complex traits and identifying potential genetic causes of diseases and therapeutic targets.

2.2.1 Gene expression and transcriptome sequencing

The human genome contains approximately 20,000 protein-coding genes and 22,000 non-coding RNA genes, and it is estimated that at least 80% of the genome

has some functional or structural role (Jackson et al., 2018). Transcription is a tightly regulated process that ensures genes are expressed at the right time, in the right cell type, and at appropriate levels to maintain cellular and organismal function (Pai et al., 2015). The measurement of gene expression allows researchers to quantify the activity of genes under various conditions, providing a snapshot of cellular function.

One of the most advanced techniques for measuring gene expression is transcriptome sequencing, commonly referred to as RNA sequencing (Stark et al., 2019). It is a high-throughput method that uses next-generation sequencing technologies to capture and quantify the complete set of RNA transcripts present in a sample. By mapping sequencing reads back to a reference genome or transcriptome, researchers can obtain precise information about the abundance and structure of transcripts, including alternative splicing events, non-coding RNAs, and novel transcripts.

2.2.2 Genetic regulation of gene expression

Gene expression regulation is a complex and dynamic process that controls cellular function across different tissues, developmental stages, and environmental conditions (The GTEx Consortium, 2020). This regulation involves various interconnected mechanisms, each playing a role in precisely adjusting gene activity.

At the transcriptional level, promoters and enhancers are crucial regulatory elements, coordinating input from transcription factors to regulate gene expression (Schoenfelder & Fraser, 2019). Post-transcriptional mechanisms introduce additional complexity to gene regulation (Baralle & Giudice, 2017). Controlled by RNA-binding proteins, alternative splicing significantly increases proteome diversity from a limited gene set. Epigenetic modifications, including DNA methylation, histone modifications, chromatin remodeling, and non-coding RNA, offer a flexible and potentially inheritable way to regulate gene expression (Wu et al., 2023). This adaptability enables cells to respond to environmental signals and plays a role in age-related shifts in gene expression without making changes in DNA sequence.

2.2.3 Studying genetic regulation of gene expression traits

Variants that affect gene expression are called expression quantitative trait loci (eQTLs) (Nica & Dermitzakis, 2013). Studying eQTLs provides insights into sequence variation and offers evidence of how genes are regulated, which can be further

utilized to understand the impact of gene regulation on disease traits. A gene with at least one significant eQTL associated with its expression is called an eGene (The GTEx Consortium, 2020).

Regulatory variants affect gene expression in multiple ways, such as modulating chromatin accessibility, enhancer activity, or transcription factor binding (Albert & Kruglyak, 2015). Commonly, eQTLs that are located within 1Mb from the transcription start site (TSS) of the gene they regulate are called cis-eQTLs (figure 1) (Nica & Dermitzakis, 2013). On the other hand, the variants located in the different chromosomes, or at least 5Mb from the TSS, are called trans-eQTLs. In this study, we focus only on cis-eQTLs.

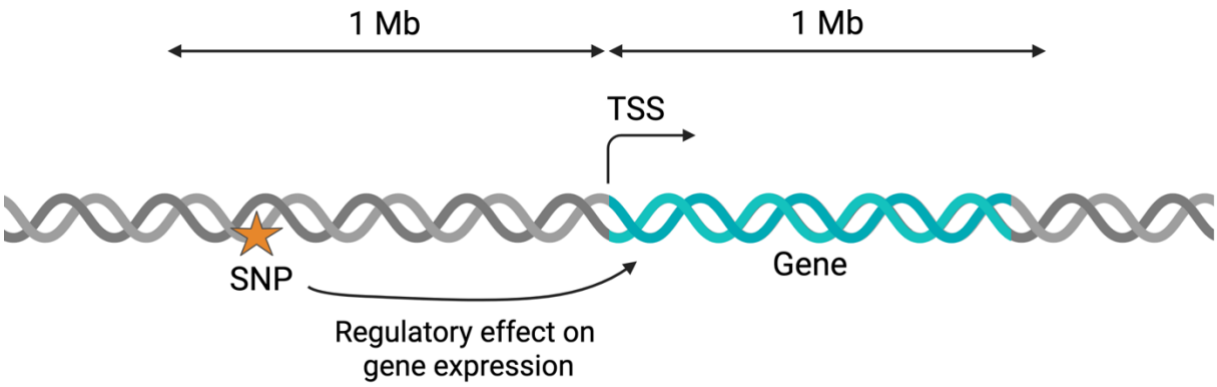

Figure 1: Illustration of a cis-eQTL effect. A genetic variant (single nucleotide polymorphism, SNP) positioned within one megabase of the gene's transcription start site (TSS) modulates the gene's expression. The arrow represents the regulatory effect of the variant on gene expression. Created with BioRender.com.

When studying eQTLs, researchers genotype individuals and measure their gene expression using microarrays or, more commonly today, RNA sequencing. To achieve good statistical power, at least a few hundred individual samples are needed (Aguet et al., 2023). The most extensive studies have included tens of thousands of samples, most commonly tissue biopsies collected from living or recently deceased individuals. In these studies, association analyses are typically performed for all genes that are detectably expressed in the sampled tissue.

eQTLs are discovered by analyzing the relationship between genotypes and gene expression levels through association or linkage analysis (Aguet et al., 2023). In a statistical test, individuals are grouped by their allele and compared if the gene expression of a certain gene differs significantly from the other group (Albert &

Kruglyak, 2015). If so, the variant or variant in linkage disequilibrium affects the gene's expression.

2.2.4 Fine-mapping and the role of credible sets

A newly developed method for identifying most causal variants affecting gene expression is genetic fine-mapping (Wang et al., 2020). Its goal is to identify the most likely causal variants associated with the trait of interest. A key challenge in this process is distinguishing true causal variants from those only correlated due to linkage disequilibrium. The 'Sum of Single Effects' (SuSiE) fine-mapping method (Wang et al., 2020) addresses this challenge by using credible sets—a collection of variants designed to capture the most likely causal variant. SuSiE identifies all credible sets for a gene supported by the data while minimizing the size of each set to ensure precision.

Another method for identifying significant eQTLs is the permutation approach (Delaneau et al., 2017). This method involves randomly shuffling gene expression data while maintaining the structure of genetic variants, creating a null distribution of associations. By comparing observed associations to this null distribution, the method calculates adjusted P-values, effectively controlling for false positives. The approach is computationally efficient, requiring fewer permutations than traditional methods, which allows for rapid processing of large-scale genomic datasets. This permutation technique enables researchers to identify statistically significant relationships between genetic variants and gene expression levels, providing a robust foundation for understanding the genetic basis of gene regulation.

2.3 Genome-wide association studies (GWAS)

Before the development of Genome-Wide Association Studies (GWAS), genetic research focused on specific candidate genes that were thought to influence diseases or traits (Jackson et al., 2018). This approach was limited because it only targeted genes already known to be involved. GWAS changed this with high-throughput genotyping, allowing researchers to scan the entire genome, looking at genetic variations across the entire DNA sequence of individuals (Uffelman et al., 2021). By analyzing large sample sizes, GWAS can identify common genetic variants linked to diseases or traits, even those with small effects. This approach helps to understand

the complex genetic architecture behind traits influenced by many genes rather than just a few.

2.3.1 Studying complex disease using association

Although the first GWAS was conducted in 2005 (Klein et al., 2005), it is commonly thought that the first modern large-scale GWAS was the Wellcome Trust Case-Control Consortium in 2007. By studying seven diseases with different aetiologies simultaneously, it aimed to gain insights into the specific genetic contributions to each disease and the variations in allelic architecture across them. It also addressed key methodological challenges in GWAS, such as study design, quality control, and data analysis. This template for modern-day study was enabled by three factors: Firstly, advances in genotyping technologies allowed more than 500,000 SNPs to be genotyped across genomes in a single run. Secondly, data provided by the HapMap project (The International HapMap Consortium, 2005) helped researchers select relevant SNPs to focus on. Lastly, researchers combined their efforts by contributing case data to consortia, enabling them to reach the large sample sizes necessary for statistical power, with funding from various sources supporting these large-scale studies. This collaborative approach continues today in initiatives such as the FinnGen study (Kurki et al., 2023), a large-scale Finnish genomics project that has analyzed genetic data from over 500,000 biobank participants and linked it to nationwide health records through a collaboration between Finnish research institutions, biobanks, and international industry partners.

In addition to these foundational factors, the success of a GWAS for a specific trait or disease depends on other elements, such as the frequency and effect size of genetic variants and the complexity of diagnosing or measuring the trait (Visscher et al., 2017).

The experimental workflow of a GWAS involves several key steps (Uffelman et al., 2021). First, DNA and phenotypic information (such as disease status, age, and sex) are collected from individuals. The next step is genotyping, where each individual's DNA is analyzed using GWAS arrays or sequencing strategies to identify genetic variations, most commonly SNPs, but also copy-number and sequence variants. In recent years, large, open-access population biobanks have provided data from thousands of genotyped individuals who have been extensively phenotyped through

questionnaires, lab measurements, or electronic health records. While most biobanks have relied on imputed genotype data for common variants, whole exome, and whole genome sequencing are increasingly used as their costs decrease. Imputation fills gaps where genetic variants have not been directly genotyped by predicting unobserved variants based on haplotype structures inferred from observed SNPs and reference panels with fully sequenced genomes (Visscher et al., 2017). This approach enhances the study's statistical power by increasing the number of variants available for analysis.

After genotyping, quality control measures ensure data accuracy by removing errors or inconsistencies (Uffelman et al., 2021). Statistical tests, such as the mixed-model logistic regression method SAIGE (Zhou et al., 2018), are performed to identify associations between specific genetic variants and the traits or diseases under investigation. The results are typically reported as genomic risk loci — regions containing groups of SNPs in linkage disequilibrium that are jointly associated with a particular trait.

