Colocalization of GWAS and eQTL Signals Detects Target Genes

Farhad Hormozdiari,[1] Martijn van de Bunt,[2,3] Ayellet V. Segrè,[4] Xiao Li,[4] Jong Wha J. Joo,[1] Michael Bilow,[1] Jae Hoon Sul,[5,6] Sriram Sankararaman,[1,8] Bogdan Pasaniuc,[7,8] and Eleazar Eskin[1,8,*]

The vast majority of genome-wide association study (GWAS) risk loci fall in non-coding regions of the genome. One possible hypothesis is that these GWAS risk loci alter the individual's disease risk through their effect on gene expression in different tissues. In order to understand the mechanisms driving a GWAS risk locus, it is helpful to determine which gene is affected in specific tissue types. For example, the relevant gene and tissue could play a role in the disease mechanism if the same variant responsible for a GWAS locus also affects gene expression. Identifying whether or not the same variant is causal in both GWASs and expression quantitative trail locus (eQTL) studies is challenging because of the uncertainty induced by linkage disequilibrium and the fact that some loci harbor multiple causal variants. However, current methods that address this problem assume that each locus contains a single causal variant. In this paper, we present eCAVIAR, a probabilistic method that has several key advantages over existing methods. First, our method can account for more than one causal variant in any given locus. Second, it can leverage summary statistics without accessing the individual genotype data. We use both simulated and real datasets to demonstrate the utility of our method. Using publicly available eQTL data on 45 different tissues, we demonstrate that eCAVIAR can prioritize likely relevant tissues and target genes for a set of glucose- and insulin-related trait loci.

Introduction

Genome-wide association studies (GWASs) have successfully detected thousands of genetic variants associated with various traits and diseases.[1–4] The vast majority of genetic variants detected by GWASs fall in non-coding regions of the genome, and it is unclear how these non-coding variants affect traits and diseases.[5] One potential approach to identifying the mechanism of these non-coding variants in disease is through integration of expression quantitative loci (eQTL) studies and GWASs.[5] This approach is based on the concept that a GWAS variant, in some tissues, affects expression at a nearby gene and that both the gene and the tissue might play a role in the disease mechanism.[6,7]

Unfortunately, integrating GWASs and eQTL studies is challenging for two reasons. First, the correlation structure of the genome, known as linkage disequilibrium (LD),[8] produces an inherent ambiguity in interpreting results of genetic studies. Second, some loci harbor more than one causal variant for any given disease. We know that marginal statistics of a variant can be affected by other variants in LD.[8–11] For example, the marginal statistics of two variants in LD can capture a fraction of the effect of each other. Although GWASs have benefited from LD in the human genome by tagging only a subset of common variants to capture a majority of common variants, a fine-mapping process, which attempts to detect true causal variants that are responsible for an association signal at the locus, becomes more challenging. Colocalization determines whether a single variant is responsible for both GWAS and eQTL signals in a locus. Thus, colocalization requires correctly identifying the causal variant in both studies.

Recently, researchers proposed a series of methods[6,12–17] to integrate GWASs and eQTL studies. PrediXscan[7] and TWAS,[17] which impute gene expression and then associate the imputed expression with the trait, are examples of such methods. However, these methods do not provide a basis for determining colocalization of GWAS causal variants and eQTL causal variants. Another class of methods integrates GWASs and eQTL studies to provide insight about the colocalization of causal variants. For example, regulatory trait concordance (RTC)[13] detects variants that are causal in both studies while accounting for LD. RTC is based on the assumption that removing the effect of causal variants from eQTL studies will reduce or eliminate any significant association signal at that locus. Thus, when the GWAS causal variant is colocalized with the eQTL causal variant, re-computing the marginal statistics for the eQTL variant by conditioning on the GWAS causal variant will remove any significant association signal observed in the locus. Sherlock,[12] another method, is based on a Bayesian statistical framework that matches GWAS association signals with eQTL signals for a specific gene in order to detect whether the same variant is causal in both studies. Similar to RTC, Sherlock accounts for the uncertainty of LD. QTLMatch[16] is another proposed method of detecting cases where the most significant GWAS and

[1]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA 90095, USA; [2]Oxford Centre for Diabetes, Endocrinology, & Metabolism, University of Oxford, Oxford OX3 7LJ, UK; [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK; [4]Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [5]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, CA 90095, USA; [6]Semel Center for Informatics and Personalized Genomics, University of California, Los Angeles, Los Angeles, CA 90095, USA; [7]Department of Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA; [8]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA
*Correspondence: eeskin@cs.ucla.edu
http://dx.doi.org/10.1016/j.ajhg.2016.10.003.

eQTL variants are colocalized as a result of a causal relationship or coincidence. COLOC,[14,15] a method expanded from QTLMatch, is the state-of-the-art method that colocalizes GWAS and eQTL signals. COLOC utilizes an approximate Bayes factor to estimate the posterior probabilities that a variant is causal in both GWASs and eQTL studies. Unfortunately, most existing colocalization methods that utilize summary statistics assume the presence of only one causal variant in any given locus for both GWASs and eQTL studies. As we show below, this assumption reduces the accuracy of results when the locus contains multiple causal variants.

In this paper, we present a probabilistic model for integrating GWAS and eQTL data. For each study, we use only the reported summary statistics and simultaneously perform statistical fine-mapping to optimize integration. Our approach, eCAVIAR (eQTL and GWAS Causal Variant Identification in Associated Regions), extends the CAVIAR[18] framework to explicitly estimate the posterior probability that the same variant is causal in both a GWAS and eQTL study while accounting for the uncertainty of LD. We apply eCAVIAR to colocalize variants that pass the genome-wide significance threshold in a GWAS. For any given peak variant identified in a GWAS, eCAVIAR considers a collection of variants around that peak variant as one single locus. This collection includes the peak variant itself, M variants upstream of this peak variant, and M variants downstream of this peak variant (e.g., M can be set to 50). Then, for all variants in a locus, we consider their marginal statistics obtained from the eQTL study in all tissues and all genes. We consider only genes and tissues in which at least one of the genes is an eGene.[19,20] eGenes are genes that have at least one significant variant (p value $\leq 10^{-5}$ when corrected for multiple hypothesis) associated with the expression of that gene. We assume that the posterior probability that the same variant is causal in both a GWAS and eQTL study is independent. Thus, this posterior probability is equal to the product of posterior probabilities that a given variant is causal in a GWAS and eQTL study. We refer to the amount of support for a variant responsible for the associated signals in both studies as the colocalization posterior probability (CLPP).

Our framework allows for multiple variants to be causal in a single locus, a phenomenon that is widespread in eQTL data and referred to as allelic heterogeneity (AH). Our approach can accurately quantify the amount of support for a variant responsible for the associated signals in both studies and identify scenarios where there is support for an eQTL-mediated mechanism. Moreover, we can identify scenarios where the variants underlying both studies are clearly different. Utilizing simulated datasets, we show that eCAVIAR has high accuracy in detecting target genes and relevant tissues. Furthermore, the amount of CLPP depends on the complexity of the LD.

We applied our method to colocalize the MAGIC (Meta-analyses of Glucose and Insulin-Related Trait Consortium)[21–24] GWAS dataset and publicly available eQTL data on 45 different tissues. We obtained 44 tissues from the Genotype-Tissue Expression (GTEx) eQTL dataset (release v.6, dbGaP: phs000424.v6.p1)[19] and one pancreatic islet tissue from the van de Bunt et al. study.[25] Our results provide insight into disease mechanisms by identifying specific GWAS loci that share a causal variant with eQTL studies in a tissue. In addition, we have identified several loci where GWAS and eQTL causal variants appear to be different, suggesting that the genetic factors underlying disease mechanisms are more complex than previously thought.

Material and Methods

CAVIAR Model for Fine-Mapping

Standard GWAS and Indirect Association

We collect quantitative traits for N individuals and genotypes for all individuals at M SNPs (variants). In this case, we collect data for one phenotype and the expression of multiple genes. We assume that both the phenotype and gene expression have at least one significant variant. To simplify the description of our method, we assume that the number of individuals and the pairwise Pearson's correlations of genotype (LD) in both the GWAS and eQTL study are the same. (In Appendix A, we describe a more general model where the number of individuals and LD in both the GWAS and eQTL study are not the same.) Let $Y^{(\mathrm{p})}$ indicate an $N \times 1$ vector of the phenotypic values where $y_j^{(\mathrm{p})}$ denotes the phenotypic value for the j^{th} individual. We use $Y^{(\mathrm{e})}$ to indicate an $N \times 1$ vector of gene expression collected for one gene of interest, for which there exists one significant variant associated with the expression of that gene. Let G indicate an $N \times M$ matrix of genotype information where G_i is an $N \times 1$ vector of minor allele counts for all N individuals at the i^{th} variant. In this setting, g_{ji} indicates the j^{th} element from vector G_i, or the minor allele count for the j^{th} individual. In diploid genomes, such as those of humans, we can have three possible minor allele counts: $g_{ji} = \{0, 1, 2\}$. We standardize both the phenotypes and the genotypes to mean 0 and variance 1, where X is the standardized matrix of G. Let X_i denote an $N \times 1$ vector of standardized minor allele counts for the i^{th} variant. We assume an "additive" Fisher's polygenic model, which is widely used by the GWAS community. In Fisher's polygenic model, the phenotypes follow a normal distribution. The additive assumption implies that each variant contributes linearly to the phenotype. Thus, we consider the following linear model:

$$Y^{(\mathrm{p})} = \mu^{(\mathrm{p})} 1 + \sum_{i=1}^{M} \beta_i^{(\mathrm{p})} X_i + \boldsymbol{e}^{(\mathrm{p})},$$

$$Y^{(\mathrm{e})} = \mu^{(\mathrm{e})} 1 + \sum_{i=1}^{M} \beta_i^{(\mathrm{e})} X_i + \boldsymbol{e}^{(\mathrm{e})},$$

where $\mu^{(\mathrm{p})}$ is the phenotypic mean and $\mu^{(\mathrm{e})}$ is the gene-expression mean. Let $\beta_i^{(\mathrm{p})}$ and $\beta_i^{(\mathrm{e})}$ be the effect size of the i^{th} variant toward the phenotypes and gene expression, respectively. In addition, $\boldsymbol{e}^{(\mathrm{p})}$ is the environment and measurement error toward the collected phenotype, and $\boldsymbol{e}^{(\mathrm{e})}$ is the environment and measurement error toward the gene expression. In this model, we assume

that $\boldsymbol{e}^{(p)}$ is a vector of independent and identically distributed and normally distributed random variables. Let $\boldsymbol{e}^{(p)} \sim N(0, \sigma_e^{(p)2}\boldsymbol{I})$, where $\sigma_e^{(p)}$ is a covariance scalar and $\boldsymbol{I}$ is an $N \times N$ identity matrix. In our setting, we have the marginal statistics of M variants for the phenotype of interest and gene expression. Let $S^{(p)} = \{s_1^{(p)}, s_2^{(p)}, \cdots s_M^{(p)}\}$ indicate the marginal statistics for the phenotype of interest, and let $S^{(e)} = \{s_1^{(e)}, s_2^{(e)}, \cdots s_M^{(e)}\}$ indicate the marginal statistics for gene expression. The joint distribution of the marginal statistics, given the true effect sizes, follows a multivariate normal (MVN) distribution and is similar to that found in our previous works.[18,26–28] Thus, we have

$$\begin{aligned}\left(S^{(p)} \mid \Lambda^{(p)}\right) &\sim \mathcal{N}\left(\Sigma\Lambda^{(p)}, \Sigma\right),\\ \left(S^{(e)} \mid \Lambda^{(e)}\right) &\sim \mathcal{N}\left(\Sigma\Lambda^{(e)}, \Sigma\right),\end{aligned} \qquad \text{(Equation 1)}$$

where Σ is the pairwise Pearson's correlations of genotypes. Let $\Lambda^{(p)} = \{\lambda_1^{(p)}, \lambda_2^{(p)}, \cdots \lambda_M^{(p)}\}$ and $\Lambda^{(e)} = \{\lambda_1^{(e)}, \lambda_2^{(e)}, \cdots \lambda_M^{(e)}\}$ be the true standardized effect size for all the variants of the desired phenotype and gene expression, respectively. The true effect size is zero for a non-causal variant and non-zero for a causal variant. Let $\Sigma\Lambda^{(p)}$ and $\Sigma\Lambda^{(e)}$ be the LD-induced non-centrality parameter (NCP) for the desired phenotype and gene expression, respectively.

