Allele-Specific QTL Fine Mapping with PLASMA

Austin T. Wang,[1,2,3,*] Anamay Shetty,[3,4] Edward O'Connor,[3] Connor Bell,[3] Mark M. Pomerantz,[3] Matthew L. Freedman,[3,5,6] and Alexander Gusev[3,5,7,*]

Although quantitative trait locus (QTL) associations have been identified for many molecular traits such as gene expression, it remains challenging to distinguish the causal nucleotide from nearby variants. In addition to traditional QTLs by association, allele-specific (AS) QTLs are a powerful measure of *cis*-regulation that are concordant with traditional QTLs but typically less susceptible to technical/environmental noise. However, existing methods for estimating causal variant probabilities (i.e., fine mapping) cannot produce valid estimates from asQTL signals due to complexities in linkage disequilibrium (LD). We introduce PLASMA (Population Allele-Specific Mapping), a fine-mapping method that integrates QTL and asQTL information to improve accuracy. In simulations, PLASMA accurately prioritizes causal variants over a wide range of genetic architectures. Applied to RNA-seq data from 524 kidney tumor samples, PLASMA achieves a greater power at 50 samples than conventional QTL-based fine mapping at 500 samples, with more than 17% of loci fine mapped to within five causal variants, compared to 2% by QTL-based fine mapping, and a 6.9-fold overall reduction in median credible set size compared to QTL-based fine mapping when applied to H3K27AC ChIP-seq from just 28 prostate tumor/normal samples. Variants in the PLASMA credible sets for RNA-seq and ChIP-seq were enriched for open chromatin and chromatin looping, respectively, at a comparable or greater degree than credible variants from existing methods while containing far fewer markers. Our results demonstrate how integrating AS activity can substantially improve the detection of causal variants from existing molecular data.

Introduction

A major open problem in genetics is understanding the biological mechanisms underlying complex traits, which are largely driven by non-coding variants. A widely adopted approach for elucidating these regulatory patterns is the identification of disease variants that also modify molecular phenotypes (such as gene expression).[1–4] These variants, known as quantitative trait loci (QTLs), are typically single nucleotide polymorphisms (SNPs) that exhibit a statistical association with overall gene expression abundance.[5–8] Although QTL association analysis is now mature, it remains challenging for scientists to identify the precise variants that causally influence the molecular trait (as opposed to variants in linkage disequilibrium [LD] with causal variants), a task known as fine mapping.[9] Because only a small subset of QTL-associated markers are estimated to be causal,[10,11] direct experimental validation is prohibitive and has motivated statistical fine-mapping solutions.[12] The aim of statistical fine mapping is to quantify the probability of each marker being causal, allowing one to prioritize the most likely causal markers and thus formally quantify the effort needed for experimental validation. Recent statistical fine-mapping methods operate on summary QTL statistics and can handle multiple causal variants by modeling the local LD structure.[13–16] These models have two outputs to help guide the prioritization of putative causal SNPs. First, a Posterior Inclusion Probability (PIP), which corresponds to the marginal probability of causality for the given marker, is calculated for each marker. Second, a n%-confidence credible set is created: a set of markers with an n% probability of containing all the causal markers. Although QTL studies have enough power to identify thousands of associations, they are typically insufficient for fine mapping below dozens of credible variants, even for very large studies.[5,17] The need for large studies severely limits QTL analyses of expensive assays such as ChIP or single-cell RNA-seq or of difficult-to-collect tissues.

Here, we sought to improve molecular fine mapping by leveraging an intra-individual allele-specific (AS) signal, which is a measure of *cis*-regulatory activity that is independent of total inter-individual variation. For heterozygous variants residing in expressed exons, it is often possible to map expressed reads to each allele and quantify the extent to which molecular activity is allele specific.[6,18–21] AS analysis allows for a precise comparison of the effects on molecular activity that are specific to each allele (*cis*-effects), while controlling for effects affecting both alleles (*trans*-effects). Thus, AS data are inherently less noisy than regular QTL data, which captures total phenotype regardless of source. Allele-specific data has furthermore been used to quantify *cis*-regulation, implying that both AS and regular QTL features represent the same underlying *cis*-regulatory patterns.[22] Several methods have recently been developed to robustly identify asQTLs,[19,20,23] but the calculated association statistics follow a different distribution than QTL summary statistics and cannot be directly integrated into existing fine-mapping software to produce valid posterior measures and credible sets.

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02142, USA; [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA; [3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA; [4]Cambridge University, Cambridge CB2 1TN, UK; [5]The Eli and Edythe L. Broad Institute, Cambridge, MA 02142, USA; [6]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA; [7]Brigham & Women's Hospital, Division of Genetics, Boston, MA 02215, USA
*Correspondence: atwang@mit.edu (A.T.W.), alexander_gusev@dfci.harvard.edu (A.G.)
https://doi.org/10.1016/j.ajhg.2019.12.011.

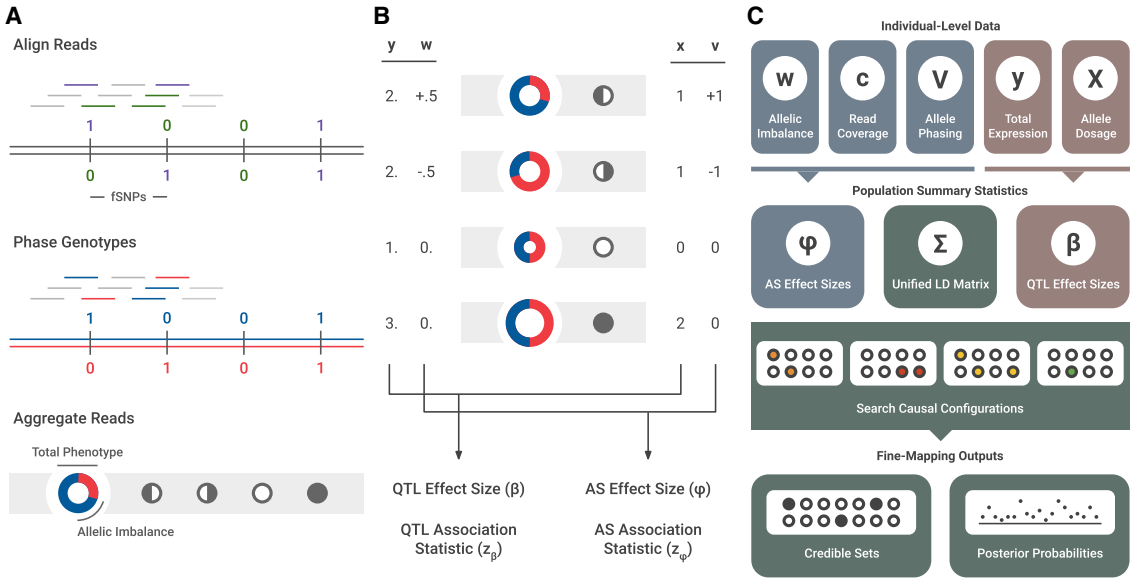

Figure 1. Overview of the PLASMA Method

(A) Pre-processing of sequence-based data. First, reads are mapped to the sample's genotype. Reads intersecting markers are colored. Then, the sample's genotype is phased. Reads intersecting heterozygous markers can then be mapped to a particular haplotype. Lastly, reads across the locus are aggregated in an allele-specific manner. To visualize this data, the expression is represented by a ring chart, and the genotypes by pedigree symbols. In the ring chart, the diameter signifies the total read count, and the colors signify the proportion of reads coming from each haplotype. For the pedigree symbols, a white circle signifies a wild-type homozygote, a shaded circle signifies an alternative homozygote, and a half-shaded circle signifies a heterozygote. In heterozygotes, the direction of shading corresponds to the direction of heterozygosity (phasing).

(B) Visual representation of QTL and AS statistics under a single causal variant, where the alternative allele increases expression. The total expression (y) is determined by the allelic dosage (x), whereas the allelic imbalance (w) is determined by the phasing (v). These two sets of data are used to calculate QTL and AS association statistics (z_β and z_ϕ).

(C) Diagram of PLASMA's fine-mapping process. First, QTL and AS statistics are calculated from read data. Then, these statistics, along with an LD matrix, are used to generate probabilities for causal configurations. By searching through the space of these causal configurations, the model produces credible sets and posterior probabilities for each marker.

To combine the established statistical models of QTL analysis with the power of AS analysis, we introduce PLASMA (Population Allele-Specific Mapping), a novel fine-mapping method that gains power from both the number of individuals and the number of allelic reads per individual. By modeling each locus across individuals in an allele-specific and LD-aware manner, PLASMA achieves a substantial improvement over existing fine-mapping methods with the same data. We demonstrate through simulations that PLASMA successfully detects causal variants over a wide range of genetic architectures. We applied PLASMA to diverse RNA-seq data and ChIP-seq data, which showed a significant improvement in power over conventional QTL-based fine mapping.

Material and Methods

Overview of PLASMA

PLASMA's inputs are determined from a given individual-level sequencing-based molecular phenotype (gene or peak) and the corresponding local genotype SNP data (Figure 1A). For each sample, we assumed the variant data were phased into haplotypes and expression reads had been mapped to each variant. Reads intersecting heterozygous markers (signified as fSNPs, or feature SNPs, indicated with green or purple on the figure) were then assigned to a particular haplotype, indicated as blue or red on the figure. These reads were then aggregated in a haplotype-specific manner to produce a total expression phenotype and an allelic imbalance phenotype. This aggregation of reads is analogous to the way existing methods such as RASQUAL and WASP calculate allelic fractions and total fragment counts.[19,20] The total expression phenotype (y) is simply the total number of mapped reads. The allelic imbalance phenotype (w) is defined as the log read ratio between the haplotypes. This log-odds-like phenotype has previously been used to analyze asQTL effect sizes, showing consistency with conventional QTL analysis.[22] To mitigate the effects of mapping bias, we ran state-of-the-art mapping bias and QC pipelines on all RNA-seq and ChIP-seq data prior to analysis.[19]

PLASMA integrates two statistics computed for each marker to perform fine mapping: a QTL association statistic (z_β) based on the total phenotype and an AS association statistic (z_ϕ) based on the allelic imbalance phenotype. Figure 1B shows how a causal marker influences total expression and allelic imbalance and how this effect influences the statistics for the marker. Here, the causal marker's alternative allele causes higher expression compared to that of the wild-type (WT) allele. Increasing the dosage (x) of the alternative allele increases the total expression (y) at the locus. The effect size (β), consistent with a typical QTL analysis, quantifies the association between a marker's allelic dosage and the total expression at the locus with a linear relationship with residuals ϵ:

$$\mathbf{y} = \mathbf{x}_i\beta_i + \boldsymbol{\epsilon}_i \qquad \text{(Equation 1)}$$

From this effect size, PLASMA calculates z_β, the QTL association statistic. Note that this statistic is not dependent on haplotype-specific data.

On the other hand, looking at the heterozygotes, the haplotype possessing the alternative allele has a higher expression than the haplotype possessing the wild-type allele. In other words, the direction of imbalance of expression (w) is the same as the phasing (v) of the allele. The ϕ effect size quantifies the association between a marker's phasing with the imbalance of expression. An important departure from existing methods is that PLASMA models a linear relationship between the phase of a causal marker and the log read ratio, rather than directly relating the genotype to the allelic fraction in a non-linear manner with residuals ζ:

$$\mathbf{w} = \mathbf{v}_i\phi_i + \boldsymbol{\zeta}_i \qquad \text{(Equation 2)}$$

To calculate the AS association statistic z_ϕ, PLASMA models the quality of each sample, taking into account each sample's read coverage and read overdispersion (Figure 1C).

These QTL and AS association statistics, together with the local LD matrix, are then jointly used to fine map the locus (Figure 1C). Since PLASMA models both z_β and z_ϕ as a linear combination of genotypes, z_β and z_ϕ have identical LD (see Supplemental Material and Methods for proof). PLASMA assumes that the QTL and AS statistics measure the same underlying *cis*-regulatory signal and are thus expected to have the same direction and same causal variants (but see Discussion for possible model violations). Although they both measure regulatory effects, the two statistics have independent noise because the haplotype-level variance within individuals is considered only in AS analysis, allowing them to be used jointly in fine mapping. Furthermore, PLASMA accepts, as a hyperparameter, a correlation between QTL and AS effects, allowing the two sets of statistics to utilize a joint probability distribution (though our analyses show that setting this hyperparameter to zero yields the most power). The distribution is used to assign a probability to a given causal configuration, a binary vector signifying the causal status of each marker in the locus. Although the correlation between QTL and AS causal effects can vary based on the hyperparameter specification, PLASMA assumes that the AS and QTL phenotypes have the same causal variants. PLASMA searches through the space of possible causal configurations, within a constraint on the number of causal variants. This procedure is related to that in CAVIAR, CAVIARBF, and FINEMAP,[13–15] but generalized to the two correlated expression phenotypes. From these scored configurations, PLASMA computes a posterior inclusion probability (PIP) for each marker, indicating the marginal probability that a marker is causal, and a ρ-level credible set containing the causal variant with ρ probability.

