**Title: Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis**

**Authors:** Fan Yang[1], Jiebiao Wang[1], the GTEx consortium, Brandon L. Pierce[1,2,3], and Lin S. Chen[1]

**Affiliations:** [1]Department of Public Health Sciences, [2]Department of Human Genetics, and [3]Comprehensive Cancer Center, The University of Chicago, Chicago, IL 60637

**Corresponding Authors:**

Dr. Lin Chen

The University of Chicago

5841 South Maryland Ave, W258, MC2000

Chicago, IL 60637

Phone/ Fax: 773-702-1626 / 773-834-0139

Email: lchen@health.bsd.uchicago.edu

AND

Dr. Brandon Pierce

The University of Chicago

5841 South Maryland Avenue, W264, MC2000

Chicago, IL 60637

Phone / Fax: 773-702-1917 / 773-834-0139

Email: brandonpierce@uchicago.edu

**Running Title:**

Genomic mediation analyses of GTEx data

**Key Words:** expression quantitative trait loci (eQTL), trans-eQTL, cis-eQTL, genomic mediation analysis, adaptive selection, confounding

**Funding:** This work was supported by National Institutes of Health grants R01 GM108711, U01 HG007601, and R01 MH101820.

**Acknowledgements:** We would like to thank Alexis Battle for providing the estimates of mappability used in this work. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplements to University of Miami grants DA006227 & DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 & MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania (MH101822). The data used for the analyses described in this manuscript were obtained from: the GTEx Portal in January of 2015.

**ABSTRACT**

The impact of inherited genetic variation on gene expression in humans is well-established. The majority of known expression quantitative trait loci (eQTLs) impact expression of local genes (cis-eQTLs); more research is needed to identify effects of genetic variation on distant genes (trans-eQTLs) and understand the biological mechanisms. One common trans-eQTLs mechanism is "mediation" by a local (cis) transcript. Thus, mediation analysis can be applied to genome-wide SNP and expression data in order to identify transcripts that are "cis-mediators" of trans-eQTLs, including those "cis-hubs" involved in regulation of many trans-genes. Identifying such mediators helps us understand regulatory networks and suggests biological mechanisms underlying trans-eQTLs, both of which are relevant for understanding susceptibility to complex diseases. The multi-tissue expression data from the Genotype-Tissue Expression (GTEx) program provides a unique opportunity to study cis-mediation across human tissue types. However, the presence of complex hidden confounding effects in biological systems can make mediation analyses challenging and prone to confounding bias, particularly when conducted among diverse samples. To address this problem, we propose a new method: Genomic Mediation analysis with Adaptive Confounding adjustment (GMAC). It enables the search of a very large pool of variables, and adaptively selects potential confounding variables for each mediation test. Analyses of simulated data and GTEx data demonstrate that the adaptive selection of confounders by GMAC improves the power and precision of mediation analysis. Application of GMAC to GTEx data provides new insights into the observed patterns of cis-hubs and trans-eQTL regulation across tissue types.

## INTRODUCTION

Recent studies of the effects of genetic variation on expression of distant genes (trans-eQTLs) have revealed that many trans-eQTL effects are "mediated" by the local (cis-) gene transcripts near the eQTLs [1, 2]. In other words, some cis-eQTLs are also trans-eQTLs because the variation in the expression of the cis-genes effects the expression of a trans-gene or genes. By studying the cis- to trans-gene transcript mediation patterns, one may identify the cis-genes that regulate trans-eQTLs, including the "cis-hubs" that regulate many trans-genes [3, 4]. Characterizing these regulatory relationships will allow us to better understand regulatory networks and their roles in complex diseases [5], as it is well-known that SNP influencing human traits tend to be eQTLs [6]. Analyses of cis-mediation will also provide us with a better understanding of the biological mechanisms underlying trans-eQTLs [7].

The expression levels of a given gene can vary substantially across human cell types, and the regulatory relationships between SNPs and gene expression levels may also depend on cell type [8, 9]. To date, most large-scale eQTL studies have been conducted using RNA extracted from peripheral blood cells, which are mixtures of different cell types and may not be informative for gene regulation in other human tissues. In order to study gene expression and regulation in a variety of human tissues, the National Institutes of Health common-fund GTEx (Genotype-Tissue Expression) project has collected expression data on 44 tissue types from hundreds of post-mortem donors [10, 11]. This rich transcriptome data, coupled with data on inherited genetic variation, provides an unprecedented opportunity to study gene expression and regulation patterns from both cross-tissue and tissue-specific perspectives.

In prior studies, mediation tests have been applied to genome-wide expression data from blood cells to examine whether the effects of trans-eQTLs are mediated by cis-gene transcripts, i.e., cis-gene expression levels regulate trans-gene expression levels [1, 4, 12]. One major challenge in such mediation analyses or gene regulatory analysis is the presence of unmeasured or unknown confounding effects, as it is well-

known that "mediator-outcome confounding" (or in this case, confounding of the association between the cis and trans genes affected by an eQTL) can bias estimates obtained from mediation analysis [13-15]. It is well recognized that transcriptional variation can be affected by many factors including genetic, environmental, demographic, technical, as well as biological factors. The presence of unmeasured or unknown confounding effects may induce inflated rates of false detection of mediation relationships or jeopardize the power to detect real mediation, if those effects are not well accounted for. Given that eQTL analyses are conducted in the context of complex biological systems, there are a wide array of biological variables that could bias mediation estimates, a problem that may be exacerbated by the diversity of GTEx participants, with respect to ethnicity, age, and cause of death. Given these challenges, it is desirable to have methods that consider a large pool of potential confounding variables.

To adjust for unmeasured or unknown confounding effects in genomics studies, existing literature focused on the construction of sets of "hidden" variables that capture a substantial amount of the variation in a large set of variables. Common approaches for detecting hidden variables in expression data include principal components analysis (PCA) [16], surrogate variable analysis (SVA) [17], and the Probabilistic Estimation of Expression Residuals (PEER) method [18]. A commonality of those approaches is that they model the effects of hidden confounding factors and summarize those effects into a set of constructed variables. The constructed variables are sorted decreasingly by their estimated impacts on expression variability, and the top constructed variables are selected and adjusted as a set of covariates to eliminate major confounding effects. For example, in GTEx eQTL analyses [10] (cite Jo et al., GTEx companion paper, unpublished) the top PEER factors were estimated for each tissue type, with the number of selected PEER factors depending on tissue sample size. Up to 35 PEER factors are selected for tissues with large sample sizes. One aspect that is largely ignored is that not all the potential gene pairs (or pairs of regulator and regulated genes) are affected by the same set of hidden confounders. When studying mediation or gene regulation in the genome, potentially there are many thousands of trios representing a cis-mediated trans-eQTL, consisting of a genetic variant, a cis-gene transcript, and a trans-

gene transcript in a specific tissue type. Adjusting a universal set of variables for all mediation trios is not only inefficient but also may limit our ability to consider a larger pool of potential confounding variables in genomic mediation analyses.

We propose to adaptively select the variables to adjust for each trio given a large set of constructed or directly measured potential confounding variables. This strategy supplements existing confounding adjustment approaches that focus on the construction of variables for capturing confounding effects, and enlarges the pool of variables to be considered. Additionally, by leveraging the cis genetic variant as an 'instrumental variable', we are able to select the variables capturing confounding effects rather than variables only correlated with cis- and trans-genes. We further propose a mediation test with non-parametric p-value calculation, adjusting for the adaptively selected sets of confounders. We term the proposed algorithm Genomic Mediation analysis with Adaptive Confounding adjustment (GMAC). Figure 1 provides a graphical illustration of the main steps in GMAC. The GMAC algorithm improves the efficiency and precision of confounding adjustment and the subsequent genomic mediation analyses. We applied GMAC to each of the 44 tissue types of GTEx data (accession number: GTEx_phs000424) in order to study the trans-regulatory mechanism in human tissues. Our algorithm identifies genes that mediate trans-eQTLs in multiple tissues, as well as "cis-hubs" that mediate the effects of a trans-eQTL on multiple genes.

