



Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/



Review

eQTL studies: from bulk tissues to single cells



Jingfei Zhang ^a, Hongyu Zhao ^{b,*}

^a Information Systems and Operations Management, Emory University, Atlanta, GA 30322, USA

^b Department of Biostatistics, Yale School of Public Health, New Haven, CT 208034, USA

ARTICLE INFO

Article history:

Received 21 February 2023

Received in revised form

2 May 2023

Accepted 4 May 2023

Available online 18 May 2023

Keywords:

eQTL

Bulk samples

Single cell

Tissues

Cell-type-specific

Context-dependent

ABSTRACT

An expression quantitative trait locus (eQTL) is a chromosomal region where genetic variants are associated with the expression levels of specific genes that can be both nearby or distant. The identifications of eQTLs for different tissues, cell types, and contexts have led to a better understanding of the dynamic regulations of gene expressions and implications of functional genes and variants for complex traits and diseases. Although most eQTL studies have been performed on data collected from bulk tissues, recent studies have demonstrated the importance of cell-type-specific and context-dependent gene regulations in biological processes and disease mechanisms. In this review, we discuss statistical methods that have been developed to enable the detection of cell-type-specific and context-dependent eQTLs from bulk tissues, purified cell types, and single cells. We also discuss the limitations of the current methods and future research opportunities.

Copyright © 2023, The Authors. Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, and Genetics Society of China. Published by Elsevier Limited and Science Press. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Introduction

Recent years have seen significant progress in identifying genomic regions associated with complex traits and diseases through genome-wide association studies (GWAS), where tens of thousands of genomic regions have been associated with thousands of traits, including many complex diseases (see the curated GWAS results at <https://www.ebi.ac.uk/gwas/>). One challenge of interpreting GWAS findings is that most of associated genetic variants, for example, single nucleotide polymorphisms (SNPs), are in intergenic regions, making it difficult to infer functional genes and variants in these regions. Many efforts have been made to annotate the human genome through experimental studies, e.g., ENCODE Project (The ENCODE Project Consortium et al., 2020), Roadmap Epigenomics Project (Roadmap Epigenomics et al., 2015), and psychENCODE (PsychENCODE Consortium et al., 2015), and computational approaches, e.g., CADD (Kircher et al., 2014), GWAVA (Ritchie et al., 2014), GenoCanyon (Lu et al., 2015), GenoSkyline (Lu et al., 2016), EIGEN (Ionita-Laza et al., 2016), GenoSkyline-Plus (Lu et al., 2017), and STARR (Li et al., 2022) to infer the functional roles of different SNPs and other variants, including expression quantitative trait locus (eQTL) studies (The GTEx Consortium, 2020), where the goal is to infer genetic variants affecting genetic regulation by

associating genotypes with gene expression levels across a sample of individuals. Because eQTL studies measure expression levels of all the genes in the genome, they provide an unbiased view of the regulation of gene expression. Using results from eQTL studies in lymphoblastoid cell lines from HapMap samples, it was shown that SNPs associated with complex traits are significantly more likely to be eQTLs identified than minor-allele-frequency-matched SNPs (Nicolae et al., 2010). Another study assessed the enrichment and depletion of variants in different annotation classes (Kindt et al., 2013), including genic regions, regulatory features, measures of conservation, and patterns of histone modifications. It was found that annotations associated with chromatin state and eQTLs were the most enriched groups. These early results stimulated many large community efforts to collect gene expression and genotype data for eQTL studies, and the accumulation of eQTL results parallels the great success of GWAS. Assume an SNP is associated with a complex trait as well as the expression level of a specific gene. In that case, this gene may be implicated as a possible candidate gene for the trait. Several methods have been developed to formalize this idea for colocalization analysis that aims to find the SNPs that are associated with both expression and complex traits (Hormozdiari et al., 2016; Wen et al., 2016; Giambartolomei et al., 2018). Transcriptome-wide association analysis methods have been developed to use eQTL data to predict the expression levels and associate the predicted (imputed) expression levels with the observed complex traits (Gamazon et al., 2015; Gusev et al., 2016; Hu et al., 2019). Mendelian randomization methods have also been

* Corresponding author.

E-mail address: hongyu.zhao@yale.edu (H. Zhao).

Table 1

Representative resources for eQTL studies.

Resource	Sample types	Sample sizes	Link
GTEX	Bulk tissues	73 to 706 across 49 tissues	https://gtexportal.org/home/
eQTL catalog	Bulk tissues from 29 studies	73 to 838 across studies	https://www.ebi.ac.uk/eqlt/
eQTLGen	Blood	31,684 across 37 datasets	https://www.eqtldgen.org
OneK1K	Single cells from PBMC	1.27 million cells from 982 donors	https://onek1k.org

Table 2

Representative statistical methods for detecting cell-type-specific and context-dependent eQTLs.

Samples	Methods	Key ideas	Pros	Cons
Bulk	Westra et al. (2015) Zhernakova et al. (2017) Avila Cobos et al. (2020) Aguirre-Gamboa et al. (2020)	Detect interactions effects between candidate eQTL genotypes and cell-type-specific proxy markers (e.g., cell type proportions) on gene expression levels in bulk tissues	Applicable to large collection of eQTL studies based on bulk samples	Limited resolution for cell types and dependence on informative and robust cell-type-specific proxy markers
Single cells	Cuomo et al. (2022) Strober et al. (2022)	Detect differential effects of candidate eQTL genotypes on gene expression levels for different cell types and/or contexts inferred from single-cell expression data	High-resolution cell types and different molecular contexts	Limited number of subjects available and sparsity in single-cell gene expression data

proposed to investigate whether the expression trait is a causal factor for a complex trait of interest ([Richardson et al., 2020](#); [Yuan et al., 2020](#); [Zhou et al., 2020](#); [Liu et al., 2021](#)) (Tables 1 and 2).

The most well-known eQTL study is the Genotype-Tissue Expression (GTEx) project where dozens of tissues from hundreds of individuals were analyzed to identify tissue-specific eQTLs ([The GTEx Consortium, 2020](#)). The GTEx project has proved to be a valuable resource for the research community. Version 8 of the GTEx analyzed 15,201 RNA-sequencing samples from 49 tissues of 838 postmortem donors. It was found that cis-eQTLs showed 1.46-fold enrichment in the GWAS catalog (<https://www.ebi.ac.uk/gwas/>) where significant GWAS association results are collected. The cross-tissue eQTL similarities were consistent with tissue relatedness, with tissues from the brain region forming one cluster, and other organs being more similar to each other, with the exceptions of testis, lymphoblastoid cell lines, whole blood, and liver that are distinct from other tissues. BLUEPRINT collects genetic, epigenetic, and transcriptomic profiling in three immune cell types to investigate the contributions of different factors in gene expression ([Chen et al., 2016](#)). eQTL catalog is a resource developed by reprocessing data from dozens of studies with more than 30,000 samples, where summary statistics are available for many cell types and tissues ([Kerimov et al., 2021](#)). The results from these studies and resources thus generated have demonstrated the values of eQTL information in inferring causal genes and variants at GWAS loci.

Most eQTL studies to date have been performed on bulk samples, where the estimated effect size of an SNP represents the average effect across different cell types, and the cell-type origin (origins) of the inferred eQTLs is (are) often unknown for a bulk sample consisting of distinct cell types. Despite some successes in using eQTL results to infer disease-causing genes and variants, recent studies based on both modeling ([Yao et al., 2020](#)) and carefully chosen gene-trait pairs ([Connally et al., 2022](#)) have shown that the known eQTLs, which are mostly derived from the analysis of bulk tissues, only explain a very small proportion of the GWAS signals, where GWAS hits colocalize with eQTL SNPs. There is growing evidence (as summarized below) that eQTL effects are often cell-type-specific and/or context-dependent, and many of the eQTLs uniquely identified through cell-type-specific and context-dependent analysis (either experimentally or computationally) colocalize with GWAS results ([Aguirre-Gamboa et al., 2020](#); [Donovan et al., 2020](#); [Patel et al., 2021](#)), suggesting the importance of cell-type-specific and context-

dependent eQTLs for interpreting and understanding GWAS signals. Therefore, there is a great need to identify these additional eQTLs missed from tissue-based analysis to expand the space of eQTLs and make more informed inferences on disease-causing genes and variants.

