Genetic architecture of gene expression regulation via orthogonal tissue decomposition

Heather E. Wheeler[1,2*], GTEx Consortium, Nicholas Knoblauch[3], Nancy J. Cox[4], Dan L. Nicolae[2], Hae Kyung Im[2*]

[1]Departments of Biology and Computer Science, Loyola University Chicago; [2]Department of Medicine, University of Chicago, Chicago, IL;
[3]Committee on Genetics, Genomics, and Systems Biology, University of Chicago; [4]Division of Genetic Medicine, Vanderbilt University

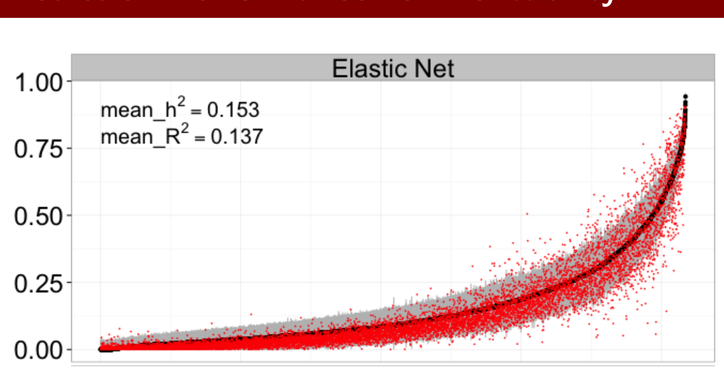
Summary

For many complex traits, gene regulation is likely to play a crucial mechanistic role given the consistent enrichment of expression quantitative trait loci (eQTLs) among trait-associated variants. In order to fully harness gene regulation mechanisms in future studies of complex traits, we sought to better understand the underlying genetic architecture of the gene regulation traits themselves. We show that local heritability of gene expression (variance in gene expression due to genetic variation within 1Mb of a gene) can be accurately estimated across tissues, but distal heritability cannot be reliably estimated at sample sizes less than 1000. Using both the elastic net and Bayesian sparse linear mixed modeling, we show that for local gene regulation, the genetic architecture is mostly sparse rather than polygenic. Using genome and transcriptome data from the diverse set of tissues available in The GTEx Project, we developed a model called Orthogonal Tissue Decomposition (OTD), which partitions gene expression into cross-tissue and tissue-specific components. Estimates of the local heritability of OTD-generated cross-tissue gene expression have larger magnitude and smaller standard errors compared to single tissue estimates due to the borrowing of information across all samples. We show that genes with high cross-tissue heritability are more likely to have cross-tissue eQTLs, confirming that OTD is capturing the cross-tissue component of gene expression. We also found evidence that genes with large tissue-specific heritability are enriched in common complex disease genes discovered via GWAS. Cross-validated predictors of cross-tissue and tissue-specific expression built here have been added to the PrediXcan (Gamazon, Wheeler, Shah et al. Nature Genetics 2015) database for future investigations of complex traits. In conclusion, local heritability estimates can be reliable obtained across tissues but not the distal component. For local regulation, the architecture is sparse rather than polygenic. Finally, an interesting enrichment of high tissue specific heritability of disease-associated genes merits further investigation.