Modeling QTL and AS Summary Statistics

Marginal QTL effect sizes for a given locus are calculated under the conventional linear model of total gene expression, with the allelic dosage ($\mathbf{x}$) as the independent variable and the total expression ($\mathbf{y}$) as the independent variable. Let us consider a QTL study of a given locus with n individuals and m markers. Let $\mathbf{y}$ be an ($n \times 1$) vector of total expression across the individuals, re-centered at zero. Given a marker i, let $\mathbf{x}_i$ be a zero-recentered vector of dosage genotypes. The genetic effect β_i of marker i on total gene expression is defined as follows:

$$\mathbf{y} = \mathbf{x}_i\beta_i + \boldsymbol{\epsilon}_i \qquad \text{(Equation 3)}$$

The empirical value of β_i is determined with the maximum likelihood estimator, equivalent to the ordinary-least-squares linear regression estimator:

$$\widehat{\beta}_i = \left(\mathbf{x}_i^\top \mathbf{x}_i\right)^{-1}\mathbf{x}_i^\top \mathbf{y} \qquad \text{(Equation 4)}$$

The QTL summary statistic (Wald statistic) for marker i is defined as:

$$\widehat{z}_{\beta,i} = \frac{\widehat{\beta}_i}{\sqrt{\left(\mathbf{x}_i^\top \mathbf{x}_i\right)^{-1}\widehat{\sigma}^2_{y,i}}} \qquad \text{(Equation 5)}$$

where $\widehat{\sigma}^2_{y,i}$ is calculated from the residuals.

AS effect sizes are calculated under a weighted linear model, with the phasing ($\mathbf{v}$) as the independent variable and the allelic imbalance ($\mathbf{w}$) as the dependent variable. PLASMA models allele-specific expression under the observation that a *cis*-regulatory variant often has a greater influence on the gene allele of the same haplotype. A marker's phase v is 1 if haplotype A contains the alternative marker allele, -1 if haplotype B contains the alternative marker allele, and 0 if the individual is homozygous for the marker. Let w be the log expression ratio between haplotypes A and B, ϕ_i be the AS effect size of variant i, and ζ_i be the residual, interpreted as the log baseline expression ratio between haplotypes A and B. Additionally, a sampling error $\tau_j = \widehat{w}_j - w_j$ is defined for each individual, quantifying the quality of data from the sample. The genetic effect of marker i on allele-specific expression is as follows:

$$\widehat{\mathbf{w}} = \mathbf{v}_i\phi_i + \boldsymbol{\zeta}_i + \boldsymbol{\tau} \qquad \text{(Equation 6)}$$

Experimentally derived AS data, such as RNA-seq data, yield reads that are mapped to a particular haplotype. For a given individual j, let $c_{A,j}$ be the allele-specific read count from haplotype A. The allele-specific read count is modeled with a beta-binomial distribution, given the total mapped read count c_j:

$$c_{A,j} \sim \mathrm{BB}\left(\alpha_j, \beta_j, c_j\right) \qquad \text{(Equation 7)}$$

This beta binomial model is used to estimate the variance of the sampling error τ_j:

$$\widehat{\sigma}^2_{c,j} = \frac{2}{c_j}\left(1 + \cosh\left(\widehat{w}^*_j\right)\right)\left(1 + \rho_{e,j}\left(c_j - 1\right)\right) \qquad \text{(Equation 8)}$$

where $\rho_{e,j}$ is the overdispersion and w^*_j is an adjusted estimator of w_j to reduce the bias of $\widehat{\sigma}^2_{c,j}$. (Full derivation in Supplemental Material and Methods).

Due to heteroscedasticity among individuals, the AS effect size ϕ_i is estimated in a weighted manner, giving larger weights to individuals with lower estimated sampling error. Given individual j, the weight for j is set as the inverse of the estimated sampling error variance:

$$\omega_j = \frac{1}{\widehat{\sigma}^2_{c,j}} \qquad \text{(Equation 9)}$$

Let weight matrix $\boldsymbol{\Omega}$ be a diagonal matrix with $\boldsymbol{\Omega}_{j,j} = \omega_j$. We use the weighted-least-squares estimator for ϕ_i:

$$\widehat{\phi}_i = \left(\mathbf{v}_i^\top \boldsymbol{\Omega}\mathbf{v}_i\right)^{-1}\boldsymbol{v}_i^\top \boldsymbol{\Omega}\widehat{\mathbf{w}} \qquad \text{(Equation 10)}$$

With this estimator, the AS association statistic for marker i is calculated as the AS effect size divided by the estimated variance

of the effect size (full derivation in Appendix and Supplemental Material and Methods):

$$\hat{z}_{\phi,i} = \frac{\hat{\phi}_i}{\sqrt{\left(\mathbf{v}_i^\top \boldsymbol{\Omega} \mathbf{v}_i\right)^{-2}\left(\left(\mathbf{v}_i^\top \boldsymbol{\Omega}^2 \mathbf{v}_i\right)\hat{\sigma}^2_{w,i} + \mathbf{v}_i^\top \boldsymbol{\Omega} \mathbf{v}_i\right)}} \quad \text{(Equation 11)}$$

In the case of computationally phased data, there may exist phasing errors that would decrease the accuracy of the estimated effect sizes ($\hat{\phi}$). With imperfect phasing, the observed phasing $\hat{\mathbf{v}}_i$ may differ from the true phasing $\mathbf{v}_i$, a modified equation may be used to calculate the AS z-score given the per-SNP probability of mis-phasing ψ_i:

$$\hat{z}_{\phi,i} = \frac{\hat{\phi}_i}{\sqrt{\left(\hat{\mathbf{v}}_i^\top \boldsymbol{\Omega} \hat{\mathbf{v}}_i\right)^{-2}\left(\left(\hat{\mathbf{v}}_i^\top \boldsymbol{\Omega}^2 \hat{\mathbf{v}}_i\right)\left(\hat{\sigma}^2_{w,i} + 4\psi_i\phi_i^2\right) + \hat{\mathbf{v}}_i^\top \boldsymbol{\Omega} \hat{\mathbf{v}}_i\right)}} \quad \text{(Equation 12)}$$

Fine-mapping simulation results under imperfect phasing are presented in Supplemental Material and Methods and in Figure S5.

Inference of Credible Sets and Posterior Probabilities

PLASMA defines a joint generative model for total (QTL) and haplotype-specific (AS) effects on expression. Let $\hat{\mathbf{z}}$ be the combined vector with dimension 2m of AS association statistics and QTL association statistics:

$$\hat{\mathbf{Z}} = \begin{bmatrix} \hat{\mathbf{Z}}_\phi \\ \hat{\mathbf{Z}}_\beta \end{bmatrix} \quad \text{(Equation 13)}$$

Let $\mathbf{R}_z$ be the genotype LD matrix, and $r_{\beta\phi}$ be a hyperparameter describing the overall correlation between the QTL and AS summary statistics calculated across all loci. Let the combined correlation matrix $\mathbf{R}$ as:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_z & r_{\beta\phi}\mathbf{R}_z \\ r_{\beta\phi}\mathbf{R}_z & \mathbf{R}_z \end{bmatrix} \quad \text{(Equation 14)}$$

The joint z-scores are modeled under a multivariate normal distribution, with covariance $\mathbf{R}$:

$$\hat{\mathbf{z}} \sim \mathcal{N}_{2m}(\mathbf{z}, \mathbf{R}) \quad \text{(Equation 15)}$$

PLASMA utilizes a likelihood function that gives the probability of statistics $\hat{\mathbf{z}}$, given a causal configuration. Let a causal configuration $\mathbf{c}$ be a vector of causal statuses corresponding to each marker, with 1 being causal and 0 being non-causal. PLASMA assume that the causal configuration is the same for the QTL and AS signals.

Let hyperparameters $\sigma^2_{c,\phi}$ and $\sigma^2_{c,\beta}$ be the variance of AS and QTL causal effect sizes, respectively. Furthermore, let the jointness parameter $r_{c,\beta\phi}$ be the underlying correlation of the causal QTL and AS effect sizes. (This is not to be confused with $r_{\beta\phi}$, which concerns the correlation between the observed association statistics. See Supplemental Material and Methods for a mathematical relationship between these two hyperparameters.) These three hyperparameters are closely related to the heritability of gene expression (see Supplemental Material and Methods). Let $\Sigma_\mathbf{c}$ be the covariance matrix of causal effect sizes given a causal configuration:

$$\sum\nolimits_c = \begin{bmatrix} \mathrm{diag}(\mathbf{c})\sigma^2_{c,\phi} & \mathrm{diag}(\mathbf{c})r_{c,\beta\phi}\sigma_{c,\phi}\sigma_{c,\beta} \\ \mathrm{diag}(\mathbf{c})r_{c,\beta\phi}\sigma_{c,\phi}\sigma_{c,\beta} & \mathrm{diag}(\mathbf{c})\sigma^2_{c,\beta} \end{bmatrix} \quad \text{(Equation 16)}$$

PLASMA's likelihood for a causal configuration is defined as:

$$\mathcal{L}(\mathbf{c}; \hat{\mathbf{z}}) = \mathcal{N}_{2m}\left(0, \mathbf{R} + \mathbf{R}\sum\nolimits_\mathbf{c}\mathbf{R}\right) \quad \text{(Equation 17)}$$

Let γ be the prior probability that a single variant is causal and $1-\gamma$ as the probability that a variant is not causal. The prior probability of a configuration consisting of m variants is defined as:

$$P(\mathbf{c}) = \prod_{i=1}^{m} \gamma^{c_i}(1-\gamma)^{1-c_i} \quad \text{(Equation 18)}$$

With the prior and likelihood, the posterior probability of a causal configuration, normalized across the set of all possible configurations, $\mathbb{C}$ is calculated as:

$$P(\mathbf{c} \mid \hat{\mathbf{z}}) = \frac{P(\hat{\mathbf{z}} \mid \mathbf{c})P(\mathbf{c})}{\sum_{\mathbf{c}^* \in \mathbb{C}} P(\hat{\mathbf{z}} \mid \mathbf{c}^*)P(\mathbf{c}^*)} \quad \text{(Equation 19)}$$

PLASMA defines the ρ-level credible set $\mathbb{K}$ as the smallest set of markers with a ρ_c probability of including all causal markers. Let $\mathbb{C}_\mathbb{K}$ be the set of all causal configurations whose causal markers is a subset of $\mathbb{K}$, excluding the null set. The credible set confidence level ρ_c is calculated as the sum of the probabilities of the configurations in $\mathbb{C}_\mathbb{K}$:

$$\rho_c = \sum_{\mathbf{c} \in \mathbb{C}_\mathbb{K}} \Pr(\mathbf{c} \mid \hat{\mathbf{z}}) \quad \text{(Equation 20)}$$

Additionally, PLASMA defines a marker's posterior inclusion probability (PIP) as the probability that a single given marker is causal, marginalized over all other markers. This probability is calculated by summing over all configurations containing the marker.

To reduce the number of configurations to evaluate in the case of multiple causal variants, PLASMA uses the heuristic that configurations with significant probabilities tend to be similar to each other. PLASMA uses a shotgun stochastic search procedure to find all configurations with a significant probability. For each iteration of the algorithm, the next configuration is drawn randomly from the neighborhood of similar configurations, weighted by the posterior probability of each candidate. The search is terminated under the presumption that all configurations with nonzero probability have been uncovered.

Given the large number of configurations evaluated, it is impractical to calculate the best possible credible set satisfying ρ_c. Instead, PLASMA uses a greedy approximation algorithm. At each step, before ρ_c is reached, the algorithm adds the marker that increases the confidence the most.

The Jointness Parameter in PLASMA

Although PLASMA always assumes the same causal variants for QTL and AS, the correlation between QTL and AS causal effects can be set in PLASMA-J with a jointness hyperparameter $r_{c,\beta\phi}$. A high value (near 1) assumes that the QTL and AS causal effects tend to be consistent in magnitude, while a low value (near zero) assumes more disparity. Note that this is unrelated to the choice of causal variants, and PLASMA assumes that QTL and AS share the same causal variants regardless of the jointness parameter.

A previous analysis comparing QTL effects with a similar formulation of AS effects has uncovered a highly nonlinear relationship, especially with QTL effects calculated using untransformed total expression data.[22] As a further complication, this relationship between QTL and AS effects is shown to be highly dependent on allele frequency. Thus, even under the assumption that QTL and

AS signals share a causal variant, there is no guarantee of a strong linear correlation between QTL and AS effect sizes. Due to this uncertainty, the jointness parameter to zero by default, making no assumption on the relationship between QTL and AS effect sizes.

