Prediction of gene expression measurements to understand the origin of complex diseases

Akanksha Yadav

Lab Rotation - I Report

Supervisor: Prof. Johannes Söding
Co-Supervisor: Franco Simonetti

January 06, 2020 - February 28, 2020

Summary

The basic motivation behind predicting gene expression pattern is to be able to determine an individual's predisposition to disease risks among other phenotypic traits, given only the genotype. The underlying paradigm of such a study is that a difference in phenotype, is at least in part, attributable to a difference in genotype which manifests via differential gene expression among individuals.

PrediXcan [2] is one such gene-based method used to test associations between predicted gene expression and complex phenotypes which in turn helps to gain insight into corresponding molecular mechanisms of action. The expression values are estimated from a multiple linear regression model built using a reference data set containing known genotype and expression values for a set of individuals. The method leverages additive effect of strong to medium-strength signals from genetic variants (SNPs) known to influence gene expression (eQTLs) in model building and outperforms the ability of typical GWAS studies to identify causal SNPs/genes. But the method is restricted to building models from a particular set of SNPs that lie in the vicinity of the gene of interest (1MB upstream or downstream) termed as cis-eQTLs.

Trans-eQTLs are distal to the genes they affect, have low effect sizes and vary across tissues. Because of the issue of low statistical power in identifying them from millions of SNPs which could potentially be true trans-eQTLs, they have been often ignored from such studies. We have come up with a method to find sets of causal trans-eQTLs for different genes in a particular tissue and hypothesize the improvement of single-gene models by inclusion of these trans-eQTLs along with the cis-eQTLs. This report discusses modification of the PrediXcan method and subsequent analysis of model improvement upon including trans-eQTLs, along with addressing issues in the evaluation.

1 Introduction

The human genome is around 99% identical and from the remaining percentage of genetic variance at single nucleotide positions (SNPs), we expect to be able to explain phenotypic traits ranging from differences in colour of the eye to potential disease risks among individuals. Since the sequencing of the human genome has been made possible, there has been an unprecedented growth in studies trying to link genomic loci with occurence of complex human diseases. The persistent goal has been to understand the human genome, discover pathways underlying diseases and to move into an era of personalized medicine.

1.1 Fundamental concepts

A single nucleotide polymorphism, or **SNP** (pronounced "snip"), is a variation at a single nucleotide position in a DNA sequence among individuals. In our study, if more than 1% of a given population does not carry the same nucleotide at a specific position, then this variation is classified as a SNP.

1.1.1 Genome-wide association studies (GWAS)

A GWAS is an experimental setup used to test associations between genetic variants (SNPs) with disease traits of interest. The past decade has seen various GWAS studies [6] being conducted for various diseases but the extrapolation to discovering associated pathways has been an issue. The limitations exist on multiple levels in terms of the biology and statistics - (i) the genetic architecture behind a complex trait is polygenic in nature, which is not captured by single variant-single trait association tests (ii) small effect size cannot be captured due to multiple testing problems. (iii) linkage disequilibrium makes it difficult to identify causal SNPs. Also, the downstream analysis has been difficult as SNPs often lie in non-coding regions of the genome and are involved in regulation of genes that are not in close proximity and hence target genes are difficult to identify.

1.1.2 Expression quantitative trait loci (eQTLs)

eQTLs are a class of variants that have an affect on gene expression. Using these particular set of SNPs as explanatory variables in predicting disease risks can provide the missing molecular insight in previous GWAS studies. Also, their potential biological relevance in regulation of gene expression [1] increases interests to study them. Recent efforts from the GTEx consortium have produced large-scale data for genotype and expression values across tissues for a group of individuals.[3] Such projects have brought into the field a new kind of method termed as transcriptome-wide association analysis, which trains an external imputation model, predicts gene expression from the genotype of GWAS samples and performs gene-level association test between predicted expression and traits of interest. [2, 4]

1.1.3 Cis- v/s Trans-eQTLs

Cis-eQTLs commonly refer to those genetic variants that act on genes nearby (distance $<$ 1Mb), while trans-eQTLs influence genes farther away or those situated on a different chromosome. Cis-eQTLs are usually located near transcription start site (TSS) of genes they affect and hence likely to have direct effects on expression. While cis-eQTLs are known to have larger effect sizes, and hence easier to discover, the proportion of variance in gene expression explained by them is limited. On the other hand, trans-eQTLs often lie in regulatory regions of the chromosome and occur in hubs, hence in principle, they could possibly explain part of the missing variance. The issue, though, has been in identifying the causal trans-eQTLs from among the vast multitude of SNPs, for each gene. Smaller effect sizes have also made their discovery a challenging statistical problem. Hence, there is need for a robust method to be able to detect trans-eQTLs with good sensitivity.

