A Comparison of GMAC and ADDIS

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 ackowledged 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 preforms 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 $\mathbf{X_{ij}}$:

$$T_j = \beta_0 + \beta_1 C_i + \beta_2 L_i + \mathbf{\Gamma X}_{ij} + \epsilon \tag{1}$$

The mediation statistic is the observed t-value of the cis-gene coefficient β_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 (Yang et al. 2017).

Consistent with Yang et al. (2017), the analysis was preformed 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 preformed via multiple correspondence analysis (MCA) prior to its use in GMAC (Josse, Husson, and others 2016). The analysis was preformed 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% (Yang et al. 2017). 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

• 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 β_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:

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

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 preformed on the β_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 β_1 coefficient obtained from the regression:

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

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:

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

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)**:

$$T_j = \beta_0 + \beta_1 C_i + \beta_2 L_i + \mathbf{\Gamma} \mathbf{X}_{ij} + \mathbf{\Gamma}_G \mathbf{G}_{ij} + \epsilon \tag{5}$$

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 and final 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

$$T_j = \beta_0 + \beta_1 C_i + \beta_2 L_i + \mathbf{\Gamma}_z \mathbf{Z}_{ij} \tag{6}$$

where $\mathbf{Z}_{ij}$ is a subset of both bfX_{ij} and W_{ij} representing only the confounders selected under the MRPC methodology. The goal was to analyze the power of MRPC when infer the larger GMAC model under the assumption that the GMAC model is the true.

• GMAC Self Simulation (GSS)

We created a fourth simulation scenario to test the importance of permutation on the analysis result. Each of the above simualtions uses variations understanding the inferential power of MRPC/GMAC when we know we have the incorrect model. 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:

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

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

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.

To uncover the computational differences between MRPC and GMAC, we focused on trios with conflicting results between the two methods (trios inferred M0 or M3 under MRPC). The primary differences we observed between the two algorithms for these trios were that the inclusion of a larger set of confounding variables by GMAC often had the effect of strengthening the association between the cis and trans genes. That is, let $\mathbf{Z}_{ij}$ denote the set of confounding variables used under MRPC and $\mathbf{X}_{ij}$ the larger set used by GMAC such that the columns $\{z_{ij}\} \subset \{x_{ij}\}$. Because the confounding variables under GMAC are selected such that they have a significant association with either the cis trans trans genes, the partial correlation between the cis gene and trans gene tends to strengthen as the column dimension of $\mathbf{Z}_{ij}$ approaches the column dimension of $\mathbf{X}_{ij}$. The result for GMAC is that under the mediation test, relatively weak associations ($0.1 \leq \rho \leq 0.2$) can be deemed significant. Conversely, for MRPC, as the size of the network increases, the method becomes increasingly conservative. Therefore, when $\mathbf{Z}_{ij} = \mathbf{X}_{ij}$ MRPC tends to infer the null model unless the association between two nodes in the network is substantial.

3.2 Simulation Results

To observe the predictive differences between the MRPC-ADDIS and GMAC methods, we simulated the GMAC mediation test under two models for the trans gene: (i) when the true model generating the trans gene is smaller than the analysis model (STM), and (ii) when the model generating the trans gene is larger than the analysis model (LTM) see section **2.3** for details. The STM simulated mediation test from (i) represents the case when the analysis model includes additional confounders not involved in the true trans gene generating process (i.e when we include more confounders than needed). Alternatively, the LTM simulated mediation test from (ii) represents a scenario similar to the test preformed by MRPC-ADDIS where the number of confounders included in the mediation test is a subset of the confounders in the model generating the trans gene with the greatest contribution to the response (i.e when we don't include enough confounders).

In general, over-specifying or under-specifying the model does little to change the inference of the mediation test (**Figure 4**) when the number of confounders is already large (relative to the MRPC model). However,

The blue dots in **Figure 4** distinguish specific trios which contain an eQTL with a rare allele present in the sample (typically with an allele frequency $\approx 1\%$).

Tables and Figures

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

	Adipose Subcutaneous	Artery Tibial	Muscle Skeletal	Skin Sun Exposed	Whole 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 unqiue 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 Inferred	Percentage In 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	Total Trios	Total Cis Mediated	Total Trans Mediated	Unique Cis Only	Unique Trans Only	Unique Both	Unique Total
AdiposeSubcut	11850	374	282	122	30	252	404
ArteryTibial	11471	334	226	129	21	205	355
MuscleSkeletal	10257	401	314	136	49	265	450
SkinSunExposed	13045	404	333	110	39	294	443
WholeBlood	8823	451	398	110	57	341	508