2.3.2 Functional interpretation of GWAS

To improve the reliability and statistical power of GWAS findings, researchers often conduct meta-analyses to combine data from multiple studies, which strengthens the robustness of the results and aids in their functional interpretation (Uffelman et al., 2021). These results can lead to significant applications. Successful drug targets have been discovered directly from GWAS, such as variants in the *SLC30A8* gene associated with type 2 diabetes, which guided the development of treatments (Visscher et al., 2017). Another important application is the development of polygenic risk scores, which aggregate the effects of multiple variants to predict individual susceptibility to complex traits (Khera et al., 2018). Polygenic risk scores have shown promise for early identification of high-risk individuals for targeted prevention or intervention strategies. According to studies, they can even achieve cumulative effect sizes similar to those observed in monogenic diseases, underscoring their potential clinical utility (Sirugo et al., 2020). However, the clinical application of polygenic risk scores raises ethical considerations, such as potential differences in predictive accuracy across populations due to varying genetic ancestry.

Despite these direct applications, interpreting the biological significance of GWAS findings often requires additional functional analyses. Unlike studies of Mendelian traits, where the causal variant, target gene, and protein-altering mechanism are often identified simultaneously, GWAS typically pinpoints a trait and a broader genomic region containing potential causal variants (Visscher et al., 2017). If these variants lie in non-coding regions of the genome, GWAS may lack detailed information about the mechanisms by which the variant influences the trait. Integrating GWAS data with other genomic resources, such as eQTLs and epigenomic annotations, helps connect statistical associations to biological mechanisms, providing insights into their functional roles in disease pathways.

2.4 Integrating transcriptome data with GWAS

2.4.1 Colocalization studies

Most of the significant GWAS risk loci have been discovered in non-coding regions of the genome and do not directly explain how the variant is associated with the trait (Aguet et al., 2023). These regions most likely have regulatory functions on the gene and affect the gene expression rather than protein composition. In colocalization studies, GWAS and eQTL mapping are combined to gain a deeper understanding of a gene variant's functionality (Hormozdiari et al. 2016). A gene's expression is likely to play a role in the disease mechanism if the same variant associated with the GWAS locus also influences the expression of the gene. This requires that the causal variants have been identified correctly in both studies.

One of the key methods used in colocalization analysis is COLOC, a Bayesian approach that evaluates the likelihood of a shared causal variant between two traits (Wallace, 2020). COLOC assesses five hypotheses: no association with either trait (H0), association with one trait but not the other (H1 and H2), association with both traits due to independent variants (H3), and association with both traits due to a shared variant (H4). The method calculates posterior probabilities for each hypothesis, with evidence for H4 indicating colocalization. COLOC is widely used due to its ability to integrate GWAS summary statistics, making it accessible for large-scale analyses. Initially, COLOC assumed a single causal variant per trait in the analyzed region, which limited its accuracy in complex genetic architectures. However, recent advancements have addressed this limitation by integrating COLOC

with SuSiE, a fine-mapping framework that allows for the detection of multiple causal variants. Using methods like COLOC, researchers can gain deeper insights into how genetic variants affect gene expression and contribute to disease risk (Hormozdiari et al. 2016). This has important implications for identifying therapeutic targets and clarifying the biological pathways underlying complex diseases.

2.4.2 Challenges and opportunities in colocalization analysis

Almost 90% of GWAS hits are located in non-coding regions of the genome (Watanabe et al., 2019). However, currently known eQTLs fail to explain the majority of these associations (Mostafavi et al., 2023). This discrepancy could arise from several factors. First, GWAS variants may act as context-dependent eQTLs, functioning only in specific tissues (The GTEx Consortium, 2020), developmental stages (Nguyen et al., 2023), or under certain stimuli (Lea et al., 2022). Second, ancestry mismatches between GWAS and eQTL cohorts can significantly reduce colocalization power, as population differences in allele frequencies and linkage disequilibrium limit cross-ancestry generalizability (Chen et al., 2024). Lastly, methodological limitations persist, including insufficient statistical power and deficiencies in colocalization analysis (Hormozdiari et al., 2016).

Diversity across populations is a critical factor to consider when designing future genetic studies. Research involving individuals of European ancestry is disproportionately represented, with a significant gap in studies focusing on individuals of African ancestry (Fatumo et al., 2022).

Beyond eQTLs, other QTL datasets offer valuable insights that can enhance the interpretation of GWAS findings by revealing various molecular mechanisms that link genetic variation to complex traits (Aguet et al., 2023). For instance, splicing QTLs (sQTLs) reveal how genetic variation influences RNA splicing, which creates variation in the diversity of transcript isoforms and contributes to gene expression differences. Methylation QTLs (meQTLs) connect genetic variants to changes in DNA methylation by identifying the proportion of methylated cytosines at targeted sites across the genome, a process through which cytosine methylation regulates gene expression and reflects epigenetic processes. Protein QTLs (pQTLs) directly link genetic variants to protein levels, offering key insights into how genetic variation influences protein levels and their regulation. Additionally, chromatin accessibility

QTLs (caQTLs) reveal how genetic variation influences chromatin accessibility and activity, providing insight into the molecular mechanisms that regulate gene expression.

3 Aims

Recent genetic discoveries from the FinnGen study involve variants that are enriched in the Finnish population (Kurki et al., 2023), complicating the analysis of functional effects in data from non-Finnish groups. We believe that combining FinnGen findings with functional genomics data from Finnish individuals will help us understand how these genetic variants are linked to diseases. For this purpose, we are using Finnish liver transcriptome data from 131 individuals collected for the FinnLiver dataset. Since the transcriptome data originates from the liver tissue, it is expected that the most significant findings from the analyses will be related to cardiometabolic diseases and traits.

This thesis has four aims:

1. Determining how gene expression traits are regulated by genetic factors in the liver tissue.
2. Describing results from colocalization integration of FinnLiver transcriptomics with the disease and trait associations discovered in FinnGen.
3. Providing new insights into disease mechanisms by identifying and describing specific GWAS loci that share a causal variant with eQTL data from FinnLiver.
4. Finally, comparing obtained results with colocalization results from GTEx liver to assess if genetic variants enriched in the Finnish population explain some of the differences in colocalizing signals between the datasets.

4 Material and methods

4.1 Data sets

4.1.1 FinnLiver

The FinnLiver data is a set of liver tissue RNA sequencing and array genotyping from 131 Finnish individuals with ages ranging from 26 to 67, with a mean of 50 years. Liver biopsies were taken during bariatric surgery, which means all individuals were obese at sample collection. Liver samples were collected by Professor Hannele Yki-Järvinen from the University of Helsinki.

Transcriptome and genotyping data were generated together with funding from the FinnGen project. Array genotyping has been conducted with Illumina GlobalScreeningArray-24v3-0 using a Multi-Disease bead chip customized with 34191 Finnish variants. Genotype imputation was performed using SiSU v4.2, which is the same reference panel as used in FinnGen. Poly-A stranded Illumina RNA sequencing was performed with an average of 150M reads per sample, and 150 bp reads at the FIMM Technology Centre.

The eQTL Catalogue (Kerimov et al., 2021) pipeline processes gene expression and genotype data through several key steps: data preparation and cleaning, expression normalization, association testing between genetic variants and gene expression, and result summarization. The pipeline outputs a list of identified eQTLs, including SuSiE fine-mapped credible sets and those identified using permutation methods. The pipeline is described in more detail at the eQTL Catalogue website: https://www.ebi.ac.uk/eqtl/Methods/.

4.1.2 GTEx

The Genotype-Tissue Expression (GTEx) project aims to improve the understanding of mechanisms of complex traits and diseases by building a comprehensive, publicly available database for genetic effects on gene expression (The GTEx Consortium, 2020). The project, launched in 2010, was set to collect samples from over 50 tissues and over 1000 post-mortem donors while creating optimized standards and protocols for tissue collection, donor recruiting, sample handling, and data sharing. The GTEx version 8 dataset includes 838 donors, 17,382 samples from 52 different tissues, and

two cell lines that passed through quality control. Based on ancestry, the donors were European American 715 (85.3%), African American 103 (12.3%), Asian American 12 (1.4%), and 16 (1.9%) Hispanic or Latino ethnicity. Of the 838 donors, 557 (66.4%) were male and 281 (33.5%) female.

In this analysis, we focus on samples collected from the liver tissue (N=208). The RNA-seq data is processed via the same eQTL catalog pipeline as FinnLiver.

The key differences to FinnLiver are:

1. Genetic ancestry of the donors
2. Samples collected from deceased individuals within 24 hours of death
3. WGS for genotyping, i.e., ~all variation captured
4. RNA-seq with 76bp paired-end reads and ~50M read depth

4.1.3 FinnGen

The FinnGen study is a comprehensive public-private research initiative designed to advance our understanding of disease genetics (Kurki et al., 2023). It is a collaboration of nine Finnish biobanks, various research institutes, universities, and university hospitals, as well as 13 international pharmaceutical companies and the Finnish Biobank Cooperative (FINBB). The study leverages nationwide health registries that contain data from nearly every Finnish resident, including details about hospitalizations, prescription drug purchases, medical procedures, and deaths. Its latest release (R12) includes genotypic, phenotypic, and endpoint data for 520,210 individuals from Finland (FinnGen, 2024).

Publicly available results for FinnGen GWAS R12 are available on the FinnGen browser (r12.finngen.fi).

FinnGen GWAS results (R12) were used for colocalization analysis with 500,348 samples, of which 282,064 were females and 218,284 were males. Colocalization analysis is described in 3.1.4 Colocalization analysis.

FinnGen variant annotations were used for functional annotation analysis to describe variants (available for FinnGen researchers).

FinnGen disease endpoints were used for disease and trait categorization for GWAS traits (available at https://www.finngen.fi/en/researchers/clinical-endpoints, corresponding version DF12).

