

# 5/17/2021 GMAC analysis and pvalue plots

Jarred Kvamme, University of Idaho

9/22/2021

## 1. Overview

Both MRPC-LOND and MRPC-ADDIS techniques inferred a large number of trans mediated trios. The trans-mediation model has been previously identified, but is not the commonly acknowledged mode of mediation. Since this result is surprising relative to the existing literature, we sought to apply another method for inferring mediation on a subset of GTEx trios analyzed herein by MRPC. The Genomic Mediation analysis with Adaptive Confounding (GMAC) algorithm allows for a unique selection of a subset of potential confounders,  $\mathbf{X}_{ij}$  from a larger covariate pool,  $\mathbf{H}$ , for each trio. By taking advantage of the Principle of Mendelian Randomization, the authors filter  $\mathbf{H}$  by removing common child and intermediate confounding variables (e.g variables associated with the eQTL as well as the cis/trans genes). Post-filtering, GMAC performs a mediation test on the edge between the cis gene and trans gene via the regression of the trans-gene  $T_j$  on the cis-eQTL  $L_i$ , cis-gene  $C_i$ , and the set of adaptively selected confounders  $\mathbf{X}_{ij}$ :

$$T_j = \beta_0 + \beta_1 C_i + \beta_2 L_i + \mathbf{F}\mathbf{X}_{ij} + \epsilon \quad (1)$$

The mediation statistic is the observed  $t$ -value of the cis-gene coefficient  $\beta_1$ . A null distribution for no-mediation is constructed by iteratively permuting the values of the cis-transcript within each genotype and repeating the above regression. The authors argue that the permutation of the cis-transcript within the genotypes of the cis-eQTL removes the association between the cis and trans gene transcripts while preserving the higher order associations with the cis-eQTL. The resulting permutation test for mediation compares the observed relationship between the trans and cis gene to a null distribution constructed from a model with no association and assuming that possible confounding has been well adjusted via the selected covariates.

It is important to note that the above mediation test describes only the association between cis-gene and trans-gene transcripts ( $C_i \leftrightarrow T_j$ ) and does not consider possible effects between the cis-eQTL and the cis-gene transcript ( $L_i \rightarrow C_i$ ), or the cis-eQTL and trans-gene transcript ( $L_i \rightarrow T_j$ ).

## 2. Methods

### 2.1 Applying GMAC to GTEx Trios

To compare the GMAC and MRPC algorithms, we applied the GMAC algorithm to the top five GTEx tissues by sample size. Following with the creators of GMAC, we used the full set of principle components retained from the PCA of the expression matrix as the covariate pool, and three additional known confounders: the PCR used, the platform used, and sex of the individual in each sample [yang2017identifying].

Consistent with yang2017identifying, the analysis was performed using a common child and intermediate variable filtering FDR of 10% and a confounder selection FDR of 5% for each trio. Each trio supplied to

GMAC consisted of the cis-QTL and the PEER normalized cis and trans gene transcripts with the highest association to the eQTL. To mitigate missing values in the eQTL matrix, multiple imputation of the matrix of unique cis-eQTLs was performed via multiple correspondence analysis (MCA) prior to its use in GMAC [JOSSE2016MISSMDA]. The analysis was performed twice on each trio, first with the cis gene as the mediator and second with the trans gene as the mediator. This allowed for GMAC inferred trios to be decomposed into the three groupings used under MRPC: 1) Cis-gene mediation, 2) Trans-gene mediation, 3) both (undirected).

## 2.2 Comparison of GMAC and MRPC Results

After applying GMAC to each tissue, the false discovery rate among the retained mediation p-values was controlled at the more liberal rate of 10% [YANG2017IDENTIFYING]. Each trio determined to have significant mediation after FDR filtering was compared with the regulatory network type inferred by MRPC-ADDIS. MRPC-ADDIS can infer three types of regulatory networks that contain an edge between the cis and trans gene (M1, M2, or M4). Since GMAC considers only the presence of the edge and not its direction, trios inferred to be one of M1, M2, or M4 under ADDIS, that were also significant under GMAC, were considered consistent (e.g.  $C_i \rightarrow T_j$ ;  $T_j \rightarrow C_i$ ;  $C_i \leftrightarrow T_j$  are synonymous under GMAC).

## 2.3 Simulations Using Real Trios

- The Small True Model Simulation (STM)

To further understand the conflict in edge determination between MRPC-ADDIS and GMAC, we simulated the mediation test - the test for the  $\beta_1$  coefficient in the presence of all adaptively selected confounders,  $\mathbf{X}_{ij}$ , as described by the regression in **eq (1)** - under two different scenarios. (i) To observe the predictive power of the mediation test when the trans gene comes from a set of explanatory variables smaller than those in  $\mathbf{X}_{ij}$ , We simulated the trans gene of each trio using the linear relationship:

$$T_j^* = \hat{\beta}_0 + \hat{\beta}_1 C_i + \hat{\beta}_2 L_i + \hat{\mathbf{\Gamma}}_W \mathbf{W}_{ij} + \epsilon \quad (2)$$

where  $T_j^*$  is the simulated trans gene, the coefficients are replaced by their estimates from the regression in **eq (1)**, the errors are  $\epsilon \sim N(0, \hat{\sigma})$ , and  $\mathbf{W}_{ij}$  is a subset of confounders in  $\mathbf{X}_{ij}$  representing the “highly” significant confounders from **eq (1)** ( $p < 0.001$ ). Note that if the GMAC inferred mediation type was trans gene mediation only then the cis gene was simulated and the mediation test was performed on the  $\beta_1$  coefficient from regression of the cis gene:  $C_i = \beta_0 + \beta_1 T_j + \beta_2 L_i + \mathbf{\Gamma} \mathbf{X}_{ij} + \epsilon$ .

