Leveraging allelic imbalance to refine fine-mapping for eQTL studies

Jennifer Zou[1☯], **Farhad Hormozdiari**[2,3], **Brandon Jew**[4], **Stephane E. Castel**[5,6], **Tuuli Lappalainen**[5,6], **Jason Ernst**[1,7], **Jae Hoon Sul**[8]*, **Eleazar Eskin**[1,9]*

1 Computer Science Department, University of California Los Angeles, Los Angeles, California, United States of America, **2** Genetic Epidemiology and Statistical Genetics Program, Harvard University, Cambridge, Massachusetts, United States of America, **3** Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, **4** Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, California, United States of America, **5** New York Genome Center, New York, New York, United States of America, **6** Department of Systems Biology, Columbia University, New York, New York, United States of America, **7** Department of Biological Chemistry, University of California Los Angeles, Los Angeles, California, United States of America, **8** Department of Psychiatry and Biobehavioral Sciences, University of California Los Angeles, Los Angeles, California, United States of America, **9** Department of Human Genetics, University of California Los Angeles, Los Angeles, California, United States of America

☯ These authors contributed equally to this work.
* JaeHoonSul@mednet.ucla.edu (JHS); eeskin@cs.ucla.edu (EE)

Abstract

Many disease risk loci identified in genome-wide association studies are present in non-coding regions of the genome. Previous studies have found enrichment of expression quantitative trait loci (eQTLs) in disease risk loci, indicating that identifying causal variants for gene expression is important for elucidating the genetic basis of not only gene expression but also complex traits. However, detecting causal variants is challenging due to complex genetic correlation among variants known as linkage disequilibrium (LD) and the presence of multiple causal variants within a locus. Although several fine-mapping approaches have been developed to overcome these challenges, they may produce large sets of putative causal variants when true causal variants are in high LD with many non-causal variants. In eQTL studies, there is an additional source of information that can be used to improve fine-mapping called allelic imbalance (AIM) that measures imbalance in gene expression on two chromosomes of a diploid organism. In this work, we develop a novel statistical method that leverages both AIM and total expression data to detect causal variants that regulate gene expression. We illustrate through simulations and application to 10 tissues of the Genotype-Tissue Expression (GTEx) dataset that our method identifies the true causal variants with higher specificity than an approach that uses only eQTL information. Across all tissues and genes, our method achieves a median reduction rate of 11% in the number of putative causal variants. We use chromatin state data from the Roadmap Epigenomics Consortium to show that the putative causal variants identified by our method are enriched for active regions of the genome, providing orthogonal support that our method identifies causal variants with increased specificity.

Author summary

In recent years, many studies have identified genetic variants that are associated with the expression of genes (eQTLs). While thousands of eQTLs have been identified, not all associated variants cause changes in gene expression. This is in part due to the complex patterns of genetic correlation in the human genome. If a region of the genome contains many genetic variants that are highly correlated with each other, non-causal genetic variants close to a causal variant are also correlated with gene expression. Statistical fine-mapping is the process of identifying true causal variants from a set of candidate variants. In regions with high genetic correlation, previous fine-mapping methods may not be able to differentiate causal variants from nearby variants. We propose a method that utilizes a complementary source of information called allelic imbalance (AIM). We show that by combining eQTL and AIM data, we can identify the true causal variants more efficiently and substantially decrease the number of putative causal variants for downstream analysis.

Introduction

Understanding the regulation of gene expression by genetic variants is essential for identifying the biological mechanisms of gene expression and complex traits. In expression quantitative trait loci (eQTL) studies, a statistical test is performed to find genetic variants that are significantly correlated with gene expression (eQTLs). Identifying eQTLs is critical in genetic studies not only because they influence gene expression but also because they are enriched in disease risk loci [1–5]. In recent years, many studies have identified eQTLs in different organisms and tissues [6–10].

Once eQTLs are identified, the next step is to isolate the causal variants that influence gene expression. We define causal variants to be variants that are responsible for the observed peak of association in a locus. Not all eQTLs are causal, and the process of identifying causal variants from several candidate variants is called fine-mapping. Fine-mapping in eQTL studies has two main challenges. The first challenge is the complex linkage disequilibrium (LD) structure present in the human genome. If a region of the genome contains many genetic variants that are highly correlated (or "in high LD") with each other, non-causal genetic variants close to a causal variant appear to be correlated with gene expression [11–13]. The second challenge is that there may be multiple causal variants in a region [14, 15], increasing the complexity of fine-mapping algorithms. A few fine-mapping approaches have been developed to address these challenges [16–22]. These methods attempt to calculate a posterior probability for each variant and select a set of variants called a "causal set" that contains all causal variants with high probability (e.g., 95%) while minimizing the number of variants in the causal set. Although these previous methods are accurate in including all true causal variants in the causal set, they may also include many non-causal variants in regions with high LD, which increases the number of variants that need to be validated in downstream studies.

In addition to total expression, there is another source of information called allelic imbalance (AIM) that we can utilize to identify genetic variants that regulate gene expression [23–28]. A diploid individual has AIM for a specific gene if the amount of gene expression from one chromosome is not equal to the amount of gene expression from the other chromosome [29, 30]. Because causal variants for AIM change expression levels of a gene in an allele-specific manner, they provide insight into the regulation of gene expression. This is different from eQTL mapping using total expression, which identifies genetic variants associated with a

significant difference in total expression. Thus, eQTL mapping and AIM analysis are complementary approaches to studying the genetic basis of gene expression, and combining these two types of data may improve statistical fine-mapping.

We propose a method that combines both total expression and AIM data to improve fine-mapping in eQTL studies. We first compute an eQTL summary statistic for total expression and an AIM summary statistic for allelic imbalance. We then aggregate the two statistics utilizing meta-analysis. The meta-analysis statistic is then incorporated into a fine-mapping approach for GWAS called CAVIAR [18], which models multiple causal variants and takes into account the LD structure among genetic variants. Our simulation results show that our method identifies causal variants with the correct true positive rate or recall, and more importantly, it generates causal sets that are noticeably smaller than sets generated using only the eQTL statistics. This means that our approach includes fewer non-casual variants, yielding higher specificity. We apply our method to RNA-seq data from ten tissues in the Genotype-Tissue Expression (GTEx) dataset [8, 10] and find that the causal set size is also reduced when compared to an approach that only uses eQTL statistics. Since the true causal variants for gene expression are unknown in most cases, we use orthogonal gene annotation and chromatin state data from the Roadmap Epigenomics Consortium to show that the putative causal variants identified by our method are enriched for active regions of the genome. We also show that variants that are predicted to be causal by using only eQTL statistics but not by our approach are depleted for active regions of the genome, providing orthogonal support that our method identifies causal variants with increased specificity.

Results

Method overview

The main goal of fine-mapping methods is to identify a small set of variants that include all true causal variants with high probability. A naive approach would be to conduct an eQTL mapping and to use these summary statistics as input to fine-mapping algorithms for GWAS such as CAVIAR [18]. Our method improves this fine-mapping algorithm by including AIM summary statistic data, which identifies variants associated with changes in AIM. This is different from eQTL mapping, which identifies genetic variants associated with a significant difference in total expression. Thus, the AIM analysis may provide additional information on which variants are causal, which is not captured in the eQTL mapping.

In terms of a statistic for each analysis, an eQTL mapping computes a statistic that measures correlation between total expression levels of a gene and genotypes of a variant. On the other hand, an AIM analysis computes a statistic that measures correlation between AIM status (e.g., 0 for balanced expression and 1 for AIM) and heterozygous status of a variant (e.g., 0 for homozygous genotype and 1 for heterozygous genotype). The AIM status can only be computed in individuals with heterozygous variants in the coding region of the gene. Thus, only individuals with heterozygous variants in the coding region are used to calculate the AIM statistic for a variant. We refer to the heterozygous coding variants used to call AIM as aiSNPs and the variants tested for association as cis-regulatory variants or cis-eQTLs. While it is possible for nearby variants to have a trans-regulatory effect, we focus on cis-regulatory variants that have a direct effect on the expression of the gene being tested [31, 32]. These variants are more likely to have higher agreement between eQTL statistics and AIM statistics. For example, in Fig 1A, individual 2 has a heterozygous genotype at the causal cis-regulatory variant and has AIM for the gene, while other individuals with homozygous genotypes have balanced expression. The causal cis-regulatory variant has heterozygosity patterns that perfectly match the
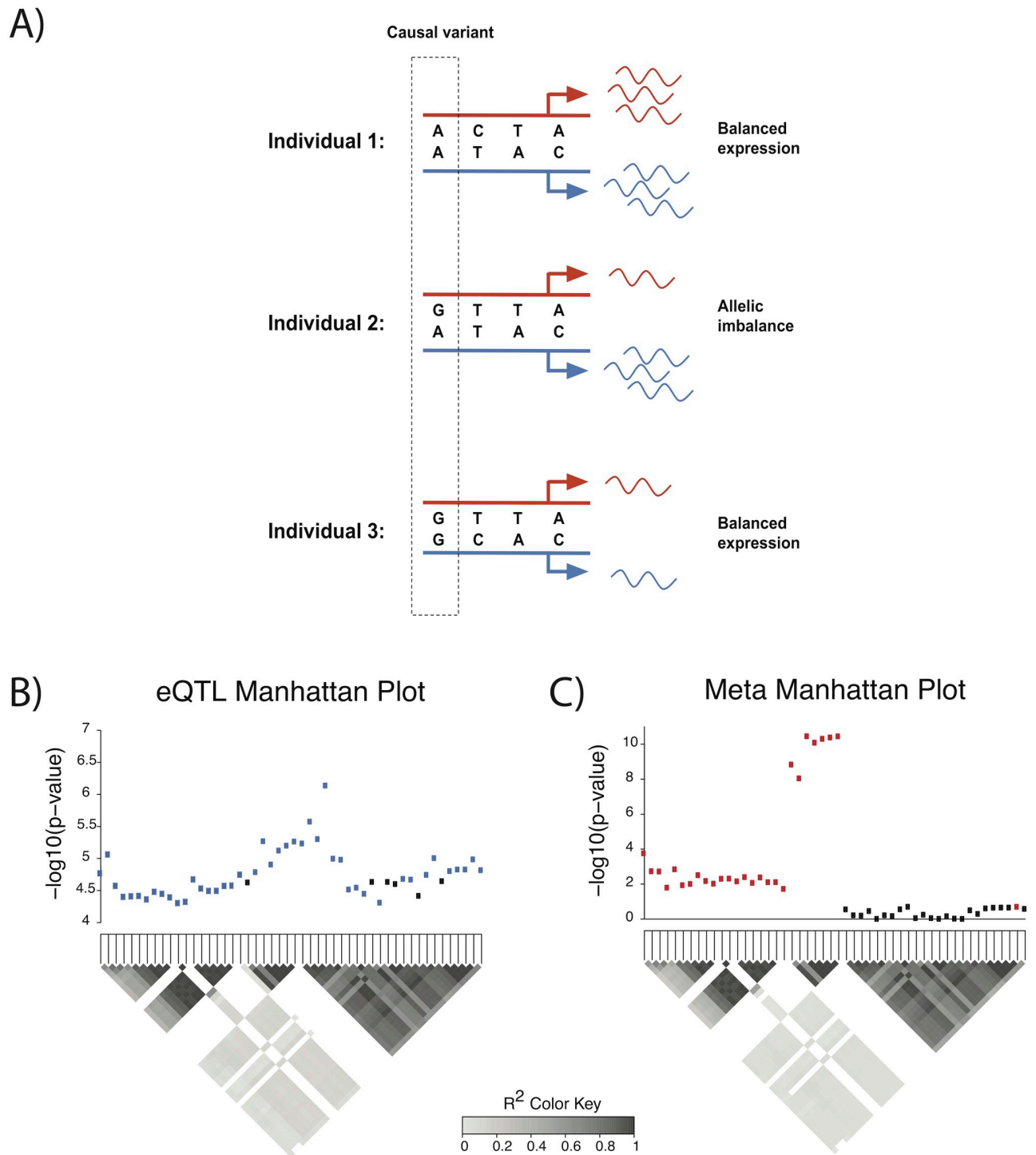
Fig 1. Integrating AIM and eQTL summary statistics improves fine-mapping of cis-regulatory variants. A) Individuals 1 and 3 have balanced expression, and individual 2 has AIM. The first variant has a heterozygosity pattern across individuals that perfectly matches the AIM status across individuals, making it more likely to regulate gene expression than other cis-regulatory variants shown. B and C) Manhattan plots for gene TPO showing -log_{10} p-values for eQTL and meta-statistics summary statistics, respectively. The heatmaps below each plot show the pairwise correlations due to LD. The blue points correspond to variants included in the causal set obtained using only eQTL summary statistics, and the red points correspond to variants included in the causal set obtained using the meta-statistics.