4.1.4 Colocalization analysis

Colocalization analysis was conducted by Juha Mehtonen from FIMM using the FinnGen colocalization pipeline for fine-mapped eQTLs from FinnLiver using SuSiE credible sets and 2,502 GWAS associations from FinnGen.

GTEx and FinnGen colocalization results were available from FinnGen conducted with the same pipeline and FinnGen DF12.

As a result, there was data with unfiltered colocalizations, including all pairs of eQTL and GWAS results, that had at least one shared variant in the credible set. In addition to that, the pipeline created pre-filtered colocalization results with the following thresholds:

1. PP.H4 > 0.5
2. cs_log10bf_thresh1 > 0.9
3. cs_log10bf_thresh2 > 1.0.

PP.H4 (Posterior Probability for Hypothesis 4): Used in the COLOC method, PP.H4 represents the posterior probability that a single shared variant is responsible for both GWAS and eQTL signals.

The thresholds cs_log10bf_thresh1 > 0.9 and cs_log10bf_thresh2 > 1.0, based on the SuSiE fine-mapping method, represent the minimum log10 Bayes factor required to declare additional independent signals confidently. A cs_log10bf > 0.9 (applied to FinnLiver eQTL data) indicates that a model with an additional signal is approximately 8 times more likely than a simpler model with one fewer signal, while a cs_log10bf > 1.0 (applied to FinnGen GWAS data) requires at least 10 times greater likelihood for an additional signal.

4.1.5 Other data

GnomAD constraint scores were used for gene constraint analysis (Karczewski et al., 2020; available at https://gnomad.broadinstitute.org/downloads#v4-constraint).

4.2 Computational analysis

The analytical methods used in this thesis primarily involve the interpretation of datasets described in section 3.1 Data sets. Key points in the study include ensuring data purity, determining thresholds for significant results, and drawing conclusions from the data using figures and tables as guideposts.

Analysis of the results conducted for FinnLiver eQTLs, FinnGen GWAS and FinnLiver eQTL colocalization and FinnGen GWAS and GTEx colocalization were conducted using R (version 4.3.2 GUI 1.80 Big Sur Intel build (8281)) in RStudio (version 2024.04.1+748) interface, with packages: tideverse (v. 2.0.0), scales (v. 1.3.0), ggvenn (v. 1.0.10), viridisLite (v. 0.4.2), grid (v. 4.3.2) and gridExtra (v. 2.3).

4.2.1 Comparing statistical methods for identifying significant eQTLs

We evaluated two statistical methods–fine-mapping and permutation–to identify significant eQTLs. To determine a comparable threshold for permutation results, we first identified the number of significant genes from the fine-mapping method. We then examined the p-beta value corresponding to the 1,836th most significant gene in the permutation results. This value (≤ 0.000816386) was then rounded to 0.001 for simplicity. We used this p-beta threshold of < 0.001 for permuted results to enable a direct comparison between the two methods in identifying significant genes.

4.2.2 Excluding colocalizations that are not truly colocalized

From 1,243 initial FinnGen and FinnLiver colocalizations in the pre-filtered data, we excluded 298 cases due to low-purity GWAS loci or eQTL. Among the remaining results, 282 colocalizations showed no credible set overlap between GWAS and eQTL and were also excluded, as the absence of shared variants in the credible sets indicates that these signals are highly unlikely to share the same causal variants and are therefore not truly colocalized.

CLPA: Causal Posterior Agreement is a colocalization metric specifically designed to evaluate the concordance between fine-mapping results from two distinct studies or phenotypes. Unlike the Causal Posterior Probability (CLPP), CLPA is uniquely independent of credible set size, making it a more robust measure in scenarios where credible set sizes vary or are influenced by linkage disequilibrium structures.

PP.H4: described in 3.1.4 Colocalization analysis

4.2.3 Disease endpoints for colocalized GWAS loci traits

To categorize colocalizations by disease type, we merged FinnGen disease endpoints matching the GWAS loci traits (3.1.3 FinnGen). For 39 traits, no matching disease endpoint was available, and it was manually assigned to them. These corresponded to 29 unique genes and 16 unique GWAS traits.

5 Results

5.1 Identification and characterization of eQTLs

Our study conducted an in-depth analysis of eQTLs, examining the results of 17,584 genes assessed for potential eGene status. Our comparison of the results of fine-mapping and permutation methods revealed that both methods largely identified the same significant genes, with an overlap of 95.4% (figure 2). Given this high concordancy, we confidently proceeded with the eQTLs found by the fine-mapping method for further analyses. This streamlined approach allowed for a more focused and thorough investigation. Ultimately, the fine-mapping method identified significant eQTLs in 1,836 unique genes, representing 10.44% of the total genes analyzed.

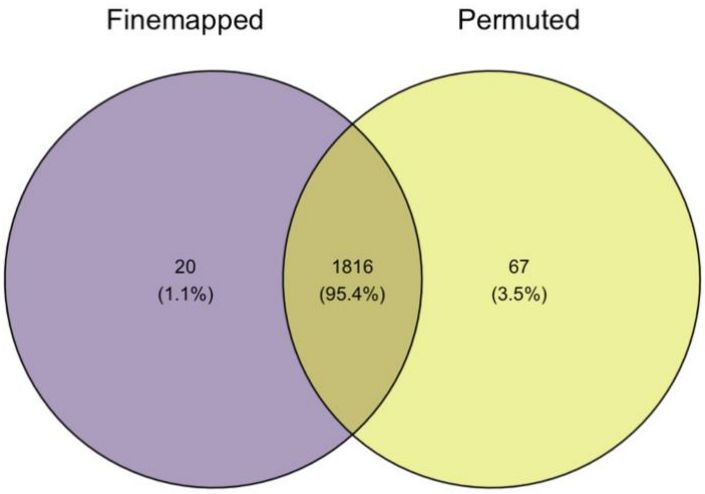

Figure 2: Significant eGenes picked up using fine-mapping and/or permuting statistical methods. 1,836 genes identified as significant using the fine-mapping method were selected for further analysis. This excluded 67 eGenes that were found significant only with the permuting method. In total, 15,748 genes were excluded from further analysis.

As we explored our eGene profile further, we discovered an intriguing pattern in signal distribution (figure 3). While the majority of eGenes, 1,613, exhibited only a single significant credible set, a subset revealed a more complex picture. We observed 198 with two significant signals, 21 with three, three with four, and one exceptional case with five significant credible sets. In total, we identified 2,089 significant signals across all eGenes. By taking into account all the credible sets and not just primary

signals, we extended our results by 12.11%. This approach allowed us to identify 253 additional signals that would have been overlooked if we had focused solely on primary signals.

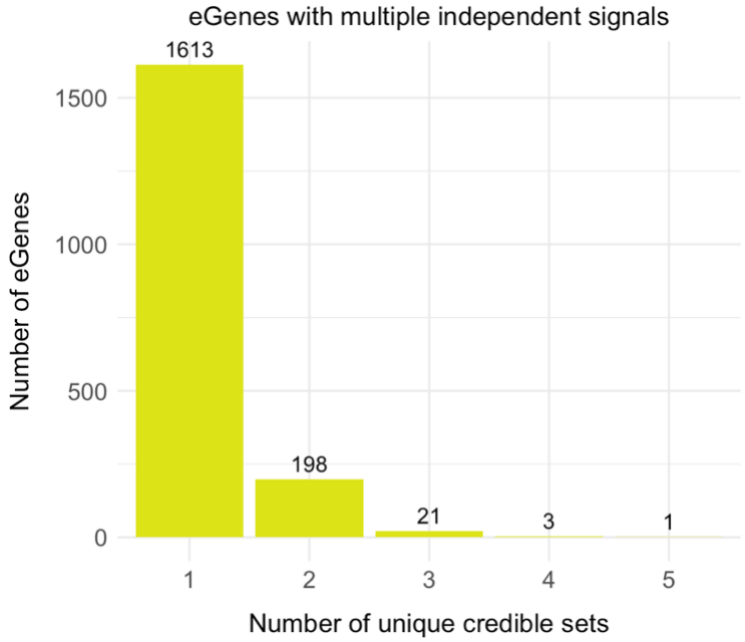

Figure 3: Number of credible sets in eGenes. Nearly 90% of the eGenes had only one significant credible set. 198 eGenes had two credible sets, and 25 eGenes had more than two.

To examine potential factors associated with eGene status, we analyzed gene expression and constraint levels, as these may play a role in eQTL formation. Genes were divided into ten bins based on these factors. Our analysis uncovered a trend: the proportion of eGenes increased along with higher gene expression and LOEUF levels, where high LOEUF indicates lower constraint (figure 4). We also observed that 4,752 genes lacked constraint information, which could limit the robustness of interpretations regarding the relationship between constraint and eGene status. 472 of these genes were identified as eGenes.