1.2 PrediXcan model

The model building comprises two parts as summarized in Fig.1
First involves the building of a predictive model from reference transcriptome datasets as reflected in the top panel, containing genotypes of n individuals at various (M) SNPs. The model is built fitting multiple regression elastic-net models for every gene with genotypes (encoded as 0,1,2) as variables to predict the gene's expression. (T is the expression matrix and X_k is the number of reference alleles for SNP k in the middle panel). The genotypes in above mentioned models were subsetted for the SNPs within 1Mb of the gene of interest. The weights obtained for the SNPs in each single-gene models were curated in a PredictDB database for p tissues.
The second part (bottom panel) uses the weights from the database to predict gene expression for a new individual given the genotype available from typical GWAS studies. The imputed expression values are then tested for association with disease traits and the gene-trait associations allow to discover disease pathways, along with illustrating a biological role of the SNPs associated.
Specifics of the PredictDB pipeline -
(i) Only cis-eQTLs (within 1Mb upstream and downstream of the gene) were used to build models
(ii) Elastic net models were built with fixed weights of L1 and L2 regularization
(iii)λ, the complexity parameter is optimised using a 10-fold cross validation and the final model corresponds to the value of λ for which mean squared error (MSE) is minimum:

$$\min_{\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1 \right] \tag{1}$$

(iv) Original PredictDB pipeline made use of only European samples for model building
(v) Covariates correction (to remove confounding effects) was performed by regressing out principal components, PEER factors, sex, sequencing platform from the expression values and taking the residuals as the new expression.

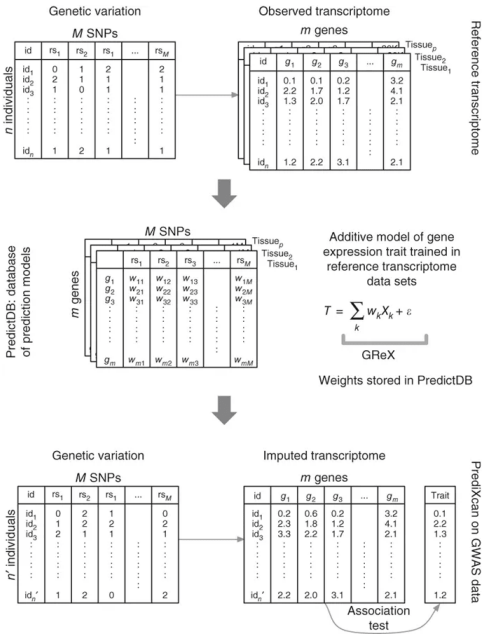

Figure 1: The workflow shows the steps used in developing the PrediXcan method - *image taken from* [2]

2 Methods

The following section gives an overview of the methodologies applied in the subsequent study and doesn't necessarily indicate the exact code or scripts written for the same. Although, a part of the code was modified from `https://github.com/hakyimlab/PredictDB_Pipeline_GTEx_v7` and we thank the authors for easy access.

As a preamble to this study, we have come up with a method to discover trans-eQTLs given a set of genotype and gene expression for different tissues. Based on this premise, we wanted to gauge model performance in terms of the proportion of variance in gene expression explained upon inclusion of our discovered trans-eQTLs. Current methods, owing to the difficulty in finding statistically significant trans-eQTLs, rely on predicting gene expression from the cis-eQTLs alone, which have larger effect sizes and are limited in number within the 1Mb window size of the gene under study.

The method introduced above has been long-standing in the field and traditionally developed taking only cis-eQTLs into account. The idea, hence, was to modify the model building pipeline to introduce our trans-eQTLs and compare it with existing models that only incorporate cis-eQTLs.

2.1 Datasets

All the data used for the study has been obtained from the current release of the GTEx project (V8, dbGAP accession code phs000424.v8.p2) This includes the genotype and expression matrices for 838 (total) samples across 54 tissues, although the number of available samples with genotype and tissue-specific gene-expression varies across tissues.

