Scalable unified framework of total and allele-specific counts for cis-QTL, fine-mapping, and prediction

Yanyu Liang[1,*] François Aguet[2] Alvaro Barbeira[1] Kristin Ardlie[2]
Hae Kyung Im[1,*]

April 22, 2020

1 Section of Genetic Medicine, The University of Chicago, Chicago, IL, USA

2 The Broad Institute of MIT and Harvard, Cambridge, MA, USA

***** Correspondence to yanyul@uchicago.edu and haky@uchicago.edu

Abstract

Genome-wide association studies (GWAS) have been highly successful in identifying genomic loci associated with complex traits. However, identification of the causal genes that mediate these associations remains challenging, and many approaches integrating transcriptomic data with GWAS have been proposed. However, there currently exist no computationally scalable methods that integrate total and allele-specific gene expression to maximize power to detect genetic effects on gene expression. Here, we describe a unified framework that is scalable to studies with thousands of samples. Using simulations and data from GTEx, we demonstrate an average power gain equivalent to a 29% increase in sample size for genes with sufficient allele-specific read coverage. We provide a suite of freely available tools, mixQTL, mixFine, and mixPred, that apply this framework for mapping of quantitative trait loci, fine-mapping, and prediction.

1 Introduction

Genome-wide association studies (GWAS) have identified tens of thousands of genomic loci associated with complex traits. A large majority of these loci lie in non-coding regions of the genome, which hinders identification of the underlying molecular mechanisms and causal genes. Multiple methods have been developed to integrate GWAS results with expression quantitatite trait loci (eQTLs), to test whether complex trait associations are mediated through regulation of gene expression. Two strategies are commonly employed: 1) association-based approaches including PrediXcan [Gamazon et al., 2015], fusion [Gusev et al., 2016], and smr [Zhu et al., 2016]; and 2) colocalization-based approaches including coloc [Giambartolomei et al., 2014], eCAVIAR [Hormozdiari et al., 2016], and enloc [Wen et al., 2017]. These approaches rely on high-quality eQTL mapping, fine-mapping, and gene expression predictions.

In cis-eQTL analysis, allele-specific expression (ASE), i.e., the relative expression difference between the two haplotypes, captures the genetic effect of nearby variants. ASE provides additional signal to total read count, and several methods have been proposed to combine total and allele-specific read count for QTL mapping, such as TReCASE [Sun, 2012], WASP [Van De Geijn et al., 2015], and RASQUAL [Kumasaka et al., 2016]). However, these methods are computationally too costly to be applied to sample sizes beyond a few hundred and as a result have not been applied to large-scale studies like GTEx, which includes over 17,000 samples across 49 tissues. Recently, two fine-mapping approaches have been proposed utilizing effect size estimates obtained from both ASE and eQTL mapping via meta-analysis [Zou et al., 2019; Wang et al., 2020]. However, no existing methods, to our knowledge, provides a unified framework of total and allele-specific counts with explicit multi-SNP modeling for QTL mapping, fine-mapping, and prediction.

By assuming a log-linear model for transcript expression levels with independent reads from each haplotype and weak genetic effects, as proposed by [Mohammadi et al., 2017], we derive two approximately independent equations for allelic imbalance (read count difference between the two haplotypes) and total read count. This enables us to develop computationally efficient algorithms for cis-QTL mapping, fine-mapping, and prediction. We demonstrate the resulting gain in performance through simulations under a range of different settings, applications to GTEx v8 data [Aguet et al., 2019], and comparisons to a large-scale eQTL meta-analysis from eQTLGen [Võsa et al., 2018].

The software, simulation, data preprocessing, and analysis pipeline can be found at https://github.com/hakyimlab/mixqtl and https://github.com/liangyy/mixqtl-pipeline. A computationally efficient GPU-based implementation of mixQTL has been embedded in tensorQTL https://github.com/broadinstitute/tensorqtl.

2 Results

Overview of the statistical model

To develop a computationally efficient approach that integrates total and allele-specific count data, we assumed multiplicative cis-regulatory effects and noise, similarly to the model proposed in [Mohammadi et al., 2017]. For a given gene, we modeled the haplotypic read count Y_i^h, which is the number of reads from haplotype h of individual i as

$$Y_i^h = L_i \cdot \theta_{0,i} \cdot \exp(\beta \cdot X_i^h) \cdot \exp(\epsilon_i^h), \tag{1}$$

where L_i is the library size for individual i, $\theta_{0,i}$ is the baseline abundance (for a haplotype with the reference allele), $\exp(\beta)$ is the cis-regulatory effect (allelic fold change due to the presence of the alternative allele), X_i^h indicates the dosage of the affecting variant (0 if the individual has the reference allele, and 1 if they have the alternative one), and $\exp(\epsilon_i^h)$ is the multiplicative noise.

Calculating the total read count as the sum of the two haplotypic counts and assuming weak cis-regulatory effects, we derived an approximately linear model for the logorithm of the haplotypic and total read counts (see details in Methods and Supplementary Note 7). In practice, we only observe the allele-specific reads that include a heterozygous site, which is a fraction of the total haplotypic count denoted as $Y_i^{(h)\text{obs}} = \alpha_i \cdot Y_i^h$. To take this partial readout into account, we modeled the observed total and allele-specific counts as

$$\begin{aligned}
\log Y_i^{(1)\text{obs}} &= \log L_i + \log \alpha_i + \log \theta_{0,i} + X_i^1 \beta + \epsilon_i^{(1)} \\
\log Y_i^{(2)\text{obs}} &= \log L_i + \log \alpha_i + \log \theta_{0,i} + X_i^2 \beta + \epsilon_i^{(2)} \\
\log \frac{Y_i^{\text{total}}}{2} &\approx \log L_i + \log \theta_{0,i} + \frac{X_i^1 + X_i^2}{2} \beta + \epsilon_i^{\text{trc}}
\end{aligned} \tag{2}$$

where the error terms are $\epsilon_i^{\text{trc}} \sim N(0, \frac{\sigma^2}{Y_i^{\text{total}}})$, $\epsilon_i^{(h)} \sim N(0, \frac{\sigma^2}{Y_i^{(h)\text{obs}}})$ and the errors of the two haplotypes are independent: $\epsilon^{(1)} \perp\!\!\!\perp \epsilon^{(2)}$.

We further simplified the models by combining the two allele-specific counts, defining the baseline abundance variation as a random effect z_i ($\log \theta_{0,i} =$ population mean $+ z_i$), and dropping ϵ_i^{trc}

from the total count since this "techinical" noise which scales as the inverse of the read count is small compared to the "biological" variability, z_i, (See Methods section and Supplementary Note 10.1) to obtain our final model

$$\log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}} = \qquad (X_i^1 - X_i^2)\beta + \epsilon_i^{\text{asc}} \text{ (allelic imbalance eq.)} \tag{3}$$

$$\log \frac{Y_i^{\text{total}}}{2L_i} \approx \mu_0 + \frac{X_i^1 + X_i^2}{2}\beta \quad + z_i \quad \text{(total read count eq.)} \tag{4}$$

where $z_i \sim N(0, \sigma_0^2)$ and $\epsilon_i^{\text{asc}} \sim N(0, \sigma^2 \cdot (\frac{1}{Y_i^{(1)\text{obs}}} + \frac{1}{Y_i^{(2)\text{obs}}}))$ and $z_i \perp\!\!\!\perp \epsilon^{\text{asc}}$ (baseline abundance is independent of the multiplicative error).

This single SNP model extends to multiple SNPs in a straightforward manner by using a vector of allelic dosages $(X_1, \cdots, X_p)$ and genetic effects $(\beta_1, \cdots, \beta_p)$ instead of the scalar values above. Here, p represents the number of genetic variants in the cis-window of the gene under consideration (Supplementary Notes 9 and 11).

For cis-QTL mapping, we took advantage of the approximate independence of the allelic-imbalance and the total read counts in equations (3) and (4), solving them as separate linear regressions (for computational efficiency) and combining the results via inverse-variance weighted meta-analysis. We call this method mixQTL.

For the fine-mapping and prediction problems, we also leveraged the approximate independence of the allelic-imbalance and total read count equations. We used a two-step approach in which we first scale the two equations so that they become independent data points with equal variances. In a second step, we combined these data points into an augmented dataset and applied the existing algorithms SuSiE [Wang et al., 2019] and elastic net [Friedman et al., 2010]. We term these methods mixFine and mixPred, for fine-mapping and prediction, respectively.

Simulation of total and allele-specific reads

To assess the benefits of this unified framework relative to using total read count or allele-specific expression only, we simulated haplotypic reads according to the framework illustrated in Figure 1, with additional details in Methods (6.7) and Supplementary Notes 12.

For all simulation settings, we set an average library size of 94 million reads (to match closely to GTEx v8 library size) and used two expression levels (expected value of $\theta_{0,i}$ in Eq 1): 10 and

1 read per million, corresponding to $\theta = 10^{-5}$ and 10^{-6}. The fraction of allele-specific reads was kept at the similar levels across simulations by using the same distribution of polymorphic sites per individual.

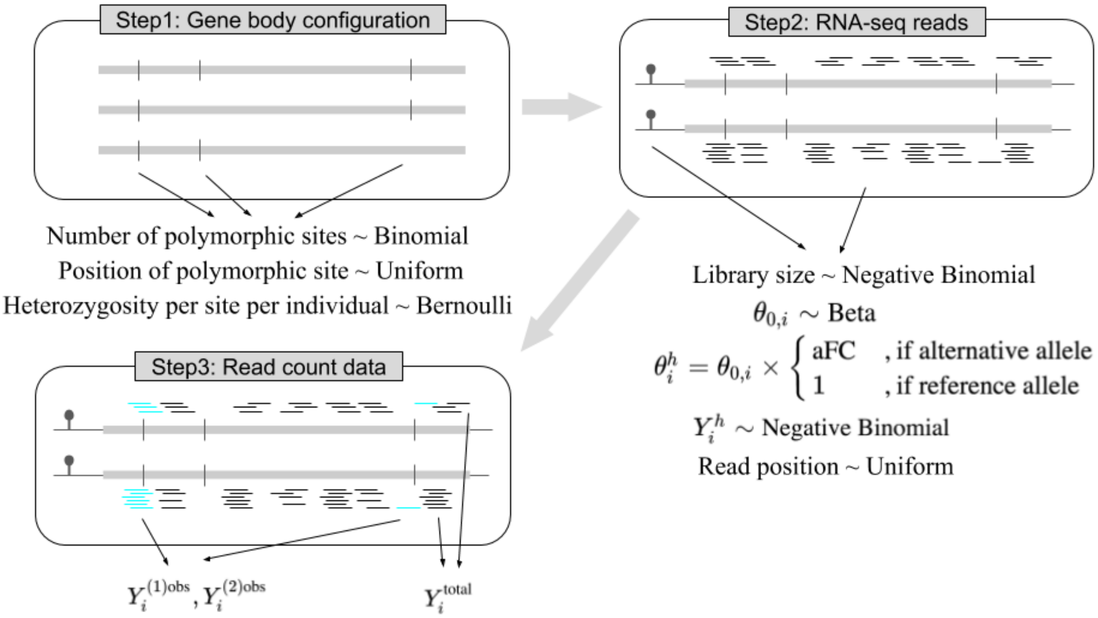

Step1: Gene body configuration
Number of polymorphic sites ~ Binomial
Position of polymorphic site ~ Uniform
Heterozygosity per site per individual ~ Bernoulli
Step2: RNA-seq reads
Library size ~ Negative Binomial
$\theta_{0,i} \sim$ Beta
$\theta_i^h = \theta_{0,i} \times \begin{cases} \text{aFC} & \text{, if alternative allele} \\ 1 & \text{, if reference allele} \end{cases}$
$Y_i^h \sim$ Negative Binomial
Read position ~ Uniform
Step3: Read count data
$Y_i^{(1)\text{obs}}, Y_i^{(2)\text{obs}}$
Y_i^{total}

Figure 1: Simulation scheme for total and allele-specific read counts. Step 1 simulates a gene body configuration by first simulating the number of polymorphic sites of the gene followed by positioning these polymorphic sites uniformly across the gene body. For each individual, the actual heterozygosity of these polymorphic sites are drawn from Bernoulli distribution. Step 2 simulates the haplotypic reads by first simulating Negative Binomial library size L_i, Beta baseline abundance $\theta_{0,i}$, and the genetic effect β. These parameters determine the expected count for each transcript. Then, the actual haplotypic read count Y_i^h is generated using a Negative Binomial distribution given the expected count where the reads are distributed uniformly across the gene body. In Step 3, the gene-level allele-specific counts $Y_i^{(h)\text{obs}}$ are determined by counting the reads that overlap heterozygous sites, in which aFC is the allelic fold change which equals to e^β in our parameterization. For convenience, we used natural log rather than base 2 log. Y_i^{total} is calculated as the sum of the two haplotypic counts Y_i^1 and Y_i^2.

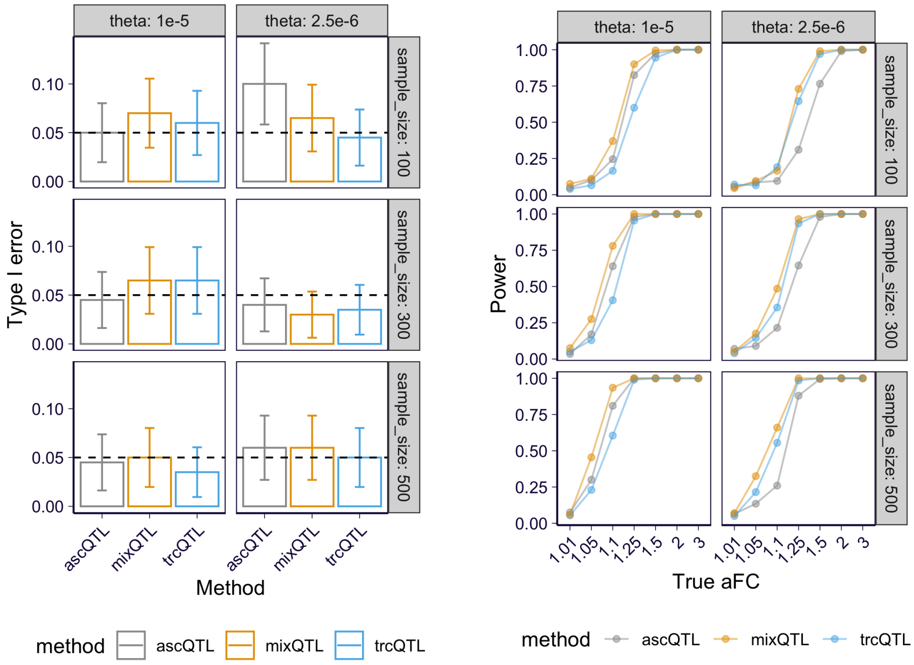
To compare the computational cost of mixQTL to RASQUAL and WASP, we tested their performance on simulated data with 100 samples. As shown in Supplementary Figure S4, type I error and power were similar for all three methods and mixQTL was 10 to 43 times as fast as the others.

Combining total and allele-specific read counts improves cis-eQTL mapping

To assess the gain in power of combining total and allele-specific read counts, we simulated 200 replicates with allelic fold change varying among 1, 1.01, 1.05, 1.1, 1.25, 1.5, 2, 3. We compared mixQTL with two methods: using either only allele-specific counts (ascQTL) or total counts (trcQTL). See details in Supplementary Note 10.1.

All three methods had calibrated type I errors (Figures 2A and S1). And mixQTL outperformed both trcQTL and ascQTL in all simulation settings, demonstrating the benefits of combining total and allele-specific counts in cis-eQTL mapping (Figures 2B and S2).

The degree of improvement varied with the number of reads and sample size. The power of ascQTL was sensitive to the number of allele-specific reads, as expected. As shown in Figure 2B, ascQTL yielded much higher power in the case of relatively large θ (on the left) compared with small θ (on the right). In contrast, trcQTL was less sensitive to the number of reads observed under the range of read counts in our simulation settings. Such sensitivity differences between ascQTL and trcQTL are consistent with the nature of count data, where the magnitude of the noise is inversely related to the count.



(A) (B)

Figure 2: QTL mapping performance for mixQTL and approaches based on either total reads (trcQTL) or allele-specific reads (ascQTL) on simulated data. Each panel presents the results for two relative abundances of the gene, θ, and three sample sizes. **(A)** Type I error (y-axis) at a 5% significance level across methods (x-axis). The dashed line represents the desired error rate under the null hypothesis. The error bar indicates the 95% confidence interval (CI) in observed error rate estimated from 200 replicates. **(B)** Power (y-axis) at a 5% significance level across methods under a range of true aFC values (x-axis). Power is defined as the fraction of eQTLs passing the significance threshold.

