bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Widespread allelic heterogeneity in complex traits

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

[1]Department of Computer Science, University of California, Los Angeles, CA 90095, USA

[2]Bioinformatics IDP, University of California, Los Angeles, CA 90095, USA

[3]Cancer Program, The Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA

[4]Neurogenetics Program, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles, CA 90095, USA

[5]Department of Pathology and Laboratory Medicine, University of California, Los Angeles, CA 90095, USA

[6]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA

[7]Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem

[8]These authors contributed equally to this work

*Corresponding author: sagiv@vms.huji.ac.il, eeskin@cs.ucla.edu

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

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 GWAS and four expression quantitative trait loci (eQTL) datasets. We identified a total of 4152 loci with strong evidence of AH. The proportion of all loci with identified AH is 4-23% in eQTLs, 35% in GWAS of High-Density Lipoprotein (HDL), and 23% in schizophrenia. For eQTLs, we observed a strong correlation between sample size and the proportion of loci with AH (R^2=0.85, P = 2.2e-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.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 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 a causal sequence variant was 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 be 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 is causal variants that have a biological effect on the trait and are responsible for the association signal. The second category is variants that are statistically associated with the trait due to high LD with the causal variants. The third category is 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 vs. 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 specific Mendelian disease. For example, approximately 100 independent mutations are known to exist at the cystic fibrosis locus [14], and even more are present at loci causing inherited haemoglobinopathies [15]. In contrast to Mendelian traits, the extent of AH at loci contributing to common, complex disease

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 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 new method to quantify the number of independent causal variants at a locus responsible for the observed association signals in GWAS. Our method is incorporated into the CAusal Variants Identification in Associated Regions (CAVIAR) 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 portions 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

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

depression disorder (MDD) [22], where we observed 23%, 35%, and 50% of loci (respectively) with strong evidence for AH.

2 Results

2.1 Overview of identifying allelic heterogeneity

Our method utilizes the shape of marginal statistics (statistics obtained from GWAS such as z-score) and the patterns of LD to detect whether or not a locus harbors 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 [23, 24]. We use the fact that the joint distribution of the marginal statistics follows a multivariate normal distribution (MVN) [18, 25–27] 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 only using 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.

2.2 CAVIAR is more accurate than existing methods

In order to assess the performance of our method, we generated simulated data sets. We used HAPGEN2 [28], 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 simulated phenotypes using the linear mixed model (described in the Method section).

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

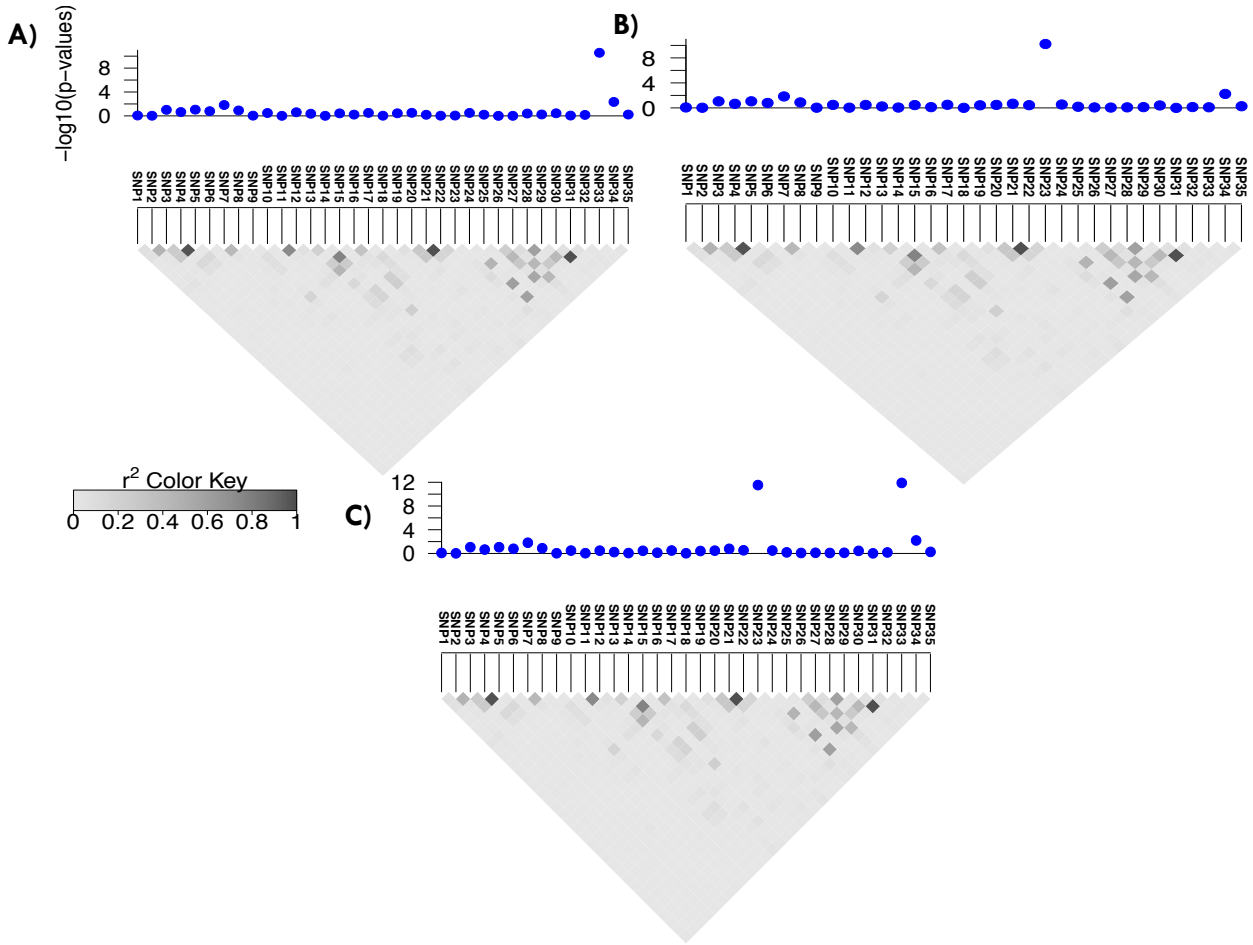

Figure 1. Overview of CAVIAR for Detecting Allelic Heterogeneity Regions. Panels (A) and (B) indicate the marginal statistics for a locus where we have implanted one causal variant. In panel (A), SNP23 is causal and in panel (B), SNP33 is causal. Panel (C) is the same locus where both SNP23 and SNP33 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 grey 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.