The genotype was obtained from dbGAP in the form of a vcf file with all 838 samples. No filtering was done on the basis of race. No imputation was done as the data was from

whole genome sequencing. Only SNPs with a minor allele frequency >= 0.01, biallellic, and those that are unambiguously stranded i.e, the variant is not the base's complement were selected. Genotypes were encoded as 0,1 or 2 denoting the count of the effect allele. The expression matrix was procured from the haplotype expression matrices produced with phASER. In this study, phASER expressions have been pre-processed in several ways. The idea behind it was that the trans-eQTLs were discovered using KNN (k-nearest neighbours) -correction on the phASER expression and we wanted to see how well the results are replicable for other expressions that use linear models for covariates correction.
The five set of expressions that have been later on referred to using short-hand notations are as follows:
tpms.qcfilter- This is the phASER expression where read counts were converted to TPMs(Transcripts Per Kilobase Million) with usual quality control filter as in GTEx
tpms.cclm- This is the phASER expression with covariates corrected using a linear model. Here covariates refer to the top 5 genetic principal components, sequencing platform, sex, age, age^2, age^3, PMI(post-mortem interval in minutes)
tmm.cclm- This is the above expression further processed using a method of normalization called Trimmed Mean of M-values (TMM)
tpms.cclm.peer- This is the tpms.cclm expression further corrected for PEER (probabilistic estimation of expression residuals) factors which is supposed to correct for hidden covariates
tpms.knncor- This is the original phASER expression after the quality control, corrected using K=30 nearest neighbours calculated using a distance matrix obtained from the expression matrix itself.
Gene annotation was derived from gencode v26, using GTEx's collapsed gene model. SNP annotation was obtained from the look-up table containing rsIDs from dbSNP 151.

2.2 Workflow

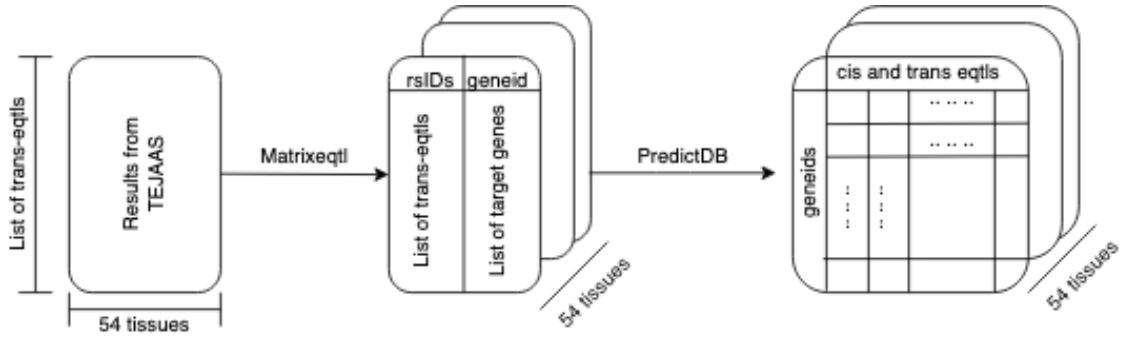

Figure 2: Schematic of the workflow in the current study

This section consists of a brief outline to the work done during the project. A considerable amount of time was dedicated to setting up the pipeline itself, often modifying and fixing any bugs, the details of which are omitted.
As indicated in Fig. 2, the first box corresponds to the set of trans-eQTLs obtained

for each tissue from **Tejaas**, an algorithm developed by the lab for sensitive detection of trans-eQTLs. The first step was to find the target genes corresponding to the trans-eQTLs for each tissue. This was achieved by using a software named **Matrix eQTL** [5], which tests for association between each SNP and each transcript by an additive linear model. Using this data, the PredictDB pipeline was run twice - once, to build the model from cis-eQTLs alone and next, to build the model for the target genes obtained in the prior step using using both, the previously included cis-eQTLs and the newly discovered trans-eQTLs.
Due to the limited period of time and numerous factors under consideration, the above analysis was done for only three tissues: Adipose Subcutaneous (AS), Muscle Skeletal (MS), and Artery Aorta (AA). The steps of analysis are as follows:
(i) Running Matrix eQTL using the trans-eQTLs and 5 differently pre-processed expressions (tpms.qcfilter, tpms.cclm, tmm.cclm, tpms.cclm.peer, tpms.knncor) and obtaining corresponding target genes for each type of expression for the three tissues
(ii) Running PredictDB for the 5 expressions for each tissue, that by default, creates models from cis-eQTLs only (by subsetting genotype data based on gene start position and position of SNPs within 1Mb window) - "cis-only" run
(iii) Running PredictDB for the same 5 expressions for each tissue, but including the trans-eQTLs for the target genes obtained in step (i) for the corresponding expression and tissue. This creates models from both cis and trans-eQTLs for the particular genes - "cis+trans" run

