bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# TYPES OF CIS- AND TRANS-GENE REGULATION OF EXPRESSION QUANTITATIVE TRAIT LOCI ACROSS HUMAN TISSUES

JARRED KVAMME[1†], MD BAHADUR BADSHA[2,6†], EVAN A. MARTIN[1,7], JIAYU WU[3], MOHAMED MEGHEIB[1], XIAOYUE WANG[3], AUDREY QIUYAN FU[1,2,4,5*]

[1]*The Bioinformatics and Computational Biology Program, University of Idaho;* [2] *Institute for Modeling Collaboration and Innovation, University of Idaho;* [3] *Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences;* [4] *Institute for Interdisciplinary Data Sciences, University of Idaho;* [5] *Department of Mathematics and Statistical Science, University of Idaho;* [6] *Present address: Center for Applied Bioinformatics, St. Jude Children's Research Hospital;* [7] *Present address: Pacific Northwest National Laboratory. †These authors contributed equally to this work. * Correspondence:* `audreyf@uidaho.edu`.

## ABSTRACT

Expression quantitative trait loci (eQTLs) have been identified for most genes in the human genome across nearly 50 tissues or cell types. While most of the eQTLs are near the associated genes, some can be far away or on different chromosomes, with the regulatory mechanisms largely unknown. Here, we study cis- and trans-regulation of eQTLs across multiple tissues and cell types. Specifically, we constructed trios consisting of an eQTL, its cis-gene and trans-gene and inferred the regulatory relationships with causal network inference. We identify multiple types of regulatory networks for trios: across all the tissues, more than half of the trios are inferred to be conditionally independent, where the two genes are conditionally independent given the genotype of the eQTL (cis-gene $\leftarrow$ eQTL $\rightarrow$ trans-gene). Around 1.5% of the trios are inferred to be mediation (eQTL $\rightarrow$ mediator $\rightarrow$ target), around 1.3% fully connected among the three nodes, and just a handful v-structures (eQTL $\rightarrow$ gene 1 $\leftarrow$ gene 2). Unexpectedly, across the tissues, on average more than half of the mediation trios have the trans-gene as the mediator. Most of the mediators (cis and trans) are tissue specific. Furthermore, cis-gene mediators are significantly enriched for protein-coding genes compared with the genome average, whereas trans-gene mediators are significantly enriched for pseudogenes and depleted for long non-coding RNAs (lncRNAs).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



## 1. INTRODUCTION

Gene expression is regulated by genetic variants, among many factors. Expression Quantitative Trait Loci (eQTLs) have been identified to be widespread in the genome by several other consortia [1, 2, 3, 4, 5, 6]. In particular, the Genotype-Tissue Expression (GTEx) Consortium examines the expression profiles and identifies regulatory variants in around 50 tissues and cell types. They have identified at least one cis-eQTL (i.e., the eQTL is near its target gene, which is the cis-gene of this eQTL) for nearly all the protein-coding genes and nearly 70% of the long intergenic noncoding RNA (lincRNA) genes, as well as a small number of interchromosomal trans-eQTLs (i.e., the eQTL is far away from its target gene, which is the trans-gene of the eQTL). On the other hand, trans-eQTLs are enriched in the GWAS (genomewide association study) Catalog variants [2], suggesting potentially important roles of trans-eQTLs in complex traits and diseases.

However, the relationship between the cis- and trans-gene of the same eQTL remains unclear. A popular model is cis-gene mediation, where eQTL regulates the cis-gene (e.g., a transcription factor), which acts as the mediator and regulates the trans-gene [3, 7, 8, 9, 10, 11, 2, 12], possibly through regulatory elements (such as cis-regulatory domains, which may contain transcription factor binding sites) on the genome [11]. However, other modes of relationships have also been observed: Bryois et al. (2014) [3] analyzed 869 lymphoblastoid cell lines with a highly conservative approach and identified 49 cases where the eQTL regulates the cis-gene and the trans-gene independently, and 2 cases where the eQTL affects the trans-gene, which in turn affects the cis-gene, in addition to 19 cis-gene mediation trios. Delaneau et al. (2019) also estimated a high probability of 0.86 for an eQTL to regulate the cis- and trans-gene independently [11]. Furthermore, there is experimental evidence supporting functionally important inter-chromosomal interactions exist. For example, enhancers located on multiple chromosomes may converge and regulate the expression of olfactory receptor genes in cis and in trans [13, 14], suggesting that trans-regulation may be complex and much of it remains unknown.

Analyses of the multiple tissues and cell types of the GTEx consortium have further demonstrated tissue sharing and tissue specificity in the effect of eQTLs: around 80% of all the eQTLs found are shared in more than five tissues, and 25-30% of all the eQTLs are shared in nearly all the GTEx tissues [2]. To better understand regulation of cis- and trans-genes across tissues, we take a network approach to classify possible causal relationships in trios of an eQTL and its cis- and trans-target genes using 48 tissues and cell types in the GTEx version 8 data [2]. Our analysis goes beyond cis-mediation and also accounts for potential confounding from other genes. Our analysis further examines tissue sharing and specificity of these causal relationships.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



## 2. Results

2.1. **Multiple types of regulatory networks are identified for trios.** Using our MRPC (incorporating the principle of Mendelian Randomization into the PC algorithm [15]; the PC algorithm is named after developers Peter Spirtes and Clark Glymour [16]) method (version 3.0.0; `https://cran.r-project.org/web/packages/MRPC/index.html`) and accounting for potential confounding variables, we were able to identify multiple types of regulatory networks for trios (see Section "MRPC analysis accounting for confounding" in Methods for a detailed explanation of why we used MRPC for inference). The more interesting types among them are mediation ($M_1$), v-structure ($M_2$), conditional independence ($M_3$), and fully connected ($M_4$) (Figure 1A). In these networks, the two genes have correlated expression levels, although the correlation arises from different regulatory mechanisms. Specifically, in the mediation network, the eQTL acts as the instrumental variable, and one gene acts as the mediator between the eQTL and the other gene. When conditioned on the mediator, the downstream target is determined only by the mediator and is independent of the eQTL. In the network of conditional independence, the two genes are regulated by the same eQTL, inducing correlation between the two genes, even though there is no direct relationship between them. In the fully connect network, the two genes are not only regulated by the same eQTL, but also influence by additional, unknown processes between the two genes that lead to extra correlation between them. By contrast, in the v-structure network, both the eQTL and a gene are the parents of the other gene. Whereas the eQTL and the "parent" gene are independent, their information becomes dependent when the expression of the "child" gene is given. Apart from these networks, there are also inferred networks with only one edge – these are under the null model ($M_0$) category or the "other" category.

Across tissues, the number of trios we selected for network inference varies from 1,729 to 13,224 (median: 6,230; mean: 6,738; Supplementary Table 1). This number generally increases with the sample size (median: 235; mean: 315.2; Supplementary Figure 1). We further derived principal components (PrCs) from the whole-genome gene expression data in each tissue. Using Holm's method to control the familywise error rate across all the p-values at 5% [17], we identified PrCs that are highly significantly associated with the eQTLs or genes in all the trios of that tissue. We then applied MRPC to each trio, together with the associated PrCs. The number of PrCs included in the trio analysis varies from 0 to 13 across tissues (median: 0-2; Supplementary Table 2).

Among these trios (Figure 2, Supplementary Table 1), over half are inferred to be conditionally independent (median: 56.5%; mean: 55.3% across tissues), around 1.5% to be mediation (median: 1.5%; mean: 1.6% across tissues), about 1.3% to be fully connected (median: 1.3%; mean: 1.3% across tissues), and just a handful v-structures (median: 4; mean: 4.5 across tissues). Examples are provided in Figure 3; the same eQTLs or genes may appear in multiple tissues.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



2.2. **Cis- versus trans-gene mediation.** The mediation network type is of particular interest, as in general the hypothesis is that an eQTL regulates its trans-gene target through the cis-gene, which acts as a mediator. We term this the “cis-gene mediation" (Figure 1B). A trans-gene may also act as a mediator; we term this type of mediation the “trans-gene mediation" (Figure 1B). Several studies have examined cis-gene mediation [3, 10, 2]. In particular, Yang et al. [10] systematically identified trios of cis-gene mediation across GTEx tissues, using their GMAC (Genomic Mediation analysis with Adaptive Confounding adjustment) method, which is also based on the principle of Mendelian randomization and accounts for confounding variables. However, Yang et al. focused on trios with strong associations and examined a much smaller number of trios in each tissue (median: 1,112.5; mean: 1,473; around one fifth of the number of trios we examined). On the other hand, at 5% FDR, they identified more trios of cis-gene mediation than we did in more than half of the tissues: specifically, they identified a median of 102.5 (mean: 140.0) cis-gene mediation trios, whereas we identified a median of 98.5 (mean: 91.7) mediation (of both types) trios across all tissues. Take whole blood for example, they identified 281 trios of cis-gene mediation, whereas we identified 131 mediation trios of both types.

