A unified framework for cell-type-specific eQTL prioritization by integrating bulk and scRNA-seq data

Authors

Xinyi Yu, Xianghong Hu, Xiaomeng Wan, ..., Mingxuan Cai, Tianwei Yu, Jiashun Xiao

Correspondence

yutianwei@cuhk.edu.cn (T.Y.), xiaojiashun@sribd.cn (J.X.)

This paper introduces IBSEP, a statistical framework that integrates bulk and scRNA-seq data for prioritizing cell-type-specific eQTLs. By overcoming the limitations while leveraging the advantages associated with each technique, IBSEP enhances the identification of genetic variants linked to complex diseases, revealing new insights into transcriptional regulation at the cellular level.

Yu et al., 2025, The American Journal of Human Genetics *112*, 1–21
February 6, 2025
https://doi.org/10.1016/j.ajhg.2024.12.018

Please cite this article in press as: Yu et al., A unified framework for cell-type-specific eQTL prioritization by integrating bulk and scRNA-seq data, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2024.12.018

ARTICLE

A unified framework for cell-type-specific eQTL prioritization by integrating bulk and scRNA-seq data

Xinyi Yu,[1,2] Xianghong Hu,[3] Xiaomeng Wan,[3] Zhiyong Zhang,[1,2] Xiang Wan,[1] Mingxuan Cai,[4] Tianwei Yu,[1,2,*] and Jiashun Xiao[1,*]

Summary

Genome-wide association studies (GWASs) have identified numerous genetic variants associated with complex traits, yet the biological interpretation remains challenging, especially for variants in non-coding regions. Expression quantitative trait locus (eQTL) studies have linked these variations to gene expression, aiding in identifying genes involved in disease mechanisms. Traditional eQTL analyses using bulk RNA sequencing (bulk RNA-seq) provide tissue-level insights but suffer from signal loss and distortion due to unaddressed cellular heterogeneity. Recently, single-cell RNA-seq (scRNA-seq) has provided higher resolution, enabling cell-type-specific eQTL (ct-eQTL) analyses. However, these studies are limited by their smaller sample sizes and technical constraints. In this paper, we present a statistical framework, IBSEP, which integrates bulk RNA-seq and scRNA-seq data for enhanced ct-eQTL prioritization. Our method employs a hierarchical linear model to combine summary statistics from both data types, overcoming the limitations while leveraging the advantages associated with each technique. Through extensive simulations and real data analyses, including peripheral blood mononuclear cells and brain cortex datasets, IBSEP demonstrated superior performance in identifying ct-eQTLs compared to existing methods. Our approach unveils transcriptional regulatory mechanisms specific to cell types, offering deeper insights into the genetic basis of complex diseases at a cellular resolution.

Introduction

Genome-wide association studies (GWASs) have made significant achievements in identifying genomic regions associated with complex traits/diseases, yet the biological interpretation of GWAS results remains challenging, as the majority of risk variants are located in non-coding regions of the genome.[1] Among numerous studies annotating genetic variations, expression quantitative trait locus (eQTL) studies link the risk variations to their potentially regulating target genes, aiding in identifying genes that may participate in disease mechanisms.[2] Studies have shown that genetic variations associated with complex traits are enriched in identified eQTLs.[3] Given the significant relevance of eQTL studies in interpreting GWAS results, researchers worldwide have accumulated substantial gene expression and genotype data for eQTL studies,[4,5] laying the foundation for uncovering the genetic basis of complex traits.[6]

So far, most eQTL studies have been based on bulk RNA sequencing (bulk RNA-seq) at the tissue level, measuring the average gene expression levels from hundreds to millions of cells within a tissue. One of the most prominent eQTL studies comes from the Genotype-Tissue Expression (GTEx) project, which analyzed 15,201 RNA-seq samples from 49 tissues of 838 donors, identifying tissue-level eQTLs in 94.7% of protein-coding genes.[4] Despite the fruitful discoveries of tissue-level eQTLs, they still only explain a moderate fraction of GWAS signals.[7] In recent years, increasing evidence suggests that transcriptional regulation mechanisms are dynamic and highly dependent on cell type or cell state.[8] Some eQTLs have been demonstrated to be detectable only in specific cell types,[8] indicating the critical importance of cell-type-specific eQTLs (ct-eQTL) prioritization for better interpretation and understanding of GWAS results.

Studies of ct-eQTLs can be divided into two categories: those based on traditional bulk RNA-seq data and those based on single-cell RNA-seq (scRNA-seq) data. Given the abundant bulk RNA-seq data, ct-eQTLs can be identified by detecting the interaction effects of single-nucleotide polymorphisms (SNPs) and cell type proportions on gene expression, known as interaction eQTLs (ieQTLs). For instance, Zhernakova et al.[9] introduce a "genotype-cell type proportion" interaction term into the eQTL linear model and observed that 12% of *cis*-regulated genes showed context-dependent eQTL effects. Further studies include Decon-eQTL,[10] which included multiple cell types in the interaction term, and CSeQTL,[11] which models gene expression using binomial or Poisson distribution. While this approach can be directly applied to existing bulk RNA-seq data, its statistical power is unsatisfactory, particularly when locating eQTLs specific to rare cell types. The other type of approach directly utilizes scRNA-seq data to locate ct-eQTLs. For instance, Van et al. analyzed peripheral blood mononuclear cell (PBMC) scRNA-seq data from 45 donors and identified a total of 379 eQTLs across six cell types, including

[1]Shenzhen Research Institute of Big Data, Shenzhen 518172, China; [2]School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen 518172, China; [3]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong SAR, China; [4]Department of Biostatistics, City University of Hong Kong, Hong Kong SAR, China
*Correspondence: yutianwei@cuhk.edu.cn (T.Y.), xiaojiashun@sribd.cn (J.X.)
https://doi.org/10.1016/j.ajhg.2024.12.018.

Please cite this article in press as: Yu et al., A unified framework for cell-type-specific eQTL prioritization by integrating bulk and scRNA-seq data, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2024.12.018

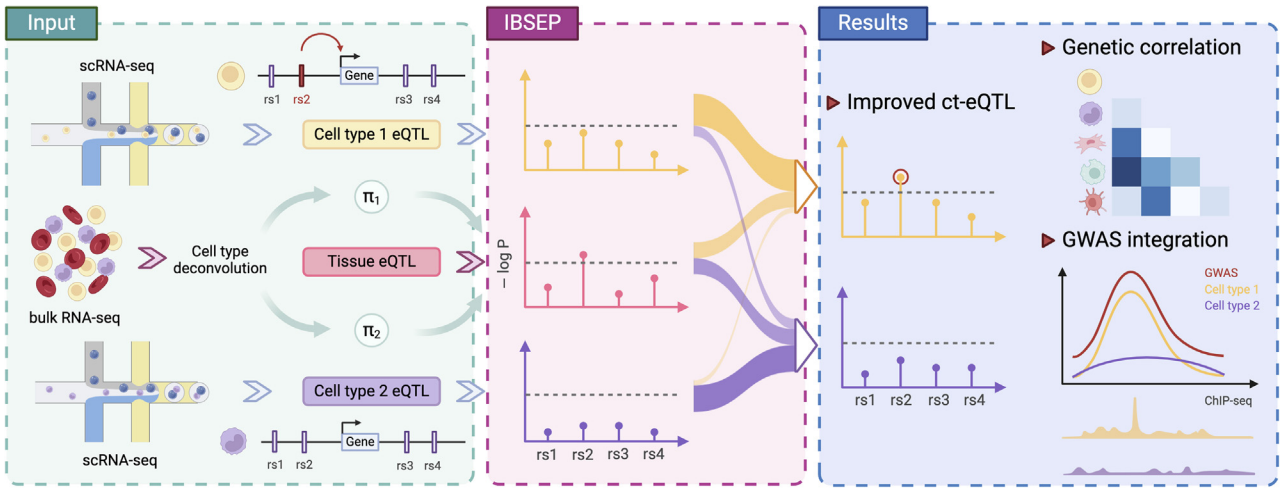

Figure 1. Overview of IBSEP
IBSEP is designed to improve ct-eQTL prioritization by integrating bulk RNA-seq and scRNA-seq data, revealing the heterogeneity of transcriptional regulation among different cell types. The workflow of IBSEP begins by estimating the cell type proportions for bulk RNA-seq samples. Then, it takes summary statistics of cell-type-level eQTLs from scRNA-seq data, tissue-level eQTLs from bulk RNA-seq data, and the estimated cell type proportions as input to a hierarchical linear model and output improved ct-eQTL summary statistics. Finally, by integrating with disease GWAS results using colocalization analysis, the improved ct-eQTL results can provide a comprehensive and in-depth understanding of the genetic basis and pathogenic mechanisms of complex diseases at the resolution of cell type.

48 newly discovered ct-eQTLs[12]; Bryois et al. used brain cortex scRNA-seq data from 192 individuals and identified a total of 7,607 eQTLs across eight major brain cell types, with 46% showing cell-type-specific associations, particularly in microglia cells.[13] However, due to the high cost and technical constraints (e.g., dropouts) of scRNA-seq, these studies typically have small-scale and low-quality gene expression counts, limiting the power of ct-eQTL mapping. In summary, tissue-level sequencing data have a larger sample size but low resolution, while single-cell-level sequencing data have high resolution but a smaller scale.

Although the above methods can be utilized for ct-eQTL mapping, they are not yet perfect in terms of effectiveness, applicability, and scalability. On the one hand, both bulk RNA-seq and scRNA-seq studies on ct-eQTLs are being conducted independently. There is a notable absence of methods capable of bridging the platform disparity between bulk RNA-seq and scRNA-seq technologies to effectively integrate these datasets for the identification of ct-eQTLs.[14] On the other hand, existing ct-eQTL methods requires individual-level genotype and transcriptomic data and thus cannot fully make use of widely available summary statistics.[9–11,15,16] Besides, current ct-eQTL analysis ignored the widespread genetic correlations among cell types, especially biologically close ones, and performed ct-eQTL mapping on a single cell type at a time, leading to a suboptimal statistical power.[13,17]

In this paper, we develop a unified statistical framework for integrating bulk RNA-seq and scRNA-seq data for ct-eQTL prioritization (IBSEP). The keys to the success of IBSEP are 3-fold. First, by integrating tissue-level and cell-type-level eQTL data into a unified hierarchical linear model, it can effectively leverage the advantages of each type of data to enhance the statistical power of ct-eQTL prioritization. Second, by directly modeling the eQTL summary statistics, it can use these summary statistics as input, enabling greater utility in large-scale data integration analyses. Third, by incorporating the *cis*-coheritability between SNPs across cell types, it can consider the complex structural relationships between cell types, revealing both shared and specific patterns of transcriptional regulation across different cell types. Through comprehensive simulation studies, we demonstrated that IBSEP outputs well-calibrated p values and achieves significant power gains. Then, by separately integrating blood[18,19] and brain[13] scRNA-seq datasets with corresponding GTEx bulk RNA-seq datasets,[4] IBSEP unveiled more ct-eQTLs and the underlying cell-type-specific transcriptional regulatory mechanisms. Further colocalization analysis of GWAS risk SNPs and IBSEP ct-eQTLs discovered target genes mediating immune and brain disease associations. These findings provide insights for uncovering the underlying mechanisms of gene transcription regulation associated with diseases at the resolution of cell types.

Material and methods

Method overview

Current independent ct-eQTL studies overlook the inherent correlation between bulk RNA-seq and scRNA-seq data, leading to large room for improvement in statistical power for ct-eQTL prioritization. IBSEP addresses the challenges of ct-eQTL prioritization from a new perspective (Figure 1). Essentially, bulk-level gene expression of a tissue sample can be viewed as a weighted average of gene expression at the cell type level, with the

Please cite this article in press as: Yu et al., A unified framework for cell-type-specific eQTL prioritization by integrating bulk and scRNA-seq data, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2024.12.018

weights being the cell type proportions of the tissue sample. Given this fact, we designed a unified statistical framework to integrate bulk RNA-seq and scRNA-seq data. However, the variation in cell type proportions within tissue samples complicates the integration of these models, posing significant challenges for analyzing summary-level eQTL datasets. By using the law of large numbers, we prove that only the average cell type proportions are required in the integrative analysis. This enables us to establish a simple yet effective hierarchical linear model to use summary statistics instead of individual-level data. In short, IBSEP takes summary statistics of ct-eQTLs from scRNA-seq data with a smaller sample size, summary statistics of tissue-level eQTLs from bulk RNA-seq data with a larger sample size, and the average cell type proportion in tissue samples as model inputs and outputs improved ct-eQTL summary statistics. IBSEP significantly enhances the statistical power of ct-eQTL mappings while producing well-controlled type I errors. With our model design innovations, IBSEP shows high computational efficiency by offering a closed-form solution at each step. Moreover, the output of IBSEP can be readily applied in downstream analyses, such as colocalization with GWAS risk variants, to more effectively discover gene regulation underlying the disease mechanisms at the cell type level.

The IBSEP model

We begin our formulation with the individual-level scRNA-seq and bulk RNA-seq data. For the tissue-level data of a target gene, suppose we have collected a dataset $\{\mathbf{y}_t, \mathbf{X}_t\}$, where $\mathbf{y}_t \in \mathbb{R}^{N_t}$ is the normalized gene expression vector and $\mathbf{X}_t \in \mathbb{R}^{N_t \times M}$ is the standardized genotype matrix. M is the number of SNPs within the target gene (e.g., 200), and N_t is the number of tissue samples (e.g., 1,000). Without loss of generality, we assume the covariates have been properly adjusted. More detailed treatment of covariate adjustments can be found in our previous work.[20,21] The following linear model relates gene expression and genotype:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta}_t + \epsilon_t, \epsilon_t \sim \mathcal{N}\left(0, \left(1 - h_t^2\right)\boldsymbol{I}_{N_t}\right), \quad \text{(Equation 1)}$$

where $\boldsymbol{\beta}_t = [\beta_{t1}, \ldots, \beta_{tM}]^T, \beta_{tj} \sim \mathcal{N}(0, h_t^2/M)$ are the *cis*-SNP effect sizes, h_t^2 is the *cis*-heritability of the target gene, and ϵ_t is the independent error term.

Without loss of generality, we consider only two cell types in the tissue. Suppose that by using scRNA-seq, we have obtained a cell-type-level eQTL dataset $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{X}_c\}$ from the same population, where $\mathbf{y}_1 \in \mathbb{R}^{N_c}$ and $\mathbf{y}_2 \in \mathbb{R}^{N_c}$ are the normalized mean gene expression vectors of cell type 1 and cell type 2, respectively, $\mathbf{X}_c \in \mathbb{R}^{N_c \times M}$ is the standardized genotype matrix, and N_c is the sample size of scRNA-seq data, which is typically smaller than 200. Combining tissue-level and cell-type-level data, we consider the Bayesian hierarchical linear model:

$$\begin{aligned}
\mathbf{y}_1 &= \mathbf{X}_c\boldsymbol{\beta}_1 + \epsilon_1, \epsilon_1 \sim \mathcal{N}\left(0, \left(1 - h_1^2\right)\boldsymbol{I}_{N_c}\right), \\
\mathbf{y}_2 &= \mathbf{X}_c\boldsymbol{\beta}_2 + \epsilon_2, \epsilon_2 \sim \mathcal{N}\left(0, \left(1 - h_2^2\right)\boldsymbol{I}_{N_c}\right), \\
\mathbf{y}_t &= \mathbf{X}_t\boldsymbol{\beta}_t + \epsilon_t = \boldsymbol{\pi}_1 \odot \mathbf{X}_t\boldsymbol{\beta}_1 + \boldsymbol{\pi}_2 \odot \mathbf{X}_t\boldsymbol{\beta}_2 + \epsilon_t,
\end{aligned} \quad \text{(Equation 2)}$$

where $\boldsymbol{\beta}_1 = [\beta_{11}, \ldots, \beta_{1M}]^T, \beta_{1j} \sim \mathcal{N}(0, h_1^2/M)$ and $\boldsymbol{\beta}_2 = [\beta_{21}, \ldots, \beta_{2M}]^T, \beta_{2j} \sim \mathcal{N}(0, h_2^2/M)$ are the *cis*-SNP effect sizes of cell type 1 and cell type 2, respectively; h_1^2 and h_2^2 are the *cis*-heritability of this gene in the two cell types; ϵ_1 and ϵ_2 are the independent error terms; $\boldsymbol{\pi}_1 = [\pi_{11}, \ldots, \pi_{1N_t}]^T$ and $\boldsymbol{\pi}_2 = [\pi_{21}, \ldots, \pi_{2N_t}]^T$ are the proportions of the two cell types in the tissue samples; and $\odot$ represents an element-wise product. Thus, tissue-level gene expression $\mathbf{y}_t$ can be mathematically decomposed into a weighted average of cell type 1 ($\mathbf{X}_1\boldsymbol{\beta}_1$) and cell type 2 ($\mathbf{X}_2\boldsymbol{\beta}_2$) where the weights are the cell type proportions $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$. We aimed to estimate the *cis*-SNP effect sizes ($\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$) by integrating tissue-level and cell-type-level data. However, the variation in cell type proportions within tissue samples presents significant challenges for integrating these two sets of models. On the one hand, the need to account for the heterogeneity of samples makes it difficult to apply to summary-level eQTL data. On the other hand, it is challenging to design an efficient and reliable algorithm that can simultaneously analyze multiple cell types with complex genetic correlations in real ct-eQTL analysis.