Combining total and allele-specific read count improves fine-mapping

To mimic LD structure realistically in our simulations, we used the genotypes of European individuals from the 1000 Genomes projects phase 3 [1000 Genomes Project Consortium, 2015] within 1MB cis-windows of 100 randomly selected genes. We applied mixFine and trcFine (which uses total read count only; Supplementary Notes 11.3) to the simulated data and characterized the fine-mapping results with two metrics: 1) power curve, defined as the proportion of detected variants among causal ones versus the number of detected variants, where detection means the variant has posterior inclusion probability (PIP) > some threshold (which is varied to get the desired number of detected SNPs); 2) the size of 95% credible set (CS) which contains the causal variant.

The PIP of both trcFine and mixFine were consistent with the proportion of true causal variants within each bin of 0.1 length (Figure 3A). By combining total and allele-specific reads, mixFine achieved higher power than trcFine (Figures 3B and S5) across all simulation settings. mixFine achieved the highest improvement relative to trcFine at high expression level, θ, corresponding to high-quality allele-specific signals. The gain in power decreased with larger sample sizes.

The increased power was also reflected in the number and size of 95% CSs containing the true signals. As shown in Figures 3C and S6, mixFine identified more true positive 95% CSs and these 95% CSs were generally smaller than the ones of trcFine demonstrating that mixFine can pinpoint causal SNPs more accurately.

Overall, the combined method was more powerful for identifying causal variants, which is consistent with recent reports [Zou et al., 2019; Wang et al., 2020].

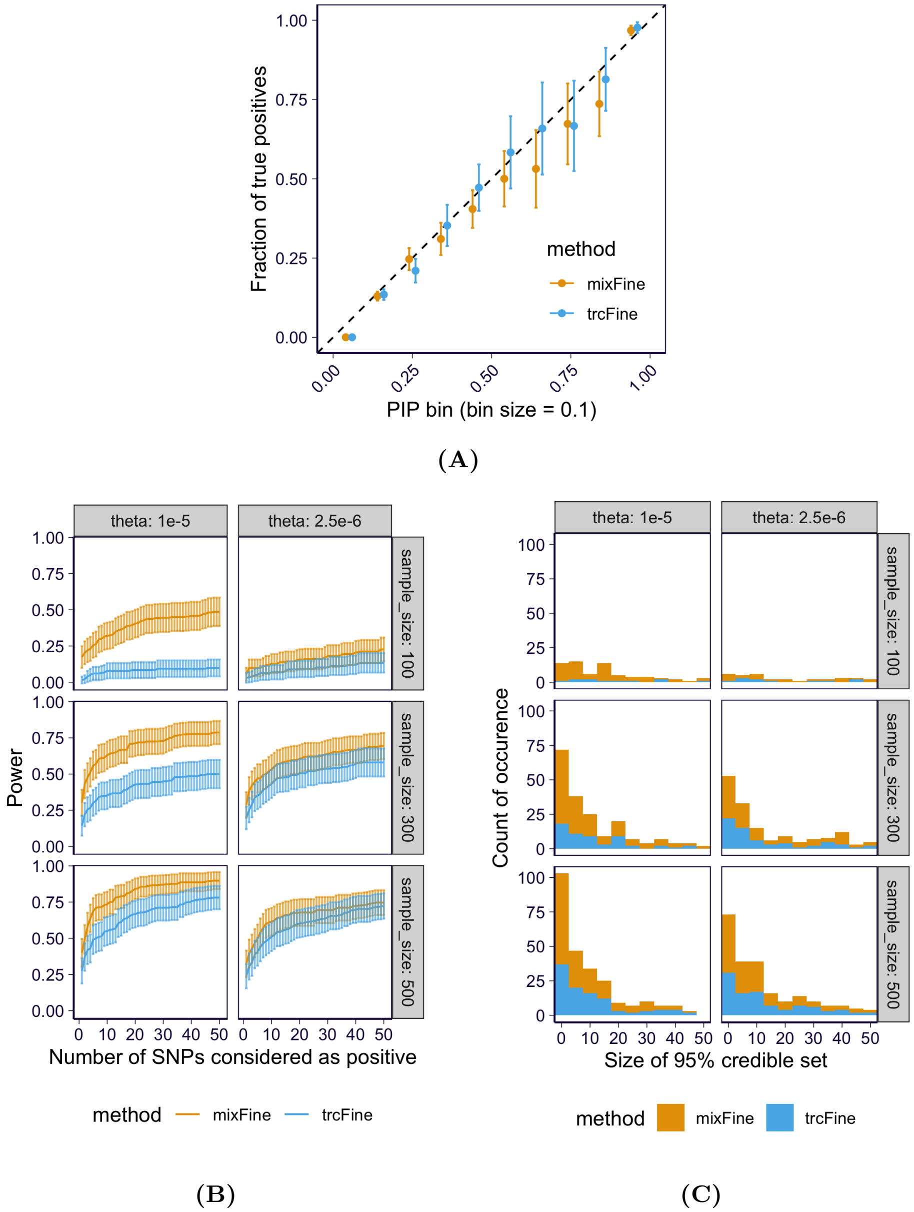
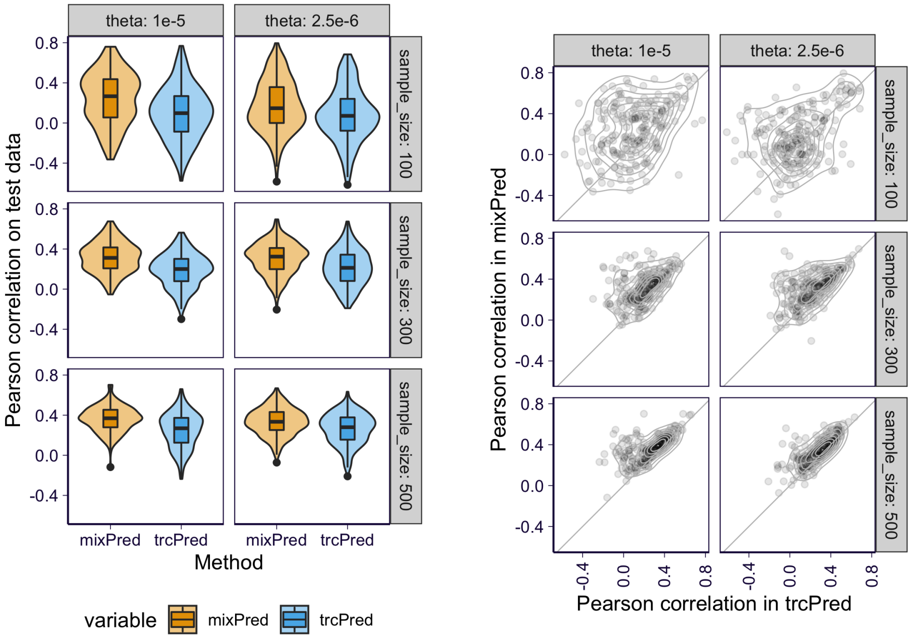


Figure 3: Fine-mapping performance of the combined (mixFine) and total read-based (trcFine) approaches on simulated data. **(A)** The observed true positive rate within SNPs binned by PIP are shown (aggregated across all simulation settings) for both mixFine (orange) and trcFine (blue). **(B)** The power at a PIP cutoff (on y-axis) is plotted against the number of variants passing the PIP cutoff (on x-axis) for mixFine and trcFine. The solid curves indicate the mean power (recall rate) among 100 simulation replicates and the error bars indicate the 95% CI. **(C)** The distribution of the size of 95% CS that contain the causal variant for mixFine and trcFine across all 100 simulation replicates. The counts in each bin are stacked.

Combining total and allele-specific read count improves prediction

Using the data from the fine-mapping simulation, we tested the performance of mixPred and trcPred (Supplementary Notes 11.3) on held-out test data. Specifically, we split each simulation replicate into training (4/5) and test (1/5) sets. We trained prediction models using training data and evaluated the prediction performance on test data using Pearson correlation between predicted and true response. For each data set, we repeated the splitting-training-evaluation procedure twice to reduce the stochasticity introduced by splitting.

Overall, mixPred achieved higher prediction accuracy than trcPred (Figure 4 and Supplementary Figure S7 and S8). The gain in performance was more apparent when the expression level θ was higher and as a consequence the allele-specific count was larger.



(A)

(B)

Figure 4: Prediction performance of the combined (mixPred) and total read-based (trcPred) methods on simulated data. (A) The overall distribution of Pearson correlations between predicted and observed total count abundance in log-scale, i.e., $\log(Y_i^{\text{total}}/L_i)$, for mixPred (orange) and trcPred (blue) across all data splits are shown. **(B)** For each split, the prediction performance of mixPred (y-axis) is plotted against the prediction performance of trcPred (x-axis).

mixQTL outperforms standard eQTL mapping in GTEx data

Next, we compared mixQTL to the standard eQTL mapping approach (denoted here simply as eQTL) used by the GTEx consortium [Aguet et al., 2019], using 670 whole blood RNA-seq samples from the v8 release. We included variants within a ±1 Mb cis-window around the transcription start site of each gene, and limited our analysis to genes passing the following two criteria: 1) at least 15 samples having at least 50 allele-specific counts for each haplotype; and 2) at least 500 samples having a total read count of at least 100. 28% of genes passed these filters, corresponding to 5,734 genes in total. For genes with below-threshold allele-specific counts, the calculation is performed using total read counts only, such that all genes considered using the standard approach are also tested in mixQTL. Performance for these genes was similar to the standard eQTL approach (Supplementary Figure S9). We then stratified genes that passed the filtering criteria by their median expression level (read counts) into low, medium, and high expression tertiles.

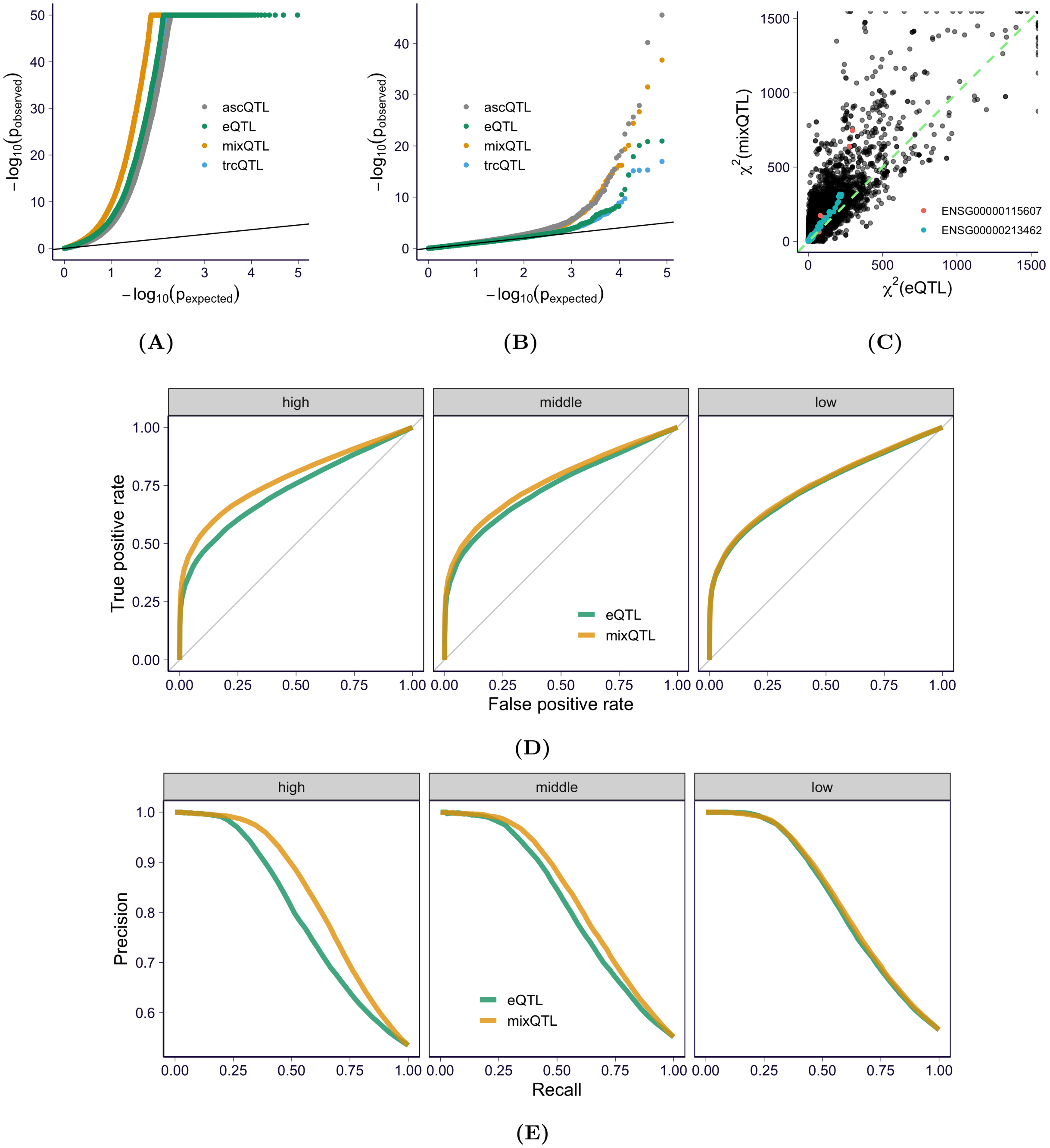
All three approaches mixQTL, aseQTL, and trcQTL were relatively well-calibrated when permuting data in four randomly selected genes (Supplementary Figure S10). The estimated effect sizes were consistent with allelic fold change estimates from the main GTEx v8 analysis (Supplementary Figure S11).

To further compare the performance of the methods, we used eQTLGen [Võsa et al., 2018], a large-scale meta-analysis of over 30,000 blood samples, as our “ground truth” eQTL discovery reference (Supplementary Notes 14). We selected a random subset of 100,000 variant/gene pairs tested by eQTLGen with FDR < 0.05 as the set of “ground truth” eQTLs. We also selected a random set 100,000 variant/gene pairs with p > 0.50 as a background set of “non-significant” eQTLs. Only 96,660 and 78,691 of the “ground truth” and “non-significant” pairs were found in the GTEx data.

For the “ground truth” eQTLs, mixQTL yielded more significant p-values compared to the standard eQTL, ascQTL, and trcQTL approaches (Fig. 5). The “non-significant” variant/gene pairs showed moderate enrichment for small p-values for all methods (Figure 5B), likely reflecting a combination of false negatives in eQTLGen and potential false positives in our analysis. Overall, we found that mixQTL achieves increased power compared to standard eQTL mapping on real data for the set of genes with sufficient total and allele-specific read counts.

As an intuitive measure of improved performance, we estimated the effective sample size gain of mixQTL compared to standard eQTL mapping as the median of the ratio between mixQTL χ^2 statistics and eQTL χ^2 statistics. mixQTL showed a 29% increase in effective sample size compared to the standard eQTL mapping approach (Figure 5C).

To account for the trade-off between true and false positive rates, as well as between precision and power, we used receiver operating characteristic (ROC) and precision-recall (PR) curves to compare the performance of mixQTL and standard eQTL approaches using the eQTLGen "ground truth" and "non-significant" eQTLs. We found that mixQTL achieves higher performance in both ROC (Figure 5D) and PR curves (Figure 5E). Consistent with simulation results, this gain is more significant for genes with higher expression levels.



(A)

(B)

(C)

(D)

(E)

Figure 5: Performance of mixQTL on GTEx v8 whole blood RNA-seq. (A) QQ-plot of nominal p-values for a random subset of cis-eQTLs (FDR < 0.05) reported in eQTLGen. **(B)** QQ-plot of nominal p-values for a random subset of variant/gene pairs with p-value > 0.5 in eQTLGen. **(C)** χ^2 statistics from eQTL analysis (x-axis) and mixQTL analysis (y-axis) among a random subset of cis-eQTLs (FDR $<$ 0.05) reported in eQTLGen. The slope indicates the relative effective sample size increase. Two randomly selected genes are highlighted in red and green, respectively. **(D, E)** ROC and PR curves for mixQTL and the standard eQTL method measured in eQTLGen. Each panel shows the results of genes stratified by expression level tertiles.

