Cell type-specific Expression Quantitative Trait Loci

Paul Little[1,*], Vasyl Zhabotynsky[2], Yun Li[2,3], Danyu Lin[2,4], Wei Sun[1,2,5*]

Last updated: March 31, 2022

1. Biostatistics Program, Public Health Science Division, Fred Hutchinson Cancer Research Center
2. Department of Biostatistics, University of North Carolina at Chapel Hill
3. Department of Genetics, University of North Carolina at Chapel Hill
4. Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill
5. Department of Biostatistics, University of Washington
All correspondence should be addressed to Paul Little plittle@fredhutch.org or Wei Sun wsun@fredhutch.org.

ORCID for Paul Little: 0000-0002-9789-8501, ORCID for Wei Sun: 0000-0002-6350-1107

Abstract

Genetic regulation of gene expression can vary across cell types. To map cell type-specific gene expression quantitative trait loci (eQTLs), a popular approach is to assess the interaction between genotype and cell type proportions in a linear model. A linear model requires transformation of RNA-seq count data, which distorts the relation between gene expression and cell type proportions and it could lead to reduced power and inflated false positive findings. We have developed a statistical method, CSeQTL, to perform cell type-specific eQTL mapping for bulk RNA-seq data while directly modeling RNA-seq count data. CSeQTL utilizes both total expression and allele-specific expression to improve the power of eQTL mapping. As demonstrated through simulations and applications on human whole blood and brain RNA-seq data, CSeQTL controls false positive probability and reports much more findings than the linear model approach. We have conducted a comprehensive evaluation of the overlap between cell type-specific eQTLs and the genetic loci associated with 21 categories of human traits and identified several interesting findings. For example, the genetic loci associated with substance addiction are enriched among the eQTLs in the excitatory neurons of schizophrenia subjects, but not in healthy controls.

Keywords: Cell types; eQTL mapping; Genetics; GTEx; Immunity; RNA-seq; scRNA-seq; Schizophrenia; Whole blood.

Introduction

The study of gene expression is critical to understanding cellular and molecular biology. Recent studies of gene expression by single-cell RNA sequencing (scRNA-seq) reveal that gene expression can vary substantially across cell types, and cell type compositions vary across tissue samples [1]. Therefore variation of gene expression across bulk tissue samples reflect the variation due to cell type-specific expression as well as cell type compositions [2]. Studying gene expression quantitative trait loci (eQTLs) (i.e., genetic variants associated with gene expression) is an effective approach to investigate the genetic basis of complex traits [3,4]. Since genetic loci implicated in a disease may be associated with specific cell types [5, 6], cell type-specific eQTLs can provide additional insights for the genetic basis of complex traits [7–9].

Most eQTL studies use RNA-seq data from bulk samples and ignore the heterogeneity of cell type compositions. Gene expression variation due to cell type composition may be partially captured by principal components or latent factors such as the probabilistic estimation of expression residuals (PEER) factors [10, 11]. A few studies have conducted cell type-specific eQTL mapping using bulk tissue gene expression by testing the interaction between genotype and cell type proportion in a linear model [3, 4, 9, 12–16]. Such an "interaction eQTL" is not equivalent to a cell type-specific eQTL since it assesses the difference of eQTL effect sizes between cell types rather than eQTL strength per cell type. It is straightforward to assess cell type-specific eQTL effect using such a linear model, by adding up the baseline eQTL effect (in a reference cell type) and the interaction term, however, a linear model is not appropriate for RNA-seq count data due to the strong association between mean value and variance in count data. A log or normal quantile transformation of count data may stabilize the variance, but then the relation between transformed gene expression and cell type proportion is not linear anymore. A more appropriate eQTL modeling choice is to use a negative binomial model. In addition to total read count (TReC), allele-specific expression (ASE) measured by allele-specific read count (ASReC) is available from RNA-seq data. Combining TReC and ASE (with negative binomial and beta-binomial models respectively) can improve the power for *cis*-eQTL mapping [17]. The term *cis*-eQTL is often used to refer to a local eQTL that is close to its target gene, though here we use a more traditional *cis*-eQTL definition such that a *cis*-eQTL leads to allelic imbalance of gene expression. For example, an alternative allele in the paternal copy of the chromosome only alters paternal expression but not maternal expression. The vast majority of the eQTLs that can be identified with sample sizes of around a few hundreds are local eQTLs and most of them are *cis*-eQTLs.

Recently, we have also developed a method named pTReCASE for eQTL mapping using RNA-seq data from tumor samples where we treated tumor and non-tumor cells as two cell types with known cell type composition [18]. However, pTReCASE is limited to a simple situation with two cell types where both cell

types have sufficient expression and their cell type proportions have enough variation across samples. In a more general situation with an arbitrary number of cell types, several challenges arise. A cell type may have nearly constant proportions across samples, which leads to an identifiability issue to estimate the cell type-specific eQTL effect for that cell type. A gene's expression may be zero or near zero in a subset of cell types, and thus it is challenging or impossible to estimate eQTL effects in such cell types. To address these challenges, we start from scratch and design a flexible and robust computational framework to handle these issues.

With the rise of scRNA-seq data, bulk RNA-seq continues to provide multiple benefits for the study of cell type-specific eQTLs. First, scRNA-seq can be very expensive for studies with large samples sizes and its application is also limited by the requirement of high quality samples. Second, scRNA-seq may not give an unbiased sampling of all the cell types in a tissue sample. Third, inherent sparsity of scRNA-seq data often makes it challenging to assign the cell types of many cells. Our method allows a new paradigm of study design: collecting scRNA-seq data in a subset of samples along with bulk RNA-seq in all samples. The scRNA-seq data can be used to construct a cell type-specific gene expression reference, and then after estimating cell type proportions using such reference, the bulk RNA-seq data can be used for cell type-specific eQTL mapping.

Results

A brief introduction of CSeQTL

Our method is called **C**ell type-**S**pecific **eQTL** or CSeQTL for short. CSeQTL jointly models TReC and ASReC outcomes as a function of covariates, cell type composition, and the genotype of one SNP (Single Nucleotide Polymorphism). More specifically, TReC and ASReC are modeled by a negative binomial (or Poisson) and a beta-binomial (or binomial) distribution, respectively, with shared parameters for genetic effects [17]. Unlike TReCASE and pTReCASE methods, CSeQTL needs to handle the challenging situations where cell type-specific gene expression may be zero or close to zero, or the proportion of one or more cell types could be close to zero or lack variation. All these challenging situations render the eQTL effects un-identifiable. We address these issues by multiple computational solutions, including trimming outliers of TReC to increase the robustness of our estimates and iteratively detecting and removing non-expressed cell types. We compare CSeQTL versus a linear model approach that we refer to as OLS (ordinary least squares) method, which is implemented as follows. First, for each gene, apply an inverse normal quantile transformation for read-depth normalized TReC. Next, define a reference cell type (usually the one with

highest proportions on average) and fit a linear model with transformed gene expression against the following covariates: cell type proportions of all but the reference cell type, SNP genotype, and interaction terms between genotype and cell type proportions, as well as other covariates such as age, sex, batches etc.. With this model setup, the cell type-specific eQTL effect for the reference cell type is the genotype main effect, and the cell type-specific eQTL effect is the summation of the genotype's main effect and the effect size of the corresponding interaction with genotype.