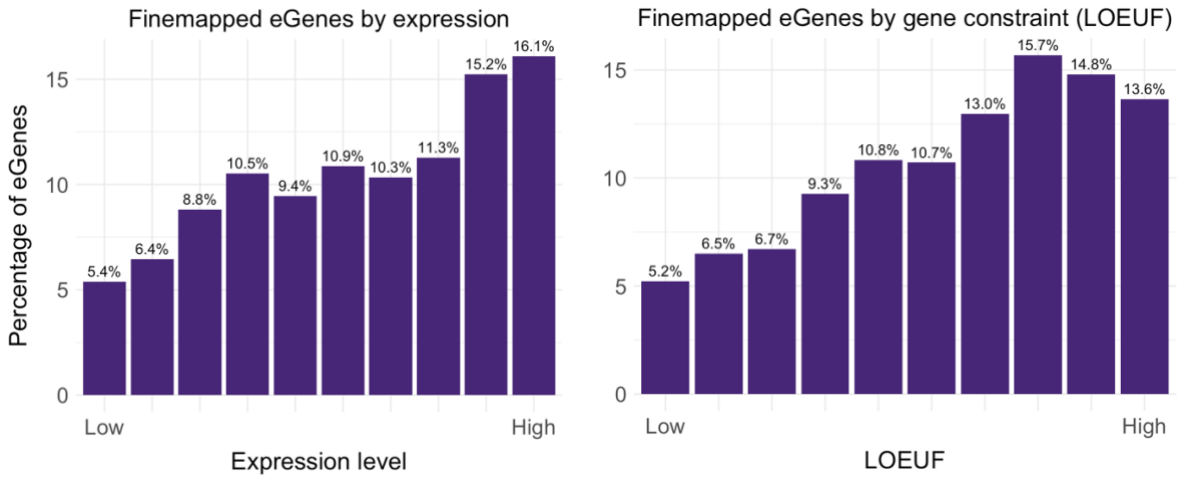

Figure 4: Positive correlation between being an eGene and expression / LOEUF (The Loss-of-Function Observed/Expected Upper Fraction) level. The percentage of eGenes increases as the expression and LOEUF level of the gene increases.

To gain insights into the functional relevance of identified eQTLs, we merged our results with FinnGen variant annotations. Our analysis revealed that 2% of eQTLs contained coding variants in their credible sets. Within this subset, synonymous and missense variants were the most common types (figure 5). We observed that coding variants appeared more frequently in smaller credible sets compared to larger ones (figure 6).

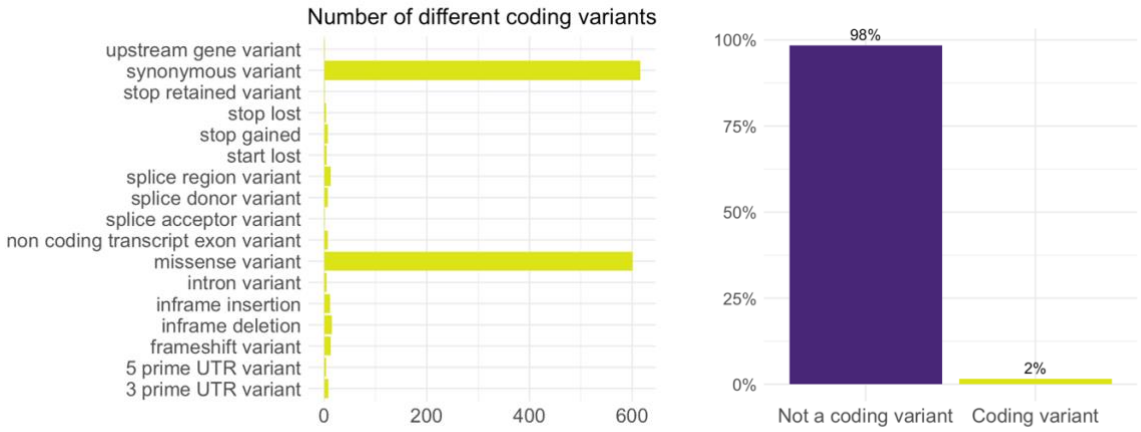

Figure 5: Coding variant status and distribution of coding variant types. Left: The most common types of variants are synonymous and missense variants. Right: Only 2 % of the significant variants are coding variants.

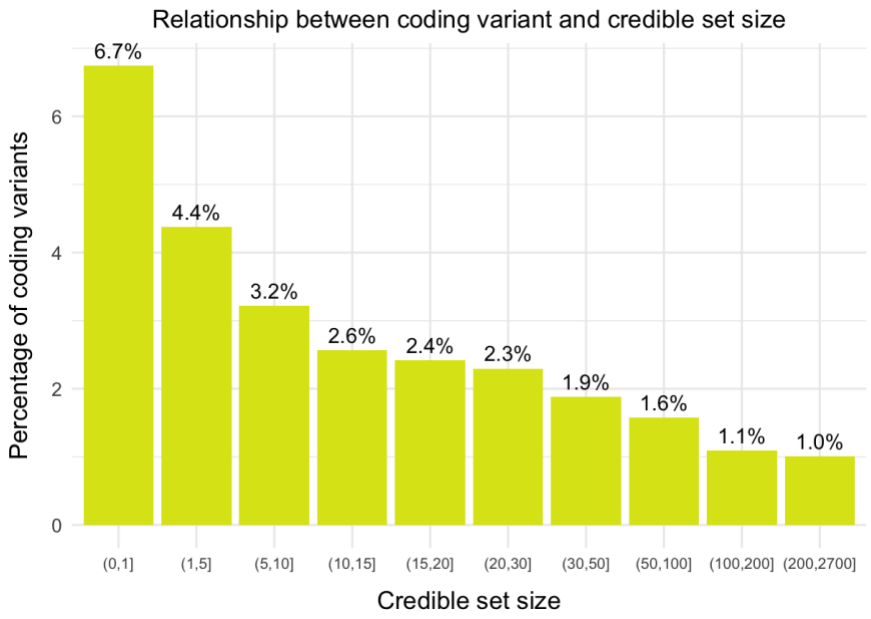

Figure 6: Relationship between being a coding variant and the size of the credible set. The smaller the credible set is, the higher the percentage of being a coding variant.

5.2 Colocalization analysis reveals associations with circulatory and metabolic diseases

To investigate potential shared genetic effects between complex traits and gene expression, we examined colocalizations between GWAS loci and eQTL credible sets. In total, 16,388 GWAS loci and 2,346 eQTL credible sets were analyzed. After excluding results with low purity or lacking credible set overlap, 663 colocalizations remained.

Figure 7 illustrates the relationship between PP.H4, CLPA values, and credible set sizes. It demonstrates PP.H4's sensitivity to credible set size, with higher values for smaller sets, while CLPA shows no clear correlation with credible set size. We also observed a positive correlation between PP.H4 and CLPA values themselves (figure 8). We examined these relationships to guide our threshold selection for significant colocalizations. To account for both metrics' characteristics, we assessed colocalization significance using both PP.H4 and CLPA values, applying thresholds of PP.H4 ≥ 0.9 and CLPA ≥ 0.5. This combined approach resulted in 180 significant colocalizations, with CLPA alone removing 128 and PP.H4 removing 100 colocalizations.

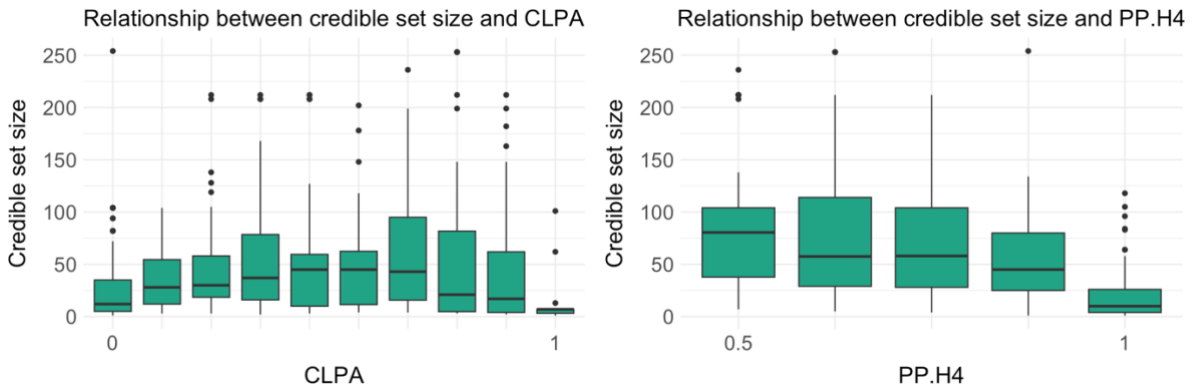

Figure 7: Relationship between eQTL credible set size and CLPA/PP.H4 values FinnGen and FinnLiver colocalizations. Left: The data does not show any relationship between credible set size and CLPA value. Right: The PP.H4 values are the highest (most significant) with the smallest credible sets.

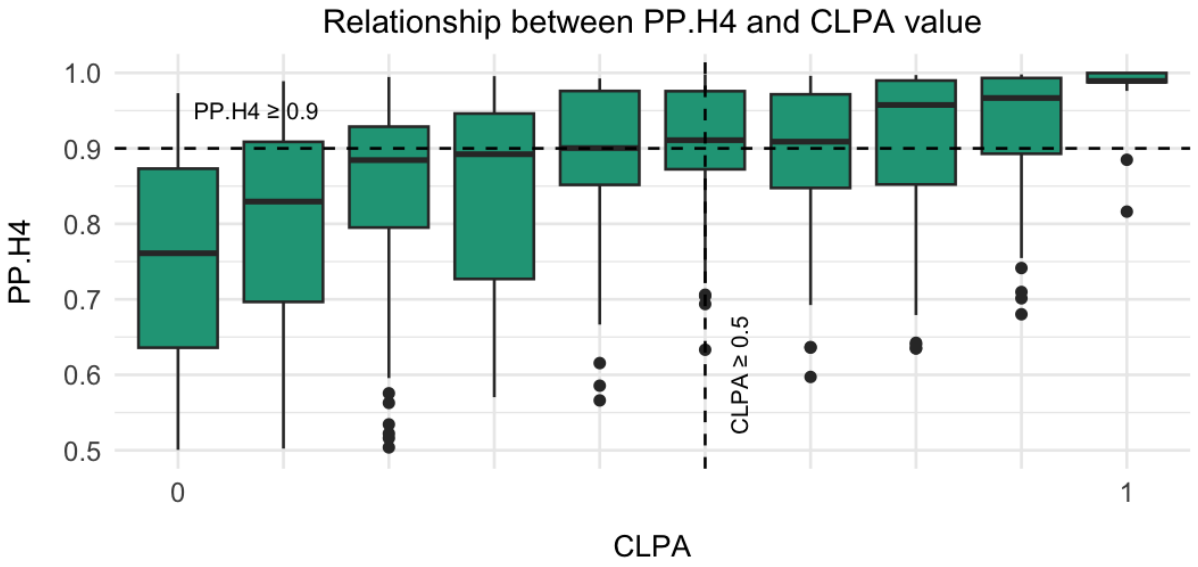

Figure 8: Relationship between PP.H4 and CLPA values in FinnGen and FinnLiver colocalizations. We observe a positive correlation between PP.H4 and CLPA.