To facilitate the identifications of cell-type-specific and context-dependent eQTLs, statistical methods have been developed for both bulk samples through digital deconvolution analysis, and for single-cell data, which offer finer cell-type resolutions and can capture dynamic effects of eQTLs. We illustrate three different data types that can be used for inferring cell-type-specific and context-dependent eQTLs (Fig. 1). In this review, we discuss existing statistical methods that use bulk tissues and single-cell data to identify cell-type-specific and context-dependent eQTLs, showing high-level analysis pipelines for bulk tissues consisting of distinct cell types and single cells (Fig. 2), with details in the next section. We then summarize results from empirical studies using bulk samples, purified cells, and single-cell data. We conclude with the limitations of the existing computational methods and future methodological needs.

Analytical approaches for cell-type-specific and context-dependent eQTL inference

eQTL inference using bulk samples

Early eQTL studies collected gene expression data using microarrays, where gene expression levels need to be normalized to remove batch effects, and the normalized data are analyzed to identify eQTLs. Consider a study with N subjects, S SNPs, and G genes. For the n th subject, y_{ng} denotes the expression level of the g th gene, and x_{ns} denotes the genotype of the s th SNP. For a SNP with two alleles, say A and a, its three genotypes AA, Aa, and aa can be coded as 2, 1, and 0, respectively. We can study the relationship between the observed gene expression level y_{ng} and genotype x_{ng} through the following regression model

$$y_{ng} = \beta_g + \beta_{gs}x_{ns} + \varepsilon_{ngs}, \quad (1)$$

where β_g is the intercept, β_{gs} is the effect of the s th SNP on the expression of the g th gene, and ε_{ngs} is the error term, often assumed to follow a normal distribution. A more comprehensive model may

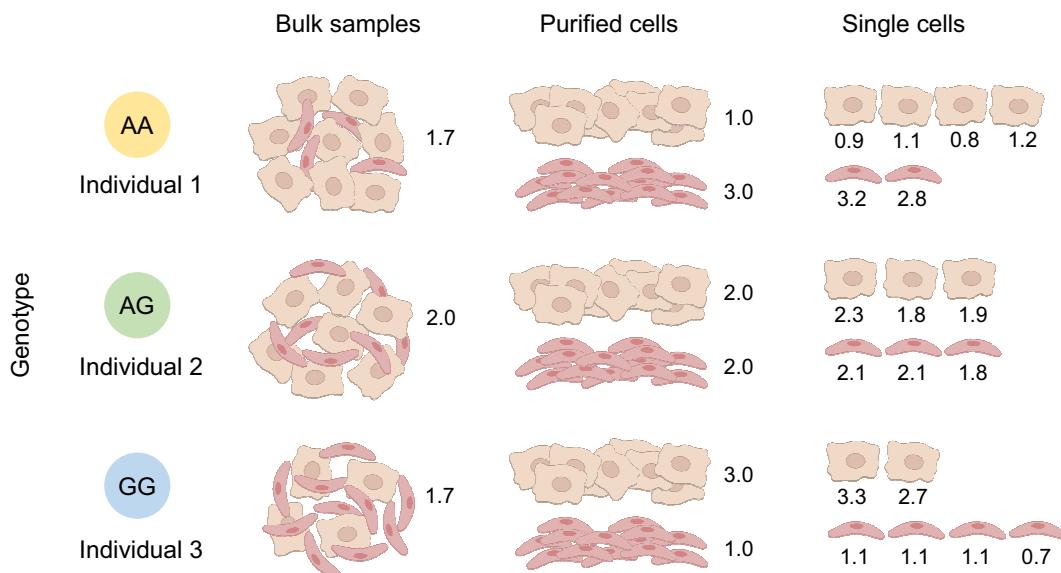


Fig. 1. Illustration of eQTL analysis at different resolutions: single cells, purified cells, and bulk samples. Shown are data from three individuals with genotypes of AA, AG, and GG, respectively. Two cell types make up the bulk samples, the oval-shaped cells, and the triangle-shaped cells. For single-cell data, we can observe expression level at the single-cell level. For example, for the first individual with genotype AA, there are four oval-shaped cells with expression levels of 0.9, 1.1, 0.8, and 1.2, and two triangle-shaped cells with expression levels of 3.2 and 2.8, respectively. eQTL analysis can be performed for two cell types separately using single cells across these three individuals to correlate genotypes with observed single-cell level gene expression data. For data from purified cells, we observe aggregated gene expression levels for different cell types but without individual cell level measurements. The average expression level for the oval-shaped cells is 1, 2, and 3, respectively, for the three individuals. For data from bulk samples, we can no longer distinguish contributions from two distinct cell types. The average expression level for the three individuals is 1.7, 2.0, and 1.7, respectively. For single-cell data, not only we can study the association between genotypes and cell-type-specific expressions, but also we can correlate genotypes with cell-type proportions. Through deconvolution methods, the bulk samples may be deconvoluted to different cell types to allow cell-type-specific eQTL analysis with estimated cell type proportions from different individuals.

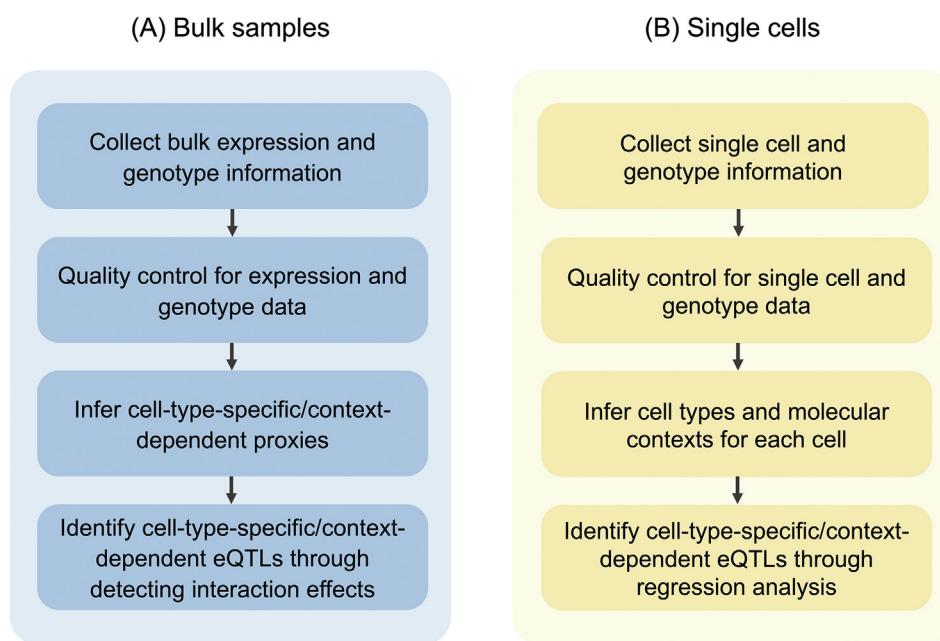


Fig. 2. General pipeline for (A) bulk-sample-based and (B) single-cell-based analysis to identify cell-type-specific and context-dependent eQTLs.

also include other covariates, such as age and sex. Testing the null hypothesis that the s th SNP does not affect the expression level of the g th gene is equivalent to testing $\beta_{gs} = 0$. A typical eQTL study considers more than 20,000 genes and up to millions of SNPs. Because of the large number of SNPs to be tested, researchers often

focus on cis-eQTLs for a given gene, which are SNPs in close physical proximity, say within one million base pairs of the candidate gene. In contrast, trans-eQTLs correspond to SNPs that are on different chromosomes or further away from the gene of interest on the same chromosome. Most eQTL findings have been for cis-eQTLs

largely due to statistical power differences in detecting cis-eQTLs and trans-eQTLs. With a few hundred samples, which is the typical size of an eQTL study, there is limited power to do a genome-wide association study required to identify trans-eQTLs, which often have smaller effect sizes than cis-eQTLs. Even for cis-eQTL analysis, hundreds or thousands of SNPs often need to be considered, and multiple comparison adjustments must be done to appropriately control false-positive findings. Several computational tools have been developed and commonly used for eQTL analysis in bulk samples, such as MatrixEQTL (Shabalin, 2012) and FastQTL (Ongen et al., 2016).

