UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Kateryna Peikova

Genetic effects on gene expression across cell types, tissues and biological contexts

Master's Thesis (30 ECTS)

Supervisor: Kaur Alasoo, PhD

Tartu 2020

Genetic effects on gene expression across cell types, tissues and biological contexts

Abstract:
The human body consists of many tissues (e.g. brain, blood, skin or fat) which in turn are made of many different component cell types (e.g. neurons, monocytes, fibroblasts or adipocytes). The identities and functions of different cell types are defined by the different sets of genes that they express. Similarly, genetic differences between individuals can alter gene expression levels and in turn influence one's risk of developing various complex diseases. Specific genetic variants associated with gene expression levels are referred to as expression quantitative trait loci (eQTLs). While multiple studies have demonstrated that the eQTL effect sizes vary between cell types and tissues, the magnitude of this variation has remained unclear. Although small studies focusing on purified cell types have generally reported large differences in eQTL effect sizes between cell types, the largest analysis of gene expression across 49 human tissues by the GTEx project found a high level of eQTL sharing between tissues. Furthermore, different analytical choices have made it difficult to compare results from different studies. Fortunately, the eQTL Catalogue project has recently released uniformly processed eQTL summary statistics from 19 individual studies. In this thesis, we used the eQTL Catalogue summary statistics to estimate the sharing of eQTLs across up to 46 individual cell types and tissues. Consistent with previous reports, we find high levels of eQTL sharing between tissues. In contrast, there was much less sharing between purified cell types. This suggests that high tissue-level sharing is driven by sharing of cell types between tissues and averaging of effect sizes across many different component cell types. This was further supported by factor analysis, which revealed that eQTL effect sizes in tissues were comprised of multiple shared and cell-type-specific components. Finally we tried use the cell-type-specific eQTL components to interpret complex disease associations, but did not find compelling evidence for specific enrichments. Our results indicate that much larger datasets from purified cell types are needed to completely interpret eQTL signals detected in complex tissues.

Keywords:
eQTL, GWAS, matrix factorization, gene expression

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Geneetiliste variantide mõju geeniekspressioonile erinevates rakutüüpides, kudedes ja bioloogilistes tingimustes

Lühikokkuvõte: Inimese keha koosneb mitmetest kudedest (näiteks aju, veri, nahk või rasvkude) ning koed omakorda koosnevad paljudest erinevatest rakutüüpidest (näiteks neuronid, monotsüüdid, fibroblastid või rasvarakud). Iga rakutüübi unikaalse olemuse ja funktsiooni määrab ära nende geenide hulk, mis on just selles rakutüübis ekspresseeritud. Samuti mõjutavad geenide ekspressiooni indiviidide vahelised geneetilised erinevused mis võivad omakorda suurendada mõne komplekshaiguse tekkeriski. Geeniekpsressiooniga seotud geneetilisi variante kutsutakse geenekspressiooni kvantitatiivse tunnuse lookusteks (ingl k *expression quantitative trait locus* - eQTL). Mitmed uuringud on küll näidanud, et eQTLide efektisuurused on erinevates rakutüüpides ja kudedes erinevad, aga sellise variatsiooni suurus ja olulisus on veel ebaselge. Kuigi eraldatud rakutüüpe vaadelnud väiksed uuringud on tuvastanud eQTLide efektisuuruste vahel suuri erinevusi, siis suurim 49 erinevat kude hõlmanud GTEx uuring leidis, et eQTLide efektisuurused on paljudes kudedes väga sarnased. Lisaks teeb erinevate uuringute tulemuste võrdlemise keeruliseks erinevused kasutatud metodoloogias. Õnneks on *eQTL Catalogue* projekt teinud hiljuti vabalt kättesaadavaks ühiselt analüüsitud kokkuvõttestatistikud 19 erinevast uuringust. Käesolevas töös kasutasime me *eQTL Catalogue* kokkuvõttestatistikuid, et hinnata, mil määral varieeruvad eQTLide efektisuurused kuni 46 erineva koe ja rakutüübi vahel. Kooskõlas varasema GTEx uuringiga tuvastasime me, et eQTLide efektisuurused erievates kudedes on üsna sarnased. Samas aga olid erinevused eraldatud rakutüüpide vahel palju suuremad. Need tulemused viitavad, et suurt rakutüüpide vahelist sarnasust põhjustavad jagatud rakutüübid, mis on olemas mitmes erinevas koes, ning efektisuuruste keskmistamine üle paljude koe koosseisus olevate rakutüüpide. Seda järeldust toetab ka faktoranalüüs, mis tuvastas, et eQTLide efektisuurused kudedes koosnevad mitmest rakutüübispetsiifilisest ja jagatud komponendist. Meie tulemused näitavad, et keerukates kudedes leiduvate seoste tõlgendamiseks on vaja palju suuremaid rakutüübispetsiifilisi andmestikke.

Võtmesõnad:
eQTL, GWAS, faktoranalüüs, geeniekspressioon

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Contents

Introduction

Underlying biology

The morphological and functional diversity of organisms are mostly defined by the genomic information stored in the the nucleus in the form of deoxyribonucleic acid (DNA). DNA encodes genes, and their expression happens in two main steps. First, DNA is copied to ribonucleic acid (RNA); this process is called transcription. The next step is a translation when proteins are synthesized from RNA. Proteins are the main components in the metabolic reactions of the cell. The DNA sequence contains parts that code proteins, as well as non-coding regulatory sites. The regulatory regions control where (which cell types or tissues) and when (developmental stage, response to a stimulus, disease progression, etc.) the genes are expressed as well as how strongly they are expressed (how many RNA molecules are transcribed from a given gene). The enhancers and promoters are the regions where transcription factor proteins are bound, and further, the gene expression is regulated. Gene expression defines metabolism, molecular functionality and, in general, type of a cell. The change in gene expression mechanisms leads to a change in phenotype and may cause diseases.

People have many differences in the DNA code due to mutations and recombinations. These regions in the DNA that differ between people are called *genetic variants*. The dbSNP database of genetic variants build 154 [5] contains more than 730 million of unique variations among humans. These variations are fundamental factors that make people physiologically different. Genetic variants can be used for ancestry testing or association tests with traits or diseases.

Genome-wide association studies

Genome-wide association studies (GWAS) are the most common studies to associate genomic variants with particular traits for almost 20 years. Discovering genetic regions that are linked to a disease can help to understand biological mechanisms (e.g. what proteins are involved in a disease) behind it. However, GWAS studies have some limitations. A GWAS usually results in regions that contain multiple variants (that are in *linkage disequilibrium* with each other) or genes [1]. Very often multiple variants hold a strong association with a trait because of the high correlation between these variants. It is not easy to find out which of them is a true causal variant. Additionally, if the variant region spans several genes, it is hard to identify which of them is actually affected by the mutation. For instance, a recent GWAS for breast cancer identified 46 risk variants, 21 of which had two or more predicted target genes [6]. Alternatively, a gene can be located further from the tested variant region, for example, FTO locus associated with body mass index, where the causal genes IRX3/IRX5 are far away [4]. Identifying the true causal gene is crucial for studying the genetic mechanisms of a disease.

Expression quantitative trait loci

Genetic variants can also be associated with gene expression. Variant regions that are significantly associated with the expression level of a specific gene are called *expression quantitative trait loci* (eQTL). It is said that eQTL shows a variant effect on a particular gene. The eQTL explains a fraction of gene expression variation. Formally, it is a statistical association between a genomic region and the expression level of a gene (see Figure 1). Expression level indicates how many mRNA (messenger RNA) molecules are transcribed from a region. The association is usually measured as a linear regression, where eQTL effect size is a regression slope β. There are two main types of eQTLs: *Cis*- and *Trans*-. *Cis*-eQTLs are variants that are located within $+/-1$ megabasepair from the gene promoter (see Figure 1). *Trans*-eQTLs can be situated in larger distances. EQTL is a significant term to study how genetic variation affects phenotypes and leads to diseases. Even though that DNA molecule is the same in all cells of the organism, cell types differ in their structure and functions. Gene expression is a mechanism of phenotype variability. Gene expression level can also change to various biological contexts. Many eQTLs can be discovered only in particular cell types and tissues or under a stimulus, that makes them specific. Combining eQTL data with GWAS associated variants can help to identify which genes are regulated by them. Later these genes can be targets for treatment research. Discovering shared and condition-specific eQTLs can help to find rare gene expression associations that are specific to the cell type or contexts where these variants are active, but underrepresented or missed in tissues.

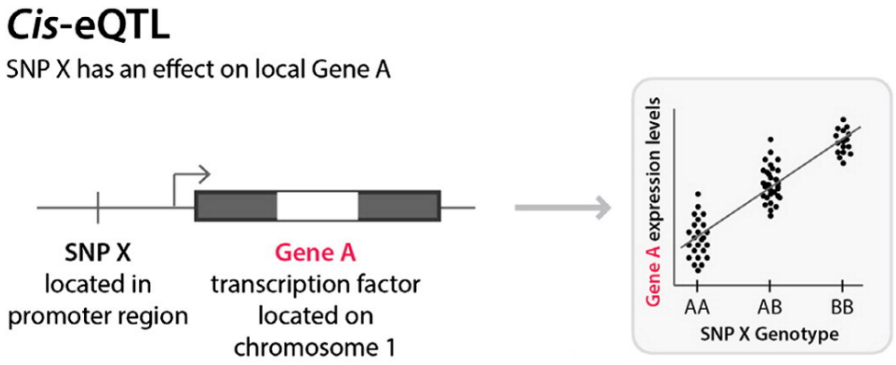

Figure 1. Cis-eQTL association [31]. Single-nucleotide polymorphism (SNP) variant is closely located to the gene A. The expression level of the gene A is strongly associated with variant genotype, therefore the gene-variant pair forms an cis-eQTL.

eQTL summary statistics

Often the eQTL data for the whole genome is given in the form of association summary statistics for each significant variant-gene pair. There are many methods on how to discover eQTLs. Many of them are improved and optimised versions of linear regression approach (see Figure 2). The association between variant and gene expression is assumed to be linear. The expression samples are grouped by the allele (particular version of a mutation). If a gene is higher expressed in one of the groups, the variant is associated with a given gene and expected to influence the expression of it. The effect size is assumed to be a regression slope β. The statistical test is performed, and adjusted p-value for every pair is provided as well. Additionally, some methods can return standard errors. The eQTL mapping is done for a set of genes across the whole genome. Finally, the summary statistics include effect size, p-value and standard error for one cell type or tissue. Recently several datasets were developed to combine summary statistics over multiple tissues and cell types.

GTEx and eQTL Catalogue

Availability of a dataset that contains eQTL summary statistics across many tissues and cell types is crucial for understanding general and specific patterns of gene expression regulation. The Genotype-Tissue Expression (GTEx) project [16] was developed to study eQTLs across many tissues and link them with GWAS disease studies. Despite the large project scale, GTEX is missing many tissues as well as pure cell types. The new dataset eQTL Catalogue was developed recently [14]. The eQTL Catalogue contains a large number of cell types and tissues that come from a wide number of studies with different sequencing methods, sample sizes and conditions. An additional advantage is uniformly processed data that contains summary statistics for the same set of genes across all of the tissues, as well as, the consistent set of variants, that makes it convenient to analyse. For every cell type or tissue, eQTL Catalogue has a summary statistics that contains eQTL effect sizes, standard errors, p-values and additional metadata. The main differences between GTEx and eQTL Catalogue are sets of tissues and cell types, data processing and sequencing techniques. Tissues make up most of the assays in the GTEx. The eQTL Catalogue, on the other hand, has purified cell types, that are of high interest for similarity analysis. What is more, the GTEx dataset contains only healthy normal tissues, eQTL Catalogue has cell types that were stimulated (and hence may contain immune response). While all tissues in GTEx were profiled using RNA sequencing (RNA-seq), eQTL Catalogue has both microarray and RNA-seq sequenced data.

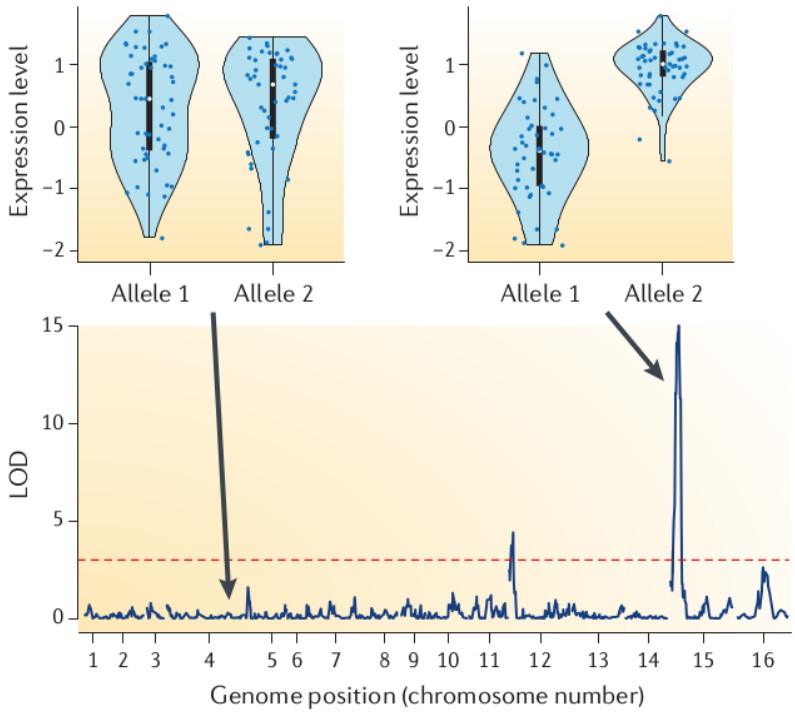

Figure 2. Genome scan for mRNA levels of the yeast TPO1 gene in a cross between two yeast strains [1]. Two examples showing the expression distribution between alleles. The regression has a larger slope (effect size) in the second region and therefore the association is more significant (the logarithm of the odds (LOD) is used to measure the association strength).

Fine-mapping

Choosing causal variants is essential for studying the genetic mechanisms behind diseases. Identifying the true causal variant, however, is not easy because genetic variants are often inherited together causing correlations between their genotype values (in technical terms, the affected genetic variants are said to be in high *linkage disequilibrium* (LD) with each other). Reducing the number of analysed variants to sufficient enough to capture the strongest effects across tissues and cell types and also removing the high degree correlation is critical. However, this is a tricky task due to variants being in LD with the neighbouring variants.

As closely located variants can be in high LD and it is difficult to define which one of them is a causal variant, we used fine-mapped eQTLs derived with the "Sum of Single Effects" (SuSiE) model [30] from the eQTL Catalogue. The purpose of this approach is to reduce the number of variants of interest to small sets (credible sets)

of highly correlated. A credible set is a subset of variants (see Figure 3) that includes at least one causal effect variant (non-zero regression coefficient) for a gene with the probability α (coverage of the credible set) or larger. Credible sets ($\alpha = 0.95$) were identified for each cell type separately. Fine-mapped results can miss some genes across tissues. Additionally, one gene can have several credible sets. Counts of unique genes and credible sets differ across tissue (see Figure 4). The number of distinct genes differs between 350 and 6500 with the number of unique credible sets slightly exceeding it. These reasons introduce a challenge to collectively analysing different cell types and tissues.

