Efficient count-based models improve power and robustness for large-scale single-cell eQTL mapping

Zixuan Eleanor Zhang[1,**], Artem Kim[1], Noah Suboc[1], Nicholas Mancuso[1,2,3,*,**], Steven Gazal[1,2,3,*,**]

1. Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California
2. Department of Quantitative and Computational Biology, University of Southern California
3. Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California

* contributed equally
** corresponding authors

Abstract

Population-scale single-cell transcriptomic technologies (scRNA-seq) enable characterizing variant effects on gene regulation at the cellular level (e.g., single-cell eQTLs; sc-eQTLs). However, existing sc-eQTL mapping approaches are either not designed for analyzing sparse counts in scRNA-seq data or can become intractable in extremely large datasets. Here, we propose jaxQTL, a flexible and efficient sc-eQTL mapping framework using highly efficient count-based models given pseudobulk data. Using extensive simulations, we demonstrated that jaxQTL with a negative binomial model outperformed other models in identifying sc-eQTLs, while maintaining a calibrated type I error. We applied jaxQTL across 14 cell types of OneK1K scRNA-seq data (N=982), and identified 11-16% more eGenes compared with existing approaches, primarily driven by jaxQTL ability to identify lowly expressed eGenes. We observed that fine-mapped sc-eQTLs were further from transcription starting site (TSS) than fine-mapped eQTLs identified in all cells (bulk-eQTLs; P=1x10^{-4}) and more enriched in cell-type-specific enhancers (P=3x10^{-10}), suggesting that sc-eQTLs improve our ability to identify distal eQTLs that are missed in bulk tissues. Overall, the genetic effect of fine-mapped sc-eQTLs were largely shared across cell types, with cell-type-specificity increasing with distance to TSS. Lastly, we observed that sc-eQTLs explain more SNP-heritability (h^2) than bulk-eQTLs (9.90 ± 0.88% vs. 6.10 ± 0.76% when meta-analyzed across 16 blood and immune-related traits), improving but not closing the missing link between GWAS and eQTLs. As an example, we highlight that sc-eQTLs in T cells (unlike bulk-eQTLs) can successfully nominate *IL6ST* as a candidate gene for rheumatoid arthritis. Overall, jaxQTL provides an efficient and powerful approach using count-based models to identify missing disease-associated eQTLs.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Introduction

Large gene expression quantitative trait loci (eQTLs) studies have facilitated interpreting genetic variants identified in genome-wide association studies (GWAS) through colocalization [1–4] or transcriptome-wide association studies (TWAS) [5–9]. These approaches have been largely dependent on eQTLs discovered from bulk-RNA sequencing (bulk-eQTLs) on tissue samples [10,11] or on a limited number of cell types [12–14]. However, limited overlap between bulk-eQTLs and GWAS risk loci [4,10,15–18] has hindered the functional interpretation of genetic risk variants and their translation to therapeutic development for human diseases. Multiple hypotheses could explain this “missing link” between GWAS and eQTLs, including the lack of disease-relevant cell types/contexts, cell-type-specific eQTL effects diluted in bulk samples, and limited statistical power to detect weak-effect eQTLs [15,16].

Recent and ongoing generation of large scale single-cell RNA sequencing (scRNA-seq) datasets allow direct interrogation of these hypotheses by quantifying gene expression across heterogeneous cell types for a large number of individuals [19,20]. For example, the OneK1K project has released scRNA-seq data from 1.27 million peripheral blood mononuclear cells (PMBCs) of 982 donors [19], with plans to profile 50 million cells in 10,000 donors (TenK10K) [21]. A current challenge is thus to efficiently identify single-cell (sc-)eQTLs from sparse counts data in these extremely large datasets. Previous sc-eQTL studies [22–26] have leveraged pseudobulk data and used tools that are designed for bulk-eQTL mapping (e.g., Matrix eQTL [27], FastQTL [28], and tensorQTL [29]). These tools fit linear models after data normalization on the gene expression matrix [30,31]. Although the model fitting step is computationally efficient, the eQTL effect on gene expression is less interpretable due to the data transformation (e.g., inverse rank transform). Moreover, for sparse read counts observed in scRNA-seq, transformations are less effective due to sheer number of zeros [32,33]. Recent studies have proposed modelling the expression of single cells by fitting mixed effect models, either using off-the-shelf R functions [34] or under bespoke software such as CellRegMap [35] and SAIGE-QTL [36]. While these approaches improve upon bulk-eQTL mapping approaches, they can become computationally intractable for extremely large single-cell datasets currently being generated at a population scale. In addition to these computational challenges, the characterization of sc-eQTLs across cell types is further complicated by the differential statistical power induced by differences in cell abundances. For example, recent work reported sc-eQTLs were largely cell-type-specific [19], in contrast to higher levels of sharing across cell types when eQTLs were identified from sorted RNA-seq data [37]. Therefore, sc-eQTL mapping and characterization stands to benefit from scalable and statistically powerful software.

To address these limitations, we propose jaxQTL, an efficient software to perform large-scale sc-eQTL mapping using flexible, count-based models. Under simulations, we found that a negative binomial (negbinom) model outperforms linear and Poisson models in identifying sc-eQTLs while maintaining calibrated type I errors. By analyzing OneK1K, we found that jaxQTL with a negative binomial model identifies more eGenes than other models and existing softwares, such as tensorQTL and SAIGE-QTL. Importantly, we found that sc-eQTLs effects were largely consistent across cell types, with cell-type-specificity increasing with distance to transcription start site. Finally, we found that sc-eQTLs explained a greater fraction of heritability for GWAS immune traits compared with bulk-eQTLs, thus improving but not closing the missing link between GWAS and eQTLs. Taken together, our results demonstrate that jaxQTL is a scalable tool in identifying sc-eQTLs by analyzing large single-cell datasets to improve the biological interpretation of genetic risk at disease-relevant cell types.

Results

Overview of jaxQTL

We provide a brief overview of jaxQTL model assumptions and inferential pipeline. Given pseudobulk counts y_c for a focal gene in a cellular context c (i.e. summed across all cells of type c), covariates X (e.g., age, sex, genotyping principal components), and a cis-genetic variant g, jaxQTL implements a generalized linear model (GLM) according to,

$$E[y_c \mid X] = h(X\beta + g\beta_g + l_c),$$

where β are the covariate effects, β_g is the allelic effect, l_c is an offset adjusting for differences in library size, and $h(\cdot)$ is a function which maps linear predictions to expected values matching distribution assumptions (e.g., negbinom, Poisson). For example, if we assume a Poisson or negative binomial distribution, then $h := exp(\cdot)$, with effect sizes reflecting a change in the transcription rate (or proportion if including library size offsets). This is in contrast to linear regression performed on the rank-inverse normal transformed counts, where effect sizes have no direct interpretation of the expression values, but rather reflect a change in rankings.

Performing cis-association scans using this approach is computationally prohibitive, due to the sheer number tests required for each gene and cell-type. To address this fundamental limitation, jaxQTL leverages three key insights. First, jaxQTL performs *just-in-time* (JIT) compilation provided by the JAX framework (**Web Resources**), to translate high-level Python into machine-level instructions optimized for a specific parallelized architectures (e.g., CPU, GPU, or TPU) with no additional work required from the user other than a runtime flag. Second, jaxQTL performs a score-test[38] for all cis-genetic variants simultaneously using an optimized block-matrix approach, rather than sequentially. Lastly, jaxQTL implements multiple recent advances to compute p-values efficiently, which provide trade-offs between additional scalability and statistical power (e.g., Beta-approximation [28] to permutations or ACAT-V [39]; **Figures S1, S2; Methods**). We provide a table summarizing the capabilities of jaxQTL alongside other softwares (**Table S1**) and have released jaxQTL as open-source software (see **Code availability**).

Negative binomial outperforms other models in identifying sc-eQTLs in realistic simulations

We assessed the type I error and power of different models implemented in jaxQTL (jaxQTL-linear, jaxQTL-negbinom, and jaxQTL-Poisson) and softwares (SAIGE-QTL and tensorQTL) by simulating single-cell read counts from a Poisson mixed effect (PME) generative model using parameters that reflect observed expression and overdispersion in OneK1K [30,40–46] (**Figure S3, S4**). We evaluated model performance across varying cell type proportions by sampling individual library sizes across three cell types representing high, medium, and low library sizes (CD4+ naïve and central memory T (CD4$_{NC}$) cells, immature and naïve B (B$_{IN}$) cells, and Plasma cells, respectively; **Table S2**). We varied sample-coverage (i.e., the percentage of non-zero expression read counts) across simulations to account for gene expression intensity across individuals.

First, all models exhibited calibrated type I error rates when simulating from the single-cell PME model (**Figure 1A**), except for pseudobulk jaxQTL-Poisson which displayed increased false positives likely due to its over-conservative standard errors. We observed largely similar conclusions for jaxQTL when varying heritability, random intercept variance $\sigma^2{}_u$ (modeling similarity of cell read counts within the same person; see **Methods**), sample size, and minor allele frequency (MAF) parameters (**Figures S5-S8**). Importantly, jaxQTL-negbinom and linear models remain calibrated across cell type abundances and sample sizes, unlike SAIGE-QTL which exhibited increased false positives in rarer cell types when sample sizes are small (N < 200; **Figure S7**). SAIGE-QTL and jaxQTL-negbinom had slight inflation when $\sigma^2{}_u \approx 1$, however this scenario is unlikely to occur in practice (**Figure S3**).

Next, we observed that jaxQTL-negbinom had improved power compared with jaxQTL-linear and tensorQTL, especially for lower coverage genes. Across three cell types, genes with higher coverage exhibited greater statistical power to identify their sc-eQTLs. Specifically, for large cell type proportions, jaxQTL-negbinom outperformed jaxQTL-linear for genes with >95% coverage (P = 7.53 x 10^{-4}). For medium and rare cell types, jaxQTL-negbinom exhibited greater power over jaxQTL-linear down to ~70% coverage (P = 2.02 x 10^{-4} and 3.91 x 10^{-27} respectively), highlighting the benefit of count-based models for lower expressed genes and rarer cell types. While both jaxQTL-linear and tensorQTL fit linear models of gene expression, differences in power can be explained by jaxQTL score test versus tensorQTL Wald test. We obtained similar conclusions when varying heritability, random intercept variance $\sigma^2{}_u$, sample size, and MAF parameters (**Figures S5-S8**).

After assessing model performances under the PME model with non-zero random intercept variance $\sigma^2{}_u$, we repeated our analyses by sampling read counts from PME models with $\sigma^2{}_u$= 0, which reflects a standard Poisson model (**Figure S9**). As expected, the performance of jaxQTL-negbinom and SAIGE-QTL closely resembled jaxQTL-Poisson (see **Supplemental Note**). Again, count-based models outperformed jaxQTL-linear and tensorQTL, notably for genes in rarer cell types. All models were well-calibrated under the null.

Altogether, jaxQTL-negbinom provides a calibrated and powerful pseudobulk model for single-cell data that performs comparably to the PME model of SAIGE-QTL. Our empirical results can be in part explained by the structural similarity of the variance under negative binomial and PME models of pseudobulk (see **Supplemental Note**).

jaxQTL improves power for eGene discovery in the OneK1K dataset

To benchmark jaxQTL in identifying eGenes on real datasets, we applied jaxQTL on single-cell data of 14 PBMC cell types from N = 982 individuals in OneK1K [19]. We defined eGene as genes with at least one sc-eQTL in a cell type, i.e., gene-cell-type pairs. Before comparing different sc-eQTL models, we investigated the calibration of gene-level P values obtained by the Beta-approximation approach that permutes the gene expressions observed in OneK1K data. Across the three representative cell types, we found that gene-level P values from jaxQTL-linear and jaxQTL-negbinom were well-calibrated (**Figure S10**), however gene-level P values for jaxQTL-Poisson were inflated due to overcorrection by the permutation method on its variant P values.

After confirming the gene-level P values were calibrated, we compared the statistical power in identifying eGenes across different models using jaxQTL (**Figure 2A; Table S3**). Across 14 cell types, jaxQTL-negbinom identified 14% more eGenes compared with jaxQTL-linear (18,907 vs. 16,654 eGenes, $P = 1 \times 10^{-35}$), and 21% more compared with jaxQTL-Poisson (15,634 eGenes, $P = 5 \times 10^{-75}$). The number of eGenes found per cell type was highly correlated with cell type proportions, which reflects differential statistical power (Pearson $\rho = 0.97$; **Figure S11**). Focusing on jaxQTL-negbinom and jaxQTL-linear, we found the negbinom model provided higher χ^2 test statistics for lead SNP-eGene pairs across cell types (median $\chi^2 = 43.60$ vs. 38.46, $P = 3 \times 10^{-24}$; **Figure S12**). eGenes identified between models show substantial overlap (**Figure S13A**). Consistent with simulation results, eGenes identified exclusively by jaxQTL-negbinom had lower coverage (median 79% vs. 94%, $P = 2 \times 10^{-115}$; **Figure S13B**) than eGenes also identified with jaxQTL-linear, confirming that the negbinom model is more powerful for genes with lower expression. The reduced power of the jaxQTL-Poisson was caused by the penalty on its inflated type I error when computing the gene-level *P* values using the permutation approach. Given the improvement of the negbinom model over Poisson and linear models, to simplify our presentation we refer to jaxQTL-negbinom as jaxQTL for the remainder of the manuscript.