The regression setting in (1), where the errors are assumed to be Gaussian, is reasonable for microarray-based gene expression measurements. However, with gene expression data collected through RNAsequencing, such as those from the GTEx project, the measured gene expression level is the total number of sequence reads mapped to a specific gene, which needs to be adjusted for total sequencing depth and other factors. These data may be better modeled by other distributions, for example, negative binomial, while accounting for factors that may impact the observed sequencing reads. In this case, a generalized linear regression model may be more appropriate than (1) and may also have better statistical power, although it may be computationally more expensive.

For RNA sequencing, there is added benefit of observing alternative splicing and allele-specific expression. In the case of allele-specific expression, consider the presence of a SNP in the transcribed region of a gene with two alleles *B* and *b*, and the simple scenario that all the sequence reads contain this SNP. For heterozygous individuals with genotype *Bb*, a sequence read covering this SNP may either have *B* or *b*. In one extreme case, all the sequence reads may only contain *B* but not *b*. Even in the absence of measured total gene expression levels for homozygous individuals with genotypes *BB* and *bb*, the imbalance between the mapped sequence reads having *B* and *b* suggests the presence of cis-eQTLs, either the SNP with alleles *B* and *b* itself or some SNP with perfect dependence with this SNP, that regulates the expression level of this gene. Statistical models have been proposed to explicitly incorporate this allelic-specific expression to identify cis-eQTLs, including TReCASE (Sun, 2012), RASQUAL (Kumasaka et al., 2016), and mixQTL (Liang et al., 2021). It was found that considering allelic-specific expression could identify 20%–100% more genes with eQTLs across 28 tissues in the GTEx project than only considering total expression levels using TreCASE (Zhabotynsky et al., 2022), and the power gain of mixQTL was equivalent to a 29% increase in sample size for genes with sufficient allele-specific read coverage (Liang et al., 2021).

The analysis of tissue-level data also allows for the investigation of context-dependent eQTLs if the context can be well defined. For example, 369 sex-biased eQTLs were inferred through separate analyses of male and female GTEx samples (The GTEx Consortium, 2020), where the sex of an individual may be considered a context. Furthermore, 178 population-biased eQTLs were also implicated, where population origin may be regarded as another context. Other context-dependent effects can be considered by including an interaction term between the context variable and the SNP genotype in the regression model (1).

Cell-type-specific eQTL inference using bulk samples

Several studies have been published that used purified cells of different cell types to infer cell-type-specific eQTLs (ct-eQTLs). As the gene expression data in these samples are collected in the same manner as bulk tissue samples, the same statistical methods for bulk samples can be applied to infer ct-eQTLs from these data. However, the sample size tends to be smaller, and the measurement noises

may be higher. In addition, the purified cell types may be contaminated with other cell types.

Without collecting data from purified cell types, Westra et al. (2015) proposed identifying ct-eQTLs by investigating whether there is an interaction effect between the surrogate score for a cell type and candidate SNP's genotype on bulk gene expression levels from the collected samples. More formally, this model can be written as

$$Y_{ng} = \beta_g + \beta_{gs}x_{ns} + \beta_{gm}m_n + \beta_{g,sm}(x_{ns} \times m_n) + \varepsilon_{ngs}, \quad (2)$$

with two additional terms $\beta_{gm}m_n$ and $\beta_{g,sm}(x_{ns} \times m_n)$ compared to model (1), where m_n is a proxy marker for the cell type of interest in the n th individual, β_{gm} is the effect of the proxy marker on the expression level of the g th gene, and $\beta_{g,sm}$ is the interaction effect between genotype x_{ns} and proxy marker m . A significant interaction effect, that is, $\beta_{g,sm} \neq 0$, is interpreted as the cell-type-specific effect of SNP s on expression of gene g . The same approach was used to study context-dependent eQTLs by Zhernakova et al. (2017).

Instead of deriving cell-type-specific proxy markers or enrichment scores, the estimated cell type proportions can also be used as a proxy for a given cell type. Recent years have seen the developments of many methods to deconvolute bulk RNA-seq samples to infer proportions of different cell types and cell-type-specific expression levels (Avila Cobos et al., 2020). For the n th subject, π_{nk} denotes the estimated proportion of the k th cell type for this individual, where there is a total of K cell types. We can use the following regression model to detect ct-eQTLs for the k th cell type.

$$Y_{ng} = \beta_g + \beta_{gs}x_{ns} + \beta_{gk}\pi_{nk} + \beta_{g,sk}(X_{ns} \times \pi_{nk}) + \varepsilon_{ngs}. \quad (3)$$

In this model, β_{gk} is the cell type proportion effect from the k th cell type, and $\beta_{g,sk}$ is the interaction effect between the s th SNP and the proportion of k th cell type. A non-zero $\beta_{g,sk}$ suggests a cell-type-specific effect for the s th SNP.

The formulations in (2) and (3) consider one cell type at a time and ignore the contributions of possible cell-type-specific effects from other cell types, both in terms of proportions and expression profiles, leading to a potential loss of information. Moreover, models (2) and (3) only consider a tissue and cell type pair at a time and may not attribute a non-zero $\beta_{g,sk}$ to the correct cell type. For example, consider the case of two cell types, where $k = 1$ or 2 . If $\beta_{g,s1} > 0$, then $\beta_{g,s2} < 0$ due to the constraint that $\pi_{n1} + \pi_{n2} = 1$. Furthermore, the power differs across cell types, with a higher statistical power in detecting ct-eQTLs for more abundant cell types. A more comprehensive model that takes into account all cell types simultaneously can be formulated as

$$Y_{ng} = \beta_g + \beta_{gs}x_{ns} + \sum_{k=1}^K \beta_{gk}\pi_{nk} + x_{ns} \left(\sum_{k=1}^K \beta_{g,sk} \times \pi_{nk} \right) + \varepsilon_{ng}, \quad (4)$$

subject to the constraint that $\sum_k \pi_{nk} = 1$. Correspondingly, the s th SNP is a ct-eQTL for the k th cell type if $\beta_{g,sk} \neq 0$. Another way to parametrize this model is in the form of

$$Y_{ng} = \sum_{k=1}^K (\beta_{gk} + \beta_{g,sk} \times x_{nk})\pi_{nk} + \varepsilon_{ng}. \quad (5)$$

Note that the $\beta_{gk} + \beta_{g,sk} \times x_{nk}$ term in model (5) is essentially the cell-type-specific gene expression for the k th cell type in sample n , essentially the same model considered in Decon-eQTL (Aguirre-Gamboa et al., 2020).

