An atlas of the genetic architecture of gene expression traits across the entire human body

Heather E. Wheeler[1,2], GTEx Consortium, Kaanan P. Shah[3], . . ., Nancy J. Cox[4], Dan L. Nicolae[3], Hae Kyung Im[3]

[1]Department of Biology and [2]Department of Computer Science, Loyola University Chicago, [3]Section of Genetic Medicine, Department of Medicine, University of Chicago, [4]Division of Genetic Medicine, Vanderbilt University 2016-01-26 16:16:19

Abstract

For most complex traits, gene regulation is known to play a crucial mechanistic role as demonstrated by the consistent enrichment of expression quantitative trait loci (eQTLs) among trait-associated variants. For this reason, the GTEx project has generated RNA-Seq data on hundreds of individuals across more than 40 tissues providing a comprehensive atlas of gene expression traits. Here, we systematically examined the local versus distant heritability as well as the sparsity versus polygenicity of protein coding gene expression traits in tissues across the entire human body. To determine tissue context specificity, we decomposed the expression levels into cross-tissue and tissue-specific components via orthogonal tissue decomposition (OTD). Regardless of tissue type, we found that local heritability can be well characterized with current sample sizes. Unless strong functional priors are used, the heritability due to distant variants cannot be estimated. We also find that the distribution of effect sizes is more consistent with a sparse architecture across all tissues. We also show that the cross-tissue and tissue-specific expression phenotypes constructed with our OTD model recapitulate complex Bayesian multi-tissue analysis results demonstrating that they reflect the expected biology. Finally, this knowledge was applied to develop prediction models of gene expression traits for all

tissues. The prediction models, heritability for each tissue, and prediction performance R^2 are made publicly available (https://github.com/hakyimlab/PrediXcan).

Introduction

Regulatory variation plays a key role in the genetics of complex traits [1–3]. While many common diseases have are likely polygenic [4–6], it is unclear whether gene expression levels are also polygenic or instead have simpler genetic architectures. It is also unclear how much these expression architectures vary across genes [7]. Most human expression quantitative trait loci (eQTL) studies have focused on how local genetic variation affects gene expression in order to reduce the multiple testing burden that would be required for a global analysis [7,8]. Furthermore, when both local and distal eQTLs are reported [9–11], effect sizes and replicability are much higher for local eQTLs. Indeed, while the heritability of gene expression attributable to local genetic variation has been estimated accurately, large standard errors have prevented accurate estimation of the contribution of distal genetic variation to gene expression variation [11,12].

We assessed the ability of various models, with different underlying assumptions, to predict gene expression in order to both understand the underlying genetic architecture of gene expression and to further optimize predictors for our complex trait prediciton method, PrediXcan [13]. In our PrediXcan paper, we showed that a polygenic score model was suboptimal to more sparse models like the elastic net model with $\alpha = 0.5$. However, a top eQTL only model did not do as well as the elastic net either [13], suggesting that for many genes, the genetic architecture is sparse, but not regulated by one SNP. Gene expression traits with sparse architecture should be better predicted with models such as LASSO (Least Absolute Shrinkage and Selection Operator), which prefers solutions with fewer parameters, each of large effect [14]. Conversely, highly polygenic traits should be better predicted with ridge regression or similarly polygenic models that prefer solutions with many parameters, each of small effect [15–17]. To obtain a more thorough understanding of gene expression achitecture, we used the hybrid approaches of the elastic net and BSLMM (Bayesian Sparse Linear Mixed Model) [18] to quantify sparse and polygenic effects.

Most previous human eQTL studies were performed in whole blood or lymphoblastoid cell lines due to ease

of access or culturabilty [9,19,20]. The Genotype-Tissue Expression (GTEx) Project aims to examine the genetics of gene expression more comprehensively and recently published a pilot analysis of eQTL data from 1641 samples across 43 tissues from 175 individuals confirming that eQTLs are highly shared across tissues [21]. We also showed that gene expression predictors perform well across tissues [13]. In order to harness this cross-tissue effect for prediction and to better understand the genetic architecture of tissue-specific and cross-tissue gene regulation, we developed a mixed effects model called orthogonal tissue decomposition (OTD) to determine the cross-tissue and tissue-specific components of gene expression in the rich GTEx dataset. We modeled the underlying genetic architecture of the cross-tissue and tissue-specific gene expression components and developed predictors for use in PrediXcan [13].

Results

Local genetic variation can be well characterized for all tissues

We estimated the heritability of gene expression in whole blood from the Depression Genes and Networks (DGN) cohort (n=922) [20] using a mixed-effects model (see Methods) and calculated variances using restricted maximum likelihood as implemented in GCTA [22]. We fit two joint models, each with the same local genetic relationship matrix (GRM) and a two different distal GRMs. The local GRM was derived from SNPs within 1 Mb of each gene in both models. In model 1, the distal GRM was derived from all SNPs on non-gene chromosomes. In model 2, the distal 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) [23]. The mean local h^2 was similar in both models, 0.123 in model 1 and 0.130 in model 2 (Fig 1). Of expressed genes, 52.6% and 54.6% had a positive 95% confidence interval (CI) for the local estimate in models 1 and 2, respectively. The mean distal h^2 was 0.123 with a standard error (SE) of 0.284 in model 1, while the mean distal h^2 was 0.076 with a SE of 0.142 in model 2. Using functional priors (known eQTLs) to define distal h^2 in model 2 increased the percentage of genes with a positive CI from 3.2% to 4.2% (Fig 1). In model 2, the maximum local h^2 was 0.93 with a standard error (SE) of 0.009 while the maximum distal h^2 was 0.91 with a SE of 0.16.

We also estimated the heritability of gene expression in the 9 tissues from the GTEx Project with the largest sample sizes, which ranged from 190-361 samples (see Methods). Mean local h^2 estimates ranged from 0.028-0.044 with 4.4-8.6% of genes having a positive CI (Fig 2). While the reduced sample sizes reduced the overall mean, maximum local h^2 estimates in each tissue ranged from 0.79-0.91, similar to the maximum seen in DGN. As expected, distal h^2 (eQTLs on non-gene chromosomes) could not be reliably estimated in these smaller sample sizes as less than 1% of genes in each tissue had a positive CI in the joint model (Fig 3-[**SUP**])). Thus, we focus on the local architecture in subsequent sections.

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

We performed 10-fold cross-validation using the elastic net [24] 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 $\alpha = 0.5$ through $\alpha = 1$ in the DGN cohort (Fig 4). An $\alpha = 0.05$ was also clearly suboptimal for gene expression prediction in the nine GTEx tissues, while models with $\alpha = 0.5, 0.95,$ or 1 had similar predictive power (Fig 5-[**SUP**]). This suggests that for most genes, the effect of local genetic variation on gene expression is sparse rather than polygenic.

To further examine sparsity and polygenicity, we used BSLMM [18] to define the total proportion of variance in expression explained by sparse and polygenic effects together (PVE) and the proportion of this genetic variance explained by sparse effects (PGE) for local SNPs in each gene in the DGN cohort. The PVE can be thought of as a Bayesian estimate of chip heritability and, indeed, there is a strong correlation between BSLMM-estimated PVE and GCTA-estimated h^2 (Fig 6A, R=0.96). For genes with large PVE, the PGE also was large, indicative of a sparse genetic architecture. For example, all genes with PVE > 0.50 had PGE > 0.82 and their median PGE was 0.989 (Fig 6B). The median PGE for genes with PVE > 0.1 was 0.949.

Fittingly, for most (96.3%) of the genes with PVE estimates > 0.10, the median number of SNPs included in the model was no more than 10 (Fig 6C).

Interestingly, when we applied BSLMM to the GTEx data, we found that many genes had strikingly larger BSLMM-estimated PVE than GCTA-estimated h^2 (Fig 7). This likely reflects the increased power of the BSLMM method at the lower sample sizes ($n \leq 361$) present in GTEx to estimate variance explained when the trait is more sparse than polygenic. GCTA assumes an underlying polygenic model, but as we saw in DGN, BSLMM-estimated PVE and GCTA-estimated h^2 are more correlated when the sample size is larger (n=922, Fig 6A). As we observed in DGN, genes with larger PVE estimates were more likely to have a PGE estimate approaching 1 with a lower credible set greater than 0.01 in each of the nine GTEx tissues (Fig 8).

Cross-tissue expression phenotype has increased predictive power and recapitulates known multi-tissue eQTL target genes

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 9-[**SUP**]). The percentage of GCTA 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 10). Similarly, the percentage of BSLMM PVE estimates with a lower credible set greater than 0.01 was 49% for cross-tissue expression, but ranged from 24-27% for tissue-specific expression (Fig 11).