CAVIAR Generative Model for a Single Phenotype

We introduce a new variable, $C^{(p)}$, which is an $M \times 1$ binary vector. We refer to this binary vector as the causal status. The causal status indicates which variants are causal and which are not. We set $c_i^{(p)}$ to 1 if the i^{th} variant is causal; otherwise, we set it to 0. In CAVIAR,[18,27] we introduce a prior on the vector of effect sizes by utilizing the MVN distribution. Given the vector of causal status, we define this prior on the vector of effect sizes as

$$\left(\Lambda^{(p)} \mid C^{(p)}\right) \sim \mathcal{N}\left(0, \sigma^{(p)2}\Sigma_c^{(p)}\right), \qquad \text{(Equation 2)}$$

where $\Sigma_c^{(p)}$ is a diagonal matrix and $\sigma^{(p)2}$ is a constant that indicates the variance of our prior over the GWAS NCPs. We set $\sigma^{(p)2}$ to 5.2.[18,27] The diagonal elements of $\Sigma_c^{(p)}$ are set to 1 or 0 such that for variants selected as causal in $C^{(p)}$, their corresponding diagonal elements in $\Sigma_c^{(p)}$ are set to 1; otherwise, we set them to 0. CAVIAR uses this prior as a conjugate prior to compute the likelihood of each possible causal status. The joint distribution of the marginal statistics given the causal status is as follows:

$$\left(S^{(p)} \mid C^{(p)}\right) \sim \mathcal{N}\left(0, \Sigma + \sigma^{(p)2}\Sigma\Sigma_c^{(p)}\Sigma\right). \qquad \text{(Equation 3)}$$

In a similar way, for the gene of interest for which we perform eQTL mapping, we have

$$\left(\Lambda^{(e)} \mid C^{(e)}\right) \sim \mathcal{N}\left(0, \sigma^{(e)2}\Sigma_c^{(e)}\right), \qquad \text{(Equation 4)}$$

where $\Sigma_c^{(e)}$ is a diagonal matrix and $\sigma^{(e)2}$ is set to 5.2.[18,27] The diagonal elements of $\Sigma_c^{(e)}$ are set to 1 or 0. For variants selected as causal in $C^{(e)}$, their corresponding diagonal elements in $\Sigma_c^{(e)}$ are set to 1; otherwise, we set them to 0.

eCAVIAR Computes the Colocalization Posterior Probability for a GWAS and eQTL Study

Given the marginal statistics for a GWAS and eQTL study, which are denoted by $S^{(p)}$ and $S^{(e)}$, respectively, we want to compute the CLPP. CLPP is the probability that the same variant is causal in both studies. For simplicity, we compute the CLPP for the i^{th} variant. We define the CLPP for the i^{th} variant as $P(c_i^{(p)} = 1, c_i^{(e)} = 1 \mid S^{(p)}, S^{(e)})$, and we use ϕ_i to indicate the CLPP for the i^{th} variant. We utilize the law of total probability to compute the summation probability of all causal statuses where the i^{th} variant is causal in both the GWAS and eQTL study and other variants can be causal or non-causal. Thus, the above equation can be extended as follows:

$$\begin{aligned}\phi_i &= P\left(c_i^{(p)} = 1, c_i^{(e)} = 1 \mid S^{(p)}, S^{(e)}\right)\\ &= \sum_{C_{/i}^{*(p)} \in \{0,1\}^{M-1}} \sum_{C_{/i}^{*(e)} \in \{0,1\}^{M-1}} P\left(C_{/i}^{(p)} = C_{/i}^{*(p)}, C_{/i}^{(e)} = C_{/i}^{*(e)}, c_i^{(p)} = 1, c_i^{(e)} = 1 \mid S^{(p)}, S^{(e)}\right)\\ &= \sum_{C^{*(p)} \in \{0,1\}^M} \sum_{C^{*(e)} \in \{0,1\}^M} P\left(C^{(p)} = C^{*(p)}, C^{(e)} = C^{*(e)} \mid S^{(p)}, S^{(e)}\right) \mathbb{I}\left(c_i^{*(p)} = 1, c_i^{*(e)} = 1\right),\end{aligned} \qquad \text{(Equation 5)}$$

where $C_{/i}^{(p)}$ and $C_{/i}^{(e)}$ are vectors of causal status for all variants, excluding the i^{th} variant for the phenotype of interest and gene expression. Let $\mathbb{I}()$ be an indicator function defined as follows:

$$\mathbb{I}\left(c_i^{*(p)} = 1, c_i^{*(e)} = 1\right) = \begin{cases} 1 & c_i^{*(p)} \text{ and } c_i^{*(e)} \text{ are causal} \\ 0 & o/w \end{cases}. \qquad \text{(Equation 6)}$$

Utilizing the Bayes' rule, we compute the CLPP as follows:

$$\phi_i = \frac{\sum_{C^{*(p)}} \sum_{C^{*(e)}} P\left(S^{(p)}, S^{(e)} \mid C^{(p)} = C^{*(p)}, C^{(e)} = C^{*(e)}\right) P\left(C^{*(p)}, C^{*(e)}\right) \mathbb{I}\left(c_i^{*(p)} = 1, c_i^{*(e)} = 1\right)}{\sum_{C^{*(p)}} \sum_{C^{*(e)}} P(S^{(p)}, S^{(e)} \mid C^{(p)} = C^{*(p)}, C^{(e)} = C^{*(e)}) P(C^{*(p)}, C^{*(e)})}, \qquad \text{(Equation 7)}$$

where $P(C^{*(p)}, C^{*(e)})$ is the prior probability of the causal status of $C^{*(p)}$ and $C^{*(e)}$ for the GWAS and eQTL study, respectively. We assume that the prior probability over the causal status for the GWAS and eQTL study is independent: $P(C^{*(p)}, C^{*(e)}) = P(C^{*(p)})P(C^{*(e)})$. To compute the prior of causal status, we use the same assumptions that are widely used in fine-mapping methods,[18,27,29]

whereby the probability of causal status follows a binomial distribution where the probability that a variant is causal is equal to γ. Thus, this prior is equal to $P(C^{*(p)}) = \prod_{i=1}^{M} \gamma^{c_i^{*(p)}} (1-\gamma)^{1-c_i^{*(p)}}$, and γ is set to 0.01.[18,30–32]

GWASs and eQTL studies are usually performed on independent sets of individuals. Furthermore, given the causal status in both studies, the marginal statistics for these two studies are independent. We have $P(S^{(p)}, S^{(e)} \mid C^{*(p)}, C^{*(e)}) = P(S^{(p)} \mid C^{*(p)}) P(S^{(e)} \mid C^{*(e)})$. Thus, we simplify Equation 7 and compute the CLPP as follows:

$$\phi_i = \frac{\sum_{C^{*(p)}} P(S^{(p)} \mid C^{(p)} = C^{*(p)}) P(C^{*(p)}) \mathbb{I}\left(c_i^{*(p)} = 1\right)}{\sum_{C^{*(p)}} P(S^{(p)} \mid C^{(p)} = C^{*(p)}) P(C^{*(p)})} \times \frac{\sum_{C^{*(e)}} P(S^{(e)} \mid C^{(e)} = C^{*(e)}) P(C^{*(e)}) \mathbb{I}\left(c_i^{*(e)} = 1\right)}{\sum_{C^{*(e)}} P(S^{(e)} \mid C^{(e)} = C^{*(e)}) P(C^{*(e)})}. \quad \text{(Equation 8)}$$

According to the above equation, the probability that the same variant is causal in both the GWAS and eQTL study is independent. This probability is equal to the multiplication of two probabilities: (1) the probability that the variant is causal in the GWAS and (2) the probability that the same variant is causal in the eQTL study. Thus, we compute the CLPP as $P(c_i^{(p)} = 1, c_i^{(e)} = 1 \mid S^{(p)}, S^{(e)}) = P(c_i^{(p)} = 1 \mid S^{(p)}) \times P(c_i^{(e)} = 1 \mid S^{(e)})$, where $P(c_i^{(p)} = 1 \mid S^{(p)})$ and $P(c_i^{(e)} = 1 \mid S^{(e)})$ are computed from the first and second parts of Equation 8, respectively.

Detecting Target Genes and Relevant Tissues

In the previous sections, we described the process of computing the CLPP score for each variant in a locus for a given eGene in a tissue. In this section, we describe a systematic way to detect the target genes and relevant tissues.

We compute the CLPP score for every GWAS significant variant. Thus, for a given GWAS variant, an eGene that has a CLPP score above the colocalization cutoff is considered a target gene. In addition, we consider tissues from which the target genes are obtained as the relevant tissues. Moreover, we can rank the relevant tissues and target genes for a given GWAS significant variant according to their CLPP scores. Thus, we utilize the magnitude of CLPP to rank the tissues and genes on the basis of their importance for a given GWAS risk locus.

Generating Simulated Datasets

Simulating Genotypes

We first simulate genotype data starting from the real genotypes obtained from the European population in the 1000 Genomes data.[33,34] In order to simulate the genotypes, we utilize HAPGEN2[35] software, which is widely used to generate genotypes. We focus on chromosome 1 and the GWAS variants obtained from the NHGRI catalog.[36] We consider 200-kb windows around the lead SNP to generate a locus. Then, we filter out monomorphic SNPs and SNPs with a low minor allele frequency (MAF $\leq$ 0.01) inside a locus.