The IBSEP model with summary-level data

To illustrate the challenges, assume the individual data $\{\mathbf{X}_t, \mathbf{y}_t, \mathbf{X}_c, \mathbf{y}_1, \mathbf{y}_2\}$ are unavailable, but we have access to the summary statistics $\{\widehat{b}_t, \widehat{s}_t\} = \{b_{tj}, s_{tj}\}_{j=1}^M$, $\{\widehat{b}_1, \widehat{s}_1\} = \{b_{1j}, s_{1j}\}_{j=1}^M$, and $\{\widehat{b}_2, \widehat{s}_2\} = \{b_{2j}, s_{2j}\}_{j=1}^M$:

$$\begin{aligned}
\widehat{b}_{1j} &= \mathbf{x}_{cj}^T\mathbf{y}_1 \Big/ \mathbf{x}_{cj}^T\mathbf{x}_{cj} = \mathbf{x}_{cj}^T\mathbf{y}_1 \Big/ N_c, \widehat{s}_{1j} \\
&= \sqrt{\left(\mathbf{y}_1 - \mathbf{x}_{cj}\widehat{b}_{1j}\right)^T\left(\mathbf{y}_1 - \mathbf{x}_{cj}\widehat{b}_{1j}\right) \Big/ \left(N_c\mathbf{x}_{cj}^T\mathbf{x}_{cj}\right)}, \\
\widehat{b}_{2j} &= \mathbf{x}_{cj}^T\mathbf{y}_2 \Big/ \mathbf{x}_{cj}^T\mathbf{x}_{cj} = \mathbf{x}_{cj}^T\mathbf{y}_2 \Big/ N_c, \widehat{s}_{2j} \\
&= \sqrt{\left(\mathbf{y}_2 - \mathbf{x}_{cj}\widehat{b}_{2j}\right)^T\left(\mathbf{y}_2 - \mathbf{x}_{cj}\widehat{b}_{2j}\right) \Big/ \left(N_c\mathbf{x}_{cj}^T\mathbf{x}_{cj}\right)}, \\
\widehat{b}_{tj} &= \mathbf{x}_{tj}^T\mathbf{y}_t \Big/ \mathbf{x}_{tj}^T\mathbf{x}_{tj} = \mathbf{x}_{tj}^T\mathbf{y}_t \Big/ N_t, \widehat{s}_{tj} \\
&= \sqrt{\left(\mathbf{y}_t - \mathbf{x}_{tj}\widehat{b}_{tj}\right)^T\left(\mathbf{y}_t - \mathbf{x}_{tj}\widehat{b}_{tj}\right) \Big/ \left(N_t\mathbf{x}_{tj}^T\mathbf{x}_{tj}\right)}.
\end{aligned} \quad \text{(Equation 3)}$$

Clearly, the eQTLs and ct-eQTLs obtained through independent analyses fail to fully leverage the correlations between $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_t$, resulting in suboptimal statistical power. To bridge the tissue-level and cell-type-level eQTL data in mathematical models, we define the true marginal effect sizes as

$$\begin{aligned}
b_{1j} &= \mathbb{E}\left[\widehat{b}_{1j} \middle| \boldsymbol{\beta}_1\right] = \mathbb{E}\left[\mathbf{x}_{cj}^T\left(\mathbf{X}_c\boldsymbol{\beta}_1 + \epsilon_1\right) \Big/ N_c \middle| \boldsymbol{\beta}_1\right] = \sum_{k=1}^{M} r_{jk}\beta_{1k}, \\
b_{2j} &= \mathbb{E}\left[\widehat{b}_{2j} \middle| \boldsymbol{\beta}_2\right] = \mathbb{E}\left[\mathbf{x}_{cj}^T\left(\mathbf{X}_c\boldsymbol{\beta}_2 + \epsilon_2\right) \Big/ N_c \middle| \boldsymbol{\beta}_2\right] = \sum_{k=1}^{M} r_{jk}\beta_{2k}, \\
b_{tj} &= \mathbb{E}\left[\widehat{b}_{tj} \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right] = \mathbb{E}\left[\mathbf{x}_{tj}^T\left(\boldsymbol{\pi}_1 \odot \mathbf{X}_t\boldsymbol{\beta}_1 + \boldsymbol{\pi}_2 \odot \mathbf{X}_t\boldsymbol{\beta}_2 + \epsilon_t\right) \Big/ N_t \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right],
\end{aligned} \quad \text{(Equation 4)}$$

where r_{jk} is the correlation between SNP j and SNP k. In the last line of Equation 4, it can be observed that the differences in $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ across different samples make it infeasible to establish a direct relationship between the tissue-level marginal effects $\mathbf{b}_t$ and the cell-type-level marginal effects $\{\mathbf{b}_1, \mathbf{b}_2\}$.

Fortunately, we found that when integrating summary-statistic-level eQTL data, differences in cell type proportions in tissue samples can be disregarded. The cell type proportion vectors $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ in tissue-level RNA-seq data can be approximated by their mean $\overline{\pi}_1$, $\overline{\pi}_2$, respectively. Specifically, we proved that when the tissue sample size is large enough (e.g., $N_t > 100$), the estimate of marginal effect size $\widehat{b}_{tj}$ for *cis*-SNP j at the tissue level is approximately independent of the cell type proportion composition variations across individuals. In other words, the summary statistics $\widehat{b}_{tj}$ can be

approximated by the mean cell type proportions $(\overline{\pi}_1, \overline{\pi}_2)$ instead of requiring individual cell type proportions $(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$. The proof is provided in supplemental methods section 3.3, and the experimental validation is given in supplemental methods section 3.4. Given this point, we update Equation 2 as follows:

$$\mathbf{y}_t = \boldsymbol{\pi}_1 \odot \mathbf{X}_t\boldsymbol{\beta}_1 + \boldsymbol{\pi}_2 \odot \mathbf{X}_t\boldsymbol{\beta}_2 + \epsilon_t \approx \mathbf{X}_t(\overline{\pi}_1\boldsymbol{\beta}_1 + \overline{\pi}_2\boldsymbol{\beta}_2) + \epsilon_t, \quad \text{(Equation 5)}$$

and the marginal effect b_{tj}:

$$b_{tj} = \mathbb{E}\left[\widehat{b}_{tj}\middle|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right] \approx \mathbb{E}\left[\mathbf{x}_{tj}^T(\mathbf{X}_t(\overline{\pi}_1\boldsymbol{\beta}_1 + \overline{\pi}_2\boldsymbol{\beta}_2) + \epsilon_t) \middle/ N_t\middle|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right]$$

$$= \overline{\pi}_1 \sum_{k=1}^{M} r_{jk}\beta_{1k} + \overline{\pi}_2 \sum_{k=1}^{M} r_{jk}\beta_{2k} = \overline{\pi}_1 b_{1j} + \overline{\pi}_2 b_{2j}. \quad \text{(Equation 6)}$$

Based on this, we can finally establish a simple yet effective linear model to integrate tissue-level data and cell-type-level data.

In order to better utilize the correlation among $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_t$, we impose the following probabilistic structure for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$:

$$\mathbb{E}\begin{pmatrix}\beta_{1j}\\ \beta_{2j}\end{pmatrix} = 0, \mathrm{Var}\begin{pmatrix}\beta_{1j}\\ \beta_{2j}\end{pmatrix} = \boldsymbol{\Omega} = \begin{pmatrix}\omega_1 & \omega_{12}\\ \omega_{12} & \omega_2\end{pmatrix}, j = 1, \ldots, M, \quad \text{(Equation 7)}$$

where $\boldsymbol{\Omega}$ captures the genetic covariance of the two cell types in *cis*-SNPs. The diagonal elements represent the per-SNP *cis*-heritability for cell types 1 and 2, and the off-diagonal elements represent the per-SNP *cis*-coheritability between the two cell types.

The estimates of marginal effect sizes $\{\widehat{b}_{1j}, \widehat{b}_{2j}, \widehat{b}_{tj}\}$ for *cis*-SNP j can be decomposed as

$$\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\\ \widehat{b}_{tj}\end{pmatrix} = \begin{pmatrix}b_{1j}\\ b_{2j}\\ b_{tj}\end{pmatrix} + \begin{pmatrix}e_{1j}\\ e_{2j}\\ e_{tj}\end{pmatrix} = \underbrace{\begin{pmatrix}1 & 0\\ 0 & 1\\ \overline{\pi}_1 & \overline{\pi}_2\end{pmatrix}}_{\mathbf{A}}\begin{pmatrix}b_{1j}\\ b_{2j}\end{pmatrix} + \begin{pmatrix}e_{1j}\\ e_{2j}\\ e_{tj}\end{pmatrix}, j = 1, \ldots, M, \quad \text{(Equation 8)}$$

where e_{1j}, e_{2j}, and e_{tj} are the independent estimation errors with $\mathrm{Var}(e_{1j}) = \widehat{s}_{1j}^2$, $\mathrm{Var}(e_{2j}) = \widehat{s}_{2j}^2$, and $\mathrm{Var}(e_{tj}) = \widehat{s}_{tj}^2$.

With the model specification, we can derive the covariance of marginal effect sizes (more details are available in supplemental methods section 3.1). In a compact form, we have

$$\mathrm{Cov}\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\\ \widehat{b}_{tj}\end{pmatrix} = \mathrm{Cov}\begin{pmatrix}b_{1j}\\ b_{2j}\\ b_{tj}\end{pmatrix} + \mathrm{Cov}\begin{pmatrix}e_{1j}\\ e_{2j}\\ e_{tj}\end{pmatrix} = \mathbf{A}\underbrace{l_j\boldsymbol{\Omega}}_{\boldsymbol{\Omega}_j}\mathbf{A}^T + \underbrace{\begin{pmatrix}\widehat{s}_{1j}^2 & 0 & 0\\ 0 & \widehat{s}_{2j}^2 & 0\\ 0 & 0 & \widehat{s}_{tj}^2\end{pmatrix}}_{\widehat{\mathbf{s}}_j^2}, \quad \text{(Equation 9)}$$

where $l_j = \sum_k r_{jk}^2$ is the linkage disequilibrium (LD) score for SNP j.

To account for the hidden confounding biases in GWAS summary statistics, we generalized the linear model (Equation 5) based on the genetic drift model used in LDSC[22] and obtain the modified covariance of marginal effects (supplemental methods section 3.2). Now, Equation 9 becomes

$$\mathrm{Cov}\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\\ \widehat{b}_{tj}\end{pmatrix} = \mathbf{A}\boldsymbol{\Omega}_j\mathbf{A}^T + \widehat{\mathbf{S}}_j\mathbf{C}\widehat{\mathbf{S}}_j, \mathbf{C} = \begin{pmatrix}c_1 & c_{12} & c_{13}\\ c_{12} & c_2 & c_{23}\\ c_{13} & c_{23} & c_3\end{pmatrix}, \quad \text{(Equation 10)}$$

where elements in $\mathbf{C}$ are inflation constants that adjust for the confounding biases due to the geographic structure, sample overlap, etc.

We summarize the reasons and benefits of the proposed model. First, by integrating tissue-level marginal effect b_{tj} and the cell-type-level marginal effect b_{1j}, b_{2j} using a simple linear model, we can effectively combine high-quality large-scale bulk RNA-seq data with small-sample scRNA-seq data, fully leveraging the advantages of each type of data to enhance the statistical power of ct-eQTL prioritization. Second, by directly modeling the eQTL summary statistics, we can use these summary statistics as input, providing greater utility in large-scale data integration analyses. Third, by introducing off-diagonal elements $\boldsymbol{\Omega}$ to represent the average *cis*-coheritability between SNPs across cell types, we can consider the complex structural relationships between cell types, revealing both shared and specific patterns of transcriptional regulation across different cell types.

The IBSEP estimator

Based on the proposed statistical model for integrating ct-eQTL analysis, the following focuses on the development of efficient algorithms tailored to it. From Equations 8 and 10, we first obtain the conditional mean of the observed marginal effects (supplemental methods section 3.5):

$$\mathbb{E}\left[\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\\ \widehat{b}_{tj}\end{pmatrix}\middle| b_{1j}\right] = \underbrace{\begin{pmatrix}1\\ \omega_{j,12}/\omega_{j,11}\\ \overline{\pi}_1 + \overline{\pi}_2\omega_{j,12}/\omega_{j,11}\end{pmatrix}}_{\boldsymbol{\lambda}_1} b_{1j}, \quad \text{(Equation 11)}$$

and the conditional variance:

$$\mathrm{Var}\left[\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\\ \widehat{b}_{tj}\end{pmatrix}\middle| b_{1j}\right] = \mathbf{A}\left(\boldsymbol{\Omega}_j - \frac{\boldsymbol{\omega}_{j,\cdot 1}\boldsymbol{\omega}_{j,\cdot 1}^T}{\omega_{j,11}}\right)\mathbf{A}^T + \widehat{\mathbf{S}}_j\mathbf{C}\widehat{\mathbf{S}}_j : = \boldsymbol{\Lambda}_1^{-1}, \quad \text{(Equation 12)}$$

where $\boldsymbol{\omega}_{j,\cdot 1} = [\omega_{j,11}, \omega_{j,12}]^T$ is the first column of $\boldsymbol{\Omega}_j$. Based on Equation 11, we can define $\mathbf{m}(b) := \begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\\ \widehat{b}_{tj}\end{pmatrix} - \boldsymbol{\lambda}_1 b$ and give the first-order moment condition:

$$\mathbb{E}\left[\mathbf{m}(b)|_{b=b_{1j}}\right] = \mathbb{E}\left[\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\\ \widehat{b}_{tj}\end{pmatrix} - \boldsymbol{\lambda}_1 b_{1j}\right] = 0. \quad \text{(Equation 13)}$$

In the framework of a generalized method of moments (GMM),[23] we treat the true marginal effect b_{1j} of cell type 1 as an estimable parameter and obtain its best linear unbiased estimator by solving the moment condition in Equation 13:

$$\widehat{b}_{1j}^{\mathrm{IBSEP}} = \arg\min_b \mathbf{m}(b)^T\boldsymbol{\Lambda}_1\mathbf{m}(b) = \underbrace{\left(\boldsymbol{\lambda}_1^T\boldsymbol{\Lambda}_1\boldsymbol{\lambda}_1\right)^{-1}\boldsymbol{\lambda}_1^T\boldsymbol{\Lambda}_1}_{\mathbf{w}_1^T}\underbrace{\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\\ \widehat{b}_{tj}\end{pmatrix}}_{\widehat{\mathbf{b}}_{\cdot j}} = \mathbf{w}_1^T\widehat{\mathbf{b}}_{\cdot j}. \quad \text{(Equation 14)}$$

Please cite this article in press as: Yu et al., A unified framework for cell-type-specific eQTL prioritization by integrating bulk and scRNA-seq data, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2024.12.018

According to the asymptotic normality of a GMM, its corresponding variance is

$$\mathrm{Var}\left(\widehat{b}_{1j}^{\mathrm{IBSEP}}\right)=\left(\frac{\partial \mathbf{m}^T}{\partial b}\mathbf{\Lambda}_1\frac{\partial \mathbf{m}}{\partial b}\right)^{-1}=\left(\boldsymbol{\lambda}_1^T\mathbf{\Lambda}_1\boldsymbol{\lambda}_1\right)^{-1}. \quad \text{(Equation 15)}$$

Similarly, we can obtain the $\widehat{b}_{2j}^{\mathrm{IBSEP}}$ and its variance $\mathrm{Var}(\widehat{b}_{2j}^{\mathrm{IBSEP}})$. Theoretical properties of the IBSEP estimator are given in supplemental methods section 3.6.

Parameter estimation and ct-eQTL prioritization

The above IBSEP estimator involves a few unknown parameters to be estimated, including mean cell type proportions $\overline{\pi}_1,\overline{\pi}_2$, per-SNP *cis*-heritability of the two cell types (ω_1,ω_2), the per-SNP *cis*-coheritability between the two cell types (ω_{12}), and inflation constants in $\mathbf{C}$. To estimate the mean cell type proportions, we feed the bulk RNA-seq into a commonly used cell type deconvolution method, CIBERSORTx,[24] to obtain individual cell type proportion estimations of $\boldsymbol{\pi}_1,\boldsymbol{\pi}_2$ and then average them to get $\overline{\pi}_1,\overline{\pi}_2$.