3 Results

This section showcases preliminary analysis results from the methods outlined above.
One of the main tasks after the above execution of pipeline, was to be able to summarize and visualize the difference in the two models in hand, *cis-only* and *cis+trans*. Ideally one would like to test the models on an external dataset ("unseen" by the model), but to get the first looks it suffices to see some summary statistics from the model building itself. For this purpose, we were particularly interested in a output parameter derived from a 5-fold nested cross validation.
The nested cross validation works as follows: (i) split data into 5-folds (ii) for each fold, hold out one fold and train a model on the rest using a 10-fold cross validation to optimize the λ parameter. (iii) test on the hold-out fold and retain summary statistics for model performance.

The summary statistics that are referred to in the following text are as follows:
ρ^2_{avg}- average correlation between predicted and observed on the hold out folds when doing nested cross-validation, squared
R^2- average coefficient of determination when predicting values of the hold out fold during nested cross validation. Coefficient of determination is defined as
$1 - \sum(y_{\text{observed}} - y_{\text{predicted}})^2 / \sum(y_{\text{observed}} - \overline{y}_{\text{observed}})^2$

From the *cis+trans* model three cases arise -
(i) genes for which trans-eQTLs were discovered and also included in the model (i.e having non-zero weights) - "*Trans-positive*"
(ii) genes for which trans-eQTLs were discovered but none of them got included in the model
(iii) genes for which trans-eQTLs were not discovered
(ii) and (iii) combined denoted as "*Trans-negative*"

