Genetic architecture of transcriptome regulation and orthogonal tissue decompositon

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

2015-05-26 09:50:46 [1]Department of Medicine, University of Chicago, [2]Committee on Genetics, Genomics, and Systems Biology, University of Chicago, [3]Division of Genetic Medicine, Vanderbilt University

Abstract

Lorem ipsum dolor sit amet, est ad doctus eligendi scriptorem. Mel erat falli ut. Feugiat legendos adipisci vix at, usu at laoreet argumentum suscipiantur. An eos adhuc aliquip scriptorem, te adhuc dolor liberavisse sea. Ponderum vivendum te nec, id agam brute disputando mei.

Introduction

cite Kruglyak review, Price PGen 11, Wright NatGen 14

Results

Local genetic variation explains a large proportion of gene expression variance

We estimated the heritability of gene expression in whole blood from the Depression Genes and Networks (DGN) cohort (n=922) [1] using a mixed-effects model (see Methods) and calculated variances using restricted maximum likelihood as implemented in GCTA [2]. We fit a joint model with a local and a global genetic relationship matrix (GRM). The local GRM was derived from SNPs within 1 Mb of each gene and the global GRM was derived from SNPs that are located on non-gene chromosomes and are eQTLs in the Framingham Heart Study (FHS) cohort (n=5257, FDR < 0.05) [3]. The mean local h^2 was 0.130 and 54.6% of genes had a positive 95% confidence interval (CI), while the mean global h^2 was 0.076 and just 4.2% of genes had a positive CI (Fig 1). The maximum local h^2 was 0.93 with a standard error (SE) of 0.009 while the maximum global h^2 was 0.91 with a SE of 0.16. Similar results were observed for the 1194 genes with *trans*-eQTLs (FHS FDR < 0.05) when the global GRM was limited to known *trans*-eQTLs (Fig 2). That is, the mean local h^2 was 0.133 and 61.3% of genes had a positive 95% confidence interval (CI), while the mean *trans* h^2 was just 0.021 and 4.2% of genes tested had a positive CI.

Cross-tissue and tissue-specific gene expression by orthogonal tissue decomposition

In order to better understand the context specificity of gene expression regulation, we developed a method called orthogonal tissue decomposition (OTD), which uses a mixed effects model to generate cross-tissue and tissue-specific gene expression levels (see Methods). Using a marginal model with just the local GRM, we estimated the local h^2 of cross-tissue gene expression and tissue-specific gene expression in the nine tissues with the most samples. The cross-tissue heritabilities were larger and the standard errors were smaller than the tissue-specific estimates (Fig 3). The percentage of h^2 estimates with positive CIs was much larger for cross-tissue expression (17.3%) than the tissue-specific expressions (all less than 3%, Fig 4).

We also compared the cross-tissue h^2 from the OTD to h^2 estimates from the pre-OTD measures of gene expression in each of the nine tissues, which we term tissue-wide expression. Again, the cross-tissue heritabilities were larger and the standard errors were smaller than the tissue-wide estimates (Fig 5), though less striking than the tissue-specific comparison. The percentage of tissue-wide h^2 estimates with positive CIs ranged from 4.4-8.6% and thus were all larger than the tissue-specific postive CI percentages, but smaller than the cross-tissue percentage (Fig 6).

The effect of local genetic variation on gene expression is sparse rather than polygenic

We performed 10-fold cross-validation using the elastic net to test the predictive performance of local SNPs for gene expression across a range of mixing parameters, α. The α that gives the largest cross-validation R^2 informs the sparsity of each gene expression trait. That is, at one extreme, if the optimal $\alpha = 0$ (equivalent to ridge regression), the gene expression trait is highly polygenic, whereas if the optimal $\alpha = 1$ (equivalent to LASSO), the trait is highly sparse. We found that for most gene expression traits, the cross-validated R^2 was suboptimal for $\alpha = 0$ and $\alpha = 0.05$, but nearly identically optimal for $\alpha = 0.5$ and $\alpha = 1$ in the DGN cohort (Fig 7). Therefore, the effect of local genetic variation on gene expression is sparse rather than polygenic.

