Using information across tissues and genes to predict gene expression in Transcriptome-wide Association Studies (TWAS)

Fabio Morgante[1], Gao Wang[2], Matthew Stephens[2,3] and Yang I Li[1,2]

[1]Section of Genetic Medicine - Department of Medicine, [2]Department of Human Genetics, [3]Department of Statistics
The University of Chicago, Chicago, IL

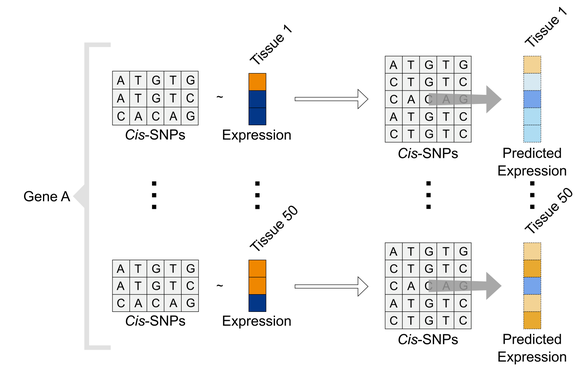
Prediction methods in TWAS

Single-tissue

- TWAS imputes gene expression based on genotypes in samples of genotyped and phenotyped individuals by leveraging genotyped expression reference panels (*e.g.*, GTEx). Predicted gene expression is then tested for association with traits of interest.
- Methods initially developed for TWAS (*e.g.*, PrediXcan, FUSION) perform gene expression imputation in one tissue at a time.
- However, these methods do not exploit patterns of expression Quantitative Trait Loci (eQTL) sharing among tissues.

Multi-tissue

- Multivariate methods that perform gene expression imputation in all tissues simultaneously (*e.g.*, fqtl, UTMOST) have been shown to provide greater accuracy than single-tissue methods.
- However, these methods analyze every gene separately: they do not exploit the information that multiple genes can provide.



Adapted from Ref. 4. Image credit: Maria E. Adonay

Our proposed multi-tissue prediction method (M&M)

- We have developed a multivariate Bayesian variable selection method for fine-mapping. Here, we evaluate its prediction performance.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$
$$\mathbf{B} = \sum_{l=1}^{L} \mathbf{B}_l$$
$$\mathbf{B}_l = \gamma_l \mathbf{b}_l^T$$
$$\gamma \sim Mult(1, \pi)$$
$$\mathbf{b} \sim \sum_{k=1}^{K} \omega_k N_R(\mathbf{0}, \mathbf{U}_k)$$
$$\mathbf{E} \sim MN_{N \times R}(\mathbf{0}, \mathbf{I}, \mathbf{V})$$

- Main idea: Build a mixture with several covariance matrices, $\mathbf{U}_k$, that captures patterns of sharing and specificity among tissues. Then, let the data inform us which components are important by estimating their weights, ω_k.
- Covariance matrices can be either data-driven (*i.e.*, estimated from the data considering information from multiple genes) or canonical (*e.g.*, identity).

For more details about the method, visit poster 2940W.

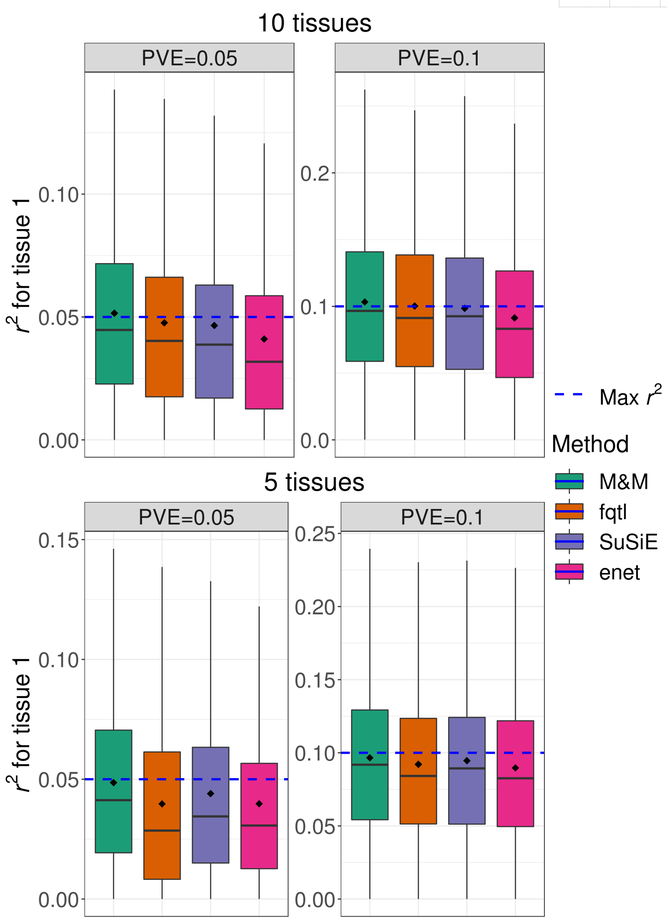
Comparison of methods via simulation

- We simulated gene expression from actual GTEx genotypes:
  - 500 randomly sampled genes with ~600 individuals.
  - 1000 randomly sampled cis-SNPs (MAF>0.01).
  - Either 5 or 10 tissues.
  - ~50% of genes = 1 causal variant (CV), ~30% of genes = 2 CVs and ~20% of genes = 3 CVs.
  - Effects were sampled from a multivariate normal distribution with different covariance structure across tissues:
    1. Independent effects.
    2. Shared effects.
    3. Effects in only one tissue.
  - Per tissue proportion of variance explained (PVE) by CV(s) = 0.05 or 0.1.
  - Independent residuals among tissues, for simplicity.
- We compared our method to other methods:
  - Factored QTL (fqtl): multivariate regression method developed for TWAS[1].
  - Sum of Single Effects (SuSiE): univariate regression method developed for fine-mapping[2].
  - Elastic Net (enet): univariate regression method used in PrediXcan[3] and FUSION[4].
- Prediction setting details:
  - Model trained on 80% of the data, prediction performance evaluated in the remaining 20%.
  - Accuracy evaluated as r^2(true, predicted).