We next compared jaxQTL performance against tensorQTL [29] (a commonly used software for bulk-eQTL mapping using a linear model) and SAIGE-QTL [36] (a recent software for sc-eQTL mapping using a Poisson mixed-effect model) (**Figure 2B; Table S4**). Across 14 cell types, jaxQTL identified 16% more eGenes compared with tensorQTL ($P = 2 \times 10^{-47}$) and 11% more eGenes compared with SAIGE-QTL ($P = 3 \times 10^{-24}$), thus demonstrating jaxQTL increased power to identify eGenes in both rare and common cell types. The advantage of jaxQTL over SAIGE-QTL can be partially explained by gene-level calibration methods (permutation vs. ACAT-V), as performance gap decreased when applying ACAT-V in jaxQTL (5% more eGenes; **Figure 2B**). We note that our tensorQTL results identified more eGenes than tensorQTL in ref. [36], likely due to different procedures to create pseudobulk data (see **Methods**). Finally, we confirmed that jaxQTL-linear results agreed with our tensorQTL results since the Wald test and score test are asymptotically equivalent (97% overlap; **Figure S14**), with differences up to gene-level *P* value obtained by the permutation approach.

Next, we evaluated the computational performance of jaxQTL in cis-eQTL mapping compared with existing approaches using 50 randomly selected genes from chromosome 1 in OneK1K (see **Methods**; **Figure S15**). The average run time of jaxQTL on GPU/TPU across 3 cell types is 3.7x faster compared with SAIGE-QTL (12 vs. 44 mins) and 9.2x slower compared with tensorQTL (1.3 mins). To demonstrate the impact of sample size on run time, we simulated data for varying sample size by downsampling from N=100 to 700. The average run time is 39 mins for SAIGE-QTL, 15 mins for jaxQTL on GPU/TPU, and 2 mins for tensorQTL. To mimic TenK10K data[21], we performed upsampling to simulate data for N=10,000 (see **Methods**). Focusing on a dominant cell type such as $CD4_{NC}$ cells, jaxQTL on GPU was at least 1,560x faster (30 mins) compared with SAIGE-QTL, highlighting the efficiency of jaxQTL when applied to ever-increasing population-scale single-cell data. We note that the runtime of jaxQTL is dominated by performing permutations. Importantly, when performing ACAT-V to compute gene-level P values on CPU, jaxQTL was 1.3 - 10,596x times faster than SAIGE-QTL for N=100 to 10,000, and comparable with the linear model of tensorQTL (with permutations).

In summary, jaxQTL outperforms other models and methods in identifying eGenes in OneK1K, highlighting that pseudobulk-approaches for scRNA-seq are powerful (even for rarer cell types) when appropriately modeling count data. In addition, we observed that its computation time can scale to scRNA-seq datasets with thousands of individuals approaching that of classical linear models.

jaxQTL results replicate across datasets and ancestries

To verify that increased eGene detection from jaxQTL is not driven by false positives, we first replicated our sc-eQTLs results in 88 European- and 88 Asian-ancestry individuals from CLUES PBMC scRNA-seq study (**Figure 3**; **Figure S16, S17; Table S5**)[47]. Of the lead SNP-eGene pairs found in matched CLUES cell types, 40-86% can be replicated in a European cohort and 23-74% in an Asian cohort at FDR < 0.05 with concordant directional effect. Additionally, we replicated 75-92% sc-eQTLs in EUR whole blood samples from GTEx (N=588)[10] and 50%-78% in FACS-sorted immune cell types from DICE study (N=91)[48] (**Figure S18; Table S6**). We also observed consistent direction of sc-eQTL effects when comparing with shared lead SNP-eGene results in original OneK1K results (**Figure S19**). Lastly, we recapitulated the depletion of selection constraint and short enhancer domains in eGenes [10,49,50] (**Figure S20;** see **Web resources**). Consistent with refs. [49,51–53], we observed genes depleted of loss-of-function mutations (pLI > 0.9) had smaller sc-eQTL effect sizes (**Figure S21**), and that the effect size of lead SNPs of eGenes was smaller at lower allele frequency (**Figure S22**), confirming selection constraints on gene expressions in immune cell types. Altogether, these results demonstrate that jaxQTL results replicate across datasets and recapitulate known findings from bulk-eQTL studies.

sc-eQTLs are more enriched in cell-type-matched CREs than bulk-eQTLs

To characterize sc-eQTLs and their potential downstream role in human diseases, we performed fine-mapping for eGenes identified by jaxQTL on OneK1K using the SuSiE summary statistics approach [54]. Briefly, SuSiE performs Bayesian variable selection to identify likely causal SNPs in the form of credible sets and provide posterior inclusion probabilities (PIPs) to quantify uncertainty in its selection (see **Methods**). After restricting to the 18,281 non-MHC and non-MAPT eGenes identified by jaxQTL, SuSiE reported 95% credible sets (CS) for 12,978 eGenes across 14 cell types (6,776 unique eGenes). The average number of CSs per eGene was 1.15 with a median size of 16 SNPs, with 88% of eGenes explained by a single causal variant. We observed that the average number of CSs tracked with cell type proportion (P = 1.49 x 10^{-5}; **Figure S23**), suggesting the fine-mapping results were likely biased by lower statistical power to pinpoint independent causal eQTLs in rarer cell types. To establish a baseline, we also performed cis-eQTL and fine-mapping analyses using “bulk” gene expression (i.e., summed over cell types). As expected, we found that sc-eQTLs successfully identified a CS for more unique eGenes than bulk-eQTLs (6,776 vs 6,338, P = 9.5 x 10^{-23}).

To characterize the functional architecture of fine-mapped sc-eQTLs, we performed enrichment analysis using cell-type-agnostic annotations from S-LDSC baseline model[55] (see **Methods**). Consistent with bulk-eQTLs [56], fine-mapped sc-eQTLs (PIP ≥ 0.5) across cell types were highly enriched in promoter-like regions, enhancers, and evolutionarily conserved regions (**Figure 4A**). Overall, we observed greater enrichments for these annotations when using sc-eQTLs compared with bulk-eQTLs, however these differences were not significant, likely resulting from baseline annotations not reflecting cell-type-specificity. Additionally, we observed fine-mapped sc-eQTLs were less likely to be near promoter regions (P = 6.71 x 10^{-4}) and more distal (P = 1.07 x 10^{-4}) when compared with bulk-eQTLs (**Figure 4B**), suggesting that single-cell eQTL mapping can better prioritize distal regulatory elements.

Next, we sought to compare the enrichment of cell-type-specific candidate cis-regulatory elements (cCREs), derived using single-cell omics or isolated cell types, between fine-mapped sc-eQTLs and bulk-eQTLs. First, we confirmed that sc-eQTLs were enriched in cell-type-matched cCREs reflecting open chromatin, enhancers, and promoters (**Figure 4C; Methods**). We observed that sc-eQTLs were highly enriched in enhancer-gene links in matched cell types. Importantly, sc-eQTLs were more enriched in cell-type matched open chromatin (P = 6.77 x 10^{-9}), enhancers (P = 3.40 x 10^{-10}) and promoters (P = 1.26 x 10^{-6}) compared with bulk-eQTLs after meta-analysis across cell types (**Figure 4C**). Lastly, we further confirmed the accuracy of sc-eQTLs linked to their target genes by observing stronger enrichment in cell-type-matched enhancer-gene links identified by SCENT [57] compared to bulk-eQTLs (P = 3.38 x 10^{-4}; **Figure S24**), highlighting the necessity of conducting eQTL mapping using single-cell data to capture signals masked by bulk approach.

In summary, our analyses suggest that sc-eQTLs are enriched for distal cell-type-specific CREs that are likely missed by bulk-eQTL approach.

sc-eQTL location predicts cell-type specificity

Understanding how sc-eQTLs are shared across cell types is challenging due to differential statistical power between cell types. Specifically, simply counting sc-eQTLs based on their significance or fine-mapping results would conclude to a high cell-type specificity of sc-eQTLs (**Figure S25**; results similar to the ones reported by ref. [19]), but ignores the pervasive correlation of eQTL effects between cell types (**Figure S26**; as observed using eQTLs from bulk cell type samples in ref. [37]). To mitigate this, we analyzed fine-mapped sc-eQTLs using *mashr*, which provides posterior effect estimates and significance of effect after accounting for the effect size correlation between cell types and residual correlation due to sample overlap [58] (2,012 sc-eQTLs with PIP ≥ 0.5 and significant at a local false sign rate (LFSR) < 0.05 in at least one cell type).

Using the *mashr* estimated effect sizes, we found that 67% of fine-mapped sc-eQTLs were shared by sign across all 14 cell types (**Figure 5A**), suggesting their directional effect on gene expression was consistent. In contrast, eQTL effect size magnitudes were less shared due to effect size heterogeneity, with 15% universally shared and 9% specific (**Figure 5A**), consistent with previous findings [10,58,59]. Cell-type-specific sc-eQTLs were most common in monocytes, reflecting their difference from lymphocytes such as B cells and T cells (**Figure S27**).

We found cell-type-specific sc-eQTLs identified were more distal from TSS compared with shared sc-eQTLs (mean 102,164 vs. 53,733 bp, P = 2.17 x 10^{-6}; **Figure 5B**), consistent with previous work showing that distal CREs were more likely to be cell-type-specific [60,61]. Lastly, we calculated the pairwise sharing of sc-eQTL by magnitude and found cell type sharing outside expected subtype groups (**Figure S28**). For example, we found 84% shared sc-eQTLs between effector memory CD8+ T cells and natural killer (NK) cells, suggesting their shared cytotoxic effector mechanisms between adaptive and innate immunity [62].

To validate the specificity of cell-type-specific sc-eQTLs identified by *mashr* (in comparison with a baseline simple counting approach), we first calculated the replication rate and observed that cell-type-specific eQTLs after *mashr* analysis were less replicated in eQTLGen (0.60 vs. 0.77, P = 3.11 x 10^{-7}) and GTEx (0.33 vs. 0.47, P = 5.19 x 10^{-4}) respectively, consistent with the expectation that sc-eQTL effects private to a single cell type were likely masked by the bulk-eQTL study. Second, we found the cell-type-specific sc-eQTLs identified by *mashr* were more distal from TSS compared to the simple counting approach (mean 102,164 vs. 63,873 bp, P = 9.56 x 10^{-4}), suggesting *mashr* was more powerful in identifying distal cell-type-specific sc-eQTLs. Finally, we identified scATAC-seq peaks exclusive to each cell type and calculated the enrichment of cell-type-specific open chromatins. We observed the specific eQTLs identified by *mashr* were more enriched in cell-type-specific open chromatin in rarer cell types (**Figure S29**), such as non-classical monocytes ($\mathrm{Mono_{NC}}$), NK recruiting ($\mathrm{NK_R}$) cells, CD4+ T cells expressing SOX4 ($\mathrm{CD4_{SOX4}}$), and Plasma, which reflected eQTL sharing results (**Figure S27**).

In summary, our analyses suggest that the genetic effect of sc-eQTLs are largely shared across cell types. sc-eQTLs closer to TSS are more likely to be shared while distal sc-eQTLs are likely cell-type-specific due to the different regulatory elements involved in transcription.

sc-eQTLs reveal cell types associated with GWAS immune traits

After observing that sc-eQTLs improved our ability to identify cell-type-specific eQTLs, we next investigated whether sc-eQTLs can improve the interpretation of GWAS findings, which SNP-heritability (h^2) tend to be concentrated in SNPs within cCREs [55,63,64] and genes active in disease-relevant cell types [65–67]. To quantify the extent to which sc-eQTLs can characterize GWAS findings, we first evaluated the fraction of h^2 explained by SNP-annotations constructed from fine-mapped sc-eQTLs in all PBMC cell types and meta-analyzed S-LDSC results across 16 immune diseases and blood traits (**Table S7**). We found that the union of fine-mapped sc-eQTLs across 14 cell types (3.5% of common SNPs) were enriched in h^2 (2.82 ± 0.25), and explained 9.90 ± 0.88% of h^2. Importantly, fine-mapped sc-eQTLs explained more h^2 than bulk-eQTLs (6.10 ± 0.76%; **Figure 6A**) while maintaining comparable h^2 enrichment (2.76 ± 0.34; **Figure S30**), suggesting that sc-eQTLs increase the number of GWAS variants that can be functionally characterized.

We further investigated whether h^2 was enriched within sc-eQTLs from disease-relevant cell types (**Figure 6B**). We ran S-LDSC on sc-eQTL annotations built from each cell type, and individually looked at their effects while conditioning on the union of fine-mapped sc-eQTLs. Overall, identified cell types were consistent with known biology and previous studies leveraging cCREs and genes differentially expressed. For example, we identified monocyte cell types for monocyte percentage (min P = 1.80 x 10^{-3} for $\mathrm{Mono_{NC}}$) and major sub-cell-types of B cells and T cells for lymphocyte percentage (min P = 4.26 x 10^{-4} for $\mathrm{CD4_{NC}}$). For immune diseases, we identified various T cell subtypes for celiac disease [66] (min P = 7.65 x 10^{-3} for $\mathrm{CD4_{NC}}$), inflammatory bowel disease [66] (min P = 1.34 x 10^{-4} for naïve and central memory CD8+ T cells ($\mathrm{CD8_{NC}}$)), hypothyroidism [67] (min P = 9.91 x 10^{-3} for $\mathrm{CD8_{ET}}$), primary biliary cirrhosis [66] (min P = 3.45 x 10^{-2} for $\mathrm{CD8_{NC}}$), and rheumatoid arthritis [66,68] (min P = 3.43 x 10^{-3} for $\mathrm{CD8_{ET}}$), as well as the role of NK cells for eczema [69](min P = 1.12 x 10^{-2} for NK).