CSeQTL controls type I error and has much higher power than OLS

We conducted simulations to evaluate type I error and power of CSeQTL in a variety of settings. First, we varied the base-line expression (i.e., gene expression of the reference allele) across cell types to illustrate the impact of base-line expression on cell type-specific eQTL mapping. Second, we considered three scenarios of the variation in cell type composition for three cell types, referred to as CT1, CT2, and CT3 (Figure 1(a)). In scenario 1, identically and independently distributed cell type proportions were generated and then normalized to sum to one. This is an ideal, but unrealistic situation. In a more realistic scenario 2, on average CT2 has highest abundance, followed by CT1 and then CT3. CT3 proportion also has smaller variance, which makes it harder to map cell type specific eQTL for CT3. Lastly in scenario 3, we first simulated proportions as scenario 2 and then added outlier proportions to mimic observations in real data. A secondary set of simulations explored how performances would be affected given a noisy version of cell type proportion estimates (see Supplementary Figure 1 for an example).

We specified the expected expression for the reference allele of CT1 to be 500 and set the ASReC to be 5% of the TReC. We set the fold change of reference allele gene expression of CT2 or CT3 versus CT1 to be 0.1, 1.0, or 10. Following the TReCASE model, TReC and ASReC were simulated conditional on phased SNP genotypes, cell type proportions, expected expressions per allele and cell type, and other covariates. The sample size was 300. All the eQTLs were set to be *cis*-eQTLs that influenced both TReC and ASReC and we specified eQTL effect size by fold change of alternative allele B versus reference allele A. In a complete null situation, all cell type-specific eQTL effects were set to be 1.0 (Figure 1(b)). In another mixed null/alternative situation, we allowed CT1's eQTL effect to vary from $\exp(-1)$ to $\exp(1)$, CT2's eQTL effect was set to be 1 (i.e., no eQTL effect), and CT3's eQTL effect was set to be 1.5. This design allowed us to assess power in CT1 and CT3 and type I error in CT2 simultaneously (Figure 1(c)).

Under the both complete null and mixed null/alternative situations, CSeQTL controls type I error but OLS has apparent type I error inflation in several configurations (Figure 1(b-c)). Next we focus on the mixed null/alternative situation where CSeQTL has higher power than OLS in most configurations. Under

scenario 1 when three cell types have the same distribution of proportions, CSeQTL generally has higher power than OLS to detect eQTLs for CT1 and CT3. When the base-line expression of CT3 is low (0.1 fold of CT1), OLS's power in CT3 is positively correlated with CT1's eQTL effect size even though CT3 has a constant effect size throughout, implying "leaking" of eQTL effect from CT1 to CT3 that is likely due to transformation of gene expression. OLS suffers from type I error inflation in cases where CT2 has lower base-line expression, which underscores the difficulty to estimate eQTL effects when cell type-specific gene expression level is low. In scenario 2 where CT1 has highest proportion and CT3 has lowest proportion, power to detect eQTLs is reduced across models and cell types when compared with scenario 1. CSeQTL's power to detect CT1 eQTLs are much higher than OLS. CT3 eQTLs are detectable by either method if its baseline expression is high and in that case (2nd row and 2nd column of Figure 1(c)), CSeQTL has much higher power than OLS, e.g., $> 80\%$ power vs. $< 20\%$ power in most cases. OLS still has type I error inflation for CT2, though in a smaller magnitude than scenario 1. Finally in scenario 3, the introduction of outlier cell type proportions substantially increases the type I error inflation of OLS for CT2 when baseline expression of CT2 is low. Additional simulation results as well as an additional set of results using a noisy version cell type proportions are presented in Supplementary Materials (Supplementary Figures 2-5).

Our implementation allows trimming of outliers whose Cook's distance is larger than a threshold, following the approach used by DESeq2 [19]. The Cooks' distance is calculated based on the null model (no eQTL) and the value of outliers are imputed with the null model's predicted outcome (See Online Methods Section 3 for more details). This trimming procedure may slightly reduce the power, but helps to guard against type I error. The impact of trimming is more apparent in one dataset (GTEx brain samples) in real data analysis, which is discussed in the next section.