## RESULTS

### ***GMAC improves power and precision of analysis of GTEx data***

We performed genomic mediation analysis with data from each tissue type in GTEx. Taking the tissue Adipose Subcutaneous as an example, there are 298 samples for this tissue type and gene-level expression measures for 27,182 unique transcripts are available after quality control. We detect a cis-eQTL for 8,500 of these transcripts, corresponding to 8,216 unique cis-eSNPs for subsequent analysis. We applied Matrix

eQTL [19] to the 8,216 SNPs and the 27,182 gene expression levels to calculate the pair-wise trans-associations. At the p-value cutoff of $10^{-5}$, there are 3,169 significant pairs of SNP and trans-gene transcripts. Since some cis-eSNPs are the lead cis-eSNPs for multiple local gene transcripts, those significant SNP and trans-gene pairs entailed a total of 3,332 trios (i.e., SNP-cis-trans) for this tissue type. We applied GMAC to the 3,332 trios in this tissue type to test for mediation, and obtained the mediation p-values for those trios. We considered all PCs constructed from the expression data of each tissue type as potential confounders, with the number of PCs equal to the sample size for each tissue minus 1. We analyzed trios for mediation in a similar fashion for all other GTEx tissue types.

At the 5% false discovery rate (FDR) [20] level, we identified 6,145 instances of significant mediation out of 64,824 trios tested in the 44 tissue types. These trios represent potential examples of cis-mediation of trans-eQTLs within a specific tissue. Table 1 lists the number of significant mediation trios at 5% FDR and the number of trios with suggestive mediation (p-value < 0.05), as well as the total number of trios with significant cis- and trans-associations for all tissue types. The number of confounders selected for each mediation test ranged from 0 to 22 across all tissue types, with a mean of 7.695 and a median of 8. The median number of confounders selected for each tissue type ranged from 3 to 12, while the pool of variables (PCs) from which we selected confounders from ranged from 69 to 360. Supplementary Table 2 presents the descriptive statistics for the number of selected confounders for all the trios in each tissue type. It is clear that with GMAC, on average we adjust a much fewer number of confounding variables in the mediation tests, and greatly improving the efficiency of the analyses.

Again taking the tissue Adipose Subcutaneous as an illustration, in Figure 2 we plotted the negative log base 10 of the mediation p-values versus the percentage of reduction in trans-effects after adjusting for a potential cis-mediator, based on mediation tests without adjusting for hidden confounders (Figure 2A) and mediation tests by GMAC considering all PCs as potential confounders (Figure 2B). The percentage of reduction in trans-effects is calculated by $(\beta_2^m - \beta_2)/\beta_2^m$, where $\beta_2^m$ is the marginal trans-effect of the

eQTL on the trans-gene expression levels, and $\beta_2$ is the trans-effect after accounting for cis-mediation. For trios representing true cis-mediation, we expect the trans-effects to be substantially reduced after adjusting for the mediator. That is, we expect the trios with very significant mediation p-values to have positive % reduction in the trans-effect. In Figure 2A, we observed many trios with significant mediation p-values, but for a substantial number of these trios, the percentages of reduction in trans-effects are negative. This pattern is expected in the presence of unadjusted confounders, so these trios represent false positives. Thus, mediation analyses of GTEx data without adjusting for hidden confounding effects will lead to many spurious findings.

In addition to our main analysis based on GMAC (adaptively selecting confounders from all expression PCs), we also conducted mediation tests adjusting for only the 35 PEER factors used in the GTEx eQTL analyses (cite Jo et al., GTEx companion paper, unpublished) At an FDR of 5%, 3,356 out of 64,824 trios from all tissue types were significant. Using GMAC adjusting for adaptively selected PEER factors, 5,131 trios were significant at the 5% FDR level. The comparison of adjusting for all (up to 35) PEER factors versus GMAC (considering a larger pool of potential confounders with up to 360 PCs) demonstrates that adaptive selection enables more efficient adjustment of confounding effects with a much fewer number of selected confounding variables (Supplementary Table 2) and improves power to detect mediation. Furthermore, using GMAC to adaptively select confounders from all PCs identifies 6,145 significant trios, suggesting an increase in power. Meanwhile, all the three methods, 1) GMAC with adaptively-selected PCs, 2) GMAC with adaptively-selected PEER factors, and 2) adjusting for all PEER factors, would yield reasonable mediation estimates (i.e., percentages of reduction in trans-effects versus mediation p-values), as compared to no confounder adjustment (see Supplementary Figure 1).

The majority of the cis-mediators and trans target genes observed among our trios showing mediation have high mappability scores (Supplementary Figure 2A and 2D). However, non-uniquely mapping reads can result in false positive eQTLs, so we consider the mappability of each gene as a quality control filter

for studying specific examples of cis-mediation (see Methods). Examining the mappability for genes involved in cis-mediation, we observed that cis-genes showing evidence of cis-mediation for multiple trans genes were enriched for cis-genes with low mappability scores (Supplementary Figure 2). Similarly, genes showing evidence of cis-mediation across many different tissue types were also enriched for genes showing low mappability scores (Supplementary Figure 2). This finding demonstrates that transcripts that do not uniquely map to the genome are an important source of false positives when conducting genomic mediation analysis. More specifically, we find that analyzing low-mappability genes can lead to the identification of spurious cis-hubs and cross-tissue cis-mediators.

We attempted to identify "cis-hubs" with high mappability in the GTEx data, defined as a transcript that appears to mediate the effect of a nearby eSNP on expression of multiple distant (i.e., trans) gene transcripts. Restricting our analysis to cis and trans genes with mappability >0.95, we observed 685 cis-genes with at least two trans targets (considering all tissues), representing 21% of the 3,168 cis-genes observed among the trios with a mediation P <0.05 (Table 2). In addition, we attempted to identify cis-genes that have at least one trans target in multiple tissues. Restricting to high mappability genes, we observed 531 cis-genes with trans targets in more than one tissue, representing 17% of the 3,168 cis-genes observed among the trios with a mediation P <0.05 (Table 2). We observed only six examples of cis-genes that had the same trans targets in multiple tissues. In other words, vast majority of cis-hubs observed were of two distinct types: 1) those that mediated the effect of a trans-eQTL on multiple trans-genes within a single tissue type, and 2) those that were mediators in multiple tissues, but with unique trans targets in each tissue type. All instances of cis-mediation of trans-eQTLs with a mediation P-value <0.1 (16,648 trios) are listed in Supplementary Table 1, including trios containing transcripts with low mappability.

***Examples of mediation across tissues***

In analyses restricting to cis and trans genes with mappability scores >0.95, one biologically interpretable example of a cis-gene that appears to mediate the effects of a trans-eSNPs in multiple tissues is the *IFI44L* gene on chromosome 1 (Figure 3A). *IFI44L* is a cis-eGene in two GTEx tissues (cerebellar hemisphere and tibial nerve), and the cis-eSNPs associated with *IFI44L* expression are also associated with expression of multiple genes in trans in both cerebellar and tibial nerve tissue. *OAS1* is a trans target of these SNPs in both tissues, while other trans targets are observed in only cerebellar (*AGRN* and *PARP12*) or tibial nerve (*RSAD2*, *OAS2*, and *EPSTI1*). Below the mappability threshold of 0.95, we observe an additional potential trans targets of *IFI44L*, present in both cerebellar and tibial nerve tissue, IFIT3 (mappability=0.87). These relationships are depicted in Figure 3A.

Interestingly, if we expand our analysis to include cis and trans genes with mappability >0.90, we detect *IFI44* (mappability of 0.93) as a cis-mediator regulating a nearly identical similar set of trans genes across three tissues: cerebellar hemisphere (*OAS1*, *IFIT2*, *AGRN*, and *PARP12*), tibial nerve (*OAS1*, *IFIT3*, *RSAD2*, and *EPSTI1*), and sun-exposed skin (*IFIT1*) (Figure 3B). *IFI44* resides adjacent to *IFI44L* on 1p31.1, and these genes appear to be regulated by the same SNP in each tissue, making it unclear which of the two genes is truly a cis-mediator of the observed trans-eQTLs. *IFI44* and *IFI44L* are paralogs, so it is also possible that sequence similarity between these two genes causes our RNA-seq-based gene expression measurements for *IFI44* (and/or *IFI44L*) to reflect the expression variation of both genes to some extent. The causal cis-eSNP for *IFI44* (and/or *IFI44L*) appears to be different in different tissues, as the LD between the lead cis-eSNPs in cerebellar (rs12129932) and the lead eSNP in tibial nerve (rs74998911) is quite low ($r^2$<0.01 in EUR 1000 Genomes data).

