Widespread Allelic Heterogeneity in Complex Traits

Farhad Hormozdiari,[1,2] Anthony Zhu,[1] Gleb Kichaev,[3] Chelsea J.-T. Ju,[1] Ayellet V. Segrè,[4] Jong Wha J. Joo,[1,5] Hyejung Won,[6] Sriram Sankararaman,[1,7] Bogdan Pasaniuc,[7,8] Sagiv Shifman,[9,10,*] and Eleazar Eskin[1,7,10,*]

Recent successes in genome-wide association studies (GWASs) make it possible to address important questions about the genetic architecture of complex traits, such as allele frequency and effect size. One lesser-known aspect of complex traits is the extent of allelic heterogeneity (AH) arising from multiple causal variants at a locus. We developed a computational method to infer the probability of AH and applied it to three GWASs and four expression quantitative trait loci (eQTL) datasets. We identified a total of 4,152 loci with strong evidence of AH. The proportion of all loci with identified AH is 4%–23% in eQTLs, 35% in GWASs of high-density lipoprotein (HDL), and 23% in GWASs of schizophrenia. For eQTLs, we observed a strong correlation between sample size and the proportion of loci with AH ($R^2 = 0.85$, $p = 2.2 \times 10^{-16}$), indicating that statistical power prevents identification of AH in other loci. Understanding the extent of AH may guide the development of new methods for fine mapping and association mapping of complex traits.

Introduction

Genome-wide association studies (GWASs) have successfully identified many loci associated with various diseases and traits.[1–4] Unfortunately, interpreting the detected associated genes is challenging due to two facts. First, most of the associated variants fall in non-coding regions of the genome.[5–10] Second, for only a handful of GWAS loci was a causal sequence variant detected that underlies the trait or disease susceptibility. Therefore, it is challenging to identify the relevant genes, which is the first step to understanding the biological mechanisms of the disease.

Detecting the causal variants is complicated by the fact that most significant associated variants may not be causal, but instead are in linkage disequilibrium (LD) with unknown functional variants. In general, sequence variants, with respect to a trait, can be grouped into three categories. The first category contains causal variants that have a biological effect on the trait and are responsible for the association signal. The second category contains variants that are statistically associated with the trait due to high LD with the causal variants. The third category contains variants that are not statistically associated with the trait and are not causal. Fine-mapping methods aim to distinguish between the two first categories (causal variants versus correlated variants). One way to link the causal variant with a particular gene is by colocalization methods that determine whether a single variant is responsible for both variation in the trait and variation in expression of a gene at the same locus (expression quantitative trait loci, eQTLs).

Fine-mapping and colocalization methods are designed to identify the causal variant and the associated gene at a locus, but in many cases, they assume a single causal variant. In the presence of multiple causal variants, those fine-mapping and colocalization methods[11–13] will have a lower accuracy to detect the true causal variants and genes. Thus, a fundamental question is how many different causal variants are present in a locus.

The presence of multiple causal variants at the same locus that influence a particular disease or trait is known as allelic heterogeneity (AH). AH is very common for Mendelian traits. Clearly, many different mutations in the same gene may cause loss or gain of function leading to a specific Mendelian disease. For example, approximately 100 distinct mutations are known to exist at the cystic fibrosis locus,[14] and even more are present at loci causing inherited hemoglobinopathies.[15] In contrast to Mendelian traits, the extent of AH at loci contributing to common, complex disease is almost unknown. Hermani et al.[16] have shown in their study the existence of multiple reproducible epistatic effects that influence gene expression. However, a recent study has shown that most of these epistatic effects can be explained by a third variant in that locus.[17] Thus, it is of utmost importance to detect loci that harbor AH in order to avoid considering them as epistatic interactions.

Identifying the number of causal variants in complex traits is difficult because of extensive LD and small effect sizes. The standard approach to identify AH is to use conditional analysis. In conditional analysis, the independent association of multiple SNPs is tested after conditioning

[1]Department of Computer Science, University of California, Los Angeles, CA 90095, USA; [2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [3]Bioinformatics IDP, University of California, Los Angeles, CA 90095, USA; [4]Cancer Program, The Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA; [5]Department of Computer Science Engineering, Dongguk University-Seoul, 04620 Seoul, South Korea; [6]Neurogenetics Program, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA; [7]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA; [8]Department of Pathology and Laboratory Medicine, University of California, Los Angeles, CA 90095, USA; [9]Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 91904, Israel
[10]These authors contributed equally to this work
*Correspondence: sagiv@vms.huji.ac.il (S.S.), eeskin@cs.ucla.edu (E.E.)
http://dx.doi.org/10.1016/j.ajhg.2017.04.005.

on other SNPs, which are more significant. The conditional analysis lacks power because it requires multiple variants to have a strong effect and to be independently significant. When several associated variants are highly correlated, the conditional analysis will not detect multiple independent associations, but we cannot rule out the existence of AH. Thus, the extent of AH in complex traits is unknown. AH could substantially influence our ability to explain the missing heritability and to identify causal genes.

We developed and applied a method to quantify the number of independent causal variants at a locus responsible for the observed association signals in GWASs. Our method is incorporated into the CAVIAR (causal variants identification in associated regions) software.[18] The method is based on the principle of jointly analyzing association signals (i.e., summary level Z-scores) and LD structure in order to estimate the number of causal variants. Our method computes the probability of having multiple independent causal variants by summing the probability of all possible sets of SNPs for being causal. We compared results from our method to results produced using the standard conditional method (CM),[19] which tests for independent association of a variant after conditioning on its significantly associated neighbors. Using simulated datasets, we illustrate that CAVIAR tends to outperform CM. We observed a very low false positive rate for CAVIAR to detect loci with AH even when the true causal variant is not included in our dataset. We applied CAVIAR to both eQTL and GWAS datasets. Our results indicate that in the Genotype-Tissue Expression (GTEx) dataset,[20] 4%–23% of eGenes harbor AH. We observed a high correlation between the number of loci with AH and sample size. In addition, we replicated a significant fraction of the loci with AH in three other existing eQTL studies. We also applied CAVIAR to three GWAS datasets: schizophrenia (SCZ),[3] high-density lipoprotein (HDL),[21] and major depression disorder (MDD),[22] where we observed 23%, 35%, and 50% of loci (respectively) with strong evidence for AH.

Subjects and Methods

Joint Distribution of Observed Statistics in Standard GWASs

In this section, we provide a brief description of statistical tests that are performed in GWASs and the joint distribution of computed marginal statistics. These statistics are used as an input to CAVIAR. We consider GWASs on a quantitative trait for n individuals. Let Y be a vector of $(n \times 1)$ where y_i indicates phenotypic value for the i^{th} individual. Moreover, we genotype all the individuals for all the m variants. Let $G \in \{0,1,2\}^{\{n \times m\}}$ be a matrix of genotypes for all the individuals where g_{ij} is the minor allele count for the i^{th} individual at the j^{th} variant. We standardize the minor allele count for each variant to have mean of 0 and variance of 1. We use X_j to indicate the standardized minor allele count for all the individuals at the j^{th} variant. We have $1^T X_i = 0$ and $X_i^T X_i = n$ as the genotypes are standardized.

We assume that phenotypic values follow the linear model and the c variant is the causal variant. Thus, we have:

$$Y = \mu 1 + \beta_c X_c + \boldsymbol{e}$$

where μ is the phenotypic mean value, β_c is the effect size of the variant X_c, $\mathbf{1}$ is a $(n \times 1)$ vector of ones, and $\boldsymbol{e}$ is a $(n \times 1)$ vector to model the environment contribution and error in the measurement. In this model, we assume the error is i.i.d. and follows a normal distribution with mean of zero and variance of $\sigma_e^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is a $(n \times n)$ matrix of identity and σ_e is the variance scalar of $\boldsymbol{e}$. From the model explained above, we have $Y \sim N(\mu 1 + \beta_c X_c, \sigma_e^2 \boldsymbol{I})$. We use the maximum likelihood to compute the estimated β_c and σ_e which are denoted by $\widehat{\beta}_c$ and $\widehat{\sigma}_e$. Thus, we have:

$$\mathscr{L}(Y \mid \mu, \beta_c, \sigma_e^2) = \mid 2\pi\sigma_e^2 \boldsymbol{I} \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_e^2}(Y - \mu 1 - \beta_c X_c)^T \times (Y - \mu 1 - \beta_c X_c) \right),$$

$$\widehat{\mu} = \frac{1}{n} 1^T Y, \widehat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma_e^2}{n}\right),$$

$$\widehat{\beta}_c = \frac{X_c^T \boldsymbol{y}}{n}, \sqrt{n}\,\frac{\widehat{\beta}_c}{\sigma_e} \sim \mathcal{N}\left(\frac{\beta_c}{\sigma_e}\sqrt{n}, 1\right). \quad \text{(Equation 1)}$$