After generating the phenotypes, we performed a t-test to generate the marginal statistics for each variant. We applied CAVIAR to all the simulated data sets to detect loci that harbor AH. We generated two datasets. In the first datasets, we set the non-centrality parameter (NCP) to have 10%, 30%, 50% or 70% power to detect the causal variants. In the second datasets, we set the NCP of the causal variants to have 20%, 40%, 60%, or 80% power to detect the causal variants. We compared our results with the conditional method (CM). We use false 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

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correctly detected. We found that our method has higher TP compared to CM for the same FP rate. Figure 2 and Supplementary Figure 1 illustrate the ROC curves for the first and second simulated datasets.

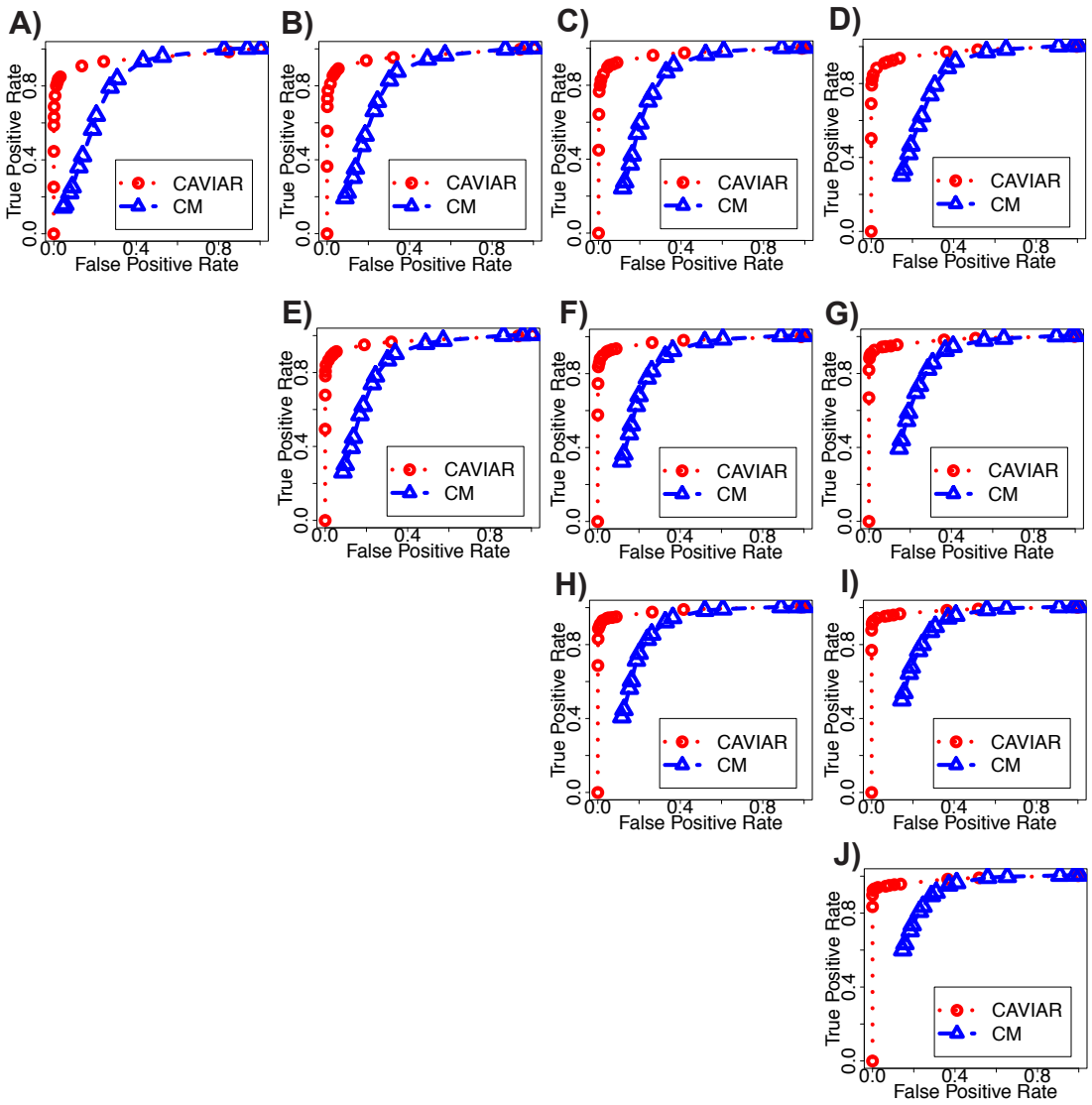

Figure 2. ROC curve for our method (CAVIAR) and CM. We implant one causal variant to compute the false positive (FP) rate. FP indicates loci that harbor one causal variant; however, these loci are detected as AH. We implant two causal variants to compete the true positive (TP) rate. TP indicates loci that harbor AH and are detected correctly. We range the NCP such that the power at the causal variant is 20%, 40%, 60%, and 80% at the genome significant level 10^{-8}. In (A) , (B), (C), and (D) NCP for the first causal SNP is set to 20% power. In (A) , (B), (C), and (D) NCP for the second causal SNP is set to 20%, 40%, 60%, and 80%, respectively. In (E), (F), and (G) NCP for the first causal SNP is set to 40% power. In (E), (F), and (G) NCP for the second causal SNP is set to 40%, 60%, and 80%, respectively. In (H) and (I) NCP for the first causal SNP is set to 60% power. In (H) the NCP of the second causal SNP is set to 60% power and in (I) the NCP of the second causal SNP is set to 80%. in (J) the NCP for both causal SNPs are set to 80% power.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.3 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 variants 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 2). Our conclusion is that CAVIAR may fail to detect AH in some loci, but a very small proportion of loci where we detected AH do not harbor AH.