To empirically evaluate the effect of the jointness parameter on fine-mapping performance, PLASMA-J was run with different values of the jointness parameter on simulated loci. Figure S2 shows the distribution of PLASMA-J credible sets with different values of jointness, ranging from 0 to 0.99. In the one causal variant case, results are largely invariant to the parameter below a value of 0.99.

Generation of Simulated Loci

Genotype data were sampled from phased SNP data using the CEU population in the 1000 Genomes Project. First, a contiguous section of markers is randomly chosen. Next, a random selection of samples are randomly selected from the section. The genotypes corresponding to the chosen samples yield two haplotype matrices, denoted as $\mathbf{H}_a$ and $\mathbf{H}_b$.

Among the markers, the desired number of causal markers is randomly selected. In the case of multiple causal variants, each causal marker is assigned a relative effect size, sampled from a normal distribution with zero mean and unit variance. For each individual, the ideal un-scaled gene expression for each haplotype $\mathbf{q}_a$ and $\mathbf{q}_b$ is determined by multiplying the relative effect sizes with each haplotype matrix.

Read count data are simulated with this haplotype-specific expression. In real data, only a fraction of the reads can be mapped to a specific haplotype. Due to this difference between total reads and mapped reads, the allelic imbalance and the total read count (QTL) are calculated separately.

To calculate total read count data, the total ideal un-scaled expression $\mathbf{q}_t$ is defined as $\mathbf{q}_a + \mathbf{q}_b$, the sum of the haplotype-specific un-scaled gene expression. Gaussian-distributed noise is then added so that the variance of $\mathbf{q}_t$ is consistent with the total variance across samples as specified by the QTL heritability. Finally, this final expression is scaled so that the total expression across samples is of unit variance. Total read counts are not explicitly generated, since a multiplicative factor across samples does not influence the QTL association statistics calculated by the model. This is reflective of typical QTL study protocols which aggressively rank/quantile normalize the data to fit a normal distribution.

To calculate allele-specific read counts, heritability, mean read coverage, and the total variance of the AS phenotype are taken into account. The ideal allelic imbalance phenotype is determined as $\text{logit}(\mathbf{q}_a / \mathbf{q}_b)$ (calculated element-wise). Gaussian-distributed noise is then added so that the signal-to-noise ratio of the phenotype's variance is consistent with the specified AS heritability. This noisy phenotype is then scaled to the specified total variance. The read coverage for each sample is then drawn from a Poisson distribution, given the mean read coverage. Lastly, allele-specific read counts are generated from these phenotypes, with the counts for each sample being drawn from a beta-binomial distribution.

Comparison of Existing Models with PLASMA

Our analyses benchmark PLASMA against existing fine-mapping methods. Two distinct versions of PLASMA are tested, "PLASMA-J" and "PLASMA-AS." The PLASMA-J (Joint-Independent) version looks at both AS and QTL statistics, assuming a shared set of AS and QTL causal variants, and also that the AS and QTL causal effects are uncorrelated. The "PLASMA-AS" version is restricted to only AS data. As a baseline, we compare PLASMA to a QTL-Only version of PLASMA and to the CAVIAR method (expected to be equivalent to PLASMA QTL-Only).[13] The behavior and performance of CAVIAR is representative of similar QTL-based methods such as CAVIARBF, FINEMAP, and PAINTOR without functional annotation data.[14–16] The versions of PLASMA are furthermore compared against the only other publicly released fine-mapping method (to our knowledge) that integrates AS data described in the pre-print of Zou et al.[24] This unnamed method, denoted as "AS-Meta," utilizes the association between SNP heterozygosity and a binary indicator of allelic imbalance. By binarizing allelic imbalance, AS-Meta is expected to lose power relative to treating imbalance as a quantitative phenotype but may be more robust to spurious AS signal. Furthermore, AS-Meta utilizes only indicators of heterozygosity, rather than marker phasing. AS-Meta can therefore be used with unphased genotypes, but at the expense of being unable to leverage the direction of the allelic effect. Lastly, as an additional comparison with an AS-based method, we analyze the performance of RASQUAL, a method for inferring allele-specific genetic effects using both allelic and total expression signal. Note that RASQUAL computes allele-specific effect sizes for each marker only and is not intended to compute credible sets or posterior marginal probabilities. Traditional fine mapping on RASQUAL statistics is made possible by converting RASQUAL chi-square statistics back to quasi-z-scores with sign based on the direction of the RASQUAL effect-size. These statistics are then fed into standard QTL-only fine mapping to obtain credible sets and posterior probabilities. We denote the modification of RASQUAL as "RASQUAL+." This process is comparable to fine mapping using a combined AS/QTL effect, rather than modeling QTL and AS effects separately.

Quality Control of Genotype Data

For TCGA data, germline genotype calls are downloaded from the Genomic Data Commons. For PrCa ChIP samples, germline genotypes are called from blood. Genotypes are then imputed to the Haplotype Reference Consortium[25] using the Michigan Imputation Server[26] and restricted to variants with INFO greater than 0.9 and MAF greater than 0.01. Variants are further restricted to QC-passing SNPs from Moyerbrailean et al.[1] which represent common, well-mapped variants from the 1000 Genomes project.

Quality Control of RNA-Seq Data

Raw RNA-seq BAM files are downloaded from the Genomic Data Commons. Initial RNA-seq mapping and alignment are performed following TCGA parameters for the STAR aligner.[27] Mapping bias is accounted for by re-mapping using the WASP pipeline[19] and the STAR aligner with the same parameters. Reads are randomly de-duplicated as recommended by the WASP pipeline.

Somatic copy number calls are downloaded from FireBrowse and local beta-binomial overdispersion parameters are estimated for each contiguous region of copy number change.

Quality Control of ChIP-Seq Data

ChIP-seq reads are aligned using bwa and default parameters,[28] and peaks are called using MACS2 and default parameters (with DNA-seq input provided as control).[29] Peaks are then unified across all samples. Mapping bias is accounted for by re-mapping using the WASP pipeline and the bwa aligner with the same parameters. Reads are randomly de-duplicated as recommended by the WASP pipeline. Beta binomial overdispersion parameters are

estimated globally for each sample as somatic copy number was expected to be minimal.

Allele-Specific Quantification

The StratAS algorithm is used to quantify allele-specific signal and identify initially significant features for fine mapping.[23] For each peak/gene (the feature) and individual, all reads at heterozygous SNPs in the feature are aggregated to compute the haplotype-specific read counts and summed across the two haplotypes of each individual to compute the QTL read counts. Each QC passing variant within 100 kb of the feature are then tested for an allele-specific association with the feature and features significant at a genome-wide false discovery rate (FDR) of 5% are retained for fine mapping.

Functional Enrichment Analysis

For QTLs fine mapped from RNA-seq, we select regions of accessible chromatin in the most relevant tissue as reference the functional feature, reasoning that high-confidence causal variants should be more abundant in accessible regions. For QTLs fine mapped from ChIP-seq, we select chromosome looping anchors from Hi-ChIP in the relevant tissue as the reference functional feature, reasoning that high-confidence causal variants should be more abundant in regions that are in conformation with promoters.

Enrichment is then estimated by computing the proportion of markers in credible sets that intersect with the functional feature. Controls are calculated as the intersection between all tested markers and the functional feature. Odds ratios and p values are computed with Fisher's exact test.

Results

Simulation Framework

We evaluate PLASMA with a framework that simulates the expression of whole loci in an allele-specific manner. This simulation framework jointly simulates total reads and allele-specific read counts, under given values of the number of causal variants, the QTL heritability, the AS heritability, the variance of the AS phenotype across samples, and the expected read coverage (see Material and Methods). The variance and heritability of the AS phenotype are handled by two separate parameters, where the former describes the total spread of allelic imbalance and the latter specifies the fraction of the variance that is due to genetic effects. This allows us to investigate cases where a significant amount of observed imbalance is caused by non-genetic variance in the allelic expression. To quantify the total variance of the AS phenotype in the population, we define the "standard allelic deviation" (d) as the standard deviation of the AS phenotype w, quantified on the allelic fraction scale (between 0.5 and 1). Importantly, this quantity is independent of the genetic effect, which is controlled by the heritability parameter. Simulations were performed using real phased haplotype data from the 1000 Genomes Project European samples. Parameter settings for simulation analyses are shown in Table S1.

As the performance of standard QTL association models is well established, we first focused on performance of our proposed AS statistic. Figure S3A shows how the mean z_ϕ varies as a function of standard allelic deviation and mean read coverage at a fixed AS heritability of 0.5. Second, Figure S3B shows how the mean z_ϕ varies as a function of standard allelic deviation and heritability with mean coverage fixed at 100. The statistic is the greatest at high read coverage and high heritability, consistent with the degree of experimental and intrinsic signal available to the model. These results hold even at low AS variance ($d = 0.6$) and show that PLASMA does not conflate high AS variance (standard allelic deviation) with high signal (coverage or heritability). This robustness to variance in the AS phenotype makes the model resistant to false positives driven by non-genetic sources of allelic variance. At very high variance ($d > 0.8$), z_ϕ shows a sharp decrease. This decrease in signal is due to an increase in the sampling error of the AS phenotype (w) at high overall variance, as shown in Equation A27 (see Supplemental Material and Methods for a mathematical relationship between total variance and sampling error).

Comparison with Existing Methods in Simulation

First, we evaluate how well each PLASMA prioritizes candidate causal markers using simulated loci with one causal variant, compared to existing QTL and AS-based methods. We define the "inclusion curve" for each model, where markers are ranked by posterior probability and added one by one to a cumulative set (note that this set is not dependent on the definition of a credible set). The x axis represents the cumulative number of markers chosen, and the y axis represents the "inclusion rate," the proportion of true causal markers among the chosen markers. Figures 2A and 2D show inclusion plots at low and high AS variance, respectively. As expected, FINEMAP, QTL-Only, and the CAVIAR methods are indistinguishable and do not vary with AS variance. Due to this similarity in results, FINEMAP is used as the primary QTL-based methods in subsequent analyses. Furthermore, PLASMA-J and PLASMA-AS perform similarly at both levels of AS variance. Additionally, AS-Meta's performance exhibits a dependency on the degree of AS variance. Lastly, RASQUAL+ at high AS variance does significantly improve over QTL-based methods, but not as well as PLASMA. At low AS variance (with same amount of signal and noise), RASQUAL+ performs considerably worse, indicating that RASQUAL+ is more sensitive to the genetic architecture of the locus than PLASMA is.

Second, we evaluate the ability of each model to rule out likely non-causal markers in simulated loci with one causal variant. We conduct a direct comparison of the distributions of the 95% confidence credible sets, with smaller sets indicating higher specificity. Figures 2A and 2D show distribution plots at low and high AS variance, respectively. At low variance, PLASMA-J offers the smallest median credible 95% set size (3.0), followed by PLASMA-AS (3.0), then AS-Meta (55.0), and lastly the QTL-based methods: FINEMAP (89.0), CAVIAR (89.0), and QTL-Only (91.0).

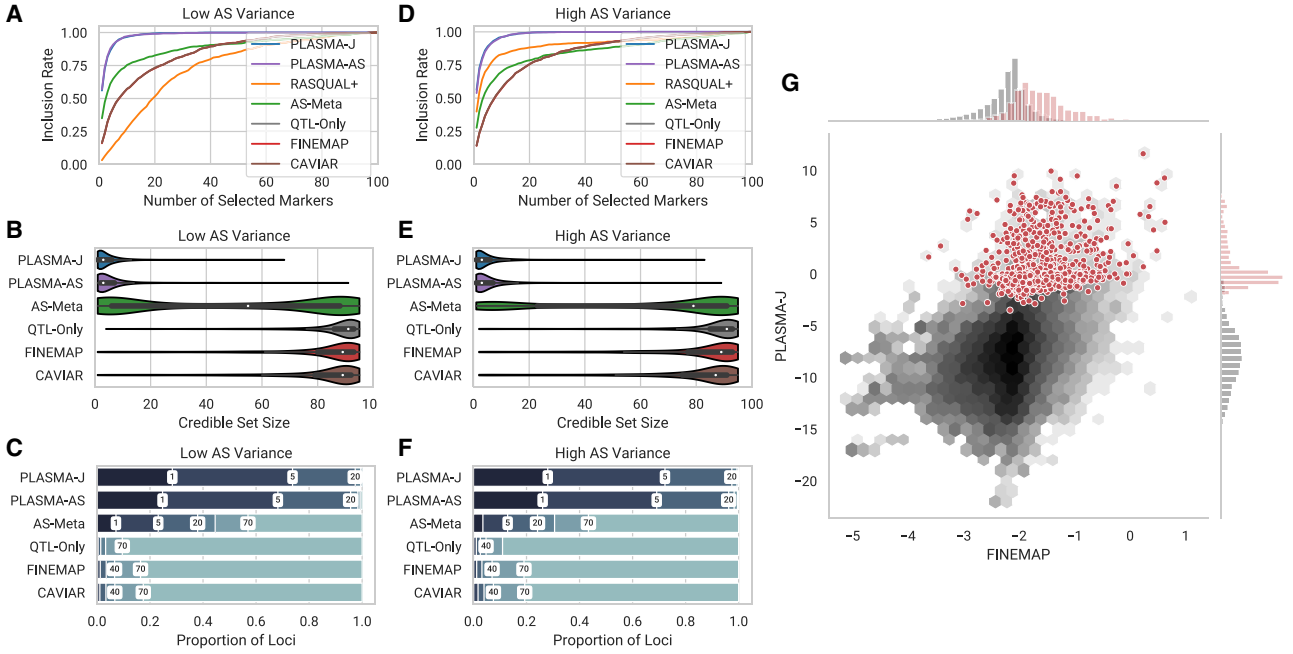

Figure 2. Comparison of Fine-Mapping Methods on Two Sets of Simulated Loci of 100 Markers and 100 Samples Each
These two sets differ in the variance of the AS phenotype, but with a fixed mean coverage and AS heritability.
(A and D) Inclusion curve at low and high AS variance, respectively (0.6, 0.8 standard allelic deviation). The y axis shows the expected proportion of markers included in the credible set, and the x axis shows the number of selected markers by posterior probability.
(B and E) Distribution of 95% confidence credible set sizes at low and high AS variance, respectively.
(C and F) Proportions of loci with 95% credible set sizes within given thresholds at low and high AS variance, respectively. Thresholds used are 1, 5, 20, 40, 70, and 100 markers.
(G) A per-snp fine-mapping comparison between PLASMA-J and FINEMAP across 500 simulated loci with one causal variant each. The axes denote the posterior log-odds of causality for FINEMAP and PLASMA-J, respectively. The black hexagons represent the joint distribution of all markers, while the red dots represent specifically the causal markers. Univariate histograms for PLASMA-J and FINEMAP are plotted along the margins.