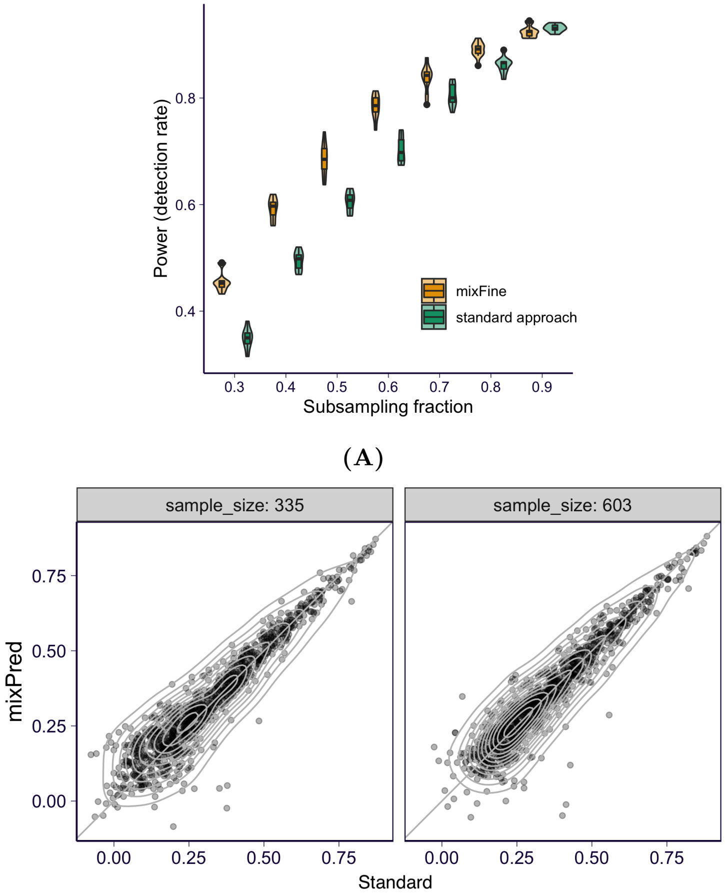
Fine-mapping and prediction model building in GTEx data

We applied mixFine to the GTEx v8 whole blood RNA-seq data, using the same subset of genes with high expression and allelic counts that were used for QTL mapping above. We compared mixFine to the SuSiE fine-mapping approach [Wang et al., 2019] applied to the inverse normal transformed expression values used for standard eQTL mapping [Aguet et al., 2019]. We corrected for sex, 5 genetic principal components, WGS platform, WGS library prep protocol (PCR), and 60 PEER factors. We refer to the latter as the "standard approach" below for simplicity.

To compare the power of causal variant detection, we performed a subsampling analysis on a random subset of 1,000 genes. First, we defined "consensus SNPs" as the variants with PIP > 0.5 in both mixFine and the "standard approach" using all samples. Similarly, a variant was defined as "top SNP" if it was the most significant variant within the 95% CS for both mixFine and the "standard approach". Then, we compared how well the "consensus SNPs" and "top SNPs" were detected by mixFine and the standard fine-mapping approach using only a subset of samples. We subsampled to 90%, 80%, $\cdots$, 30% of samples, and repeated each random subsampling step 10 times.

At each subsampling level, mixFine, on average, detected more "consensus SNPs" than the standard approach (Figure 6A) and performance improved most on the more highly expressed genes (top tertile) (Figure S12). Moreover, mixFine detected "top SNPs" in 95% CSs with average size $= 9.6$ variants whereas the corresponding 95% CS of standard approach had average size $= 13.6$ variants (Figure S13). These results indicate that, when sufficient counts are available, mixFine, the multi-SNP model combining total and allele-specific counts, can better pinpoint causal cis-eQTLs than the standard approach on real data.

To compare the performance of mixPred and the standard method on real data, we implemented a cross-validated evaluation pipeline where we split the full data into k folds. At each fold, we trained the prediction model using the remaining $(k-1)$ folds and evaluated the performance (by Pearson correlation between predicted and observed $\log(Y_i^{\text{total}}/L_i)$) on the held out fold. We applied this evaluation pipeline to mixPred and the standard approach (based on inverse normalized expression) on the same 1000 genes as the subsampling analysis with k equals to 2 and 10 (corresponds to sample size $= 335$ and 603). Both mixPred and the standard approach achieved higher prediction performance as sample size increased, suggesting that sample size was not saturated and was a limiting factor of the prediction performance (Supplementary Figure S14). At the same sample size, we observed, on average, higher performance in mixPred as compared to the standard approach, and the performance gain was more obvious for smaller sample sizes (Figure 6B). These results indicate that mixPred can improve over the standard approach for building prediction models by leveraging allele-specific counts as extra observations.



(A)

(B)

Figure 6: Performance of mixFine and mixPred on GTEx v8 whole blood RNA-seq. (A) Fraction of detected "consensus SNPs" as a function of subsampling level, for mixFine and the standard approach. **(B)** Median Pearson correlation across k held-out folds for mixPred vs. the standard method, for $k = 2$, corresponds to training sample size $= 335$ (left panel) and $k = 10$, corresponding to training sample size $= 603$. Each point corresponds to a gene.

3 Discussion

We proposed a unified framework that integrates both allele-specific and total read counts to estimate genetic cis-regulatory effects, resulting in improved eQTL mapping, fine-mapping, and prediction of gene expression traits. Our suite of tools (mixQTL, mixFine, and mixPred) can be scaled to much larger sample sizes (thousands) due to the underlying log-linear approximation. By assuming multiplicative genetic effects, we transform the observed read counts into two approximately independent quantities: allelic imbalance and total read count. We take advantage of this independence to develop computationally efficient approaches that integrate both allele-specific and total reads.

Specifically, mixQTL estimates the genetic effect separately for the allelic imbalance and the total read counts, and combines the resulting statistics via meta-analysis. These calculations have computationally efficient closed-form solutions, enabling their use in the permutation schemes applied to compute FDR in eQTL mapping [Shabalin, 2012; Ongen et al., 2015; Taylor-Weiner et al., 2019].

Furthermore, the simple multi-SNP extension and the independence of the terms enable use of a two-step inference procedure. In the first step, the allelic imbalance and total read count are scaled so that the error terms have the same variance. And in the second step, given their approximate independence, the pair of equations (from allelic imbalance and total counts) can simply be input into existing fine-mapping and prediction algorithms. We showed through extensive simulations and applications to GTEx v8 data that our suite of methods outperforms current methods that use only total read count. Given the straightforward extension of current approaches with the models proposed here, as well as their computational efficiency, we anticipate that combining total and allele-specific read counts will find widespread use for eQTL mapping, fine-mapping, and prediction of gene expression.

4 Acknowledgement

We used data from the GTEx project (dbGap accession number phs000424.v8.p2). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI

Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc.(HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951,MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819),Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000NNN.vN.pN

HKI, YL, and AB were partially funded by R01MH10766 and P30 DK20595 (Diabetes Research and Training Center).

5 Disclosure

F.A. is an inventor on a patent application related to TensorQTL; H.K.I. has received speaker honoraria from GSK and AbbVie.

Code and data availability

Software mixQTL, mixFine, and mixPred https://github.com/hakyimlab/mixqtl

Reproducible pipeline https://github.com/liangyy/mixqtl-pipeline

GPU-based implementation embedded in tensorQTL https://github.com/broadinstitute/tensorqtl.

References

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

F. Aguet, A. N. Barbeira, R. Bonazzola, A. Brown, S. E. Castel, B. Jo, S. Kasela, S. Kim-Hellmuth, Y. Liang, M. Oliva, et al. The gtex consortium atlas of genetic regulatory effects across human tissues. *BioRxiv*, page 787903, 2019.

E. Evangelou and J. P. Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6):379–389, 2013.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E. Eyler, J. C. Denny, D. L. Nicolae, N. J. Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091, 2015.

C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, and V. Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics*, 10(5):e1004383, 2014.

A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. Penninx, R. Jansen, E. J. De Geus, D. I. Boomsma, F. A. Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245, 2016.

F. Hormozdiari, M. Van De Bunt, A. V. Segre, X. Li, J. W. J. Joo, M. Bilow, J. H. Sul, S. Sankararaman, B. Pasaniuc, and E. Eskin. Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

N. Kumasaka, A. J. Knights, and D. J. Gaffney. Fine-mapping cellular qtls with rasqual and atac-seq. *Nature genetics*, 48(2):206, 2016.

C. H. Lee, S. Cook, J. S. Lee, and B. Han. Comparison of two meta-analysis methods: inverse-variance-weighted average and weighted sum of z-scores. *Genomics & informatics*, 14(4):173, 2016.

Y. Lee, L. Francesca, R. Pique-Regi, and X. Wen. Bayesian multi-snp genetic association analysis: Control of fdr and use of summary statistics. *bioRxiv*, page 316471, 2018.

P.-R. Loh, G. Kichaev, S. Gazal, A. P. Schoech, and A. L. Price. Mixed-model association for biobank-scale datasets. *Nature genetics*, 50(7):906–908, 2018.

P. Mohammadi, S. E. Castel, A. A. Brown, and T. Lappalainen. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome research*, 27(11):1872–1884, 2017.

H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, and O. Delaneau. Fast and efficient qtl mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 2015.

A. A. Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10): 1353–1358, 2012.

O. Stegle, L. Parts, R. Durbin, and J. Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS computational biology*, 6(5): e1000770, 2010.

W. Sun. A statistical framework for eqtl mapping using rna-seq data. *Biometrics*, 68(1):1–11, 2012.

A. Taylor-Weiner, F. Aguet, N. J. Haradhvala, S. Gosai, S. Anand, J. Kim, K. Ardlie, E. M. Van Allen, and G. Getz. Scaling computational genomics to millions of individuals with gpus. *Genome biology*, 20(1): 1–5, 2019.

B. Van De Geijn, G. McVicker, Y. Gilad, and J. K. Pritchard. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11):1061, 2015.

U. Võsa, A. Claringbould, H.-J. Westra, M. J. Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Kasela, et al. Unraveling the polygenic architecture of complex traits using blood eqtl meta-analysis. *bioRxiv*, page 447367, 2018.

A. T. Wang, A. Shetty, E. O'Connor, C. Bell, M. M. Pomerantz, M. L. Freedman, and A. Gusev. Allele-specific qtl fine mapping with plasma. *The American Journal of Human Genetics*, 106(2):170–187, 2020.

G. Wang, A. K. Sarkar, P. Carbonetto, and M. Stephens. A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv*, page 501114, 2019.

X. Wen, R. Pique-Regi, and F. Luca. Integrating molecular qtl data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS genetics*, 13(3):e1006646, 2017.

Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher, et al. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics*, 48(5):481, 2016.

J. Zou, F. Hormozdiari, B. Jew, S. E. Castel, T. Lappalainen, J. Ernst, J. H. Sul, and E. Eskin. Leveraging allelic imbalance to refine fine-mapping for eqtl studies. *PLoS Genetics*, 15(12), 2019.

6 Material and Methods

6.1 Notation and terminology

Notation	Description	Synonym in text	Observable
i	Individual index.	-	-
h	Haplotype index, with $h = 1, 2$ for diploid.	-	-
X_i^h	Alternative allele count (0 or 1) of the variant linking to the gene haplotype h.	allelic dosage	Yes
L_i	The total number of reads in the RNA-seq library.	library size	Yes
Y_i^h	Count of reads originated from gene haplotype h.	haplotypic (read) count	No
$Y_i^{(h)\text{obs}}$	Allele-specific read count that gets aligned to the gene haplotype h.	allele-specific (read) count	Yes
Y_i^{total}	Total count of reads originated from any of the two gene haplotypes (sum).	total (read) count	Yes
$\theta_{0,i}$	The abundance of the gene haplotype relative to the total transcriptome when the linked causal variants are all in reference alleles	baseline (relative) abundance	No
θ_i^h	The abundance of the gene haplotype h relative to the total transcriptome in individual i	(relative) abundance; expression level†	No
β	The log fold change of gene haplotype abundance when linking to alternative allele relative the reference allele	allelic fold change (aFC) in natural log scale	No
$\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}$	The ratio of the allele-specific counts between two haplotypes	allelic imbalance	Yes
Y_i^{trc}	Shorthand of the term $\log \frac{Y_i^{\text{total}}}{2L_i}$.	-	-
Y_i^{asc}	Shorthand of the term $\log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}$	-	-
θ	Only used in simulation where $\theta = \text{E}(\theta_{0,i})$	expression level*	-

Table 1: Summary of notation and terminology used in the paper. The **Description** column contains a brief definition of each **Notation**, and the **Synonym in text** column contains the corresponding terminology used in the text. The **Observable** column indicates whether the entity is an observable variable or not. ($\dagger$, $\star$: expression level does not strictly refer to θ_i^h or $\text{E}(\theta_{0,i})$, but more generally to the abundance of the gene transcripts relative to the transcriptome.)

6.2 Statistical model of cis-regulation

For individual i, let X_i^1 and X_i^2 be the number of alternative alleles in each of the two haplotypes at the variant of interest. Let Y_i^1 and Y_i^2 be the number of reads mapped to each of the two haplotypes (i.e., haplotypic counts; in practice, these quantities are unobserved) and L_i the library size for individual i. As proposed in [Mohammadi et al., 2017], we use the concept of allelic fold change (aFC) to represent the genetic effect on cis-expression. We denote $\theta_{0,i}$ as the baseline abundance of the transcripts originating from each of the gene haplotype without considering genetic effect. Let β be the genetic effect of a variant of interest, which is defined as the log fold change relative to the reference allele. Then, the transcript abundance of each haplotype h after accounting for the genetic effect is $\theta_i^h = \theta_{0,i} \times g(\beta, X_i^h)$ where $g(\beta, X_i^h)$ is e^β if X_i^h is the alternative allele; otherwise $g(\beta, X_i^h) = 1$. We model read count Y_i^h as

$$\log Y_i^h | L_i, \theta_i^h \sim N(\log(L_i \theta_i^h), \tau_i^h). \tag{5}$$

In an RNA-seq experiment, a fraction of reads contribute to allele-specific read counts. Let α_i denote the fraction of allele-specific reads in individual i, which depends on the number of heterozygous sites within the transcript. Instead of observing haplotypic counts Y_i^1 and Y_i^2, we observe total read count Y_i^{total} and gene-level allele-specific read counts $Y_i^{(1)\text{obs}}$ and $Y_i^{(2)\text{obs}}$. Similarly, we further assume that the baseline abundance of allele-specific reads per haplotype is $\theta_{0,i} \times \alpha_i$, so we have

$$\begin{aligned}
\log Y_i^{(1)\text{obs}} | L_i, \theta_i^1, \alpha_i &\sim N(\log(\alpha_i L_i \theta_i^1), \tau_i^{(1)}) && (6)\\
\log Y_i^{(2)\text{obs}} | L_i, \theta_i^2, \alpha_i &\sim N(\log(\alpha_i L_i \theta_i^2), \tau_i^{(2)}) \\
\log Y_i^{\text{total}} | L_i, \theta_i^1, \theta_i^2 &= \log(Y_i^1 + Y_i^2) | L_i, \theta_i^1, \theta_i^2 && (7)\\
&\sim N(\log[L_i(\theta_i^1 + \theta_i^2)], \tau_i) && (8)
\end{aligned}$$

6.3 Linearizing the model by approximation

Based on the model described in Section 6.2 along with approximations under weak effect assumptions, we propose the following linear mixed effects model (see Supplementary Notes 8 for derivation):

$$\underbrace{\log \frac{Y_i^{\text{total}}}{2L_i}}_{Y_i^{\text{trc}}} = \mu_0 + z_i + \underbrace{\frac{X_i^1 + X_i^2}{2}}_{X_i^{\text{trc}}} \beta + \epsilon_i^{\text{trc}} \tag{9}$$

$$\underbrace{\log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}}_{Y_i^{\text{asc}}} = \underbrace{(X_i^1 - X_i^2)}_{X_i^{\text{asc}}} \beta + \epsilon_i^{\text{asc}} \tag{10}$$

$$z_i \sim N(0, \sigma_0^2), \;\; \epsilon_i^{\text{trc}} \sim N(0, \frac{\sigma^2}{Y_i}), \;\; \epsilon_i^{\text{asc}} \sim N(0, \underbrace{\frac{\sigma^2 Y_i^{(1)} Y_i^{(2)}}{Y_i^{(1)} + Y_i^{(2)}}}_{\sigma^2 / w_i}), \tag{11}$$

where z_i is the individual-level random effect capturing the between-individual variation of $\theta_{i,0}$. Notice that the individual-level random effect cancels out when we take the difference between the two log-scale allele-specific read counts (i.e., allelic imbalance in log-scale). The scaling of ϵ^{trc} and ϵ^{asc} in Eq 11 is to ensure that variance of read count scales linearly with the magnitude of read count (see Supplementary Notes 7.2). In other words, this model ensures $\text{Var}(Y) \approx \text{constant} \times \text{E}(Y)$, such that over-dispersion is implicitly taken into account.

Since σ^2/Y_i is typically much smaller than σ_0^2 (see Supplementary Notes 10), we can further simplify Eq 9, 10 as

$$\begin{aligned} Y_i^{\text{trc}} &= \mu_0 + X_i^{\text{trc}}\beta^{\text{trc}} + z_i \quad, z_i \quad \sim N(0, \sigma_0^2) && (12) \\ Y_i^{\text{asc}} &= \qquad X_i^{\text{asc}}\beta^{\text{asc}} + \epsilon_i^{\text{asc}}, \epsilon^{\text{asc}} \sim N(0, \sigma^2/w_i) && (13) \end{aligned}$$

Eqs 12, 13 are applicable to both single-SNP and multi-SNP scenarios. In the single-SNP case, X_i and β are scalars, and in the multi-SNP case, X_i and β are vectors including all SNPs within the cis-window (see Supplementary Notes 9).