Citations

The relationship was first described by Reference 4. However, there are also opinions that the relationship is spurious [5]. We used R for our calculations [6], and we used package `knitcitations` [7] to make the bibliography.

Discussion

1. local + trans heritability (others have done this)
2. trans heritability estimates not reliable, proportion not reliable
3. orthogonal tissue decomposition
4. cross tissue + tissue specific heritability – estimates higher and se lower for cross-tissue The tissue availability is unbalanced because of the difficulties of sample collection and the uneven quality of the tissues. Furthermore, by using a mixed effects model to create cross-tissue expression, we borrow information across tissues, which should increase our power to detect associations and achieve better predictive models.
5. elastic net mixing parameter (alpha) as measure of polygenicity/sparsity

Future

1. simulation to show sparsity well represented by alpha
2. use number of PC's (computed using only local snps) that maximize prediction performance. This will count independent signals.
3. FHS heritability. could this improve trans heritability?

Methods

Genomic and Transcriptomic Data

DGN Dataset

We obtained whole blood RNA-Seq and genome-wide genotype data for 922 individuals from the Depression Genes and Networks (DGN) cohort [1], all of European ancestry. For our analyses, we used the HCP (hidden covariates with prior) normalized gene-level expression data used for the *trans*-eQTL analysis in Battle et al. [1] and downloaded from the NIMH repository. Approximately 650K SNPs (minor allele frequency [MAF] > 0.05, Hardy-Weinberg Equilibrium [P > 0.05], non-ambiguous strand [no A/T or C/G SNPs]) comprised the input set of SNPs for imputation, which was performed on the University of Michigan Imputation-Server (https://imputationserver.sph.umich.edu/start.html) [8,9] with the following parameters: 1000G Phase 1 v3 ShapeIt2 (no singletons) reference panel, SHAPEIT phasing, and EUR population. Approximately 1.9M non-ambiguous strand SNPs with MAF > 0.05, imputation $R^2 > 0.8$ and, to reduce computational burden, inclusion in HapMap Phase II were retained for subsequent analyses.

GTEx Dataset

Partitioning local and global heritability of gene expression

Orthogonal tissue decomposition

We will use a mixed effects model to estimate these components and enforce orthogonality of the components. This approach is an extension of our method to develop an intrinsic growth phenotype13.

Determining polygenicity versus sparsity using the elastic net

Equations

The deterministic part of the model is defined by this **in-line equation** as $\mu_i = \beta_0 + \beta_1 x$, and the stochastic part by the **centered equation**:

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_i)^2/(2\sigma^2)}$$

Tables

Table 1: This is a GLM summary table.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.06	0.11	0.55	0.59
x	2.02	0.11	18.07	0.00