Unexpectedly, our inference results include a large number of trios of trans-gene mediation (Figure 4; Supplementary Tables 3, 4 and 5; see examples in Figure 3A). In fact, often more than half (59% across all tissues; median: 60%; mean: 59%) of the mediation trios are of trans-gene mediation. Specifically, among the mediators, there are 1,005 cis-genes, 1,898 trans-genes, and 91 genes that are both cis- and trans-genes for an eQTL (Supplementary Table 6). Some of these mediators are shared in more than one tissue; that is, the same gene is a mediator in a trio in one tissue, and also a mediator in a different trio in another tissue (Figure 5; Supplementary Tables 4 and 5). Among the 1,005 cis-gene mediators, 213 (21%) are shared in more than one tissue, and the distribution of shared tissues follows an exponential decay. Among the 1,898 trans-genes mediators, by contrast, only 94 (5%) are shared in more than one tissue, and the distribution follows a faster exponential decay.

We examined a few examples in detail (Figure 3). The protein-coding gene ICOSLG (Inducible T Cell Costimulator Ligand) on chromosome 21 is inferred to be the mediator to ENSG00000277117, a novel gene similar to itself, across multiple tissues (Figure 3A). The median of the absolute correlations is 0.44 between the eQTL and ICOSLG, and is only 0.002 between the same eQTL and the trans-gene, supporting the cis-gene being the mediator. Similarly, the lncRNA MAP3K2-DT (MAP3K2 Divergent Transcript, where MAP3K2 stands for Mitogen-Activated Protein Kinase Kinase Kinase 2) on chromosome 2 is a mediator to a number of genes in multiple tissues (Figure 3A). The median absolute correlation is 0.74 between the eQTL and the cis-gene, and is also only 0.002 between the eQTL and the corresponding trans-gene.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



The correlation patterns are reversed in trans-gene mediating trios. For example, the pseudogene MTND1P23 (MT-ND1 Pseudogene 23, where MT-ND1 stands for Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunit 1) is inferred to be the mediator in many trios across a large number of tissues, including whole blood, nerve tibial, skin and so on (Figure 3B). The cis-genes in these trios are often protein-coding genes: for example, in whole blood, all the cis-genes displayed in Figure 3B are protein-coding genes, except for RPL3P7 (Ribosomal Protein L3 Pseudogene 7), which is also a pseudogene. In whole blood, the median absolute correlation is 0.30 between the eQTL and MTND1P23, and is 0.09 between the eQTL and the corresponding cis-gene. In nerve tibial, the medians are 0.38 and 0.08, respectively. Therefore, the data is consistent with the trans-gene mediation model, where the association with the eQTL attenuates when the regulatory signal moves from the eQTL to the mediator and next to the target.

2.3. **Gene type enrichment among cis-gene and trans-gene mediators.** We next investigated whether the mediators are enriched for a specific type of genes (Figure 6; Supplementary Table 7). We used the Ensembl human gene annotations and grouped the gene types into four broad categories: protein-coding genes, pseudogenes, lncRNAs, and others. Multiple RNAs, such as miRNA, snoRNA, etc., were lumped into the “other" category due to low occurrences. The percentages of protein-coding genes and lincRNA are higher among cis-gene mediators (52% and 26%, respectively) than among the trans-gene mediators (27% and 20%), whereas the percentage of pseudogenes is lower in cis-gene mediators (19%) than in trans-gene mediators (31%) (Supplementary Table 8). When compared with the whole genome, we observe statistically significant enrichment of protein-coding genes among cis-gene mediators, enrichment of pseudogenes among trans-gene mediators, and depletion of lincRNAs among trans-gene mediators, all with chi-square test p-values less than $10^{-6}$ (Supplementary Table 8).

2.4. **Analysis of HiC sequencing data for trios of trans-gene mediation.** We also examined the HiC sequencing data for physical interactions in multiple cell types (lung, skin, lymphoblastoid cells, and fibroblast cells; [18, 19, 20]; see “HiC sequencing data analysis" in Methods) for potential evidence that may support trans-gene mediation. HiC sequencing reads indicate that two loci on the genome are physically close in space, even when the loci are far apart on the linear genome [21]. Studies of promoter-enhancer interactions have suggested that enhancers that may be located far away from a gene can be brought close to the promoter of this gene through chromatin looping, thus facilitating the regulation of gene expression [13, 14]. Here, we were interested in regulation of an eQTL and its trans-gene mediators. The hypothesis is that trans-regulation may involve physical contact between the eQTL and the gene, which may be captured by HiC sequencing. However, we did not observe significant enrichment of HiC reads for any trio of trans-gene mediation (Supplementary Tables 9).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



For example, in fibroblasts (Supplementary Tables 9), we were able to extract as many as 226 HiC reads between a 200kb neighborhood of the eQTL at chr1:46,309,111 and a 200kb neighborhood of its trans-gene UQCRHL also on chr1 at chr1:15,807,169-15,809,348 (30.5 Mb away from the eQTL). When we performed Monte Carlo simulation to generate the null distribution of the HiC reads for chromosome 1, we obtained a Monte Carlo p-value of 0.07 (with a q-value of 0.54). In skin that is unexposed to the Sun (Supplementary Tables 9), between the eQTL at chr17:28,499,368 and its trans-gene TSPO at chr22:43,151,547-43,163,242, we were able to extract only 2 reads, although low numbers of interaction reads are common between two chromosomes. The Monte Carlo p-value was 0.27 with a q-value of 1.00.

2.5. **Sensitivity analysis of the results to the choice of FDR control methods in MRPC.** MRPC allows for different FDR control methods to be incorporated into network inference (see “MRPC analysis accounting for confounding" in Methods). The LOND (the significance Level based On the Number of Discoveries so far) method ([22]; referred to as “MRPC-LOND" hereinafter) is known to be conservative: our simulation studies show that it tends to infer fewer edges than there are, especially in larger networks [15, 23]. Do our conclusions change if the FDR control was less stringent and MRPC inferred more edges? To examine this question, we re-analyzed the GTEx data, using the ADDIS (ADaptive DIScarding) method [24] for FDR control; this option is also implemented in our MRPC package (referred to as “MRPC-ADDIS" hereinafter). Our simulation showed that MRPC-ADDIS was slightly less conservative than MRPC-LOND on three-node networks, although the difference is more pronounced on a larger network (Supplementary Figure 2). As expected, MRPC-ADDIS identified more trios (in about 70% of all tissues) with an interesting network structure (i.e., $M_1$-$M_4$) (Supplementary Figures 3; Supplementary Table 10). Among all the trios, MRPC-ADDIS inferred a similar percentage of conditional independence and mediation trios: over half are inferred to be conditionally independent (median: 55.5%; mean: 55.0% across tissues), and around 1.9% to be mediation (median: 1.8%; mean: 2.0% across tissues). More trios were inferred to be fully connected or v-structures: about 4.5% to be fully connected (median: 4.5%; mean: 4.6% across tissues), and more v-structures (median: 22.5; mean: 25.88 across tissues).

Furthermore, MRPC-ADDIS identified more trios mediated by trans-genes, accounting for about 70% of all mediation trios (compared to 60% with MRPC-LOND) (Supplementary Figures 4; Supplementary Tables 3, 11 and 12). This comparison provides additional evidence that the strong presence of trans-gene mediation is less likely a computational artifact. Similar to the MRPC-LOND results, cis- and trans-mediators inferred by MRPC-ADDIS are also often identified in multiple tissues, and the histograms of shared tissues follow a similarly exponential decay (Supplementary Figure 5). We also observe statistically significant enrichment of protein-coding genes among cis-gene mediators, and

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



enrichment of pseudogenes and depletion of lincRNAs among trans-gene mediators (Supplementary Figure 6; Supplementary Tables 6, 7 and 8). Nonetheless, unsurprisingly, HiC analysis on these mediation trios did not yield positive results, either (Supplementary Table 13).

2.6. **Comparing MRPC to GMAC on the mediation trios.** To further validate the observation of many trans-gene mediation trios, we applied the GMAC method [10] (version 3.0; `https://cran.r-project.org/web/packages/GMAC/index.html`) introduced above to all the candidate trios in five GTEx tissues with the largest sample sizes: adipose subcutaneous, tibial artery, muscle skeletal, sun exposed skin, and whole blood. GMAC aims to infer in a trio whether the data supports the candidate mediator having an effect on the target gene, even if the eQTL may also have a direct effect on the target. The candidate mediator can be either a cis-gene or trans-gene, depending on how the two genes are ordered in the input. In our application, we applied GMAC to each trio twice: first treating the cis-gene as a potential mediator, and next treating the trans-gene as a potential mediator. With this analysis, we were interested in testing whether a method different from MRPC also identified many trans-gene mediation trios.

At an FDR of 10%, GMAC identified on average 393 (median: 401) cis-gene mediation trios and 311 (median: 314) trans-gene mediation trios across the five tissues (Supplementary Table 14). Most of these trios are the same trios, and GMAC inferred mediation when the cis-gene was the mediator and also when the trans-gene was the mediator. This means that these trios follow an $M_4$ model in our framework where the edge between the two genes is bidirected (Figure 1). Nevertheless, this result is consistent with ours and confirms that trans-gene mediation is at least as common as cis-gene mediation.