6.4 Numerically efficient QTL mapping leveraging approximate independence of allelic imbalance and total read count

The likelihood function corresponding to the proposed model in Eqs 12, 13 takes the form

$$\prod_i \Pr(Y_i^{\text{total}}|\mu_0, \sigma_0^2, \sigma^2, \beta) \cdot \Pr(\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}|\sigma^2, \beta),$$

factoring into total read count and allelic imbalance components. (see Supplementary Notes 8.2). This means that the likelihood for total read count and the ratio of allele-specific read counts provide approximately independent information on β, and enables us to solve each component separately and combine the results via meta-analysis (standard approach with independent studies [Evangelou and Ioannidis, 2013]). Specifically, we fit β^{trc} and β^{asc} using total and allele-specific observations as two separate linear regression problems, and meta-analyze the results using inverse-variance weighting (see details in Supplementary Notes 10.2).

6.5 Two-step inference procedure for multi-SNP model

The prediction and fine-mapping problems both rely on the linearized model Eq 12, 13, but with different objectives. For prediction, the objective is to find the best predictor, whereas for fine-mapping, the objectiveis to infer whether β_k is non-zero. Existing solvers for both prediction and fine-mapping use total read information only and assume that data (X, y) follow the model $y = X\beta + \epsilon$, where the noise term ϵ is independent across the rows of the data matrix. We will refer to this model as the 'canonical' linear model. We propose a two-step inference procedure that first processes the data such that it approximates $y = X\beta + \epsilon$, and then uses existing solvers for prediction and fine-mapping problems, respectively.

For the first step, we process total and allele-specific reads separately to fit the 'canonical' linear model. Specifically, we estimate σ^2 from $(Y^{\text{asc}}, X^{\text{asc}})$ based on Eq 13 using elastic net with cross-validation. And similarly, based on Eq 12, we estimate σ_0^2 from $(Y^{\text{trc}}, X^{\text{trc}})$ and obtain the intercept μ_0 by running fine-mapping with $(Y^{\text{trc}}, X^{\text{trc}})$. Then, we shift Y^{trc} by $\hat{\mu}_0$ and scale $(Y^{\text{trc}}, X^{\text{trc}})$ by $1/\hat{\sigma}_0$. And similarly, we scale $(Y^{\text{asc}}, X^{\text{asc}})$ by $w/\hat{\sigma}$. These linear transformations ensure that the transformed $(\tilde{Y}^{\text{trc}}, \tilde{X}^{\text{trc}})$ and $(\tilde{Y}^{\text{asc}}, \tilde{X}^{\text{asc}})$ both approximately follow $Y = X\beta + \epsilon$. The implementation details are described in Supplementary Notes 11. At the second step, we concatenate the transformed data from both total and allele-specific read counts as $(\tilde{Y}, \tilde{X})$, which is compatible with existing solvers for prediction and fine-mapping problems.

6.6 Adjusting for covariates

When analyzing real data, we need to take covariates such as sex, batch effect, population stratification into account. Here, we adapt the procedure which has been proposed previously [Mohammadi et al., 2017]. We regress out the effect of covariates beforehand and use the residual as the response in both QTL mapping and fitting multi-SNP model. Specifically, let $c_1, \cdots, c_K$ denote the K covariates to be considered. We first regress Y^{trc} against $c_1, \cdots, c_K$ jointly and select the covariates with nominally significant coefficients ($p < 0.05$). Then we regress Y^{trc} against the selected covariates jointly and set the residuals as the adjusted Y^{trc} for QTL mapping and multi-SNP inference downstream.

6.7 Simulation scheme

We simulate RNA-seq reads with total and allele-specific readouts as sketched in three steps in Figure 1. In step 1, we specify, for each individual i, the position of heterozygous sites within the gene body. The expected read count from each haplotype transcripts, $\mathrm{E}(Y_i^h)$, is determined by the RNA-seq library size L_i, the baseline abundance of the transcript $\theta_{0,i}$, and the genetic effect β. In step 2, given the expected haplotypic count, we draw Y_i^h from Negative Binomial to model the variation among count data. In step 3, we position the reads randomly along the gene body and readout observed allele-specific count $Y_i^{(h)\text{obs}}$ by counting the number of reads overlapping heterozygous sites simulated in step 1. The total read count readout is $Y_i = Y_i^1 + Y_i^2$, which is independent of the number of heterozygous sites.

To survey a wide range of parameters, we simulate data with a grid of parameters. We vary sample size among 100, 200, ..., 500. At library size around 90 million, we vary the level of $\theta_{0,i}$ to cover the gene with different expression levels, among $5 \times 10^{-5}, 2.5 \times 10^{-5}, 1 \times 10^{-5}, 2.5 \times 10^{-6}, 1 \times 10^{-6}$. The genetic effect, aFC, is set to 1 (null), 1.01, 1.05, 1.1, 1.25, 1.5, 2, 3 in the single SNP model. For the multi-SNP scenario, we set the number of causal SNPs between 1 and 3 with heritability from 0.2 to 0.55. The number of polymorphic sites within the gene body is centered around 10 with minor allele frequency from 0.05 to 0.3. A detailed description and parameter settings are provided in the Supplementary Notes 12.

6.8 Analysis of GTEx v8 data

We downloaded the phased genotypes, total read count matrix, and variant-level allele-specific read counts in whole blood from GTEx release 8 [Aguet et al., 2019] via dbGaP (accession number phs000424.v8.p1). To obtain gene-level read counts, we summed over allele-specific counts at all the heterozygous sites for each gene haplotype. We also obtained library size, sex, and genotype PCs from GTEx v8. For comparisons with the inverse normalization-based approach, we also downloaded normalized expression matrices.

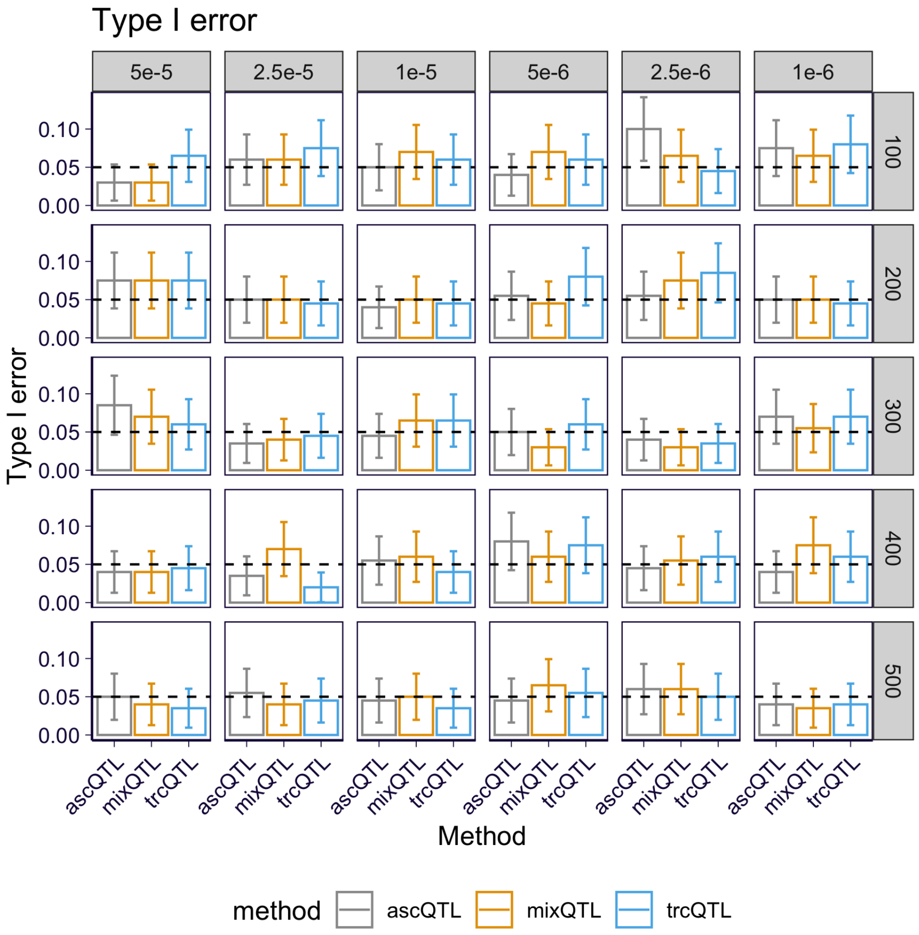
Similarly to the GTEx v8 report [Aguet et al., 2019], we restricted the analysis to the cis-regulatory window defined as 1Mbp up/downstream of the transcription start site of each gene.

To obtain the PEER factors for mixQTL analysis, we ran `peertool` [Stegle et al., 2010] on matrix with entry $\log(\frac{Y_{i,g}}{2L_i})$ for individual i and gene g (impute value by k-nearest neighbor if $Y_{i,g}$ is zero which is done by `impute::impute.knn` in R).

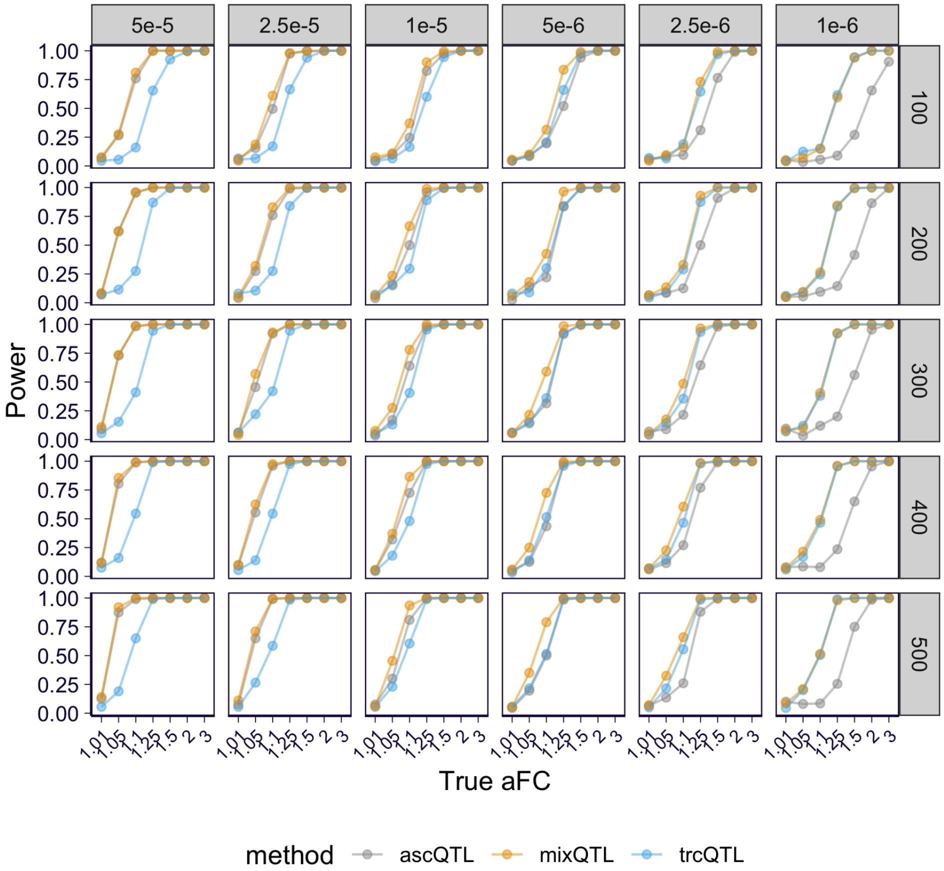
We considered very large allele-specific counts to be outliers likely due to alignment artifacts and removed individuals with allele-specific read counts greater than 1000. To further limit the undue influence of large count outliers on the estimated log fold-change, $\hat{\beta}^{\text{asc}}$, we set the largest weight $\left(\frac{1}{Y^{(1)\text{obs}}} + \frac{1}{Y^{(2)\text{obs}}}\right)^{-1}$ to be at most K fold to the smallest one, where $K = \min(10, \text{sample size}/10)$.

Specific analysis focused on high or low expression were performed with different gene filtering criteria as stated in the Results section.

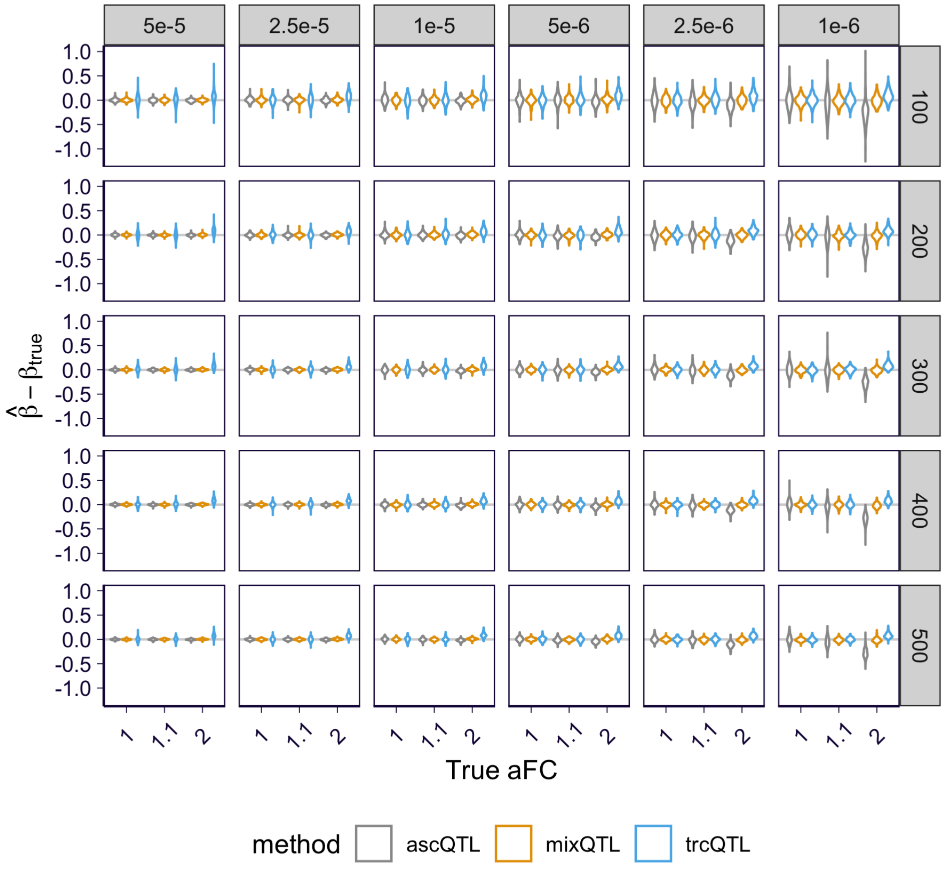
Supplementary Figures



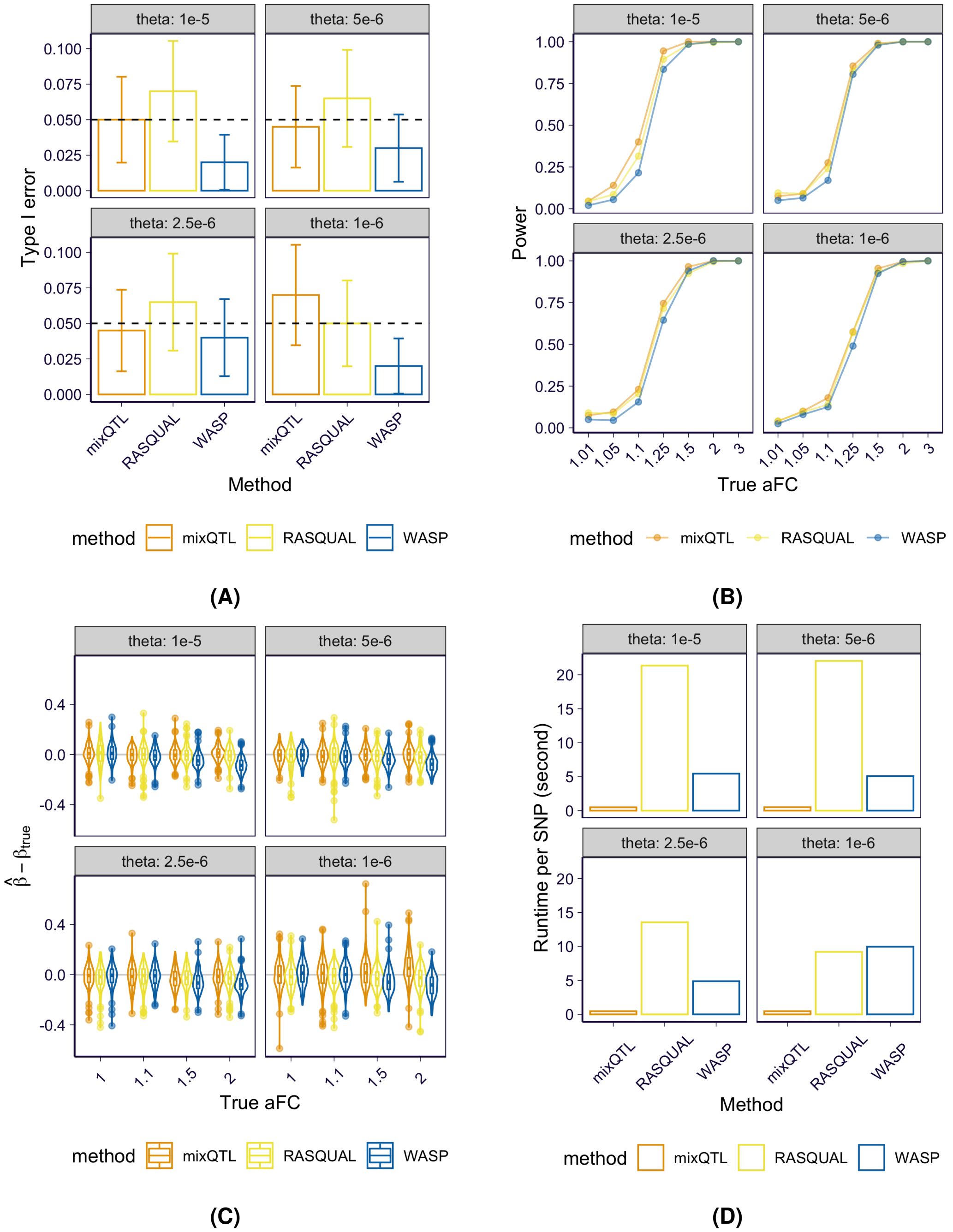
Supplementary Fig. S1. Type I error of mixQTL, ascQTL, and trcQTL on the full grid of simulations. Each panel shows results on data simulated under a pair of θ (by column) and sample size (by row).