In practice, ct-eQTL analysis based on the above models often uses transformed gene expression data instead of read counts. This

may distort the association between the observed gene expression level and cell type compositions, leading to reduced power and inflated false positives. Recently, Little et al. (2022) proposed ct-eQTL to jointly model total read counts and allele-specific counts by a negative binomial (or Poisson) and a beta-binomial (or binomial) distribution with the consideration of covariates, cell type composition, and SNP genotype. ct-eQTL also includes an allele-specific expression to further increase the power to detect cell type-specific eQTLs. Empirical studies showed higher power of ct-eQTL than linear model-based methods. DeCAF is a linear model-based method that considers both total expression levels and allele-specific expression (Kalita and Gusev, 2022).

Although the above approaches are intuitive, applying them to infer ct-eQTLs in practice has challenges. First, there are uncertainties in the estimated proxy markers and cell type proportions, and these need to be appropriately incorporated into the analysis. However, this issue has only recently been studied (Xie and Wang, 2022), and the impact of incorporating these uncertainty estimates in ct-eQTL inference needs to be explored. Second, although ct-eQTLs may be inferred for all cell types in principle, it would be relatively easier for more abundant cell types than for less abundant or rare cell types. Third, there have to be sufficient variations in cell type compositions across subjects to allow ct-eQTL inference. For example, in the extreme case that all the subjects have identical cell type proportions, the parameters in the above models are not identifiable. Fourth, the above formulation does not consider similarity among some cell types, although methods have been proposed to consider cell lineage (Yankovitz et al., 2021).

Cell-type-specific eQTLs from single-cell data

In addition to bulk data, single-cell data are increasingly used for ct-eQTL inference (Jonkers and Wijmenga, 2017; van der Wijst et al., 2018; Liu et al., 2021; Neavin et al., 2021). Most published single-cell-based ct-eQTL analyses are performed by analyzing pseudo-bulk RNA-seq data for different cell types, where the single-cell data are first annotated to distinct cell types, and the cells annotated to the same cell types from a specific subject are combined to derive cell-type-specific gene expression levels. eQTL methods for bulk samples can then be applied to detect ct-eQTLs. For example, Yazar et al. (2022) grouped cells of the same type for each individual and adjusted for covariate effects before performing Spearman rank correlation analysis. However, the sample size is still much more limited for single-cell data compared to that of bulk samples, and there are many ongoing efforts for single-cell-based genetic association analysis, for example, the single-cell eQTLGen consortium (van der Wijst et al., 2018) and the OneK1K cohort (Yazar et al., 2022).

Instead of aggregating all the cells in a given individual, Nathan et al. (2022) used Poisson mixed effects regression to model the effects of SNPs, cell states (which can be both discrete and continuous), batch structure, and other covariates (such as sex, age, genotype principal components and gene expression principal components, and percentage of mitochondrial unique molecular identifiers [UMIs]) on the observed gene expression level measured by UMI counts at the single-cell level. The effect of an SNP is modeled as a fixed effect in the analysis. When the Poisson mixed effects model was compared with the computationally more expensive negative binomial mixed effects model, it was found that the Poisson model was adequate for the single-cell data analyzed.

Although single cells can be grouped into pre-defined cell types for ct-eQTL analysis, the very high resolution at the single-cell level offers the opportunity for more refined analysis, where the individual cells can be characterized by a vector of continuous contexts. For example, principal component analysis can be performed for the highly variable genes across all the cells based on normalized gene

expression data, and the top principal components for a single cell may be taken as the cellular states for this cell. After the cellular states are defined for a single cell, the effects of an SNP on gene expression may be studied in the context of these cellular states to see whether the effects may vary depending on different states. Assuming we have N subjects, with m_n cells collected from the n th subject, and a total of C different cellular contexts are defined for each cell. Let the states of i th cell for the n th subject be denoted by a vector of contexts $(h_{ni1}, h_{ni2}, \dots, h_{nIC})$ of dimension C . Cuomo et al. (2022) proposed a cellular regulatory map model, called CellRegMap, as

$$y_{ngi} = \beta_g + \beta_{gs}x_{ns} + \beta_{g,si}x_{ns} + u_{ng} + c_{ngi} + \epsilon_{ngi}, \quad (6)$$

where y_{ngi} represents the measured expression level of the g th gene in the i th cell of the n th subject, x_{ns} is the genotype of the s th SNP of the n th individual, β_g is the baseline expression level, β_{gs} represents the persistent effect of the s th SNP across all the cells in different subjects, $\beta_{g,si}$ is the cell-specific effect on the g th expression level, u_{ng} accounts for the fact that the m_n cells are from the same subject, c_{ngi} accounts for the cell context effects, and ϵ_{ngi} is the error term. CellRegMap adopts an overall random effects model approach where $\beta_{g,si} \sim N(0, \sigma_{S \times C}^2 \Sigma)$, $u_{ng} \sim N(0, \sigma_R^2 \Sigma)$, $c_{ngi} \sim N(0, \sigma_C^2 \Sigma)$, and $\epsilon_{ngi} \sim N(0, \sigma_e^2)$. The matrix Σ is defined by the cellular context vectors $\Sigma = HH^T$. CellRegMap uses a score test to investigate whether an SNP has a context-dependent effect on gene expression level with the null hypothesis $\beta_{g,si} = 0$. This model can also be used to test the main effect and estimate the allelic effects of single cells for each gene-SNP pairs based on the best linear unbiased predictor. In practice, it is important to define cellular contexts, and CellRegMap used MOFA to define cellular states (Argelaguet et al., 2018), where latent factors are inferred from single-cell data that explain variation in gene expression in the data. Because of the computational issues and the assumption of normal errors, the single cells were aggregated into meta cells because of the sparsity in single-cell data in real data analysis. In addition, only specific gene-SNP pairs were considered due to statistical power concerns.

Strober et al. (2022) proposed a similar approach, called single-cell unsupervised regulation of gene expression (SURGE), where a continuous representation of the cell contexts is learned through a probabilistic model with matrix factorization. The model has a form similar to that of CellRegMap as follows:

$$y_{ngi} = \beta_g + \beta_{gs}z_{ns} + \sum_{c=1}^C h_{nic}\beta_{gsc}z_{ns} + u_{ng} + \epsilon_{ngi}, \quad (7)$$

where y_{ngi} is standardized gene expression level for the g th gene in the i th cell of the n th subject, and z_{ns} is the standardized genotype for the s th SNP of the n th individual. The other parameters have the same meaning as the CellRegMap model, but the context vector h_{nic} is latent and learned from the model instead of prespecified. SURGE has the following assumptions about the model parameters: $\beta_{gs} \sim N(0, 1)$, $h_{nic} \sim N(0, \sigma_c^2)$, $\beta_{gsc} \sim N(0, 1)$, $1/\sigma_c^2 \sim Gamm(\alpha_0, \beta_0)$, $u_{ng} \sim N(0, \psi_{gs}^2)$, $\epsilon_{ngi} \sim N(0, \sigma_{gs}^2)$, $1/\psi_{gs}^2 \sim Gamm(\alpha_0, \beta_0)$, and $1/\sigma_{gs}^2 \sim Gamm(\alpha_0, \beta_0)$. SURGE approximates the posterior distribution of all latent variables using mean-field variational inference. Similar to CellRegMap, only eQTLs identified in previous studies were used for analysis due to statistical power concerns, and the reliance on the normal distribution assumption of the error terms limits its direct application to single-cell data. As a result, single cells have to be aggregated to meta cells before the model is applied.

Empirical results on cell-type-specific and context-dependent eQTL analyses

This section reviews the growing evidence of cell-type-specific and context-dependent eQTLs using bulk samples, purified cells, and single cells.