## 3. Discussion

Using our causal network inference method, we identified multiple types of regulatory relationships in trios of an eQTL and its cis- and trans-genes, which provides a more comprehensive picture of the complex relationships in trios. Across the tissues, more than half of the trios are inferred to be conditionally independent, and around 1.5% of the trios are inferred to be mediation. Interestingly, on average more than half of the mediation trios have the trans-gene as the mediator. Furthermore, cis-gene mediators are enriched for protein-coding genes compared with the genome average, whereas trans-gene mediators are enriched for pseudogenes and depleted for lincRNAs.

An edge inferred by our statistical method indicates that the relationship is strong enough not to be explained away by other factors we have considered. Such a strong relationship, albeit still a statistical result, is therefore more likely to reflect a genuine regulatory relationship. On the other hand, it is very likely that an inferred edge condenses a complex regulation process, which may involve many genes or processes other than transcription.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



However, although the number of trios we examined is on the order of thousands across tissues, this is still a very small number of possible trios for the human genome. The trios we considered here generally have strong association. This is because when multiple SNPs were identified to be eQTLs for the same cis-gene, we used the SNP with the smallest p-value. As a result, we may have missed many eQTLs that could have slightly weaker association than the chosen one, but may have stronger association with trans-genes. The distribution of the regulation types may be biased due to this omission, but the distribution from our analysis is quantitatively comparable to that from other studies [3, 11], with the majority of the trios being the conditional independence type, and a small percentage of mediation trios.

Although surprising initially, the observation of a large number of trans-gene mediating trios is consistent with existing literature on the prevalence of trans-regulation. As summarized by Liu et al. (2020) [25], trans-eQTLs contribute to 60-90% of the heritability in gene expression across multiple studies [26, 27, 28, 29], suggesting that eQTLs often regulate genes that are far away. Trans-gene mediation has also been identified before by other studies. For example, [3] applied the CIT method to the lymphoblastoid cell lines (LCLs) from a cohort of 869 individuals and identified 19 trios of cis-gene mediation, 2 trios of trans-gene mediation, as well as 49 conditional independent trios. The number of trans-gene mediation trios is low in this study, but the number of conditional independent trios is also low, suggesting lower power to detect any type. Overall, the observation that the conditional independent trios are much more frequent than mediation trios is consistent with our findings. Additionally, we detected the enrichment of pseudogenes among trans-gene mediators. This is also consistent with functional studies that report pseudogenes acting as the trans-regulator of protein-coding genes, although such studies are still scarce: for example, the pseudogene BRAFP1 (BRAF Pseudogene 1) on chromosome X can regulate the protein-coding gene BRAF (B-Raf Proto-Oncogene, Serine/Threonine Kinase) on chromosome 7 in cancer cells [30], and the pseudogene HBBP1 (Hemoglobin Subunit Beta Pseudogene 1) on chromosome 11 can regulate the transcription factor TAL1 (TAL BHLH Transcription Factor 1, Erythroid Differentiation Factor) on chromosome 1 during erythropoiesis [31]. Although these four genes appeared in tested trios for some of the GTEx tissues in our analysis, no trios included the two specific pairs, likely due to two reasons: i) the rather stringent criteria which we used for identifying trans-genes and subsequently constructing the trios; and ii) the lack of suitable cell types in GTEx.

What is the potential mechanism for trans-gene mediation? GTEx found enrichment of trans-eQTLs at CTCF (CCCTC-Binding Factor) binding sites, and hypothesized that such eQTLs may disrupt CTCF binding, which influences the spatial chromatin interaction and therefore gene expression [2]. Our analysis of the HiC data was inspired by the studies on distal enhancers for olfactory receptor genes [13, 14], and assumed direct contact between an eQTL and its trans-gene. Not observing significant HiC enrichment in our analysis is not necessarily evidence against trans-mediation. The number of reads connecting two

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



chromosomes is generally low, and so is the number of reads connecting two distal regions on the same chromosome. The inconclusive result therefore may have two interpretations: i) There is genuinely no physical interaction; and ii) Physical interaction is not captured by current technology. We currently do not have knowledge of which interpretation is more likely. Furthermore, because of the sparse read counts between distal locations in existing HiC data, we examined a wide neighborhood of each eQTL in our analysis, and this neighborhood of 200kb would generally include the cis-gene of the eQTL. The reads we identified between an eQTL and a trans-gene could also be between the cis-gene and the trans-gene, or between other genes in the two neighborhoods. Võsa et al (2021) [1] also examined the HiC enrichment between eQTLs (or equivalently cis-genes) and trans-genes, and found significant enrichment compared to all possible gene pairs at the genome level ($p = 2.4 \times 10^{-153}$), which provides global support for spatial interaction, although it remains difficult to pinpoint which pairs are enriched for such interaction.

We validated the common presence of trans-gene mediation with the GMAC method. However, it is important to note that GMAC detects mediation by testing for association between two genes in a trio, in the presence of an eQTL and confounder variables. This means that GMAC tests only for the edge between the two genes. As long as this edge is present, GMAC interprets it as mediation; the presence of other edges is less relevant [10]. Therefore, mediation under GMAC corresponds to $M_1$, $M_2$, or $M_4$ (all have an edge between two genes) under our framework, whereas lack of a mediation relationship under GMAC corresponds to $M_0$ or $M_3$ (no edge between two genes) under our framework (Figure 1). The comparison of MRPC and GMAC raises the question: what does mediation mean? GMAC allows an eQTL to regulate a target gene directly *and* through a mediator, whereas MRPC requires that all the eQTL effect goes through the mediator and that there is no edge connecting the eQTL and the target gene. We may consider the former *partial* mediation and the latter *complete* mediation. Partial mediation, however, is a challenge in statistical inference. Consider the following two models: i) $V \to T_1, V \to T_2, T_1 \to T_2$; and ii) $V \to T_1, V \to T_2, T_1 \leftarrow T_2$. These two graphical models always have the same likelihood, which means that they are Markov equivalent [32] and cannot be distinguished without additional biological information. Both models are interpreted as mediation by GMAC, but are considered to be statistically unidentifiable under MRPC and will be inferred as $M_4$ ($V \to T_1, V \to T_2, T_1 - T_2$; Figure 1) instead.

## 4. METHODS

4.1. **The GTEx genotype and gene expression data used for this study.** We used the association test results from GTEx (in *.egenes.txt.gz, where * refers to a tissue or cell type name) to identify the top eQTL for individual genes [33]. In the files *.egenes.txt.gz, a single variant is reported for each gene, even though the association may not be strong. We extracted the genes with an association q-value $\leq 0.05$, which resulted in 4,934 genes.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



4.2. **PEER normalization of GTEx gene expression data.** We used the transcript TPM (transcript per million) data for gene expression in 48 tissues (or cell types). A handful of tissues were removed due to a low sample size ($< 100$). The sample size of the tissues analyzed here ranges from 114 to 706, with a mean of 315 and a median of 235. Following the standard procedure adopted by the GTEx consortium, we performed PEER normalization (probabilistic estimation of expression residuals) [34] on the whole-genome gene expression data for each tissue. We included the covariates provided by GTEx: sex, platform, PCR, and the top five principal components from the genome-wide genotype data, which may contain signals on the potential population structure. We added age to this list of covariates. Additionally, we included different numbers of PEER factors for each tissue, depending on the sample size: 15 factors for $< 150$, 30 for $150-250$, 45 for $250-350$, and 60 for $\geq 350$.

4.3. **Selection and identification of trios.** Using the eQTLs reported by GTEx and the PEER-normalized gene expression described in the two sections above, we next ran the R package MatrixEQTL [35] to look for trans-genes located 1 Mb away from the eQTLs with a p-value $< 10^{-5}$. Multiple trans-genes may be identified for the same eQTL. We constructed trios for each eQTL with a cis-gene and a trans-gene. Different trios may have the same eQTL, or the same gene. A gene may also be a cis-gene in one trio but a trans-gene in another.

4.4. **Identification of associated confounding variables.** We performed principal component analysis on the PEER-normalized gene expression in each tissue and then tested the significance of the three variables in each candidate trio to each PrC using a simple regression and obtained a p-value. We used Holm's method to control the familywise error rate across all the p-values at 5% [17]. Each PrC reflects potential impact from a large number of genes and represents the influence from the larger gene regulatory network to which a trio may belong. We then included these PrCs as additional nodes in the MRPC analysis of the trio. Due to the strong control of Holm's method, the median number of PrCs included in the end varies between 0 and 2 across tissues.

4.5. **MRPC analysis accounting for confounding.** We used our R package MRPC [15, 23] to perform the network inference for each trio in each tissue, including the associated principal components as additional nodes. MRPC builds on the classical PC algorithm (named after its developers Peter Spirtes and Clark Glymour; [16]) for inference of directed acyclic graphs, and incorporates the principle of Mendelian randomization (PMR) [36]. MRPC therefore combines the strengths of two classes of methods for causal network inference. As discussed in our earlier work [15, 23], one class is the general-purpose network inference methods, such as the PC algorithm and its variants (e.g., methods implemented in R packages bnlearn [37] and pcalg [38]). Existing methods in this class are computationally efficient but difficult to modify to account for the PMR. The other class

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



is developed for genomic data and explicitly accounts for the PMR (e.g., CIT [39], QPSO [40], and findr [41]). However, the types of causal networks detected by existing methods are often limited to a subset of the five basic models in Figure 1.