For the estimation of the parameters $\{\omega_1,\omega_2,\omega_{12},\mathbf{C}\}$, we consider following regressions derived from Equation S7:

$$\begin{aligned}
z_{1j}^2 &:= \widehat{b}_{1j}^2\big/\widehat{s}_{1j}^2 \sim \omega_1 l_j\big/\widehat{s}_{1j}^2+c_1,\\
z_{2j}^2 &:= \widehat{b}_{2j}^2\big/\widehat{s}_{2j}^2 \sim \omega_2 l_j\big/\widehat{s}_{2j}^2+c_2,\\
z_{tj}^2 &:= \widehat{b}_{tj}^2\big/\widehat{s}_{tj}^2 \sim \left(\overline{\pi}_1^2\omega_1+\overline{\pi}_2^2\omega_2+2\overline{\pi}_1\overline{\pi}_2\omega_{12}\right)l_j\big/\widehat{s}_{tj}^2+c_3,\\
z_{1j}z_{2j} &:= \widehat{b}_{1j}\widehat{b}_{2j}\big/\left(\widehat{s}_{1j}\widehat{s}_{2j}\right) \sim \omega_{12}l_j\big/\left(\widehat{s}_{1j}\widehat{s}_{2j}\right)+c_{12},\\
z_{1j}z_{tj} &:= \widehat{b}_{1j}\widehat{b}_{tj}\big/\left(\widehat{s}_{1j}\widehat{s}_{tj}\right) \sim \left(\overline{\pi}_1\omega_1+\overline{\pi}_2\omega_{12}\right)l_j\big/\left(\widehat{s}_{1j}\widehat{s}_{tj}\right)+c_{13},\\
z_{2j}z_{tj} &:= \widehat{b}_{2j}\widehat{b}_{tj}\big/\left(\widehat{s}_{2j}\widehat{s}_{tj}\right) \sim \left(\overline{\pi}_2\omega_{12}+\overline{\pi}_2\omega_2\right)l_j\big/\left(\widehat{s}_{2j}\widehat{s}_{tj}\right)+c_{23}.
\end{aligned} \quad \text{(Equation 16)}$$

We separately regress the squared Z scores z_{1j}^2 on $l_j/\widehat{s}_{1j}^2$, z_{2j}^2 on $l_j/\widehat{s}_{2j}^2$, and $z_{1j}z_{2j}$ on $l_j/(\widehat{s}_{1j}\widehat{s}_{2j})$. The slopes $\widehat{\omega}_1,\widehat{\omega}_2,\widehat{\omega}_{12}$ and intercepts $\widehat{c}_1,\widehat{c}_2,\widehat{c}_{12}$ are estimates of $\omega_1,\omega_2,\omega_{12}$ and c_1,c_2,c_{12}, respectively. Also, the intercepts $\widehat{c}_3,\widehat{c}_{13},\widehat{c}_{23}$ of regressing $z_{tj}^2,z_{1j}z_{tj},z_{2j}z_{tj}$ on $l_j/\widehat{s}_{tj}^2, l_j/(\widehat{s}_{1j}\widehat{s}_{tj}), l_j/(\widehat{s}_{2j}\widehat{s}_{tj})$ give the estimates of c_3,c_{13},c_{23}, respectively. In practice, if there are no sample overlap, then the matrix $\mathbf{C}$ can be set as a diagonal matrix to alleviate the burden of parameter estimation, making the estimation of (co-)heritability more stable. Besides, since the *cis*-SNPs of a gene are typically in a small number (tens to hundreds), to avoid unreasonable coheritability estimates, we adopt a conservative approach: setting $\widehat{\omega}_{12}=0$ if the p value of the Wald test $\left(\frac{\widehat{\omega}_{12}}{se(\widehat{\omega}_{12})}\right)$ is smaller than a threshold (e.g., 0.05).

Finally, we directly plug the estimated parameters into Equations 14 and 15 to obtain IBSEP estimators $\{\widehat{b}_{1j}^{\mathrm{IBSEP}},\widehat{b}_{2j}^{\mathrm{IBSEP}}\}$ and their variances $\mathrm{Var}(\widehat{b}_{1j}^{\mathrm{IBSEP}}),\mathrm{Var}(\widehat{b}_{2j}^{\mathrm{IBSEP}})$. Cell-type-specific *cis*-SNPs with non-zero effects on the target gene can be detected using the Wald test based on IBSEP Z scores $\{z_{1j}^{\mathrm{IBSEP}}=\widehat{b}_{1j}^{\mathrm{IBSEP}}/\sqrt{\mathrm{Var}(\widehat{b}_{1j}^{\mathrm{IBSEP}})},z_{2j}^{\mathrm{IBSEP}}=\widehat{b}_{2j}^{\mathrm{IBSEP}}/\sqrt{\mathrm{Var}(\widehat{b}_{2j}^{\mathrm{IBSEP}})}\}$.

Compared methods in simulation

In the simulation of type I error and power evaluation, the compared methods include "cell type 1," "cell types 1&2," and "tissue ieQTL." "Cell type 1" only uses the summary statistics of cell type 1 $\{\widehat{b}_{1j},\widehat{s}_{1j}^2\}$. "Cell types 1&2" uses the summary statistics of cell type 1 and cell type 2 $\{\widehat{b}_{1j},\widehat{s}_{1j}^2,\widehat{b}_{2j},\widehat{s}_{2j}^2\}$, which can be viewed as a special case of IBSEP without tissue eQTLs. The estimator of "cell types 1&2" is derived in a similar manner as the full version of IBSEP. Without tissue-level eQTL summary statistics, the conditional mean of the estimated marginal effects and the conditional variance become

$$\mathbb{E}\left[\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\end{pmatrix}\middle| b_{1j}\right]=\underbrace{\begin{pmatrix}1\\ \omega_{j,12}/\omega_{j,11}\end{pmatrix}}_{\boldsymbol{\lambda}_1}b_{1j}, \quad \text{(Equation 17)}$$

$$\mathrm{Var}\left[\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\end{pmatrix}\middle| b_{1j}\right]=\mathbf{A}\left(\boldsymbol{\Omega}_j-\frac{\boldsymbol{\omega}_{j,\cdot 1}\boldsymbol{\omega}_{j,\cdot 1}^T}{\omega_{j,11}}\right)\mathbf{A}^T+\widehat{\mathbf{S}}_j\mathbf{C}\widehat{\mathbf{S}}_j:=\mathbf{\Lambda}_1^{-1}, \quad \text{(Equation 18)}$$

where

$$\mathbf{A}=\begin{pmatrix}1&0\\0&1\end{pmatrix},\widehat{\mathbf{S}}_j=\begin{pmatrix}\widehat{s}_{1j}^2&0\\0&\widehat{s}_{2j}^2\end{pmatrix},\mathbf{C}=\begin{pmatrix}c_1&c_{12}\\c_{12}&c_2\end{pmatrix}. \quad \text{(Equation 19)}$$

The IBSEP-ct estimator and corresponding variance are as follows:

$$\begin{aligned}
\widehat{b}_{1j}^{\mathrm{IBSEP-ct}} &= \arg\min_b \mathbf{m}(b)^T\mathbf{\Lambda}_1\mathbf{m}(b)\\
&= \underbrace{\left(\boldsymbol{\lambda}_1^T\mathbf{\Lambda}_1\boldsymbol{\lambda}_1\right)^{-1}\boldsymbol{\lambda}_1^T\mathbf{\Lambda}_1}_{\mathbf{w}_1^T}\underbrace{\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\end{pmatrix}}_{\widehat{\mathbf{b}}_{\cdot j}}=\mathbf{w}_1^T\widehat{\mathbf{b}}_{\cdot j}.
\end{aligned} \quad \text{(Equation 20)}$$

$$\mathrm{Var}\left(\widehat{b}_{1j}^{\mathrm{IBSEP-ct}}\right)=\left(\frac{\partial \mathbf{m}^T}{\partial b}\mathbf{\Lambda}_1\frac{\partial \mathbf{m}}{\partial b}\right)^{-1}=\left(\boldsymbol{\lambda}_1^T\mathbf{\Lambda}_1\boldsymbol{\lambda}_1\right)^{-1}. \quad \text{(Equation 21)}$$

For "tissue ieQTL," we follow the commonly used approach to test the interaction between tissue-level gene expression and cell type proportion.[9] The linear mixed model with an interaction term is

$$\mathbf{y}_t=\beta+\beta_j\mathbf{x}_j+\gamma_k\boldsymbol{\pi}_k+\beta_{jk}(\mathbf{x}_j\times\boldsymbol{\pi}_k)+\epsilon_t, \quad \text{(Equation 22)}$$

where $\mathbf{x}_j\in\mathbb{R}^{N_t}$ is the genotype vector of SNP j, $\boldsymbol{\pi}_k\in\mathbb{R}^{N_t}$ is the proportion vector of cell type k, and $(\mathbf{x}_j\times\boldsymbol{\pi}_k)$ is the interaction term. We test $\beta_{jk}=0$: $\beta_{jk}\neq 0$ means SNP j is an ieQTL of cell type k. In the simulation study of two cell types, we consider cell type 1 as the target cell type and performed "tissue ieQTL" for cell type 1 only.

Data preparation in real data analysis for blood

Due to the close association between gene expression in blood cells and many diseases and the convenience of obtaining blood samples, there are rich data resources on tissue-level eQTL and cell-type-level eQTL studies in blood. The analysis below involves two cell-type-level PBMC eQTL datasets and two neutrophil eQTL datasets: Oelen et al. sequenced and analyzed a cell-type-level eQTL dataset of approximately 1 million PBMCs from 120 individuals (denoted as 1M sc-blood dataset), from which we obtained summary statistics for six cell types, including B cells, CD4 T cells, CD8 T cells, monocytes, natural killer (NK) cells, and dendritic cells (DCs)[18]; the OneK1K eQTL dataset by Yazar et al. was derived from the analysis of scRNA-seq data from 982 individuals, covering 14 PBMC sub-cell types (which we grouped into the aforementioned six cell types)[17]; Naranbhai et al. sequenced RNA expressions of neutrophils isolated from 93 healthy individuals by microarray and performed ct-eQTL mapping (denoted as the Neutro2015 dataset)[19]; and the BLUEPRINT neutrophil eQTL summary statistics were computed from the transcriptional

profiles of 196 individuals.[25] Additionally, we collected eQTL results for whole-blood tissue from the GTEx project, which was obtained from the analysis of bulk RNA-seq from 670 samples.[4]

For each dataset, variants in the 1000 Genomes Project reference panel with a minor-allele frequency (MAF) $< 1\%$ were excluded. We merged the summary statistics of GTEx whole blood, each cell type of 1M, Neutro2015, and each sub-cell type of OneK1K and BLUEPRINT with the reference panel and then aligned the effect alleles. We retained genes that were shared in all cell-type-level eQTL datasets and GTEx. Since 1M only considered SNPs within a 100 kb distance of each gene midpoint, we excluded genes with fewer than 50 SNPs, resulting in 5,121 genes. For each SNP, we calculated its LD score with SNPs mapped to the same gene using the LDSC package. We ran CIBERSORTx with the default settings and utilized the provided LM22 signature matrix to obtain the individual cell type proportion estimations.[26]

Data preparation in real data analysis for brain

For brain ct-eQTL analysis, the largest existing dataset of brain cell-type-level eQTLs was provided by Bryois et al. (denoted as sc-brain dataset),[13] obtained by analyzing scRNA-seq from 192 independent individuals of 8 cell types, including excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, microglia, oligodendrocyte progenitor cells (OPCs), endothelial cells, and pericytes. Besides, we collected a GTEx brain cortex eQTL dataset obtained by bulk RNA-seq from 205 individual subjects.[4]

Analogously, we merged the summary statistics of the GTEx brain cortex and each cell type of sc-brain with the 1000 Genomes Project reference panel. Considering the significant differences in the genes contained within different cell types of sc-brain (12,831 genes in excitatory neurons and 5,871 genes in pericytes), we did not take the intersection of genes across all eight cell types. As sc-brain mapped *cis*-eQTLs within a 1 Mb window of the transcription start site (TSS) of a gene, each gene typically contains thousands of SNPs. We therefore did not further filter genes. We ran CIBERSORTx to create the signature matrix for the 8 cell types using a Smart-seq reference[27] and then estimated the cell type proportions in GTEx brain cortex samples.

Colocalization

The colocalization analysis was conducted following the procedure described in Julien et al.[13] For brain-related diseases/traits, we defined the coordinates of SNPs in LD ($r^2 > 0.1$ in the 1000 Genome European [EUR] cohort) with the reported lead SNPs as loci using LDlinkR.[28] For blood-related diseases/traits, we defined lead SNPs as the SNPs passing the genome-wide significant threshold ($p < 5 \times 10^{-8}$) with the most significant p value in each cytoband. For each locus, we tested the colocalization between the GWAS and eQTL signals for genes with at least 10 SNPs overlapping with this locus using the "coloc.abf" function of the *Coloc* R package.[29] coloc.abf outputs "PP.H4.abf," which is the posterior probability that the gene and the disease share a same causal SNP. We defined colocalized genes as those with PP.H4.abf greater than 0.7 and colocalized loci as those containing at least one colocalized gene.

Results

Simulation study

One of the key innovations of IBSEP is the theoretical justification that only the average cell type proportions of tissue samples are required in the integration analysis (supplemental methods section 3.3). We conducted simulations to verify this approximation using real genotypes from UK Biobank (UKBB) individuals. We compared the marginal effect size estimates calculated from tissue models with individual ($\widehat{\boldsymbol{b}}_{ind}$) and average ($\widehat{\boldsymbol{b}}_{avg}$) cell type proportions under different samples sizes and cell type proportion variances (supplemental methods section 3.4). Consistent with the theoretical result, the Pearson correlation coefficient (PCC) between $\widehat{\boldsymbol{b}}_{ind}$ and $\widehat{\boldsymbol{b}}_{avg}$ was already high with a small sample size (e.g., $N_t = 100$), and it approached 1 as N_t grew (Figure S1). Correspondingly, the mean absolute error (MAE) between $\widehat{\boldsymbol{b}}_{ind}$ and $\widehat{\boldsymbol{b}}_{avg}$ decreased as N_t increased. Besides, $\widehat{\boldsymbol{b}}_{avg}$ became closer to $\widehat{\boldsymbol{b}}_{ind}$ in terms of both PCC and MAE, with smaller cell type proportion variance (Figure S2). We also extended the above simulation study to multiple cell type settings and verified that the PCC and MAE between $\widehat{\boldsymbol{b}}_{ind}$ and $\widehat{\boldsymbol{b}}_{avg}$ were almost equal to one and zero, respectively, across various cell type proportion variances and sample sizes (Figures S4 and S5).