The association statistic for SNP c, denoted by S_c, follows a non-central t-distribution, which is the ratio of a normally distributed random variable to the square root of an independent chi-square distributed random variable. As shown in previous works, if the number of individuals are large enough,[18] we can assume the marginal statistics follows a normal distribution with mean equal to $\lambda_c = \frac{\widehat{\beta}}{\widehat{\sigma}}\sqrt{n}$ and variance of 1:

$$S_c \sim t_{\lambda_c, n} \approx \mathcal{N}(\lambda_c, 1). \quad \text{(Equation 2)}$$

We assume that variant i is correlated with the causal variant c and the correlation between the two variants is r where $r = \frac{1}{n} X_i^T X_c$. Then, the marginal statistic estimated at variant i is equal to the marginal statistics for the causal variant that is scaled by r. We compute the covariance of the marginal statistics between two variants where the LD between the two variants is r.

$$\mathrm{Cov}\left(\sqrt{n}\,\frac{\widehat{\beta}_i}{\sigma}, \sqrt{n}\,\frac{\widehat{\beta}_c}{\sigma}\right) = \frac{1}{n\sigma^2} X_i^T \mathrm{Var}(Y) X_c = r.$$

We compute the joint distribution of the marginal statistics for two variants i and j as follows:

$$\begin{bmatrix} S_i \\ S_j \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \lambda_i \\ \lambda_j \end{bmatrix}, \begin{bmatrix} 1 & r_{ij} \\ r_{ij} & 1 \end{bmatrix} \right), \quad \text{(Equation 3)}$$

where $\lambda_j = r\lambda_i$ if the variant i is causal and the variant j is not causal or $\lambda_i = r\lambda_j$ if the variant j is causal and the variant i is not causal.

Computing the Likelihood of Causal Configuration

We can extend the joint distribution of marginal statistics for two variants to a more general case. Assume we have m variants and the pairwise correlation between each variant is denoted by Σ. Let $S = [S_1, S_2, S_3, \cdots S_m]^T$ be the vector of marginal statistics

obtained for each variant. The joint distribution of the marginal statistics for all the m variants is computed as

$$(S \mid \Lambda) \sim \mathcal{N}(\Sigma\Lambda, \Sigma), \qquad \text{(Equation 4)}$$

where Λ is $(m \times 1)$ is a vector of normalized effect sizes and λ_i is the normalized effect size of the i^{th} variant. We introduce a parameter C that indicates the causal status of each variant. Each variant can have two possible causal status. We have $c_i = 1$ if the i^{th} variant is causal and $c_i = 0$ if the i^{th} variant is not a causal variant. We define a prior probability on the vector of effect size Λ for a given causal status using a multivariate normal distribution,

$$(\Lambda \mid C) \sim \mathcal{N}(0, \Sigma_c), \qquad \text{(Equation 5)}$$

where Σ_c is a $(m \times m)$ matrix. The off diagonal elements of Σ_c are set to zero. The diagonal elements are set to σ or zero. We set the i^{th} diagonal element to σ if the i^{th} variant is causal and we set the i^{th} diagonal element to zero if the i^{th} variant is not causal. Thus, the joint distribution follows a multivariate normal distribution,

$$(S \mid C) \sim \mathcal{N}(0, \Sigma + \Sigma\Sigma_c\Sigma)\,. \qquad \text{(Equation 6)}$$

Computing the Number of Independent Causal Variants in a Locus

In this section, we provide the formula to compute the probability of having i causal variants in a locus. We compute the probability of having i causal variants in a locus by summing over all the possible causal configurations where only i variants are causal. Let N_c indicates the number of causal variants in a locus. We have,

$$\Pr(N_c = i \mid S) = \frac{\sum_{|\boldsymbol{c}|=i} P(S \mid C)P(C)}{\sum_{C \in \mathcal{C}} P(S \mid C)P(C)} \qquad \text{(Equation 7)}$$

where $P(C)$ is the prior on the causal configuration C, $\mathcal{C}$ is the set of all possible causal configurations, including the configuration of all the variants that are not causal, and $|C|$ indicates the number of causal variants in the causal configuration C. The numerator in the above equation considers all possible causal configurations that have i causal variants, and the denominator is a normalization factor to ensure that the probability definition holds.

In this paper, we use a simple prior for a causal status. We assume that the probability of a variant to be causal is independent from other variants and that the probability of a variant to be causal is γ. Thus, we compute the prior probability as $\mathrm{P}(C^*) = \prod \gamma^{|c_i|}(1-\gamma)^{1-|c_i|}$. We utilize different values for γ as shown in Figure S2. In our experiment, we set γ to 0.001.[23–25] It is worth mentioning that, although we use a simple prior for our model, we can incorporate external information such as functional data or knowledge from previous studies. As a result, we can have variant-specific prior where γ_i indicates the prior probability for the i^{th} variant to be causal. Thus, we can extend the prior probability to a more general case, $\mathrm{P}(C^* \mid \Gamma = [\gamma_1, \gamma_2, \cdots \gamma_\ell]) = \prod \gamma_i^{|c_i|}(1-\gamma_i)^{1-|c_i|}$.

Reducing the Computational Complexity for Computing the Likelihood of a Causal Status

Unfortunately, the time complexity to compute the likelihood of a causal status using Equation 6 is $O(m^3)$. In this section, we provide a sped-up process that reduces the time complexity to $O(m^2k)$, where k is the number of causal variants (k is the number of variants that their causal status is 1). The number of causal variants is smaller than the total number of variants in a locus ($k \ll m$). Thus, we can speed up the likelihood computation by a factor of (m/k).

According to Equation 6, to compute the likelihood of a causal status, we must compute the quantity

$$(2\pi)^{-\frac{m}{2}} \det \mid \Sigma + \Sigma\Sigma_c\Sigma \mid^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} S^T (\Sigma + \Sigma\Sigma_c\Sigma)^{-1} S \right\} \qquad \text{(Equation 8)}$$

where $\det|.|$ denotes the determinant of a matrix. First, we speed up the exponential part. Thus, we have:

$$S^T(\Sigma + \Sigma\Sigma_c\Sigma)^{-1}S = S^T\Sigma^{-1}(I_{m\times m} + \Sigma_c\Sigma)^{-1}S \qquad \text{(Equation 9)}$$

where $S^T\Sigma^{-1}$ is independent from the causal status and can be computed once and used many times. As a result, the most time-consuming part is to compute $(I_{m\times m} + \Sigma_c\Sigma)^{-1}$. We set elements of $U_{m\times k}$ and $V_{k\times m}$ such that $\Sigma_c\Sigma = UV$. Let α_i indicate the index of i^{th} causal variant. We set elements of V as follows: $V(i,j) = r_{\alpha_i, j}$. Let $U(\alpha_i, i) = \sigma$, and we set the rest of elements of U to zero. Thus, we have:

$$(I_{m\times m} + \Sigma_c\Sigma)^{-1} = (I_{m\times m} + UV)^{-1} \qquad \text{(Equation 10)}$$

We use the Woodbury matrix identity formula to compute the inverse. We have

$$(A + UEV)^{-1} = A^{-1} - A^{-1}U\left(E^{-1} + VA^{-1}U\right)^{-1}VA^{-1}. \qquad \text{(Equation 11)}$$

We set A to $I_{m\times m}$ and E to $I_{k\times k}$, so the left side of the Woodbury matrix identity formula converts to $(I_{m\times m} + \Sigma_c\Sigma)^{-1}$. From the right side of the Woodbury matrix identity formula, we have:

$$(I_{m\times m} + \Sigma_c\Sigma)^{-1} = (I_{m\times m} + UV)^{-1} \qquad \text{(Equation 12)}$$

$$= I^{-1}_{m\times m} - I^{-1}_{m\times m}U\left(I^{-1}_{k\times k} + VI^{-1}_{m\times m}U\right)^{-1}VI_{m\times m} \qquad \text{(Equation 13)}$$

$$= I_{m\times m} - U(I_{k\times k} + VU)^{-1}V \qquad \text{(Equation 14)}$$

where $(I_{k\times k} + VU)^{-1}$ requires inverting a $(k \times k)$ matrix that is much faster than inverting a $(m \times m)$ matrix.

Similarly, the naive method to compute $\det|\Sigma + \Sigma\Sigma_c\Sigma|$ requires $O(m^3)$. We utilize the Sylvester's determinant identity that is as follows:

$$\det \mid I_{m\times m} + UV \mid = \det \mid I_{k\times k} + VU \mid . \qquad \text{(Equation 15)}$$

Thus, instead of computing the determinant of a $(m \times m)$ matrix, we can compute the same value by computing the determinant of a $(k \times k)$ matrix. We have

$$\det|\Sigma + \Sigma\Sigma_c\Sigma| = \det|\Sigma| \quad \det|I_{m\times m} + \Sigma_c\Sigma| \qquad \text{(Equation 16)}$$

$$= \det|\Sigma| \quad \det|I_{m\times m} + UV| \qquad \text{(Equation 17)}$$

$$= \det|\Sigma| \quad \det|I_{k\times k} + VU| . \qquad \text{(Equation 18)}$$

In the above equation, we can compute $\det|\Sigma|$ once and use it for different causal statuses. In addition, the computational complexity of $\det|I_{k\times k} + VU|$ is $O(k^3)$. Thus, the time complexity to compute $\det|\Sigma + \Sigma\Sigma_c\Sigma|$ is $O(k^3)$.

