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]

2015-06-01 10:17:11 [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 [4] 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.

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. The 922 individuals were unrelated (all pairwise $\hat{\pi} < 0.05$) and thus all included in downstream analyses. Imputation of approximately 650K input SNPs (minor allele frequency [MAF] > 0.05, Hardy-Weinberg Equilibrium [$P > 0.05$], non-ambiguous

strand [no A/T or C/G SNPs]) was performed on the University of Michigan Imputation-Server (https://imputationserver.sph.umich.edu/start.html) [5,6] 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

We obtained RNA-Seq gene expression levels from 8555 tissue samples (53 unique tissue types) from 544 unique subjects in the GTEx Project [7] data release on 2014-06-13. Of the individuals with gene expression data, genome-wide genotypes (imputed with 1000 Genomes) were available for 450 individuals. While all 8555 tissue samples were used in the OTD model (described below) to generate cross-tissue and tissue-specific components of gene expression, we used the nine tissues with the largest sample sizes when quantifying tissue-specific effects. Tissues and sample sizes (both RNA-seq and genotypes available) included cross-tissue (n=450), skeletal muscle (n=361), whole blood (n=339), skin from the sun-exposed portion of the lower leg (n=303), subcutaneous adipose (n=298), tibial artery (n=285), lung (279), thyroid (n=279), tibial nerve (n=256) and left ventricle heart (n=190). Approximately 2.6M non-ambiguous strand SNPs included in HapMap Phase II were retained for subsequent analyses.

Partitioning local and global heritability of gene expression

To investigate the proximity of gene expression regulation to each gene, we partitioned the proportion of gene expression variance explained by SNPs in the DGN cohort into two components: local (SNPs within 1Mb of the gene) and global (eQTLs on non-gene chromosomes) as defined by the GENCODE [8] version 12 gene annotation. We calculated the proportion of the variance (narrow-sense heritability) explained by each component using the following mixed-effects model:

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

Assuming a random effects for $w_{k,g} \approx N(0, \sigma_w^2)$ and $\epsilon \approx N(0, \sigma_\epsilon^2 I_n)$, where I_n is the identity matrix, we calculated the total variability explained by local and global components by estimating σ_w^2 with restricted maximum likelihood (REML) using GCTA software [2]. For heritability analyses in the GTEx cohort, we removed the *global* term from the model and only estimated marginal *local* h^2 due to the smaller sample sizes of both cross-tissue and tissue-specific expression levels compared to DGN.

Orthogonal tissue decomposition

To better understand the context specificity of gene expression regulation, we developed a method called orthogonal tissue decomposition (OTD). This approach is an extension of our method to develop an intrinsic growth phenotype [9]. We applied OTD to GTEx Project [7] data and decomposed the expression of each gene into cross-tissue and tissue-specific components. The tissue availability is unbalanced across individuals because of the difficulties of sample collection and the uneven quality of the tissues. OTD decomposes the expression traits into orthogonal components as represented by the following model:

$$Y_i = T_{i,cross} + T_{i,tissue}$$

Specifically, to generate cross-tissue and tissue-specific expression levels, we used the `lmer` function in the R [10] package `lme4` [11,12] to fit the following mixed-effects model:

```
fit <- lme4::lmer(expression ~ (1|SUBJID) + TISSUE + GENDER + PEERs)
```

The model included tissue-wide gene expression levels in 8555 GTEx tissue samples from 544 unique subjects. A total of 17,647 Protein-coding genes (defined by GENCODE [8] version 18) with a mean gene expression level across tissues greater than 0.1 RPKM (reads per kilobase of transcript per million reads mapped) were included in the model. `SUBJID` was a random effect and the covariates `TISSUE`, `GENDER`, and `PEERs` were fixed effects used to predict tissue-wide expression levels (`expression` in the model). `PEERs` included the top 15 PEER factors estimated across all tissues using the R package `PEER` [13] to control for batch effects and experimental confounders. Cross-tissue expression was defined as the random effects from the model (`ranef(fit)`) and tissue-specific expression as the residuals (`resid(fit)`).