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 whole tissue expression. Again, the cross-tissue heritabilities were larger and the standard errors were smaller than the whole tissue estimates (Fig 12-[**SUP**]), though less striking than the tissue-specific comparison. The percentage of whole tissue 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 2). Cross-tissue BSLMM PVE estimates had lower error than whole tissue PVE (Fig 8, Fig 11). Like whole tissue expression, cross-tissue and tissue-specific expression showed better predictive performance using the elastic-net when $\alpha \geq 0.5$ than when $\alpha = 0.05$ (Fig 13-[**SUP**]). Cross-tissue predictive performance exceeded that of both tissue-specific and whole tissue expression as indicated by higher cross-validated R^2 (Fig 5-[**SUP**], Fig 13-[**SUP**]).

We compared our OTD results to those from a joint multi-tissue eQTL analysis method [25], which was previously performed on a subset of the GTEx data [21]. The results of this analysis include eQTL posterior probabilities for each of the nine tissues, which can be interpreted as the probability a SNP is an eQTL in tissue x given the data. Using the top eQTL for each gene, we defined an entropy statistic (see Methods) that combines the nine posterior probabilities into one value such that higher entropy values mean the gene is more likely to be regulated by the same eQTL across all nine tissues, rather than in just a subset of the nine. We observed a strong correlation between entropy and both the cross-tissue expression heritability (R = 0.082, Fig 14A) and PVE (R = 0.12, Fig 14B) estimates, using the cross-tissue expression derived from the OTD. Thus, 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. Also, the correlation between tissue-specific OTD gene expression PVE and the posterior probability that the gene has an eQTL in that tissue is strongest in each respective tissue, confirming that OTD also captures tissue-specific components of gene expression (Fig 15).

Discussion

Because regulatory variation plays a key mechanistic role in the genetics of complex traits [1–3], we sought to comprehensively characterize the genetic architecture of gene expression across tissues. We accurately quantify the local heritability of gene expression in DGN whole blood and nine GTEx tissues. In DGN, the mean local h^2 was 0.13, similar to that found in family studies of blood expression, where mean h^2 ranged from 0.07-0.11 [11,12]. While we found that functional priors (known trans-eQTLs) can reduce the error of the estimate by reducing the number of genetic markers included in the genetic relationship matrix, larger

sample sizes (n > 1000) are needed to accurately estimate distal heritability.

Using the hybrid polygenic-sparse approach of BSLMM (Bayesian Sparse Linear Mixed Model) [18], we show that the local architecture of gene expression is sparse (high PGE) for most heritable genes in both DGN and GTEx. Using the elastic net [24], we observed improved cross-validated expression prediction for $\alpha \geq 0.5$ across tissues, confirming the sparsity result. This result demonstrates that sparse effects can be identified with sample sizes in the hundreds rather than the thousands and is supported by many prior studies with sample sizes near 100 that identified replicable eQTLs near the transcription start sites of genes[9]. Conversely, for traits that are highly polygenic, e.g. height, BMI, schizophrenia, and bipolar disorder, thousands to tens of thousands of samples are needed to identify significant genetic signals [26]. Therefore, the distal contributions to expression h^2 are likely to be more polygenic because they could not be accurately estimated here with sample sizes in the hundreds.

Our BSLMM analysis quantified the optimal number of SNPs to include for each gene. For example, the median number of SNPs for *ERAP2* was 1. When we previously plotted out-of-sample observed vs. predicted expression using elastic net ($\alpha = 0.5$) generated predictors for this gene, we saw three clusters, corresponding to each of the three genotypes for the causal variant [13]. Similarly, BSLMM estimated 2, 3, and 5 SNPs for *PEX6*, *NUDT2*, and *ERAP1*, respectively, consistent with the out-of-sample observed vs. predicted expression plots in our PrediXcan paper (see Figure 5 in [13]). Potentially due to variations in imputation quality of the input SNPs for expression prediction, it is useful to include more than the likely causal variants (i.e. elastic net rather than BSLMM) in the prediction for robustness. In addition, the elastic net is amenable cross-validation, while genome-wide cross-validation with BSLMM is impractical at current runtimes.

We developed a mixed effects model called orthogonal tissue decomposition (OTD) to determine the cross-tissue and tissue-specific components of gene expression in the GTEx dataset. Previous studys have shown that many eQTLs are shared across tissues [21,25]. In addition, because expression data from multiple tissues were available from the same individuals in GTEx, we could effectively use the multiple tissue samples as subject replicates in our OTD model. However, the tissue availability is unbalanced across individuals because of the difficulties of sample collection and the uneven quality of the tissues. By combining all available expression data in our OTD model, we found that estimates of the local heritability of cross-tissue gene

expression have larger magnitude and improved standard errors compared to single tissue estimates due to the borrowing of information across all samples. Thus, OTD effectively increases power to estimate heritability. Comparing our OTD results to a previously performed joint multi-tissue eQTL analysis method [25], 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 confirmed that OTD also captures tissue-specific components of gene expression by showing the correlation between tissue-specific OTD gene expression PVE and the posterior probability that the gene has an eQTL in that tissue is strongest in each respective tissue. Interestingly, whole blood and thyroid appear to be outliers in that they each have a negative correlation with all the other tissues (Fig 15). In the GTEx pilot analysis, whole blood had the lowest levels of eQTL sharing with other tissues and thyroid had the largest number of cis-eQTL genes, which implies a higher number of tissue-specific eQTLs [21].

In this paper, we quantitate the genetic architecture of gene expression and develop predictors across tissues. We show that local heritability can be accurately estimated across tissues, but distal heritability cannot be reliably estimated at current sample sizes. Using two different approaches, the elastic net and BSLMM, we show that for local gene regulation, the genetic architecture is mostly sparse rather than polygenic. Using new expression phenotypes generated in our OTD model, we show that cross-tissue predictive performance exceeded that of both tissue-specific and whole tissue expression as indicated by higher elastic net cross-validated R^2. Predictors generated in this study of gene expression architecture have been added to our PrediXcan database (https://github.com/hakyimlab/PrediXcan) [13] for use in future studies of complex trait genetics.

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 [20], 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. [20] 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) [27,28] 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 [21] 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 ($n = 279$), 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 distal 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 distal (eQTLs on non-gene chromosomes) as defined by the GENCODE [29] version 12 gene annotation. We calculated the proportion of the variance (narrow-sense heritability) explained by each component using the following mixed-effects model:

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

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 distal components by estimating σ_w^2 with restricted maximum likelihood (REML) using GCTA software [22]. For heritability analyses in the GTEx cohort, we removed the *distal* 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.

Determining polygenicity versus sparsity using the elastic net

We applied the elastic net [24] 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` [30,31] 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/

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

over a grid of values of λ covering the entire range [30,31]. 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) [14] at one extreme and ridge regression ($\alpha = 0$) [15] 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 whole tissue, cross-tissue, and tissue-specific expression in the GTEx cohort, we tested $\alpha = 0.05, 0.5, 0.95, 1$.

Quantifying sparsity with Bayesian Sparse Linear Mixed Models (BSLMM)

We used BSLMM [18] to model the effect of local genetic variation (SNPs within 1 Mb of gene) on the genetic architecture of gene expression. The BSLMM consists of a standard linear mixed model, with one random effect term, and with sparsity inducing priors on the regression coefficients [18]. We used the software GEMMA [32] to implement BSLMM for each gene using the following parameters:

```
gemma -g [localGenoFile] -p [geneExpFile] -a [snpAnnotFile] -bslmm 1 -s 100000 -o [outFile]
```

The `-bslmm 1` option specifies a linear BSLMM and the `-s 100000` option specifies the number of sampling steps per gene. The BSLMM estimates the PVE (the total proportion of variance in phenotype explained by

the sparse effects and random effects terms together) and PGE (the proportion of genetic variance explained by the sparse effects terms). From the second half of the sampling iterations for each gene, we report the median and the 95% credible sets of the PVE, PGE, and the $|\gamma|$ parameter (the number of SNPs with non-zero coefficients).

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 [33]. We applied OTD to GTEx Project [21] 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 [34] package `lme4` [35,36] to fit the following mixed-effects model:

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

The model included whole tissue gene expression levels in 8555 GTEx tissue samples from 544 unique subjects. A total of 17,647 Protein-coding genes (defined by GENCODE [29] 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 whole tissue expression levels (`expression` in the model). `PEERs` included the top 15 PEER factors estimated across all tissues using the R package `PEER` [37] 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)`).