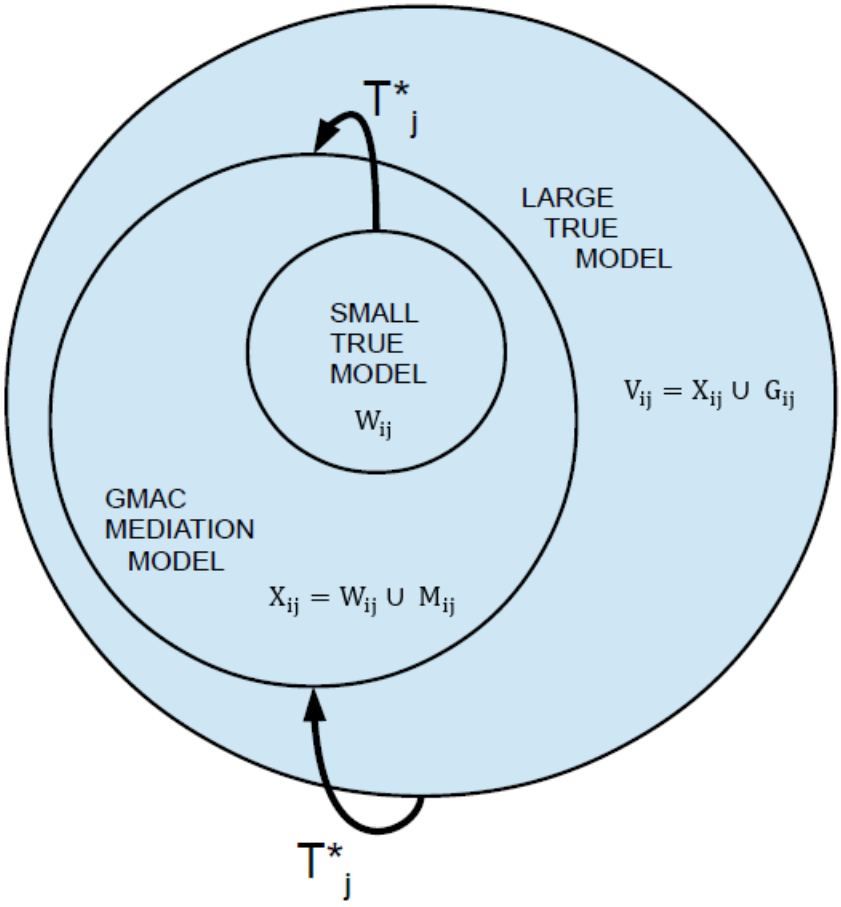


Figure 1: A visual depiction of the relationship between the simulation models for the trans gene and the regression used for the mediation test

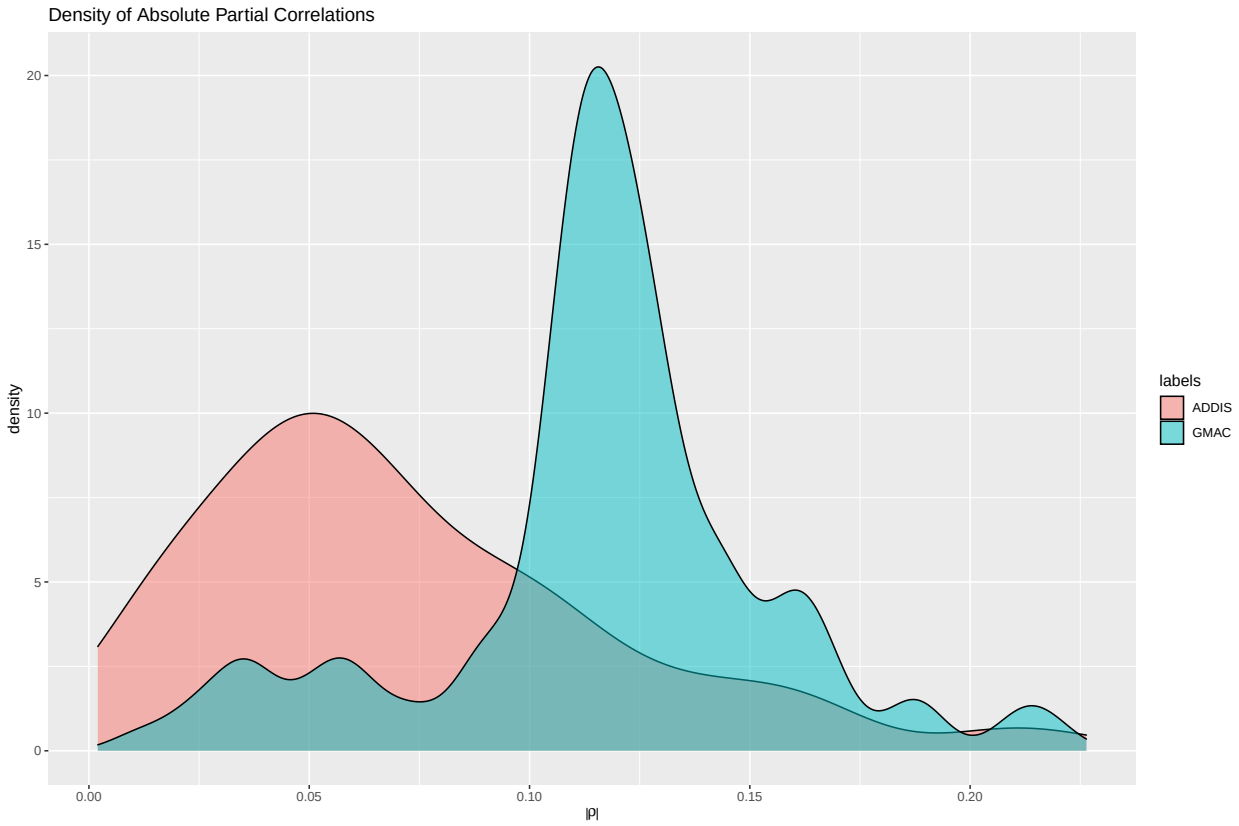
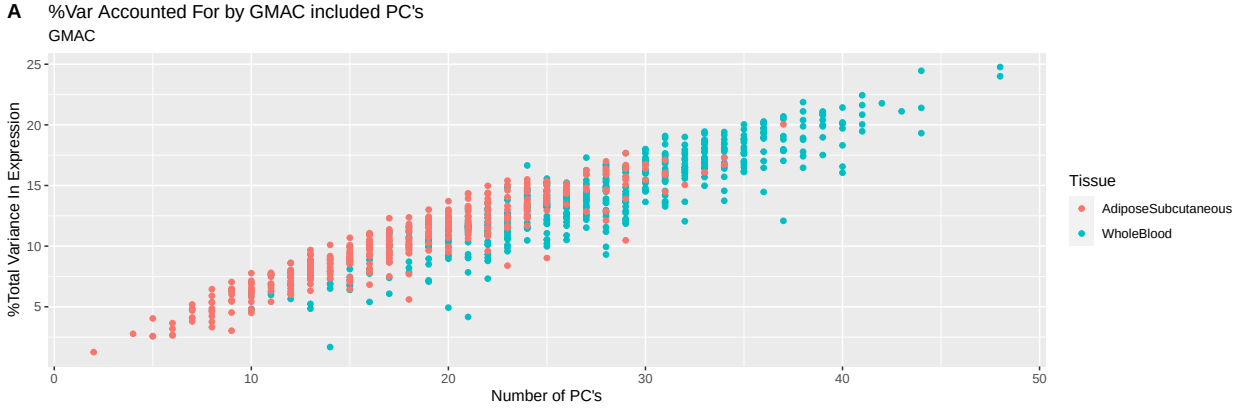