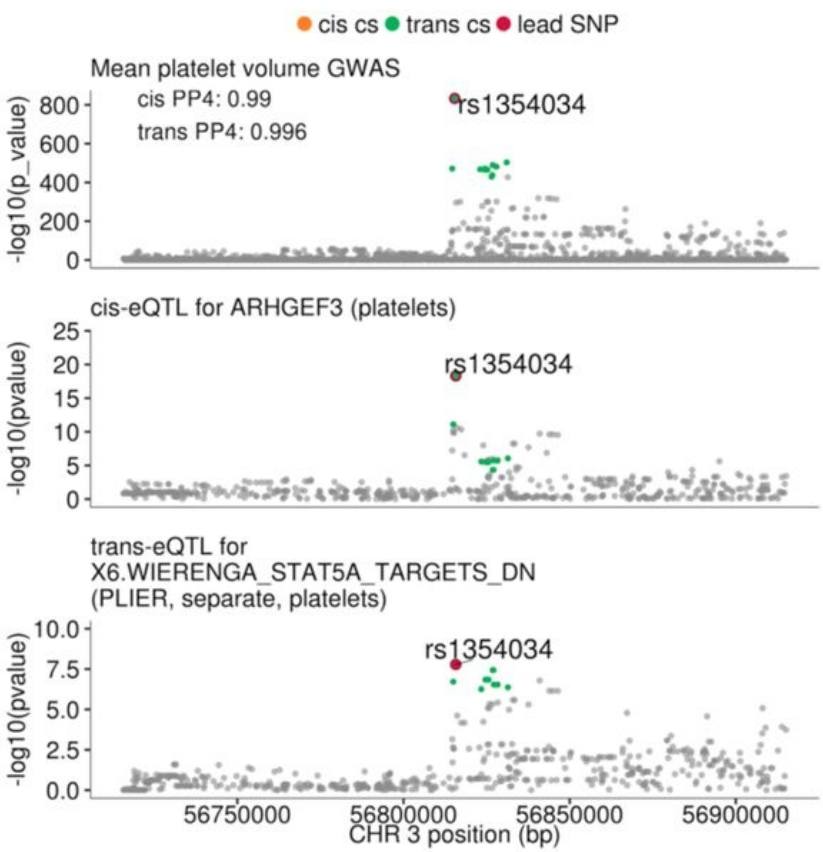

Figure 3. Colocalisation between GWAS signal for mean platelet volume [15]. Illustration of linkage disequilibrium between variants. Many variants that have a high log p-value are correlated with each other. The variants from credible sets are colored green and yellow. There are many variants near a lead variant that are significant. The fine-mapping usually results in small subset of correlated variants.

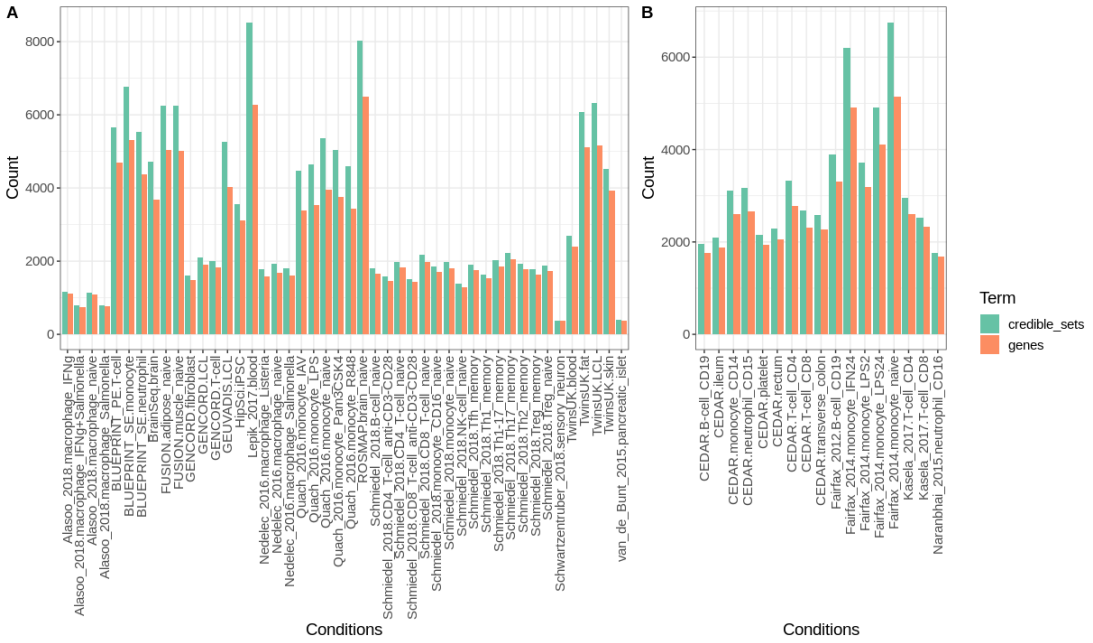

Figure 4. Fine-mapped genes and credible set counts for microarray and RNA-seq datasets. Some of the genes have multiple credible sets, the set of fine-mapped genes is not consistent across all cell types and tissues.

Methods for jointly analysing summary statistics from multiple tissues.

Several methods for estimating similarity and learning eQTL patterns between tissues were applied on the GTEx dataset. Generally, they try to find what effects are similar or shared among the tissues and which ones are specific to a tissue or cell type.

Meta-Tissue In the original GTEx paper [8], they searched for eQTLs shared among tissues using Meta-Tissue approach [24]. Meta-Tissue uses mixed-effects models of gene expression across multiple tissues and applies meta-analysis to combine results to discover eQTLs. The advantage of the method is taking into account the expression correlation between tissues. This approach helps to discover eQTLs that are shared across tissues but might be lost in the single-tissue analysis due to lack of statistical power. They estimate the presence of an effect in a given tissue using m-value. This statistics is a posterior probability that effect exists in a study. In the [8] the eQTL similarity between tissues was estimated through pairwise-tissue Spearman's correlation of posterior effects obtained with Meta-tissue. Group of brain tissues had the highest correlation coefficient. The large similarity (on average Spearman's 0.8 coefficient) was

also found in smaller groups of arterial and muscle-heart tissues. Authors also measured the sharing of eQTLs with Jaccard index. They discovered a trend where the tissues with higher effects correlation have more shared eQTLs in common.

Multivariate adaptive shrinkage (Mash) Another model called Multivariate adaptive shrinkage (Mash) [28] was developed to assess the eQTL effects sharing between GTEx tissues. It is a multivariate Gaussian mixture model. It is supposed to improve the eQTL effect estimates based on the effect sizes distributions across the cell types and tissues. The Mash provides a new method to measure the pairwise sharing between tissues.

To analyse tissues from the GTEx dataset eQTL effect sizes and standard errors (were provided as weights to the model) were used with Mash. They analysed cis-eQTLs of 16k genes from 44 tissues. 76% of single nucleotide polymorphisms (SNPs) turned out to be significant in at least one tissue. The sharing scores resulted in several groups of similar tissues: two brain groups, skin, adipose tissues, heart, artery and gastrointestinal groups. In the brain tissues sharing by sign is higher than 96% and by magnitude is 76%. Across all tissues, sharing by sign is at least 85% and 36% by magnitude on average.

Semi-nonnegative sparse matrix factorization Weighted semi-nonnegative sparse matrix factorization (sn-spMF) [10] was developed to decompose effects across multiple conditions (i.e. tissues) to latent factors. An interpretation of resulting factors is that each eQTL has a loading to them. Each tissue has some loadings to factors as well, resulting in that factors become either tissue-specific or shared. The sn-spMF was applied to the GTEx v8 data containing effect estimates of 49 human tissues. The final matrix contains 24 factors. One of them appeared universal for all tissues. Other factors are rather tissue-specific. 8 factors belong to multiple tissues, and 14 are single-tissue-specific. 41 out of 49 tissues have loadings on at least one tissue-specific factor. In other words, 8 tissues have loadings only on the universal factor. Authors mentioned that these tissues have a relatively small sample size. 20% of eQTLs have universal factor weights being nonzero. The number of eQTLs assigned to other factors varies from 1.5% to 8.1%. 53% of gene-variants pairs have significant loadings on at least one tissue-specific factor.

Aims and goals

In this project, we aim to assess developed methods used with GTEx dataset on the eQTL Catalogue data and discover results brought by the new dataset, that contains many purified cell types. The eQTL similarity between cell types, tissues and particular conditions is essential to evaluate how similar the same tissues are (or cell types and biological contexts) when coming from different studies or how close the eQTL effects are in the stimulated and naive (unstimulated) cell types. Moreover, we assessed how similar tissues between each other are compared to pure cell types. We tried to answer

these questions by simple statistical approaches as correlation coefficients, as well as applying a more complex model Mash [28]. We discuss it in the first chapter.

Additionally, we want to study how genetic variations that regulate gene expression can be specific to a cell type, tissue or condition, or shared among them (and how many eQTLs are shared). We try to discover underlying patterns in genetic effects across tissues and cell types. We describe the clustering eQTLs by their effect and matrix-factorization-based method sn-spMF [10] to capture latent factors in the regulation of gene expression among multiple tissues and cell types in the second chapter. The association of eQTL loci with GWAS traits is another question approached in this thesis and reported in the third chapter. We performed GWAS disease enrichment for eQTL regions belonging to particular latent factors discovered with matrix factorization approach.

1 Assessing eQTL sharing between the datasets in the eQTL Catalogue

Different cell types and tissues can share some eQTL effects and have specific ones. It is not a trivial task to measure how similar cell types or tissues are between each other. We tried to examine if biologically related cell types share more eQTLs with each other compared to other cell types and tissues. Furthermore, we compared if sharing between different tissues are stronger than between different cell types. Additionally, we investigated if the same cell types coming from different studies have a larger eQTL sharing degree.

1.1 Data

The gene expression data for the analysis was obtained from the eQTL Catalogue [14]. The project includes 14 RNA-seq (RNA sequencing) datasets and 5 microarray datasets with 46 and 17 cell types, tissues or biological contexts respectively (starting from here, we will use these terms interchangeably, as technically they are simply conditions in the context of our analysis). Microarray and RNA-seq are two distinct technologies for measuring gene expression. The downstream methods in this thesis were applied separately for microarray and RNA-seq data, mostly due to the following reasons. The gene sets are inconsistent between datasets. Due to technical differences, the scale of effect and measurement noise bias may vary between datasets as well. However, it might be possible in the future to combine microarray and RNA-seq data in a single analysis using additional preprocessing and meta-analysis.

The eQTL Catalogue summary statistics contain information about effect sizes (the strength of association) of genetic variants on expression levels of various genes. The effect sizes are measured as B's, and standard errors (SE) with p-values are provided. The eQTL mapping (discovery of eQTLs) was done for each tissue, cell type or condition separately. Only local cis-eQTL effects were used in the further analysis.

1.2 Methods

The overall analysis workflow consists of several parts (see Figure 5) and combines data from multiple sources. The fine-mapped variants and eQTL effects were obtained from the eQTL Catalogue. Using fine-mapped variants we reduced amount of effects analysed in this work. Then, selected effects were provided into Mash model to estimate similarity and sn-spMF method (see Chapter 2) to decompose effects to factor matrix. Later the variant call data from 1000 Genome project phase 3 [25] were annotated with decomposed factors and diseases were enriched using *ldsc* method (see Chapter 3).

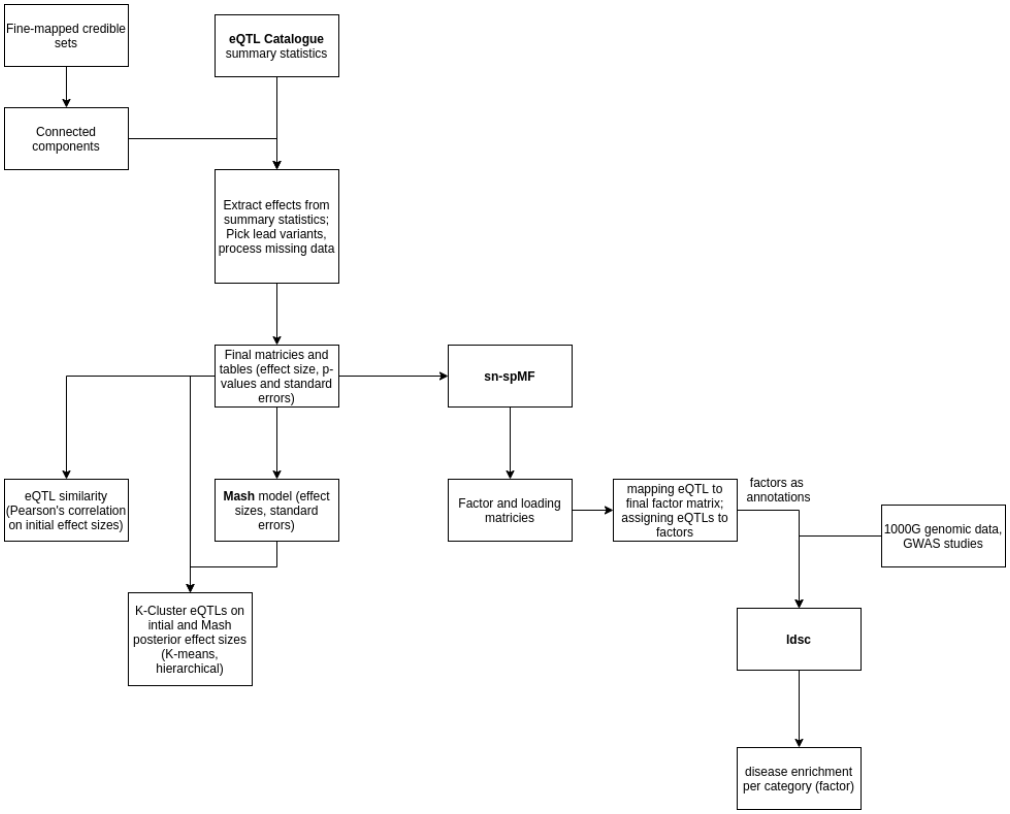

Figure 5. The analysis workflow.

Most of the approaches that aim to estimate condition-specific effects and patterns of sharing require consistent sets of variant-gene pairs across all of the conditions. There are some methods such as [10] that allow missing values as well. In our case, these conditions are cell types, tissues as well as those together with biological contexts. The pick variant, a variant with the highest posterior inclusion probability (pip-value), is not always the same across all conditions. To reduce the number of variants and diminish the amount of correlation among effects, we came up with connected components strategy.

1.2.1 Connected components

The fine-mapped variants can be different for a given gene across all conditions (see Figure 6A). To aggregate credible sets across tissues and capture lead effects in tissues for each gene, we built connected components out of fine-mapped credible sets. We selected those variants that appear in all tissues. As variant sets differ between conditions, for each gene, we combined the credible sets from all tissues into a graph and identified connected components in it. Each credible set in such interpretation is a vertex. Connected

components of credible sets can be informally defined as follows: if two credible sets share at least one variant we consider them connected; in other words, there is an edge between them.