After validating the foundation of our model, we conducted a series of experiments to evaluate the performance of IBSEP. Firstly, we examined whether IBSEP could control type I errors under different scenarios. For simplicity but without loss of generality, we only considered two cell types (cell type 1 and cell type 2). We randomly selected 100 samples from the UKBB as the genotypes for scRNA-seq data ($\mathbf{X}_c$) and another 1,000 samples for bulk RNA-seq data ($\mathbf{X}_t$). Specifically, we used the first 1,000 HapMap3-matched SNPs from chromosome 20 in our simulation study. We treated these SNPs as local SNPs mapped to a target gene and performed ct-eQTL analysis. To mimic the complicated *cis*-SNP distributions in a realistic biological context, we treated cell type 1 as the target cell type and partitioned these SNPs into two segments. The first segment, taking up 80%, was considered the null region, which contained *cis*-SNPs in cell type 2 only, and the remaining 20% was considered the non-null region, which contained shared *cis*-SNPs between the two cell types. We randomly selected $p_{cis} \in \{0.5\%, 2\%, 5\%\}$ SNPs in cell type 2 as *cis*-SNPs (Figure 2A), with their true effects following

$$\begin{pmatrix} \beta_{1j} \\ \beta_{2j} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} h_1^2 & r_g h_1 h_2 \\ r_g h_1 h_2 & h_2^2 \end{pmatrix} \Big/ Mp_{cis}\right), j \in \mathcal{I}_{cis}, \quad \text{(Equation 23)}$$

where $\mathcal{I}_{cis}$ collected all indices of *cis*-SNPs, and the effects of the remaining SNPs were set to 0. We set *cis*-heritability $h_1^2 = 0, h_2^2 \in \{1\%, 5\%, 10\%\}$ and genetic correlation $r_g = 0$ for *cis*-SNPs in the null region. Regarding the *cis*-SNPs in the non-null region, we set $h_1^2 = h_2^2 \in \{1\%, 5\%, 10\%\}$ and $r_g = 0.3$. We generated gene expression levels for both cell types according to

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}_c\boldsymbol{\beta}_1 + \epsilon_1, \\ \mathbf{y}_2 &= \mathbf{X}_c\boldsymbol{\beta}_2 + \epsilon_2, \end{aligned} \quad \text{(Equation 24)}$$

Please cite this article in press as: Yu et al., A unified framework for cell-type-specific eQTL prioritization by integrating bulk and scRNA-seq data, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2024.12.018

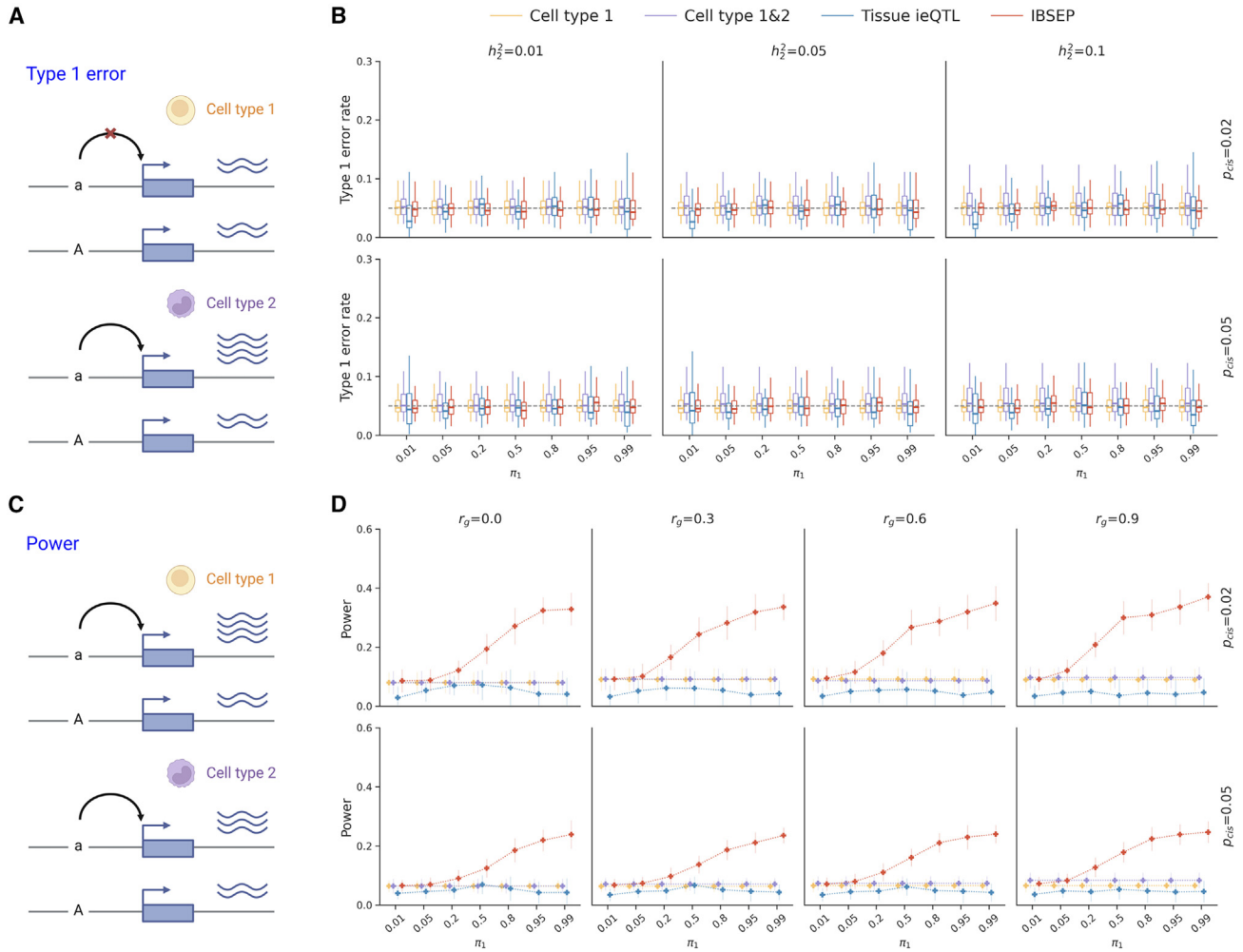

Figure 2. Simulation result

(A) Diagram of simulation on type I error rate assessment.

(B) Type I error rates of IBSEP and compared methods under different *cis*-heritability of cell type 2 (h_2^2), fractions of *cis*-SNPs (p_{cis}), and mean cell type 1 proportion in tissue samples (π_1).

(C) Diagram of simulation on power assessment.

(D) Statistical power of IBSEP and compared methods under different genetic correlations between the two cell types (r_g), fractions of *cis*-SNPs (p_{cis}), and mean cell type 1 proportion in tissue samples (π_1).

Error bars represent the standard errors of power evaluated on 50 replications.

where ϵ_1 and ϵ_2 are the independent error terms. To generate tissue-level gene expression, we first generated the individual proportions of cell type 1 in the tissue samples from a beta distribution $\pi_{1i} \sim \text{Beta}(\alpha\pi_1, \alpha(1-\pi_1)), i = 1, \ldots, N_t$, where $\pi_1 \in \{0.01, 0.05, 0.2, 0.5, 0.8, 0.95, 0.99\}$ is the true mean proportion of cell type 1 and N_t is the tissue sample size; then, the proportion of cell type 2 is given as $1 - \pi_{1i}$, and we set $\alpha = 5$ to allow a greater variance of $\boldsymbol{\pi}_1$. Finally, we generated gene expression levels for the tissue samples based on

$$\mathbf{y}_t = \boldsymbol{\pi}_1 \odot \mathbf{X}_t\boldsymbol{\beta}_1 + \boldsymbol{\pi}_2 \odot \mathbf{X}_t\boldsymbol{\beta}_2 + \epsilon_t. \qquad \text{(Equation 25)}$$

Four methods for eQTL analysis of cell type 1 were performed: only using data of cell type 1 ("cell type 1"), using data of cell types 1 and 2 (denoted as "cell types 1&2"), ieQTL analysis using tissue data ("tissue ieQTL"), and IBSEP. We reported the fraction of SNPs with a p value less than 0.05 in the null region of cell type 1 as the type I error rate. As shown in Figures 2B and S6, all four methods produced well-controlled type I error rates regardless of the heritability of cell type 2, the fraction of *cis*-SNPs, or the cell type proportions.

Moreover, we conducted additional experiments to test the robustness of IBSEP under model mis-specification and inaccurate parameter estimation. In reality, there may be some other cell types present in tissue samples that are either unknown or not available in single-cell datasets. Building upon the aforementioned simulation on type I error, let us assume the existence of cell type 3 in tissue samples, which is not available in the summary statistics of single-cell data. Nevertheless, IBSEP with the incorrect "2-cell-type" assumption can still control the type I error rate as long as the proportion of cell type 3 was not substantial. As expected, IBSEP, with the correct

"3-cell-type" assumption, can control the type I error rate when single-cell data for cell type 3 were available (Figure S11). Another realistic situation is when the cell type deconvolution method inaccurately estimates individual cell proportions in tissue samples, leading to biases between the estimated and true mean cell type proportions. We simulated various scenarios of inaccurate average cell type estimation and found that the bias in mean cell type proportion estimation has little impact on IBSEP's type I error control (Figure S12).

Next, we evaluated the power of IBSEP and compared methods. The procedures for generating genotypes and gene expressions are the same as above. We assumed that cell type 1 and cell type 2 shared *cis*-SNPs across the entire region, with their effect sizes following Equation 23 (Figure 2C). Different from the type I error rate evaluation, we fixed $h_1^2 = h_2^2 = 0.05$ while varying the genetic correlation between two cell types, $r_g \in \{0, 0.3, 0.6, 0.9\}$. The power was calculated as the fraction of *cis*-SNPs with a p value less than 0.05 in the target cell type 1. As shown in Figure 2D, IBSEP was the overall winner compared with other methods. Across different fractions of *cis*-SNPs and genetic correlations, the statistical power of IBSEP increased with the proportion of cell type 1 in the tissue samples. This is because as the proportion of cell type 1 in the tissue samples increases, IBSEP can borrow more information from the tissue samples, leading to a more pronounced improvement in statistical power compared to the baseline model ("cell type 1"). "Cell types 1&2" leveraged the genetic correlation between the two cell types, resulting in a slight improvement over the baseline model, and the degree of improvement became more significant with higher r_g. However, when cell type 1 comprised a substantial proportion of the tissue samples (greater than 5%), IBSEP exceeded "cell types 1&2." This indicates that large-scale tissue samples can greatly assist in ct-eQTL prioritization. Lastly, we observed that although the "tissue ieQTL" was solely based on tissue samples, it can still match the statistical power of the baseline model under an ideal case where $\boldsymbol{\pi}_1$ exhibits both the largest variance and relatively large abundance. In the majority of scenarios, ieQTL mapping needs to be performed with caution (Figures S19 and S20), and its necessary conditions are discussed in supplemental methods section 3.11.

In addition to the above simulations with two cell types, we also thoroughly evaluated type I error rates and the power of IBSEP in the multiple-cell-type setting. To mimic the real situation, we considered six cell types, with the most abundant cell type having an average proportion of 0.5 and the rarest cell type having an average proportion of 0.02 in tissue samples. The detailed simulation design can be found in supplemental methods section 3.7. Across various scenarios, IBSEP well controlled the type I error rates of all the cell types (Figures S8 and S9). Overall, IBSEP had pronounced improvements in statistical power compared to the baseline model and "tissue ieQTL," especially for cell types comprising a substantial proportion (greater than 5%) in tissue samples (Figure S10).

To further explore IBSEP's characteristics and provide more insights and guidance of IBSEP, we conducted a series of sensitivity analyses. First, regarding IBSEP's efficacy at low-cell-type proportions, we determined the minimum cell type proportion required to observe a benefit from including bulk eQTL summary statistics. As shown in Figure S13, the minimum cell type proportion decreased as the tissue sample size N_t increased (supplemental methods section 3.8). Second, considering that cell type abundance in tissue samples had an important impact on IBSEP, we determined the minimum sample size of bulk data required to help improve ct-eQTL prioritization. As expected, in the two-cell-type setting with the sample size of scRNA-seq data $N_c = 100$, when the mean fraction π_1 of the rare cell type was as low as 0.01, the minimum N_t required for IBSEP to show a robust improvement over the baseline model was approximately 3,000. As the proportion increased to 0.1, the minimum sample size decreased to 400, and for a relatively abundant cell type with $\pi_1 = 0.3$, it was further reduced to 200 (Figure S14). More details are provided in supplemental methods section 3.9. Third, we evaluated the sensitivity of IBSEP to the systematic donor-level differences between bulk and scRNA-seq data, particularly in the context of varying ancestry backgrounds. We compared the performance of IBSEP by integrating the bulk RNA-seq data from EUR with scRNA-seq data from either the same population or a different one, such as East Asian (EAS). Not surprisingly, IBSEP's statistical power was slightly attenuated in the presence of ancestry disparities. However, IBSEP still enhanced ct-eQTL prioritization for the ancestry group different from the bulk data as long as the mean cell type proportion π_1 was not rare and the tissue sample size N_t was sufficient (supplemental methods section 3.10).

Real data analysis

We applied IBSEP for ct-eQTL mapping and downstream analysis in two tissues: blood and brain cortex. For both tissues, we obtained tissue-level eQTLs from GTEx and cell-type-level eQTLs from scRNA-seq and then integrated them using IBSEP to obtain the improved ct-eQTL results. Compared to the original cell-type-level eQTL results, IBSEP ct-eQTLs achieved significant improvements in statistical power. Specifically, IBSEP identified more ct-eQTLs, a substantial portion of which could be replicated in a larger cell-type-level eQTL dataset or validated by biological evidence from epigenomic data. Furthermore, colocalization analysis between IBSEP ct-eQTLs and disease risk variants revealed more genes that potentially mediate disease development, further indicating that IBSEP has detected more eQTLs relevant to disease mechanisms in the cell type level.

IBSEP in blood ct-eQTL analysis

Blood ct-eQTL prioritization

We combined 1M with Neutro2015 as a cell-type-level eQTL dataset containing six PBMC cell types and neutrophils and

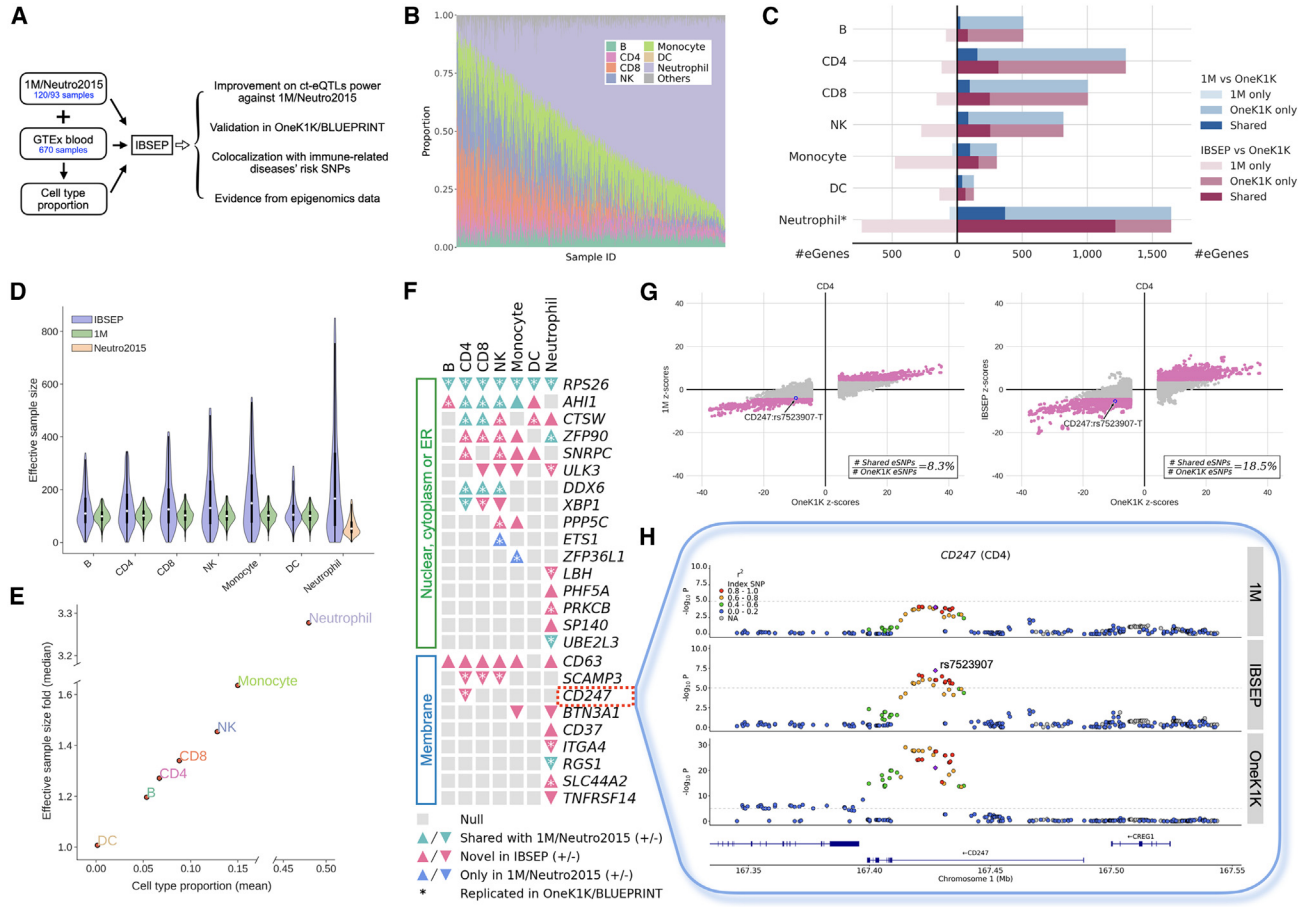

Figure 3. Results of blood ct-eQTL prioritization

(A) Workflow of IBSEP's blood ct-eQTL analysis.

(B) Individual cell type composition estimates for GTEx blood samples by CIBERSORTx.

(C) Comparison of the number of eGenes discovered by 1M/IBSEP with OneK1K in each PBMC cell type and those discovered by Neutro2015/IBSEP with BLUEPRINT in neutrophils.

(D) Distribution of effective sample sizes of 1M and IBSEP for genes in each PBMC cell type and those of Neutro2015 and IBSEP in neutrophils.

(E) Comparison between the mean cell type proportion and the incremental fold of median effective sample size.

(F) ct-eQTLs of genes related to nuclei, cytoplasm, or the ER and membrane encoding discovered by 1M/Neutro2015 and IBSEP. Up/down triangles represent the up-/down-regulatory functions of the effect allele of the lead ct-eQTL, respectively.