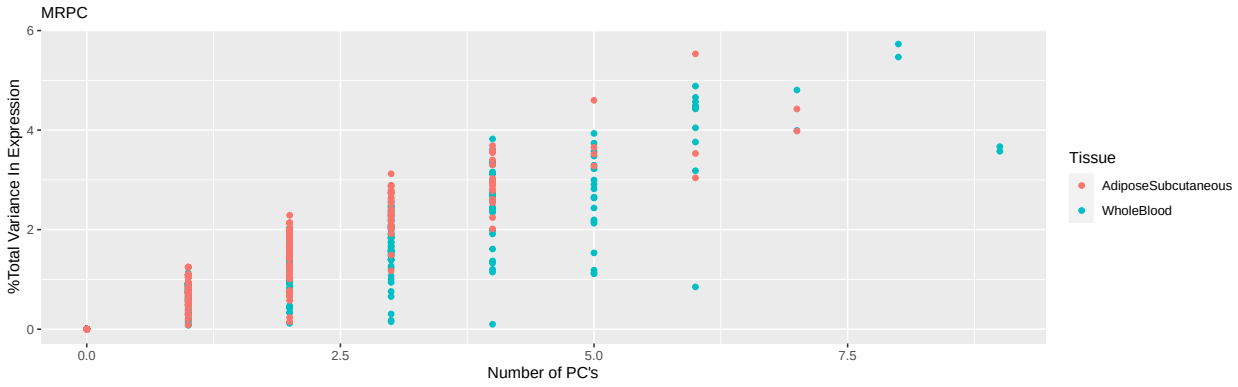
Figure 2: A density plot for the absolute partial correlations to demonstrate the strengthening of the adjusted relationship under the GMAC mediation model.

variance across tissues plots

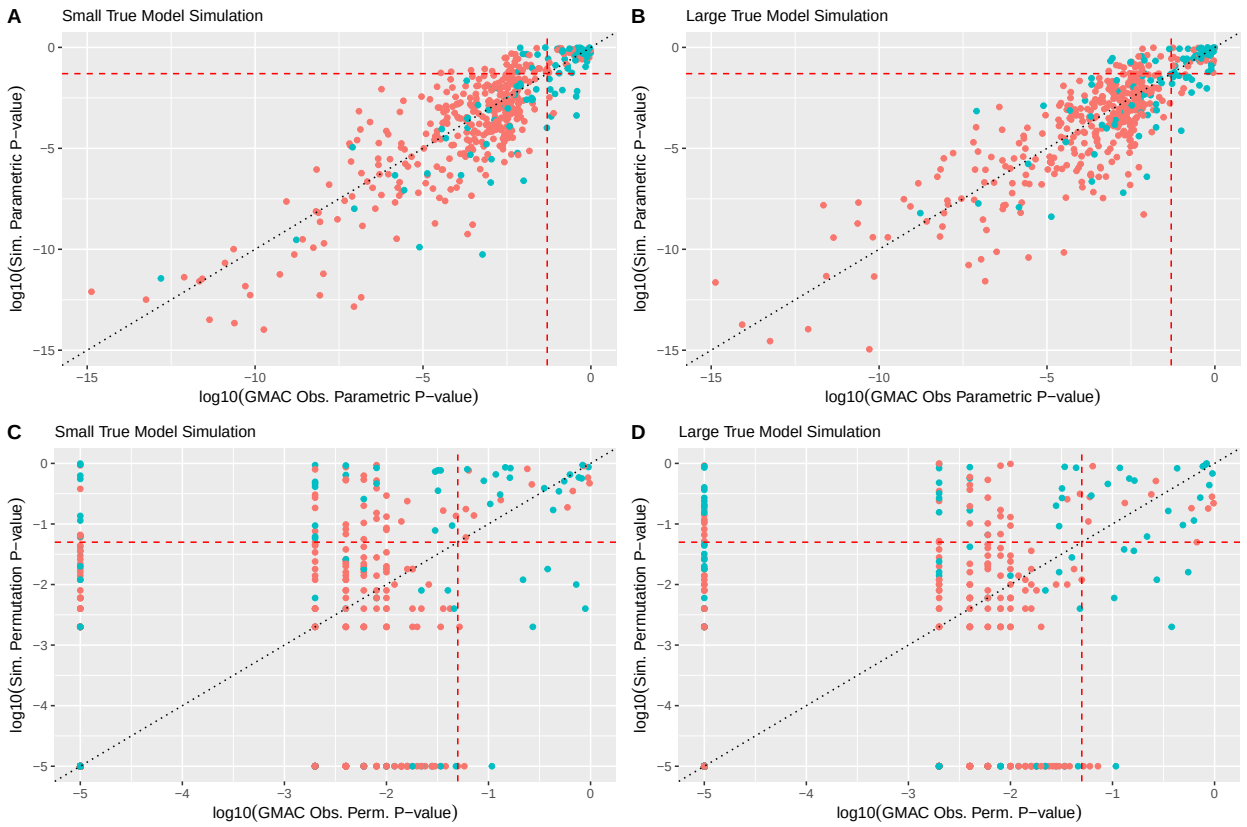
A %Var Accounted For by GMAC included PC's