2.4 CAVIAR is robust to different input parameters

There are two main input parameters to CAVIAR, excluding the summary statistics of a locus: the prior probability that a variant is causal (γ) and LD structure. The LD structure can be computed from raw genotypes when it is available. However, in most datasets, we do not have access to the raw genotypes. Thus, we approximate the LD utilizing the 1000G [23, 24] or HapMap [29] dataset.

We simulated marginal statistics in a way similar to previous sections. Then, 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 is shown in Supplementary Figure 2.

In the next 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. After, we simulate the marginal statistics, we 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 that

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

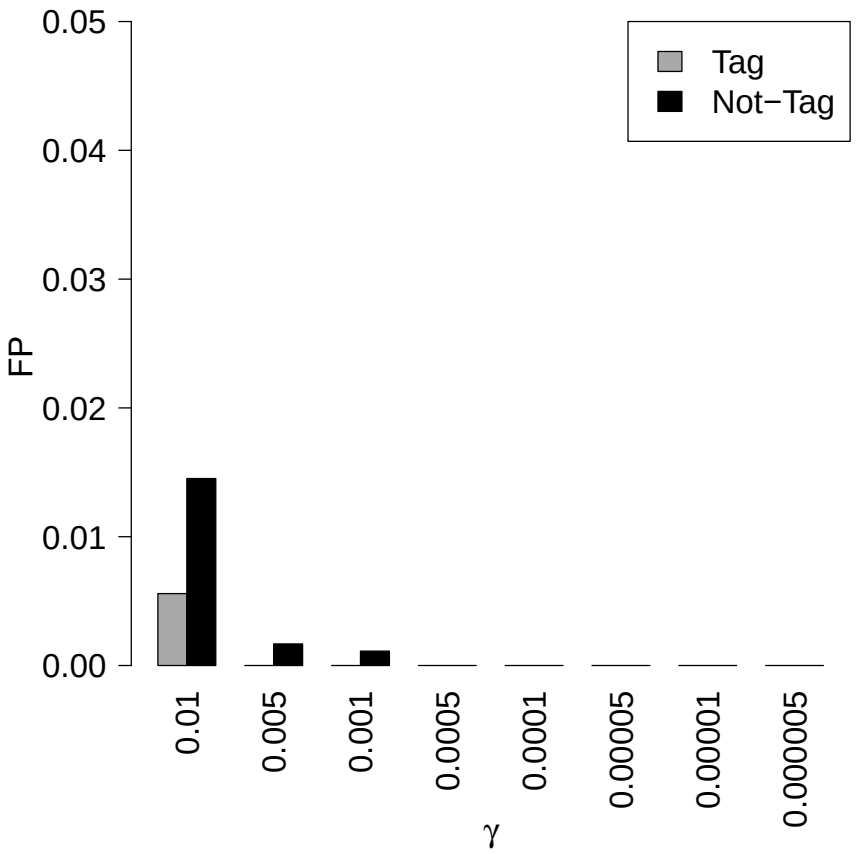

Figure 3. CAVIAR has low FP even when the true causal variant is not collected (untagged). Thus, most loci that are detected by CAVIAR to harbor AH are most probably true. X-axis indicates the prior probability of causal variant (γ). We set γ to 0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005, 0.000001, and 0.000005.

harbor one causal variant to compute the FP, and we simulated 10,000 loci that harbor 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 observe low FP and high TP for CAVIAR results. The result for this experiment is shown in Supplementary Figure 3.

2.5 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 new method is that it can accurately detect the number of causal variants in a locus. We simulated the phenotypes similar to the previous sections. In these experiments, we implanted one, two, and three causal variants in a locus. We set the NCP of the

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

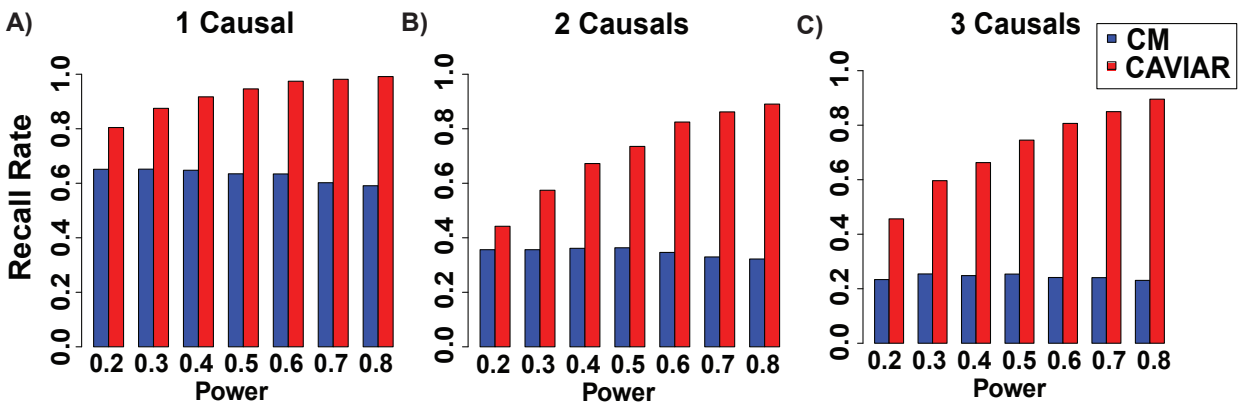

Figure 4. CAVIAR is more accurate than CM to detect the number of causal variants. The x-axis is the power of causal variants and the y-axis is the accuracy to detect the number of causal variants in a locus. We implanted one, two, and three causal variants. (A-C) Recall rate of each method for different number of causal variants, (A) one causal variant (B) two causal variants (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.

causal variant such that the statistical power is 20%, 30%, 40%, 50%, 60%, 70%, and 80%.

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. 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 $\frac{\alpha}{m}$ (Bonferroni correction) where m is the total number of variants in a locus and α is 0.05. We compute recall rate as the fraction of simulations where any methods correctly predicated 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 CAVIAR has much higher recall rate in detecting the true number of causal variants in a locus than CM.