We refer to the trans-gene generating function in **(2)** as the Small True Model (STM) as the simulated trans gene comes from a model that is a subset of the explanatory variables in the analysis model described in **(1)**. The simulated mediation test then replaces  $T_j$  in **(1)** with the simulated trans gene  $T_j^*$ . Therefore, the simulated mediation test under the STM can be decomposed as the test on the  $\beta_1$  coefficient obtained from the regression:

$$T_j^* = \beta_0 + \beta_1 C_i + \beta_2 L_i + \mathbf{\Gamma}_{1,W} \mathbf{W}_{ij} + \mathbf{\Gamma}_{2,M} \mathbf{M}_{ij} + \epsilon \quad (3)$$

where  $\mathbf{M}_{ij}$  represents the additional confounders in  $\mathbf{X}_{ij}$  that are not included in  $\mathbf{W}_{ij}$  see **Figure 1**.

- The Large True Model Simulation (LTM)

Conversely, a second simulation model was implemented to observe the power of the mediation test when the generating model for the trans gene is larger than the model used to infer the mediation relationship. In this scenario, the trans gene is simulated via:

$$T_j^* = \hat{\beta}_0 + \hat{\beta}_1 C_i + \hat{\beta}_2 L_i + \hat{\Gamma}_V \mathbf{V}_{ij} + \epsilon \quad (4)$$

which we refer to as the Large True Model (LTM). Note that the coefficient estimates in (4) come from the regression of the original data on a larger set than in **eq. (1)**:

$$T_j = \beta_0 + \beta_1 C_i + \beta_2 L_i + \Gamma \mathbf{X}_{ij} + \Gamma_G \mathbf{G}_{ij} + \epsilon \quad (5)$$

where  $\mathbf{G}_{ij}$  represents the additional explanatory variables in  $\mathbf{V}_{ij}$  not in  $\mathbf{X}_{ij}$  (see **Figure 1**). The additional variables represented by  $\mathbf{G}_{ij}$  were randomly selected from the confounder pool  $\mathbf{H}$  with equal probability. The four sets of  $p$ -values (GMAC mediation  $p$ -value, GMAC permutation  $p$ -value, the STM mediation  $p$ -value, and the LTM mediation  $p$ -value) were visually inspected to determine the effect of model mis-specification on the power of the mediation and permutation test(s).

- True GMAC Model Simulation (TGM)

We preformed a third simulation of the trans gene using the analysis model in GMAC given by (1) as the trans gene generating process. We then applied the MRPC model

$$T_j = \beta_0 + \beta_1 C_i + \beta_2 L_i + \Gamma_z \mathbf{Z}_{ij} \quad (6)$$

where the column dimension  $\mathbf{Z}_{ij}$  is a less than or equal to the column dimension of  $\mathbf{W}_{ij}$  and represents only the confounders selected under the MRPC PC-selection methodology. The goal was to analyze the power of MRPC in detecting the mediation edge when the data is generated from the much larger GMAC model.

- GMAC Self Simulation (GSS)

We created a fourth simulation scenario to test the importance of permutation on the analysis result. Each of the above simulations seeks to understanding the inferential power of MRPC/GMAC under model mis-specification. In this scenario we take a different approach and use the model suggested by GMAC for each trio given by **eqn. (1)** to generate the trans gene such that:

$$T_j^* = \hat{\beta}_0 + \hat{\beta}_1 C_i + \hat{\beta}_2 L_i + \hat{\Gamma} \mathbf{X}_{ij} + \epsilon$$

where  $\epsilon \sim N(0, \hat{\sigma})$ . We then apply the correct model (1) to each simulated trio under this scenario to obtain the parametric and permutation inferences when we know the analysis model is correct.

## 2.4 Simulations Using Synthetic Trios

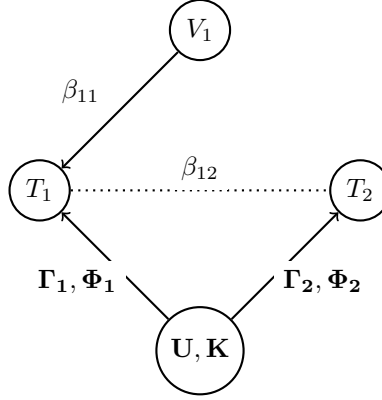
- **Purpose:** We simulated 10,000 trios of all model types (M0 - M4) with  $u$  “unknown” confounding variables (PCs) represented by the matrix  $\mathbf{U}$  and 3 simulated “known” confounders represented by the matrix  $\mathbf{K}$ .

(1.) - We simulate the variant  $V_1 = [v_1, v_2, \dots, v_n]$  from a multinomial distribution under Hardy-Weinberg equilibrium:

$$Pr(v_i = 0) = (1 - \theta)^2; \quad Pr(v_i = 1) = 2\theta(1 - \theta); \quad Pr(v_i = 2) = \theta^2$$

(2.) - We denote the  $j^{th}$  molecular phenotype as  $T_j$  and its set of parents by  $\mathbf{Parents}(T_j)$ . The set of parent nodes  $\mathbf{Parents}(T_j)$  may be empty, contain  $V$ -nodes,  $T$ -nodes, or a combination of both (Badsha and Fu, 2019). We model  $T_j$  as following a normal distribution where

$$T_j \sim N \left( \beta_0 + \sum_{i \in \text{Parents}(T_j)} \beta_i V_i + \left( \sum_{k \in \text{Parents}(T_j)} \beta_k T_k \right) + \mathbf{\Gamma} \mathbf{U} + \mathbf{\Phi} \mathbf{K}, \sigma_j \right) \quad (7)$$