Simulating Summary Statistics Directly from LD Structure

We generate an LD matrix for a locus by computing the Pearson's correlations of each pair of variants from the genotypes. Then, we generate marginal summary statistics for each locus by assuming that the marginal summary statistics follow the MVN distribution utilized in our previous studies.[18,26–28,37,38] We measure the strength of a causal variant on the basis of NCPs. We set the NCP of the causal variant to obtain a certain statistical power. The NCPs of the non-causal variants are set to 0. The statistical power is the probability of detecting a causal variant under the assumption that the causal variant is present. The statistical power is computed as follows:

$$\text{power} = 1 - \frac{1}{\sqrt{2\pi}} \int_{\Phi^{-1}(\alpha/2)+\lambda}^{\Phi^{-1}(1-\alpha/2)+\lambda} e^{-\frac{1}{2}x^2} dx = \Phi\left(\Phi^{-1}(\alpha/2) + \lambda\right) + 1 - \Phi\left(\Phi^{-1}(1-\alpha/2) + \lambda\right),$$

where α is the significant threshold. Moreover, Φ and Φ^{-1} denote the cumulative density function (CDF) and the inverse of CDF, respectively, for the standard normal distribution. In our experiment, the NCP is computed for the genome-wide significance level ($\alpha = 10^{-8}$). We use a binary search to compute the NCP for a desired statistical power.

Simulating Summary Statistics with a Linear Additive Model

We utilize 100 variants in a locus to generate the simulated phenotypes from the simulated genotypes. We simulate the phenotypes by assuming the following linear additive model:

$$Y = \sum_{i=1}^{M} \beta_i X_i + \boldsymbol{e}, \quad \text{(Equation 9)}$$

where $\boldsymbol{e} \sim N(0, \sigma_e^2)$. We generate the effect size of the causal variant from a normal distribution with mean 0 and variance σ_g^2 / M_c, where M_c indicates the number of causal variants in a locus. Furthermore, we set the effect size to 0 for variants that are not causal. Thus, the effect size for each variant is simulated as follows:

$$\begin{cases} \beta_i = 0 & \text{if the } i^{\text{th}} \text{ variant is non-causal} \\ \beta_i \sim N\left(0, \sigma_g^2 / M_c\right) & \text{if the } i^{\text{th}} \text{ variant is causal} \end{cases}.$$

After simulating the phenotype for all the individuals, we utilize linear regression to estimate the effect sizes and the marginal statistics for all M variants in a locus. In our simulations, M is equal to 100.

Results

Overview of eCAVIAR

The goal of our method is to identify target genes and the most relevant tissues for a given GWAS risk locus while accounting for the uncertainty of LD. Target genes are genes with expression levels that affect the phenotype (e.g., disease status) of interest. Our method detects the target gene and the most relevant tissue by utilizing our proposed quantity of CLPP. eCAVIAR estimates the CLPP, which is the probability that the same variant is causal in both a GWAS and eQTL study. eCAVIAR computes the CLPP by utilizing the marginal statistics (e.g., Z score) obtained from GWAS and eQTL analyses, as well as the LD structure of genetic variants in each locus. LD can be computed from genotype data or approximated from existing datasets, such as the 1000 Genomes data[33,34] or HapMap.[39] We show in the Material and Methods that the marginal statistics of both the GWAS and eQTL study follow a MVN distribution given the causal variants and effect sizes for both studies. We use the MVN distribution to estimate the CLPP. We show that the CLPP is equal to the product of the posterior probability that the variant is causal in the GWAS and the posterior probability that the variant is causal in the eQTL study.

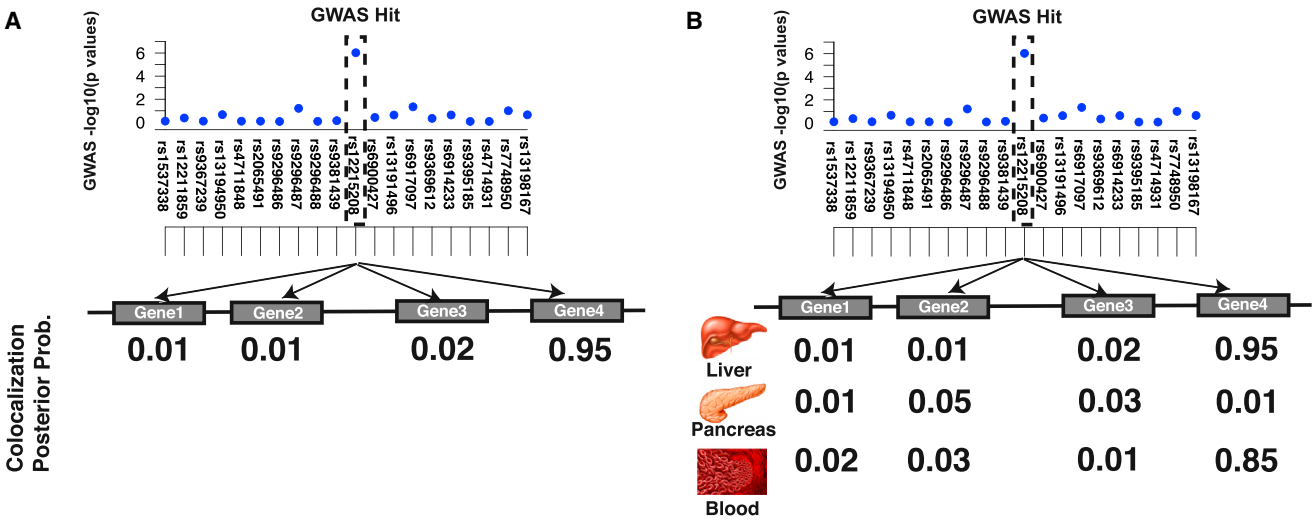

Figure 1. Overview of Our Method for Detecting the Target Gene and Most Relevant Tissue
We compute the CLPP for all genes and all tissues.
(A) A simple case where we have only one tissue and want to find the target gene. We consider all genes for this GWAS risk locus and observe that gene 4 has the highest CLPP. Thus, the target gene is gene 4.
(B) We have three tissues and utilize the quantity of CLPP. Thus, the target gene is gene 4 again. Moreover, in this example, liver and blood are considered the relevant tissues for this GWAS risk locus, whereas the pancreas is not relevant.

Calculating the posterior probability of a causal variant is computationally intractable. Therefore, we assume the presence of at most six causal variants in a locus.

The estimated CLPP for a GWAS risk locus and a gene, which is obtained from eQTL studies, can be used for inferring specific disease mechanisms. First, we identify genes that have expression levels affected by a GWAS variant. These genes are referred to as target genes. Second, we identify in which tissues the eQTL variant has an effect. To identify target genes, we compute the CLPP for all genes in the GWAS risk locus. Genes that have a significantly higher CLPP are selected as target genes (Figure 1A). Similarly, we compute the CLPP for all tissues and identify relevant tissues as those with comparatively high CLPP values (Figure 1B). Figure 1B shows that the GWAS risk locus affects gene 4, and the relevant tissues are liver and blood. However, Figure 1B indicates that the pancreas is not a relevant tissue for this GWAS risk locus. Another application of CLPP is to identify loci where the causal variants between a GWAS and eQTL study are different. We can identify these loci if the CLPP is low for all variants in the loci and if there are statistically significant variants in both the GWAS and eQTL study.

To better motivate the behavior of CLPP, we consider the following four scenarios in Figure 2. In the first scenario, the same variant has effects in both the GWAS and eQTL study. Thus, its CLPP is high (Figure 2A). In the second scenario, we consider that the variant is associated with a phenotype in the GWAS and not associated with gene expression. In this case, the quantity of CLPP is low (Figure 2B). In the third scenario, we consider that the variant is not associated with a phenotype in the GWAS. However, it is associated with expression of a gene. In this case, the CLPP is not computed for this variant. Rather, we compute the CLPP for GWAS risk loci that are considered significant. In the fourth scenario, we have a variant that appears significant in both the GWAS and eQTL study. However, other variants in both studies are also significant because of high LD with the causal variant. The complex LD (see Figure 2C) of these variants results in a low CLPP. Here, we remain uncertain about which variants are actual causal variants. Finally, Figure 2D illustrates an example in which there is more than one causal variant. This demonstrates that assuming the presence of a single causal variant can result in underestimation of CLPP. In this example, we have a locus with 35 variants (SNPs), and we have two causal variants (SNP6 and SNP26) that are not in high LD with each other. If we assume that we have only one causal variant, there are 35 possible causal variants for this locus, and most of the causal variants have a very low likelihood. The likelihoods of selecting either SNP6 or SNP26 as causal are similar, and they are higher than the likelihoods of selecting any other variant as causal. In this example, the estimated posterior probability that SNP6 or SNP26 is causal is 50%. Thus, the estimated CLPP for SNP6 or SNP26 is 25%. However, if we allow more than one causal variant in the locus, all sets of causal variants have very low likelihood values, except the set in which both SNP6 and SNP26 are selected as causal. In this case, the posterior probability that SNP6 or SNP26 is causal is close to 1. In this case, we assume that we have more than one causal variant in this locus given that the CLPP values of SNP6 and SNP26 are close to 1.

eCAVIAR Accurately Computes the CLPP

In this section, we use simulated datasets to assess the accuracy of our method. We simulated summary statistics by utilizing the MVN distribution used in our previous studies.[18,26–28,38] More details on simulated data are