There is some variation due to differences in calibration among the methods, but all QTL-based methods have recall above 0.95. PLASMA appears robust to changes in AS variance; at high AS variance, medians are 3.0 for PLASMA-J and 3.0 for PLASMA-AS. In contrast, the performance of AS-Meta varies significantly with the degree of AS variance, even when the underlying signal (coverage and heritability) is constant, with a median set size of 79.0 at high variance. This sensitivity may be due to the fact that AS-Meta does not incorporate marker phasing and thus must rely solely on the intensity rather than the direction of imbalance. Here, RASQUAL+ does not generate meaningful credible sets, with 95% credible set recall being 0.06 and 0.58 for low and high AS variance, respectively. RASQUAL+ is therefore not included in further fine-mapping analyses, though we underscore that RASQUAL remains an effective tool for QTL discovery.

Third, we directly compare how PLASMA-J and FINEMAP prioritize a common set of variants pooled from 500 loci, each with 100 total markers and one causal marker. Figure 2C shows a joint histogram of log posterior marginal odds of these 50,000 variants, with causal variants highlighted in red. Distributions of posterior log-odds for each method are shown as univariate histograms along each axis. As expected, PLASMA and FINEMAP posterior log-odds are positively correlated. Comparing the distribution of the odds of causal variants to those of the rest, it is furthermore evident that PLASMA more confidently assigns probabilities of causality and can much more cleanly segregate causal from non-causal variants.

Lastly, we run the AS-based methods across a wide range of coverage and heritability conditions, recording the mean 95% confidence credible sets, shown in Figure 3. Figures 3A–3C show mean credible set sizes as a function of AS variance and coverage, and Figures 3D–3F show mean credible set sizes as a function of AS variance and AS heritability. In terms of the range of set sizes, PLASMA-J performs the best (3.2 markers on average at best conditions), followed by the PLASMA-AS (3.4 at best conditions), and lastly the AS-Meta method (31 at best conditions). Generally speaking, all methods show results consistent with the behavior of z_ϕ in Figure S3. Although increasing either coverage or heritability results in smaller set sizes, increasing coverage beyond 100 gives diminishing returns as the observed expression levels approach the true expression levels. As expected, AS-Meta tends to struggle at low AS variance, especially apparent at a standard allelic deviation of 0.55, with a mean set size of 78 at best. This may be due to the large majority of samples falling under the threshold for allelic imbalance at 0.65. To verify that

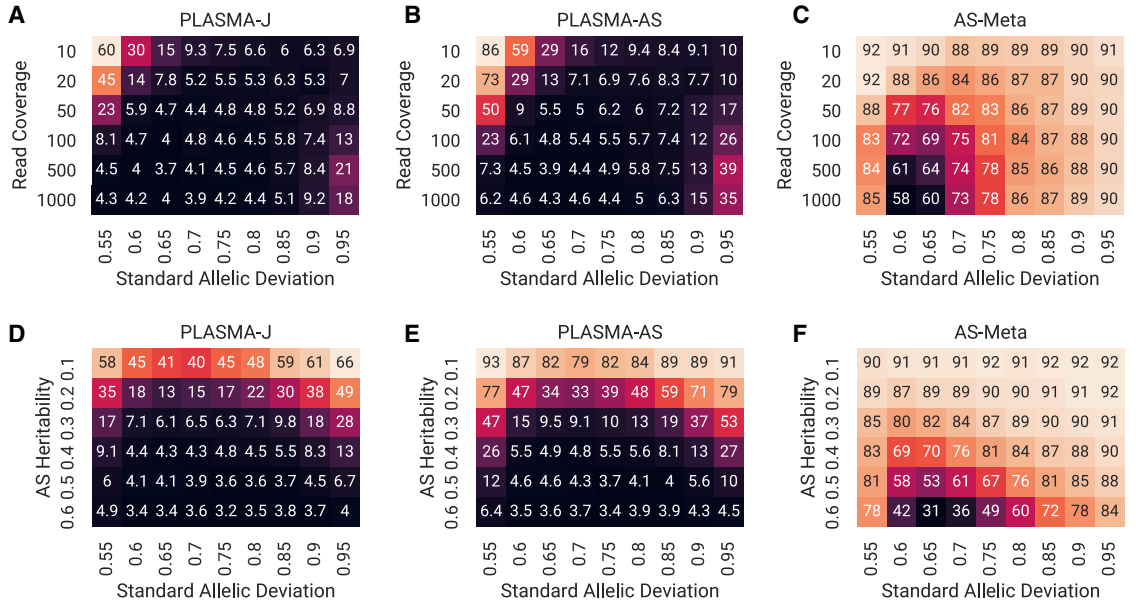

Figure 3. Comparison of the Mean 95% Confidence Credible Set Sizes across Different AS Fine-Mapping Methods
Each square is the mean set size calculated over 500 simulated loci of 100 markers and 100 samples each, with one causal variant. (A–C) Mean credible set sizes as a function of standard allelic deviation and mean read coverage, with AS heritability set to 0.4. (D–F) Mean credible set sizes as a function of standard allelic deviation and AS heritability, with mean read coverage set to 100 reads.

PLASMA is calibrated across the full range of conditions, Figure S4 shows that the 95% credible set sizes have at least a 95% chance of including the causal variant.

Inference of Multiple Causal Variants

To demonstrate PLASMA beyond a one-causal-variant assumption, we fine mapped sets of simulated loci with 2 causal variants with each version of PLASMA. Figure 4A shows the inclusion curves for each version of PLASMA along with FINEMAP. For these curves, inclusion is defined as the expected proportion of causal variants selected, where an inclusion of 1.0 indicates the identification of both causal variants. Here, PLASMA-J and PLASMA-AS deliver an improvement over conventional QTL fine mapping. Compared to single causal variant fine mapping, all methods display a decrease in power, which is consistent with results in earlier QTL fine-mapping analysis,[13,14] where capturing all causal variants becomes increasingly difficult as the number of causal variants increase. The lower power for fine mapping multiple causal variants may be due to the stringent requirement that a model must identify all causal variants in a locus for an inclusion of 1.0. To evaluate the ability of the models to detect the top causal variant, we relax this requirement from identifying all causal variants per locus to at least one causal variant per locus. Inclusion plots for this scenario are shown in Figure 4B, with PLASMA greatly improving the prioritization of the lead causal variant over existing methods.

We next considered credible set sizes which, unlike the inclusion curves, require accurate calibration to be comparable. Previous analyses have shown that proper calibration of fine-mapping methods is more challenging in the presence of multiple causal variants.[30] Unlike the single causal variant case, where all PLASMA model hyperparameters were inferred from simulation parameters, the causal variance hyperparameters in this case were manually calibrated. This need for calibration may be due to linkage disequilibrium obfuscating the relationship between causal effect sizes and total heritability at a locus, and further complicated by the imperfect estimation of linkage disequilibrium at 100 samples.[31] (See Supplemental Material and Methods for information about hyperparameter estimation.) The PLASMA results shown in this section are calibrated such that the recall rates for 95% confidence credible sets are 0.95, 0.96 for PLASMA-J and PLASMA-AS, respectively. This calibration yields median credible set sizes of 86.0 and 90.0 for PLASMA-J and PLASMA-AS, respectively. Like PLASMA, FINEMAP requires user-defined hyperparameters on the prior number of causal variants and on the causal effect sizes. These FINEMAP parameters were set to be equivalent to corresponding calibrated PLASMA parameters. Despite this conservative parameter setting, FINEMAP is overconfident on this dataset with a recall rate of 0.86, so the generated credible sets for FINEMAP are not directly comparable to those of PLASMA.

Fine Mapping of TCGA Kidney RNA-Seq Data

To evaluate our method on real data, we fine mapped gene expression data from 524 human kidney tumor samples and 70 matched normal samples collected by TCGA.[32] The data were processed through a rigorous QC pipeline to account for mapping biases based on established best practices.[19,22] Figures 5A and 5C show credible set size

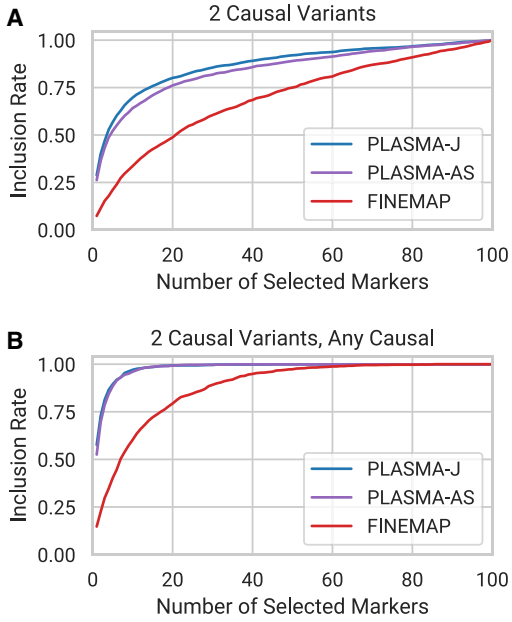

Figure 4. Comparison of Fine-Mapping Methods on Sets of Simulated Loci with Two Causal Variants

Each locus is of 100 markers and 100 samples each, with AS heritability set to 0.4, QTL heritability set to 0.05, and mean read coverage set to 100 reads.

(A) Inclusion curve for including both causal variants.

(B) Inclusion curve for including one or both causal variants.

distribution plots for tumor and normal data, respectively, under a 1 causal variant assumption. We furthermore analyze how well often each method is able to narrow down credible sets under a certain size in simulated loci with one causal variant, shown in Figures 5B and 5D. Among the tumor samples (N = 524; 5,652 loci), 27.9% of loci are fine mapped within 10 variants with PLASMA-J, while 3.4% of loci are fine mapped within 10 variants with FINEMAP. Furthermore, 263 of these loci can be fine mapped down to a single causal variant by PLASMA-J. PLASMA-J, moreover, achieves a median credible set size for 32 variants, whereas FINEMAP achieves a median credible set size of 167 variants. FINEMAP also significantly improves over AS-Meta, which has 6.6% of loci fine mapped within 10 causal variants, and a median credible set size of 166. Results for normal samples (n = 70; 2,034 loci) have a similar trend, with 23.2%, 2.5%, and 1.3% of loci fine mapped within 10 causal variants, for PLASMA-J, AS-Meta, and FINEMAP, respectively. Corresponding median credible set sizes are 32, 248, and 252 variants, for PLASMA-J, AS-Meta, and FINEMAP, respectively. The somewhat lower power for all models is due to having fewer normal samples than tumor samples. To show that these credible set sizes are robust, our choice of heritability hyperparameters, fine-mapping analyses were repeated on the full set of tumor genes with the AS heritability hyperparameter set to 0.05 instead of 0.5. A comparison of the credible set sizes with those from the original parameters are shown in Figure S6.

