MODELING THE IMPACT OF GENETIC VARIATION ON GENE EXPRESSION

by
Benjamin J. Strober

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

November 2021

Abstract

A complete, mechanistic understanding of the consequences of genetic variation could provide immense insights into disease development and, ultimately, human health. A powerful approach for filling in the missing links between genetic variation and higher order traits is to use molecular phenotypes, such as gene expression levels, as an intermediate phenotype. This dissertation advances our understanding of how the genetic regulation of gene expression changes as a function of cellular context or environment. Secondly, this dissertation provides a novel approach to identify functional rare variants via the incorporation of gene expression data. These advances have been achieved via the completion of four major projects.

My first project quantified how genetic effects on gene expression, as measured by expression quantitative trait loci (eQTL), vary between tissues of the human body. For my second project, we identified changes in genetic regulation of gene expression along the continuous process of cellular differentiation by detecting dynamic eQTLs, or eQTLs whose effect size changed according to differentiation progress. Broadly, these first two projects identified instances of context-specific eQTLs, as well as describe their biological relevance. However, the relevant contexts, such as cell type or state, that actually modulate genetic effects may not be known a priori. Therefore, in the third project, we developed SURGE, a novel probabilistic model to learn a continuous representation of the cellular contexts that modulate genetic effects. In my fourth project we assessed how rare genetic variants contribute to extreme patterns of gene expression. We developed a probabilistic model, SPOT, to identify instances of abnormal splicing patterns. We also developed Watershed, a novel probabilistic model that identifies functional rare genetic variants by integrating patterns of gene expression data and genomic annotation data.

Thesis Readers

Alexis Battle, Ph.D. (advisor). Associated Professor, Dept. of Biomedical Engineering, Dept. of Computer Science, Dept. of Genetic Medicine. Johns Hopkins University

Archana Venkataraman, Ph.D. John C. Malone Assistant Professor, Dept. of Electrical and Computer Engineering. Johns Hopkins University

David Lee Valle, M.D. Professor, Dept. of Genetic Medicine. Johns Hopkins University

Acknowledgements

I would like to first and foremost thank my graduate advisor, Alexis Battle, for her guidance and support over the past six years. Alexis has been an inspiration to me ever since I first rotated in her lab. I am constantly amazed by her ability formulate clear scientific questions and elegantly communicate scientific thoughts to others. Over the course of graduate school, Alexis has taught me many things, both technical skills and life-lessons, but most importantly she has taught me how to be a scientist. I cannot thank her enough.

I am also thankful for the numerous scientific collaborations I have had with other labs during my graduate studies. I believe scientific collaboration is essential for modern scientific advances. Most notably, I would like to thank Yoav Gilad, Katie Rhodes and Reem Elorbany at University of Chicago for a fantastic collaboration while we studied dynamic genetic effects during cellular differentiation. Not only was this collaboration successful, but it was also fun. Similarly, I would like to thank Nicole Ferraro and Stephen Montgomery from Stanford University for a great collaboration studying the effects of rare genetic variants on gene expression.

Next, I would like to thank my committee members, Archana Venkataraman and David Valle for their thoughtful insight and guidance on my projects.

I would also like to thank all current and previous members of the Battle lab for their support. You have all made the Battle lab a fun and supportive place to work. In particular, I would like to thank Karl Tayeb, Princy Parsana, Josh Popp, and Ashton Omdahl for always being willing to brainstorm research ideas and for being good friends.

I am immensely grateful for the close-knit group of friends I have had during my graduate studies. You guys have made this process fun. I am thankful for fellow BME PhD students, Michael Ketcha and Meg Chow. We've shared outdoor adventures, nights at the Lithuanian Dance Hall, and annual sleepover parties with pillow forts. Next, I want to thank Matt Sheckells and Andrew Spielvogel for their companionship. Our daily lunches were sometimes the only reason I went to work. And finally, I'd like to thank my former landlord, housemate and friend, Mike Peven for 'guys night's.

I would also like to thank my family. My parents, Sam and Lynn, have provided immense support my entire life. My dad, through example, taught me the importance of hard work and dedication to your trade, as well as the importance of having fun outside of work. My mom has instilled in me a love of science and math from a very young age and has been unbelievably encouraging during my graduate studies. My brother, Jordan, who coincidentally started a PhD the same year I did, has been a role-model of mine my entire life.

I would like to thank my partner Dr. Steph Cabral. Steph and I met during undergraduate at Johns Hopkins while both studying Biomedical Engineering. And while I completed graduate school, Steph completed medical school. During this time, Steph has been my best friend, primary climbing partner, and cat co-parent. I cannot thank her enough for her constant encouragement.

And last but not least, I'd like to thank my cat and faithful companion Margo Strober-Cabral. She has been a beacon of support throughout this process.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

The development of DNA sequencing technologies over the past 20 years has enabled the population-scale measurement of DNA sequence across the entire genome. Interestingly, approximately 99.9% of DNA is identical across humans, with only 0.1% of DNA sequence being variable. We will refer to DNA positions that have variable sequence across humans as a genetic variants.

Sequencing technologies have also furthered our understanding of genes and gene functions. Genes are segments of DNA sequence that contain the necessary information for the production of proteins. The basic function of a gene can be understood through the central dogma of molecular biology. DNA sequence corresponding to a gene is first transcribed into messenger RNA, or gene expression, and then the gene expression is translated into proteins. It is well-studied that gene expression is a highly regulated process. The amount of expression produced by a given gene is tightly controlled by binding of proteins to genetic regulatory elements, such as promotors and enhancers, in DNA sequence nearby the gene (Figure 1-1; 2). Tight regulation of gene expression allows for cells to express different genes in response to different environmental conditions or cellular contexts. Importantly, modern RNA-sequencing technologies allow for population-scale measurement of the relative abundance of the gene expression across all genes in the genome.

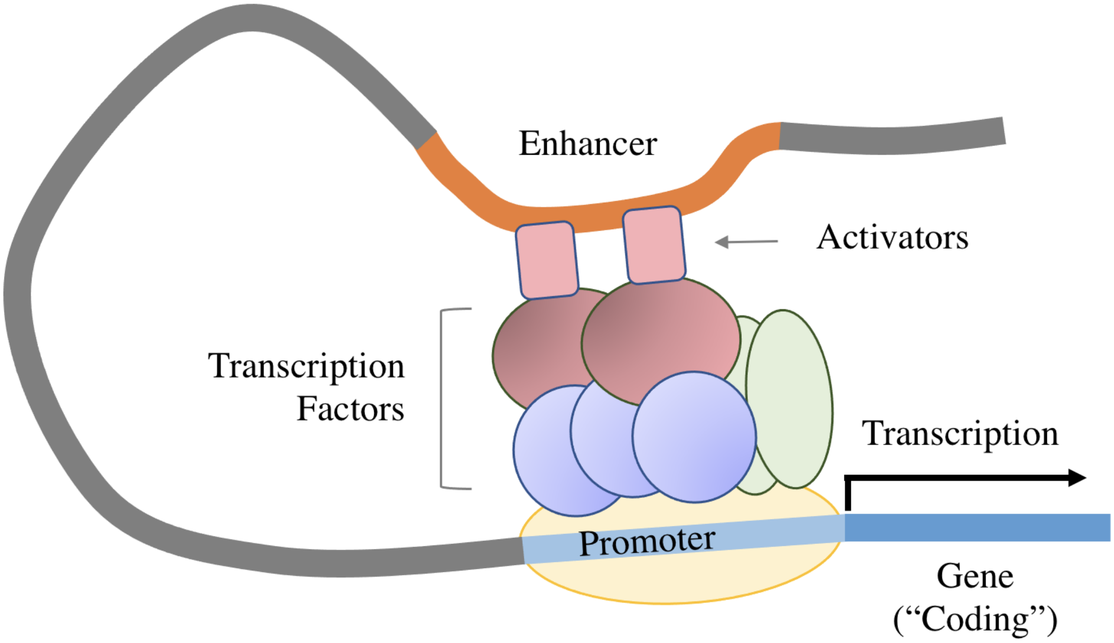


Figure 1-1: Illustration of gene expression regulation

Schematic illustration representing regulation of gene expression through proteins (transcription factors and activators) binding to genetic regulatory elements (promotors and enhancers, respectively) nearby a gene.

Naturally, a complete, mechanistic understanding of the consequences of genetic variation could provide immense insights into disease development and, ultimately, human health. The advent of population-scale detection of genetic variants has subsequently catalyzed the use of genome wide association studies (GWAS) to identify associations between genetic variations and complex traits or disease (1). However, most genetic variants found to be associated with complex traits are not located within protein-coding genes (2), thereby providing no mechanistic insight as to how the genetic variant regulates the complex trait.

These non-coding variants are thought to contribute, in part, to higher order traits through the regulation of expression of nearby genes via the disruption or creation of genetic regulatory elements (2). A natural, and well-studied, approach for filling in mechanistic links between genetic variation and higher order traits is to use molecular phenotypes, such as gene expression, as an intermediate phenotype (3,4). Here, we will review two distinct approaches that use gene expression measurements, most commonly quantified through RNA sequencing, to help prioritize variants with functional effects on gene expression. The first approach, expression quantitative trait loci (eQTL) analysis, is generally limited in application to common variants, or variants with minor allele frequency greater than 5%. The second approach, RNA sequencing outlier analysis, is designed to detect rare variants, or variants with allele frequency less than 1%, with functional effects on gene expression.

Common variant prioritization with eQTL analysis

Variants significantly associated with mRNA expression are known as eQTLs. Those eQTLs that affect nearby genes are called *cis*-eQTLs, while those affecting distal genes are *trans*-eQTLs. eQTL analysis has shown that many disease-associated (GWAS) loci influence the regulation of nearby genes (ie. the GWAS loci are also eQTLs) (*5)*. However, a substantial fraction of disease-associated loci still remain unexplained (*6*). There are likely a multitude of reasons why there is not stronger overlap between eQTL and GWAS signals, including incomplete power to detect small effect size genetic associations (*6*) and genetic variants acting through a mechanism other than regulation of the absolute expression levels of genes (*4*). Another possible explanation, which we will focus on in this thesis, is genetic regulatory effects can be cell-type or context dependent (*7, 8*). For example, the discovery of a variant that regulates cardiomyocyte expression levels of a particular gene, but does not regulate endothelial cell expression levels for the same gene. One hypothesis of how context-specific eQTLs arise is through the variant disrupting or enhancing the binding affinity of a context-specific transcription factor to a transcription factor binding site.

A main focus of this thesis was studying how genetic regulation of gene expression (ie. eQTLs) change as a function of context and developing new statistical methods to model context specific changes in the genetic regulation of gene expression. In Chapter 2, we quantify levels of tissue specificity of cis and trans eQTLs across 49 human tissues. In Chapter 3, we model the dynamic genetic regulation of gene expression throughout the process of cellular differentiation. In Chapter 4, we develop a novel methodological approach to identify latent contexts that drive changes in genetic regulation of gene expression.

Rare variant prioritization with RNA sequencing outlier analysis

The human genome contains tens of thousands of rare (minor allele frequency <1%) variants, some of which may contribute to disease risk. Unlike common variants, standard approaches such as eQTL analysis cannot be used to identify functional rare variants because the underlying regression models are underpowered to detect associations when only one or a few samples contain the rare allele for the variant of interest. One approach to identify functional rare variants is through outlier calling.

Outlier calling is motivated by the simple hypothesis that a functional variant, regardless of allele frequency, will cause a disruption at the cellular level. More specifically, a functional *rare* variant will result in abnormal or extreme expression of a nearby gene, relative to the general population. Therefore, RNA sequencing outlier analysis is simply identifying individuals that have extreme expression levels for a particular gene (outliers), and using that outlier status to prioritize rare variants nearby the gene in the outlier individual. Previous work (9), has shown that individuals with extreme total expression levels (eOutliers) are enriched for nearby rare variants. In Chapter 5 of this thesis, we extend what it means to be extreme with respect to gene expression levels from just total expression levels (eOutliers) to include splicing (sOultiers) and allele specific expression (aseOutliers).

Chapter 2 Quantify tissue-specificity of cis and trans eQTLs

Contributions

This chapter describes analysis of tissue-specificity of cis and trans-eQTLs as a part of the GTEx version 6p eQTL project. I co-led this project along with Francois Aguet, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, YoSon Park , Princy Parsana, Ayellet V. Segre, and Zachary Zappala. This work was published in (10). My main contribution to the published manuscript includes:

- Analysis of tissue-specificity of cis and trans eQTLs using replication rate modeling and Meta-Tissue (11)
- Enrichment analysis of cis and trans eQTLs within cis-regulatory elements

The text of this chapter is a modification of the published work (10), focusing on results relevant to my contribution. The text was written together by Francois Aguet, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, YoSon Park , Princy Parsana, Ayellet V. Segre, Benjamin J. Strober, Zachary Zappala, Alexis Battle, Christopher D. Brown, Barbara E. Engelhardt, and Stephen B. Montgomery. The full list of collaborators involved in this project is available in (10).

Abstract

Characterization of the molecular function of the human genome and its variation across individuals is essential for identifying the cellular mechanisms that underlie human genetic traits and diseases. The Genotype-Tissue Expression (GTEx) project aims to characterize variation in gene expression levels across individuals and diverse tissues of the human body, many of which are not easily accessible. Here we describe genetic effects on gene expression levels across 44 human tissues. We find that local genetic variation affects gene expression levels for the majority of genes, and we further identify inter-chromosomal genetic effects for 93 genes and 112 loci. On the basis of the identified genetic effects, we characterize patterns of tissue specificity, compare local and distal effects, and evaluate the functional properties of the genetic effects. We discovered trans eQTLs exhibit stronger levels of tissue-specificity than cis-eQTLs.

Introduction

The human genome encodes instructions for the regulation of gene expression, which varies both across cell types and across individuals. Recent large-scale studies have characterized the regulatory function of the genome across a diverse array of cell types, each from a small number of samples (12). Measuring how gene regulation and expression vary across individuals has further expanded our understanding of the functions of healthy tissues and the molecular origins of complex traits and diseases (2,3,13). However, these studies have been conducted in limited, accessible cell types, thus restricting the utility of these studies in informing regulatory biology and human health.

The Genotype-Tissue Expression (GTEx) project was established to characterize human transcriptomes within and across individuals for a wide variety of primary tissues and cell types. Here, we report on a major expansion of the GTEx project that includes publicly available genotype, gene expression, histological and clinical data for 449 human donors across 44 (42 distinct) tissues. This enables the study of tissue-specific gene expression and the identification of genetic associations with gene expression levels (expression quantitative trait loci, or eQTLs) across many tissues, including both local (cis-eQTLs) and distal (trans-eQTLs) effects.

In this study, we associate genetic variants with gene expression levels from the GTEx v6p release. We found pervasive *cis*-eQTLs, which affect the majority of human genes. In addition, we identify *trans*-eQTLs across 18 tissues and highlight their increased tissue specificity relative to *cis*-eQTLs.

Results

We identified associations between the expression levels of all expressed genes (eGenes) and genetic variants (eVariants) located within 1 Mb of the target gene's transcription start site (TSS), which we refer to as cis-eQTLs for convenience, without requiring evidence of allelic effects at each locus. However, the majority of *cis*-eQTLs do exhibit allele specific expression. We applied a linear model controlling for ancestry, sex, genotyping platform and latent factors (14) in the expression data for each tissue that may reflect batch or other technical variables (see Methods). Considering all tissues, we found a total of 152,869 *cis*-eQTLs for 19,725 genes, representing 50.3%

and 86.1% of all known autosomal long intergenic noncoding RNA (lincRNA) and protein-coding genes, respectively (Figure 2-1).

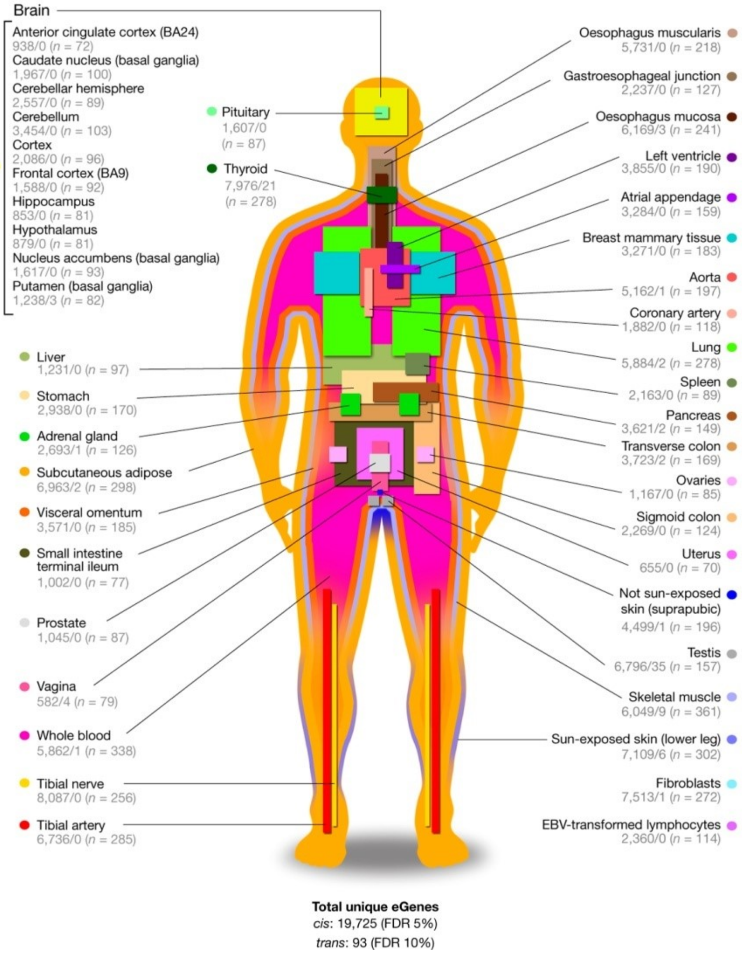
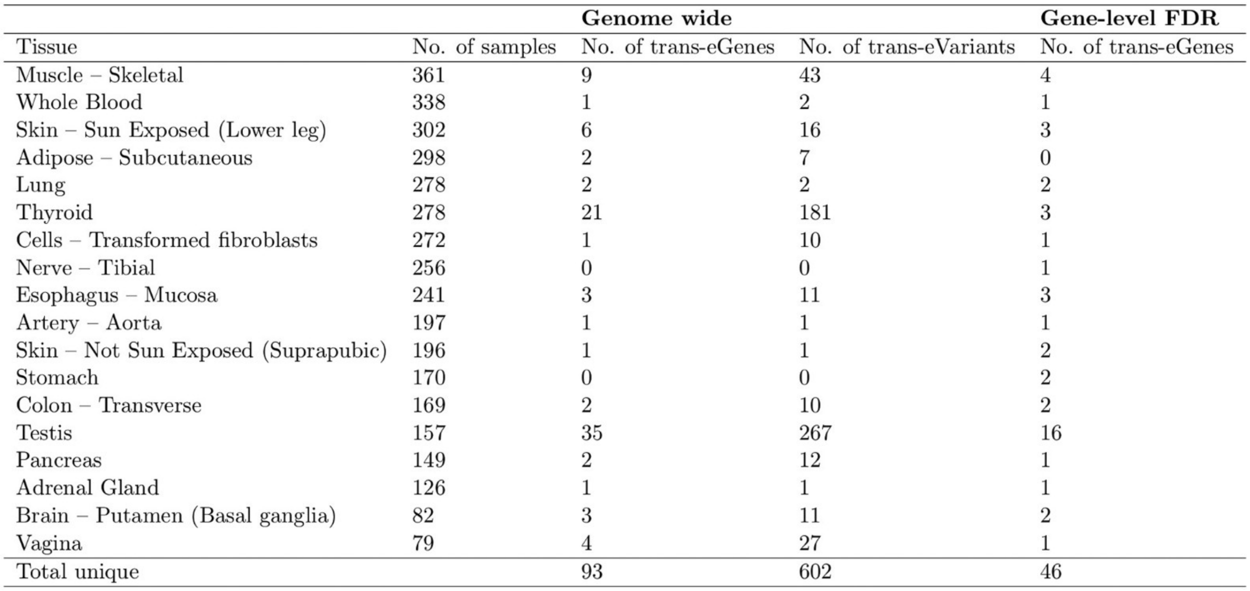


Figure 2-1: Illustration of GTEx v6p sample collection

Illustration of the 44 tissues and cell lines included in the GTEx v6p project with the associated number of cis- (left) and trans-eGenes (right) and sample sizes.

To identify trans-eQTLs, we tested for association between every protein-coding or lincRNA gene and all autosomal variants where the gene and variant were on different chromosomes. To minimize false positives in trans-eQTL detection, we controlled for the same observed and inferred confounders as in the *cis*-eQTL analysis, and further removed genes with poor mappability, variants in repetitive regions, and trans-eQTLs between pairs of genomic loci with evidence of RNA-seq read cross-mapping due to sequence similarity (see Methods). Applying this approach, we identified 673 trans-eQTLs at a 10% genome-wide FDR. This includes 112 distinct loci ($R^2 \leq 0.2$) and 93 unique genes (94 total gene associations, including a trans-eGene detected in both testis and thyroid) in 16 tissues. (Table 2-1).



		Genome wide		**Gene-level FDR**
Tissue	No. of samples	No. of trans-eGenes	No. of trans-eVariants	No. of trans-eGenes
Muscle – Skeletal	361	9	43	4
Whole Blood	338	1	2	1
Skin – Sun Exposed (Lower leg)	302	6	16	3
Adipose – Subcutaneous	298	2	7	0
Lung	278	2	2	2
Thyroid	278	21	181	3
Cells – Transformed fibroblasts	272	1	10	1
Nerve – Tibial	256	0	0	1
Esophagus – Mucosa	241	3	11	3
Artery – Aorta	197	1	1	1
Skin – Not Sun Exposed (Suprapubic)	196	1	1	2
Stomach	170	0	0	2
Colon – Transverse	169	2	10	2
Testis	157	35	267	16
Pancreas	149	2	12	1
Adrenal Gland	126	1	1	1
Brain – Putamen (Basal ganglia)	82	3	11	2
Vagina	79	4	27	1
Total unique		93	602	46

Table 2-1:Number of identified trans-eQTLs

Each tissue with non-zero values is included as a row; the columns include the number of samples for that tissue, followed by the number of unique trans-eGenes and trans-eVariants identified in the genome-wide tests, and the number of unique trans-eGenes found using gene-level FDR calibration. Ultimately, the set of 673 trans-eQTLs identified in the genome-wide approach yielded 602 unique trans-eVariants.

The extensive and diverse tissue sampling allowed us to develop a global view of how genetic effects vary between tissues of the human body by evaluating the sharing of eQTLs across tissues. We performed a meta-analysis across all 44 tissues for both cis- and trans-eQTLs to assess eQTL sharing between tissues. To do so, we applied Meta-Tissue (11), a linear mixed model that allows for heterogeneity in effect sizes across tissues and controls for correlated expression measurements that result from collecting multiple tissues from the same donors. For each eQTL, we estimated the posterior probability that the effect is shared in each tissue (*m* value). For both *cis*- and trans-eQTLs, we observed patterns that reflected relationships between related tissues and concordance between cis and trans in estimates of tissue similarity (Figure 2-3a). The strongest broad pattern observed was the high correlation among brain tissues (median Spearman's *ρ* of 0.584 (cis) and 0.241 (trans)) and among non-brain tissues (median Spearman's *ρ* of 0.606 (cis) and 0.165 (trans)), with much lower correlation observed between these two groups (median Spearman's *ρ* of 0.499 (cis) and 0.096 (trans)). Within non-brain tissues, we observed strong correlation among closely related tissues, such as arterial tissues (median Spearman's *ρ* of 0.743 (cis) and 0.264 (trans)), skeletal muscle and heart tissues (median Spearman's *ρ* of 0.672 (cis) and 0.184 (trans)), and skin tissues (Spearman's *ρ* of 0.804 (cis) and 0.365 (trans)). Overall, the median pairwise correlation between tissues was 0.547 (cis) and 0.138 (trans).

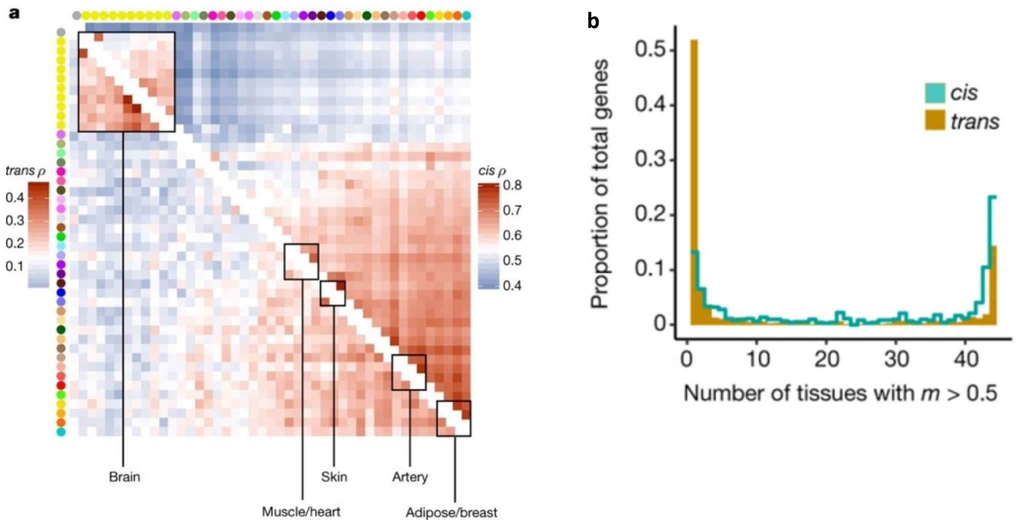


Figure 2-2: Tissue-specificity of eQTLs

(a) Similarity (Spearman's ρ) of Meta-Tissue effect sizes between tissues for cis- (upper triangle, 5% FDR) and trans- (lower triangle, 50% FDR) eQTLs. Tissues (by colours as in Figure 2-1) are ordered by

agglomerative hierarchical clustering of the cis-eQTL results. (b) Distribution of the number of tissues having Meta-Tissue m > 0.5 for the top variant for each trans-eGene at 50% FDR, and FDR-matched, randomly selected cis-eGenes (also 50% FDR). cis-eGenes were matched for discovery tissue to the trans-eGenes.

Overall, we observed much greater tissue specificity for trans-eQTLs than a set of FDR-matched *cis*-eQTLs (Figure 2-2b); this observation was robust to choices of *m* value threshold and selection criteria for matching *cis*-eQTLs (Appendix A: Figure S1a-d). While 3.8% of *trans*-eQTLs were shared across three or more tissues at $m > 0.9$, 25.3% of FDR-matched *cis*-eQTLs were shared. Our estimate of increased tissue specificity for trans-eQTLs agreed with the minimal sharing of *trans* effects reported in previous eQTL studies with fewer tissues (15), and greatly exceeds what would be expected on the basis of replication between tissues for cis-eQTLs of matched minor allele frequency (MAF) and effect size (Wilcoxon rank sum test; $P \leq 2.2 \times 10^{-16}$ for all choices of replication FDR; Appendix A: Figure S1e).

Next, we quantified the enrichment of trans-eVariants in promoter and enhancer regions using tissue-specific annotations from the Roadmap Epigenomics project (12) (Appendix A: Supplementary table 1). trans-eVariants (10% FDR) were enriched in cell-type matched enhancers (median Fisher's exact test, $P \leq 2.2 \times 10^{-3}$) but not strongly enriched for promoters (median $P \leq 0.22$), compared to randomly selected variants matched by distance to nearest TSS, MAF and chromosome (Figure 2-4). trans-eVariants were more enriched than *cis*-eVariants at matched FDR (Wilcoxon rank sum test, promoter: $P \leq 4.6 \times 10^{-7}$; enhancer: $P \leq 2.2 \times 10^{-16}$). Stronger effect sizes are needed to detect trans-eVariants at the same FDR, but even comparing to a matched number of the strongest cis-eVariants, we observed greater enrichment in enhancer (but not promoter) regions among trans-eVariants, consistent with greater tissue-specificity of enhancer activity and trans-eVariants (16) (Figure 2-3).

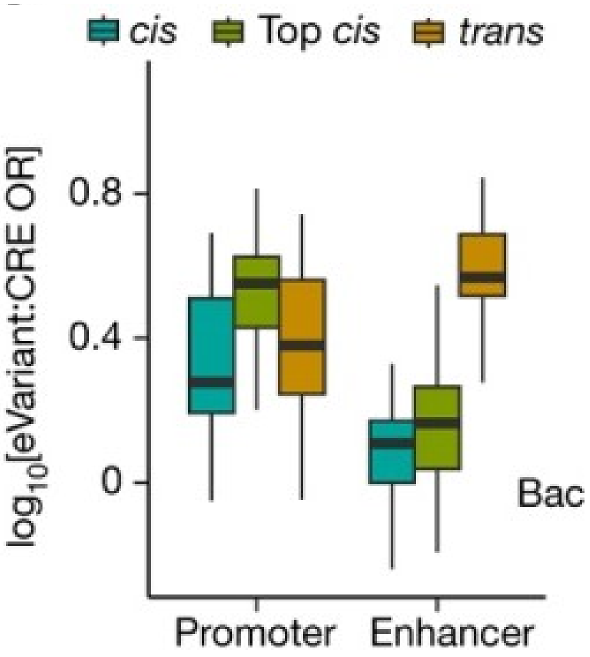


Figure 2-3: cis regulatory element (CRE) enrichment of eQTLs

CRE enrichment (y-axis) of trans-eVariants (10% FDR), cis-eVariants (10% FDR, to match trans-eVariants), and top most significant cis-eVariants. Box plots show promoter and enhancer enrichment (x-axis) in matched cell-type CRE annotations compared to MAF- and distance-matched background variants

Methods

cis-eQTL mapping

We conducted *cis*-eQTL mapping within the 44 tissues with at least 70 samples each. Only genes with ten or more donors with expression estimates > 0.1 RPKM and an aligned read count of six or more within each tissue were considered significantly expressed and used for cis-eQTL mapping. Within each tissue, the distribution of RPKMs in each sample was quantile-transformed using the average empirical distribution observed across all samples. Expression measurements for each gene in each tissue were subsequently transformed to the quantiles of the standard normal distribution. The effects of unobserved confounding variables on gene expression were quantified with PEER (14), run independently for each tissue. Fifteen PEER factors were identified for tissues with fewer than 150 samples; 30 for tissues with sample sizes between 150 and 250; and 35 for tissues with more than 250 tissues.

Within each tissue, cis-eQTLs were identified by linear regression, as implemented in FastQTL (17), adjusting for PEER factors, sex, genotyping platform, and three genotype-based principal components (PCs). We restricted our search to variants within 1 Mb of the TSS of each gene and, in the tissue of analysis, minor allele frequencies ≥0.01 with the minor allele observed in at least 10 samples. Nominal *P* values for each variant–gene pair were estimated using a two-tailed *t*-test. The significance of the most highly associated variant per gene was determined from empirical *P* values, extrapolated from a Beta distribution fitted to adaptive permutations with the setting –permute 1000 10000. These empirical *P* values were subsequently corrected for multiple testing across genes using Storey's *q* value method (18). To identify the list of all significant variant–gene pairs associated with eGenes, variants with a nominal *P* value below the gene-level threshold were considered significant and included in the final list of variant–gene pairs.

trans-eQTL mapping

Matrix eQTL (19) was used to test all autosomal variants (MAF > 0.05) using the same expression filters as cis-eQTL mapping, but restricted to variants and genes lying on different chromosomes, in each tissue independently using an additive linear model. For trans-eQTL mapping, we tested variants for association with expression of only protein coding or lincRNA genes. We included as covariates the three genotype PCs, genotyping platform, sex, and PEER factors estimated from expression data in Matrix eQTL when performing association testing. The correlation between variant and gene expression levels was evaluated using the estimated *t* statistic from this model, and corresponding FDR was estimated using Benjamini–Hochberg FDR correction (20) separately within each tissue and also using permutation analysis. For all trans association tests, we applied stringent quality control to account for potential false positives due to RNA-seq read mapping errors, repeat elements, and population stratification.

Functional enrichment

We annotated discovered eVariants using chromatin state predictions from 128 cell types or cell lines sampled by the Roadmap Epigenomics project (12). Genome segmentation was performed for each cell type or cell line using a 15-state hidden Markov model (HMM) over 400 bp windows. Several of the learned states are labelled as enhancers, promoters, and repressed regions. For the standard 15-state Roadmap segmentations, regulatory elements are labelled independently for each cell type. For enrichment analyses, we constructed background variants sets that matched eVariants to randomly selected variants based on chromosome, distance to nearest TSS, and MAF.

trans-eQTL analysis was restricted to protein-coding genes and to GTEx tissues that are composed of at least one Roadmap Epigenomics cell type (26 tissues), which included 85 eVariants and 23 eGenes (10% FDR). We quantified enrichment of the *trans* variants relative to random variants in both enhancer and promoter elements in the GTEx discovery tissue's matched Roadmap cell type (Appendix A: Supplementary table 1). We then performed the same analysis with randomly matched cis-eGenes. Matching cis-eGenes were selected as follows: for each of the 23 trans-eGenes *g*, each having N_g associated eVariants (10% FDR), we randomly selected a cis-eGene that also had at least N_g associated variants (10% FDR). We then selected the top N_g variants associated with this gene based on *P* value. We then performed the same analysis using random sets of the strongest cis-eGenes, rather than random eGenes. Matching the strongest cis-eGenes was performed as follows: for each of the 23 trans-eGenes *g*, each having N_g associated eVariants (10% FDR), we randomly selected a cis-eGene amongst the ten strongest cis-eGenes in that tissue, based on the *P* value of the strongest associated variant that also had at least N_g associated variants (10% FDR). We then selected the top N_g associated variants with this gene based on *P* value. Selecting 23 random *cis*-eGenes a single time yields unstable results, so we ran cis-eGene selection and enrichment 70 times with different selections. This was done for both random cis-eGenes and random selections amongst the strongest cis-eGenes. We rank-ordered the 70 trials for both promoters and enhancers based on average odds ratio enrichment relative to background. We then used the trial that was closest to median rank for plotting both promoter and enhancer enrichment results.

Discussion

Since the initial sequencing of the human genome, extensive effort has been devoted to the characterization of genome function and phenotypic consequences of genetic variation. Describing the effects of genetic variation on gene expression levels across tissues is a critical but challenging component of this goal. Here, we describe advances enabled by the GTEx project v6p data, which provide a comprehensive survey of gene expression and the impact of genetic variation on gene expression across diverse human tissues. We report widespread *cis*-eQTLs in 44 tissues and *trans*-eQTLs in 18 tissues. *cis*-acting genetic variants tend to affect either most tissues or a small number of tissues. By contrast, identified *trans*-eQTL effects tend to be tissue-specific and correspondingly show greater enrichment in enhancer regions.

This chapter provides a quantitative analysis of how genetic regulation of gene expression changes across tissues in the human body. Yet, this analysis was limited to assaying gene expression from tissues at a single time point, and does not consider the developmental trajectory of the cell types constituting the assayed tissue. In Chapter 3, we evaluate how dynamic gene expression data can add another dimension to eQTL analysis

Chapter 3 Dynamic genetic regulation of gene expression during cellular differentiation

Contributions

This project resulted from joint work with Reem Elorbany and Katie Rhodes. Reem and Katie performed the experiments, and I performed the computational data analysis. Karl Tayeb developed split-GPM with my input. This project was jointly supervised by Yoav Gilad and Alexis Battle. The work described in this chapter was published in (21). The text of this chapter is a slight modification of the published work.

Abstract

Genetic regulation of gene expression is dynamic, as transcription can change during cell differentiation and across cell types. We mapped expression quantitative trait loci (eQTLs) throughout differentiation to elucidate the dynamics of genetic effects on cell type specific gene expression. We generated time-series RNA-sequencing data, capturing 16 time points from induced pluripotent stem cells to cardiomyocytes, in 19 human cell lines. We identified hundreds of dynamic eQTLs that change over time, with enrichment in enhancers of relevant cell types. We also found nonlinear dynamic eQTLs, which affect only intermediate stages of differentiation, and cannot be found by using data from mature tissues. These fleeting genetic associations with gene regulation may represent a new mechanism to explain complex traits and disease. We highlight one example of a nonlinear eQTL that is associated with body mass index.

Introduction

Genetic variants that alter gene regulation play an essential role in the genetics of human disease and other complex phenotypes (2,4). Large studies have identified thousands of genetic loci associated with complex diseases, most of which are in non-coding regions of the genome and therefore are putatively involved in gene regulation (*2*). Expression quantitative trait locus (eQTL) analysis has shown that many disease-associated loci influence the regulation of nearby genes (5, 22) but still, a substantial fraction of disease-associated loci remain unexplained (*6*).

Much effort has been dedicated to map and identify eQTLs across tissues and cell types, as regulatory impact of disease-associated loci may be most evident in cell types relevant to each disease. Regulatory genetic effects can be also timepoint-specific or environment-dependent (*7,10*), and may influence temporal programs of gene regulation. Yet, almost all studies of the genetics of gene regulation, including the multi-tissue GTEx project described in Chapter 2 (*10*), involve data collected at a single time point, usually from adult individuals. Dynamic gene expression data can add another dimension to eQTL analysis, allowing identification of genetic variants with transient effects that may not have been found in analysis of static data.

Results

We took advantage of a panel of induced pluripotent stem cell (iPSC) lines from 19 individuals to investigate high-resolution temporal genetic effects on gene regulation over time during cardiomyocyte differentiation. Specifically, we collected gene expression data throughout the differentiation from iPSCs to cardiomyocytes in 19 well-characterized, human Yoruba HapMap cell lines (23). For each cell line, RNA was extracted and sequenced every 24 hours for 16 days, to capture the entire differentiation process; in total, we sequenced 297 RNA samples (Appendix B: Figure S1-2). Combined with available whole genome sequences and genotype data for each cell line, these data provide a resource with which to investigate how gene expression and genetic regulation change throughout cardiomyocyte differentiation with high temporal resolution.

Quality controls and filtering yielded 16,319 genes for downstream analysis (see Methods). Following standardization and normalization of the RNA sequencing data (see Methods), we evaluated the contribution of potential confounders to overall variation in our data, confirming that our study design was effective (Appendix B: Figure S3). We also used replicates from an independent differentiation to confirm that the gene expression patterns we observed in our iPSCs and iPSC-derived cardiomyocytes are robust with respect to variance that may be associated with the differentiation procedure (Appendix B: Figure S4) (23) (see Methods).

We evaluated the efficiency of our differentiation by FACS (Appendix B: Supplementary table 1), and by considering the time course expression of known cell type specific marker genes (24,25) (Appendix B: Figure S5). As expected, cardiomyocyte purity and the expression of lineage marker genes are variable across our samples. This variability between cell lines was observed across the entire time course, though the effect of differentiation time is the primary source of variation in the data (Figure 3-1A, Appendix B: Figures S3, S6).

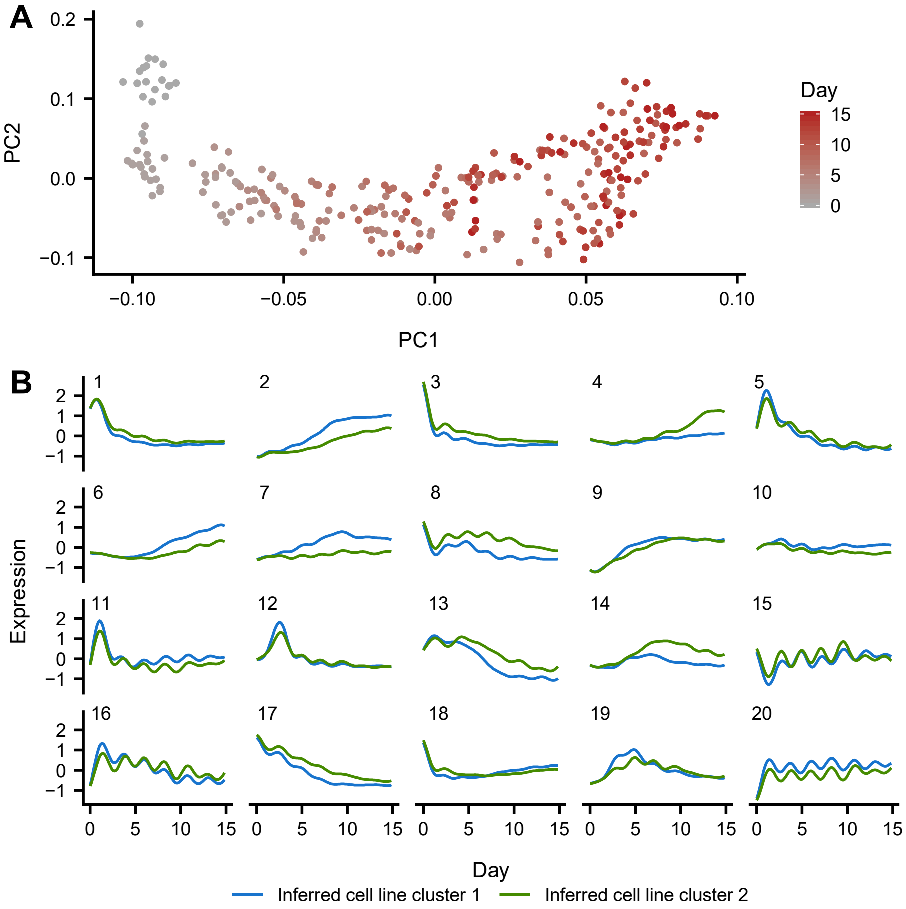


Figure 3-1: Gene expression trends throughout cardiomyocyte differentiation

(A) The first two gene expression principal component loadings for all 297 RNA-seq samples across cell lines, where each sample is colored by day of collection. (B) Predicted cell line cluster expression trajectories for 20 gene clusters according to split-GPM. Many gene clusters (8, 11, 15, 16, and 20) exhibit periodic expression trajectories that correspond with cell culture media changes.

We characterized global patterns of gene expression across time by applying split-GPM, an unsupervised probabilistic model that infers time course trajectories of gene expression using Gaussian processes, while simultaneously performing clustering of genes and cell lines (see Methods). Using this approach, we identified two clusters of cell lines that displayed broad differences in the expression patterns of multiple clusters of genes; within each gene cluster, genes exhibit shared expression changes over time. The assignment of cell lines to clusters is robust with respect to the parameters we tested, such as the number of gene clusters we infer (Appendix B: Figure S7).

The two cell line clusters we identified differ in the efficiency of cardiomyocyte differentiation. Cell lines in the first (larger) cluster display greater Troponin expression levels in the final six timepoints of differentiation (p=.014, Wilcoxon rank-sum test). The expression of a group of genes enriched for myogenesis also increases by a greater magnitude over time in cell lines in the first cluster (Bonferroni p=9.29e-14; gene cluster 2 in Figure 3-1B) (26). Cell lines in the second, smaller cluster, show high expression of genes related to KRAS activation (Bonferroni p=0.005; gene cluster 4 in Figure 3-1B), which is associated with increased self-renewal of undifferentiated iPSCs and decreased neuronal differentiation propensity (27). Other gene clusters illuminate broad changes in gene expression over time such as a transient rise in MYC and E2F target genes in the early days of differentiation (gene cluster 13 in Figure 3-1B; Appendix B: Supplementary table 2). Together, this analysis documents patterns of gene expression trajectories over time and captures differences among our cell lines that are not obvious from the individual time point data alone.

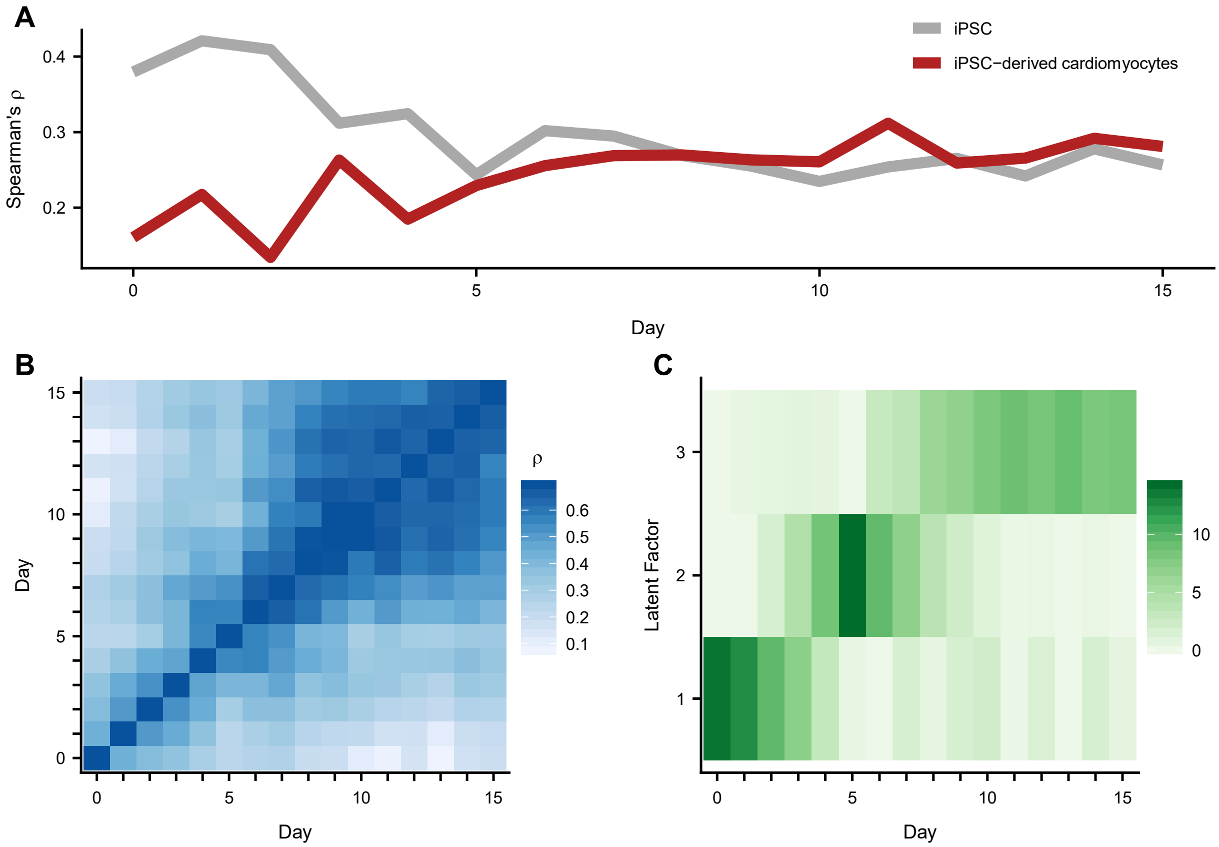


Figure 3-2: eQTL patterns during cardiomyocyte differentiation

We limit to genes with at least one significant eQTL (WASP combined haplotype test; eFDR <= .05) across time points. If a gene has more than one significant eQTL, we select a single variant for that gene with the smallest geometric mean p-value across all 16 time points. (A) Spearman correlation of p-values between eQTLs from each day (x-axis) and existing iPSC (grey) and iPSC-derived cardiomyocyte (red) eQTLs. (B) Spearman correlation of eQTL p-values for each pair of days. (C). Factors identified via sparse matrix factorization of eQTL -log$_{10}$ p-values using 3 latent factors and a L1 penalty of .5.

Next, we evaluated the impact of genetic variation on gene regulation in our system. We used WASP (28) to identify cis-eQTLs in the data from each time point, independently (see Methods). To control for latent confounders in the independent analysis of data from each time point, we included the first three expression PCs using data from samples of the corresponding time point as covariates (Appendix B: Figures S8, S9). At an empirical false discovery rate (eFDR) of 5%, we identified a median of 111 genes (range 71 – 231) with at least one eQTL in each time point (Appendix B: Figures S9, S10). As expected, the eQTLs we identified early in the time course replicated in data from iPSCs, whereas eQTLs from later time points were better supported by data from iPSC-derived cardiomyocytes (both $p < 0.001$, linear regression; Figure 3-2A) (23).

We computed the correlation of the significant eQTL summary statistics for each pair of time points (Figure 3-2B). We observed that correlation between eQTL summary statistics increases as the distance between time points decreases ($p <= 2e-16$, linear regression). Though this observation is intuitive, it indicates that the dynamic impact of genetic variation on gene regulation in our data is not random, and is related to the temporal process of cardiomyocyte differentiation.

To more formally quantify the temporal structure of genetic regulation throughout differentiation, we performed sparse non-negative matrix factorization on the matrix of significant eQTL summary statistics from all time points (see Methods). The learned factors capture genetic signal that is largely specific to a subset of differentiation time (Figure 3-2C), a pattern that is robust with respect to the number of latent factors or sparse prior choice (Appendix B: Figure S11).

Our analysis indicates that temporal structure dominates the patterns of genetic association with gene expression in our data. However, the observation that most significant non-dynamic eQTLs can be identified in only a few time points (median of 2; Appendix B: Figure S12) is most likely explained by incomplete power to identify eQTLs in each time point independently. To robustly identify dynamic eQTLs whose effect varies significantly over time, leveraging power across all time points (Figure 3-3A), we used a Gaussian linear model applied jointly to data from the entire experiment. Specifically, we quantified the effect of interactions between genotype and differentiation time on gene expression, controlling for linear effects of both differentiation time and genotype. In addition, we accounted for the systematic differences in differentiation trajectories identified between cell lines (Figure 3-1B, Appendix B: Figures S13-S16, Supplementary table 3) (see Methods), which would otherwise lead to false positives in our analysis. Using this approach, we identified 550 genes with a significant dynamic eQTL (eFDR <= .05; Appendix B: Figures S17-S20, Supplementary table 4).

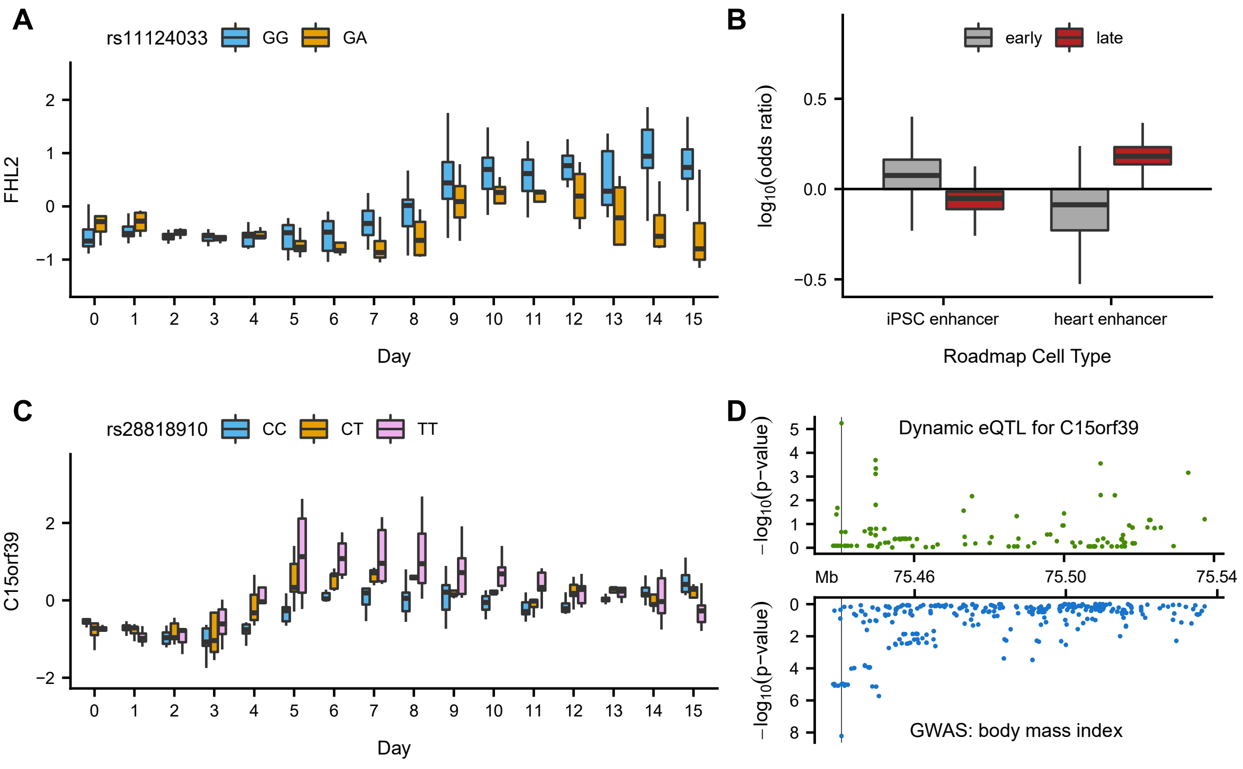


Figure 3-3: Dynamic eQTLs detect genetic regulatory changes caused by cardiomyocyte differentiation. (A) Linear interaction association between genotype (color) of rs11124033 and time point (x-axis) on residual gene expression (cell line effects regressed on expression) of FHL2 (y-axis). (B) Enrichment of dynamic eQTLs within cell type specific chromHMM enhancer elements relative to 1000 sets of randomly selected matched background variants. Dynamic eQTLs were classified as early or late (C) Nonlinear interaction association between genotype (color) of rs28818910 and time point (x-axis) on residual gene expression of C15orf39 (y-axis). (D) Nonlinear interaction association significance of all variants tested within 50 KB of the C15orf39 transcription start site with expression of C15orf39 (green) and GWAS significance for BMI of variants in the same window (blue). Vertical line depicts genomic location of the most significant nonlinear dynamic eQTL (rs28818910) for C15orf39.

We classified the 550 dynamic eQTL as *early* (eQTL effect size decreasing over time), *late* (eQTL effect size increasing over time), or *switch* (eQTL effect size exhibiting different directions of effect over time; Appendix B: Figure S21) (see Methods). We found that the early dynamic eQTLs are enriched for chromHMM enhancer elements annotated in iPSC Roadmap cell types but not in heart-related cell types (29). In turn, late dynamic eQTLs are enriched for chromHMM enhancer elements annotated in heart-related Roadmap cell types but not in iPSCs (Figure 3-3B, Appendix B: Figure S22). These observations indicate that dynamic eQTL mapping can capture temporal changes in cellular gene regulation reflecting changes in regulatory element activity as the cell cultures differentiate.

The observation that we are able to capture the function of cell-type-specific regulatory elements prompted us to consider dynamic eQTLs in other contexts. We found that dynamic eQTLs are enriched for genes with roles in myogenesis (Bonferroni p = .0019, Fisher's exact) (26), and also show significant enrichment for genes related to dilated cardiomyopathy (p = .001, Fisher's exact) (see Methods) (30). Two significant dynamic eQTLs in particular, rs7633988 and rs6599234 (in strong LD, $R^2 = 0.93$), are GWAS variants for QRS duration and QT interval, respectively (Appendix B: Figure S23) (31, 32). Both variants show an association with the expression levels of *SCN5A*, which is involved in the creation of sodium channels and is in the dilated cardiomyopathy gene set (33). Another dynamic eQTL, rs11124033, associated with the expression of *FHL2* (Figure 3-3A), is also associated with dilated cardiomyopathy. This variant lies in a Roadmap chromHMM promoter element annotated in heart-related cell types but not in iPSCs (29). Interestingly, none of these examples were identified as eQTLs in the non-dynamic QTL analysis of each time point from our dataset or in the GTEx heart tissue data (10).

Finally, we sought to identify a wider range of dynamic regulatory patterns, including nonlinear associations such as when a genetic effect increases in magnitude in the middle of the time course before decreasing or disappearing. To identify nonlinear dynamic eQTLs we expanded our linear model using a second order polynomial basis function (see Methods). We acknowledge that our study is underpowered to expand to a more general class of nonlinear dynamic eQTLs that do not assume a continuous effect of differentiation time (Appendix B: Figure S24) (see Methods).

We identified 693 genes with a nonlinear dynamic eQTL (eFDR <= .05; Appendix B: Figures S17B, S19B), 28 of which have their strongest genetic effect in the middle of the differentiation time course (middle dynamic eQTLs; Appendix B: Figures S25) (see Methods). It is worth noting that 25 of these middle dynamic eQTL genes and their strongest associated variant are not identified as eQTLs in our non-dynamic QTL analysis in either iPSCs (day 0) or cardiomyocytes (day 15).

In one example of a non-linear dynamic eQTL, rs8107849 is associated with the expression of *ZNF606* with a larger magnitude of effect during days 4 through 11 (Appendix B: Figure S26). The rs8107849 locus does not lie in iPSC or heart-related chromHMM regulatory regions and was not identified in our analysis as a non-dynamic eQTL in any time point. While *ZNF606* is known to have a role in differentiation of chondrocytes (34), it is possible this is a conserved process involved in the differentiation of additional cell types, including cardiomyocytes. Another nonlinear dynamic eQTL reveals an association between rs28818910 and *C15orf39*. The rs28818910 variant is also associated with BMI (p < 6.07 e-9, reported; Figures 3-3C, 3-3D) (35) and weakly associated with red blood cell count (p < 1.48 e-6, reported) (35). This dynamic eQTL and both traits show similar patterns of association across the region (Appendix B: Figure S27). The rs28818910 locus is associated with inter-individual differences in gene expression only during intermediate stages of differentiation; it does not lie in annotated regulatory elements of either iPSCs or cardiomyocytes and is not identified as an eQTL in iPSCs, mature cardiomyocytes, or either of the two GTEx heart tissues. Thus, this is an example of a temporary, dynamic regulatory effect that may have phenotypic consequences.

Methods

Genotype data

We used previously collected and imputed genotype data for the 19 Yoruba individuals from the HapMap and 1000 Genomes Project.

RNA-seq quantification

All RNA-seq samples were aligned to the human genome (GRCh37) using Subread. We counted reads and estimated gene level expression with reads per kilobase million (RPKM) using the `edgeR` R package. We then filtered to genes that were protein-coding, autosomal, and had at least 10 samples such that RPKM >= .1 and raw read counts >= 6. This yielded 16,319 genes. The RPKM distribution in each sample was then quantile normalized and each gene, across all samples, was standardized (mean 0, standard deviation 1).

Biological replication

We computed replication of day 0 cell lines within previously generated iPSC lines (*9*) and replication of day 15 cell lines within previously generated iPSC-derived cardiomyocyte cell lines (23). Notably, the samples from Banovich et al. were also generated in the Gilad lab and use the same panel of iPSCs. Count data from all 4 data sets was re-processed under a uniform pipeline:

1. Count data was $\log_2$(count+1) transformed
2. Each gene was standardized to have mean zero and standard deviation 1
3. Top gene expression PCs (in each data set separately) were regressed out.

We regressed out the top 3 PCs in the day 0 and day 15 data sets, top 10 PCs in the Banovich et al iPSC data set, and top 3 PCs in the Banovich et al. iPSC-derived cardiomyocyte data set. The choice of 3 PCs was selected to match the number of PCs in the non-dynamic eQTL analysis. The choice of 10 PCs in the Banovich et al. iPSC data set was selected to match their analysis.

Cell line clustering model (split-GPM)

We applied a generative model that assumes a joint clustering over the 19 cell lines and 16,319 genes. That is, the model encodes a global assignment of each of G genes to L gene clusters and assignment of each of N cell lines to K cell line clusters. For each cell line cluster, each gene cluster specifies a Gaussian process (GP) representing a latent gene expression trajectory across time. Thus, the model identifies groups of cell lines with globally different behavior, and groups of genes with similar expression trajectories within each cell line cluster.

Let y_{ng}be the observed gene expression trajectory for gene g in cell line n at times t_{ng}. Our observations are generated as follows:

$$\Phi_n \sim Categorical(\pi)$$

$$\Lambda_g \sim Categorical(\psi)$$

$$f^{kl} \sim GP\big(0, K(\theta)\big)$$

$$y_{ng} | \ \Phi_n = k, \Lambda_g = l, f^{kl}\ , t_{ng} \sim N(f^{kl}\big(t_{ng}\big), \sigma^2 I)$$

$\pi \in R^K \geq 0\ s.t. \sum_{k=1}^{K} \quad \pi_k = 1,\ \ \psi \ \in R^L \geq 0\ s.t. \sum_{l=1}^{L} \quad \psi_l = 1$ are cell line cluster mixture weights and gene cluster mixture weights respectively, θ are GP kernel hyperparameters and σ^2 is a global variance parameter. f^{kl} is a function drawn from a gaussian process, while $f^{kl}(t)$ is the function evaluated at points t.

We collect $\{\Phi_n\}_{n=1,\ldots N}$ into an NxK binary matrix $\Phi\ s.t. \Phi_{nk} = 1 \ \Leftrightarrow \ \Phi_n = k.$ Likewise, we collect $\{\Lambda_g\}_{g=1\ldots G}$ into a GxL binary matrix $\ s.t.\ \Lambda_{gl} = 1 \ \Leftrightarrow \lambda_g = l$. The observed data points are conditionally independent given the functions and assignments. Our full likelihood is:

$$p(\{y_{ng}\} |\ \{f^{kl}\}, t_{ng}, \Phi, \Lambda) = \prod_{n,g,k,l}^{N,G,K,L} \quad N(y_{ng} | f^{kl}\big(t_{ng}\big), \sigma^2)^{1(\Phi_{nk})1(\Lambda_{gl})}$$

split-GPM approximate inference

Exact computation of the posterior $p(\{f^{kl}\}, \Phi, \Lambda, | \{y_{ng}\}, \{t_{ng}\})$ is intractable so we resort to a variational approximation that factorizes and minimizes the KL-divergence of the true posterior:

$$q(\{f^{kl}\}, \Lambda, \Gamma) = \prod_{k,l}^{K,L} \quad q(f^{kl}) \prod_{n}^{N} \quad q(\Phi_n) \prod_{g}^{G} \quad q(\Lambda_g)$$

$$f^{kl} \sim GP(0, K(\theta))$$

$$\Phi_n \sim Categorical(\widehat{\Phi}_n)$$

$$\Lambda_g \sim Categorical\left(\widehat{\Lambda_g}\right)$$

To update the assignments, we iteratively update Φ and Λ until convergence or until a fixed number of iterations is reached.

$$ELBO(q) = E_q[log\, p(\{\{y_{ng}\}|\{f^{kl}\}, \{t_{ng}\}, \Phi, \Lambda)] + E_q[log\, p(\{f^{kl}\}, \Phi, \Lambda) \\ - \; E_q[log\, q(\{f^{kl}\}, \Phi, \Lambda)]$$

We iteratively estimate assignment variables and trajectory estimates, then perform gradient based optimization with respect to the kernel parameters. This approximation requires $K \cdot L$ GP regressions, each computed over every data point. To make the problem tractable we further approximate each GP via SVGP. In this analysis, we train a model with $K = 2$ cell line clusters, $L = 20$ gene clusters and an RBF kernel with shared length-scale and variance parameters for all $K \cdot L$ clusters.

Non-dynamic cis-eQTL calling per time point

Separately, each time point has a small sample size (maximum of 19 samples). Therefore, we used the WASP combined haplotype test (CHT) to increase power, integrating both total expression and allelic imbalance data into the same test, to detect cis-eQTLs in each of the 16 time points, independently. In order to increase accuracy of allele-specific expression estimates, RNA-seq data was re-quantified for eQTL calling by filtering Subread mapped reads using the WASP mapping pipeline under default settings in order to reduce biases in allelic mapping. We tested cis-eQTL association for variants within 50 KB of each gene's transcription start site. Further, we tested the same set of variant-gene pairs in all time points, limiting to variant-gene pairs that passed the following filters in all 16 time points:

1. Variant has minor allele frequency >= .1
2. Gene passes all filters described in "RNA-seq quantification" section
3. Gene has >= 100 reads mapped summed across all cell lines
4. Exon of the gene contains a heterozygous variant in at least 5 cell lines
5. Sum of reads mapping to minor allele across all cell line, heterozygous variant pairs >= 25

These filters yielded 1,009,173 variant-gene pairs (6,362 unique genes) tested in each time point. The same variant-gene pairs were tested in each time point to reduce bias when comparing genetic regulatory effects between time points. We included the first three raw read count expression PCs from samples belonging to the corresponding time point as covariates. The choice to control for three PCs was motivated by maximizing the number of significant non-dynamic eQTLs detected in each time step (Appendix B: Figure S9B). We ran one permutation of the CHT genome-wide. It is worth noting that the CHT is not well calibrated (Appendix B: Figure S10**)**. Multiple testing correction was performed using empirical FDR (eFDR) to assess genome-wide significance based on a vector of observed p-values and a vector of null (permuted) p-values. An empirical approach to FDR correction should account and control for the lack of calibration observed when the CHT was applied to our data.

Sparse non-negative matrix factorization

We performed sparse, non-negative matrix factorization of eQTL statistics for all time points to identify broad patterns in eQTL effects. Here, we limited to genes with at least one significant eQTL (eFDR <= .05) across time points. If a gene had more than one significant eQTL, we selected a single variant for that gene with the smallest geometric mean p-value across all 16 time points. We then filled in a matrix, X, where each row represents one gene, each column represents a time point, and each element represents the -$\log_{10}$ p-value corresponding to the row's gene and the column's time point. We then performed sparse non-negative matrix factorization on X (dim NxT) using the python function `sklearn.decomposition.NMF`. With K latent factors, this will reduce X into the product of a loadings matrix (L; dim NxK) and a factor matrix (F; dim KxT). F captures shared patterns of eQTL effect sizes across time while L reflects which factors are relevant for each eQTL. All default settings were used except we set `l1_ratio=1` to enforce an element-wise L1 penalty. We ran this analysis for a range of number of latent factors and L1 penalties (alpha) (Appendix B: Figure S11).

Linear dynamic eQTLs

Linear dynamic eQTLs are cis-eQTLs whose effects are linearly modulated by differentiation time. We detected linear dynamic eQTLs with a gaussian linear model that quantified the interaction between genotype and differentiation time on gene expression, while controlling for the linear effects of both genotype and differentiation time. We also controlled for linear effects of the first five cell line collapsed PCs (see below) and, critically, the linear effects of the interaction between the first five cell line collapsed PCs and differentiation time.

We built a separate linear model for each tested variant-gene pair. Specifically, let t denote the time point of the current sample, c denote the cell line of the current sample, T denote the total number of time points, and C denote the total number of samples. $E \in R^{CxT}$denotes the standardized expression matrix for the current gene, $G \in R^{C}$denotes the dosage based genotype vector for the current variant, and $PC^{K} \in R^{C}$denotes the Kth cell line collapsed PC vector. We modeled the expression levels as follows:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 t + \beta_3 PC_c^1 + \beta_4 PC_c^1 t + \cdots + \beta_{11} PC_c^5 + \beta_{12} PC_c^5 t + \beta_{13} G_c t, \sigma)$$

We used R `lm` to quantify the significance of the interaction between genotype and time (β_{13}). We computed a null distribution by randomly permuting the time point variable that was used for the term capturing the interaction between genotype and time (β_{13}), while keeping the time point variable in all other terms not permuted. An independent permutation was used for every tested variant gene pair. Using this permutation run, we computed significance with eFDR.

We tested the same set of variant-gene pairs that was tested in the non-dynamic eQTL calling analysis. This was done to reduce bias when comparing non-dynamic eQTLs and dynamic eQTLs.