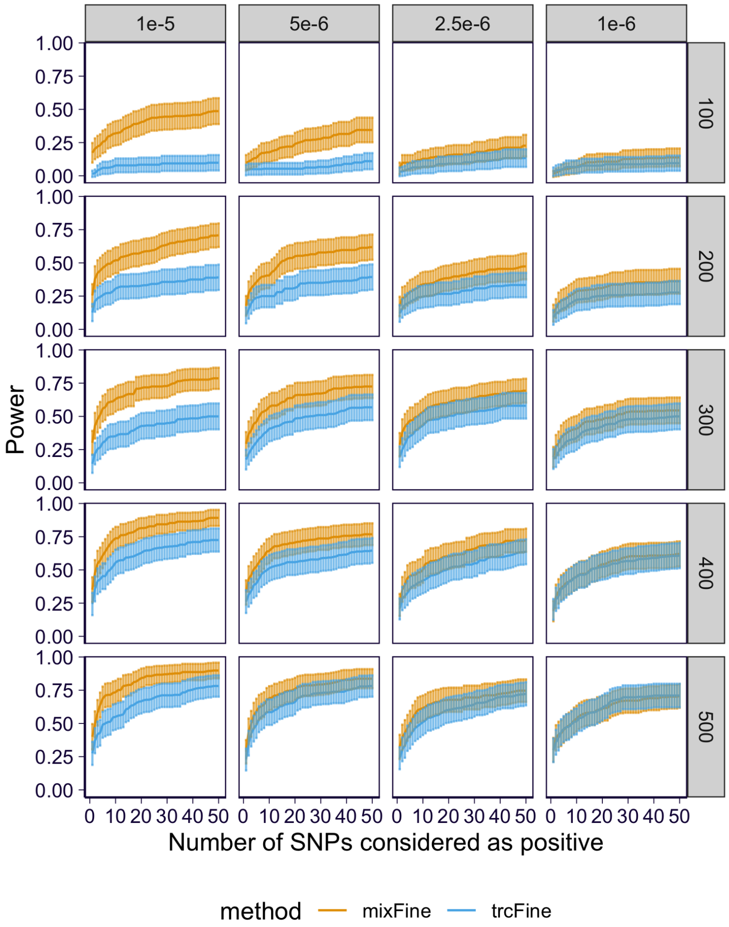
Supplementary Fig. S2. Power of mixQTL, ascQTL, and trcQTL on the full grid of simulations. Each panel shows results on data simulated under a pair of θ (by column) and sample size (by row).



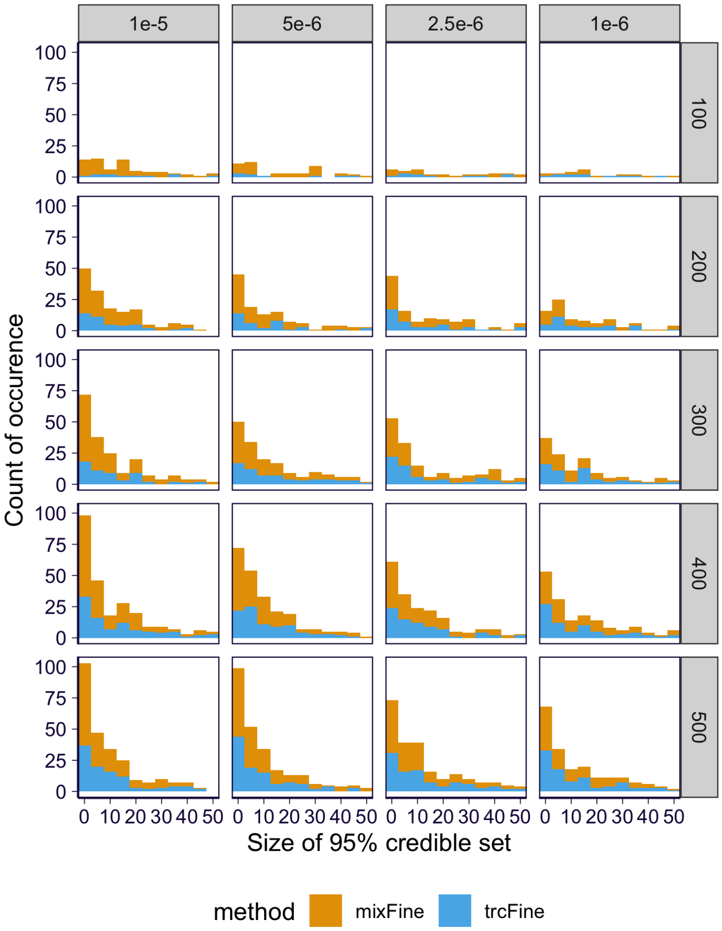
Supplementary Fig. S3. Difference between $\hat{\beta}$ and true β of mixQTL, ascQTL, and trcQTL on the full grid of simulations. Each panel shows results on data simulated under a pair of θ (by column) and sample size (by row).



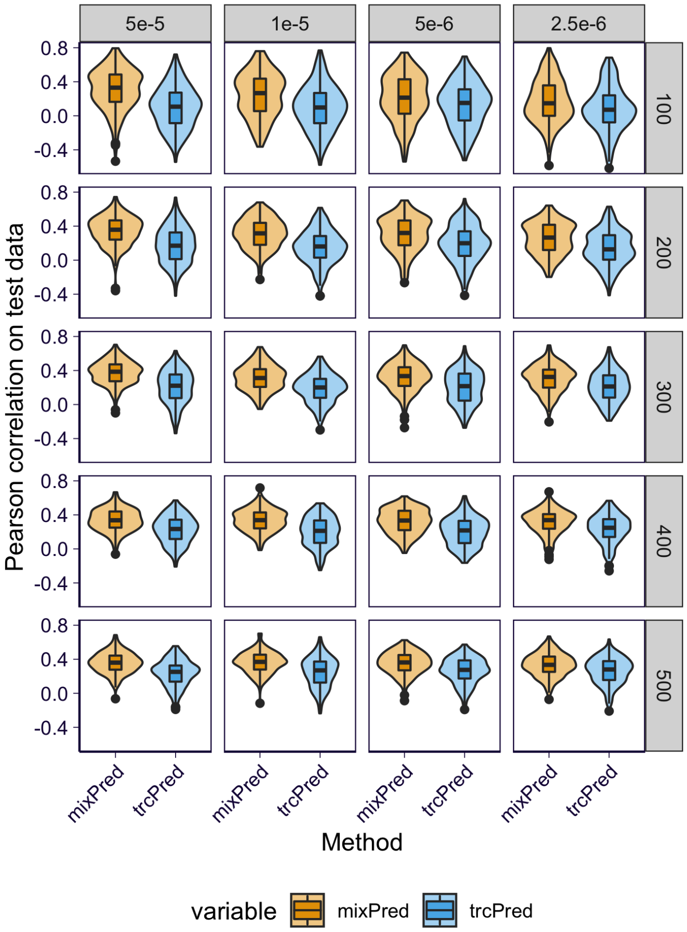
Supplementary Fig. S4. Performance of WASP, RASQUAL, and mixQTL on simulated data. Here, we show results (by panel) for a range of read depths: $\theta = 1 \times 10^{-5}, 5 \times 10^{-6}, 2.5 \times 10^{-6}, 1 \times 10^{-6}$ (defined as the average baseline abundance in the simulation that controls the number of reads) with sample size equal to 100. **(A)** Type I errors computed from 200 replicates. **(B)** Power (y-axis) calculated at $\alpha = 0.05$ for a range of true aFC values (x-axis). **(C)** Difference between estimated aFC and true aFC for a range of true effect sizes (x-axis). **(D)** Per-test computation times. The computation of overdispersion parameters was included.



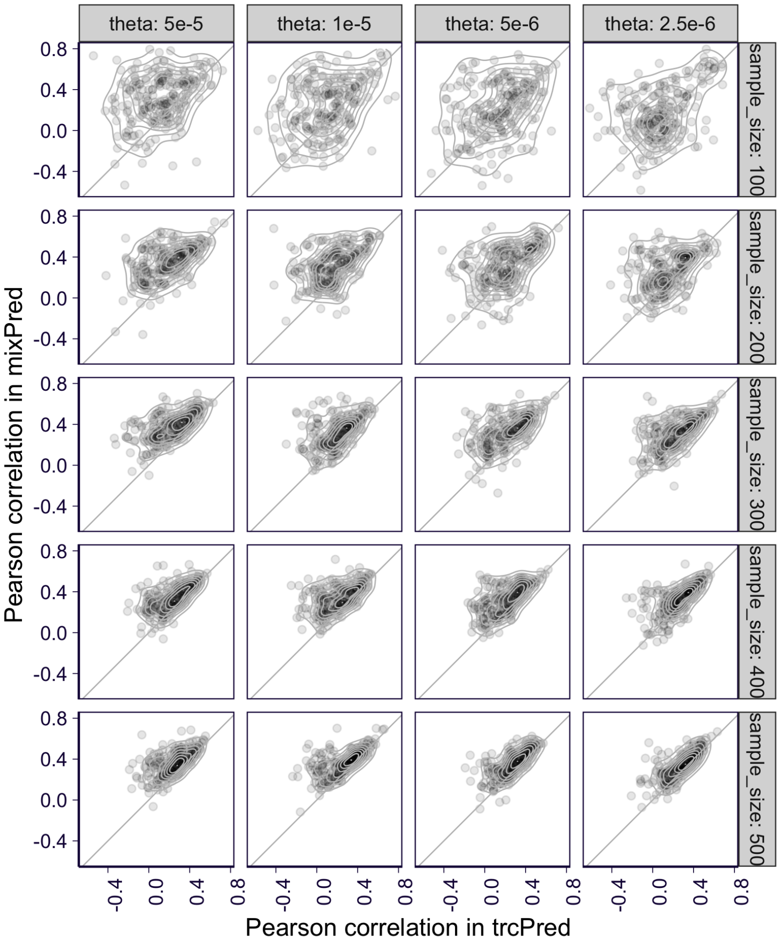
Supplementary Fig. S5. Power curves of mixFine and trcFine on the full grid of simulations. Each panel shows results on data simulated under a pair of θ (by column) and sample size (by row).



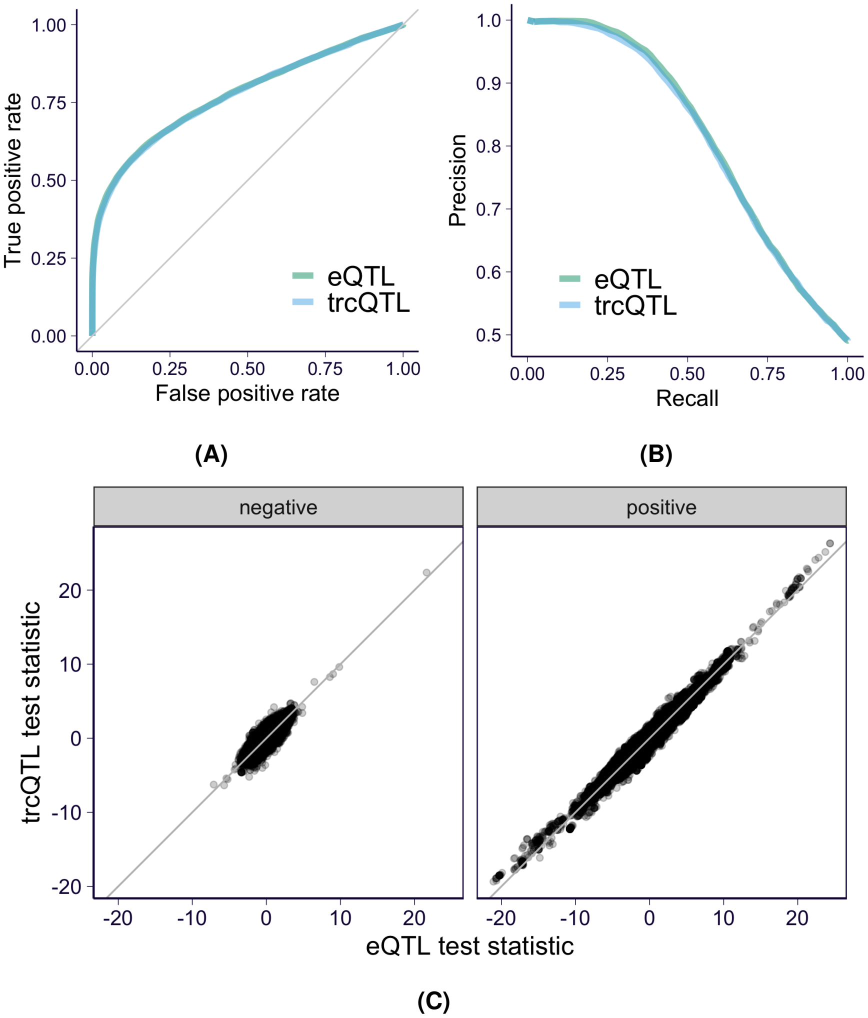
Supplementary Fig. S6. Distribution of the positive 95% CS's which contain causal variants in mixFine and trcFine on the full grid of simulations. Each panel shows results on data simulated under a pair of θ (by column) and sample size (by row).



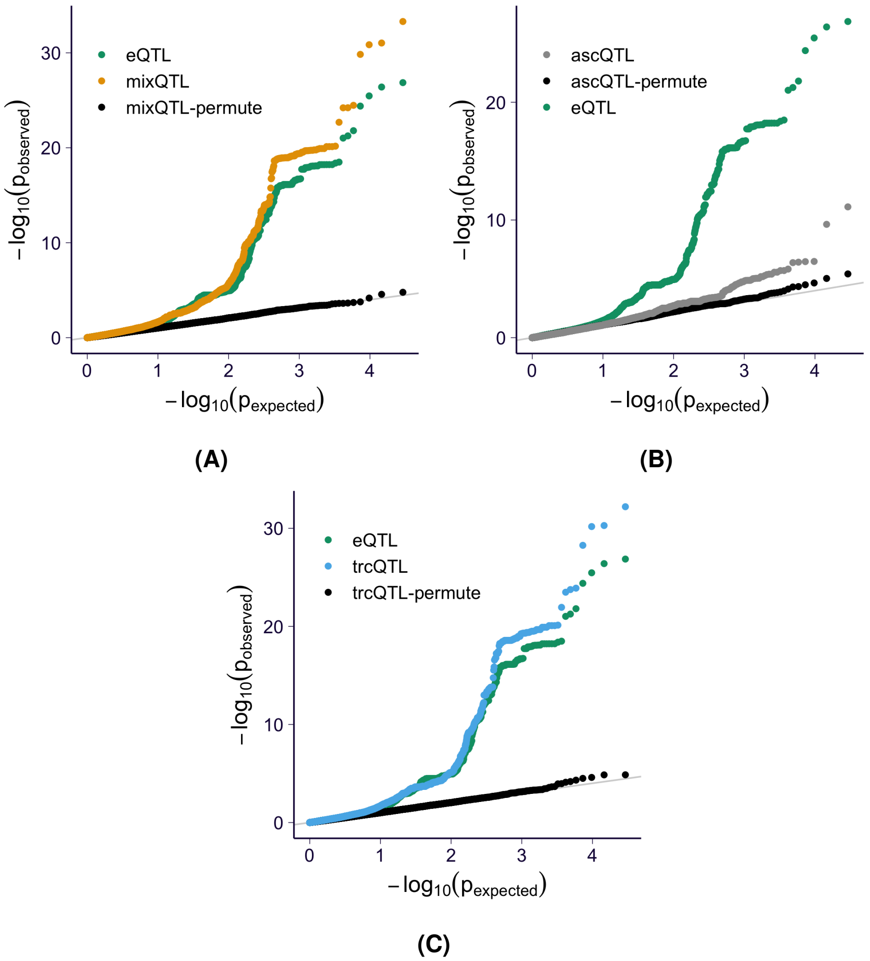
Supplementary Fig. S7. Distribution of Pearson correlations between predicted and observed expression level (in the scale $\log(Y_i^{\textbf{total}}/L_i)$) for mixPred and trcPred on the full grid of simulations. Correlation is calculated on held-out test data. Each panel shows results on data simulated under a pair of θ (by column) and sample size (by row).