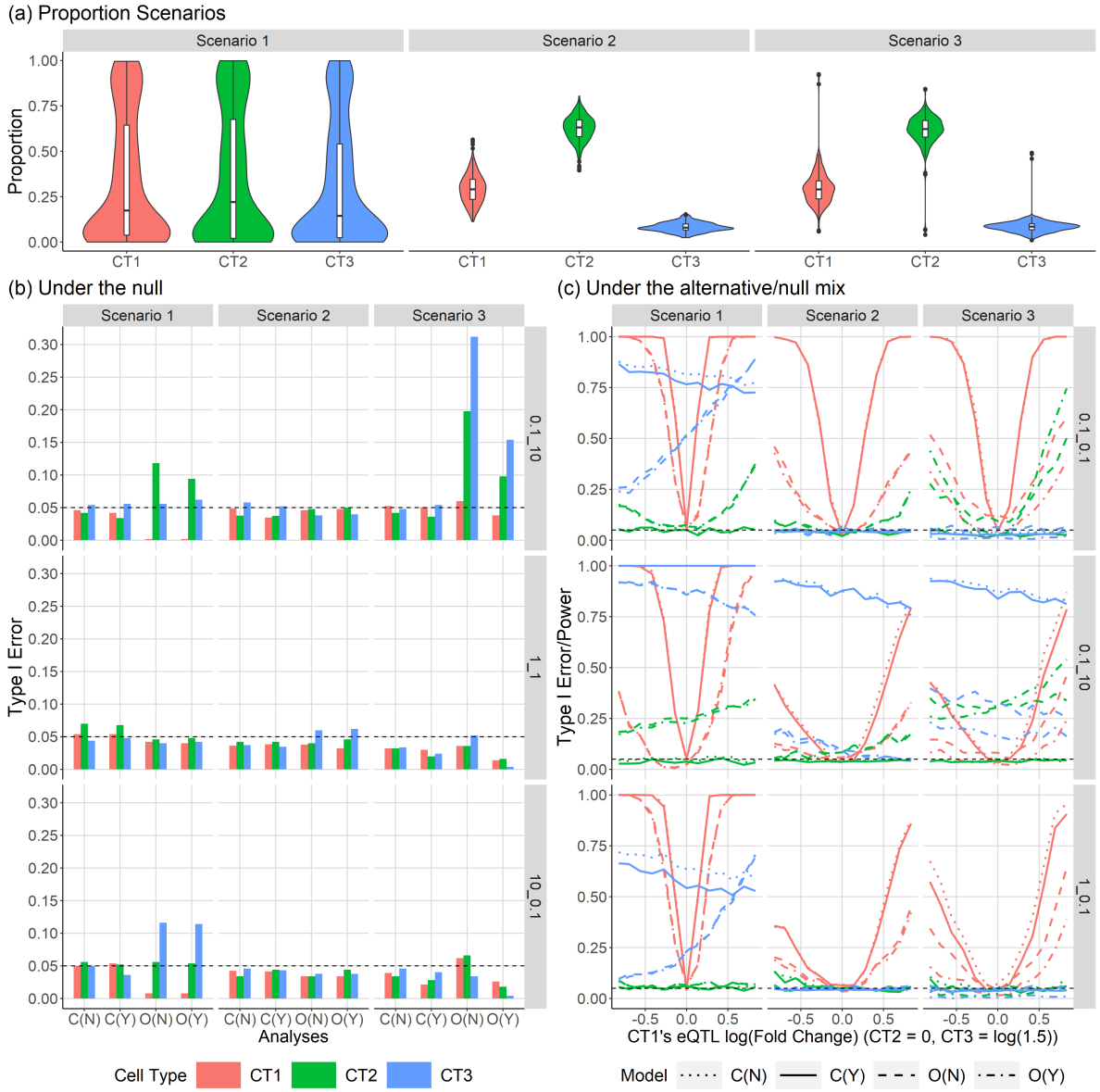

Figure 1: **Summary of the results from simulation studies.** (a) Simulated cell type proportions for three scenarios. Scenario 1: equally abundant and highly variable cell type proportions. Scenario 2: variable abundance and smaller variance. Scenario 3: modification of scenario 2 by adding outliers of cell type proportions. (b) Simulation results under the global null (i.e., no eQTL for any cell type) for different scenarios and methods (columns of the plots) and reference allele expression configurations (rows of the plots). Each reference allele configuration is denoted by fold change of reference allele expression. in CT2 and CT3 versus CT1. For example, "10_0.1" indicates fold changes of 10 in CT2 over CT1 and 0.1 in CT3 over CT1. X-axis labels C(N), C(Y), O(N), and O(Y) correspond to CSeQTL w/o trimming, CSeQTL w/ trimming, OLS w/o trimming, and OLS w/ trimming, respectively. (c) Simulation under the mixture of null and alternative hypothesis by models, scenarios, and reference allele expression configurations.

CSeQTL identifies many more eQTLs than the linear model in human brain and blood

We applied CSeQTL to analyze bulk RNA-seq data from human brain generated by CommonMind Consortium (CMC) (254 schizophrenia patients and 283 controls) [20, 21], bulk RNA-seq data from human brain (frontal cortex, n=175) and whole blood (n=670) from Genotype Tissue Expression's (GTEx) [20–22], as well as cell type-specific bulk RNA-seq data from a few immune cell types, which were generated by the BLUEPRINT project [23]. The results for BLUEPRINT data were used to evaluate our findings from human blood. To calculate cell type compositions in human brain, we estimated cell type-specific gene expression using scRNA-seq data collected from middle temporal gyrus (MTG) of human brains [24]. For whole blood, we used a popular reference known as LM22, which contains cell type-specific expression for 22 cell types [25]. Since LM22 was derived from microarray, we applied a transformation to align it with RNA-seq data. See Supplementary Materials Sections 2 for more details.

With each of the four cohorts (CMC brain, GTEx brain, GTEx whole blood, and BLUEPRINT), we obtained phased genotypes, TReC, ASReC based on phased genotypes, observed covariates, and latent batch covariates (See Supplementary Materials Sections 4 for more details). Given TReC and cell type-specific reference for brain (MTG) and blood (LM22), we estimated cell type proportions using ICeDT [26] (Figure 2). With CMC, individuals were stratified by disease status (schizophrenia vs. control). The distribution of cell type proportions were comparable between schizophrenia's patients and healthy controls from CMC and GTEx's brain samples (Figure 2(a)). Excitatory neurons had highest proportions followed by astrocytes, inhibitory neurons, oligodendrocytes, oligodendrocyte precursor cells (OPC). Microglia was the cell type with lowest proportions as well as smallest variation, hence it would be very challenging to detect cell type-specific eQTLs in microglia. For the 22 blood cell types, we collapsed them into 7 cell types since some cell types had limited prevalence and variability (Supplementary Figure 8) which limited the power for cell type-specific eQTL mapping (See online methods for more details). For GTEx's whole blood samples, neutrophil was the dominant cell type with highest proportions and largest variance (Figure 2(b)).

We conducted both cell type-specific eQTL mapping as well as traditional bulk eQTL mapping that assessed aggregated eQTL effects in bulk tissue samples. In addition to SNP genotype, a set of covariates were included for both cell type-specific and bulk eQTL mapping: library size, observed covariates such as age, sex, known batch effects, and genotype principal components (PCs). For bulk eQTL mapping, we also included cell type proportions as covariates. Additionally, we added latent factors estimated from gene expression data, which were calculated by PCs of residualized gene expression data accounting for all the aforementioned covariates. When using OLS for cell type-specific eQTL mapping, we included interaction