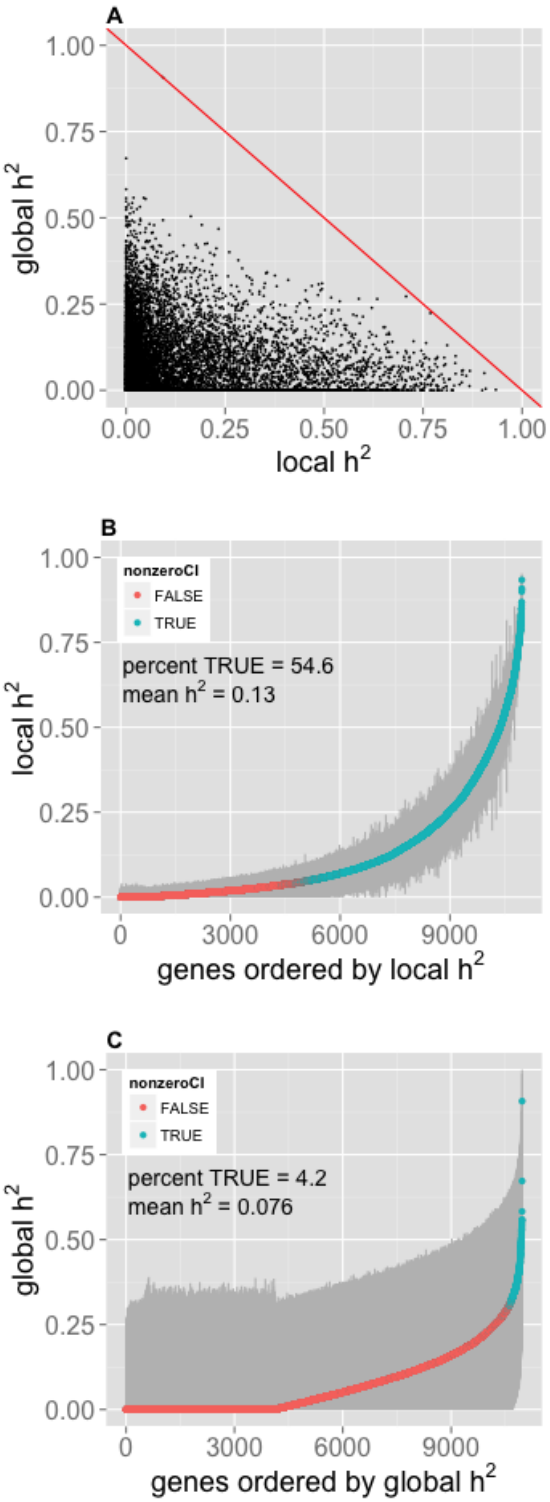

Figure 1: DGN whole blood expression joint heritability (h^2). Local (SNPs within 1 Mb of each gene) and global (SNPs that are eQTLs in the Framingham Heart Study on other chromosomes [FDR < 0.05]) h^2 for gene expression were jointly estimated. (**A**) Global h^2 compared to local h^2 per gene. (**B**) Local and (**C**) global 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.

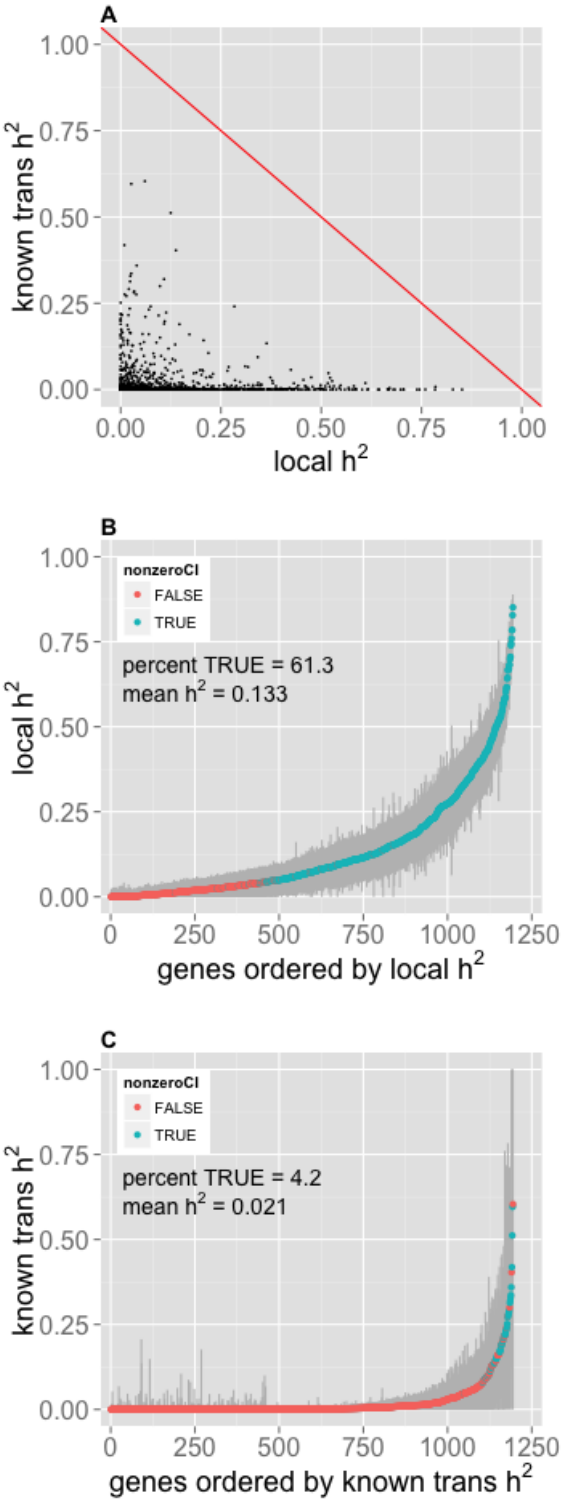

Figure 2: DGN whole blood expression joint heritability (h^2) with known trans-eQTLs. Local (SNPs within 1 Mb of each gene) and known trans (SNPs that are trans-eQTLs in the Framingham Heart Study for each gene [FDR < 0.05]) h^2 for gene expression were jointly estimated. (**A**) Known trans h^2 compared to local h^2 per gene. (**B**) Local and (**C**) known trans 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.