https://doi.org/10.1371/journal.pgen.1008481.g001

AIM patterns, and the AIM statistic for this variant is higher than any of the other non-causal variants. We compute the AIM statistic as discussed in the Methods section.

Because our AIM statistic measures cis-regulatory strength of variants in a complementary way to eQTL studies, it can be used to improve traditional eQTL fine-mapping studies. For example, a certain gene may contain many highly correlated genetic variants with low p-values using eQTL mapping, which makes it difficult to pinpoint causal variants. In Fig 1B, there are 50 significant eQTLs for gene TPO. While it is difficult to isolate the true causal variants using total expression data alone, AIM statistics may provide more information on which of the 50 variants are more likely to be causal. Fig 1C shows the corresponding meta-analyzed summary statistics ("meta-statistics"), which combine the eQTL and AIM summary statistics. There is a block of variants on the right of the Manhattan plot that have substantially lower meta-statistics than eQTL statistics. Using only the eQTL summary statistics we obtain a causal set of 44 variants. However, when we combine both eQTL and AIM information, we obtain a causal set size of 27, which is a 39% reduction in set size. Hence, by combining eQTL and AIM statistics, we can potentially deprioritize non-causal variants in high LD regions and detect the true causal variants more accurately.

When performing fine-mapping with the meta-statistics, we assume that causal variants in the eQTL mapping also cause changes in AIM status. This assumption is likely to be valid under an additive model for genetic effects when the causal variant is a cis-regulatory variant. Under these assumptions, an allele that increases the total expression level of a gene also causes individuals with heterozygous genotypes to have different expression levels from each chromosome (AIM). With this assumption, we combine eQTL and AIM summary statistics of variants in close proximity to a gene by performing meta-analysis on the two statistics. Meta-analysis is a popular statistical approach that combines results of multiple studies to increase statistical power. In GWAS, meta-analysis has discovered many associations that were not identified by each individual study [33, 34]. We perform fixed effect meta-analysis between the AIM and eQTL statistics to better identify causal cis-eQTLs that influence gene expression. We utilize fixed effect meta-analysis, rather than random effects meta-analysis because we only have two data sets. While random effects models could potentially be applied, these approaches are most effective when there are more than two studies in consideration [35, 36]. While traditional applications of fixed effect meta-analysis have the same phenotype and rely on independent samples to achieve independence between the statistics, our framework uses two orthogonal types of data: total expression and allelic imbalance [37, 38]. By using these two types of data, we have an additional issue not present in traditional meta-analysis. Unlike traditional meta-analysis, the dependent and independent variables for the two studies are different, and the two statistics may have conflicting signs. This is not ideal because we want to fine-map variants that have high likelihood to have cis-regulatory impact, regardless of sign. To address this, we compute two forms of the meta-statistic ($S^{M1} = \frac{S^A + S^E}{\sqrt{2}}$ and $S^{M2} = \frac{S^A - S^E}{\sqrt{2}}$) for each gene to use as input to our fine-mapping framework (Methods, S6 Fig).

To improve identification of causal variants from these statistics, we extend the previous fine-mapping framework for GWAS called CAVIAR [18]. CAVIAR takes as input a set of summary statistics for variants in a locus and a pairwise correlation matrix between these variants calculated from genotype data. Using this data, CAVIAR identifies a minimal subset of variants that contains all causal variants with probability ρ. We refer to this set as the ρ causal set, and this can be interpreted as a confidence interval on the set. We extend the CAVIAR framework to incorporate both eQTL and AIM summary statistic data by utilizing fixed effect meta-analysis. By integrating these two types of data, our method attempts to reduce the size of the causal set, increasing specificity compared to the traditional CAVIAR approach that

uses only total expression data. In practice, we apply our fine-mapping framework to two forms of the meta-statistic for each gene and select the causal set that is smaller (Methods).

## Meta-analysis achieves correct recall rate

We generate simulated data to assess the performance of our approach. In this simulation, we use real genotypes from the whole blood tissue in the GTEx dataset that contains 325 samples. For each gene, we identify the top 50 genetic variants with the highest eQTL statistics. We use these top 50 genetic variants to calculate pairwise correlation matrices as described in the Methods section. We then choose one causal variant randomly from the 50 variants and assign an effect size to this variant such that we have 50% power to detect its association. Given this effect size for the causal variant, we generate eQTL statistics for all variants by using their genotype correlation (LD) to the causal variant and by assuming that the statistics follow a multivariate normal distribution (MVN). Using the same effect size and implanted causal variant, we generate AIM statistics for all variants by using their heterozygosity correlation to the causal variant and the same MVN assumption. We use fixed effect meta-analysis to combine these simulated statistics into meta-statistics, which we use as input to our fine-mapping framework. For comparison, we apply the CAVIAR framework using only the simulated eQTL statistics. We also generate simulations where the number of causal variants is 2 and 3 and the power to detect a causal variant is 80%. We use 1,000 randomly chosen genes for these simulations.

One important metric to measure the performance of fine-mapping approaches is determining how accurately they identify implanted causal variants. For this measure, we define recall rate to be the proportion of genes in which all true causal variants are included in the causal set generated from the fine-mapping approaches. Since our method and the original CAVIAR method are designed to include all causal variants with 95% probability in the causal sets, the recall rate should be close to 95%. For all combinations of power and number of causal variants, simulation results show that our method achieves very similar recall rates compared to the original CAVIAR method (S1 and S2 Tables) that has been shown to have well-calibrated recall rates compared to competing fine-mapping algorithms when there are multiple causal variants [18]. These results demonstrate that our method also detects all true causal variants accurately even when there are several causal variants near a gene and when the power to detect a causal variant is not very high.

Meta-analysis reduces the size of causal set

A naive approach to improve recall rate would be to include as many variants as possible in the causal set. For instance, if we include all variants near a gene in the causal set, the recall rate would always be 1, as the causal set is guaranteed to contain all causal variants. This, however, would not be cost effective because downstream validation of these variants using functional assays is often costly. Therefore, the size of the causal set needs to be minimized while retaining high recall rate. While previous fine-mapping approaches such as CAVIAR attempt to minimize the causal set size, they may not accurately differentiate between causal and non-causal variants if they are in regions of high LD. Our method incorporates additional information about causal variants, AIM statistics, into our fine-mapping framework to improve our ability to determine whether variants are causal or not.