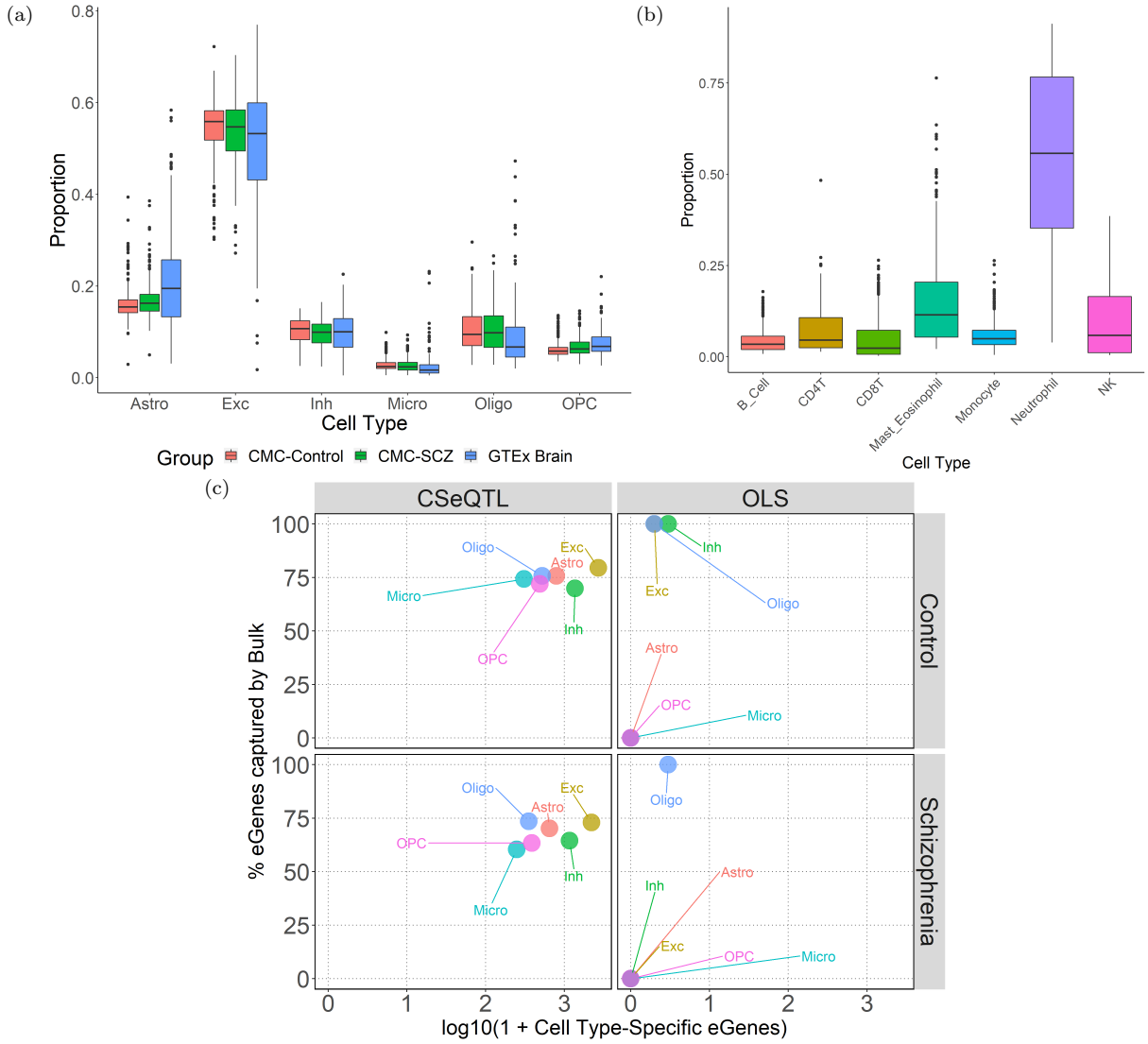

Figure 2: **Summary of cell type compositions and the number of eGenes.** (a) Cell type proportion estimates of six cell types astrocytes (Astro), excitatory neuron (Exc), inhibitory neuron (Inh), microglia (Micro), oligodendrocytes (Oligo), and oligodendrocyte precursor cells (OPC) from the brain samples of schizophrenia patients and controls from the CommonMind Consortium as well as samples from GTEx Brain. (b) Cell type proportion estimates of 7 cell types from whole blood samples of GTEx. Cell type proportions were first estimated for 22 cell types (Supplementary Figure 8) and then collapsed to 7 cell types to avoid individual cell types with very low proportions and variance. (c) A summary of the number of detected eGenes per stratum by method and case/control status. The X-axis is the (number of cell type-specific eGenes + 1) in log (base 10) scale. The Y-axis denotes the percentage of cell type-specific eGenes that overlap with eGenes from bulk eQTL mapping.

terms between genotype and proportions. Unlike previous analysis of CMC data [20] that pooled controls and schizophrenic cohorts to perform eQTL mapping, we ran stratified analyses allowing for the detection of disease status specific and cell type-specific eQTLs and performed additional downstream comparisons. Genes were excluded from the analysis if the 75th quantile of TReC was less than 50. SNPs considered per gene were among those located 50 kilobases before the transcription start site and 50 kilobases after the transcription end site. SNPs per gene were then excluded if the minor allele frequency fell below 5%.

We applied trimming to each gene using the Cook's distance derived from the respective model with a cutoff of standardized Cook's distance 10 for GTEx's brain data, which has smallest sample size and highest type I error inflation without trimming, and 20 for the other three cohorts. To explore the appropriate trimming thresholds per dataset, we permuted the order of samples within the phased genotype matrix to preserve SNP correlation and ran TReC-only eQTL mapping for all the genes on chromosome 1 with thresholds of 10, 15, 20, or without trimming. We selected the threshold per dataset to ensure type I error was controlled per cell type ($\alpha \approx$ proportion of nominal p-values across gene/SNP pairs $< \alpha$) with $\alpha = 0.05$. The selected cutoffs appeared to be associated with each cohort's sample sizes.

For eQTL mapping, two layers of multiple testing need to be accounted for: (1) testing across multiple local SNPs per gene and (2) testing across genes. For each gene, we assessed the significance of its minimum p-value across all local SNPs by calculating the corresponding permutation p-value. A brute-force implementation, which requires permuting the data many times and running CSeQTL on each permuted dataset, is computationally prohibitive. Instead, we employed a computationally efficient method named geoP [11, 27] that calculates the permutation p-value by estimating the effect number of independent tests. To account for multiple testing across genes, we selected a permutation p-value cutoff to control the false discovery rate (FDR). This two layer multiple testing correction procedure was applied to the bulk eQTL mapping as well as each cell type for cell type-specific eQTL mapping.

When performing bulk eQTL, CSeQTL is equivalent to the TReCASE method [28] that combines total expression and allele-specific expression to improve the power of eQTL mapping. Consistent with our previous results [11, 28], CSeQTL has much higher power than OLS for bulk eQTL mapping. For example, considering the results for CMC schizophrenia samples (n=250), for bulk eQTL mapping, CSeQTL and OLS identified around 10,000 and 3,000 eGenes (genes with at least one significant eQTL) respectively. For cell type-specific eQTLs, the difference is even more striking. Without trimming, CSeQTL identified 230 (microglia) to 2,544 (excitatory neurons) eGenes across cell types, while OLS identified 6 to 11 eGenes per cell type with 11 eGenes in microglia, the cell type where eQTL mapping is most challenging due to lowest cell type proportions and smallest variance of cell type proportions. After trimming outliers, CSeQTL identified similar number of cell type-specific eQTLs per cell type and OLS only identified 2 eQTLs in oligodendrocytes,

suggesting the OLS findings before trimming were false positives due to outliers in the data (Figure 2(c), Supplementary Table 2). The results were similar for CMC control samples (n=275), where CSeQTL could identify hundreds to thousands of eGenes per cell type while OLS only found a few (Supplementary Table 2).

For GTEx brain data, which had a relatively small sample size of 174, trimming of outliers had a large effect. For example, for microglia, the most challenging cell type to map eQTLs, CSeQTL and OLS identified 1,503 and 776 eQTLs before trimming, but 348 and 0 eQTLs after trimming (Supplementary Figure 14 and Supplementary Table 3). This result highlights that CSeQTL still has much higher power than OLS when sample size is smaller, but should be used with caution.

For BLUEPRINT with purified bulk tissue, CSeQTL again detected many more eGenes and eQTLs than the linear model along with a good deal of overlap among eGenes per cell type between models (Supplementary Table 4 and Supplementary Figure 18). The degree of shared eGenes between models and the three cell types is also provided which reveals more eGenes exclusively detected by CSeQTL over the linear model (Supplementary Figure 17).

GTEx whole blood is the only tissue among those we analyzed where OLS identified a substantial number of cell type-specific eQTLs after trimming. It identified 1,295 eGenes in neutrophil, and 10 or less eGenes in other cell types. In contrast, CSeQTL identified >4,000 eGenes in neutrophil, and hundreds of eGenes in other cell types (Supplementary Figures 20 and 22 and Supplementary Table 5).