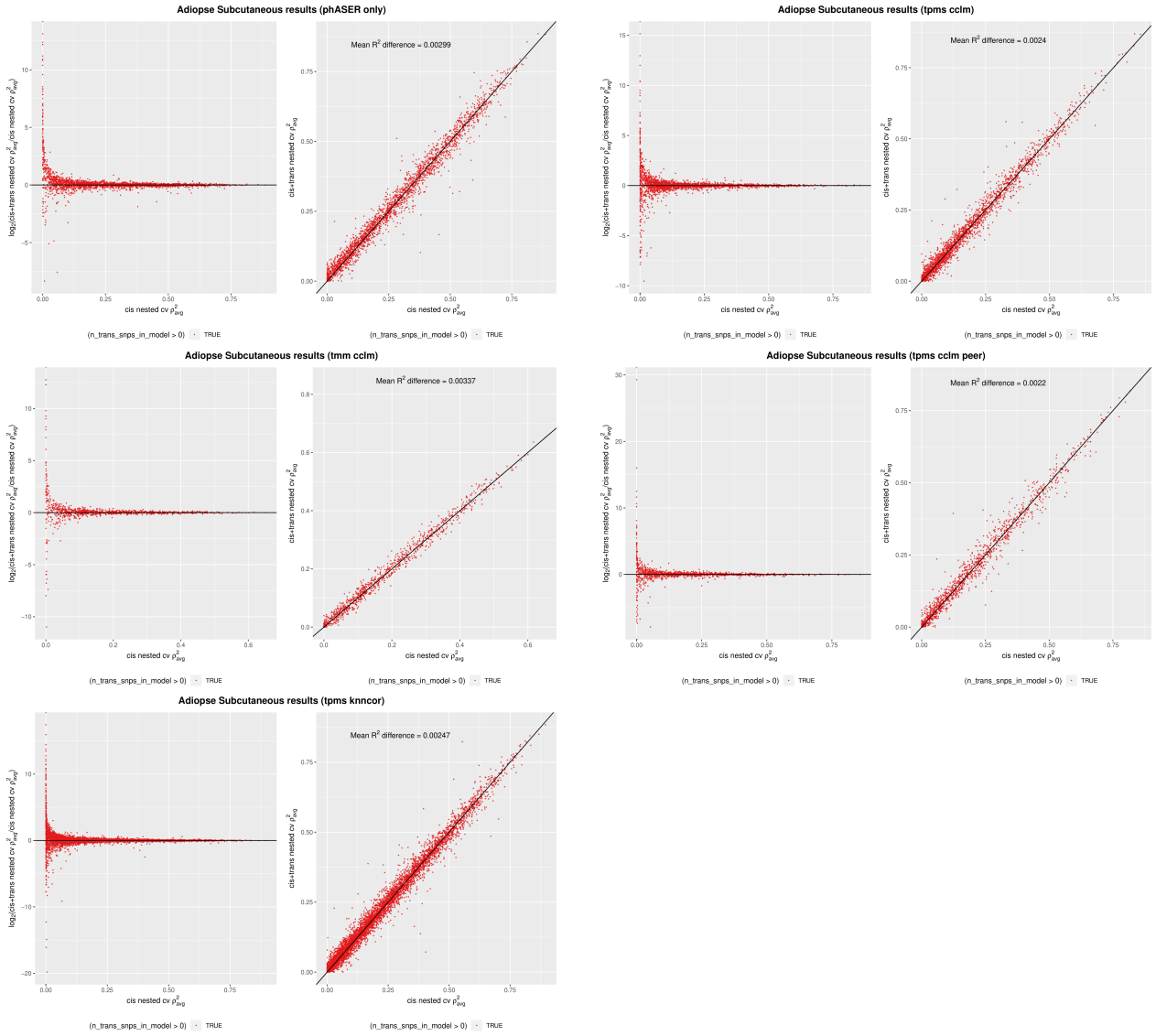

Figure 3: Performance improvement summary for Adipose Subcutaneous tissue (without any filter)

Fig.3 summarizes the performance improvement for Adipose Subcutaneous tissue for the 5 expressions indicating only "*Trans-positive*" genes, as labelled. The left panel in each expression shows log_2 ratio of ρ^2_{avg} between *cis+trans* and *cis* model for each gene (red points) plotted against *cis* ρ^2_{avg}. The right panel gives an alternative view for the same

with *cis+trans* plotted against *cis* ρ^2_{avg}. The mean difference of ρ^2_{avg} across the "*Trans-positive*" genes seems small. Fig.4 summarizes the mean difference of ρ^2_{avg} for the three tissues.

Since, the absolute difference in ρ^2_{avg}, as seen above, is quite small, no solid conclusion

	Type.of.expression	AS	MS	AA
1	tpms_qcfilter	0.00299	0.00464	0.00821
2	tpms_cclm	0.00240	0.00471	0.00510
3	tmm_cclm	0.00337	0.00392	0.00638
4	tpms_cclm_peer	0.00220	0.00490	0.00742
5	tpms_knncor	0.00247	NA	NA

Figure 4: Mean ρ^2_{avg} difference for Adipose Subcutaneous(AS), Muscle Skeletal(MS), Artery Aorta(AA) tissues

could be drawn based on the above observation. Fig. 5 compares the ρ^2_{avg} distribution for *Trans-positive* and *Trans-negative* model definitions. This indicates most genes are predicted with ρ^2_{avg} in the range 0.05-0.2 and for the *Trans-positive* models, the ρ^2_{avg} distributions are not exactly identical between *cis* and *cis+trans*. This appears most apparent for "tpms.cclm.peer" expression. If the trans-eQTLs were completely random, then the distribution should have been identical between *cis* and *cis+trans* as in the case of *Trans-negative* models.