Supplementary Fig. S8. Pairwise comparison of prediction performance of mixPred and trcPred on the full grid of simulations. Correlation of predicted versus observed expression level (in the scale $\log(Y_i^{\text{total}}/L_i)$) is calculated on held-out test data. The prediction performance of mixPred (y-axis) is plotted against the prediction performance of trcPred (x-axis) for each split. Each panel shows results on data simulated under a pair of θ (by column) and sample size (by row).



Supplementary Fig. S9. Performance of trcQTL and the standard eQTL approach on genes with low total read counts. Genes with low total counts are defined as having no more than 50 total read counts in any one sample. In GTEx v8 whole blood samples, we extracted 912 genes with low total counts and calculated trcQTL estimates for variants in the corresponding cis-windows. To compare the power of trcQTL and eQTL, we used the 85,129 variant/gene pairs with FDR $<$ 0.05 in eQTLGen as a "ground truth" set. We also randomly selected 88,242 variant/gene pairs from the pairs with p-value $>$ 0.5 in eQTLGen as a negative set. **(A,B)** ROC and PR curves for trcQTL and the standard eQTL method. **(C)** Test statistics for the standard eQTL method (x-axis) and trcQTL (y-axis). The variant/gene pairs in the eQTLGen negative set are shown in the left panel, and pairs in the "ground truth" set in the right panel.

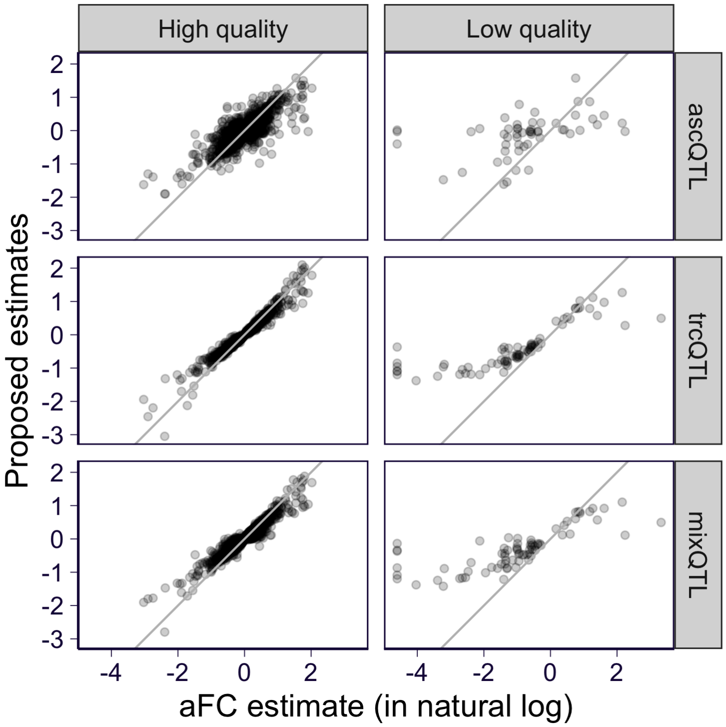


(A)

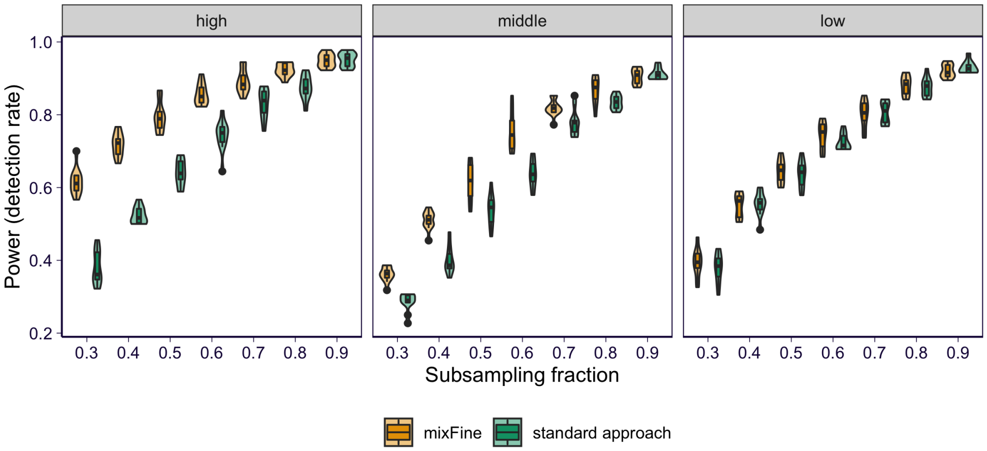
(B)

(C)

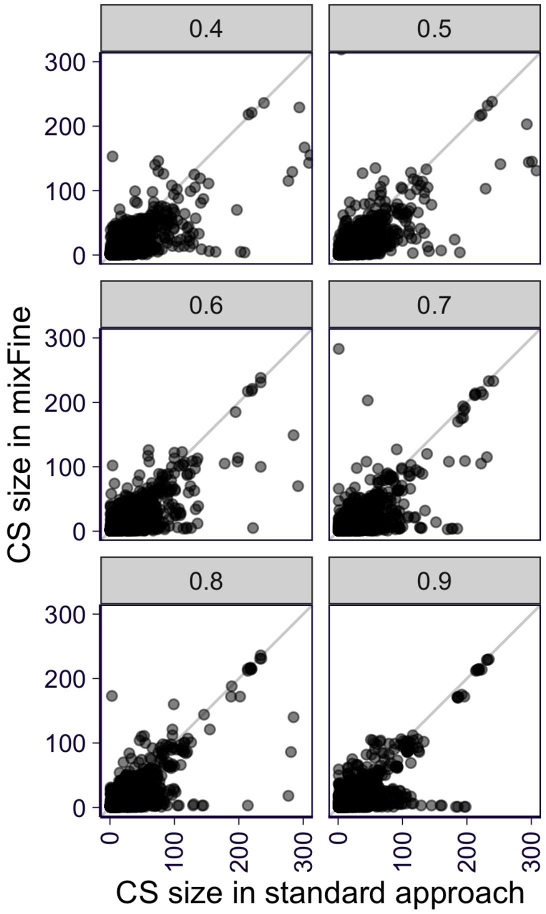
Supplementary Fig. S10. QQ-plot of nominal p-values from ascQTL and trcQTL on four randomly selected genes in GTEx v8 whole blood RNA-seq. The nominal p-values of trcQTL and ascQTL are compared against the standard eQTL method for four randomly selected genes ENSG00000000457, ENSG00000001461, ENSG00000002834, and ENSG00000277734. The results of ascQTL and trcQTL on permuted genotypes are shown in black. **(A)** Results from mixQTL. **(B)** Results from ascQTL. **(C)** Results from trcQTL.



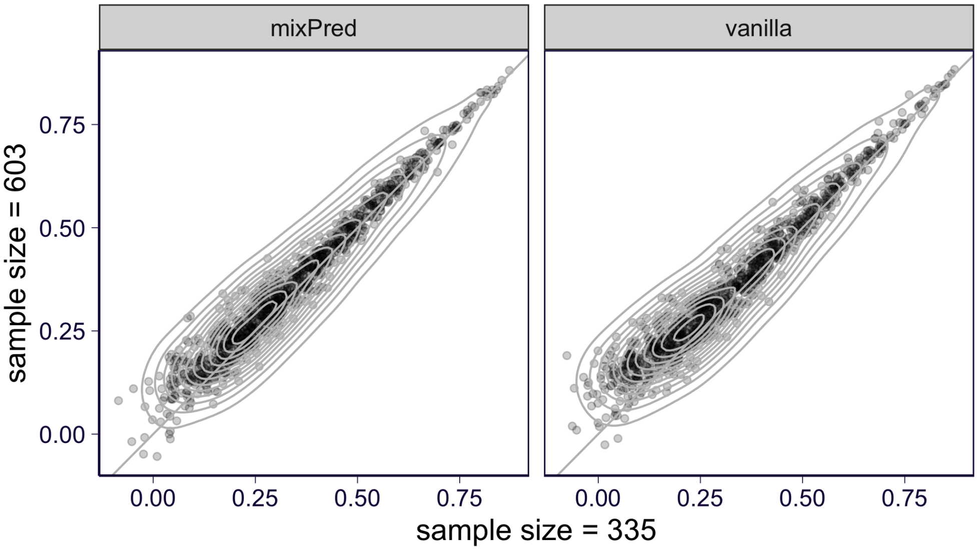
Supplementary Fig. S11. Comparison of aFC estimates from GTEx v8 and the estimated allelic fold change of ascQTL, trcQTL, and mixQTL. The estimates of the top variants in the eGenes of GTEx v8 whole blood are shown (based on eQTL results). On the x-axis, the aFC estimate reported by GTEx v8 is shown (the reported value is in $\log_2$ and, for visualization, we rescale it to natural log scale by multiplying the value with $\log(2)$). On the y-axis, the estimated allelic fold changes (in natural log scale) of ascQTL, trcQTL, and mixQTL are shown. The variant/gene pairs are stratified on the basis of the quality of aFC estimate, which is defined as 'high quality' if the 95% CI of $\log_2$ aFC is smaller than 1 and the low and high boundaries of the 95% CI are not more extreme than $-\log_2(50)$ and $\log_2(50)$, and as 'low quality' otherwise.



Supplementary Fig. S12. Performance of mixFine on GTEx v8 whole blood RNA-seq stratified by expression level. At each subsampling level (x-axis), the fraction of “consensus SNPs” being detected is shown on the y-axis. Each panel shows the results of genes stratified by expression level tertiles.



Supplementary Fig. S13. Performance of mixFine on GTEx v8 whole blood RNA-seq on pinpointing the "top" SNPs. At each subsampling level (shown in each panel), we compare mixFine (y-axis) and the standard method (x-axis) on the size of 95% CS's which are paired by sharing the same "top SNP".



Supplementary Fig. S14. Performance of mixPred and the standard method on GTEx v8 whole blood RNA-seq with different sample sizes. For each gene, the median prediction performance across the k held out folds is shown. Here the results with sample size = 335 ($k = 2$) are shown on x-axis and the results with sample size = 603 ($k = 10$) are shown on y-axis. The left panel shows the results of mixPred and the right panel shows the results of the standard method.

Supplementary Notes

7 Statistical model for read count

Here we introduce the statistical model of read count in this paper. For completeness, we opt for keeping some text that overlaps with main text. Recall that i indexes individual and h indexes haplotypes. X_i^h is the phased genotype of the corresponding individual i haplotype h. Y_i^{total} is the total read count within the gene body and L_i is the library size. $Y_i^{(h)\text{obs}}$ is the allele-specific read count of the corresponding haplotype transcript h and Y_i^h is the actual (though unobserved) read count of the haplotype transcript h. α_i is the expected fraction of allele-specific reads in individual i. Additionally, the cis-genetic effect a single SNP on haplotype h is represented as $g(\beta, X_i^h)$ where

$$g(\beta, X_i^h) = \begin{cases} 1 & \text{, if } X_i^h = 0 \\ e^{\beta} & \text{, if } X_i^h = 1 \end{cases} \tag{14}$$

$$= e^{X_i^h \beta} \tag{15}$$

We assume multiplicative effect when there are multiple causal SNPs. And the effect of multiple SNPs $j = 1, \cdots, p$ is

$$\prod_{j=1}^{p} g(\beta_j, X_{ij}^h) = e^{\sum_j X_{ij}^h \beta_j} \tag{16}$$

$$= e^{\boldsymbol{X}_i^h \boldsymbol{\beta}} \tag{17}$$

$$:= g(\boldsymbol{\beta}, \boldsymbol{X}_i^h) \tag{18}$$

7.1 Overview

We model haplotypic count Y_i^h as lognormal distribution as follow.

$$\log Y_i^h \sim N(\log(L_i\theta_i^h), \tau_i^h) \tag{19}$$

$$\theta_i^h = \theta_{0,i} \times g(\beta, \boldsymbol{X}_i^h), \tag{20}$$

$\theta_{0,i}$ is the baseline abundance of haplotype transcript without considering genetic effect (*i.e.* it represents the abundance when the affecting SNP is reference allele).

In practice, we do not observe Y_i^h but allele-specific read count $Y_i^{(h)\text{obs}}$. So, we further assume that the baseline abundance of corresponding allele-specific reads are $\theta_{0,i}^{(1)} = \theta_{0,i}^{(2)} = \alpha_i\theta_{0,i}$. And by definition, total read count $Y_i^{\text{total}} = Y_i^1 + Y_i^2$. So, similar to Eq 19, 20, $Y_i^{(h)\text{obs}}$ and Y_i^{total} follow

$$\log Y_i^{(h)\text{obs}} \sim N(\log(L_i\theta_i^{(h)}), \tau_i^{(h)}) \tag{21}$$

$$\log Y_i^{\text{total}} \sim N(\log(L_i\theta_i), \tau_i) \tag{22}$$

$$\theta_i^{(h)} = \alpha_i\theta_{0,i} \times g(\beta, \boldsymbol{X}_i^h) \tag{23}$$

$$\theta_i = \theta_{0,i} \times [g(\beta, \boldsymbol{X}_i^1) + g(\beta, \boldsymbol{X}_i^2)] \tag{24}$$

7.2 Parameterizing τ to weight total and AS count properly

Note that lognormal distribution has the following property.

$$\log X \sim N(\mu, \tau) \tag{25}$$

$$X \sim \text{lognormal}(\mu, \tau) \text{ , by definition of lognormal} \tag{26}$$

$$\text{E}(X) = e^{\mu + \frac{\tau}{2}} \tag{27}$$

$$\text{Var}(X) = (e^{\tau} - 1)(e^{2\mu+\tau}) \tag{28}$$

When modeling read count, given the mean, we would like the variance to scale linearly with the mean (as assumed in RASQUAL [Kumasaka et al., 2016]). In other word, we want to ensure that $\text{Var}(X)/\text{E}(X)$, also known as over-dispersion parameter, is roughly a constant. From Eq 27, 28 we have $\text{Var}(X) = (e^{\tau} - 1)\text{E}(X)^2$. For count data, since τ is capturing the variation of count in log-scale, τ is typically close to 0. So $e^{\tau} - 1 \approx \tau$ and $\text{Var}(X) \approx \tau\text{E}(X)^2$. This result suggests that to ensure $\text{Var}(X)/\text{E}(X) = \text{constant}$, τ should be approximately proportional to $1/\text{E}(X)$. So, for the distribution of $Y \sim \text{lognormal}(\log(L\theta), \tau)$, we impose the constraint on τ such that $\tau \approx \sigma^2/\text{E}(Y)$. In practice, $\text{E}(Y)$ is unknown so that we plug-in Y in replace of $\text{E}(Y)$.

8 Single-SNP model

On the basis of the model described in Supplementary Notes 7.1, we propose the single-SNP model where we focus on one ”test SNP” X_i^h instead of the whole phased haplotype $\boldsymbol{X}_i^h$. Hence, the cis-genetic effect of interest is $g(\beta, X_i^h)$.

8.1 From likelihood to linear mixed model

Here, we model cis-genetic effect of test SNP as allelic fold change (aFC) [Mohammadi et al., 2017]. So β is log-scale aFC in $g(\beta, X_i^{(h)}) = e^{X_i^{(h)}\beta}$. From Eq 21, 23, we have (for $h = 1, 2$)

$$
\begin{aligned}
\log Y_i^{(h)\text{obs}} &= \log L_i + \log \theta_i^{(h)} + \epsilon_i^{(h)} && (29) \\
&= \log L_i + \log \alpha_i + \log \theta_i^{h} + \epsilon_i^{(h)} && (30) \\
&= \log L_i + \log \alpha_i + \log \theta_{0,i} + \log(e^{X_i^h \beta}) + \epsilon_i^{(h)} && (31) \\
&= \log L_i + \log \alpha_i + \log \theta_{0,i} + X_i^h \beta + \epsilon_i^{(h)} && (32) \\
\epsilon_i^{(h)} &\sim N(0, \frac{\sigma^2}{Y_i^{(h)}}), && (33)
\end{aligned}
$$

where the error term scaling in Eq 33 follows from the discussion in Supplementary Notes 7.2. To further simplify the term $\log \theta_{0,i}$, as the variation of baseline abundance among individuals, we assume $\log \theta_{0,i} \sim N(\mu_0, \sigma_0^2)$. So that Eq 32, 33 can be further written as

$$
\begin{aligned}
\log Y_i^{(h)\text{obs}} &= \mu_0 + \log L_i + \log \alpha_i + z_i + X_i^h \beta + \epsilon_i^{(h)} && (34) \\
\epsilon_i^{(h)} &\sim N(0, \frac{\sigma^2}{Y_i^{(h)\text{obs}}}), \; z_i \sim N(0, \sigma_0^2), && (35)
\end{aligned}
$$

which is the approximated likelihood function for allele-specific counts $Y_i^{(1)\text{obs}}$ and $Y_i^{(2)\text{obs}}$. Such likelihood function is equivalent to linear mixed effects model.