Tissue analysis

Whole blood

Among the first analysis of cis-eQTLs using bulk samples, [Westra et al. \(2015\)](#) analyzed whole blood gene expression data of 5683 individuals from seven cohorts to infer cell-type-specific cis-eQTLs. A total of 1115 cis-eQTLs (8.5% of the significant cis-eQTLs from prior eQTL analysis for the whole tissue) were found to have significant interaction effects with neutrophil proxy. The results were replicated in six individual purified cell-type eQTL datasets. More importantly, the authors showed SNPs associated with Crohn's disease preferentially affect gene expression within neutrophils, demonstrating the insights gained from cell-type-specific eQTL analysis. [Zhernakova et al. \(2017\)](#) performed eQTL and context-dependent eQTL analysis on RNA-seq data of peripheral blood from 2116 unrelated individuals, identifying 23,060 genes with eQTLs, among which 2743 (12%) showed context-dependent effects.

GTEX data

Cell-type-specific analysis was performed on the GTEx data ([Kim-Hellmuth et al., 2020](#)). The authors estimated cell type enrichment for seven cell types (adipocytes, epithelial cells, hepatocytes, keratinocytes, myocytes, neurons, and neutrophils) across 35 tissues. Between 43 pairs of tissues and cell types, they identified eQTLs specific to at least one cell type by testing for interaction effects between SNP and cell type enrichment on the observed expression levels. They found that these cell-type-interaction QTLs, called ieQTLs, are enriched for genes with tissue-specific eQTLs and generally not shared across unrelated tissues. Furthermore, these ieQTLs are enriched for complex trait associations and had colocalization signals for hundreds of undetected loci in bulk tissue.

Cultured and purified cells

Brain cells

[Aygun et al. \(2021\)](#) used a cell-type-specific in vitro model system including 85 neural progenitors and 74 virally labeled and sorted neuronal progeny for eQTL analysis. They identified 2079 and 872 eQTLs in progenitors and neurons, respectively, with 66% and 47% of these eQTLs not identified in fetal bulk brain eQTLs from a largely overlapping sample or in adult data from GTEx. These eQTLs had cell-type-specific colocalizations with GWAS hits for neuropsychiatric disorders and other brain-related traits.

Microglia in the brain play critical roles in immune defense and development and are implicated in neurodegenerative disorders. [Young et al. \(2021\)](#) gathered gene expression profiles in primary microglia isolated from 141 patients undergoing neurosurgery. A total of 585 microglia eQTLs were identified. Through joint analysis with monocytes and IPSDMac, 855 microglia eQTLs were inferred, with 108 microglia specific, and 449 shared across three cell types. For colocalization with GWAS hits, there was an excess of colocalized microglial eQTLs for Alzheimer's disease, Parkinson's disease, and inflammatory bowel disease.

Melanocyte cultures

Because melanocytes give rise to melanoma but account for less than 5% of human skin biopsies, [Zhang et al. \(2018\)](#) performed eQTL

analysis in primary melanocyte cultures from 106 newborn males to identify eQTLs in melanocytes. The identified melanocyte eQTLs differed considerably from those from the GTEx tissues, including the skin. Novel risk genes for melanoma were implicated using the transcriptome-wide association study based on this dataset.

Immune cells

The DICE project isolated 13 immune cell types from 106 leukapheresis samples of 91 healthy subjects ([Schmiedel et al., 2018](#)). It was found that eQTLs are highly cell-type specific, and sex has a major effect on gene expression. In the ImmuNexUT study, with samples from 79 healthy controls and 337 patients diagnosed with different immune-mediated diseases, [Ota et al. \(2021\)](#) purified 28 immune cell types from these individuals with a total of 9852 samples and performed cell-type-specific eQTL analysis. They identified a median of 7092 genes with eQTLs in each cell type, 2.2-fold more than that identified in the DICE study ([Schmiedel et al., 2018](#)). They further identified eQTLs that were only present in patients.

Single-cell analysis

Peripheral blood mononuclear cells (PBMC)

In a proof-of-concept study, [van der Wijst et al. \(2018\)](#) analyzed 25,000 single-cell RNA-seq data from 45 donors. They identified 379 unique cis-eQTLs involving 287 unique eGenes across six cell types. A total of 48 cis-eQTLs were only identified from cell-type-specific analysis. The authors also demonstrated the benefit of performing cell subtype analysis for cMonocytes and ncMonocytes.

[Oelen et al. \(2022\)](#) exposed Peripheral blood mononuclear cells (PBMC) samples from 120 individuals to three pathogens. They sequenced these samples in an unstimulated condition and after 3 h and 24 h in vitro stimulation for the three pathogens. They identified cell-type-specific eQTLs, with the number of such eQTLs correlated with the cell-type abundance. Furthermore, the effects of eQTLs differed across pathogen stimulations, and the strongest enrichment for GWAS signals was observed for eQTLs that were identified from stimulation experiments.

The investigators of the OneK1K cohort analyzed 1.27 million PBMC single-cell RNA-seq data from 982 donors of Northern European ancestry and performed eQTL analyses on 14 immune cell types ([Yazar et al., 2022](#)). A total of 26,597 cis-eQTLs were identified, with most having cell-type-specific effects. Dynamic effects were also observed based on the pseudo-time trajectory for the B cell landscape. In addition to cis-eQTLs, 990 trans-eQTLs were identified, with most genes regulated by trans-eQTLs being specific for a single-cell type, and none were ubiquitous across cell types. Colocalization analysis between eQTLs and GWAS signals suggested that 60% of colocalizing genes were detected upon activation, and colocalization is very cell type-specific.

Induced pluripotent stem cells (iPSCs)

[Neavin et al. \(2021\)](#) gathered single-cell RNA-seq data from 64,018 fibroblasts of 79 donors and performed single-cell eQTL analysis. For the six types of fibroblasts and four types of induced pluripotent stem cells (iPSCs), most of the detected eQTLs in fibroblasts were specific to one cell type. Only 41% of the 45,503 eQTLs identified in the six fibroblast types were significant in GTEx.

Using 125 iPSC lines derived from 125 donors, [Cuomo et al. \(2020\)](#) collected single-cell gene expression data from 36,044 cells at four differentiation time points using full-length RNA-sequencing and the expression levels of selected cell surface markers through cell sorting. Substantial regulatory changes were observed, with over 30% of eQTLs being specific to a single stage. Hundreds of eQTLs at the mesendo and defendo stages were new. This study also tested

for associations between pseudo-time and the genetic effect size using a linear model, as well as identified 899 time-dynamic eQTLs.

T cells

Soskic et al. (2022) analyzed 655,349 CD4⁺ T cells from 119 healthy donors, both for unstimulated cells and at three time points after cell activation. Different numbers of genes showing eQTL effects were detected at different time points with hundreds of them only detected at specific cell states. Using pseudo-time trajectory information, 2265 genes were found to have dynamic eQTL effects, representing about one-third of the genes. Colocalized genes with GWAS signals were enriched in time-dependent eQTLs.

With 89 healthy donors, Schmiedel et al. (2022) performed eQTL analysis for more than one million activated CD4⁺ T cells classified into 19 distinct CD4⁺ T cell subsets. The effects of many eQTLs were strongly manifested only in specific cell types in an activation-dependent manner, and significant sex effects were also observed.

Nathan et al. (2022) performed single-cell eQTL analysis using gene expression data from more than 500,000 unstimulated memory T cells from 259 Peruvian individuals. They found that the effects of one-third of cis-eQTLs were mediated by continuous multimodally defined cell states, with independent eQTLs at some loci having opposing cell-state relationships.