We compare the size of the causal set between our approach and the original CAVIAR approach in the following manner. For each gene, we calculate the *reduction rate* as $r = (N_{eQTL} - N_{Meta})/N_{eQTL}$ where N_{Meta} is the size of the causal set from our meta-analysis approach and N_{eQTL} is the size of the causal set from the original CAVIAR approach using only eQTL
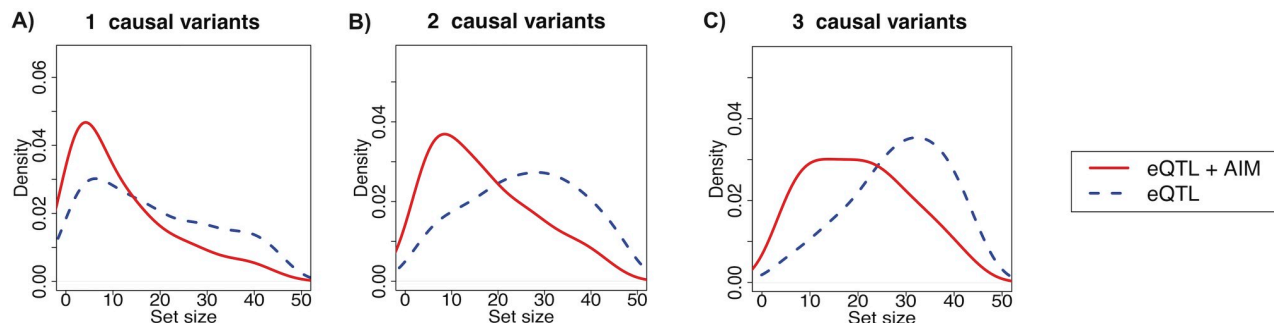
Fig 2. Distribution of causal set sizes in simulations. The power to detect causal variants is 50% in all scenarios. "eQTL+AIM" shows the distribution of causal set sizes using our approach that uses meta-statistics, while "eQTL" shows the distribution of causal set sizes using the naive approach that uses eQTL statistics. We consider three scenarios where we have A) one causal variant, B) two causal variants, and C) three causal variants.

https://doi.org/10.1371/journal.pgen.1008481.g002

statistics. A positive reduction rate means our approach has the smaller causal set size than the original CAVIAR approach, while a negative reduction rate means it has the larger causal set. For each combination of power level and number of causal variants, we compute the median reduction rate across all genes.

Using the same simulation data as in the previous recall rate simulation, we show that the size of the causal set from our approach is noticeably smaller than that from the CAVIAR approach that uses only eQTL statistics (Fig 2). The median reduction rate is 24%, 31%, and 27% when the number of causal variants is 1, 2, and 3, respectively at 50% power to detect causal variants. At 80% power to detect causal variants, the median reduction rate is 27%, 31%, and 29% when the number of causal variants is 1, 2, and 3, respectively. This result also shows that we can achieve high reduction rates when there are multiple causal variants in a locus (Fig 2). This is advantageous because it is harder to predict causal variants when there are several causal variants that may have high LD with other non-causal variants. By incorporating information about causal variants for AIM, our framework is able to exclude non-causal variants from the causal set, reducing the causal set size.

Different effect sizes yield modest decrease in reduction rate

As our method utilizes fixed effect meta-analysis to combine eQTL and AIM statistics, it assumes that effect sizes in the eQTL studies and AIM studies are the same. To measure the performance of our approach when this assumption is violated, we consider scenarios where the true effect size in eQTL studies is not equal to that in AIM studies. Using the same 1000 randomly chosen genes from the previous simulations, we fix the effect size of eQTL causal variants to 5.2 (50% power) and vary the effect size of AIM causal variants from 2.0 to 4.0. We assume that AIM has smaller effect size than eQTL because in real data, incorrect mapping of reads, inaccurate phasing, reads overlapping multiple intronic variants, or other technical issues may deteriorate AIM signals and reduce our power to detect effect of AIM [39–41]. Although we decrease the effect size of the AIM summary statistics in these simulations, the argument is symmetric, and decreasing the eQTL statistics would have similar results. We generate AIM summary statistics and eQTL summary statistics using the same simulation framework as previously described.

We observe that the median reduction rate decreases as the fixed effect assumption is further violated, which is expected (Fig 3). The results also show that as long as AIM effect sizes are not prohibitively low compared to the eQTL effect sizes, our framework can yield a positive reduction rate. For all numbers of causal variants, the reduction rate is positive when the AIM
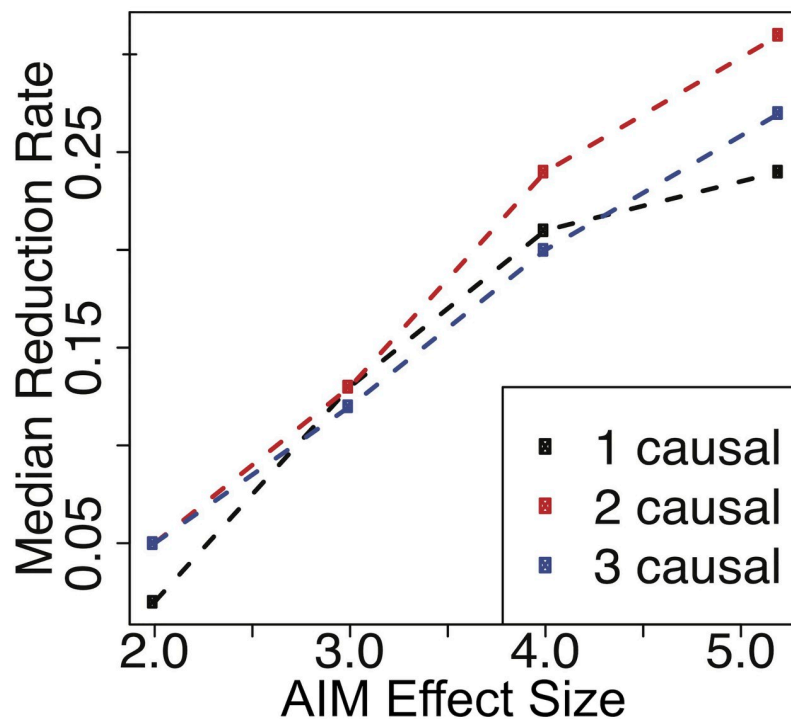
Fig 3. Median reduction rate using simulated data with different effect sizes between eQTL and AIM statistics.
eQTL effect sizes are fixed at 5.2 (50% power), while AIM effect sizes range from 2.0 to 5.2. Median reduction rates are
shown for 1, 2, and 3 causal variants.

https://doi.org/10.1371/journal.pgen.1008481.g003

effect size is greater than 2.0 (38% of eQTL effect size). In real data, although true effect sizes of
eQTL and AIM cis-regulatory variants are unknown, it is unlikely to observe such a large dis-
crepancy between the two association statistics. This also means that we expect slightly reduced
reduction rate in real data compared to those observed in simulation, as there may be some
genes in which effect sizes of eQTL and AIM are not exactly the same. As long as the difference
in effect sizes is not substantial, combining AIM and total expression information is still bene-
ficial and the number of putative causal variants can be decreased.

We observe that that the recall is generally high, although slightly lower in the most extreme
cases (S3 Table). It is possible that this reduced recall is due to the prior we set on the true effect
sizes. We compared our recall for these simulations with the recall obtained using an alterna-
tive approach used by PAINTOR 3.0, where the prior is estimated from the data using Empiri-
cal Bayes. This approach did not yield substantially better recall in the most extreme cases and
had notably worse performance when there were multiple causal variants (S4 Table). Never-
theless, both cases had high recall rate in most cases. Based on these results and prior work
[42], we expect that the prior does not substantially impact the final results, and the main effect
of unequal effect sizes is the decrease in efficiency of fine-mapping.

Meta-analysis improves fine-mapping in GTEx data

We then apply our approach to RNA-seq data from the GTEx Consortium [10]. We apply the
framework to ten randomly selected tissues, ranging in sample sizes from 101 to 491. Unlike
simulated data, real AIM data is not perfect, as it contains mapping errors and non-negligible
noises that may reduce AIM signals significantly for certain genes. In particular, we observe

that variants with small numbers of reads have low concordance of allelic ratios between variants used to call AIM (S1 Fig). Genes with low read counts may contain very low AIM statistics for genetic variants even though those genetic variants have high eQTL statistics. As shown in simulations, our method may not be accurate when the fixed effect assumption is violated. Thus, we require individuals to have at least 20 reads mapped to each gene. While this requirement increases the quality of AIM calls for each individual, it may reduce the number of genes for which this method can be applied (Table 1). We choose a conservative threshold for the minimum number of reads that still allows us to apply this framework to a large number of genes.

We set the recall rate of our method and the original CAVIAR approach to 95%, meaning that both approaches generate causal sets that include all causal variants with 95% probability. Our results in the GTEx dataset show that our approach is effective at identifying smaller causal sets compared to the original CAVIAR approach, which only uses total expression. The median reduction rate across all tissues and genes is 11%, and the median reduction rate across genes for each tissue ranges from 9% to 11% (Table 1). We report the median reduction rate to show that our method can provide some modest improvement in most cases, but in optimal cases, our method can reduce the causal set size substantially (S3 Fig). Our method achieves similar reduction rates for all tissues, even those with small sample sizes. Across tissues, the reduction rate is positive for 80% of genes, which indicates that our approach is able to reduce the size of the causal set for a majority of genes. This also implies that although the AIM data can be more noisy than total expression data in some cases, it still provides valuable, complementary signal that can be used to improve fine-mapping. The distribution of causal set size using the two sets of statistics is shown in Fig 4A.

In this analysis, we identify AIM status of each individual by binarizing the AIM data with an empirically determined threshold that corresponds to two standard deviations from the null allelic ratio of 0.5 (Methods; S2 Fig). We compare this binarization strategy with one in which the AIM status is determined using a binomial test. When using the adipose subcutaneous data and the binomial binarization in our fine-mapping framework, we observe a slight decrease in median reduction rate (from 11% to 7.2%). Furthermore, the number of genes for which the fine-mapping can be applied decreases from 4023 to 420. This decrease is due to the fact that applying a two-sided binomial test to binarize AIM status is a more stringent criteria than using the empirically determined threshold. When using the binomial threshold, many genes have no individuals with AIM. We are unable to calculate AIM summary statistics and perform fine-mapping for these genes. Therefore, we recommend using the empirically determined threshold to binarize AIM status.

Table 1. Reduction rate for 10 GTEx tissues. Let $\bar{r}$ be the median reduction rate, N be the sample size, and M be the number of genes tested.