Determining polygenicity versus sparsity using the elastic net

We applied the elastic net [4] to model the effect of local genetic variation (SNPs within 1 Mb of gene) on the genetic architecture of gene expression. We used the `cv.glmnet` function in the R package `glmnet` [14,15] to perform 10-fold cross-validation of the elastic net across a range of mixing paramaters (α) to find the α that maximized predictive performance, measured by Pearson's R^2. Specifically, `glmnet` solves the following problem:

[[Haky, should/can we simplify this equation? I just took it from http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html, but they don't explain all the terms]]

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1 \right],$$

over a grid of values of λ covering the entire range [14,15]. This tuning parameter λ controls the overall strength of the penalty.

The elastic net penalty is controlled by mixing parameter α, which spans LASSO ($\alpha = 1$, the default) [16] at one extreme and ridge regression ($\alpha = 0$) [17] at the other. The ridge penalty shrinks the coefficients of correlated SNPs towards each other, while the LASSO tends to pick one of the correlated SNPs and discard the others. Thus, an optimal prediction R^2 for $\alpha = 0$ means the gene expression trait is highly polygenic, while an optimal prediction R^2 for $\alpha = 1$ means the trait is highly sparse. An optimal prediction R^2 in between (e.g. $\alpha = 0.5$) means the trait has a mixed genetic architecture.

In the DGN cohort, we tested 21 values of the mixing parameter ($\alpha = 0, 0.05, 0.1, ..., 0.90, 0.95, 1$) for optimal prediction of gene expression of the 341 genes on chromosome 22. For the rest of the autosomes in DGN and for cross-tissue, tissue-specific, and tissue-wide expression in the GTEx cohort, we tested $\alpha = 0.05, 0.5, 0.95, 1$. **[[Nick, is this correct?]]**

Enrichment analysis

- For top CT and TS genes:

1. GO enrichment
2. GWAS catalog enrichment (i.e. top T2D, T1D, schizo, etc. genes)

Tables

Table 1: This is a GLM summary table.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.13	0.1	1.27	0.21
x	2.03	0.1	19.64	0.00

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. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). Wiley-Blackwell; 2005;67: 301–320. doi:10.1111/j.1467-9868.2005.00503.x

5. 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

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

7. Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, et al. The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. American Association for the Advancement of Science (AAAS); 2015;348: 648–660. doi:10.1126/science.1262110

8. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Research. Cold Spring Harbor Laboratory Press; 2012;22: 1760–1774. doi:10.1101/gr.135350.111

9. Im HK, Gamazon ER, Stark AL, Huang RS, Cox NJ, Dolan ME. Mixed effects modeling of proliferation rates in cell-based models: Consequence for pharmacogenomics and cancer. Akey JM, editor. PLoS Genetics. Public Library of Science (PLoS); 2012;8: e1002525. doi:10.1371/journal.pgen.1002525

10. 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/

11. Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using eigen and s4 [Internet]. 2014. Available: http://CRAN.R-project.org/package=lme4

12. Bates D, Maechler M, Bolker BM, Walker S. lme4: Linear mixed-effects models using eigen and s4 [Internet]. 2014. Available: http://arxiv.org/abs/1406.5823