(G) Comparison of *Z* scores of eSNPs between OneK1K and 1M (left) or IBSEP (right) in CD4 cells.

(H) Locuszoom plots of 1M, IBSEP, and OneK1K around *CD247* in CD4 cells.

performed integrative analysis of it with the GTEx dataset to obtain the IBSEP statistics for these seven cell types. Because of the larger sample sizes, OneK1K and BLUEPRINT served as the validation datasets for PBMCs and neutrophils, respectively. Firstly, using CIBERSORTx,[24] we estimated the proportions of the seven cell types in each GTEx blood sample and then calculated the average cell type proportions. Subsequently, we input the cell-type-level eQTL summary statistics from 1M and Neutro2015, the tissue-level eQTL summary statistics from GTEx, and the average cell type proportions into IBSEP to obtain the improved ct-eQTL summary statistics (Figure 3A). We conducted a comprehensive and detailed comparison of the ct-eQTL results from 1M/Neutro2015, OneK1K/BLUEPRINT, IBSEP, and the tissue-level eQTL results from GTEx; the detailed results are available in Tables S5 and S6. Among the seven studied cell types, neutrophils, as the most abundant cell type in whole-blood samples, account for approximately 48% on average; monocytes, NK cells, T cells, and B cells together make up a large portion of PBMCs, while DCs account for a minor proportion (Figures 3B and S21; Tables S2 and S3). Our cell type proportion estimations align with established scientific knowledge about cell composition in blood samples.[30,31]

Throughout real data analysis, SNPs with p values less than 10^{-5} are defined as eSNPs, and genes containing at least one eSNP are referred to as eGenes. The same p value threshold was applied across all datasets and IBSEP, and a comparison with the commonly adopted eGene/eSNP definitions through false discovery rate (FDR) control is provided in supplemental methods section 3.12. Figure 3C shows the comparison of eGenes identified by

1M/Neutro2015, OneK1K/BLUEPRINT, and IBSEP (Table S4). First and foremost, it is evident that IBSEP detected significantly more eGenes than 1M/Neutro2015 in all cell types, and a substantial proportion of them can be found in the validation dataset (OneK1K/BLUEPRINT). Specifically, for the six PBMC cell types, due to the small sample size, 1M only replicated a small portion of the OneK1K eGenes: from the lowest 4.7% in B cells to the highest 32.5% in monocytes. IBSEP significantly increased the replication rate, for example, from 12% to 24.5% in CD4 cells and from 10.4% to 31.1% in NK cells. For neutrophils, IBSEP even raised the replication rate from 22.4% to 73.9%. Among the IBSEP eGenes that were not replicated in OneK1K/BLUEPRINT, most were identified by GTEx (Figure S25), indicating that the discovery of these eGenes can be attributed to the tissue-level, high-quality GTEx sources, and IBSEP allows for them to be specified at the cell type level. Most importantly, although the sample size of GTEx is almost six times larger than that of 1M/Neutro2015, IBSEP's results were not dominated by GTEx. This is evidenced by the significantly higher proportion of IBSEP eGenes replicated in OneK1K/BLUEPRINT compared to those from GTEx across all the cell types (Figure S26). For instance, OneK1K identified 1,296 eGenes for CD4 cells, and randomly selecting 1,296 genes from all genes would yield an average overlap of only 25% with OneK1K eGenes. With GTEx eGenes, this proportion increased to 39%, reflecting the higher likelihood that tissue-level eGenes correspond to cell-type-level eGenes. In contrast, an impressive 73% of CD4 eGenes identified by IBSEP were replicated in OneK1K, demonstrating that IBSEP indeed captured information beyond tissue-level resolution.

Moreover, we counted how many cell types the 1M/IBSEP/OneK1K eGenes were displayed in (restricted to the six PBMC cell types). As shown in Figure S27, 1M, IBSEP, and OneK1K exhibited similar distributions: eGenes displayed in only 1–2 cell types accounted for the majority, while only a small portion of eGenes were shown in 5–6 cell types. This suggests that the cell type specificity of single-cell eQTLs was well preserved by IBSEP. In addition to eGenes, we also inspected eSNPs of IBSEP. Comparing the Z scores of 1M/Neutro2015 and IBSEP with OneK1K/BLUEPRINT (Figure S24), we have the following observations. First, the sign of IBSEP eSNP Z scores remained in high concordance with the validation datasets, indicating that IBSEP did not distort the effect direction of significant SNPs. Second, compared with 1M/Neutro2015, IBSEP helped more SNPs to reach the significance threshold and thus retrieved more eSNPs in OneK1K/BLUEPRINT. Third, IBSEP exhibited the most notable improvement of Z scores for neutrophils, which was consistent with the greatest gain in eGene count for neutrophils. To further quantify IBSEP's improvement in statistical power, we calculated the effective sample sizes of 1M/Neutro2015 and IBSEP by gene and cell type using the χ^2 statistics (Figure 3D). Overall, the median effective sample size of IBSEP was 1.12 (DC) to 1.76 (monocyte) times larger than that of 1M and 3.96 times larger than that of Neutro2015. In other words, the sample size of 1M would need to be increased by 12%–76% to achieve an equivalent power gained by IBSEP or 296% for Neutro2015. The improvement in statistical power of IBSEP primarily stemmed from tissue-level eQTLs with larger sample sizes and higher data quality, as the increase in effective sample size was proportional to the corresponding cell type proportions in tissue samples (Figure 3E).

The ultimate goal of identifying ct-eQTLs is to establish links between diseases and their putative target genes, thereby revealing cellular-level disease regulatory mechanisms. Next, we focused on genes involved in two pathways: nuclei, cytoplasm, or endoplasmic reticulum (ER) protein encoding and membrane protein encoding, which are closely related to autoimmune diseases. As shown in Figure 3F, due to its small sample size, 1M/Neutro2015 only identified 8 eGenes associated with nuclei, cytoplasm, or the ER and one for membrane. In contrast, IBSEP identified 25 eGenes, most of which could be replicated in OneK1K/BLUEPRINT. Locuszoom plots of several example eGenes provide insight into IBSEP's mechanisms. For instance, *CD247* (MIM: 186780) encodes the CD3ζ chain, a crucial component of the T cell receptor (TCR) complex, essential for modulating T cell activity and proliferation, thus playing a vital role in the adaptive immune system.[32,33] Very consistently, IBSEP newly identified *CD247* as an eGene for CD4 cells, and it is replicable in OneK1K. Figure 3H shows that in 1M, the p values of rs7523907-T and nearby SNPs were close to significance, and IBSEP helped to reduce these p values, causing them to exceed the significance threshold. As shown in Figure 3G, both IBSEP and OneK1K demonstrated that rs7523907-T is associated with the down-regulation of *CD247* expression, indicating the reliability of IBSEP. From the perspective of autoimmune disease GWASs, rs7523907-T has been reported to be associated with an increased risk of asthma.[34] The gene behind rs7523907-T, *CD247*, was linked to multiple autoimmune conditions, including systemic sclerosis[35,36] and hematological phenotypes.[37,38] Similarly, IBSEP boosted rs62136101-C of *PPP5C* (MIM: 600658) in NK to be an eSNP, which turned out to be a highly significant SNP in OneK1K (Figure S28). As another example, *ZFP90* (MIM: 609451), a gene encoding a member of the zinc-finger protein family, affects a range of cellular processes, such as differentiation, development, and immune responses. Concordant with its widespread biological role, IBSEP identified *ZFP90* as an eGene across several cell types, including CD4 cells, CD8 cells, NK, monocytes, and neutrophils, suggesting its lower specificity to any single cell type. Examining the Locuszoom plots of *ZFP90* in CD4 cells and NK, we observed that although this gene was easily discovered by OneK1K, it failed to be identified in 1M (Figures S29 and S30). By leveraging higher-power, tissue-level eQTL summary statistics from the same population, IBSEP overall increased the significance of potential

eSNPs while preserving the *p* value patterns within the *cis* window, allowing *ZFP90* to be prioritized on the cell type level.

Colocalization between blood ct-eQTLs and genetic risk SNPs of immune-related phenotypes

Finally, we applied *Coloc*[29] to conduct colocalization analysis between ct-eQTLs and risk variants of six traits/diseases, including two hematological phenotypes, monocyte count and neutrophil count from UKBB, and four common autoimmune diseases, rheumatoid arthritis (RA [MIM: 180300]),[39] multiple sclerosis (MS [MIM: 126200]),[40] inflammatory bowel disease (IBD [MIM: 266600]),[41] and asthma (MIM: 600807) from the UKBB. *Coloc* is a Bayesian method that compares a gene's eQTL *Z* scores and the corresponding GWAS *Z* scores to provide the posterior probability that the gene and the phenotype share the same causal SNP. We defined colocalized genes as those with a posterior probability greater than 0.7 and the colocalized loci as those with at least one colocalized gene. As expected, for each phenotype and each cell type, more colocalized loci/genes were identified using IBSEP's ct-eQTL results compared to 1M/Neutro2015 (Figure 4A). This can be attributed to IBSEP's advantage in the power of ct-eQTL prioritization, and a higher *Z* score of the causal SNP leads to an increased posterior probability. A clear example is neutrophils, where the results of ct-eQTL prioritization showed that the amount of eGenes identified by IBSEP was 3–11.7 times more than that found in PBMC cell types. Consequently, among the six phenotypes studied, the loci/genes colocalized by IBSEP in neutrophils were also several times greater than those in PBMC cell types (Table S11). Another interesting finding is that the increment in the number of colocalized loci across various cell types was more pronounced in monocyte count and neutrophil count than in autoimmune diseases. This is possibly because monocyte count and neutrophil count are continuous traits and the larger sample size for association mapping contributes to higher GWAS power. Once the power of ct-eQTLs is enhanced, more potential colocalized genes will be uncovered.

We further inspect the monocyte count, which exhibits rich results in the colocalization analysis. The monocyte count is a vital index in hematology: monocyte elevations usually indicate the occurrence of inflammation.[42] Therefore, identifying the genetic influences underlying the monocyte count is essential to characterizing the immune responses.[43] Figure 4B shows that IBSEP identified many colocalized genes not found by 1M/Neutro2015, a substantial portion of which could be replicated in validation datasets. Among the IBSEP colocalized genes, a large portion was only observed in no more than three cell types, while a small fraction was shared across multiple cell types. A notable example is *OSER1*, which was colocalized by IBSEP in neutrophils and PBMCs but not DCs. BLUEPRINT supported the colocalization of *OSER1* in neutrophils. As shown in Figure 4C, the peaks of the monocyte count GWAS and the BLUEPRINT *OSER1* eQTLs colocalized at rs2143606, located in the promoter of *OSER1*. However, the H3K27ac chromatin immunoprecipitation (ChIP)-seq track[44] in this region demonstrates that epigenomic signals are not only present in neutrophils but also in several other cell types (B cells, CD4 cells, CD8 cells, NK cells, and monocytes), indicating that the effect of rs2143606 on the monocyte count through *OSER1* is likely not neutrophil specific. Indeed, IBSEP successfully colocalized *OSER1* across these cell types. As an illustration, in monocytes, IBSEP prioritized rs2143606 with a sufficiently significant *p* value, making *OSER1* a colocalized gene for monocytes.

In contrast, *MARK3* (MIM: 602678) represents a more cell-type-specific colocalized gene identified by IBSEP in B cells, monocytes, and neutrophils. Figure 4D shows that monocyte count GWAS *p* values were most significant at a 10 kb high LD region (chr14:103.84–103.85 Mb) upstream of *MARK3*'s promoter. The H3K27ac peaks at the promoter region were most prominent in B cells, monocytes, and neutrophils, suggesting that the *cis*-SNPs in this region may influence the monocyte count by regulating the expression of *MARK3* in these cell types. We take monocytes as an example to explain why *MARK3* was colocalized by IBSEP rather than 1M. Clearly, the mismatch between the top, but not significant, SNP in 1M and the most significant region of GWAS hindered *MARK3* from being colocalized by 1M. In contrast, the improved ct-eQTLs of this gene provided by IBSEP made the high LD region stood out to successfully colocalize with GWAS. For other five traits/diseases, the comparison results of cell-type-level colocalized loci/genes between 1M/Neutro2015 and IBSEP are displayed in Figures S34–S38.

IBSEP in brain ct-eQTL analysis

Brain ct-eQTL prioritization

In addition to blood tissues, brain cell-type-level eQTL analysis is another hot research topic. The human brain is a highly complex organ containing billions of neurons and various types of glial cells. Therefore, the mechanisms of brain disorders are complex, involving multiple cell types, each with its own distinct role, necessitating the study of brain ct-eQTLs. As the sample size of the existing cell-type-level eQTL (sc-brain) dataset for brain is not large enough, to identify more potential brain ct-eQTLs, we applied IBSEP to integrate these brain ct-eQTL results and the GTEx brain cortex eQTL dataset, following a similar procedure to that for the blood tissue (Figure 5A). Detailed brain ct-eQTL results are available in Table S10. First, we obtained the bulk RNA-seq data of GTEx brain cortex samples and estimated the proportions of the above-mentioned 8 cell types (Figure 5B; Tables S7 and S8) and then applied the averages for integration (Figure S39). Reasonably, neuronal cells (excitatory neurons and inhibitory neurons) account for approximately 41% on average, while glial cells (astrocytes, oligodendrocytes, microglia, and OPCs) account for approximately 51% on average,

Please cite this article in press as: Yu et al., A unified framework for cell-type-specific eQTL prioritization by integrating bulk and scRNA-seq data, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2024.12.018

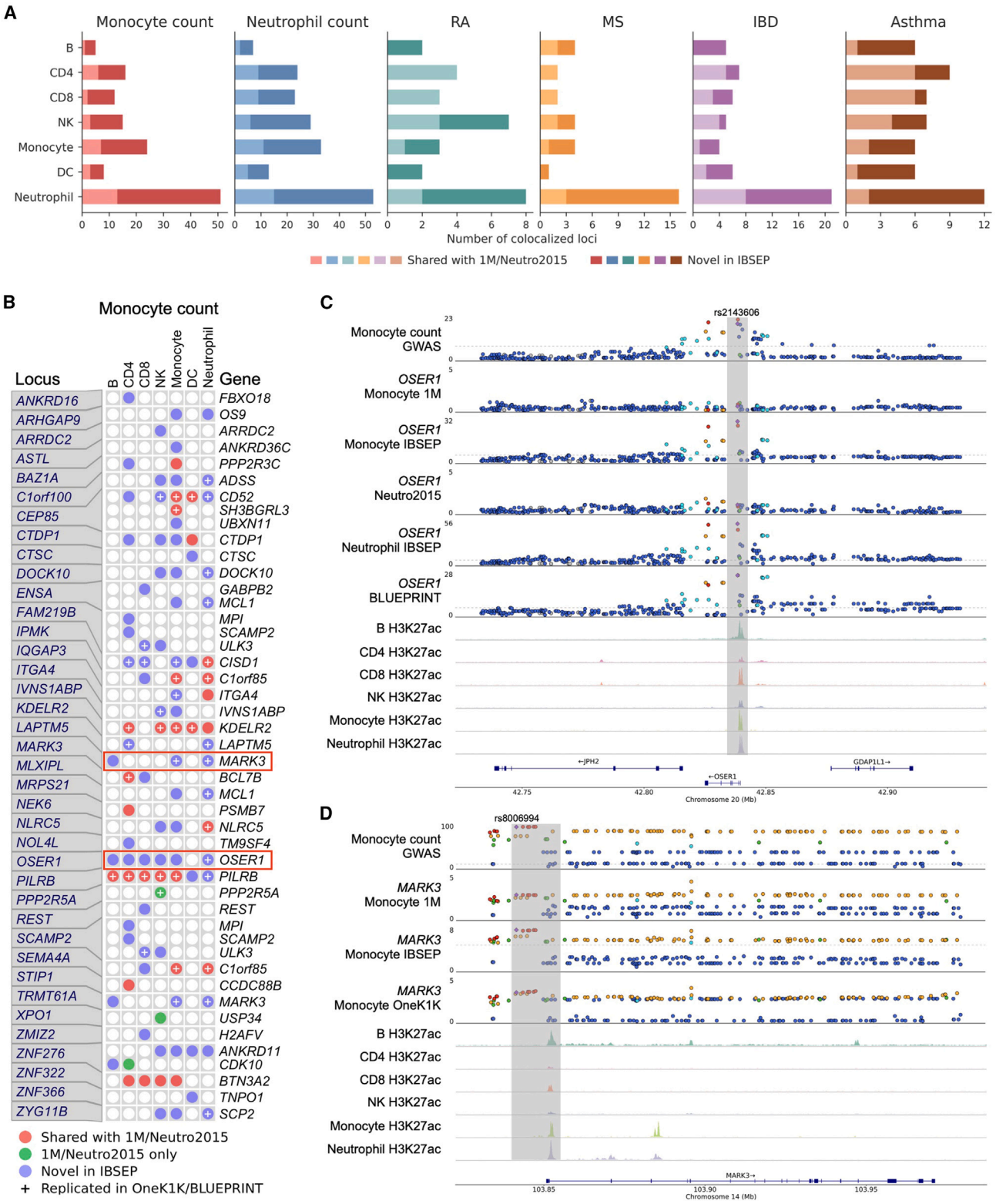

Figure 4. Results of colocalization between blood ct-eQTLs and genetic risk SNPs of immune-related phenotypes

(A) Comparison of the number of colocalized genes identified by 1M/Neutro2015 and IBSEP across six phenotypes and six blood cell types.
(B) Cell-type-level colocalized genes of monocyte count found by 1M/Neutro2015 and IBSEP. For better visualization, genes colocated only in neutrophils are not displayed.
(C) Locuszoom plots of monocyte count GWAS, 1M, Neutro2015, IBSEP, and BLUEPRINT around *OSER1* and corresponding H3K27ac tracks.
(D) Locuszoom plots of monocyte count GWAS, 1M, IBSEP, and OneK1K around *MARK3* and corresponding H3K27ac tracks.
In the Locuszoom plots of GWAS and eQTL, the dashed lines represent −log *p* values of 8 and 5, respectively.