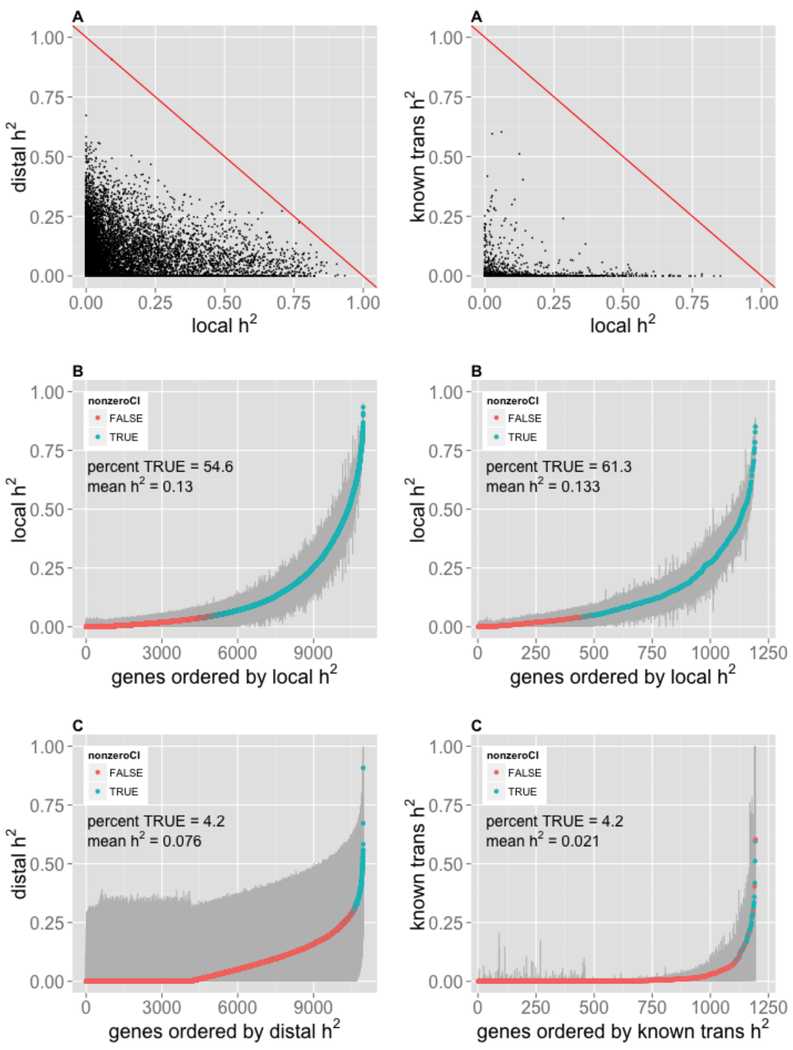
Linear Model for Gene Expression Trait

$$Y_g = \sum_{k=\in local} w_{k,g} X_k + \sum_{k=\in distal} w_{k,g} X_k + \epsilon$$

Prediction Performance vs. Heritability

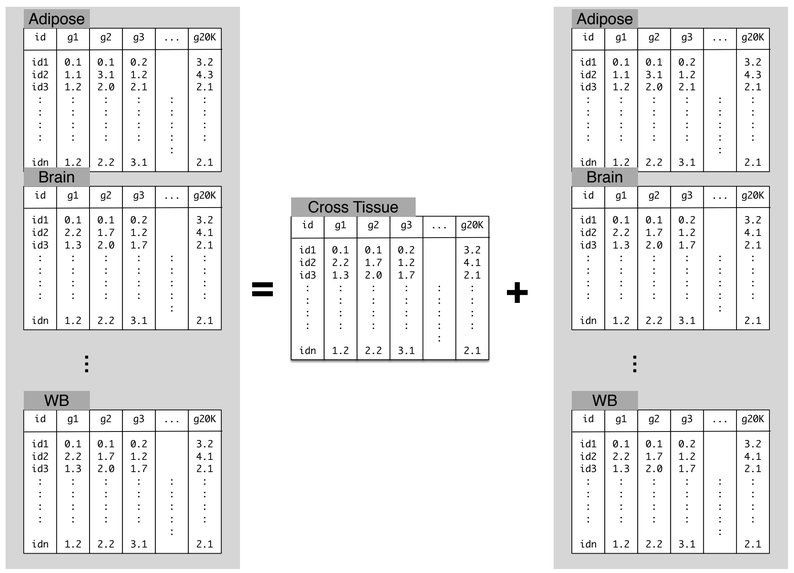


Local and Distal Joint Heritability (h^2)