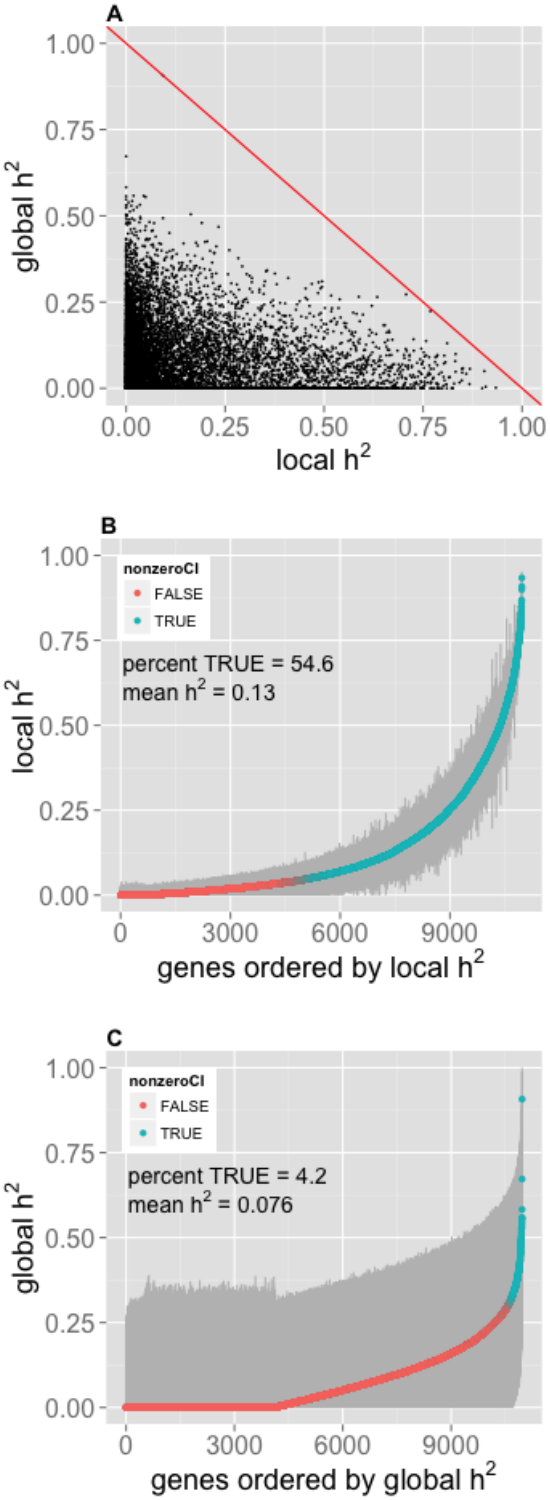

Figure 1: DGN whole blood expression joint heritability (h^2). Local (SNPs within 1 Mb of each gene) and global (SNPs that are eQTLs in the Framingham Heart Study on other chromosomes [FDR < 0.05]) h^2 for gene expression were jointly estimated. (**A**) Global h^2 compared to local h^2 per gene. (**B**)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