Furthermore, we can linearize the likelihood of total read count Y_i^{total} in similar fashion. From Eq 22, 24 , we have

$$
\begin{aligned}
\log Y_i^{\text{total}} &= \mu_0 + \log L_i + z_i + \log(\theta_i^1 + \theta_i^2) + \epsilon_i && (36) \\
&= \mu_0 + \log L_i + z_i + \log(e^{X_i^1 \beta} + e^{X_i^2 \beta}) + \epsilon_i && (37) \\
\epsilon_i &\sim N(0, \frac{\sigma^2}{Y_i^{\text{total}}}), \; z_i \sim N(0, \sigma_0^2) && (38)
\end{aligned}
$$

Here we linearize $\log(e^{X_i^1 \beta} + e^{X_i^2 \beta})$ under the weak-effect assumption as follow

$$
\begin{aligned}
\log(e^{X_i^1 \beta} + e^{X_i^2 \beta}) &= \log[(X_i^1 e^{\beta} + 1 - X_i^1) + (X_i^2 e^{\beta} + 1 - X_i^2)] && (39) \\
&= \log(2 + X_i e^{\beta} - X_i) \quad \text{, let } X_i = X_i^1 + X_i^2 && (40) \\
&= \log[2 + X_i(e^{\beta} - 1)] && (41) \\
&= \log 2 + \frac{1}{2}(e^{\beta} - 1)X_i + o(X_i(e^{\beta} - 1)) && (42) \\
&\approx \log 2 + \frac{1}{2} X_i \beta \quad \text{, when } \beta \text{ is close to 0} && (43)
\end{aligned}
$$

So that Eq 37 can be approximated as

$$
\log \frac{Y_i^{\text{total}}}{2} \approx \mu_0 + \log L_i + z_i + \frac{X_i^1 + X_i^2}{2} \beta + \epsilon_i \tag{44}
$$

In summary, combining Eq 34 ,38, 35, 44, we have a linear mixed effects model unifying total and allele-specific read counts after linearization along with other approximations. And it also serves as an approximated likelihood for total and allele-specific reads, in which we can see that these read counts are not independent since they share the same random effect z_i.

8.2 Simplifying the model

Note that α_i is not observed so that we are unable to solve the model proposed in Supplementary Notes 8.1 in a computationally efficient manner. Here we address this problem by re-parameterizing the model. In principle, conditioning on genetic effect β, the ratio of allele-specific reads should be independent to the observations on the total read counts. This intuition motivates us to model the ratio of $Y_i^{(1)\text{obs}}$ and $Y_i^{(2)\text{obs}}$ rather than each of them separately. Mathematically, we subtract $\log Y_i^{(2)\text{obs}}$ from $\log Y_i^{(1)\text{obs}}$, which gives

$$\log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}} = (X_i^1 - X_i^2)\beta + \epsilon_i^{\text{asc}} \tag{45}$$

$$\epsilon_i^{\text{asc}} \sim N(0, \sigma^2(\frac{1}{Y_i^{(1)\text{obs}}} + \frac{1}{Y_i^{(2)\text{obs}}})), \tag{46}$$

where both z_i and α_i cancel out. This result naturally shows that the likelihood function of Y_i^{total} and $\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}$ takes the form:

$$\mathcal{L}(\boldsymbol{Y}^{\text{total}}, \frac{\boldsymbol{Y}^{(1)\text{obs}}}{\boldsymbol{Y}^{(2)\text{obs}}}; \mu_0, \sigma_0^2, \sigma^2, \beta) = \prod_i \Pr(Y_i^{\text{total}}|\mu_0, \sigma_0^2, \sigma^2, \beta) \Pr(\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}|\sigma^2, \beta) \tag{47}$$

$$= \underbrace{\prod_i \Pr(Y_i^{\text{total}}|\mu_0, \sigma_0^2, \sigma^2, \beta)}_{\text{total read count likelihood}} \underbrace{\prod_i \Pr(\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}|\sigma^2, \beta)}_{\text{allele-specific read count likelihood}} \tag{48}$$

$$:= \mathcal{L}^{\text{trc}}(\boldsymbol{Y}^{\text{total}}) \times \mathcal{L}^{\text{asc}}(\frac{\boldsymbol{Y}^{(1)\text{obs}}}{\boldsymbol{Y}^{(2)\text{obs}}}) \tag{49}$$

With the simplification shown in Eq 45, the model used for inference can be summarized as follow

$$\log \frac{Y_i^{\text{total}}}{2L_i} = \mu_0 + z_i + \frac{X_i^1 + X_i^2}{2}\beta \quad + \epsilon_i^{\text{trc}} \tag{50}$$

$$\log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}} = \qquad (X_i^1 - X_i^2)\beta + \epsilon_i^{\text{asc}} \tag{51}$$

$$z_i \sim N(0, \sigma_0^2),\ \epsilon_i^{\text{trc}} \sim N(0, \frac{\sigma^2}{Y_i^{\text{total}}}),\ \epsilon_i^{\text{asc}} \sim N(0, \frac{\sigma^2 Y_i^{(1)\text{obs}} + Y_i^{(2)\text{obs}}}{Y_i^{(1)\text{obs}} Y_i^{(2)\text{obs}}}) \tag{52}$$

9 Generalizing to multi-SNP model

The linearized model described in Eq 50, 51, 52 is easily extensible to multi-SNP scenario since we assume multiplicative genetic effect, as described in Supplementary Notes 18. To see the extension, all we need to examine is how $\log \theta_i^h$ and $\log(\theta_i^1 + \theta_i^2)$ as compared to the single SNP case since the rest of the terms stay the same.

$$
\begin{aligned}
\log \theta_i^h &= \log \theta_{0,i} + \log g(\boldsymbol{\beta}, \boldsymbol{X}_i^h) && (53)\\
&= \log \theta_{0,i} + \log e^{\boldsymbol{X}_i^h \boldsymbol{\beta}} && (54)\\
&= \log \theta_{0,i} + \boldsymbol{X}_i^h \boldsymbol{\beta} && (55)\\
\log(\theta_i^1 + \theta_i^2) &= \log \theta_{0,i} + \log\{\prod_j [1 + (e^{\beta_j} - 1) X_{ij}^1] + \prod_j [1 + (e^{\beta_j} - 1) X_{ij}^2]\}, && (56)\\
&\quad \text{similar to Eq 39} && (57)\\
&\approx \log \theta_{0,i} + \log[1 + \sum_j (e^{\beta_j} - 1) X_{ij}^1 + 1 + \sum_j (e^{\beta_j} - 1) X_{ij}^2], && (58)\\
&\quad \text{high orders term like } (e^{\beta_j} - 1) X_{ij}^1 (e^{\beta_{j'}} - 1) X_{ij'}^1 \text{ are ignored} && (59)\\
&= \log \theta_{0,i} + \log(2 + \sum_j (e^{\beta_j} - 1) X_{ij}) \text{ , } X_{ij} := X_{ij}^1 + X_{ij}^2 && (60)\\
&\approx \log \theta_{0,i} + \log 2 + \frac{1}{2} \boldsymbol{X}_i \boldsymbol{\beta} \text{ , follows similarly as Eq 42, 43} && (61)
\end{aligned}
$$

So, we can simply plug-in the multi-SNP version of $\log \theta_i^h$ and $\log(\theta_i^1 + \theta_i^2)$ to Eq 30 and 36 respectively and the similar conclusion follows with $\boldsymbol{X}$ and $\boldsymbol{\beta}$ in replace of X and β.

10 QTL mapping procedure

In the following, we describe the mixQTL procedure to map cis-eQTLs under the model proposed in Eq 50, 51, 52.

10.1 Converting the problems into two linear regressions

Instead of solving the proposed mixed effects model using numerical solver, we propose a meta-analysis procedure. In this procedure, we solve Eq 50 and 51 separately and meta-analyze the estimates afterwards.

Specifically, to solve Eq 50, we first recognize that σ_0^2 is much larger than $\sigma^2 / Y_i^{\text{total}}$. This is due to the following three facts: 1) σ^2 has the scale of 1 ($\sigma^2 = 1$ corresponds to Poisson); 2) Y_i^{total} is total count which is typically hundreds to thousands; and 3) $e^{\sigma_0} - 1$ is roughly the scale of the ratio between $\theta_{0,i}$ and the population mean abundance ($\text{E}(\theta_{0,i})$), which makes $\sigma_0 \sim 0.5$ (corresponds to $\theta_{0,i}$ to $\text{E}(\theta_{0,i})$ ratio being from 0.6 to 1.6) a reasonable estimate. So, we further simplify Eq 50 by ignoring the noise term from ϵ_i^{trc}. Such simplification results in the following linear model

$$
Y_i^{\text{trc}} = \mu_0 + X_i^{\text{trc}} \beta^{\text{trc}} + z_i, \; z_i \sim N(0, \sigma_0^2) \; , \tag{62}
$$

where $X^{\text{trc}} := \frac{X^1 + X^2}{2}$, $Y^{\text{trc}} = \log \frac{Y_i^{\text{total}}}{2L_i}$. Eq 62 itself can be used for QTL mapping and we call this approach trcQTL in the paper.

For solving Eq 51, notice that it is weighted simple linear regression with the form

$$
Y_i^{\text{asc}} = X_i^{\text{asc}} \beta^{\text{asc}} + \epsilon_i^{\text{asc}}, \; \epsilon_i^{\text{asc}} \sim N(0, \sigma^2 / w_i) \; , \tag{63}
$$

where $Y_i^{\text{asc}} = \log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}$, $X_i^{\text{asc}} = X_i^1 - X_i^2$, $w_i = \frac{Y_i^{(1)\text{obs}} Y_i^{(2)\text{obs}}}{Y_i^{(1)\text{obs}} + Y_i^{(2)\text{obs}}}$. We call QTL mapped by Eq 63 ascQTL.

Note that we can combine Eq 62 and 63 and solve them jointly in close form. But here we still prefer meta-analysis for two reasons: 1) it allows combining summary statistics across studies; and 2) it allows the over-dispersion in total and allele-specific read counts to be different which is more realistic in practice since total and allele-specific read counts may go through different pre-processing steps.

Since the inference of linear regression has analytical solution which only involves $X^T X$ and $X^T Y$, we can solve it quickly and in a parallel way as proposed by Matrix eQTL [Shabalin, 2012]. We sketch the pseudocode on calculating trcQTL and ascQTL estimates in matrix form in Supplementary Notes 13.

10.2 Meta-analysis for QTL mapping

Once we obtain estimated $\widehat{\beta}^{\text{trc}}$ and $\widehat{\beta}^{\text{asc}}$, we can use these estimates to approximate $\mathcal{L}^{\text{trc}}$ and $\mathcal{L}^{\text{asc}}$ in Eq 49. Specifically, when sample size is large,

$$\mathcal{L}^{\text{trc}}(Y_i^{\text{total}}|\beta) \approx N(\beta; \widehat{\beta}^{\text{trc}}, \text{se}(\widehat{\beta}^{\text{trc}})) \tag{64}$$

$$\mathcal{L}^{\text{asc}}(\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}|\beta) \approx N(\beta; \widehat{\beta}^{\text{asc}}, \text{se}(\widehat{\beta}^{\text{asc}})) \tag{65}$$

So that the joint likelihood, as factorized in Eq 48, is simply $N(\beta; \widehat{\beta}^{\text{trc}}, \text{se}(\widehat{\beta}^{\text{trc}})) \times N(\beta; \widehat{\beta}^{\text{asc}}, \text{se}(\widehat{\beta}^{\text{asc}}))$. As shown previously [Lee et al., 2016], maximizing the approximate joint likelihood is equivalent to inverse-variance meta-analysis, which takes the form

$$\widehat{\beta}^{\text{mix}} = \frac{w^{\text{trc}}\widehat{\beta}^{\text{trc}} + w^{\text{asc}}\widehat{\beta}^{\text{asc}}}{w^{\text{trc}} + w^{\text{asc}}} \tag{66}$$

$$\text{se}(\widehat{\beta}^{\text{mix}}) = \sqrt{\frac{1}{w^{\text{trc}} + w^{\text{asc}}}} , \tag{67}$$

where $w^{\text{trc}} = 1/\text{se}(\widehat{\beta}^{\text{trc}})^2$ and $w^{\text{asc}} = 1/\text{se}(\widehat{\beta}^{\text{asc}})^2$.

11 Inference procedure for multi-SNP model

With the simplification made in Supplementary Notes 10.1, the multi-SNP model can be written as

$$Y_i^{\text{trc}} = \mu_0 + \boldsymbol{X}_i^{\text{trc}}\beta + z_i \quad , z_i \quad \sim N(0, \sigma_0^2) \tag{68}$$

$$Y_i^{\text{asc}} = \quad \boldsymbol{X}_i^{\text{asc}}\beta + \epsilon_i^{\text{asc}}, \epsilon^{\text{asc}} \sim N(0, \sigma^2/w_i) \; . \tag{69}$$

11.1 Motivating two-step inference procedure

Here we focus on two inference problems under the multi-SNP model: 1) construct genetic predictor of expression; and 2) infer whether β_k is non-zero, *i.e.* causal SNP. Problem 1) is prediction problem in machine learning context and in terms of building genetic predictor, elastic net has been used for this task [Gamazon et al., 2015]. For problem 2), the inference problem is formulated into a Bayesian variable selection problem and efficient solvers such as susieR [Wang et al., 2019] and DAP-G [Lee et al., 2018] have been developed in the context of eQTL analysis.

However, the existing methods only use total read information (typically inverse normalized expression) and they assume the inversely normalized expression Y and genotype vector $\boldsymbol{X}$ follow $Y \sim N(\boldsymbol{X}\beta, \nu)$. The modeling assumption is very close to Eq 68, 69 but it requires equal variance in error term and shared intercept across all observations. To apply the existing tools, we need to bypass the gap between our model and their modeling assumption. For this reason, we propose a two-step inference procedure to perform inference for multi-SNP model. In step 1, we infer μ_0, σ_0^2, and σ^2 and transform the data such that they approximately follow $Y \sim N(\boldsymbol{X}\beta, \nu)$. And in step 2, we apply the transformed data to existing solvers for both prediction and fine-mapping problems.

11.2 Inferring μ_0, σ_0^2, and σ^2

To estimate σ^2 from Eq 69 is equivalent to estimate the mean squared error (MSE) of the model. To avoid overfitting β and underestimating MSE, we apply 4-fold cross-validation using LASSO to get effect size estimate $\widehat{\beta}^{\text{asc,lasso}}$. And $\widehat{\sigma}^2 = \frac{1}{N^{\text{asc}}-1}\|\boldsymbol{Y}^{\text{asc}} - \boldsymbol{X}^{\text{asc}}\widehat{\beta}^{\text{asc,lasso}}\|_2^2$ where N^{asc} is the number of allelic imbalance observations. Any alternative approach would be to treat β_j^{asc} as random effect and estimate σ^2 as one of the variance components. But here we implement the former.

We apply similar approach to estimate σ_0^2 using LASSO. To obtain μ_0, we fit Bayesian variable selection model using susieR with the total read count observations Y_i^{trc}, $\boldsymbol{X}_i^{\text{trc}}$, $i = 1, \cdots, N^{\text{trc}}$. The output intercept $\widehat{\mu}_0$ is the estimate.

11.3 Data transformation and inference

Once we obtain $\widehat{\mu}_0$, $\widehat{\sigma}_0^2$, and $\widehat{\sigma}^2$, we shift and re-scale the total and allelic imbalance observations by

$$\widetilde{Y}_i^{\text{trc}} = \frac{Y_i^{\text{trc}} - \widehat{\mu}_0}{\widehat{\sigma}_0}, \quad \widetilde{\boldsymbol{X}}_i^{\text{trc}} = \frac{\boldsymbol{X}_i^{\text{trc}}}{\widehat{\sigma}_0} \tag{70}$$

$$\widetilde{Y}_i^{\text{asc}} = \frac{Y_i^{\text{asc}}}{\widehat{\sigma}}, \qquad \widetilde{\boldsymbol{X}}_i^{\text{asc}} = \frac{\boldsymbol{X}_i^{\text{asc}}}{\widehat{\sigma}}, \tag{71}$$

where the transformed data (on the left-hand side) is used for downstream analysis on performing prediction and fine-mapping.