Tissue	$\bar{r}$	N	M
Adipose Subcutaneous	0.11	385	4023
Artery Aorta	0.11	267	2842
Brain Caudate Basal Ganglia	0.11	144	1272
Brain Hippocampus	0.10	111	680
Heart Atrial Appendage	0.11	264	1907
Cells EBV-transformed Lymphocytes	0.09	117	470
Muscle Skeletal	0.11	491	3414
Thyroid	0.10	399	1470
Uterus	0.09	101	600
Whole Blood	0.11	369	3452

https://doi.org/10.1371/journal.pgen.1008481.t001

A)

B)
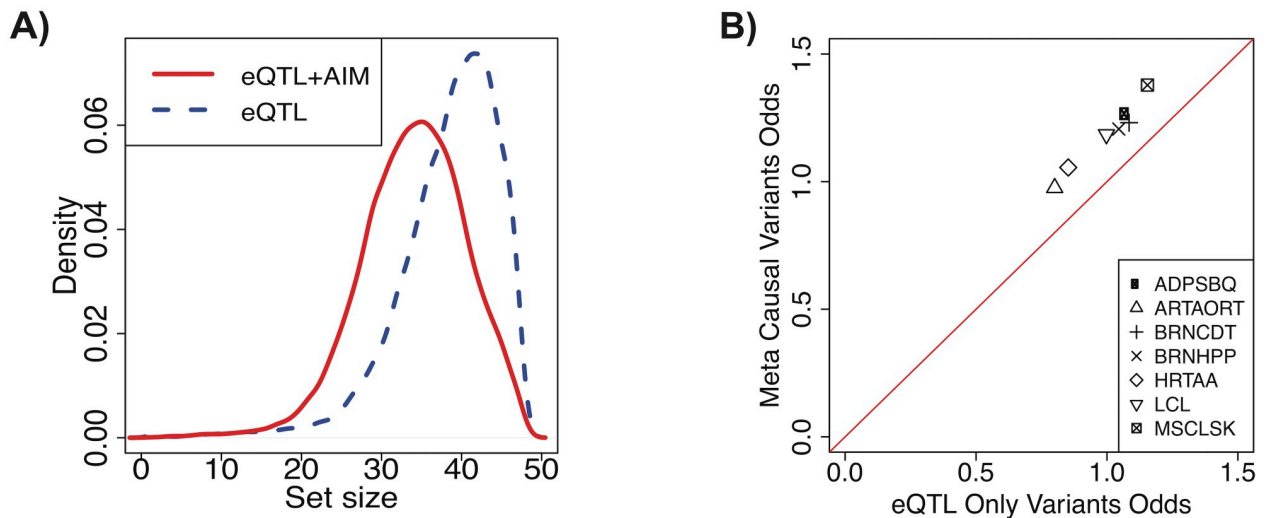
Fig 4. Integrating AIM and eQTL statistics improves specificity of fine-mapping in GTEx data. A) The distribution of causal set sizes from the GTEx data for our approach using meta-statistics ("eQTL+AIM") and the previous approach using only eQTL statistics ("eQTL"). B) Enrichment for active regions of the genome was computed in a tissue-specific manner. For each tissue, the "meta causal variants" includes all variants in the meta causal sets for all genes tested. The "eQTL only variants" includes all variants contained in the eQTL causal sets but excluded from the meta causal sets. The x-axis shows the odds of being in an active region of the genome for the eQTL only causal variants, and the y-axis shows the odds of being in an active region of the genome for the meta causal variants. Using a Fisher's exact test, the odds are significantly greater for the meta causal variants in all tissues. The tissue abbreviations are for Adipose Subcutaneous, Artery Aorta, Brain Caudate Basal Ganglia, Brain Hippocampus, Heart Atrial Appendage, Cells EBV-transformed Lymphocytes, and Muscle Skeletal respectively.

To assess the consistency of the eQTL and AIM summary statistics, we perform colocalization analysis between the two sets of statistics in the adipose subcutaneous tissue. Colocalization determines whether a single variant is responsible for the observed patterns of signal in two sets of summary statistics. We use colocalization to assess the agreement between the AIM and eQTL statistics since only the causal variants must agree for our framework to improve fine-mapping. In genes with some variants with high colocalization between the AIM and eQTL statistics, our method can potentially prioritize the causal variants and deprioritize the noncausal variants that are less likely to colocalize. We compute a colocalization posterior probability (CLPP) for each variant using the eCAVIAR method [43]. The CLPP corresponds to the amount of support for a variant being responsible for the associated signals in both studies. We observe a high Spearman correlation between the maximum CLPP for a gene and the reduction rate ($\rho = 0.31$), which indicates a strong correlation between colocalization of causal variants and fine-mapping performance. This also confirms that our method performs best under a model of additive cis-regulatory effects where the causal variants between total expression and AIM are the same. For each gene, we select the variant with the largest CLPP and determine whether it passes a colocalization threshold of 0.01, which is the threshold recommended by the eCAVIAR method. Of the genes we analyzed, we identify 1331/4023 genes with at least one variant meeting this colocalization threshold. These results imply that for these genes, the effect sizes are similar between the causal variants for AIM and total expression. When we consider only the 1331 genes with at least one colocalized variant, we achieve a slightly improved median reduction rate of 16%.

Variants in causal sets are enriched for active regions of the genome

We have shown statistically that our framework reduces the causal set size compared to the method that only uses total expression data. To further validate whether the reduction in the

causal set size is biologically relevant, we use orthogonal data sets that have finer resolution than typical LD blocks, such as gene annotations and chromatin state annotations. Since most variants input into our fine-mapping framework are in active regions of the genome to begin with, we test not only for enrichment in the meta causal set but also compared to the eQTL causal set. This is a more stringent test that explicitly shows improvement in specificity.

The first annotation we examine is the distance to the transcription start site (TSS) of a gene for causal variants. It has been observed that eQTLs with the strongest effect sizes are highly enriched near the TSS [44]. This observation may be due to the fact that cis-regulatory variants may potentially disrupt promoter or enhancer DNA regions, mediating a change in gene expression. Therefore, we expect that true causal cis-regulatory variants will be closer to the TSS than non-causal variants. We look at the distance to the TSS of variants in two causal sets; 1) variants in the causal set from our meta-analysis framework based on AIM and eQTL information ("meta causal set"), and 2) variants that are not present in our meta causal set, but present in the causal set from the original approach that uses only eQTL information ("eQTL only causal set"). We find that variants in the eQTL only causal set are significantly farther from the TSS than variants in the meta causal set across all genes and tissues (Mann U Whitney, $p < 2.2e - 16$, S4 Fig). This indicates that variants that are excluded from the meta causal set but present in the causal set using eQTL information have weaker evidence of being true causal variants than variants in the meta causal set. While the tendency to see variants falling closer to the TSS could be explained by simply choosing variants with higher effect sizes, our method does not merely choose variants with the highest effect sizes. We created an alternative set of variants for each gene in the adipose subcutaneous tissue constructed from the top eQTLs with the same size as the corresponding meta causal sets. We call these sets the TopN sets. We then computed the number of variants in the causal set that were not present in the TopN set. The majority of genes have multiple putative causal variants that are not in the TopN sets (S5 Fig). Thus, our approach does not decrease the causal set size by simply discarding weaker regulatory variants.

We also compute enrichment in chromatin states, which are combinatorial patterns of epigenetic marks that have been shown to be associated with distinct biological functions [45]. Causal variants are more likely to be within active chromatin states. We define active chromatin states to be regions of the genome that are annotated as promoters, enhancers, transcribed regions, open chromatin, or zinc finger genes by the ChromHMM software [46, 47]. All other regions of the genome are labeled as non-active regions of the genome. We obtain these annotations for 7 tissues that overlapped the Roadmap Epigenomics Project (Adipose Subcutaneous, Artery Aorta, Brain Caudate Basal Ganglia, Brain, Hippocampus, Heart Atrial Appendage, Cells, EBV-transformed Lymphocytes, Muscle Skeletal) and overlap the variants for each tissue with cell-type specific chromatin state annotations. We again compare variants in the meta causal set with those in the eQTL only causal set, and perform a Fisher's exact test, where the null hypothesis is that variants in the two causal sets are equally likely to fall within active regions of the genome. In these 7 tissues, we find that the variants in the meta causal set are significantly enriched for active regions compared to those in the eQTL only causal set (Fig 4b, S5 Table). This enrichment for active regions of the genome in the meta causal set suggests that using both AIM and total expression data increases the specificity of statistical fine-mapping.

In both the TSS and chromatin state analyses, we compare the meta causal set with the eQTL only causal set to show that the variants excluded from the meta causal set are likely non-causal. We also perform two additional comparisons. The first is a direct comparison between the eQTL causal sets and the meta causal sets. These two sets of variants contain many overlapping variants, and therefore, the differences are smaller than the previous comparison. Variants in the meta causal set are still significantly closer to the TSS than those in the

eQTL causal set (Mann U Whitney, $p < 2.2e - 16$). Using the same chromatin state annotations, we perform a Fisher's exact test, where the null hypothesis is that variants in the meta causal set and the eQTL causal set are equally likely to fall within an active region of the genome. Across the 7 tissues, the variants in the meta causal set are more likely to fall in active regions of the genome compared to those in the eQTL causal set, but this result is only significant in 3 of the tissues due to substantial overlap of variants in the causal sets (S7 Table). The second comparison is between the set of variants only found in the meta causal set ("meta only causal set") and the set of variants only found in the eQTL causal set ("eQTL only causal set"). We expect that the variants found in the meta only causal set are more likely to be causal than those in the eQTL only causal set, since the AIM information may provide evidence for additional causal variants that were not identified using only total expression information. Variants in the meta only causal set are significantly closer to the TSS than those in the eQTL only causal set (Mann U Whitney, $p < 2.2e - 16$). In the chromatin state analysis, the odds ratio is greater than one in 6 out of 7 tissues, indicating that the variants in the meta only set are slightly enriched for being in active regions of the genome in these tissues (S6 Table). However, due to the small number of variants in the meta only set, this result is not significant in any of the tissues. The small number of variants in the meta only causal sets also indicates that there are relatively few cases where the meta analysis implicates a variant that was not previously supported by the total expression data alone.