### 3. Results

#### 3.1 Comparing GMAC and MRPC

In light of the surprising number of trans-gene mediation trios inferred by MRPC, we sought to compare our results with GMAC by applying the GMAC method to the top five GTEx tissues by sample size. It is important to note that the test for mediation used by GMAC describes only the association between cis-gene and trans-gene transcripts ( $C_i \leftrightarrow T_j$ ) and does not consider the possible effects between the cis-eQTL and the cis-gene ( $L_i \rightarrow C_i$ ), or the cis-eQTL and trans-gene ( $L_i \rightarrow T_j$ ). Therefore, since GMAC considers only the presence of the mediation edge, trios inferred to be one of M1, M2, or M4 under ADDIS, that were also significant under GMAC, were considered consistent (e.g  $C_i \rightarrow T_j$ ;  $T_j \rightarrow C_i$ ;  $C_i \leftrightarrow T_j$  are synonymous under GMAC).

At the 10% false discovery rate, GMAC identified 2,160 trios with an edge between the cis and trans genes out of 55,446 total trios tested across the five selected tissues: Adipose subcutaneous, Tibial artery, Muscle skeletal, Sun exposed skin, and Whole blood. Of the trios with mediation edges, 653 were identified as the cis gene mediating the trans gene, 245 as trans gene mediating the cis gene and 1,345 as both (29.1%, 10.9%, and 60% respectively). As can be seen from **Table 2**, the consistency in inferred mediation edges between the two methods varied between 39% and 50% of the trios across tissues.

#### 3.2 Simulation Results

- - We noticed that the mediation parametric p-value can vary quite a bit. we ran each simulation scenario 1000 times on each trio and obtained the median p-value from each of the 4 scenarios. We used this as a robust estimate of the parametric p-value
- - The nominal p-value and mediation p-value give more or less the same value except for rare variant trios
- - We noticed that Badsha implemented the PC selection code for MRPC in a way that led to the qvalues of each correlation test being applied after the p-values had already been bonferroni adjusted. This lead to extremely few PCs being selected for each trio. After implementing the correct PC selection process, MRPC and GMAC select equivalent numbers of PCs

- - Figure 1 panel D supports that the nominal p-value appears to be better at determining the presence of a mediation edge for rare variants (points on the nominal p axis are all significant where as some trios/rare variants are insignificant on the parametric axis). This supports the use of the permuted regression test when the instrumental variable/ cis-eQTL contains a rare allele
- - Figures 4-5 and Tables 4-9 represent two cases where the marginal tests performed by MRPC falsely remove an edge in the graph that is supported by higher order tests. These cases justify the need to bypass the marginal tests and proceed directly to conditional tests which have increased power in these scenarios.

# 1 Tables and Figures

Table 1: Descriptive statistics for the distribution of missing values across the eQTL’s for each tissue used in GMAC

|         | Adipose<br>Subcutaneous | Artery Tibial | Muscle Skeletal | Skin Sun Exposed | Whole<br>Blood |
|---------|-------------------------|---------------|-----------------|------------------|----------------|
| Min.    | 0.000000                | 0.000000      | 0.000000        | 0.000000         | 0.000000       |
| 1st Qu. | 0.000000                | 0.000000      | 0.000000        | 0.000000         | 0.000000       |
| Median  | 0.000000                | 0.000000      | 0.000000        | 0.000000         | 0.000000       |
| Mean    | 0.006365                | 0.006625      | 0.006560        | 0.006701         | 0.006103       |
| 3rd Qu. | 0.003442                | 0.003425      | 0.002833        | 0.003306         | 0.002985       |
| Max.    | 0.156627                | 0.159247      | 0.158640        | 0.160331         | 0.155224       |

Table 2: The breakdown of unique trios with inferred significant cis or trans mediation under GMAC across their respective ADDIS inferred regulatory networks. The column “Percentage In Common” is the proportion of significant trios that also contained a mediation edge in the regulatory network inferred under ADDIS

| Tissue              | M0  | M1 | M2 | M3  | M4  | Other | Total GMAC<br>Inferred | Percentage In<br>Common |
|---------------------|-----|----|----|-----|-----|-------|------------------------|-------------------------|
| AdiposeSubcutaneous | 107 | 47 | 12 | 102 | 134 | 2     | 404                    | 0.4777                  |
| ArteryTibial        | 88  | 37 | 10 | 108 | 112 | 0     | 355                    | 0.4479                  |
| MuscleSkeletal      | 126 | 37 | 8  | 145 | 132 | 2     | 450                    | 0.3933                  |
| SkinSunExposed      | 118 | 42 | 12 | 132 | 139 | 0     | 443                    | 0.4357                  |
| WholeBlood          | 107 | 55 | 29 | 142 | 171 | 4     | 508                    | 0.5020                  |

Table 3: Breakdown of trios with inferred mediation under GMAC across both cis and trans mediation types. The column “Unique Both” represents the intersection of the columns “Total Cis Mediated” and “Total Trans Mediated”

| Tissue         | Sample<br>Size | Tested<br>Trios | Cis-Gene<br>Mediation<br>Trios | Trans-Gene<br>Mediation Trios | Unique<br>Cis | Unique<br>Trans | Unique<br>Both | Unique<br>Total |
|----------------|----------------|-----------------|--------------------------------|-------------------------------|---------------|-----------------|----------------|-----------------|
| AdiposeSub     | 581            | 11850           | 374                            | 282                           | 122           | 30              | 252            | 404             |
| ArteryTibial   | 584            | 11471           | 334                            | 226                           | 129           | 21              | 205            | 355             |
| MuscleSkeletal | 706            | 10257           | 401                            | 314                           | 136           | 49              | 265            | 450             |
| SunExp.Skin    | 605            | 13045           | 404                            | 333                           | 110           | 39              | 294            | 443             |
| WholeBlood     | 670            | 8823            | 451                            | 398                           | 110           | 57              | 341            | 508             |