To investigate how the methods perform at lower sample sizes, we randomly subsample individuals prior to fine mapping. Figure 5E plots the credible set size distributions for PLASMA-J, AS-Meta, and FINEMAP at various sample sizes of kidney tumor data. In terms of loci fine mapped to credible set sizes within 10 variants, PLASMA with 50 samples (484 loci within 10 causal variants) has significantly greater power than FINEMAP with 500 samples (193 loci within 10 causal variants) or AS-Meta with 500 samples (371 loci within 10 causal variants). Additionally, in terms of median credible set size, PLASMA with 10 samples (170 median) has about the same power as FINEMAP with 500 samples (167 median). At a given sample size, PLASMA is thus better able to prioritize variants that will be ranked highly in larger studies. Furthermore, as sample size increases, PLASMA increases in power relative to other methods. In tumor samples PLASMA yields a 1.3-fold decrease in median credible set size over FINEMAP at 10 samples, but a 6.9-fold decrease at 500 samples, indicating that PLASMA scales more effectively with sample size than conventional QTL fine mapping. Nevertheless, PLASMA yields a substantial reduction of credible set sizes even with sample sizes as low as 10, with a median credible set size of 170, compared to a median set size of 219 with FINEMAP. An analogous down-sampling analysis on the normal data is shown in Figure S7. There, PLASMA has higher power for normal samples than for tumor samples, which may be due to the lower variance in the normal data.

Next, we look at how causal variant prioritization is impacted by sample size in the down-sampled analysis. Because the true causal variants in each locus is not known, we use a proxy of markers with a posterior probability of at least 0.1 when fine mapped with FINEMAP on all samples. Note that this will strongly bias the credible set in favor of FINEMAP and thus do not compare this proxy to FINEMAP credible sets. In Figure S8, PLASMA is again more effective than AS-Meta at each sample size at prioritizing causal variants.

To explore multiple causal variant fine mapping on real data, we run PLASMA and FINEMAP assuming up to three causal variants on the full tumor and normal kidney RNA-seq dataset. Figure S9 shows multiple causal variant fine-mapping results for kidney tumor and normal RNA-seq data. As with the simulations, all methods increase in credible set sizes relative to single-causal-variant fine mapping. On tumor data, PLASMA-J, PLASMA-AS, and FINEMAP report a median credible set size of 93, 172, and 150, respectively, with the caveat of possibly unstable calibration for multiple causal variants (as seen in simulations). Interestingly, PLASMA-AS displays a larger power drop than FINEMAP does. This difference suggests that allelic imbalance may be less informative when fine mapping with multiple causal variants. Nevertheless, PLASMA-J performs substantially better than either, suggesting that the joint model is able to combine power from both QTL and AS signals. Regardless, it appears that the majority of

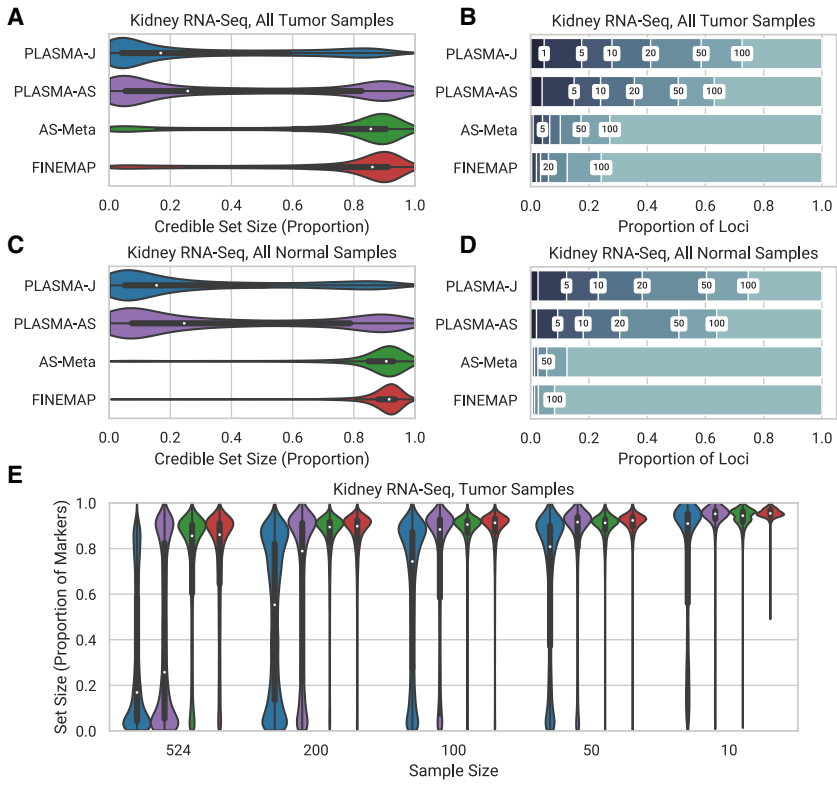

Figure 5. Comparison of the Distribution of 95% Confidence Credible Set Sizes across Loci in Kidney Tumor and Normal Samples, with an Allelic Imbalance False Discovery Rate of 0.05

Expression was measured as RNA-seq read counts mapped to phased genotypes. Fine mapping was conducted with 500 tumor samples and 70 normal samples. Every SNP within 100 kb of each locus was included in the fine-mapping input.

(A and C) Distribution of credible set sizes for tumor and normal samples, respectively, under a 1-causal-variant assumption.

(B and D) Proportion of loci fine mapped to 95% credible sets within given thresholds. Thresholds used are 1, 5, 10, 20, 50, and 100.

(E) 95% credible set size distributions for randomly down-sampled kidney tumor datasets with decreasing sample sizes.

loci contain a single causal variant, with FINEMAP estimating this fraction at 68.8%.

Lastly, we look at how PLASMA prioritizes experimentally verified causal variants at GWAS risk loci. Figure 6 shows the strength AS and QTL associations for DPF3 and SCARB1, genes in two kidney GWAS loci that have verified causal variants.[23,33] At each sample size threshold, the AS statistic generally more confidently identifies the true causal variant than the QTL statistic. In the case of DPF3, the AS statistic is able to prioritize the true causal variant at a substantially lower sample size than the QTL statistic. Moreover, the 95% credible sets from the PLASMA-AS model are smaller than those from the QTL-Only model at a given sample size. By producing a more accurate and confident prioritization of causal variants, PLASMA can substantially reduce the difficulty of experimentally validating causal variants.

Fine Mapping of Prostate H3k27ac ChIP-Seq Data

To evaluate PLASMA with a different molecular phenotype, we fine mapped H3k27ac activity measured by ChIP-seq from 24 human prostate tumor samples and 24 matched normal subjects. Although this study measures chromatin activity rather than expression, the nature of the data is nearly identical to that of RNA-seq and is processed analogously by our QC pipeline and by PLASMA. Instead of fine mapping eQTLs around gene loci, we fine mapped chromatin QTLs (cQTLs) around chromatin peaks. Figure 7 shows distribution plots for tumor data (1,375 peaks) and normal data (908 peaks) under a 1 causal variant assumption. Among the tumor data, 14.5% of peaks are fine mapped within 50 variants with PLASMA-J, while 1.9% of loci are fine mapped within 50 variants with FINEMAP. Furthermore, PLASMA achieves a median credible set size of 236, compared to QTL-Only fine mapping achieving a size of 318. PLASMA also outperforms AS-Meta, with 1.9% of loci fine mapped within 50 causal variants (no gain over FINEMAP) and a median credible set size of 310. Results from normal samples are comparable, with 5.2%, 2.5%, and 2.3% of loci fine mapped within 50 causal variants for PLASMA-J, AS-Meta, and FINEMAP, respectively. These methods achieve a median credible size of 296, 313, and 319 variants, respectively. Overall, these ChIP fine-mapping results are roughly in line with those from RNA-seq fine mapping.

PLASMA Increases Functional Enrichment of Credible Set Markers

To evaluate PLASMA's ability to select markers in functional regions using kidney RNA-seq data, we look for enrichment of prioritized variants at open chromatin regions measured with DNase-seq in a kidney cell line.[34] Since chromatin accessibility is an indicator of transcription factor binding and regulation,[35] an enrichment of credible set markers for open chromatin would indicate that the fine-mapping procedure is prioritizing markers in functionally relevant regions. For instance, the causal variant in the DPF3 locus lies within a DNase-seq peak (Figure 6A). Note that quantifying overlapping with an independent functional feature such as open chromatin imposes no assumptions on the ground truth, in contrast to comparing to external QTL/GWAS data which may be biased toward conventional QTL analysis. The null distribution is defined as the credible set markers being located

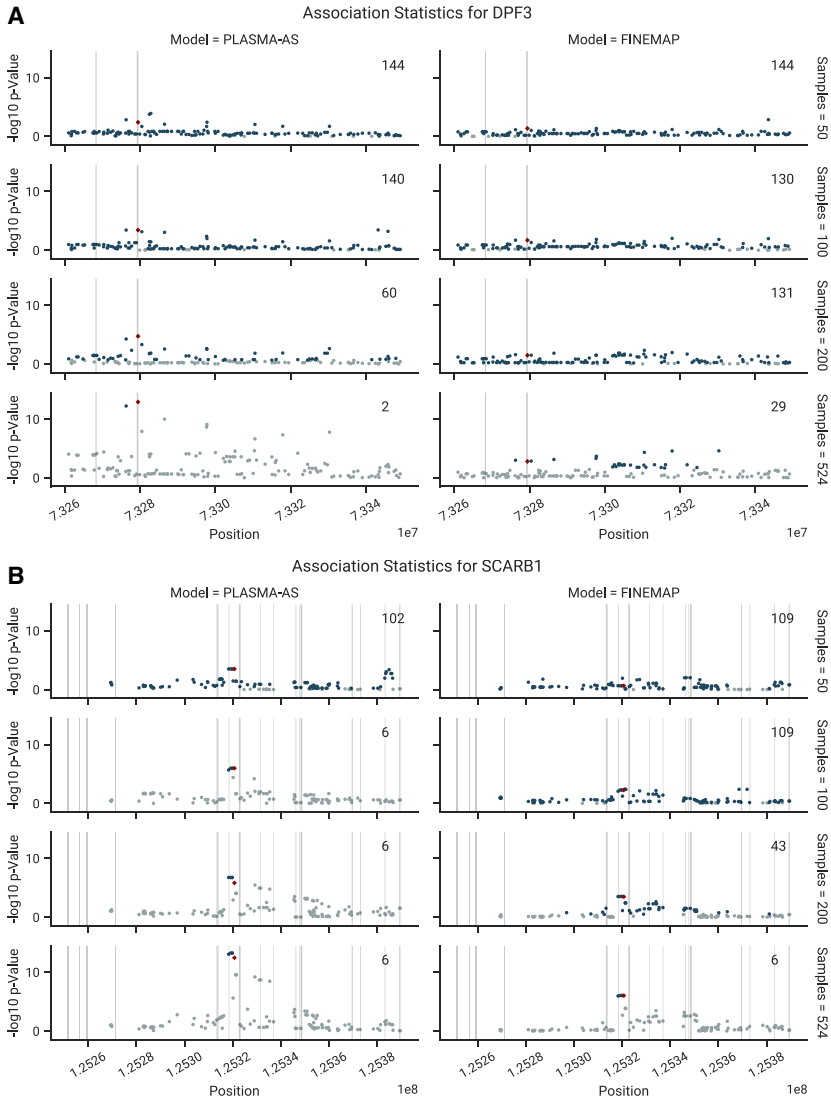

Figure 6. Comparison of AS and QTL Associations in the Experimentally Verified Loci, as a Function of Sample Size

Regions of 70 kb are shown around the causal marker. AS and QTL p values were calculated using from z_ϕ and z_β, respectively. Fine mapping was conducted with the PLASMA-AS and QTL-Only models, respectively. 95% credible set sizes for the whole locus are displayed in the top right of each subplot. Markers in the 95% credible set are shown in dark blue, while markers not in the sets are shown in light blue. The experimentally verified causal marker is shown in red. Regions of open chromatin (DNase-seq peaks) are shaded in gray.
(A) Associations in the DPF3 locus.
(B) Associations in the SCARB1 locus.

independently of open chromatin and use Fisher's exact test to calculate enrichment as a function of minimum causal variant probability. Figures 8, S10A, S10C, and S10D show the odds ratios and p values (computed by Fisher's exact test), respectively, as a function of posterior probability threshold from each fine-mapping method. The credible set markers produced by PLASMA, for the most part, display a significantly stronger enrichment with open chromatin compared to existing methods. For instance, at the p = 0.1 threshold for tumor samples, PLASMA's credible set markers achieve a p value of 9.26×10^{-52} and an odds ratio of 2.16. In comparison, credible sets from QTL-Only fine mapping at that threshold achieves a p value of 2.02×10^{-7} and an odds of 1.62. This enrichment shows that even with far smaller credible sets, PLASMA is able prioritize markers that fall in regions of likely functional significance. The difference between PLASMA and existing methods is greatest at higher posterior probability thresholds. PLASMA may be assigning a more meaningful number of markers with such high posterior probabilities, compared to existing methods that are rarely so confident about a marker's causal status.