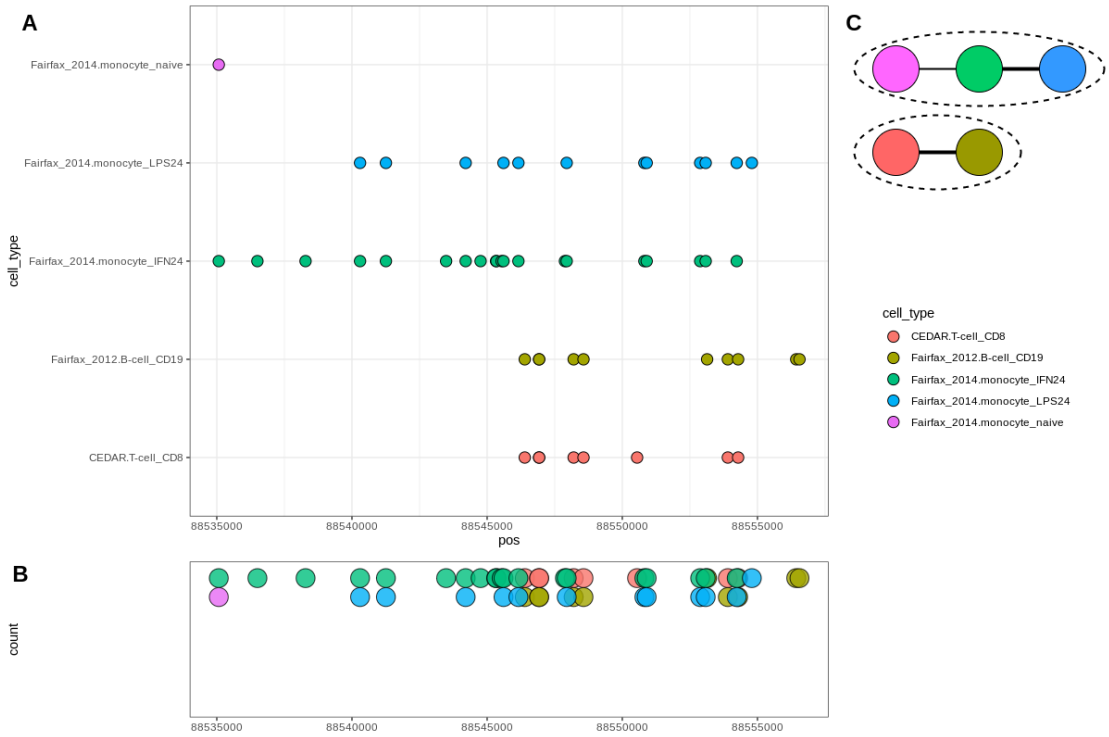

Figure 6. Connected components strategy. **A**. Fine-mapped variants across tissues. There is no variant shared across all of the cell types. **B** Shared variants in credible sets. **C** Resulting two connected components.

The number of connected components, as well as its size, tend to be larger with a higher number of cell types. Additionally, some of the genes can have several connected components (see Figure 6C). We did not include variants from credible sets consisting of more than 50 variants and with Z-score smaller than 3 to avoid data ambiguity and remove unconfident effects. For each connected component, we obtained summary statistics from the eQTL catalogue. Ideally, each connected component represents a distinct genetic effect. Probably due to polygenicity (one gene may be influenced by multiple variants) and fine-mapping differences between tissues, on average, each gene had two connected components. For later analyses, we tried two strategies: subsetting one variant per gene to reduce the amount of correlation in the effects and preserving multiple effects per gene. Overall, for microarray data, we obtained 24317 connected components of 13840 unique genes. In the RNA-seq dataset, we found 55564 connected

components with 25295 distinct genes.

1.2.2 Dealing with missing data

Due to filtering by minor allele frequency and cleaning in the eQTL Catalogue, not all fine-mapped variants were present in the summary statistics from all cell types and tissues. There are some of the connected components for which none of the credible variants was present in all cell types. Some of the methods we used do not support missing effects across one or more conditions. To include most of the connected components in our analysis, we either used averaged effects data or, if the method allows, left in the values missing. As a *lead variant* (variant representing a connected component), we declared the one that is present in at least 95% of the cell types and tissues. In the case of a variant present in most, but not all of the cell types (95%, 16 for microarray and 44 for RNA-seq) we added variants for missing tissues with averaged values for p-value, standard error and effect size across present tissues. If several such variants were present in eQTL Catalogue summary statistics, we picked a variant with the smallest p-value across all conditions. As a result, we obtained $\approx$ 20k and $\approx$ 17k of lead gene-variant pairs allowing multiple effects per gene for microarray and RNA-seq datasets respectively. As RNA-seq dataset contains more cell types, more variants are missing in some of them, resulting in a smaller number of eQTLs being present across all of the conditions.

1.2.3 Effect sizes correlation

Calculating the correlation coefficient is a baseline approach for comparing how similar the conditions or cell types are. We calculated pairwise Pearson's and Spearman's correlation coefficients of effect sizes across all of the conditions, to evaluate effect similarities between them. Initial effect sizes of lead gene-variants pairs were used to calculate cell type correlations. We applied hierarchical clustering on a pairwise correlation matrix to group tissues. On average Spearman's coefficients were lower than corresponding Pearson's. However, hierarchical clustering resulted in the same tree structure for both Spearman's and Pearson's correlation metrics (see Figures 7 and 8).

Microarray CD4+ and CD8+ T cells from the CEDAR dataset [27] are the most similar cell types based on raw eQTL effect sizes (Pearson's coefficient is 0.84). The next group is stimulated IFN24, LPS24 and LPS2 monocytes from Fairfax dataset [7] with the average Pearson's coefficient of 0.8 (see Figure 7). We expect that genetic effects within the same cell types with different stimuli will be similar. We cut 5 subtrees from hierarchical clustering (see Figure 7). One cluster combines intestinal cell types; three more clusters separate monocytes, neutrophils and T-cells with B-cells. The platelet cell type is not clustered together with any other one. Using Spearman's correlation led to the same clustering results (see Figure 8).

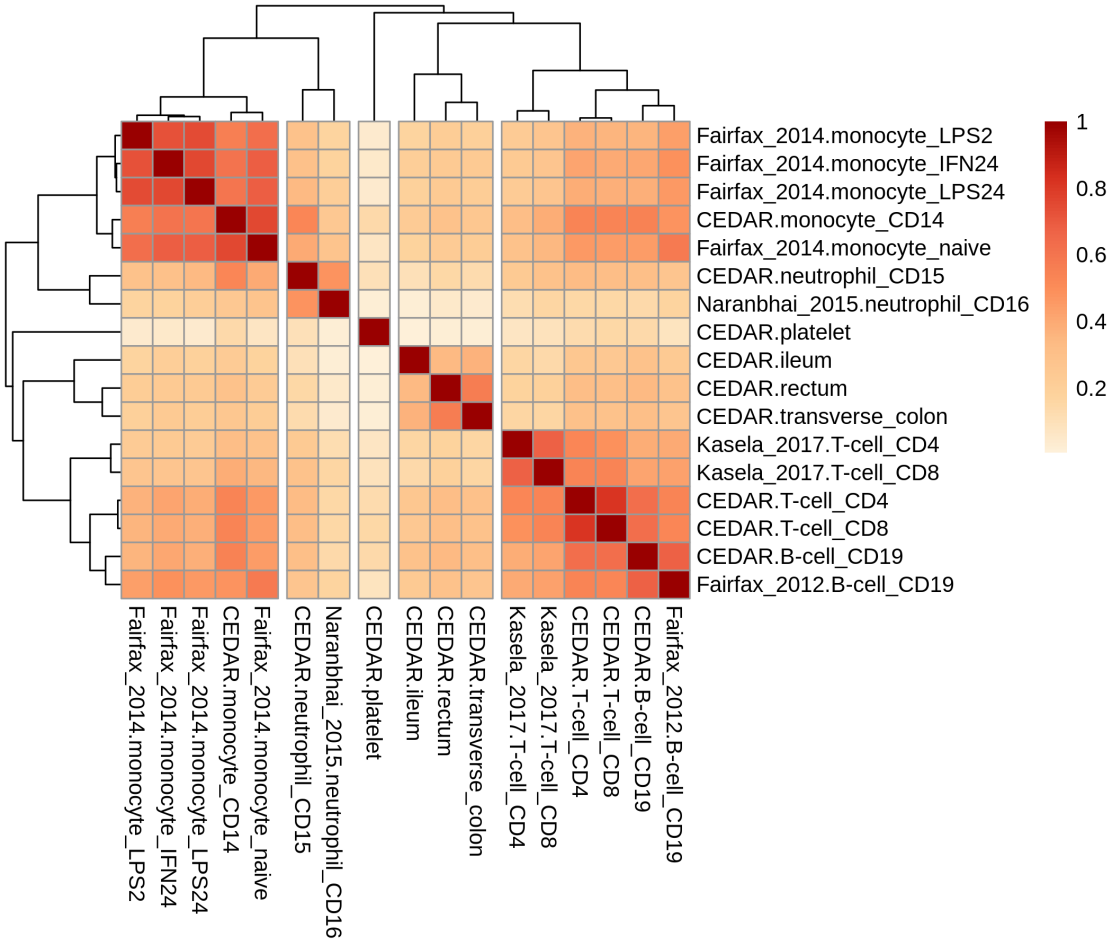

Figure 7. Pearson's correlation eQTL effect sizes for 20k of gene variant pairs. Columns are cut into 5 clusters from the hierarchical clustering.

RNA-seq As RNA-seq dataset has more cell types and tissues, it leads to more complex relativity of cell types. Whole blood tissues are more similar to each other than to all other tissues (see Figure 9). T-cells from the BLUEPRINT study [3] clustered together with multiple T-cell subtypes from the Schmiedel 2018 study [21]. Adipose, fat and skin tissues were clustered together as well. Naive monocytes coming from two different studies formed a separate cluster as well as macrophages. In general, the same or very similar cell types and tissues have relatively high correlation coefficients and are clustered together even though they came from different studies. The tree structure depends on the set of the cell types and tissues and defines the relative relationship between them. As to correlation values, the stimulated and naive monocytes group from the Quach [18] dataset has a pairwise Pearson's coefficient from 0.91 to 0.96. T helper cells from Schmiedel

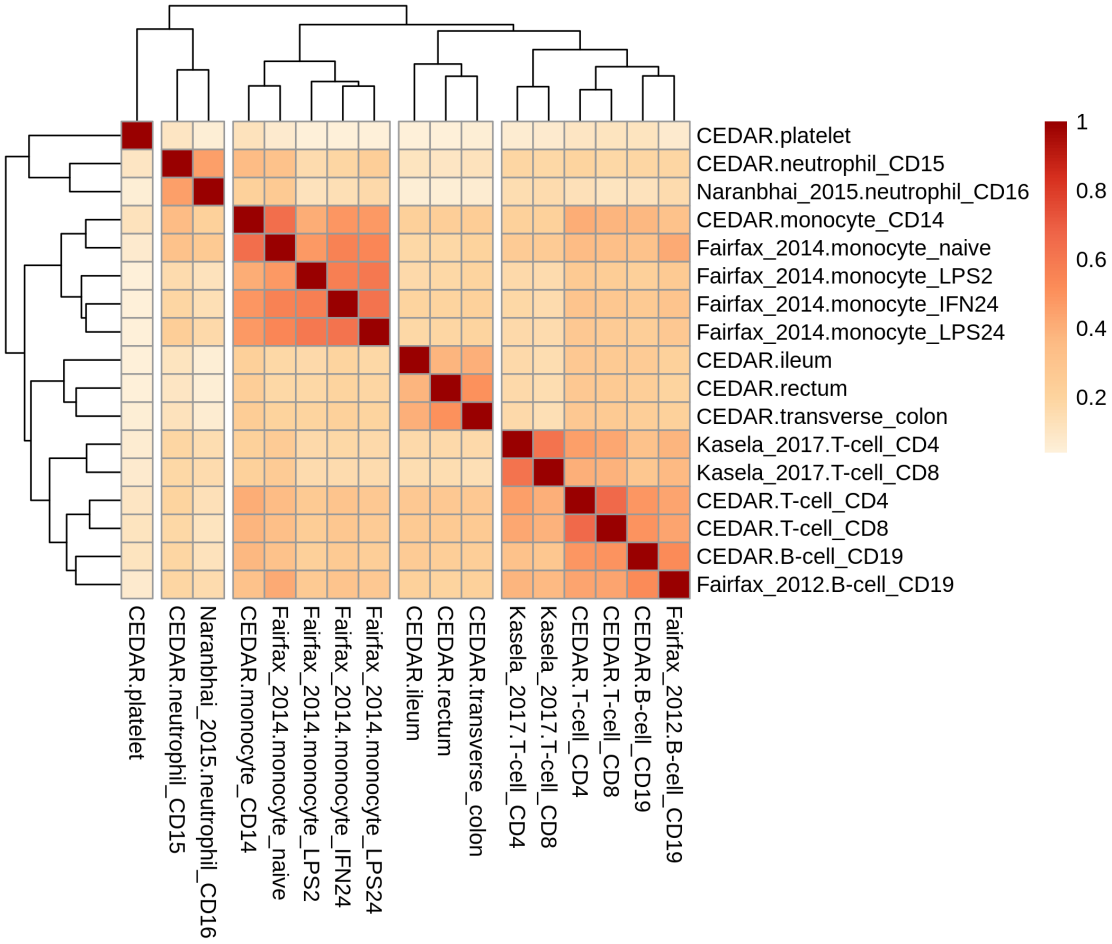

Figure 8. Spearman's correlation heatmap. Columns are cut into 5 clusters from the hierarchical clustering.

2018 have at least 0.9 Pearson's correlation. Compared to brain tissues Spearman's correlation of posterior Meta-Tissue effects (0.8 reported in the GTEx dataset), the RNA-seq brain studies have 0.71 correlation value. The adipose and fat tissues are 0.81 correlated, that is similar to the GTEx results [8].