Significant colocalizations were categorized by disease type (figure 9). Some GWAS traits corresponded to multiple disease endpoints, resulting in certain colocalizations appearing in multiple disease categories. The most prevalent disease endpoints were 'diseases of the circulatory system' (58 colocalizations), 'endocrine, nutritional and metabolic diseases' (46 colocalizations), and 'diseases of the digestive system' (19 colocalizations). Anthropometric traits accounted for 18 colocalizations, while the remaining 10 categories each had fewer than 10 colocalizations.

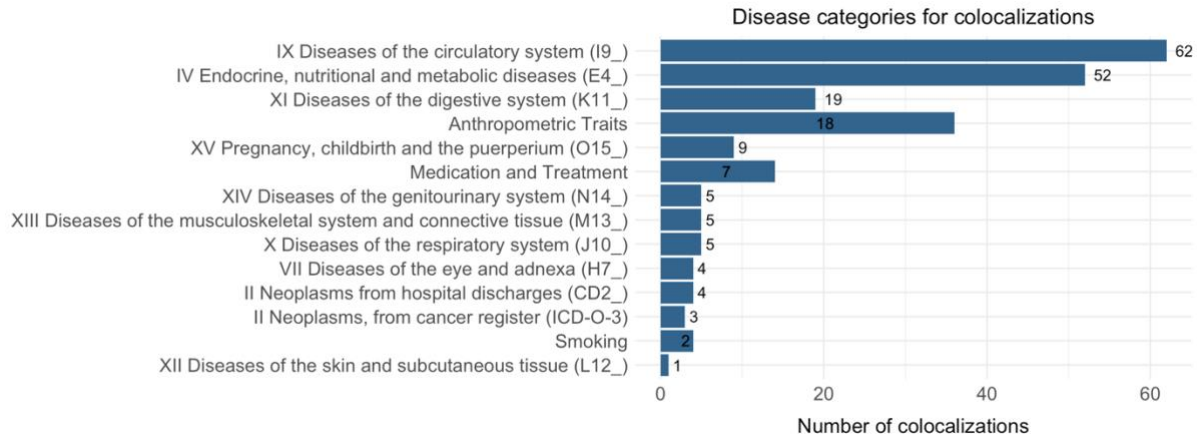

Figure 9: Associated disease categories for GWAS (Genome-wide association study) traits in significant colocalizations. Some traits are associated with multiple disease categories, and their colocalizations are presented in more than one category. Of the total 14 categories, three major ones are 'Diseases of the circulatory system', 'Endocrine, nutritional and metabolic diseases', and 'Diseases of the digestive system'.

We selected two categories, 'diseases of the circulatory system' and 'endocrine, nutritional and metabolic diseases', for further analysis, as these were the most common categories and are most relevant to our liver-derived sample. We identified genes that colocalized exclusively with GWAS traits belonging to a single disease endpoint: three genes (*ACTR1B, PPIL3*, and *CTSH*) with 'endocrine, nutritional and metabolic diseases', and six genes (*ARHGAP10, FADS1, SKI, DMTN, SLAC22A1,* and *RP11-118B18.2*) with 'diseases of the circulatory system' (figure 10).

We also identified three instances where multiple genes colocalized with the same GWAS locus (table 1). *FADS1, FADS2,* and *FADS3* all colocalized with venous thromboembolism and had the same credible set lead variant. *CELSR2*, *PSMA5*, *PSRC1,* and *SORT1* all colocalized with multiple metabolic and cardiovascular-related traits and diseases such as pure hypercholesterolemia and coronary atherosclerosis.


Disease categories and associated genes
Count
15
10
5
0
15
2
2
17
16
7
1
3
2
14
1
4
3
3
RP4-781K5.7
PSMB3
RP4-591C20.9
RP11-118B18.2
RP11-345M22.1
CTD-3098H1.2
RP11-10A14.5
CFHR1
STMP1
UGT1A1
PPIL3
LINC01750
AMA5-AS1
APOC4-APOC2
FADS3
SLC35E2B
LITAF
ZNF197
SLC22A1
TLR1
SPTLC3
C15orf40
SCARA5
SLC22A11
RPSA
VKORC1
SCG5
LIPC
SLC2A2
DBF4B
CYP3A7
ADAR
DMTN
SKI
AP3S2
PRSS53
FADS1
SLC22A3
CELSR2
PSMA5
SCRN2
MAN2C1
MFSD13A
FADS2
SORT1
PSRC1
C1QTNF6
APOE
XRCC3
PAX8
EEF1AKMT3
SPP1
CFHR3
ACTR1B
LTBR
NRXN2
STN1
TMEM147
SULT2A1
CTSH
CWF19L1
WHRN
SRRT
ARHGAP10
RASGRP2
1
2
3
4
Anthropometric Traits
II Neoplasms from hospital discharges (CD2_)
II Neoplasms, from cancer register (ICD-O-3)
IV Endocrine, nutritional and metabolic diseases (E4_)
IX Diseases of the circulatory system (I9_)
Medication and Treatment
Smoking
VII Diseases of the eye and adnexa (H7_)
X Diseases of the respiratory system (J10_)
XI Diseases of the digestive system (K11_)
XII Diseases of the skin and subcutaneous tissue (L12_)
XIII Diseases of the musculoskeletal system and connective tissue (M13_)
XIV Diseases of the genitourinary system (N14_)
XV Pregnancy, childbirth and the puerperium (O15_)

Figure 10: Disease endpoints and their associated genes. Above the heatmap is the number of unique genes per disease endpoint, and on the right is the number of disease endpoints associated with the gene.

Table 1: Genes that colocalized to the same GWAS locus.

GWAS region:	chr1:107774968-110775908	chr11:60276489-63276489	chr19:41771969-46409976
Genes:	*CELSR2*	*FADS1*	*APOE*
	PSMA5	*FADS2*	*APOC4-APOC2*
	PSRC1	*FADS3*	-
	SORT1	-	-

5.3 Comparative analysis reveals differences in GTEx and FinnLiver colocalizations

Pre-filtered results of GTEx and FinnGen colocalizations yielded 998 findings. After applying the same thresholds used in FinnGen and FinnLiver colocalizations (PP.H4≥0.9 and CLPA≥0.5) and removing low-purity results, we identified 95 significant colocalizations. This number is approximately half of the significant colocalizations found in the FinnLiver GWAS analysis. CLPA alone removed 229 colocalizations, and PP.H4 removed 40, with CLPA excluding a substantially larger proportion than in the FinnLiver colocalizations.

The GTEx and FinnGen colocalization results showed similar patterns to the FinnLiver and FinnGen colocalization. Figure 11 illustrates that CLPA values do not correlate with credible set size, while PP.H4 demonstrates a negative correlation with credible set size. The PP.H4 and CLPA relationship also showed a positive correlation; however, it was not as clear as in FinnLiver colocalizations (figure 12). This consistency across datasets supports the robustness of our thresholding approach.

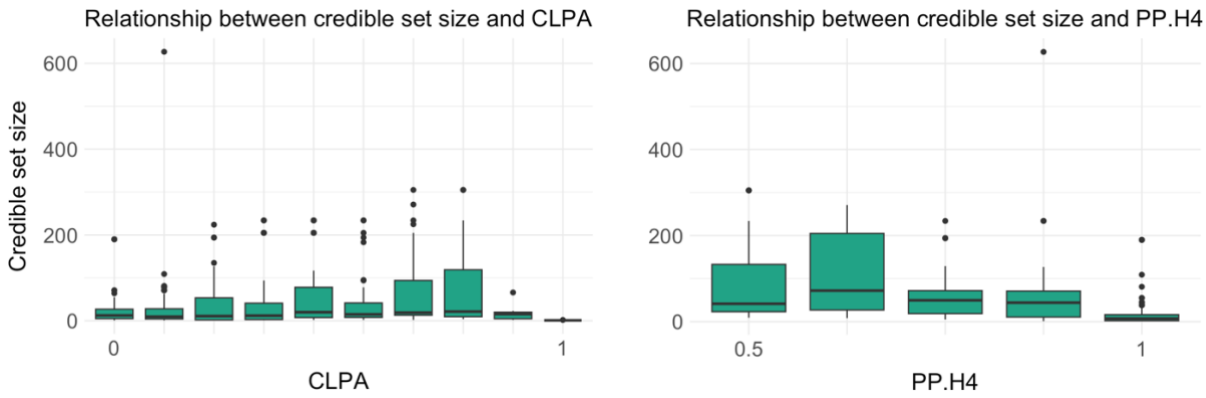

Figure 11: Relationship between eQTL credible set size and thresholds in FinnGen and GTEx colocalizations. The colocalizations behave the same way in both FinnLiver and GTEx colocalizations with FinnGen. Left: The data does not show any relationship between credible set size and CLPA value. Right: The PP.H4 values are the highest with the smallest credible sets.