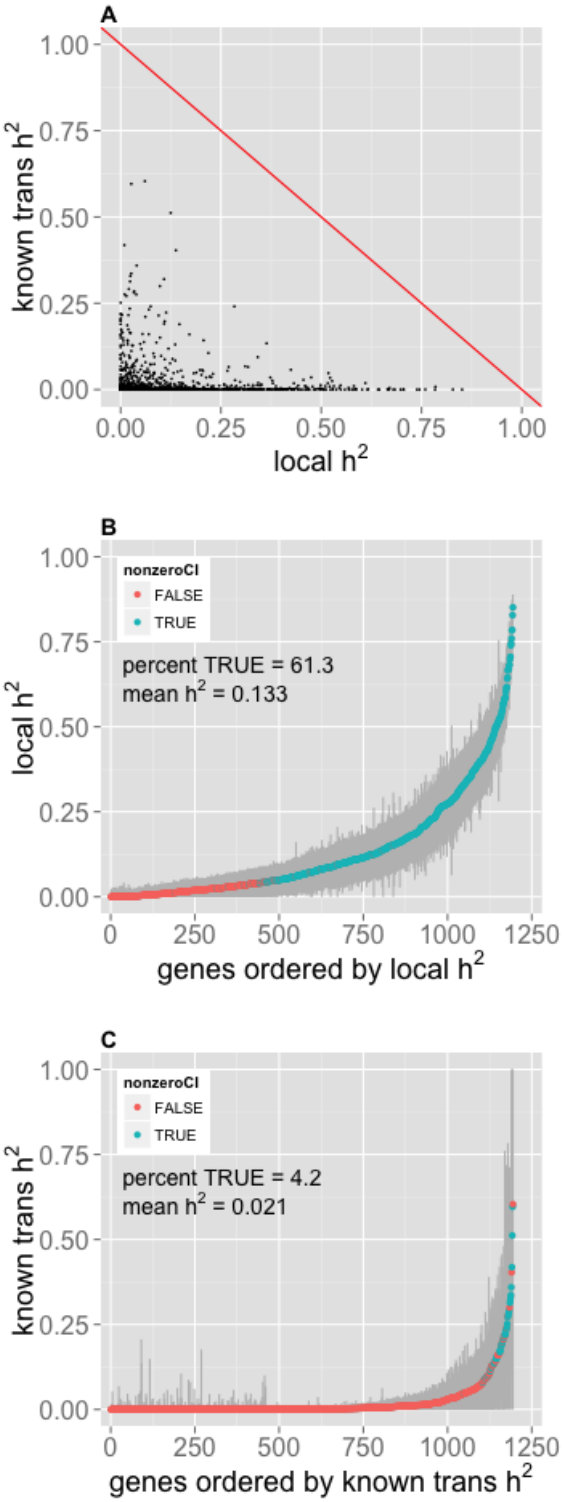

Figure 2: DGN whole blood expression joint heritability (h^2) with known trans-eQTLs. Local (SNPs within 1 Mb of each gene) and known trans (SNPs that are trans-eQTLs in the Framingham Heart Study for each gene [FDR < 0.05]) h^2 for gene expression were jointly estimated. (**A**) Known trans h^2 compared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