Conditional Method

A standard method to detect allelic heterogeneity (AH) is the conditional method (CM). CM is an interactive process. In each iteration of CM, we identify the SNP with the most significant association statistics. Then, conditioning on that SNP and all SNPs that are selected in previous steps, we re-compute the marginal statistics of all remaining variants in the locus. After the first iteration of CM, we consider a locus to have AH when the re-computed marginal statistics for at least one of the variants is more significant than a predefined threshold. Similarly, we consider a locus to not have AH when the re-computed marginal statistics of all variants is not significant. The predefined threshold is referred to as the stopping threshold for CM. We set the stopping threshold to be $(0.05/m)$ (stopping threshold is based on p value). Thus, we perform CM until there is no significant variant in a locus. This standard method can be applied to either summary statistics or individual-level data. GCTA-COJO[19] performs conditional analysis while utilizing the summary statistics.

When applying CM to individual-level data, we re-compute the marginal statistics by performing linear regression where we add the set of variants that are selected as covariates. We utilize the LD between the variants, which we obtain from a reference dataset, when applying CM to summary statistics data. In this case, we re-compute the marginal statistics for the i^{th} variant as follows:

$$Z_i^{new} = \frac{(z_i - z_j r_{ij})}{\sqrt{1 - r_{ij}^2}} \qquad \text{(Equation 19)}$$

when we have selected the j^{th} variant as causal. Let z_i indicate the marginal statistics for the i^{th} variant and r_{ij} the genotype correlations between the i^{th} and j^{th} variants.

Datasets

We obtained the summary statistics for GTEx eQTL dataset (release v6p, dbGaP: phs000424.v6.p1). We estimated the LD structure using the available genotypes in the GTEx dataset. We considered 44 tissues and applied our method to all eGenes in order to detect loci that harbor allelic heterogeneity. eGenes are genes that have at least one significant eQTL (p value $< 10^{-5}$). The number of individuals ranges from 70 to 361 depending on the tissues of interest.

We obtained the summary statistics of blood eQTL for 373 European individuals from the GEUVADIS (Genetic European Variation in Disease) website. We approximated LD structure from the 1000G European population. We applied our method to the 2,954 eGenes in GEUVADIS to detect AH loci.

We obtained the summary statistics from the MuTHER (Multiple Tissue Human Expression Resource) website and utilized the skin and fat (adipose) tissues. MuTHER dataset consist of 856 individuals. We then approximated LD from the 1000G European population. We obtained 1,433 eGenes for skin and 2,769 eGenes for adipose.

We used the high-density lipoprotein cholesterol (HDL-C) trait.[21] We considered only the GWAS hits, which are reported in a previous study. We applied ImpG-Summary[26] to impute the summary statistics with 1000G as the reference panel. We used the 1000G European population to estimate the LD. HDL-C consists of around 100,000 individuals. We used the 37 loci that is reported in previous study,[21] which have at least one significant variant (genome-wide significance level of 5×10^{-8}), as an input to CAVIAR. Following common protocol in fine-mapping methods, we apply CAVIAR to loci that have at least one statistically significant variant.

We analyzed the recent GWASs on major depression disorder (MDD)[22] and schizophrenia.[3] The major depression disorder study has 2 and the schizophrenia study has 108 loci identified to contain at least one significant variant. We utilized the summary statistics provided by each study and approximated the LD using the 1000G populations. SCZ consists of 36,989 case subjects and 113,075 control subjects, and MDD consists of 5,303 case subjects and 5,337 control subjects.

Data Simulation

Simulating Genotypes

We first simulated genotypes using HAPGEN2,[27] where we utilized the 1000G CEU population as initial reference panels. We simulated 1,000 individuals. We focus on the chromosome 1 and the GWAS variants that are obtained from the NHGRI catalog.[28] We filter out monomorphic SNPs and SNPs with low minor alley frequency (MAF ≤ 0.01).

Simulated Summary Statistics with No Epistasis Interaction

We simulated the summary statistics using LD structure and effect size for causal variants. It is known that summary statistics follows a MVN distribution ($S \sim N(\Sigma\Lambda, \Sigma)$). The mean of statistics is $\Sigma\Lambda$ and the variance of statistics is Σ. As mentioned above, Σ is the LD structure and $\Lambda = [\lambda_1, \lambda_2, ..\lambda_m]$ is the true effect size used to simulate data. For variants that are non-causal, we set the effect size to zero. This process of simulating summary statistics is used in our previous studies.[18,29–33] Effect size for the causal variants is set to obtain the desired power. In our simulated datasets, effect size is computed for the genome-wide significant level ($\alpha = 10^{-8}$). We use binary search to compute the value of the effect size for a desired statistical power. The statistical power is defined as

$$\begin{aligned}\text{Power} &= 1 - \frac{1}{\sqrt{2\pi}} \int_{\Phi^{-1}(\alpha/2)+\lambda_i}^{\Phi^{-1}(1-\alpha/2)+\lambda_i} e^{-\frac{1}{2}x^2} dx \\ &= \Phi\big(\Phi^{-1}(\alpha/2) + \lambda_i\big) + 1 - \Phi\big(\Phi^{-1}(1-\alpha/2) + \lambda_i\big)\end{aligned}$$

where α is the significant threshold. Moreover, Φ and Φ^{-1} denote the cumulative density function (CDF) and inverse of CDF for the standard normal distribution.

We use false positive (FP) and true positive (TP) as metrics to compare different methods. FP indicates the fraction of loci that harbor one causal variant and are incorrectly detected as loci that harbor AH. TP indicates the fraction of loci that harbor AH and are correctly detected.

Simulating Summary Statistics with Epistasis Interaction

We simulated the genotypes similar to the case where we have no epistasis interaction, which is mentioned above. Then, we simulated phenotypes using the following model:

$$Y = x_i x_j \beta_{ij} + e, \qquad \text{(Equation 20)}$$

where x_i and x_j indicate the standardized genotypes for i^{th} and j^{th} variants, respectively. Moreover, β_{ij} is the epistasis interaction effect size. We set β_{ij} such that we obtained the desired heritability for the simulated phenotype. Then, we computed the marginal statistics for each variant utilizing the linear regression single variant testing.

Results

Overview of Identifying Allelic Heterogeneity

Our method utilizes the shape of marginal statistics (statistics obtained from GWASs such as z-score) and the patterns