Our results demonstrated limited overlap between the eGenes of different cell types (Supplementary Figures 11, 12, 15, 21), though the majority of the cell type-specific eQTLs (e.g., 50 to 90%) overlap with the bulk eQTLs. These results suggest that eQTL strength can be highly variable across cell types, though it could also be partly due to variation of statistical power across cell types. Very low consistency between cell type-specific and bulk eQTL may indicate false discoveries in cell type-specific eQTLs. For example, for GTEx brain study, before trimming, OLS identified 10,264 eQTLs in microglia for 776 eGenes, while only 0.03% overlap with bulk eQTLs, and none of these 10,000+ eQTLs remained significant after trimming outliers (Supplementary Table 3).

Validation of CSeQTL findings across studies

We validated CSeQTL findings across studies by comparing CMC schizophrenia vs. control or GTEx brain and GTEx whole blood vs. BLUEPRINT which had gene expression from three purified cell types: CD4+ T, monocyte, and neutrophil. Hereafter we focused on the eQTL results after trimming outliers since earlier results demonstrated it could reduce the number of potential false positives, particularly in GTEx brain

data.

Since OLS found few cell type-specific eQTLs, we compared the results between CSeQTL and OLS by evaluating the correlation between log transformed minimum p-value per gene. When comparing CMC schizophrenia vs. CMC control, CSeQTL results had much stronger correlations than OLS across all cell types (Figure 3, Supplementary Table 6). When comparing results from CMC schizophrenia/controls versus GTEx brain, we observed significant positive correlation for excitatory neurons and astrocyte for either all genes, all eGenes, or all non-eGenes, and CSeQTL results had stronger correlations than OLS. The correlation were not significant for other cell types when considering all eGenes. It was likely due to the fact that GTEx brain study had smaller sample size, hence less accurate cell type-specific eQTL findings. For the findings from GTEx whole blood, we compared them with the results from BLUEPRINT, which had bulk RNA-seq data from three purified cell types: CD4+ T, monocyte, and neutrophil. Among the eGenes, the correlations were positive and significant for CSeQTL findings. OLS had cell type-specific eQTLs only for neutrophils, for which the across study correlation was higher for CSeQTL findings.

Next we focused on CSeQTL results and evaluated the distribution of cell type-specific eQTLs with respect to functional annotations of genomic regions (e.g., enhancers, promoters, 3' UTR, 5' UTR, etc.) by Torus [29] (Supplementary Figures 23-24). For brain tissues, the functional enrichment of eQTLs for excitatory neuron, which was the most abundant cell type, was similar to the functional enrichment of bulk eQTLs. The lack of significant functional enrichment in other cell types could be partially due to smaller number of cell type-specific eQTLs. Comparing brain samples of schizophrenia patients versus controls (either CMC controls and GTEx controls), enrichment of eQTLs at 5' UTR and non-coding (NC) transcript were observed in both control groups but were absent in schizophrenia patients. Since neutrophil was the dominant cell type in whole blood, as expected, functional enrichment in neutrophil and whole blood was highly consistent. For CD4+ T and monocyte, the overall enrichment patterns were consistent between CSeQTL results from GTEx whole blood and purified cell types from BLUEPRINT, with one noticeable exception that CSeQTL lacked the enrichment on promoters for CD4+ T and monocyte. These results suggest that despite the small proportions of CD4+ T and monocyte in whole blood and limited power of CSeQTL, there was consistent functional enrichment of the eQTLs from CSeQTL and purified cell types. For those cell types with lower proportions in whole blood, the enrichment of functional elements were insignificant for most categories, though one category "promoter flanking" standed out to be significant for most cell types, suggesting stronger cell type-specific eQTL signals for whole blood tended to be located in such regions.

Another way to compare cell type-specific eQTLs is to examine their locations relative to the corresponding genes [20]. We summarized the relative locations of the minimum p-value SNPs (minP-SNP) for the

CSeQTL eGenes (Supplementary Figures 10, 13, 16, 19), by calculating the distance from a minP-SNP to the center of the corresponding gene and normalizing/dividing it by half of the length of the corresponding gene. Hence $[-1, 1]$ was the range of gene body and normalized distances beyond -15 and 15 were trimmed. The distribution of the minP-SNPs between CMC schizophrenia and controls were similar (Supplementary Figure 10). The locations of minP-SNPs from bulk eQTL mapping showed enrichment around transcription start site (TSS) or transcription end site (TES), though such patterns were not as clear in the eQTLs of each brain cell type in CMC cohorts. A potential reason was that the eQTLs around TSS and TES were more likely to affect gene expression in most cell types since most cell types might use the same TSS and TES for each gene. In GTEx brain results, more cell type-specific eQTLs tended to be located beyond 15 unites of half gene length (e.g., 13% to 20% in GTEx brain vs. 5% to 12% in CMC cohorts, Supplementary Figures 10, 13), which might be due to limited sample size hence higher uncertainty to locate the eQTLs. For the three purified cell types from BLUEPRINT (Supplementary Figure 16), the enrichment of eQTLs around TSS was stronger than TES, which was different from bulk eGenes in brain tissues where the enrichment in TES was slightly higher. Similar patterns of eQTL locations were observed for the same three cell types in GTEx whole blood samples, providing additional support for the accuracy of CSeQTL results (Supplementary Figure 19).

CSeQTL helps interpret GWAS findings

EQTLs are often used to study the genetic basis of complex traits by examining their overlap with genetic loci identified from Genome-Wide Association Studies (GWAS). Here we systematically evaluated the overlap between cell type-specific eQTLs and GWAS hits of either all the traits included in the GWAS catalog [30] or 21 categories of traits (Figure 4 and Supplementary Figure 5). We derived these 21 categories by grouping similar traits to study whether cell type-specific eQTLs were associated with some but not all traits. Calculations of GWAS enrichment and confidence intervals followed the procedure presented in Vasyl et al. [11].

GWAS hits of several categories were enriched in the bulk eQTLs of all three brain datasets, though the degree of enrichment (measured by log fold change in Figure 4) were small. Two categories with highest enrichment were Education/wealth and psychiatric/neurologic. When considering cell type-specific eQTLs by CSeQTL, due to the smaller number of eQTLs, the confidence to estimate enrichment was often low, which led to wider confidence intervals. Despite such limitation, we observed a very interesting finding that GWAS hits of psychiatric/neurologic traits and substance addiction were enriched in the cell type-specific eQTLs for excitatory neurons for schizophrenia patients, but not for the two control groups (CMC controls

and GTEx brain samples), suggesting potential disease-specific and cell type-specific eQTL signals. The GWAS loci of substance addiction were not enriched among the bulk eQTLs, and thus was an example where new insight could be discovered by cell type-specific eQTL mapping.

Blood is arguably most accessible tissue and thus molecular biomarkers (e.g., cell type-specific gene expression) in blood can be very valuable to understand the mechanism that connect genetic variants and complex traits. Our cell type-specific eQTL results provided a comprehensive resource for such studies. A few examples were immediately clear from our enrichment analysis (Figure 5). One was that GWAS hits of colon related diseases were enriched among the eQTLs of natural killer (NK) cells, but not other cell types. Such enrichment might be associated with the functional roles of NK cells in colon, such as response to intestinal bacterial by recruiting NK cells from blood [31]. Morphology (hair and balding morphology) GWAS loci were also enrichment in NK eQTLs, corroborating earlier studies on the association between hair follicle immune privilege and NK cell activity [32, 33]. Another example is the association between liver disease GWAS hits and the eQTLs in CD8+ T cells and earlier studies have reported that CD8+ T cells were associated with liver damage, hepatitis, immunopathology, and liver cancer [34–38].