Comparison of OTD PVE to multi-tissue eQTL results

Using results from a joint multi-tissue eQTL analysis method [25] performed with a subset of the GTEx data (maximum n=175 in the nine tissues of the pilot phase, see [21]), we defined an entropy statistic to compare these results to those from our OTD method. The results of the multi-tissue analysis include eQTL posterior probabilities for each of the nine tissues, which can be interpreted as the probability a SNP is an eQTL in tissue t given the data. Using the top eQTL for each gene g, we defined the entropy S_g as:

$$S_g = - \sum_t p_{t,g} \log p_{t,g}$$

where $p_{t,g}$ is the eQTL probability in tissue t normalized to 1 for each gene g. Thus, eQTLs with higher entropy statistics are more likely to be cross-tissue eQTLs, rather than only regulating gene expression in one or a few tissues. We calculated the Pearson correlation between S_g and the cross-tissue expression heritability and PVE for each gene to verify that our OTD method captures cross-tissue effects. We also calculated a Pearson correlation matrix between the posterior probabilities in each tissue from the multi-tissue eQTL method and the tissue-specific gene expression PVE from the OTD method.

Figures

Supplemental Figures

Acknowlededgments
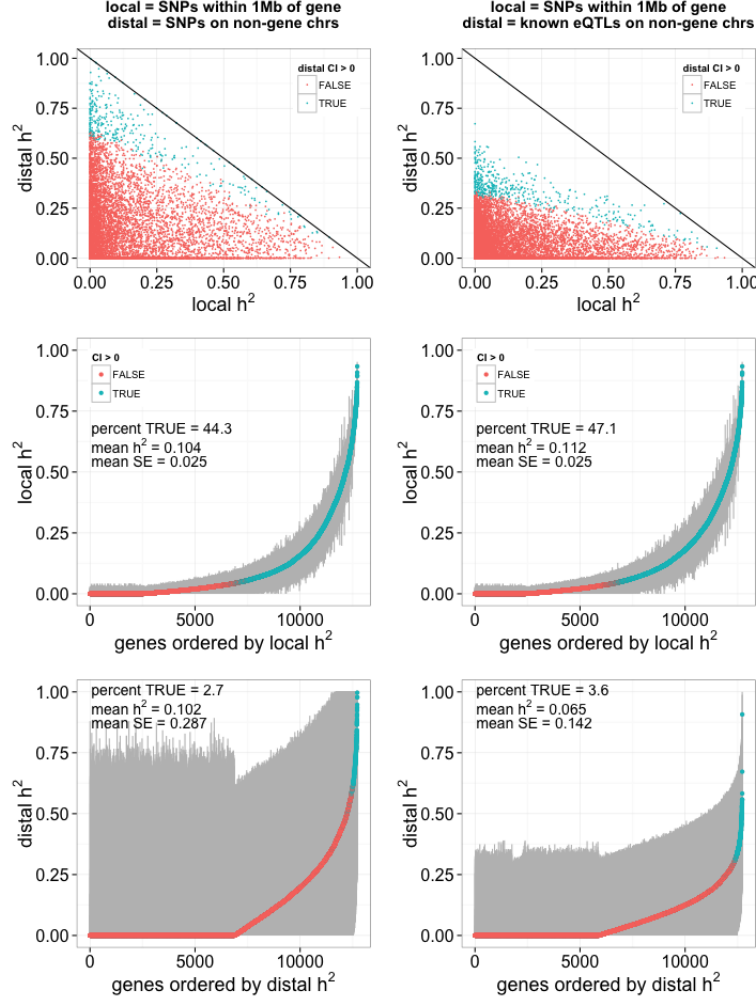