Altogether, these results highlight that sc-eQTLs can improve our ability to characterize GWAS findings by identifying new eQTLs in disease relevant cell types.

sc-eQTLs prioritize candidate genes missed by bulk-eQTLs

To demonstrate that sc-eQTLs can prioritize GWAS candidate genes that would have been missed by bulk sequencing, we present OneK1K sc-eQTLs and bulk-eQTLs results at a leading genetic risk loci associated with rheumatoid arthritis (RA) [70] in *ANKRD55-IL6ST* region [71,72] (**Figure 7**). We identified the GWAS leading SNP rs7731626 (chr5:55444683:G>A) as the top candidate causal sc-eQTLs for *IL6ST* in CD4_{NC} (PIP = 1) and CD8_{NC} T cells (PIP = 0.98), and observed significant colocalization with RA for these two cell types (PP.H4 = 1, and PP.H4 = 0.99, respectively). In contrast, this SNP became null in bulk-eQTL results as its eQTL effect on *IL6ST* was diluted when cell types were lumped together. We further confirmed that rs7731626 is likely to be within an enhancer acting on *IL6ST* regulation in T cells by observing contact between rs7731626 and *IL6ST* promoter exclusively in T cells using promoter capture Hi-C (PCHi-C) experimental data [73], as well as H3K27ac peaks (capturing enhancer activity) at this locus in naive CD4+ and CD8+ T cells [74,75] (**Figure 7**; see **Code and Data Availability**); the link between rs7731626 and *IL6ST* in T cells has also been established by other single-cell multi-omic data approaches [34,76–78]. We replicated the rs7731626-*IL6ST* association in CD4+ and CD8+ T cells in CLUES European- and Asian-ancestry individuals (N=88, p=0.03 respectively), except rs7731626 in CD8+ T cells among Europeans (p=0.3), likely due to lower sample size for detecting weaker effect as evidenced by consistent effect direction. Besides, we also identified rs7731626 as both a bulk-eQTL and sc-eQTL for *ANKRD55* and similar colocalization with RA (**Figure S31**), illustrating that while bulk-eQTL approaches can identify strong sc-eQTL signal, they can nominate an incomplete list of candidate genes.

Additional colocalization results between RA GWAS and OneK1K sc-eQTLs are presented in **Figure S32**. We identified 43 eGene colocalizing with RA (PP.H4 > 0.9), primarily in T and B cells (consistent with literature [79]). We notably found that *EOMES* eQTLs colocalized with RA in CD8+ T cells (PP.H4 = 0.97), consistent with the role of *EOMES* as a key TF for mediating immunity function in effector CD8+ T cells[80]. Similarly, *CD40* eQTLs colocalized with RA in Plasma cells, reflecting that *CD40* signaling pathway plays an essential role in immune response in B cells [81,82].

Altogether, these results illustrate that sc-eQTL analysis can reveal new candidate genes for diseases that are masked by bulk approach.

Discussion

We developed jaxQTL, an efficient and powerful approach for large-scale eQTL mapping on single-cell data using count-based models. Through simulation and real data analyses, we showed that the negative binomial model was the most powerful and well-calibrated model compared with the linear and Poisson models. In an application to OneK1K, we found that jaxQTL was more powerful in identifying eGenes compared to tensorQTL (linear model) and SAIGE-QTL (Poisson mixed model) while exhibiting comparable or better runtimes when performed using GPUs/TPUs.

We further leveraged jaxQTL results to characterize eGenes and their corresponding fine-mapped sc-eQTLs. First, we found that eGenes are depleted of loss-of-functions variants and large-effect eQTLs, consistent with previous works on bulk-eQTLs [49,51–53]. Second, we showed that sc-eQTLs are more distal to TSS and more enriched in cell-type matched CREs compared with bulk-eQTLs, and that sc-eQTLs effects are largely consistent across cell types (as observed using eQTLs from bulk cell type samples in ref.[37]), with cell-type-specificity increasing with distance to TSS. These results summarize that while bulk-eQTLs identify primarily proximal regulatory effects with low cell-type-specificity (e.g., promoter), sc-eQTLs allow to identify additional distal regulatory effects with medium to high cell-type-specificity (e.g., enhancer). Finally, we demonstrated that sc-eQTLs explain more heritability than bulk-eQTLs for GWAS traits, suggesting that the GWAS risk variants were partially driven by eQTLs with medium to high cell-type-specificity. We used an example of *ANKRD55-IL6ST* loci to demonstrate that sc-eQTLs can prioritize RA-associated gene *IL6ST* missed by bulk approach. This candidate gene is well-documented for its functional role on the key therapeutic target cytokine *IL6* [83–85], and is further supported by other genetic evidence such as its higher PoPs prioritization scores [86] compared with the closest gene *ANKRD55* (top 1% vs 13% percentile respectively).

jaxQTL has several advantages compared to existing sc-eQTL methods. First, jaxQTL requires no data transformation on gene expression outcomes such that it provides an eQTL effect estimate on the interpretable count data scale. Second, we showed through simulations that jaxQTL outperformed the linear model (used by tensorQTL) in identifying eQTLs, especially for lower expressed genes and rare cell types. This will be well suited for rare cell types and precise cell states in short-read scRNA-seq and low counts of isoforms transcripts in long-read scRNA-seq. Lastly, jaxQTL leverages GPU/TPU to maximize efficiency in sc-eQTL mapping at population scale (N>100) while performing the permutations requested in eQTL best practices [28]. Although SAIGE-QTL similarly used a count-based model without permutation calibration, jaxQTL on GPU/TPU was on average 3.7x faster than SAIGE-QTL with sample size observed (N=982) in OneK1K and 1,560x faster in simulated data for dominant cell type with sample size N=10,000 (mimicking the upcoming TenK10K data). Importantly, when performing ACAT-V to compute gene-level P values, jaxQTL was 1.3 - 10,596x times faster than SAIGE-QTL for N=100 to 10,000, and displayed times comparable with the linear model of tensorQTL (with permutations).

Our findings have several implications for further single-cell sequencing studies and their integration with GWAS. First, we demonstrated the benefits of the negative binomial model over the Poisson mixed model for sc-eQTLs mapping. jaxQTL can be applied to other count data such as peak reads from scATAC-seq, or extend to accommodate other molecular outcomes such as isoform ratios, while efficiently accounting for the larger number of tests to perform in these datasets (e.g. ~300K of ATAC-seq peaks in scATAC-seq [87] vs. ~20K genes in scRNA-seq). We thus recommend considering jaxQTL with the negative binomial models for further analyses of population-scaled scATAC-seq and/or multiome datasets to identify chromatin activity QTLs. Second, jaxQTL is scalable for identifying sc-eQTLs in extremely large scRNA-seq datasets. As advances in the ability to multiplex samples in single-cell sequencing assays is allowing generating scRNA-seq and multiome datasets in hundreds to thousands of samples, we expect jaxQTL to be leveraged for analyzing datasets beyond OneK1K and TenK10K. Third, we highlighted that sc-eQTL effects are more shared across cell types than previously reported [19]. This property might motivate new sc-eQTL mapping and fine-mapping methods jointly integrating all cell types to increase power. Finally, we observed that while OneK1K sc-eQTLs explained a higher proportion of heritability than bulk-eQTLs for GWAS of immune traits, they did not close the missing link between GWAS and eQTLs. While generation of larger scRNA-seq datasets will improve the detection of distal sc-eQTLs with high cell-type-specificity, closing this gap might involve identifying sc-eQTLs with *trans*-effects, generating scRNA-seq data from disease-relevant contexts (such as stimulating condition [88] and developmental stage[89]), or generating other types of single-cell QTLs (such as splice QTLs[90], chromatin activity QTLs[87] and methylation-QTL [91]). In all those scenarios, jaxQTL can be easily extended to efficiently analyze those datasets.

We note several limitations of our work. First, jaxQTL power is dependent on cell type abundance, which will limit sc-eQTL power for rare cell types and precise cell states. Despite this, jaxQTL identified more eGenes within rare cell types than existing methods. Second, pseudobulk approaches aggregate read counts over discrete cell types, which may fail to capture dynamic contexts, thus limiting the identification of dynamic sc-eQTLs[34,35] (i.e. variants impacting gene expression within a cell type whose effects vary dynamically along a continuous state). Identifying dynamic sc-eQTLs with jaxQTL could be performed by identifying sc-eQTLs in more precise cell states, but further work would be required to evaluate this approach. Third, our analyses were restricted to PBMCs and immune-related diseases, and it is unclear if our conclusions translate to sc-eQTLs from different tissues and disease types. However, ongoing release of sc-RNAseq data from brain samples will allow characterizing brain sc-eQTLs and their role in psychiatric traits[20]. Fourth, our analyses were also restricted to European individuals. Assessing the transportability of our conclusions in non-European populations is a critical future research direction, as different environments and genetic backgrounds impact gene regulation and disease effect sizes [92]. While the recent release of a population-scaled scRNA-seq from East-Asian individuals will allow identifying sc-eQTLs and characterizing their role in East-Asian populations[90], there is a need to generate datasets in more diverse populations. Despite these limitations, jaxQTL provides an efficient and flexible framework for eQTL mapping on single-cell data using a count-based model. Our findings enable rich biological and mechanistic interpretation for disease risk loci at the cell-type level and nominate therapeutic targets for complex diseases.

Declaration of interests

S.G. reports consulting fees from Eleven Therapeutics unrelated to the present work. The other authors declare no competing interests.

Acknowledgments

This work was funded in part by the National Institutes of Health (NIH) under awards P01CA196569, R01HG012133, R01CA258808, R35GM147789, R01GM140287, and R00HG010160. We thank Dr. Joseph Powell for kindly sharing their processed OneK1K gene expression count matrix and genotype data. We thank the ENCODE consortium for generating experimental data for epigenetic markers by cell types. We thank members from the Impact of Genomic Variation on Function (IGVF) consortium QTL focus group for insightful discussions on these results. We thank participants for providing feedback for this work during the Computational Genomics Summer Institute (CGSI) program at UCLA under award GM135043. We thank Dr. Bolun Li for providing user feedback on jaxQTL software.

Author Contributions

Z.Z., N.M., and S.G. conceived the study. Z.Z., S.G., and N.M. developed the method. Z.Z. performed analyses. N.S. and A.K. prepared data and performed analyses. All authors edited and approved the manuscript.

Web Resources

JAX: https://github.com/google/jax
tensorQTL: https://github.com/broadinstitute/tensorQTL
SAIGE-QTL: https://github.com/weizhou0/qtl
bedtools: https://bedtools.readthedocs.io/en/latest/
qvalue R package: https://github.com/StoreyLab/qvalue
susieR package: https://github.com/stephenslab/susieR
GTEx pipeline: https://github.com/broadinstitute/gtex-pipeline/tree/master
SLDSC software: https://github.com/bulik/ldsc
1000 Genome annotations: https://alkesgroup.broadinstitute.org/LDSCORE/
PLINK software: https://www.cog-genomics.org/plink/2.0/
mashr: https://stephenslab.github.io/mashr/index.html

GTEx v8 eQTL summary statistics (EUR): https://gtexportal.org/home/downloads/adult-gtex/qtl
DICE cis-eQTL summary statistics: https://dice-database.org/
eQTL catalogue: https://www.ebi.ac.uk/eqtl/
CLUES data: Gene expression data are available in the Human Cell Atlas Data Coordination Platform (https://explore.data.humancellatlas.org/projects/9fc0064b-84ce-40a5-a768-e6eb3d364ee0/project-

matrices) and at GEO accession number GSE174188. Genotypes are available at dbGap accession number phs002812.v1.p1

GTEx v8 SuSiE fine-mapping results: https://www.finucanelab.org/data
Gene GTF file (release 84):
https://ftp.ensembl.org/pub/grch37/release-84/gtf/homo_sapiens/Homo_sapiens.GRCh37.82.gtf.gz
pLI and LOEUF score from gnomad:
https://storage.googleapis.com/gcp-public-data--gnomad/release/4.0/constraint/gnomad.v4.0.constraint_metrics.tsv

Data and Code Availability

Single-cell eQTL summary statistics results produced by jaxQTL and fine-mapping results are available on Zenodo:10.5281/zenodo.14624945

jaxQTL software: https://github.com/mancusolab/jaxQTL
jaxQTL analysis code: https://github.com/mancusolab/jaxqtl_analysis
Original Onek1k data are available at:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5901755

We downloaded the call sets from the ENCODE portal (https://www.encodeproject.org/) with the following identifiers: ENCFF313TWH (CD4 T cell), ENCFF635YOQ (CD8 T cell), ENCFF071MEQ (B cell), ENCFF814VKT (NK cell), and ENCFF468QFA (Monocyte).

Online Methods

GLM and count-based eQTL models

Here, we describe the core generalized linear model (GLM) for a focal gene within a focal cell type c, assuming that cell type labels for each cell have been provided. Specifically, jaxQTL models the conditional expectation of pseudobulk counts y_c (i.e. sum of read counts across cells within the cell type), given covariates X (e.g., age, sex, genotyping principal components) and a cis-genetic variant g, as

$$E[y_c \mid X, g] = h(X\beta + g\beta_g + l_c),$$

where β are the covariate effects, β_g is the allelic effect, l_c is an offset adjusting for differences in library size, and $h(\cdot)$ is a function which maps linear predictions to expected values. Additionally, jaxQTL models the conditional variance as,

$$Var[y_c \mid X, g] = Var_\alpha(h(X\beta + g\beta_g + l_c)),$$

where $Var_\alpha(\cdot)$ corresponds to a variance function determined by the specified likelihood (e.g., Poisson, negative binomial) allowing for an overdispersion parameter α if required (e.g., negative binomial) [45,93].