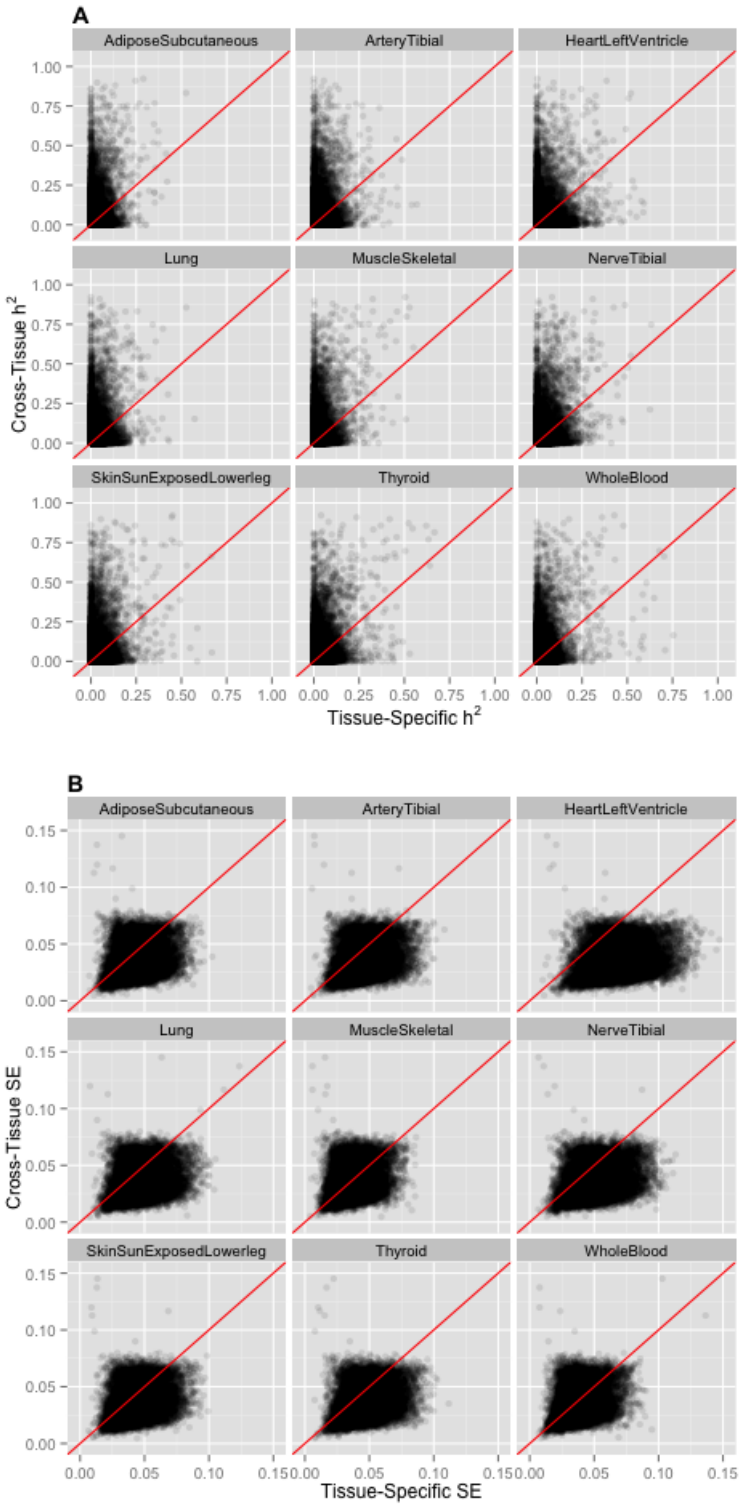

Figure 3: Cross-tissue and tissue-specific comparison of heritability (h^2, **A**) and standard error (SE, **B**) estimation. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-specifc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