Regardless of the uncertainty whether *IFI44L* or *IFI44* is the true cis-hub of this trans-eQTL, nearly all of the genes involved in the putative regulatory pathways identified here are interferon-regulated/inducible genes, namely *OAS1*, *OAS2*, *IFIT1*, *IFIT3*, *IFI44*, *IFI44L*, *RSAD2*, and *AGRN* [21, 22]. These genes

have been previously reported to be co-expressed and/or co-regulated in various human cell types, including interferon-exposed fibroblasts and mammary epithelial cell lines [21], virus-infected airway epithelial cells cultures [23], peripheral blood of individuals with acute respiratory infections [24] as well as in both normal and cancerous human tissue (Cancer Cell Metabolism Gene DB, https://bioinfo.uth.edu/ccmGDB/). This previously-reported co-expression findings also extend to *EPSTI1* [21], the one gene we find to be a trans-target of *IFI44L* (and/or *IFI44*) that does not have a well-established function in immune response, providing additional evidence of an immune-related function for this gene.

Variation in the *IFI44L* gene is associated with risk for MMR (measles, mumps, and rubella) vaccination-related febrile seizures, with a missense variant in *IFI44L* showing the strongest association [25]. Variation in *IFI44L* has also been implicated in schizophrenia risk [26] as well as bipolar disorder [27]. These findings suggest that the putative cross-tissue cis-hub identified here may be relevant to multiple neurological and psychological disorders, particularly those with etiologies related to immune function.

***Comparison of GMAC with other methods using simulated data***

We evaluate the performance of the proposed GMAC in various simulated data scenarios. For each scenario described below, we simulated 1000 mediation trios ($L_i$, $C_i$, $T_j$) for a sample size n = 350, similar to the sample size of the GTEx data. Each mediation trio consists of a gene transcript $i$ ($C_i$), its cis-associated genetic locus ($L_i$), and a gene transcript $j$ ($T_j$) in trans-association with the locus. Note that in the mediation analysis in this work (simulations and real data analysis), we consider only the trios with evidence of cis and trans- associations, $L_i \rightarrow C_i$ and $L_i \rightarrow T_j$. We are interested in testing whether an observed trans-eQTL association is mediated by the cis-gene transcript, i.e., $L_i \rightarrow C_i \rightarrow T_j$. We compared GMAC with other methods in different scenarios, including in the presence of confounders, common child variables, and intermediate variables. A common child variable is a variable that is affected by both

$C_i$ and $T_j$, and an intermediate variable is a variable that is affected by $C_i$ and affecting $T_j$, that is, at least partially mediating the effects from $C_i$ to $T_j$. Figure 1C illustrates a common child variable and Figure 1D shows an example of an intermediate variable.

*Comparison with other methods under the null in the presence of common child variables*

We first consider a scenario in which there is one common child variable for each pair of cis and trans gene transcripts (Figure 1C). In this scenario, adjusting common child variables in mediation analyses would 'marry' $C_i$ and $T_j$ and make $C_i$ appearing to be regulating $T_j$ even if there is no such effect (i.e., "collider bias") [28] increasing the false positive rate for detecting mediation.

We simulated a pool of independent and normally distributed variables **H**, with dimensionality being the same as sample size 350. Note that GMAC allows the dimensionality of candidate variables to be greater than sample size. For each of the 1000 mediation trios, we simulated the genetic locus $L_i$ under the Hardy-Weinberg Equilibrium assumption with a minor allele frequency of 0.1. Given $L_i$, the cis-gene transcript $C_i$ and trans-gene transcript $T_j$ are generated according to the models: $C_i = \beta_{i0c} + \beta_{i1c} L_i + \epsilon_{ic}$ and $T_j = \beta_{i0t} + \beta_{i1t} L_i + \epsilon_{it}$. In this scenario, the trans-effect is not mediated by the cis-gene transcript. We let the parameters in the above models vary across the 1000 trios with $\beta_{i1c}$ sampled uniformly from 0.5 to 1.5, and the rest sampled uniformly from 0.5 to 1.0. The error terms $\epsilon_{ic}$ and $\epsilon_{it}$ are normally distributed. For each mediation trio, one candidate variable in **H** is randomly chosen to be the common child variable, $Z_j$, and the effects of cis- and trans- gene transcripts on $Z_j$ are sampled uniformly from 1 to 1.5.

We compared the results based on the following methods: 1) Oracle adjustment, which correctly adjusts for no variables in **H** in the mediation test in this scenario; 2) The GMAC algorithm; and 3) Adjustment for child, which incorrectly adjusts for the common child variable, $Z_j$. Table 3A shows the true Type I

error rates at the significance levels of 0.01 and 0.05. As expected, adjusting for child would "marry" the cis- and trans-genes in the mediation test and would result in inflated rates of false positive findings.

*Comparison with other methods under the null in the presence of confounders*

We also consider a scenario in which the data is generated under the null in the presence of confounders (Figure 1B). Each candidate variable has a 95% probability of being an unrelated variable for all trios, and a 5% probability of being a confounder in the cis-trans genes relationship for a randomly chosen proportion of trios where the proportion follows a uniform distribution from 0 to 0.2. This specification results in on average 1.85 confounders for each trio in our simulated data. Suppose for the i$^{th}$ trio, there are $n_i$ number of variables in **H** selected to be confounders, which are $X_{i1}, \dots, X_{in_i}$. The cis-gene transcript $C_i$ and trans-gene transcript $T_j$ are generated according to the regression models $C_i = \beta_{i0c} + \beta_{i1c} L_i + \alpha_{i1}X_{i1} + \cdots + \alpha_{in_i}X_{in_i} + \epsilon_{ic}$ and $T_j = \beta_{i0t} + \beta_{i1t} L_i + \gamma_{i1}X_{i1} + \cdots + \gamma_{in_i}X_{in_i} + \epsilon_{it}$. In this scenario, there are no cis- to trans-gene mediation effects. We let the parameters in the above models vary across the 1000 trios with similar parameter specification as before.

We compared the results based on the following methods: 1) Oracle adjustment, which correctly adjusts for the true confounders in the mediation test in this scenario; 2) The GMAC algorithm; and 3) No adjustment, which incorrectly adjusts for no confounders in the mediation test. Table 3B showed that failure to adjust for confounding leads to an inflated type I error rates. In contrast, our proposed GMAC algorithm well controls the type I error rates. And 1761 out of 1847 generated confounders across the 1000 mediation trios are correctly selected in the above simulation setup.

*Comparison with other methods under the alternative in the presence of intermediate variables*

We consider another scenario in which there is one intermediate variable for each cis-trans relationship (Figure 1D). For each mediation trio, we simulated the genetic locus and the cis-gene transcript as before

and further simulated a child variable, $W_i$of the cis-gene transcript. The trans-gene transcript, $T_j$ is then simulated to be affected by $W_i$, according to $T_j = \beta_{i0t} + \beta_{i1t} L_i + \gamma_i W_i + \epsilon_{it}$. Therefore, the cis- affects the trans-gene transcript via the intermediate variable$W_i$, and the mediation effects from cis to trans gene transcript is non-zero in this scenario.

We compared the results based on the following methods: 1) Oracle adjustment, which correctly adjusts for zero variables in **H** in the mediation test in this scenario; 2) The GMAC algorithm; and 3) Adjustment for intermediate variable, which incorrectly adjusts for the corresponding intermediate variable in the mediation test. Table 3C shows that when the power to detect mediation is high (by Oracle and GMAC), incorrectly adjusting for an intermediate variable reduces the power to detect mediation. In comparison, GMAC correctly filters out most of the true intermediate variables in the mediation tests, and maintains power comparable to oracle adjustment.

*Comparison with other methods under the alternative in the presence of confounders*

To compare with the existing approach that adjusts for a universal set of variables, we consider a scenario in which the dimensionality of candidate variables **H** is 100. For each trio, up to five candidate variables are randomly selected to confound the cis trans gene relationship. We set the effect of cis transcript on trans transcript to be 0.1, i.e., non-zero mediation effects.

We compared the results based on the following methods: 1) Oracle adjustment, which correctly adjusts for the true confounders in the mediation test; 2) The GMAC algorithm; and 3) Universal adjustment, which adjusts for all variables in **H** in the mediation tests for all trios. Table 3D shows that GMAC has comparable or even better power than oracle adjustment in this scenario. This is because that the unidentified and therefore unadjusted confounders (2908 out of 3023 generated confounders across the 1000 mediation trios are correctly selected in the above simulation setup) may generate a positive

association between $C_i$ and $T_j$. That could strengthen the positive link from $C_i$ to $T_j$ and could result in higher power to detect mediation. In comparison, adjusting for all variables in the pool of confounders is inefficient and reduces power to detect mediation.