1.2.4 Multivariate adaptive shrinkage (Mash)

The similarity between eQTL effects can be measured with the Multivariate adaptive shrinkage (Mash) model [28]. The simple pairwise correlation ignores the uncertainty in effect sizes and thus can underestimate the similarity of the cell types. The advantage of the Mash model over simple pairwise correlation is that at first, it improves effect size

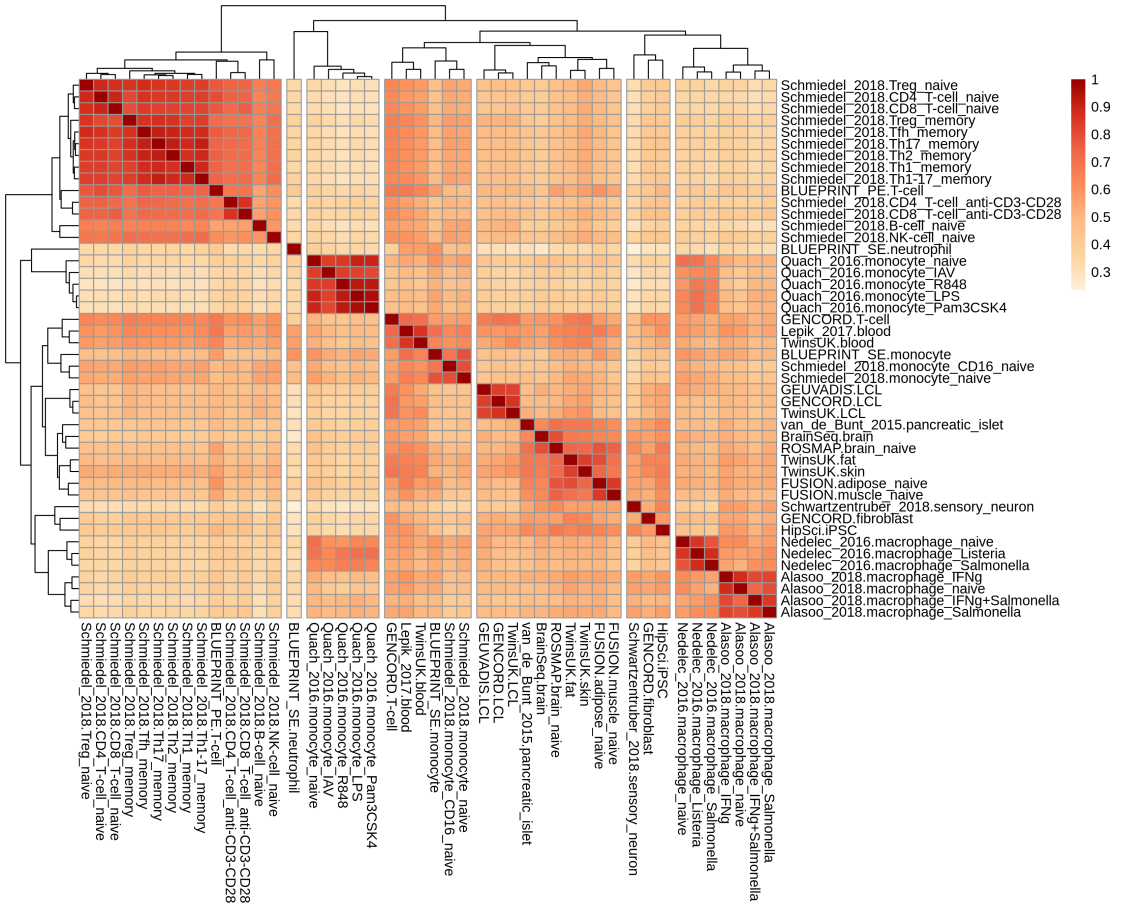

Figure 9. Pearson's correlation heatmap for RNA-seq dataset. Columns are cut into 7 clusters from the hierarchical clustering.

and standard error estimates and then measures the similarity. The similarity between two cell types is measured as the ratio of posterior effects shared by magnitude and sign, which may provide more reliable sharing estimates. We fit the Mash model on the same set of eQTL lead variants that we used in the correlation analysis to estimate eQTL sharing between cell types and tissues.

The input of the Mash is flexible. The summary data can be either Z-scores, effect sizes with p-values or effect sizes with standard errors. Effect estimates and standard errors are preferred as input and are supposed to give the least noisy results. The first step in the Mash method estimates covariance matrices and mixture proportions (see Figure 10). Each covariance matrix represents possible patterns in the data. There are two types of covariance matrices: canonical (simple) and data-driven. To estimate mixture proportions (weight of a scaled covariance matrix) from the candidate covariance

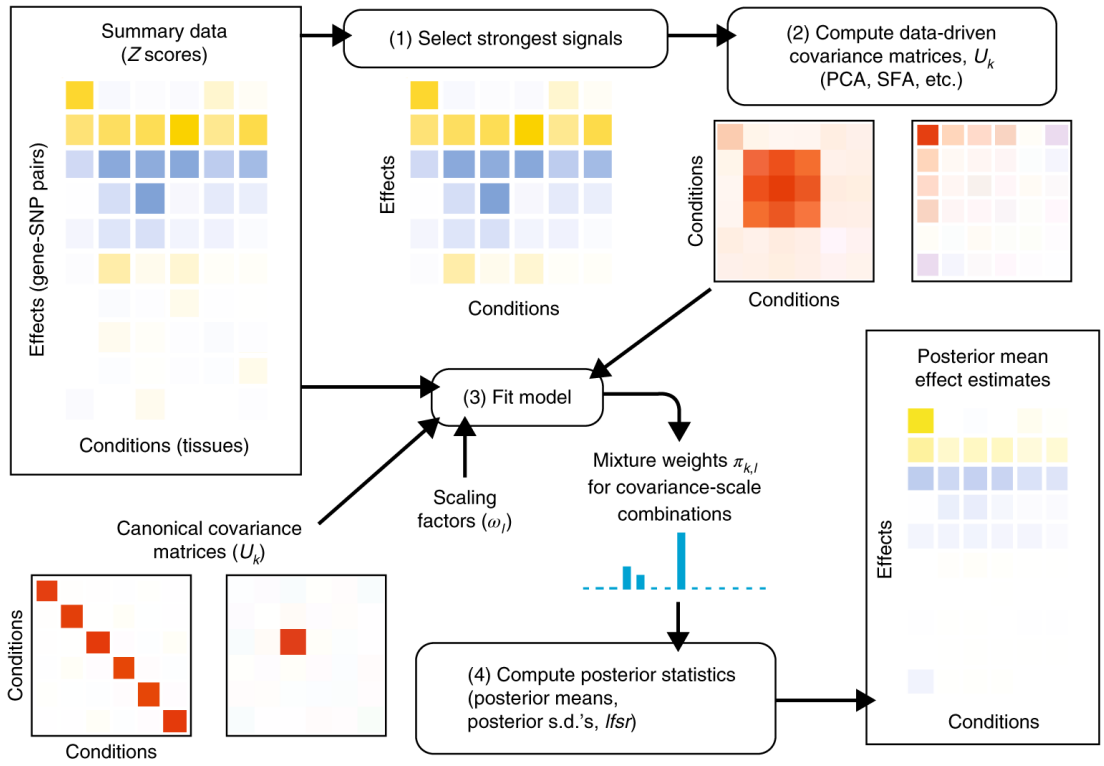

Figure 10. Mash model [28].

matrices and effects they use maximum likelihood, so irrelevant matrices are assigned with small weights because they are not supported by the data. In the second step, they estimate posterior effects through Bayes' theorem from mixture proportions and weighted covariance matrices from the first step. The Mash paper advises applying "condition-by-condition" before fitting the model to select strong signals and build data-driven covariance matrices out of those strong signals. Probably this step is somewhat unnecessary in our case, as the connected components approach that we used to select lead eQTLs ensures that the effects are strong. PCA is performed, and data-driven candidate covariance matrices are built based on the first five principal components. Then extreme deconvolution [2] is applied on PCA covariance matrices to refine estimates. The new covariance matrices result from fitting the data with Gaussian mixtures. The initial effects matrices (betas and standard errors) with covariance matrices are provided to the Mash model. The method fits the mixture model,

$$p(b;\pi,U) = \sum_{k=1}^{K}\sum_{l=1}^{L} \pi_{k,l} N_R(b;0,\omega_l U_k) \tag{1}$$

Where π is a mixture proportion, $N_R(\cdot; \mu, \sigma)$ denotes the multivariate normal density in R dimensions with mean value and variance-covariance matrix Σ. It estimates the relevance of every matrix to the data using maximum likelihood and assigns small weights to those that are not supported by data. We noticed that the fitted model with only data-driven covariance provides less uniform pairwise sharing estimation. The method applies Empirical Bayes approach (ash) [23] (advantages of ash: assumes that distribution is unimodal, uses effect sizes and standard errors for estimation instead of just one p-value or z-score) on effect sizes and standard errors to correct estimations for each lead eQTL in every cell type. The method shrinks the effects based on the prior distribution of the effects and standard errors. The smaller the error is, the more meaningful the effect is, so less shrinkage is applied. As a result, posterior mean effect estimates with standard errors are calculated. Matrix of sharing by the magnitude of eQTL-gene pairs among tissues is computed. Two conditions share an SNP if the effect sign and magnitude (within a factor of 2) is the same for these conditions.

1.3 Results

Microarray We calculated eQTL sharing by magnitude and sign for microarray data and discovered 6 groups of cell types and tissues (see Figure 11). Naive monocytes and monocytes with immune response formed two clusters. There intestinal tissues cluster remains the same as in the simple correlation approach. Remaining clusters are T- and B-cells, neutrophils and platelet cell type. As to specific sharing scores, CEDAR CD4 and CD8 T-cells share 99.5% of eQTLs by magnitude and sign. Corresponding T-cells from the Kasela study [13] share 99.1% of the effects. The next pairs are CEDAR ileum-rectum cell types and LPS stimulated monocytes from the Fairfax study that both have the similarity of 95%. The intestine group of cell types share at least 93% of effects between them. Overall in the microarray set of cell types, the least sharing is between CEDAR platelets and Fairfax monocytes (less than 20%). In general, pure cell types have a small sharing outside of their cluster. Overall, compared with the simple effect size correlation, the main clusters remain almost the same (see Figure 11). The eQTL sharing measure from the Mash model shows higher similarity scores between the same tissues than a simple correlation approach.

RNA-seq The RNA-seq dataset showed higher similarity values for the biologically related groups of cell types and tissues. Macrophages coming from different studies were grouped together (see Figure 12). Naive monocytes, blood tissues and neutrophils were combined into separate cluster if the tree is cut into 7 clusters. The similarity values are quite close to the obtained by the regular correlation coefficient. Th-cells are at least 0.99 similar by the sign and magnitude of the posterior effect. The stimulated R848, Pam3CSK4 and LPS monocytes from the Quach dataset are 99% similar, as well as CD4

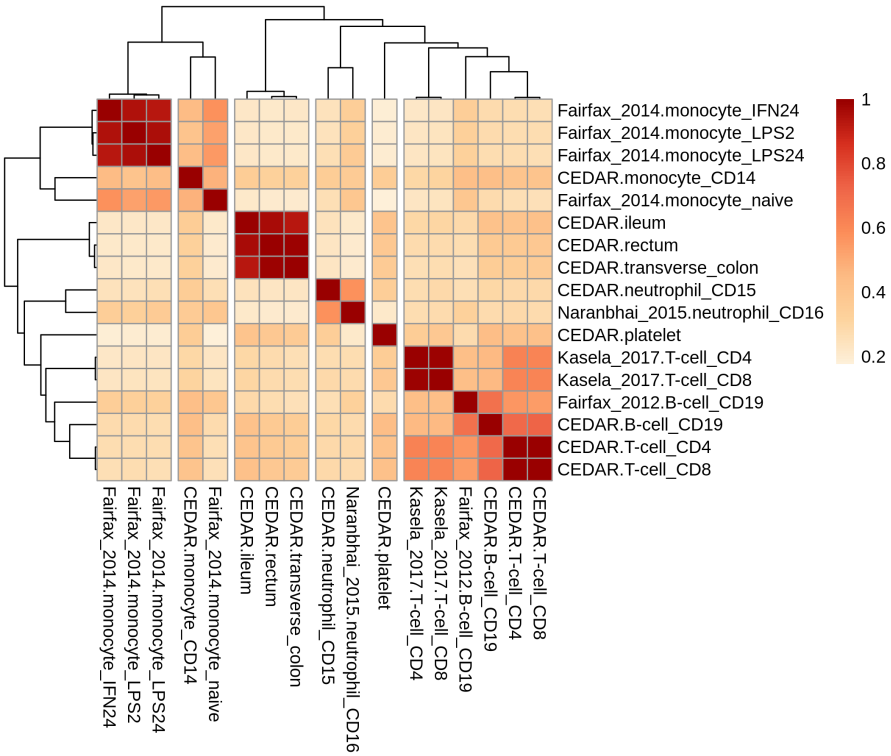

Figure 11. Mash eQTL sharing between microarray cell types and tissues. The model is fit with βs (effect size), SE (standard error) and PCA (Principal Components Analysis) data-driven covariance matrices.

and CD8 T-cells from Schmiedel. Even though the members of clusters are somewhat arbitrary, we can distinguish 6 clusters combining pure cell types and one cluster for tissues. The purified cell types have little sharing degree, while tissues (brain, skin, fat, muscle) are more similar between each other. The high similarity between these tissues can be due to a bulk tissue being rather average of the effect signals coming from cell types comprising a tissue. A tissue consists of multiple cell types in different ratios, where many of eQTL effects can be very low or missed, as a cell type can be underrepresented in a tissue.

One of the advantages of this method is that the model can take sparse input effects (many zero effects) with correlations among non-zero effects. The interpretation of the sparsity is condition-specific effects, while the presence of correlation accounts for shared effects. Also, only the subset of the strong variant gene pairs can be used to estimate the covariance and the remaining eQTLs can be fit to a model to improve effect sizes and estimate cell types sharing. The drawback of the Mash model is that conditions cannot have missing values.

Figure 12. Mash eQTL sharing between RNA-seq cell types and tissues. The model is fit with βs, SE and PCA data-driven covariance matrices.

Mash provides a powerful tool to improve effect sizes and estimate similarities between conditions. Both for microarray and RNA-seq datasets, cell types, tissues and conditions were grouped, preserving true biological similarity between them. Same cell types from different studies were indeed closer to each other by effect size sign and magnitude than to other cell types. However, Mash also captured the study effects by giving cell types coming from the same study high similarity scores.

Technical details All data manipulation was done using the R language. The connected components strategy was implemented with graph R package *igraph*. The querying of eQTL Catalogue summary statistics was done in parallel in the *SLURM* system using *scanTabix* utility from *Rsamtools*.

2 Identifying factors underlying eQTL sharing between datasets

Another approach to discover sharing patterns between cell types is to perform clustering of variant gene pairs. The grouping of eQTLs is useful in performing disease enrichment analysis. We tried to discover clusters of eQTLs that are similar by their effects across cell types and tissues. We applied a more complicated approach than thresholding on effect sizes to assign an eQTL to cell type. At first we tried simple clustering algorithms on raw (and posterior Mash) eQTL effects as well as tried a more advanced matrix factorization method.

2.1 Methods

2.1.1 Clustering eQTLs

The eQTLs may form clusters based on the effect sizes they have across tissues [11]. Clustering can be useful in applying disease enrichment for variants that belong to specific groups of eQTLs. Mainly, we assessed if variant-gene pairs from connected components can form distinct clusters.

To cluster eQTLs, we used both raw effect sizes and posterior means obtained from Mash model. When it comes to the meta-analysis of eQTLs across several conditions, we consider the sign of the effect in specific tissue being random. For every prior eQTL effect, we multiplied the effect size by the sign of the absolute largest effect across all tissues. The intuition behind this strategy is to ensure the strongest effect for the gene-variant pair is always positive. We calculated distances between eQTLs by inverting pairwise correlation coefficients. Based on these distances, we performed hierarchical clustering and cut the hierarchical tree to obtain 10 clusters. Even though clusters, in general, contained similar eQTLs, they remained quite obscure (see Figure 13 and 14). Besides effect noise influencing this approach, we don't capture underlying patterns in tissues and cell type expressions. For instance, microarray data contains too many biologically related cell types (monocytes, neutrophils and platelets that are blood cells make up 8 cell types out 17 from microarray dataset). The possible reasons for weak results are the difference in effect magnitude and the presence of eQTLs with very large effects. Another reason can come with bulk tissues having averaged effect sizes across cell types the tissue contains. One variant-gene pair can have strong effects in multiple cell types and conditions. Additionally, different cell types can be overall similar in genetic effects. Some eQTLs have a large effect in platelet, monocyte and neutrophil cell types. These cells are all components of the blood. Other eQTLs may have large effects in two or less named blood cell types. Because most of the variants have an effect in more than one condition and often the effect estimates are noisy, it becomes difficult to differentiate variant-gene pairs into large groups using simple correlation and clustering

approaches. As to the RNA-seq dataset, the number of conditions is very large. There are too many possible latent components (combinations of strong effects across cell types) that lead to even noisier results.