Results of the simulations

Independent effects among all tissues

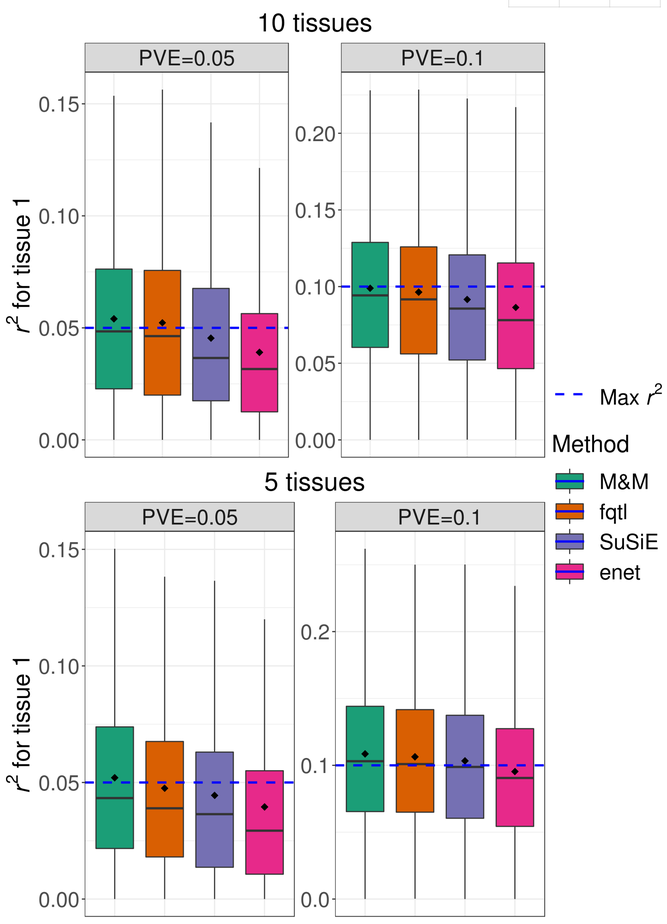
1	0	0
0	1	0
0	0	1



All the methods performed similarly, although M&M performed slightly better with lower PVE

Shared effects among all tissues

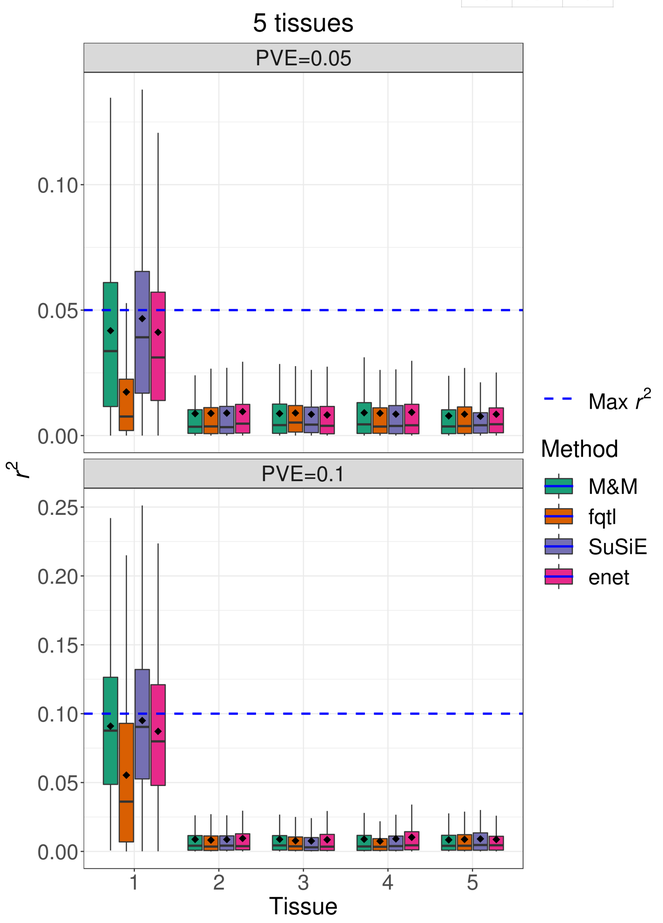
1	1	1
1	1	1
1	1	1



Multi-tissue methods performed better than single-tissue methods, especially with lower PVE

Effects in only one tissue

1	0	0
0	0	0
0	0	0



M&M retained performance in the presence of noise from tissues with no genetic effects

Main conclusions

- Multi-tissue methods achieve higher accuracy than single-tissue methods, especially with a larger number of tissues with genetic effects and lower PVE.
- Our method performed well across all the simulated scenarios, even when the other multi-tissue method did not.
- Our method gives interpretable results due to concurrent fine-mapping (visit poster 2940W for details).

References

1. Gamazon *et al.* (2015). A gene-based association method for mapping traits using reference transcriptome data. Nature genetics, 47(9), 1091.
2. Gusev *et al.* (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics, 48(3), 245.
3. Park *et al.* (2017). Multi-tissue polygenic models for transcriptome-wide association studies. bioRxiv, 107623.
4. Wang *et al.* (2018). A simple new approach to variable selection in regression, with application to genetic fine-mapping. bioRxiv, 501114.