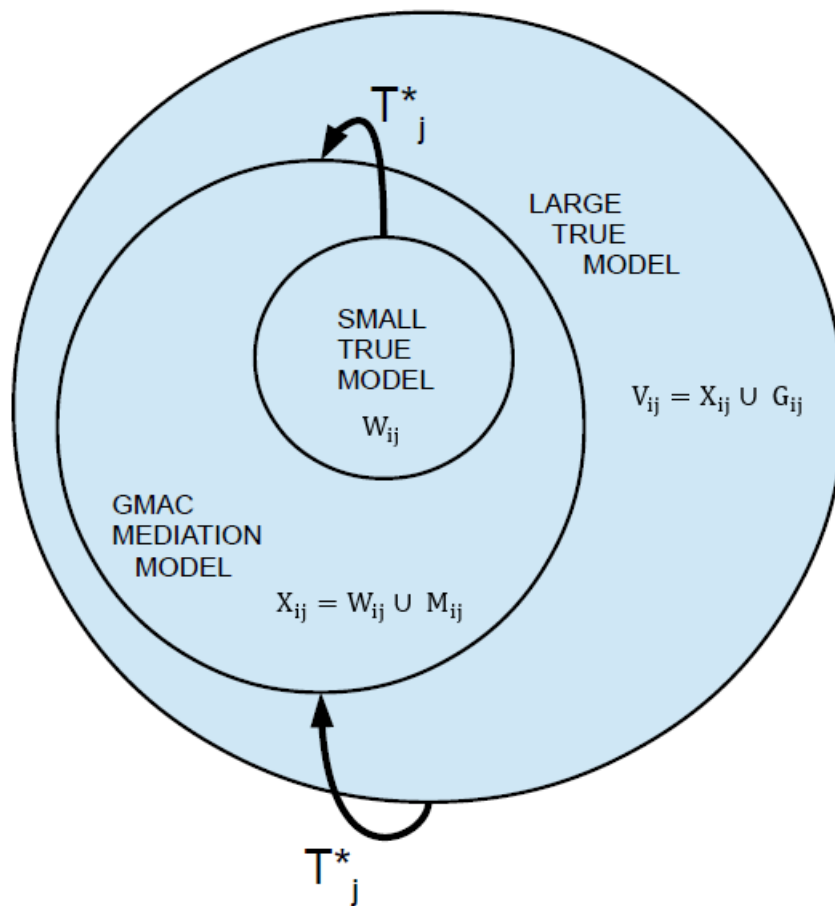


Figure 1: A visual depiction of the relationship between the STM and LTM simulation models for the trans gene and the regression used for the mediation test

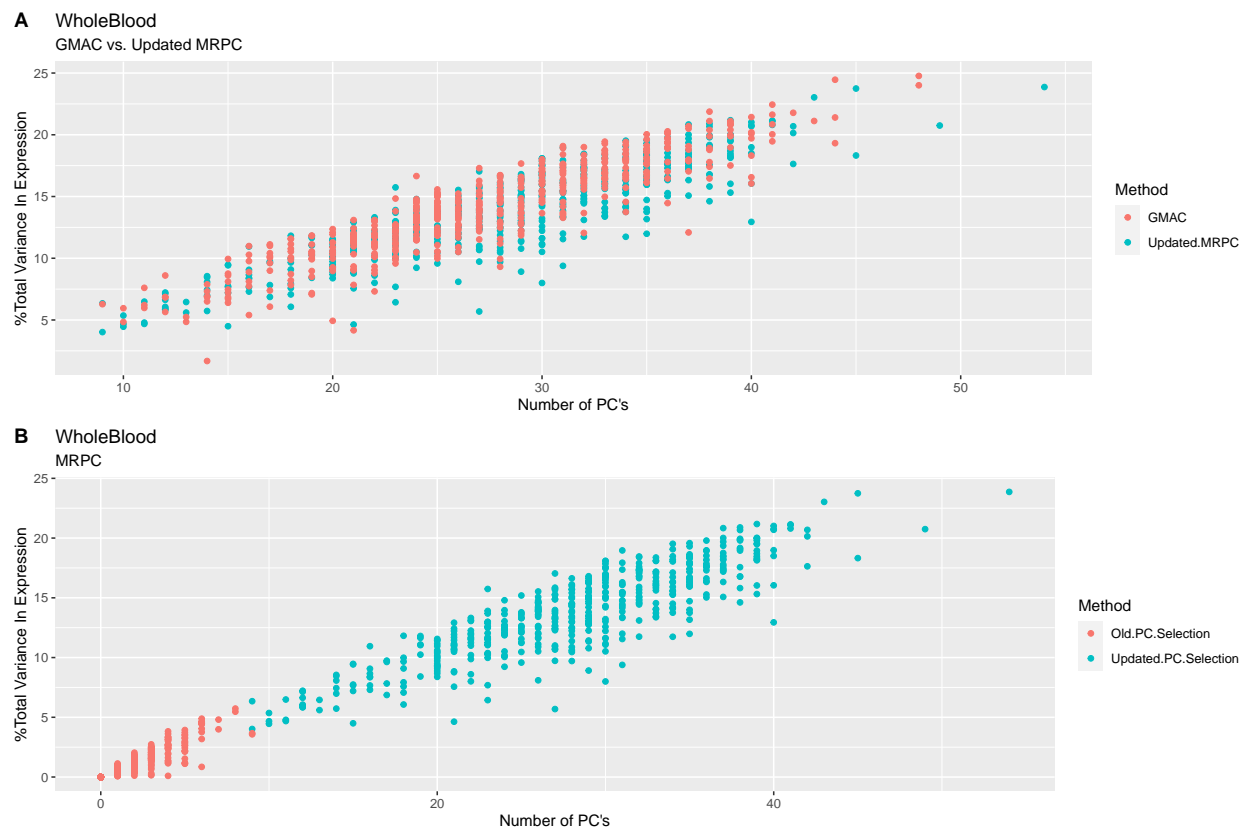


Figure 2: A) The number of PC's included for each trio in whole blood after using the corrected PC selection for MRPC with the number of PC's selected by GMAC. B) compares the number of PC's selected for each trio under Badsha's original code vs the corrected version of PC selection