2.6 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

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

simulated phenotypes using the linear additive model (described in the Method section). We vary the number of individuals in each dataset among 500, 1000, 1500, 2000, 2500, and 3000. 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 extremely low false positive 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 0.02 are not shown as the FP is zero for these values. As a result, CAVIAR only detects 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 the loci detected by CAVIAR that harbor AH are not artifacts of epistatic interaction.

2.7 Prevalence of allelic heterogeneity in eQTL datasets

We used four datasets to examine the extent of AH in eQTL datasets. We utilize version v6p of data, which consists of 44 tissues. For each tissue, we have around 22,000 genes (probes). For each gene in each tissue, we have access to the marginal statistics for the cis-eQTL analysis, which we obtained from the GTEx project [20] and genotype data that is used to compute the LD pattern for each gene. 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 eGene. We applied our method to detect AH loci only to eGene. 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).

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 (R^2=0.85, P = 2.2e-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 [30] and Wester et al. (2013)

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[31]. We tested the overlap between genes with AH for skin and adipose tissues based on the GTEx [20] and MuTHER [32] 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=7.9e-97), skin (Figure 6C, P=4.9e-63), and adipose (Figure 6D, P=1.1e-69) tissues.

Tissue	**#Individual**	**#AH (CAVIAR)**	**#AH/ #eGene**	**#AH (CM)**	**#Overlap between(%) CAVIAR & CM**
Vagina	79	63	0.0410	26	13 (50.0%)
Brain-Anterior-cingulate-cortex-BA24	72	73	0.0418	27	10 (42.4%)
Small-Intestine-Terminal-Ileum	77	87	0.0439	32	17 (53.1%)
Brain-Hypothalamus	81	81	0.0462	31	17(54.8%)
Uterus	70	71	0.0472	33	14(42.4%)
Brain-Putamen-basal-ganglia	82	102	0.0475	52	24 (46.1%)
Brain-Hippocampus	82	82	0.0478	35	13 (37.1%)
Liver	97	120	0.0558	40	18 (45.0%)
Prostate	87	120	0.0574	31	21 (67.7%)
Brain-Nucleus-accumbens-basal-ganglia	93	162	0.0624	70	41 (58.5%)
Brain-Frontal-Cortex-BA9	92	159	0.0624	70	40 (57.1%)
Ovary	85	112	0.0630	58	30 (51.7%)
Pituitary	87	181	0.0668	71	39 (54.9%)
Brain-Cerebellar-Hemisphere	89	241	0.0697	115	59 (51.3%)
Brain-Caudate-basal-ganglia	100	206	0.0700	95	54 (56.8%)
Spleen	89	224	0.0713	107	59 (55.1%)
Brain-Cortex	96	236	0.0784	91	51 (56.0%)
Artery-Coronary	124	230	0.0793	119	65 (54.6%)
Colon-Sigmoid	124	282	0.0868	130	87 (66.4%)
Cells-EBV-transformed-lymphocytes	114	287	0.0875	126	81 (64.2%)
Esophagus-Gastroesophageal-Junction	127	294	0.0909	138	100 (72.4%)
Brain-Cerebellum	103	393	0.0918	181	119 (65.7%)
Adrenal-Gland	126	358	0.0984	189	110 (58.2%)
Stomach	170	433	0.1080	146	103 (70.5%)
Pancreas	149	526	0.1203	235	153 (65.1%)
Colon-Transverse	124	576	0.1207	279	203 (72.7%)
Heart-Atrial-Appendage	159	522	0.1250	192	125 (65.1%)
Breast-Mammary-Tissue	183	590	0.1282	246	178 (72.3%)
Adipose-Visceral-Omentum	185	611	0.1325	296	223 (75.3%)
Heart-Left-Ventricle	190	627	0.1385	278	205 (73.7%)
Testis	157	1220	0.1464	349	257 (73.6%)
Artery-Aorta	197	898	0.1535	435	338 (77.7%)
Skin-Not-Sun-Exposed-Suprapubic	197	835	0.1554	180	138 (76.6%)
Esophagus-Muscularis	218	1153	0.1792	515	411 (79.8%)
Esophagus-Mucosa	241	1228	0.1792	605	480 (79.3%)
Lung	278	1356	0.1929	662	537 (81.1%)
Adipose-Subcutaneous	298	1669	0.2138	870	714 (82.6%)
Muscle-Skeletal	361	1452	0.2171	715	614 (85.8%)
Whole-Blood	338	1489	0.2182	792	656 (82.8%)
Skin-Sun-Exposed-Lower-leg	302	1780	0.2199	549	467 (85.0%)
Artery-Tibial	285	1647	0.2212	796	653 (82.0%)
Cells-Transformed-fibroblasts	272	1841	0.2325	965	798 (82.6%)
Thyroid	278	2088	0.2337	932	786 (84.3%)
Nerve-Tibial	256	2012	0.2386	1075	881 (81.9%)
Total	7014	28717			

Table 1. List of 44 tissues in GTEx. 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 conditional method (CM) that harbor AH.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.8 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]. Out of 37 loci, 13 (35%) showed evidence for AH with probability $\geq 80\%$ (see Supplementary Table 1). 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 Supplementary Table 2). One example of AH in SCZ is the locus on chromosome 18 that includes the $TCF4$ gene (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 Supplementary Figures 4-39 for HDL and Supplementary Figures 40-167 for SCZ).