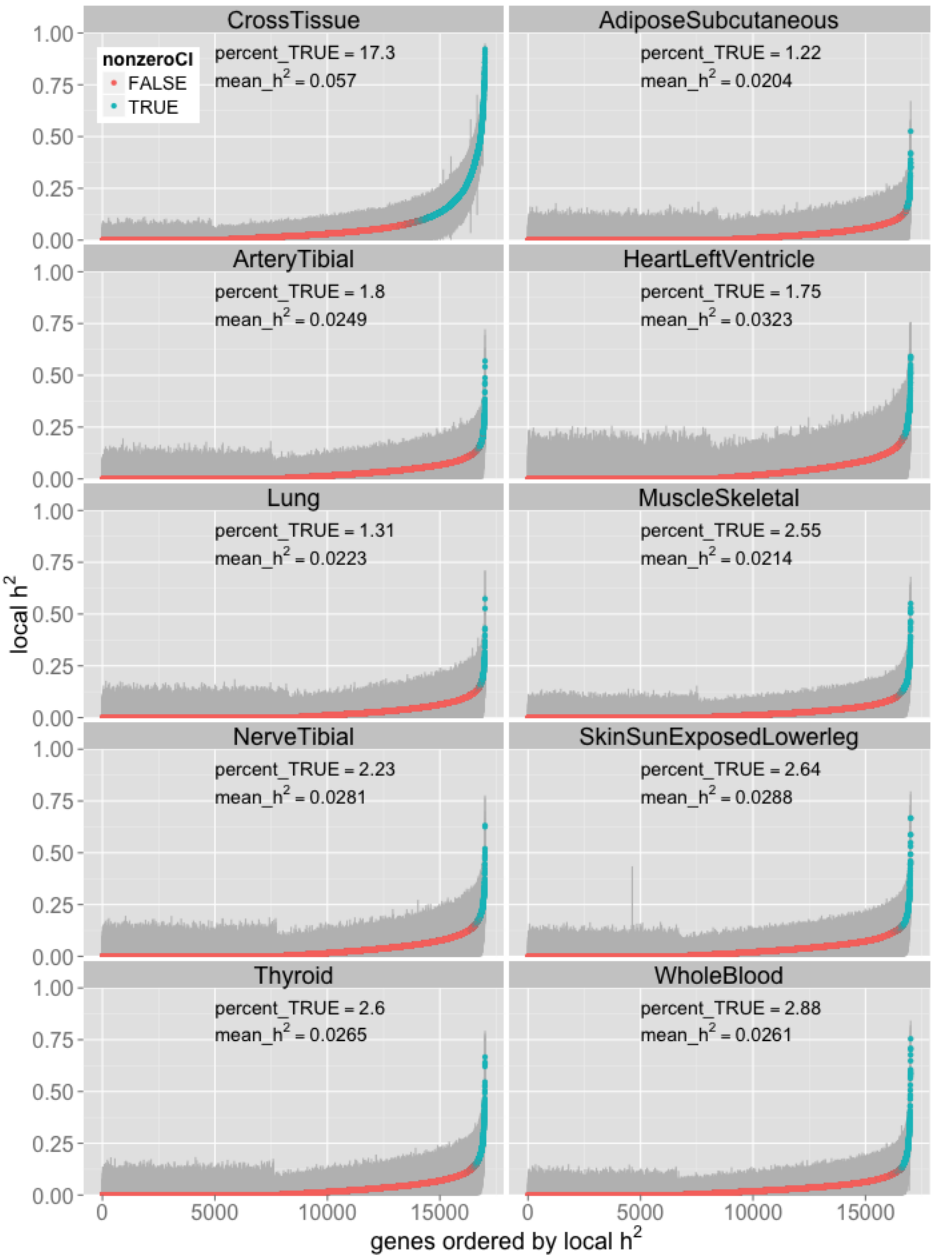

Figure 4: Cross-tissue heritability (h^2) compared to tissue-specific h^2. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-specifc local h^2 is estimated using the tissue-specific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