Please cite this article in press as: Yu et al., A unified framework for cell-type-specific eQTL prioritization by integrating bulk and scRNA-seq data, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2024.12.018

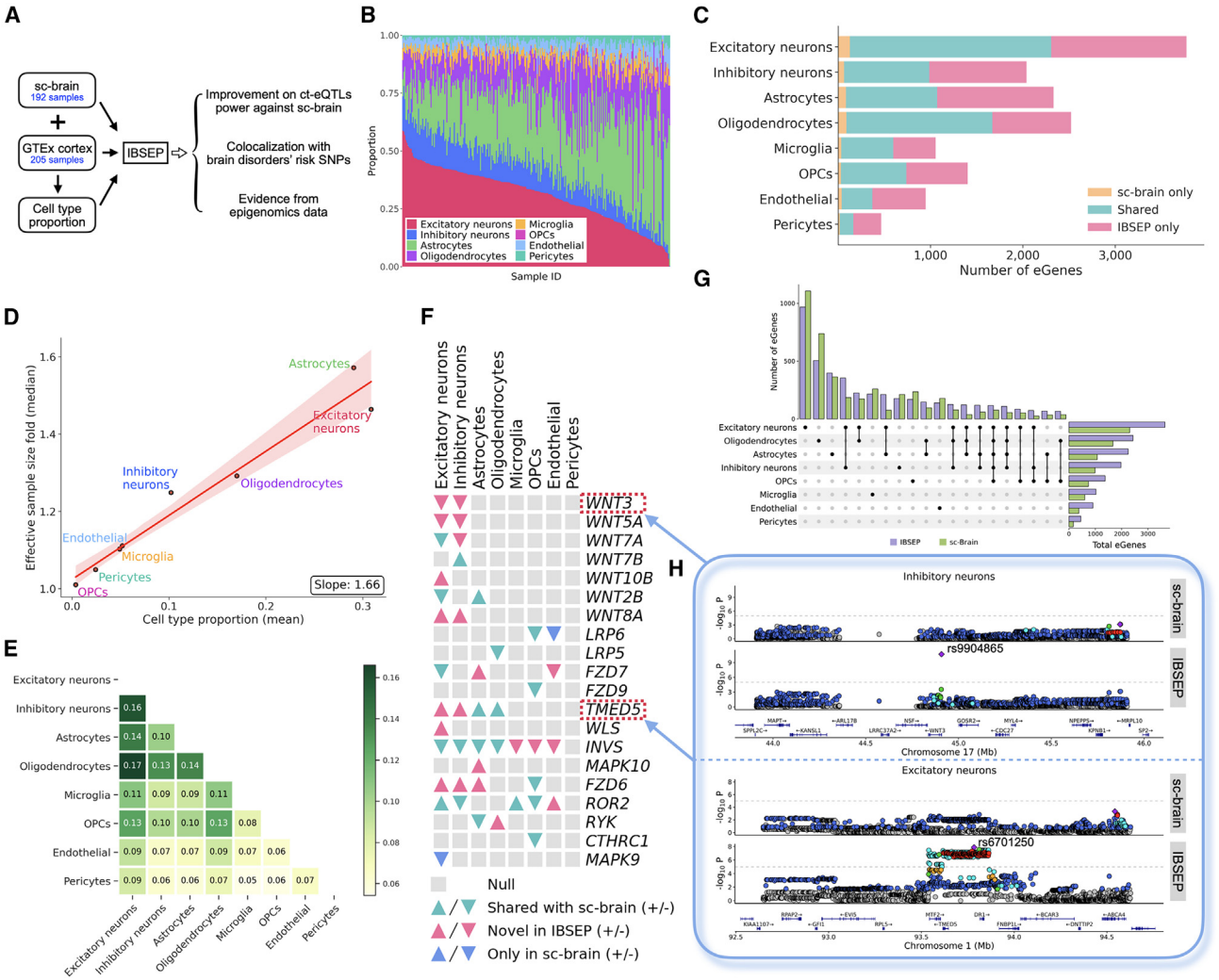

Figure 5. Results of brain ct-eQTL prioritization

(A) Workflow of IBSEP's brain ct-eQTL analysis.
(B) GTEx brain cortex samples' individual cell type proportion estimation.
(C) Comparison of the number of eGenes discovered by sc-brain and IBSEP in each cell type.
(D) Comparison between the mean cell type proportion and the incremental fold of median effective sample size.
(E) Heatmap shown the fraction of genes with significant genetic correlation between cell type pairs.
(F) ct-eQTLs of genes in Wnt pathway discovered by sc-brain and IBSEP. Up/down triangles represent the up-/down-regulatory functions of the effect allele of the lead ct-eQTL, respectively.
(G) UpSet plot of the number of eGenes discovered by IBSEP and sc-brain across various cell types and their intersections.
(H) Locuszoom plots of sc-brain and IBSEP around *WNT3* and *TMED5*.

with only a small proportion of cells being endothelial and pericytes.

Figure 5C compares the number of eGenes discovered by sc-brain and IBSEP across the eight cell types (Table S9). sc-brain identified 164–2,307 eGenes, while IBSEP achieved significant power gains by identifying 449–3,644 eGenes, with a considerable proportion being newly discovered in each cell type. For instance, IBSEP newly identified 1,464 eGenes in excitatory neurons, which represents the largest proportion in GTEx samples, and thus, more information could be leveraged from GTEx data; for microglia that only represent a smaller proportion in GTEx samples, IBSEP identified 457 eGenes not found by sc-brain and that are thus considered novel eGenes. Despite the sample size of GTEx not being significantly larger than sc-brain, IBSEP still demonstrates a substantial improvement in statistical power due to its high-quality RNA-seq data. To investigate the origin of these novel eGenes discovered by IBSEP, we conducted ct-eQTL analysis using summary statistics from sc-brain only (denoted as IBSEP-ct); IBSEP-ct only utilized genetic correlations between cell types without leveraging tissue-level information. Consistent with the simulation results, although the statistical power of IBSEP-ct was slightly lower than that of IBSEP, it still identified more eGenes compared to sc-brain (Figure S43). This is because there are widespread genetic correlations among cell types, especially biologically close ones, such as excitatory neurons with inhibitory neurons, allowing IBSEP-ct

to effectively utilize this to improve statistical power (Figure 5E). Besides, we observed that 81%–96% of eGenes newly discovered by IBSEP were included in GTEx or IBSEP-ct (Figures S41 and S44), and the number of novel eGenes present exclusively in GTEx but not IBSEP-ct was highly positively correlated with the mean cell type proportion in GTEx samples (Figure S45). This combined evidence indicates that the novel eGenes identified by IBSEP can be attributed to both inter-cell-type genetic correlations and tissue-level eQTLs.

When sorting the number of eGenes by cell type count, we observed that the ranks were highly consistent between IBSEP and sc-brain: the number of eGenes decreases as the number of shared cell type increases (Figure S46). However, compared to sc-brain, IBSEP increased the proportion of eGenes observed in multiple cell types, which is a natural outcome of statistical power improvement.[17] More specifically, as shown in Figure 5G, eGenes specific to excitatory neurons, oligodendrocytes, and astrocytes rank in the top three in both sc-brain and IBSEP, which is related to the abundance of these three cell types in scRNA-seq and GTEx samples. Among eGenes observed in two cell types, the top-ranking pairs include excitatory neurons-inhibitory neurons, excitatory neurons-oligodendrocytes, excitatory neurons-astrocytes, and oligodendrocytes-astrocytes, which aligns with their higher statistical power and biological correlation between these cell type pairs. We further calculated the effective sample size of sc-brain and IBSEP in each cell type. As expected, the effective sample sizes of IBSEP were generally higher than those of sc-brain in all cell types (Figure S47). Specifically, the median effective sample size of IBSEP was 1.28–1.74 times that of sc-brain, and these folds were significantly positively correlated with the average proportion of cell types in GTEx samples (Figure 5D).

The development and function of the nervous system are regulated by multiple signaling pathways, among which the Wnt pathway is considered to be one of the most important pathways in neurons for maintaining the normal function and structure of the nervous system.[45,46] Using a public human brain single-cell atlas,[27] we observed that neurons exhibit higher Wnt pathway gene expressions than other cell types (Figure S48). Therefore, we focused on examining the ct-eQTLs of genes related to the Wnt pathway and compared their ct-eQTL results between IBSEP and sc-brain. As shown in Figure 5F, IBSEP identified 19 eGenes associated with the Wnt pathway, 6 of which are not reported by sc-brain. Most importantly, among these eGenes discovered by IBSEP, excitatory neurons or inhibitory neurons showed significant enrichment compared with other cell types, which was in line with crucial biological functions of the Wnt pathway in neurons. Specifically, IBSEP identified 7 genes in the *WNT* family, all of which were displayed in neurons. The *WNT* gene family is core to the Wnt pathway and associated with various brain disorders.[47,48] For example, *WNT3* (MIM: 165330) is linked to several disorders, including depression, Alzheimer disease (AD [MIM: 104300]), and Parkinson disease (MIM: 168601).[49–51] IBSEP found *WNT3* to be an eGene for excitatory neurons and inhibitory neurons, whereas sc-brain failed to identify it. As shown in the top plot of Figure 5H, the ct-eQTL signals for *WNT3* in inhibitory neurons are weak in sc-brain. By contrast, IBSEP successfully identified a significant eQTL, rs9904865, in the promoter region of *WNT3*. Another example is *TMED5* (MIM: 616876), a gene reported to be associated with educational attainment and AD.[52,53] sc-brain identified *TMED5* as an eGene for astrocytes and oligodendrocytes only, while IBSEP additionally found it to be an eGene in neurons. As shown in the bottom plot of Figure 5H, there is no significant signal for *TMED5* in sc-brain on excitatory neurons, whereas IBSEP showed strong signals around *TMED5* spanning an ~300 kb region.

Colocalization between brain ct-eQTLs and genetic risk SNPs of brain disorders

To explore the role of specific cell types in brain disorders, we conducted colocalization analysis for brain ct-eQTLs and risk loci from four brain disorder GWASs, including AD,[54] bipolar disorder (BIP [MIM: 125480]),[55] major depression disease (MDD [MIM: 608516]),[56] and schizophrenia (SCZ [MIM: 181500]).[57] The power of the colocalization analysis was strongly enhanced with IBSEP results: using sc-brain's ct-eQTL results, we can only identify 24–120 colocalized loci across the four diseases, while leveraging IBSEP's ct-eQTL results revealed 34–191 colocalized loci (Figure 6A; Table S12). Furthermore, for each disease, IBSEP maintained a generally consistent relative abundance of colocalized loci across cell types compared to sc-brain. For instance, regarding AD, both IBSEP and sc-brain identified most colocalized loci in microglia, in concordance with the accumulated scientific understanding of the critical role of microglia in AD[58]; for the other three psychiatric disorders, both IBSEP and sc-brain found more colocalized loci in neurons (Figures S49 and S50).

As a concrete example, sc-brain and IBSEP identified 23 and 29 colocalized genes in AD, respectively, which exhibited a high degree of cell type specificity (Figure 6B). Among the colocalized genes identified by IBSEP, 18 genes were exclusively expressed in one cell type, while 7 and 4 genes were present in 2 and 3 cell types, respectively. Microglia play a crucial role in clearing amyloid-beta (Aβ) plaques, abnormal aggregation of which is a major pathological feature of AD. IBSEP identified several genes that are colocalized exclusively in microglia. For instance, *BIN1* (MIM: 601248) has been found to be a significant risk locus for AD,[59] with its expression levels influenced by a microglia-specific enhancer.[60] Another microglia-specific colocalized gene shared by IBSEP and sc-brain was *INPP5D* (MIM: 601582), whose expression levels positively correlate with amyloid plaque density and increase predominantly in plaque-associated microglia with the progression of AD.[61] In addition to microglia, AD is also associated with other cell types. For instance, astrocytes play crucial roles in neuronal

Please cite this article in press as: Yu et al., A unified framework for cell-type-specific eQTL prioritization by integrating bulk and scRNA-seq data, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2024.12.018

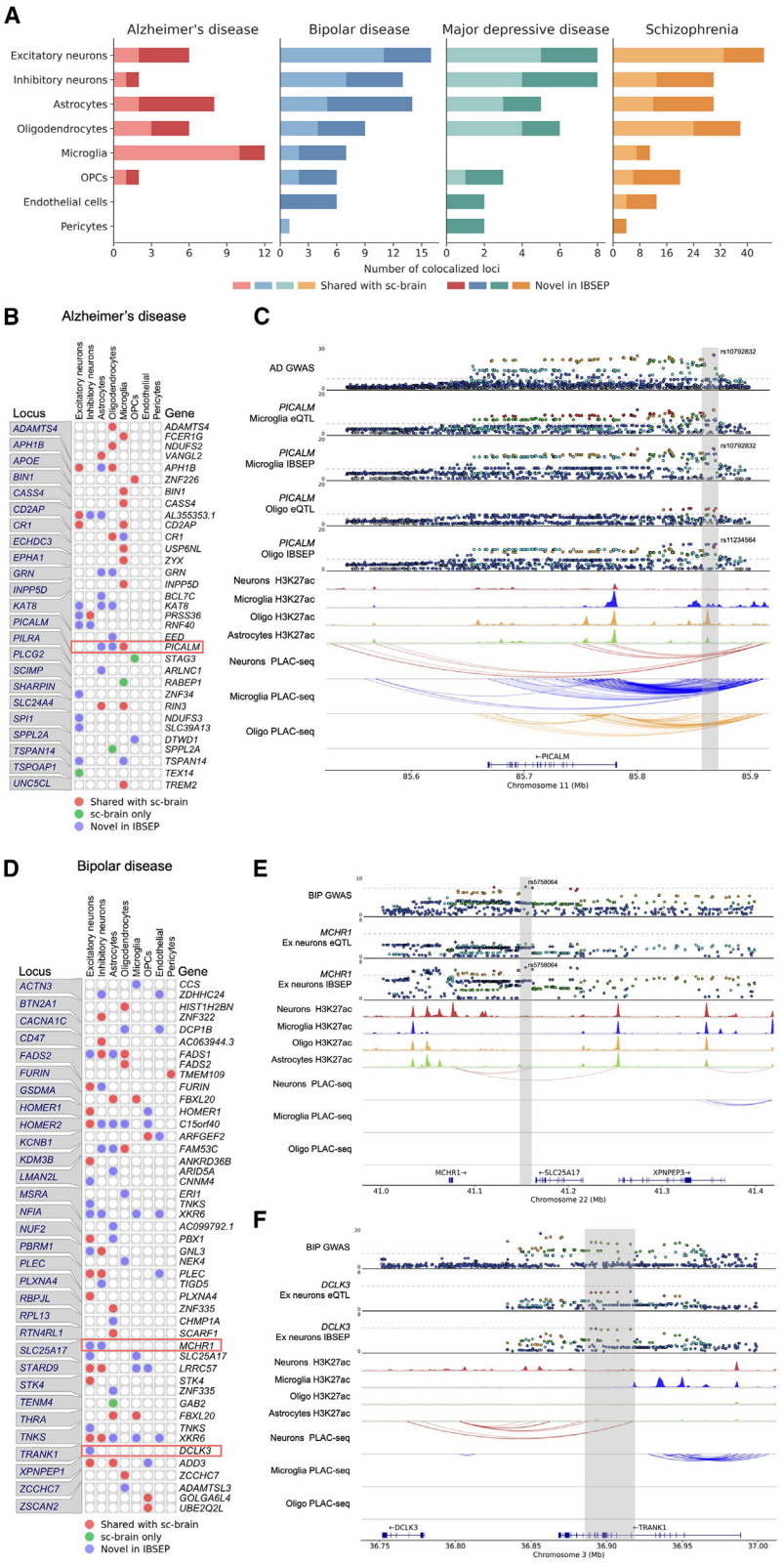

Figure 6. Results of colocalization between brain ct-eQTLs and genetic risk SNPs of brain disorders

(A) Comparison of the number of colocalized genes identified by sc-brain and IBSEP across four brain disorders and eight brain cell types.