Figure 3: **Comparing findings between data cohorts for shared cell types and between models.** The y-axis corresponds to correlation between negative log base 10 transform of nominal p-values per gene's minimum p-value SNP analyzed between a pair of cohorts. 95% confidence intervals are also plotted. For Blueprint and GTEx Whole Blood, CD4 T cells, monocytes, and neutrophils are presented. For Commonmind Consortium and GTEx Brain Frontal Cortex, bulk and the six cell types are presented. Pearson correlation was calculated among all shared genes analyzed ("All Genes"), among shared eGenes ("eGenes"), and among all shared non-eGenes ("Non-eGenes").

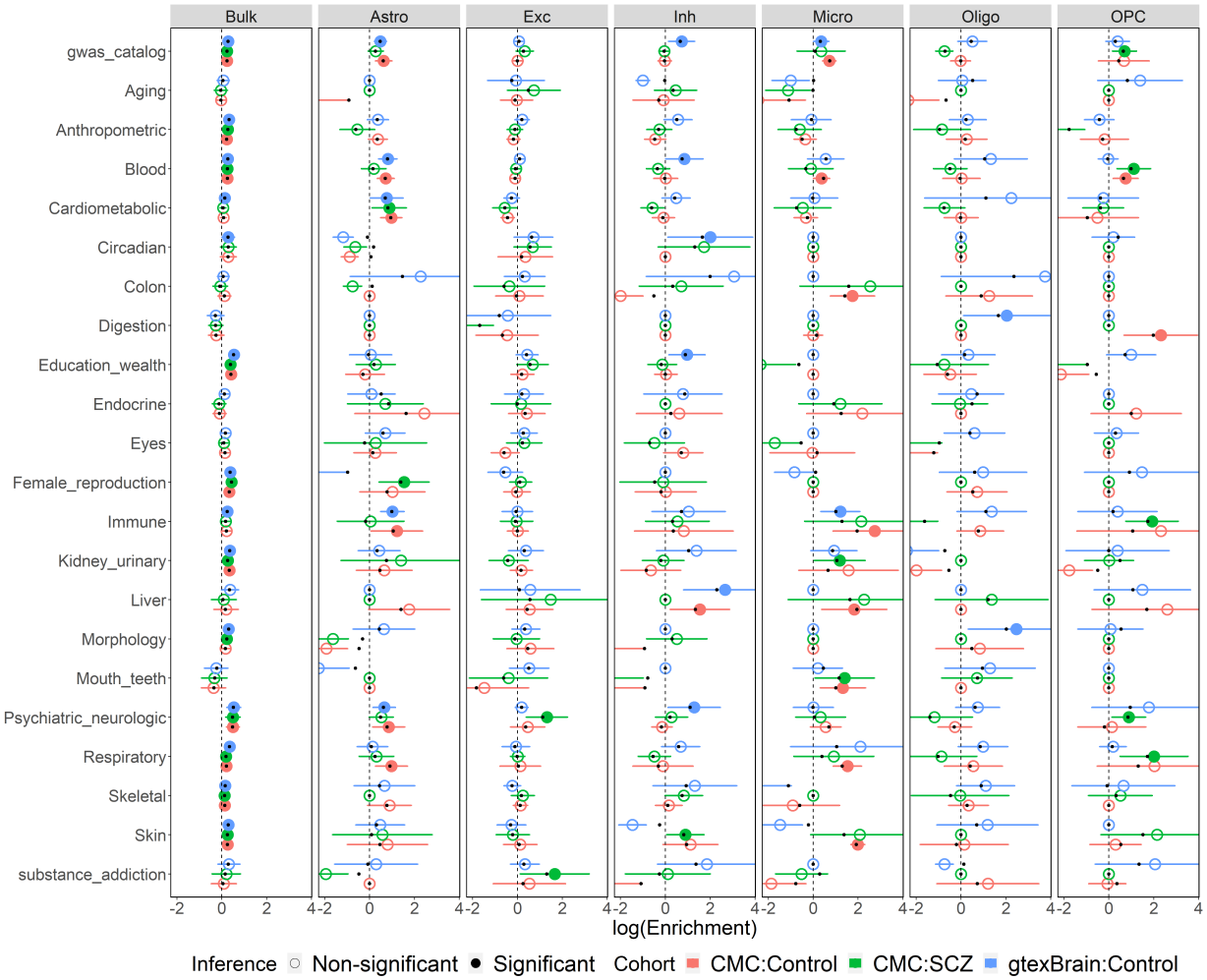

Figure 4: **GWAS enrichment for CMC and GTEx Brain.** Black points correspond to log enrichment using all available loci while open and filled circles correspond to jackknife estimates of log enrichment. Jackknife-based 95% confidence intervals are provided. Filled circles correspond to the lower bound of confidence intervals larger than zero or significant enrichment.