3 Methods

3.1 Joint Distribution of Observed Statistics in Standard GWAS

In this section, we provide a brief description of statistical tests that are performed in GWAS and the joint distribution of computed marginal statistics. These statistics are used as an input to CAVIAR. We consider, we perform GWAS on a quantitative traits 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 zero and variance one. We use X_j to indicate the standardized minor allele count for all the individuals at the j-th variant. We have $\mathbf{1}^T X_i = 0$ 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 :

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

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 are 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 $Y \sim N(\mu \mathbf{1} + \beta_c X_c, \sigma_e^2 \boldsymbol{I})$. We use the maximum likelihood to compute the estimated β_c and σ_e which are denoted by $\hat{\beta}_c$ and $\hat{\sigma}_e$. Thus, we have:

$$\begin{aligned} &\mathcal{L}\left(Y|\mu, \beta_c, \sigma_e^2\right) = |2\pi\sigma_e^2 \boldsymbol{I}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2\sigma_e^2}\left(Y - \mu\mathbf{1} - \beta_c X_c\right)^T \left(Y - \mu\mathbf{1} - \beta_c X_c\right)\right), \\ &\hat{\mu} = \tfrac{1}{n}\mathbf{1}^T Y,\ \hat{\mu} \sim \mathcal{N}\left(\mu, \tfrac{\sigma_e^2}{n}\right), \\ &\hat{\beta}_c = \tfrac{X_c^T \boldsymbol{y}}{n},\ \sqrt{n}\tfrac{\hat{\beta}_c}{\sigma} \sim \mathcal{N}\left(\tfrac{\beta_c}{\sigma_e}\sqrt{n}, 1\right). \end{aligned} \tag{1}$$

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-squared distributed random 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 $\lambda_c = \frac{\hat{\beta}}{\hat{\sigma}}\sqrt{n}$ and variance of one,

$$S_c \sim t_{\lambda_c, n} \approx \mathcal{N}\left(\lambda_c, 1\right). \tag{2}$$

We assume that variant i is correlated with the causal variant c and the correlation between the two variants are r where $r = \frac{1}{n} X_i^T X_c$. Then, the marginal statistics 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 are r.

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

is causal and the variant i is not causal.

3.2 Computing the Likelihood of Causal Configuration

We can extend the joint distribution of marginal statistics for two variants to more general case. Assume we have m variants and the pair-wise 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 follow:

$$(S|\Lambda) \sim \mathcal{N}(\Sigma\Lambda, \Sigma), \tag{4}$$

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 new 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,

$$(\Lambda|C) \sim \mathcal{N}(0, \Sigma_c), \tag{5}$$

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 i-th diagonal element to zero if the i-th variant is not causal. Thus, the joint distribution follows a multivariate normal distribution,

$$(S|C) \sim \mathcal{N}(0, \Sigma + \Sigma\Sigma_c\Sigma), \tag{6}$$

3.3 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 causal variants in a locus. We have,

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

where $P(C)$ is the prior on the causal configuration C, $\mathcal{C}$ is the set of all possible causal configurations including the configuration all the variants 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 the probability definition holds.

In this paper, we use a simple prior for a causal status. We assume the probability of a variant to be causal is independent from other variants and the probability of a variant to be causal is γ. Thus, we compute the prior probability as, $\mathrm{P}\left(C^*\right) = \prod \gamma^{|c_i|}(1-\gamma)^{1-|c_i|}$. We utilize different values for γ as shown in Supplementary Figure 2. In our experiment, we set γ to 0.001 [33–35]. It is worth mentioning, 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, $\mathrm{P}\left(C^*|\Gamma = [\gamma_1, \gamma_2, \cdots \gamma_\ell]\right) = \prod \gamma_i^{|c_i|}(1-\gamma_i)^{1-|c_i|}$.

3.4 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 speed up process that reduces the time complexity to $O(m^2k)$ where k is the number of causal variants for a causal status. The number of causal variants is smaller than the total number of variants in a locus ($k << m$). Thus, we manage to speed up the likelihood computation by a factor of $\frac{m}{k}$.

According to Equation (6) to compute the likelihood of a causal status, we require to compute the following quantity:

$$(2\pi)^{-\frac{m}{2}} \det|\Sigma + \Sigma\Sigma_c\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}S^T(\Sigma + \Sigma\Sigma_c\Sigma)^{-1}S\} \tag{8}$$

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

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

where $S^T\Sigma^{-1}$ is independent from the causal status and can be computed once and used many times. As a result, the expensive computational part is to compute $(I_{m\times m} + \Sigma_c\Sigma)^{-1}$. We set elements of U and V 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:

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

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

$$(A + UEV)^{-1} = A^{-1} - A^{-1}U(E^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{11}$$

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

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

$$= I_{m\times m}^{-1} - I_{m\times m}^{-1}U(I_{k\times k}^{-1} + VI_{k\times k}^{-1}U)^{-1}VI_{n\times n} \tag{13}$$

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

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.

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

$$\det|I_{m\times m} + UV| = \det|I_{k\times k} + VU| \tag{15}$$

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:

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\det|\Sigma + \Sigma\Sigma_c\Sigma| = \det|\Sigma| \;\; \det|I_{m\times m} + \Sigma_c\Sigma| \tag{16}$$

$$= \det|\Sigma| \;\; \det|I_{m\times m} + UV| \tag{17}$$

$$= \det|\Sigma| \;\; \det|I_{k\times k} + VU| \tag{18}$$

In 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 $O(k^3)$.

3.5 Conditional Method (CM)

A standard method to detect allelic heterogeneity (AH) is the conditional method (CM). In CM, we identify the SNP with most significant association statistics. Then, conditioning on that SNP, we re-compute the marginal statistics of all the remaining variants in the locus. 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 fall below the predefined threshold. The predefined threshold is referred to as the stopping threshold for CM. 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 ith variant as follows:

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

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

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.6 Datasets

Genotype-Tissue Expression (GTEx): We obtained the summary statistics for GTEx eQTL dataset (Release v6p, dbGaP Accession 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, genes that have at least one significant eQTL, in order to detect loci that harbor allelic heterogeneity.

Genetic European Variation in Disease (GEUVADIS): We obtained the summary statistics of blood eQTL for 373 European individuals from the GEUVADIS website. We approximated LD structure from the 1000G CEU population. We applied our method to the 2954 eGenes in GEUVADIS to detect AH loci.

Multiple Tissue Human Expression Resource (MuTHER): We obtained the summary statistics from the MuTHER website. We utilized the skin and fat (adipose) tissues. We then approximated LD from the 1000G CEU population. We obtained 1433 eGenes for skin and 2769 eGenes for adipose.