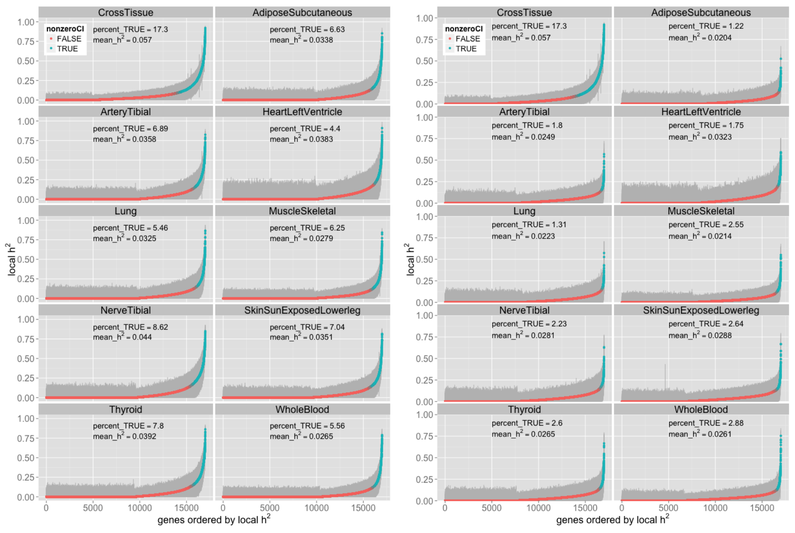


DGN whole blood expression joint (h^2). Left: Local (SNPs within 1 Mb of each gene) and distal (SNPs that are eQTLs in the Framingham Heart Study on other chromosomes [FDR < 0.05]) h^2 for gene expression were jointly estimated. (**A**) Distal h^2 compared to local h^2 per gene. (**B**) Local and (**C**) distal gene expression h^2 estimates ordered by increasing h^2. The 95% confidence interval (CI) of each h^2 estimate is in gray and genes with a lower bound greater than zero are in blue. Right: Joint h^2 estimated with local (SNPs within 1 Mb of each gene) and known trans-eQTLs (SNPs that are trans-eQTLs in the Framingham Heart Study for each gene [FDR < 0.05]).

Orthogonal Tissue Decomposition (OTD)