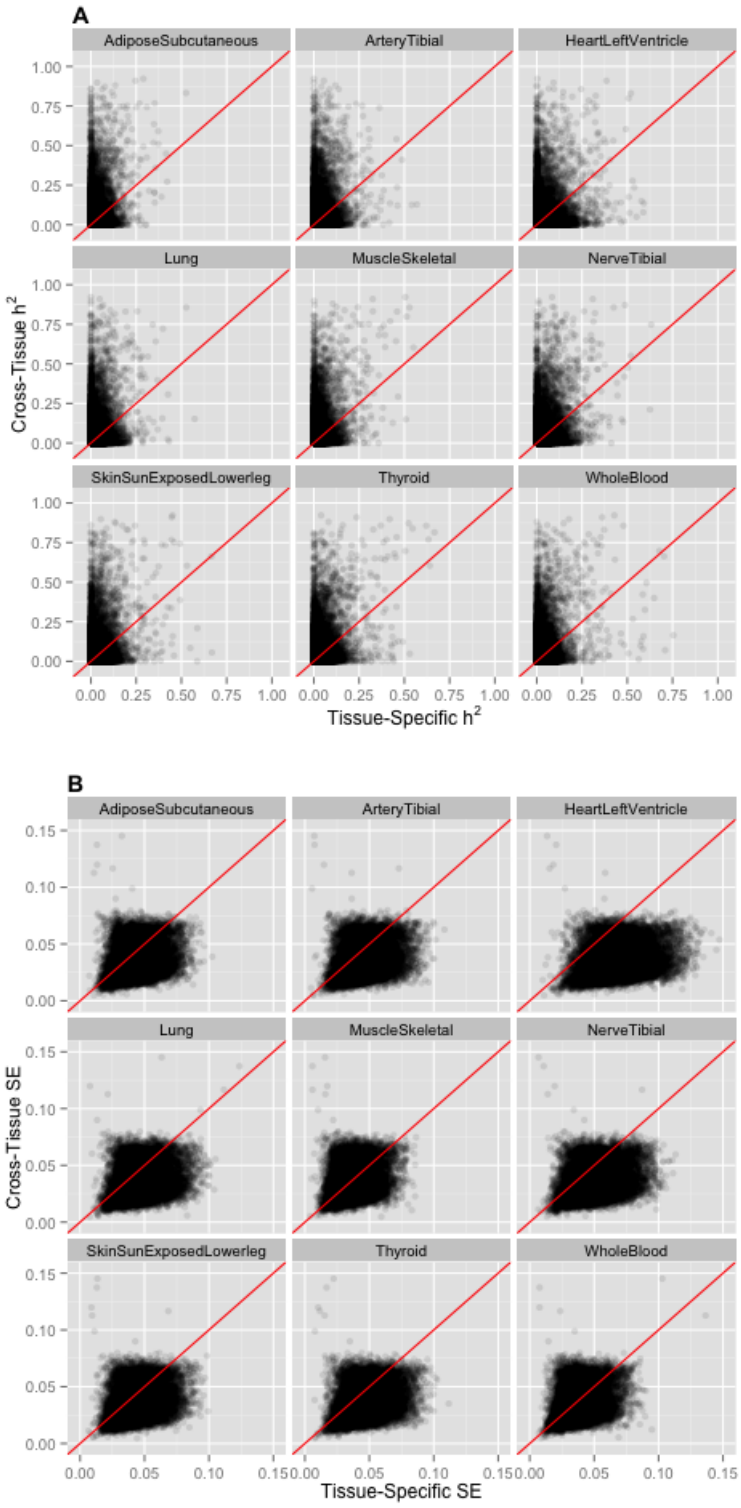

Figure 3: Cross-tissue and tissue-specific comparison of heritability (h^2, **A**) and standard error (SE, **B**) estimation. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-specifc local h^2 is estimated using the tissue-specific component (residuals) of the mixed effects model for gene expression for each respective tissue and SNPs within 1 Mb of each gene.