Figure 1: DGN whole blood expression joint heritability (h^2). Local (SNPs within 1 Mb of each gene) and distal (Left: SNPs on non-gene chromsomes. Right: SNPs that are eQTLs in the Framingham Heart Study on other chromosomes [FDR < 0.05]) h^2 for gene expression were jointly estimated. (**Top**) Distal h^2 compared to local h^2 per gene in each model. (**Middle**) Local and (**Bottom**) 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.
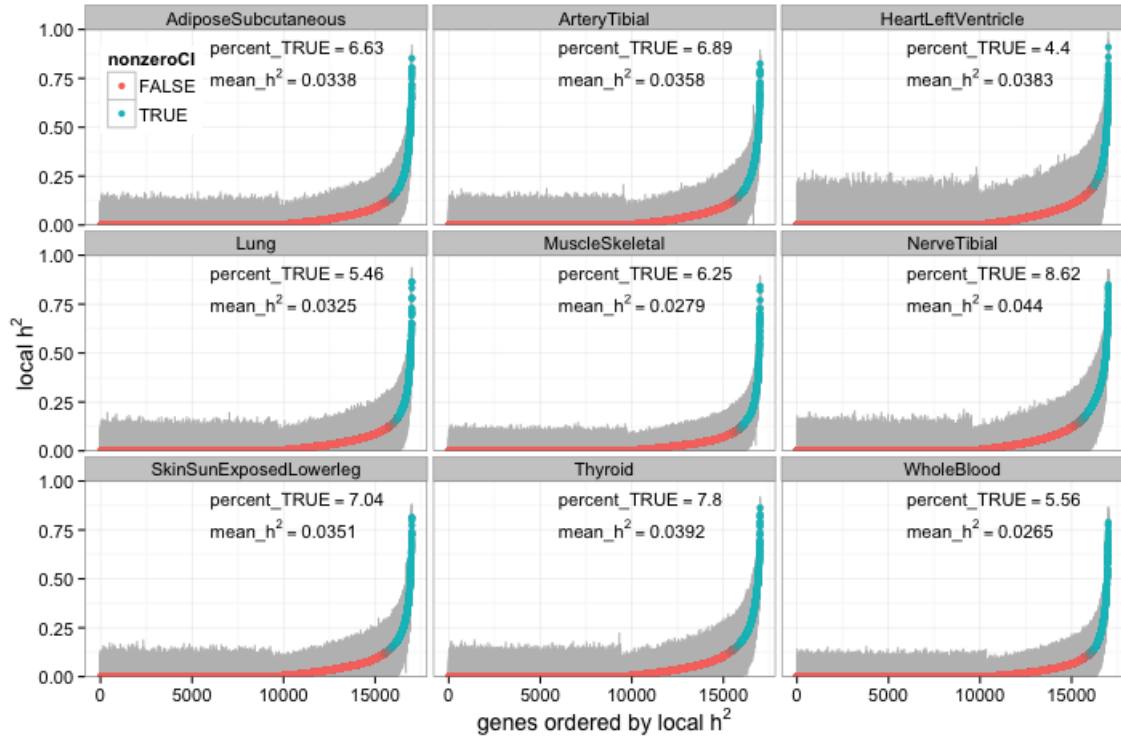
Figure 2: GTEx whole tissue local heritability (h^2) estimation. Local (SNPs within 1 Mb of each gene) gene expression h^2 estimates ordered by increasing h^2. The 95% confidence interval (CI) of each h^2 estimate is in gray and genes with a lower bound greater than zero are in blue
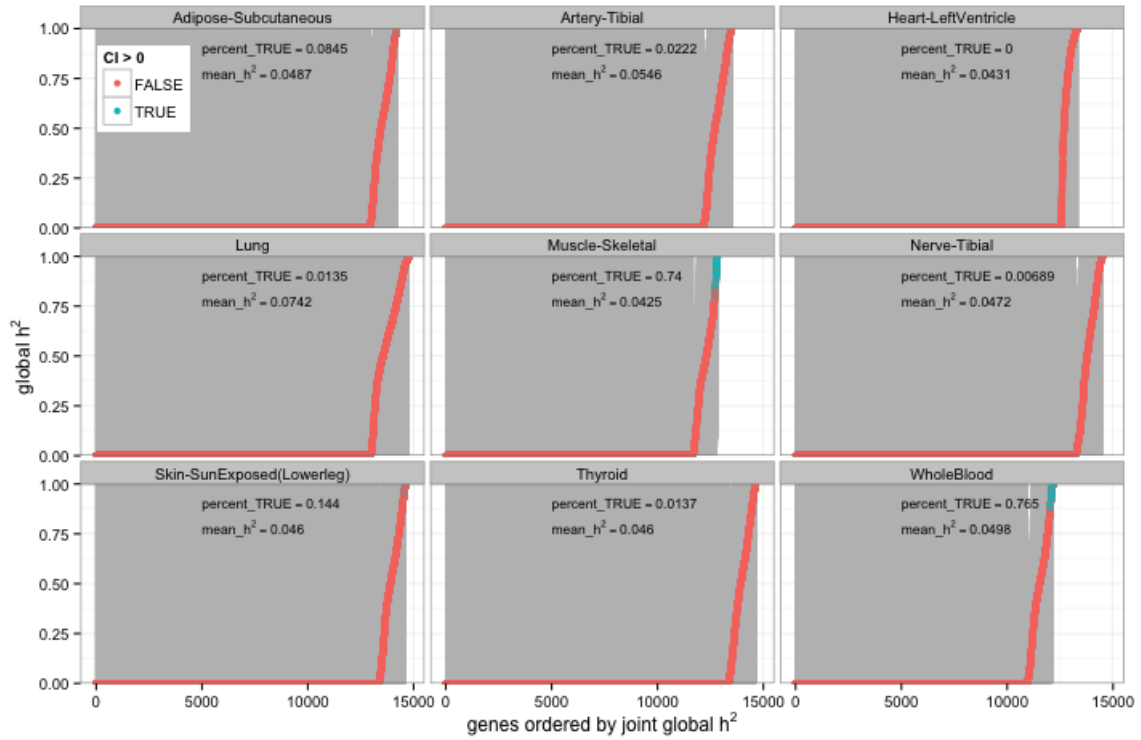
Figure 3: GTEx whole tissue distal heritability (h^2) estimation. Distal (SNPs that are eQTLs in the Framingham Heart Study on other chromosomes [FDR < 0.05]) gene expression h^2 estimates from a joint model are ordered by increasing h^2. The 95% confidence interval (CI) of each h^2 estimate is in gray and genes with a lower bound greater than zero are in blue
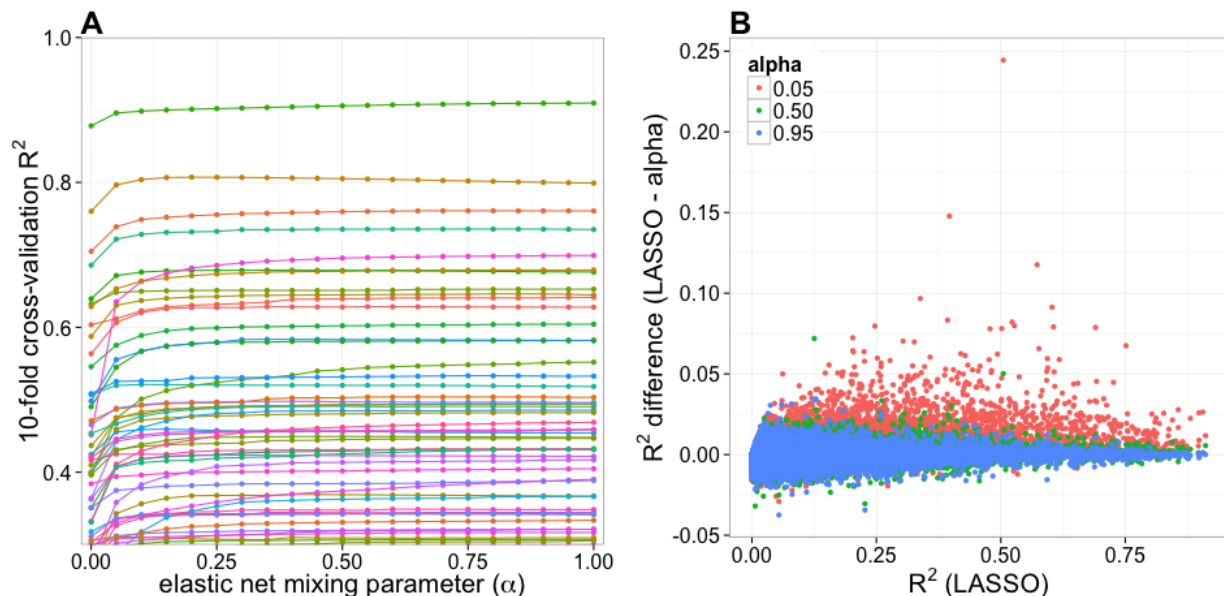
Figure 4: DGN cross-validated predictive performance across the elastic net. (**A**) 10-fold cross-validated R^2 of predicted vs. observed expression in DGN whole blood compared to a range of elastic net mixing parameters (α) for genes on chromosome 22 with $R^2 > 0.3$. (**B**) Predictive R^2 difference between LASSO ($\alpha = 1$) and several other values of α compared to LASSO predictive R^2 for 13171 autosomal genes.