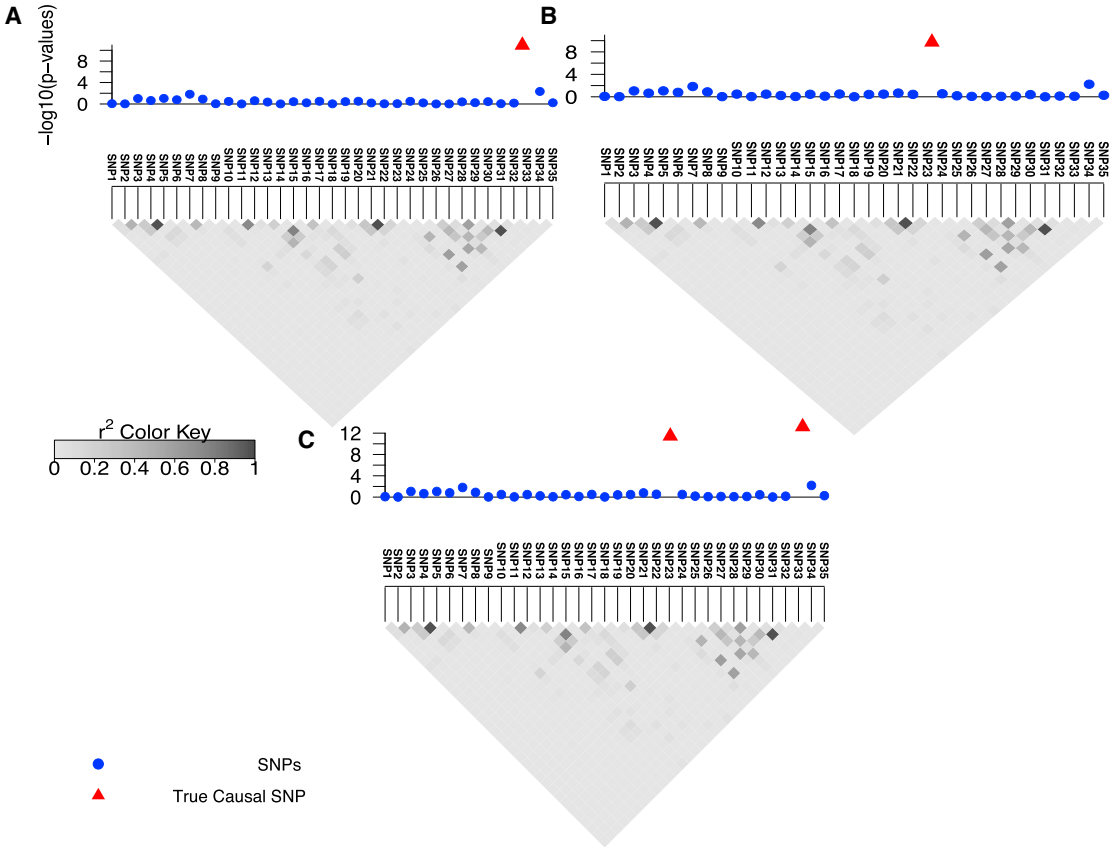

Figure 1. Overview of CAVIAR for Detecting Allelic Heterogeneity Regions
(A and B) The marginal statistics for a locus where we have implanted one causal variant. In (A), SNP33 is causal and in (B), SNP23 is causal.
(C) The same locus where both SNP23 and SNP33 are causal.
In these figures, the x axis is the negative logarithm of the p values for each locus to indicate the strength of the marginal statistics. The gray triangle below each figure indicates the LD pattern. Each square indicates the correlation between two variants, and the magnitude of the correlation is shown by the color intensity of the square. The darker the square, the higher the correlation between two variants.

of LD to detect whether or not there is evidence for AH. In Figures 1A and 1B, we illustrate a simple example with no AH. However, we can guess that the region shown in Figure 1C harbors AH; here, we observe two high independent peaks in the region. Unfortunately, detecting AH is less intuitive in regions that are more complicated.

The input to our method is the LD structure of the locus and the marginal statistics for each variant in the locus. The LD between each pair of variants is computed from genotyped data or is approximated from HapMap or 1000G data.[34,35] We use the fact that the joint distribution of the marginal statistics follows a multivariate normal distribution (MVN)[18,30,36,37] to compute the posterior probability of each subset of variants being causal, as described below. Then, we compute the probability of having *i* independent causal variants in a locus by summing the probability of all possible subsets of size *i* (sets that have *i* causal variants). We consider a locus to be AH when the probability of having more than one independent causal variant is more than 80%.

We would like to emphasize that using only summary statistics and LD information is insufficient for differentiating tightly linked variants. For example, if two variants are in perfect pairwise LD (correlation of 1), it is impossible to detect with just the summary statistics whether only one of the variants is causal or both are causal. Therefore, our estimates are just a lower bound on the amount of loci with AH for a given complex trait.

CAVIAR Is More Accurate than Existing Conditional Method

In order to assess the performance of our method, we generated simulated datasets. We used HAPGEN2,[27] a widely used software, to generate a case-control study using the European population obtained from the 1000G. Then, we implanted one or two causal variants in a region and generated marginal statistics for each variant (described in the Subjects and Methods). We applied CAVIAR to all the simulated datasets to detect loci that harbor AH. We generated two datasets. In the first dataset, we set power to detect the causal variants at 10%, 30%, 50%, or 70%. In the second dataset, we set power to detect the causal variants at 20%, 40%, 60%, or 80%. We compared our results with the conditional method (CM). We use false

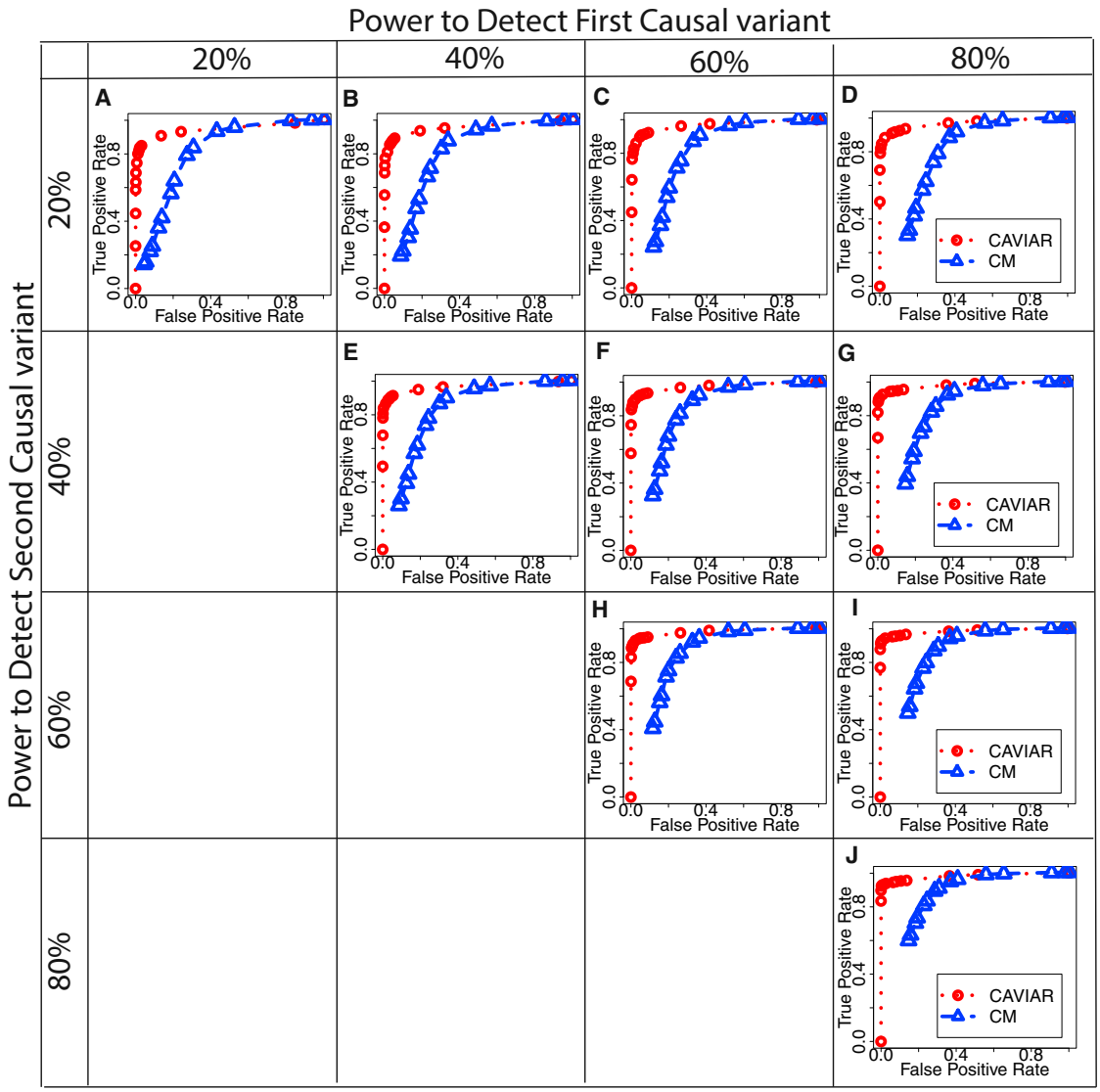

Figure 2. ROC Curve for CAVIAR and CM

We implant one causal variant to compute the false positive (FP) rate. FP indicates loci that harbor one causal variant; however, these loci are detected as AH. We implant two causal variants to compute the true positive (TP) rate. TP indicates loci that harbor AH and are detected correctly. We range the effect size such that the power at the causal variant is 20%, 40%, 60%, and 80% at the genome significant level 10^{-8}. We obtain these results from simulated data with no epistasis interaction. We simulated data using 1,000 individuals and set γ to 0.001.

positive (FP) and true positive (TP) as metrics for comparison. FP indicates the fraction of loci with one causal variant that are incorrectly detected as loci with AH. TP indicates the fraction of loci with AH that are correctly detected. We found that our method has higher TP compared to CM for the same FP rate. Figures 2 and S1 illustrate the ROC curves for the first and second simulated datasets. In all these experiments, we set γ to 0.001.

It is possible that the most significant variant is not causal.[18,30] In some cases, there may exist more than one causal variant in a locus; however, a variant that is in tight LD with the two causal variants tends to have higher marginal statistics.[18] This phenomenon is recently known as "ghost effect."[38] In fact, in the above simulation when we simulate more than one causal variant in a locus, we observed a ghost effect in few loci. However, the chance of a ghost effect occurring in simulated data is small. Thus, we extend our simulation to address this issue. We simulated data where we have more than one causal variant and generated cases only where a ghost effect occurs in a locus. We simulated cases where the power to detect the first causal variant is 50% and the power to detect the second causal variant ranges between 10%, 30%, 50%, and 70%. We observed that CAVIAR tends to have high recall rate even when the locus has ghost effect (see Table S1).