High-Density Lipoprotein Cholesterol (HDL-C): We used the High-Density Lipoprotein Cholesterol (HDL-C) trait [21]. We only considered the GWAS hits, which are reported in a previous study. We applied ImpG-Summary [36] to impute the summary statistics with 1000G as the reference panel. We identified 37 loci that have at least one causal variant. Following common protocol in fine-mapping methods, we assumed at least one causal variant. Then, we applied our method to each locus.

Psychiatric diseases: We analyzed the recent GWAS on major depression disorder and schizophrenia. 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 CEU population.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.7 Data simulation

Simulated data with no epistasis interaction: We first simulated genotypes using HAPGEN2[28], where we utilized the 1000G CEU population as initial reference panels. Then, we simulated phenotypes using the Fisher's polygenic model, where the effects of causal variants are obtained from the normal distribution with a mean of zero. Let Y indicate the phenotypes and X indicate the standardized genotypes. In addition, $\boldsymbol{\beta}$ is the vector of effect sizes where β_i is the effect of the i-th variant. Thus, we have:

$$Y = X\boldsymbol{\beta} + e \tag{20}$$

, where e models the environment and measurement noise. Under the Fisher's polygenic model, the effect size of the causal variants is obtained from $N(0, (\sigma_g^2)/N_c)$, where N_c is the number of causal variants and σ_g is the genetic variation. In addition, the effect size for variants that are non-causal is zero. We set the effect size in order to obtain the desired statistical power. We implanted one, two, or three causal variants in our simulated datasets.

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.

Simulated data 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:

$$Y = x_i x_j \beta_{ij} + e \tag{21}$$

, where x_i and x_j indicate the standardized genotypes for ith and jth 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.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4 Discussion

We have proposed a novel probabilistic method to detect loci with AH. Our results show that our method is more accurate than existing methods. One of the main benefits of our method is that it only requires summary statistics. Summary statistics of a GWAS study or eQTL study are widely available; thus, 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. Similarly, attempts to explain GWAS using eQTLs data should be more successful with methods that assume that some loci may include multiple causative variants (e.g. eCAVIAR [37] and RTC [38]).

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 (LMM) [39–44] 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 work 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.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5 Acknowledgments

FH, JWJJ, and EE are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, 1320589 and 1331176, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-ES021801, R01-MH101782 and R01-ES022282. EE is supported in part by the NIH BD2K award, U54EB020403. AVS is supported by a contract (HHSN268201000029C) to the Laboratory, Data Analysis, and Coordinating Center (LDACC) at The Broad Institute, Inc. SS was supported in part by NIH grant R00-GM 111744-03. GK is supported by the Biomedical Big Data Training Program (NIH-NCI T32CA201160). We acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691).

6 Web Resources

CAVIAR is available `http://genetics.cs.ucla.edu/caviar/`

GTEx dataset (Release v6, dbGaP Accession phs000424.v6.p1) is available at `http://www.gtexportal.org`.

GEUVADIS dataset is available at `ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/`.

MuTHER dataset is available at `http://www.muther.ac.uk/Data.html`.

Blood eQTL dataset is available at `http://genenetwork.nl/bloodeqtlbrowser/`.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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*, 661–78.

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. (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. (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. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nature Genetics *41*, 703–7.

5. Zhang, F. and Lupski, J. R. (2015). Non-coding genetic variants in human disease: Figure 1. Hum. Mol. Genet. *24*, 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. The American Journal of Human Genetics *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*, 747–53.

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

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

association loci for human diseases and traits. Proceedings of the National Academy of Sciences *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. (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. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nature Genetics *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. (2012). Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. Human molecular genetics pp. dds098.

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 Genetics *10*, e1004383.

14. Estivill, X., Bancells, C., and Ramos, C. (1997). Geographic distribution and regional origin of 272 cystic fibrosis mutations in european populations. Human Mutation *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. Human Mutation *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.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics *44*, 369–375.

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. (2015). 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. (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature *523*, 588–591.

23. 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

24. 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

25. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B., and Eskin, E. (2015). Identification of causal genes for complex traits. Bioinformatics *31*, i206–i213.

26. Joo, J. W. J., Hormozdiari, F., Han, B., and Eskin, E. (2016). Multiple testing correction in linear mixed models. Genome Biology *17*, 62.

27. Han, B. and Eskin, E. (2012). Interpreting meta-analyses of genome-wide association studies. PLoS Genetics *8*, e1002555.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.

29. 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. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–8.

30. Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

31. 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. Nature Genetics *45*, 1238–1243.

32. Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., et al. (2011). The architecture of gene regulatory variation across multiple human tissues: The MuTHER study. PLoS Genetics *7*, e1002003.

33. Darnell, G., Duong, D., Han, B., and Eskin, E. (2012). Incorporating prior information into association studies. Bioinformatics *28*, i147–i153.

34. Eskin, E. (2008). Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. Genome Research *18*, 653–660.

35. 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.

36. 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.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

37. 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. The American Journal of Human Genetics.

38. 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 Genetics *6*, e1000895.

39. 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. Nature Genetics *42*, 348–54.

40. 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.

41. 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. Nature Methods *8*, 833–835.

42. 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. Nature Genetics *47*, 284–290.

43. Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nature Genetics *44*, 821–824.

44. Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2005). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics *38*, 203–208.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

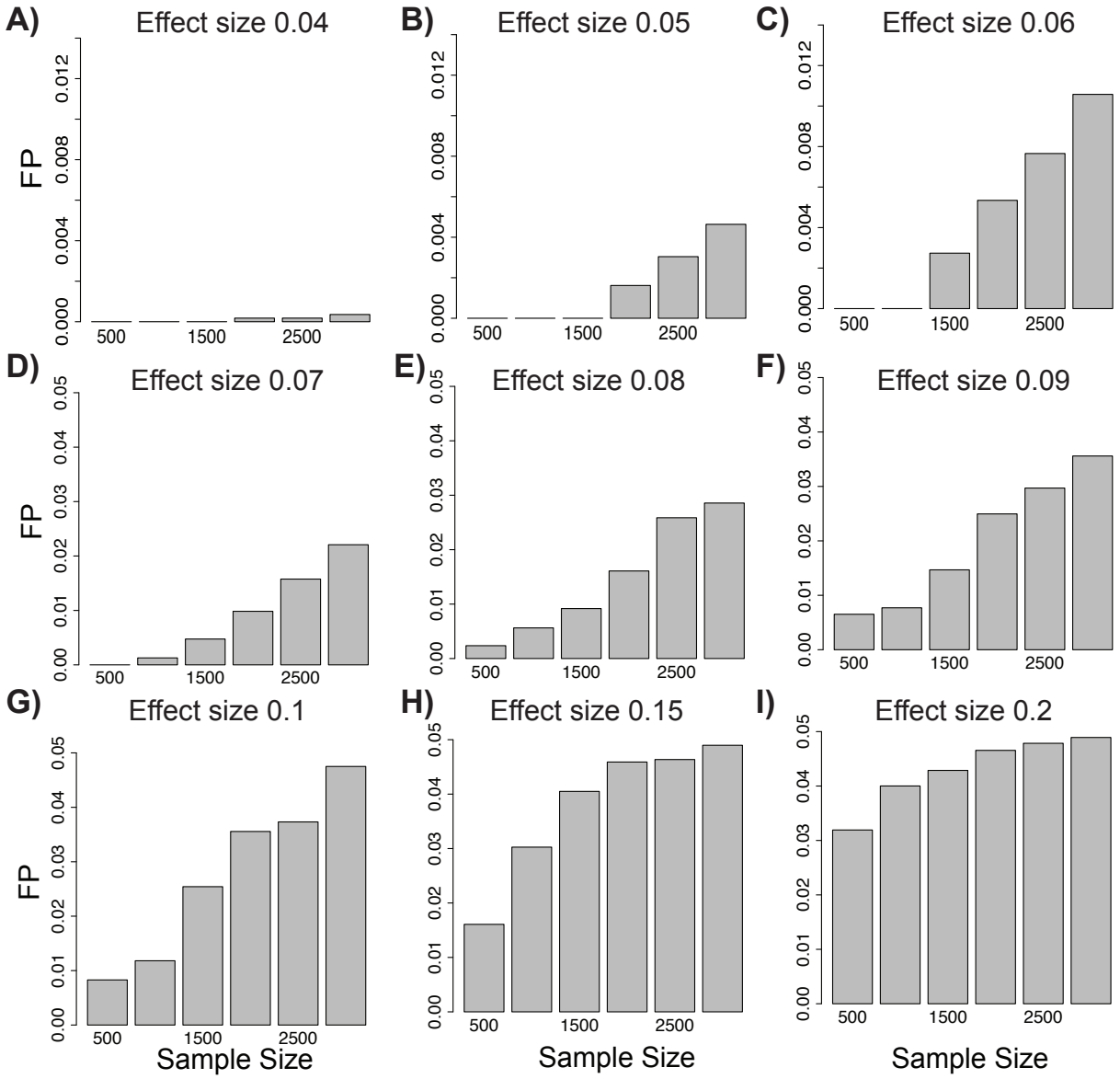

Figure 5. CAVIAR distinguishes between epistatic interaction and allelic heterogeneity. The x-axis is the sample size that we vary between 500, 1000, 1500, 2000, 2500, and 3000 individuals and the y-axis is the false positive (FP) rate. We simulated datasets where we have epistatic interaction and compute the FP as the number of cases where CAVIAR incorrectly detects these loci to harbor AH. (A-I) illustrate the FP for different effect sizes of the epistatic interaction.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