Overall, we tried several methods: hierarchical clustering on prior and Mash posterior effect estimates, k-means clustering on both prior (see Figures 13 and 14) and posterior effect sizes. All of the approaches gave relatively poor results probably showing that such simple methods are not suitable for strongly noisy data. Additional limitation of mentioned approaches is ignoring of standard errors of the effects. However, Mash model takes into account standard errors in fitting the model and later estimating the posterior effects.

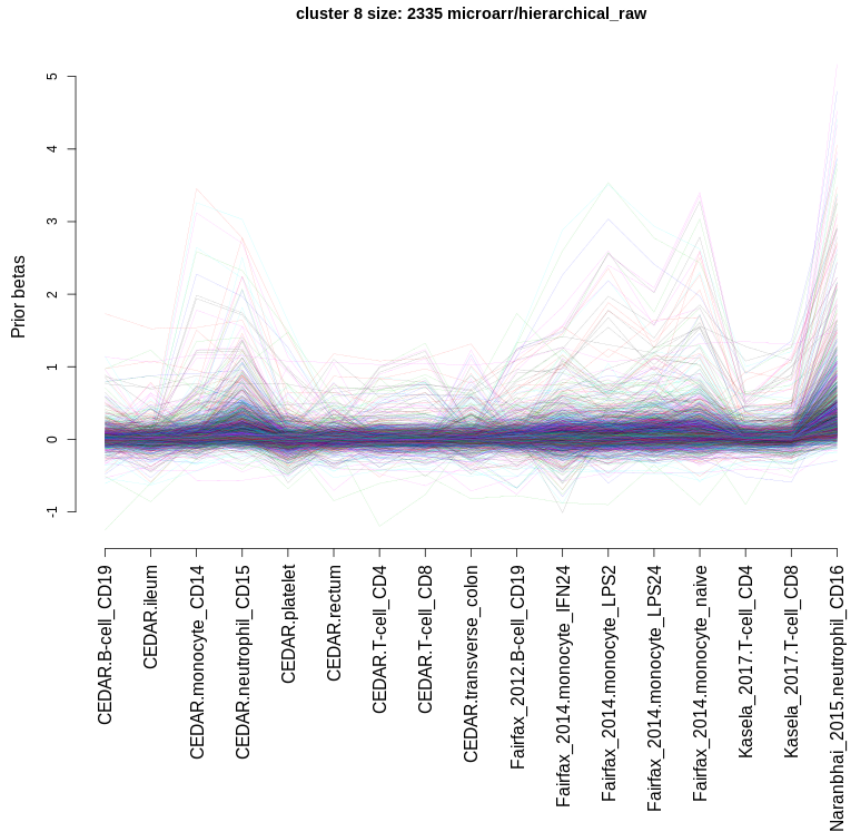

Figure 13. Cluster in microarray eQTLs with high effect sizes for blood cell types: monocyte and neutrophils.

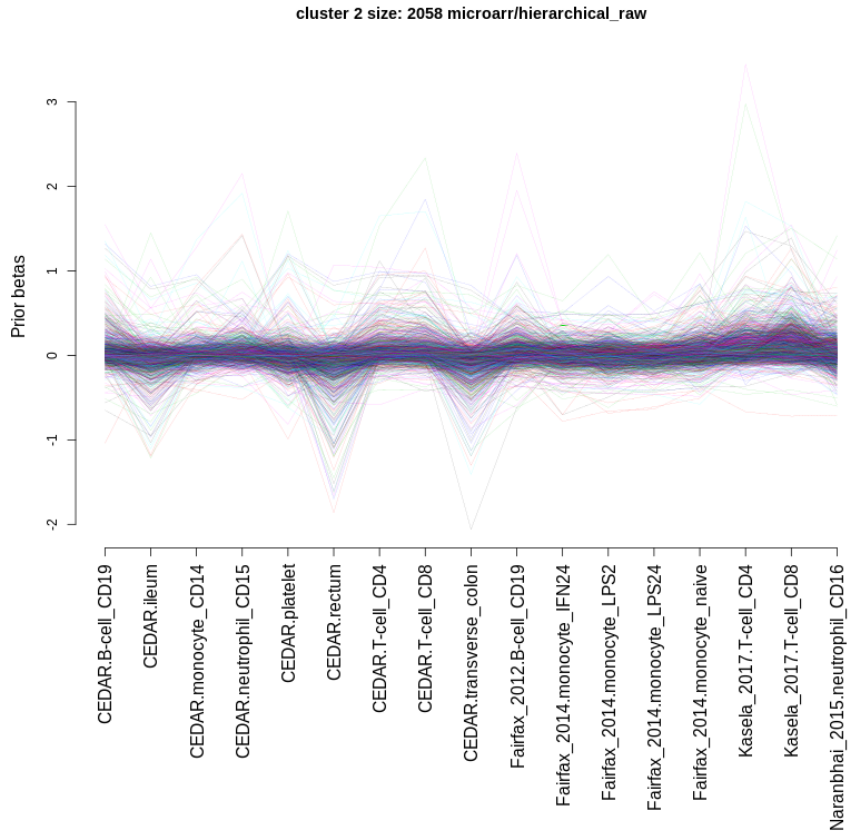

Figure 14. Cluster (K-means clustering) in microarray eQTL effect sizes showing negative effects in intestinal tissues.

2.1.2 Matrix factorization

Another approach to characterise regulatory variation across multiple conditions is a matrix factorization. The key idea of matrix factorisation is to decompose larger matrix of terms into two smaller matrices (see Figure 15) under some restrictions (e.g. non-negativity, sparsity, etc), so their product will give the initial matrix. Such methods allow to discover underlying patterns of eQTLs effects. The advantage of this approach is independence on the number of conditions used in the analysis, as a resulting factor matrix usually has a smaller number of factors than there are conditions. The regularization parameters of matrix factorization influence the resulting matrix shape and sparsity. The factor matrix may show hidden structure of latent factors to which the gene-variants pairs can be assigned based on the loadings.

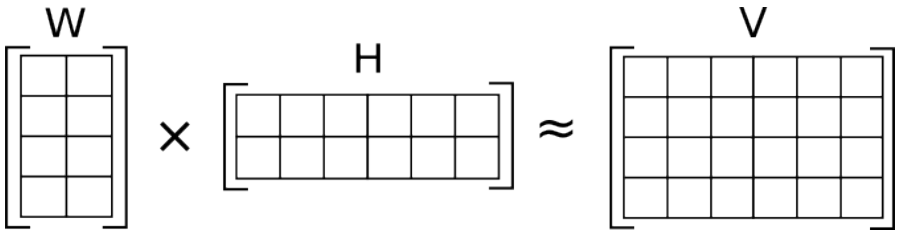

Figure 15. Matrix factorization

2.1.3 Semi-nonnegative sparse matrix factorization

The sn-spMF (weighted semi-nonnegative sparse matrix factorization) model is a type of constrained matrix factorization [10]. The method assumes that eQTL effect sizes across cell types and tissues are a weighted linear combination of factors. Method decomposes eQTL effects across tissues into two matrices: loading matrix $L_{N\times K}$ and factor matrix $F_{T\times K}$. N, T and K are numbers of eQTLs, cell types and latent factors (rank of the decomposition) respectively. We want to learn the factor matrix F, so $X \approx LF^T$ and satisfies the sparsity regularization. Regularization parameter α makes sure that loading matrix L stays sparse and parameter λ ensures the factor matrix F is also sparse. Matrix decomposition allows also to easily map new gene-variant pairs to previously obtained factor matrices using just a weighted linear regression. Factor weights correspond to strength of factor explaining the eQTL effect. As the resulting factor-cell-type matrix is very sparse, the factor weight also agrees with cell types that have a loading on that factor. The model deals with the effect sign being arbitrary through a nonnegative factor matrix. The optimisation includes minimizing weighted squared error and regularization (sparsity in factor and loading matrices). They update factor and loading matrices with alternating least squares with gradient descent. The objective function is as below:

$$\min_{F,L} \frac{1}{2D}\|(X - LF^T) \odot W\|_2^F + \alpha\|L\|_1 + \lambda\|F\|_1 \tag{2}$$

Where F is nonnegative, W is a reciprocal of standard errors.

2.2 Results

2.2.1 Parameters search

The model takes as an input sparsity parameters and initial number of factors. For choosing hyper-parameters we run two-level grid-search as recommended in the sn-spMF tutorial. The first step serves for narrowing down the parameters step, while during the second step the best set of parameters is chosen. We picked parameters that led to higher matrix sparsity and a lower correlation between factors as well as

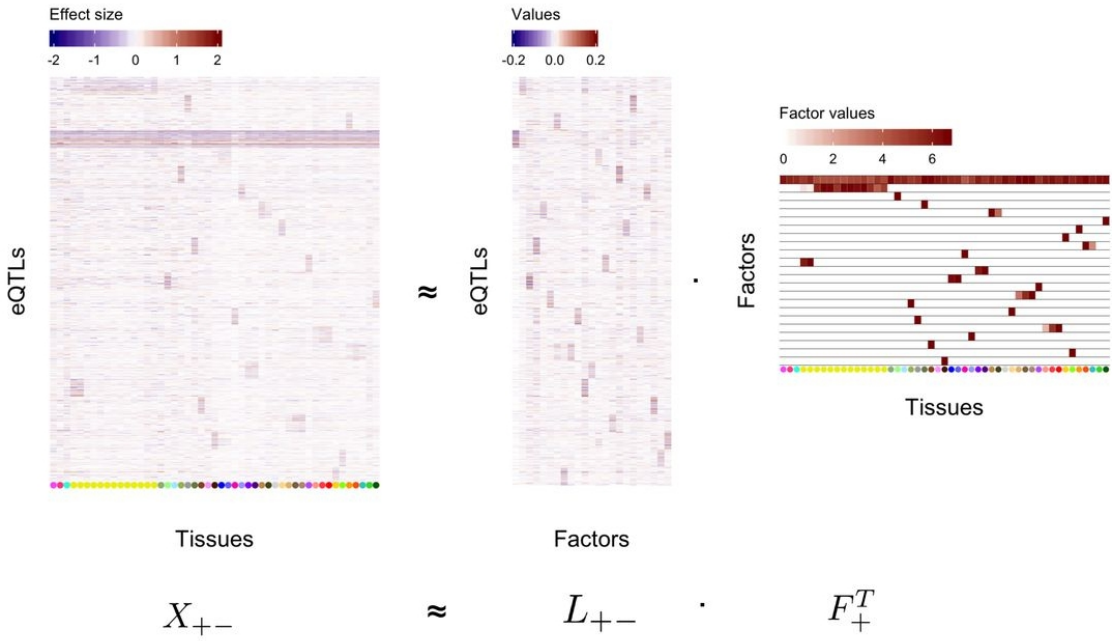

Figure 16. Learning factors underlying eQTL effects from GTEx [10].

higher cophenetic correlation of the factor matrix (stability of the matrix). Cophenetic correlation measures how often two tissues assigned to the same factor out of randomly initialized runs (consensus matrix) of the decomposition. In other words, it measures the rank of the dispersion of consensus matrix. Initial parameter limitation of the model is a starting number of factors cannot be greater than the number of conditions.

For the factorization approach we tried several setups as input data, picking one effect per gene as well as preserving multiple effects for one gene. The results were very similar between these two cases. For the microarray data, we run factorization with 10 and 14..17 factors. The regularization alpha and lambda parameters were in range 500..1400. The final parameters are $K = 15$, $\alpha = 800$ and $\lambda = 300$ that resulted in the factor matrix with 10 latent factors (see Figure 17). Cophenetic coefficient of the resulting matrix is 0.91 and the correlation between factors is 3.24. There is a factor for intestinal cell types, two T-cell factors (representing T-cells from different studies), neutrophil-specific factor, B-cell factor, monocyte-specific factor and two factors specific for immune response in monocytes. Also, there is a shared factor that combines effects across monocytes, T-cells, B-cells and small effects from intestinal cells.

For the RNA-seq dataset we tested 20, 30, 35...40 as the initial number of factors. We tested regularization parameters in range 200..1600. We tested two sets of input

data: lead variants that were present in 95% of the conditions and all lead variants from connected components with missing values. Nevertheless, if input effects included several variants per gene, the resulting factor matrix remained very similar. The data with missing values was accepted as the final version. We performed mapping of credible set variants on the final matrix with parameters $K = 39$, $\alpha = 700$ and $\lambda = 700$. As a result of analysis with mentioned parameters, the matrix converged to the 13 latent factors (see Figure 18). The total correlation between factors was 3.57 and cophenetic coefficient 0.9. The factor matrix distinguishes groups of cells quite well. There are a brain, muscle, skin and fat-specific factors that capture all cell types and tissues of that kind (see Figure 18). Naive monocytes and LCL cell types are represented by separate factors as well. The large T-cell factor also captures small signals from blood cell types (monocytes and neutrophils). There are two large factors T-cell (factor 7) and universal factor (factor 1) that generally contain effects from cell types found in blood. For instance, whole blood tissues (Lepik and TwinsUK) have loadings both on these factors, showing the underlying pattern of bulk blood tissue consisting of many cell types. Another example of Nedelec macrophages having loadings on monocytes-dominated factor and partly universal factor 1.

2.2.2 Mapping variants to factors

After obtaining loadings matrix from the subset of lead gene-variant pairs, we mapped all eQTLs from credible sets (filtered by the size and z-score) to factors. Weighted linear regression. Mapping results into each pair having factor loadings and p-values across latent factors.

The advantage of matrix factorization model compared to Mash method is ability to capture hidden effect patterns across conditions. Hence, Mash doesn't provide the straightforward method to discover cell type-specific or shared eQTLs. The sn-spMF method on the other hand offers a way of assigning the eQTLs to specific factors based on its loadings.