We fit a null GLM using iteratively reweighted least squares (IRWLS). Namely, assuming there is no genotype effect (i.e. $\beta_g = 0$), the update step for covariate effects β at the $t+1$ iteration can be computed by solving a linear system given by,

$$X^T W_t X \beta_{t+1} = X^T W_t (e_t + X\beta_t),$$

where $e_t \propto (y - \mu_t)$ are the "working residuals", $\mu_t = h(X\beta_t + l_c)$, and W_t are the GLM weights proportional to the reciprocal of the conditional variance. To solve this linear system, jaxQTL allows for different solvers (e.g., QR, Cholesky, and conjugate gradient), however in practice we found Cholesky to outperform other approaches.

While both the Wald test and score test are implemented in the software, jaxQTL employs the score test in assessing the nonzero cis-SNP effect β_g on y_c for its improved computational efficiency[94]. Specifically, we first fit the null GLM model described above. Next, jaxQTL uses a block matrix approach to efficiently compute score test statistics for all p cis-SNPs in a focal gene. Let $d = diag(\hat{G}^T W \widehat{G})$ be a length p vector, then the vector of test statistics Z is given by,

$$Z = (\hat{G}^T W e) \oslash d^{1/2},$$

where $\hat{G}$ is the weighted residualized genotype obtained by $\hat{G} = G - X(X^T W X)^{-1} X^T W G$, $d^{1/2}$ is the element-wise square-root, and $\oslash$ is element-wise division.

To adjust for multiple testing corrections in eGene discovery, jaxQTL provides gene-level *P* values calibrated by a permutation-based Beta-approximation approach, which is similarly implemented in FastQTL and tensorQTL software [28,29]. To infer beta distribution parameters, we applied natural gradient descent using second order approximation to ensure parameters stay on the manifold (i.e. > 0)[95]. Lastly, we controlled the false discovery rate (FDR) and identified eGenes using these gene-level *P* values.

We implemented jaxQTL in Python to enable *just-in-time* (JIT) compilation through the *JAX* package (**Web Resources**), which generates and compiles heavily optimized C++ code in real-time and operates seamlessly on CPU, GPU, or TPU (**Code Availability**).

Analysis of OneK1K single-cell data

We obtained 1,267,768 PBMC blood cells for 14 cell types from *N* = 982 healthy individuals of European ancestry in the OneK1K cohort[19]. Each donor has an average of 1,291 cells (range 62-3,501). Each cell type has a varying number of donors due to sampling variance (**Table S2**). For sc-eQTL mapping, we created pseudobulk count data for each of the pre-annotated 14 cell types. After retaining genes with sample-coverage (i.e., fraction of non-zero expression read counts) in at least 1% of the population for each cell type, we performed sc-eQTL mapping for an average of 16,096 genes per cell type (**Table S2**). To establish a baseline for comparison, we created "bulk" data by summing all read counts across cell types for every gene and identified bulk-eQTLs using jaxQTL.

We first aimed to benchmark between different sc-eQTL models, including linear, negbinom, and Poisson. For negbinom and Poisson, we calculated individual library size in each cell type, i.e., the offset term in the model, by summing read counts across all genes per individual. For the linear model, we normalized gene expression read counts between individuals by TMM approach and then normalized across individuals by rank-based inverse normal transformation, as performed in GTEx [10]. We implemented all three models using a score test in jaxQTL.

For genotype data, we retained 5,313,813 SNPs with imputation INFO score >0.8, MAF>0.05, and Hardy Weinberg equilibrium (HWE) *P*>1e-6. Genotype PCs were calculated using 459,603 LD-pruned SNPs with INFO>0.9, MAF>0.01, and HWE *P*>1e-6. Across all sc-eQTL models, we adjusted covariates including age, sex, first six genotype PCs, and two expression PCs computed in each cell type. We defined the cis-window size as ±500kb (total 1Mb) around TSS. We downloaded the gene annotation GTF file (Homo_sapiens.GRCh37.82) and collapsed it to a single transcript model using "collapse_annotation.py" from GTEx analysis pipeline (**Web Resources**)[10]. We controlled gene-level FDR at 0.05 per cell type using the qvalue method on gene-level *P* values through *qvalue* R package (**Web Resources**) [96]. eGenes were identified by qvalue < 0.05. We used genome build hg19 for all variants and gene annotations.

To benchmark with other existing software, we compared the eGenes results of jaxQTL against tensorQTL [29] and SAIGE-QTL [36]. We first restricted genes to sample-coverage > 10% as in ref. [36] and obtained their SAIGE-QTL eGenes results [36]. Then we applied FDR control on this subset of genes to call eGenes for jaxQTL-negbinom. Similarly, we performed sc-eQTL mapping using tensorQTL on this set of genes (**Web Resources**). We note that the tensorQTL results we report are different from tensorQTL results reported in ref. [36], as we sum pseudobulk counts (following GTEx recommended guidelines to transform counts from bulk RNA-seq [10]) rather than averaging them. All reported P values are two-sided unless specified otherwise.

Simulations

To evaluate the performance of the jaxQTL-linear, jaxQTL-negbinom, jaxQTL-Poisson, tensorQTL (linear), and SAIGE-QTL (Poisson mixed effect) models, we first simulated read counts y_{ij} for individual i in cell j for a focal gene under the Poisson mixed effect model, given by:

$$log(E(y_{ij}|\ g_i, u_i, l_{ij}))\ = \beta_0 + u_i +\ g_i\,\beta_g + log(l_{ij}),$$

where β_0 is a baseline intercept, $u_i\ \sim N(0, \sigma^2{}_u)$ is the random intercept for individual i that induces within-sample correlation across cells, g_i is genotype with effect-size $\beta_g \sim N(0, h^2{}_{cis})$ where $h^2{}_{cis}$ is cis-SNP heritability. To reflect the cell-wise and individual-wise read counts (i.e., library size l_{ij}) observed in scRNA-seq data, we sampled l_{ij} from empirical values observed in OneK1K data. To fit pseudobulk linear, negbinom, and Poisson models, we created pseudobulk counts as $y_i = \sum_j y_{ij}$ and library size as $l_i = \sum_j l_{ij}$. We varied the baseline expression β_0 , cis-SNP heritability $h^2{}_{cis}$, the random intercept variance $\sigma^2{}_u$, MAF, and sample size N. Given fixed β_0 values, we obtained varying sample-coverage across simulation replicates and calculated the average simulated sample-coverage (**Figure 1**; **Figure S4-8**).

To evaluate the performance under model misspecification, we simulated single-cell read counts from the standard Poisson model assuming no within-individual correlation between cells, i.e., $\sigma^2{}_u = 0$. In each scenario, we performed a score test for association between simulated gene expression and genotype after fitting jaxQTL-linear, jaxQTL-Poisson, jaxQTL-negbinom, and SAIGE-QTL models. Different from the score test in jaxQTL-linear, tensorQTL reports Wald test statistics. For the linear models (jaxQTL-linear and tensorQTL), we normalized pseudobulk counts by rank-based inverse normal transformation [10,97]. Each simulation had 500 replicates.

Replication of sc-eQTLs

To validate sc-eQTLs identified by jaxQTL-negbinom, we performed replication analysis using jaxQTL on two independent cohorts from CLUES study [47]. We obtained scRNA-seq for N=256 individuals of European and Asian ancestry (see **Web resources**). We removed 50 control individuals from the ImmVar study, 2 outliers detected through a PCA analysis and 2 male individuals from the remaining based on ref [92]. After intersecting with genotype data, we retained 88 European- and 88 Asian-ancestry individuals for replication analysis. Of these, 65 and 67 individuals were diagnosed with systemic lupus erythematosus (SLE) but were not in active state of disease flare. We matched 7 cell types in CLUES with 14 cell types in OneK1K. We performed analysis in European and Asian individuals separately. For lead SNP-eGene pairs identified by jaxQTL-negbinom, we fitted negbinom model using jaxQTL in CLUES cell types. We adjusted for age, sex, first six genotype PCs, SLE status, and batch numbers in sc-eQTL model. For each cell type, we reported the fraction of pairs replicated at FDR < 0.05 using *qvalue* R package[96]. To compare sc-eQTL effect size estimated by jaxQTL in CLUES and OneK1K samples adjusting for fitted count scale differences, we calculated an adjusted slope estimate as $\tilde{\beta} \propto \hat{\beta}\ w$ where $w \approx \sqrt{2p(1-p)}$ after accounting for weights in the GLM. Lastly, we compared results reported in the original OneK1K linear-model based analyses to demonstrate directional consistency in sc-eQTL effect estimates[19].

We further investigated the replication of lead SNP-Gene pairs in eQTLs identified by previous bulk-eQTL and sc-eQTL studies (see **Web resources**). For bulk-eQTL studies, we downloaded 1) cis-eQTL results from European whole blood samples in GTEx v8 (N=588)[10]; 2) cis-eQTL summary statistics results from FACS-sorted PBMC immune cell types in DICE study (N=91) [48] . For sc-eQTL study, we downloaded cis-eQTL summary statistics from PBMC scRNA-seq data in CLUES study curated by eQTL catalogue (N=193) [47,98]. All these prior results are based on the linear model approach when performing eQTL mapping. Again we reported the fraction of pairs replicated at FDR < 0.05 within each cell type.

Computational runtime

To evaluate the computational runtime of jaxQTL on cis-eQTL mapping in comparison with other software, we randomly selected 50 genes from chromosome 1 with sample-coverage > 1% in $\mathrm{CD4_{NC}}$, $\mathrm{B_{IN}}$, and Plasma cells observed in OneK1K. To benchmark the performance across different sc-eQTL sample sizes, we performed downsampling to create gene expressions for N=100, 300, 500, and 700 individuals. Moreover, we performed upsampling for the expected TenK10K cohort[21]. Specifically, we sampled N=10,000 individuals using the single-cell data matrix. Assuming each individual has a total of 5,000 cells as expected in TenK10K, we sampled the number of cells per person proportionally as observed for these three cell types. Then we created pseudobulk data for jaxQTL and tensorQTL.

Fine-mapping on sc-eQTLs

To identify causal eQTL for every eGene, we performed fine-mapping using SuSiE summary statistics approach for eGenes identified by sc-eQTL and bulk-eQTL approach (both jaxQTL-negbinom). We excluded eGenes in the *MHC* region (chr6: 25Mb-34Mb) with complex LD patterns and the *MAPT* region (chr17: q21.31) with complex inversion and duplication [99,100]. For optimal statistical power, we first used jaxQTL-negbinom to compute all pairwise summary statistics for cis-SNPs in every eGene. We calculated the in-sample LD correlation matrix for cis-SNPs $\hat{R}$ after projecting out the covariates effect under the GLM weights. Specifically for negbinom model results, we calculated a weighted residualized G by $\hat{G} = G - X(X^TWX)^{-1}X^TWG$, followed by computing $\hat{R} = D^{-1/2}\hat{G}^TW\hat{G}D^{-1/2}$, where $D = diag(\hat{G}^TW\hat{G})$ and W is the individual weights calculated after fitting the null model.

Enrichment analysis on sc-eQTLs

We downloaded annotations from the LDSC baseline model and selected 12 annotations of promoter-like regions, enhancers, conserved regions, and epigenetic markers (**Web Resources**). For cell-type matched candidate cis-regulatory elements (CREs), we downloaded CRE peaks in PBMC cells identified by scATAC-seq[101]. We extracted CREs from cell types matched with cell types in OneK1K based on labels and marker genes (**Table S8**). For enhancers and promoters, we collected 33 samples for 6 cell types from EpiMap and used *bedtools* to merge peak regions from different samples in the same cell type (**Table S9**). Lastly, to interrogate the accuracy of sc-eQTL linked to target genes, we obtained the enhancer-gene links identified by SCENT[57] in B cells, T/NK cells, and Myeloid cells. Cell types were matched based on labels (**Table S10**). We removed ENCODE promoter-like regions from SCENT peaks to retain putative enhancer regions.

For enrichment analysis on scATAC-seq, EpiMap enhancers, and promoters, we created annotations for cis-SNPs taking the value of 1 if falling within the CRE region and 0 otherwise. For every eGene, we performed logistic regression similar to torus[102] by fitting $logit(PIP_{k,j}) = \beta_0 + \beta_{k,a} a_{k,j}$, where j denotes cis-SNPs in eGene k and their annotation a. To obtain a single enrichment score for every annotation in a cell type, we performed a fixed effect meta-analysis using fitted slopes and their standard errors across all eGenes. Specifically, the meta-analyzed slope over eGenes is $(\sum_k \beta_{k,a} / W_{k,a})/(\sum_k W_{k,a})$ with variance $1/\sum_k W_{k,a}$, where $W_{k,a} = 1/SE(\beta_{k,a})^2$. When comparing sc-eQTLs against bulk-eQTLs enrichment, we meta-analyzed summary statistics across cell types.

To calculate the enrichment of sc-eQTLs in enhancer-gene pairs identified by SCENT, we defined the enrichment score for every eGene as:

$$Score_{eGene} = \frac{Number\ of\ causal\ SNP\ in\ annotation_{eGene} \ / \ Number\ of\ cis-SNP\ in\ annotation_{eGene}}{Number\ of\ causal\ SNP_{eGene} / Nnumber\ of\ cis-SNP_{eGene}}$$

Then we calculated the enrichment for each cell type by taking an average of $Score_{eGene}$. To evaluate the uncertainty of this mean enrichment score, we calculated the variance of this mean enrichment by bootstrapping 1,000 iterations.