## DISCUSSION

In this paper, we have developed the GMAC algorithm for conducting mediation analysis to identify cis-transcripts that mediate the effects of trans-eQTLs on distant genes. We address a central problem in mediation analysis, "mediator-outcome confounding", by developing an algorithm that can a) search a very large pool of variables (surrogate and/or measured) for variables likely to have confounding effects and b) adaptively adjust for such variables in each mediation test conducted. Analyses of simulated data show that the GMAC algorithm improves the power to detect true mediation compared with existing methods, while controlling the true false discovery rate. We have applied this method to gene expression data from 44 human tissues from the GTEx Project, allowing us to identify genes that mediate the effects of trans-eQTLs in multiple tissue types. Over 20% of cis-mediators we observe appear to mediate the effects of a trans-eQTL on multiple genes, but the vast majority of these cis-hubs are either tissue-specific (i.e., mediating multiple trans-genes in a single tissue type) or have unique trans targets in each tissue type. We provided one example of a biologically plausible multi-tissue cis-hub, whereby a cis-mediator of a trans-eQTLs appears to have common trans targets across multiple tissue types. The cis-hub identified (*IFI44L*) has potential relevance for neurological and psychological disorders, particularly those with etiologies related to immune function, demonstrating the potential value of our approach for understanding disease-relevant pathways.

One innovative aspect of this work is our algorithm that rigorously addresses the problem of "mediator-outcome confounding" in the context of genomic mediation analysis. In eQTL-based mediation analysis, potential confounders of the cis-trans association include demographic and environmental factors, as well

as a wide array of biological phenomenon, such as expression of specific genes or other biological processes that may be represented by the expression of sets of genes. Neglecting to control for such confounding variables can lead to substantial bias in estimates of mediation, resulting in spurious findings, as we have described previously [1]. Considering the complexity of the biological systems under study, as well as the diversity of the GTEx donors, a careful control for such confounding variables is critically important.

Most existing methods control for confounding variables by constructing a set of variables that represent the largest components of variation in the transcriptome and adjusting for the selected set for all tests conducted. In contrast, GMAC adaptively selects a set of confounding variables for each mediation trio, enabling large-scale genomic mediation analyses adjusting only for the confounding variables that could potentially bias a specific mediation estimate. The strategy of selecting only potential confounders for adjustment purposes is important (as opposed to adjusting for all known covariates) for three reasons: 1) Adjusting fewer variables (i.e., fewer degrees of freedom; see Supplementary Table 2) increases power; 2) The number of variables from which one selects covariates could be extremely large (e.g., all expressed genes), making adjustment for all covariates impossible, and 3) inadvertently adjusting for "common child" or intermediate variables can result in substantial biases. In this work we select potential confounders from all expression PCs, but one could also select from among transcripts that are not well-represented by PCs. By efficiently selecting confounders from a very large pool of potential variables, GMAC improves both power and precision in mediation analyses.

There are several limitations of our approach and its application to GTEx data. First, when working with real genomic data, we can never be sure that we have measured and accounted for all possible mediator-outcome confounding. Potential confounders include participant characteristics, environmental factors, tissue micro-environmental factors, as well as a wide array of biological factors which may or may not be captured by the expression data being analyzed. Second, in the analysis presented here, we only consider

the trios with both strong cis- and trans-eQTL effects. For any given tissue type we are analyzing, the sample size is small for robust genome-wide detection/analysis of trans-eQTLs. As such, the mediation trios we considered are only a subset of the true mediation trios in the genome. And the small sample sizes may also result in underpowered mediation tests. As the sample size of GTEx increases, future studies will have increased power to identify cis-mediators using GMAC. Third, for some of the trios we analyze for mediation, the causal variant for the cis-eQTL may not be the causal variant underlying the trans-eQTL. Rather the causal variants may be in close proximity to one another and in LD. In these cases, the power to detect mediation could be low compared to analyses of the true causal variant. Fourth, we did not consider the full complexity of gene isoforms and splice variants in this work; future studies should consider the possibility of mediation relationships that are isoform-specific. Lastly, some trans-eQTLs may not be mediated by variation in the expression of a cis-gene. Other potential mediating mechanism could include variation in coding sequence, physical inter-chromosomal interaction, or variation in non-coding RNA. Our work is not intended to identify and analyze such trans-eQTLs, as we perform trans-eQTL analyses using only SNPs known to be cis-eSNPs.

It is important to note that our expectation is that most trans-eQTLs are fully-mediated by a transcript that is regulated in cis by the causal trans-eQTL variant. We did not observe "complete mediation" (i.e., % mediation = 100%) for the majority of the significant mediation P values we observed. However, as we have explained and demonstrated previously [1], full mediation will be observed as partial mediation in the presence of mediator measurement error and/or imperfect LD between the causal variant and the variant used for analysis purposes. Thus, considering RNA quantification is not error free and causal variants are often unknown, we expect to often observe partial mediation when full mediation is present.

We also demonstrate that it is critical to consider mappability for both cis and trans genes involved in mediation analysis. For genes containing sequences that do not uniquely map to the human transcriptome, it is possible that gene expression measures may be comprised of signals coming from multiple genes,

which can produce false positives in mediation analysis, including spurious detection of cis-hubs and cross-tissue cis-mediators.

We have developed R GMAC package to perform the proposed genomic mediation analysis with adaptive selection of confounding variables. It tests for mediation effects for a set of user specified mediation trios or all the (eQTL, cis and trans gene) trios in the genome; it considers either a user provided pool of potential confounding variables, real or constructed by other methods, or all the PCs based on expression data as the potential confounder pool; and it returns mediation p-values and provides diagnostic checks of model assumptions. The software will be available through R CRAN.

Our application of the GMAC algorithm to the multi-tissue expression data from GTEx provides a unique cross-tissue perspective on cis-mediation of trans-regulatory relationships across human tissues. This mutli-tissue perspective is important because observing mediation relationships that are consistent across multiple tissues provides confidence that a significant mediation P-value reflects a true instance of mediation. For the "cis-hub" genes and genes that appear to be cis-mediators in multiple tissues, further investigation is warranted, as these genes may have many regulatory relationships that we are not powered to detect in this work. Thus, a multi-tissue mediation analysis approach has the potential to increase power to identify true mediators while controlling for false positives. In future work, attempts at joint analyses of multiple tissue types may provide a more complete picture of the cross-tissue and tissue-specific trans-regulatory mechanisms. The GMAC approach described here will be a valuable tool for such studies, as well any future studies that aims to understand the relationships among cis- and trans-eQTLs and characterize the biological mechanisms and networks involved in human disease biology.

# METHODS

***Bio-specimen collection and processing of GTEx data***

A total of 7,051 tissues samples were obtained from 44 distinct tissue types from 449 post-mortem tissue donors. Among donors, 65.6% were male and 34.4% were female. They were from multiple ethnicity groups with 84.3% white, 13.7% African American, 1% Asian, and 1% unreported ethnicity. Those donors spanned a wide age range (20-70 years). More than half of the donors died from traumatic injury and these individuals tended to be of younger ages. Donor enrollment and consent processes have been described elsewhere [10, 11]. Biospecimen collection and processing has been described previously in detail [10, 11]. Briefly, each tissue sample was preserved in PAXgene tissue kit and the stored as both frozen and paraffin embedded tissue. Total RNA was isolated from PAXgene fixed tissues samples using the PAXgene Tissue mRNA kit. For whole blood, Total RNA was isolated from samples collected and preserved in PAXgene blood RNA tubes.

Blood samples were used as the primary source of DNA. Genotyping was conducted using the Illumina Human Omni5-Quad and Infinium ExomeChip arrays. Standard QC procedures were performed using the PLINK software [29] and genotype imputation was performed using the IMPUTE2 software [30] and reference haplotypes from the 1000 Genomes Project. Principal components (PC) analysis was used to generate variables representing ancestry [16]. The first three PCs were included as covariates in all analyses, and these were sufficient to represent the major population groups present in the GTEx dataset (Caucasian, African American, and Asian individuals).

***Quantification of gene expression levels***

RNA-seq data was generated for RNA samples with a RIN value of 6 or greater. Non-strand specific RNA sequencing was performed using an automated version of the Illumina TruSeq RNA sample

preparation protocol. Sequencing was done on an Illumina HiSeq 2000, to a median depth of 78M 76 bp paired-end reads per sample.