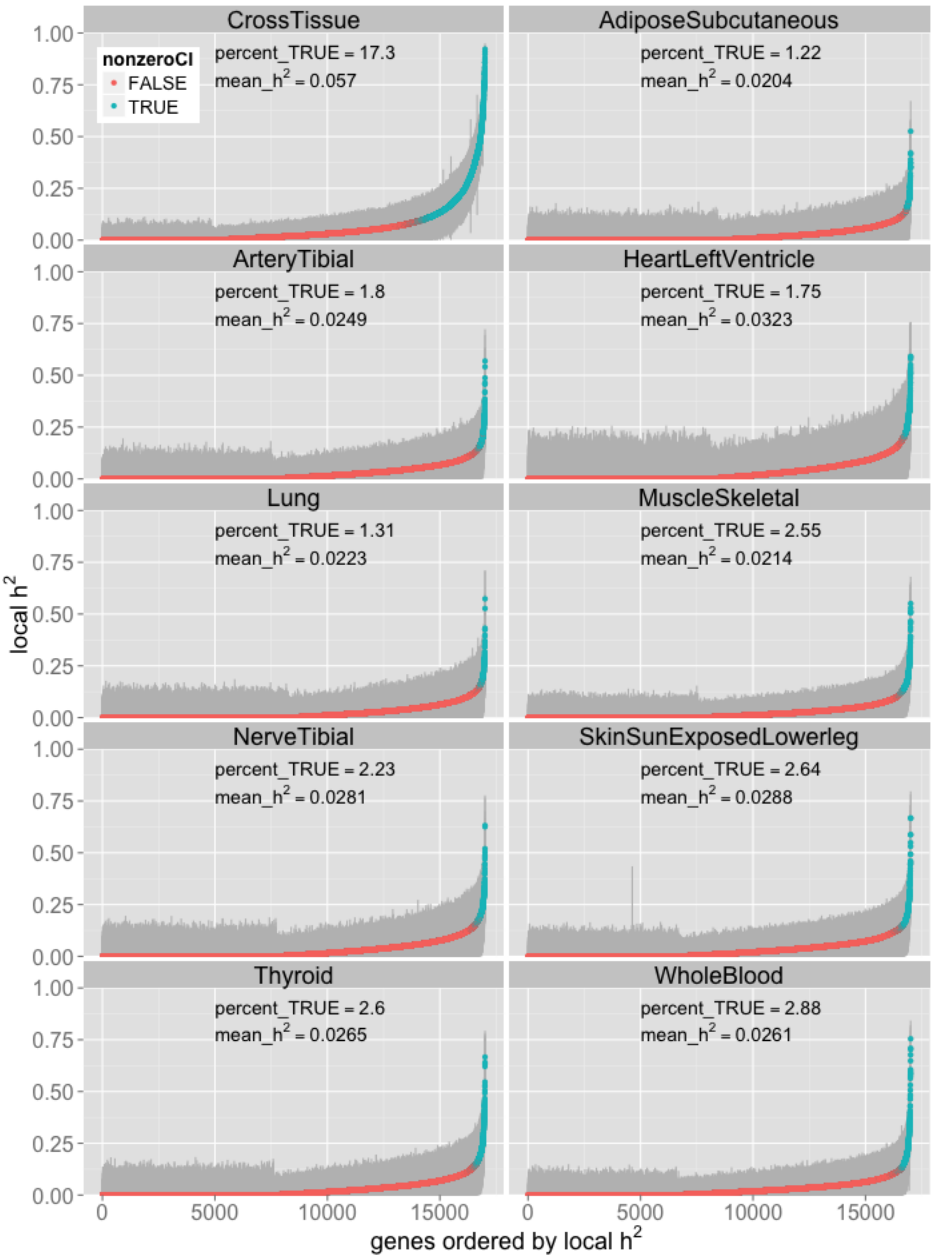

Figure 4: Cross-tissue heritability (h^2) compared to tissue-specific h^2. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-specifc local h^2 is estimated using the tissue-specific component (residuals) of the mixed effects model for gene expression for each respective tissue and SNPs within 1 Mb of each gene.