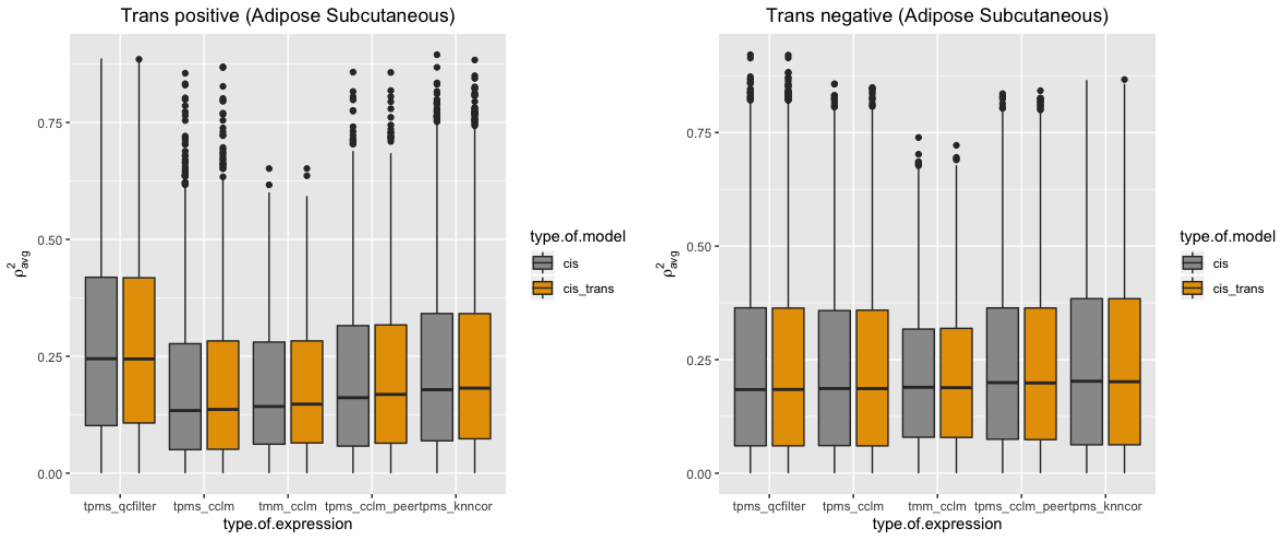

Figure 5: ρ^2_{avg} distribution for Adipose Subcutaneous tissue (without any filter)

To be able to figure out if the differences were completely random, we decided to look at the noise across runs and set a reasonable threshold for different parameters. To come up with a good filtering criteria, a few combinations of noise filtering ability versus loss of improvement signal was observed empirically. For this purpose, we used the *Trans-negative* models for which an improvement could occur due to different cis-eQTLs being

incorporated in two runs of the same data. Fig. 6 indicates log_2 ratio of improvement which should be as close to zero as possible in the case of *Trans-negative* model. For genes with $\rho^2_{cis,avg} < 0.01$, the improvement ratio blows up for very small changes in ρ^2_{avg} in either direction and shows false improvement when it is not the case. Hence, $\rho^2_{cis,avg} > 0.01$ should be used as a cut-off for evaluating model performance. Based on

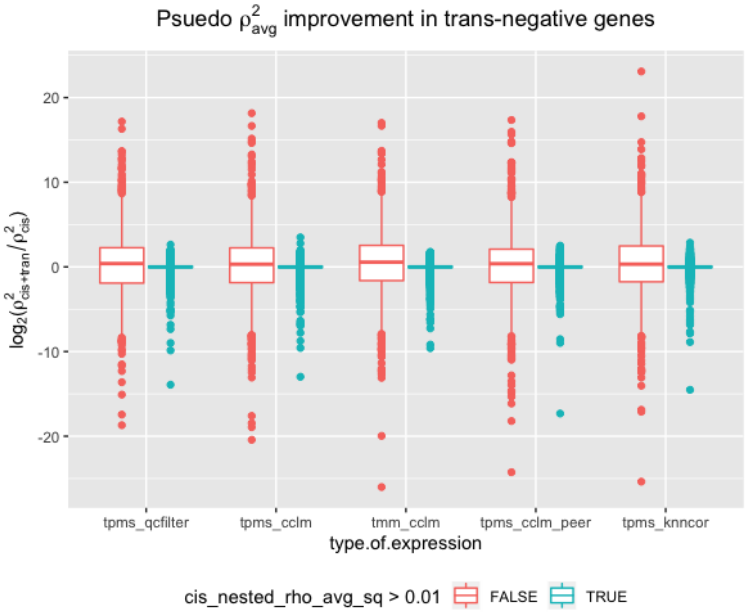

Figure 6: $log_2(\rho^2_{\text{cis+trans}}/\rho^2_{\text{cis}})$ distribution for Adipose Subcutaneous tissue - *Trans-negative* with model separation on the basis of $\rho^2_{cis,avg} > 0.01$

the filtering criteria for both *Trans-positive* and *Trans-negative* models, a contour plot in Fig.7 indicates its effects. ρ^2_{avg} difference is plotted against $\rho^2_{\text{cis,avg}}$ and hence density above the x axis, would be indicative of improvement.