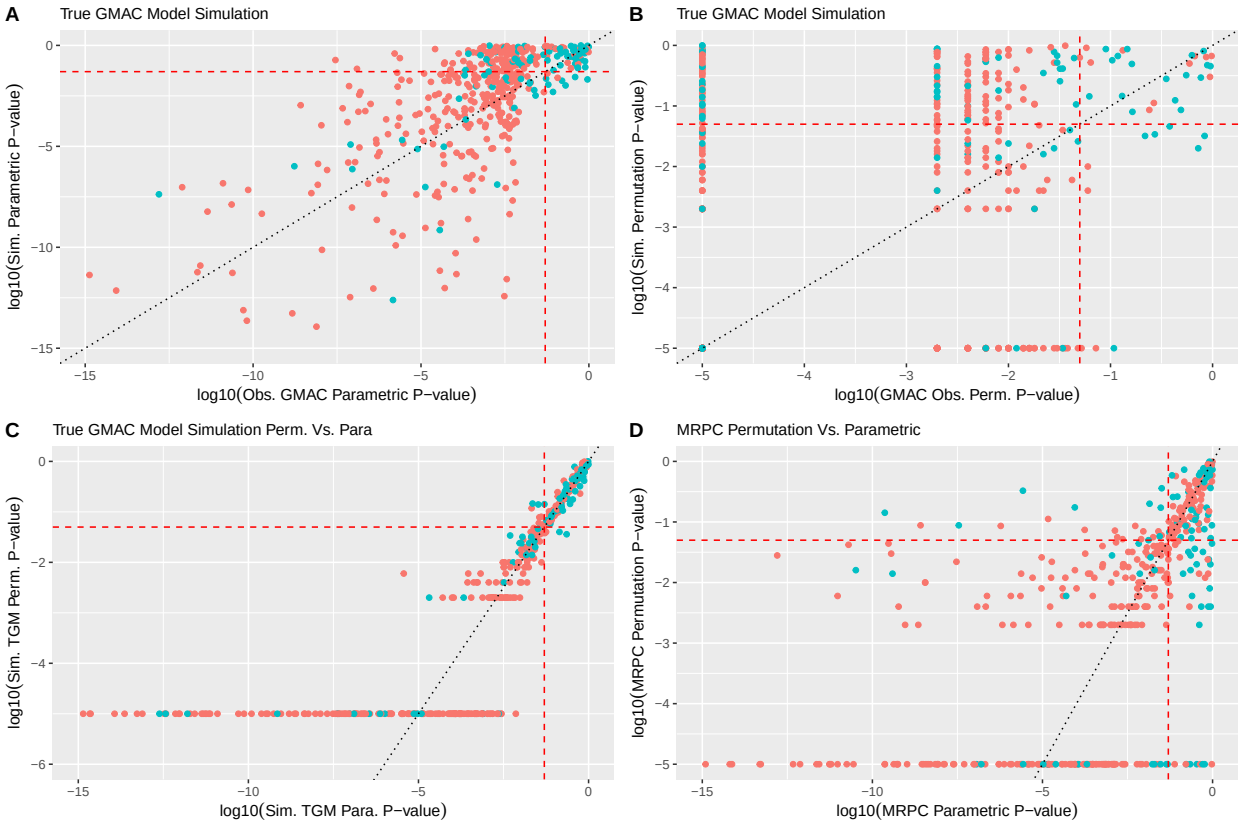
B %Var Accounted For by MRPC included PC's