(B) Cell-type-level colocalized genes of Alzheimer disease found by sc-brain and IBSEP.

(C) Locuszoom plots of Alzheimer disease GWAS, sc-brain, and IBSEP around *PICALM* and corresponding epigenomic tracks.

(D) Cell-type-level colocalized genes of bipolar disease found by sc-brain and IBSEP.

(E) Locuszoom plots of bipolar disease GWAS, sc-brain, and IBSEP around *MCHR1* and corresponding epigenomic tracks.

(F) Locuszoom plots of bipolar disease GWAS, sc-brain, and IBSEP around *DCLK3* and corresponding epigenomic tracks.

In the Locuszoom plots of GWAS and eQTL, the dashed lines represent −log *p* values of 8 and 5, respectively.

metabolism and clearance of Aβ, and the development of astrocytes with pathological phenotypes may contribute to the progression of AD. Both IBSEP and sc-brain have colocalized *RIN3* (MIM: 610223) in astrocytes and microglia, which has been reported as a potential contributor to the development of AD.[62] Previous studies have found that *RIN3* plays a significant role in the early endocytic pathway, which is closely related to microglial function.[63] Although the specific interactions and functions between *RIN3* and astrocytes have not been fully elucidated, ct-eQTLs detected by ISBEP may shed light on the genetic etiology of AD associated with *RIN3*. Another frequently discussed gene, *PICALM* (MIM: 603025), has been reported to be highly expressed in microglia in late-onset AD.[64] Both sc-brain and IBSEP identified *PICALM* colocalized in microglia. Additionally, IBSEP found *PICALM* colocalized in oligodendrocytes and astrocytes as well. In Figure 6C, significant GWAS and eQTL signals were observed ∼80 kb upstream of *PICALM*, with clear peaks in H3K27ac ChIP-seq[65] in microglia as well as oligodendrocytes and astrocytes. Proximity ligation-assisted ChIP-seq (PLAC-seq)[60] in both microglia and oligodendrocytes shows interaction with the gene's starting position in this region, suggesting that it may be an active enhancer for *PICALM* in multiple cell types rather than microglia specific. On the contrary, there are no significant epigenomic signals in neurons in this region, consistent with neither IBSEP nor sc-brain not colocalizing *PICALM* in neurons.

For BIP, sc-brain identified 27 colocalized genes, while IBSEP increased this number to 41. The pathogenesis of BIP is complex, involving multiple cell types, particularly neurons and glial cells, with the former's abnormal activity potentially leading to mood fluctuations and the latter

playing a crucial role in maintaining neuronal function and regulating neurotransmitter levels.[66–68] Among the 41 colocalized genes identified by IBSEP, 27 were observed in excitatory or inhibitory neurons, and 19 were associated with astrocytes or oligodendrocytes, highlighting the importance of these cell types in the pathogenesis of BIP (Figure 6D). Mullins et al. conducted mediation analysis using the same GWAS summary statistics with the tissue-level brain eQTLs from PsychENCODE Consortium,[69] identifying several genes potentially mediating BIP.[55] Some of these genes were also colocalized by both IBSEP and sc-brain, including *GNL3* (MIM: 608011), *ADD3* (MIM: 601568), *FURIN* (MIM: 136950), *LRRC57*, and *STK4* (MIM: 604965). Most importantly, IBSEP can identify the cell types in which these genes act as potential mediators, enabling us to uncover biological pathogenic mechanisms at a fine-grained scale. For instance, *FURIN* is reported to be associated with multiple psychiatric disorders.[57,70,71] The most significant SNP of *FURIN*, rs4702, was identified by a BIP GWAS as well as IBSEP in neurons. In addition, IBSEP further colocalized *MCHR1* (MIM: 601751) and *DCLK3* (MIM: 613167), which are also on the list of genes associated with BIP[55] but not discovered by sc-brain. The *MCHR1* gene, possibly involved in mood regulation and circadian rhythm disruption, is associated with anxiety-like behaviors.[72,73] IBSEP identified it as a co-localized gene in both excitatory neurons and inhibitory neurons. As shown in Figure 6E, the most significant SNP, rs5758064, of IBSEP in excitatory neurons is also the lead SNP for this locus in the BIP GWAS. Epigenomic signals indicate that rs5758064 may correspond to a neuron-specific enhancer interacting with *MCHR1*. *DCLK3* plays a critical role in the development and maintenance of the nervous system, which encodes a kinase regulating multiple signaling pathways within cells.[74] In the region of 36.88–36.92 Mb of chromosome 3, both excitatory neuron eQTLs and the BIP GWAS exhibited significant signals (Figure 6F). The PLAC-seq shows an interaction between *DCLK3* and this region, suggesting it may correspond to a neuron-specific enhancer for *DCLK3*.

Discussion

ct-eQTL mapping is critical for understanding the genetic regulation of gene expression in specific cellular contexts, which in turn informs the mechanisms underlying complex diseases. This paper introduces a framework, IBSEP, designed to enhance the detection and prioritization of ct-eQTLs by integrating bulk RNA-seq and scRNA-seq data. By leveraging the strengths of both data types, IBSEP overcomes the limitations of each, resulting in improved accuracy and detection power for ct-eQTLs. Through comprehensive simulations, we showed that our method well controls the type I error rate while having greater statistical power compared to existing approaches and that it is robust across various genetic architecture settings. Our analyses of blood and brain cortex datasets demonstrate that IBSEP greatly enhances the power of ct-eQTL prioritization, uncovering transcriptional regulatory mechanisms specific to different cell types. These findings provide valuable insights into the genetic architecture of complex traits at a cellular resolution.

Genetic variations can influence RNA levels and regulate various molecular traits, such as histone modifications, chromatin accessibility, allele-specific expression, alternative splicing, DNA methylation, and protein expression.[75] QTL analyses based on histone modifications and chromatin accessibility can reveal variant loci that affect transcription factor binding and nucleosome positioning. Detection of protein QTLs (pQTLs) can identify variant loci impacting transcriptional and post-transcriptional mechanisms. However, these molecular QTL (xQTL) studies face similar challenges to eQTL studies, with most remaining at the tissue level due to the high costs and technical constraints limiting cell-type-specific QTL research. For instance, the recently released UKBB Pharma Proteomics Project includes 2,923 plasma protein traits from 54,219 samples.[76,77] In contrast, a recent cell-type-specific pQTL study had only 303 samples,[78] resulting in many potential ct-pQTLs being undetected due to insufficient statistical power. As a general framework for integrating tissue-level and cell-type-level QTL data, the principles and methodologies of IBSEP can be extended to these molecular traits straightforwardly. By incorporating these additional layers of molecular data, researchers can develop a more comprehensive understanding of how genetic variants influence a wide array of biological processes. This expansion is crucial for capturing the full scope of gene regulation and its effects on phenotype and disease. The integration of diverse molecular traits with ct-eQTL analysis will provide a more detailed and nuanced view of genetic regulation.

Due to the genetic heterogeneity among different populations, QTLs identified in studies of genetic mechanisms of specific molecular traits can vary across populations.[79,80] Unfortunately, most current eQTL samples come from EUR populations. As many have noted, genomic discoveries in EURs cannot be directly translated to non-EUR individuals,[81] resulting in severe underrepresentation of non-EUR populations in eQTL research. This exacerbates inequalities in genetic research and healthcare for non-EUR populations. In fact, population-specific genetic variants and environmental contexts can significantly affect the manifestation of eQTLs. Studying diverse populations can reveal the different transcriptional regulatory patterns and different genetic etiology for diseases.[82] Therefore, in future research, we plan to extend the integrated analysis method applied in IBSEP to diverse ethnic groups. We aim to improve the statistical power of ct-eQTL localization in non-EUR populations and identify both shared and population specific ct-eQTLs, thereby improving the generalizability and applicability of findings and ensuring that medical insights and treatments are relevant and effective for a broad range of individuals.

Following eQTL detection, statistical fine-mapping can identify candidate causal variants likely to drive variation in expression within specific cell types, improving the translation of findings of genetic studies.[83,84] Particularly, Mendelian randomization (MR) is widely used to investigate the causal relationship between exposure factors and disease outcomes.[85] In the context of ct-eQTLs, the expression level of a gene in a specific cell type is considered the exposure factor, while the complex disease is regarded as the outcome. The candidate causal variants inferred from ct-eQTL results could be used as instrumental variables, which typically come from a small region of the genome surrounding the target gene, known as *cis*-variants, and therefore, this type of analysis is referred to as *cis*-MR analysis. With such a causal inference framework, *cis*-MR can use candidate causal variants to identify drug targets that increase disease risk through alterations in gene expression, enabling pinpoint causal variants with greater accuracy, and to understand the biological mechanisms linking genetic variants to disease in distinct cellular resolution.

Although IBSEP provides an effective way to leverage the advantages of each type of data to improve the statistical power for ct-eQTL prioritization, we should be aware of some potential limitations. First, in the estimation of cell type proportions ($\boldsymbol{\pi}$) for bulk RNA-seq samples, typically count-scale gene expressions data were used. However, in the hierarchical linear model applied by IBSEP (Equation 2), we directly plugged in $\boldsymbol{\pi}$ into the quantile normalized gene expressions models. This disparity arises from the different purposes and contexts in which $\boldsymbol{\pi}$ is utilized. The initial estimation of cell type proportions using count data allows for a robust and unbiased assessment of the relative abundance of each cell type. The purpose of incorporating these pre-estimated cell type proportions in IBSEP is merely to ensure that the model captures the relative differences of cell type compositions in tissue-level data more effectively. Besides, we note that even in the presence of substantial cell type proportion estimation errors, IBSEP still can well control the type I error control rate (Figure S12). Second, the current implementation of IBSEP requires that both RNA-seq and scRNA-seq data originate from the same population. This is not a significant limitation at present, as most eQTL data have been derived from EUR populations. Nevertheless, as genetic studies increasingly focus on non-EUR populations, it is essential to account for population heterogeneities, such as LD mismatch, before extending IBSEP to eQTL data from multiple ancestries. Third, IBSEP can leverage the widespread genetic correlations among cell types, especially biologically close cell types, for certain genes to further boost statistical power (Figure 5E). However, the estimation of coheritability using LDSC regression[22] may be unstable due to the limited number of *cis*-SNPs in a given gene. Therefore, we adopt a conservative approach by reducing the coheritability estimates to zero if they do not significantly deviate from zero. Fourth, IBSEP is designed to identify ct-eQTLs across discrete cell types, such that the tissue-level gene expressions from bulk RNA-seq can be decomposed into the weighted average of gene expression at the cell type level. Recently, a few single-cell eQTL studies have begun to focus on continuous phenotypes or intermediate states instead of discrete lineages, aiming to map "dynamic" eQTLs that exhibit dynamical effects along a continuous axis. This strategy have been successfully applied to continuous trajectories within differentiating induced pluripotent stem cells (iPSCs)[86] and T cells.[87] How to improve the IBSEP model to accommodate these granular, single-cell resolution data is an interesting direction for future work.

Data and code availability

The cell-type-level and tissue-level eQTL summary statistics, GTEx bulk tissue expressions, and GWAS summary statistics were obtained from the links summarized in Table S1. The IBSEP software and scripts for reproducing the analyses are available at https://github.com/xinyiyu/IBSEP.

Acknowledgments

This work was supported in part by NSFC (grant no. 12401384), the Shenzhen Research Institute of Big Data Internal Project (J00220230008), and the Shenzhen Science and Technology Program (grant no. RCBS20231211090613024). This research has been conducted using the UKBB Resource under application no. 96744.

Declaration of interests

The authors declare no competing interests.

Web Resources

1M, https://eqtlgen.org/sc/datasets/1m-scbloodnl.html
1000 Genomes, https://www.internationalgenome.org/
CIBERSORTx, https://cibersortx.stanford.edu/
Coloc, https://github.com/chr1swallace/coloc
GTEx, https://gtexportal.org/
IBSEP, https://github.com/xinyiyu/IBSEP
LDlinkR, https://github.com/CBIIT/LDlinkR
LDSC, https://github.com/bulik/ldsc
OMIM, https://www.omim.org
OneK1K, https://onek1k.org
PLINK, https://www.cog-genomics.org/plink
UKBB, https://www.ukbiobank.ac.uk

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2024.12.018.

Received: August 1, 2024
Accepted: December 18, 2024
Published: January 16, 2025

Please cite this article in press as: Yu et al., A unified framework for cell-type-specific eQTL prioritization by integrating bulk and scRNA-seq data, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2024.12.018

References

1. Elliot, S., Abayomi, M., Ala, A., Annalisa, B., Maria, C., Laurent, G., Tudor, G., Osman, G., Hall, P., Hayhurst, J., et al. (2023). The nhgri-ebi gwas catalog: knowledgebase and deposition resource. Nucleic. acids. research *51*, D977–D985.
2. Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. *16*, 197–212.
3. Kindt Alida, S.D., Pau, N., Semple Colin, A.M., and Haley Chris, S. (2013). The genomic signature of trait-associated variants. BMC Genom. *14*, 1–15.
4. Consortium, G.T.E. (2020). The gtex consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.
5. Urmo, V., Annique, C., Harm-Jan, W., Marc Jan, B., Patrick, D., Biao, Z., Holger, K., Ashis, S., Roman, K., Seyhan, Y., et al. (2021). Large-scale cis-and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310.
6. Cuomo, A.S.E., Nathan, A., Raychaudhuri, S., MacArthur, D.G., and Powell, J.E. (2023). Single-cell genomics meets human genetics. Nat. Rev. Genet. *24*, 535–549.
7. Connally, N.J., Nazeen, S., Lee, D., Shi, H., Stamatoyannopoulos, J., Chun, S., Cotsapas, C., Cassa, C.A., and Sunyaev, S.R. (2022). The missing link between genetic association and regulatory function. Elife *11*, e74970.
8. Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O., and Knight, J.C. (2012). Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of hla alleles. Nat. Genet. *44*, 502–510.
9. Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., Peter, Van't H., 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.
10. Raúl, A.-G., de, K.N., di, T.J., Annique, C., van der, W.M.G.P., Dylan de, V., Harm, B., Roy, O., Urmo, V., Zorro, M.M., et al. (2020). Deconvolution of bulk blood eqtl effects into immune cell subpopulations. BMC Bioinf. *21*, 1–23.
11. Paul, L., Si, L., Vasyl, Z., Yun, L., Lin, D.-Y., and Sun, W. (2023). A computational method for cell type-specific expression quantitative trait loci mapping using bulk rna-seq data. Nat. Commun. *14*, 3030.
12. van der Wijst, M.G.P., Brugge, H., de Vries, D.H., Deelen, P., Swertz, M.A., LifeLines Cohort Study; and BIOS Consortium, and Franke, L. (2018). Single-cell rna sequencing identifies celltype-specific cis-eqtls and co-expression qtls. Nat. Genet. *50*, 493–497.
13. Julien, B., Daniela, C., Will, M., Lynette, F., Eduard, U., Ward, O., Alejandro, I.V., Suresh, S., Erik, N., Manuel, 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.
14. Zhang, J., and Hongyu, Z. (2023). eqtl studies: from bulk tissues to single cells. Journal of Genetics and Genomics *50*, 925–933.
15. Strober, B.J., Tayeb, K., Popp, J., Qi, G., Gordon, M.G., Perez, R., Ye, C.J., and Battle, A. (2024). Surge: uncovering context-specific genetic-regulation of gene expression from single-cell rna sequencing using latent-factor models. Genome Biol. *25*, 28.
16. Cuomo, A.S.E., Heinen, T., Vagiaki, D., Horta, D., Marioni, J.C., and Stegle, O. (2022). Cellregmap: a statistical framework for mapping context-specific regulatory variants using scrna-seq. Mol. Syst. Biol. *18*, e10663.
17. Seyhan, Y., Jose, A.-H., Kristof, W., Anne, S., Grace, G.M., Stacey, A., Lu, Q., Antonia, R., Taylor Thomas, R.P., Clarke, L., et al. (2022). Single-cell eqtl mapping identifies cell type–specific genetic control of autoimmune disease. Science *376*, eabf3041.
18. Roy, O., de, V.D.H., Harm, B., Grace, G.M., Martijn, V., Ye, C.J., Harm-Jan, W., Lude, F., et al.; consortium single-cell eQTLGen; and BIOS Consortium (2022). Single-cell rna-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. Nat. Commun. *13*, 3267.
19. Vivek, N., Fairfax, B.P., Seiko, M., Peter, H., Wong, D., Ng, E., Hill, A.V.S., and Knight, J.C. (2015). Genomic modulators of gene expression in human neutrophils. Nat. Commun. *6*, 7545.
20. Cai, M., Xiao, J., Zhang, S., Wan, X., Zhao, H., Chen, G., and Yang, C. (2021). A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. Am. J. Hum. Genet. *108*, 632–655.
21. Xiao, J., Cai, M., Yu, X., Hu, X., Chen, G., Wan, X., and Yang, C. (2022). Leveraging the local genetic structure for trans-ancestry association mapping. Am. J. Hum. Genet. *109*, 1317–1337.
22. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
23. Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. Econometrica *50*, 1029–1054.
24. Steen Chloé, B., Long, L.C., and Alizadeh Ash, A. (2020). Newman Aaron M. Profiling cell type abundance and expression in bulk tissues with cibersortx. In Stem Cell Transcriptional Networks: Methods and Protocols, pp. 135–157.
25. Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, 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.e24.
26. Kasela, S., François, A., Kim-Hellmuth, S., Brown, B.C., Nachun, D.C., Tracy, R.P., Durda, P., Liu, Y., Taylor, K.D., Johnson, W.C., et al. (2024). Interaction molecular qtl mapping discovers cellular and environmental modifiers of genetic regulatory effects. Am. J. Hum. Genet. *111*, 133–149.
27. Jorstad, N.L., Close, J., Johansen, N., Yanny, A.M., Barkan, E.R., Travaglini, K.J., Bertagnolli, D., Campos, J., Casper, T., Crichton, K., et al. (2023). Transcriptomic cytoarchitecture reveals principles of human neocortex organization. Science *382*, eadf6812.
28. Myers, T.A., Chanock, S.J., and Machiela, M.J. (2020). Ldlinkr: an r package for rapidly calculating linkage disequilibrium statistics in diverse populations. Front. Genet. *11*, 513535.
29. Claudia, G., Damjan, V., Schadt, E.E., Lude, F., Hingorani Aroon, D., Wallace, C., and Vincent, P. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.
30. Laura, D., and Laura, D. (2005). Blood Groups and Red Cell Antigens, *2* (NCBI Bethesda).