Cell line confounder estimation using cell line collapsed PCA

Different cell lines can display broadly different patterns of expression across the entire time course, including not only consistent shifts upward or downward in expression of subsets of genes, but different slopes and more generally different expression trajectory shapes (Figure 3-1B). Variability in slope is of particular concern for detection of dynamic eQTLs – if a subset of cell lines display different slopes over time for many genes, this would lead directly to false positive dynamic eQTLs. Specifically, these cell line subsets reflecting confounders could by chance correspond to the same grouping as genotype across numerous SNPs given the large number of SNPs compared to cell lines. This would then produce apparently large effect $\beta_{13}G_ct$ terms in the dynamic eQTL linear model, and thus numerous false positives. To combat this problem, we used a PCA-based approach we refer to as “cell line collapsed PCA” to identify broad, cell line specific patterns across the entire time course. To do so, we simply rearranged the gene expression matrix from the standard RNA-seq quantification (RPKM levels across 297 samples by 16,319 genes) such that each row was now expression from one cell line and each column was a gene at a single time point. We excluded time points that were not fully observed (days 2, 4, and 13) to avoid missing entries, yielding a final matrix of size 19 by 212,147 (Appendix B: Figure S13). After standardizing each column, we applied PCA to this matrix to learn a low dimensional representation. Here, each cell line has a shared loading across all time points, and PCs reflect trajectories across all genes, rather than a standard application of PCA with loadings for each sample (a cell line, time point pair).

To ensure that we effectively controlled for the potential confounding effects of cell lines displaying broad trajectory differences over time, we calculated the frequency at which each pair of cell lines share the same genotype across all significant dynamic eQTLs. As noted above, a confounder would cause subsets of cell line to have the same eQTL SNP genotype more often than expected by chance alone, corresponding to cell line clusters with broad differences. In fact, when we do not include cell line collapsed PC loadings in our model, we do see an abundance of such likely false positives (Appendix B: Supplementary table S3). After controlling for 5 cell line collapsed PCs, the cell lines do not share the same genotype across significant dynamic QTLs more often than background (Appendix B: Figure S16), confirming that cell line PCs help address confounding effects of individual cell line trajectories.

An alternative approach of using pseudo-time, rather than actual time in association testing, does not fully address the problem mentioned here – cell lines don't simply progress faster or slower along the same ultimate trajectory, but seem to deviate in a more complex pattern. Here, this pattern appears to correspond to cell type purity, but more generally, differentiation or any temporal response that follows branching trajectories that can't be captured by a single monotonic pseudo-time term could lead to similar false positives.

We controlled for the first five cell line collapsed PCs and their interaction with differentiation time when detecting both linear and nonlinear dynamic eQTLs. While there does not exist an optimal method to select the number of cell line collapsed PCs, we selected 5 cell line collapsed PCs that: (a) capture most of the variance in gene expression (Appendix B: Figure S14a), (b) ensure cell lines do not share the same genotype across significant dynamic QTLs more often than background (Appendix B: Figure S16), and (c) result in consistency between non-dynamic eQTLs and dynamic eQTLs (Appendix B: Figures S21, S25).

Simulating expression samples for linear dynamic eQTL power analysis

Using the same notation as defined in the "Linear dynamic eQTLs" section, we define the alternate model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t + \beta_3 (t * G_c), \sigma)$$

And the null model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t, \sigma)$$

For each setting of number of cell lines, t-statistic and minor allele frequency, we simulated 10,000 independent tests (variant-gene pairs) where a specified proportion of those tests follow the null and alternate models. We made the simplifying assumption that each cell line contained 16 time points (T=16). For each test:

1. The genotype vector (G_c) was randomly generated assuming a specified minor allele frequency. Specifically, both alleles of the variant were drawn independently and both alleles were forced to have the specified minor allele frequency
2. β_1 was randomly generated for each test from a separate gaussian distribution with mean 0 and standard deviation of .1
3. β_2 was randomly generated for each test from a separate gaussian distribution with mean 0 and standard deviation of .1
4. β_3 was equal to the t-statistic multiplied by σ. For convenience, σ was fixed to be .1
5. E_{ct} was randomly drawn
6. p-values were computed using the linear model described in the "Linear dynamic eQTLs" section excluding any fixed effects containing cell line collapsed PCs

Significance of simulated tests was assessed at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data).

Nonlinear dynamic eQTLs

To detect dynamic eQTLs whose effect size changes non-linearly with time, we used a second order polynomial basis function over time, which alters the above linear dynamic eQTL model as follows:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 t + \beta_3 t^2 + \beta_4 PC_c^1 + \beta_5 PC_c^1 t + \beta_6 PC_c^1 t^2 + \cdots + \beta_{16} PC_c^5 + \beta_{17} PC_c^5 t + \beta_{18} PC_c^5 t^2 + \beta_{19} G_c t + \beta_{20} G_c t^2, \sigma)$$

We quantify the joint effect of the two interaction terms between genotype and time (β_{19} and β_{20}) with a likelihood ratio test with two degrees of freedom using the R `lmtest` package. We computed a null distribution by randomly permuting the time point variable that was used for the two terms capturing the interaction between genotype and time (β_{19} and β_{20}), while keeping the time point variable in all other terms not permuted. An independent permutation was used for every tested variant gene pair. It is worth noting that the nonlinear dynamic eQTLs are not well calibrated (Appendix B: Figure S18**).** Using this permutation run, we computed significance using eFDR. An empirical approach to FDR correction should account and control for the observed lack of calibration of this test.

Simulating expression samples for nonlinear dynamic eQTL power analysis

Linear dynamic eQTLs allow us to capture dynamic eQTLs whose effect size changes linearly with differentiation time. Nonlinear dynamic eQTLs allow us to capture dynamic eQTLs whose effect size changes as a quadratic function of differentiation time. However, both of these approaches are unable to capture arbitrary nonlinear functions of differentiation time. A statistical test that could capture arbitrary nonlinear functions of differentiation time is an ANOVA analysis where time is fit as a factor with 16 levels (ANOVA eQTLs). Here, we simulate several nonlinear dynamic eQTLs and access detection power using three different dynamic eQTL methods:

1. Linear dynamic eQTLs
2. Nonlinear dynamic eQTLs
3. ANOVA dynamic eQTLs

Using a similar notation as defined in the “Linear dynamic eQTLs” section, we define the alternate model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t_{new} + \beta_3 (t_{new} * G_c), \sigma)$$

And the null model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t_{new}, \sigma)$$

Here, $t_{new}$is a transformation of t. We used four arbitrary transformations of t:

1. $t_{new} = t(t - 10)$
2. $t_{new} = t(t - 7)(t - 15)$
3. $t_{new} = \sin{(pi * \frac{t}{5})}$
4. $t_{new} = I[t > 7]$

Transformed differentiation time (t_{new}) was scaled to have the same standard deviation as the original values of differentiation time. For each setting of number of cell lines, t-statistic and time transformation, we simulated 10,000 independent tests (variant-gene pairs) where 30% of those tests follow the alternate model and 70% follow the null model. We made the simplifying assumption that each cell line contained 16 time points (T=16). For each test:

1. The genotype vector (G_c) was randomly generated assuming a minor allele frequency of .4. Specifically, both alleles of the variant were drawn independently and both alleles were forced to have a minor allele frequency of .4.
2. β_1 was randomly generated for each test from a separate gaussian distribution with mean 0 and standard deviation of .1
3. β_3 was equal to the t-statistic multiplied by σ. For convenience, σ was fixed to be .1
4. E_{ct} was randomly drawn
5. p-values were computed using the three statistical models described above

Significance of simulated tests was assessed at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data).

Linear dynamic eQTL classifications

We classified the linear dynamic eQTLs as *early* (when the eQTL effect size decreased over time), *late* (when the eQTL effect size increased over time), or *switch* (when the eQTL effect size changes sign over the time course. To do so, we computed predicted eQTL effect size at day 0 and day 15 according to the fitted linear dynamic eQTL model: Let $\hat{E}_{vg}(t = x, G = y)$ be the predicted expression (according to the fitted dynamic eQTL model) of gene g at time x for a sample with genotype dosage y for variant v. We defined the eQTL effect size ($\beta_{vg}(t = x)$) of variant v on gene g at time x as:

$$\beta_{vg}(t = x) = \hat{E}_{vg}(t = x, G = 0) - \hat{E}_{vg}(t = x, G = 2)$$

If the sign of $\beta_{vg}(t = 0)$ is equal to the sign of $\beta_{vg}(t = 15)$, we assigned that dynamic eQTL to:

1. early if $\left|\beta_{vg}(t = 0)\right| \geq \left|\beta_{vg}(t = 15)\right|$
2. late if $\left|\beta_{vg}(t = 0)\right| < \left|\beta_{vg}(t = 15)\right|$

If the sign of $\beta_{vg}(t = 0)$ is not equal to the sign of $\beta_{vg}(t = 15)$, we assigned that dynamic eQTL to:

1. early if $\left|\beta_{vg}(t = 0)\right| \geq \left|\beta_{vg}(t = 15)\right|$ and $\left|\beta_{vg}(t = 15)\right|$ < thresh
2. late if $\left|\beta_{vg}(t = 0)\right| < \left|\beta_{vg}(t = 15)\right|$ and $\left|\beta_{vg}(t = 0)\right|$ < thresh
3. switch if $\left|\beta_{vg}(t = 0)\right| \geq$ thresh and $\left|\beta_{vg}(t = 15)\right| \geq$ thresh

We assigned thresh = 1.

Nonlinear dynamic eQTL classifications

We classified the nonlinear dynamic eQTLs as early (when the eQTL effect size decreased over time), late (when the eQTL effect size increased over time), switch (when the eQTL effect size changes sign over the time course, or middle (when the eQTL is strongest in the middle of the time course). To do so, we computed predicted eQTL effect size at t=0, t=7.5, and t=15 according to the fitted nonlinear dynamic eQTL model:

$$\beta_{vg}(t = 0) = \hat{E}_{vg}(t = 0, G = 0) - \hat{E}_{vg}(t = 0, G = 2)$$

$$\beta_{vg}(t = 7.5) = \hat{E}_{vg}(t = 7.5, G = 0) - \hat{E}_{vg}(t = 7.5, G = 2)$$

$$\beta_{vg}(t = 15) = \hat{E}_{vg}(t = 15, G = 0) - \hat{E}_{vg}(t = 15, G = 2)$$

If $\beta_{vg}(t = 7.5) \geq \beta_{vg}(t = 0)$ and $\beta_{vg}(t = 7.5) \geq \beta_{vg}(t = 15)$, we assigned the dynamic eQTL to middle.

If the sign of $\beta_{vg}(t = 0)$ is equal to the sign of $\beta_{vg}(t = 15)$, we assigned that dynamic eQTL to:

1. early if $|\beta_{vg}(t = 0)| \geq |\beta_{vg}(t = 15)|$
2. late if $|\beta_{vg}(t = 0)| < |\beta_{vg}(t = 15)|$

If the sign of $\beta_{vg}(t = 0)$ is not equal to the sign of $\beta_{vg}(t = 15)$, we assigned that dynamic eQTL to:

1. early if $|\beta_{vg}(t = 0)| \geq |\beta_{vg}(t = 15)|$ and $|\beta_{vg}(t = 15)| <$ thresh
2. late if $|\beta_{vg}(t = 0)| < |\beta_{vg}(t = 15)|$ and $|\beta_{vg}(t = 0)| <$ thresh
3. switch if $|\beta_{vg}(t = 0)| \geq$ thresh and $|\beta_{vg}(t = 15)| \geq$ thresh

We assigned thresh = 1.

ChromHMM enrichment analysis

We computed enrichment of dynamic eQTLs within cell type specific chromHMM (15 state model) enhancer elements relative to 1,000 sets of randomly selected background variants matched for distance to transcription start site and minor allele frequency. We considered the following four chromHMM states to represent enhancer elements:

1. EnhG (state 6)

2. Enh (state 7)
3. BivFlnk (state 11)
4. EnhBiv (state 12)

We used the following five Roadmap cell types to represent iPSCs:

1. E018: iPS-15b Cells
2. E019: iPS-18 Cells
3. E020: iPS-20b Cells
4. E021: iPS DF 6.9 Cells
5. E022: iPSC DF 19.11 Cells

And the following five Roadmap cell types to represent heart-related cells:

1. E065: Aorta
2. E083: Fetal heart
3. E095: Left ventricle
4. E104: Right atrium
5. E105: Right Ventricle

To compute enrichment within iPSC specific enhancer elements, we limited to enhancer elements found in at least one of the 5 iPSC cell types and none of the heart-related cell types. Likewise, for enrichment with heart specific enhancer elements, we limited to enhancer elements found in at least one of the 5 heart-related cell types and none of the iPSC related cell types. Odds ratios were smoothed by adding smoothing constant of 1 to each overlap count.

Dilated cardiomyopathy gene set enrichment analysis

We define the dilated cardiomyopathy gene set as the union of all genes in Supplementary Table 3 of Burke et al. (30). Enrichment was computed via Fisher's exact test.

Code availability

All custom scripts used for this analysis can be found at https://github.com/BennyStrobes/ipsc_cardiomyocyte_differentiation. The split-GLM (developed by Karl Tayeb) package can be found at https://github.com/karltayeb/ipsc_gp_clustering.

Discussion

In summary, our time course study design allowed us to identify hundreds of dynamic eQTLs throughout the differentiation of human iPSCs to cardiomyocytes. Dynamic eQTLs, in particular those with nonlinear effects, may often be transient and will not be found in studies that only consider gene expression data from either stem cells or mature tissues and cell types. Many of our dynamic eQTLs lie in regions without known regulatory annotations, as functional studies have focused on static cell types. Thus, these loci are candidates for novel regulatory effects, which may be followed up with further functional validation in relevant intermediate time points. Dynamic genetic effects identified in our study, or in future time series genomic datasets, provide a novel resource for investigating mechanisms underlying disease associations that cannot be characterized based on studies of terminal cell types.

Chapter 2 and Chapter 3 of this thesis attempt to characterize how the genetic regulation of gene expression changes in different contexts: tissue type and stage of cellular differentiation, respectively. However, both chapters required a priori knowledge of which context to test for interaction with genetic regulation of gene expression. In Chapter 4, we development a new statistical to methodology to identify context-specific eQTLs without having to specify a context.

Chapter 4 Uncovering context-specific genetic regulation of gene expression from single-cell RNA-sequencing using latent-factor models

Contributions

I led this project under the supervision of Alexis Battle. The idea for the project was conceived by Alexis Battle and myself. I developed the SURGE model and performed the analysis.

Abstract

Identification of genetic variants associated with gene expression, or expression quantitative trait loci (eQTLs), can be used to better understand the regulatory mechanisms linking genetic variation with cellular and high-level phenotypes including disease. However, genetic regulation of gene expression is a complex process, with genetic effects known to vary across contexts such as developmental time points, cell types, and environmental conditions. Indeed, eQTLs from adult bulk tissue samples fail to explain the majority of known disease loci. It is therefore critical to identify eQTLs from more diverse contexts in order to properly characterize the molecular mechanisms underlying disease associated loci. Recent work has shown single-cell RNA-sequencing (scRNA-seq) provides unique data to uncover context-specific eQTLs; such higher-resolution data will naturally span diverse cell types and cellular states, many of which would not be observable from bulk RNA-seq. However, the relevant factors, such as cell type or state, that actually modulate genetic effects may not be known a priori.

Furthermore, an individual cell may be defined by multiple, overlapping contexts, such as a particular cell type, cell state, and perturbation response affecting partially overlapping sets of cells. Therefore, we developed SURGE, a novel probabilistic model that uses matrix factorization to jointly learn a continuous representation of the cellular contexts that modulate genetic effects. This includes the extent of relevance of each context to each cell or sample, and the corresponding eQTL effect sizes specific to each learned context, allowing for discovery of context-specific eQTLs without pre-specifying subsets of cells or samples. In a proof of concept using bulk expression data over 49 tissues from the GTEx project, SURGE automatically learns factors capturing tissue and cell type composition differences, in addition to two factors reflecting individual ancestry. We applied SURGE to a single-cell eQTL data set consisting of multiplexed single-cell RNA-sequencing data from over 750,000 peripheral blood mononuclear cells from 119 individuals. SURGE automatically identifies cell-type specific eQTLs from this data, identifying factors capturing continuous representations of distinct blood cell types and grouping biologically related cell types into the same factor. In summary, we provide a novel approach to automatically uncover cell types and contexts that modulate genetic regulation of gene expression, enabling the unbiased discovery of diverse context-specific eQTLs from single cell, time course, and multi-condition data, and expanding our ability to explain mechanisms underlying disease-associated loci.

Introduction

A complete, mechanistic understanding of the genetic basis of complex, multi-factorial traits could provide immense insights into disease development and, ultimately, human health. A powerful approach to filling in the missing links between genotype and complex traits is to use molecular traits, such as gene expression levels, as an intermediate phenotype. Genetic variants significantly associated with mRNA expression are known as expression quantitative trait loci (eQTL) (5, 18). Unfortunately, characterizing the impact of noncoding variants is far from complete. As we explored in Chapter 2 and 3 of this thesis, this complexity arises in part because the effects of genetic variation on gene expression vary considerably between different cellular contexts, such as cell types, developmental stage, or condition (11, 31) (Figure 4-1A).

It is therefore critical to identify eQTLs from diverse contexts in order to properly characterize the molecular mechanisms underlying disease associated loci. Indeed, eQTLs from adult bulk tissue samples fail to explain the majority of known disease loci (36). Recent work has shown single-cell RNA-sequencing (scRNA-seq) provides unique data to uncover context-specific eQTLs; such higher-resolution data will naturally span diverse cell types and cellular states, many of which would not be observable from bulk RNA-seq (37,38) (Figure 4-1A).

However, the relevant factors, such as cell type or state, that actually modulate genetic effects may not be known a priori. Furthermore, an individual cell may be defined by multiple, overlapping contexts, such as cell type and a perturbation response affecting partially overlapping sets of cells (37, 39). Therefore, we developed SURGE (Single cell Unsupervised Regulation of Gene Expression), a novel probabilistic model that uses matrix factorization to jointly learn a continuous representation of the cellular contexts defining each measurement, and the corresponding eQTL effect sizes specific to each learned context, allowing for discovery of context-specific eQTLs without pre-specifying subsets of cells or samples. First, we validate SURGE on simulated data. Next in a proof of concept experiment we apply SURGE to bulk gene expression measurements from ten GTEx tissues (50) to uncover the relevant contexts underlying eQTL regulatory patterns in bulk RNA-seq data. Finally, we use SURGE to identify context-specific eQTLs from 1.2 million peripheral blood mononuclear cells (PBMC) spanning 224 genotyped individuals (40) and we apply colocalization analysis and stratified LD-score regression (S-LDSC; 41-43) to demonstrate the disease relevance of the context-specific eQTLs identified with SURGE.

Results

A standard approach to identify context-specific eQTLs is to quantify the effect of the interaction between genotype and pre-specified cellular context on gene expression levels using a linear model (interaction-eQTLs; 31). However, this approach and, to our knowledge, all existing approaches used to identify context-specific genetic regulation of gene expression require pre-specifying which contexts to test for interaction, therefore inhibiting eQTL discovery in novel, previously unstudied cellular contexts or uncharacterized cell types.

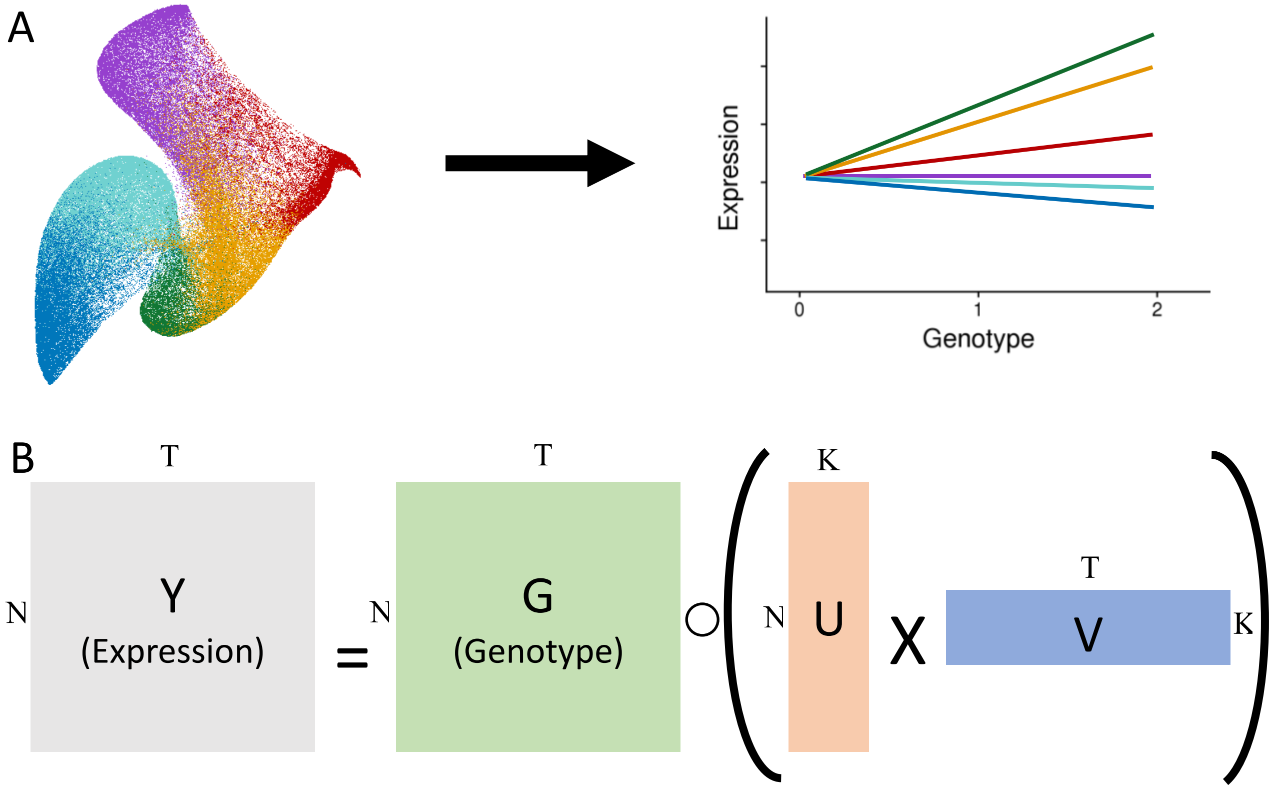


Figure 4-1: SURGE model overview

(A). Schematic example of an interaction eQTL where the eQTL effect size (right) changes a function of cellular context (colors). Single cell UMAP plot (left) generated by (44). (B) SURGE is a novel probabilistic model that uses matrix factorization to jointly learn a continuous representation of the cellular contexts defining each measurement (U), and the corresponding eQTL effect sizes specific to each learned context (V) based on observed expression (Y) and genotype data (G). Assuming there are N samples, T genome-wide independent variant-gene pairs, and K latent contexts, the observed expression matrix (Y) is of dimension NXT, the observed genotype matrix (G) is of dimension NXT, the SURGE latent context matrix (U) is of dimension NXK, and the context-specific eQTL effect size matrix is of dimension KXT.

To address this issue, we expanded upon the traditional interaction-eQTL model with the development of SURGE. SURGE is able to uncover context-specific eQTLs without pre-specifying the contexts of interest. SURGE achieves this goal by leveraging information across genome-wide independent variant-gene pairs to jointly learn both a continuous representation of the SURGE latent contexts, or the cellular contexts defining each measurement and the corresponding eQTL effect sizes specific to each SURGE latent context (Figure 4-1B, Appendix C: Figure S1; see Methods). Importantly, the SURGE framework allows for any individual measurement to be defined by multiple, overlapping contexts. From an alternative but equivalent lens, SURGE discovers the latent contexts whose linear interaction with genotype explains the most variation in gene expression levels across genome-wide independent variant-gene pairs (see Methods). From this perspective, SURGE enables unsupervised discovery of the principal axes of genetic regulation of gene expression within an eQTL data set. Additionally of note, SURGE controls for the effect of known covariates and sample repeat structure induced by assaying multiple measurements from the same individual on gene expression when identifying context-specific eQTLs, which would otherwise lead to false positive eQTLs (see Methods). Finally, built into SURGE's optimization procedure is the automatic selection of the number of relevant latent contexts. The user simply has to initialize the number of latent contexts to be large and greater than the likely number of underlying latent contexts present in the eQTL data set, and SURGE will remove unnecessary contexts during optimization (44; see Methods; Appendix C: Figure S2).

For proof of concept, we applied SURGE to model RNA-sequencing samples from 10 GTEx version 8 tissues (see Methods). Here, each RNA sample was extracted from a specific tissue, and while tissue identity information was not provided to SURGE, 5 of the 7 SURGE latent contexts captured differences in tissue type between the samples (Figure 4-2A, Appendix C: Figure S3). SURGE latent context 1 (latent contexts ordered by PVE, see Methods), for example, isolates RNA samples from Muscle-Skeletal tissue; RNA samples derived from Muscle-Skeletal tissue have an average latent context 1 value of 2.647 (sdev .538), while RNA samples from other tissues have an average latent context 1 value of -.542 (sdev 0977). Furthermore, we discovered SURGE latent context 3 and 4 cluster samples according to their known ancestry; samples from African Ancestry donors were strongly loaded on both latent context 3 and 4 (Figure 4-2B, Appendix C: Figure S4).

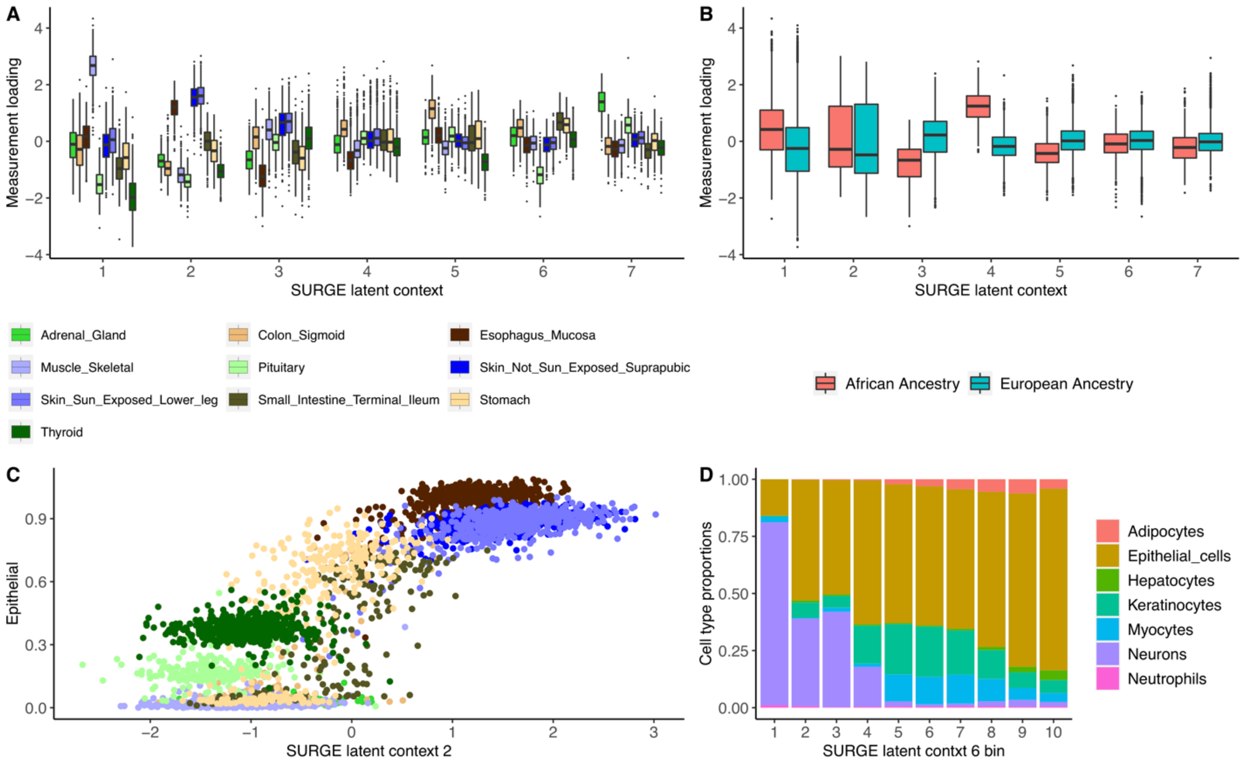


Figure 4-2: SURGE applied to GTEx v8 bulk RNA seq samples

(A,B) SURGE latent context loadings of GTEx v8 RNA-seq samples (y-axis) stratified by (A) known tissue identity and (B) known ancestry for each of the 7 identified SURGE latent contexts. (C) Scatter plot of SURGE latent context 2 loadings (x-axis) and xCell Epithelial cell type proportions estimates (y-axis) for GTEx v8 RNA-seq samples colored by known tissue identity. (D) GTEx v8 RNA-seq samples are

separated into 10 equally-sized bins according to their value on SURGE latent context 6. The stacked bar plot depicts the average cell-type composition according to xCell estimates across all samples (y-axis) in each of the 10 bins (x-axis).

Next, we intersected the learned SURGE latent contexts with previously computed computational estimates of each RNA sample's cell type composition according to xCell (45, 46). We found that the SURGE latent contexts were not simply identifying differences in tissue identity between the samples, but learning changes in cell type composition of samples both across tissues and within a single tissue (Figure 4-2C, Figure 4-2D, Appendix C: Figures S5, S6). SURGE latent context 2, for example, is highly correlated with epithelial cell type levels across samples from all ten tissues (Figure 4-2C). Moreover, many of the SURGE latent contexts capture complex multi-cell type composition continuums, not simply the change in proportions of a single cell type (Figure 4-2D, Appendix C: Figure S5). This holds true even when SURGE is applied to RNA samples from a single tissue (see Methods, Appendix C: Figure S6). As expected, we observe greater power to detect context-specific eQTLs when SURGE latent contexts are used as opposed to using cell type composition estimates from xCell (46, see Methods, Appendix C: Figure S7). In summary, the SURGE identifies tissue-type, cell-type, and ancestry as the primary axes of genetic regulation of gene expression within GTEx eQTL data.

Next, we applied SURGE to a recently generated single cell eQTL data set consisting of 1.2 million PBMC spanning 224 genotyped individuals (40). Notably, 141 of these individuals have systemic lulus erythematosus (SLE) while the remainder are healthy.

To mitigate the sparsity characteristic of 10X sequencing data, we aggregated cell level expression data across highly correlated cells to generate 22870 pseudocells (see Methods; Appendix C: Figure S8; 47), aggregating on average 22 cells per pseudocell. Here, SURGE identified 3 latent contexts that capture continuous representations of distinct blood cell types while integrating biologically related cell types along a gradient within a single latent context (Figure 4-3A, Appendix C: Figures S9-S11). SURGE latent context 3, for example, is strongly loaded on B-cells, while still identifying fine-resolution differences distinguishing naïve B-cells from plasma-derived B cells (Appendix C: Figure S10). Additionally, SURGE latent context 1 identified subtle differences that isolated monocytes derived from healthy individuals from monocytes derived from SLE individuals (Appendix C: Figure S12; $p < 1.4e-32$, Wilcoxon rank sum test).

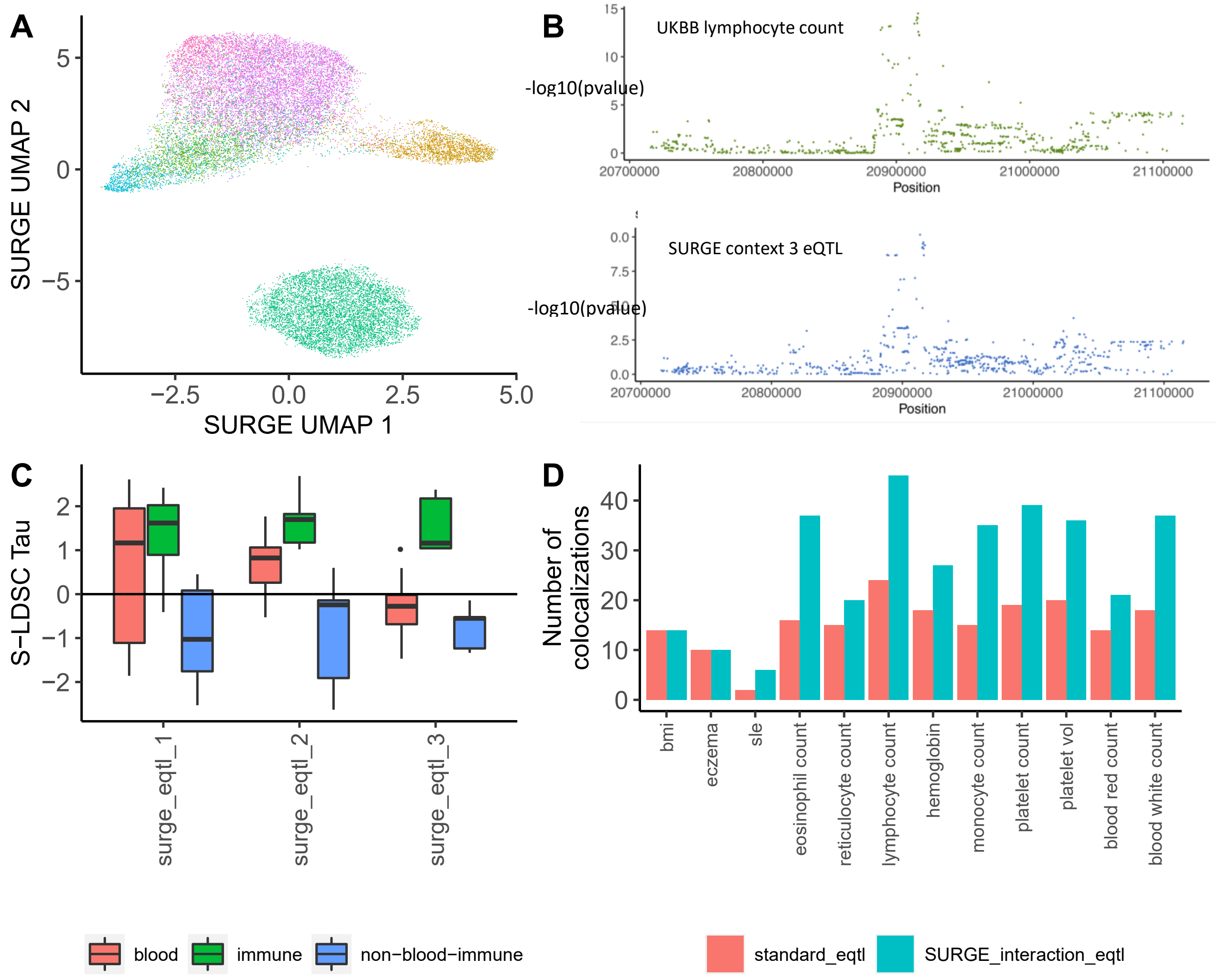


Figure 4-3: SURGE applied to PBMC single-cell eQTL data

(A). UMAP-projected SURGE latent contexts of pseudocells colored by marker-gene derived cell types. Mapping from color to cell type can be found in Appendix C: Figure S10. (B) Colocalization between SURGE context 3 specific eQTL variant rs1253904 for CDA and GWAS signal for lymphocyte count. (C). S-LDSC estimates of effect sizes (Tau; y-axis) corresponding to variant annotations derived from SURGE interaction eQTLs (x-axis) for traits belonging to different categories of traits (color). Trait category of "blood" consists of GWAS for eosinophil count, reticulocyte count, lymphocyte count, corpuscular hemoglobin, monocyte count, platelet count, blood platelet volume, red blood count, and white blood count. Trait category of "immune" consists of GWAS for Celiac, Crohns, IBD, Lupus, Multiple-sclerosis,

PBC, Rheumatoid Arthritis, Eczema, and Ulcerative Colitis. Trait category of "non-blood-immune" consists of GWAS for Alzheimer, Bipolar, CAD, Schizophrenia, BMI, height, and type-2 Diabetes. surge_eqtl_x corresponds to a binary annotation isolating all variants with SURGE interaction eQTL pvalue < 1e-5 with respect to SURGE latent context x. S-LDSC was run for each eQTL study independently while controlling for BaselineLD annotations. (D) Number of colocalizations identified (PPH4 > .95; y-axis) between various densely genotyped GWAS studies (x-axis) and eQTLs identified from pseudocells. Number of colocalizations with standard eQTLs shown in red and number of unique colocalizations aggregated across SURGE interaction eQTLs identified in 3 latent contexts shown in blue.

Finally, we sought to evaluate the relationship between context-specific eQTLs identified using SURGE and disease-associated loci. Using coloc (41), we identified hundreds of colocalizations between SURGE context-specific eQTLs and GWAS loci (Figure 4-3B, Appendix C: Figure S13). For example, SURGE context 3 specific eQTL variant rs1253904 for CDA colocalized with a GWAS signal for lymphocyte count (Figure 4-3B). Furthermore, based on S-LDSC (42), SURGE context-specific eQTLs showed specific enrichment for trait heritability of immune and blood-related traits (average S-LDSC enrichment 6.12 and 3.35, respectively), but were not enriched among equivalently heritable traits unrelated to the immune system or blood morphology (Figure 4-3C, Appendix C: Figure S14). Relative to standard eQTLs, SURGE context-specific eQTLs consistently explained significantly higher proportions of total trait heritability (Appendix C: Figure S15), suggesting that SURGE is identifying trait-relevant loci that are not identified with standard eQTL analysis. Similarly, we identified significantly more trait colocalizations with SURGE context-specific eQTLs relative to using standard eQTLs (Figure 4-3D).

Methods

SURGE model overview

The SURGE model is defined according to the following probability distributions:

$$y_{nt} \sim N(\mu_t + \sum_l X_{nl} W_{lt} + \prod_i I[n \in i]\alpha_{it} + G_{nt}F_t + G_{nt}(\sum_k U_{nk}V_{kt}), \sigma_t^2)$$

$$U_{nk} \sim N(0, \gamma_k^2)$$

$$V_{kt} \sim N(0,1)$$

$$\gamma_k^2 \sim InverseGamma(\alpha_0, \beta_0)$$

$$F_t \sim N(0,1)$$

$$\alpha_{it} \sim N(0, \psi_t^2)$$

$$\psi_t^2 \sim InverseGamma(\alpha_0, \beta_0)$$

$$\sigma_t^2 \sim InverseGamma(\alpha_0, \beta_0)$$

Here, n indexes RNA samples, t indexes independent variant-gene pairs being tested for eQTL analysis, and i indexes individuals. We use the notation $n \in i$ to represent the instance where RNA sample n is drawn from the individual i. y_{nt} is the observed normalized gene expression (mean 0 and variance 1 for each test t) level of the gene corresponding to test t in sample n. G_{nt} is the observed, standardized (mean 0 and variance 1 for each test t) genotype of the variant corresponding to test t in sample n. X_{nl} is the observed value of covariate l for sample n . SURGE infers the values of:

- F_t: the eQTL effect size of test t that is shared across samples
- V_{kt}: the eQTL effect size of test t for latent context k
- U_{nk}: the latent context value of sample n on factor k
- μ_t: the intercept of each test

- W_{lt}: The effect size of covariate l on the gene corresponding to test t
- α_{it}: the random effect intercept for each individual for each test
- γ_k^2: The variance of the values in latent context k
- ψ_t^2: The variance of intercepts corresponding to each individual in test t
- σ_t^2: The residual variance in gene expression levels in test t

a_0, and β_0 are model hyper-parameters set to provide non-informative priors while stabilizing optimization. In practice we set α_0 to 1e-16 and β_0 to 1e-16. A mean-zero gaussian prior is placed on U_{nk} in order to produce interpretable assignments of samples to factors. The level of regularization of that prior is learned separately for each latent context (γ_k^2), allowing SURGE to zero-out (γ_k^2 approaches 0) irrelevant contexts and automatically learn the effective number of latent contexts.

SURGE optimization

All latent variables [Z =(F_t, V_{kt}, U_{nk}, μ_t, W_{lt}, α_{it}, γ_k^2, ψ_t^2, σ_t^2)] are learned using mean-field variational inference (48). The goal of variational inference is to minimize the KL-divergence between $q(Z)$ and $p(Z|Y,G,X)$, which can be written as $KL(q(Z)||p(Z|Y,G,X)$. Here, $q(Z)$ is a simple, tractable distribution that is used to approximate $p(Z|Y,G,X)$. We used the “mean-field approximation” for $q(Z)$ such that all latent variables are independent of another. More specifically:

$$\log q(Z) =$$

$$\sum_t \sum_k logN(V_{kt}|\mu_{V_{kt}}, \sigma^2_{V_{kt}}) +$$

$$\sum_{t}\sum_{i} logN(\alpha_{it}\,|\mu_{\alpha_{it}}, \sigma^2_{\alpha_{it}}) +$$

$$\sum_{t}\sum_{l} \log N\left(W_{lt}|\mu_{W_{lt}}, \sigma^2_{W_{lt}}\right) +$$

$$\sum_{t} [\, logN\left(F_t|\mu_{F_t}, \sigma^2_{F_t}\right) + logN\left(\mu_t|\mu_{\mu_t}, \sigma^2_{\mu_t}\right) + logIG\left(\psi^2_t|\alpha_{\psi_t}, \beta_{\psi_t}\right) + logIG\left(\sigma^2_t|\alpha_{\sigma_t}, \beta_{\sigma_t}\right)] +$$

$$\sum_{k} logIG(\gamma^2_k|\alpha_{\gamma_k}, \beta_{\gamma_k}) +$$

$$\sum_{n}\sum_{k} log\ N(U_{nk}|\mu_{U_{nk}}, \sigma^2_{U_{nk}})$$

Where $N(x|\mu, \sigma^2)$ is a univariate normal distribution parameterized by mean μ and variance σ^2 and $IG(X|\alpha, \beta)$ is a univariate inverse-gamma distribution parameterized by α and β.

It can be shown that minimizing the KL-divergence $KL(q(Z)||p(Z|Y, G, X)$ is equivalent to maximizing the evidence lower bound (ELBO):

$$E_q[logp(G, Y, X, Z)] - E_q[logq(Z)]$$

Therefore, we will frame SURGE optimization from the perspective of maximizing the ELBO with respect to the parameters defining $q(Z)$, or the variational parameters. Noteworthy is $p(G, Y, X, Z)$ is explicitly defined in “SURGE model overview”. The approach we take to maximize the ELBO is through coordinate ascent (48), iteratively optimizing one latent variable of the mean field density while holding all other latent variables fixed. Accordingly, the ELBO is guaranteed to monotonically increasing after each variational update, and in the case of the SURGE model, each update results in closed form updates.

Optimization of variational parameters is performed as follows: we randomly initialize all variational parameters (see below section entitled “Random initialization for SURGE optimization”) and then iteratively loop through all latent variables in Z and update the variational parameters corresponding to that latent variable until we reach convergence.

To assess convergence, we divide the change in ELBO from previous iteration by by the current value of the ELBO. We consider the model converged when this fraction is less than 1e-8.

Random initialization for SURGE optimization

It is important to note that mean-field variational inference is not guaranteed to converge to the global optima of the ELBO. To mitigate the effects of local optima, we recommend optimizing multiple models with different random initializations and using the parameters learned from the model that achieves the largest ELBO.

Percent variance explained of SURGE latent contexts

Following the approach taken by (49), we define the “Percentage Variance Explained” (PVE) of the k^{th} latent context as:

$$pve_k = \frac{s_k}{(\sum_k s_k) + (N * \sum_t \sigma_t^2)}$$

$$s_k = \sum_n \sum_t G_{nt} U_{nk} V_{kt}$$

As stated in (49), this approach is a measure of the amount of signal in data set that is identified by the k^{th} latent context. However, the name “percentage variance explained” should be considered loosely as the factors are not orthogonal.

Removing irrelevant latent contexts

Upon model convergence, we remove latent contexts with PVE $\leq 1e^{-5}$

Selection of variant-gene pairs used for optimization

SURGE optimization (ie. learning the SURGE latent contexts) requires an input expression matrix and genotype matrix. As specified above, both matrices should be of dimension $N\mathrm{X}T$, where N is the number of RNA samples and T is the number of genome-wide independent variant gene pairs. We desire each variant-gene pair to be independent of one another because we want the SURGE to capture eQTL patterns that are persistent across the genome, not specific to a single gene or variant.

Therefore, to encourage the expression and genotype data consists of independent variant-gene pairs we limit there to be a single variant-gene pair selected for each gene and limit there to be a single variant-gene pair selected for each variant.

Furthermore, it has been shown that context-specific eQTLs are more likely to be standard eQTLs than not. Therefore, we limit variant-gene pairs used for SURGE optimization to those that are standard eQTLs within the data set.

SURGE interaction-eQTLs

SURGE optimization on a subset of genome-wide independent variant-gene pairs will result in estimates of the SURGE latent contexts (U) as well as eQTL effect size estimates for each of the SURGE latent contexts for only the genome-wide independent variant gene pairs (V). It is of interest, however, to call interaction eQTLs with respect to each of the SURGE latent contexts for *all* variant gene-pairs, not just a subset of variant gene pairs that are independent.

Therefore, to identify SURGE interaction-eQTL for a particular variant-gene pair we treat the SURGE latent contexts (U: dim NXK) and generate a separate linear mixed model for each tested variant-gene pair. The linear mixed model is as follows:

$$y_n \sim N(\mu + \sum_i \alpha_i I[n \in i] + \sum_l W_l X_{nl} + \beta_g G_n + \sum_k \beta_k U_{nk} + \sum_k \beta_{gxk} G_n U_{nk}, \sigma^2)$$

$$\alpha_i \sim N(0, \psi^2)$$

Here:

- y_n is the observed expression level of the gene corresponding to the variant-gene pair in sample n
- g_n is the observed genotype of the variant corresponding to the variant-gene pair in sample n

- X_{nl} is the observed value of covariate l in sample n
- μ is the intercept
- α_i is the random effect intercept for individual i. We use the notation $n \in i$ to represent the case where sample n is drawn from individual i
- W_l is the fixed effect for covariate l
- β_g is the fixed effect for genotype
- β_k is the fixed effect of the k^{th} latent context
- β_{gxk} is the fixed effect of the interaction between the k^{th} latent context and genotype

We use the R package 'lme4' to quantify the significance of all K interaction terms: $\beta_{gx1}, \dots, \beta_{gXk}, \dots, \beta_{gxK}$. Intuitively, if the k^{th} interaction term (β_{gxk}) is significant, it implies that the eQTL effect size of this variant-pairs significantly changes along latent context k.

Simulation experiments

To assess SURGE's ability to accurately capture contexts underlying context-specific eQTLs we performed the following simulation experiment:

We randomly generated genotype and expression matrices across 1000 variant-gene pairs and N RNA samples. For each simulated variant-gene pair, we simulated the genotype vector (G) across the N samples according to the following probability distributions:

$$G_n \sim Binomial(2, \text{allele\_frequency})$$

$$\text{allele\_frequency} \sim Uniform(.05, .95)$$

Then, we simulated the expression vector (y) across the N samples using that variant-gene pair's simulated genotype vector according to the following probability distributions:

$$y_n \sim N(\mu + \beta G_n + \sum_k G_n U_{nk} V_k \theta_k, 1)$$

$$\mu \sim N(0,1)$$

$$\beta \sim N(0,1)$$

$$U_{nk} \sim N(0,1)$$

$$V_k \sim N(0, \gamma)$$

$$\theta_k \sim Bernoulli(p)$$

In this simulation, we evaluate SURGE's ability to re-capture the simulated latent contexts (U) (Appendix C: Figure S1) as a function of the simulation hyper-parameters:

- The number of latent contexts (K)
- The sample size (N)
- The strength of the interaction terms (γ)
- The fraction of tests that are context-specific eQTLs for a particular context (p)

We also access SURGE's ability to accurately estimates the number of relevant contexts (K) (Appendix C: Figure S2).

Application of SURGE to GTEx samples from 10 tissues: expression quantification

To normalize expression from samples from 10 GTEx tissues (Adrenal gland, Colon-sigmoid, Esophagus-Mucosa, Muscle-Skeletal, Pituitary, Skin-not-sun-exposed, Skin-sun-exposed, small-intestine-terminal-ileum, Stomach, Thyroid), we concatenated log-TPM expression measurements across all samples used in the GTEx v8 eQTL analysis for one of those tissues (50). We also limited to genes that were tested for eQTLs in the GTEx v8 analysis (50) in all 10 tissues. Next, we quantile normalized this matrix to ensure each sample had an equivalent distribution across genes and then standardized each gene (mean 0 and standard deviation 1).

Application of SURGE to GTEx samples from 10 tissues: standard eQTL calling

We first tested for standard eQTLs, or association between genotype and the concatenated expression vector described above in "Application of SURGE to GTEx samples from 10 tissues: expression quantification". For this analysis, we limited to genes that passed filters described in "Application of SURGE to GTEx samples from 10 tissues: expression quantification". We then limited to variants with MAF >= .05 (according cross-tissue concatenation of genotype vector) that were less than 50KB from the transcription start site of a gene. We controlled for the effects of 80 expression PCs and 4 genotype PCs. We assessed genome-wide significance according to a gene-level Bonferonni correction, followed by a genome-wide Benjamini-Hochberg correction.

Application of SURGE to GTEx samples from 10 tissues: SURGE optimization

To select a subset of variant-gene pairs to be used for SURGE model optimization, we first limited to variant-gene pairs that were standard eQTLs (FDR <= .05; see "Application of SURGE to GTEx samples from 10 tissues: standard eQTL calling"). This was done to ensure a higher fraction of the variant-gene pairs used for SURGE optimization were context-specific eQTLs as it is known standard eQTLs are more likely to be context-specific eQTLs than variant-gene pairs that are not standard eQTLs. Furthermore, we limited to the most significant variant per gene amongst the 2000 most significant genes and removed a variant-gene pair if the variant was already in the training set for its association with a more significant gene. This yielded 1,996 genome-wide independent variant-gene pairs used for SURGE optimization. We than ran SURGE under default parameter settings over these genome-wide independent variant-gene pairs. We included 80 expression PCs and 4 genotype PCs as covariates in SURGE. The converged SURGE model resulted in 7 latent contexts with PVE $> 1e^{-5}$.

Application of SURGE to GTEx samples from a single tissue

To run SURGE on GTEx samples from a single GTEx tissue, we took a very similar approach to that described in "Application of SURGE to GTEx samples from 10 tissues: expression quantification", "Application of SURGE to GTEx samples from 10 tissues: standard eQTL calling", and "Application of SURGE to GTEx samples from 10 tissues: SURGE optimization". The only difference is that we now limit to samples from the tissue of interest. Furthermore, we now only control for 60 expression PCs and 2 genotype PCs during standard eQTL calling and SURGE optimization.

Application of SURGE to PBMC single cell eQTL data: pseudocell expression quantification

We imported raw, un-normalized UMI counts from (40). We used SCRAN (51) to generate log-normalized counts for each cell. We removed genes that were expressed in fewer than .5% of cells. We then limited to the top 6000 highly variable genes via the Scanpy function “highly_variable_genes” (52). We then removed the effects of sequencing batch using Combat (53) as implemented in Scanpy. We then scaled each gene to have mean 0 and variance 1, with a maximum absolute value of 10 to mitigate outlier effects as implemented by “scanpy.pp.scale”.

Next, we sought to generate pseudocells that represented groupings of highly correlated cells within an individual. We first removed individuals from this analysis with fewer than 2500 cells. Next we performed Leiden clustering as implemented by Scanpy (54) independently in each individual using all default parameters, except we used a fine-grained cluster resolution of 10. Here, each leiden cluster corresponds to a pseudocell. We took the average expression across all cells assigned to the pseudocell to estimate the expression profile of the pseudocell. Finally, we standardized each gene (across pseudocells) to have mean 0 and standard deviation 1, again capping the absolute value of standardized scores to be 10 to mitigate outlier effects.

Application of SURGE to PBMC single cell eQTL data: standard eQTL calling

We first tested for standard eQTLs, or association between genotype and the expression vector across pseudocells described above in “Application of SURGE to Ye-lab generated single cell eQTL data: pseudocell expression quantification”. For this analysis, we limited to genes that passed filters described in “Application of SURGE to Ye-lab generated single cell eQTL data: pseudocell expression quantification”. We then limited to variants with MAF >= .05 that were less than 200KB from the transcription start site of a gene. We controlled for the effects of 30 expression PCs and 2 genotype PCs. We assessed genome-wide significance according to a gene-level Bonferonni correction, followed by a genome-wide Benjamini-Hochberg correction (55).

Application of SURGE to PBMC single cell eQTL data: SURGE optimization

To select a subset of variant-gene pairs to be used for SURGE model optimization, we first limited to variant-gene pairs that were standard eQTLs (FDR <= .05; see “Application of SURGE to Ye-lab generated single cell eQTL data: standard eQTL calling”). This was done to ensure a higher fraction of the variant-gene pairs used for SURGE optimization were context-specific eQTLs as it is known standard eQTLs are more likely to be context-specific eQTLs than variant-gene pairs that are not standard eQTLs. Furthermore, we limited to the most significant variant per gene amongst the 2000 most significant genes and removed a variant-gene pair if the variant was already in the training set for its association with a more significant gene. We than ran SURGE under default parameter settings over these genome-wide independent variant-gene pairs. We included 30 expression PCs and 2 genotype PCs as covariates in SURGE. The converged SURGE model resulted in 3 latent contexts with PVE $> 1e^{-5}$.

Discussion

Here, we presented SURGE, a novel probabilistic model that identifies context-specific eQTLs from single-cell data without pre-specifying contexts or subsets of cells or samples. SURGE leverages information from variant-gene pairs across the entire genome to learn a continuous representation of the cellular contexts defining each measurement, and the corresponding eQTL effect sizes specific to each learned context. Importantly, SURGE allows for unsupervised discovery of the principal axes of genetic regulation of gene expression within an eQTL data set, identifying cell-type, tissue-type, and ancestry when applied to GTEx tissue samples and highly resolved blood cell-types when applied to blood-derived single cells. Ultimately, SURGE identified many trait-relevant loci that could not be detected through traditional eQTL approaches. Thus, these loci are candidates for novel regulatory effects, which may be followed up with further functional validation in relevant contexts.

This chapter of the thesis, along with Chapter 2 and 3, provide significant methodological advancements in the analysis of context-specific genetic regulation of gene expression. These analyses are limited, however, to the study of common genetic variants. We exclude the analysis of rare genetic variants because eQTL studies are statistical underpowered to detect associations when the minor allele frequency of the genetic variant is low. For the remainder of this thesis, we transition to developing alternative approaches suitable for understanding the impact of *rare* genetic variants on gene expression.

Chapter 5 Transcriptomic signatures across human tissues identify functional rare genetic variation

Contributions

This chapter describes the utility of using multiple transcriptomic signatures to inform functional rare genetic variation across human tissues. I co-led this project along with Nicole Ferraro. This work was jointly supervised by Pejman Mohammadi, Stephen B. Montgomery, and Alexis Battle. This work was published in (56), as a part of the GTEx version 8 project. My main contribution to the published manuscript includes:

- Developing SPOT and conducting sOutlier analysis
- Develop Watershed

The text of this chapter is a modification of the published work (56), focusing on results relevant to my contribution. The text was written together by Nicole Ferraro, Benjamin J. Strober, Jonah Einson, Pejman Mohammadi, Stephen B. Montgomery, and Alexis Battle. The full list of collaborators involved in this project is available in (56).

Abstract

Rare genetic variants are abundant across the human genome, and identifying their function and phenotypic impact is a major challenge. Measuring aberrant gene expression has aided in identifying functional, large-effect rare variants (RVs). Here, we expanded detection of genetically driven transcriptome abnormalities by analyzing gene expression, allele-specific expression, and alternative splicing from multitissue RNA-sequencing data, and demonstrate that each signal informs unique classes of RVs. We

developed Watershed, a probabilistic model that integrates multiple genomic and transcriptomic signals to predict variant function, validated these predictions in additional cohorts and through experimental assays, and used them to assess RVs in the UK Biobank, the Million Veterans Program, and the Jackson Heart Study. Our results link thousands of RVs to diverse molecular effects and provide evidence to associate RVs affecting the transcriptome with human traits.

Introduction

The human genome contains tens of thousands of rare [minor allele frequency (MAF) <1%] variants (57), some of which contribute to rare and common disease risks (58). Unlike the common genetic variants discussed in Chapters 2, 3, and 4, identifying functional rare variants (RVs), especially in the noncoding genome, remains difficult because of their low frequency and the lack of a regulatory genetic code. Outlier gene expression aids in identifying functional, large-effect RVs (59). Furthermore, transcriptome sequencing provides diverse measurements beyond gene expression level, including allele-specific expression (ASE) and alternative splicing, that have yet to be systematically evaluated and integrated into variant effect prediction (60).

Using 838 samples with both whole-genome and transcriptome samples in the Genotype-Tissue Expression (GTEx) project version 8 (v8), we assessed how rare genetic variants contribute to outlier patterns in total expression (hereafter referred to simply as “expression”), allelic expression, and alternative splicing deep into the allele frequency (AF) spectrum. We integrated these three transcriptomic signals across 49 tissues, along with diverse genomic annotations to prioritize high-impact RVs, and assessed their relationship to complex traits in the UK Biobank (UKBB) (61), the Million

Veterans Program (MVP) (62), and the Jackson Heart Study (JHS) (63). We further identified dozens of candidate RVs influencing well-studied disease genes, including *APOE*, *FAAH*, and *MAPK3*.

Results

Detection of aberrant gene expression across multiple transcriptomic phenotypes

We quantified three transcriptional phenotypes for each gene to capture a wide range of functional effects caused by regulatory genetic variants. Briefly, to identify expression outliers (eOutliers), we generated *Z* scores from corrected expression data per tissue to determine whether a gene in an individual has extremely high or low expression (Appendix D: Figure S1) (21). To identify genes with excessive allelic imbalance [allele-specific expression (ASE) outliers (aseOutliers)] we used ANEVA-DOT (analysis of expression variation–dosage outlier test; Appendix D: Figures S2, S3) (64) (see Methods). This method uses estimates of genetic variation in dosage of each gene in a population to identify genes for which an individual has a heterozygous variant with an unusually strong effect on gene regulation (64). Splicing outliers (sOutliers) were detected using SPOT (splicing outlier detection), an approach introduced here that fits a Dirichlet-Multinomial distribution directly to counts of reads split across alternatively spliced exon-exon junctions for each gene. SPOT then identifies individuals that deviate significantly from the expectation on the basis of this fitted distribution (Appendix D: Figures S4-S6) (see Methods). Each of the three methods was applied across all GTEx samples. An individual was called a multitissue outlier for a given gene if its median outlier statistic across all measured tissues exceeded a chosen threshold (Figure 5-1A) (see Methods). Using this multitissue approach for each phenotype, we found that each individual had a median of four eOutlier, four aseOutlier, and five sOutlier genes.

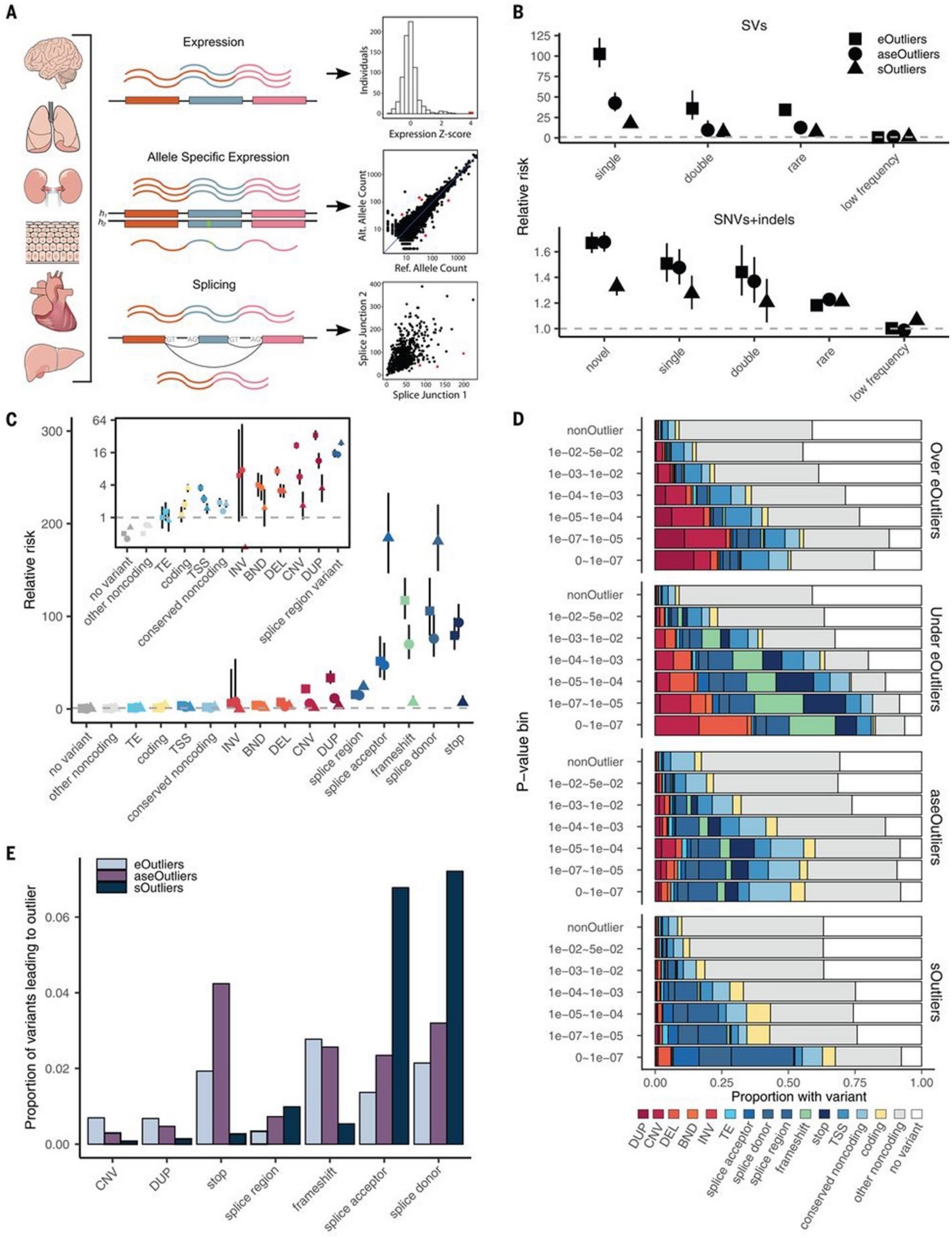


Figure 5- 1: Impact of rare variation on diverse outlier signals

(A) RNA-seq data in 838 individuals were combined across 49 tissues and used to identify shared tissue

expression, ASE, and alternative splicing outliers. (B) Relative risk of new (not in gnomAD), singleton, doubleton, rare (MAF <1%), and low-frequency (MAF 1 to 5%) variants in a 10-kb window around the outlier genes across all data types compared with nonoutlier individuals for the same genes. Outliers were defined as those with values >3 SDs from the mean (|median Z| *> 3) or, equivalently, a median* P *< 0.0027. Bars represent the 95% confidence interval. (C) Assigning each outlier its most consequential nearby RV, the relative risk for different categories of RVs falling within 10 kb of each outlier type. The inset panel shows enrichments for a subset of variant categories on a log(2)-transformed* y*-axis scale for better visibility. (D) Proportion of outliers at a given threshold that have a nearby RV in the given category. eOutlier* |*median* Z *scores*| *were converted to* P *values using the cumulative probability density function for the normal distribution. TE, transposable element; INV, inversion; BND, break end; DEL, deletion; DUP, duplication. (E) Proportion of RVs in a given category that lead to an outlier at a* P*-value threshold of 0.0027 across types.*

Genes with aberrant expression, ASE, and splicing are enriched for functionally distinct RVs

We observed that multitissue outliers for any of the three transcriptomic phenotypes were significantly more likely to carry a RV (MAF <1%) in the gene body or ±10 kb than individuals without outliers, assessed among 714 individuals with European ancestry. These enrichments were progressively more pronounced for rarer variants and were stronger for structural variants (SVs) than for single-nucleotide variants (SNVs) and indels (Figure 5-1B). These trends were not reliant on the specific choice of the threshold used to define outliers (Appendix D: Figures S7, S8).

We found only 35 cases in which an individual gene was a multitissue outlier for all three transcriptional phenotypes. All but one of these had a nearby RV, and most were annotated as splice variants. Among genes that were outliers for two transcriptional phenotypes in an individual (*n* = 465), the greatest overlap occurred between aseOutliers and eOutliers (*n* = 319; Appendix D: Figure S9A). We found that aseOutliers with modest expression changes (1 < |median *Z*| < 3) showed stronger enrichment for nearby RVs than those without any expression change (Appendix D: Figure S9), highlighting an important benefit of combining these phenotypes to discover diverse RV effects. We found that genes for which no outlier individuals were identified were enriched for Gene Ontology biological process terms relating to sensory perception and detection of chemical stimuli for all outlier types (Appendix D: Figure S10) (see Methods), which is consistent with enrichments seen for genes that do not have any cis-expression quantitative trait loci (eQTLs) discovered in GTEx (11).

We found that different types of genetic variants contribute to outliers for the three molecular phenotypes, although rare splice donor variants were enriched near all outlier types (Figure 5-1C). The largest differences in variant type enrichment among the three outlier types were copy number variations (CNVs) and duplications, which were almost exclusively associated with eOutliers, and splice acceptor variants, which were enriched considerably more within sOutliers (Appendix D: Figure S11).

For all phenotypes, the proportion of outliers with a nearby RV of any category increased with threshold stringency (Figure 5-1D). For eOutliers, aseOutliers, and sOutliers, at the strictest threshold of median outlier $P < 1.1 \times 10^{-7}$, most individuals were carrying at least one RV nearby the outlier gene (82 to 94%). When looking further at RVs with functional annotations (from the annotations listed in Figure 5-1C), we found that underexpressed eOutliers were the most interpretable, with 88% of outlier-associated RVs having an additional functional annotation, whereas aseOutliers had the lowest proportion at 56% (Figure 5-1D). This analysis provides further insight into expectations for causal RV types when an outlier effect of a specific magnitude is observed in an individual.

Conversely, a large proportion of genes with nearby rare genetic variants did not appear as outliers, even for the most predictive classes such as loss-of-function variants. The largest proportion of variants leading to any outlier status were rare splice donor and splice acceptor variants, of which only 7.2 and 6.8%, respectively, led to an sOutlier (Figure 5-1E and Appendix D: Figure S11). Overall, whereas some transcriptomic effects may have been missed, the low frequency with which RVs of these classes led to large transcriptome changes reinforces the utility of incorporating functional data in variant interpretation even for specific variant classes already used in clinical interpretation.

Genomic position of RVs predicts the impact on expression

Although we primarily assessed RVs that occur either within an outlier gene or in a 10-kb surrounding window, gene regulation can occur at greater distances (65). Because we observed the strongest enrichments for the lowest-frequency variants, we intersected singleton variants [(SVs); i.e., those appearing only once in GTEx and SNVs and/or indels that do not appear in the Genome Aggregation Database (gnomAD) (66)] with 200-kb-length windows exclusive of other windows and upstream from outlier genes and compared their frequency in outlier versus nonoutlier individuals. SNV enrichments dropped off quickly at greater distances from the gene but remained weakly enriched for eOutliers out to 200 kb. The same was true for rare indels, with enrichment at 200 kb only for sOutliers. SVs remained enriched at much longer distances, being enriched 2.33-fold as far as 800 kb to 1 Mb upstream and up to 600 kb downstream of the gene body (Figure 5-2A and Appendix D: Figure S12A).