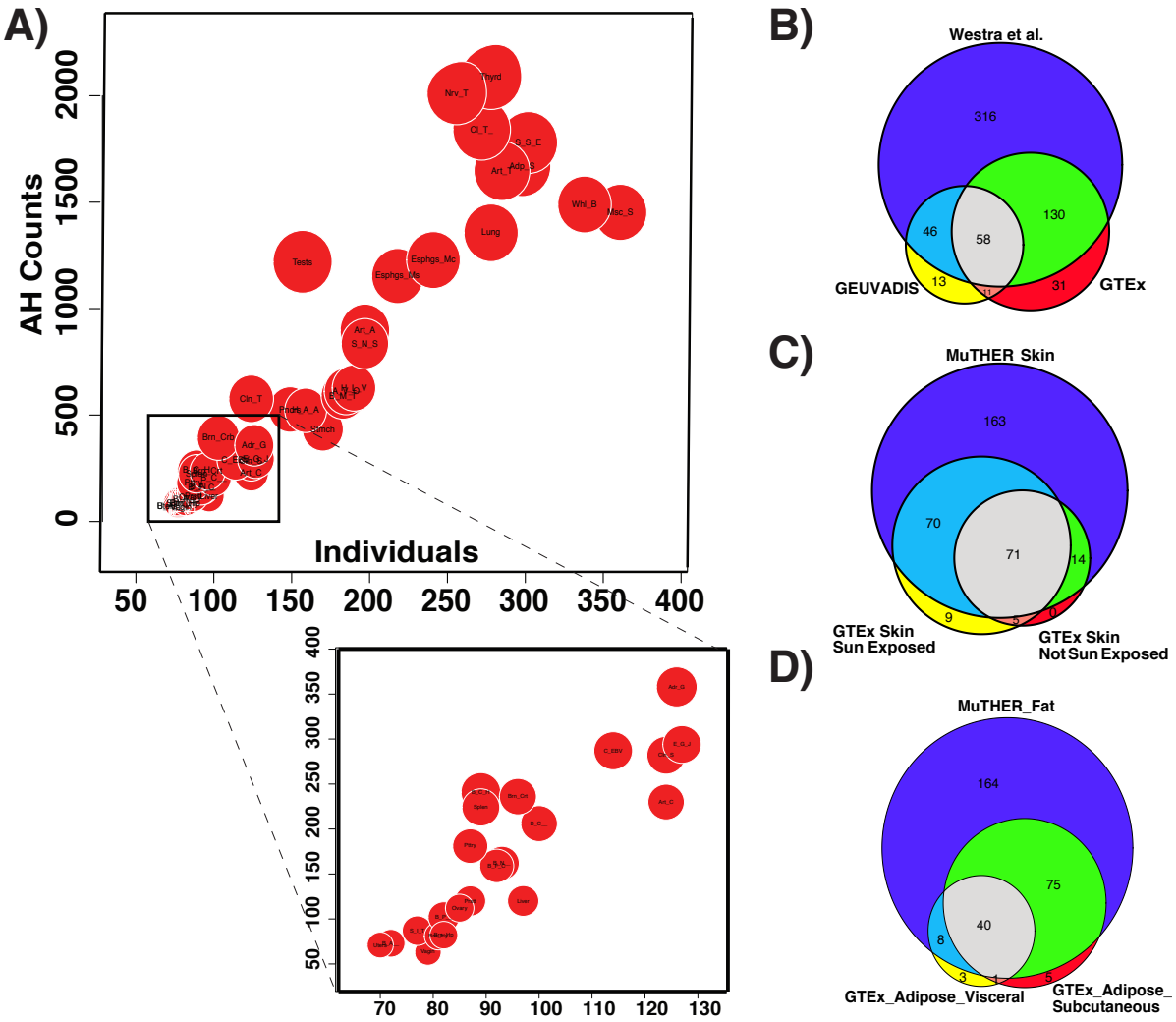

Figure 6. Levels of allelic heterogeneity in eQTL studies. (A) Linear relationship between the amount of AH and sample size. Each red circle indicates a different type of tissue from the GTEx dataset. The size of each red circle is proportional to the number of genes that harbor a significant eQTL (eGenes). (B-D) Significant overlap between AH estimations for different eQTL datasets, shown for (B) blood (P=7.9e-97), (C) skin (P=4.9e-63), and (D) adipose (P=1.1e-69).

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted December 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

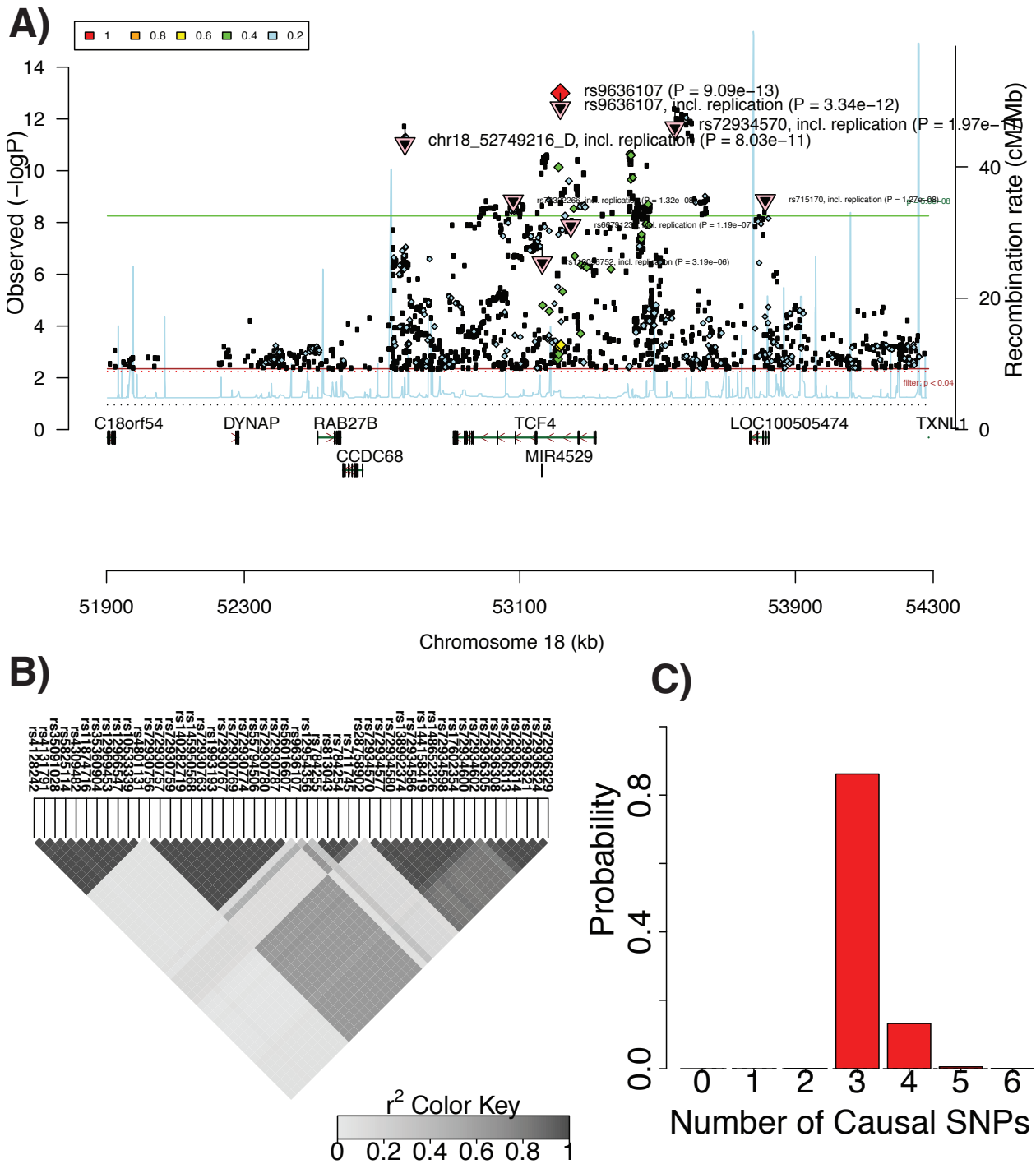

Figure 7. Allelic heterogeneity in the TCF4 locus associated with schizophrenia. (A) Manhattan plot obtained from Ricopili (http://data.broadinstitute.org/mpg/ricopili/) consists of all the variants in a 1Mbp window centered on the most significant SNP in the locus (rs9636107). This plot indicates multiple 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 of the estimated number of causal variants.