Grants

GTEx data

Figure 5: GTEx whole tissue cross-validated predictive performance across the elastic net. Predictive R^2 difference between LASSO ($\alpha = 1$) and several other values of α compared to LASSO predictive R^2 for 18473 autosomal genes per tissue.
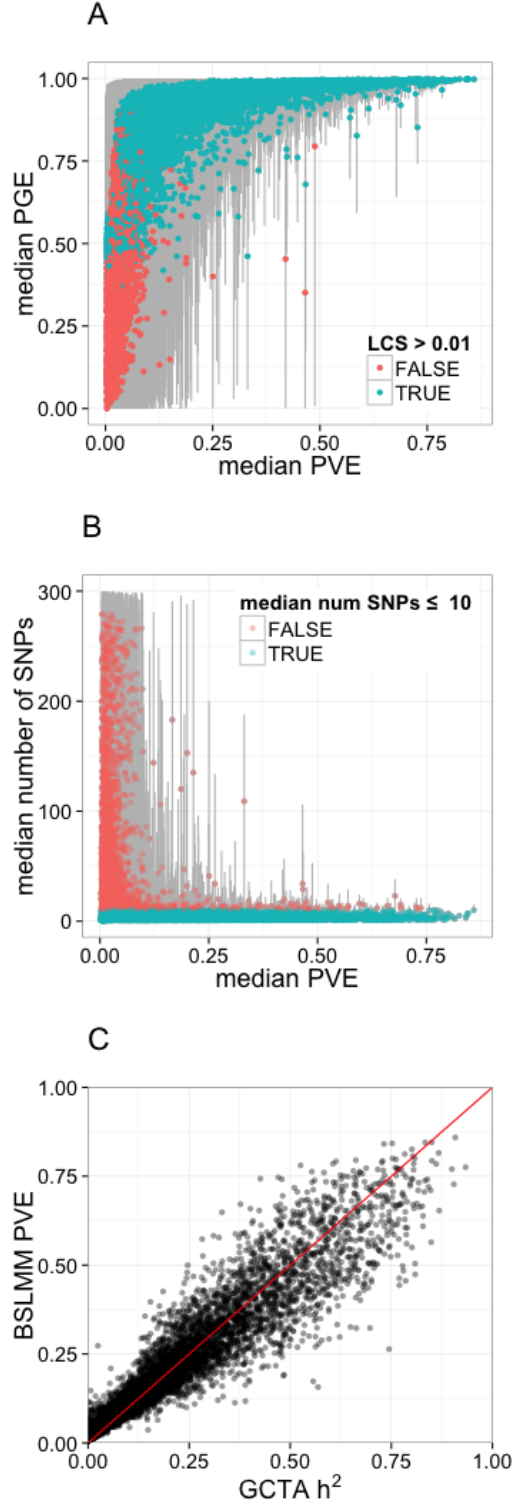
Figure 6: Bayesian Sparse Linear Mixed Models reveal the sparsity of gene expression architecture. (**A**) BSLMM-estimated PVE (total proportion of variance expla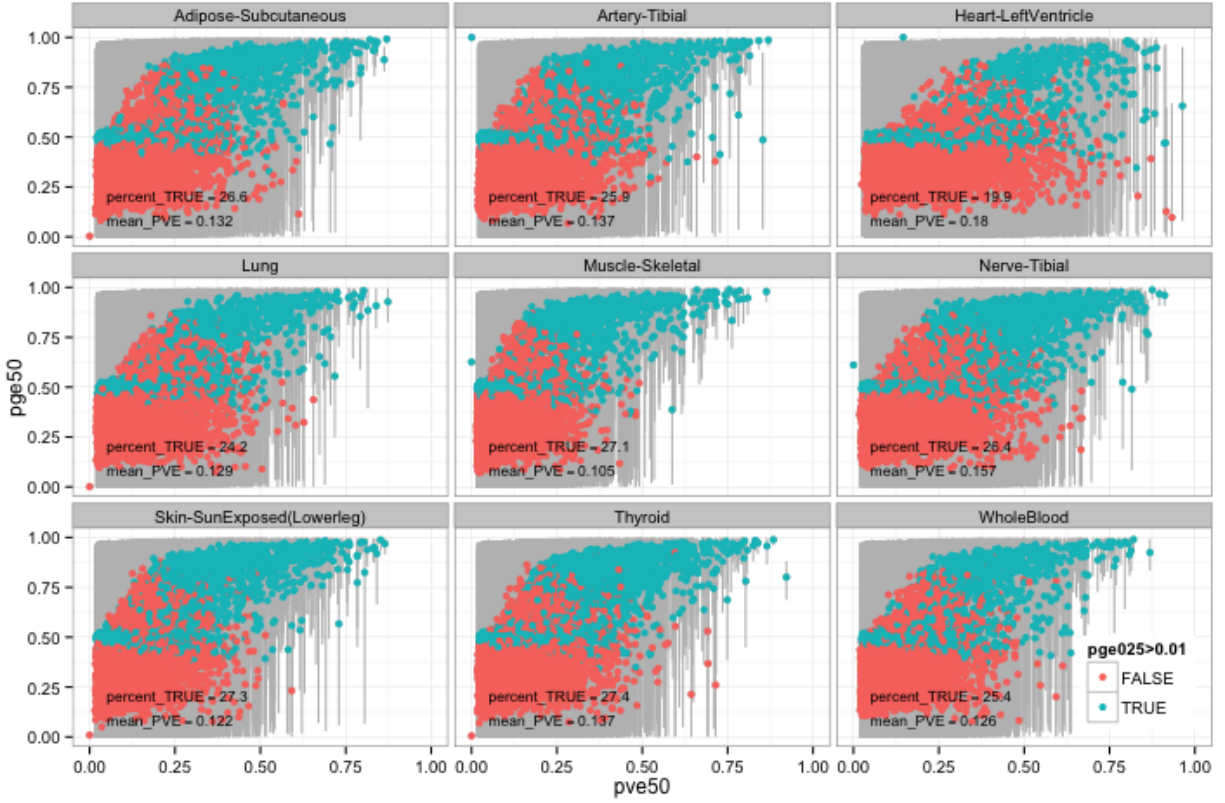ined) compared to GCTA-estimated heritability per gene (R=0.96) (**B**) Comparison of median PGE (proportion of PVE explained by sparse effects) to median PVE (total proportion of variance explained) for expression of each gene. The 95% credible set of each PGE estimate is in gray and genes with a lower 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.
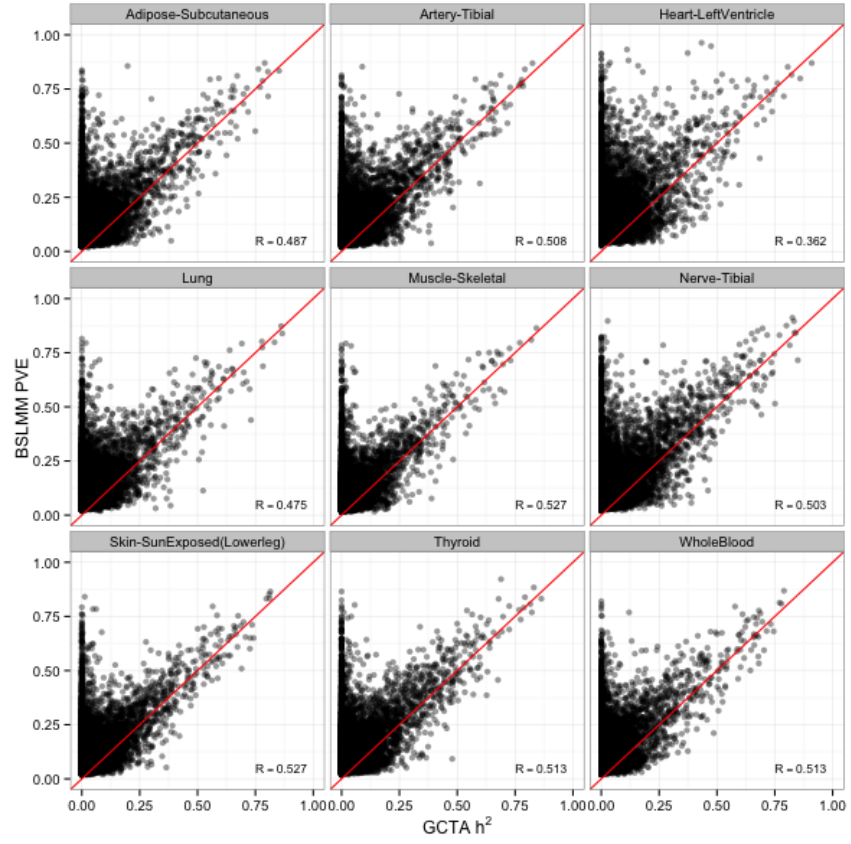
Figure 7: GTEx whole tissue expression BSLMM-estimated PVE (total proportion of variance explained) compared to GCTA-estimated heritability per gene. R = Pearson correlation.