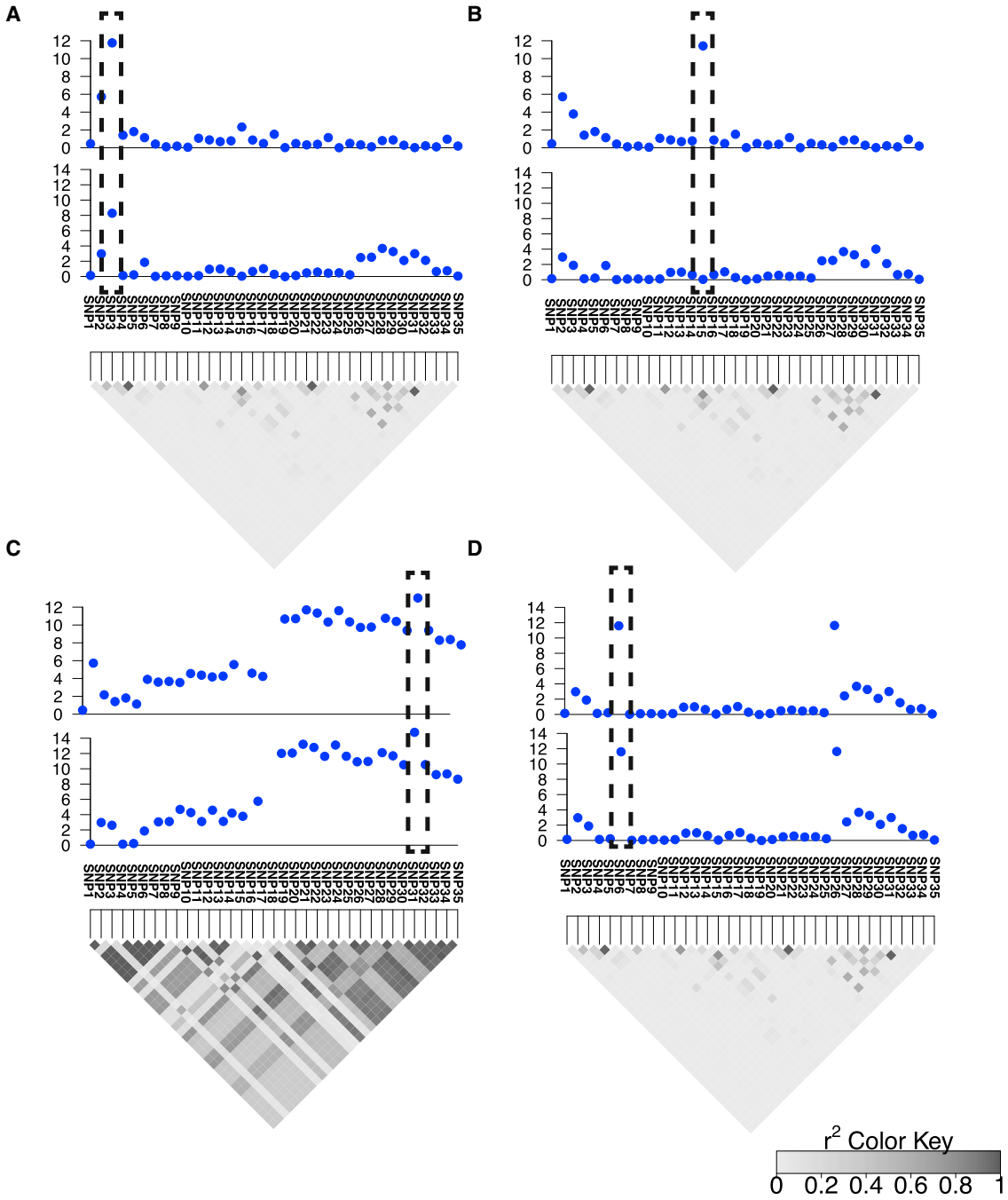

Figure 2. Overview of eCAVIAR

Broadly, eCAVIAR aligns the causal variants in an eQTL study and GWAS. The x axis is the variant (SNP) location, and the y axis is the significance score (−log of p value) for each variant. The gray triangle indicates the LD structure, and every diamond in this triangle indicates the Pearson's correlation. The darker the diamond, the higher the correlation; and the lighter the diamond, the lower the correlation between the variants.

(A) In the case where the causal variants are aligned, the colocalization posterior probability (CLPP) is high for the variant that is embedded in the dashed black rectangle.

(B) However, in the case where the causal variants are not aligned (the causal variants are not the same variants), the quantity of CLPP is low for the variant that is embedded in the dashed black rectangle.

(C) In this case, the LD is high, which implies that the uncertainty is high as a result of LD, and the CLPP value is low for the variant that is embedded in the dashed black rectangle.

(D) A case where a locus has two independent causal variants. If we consider that we have only one causal variant in a locus, then the CLPP of the causal variants is estimated to be 0.25. However, if we allow more than one causal variant in the locus, eCAVIAR estimates the CLPP to be 1.

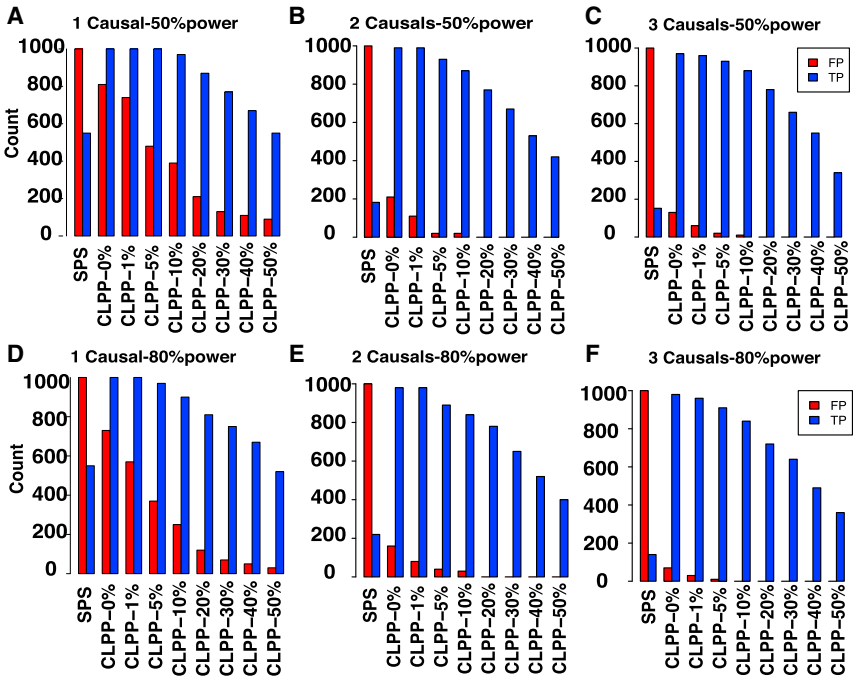

Figure 3. eCAVIAR Is Robust to the Presence of AH

We simulated marginal statistics directly from the LD structure for an eQTL study and GWAS. In both studies, we implanted one, two, or three causal variants on which the statistical power was 50% (A–C, respectively) or 80% (D–F, respectively). eCAVIAR had a low TP for a high cutoff and a low FP. This indicates that eCAVIAR has high confidence in detecting a colocalized locus in both the GWAS and eQTL study, even in the presence of AH.

provided in the Material and Methods. In one set of simulations, we fixed the effect size of a genetic variant so that the statistical power for the causal variant was 50%. In another set, we fixed the effect size so that the power was 80%. We considered two cases. The first case included only one causal variant in both studies. The second case included more than one causal variant in these studies. For both scenarios, we simulated two datasets. In the first dataset, we implanted a shared causal variant. We generated 1,000 simulated studies, which we used to compute the true-positive rate (TP). In the second dataset, we implanted a different causal variant in the eQTL study and GWAS. We filtered out cases where the most significant variant was different between the two studies. As in the previous case, we generated 1000 simulated studies.

eCAVIAR Is Accurate in the Case of One Causal Variant

We applied eCAVIAR to the simulated datasets and computed the CLPP for each case. We used different cutoffs to determine whether or not a variant was shared between the two studies. For each cutoff, we computed the false-positive rate (FP) and TP. The baseline method detects the most significant variant in a study as the causal variant. Thus, in the baseline method, we have colocalization when the eQTL study and GWAS share the same most significant variant. We refer to this method as the shared peak SNP (SPS) method. The results are shown in Figures 3A and 3D. Moreover, the same results are plotted in a receiver operating characteristic (ROC) curve (Figure S1). Our method has a higher TP and lower FP than SPS. However, eCAVIAR has a low TP when the cutoff for CLPP is high. Furthermore, eCAVIAR has an extremely low FP. Our results imply that eCAVIAR has high confidence for selecting loci to be colocalized between a GWAS and eQTL study. eCAVIAR is conservative in selecting a locus to be colocalized. Given the high cutoff of CLPP, eCAVIAR can miss some true colocalized loci. However, loci that are selected by eCAVIAR to be colocalized are likely to be predicted correctly.

The computed CLPP depends on the complexity of the LD at the locus. We applied eCAVIAR to the simulated datasets and computed the CLPP (Figure S2). Here, the average quantity of CLPP decreased as we increased the Pearson's correlation (r) between paired variants. This effect increased the complexity of LD between the two variants. Furthermore, the 95% confidence interval for the computed quantity increased as we increased the Pearson's correlation. This result implies that the computed CLPP can be small for a locus with complex LD, even when a variant is colocalized in both a GWAS and eQTL study.

eCAVIAR Is Robust to the Presence of AH

The presence of more than one causal variant in a locus is a phenomenon referred to as AH. AH can confound the association statistics in a locus, and colocalization for a locus harboring AH is challenging. In order to investigate the effect of AH, we performed the following simulations. We implanted two or three causal variants in both the GWAS and eQTL study, and we then generated the marginal statistics by using the MVN distribution mentioned in the previous section. Next, we computed the TP and FP for eCAVIAR and SPS (see Figure 3). In the case of eCAVIAR, colocalization is considered true when all colocalized variants are detected. However, for SPS, colocalization is considered true when at least one of the colocalized variants is detected. Figures 3A, 3B, and 3C illustrate the results of one, two, and three causal variants, respectively, when the statistical power was 50%. Similarly, Figures 3D, 3E, and 3F illustrate the results of one, two, and three causal variants, respectively, when the statistical power was 80%. Interestingly, SPS had a very low TP when there were two or three causal variants (see Figure 3). This implies that SPS is not accurate when AH is present. Similar to cases with one single casual variant (see Figures 3A and 3D), eCAVIAR had a very low FP when there were two or three causal variants (see Figures 3B, 3C, 3E, and 3F). This implies that eCAVIAR has high confidence in detecting a locus to be colocalized between a GWAS and eQTL study.

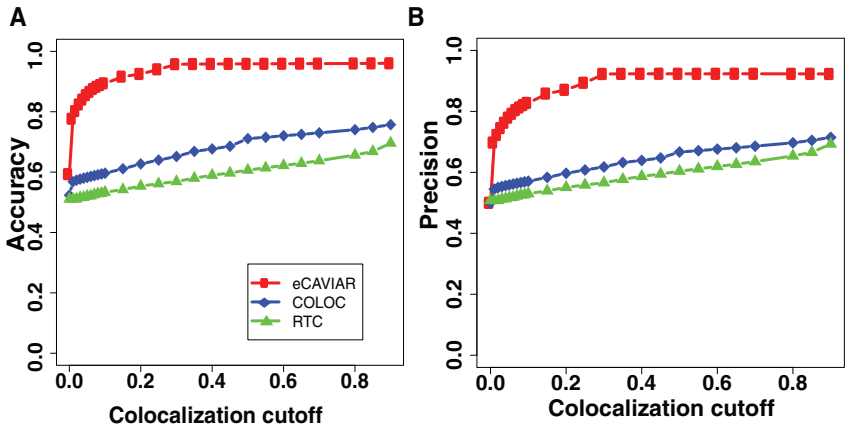

Figure 4. eCAVIAR Is More Accurate Than Existing Methods for Regions with One Causal Variant

We compare the accuracy and precision of eCAVIAR with those of the two existing methods (RTC and COLOC). The x axis is the colocalization cutoff threshold. In these datasets, we implanted one causal variant, and we utilized simulated genotypes. We simulated the genotypes by using HAPGEN2[35] software. We used the European population from 1000 Genomes data[33,34] as the starting point to simulate the genotypes. The accuracy and precision of all three methods are shown in (A) and (B), respectively. We computed the TP (true-positive rate), TN (true-negative rate), FN (false-negative rate), and FP (false-positive rate) for the set of simulated datasets for which we generated the marginal statistics in a linear model. Accuracy = (TP + TN)/(TP + FP + FN + TN), and precision = TP/(TP + FP). We set the non-colocalization cutoff threshold to 0.001. We observed that eCAVIAR and COLOC had higher accuracy and precision than RTC.