Similarly, to validate the credible sets computed from prostate ChIP-seq data, we look for enrichment of credible set markers at chromatin looping anchors measured by Hi-ChIP in a prostate cell line. Regulatory elements overlapping loops are more likely to be involved in *cis*-regulation and we reasoned that they should therefore be enriched for true causal cQTLs.[36,37] Again, we note that this functional feature is independent of the QTL signal or locus LD and is not biased toward a QTL or AS model. Figures S10B and S10E show the odds ratios and p values, respectively, across models as a function of posterior probability threshold (computed by Fisher's exact test). The credible set markers produced by PLASMA display a significantly stronger enrichment with looping anchors compared to the other methods. For instance, at the p = 0.1 threshold, PLASMA's credible sets achieve a p value of 1.05×10^{-6} and an odds of 1.77. In contrast, credible set markers from FINEMAP at that threshold achieves a non-significant p value of 0.80 and an odds of 0.72.

Discussion

We present PLASMA, a statistical fine-mapping method that utilizes allele-specific expression and phased genotypes to select candidate causal variants. By modeling gene expression at a locus in an allele-specific manner, PLASMA scales in power both across individuals and across read counts. Through read-count-level simulations of loci, we show that PLASMA performs robustly across a wide range of realistic conditions and consistently outperforms

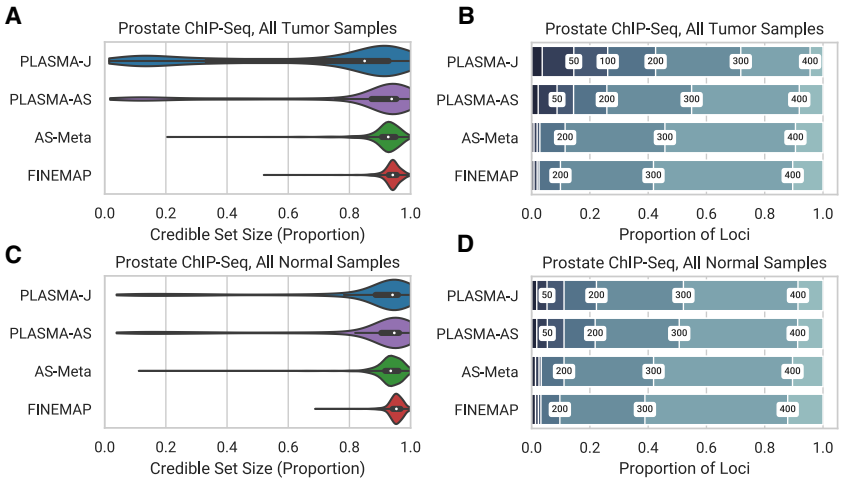

Figure 7. Comparison of 95% Confidence Credible Set Sizes for Peaks in Prostate Tumor and Normal Cells, with an Allelic Imbalance False Discovery Rate of 0.05
Presence of H3k27ac histone marks was quantified as ChIP-seq read counts mapped to phased genotypes.
(A and C) Distribution of credible set sizes for tumor and normal samples, respectively, under a 1-causal-variant assumption.
(B and D) Proportion of peaks fine mapped to 95% credible sets within given thresholds. Thresholds used were 20, 50, 100, 200, 300, and 400.

existing statistical fine-mapping methods, including cases where a significant amount of observed imbalance is caused by non-genetic factors. We further demonstrate this increased power on experimental data by applying PLASMA to a large RNA-seq study, as well as a smaller ChIP-seq study. In both cases, PLASMA achieves substantially smaller credible set sizes compared to existing fine-mapping methods, greatly increasing the number of loci amenable to experimental causal variant validation. Lastly, we show that even with these greatly reduced (more specific) credible set sizes, PLASMA achieves an equivalent or superior degree functional enrichment as existing methods. These results not only present PLASMA as a powerful tool for prioritizing causal variants, but also demonstrate how AS analysis can be directly integrated into statistical fine mapping. A key benefit of PLASMA is its ability to utilize existing, conventional sequencing-based QTL data, such as RNA-seq, CHiP-seq, and ATAC-seq at low sample size. This allows researchers to gain significant insight simply by revisiting past QTL studies, especially those with sample sizes too low for conventional QTL fine mapping.

Although it is evident that an AS analysis with PLASMA confers more signal than an equivalently sized QTL analysis, AS analysis presents additional obstacles and potential confounders. First, unlike conventional QTL fine-mapping methods that rely only on allelic dosage, PLASMA additionally utilizes genotype phasing, making phasing accuracy a potential concern. However, since PLASMA focuses on *cis*-regulation, the genotypes observed span no more than several hundred kilobases per locus, well within the high accuracy range of modern phasing algorithms.[38] Second, PLASMA depends on having heterozygous individuals in the tested feature and SNP in order to leverage AS signal. In our analyses we focused on features that were testable by AS (10,946 of 19,645 total genes, 113,459 of 525,629 total peaks). However, even in the complete absence of heterozygotes, PLASMA can still conduct conventional fine mapping based on dosage and total expression. Recent technologies that could potentially offer greater signal include RNA-seq with unspliced transcripts[39] and direct allele-specific measurement of expression using single-cell RNA-seq.[40] Third, PLASMA assumes the same causal configuration underlying both the AS and QTL effects (and is thus able to combine the signals) but the causal effects may differ due to real biological confounding. For example, *cis* effects on gene A followed by (local) *trans* effects of gene A on gene B would be identified as a QTL association, but would not exhibit AS association. This would be a model violation for PLASMA and produce larger credible set sizes. Although PLASMA can consider correlations between causal AS and QTL affect sizes, this parameter is hard to estimate, and we find in real data that the model with correlation set to zero (PLASMA-J) exhibited greater power than a non-zero constant. Future work is required to fully elucidate the relationship between allele-specific and total effects, which likely differs across genes. Fourth, genomic imprinting (where either the maternal or paternal copy of the gene is silenced) or random monoallelic expression would produce the appearance of allelic imbalance within affected individuals in the absence of true *cis*-regulatory signal.[20] Although PLASMA does not explicitly model such biases, a bias that is independent of genotype will only cause a reduction in power and not produce false positives. A potential extension would be to model such violations or discrepancies between the QTL and AS models directly, following the lines of methods such as RASQUAL.[20] Fifth, PLASMA currently does not incorporate covariate analysis in the allele-specific model (though the intra-individual nature of the test controls for false positives), which could additionally be used to model environmental confounders and increase power.[41] AS covariate analysis could potentially be achieved through a multivariate likelihood ratio test as in WASP.[19]

PLASMA's approach in combining QTL and AS signals opens up possible future work in two distinct directions. The first direction would be to build upon the generative fine-mapping model to incorporate additional sources of signal. For example, one can incorporate epigenomic annotation data by setting the priors for causality for each

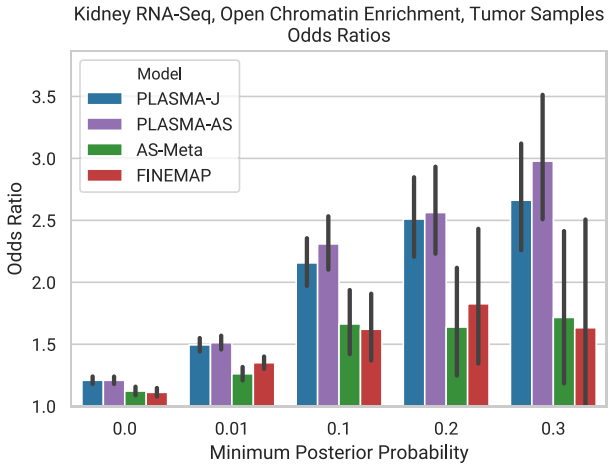

Figure 8. Enrichment of PLASMA Kidney RNA-Seq Credible Sets on Open Chromatin
The x axis represents a minimum posterior probability threshold on the markers within credible sets, with a higher threshold indicating more confident but fewer markers. The markers in these thresholded credible sets were intersected with the functional feature in question. 95% confidence intervals were calculated with Fisher's exact test, with a control of all markers in a locus.

marker. Approaches used in existing QTL-based methods such as PAINTOR and RiVIERA-MT[16,42] could be transferred to PLASMA with relatively little difficulty. Another possibility would be to conduct N-phenotype colocalization by utilizing additional phenotypes in addition to the AS and QTL phenotypes. Generalizing from two to multiple phenotypes would be straightforward and could utilize the colocalization algorithm first introduced in eCAVIAR.[2] A second, more general direction would be to adapt QTL-based population genetics methods to utilize AS summary statistics. Since both QTL and AS statistics can be characterized as linear combinations of haplotype-level genotypes, they share many distributional properties, including LD, allowing them to be easily interchangeable in many circumstances. One such application would be gene expression prediction for transcriptome-wide association studies (TWASs),[43] where the increased signal of AS statistics could increase power to identify gene-phenotype relationships. Broadly speaking, the allele-specific model and association statistics that PLASMA introduces will be relevant to any analysis of small sample size or limited tissue.

Appendix A

Modeling Genetic Effects on Total Expression

We calculate marginal effect sizes for a given locus under the conventional linear model of total gene expression. Let us consider a QTL study of a given locus with n individuals and m markers. Let $\mathbf{y}$ be an $(n \times 1)$ vector of total expression across the individuals, recentered at zero. Given a marker i, let $\mathbf{x}_i$ be an $(n \times 1)$ zero-recentered vector of genotypes. We define β_i, the genetic effect of marker i on total gene expression as follows:

$$\mathbf{y} = \mathbf{x}_i\beta_i + \boldsymbol{\epsilon}_i \qquad \text{(Equation A1)}$$

We model the residuals $\boldsymbol{\epsilon}_i$ as normally distributed with variance $\sigma^2_{y,i}$.

Calculation of QTL Summary Statistics

We use the maximum likelihood estimator of β_i, equivalent to the ordinary-least-squares linear regression estimator:

$$\widehat{\beta}_i = \left(\mathbf{x}_i^\top \mathbf{x}_i\right)^{-1}\mathbf{x}_i^\top \mathbf{y} \qquad \text{(Equation A2)}$$

Under the null model where i is not causal, i does not explain any amount of variation of the phenotype, and the variance of $\mathbf{y}$ is simply $\sigma^2_{y,1}$. Thus, under the null:

$$\begin{aligned}\mathrm{Var}\left(\widehat{\beta}_i\right) &= \left(\mathbf{x}_i^\top \mathbf{x}_i\right)^{-2}\mathrm{Var}\left(\mathbf{x}_i^\top \mathbf{y}\right)\\ &= \left(\mathbf{x}_i^\top \mathbf{x}_i\right)^{-2}\left(\mathbf{x}_i^\top \mathbf{x}_i\right)\mathrm{Var}(y)\\ &= \left(\mathbf{x}_i^\top \mathbf{x}_i\right)^{-1}\mathrm{Var}(y)\\ &= \left(\mathbf{x}_i^\top \mathbf{x}_i\right)^{-1}\sigma^2_{y,i}\end{aligned} \qquad \text{(Equation A3)}$$

We estimate $\sigma^2_{\beta,i}$ from the residuals:

$$\sigma^2_{\beta,i} = \frac{\boldsymbol{\epsilon}_i^\top \boldsymbol{\epsilon}_i}{n-1} \qquad \text{(Equation A4)}$$

We thus define our QTL summary statistic (Wald statistic) for marker i as:

$$\widehat{z}_{\beta,i} = \frac{\widehat{\beta}_i}{\sqrt{\left(\mathbf{x}_i^\top \mathbf{x}_i\right)^{-1}\sigma^2_{y,i}}} \qquad \text{(Equation A5)}$$

We assume that the number of individuals is enough such that the observed statistic is normally distributed with unit variance:

$$\widehat{z}_{\beta,i} \sim \mathcal{N}(z_{\beta,i}, 1) \qquad \text{(Equation A6)}$$

In the case where $\mathbf{x}_i$ is of unit variance, the statistic simplifies to:

$$\widehat{z}_{\beta,i} = \frac{\widehat{\beta}_i\sqrt{n}}{\sqrt{\widehat{\sigma}^2_{y,i}}} \qquad \text{(Equation A7)}$$

Modeling Haplotype-Specific Effects on Expression

We model allele-specific expression under the observation that a *cis*-regulatory variant often has a greater influence on the gene allele of the same haplotype. Under this model, an individual who is heterozygous for one or more *cis*-regulatory markers will show an imbalance in expression between the alleles.