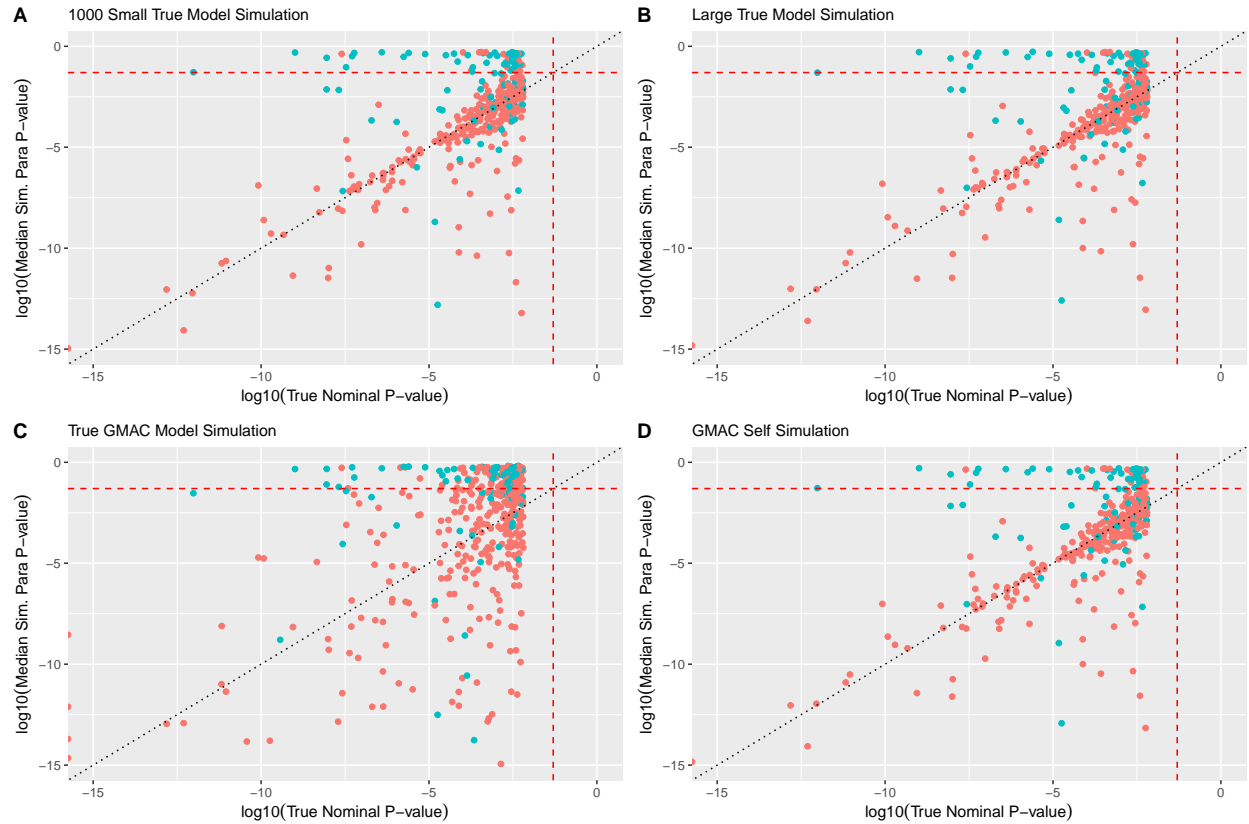
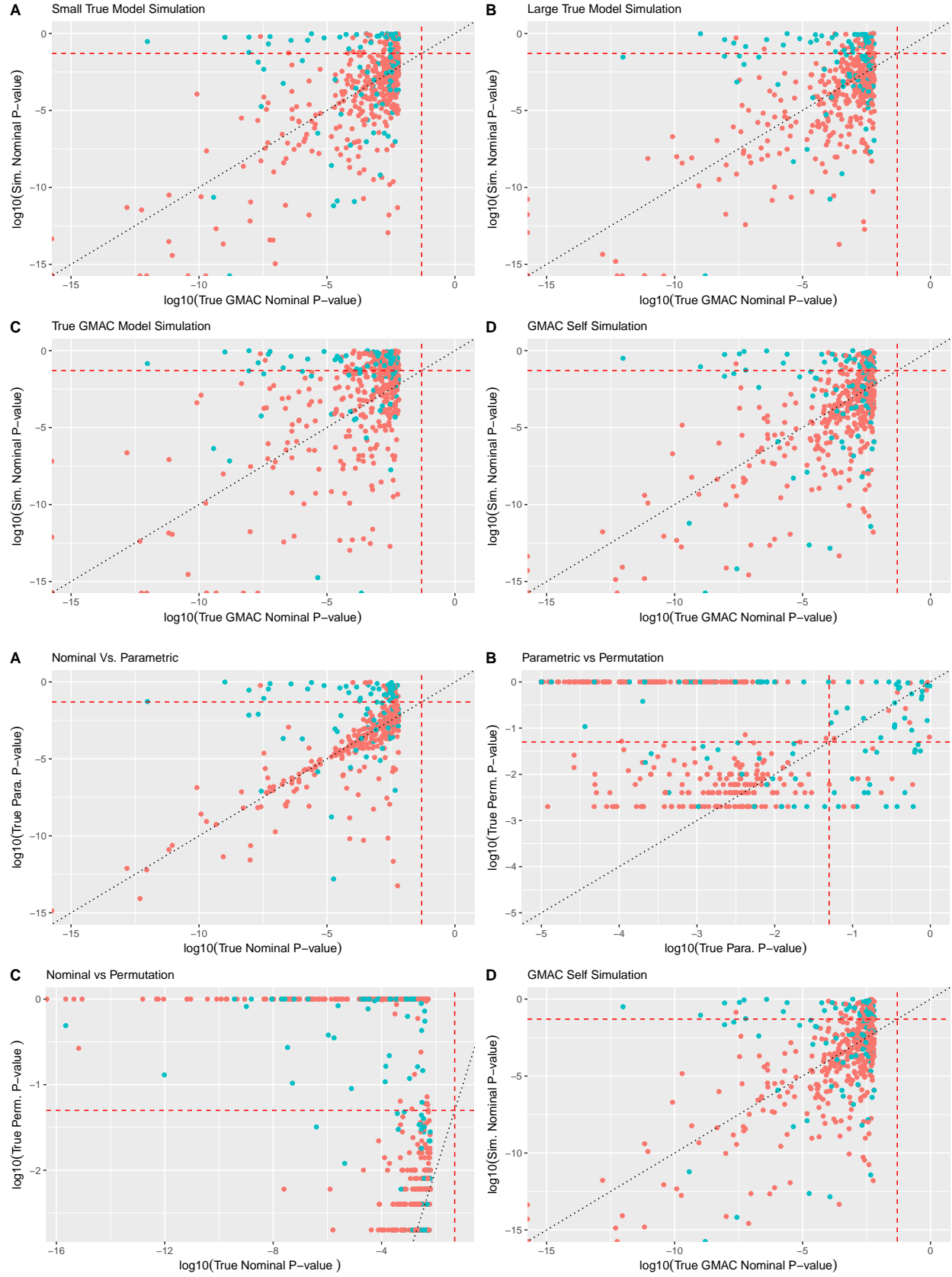