Cell type sharing of sc-eQTLs

To investigate cell type specificity or sharing of sc-eQTLs, we performed the *mashr* analysis on 2,256 fine-mapped sc-eQTL (PIP≥0.5) with complete summary statistics across 14 cell types [58] (**Web Resources**). We first constructed a “finemap sc-eQTL” Z score matrix of size 2,256 x 14. To estimate residual covariance between cell types due to sample overlap, we constructed a null sc-eQTL matrix (21,542 x 14) by randomly sampling from 2 SNPs in every gene with max |Z| < 2 across all cell types.

Following the instructions described elsewhere [58], we used a data-driven approach to estimate the covariance matrix of the sc-eQTL effect. In brief, we first used the “finemap sc-eQTL” matrix to create 27 candidate covariance matrices including empirical covariance of Z score, 5 rank-1 approximation to the covariance matrix, rank-5 approximation, and 19 canonical covariance matrices created by *cov_canonical()*. Then we applied *cov_ed()* to estimate the covariance pattern by extreme deconvolution. Lastly, we estimated the mixture weight by fitting the *mashr* model on the null sc-eQTL matrix using 27 covariance patterns and residual covariance. Lastly, we fitted *mashr* on the “finemap sc-eQTL” matrix to obtain posterior effect estimates and local false sign rate (LFSR) using the mixture estimates from above. Since we used the Z score model of *mashr*, we converted the posterior estimate back to the effect scale by multiplying their standard errors as described elsewhere [58].

To count for sc-eQTL sharing, we first selected 2,012 sc-eQTLs with LFSR < 0.05, which was similar to FDR control. For each significant sc-eQTL, we called the cell type with the strongest *mashr* effect size as discovery cell type. We considered two types of eQTL sharing: 1) “share by sign” means the other cell type shared the sign of effect with the discovery cell type; 2) “share by magnitude” means conditioning on “share by sign”, the magnitude was within a factor of 2 compared to the discovery cell type.

For enrichment analysis of scATAC-seq peaks, we first used *bedtools subtract -A* recursively to identify peaks exclusive to each cell type, i.e., cell-type-specific peaks. Then we calculated the enrichment score using:

$$Enrichment_{CT} = \frac{Number\ of\ CT{-}specific\ eQTLs\ in\ Annot_{CT} /\ Number\ of\ analyzed\ eQTLs\ in\ Annot_{CT}}{Number\ of\ CT{-}specific\ eQTLs\ /\ Total\ number\ of\ analyzed\ eQTLs},$$

where CT refers to cell type and Annot$_{CT}$ is cell-type-specific peaks. To obtain standard errors for the enrichment, we performed bootstrapping with 1,000 iterations on the cell type labels for CT-specific eQTLs.

Integration sc-eQTLs with GWASs

To assess the overlap between sc-eQTL and GWAS risk variants, we performed an S-LDSC analysis on 16 GWAS results for blood and immune-related traits (**Web resources; Table S7**). Firstly, we created annotations using SNPs in credible sets of fine-mapped eGenes from 14 cell types (as recommended in ref. [91]). We constructed three sets of annotations for 1) cell-type sc-eQTL: a union of credible sets of eGenes per cell type; 2) sc-eQTL_union: a union of credible sets from 1) across all cell types; 3) bulk-eQTL: credible sets of eGenes in bulk-eQTL results. Then we annotated SNPs using European individuals from 1000 Genome Project and performed S-LDSC analysis using these annotations (**Web resources**). To estimate heritability, we used baseline-LD v2.2 model (96 annotations) as recommended in ref.[103] to obtain estimates reducing biases from MAF- and LD-dependent architectures. To identify the likely causal cell types associated with each GWAS trait, we fitted the baseline model v1.2 (53 annotations) to optimize statistical power as recommended in ref.[104]. To account for background non-cell-type-specific eQTLs in the baseline model, we construct an additional annotation by taking a union of fine-mapped SNPs in 95% credible sets from SuSiE results across 49 GTEx tissues (**Web resources**) and sc-eQTL_union from our OneK1K results in order to identify cell-type-specific effects. We focused on two metrics: 1) the proportion of heritability explained by each annotation $h^2(C)$ from the baseline-LD v2.2 model result, and 2) the standardized coefficient ${\tau_c}^*$ calculated by:

$${\tau_c}^* = \tau_c\ sd(\tau_c)/(h^2/M),$$

where τ_c, $sd(\tau_c)$, h^2 were estimated from the baseline model and M was the total number of SNPs that ${h^2}_g$ was computed on (M = 5,961,159) using 1000 Genome Project and baseline v1.2 model. Here τ^* is the change in per-SNP heritability with one standard deviation increase in annotation, which makes it comparable between annotations and GWAS traits. The P values are one-sided hypothesis test for $\tau^* >$ 0.

Figures

Figure 1: Negative binomial outperforms other models in identifying sc-eQTLs in realistic simulations

We simulated single-cell read counts using library size observed in three cell types (CD4$_{NC}$, B$_{IN}$, Plasma) representing different levels of cell type proportions (high, medium, low). We reported the type I error rate ($h^2{}_{cis} = 0$) **(A)** and power **(B)** of jaxQTL-linear, jaxQTL-negbinom, jaxQTL-Poisson, SAIGE-QTL, and tensorQTL models across different sample-coverage (i.e., percentage of non-zero expression read counts). We fixed cis-heritability $h^2{}_{cis}$ = 0.05, random intercept variance $\sigma^2{}_u$ (modeling similarity of cell read counts within the same person) = 0.2, sample size = 1,000, and MAF = 0.2; results when varying these parameters are reported in **Figures S5-S8**. Error bars represent 95% confidence intervals (CIs) estimated from 500 replicates. The dashed line in **(A)** represents a type I error of 0.05. SAIGE-QTL assumes single-cell counts while the rest assumes pseudobulk counts.

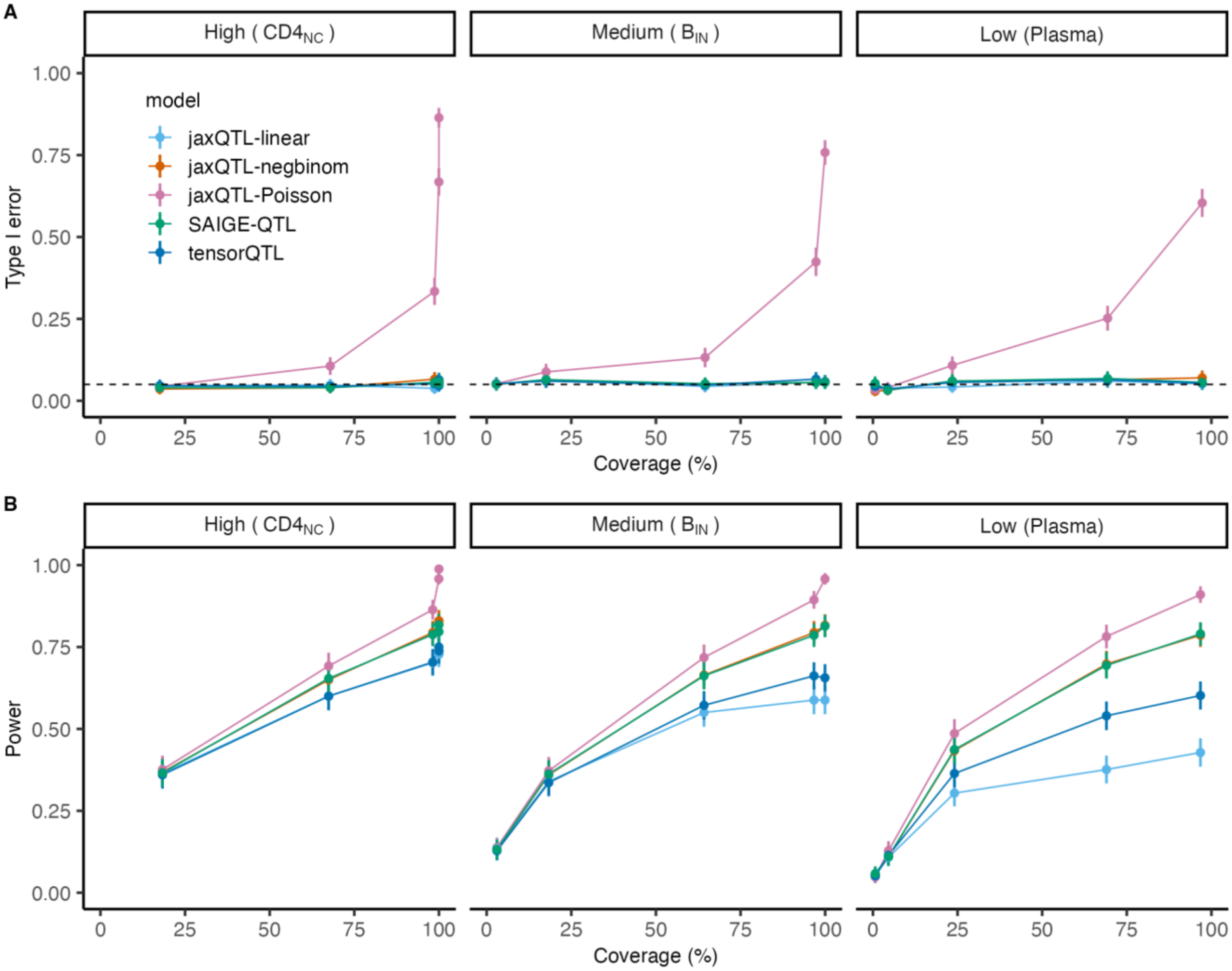


Figure 2: jaxQTL improves power for eGene discovery in the OneK1K dataset.

We compared eGene findings in OneK1K across models and software at FDR < 0.05. **(A)** For model comparison, we reported the number of eGenes identified by jaxQTL-negbinom, jaxQTL-linear, and jaxQTL-Poisson for genes with sample-coverage > 1%. **(B)** For software comparison, we reported the number of eGenes identified by jaxQTL (negbinom), jaxQTL (use ACAT-V instead of permutation method), tensorQTL, and SAIGE-QTL for genes with sample-coverage > 10% of individuals. The asterisks denote the cell type in which jaxQTL (negbinom) identified more eGenes compared with the linear model in **(A)** or SAIGE-QTL in **(B)** after Bonferroni correction. See details of description on cell types in **Table S2**. Numerical results are reported in **Table S3, S4**.

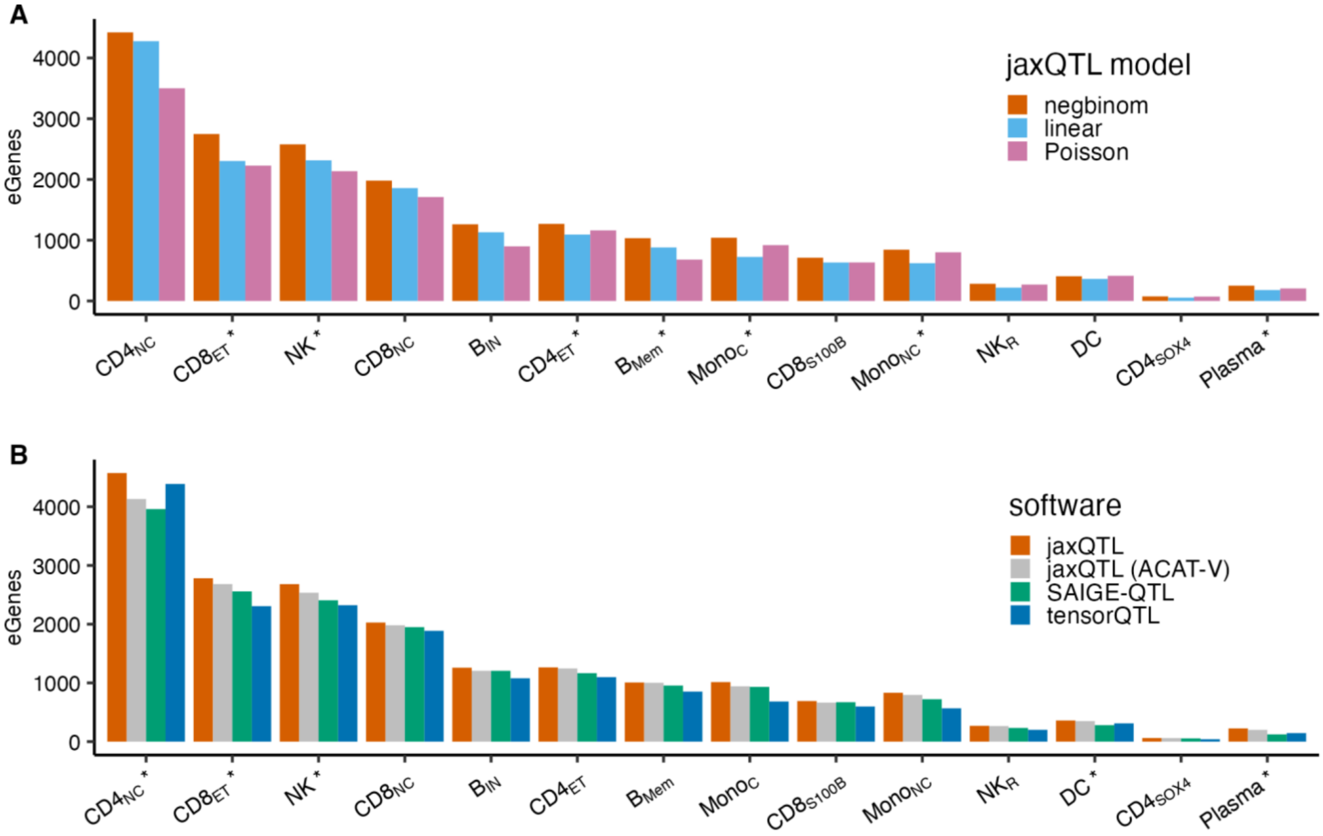


Figure 3: jaxQTL sc-eQTLs replicate in European and Asian samples

We performed replication analysis for 18,907 sc-eQTLs identified by jaxQTL in CLUES study. **(A)** Of the lead SNP-eGene pairs found in matched cell types (panels) among 88 European- and 88 Asian-ancestry individuals (14,229 and 13,579 sc-eQTLs respectively), we reported the replication rate at FDR < 0.05 by ancestry. **(B)** We plotted the adjusted sc-eQTL effect estimated by jaxQTL in CLUES versus OneK1K samples (see **Methods**). Colored points are pairs replicated at FDR < 0.05 in CLUES samples and grey points are otherwise. 35 and 20 pairs with absolute adjusted sc-eQTL effect estimate > 2 were truncated for visualization (see complete results in **Figure S16, S17** and **Table S5**). The colored line in **(B)** is a fitted linear regression line with a 95% confidence band. T4: CD4+ T cells; T8: CD8+ T cells; NK: natural killer cells; cM: CD14+ conventional monocytes; ncM: CD16+ unconventional monocytes; cDC: conventional dendritic cells.