Comparisons with alternative AIM statistics and fine-mapping frameworks

Our method is unique in that it jointly models many variants within a single locus and outputs a minimal causal set that contains all causal variants with high probability. Other methods that combine total expression with AIM data to find cis-regulatory variants perform a marginal test on each SNP and output a likelihood ratio statistic for each variant tested [37, 38, 48]. This is a distinct type of output from the causal sets obtained in our method. In regions of high LD, these existing methods may identify many variants that are significantly associated with gene expression. One way to obtain a minimal causal set from existing statistics is to use these statistics as input to the CAVIAR fine-mapping framework. To test whether existing likelihood ratio statistics can be used in our framework, we compute statistics for cis-regulatory variants using RASQUAL and TReCASE, which jointly model total expression using a negative binomial distribution and allelic imbalance using a beta binomial distribution [37]. These methods perform a likelihood ratio test for each cis-regulatory variant, which outputs a chi-squared statistic. We call the statistics obtained from these joint likelihood-based methods "joint statistics." We compute joint statistics using the RASQUAL and TReCASE methods for the 4023 genes we tested in the adipose subcutaneous tissue. We only apply our fine-mapping framework to genes with joint chi-squared statistics for all variants tested in our framework. This is necessary because the reduction rates cannot be compared if the number of variants used as input is different between the methods. For RASQUAL, we test 2831 genes that meet this requirement. For TReCASE, we also require that all genes have a chi-squared statistic for the TReC (total expression) and ASE models, which leaves 649 genes (S1 Text). We convert the joint chi-squared statistics into z-scores, which we use as input to the CAVIAR fine-mapping framework. When comparing the causal set sizes obtained using the joint statistics to causal set sizes using only eQTL summary statistics, RASQUAL and TReCASE have median reduction rates of 0% and -15%, respectively. This is substantially lower than the median reduction rate of 11% using our method for the same subsets of genes. For TReCASE, we performed an additional comparison with fine-mapping using the TReC (eQTL) chi-squared statistic, which is output from the software. This

comparison more directly highlights the change in fine-mapping when including AIM data, rather than comparing the differences in statistics used as input. Using this second baseline, TReCASE obtains a median reduction rate of 0%.

Although there is little work on fine-mapping using allelic imbalance data, there are many alternative fine-mapping algorithms for GWAS data and eQTL data. Two alternative fine-mapping frameworks that could be applied in this context are DAP-G and PAINTOR 3.0. We apply these two fine-mapping frameworks to the 4023 genes we reported in the adipose subcutaneous tissue. DAP-G has a similar methodology as CAVIAR but approximates the posterior by only considering a subset of causal status configurations that contribute highly to the posterior. DAP-G also has the option of inputting a prior probability for each variant. This method can be utilized in a two-step fine-mapping approach. We apply the DAP-G method to our AIM statistics to obtain a posterior probability for each variant. We then use the posterior probabilities as a prior for fine-mapping the eQTL summary statistics using DAP-G. This approach obtains a median reduction rate of 0%. However, the number of models used to approximate the posterior decreased on average by 96% compared to fine-mapping the eQTL summary statistics using DAP-G without a prior. Another fine-mapping method that can be applied is PAINTOR 3.0, which was originally intended for fine-mapping with multiple ancestries. PAINTOR 3.0 jointly models two sets of summary statistics, treating each set as conditionally independent given the causal variants. While our approach should theoretically be a close approximation of this joint model (Methods section), the median reduction rate when jointly modeling the two statistics is -4.5%, indicating an increase in causal set size.

Discussion

We developed a novel fine-mapping approach for eQTL studies that utilizes AIM information from RNA-seq data. Based on the insight that causal variants for total gene expression also influence gene expression in an allele-specific manner, we developed a new statistic that aggregates AIM and eQTL statistics through fixed effect meta-analysis. We then incorporated the meta-analysis statistic into the fine-mapping algorithm for GWAS called CAVIAR, which takes into account LD structure among variants and multiple causal variants in a region. We used simulations to show that our approach achieves the correct recall rate and reduces the causal set size considerably compared to the fine-mapping approach that uses only eQTL statistics. We also showed through simulations that our method is effective at reducing causal set size even when the fixed effect assumption is not met. When applying our method to the GTEx data set, we observed a median reduction rate of 11% in causal set size across the ten tissues studied. Furthermore, the variants included in the causal set from our approach were enriched for active regions of the genome compared to those excluded from the causal set. These results demonstrated that eQTL and AIM statistics may be integrated to improve fine-mapping and to effectively reduce the number of variants tested in downstream studies.

We also compared our approach to multiple alternative approaches to improve fine-mapping in the adipose subcutaneous tissue. We compared our meta-statistic to two alternative statistics computed using the RASQUAL and TReCASE methods, which obtain median reduction rates of 0% and -15%, respectively. The difference between these two approaches and our approach may be due to the increase in power observed when using RASQUAL and TReCASE compared to a traditional eQTL association analysis. While increasing power is optimal for increasing the number of significant associations, it may not improve our ability to differentiate between causal variants and non-causal variants in LD. In fact, increasing the power may

make more variants highly significant and increase the size of the causal sets. Another potential explanation is that our binarization procedure is more robust to noisy AIM data. Although our binarization procedure may result in a decrease in effect size, our fine-mapping method relies more on the relative effect sizes between variants tested than the effect sizes of individual variants. We also compared our method with alternative fine-mapping methods. DAP-G and PAINTOR 3.0 yielded reduction rates of 0% and -4.5%, respectively. Although the DAP-G was unable to use AIM data to reduce causal set size, using the AIM data to create a prior reduced the number of causal sets considered when approximating the posterior. This indicates that the AIM data was valuable for reducing the computational cost of performing fine-mapping. The differences between our results and the PAINTOR 3.0 results may be in part due to the fact that the PAINTOR framework does not model the uncertainty in the effect size as in CAVIAR, which decreases its performance when external annotations are not used [42]. This decrease in performance may be exacerbated when the effect size varies greatly across variants or when the effect size is not estimated well in a locus.

One of the difficulties in identifying causal variants using AIM information is accurate identification of AIM calls from RNA-seq data. Previous studies have shown that noise caused by mapping errors, small numbers of reads, or inconsistencies between two variants used to call AIM in one locus can impact our ability to accurately detect AIM [39–41]. These problematic AIM calls can decrease signal substantially, and our simulations show that this adversely affects the efficiency of our fine-mapping algorithm. Due to the low concordance in AIM calls from heterozygous coding variants of the same gene, we developed a framework that binarizes AIM status before computing an association statistic. Additionally, when binarizing AIM status, we only use allele-specific reads from the heterozygous SNP with the most reads. These steps make our statistic more robust to noisy data. However, this approach does not utilize all the read data and may be less sensitive than alternative approaches that directly model the likelihood of read data.

Some additional limitations of our analysis of the GTEx data include the small number of genes for which our method can be applied. Since we cannot compute a statistic when no individuals are labeled as having AIM, the genes for which we can compute a meta-statistic are a subset of the genes with AIM data. The number of genes with sufficient data is especially low in tissues where the sample size is low. Although the genes we do analyze have higher expression on average, there may be additional functionally relevant genes for which we cannot apply our method. Another limitation is that we only use the top 50 eQTL variants as input to our fine-mapping framework. This is due to computational limitations of our Bayesian fine-mapping framework. Since we restrict our analysis to the top 50 eQTLs, it is possible for our method to miss some causal variants. Our causal sets are only guaranteed to contain the causal variants among the input with high probability. Fine-mapping complete regions with high LD remains an active area of research (S1 Text). However, in some targeted fine-mapping analyses, it may be possible to run our method for a larger number of variants.

One key advantage of our approach is that additional biological data does not need to be collected. Researchers can perform AIM calling on existing RNA-seq data to apply our approach. Hence, our approach enables more accurate detection of causal variants regulating gene expression without the need for additional experiments. While fine-mapping using other sources of data, such as epigenetic data and transcription factor binding data, may also improve specificity, these types of data are costly to collect in many individuals. Due to the number of existing RNA-seq data sets with a large number of individuals, combining allelic imbalance data with total expression data is a cost-effective way to improve fine-mapping and reduce the number of putative causal variants for followup studies.

Methods

Overview of CAVIAR generative model

This section is a brief overview of the CAVIAR fine-mapping framework [18]. Let $S = [s_1, s_2, \cdots s_m]$ indicate the observed marginal statistics (e.g., z-scores) for a set of m variants, where s_i is the observed marginal statistic of i-th variant. We assume the computed marginal statistics follow an MVN distribution [18],

$$(S|\Lambda) \sim \mathcal{N}(\Sigma\Lambda, \Sigma), \tag{1}$$

where Σ is the LD matrix between pairs of variants and $\Lambda = [\lambda_1, \lambda_2, \cdots \lambda_m]$ is a vector of the true effect sizes [18]. Let $C = [c_1, c_2, \cdots c_m]$ be a vector of zeros and ones that indicate the causal status of each variant. We define the prior probability on Λ for a given C as:

$$(\Lambda|C) \sim \mathcal{N}(0, \Sigma_c), \tag{2}$$

where $\Sigma_c = \sigma^2 \text{diag}(C)$, and σ is a constant which indicates the variance of our prior over the true effect sizes. Similar to prior studies, we set σ to 5.2 [18, 22, 49].

The likelihood of the statistics given the causal status (Eq 3) is also MVN, and we can use this distribution to compute the joint likelihood of the marginal statistics given a causal status.

$$(S|C) \sim \mathcal{N}(0, \Sigma + \Sigma\Sigma_c\Sigma), \tag{3}$$

Given a set of possible causal statuses $\mathbb{C}$, the posterior probability of a causal status $C^* \in \mathbb{C}$ can be expressed as $P(C^*|S) = \frac{P(S|C^*)P(C^*)}{\sum_{C \in \mathbb{C}} P(S|C)P(C)}$. Using this fine-mapping framework, we can compute the posterior probability of a causal status. Unfortunately, considering all possible causal sets is computationally intractable. To make this feasible, we assume that the maximum number of causal variants within a region is six [18, 22, 49]. We also use a greedy algorithm that eliminates the need to consider all possible subsets. To identify the minimal ρ causal set, at each iteration of the greedy algorithm, we select the variant that increases the total posterior probability the most. Variants are added until the posterior probability of the causal set is at least ρ fraction of the total posterior probability of the data.