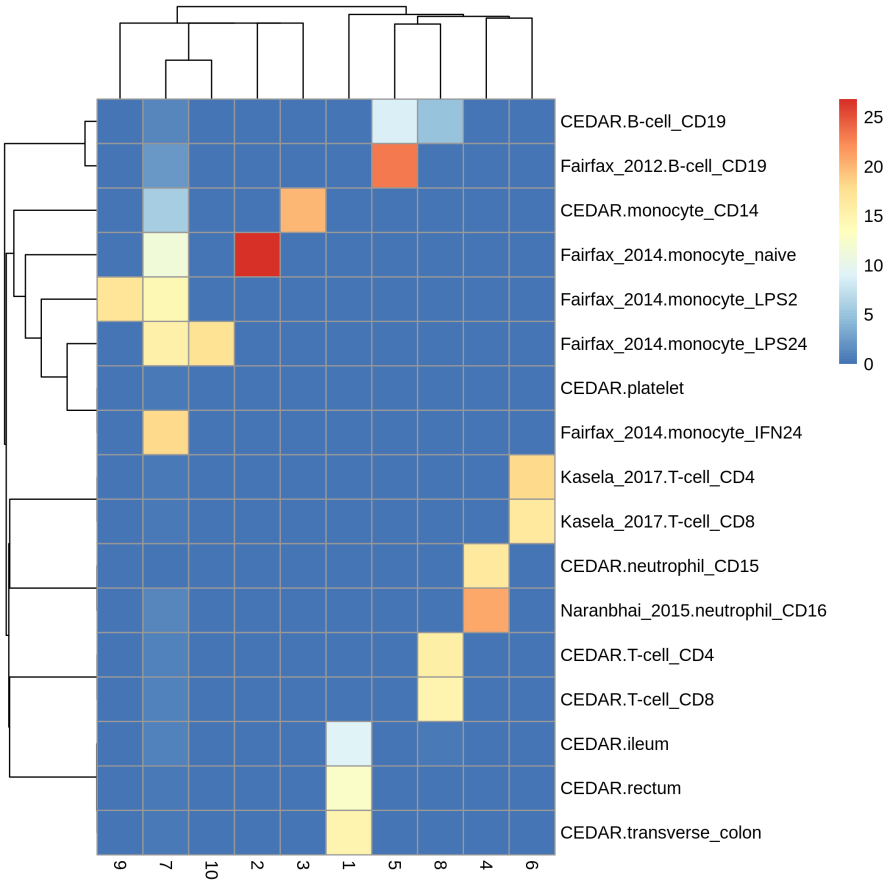

Figure 17. Final sn-spMF factor matrix for microarray dataset.

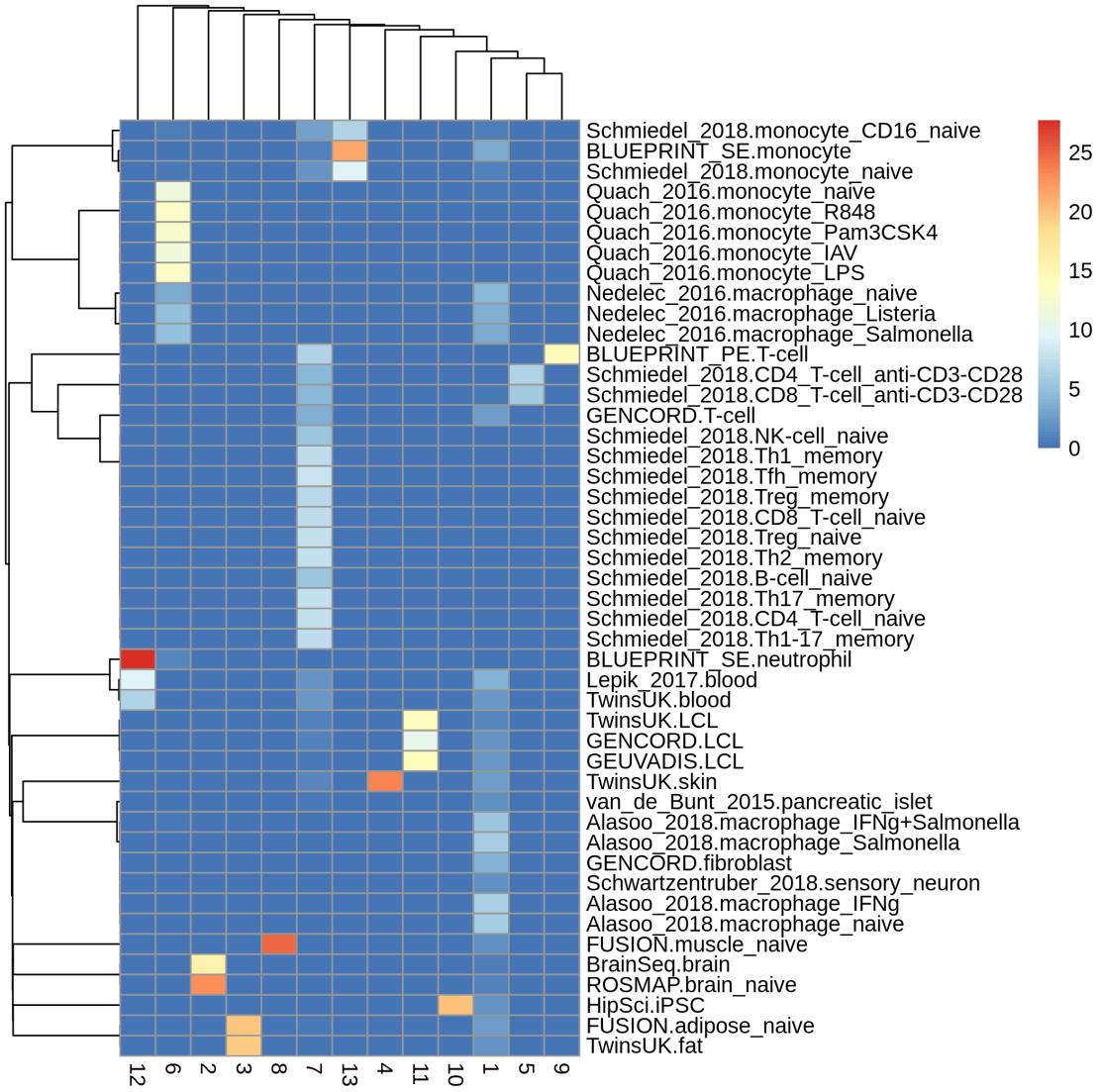

Figure 18. Final sn-spMF factor matrix for RNA-seq dataset.

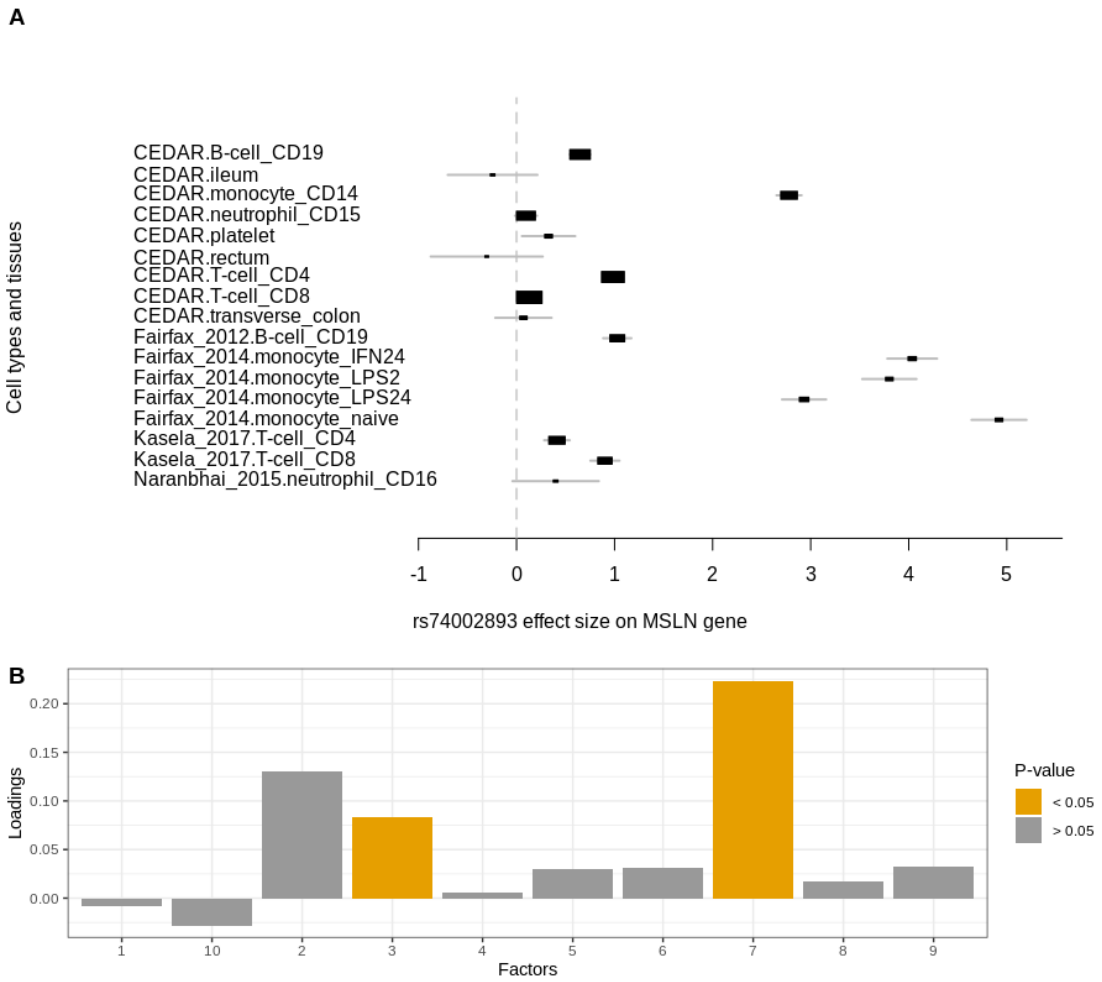

Figure 19. A. eQTL effect sizes across microarray tissues and cell types. The effect sizes are large for monocytes. B. The sn-spMF factor loadings for a given eQTL. The factor 7 which is specific to monocytes (see Figure 19) has a high loading.

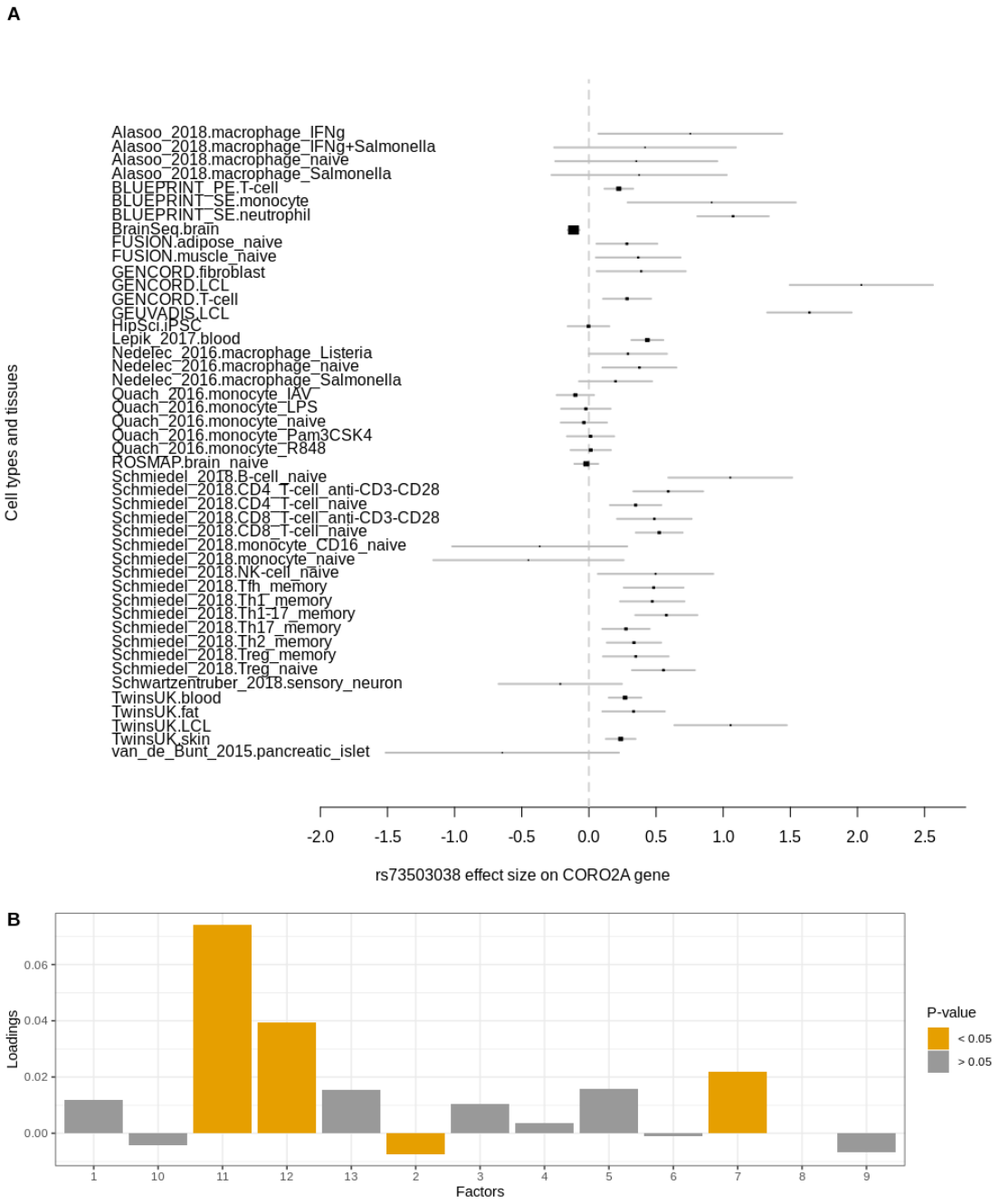

Figure 20. A. eQTL effect sizes across RNA-seq tissues and cell types. The effect sizes are large for Lymphoblastoid Cell lines (LCL). B. The sn-spMF factor loadings for the rs73503038 - CORO2A eQTL. Factors 11 (LCL-specific) has the largest loading.

3 Linking eQTL factors to specific disease with enrichment analysis

Latent factors in gene expression regulation can be used for variants annotation in disease enrichment. Disease heritability is a way to measure the genetic contribution to disease. It is a ratio of disease trait variation due to genetic factors [29]. The disease is often a trait in GWAS analysis. Many of them are polygenic (there are multiple genes and/or variants that contribute to the disease), so it is necessary to use methods that allow genetic correlations.

3.1 Methods

3.1.1 LD Score regression

LD Score regression is a model that aims to make use of the LD between variants. They estimate χ^2 association statistics through regression of GWAS heritability and variants ld-scores [20]:

$$E[\chi^2|l_j] = Nh^2 l_j/M + Na + 1 \tag{3}$$

where M is number of variants, N denotes the sample size, h^2/M is the average heritability and a measures the contribution of confounding biases [20].

3.1.2 Estimating partitioned heritability with stratified LD Score regression

The stratified LD Score regression aggregates effects across all of the variants that are in LD with a given variant to test association with a specific GWAS trait [19]. They assume that association statistics are higher for the variants that have higher correlation coefficients. If all variants are split into categories, then the category is more enriched for heritability if it correlates with (includes) variants that have higher association statistics. While a category that is in LD with variants that on average have smaller association statistics will not contribute to the heritability. We estimated disease enrichment for categories corresponding to latent factors discovered with matrix factorization.

LD score estimation The 1000 genomes phase 3 VCF files were obtained to run LD-score regression. The preprocessing of genotype data included subsetting European samples from VCF files, converting the files into binary plink format and filtering biallelic variants with a minor allele frequency 0.05. Additionally, allele frequencies were estimated, and centimorgans were mapped. We computed univariate LD Scores with 1 cM window on genotype data for every chromosome for later heritability estimation.