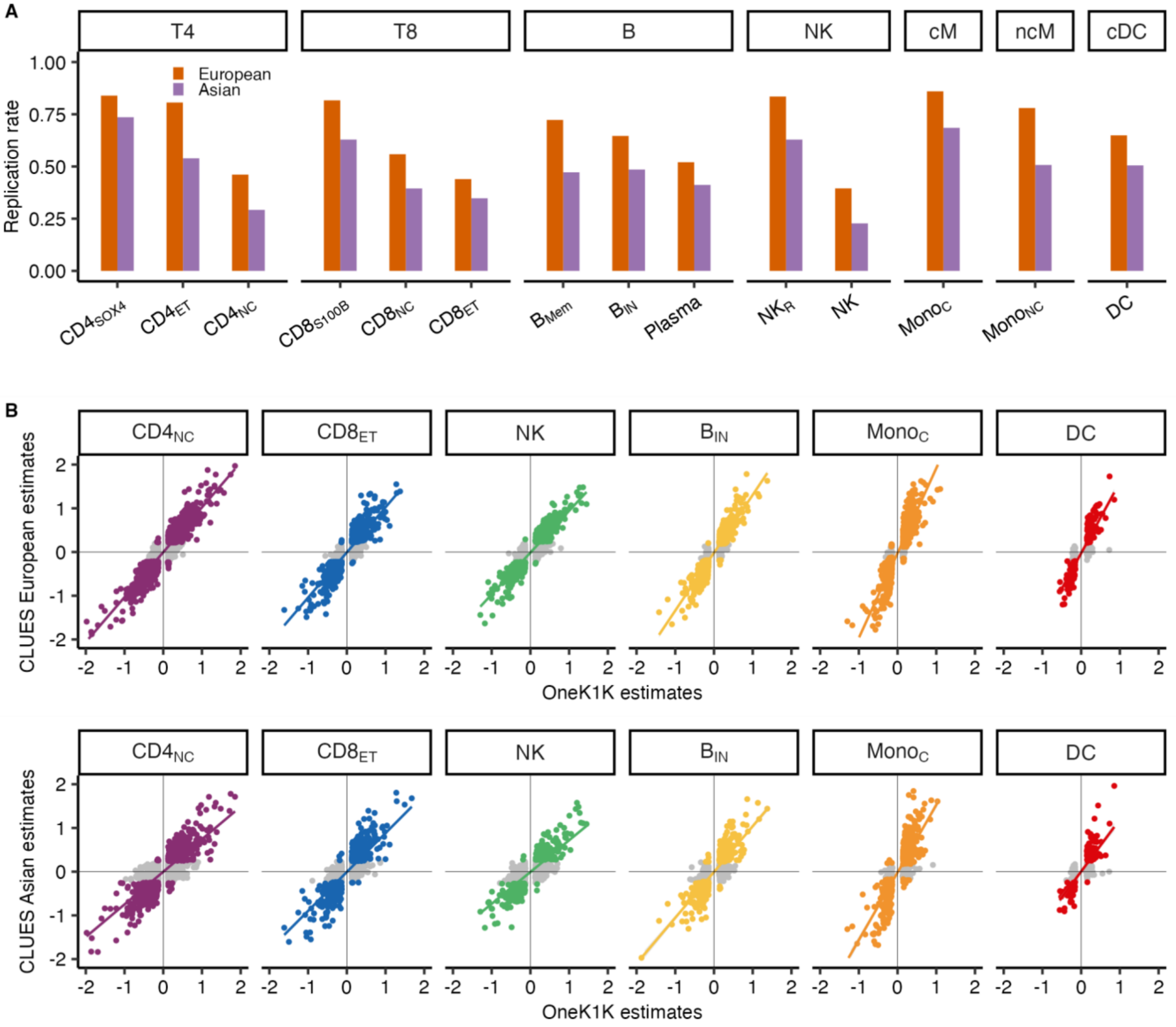


Figure 4: sc-eQTLs are more enriched in cell-type-matched CREs than bulk-eQTLs

We performed enrichment analysis of sc-eQTLs and bulk-eQTLs using their fine-mapping results and diverse annotations. **(A)** We report the odds ratio for eQTLs enrichment within 12 S-LDSC representative baseline annotations (see **Methods**). **(B)** For fine-mapped sc-eQTLs with PIP≥0.5, we report the fraction of fine-mapped eQTLs falling in three distance to TSS bins. **(C)** We report the enrichment in sc-eQTLs per cell type and bulk-eQTLs within 3 types of cell-type-specific functional annotations; we report meta-analysis results across 14 cell types at the bottom of each dataset. Error bars represent 95% CIs. Numerical results are reported in **Tables S11-14**.

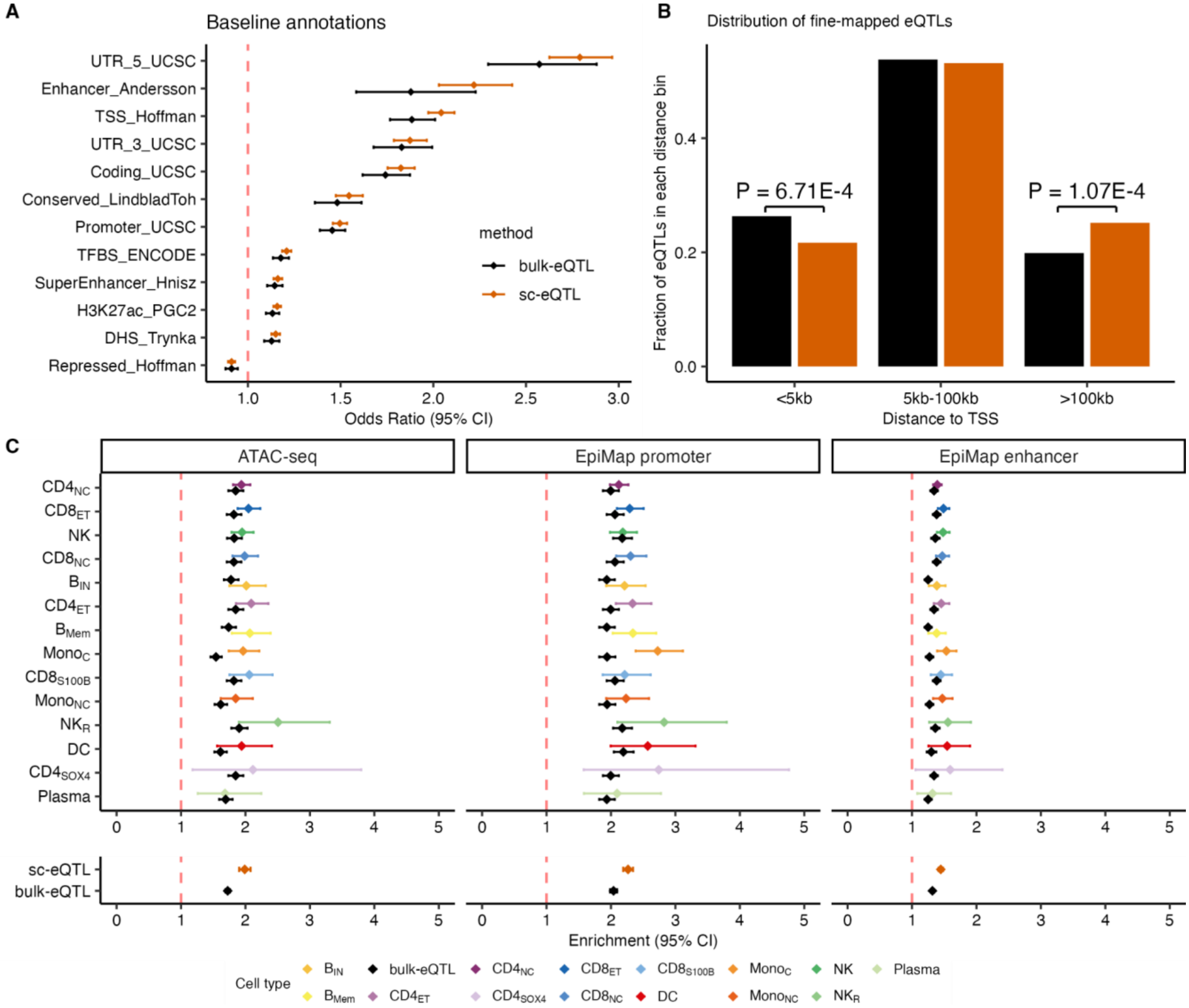


Figure 5: sc-eQTL location predicts cell-type specificity

We investigated sc-eQTL sharing across cell types by performing *mashr* analysis [58] on 2,012 OneK1K sc-eQTLs fine-mapped in at least one cell type; as a baseline, we also investigated sc-eQTL sharing by simple counting approach. **(A)** We report the fraction of sc-eQTL shared across different numbers of cell types using three approaches: a simple counting approach, *mashr* estimates of sharing by magnitude (i.e., the magnitude of effect size is within a factor of 2 compared to the strongest signal), *mashr* estimates of sharing by sign (i.e. the sign of effect size is shared with this discovery cell type, i.e., the cell type with the strongest signal). **(B)** We report the distance to TSS for sc-eQTLs identified as cell-type-specific or shared in at least 2 cell types using a simple counting approach and *mashr*. The median value of distances is displayed as a band inside each box; boxes denote values in the second and third quartiles; the length of each whisker is 1.5 times the interquartile range, defined as the height of each box. Numerical results are reported in **Tables S15**.

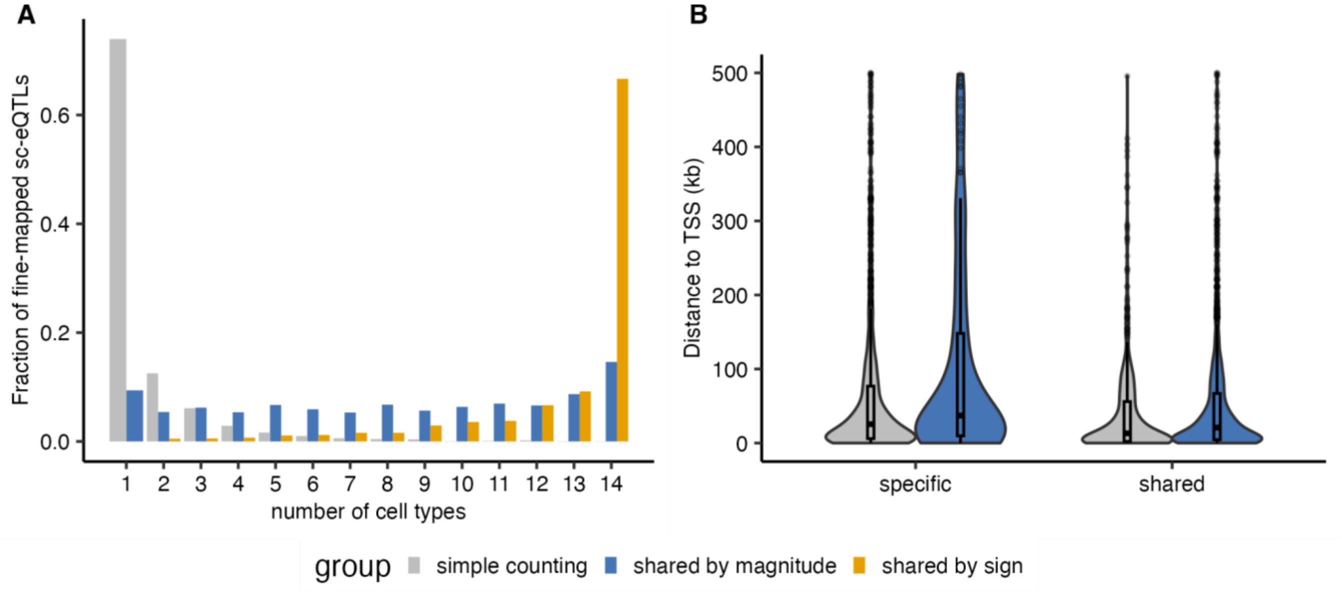


Figure 6: sc-eQTLs explain more heritability than bulk-eQTLs for immune-related GWAS traits.