Computing eQTL association statistics

The following is a general eQTL association framework. We use eQTL association statistics computed by the GTEx Consortium [50] and details regarding the computation of the statistics can be found on the GTEx Portal (https://gtexportal.org/home/documentationPage#staticTextAnalysisMethods). Let Y be the normalized total expression values for n individuals in a single gene, and let X_i be the normalized genotypes of variant i for all individuals. Let Z be a matrix of covariates. Suppose $Y = \mu + \beta_i X_i + \gamma Z + \epsilon$, where μ is the phenotypic mean in the population, β_i is the effect size of variant i, γ is the effect size of covariates, and ϵ models environmental and measurement noise. We use the maximum likelihood estimates $\hat{\mu} = \frac{1}{n} 1^T Y$ and $\hat{\beta}_i = (X_i^T X_i)^{-1} X_i^T Y$. The error can be calculated as $\hat{\epsilon} = Y - \hat{\mu} 1 - \hat{\beta}_i X_i - \hat{\gamma} Z$, and the standard deviation is calculated as $\hat{\sigma} = \sqrt{\frac{\hat{\epsilon}^T \hat{\epsilon}}{n-2}}$. Using these, the association statistic for variant i is calculated as $s_i^E = \frac{\hat{\beta}_i}{\hat{\sigma}} \sqrt{n}$.

We use the top 50 eQTL summary statistics for each gene as input to our framework. Assuming that we have enough individuals, the marginal statistics have the same posterior distribution used by CAVIAR (Eq 3). The distribution of the eQTL summary statistics given a

causal status is as follows:

$$(S^E | C) \sim \mathcal{N}(0, \mathbf{\Sigma^E} + \mathbf{\Sigma^E}\mathbf{\Sigma_c}\mathbf{\Sigma^E}), \qquad (4)$$

where S^E is a vector of the observed marginal statistics for a gene and $\mathbf{\Sigma^E}$ is the LD matrix between pairs of variants in the eQTL study.

Computing AIM association statistics

Using phased genotype data, AIM status for a gene in an individual can be directly computed from RNA-seq data by using heterozygous coding SNPs to map reads to one of the haplotypes. If a gene has multiple heterozygous coding SNPs, we use only the SNP with the most reads to avoid double counting reads overlapping multiple SNPs. The proportion of reads mapping to each haplotype can be used as a proxy for relative contribution from each haplotype to total expression. The genotype phasing and read mapping was performed by the GTEx Consortium and accounts for genotyping error, reference bias, and other sources of technical variation [41]. Let c_1 and c_2 indicate the number of reads supporting the first and second haplotype, respectively. We calculate the allelic ratio (AR) for an individual as $AR = \frac{c_1}{c_1+c_2}$. We consider an individual to have AIM for a gene when the allelic ratio is less than .35 or greater than .65; otherwise, we label the gene in the individual as having balanced expression. This threshold was obtained by taking two standard deviations from the the null (AR = 0.5), where the standard deviation is computed from the empirical distribution (S2 Fig). We compare this binarization framework to one using a binomial test for each individual. If the allelic ratio for an individual is significantly different from 0.5, the individual is labeled as having AIM. Otherwise, the individual is labeled as having balanced expression. Both methods are valid binarization strategies. However, using the empirical threshold yielded a larger reduction rate and allowed us to apply our fine-mapping framework to more genes. This is because applying a two-sided binomial test to binarize the allelic imbalance status across individuals typically generates higher p-values than a 0.05 significance threshold. For example, if an individual has 6 reads mapping to one haplotype of a gene and 14 reads mapping to the other haplotype, the allelic ratio is 0.3. Using the empirically determined threshold of .35, this individual has AIM. However, using a binomial test, this individual is labelled as having balanced expression (p-value = .1153). When using the binomial threshold, fewer individuals are labelled as having allelic imbalance, despite displaying some imbalance between the two alleles. When none of the individuals are labelled as having allelic imbalance for a gene, we are unable to calculate AIM summary statistics and perform fine-mapping for that gene.

If an individual is heterozygous for a causal variant, we expect the expression from each allele to be different. On the other hand, if an individual is homozygous for a causal variant, we expect the expression for each allele to be comparable. AIM association statistics measure the correlation between AIM status (e.g., 0 for balance expression and 1 for AIM) and heterozygous status of a variant (e.g. 0 for homozygous genotype and 1 for heterozygous genotype).

The association statistics are calculated in a way similar to case and control GWAS association statistics. Let n_1 be the number of individuals with AIM in the study, and let n_2 be the number of individuals with balanced expression in the study. Let p_1 be the proportion of individuals with AIM who are heterozygous for the SNP. Let p_2 be the proportion of individuals with balanced expression who are heterozygous for the SNP. The difference between these two proportions is also normally distributed. Under the null hypothesis ($p_1 = p_2$), we have:

$$\hat{p}_1 - \hat{p}_2 \sim N(0, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}). \qquad (5)$$

Let p be the frequency of heterozygous individuals in the population. With the simplifying assumption used in standard GWAS with unequal case and control size that $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \approx \frac{p(1-p)(n_1+n_2)}{n_1 n_2}$, the null hypothesis becomes $\hat{p}_1 - \hat{p}_2 \sim N(0, \frac{p(1-p)(n_1+n_2)}{n_1 n_2})$. We set the AIM association statistic to be:

$$s = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{2\hat{p}(1-\hat{p})}} \sqrt{\frac{2n_1 n_2}{(n_1 + n_2)}}, \tag{6}$$

which follows the standard normal distribution. Software to compute the AIM summary statistics from AIM count data is available at https://github.com/jzou1115/aim. We compute AIM summary statistics for the top 50 eQTL summary statistics for each gene. We assume that the AIM summary statistics follow MVN, and the marginal statistics have the same posterior distribution used by CAVIAR. We compute the heterozygosity matrix from genotype data for all variants in that locus. Thus, similar to the posterior of CAVIAR (Eq 3) we have:

$$(S^A | C) \sim \mathcal{N}(0, \mathbf{\Sigma^A} + \mathbf{\Sigma^A}\mathbf{\Sigma_c}\mathbf{\Sigma^A}), \tag{7}$$

where S^A is the observed marginal statistics for the gene we are interested and $\mathbf{\Sigma^A}$ is the heterozygosity matrix.

Utilizing fixed effect meta-analysis for joint analysis

We model the joint distribution of both eQTL and AIM summary statistics, assuming that the statistics are conditionally independent given the true effect sizes. The posterior can be computed as

$$\begin{pmatrix} S^A \\ S^E \end{pmatrix} \bigg| C \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma^A} + \sigma_a^2 \mathbf{\Sigma^A}\mathbf{\Sigma^c}\mathbf{\Sigma^A} & 0 \\ 0 & \mathbf{\Sigma^E} + \sigma_e^2 \mathbf{\Sigma^E}\mathbf{\Sigma_c}\mathbf{\Sigma^E} \end{pmatrix} \right), \tag{8}$$

An alternative approach is to do a fixed effect meta-analysis followed by fine-mapping. For now we assume both studies have the same number of individuals. Thus, the computed combined marginal statistics S^M can be calculated as $S^M = \frac{S^A + S^E}{\sqrt{2}}$. This combined statistic is the sum of two Gaussian random variables. Therefore, the posterior can be expressed as

$$(S^M | C) \sim N\left(0, \frac{1}{2}(\mathbf{\Sigma^E} + \mathbf{\Sigma^A}) + \frac{1}{2}(\sigma_a^2 \mathbf{\Sigma^A}\mathbf{\Sigma_c}\mathbf{\Sigma^A} + \sigma_e^2 \mathbf{\Sigma^E}\mathbf{\Sigma_c}\mathbf{\Sigma^E})\right) \tag{9}$$

Our meta-analysis framework is an approximation for modeling the joint distribution of AIM scores and eQTL scores. Since the likelihood ratio of the joint model follows the chi-squared distribution with two degrees of freedom, the significance threshold for the joint model for one variant is a circle. The significance threshold for each of the meta-statistics is a set of parallel lines, making the combined meta-analysis significance threshold a square. A geometric interpretation of the significant thresholds under the different models is shown in S6 Fig, and it is evident that the meta-analysis framework is a close approximation to using the joint distribution.

Calculating pairwise correlation matrices

To calculate pairwise correlation matrices, we use genotype data from the whole blood tissue in GTEx (Release v6, dbGaP Accession phs000424.v6.p1 available at: http://www.gtexportal.org) that contains 325 samples. For each gene, we compute the eQTL statistic for every

cis variant within 1MB from the transcription start site and identify the top 50 genetic variants with the highest eQTL statistics. We use these top 50 genetic variants to compute the pairwise correlation matrices. For eQTL statistics, we calculated $\Sigma^{\mathbf{E}}$ as the pairwise correlation matrices between genotypes of individuals. For AIM statistics, we calculated $\Sigma^{\mathbf{A}}$ as the pairwise correlation between the heterozygosity calls of individuals. Since the statistics are normally distributed, we calculated the meta-analysis correlation matrix as $\Sigma^{\mathbf{M}} = \frac{1}{2}(\Sigma^{\mathbf{E}} + \Sigma^{\mathbf{A}})$.

Generating simulated datasets

Given a genotype LD matrix ($\Sigma^{\mathbf{E}}$) and a heterozygosity LD matrix ($\Sigma^{\mathbf{A}}$) for a gene and a vector C that indicates the causal status of each SNP (e.g., 0 when the variant is not causal and 1 when the variant is causal), we can simulate the association statistics by sampling $S^E \sim N(\lambda_c \Sigma^{\mathbf{E}} C, \Sigma^{\mathbf{E}})$ and $S^A \sim N(\lambda_c \Sigma^{\mathbf{A}} C, \Sigma^{\mathbf{A}})$, where λ_c is the non-centrality parameter (NCP). We set λ_c such that we have the desired power. We first set the power level to 50% and 80% for both AIM and eQTL statistics. Simulations at these power levels indicate that our fine-mapping method is effective for the strongest eQTL loci. We then explore cases where the power level is different between AIM and eQTL statistics and show that our method can also reduce causal set size when the fixed effect assumption is broken. We calculate the simulated statistics for the meta-analysis in two ways: 1) $S^{M1} = \frac{S^A + S^E}{\sqrt{2}}$ and 2) $S^{M2} = \frac{S^A - S^E}{\sqrt{2}}$. We perform one simulation for each gene in the GTEx data to observe a natural range of correlation between variants. Simulations to generate S^E, S^A, S^{M1}, and S^{M2} were performed for up to three causal variants.