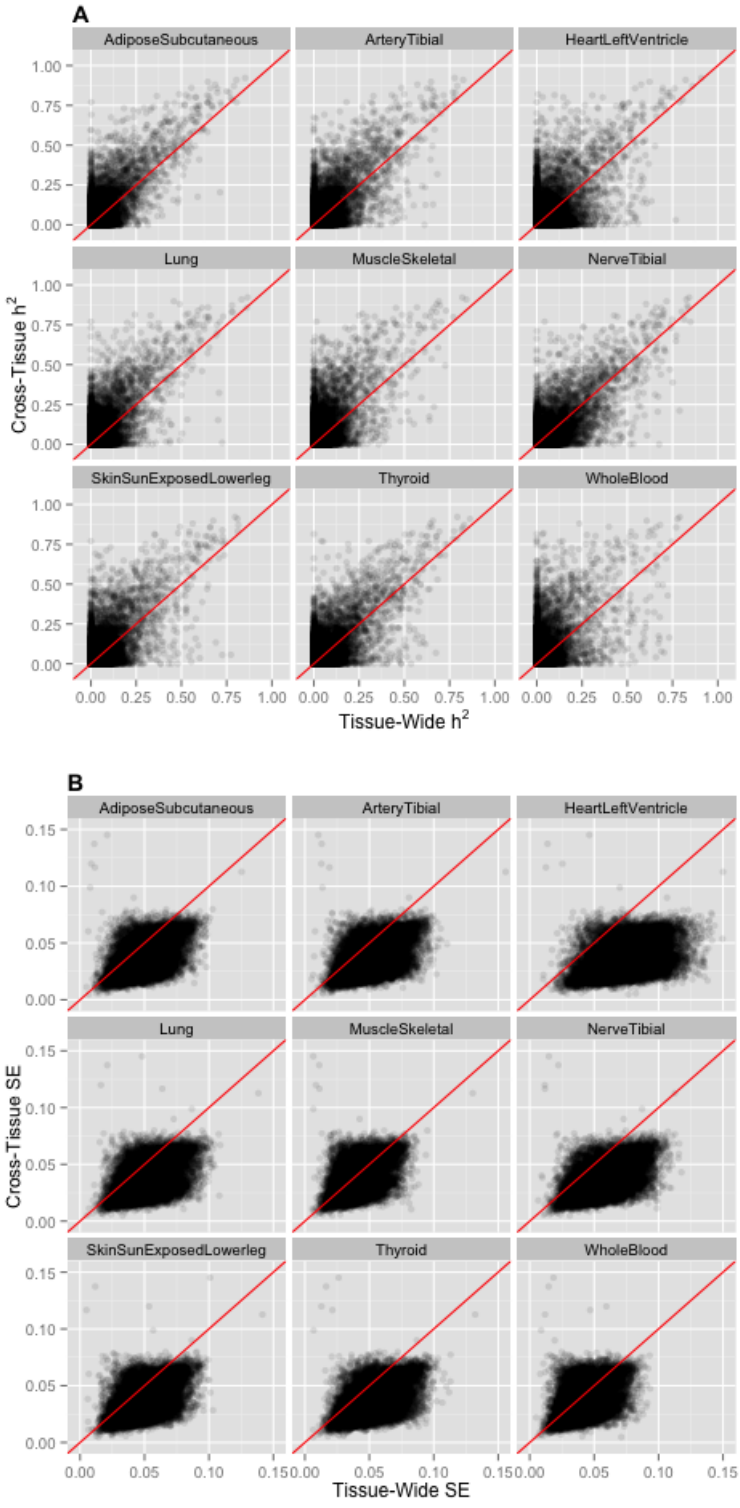

Figure 5: Cross-tissue and tissue-wide comparison of heritability (h^2, **A**) and standard error (SE, **B**). Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-wide local h^2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

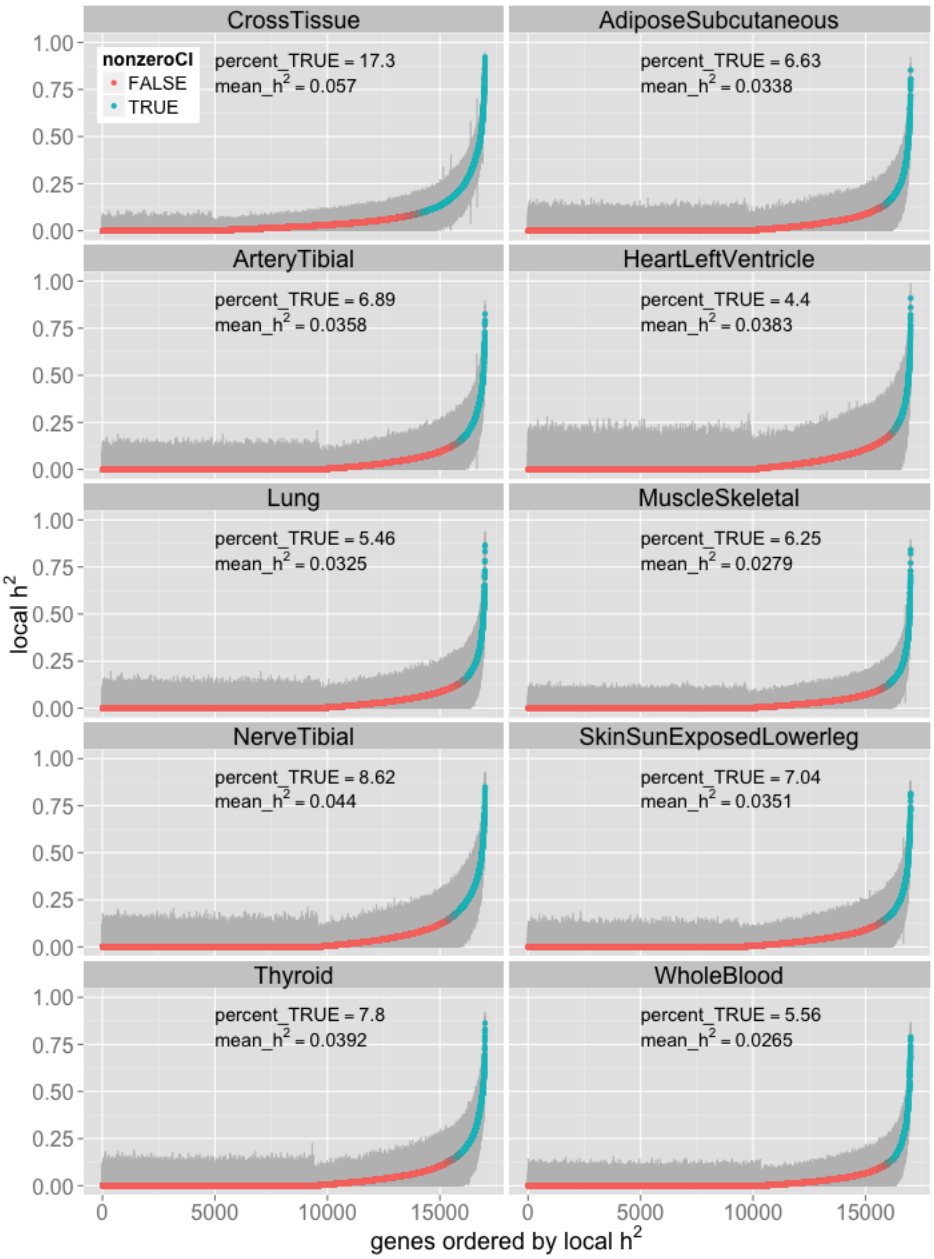

Figure 6: Cross-tissue heritability (h^2) compared to tissue-wide h^2. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-wide local h^2 is estimated using the measured gene expression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