RVs in promoter regions have been previously linked to outlier expression (21). To extend these observations and to assess the types of transcription factor (TF)–binding sites that could lead to outliers, we tested enrichment of rare transcription start site (TSS) proximal variants in specific TF motifs near under- and over-eOutliers. For under-eOutliers, we saw an enrichment of variants in *GABP*, a TF that activates genes that control the cell cycle, differentiation, and other critical functions (67). For over-eOutliers, we saw an enrichment of RVs intersecting the *E2F4* motif, a TF that has been reported as a transcriptional repressor (68). In both under- and over-eOutliers, we saw RVs in *YY1*, which can act as either an activator or repressor, depending on context (69), and has been associated with *GABP* in coregulatory networks (Figure 5-2B and Appendix D: Figure S12B) (70). Thus, these naturally occurring RV perturbations can provide information about how specific TFs can strongly up- or down-regulate their target genes.

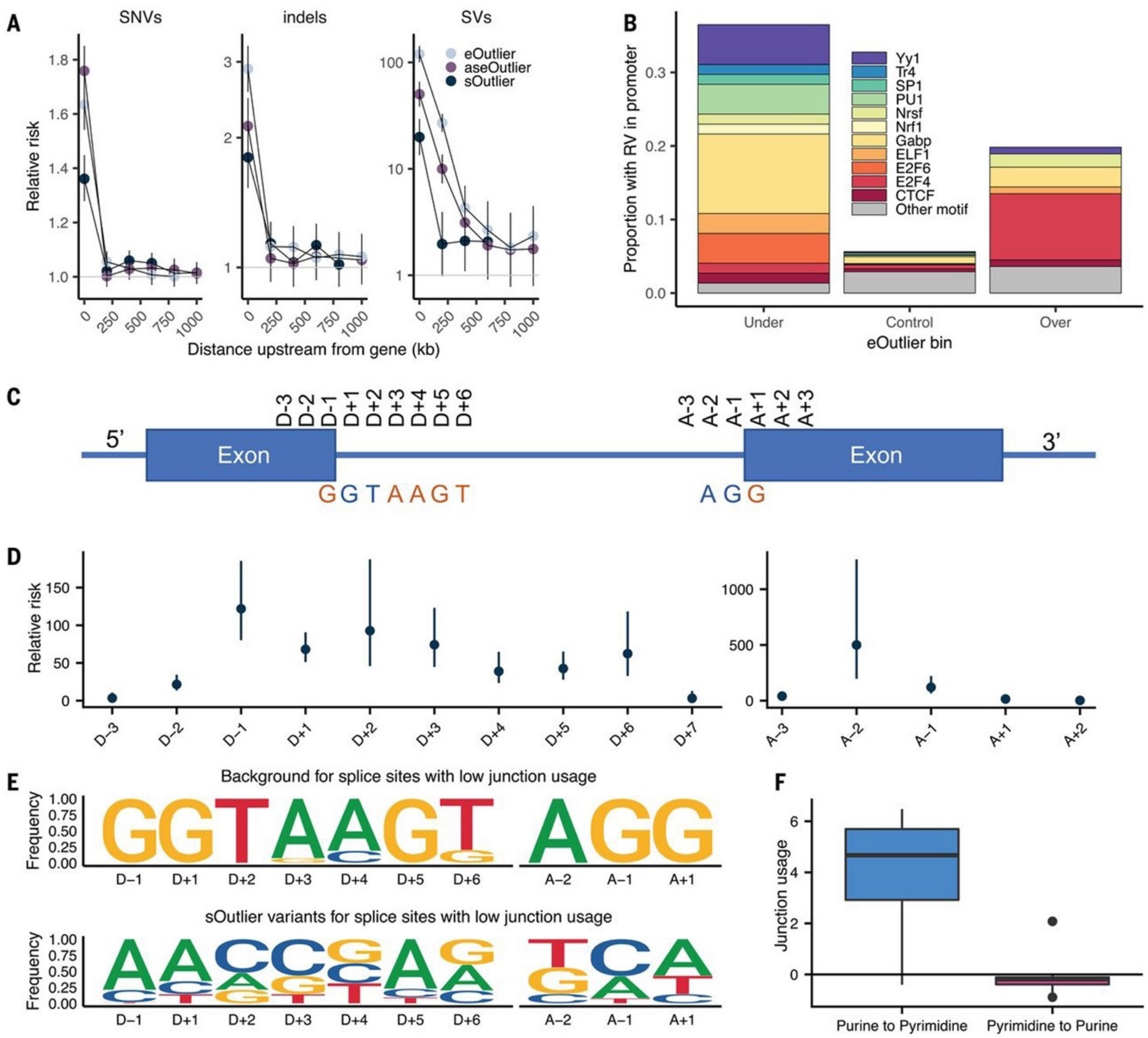


Figure 5-2: Impact of rare variation on splicing patterns

*(A) Relative risk of SNVs and indels (not found in gnomAD), and SVs (singleton in GTEx) at varying distances upstream of outlier genes (bins exclusive) across data types. (B) Proportion of eOutliers with TSS RVs in promoter motifs within 1000 bp. Under and over bins are defined with a median Z score threshold of 3, and controls are all individuals with a median Z score <3 for the same set of outlier genes. (C) Graphic summarizing positional nomenclature relative to observed donor and acceptor splice sites. (D) Relative risk (y-axis) of an sOutlier (median LeafCutter cluster $P < 1 \times 10^{-5}$) RV being located at a specific position relative to the splice site (x-axis) compared with nonoutlier RVs. Relative risk calculation was done separately for donor and acceptor splice sites. (E**)** Independent position weight matrices*

showing mutation spectra of sOutlier (median LeafCutter cluster $P < 1 \times 10^{-5}$*) RVs at positions relative to splice sites with negative junction usage (i.e., splice sites used less in outlier individuals than in nonoutliers). (F) Junction usage of a splice site is the natural log of the fraction of reads in a LeafCutter cluster mapping to the splice site of interest in sOutlier (median LeafCutter cluster* $P < 1 \times 10^{-5}$*) samples relative to the fraction in nonoutlier samples aggregated across tissues by taking the medianJunction usage (*y*-axis) of the closest splice sites to RVs that lie within a polypyrimidine tract* $(A - 5, A - 35)$ *binned by the type of variant (*x*-axis).*

RVs in splicing consensus sequence drive splicing outliers

Previous studies have shown RVs disrupting splice sites result in outlier alternative splicing patterns (71). We used sOutlier calls made for each LeafCutter cluster (16, 35) to assess enrichment of splicing-related variants more precisely. We observed extreme enrichment of RVs near splice sites in sOutliers. An sOutlier was 333 times more likely than a nonoutlier to harbor a RV within a 2-bp window around a splice site (Appendix D: Figure 13A) (see Methods), with signal decaying at greater distances but still enriched up to 100 bp away (relative risk = 7.43). To obtain base pair resolution enrichments, we computed the relative risk of sOutlier RVs located at specific positions relative to observed donor and acceptor splice sites (see Methods). Ten positions near the splice site showed significant enrichment for RVs in sOutliers compared with controls (Figure 5-2C,D). These positions corresponded precisely to positions that have also been shown to be intolerant to mutations because of their conserved role in splicing (we will refer to these positions as the splicing consensus sequence) (34). Among the most enriched positions within the splicing consensus sequence were the four essential splice site positions (D + 1, D + 2, A − 2, A − 1) (72), which showed an average relative risk of 195.

sOutliers further captured the transcriptional consequences both for variants that disrupted a reference splicing consensus sequence and those that created a new splicing consensus sequence. Individuals with sOutlier variants in which the rare allele deviated away from the splicing consensus sequence showed decreased junction usage of the splice site near the variant, whereas individuals with variants in which the rare allele created a splicing consensus sequence showed increased junction usage of the splice site near the variant relative to nonoutliers (Figure 5-2E and Appendix D: Figures S13B and S14) (see Methods). We saw a related enrichment pattern after separating annotated and new (unannotated) splice sites (Appendix D: Figure S15). sOutliers were also enriched for RVs positioned within the polypyrimidine tract (PPT), a highly conserved, pyrimidine-rich region, ~5 to 35 bp upstream from acceptor splice sites (73). A RV was 6.25 times more likely to be located in the PPT near an sOutlier relative to a nonoutlier. sOutliers with a RV that changed a position in the PPT from a pyrimidine to a purine (i.e., disrupting an existing PPT) showed decreased junction usage of the splice site near the variant, whereas the inverse was true for variants that changed a position in the PPT from a purine to pyrimidine (Figure 5-2F and Appendix D: Figure S16).

Prioritizing RVs by integrating genomic annotations with diverse personal transcriptomic signals

To incorporate diverse transcriptome signals into a method to prioritize RVs, we developed Watershed, an unsupervised probabilistic graphical model that integrates information from genomic annotations of a personal genome with multiple signals from a matched personal transcriptome. Watershed provides scores that can be used for personal genome interpretation or for cataloging potentially impactful rare alleles, quantifying the posterior probability that a variant has a functional effect on each transcriptomic phenotype based on both whole-genome–sequencing (WGS) and RNA-sequencing (RNA-seq) signals (Figure 5-3A). The Watershed model can be adapted to any available collection of molecular phenotypes, including different assays, different tissues, or different derived signals. Further, Watershed automatically learns Markov random field (MRF) edge weights reflecting the strength of the relationship between the different tissues or phenotypes included that together allow the model to predict functional effects accurately.

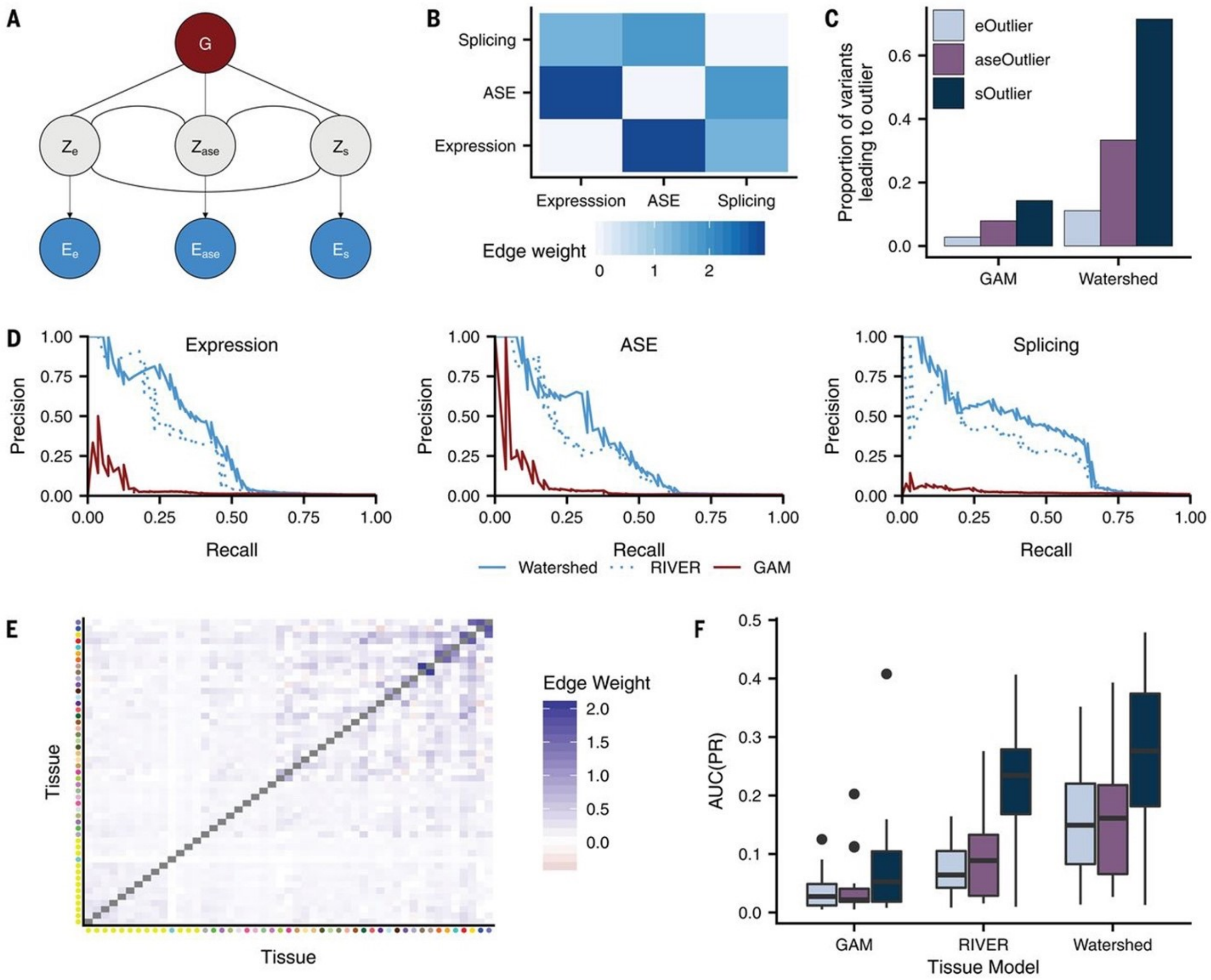


Figure 5-3: Watershed prioritizes functional rare genetic variants

(A) Graphic summarizing plate notation for the Watershed model when it is applied to three median outlier signals (expression, ASE, and splicing). (B) Symmetric heatmap showing learned Watershed edge parameters (weights) between pairs of outlier signals after training Watershed on three median outlier signals. (C) The proportion of RVs with Watershed posterior probability >0.9 (right) and with GAM probability greater than a threshold set to match the number of Watershed variants for each outlier signal (left) that lead to an outlier at a median P*-value threshold of 0.0027 across three outlier signals (colors). Watershed and GAM models were evaluated on held-out pairs of individuals. (D) Precision-recall curves comparing performance of Watershed, RIVER, and GAM (colors) using held-out pairs of individuals for three median outlier signals. (E) Symmetric heatmap showing learned tissue-Watershed edge parameters (weights) between pairs of tissue outlier signals after training tissue-Watershed on eOutliers across single*

tissues. (F) Area under precision recall curves [AUC(PR); y*-axis] in a single tissue between tissue-GAM, tissue-RIVER, and tissue-Watershed* (x*-axis) when applied to outliers across single tissues in all three outlier signals (colors). Precision recall curves in each tissue were generated using held-out pairs of individuals.*

We first applied Watershed to the GTEx v8 data using the three outlier signals examined here, expression, ASE, and splicing (Figure 5-3A) (see Methods), for which each was first aggregated by taking the median across tissues for the corresponding individual. In agreement with existing evidence of similarity between outlier signals (Appendix D: Figure S9), the learned Watershed edge parameters were strongest between ASE and expression, followed by ASE and splicing, but strictly positive for all pairs of outlier signals (i.e., each outlier signal was informative of all other signals; Figure 5-3B). To evaluate our model, we used held-out pairs of individuals that shared the same RV, making Watershed predictions in the first individual and evaluating those predictions using the second individual's outlier status as a label (see Methods). Watershed outperforms methods based on genome sequence alone [our genomic annotation model (GAM) and combined annotation-dependent depletion (CADD); Figure 5-3C and Appendix D: Figure S17] (74). We also compared performance of Watershed with RIVER [RNA-informed variant effect on regulation (21)], a simplification of the Watershed model in which each outlier signal is treated independently. We found that explicitly modeling the relationship between different molecular phenotypes provided a performance gain for Watershed (Figure 5-3D, Appendix D: Figures S18, S19, and Supplementary table 1) (see Methods). We observed that even the most predictive genomic annotations only resulted in eOutliers, aseOutliers, and sOutliers 2.8, 7.9, and

14.3% of the time, respectively (Figure 5-3C). However, integrating transcriptomic signals with genomic annotations from Watershed (at a posterior threshold of 0.9) detected SNVs that resulted in eOutliers, aseOutliers, and sOutliers with greater frequency 11.1, 33.3, and 71.4% of the time, respectively (Figure 5-3C and Appendix D: Figure S20).

We further extended the Watershed framework to prioritize variants on the basis of their predicted tissue-specific impact. We trained three "tissue-Watershed" models (one for each of expression, ASE, and splicing separately), in which each model considers the effects in all tissues jointly, sharing information in the MRF, and ultimately outputs 49 tissue-specific scores for each RV (Appendix D: Figures S19, S21) (see Methods). We observed that the parameters learned for each of the three tissue-Watershed models resembled known patterns of tissue similarity (Figure 5-3E and Appendix D: Figure S22). Further, using held-out individuals, the tissue-Watershed model outperformed a RIVER model in which each tissue is treated completely independently ($P = 2.00 \times 10^{-5}$, $P = 2.00 \times 10^{-5}$, and $P = 5.90 \times 10^{-3}$ for expression, ASE, and splicing, respectively; one-sided binomial test; Figure 5-3F and Appendix D: Figures S23, S24) and a collapsed RIVER model trained with single median outlier statistics ($P = 0.0577$, $P = 0.251$, and $P = 0.00128$ for expression, ASE, and spicing, respectively; one-sided binomial test; Appendix D: Figures S25, S26). Critically, integrative models that incorporated transcriptomic signal and genomic annotations from a single tissue still outperformed methods based only on genome sequence annotations (Figure 5-3F), supporting the benefit of collecting even a single RNA-seq sample to improve personal genome interpretation.

Replication and experimental validation of predicted RV transcriptome effects

We first assessed the replication of “candidate causal RVs” previously identified by the SardiNIA Project (75), using GTEx Watershed prioritization. Of five SardiNIA candidate causal RVs that were also present in a GTEx individual, four had high (>0.7) GTEx Watershed expression posterior probabilities (Appendix D: Supplementary Table 2). Next, we tested replication of GTEx RVs, prioritized by Watershed, in an independent cohort evaluating 97 whole-genome and matched transcriptome samples from the Amish Study of Major Affective Disorders (ASMAD) (76). We evaluated GTEx RVs also present in this cohort at any frequency, quantifying eOutlier, aseOutlier, and sOutlier signal in each ASMAD individual harboring one of the GTEx variants (see Methods). For all three phenotypes, ASMAD individuals with variants having high (>0.8) Watershed posterior probability based on GTEx data had significantly more extreme outlier signals at nearby genes compared with individuals with variants having low (<0.01) GTEx Watershed posterior probability (expression: $P = 2.729 \times 10^{-6}$, ASE: $P = 2.86 \times 10^{-3}$, and splicing: $P = 5.86 \times 10^{-13}$; Wilcoxon rank-sum test; Appendix D: Figure S27). Every variant with a high GTEx Watershed splicing posterior probability (>0.8) resulted in an sOutlier ($P \leq 0.01$) in the ASMAD cohort. Furthermore, ASMAD individuals with variants having high (>0.8) GTEx Watershed posterior probability had significantly larger outlier signals relative to equal size sets of variants prioritized by GAM (expression: P =

0.00129, ASE: *P* = 0.0287, and splicing: *P* = 0.00058; Wilcoxon rank-sum test; Appendix D: Figure S27). Overall, RVs prioritized by Watershed using GTEx data displayed evidence of functional effects in ASMAD individuals.

We further applied both a massively parallel reporter assay (MPRA) and a CRISPR-Cas9 assay to assess the impact of Watershed-prioritized RVs. We experimentally tested the regulatory effects of 52 variants with moderate Watershed expression posterior (≥0.5) and 98 variants with low Watershed expression posterior (<0.5) using MPRA (see Methods). We observed increased effect sizes for RVs with high Watershed expression posterior relative to variants with low expression posterior (*P* = 0.025; one-sided Wilcoxon rank-sum test; Appendix D: Figure S28). Next, we assessed the functional effects of 20 variants by editing them into inducible-Cas9 293T cell lines. These included 14 rare stop-gained variants and six non-eQTL common variants as negative controls. Of the 14 rare stop-gained variants, 13 had expression or ASE Watershed posterior >0.8, with the remaining variant [previously tested in (77)] having a posterior of 0.22. All control variants had Watershed posteriors <0.03. Of the 13 variants with a Watershed posterior >0.8, 12 showed a significant decrease in expression of the rare allele (*P* < 0.05, Bonferroni corrected; Appendix D: Figure S29) (see Methods).

Aberrant expression informs RV trait associations

We found that each individual had a median of three eOutliers, aseOutliers, and sOutliers (median outlier *P* < 0.0027) with a nearby RV. When filtering by moderate Watershed posterior probability (>0.5) of affecting expression, ASE, or splicing, individuals had a median of 17 genes with RVs predicted to affect expression, 27 predicted to affect ASE, and nine predicted to affect splicing (Figure 5-4A). From the set of outlier calls, we found multiple instances of RVs influencing well-known and well-studied genes, including *APOE* and *FAAH*. In particular, for *APOE*, which has been associated with numerous neurological diseases and psychiatric disorders (78), we found two aseOutlier individuals both carrying a rare, missense variant, rs563571689, with ASE Watershed posteriors >0.95, not previously reported. For *FAAH*, which has been linked to pain sensitivity in numerous contexts (79), we found two eOutlier individuals with a rare 5′ untranslated region variant, rs200388505, with ASE and expression Watershed posteriors >0.9.

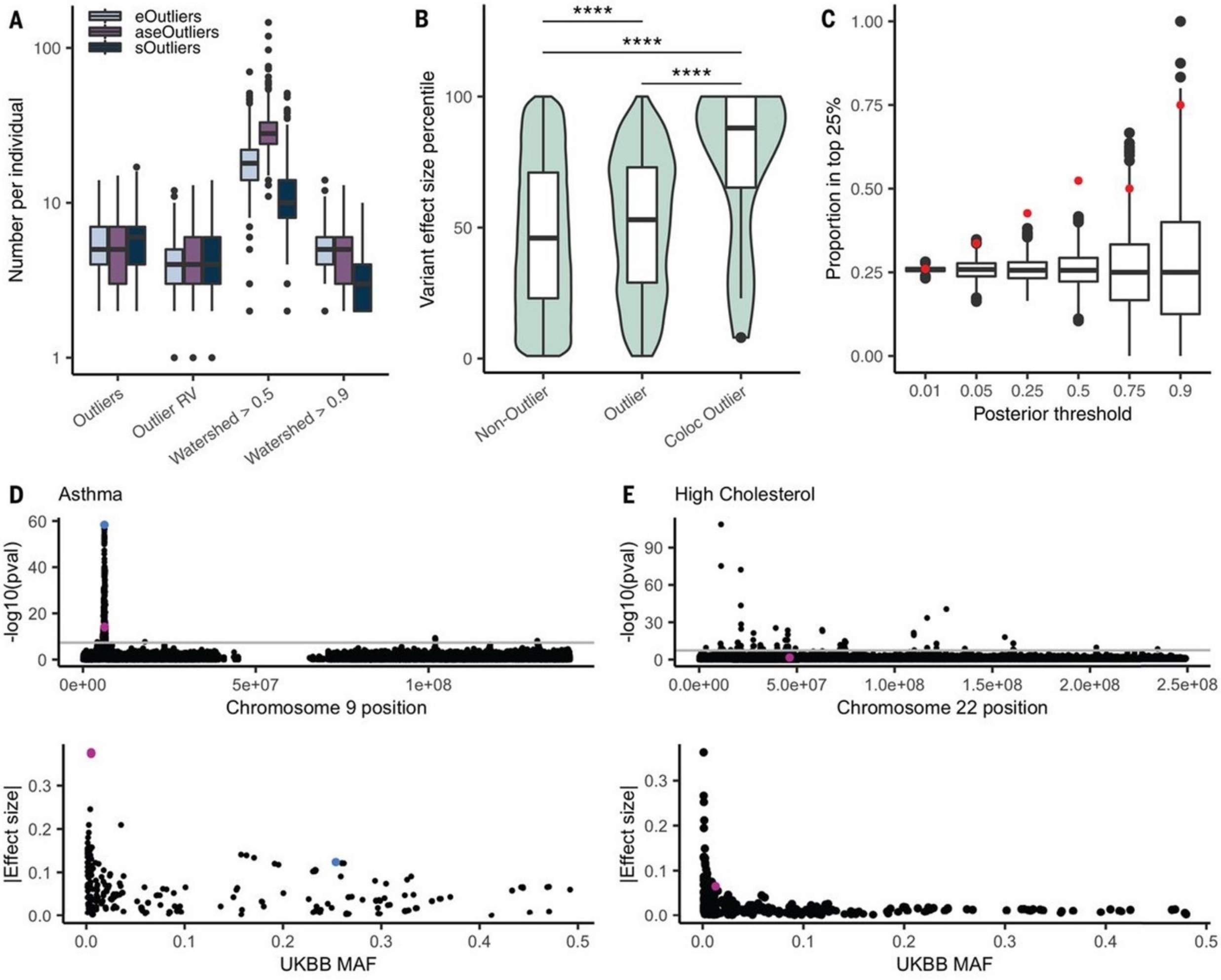


Figure 5-4: Functional rare variants identify trait relevant loci

(A) Distribution of the number of outlier genes, outlier genes with a nearby RV, and genes with a high Watershed posterior variant per data type. We added one to all values so that individuals with 0 are included. (B) Distribution of effect sizes, transformed to a percentile, for the set of GTEx RVs that appear in UKBB and are not outlier variants, those that are outlier variants, and those outlier variants that fall in colocalizing genes for the matched trait across 34 traits. Percentiles were calculated on the set of rare GTEx variants that overlap UKBB. The set of genes was restricted to those with at least one outlier individual in any data type and a nearby variant included in the test set (4787 variants and 1323 genes). P values were calculated from a one-sided Wilcoxon rank-sum test. (C) Proportion of variants filtered by Watershed posterior that fell in the top 25% of effect sizes for a colocalized trait (red) and the proportion of randomly selected variants of an equal number that also fall in these regions over 1000 iterations (black). (D) Manhattan plot (top) across chromosome 9 for asthma in the UKBB, filtered for

non–low-confidence variants, with two high-Watershed variants, rs149045797 and rs146597587, shown in pink and the lead colocalized variant, rs3939286, shown in blue. The variants' effect size ranks were similarly high for both self-reported and diagnosed asthma, but the summary statistics are shown for asthma diagnosis here. The UKBB MAF versus absolute value of the effect size for all variants within 10 kb of the Watershed variant is also shown (bottom). (E) Manhattan plot across chromosome 22 for self-reported high cholesterol in the UKBB, filtered to remove low confidence variants, with the high-Watershed variant rs564796245 shown in pink. The UKBB MAF versus absolute value of the effect size for all variants within 10 kb of the Watershed variant is also shown (bottom).

To assess whether identified rare functional variants from GTEx associate with traits, we intersected this set with variants present in the UKBB (61). We focused on a subset of 34 traits for which GWAS association for a UKBB trait had evidence of colocalizations with eQTLs and/or alternative splicing QTLs (sQTLs) in any tissue (see Methods). GTEx has demonstrated that genes with RV associations for a trait are strongly enriched for their eQTLs colocalizing with GWAS signals for the same trait, indicating that QTL evidence can be used to guide RV analysis. Furthermore, RVs near GTEx outliers had larger trait association effect sizes than background RVs near the same set of genes in the UKBB data (*P* = 3.51 × 10^{-9}; one-sided Wilcoxon rank-sum test), with a shift in median effect size percentile from 46 to 53%. Notably, outlier variants that fell in or nearby genes with an eQTL or sQTL colocalization had even larger effect sizes (median effect size percentile 88%) than nonoutlier variants (*P* = 1.93 × 10^{-5}; one-sided Wilcoxon rank-sum test) or outlier variants falling near any gene not matched to a colocalizing trait (*P* = 4.88 × 10^{-5}; one-sided Wilcoxon rank-sum test; Figure 5-4B).

Although most variants tested in UKBB had low Watershed posterior probabilities of affecting the transcriptome (Appendix D: Figure S30A), we hypothesized that filtering for those variants that do have high posteriors would yield variants in the upper end of the effect size distribution for a given trait. For each variant tested in UKBB, we took the maximum Watershed posterior per variant and compared this with a genomic annotation-defined metric, CADD (74). We found that Watershed posteriors were a better predictor of variant effect size than CADD scores for the same set of RVs in a linear model. Across different Watershed posterior thresholds, we found that the proportion of variants falling in the top 25% of RV effect sizes in colocalized regions exceeded the proportion expected by chance (Figure 5-4C). Whereas filtering by CADD score did return some high effect size variants, this proportion declined at the highest thresholds (Appendix D: Figure S30D). Furthermore, there was very little overlap between variants with high Watershed posteriors and high CADD variants (Appendix D: Figure S30D), with CADD variants more likely to occur in coding regions and Watershed variants more frequent in noncoding regions (Appendix D: Figure S30D). Thus, the approaches largely identified distinct and complementary sets of variants for these traits.

We identified 33 rare GTEx variant trait combinations in which the variant had a Watershed posterior >0.5 and fell in the top 25% of variants by effect size for the given trait. We highlight two such examples, for asthma and high cholesterol (Figure 5-4D,E), showing that although RVs usually do not have the frequency to obtain genome-wide significant *P* values, when they are prioritized by the probability of affecting expression, we could identify those with greater estimated effect sizes on the trait. In the case of asthma, the RV effect sizes in UKBB were three times greater than the lead colocalized variant. These variants included rs146597587, which is a high-confidence loss-of-function splice acceptor with an overall gnomAD AF of 0.0019, and rs149045797, an intronic variant with a frequency of 0.0019, both of which were associated with the gene *IL33*, the expression of which has been implicated in asthma (80). Previous work has identified the protective association between rs146597587 and asthma (81), and we found that this is potentially mediated by outlier allelic expression of *IL33* leading to moderate decreases in total expression, with median *Z* scores ranging from −1.08 to −1.77 in individuals with the variant, and median single-tissue *Z* scores across the six individuals exceeding −2 in 10 tissues. An asthma association had also been reported recently for the other high Watershed asthma-associated variant rs149045797 and was in perfect linkage disequilibrium with rs146597587 (82). An additional high Watershed variant, rs564796245, an intronic variant in *TTC38* with a gnomAD AF of 0.0003, had a high effect size for self-reported high cholesterol in the UKBB but was not previously reported. We were able to test this variant against four related blood lipids traits in the MVP (83). We found that for these traits, which included high-density lipoprotein (HDL), low-density lipoprotein, total cholesterol, and triglycerides, among rare (gnomAD AF <0.1%) variants within a 250-kb window of rs564796245, this variant was in the top 5% of variants by effect size; for HDL specifically, it was in the top 1% (Appdenix D: Figure S31). We also assessed this variant's association with the same four traits in the JHS (63), an African American cohort in which four individuals carried the RV. Here, we found that the direction of effect was consistent with MVP and UKBB for all four traits, and the variant fell in the top 28th to 38th percentile of all rare (gnomAD AF <0.1%) variants in this region (Appendix D: Figure S32). Only four of the variants tested in UKBB had Watershed posterior probabilities >0.9 for colocalized genes, but of those, three showed high effect sizes for a relevant trait.

Methods

GTEx data

All human donors were deceased, and informed consent was obtained via next-of-kin consent for the collection and banking of deidentified tissue samples for scientific research. The research protocol was reviewed by Chesapeake Research Review Inc., Roswell Park Cancer Institute’s Office of Research Subject Protection, and the institutional review board of the University of Pennsylvania. We used the RNA-sequencing, allele-specific expression, and whole-genome sequencing (WGS) data from the v8 release of the GTEx project and assessed expression data across the 49 biological tissues with at least 70 samples. Sample size varied across tissues, with average missingness of ~50%. Self-reported ancestry for these individuals spanned three of the major continental populations with the majority (n=714 with WGS) comprising individuals of predominantly European ancestry, 121 individuals with African ancestry, 11 with Asian ancestry, and 12 unknown or other.

Rare variant annotations

We retained all SNVs and indels that passed quality control in the GTEx VCF, variant calling described in, using the hg38 genome build. Structural variants were called according to on the subset of individuals available from V7 with GenomeSTRiP GSCNQUAL set to limit the false discovery rate (FDR) for each variant type. Genome STRiP's IntensityRankSumAnnotator was used to evaluate FDR based on available Illumina Human Omni 5M gene expression array data. GSCNQUAL was limited to ≥ 1 for GenomeSTRiP deletions and ≥ 8 for multi-allelic copy number variants, corresponding to an FDR of 10%. The GSCNQUAL cutoff for GenomeSTRiP duplications was set at ≥ 17, the point where the FDR plateaued at 15.1% and did not fluctuate more than ±1% for over 50 steps in increasing GSCNQUAL score. Additionally, the Mobile Element Locator Tool (MELT) version 2.1.4 was run using MELT-SPLIT to identify ALU, SVA, and LINE1 insertions into the test genomes. MELT calls that were categorized as "PASS" in the VCF info field, had an ASSESS score ≥ 3, and SR count ≥ 3 were retained. Structural variant (SV) calls were then lifted to the hg38 genome build using liftOver from the Genome Browser.

We defined rare variants as those with < 1% MAF within GTEx and, for SNVs and indels, also occurring at < 1% frequency in non-Finnish Europeans within gnomAD Novel variants were those that occurred in GTEx but were not found in gnomAD. GTEx singletons had an average allele frequency of 0.0030 in gnomAD and doubletons had an average frequency of 0.0096.

Annotation of protein-coding regions and transcription factor binding site motifs was generated by running Ensembl VEP (version 88). Loss of function (LoF) annotation was generated using loftee. Conservation scores (Gerp, PhyloP, PhastCons) were downloaded from UCSC genome browser and CADD scores were extracted from a pre-compiled annotation file (https://cadd.gs.washington.edu/download) using variant scores from the hg38 genome build.

Expression outlier calling

Within each tissue, we log2-transformed the expression values ($\log_2$(TPM + 2)), where TPM is the number of transcripts per million mapped reads, generated by RNA-SeQC using the GENCODE v26 gene annotation, available through the GTEx portal. We subsetted to autosomal lincRNA and protein-coding genes and restricted to genes with at least 6 reads and TPM > 0.1 in at least 20% of individuals. We scaled the expression of each gene to mean of 0 and standard deviation of 1 to avoid the deflation of outlier values caused by quantile normalization. As we expected unmeasured technical confounders to impact expression, for each tissue we estimated hidden factors for the transformed expression matrix using PEER. The number of PEER factors retained was based on sample size and matched the values chosen in the GTEx eQTL analyses, which were 15 for sample sizes less than or equal to 150, 30 for less than 250, 45 for less than 350, and 60 otherwise. We obtained expression residuals by regressing out PEER factors, the top three genotype principal components, sex, and the genotype of the strongest cis-eQTL per gene in each tissue using the following linear model:

$$Y_g = \mu_g + \sum_{n=1}^{N} \alpha_{g,n} P_n + \sum_{k=1}^{3} \beta_{g,k} G_k + \gamma_g S + \delta_g Q + \varepsilon_g$$

where Y_g is the transformed expression of gene g, μ_g is the mean expression level for the gene, P_n is the nth PEER factor, G_k are the top k genotype principal components, S

is the sex covariate, and Q is the genotype of the strongest cis-eQTL for gene g. We then re-scaled the expression residuals ε_g for each gene, to obtain corrected expression Z-scores for each individual per gene per tissue.

For each gene, we calculated an individual's median Z-score across all tissues for which data were available, restricting to individuals with measurements in at least five tissues. To account for situations where widespread extreme expression might occur in an individual due to non-genetic influences, we excluded 39 individuals where the proportion of tested genes that were multi-tissue outliers at a threshold of |median Z-score| > 3 exceeded 1.5 times the interquartile range of the distribution of proportion outlier genes across all individuals. We then use the median Z-scores per individual across tissues to determine eOutliers and used a threshold of |median Z| > 3 or an equivalent median p-value of 0.0027 for aseOutliers and sOutliers to determine the outlier set of genes. This threshold was chosen to balance the number of outliers identified with increases in nearby rare variant enrichments, though the conclusions are robust to threshold choice (Appendix D: Figures S1D, S7). Controls were defined as any individual with a |median Z-score| of less than 3 (or another threshold as indicated) for the same set of genes as those with any outlier individual. We allowed a gene to have multiple outlier individuals and an individual could be an outlier for multiple genes. Code for generating eOutlier calls was modified from scripts available at https://github.com/joed3/GTExV6PRareVariation.

ASE outlier calling

Allelic expression (ASE) data was produced as described in (64). We used the Analysis of Expression VAriation Dosage Outlier Test (ANEVA-DOT) to identify genes in each individual that showed an excessive imbalance of ASE, relative to the population. Briefly, ANEVA-DOT relies on tissue-specific estimates of genetic variation in gene dosage, V^G, derived by Analysis of Expression VAriation (ANEVA) on a reference population ASE data to identify genes in individual test samples that are likely affected by rare variants with unusually large regulatory effects. We calculated reference V^G estimates from GTEx v8 data from 15,201 RNA-seq samples spanning 49 tissues and 838 individuals with WGS data. Across all analyzed tissues we estimated V^G a total of 2,727,867 times using all available autosomal aeSNPs (variants used to assess allelic expression) with at least 30 reads in 6 individuals. These estimates are publicly available at https://doi.org/10.5281/zenodo.3897759, version 2.31. We used the ANEVA-DOT tool R package (https://doi.org/10.5281/zenodo.3406690) to calculate a p-value for every gene-individual pair with allelic expression data and a corresponding V^G estimate (Appendix D: Figure S3). The p-value can be interpreted as the result of a binomial test of allelic imbalance, that is overdispersed for each gene individually according to its expected dosage variation in a given tissue in the population. Genes with significant ANEVA-DOT p-values are referred to as aseOutliers in this text. We tested all tissues available for each GTEx v8 individual, using only genes with a minimum coverage of 8 reads spanning an aeSNP and with V^G estimates available (49 tissues, median genes per tissue = 4899, Appdendix D: Figure S2). For each gene expressed we considered the aeSNP with the highest coverage in an individual.

For all single-tissue analyses, we removed global outlier genes and individuals from each tissue group independently. Global outlier genes are likely to be ASE outliers at 5% FDR in more than 1% of tested individuals per tissue, as has been previously described in (64). These genes are likely to have poor V^G estimates due to the presence of different ASE patterns within the gene or other global biological factors. Outlier individuals were also defined as in ((64)), and were removed from downstream single tissue analysis. These samples contain an unusually high number of outliers ($n >$ Q3+1.5*IQR) at 5% FDR, and are likely to be caused by technical errors. Tissue specific lists of global outlier genes and individuals for outlier threshold of 5% FDR are available here: https://doi.org/10.5281/zenodo.3899574. In all other analyses unless otherwise specified, we did not apply an FDR control procedure and instead imposed a higher threshold for declaring significance, to be consistent with expression and splicing outliers. For cross-tissue analyses, we calculated median ANEVA-DOT p-values for genes which were expressed in more than 5 tissues, without removing known global outliers first. Therefore, to account for genes with poor V^G, we applied the filtering steps described in (64) on the resulting individual-level median p-values. Briefly, we removed individuals with too few genes tested ($n <$ Q1-1.5*IQR), removed individuals with too many outliers ($n >$ Q3+1.5*IQR), and removed genes which appeared as outliers too many times across individuals with a score available (genes that are likely to be called as outliers in more than 1% of cases, Appendix D: Figure S2). To define multi-tissue outliers, we used a threshold of median p-value < 0.0027, equivalent to |median Z| > 3, to determine outlier status.

Split read count quantification and processing

LeafCutter (64) provided an annotation-free approach for RNA splicing quantification allowing us to capture split reads overlapping rare exon-exon junctions. Junctions were extracted from WASP-corrected BAM files with a modified version of the “bam2junc.sh” script from LeafCutter that only retained reads that passed WASP filters (9). Next in each tissue separately, junction reads were clustered using the “leafcutter_cluster.py” script from LeafCutter, with the options “--maxintronlen 500000” and “mincluratio 0”. LeafCutter assigns exon-exon junctions into mutually exclusive sets, termed clusters. Each exon-exon junction in a cluster had to share a splice site with at least one other exon-exon junction in that cluster, but could not share a splice site with an exon-exon junction from another cluster. A cluster had to contain at least two exon-exon junctions.

Next, in each tissue separately, we applied the following series of custom filters to the LeafCutter results in order to remove exon-exon junctions with low expression while retaining rare exon-exon junctions:

1. Removed exon-exon junctions where no sample has >= 15 split reads
2. Re-defined LeafCutter cluster assignments after removal of exon-exon junctions (according to the above filter) and removed exon-exon junctions that no longer shared a splice site with any other exon-exon junction.
3. Removed all exon-exon junctions belonging to a LeafCutter cluster where more than 10% of the samples had less than 3 reads summed across all exon-exon junctions assigned to that LeafCutter cluster.

Next, we merged LeafCutter cluster assignments across all 49 tissues to make a specific LeafCutter cluster comparable across multiple tissues. For this, we re-defined LeafCutter cluster assignments using the union of all exon-exon junctions that passed the above filters across 49 tissues. Lastly, we mapped our LeafCutter clusters to genes by intersecting splice sites, defining a Leafcutter cluster with splice sites of annotated exons. We limited to genes used in expression outlier calling (described in “Expression outlier calling” section). If an annotated splice site was in a LeafCutter cluster, we considered the LeafCutter cluster mapped to the gene. It was therefore possible for a LeafCutter cluster to map to multiple genes. We filtered LeafCutter clusters, and their corresponding exon-exon junctions, to those that were mapped to at least one gene. Finally, we removed any LeafCutter clusters with more than 20 exon-exon junctions due to computational limitations of SPOT.

SPOT: Overview

sOutliers were identified separately for each LeafCutter cluster in each tissue using Splicing Outlier deTection (SPOT). For a given LeafCutter cluster in a given tissue, we defined a matrix, X (dim NxJ), where each row corresponds to one of N samples, each column corresponding to one of J exon-exon junctions, and each element was the number of raw split read counts corresponding to that row’s sample and that column’s exon-exon junction. We were able to compute a p-value representing how abnormal a given sample’s splicing patterns were for the given LeafCutter cluster as follows:

1. Fitted parameters of Dirichlet-Multinomial distribution based on observed data X in order to capture the distribution of split read counts mapping to this LeafCutter cluster
2. Used fitted Dichlet-Multinomial distribution to compute the Mahalanobis distance for each of the N samples
3. Computed Mahalanobis distance for 1,000,000 samples simulated from the fitted Dirichlet-Multinomial and use these 1,000,000 Mahalanobis distances as an empirical distribution to assess the significance of the N real Mahalanobis distances

SPOT: Dirichlet-Multinomial parameter estimation

We defined a Dirichlet-Multinomial (DM) probability distribution based on data from N samples to capture the probability that a split read would map to each of the J junctions in the Leafcutter cluster:

Let x_{nj} be the raw number of split reads mapped to the j^th junction in the n^th sample and $t_n = \sum_J x_{nj}$ be the total number of split reads mapped to any junction in this LeafCutter cluster in the nth sample. Then

$x_{n1}, \ldots, x_{nJ} | t_n \sim DM(t_n, \alpha_1 p_1, \ldots, \alpha_J p_J$ where $p_j = \frac{\exp{(c_j)}}{\sum_k^J \exp{(c_k)}}$

We used the following non-informative Gamma prior distribution to stabilize optimization:

$$\alpha_j \sim Gamma(1 + 1e^{-4}, 1e^{-4})$$

We then performed maximum likelihood estimation (via LBFGS as implemented in STAN) to learn the optimal parameter settings of α and c_j

SPOT: Mahalanobis distance

The Mahalanobis distance is the multivariate generalization of how many standard deviations a point is from the mean taking into account the covariance structure. After learning the parameters of the Dirichlet-Multinomial distribution for a specific LeafCutter cluster (ie $\widehat{\alpha_1},\ldots,\widehat{\alpha_J}$and $\widehat{c_1},\ldots,\widehat{c_J}$; see "SPOT: Dirichlet-Multinomial parameter estimation"), we were able to compute the mean vector (μ_n) and covariance matrix (Σ_n) for a specific sample n, according to the Dirichlet-Multinomial. Using $\mu_n$and $\Sigma_n$we were able to compute the Mahalanobis distance of sample n (MD_n). The covariance matrix of the Dirichlet-Multinomial (Σ_n) is of rank $J-1$ because one of the dimensions is always a linear combination of the other $J-1$ dimensions. As such, we approximated ${\Sigma_n}^{-1}$with the pseudo-inverse of $\Sigma_n$when computing the Mahalanobis distance.

SPOT: Empirical distribution to assess significance

For a given LeafCutter cluster, we have already computed the Mahalanobis distance of each of the N samples according to the fitted Dirichlet-Multinomial distribution for that LeafCutter cluster. However, the Mahalanobis distance is biased by the dimensionality of the space (i.e. the number of junctions assigned to the LeafCutter cluster). In order to convert the Mahalanobis distance to a test statistic that was not biased by dimensionality, we simulated an empirical distribution of Mahalanobis distances for each LeafCutter cluster. Specifically, for one LeafCutter cluster we drew 1,000,000 random samples from the fitted Dirichlet-Multinomial distribution assuming each of these random samples has 20,000 reads mapped to the LeafCutter cluster ($t_n = 20000$). We then computed the Mahalanobis distance of each of these 1,000,000 samples and used the 1,000,000 Mahalanobis distances as an empirical distribution that converted our N Mahalanobis distances (from the real data) into p-values.

SPOT: Gene level correction

To compute a splicing outlier p-value for a gene associated with C LeafCutter clusters, we first computed minimum p-value across all C clusters for the gene. However, the minimum of a list of p-values is not a valid p-value. To address this, we computed the probability of observing a minimum p-value according to a probability density function defining the minimum across C independent uniform random variables between 0 and 1:

$$p(min(pvalue_1, \ldots, pvalue_C) <= z) = 1 - (1 - z)^C$$

This approach made the conservative, simplifying assumption that all clusters mapped to a gene were independent of one another.

We excluded individuals (global outliers) where the proportion of tested genes that were multi-tissue outliers (at a threshold of median p-value < .0027) exceeded 1.5 times the interquartile range of the distribution of proportion outlier genes across all individuals.

SPOT: Robustness to hyperparameter choice

SPOT, under default settings, makes the assumption that each random sample, used in generating an empirical distribution for each LeafCutter cluster, has 20,000 total reads mapped to that cluster (see “SPOT: Empirical distribution to access significance”). To understand if our sOutlier p-values were sensitive to the choice of 20,000 total reads, we re-computed sOutlier calls in Muscle-Skeletal tissue using SPOT with 10,000 total reads and 100,000 total reads (Appendix D: Figure S6). sOutlier p-values generated from SPOT under default settings (20,000 reads) are highly correlated to sOutlier p-values generated from SPOT using 10,000 reads (Spearman’s ρ = .997) and 100,000 reads (Spearman’s ρ = .997). Only .052% and .046% of sample-LeafCutter cluster pairs had a -log10(p-value) change greater than 1 between SPOT under default settings compared to SPOT run with 10,000 and 100,000 reads, respectively. All of the sample-LeafCutter cluster pairs that had a -log10(p-value) change greater than 1 correspond to LeafCutter clusters where more than 95% of the total observed reads mapping to the cluster, summed across samples, map to a single exon-exon junction. These rare instances of divergence in sOutlier p-values between SPOT under different hyperparameter settings are caused by numerical instability in computing the pseudo-inverse (See “SPOT: Mahalanobis distance”) when distributions are heavily skewed towards a particular junction.

SPOT, under default settings, uses a Gamma prior (on each α_j) when fitting a Dirichlet-Multinomial distribution to each LeafCutter cluster (See “SPOT: Dirichlet-Multinomial parameter estimation”). This prior is intended to stabilize the LBFGS-based optimization routine, while having minimal consequences on parameter estimates. To see if the prior had minimal impact on parameter estimates, we re-computed sOutlier calls in Muscle-Skeletal tissue using a version of SPOT where no prior was used (Appendix D: Figure

S6). To encourage SPOT with no prior to converge to a reasonable estimate, we performed Dirichlet-Multinomial parameter estimation 10 times (with 10 random initializations) and selected the Dirichlet-Multinomial parameter estimate whose expected value had the smallest Euclidean norm with expected value of the maximum likelihood estimate of a Multinomial distribution fitted to the same data. sOutlier p-values generated from SPOT using default settings (ie. with the prior) are highly correlated to sOutlier p-values generated from SPOT when no prior is used (Spearman's ρ = .997) .Only .049% of sample-LeafCutter cluster pairs had a -log10(p-value) change greater than 1 between SPOT under default settings compared to SPOT with no prior. Similar to the above comparison of SPOT using variable number of simulated reads, these rare instances of divergence in sOutlier p-values between SPOT with and without a prior are caused by numerical instability in computing the pseudo-inverse when distributions are heavily skewed towards a particular junction.

Enrichment calculations

We calculated relative risk enrichments as the proportion of outliers with a given variant type nearby the outlier gene over the proportion of non-outlier individuals with the given variant type nearby the same set of genes. We included 95% confidence intervals estimated via a normal approximation. When assessing rare variant enrichments overall and by category, we used a 10kb +/- window around the gene body. When considering variant categories per outlier, if more than one rare variant was present nearby the outlier gene, we assigned each gene-individual to a single variant category based on the following ordering: duplications (DUP), copy number variations (CNV), deletions

(DEL), breakend (BND), inversions (INV), transposable elements (TE), splice, frameshift, stop, transcription start site (TSS), conserved non-coding, coding, or other non-coding, and subsetted to the 527 individuals with structural variant calls. Unless otherwise specified, we used a threshold of median p-value < 0.0027 (chosen to match |median Z-score| > 3) to define multi-tissue outliers, though provide results over a range of thresholds (Appendix D: Figure S1). A categorical model of outlier status was used as opposed to a continuous model because small changes in continuous outlier p-values do not reliably reflect true biological effects due to technical variation from RNA-sequencing, as well as to demonstrate the impact of thresholding choices for downstream applications. Additionally, this allows for matching the genes included in both the outlier and control category, defining an appropriate background distribution for statistical hypothesis testing so that we are not simply identifying differences between genomic regions rather than individual genetic effects on a given gene's expression. When considering variants in different windows upstream from the gene, we constructed exclusive distance ranges from each gene, beginning with the gene body + 10kb window used previously, and then we intersected rare variants with windows 1bp-200kb, 200kb-400kb, 400kb-600kb, 600kb-800kb, and 800kb-1Mb upstream from the set of outlier genes.

Alternative splicing enrichment calculations

We performed several enrichment analyses specific to splicing outliers to better characterize the variants underlying splicing outliers. For all of these analyses, we used sOutlier calls at the LeafCutter cluster level (instead of the gene level) in order to get more accurate enrichments. We excluded individuals identified as global outliers at the gene level (see “SPOT: gene level correction”). We limited enrichment analysis to SNVs. We used a stringent median p-value threshold of 1 x 10^{-5} in order to isolate the highest confidence instances of outlier splicing, according to SPOT. In Appendix D: Figure S13A, we show the relative risk of rare variants nearby splice sites is robust to a range of median p-value thresholds and becomes more enriched at more stringent p-value thresholds.

1. Relative risk of rare variant in window around splice site. We computed the relative risk of rare variants being located at various windows around splice sites for outlier clusters relative to non-outlier clusters. For example, if the window was [0,2], we mapped a variant to a cluster if that variant were less than or equal to two base pairs away from observed donor and acceptor splice sites ([D-2, D+2] and [A-2, A+2] based on notation in Figure 5-2C) for that cluster. Relative risk was then calculated as the proportion of outlier (LeafCutter cluster, individual) pairs with a mapped rare variant over the proportion non-outlier (LeafCutter cluster, individual) pairs with a mapped rare variant, while limiting analysis to LeafCutter clusters with at least one outlier individual. We included 95% confidence intervals estimated via a normal approximation.
2. Relative risk of rare variant at position relative to splice site. We first mapped rare variants to clusters if the rare variants were less than or equal to 1000 base pairs from an observed donor or acceptor splice site ([A-1000, A+1000] and [D-1000, D+1000] based on notation in Figure 5-2C). We then mapped each variant to its nearest splice site in that cluster and calculated its position relative to that splice

site. Then, to compute the positional relative risk at position D-1 (for example), we computed the fraction of outlier variants mapped to a donor splice site that were at position D-1 over the fraction of non-outlier variants mapped to a donor splice site that were at position D-1. We added a constant of 1 to all counts in the contingency table to stabilize enrichments. We included 95% confidence intervals estimated via a normal approximation.

3. Junction Usage for splicing median p-value outliers. We used the "junction usage" statistic to quantify whether an individual used a splice site more or less than the background population. A positive junction usage value intuitively means the individual uses the splice site more than the background population, while a negative junction value means an individual uses a splice site less than the background population. More concretely to compute the junction usage for an individual i and junction j, we first computed the following ratio in each tissue (in which that individual i is expressed) separately: $\frac{Fraction\ of\ reads\ in\ cluster\ mapping\ to\ junction\ j\ for\ individual\ i}{Fraction\ of\ reads\ in\ cluster\ \ mapping\ to\ junction\ j\ for\ non{-}outliers\ individuals}$

   We added a constant of 1 to the above contingency table to stabilize enrichments. The "junction usage" statistic is simply the natural logarithm of the median of the above statistic across all tissues in which individual i is expressed.

Watershed model overview

Watershed is a hierarchical Bayesian model that predicts regulatory effects of rare variants on a specific outlier signal based on the integration of multiple transcriptomic signals along with genomic annotations describing the rare variants. Watershed models instances of (gene, individual) pairs to predict the regulatory effects of rare variants nearby the gene. The Watershed model for a particular (gene, individual) pair, assuming K outlier signals, consists of three layers (Figure 5-3a):

1. A set of variables **G** $= G_1, \dots, G_P$representing the P observed genomic annotations aggregated over all rare variants in the individual that are nearby the gene.
2. A set of binary latent variables **Z** = $Z_1, \dots, Z_K$representing the unobserved functional regulatory status of the rare variants on each of the K outlier signals. Let Z^sbe the set of all possible values that **Z** can take on. The size of $Z^S$is 2^K.
3. A set of categorical nodes **E** = $E_1, \dots, E_K$that represents the observed outlier status of the gene for each of the K outlier signals. We allow for missingness in **E**.

A fully connected conditional random field (CRF) is defined over variables Z given G, where we let W represent the set edges among Z. Variables E$_i$ are each connected only to the corresponding latent variable Z$_i$. Specifically, the following conditional distributions together define the full Watershed model:

A. $Z \mid G \sim CRF(\alpha, \beta_1, \dots, \beta_k, \theta)$

B. $E_k \mid Z_k$ ~Categorical(ϕ_k) $\forall\ k \in K$

C. ϕ_k ~Dirichlet$(C, \dots, C)$

D. β_k ~Normal$(0, \frac{1}{\lambda})$

where,

- $\beta_k \in R^P \forall\ k\ \in K$are the parameters defining the contribution of the genomic annotations to the CRF for each outlier signal (k)
- $\alpha \in R^K$are the parameters defining the intercept of the CRF for each outlier signal (k)
- $\theta \in\ R^{(Kchoose2)}$ are the parameters defining the edge weights between pairs of outlier signals (Notational note: $\theta_{tq}\ =\ \theta_{qt}$)
- $\phi_k \forall\ k\ \in K$are the parameters defining the categorical distributions of each outlier signal
- $C$and λ are hyper-parameters of the model

Explicitly, our CRF probability distribution is defined as:

$P(Z \mid G, \beta_1, \ldots, \beta_K, \alpha, \theta)\ =\ exp(\sum_{k \in K} \quad \alpha_k Z_k\ +\ \sum_{(t,q) \in W} \quad \theta_{tq} Z_t Z_q + \sum_{k \in K} \quad \beta_k G\ Z_k\ -\ A(G, \theta, \beta_1, \ldots, \beta_K))$

where $A(G, \theta, \beta_1, \ldots, \beta_K)\ =\ log(\sum_{Z^* \in Z^S} \quad exp\ (\sum_{k \in K} \quad \alpha_k {Z^*}_k \quad + \sum_{(t,q) \in W} \quad \theta_{tq} {Z^*}_t {Z^*}_q + \sum_{k \in K} \quad \beta_k G\ {Z^*}_k))$

Because **Z** is unobserved, the Watershed log-likelihood objective over instances $n = 1, \ldots, N$:

$$\sum_{n=1}^{N} \quad log \sum_{Z^* \quad \in Z^S} \quad P(E_n, G_n, Z^* \quad \mid \beta_1, \ldots, \beta_K, \alpha, \theta, \phi_1, \ldots, \phi_K)$$

is non-convex. We therefore optimize model parameters using Expectation-Maximization (EM) as described in the following sections.

Watershed exact inference optimization routine

When the number of outlier signals (K) is small (an approximate rule being 4 or less), Watershed parameters can be optimized using exact inference updates within EM as follows:

In the E-step for instances $n = 1, \dots, N$: we compute posterior distributions over the latent variables ($Z^{(n)}$), conditioned on the current model parameters ($\beta_1, \dots, \beta_K, \alpha, \theta, \phi_1, \dots, \phi_K$) and the observed data ($G^{(n)}$ and $E^{(n)}$). For example, the joint posterior probability of $Z^{(n)} = Z$ for the nth instance can be computed as:

$$\omega^{(n)}(Z^{(n)} = Z)$$

$$= exp(\sum_{k \in K} (\alpha_k Z_k + \beta_k G^{(n)} Z_k + I(E_k{}^{(n)}) log(P(E_k{}^{(n)} | Z_k))$$

$$+ \sum_{(t,q) \in W} \theta_{tq} Z_t Z_q$$

$$- A(G^{(n)}, E^{(n)}, \theta, \beta, \alpha, \theta, \phi)$$

$$A(G^{(n)}, E^{(n)}, \theta, \beta, \alpha, \phi)$$

$$= log(\sum_{Z^* \in Z^S} exp\,(\sum_{k \in K} (\alpha_k Z^*{}_k + \beta_k G^{(n)} Z^*{}_k$$

$$+ I(E_k{}^{(n)}) log(P(E_k{}^{(n)} | Z^*{}_k))\ )$$

$$+ \sum_{(t,q)\in W} \theta_{tq} {Z^*}_t {Z^*}_q))$$

where,

$I({E_k}^{(n)})$ is an indicator function for whether ${E_k}^{(n)}$is observed. Given the joint posterior probability distribution, we can marginalize (sum over) specific dimensions (outlier signals) to obtain:

1. Marginal posterior distributions for each dimension i (where $Z^W$is the set of all possible values that **Z** can take on excluding dimension i):

$$\omega^{(n)}{}_{single}(Z_i) = \sum_{Z^* \in Z^W} \omega^{(n)}(Z^*)$$

2. Pairwise marginal posterior distributions for each pair of dimensions i, j (where $Z^W$is the set of all possible values that **Z** can take on (excluding dimension i and dimension j)):

$\omega^{(n)}{}_{pair}(Z_i, Z_j) = \sum_{Z^* \in Z^W} \omega^{(n)}(Z^*)$

Both the marginal posterior distributions and the pairwise marginal posterior distributions are used in the M-step as follows. We update β, α, and $\theta$by optimizing the conditional random field as follows:

$$argmax_{\beta,\alpha,\theta} \sum_{n=1}^{N} \sum_{Z^* \in Z^S} log(P(Z^* \mid G^{(n)}, \beta \ , \alpha, \theta))\omega^{(n)}(Z^*) - \frac{\lambda}{2}||\beta||_2 - \frac{\lambda}{2}||\theta||_2$$

Here λ is an L2 penalty hyper-parameter derived from the Gaussian priors on $\beta$and θ. We optimized this objective function by running L-BFGS on the closed-form gradient updates.

In the second part of the M-step, we update $\phi_k \forall\ k\ \in K$as follows:

$$\phi_k(s,t)\ = \sum_{n=1}^{N}\quad I(E_k{}^{(n)} =\ t)\ \omega^{(n)}{}_{single}(Z_k{}^{(n)}\ =\ s)\ +\ C$$

where,

I is an indicator operator, $t$is the categorical value of expression $E_k{}^{(n)}$, $s$is the possible binary values of $Z_k{}^{(n)}$, and C is the hyperparameter based on the Dirichlet prior on ϕ.

Once the EM algorithm has converged, we use the marginal posterior distributions for each dimension i in each instance n ($\omega^{(n)}{}_{single}(Z_i\ = 1)$) as estimates of probability that the nth (gene, individual) pair has a nearby variant that has a functional effect on the gene (with respect to outlier dimension i).

Watershed approximate inference optimization routine

When the number of outlier signals (K) is large (an approximate rule being 5 or more), it becomes computationally intractable to optimize Watershed parameters using exact inference updates, so we use approximate inference updates within EM as follows:

For the E-step, we wish to compute approximate estimates of the following posterior probability distribution:

$$\omega^{(n)}(Z^{(n)} = Z)$$

$$= exp(\sum_{k \in K} (\alpha_k Z_k + \beta_k G^{(n)} Z_k + I(E_k{}^{(n)}) log(P(E_k{}^{(n)} | Z_k))$$

$$+ \sum_{(t,q) \in W} \theta_{tq} Z_t Z_q$$

$$- A(G^{(n)}, E^{(n)}, \theta, \beta, \alpha, \theta, \phi)$$

$$A(G^{(n)}, E^{(n)}, \theta, \beta, \alpha, \phi)$$

$$= log(\sum_{Z^* \in Z^S} exp(\sum_{k \in K} (\alpha_k Z^*{}_k + \beta_k G^{(n)} Z^*{}_k$$

$$+ I(E_k{}^{(n)}) log(P(E_k{}^{(n)} | Z^*{}_k))$$

$$+ \sum_{(t,q) \in W} \theta_{tq} Z^*{}_t Z^*{}_q))$$

To approximate this function $\omega^{(n)}(Z^{(n)})$, we use the Mean-Field Approximation (a subclass of Variational Inference) and optimize $q^{(n)}(Z^{(n)})$ to minimize the KL-divergence between $q^{(n)}(Z^{(n)})$ and $\omega^{(n)}(Z^{(n)})$

where,

$q^{(n)}(Z^{(n)}) = \prod_{k \in K} q_k{}^{(n)} (Z_k{}^{(n)})$where $q_k{}^{(n)} (Z_k{}^{(n)}) = (\mu_k{}^{(n)})^{z_k{}^{(n)}} (1 - \mu_k{}^{(n)})^{(1 - z_k{}^{(n)})}$

To minimize the KL-divergence for a given sample n, we perform coordinate descent on each $\mu_k{}^{(n)}$while holding all other dimensions (values of$\mu_j{}^{(n)}$) constant. Given that$N(k)$represents the set of all nodes that share an edge with node k, the variational update for each $\mu_k{}^{(n)}$is then:

$$\mu_k{}^{(n)(update)} = \frac{exp(a_k + I(E_k{}^{(n)}) log(P(E_k{}^{(n)} \mid Z_k{=}1)))}{exp\ (I(E_k{}^{(n)}) log(P(E_k{}^{(n)} \mid Z_k{=}0)) + exp(a_k + I(E_k{}^{(n)}) log(P(E_k{}^{(n)} \mid Z_k{=}1)))}$$

where $a_k = \alpha_k + \beta_k G^{(n)} + \sum_{j \in N(k)} \quad \theta_{kj} {\mu_j}^{(n)}$

More specifically, for one instance n, we iteratively do the following until convergence:

1. Loop through all K dimensions in a random order, and update each ${\mu_k}^{(n)}$given the most recent values of ${\mu_j}^{(n)} \forall\, j \in N(k)$. Since coordinate ascent is not guaranteed to reach the global optimum, we used damped updates for each ${\mu_k}^{(n)} \forall\, k \ \in K$in order to decrease the chance of getting stuck at a local optimum:
    a. ${\mu_k}^{(n)(iter\ i+1)} = \ (1-\eta) * {\mu_k}^{(n)(iter\ i)} \ + \ (\eta) * {\mu_k}^{(n)(update)}$
    b. We use a damping value (η) of 0.8.
2. Compute the average difference, across all K dimensions, between the values of ${\mu_k}^{(n)}$from the current iteration and values of ${\mu_k}^{(n)}$ from the previous iteration. Converge if the average difference is less than 1x 10^{-8}.

Using the same notation as in “Watershed exact inference optimization routine”, Mean Field allows us to approximate the following expectations using converged estimates of ${\mu_k}^{(n)}$:

1. $\omega^{(n)}(Z^{(n)}) \ \approx \ \prod_{k \in K} \quad ({\mu_k}^{(n)})^{{z_k}^{(n)}} (1 - {\mu_k}^{(n)})^{(1-{z_k}^{(n)})}$
2. ${\omega^{(n)}}_{pair}({Z_i}^{(n)}, {Z_j}^{(n)}) \approx ({\mu_i}^{(n)})^{{z_i}^{(n)}} (1 - {\mu_i}^{(n)})^{(1-{z_i}^{(n)})} \, ({\mu_j}^{(n)})^{{z_j}^{(n)}} (1 - {\mu_j}^{(n)})^{(1-{z_j}^{(n)})}$
3. ${\omega^{(n)}}_{single}({Z_i}^{(n)}) \ \approx ({\mu_i}^{(n)})^{{z_i}^{(n)}} (1 - {\mu_i}^{(n)})^{(1-{z_i}^{(n)})}$

We use both the approximate marginal posterior distributions and the approximate pairwise marginal posterior distributions in the M-step. However, when the number of dimensions (K) is large, optimization of the parameters (β, α, and θ) defining the conditional random field becomes intractable. Therefore, we approximated the CRF objective function with the Pseudolikelihood of the CRF. Given variational estimates of ${\mu_i}^{(n)}({Z_i}^{(n)})$ for all values of dimensions (i) and all samples (n), the (log) Pseudolikelihood objective function (including priors on coefficients) is given by:

$$\sum_{n=1}^{N} \quad \sum_{k \in K} \quad (\alpha_k {\mu_k}^{(n)} + \beta_k G^{(n)} {\mu_k}^{(n)} + \sum_{j \in N(k)} \quad \theta_{kj} {\mu_k}^{(n)} {\mu_j}^{(n)} - A(k, n, \theta, \beta, \alpha))$$
$$- \frac{\lambda}{2} ||\beta||_2 - \frac{\lambda}{2} ||\theta||_2$$

$$A(k, n, \theta, \beta, \alpha) = log(\sum_{z=0}^{1} \quad exp(\alpha_k z + \beta_k G^{(n)} z + \sum_{j \in N(k)} \quad \theta_{kj} \, z \, {\mu_j}^{(n)}))$$

We computed closed form gradient updates of the above objective function and then optimized it using L-BFGS.

In the second part of the M-step, we update $\phi_k \forall \, k \; \in K$as follows:

$$\phi_k(s, t) = \sum_{n=1}^{N} \quad I({E_k}^{(n)} = t) \, \omega^{(n)}{}_{single}({Z_k}^{(n)} = s) + C$$

Where I is an indicator operator, $t$is the categorical value of expression ${E_k}^{(n)}$, $s$is the possible binary values of ${Z_k}^{(n)}$, and C is the hyperparameter based on the Dirichlet prior on ϕ.

Once the EM algorithm has converged, we use marginal posterior distributions for each dimension i, in each instance n ($\omega^{(n)}{}_{single}(Z_i = 1)$) as estimates of probability that the nth (gene, individual) pair has a nearby variant that has a functional effect on the gene (with respect to outlier dimension i).

GAM and RIVER

The genomic annotation model (GAM) is L2-regularized logistic regression using genomic annotations (**G**) as features and the binary outlier status of a specific outlier signal as the response variable. One GAM model was trained for each outlier signal.

The only difference between Watershed and RIVER is that in RIVER θ is fixed to be a vector of zeros. This allows RIVER to be optimized precisely as described in "Watershed exact inference optimization routine" assuming $\theta$is fixed to be zero. It is important to note that RIVER has changed slightly since its initial development (21) in the following way: we now use a categorical distribution (ϕ) with three categories instead of two to model $E \mid Z$. This change in RIVER was made in order to make it directly comparable to Watershed.