RNA-seq data was aligned to the human genome using Tophat. Gene-level and exon-level expression was estimated in RPKM units using RNA-SeQC. Only gene-level expression values were used for this work. RNA-seq expression samples that passed various quality control measures (as previously described) were included in the final analysis dataset.

***Mappability of transcripts***

Because non-uniquely mapping reads can result in false positive eQTLs, we use the mappability of each gene as a quality control filter, as described in Jo et al (the GTEx "trans paper"). The mappability was calculated as follows: Mappability of all k-mers in the reference human genome (hg19) computed by ENCODE [31] was downloaded from the UCSC genome browser (accession: wgEncodeEH000318, wgEncodeEH00032) [32]. The exon- and UTR-mappability of a gene was computed as the average mappability of all k-mers in exons and UTRs, respectively. We used k=75 for exonic regions, as it is the closest to GTEx read length among all possible k's. UTRs are generally quite small, so k=36 was used, the smallest among all possible k's. Mappability of a gene was computed as the weighted average of its exon-mappability and UTR-mappability, with the weights being proportional to the total length of exonic regions and UTRs, respectively.

***The selection of trios for mediation tests***

In the genomic mediation analysis presented in this work, we consider only the trios with evidence of cis and trans- associations, $L_i \rightarrow C_i$ and $L_i \rightarrow T_j$. We allow the eQTL to affect the trans-gene transcript via other pathways independent of the cis gene. Figure 1B illustrates the type of mediation relationship we would like to detect. Since genetic loci are 'Mendelian randomized' [33], without loss of generality we assume the confounders are not associated with $L_i$.

For each GTEx tissue type, we used the gene-level RPKM (from RNA-SeQC) as the gene expression values for each gene. We identified the cis-eQTLs using standard methods, restricting to genes for which at least 10 samples had RPKM > 0.1 and raw read counts >6 (cite Aguet et al. 2016, GTEx companion paper, unpublished) [10]; the complete cis-eQTL list is available through dbGaP. For genes with significant evidence of a cis-eQTL, we then selected one cis-eSNP for each gene (i.e., the high-quality SNP with the smallest P-value) and only those cis-eSNPs were included in the subsequent trans-eQTL and mediation analyses, as we require the presence of both cis- and trans- associations for testing mediation. For each tissue, we conducted genome-wide trans-eQTL analyses restricting to the cis-eSNPs described above and examining association for all genes located at least 1Mb away from the cis-eSNPs. For mediation analyses, we only considered the pairs of eQTL and trans transcript with suggestive trans-associations at the p-values threshold of $10^{-5}$ in the specific tissue type. While some false positives trans-eQTLs will pass this threshold, it will also allow for mediation analysis of true trans-eQTLs that cannot be detected at more stringent thresholds due to our limited sample size.

In both the cis- and trans-eQTL analyses, for the tissue types with sample sizes greater than 250, thirty-five tissue-specific PEER factors were constructed and adjusted; for the tissue types with sample sizes between 150 to 250, thirty PEER factors were adjusted; and for the tissue types with sample size less than 150, fifteen PEER factors were adjusted. To identify trans-eQTLs, we estimated the association for each cis-eSNP with expression of all genes at least 1 Mb away from the SNP using the Matrix eQTL software [19].

***Adaptive filtering to eliminate variables from the pool of potential confounders***

Let **H** be the pool of candidate confounding variables (constructed or real variables). Unlike SVA [17], PCA [16] or PEER factor analysis [18] that use the top constructed variables, here we consider a full set of confounder variables and the dimensionality of **H** may exceed the sample size (for example, when

there are a large number of real covariates or we may consider others genes in the genome as potential confounders for each trio).

For each trio ($L_i$, $C_i$, $T_j$), we propose to first filter out potential common child variables, $\mathbf{Z_{ij}}$, of $C_i$ and $T_j$, and intermediate variables, $\mathbf{W_{ij}}$, from $C_i$ to $T_j$. The confounding variables, common child variables and intermediate variables share a commonality --- they are correlated with both $C_i$ and $T_j$. However, adjusting common child variables in mediation analyses would 'marry' $C_i$ and $T_j$ and make $C_i$ appearing to be regulating $T_j$ even if there is no such effect (i.e., "collider bias") [28] increasing the false positive rate for detecting mediation. Adjusting for intermediate variables in a test for mediation would prevent the detection of the true mediation effect from $C_i$ to $T_j$ and hurt the power to detect true mediation. Existing methods to select confounders are often based on the correlation between each candidate variable and gene expression levels (pairs of cis- and trans-genes or the expression data matrix), which would not distinguish confounders and common child/intermediate variables.

We argue that it is possible to filter common child and intermediate variables by utilizing the randomness in the inheritance of genetic loci. Given the cis-association ($L_i \rightarrow C_i$), both common child and intermediate variables are affected by the cis-gene transcript, and as such are associated with the locus, $L_i$. On the other hand, the confounders are assumed to be not associated with $L_i$, since the genotypes are Mendelian randomized [33]. Therefore, for each trio we propose to filter the variables that are associated with $L_i$, at a liberal significance threshold of 10% FDR [20] from the pool **H**, and only consider the retained variables in the subsequent adaptive adjustment and mediation test.

***Adaptive selection of potential confounding variables***

Different mediation trios may be affected by a different subset of variables in **H**. We propose to adaptively select the confounder sets for each trio using a stratified FDR approach [34].

Specifically, for each trio, we first obtain the p-values of association for each candidate variable to the pair of expression levels, $C_i$ and $T_j$. Here the candidate variables are the ones retained after filtering child and intermediate variables specific to the trio. We use a linear regression with each candidate variable as the response and $C_i$ and $T_j$ as predictors, and obtain the p-value of the overall F-test for testing whether the candidate variable is associated with at least one of the cis and the trans-gene expression levels.

For each candidate confounding variable we then apply a predefined FDR threshold (5%) to the p-values corresponding to the joint associations of this variable to all the potential mediation trios, and we select the significant ones. We repeat this procedure for all candidate variables. Note that a confounder would be associated with both the cis- and trans-gene transcript. By using an F-test to test the joint association to either cis- or trans-gene, we obtain a superset of the confounder set. In calculating the FDR, a key parameter to be estimated is $\pi_0$, the proportion of true null hypotheses. For real variables, we estimate $\pi_0$ using the R qvalue package [20]. For PCA analysis, we estimated $\pi_0$ as one minus the percentage of variation each PC explained in the overall expression matrix.

As shown in Figure 1E, by applying the same FDR threshold to each candidate confounding variable to all trios, we identified the significant "pair"-wise associations of candidate confounders to mediation trios for all variables and all trios. It can be shown that under pertinent assumptions (confounders being independent of each other), the overall FDR is controlled at the FDR significance level. When performing the mediation test for each trio, we propose to consider only the subsets in **H** that are significantly associated with that trio.

***Mediation test and p-value calculation***

In our genomic mediation analysis, we consider only the trios with evidence of both a cis- and trans-eQTL associations. Consider one potential mediation trio, ($L_i$, $C_i$, $T_j$) with the adaptively selected set of potential confounding variables for this trio, $\mathbf{X}_{ij}$. Here $\mathbf{X}_{ij}$ is a subset of variables in **H**, often with much

lower dimensionality. Those variables are significantly associated with at least one of $C_i$ and $T_j$ (i.e., the ones checked in the row corresponding to the trio in Figure 1E.)

Given cis- and trans-associations, we propose to test for non-zero mediation effects from the cis-gene transcript to the trans-gene transcript based on the following regression:

$$T_j = \beta_0 + \beta_1 C_i + \beta_2 L_i + \boldsymbol{\Gamma} \mathbf{X_{ij}} + \varepsilon \qquad (1)$$

We are interested in testing non-zero mediation effects captured by $\beta_1$. Under the null hypothesis, after adjusting for confounders, there is no mediation effect given the effect from $L_i$ to $C_i$ and the direct effect from $L_i$ to $T_j$. Under the alternative, in the presence of effect from $L_i$ to $C_i$ and the potential direct effect from $L_i$ to $T_j$ (the dashed arrow in Figure 1B), the effect from $C_i$ to $T_j$ is non-zero, (i.e., $\beta_1 \neq 0$). We can obtain the Wald statistic for testing $\beta_1$ from the regression as the mediation test statistic.