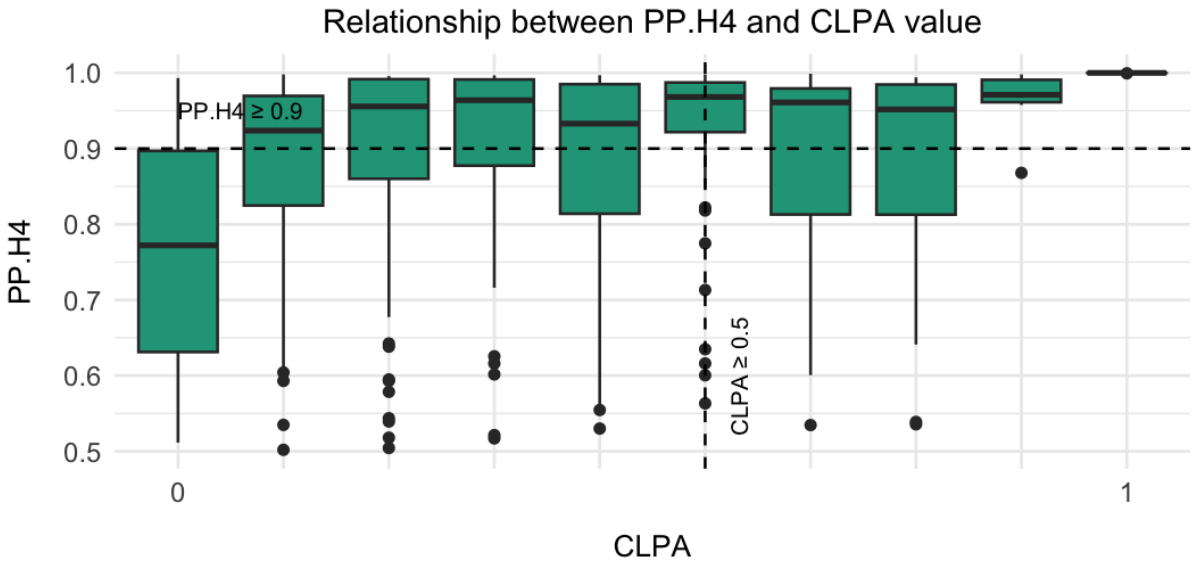

Figure 12: Relationship between PP.H4 and CLPA thresholds in FinnGen and GTEx colocalizations. We observe a positive correlation between PP.H4 and CLPA.

Analyzing the disease endpoints, we found that 'Diseases of the circulatory system' was the most prevalent category, aligning with our liver-focused study. The second most common group was 'Diseases of the eye and adnexa' (figure 13), which differs from the most common categories identified in FinnLiver colocalizations.

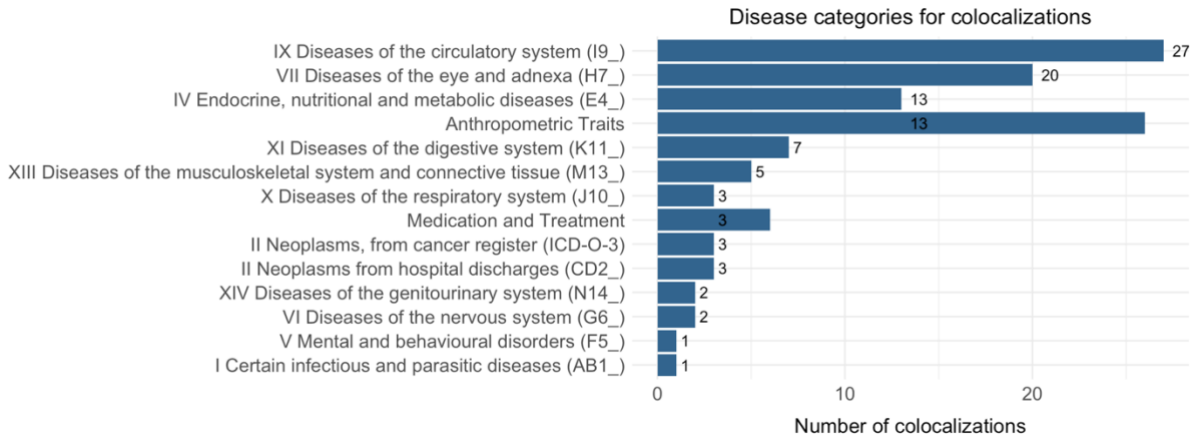

Figure 13: Associated disease categories for GWAS (Genome-wide association study) traits in significant colocalizations in GTEx. Some traits are associated with multiple disease categories and are presented in more than one disease category. Of the total 14 categories, two major categories are 'Diseases of the circulatory system' and, surprisingly, 'Diseases of the eye and adnexa'.

For further analysis, we focused on liver-related disease categories: 'diseases of the circulatory system' (FinnLiver: 54, GTEx: 24) and 'endocrine, nutritional and metabolic diseases' (FinnLiver: 42, GTEx: 10), with some colocalizations in both categories (FinnLiver: 4, GTEx: 1). Of the total 135 colocalizations in these categories, 128 were identified only by one dataset, with 99 unique to FinnLiver and 29 to GTEx. This indicates a limited overlap between the two datasets in these disease categories. Genes that colocalized significantly in only one dataset are presented in tables 2 and 3. The tables also show that these colocalizations were primarily driven by common variants.

FinnLiver identified two variants associated with *ACTR1B* and *DMTN* that were not investigated for colocalization or eQTL effects in GTEx. Among genes associated with only one disease trait, *ACTR1B* and *PPIL3* (also known as *CYPJ*) colocalized with Hypothyroidism, strict autoimmune (E4_HYTHY_AI_STRICT) in the FinnLiver dataset. Increased expression of *ACTR1B* was positively associated with disease development or progression, whereas higher expression of *PPIL3* showed a protective effect. *PPIL3* also exhibited a stronger effect size compared to *ACTR1B*.

Table 2: Genes that colocalized only in the FinnLiver dataset.

Variant N/A: Variant not included in eQTL analysis.

Gene name	Gene (ENS)	MAF GTEx	MAF FinnLiver	beta GTEx	beta FinnLiver	Lead Variant (eQTL credible set)
PSRC1	ENSG00000134222	0.237981	0.225191	1.273360	1.29204	chr1_109274968_G_T
SORT1	ENSG00000134243	0.240385	0.225191	1.303190	1.49416	chr1_109274570_A_G
PSMA5	ENSG00000143106	0.237981	0.225191	0.200399	0.894702	chr1_109274968_G_T
CELSR2	ENSG00000143126	0.237981	0.225191	1.119850	1.372980	chr1_109274968_G_T
UGT1A1	ENSG00000241635	0.302885	0.400763	-0.265328	-1.02891	chr2_233755940_ATC_A
LINC01229	ENSG00000260876	0.319712	0.316794	-0.780129	-0.839781	chr16_79722300_A_T
APOE	ENSG00000130203	0.259615	0.255725	-0.151460	-0.923941	chr19_44935906_C_G
ACTR1B	ENSG00000115073	variant N/A	0.251908	variant N/A	0.992563	chr2_97646864_A_AT
PPIL3/CYPJ	ENSG00000240344	0.247596	0.263359	-0.788865	-1.21038	chr2_200877622_C_T
FADS2	ENSG00000134824	0.372596	0.438931	-0.0391377	-0.453825	chr11_61776489_T_C
APOC4-APOC2	ENSG00000224916	0.451923	0.469466	0.117036	0.30441	chr19_44927023_C_G

ARHGAP10	ENSG00000071205	0.274038	0.263359	0.765154	1.31499	chr4_148061408_G_A
SKI	ENSG00000157933	0.5	0.461832	-0.174269	-0.426897	chr1_2271528_T_C
PPP1R3B-DT	ENSG00000248538	0.09375	0.125954	0.293188	1.08092	chr8_9325592_A_AAGAAGAAAGGG
FADS1	ENSG00000149485	0.372596	0.438931	-0.2305290	-0.602363	chr11_61776489_T_C
FADS3	ENSG00000221968	0.372596	0.438931	-0.3101580	-0.476325	chr11_61776489_T_C
DMTN	ENSG00000158856	variant N/A	0.20229	variant N/A	-1.27313	chr8_22049278_G_T
CTSH	ENSG00000103811	0.105769	0.0916031	-0.629268	-1.40921	chr15_78942128_T_C
WHRN	ENSG00000095397	0.185096	0.259542	0.475576	0.711209	chr9_114404267_A_G
AP3S2	ENSG00000157823	0.28125	0.259542	-0.624909	-0.869209	chr15_89905903_A_G

Table 3: Genes that colocalized only in the GTEx dataset.

Gene N/A: Gene not included in eQTL analysis.

Gene name	**Gene (ENS)**	**MAF GTEx**	**MAF FinnLiver**	**beta GTEx**	**beta FinnLiver**	**Variant**
DINOL	ENSG00000285244	0.201923	gene N/A	0.533309	gene N/A	chr6_36680287_CGCGT_C
SYPL2	ENSG00000143028	0.245098	0.217557	-0.533201	-0.741251	chr1_109275536_C_CT
TGFB1	ENSG00000105329	0.141827	0.21374	0.502741	0.475904	chr19_41284181_GTTATGGTA_G
PFDN1	ENSG00000113068	0.444712	0.438931	0.590329	0.37105	chr5_140335105_T_C
SLC12A2-DT	ENSG00000245937	0.211538	0.179389	-0.809434	-0.559533	chr5_128053152_G_A
SPTY2D1	ENSG00000179119	0.310096	0.412214	-0.500078	-0.424876	chr11_18618930_C_T
CAMK2B	ENSG00000058404	0.141827	0.232824	-0.484003	-0.500724	chr7_44322946_C_T

6 Discussion and conclusions

6.1 Characteristics of FinnLiver eQTLs

Multiple eQTL studies have been conducted across various tissue types. For instance, the analysis of GTEx Consortium version 8 data alone included over 15,000 samples from 49 different tissues (The GTEx Consortium, 2020). Our aim was to investigate how gene expression traits are regulated by genetic factors in the liver tissue and to determine whether our FinnLiver eQTLs exhibit similar patterns to those described in previous studies. For this purpose, we analyzed the relationships between our eQTLs and gene expression, gene constraint levels, and functional annotations.