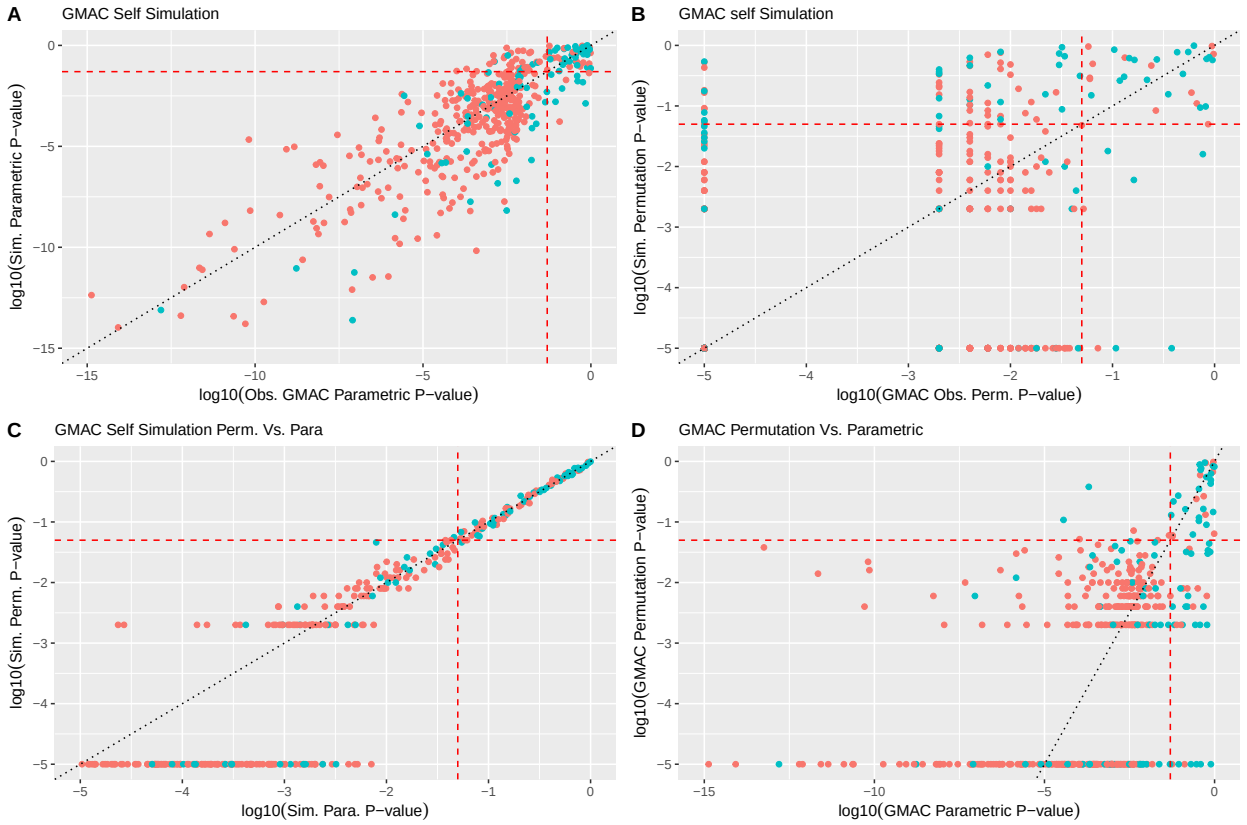
LTM and STM plots



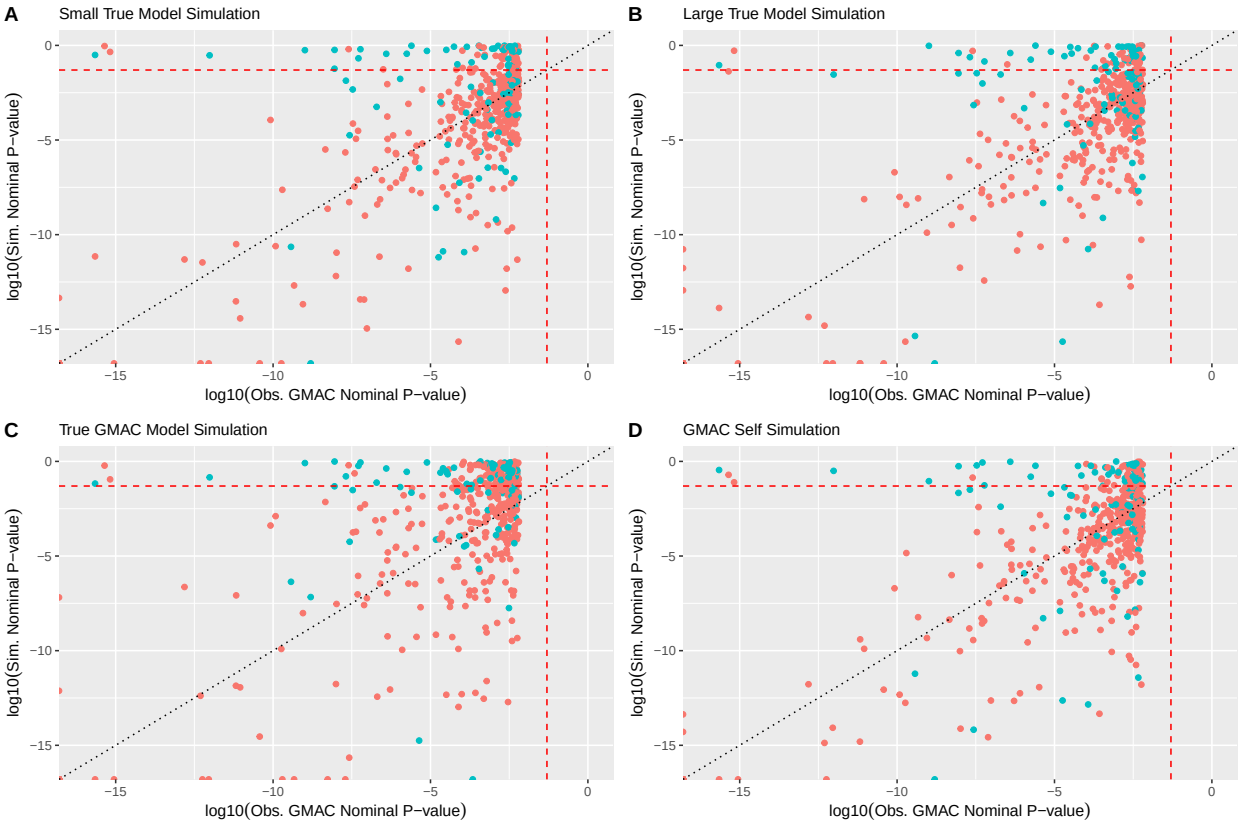
TGM and MRPC plots

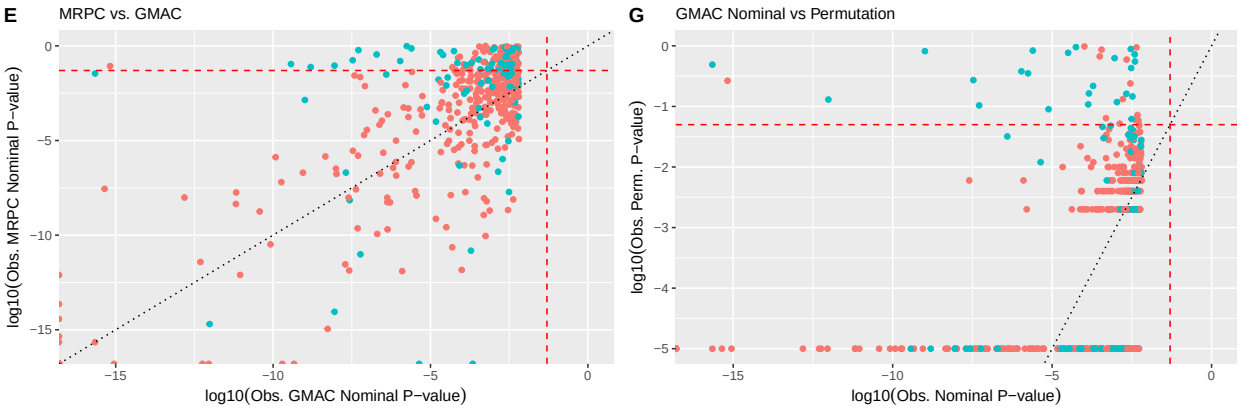


GSS and GMAC plots

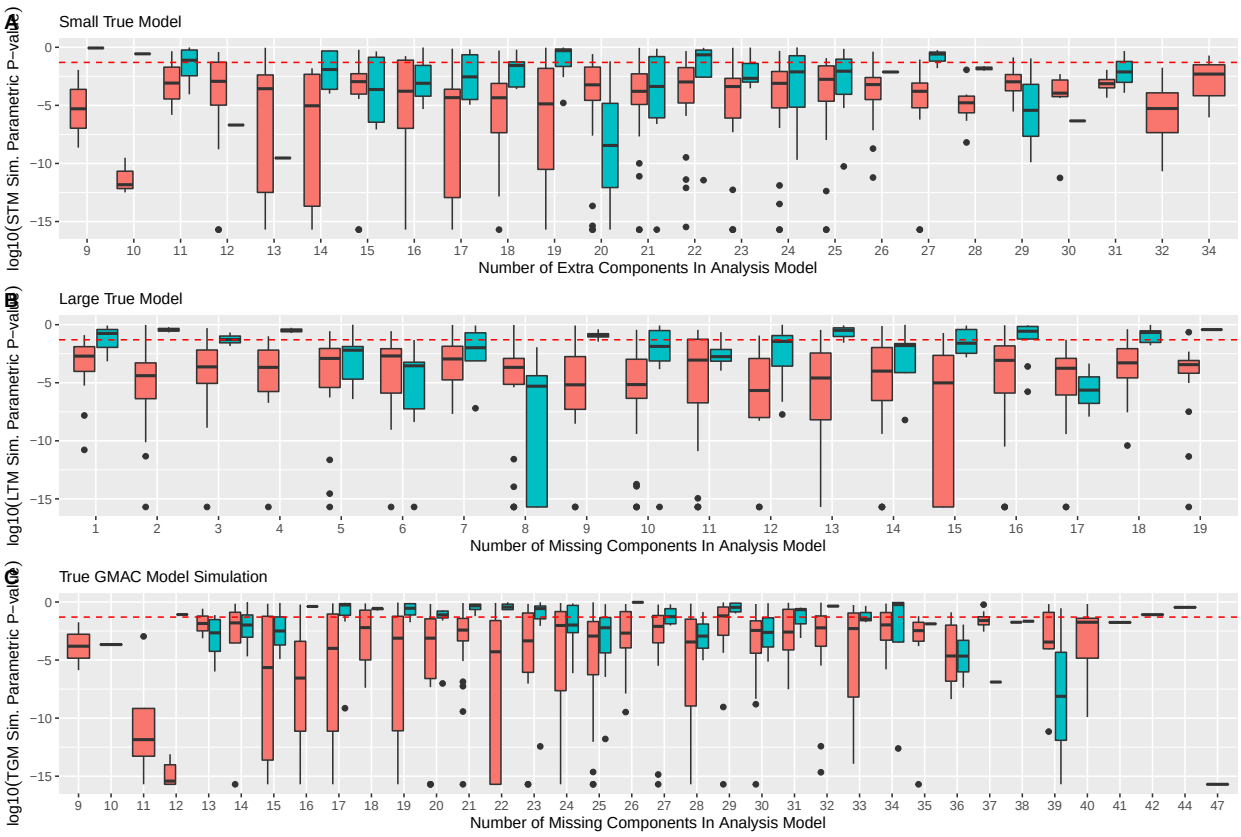


Nominal P-value plots