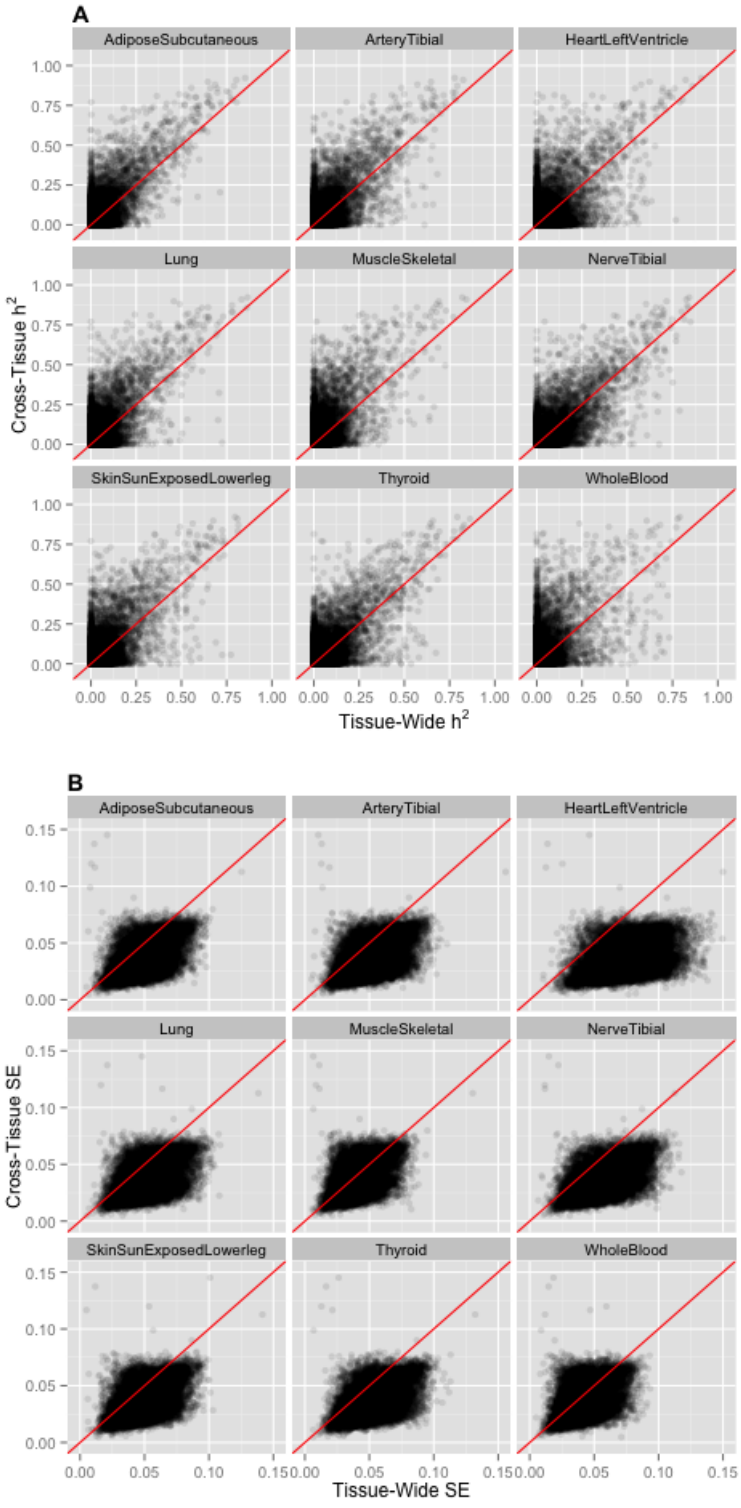

Figure 5: Cross-tissue and tissue-wide comparison of heritability (h^2, **A**) and standard error (SE, **B**). Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-wide local h^2 is estimated using the measured gene expression for each respective tissue and SNPs within 1 Mb of each gene.