Comparing our results with the liver tissue eQTL mega-analysis by Strunz et al. (2018) reveals interesting contrasts. Despite their approximately fourfold larger sample size and greater statistical power, we identified a similar proportion of eGenes. This may suggest that our fine-mapping approach using credible sets offers an advantage in identifying likely causal variants, especially compared to their analysis based on individual variants and a combination of microarray and RNA-sequencing data. Additionally, the homogeneity of the FinnLiver dataset may have enhanced our ability to detect relevant genetic signals.

Genes with low LOEUF values are highly constrained and less tolerant to mutations, often playing essential roles in vital biological processes (Karczewski et al., 2020). Changes in their expression or function can have detrimental effects. Because eQTLs influence gene expression, they are more commonly associated with genes that are more mutation-tolerant (Mostafavi et al., 2023). Our results show a higher proportion of eGenes with high LOEUF values, indicating lower constraint and greater tolerance to LoF mutations, consistent with previous findings (Umans et al., 2021). Prioritizing eGenes under stronger constraint may help identify functionally important genes.

Regulatory variants influence gene expression, often through regions such as promoters and enhancers located upstream of coding sequences (Gallagher & Chen-Plotkin, 2018). Our finding that only 2% of significant eQTLs are coding variants aligns with this. Large-scale studies, including GTEx (The GTEx Consortium, 2020) and FinnGen (Kurki et al., 2023), also show that most eQTLs are non-coding. The

most common coding variants in our data are missense and synonymous. While synonymous variants do not change the amino acid sequence, they can affect mRNA stability or splicing, thereby potentially influencing gene expression. (Gaither et al., 2021). Missense variants alter the amino acid sequence and can affect protein structure or interactions (Jänes et al., 2024). While their primary impact is at the protein level, they may also influence mRNA or other molecular processes. Given these properties, the presence of such variants in our credible sets is not unexpected.

In a recent study, primary eQTL signals were found to have larger effect sizes, higher MAFs, and to be located closer to the TSS compared to non-primary signals (Brotman et al., 2025). It was also shown that more eQTLs are identified for genes with higher expression levels. Consistent with this, we found that genes with higher expression in our dataset were more likely to be identified as eGenes. The same study reported that 46% of eQTL-GWAS colocalizations would have been missed if only primary eQTLs had been considered. This underscores the importance of applying analytical methods capable of identifying multiple causal variants within a gene rather than focusing only on a single variant. Using the SuSiE model, we were able to capture multiple causal signals within genes effectively.

While we did not explore differences between genes with multiple distinct signals in detail, the number of genes with two or more significant credible sets in our dataset is sufficient to support further investigation. Studying these differences could provide insights into the regulatory mechanisms underlying gene expression, such as whether genes with multiple signals are enriched in specific biological pathways, exhibit tissue-specific regulation, or are associated with particular phenotypes or diseases. Such analyses could also help clarify how primary and non-primary signals interact to influence gene function and trait associations.

6.2 Describing FinnGen colocalization results with FinnLiver and GTEx

Given that the transcriptomic data was derived from liver tissue, we expected a predominance of significant colocalizations associated with cardiometabolic diseases and traits. Our colocalization results confirmed this hypothesis, with 104 out of 184 GWAS trait hits linked to cardiometabolic conditions. This outcome reinforces the importance of selecting tissue samples that are most relevant to the disease under investigation.

When comparing our colocalization results in categories' diseases of the circulatory system' and 'endocrine, nutritional, and metabolic diseases' to those reported in the FinnGen Browser, we identified a few colocalizations not previously reported there. These included *UGT1A1* with Gilbert syndrome and disorders of porphyrin and bilirubin metabolism, *SKI* with statin medication use, *FADS3* with venous thromboembolism, and *PSMA5* with multiple metabolic and cardiovascular traits. All of these genes have previously been associated with the respective traits through GWAS, but our gene expression analysis provides additional support for their potential causality.

Findings, where multiple genes are colocalized to the same GWAS loci indicate that colocalization alone may not always pinpoint the most causal gene. *FADS*1-2-3 region is associated with cardiometabolic diseases such as type 2 diabetes and coronary artery disease (Yuan et al., 2019; Lattka et al., 2010). *FADS1* has been linked to a reduced risk of venous thromboembolism, likely through its role in polyunsaturated fatty acid metabolism, a pathway in which *FADS2* is also involved. While *FADS3* colocalizes to the same locus, its function is less well understood.

For the four genes colocalized with the known lipid locus at 1p13, a previous study identified *SORT1* as the true causal gene among *CELSR2, PSRC1,* and *PSMA5* (Musunuru et al., 2010). The study found that a specific non-coding variant increases hepatic expression of *SORT1*, which in turn reduces very low-density lipoprotein secretion, thereby lowering plasma LDL cholesterol levels. These findings highlight *SORT1's* critical role in lipoprotein metabolism and its potential as a therapeutic target for reducing LDL cholesterol and cardiovascular disease risk. Both *CELSR2* and *PSRC1*, along with *SORT1*, showed significant differences in gene expression between liver and adipose tissue. However, *PSMA5* did not exhibit any changes in gene expression between tissues, and its causality was not further investigated in that study.

6.3 Does the origin of data explain differences in colocalization signals between datasets?

Comparison of FinnLiver and GTEx liver eQTLs in relation to FinnGen GWAS results revealed substantial differences in the number of significant colocalizations. Despite its smaller sample size (131 individuals compared to 208 in GTEx), FinnLiver showed

nearly twice as many significant colocalizations with FinnGen traits. The striking lack of overlap in colocalization signals between FinnLiver and GTEx for liver-related disease categories further highlights these differences. However, because most of the colocalizations are driven by common variants, differences in data origin alone are unlikely to explain this discrepancy fully. A more likely explanation is that our strict significance thresholds may have prevented some GTEx colocalizations from reaching statistical significance. Additionally, differences in tissue quality may contribute, as GTEx liver samples were collected post-mortem. A study analyzing post-mortem GTEx samples during the project's interim phase found that gene expression can be affected by mRNA degradation within 24 hours after death across multiple tissue types (Chu et al., 2017). However, the sensitivity to degradation varied between tissues and even among regions within the same tissue.

The genes that colocalized only in the FinnLiver dataset showed similar MAFs between datasets but consistently higher eQTL effect sizes in FinnLiver. This suggests that while the underlying genetic variants are comparable across datasets, the associations appear stronger in the Finnish cohort. However, previous studies have shown that true population differences in eQTL effect sizes are rare and are unlikely to have a biological basis (Taylor et al., 2024). In our case, the observed effect size differences are more plausibly explained by the more homogeneous phenotype in FinnLiver, where a uniform obesity profile may enhance the detectability of genetic effects.

Two significantly colocalized variants that are associated with ACTR1B and DMTN were not present in the GTEx liver eQTL dataset. Both variants are found in the European non-Finnish population but show relatively different MAFs compared to the Finnish population (chr2:97646864_A>AT — non-Finnish European MAF: 0.09, Finnish MAF: 0.25; chr8:22049278_G>T — non-Finnish European MAF: 0.17, Finnish MAF: 0.24) (Karczewski et al., 2020). This indicates that while these variants are enriched in the Finnish population, they are not absent from other European populations. Their absence in the GTEx eQTL dataset is therefore likely due to technical factors. Supporting this, Brotman et al. (2025), in a leave-one-study-out meta-analysis, found that excluding GTEx had the least impact on eQTL discovery—despite its high sequencing depth—and suggested that this may be explained by physiological or technical differences.

6.4 Limits and future prospects

Although individuals in the FinnLiver and FinnGen datasets share similar Finnish ancestry, we did not find colocalizations involving Finnish-specific variants. A likely explanation for this is the relatively small sample size of the FinnLiver dataset, which may have limited the statistical power to detect such variants. While enriched in Finland compared to other populations, many of these variants are still low-frequency within Finland itself. Additionally, the uniform clinical background of the liver samples collected during bariatric surgery under consistent conditions may have influenced the eQTL signals detected in this study.

A potential next step in our analysis could have been to include sQTLs to explore whether they might explain regulatory variants more effectively than eQTLs. sQTLs reflect a different aspect of gene regulation by capturing variation in splicing, and they have been shown to reveal variants that are often distinct from those identified through eQTLs (Qi et al., 2022). Another valuable extension of our study would be to investigate genes with multiple significant signals, as well as the differences between primary and secondary signals. This could help clarify how these signals contribute differently to gene regulation and provide insight into where future analyses should focus to more effectively identify and interpret complex associations.

Our results suggest that future studies should match the tissue of origin to the primary site of disease when designing genetic studies, as this can strongly influence the ability to detect meaningful associations. Since gene expression not only differs between tissues but also changes across developmental stages and even throughout the cell cycle within the same tissue, future research could gain important insights by examining gene expression at different stages of development.

7 Acknowledgments

I am grateful to Taru Tukiainen and her group for the opportunity to join their team and for their guidance and support.

I would also like to acknowledge Juha Mehtonen at FIMM, as well as the participants and investigators of the FinnGen study, for their contributions.

Finally, I would like to thank my parents and my partner for their endless support and encouragement throughout this process.

I also want to acknowledge that artificial intelligence was used for two purposes in this thesis:

1. To assist with debugging code in computational analysis
2. To aid in proofreading, improving flow, and refining the text during thesis writing.

The AI tools employed were ChatGPT (GPT-4), Perplexity AI (default model), and Grammarly.