To calculate the p-value for mediation for each trio, we propose to permute the cis-gene expression levels within each genotype group and obtain the null mediation statistics based on the trios with the same locus and trans- gene but permuted cis-expression levels, ($L_i$, $C_{i0}$, $T_j$). We assume that confounding effects have been well adjusted. Given cis- and trans-associations, under the null there is no mediation. By permuting the cis-gene expression levels within each genotype group, one maintains the cis-associations while breaks the potential mediation effects from the cis- to the trans-gene transcript. That is, conditioning on the genetic locus, the permuted cis-gene expression is not correlated with the trans-gene expression levels, i.e., no mediation. Figure 1F shows the expression variation patterns of a hypothetical mediation relationship $L_i \rightarrow C_i \rightarrow T_j$ on the left panel, and a null relationship entailed by ($L_i$, $C_{i0}$, $T_j$) with $L_i \rightarrow C_{i0}$ and $L_i \rightarrow T_j$ but no mediation. A p-value of mediation is calculated for each trio by comparing the observed statistic versus the null ones.

***A summary of the GMAC algorithm***

In summary, in order to identify cis-mediators of trans-eQTLs across the entire genome, we propose the GMAC algorithm. Specifically,

- Step 0. We focus on only the trios ($L_i$, $C_i$, $T_j$) in the genome showing both cis- and trans-eQTL associations, i.e., $L_i \rightarrow C_i$ and $L_i \rightarrow T_j$.
- Step 1. Filter common child and intermediate variables. Given a pool of candidate variables **H** consisting of either real covariates, constructed surrogate variables, or both, for each trio ($L_i$, $C_i$, $T_j$) we calculate the marginal associations of variables in **H** to $L_i$ and filter the ones with significant associations at the 10% FDR level. As shown in Figure 1B-D, common child and intermediate variables are directly associated with $L_i$, while confounders are assumed to be unassociated with $L_i$. Let $\mathbf{H_{ij}}$ denote the retained pool of candidate variables specific to the trio ($L_i$, $C_i$, $T_j$).
- Step 2. Adaptively select confounders. For each trio and each of its potential confounding variables in $\mathbf{H_{ij}}$, we calculate the p-value of the F-test to assess the association of the variable to at least one of the cis- and trans- transcripts. Considering the p-values for one potential confounding variable to all trios as one stratum, we apply a 5% FDR significance threshold to each stratum (each column in Figure 1E). The significant variables corresponding to a trio (each row in Figure 1E) will be selected in the mediation analyses as the adaptively selected confounders specific to that trio. Let $\mathbf{X_{ij}}$ denote the list of adaptively selected confounder variables for the trio, ($L_i$, $C_i$, $T_j$).
- Step 3.Test for mediation. For each trio and its adaptively selected confounder set, we calculate the mediation statistic as the Wald statistic for testing the indirect mediation effect $H_0$: $\beta_1 = 0$ based on the regression entailed by equation (1). We perform within-genotype group permutation on the cis-gene transcript at least 10,000 times and re-calculate each null mediation statistic based on the locus, a permuted cis-gene transcript, and the trans-gene transcript, ($L_i$, $C_{i0}$, $T_j$). We calculate the *p*-value of mediation for the trio ($L_i$, $C_i$, $T_j$) by comparing the observed mediation statistic with the null statistics.

The proposed algorithm is superior to existing approaches for mediation analysis that adjust a universal set of variables for all trios. GMAC avoids the adjustment of common child variables, intermediate variables and unrelated variables in genomic mediation analysis, and it is able to search a much larger pool of variables for potential confounders, not just those captured by the top few surrogate variables or PCs.

# Tables

**Table 1. A description of GTEx tissue types and the number of significant instances of mediation (i.e., SNP-cis-trans trios) identified by GMAC (Genomic Mediation Analysis with adaptive Confounder adjustment).**

| Tissue name | Tissue sample size | # Trios tested | # Trios with suggestive mediation (P <0.05) | # Trios significant at 5% FDR |
|---|---|---|---|---|
| Muscle Skeletal | 361 | 2387 | 496 | 264 |
| Whole Blood | 338 | 2274 | 508 | 281 |
| Skin Sun Exposed Lower leg | 302 | 3273 | 629 | 330 |
| Adipose Subcutaneous | 298 | 3332 | 640 | 325 |
| Artery Tibial | 285 | 2699 | 527 | 281 |
| Lung | 278 | 2762 | 543 | 323 |
| Thyroid | 278 | 3894 | 696 | 376 |
| Cells Transformed fibroblasts | 272 | 3000 | 642 | 340 |
| Nerve Tibial | 256 | 3812 | 677 | 326 |
| Esophagus Mucosa | 241 | 2640 | 465 | 242 |
| Esophagus Muscularis | 218 | 2431 | 447 | 230 |
| Artery Aorta | 197 | 2009 | 368 | 186 |
| Skin Not Sun Exposed Suprapubic | 196 | 1961 | 365 | 177 |
| Heart Left Ventricle | 190 | 1290 | 242 | 115 |
| Adipose Visceral Omentum | 185 | 1410 | 257 | 125 |
| Breast Mammary Tissue | 183 | 1422 | 254 | 126 |
| Stomach | 170 | 1153 | 235 | 107 |
| Colon Transverse | 169 | 1585 | 309 | 161 |
| Heart Atrial Appendage | 159 | 1221 | 243 | 103 |
| Testis | 157 | 3896 | 607 | 267 |
| Pancreas | 149 | 1270 | 208 | 102 |
| Esophagus Gastroesophageal Junction | 127 | 857 | 148 | 74 |
| Adrenal Gland | 126 | 981 | 185 | 94 |
| Colon Sigmoid | 124 | 968 | 220 | 108 |
| Artery Coronary | 118 | 874 | 191 | 95 |
| Cells EBV-transformed lymphocytes | 114 | 856 | 152 | 78 |
| Brain Cerebellum | 103 | 1295 | 187 | 84 |
| Brain Caudate basal ganglia | 100 | 763 | 139 | 61 |
| Liver | 97 | 496 | 87 | 41 |
| Brain Cortex | 96 | 754 | 134 | 45 |
| Brain Nucleus accumbens basal ganglia | 93 | 592 | 97 | 43 |

| | | | | |
|---|---|---|---|---|
| Brain Frontal Cortex BA9 | 92 | 595 | 102 | 52 |
| Brain Cerebellar Hemisphere | 89 | 1072 | 222 | 116 |
| Spleen | 89 | 825 | 157 | 68 |
| Pituitary | 87 | 732 | 132 | 61 |
| Prostate | 87 | 474 | 101 | 54 |
| Ovary | 85 | 469 | 95 | 43 |
| Brain Putamen basal ganglia | 82 | 481 | 94 | 35 |
| Brain Hippocampus | 81 | 343 | 93 | 47 |
| Brain Hypothalamus | 81 | 342 | 74 | 41 |
| Vagina | 79 | 248 | 58 | 25 |
| Small Intestine Terminal Ileum | 77 | 434 | 82 | 39 |
| Brain Anterior cingulate cortex BA24 | 72 | 365 | 81 | 29 |
| Uterus | 70 | 287 | 62 | 25 |

**Table 2.** Frequency of cis-genes that mediate the effect of a trans-eQTL on multiple trans-genes or in multiple tissue types

| **Observed number of trans targets for each cis-gene** | | **Number of tissues for which each cis-gene is a mediator** | |
|---|---|---|---|
| Number of trans targets | Cis-gene count | Number of Tissues | Cis-gene count |
| 1 | 2,510 | 1 | 2,637 |
| 2 | 482 | 2 | 420 |
| 3 | 123 | 3 | 91 |
| 4 | 33 | 4 | 16 |
| 5-6 | 12 | 5 | 3 |
| 7-14 | 8 | 6 | 1 |

**Table. 3.** Comparison of the Type I error rate and power of GMAC compared to other methods for mediation analysis under the null (A and B) and the alternative (C and D) hypotheses, based on simulated data

A. Type I error in the Presence of a Common Child

| Significance Level | Oracle Adjustment | GMAC Algorithm | Child Adjustment |
|---|---|---|---|
| 0.01 | 0.011 | 0.010 | 0.287 |
| 0.05 | 0.049 | 0.050 | 0.413 |

B. Type I error in the Presence of Confounders

| Significance Level | Oracle Adjustment | GMAC Algorithm | No Adjustment |
|---|---|---|---|
| 0.01 | 0.007 | 0.008 | 0.459 |
| 0.05 | 0.045 | 0.048 | 0.585 |