CAVIAR running time depends on the number of variants (m) in a locus and the maximum number of causal variants allowed. In our simulation, we set the maximum number of causal variants as six. In a locus with 50 variants

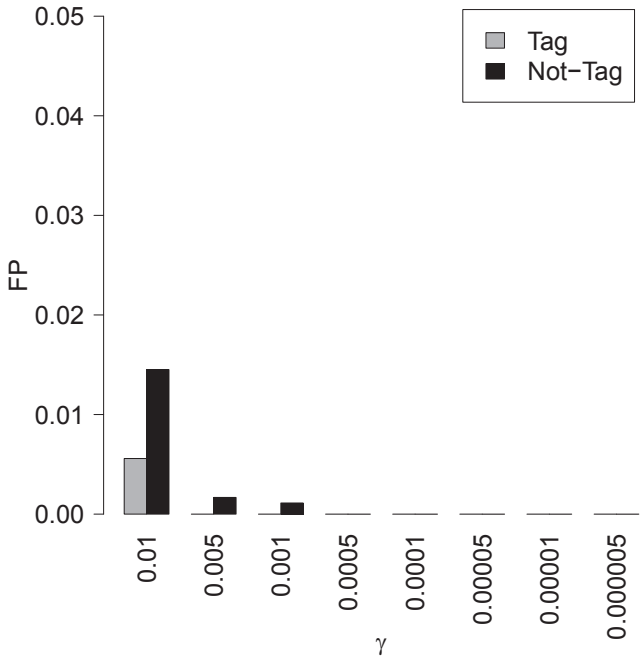

Figure 3. CAVIAR Has Low FP Even When the True Causal Variant Is Not Collected
Thus, most loci that are detected by CAVIAR to harbor AH are most probably true. x axis indicates the prior probability of causal variant (γ). We set γ to 0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005, 0.000001, and 0.000005.

where the maximum number of causal variants is 6, CAVIAR takes 26 min to finish. CAVIAR runs in less than 3 min for loci with 50 variants where the maximum number of causal variants is set to 5 or fewer (see Table S2). These results are obtained by running CAVIAR on a MacBook Air with 1.4 Ghz Intel Core i5 CPU. CAVIAR memory footprint is less than 1 Gb.

CAVIAR Has a Low False Positive even When the Causal Variant Is Not Included

In the previous section, we show CAVIAR has an extremely low FP and a high TP rate in detecting loci with AH. In the simulated data, the causal variant was included. However, in real datasets we cannot guarantee that information will exist for all the causal variants. One possible reason for detecting a locus as AH could be that the actual causal variant is not included or tagged in the data. In this section, we show that loci detected by CAVIAR with AH are rarely due to the fact that the actual causal variant is not included.

We simulate datasets where we implant one causal variant in a locus and generate marginal statistics in a method similar to the previous section. Next, we remove the causal variant from our analysis and use the remaining variants in the locus as an input to CAVIAR. We observe the FP is extremely low (FP < 0.015), even when the causal variant was not included in the analyzed data (see Figure 3). Our conclusion is that CAVIAR may fail to detect AH in some loci, but a very small proportion of loci where we detected AH are incorrect.

Setting CAVIAR Parameters to Detect Loci that Harbor AH

There are two main input parameters for CAVIAR, excluding the summary statistics of a locus: the prior probability that a variant is causal (γ) and AH posterior probability (AHPP) threshold. We consider a locus to harbor AH when the summation of posterior probability for two and more causal variants exceed AHPP threshold. To select γ and AHPP, we simulated data as mentioned above for different ranges of causal variants. We simulated datasets where the power to detect the causal variants is 10%, 30%, 50%, or 70%. In all these simulations, we implanted one or two causal variants and computed the FP and TP. The results of these experiments are shown in Figures S2–S11. We observed that to obtain extremely low FP and high TP, the best range for γ is 0.01 to 0.0001, where AHPP is set to 0.7 to 0.8. We set the γ to 0.001 from our previous results in fine-mapping.[18,30] Thus, we can apply CAVIAR to perform fine-mapping as well as to detect the loci that harbor AH. Furthermore, to be extremely stringent, we set AHPP to 0.8.

In addition, when we set AHPP to 0.8, we vary γ among 0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005, 0.000001, and 0.000005. We observed that for different values of γ, CAVIAR tends to have extremely low FP (FP ≤ 0.001) and high TP. The result for this experiment, when the power to detect the causal variants is 50%, is shown in Figure S12.

CAVIAR Is Robust to Out-of-Sample LD Structure

The LD structure can be computed from raw genotypes when they are available (in-sample LD). However, in most datasets, we do not have access to the raw genotypes. Thus, we approximate the LD utilizing the 1000G[34,35] or HapMap[39] datasets. Estimating the LD structure from a reference panel is known as out-of-sample LD.

In this experiment, we want to investigate the effect of using misspecified LD structures. We simulated the marginal statistics by utilizing the LD structure (LD matrix) obtained from HAPGEN2. Next, we simulate the marginal statistics and generate a misspecified LD by adding standard Gaussian noise, $N(0,\tau)$, to each element of the original LD matrix. We use the simulated marginal statistics and the misspecified LD as an input to CAVIAR. We simulated 10,000 loci with one causal variant to compute the FP, and we simulated 10,000 loci with two causal variants to measure TP. We vary the variance of the Gaussian noise (τ) between 0.01, 0.02, 0.05, 0.1, 0.2, 0.25, and 0.3. In this experiment, we observed low FP and high TP for CAVIAR results. The result for this experiment is shown in Figure S13. It is worth mentioning that the probability of observing a τ greater than 0.1 is extremely low.

CAVIAR Accurately Detects the Number of Causal Variants in a Locus when All the Variants Are Collected

In the previous sections, we have shown that CAVIAR is accurate in detecting loci that harbor AH. An additional benefit of our method is that it can accurately detect the number of causal variants in a locus. We simulated the

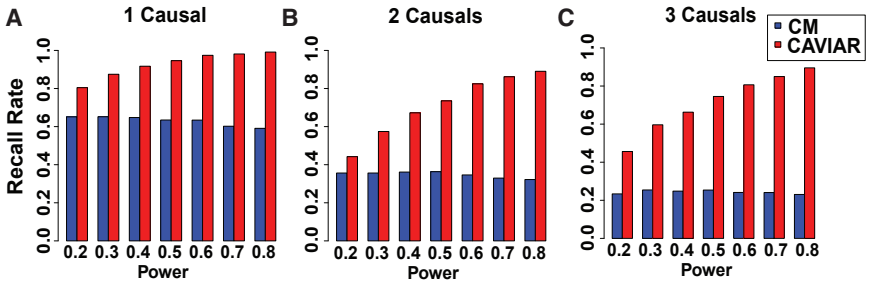

Figure 4. CAVIAR Is More Accurate than CM to Detect the Number of Causal Variants

The x axis is the power of causal variants, and the y axis is the accuracy to detect the number of causal variants in a locus. We implanted one, two, and three causal variants. We compute the recall rate as the fraction of simulations where the number of causal variants in a locus is predicted correctly. Recall rate of each method for different number of causal variants: (A) one causal variant, (B) two causal variants, and (C) three causal variants. We vary the statistical power to detect the causal variant among 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8.

phenotypes similar to the previous sections. In these experiments, we implanted one, two, and three causal variants in a locus. We set the effect size of the causal variant such that the statistical power to detect all causal is 20%, 30%, 40%, 50%, 60%, 70%, and 80%. In these experiments, power is the total probability to detect all causal variants. In this case, we have k causal variants, and we want the statistical power to be p. We set the effect size of k causal variants such that the summation of the power for all causal variants is equal to p. We use 1,000 individuals in our experiments.

As CAVIAR provides probability values for different number of causal variants, we consider a locus to have i independent causal variants where the probability of having i causal variants is the maximum probability for different numbers of causal variants. We allow up to six causal variants when applying CAVIAR. In the case of CM, the number of causal variants in a locus is equivalent to the number of conditional steps that we perform until the p value of all variants is higher than (α/m) (Bonferroni correction), where m is the total number of variants in a locus and α is 0.05. We compute the recall rate as the fraction of simulations where a method correctly predicted the number of causal variants in a locus. In Figure 4, we plot the recall rate of CAVIAR and CM in detecting the number of causal variants. We observe that, in comparison to CM, CAVIAR has a much higher recall rate in detecting the true number of causal variants in a locus.