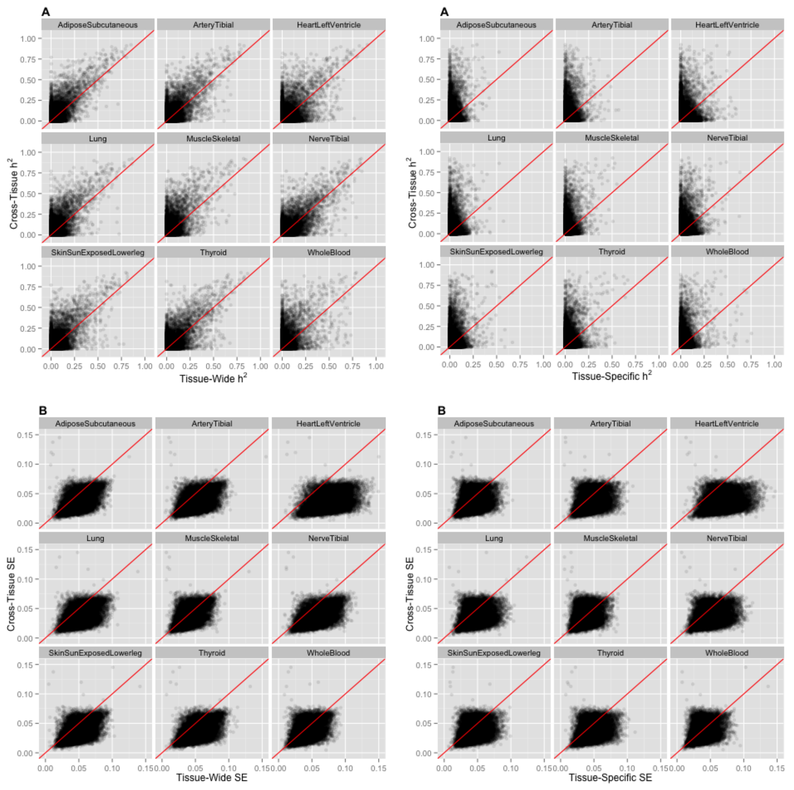


Cross-tissue, Tissue-wide and Tissue-specific h^2



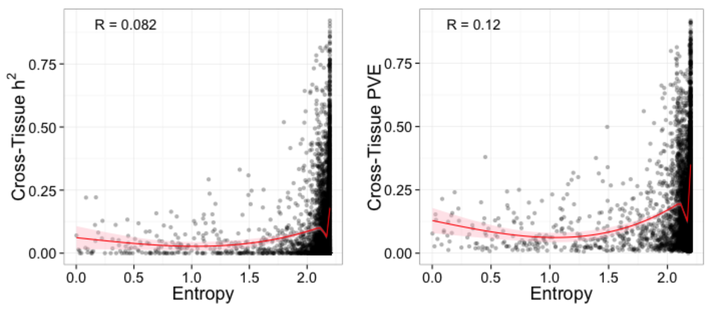
Left. Cross-tissue compared to tissue-wide h^2. Right. Cross-tissue compared to tissue-specific h^2.

Cross-tissue h^2 Estimates Have Lower Error



Cross-tissue heritability (h^2) compared to tissue-wide h^2. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-wide local h^2 is estimated using the measured gene expression for each respective tissue and SNPs within 1 Mb of each gene.

OTD Cross-tissue Expression Correlates with Multi-tissue eQTL Results

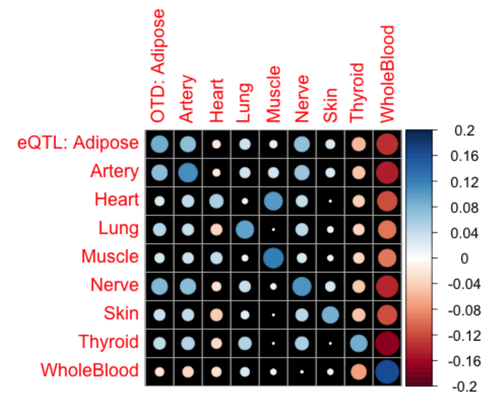


Entropy of the posterior probabilities from the Flutre et al. (PLOS Genetics 2013) multi-tissue eQTL method compared to the estimates of (left) heritability and (right) PVE of cross-tissue gene expression derived from the orthogonal tissue decomposition. The generalized additive model smoothing line is in red.

Cross-tissue Expression Has Larger Effective Sample Size

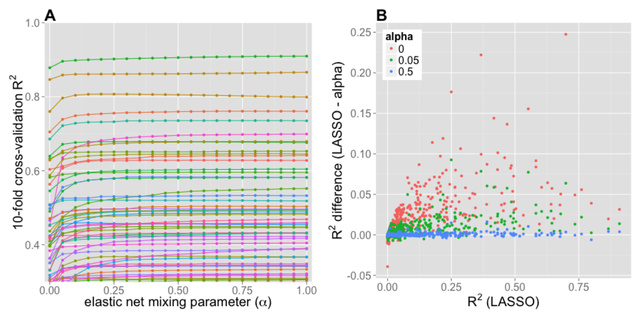
Tissue	n
Cross-Tissue	**450**
Muscle-Skeletal	361
WholeBlood	339
Skin-SunExposed(Lowerleg)	303
Adipose-Subcutaneous	298
Artery-Tibial	285
Lung	279
Thyroid	279
Nerve-Tibial	256
Heart-LeftVentricle	190