Application to GTEx data

Genotype data was collected using Illumina OMNI 5M SNP Arrays. To increase eQTL discovery power, genotypes were imputed from the 1000 Genomes Project Phase I version 3 reference panel using the IMPUTE2 software [51]. To measure the total expression level of each gene, RNA-seq data was aligned to hg19 using Tophat v1.4.1 [52] and gene-level expression quantification was performed using RNA-SeQC [53] according to the GTEx protocol [8]. Reads were assigned to each allele of an individual using personalized genomes and the official GTEx protocol described in [41].

To reduce the affect of noise in AIM calls, we require individuals to have at least 20 reads mapped to each gene. The read threshold is directly proportional to the quality of AIM calls for each individual (S1 Fig). However, using a higher read threshold decreases the number of individuals that can be used for the analysis, which decreases the quality of our results. We found a read threshold of 20 maximizes the reduction rate in the real data. We apply this framework to genes with at least one significant eQTL identified in the GTEx data set and at least one individual with a heterozygous variant in the coding region. The significant eQTL requirement restricts the set of genes tested to those likely to be regulated by cis-regulatory variants. The heterozygous variant in the coding region is used to call AIM. For each gene, we selected the top 50 variants that have the highest eQTL statistics. We calculate the meta-statistics and heterozygosity matrix for these variants, which we use as input to our fine-mapping framework. Since it is possible for AIM and eQTL summary statistics to be in different directions in the real data, we calculate two meta-statistics, $S^{M1} = \frac{S^A + S^E}{\sqrt{2}}$ and $S^{M2} = \frac{S^A - S^E}{\sqrt{2}}$. We perform fine-mapping using each set of meta-analysis statistics separately. As a post-processing step, we choose the smallest causal set.

Fine-mapping with TReCASE and RASQUAL statistics

To assess whether existing statistics, such as likelihood ratio statistics, can be used in our fine-mapping framework, we computed TReCASE and RASQUAL statistics [37, 38]. We used BAM files processed by the GTEx Consortium for ASE analysis as input to these programs [8]. For the TReCASE analysis, we performed an additional QC step using the prepareBam function of the asSeq software package, as suggested by the documentation. For both analysis, we used the default settings to compute the chi-squared statistics.

We used these statistics as input to the CAVIAR fine-mapping framework. For each method, we attempted to compute chi-squared statistics for 4023 genes in the adipose subcutaneous tissue reported in our results. We only performed fine-mapping in genes that converged and obtained joint statistics for all cis-eQTLs tested in our framework. In order to fit our model's assumptions, we transformed the chi-squared statistics into z-scores by taking the square root. Since our method is sensitive to the direction of effect sizes, we used the sign from the eQTL summary statistic together with the magnitude obtained from the chi-squared statistic. This modification was necessary for our framework since the LD matrix allows for variants to be negatively correlated. We used these modified statistics along with the previously computed LD matrices and heterozygosity matrices as input to our framework and calculated the reduction in causal set size for each gene relative to an approach that only uses eQTL statistics. For TReCASE, we peformed an additional comparison to the reduction rate relative to the causal set obtained using the TReC (eQTL) chi-squared statistics for the variants.

Supporting information

S1 Text. Supplementary methods.
(DOCX)

S1 Fig. Higher read count thresholds yield higher concordance between AIM measurements. We calculated the correlation between allelic ratios obtained by pairs of SNPs within the same gene for each individual in the adipose subcutaneous tissue (Read threshold = 0). We then recomputed the correlations using a range of read thresholds. As the read threshold is increased, the correlation improves. However, increasing the read threshold also reduces the number of individuals that can be included in the analysis. Due to this trade off, in our analysis of the real data, we use a read threshold of 20.
(TIF)

S2 Fig. Distribution of allelic ratios across all samples and genes in the adipose subcutaneous tissue. The allelic ratio (AR) was calculated for SNPs in the adipose subcutaneous tissue. The distribution of AR is approximately normal. We calculated the empirical standard deviation of this distribution and created thresholds two standard deviations from the null (AR = 0.5). Individuals with $AR < 0.35$ or $AR > 0.65$ were labeled as having AIM. Others were labeled as having balanced expression. Although this empirical threshold is less stringent than using a binomial test, in practice, this method yields higher reduction in causal set size than an approach using a binomial test to binarize AIM status.
(TIF)

S3 Fig. Distribution of reduction rate. While the median reduction rate across all genes and all tissues is 0.11, there are some genes that have substantially higher reduction rate in set size when using our approach. Only 16004/20130 genes with reduction rate greater than zero are shown.
(TIF)

S4 Fig. Enrichment of putative causal variants for being near transcription start sites (TSS). For each tissue, the "meta causal variants" includes all variants in the meta causal sets for all genes tested. The "excluded variants" includes all variants contained in the eQTL causal sets but excluded from the meta causal sets. The meta causal variants (red) are on average closer to the TSS than the excluded variants (blue).
(TIF)

S5 Fig. Comparison of causal sets with top eQTL sets. We constructed sets of the top eQTLs for each gene with the same size as the meta causal sets. We call these sets of variants the TopN sets. We then computed the number of variants in the causal set that were not present in the TopN set for each gene. The distribution of the number of putative causal variants that are not in the TopN sets is shown below. This implies that our fine-mapping framework does not merely discard large blocks of variants with weaker regulatory effect sizes.
(TIF)

S6 Fig. Meta-analysis is a close approximation for joint analysis. The significance threshold for a joint analysis of one variant is a circle. In our framework, we calculate two meta-statistics (S^{M1} and S^{M2}) and apply our method using each statistic separately. In a test for one variant, the significance threshold for each meta-statistic is a set of parallel lines, and the combination of these two sets of parallel lines forms a square. Therefore, our framework is a close approximation to using the joint distribution.
(TIFF)

S1 Table. Recall rates for eQTL method.
(XLSX)

S2 Table. Recall rates for eQTL + AIM method.
(XLSX)

S3 Table. Recall rates with unequal effect sizes. eQTL effect sizes are fixed to 5.2 (50% power), while AIM effect sizes vary from 0 to 4.
(XLSX)

S4 Table. Recall rates with unequal effect sizes using estimated prior from PAINTOR 3.0. eQTL effect sizes are fixed to 5.2 (50% power), while AIM effect sizes vary from 0 to 4.
(XLSX)

S5 Table. Enrichment of variants in meta causal sets for active regions of the genome relative to variants only in eQTL causal sets. Enrichment for active regions of the genome was computed in a tissue-specific manner. For each tissue, the "meta causal variants" includes all variants in the meta causal sets for all genes tested. The "eQTL only variants" includes all variants contained in the eQTL causal sets but excluded from the meta causal sets. "Meta Odds" column contains the odds of a variant in the meta causal variants being in an active region, and the "eQTL Only Odds" column contains the odds of a variant in the excluded variants being in an active region. The "Odds Ratio" column contains the ratio of "Meta Odds" and "eQTL Only Odds". The p-value for the odds ratio was calculated using Fisher's exact test. The odds ratios are highly significant in all tissues.
(XLSX)

S6 Table. Enrichment of variants only in meta causal sets for active regions of the genome relative to variants only in eQTL causal sets. Enrichment for active regions of the genome was computed in a tissue-specific manner. For each tissue, the "meta only variants" includes all variants contained in the meta causal sets that are excluded from the eQTL causal sets. The

"eQTL only variants" includes all variants contained in the eQTL causal sets but excluded from the meta causal sets. "Meta Only Odds" column contains the odds of a variant in the meta only causal variants being in an active region, and the "eQTL Only Odds" column contains the odds of a variant in the eQTL only variants being in an active region. The "Odds Ratio" column contains the ratio of "Meta Only Odds" and "eQTL Only Odds". The p-value for the odds ratio was calculated using Fisher's exact test. The odds ratio is greater than one in 6/7 tissues, indicating that the variants in the meta only set are slightly enriched for being in active regions of the genome in these tissues. However, this result is not significant due to the small number of variants in the meta only set.
(XLSX)

S7 Table. Enrichment of variants in meta causal sets for active regions of the genome relative to variants in eQTL causal sets. Enrichment for active regions of the genome was computed in a tissue-specific manner. For each tissue, the "meta variants" includes all variants contained in the meta causal sets, and the "eQTL variants" includes all variants contained in the eQTL causal sets. "Meta Odds" column contains the odds of a variant in the meta causal variants being in an active region, and the "eQTL Odds" column contains the odds of a variant in the eQTL variants being in an active region. The "Odds Ratio" column contains the ratio of "Meta Odds" and "eQTL Odds". The p-value for the odds ratio was calculated using Fisher's exact test. The odds ratio is greater than one in all tissues, indicating that the variants in the meta only set are slightly enriched for being in active regions of the genome in these tissues. Due to the high overlap between the two sets, this result is only significant in 3/7 tissues.
(XLSX)

Author Contributions

Conceptualization: Eleazar Eskin.

Data curation: Jennifer Zou, Brandon Jew, Stephane E. Castel, Tuuli Lappalainen.

Methodology: Jennifer Zou, Farhad Hormozdiari, Jason Ernst, Jae Hoon Sul, Eleazar Eskin.

Software: Jennifer Zou.

Writing – original draft: Jennifer Zou, Farhad Hormozdiari, Jae Hoon Sul.

Writing – review & editing: Jennifer Zou, Jason Ernst, Jae Hoon Sul, Eleazar Eskin.

References

1. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, et al. Genetics of gene expression and its effect on disease. Nature. 2008; 452(March).

2. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genetics. 2010; 6(4):e1000895. https://doi.org/10.1371/journal.pgen.1000895 PMID: 20369022

3. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genetics. 2010; 6(4):e1000888. https://doi.org/10.1371/journal.pgen.1000888 PMID: 20369019

4. Davis LK, Yu D, Keenan CL, Gamazon ER, Konkashbaev AI, Derks EM, et al. Partitioning the Heritability of Tourette Syndrome and Obsessive Compulsive Disorder Reveals Differences in Genetic Architecture. PLoS Genetics. 2013; 9(10):e1003864. https://doi.org/10.1371/journal.pgen.1003864 PMID: 24204291

5. Torres JM, Gamazon ER, Parra EJ, Below JE, Valladares-Salgado A, Wacher N, et al. Cross-Tissue and Tissue-Specific eQTLs: Partitioning the Heritability of a Complex Trait. The American Journal of Human Genetics. 2014; 95(5):521–534. https://doi.org/10.1016/j.ajhg.2014.10.001 PMID: 25439722

6. Brem RB, Clinton R. Genetic Dissection of Transcriptional Regulation in Budding Yeast. Science. 2002; 296(April). https://doi.org/10.1126/science.1069516 PMID: 11923494

7. Consortium TG. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45(6):580–585. https://doi.org/10.1038/ng.2653

8. Consortium TG. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015; 348(6235):648–660. https://doi.org/10.1126/science.1262110

9. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nature Genetics. 2016; 49(1):139–145. https://doi.org/10.1038/ng.3737 PMID: 27918533

10. Consortium TG. Genetic effects on gene expression across human tissues. Nature. 2017; 550:204–213. https://doi.org/10.1038/nature24277

11. Wellcome Trust Case Control Consortium T. Bayesian refinement of association signals for 14 loci in 3 common diseases. Nature Genetics. 2012; 44(12).

12. Malo N, Libiger O, Schork NJ. Accommodating Linkage Disequilibrium in Genetic-Association Analyses via Ridge Regression. Am J Hum Genet. 2008; 82(February):375–385. https://doi.org/10.1016/j.ajhg.2007.10.012 PMID: 18252218

13. Yang J, Ferreira T, Morris AP, Medland SE, Investigation G, Madden PAF, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Publishing Group. 2012; 44(4):369–375.

14. Jansen R, Hottenga JJ, Nivard MG, Abdellaoui A, Laport B, de Geus EJ, et al. Conditional eQTL Analysis Reveals Allelic Heterogeneity of Gene Expression. Human Molecular Genetics. 2017. https://doi.org/10.1093/hmg/ddx043 PMID: 28165122

15. Hormozdiari F, Zhu A, Kichaev G, Ju CJ, Segre AV, Joo JWJ, et al. Widespread Allelic Heterogeneity in Complex Traits. AJHG. 2017; 100(5):789–802. https://doi.org/10.1016/j.ajhg.2017.04.005

16. Brown AA, Vinuela A, Delaneau O, Spector TD, Small KS, Dermitzakis ET. Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. Nature Genetics. 2017. https://doi.org/10.1038/ng.3979

17. Servin B, Stephens M. Imputation-based analysis of association studies: Candidate regions and quantitative traits. PLoS Genetics. 2007; 3(7):1296–1308. https://doi.org/10.1371/journal.pgen.0030114

18. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying Causal Variants at Loci with Multiple Signals of Association. Genetics. 2014; 198(2):497–508. https://doi.org/10.1534/genetics.114.167908 PMID: 25104515

19. Farh KKH, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2014; 518(7539):337–343. https://doi.org/10.1038/nature13835 PMID: 25363779

20. Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, et al. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. Genetics. 2015; 200(3):719–736. https://doi.org/10.1534/genetics.115.176107 PMID: 25948564

21. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016; 32 (10):1493–1501. https://doi.org/10.1093/bioinformatics/btw018 PMID: 26773131

22. Hormozdiari F, Kichaev G, Yang WY, Pasaniuc B, Eskin E. Identification of causal genes for complex traits. Bioinformatics. 2015; 31(12):i206–i213. https://doi.org/10.1093/bioinformatics/btv240 PMID: 26072484

23. Pastinen T, Hudson TJ. Cis-acting regulatory variation in the human genome. Science (New York, NY). 2004; 306(5696):647–650. https://doi.org/10.1126/science.1101659

24. Hasin-Brumshtein Y, Hormozdiari F, Martin L, van Nas A, Eskin E, Lusis AJ, et al. Allele-specific expression and eQTL analysis in mouse adipose tissue. BMC Genomics. 2014; 15(1):471. https://doi.org/10.1186/1471-2164-15-471 PMID: 24927774

25. Baran Y, Subramaniam M, Biton A, Tukiainen T, Tsang EK, Rivas MA, et al. The landscape of genomic imprinting across diverse adult human tissues. Genome Research. 2015; 25(7):927–936. https://doi.org/10.1101/gr.192278.115 PMID: 25953952

26. Mohammadi P, Castel SE, Brown AA, Lappalainen T. Quantifying the regulatory effect size of cis -acting genetic variation using allelic fold change. Genome Research. 2017; p. 1–13.

27. Yan H. Allelic Variation in Human Gene Expression. Science. 2002; 297(5584):1143–1143. https://doi.org/10.1126/science.1072545 PMID: 12183620

28. Verlaan DJ, Ge B, Grundberg E, Hoberman R, Lam KCL, Koka V, et al. Targeted screening of cis- regulatory variation in human haplotypes. Genome Research. 2009; p. 118–127. https://doi.org/10.1101/gr.084798.108 PMID: 18971308

29. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. Nature Reviews Genetics. 2010; 11(8):533–538. https://doi.org/10.1038/nrg2815 PMID: 20567245

30. Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, et al. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. Nature methods. 2009; 6(8):613–618. https://doi.org/10.1038/nmeth.1357 PMID: 19620972

31. Cowles CR, Hirschhorn JN, Altshuler D, Lander ES. Detection of regulatory variation in mouse genes. Nature Genetics. 2002; 32(3):432–437. https://doi.org/10.1038/ng992 PMID: 12410233

32. Nica AC, Dermitzakis ET. Expression quantitative trait loci: Present and future. Philosophical Transactions of the Royal Society B: Biological Sciences. 2013; 368 (1620). https://doi.org/10.1098/rstb.2012.0362

33. Nagel M, Jansen PR, Stringer S, Watanabe K, Leeuw CAD, Bryois J, et al. Meta-analysis of genome-wide association studies for neuroticism in 449, 484 individuals identifies novel genetic loci and pathways. Nature genetics. 2018; 50:920–927. https://doi.org/10.1038/s41588-018-0151-7 PMID: 29942085

34. Lam M, Trampush JW, Yu J, Glahn DC, Malhotra AK, Lam M, et al. Large-Scale Cognitive GWAS Meta-Analysis Reveals Tissue-Specific Neural Expression and Potential Nootropic Drug Targets Resource Large-Scale Cognitive GWAS Meta-Analysis Reveals Tissue-Specific Neural Expression and Potential Nootropic Drug Targets. Cell Reports. 2017; 21(9):2597–2613. https://doi.org/10.1016/j.celrep.2017.11.028 PMID: 29186694

35. Gonnermann A, Framke T, Großhennig A, Koch A. No solution yet for combining two independent studies in the presence of heterogeneity. Statistics in Medicine. 2015; 34(16):2476–2480. https://doi.org/10.1002/sim.6473 PMID: 26040434

36. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods. 2010; 1(2):97–111. https://doi.org/10.1002/jrsm.12 PMID: 26061376

37. Kumasaka N, Knights AJ, Gaffney DJ. technical reports Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nature Genetics. 2016; 48(2). https://doi.org/10.1038/ng.3467

38. Hu Y, Sun W, Tzeng J, C P. Proper Use of Allele-Specific Expression Improves Statistical Power for cis-eQTL Mapping with RNA-Seq Data. J Am Stat Assoc. 2015; 110(511):962–974. https://doi.org/10.1080/01621459.2015.1038449 PMID: 26568645

39. Harvey CT, Moyerbrailean GA, Davis GO, Wen X, Luca F, Pique-regi R. Genetics and population analysis QuASAR: quantitative allele-specific analysis of reads. Bioinformatics. 2015; 31(December 2014):1235–1242. https://doi.org/10.1093/bioinformatics/btu802 PMID: 25480375

40. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009; 25(24):3207–3212. https://doi.org/10.1093/bioinformatics/btp579 PMID: 19808877

41. Pirinen M, Lappalainen T, Zaitlen NA, GTEx Consortium, Dermitzakis ET, Donnelly P, et al. Assessing allele-specific expression across multiple tissues from RNA-seq read data. Bioinformatics. 2015; 31(15):2497–2504. https://doi.org/10.1093/bioinformatics/btv074 PMID: 25819081

42. Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, et al. Fine mapping causal variants with an approximate bayesian method using marginal test statistics. Genetics. 2015; 200(3):719–736. https://doi.org/10.1534/genetics.115.176107 PMID: 25948564

43. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. American Journal of Human Genetics. 2016; 99(6):1245–1260. https://doi.org/10.1016/j.ajhg.2016.10.003 PMID: 27866706

44. Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JBM, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature Genetics. 2007; 39(10):1208–1216. https://doi.org/10.1038/ng2119 PMID: 17873875

45. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature Biotechnology. 2010; 28(8):817–825. https://doi.org/10.1038/nbt.1662 PMID: 20657582

46. Ernst J, Kellis M. ChromHMM: automating chromatin- state discovery and characterization. Nature Methods. 2012; 9(3):215–216. https://doi.org/10.1038/nmeth.1906 PMID: 22373907

47. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518(7539):317–330. https://doi.org/10.1038/nature14248 PMID: 25693563

48. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nature Methods. 2015; 12(11):1061–3. https://doi.org/10.1038/nmeth.3582 PMID: 26366987

49. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. The American Journal of Human Genetics. 2016; 99(6):1245–1260. https://doi.org/10.1016/j.ajhg.2016.10.003 PMID: 27866706

50. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017; 550(7675):204–213. https://doi.org/10.1038/nature24277

51. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genetics. 2009; 5(6). https://doi.org/10.1371/journal.pgen.1000529 PMID: 19543373

52. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25(9):1105–1111. https://doi.org/10.1093/bioinformatics/btp120 PMID: 19289445

53. DeLuca D, Levin J, Sivachenko A, Fennell T, Nazaire M, Williams C, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012; 11(28):1530–2. https://doi.org/10.1093/bioinformatics/bts196