The PC algorithm consists of two main steps: inferring a graph skeleton, where the key edges are retained but undirected; and determining the direction of the edges. Our MRPC improves both steps and achieves better power and lower FDR on small networks [15, 23]. The improvement in Step 1 is further explained below. Step 2 in MRPC incorporates the PMR, which takes advantage of the additional information in eQTLs. Under the PMR, the genotypes can be reasonably assumed to be randomly allocated in a natural population, and can therefore be viewed as randomization of the individuals. Since the genotypes influences the phenotypes, but not the other way around, the PMR then views an eQTL as an instrumental variable for causal inference. Causal inference on a trio aims to infer a three-node network for the eQTL and the two genes, with an edge pointing from the eQTL to one or both genes (Figure 1)

Step 1 in MRPC, as well as other PC-like algorithms, starts from a fully connected network and performs a series of tests on each edge to see whether the two nodes are marginally correlated or conditionally correlated, given one other node, or two other nodes, or any subset of other nodes. If a test produces an insignificant p-value, it means that the correlation between the two nodes may be not strong enough or can be explained away by other nodes. The edge would be removed and never tested again. Hypothesis testing in PC-like algorithms is therefore online, meaning that the number of statistical tests to be performed is unknown in advance, and that the threshold for a p-value to be considered significant cannot be fixed beforehand. Several methods have been developed to control the overall FDR in this online setting. We have implemented such a method called LOND [22] in MRPC, and demonstrated that LOND achieved better power and lower actual FDR than existing methods on small networks through extensive simulations [15, 23].

However, LOND may be too conservative and leads to the true edges being missed, especially in larger networks [23]. We have therefore also implemented the ADDIS method [24], another less-conservative online FDR control method, in MRPC. We used both LOND and ADDIS in MRPC to control the FDR at 5% for each trio. Our simulation (Supplementary Figure 2) showed that LOND and ADDIS have similar performance on three-node networks such as $M_1$, although ADDIS achieves higher precision (i.e., 1-FDR) and higher power on larger networks, especially when the sample size is a few hundred.

4.6. **Comparison with GMAC.** To compare the GMAC and MRPC methods, we applied GMAC to each of the top five GTEx tissues by sample size. Following the instructions in the GMAC package and consistent with the application in the GMAC paper [10], we used all the principle components of the genomewide expression matrix as the covariate pool, and three additional known covariates: the PCR used, the platform used, and sex of the individual in each sample.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



For each trio, GMAC identified and removed covariates that were a common child or intermediate variable to the two genes at an FDR of 10%, and identified confounders (defined as a parent node to the two genes in GMAC) at an FDR of 5%. The input to GMAC consisted of the e-QTL and the PEER-normalized expression values of the cis- and trans-gene. Since the genotypes are missing in some individuals, we performed imputation using multiple correspondence analysis (MCA; [42]) prior to the GMAC analysis. We ran GMAC twice on each trio, first with the cis gene as the potential mediator and second with the trans gene as the potential mediator. GMAC output a p-value for each trio, and Yang et al. in their original study then used these p-values in two ways to select mediation trios: i) using these unadjusted p-values and setting $p < 0.05$; and ii) applying the q-value method and setting $q < 0.05$ [10]. Here, we took the middle road and set $q < 0.1$ for each tissue.

4.7. **Gene type enrichment analysis among mediators.** Both cis-genes and trans-genes may be inferred to be the mediator. We used the Ensembl human gene annotations (GRCh38/hg38) and grouped the gene types into four broad categories: protein-coding genes, pseudogenes, lncRNAs, and others. Multiple RNAs, such as miRNA, snoRNA, etc., were lumped into “other" due to low occurrences. We summarized the counts for cis-gene mediators, and separately for trans-gene mediators. Since we are interested in which type was enriched in cis-genes and in trans-genes, we next performed a chi-square test with one degree of freedom for each type separately. Take the protein-coding genes for example. We constructed a $2 \times 2$ contingency table where the rows indicate whether a mediator is a cis-gene or a trans-gene, and the columns indicate whether the mediator gene is a protein-coding gene or not. We also performed the chi-square test to examine whether the enrichment (or depletion) of a certain gene type is higher (or lower) than the genome average. For this test, each row in the $2 \times 2$ contingency table shows the proportions of a gene type (e.g., protein-coding vs not protein-coding) among the gene set of interest (e.g., cis-genes) versus the genome level.

4.8. **HiC sequencing data analysis.** We downloaded four HiC-sequencing datasets from the ENCODE consortium: lung (ENCFF366ERB; [18, 19]), skin (ENCFF569RJM; [18, 19]), lymphoblastoid cells (ENCFF355OWW; [18, 19]), and fibroblast cells (ENCFF768UBD; [20]). We again used the Ensembl human gene annotations (GRCh38/hg38) to determine the positions of genes. We used the package *StrawR* [43] to extract reads from positions along the chromosomes corresponding to the SNP and trans-mediated gene. For all extractions, the resolution, defined as the bin size in the package, was set to 10 kb, which was the finest resolution shared by all four tissues.

We calculated a Monte Carlo p-value to identify whether an observed number of interactions between a SNP and a trans-gene mediator was significant. These p-values were formulated from the upper tail probability of the observed number of reads relative to the empirically generated null distribution. To construct the empirical null distribution for each pair, we randomly drew 10,000 pairs of neighborhoods of 200 kb uniformly located

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



on both the chromosome of the SNP and the chromosome of the trans-gene. The numbers of interaction reads in these neighborhood pairs then constitute the null distribution. The Monte Carlo p-value is then the proportion of reads exceeding the observed number of interaction reads. To account for multiple testing, we applied Holm's method [17] to control the family-wise error rate at 5%, and the Benjamini and Yekutieli method [44] and the q-value method [45] to control the false discovery rate at 5%.

## FUNDING

This research is supported by the National Institutes of Health (NIGMS P20GM104420).

## ACKNOWLEDGEMENTS

We acknowledge the extensive computing support provided by the Research Computing and Data Services from the Institute for Interdisciplinary Data Sciences at the University of Idaho.

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/Leidos Biomedical Research, Inc. 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 The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc.(HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951,MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The datasets used for the analyses described in this manuscript were obtained from dbGaP at `http://www.ncbi.nlm.nih.gov/gap` through dbGaP accession number phs000424.v8.pht002741.v8.p2.

## AUTHOR CONTRIBUTIONS

A.Q.F. conceived the study. M.B.B., E.A.M. and A.Q.F. developed the tools. M.M. contributed to tool development. J.K., M.B.B., and A.Q.F. analyzed the data. J.W. and X.W. contributed to interpretation of data. A.Q.F. wrote the manuscript with input from the co-authors.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

## REFERENCES

[1] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Harm Brugge, et al. Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics*, 53(9):1300–1310, 2021.

[2] The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[3] Julien Bryois, Alfonso Buil, David M Evans, John P Kemp, Stephen B Montgomery, Donald F Conrad, Karen M Ho, Susan Ring, Matthew Hurles, Panos Deloukas, et al. Cis and trans effects of human genomic variants on gene expression. *PLoS Genetics*, 10(7):e1004461, 2014.

[4] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, Rui Mei, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, 24(1):14–24, 2014.

[5] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506, 2013.

[6] Joseph E Powell, Anjali K Henders, Allan F McRae, Anthony Caracella, Sara Smith, Margaret J Wright, John B Whitfield, Emmanouil T Dermitzakis, Nicholas G Martin, Peter M Visscher, et al. The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. *PLoS One*, 7(4):e35430, 2012.

[7] Brandon L Pierce, Lin Tong, Lin S Chen, Ronald Rahaman, Maria Argos, Farzana Jasmine, Shantanu Roy, Rachelle Paul-Brutus, Harm-Jan Westra, Lude Franke, et al. Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 south asians. *PLoS genetics*, 10(12):e1004818, 2014.

[8] Holger Kirsten, Hoor Al-Hasani, Lesca Holdt, Arnd Gross, Frank Beutner, Knut Krohn, Katrin Horn, Peter Ahnert, Ralph Burkhardt, Kristin Reiche, et al. Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci. *Human Molecular Genetics*, 24(16):4746–4763, 2015.

[9] Chen Yao, Roby Joehanes, Andrew D Johnson, Tianxiao Huan, Chunyu Liu, Jane E Freedman, Peter J Munson, David E Hill, Marc Vidal, and Daniel Levy. Dynamic role of trans regulation of gene expression in relation to complex traits. *The American Journal of Human Genetics*, 100(4):571–580, 2017.

[10] Fan Yang, Jiebiao Wang, The GTEx Consortium, Brandon L Pierce, and Lin S Chen. Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Research*, 27(11):1859–1871, 2017.

[11] Olivier Delaneau, M Zazhytska, Christelle Borel, G Giannuzzi, Guillaume Rey, Cédric Howald, S Kumar, Halit Ongen, Konstantin Popadin, D Marbach, et al. Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science*, 364(6439), 2019.

[12] Fan Yang, Kevin J Gleason, Jiebiao Wang, Jubao Duan, Xin He, Brandon L Pierce, and Lin S Chen. CCmed: cross-condition mediation analysis for identifying replicable trans-associations mediated by cis-gene expression. *Bioinformatics*, 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



[13] Elizaveta Bashkirova and Stavros Lomvardas. Olfactory receptor genes make the case for inter-chromosomal interactions. *Current Opinion in Genetics & Development*, 55:106–113, 2019.