(A) We report the proportion of heritability (h^2) explained by SNP-annotations built from the union of sc-eQTLs and bulk-eQTLs for 16 GWAS blood and immune-related diseases. Error bars denote 1 standard error of the corresponding estimates. **(B)** We report S-LDSC standardized effect size (τ^*) and its associated *P* values obtained for 13 traits and 12 eQTL annotations; τ^* represents the proportionate change in per-SNP h^2 associated with 1 standard deviation change of annotation value after conditioning to baseline SNP-annotations. Only trait-annotation pairs with significant τ^* (after FDR correction) were plotted. The size of the dot is proportional to the standardized effect size τ^*, and the colorness of the dot is proportional to -log$_{10}$(P). Numerical results are reported in **Tables S16, S17**.

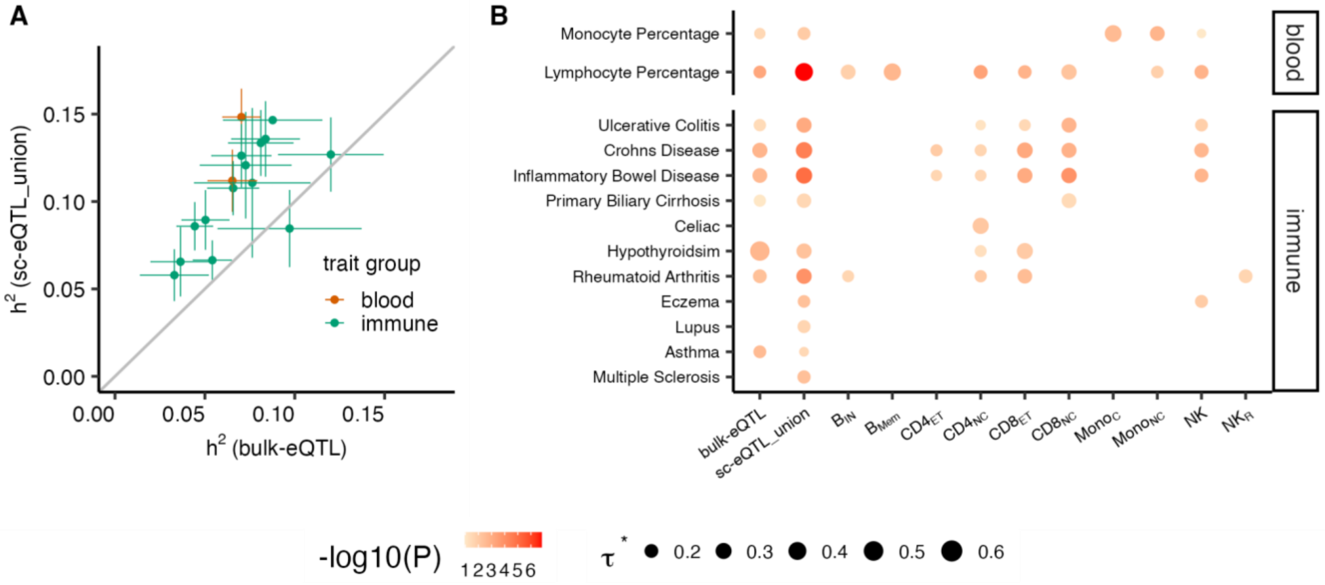
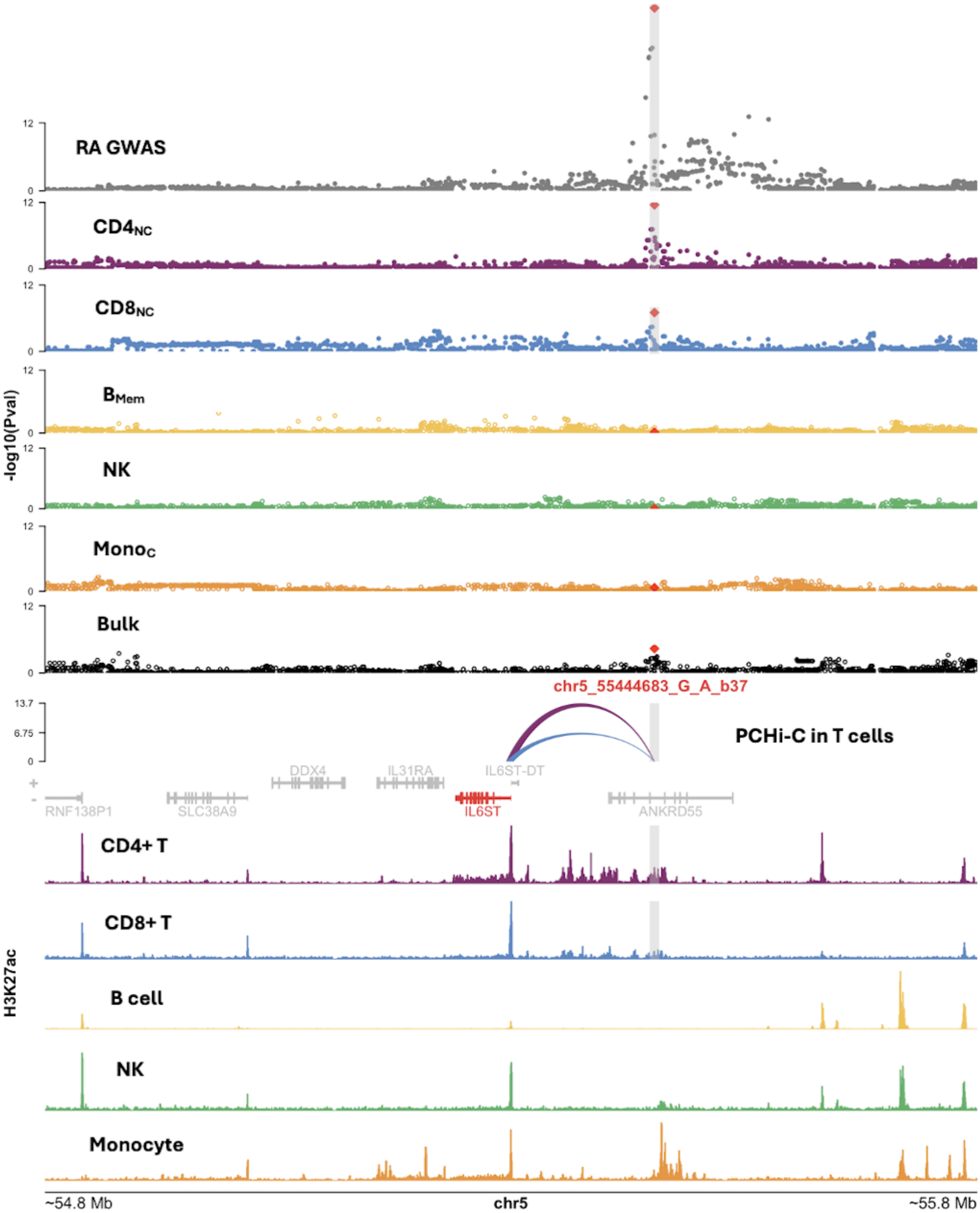


Figure 7: sc-eQTLs prioritize candidate genes missed by bulk-eQTLs.

We present an example showing that OneK1K sc-eQTLs in $CD4_{NC}$ and $CD8_{NC}$ can nominate *IL6ST* as a candidate gene for RA, while OneK1K bulk-eQTLs cannot. We report RA GWAS results at the locus highlighting GWAS leading SNP rs7731626 (chr5:55444683:G>A) (1st row), significant *IL6ST* sc-eQTLs in $CD4_{NC}$ and $CD8_{NC}$ T cells (2nd and 3rd row), non-significant (represented by open circle) sc-eQTLs in B_{Mem}, NK and $Mono_{C}$ cells (4th to 6th row), non-significant bulk-eQTLs (7th row), PCHi-C links [73] between rs7731626 loci and TSS of *IL6ST* observed in CD4+ and CD8+ T cells (height of the arch is proportional to the the score of the link; 8th row), and H3K27ac peaks observed in ENCODE samples corresponding to 5 cell types [74,75] (height of the bar is proportional to the peak intensity; 9th to 13th row). In Manhattan plots, we report -$\log_{10}(P)$ of all SNPs within ± 500kb from TSS of the *IL6ST* gene. Different colors were used to represent matching cell types. The grey shade represent ±5kb away from rs7731626 SNP in RA GWAS and plots related to T cells.


RA GWAS
CD4NC
CD8NC
BMem
NK
MonoC
Bulk
-log10(Pval)
12
0
chr5_55444683_G_A_b37
13.7
6.75
0
PCHi-C in T cells
+
-
DDX4
IL31RA
IL6ST-DT
RNF138P1
SLC38A9
IL6ST
ANKRD55
H3K27ac
CD4+ T
CD8+ T
B cell
NK
Monocyte
~54.8 Mb
chr5
~55.8 Mb

778

References

1. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.

2. Lee, D., Williamson, V.S., Bigdeli, T.B., Riley, B.P., Fanous, A.H., Vladimirov, V.I., and Bacanu, S.-A. (2015). JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. Bioinformatics *31*, 1176–1182.

3. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. Am. J. Hum. Genet. *99*, 1245–1260.

4. Chun, S., Casparino, A., Patsopoulos, N.A., Croteau-Chonka, D.C., Raby, B.A., De Jager, P.L., Sunyaev, S.R., and Cotsapas, C. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat. Genet. *49*, 600–605.

5. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.

6. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.

7. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. Am. J. Hum. Genet. *100*, 473–487.

8. Mancuso, N., Freund, M.K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., and Pasaniuc, B. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. Nat. Genet. *51*, 675–682.

9. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nat. Genet. *51*, 592–599.

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

11. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., 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.

12. 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.

13. Lappalainen, T., Sammeth, M., Friedländer, M.R., ’t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome

sequencing uncovers functional variation in humans. Nature *501*, 506–511.

14. Taylor, D.L., Jackson, A.U., Narisu, N., Hemani, G., Erdos, M.R., Chines, P.S., Swift, A., Idol, J., Didion, J.P., Welch, R.P., et al. (2019). Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. Proc. Natl. Acad. Sci. U. S. A. *116*, 10883–10888.

15. Mostafavi, H., Spence, J.P., Naqvi, S., and Pritchard, J.K. (2023). Systematic differences in discovery of genetic effects on gene expression and complex traits. Nat. Genet.

16. Umans, B.D., Battle, A., and Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? Trends Genet. *37*, 109–124.

17. 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*,.

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

19. Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M.G., Andersen, S., Lu, Q., Rowson, A., Taylor, T.R.P., Clarke, L., et al. (2022). Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. Science *376*, eabf3041.

20. Emani, P.S., Liu, J.J., Clarke, D., Jensen, M., Warrell, J., Gupta, C., Meng, R., Lee, C.Y., Xu, S., Dursun, C., et al. (2024). Single-cell genomics and regulatory networks for 388 human brains. Science *384*, eadi5199.

21. The Garvan Institute of Medical Research International partnership to map 50 million human cells and uncover genetic fingerprints of disease.

22. van der Wijst, M., de Vries, D.H., Groot, H.E., Trynka, G., Hon, C.C., Bonder, M.J., Stegle, O., Nawijn, M.C., Idaghdour, Y., van der Harst, P., et al. (2020). The single-cell eQTLGen consortium. Elife *9*, e52155.

23. van der Wijst, M.G.P., Brugge, H., de Vries, D.H., Deelen, P., Swertz, M.A., LifeLines Cohort Study, BIOS Consortium, and Franke, L. (2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat. Genet. *50*, 493–497.

24. de Vries, D.H., Matzaraki, V., Bakker, O.B., Brugge, H., Westra, H.-J., Netea, M.G., Franke, L., Kumar, V., and van der Wijst, M.G.P. (2020). Integrating GWAS with bulk and single-cell RNA-sequencing reveals a role for LY86 in the anti-Candida host response. PLoS Pathog. *16*, e1008408.

25. Oelen, R., de Vries, D.H., Brugge, H., Gordon, M.G., Vochteloo, M., single-cell eQTLGen consortium, BIOS Consortium, Ye, C.J., Westra, H.-J., Franke, L., et al. (2022). Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. Nat. Commun. *13*, 3267.

26. van Blokland, I.V., Oelen, R., Groot, H.E., Benjamins, J.W., Pekayvaz, K., Losert, C., Knottenberg, V., Heinig, M., Nicolai, L., Stark, K., et al. (2024). Single-cell dissection of the immune response after acute myocardial infarction. Circ. Genom. Precis. Med. *17*, e004374.

27. Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations.

Bioinformatics *28*, 1353–1358.

28. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T., and Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics *32*, 1479–1485.

29. Taylor-Weiner, A., Aguet, F., Haradhvala, N.J., Gosai, S., Anand, S., Kim, J., Ardlie, K., Van Allen, E.M., and Getz, G. (2019). Scaling computational genomics to millions of individuals with GPUs. Genome Biol. *20*, 228.

30. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296.

31. Ahlmann-Eltze, C., and Huber, W. (2023). Comparison of transformations for single-cell RNA-seq data. Nat. Methods 1–8.

32. Warton, D.I. (2018). Why you cannot transform your way out of trouble for small counts. Biometrics *74*, 362–368.

33. Beasley, T.M., Erickson, S., and Allison, D.B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? Behav. Genet. *39*, 580–595.