4 Discussion

Based on the results obtained so far, it has been difficult to clearly discern and quantify any model improvement upon including trans-eQTLs. That being said, we also came across potential shortcomings in the method and the pipeline used for this study.

(i)*A good filtering criteria* It was observed that across different runs with the same data, the models built vary in terms of ρ^2_{avg}, and hence it would be difficult to quantify model improvement if it were indeed smaller than the variation. This variability in the models could be in part due to the low number of data points for some races (non-Europeans) which could lead to some bias in the test/training data sets.
This could be addressed at various levels, using strict filtering criteria with a cut-off for absolute improvement, or based on the minimum number of trans-eQTLs incorporated in the model. The same pipeline could also be run multiple times to set a significance threshold or data could be homogenized for including only Europeans samples and tested

again.

(ii) *True or false-positive trans-eQTLs* A good way to be able to see if the trans-eQTLs are actually able to explain better the variance, would be to take a set of random trans-eQTLs replacing the actual set and then compare the model performance statistics again. This could give a better idea in terms of the sensitivity.

(iii) *The right target genes* Based on the results from Matrix eQTL there is quite some diversity among the target genes discovered for the same set of trans-eQTLs, from different expression matrices. Although it seems counter-intuitive, we are basically varying the dependent variable keeping the independent variable same, hence there is variability in SNP-gene pairs obtained. A simple linear regression model could be modified in this case and a reliable cut-off should be chosen for the SNP-gene associations. Since a set of trans-eQTLs often exert a joint effect on a set of genes, the best way to find the target genes would be to take into account both group-level and individual associations between SNPs and genes possibly using a LASSO model.

(iv) *Is the model building accurate enough* Although many other factors could be in play, one consideration is that the model was optimized for cis-eQTLs from the European samples alone. Lately, many reports indicate that the method performs poorly across races. Also, the expression matrices used in our study are often not used by others and hence there is some difference to begin with. Also, we do not filter for any SNPs based on ethnicity. A good way to ensure everything is in place, is to replicate published results from the same datasets as used by others and following the same protocols.

(v) *Is there a best expression matrix* We have tested, in a non-exhaustive way, 5 different kinds of expression matrices with different covariates correction. Although the expression without any correction (tpms.qcfilter) seems to give good results, it most likely due to overtraining and learning confounding effects which would perform poorly on external dataset. The best among other seems to be tpms.cclm.peer for which a good number of SNP-genes pairs are obtained which also show up in the model.

This study doesn't lead to absolute results but opens up many questions that can be looked into and improved upon in the future. An exciting venture would be to predict these models on external datasets and look for disease associations.