OTD Tissue-specific Expression Correlates with Multi-tissue eQTL Results



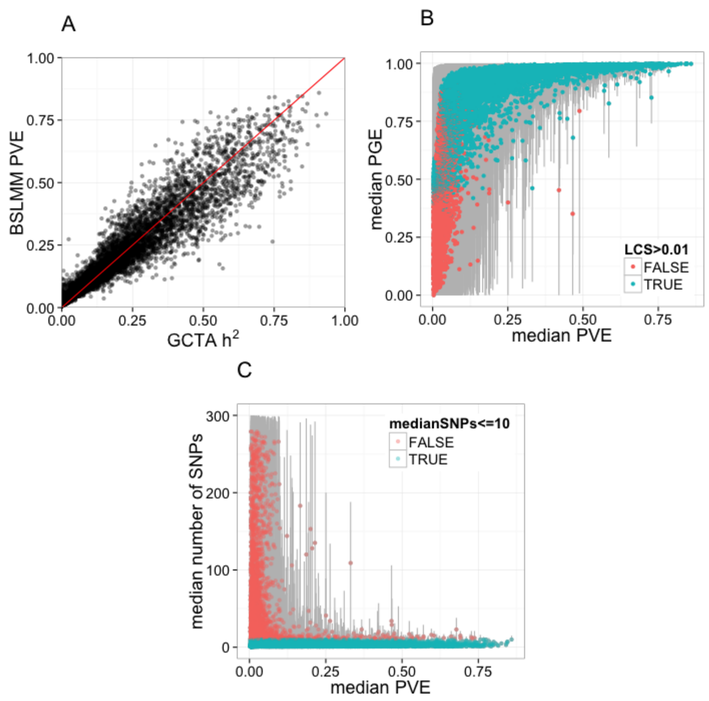
Pearson correlation (R) between the posterior probability the top multi-tissue eQTL regulates its gene in a given tissue (eQTL, Flutre et al. method, PLOS Genetics 2013) and the percent variance explained (PVE) of tissue-specific gene expression from the orthogonal tissue decomposition (OTD). Area of each circle is proportional to the absolute value of R.

Elastic Net Consistent with Sparse Architecture



Cross-validated predictive performance across the elastic net. (A) 10-fold cross-validated R^2 of predicted vs. observed expression in DGN whole blood compared to a range of elastic net mixing parameters (α) for genes on chromosome 22 with $R^2 > 0.3$. (B) Predictive R^2 difference between LASSO ($\alpha = 1$) and several other values of α compared to LASSO predictive R^2 for 341 genes on chromosome 22.

BSLMM Consistent with Sparse Archetecture



(A) Bayesian Sparse Linear Mixed Model (BSLMM)-estimated PVE (total proportion of variance explained) compared to GCTA-estimated heritability per gene (R=0.96) (B) Comparison of median PGE (proportion of PVE explained by sparse effects) to median PVE (total proportion of variance explained) for expression of each gene. The 95% credible set of each PGE estimate is in gray and genes with a lower credible set (LCS) greater than 0.01 are in blue. (C) Comparison of the median number of SNPs included in the model of each gene to median PVE. The 95% credible set of each SNP-number estimate is in gray and genes with a median of 10 or fewer SNPs are in blue.

Acknowledgements

NCI/K12CA139160; PAAR (NIH/NIGMS grant UO1GM61393); PGRN Statistical Analysis Resource (U19 HL065962); Genotype-Tissue Expression project (GTeX) (R01 MH101820 and R01 MH090937); University of Chicago DRTC (Diabetes Research and Training Center; P30 DK20595, P60 DK20595); The Conte Center for Computational Neuropsychiatric Genomics (P50MH094267); Integrated GWAS of complex behavioral and gene expression traits in outbred rats (P50DA037844).

Data Sources
GTEx (`gtexportal.org`), DGN (Depression Genes and Networks, `nimhgenetics.org`)

References

Gamazon ER*, Wheeler HE*, Shah KP* et al. A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics. 2015; 47: 1091-1098. doi:10.1038/ng.3367 *Contributed equally.

Flutre T et al. A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genetics. 2013; 9: e1003486. doi:10.1371/journal.pgen.1003486

*correspondence to hwheeler1@luc.edu or haky@uchicago.edu