References

Aguet, F., Alasoo, K., Li, Y.I., Battle, A., Im, H.K., Montgomery, S.B., Lappalainen, T. (2023). Molecular quantitative trait loci. *Nature Reviews Methods Primers*, 3,4. https://doi.org/10.1038/s43586-022-00188-6

Albert, F., Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics,* 16,197-212. https://doi.org/10.1038/nrg3891

Argentieri, M.A., Amin, N., Nevado-Holgado, A.J., Sproviero, W., Collister, J.A., Keestra, S.M., Kuilman, M.M., Ginos, B.N.R., Ghanbari, M., Doherty, A., Hunter, D.J., Alvergne, A., van Duijn, C.M. (2025). Integrating the environmental and genetic architectures of aging and mortality. *Nature Medicine,* 31,1016-1025. https://doi.org/10.1038/s41591-024-03483-9

Baralle, F., Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology,* 18,437-451. https://doi.org/10.1038/nrm.2017.27

Brotman, S. M., El-Sayed Moustafa, J. S., Guan, L., Broadaway, K. A., Wang, D., Jackson, A. U., Welch, R., Currin, K. W., Tomlinson, M., Vadlamudi, S., Stringham, H. M., Roberts, A. L., Lakka, T. A., Oravilahti, A., Fernandes Silva, L., Narisu, N., Erdos, M. R., Yan, T., Bonnycastle, L. L., … Scott, L. J. (2025). Adipose tissue eQTL meta-analysis highlights the contribution of allelic heterogeneity to gene expression regulation and cardiometabolic traits. *Nature Genetics*, 57, 180-192. https://doi.org/10.1038/s41588-024-01982-6

Chen, Y., Liu, S., Ren, Z., Wang, F., Liang, Q., Jiang, Y., Dai, R., Duan, F., Han, C., Ning, Z., Xia, Y., Li, M., Yuan, K., Qiu, W., Yan, X.-X., Dai, J., Kopp, R. F., Huang, J., Xu, S., … Chen, C. (2024). Cross-ancestry analysis of brain QTLs enhances interpretation of schizophrenia genome-wide association studies. *The American Journal of Human Genetics,* 111(11),2444-2457. https://doi.org/10.1016/j.ajhg.2024.09.001

Delaneau O., Ongen H., Brown A. A., Fort, A., Panousis, N.I., Dermitzakis, E.T. (2017). A complete tool set for molecular QTL discovery and analysis. *Nature Communications,* 8,15452. https://doi.org/10.1038/ncomms15452

Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A.R., Kuchenbaecker, K. (2022). A roadmap to increase diversity in genomic studies. *Nature Medicine*, 28,243-250. https://doi.org/10.1038/s41591-021-01672-4

FinnGen. (2024). FinnGen Documentation of R12 release. https://finngen.gitbook.io/documentation/

Gaither, J.B.S., Lammi, G.E., Li, J.L., Gordon, D.M., Kuck, H.C., Kelly, B.J., Fitch, J.R. White P. (2021). Synonymous variants that disrupt messenger RNA structure are significantly constrained in the human population. *GigaScience*, 10(4), giab023. https://doi.org/10.1093/gigascience/giab023

Gallagher, M., Chen-Plotkin, A. (2018). The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics*, 102(5),717-730. https://doi.org/10.1016/j.ajhg.2018.04.002

The GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369,1318-1330. https://doi.org/10.1126/science.aaz1776

Hormozdiari, F., Van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics*, 99(6),1245-1260. https://doi.org/10.1016/j.ajhg.2016.10.003

Huang, Q.Q., Wigdor, E.M., Malawsky, D.S. Campbell, P., Samocha, K.E., Chundru, V.K., Danecek, P., Lindsay, S., Marchant, T., Koko, M., Amanat, S., Bonfanti, D., Sheridan, E., Radford, E.J., Barrett, J.C., Wright, C.F., Firth, H.V., Warrier,

V., Young, A.S., … Martin, H.C. (2024). Examining the role of common variants in rare neurodevelopmental conditions. *Nature*, 636, 404–411. https://doi.org/10.1038/s41586-024-08217-y

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, 437,1299–1320. https://doi.org/10.1038/nature04226

Jackson, M., Marks, L., May, G.H.W., Wilson, J.B. (2018). The genetic basis of disease. *Essays in Biochemistry,* 62(5),643-723. https://doi.org/10.1042/EBC20170053

Jänes, J., Müller, M., Selvaraj, S., Manoel, D., Stephenson, J., Gonçalves, C., Lafita, A., Polacco, B., Obernier, K., Alasoo, K., Lemos, M.C., Krogan, N., Martin, M., Saraiva, L.R., Burke, D., Beltrao, P. (2024). Predicted mechanistic impacts of human protein missense variants. *bioRxiv* [Preprint], 596373. https://doi.org/10.1101/2024.05.29.596373

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D., Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., England, E.M., Seaby, E.G., Kosmicki, J.A., Walters, R.K., … MacArthur, D.G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581,434–443. https://doi.org/10.1038/s41586-020-2308-7

Kerimov, N., Hayhurst, J. D., Peikova, K., Manning, J. R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M. P., Kuzmin, I., Trevanion, S. J., Burdett, T., Jupp, S., Parkinson, H., Papatheodorou, I., Yates, A. D., Zerbino, D. R., & Alasoo, K. (2021). A compendium of uniformly processed human gene expression and splicing QTLs. *Nature Genetics*, 53, 1290–1299. https://doi.org/10.1038/s41588-021-00924-w

Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., Kathiresan, S. (2018). Genome-wide

polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics,* 50,1219–1224. https://doi.org/10.1038/s41588-018-0183-z

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C., Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science,* 308,385–389. https://doi.org/10.1126/science.1109557

Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K.M., Reeve, M.P., Laivuori, H., Aavikko, M., Kaunisto, M.A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Della Briotta Parolo, P., Lehisto, A.A., Kanai, M., Mars, N., Rämö, J., … Palotie, A. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613,508–518. https://doi.org/10.1038/s41586-022-05473-8

Lattka, E., Illig, T., Heinrich, J., Koletzko, B. (2010). Do FADS genotypes enhance our knowledge about fatty acid related phenotypes? *Clinical Nutrition,* 29,3,277-287. https://doi.org/10.1016/j.clnu.2009.11.005.

Lea, A., Peng, J., Ayroles, J. (2022). Diverse environmental perturbations reveal the evolution and context-dependency of genetic effects on gene expression levels. *Genome Research,* 10,1826–1839. https://doi.org/10.1101/2021.11.04.467311

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M., Pirruccello, J.P., Muchmore, B., Prokunina-Olsson, L., Hall, J.L., Schadt, E.E., Morales, C.R., Lund-Katz, S., Phillips, M.C., Wong, J., … Rader, D.J. (2010). From non-coding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, 466(7307),714-719. https://doi.org/10.1038/nature09266

Mostafavi, H., Spence, J.P., Naqvi, S., Pritchard, J.K. (2023). Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nature Genetics*, 55(11),1866-1875. https://doi.org/10.1038/s41588-023-01529-1

Nguyen, J.P., Arthur, T.D., Fujita, K., Salgado, B.M., Donovan, M.K.R., iPSCORE Consortium, Matsui, H., Kim, J.H., D'Antonio-Chronowska, A., D'Antonio, M., Frazer, K.A. (2023). eQTL mapping in fetal-like pancreatic progenitor cells reveals early developmental insights into diabetes risk. *Nature Communications,* 14(1),6928. https://doi.org/10.1038/s41467-023-42560-4

Nica, A., Dermitzakis, E. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368,1620. https://doi.org/10.1098/rstb.2012.0362

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526,68-74. https://doi.org/10.1038/nature15393

Pai, A., Pritchard, J., Gilad, Y. (2015). The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genetics,* 11,1. https://doi.org/10.1371/journal.pgen.1004857

Qi, T., Wu, Y., Fang, H., Zhang, F., Liu, S., Zeng, J., Yang, J. (2022). Genetic control of RNA splicing and its distinct role in complex trait variation. *Nature Genetics*, 54, 1355–1363. https://doi.org/10.1038/s41588-022-01154-4

Schoenfelder, S., Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics,* 20,437–455. https://doi.org/10.1038/s41576-019-0128-0

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

Stark, R., Grzelak, M., Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics,* 20,631–656. https://doi.org/10.1038/s41576-019-0150-2

Strunz, T., Grassmann, F., Gayán, J., Nahkuri, S., Souza-Costa, D., Maugeais, C., Fauser, S., Nogoceke, E., Weber, B.H.F. (2018). A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Scientific Reports,* 8,5865. https://doi.org/10.1038/s41598-018-24219-z

Taylor, D.J., Chhetri, S.B., Tassia, M.G., Biddanda, A., Yan, S.M., Wojcik, G.L., Battle, A., McCoy, R.C. (2024). Sources of gene expression variation in a globally diverse human cohort. *Nature,* 632,122–130. https://doi.org/10.1038/s41586-024-07708-2

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*, 1,59. https://doi.org/10.1038/s43586-021-00056-9

Umans, B., Battle, A., Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? *Trends in Genetics*, 37(2)109-124. https://doi.org/10.1016/j.tig.2020.08.009

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics,* 101(1),5-22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genetics*, 16,4. https://doi.org/10.1371/journal.pgen.1008720

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

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51,1339–1348. https://doi.org/10.1038/s41588-019-0481-0

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447,661–678. https://doi.org/10.1038/nature05911

Wu, Y.-L., Lin, Z.-J., Li, C.-C., Lin, X., Shan, S.-K., Guo, B., Zheng, M.-H., Li, F., Yuan, L.-Q., Li, Z.-h. (2023). Epigenetic regulation in metabolic diseases: mechanisms and advances in clinical study. *Signal Transduction and Targeted Therapy,* 8,98. https://doi.org/10.1038/s41392-023-01333-7

Yuan, S., Bäck, M., Bruzelius, M., Mason, A.M., Burgess, S., Larsson, S. (2019). Plasma Phospholipid Fatty Acids, FADS1 and Risk of 15 Cardiovascular Diseases: A Mendelian Randomisation Study. *Nutrients*, 11(12),3001. https://doi.org/10.3390/nu11123001.

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*, 50,1335–1341. https://doi.org/10.1038/s41588-018-0184-y