Figure 8: GTEx whole tissue expression 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.
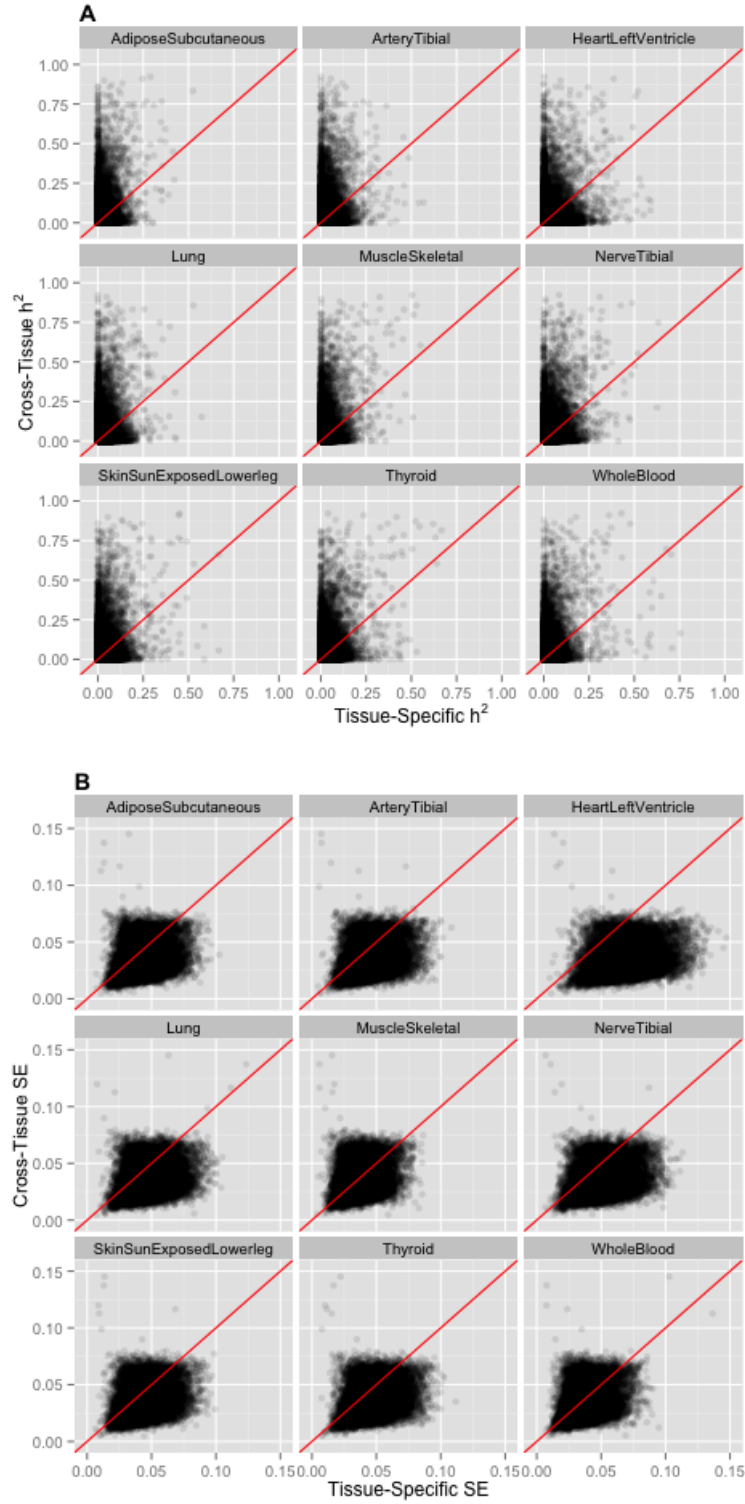
Figure 9: Cross-tissue and tissue-specific comparison of heritability (h^2, **A**) and standard error (SE, **B**) estimation. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-specifc 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.
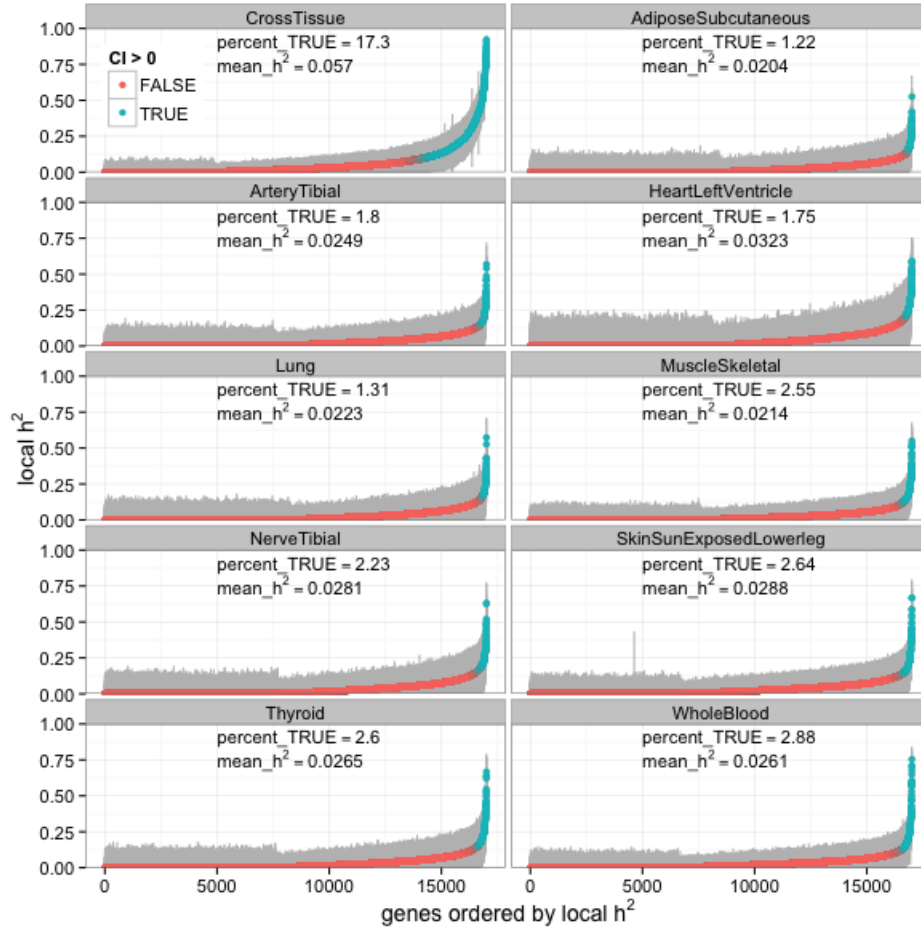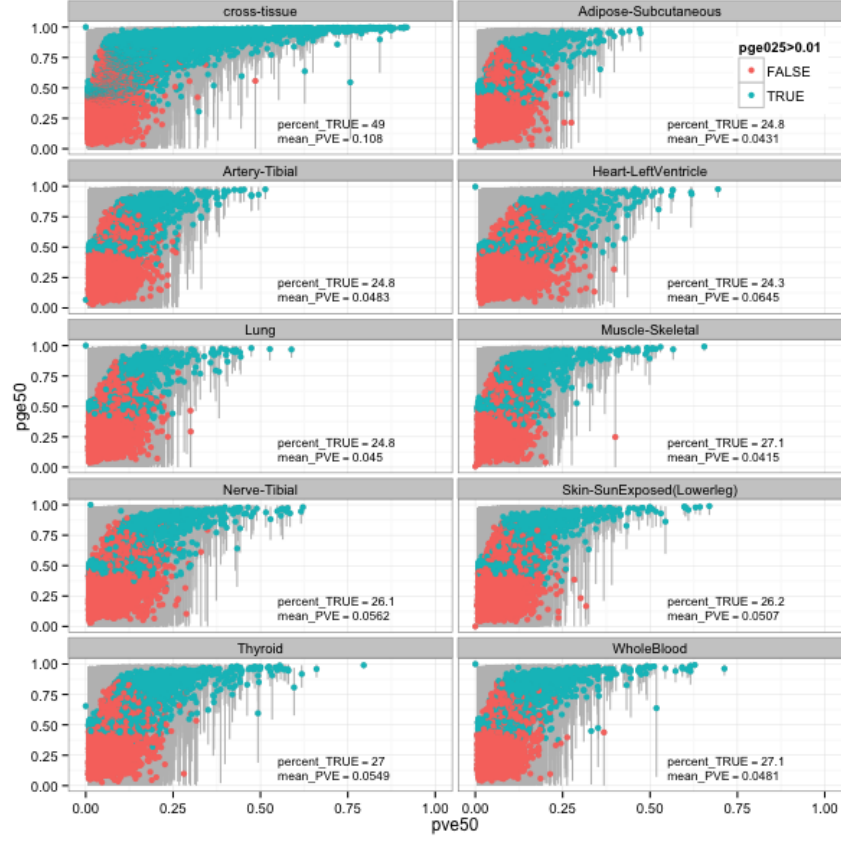
Figure 10: Cross-tissue heritability (h^2) compared to tissue-specific h^2. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-specifc local h^2 is estimated using the tissue-specific component (residuals) of the mixed effects model for gene expression for each respective tissue and SNPs within 1 Mb of each gene.