Applying Watershed to jointly model ASE, splicing, and expression

We first applied Watershed to the GTEx v8 data using 3 outlier signals: median ASE, splicing, and expression. Recall, Watershed requires a set of genomic annotations (**G**) and a corresponding set of categorical outlier signals (**E**) over (gene, individual) instances. We first limited to a set of (gene, individual) pairs with a rare variant that fell within the gene body or +/- 10kb of each gene and that passed the following set of filters in all 3 outlier signals:

1. The individual was not a global outlier
2. The gene has measured outlier signal for the corresponding individual
3. The gene has at least one individual that is an outlier (median p-value < .01)

This yielded a set of 36,702 (gene, individual) pairs that we used for training and evaluating the Watershed framework.

To generate the genomic annotations (**G**) for each (gene, individual) pair, we limited to SNVs that fell within the gene body or +/- 10kb of each of the gene and then extracted 47 genomic annotations describing each of the SNVs including regulatory element annotations, conservation scores, and derived genomic scores from other models such as CADD. If a (gene, individual) pair had more than one SNV mapped to the gene, the genomic annotations were aggregated across the SNVs with simple transformations to generate gene-level genomic annotations. The resulting gene-level genomic annotations were standardized (mean 0 and standard deviation 1) before running Watershed. 1.93 x 10^{-5}

We generated the categorical outlier signals (**E**) for each (gene, individual) pair using 3 categories per outlier signal. It is important to note that because of the filters described above there is no missingness in **E**. For aseOutliers and sOutliers, we assigned a gene with median p-value (p) to:

1. Category 1 if $-log_{10}(p + 10^{-6}) < 1$
2. Category 2 if $1 <= -log_{10}(p + 10^{-6}) < 4$
3. Category 3 if $-log_{10}(p + 10^{-6}) >= 4$

For eOutliers, we assigned a gene with median p-value (p) and median Z-score (z) to:

1. Category 1 if $-log_{10}(p\ +\ 10^{-6})\ >1$and z < 0
2. Category 2 if $-log_{10}(p\ +\ 10^{-6})\ <=\ 1$
3. Category 3 if $-log_{10}(p\ +\ 10^{-6})\ >1$and z > 0

We note that these thresholds are arbitrary, but were selected to distinguish non-outliers, moderate outliers, and extreme outliers for aseOutliers and sOutliers, and distinguish non-outliers, under-expression outliers, and over-expression outliers for eOutliers.

To train and evaluate Watershed, we identified the 3,411 cases where two or more individuals had the same rare SNV(s) near a particular gene. We held out those instances and trained Watershed on the remaining instances. For training, we set the hyperparameter C equal to 30, motivated by the number of training instances. To select the hyperparameter λ, we trained and evaluated GAM on the training data for each outlier signal independently (assigning a sample an outlier label if outlier p-value < .01) with 5-fold cross validation while running a gridsearch on λ=.1,.01,.001. We selected the λ with the largest median area under the precision recall curve (AUPRC) across the 5 folds. Each precision recall curve aggregated predictions across the three outlier signals. The optimal λ was found to be 0.001. Before running Watershed, we initialized $\alpha_k$and $\beta_k$to be the intercept and slope parameters, respectively, of GAM (when $\lambda = 0.001$) trained on the full training data for outlier signal k. θ was initialized to all zeros. $\phi_k$was initialized using the MAP updates described in “Watershed exact inference optimization routine", except we used the GAM (when $\lambda = 0.001$) posterior probabilities to approximate $\omega^{(n)}{}_{single}({Z_k}^{(n)} \ = \ s)$.

We evaluated various trained models (Watershed, RIVER, GAM, CADD) using the 3,411 cases where two individuals had the same rare SNV(s) near a particular gene (we will refer to these instances as N2 pairs). Specifically, we estimated the posterior probability of a functional rare variant (according to each of the models) in the first individual from the pair, allowing Watershed to use all data available for that individual. We then used the outlier status of the second individual as a 'label' for evaluation. In order to make the fraction of outlier instances comparable between different outlier signals, we defined a (gene, individual) pair to be an outlier for a specific outlier signal if its outlier p-value was ranked amongst the 1% most significant p-values for that outlier signal (across training and N2 pair instances). For an N2 pair, we did this evaluation in both directions: predict on the first individual and evaluate on the second, as well as predict on the second individual and evaluate on the first. Importantly, none of the N2 pairs were used in training any of the models.

Watershed with data generated using various filters

Recall from the previous section ("Applying Watershed to jointly model ASE, splicing, and expression"), Watershed training data was generated through the following approach: we limited to a set of (gene, individual) pairs with a rare variant that fell within the gene body or +/- 10kb of each gene and that passed the following set of filters in all 3 outlier signals:

1. The individual was not a global outlier
2. The gene has measured outlier signal for the corresponding individual
3. The gene has at least one individual that is an outlier (median p-value < 0.01)

These strict thresholds were set in order to reduce the imbalance between outliers and non-outliers in the training data set. We next assessed how sensitive Watershed was to these filters by training Watershed with three different training data sets generated by relaxing the above third filter as follows:

- All 3 outlier signals have at least one individual that is an outlier (median p-value < 0.05)
- All 3 outlier signals have at least one individual that is an outlier (median p-value < 0.1)
- At least 1 outlier signal has at least one individual that is an outlier (median p-value < 0.01)

We evaluated various trained models (Watershed, RIVER, GAM) using held out pairs of individuals generated under the default filtering in order to make precision-recall curves comparable to those in Figure 5-3D (Appendix D: Figures S18A-C, Table S2). We found the improvements of Watershed over RIVER decreased when using training data generated under more relaxed thresholds, while the improvements of Watershed and RIVER relative to GAM remained. The increased class imbalance (resulting from the relaxed thresholds) caused the fraction of positive training instances to decrease. This further imbalance resulted in Watershed learning considerably smaller magnitude edge weights, increasing the similarity of the Watershed model with the RIVER model.

We therefore recommend using training data generated through our default filtering approach when running Watershed.

We further accessed sensitivity of our analysis to these filters by training Watershed with training data generated through our default filtering approach, while evaluating Watershed on three different sets of held out pairs of individuals generated by relaxing the above third filter as follows:

- All 3 outlier signals have at least one individual that is an outlier (median p-value < 0.05)
- All 3 outlier signals have at least one individual that is an outlier (median p-value < 0.1)
- At least 1 outlier signal has at least one individual that is an outlier (median p-value < 0.01)

Importantly, improvements of Watershed over both RIVER and GAM were robust to relaxing these thresholds . Specifically, the difference in AUPRC between Watershed and RIVER, when evaluating performance on default held out pairs of individuals, is strictly bounded above zero for splicing, but for other phenotypes there is some overlap. But the difference in AUPRC between Watershed and RIVER is strictly bounded above zero for all phenotypes when evaluating on a larger set of held out pairs of individuals selected with less stringent filters (Appendix D: Table S2, Fig S21).

Applying Watershed to jointly model outlier signals from each tissue (tissue-Watershed)

Next, we trained three independent tissue-Watershed models (one each for ASE, splicing, and expression) where each model considered effects in all tissues, giving 49 phenotypes, corresponding to 49 Z and E variables each. In order for these models to be comparable to the model described in “Applying Watershed to jointly model three outlier types”, we used the same set of (gene, individual) pairs. We therefore used the same extracted and processed genomic annotations (**G**).

We generated the categorical outlier signals (**E**) for each (gene, individual) pair in a particular tissue (for a particular outlier signal) using 3 categories. It is important to note that, unlike the first application of Watershed to three median signals, there is now missingness in **E** as a (gene, individual) pair does not have measured outlier signal across all 49 tissues in GTEx. For ASE and splicing outliers, for a particular tissue, we assigned a gene with p-value (p) to:

1. Category 1 if $-log_{10}(p + 10^{-6}) < 1$
2. Category 2 if $1 <= -log_{10}(p + 10^{-6}) < 4$
3. Category 3 if $-log_{10}(p + 10^{-6}) >= 4$

For expression, outliers, for a particular tissue, we assigned a gene with p-value (p) and Z-score (z) to:

1. Category 1 if $-log_{10}(p + 10^{-6}) > 1$and z < 0
2. Category 2 if $-log_{10}(p + 10^{-6}) <= 1$
3. Category 3 if $-log_{10}(p + 10^{-6}) > 1$and z > 0

To train and evaluate tissue- Watershed, we identified the 3,411 cases where two individuals had the same rare SNV(s) near a particular gene. We held out those instances and trained Watershed on the remaining instances. For training, we set the hyperparameter C equal to 10, motivated by the number of training instances with observed outlier calls. We selected $\lambda = 0.001$based on cross-validation in "applying Watershed to jointly model three outlier types". We initialized $\alpha_t$and $\beta_t$to be the intercept and slope parameters, respectively, of GAM (when $\lambda = 0.001$) trained on the full training data from tissue t. θ was initialized to all zeros. $\phi_t$was initialized using the MAP updates described in "Watershed exact inference optimization routine", except we used the GAM (when $\lambda = 0.001$) posteriors to approximate $\omega^{(n)}{}_{single}(Z_k{}^{(n)} = s)$.

We took a very similar approach as described in "Applying Watershed to jointly model ASE, splicing and expression" to evaluate various trained models (tissue-Watershed, tissue-RIVER, tissue-GAM). In this setting however, both model predictions and outlier labels were in a single tissue as opposed to the median across tissues. As **E** contains missingness in this setting, we required both individuals in the N2 pair to have observed outlier signal for the gene of interest in the corresponding tissue.

Non-parametric bootstrapping of change in area under precision recall curves

We utilize non-parametric bootstrapping to assess the significance of the difference in area under a precision recall curve for two different models (assume the two models are called “model 1” and “model 2”, respectively). Assume there are Nobservations involved in generating the precision-recall curves, meaning there exist Npredictions from model 1, Npredictions from model 2, and Nbinary labels. We can then compute the area under the precision recall curve for model 1 and model 2 ($auprc_1$ and $auprc_2$, respectively), as well as the difference between the areas ($\Delta auprc) = auprc_1 - auprc_2$). Next, we generate Bnon-parametric bootstrapped samples of $\Delta auprc$. To generate one non-parametric bootstrapped sample (b) of $\Delta auprc$ we:

1. Randomly sample, with replacement N observations from the original N observations
2. Generate ${auprc_1}^{(b)}$ and ${auprc_2}^{(b)}$according to the sub-sampled observations.
3. Compute $\Delta auprc^{(b)} = {auprc_1}^{(b)}$- ${auprc_2}^{(b)}$

We can compute a 95% confidence interval on $\Delta auprc$ using the Bbootstrapped samples by first computing the .025 quantile and .975 quantile (across the B bootstrapped samples) of $\Delta auprc^{(b)} - \Delta auprc$($\delta_{.025}$and $\delta_{.975}$, respectively). The 95% confidence interval is then $[\Delta auprc - \delta_{.975}, \Delta auprc - \delta_{.025}]$.

Rare variant Watershed posterior predictions with trained Watershed model

We used the Watershed model that was previously trained on the 34,837 (gene, individual) pairs described in “Applying Watershed to jointly model ASE, splicing, and expression” to make Watershed posterior predictions on the remainder of rare variants in GTEx. To make genomic annotations comparable, the genomic annotations describing the SNVs we wish to predict on were standardized according to the mean and standard deviation of the genomic annotations from “Applying Watershed to jointly model ASE, splicing, and expression”. It is important to note that the Watershed model was trained across (gene, individual) pairs and predictions were made across (gene, SNV, individual) triplets.

Note on applying Watershed to new data sets

While we are restricted here to making predictions of variant effect on transcriptomic signals, our framework, including enrichment analysis and Watershed, could be straightforwardly applied to ribosome profiling data and/or mass-spectrometry based protein measurements by researchers using a cohort with WGS or exome sequencing to capture post-translational and structural changes.

Replication in ASMAD Cohort

As previously reported (76), 394 family members were genotyped on Illumina Omni 2.5 arrays and 80 individuals were subjected to whole genome sequencing by Complete Genomics. Genotyping was performed at the Center for Applied Genomics and the Children's Hospital of Pennsylvania. Genotype based identity by state metrics validate all familial relationships in the pedigree. All variants with Mendelian inconsistencies or missing in more than 1% of individuals were removed. Haplotypes were phased using SHAPEIT2 with duoHMM. Imputation was performed using IMPUTE2 and the TopMed Anabaptist reference panel of haplotypes. LCL lines from 100 individuals of the pedigree were obtained from the Coriell Institute. These individuals represent the 80 individuals who had been whole genome sequenced, plus an additional 20 closely related individuals.

Total RNA was extracted from LCL cultures using RNAeasy. Paired end RNA sequencing libraries were constructed using the Illumina [TruSeq stranded mRNA library prep kit] (http://www.illumina.com/products/truseq_stranded_mrna_library_prep_kit.html) with 100 independent index barcodes. Paired, 125bp reads were generated on an Illumina HiSeq2500 at the Next Generation Sequencing Core Facility at the University of Pennsylvania. Read level quality was assessed using FastQC. Reads were trimmed to remove Illumina adapters and low quality sequence using TrimGalore! ('stringency 5, length 50, q 20'). Reads were aligned to the human genome (hg38) with GENCODE gene annotations (v24) using the STAR aligner in 2-Pass mode. Gene level read counts were quantified using Feature Counts. After genotype and RNAseq quality control, 97 samples were included for further analysis.

To control for reference mapping bias and remove reads derived from PCR duplication, reads aligned to the human genome were processed using WASP. At each heterozygous site, reference and alternate allele read depth was quantified using PySam. Overlapping read pairs were only counted once. Splicing clusters were identified within each sample using Leafcutter.

aseOutlier calls in the ASMAD cohort were generated as follows. Allele specific read counts were generated with quasar. ASE snps were annotated by overlapping with coding regions of the genome. Then, for all ASE snps which overlapped a gene, the one with the highest read coverage was used to represent that gene's ASE counts. Mono-allelic sites and sites with fewer than 5 reads per allele were discarded. Genes which appeared as frequent outliers in GTEx LCL samples (available at https://doi.org/10.5281/zenodo.3899574) were removed as well. ANEVA-DOT was then run on all available genes per individual, using LCL V^G scores from GTEx as the reference. The results across all 97 available samples were compared, and individuals with more than 61 ASE outliers, after FDR correction (11 in total), were removed from downstream analysis. On average an individual in the ASMAD cohort had 176 ASE outlier genes, before FDR correction.

We next called sOutliers in the ASMAD cohort. As there are relatively few ASMAD RNA-seq samples (n=97), we used Dirichlet-Multinomial parameter estimates for each LeafCutter cluster learned from GTEx Cells EBV-transformed lymphocyte samples and then assessed how extreme each ASMAD sample was according that pre-trained distribution. More specifically, we first filtered ASMAD exon-exon junction counts to exon-exon junctions that passed the filters involved in processing GTEx Cells EBV-transformed lymphocytes (see "Split read count quantification and processing"). Then for each Leafcutter Cluster tested with SPOT in the GTEx Cells EBV-transformed lymphocytes tissue, we:

1. Retrieved Dirichlet-Multinomial parameter estimates for this LeafCutter cluster from when SPOT was trained using GTEx Cells EBV-transformed lymphocytes samples.

2. Generated a junction count matrix for the ASMAD samples. This junction count matrix will be of dimension NXJ where N is the number of ASMAD samples and J is the number of junctions assigned to this tissue in GTEx Cells EBV-transformed lymphocytes. If a particular junction in this cluster is not expressed in the ASMAD cohort, the column corresponding to this junction in the matrix will be filled in with zeros.
3. Used the GTEx-fitted Dirichlet-Multinomial distribution (from step 1) to compute the Mahalanobis distance of each of the N ASMAD samples.
4. Computed Mahalanobis distance for 1,000,000 samples simulated from the fitted Dirichlet-Multinomial and used these 1,000,000 Mahalanobis distances as an empirical distribution to assess the significance of the N real Mahalanobis distances.

We then converted from ASMAD sOutlier p-values at the LeafCutter cluster level to sOutlier p-values at the gene level using the approach described in “SPOT: Gene level correction”. We excluded individuals (global outliers) where the proportion of tested genes that were outliers (at a threshold of p-value < .0027) exceeded 1.5 times the interquartile range of the distribution of proportion outlier genes across all individuals.

Finally, we called eOutliers in the ASMAD cohort. As there are relatively few ASMAD RNA-seq samples (n=97), we concatenated ASMAD samples and GTEx Cells EBV-transformed lymphocyte samples and called eOutliers across the concatenated samples. More specifically, we first computed the TPM of each sample-gene pair independently for the ASMAD samples and the GTEx Cells EBV-transformed lymphocyte samples (using transcript lengths specific to each study). Next we concatenated the two TPM matrices into one large TPM matrix of dimension NXG where $N$is the sum of the number of samples in ASMAD and the number of samples in GTEx Cells EBV-transformed lymphocyte tissue, and G is the number of genes used in the GTEx Cells EBV-transformed lymphocyte eOutlier analysis. We then filtered to genes where at least 10% of the total samples (N) have greater than or equal to 6 raw counts and have greater than .1 TPM. We next log2-transformed the expression values ($\log_2$(TPM + 2)). We then scaled the expression of each gene to have mean 0 and standard deviation of 1, regressed out the top 30 principal components, and finally standardized each gene, again, to have mean 0 and standard deviation 1. We excluded individuals (global outliers) where the proportion of tested genes that were outliers (at a threshold of |Z-score| > 3) exceeded 1.5 times the interquartile range of the distribution of proportion outlier genes across all individuals.

UKBB and MVP GWAS integration

We assessed GWAS summary statistics from the UK Biobank (UKBB) phase 2, as made available by the Neale lab (http://www.nealelab.is/uk-biobank/). We subsetted the variants, either genotyped or imputed, in UKBB phase 2 to those SNVs that also appeared in any GTEx individuals and had a frequency of < 1% in GTEx, which resulted in 45,415 SNVs, filtered to those not flagged as low confidence due to very low allele counts. Because we are targeting rare variants occurring at frequencies too low to obtain a trait association with genome-wide significance, we focused on the effect size estimates and did not filter by p-value. We defined outlier variants in this context as any rare variant appearing near an eOutlier, sOutlier, or aseOutlier in GTEx and also appearing in UKBB. We defined non-outlier variants as rare GTEx variants appearing in UKBB, but not falling near an outlier of any type, though within 10kb of a gene for which any individual was an outlier. We subsetted to 34 traits tested for colocalization between the UKBB GWAS and GTEx eQTL/sQTL studies. When filtering to colocalized regions, we included as a colocalization event any gene that had a colocalization posterior probability > 0.5, for both eQTLs and sQTLs. We combine both enloc and coloc (41) results for eQTL colocalization and enloc results for sQTL colocalization. This resulted in 5,386 gene-trait pairs with significant co-localizations across 34 UKBB traits (Table S9). We transformed the |effect sizes| to percentiles, based on all rare GTEx SNVs that also appear in any UKBB samples tested for the included traits. When showing actual beta values for binary traits, we scaled according to the case-control ratio μ for the given trait, dividing the effect size estimates by $\mu * (1-\mu)$.

We filtered the set of GTEx rare variants in UKBB to those in colocalized regions, defined as being in a colocalized gene or within 10kb, and by the maximum Watershed posterior for that variant-gene combination across all data types (ASE, splicing, expression) and all tested individuals. We compared this to a genomic annotation based metric, CADD. We obtain an effect size β for both Watershed posterior and CADD score in predicting variant effect size percentiles in co-localized regions using the following model: $P \sim \beta X + \varepsilon$, where P is a vector of variant effect size percentiles and X is a vector of either Watershed posteriors or CADD scores for the same variant set.

We calculated the proportion of resulting variants that fall in the top 25% of effect sizes within colocalized regions for the associated trait across a range of posterior thresholds. We compared that proportion to the set we would obtain if filtering by a CADD score chosen to return an equal number of variants, prior to intersecting with colocalized regions. Additionally, we took 1000 random samples from the set of rare variants of an equal number to the actual number obtained by filtering at each threshold and assessed the proportion of random variants that fall in the top 25% of effect sizes for each colocalized trait. For replication in the Million Veterans Program (MVP) and Jackson Heart Study (JHS), we obtained summary statistics for a 250kb region on either side of the variant of interest for four lipid associated traits. We calculated the |effect size| percentile for all rare variants (gnomAD AF < 0.1%) in that region and plot the absolute effect sizes vs the gnomAD allele frequency.

Code availability

SPOT code can be found here: https://github.com/BennyStrobes/SPOT. Watershed code can be found here: https://github.com/BennyStrobes/Watershed. Code to generate all figures in this manuscript can be found here: https://zenodo.org/record/3885823#.YWnrONnMJTY.

Discussion

RVs are abundant in human genomes, yet they have remained difficult to study systematically. Using multitissue transcriptome and whole-genome data from GTEx v8, we have been able to identify and assess the properties of RVs, including SVs, that underlie extreme changes in expression, alternative splicing, and ASE.

We observed that each signal informs distinct classes of RVs, demonstrating the benefit of integrating multiple sources of personal molecular data to improve variant interpretation. We expanded characterization of the properties of RVs in multiple contexts, including structural variants affecting multiple genes, rare splice variants that disrupt or create splicing consensus sequences, and RVs occurring in tissue-specific enhancers leading to tissue-specific eOutliers. Together, these provide a map of the properties of large-effect RVs, aiding their identification and evaluation in future studies. We note that although our approach can be used to identify some large-effect RVs underlying disease, it is unlikely to capture the full spectrum of functional RVs contributing to heritability because some effects will not manifest as clear transcriptome aberrations.

We further developed a probabilistic model for personal genome interpretation, Watershed, which improves standard methods by integrating multiple transcriptomic signals from the same individual. Relevant to ongoing efforts to identify RVs affecting human traits, we found that in RVs within trait-colocalized regions, filtering by Watershed posteriors can identify variants with larger trait effect sizes better than relying on genomic annotations alone. As further demonstrated by our discovery of outlier RVs in well-studied disease genes, application of Watershed and other integrative methods will prove increasingly helpful for cataloging and prioritizing RVs affecting traits, especially those at the lowest ends of the AF spectrum. Our results provide a means to improve the quality and extent of RV prioritization, with potential future impacts enhancing RV association testing and disease gene identification.

Chapter 6 Conclusions and future directions

My PhD work focused on modeling the impact of genetic variation on gene expression. A complete, mechanistic understanding of genetic effects on gene expression could provide immense insights into disease development and ultimately human health. Most notably, this dissertation advances our understanding of how genetic regulation of gene expression changes as a function of cellular context or environment. Secondly, this dissertation provides a novel approach to identify functional rare variants via the incorporation of gene expression data. More specifically, there are four major projects that compose this dissertation.

The first project quantified the patterns of tissue-specificity of genetic regulation of gene expression. We found cis-acting genetic variants tend to affect either most tissues or a small number of tissues. By contrast, identified trans-eQTL effects tend to be tissue-specific and correspondingly show greater enrichment in enhancer regions.

The second project composing my dissertation modeled dynamic genetic regulation of gene expression during cellular differentiation. To achieve this, our collaborators generated time-series RNA-sequencing data, capturing 16 time points from induced pluripotent stem cells to cardiomyocytes, in 19 human cell lines. Utilizing this data, we were able to identify hundreds of dynamic eQTLs that change over time, with enrichment in enhancers of relevant cell types. We found nonlinear dynamic eQTLs, which can affect only intermediate stages of differentiation, and cannot be found by using data from mature tissues. These fleeting genetic associations with gene regulation may represent a new mechanism to explain complex traits and disease.

The first two projects attempt to characterize how the genetic regulation of gene expression changes in different contexts: tissue type and stage of cellular differentiation, respectively. However, both aims require a priori knowledge of which context to test for interaction with genetic regulation of gene expression. We address this issue in the third project through the development of SURGE, a novel probabilistic model that uses matrix factorization to jointly learn a continuous representation of the cellular contexts defining each measurement, and the corresponding eQTL effect sizes specific to each learned context, allowing for discovery of context-specific eQTLs without pre-specifying subsets of cells or samples. In a proof of concept using bulk expression data over 49 tissues from the GTEx project, SURGE automatically learns factors capturing tissue and cell type composition differences, in addition to one factor reflecting individual ancestry. We applied SURGE to a single-cell eQTL data set consisting of multiplexed single-cell RNA-sequencing data from over 750,000 peripheral blood mononuclear cells from 119 individuals. SURGE automatically identifies cell-type specific eQTLs from this data, identifying factors capturing continuous representations of distinct blood cell types and grouping biologically related cell types into the same factor.

The fourth projects on the previously discovered concept that aberrant gene expression can be used to identify functional, large-effect rare variants. In this project, we expanded detection of genetically driven transcriptome abnormalities by analyzing gene expression, allele-specific expression, and alternative splicing from multitissue RNA-sequencing data, and demonstrate that each signal informs unique classes of RVs. We developed Watershed, a probabilistic model that integrates multiple genomic and transcriptomic signals to predict variant function, validated these predictions in additional cohorts.

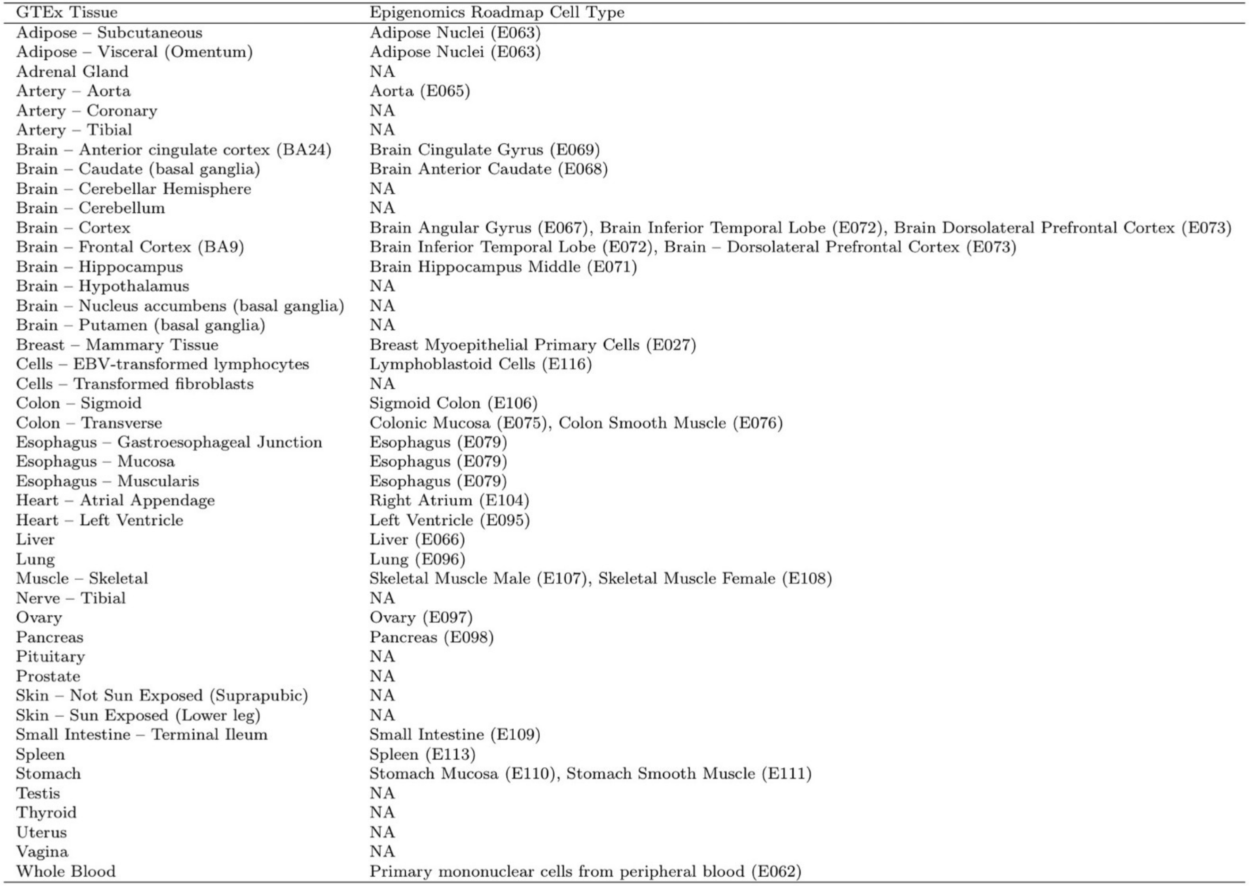
The results discussed in this dissertation prompt a number of interesting follow-up questions. Most notably, it is clear that genetic regulation of gene expression changes as a function of cellular context. However, it is unclear how relevant context-specificity of genetic regulation of gene expression is to the genetic architecture of complex traits and disease. An important next step is a rigorous evaluation of whether a substantial fraction of disease-associated loci are explained with context-specific eQTLs that could not be explained by context-agnostic eQTLs.

Another interesting direction of future research is the application of the Watershed rare variant prioritization framework to rare disease patients to attempt to identify rare variants underlying their disease. Despite the growing prevalence of sequencing technologies in the study of rare disease, the identification of the functional rare variants underlying rare disease cases still remains challenging, particularly for non-coding or regulatory variants. Current approaches based on DNA-sequencing (either whole-genome or whole-exome) alone yield a diagnostic rate of approximately 50%. It would be very exciting to see if Watershed could be used to better that diagnostic rate.

Appendix

Chapter A

Supplementary Tables



GTEx Tissue	Epigenomics Roadmap Cell Type
Adipose – Subcutaneous	Adipose Nuclei (E063)
Adipose – Visceral (Omentum)	Adipose Nuclei (E063)
Adrenal Gland	NA
Artery – Aorta	Aorta (E065)
Artery – Coronary	NA
Artery – Tibial	NA
Brain – Anterior cingulate cortex (BA24)	Brain Cingulate Gyrus (E069)
Brain – Caudate (basal ganglia)	Brain Anterior Caudate (E068)
Brain – Cerebellar Hemisphere	NA
Brain – Cerebellum	NA
Brain – Cortex	Brain Angular Gyrus (E067), Brain Inferior Temporal Lobe (E072), Brain Dorsolateral Prefrontal Cortex (E073)
Brain – Frontal Cortex (BA9)	Brain Inferior Temporal Lobe (E072), Brain – Dorsolateral Prefrontal Cortex (E073)
Brain – Hippocampus	Brain Hippocampus Middle (E071)
Brain – Hypothalamus	NA
Brain – Nucleus accumbens (basal ganglia)	NA
Brain – Putamen (basal ganglia)	NA
Breast – Mammary Tissue	Breast Myoepithelial Primary Cells (E027)
Cells – EBV-transformed lymphocytes	Lymphoblastoid Cells (E116)
Cells – Transformed fibroblasts	NA
Colon – Sigmoid	Sigmoid Colon (E106)
Colon – Transverse	Colonic Mucosa (E075), Colon Smooth Muscle (E076)
Esophagus – Gastroesophageal Junction	Esophagus (E079)
Esophagus – Mucosa	Esophagus (E079)
Esophagus – Muscularis	Esophagus (E079)
Heart – Atrial Appendage	Right Atrium (E104)
Heart – Left Ventricle	Left Ventricle (E095)
Liver	Liver (E066)
Lung	Lung (E096)
Muscle – Skeletal	Skeletal Muscle Male (E107), Skeletal Muscle Female (E108)
Nerve – Tibial	NA
Ovary	Ovary (E097)
Pancreas	Pancreas (E098)
Pituitary	NA
Prostate	NA
Skin – Not Sun Exposed (Suprapubic)	NA
Skin – Sun Exposed (Lower leg)	NA
Small Intestine – Terminal Ileum	Small Intestine (E109)
Spleen	Spleen (E113)
Stomach	Stomach Mucosa (E110), Stomach Smooth Muscle (E111)
Testis	NA
Thyroid	NA
Uterus	NA
Vagina	NA
Whole Blood	Primary mononuclear cells from peripheral blood (E062)

Supplementary table 1: Mapping from GTEx tissue type to Roadmap cell type.

Supplementary Figures

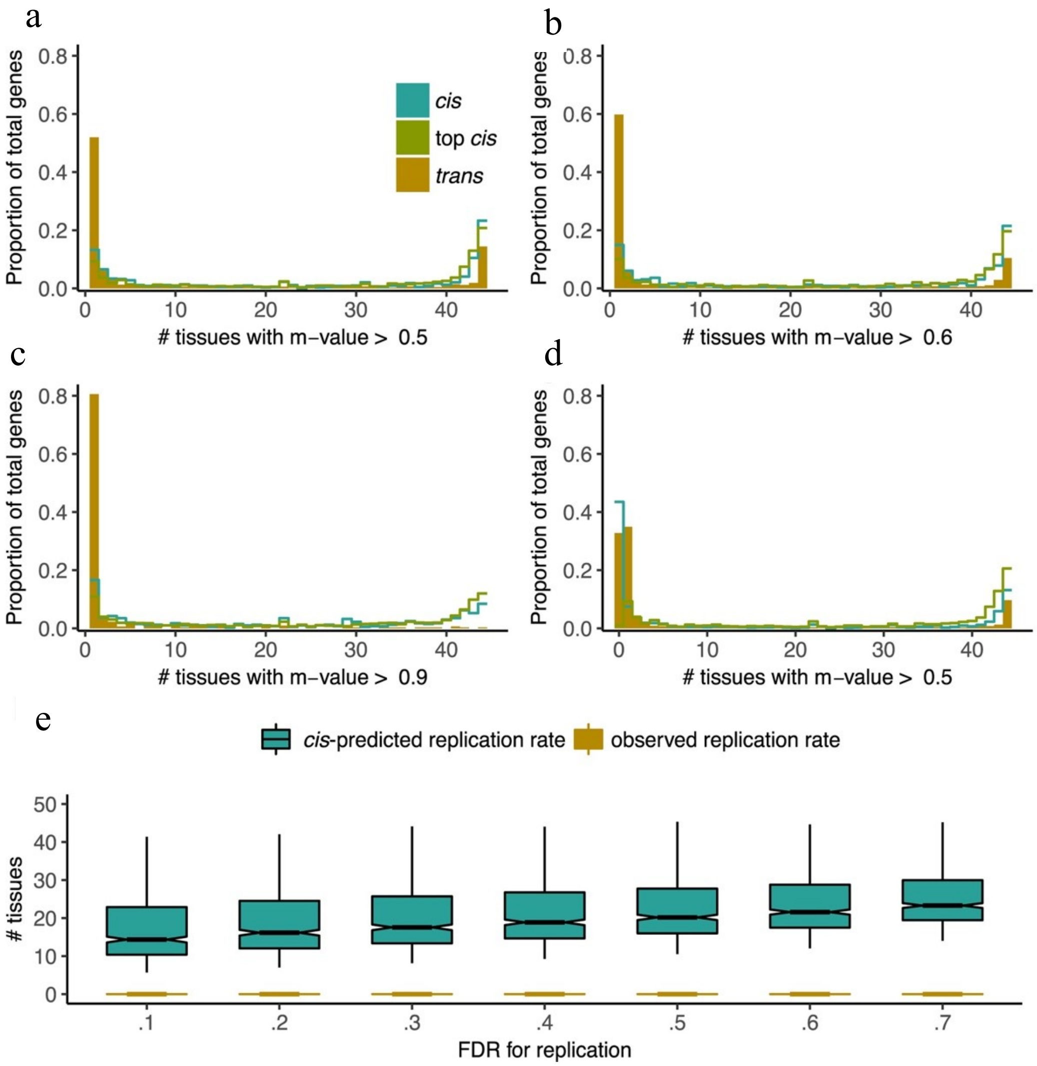


Figure S1: ***a*–*d****, Meta-analysis performed using Meta-Tissue for trans-eGenes (50% FDR), randomly selected cis-eGenes (50 % FDR), and an equal number of the top cis-eGenes by P value. Distribution of the number of tissues that have Meta-Tissue m values greater than a given threshold (****a****, 0.5;* ***b****, 0.6;* ***c****, 0.9) across variant–gene pairs that have an effect (based on m value thresholding) in at least one tissue.* ***d****, The same distribution as* ***a*** *except that variant–gene pairs with predicted effect in zero tissues (based on the number of m values > 0.5) are included. Meta-Tissue predicts that many cis-eGenes (50% FDR) and trans-eGenes (50% FDR) will have an effect in zero tissues. The number of zero predictions is largely reduced for the top most significant cis-eGenes.* ***e****, Distribution of observed replication rate between pairs of tissues for trans-eQTLs (10% FDR) versus the predicted replication rate for trans-eQTLs (10% FDR) based on a negative binomial generalized linear model trained on cis-eQTLs (10% FDR0.1). This model directly accounts for effect size and minor allele frequency. Replication rates shown for a range of FDR thresholds in replication tissue. Box plots depict the IQR, whiskers depict 1.5× IQR.*

Chapter B

Supplementary Tables

Cell Line	Percent of Live Cells Expressing TNNT2
18489	44.3
18499	24.2
18505	NA
18508	83.9
18511	NA
18517	47.8
18520	NA
18855	NA
18858	NA
18870	NA
18907	7.9
18912	47.8
19093	27
19108	NA
19127	1.1
19159	39.8
19190	63.2
19193	59.5
19209	33.4

Table S1: Flow cytometry results for each cell line at day 15 of cardiomyocyte differentiation. The percent of live cells expressing cardiac troponin (TNNT2) for every cell line at day 15 of differentiation. Cells with an NA indicate that flow cytometry was not performed on this cell line.

Hallmark gene set	Gene cluster 2	Gene cluster 4	Gene cluster 5	Gene cluster 6	Gene cluster 9	Gene cluster 11	Gene cluster 13	Gene cluster 16
TNFA signaling via NFKB	1	1	1	.000208	1	1	1	1
Mitotic spindle	1	1	1	1	.0166	1	1.80e-14	1
TGF beta signaling	1	1	1	.348	.000624	1	1	1
DNA repair	1	1	.000242	1	1	1	3.73e-7	1
G2M checkpoint	1	1	1	1	1	1	2.87e-63	.594
Myogenesis	9.29e-14	1	1	1.05e-5	1	1	1	1
Protein secretion	.00384	1	1	1	1	1	1	1
Complement	1	1.98e-5	1	1	1	1	1	1
Unfolded protein response	1	1	6.99e-5	1	1	1	1	1
MTORC1 signaling	1	1	2.07e-10	1	1	.696	1	1
E2F targets	1	1	.0111	1	1	1	5.47e-73	.0458
MYC targets V1	1	1	3.03e-25	1	1	.329	1.28e-16	1.16e-5
MYC targets V2	1	1	7.04e-21	1	1	1	.981	1
Epithelial mesenchymal transition	1	.000310	1	2.05e-5	1	1	1	1
Xenobiotic metabolism	1	.000435	1	1	1	1	1	1
Oxidative phosphorylation	1	1	1	.134	1	8.11e-11	1	1
Heme metabolism	1.24e-6	1	1	1	1	1	1	1
Coagulation	1	1.72e-16	1	1	1	1	1	1
Bile acid metabolism	1	.00392	1	1	1	1	1	1
Spermatogenesis	1	1	1	1	1	1	.00433	1
KRAS signaling up	1	.00536	1	.622	1	1	1	1

Table S2: Hallmark gene set enrichment of split-GPM gene clusters: Bonferroni corrected p-values (Fisher's exact) from gene set enrichment of gene clusters (columns) from split-GPM within Hallmark gene sets (rows). Only gene clusters and gene sets with at least one significant enrichment (Bonferroni p-value <= .05) are shown.

# of cell line collapsed PCs	**# genes with significant dynamic eQTL (eFDR <= .05)**	**# genes with significant dynamic eQTL (eFDR <= .01)**
0	2256	931
1	1943	785
2	1247	294
3	648	250
4	608	186
5	550	150
6	533	113
7	556	212
8	456	110
9	288	22
10	213	79

Table S3: Number of linear dynamic eQTLs detected. The number of genes with a significant linear dynamic eQTL (eFDR <= .05 and eFDR <= .01) as a function of the number cell line collapsed PCs used as covariates.

Supplementary Figures

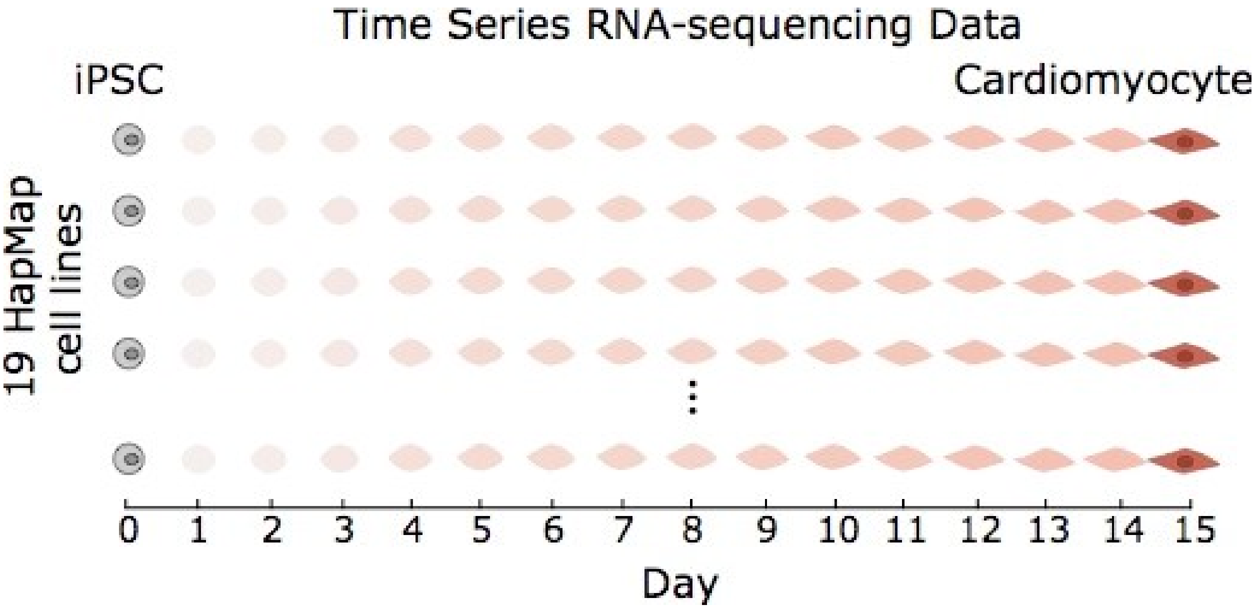


Figure S1: RNA-seq sample collection: Overview of RNA-seq sample collection. In 19 Yoruba HapMap cell lines, RNA was extracted and sequenced every 24 hours at 16 time points, generating 297 RNA-seq samples.

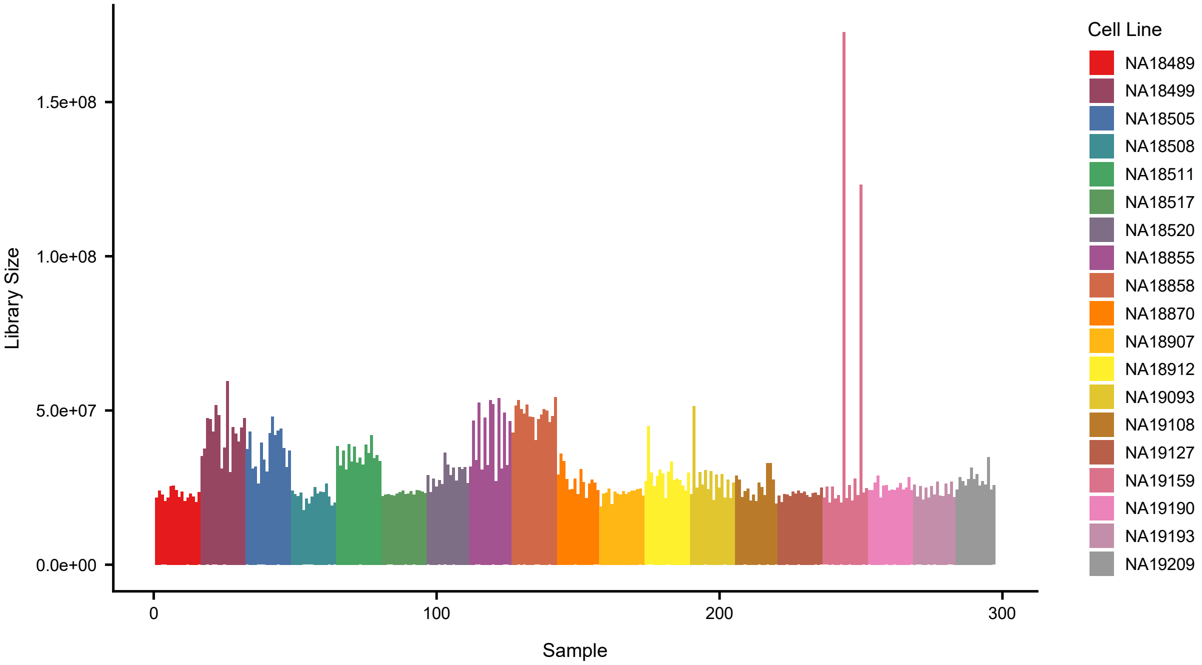


Figure S2: Library size of RNA-seq samples. The library sizes of 297 RNA-seq samples colored by their cell line identity. Within each cell line, samples are ordered along the x-axis by their differentiation time point from day 0 to 15.

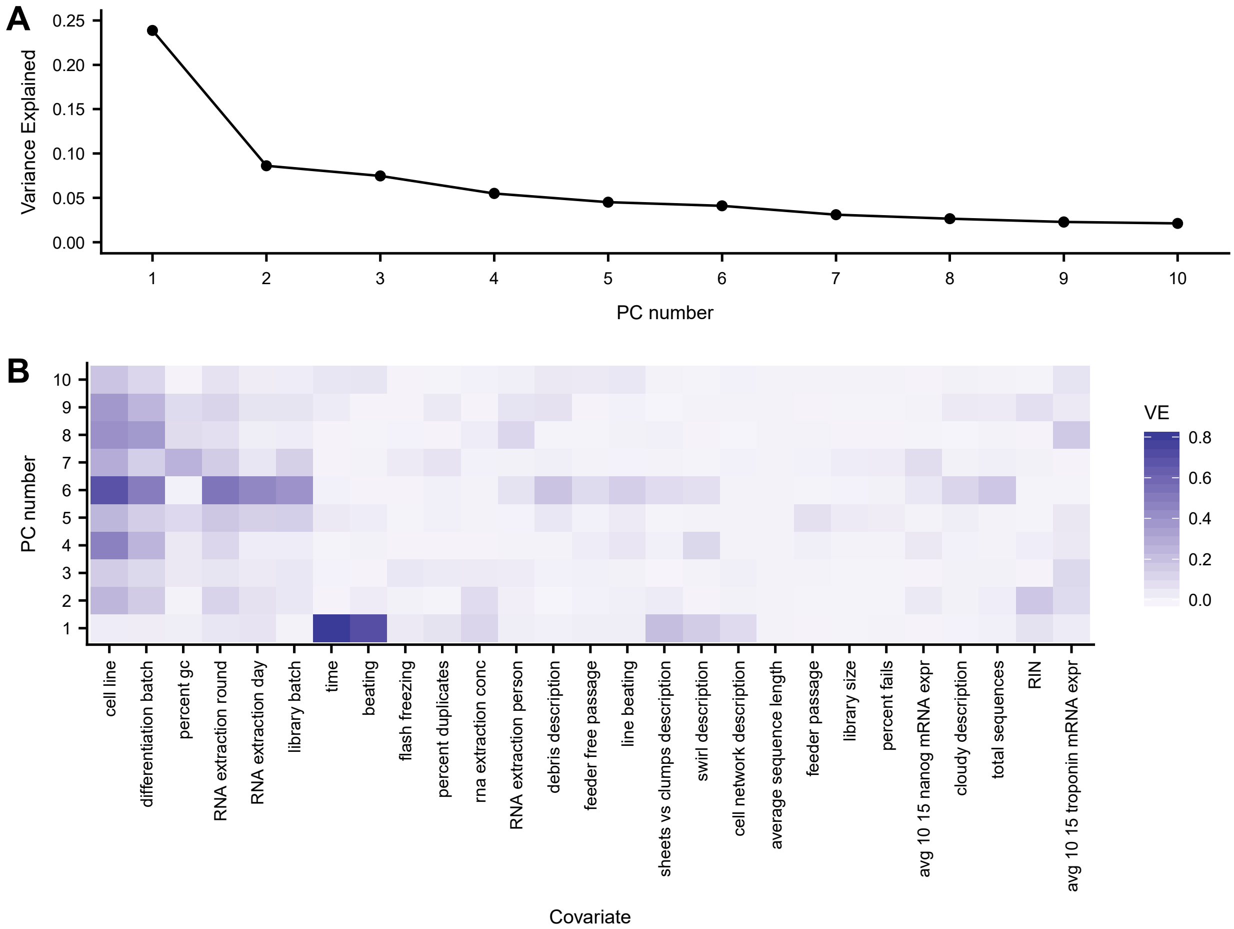


Figure S3: Explaining principal components with sample covariates. (A) Variance in gene expression explained by first 10 gene expression principal components. (B) Variance explained of each gene expression principal component using sample covariates. Adjusted R^2 was used to handle categorical sample covariates. Detailed explanation of each sample covariate can be found in Table S1.

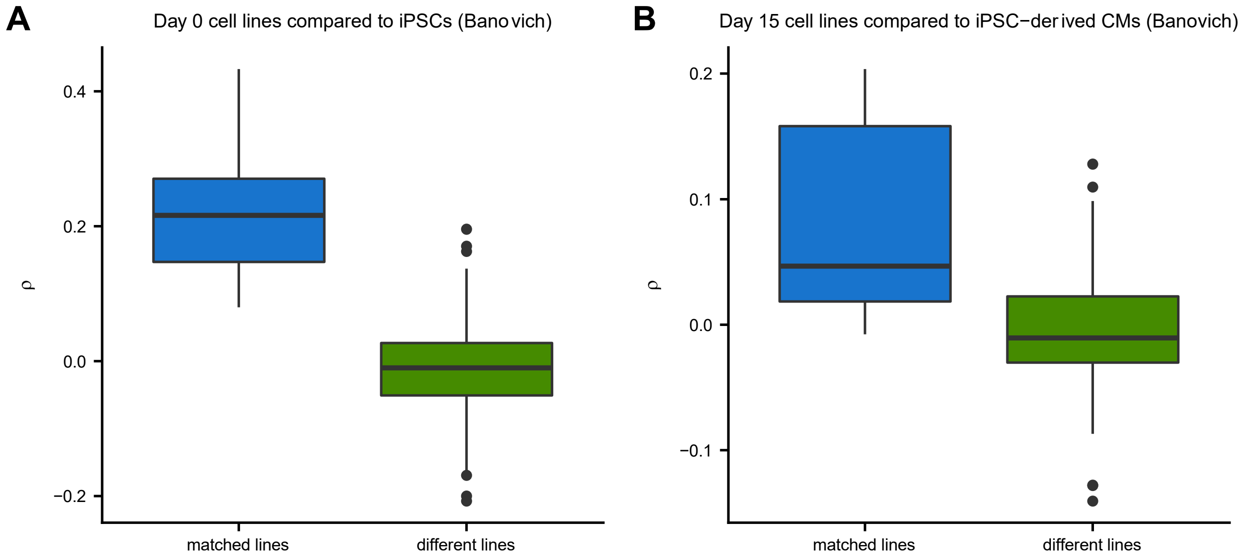


Figure S4: Biological replication of day 0 and day 15 cells. We compared day 0 and day 15 cell lines with matched iPSC lines and iPSC-derived cardiomyocyte lines, respectively, from Banovich et al. (9). This analysis was restricted to cell lines present in both data sets. Spearman correlation across genes observed in both data sets between (A) day 0 cell lines and iPSC lines and between (B) day 15 cell lines and iPSC-derived cardiomyocyte cell lines. Distribution of spearman correlations shown for matched cell lines (blue) and different cell lines (green). The correlation of gene expression is greater for matched cell lines compared to different cell lines (p < .05 for both comparisons, Wilcoxon rank-sum test).

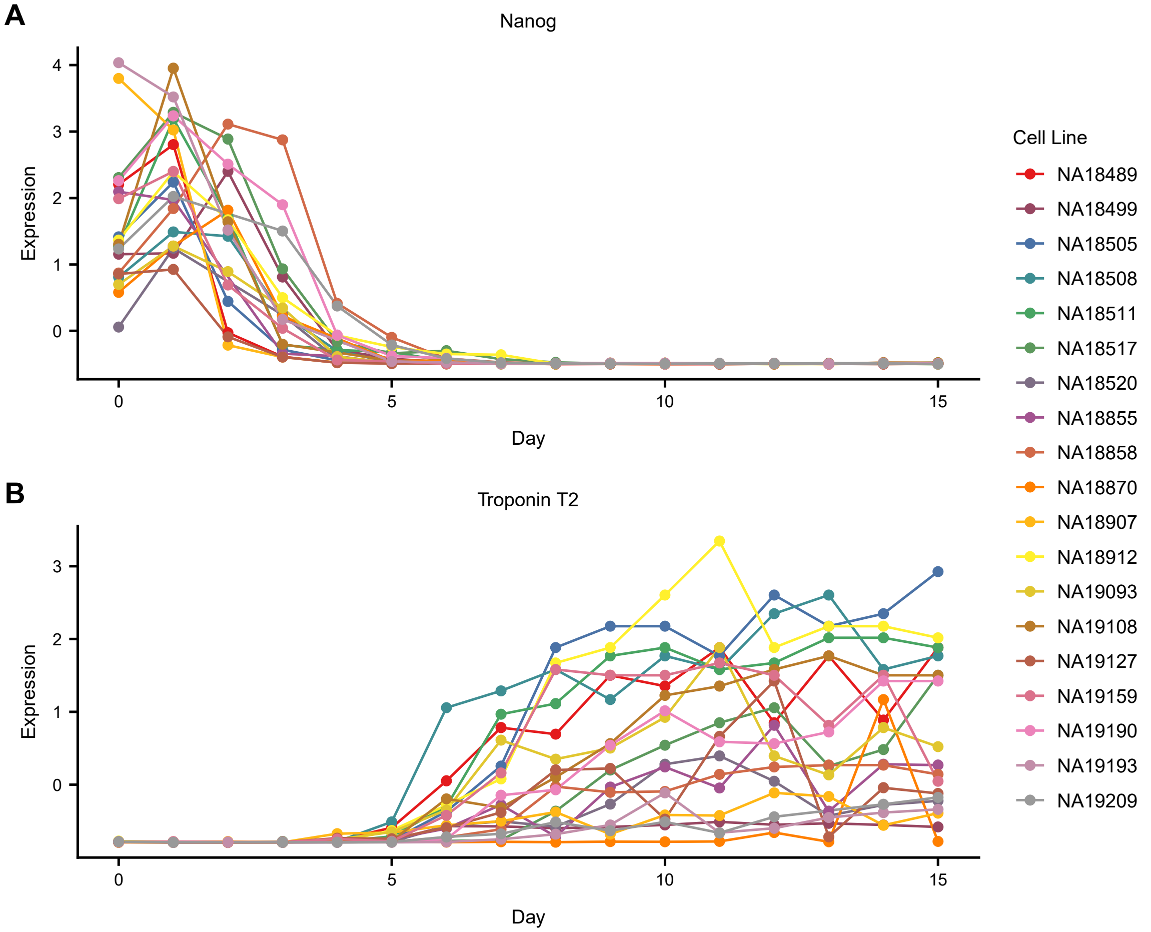


Figure S5: Expression time course of known cell type specific marker genes. Standardized gene expression levels of Nanog (A, stem cell marker gene) and Troponin T2 (B, cardiomyocyte marker gene) across 16 time points (x-axis) and 19 cell lines (colors).

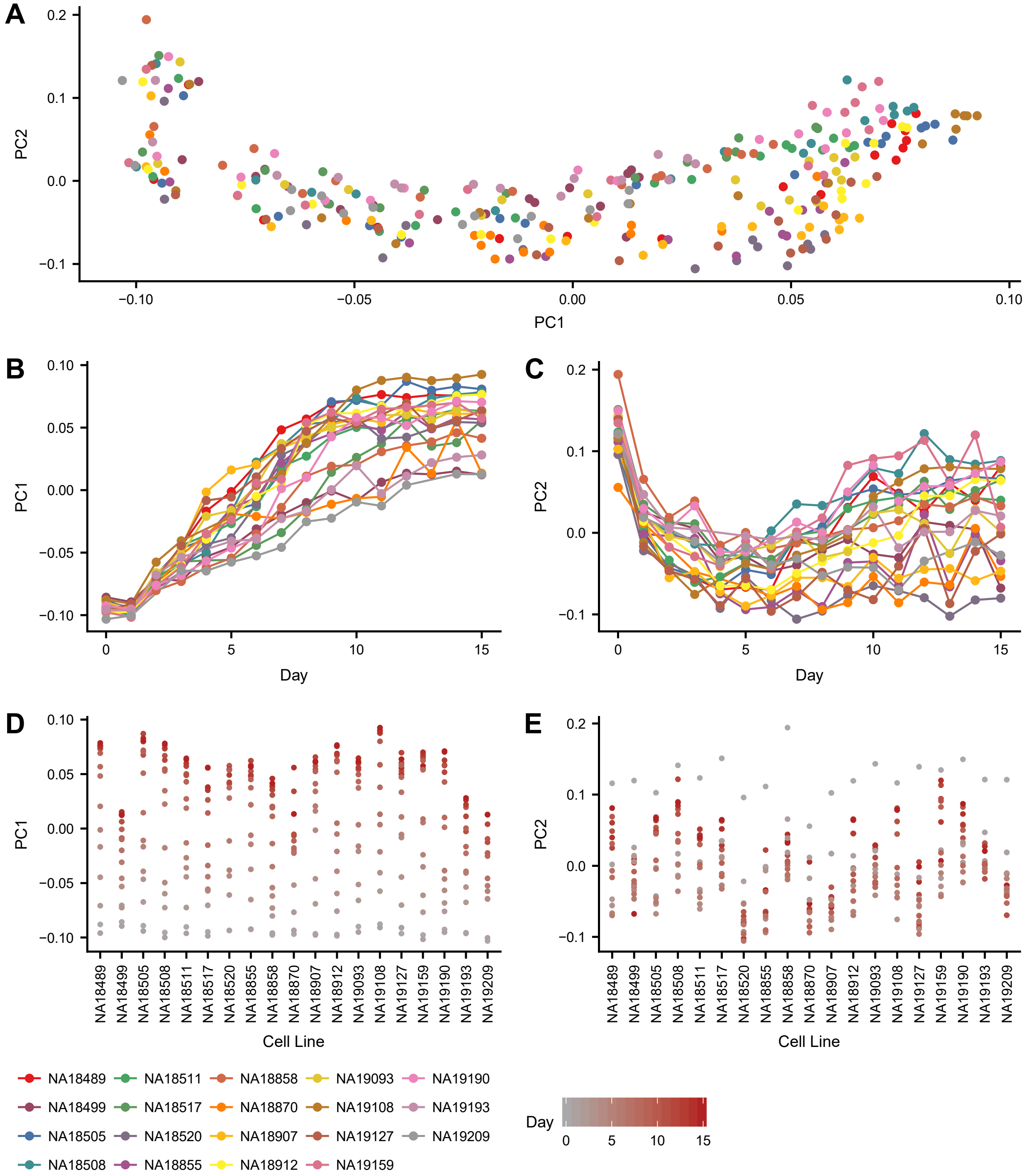


Figure S6: Principal component analysis separated by cell line identity. (A) First two gene expression principal component loadings for all 297 RNA-seq samples, where each sample is colored by its cell line identity. (B, C) Principal component 1 and 2 loadings across 16 time points (x-axis) and 19 cell lines (colors). (D, E) Principal component 1 and 2 loadings across 19 cell lines (x-axis) and 16 time points (colors).

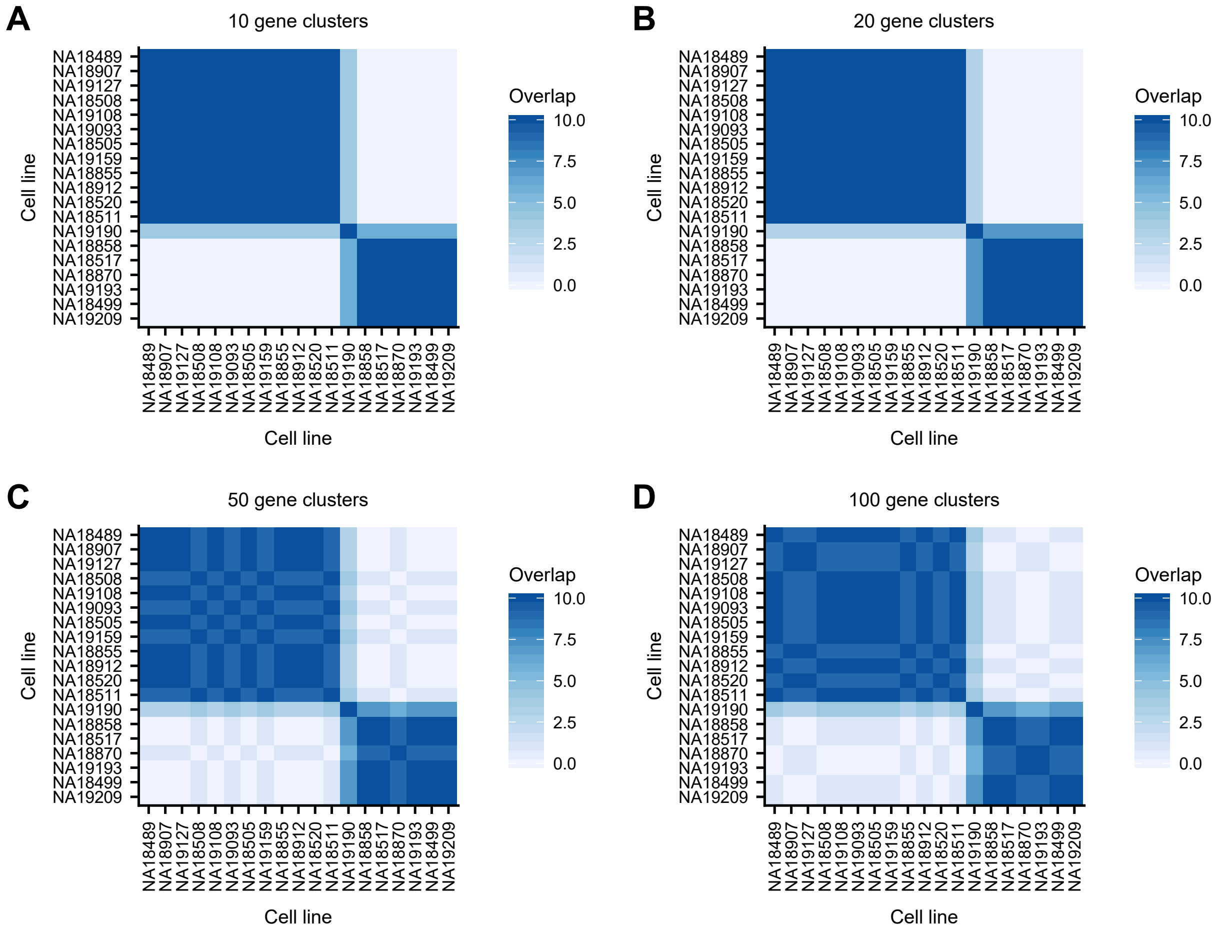


Figure S7: split-GPM cell line cluster assignment robust to hyper-parameter choice. Number of times (out of 10 split-GPM runs with independent, random initializations) that each cell line pair was assigned to the same cell line cluster when 10 (A), 20 (B), 50 (C), and 100 (D) gene clusters were used. Cell lines are ordered by their cell line collapsed PC1 loadings.

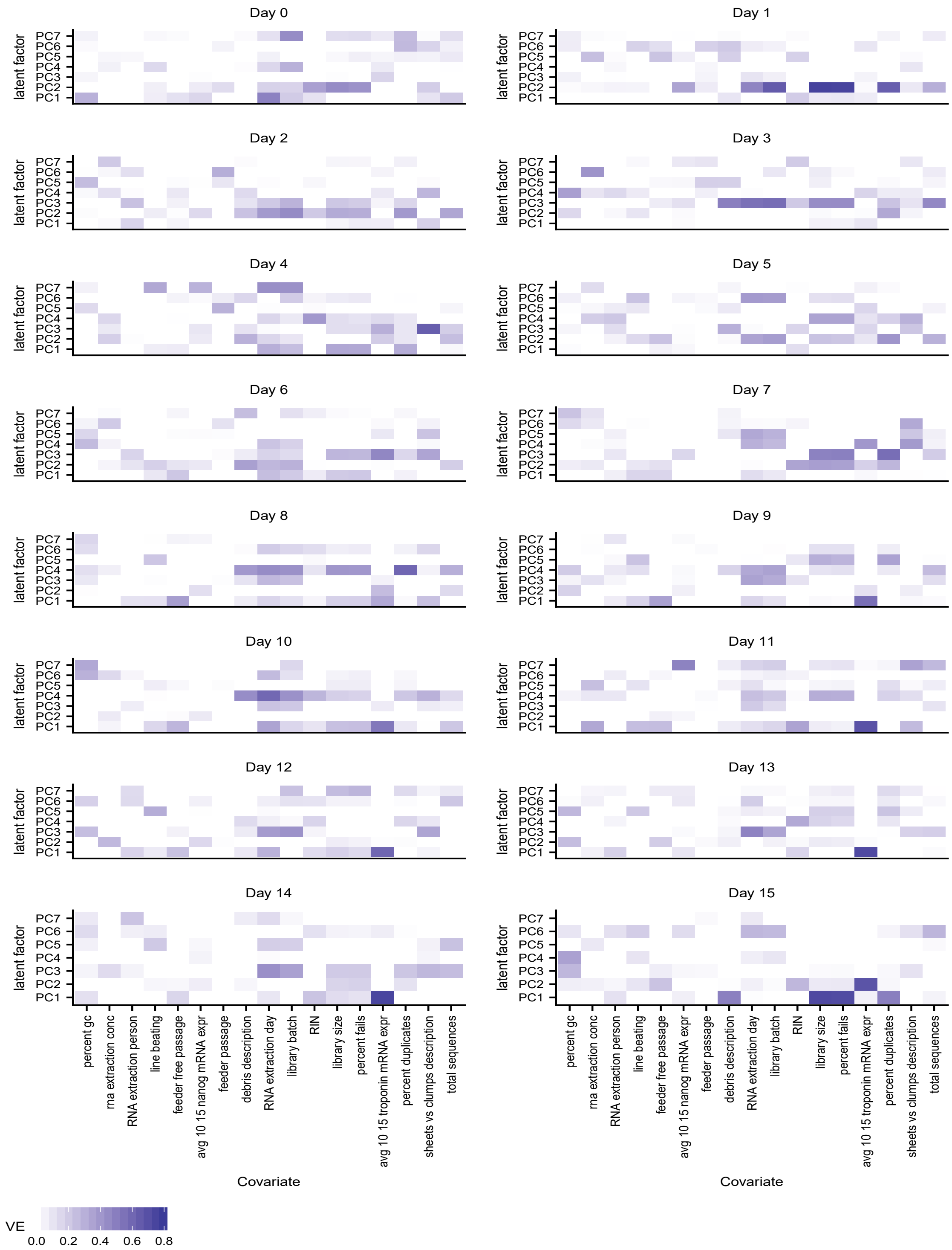


Figure S8: Explaining time step principal components with sample covariates. In each time point independently, variance explained of each raw read count expression principal components (from samples belonging to the corresponding time point) using sample covariates. Adjusted R² was used to handle categorical sample covariates. Sample categorical covariates with more than 8 categories were excluded from this analysis due to the small sample size when considering time points, independently. Detailed explanation of each sample covariate can be found in Table S1.