C. Power in the Presence of an Intermediate variable

| Significance Level | Oracle Adjustment | GMAC Algorithm | Adjusting Intermediate variable |
|---|---|---|---|
| 0.01 | 0.999 | 0.868 | 0.006 |
| 0.05 | 0.999 | 0.871 | 0.041 |

D. Power in the Presence of Confounders

| Significance Level | Oracle Adjustment | GMAC Algorithm | Adjusting All |
|---|---|---|---|
| 0.01 | 0.231 | 0.260 | 0.158 |
| 0.05 | 0.459 | 0.486 | 0.341 |

# Figures

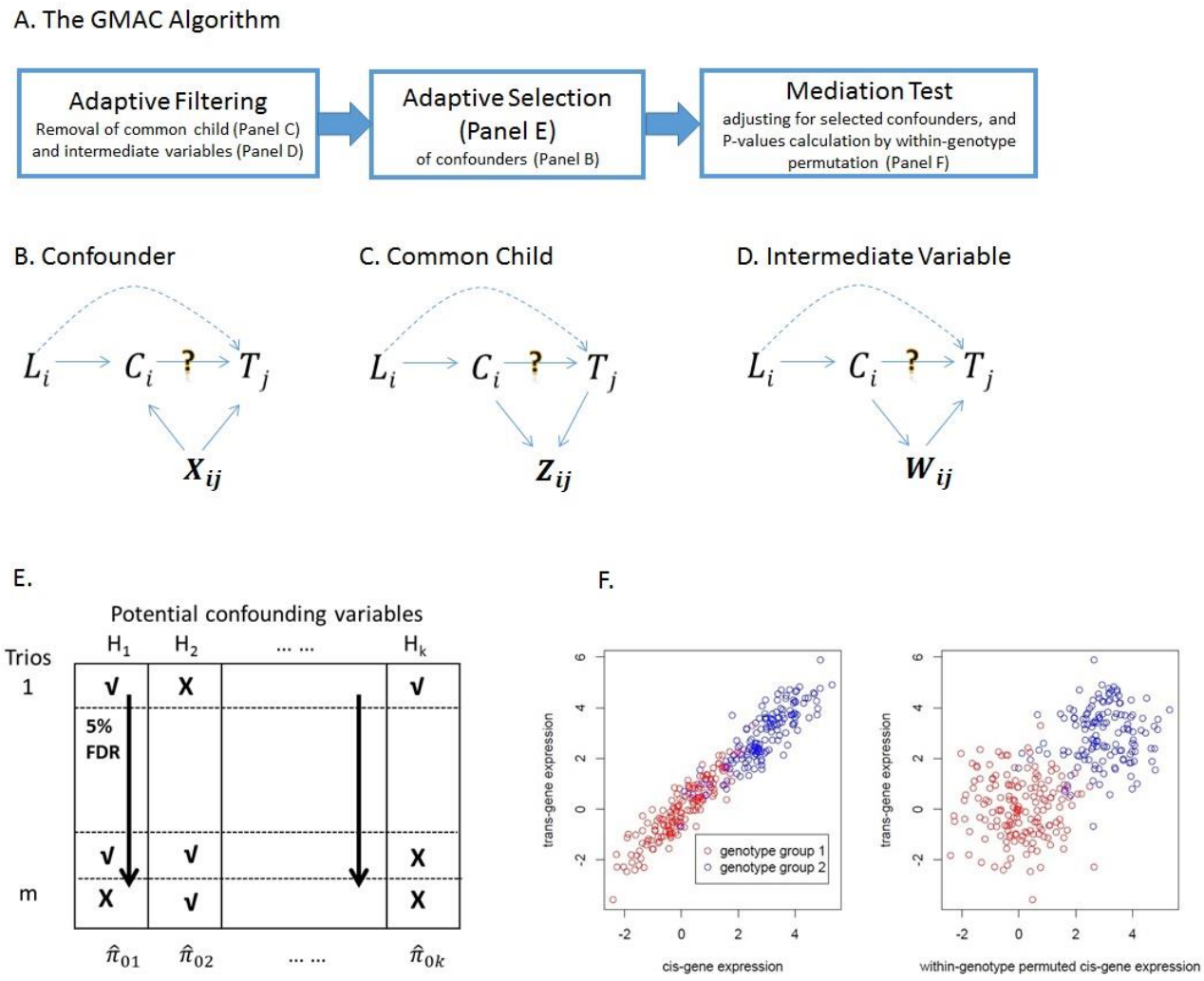


**Figure 1.** Graphical illustrations of (A) a summary of the GMAC algorithm; (B) a mediation relationship among an eQTL, $L_i$, its cis-gene transcript, $C_i$, and a trans-gene transcript, $T_j$, with confounders, $\mathbf{X_{ij}}$, allowing $L_i$ to affect $T_j$ via a pathway independent of $C_i$; (C) a mediation trio where $C_i$, and $T_j$, have common child variable(s), $\mathbf{Z_{ij}}$; (D) a mediation trio where $C_i$, affects $T_j$, through intermediate variable(s), $\mathbf{W_{ij}}$. (E) The adaptive confounder selection procedure: Based on the p-value matrix for the association of each potential confounder variable to at least one of the cis- or the trans-gene transcript, we apply a stratified FDR approach by considering the p-values for each potential confounder (each column) as a stratum, with the significant ones indicated by a check mark (√). When conducting the mediation test for each trio, we only adjust for the significant confounding variables (the ones with with √ in each row). (F) A mediation trio $L_i \rightarrow C_i \rightarrow T_j$ (left) and a trio under the null with both cis-linkage and trans-linkage but

no mediation (right). Within-genotype permutation of the cis-gene expression levels maintains the cis- and trans-linkage (different mean levels) while breaks the potential correlation between the cis- and trans-expression levels within each genotype group.

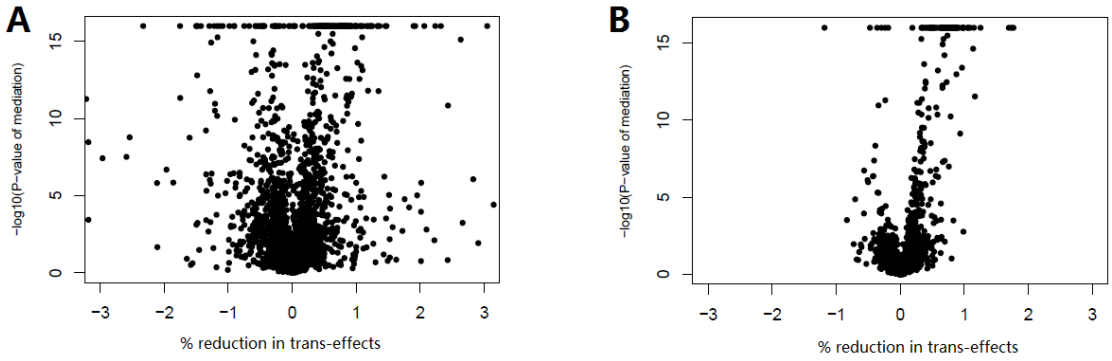


**Figure 2.** Plots of negative log base 10 of mediation p-values versus the percentage of reduction in trans-effects after accounting for cis-mediation, based on (A) mediation tests without adjusting for hidden confounders (B) mediation tests by GMAC considering all PCs as potential confounders. The percentage of reduction in trans-effects is calculated by $(\beta_2^m - \beta_2)/\beta_2^m$, where $\beta_2^m$is the marginal trans-effect of the eQTL on the trans-gene expression levels, and $\beta_2$ is the trans-effect in equation (1) which is adjusted for a potential cis-mediator. For trios with true cis-mediations, we expect the trans-effects to be substantially reduced after adjusting for the true cis-mediator. That is, we expect the trios with very significant mediation p-values to have positive % reduction in trans-effects. For results based on no adjustment of hidden confounders (A), we observed many trios with significant mediation p-values but the percentages of reduction in trans-effects are often negative. Those are suspected false positives. P-values are truncated at $10^{-16}$. The plots are based on the results from the Adipose Subcutaneous tissue.