From a quantitative perspective, let us consider a single locus in a single individual who is heterozygous for marker i. Let 0 and 1 represent the wild-type and alternative marker alleles, respectively. We define e_0 as the expression of the gene allele on the same phase as marker allele 0 and e_1 as

the expression of the gene allele on the same phase as marker allele 1. Let e_0' and e_1' be baseline expressions without the effect of marker i. We define δ_i as the cis-regulatory strength of marker allele 1 over marker allele 0 such that:

$$\frac{e_1}{e_0} = \delta_i \frac{e_1'}{e_0'} \quad \text{(Equation A8)}$$

If we define i's phase, v_i, we can arbitrarily assign haplotypes A and B. The above equation then becomes:

$$\frac{e_A}{e_B} = (\delta_i)^{v_i} \frac{e_A'}{e_B'} \quad \text{(Equation A9)}$$

The marker's phase is 1 if haplotype A contains the alternative marker allele, -1 if haplotype B contains the alternative marker allele, and 0 if the individual is homozygous for the marker.

We now re-write Equation A9 as a linear model. Let w be the log expression ratio between haplotypes A and B:

$$w = \log\left(\frac{e_A}{e_b}\right) \quad \text{(Equation A10)}$$

Let ϕ_i be the log allelic fold change (logAFC) caused by variant i:

$$\phi = \log(\delta_i) \quad \text{(Equation A11)}$$

Let ζ_i be the log baseline expression ratio between haplotypes A and B:

$$\zeta_i = \log\left(\frac{e_A'}{e_B'}\right) \quad \text{(Equation A12)}$$

With these parameters we rewrite Equation A9 as:

$$w = v_i\phi_i + \zeta_i \quad \text{(Equation A13)}$$

Given n individuals, this expression becomes:

$$\mathbf{w} = \mathbf{v}_i\phi_i + \boldsymbol{\zeta}_i \quad \text{(Equation A14)}$$

We assume that $\boldsymbol{\zeta}_i$ is drawn from a normal distribution with variance $\sigma^2_{w,i}$. Note that under this model, ϕ_i can be interpreted as the effect size of marker i on allelic imbalance, with $\boldsymbol{\zeta}_i$ as the residuals. Furthermore, assuming no haplotype bias, both $\mathbf{w}$ and $\mathbf{v}_i$ are zero-centered in expectation.

Experimentally derived AS data, such as RNA-seq data, yield reads that are mapped to a particular haplotype. Given c_A and c_B, the read counts mapped to haplotypes A and B, respectively, we define our estimator of w as:

$$\widehat{w} = \log\left(\frac{c_A}{c_B}\right) \quad \text{(Equation A15)}$$

For a given individual j, we define $c_{A,j}$ as the allele-specific read count from haplotype A. We model the allele-specific read count as drawn a beta-binomial distribution, given the total mapped read count c_j:

$$c_{A,j} \sim \mathrm{BB}(\alpha_j, \beta_j, c_j) \quad \text{(Equation A16)}$$

We define π_j as the expected proportion of read counts (allelic fraction) from haplotype A:

$$\pi_j = \frac{\mathbb{E}[c_{A,j}]}{c_j} = \frac{\alpha_j}{\alpha_j + \beta_j} \quad \text{(Equation A17)}$$

α_j and β_j can be re-parameterized in terms of π_j and the sampling overdispersion ρ_e

$$\rho_e = \frac{1}{\alpha_j + \beta_j + 1} \quad \text{(Equation A18)}$$

With this re-paramaterization, the mean and variance of $c_{A,i}$ is given as follow:

$$\mathbb{E}[c_{A,j}] = c_j\pi_j \quad \text{(Equation A19)}$$

$$\mathrm{Var}(c_{A,j}) = c_j\pi_j(1-\pi_j)(1+\rho_e(c_j-1)) \quad \text{(Equation A20)}$$

We use this beta binomial model to estimate the variance of $\widehat{w}_i$. We scale the distribution by $(1/c_i)$ to get the mean and variance for the read count proportion:

$$\mathbb{E}\left[\frac{c_{A,j}}{c_j}\right] = \pi_j \quad \text{(Equation A21)}$$

$$\mathrm{Var}\left(\frac{c_{A,j}}{c_j}\right) = \frac{1}{c_j}\pi_j(1-\pi_j)(1+\rho_e(c_j-1)) \quad \text{(Equation A22)}$$

We define w^* as the logit-transformed allelic fraction:

$$w_j^* = \mathrm{logit}(\pi_j) = \log\frac{\pi_j}{1-\pi_j} \quad \text{(Equation A23)}$$

$$\frac{dw_j^*}{d\pi_j} = \frac{1}{\pi_j(1-\pi_j)} \quad \text{(Equation A24)}$$

$$\frac{d^2w_j^*}{d\pi_j^2} = \frac{2\pi_j - 1}{\pi_j^2(1-\pi_j)^2} \quad \text{(Equation A25)}$$

We can thus find the approximate mean and variance of $\widehat{w}_j$ given $\widehat{w}_j^*$ using Taylor expansions:

$$\begin{aligned}
\mathbb{E}[\widehat{w}_j] &= \mathbb{E}\left[\mathrm{logit}\left(\frac{c_{A,j}}{c_j}\right)\right] \\
&\approx \mathrm{logit}\left(\mathbb{E}\left[\frac{c_{A,j}}{c_j}\right]\right) + \frac{1}{2}\mathrm{Var}\left(\frac{c_{A,j}}{c_j}\right)\frac{d^2}{d\pi_j^2}\mathrm{logit}(\pi_j) \\
&\approx \mathrm{logit}(\pi_j) + \frac{1}{2}\left(\frac{1}{c_j}\pi_j(1-\pi_j)(1+\rho_e(c_j-1))\right)\left(\frac{2\pi_j-1}{\pi_j^2(1-\pi_2)^2}\right) \\
&\approx \mathrm{logit}(\pi_j) + \frac{2\pi_j-1}{2c_j\pi_j(1-\pi_j)}(1+\rho_e(c_j-1)) \\
&\approx w_j^* + \frac{1}{c_j}\sinh(w_j^*)(1+\rho_e(c_j-1))
\end{aligned} \quad \text{(Equation A26)}$$

$$\begin{aligned}
\mathrm{Var}(\widehat{w}_j) &= \mathrm{Var}\left(\mathrm{logit}\left(\frac{c_{A,j}}{c_j}\right)\right)\\
&\approx \mathrm{Var}\left(\frac{c_{A,j}}{c_j}\right)\left(\frac{d}{d\pi_j}\mathrm{logit}(\pi_j)\right)^2\\
&\approx \left(\frac{1}{c_j}\pi_j(1-\pi_j)\left(1+\rho_e(c_j-1)\right)\right)\left(\frac{1}{\pi_j(1-\pi_j)}\right)^2\\
&\approx \frac{1+\rho_e(c_j-1)}{c_j\pi_j(1-\pi_j)}\\
&\approx \frac{2}{c_j}\left(1+\cosh\left(w_j^*\right)\right)\left(1+\rho_e(c_j-1)\right)
\end{aligned} \tag{Equation A27}$$

Note that w and w^* are not equivalent because $\mathbb{E}[\mathrm{logit}(c_A/c)] \neq \mathrm{logit}(\mathbb{E}[c_A/c])$. Equation A26 implies that $\widehat{w}$ is a biased estimator of w^*, especially at low read counts and/or high overdispersion. To get an estimator of w^* with reduced bias, we take the approximation that $\sinh(w^*) \approx w^*$ around zero:

$$\widehat{w}_j^* \frac{\widehat{w}_j}{1+\frac{1}{c_j}\left(1+\rho_e(c_j-1)\right)} \tag{Equation A28}$$

We use $\widehat{w}^*$ to find an estimator of $\sigma^2_{c,j}$, the variance of $\widehat{w}$:

$$\sigma^2_{c,j} = \frac{2}{c_j}\left(1+\cosh\left(\widehat{w}_j^*\right)\right)\left(1+\rho_e(c_j-1)\right) \tag{Equation A29}$$

Given our estimator $\widehat{w}_j$, we quantify the sampling error $\tau_j = \widehat{w}_j - w_j$, with $\mathbb{E}[\tau_j] = 0$ and $\mathrm{Var}(\tau_j) = \sigma^2_{c,j}$. Thus, across individuals:

$$\widehat{\mathbf{w}} = \mathbf{v}_i\phi_i + \boldsymbol{\zeta}_i + \boldsymbol{\tau} \tag{Equation A30}$$

Calculation of AS Summary Statistics

Due to heteroscedasticity among individuals, we estimate the AS effect size ϕ_i in a weighted manner, giving larger weights to individuals with lower expected sampling error. Given individual j, we define the weight for j as the inverse of the estimated read count variance:

$$\omega_j = \frac{1}{\widehat{\sigma}^2_{c,j}} \tag{Equation A31}$$

We define our weight matrix $\boldsymbol{\Omega}$ as a diagonal matrix with $\boldsymbol{\Omega}_{j,j} = \omega_j$.

We use the weighted-least-squares estimator for :

$$\widehat{\phi}_i = \left(\mathbf{v}_i^\top \boldsymbol{\Omega}\mathbf{v}_i\right)^{-1}\mathbf{v}_i^\top \boldsymbol{\Omega}\widehat{\mathbf{w}} \tag{Equation A32}$$

Under the null model where i is not causal, the variance of w_j is $\sigma^2_{w,i}$ and the variance of $\widehat{w}_j$ is $\sigma^2_{w,i} + \sigma^2_{c,j}$. We assume that the experimental errors τ and biological residuals ζ_i are uncorrelated. Thus, under the null:

$$\begin{aligned}
\mathrm{Var}(\widehat{\phi}_i) &= \mathbb{E}\left[\left(\widehat{\phi}_i - \phi_i\right)^2\right]\\
&= \left(\mathbf{v}_i^\top \boldsymbol{\Omega}\mathbf{v}_i\right)^{-1}\mathbf{v}_i^\top \boldsymbol{\Omega}(\zeta_i+\boldsymbol{\tau})(\zeta_i+\boldsymbol{\tau})^\top \boldsymbol{\Omega}^\top \mathbf{v}_i\left(\mathbf{v}_i^\top \boldsymbol{\Omega}\mathbf{v}_i\right)^{-1}\\
&= \left(\mathbf{v}_i^\top \boldsymbol{\Omega}\mathbf{v}_i\right)^{-2}\left(\mathbf{v}_i^\top \boldsymbol{\Omega}\boldsymbol{\zeta}^\top\boldsymbol{\zeta}_i\boldsymbol{\Omega}^\top\mathbf{v}_i + \mathbf{v}_i^\top\boldsymbol{\Omega}\boldsymbol{\tau}\boldsymbol{\tau}^\top\boldsymbol{\Omega}^\top\mathbf{v}_i\right)\\
&= \left(\mathbf{v}_i^\top \boldsymbol{\Omega}\mathbf{v}_i\right)^{-2}\left(\mathbf{v}_i^\top \boldsymbol{\Omega}\mathbf{I}\sigma^2_{w,i}\boldsymbol{\Omega}^\top\mathbf{v}_i + \mathbf{v}_i^\top\boldsymbol{\Omega}\mathbf{I}\mathbf{v}_i\right)\\
&= \left(\mathbf{v}_i^\top \boldsymbol{\Omega}\mathbf{v}_i\right)^{-2}\left(\left(\mathbf{v}_i^\top \boldsymbol{\Omega}^2\mathbf{v}_i\right)\sigma^2_{w,i} + \mathbf{v}_i^\top\boldsymbol{\Omega}\mathbf{v}_i\right)
\end{aligned} \tag{Equation A33}$$

We now estimate $\sigma^2_{w,i}$ from the residuals. Note that we are estimating the variance of the biological residuals $\mathrm{Var}(\zeta_i)$, which is distinct from the total residuals are $\boldsymbol{\zeta}_i + \boldsymbol{\tau}$, so we cannot directly use the variance of the total residuals. We instead use the following estimator for $\sigma^2_{w,i}$:

$$\widehat{\sigma}^2_{w,i} = \frac{\sum_{j=1}^n\left(\omega_j\left(\zeta_{i,j}+\tau_j\right)^2 - 1\right)}{\sum_{j=1}^n \omega_j} \tag{Equation A34}$$

We show that this estimator is equal to $\sigma^2_{\phi,i}$ in expectation:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}^i_{w,i}\right] &= \frac{\sum_{j=1}^n\left(\omega_j\mathbb{E}\left[\left(\zeta_{i,j}+\tau_j\right)^2\right] - 1\right)}{\sum_{j=1}^n\omega_j}\\
&= \frac{\sum_{j=1}^n\left(\omega_j\mathrm{Var}\left(\zeta_{i,j}+\tau_j\right) - 1\right)}{\sum_{j=1}^n\omega_j}\\
&= \frac{\sum_{j=1}^n\left(\omega_j\mathrm{Var}\left(\zeta_{i,j}\right) + \omega_j\mathrm{Var}\left(\tau_j\right) - 1\right)}{\sum_{j=1}^n\omega_j}\\
&= \frac{\sum_{j=1}^n\omega_j\mathrm{Var}\left(\zeta_{i,j}\right)}{\sum_{j=1}^n\omega_j}\\
&= \mathrm{Var}(\zeta_i)\\
&= \sigma^2_{w,i}
\end{aligned} \tag{Equation A35}$$

With this estimator, we define the AS association statistic for marker i as follows:

$$\widehat{z}_{\phi,i} = \frac{\widehat{\phi}_i}{\sqrt{\left(\mathbf{v}_i^\top \boldsymbol{\Omega}\mathbf{v}_i\right)^{-2}\left(\left(\mathbf{v}_i^\top \boldsymbol{\Omega}^2\mathbf{v}_i\right)\widehat{\sigma}^2_{w,i} + \mathbf{v}_i^\top\boldsymbol{\Omega}\mathbf{v}_i\right)}} \tag{Equation A36}$$

We assume that the observed statistic is normally distributed with unit variance:

$$\widehat{z}_{\phi,i} \sim \mathcal{L}(z_{\phi,i}, 1) \tag{Equation A37}$$

To gain an intuitive understanding of the association statistic, let us examine it under simplifying conditions. We assume that $\mathbf{v}_i$ is of unit variance, that read count overdispersion is negligible, and that allelic imbalance and read coverage are fixed across individuals. Under these

conditions, let $\mathbf{\Omega} = (c/k)\mathbf{I}$ for coverage c and some constant k. Equation A36 simplifies to:

$$\widehat{z}_{\phi,i} = \frac{\widehat{\phi}_i \sqrt{n}}{\sqrt{\widehat{\sigma}^2_{w,i} + \frac{k}{c}}} \qquad \text{(Equation A38)}$$

We can see that under high experimental noise (k/c), the denominator is dominated by the quality of data (read coverage). In contrast, when experimental noise is low, the denominator is dominated by $\widehat{\sigma}^2_{w,i}$, determined by the inherent heritability of the locus's AS phenotype.