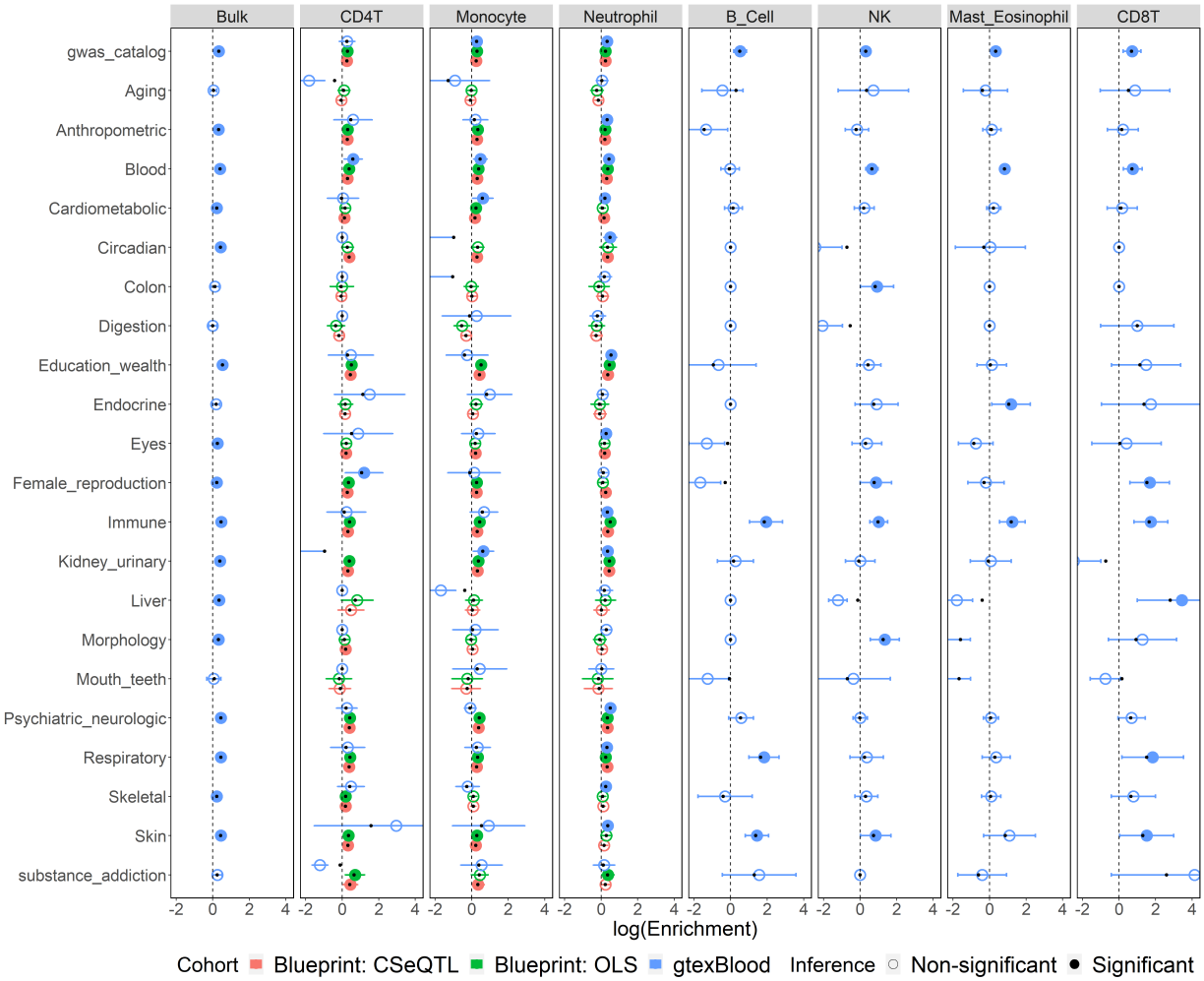

Figure 5: **GWAS enrichment for Blueprint and GTEx Whole Blood.** Black points correspond to log enrichment using all available loci while open and filled circles correspond to jackknife estimates of log enrichment. Jackknife-based 95% confidence intervals are provided. Filled circles correspond to the lower bound of confidence intervals larger than zero or significant enrichment.

Discussion

CSeQTL's framework allows mapping cell type-specific eQTLs using bulk RNA-seq data, by jointly modeling the effects of cell type composition and cell type-specific eQTLs. We have shown by both simulations and real data analyses that CSeQTL can have substantial power gain against a linear regression (ordinary least square) approach, while still maintaining type I error control. This is due to the underlying statistical model of CSeQTL. Deconvolution of gene expression to individual cell types should be done using untransformed count data [39], while eQTL mapping is often done using transformed gene expression data (e.g.. normal quantile transformation) to avoid the impact of outliers in count data. Such outliers are often due to the strong positive associations between mean and variance of count data. Our CSeQTL method satisfies these two restrictions by directly modeling count data using negative binomial distributions that accounts for the strong mean-variance dependence. In contrast, the linear regression approach uses transformed gene expression data and models cell type-specific eQTLs by adding interactions between cell type compositions and genotypes. The transformation of gene expression data distorts their associations with cell type compositions and thus can reduce power and inflate type I error. In addition, we also include allele-specific expression in our model to boost the power to detect cell type-specific eQTLs.

Model optimization of CSeQTL is very challenging because the model may not be identifiable, for example, due to a lack of variation of one cell type's proportion or very low expression of one gene in one or more cell types. A naive implementation may result in sub-optimal solutions due to non-invertible observed information matrices, negative variances, or extreme parameter estimates, which can have a profound impact on hypothesis testing. We have developed a comprehensive set of assessments to ensure the robust and optimal solution is obtained. In addition, both linear regression and CSeQTL can be sensitive to outlier data points, and we addressed this issue by trimming those outliers based on the null model without eQTL effects. As shown in our simulation and real data analyses, such trimming often have little impact, but can be very helpful for real datasets.

Our applications toward human brain and blood bulk RNA-seq data demonstrate that the linear regression method often identify 0 or a few cell type-specific eGenes (with the exception of neutrophil in whole blood) while CSeQTL can identify hundreds to thousands of cell type-specific eGenes. When examining the overlap between these cell type-specific eQTLs and GWAS findings, we have identified several interesting results, such as the overlap between the eQTLs in CD8+ T cells with GWAS hits for liver diseases, and the eQTLs in excitatory neurons of schizophrenia patients with GWAS hits for substance addiction. More importantly, our results provide a valuable resource for future studies of a specific trait and our method can be used to analyze other bulk RNA-seq datasets.

A limitation for applying our method or any cell type-specific eQTL mapping method on bulk RNA-seq data is accurate estimation of cell type composition, which in turn requires accurate cell type-specific gene expression reference. Here we have applied our method on the bulk RNA-seq data from brain and blood because these two tissues have readily available cell type-specific gene expression reference. We expect that in the near future, with the advance of the human cell atlas [1] or other similar projects, such a resource will become available in more tissue types.

Our work also enables a new type of study design to jointly model scRNA-seq and bulk RNA-seq data to study cell type-specific eQTLs. For example, scRNA-seq data can be collected in a small number of individuals, to be used as reference for cell type-specific expression. In addition, scRNA-seq data can also be used for eQTL mapping. After clustering and identification of cell types, scRNA-seq data can be converted to pseudo-bulk data of individual cell types and be used for eQTL mapping, e.g., by applying our TReCASE method [28]. The likelihood function of the TReCASE model can be combined with CSeQTL model in order to combine bulk RNA-seq and scRNA-seq data for cell type-specific eQTL mapping. Adding scRNA-seq data to bulk RNA-seq data can alleviate the challenges to using bulk RNA, e.g., limited variability in cell type proportion for one cell type. Adding bulk RNA-seq data to scRNA-seq data can reduce the cost, increase sample size, and avoid distortion of gene expression in the process of isolating single cells.

References

[1] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. Science forum: the human cell atlas. *elife*, 6:e27041, 2017.

[2] Daifeng Wang, Shuang Liu, Jonathan Warrell, Hyejung Won, Xu Shi, Fabio CP Navarro, Declan Clarke, Mengting Gu, Prashant Emani, Yucheng T Yang, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science*, 362(6420):eaat8464, 2018.

[3] GTEx Consortium et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[4] Sarah Kim-Hellmuth, François Aguet, Meritxell Oliva, Manuel Muñoz-Aguirre, Valentin Wucher, Silva Kasela, Stephane E Castel, Andrew Hamel, Ana Viñuela, Amy L Roberts, et al. Cell type specific genetic regulation of gene expression across human tissues. *bioRxiv*, page 806117, 2019.

[5] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235, 2015.

[6] NG Skene, J Bryois, TE Bakken, G Breen, JJ Crowley, HA Gaspar, P Giusti-Rodriguez, RD Hodge, JA Miller, AB Muñoz-Manchado, et al. Genetic identification of brain cell types underlying schizophrenia. *Nature Genetics*, 50(6):825, 2018.

[7] Darren J Burgess. Getting dynamic with eQTLs. *Nature Reviews Genetics*, 20(9):500–501, 2019.

[8] BJ Strober, Reem Elorbany, K Rhodes, Nirmal Krishnan, Karl Tayeb, Alexis Battle, and Yoav Gilad. Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, 364(6447):1287–1290, 2019.