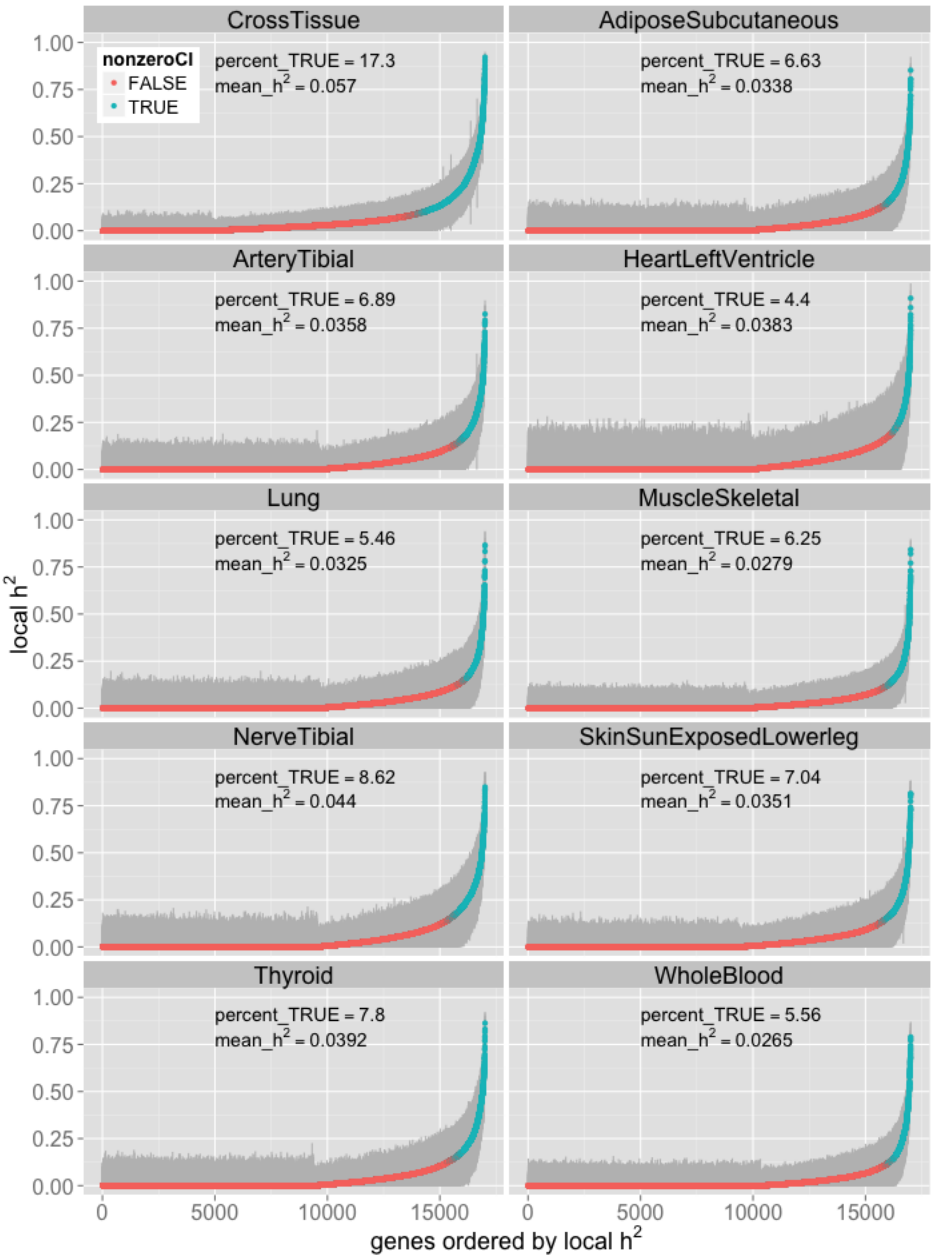

Figure 6: Cross-tissue heritability (h^2) compared to tissue-wide h^2. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-wide local h^2 is estimated using the measured gene expression for each respective tissue and SNPs within 1 Mb of each gene.