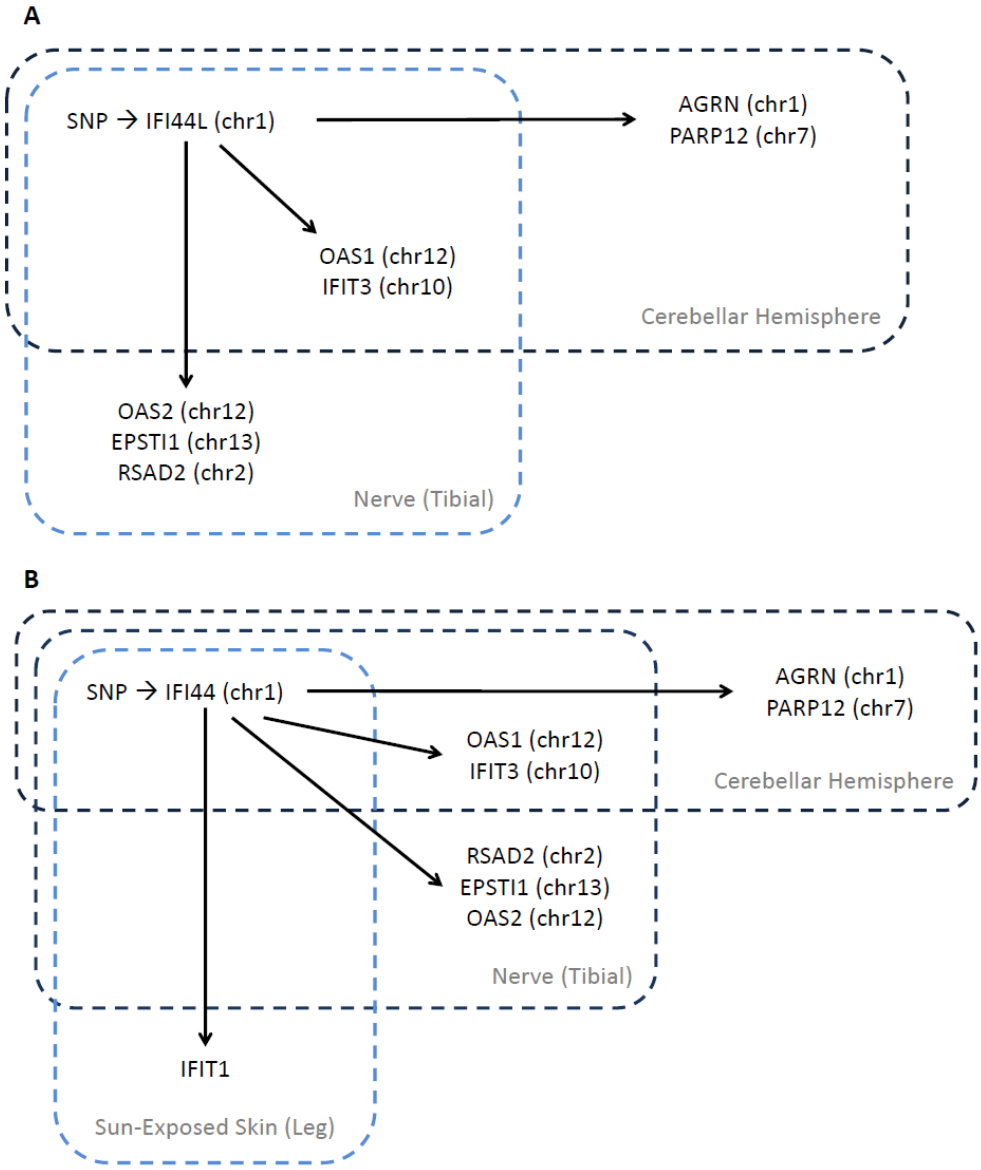


**Figure 3. A biologically interpretable example of a cis-eGene (*IFI44L*) that appears to mediate the effects of trans-eSNPs in multiple tissues.** *IFI44L* (Panel A) resides <5kb away from *IFI44* (B), and expression of these genes is associated with a common cis-eQTL that also impacts the expression of multiple genes in trans in multiple tissues. Both *IFI44* and *IFI44L* show statistical evidence of mediation for a similar set of interferon-related genes. Thus, based on this evidence, we infer that at least one of these genes is a cis-mediator, although cannot know which is (or if both are) the true mediator.

# REFERENCE


1. Pierce, B.L., et al., *Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians.* PLoS Genet, 2014. **10**(12).
2. Fehrmann, R.S.N., et al., *Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA.* PLoS Genet, 2011. **7**(8).
3. Stranger, B.E., et al., *Patterns of Cis Regulatory Variation in Diverse Human Populations.* PLoS Genet, 2012. **8**(4): p. 272-284.
4. Chen, L.S., F. Emmert-Streib, and J.D. Storey, *Harnessing naturally randomized transcription to infer regulatory relationships among genes.* Genome Biol, 2007. **8**(10).
5. Veyrieras, J.B., et al., *High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation.* PLoS Genet, 2008. **4**(10).
6. Nicolae, D.L., et al., *Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS.* PLoS Genet, 2010. **6**(4).
7. Westra, H.J., et al., *Systematic identification of trans eQTLs as putative drivers of known disease associations.* Nat Genet, 2013. **45**(10): p. 1238-U195.
8. Torres, J.M., et al., *Cross-Tissue and Tissue-Specific eQTLs: Partitioning the Heritability of a Complex Trait.* Am J Hum Genet, 2014. **95**(5): p. 521-534.
9. Wang, J.B., et al., *Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx.* Am J Hum Genet, 2016. **98**(4): p. 697-708.
10. Ardlie, K.G., et al., *The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-660.
11. Lonsdale, J., et al., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-585.
12. Battle, A., et al., *Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.* Genome Res, 2014. **24**(1): p. 14-24.
13. Robins, J.M. and S. Greenland, *Identifiability and exchangeability for direct and indirect effects.* Epidemiology, 1992. **3**(2): p. 143-55.
14. Cole, S.R. and M.A. Hernan, *Fallibility in estimating direct effects.* Int J Epidemiol, 2002. **31**(1): p. 163-5.
15. Pearl, J. *Direct and Indirect effects*. in *the 17th Conference on Uncertainy in Artificial Intelligence*. 2001. San Francisco, CA: Morgan Kaufmann.
16. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-909.
17. Leek, J.T. and J.D. Storey, *Capturing heterogeneity in gene expression studies by surrogate variable analysis.* PLoS Genet, 2007. **3**(9): p. 1724-1735.
18. Stegle, O., et al., *Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses.* Nature Protocols, 2012. **7**(3): p. 500-507.
19. Shabalin, A.A., *Matrix eQTL: ultra fast eQTL analysis via large matrix operations.* Bioinformatics, 2012. **28**(10): p. 1353-1358.
20. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-9445.
21. Cheon, H. and G.R. Stark, *Unphosphorylated STAT1 prolongs the expression of interferon-induced immune regulatory genes.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9373-9378.

22. Kyogoku, C., et al., *Cell-Specific Type I IFN Signatures in Autoimmunity and Viral Infection: What Makes the Difference?* PLoS One, 2013. **8**(12).
23. Ioannidis, I., et al., *Plasticity and Virus Specificity of the Airway Epithelial Cell Immune Response during Respiratory Virus Infection.* J Virol, 2012. **86**(10): p. 5422-5436.
24. Zaas, A.K., et al., *Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans.* Cell Host & Microbe, 2009. **6**(3): p. 207-217.
25. Feenstra, B., et al., *Common variants associated with general and MMR vaccine-related febrile seizures.* Nat Genet, 2014. **46**(12): p. 1274-82.
26. Ruderfer, D.M., et al., *Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia.* Mol Psychiatry, 2014. **19**(9): p. 1017-24.
27. Chen, D.T., et al., *Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder.* Mol Psychiatry, 2013. **18**(2): p. 195-205.
28. Greenland, S., *Quantifying biases in causal models: classical confounding vs collider-stratification bias.* Epidemiology, 2003. **14**(3): p. 300-6.
29. Purcell, S., et al., *PLINK: A tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-575.
30. Howie, B.N., P. Donnelly, and J. Marchini, *A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies.* PLoS Genet, 2009. **5**(6).
31. *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
32. Rosenbloom, K.R., et al., *ENCODE data in the UCSC Genome Browser: year 5 update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D56-63.
33. Smith, G.D. and S. Ebrahim, *'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?* Int J Epidemiol, 2003. **32**(1): p. 1-22.
34. Sun, L., et al., *Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies.* Genet Epidemiol, 2006. **30**(6): p. 519-30.