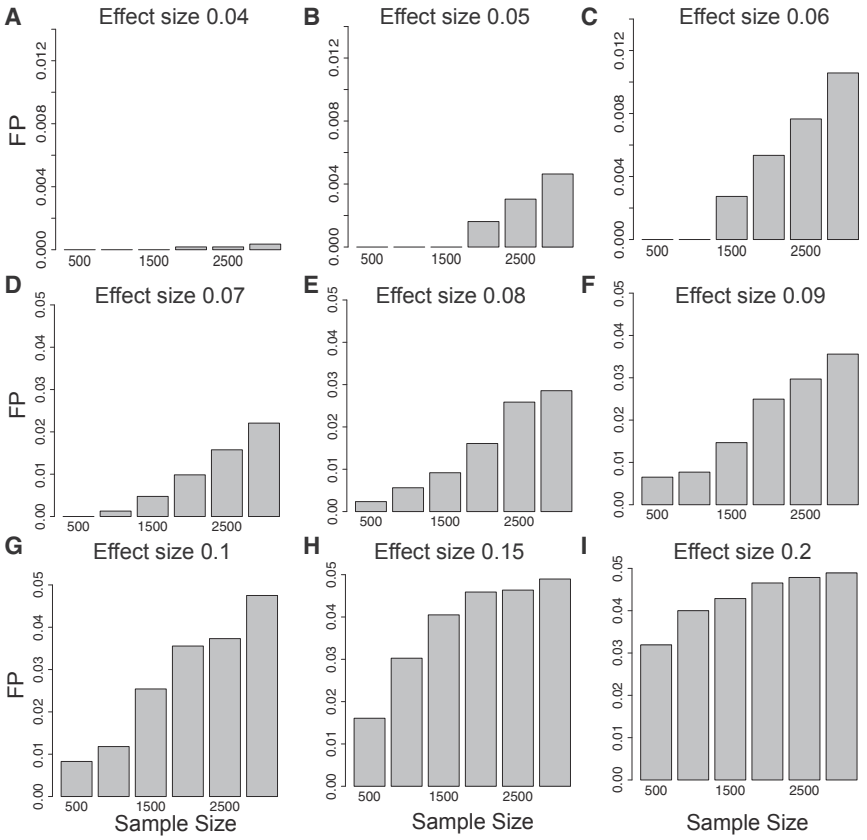

Figure 5. CAVIAR Distinguishes between Epistatic Interaction and Allelic Heterogeneity

The x axis is the sample size that we vary between 500, 1,000, 1,500, 2,000, 2,500, and 3,000 individuals. The y axis is the false positive (FP) rate. We simulated datasets where we have epistatic interaction and compute the FP as the number of cases where CAVIAR incorrectly detects these loci to harbor AH. Shown are the FP for different effect sizes of the epistatic interaction.

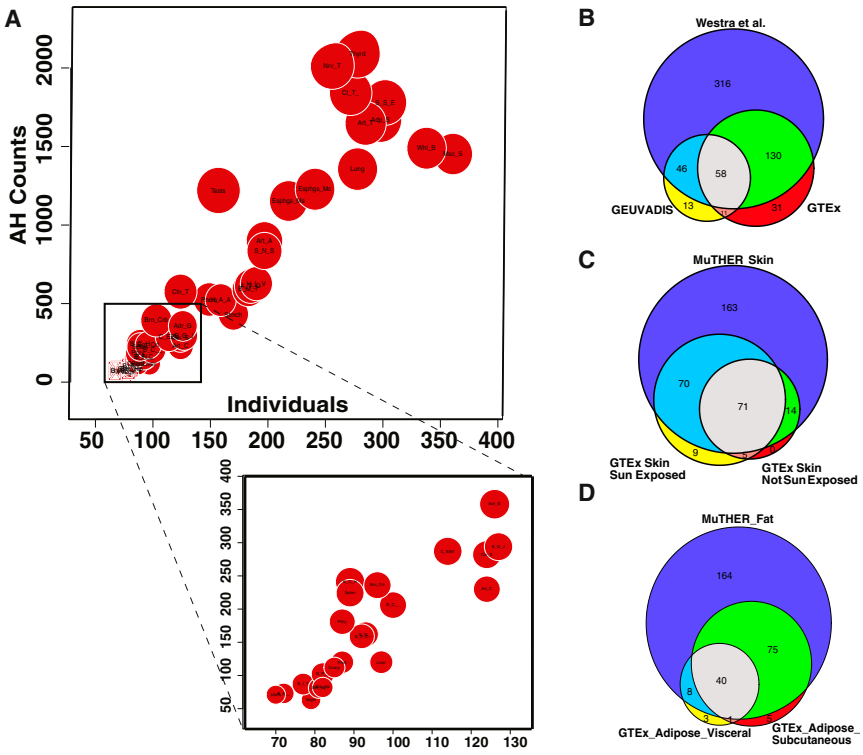

Figure 6. Levels of Allelic Heterogeneity in eQTL Studies

(A) Linear relationship between the amount of AH and sample size. Each red circle indicates a different type of tissue from the GTEx dataset. The size of each red circle is proportional to the number of genes that harbor a significant eQTL (eGenes).

(B–D) Significant overlap between AH estimations for different eQTL datasets, shown for (B) blood ($p = 7.9 \times 10^{-97}$), (C) skin ($p = 4.9 \times 10^{-63}$), and (D) adipose ($p = 1.1 \times 10^{-69}$) tissue. p values are computed using a hypergeometric test that is implemented in the SuperExactTest[43] software.

CAVIAR Distinguishes between Epistatic Interaction and Allelic Heterogeneity

It is possible to incorrectly detect a locus with AH due to the epistatic interaction in that locus. In this section, we utilize simulated data to illustrate that CAVIAR rarely detects AH in loci where the true genetics architecture is epistasis. We simulated different datasets where we implanted epistatic interactions between two randomly selected variants in a locus. Then, we generated simulated phenotypes using the linear additive model (described in the Subjects and Methods). We vary the number of individuals in each dataset among 500, 1,000, 1,500, 2,000, 2,500, and 3,000. In addition, for each dataset, we vary the effect size among 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, and 0.2. For each simulated dataset, we simulated 5,000 different marginal statistics. CAVIAR has an extremely low false positive rate in these experiments (see Figure 5). In Figure 5, we show the results for the effect size of 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, and 0.2. The results for effect size of 0.01, 0.02, and 0.03 are not shown because the FP is zero for these values. As a result, CAVIAR detects only a small fraction of epistatic interactions as AH. It is worth mentioning that the amount of epistatic interactions for different traits is very low. Thus, these results indicate that the loci that harbor AH and are detected by CAVIAR are not artifacts of epistatic interaction.

Prevalence of Allelic Heterogeneity in eQTL Datasets

We used four datasets to examine the extent of AH in eQTL datasets (GTEx,[20] GEUVADIS,[40] MuTHER,[41] and Wester et al.[42]). In the GTEx dataset,[20] we have access to 44 tissues. In each tissue, we have around 22,000 genes (probes). We obtained the marginal statistics and genotypes from the GTEx project[20] for each gene in all tissues. Then, we filtered out genes lacking at least one significant SNP. We set the significant cut-off threshold to a p value of 10^{-5}. Genes that have a significant *cis*-eQTL SNP are known as eGenes. We applied our method to detect AH loci only to eGenes. We found that 4%–23% of the eGenes show evidence for AH (with probability >80%) (Figure 6, Table 1). In addition, we applied the CM to the same set of eGenes. We observed that 50%–80% of loci detected by the CM to have AH were also detected by CAVIAR (see Table 1). Furthermore, we observed that CM detects 13.6%–44% of loci detected by CAVIAR as AH.

The number of eGenes detected in a tissue depends on the statistical power to detect a significant variant associated with the gene expression. The statistical power is highly correlated to the number of samples for that tissue. We hypothesized that there might also exist correlation between the sample size and the number of loci with AH. Indeed, we observed that the proportion of eGenes with AH for each tissue is in a linear relationship with the sample size, as shown in Figure 6A ($R^2 = 0.85$, $p = 2.2 \times 10^{-16}$). This result indicates that statistical power prevents the identification of AH at other loci.

To check the reproducibility of the AH detection, we compared the results from GTEx blood data with results from two other blood eQTL studies: GEUVADIS[40] and Wester et al.[42] We tested the overlap between genes with AH for skin and adipose tissues based on the GTEx[20] and MuTHER[41] datasets. We only considered eGenes that are common between the studies. In all comparisons, we observed a high reproducibility rate for the detection of AH in blood (Figure 6B, $p = 7.9 \times 10^{-97}$), skin (Figure 6C, $p = 4.9 \times 10^{-63}$), and adipose (Figure 6D, $p = 1.1 \times 10^{-69}$) tissues. We compute the significant overlap between AH