We generated simulated datasets where the causal variants were different between the two studies. We computed the CLPP for all variants in a region. Our experiment indicated that eCAVIAR has a high TN and an extremely low FN. eCAVIAR has a high negative predictive value (NPV): NPV = TN/(TN + FN). These results are shown in Figures S3 and S4. Thus, eCAVIAR can detect with high accuracy loci where the causal variants are different between the two studies.

eCAVIAR Is More Accurate Than Existing Methods

Here, we compare the results of eCAVIAR with those of RTC[13] and COLOC,[14] two common methods for eQTL and GWAS colocalization. The procedure in the previous section can be used to simulate datasets; however, RTC is not designed to work with summary statistics. In order to provide a dataset compatible with RTC, we simulated eQTL and GWAS phenotypes under a linear additive model in which we used simulated genotypes obtained from HAPGEN2.[35] More details on the simulated datasets are provided in the Material and Methods.

We compare the accuracy, precision, and recall rate of all three methods. Each method computes a probability that a variant is causal in both a eQTL study and GWAS. In order to determine this probability for our comparison, we need to select two cutoff thresholds. We devised one threshold for detecting variants that are colocalized in both studies and another threshold for detecting variants that are not colocalized. Here, we consider a variant to be causal in both studies if the probability of colocalization is greater than the colocalization cutoff threshold. The second cutoff threshold is used for detecting variants that are not causal in both studies. We consider a variant to be non-causal in both studies if the probability of colocalization is less than the non-colocalization cutoff threshold. In our experiment, we set the non-colocalization cutoff threshold at 0.1% and the colocalization cutoff threshold at a value ranging from 0.1% to 90%.

eCAVIAR outperformed the existing methods when the locus contained one causal variant. We observed that all three methods had a similarly high recall rate (see Figure S5). eCAVIAR had much higher accuracy and precision than RTC (see Figure 4). Next, we considered the performance of the three methods when the locus had AH. We used the same simulation described in this section, but instead we implanted two causal variants instead of one. In this setting, eCAVIAR had higher accuracy and precision than COLOC and RTC. However, RTC had a slightly higher recall rate than eCAVIAR. Moreover, RTC tended to perform better than COLOC in the presence of AH (see Figure 5). This result indicates that eCAVIAR is more accurate than existing methods—even in the presence of AH. However, if a locus contains only one causal variant, COLOC performs better than RTC. In cases with more than one causal variant, RTC performs better. These results were obtained when we set the non-colocalization cutoff threshold to 0.1%. We changed this value to 0.01% to check the robustness of eCAVIAR and observed that even when we used different values of non-colocalization, eCAVIAR outperformed existing methods (see Figures S6 and S7). In all of the above experiments, we implanted the causal variants uniformly in the locus. Next, we simulated causal variants in genomic variants enriched with functional annotations. In order to simulate the genomic enrichment, we used the same process utilized in PAINTOR.[40] We observed that eCAVIAR outperformed existing methods in these experiments (Figures S8 and S9).

Thus, eCAVIAR performs better than COLOC and RTC, the pioneering methods for eQTL and GWAS colocalization. COLOC and RTC require different input data to perform the colocalization. COLOC requires only the marginal statistics from a GWAS and eQTL study. Unlike eCAVIAR, COLOC and RTC do not require the LD structure of genetic variants in a locus. However, RTC requires individual-level data (genotypes and phenotypes) and is not applicable to datasets for which we have access to only the summary statistics.

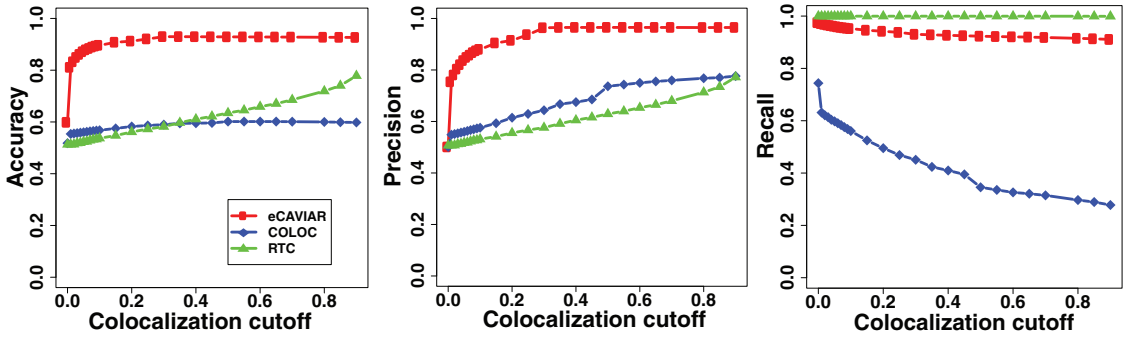

Figure 5. eCAVIAR Is More Accurate Than Existing Methods in the Presence of AH
To generate the datasets, we used a process similar to that shown in Figure 4. However, in this case, we implanted two causal variants. We simulated the genotypes by using HAPGEN2[35] software. We used the European population from 1000 Genomes data[33,34] as the starting point to simulate the genotypes. We compared the accuracy, precision, and recall rate. In these results, eCAVIAR tended to have higher accuracy and precision than RTC and COLOC. However, RTC had a slightly higher recall rate.

Effect of eQTL Sample Size on CLPP

We know that the statistical power to detect a true casual variant increases as we increase the number of samples in a GWAS. Because most GWAS sample sizes are in the order of thousands of samples, we aimed to investigate the effect of eQTL sample size on colocalization.

We simulated datasets in which we set the number of GWAS samples to 5,000. Then, we varied the number of eQTL samples from 500 to 3,500. We simulated the effect size for the causal variant in the eQTL study such that it accounted for 1%, 4%, and 10% of heritability. We computed the CLPP for different cases; the distribution of CLPP is shown in a boxplot in Figure S10. The red horizontal line indicates the 1% colocalization cutoff used for eCAVIAR. We observed that when the causal variant accounted for 1% of heritability, we required at least 2,000 eQTL samples. Conversely, when the causal variant accounted for a larger portion of heritability, eCAVIAR required fewer samples.

Using eCAVIAR to Integrate Available eQTLs for 45 Tissues and MAGIC Datasets

We utilized the MAGIC dataset and GTEx dataset[19] to detect the target gene and most relevant tissue for each GWAS risk locus. MAGIC datasets consist of eight phenotypes.[21] These phenotypes are as follows: fasting glucose (FG), fasting insulin, fasting proinsulin (FP), HOMA-B (β cell function), HOMA-IR (insulin resistance), Hb1Ac (hemoglobin A1c test for diabetes), 2 hr glucose, and 2 hr insulin after an oral glucose-tolerance test. In our analysis, we used FG and FP phenotypes containing the most significant loci. FG phenotypes had 15 variants, and FP phenotypes had ten variants reported to be significantly associated with these phenotypes by previous studies.[21,22] We considered 44 tissues included in the GTEx Portal (release v.6, dbGaP: phs000424.v6.p1).[19] In addition, we used previously published data on human pancreatic islets,[25] a key tissue in glucose metabolism that is not captured in the GTEx data. Table S1 lists tissues and the number of individuals for each tissue.

We wanted to detect the most relevant tissue and a target gene for each of the previously reported significant GWAS variants. eCAVIAR utilizes the marginal statistics of all variants in a locus obtained from a GWAS and eQTL study. We obtained each locus by considering 50 variants upstream and downstream of the reported variant. Then, we considered genes in which at least one variant is significantly associated with expression of that gene. Thus, for one GWAS variant, multiple genes in one tissue could satisfy these requirements, and we considered these pairs of variants and genes as potential colocalization loci. Tables S2 and S3 list the potential colocalization loci for FG and FP phenotypes, respectively. For any given variant, we used the CLPP to detect the most relevant tissue and a target gene. We selected the target gene and most relevant tissue as the gene and tissue, respectively, demonstrating the highest CLPP value.

Tables 1 and 2 indicate the results of eCAVIAR for FG and FP, respectively. These results show genetic variants that are causal in both the eQTL study and GWAS. We considered only variants reported to be significant with FG[21] and FP[22] phenotypes. We used a cutoff threshold of 0.01 (1%) to conclude that two causal variants are shared.

Many of the significant variants had CLPP values in a range where it is difficult to conclude whether the causal variants are shared. However, we detected a large number of loci for which the GWAS causal variants were clearly distinct from the causal variants in the eQTL data (Table 3). This included several genes that could be excluded in all tissues tested (e.g., *SEC22A* [MIM: 612442] at the rs11717195 FG locus, where there was non-colocalization).

More interesting examples could be found among genes that colocalized in one tissue yet could be excluded in many others. For example, *ADCY5* (MIM: 600293) was also at the rs11717195 FG locus. In pancreatic islet data, the GWAS variant itself colocalized with *ADCY5* eQTLs, whereas eQTLs for the same gene did not overlap the GWAS association signal in several GTEx tissues. This suggests that the phenotype influences the disease

Table 1. eCAVIAR Joint Analysis of FG and the GTEx Dataset

Chr	Position	rsID	Relevant Tissue[a]	Target Gene (MIM)
3	123,082,398	rs11717195	islet ($N = 118$)	*ADCY5* (600293)
7	15,064,309	rs2191349	islet ($N = 118$)	*DGKB* (604070)
7	44,235,668	rs4607517	colon sigmoid ($N = 124$)	*GCK* (138079)
			thyroid ($N = 278$)	
11	45,873,091	rs11605924	whole blood($N = 338$)	*MAPK8IP1* (604641)
11	47,336,320	rs7944584	nerve tibial ($N = 256$)	*CELF1* (601074)
			artery tibial ($N = 285$)	*MADD* (603584)
			islet ($N = 118$)	
			pituitary ($N = 87$)	
			artery tibial ($N = 285$)	*MDK* (162096)
			nerve tibial ($N = 256$)	*NR1H3* (602423)
			nerve tibial ($N = 256$)	*RAPSN* (601592)

The following abbreviation is used: Chr, chromosome.
[a]For each tissue, *N* indicates the number of individuals for whom we had access to summary statistics from GTEx[19] and van de Bunt et al.[25]

mechanism through a tissue-specific regulatory element that is active in islets yet inactive in other tissues.

For a majority of loci in which we identified a single causal variant in both the GWAS and eQTL study, our results implicate more than one target gene across the 45 tissues. eCAVIAR detected that three of five colocalized variants in the FG phenotype and all three variants in the FP phenotype had multiple target genes. Other eQTL studies support causal roles for *MADD* (MIM: 603584) at rs7944584 (FG) and rs10501320 (FP) in human pancreatic islets of Langerhans[25] and for *LARP6* (MIM: 611300) at rs1549318 (FP) in adipose tissue.[22] Assessing the potential candidacy of these different implicated genes will require additional sources of information, such as chromosome conformation capture (3C) experiments,[41] to demonstrate chromatin interactions between causal variants and gene promotors and/or in vitro function validation in relevant model systems. Even so, the current analysis points to many loci where no colocalizing variant can be identified. The main reason for this is probably found in the limited power of eCAVIAR at the current sample sizes for the majority of tissues, especially for those as pertinent to the phenotype as human islets (see Figure S10). Overcoming this hurdle and uncovering further mechanistic insights will require additional collection of samples.