In the case where phasing error is significant, we would expect the estimated AS effects ($\widehat{\phi}$) to have more deviation from the true effects. We derive a correction for the AS z-score, given a per-marker probability of mis-phasing ψ_i. We define $\widehat{\mathbf{v}}_i$ as the imperfect observed phasing for marker i, and we define the phasing error vector $\boldsymbol{\delta}_i$ such that $\boldsymbol{\delta}_i = \mathbf{v}_i - \widehat{\mathbf{v}}_i$. Note that each δ is a ternary $-2/0/2$ indicator, with each δ^2 being a binary 0/4 indicator of a phasing error. We assume that the occurence of a phasing error is independent of which haplotype the alternative allele is one, so that $\mathbb{E}[\boldsymbol{\delta}_i = 0]$. We now derive the variance of ϕ_i under imperfect phasing:

$$\begin{aligned}
\widehat{\phi}_i - \phi_i &= \left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)^{-1} \widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{w}} - \phi_i \\
&= \left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)^{-1} \widehat{\mathbf{v}}_i^\top \mathbf{\Omega} (\mathbf{v}_i \phi_i + \boldsymbol{\zeta}_i + \boldsymbol{\tau}) - \phi_i \\
&= \left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)^{-1} \widehat{\mathbf{v}}_i^\top \mathbf{\Omega} (\widehat{\mathbf{v}}_i \phi_i + \boldsymbol{\delta}_i \phi_i + \boldsymbol{\zeta}_i + \boldsymbol{\tau}) - \phi_i \\
&= \left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)^{-1} \widehat{\mathbf{v}}_i^\top \mathbf{\Omega} (\boldsymbol{\delta}_i \phi_i + \boldsymbol{\zeta}_i + \boldsymbol{\tau})
\end{aligned} \qquad \text{(Equation A39)}$$

$$\begin{aligned}
&\mathrm{Var}(\widehat{\phi}_i) = \mathbb{E}\left[\left(\widehat{\phi}_i - \phi_1\right)^2\right] \\
&= \left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)^{-1} \widehat{\mathbf{v}}_i^\top \mathbf{\Omega} (\boldsymbol{\delta}_i \phi_i + \boldsymbol{\zeta}_i + \boldsymbol{\tau})(\boldsymbol{\delta}_i \phi_i + \boldsymbol{\zeta}_i + \boldsymbol{\tau})^\top \mathbf{\Omega}^\top \widehat{\mathbf{v}}_i \\
&\quad \left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)^{-1} \\
&= \left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)^{-2} \left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \boldsymbol{\delta}_i^\top \boldsymbol{\delta}_i \phi_i^2 \mathbf{\Omega}^\top \widehat{\mathbf{v}}_i + \widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \boldsymbol{\zeta}_i^\top \boldsymbol{\zeta}_i \mathbf{\Omega}^\top \widehat{\mathbf{v}}_i \right. \\
&\quad \left. + \widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \boldsymbol{\tau} \boldsymbol{\tau}^\top \mathbf{\Omega}^\top \widehat{\mathbf{v}}_i\right) \\
&= \left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)^2 \left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \left(\mathbf{I} 4 \psi_i \phi_i^2 + \mathbf{I} \sigma^2_{w,i}\right) \mathbf{\Omega}^\top \widehat{\mathbf{v}}_i + \widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \mathbf{I} \widehat{\mathbf{v}}_i\right) \\
&= \left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)^{-2} \left(\left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega}^2 \widehat{\mathbf{v}}_i\right)\left(\sigma^2_{w,i} + 4\psi_i \phi_i^2\right) + \widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)
\end{aligned} \qquad \text{(Equation A40)}$$

When calculating $\mathrm{Var}(\widehat{\phi}_i)$, we approximate the ϕ_i^2 term with the observed ϕ_i^2. We thus define the corrected z-score:

$$\widehat{z}_{\phi,i} = \frac{\widehat{\phi}_i}{\sqrt{\left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)^{-2} \left(\left(\widehat{\mathbf{v}}_i^\top \mathbf{\Omega}^2 \widehat{\mathbf{v}}_i\right)\left(\widehat{\sigma}^2_{w,i} + 4\psi_i \widehat{\phi}_i^2\right) + \widehat{\mathbf{v}}_i^\top \mathbf{\Omega} \widehat{\mathbf{v}}_i\right)}} \qquad \text{(Equation A41)}$$

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2019.12.011.

Acknowledgments

We thank F. Hormozdiari for guidance on statistical fine mapping and C. Kalita for guidance on model validation. We also thank B. Pasanuic, C. Giambartolomei, and M. Kellis for helpful feedback.

A.T.W. and A.G. were supported by the Claudia Adams Barr Award, R01 CA227237, and R21 HG010748. M.M.P. was supported by the Rebecca and Nathan Milikowsky Family Foundation. M.L.F. was supported by R01CA193910, R01CA204954, R01GM107427, the Prostate Cancer Foundation Challenge Award, and the H.L. Snyder Medical Research Foundation.

Declaration of Interests

The authors declare no competing interests.

Received: June 14, 2019
Accepted: December 29, 2019
Published: January 30, 2020

References

1. Moyerbrailean, G.A., Kalita, C.A., Harvey, C.T., Wen, X., Luca, F., and Pique-Regi, R. (2016). Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? PLoS Genet. *12*, e1005875.
2. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., and Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*, 481–487.
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. Ongen, H., Brown, A.A., Delaneau, O., Panousis, N.I., Nica, A.C., Dermitzakis, E.T.; and GTEx Consortium (2017). Estimating the causal tissues for complex traits and diseases. Nat. Genet. *49*, 1676–1683.
5. Lappalainen, T. (2015). Functional genomics bridges the gap between quantitative genetics and molecular biology. Genome Res. *25*, 1427–1431.
6. Battle, A., Brown, C.D., Engelhardt, B.E., Montgomery, S.B.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; and eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.
7. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and

Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.
8. Veyrieras, J.B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. *4*, e1000214.
9. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
10. Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet. *19*, 491–504.
11. Hormozdiari, F., Zhu, A., Kichaev, G., Ju, C.J., Segrè, A.V., Joo, J.W.J., Won, H., Sankararaman, S., Pasaniuc, B., Shifman, S., and Eskin, E. (2017). Widespread Allelic Heterogeneity in Complex Traits. Am. J. Hum. Genet. *100*, 789–802.
12. Wheeler, H.E., Shah, K.P., Brenner, J., Garcia, T., Aquino-Michaels, K., Cox, N.J., Nicolae, D.L., Im, H.K.; and GTEx Consortium (2016). Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. PLoS Genet. *12*, e1006423.
13. Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M., Auton, A., Myers, S., Morris, A., et al.; Wellcome Trust Case Control Consortium (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat. Genet. *44*, 1294–1301.
14. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508.
15. Chen, W., Larrabee, B.R., Ovsyannikova, I.G., Kennedy, R.B., Haralambieva, I.H., Poland, G.A., and Schaid, D.J. (2015). Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. Genetics *200*, 719–736.
16. Benner, C., Spencer, C.C., Havulinna, A.S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics *32*, 1493–1501.
17. Kichaev, G., Yang, W.Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.
18. Gupta, R.M., Hadaya, J., Trehan, A., Zekavat, S.M., Roselli, C., Klarin, D., Emdin, C.A., Hilvering, C.R.E., Bianchi, V., Mueller, C., et al. (2017). A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. Cell *170*, 522–533.e15.
19. Knowles, D.A., Davis, J.R., Edgington, H., Raj, A., Favé, M.J., Zhu, X., Potash, J.B., Weissman, M.M., Shi, J., Levinson, D.F., et al. (2017). Allele-specific expression reveals interactions between genetic variation and environment. Nat. Methods *14*, 699–702.
20. van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods *12*, 1061–1063.
21. Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat. Genet. *48*, 206–213.
22. Moyerbrailean, G.A., Richards, A.L., Kurtz, D., Kalita, C.A., Davis, G.O., Harvey, C.T., Alazizi, A., Watza, D., Sorokin, Y., Hauff, N., et al. (2016). High-throughput allele-specific expression across 250 environmental conditions. Genome Res. *26*, 1627–1638.
23. 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.
24. Gusev, A., Spisak, S., Fay, A.P., Carol, H., et al. (2019). Allelic imbalance reveals widespread germline-somatic regulatory differences and prioritizes risk loci in Renal Cell Carcinoma. bioRxiv. https://doi.org/10.1101/631150.
25. Zou, J., Hormozdiari, F., Jew, B., Ernst, J., et al. (2018). Leveraging allele-specific expression to refine fine-mapping for eQTL studies. bioRxiv. https://doi.org/10.1101/257279.
26. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.
27. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.
28. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
29. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.
30. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2019). A simple new approach to variable selection in regression, with application to genetic fine-mapping. bioRxiv. https://doi.org/10.1101/501114.
31. Benner, C., Havulinna, A.S., Järvelin, M.R., Salomaa, V., Ripatti, S., and Pirinen, M. (2017). Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. Am. J. Hum. Genet. *101*, 539–551.
32. Creighton, C.J., Morgan, M., Gunaratne, P.H., Wheeler, D.A., et al.; Cancer Genome Atlas Research Network (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature *499*, 43–49.
33. Colli, L.M., Jessop, L., Myers, T.A., Machiela, M.J., Choi, J., Purdue, M., Brown, K., and Chanock, S.J. (2018). Functional characterization of the 14q24 renal cancer susceptibility locus implicates SWI/SNF complex member DPF3 via inhibition of apoptosis. Cancer Res *78*. abstract401. https://doi.org/10.1158/1538-7445.AM2018-401.
34. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
35. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNaseI sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390–394.
36. Grubert, F., Zaugg, J.B., Kasowski, M., Ursu, O., Spacek, D.V., Martin, A.R., Greenside, P., Srivas, R., Phanstiel, D.H., Pekowska, A., et al. (2015). Genetic Control of Chromatin States in

Humans Involves Local and Distal Chromosomal Interactions. Cell *162*, 1051–1065.

37. Mumbach, M.R., Satpathy, A.T., Boyle, E.A., Dai, C., Gowen, B.G., Cho, S.W., Nguyen, M.L., Rubin, A.J., Granja, J.M., Kazane, K.R., et al. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat. Genet. *49*, 1602–1612.
38. Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448.
39. Herzel, L., Straube, K., and Neugebauer, K.M. (2018). Long-read sequencing of nascent RNA reveals coupling among RNA processing events. Genome Res. *28*, 1008–1019.
40. van der Wijst, M.G.P., Brugge, H., de Vries, D.H., Deelen, P., Swertz, M.A., Franke, L.; LifeLines Cohort Study; and BIOS Consortium (2018). Single-cell RNA sequencing identifies cell-type-specific cis-eQTLs and co-expression QTLs. Nat. Genet. *50*, 493–497.
41. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.
42. Li, Y., and Kellis, M. (2016). RiVIERA-MT: A Bayesian model to infer risk variants in related traits using summary statistics and functional genomic annotations. bioRxiv. https://doi.org/10.1101/059345.
43. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.