Table 1. List of 44 Tissues in GTEx

Tissue	#Individual	#eGene	#AH (CAVIAR)	#AH/#eGene	#AH (CM)	#Overlap between (%) CAVIAR & CM
Vagina	79	1,535	63	0.0410	26	13 (50.0%)
Brain, anterior cingulate cortex BA24	72	1,745	73	0.0418	27	10 (42.4%)
Small intestine terminal ileum	77	1,978	87	0.0439	32	17 (53.1%)
Brain, hypothalamus	81	1,750	81	0.0462	31	17(54.8%)
Uterus	70	1,504	71	0.0472	33	14(42.4%)
Brain, putamen basal ganglia	82	2,144	102	0.0475	52	24 (46.1%)
Brain, hippocampus	82	1,713	82	0.0478	35	13 (37.1%)
Liver	97	2,148	120	0.0558	40	18 (45.0%)
Prostate	87	2,088	120	0.0574	31	21 (67.7%)
Brain, nucleus accumbens basal ganglia	93	2,596	162	0.0624	70	41 (58.5%)
Brain, frontal cortex BA9	92	2,547	159	0.0624	70	40 (57.1%)
Ovary	85	1,776	112	0.0630	58	30 (51.7%)
Pituitary	87	2,708	181	0.0668	71	39 (54.9%)
Brain, cerebellar hemisphere	89	3,455	241	0.0697	115	59 (51.3%)
Brain, caudate basal ganglia	100	2,939	206	0.0700	95	54 (56.8%)
Spleen	89	3,141	224	0.0713	107	59 (55.1%)
Brain, cortex	96	3,009	236	0.0784	91	51 (56.0%)
Artery coronary	124	2,897	230	0.0793	119	65 (54.6%)
Colon sigmoid	124	3,247	282	0.0868	130	87 (66.4%)
Cells, EBV transformed lymphocytes	114	3,280	287	0.0875	126	81 (64.2%)
Esophagus, gastresophageal junction	127	3,231	294	0.0909	138	100 (72.4%)
Brain, cerebellum	103	4,278	393	0.0918	181	119 (65.7%)
Adrenal gland	126	3,636	358	0.0984	189	110 (58.2%)
Stomach	170	4,007	433	0.1080	146	103 (70.5%)
Pancreas	149	4,372	526	0.1203	235	153 (65.1%)
Colon transverse	124	4,771	576	0.1207	279	203 (72.7%)
Heart, atrial appendage	159	4,174	522	0.1250	192	125 (65.1%)
Breast mammary tissue	183	4,600	590	0.1282	246	178 (72.3%)
Adipose visceral omentum	185	4,611	611	0.1325	296	223 (75.3%)
Heart, left ventricle	190	4,526	627	0.1385	278	205 (73.7%)
Testis	157	8,333	1,220	0.1464	349	257 (73.6%)
Artery aorta	197	5,850	898	0.1535	435	338 (77.7%)
Skin, not sun exposed, suprapubic	197	5,371	835	0.1554	180	138 (76.6%)
Esophagus muscularis	218	6,431	1,153	0.1792	515	411 (79.8%)
Esophagus mucosa	241	6,849	1,228	0.1792	605	480 (79.3%)
Lung	278	7,026	1,356	0.1929	662	537 (81.1%)
Adipose, subcutaneous	298	7,806	1,669	0.2138	870	714 (82.6%)
Muscle, skeletal	361	6,687	1,452	0.2171	715	614 (85.8%)
Whole blood	338	6,822	1,489	0.2182	792	656 (82.8%)
Skin, sun exposed, lower leg	302	8,093	1,780	0.2199	549	467 (85.0%)

(*Continued on next page*)

Table 1. ***Continued***

Tissue	#Individual	#eGene	#AH (CAVIAR)	#AH/#eGene	#AH (CM)	#Overlap between (%) CAVIAR & CM
Artery, tibial	285	7,443	1,647	0.2212	796	653 (82.0%)
Cells, transformed fibroblasts	272	7,915	1,841	0.2325	965	798 (82.6%)
Thyroid	278	8,931	2,088	0.2337	932	786 (84.3%)
Nerve tibial	256	8,429	2,012	0.2386	1,075	881 (81.9%)
Total	7,014		28,717			

Tissues are sorted based on the number of samples. #Individual indicates the number of samples for each tissue. #AH (CAVIAR) is the number of loci detected by CAVIAR that harbor AH. #AH(CAVIAR)/#eGene is the fraction of eGenes that are detected to harbor AH. #AH(CM) is the number of loci detected by the conditional method (CM) to harbor AH.

estimations for different eQTL datasets. We compute the p values using a hypergeometric test that is implemented in the SuperExactTest[43] software. In addition, Jansen et al.[44] recently performed CM on a blood eQTL dataset to detect AH genes. We computed the overlap for AH loci detected by CAVIAR in GTEx blood tissue. We observed that 492 AH loci are common between CAVIAR and the Jansen et al.[44] results. CAVIAR detected 1,489 AH loci and Jansen et al.[44] detected 2,496 AH loci (see Figure S14).

Prevalence of Allelic Heterogeneity in GWAS Datasets

To measure the level of AH in a human quantitative trait, we applied our method to a GWAS of high-density lipoprotein (HDL).[21] We obtained the summary statistics available for HDL from the ImpG-Summary[26] webpage. There are 37 loci that are reported as significant for HDL.[21] Out of 37 loci, 13 (35%) showed evidence for AH with probability $\geq$80% (see Table S3). We also studied the results of GWASs focused on two psychiatric diseases: major depression disorder (MDD)[22] and schizophrenia (SCZ).[3] For MDD, we found evidence for AH at one of two loci. For SCZ, we identified 25 loci out of 108 (23%) with high probability of AH (see Table S4). One example of AH in SCZ is the locus on chromosome 18 that includes *TCF4* (MIM: 602272) (Figure 7A). The locus contains multiple associated SNPs that are distributed in different LD blocks (Figure 7B). According to our analysis, there are three or more causal variants in this locus with high probability (Figure 7C) (for similar results in other loci, see Figures S15–S51 for HDL and Figures S52–S179 for SCZ).

Discussion

We have proposed a novel probabilistic method to detect loci with AH. Our results show that our method is more accurate than the standard conditional method (CM). One of the main benefits of our method is that it requires only summary statistics. Summary statistics of a GWAS or eQTL study are widely available, so our method is applicable to most existing datasets. We have shown that AH is widespread and more common than previously estimated in complex traits, both in GWASs and eQTL studies. Since our method is influenced by statistical power and uncertainty induced by LD, the proportions of loci with AH detected in this study are just a lower bound on the true amount of AH. Thus, our study suggests that many, and maybe even most, loci are affected by AH.

Our results highlight the importance of accounting for the presence of multiple causal variants when characterizing the mechanism of genetic association in complex traits. Falling to account for AH can reduce the power to detect true causal variants and can explain the limited success of fine mapping of GWASs. It is worth mentioning that methods for fine-mapping[38,45,46] exist that allow for more than one causal variant in a locus; however, these methods require access to raw genotypes and phenotypes. One of the advantages of CAVIAR is that it requires only the existing summary statistics (e.g., marginal statistics per-SNP obtained from GWASs). Thus, methods such as CAVIAR are applicable to most GWAS datasets. Similarly, attempts to explain GWASs using eQTL data should be more successful with methods that assume some loci may include multiple causative variants (e.g., eCAVIAR[31] and RTC[47]).

One of the limitations of our method is that we assume that the observed marginal statistics are corrected for the population using PCA-based methods. Recently, linear mixed models (LMMs)[48–53] have become a popular correction for population structures that have cryptic relationships. Thus, the current version of our method is not applicable to summary statistics that have been corrected for population structure using LMM. However, we have shown in our previous works[30,36] that the same statistical model can be extended to incorporate the summary statistics that have been corrected for population structure using LMM. Unfortunately, in this case, the study's raw genotypes and phenotypes should be available in order to perform the desired analysis.

In summary, we have developed a method to detect the presence of AH in loci of complex traits. We show that, while the method may fail to detect AH in some loci, the false positive rate is very low. Thus, when our method detects a locus to have AH with a high probability, the prediction is very reliable. Since the amount of AH detected in our study is just a lower bound on the number of loci with AH, we suggest that AH is widespread in complex traits.