Discussion

Integrating GWASs and eQTL studies provides insights into the underlying mechanism for genetic variants detected in GWASs. In this paper, we propose a quantity that can measure CLPP, the probability that the same variant is causal in both a GWAS and eQTL study, while accounting for the LD. Utilizing CLPP, we can identify target genes and relevant tissues. It is worth mentioning that we can use epigenomic data (e.g., NIH Roadmap Epigenomics[42]) to detect relevant tissues as an orthogonal analysis instead of using eCAVIAR. Moreover, eCAVIAR can detect loci where the causal variants are different between the two studies with high confidence. In our analysis, GWAS risk loci and eQTLs were different in most cases.

Because most GWAS loci are discovered to lie outside of coding regions, it is implicitly assumed that these implicated loci affect gene regulation. However, our results show a lower than expected number of variants colocalized between the GWAS and eQTL study. This points to a more complicated relationship between gene regulation and disease. It is likely that future studies will shed some light to explain this observation.

One conjecture is that the GWAS loci in fact do affect expression but are secondary signals in comparison to the stronger associations found in current eQTL studies. Because eQTL studies are including an increasing number of individuals, we will be able to prove or disprove this conjecture. Furthermore, the heterogeneity of tissues could render it hard to detect eQTLs specific to a disease-relevant cell type that composes only a fraction of the tissue. A second possibility is that GWAS variants affect other aspects of gene regulation, such as splicing or regulation at a level other than transcription regulation. Several studies have shown that alternative splicing could explain the causal mechanism of complex disease associations (e.g., a multiple-sclerosis-associated variant that leads to exon skipping in *SP140* [MIM: 608602][43]). Methods that identify variants associated with differences in relative expression of alternative transcript isoforms or exon-junction abundances are being applied to the latest version of GTEx data.[44,45] As we obtain more functional genomic information and are able to measure quantities such as

Table 2. eCAVIAR Joint Analysis of FP and the GTEx Dataset

Chr	Position	rsID	Relevant Tissue[a]	Target Gene (MIM)
11	47,293,799	rs10501320	pituitary ($N = 87$)	*ARHGAP1* (602732)
			esophagus muscularis ($N = 218$)	*C1QTNF4* (614911)
			artery tibial ($N = 285$)	*MADD* (603584)
			esophagus mucosa ($N = 241$)	
			islet ($N = 118$)	
			artery tibial ($N = 285$)	*MDK* (162096)
11	72,432,985	rs11603334	skin of sun-exposed lower leg ($N = 302$)	*ARAP1* (606646)
			pituitary ($N = 87$)	*PDE2A* (602658)
			islet ($N = 118$)	*STARD10*
15	71,109,147	rs1549318	adipose visceral omentum ($N = 185$)	*LARP6* (611300)
			cultured primary fibroblasts ($N = 272$)	
			ovary ($N = 85$)	

The following abbreviation is used: Chr, chromosome.
[a]For each tissue, *N* indicates the number of individuals for whom we had access to summary statistics from GTEx[19] and van de Bunt et al.[25]

protein abundance, we will be able to systematically catalog variants that affect regulation at levels other than transcription. A third possibility is that GWAS loci are eQTL loci only in certain conditions, such as development, where expression levels are not typically measured. Regardless, our study demonstrates strong evidence in support of the idea that most GWAS loci are not strong eQTL loci and that the mechanism by which GWAS loci affect gene regulation is more complicated than we expected.

Broadly, we have identified an analogy between colocalization and fine-mapping methods. Fine-mapping methods can be categorized into three main classes. One class relies only on the computed marginal statistics that are obtained from a GWAS or eQTL study. In this class of methods, the probability that a variant is causal depends on a variant's rank, which is obtained from the marginal statistics. Recently, Maller et. al[46] proposed a fine-mapping method that utilizes the Bayes factor. This method provides results similar to those of approaches that rank variants solely on the basis of their marginal statistics. The Maller et. al[46] method for fine-mapping is similar in nature to COLOC,[14] which is used for colocalization. The second class of methods is based on a conditional model that recomputes the marginal statistics of all variants by conditioning on variants selected as causal. The conditional method for fine-mapping and RTC[13] have some similarities in nature. The third class of methods includes CAVIAR,[18,27] CAVIARBF,[47] and FINEMAP,[29] which assume a presence of more than one causal variant in a region. These probabilistic-based methods use the MVN distribution and detect a set of variants that can capture all causal variants with a predefined probability. eCAVIAR is analogous in process to CAVIAR, CAVIARBF, and FINEMAP. However, eCAVIAR and CAVIAR-like methods try to solve different problems. CAVIAR-like methods (CAVIARBF and FINEMAP) are designed to perform fine-mapping. CAVIARBF is based on the CAVIAR statistical model that utilizes the Bayes factor to detect the causal set. FINEMAP is based on the CAVIAR statistical model that utilizes sampling techniques to speed up the computational process of detecting the causal set.

eCAVIAR is a probabilistic method that integrates GWAS and eQTL signals to detect biological mechanisms. eCAVIAR has several advantages over prior approaches. First, it can account for multiple causal variants in any given locus. Second, it leverages summary statistics without accessing the raw individual data. In addition, eCAVIAR can provide confidence levels for the colocalization of a GWAS risk variant. Utilizing the confidence level, we can categorize a variant into three categories: colocalizing variants, non-colocalizing variants, and variants whose ambiguity prevents detection of their colocalization status for the current data. eCAVIAR can be extended to utilize functional annotations to improve our results. The functional annotation can be used as a prior for a given causal status. Alternatively, we can adopt more sophisticated techniques similar to PAINTOR[40] and RiVIERA-beta,[48] which incorporate functional annotations to improve fine-mapping results. High-throughput technologies have made it possible to obtain multi-tissue eQTL studies. Leveraging multi-tissue eQTL studies such as GTEx and methods such as eCAVIAR will potentially advance discovery of new biological mechanisms for GWAS risk loci.

Appendix A: CAVIAR and eCAVIAR General Models where the GWAS and eQTL Study Have Different Numbers of Individuals

Standard GWAS Association Test

We assume that there are $N^{(p)}$ individuals in the GWAS for the phenotype of interest, and we collect the phenotypic

Table 3. Loci where the Causal Variants between eQTL Studies and GWASs Are Different

Phenotype	Chr	Position	rsID	GWAS p Value	eQTL p Value	No. of Genes	No. of Tissues
FG	2	27,741,237	rs780094	2.49×10^{-12}	2.95×10^{-55}	17	30
	2	169,763,148	rs560887	4.61×10^{-75}	1.36×10^{-14}	5	20
	3	123,065,778	rs11708067	8.72×10^{-9}	4.28×10^{-42}	5	34
	9	4,289,050	rs7034200	0.0001204	9.95×10^{-12}	8	7
	10	113,042,093	rs10885122	8.41×10^{-11}	7.73×10^{-11}	2	3
	11	61,571,478	rs174550	1.48×10^{-8}	1.03×10^{-125}	24	29
	11	92,708,710	rs10830963	1.26×10^{-68}	7.49×10^{-6}	7	6
FP	1	99,177,253	rs9727115	5.285×10^{-6}	7.04×10^{-16}	3	12
	10	114,758,349	rs7903146	3.48×10^{-18}	7.92×10^{-33}	7	26
	15	62,383,155	rs4502156	3.80×10^{-11}	8.48×10^{-14}	7	15
	17	2,262,703	rs4790333	2.15×10^{-8}	5.39×10^{-75}	21	33

The numbers of genes and tissues indicate the genes and tissues, respectively, that we applied to eCAVIAR for a GWAS risk variant. The complete lists of genes and tissues are provided in Tables S2 (FG) and S3 (FP). eCAVIAR utilizes the marginal statistics of all variants in a locus obtained from a GWAS and eQTL study. We obtain each locus by considering 50 variants upstream and downstream of the reported variant. Then, we consider genes where at least one of the variants in the locus is significantly associated with the expression of that gene. Thus, for one GWAS variant, multiple genes in one tissue can satisfy our condition. The eQTL p value indicates the most significant variant in eQTLs among all genes and all tissues. Abbreviations are as follows: Chr, chromosome; FG, fasting glucose; and FP, fasting proinsulin.

values of a quantitative trait for all individuals. Moreover, we collect the genotypes of all individuals for M variants. Let $Y^{(p)}$ be an $N^{(p)} \times 1$ vector of phenotypic values obtained from a GWAS. Let $G^{(p)}$ be an $N^{(p)} \times M$ matrix of genotypes where $G_i^{(p)}$ is an $N^{(p)} \times 1$ vector of the minor allele count for the i^{th} variant. We use $X_i^{(p)}$ to indicate the standardized vector of the minor allele count for the i^{th} variant, $G_i^{(p)}$, where $x_{ji}^{(p)}$ is the standardized genotype of the i^{th} variant for the j^{th} individual. We assume that both phenotypes and genotypes are standardized. We standardize a vector to make the mean and variance equal to 0 and 1, respectively. Thus, we have $1^T X_i^{(p)} = 0$ and $X_i^{(p)T} X_i^{(p)} = N^{(p)}$, which we can show as follows:

$$E\left(X_i^{(p)}\right) = 0 \rightarrow \frac{\sum_{j=1}^{N} x_{ji}^{(p)}}{N} = 0 \rightarrow \frac{1}{N} 1^T X_i^{(p)} = 0 \rightarrow 1^T X_i^{(p)} = 0,$$

$$\begin{aligned} \mathrm{Var}\left(X_i^{(p)}\right) = 1 &\rightarrow E\left(X_i^{(p)2}\right) - E\left(X_i^{(p)}\right)^2 = 1 \rightarrow E\left(X_i^{(p)2}\right) \\ &= 1 \rightarrow \frac{1}{N} X_i^{(p)T} X_i^{(p)} = 1 \rightarrow X_i^{(p)T} X_i^{(p)} = N^{(p)}. \end{aligned}$$

(Equation A1)

We assume the "additive" Fisher's polygenic model. In this model, each variant has a small effect toward the phenotype, where these effects are linear and additive. Thus, we have

$$Y^{(p)} = \mu^{(p)} 1 + \sum_{i=1}^{M} \beta_i^{(p)} X_i^{(p)} + \boldsymbol{e}^{(p)}, \qquad \text{(Equation A2)}$$

where $\mu^{(p)}$ is the population mean for the phenotype, $\beta_i^{(p)}$ is the effect of the i^{th} variant toward the phenotype, and $\boldsymbol{e}^{(p)}$ is the environmental and measurement noise that we assume follows a normal distribution, $\boldsymbol{e}^{(p)} \sim \mathcal{N}(0, \sigma_e^{(p)2} \boldsymbol{I})$, where $\sigma_e^{(p)2}$ is a covariance scalar. As mentioned in the Material and Methods, we test the significance of each variant one at a time. Moreover, we assume that the c^{th} variant is causal. Thus, we have the following model:

$$Y^{(p)} = \mu^{(p)} 1 + \beta_c^{(p)} X_c^{(p)} + \boldsymbol{e}^{(p)}. \qquad \text{(Equation A3)}$$