[14] Kevin Monahan, Adan Horta, and Stavros Lomvardas. LHX2-and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature*, 565(7740):448–453, 2019.

[15] Md Bahadur Badsha and Audrey Qiuyan Fu. Learning causal biological networks with the principle of Mendelian randomization. *Frontiers in Genetics*, 10:460, 2019, doi:10.3389/fgene.2019.00460.

[16] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

[17] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.

[18] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[19] Adrian L Sanborn, Suhas SP Rao, Su-Chen Huang, Neva C Durand, Miriam H Huntley, Andrew I Jewett, Ivan D Bochkov, Dharmaraj Chinnappan, Ashok Cutkosky, Jian Li, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47):E6456–E6465, 2015.

[20] Guy Nir, Irene Farabella, Cynthia Pérez Estrada, Carl G Ebeling, Brian J Beliveau, Hiroshi M Sasaki, Soun H Lee, Son C Nguyen, Ruth B McCole, Shyamtanu Chattoraj, et al. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genetics*, 14(12):e1007872, 2018.

[21] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[22] Adel Javanmard and Andrea Montanari. On online control of false discovery rate. *arXiv*, 2015. arXiv:1502.06197.

[23] Md Bahadur Badsha, Evan A Martin, and Audrey Qiuyan Fu. MRPC: An R package for inference of causal graphs. *Frontiers in Genetics*, 12, 2021.

[24] Jinjin Tian and Aaditya Ramdas. ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls. *Advances in Neural Information Processing Systems*, 32:9388–9396, 2019.

[25] Xuanyao Liu, Yang I Li, and Jonathan K Pritchard. Trans effects on gene expression can drive omnigenic inheritance. *Cell*, 177(4):1022–1034, 2019.

[26] Alkes L Price, Nick Patterson, Dustin C Hancks, Simon Myers, David Reich, Vivian G Cheung, and Richard S Spielman. Effects of cis and trans genetic ancestry on gene expression in African Americans. *PLoS Genetics*, 4(12):e1000294, 2008.

[27] Alkes L Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genetics*, 7(2):e1001317, 2011.

[28] Fred A Wright, Patrick F Sullivan, Andrew I Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, et al. Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*, 46(5):430–437, 2014.

[29] Xuanyao Liu, Hilary K Finucane, Alexander Gusev, Gaurav Bhatia, Steven Gazal, Luke O?Connor, Brendan Bulik-Sullivan, Fred A Wright, Patrick F Sullivan, Benjamin M Neale, et al. Functional architectures of local and distal regulation of gene expression in multiple human tissues. *The American Journal of Human Genetics*, 100(4):605–616, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



[30] Florian A Karreth, Markus Reschke, Anna Ruocco, Christopher Ng, Bjoern Chapuy, Valentine Léopold, Marcela Sjoberg, Thomas M Keane, Akanksha Verma, Ugo Ala, et al. The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. *Cell*, 161(2):319–332, 2015.

[31] Xiaolin Yin, Chenguang Sun, Yanan Mao, Fanqi Zhou, Yi Shao, Qian Liu, Jiayue Xu, Li Cheng, Daqi Yu, Pingping Li, et al. Genome-wide analysis of pseudogenes reveals HBBP1?s human-specific essentiality in erythropoiesis and implication in $\beta$-thalassemia. *Developmental Cell*, 56:1–16, 2021.

[32] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270. Elsevier Science Inc., 1990.

[33] The GTEx Consortium. GTEx Analysis V8 (dbGaP Accession phs000424.v8.p2). `https://gtexportal.org/home/datasets`, 2021. [Online; accessed 8-June-2020].

[34] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500, 2012.

[35] Andrey A Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[36] George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):R89–R98, 2014.

[37] Marco Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35, 2010. doi:10.18637/jss.v035.i03.

[38] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

[39] Joshua Millstein, Bin Zhang, Jun Zhu, and Eric E Schadt. Disentangling molecular relationships with a causal inference test. *BMC Genetics*, 10(1):23, 2009.

[40] Huange Wang and Fred A van Eeuwijk. A new method to infer causal phenotype networks using QTL and phenotypic information. *PLoS One*, 9(8):e103997, 2014.

[41] Lingfei Wang and Tom Michoel. Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data. *PLoS Computational Biology*, 13(8):e1005703, 2017.

[42] Julie Josse, François Husson, et al. missMDA: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

[43] Neva C Durand, James T Robinson, Muhammad S Shamim, Ido Machol, Jill P Mesirov, Eric S Lander, and Erez Lieberman Aiden. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems*, 3(1):99–101, 2016.

[44] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[45] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



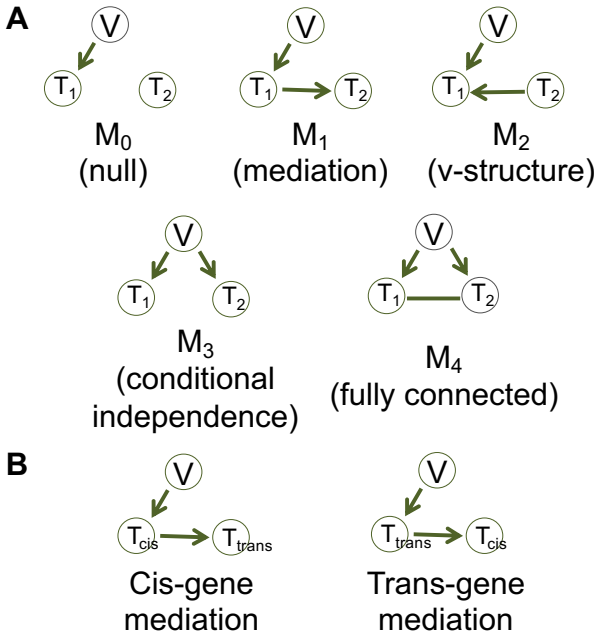



FIGURE 1. **The regulatory relationships of a trio that may be inferred by MRPC**. (A) $M_0$ through $M_4$ are the five basic causal networks that may be detected by our MRPC method. (B) The two types of the mediation model ($M_1$): cis-gene mediation and trans-gene mediation.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



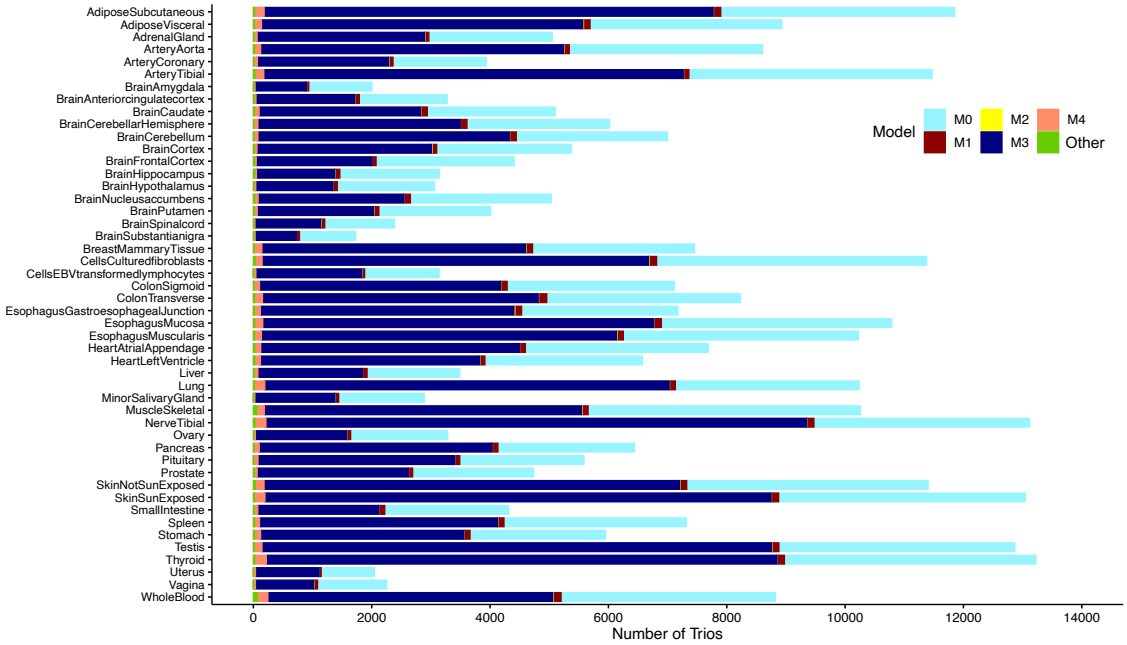



FIGURE 2. **The breakdown of inferred trio types across GTEx tissues and cell types**. Each model is represented by a unique color. “Other" refers to inferred networks that are not any of the five basic models.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