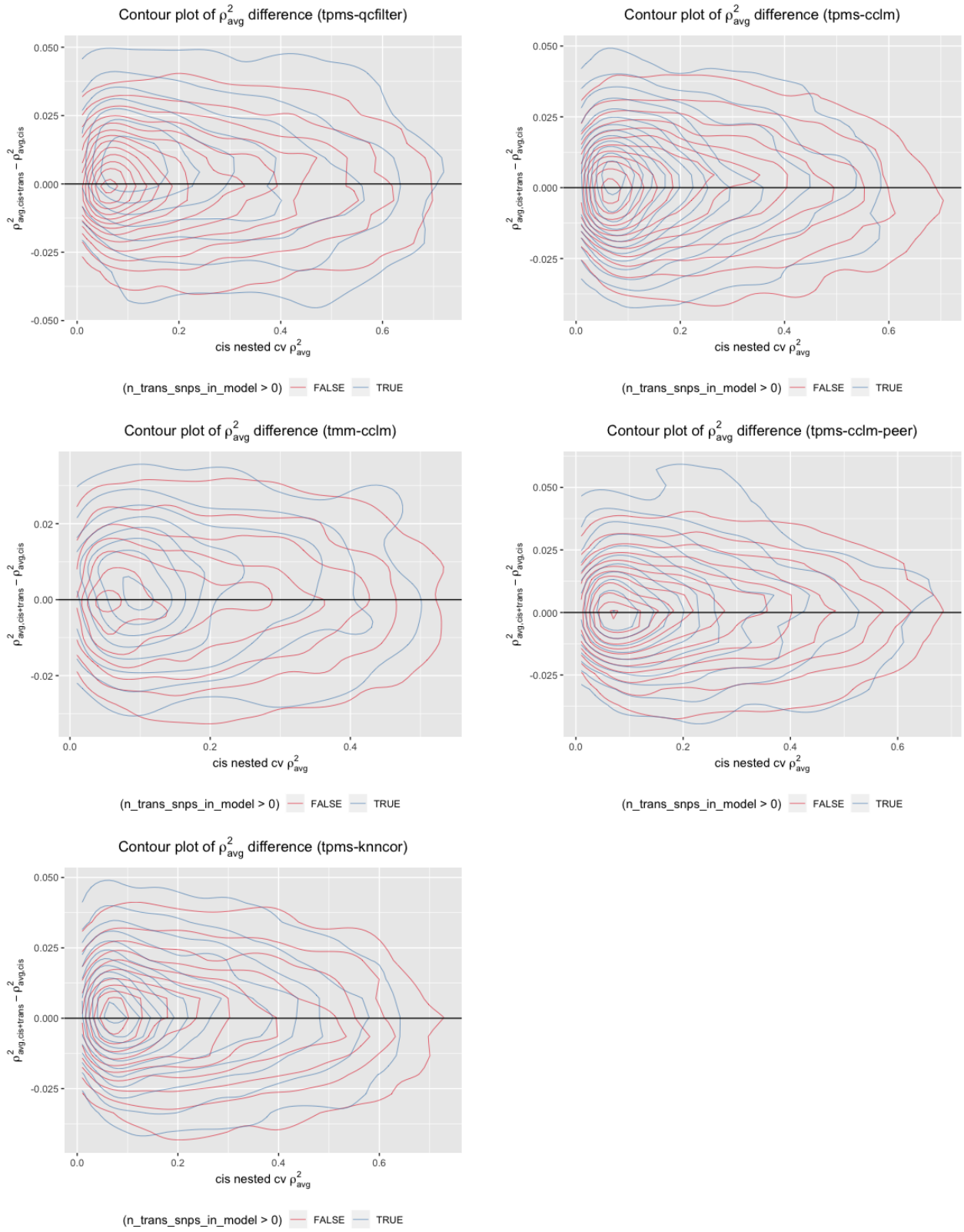

Figure 7: Performance improvement summary (contour plots) for Adipose Subcutaneous tissue (with filtering criteria $\rho^2_{cis,avg} > 0.01$)

References

[1] Frank W. Albert and Leonid Kruglyak. "The role of regulatory variation in complex traits and disease". In: *Nature Reviews Genetics* 16.4 (2015), pp. 197–212. ISSN: 14710064. DOI: 10.1038/nrg3891.

[2] Eric R. Gamazon et al. "A gene-based association method for mapping traits using reference transcriptome data". In: *Nature Genetics* 47.9 (2015), pp. 1091–1098. ISSN: 15461718. DOI: 10.1038/ng.3367.

[3] The Genotype and Tissue Expression. "The GTEx Consortium atlas of genetic regulatory effects across human tissues The Genotype Tissue Expression Consortium". In: (2019). DOI: 10.1101/787903. URL: http://dx.doi.org/10.1101/787903.

[4] Alexander Gusev et al. "Integrative approaches for large-scale transcriptome-wide association studies". In: *Nature Genetics* 48.3 (2016), pp. 245–252. ISSN: 15461718. DOI: 10.1038/ng.3506. URL: http://dx.doi.org/10.1038/ng.3506.

[5] Andrey A. Shabalin. "Matrix eQTL: Ultra fast eQTL analysis via large matrix operations". In: *Bioinformatics* 28.10 (2012), pp. 1353–1358. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts163. arXiv: 1105.5764.

[6] Peter M. Visscher et al. "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *American Journal of Human Genetics* 101.1 (2017), pp. 5–22. ISSN: 15376605. DOI: 10.1016/j.ajhg.2017.06.005. URL: http://dx.doi.org/10.1016/j.ajhg.2017.06.005.