Figure 3: All panels: comparison of the median  $p$ -value from the test for the mediation edge from 1000 iterations of each simulation scenario for each trio (y-axis) with the nominal  $p$ -value for the edge from the true model (x-axis). Each panel corresponds to a specific simulation scenario. These graphs support the need to use the permutation test when the genotype has a low frequency allele.



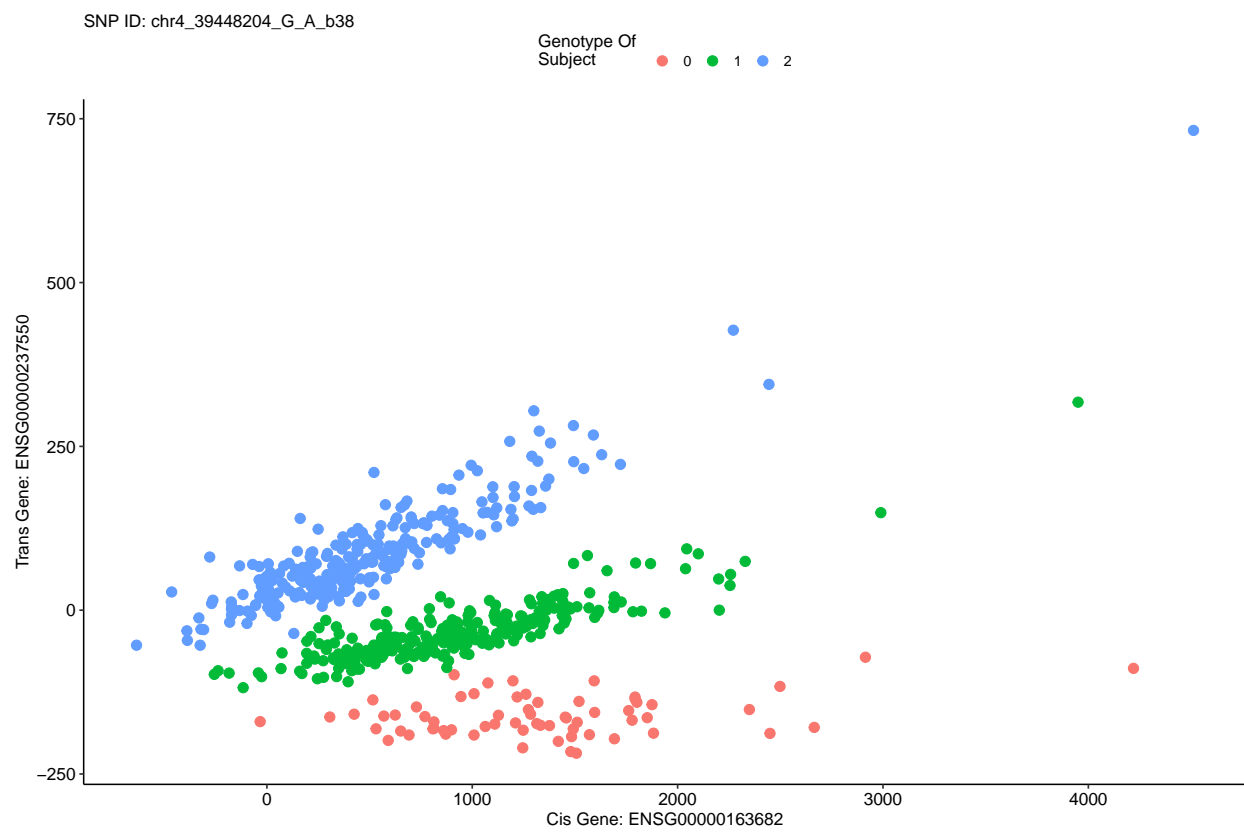


Figure 4: Trio number 9 from the GTEx tissue Adipose Subcutaneous

Table 4: Results showing the terms of interest for the Regression:  
 $\text{cis.gene} \sim \text{SNP} + \text{trans.gene} + \text{PCs}$  for trio 9 from the GTEx tissue  
 Adipose Subcutaneous