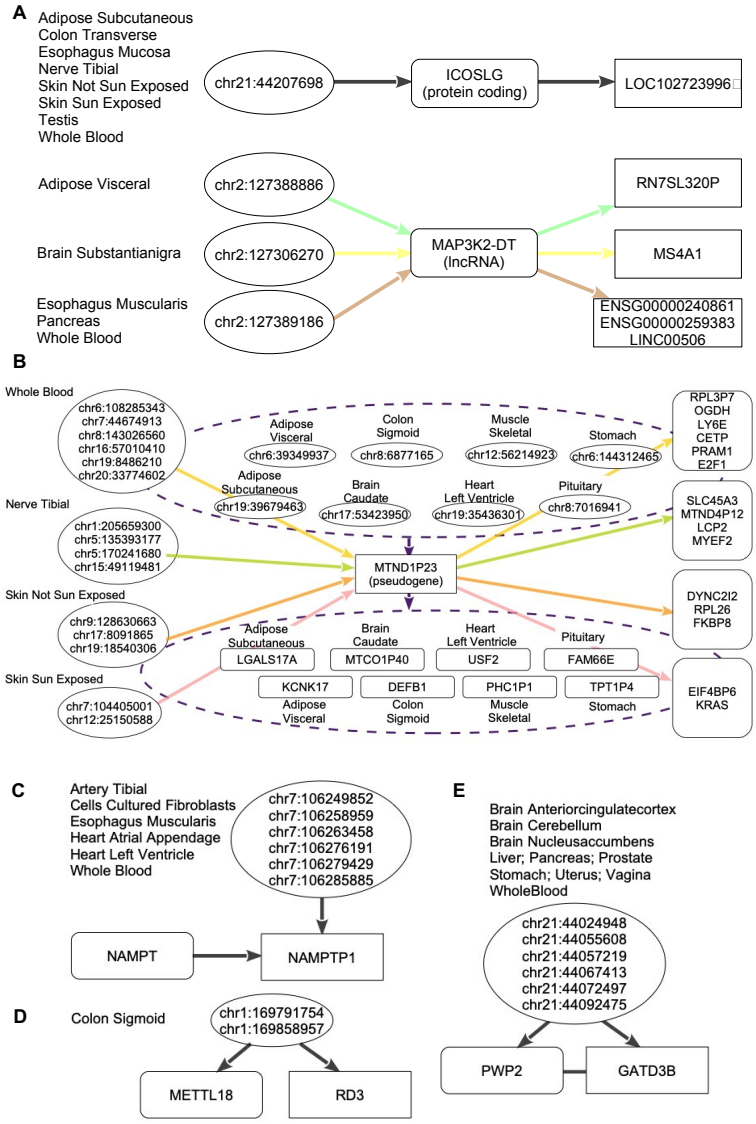



FIGURE 3. **Examples of trios inferred for different models**. Ovals indicate eQTLs, rectangles with rounded corners indicate cis-genes of the eQTLs, and rectangles with sharp corners indicate trans-genes. All the networks are inferred under both MRPC-LOND and MRPC-ADDIS. (A) Examples of cis-gene mediation. One cis-gene is a protein-coding gene, and the other lncRNA. Arrows of the same color connect the eQTLs, the cis-gene and the trans-genes from the same tissue. (B) Examples of trans-gene mediation. The common trans-gene here is a pseudogene. To reduce clutter, the dashed circle groups together multiple trios each belonging to a different tissue. (C) An example of $M_2$ (the v-structure). (D) An example of $M_3$ (a conditional independence network). (E) An example of $M_4$ (a fully connected network). (C)-(E) also demonstrate that due to linkage disequilibrium, multiple eQTLs in a small neighborhood may be identified for the same genes.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



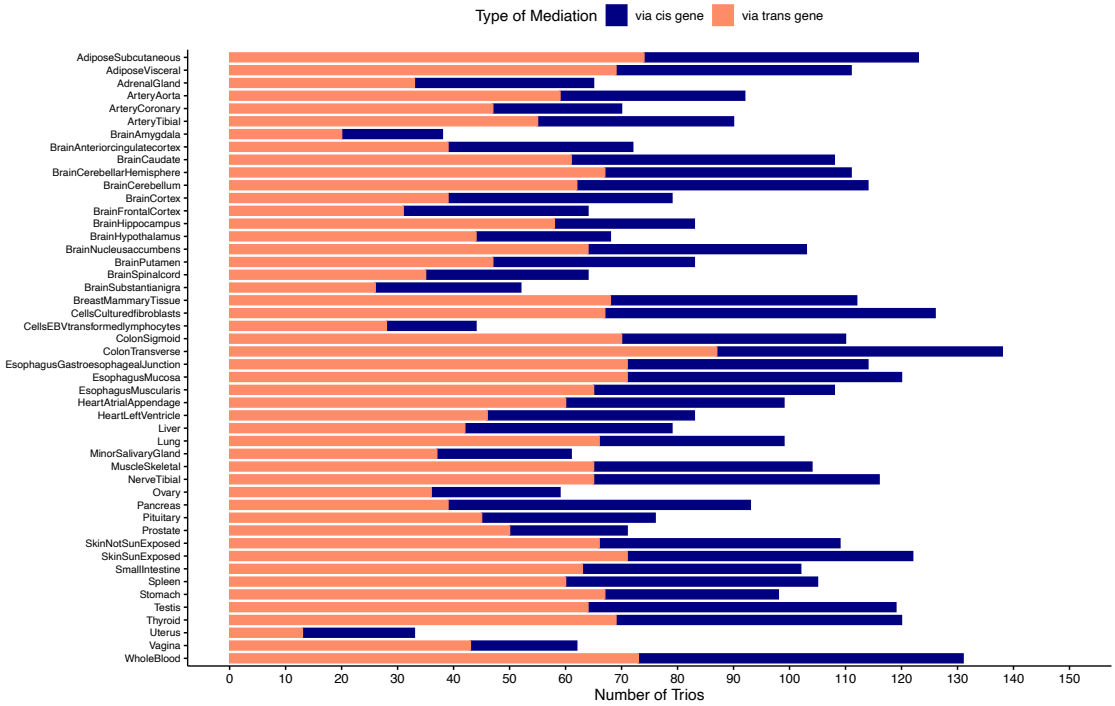



FIGURE 4. **The breakdown of inferred mediation types across GTEx tissues and cell types**

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



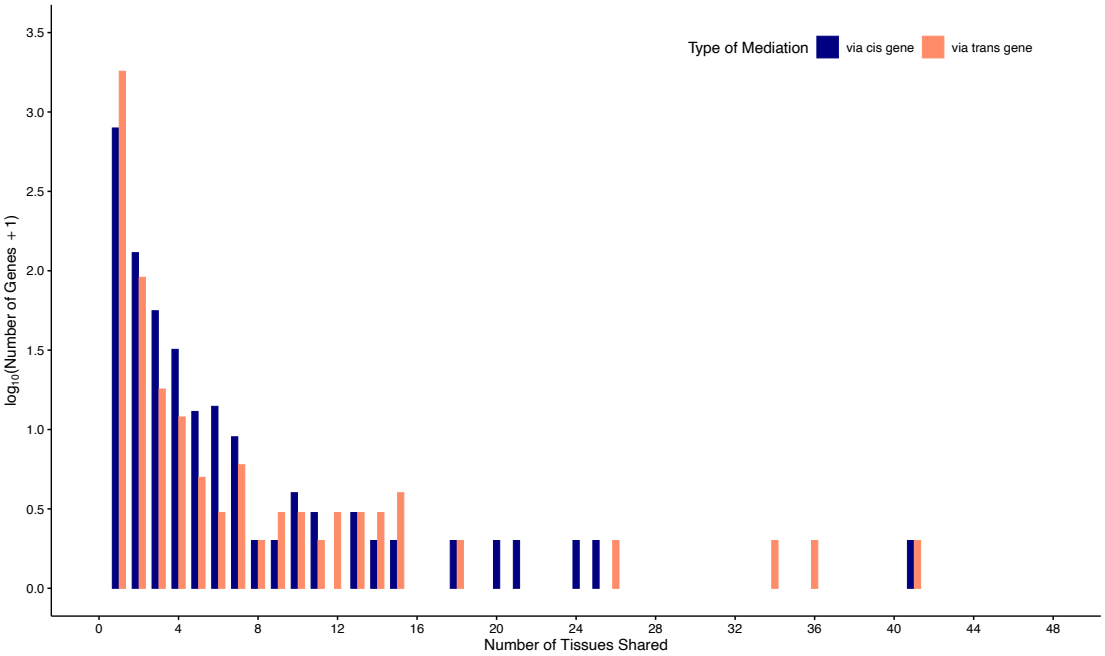



FIGURE 5. **Histograms of tissue sharing for cis-gene and trans-gene mediators**.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