Factors assignment The partitioned disease heritability requires genomic region annotations. There are 24 main baseline annotation categories [19] with additional mapped factor annotations. Factor loadings from the matrix factorization approach were used as additional categories in the annotations. The LD Score regression method supports binary and categorical annotation values. We tried several strategies on how to embed eQTLs into annotations. At first, we mapped the variants from connected components with missing values to latent factors, obtaining loadings and p-values for each eQTL. We assigned eQTL to a factor category if the loading was significant (p-value < 0.05) (see Figure 21). We assigned variants to factor categories and performed disease enrichment separately for microarray and RNA-seq factors. Several strategies for annotation scores were tested in the thesis analysis: simple binary annotation, scaled effect sizes and posterior inclusion probability from fine-mapped credible sets.

The baseline approach is a binary annotation, where each eQTL simply belongs to a specific factor (annotated as 1) or not (annotated as 0). In the case of continuous annotation, we tried setting a weight as a scaled effect size. We scaled βs eQTL-wise by the largest effect by absolute value across cell types. The enrichment values were similar for both binary and scaled betas.

Posterior inclusion probabilities (pip) from fine-mapping results were used as an annotation in recent work [12]. In this approach, we used the largest pip-value across conditions for every eQTL. Thus, an eQTL has the same annotation value for all factors, that is a drawback of this method.

Usually, universal and shared factors (that have loadings on many cell types or tissues) are assigned with largest number of eQTLs (see Figure 21). Factors that are specific to only one particular cell type on average have less assigned eQTLs (e.g. Factor 9 that corresponds to LPS2 monocytes has only 3.41% of variants).

In our analysis, usage of pip-values as annotations lead to high uncertainty in enrichment estimates. The possible reason mentioned of nosy heritability enrichment mentioned in previous works [22] and [26] is a high correlation in factor categories.

Estimating partitioned LD Scores The annotation files consisting of baseline categories and factor categories were used to estimate partitioned LD Score in every chromosome separately. The centimorgan window size was the same in partitioned and uniformal LD Scores calculation.

GWAS source The GWAS data was obtained from a unified source GWAS Atlas [17]. Overall we run enrichment tests for 12 unique GWAS traits (Alzheimer disease, Body mass index, Bipolar disorder, Mean platelet volume, Type 2 diabetes, Schizophrenia, Systemic lupus erythematosus, Rheumatoid arthritis, Inflammatory bowel disease, Height, Crohn's disease, Ulcerative colitis).

As LD Score software requires a specific format of GWAS summary statistics, we had

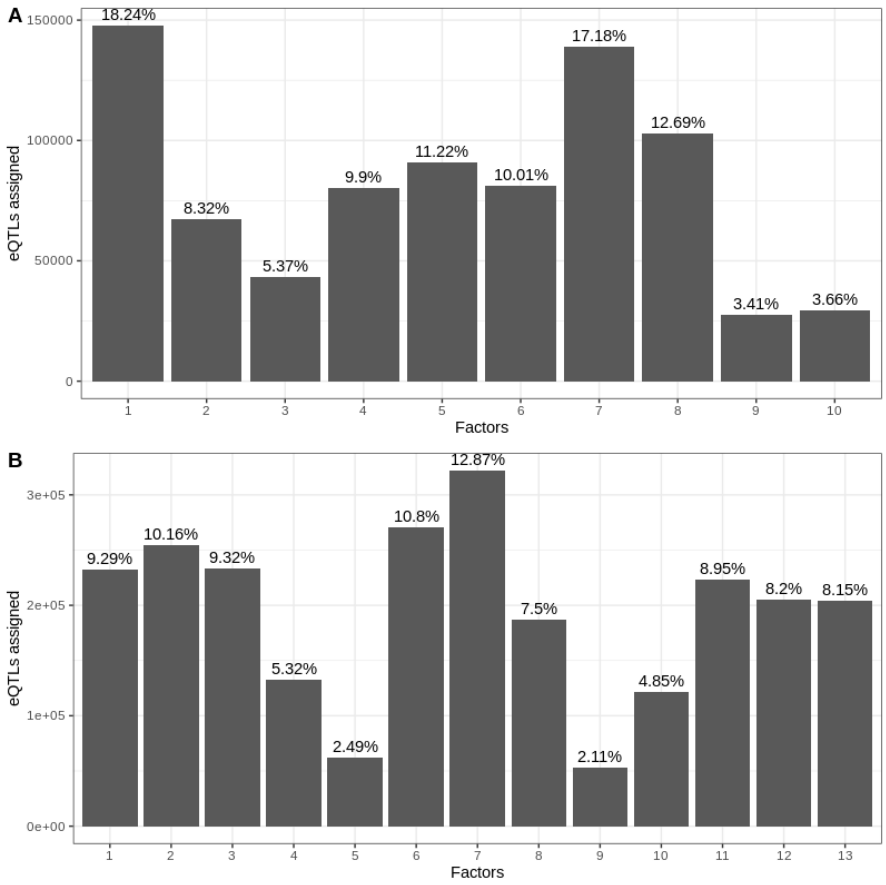

Figure 21. A. Distribution of eQTLs assigned to microarray factors (Figure 19). The factor 1 that is the only tissue-specific (other factors corresponds to cell types) factor in microarray data has been assigned the largest number of eQTLs. The second largest factor is factor 1, that is shared among many cell types. B. Distribution of eQTLs assigned to RNA-seq factors (Figure 20). Universal factor 7 has been assigned with largest number of variants. The cell-type-specific factor for T-cells from BLUEPRINT study has the smallest ratio of 2.11%.

to convert it to the required format. For each variant, we extracted effect size, standard error and log p-value from the summary statistics.

3.2 Results

None of the annotation strategies led to clear results (see Figures 22 and 23). Although we detected significant enrichment of fat tissue eQTLs (factor 3) for type 2 diabetes,

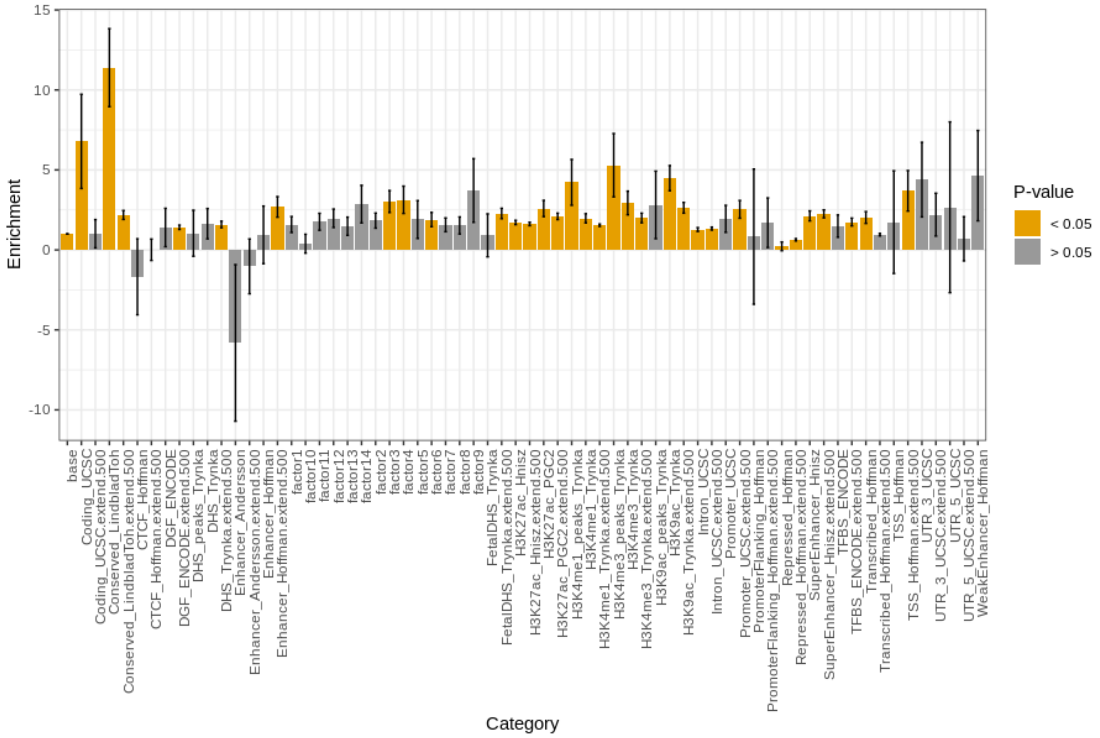

Figure 22. Type 2 diabetes enrichment for variant regions annotated by scaled RNA-seq factor loadings. Enrichment annotation was done with 24 baseline and 14 latent factors categories. Significant enrichments (factors 3, 4 and 6) are marked yellow. Significant factor categories correspond to fat-, skin- and monocyte-specific components (see Figure 20).

most of the other 14 factors showed similar absolute enrichment levels (see Figure 22).

We also tried to annotate all fine-mapped variants across all of the cell types and tissues, but having that large amount of annotated regions makes the enrichment in categories rather uniformal (see Figure 23). One possible reason for that there is a large overlap between categories introducing high correlation into the annotation. Another likely reason is most of the effects being small leading to low annotation weights.

Overall, the future research for disease enrichment with LD Score regression is needed. Moreover, datasets having purified cell types of large sample size can help to discover new cell-type-specific eQTLs and thus improve the annotation quality.

Technical details The preprocessing of genotype data was done with *bcftools* and *plink* tools. Necessary columns were extracted from GWAS VCF files to table format using *bcftools*. We used *ldsc* software for LD score estimation and heritability enrichment.

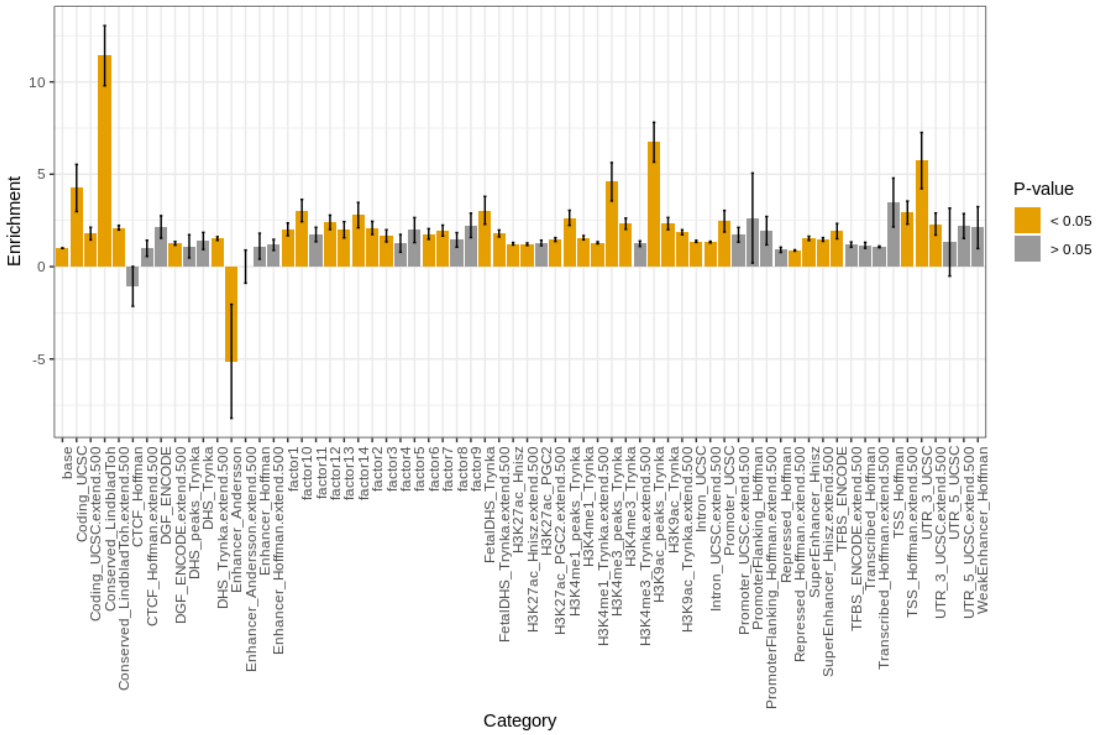

Figure 23. Schizophrenia enrichment for variant regions annotated by scaled RNA-seq factor loadings. Enrichment annotation was done with 24 baseline and 14 latent factors categories. Significant enrichments are marked yellow.

Discussion

Conclusion In this thesis, we assessed several complex methods in discovering sharing and specificity structure of genetic regulatory effects. We performed the analysis on microarray and RNA-seq summary statistics from the eQTL Catalogue. We came up with a connected components strategy on how to aggregate the eQTL effect across many conditions. The sharing results of eQTL effects agrees well with known biology (related cell types have high sharing degree and cluster together) and technical technical differences between studies seem to have smaller effects. However, the small study effects are present as well. With Mash model and factorization approach, we found that complex tissue share many eQTL effects while purified cell types have significantly less of sharing both in microarray and RNA-seq datasets. We used different clustering approaches and discovered that usually, pure cell types have a small sharing outside of their cluster. High tissue sharing and low cell type sharing supports previous findings that eQTL sharing between tissues is driven by that sharing of component cell types between tissues [9]. Furthermore, tissues capture only average eQTL effect sizes of their many component cell types. Thus, high tissue sharing might be driven by eQTLs that are present in many cell types (and thus appear to have large effects in multiple tissues) whereas highly cell-type-specific eQTLs might have weak effects in tissues due to the small prevalence of those cell types.

We discovered the eQTL effects latent factors for microarray and RNA-seq datasets. A shared factor for multiple tissues and cell types was discovered for both microarray and RNA-seq data. Multiple highly cell-type-specific factors were found, as well. Additionally, we tested the LD Score regression method for associating the variant loci with diseases. Annotation strategies did not give satisfactory results for stating significantly enriched factor categories, probably due to most eQTLs loading on more than one factor and introducing correlation between category annotations.

Future work Further analysis can include a combination of microarray and RNA-seq datasets into a single analysis. Additionally, analysing datasets with larger cell type sample sizes may help to discover new eQTLs associated with diseases. Discovering other annotation strategies for disease heritability enrichment with partitioned LD Score regression is possible plan for future work.

Acknowledgments

I would like to thank my supervisor, Kaur Alasoo, for his constant support, meaningful comments, reasonable questions, patience and optimism. I appreciate the numerous ideas he gave me during the research and his bright interpretation of any results we had, even when it seemed nothing is working. I value his readiness to discuss results and how he could always find time for me.

Also, I would like to thank Nurlan Kerimov for developing the eQTL Catalogue and making the analysis possible. I am also very grateful for the office mates who were around supporting me and making working hours more fun.

References

[1] Frank W. Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, April 2015.

[2] Jo Bovy, David W. Hogg, and Sam T. Roweis. Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations. *The Annals of Applied Statistics*, 5(2B):1657–1677, June 2011. arXiv: 0905.2979.