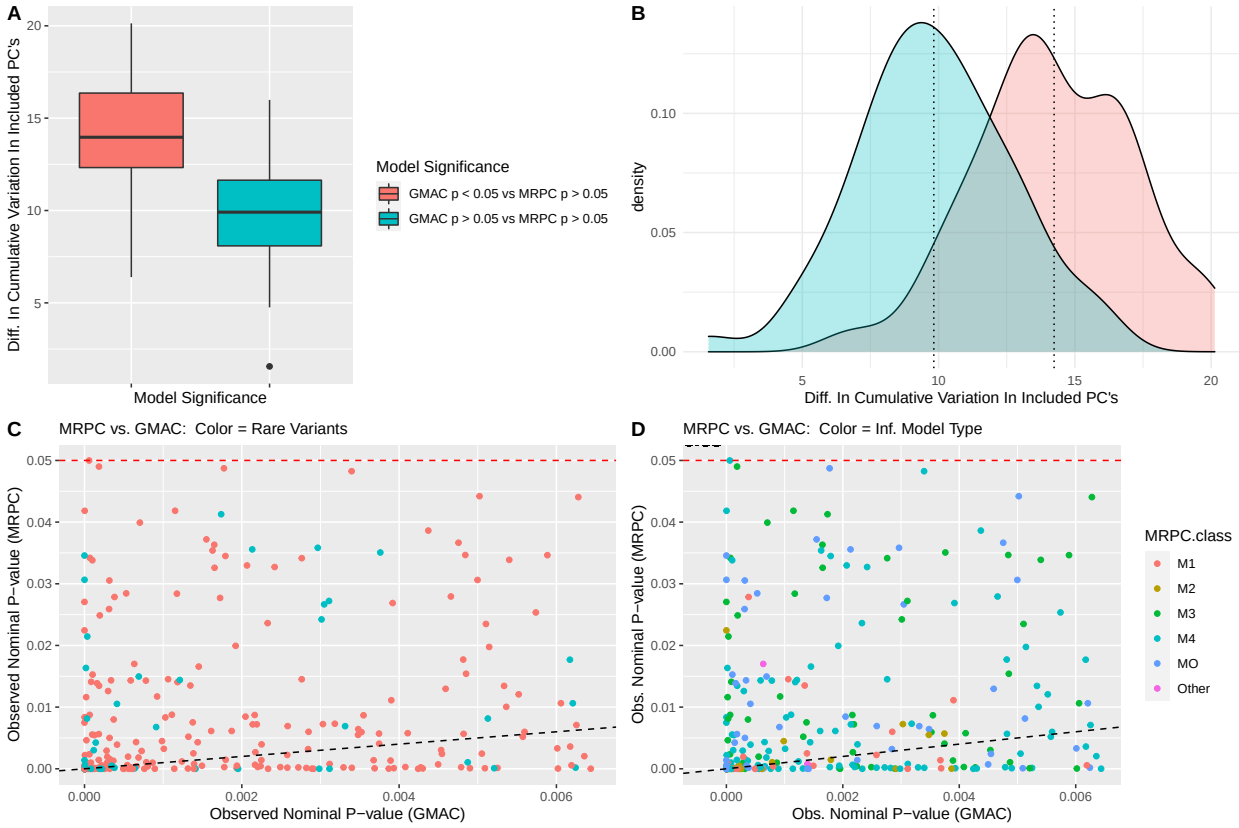



E
MRPC vs. GMAC
log10(Obs. MRPC Nominal P−value)
log10(Obs. GMAC Nominal P−value)
0
−5
−10
−15
−15
−10
−5
0
G
GMAC Nominal vs Permutation
log10(Obs. Perm. P−value)
log10(Obs. Nominal P−value)
0
−1
−2
−3
−4
−5
−15
−10
−5
0