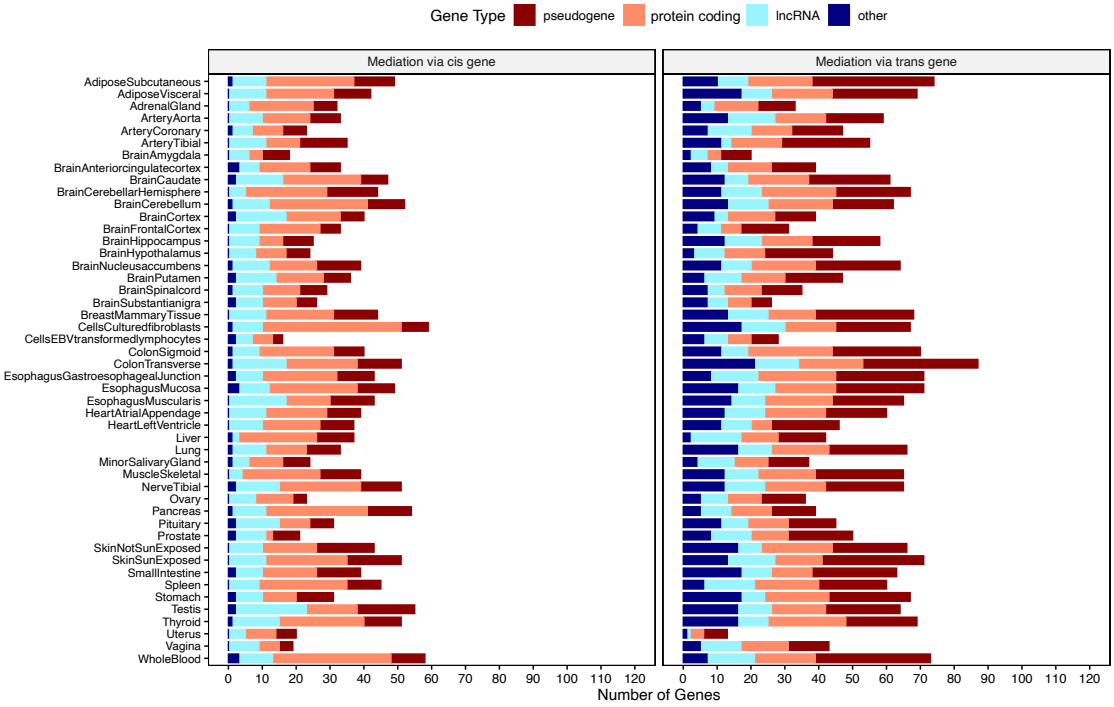



FIGURE 6. **The breakdown of types of cis-gene and trans-gene mediators across GTEx tissues and cell types**.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



SUPPLEMENTARY TABLE 1. **The breakdown of trio types inferred by MRPC-LOND across GTEx tissues and cell types**.

SUPPLEMENTARY TABLE 2. **Number of principal components (PrCs) associated with trios across tissues**.

SUPPLEMENTARY TABLE 3. **Summary of cis- and trans-gene mediation trios across GTEx tissues and cell types inferred by MRPC-LOND**.

SUPPLEMENTARY TABLE 4. **Cis-gene mediation trios with summary statistics across tissues inferred by MRPC-LOND**.

SUPPLEMENTARY TABLE 5. **Trans-gene mediation trios with summary statistics across tissues inferred by MRPC-LOND**.

SUPPLEMENTARY TABLE 6. **Counts of cis- and trans-genes as mediators**.

SUPPLEMENTARY TABLE 7. **Summary of gene types of mediators**.

SUPPLEMENTARY TABLE 8. **Gene type enrichment analysis for mediators**.

SUPPLEMENTARY TABLE 9. **HiC results in four tissues for trans-gene mediation trios inferred by MRPC-LOND**.

SUPPLEMENTARY TABLE 10. **The breakdown of trio types inferred by MRPC-ADDIS across GTEx tissues and cell types**.

SUPPLEMENTARY TABLE 11. **Cis-gene mediation trios with summary statistics across tissues inferred by MRPC-ADDIS**.

SUPPLEMENTARY TABLE 12. **Trans-gene mediation trios with summary statistics across tissues inferred by MRPC-ADDIS**.

SUPPLEMENTARY TABLE 13. **HiC results in four tissues for trans-gene mediation trios inferred by MRPC-ADDIS**.

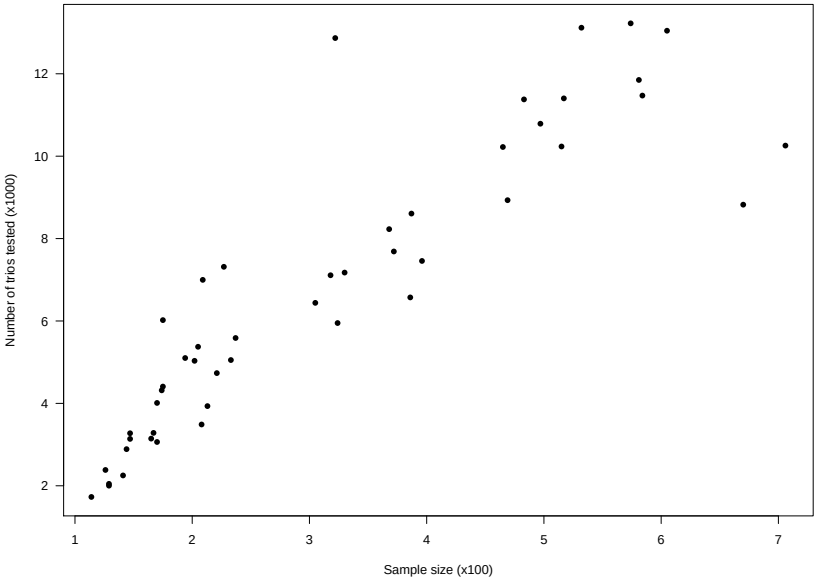
bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



SUPPLEMENTARY TABLE 14. **Mediation trios inferred by the GMAC method**.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

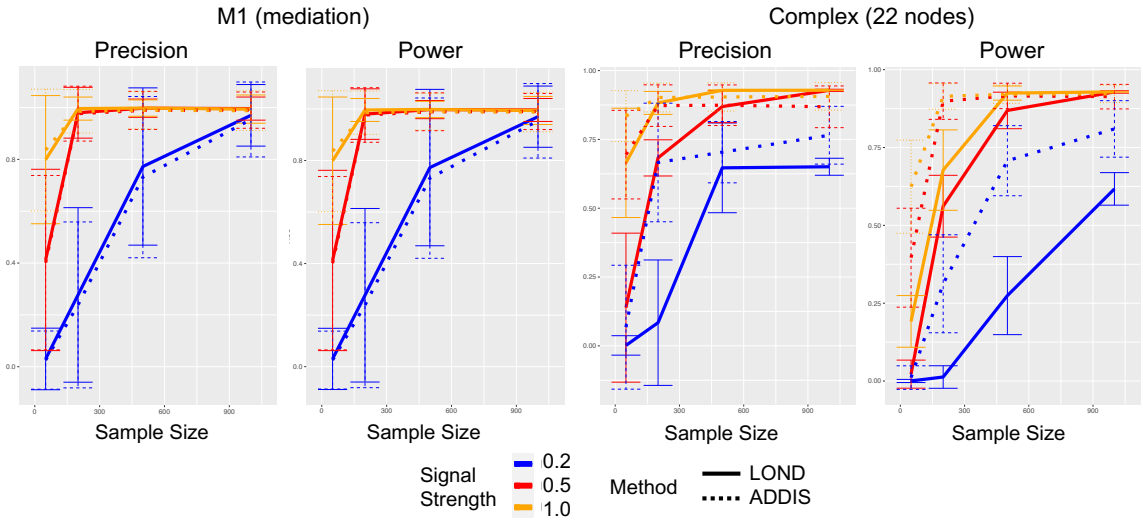





SUPPLEMENTARY FIGURE 1. **The number of trios tested versus the sample size in each tissue or cell type of the GTEx consortium**.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

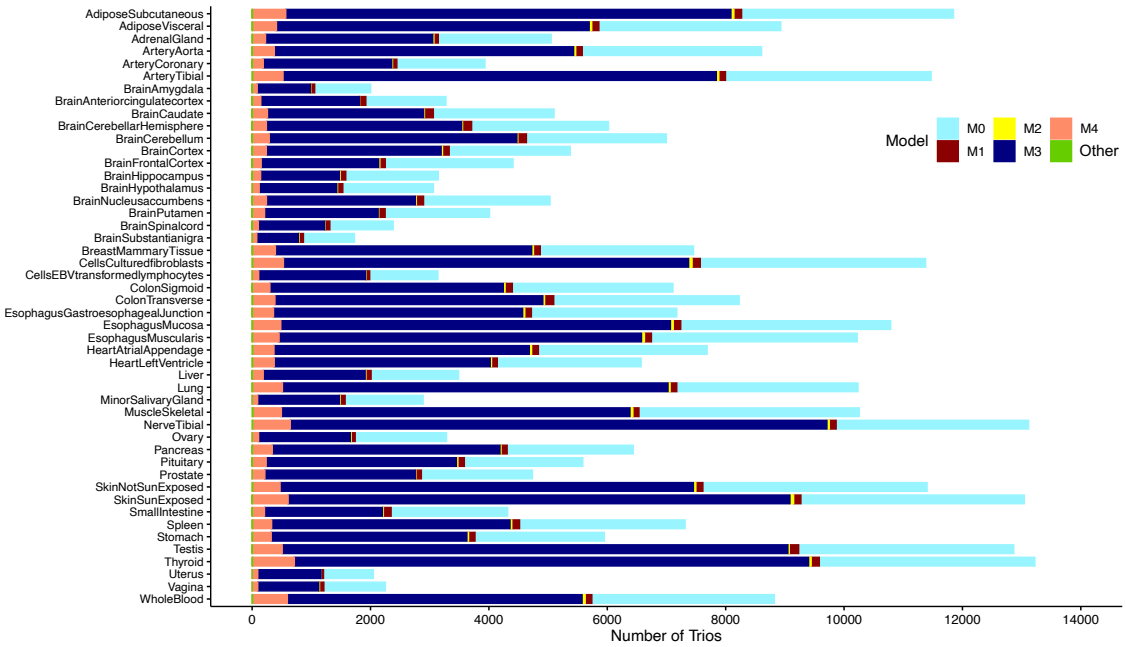





SUPPLEMENTARY FIGURE 2. **Simulation results comparing LOND and ADDIS**. See Figure 3 in [15] for the 22-node network that contains 13 genetic variants and 9 genes. Data were simulated under multiple combinations of the signal strength and sample size. The signal strength is the coefficient of the parent nodes in the linear model that generates the values for a node, whereas the sample size is the number of values at each node. We generated 1,000 independent datasets in each scenario, applied each method, and plotted the mean and a 95% interval for the precision or power across the datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