Brain cortex

In a recent study using single-cell data from the prefrontal cortex, temporal cortex, and deep white matter from 192 individuals, a total of 7607 genes were found to have eQTLs from eight cell types (Bryois et al., 2022). A majority of cell-type-specific eQTLs were replicated in tissue-level eQTLs for cortical tissue, with eQTLs for more abundant cell types more likely replicated in the tissue results. It was also found that the effect sizes estimated from tissue tend to be lower than those estimated from cell-type-specific analysis. As expected, the number of cis-eQTLs identified in a cell type was strongly correlated with the number of cells available for the corresponding cell type. The effect sizes were more similar for similar cell types with microglia being most different from others. Colocalization analysis suggests that disease risk at a given GWAS locus is usually mediated by a single gene acting in a specific cell type.

Dopaminergic neuron differentiation

Jerber et al. (2021) differentiated 215 human iPSC lines to profile over 1 million cells across four conditions, including three differentiation stages (progenitor-like, young neurons, and more mature neurons) and cells exposed to a chemical stressor. eQTL analysis was performed for 14 cell types, identifying 4828 genes with eQTLs. Compared to eQTLs identified from GTEx brain tissues, this study identified 2366 new eQTLs. As for colocalization analysis, 1284 eQTLs were colocalized, with 597 being new, and 67% of these new colocalizations were associated with eQTLs detected in later differentiation stages or upon stimulation. A colocalization using aggregated data from different cell types yielded a much smaller number of colocalizations, suggesting the importance of considering cell type specificity.

Discussion

A main driving force for many eQTL studies in recent years is their potential to offer insights into the GWAS signals, which mostly fall into non-coding regions of the human genome. Because of the importance of cell-type-specific and context-dependent eQTLs, a growing number of studies are collecting and analyzing data to facilitate such analyses. Coupled with the rise of rich data that offer cell-type-specific and context-dependent gene regulation information, there is also a need for more statistically robust and

computationally efficient methods for these analyses. In this paper, we have reviewed statistical methods that have been developed and applied to analyze different types of data for cell-type-specific and context-dependent eQTL inference, including bulk samples, purified cells, and single cells.

Despite these progresses, many issues remain to be resolved, especially in anticipation of the many population-level single-cell data to be gathered in the near future that will involve tens of thousands of individuals and tens of millions of single cells across different tissues. For example, to fully respect the nature of the single-cell data, the single data are more appropriately modeled as Poisson or negative binomial distribution. Yet, most published studies have adopted linear regression models and often aggregated multiple cells to address the sparsity of the observed single-cell data. For bulk samples using RNA sequencing, efforts have been made to use allele-specific expression to improve statistical power for identifying eQTLs, and limited work has been done for single-cell data to capitalize on allele-specific expression information. There are challenges in appropriately defining cell types and contexts for cell-type-specific and context-dependent eQTL analysis. Cuomo et al. (2022) and Strober et al. (2022) represent some initial efforts and more needs to be done to fully capture and utilize the cell state information for eQTL discoveries, including non-linear transcription programs (Wang and Zhao, 2022). In addition, as more than one SNP may jointly affect expression levels (The GTEx Consortium, 2020; Abell et al., 2022), methods that include joint effects are likely to be more powerful and can better characterize the relationship between gene expression levels and SNPs.

With many studies performed on bulk tissues, and the availability of several methods to use bulk tissue samples for cell-type-specific eQTL analysis, there is a need to effectively integrate the results from single-cell data and bulk tissue data. Furthermore, as different tissues may share cells of similar cell types and states, there is also a need to better integrate results across tissues and studies (Flutre et al., 2013; Urbut et al., 2019).

As for all genetic studies, study design is important, such as the number of samples to be collected, and, in the case of single cells, the number of cells per subject and the sequencing depth of each cell. Several software and tools have been developed to facilitate this analysis under relatively simple statistical models for data analysis (Mandric et al., 2020; Dong et al., 2021). Further developments are needed to incorporate more comprehensive statistical models for analysis, and the consideration of other information, for example, alleles-specific expression, in inferring eQTLs. Due to the limited statistical power, most studies have focused on cis-eQTLs. For example, fewer than 150 genes were found to be affected by trans-eQTLs in the GTEx project (The GTEx Consortium, 2020). There is a critical need to design statistical methods to identify trans-eQTLs, both for bulk tissues and cell-type-specific and context-dependent effects.

We have focused on the inference of eQTLs in this paper. There are many downstream applications of eQTL analysis. For example, methods need to perform colocalization analysis with cell-type-specific and context-dependent eQTL results without focusing on a specific set of SNPs through simple thresholding of statistical significance, while accounting for the linkage disequilibrium information across the SNPs in a region. Cell-type-specific and context-dependent eQTLs also offer the opportunity to improve the identifications of candidate genes through transcriptome-wide association studies at the cell-type and context-dependent level, and the recently developed methods (Okamoto et al., 2023; Song et al., 2023) for bulk samples can be extended for cell-type-specific and context-dependent analyses. There is also the need for Mendelian randomization methods to infer the causal relationship between transcript levels and complex traits and diseases (Liu et al., 2021). In this

setting, the existing methods for bulk samples may not be adequate to deal with the count nature of the single-cell data, data sparsity, and the need to integrate information from multiple data sources for more informed analysis and decisions. In addition to eQTLs that affect gene expression levels, genetic variants can also affect gene expression variances and co-variances (Hulse and Cai, 2013; Ek et al., 2018; Sarkar et al., 2019; Marderstein et al., 2021). Compared to eQTL analysis, more needs to be done to identify such variants. Even less has been explored at the cell type level, where cell-type-specific co-expressions may be inferred using either bulk samples (Su et al., 2022b) or single-cell data (Su et al., 2022a). A comprehensive catalog of eQTLs with cell-type-specific and context-dependent effects on gene expression variance and co-variance will better characterize the gene regulation of expressions and interpret GWAS results.

Although gene expression has been the focus of cell-type-specific and context-dependent analysis, other data types are being increasingly collected for similar analysis, such as methylation, chromatin accessibility, and proteomics, based on single-cell data (Wang et al., 2022). The methods and tools developed for bulk and single-cell RNA-seq data may also apply to other data types, such as recently published methylation data from the GTEx subjects (Oliva et al., 2023) and a meQTL dataset derived from primary melanocytes of 106 individuals (Zhang et al., 2021). More importantly, as different data types reflect different aspects of the same biological process, there is a need to integrate data from different modalities to assess the genetic effects of SNPs on gene expression, methylation, chromatin accessibility, protein expression, and other molecular phenotypes. Such integrated analysis will likely yield more informative annotations of the SNPs to facilitate the interpretation of the GWAS results (Hormozdiari et al., 2018).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

Zhang's research is supported by NSF DMS-2015190 and DMS-2210469. Zhao's research is supported in part by NIH R01 GM134005 and R56 AG074015.