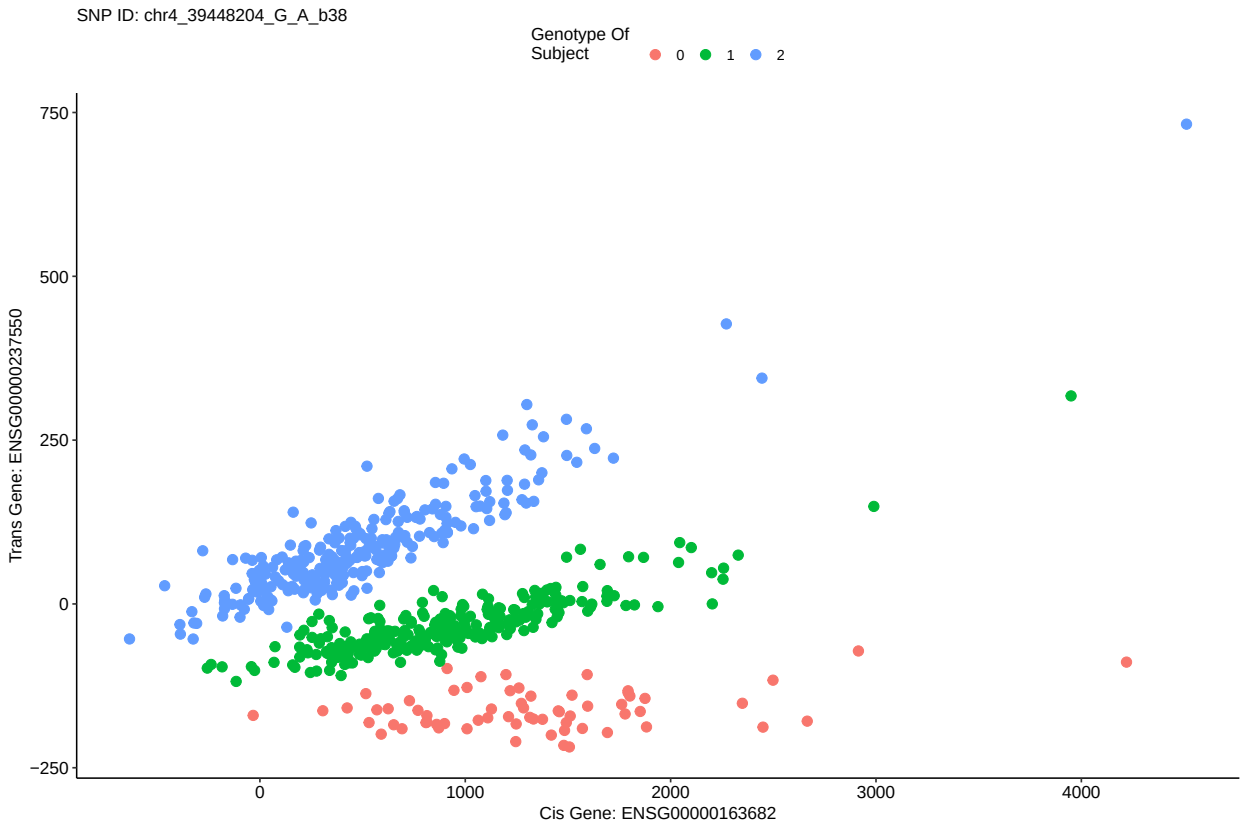
boxplots for PC inclusion differences

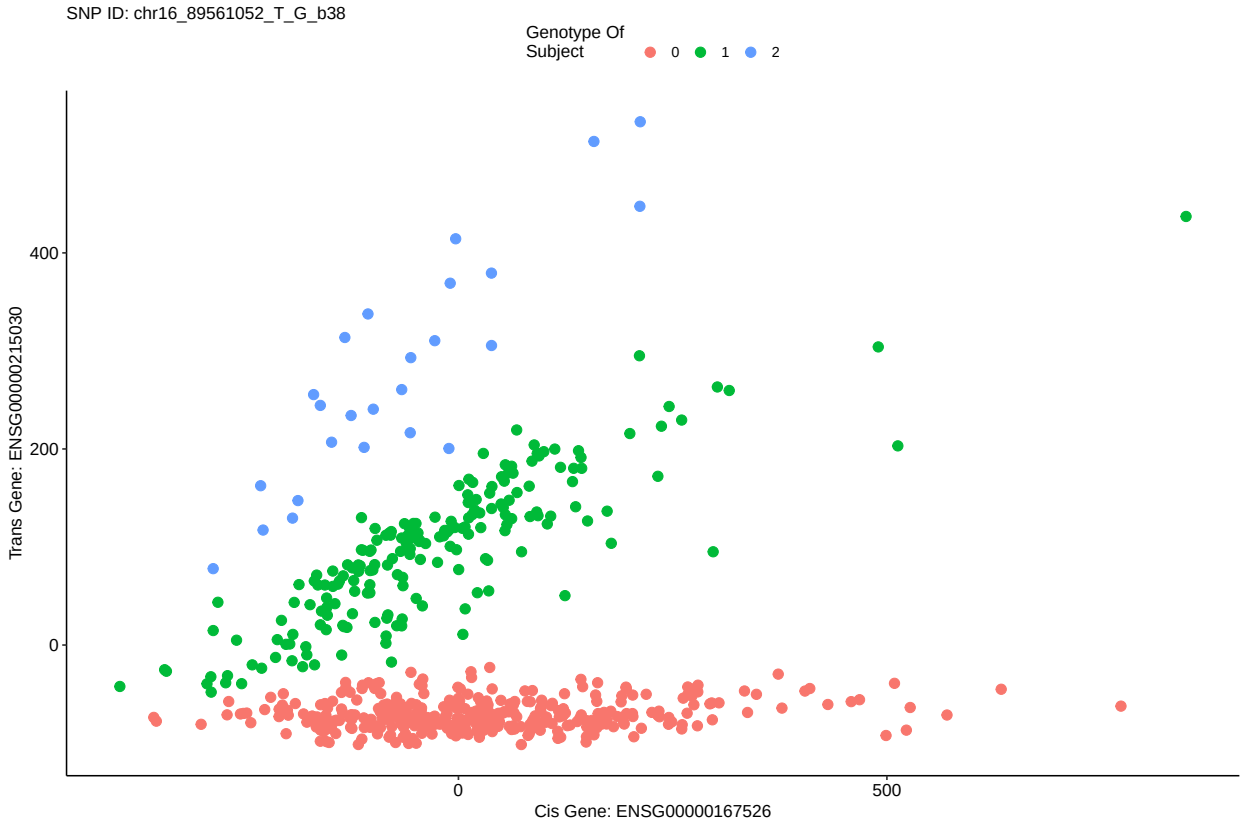


variation differences between sig and non-sig trios GMAC/MRPC



- Rare allele plots


SNP ID: chr4_39448204_G_A_b38
Genotype Of Subject
0
1
2
750
500
250
0
−250
Trans Gene: ENSG00000237550
0
1000
2000
3000
4000
Cis Gene: ENSG00000163682


SNP ID: chr16_89561052_T_G_b38
Genotype Of Subject
0
1
2
400
200
0
Trans Gene: ENSG00000215030
0
500
Cis Gene: ENSG00000167526

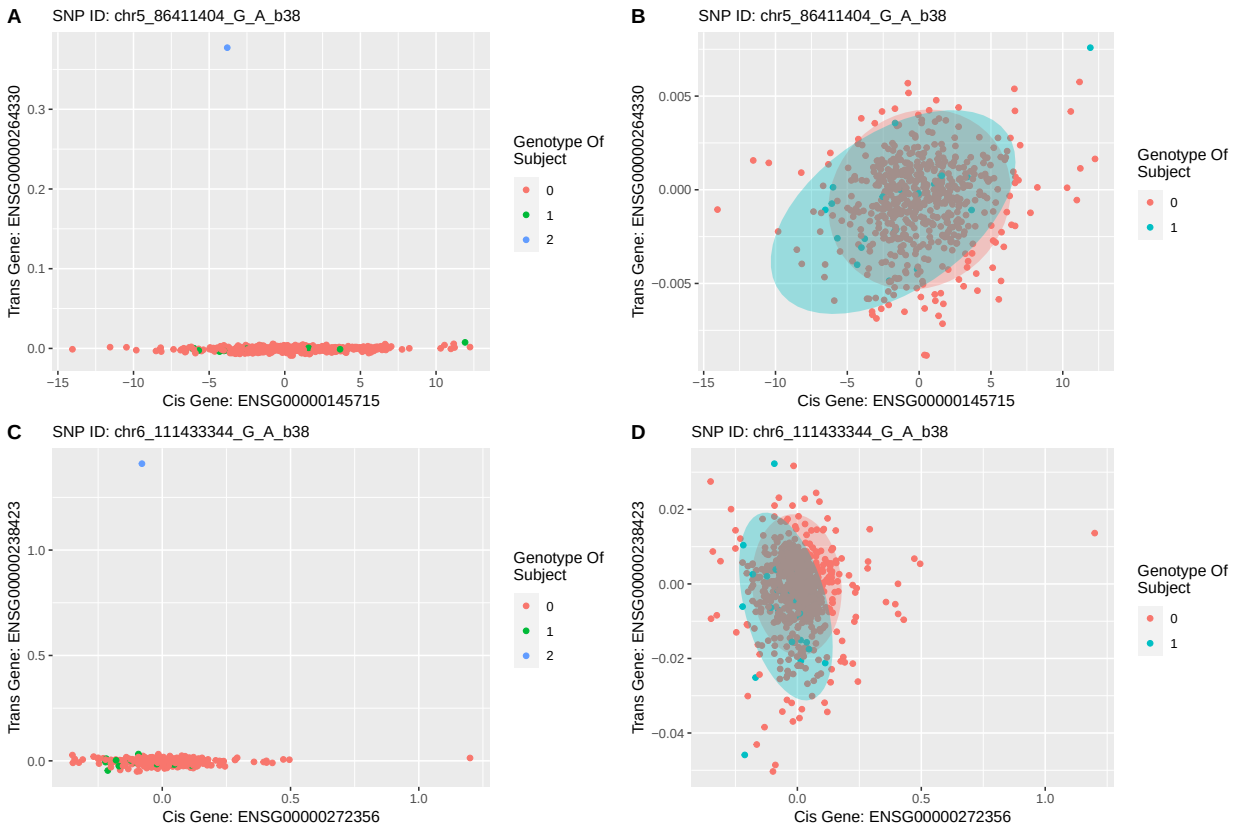


Figure 3: **A** and **C**: Example scatter plots of trios from subcutaneous adipose tissue and whole blood (respectively) with a rare allele present in the sample: Note that 0 indicates individuals homozygous for the reference allele, 1 indicates hetezygous individuals and 2 indicates individuals who 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 **C** 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 heterzygous individals.

References

Josse, Julie, François Husson, and others. 2016. "MissMDA: A Package for Handling Missing Values in Multivariate Data Analysis." *Journal of Statistical Software* 70 (1): 1–31.

Yang, Fan, Jiebiao Wang, Brandon L Pierce, Lin S Chen, François Aguet, Kristin G Ardlie, Beryl B Cummings, et al. 2017. "Identifying Cis-Mediators for Trans-eQTLs Across Many Human Tissues Using Genomic Mediation Analysis." *Genome Research* 27 (11): 1859–71.