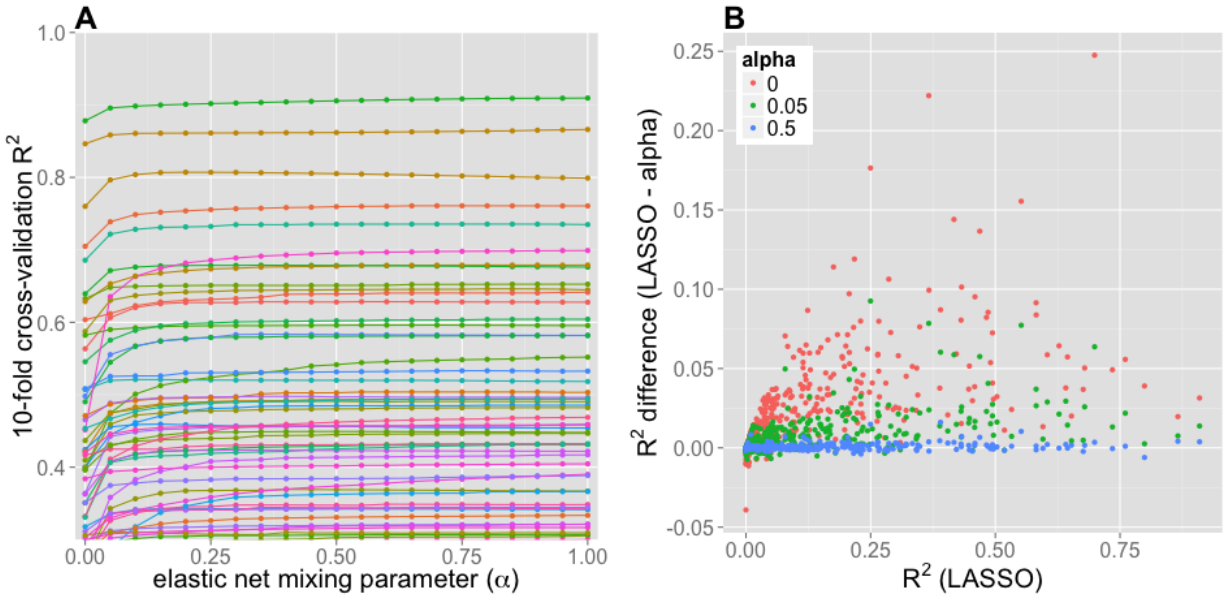

Figure 7: Cross-validated predictive performance across the elastic net. (**A**) 10-fold cross-validated R^2 of predicted vs. observed expression in DGN whole blood compared to a range of elastic net mixing parameters (α) for genes on chromosome 22 with $R^2 > 0.3$. (**B**) Predictive R^2 difference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

13. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature Protocols. Nature Publishing Group; 2012;7: 500–507. doi:10.1038/nprot.2011.457

14. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010;33: 1–22. Available: http://www.jstatsoft.org/v33/i01/

15. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for cox's proportional hazards model via coordinate descent. Journal of Statistical Software. 2011;39: 1–13. Available: http://www.jstatsoft.org/v39/i05/

16. Regression shrinkage and selection via the lasso on jSTOR [Internet]. http://www.jstor.org/stable/2346178; 2015. Available: http://www.jstor.org/stable/2346178

17. Hoerl AE, Kennard RW. Ridge regression: Applications to nonorthogonal problems. Technometrics. Informa UK Limited; 1970;12: 69–82. doi:10.1080/00401706.1970.10488635