References

- Abell, N.S., DeGorter, M.K., Gloudemans, M.J., Greenwald, E., Smith, K.S., He, Z., Montgomery, S.B., 2022. Multiple causal variants underlie genetic associations in humans. *Science* 375, 1247–1254.
- Aguire-Gamboa, R., de Klein, N., di Tommaso, J., Claringbold, A., van der Wijst, M.G., de Vries, D., Brugge, H., Oelen, R., Vosa, U., Zorro, M.M., et al., 2020. Deconvolution of bulk blood eQTL effects into immune cell subpopulations. *BMC Bioinform.* 21, 243.
- Angelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., Stegle, O., 2018. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* 14, e8124.
- Avila Cobos, F., Alquicira-Hernandez, J., Powell, J.E., Mestdagh, P., De Preter, K., 2020. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* 11, 5650.
- Aygun, N., Elwell, A.L., Liang, D., Lafferty, M.J., Cheek, K.E., Courtney, K.P., Mory, J., Hadden-Ford, E., Krupa, O., de la Torre-Ubieta, L., et al., 2021. Brain-trait-associated variants impact cell-type-specific gene regulation during neurogenesis. *Am. J. Hum. Genet.* 108, 1647–1668.
- Bryois, J., Calini, D., Macnair, W., Foo, L., Urich, E., Ortmann, W., Iglesias, V.A., Selvaraj, S., Nutma, E., Marzin, M., et al., 2022. Cell-type-specific *cis*-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat. Neurosci.* 25, 1104–1112.
- Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martin, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al., 2016. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* 167, 1398–1414.
- Connally, N.J., Nazeen, S., Lee, D., Shi, H., Stamatoyannopoulos, J., Chun, S., Cotsapas, C., Cassa, C.A., Sunyaev, S.R., 2022. The missing link between genetic association and regulatory function. *Elife* 11, e74970.
- Cuomo, A.S.E., Heinen, T., Vagiaki, D., Horta, D., Marioni, J.C., Stegle, O., 2022. CellRegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq. *Mol. Syst. Biol.* 18, e10663.
- Cuomo, A.S.E., Seaton, D.D., McCarthy, D.J., Martinez, I., Bonder, M.J., Garcia-Bernardo, J., Amatya, S., Madrigal, P., Isaacson, A., Buettner, F., et al., 2020. Publisher Correction: single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* 11, 1572.
- Dong, X., Li, X., Chang, T.W., Scherzer, C.R., Weiss, S.T., Qiu, W., 2021. powerEQTL: an R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis. *Bioinformatics* 37, 4269–4271.
- Donovan, M.K.R., D'Antonio-Chronowska, A., D'Antonio, M., Frazer, K.A., 2020. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* 11, 955.
- Ek, W.E., Rask-Andersen, M., Karlsson, T., Enroth, S., Gyllensten, U., Johansson, A., 2018. Genetic variants influencing phenotypic variance heterogeneity. *Hum. Mol. Genet.* 27, 799–810.
- Flutre, T., Wen, X., Pritchard, J., Stephens, M., 2013. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* 9, e1003486.
- Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Elyer, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L., et al., 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098.
- Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A.E., CommonMind, C., Pasaniuc, B., et al., 2018. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* 34, 2538–2545.
- Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al., 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252.
- Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H.K., Ju, C.J., Loh, P.R., Schoech, A., Reshef, Y., Liu, X., O'Connor, L., et al., 2018. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* 50, 1041–1047.
- Hormozdiari, F., van de Bunt, M., Segre, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., Eskin, E., 2016. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* 99, 1245–1260.
- Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S.M., Yu, Z., Li, B., Gu, J., Muchnik, S., et al., 2019. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.* 51, 568–576.
- Hulse, A.M., Cai, J.J., 2013. Genetic variants contribute to gene expression variability in humans. *Genetics* 193, 95–108.
- Ionita-Laza, I., McCallum, K., Xu, B., Buxbaum, J.D., 2016. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* 48, 214–220.
- Jerber, J., Seaton, D.D., Cuomo, A.S.E., Kumashiro, N., Haldane, J., Steer, J., Patel, M., Pearce, D., Andersson, M., Bonder, M.J., et al., 2021. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* 53, 304–312.
- Jonkers, I.H., Wijmenga, C., 2017. Context-specific effects of genetic variants associated with autoimmune disease. *Hum. Mol. Genet.* 26, R185–R192.
- Kalita, C.A., Gusev, A., 2022. DeCAF: a novel method to identify cell-type specific regulatory variants and their role in cancer risk. *Genome Biol.* 23, 152.
- Kerimov, N., Hayhurst, J.D., Peikova, K., Manning, J.R., Walter, P., Kolberg, L., Samovica, M., Sakhveli, M.P., Kuzmin, I., Trevanion, S.J., et al., 2021. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* 53, 1290–1299.
- Kim-Hellmuth, S., Aguet, F., Oliva, M., Munoz-Aguirre, M., Kasela, S., Wucher, V., Castel, S.E., Hamel, A.R., Vinuela, A., Roberts, A.L., et al., 2020. Cell type-specific genetic regulation of gene expression across human tissues. *Science* 369, eaaz8528.
- Kindt, A.S., Navarro, P., Semple, C.A., Haley, C.S., 2013. The genomic signature of trait-associated variants. *BMC Genom.* 14, 108.
- Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., Shendure, J., 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315.
- Kumasaka, N., Knights, A.J., Gaffney, D.J., 2016. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* 48, 206–213.
- Li, Z., Li, X., Zhou, H., Gaynor, S.M., Selvaraj, M.S., Arapoglou, T., Quick, C., Liu, Y., Chen, H., Sun, R., et al., 2022. A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nat. Methods* 19, 1599–1611.
- Liang, Y., Aguet, F., Barbeira, A.N., Ardlie, K., Im, H.K., 2021. A scalable unified framework of total and allele-specific counts for *cis*-QTL, fine-mapping, and prediction. *Nat. Commun.* 12, 1424.
- Little, P., Zhabotynsky, V., Li, Y., Lin, D., Sun, W., 2022. Cell type-specific expression quantitative trait loci. *bioRxiv*. <https://doi.org/10.1101/2022.03.31.486605>.
- Liu, L., Zeng, P., Xue, F., Yuan, Z., Zhou, X., 2021. Multi-trait transcriptome-wide association studies with probabilistic mendelian randomization. *Am. J. Hum. Genet.* 108, 240–256.
- Lu, Q., Hu, Y., Sun, J., Cheng, Y., Cheung, K.H., Zhao, H., 2015. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.* 5, 10576.
- Lu, Q., Powles, R.L., Abdallah, S., Ou, D., Wang, Q., Hu, Y., Lu, Y., Liu, W., Li, B., Mukherjee, S., et al., 2017. Systematic tissue-specific functional annotation of the