[3] Lu Chen, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yan, Kousik Kundu, Simone Ecker, Avik Datta, David Richardson, Frances Burden, Daniel Mead, Alice L. Mann, Jose Maria Fernandez, Sophia Rowlston, Steven P. Wilder, Samantha Farrow, Xiaojian Shao, John J. Lambourne, Adriana Redensek, Cornelis A. Albers, Vyacheslav Amstislavskiy, Sofie Ashford, Kim Berentsen, Lorenzo Bomba, Guillaume Bourque, David Bujold, Stephan Busche, Maxime Caron, Shu-Huang Chen, Warren Cheung, Oliver Delaneau, Emmanouil T. Dermitzakis, Heather Elding, Irina Colgiu, Frederik O. Bagger, Paul Flicek, Ehsan Habibi, Valentina Iotchkova, Eva Janssen-Megens, Bowon Kim, Hans Lehrach, Ernesto Lowy, Amit Mandoli, Filomena Matarese, Matthew T. Maurano, John A. Morris, Vera Pancaldi, Farzin Pourfarzad, Karola Rehnstrom, Augusto Rendon, Thomas Risch, Nilofar Sharifi, Marie-Michelle Simon, Marc Sultan, Alfonso Valencia, Klaudia Walter, Shuang-Yin Wang, Mattia Frontini, Stylianos E. Antonarakis, Laura Clarke, Marie-Laure Yaspo, Stephan Beck, Roderic Guigo, Daniel Rico, Joost H.A. Martens, Willem H. Ouwehand, Taco W. Kuijpers, Dirk S. Paul, Hendrik G. Stunnenberg, Oliver Stegle, Kate Downes, Tomi Pastinen, and Nicole Soranzo. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, 167(5):1398–1414.e24, November 2016.

[4] Melina Claussnitzer, Simon N. Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S. Sousa, Jacqueline L. Beaudry, Vijitha Puviindran, Nezar A. Abdennur, Jannel Liu, Per-Arne Svensson, Yi-Hsiang Hsu, Daniel J. Drucker, Gunnar Mellgren, Chi-Chung Hui, Hans Hauner, and Manolis Kellis. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine*, 373(10):895–907, September 2015.

[5] dbSNP. Database of Single Nucleotide Polymorphisms (dbSNP). Technical Report dbSNP Build ID: 154, National Center for Biotechnology Information, National Library of Medicine., Bethesda (MD).

[6] EMBRACE Collaborators, GC-HBOC Study Collaborators, GEMO Study Collaborators, ABCTB Investigators, HEBON Investigators, BCFR Investigators, Manuel A. Ferreira, Eric R. Gamazon, Fares Al-Ejeh, Kristiina Aittomäki, Irene L. Andrulis, Hoda Anton-Culver, Adalgeir Arason, Volker Arndt, Kristan J. Aronson, Banu K. Arun, Ella Asseryanis, Jacopo Azzollini, Judith Balmaña, Daniel R. Barnes, Daniel Barrowdale, Matthias W. Beckmann, Sabine Behrens, Javier Benitez, Marina Bermisheva, Katarzyna Białkowska, Carl Blomqvist, Natalia V. Bogdanova, Stig E. Bojesen, Manjeet K. Bolla, Ake Borg, Hiltrud Brauch, Hermann Brenner, Annegien Broeks, Barbara Burwinkel, Trinidad Caldés, Maria A. Caligo, Daniele Campa, Ian Campbell, Federico Canzian, Jonathan Carter, Brian D. Carter, Jose E. Castelao, Jenny Chang-Claude, Stephen J. Chanock, Hans Christiansen, Wendy K. Chung, Kathleen B. M. Claes, Christine L. Clarke, Fergus J. Couch, Angela Cox, Simon S. Cross, Kamila Czene, Mary B. Daly, Miguel de la Hoya, Joe Dennis, Peter Devilee, Orland Diez, Thilo Dörk, Alison M. Dunning, Miriam Dwek, Diana M. Eccles, Bent Ejlertsen, Carolina Ellberg, Christoph Engel, Mikael Eriksson, Peter A. Fasching, Olivia Fletcher, Henrik Flyger, Eitan Friedman, Debra Frost, Marike Gabrielson, Manuela Gago-Dominguez, Patricia A. Ganz, Susan M. Gapstur, Judy Garber, Montserrat García-Closas, José A. García-Sáenz, Mia M. Gaudet, Graham G. Giles, Gord Glendon, Andrew K. Godwin, Mark S. Goldberg, David E. Goldgar, Anna González-Neira, Mark H. Greene, Jacek Gronwald, Pascal Guénel, Christopher A. Haiman, Per Hall, Ute Hamann, Wei He, Jane Heyworth, Frans B. L. Hogervorst, Antoinette Hollestelle, Robert N. Hoover, John L. Hopper, Peter J. Hulick, Keith Humphreys, Evgeny N. Imyanitov, Claudine Isaacs, Milena Jakimovska, Anna Jakubowska, Paul A. James, Ramunas Janavicius, Rachel C. Jankowitz, Esther M. John, Nichola Johnson, Vijai Joseph, Beth Y. Karlan, Elza Khusnutdinova, Johanna I. Kiiski, Yon-Dschun Ko, Michael E. Jones, Irene Konstantopoulou, Vessela N. Kristensen, Yael Laitman, Diether Lambrechts, Conxi Lazaro, Goska Leslie, Jenny Lester, Fabienne Lesueur, Sara Lindström, Jirong Long, Jennifer T. Loud, Jan Lubiński, Enes Makalic, Arto Mannermaa, Mehdi Manoochehri, Sara Margolin, Tabea Maurer, Dimitrios Mavroudis, Lesley McGuffog, Alfons Meindl, Usha Menon, Kyriaki Michailidou, Austin Miller, Marco Montagna, Fernando Moreno, Lidia Moserle, Anna Marie Mulligan, Katherine L. Nathanson, Susan L. Neuhausen, Heli Nevanlinna, Ines Nevelsteen, Finn C. Nielsen, Liene Nikitina-Zake, Robert L. Nussbaum, Kenneth Offit, Edith Olah, Olufunmilayo I. Olopade, Håkan Olsson, Ana Osorio, Janos Papp, Tjoung-Won Park-Simon, Michael T. Parsons, Inge Sokilde Pedersen, Ana Peixoto, Paolo Peterlongo, Paul D. P. Pharoah, Dijana Plaseska-Karanfilska, Bruce Poppe, Nadege Presneau, Paolo Radice, Johanna Rantala, Gad Rennert, Harvey A. Risch, Emmanouil Saloustros, Kristin Sanden, Elinor J. Sawyer, Marjanka K. Schmidt, Rita K. Schmutzler, Priyanka Sharma, Xiao-Ou Shu, Jacques Simard, Christian F. Singer, Penny Soucy,

Melissa C. Southey, John J. Spinelli, Amanda B. Spurdle, Jennifer Stone, Anthony J. Swerdlow, William J. Tapper, Jack A. Taylor, Manuel R. Teixeira, Mary Beth Terry, Alex Teulé, Mads Thomassen, Kathrin Thöne, Darcy L. Thull, Marc Tischkowitz, Amanda E. Toland, Diana Torres, Thérèse Truong, Nadine Tung, Celine M. Vachon, Christi J. van Asperen, Ans M. W. van den Ouweland, Elizabeth J. van Rensburg, Ana Vega, Alessandra Viel, Qin Wang, Barbara Wappenschmidt, Jeffrey N. Weitzel, Camilla Wendt, Robert Winqvist, Xiaohong R. Yang, Drakoulis Yannoukakos, Argyrios Ziogas, Peter Kraft, Antonis C. Antoniou, Wei Zheng, Douglas F. Easton, Roger L. Milne, Jonathan Beesley, and Georgia Chenevix-Trench. Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. *Nature Communications*, 10(1):1741, December 2019.

[7] B. P. Fairfax, P. Humburg, S. Makino, V. Naranbhai, D. Wong, E. Lau, L. Jostins, K. Plant, R. Andrews, C. McGee, and J. C. Knight. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. *Science*, 343(6175):1246949–1246949, March 2014.

[8] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, October 2017.

[9] GTEx Consortium, Eric R. Gamazon, Ayellet V. Segrè, Martijn van de Bunt, Xiaoquan Wen, Hualin S. Xi, Farhad Hormozdiari, Halit Ongen, Anuar Konkashbaev, Eske M. Derks, François Aguet, Jie Quan, Dan L. Nicolae, Eleazar Eskin, Manolis Kellis, Gad Getz, Mark I. McCarthy, Emmanouil T. Dermitzakis, Nancy J. Cox, and Kristin G. Ardlie. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics*, 50(7):956–967, July 2018.

[10] Yuan He, Surya B. Chhetri, Marios Arvanitis, Kaushik Srinivasan, François Aguet, Kristin G. Ardlie, Alvaro N. Barbeira, Rodrigo Bonazzola, Hae Kyung Im, GTEx Consortium, Christopher D. Brown, and Alexis Battle. Mechanisms of tissue-specific genetic regulation revealed by latent factors across eQTLs. preprint, Genomics, October 2019.

[11] HIPSCI Consortium, Kaur Alasoo, Julia Rodrigues, Subhankar Mukhopadhyay, Andrew J. Knights, Alice L. Mann, Kousik Kundu, Christine Hale, Gordon Dougan, and Daniel J. Gaffney. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature Genetics*, 50(3):424–431, March 2018.

[12] Farhad Hormozdiari, Steven Gazal, Bryce van de Geijn, Hilary K. Finucane, Chelsea J.-T. Ju, Po-Ru Loh, Armin Schoech, Yakir Reshef, Xuanyao Liu, Luke

O'Connor, Alexander Gusev, Eleazar Eskin, and Alkes L. Price. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nature Genetics*, 50(7):1041–1047, July 2018.

[13] Silva Kasela, Kai Kisand, Liina Tserel, Epp Kaleviste, Anu Remm, Krista Fischer, Tõnu Esko, Harm-Jan Westra, Benjamin P. Fairfax, Seiko Makino, Julian C. Knight, Lude Franke, Andres Metspalu, Pärt Peterson, and Lili Milani. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. *PLOS Genetics*, 13(3):e1006643, March 2017.

[14] Nurlan Kerimov, James D. Hayhurst, Jonathan R. Manning, Peter Walter, Liis Kolberg, Kateryna Peikova, Marija Samoviča, Tony Burdett, Simon Jupp, Helen Parkinson, Irene Papatheodorou, Daniel R. Zerbino, and Kaur Alasoo. eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. preprint, Genomics, January 2020.

[15] Liis Kolberg, Nurlan Kerimov, Hedi Peterson, and Kaur Alasoo. Co-expression analysis reveals interpretable gene modules controlled by *trans* -acting genetic variants. preprint, Genomics, April 2020.

[16] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S

Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, A Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok, and Helen F Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, June 2013.

[17] Matthew Lyon, Shea J Andrews, Ben Elsworth, Tom R Gaunt, Gibran Hemani, and Edoardo Marcora. The variant call format provides efficient and robust storage of GWAS summary statistics. preprint, Genetics, May 2020.

[18] Hélène Quach, Maxime Rotival, Julien Pothlichet, Yong-Hwee Eddie Loh, Michael Dannemann, Nora Zidane, Guillaume Laval, Etienne Patin, Christine Harmant, Marie Lopez, Matthieu Deschamps, Nadia Naffakh, Darragh Duffy, Anja Coen, Geert Leroux-Roels, Frederic Clément, Anne Boland, Jean-François Deleuze, Janet Kelso, Matthew L. Albert, and Lluis Quintana-Murci. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell*, 167(3):643–656.e17, October 2016.

[19] ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, The RACI Consortium, Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R Day, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R B Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J Daly, Nick Patterson, Benjamin M Neale, and Alkes L Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235, November 2015.

[20] Schizophrenia Working Group of the Psychiatric Genomics Consortium, Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, March 2015.

[21] Benjamin J. Schmiedel, Divya Singh, Ariel Madrigal, Alan G. Valdovino-Gonzalez, Brandie M. White, Jose Zapardiel-Gonzalo, Brendan Ha, Gokmen Altay, Jason A. Greenbaum, Graham McVicker, Grégory Seumois, Anjana Rao, Mitchell Kronenberg, Bjoern Peters, and Pandurangan Vijayanand. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell*, 175(6):1701–1715.e16, November 2018.

[22] Blagoje Soskic, Eddie Cano-Gamez, Deborah J. Smyth, Wendy C. Rowan, Nikolina Nakic, Jorge Esparza-Gordillo, Lara Bossini-Castillo, David F. Tough, Christopher G. C. Larminie, Paola G. Bronson, David Willé, and Gosia Trynka. Chromatin

activity at GWAS loci identifies T cell states driving complex immune diseases. *Nature Genetics*, 51(10):1486–1493, October 2019.

[23] Matthew Stephens. False discovery rates: a new deal. *Biostatistics*, page kxw041, October 2016.

[24] Jae Hoon Sul, Buhm Han, Chun Ye, Ted Choi, and Eleazar Eskin. Effectively Identifying eQTLs from Multiple Tissues by Combining Mixed Model and Meta-analytic Approaches. *PLoS Genetics*, 9(6):e1003491, June 2013.

[25] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[26] The Brainstorm Consortium, Hilary K. Finucane, Yakir A. Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, Giulio Genovese, Arpiar Saunders, Evan Macosko, Samuela Pollack, John R. B. Perry, Jason D. Buenrostro, Bradley E. Bernstein, Soumya Raychaudhuri, Steven McCarroll, Benjamin M. Neale, and Alkes L. Price. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics*, 50(4):621–629, April 2018.

[27] The International IBD Genetics Consortium, Yukihide Momozawa, Julia Dmitrieva, Emilie Théâtre, Valérie Deffontaine, Souad Rahmouni, Benoît Charloteaux, François Crins, Elisa Docampo, Mahmoud Elansary, Ann-Stephan Gori, Christelle Lecut, Rob Mariman, Myriam Mni, Cécile Oury, Ilya Altukhov, Dmitry Alexeev, Yuri Aulchenko, Leila Amininejad, Gerd Bouma, Frank Hoentjen, Mark Löwenberg, Bas Oldenburg, Marieke J. Pierik, Andrea E. vander Meulen-de Jong, C. Janneke van der Woude, Marijn C. Visschedijk, Mark Lathrop, Jean-Pierre Hugot, Rinse K. Weersma, Martine De Vos, Denis Franchimont, Severine Vermeire, Michiaki Kubo, Edouard Louis, and Michel Georges. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nature Communications*, 9(1):2427, December 2018.

[28] Sarah M. Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, 51(1):187–195, January 2019.

[29] Peter M. Visscher, William G. Hill, and Naomi R. Wray. Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266, April 2008.

[30] Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine-mapping. preprint, Bioinformatics, December 2018.

[31] Harm-Jan Westra and Lude Franke. From genome to function by studying eQTLs. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10):1896–1902, October 2014.

Appendix

I. Code

The source code for the analysis described in the thesis is publicly open in the following GitHub repository:

`https://github.com/peikovakate/genetic_effects`

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Kateryna Peikova**,

1. herewith 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,

   Genetic effects on gene expression across cell types, tissues and biological contexts,

   supervised by Kaur Alasoo.

2. I grant the University of Tartu a permit to make the work specified in p. 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 3.0, 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 until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Kateryna Peikova
10/08/2020