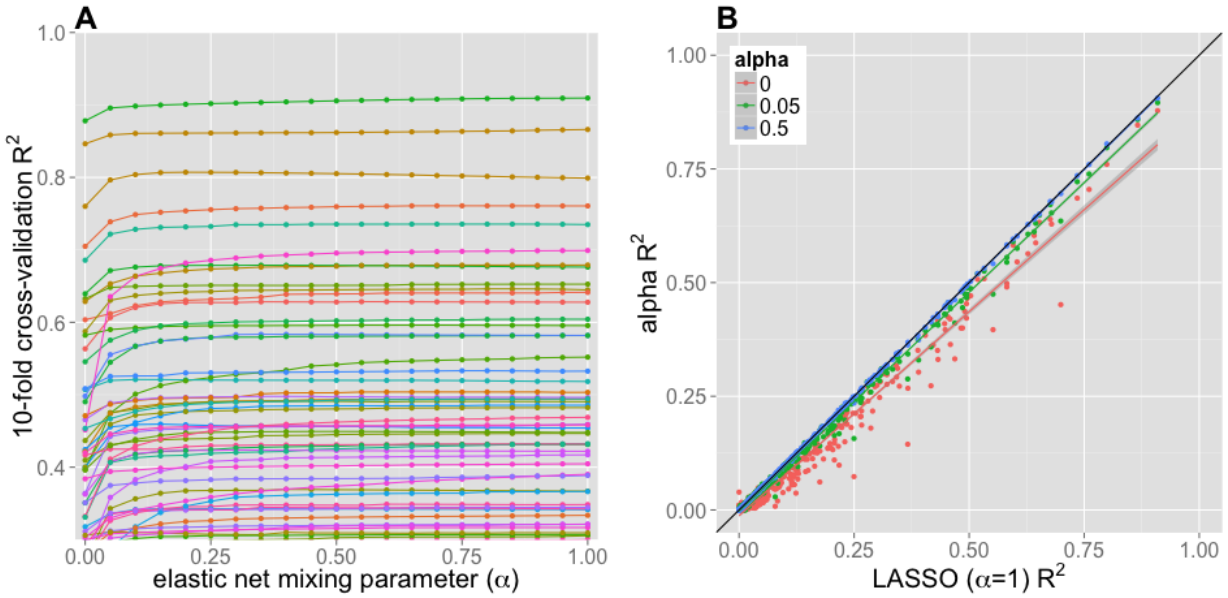

Figure 7: Cross-validated predictive performance across the elastic net. (**A**) 10-fold cross-validated R^2 of predicted vs. observed expression in DGN whole blood compared to a range of elastic net mixing parameters (α) for genes on chromosome 22 with $R^2 > 0.3$. (**B**) Predictive R^2 for several values of α compared to $\alpha = 1$ (LASSO) for 341 genes on chromosome 22.

Figures

Supplemental Figures

References

1. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research. Cold Spring Harbor Laboratory Press; 2013;24: 14–24. doi:10.1101/gr.155192.113

2. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. The American Journal of Human Genetics. Elsevier BV; 2011;88: 76–82. doi:10.1016/j.ajhg.2010.11.011

3. Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, et al. Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat Genet. Nature Publishing Group; 2015;47: 345–352. doi:10.1038/ng.3220

4. Halpern BS, Regan HM, Possingham HP, McCarthy MA. Accounting for uncertainty in marine reserve design. Ecol Letters. Wiley-Blackwell; 2006;9: 2–11. doi:10.1111/j.1461-0248.2005.00827.x

5. Keil P, Belmaker J, Wilson AM, Unitt P, Jetz W. Downscaling of species distribution models: A hierarchical approach. Freckleton R, editor. Methods Ecol Evol. Wiley-Blackwell; 2012;4: 82–94. doi:10.1111/j.2041-210x.2012.00264.x

6. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2015. Available: http://www.R-project.org/

7. Boettiger C. Knitcitations: Citations for knitr markdown files [Internet]. 2015. Available: http://CRAN.R-project.org/package=knitcitations

8. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. Nature Publishing Group; 2012;44: 955–959. doi:10.1038/ng.2354

9. Fuchsberger C, Abecasis GR, Hinds DA. Minimac2: Faster genotype imputation. Bioinformatics. Oxford University Press (OUP); 2014;31: 782–784. doi:10.1093/bioinformatics/btu704