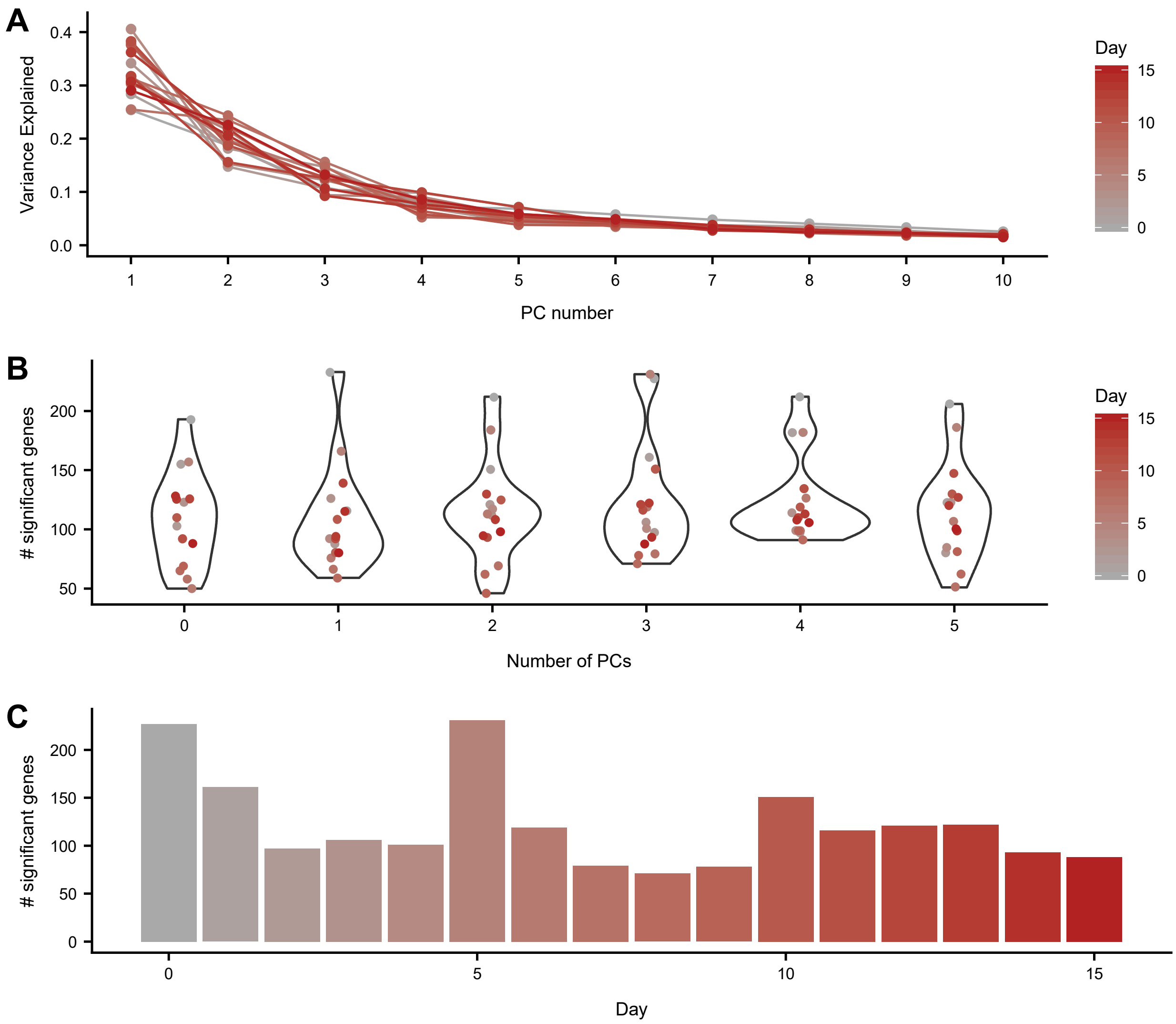


Figure S9: Number of genes with non-dynamic eQTLs. (A) Variance explained of gene expression from samples belonging to a particular time point (color) by the first 10 gene expression PCs (x-axis) computed on samples belonging to that time point. (B) The number of genes with a significant eQTL (eFDR <= .05) in each time point (color) as a function of number of expression PCs controlled for (x-axis). (C) The number of genes with a significant eQTL (eFDR <= .05) in each time point when controlling for three expression PCs.

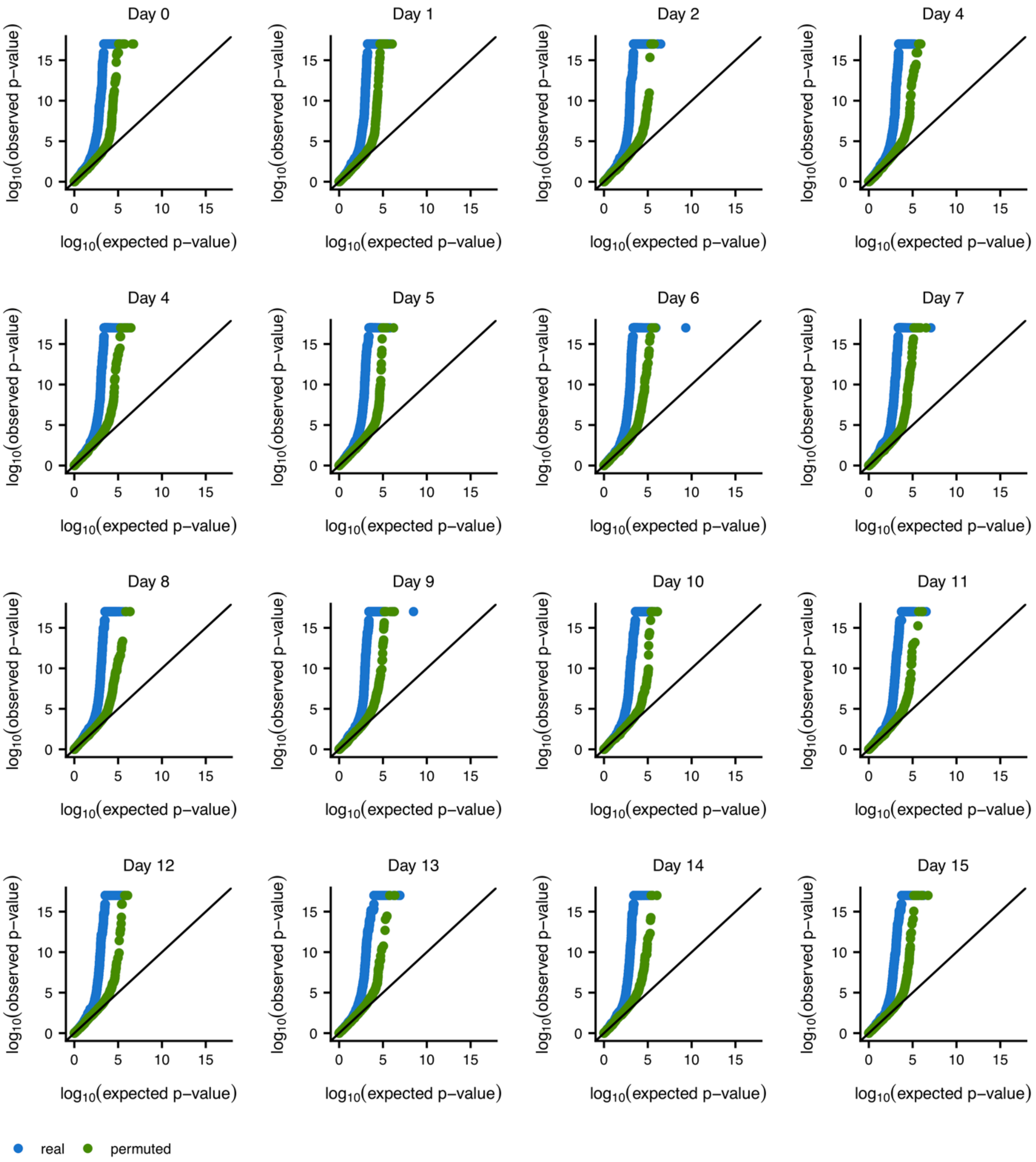


Figure S10: Q-Q plots for non-dynamic eQTLs. Q-Q plot for non-dynamic eQTLs in all 16 time steps. Blue dots correspond to p-values from actual data relative to uniformly distributed p-values, whereas green dots correspond to p-values from permuted data (using WASP's permutation strategy) relative to uniformly distributed p-values.

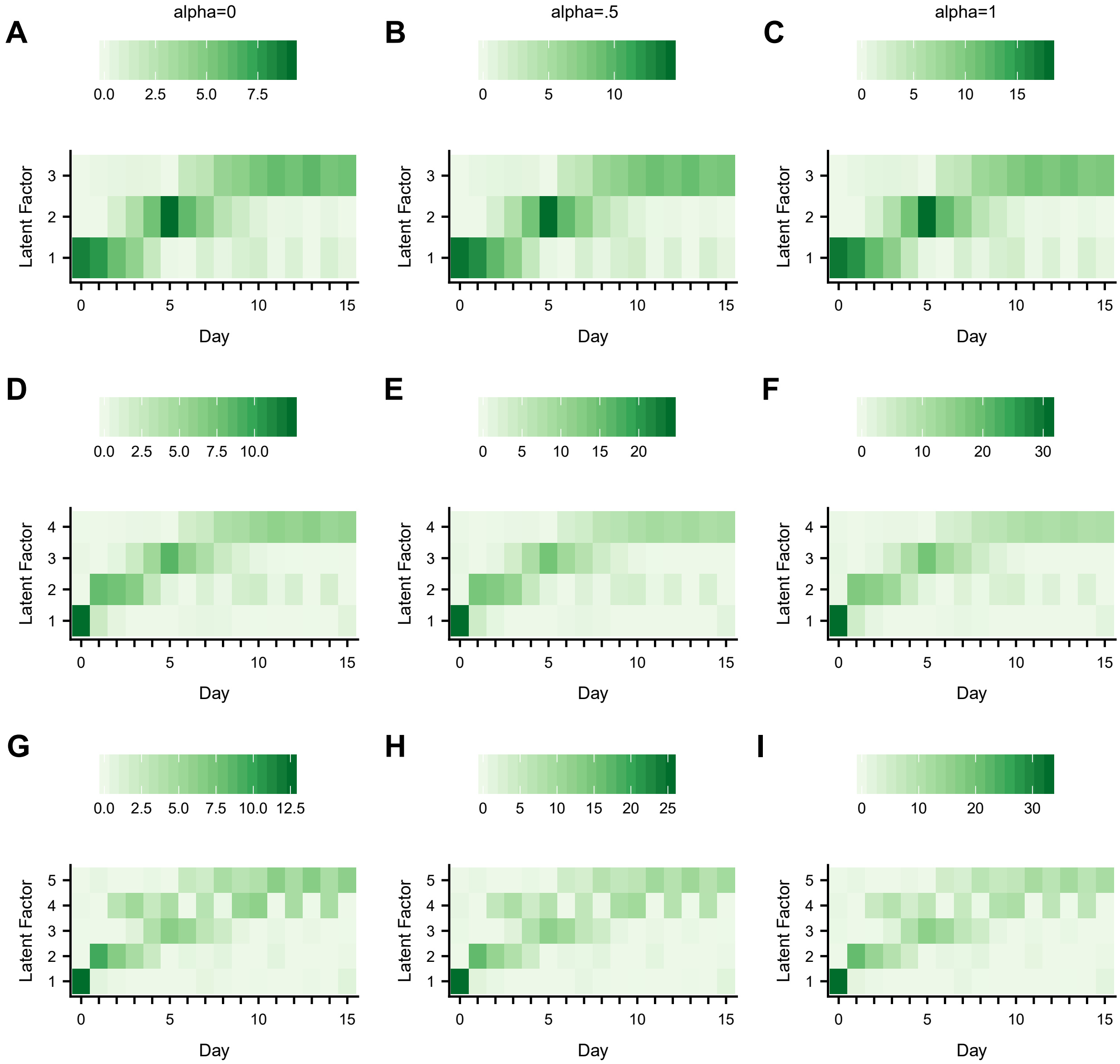


Figure S11: Matrix factorization of eQTL summary statistics. Latent factors identified via sparse non-negative matrix factorization of non-dynamic eQTL -log$_{10}$ p-values shown for a range of sparse prior choices (alpha; columns) when using 3, 4, and 5 factors (rows).

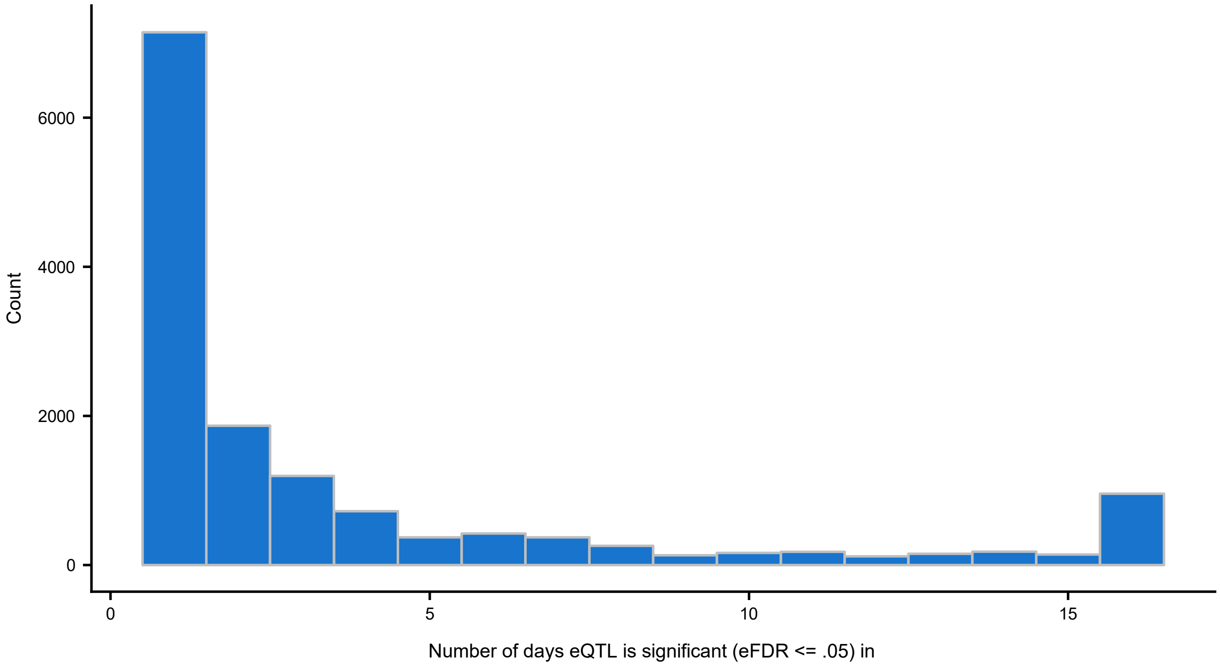


Figure S12: eQTL sharing across time points. The number of days in which each non-dynamic eQTL is significant (eFDR <= .05) for all variant-gene pairs that are significant in at least one day.

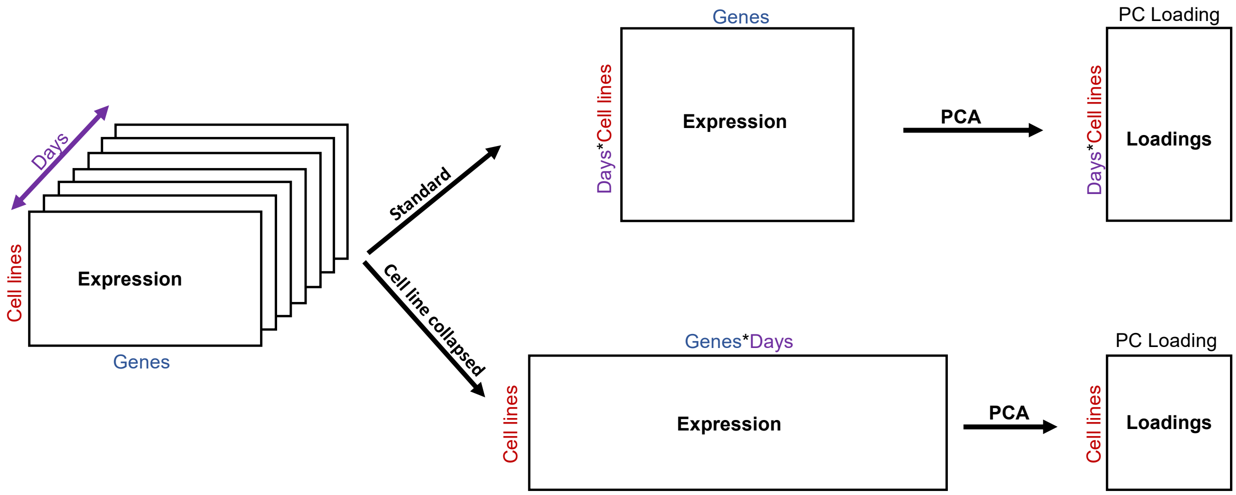


Figure S13: Overview of cell line collapsed PCA. Gene expression can be represented as a three-dimensional matrix spanning days, cell lines, and genes. For standard PCA (top row), we rearrange this gene expression matrix such that rows now correspond to cell lines at specific days (e.g., RNA-seq samples) and columns correspond to genes. Here, PCA will learn a low dimensional representation for cell lines at specific days. For cell line collapsed PCA (bottom row), we rearrange this gene expression matrix such that rows now correspond to cell lines and columns correspond to genes at specific days. Here, PCA will learn a low dimensional representation for each cell line.

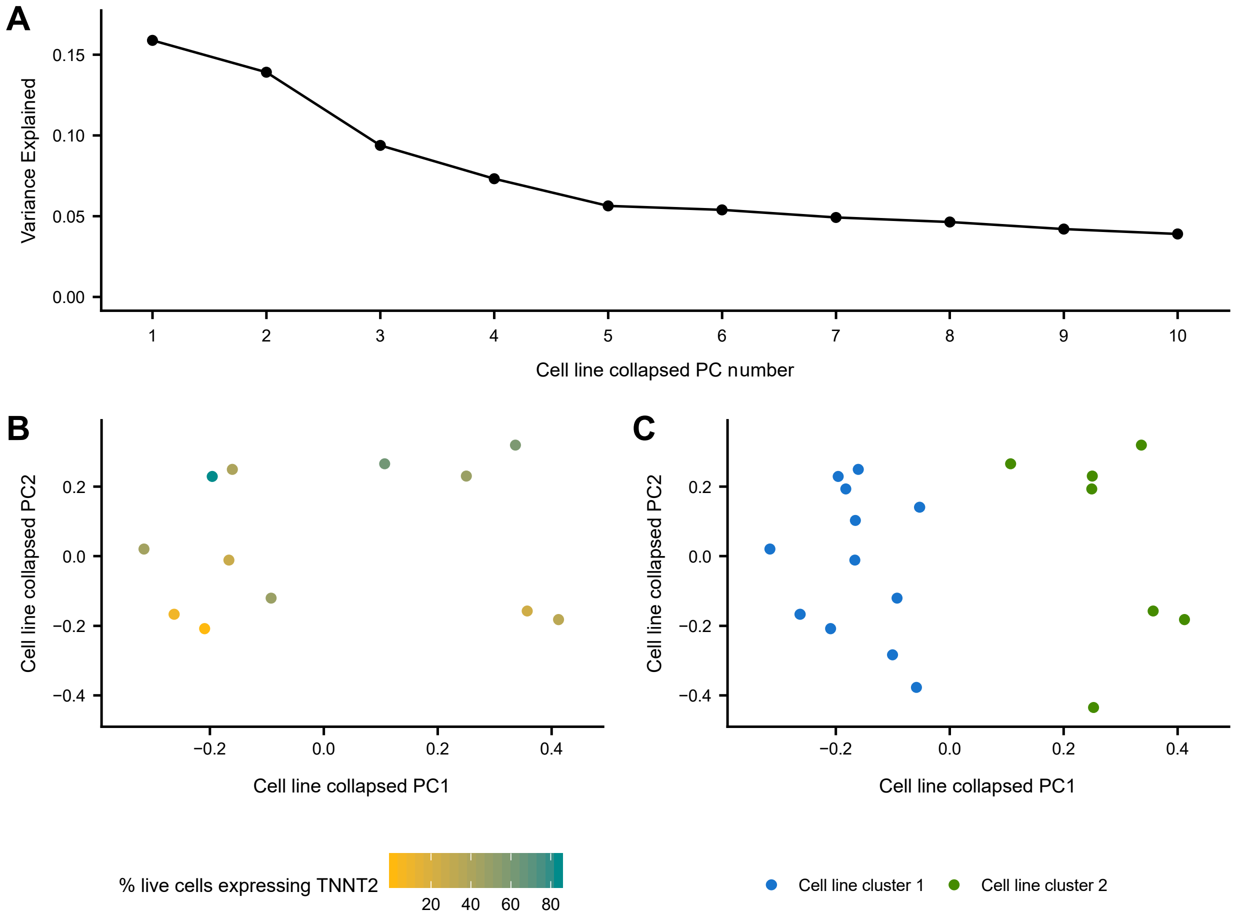


Figure S14: Analysis of cell line collapsed PCs. (A) Variance explained of gene expression by first 10 cell line collapsed principal components. (B, C) First two cell line collapsed principal components where each data point is a cell line colored by its (B) percentage of live cells expressing TNNT2 at time point 15 and (C) split-GPM cell line cluster assignment.

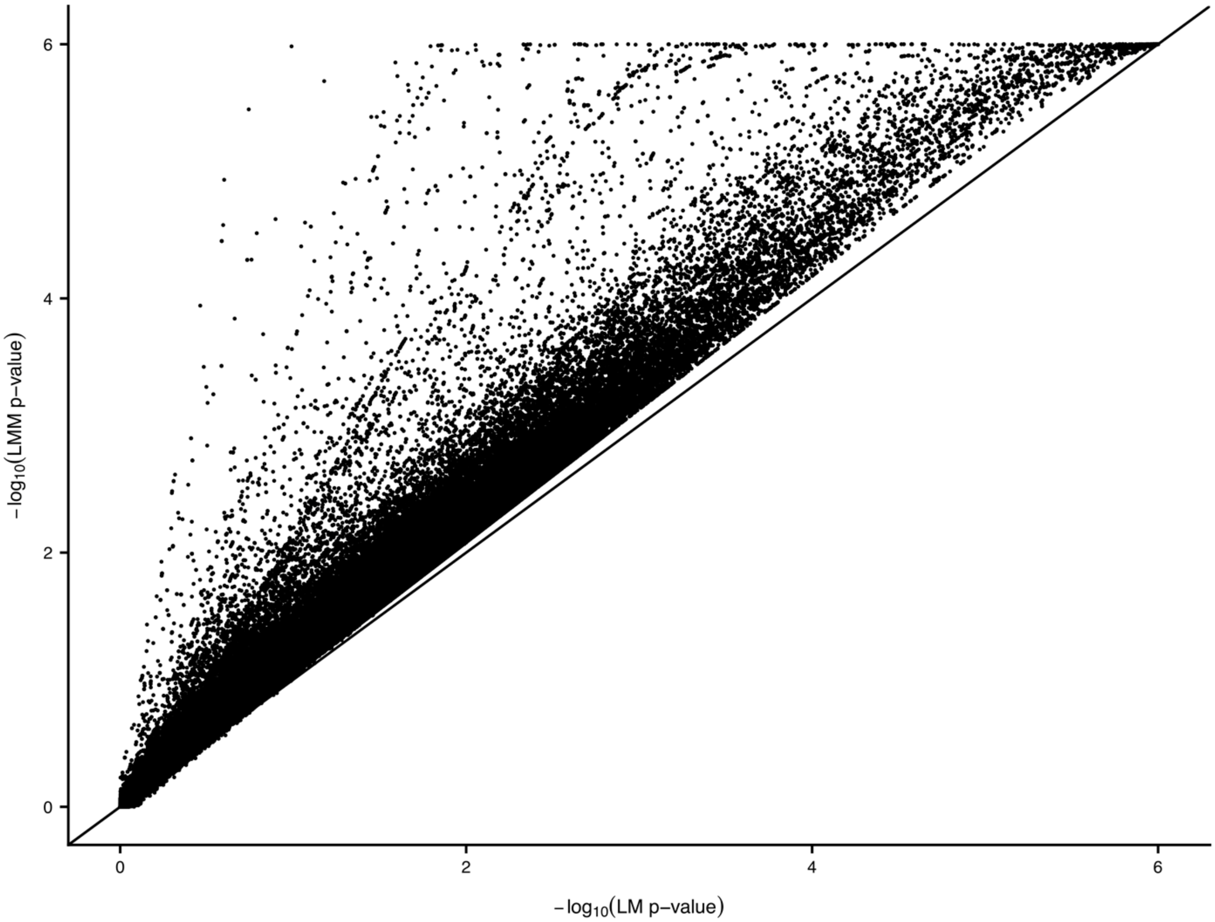


Figure S15: Detecting dynamic eQTLs with gaussian linear mixed model: Comparison of linear dynamic eQTL p-values between gaussian linear model (x-axis) and gaussian linear mixed model with cell line specific random effect (y-axis) across all tested variant-gene pairs (Pearson correlation=.983).

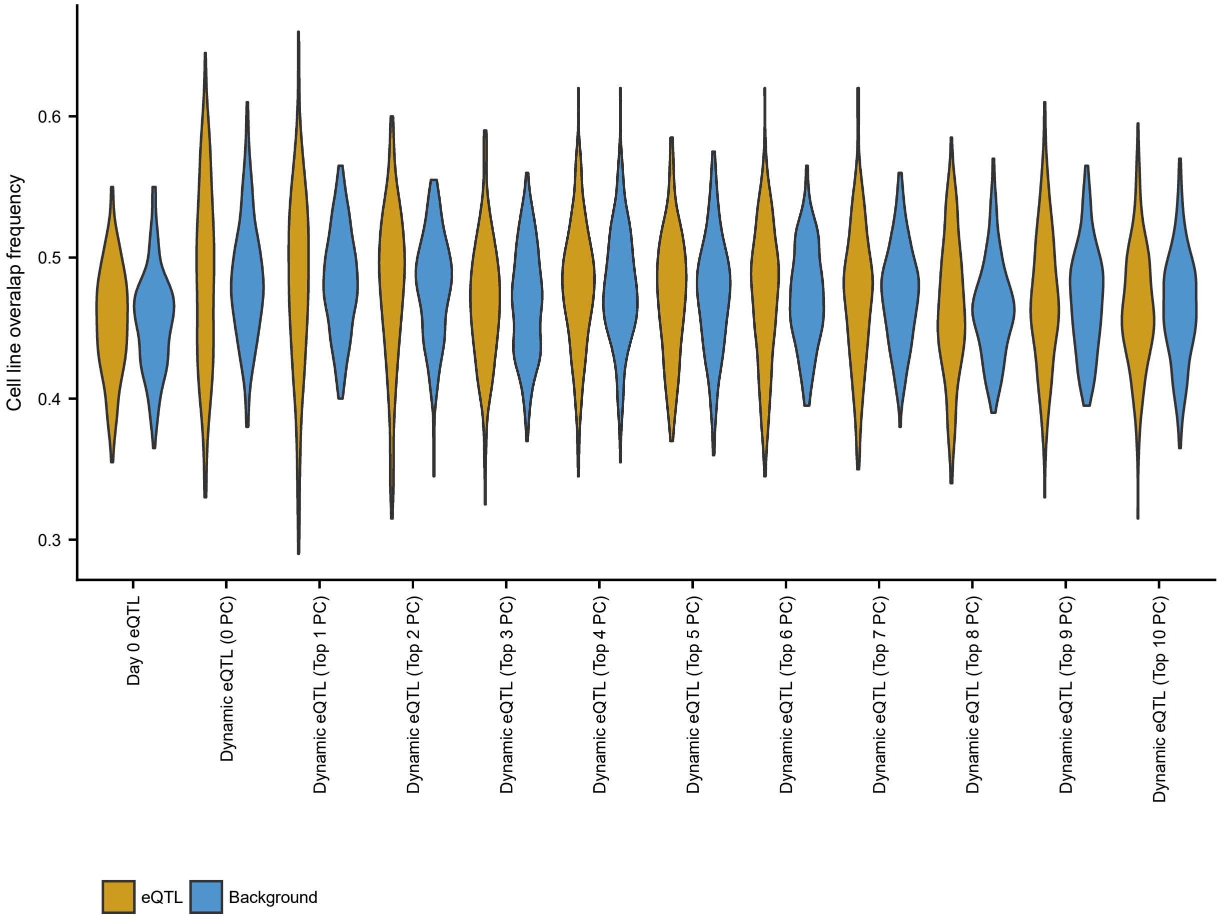


Figure S16: Frequency of cell line overlap in genotype bins. Frequency at which each cell line pair is in the same genotype bin ({0,1,2}) across the strongest associated variants of the 200 most significant eQTL genes (gold) compared to MAF-matched randomly selected background variants (blue). Analysis shown for linear dynamic eQTLs while controlling for a range of the top cell line collapsed PCs. Non-dynamic eQTLs (from day 0) are also shown as a control.

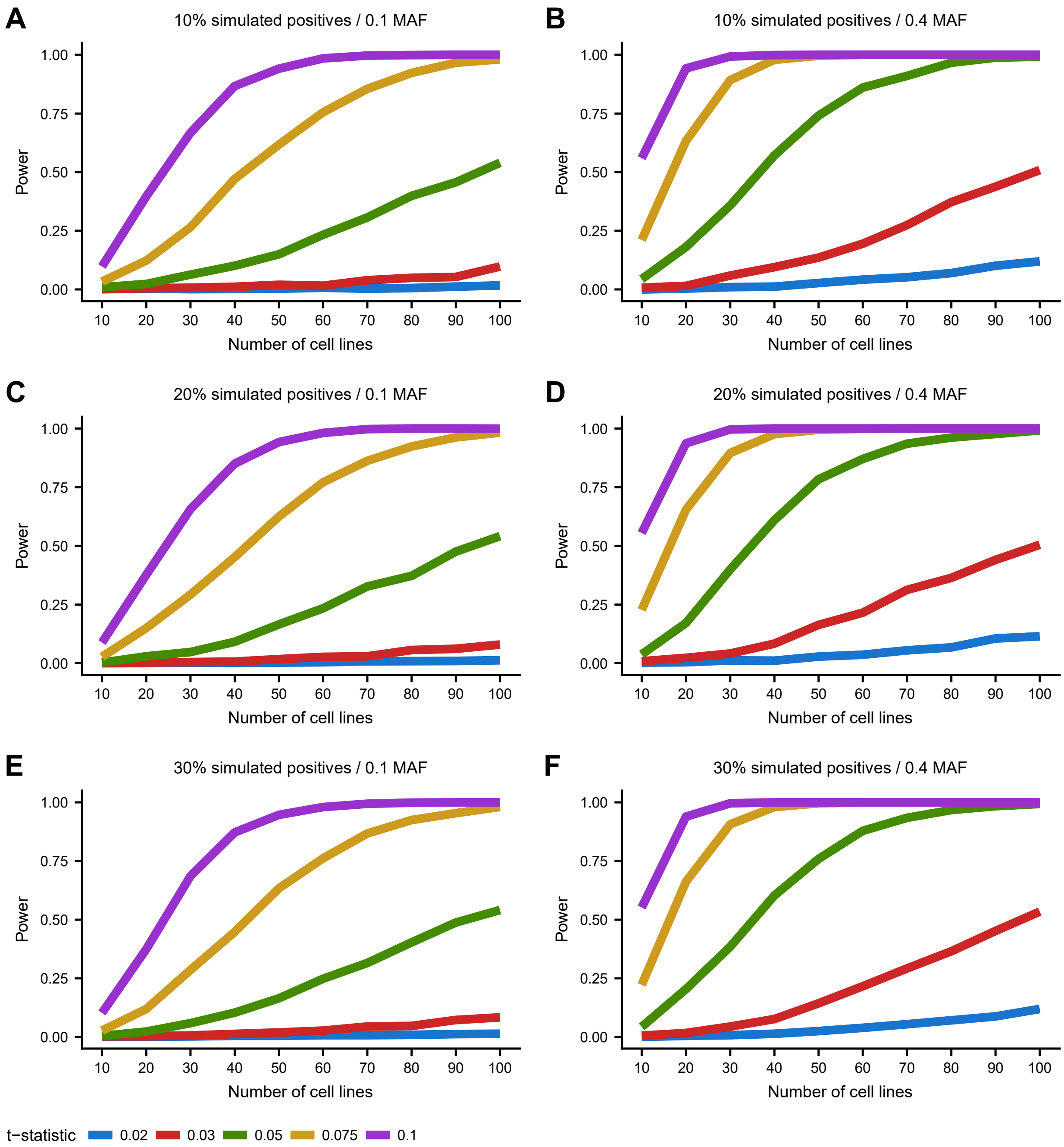


Figure S17: Simulated power analysis for linear dynamic eQTLs. Power to detect simulated linear dynamic eQTLs (y-axis) based on 10,000 simulations at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data) as a function of number of cell lines (x-axis) and t-statistic (color). t-statistic represents the ratio of the effect size of the interaction term and the standard deviation term used to simulate the expression data. We additionally vary (A-F) both the simulated MAF (columns) and the proportion of those tests that were simulated according to the alternative hypothesis (true dynamic eQTLs; rows).

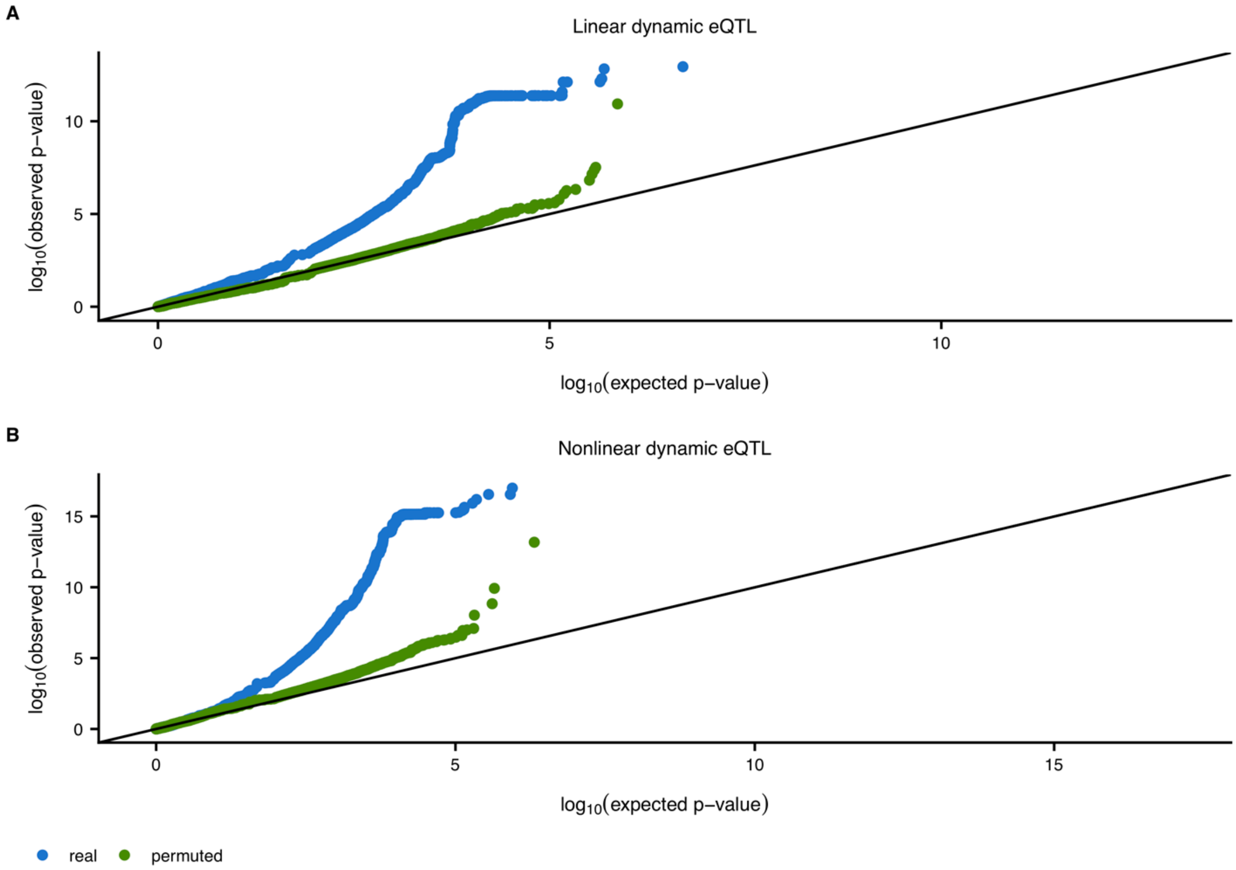


Figure S18: Q-Q plots for linear and non-linear dynamic eQTLs. Q-Q plot for (A) linear and (B) non-linear dynamic eQTLs. Blue dots correspond to p-values from actual data relative to uniformly distributed p-values, whereas green dots correspond to p-values from permuted data relative to uniformly distributed p-values.

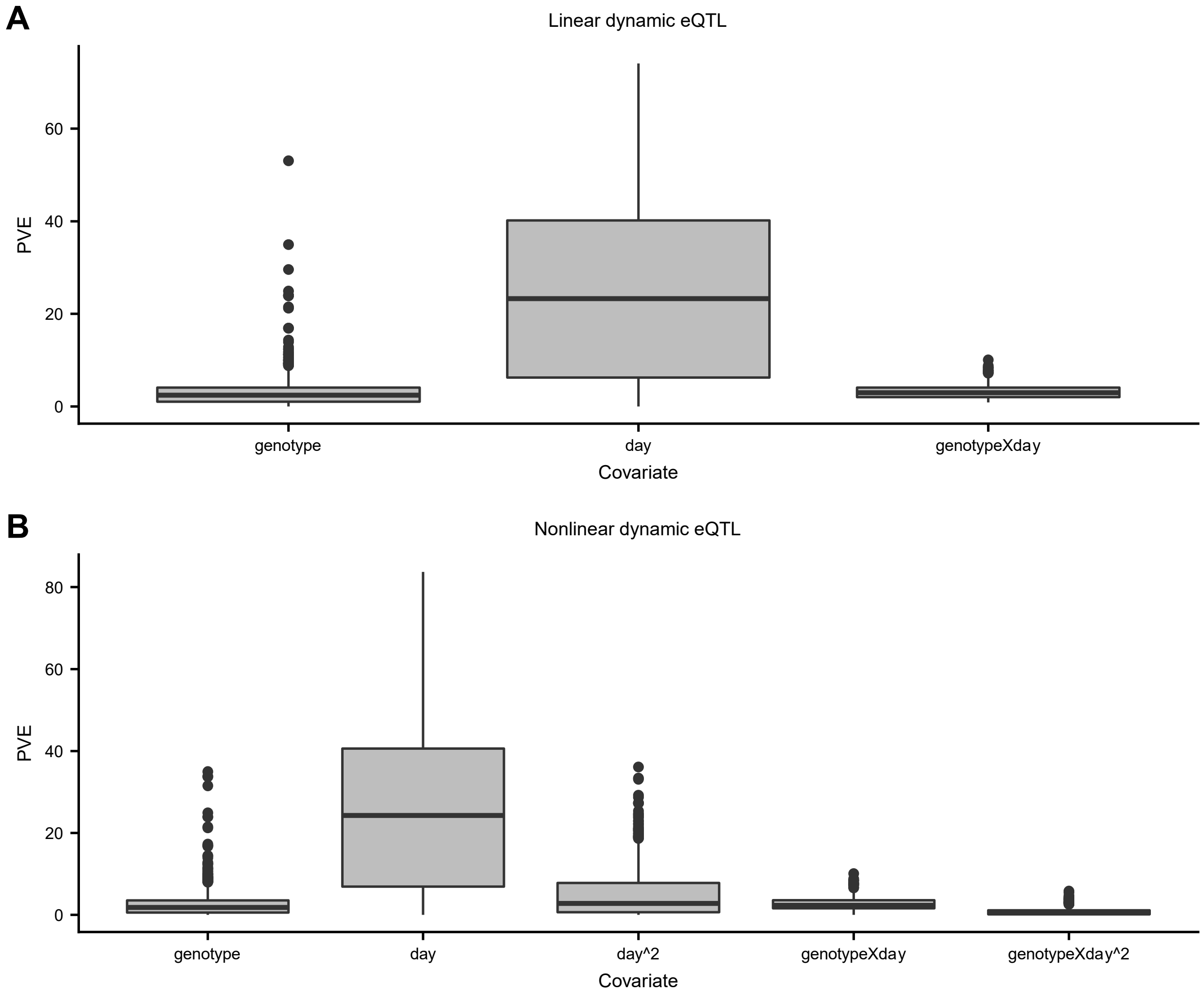


Figure S19: Percent variance explained of dynamic eQTL covariates. Distribution of percent variance explained (PVE; y-axis) of each covariate (x-axis) across significant (eFDR <= .05) (A) linear dynamic eQTLs and (B) nonlinear dynamic eQTLs. For linear dynamic eQTLs, the interaction term (genotypeXday) explains on average 3.16 % of the variance. For nonlinear dynamic eQTLs, the linear interaction term (genotypeXday) and the nonlinear interaction term (genotypeXday^2) explain on average 2.69 and 0.78 % of the variance, respectively. PVE for each covariate was estimated via ANOVA analysis which assumes an underlying order of covariates when iteratively computing the variance explained by each additional covariate. This was done to handle the covariance between covariates. For linear dynamic eQTLs, covariates were ordered as follows: all cell line collapsed PC related terms, genotype, day, and then genotypeXday. For nonlinear dynamic eQTLs, covariates were ordered as follows: all cell line collapsed PC related terms, genotype, day, day^2, genotypeXday, and then genotypeXday^2.

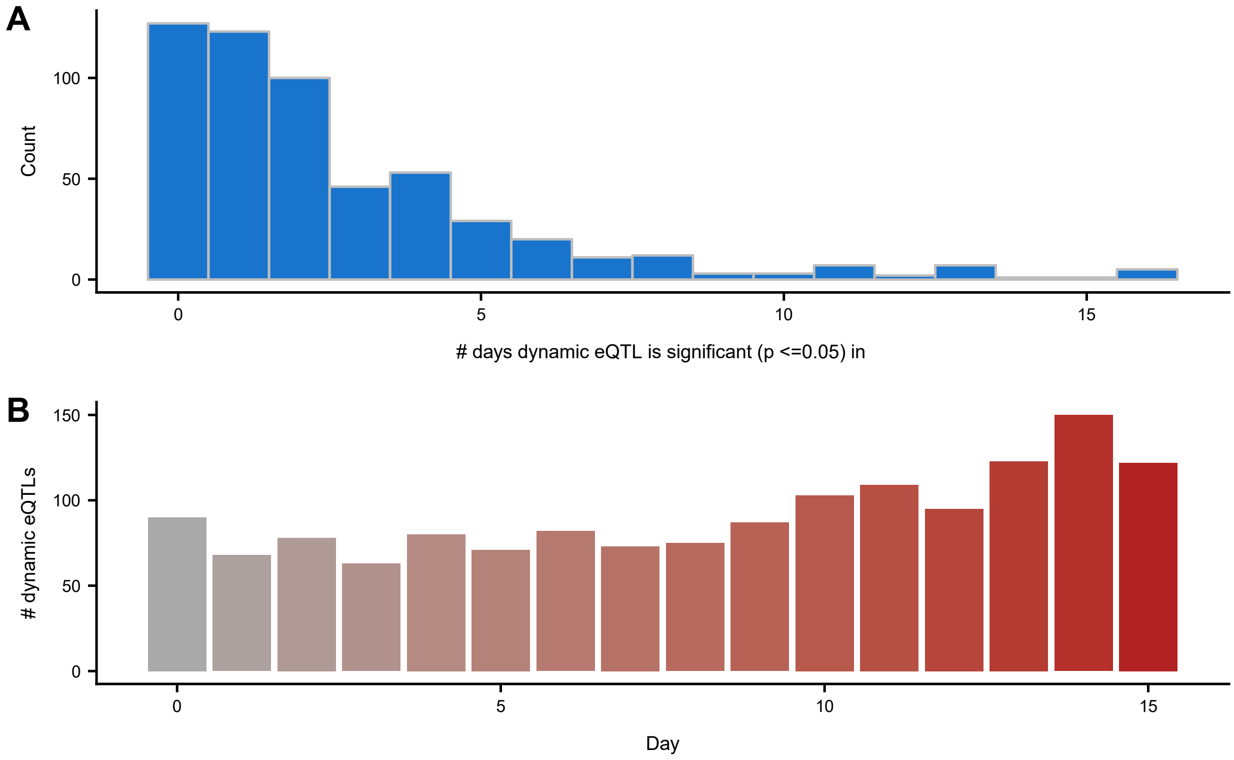


Figure S20: Comparing linear dynamic eQTLs to non-dynamic eQTLs. (A) The number of time points in which the dynamic eQTLs (most significant variant per dynamic eQTL gene) have a nominally significant (p <= .05) non-dynamic eQTL. (B) The number of dynamic eQTLs (most significant variant per dynamic eQTL gene) that are nominally significant (p <= .05) in each time point.

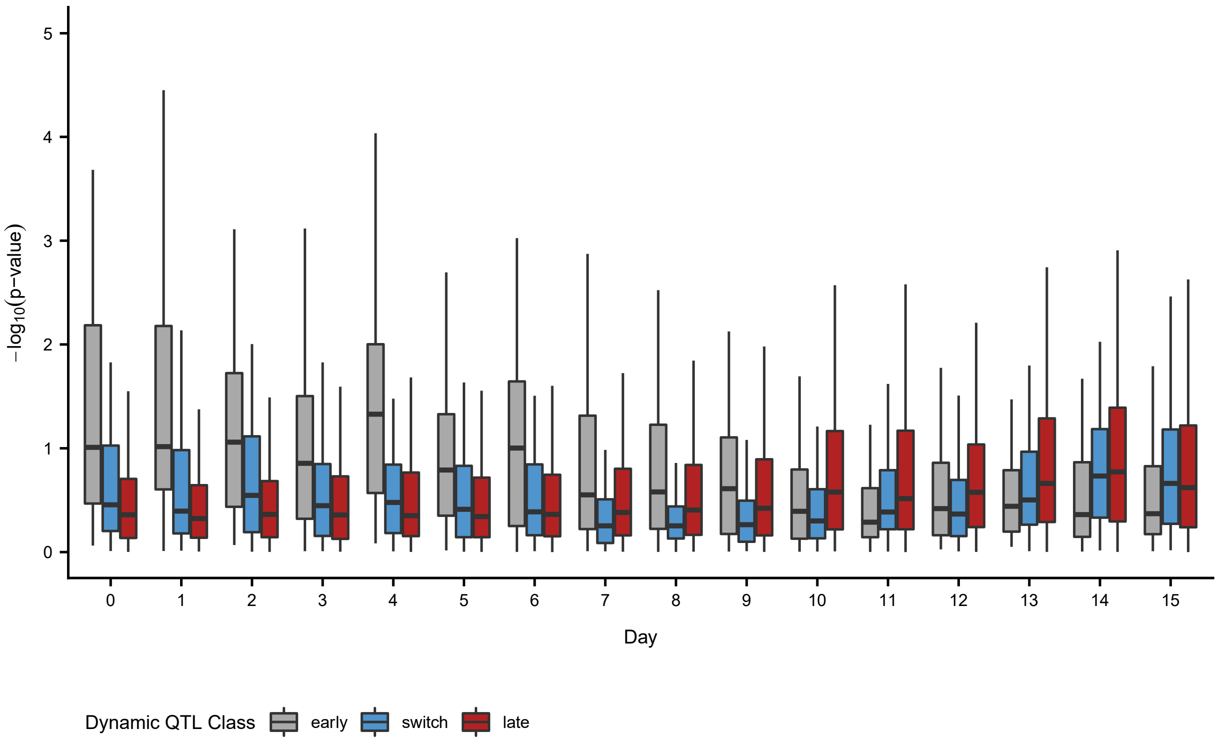


Figure S21: Comparing linear dynamic eQTLs with non-dynamic eQTLs: Non-dynamic eQTL p-values (y-axis) in all 16 time points (x-axis) of linear dynamic eQTLs (most significant variant per dynamic eQTL gene) stratified by linear dynamic eQTL classifications (early, switch, and late).

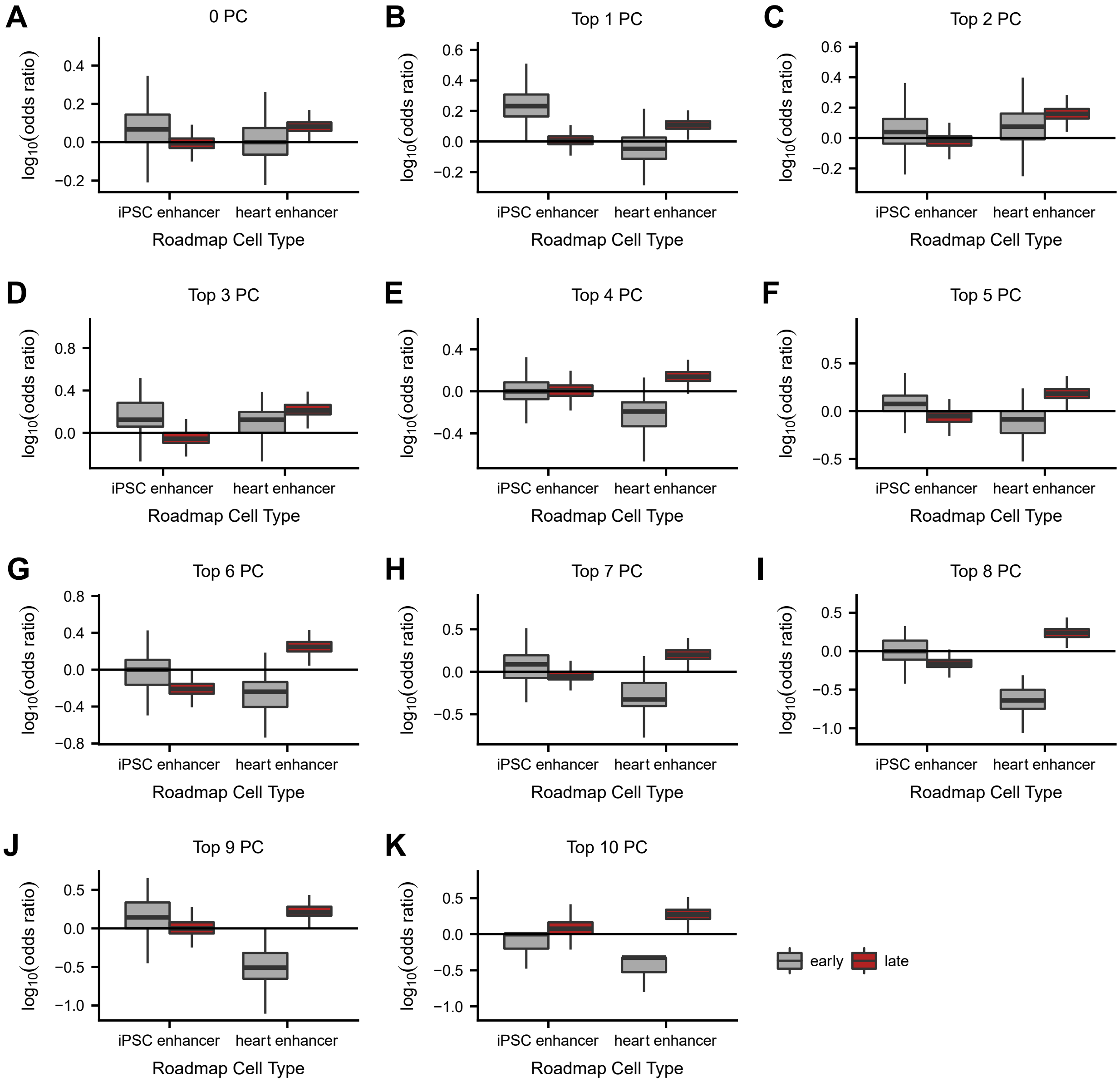


Figure S22: Dynamic eQTL enhancer enrichment. Enrichment of dynamic eQTLs within cell type specific chromHMM enhancer elements relative to 1000 sets of randomly selected background variants matched for distance to transcription start site and minor allele frequency. Dynamic eQTLs were classified as early (eQTL effect size decreasing over time) or late (eQTL effect size increasing over time). Analysis shown for linear dynamic eQTLs while controlling for a range of the top cell line collapsed PCs (A-K).

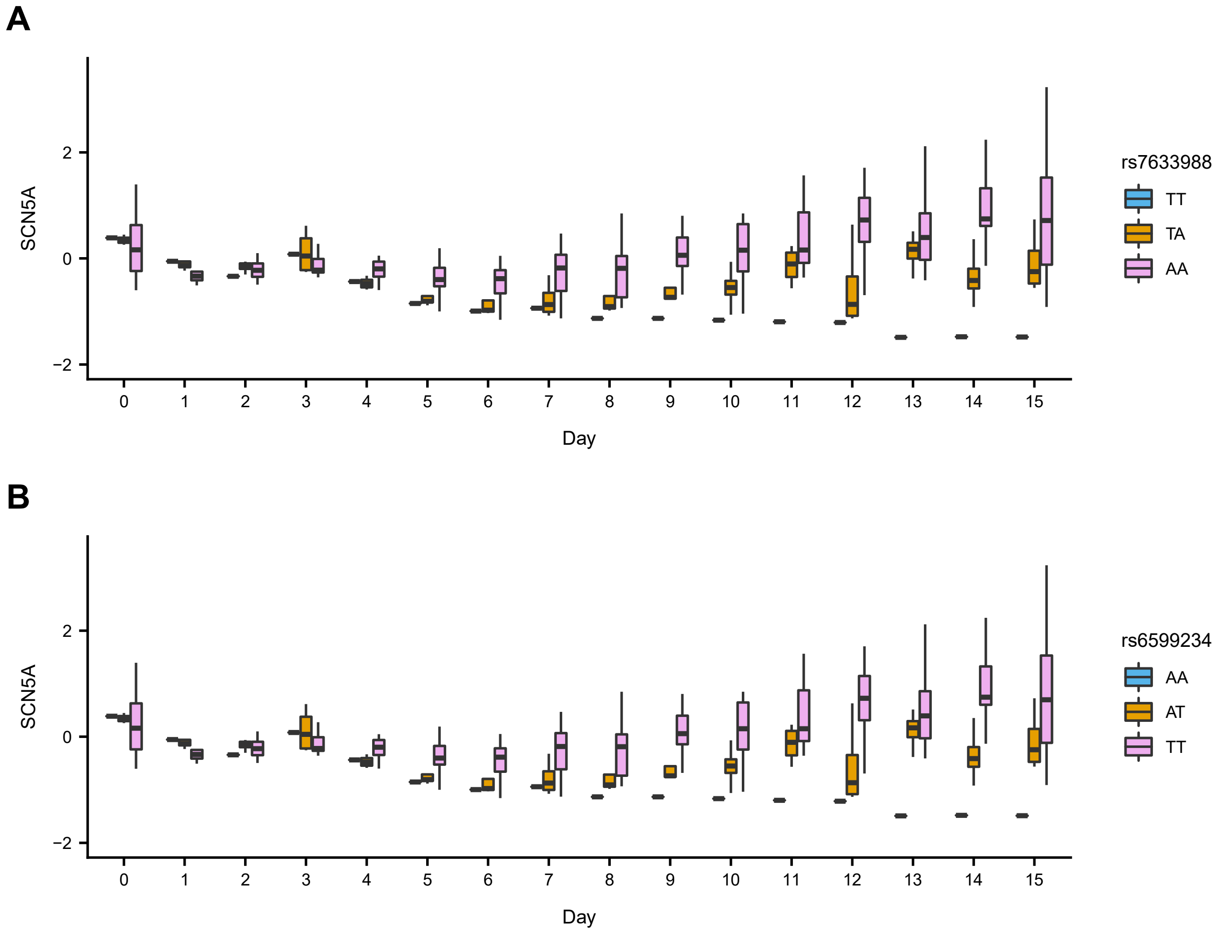


Figure S23: Two significant linear dynamic eQTLs are known GWAS variants. Linear interaction association between time point (x-axis) and genotype (color) of (A) rs7633988 and (B) rs6599234 on residual gene expression (cell line effects regressed on expression) of SCN5A (y-axis).

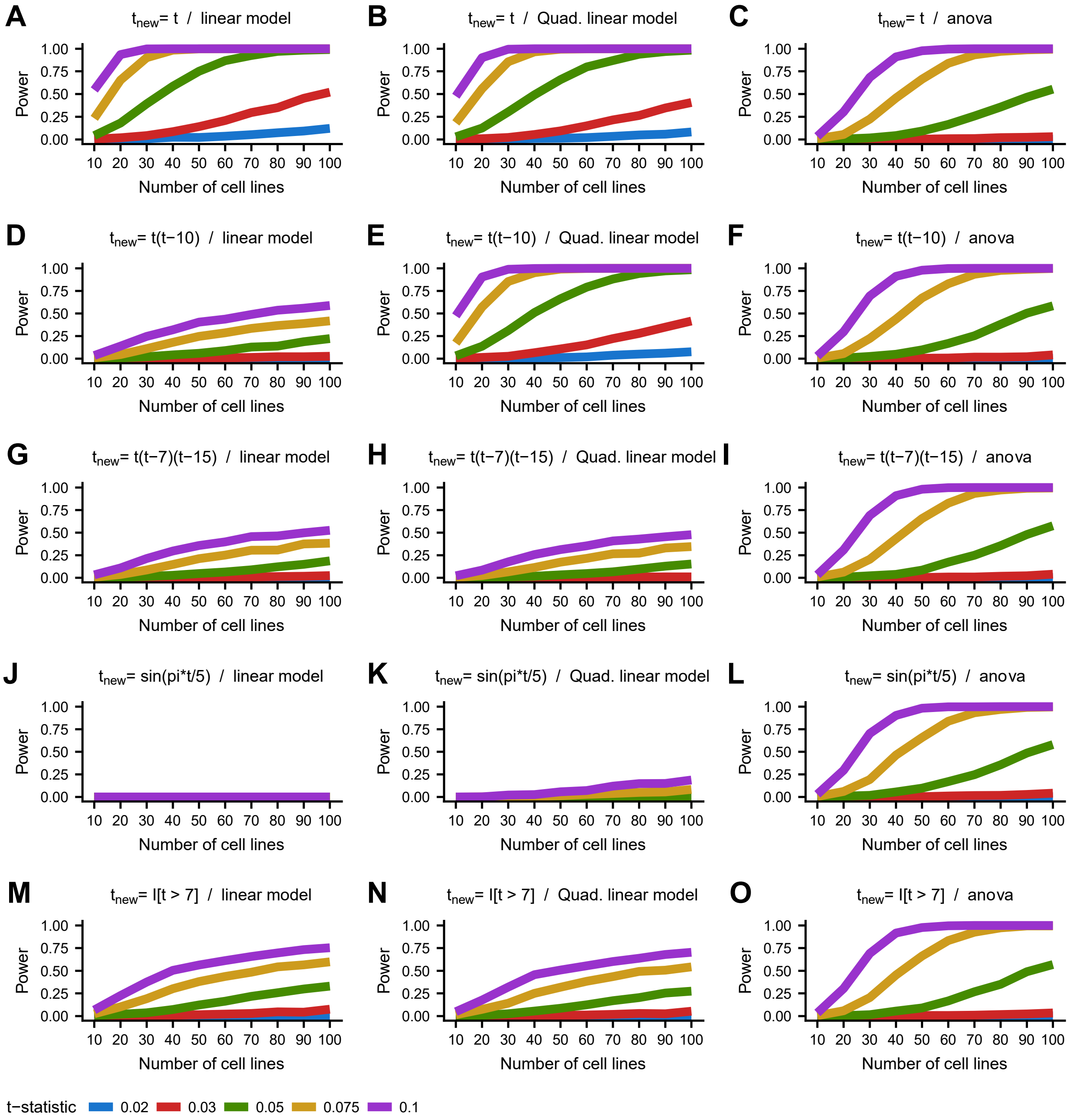


Figure S24: Non-linear simulated power analysis. Power to detect simulated dynamic eQTLs (y-axis) based on 10,000 simulations at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data) as a function of number of cell lines (x-axis) and t-statistic (color). t-statistic represents the ratio of the effect size of the interaction term and the standard deviation term used to simulate the expression data. Simulated expression was generated based on various transformations (t_{new}; rows) of the original values of differentiation time (t). Transformed differentiation time was scaled to have the same standard deviation as the original values of differentiation time. Three different statistical models were used to identify dynamic eQTLs (columns): linear model (linear dynamic eQTL), quadratic linear model (nonlinear dynamic eQTL), and categorical ANOVA analysis. The simulated MAF was .4 and 30% of all simulated tests were drawn from the alternative hypothesis.

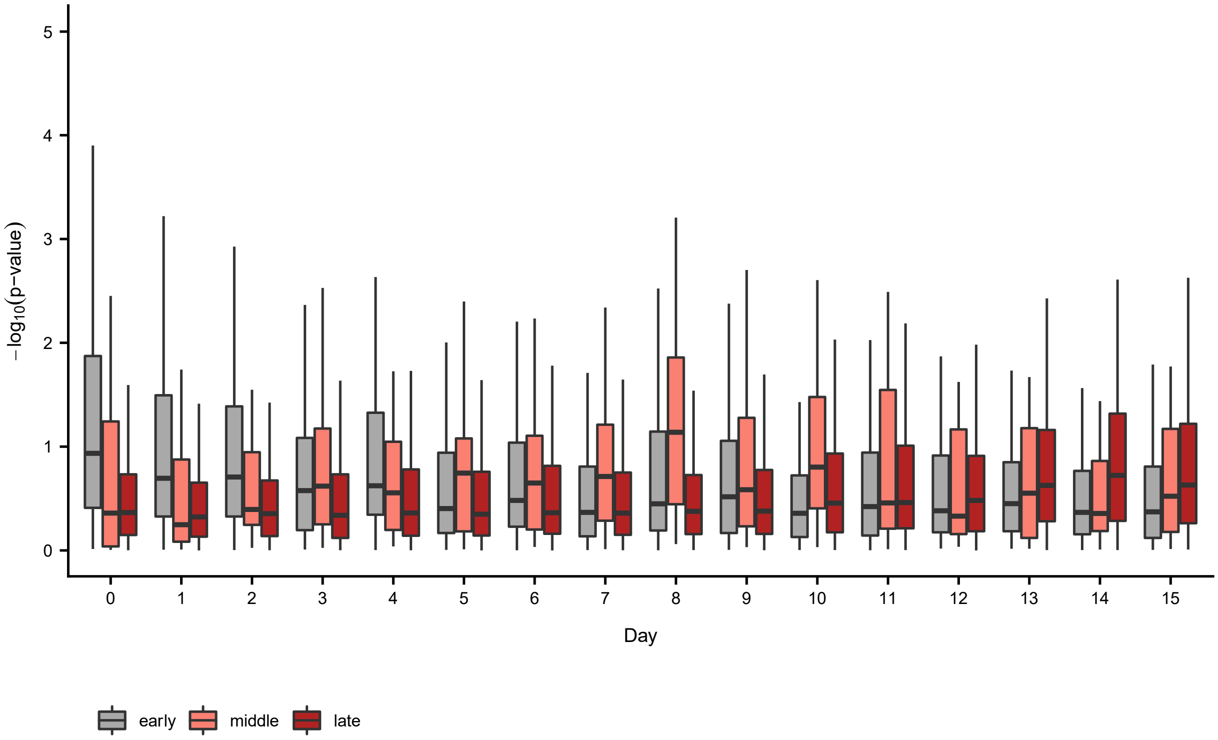


Figure S25: Comparing nonlinear dynamic eQTLs to non-dynamic eQTLs. Non-dynamic eQTL p-values (y-axis) in all 16 time points (x-axis) of nonlinear dynamic eQTLs (most significant variant per dynamic eQTL gene) stratified by nonlinear dynamic eQTL classifications (early, middle, and late).

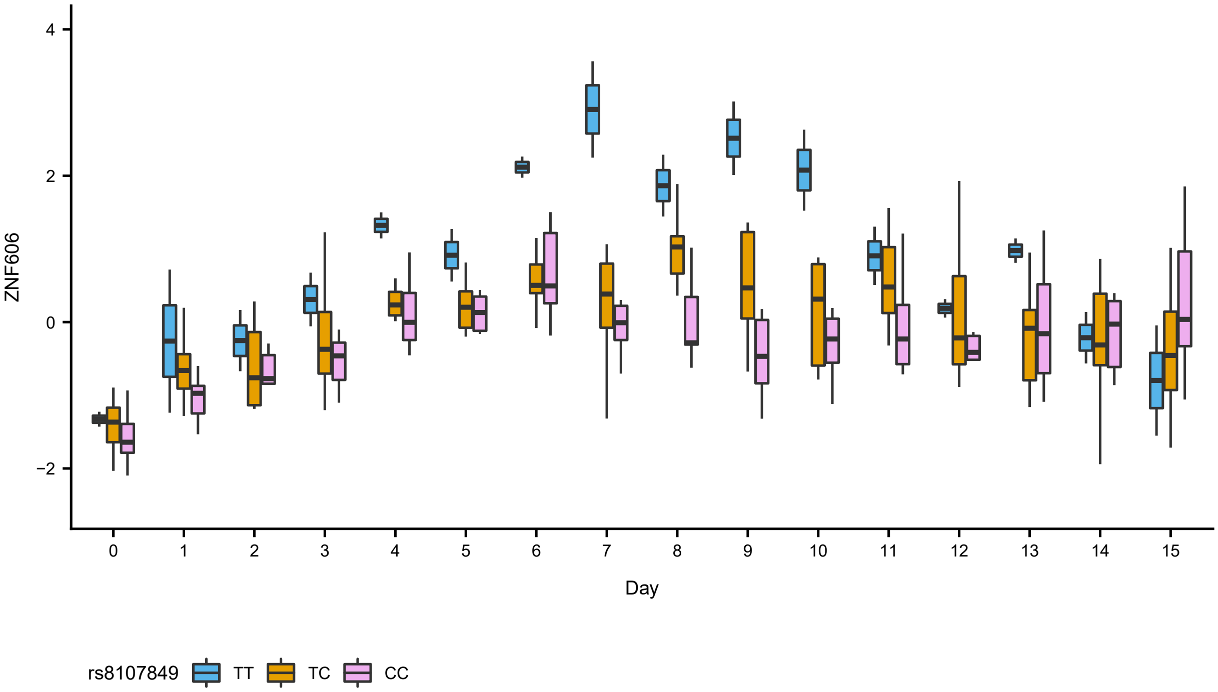


Figure S26: Middle dynamic eQTL example: Nonlinear interaction association between genotype (color) of rs8107849 and time point (x-axis) on residual gene expression (cell line effects regressed on expression) of ZNF606 (y-axis).