34. Nathan, A., Asgari, S., Ishigaki, K., Valencia, C., Amariuta, T., Luo, Y., Beynor, J.I., Baglaenko, Y., Suliman, S., Price, A.L., et al. (2022). Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. Nature *606*, 120–128.

35. 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.

36. Zhou, W., Cuomo, A.S.E., Xue, A., Kanai, M., Chau, G., Krishna, C., Xavier, R.J., MacArthur, D.G., Powell, J.E., Daly, M.J., et al. (2024). Efficient and accurate mixed model association tool for single-cell eQTL analysis. medRxiv 2024.05.15.24307317.

37. Mu, Z., Wei, W., Fair, B., Miao, J., Zhu, P., and Li, Y.I. (2021). The impact of cell type and context-dependent regulatory variants on human immune traits. Genome Biol. *22*, 122.

38. Radhakrishna Rao, C. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Math. Proc. Camb. Philos. Soc. *44*, 50–57.

39. Liu, Y., Chen, S., Li, Z., Morrison, A.C., Boerwinkle, E., and Lin, X. (2019). ACAT: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. Am. J. Hum. Genet. *104*, 410–421.

40. Ouwens, K.G., Jansen, R., Nivard, M.G., van Dongen, J., Frieser, M.J., Hottenga, J.-J., Arindrarto, W., Claringbould, A., van Iterson, M., Mei, H., et al. (2020). A characterization of cis- and trans-heritability of RNA-Seq-based gene expression. Eur. J. Hum. Genet. *28*, 253–263.

41. Ahlmann-Eltze, C., and Huber, W. (2021). glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. Bioinformatics *36*, 5701–5702.

42. Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640.

43. Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. Nat. Biotechnol. *38*, 147–150.

44. McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. *40*, 4288–4297.

45. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

46. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

47. Perez, R.K., Gordon, M.G., Subramaniam, M., Kim, M.C., Hartoularos, G.C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., et al. (2022). Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. Science *376*, eabf1970.

48. Schmiedel, B.J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A.G., White, B.M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J.A., McVicker, G., et al. (2018). Impact of genetic polymorphisms on human immune cell gene expression. Cell *175*, 1701–1715.e16.

49. Lu, Z., Wang, X., Carr, M., Kim, A., Gazal, S., Mohammadi, P., Wu, L., Gusev, A., Pirruccello, J., Kachuri, L., et al. (2024). Improved multi-ancestry fine-mapping identifies cis-regulatory variants underlying molecular traits and disease risk. medRxiv 2024.04.15.24305836.

50. Wang, X., and Goldstein, D.B. (2020). Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. Am. J. Hum. Genet. *106*, 215–233.

51. Glassberg, E.C., Gao, Z., Harpak, A., Lan, X., and Pritchard, J.K. (2019). Evidence for weak selective constraint on human gene expression. Genetics *211*, 757–772.

52. Mohammadi, P., Castel, S.E., Brown, A.A., and Lappalainen, T. (2017). Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. Genome Res. *27*, 1872–1884.

53. Taylor, D.J., Chhetri, S.B., Tassia, M.G., Biddanda, A., Yan, S.M., Wojcik, G.L., Battle, A., and McCoy, R.C. (2024). Sources of gene expression variation in a globally diverse human cohort. Nature *632*, 122–130.

54. Zou, Y., Carbonetto, P., Wang, G., and Stephens, M. (2022). Fine-mapping from summary data with the “Sum of Single Effects” model. PLoS Genet. *18*, e1010299.

55. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.

56. Liu, X., Finucane, H.K., Gusev, A., Bhatia, G., Gazal, S., O’Connor, L., Bulik-Sullivan, B., Wright, F.A., Sullivan, P.F., Neale, B.M., et al. (2017). Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues. Am. J. Hum. Genet. *100*, 605–616.

57. Sakaue, S., Weinand, K., Isaac, S., Dey, K.K., Jagadeesh, K., Kanai, M., Watts, G.F.M., Zhu, Z., Accelerating Medicines Partnership® RA/SLE Program and Network, Brenner, M.B., et al. (2024). Tissue-specific enhancer-gene maps from multimodal single-cell data identify causal disease alleles. Nat. Genet. *56*, 615–626.

58. Urbut, S.M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for

estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. *51*, 187–195.

59. Chen, M., and Dahl, A. (2024). A robust model for cell type-specific interindividual variation in single-cell RNA sequencing data. Nat. Commun. *15*, 5229.

60. Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol. *16*, 144–154.

61. Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.

62. Rosenberg, J., and Huang, J. (2018). CD8+ T cells and NK cells: Parallel and complementary soldiers of immunotherapy. Curr. Opin. Chem. Eng. *19*, 9–20.

63. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. *45*, 124–130.

64. Kim, A., Zhang, Z., Legros, C., Lu, Z., de Smith, A., Moore, J.E., Mancuso, N., and Gazal, S. (2024). Inferring causal cell types of human diseases and risk variants from candidate regulatory elements. medRxiv.

65. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. *50*, 621–629.

66. Jagadeesh, K.A., Dey, K.K., Montoro, D.T., Mohan, R., Gazal, S., Engreitz, J.M., Xavier, R.J., Price, A.L., and Regev, A. (2022). Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. Nat. Genet. *54*, 1479–1492.

67. Zhang, M.J., Hou, K., Dey, K.K., Sakaue, S., Jagadeesh, K.A., Weinand, K., Taychameekiatchai, A., Rao, P., Pisco, A.O., Zou, J., et al. (2022). Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. Nat. Genet. *54*, 1572–1580.

68. Müller-Ladner, U., Pap, T., Gay, R.E., Neidhart, M., and Gay, S. (2005). Mechanisms of disease: the molecular and cellular basis of joint destruction in rheumatoid arthritis. Nat. Clin. Pract. Rheumatol. *1*, 102–110.

69. Mack, M.R., Brestoff, J.R., Berrien-Elliott, M.M., Trier, A.M., Yang, T.-L.B., McCullen, M., Collins, P.L., Niu, H., Bodet, N.D., Wagner, J.A., et al. (2020). Blood natural killer cell deficiency reveals an immunotherapy strategy for atopic dermatitis. Sci. Transl. Med. *12*, eaay1005.

70. Ishigaki, K., Sakaue, S., Terao, C., Luo, Y., Sonehara, K., Yamaguchi, K., Amariuta, T., Too, C.L., 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.

71. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376–381.

72. Westra, H.-J., Martínez-Bonet, M., Onengut-Gumuscu, S., Lee, A., Luo, Y., Teslovich, N., Worthington, J., Martin, J., Huizinga, T., Klareskog, L., et al. (2018). Fine-mapping and functional

studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. Nat. Genet. *50*, 1366–1374.

73. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369–1384.e19.

74. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

75. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. *48*, D882–D889.

76. Mena, J., Alloza, I., Tulloch Navarro, R., Aldekoa, A., Díez García, J., Villanueva Etxebarria, A., Lindskog, C., Antigüedad, A., Boyero, S., Mendibe-Bilbao, M.D.M., et al. (2021). Genomic multiple sclerosis risk variants modulate the expression of the ANKRD55-IL6ST gene region in immature dendritic cells. Front. Immunol. *12*, 816930.

77. Lessard, S., Chao, M., Reis, K., Beauvais, M., Rajpal, D.K., Shankara, S., Sloane, J., Palta, P., Klinger, K., de Rinaldis, E., et al. (2023). Leveraging large-scale multi-omics to identify therapeutic targets from genome-wide association studies.

78. Benaglio, P., Newsome, J., Han, J.Y., Chiou, J., Aylward, A., Corban, S., Miller, M., Okino, M.-L., Kaur, J., Preissl, S., et al. (2023). Mapping genetic effects on cell type-specific chromatin accessibility and annotating complex immune trait variants using single nucleus ATAC-seq in peripheral blood. PLoS Genet. *19*, e1010759.

79. Jang, S., Kwon, E.-J., and Lee, J.J. (2022). Rheumatoid arthritis: Pathogenic roles of diverse immune cells. Int. J. Mol. Sci. *23*, 905.

80. Pearce, E.L., Mullen, A.C., Martins, G.A., Krawczyk, C.M., Hutchins, A.S., Zediak, V.P., Banica, M., DiCioccio, C.B., Gross, D.A., Mao, C.-A., et al. (2003). Control of effector CD8+ T cell function by the transcription factor Eomesodermin. Science *302*, 1041–1043.

81. Elgueta, R., Benson, M.J., de Vries, V.C., Wasiuk, A., Guo, Y., and Noelle, R.J. (2009). Molecular mechanism and function of CD40/CD40L engagement in the immune system. Immunol. Rev. *229*, 152–172.

82. Li, G., Diogo, D., Wu, D., Spoonamore, J., Dancik, V., Franke, L., Kurreeman, F., Rossin, E.J., Duclos, G., Hartland, C., et al. (2013). Human genetics in rheumatoid arthritis guides a high-throughput drug screen of the CD40 signaling pathway. PLoS Genet. *9*, e1003487.

83. Rose-John, S. (2018). Interleukin-6 family cytokines. Cold Spring Harb. Perspect. Biol. *10*, a028415.

84. Ernst, M., and Jenkins, B.J. (2004). Acquiring signalling specificity from the cytokine receptor gp130. Trends Genet. *20*, 23–32.

85. Rose-John, S., Jenkins, B.J., Garbers, C., Moll, J.M., and Scheller, J. (2023). Targeting IL-6 trans-signalling: past, present and future prospects. Nat. Rev. Immunol. *23*, 666–681.

86. Weeks, E.M., Ulirsch, J.C., Cheng, N.Y., Trippe, B.L., Fine, R.S., Miao, J., Patwardhan, T.A., Kanai,

M., Nasser, J., Fulco, C.P., et al. (2023). Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. Nat. Genet. 1–10.

87. Mu, Z., Randolph, H.E., Aguirre-Gamboa, R., Ketter, E., Dumaine, A., Locher, V., Brandolino, C., Liu, X., Kaufmann, D.E., Barreiro, L.B., et al. (2024). Impact of disease-associated chromatin accessibility QTLs across immune cell types and contexts.

88. Gutierrez-Arcelus, M., Baglaenko, Y., Arora, J., Hannes, S., Luo, Y., Amariuta, T., Teslovich, N., Rao, D.A., Ermann, J., Jonsson, A.H., et al. (2020). Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. Nat. Genet. *52*, 247–253.

89. The Developmental Genotype-Tissue Expression (dGTEx) project.

90. Tian, C., Zhang, Y., Tong, Y., Kock, K.H., Sim, D.Y., Liu, F., Dong, J., Jing, Z., Wang, W., Gao, J., et al. (2024). Single-cell RNA sequencing of peripheral blood links cell-type-specific regulation of splicing to autoimmune and inflammatory diseases. Nat. Genet. *56*, 2739–2752.

91. Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H.K., Ju, C.J.-T., Loh, P.-R., Schoech, A., Reshef, Y., Liu, X., O'Connor, L., et al. (2018). Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. Nat. Genet. *50*, 1041–1047.

92. Wang, J., Zhang, Z., Lu, Z., Mancuso, N., and Gazal, S. (2024). Genes with differential expression across ancestries are enriched in ancestry-specific disease effects likely due to gene-by-environment interactions. Am. J. Hum. Genet.

93. Hilbe, J.M. (2011). Negative binomial regression. In Negative Binomial Regression, (Cambridge: Cambridge University Press), pp. 185–220.

94. Engle, R.F. (1984). Chapter 13 Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In Handbook of Econometrics, (Elsevier), pp. 775–826.

95. Lin, W., Schmidt, M., and Khan, M.E. (2020). Handling the Positive-Definite Constraint in the Bayesian Learning Rule.

96. Storey, J.D., Bass, A.J., Dabney, A., and Robinson, D. (2015). qvalue: Q-value estimation for false discovery rate control. R Package Version *2*, 10–18129.

97. McCaw, Z.R., Lane, J.M., Saxena, R., Redline, S., and Lin, X. (2020). Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. Biometrics *76*, 1262–1272.

98. Kerimov, N., Hayhurst, J.D., Peikova, K., Manning, J.R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M.P., Kuzmin, I., Trevanion, S.J., et al. (2021). A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat. Genet. *53*, 1290–1299.

99. Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G., et al. (2005). A common inversion under selection in Europeans. Nat. Genet. *37*, 129–137.

100. Zody, M.C., Jiang, Z., Fung, H.-C., Antonacci, F., Hillier, L.W., Cardone, M.F., Graves, T.A., Kidd, J.M., Cheng, Z., Abouelleil, A., et al. (2008). Evolutionary toggling of the MAPT 17q21.31 inversion region. Nat. Genet. *40*, 1076–1083.

101. Chiou, J., Geusz, R.J., Okino, M.-L., Han, J.Y., Miller, M., Melton, R., Beebe, E., Benaglio, P., Huang, S., Korgaonkar, K., et al. (2021). Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. Nature *594*, 398–402.

102. Wen, X. (2016). Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. Aoas *10*, 1619–1638.

103. Gazal, S., Marquez-Luna, C., Finucane, H.K., and Price, A.L. (2019). Reconciling S-LDSC and LDAK functional enrichment estimates. Nat. Genet. *51*, 1202–1204.

104. van de Geijn, B., Finucane, H., Gazal, S., Hormozdiari, F., Amariuta, T., Liu, X., Gusev, A., Loh, P.-R., Reshef, Y., Kichaev, G., et al. (2020). Annotations capturing cell type-specific TF binding explain a large fraction of disease heritability. Hum. Mol. Genet. *29*, 1057–1067.