Specifically, we concatenate $\widetilde{\boldsymbol{Y}}^{\text{trc}}$ and $\widetilde{\boldsymbol{Y}}^{\text{asc}}$ into one vector $\boldsymbol{Y} \in \mathbb{R}^{(N^{\text{trc}}+N^{\text{asc}})\times 1}$ and similarly we concatenate $\widetilde{\boldsymbol{X}}^{\text{trc}}$ and $\widetilde{\boldsymbol{X}}^{\text{asc}}$ into one matrix $\boldsymbol{X} \in \mathbb{R}^{(N^{\text{trc}}+N^{\text{asc}})\times p}$ where p is the number of SNPs. To perform fine-mapping, we run `susieR::susie(X = X, Y = Y, intercept = FALSE, standardize = FALSE)` with `X` equal to $\boldsymbol{X}$ and `Y` equal to $\boldsymbol{Y}$. To build prediction model, we run `glmnet::glmnet(x = X, y = Y, lambda = lambda, alpha = 0.5)` with `x` equal to $\boldsymbol{X}$ and `y` equal to $\boldsymbol{Y}$. The hyperparamter `lambda` is selected by 5-fold nested cross-validation where at each `lambda` the 5-fold cross-validation are repeated three times and `lambda` that has lowest cross-validated mean squared error (averaged across three runs) is used. For comparison, we feed the part of total read count data ($\boldsymbol{X}^{\text{trc}}$, $\boldsymbol{Y}^{\text{trc}}$) directly into: 1) susieR for fine-mapping; and 2) elastic net for prediction. The procedure is the same but $\boldsymbol{X}$, $\boldsymbol{Y}$ are replaced by $\boldsymbol{X}^{\text{trc}}$, $\boldsymbol{Y}^{\text{trc}}$. And we call this total read count-only approach for fine-mapping and prediction as trcFine and trcPred.

12 Simulating RNA-seq reads

To examine the performance of the methods, we propose and implement a simulation scheme which generates total and allele-specific read counts. The simulation procedure includes three parts: 1) simulate gene body which will be aligned by reads; 2) randomly draw the causal variants; 3) simulate the number of reads for each haplotype transcript and place these reads to the gene body obtained in step 1). The total and allele-specific read counts can be directly read out from step 3) where the total read count is the sum of two haplotypic read counts and the allele-specific read count is the number of reads overlapping with heterozygous sites within gene body.

In step 1), we fix the length of gene body to be 10kbp. To simulate the heterozygous sites within gene body for each individual, we start with determining the position of polymorphic sites along gene body. We first sample the number of polymorphic sites from Binomial distribution, and then draw their positions and minor allele frequencies (MAFs). And finally, whether a polymorphic site is heterozygous in an individual is determined by Bernoulli distribution with MAF. The procedure is sketched as follow.

1. Number of polymorphic site within gene body $N_h \sim \text{Binomial}(L_{\text{gene}}, f^h)$, where $L_{\text{gene}} = 10^4, f^h = 0.001$.
2. Position P_m ($m = 1, \cdots, N_h$) of these polymorphic sites are sampled by $P_m \sim \text{Sample}(\{1, \cdots, L_{\text{gene}}\})$And the corresponding MAF f_m are drawn from $f_m \sim \text{Uniform}(\text{maf}^l, \text{maf}^h)$, where $\text{maf}^l = 0.05, \text{maf}^h = 0.3$.

3. For each individual i, whether the mth polymorphic site is heterozygous (denote as Z_{im}) is determined by $Z_{im} \sim \text{Bernoulli}(2f_m(1-f_m))$.

In step 2), the genetic effect equals to $e^{X_i^h \beta}$ (in single-SNP model) and $e^{\boldsymbol{X}_i^h \boldsymbol{\beta}}$ (in multi-SNP model). To do so, we need to obtain haplotype and effect size. For single-SNP model, we first sample MAF of the causal variants and obtain the two haplotypes of each individual by drawing from Bernoulli. For multi-SNP model, we use the 1000G phase3 genotypes of European individuals. In brief, we randomly select 200 genes on chromosome 22 and extract phased genotypes of 1Mbp cis-window surrounding the transcription start site of them (excluding variants with allele frequency < 0.01 or > 0.99). The genetic effect size, e^{β}, ranges among 1, 1.01, 1.05, 1.1, 1.25, 1.5, 2, 3 for single-SNP case. In multi-SNP case, the number of causal SNPs is sampled from 1, 2, 3 and the genetic effect ranges from 0.015 to 0.075 such that the heritability ranges approximately from 19.4% to 54.5%. The detailed procedure for sampling $e^{X^h\beta}$ and $e^{\boldsymbol{X}_i^h \boldsymbol{\beta}}$ is as follow.

- **Single-SNP scenario**:
  1. Sampling X_i^h: MAF of causal SNP $f^c \sim \text{Uniform}(\text{maf}^l, \text{maf}^h)$ and $X_i^h \sim \text{Bernoulli}(f^c)$ where $\text{maf}^l = 0.05, \text{maf}^h = 0.3$.
  2. Setting up β: fixed to 1, 1.01, ..., 2, 3.
- **Multi-SNP scenario**:
  1. Sampling $\boldsymbol{X}_i^h$: obtained from 1000G phased genotypes.
  2. Setting up $\boldsymbol{\beta}$: number of causal SNPs $\sim \text{Sample}(\{1, 2, 3\})$ and the genetic variation $v_g \sim \text{Uniform}(0.015, 0.075)$. The genetic effect of causal variants are determined by randomly partition the genetic variation and convert per-SNP genetic variation into effect size by $\beta_k = \sqrt{v_{g,k}/(2f_k(1-f_k))}$ where f_k is MAF of kth causal SNP.

In the step 3), the last step, we sample the reads coming from each of the haplotype transcripts. The procedure is as follow.

1. For individual i, sample library size $L_i \sim \text{NegativeBinomial}(\text{size}, \text{prob})$ where $\text{size} = 15, \text{prob} = 1.6 \times 10^{-7}$ (Negative Binomial follows parameterization in `rnbinom` in R).
2. And then, sample individual-specific baseline abundance $\theta_{0,i} \sim \text{Beta}$ where $\text{E}(\theta_{0,i})$ ranges among $5 \times 10^{-5}, 2.5 \times 10^{-5}, 1 \times 10^{-5}, 5 \times 10^{-6}, 2.5 \times 10^{-6}, 1 \times 10^{-6}$ and $\text{sd}(\theta_{0,i}) = \text{E}(\theta_{0,i})/4$ (so that the non-genetic variation is roughly $1/4^2 = 1/16$).
3. The actual relative abundance of haplotype h in individual i is $\theta_i^h = \theta_{0,i} e^{X_i^h \beta}$ or $\theta_i^h = \theta_{0,i} e^{\boldsymbol{X}_i^h \boldsymbol{\beta}}$
4. Sample actual read count for each haplotype: $Y_i^h \sim \text{NegativeBinomial}(\text{size}, \text{prob})$ where $\text{size} = 2L_i\theta_i^h, \text{prob} = \frac{2}{3}$. This corresponds to $\text{E}(Y_i^h) = L_i\theta_i^h$ and $\text{Var}(Y_i^h) = \frac{3}{2}\text{E}(Y_i^h)$.
5. Randomly place reads, Y_i^h in total, onto the corresponding gene body simulated in step 1) where the read is aligned to each position of gene body with equal probability.
6. Total count is $Y_i^{\text{total}} = Y_i^1 + Y_i^2$ and allele-specific count $Y_i^{(h)\text{obs}}$ is the number of reads (as part of Y_i^h) that overlaps with the heterozygous sites of the individual (indicated by $Z_{i\cdot}$).

13 Pseudocode on solving trcQTL and ascQTL in matrix form

We sketch the matrix operations for solving a grid of least squares problems $\boldsymbol{y}_k \sim \boldsymbol{x}_j$ for each pair of j, k where we let $Y = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_K]$ and $X = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$. To obtain nominal p-value, $K = 1$. For permutation procedure proposed in fastQTL [Ongen et al., 2015], K equals to the number of permutation and $\boldsymbol{y}_k$ is the kth permuted $\boldsymbol{y}$.

To ensure trcQTL and ascQTL ran on the same permuted $\boldsymbol{y}$, we perform permutation before removing low count observations. So that in each permutation, different individuals are removed by low-count filter. To account for this fact, we introduce mask $M \in \{0,1\}^{n \times K}$ where M_{ik} indicating if the ith individual is included in kth permutation.

For trcQTL, the corresponding least squares problem has intercept, as mentioned in Eq 62. The pseudocode to solve the grid of trcQTL problems for all cis-SNP of a gene is sketched in Algorithm 1 where $Y = \boldsymbol{Y}^{\text{trc}}$ for nominal pass and $Y_{\cdot k} = P_k \boldsymbol{Y}^{\text{trc}}$ with permutation matrix P_k for permutation pass.

Note that the pseudocode only requires basic matrix operation. The matrix operation is element-wise if not notice explicitly. The Einstein summation is represented by `einsum` with similar arguments as `numpy.einsum` in Python. For instance, `einsum('ij,jk→ik', A, B)` means that to take the inner product of the i row in A and k column in B as the element at ith row and jth column in the output matrix.

Similar to trcQTL, the corresponding least squares problem of ascQTL is weighted without intercept, as mentioned in Eq 63. The pseudocode to solve the grid of ascQTL problems for all cis-SNP of a gene is sketched in Algorithm 2 where $Y = \boldsymbol{Y}^{\text{asc}}$ for nominal pass and $Y_{\cdot k} = P_k \boldsymbol{Y}^{\text{asc}}$ with permutation matrix P_k for permutation pass. And W as the weight matrix should be permutate accordingly, *i.e.* $W_{\cdot k} = P_k \boldsymbol{w}$. And to obtain valid mixQTL estimates under permutation, P_k is required to be shared by trcQTL and ascQTL in permutation pass.

Note that both Algorithm 1 and Algorithm 2 are iteration free. And throughout the computation, only two-way tensors are involved explicitly so that the memory usage does not blow up.

Algorithm 2: Solve multiple least squares problems $y = bx + e$ with weight w in matrix form

```
Input  : $Y \in \mathbb{R}^{n \times K}$, $X \in \mathbb{R}^{n \times p}$, $M \in \{0,1\}^{n \times K}$, $W \in \mathbb{R}_+^{n \times K}$.
Output: $\widehat{B} \in \mathbb{R}^{K \times p}$ and se($\widehat{B}$) $\in \mathbb{R}^{K \times p}$ where $\widehat{B}_{kj}$, se($\widehat{B}_{kj}$) are estimates of $Y_{\cdot k} = B_{kj} X_{\cdot j} + \epsilon$ where
        data is weighted by $W_{\cdot k}$ and masked by $M_{\cdot k}$.
Function SolveMatrixLSwithWeight($Y$, $X$, $M$, $W$):
    $n$ = einsum('ik→k', $M$);
    $W$ = $WM$;
    $Y_{sqW}$ = $Y\sqrt{W}$;
    $Y$ = $YW$;
    $T$ = einsum('ij,ik→jk', $X$, $Y$);
    $S$ = $X^2$;
    $S$ = einsum('ij,ik→jk', $S$, $W$);
    $\widehat{B}$ = $T/S$;
    $Y_{sq}$ = einsum('ik,ik→k', $Y_{sqW}$, $Y_{sqW}$);
    $R_{sq}$ = $Y_{sq} - 2\widehat{B}T + \widehat{B}^2 S_{11}$;
    $\widehat{\sigma}$ = $\sqrt{R_{sq}/(n-1)}$;
    se($\widehat{B}$) = $\widehat{\sigma}/\sqrt{S}$;
    return $\widehat{B}$, se($\widehat{B}$)
End
```

14 Evaluating QTL mapping performance using eQTLGen results

To evaluate the performance of QTL mapping method, we treat eQTLGen [Võsa et al., 2018] as a silver standard, in the sense that eQTLs identified as positive in eQTLGen are treated as the true associations and the non-significant variant/gene pairs in eQTLGen are treated as true non-associations. Although 336 GTEx samples are included in eQTLGen analysis, they make up of only around 1.5% of total samples. So, eQTLGen results are unlikely driven by GTEx samples. And besides, GTEx v8 includes additional samples that are not included in eQTLGen. Therefore, eQTLGen is an approximately independent eQTL study with

Algorithm 1: Solve multiple least squares problems $y = a + bx + e$ in matrix form

```
Input  : $Y \in \mathbb{R}^{n \times K},\ X \in \mathbb{R}^{n \times p},\ M \in \{0,1\}^{n \times K}$.
Output: $\widehat{A}, \widehat{B}, \mathrm{se}(\widehat{A}), \mathrm{se}(\widehat{B}) \in \mathbb{R}^{K \times p}$ where $\widehat{A}_{kj}, \widehat{B}_{kj}, \mathrm{se}(\widehat{A}_{kj}), \mathrm{se}(\widehat{B}_{kj})$ are estimates of
        $Y_{\cdot k} = A_{kj} + B_{kj} X_{\cdot j} + \epsilon$ where data is masked by $M_{\cdot k}$.
Function SolveMatrixLSwithIntercept($Y, X, M$):
    $U$ = matrix(1, dim = dim($X$));
    $n$ = einsum('ik→k', $M$);
    $Y = YM$;
    $T_1$ = einsum('ij,ik→jk', $X$, $Y$);
    $T_2$ = einsum('ij,ik→jk', $U$, $Y$);
    $S_{11} = X^2$;
    $S_{11}$ = einsum('ij,ik→jk', $S_{11}$, $M$);
    $S_{22} = U^2$;
    $S_{22}$ = einsum('ij,ik→jk', $S_{22}$, $M$);
    $S_{12} = XU$;
    $S_{12}$ = einsum('ij,ik→jk', $S_{12}$, $M$);
    $\Delta = |S_{11} S_{22} - S_{12} S_{12}|$;
    $\widehat{B} = (S_{22} T_1 - S_{12} T_2)/\Delta$;
    $\widehat{A} = (S_{11} T_2 - S_{12} T_1)/\Delta$;
    $Y_{sq}$ = einsum('ik,ik→k', $Y$, $Y$);
    $R_{sq} = Y_{sq} - 2\widehat{B} T_1 - 2\widehat{A} T_2 + 2\widehat{B}\widehat{A} S_{12} + \widehat{B}^2 S_{11} + \widehat{A}^2 S_{22}$;
    $\widehat{\sigma} = \sqrt{R_{sq}/(n-2)}$;
    $\mathrm{se}(\widehat{B}) = \widehat{\sigma}\sqrt{S_{22}/\Delta}$;
    $\mathrm{se}(\widehat{A}) = \widehat{\sigma}\sqrt{S_{11}/\Delta}$;
    return $\widehat{A}, \widehat{B}, \mathrm{se}(\widehat{A}), \mathrm{se}(\widehat{B})$
End
```

much larger sample size (50-fold relative to GTEx v8) and diverse populations (predominantly Europeans along with other populations).

To simplify the analysis, we randomly select 100,000 eQTLGen cis-eQTLs (FDR ¡ 0.05) as the true associations in the silver standard. And we randomly collect 100,000 variant/gene pairs in eQTLGen with p-value ¿ 0.5 as the true non-associations. Among those variant/gene pairs in silver standard, 96,660 true associations and 78,691 true non-associations are included in both our mixQTL mapping pipeline and GTEx v8 analysis. So that we keep only these variant/gene pairs for downstream analysis.

14.1 Comparing the effective sample size

To compare the effective sample size between mixQTL and eQTL approaches, we performed analysis similar to [Loh et al., 2018]. Here, we utilize the fact that χ^2 statistic scales proportionally with the sample size, among those true associations. So, we can calculate the ratio χ^2_{mixQTL} over χ^2_{eQTL} for each truly associated variant/gene pair as the measure of effective sample size of mixQTL relative to eQTL approach. Specifically, we calculate the relative effective sample size using the true associations in the silver standard constructed above (as the proxy of true associations based on independent evidence). Note that the gain of power in mixQTL depends on the amount of allele-specific observations so we measured the average relative effective sample size as the median of the χ^2 ratio. Among the 96,660 variant/gene pairs collected as true associations in silver standard, we measured the median of χ^2_{eQTL} as 2.59 and the median of χ^2_{mixQTL} as 3.56. And the median of the ratio χ^2_{mixQTL} over χ^2_{eQTL} is 1.29. In other word, it suggests that the mixQTL approach (with 670 individuals) is equivalent to the eQTL approach with 863 individuals.

14.2 Drawing receiver operating characteristic and precision-recall curves

The ROC and PR curves are constructed using $-\log(p)$ as prediction score (higher means more likely to be causal). To simplify the calculation, we evaluate the performance measures at a grid of score cutoffs: 0.1, 0.2, ..., 1.9, 2, 2.2, ..., 2.8, 3, 4, ..., 50. For ROC curve, we calculate true positive rate and false positive rate at these cutoffs. And similarly, for PR curve, we calculate precision and power at these cutoffs.