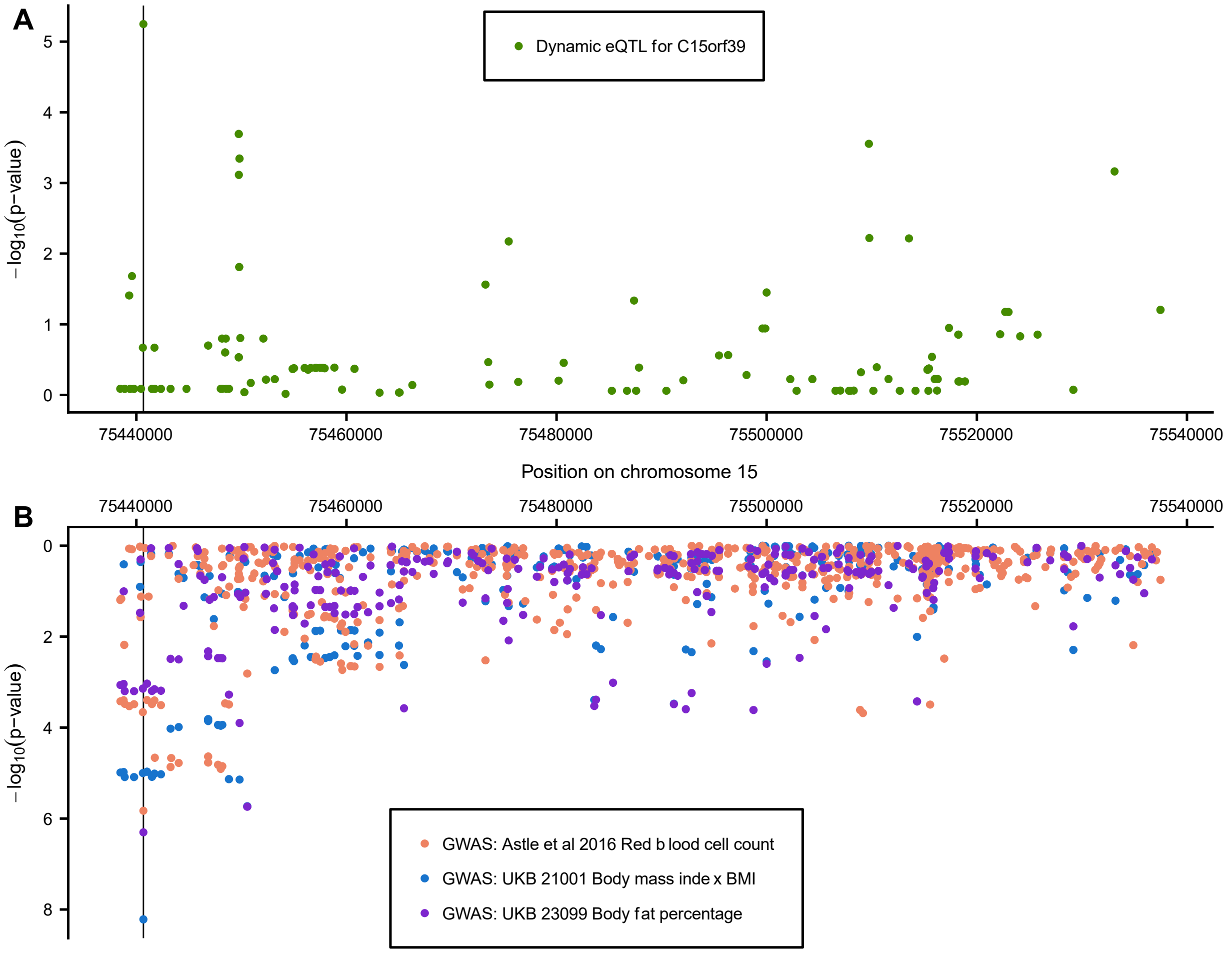


Figure S27: Nonlinear dynamic eQTL overlaps GWAS variant: (A) Manhattan plot showing interaction association p-values for C15orf39 according to nonlinear dynamic eQTL calling for all variants tested within 50KB of the C15orf39 transcription start site. (B) Manhattan plot showing GWAS p-values on the same region surrounding C15orf39 from three different GWAS studies (colors) (23, 24). Vertical line depicts genomic location of most significant nonlinear dynamic eQTL (rs28818910) for C15orf39. p-values shown for body mass index and body fat percentage are based on round 1 of UK Biobank (UKB) (23). Body mass index and body fat percentage p-values for rs28818910 according to the round 2 of UKB (31) become slightly less extreme (p=1.322e-07 and p= 2.521e-06, respectively), but are still significant after multiple testing correction for all significant (eFDR <= .05) nonlinear dynamic eQTL variants (Bonferroni p= 0.000902 and Bonferroni p=.0172, respectively).

Chapter C

Supplementary Figures

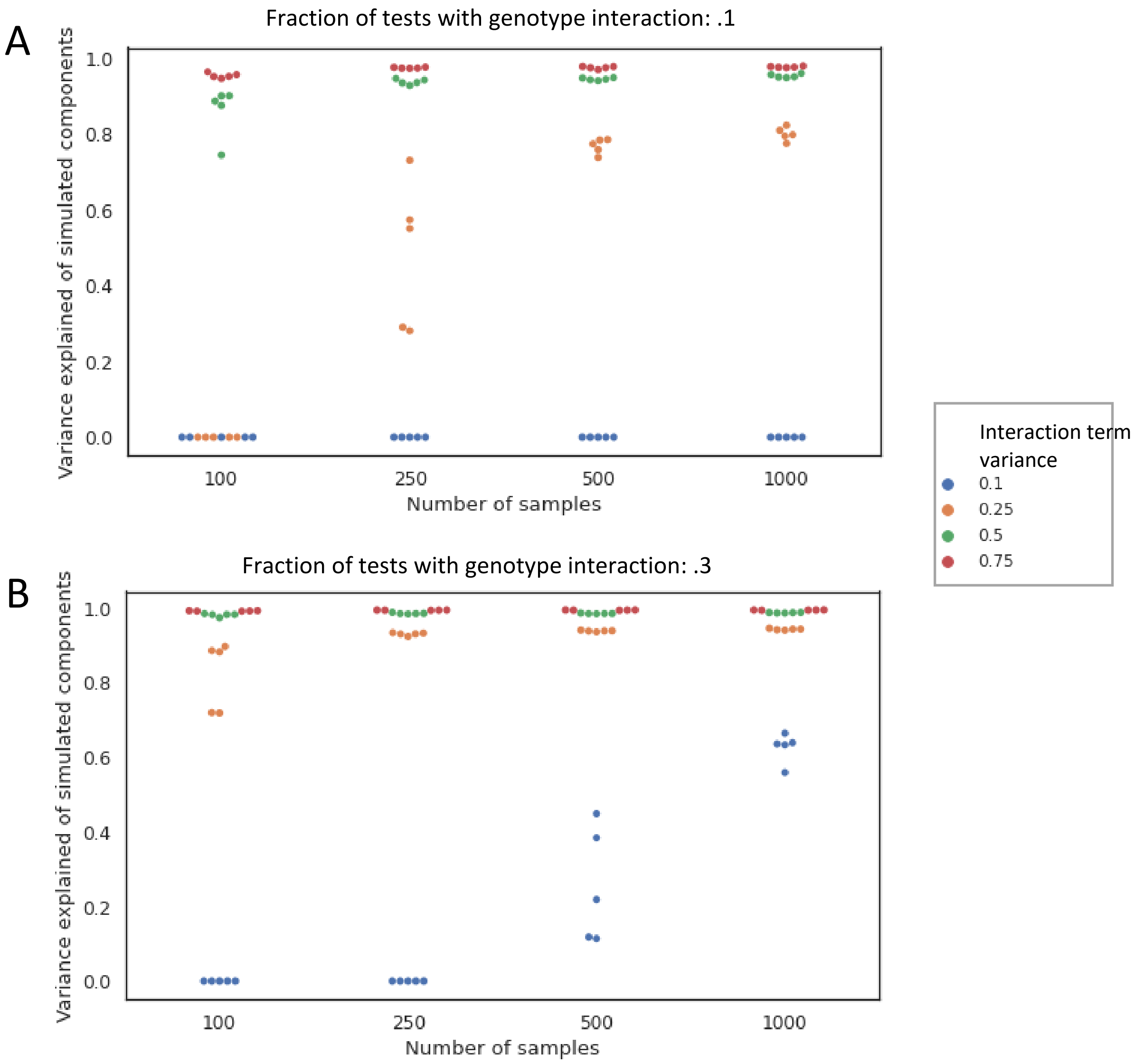


Figure S1: In this simulation, we evaluate SURGE's ability to re-capture simulated latent contexts as measured by the variance explained of the simulated components by the learned components (y-axis). In this simulation we vary the sample size (x-axis), the strength (variance) of the simulated interaction terms (colors), and the fraction of tests that are context-specific eQTLs for a particular context (A, B). For each parameter setting, we run 5 independent simulations.

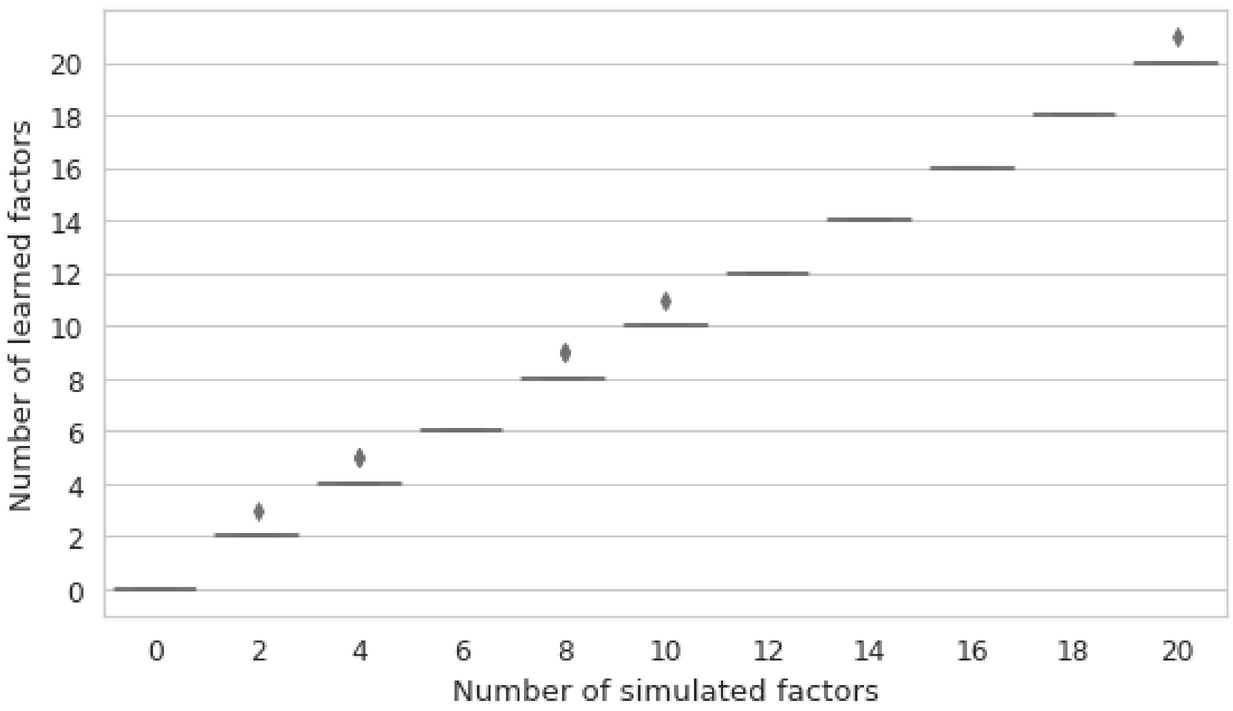


Figure S2: In this simulation, we evaluate SURGE's ability to identify the number of simulated latent contexts (x-axis) over 5 independent simulations and SURGE optimizations (y-axis). In this simulation, the sample size was fixed to 250, the strength (variance) of the simulated interaction terms was fixed to .5, and the fraction of tests that are context-specific eQTLs for a particular context was fixed to .3.

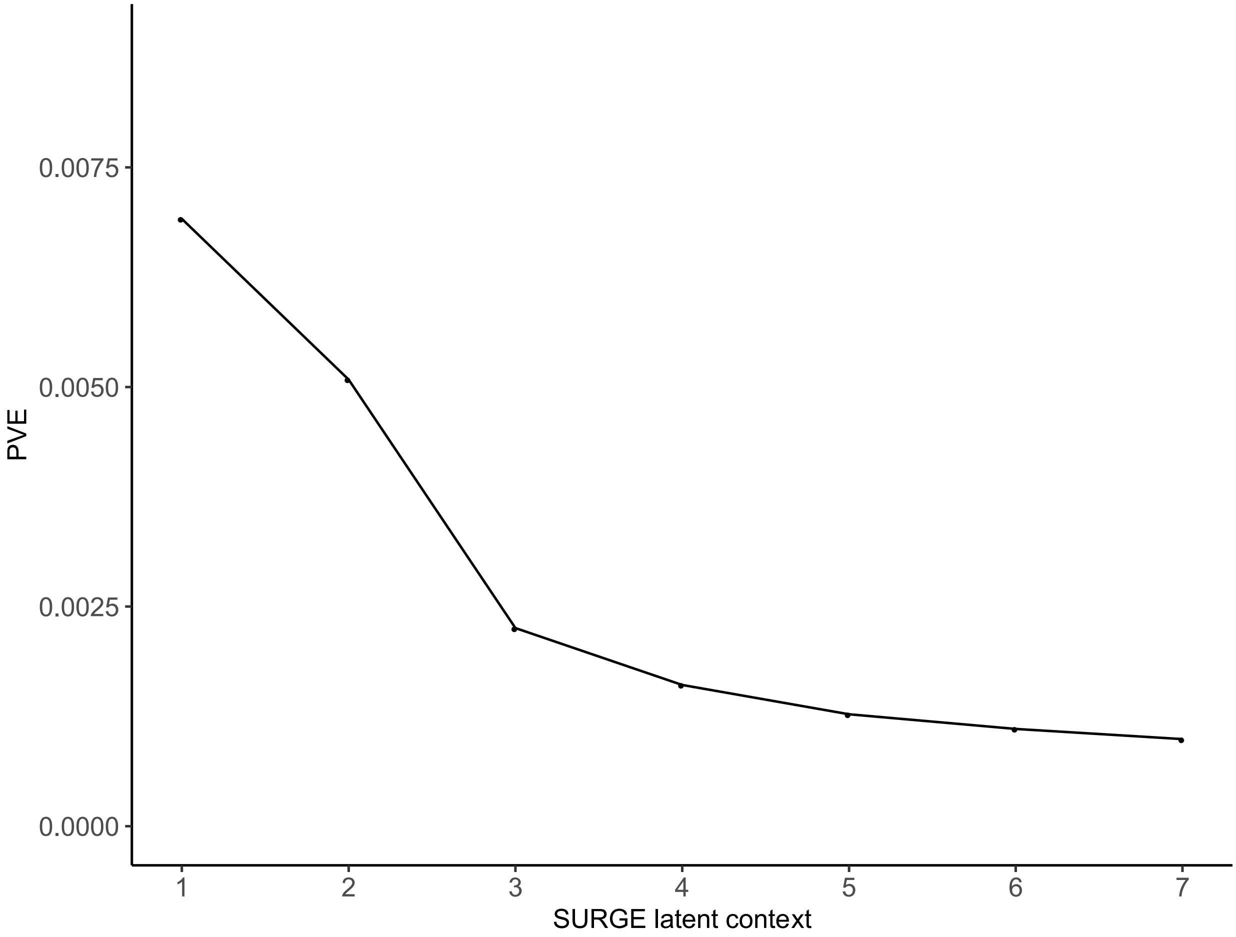


Figure S3: Percent variance explained (PVE; see Methods; y-axis) of the 7 SURGE latent contexts identified when SURGE was applied to samples concatenated across 10 GTEx v8 tissues.

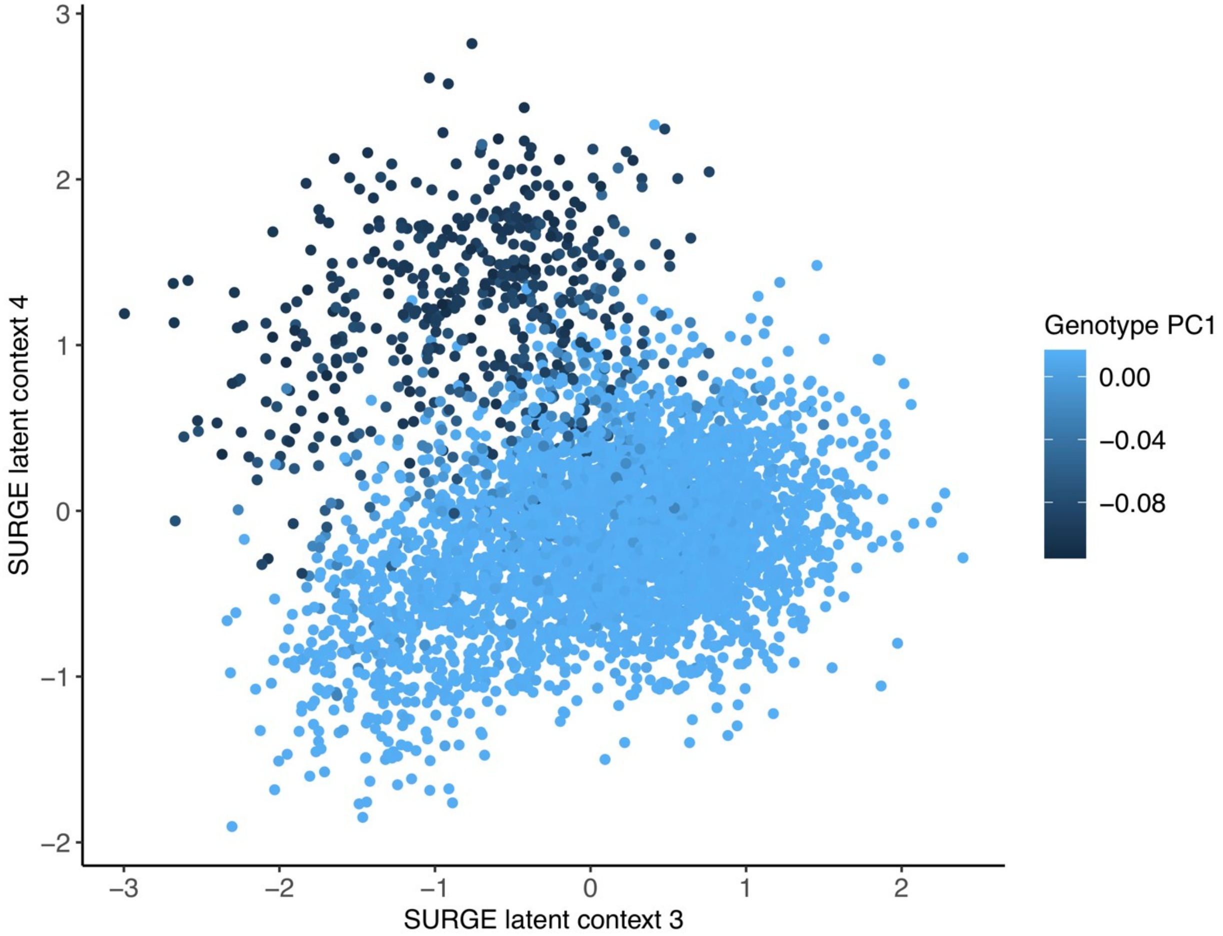


Figure S4: Scatter-plot of SURGE latent context 3 values (x-axis) by SURGE latent context 4 (x-axis) values across all GTEx version 8 samples concatenated over 10 GTEx tissues. Samples are colored by their loading on the first Genotype PC.

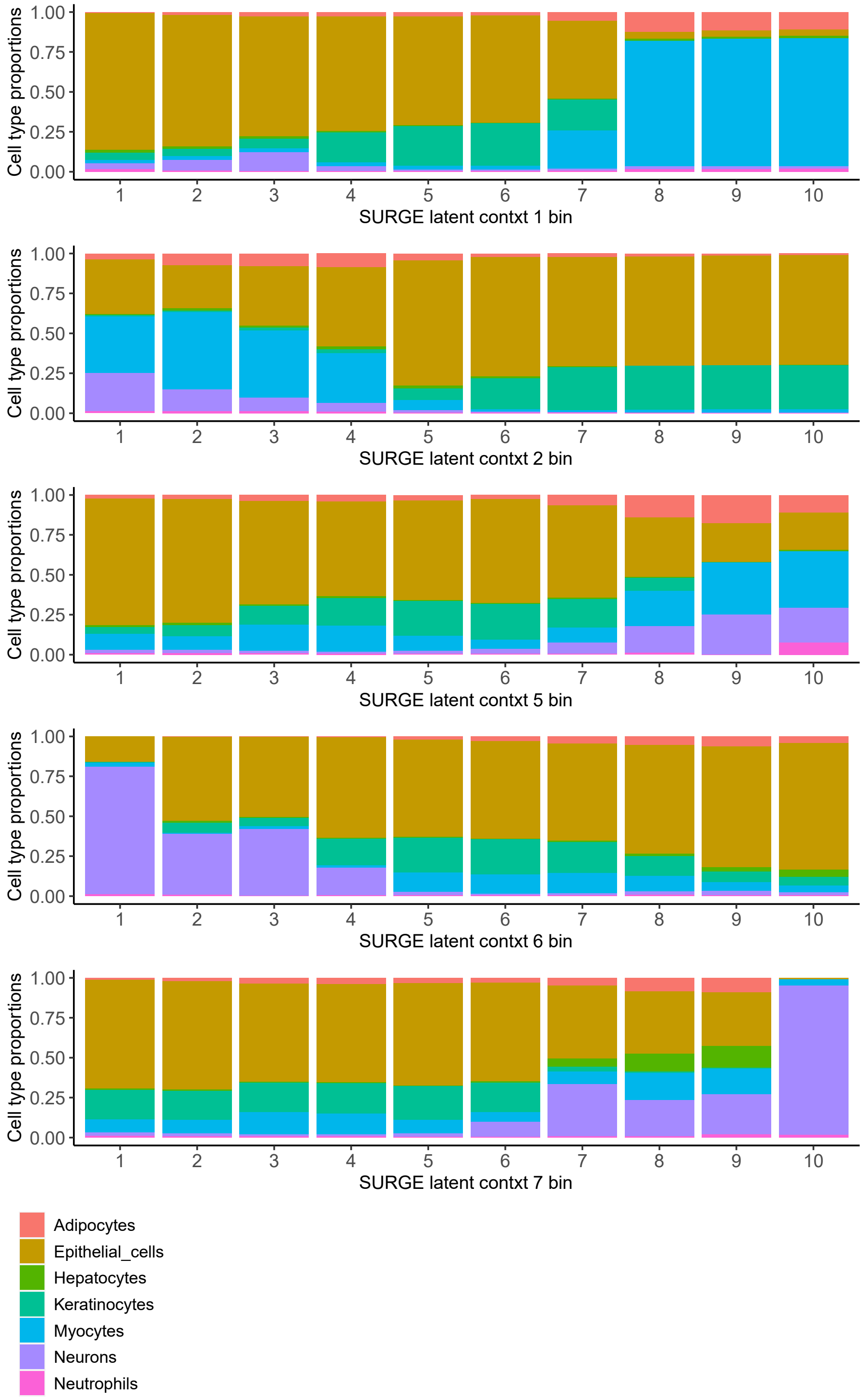


Figure S5: GTEx v8 RNA-seq samples are separated into 10 equally-sized bins according to their value on SURGE latent context 1, 2, 5, 6, and 7 (rows). The stacked bar plots depicts the average cell-type composition according to xCell estimates across all samples (y-axis) in each of the 10 bins (x-axis). These results were generated when SURGE was applied to samples from 10 GTEx v8 tissues.

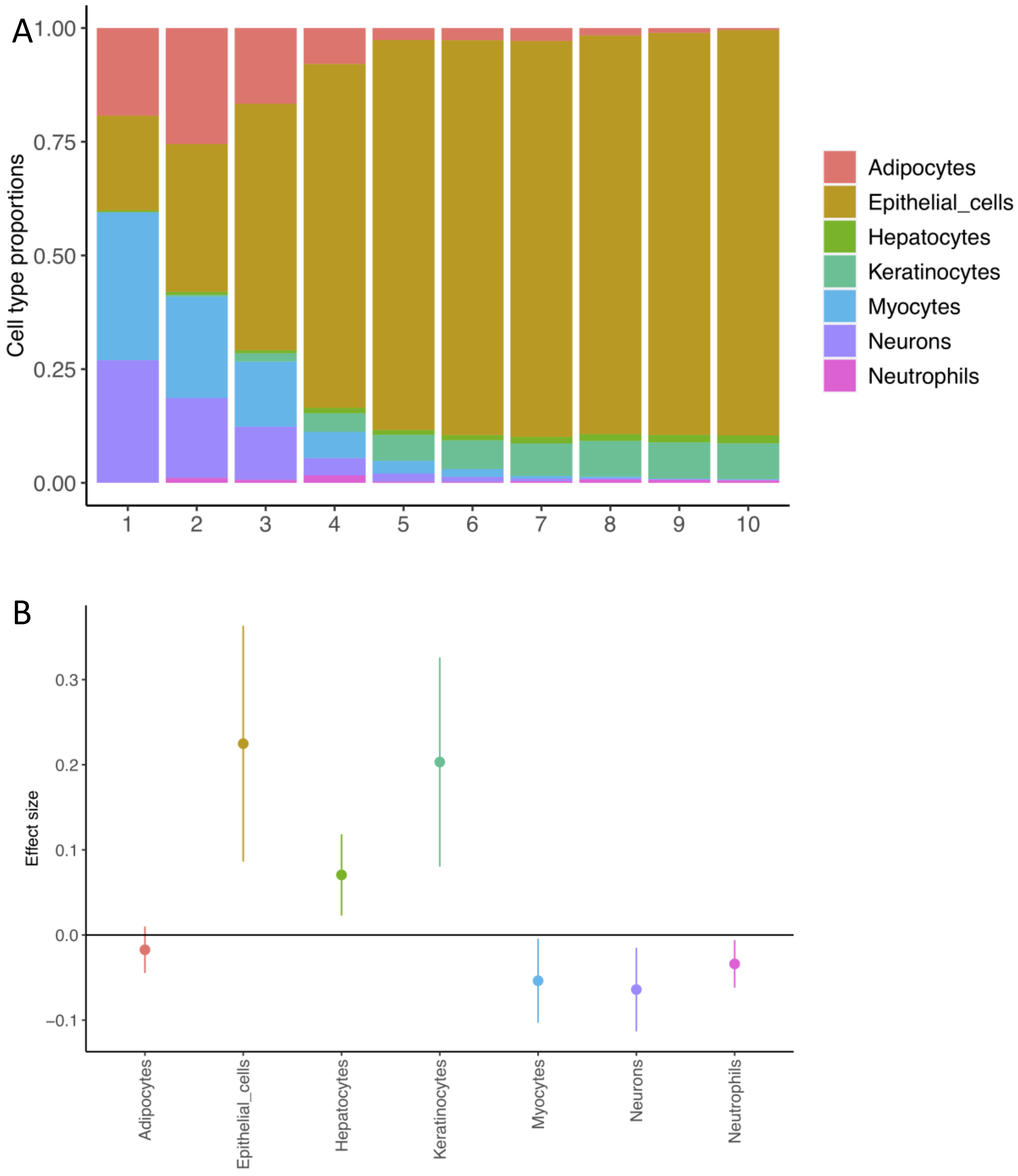


Figure S6: These results were generated when SURGE was applied to samples from only Colon-Sigmoid GTEx v8 tissue. (A) GTEx v8 Colon-Sigmoid RNA-seq samples are separated into 10 equally-sized bins according to their value on SURGE latent context 1. The stacked bar plot depicts the average cell-type composition according to xCell estimates across all samples (y-axis) in each of the 10 bins (x-axis). (B) We fit a multivariate linear model to predict SURGE latent context 1 from xCell cell type proportions across 8 cell types. This plot shows the effect sizes and standard error of the effect sizes from this multivariate linear model (y-axis) for each of the 8 cell types that were used as fixed effects in the model (x-axis). 6 of the 7 cell types are predictive of SURGE latent context 1, even when conditioned on all other cell types.

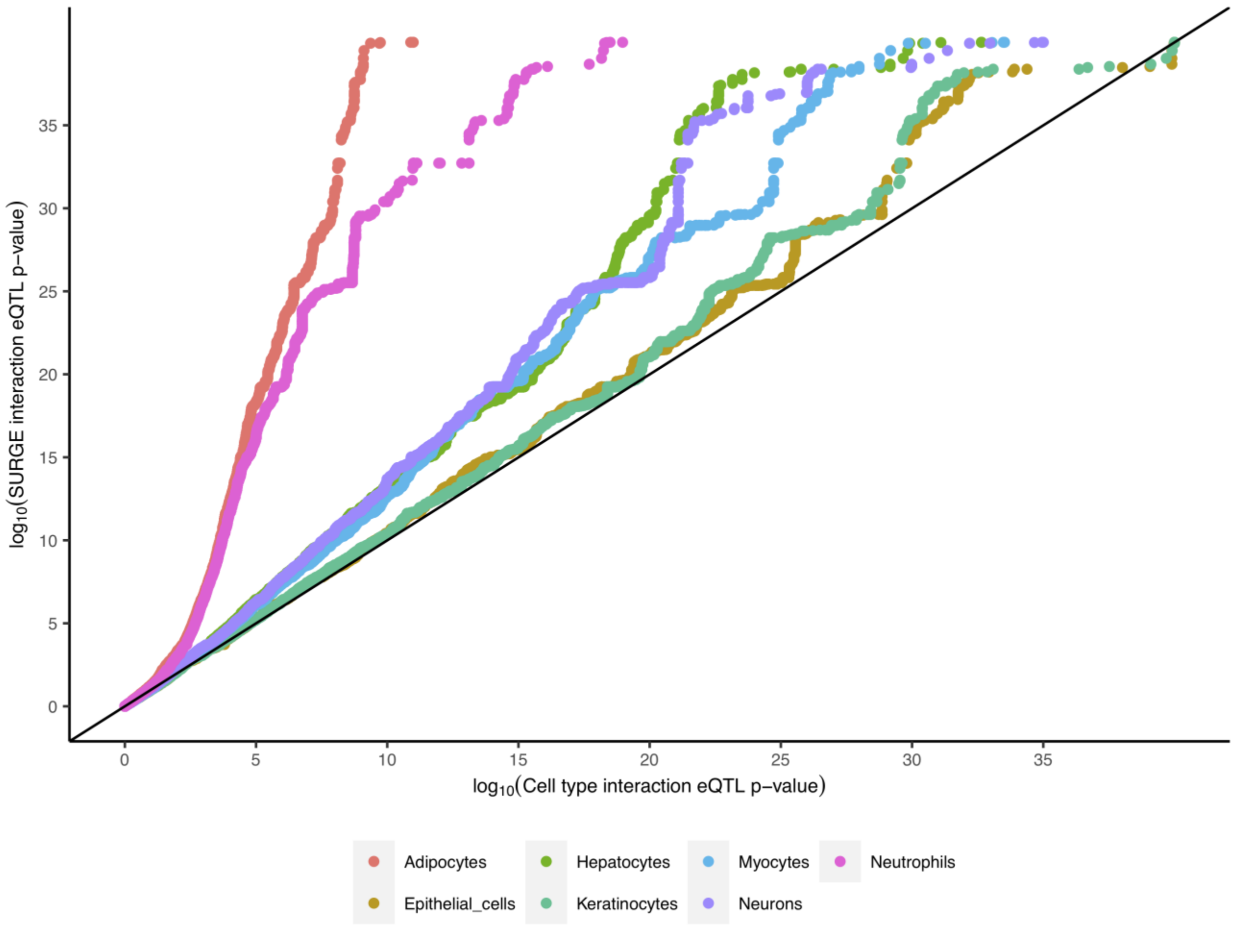


Figure S7: These results were generated when SURGE was applied to samples from only Colon-Sigmoid GTEx v8 tissue. -log10(pvalues) of SURGE context 1 interaction eQTLs (y-axis) compared to -log10(pvalues) of interaction eQTLs using xCell cell type proportion from single cell type as the context (x-axis). Results shown for all 7 xCell cell types (colors).

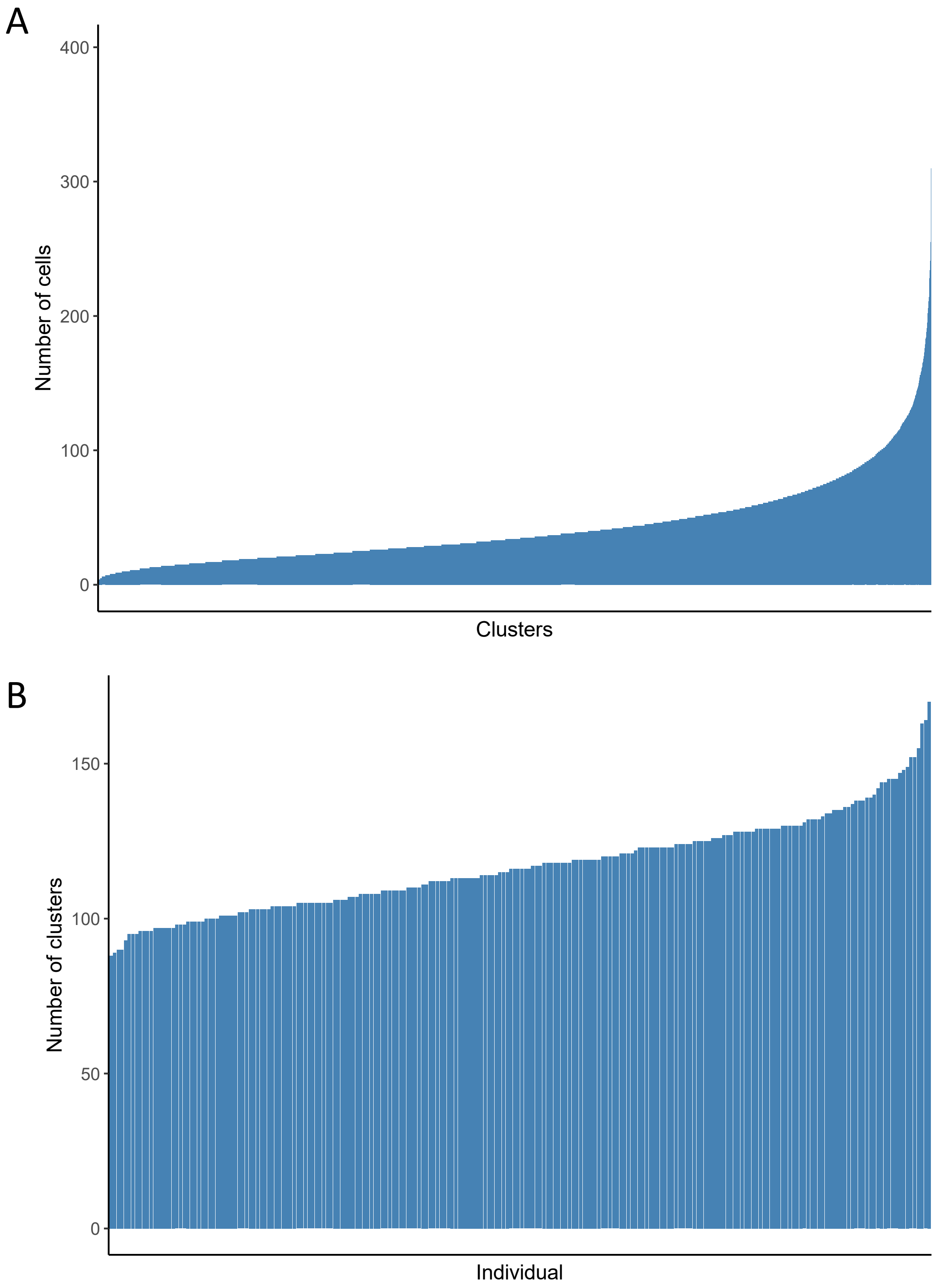


Figure S8: Pseudocell aggregation of single cell expression data. (A). Distribution of number of cells (y-axis) per pseuodcell (x-axis). (B) Distribution of number pseudocells (y-axis) per individual (x-axis).

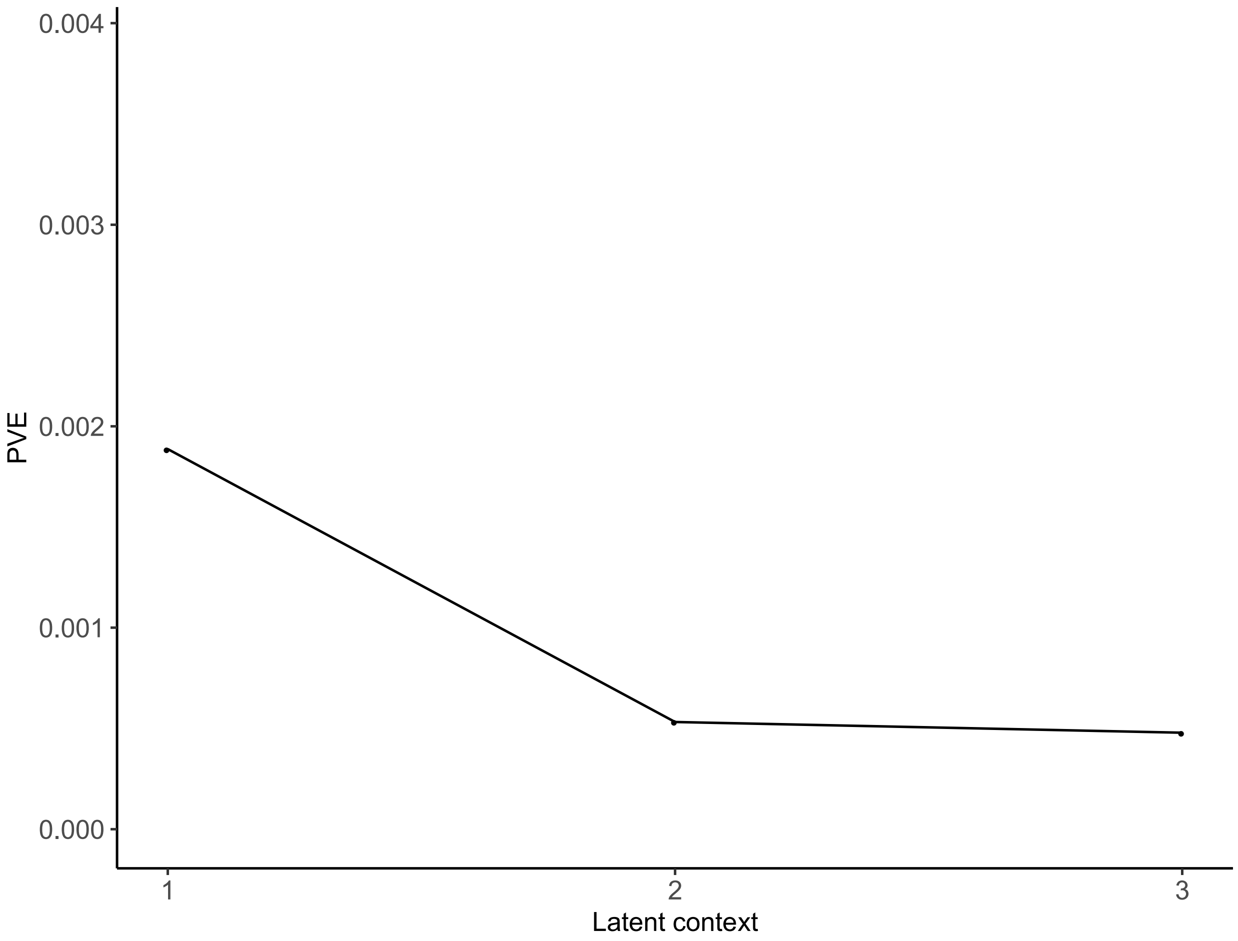


Figure S9: Percent variance explained (PVE; see Methods; y-axis) of the 3 SURGE latent contexts identified when SURGE was applied to single cell eQTL data.

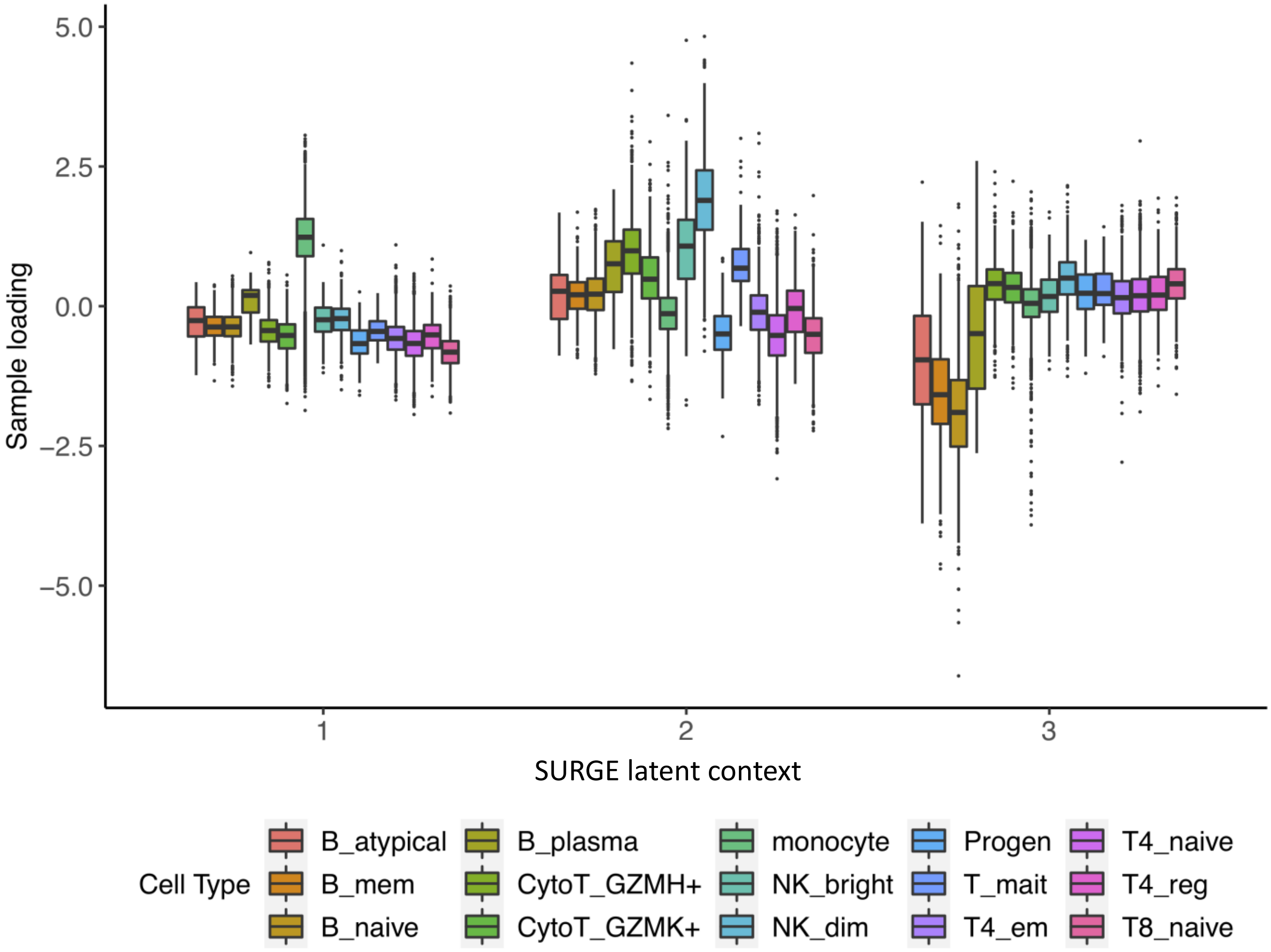


Figure S10: SURGE latent context loadings of pseudocells (y-axis) stratified by cell type according to marker gene expression profiles for each of the 3 identified SURGE latent contexts.

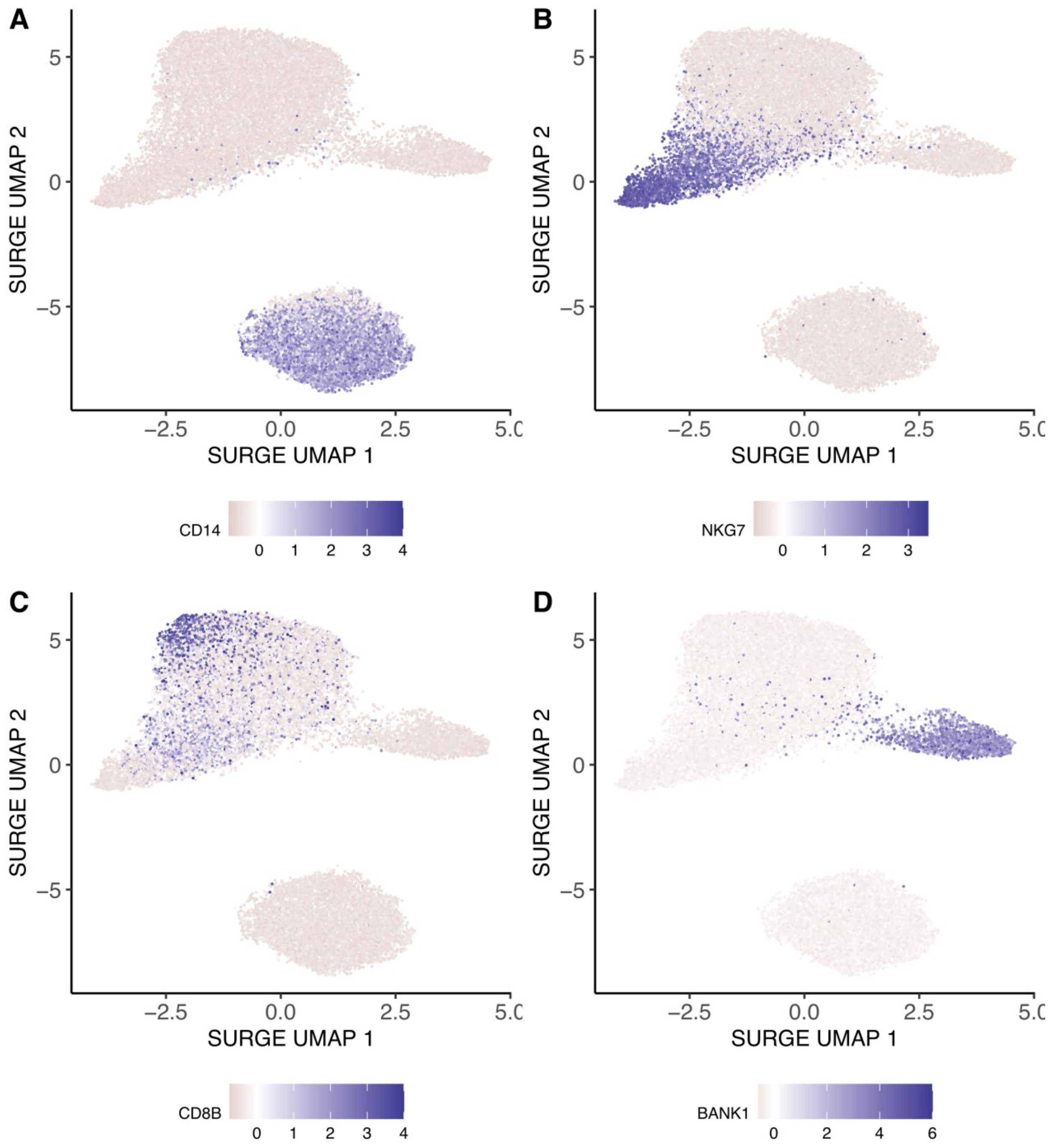


Figure S11: UMAP-projected SURGE latent context loadings of pseudocells (x and y-axis) colored by expression levels of four marker genes: (A) CD14, (B) NKG7, (C) CD8B, (D) BANK1.

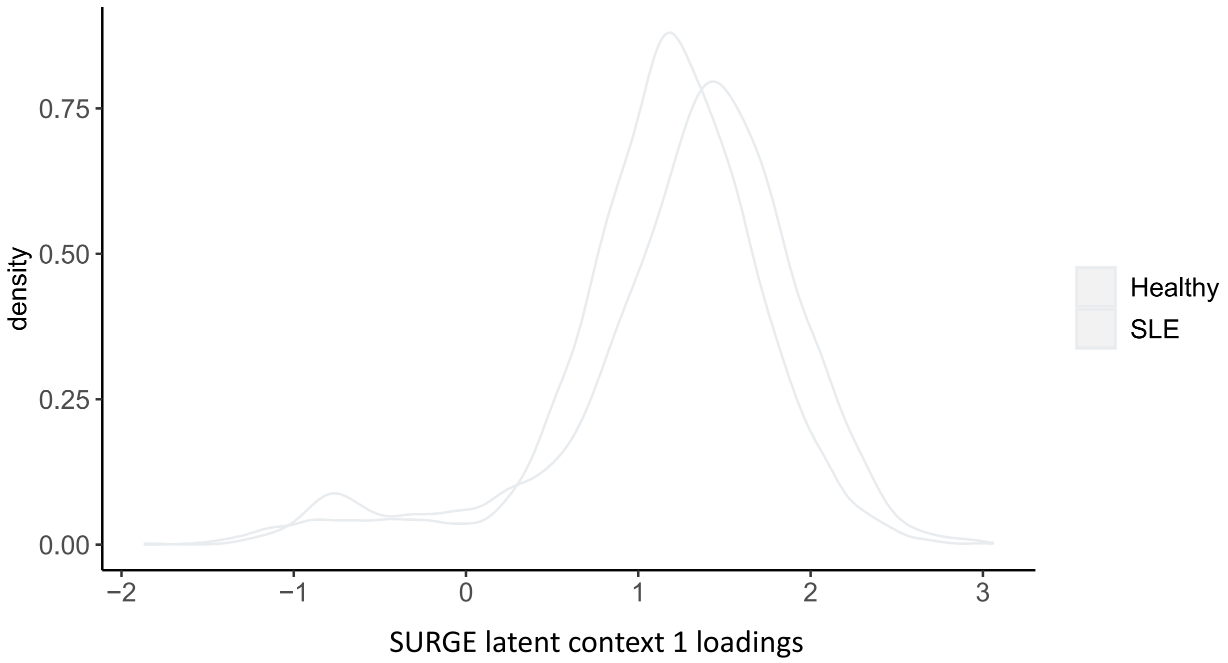


Figure S12: Density of (y-axis) SURGE latent context 1 loadings (x-axis) on pseudocells annotated as monocytes according to marker-gene expression profiles color-stratified by disease status of individuals corresponding to pseudocells.

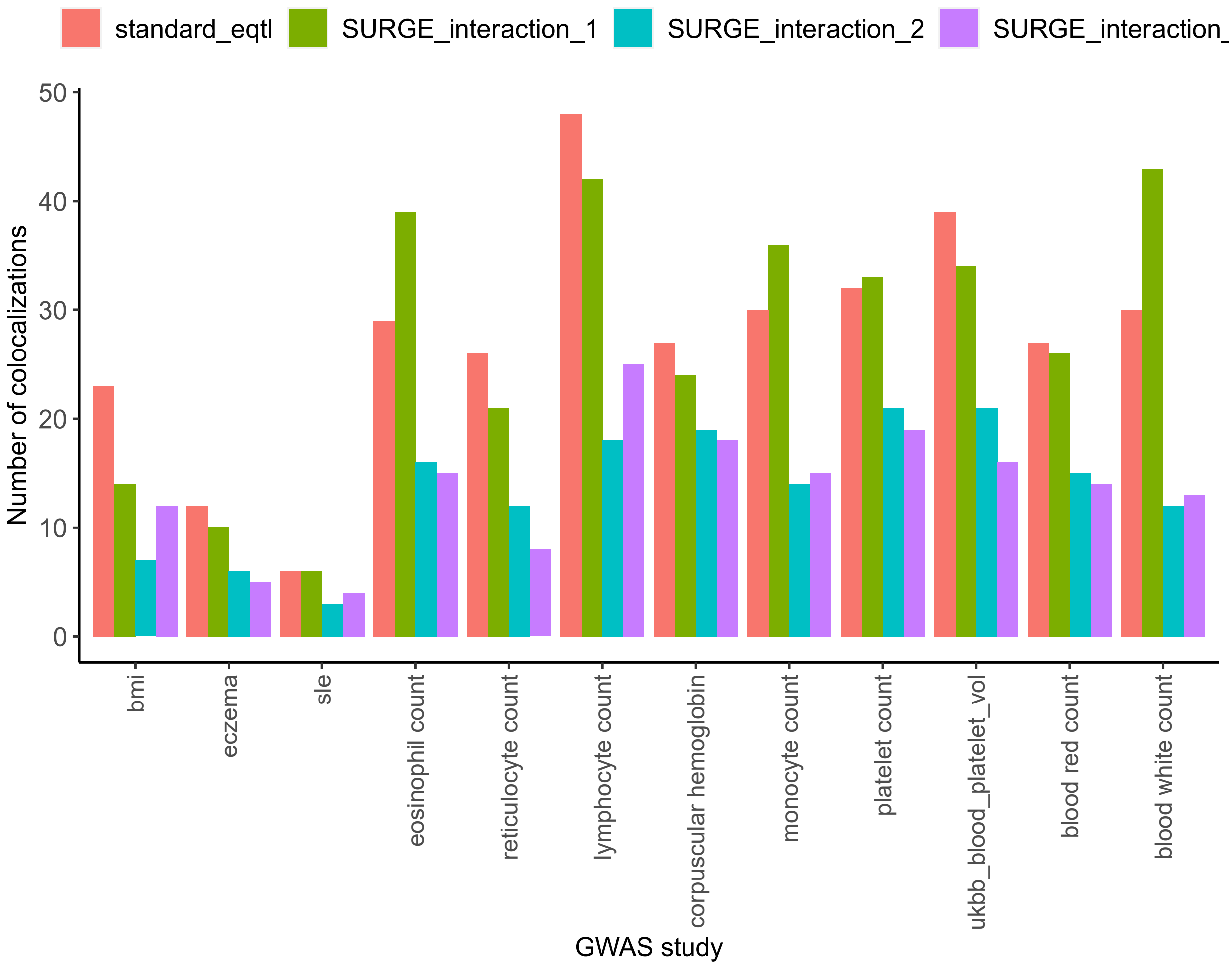


Figure S13: Number of colocalizations identified (PPH4 > .95; y-axis) between various densely genotyped GWAS studies (x-axis) and various categories of eQTLs called from pseudocells.

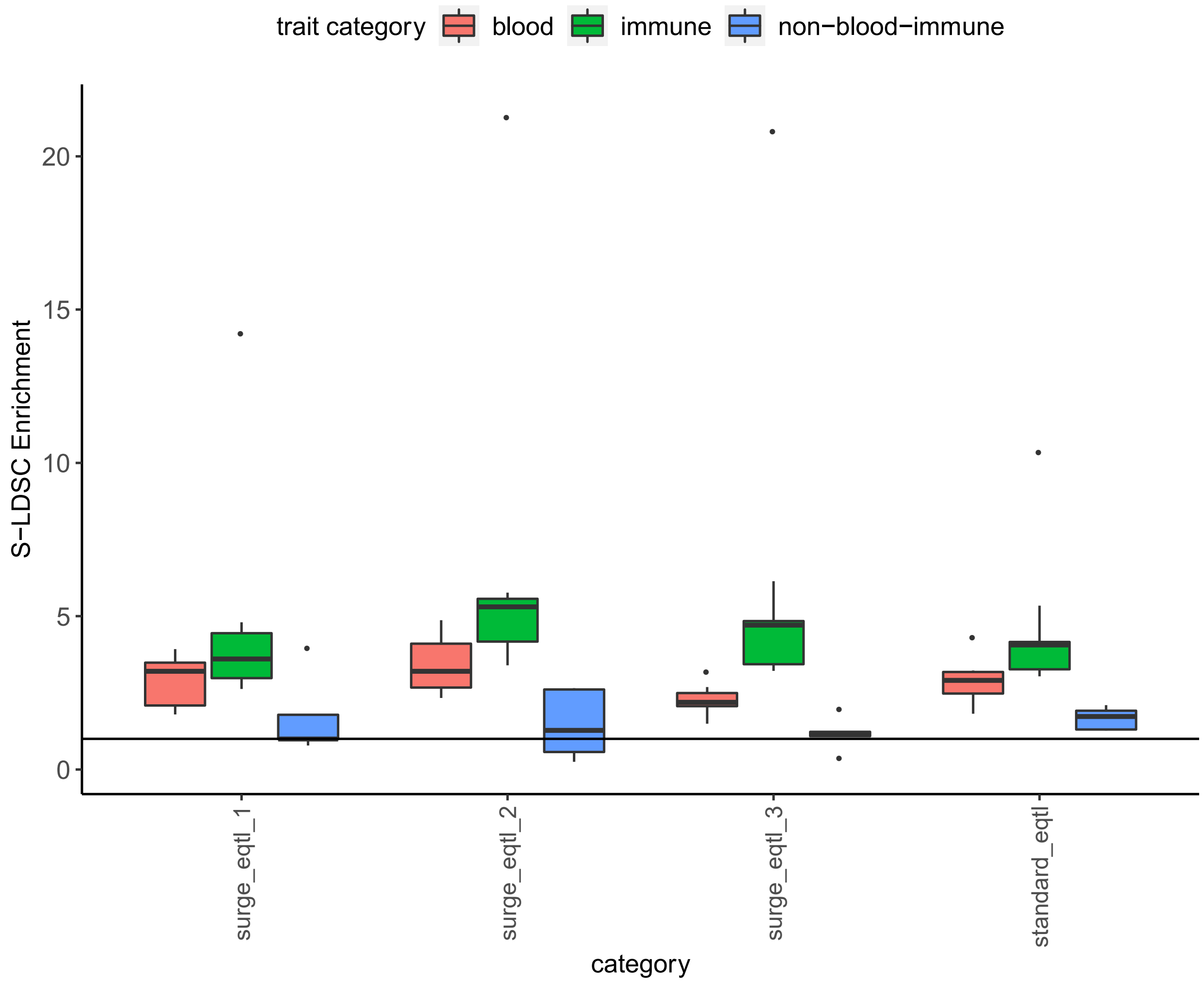


Figure S14: S-LDSC estimates of enrichment (y-axis) corresponding to variant annotations derived from SURGE interaction eQTLs and standard eQTLs (x-axis) for traits belonging to different categories of traits (color). Trait category of "blood" consists of GWAS for eosinophil count, reticulocyte count, lymphocyte count, corpuscular hemoglobin, monocyte count, platelet count, blood platelet volume, red blood count, and white blood count. Trait category of "immune" consists of GWAS for Celiac, Crohns, IBD, Lupus, Multiple-sclerosis, PBC, Rheumatoid Arthritis, Eczema, and Ulcerative Colitis. Trait category of "non-blood-immune" consists of GWAS for Alzheimer, Bipolar, CAD, Schizophrenia, BMI, height, and type-2 Diabetes. surge_eqtl_x corresponds to a binary annotation isolating all variants with SURGE interaction eQTL pvalue < 1e-5 with respect to SURGE latent context x. Standard_eqtl corresponds to the binary annotation isolating all variants with standard eQTL pvalue < 1e-5. S-LDSC was run for each eQTL study independently while controlling for BaselineLD annotations.

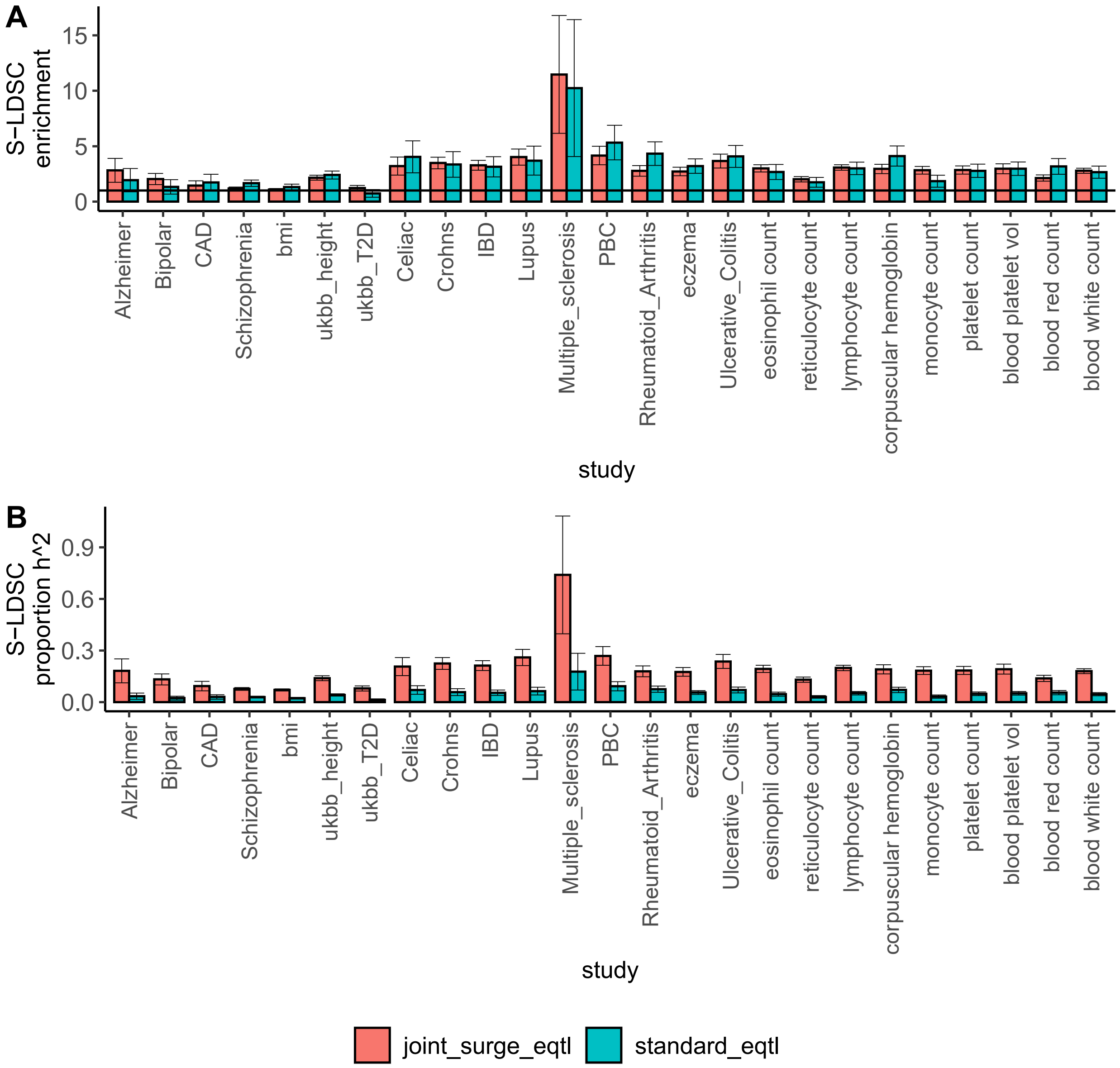


Figure S15: S-LDSC estimates of (A) enrichment (y-axis) and B proportion of total heritability explained (y-axis) for various traits (x-axis) corresponding to variant annotations derived from SURGE interaction eQTLs and standard eQTLs. joint_surge_eqtl corresponds to a binary annotation isolating all variants with SURGE interaction eQTL pvalue < 1e-5 for any of the 3 SURGE latent contexts. Standard_eqtl corresponds to the binary annotation isolating all variants with standard eQTL pvalue < 1e-5. S-LDSC was run for each eQTL study independently while controlling for BaselineLD annotations.

Chapter D

Supplementary Tables

Watershed AUC (PR) - RIVER (AUC) (PR) and corresponding 95% Confidence intervals			
Training and evaluation data	**Expression**	**ASE**	**Splicing**
Standard training data Standard evaluation data (Fig 4D)	**0.050** [-0.012, 0.11]	**0.049** [-0.045, 0.14]	**0.097** [0.034, 0.16]
Training data filter 1 Standard evaluation data (Fig S28A)	**0.056** [-0.0011, 0.11]	**0.043** [-0.046, 0.16]	**0.069** [-0.0045, 0.13]
Training data filter 2 Standard evaluation data (Fig S28B)	**0.046** [-5 x 10^{-5}, 0.087]	**-0.0096** [-0.087, 0.065]	**0.024** [0.0037, 0.042]
Training data filter 3 Standard evaluation data (Fig S28C)	**0.045** [0.00011, 0.085]	**0.0056** [-0.067, 0.075]	**0.024** [0.0062, 0.037]
Evaluation data filter 1 Standard training data (Fig S28D)	**0.033** [0.0031, 0.059]	**0.032** [0.0015, 0.054]	**0.066** [0.03, 0.099]
Evaluation data filter 2 Standard training data (Fig S28E)	**0.05** [0.028, 0.07]	**0.047** [0.022, 0.068]	**0.066** [0.039, 0.091]
Evaluation data filter 3 Standard training data (Fig S28F)	**0.066** [0.043, 0.088]	**0.033** [0.016, 0.049]	**0.076** [0.049, 0.1]

Supplementary Table 1. Change in area under precision recall curves between Watershed and RIVER. Table summarizing the difference in area under the precision recall curves (AUC (PR)) between Watershed and RIVER for each of the three outlier types. 95% confidence intervals on these statistics generated using non-parametric bootstrapping with 20,000 bootstrapped samples (see Supplementary methods). Results shown across 7 different filters placed of Watershed training training or evaluation data (rows of table; See Supplementary methods) corresponding to 7 precision recall curves described in Fig 4D and Fig S28. Standard data corresponds to filtering to genes where all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.01). Filter 1 corresponds to filtering to genes where all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.05). Filter 2 corresponds to filtering to genes where all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.1). Filter 3 corresponds to filtering to genes where at least 1 outlier signals has at least one individual that is an outlier (median p-value < 0.01).

Gene:variant pair	Median expression p-value	Median Watershed expression posterior
P2RX7: chr12:g.121133096G>T	0.0105	0.996
ZNF350: chr19:g.51986869G>A	0.0619	0.925
CADM1: chr11:g.115500916G>A	0.779	0.00249
TSSC1: chr2:g.3377790T>C	0.0323	0.757
ARMC5: chr16:g.31460010C>T	0.0186	0.973

Supplementary Table 2. Replication of SardiNIA Project “candidate causal rare variants”. The SardiNIA Project identified 30 “candidate causal rare variants” (and corresponding regulated genes). The above table shows 5 of the 30 “candidate causal rare variants” that were also present in an individual in GTEx v8, along with corresponding expression outlier p-value and Watershed expression posterior in GTEx v8 individuals. If multiple GTEx v8 individuals harbor the rare variant, we computed the median expression outlier p-value and median Watershed expression posterior across those individuals. SardiNIA Project rare variant calls were lifted to the hg38 genome build from the hg19 genome build using the Genome Browser. The variants from the SardiNIA Project were prioritized with expression outliers, followed by filtering based on genomic annotations. It is important to note that some of the genomic annotations used as input to Watershed were the same genomic annotations used by the SardiNIA Project to generate their list of “candidate causal rare variants”.