|             | Estimate   | Std. Error | t value  | Pr(> t ) |
|-------------|------------|------------|----------|----------|
| (Intercept) | 2070.6858  | 75.9773    | 27.2540  | 0        |
| trans.gene  | 5.0385     | 0.2468     | 20.4178  | 0        |
| SNP         | -1050.0066 | 35.6044    | -29.4909 | 0        |

Table 5: Results showing the terms of interest for the Regression:  
 $\text{trans.gene} \sim \text{SNP} + \text{cis.gene} + \text{PCs}$  for trio 9 from the GTEx tissue  
 Adipose Subcutaneous

|             | Estimate  | Std. Error | t value  | Pr(> t ) |
|-------------|-----------|------------|----------|----------|
| (Intercept) | -260.4547 | 10.3493    | -25.1664 | 0        |
| cis.gene    | 0.0856    | 0.0042     | 20.4178  | 0        |
| SNP         | 159.5456  | 3.0487     | 52.3325  | 0        |

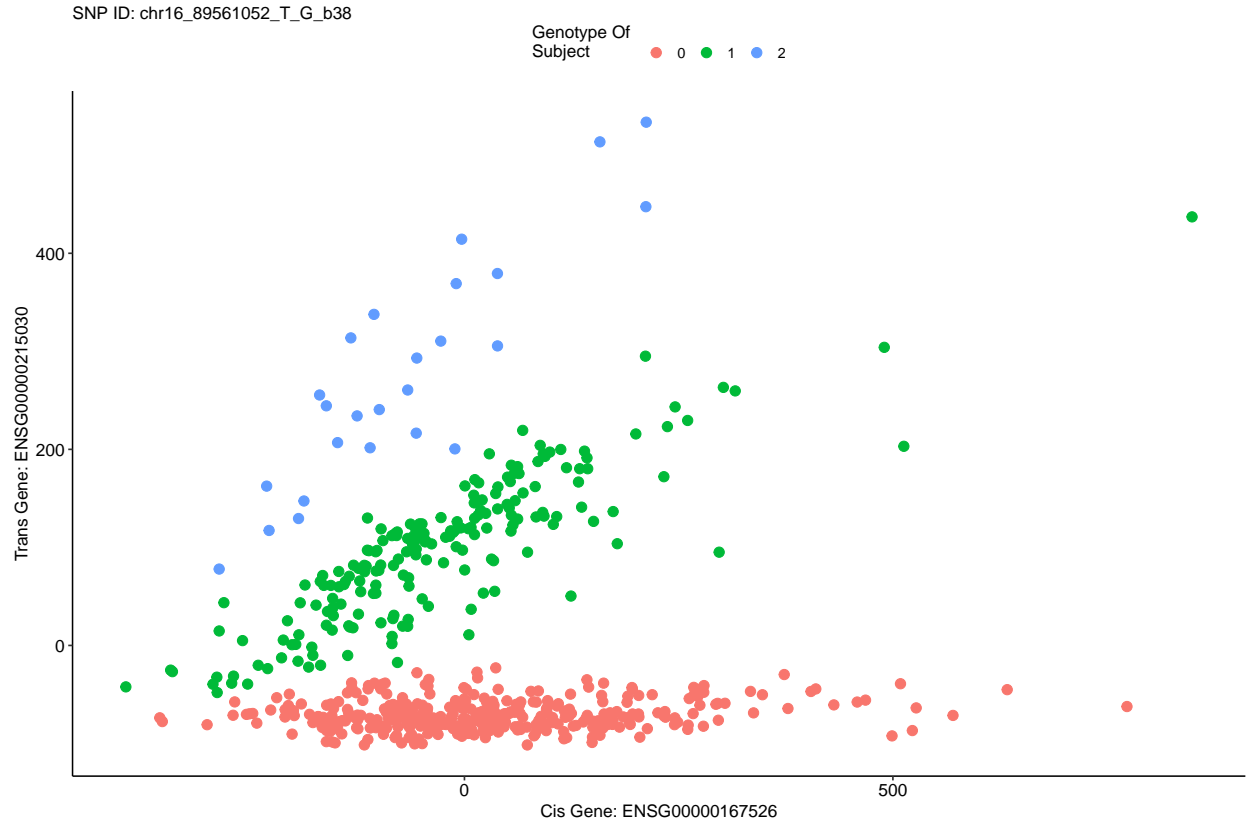


Figure 5: Trio number 1922 from the GTEx tissue Artery Tibial

Table 6: Results showing the terms of interest for the Regression:  $\text{cis.gene} \sim \text{SNP} + \text{trans.gene} + \text{PCs}$  for trio 1922 from the GTEx tissue Artery Tibial

|             | Estimate  | Std. Error | t value  | $\text{Pr}( >  t  )$ |
|-------------|-----------|------------|----------|----------------------|
| (Intercept) | 41.2423   | 27.1619    | 1.5184   | 0.1295               |
| trans.gene  | 1.0308    | 0.1027     | 10.0398  | 0.0000               |
| SNP         | -222.7941 | 19.8243    | -11.2384 | 0.0000               |