31. K. Verhoeckx, P. Cotter, I. Lopez-Exposito, C. Kleiveland, T. Lea, A. Mackie, T. Requena, D. Swiatecka, and H. Wichers. The impact of food bioactives on health: in vitro and ex vivo models. 2015. Springer.
32. Dong, D., Zheng, L., Lin, J., Zhang, B., Zhu, Y., Li, N., Xie, S., Wang, Y., Gao, N., and Huang, Z. (2019). Structural basis of assembly of the human t cell receptor–cd3 complex. Nature *573*, 546–552.
33. Wu, W., Zhou, Q., Masubuchi, T., Shi, X., Li, H., Xu, X., Huang, M., Meng, L., He, X., Zhu, H., et al. (2020). Multiple signaling roles of cd3ε and its application in car-t cell therapy. Cell *182*, 855–871.e23.
34. Zhu, Z., Zhu, X., Cong-Lin, L., Shi, H., Shen, S., Yang, Y., Hasegawa, K., Camargo, C.A., and Liang, L. (2019). Shared genetics of asthma and mental health disorders: a large-scale genome-wide cross-trait analysis. Eur. Respir. J. *54*.
35. Radstake, T.R.D.J., Gorlova, O., Rueda, B., Martin, J.E., Alizadeh, B.Z., Palomino-Morales, R., Coenen, M.J., Vonk, M.C., Voskuyl, A.E., Schuerwegh, A.J., et al. (2010). Genome-wide association study of systemic sclerosis identifies cd247 as a new susceptibility locus. Nat. Genet. *42*, 426–429.
36. Elena, L.-I., Marialbert, A.-H., Martin, K., Shervin, A., Ansuman T, S., Jeffrey, G., Maxwell R, M., Lorenzo, B., Simeón Carmen, P., Patricia, C., et al. (2019). Gwas for systemic sclerosis identifies multiple risk loci and highlights fibrotic and vasculopathy pathways. Nat. Commun. *10*, 4955.
37. Ming-Huei, C., Laura M, R., Abdou, M., Saori, S., Huffman, J.E., Arden, M., Bhavi, T., Jiang, T., Akbari, P., Dragana, V., et al. (2020). Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell *182*, 1198–1213.
38. Dragana, V., Bao, E.L., Akbari, P., Lareau Caleb, A., Abdou, M., Jiang, T., Ming-Huei, C., Raffield, L.M., Manuel, T., Huffman, J.E., et al. (2020). The polygenic and monogenic basis of blood traits and diseases. Cell *182*, 1214–1231.
39. Kazuyoshi, I., Saori, S., Chikashi, T., Yang, L., Kyuto, S., Kensuke, Y., Tiffany, A., Lai, T.C., Laufer, V.A., Scott, I.C., et al. (2022). Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. Nat. Genet. *54*, 1640–1651.
40. International Multiple Sclerosis Genetics Consortium (2019). . Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science *365*, eaav7188.
41. De Lange, K.M., Moutsianas, L., Lee, J.C., Lamb, C.A., Luo, Y., Kennedy, N.A., Jostins, L., Rice, D.L., Gutierrez-Achury, J., Ji, S.G., et al. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat. Genet. *49*, 256–261.
42. Berend, H. (2001). The differential cell count. Lab. Hematol. *7*, 89–100.
43. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. Cell *167*, 1415–1429.e19.
44. ENCODE Project Consortium (2012). An integrated encyclopedia of dna elements in the human genome. Nature *489*, 57–74.
45. Nusse, R., Fuerer, C., Ching, W., Harnish, K., Logan, C., Zeng, A., Ten Berge, D., and Kalani, Y. (2008). Wnt signaling and stem cell control. In Cold Spring Harbor symposia on quantitative biology, *73* (Cold Spring Harbor Laboratory Press), pp. 59–66.
46. Mulligan, K.A., and Cheyette, B.N.R. (2012). Wnt signaling in vertebrate neural development and function. J. Neuroimmune Pharmacol. *7*, 774–787.
47. Gao, J., Liao, Y., Qiu, M., and Shen, W. (2021). Wnt/β-catenin signaling in neural stem cell homeostasis and neurological diseases. Neuroscientist *27*, 58–72.
48. Rivka, N., and Kris, V. (2017). How wnt signaling builds the brain: bridging development and disease. Neuroscientist *23*, 314–329.
49. Mats, N., Jansen, P.R., Sven, S., Kyoko, W., De Leeuw Christiaan, A., Julien, B., Savage, J.E., Hammerschlag Anke, R., Skene, N.G., Muñoz-Manchado Ana, B., et al. (2018). Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. Nat. Genet. *50*, 920–927.
50. Céline, B., Fahri, K., Jansen Iris, E., Luca, K., Sonia, M.-G., Amin, N., Naj Adam, C., Rafael, C.-M., Benjamin, G.-B., Andrade, V., et al. (2022). New insights into the genetic etiology of alzheimer's disease and related dementias. Nat. Genet. *54*, 412–436.
51. Pankratz, N., Beecham, G.W., DeStefano, A.L., Dawson, T.M., Doheny, K.F., Factor, S.A., Hamza, T.H., Hung, A.Y., Hyman, B.T., Ivinson, A.J., et al. (2012). Meta-analysis of parkinson's disease: identification of a novel locus, rit2. Ann. Neurol. *71*, 370–384.
52. Li, Q.S., and De Muynck, L. (2021). Differentially expressed genes in alzheimer's disease highlighting the roles of microglia genes including olr1 and astrocyte gene cdk2ap1. Brain Behav. Immun. Health *13*, 100227.
53. Aysu, O., Wu, Y., Wang, N., Hariharan, J., Michael, B., Moeen, N.S., Julia, S., Hyeokmoon, K., Grant, G., Tamara, G., et al. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. Nat. Genet. *54*, 437–449.
54. Jeremy, S., Cooper, S., Liu, J.Z., Barrio-Hernandez, I., Erica, B., Natsuhiko, K., Young, A.M.H., Franklin, R.J.M., Johnson, T., Karol, E., et al. (2021). Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new alzheimer's disease risk genes. Nat. Genet. *53*, 392–402.
55. Niamh, M., Forstner, A.J., O'Connell, K.S., Brandon, C., Coleman Jonathan, R.I., Qiao, Z., Als Thomas, D., Bigdeli, T.B., Sigrid, B., Julien, B., et al. (2021). Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat. Genet. *53*, 817–829.
56. Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A., Adams, M.J., Agerbo, E., Air, T.M., Andlauer, T.M.F., et al. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet. *50*, 668–681.
57. Trubetskoy, V., Pardiñas, A.F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T.B., Bryois, J., Chen, C.-Y., Dennison, C.A., Hall, L.S., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature *604*, 502–508.
58. Hansen, D.V., Hanson, J.E., and Sheng, M. (2018). Microglia in alzheimer's disease. JCB (J. Cell Biol.) *217*, 459–472.
59. Andrew, R.J., De Rossi, P., Nguyen, P., Kowalski, H.R., Recupero, A.J., Guerbette, T., Krause, S.V., Rice, R.C., Laury-Kleintop, L., Wagner, S.L., and Thinakaran, G. (2019). Reduction

of the expression of the late-onset alzheimer's disease (ad) risk-factor bin1 does not affect amyloid pathology in an ad mouse model. J. Biol. Chem. *294*, 4477–4487.

60. Alexi, N., Holtman Inge, R., Coufal Nicole, G., Schlachetzki Johannes, C.M., Miao, Y., Rong, H., Han, C.Z., Monique, P., Xiao, J., Yin, W., et al. (2019). Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. Science *366*, 1134–1139.
61. Tsai, A.P., Lin, P.B.-C., Dong, C., Moutinho, M., Casali, B.T., Liu, Y., Lamb, B.T., Landreth, G.E., Oblak, A.L., and Nho, K. (2021). Inpp5d expression is associated with risk for alzheimer's disease and induced by plaque-associated microglia. Neurobiol. Dis. *153*, 105303.
62. Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., Van Der Lee, S.J., Amlie-Wolf, A., et al. (2019). Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates aβ, tau, immunity and lipid processing. Nat. Genet. *51*, 414–430.
63. Hiroaki, K., Kota, S., Kyoko, T., Kenji, K., Yasuhiro, A., Hiroshi, K., and Toshiaki, K. (2003). Rin3: a novel rab5 gef interacting with amphiphysin ii involved in the early endocytic pathway. J. Cell Sci. *116*, 4159–4168.
64. Kunie, A., Jean-Pierre, B., Virginie, S., Valérie, S., Michèle, A., Robert, D., Anaïs, C., Pascale, L., Jérémie, L., Véronique, S., et al. (2013). Clathrin adaptor calm/picalm is associated with neurofibrillary tangles and is cleaved in alzheimer's brains. Acta Neuropathol. *125*, 861–878.
65. Ryan, C.M., Anna, S., Soumya, K., Michael J, G., Laure, F., Granja Jeffrey, M., Louie Bryan, H., Tiffany, E., Shadi, S., Tansu, B.S., et al. (2020). Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for alzheimer's and parkinson's diseases. Nat. Genet. *52*, 1158–1168.
66. Kato, T. (2008). Molecular neurobiology of bipolar disorder: a disease of 'mood-stabilizing neurons. Trends Neurosci. *31*, 495–503.
67. Jerome, M., Qiu-Wen, W., Yongsung, K., Yu, D.X., Son, P., Yang, B., Yi, Z., Diffenderfer Kenneth, E., Zhang, J., Sheila, S., et al. (2015). Differential responses to lithium in hyperexcitable neurons from patients with bipolar disorder. Nature *527*, 95–99.
68. Watkins, C.C., Sawa, A., and Pomper, M.G. (2014). Glia and immune cell signaling in bipolar disorder: insights from neuropharmacology and molecular imaging to clinical application. Transl. Psychiatry *4*, e350.
69. Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., et al. (2018). Transcriptome-wide isoform-level dysregulation in asd, schizophrenia, and bipolar disorder. Science *362*, eaat8127.
70. Levey, D.F., Stein, M.B., Wendt, F.R., Pathak, G.A., Zhou, H., Aslan, M., Quaden, R., Harrington, K.M., Nuñez, Y.Z., Overstreet, C., et al. (2021). Bi-ancestral depression gwas in the million veteran program and meta-analysis in¿ 1.2 million individuals highlight new therapeutic directions. Nat. Neurosci. *24*, 954–963.
71. Lee, P.H., Verneri, A., Hyejung, W., Yen-Chen A, F., Jacob, R., Zhu, Z., Tucker-Drob, E.M., Nivard, M.G., Grotzinger, A.D., Danielle, P., et al. (2019). Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. Cell *179*, 1469–1482.
72. Smith, D.G., Davis, R.J., Rorick-Kehn, L., Morin, M., Witkin, J.M., McKinzie, D.L., Nomikos, G.G., and Gehlert, D.R. (2006). Melanin-concentrating hormone-1 receptor modulates neuroendocrine, behavioral, and corticolimbic neurochemical stress responses in mice. Neuropsychopharmacology *31*, 1135–1145.
73. Madhuri, R., Nadia, D., Madelyn, C., and Marco, G. (2007). A study of the involvement of melanin-concentrating hormone receptor 1 (mchr1) in murine models of depression. Biol. Psychiatr. *61*, 174–180.
74. Laurie, G., Laetitia, F., Marie-Claude, G., Lucie de, L., Maria-Angeles Carrillo-de, S., Geraldine, L., Karine, C., Lev, S., Sophie, L., Julien, F., et al. (2018). The striatal kinase dclk3 produces neuroprotection against mutant huntingtin. Brain *141*, 1434–1454.
75. eGTEx Project (2017). Enhancing gtex by bridging the gaps between genotype, gene expression, and disease. Nat. Genet. *49*, 1664–1670.
76. Dhindsa, R.S., Burren, O.S., Sun, B.B., Prins, B.P., Matelska, D., Wheeler, E., Mitchell, J., Oerton, E., Hristova, V.A., Smith, K.R., et al. (2023). Rare variant associations with plasma protein levels in the uk biobank. Nature *622*, 339–347.
77. Sun, B.B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.H., Richardson, T.G., Surendran, P., Mahajan, A., Robins, C., Vasquez-Grinnell, S.G., et al. (2023). Plasma proteomic associations with genetics and health in the uk biobank. Nature *622*, 329–338.
78. Christian, L., Pascal, P., Juliana, I.-K., Jonas, C.-A., Maija-Leena, E., Ann-Christine, S., Gunnel, N., Sandling, J.K., Ingrid, K., Olsson, T., et al. (2020). Function of multiple sclerosis-protective hla class i alleles revealed by genome-wide protein-quantitative trait loci mapping of interferon signalling. PLoS Genet. *16*, e1009199.
79. Mogil, L.S., Andaleon, A., Badalamenti, A., Dickinson, S.P., Guo, X., Rotter, J.I., Johnson, W.C., Im, H.K., Liu, Y., and Wheeler, H.E. (2018). Genetic architecture of gene expression traits across diverse populations. PLoS Genet. *14*, e1007586.
80. Linda, K., Mak, A.C.Y., Donglei, H., Celeste, E., Scott, H., Elhawary Jennifer, R., Gupta, N., Gabriel, S., Xiao, S., Keys, K.L., et al. (2023). Gene expression in african americans, puerto ricans and mexican americans reveals ancestry-specific patterns of genetic architecture. Nat. Genet. *55*, 952–963.
81. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.
82. Randolph, H.E., Fiege, J.K., Thielen, B.K., Mickelson, C.K., Shiratori, M., João, B.-B., Langlois, R.A., and Barreiro, L.B. (2021). Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. Science *374*, 1127–1133.
83. Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet. *19*, 491–504.
84. Mingxuan, C., Wang, Z., Xiao, J., Hu, X., Gang, C., and Xmap, Y.C. (2023). Cross-population fine-mapping by leveraging genetic diversity and accounting for confounding bias. Nat. Commun. *14*, 6870.
85. Xianghong, H., Mingxuan, C., Jiashun, X., Wan, X., Zhiwei, W., Hongyu, Z., and Can, Y. (2024). Benchmarking mendelian randomization methods for causal inference using genome-wide association study summary statistics. Am. J. Hum. Genet. *111*, 1717–1735.

86. 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). Single-cell rna-sequencing of differentiating ips cells reveals dynamic genetic effects on gene expression. Nat. Commun. *11*, 810.

87. Aparna, N., Samira, A., Kazuyoshi, I., Cristian, V., Tiffany, A., Yang, L., Jessica I, B., Yuriy, B., Sara, S., Price Alkes, L., et al. (2022). Single-cell eqtl models reveal dynamic t cell state dependence of disease loci. Nature *606*, 120–128.