Supplementary Figures

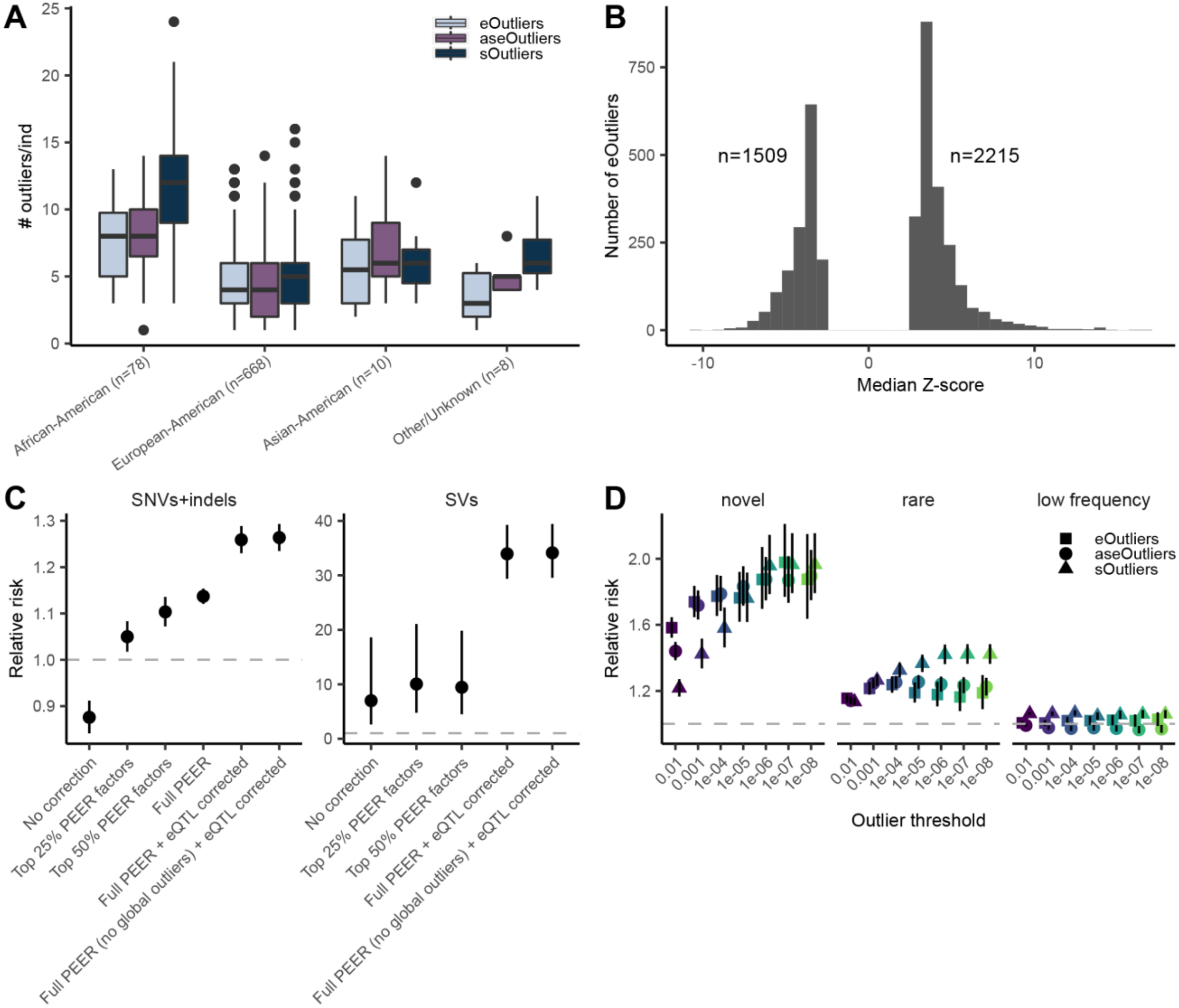


Figure S1. Outlier distribution and effect of expression data correction. (A) Number of outliers per individual across each population defined by self-reported ethnicity, at a threshold of median p-value < 0.0027. (B) Number of eOutliers split by direction of the expression effect. (C) Effect of different expression data correction procedures on relative risk of an outlier having a nearby rare variant. From left, rare (MAF < 1%) variant enrichments for eOutliers identified from uncorrected data, data corrected for first 25% of PEER factors (based on sample size), first 50% of PEER factors, full PEER factors and known covariates, all PEER factors + strongest cis-eQTL per gene, and all PEER factors learned with global outliers removed plus strongest cis-eQTL per gene. (D) Rare SNV and indel enrichments, defined as relative risk, for novel (left), rare (gnomAD AF < 1%), and low frequency (gnomAD AF > 1% and < 5%) within 10kb of outlier genes across a range of outlier thresholds (x-axis).

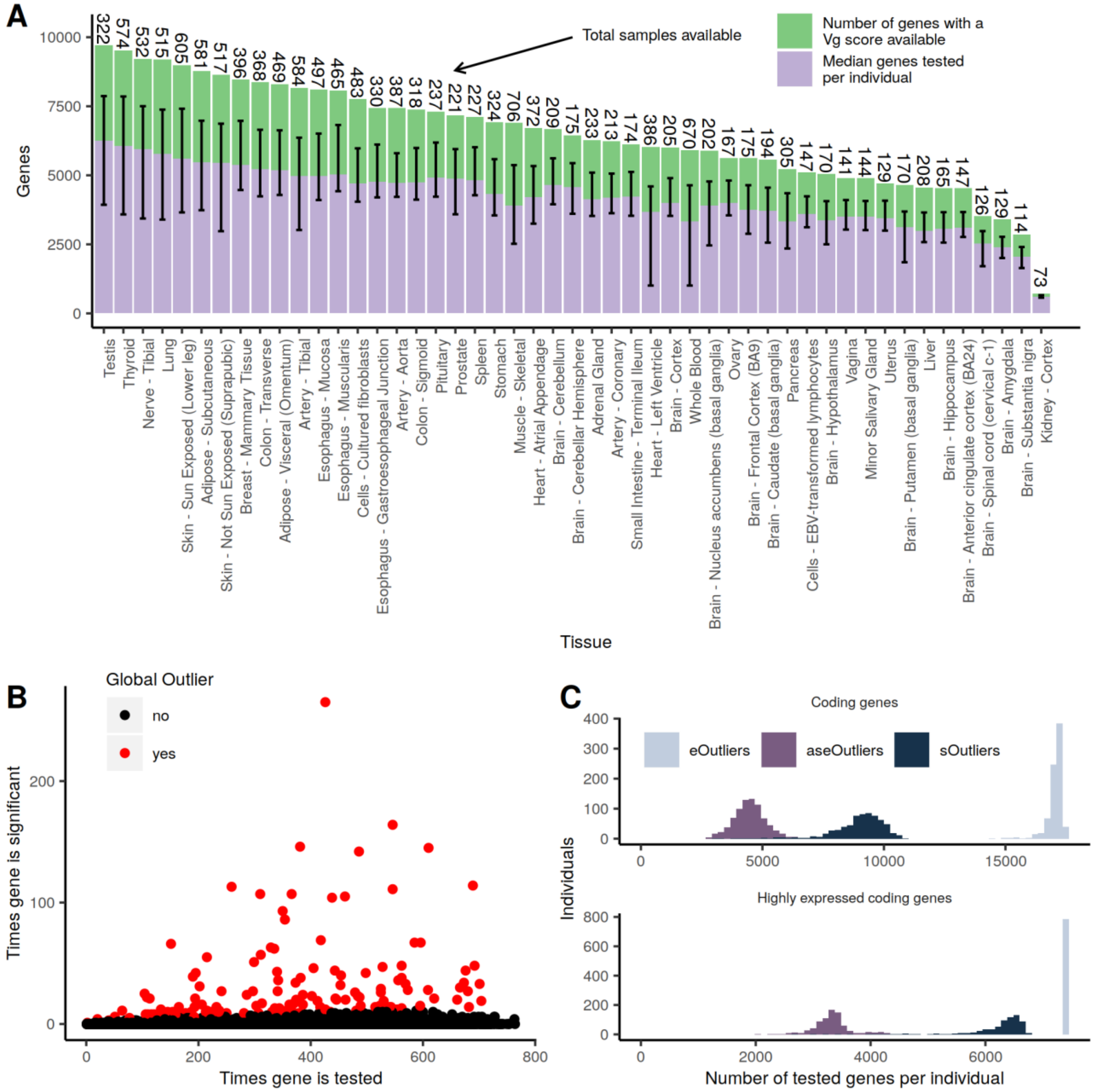


Figure S2. Quality control for ASE processing. (A) Average number of tests per individual tissue sample ± range. The total number of V^G scores available per tissue is shown above in green, with the total samples available per tissue. (B) The total number of times a gene was tested by considering its median ANEVA-DOT p-value vs the number of times it was called as an outlier. We call global outliers by drawing a 95% binomial confidence interval around the outlier frequency for each gene, and flagging all genes where the interval contains 1% or greater. Global outlier genes were removed from downstream analysis. (C) Distribution of median number of scores available across all three outlier methods, limiting to coding genes above, and coding genes with a median TPM > 10 across all individuals and tissues below.

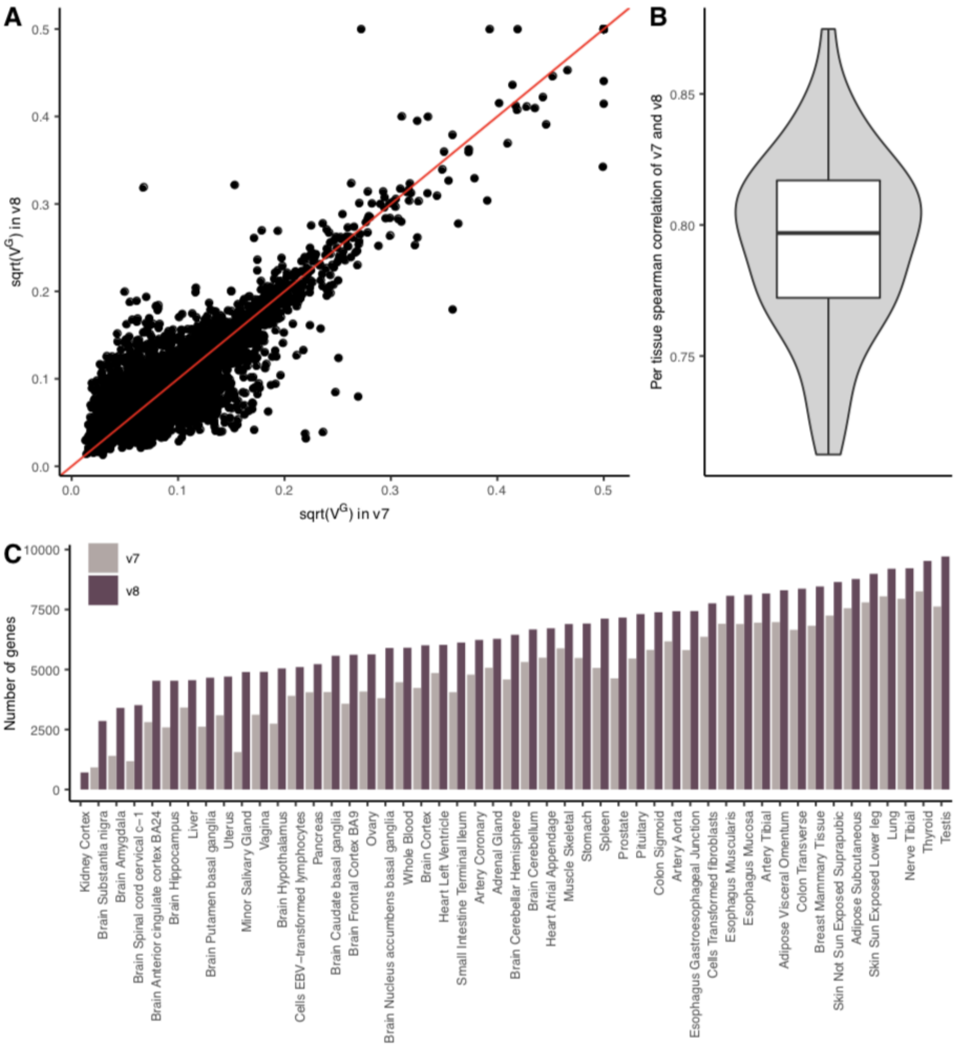


Figure S3. ANEVA estimates of genetic variance in gene expression (V^G). (A) Comparison of V^G estimates for an example tissue (Adipose subcutaneous) derived from GTEx v8 dataset compared to that of v7. The red line represents x=y. (B) Distribution of the spearman correlation coefficient between V^G estimates from v7 and v8 across all GTEx tissues. The lower and the upper whiskers indicate 1.5 interquartile range from the first and the third quartile, respectively. (C) The number of genes with V^G estimates available across GTEx tissues in each version.

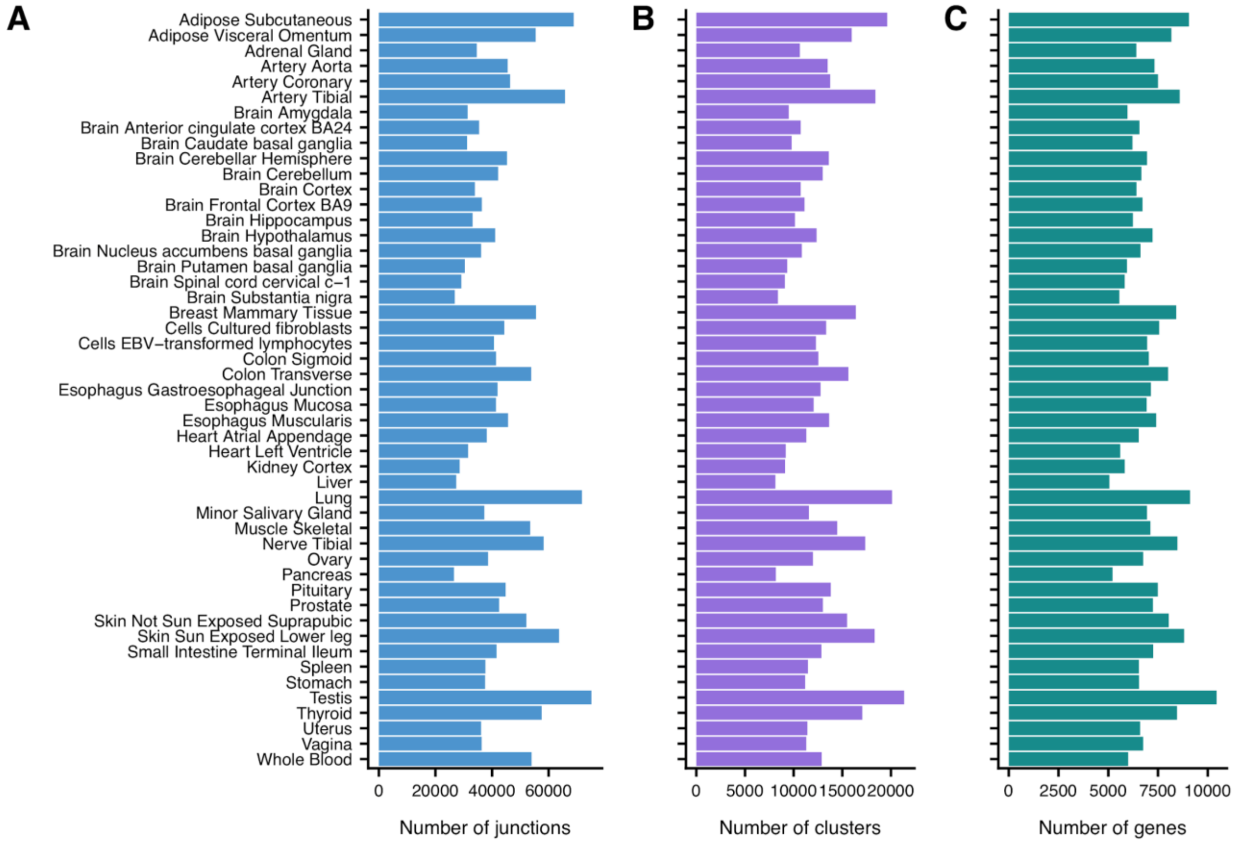


Figure S4. sOutlier split read count processing. The number of unique (A) junctions, (B) LeafCutter clusters, and (C) genes that are found in each tissue (rows) after split read count quantification and processing.

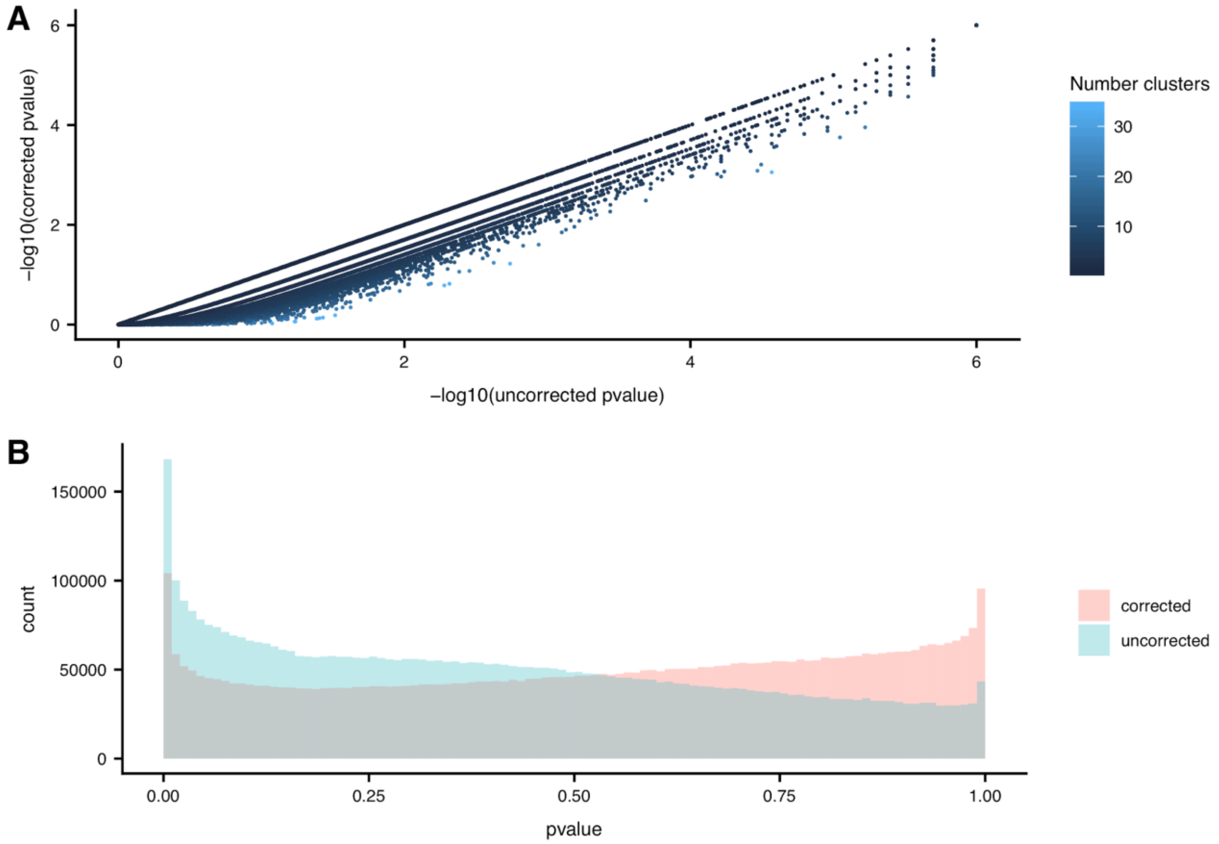


Figure S5. SPOT gene level correction. (A) Scatterplot showing the $-log_{10}$(sOutlier p-values+ 1x10^{-6}) in Muscle-Skeletal tissue at the gene level before the gene-level correction (x-axis) and after the gene level correction (y-axis) for the number of LeafCutter clusters mapped to each gene (color). (B) The distribution of sOutlier p-values in Muscle-Skeletal tissue at the gene level before the gene level correction (teal) and after the gene level correction (salmon) for the number of LeafCutter clusters mapped to each gene.

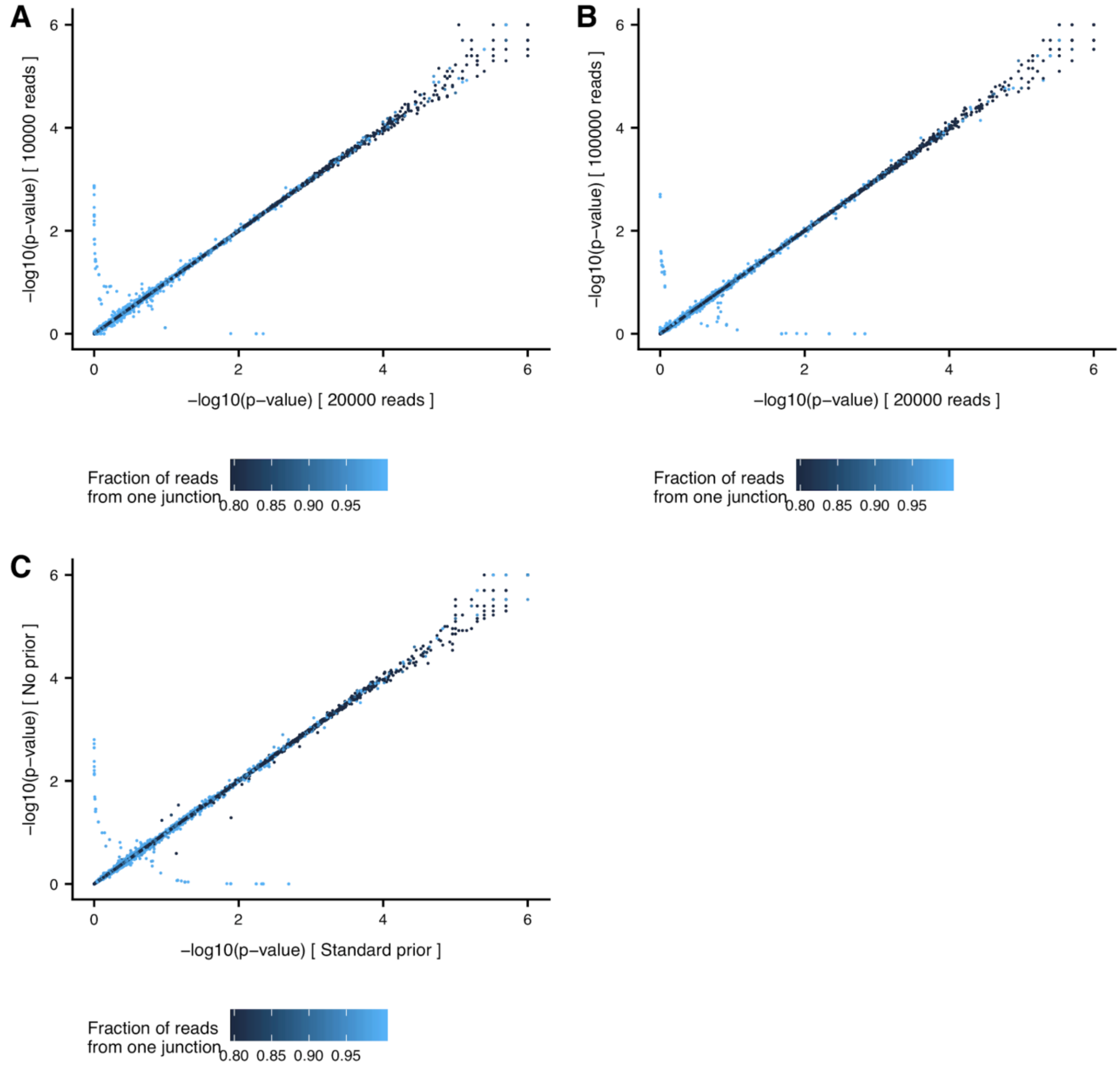


Figure S6. Robustness of SPOT to hyperparameter choice. Scatterplot showing the $-log_{10}$(sOutlier p-values + $1x10^{-6}$) of sample-LeafCutter cluster pairs in Muscle-Skeletal tissue from default implementation of SPOT (x-axis) compared to implementations of SPOT using different hyperparameter settings (y-axis; A, B, C) colored by the maximum fraction of reads mapping to a single junction (summed across samples) in the corresponding LeafCutter cluster. Any cluster with a maximum fraction of reads mapping to a single junction that is less than or equal to 80% is colored identically to better highlight differences above 80%. (A, B) Comparison of sOutlier p-values from the default implementation of SPOT (x-axis) and an implementation of SPOT where random samples used to generate the empirical distribution have 10,000 (A) and 100,000 (B) reads mapped to the cluster. (y-axis). (C) Comparison of sOutlier p-values from the default implementation of SPOT (x-axis) and an implementation of SPOT where there is no Gamma prior placed on α_j (y-axis).

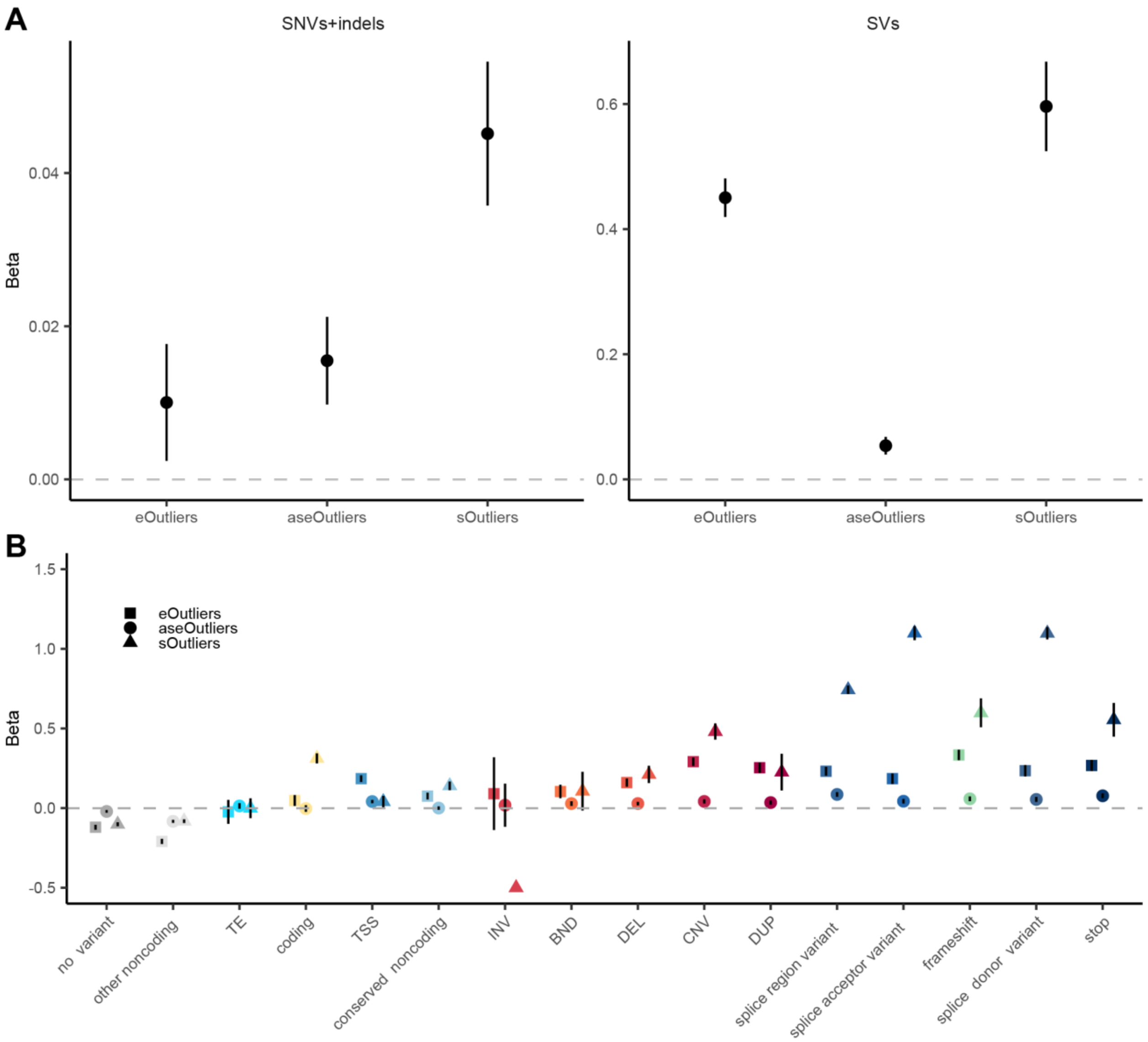


Figure S7. Association of rare variant status and continuous outlier measure. (A) Across each outlier type, the beta coefficient estimate and 95% confidence interval (y-axis) from a linear model of binary rare variant status as the outcome and continuous outlier measure, defined as the -log10(median p-value), as the predictor. Outcome is 1 if the gene has a nearby SNV or indel that is not found in gnomAD, or for SVs if it is a singleton variant within GTEx. (B) Beta coefficient estimates from similar models as in (A) but considering rare variant status across a range of categories (x-axis).

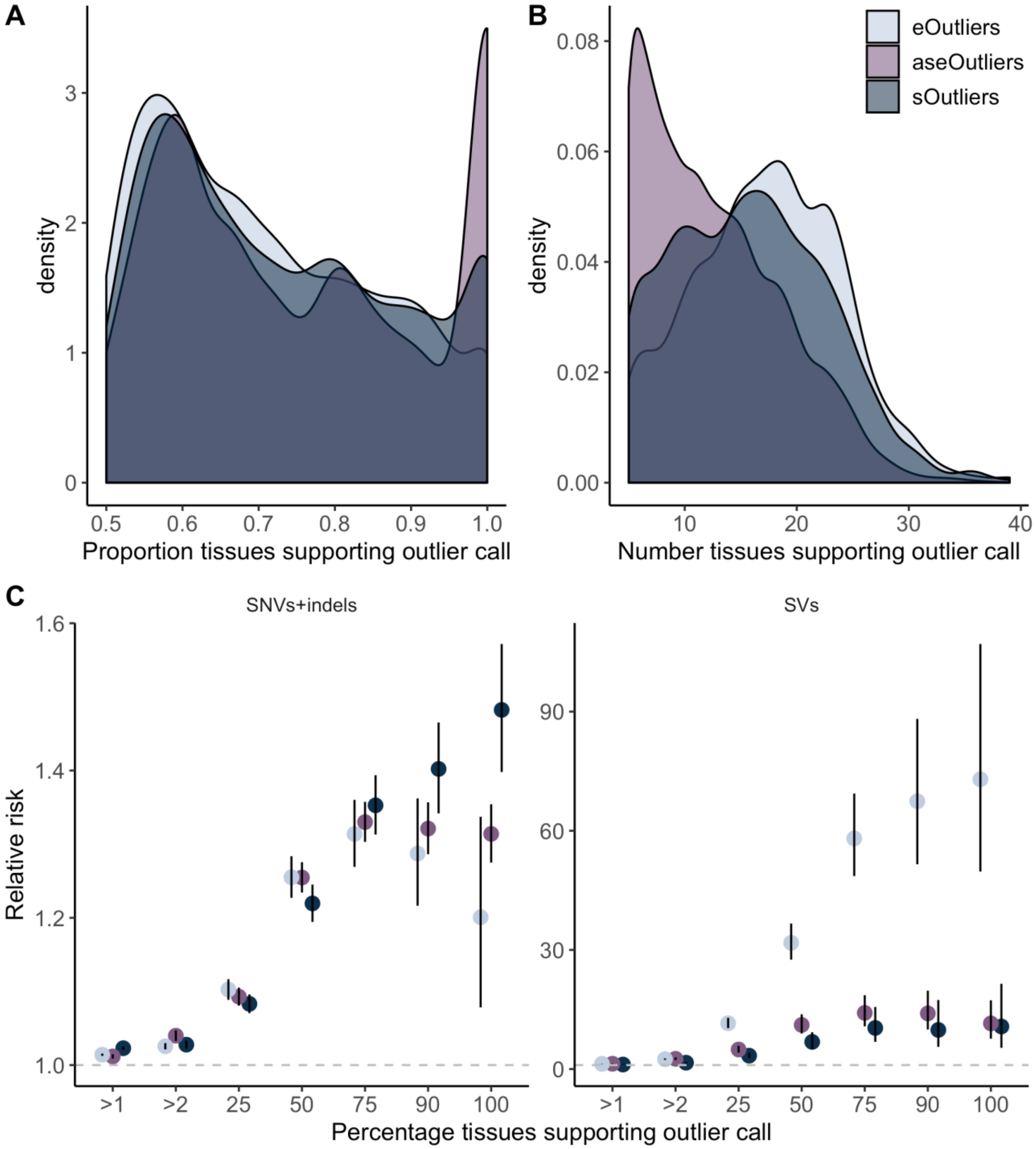


Figure S8. Number of tissues supporting outlier calls. (A) For all multi-tissue outlier calls, the proportion of tested tissues with outlier signal at the same threshold (p-value < 0.0027 or |Z| > 3). (B) For all multi-tissue outlier calls, the number of tested tissues with outlier signal at the same threshold (p-value < 0.0027 or |Z| > 3), restricted to individuals with data from at least 5 tissues. (C) The impact of the number of tissues supporting the outlier call on the relative risk of outliers having a rare variant (MAF < 1%) within 10kb. For the >1 and >2 bins, this refers to >1 or > 2 tissues, while the remaining bins are percentages of the total number tested. For SVs, sOutlier enrichments stop at the 50% bin due to small numbers at later bins.

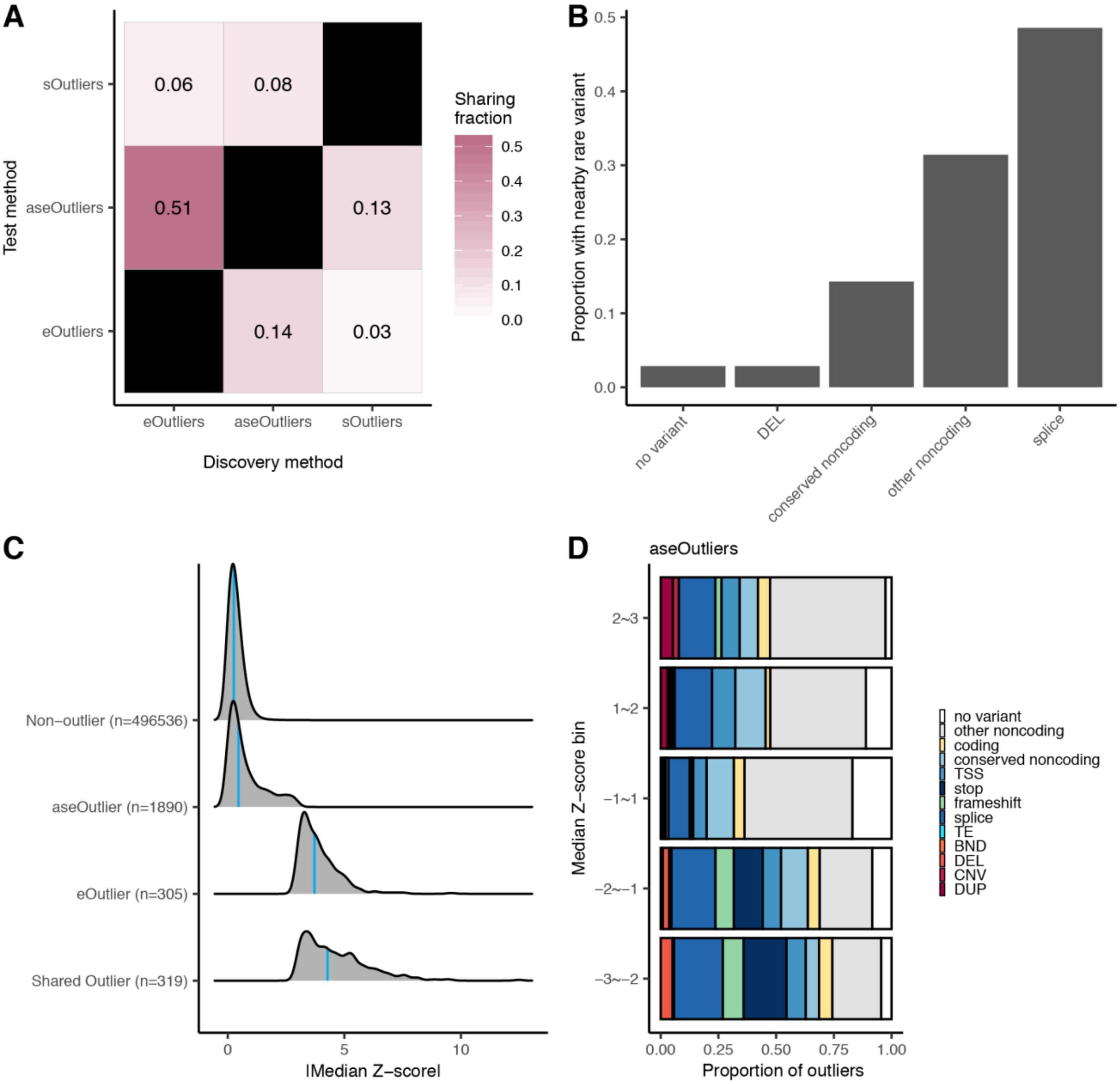


Figure S9. Comparing outliers across methods. (A) Of the set of individuals and genes tested across all data types, the fraction discovered via one method that also meet the outlier thresholds (p < 0.0027) in another method. Across all data types, 624 individuals and 8,722 genes, including 2,281,262 unique combinations, were tested by all methods. (B) The proportion of outliers shared across all methods assigned to the given rare variant category nearby the outlier gene. Of the 2,209 aseOutliers, 1,385 sOutliers, and 624 eOutliers discovered at this threshold among the shared set, 35 individual-gene pairs are found by all three methods, encompassing 31 unique genes. (C) Of the set of eOutliers and aseOutliers within this set, the distribution of |median Z-scores| for outliers in both types, expression alone, ASE alone, or non-outliers for the same set of genes. Blue lines represent the 50th percentile. (D) The proportion of aseOutliers with a nearby rare variant of a given type split by the corresponding median Z-score bin for the same individual-gene pair.

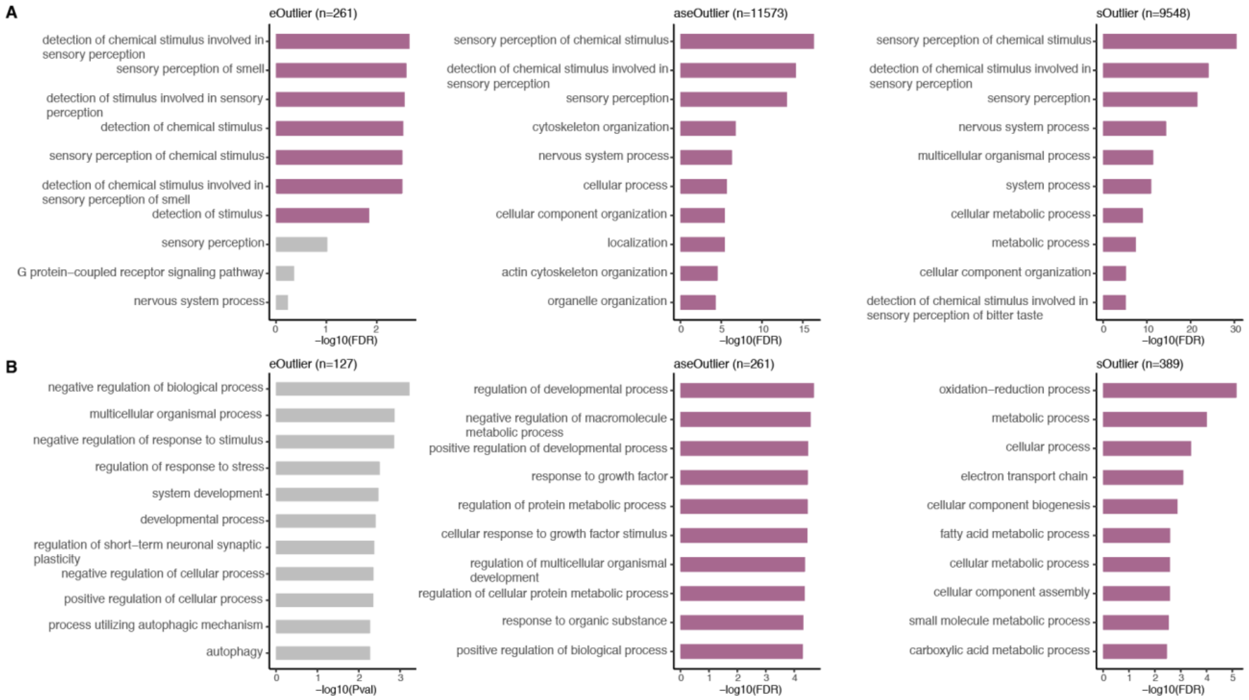


Figure S10. Gene ontology term enrichments for outlier and non-outlier genes. The top ten Gene Ontology (GO) terms enriched, by -log10(FDR-corrected p-value) on the x-axis, in the set of genes with no outliers in any tissue (A) and those associated with the most extreme outliers (B). Results are included for eOutliers on the left, aseOutliers in the center and sOutliers on the right, with the number of included genes at the top of each plot. Pink bars are significant at an FDR-corrected p-value threshold of 0.05, while the gray bars are not significant. For eOutliers in (B), all terms had an FDR corrected p-value of 1, and so nominal p-values are presented instead.

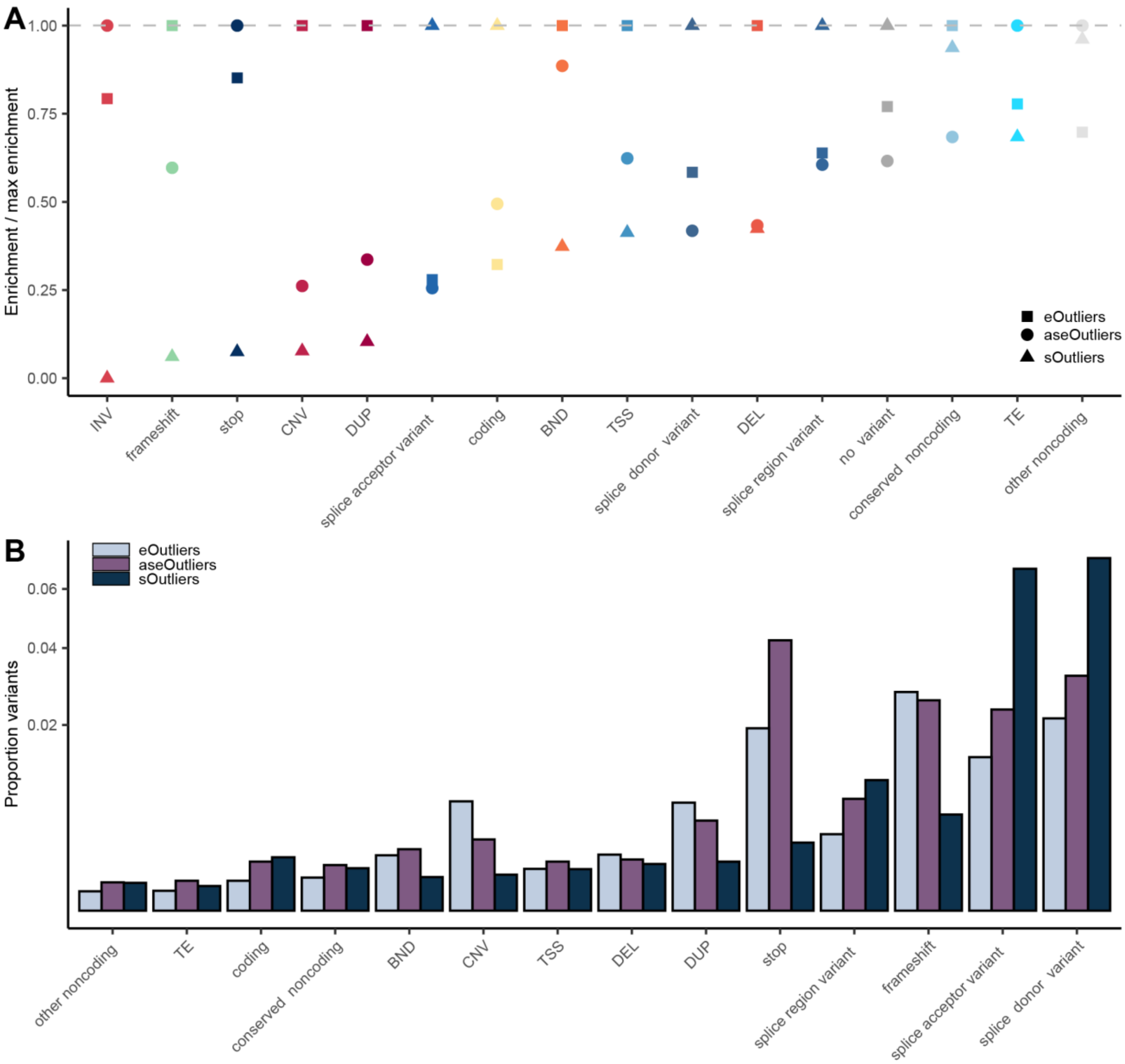


Figure S11. Comparison of variant class enrichments across methods. (A) For each variant category, the relative risk enrichment for each outlier type over the maximum enrichment for that category. (B) For each variant category, the proportion of variant occurrences leading to an outlier across all categories, with INV removed due to either very low or zero instances. Those marked ns indicate that in 1000 iterations permuting outlier status, a proportion greater than or equal to the actual proportion was found greater than 5% of the time. TSS = transcription start site, TE = transposable element, INV = inversion, BND = breakend, DEL = deletion, CNV = copy number variation, DUP = duplication.

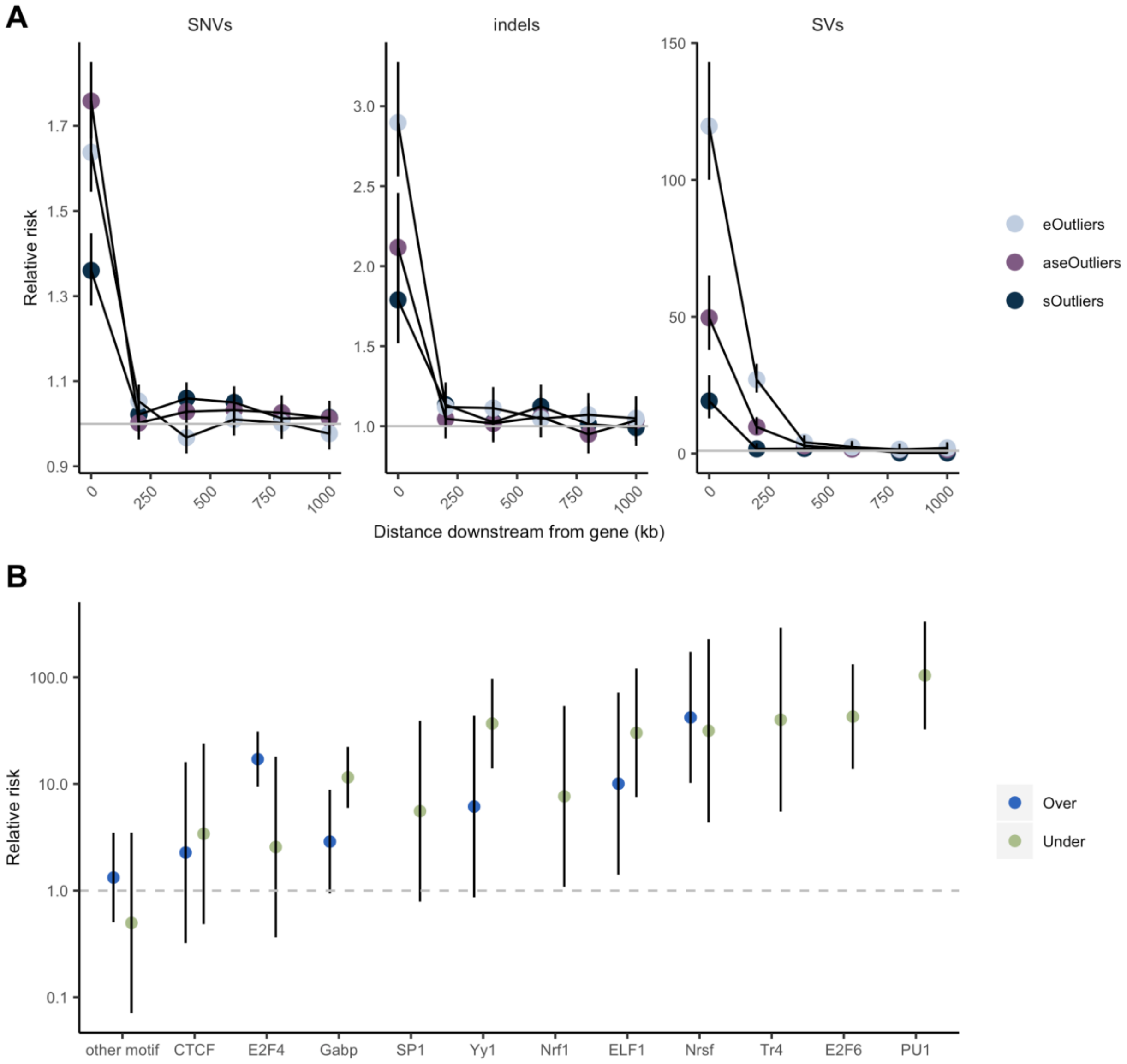


Figure S12. Rare variant enrichments at distances downstream of outlier genes and in promoter motifs. (A) Relative risk of singleton SNVs, indels, and SVs at varying distances downstream of outlier genes (bins exclusive) across data types. (B) Relative risk of rare (MAF < 1%) variants interrupting promoter motifs nearby over eOutliers (blue) or under eOutliers (green) relative to controls. For data points not included for one direction, there were not enough instances of rare variants overlapping a given motif near outliers to estimate risk.

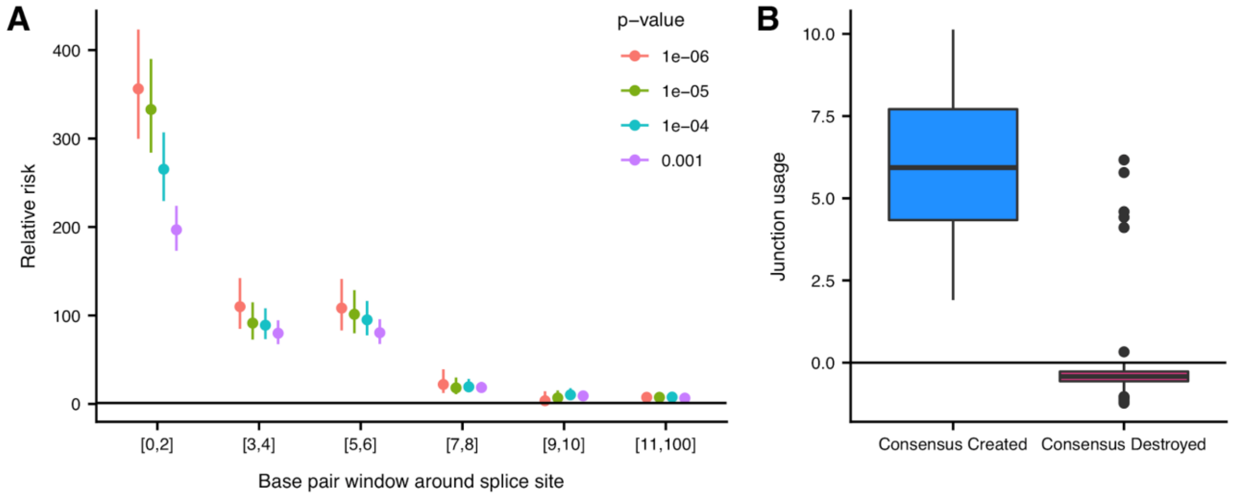


Figure S13. Enrichment of rare variants nearby splice sites in sOutliers. (A) Relative risk (y-axis) of rare variants within various window sizes around splice sites (x-axis) for sOutlier LeafCutter clusters relative to non-outlier clusters at several median LeafCutter cluster p-value thresholds (color). (B) Junction usage of a splice site is the natural log of the fraction of reads in a LeafCutter cluster mapping to the splice site of interest in sOutlier (median LeafCutter cluster p-value < 1 x 10^{-5}) samples relative to the fraction in non-outliers samples aggregated across tissues by taking the median. Junction usage (y-axis) of the closest splice sites to rare variants that lie within the splicing consensus sequence binned by the type of variant (x-axis).

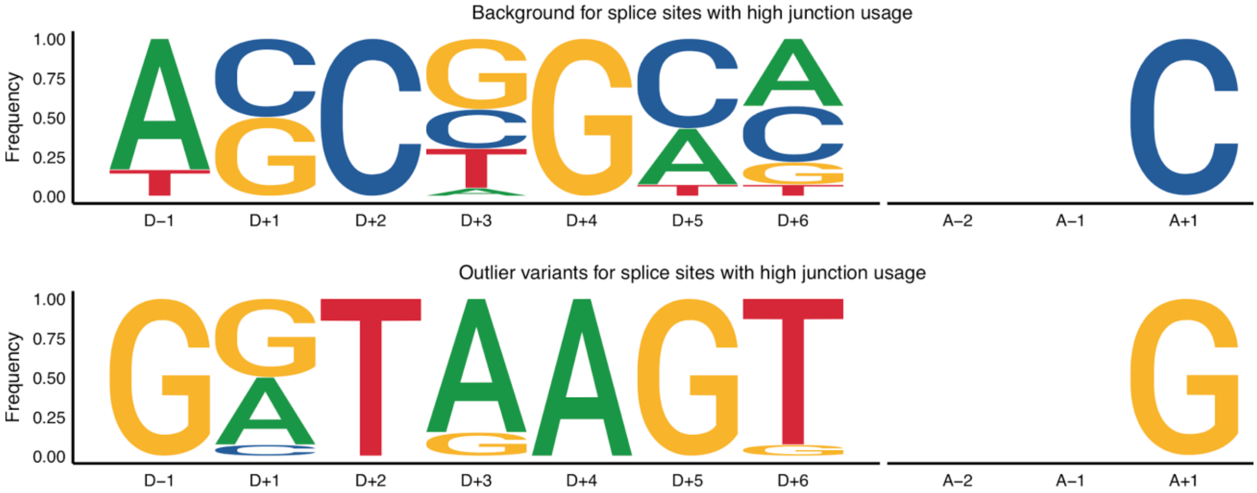


Figure S14. sOutlier variants in consensus sequence of splice sites with high junction usage. Independent position weight matrices showing mutation spectrums of sOutlier (median LeafCutter cluster p-value < 1 x 10^{-5}) rare variants at positions relative to splice sites with positive junction usage (ie. splice sites used more in outlier individuals than in non-outliers).

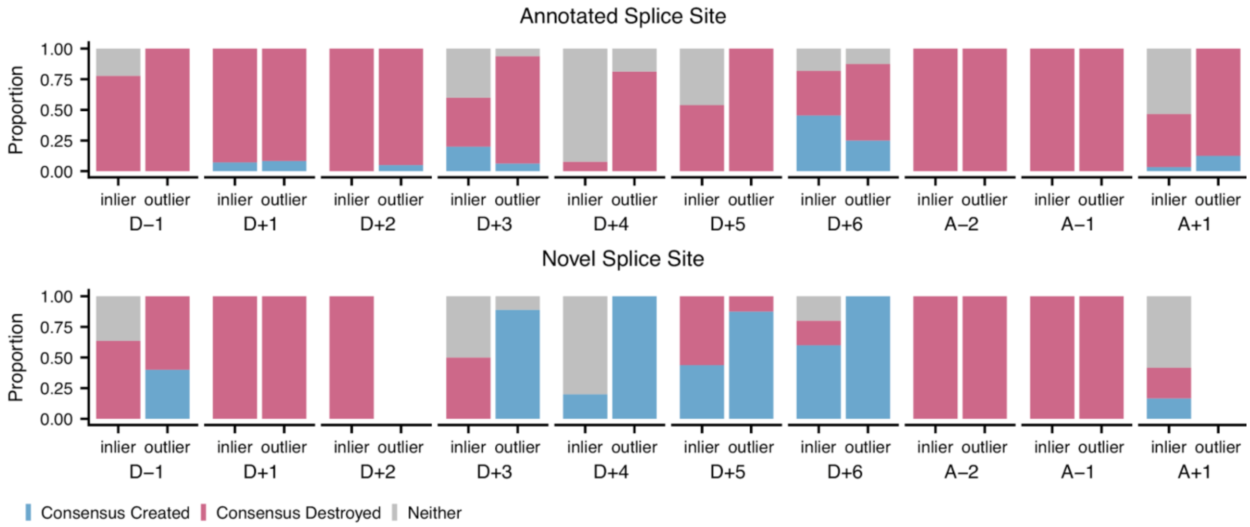


Figure S15. sOutlier variants in consensus sequence of annotated and novel splice sites. Proportion of sOutlier (median LeafCutter cluster p-value < 1 x 10^{-5}) and non-outlier variants, at each position in the splicing consensus sequence, that create the consensus sequence (blue) or destroy the consensus sequence (red) where variants are binned by whether the nearby splice site is annotated or novel (rows).

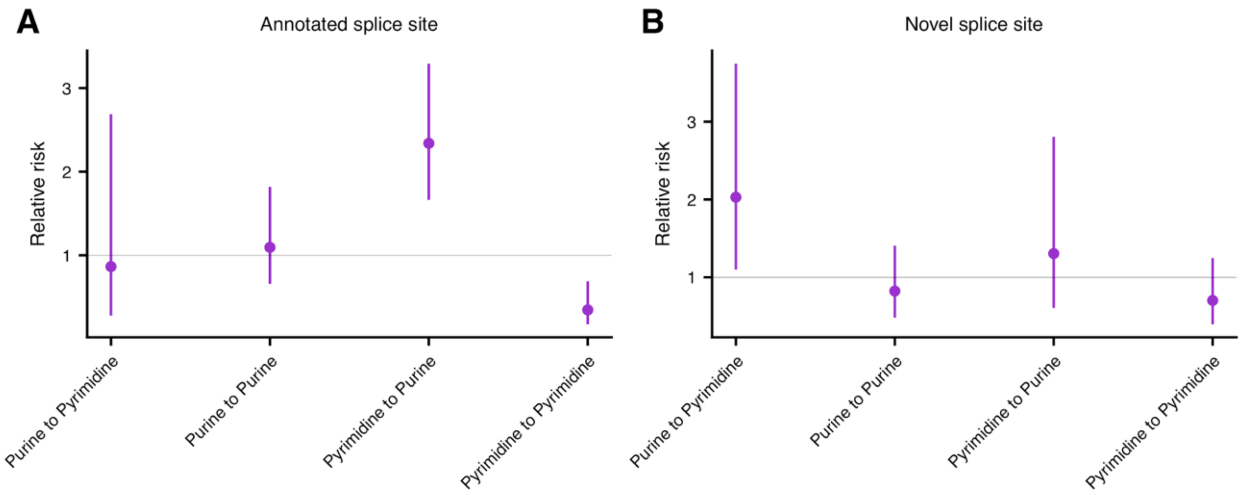


Figure S16. sOutlier variant type enrichments in PPT. Relative risk for sOutliers relative to non-outliers (median LeafCutter cluster p-value < 1 x 10^{-5}) of having a rare variant that is located in PPT (5 to 35 base pairs upstream from an acceptor splice site) having a specific mutation spectrum (x-axis). Relative risk calculation done separately for annotated (A) or novel (B) splice sites.

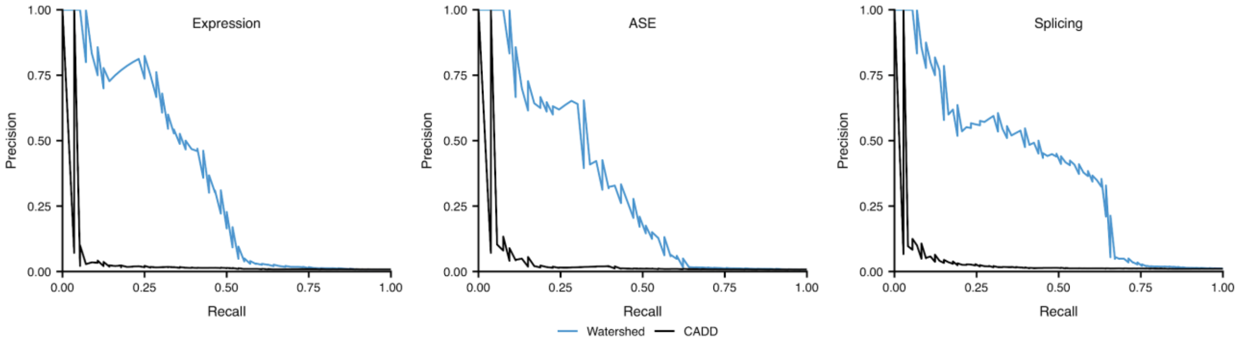


Figure S17. Precision recall curves for Watershed and CADD. Precision-recall curves comparing performance of Watershed and CADD (colors) using held out pairs of individuals for all three median outlier signals.

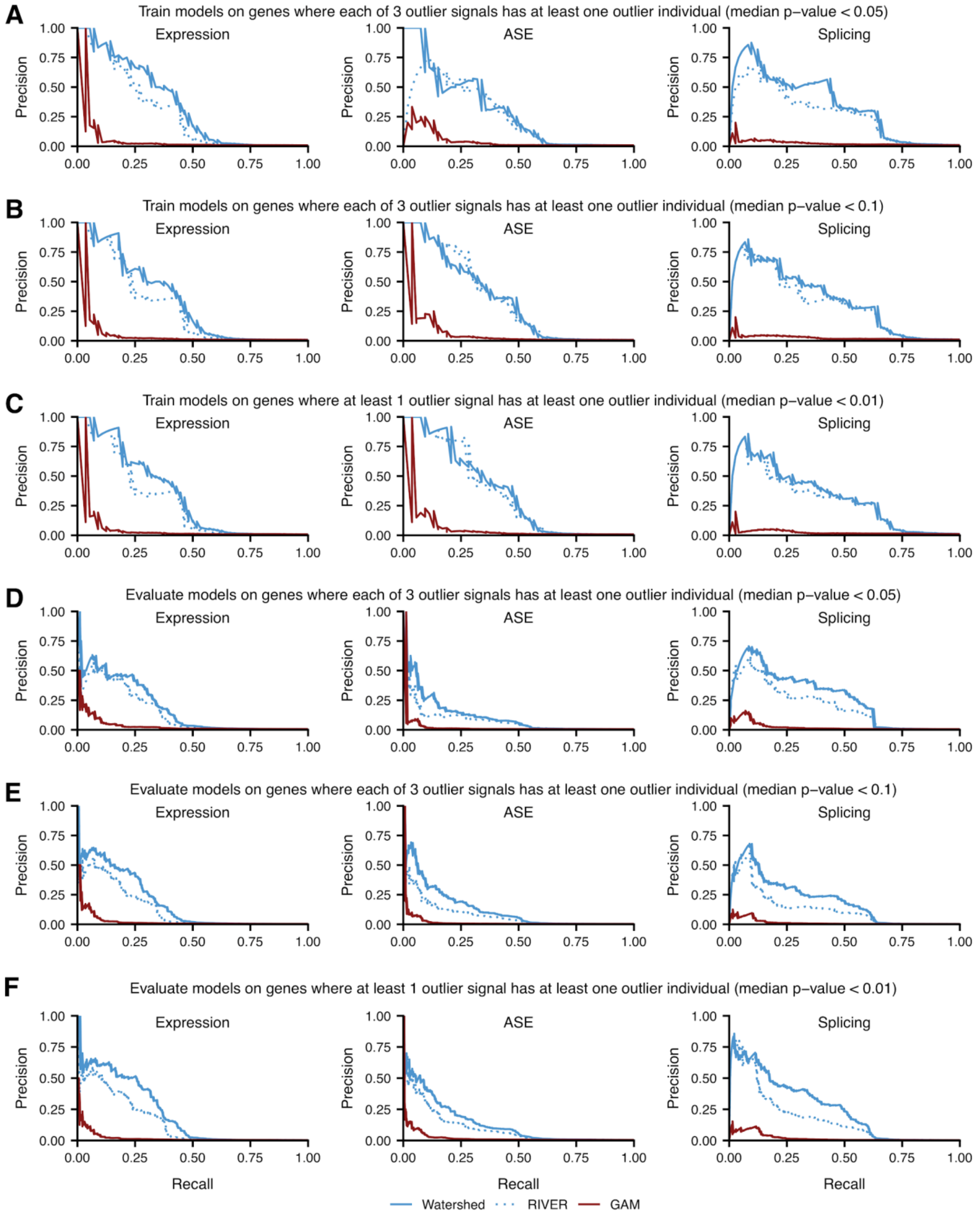


Figure S18. Watershed precision recall curves with different training or evaluation data. Precision-recall curves comparing performance of Watershed, RIVER, and GAM (colors) using held out pairs of individuals for three median outlier signals (columns) when models were trained with different training data sets (A, B, C; see Supplementary methods) or when models were evaluated with different held out pairs of individuals (evaluation data; D, E, F; see supplementary methods). Training data for Watershed,

RIVER, and GAM filtered to only include genes where (A) all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.05), (B) all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.1), (C) at least 1 outlier signal has at least one individual that is an outlier (median p-value < 0.01). Held out pairs of individuals (evaluation data) used in A, B, C were the same held out pairs of individuals used to generate precision-recall curves in Fig 4D. Held out pairs of individuals used to evaluate Watershed, RIVER, and GAM filtered to only include genes where (D) all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.05), (E) all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.1), (F) at least 1 outlier signal has at least one individual that is an outlier (median p-value < 0.01). Training data used to train models composing D, E, F was the same training data used to generate models underlying precision-recall curves in Figure 5-3D.