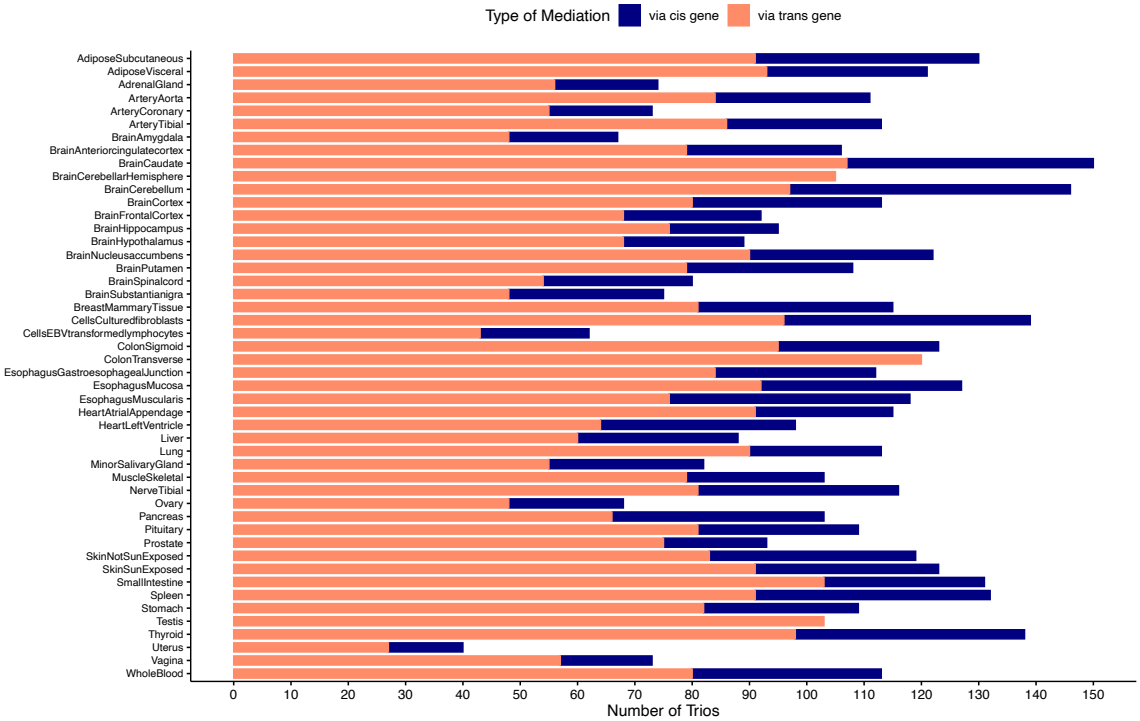





SUPPLEMENTARY FIGURE 3. **The breakdown of inferred trio types across GTEx tissues and cell types**. MRPC-ADDIS was used for inference.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

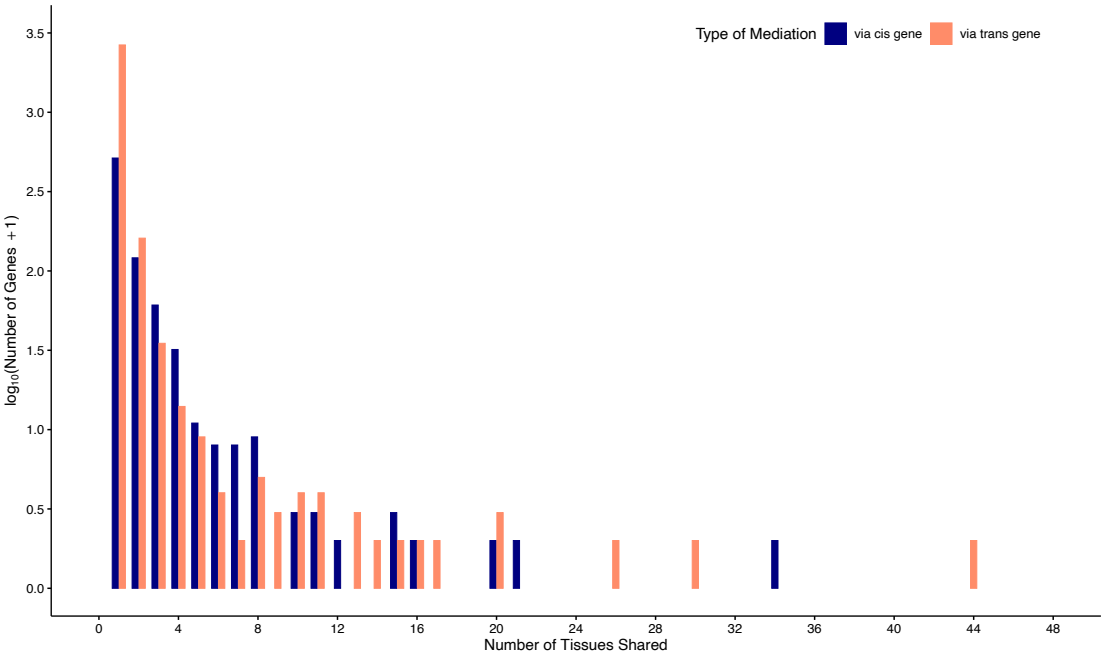





SUPPLEMENTARY FIGURE 4. **The breakdown of inferred mediation types across GTEx tissues and cell types inferred by MRPC-ADDIS**.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

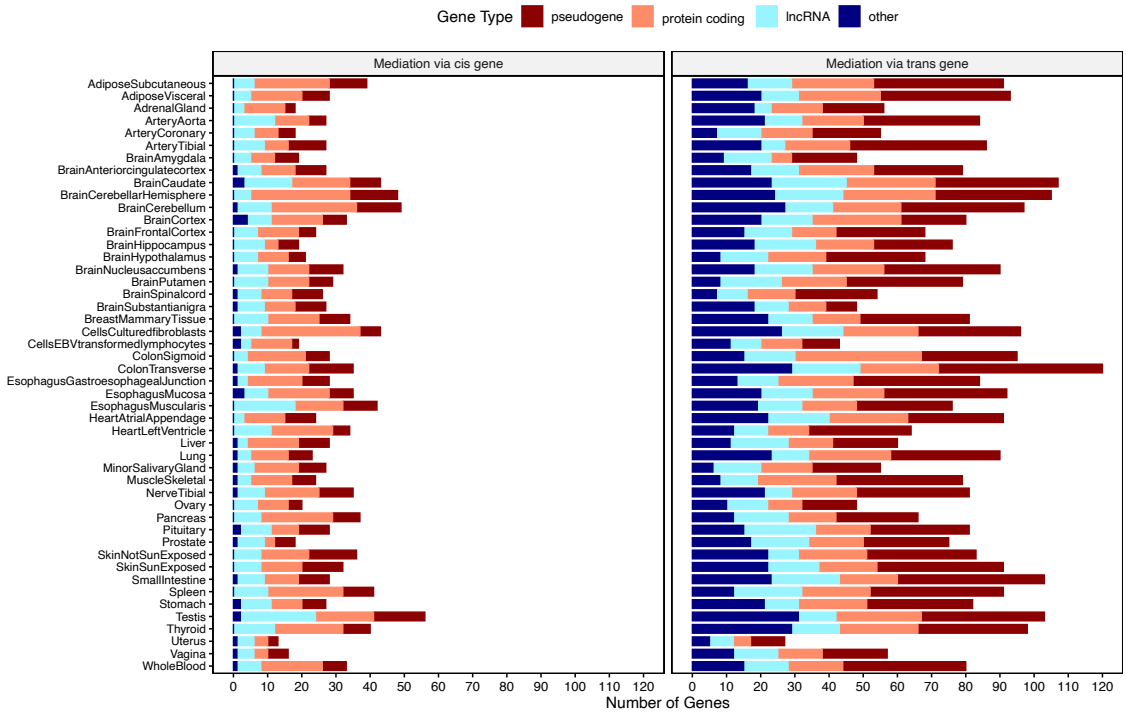





SUPPLEMENTARY FIGURE 5. **Histograms of tissue sharing for cis-gene and trans-gene mediators inferred by MRPC-ADDIS**.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477617; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.






SUPPLEMENTARY FIGURE 6. **The breakdown of types of cis-gene and trans-gene mediators inferred by MRPC-ADDIS**.