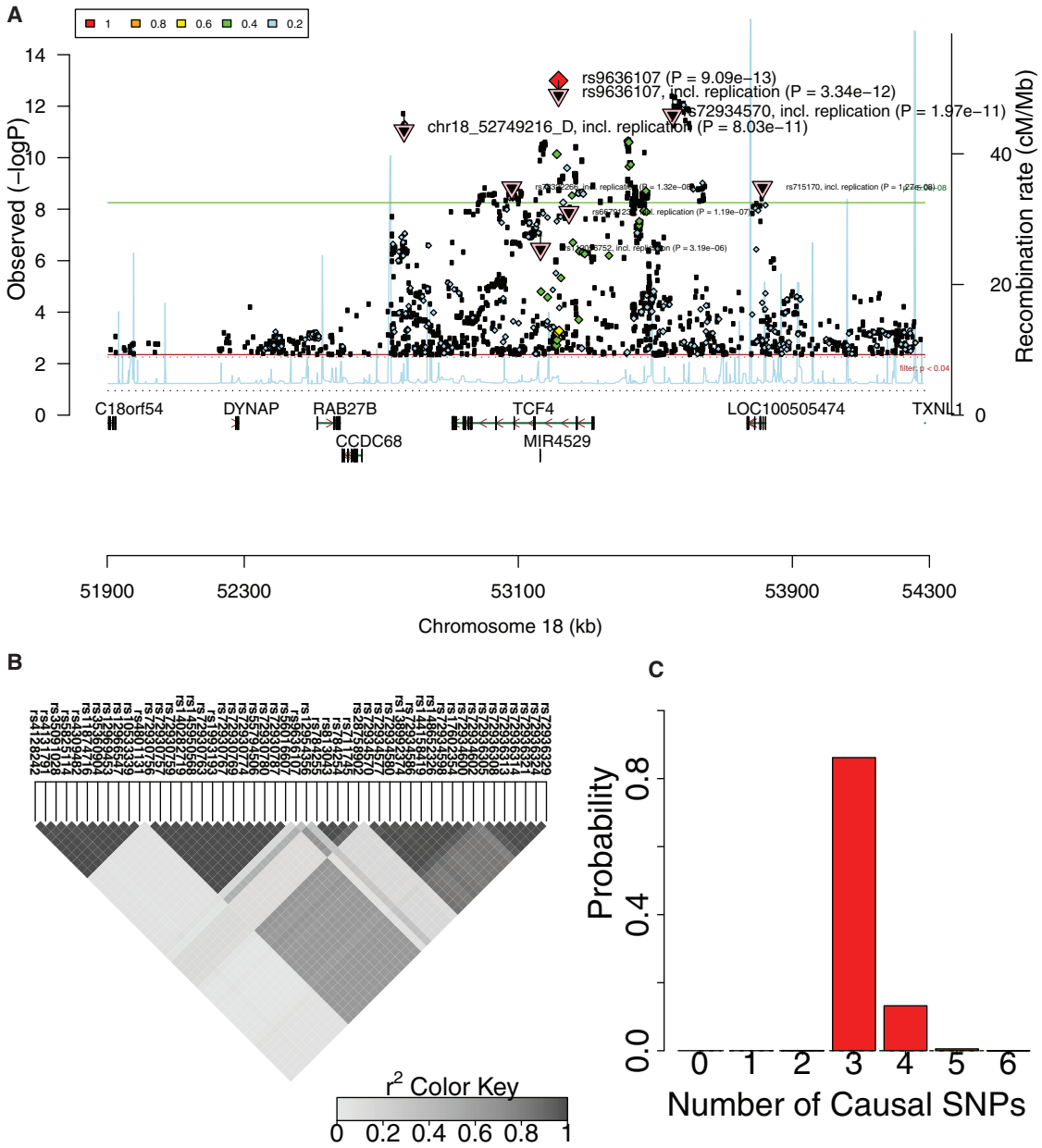

Figure 7. Allelic Heterogeneity in the *TCF4* Locus Associated with Schizophrenia
(A) Manhattan plot obtained from Ricopili consists of all the variants (7,193 variants) in a 1 Mbp window centered on the most significant SNP in the locus (rs9636107). We use PGC-SCZ52-may13 version of the data. This plot indicates multiple significant variants that are not in tight LD with the peak variant.
(B) LD plot of the 50 most significant SNPs showing several distinct LD blocks.
(C) Histogram for the probability of having different number of causal variants.

Supplemental Data

Supplemental Data include 179 figures and 4 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.04.005.

Acknowledgments

F.H., J.W.J.J., 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 the NIH BD2K award U54EB020403. A.V.S. is supported by a contract (HHSN268201000029C) to the Laboratory, Data Analysis, and Coordinating Center (LDACC) at The Broad Institute. S. Sankararaman was supported in part by NIH grant R00-GM 111744-03. G.K. is supported by the Biomedical Big Data Training Program (NIH-NCI T32CA201160). S. Shifman was supported by the National Institute for Psychobiology in Israel, founded by The Charles E. Smith Family and by the Israel Science Foundation (grant no. 688/12). We acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691).

Received: December 27, 2016
Accepted: April 7, 2017
Published: May 4, 2017

Web Resources

Blood eQTL browser, http://genenetwork.nl/bloodeqtlbrowser/
CAVIAR, http://genetics.cs.ucla.edu/caviar/
Complex Traits Genomic Group, http://www.cnsgenomics.com/software/
dbGaP, http://www.ncbi.nlm.nih.gov/gap
GEUVADIS, ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/
GTEx Portal (release v.6, dbGaP: phs000424.v6.p1), http://www.gtexportal.org/home/
HapGen2, http://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html
ImpG Summary, http://bogdan.bioinformatics.ucla.edu/software/impg/
LocusZoom, http://locuszoom.sph.umich.edu/locuszoom/
MDD summary statistics, https://www.med.unc.edu/pgc/files/resultfiles/converge.MDD.summary_stats.2Sep2015.tbl.gz
MuTHER, http://www.muther.ac.uk/Data.html
OMIM, http://www.omim.org/
Ricopili, http://data.broadinstitute.org/mpg/ricopili
SCZ summary, https://www.med.unc.edu/pgc/files/resultfiles/scz2.snp.results.txt.gz
SuperExactTest, https://cran.r-project.org/web/packages/SuperExactTest/index.html

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. 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.
3. Ripke, S., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K.-H., Holmans, P.A., Lee, P., Bulik-Sullivan, B., Collier, D.A., Huang, H., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
4. Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C., et al.; Type 1 Diabetes Genetics Consortium (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. *41*, 703–707.
5. Zhang, F., and Lupski, J.R. (2015). Non-coding genetic variants in human disease. Hum. Mol. Genet. *24* (R1), R102–R110.
6. Edwards, S.L., Beesley, J., French, J.D., and Dunning, A.M. (2013). Beyond GWASs: illuminating the dark road from association to function. Am. J. Hum. Genet. *93*, 779–797.
7. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.
8. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
9. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.
10. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al.; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
11. 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.
12. 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.
13. 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.
14. Estivill, X., Bancells, C., Ramos, C.; and The Biomed CF Mutation Analysis Consortium (1997). Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. Hum. Mutat. *10*, 135–154.
15. Hardison, R.C., Chui, D.H., Giardine, B., Riemer, C., Patrinos, G.P., Anagnou, N., Miller, W., and Wajcman, H. (2002). HbVar: A relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. Hum. Mutat. *19*, 225–233.
16. Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A.K., McRae, A.F., Yang, J., Gibson, G., Martin, N.G., Metspalu, A., et al. (2014). Detection and replication of epistasis influencing transcription in humans. Nature *508*, 249–253.
17. Wood, A.R., Tuke, M.A., Nalls, M.A., Hernandez, D.G., Bandinelli, S., Singleton, A.B., Melzer, D., Ferrucci, L., Frayling, T.M., and Weedon, M.N. (2014). Another explanation for apparent epistasis. Nature *514*, E3–E5.
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. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. *44*, 369–375, S1–S3.
20. 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.
21. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman,

D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.
22. Cai, N., Bigdeli, T.B., Kretzschmar, W., Li, Y., Liang, J., Song, L., Hu, J., Li, Q., Jin, W., Hu, Z., et al.; CONVERGE consortium (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature *523*, 588–591.
23. Darnell, G., Duong, D., Han, B., and Eskin, E. (2012). Incorporating prior information into association studies. Bioinformatics *28*, i147–i153.
24. Eskin, E. (2008). Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. Genome Res. *18*, 653–660.
25. 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.
26. Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D.P., Patterson, N., and Price, A.L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics *30*, 2906–2914.
27. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.
28. 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.
29. 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.
30. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B., and Eskin, E. (2015). Identification of causal genes for complex traits. Bioinformatics *31*, i206–i213.
31. 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.
32. 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.
33. 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.
34. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A.; and 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
35. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
36. Joo, J.W.J., Hormozdiari, F., Han, B., and Eskin, E. (2016). Multiple testing correction in linear mixed models. Genome Biol. *17*, 62.
37. Han, B., and Eskin, E. (2012). Interpreting meta-analyses of genome-wide association studies. PLoS Genet. *8*, e1002555.
38. Fang, M., and Georges, M. (2016). Bayesfm: a software program to fine-map multiple causative variants in gwas identified risk loci. bioRxiv. http://dx.doi.org/10.1101/067801.
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. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
41. Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., et al.; MuTHER Consortium (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. *7*, e1002003.
42. Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. *45*, 1238–1243.
43. Wang, M., Zhao, Y., and Zhang, B. (2015). Efficient test and visualization of multi-set intersections. Sci. Rep. *5*, 16923.
44. Jansen, R., Hottenga, J.-J., Nivard, M.G., Abdellaoui, A., Laport, B., de Geus, E.J., Wright, F.A., Penninx, B.W.J.H., and Boomsma, D.I. (2017). Conditional eQTL analysis reveals allelic heterogeneity of gene expression. Hum. Mol. Genet. *26*, 1444–1451.
45. Servin, B., and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet. *3*, e114.
46. Wallace, C., Cutler, A.J., Pontikos, N., Pekalski, M.L., Burren, O.S., Cooper, J.D., García, A.R., Ferreira, R.C., Guo, H., Walker, N.M., et al. (2015). Dissection of a complex disease susceptibility region using a bayesian stochastic search approach to fine mapping. PLoS Genet. *11*, e1005272.
47. 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.
48. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-Y.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354.
49. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics *178*, 1709–1723.
50. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods *8*, 833–835.
51. Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.
52. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.
53. Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. *38*, 203–208.