[9] CA Glastonbury, Alves AC, Moustafa JS, and Small KS. Cell-type heterogeneity in adipose tissue is associated with complex traits and reveals disease-relevant cell-specific eQTLs. *The American Journal of Human Genetics*, 104:1013–1024, 2019.

[10] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500, 2012.

[11] Vasyl Zhabotynsky, Licai Huang, Paul Little, Yi-Juan Hu, Fernando Pardo-Manuel de Villena, Fei Zou, and Wei Sun. eQTL mapping using allele-specific count data is computationally feasible, powerful, and provides individual-specific estimates of genetic effects. *PLOS Genetics*, in press, 2022.

[12] Harm-Jan Westra, Danny Arends, Tõnu Esko, Marjolein J Peters, Claudia Schurmann, Katharina Schramm, Johannes Kettunen, Hanieh Yaghootkar, Benjamin P Fairfax, Anand Kumar Andiappan, et al. Cell specific eQTL analysis without sorting cells. *PLoS genetics*, 11(5):e1005223, 2015.

[13] Dvir Aran, Zicheng Hu, and Atul J Butte. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology*, 18(1):220, 2017.

[14] Bernard Ng, Charles C White, Hans-Ulrich Klein, Solveig K Sieberts, Cristin McCabe, Ellis Patrick, Jishu Xu, Lei Yu, Chris Gaiteri, David A Bennett, et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nature neuroscience*, 20(10):1418, 2017.

[15] Margaret KR Donovan, Agnieszka DAntonio-Chronowska, Matteo DAntonio, and Kelly A Frazer. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nature communications*, 11(1):1–14, 2020.

[16] Devanshi Patel, Xiaoling Zhang, John J Farrell, Jaeyoon Chung, Thor D Stein, Kathryn L Lunetta, and Lindsay A Farrer. Cell-type-specific expression quantitative trait loci associated with alzheimer disease in blood and brain tissue. *Translational Psychiatry*, 11(1):1–17, 2021.

[17] Wei Sun and Yijuan Hu. eqtl mapping using RNA-seq data. *Statistics in biosciences*, 5(1):198–219, 2013.

[18] Douglas R Wilson, Joseph G Ibrahim, and Wei Sun. Mapping tumor-specific expression QTLs in impure tumor samples. *Journal of the American Statistical Association*, pages 1–18, 2019.

[19] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014.

[20] Menachem Fromer, Panos Roussos, Solveig K Sieberts, Jessica S Johnson, David H Kavanagh, Thanneer M Perumal, Douglas M Ruderfer, Edwin C Oh, Aaron Topol, Hardik R Shah, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience*, 19(11):1442, 2016.

[21] Gabriel E Hoffman, Jaroslav Bendl, Georgios Voloudakis, Kelsey S Montgomery, Laura Sloofman, Ying-Chih Wang, Hardik R Shah, Mads E Hauberg, Jessica S Johnson, Kiran Girdhar, et al. CommonMind

Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Scientific data*, 6(1):1–14, 2019.

[22] GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[23] Lu Chen, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yan, Kousik Kundu, Simone Ecker, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, 167(5):1398–1414, 2016.

[24] Rebecca D Hodge, Trygve E Bakken, Jeremy A Miller, Kimberly A Smith, Eliza R Barkan, Lucas T Graybuck, Jennie L Close, Brian Long, Nelson Johansen, Osnat Penn, et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature*, 573(7772):61–68, 2019.

[25] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453, 2015.

[26] Douglas R Wilson, Joseph G Ibrahim, and Wei Sun. ICeD-T provides accurate estimates of immune cell abundance in tumor samples by allowing for aberrant gene expression patterns. *BioRxiv*, page 326421, 2018.

[27] Wei Sun and Fred A Wright. A geometric interpretation of the permutation p-value and its application in eQTL studies. *The Annals of Applied Statistics*, 4(2):1014–1033, 2010.

[28] Wei Sun. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics*, 68(1):1–11, 2012.

[29] Xiaoquan Wen. Molecular QTL discovery incorporating genomic annotations using bayesian false discovery rate control. *The Annals of Applied Statistics*, 10(3):1619–1638, 2016.

[30] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[31] Feixue Wang, Jennie Ka Ching Lau, and Jun Yu. The role of natural killer cell in gastrointestinal cancer: killer or helper. *Oncogene*, 40(4):717–730, 2021.

[32] Taisuke Ito, Natsuho Ito, Matthias Saatoff, Hideo Hashizume, Hidekazu Fukamizu, Brian J Nickoloff, Masahiro Takigawa, and Ralf Paus. Maintenance of hair follicle immune privilege is linked to prevention of nk cell attack. *Journal of Investigative Dermatology*, 128(5):1196–1206, 2008.

[33] Marta Bertolini, Kevin McElwee, Amos Gilhar, Silvia Bulfone-Paus, and Ralf Paus. Hair follicle immune privilege and its collapse in alopecia areata. *Experimental dermatology*, 29(8):703–725, 2020.

[34] Michael Dudek, Dominik Pfister, Sainitin Donakonda, Pamela Filpe, Annika Schneider, Melanie Laschinger, Daniel Hartmann, Norbert Hüser, Philippa Meiser, Felix Bayerl, et al. Auto-aggressive CXCR6+ CD8 T cells cause liver immune pathology in nash. *Nature*, 592(7854):444–449, 2021.

[35] Alexandre P Bénéchet, Giorgia De Simone, Pietro Di Lucia, Francesco Cilenti, Giulia Barbiera, Nina Le Bert, Valeria Fumagalli, Eleonora Lusito, Federica Moalli, Valentina Bianchessi, et al. Dynamics and genomic landscape of CD8+ T cells undergoing hepatic priming. *Nature*, 574(7777):200–205, 2019.

[36] Yik Chun Wong, Szun Szun Tay, Geoffrey W McCaughan, David G Bowen, and Patrick Bertolino. Immune outcomes in the liver: Is CD8 T cell fate determined by the environment? *Journal of hepatology*, 63(4):1005–1014, 2015.

[37] Beena John and Ian Nicholas Crispe. Passive and active mechanisms trap activated CD8+ T cells in the liver. *The Journal of Immunology*, 172(9):5222–5229, 2004.

[38] Denitra A Breuer, Maria Cristina Pacheco, M Kay Washington, Stephanie A Montgomery, Alyssa H Hasty, and Arion J Kennedy. CD8+ T cells regulate liver injury in obesity-related nonalcoholic fatty liver disease. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 318(2):G211–G224, 2020.

[39] Yi Zhong and Zhandong Liu. Gene expression deconvolution in linear space. *Nature methods*, 9(1):8–9, 2012.

Acknowledgements

Funding

PL, VZ, and WS were supported in part by NIGMS grant R01 GM105785.

Author Contributions

WS, DL, YL supervised the project. PL and WS designed the method, acquired and preprocessed the four datasets used in this paper. PL wrote the software package to performed simulation and real data analyses. VZ provided the geoP software. PL and WS wrote the manuscript, with input from DL, YL, and VZ.