- human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. *PLoS Genet.* 13, e1006933.
- Lu, Q., Powles, R.L., Wang, Q., He, B.J., Zhao, H., 2016. Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLoS Genet.* 12, e1005947.
- Mandric, I., Schwarz, T., Majumdar, A., Hou, K., Briscoe, L., Perez, R., Subramanian, M., Hafemeister, C., Satija, R., Ye, C.J., et al., 2020. Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nat. Commun.* 11, 5504.
- Marderstein, A.R., Davenport, E.R., Kulm, S., Van Hout, C.V., Elemento, O., Clark, A.G., 2021. Leveraging phenotypic variability to identify genetic interactions in human phenotypes. *Am. J. Hum. Genet.* 108, 49–67.
- Nathan, A., Asgari, S., Ishigaki, K., Valencia, C., Amariuta, T., Luo, Y., Beynor, J.I., Baglaenko, Y., Suliman, S., Price, A.L., et al., 2022. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature* 606, 120–128.
- Neavin, D., Nguyen, Q., Daniszewski, M.S., Liang, H.H., Chiu, H.S., Wee, Y.K., Senabouth, A., Lukowski, S.W., Crombie, D.E., Lidgerwood, G.E., et al., 2021. 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.
- Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., Cox, N.J., 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6, e1000888.
- Oelen, R., de Vries, D.H., Brugge, H., Gordon, M.G., Vochteloo, M., single-cell eQTLGen consortium, BIOS Consortium, Ye, C.J., Westra, H.J., Franke, L., et al., 2022. Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. *Nat. Commun.* 13, 3267.
- Okamoto, J., Wang, L., Yin, X., Luca, F., Pique-Regi, R., Helms, A., Im, H.K., Morrison, J., Wen, X., 2023. Probabilistic integration of transcriptome-wide association studies and colocalization analysis identifies key molecular pathways of complex traits. *Am. J. Hum. Genet.* 110, 44–57.
- Oliva, M., Demanelis, K., Lu, Y., Chernoff, M., Jasmine, F., Ahsan, H., Kibriya, M.G., Chen, L.S., Pierce, B.L., 2023. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat. Genet.* 55, 112–122.
- Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T., Delaneau, O., 2016. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 32, 1479–1485.
- Ota, M., Nagafuchi, Y., Hatano, H., Ishigaki, K., Terao, C., Takeshima, Y., Yanoaka, H., Kobayashi, S., Okubo, M., Shirai, H., et al., 2021. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* 184, 3006–3021.
- Patel, D., Zhang, X., Farrell, J.J., Chung, J., Stein, T.D., Lunetta, K.L., Farrer, L.A., 2021. Cell-type-specific expression quantitative trait loci associated with Alzheimer disease in blood and brain tissue. *Transl. Psychiatry* 11, 250.
- PsychENCODE Consortium, Akbarian, S., Liu, C., Knowles, J.A., Vaccarino, F.M., Farnham, P.J., Crawford, G.E., Jaffe, A.E., Pinto, D., Dracheva, S., et al., 2015. The PsychENCODE project. *Nat. Neurosci.* 18, 1707–1712.
- Richardson, T.G., Hemani, G., Gaunt, T.R., Relton, C.L., Davey Smith, G., 2020. A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human genome. *Nat. Commun.* 11, 185.
- Ritchie, G.R., Dunham, I., Zeggini, E., Flück, P., 2014. Functional annotation of noncoding sequence variants. *Nat. Methods* 11, 294–296.
- Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al., 2015. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330.
- Sarkar, A.K., Tung, P.Y., Blischak, J.D., Burnett, J.E., Li, Y.I., Stephens, M., Gilad, Y., 2019. Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genet.* 15, e1008045.
- Schmidel, B.J., Gonzalez-Colin, C., Fajardo, V., Rocha, J., Madrigal, A., Ramirez-Suastegui, C., Bhattacharyya, S., Simon, H., Greenbaum, J.A., Peters, B., et al., 2022. Single-cell eQTL analysis of activated T cell subsets reveals activation and cell type-dependent effects of disease-risk variants. *Sci Immunol* 7, eabm2508.
- Schmidel, B.J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A.G., White, B.M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J.A., McVicker, G., et al., 2018. Impact of genetic polymorphisms on human immune cell gene expression. *Cell* 175, 1701–1715.
- Shabalina, A.A., 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, 1353–1358.
- Song, X., Ji, J., Rothstein, J.H., Alexeef, S.E., Sakoda, L.C., Sistig, A., Achacoso, N., Jorgenson, E., Whittemore, A.S., Klein, R.J., et al., 2023. MiXcan: a framework for cell-type-aware transcriptome-wide association studies with an application to breast cancer. *Nat. Commun.* 14, 377.
- Soskic, B., Cano-Gamez, E., Smyth, D.J., Ambridge, K., Ke, Z., Matte, J.C., Bossini-Castillo, L., Kaplanis, J., Ramirez-Navarro, L., Lorenc, A., et al., 2022. Immune disease risk variants regulate gene expression dynamics during CD4+ T cell activation. *Nat. Genet.* 54, 817–826.
- Strober, B.J., Tayeb, K., Popp, J., Qi, G., Gordon, M.G., Perez, R., Ye, C.J., Battle, A., 2022. Uncovering context-specific genetic regulation of gene expression from single-cell rnasequencing using latent-factor models. *bioRxiv*. <https://doi.org/10.1101/2022.12.22.521678>.
- Su, C., Xu, Z., Shan, X., Cai, B., Zhao, H., Zhang, J., 2022a. Cell-type-specific co-expression inference from single cell RNA-sequencing data. *Nat. Commun* 14, 4846.
- Su, C., Zhang, J., Zhao, H., 2022b. CSNet: Estimating cell-type-specific gene co-expression networks from bulk gene expression data. *bioRxiv*. <https://doi.org/10.1101/2021.12.21.473558>.
- Sun, W., 2012. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* 68, 1–11.
- The ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shores, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al., 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 583, 699–710.
- The GTEx Consortium, 2020. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330.
- Urbut, S.M., Wang, G., Carbonetto, P., Stephens, M., 2019. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* 51, 187–195.
- van der Wijst, M.G.P., Brugge, H., de Vries, D.H., Deelen, P., Swertz, M.A., LifeLines Cohort Study, BIOS Consortium, Franke, L., 2018. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* 50, 493–497.
- Wang, S.K., Nair, S., Li, R., Kraft, K., Pampari, A., Patel, A., Kang, J.B., Luong, C., Kundaje, A., Chang, H.Y., 2022. Single-cell multiome of the human retina and deep learning nominate causal variants in complex eye diseases. *Cell Genom.* 2, 100164.
- Wang, Y., Zhao, H., 2022. Non-linear archetypal analysis of single-cell RNA-seq data by deep autoencoders. *PLoS Comput. Biol.* 18, e1010025.
- Wen, X., Lee, Y., Luca, F., Pique-Regi, R., 2016. Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *Am. J. Hum. Genet.* 98, 1114–1129.
- Westra, H.J., Arends, D., Esko, T., Peters, M.J., Schurmann, C., Schramm, K., Kettunen, J., Yaghootkar, H., Fairfax, B.P., Andiappan, A.K., et al., 2015. Cell specific eQTL analysis without sorting cells. *PLoS Genet.* 11, e1005223.
- Xie, D., Wang, J., 2022. Robust statistical inference for cell type deconvolution. *arXiv:* 2202.06420.
- Yankovitz, G., Cohn, O., Bacharach, E., Pesheh-Yaloz, N., Steuerman, Y., Iraqi, F.A., Gat-Viks, I., 2021. Leveraging the cell lineage to predict cell-type specificity of regulatory variation from bulk genomics. *Genetics* 217, iyab016.
- Yao, D.W., O'Connor, L.J., Price, A.L., Gusev, A., 2020. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* 52, 626–633.
- Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M.G., Andersen, S., Lu, Q., Rowson, A., Taylor, T.R.P., Clarke, L., et al., 2022. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* 376, eabf3041.
- Young, A.M.H., Kumazaka, N., Calvert, F., Hammond, T.R., Knights, A., Panousis, N., Park, J.S., Schwartzentruber, J., Liu, J., Kundu, K., et al., 2021. A map of transcriptional heterogeneity and regulatory variation in human microglia. *Nat. Genet.* 53, 861–868.
- Yuan, Z., Zhu, H., Zeng, P., Yang, S., Sun, S., Yang, C., Liu, J., Zhou, X., 2020. Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nat. Commun.* 11, 3861.
- Zhabotynsky, V., Huang, L., Little, P., Hu, Y.J., Pardo-Manuel de Villena, F., Zou, F., Sun, W., 2022. eQTL mapping using allele-specific count data is computationally feasible, powerful, and provides individual-specific estimates of genetic effects. *PLoS Genet.* 18, e1010076.
- Zhang, T., Choi, J., Dilshat, R., Einarsdottir, B.O., Kovacs, M.A., Xu, M., Malasky, M., Chowdhury, S., Jones, K., Bishop, D.T., et al., 2021. Cell-type-specific meQTLs extend melanoma GWAS annotation beyond eQTLs and inform melanocyte gene-regulatory mechanisms. *Am. J. Hum. Genet.* 108, 1631–1646.
- Zhang, T., Choi, J., Kovacs, M.A., Shi, J., Xu, M., Program, N.C.S., Melanoma Meta-Analysis, Goldstein, A.M., Trower, A.J., Bishop, D.T., et al., 2018. Cell-type-specific eQTL of primary melanocytes facilitates identification of melanoma susceptibility genes. *Genome Res.* 28, 1621–1635.
- Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.J., et al., 2017. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* 49, 139–145.
- Zhou, D., Jiang, Y., Zhong, X., Cox, N.J., Liu, C., Gamazon, E.R., 2020. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nat. Genet.* 52, 1239–1246.