Table 7: Results showing the terms of interest for the Regression:  $\text{trans.gene} \sim \text{SNP} + \text{cis.gene} + \text{PCs}$  for trio 1922 from the GTEx tissue Artery Tibial

|             | Estimate | Std. Error | t value | $\text{Pr}( >  t  )$ |
|-------------|----------|------------|---------|----------------------|
| (Intercept) | -90.2013 | 9.5820     | -9.4136 | 0                    |
| cis.gene    | 0.1480   | 0.0147     | 10.0398 | 0                    |
| SNP         | 180.7298 | 3.2883     | 54.9611 | 0                    |

Table 8: Correlations between the cis and trans genes trio 9 from the GTEx tissue Adipose Subcutaneous and trio 1922 from Artery Tibial. Notice that the correlation between the cis and trans gene is quite weak despite an obvious within-genotype correlation between the two molecular phenotypes. Thus MRPC removes this edge in the marginal tests (given by the pvalues) despite being clearly supported in the higher order conditional tests.

|               | cor    | df  | t      | p-value |
|---------------|--------|-----|--------|---------|
| Trio 9 AS     | 0.0681 | 579 | 1.6417 | 0.1012  |
| Trio 1922 ArT | 0.0561 | 582 | 1.3546 | 0.1761  |

|               | cor       | df  | t        | p-value   |
|---------------|-----------|-----|----------|-----------|
| Trio 1922 ArT | 0.0560624 | 582 | 1.354617 | 0.1760653 |

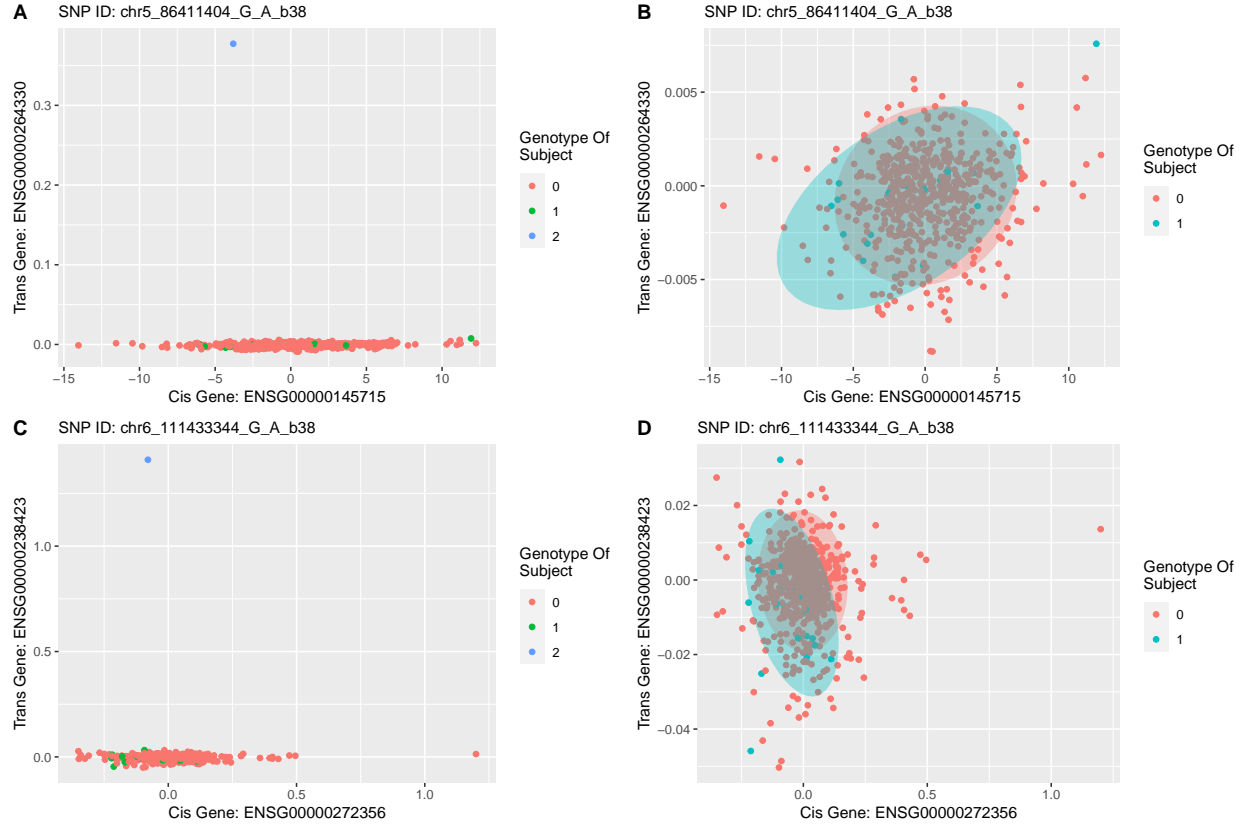


Figure 6: **A** and **C**: Example scatter plots of trios from subcutaneous adipose tissue and whole blood (respectively) with a rare allele present in the sample: Note that 0 indicates individuals homozygous for the reference allele, 1 indicates heterozygous individuals and 2 indicates individuals who are homozygous for the alternative (rare) allele. The apparent outlier represents a single individual in the sample who was homozygous for the rare allele, and **B** and **D** are the scatter plots with the point(s) for the homozygous-alternative individuals removed and a confidence ellipse calculated over the remaining homozygous-reference and heterozygous individuals.