To ease our notations, we utilize the fact that both phenotypes and genotypes are standardized. Thus, the phenotypes follow a normal distribution with mean $\beta_c^{(p)} X_c^{(p)}$ and variance $\sigma_e^{(p)2} \boldsymbol{I}$, $Y^{(p)} \sim \mathcal{N}(\beta_c^{(p)} X_c^{(p)}, \sigma_e^{(p)2} \boldsymbol{I})$. To estimate the effect size, we utilize the maximum likelihood. The likelihood is computed as follows:

$$\begin{aligned} \boldsymbol{L}\left(Y^{(p)} \mid \mu^{(p)}, \beta_c, \sigma_e^{(p)}\right) = &\frac{1}{\sqrt{2\pi}\sigma_e^{(p)}} \exp\left(-\frac{1}{2\sigma_e^{(p)2}} \right. \\ &\times \left(Y^{(p)} - \mu^{(p)} - \beta_c^{(p)} X_c^{(p)}\right)^T \\ &\left. \times \left(Y^{(p)} - \mu^{(p)} - \beta_c^{(p)} X_c^{(p)}\right) \right). \end{aligned}$$

(Equation A4)

We compute the optimal effect size that maximizes the above likelihood by computing the likelihood derivative and setting it to 0. As a result, the optimal effect size is computed as follows:

$$\begin{aligned} \frac{\partial \boldsymbol{L}\left(Y^{(p)} \mid \mu^{(p)}, \beta_c, \sigma_e^{(p)}\right)}{\partial \beta_c} = 0 &\rightarrow \widehat{\beta}_c \\ &= \left(X_c^{(p)T} X_c^{(p)}\right)^{-1} X_c^{(p)} \left(Y^{(p)} - \mu^{(p)}\right) \rightarrow \widehat{\beta}_c \\ &\sim \mathcal{N}\left(\beta_c, \frac{\sigma_e^{(p)2}}{X_c^{(p)T} X_c^{(p)}}\right). \end{aligned}$$

(Equation A5)

We calculate the marginal statistics by dividing the estimated effect size by the SD of the estimated effect size. Thus, we have

$$S_c = \frac{\widehat{\beta}_c}{\widehat{\sigma}_e^{(p)}}\sqrt{X_c^{(p)T}X_c^{(p)}} \sim \mathcal{N}\left(\frac{\beta_c}{\sigma_e^{(p)}}\sqrt{N^{(p)}}, 1\right), \quad \text{(Equation A6)}$$

where $\lambda_c = (\beta_c/\sigma_e^{(p)})\sqrt{N^{(p)}}$ is the true effect size of the c^{th} variant, which is the causal variant. The normal distribution for the marginal statistics holds under asymptotic assumptions.

Indirect GWAS Association Test

We assume that there are two variants where the c^{th} variant is causal and the i^{th} variant is not causal. To estimate the effect size, we use the same testing model as in the previous section. Thus, we have

$$Y^{(p)} = \mu^{(p)}1 + \beta_i^{(p)}X_i^{(p)} + \boldsymbol{e}^{(p)}, \quad \text{(Equation A7)}$$

where we maximize the likelihood function to obtain the optimal effect sizes. The optimal effect size for the i^{th} variant is as follows:

$$\widehat{\beta}_i = \left(X_i^{(p)T}X_i^{(p)}\right)^{-1}X_i^{(p)}\left(Y^{(p)} - \mu^{(p)}\right) \rightarrow \widehat{\beta}_i \sim \mathcal{N}\left(\beta_i, \frac{\sigma_e^{(p)2}}{X_i^{(p)T}X_i^{(p)}}\right). \quad \text{(Equation A8)}$$

We compute the marginal statistics similarly to in the previous section:

$$S_i = \frac{\widehat{\beta}_i}{\widehat{\sigma}_e^{(p)}}\sqrt{X_i^{(p)T}X_i^{(p)}} \sim \mathcal{N}\left(\frac{\beta_i}{\sigma_e^{(p)}}\sqrt{N^{(p)}}, 1\right). \quad \text{(Equation A9)}$$

The variance of the marginal statistics is 1. Thus, the correlation and covariance of marginal statistics are equal. We compute the covariance of the marginal statistics between the causal variant and the non-causal variant. We compute this correlation as follows:

$$\begin{aligned}\mathrm{Cov}(S_c, S_i) &= \mathrm{Cov}\left(\frac{\widehat{\beta}_c}{\widehat{\sigma}_e^{(p)}}\sqrt{X_c^{(p)T}X_c^{(p)}}, \frac{\widehat{\beta}_i}{\widehat{\sigma}_e^{(p)}}\sqrt{X_i^{(p)T}X_i^{(p)}}\right)\\ &= \mathrm{Cov}\left(\frac{X_c^{(p)T}Y^{(p)}}{\widehat{\sigma}_e^{(p)}}\frac{1}{\sqrt{X_c^{(p)T}X_c^{(p)}}}, \frac{X_i^{(p)T}Y^{(p)}}{\widehat{\sigma}_e^{(p)}}\frac{1}{\sqrt{X_i^{(p)T}X_i^{(p)}}}\right)\\ &= \frac{1}{\widehat{\sigma}_e^{(p)2}}\frac{X_i^{(p)}}{\sqrt{X_i^{(p)T}X_i^{(p)}}}\mathrm{Cov}\left(Y^{(p)}, Y^{(p)}\right)\frac{X_c^{(p)}}{\sqrt{X_c^{(p)T}X_c^{(p)}}}.\end{aligned} \quad \text{(Equation A10)}$$

Using the Slutsky's theorem and the fact that the number of individuals in a study is large enough, we assume that $\widehat{\sigma}_e^{(p)2}$ approaches $\mathrm{Var}(Y^{(p)})$. Thus, asymptotically we have

$$\mathrm{Cov}(S_c, S_i) = \frac{X_i^{(p)T}X_c^{(p)}}{\sqrt{X_i^{(p)T}X_i^{(p)}}\sqrt{X_c^{(p)T}X_c^{(p)}}} = r_{ci}. \quad \text{(Equation A11)}$$

This indicates that the correlation between the marginal statistics of two variants is equal to their genotype correlation. This result is known from our previous studies.[18,27,49]

Standard eQTL Association Test

We assume that in the eQTL study, we collect the expression of multiple genes for $N^{(e)}$ individuals. Let superscript (p) and (e) indicate variables related to the GWAS and eQTL study, respectively. Let $Y^{(e)}$ indicate the expression level of all $N^{(e)}$ individuals for one gene. We consider one gene to ease the presentation of the method. As illustrated in the previous GWAS section, we use the "additive" Fisher's polygenic model:

$$Y^{(e)} = \mu^{(e)}1 + \sum_{i=1}^{M}\beta_i^{(e)}X_i^{(e)} + \boldsymbol{e}^{(e)}, \quad \text{(Equation A12)}$$

where $\mu^{(e)}$ is the population mean for the expression of the gene of interest, $\beta_i^{(e)}$ is the effect of the i^{th} variant toward the gene expression, and $\boldsymbol{e}^{(e)}$ is the environmental and measurement noise that follows a normal distribution, $\boldsymbol{e}^{(e)} \sim \mathcal{N}(0, \sigma_e^{(e)2}\boldsymbol{I})$, where $\sigma_e^{(e)2}$ is a covariance scalar. As mentioned in the Material and Methods, we test the significant of each variant one at a time. Similarly, we assume that the c^{th} variant is causal. Thus, we have the following model:

$$Y^{(e)} = \mu^{(e)}1 + \beta_c^{(e)}X_c^{(e)} + \boldsymbol{e}^{(e)}. \quad \text{(Equation A13)}$$

The optimal estimated effect size is similar between the eQTL study and GWAS.

CAVIAR Model for GWASs and eQTL Studies

We know that the covariance between the estimated effect size of two variants is equal to their genotype correlation. Furthermore, the mean of the marginal statistics of the non-causal variants is equal to the mean of the marginal statistics of the causal variants scaled by the genotype correlation. Thus, we have

$$\left(S^{(p)} \mid \Lambda^{(p)}\right) \sim \mathcal{N}\left(\Lambda^{(p)}\Sigma^{(p)}, \Sigma^{(p)}\right), \quad \text{(Equation A14)}$$

where the $\Sigma^{(p)}$ matrix is the pairwise genotype correlations obtained from a GWAS. For the eQTL study, we obtain a similar equation for the joint marginal statistics:

$$\left(S^{(e)} \mid \Lambda^{(e)}\right) \sim \mathcal{N}\left(\Lambda^{(e)}\Sigma^{(e)}, \Sigma^{(e)}\right), \quad \text{(Equation A15)}$$

where the $\Sigma^{(e)}$ matrix is the pairwise genotype correlations obtained from the eQTL study. We consider $\Lambda^{(p)}$ and $\Lambda^{(e)}$ to be the true effect-size vectors for the GWAS and eQTL study, respectively. True effect sizes are nonzero for causal variants and zero for the non-causal variants. Moreover, we consider $\Lambda^{(p)}\Sigma^{(p)}$ and $\Lambda^{(e)}\Sigma^{(e)}$ to be the LD-induced effect sizes for the GWAS and eQTL study, respectively.

We introduce a MVN prior over the true effect-size vectors. The true effect sizes for variants are independent and, for causal variants, non-zero. Thus, we have the following prior:

$$\left(\Lambda^{(\mathrm{p})} \mid C^{(\mathrm{p})}\right) \sim \mathcal{N}\left(0, \sigma^{(\mathrm{p})2}\Sigma_c^{(\mathrm{p})}\right),$$

$$\left(\Lambda^{(\mathrm{e})} \mid C^{(\mathrm{e})}\right) \sim \mathcal{N}\left(0, \sigma^{(\mathrm{e})2}\Sigma_c^{(\mathrm{e})}\right), \qquad \text{(Equation A16)}$$

where $\Sigma_c^{(p)}$ is a diagonal matrix and $\sigma^{(p)2}$ is set to 5.2,[18,27] which indicates the variance of our prior over the GWAS effect sizes. The diagonal elements are set to 1 or 0. For variants that are selected as causal, we set the corresponding diagonal elements to 1; otherwise, we set them to 0.