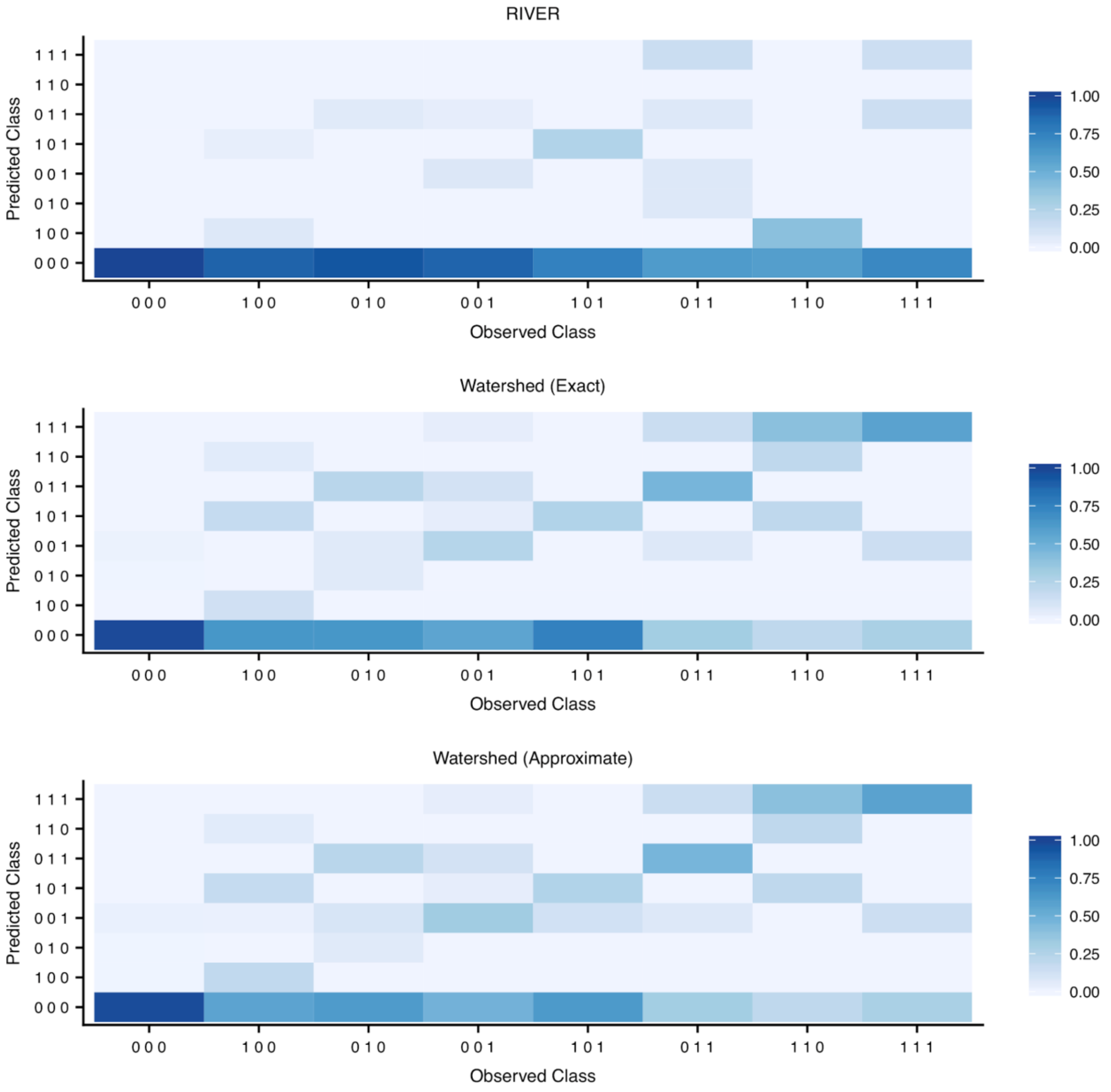


Figure S19. Watershed confusion matrices. Confusion matrices comparing performance of RIVER (top), Watershed with parameters optimized via exact inference (middle), and Watershed with parameters optimized via approximate inference (bottom) in jointly predicting outlier status of all three outlier signals (class) using held out pairs of individuals. The first element of the binary class abbreviations represents median splicing outlier status, the second element of the class abbreviations represents median expression outlier status, and the third element of the class abbreviations represents ASE outlier status. An observed class of "1 0 1" therefor corresponds to a sample that is an outlier for splicing and ASE, but not expression. The predicted class of a sample is the class (out of the 8 classes) that has the largest posterior probability. Columns in each heatmap are normalized to sum to one.

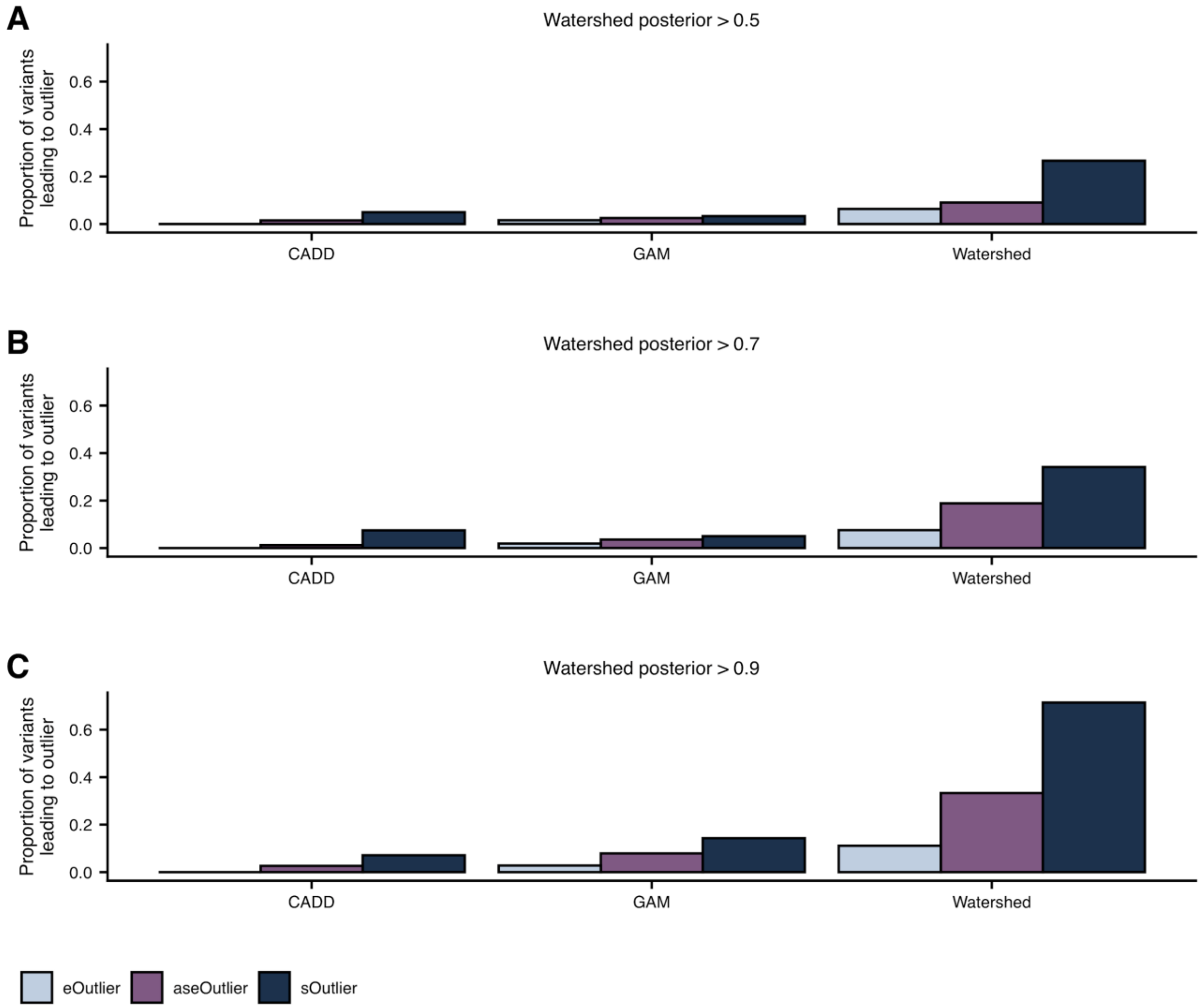


Figure S20. Prioritization of variants that lead to outliers with Watershed. The proportion of rare variants, with Watershed posterior probability greater than 0.5 (A), 0.7 (B), 0.9 (C) (right), with GAM probability greater than a threshold set to match the number of Watershed variants for each outlier signal (center), and with CADD score greater than a threshold set to match the number of Watershed variants for each outlier signal (left), that lead to an outlier at a median p-value threshold of 0.0027 across three outlier signals (colors). Watershed, GAM, and CADD models evaluated on held-out pairs of individuals.

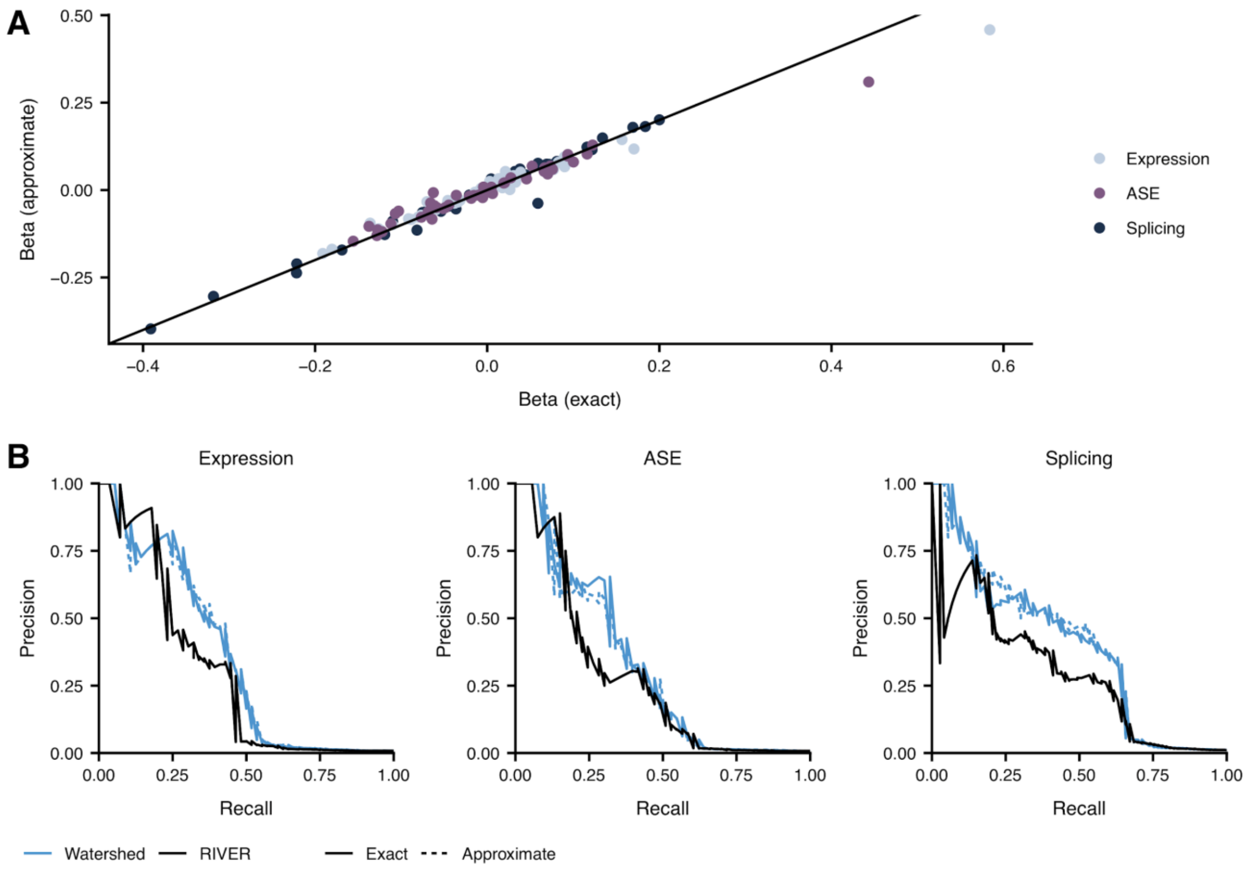


Figure S21. Comparison of exact and approximate inference in Watershed. (A) Scatterplot comparing Watershed (applied to median ASE, splicing, and expression outlier signals) genomic annotation coefficients (β) when model was optimized using exact inference (x-axis) compared to when model was optimized using approximate inference (y-axis) colored by which outlier signal the coefficient predicted. (B) Precision-recall curves comparing performance of RIVER, Watershed optimized via exact inference, and Watershed optimized via approximate inference (colors) using held out pairs of individuals for all three median outlier signals.

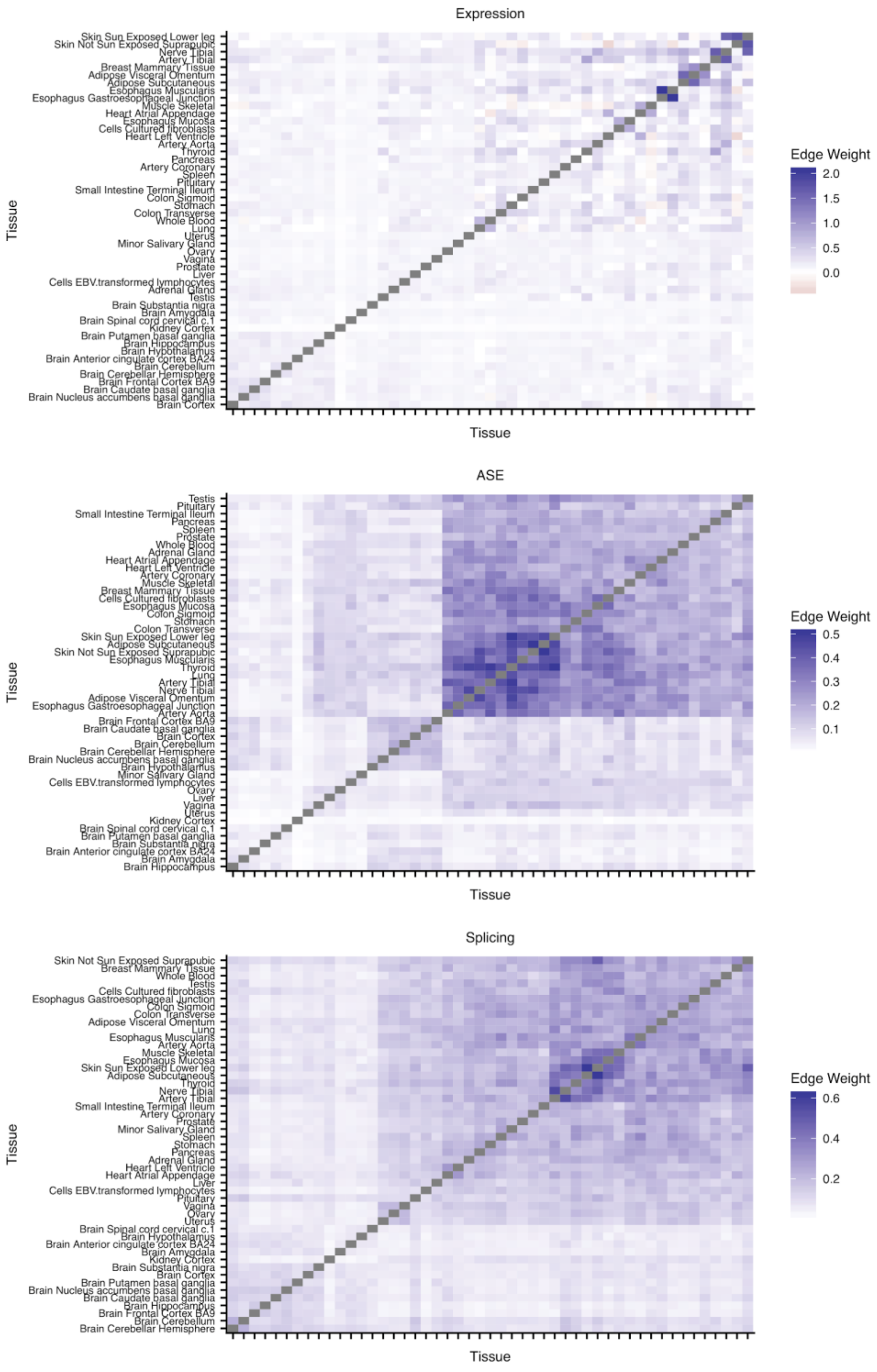


Figure S22. Tissue-Watershed edge weights. Learned tissue-Watershed edge weights (θ) between pairs of tissue- outlier signals after training tissue-Watershed on expression (top), ASE (middle), and splicing (bottom) outliers across single tissues.

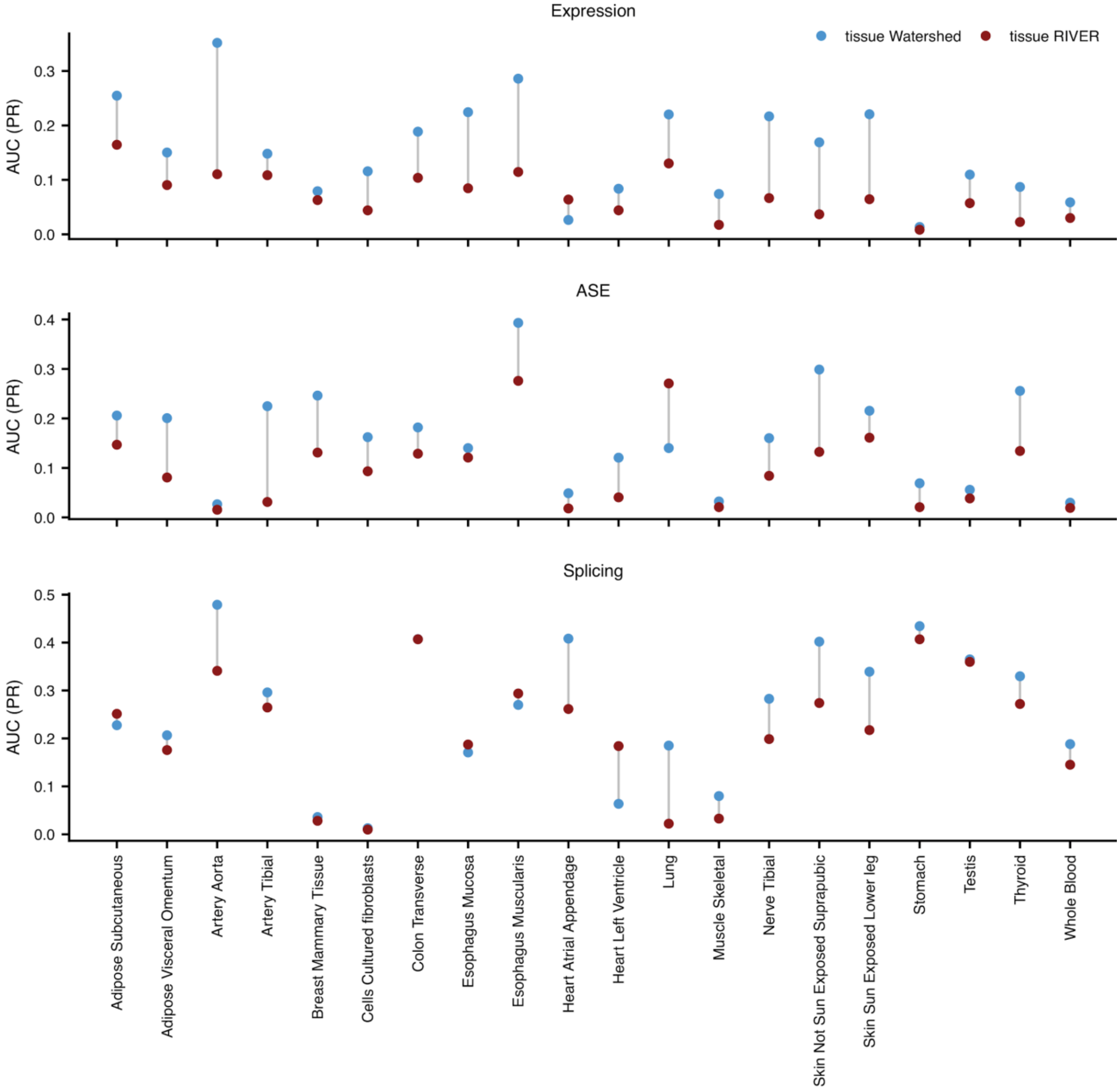


Figure S23. Area under precision recall curves in single tissues. Area under precision recall curves (AUC (PR); y-axis) in a single tissue (x-axis) for tissue-Watershed (blue) and tissue-RIVER (red) when applied outliers across single tissues for all 3 outlier types (rows). Precision recall curves in each tissue generated using held out pairs of individuals where both individuals share the same rare variant and have observed outlier signal for the gene of interest. We limit to tissues that have at least 5 held out pairs of individuals that have outlier labels in ASE, splicing, and expression.

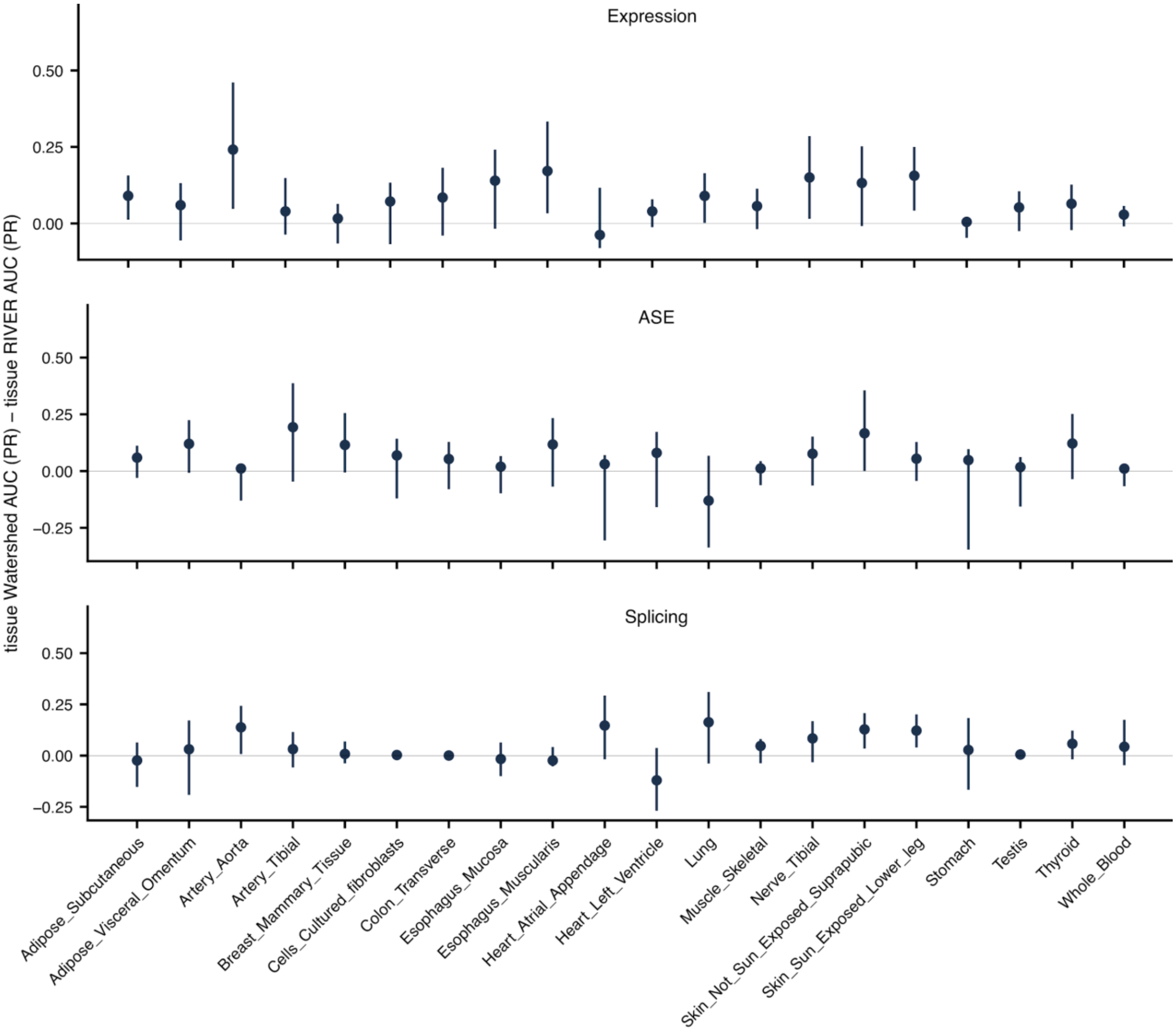


Figure S24. Difference in area under precision recall curves in single tissues. Difference in the area under the precision recall curves between tissue-Watershed and tissue-RIVER (y-axis) in a single tissue (x-axis), shown for expression, ASE, and splicing outlier signals (rows). Precision recall curves in each tissue generated using held out pairs of individuals where both individuals share the same rare variant and have observed outlier signal for the gene of interest. We limit to tissues that have at least 5 held out pairs of individuals that have outlier labels in ASE, splicing, and expression. Error bars (95% confidence interval) on these statistics generated using non-parametric bootstrapping with 20,000 bootstrapped samples (see Supplementary methods).

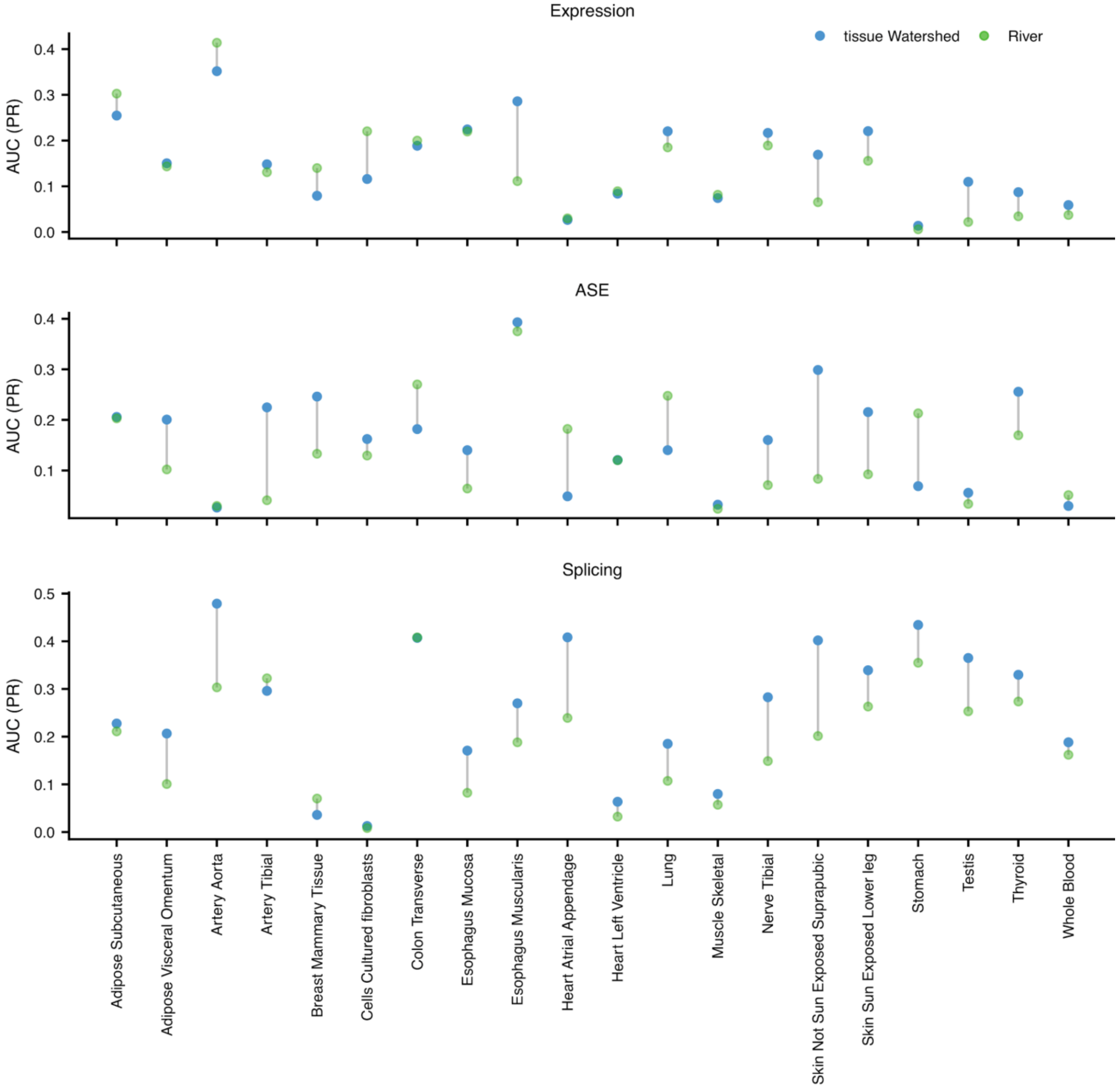


Figure S25. Area under precision recall curves in single tissues. Area under precision recall curves evaluated on outlier calls in a single tissue (x-axis) for each of the three outlier types (rows) based on a tissue-Watershed model trained across single tissues (blue) and a RIVER model trained on the median outlier signal (green). Precision recall curves in each tissue generated using held out pairs of individuals where both individuals share the same rare variant and have observed outlier signal for the gene of interest. We limit to tissues that have at least 5 held out pairs of individuals that have outlier labels in ASE, splicing, and expression.

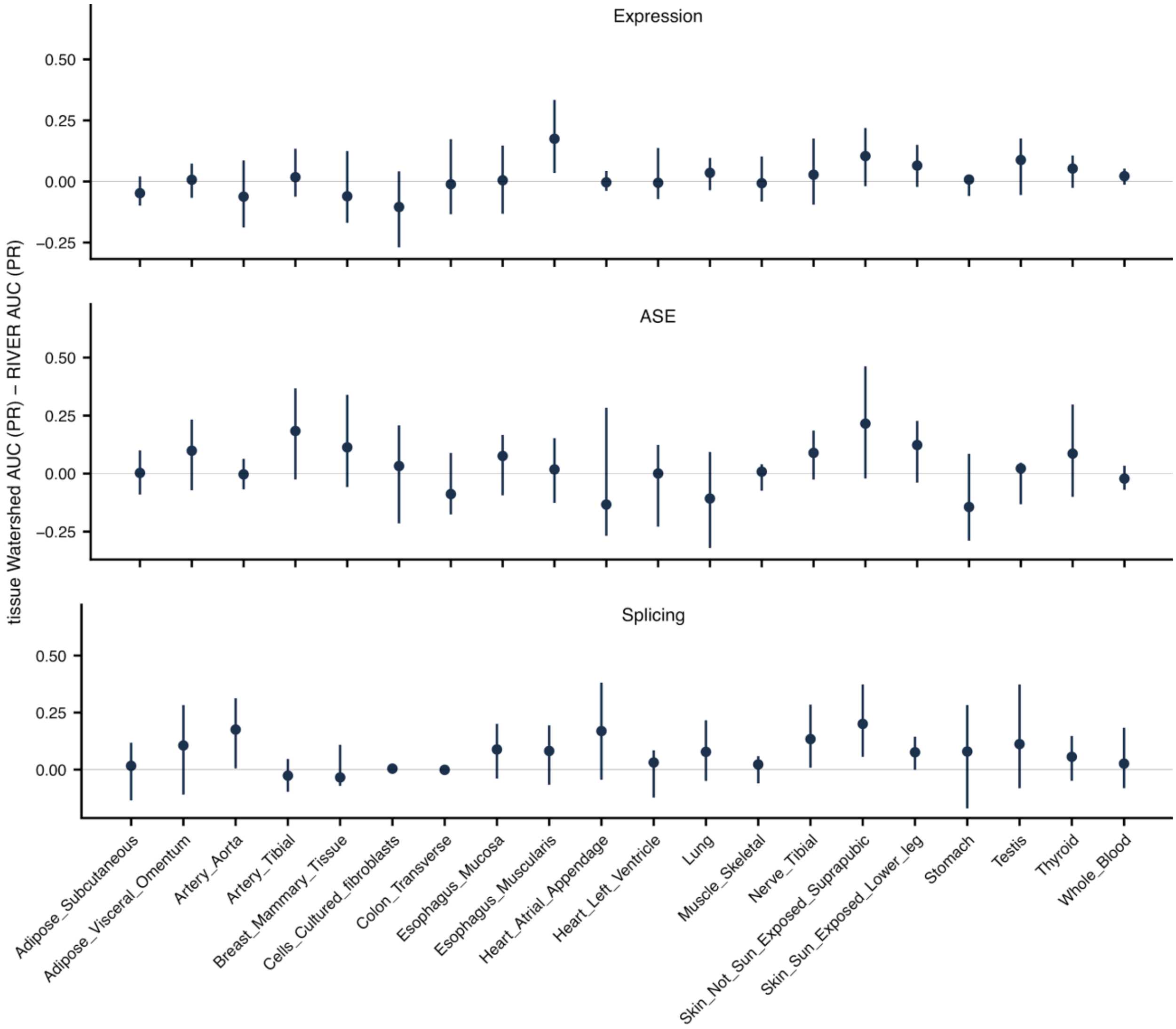


Figure S26. Difference in area under precision recall curves in single tissues. Difference in the area under the precision recall curves between tissue-Watershed and a RIVER model trained on the median outlier signal (y-axis) in a single tissue (x-axis), shown for expression, ASE, and splicing outlier signals (rows). Precision recall curves in each tissue generated using held out pairs of individuals where both individuals share the same rare variant and have observed outlier signal for the gene of interest. We limit to tissues that have at least 5 held out pairs of individuals that have outlier labels in ASE, splicing, and expression. Error bars (95% confidence interval) on these statistics generated using non-parametric bootstrapping with 20,000 bootstrapped samples (see methods).

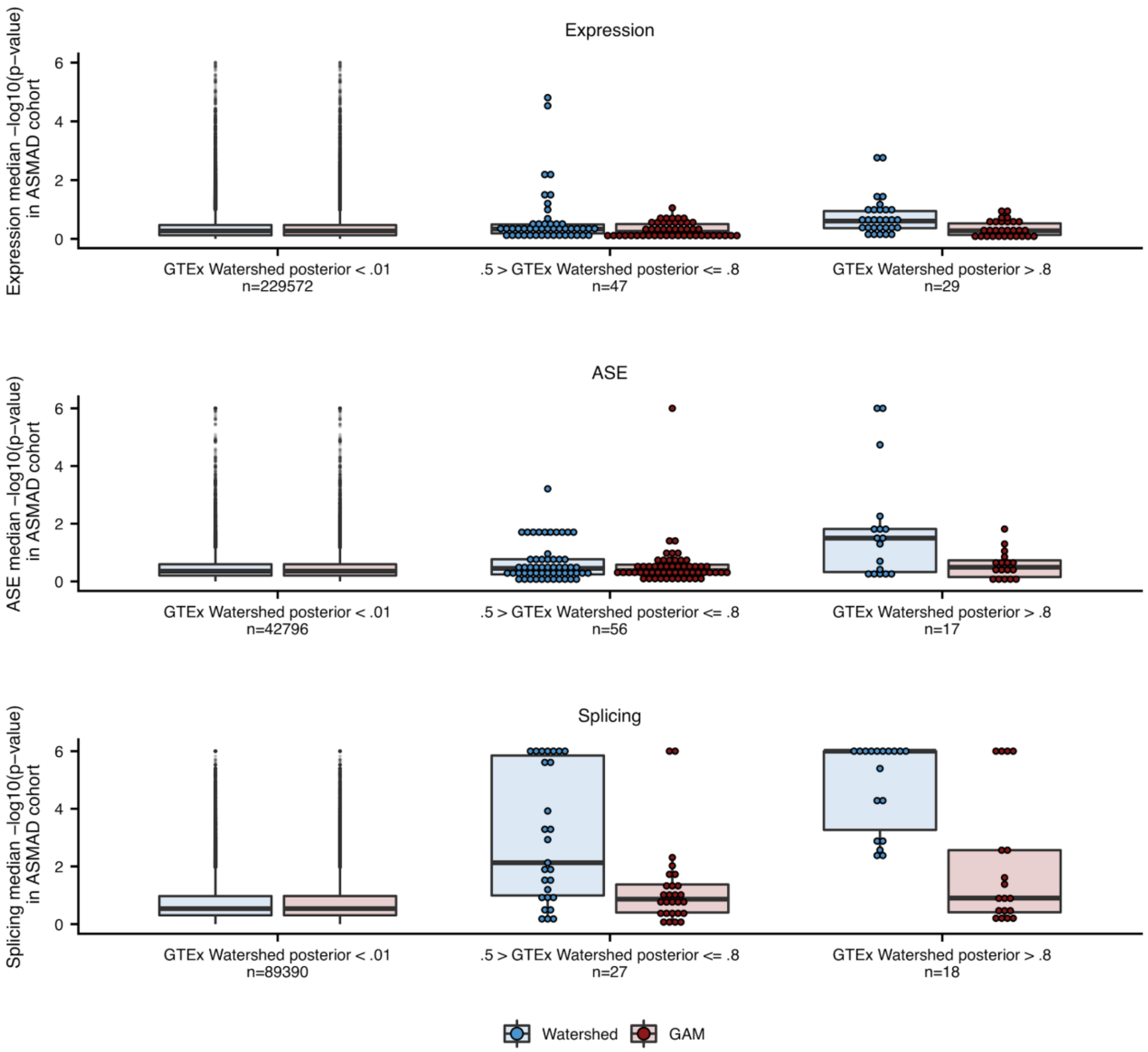


Figure S27. Replication in ASMAD cohort. Expression, ASE, and splicing outlier -log$_{10}$(p-value + 1x10^{-6}) in ASMAD cohort of genes nearby rare variants binned by: GTEx Watershed posterior probability in the corresponding outlier type (blue), and GTEx GAM posterior probability in the corresponding outlier type greater than a threshold set to match the number Watershed variants in the corresponding bin (red). This analysis is limited to GTEx rare variants present in the ASMAD cohort. The number of variant-gene pairs in each bin (n) is shown beneath the posterior threshold labels on the x-axis. If multiple GTEx individuals have the same rare variant, we report the median posterior probability across individuals. If multiple ASMAD individuals have the same rare variant, we report the median p-value across individuals. There are 10 variant-gene pairs in the GTEx Watershed posterior > .8 bin that have ASMAD splicing outlier p-value exactly equal to 0 (or equivalently -log$_{10}$(p-value + 1x10^{-6}) equal to 6). This p-value point mass at 0 is a result of SPOT calculating p-values from an empirical distribution.

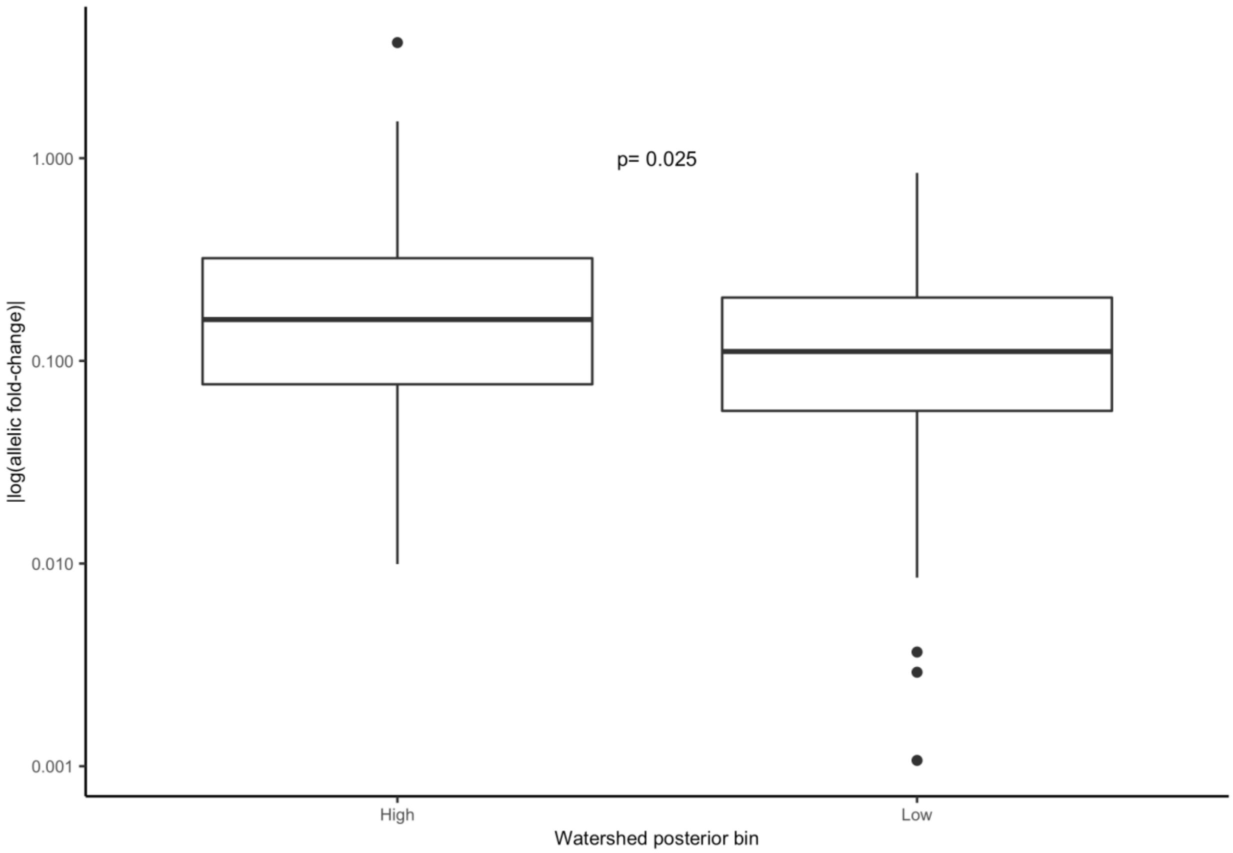


Figure S28. MPRA results. For 52 high Watershed expression (score >= 0.5) rare variants and 98 low Watershed expression (score < 0.5) variants nearby 62 eOutlier genes, the log fold-change in expression between the reference and edited alleles. p-value for the difference between Watershed bins is calculated from a one-sided Wilcoxon rank sum test.

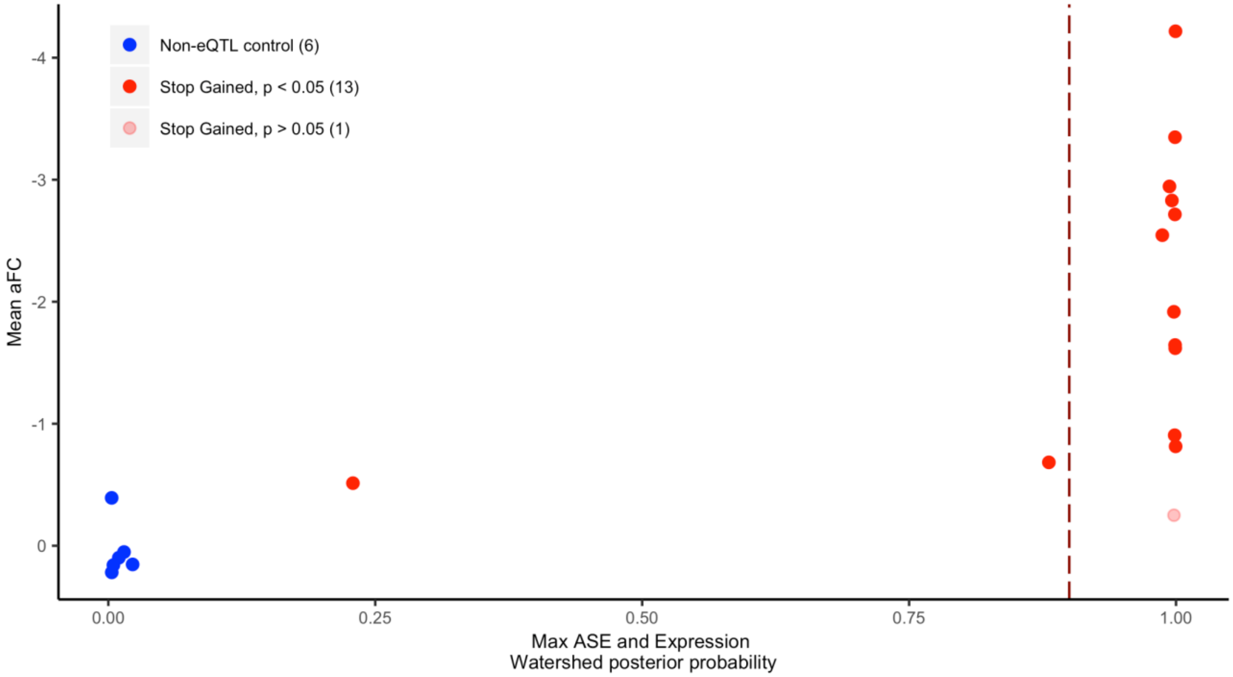


Figure S29. Experimental validation by editing 20 variants into inducible-Cas9 293T cell lines. 14 stop-gained variants were edited into cell lines, and their effect was evaluated using allelic fold change (aFC), shown on the y-axis, with the variant's maximum of ASE or expression Watershed score along the x-axis. When compared to negative control variants, 13 of the 14 edited variants caused significant aFC of their target genes (dark red). Non-eQTL control variants shown here are the 6 with Watershed scores available out of the 30 edited in total, and are not expected to induce an aFC effect.

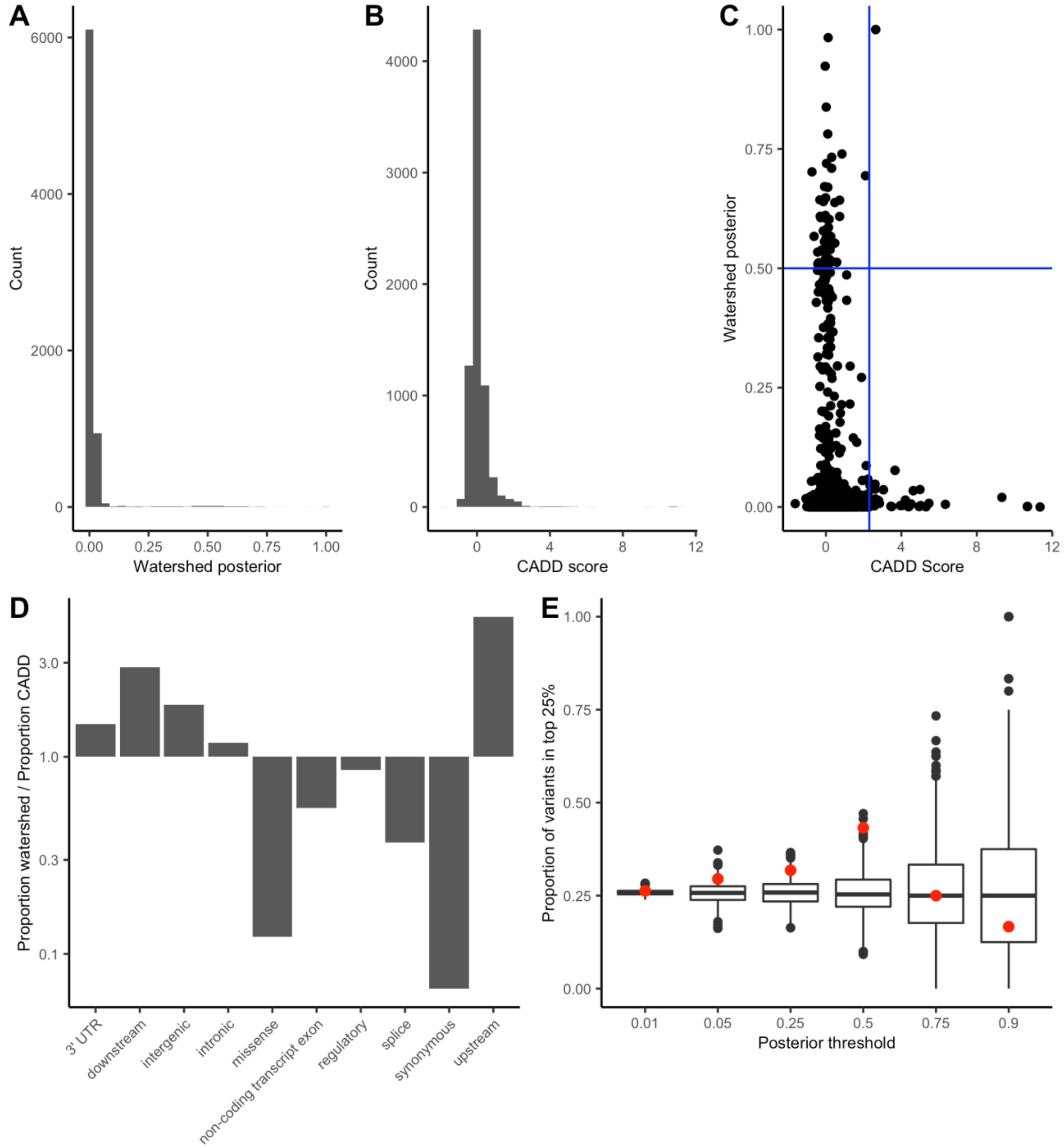


Figure S30. High CADD and Watershed variants in UKBB. (A) Distribution of the maximum Watershed posterior per variant for the set of variants in co-localized regions tested by Watershed and in UKBB. (B) Distribution of CADD scores per variant for the same set of variants in co-localized regions tested by Watershed and in UKBB. (C) The maximum Watershed posterior vs. CADD score for the tested variants in UKBB. The blue lines represent cut-offs of watershed posterior > 0.5, and the matching CADD threshold, 2.3, to obtain the same number of variants. (D) Of the high watershed and CADD variants in colocalized regions, the proportion of Watershed variants belonging to a specific category over the proportion of CADD variants in the same category. The y-axis is log-scaled, so bars below 1 indicate the category is more common in high CADD variants, and vice versa. (E) Filtering by the CADD score that returns the same number of variants as the Watershed posterior on the x-axis, and returning the

proportion that fall in the top 25% of effect sizes across traits in co-localized regions (red), and the proportion obtained by selecting a random set of tested variants equal in size (black).

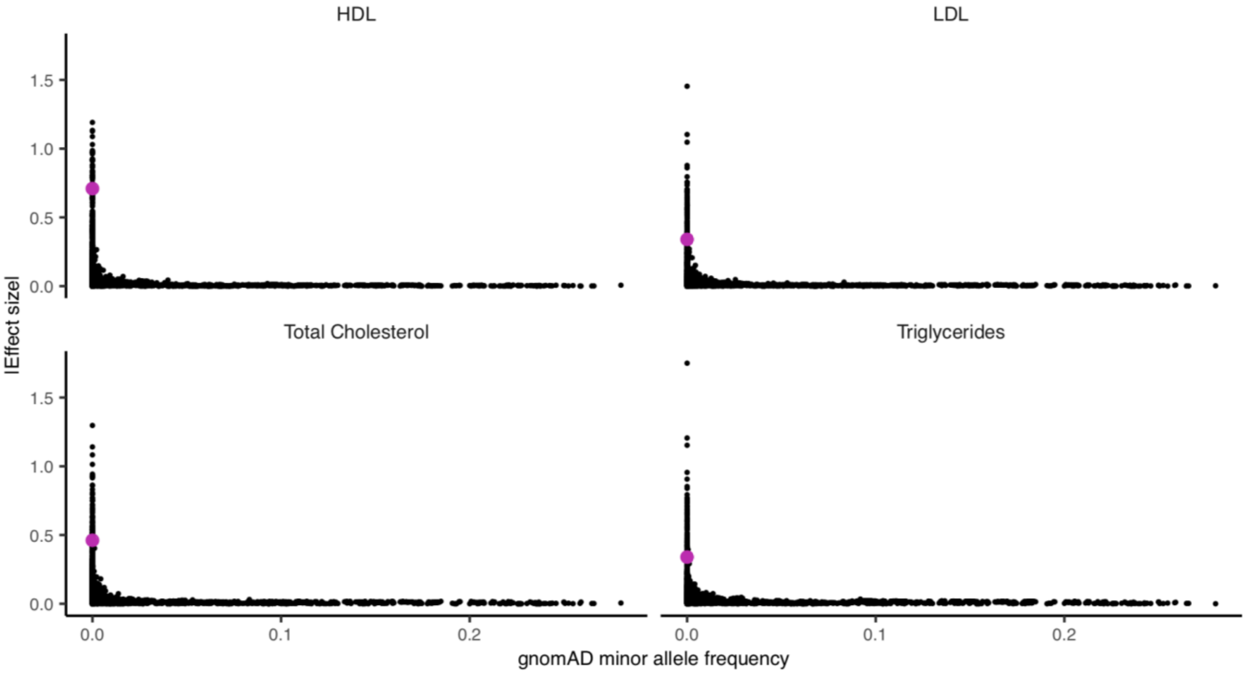


Figure S31. Distribution of rs564796245 effect sizes in MVP. All variants within a 250kb window of the high Watershed variant, in pink, rs564796245, tested for four related traits in the MVP cohort. The variant has a minor allele count of 11 in MVP, and for the set of rare variants tested in this window with a gnomAD non-Finnish European AF < 0.1%, it falls in the 99th percentile for HDL, 95th for LDL, 97th for Total Cholesterol, and 95th for Triglycerides.

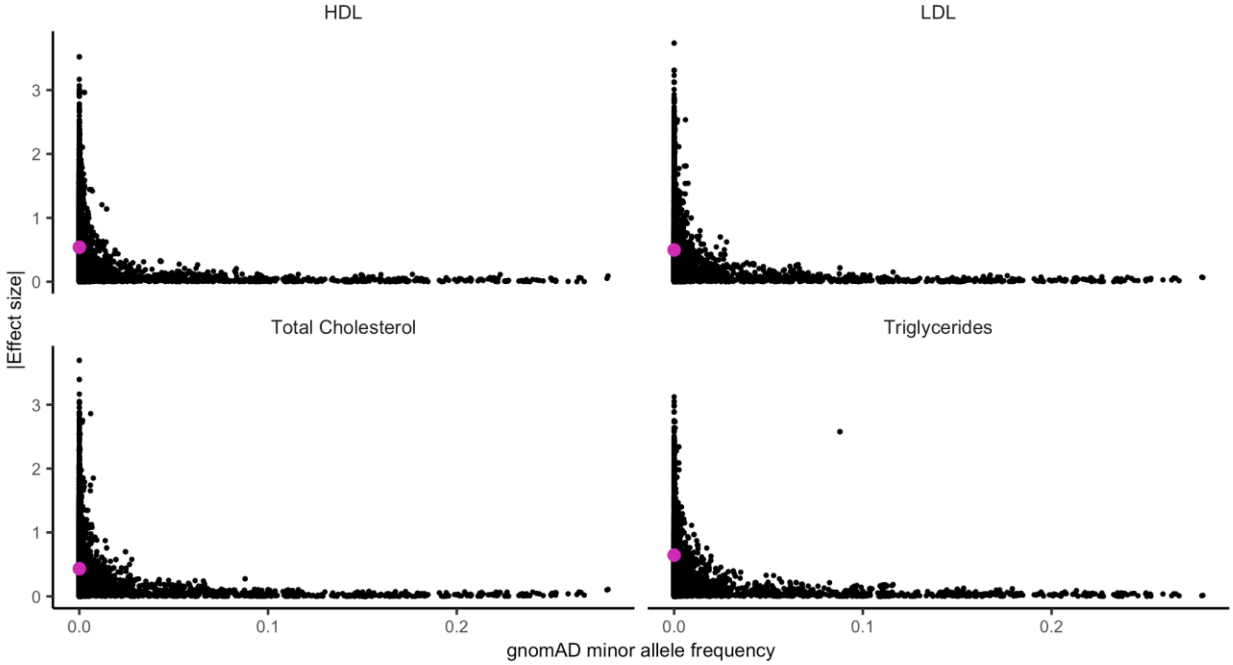


Figure S32. Distribution of rs564796245 effect sizes in JHS. All variants within a 250kb window of the high Watershed variant, in pink, rs564796245, tested for four related traits in the JHS cohort. The variant has a minor allele count of 4 in JHS, and for the set of rare variants tested in this window with a gnomAD non-Finnish European AF < 0.1%, it falls in the 69th percentile for HDL, 66th for LDL, 62nd for Total Cholesterol, and 72nd for Triglycerides.

Bibliography

1. Albert, Frank W., and Leonid Kruglyak. 2015. “The Role of Regulatory Variation in Complex Traits and Disease.” *Nature Reviews. Genetics* 16 (4): 197–212.
2. Arking, Dan E., Sara L. Pulit, Lia Crotti, Pim van der Harst, Patricia B. Munroe, Tamara T. Koopmann, Nona Sotoodehnia, et al. 2014. “Genetic Association Study of QT Interval Highlights Role for Calcium Signaling Pathways in Myocardial Repolarization.” *Nature Genetics* 46 (8): 826–36.
3. Astle, William J., Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L. Mann, Daniel Mead, et al. 2016. “The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease.” *Cell* 167 (5): 1415–29.e19.
4. Banovich, Nicholas E., Yang I. Li, Anil Raj, Michelle C. Ward, Peyton Greenside, Diego Calderon, Po Yuan Tung, et al. 2018. “Impact of Regulatory Variation across Human iPSCs and Differentiated Cells.” *Genome Research* 28 (1): 122–31.
5. Battle, Alexis, Sara Mostafavi, Xiaowei Zhu, James B. Potash, Myrna M. Weissman, Courtney McCormick, Christian D. Haudenschild, et al. 2014. “Characterizing the Genetic Basis of Transcriptome Diversity through RNA-Sequencing of 922 Individuals.” *Genome Research* 24 (1): 14–24.
6. Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society* 57 (1): 289–300.

7. Burke, Michael A., Stuart A. Cook, Jonathan G. Seidman, and Christine E. Seidman. 2016. “Clinical and Mechanistic Insights Into the Genetics of Cardiomyopathy.” *Journal of the American College of Cardiology* 68 (25): 2871–86.
8. Ernst, Jason, and Manolis Kellis. 2017. “Chromatin-State Discovery and Genome Annotation with ChromHMM.” *Nature Protocols* 12 (12): 2478–92.
9. Geijn, Bryce van de, Graham McVicker, Yoav Gilad, and Jonathan K. Pritchard. 2015. “WASP: Allele-Specific Software for Robust Molecular Quantitative Trait Locus Discovery.” *Nature Methods* 12 (11): 1061–63.
10. Grundberg, Elin, Kerrin S. Small, Åsa K. Hedman, Alexandra C. Nica, Alfonso Buil, Sarah Keildson, Jordana T. Bell, et al. 2012. “Mapping Cis- and Trans-Regulatory Effects across Multiple Tissues in Twins.” *Nature Genetics* 44 (10): 1084–89.
11. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, et al. 2017. “Genetic Effects on Gene Expression across Human Tissues.” *Nature* 550 (7675): 204–13.
12. Heintzman, Nathaniel D., Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, et al. 2009. “Histone Modifications at Human Enhancers Reflect Global Cell-Type-Specific Gene Expression.” *Nature* 459 (7243): 108–12.

13. Hong, Kyung-Won, Ji Eun Lim, Jong Wook Kim, Yasuharu Tabara, Hirotsugu Ueshima, Tetsuro Miki, Fumihiko Matsuda, Yoon Shin Cho, Yeonjung Kim, and Bermseok Oh. 2014. “Identification of Three Novel Genetic Variations Associated with Electrocardiographic Traits (QRS Duration and PR Interval) in East Asians.” *Human Molecular Genetics* 23 (24): 6659–67.
14. Joehanes, Roby, Xiaoling Zhang, Tianxiao Huan, Chen Yao, Sai-Xia Ying, Quang Tri Nguyen, Cumhur Yusuf Demirkale, et al. 2017. “Integrated Genome-Wide Analysis of Expression Quantitative Trait Loci Aids Interpretation of Genomic Association Studies.” *Genome Biology* 18 (1): 16.
15. Kim-Hellmuth, Sarah, Matthias Bechheim, Benno Pütz, Pejman Mohammadi, Yohann Nédélec, Nicholas Giangreco, Jessica Becker, et al. 2017. “Genetic Regulatory Effects Modified by Immune Activation Contribute to Autoimmune Disease Associations.” *Nature Communications* 8 (1): 266.
16. Knowles, David A., Joe R. Davis, Hilary Edgington, Anil Raj, Marie-Julie Favé, Xiaowei Zhu, James B. Potash, et al. 2017. “Allele-Specific Expression Reveals Interactions between Genetic Variation and Environment.” *Nature Methods* 14 (7): 699–702.
17. Kubara, Kenji, Kazuto Yamazaki, Yasuharu Ishihara, Takuya Naruto, Huan-Ting Lin, Ken Nishimura, Manami Ohtaka, et al. 2018. “Status of KRAS in iPSCs Impacts upon Self-Renewal and Differentiation Propensity.” *Stem Cell Reports* 11 (2): 380–94.
18. Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. ’t Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, et al. 2013.

“Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans.” *Nature* 501 (7468): 506–11.

19. Lian, Xiaojun, Cheston Hsiao, Gisela Wilson, Kexian Zhu, Laurie B. Hazeltine, Samira M. Azarin, Kunil K. Raval, Jianhua Zhang, Timothy J. Kamp, and Sean P. Palecek. 2012. “Robust Cardiomyocyte Differentiation from Human Pluripotent Stem Cells via Temporal Modulation of Canonical Wnt Signaling.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (27): E1848–57.
20. Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. “The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection.” *Cell Systems* 1 (6): 417–25.
21. Li, Xin, Yungil Kim, Emily K. Tsang, Joe R. Davis, Farhan N. Damani, Colby Chiang, Gaelen T. Hess, et al. 2017. “The Impact of Rare Variation on Gene Expression across Tissues.” *Nature* 550 (7675): 239–43.
22. Li, Yang I., Bryce van de Geijn, Anil Raj, David A. Knowles, Allegra A. Petti, David Golan, Yoav Gilad, and Jonathan K. Pritchard. 2016. “RNA Splicing Is a Primary Link between Genetic Variation and Disease.” *Science* 352 (6285): 600–604.
23. Nicolae, Dan L., Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan, and Nancy J. Cox. 2010. “Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS.” *PLoS Genetics* 6 (4): e1000888.

24. Okita, Keisuke, Tomoko Ichisaka, and Shinya Yamanaka. 2007. “Generation of Germline-Competent Induced Pluripotent Stem Cells.” *Nature* 448 (7151): 313–17.

25. Ongen, Halit, Alfonso Buil, Andrew Anand Brown, Emmanouil T. Dermitzakis, and Olivier Delaneau. 2016. “Fast and Efficient QTL Mapper for Thousands of Molecular Phenotypes.” *Bioinformatics* 32 (10): 1479–85.

26. Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. “Integrative Analysis of 111 Reference Human Epigenomes.” *Nature* 518 (7539): 317–30.

27. Rook, Martin B., Melvin M. Evers, Marc A. Vos, and Marti F. A. Bierhuizen. 2012. “Biology of Cardiac Sodium Channel Nav1.5 Expression.” *Cardiovascular Research* 93 (1): 12–23.

28. Shabalin, Andrey A. 2012. “Matrix eQTL: Ultra Fast eQTL Analysis via Large Matrix Operations.” *Bioinformatics* 28 (10): 1353–58.

29. Stegle, Oliver, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. 2012. “Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses.” *Nature Protocols* 7 (3): 500–507.

30. Storey, John D., and Robert Tibshirani. 2003. “Statistical Significance for Genomewide Studies.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (16): 9440–45.

31. Strober, B. J., R. Elorbany, K. Rhodes, N. Krishnan, K. Tayeb, A. Battle, and Y. Gilad. 2019. “Dynamic Genetic Regulation of Gene Expression during Cellular Differentiation.” *Science* 364 (6447): 1287–90.

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

33. Tam, Vivian, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. 2019. “Benefits and Limitations of Genome-Wide Association Studies.” *Nature Reviews. Genetics* 20 (8): 467–84.

34. Zhou, Zhenhai, Honggui Yu, Yunjia Wang, Qiang Guo, Longjie Wang, and Hongqi Zhang. 2016. “ZNF606 Interacts with Sox9 to Regulate Chondrocyte Differentiation.” *Biochemical and Biophysical Research Communications* 479 (4): 920–26.

35. Zhu, Zhihong, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R. Robinson, Joseph E. Powell, Grant W. Montgomery, et al. 2016. “Integration of Summary Data from GWAS and eQTL Studies Predicts Complex Trait Gene Targets.” *Nature Genetics* 48 (5): 481–87.

36. Yao, D. W., O’Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).

37. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 810 (2020).

38. Neavin, D. *et al.* Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome Biol.* **22**, 76 (2021).
39. Findley, A. S. *et al.* Functional dynamic genetic effects on gene regulation are specific to particular cell types and environmental conditions. *Elife* **10**, (2021).
40. Perez, R. K. *et al.* Multiplexed scRNA-seq reveals the cellular and genetic correlates of systemic lupus erythematosus. *In development*, (2021)
41. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
42. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
43. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
44. Elorbany, R. *et al.* Single-Cell Sequencing Reveals Lineage-Specific Dynamic Genetic Regulation of Gene Expression During Human Cardiomyocyte Differentiation. *bioRxiv* 2021.06.03.446970 (2021) doi:10.1101/2021.06.03.446970.
45. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).
46. Kim-Hellmuth, S. *et al.* Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, (2020).

47. Baran, Y. *et al.* MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* **20**, 206 (2019).

48. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017)

49. Wang, W. & Stephens, M. Empirical Bayes Matrix Factorization. *arXiv [stat.ME]* (2018).

50. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

51. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).

52. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

53. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* **2**, lqaa078 (2020).

54. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

55. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

56. Ferraro, N. M. *et al.* Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* 369, (2020).

57. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336, 740–743 (2012).

58. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* 18, 77 (2017).

59. Zeng, Y. *et al.* Aberrant gene expression in humans. *PLoS Genet.* 11, e1004942 (2015).

60. Frésard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat. Med.* 25, 911–919 (2019).

61. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209 (2018).

62. Chanfreau-Coffinier, C. *et al.* Projected Prevalence of Actionable Pharmacogenetic Variants and Level A Drugs Prescribed Among US Veterans Health Administration Pharmacy Users. *JAMA Netw Open* 2, e195345 (2019).

63. Taylor, H. A., Jr *et al.* Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. *Ethn. Dis.* 15, S6–4–17 (2005).

64. Mohammadi, P. *et al.* Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* 366, 351–356 (2019).

65. Spitz, F. Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. *Semin. Cell Dev. Biol.* 57, 57–67 (2016).

66. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443 (2020).

67. Yang, Z.-F., Mott, S. & Rosmarin, A. G. The Ets transcription factor GABP is required for cell-cycle progression. *Nat. Cell Biol.* 9, 339–346 (2007).

68. Takahashi, Y., Rayman, J. B. & Dynlacht, B. D. Analysis of promoter binding by the E2F and pRB families in vivo: distinct E2F proteins mediate activation and repression. *Genes Dev.* 14, 804–816 (2000).

69. Gordon, S., Akopyan, G., Garban, H. & Bonavida, B. Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene* 25, 1125–1142 (2006).

70. Kang, K., Choi, Y., Kim, H. H., Yoo, K. H. & Yu, S. Predicting FOXM1-Mediated Gene Regulation through the Analysis of Genome-Wide FOXM1 Binding Sites in MCF-7, K562, SK-N-SH, GM12878 and ECC-1 Cell Lines. *Int. J. Mol. Sci.* 21, (2020).

71. Zhang, S. *et al.* Base-specific mutational intolerance near splice sites clarifies the role of nonessential splice nucleotides. *Genome Res.* 28, 968–974 (2018).

72. Shapiro, M. B. & Senapathy, P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* 15, 7155–7174 (1987).

73. Coolidge, C. J., Seely, R. J. & Patton, J. G. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res.* 25, 888–896 (1997).

74. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315 (2014).

75. Pala, M. *et al.* Population- and individual-specific regulatory variation in Sardinia. *Nat. Genet.* 49, 700–707 (2017).

76. Georgi, B. *et al.* Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate. *PLoS Genet.* 10, e1004229 (2014).

77. Brandt, M., Gokden, A., Ziosi, M. & Lappalainen, T. A polyclonal allelic expression assay for detecting regulatory effects of transcript variants. *Genome Med.* 12, 79 (2020).

78. Forero, D. A. *et al.* APOE gene and neuropsychiatric disorders and endophenotypes: A comprehensive review. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 177, 126–142 (2018).

79. Habib, A. M. *et al.* Microdeletion in a FAAH pseudogene identified in a patient with high anandamide concentrations and pain insensitivity. *Br. J. Anaesth.* 123, e249–e253 (2019).

80. Préfontaine, D. *et al.* Increased IL-33 expression by epithelial cells in bronchial asthma. *J. Allergy Clin. Immunol.* 125, 752–754 (2010).

81. Smith, D. *et al.* A rare IL33 loss-of-function mutation reduces blood eosinophil counts and protects from asthma. *PLoS Genet.* 13, e1006659 (2017).

82. Olafsdottir, T. A. *et al.* Eighty-eight variants highlight the role of T cell regulation and airway remodeling in asthma pathogenesis. *Nat. Commun.* 11, 393 (2020).

83. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* 50, 1514–1523 (2018).

84. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* 50, 151–158 (2018).