Utilizing the conjugate prior, we can combine Equations A14 and A16 to obtain the joint distribution of the marginal statistics given the vector of causal status. These distributions are

$$\left(S^{(\mathrm{p})} \mid C^{(\mathrm{p})}\right) \sim \mathcal{N}\left(0, \Sigma^{(\mathrm{p})} + \sigma^{(\mathrm{p})2}\Sigma^{(\mathrm{p})}\Sigma_c^{(\mathrm{p})}\Sigma^{(\mathrm{p})}\right),$$

$$\left(S^{(\mathrm{e})} \mid C^{(\mathrm{e})}\right) \sim \mathcal{N}\left(0, \Sigma^{(\mathrm{e})} + \sigma^{(\mathrm{e})2}\Sigma^{(\mathrm{e})}\Sigma_c^{(\mathrm{e})}\Sigma^{(\mathrm{e})}\right). \quad \text{(Equation A17)}$$

To show the correctness of the above equations, we utilize the law of total expectation and law of total variance. Given two random variables A and B, the law of total expectation is as follows:

$$\mathrm{E}[A] = \mathrm{E}_B\left[\mathrm{E}_{A|B}[A \mid B]\right]. \qquad \text{(Equation A18)}$$

If we let $A = \left(S^{(\mathrm{p})} \mid C^{(\mathrm{p})}\right)$ and $B = \Lambda^{(\mathrm{p})}$, we can compute the mean of the marginal statistics given the causal status as follows:

$$\begin{aligned}\mathrm{E}\left[S^{(\mathrm{p})} \mid C^{(\mathrm{p})}\right] &= \mathrm{E}_{\Lambda^{(\mathrm{p})}}\left[E_{\left(S^{(\mathrm{p})} \mid \Lambda^{(\mathrm{p})}\right)}\left(S^{(\mathrm{p})} \mid \Lambda^{(\mathrm{p})}\right)\right] = \mathrm{E}_{\Lambda^{(\mathrm{p})}}\left[\Sigma^{(\mathrm{p})}\Lambda^{(\mathrm{p})}\right] \\ &= \Sigma^{(\mathrm{p})}\mathrm{E}_{\Lambda^{(\mathrm{p})}}\left[\Lambda^{(\mathrm{p})}\right] = 0.\end{aligned} \qquad \text{(Equation A19)}$$

To compute the variance of the joint distribution of the marginal statistics given the causal status, we use the law of total variance:

$$\mathrm{Var}[A] = \mathrm{E}_B[\mathrm{Var}[A \mid B]] + \mathrm{Var}_B[\mathrm{E}_B[A \mid B]]. \qquad \text{(Equation A20)}$$

Thus, we compute the variance of joint distribution as follows:

$$\begin{aligned}\mathrm{Var}\left[S^{(\mathrm{p})} \mid C^{(\mathrm{p})}\right] &= \mathrm{E}_{\Lambda^{(\mathrm{p})}}\left[\mathrm{Var}_{S^{(\mathrm{p})} \mid \Lambda^{(\mathrm{p})}}\right] + \mathrm{Var}_{\Lambda^{(\mathrm{p})}}\left[\mathrm{E}\left[S^{(\mathrm{p})} \mid \Lambda^{(\mathrm{p})}\right]\right] \\ &= \mathrm{E}_{\Lambda^{(\mathrm{p})}}\left[\Sigma^{(\mathrm{p})}\right] + \mathrm{Var}_{\Lambda^{(\mathrm{p})}}\left[\Sigma^{(\mathrm{p})}\Lambda^{(\mathrm{p})}\right] \\ &= \Sigma^{(\mathrm{p})} + \sigma^{(\mathrm{p})2}\Sigma^{(\mathrm{p})}\Sigma_c^{(\mathrm{p})}\Sigma^{(\mathrm{p})}.\end{aligned} \qquad \text{(Equation A21)}$$

Supplemental Data

Supplemental Data include ten figures and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.10.003.

Acknowledgments

F.H., J.W.J.J., M.B., and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, 1320589, and 1331176 and NIH grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-ES021801, R01-MH101782, and R01-ES022282. E.E. is supported in part by NIH Big Data to Knowledge (BD2K) award U54EB020403. M.v.d.B. is supported by a Novo Nordisk postdoctoral fellowship run in partnership with the University Of Oxford. A.V.S. and X.L. are supported by contract HHSN268201000029C (Broad Institute). S.S. is supported in part by NIH grant R00-GM 111744-03. We acknowledge support from the National Institute of Neurological Disorders and Stroke Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691).

Received: June 24, 2016
Accepted: October 3, 2016
Published: November 17, 2016

Web Resources

eCAVIAR, http://genetics.cs.ucla.edu/caviar/
GTEx Portal (release v.6), http://www.gtexportal.org

References

1. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
2. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.
3. Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al.; LifeLines Cohort Study (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science *340*, 1467–1471.
4. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., et al.; Multicenter Genetic Studies of Schizophrenia Consortium; Psychosis Endophenotypes International Consortium; Wellcome Trust Case Control Consortium 2 (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet. *45*, 1150–1159.
5. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.
6. Huang, Y.-T., Liang, L., Moffatt, M.F., Cookson, W.O., and Lin, X. (2015). igwas: Integrative genome-wide association studies of genetic and genomic data for disease susceptibility using mediation analysis. Genet. Epidemiol. *39*, 347–356.
7. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., and Im, H.K.; GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
8. Pritchard, J.K., and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. *69*, 1–14.

9. Abecasis, G.R., Noguchi, E., Heinzmann, A., Traherne, J.A., Bhattacharyya, S., Leaves, N.I., Anderson, G.G., Zhang, Y., Lench, N.J., Carey, A., et al. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. Am. J. Hum. Genet. *68*, 191–197.
10. Dunning, A.M., Durocher, F., Healey, C.S., Teare, M.D., McBride, S.E., Carlomagno, F., Xu, C.F., Dawson, E., Rhodes, S., Ueda, S., et al. (2000). The extent of linkage disequilibrium in four populations with distinct demographic histories. Am. J. Hum. Genet. *67*, 1544–1554.
11. Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. *22*, 139–144.
12. He, X., Fuller, C.K., Song, Y., Meng, Q., Zhang, B., Yang, X., and Li, H. (2013). Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. Am. J. Hum. Genet. *92*, 667–680.
13. Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. *6*, e1000895.
14. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.
15. Wallace, C., Rotival, M., Cooper, J.D., Rice, C.M., Yang, J.H., McNeill, M., Smyth, D.J., Niblett, D., Cambien, F., Tiret, L., et al.; Cardiogenics Consortium (2012). Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. Hum. Mol. Genet. *21*, 2815–2824.
16. Plagnol, V., Smyth, D.J., Todd, J.A., and Clayton, D.G. (2009). Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. Biostatistics *10*, 327–334.
17. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
18. 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.
19. Ardlie, K.G., Deluca, D.S., Segrè, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., et al.; GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.
20. Sul, J.H., Raj, T., de Jong, S., de Bakker, P.I., Raychaudhuri, S., Ophoff, R.A., Stranger, B.E., Eskin, E., and Han, B. (2015). Accurate and fast multiple-testing correction in eQTL studies. Am. J. Hum. Genet. *96*, 857–868.
21. Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L., et al.; DIAGRAM Consortium; GIANT Consortium; Global BPgen Consortium; Anders Hamsten on behalf of Procardis Consortium; MAGIC investigators (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat. Genet. *42*, 105–116.
22. Strawbridge, R.J., Dupuis, J., Prokopenko, I., Barker, A., Ahlqvist, E., Rybin, D., Petrie, J.R., Travers, M.E., Bouatia-Naji, N., Dimas, A.S., et al.; DIAGRAM Consortium; GIANT Consortium; MuTHER Consortium; CARDIoGRAM Consortium; C4D Consortium (2011). Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. Diabetes *60*, 2624–2634.
23. Saxena, R., Hivert, M.-F., Langenberg, C., Tanaka, T., Pankow, J.S., Vollenweider, P., Lyssenko, V., Bouatia-Naji, N., Dupuis, J., Jackson, A.U., et al.; GIANT consortium; MAGIC investigators (2010). Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nat. Genet. *42*, 142–148.
24. Soranzo, N., Sanna, S., Wheeler, E., Gieger, C., Radke, D., Dupuis, J., Bouatia-Naji, N., Langenberg, C., Prokopenko, I., Stolerman, E., et al.; WTCCC (2010). Common variants at 10 genomic loci influence hemoglobin A_1(C) levels via glycemic and nonglycemic pathways. Diabetes *59*, 3229–3239.
25. van de Bunt, M., Manning Fox, J.E., Dai, X., Barrett, A., Grey, C., Li, L., Bennett, A.J., Johnson, P.R., Rajotte, R.V., Gaulton, K.J., et al. (2015). Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. PLoS Genet. *11*, e1005694.
26. Han, B., Kang, H.M., and Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. PLoS Genet. *5*, e1000456.
27. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B., and Eskin, E. (2015). Identification of causal genes for complex traits. Bioinformatics *31*, i206–i213.
28. Kostem, E., Lozano, J.A., and Eskin, E. (2011). Increasing power of genome-wide association studies by collecting additional single-nucleotide polymorphisms. Genetics *188*, 449–460.
29. 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.
30. Eskin, E. (2008). Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. Genome Res. *18*, 653–660.
31. Darnell, G., Duong, D., Han, B., and Eskin, E. (2012). Incorporating prior information into association studies. Bioinformatics *28*, i147–i153.
32. Sul, J.H., Han, B., He, D., and Eskin, E. (2011). An optimal weighted aggregated association test for identification of rare variants involved in common diseases. Genetics *188*, 181–188.
33. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A.; 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
34. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
35. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.

36. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.
37. Hormozdiari, F., Kang, E.Y., Bilow, M., Ben-David, E., Vulpe, C., McLachlan, S., Lusis, A.J., Han, B., and Eskin, E. (2016). Imputing phenotypes for genome-wide association studies. Am. J. Hum. Genet. *99*, 89–103.
38. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E., and Halperin, E. (2010). Leveraging genetic variability across populations for the identification of causal variants. Am. J. Hum. Genet. *86*, 23–33.
39. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
40. 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.
41. Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D.R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat. Genet. *46*, 205–212.
42. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH roadmap epigenomics mapping consortium. Nat. Biotechnol. *28*, 1045–1048.
43. Matesanz, F., Potenciano, V., Fedetz, M., Ramos-Mozo, P., Abad-Grau, Mdel.M., Karaky, M., Barrionuevo, C., Izquierdo, G., Ruiz-Peña, J.L., García-Sánchez, M.I., et al. (2015). A functional variant that affects exon-skipping and protein expression of SP140 as genetic mechanism predisposing to multiple sclerosis. Hum. Mol. Genet. *24*, 5619–5627.
44. Monlong, J., Calvo, M., Ferreira, P.G., and Guigó, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. Nat. Commun. *5*, 4698.
45. Ongen, H., and Dermitzakis, E.T. (2015). Alternative splicing QTLs in european and african populations. Am. J. Hum. Genet. *97*, 567–575.
46. 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.
47. 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.
48. Li, Y., and Kellis, M. (2016). Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. Nucleic Acids Res. *44*, e144.
49. Joo, J.W.J., Hormozdiari, F., Han, B., and Eskin, E. (2016). Multiple testing correction in linear mixed models. Genome Biol. *17*, 62.