Figure 11: GTEx orthogonal tissue decomposition cross-tissue and tissue-specific expression 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.
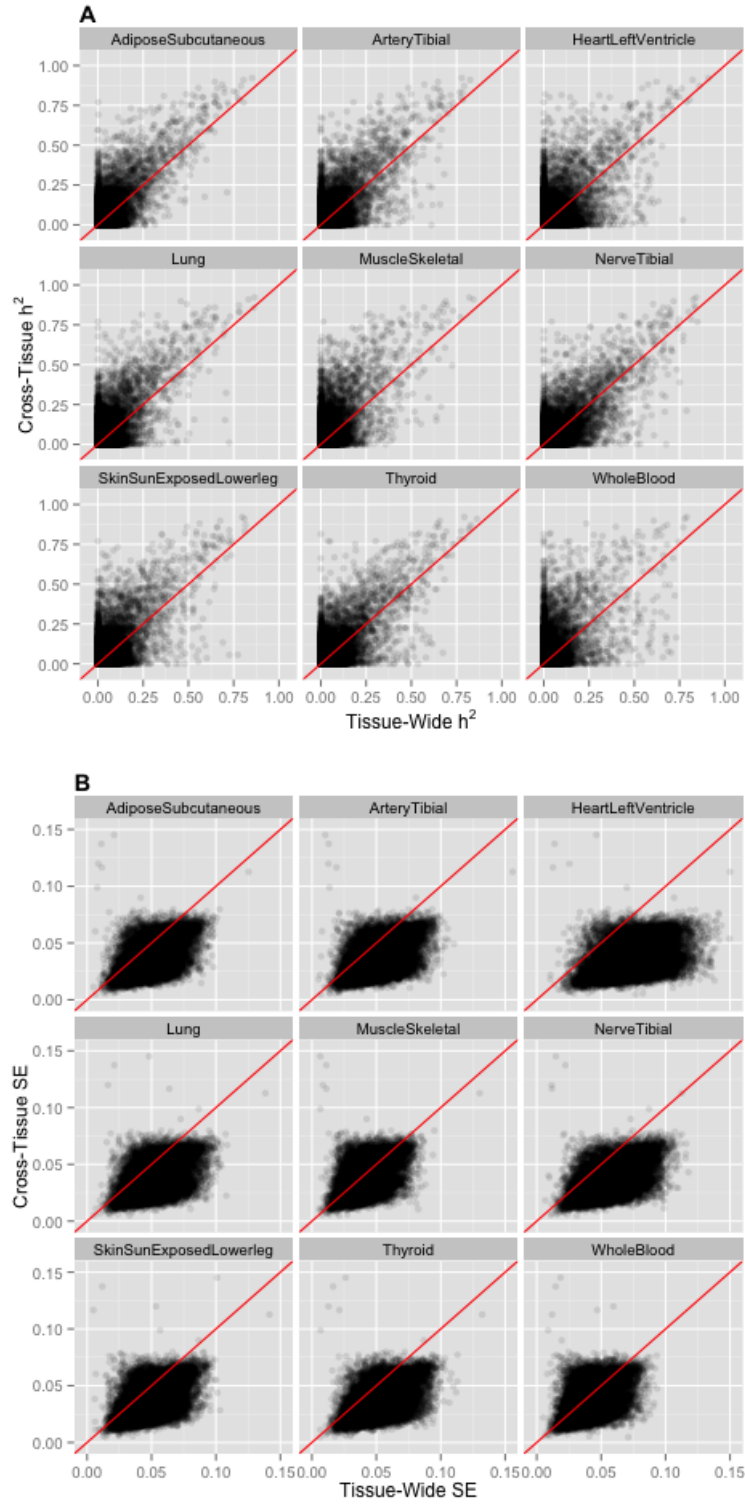
Figure 12: Cross-tissue and whole tissue comparison of heritability (h^2, **A**) and standard error (SE, **B**). Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Whole tissue local h^2 is estimated using the measured gene expression for each respective tissue and SNPs within 1 Mb of each gene.
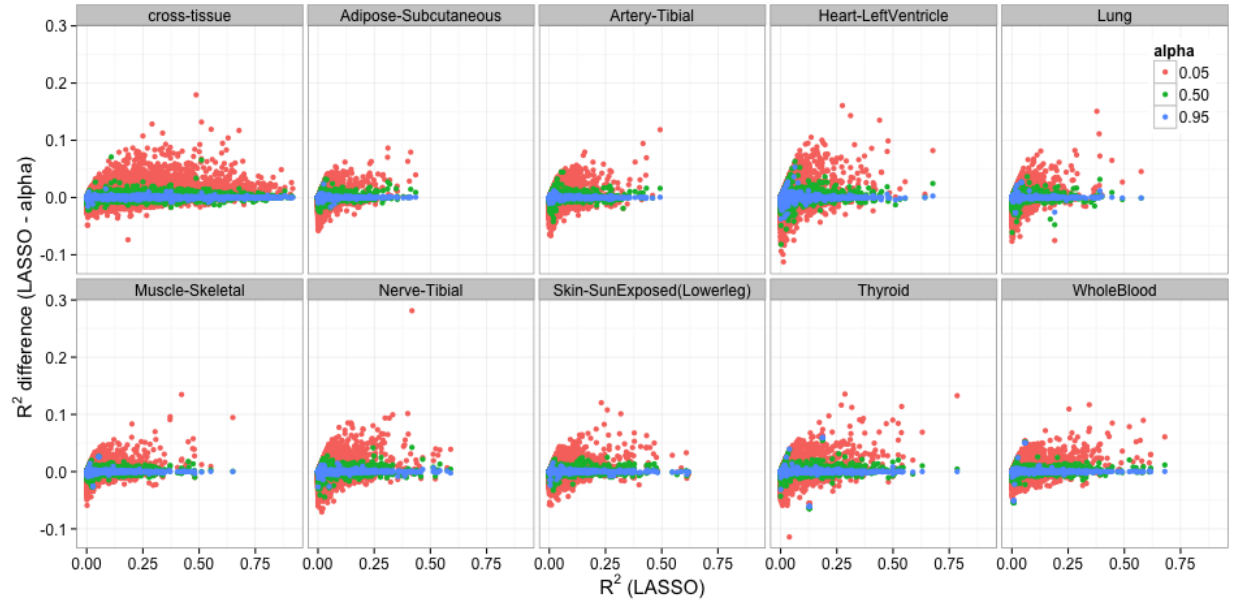
Figure 13: GTEx orthogonal tissue decomposition cross-tissue and tissue-specific expression cross-validated predictive performance across the elastic net. Predictive R^2 difference between LASSO ($\alpha = 1$) and several other values of α compared to LASSO predictive R^2 for 18473 autosomal genes per tissue.
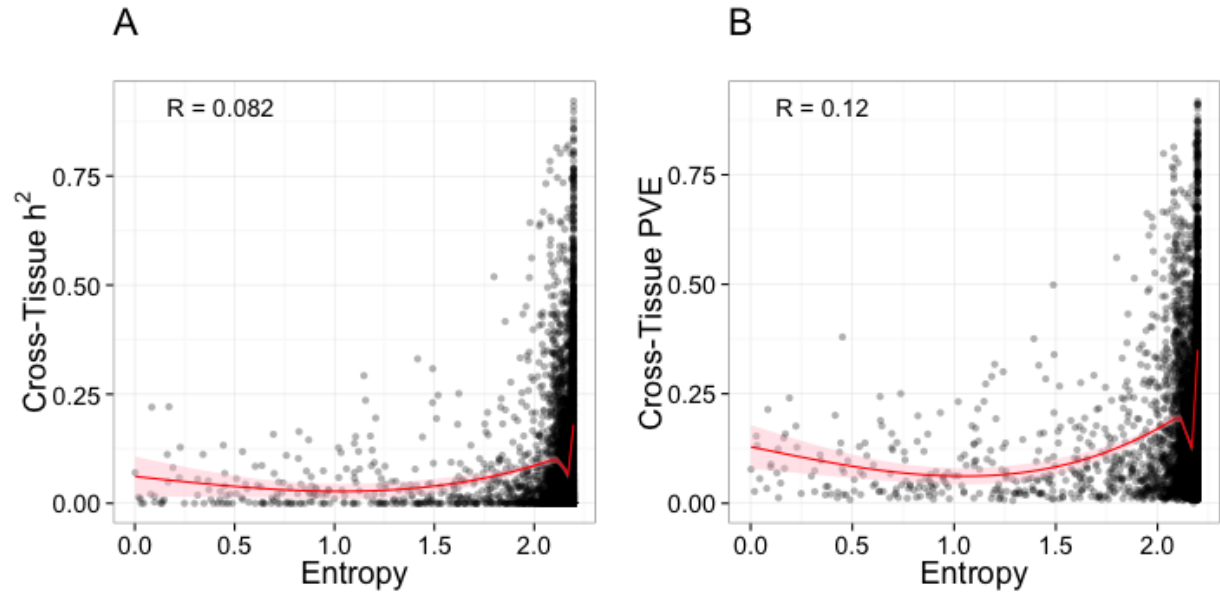

Figure 14: Entropy of the posterior probabilities from the Flutre et al. multi-tissue eQTL method compared to the estimates of (**A**) heritability and (**B**) PVE of cross-tissue gene expression derived from the orthogonal tissue decomposition. The generalized additive model smoothing line is in red.
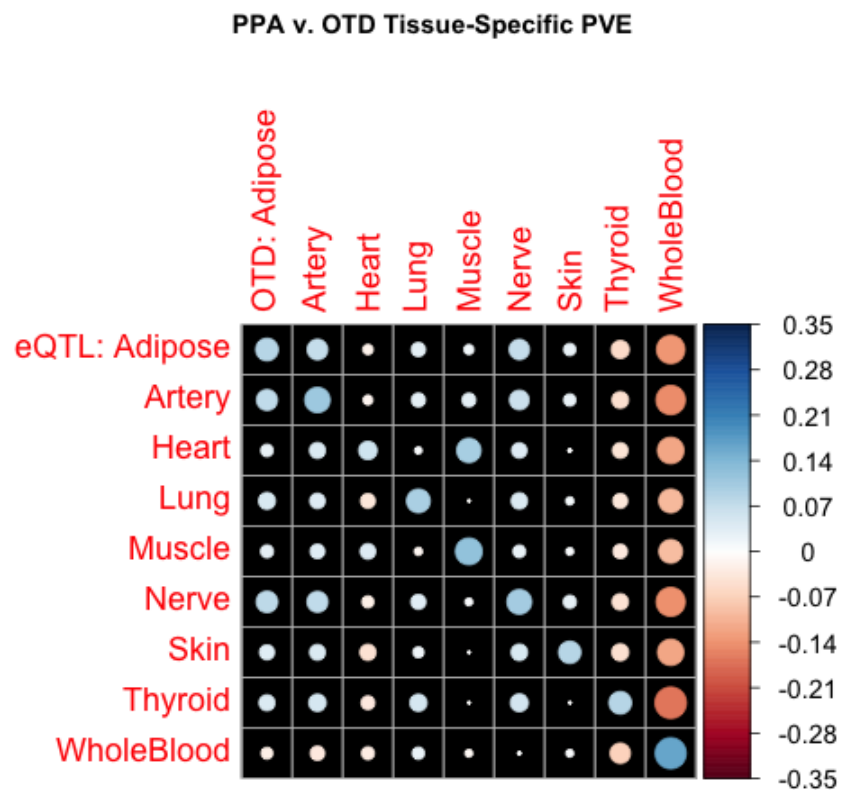
Figure 15: Pearson correlation (R) between the posterior probability the top multi-tissue eQTL regulates its gene in a given tissue (eQTL, Flutre et al. method) and the PVE of tissue-specific gene expression from the orthogonal tissue decomposition (OTD). Area of each circle is proportional to the absolute value of R.

Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc.(HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951,MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819),Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v3.p1.

DGN data

NIMH Study 7 (GenRED I) - Data and biomaterials were collected in six projects that participated in the National Institute of Mental Health (NIMH) Genetics of Recurrent Early-Onset Depression (GenRED) project. From 1999-2003, the Principal Investigators and Co-Investigators were: New York State Psychiatric Institute, New York, NY, R01 MH060912, Myrna M. Weissman, Ph.D. and James K. Knowles, M.D., Ph.D.; University of Pittsburgh, Pittsburgh, PA, R01 MH060866, George S. Zubenko, M.D., Ph.D. and Wendy N. Zubenko, Ed.D., R.N., C.S.; Johns Hopkins University, Baltimore, MD, R01 MH059552, J. Raymond DePaulo, M.D., Melvin G. McInnis, M.D. and Dean MacKinnon, M.D.; University of Pennsylvania, Philadelphia, PA, RO1 MH61686, Douglas F. Levinson, M.D. (GenRED coordinator), Madeleine M. Gladis, Ph.D., Kathleen MurphyEberenz, Ph.D. and Peter Holmans, Ph.D. (University of Wales College of Medicine); this preprint is the author/funder. It is made available under a CC-BY 4.0 International license. bioRxiv preprint first posted online June 17, 2015; doi: http://dx.doi.org/10.1101/020164; The copyright holder for University of Iowa, Iowa City, IW, R01 MH059542, Raymond R. Crowe, M.D. and William H. Coryell, M.D.; Rush University Medical Center, Chicago, IL, R01 MH059541- 05, William A. Scheftner, M.D., Rush-Presbyterian. NIMH Study 18 - Data and biomaterials were obtained from the limited access datasets distributed from

the NIH-supported "Sequenced Treatment Alternatives to Relieve Depression" (STAR*D). STAR*D focused on non-psychotic major depressive disorder in adults seen in outpatient settings. The primary purpose of this research study was to determine which treatments work best if the first treatment with medication does not produce an acceptable response. The study was supported by NIMH Contract # N01MH90003 to the University of Texas Southwestern Medical Center. The ClinicalTrials.gov identifier is NCT00021528. NIMH Study 52 (GenRED II) – Data and biomaterials in this release were collected in six projects that participated in the National Institute of Mental Health (NIMH) Genetics of Recurrent Early-Onset Depression (GenRED) project (1999-2009). The Principal Investigators and Co-Investigators were: New York State Psychiatric Institute, New York, NY, R01 MH 060912, Myrna M. Weissman, Ph.D.; Johns Hopkins University, Baltimore, MD, R01 MH059552, J. Raymond DePaulo, M.D., and James B. Potash, M.D., M.P.H.; University of Pennsylvania, Philadelphia, PA (1999-2005), and Stanford University (2006-2009), R01 MH61686, Douglas F. Levinson, M.D. (GenRED coordinator); University of Iowa, Iowa City, IW, R01 MH059542e, Raymond R. Crowe, M.D., and William H. Coryell, M.D.; Rush University Medical Center, Chicago, IL, R01 MH059541-05, William A. Scheftner, M.D.; and University of Pittsburgh, Pittsburgh, PA (1999-2003), R01 MH060866, George S. Zubenko, M.D., Ph.D., and Wendy N. Zubenko, Ed.D., R.N., C.S. NIMH Study 88 – Data was provided by Dr. Douglas F. Levinson. We gratefully acknowledge the resources were supported by National Institutes of Health/National Institute of Mental Health grants 5RC2MH089916 (PI: Douglas F. Levinson, M.D.; Coinvestigators: Myrna M. Weissman, Ph.D., James B. Potash, M.D., MPH, Daphne Koller, Ph.D., and Alexander E. Urban, Ph.D.) and 3R01MH090941 (Co-investigator: Daphne Koller, Ph.D.).

Computing resources

This work made use of the Open Science Data Cloud (OSDC) which is an Open Cloud Consortium (OCC)-sponsored project. This work was supported in part by grants from Gordon and Betty Moore Foundation and the National Science Foundation and major contributions from OCC members like the University of Chicago.

W, White KP, Atkinson M, Klampanos I, Alvarez H, Harvey C and Mambretti J, The Design of a Community Science Cloud: The Open Science Data Cloud Perspective. (2012) doi:10.1109/SC.Companion.2012.127 This work made use of the Bionimbus Protected Data Cloud (PDC), which is a collaboration between the Open Science Data Cloud (OSDC) and the IGSB (IGSB), the Center for Research Informatics (CRI), the Institute for Translational Medicine (ITM), and the University of Chicago Comprehensive Cancer Center (UCCCC). The Bionimbus PDC is part of the OSDC ecosystem and is funded as a pilot project by the NIH. https://www.bionimbus-pdc.opensciencedatacloud.org/ Heath AP, Greenway M, Powell R, Spring J, Suarez R, Hanley D, Bandlamudi C, McNerney ME, White KP and Grossman RL, Bionimbus: A Cloud for Managing, Analyzing and Sharing Large Genomics Datasets. J Am Med Inform Assoc (2014) doi:10.1136/amiajnl-2013-002155

References

1. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. Gibson G, editor. PLoS Genetics. Public Library of Science (PLoS); 2010;6: e1000888. doi:10.1371/journal.pgen.1000888

2. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. Gibson G, editor. PLoS Genetics. Public Library of Science (PLoS); 2010;6: e1000895. doi:10.1371/journal.pgen.1000895

3. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. The American Journal of Human Genetics. Elsevier BV; 2014;95: 535–552. doi:10.1016/j.ajhg.2014.10.004

4. Purcell SM, Wray NR, Stone JL, Visscher PM, Sullivan MCOPF, Sklar P, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. Nature Publishing Group; 2009; doi:10.1038/nature08185

5. Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, et al. Bayesian inference analyses

of the polygenic architecture of rheumatoid arthritis. Nature Genetics. Nature Publishing Group; 2012;44: 483–489. doi:10.1038/ng.2232

6. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics. Nature Publishing Group; 2012;44: 981–990. doi:10.1038/ng.2383

7. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. Nature Publishing Group; 2015;16: 197–212. doi:10.1038/nrg3891

8. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of cis regulatory variation in diverse human populations. Barsh GS, editor. PLoS Genetics. Public Library of Science (PLoS); 2012;8: e1002639. doi:10.1371/journal.pgen.1002639

9. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nature Genetics. Nature Publishing Group; 2007;39: 1217–1224. doi:10.1038/ng2142

10. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. Storey JD, editor. PLoS Genetics. Public Library of Science (PLoS); 2011;7: e1002078. doi:10.1371/journal.pgen.1002078

11. Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, et al. Heritability and genomics of gene expression in peripheral blood. Nature Genetics. Nature Publishing Group; 2014;46: 430–437. doi:10.1038/ng.2951

12. Price AL, Helgason A, Thorleifsson G, McCarroll SA, Kong A, Stefansson K. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. Gibson G, editor. PLoS Genetics. Public Library of Science (PLoS); 2011;7: e1001317. doi:10.1371/journal.pgen.1001317

13. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics. Nature Publishing Group; 2015;47: 1091–1098. doi:10.1038/ng.3367

14. Regression shrinkage and selection via the lasso on jSTOR [Internet]. http://www.jstor.org/stable/2346178; 2016. Available: http://www.jstor.org/stable/2346178

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

16. de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: The promise of whole-genome markers. Nat Rev Genet. Nature Publishing Group; 2010;11: 880–886. doi:10.1038/nrg2898

17. Wheeler HE, Aquino-Michaels K, Gamazon ER, Trubetskoy VV, Dolan ME, Huang RS, et al. Poly-omic prediction of complex traits: OmicKriging. Genetic Epidemiology. Wiley-Blackwell; 2014;38: 402–415. doi:10.1002/gepi.21808

18. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. Visscher PM, editor. PLoS Genetics. Public Library of Science (PLoS); 2013;9: e1003264. doi:10.1371/journal.pgen.1003264

19. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature. Nature Publishing Group; 2005;437: 1365–1369. doi:10.1038/nature04244

20. 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

21. 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

22. 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

23. 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. Nature Genetics. Nature Publishing Group; 2015;47: 345–352. doi:10.1038/ng.3220

24. 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

25. Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eQTL analysis in multiple tissues. Gibson G, editor. PLoS Genetics. Public Library of Science (PLoS); 2013;9: e1003486. doi:10.1371/journal.pgen.1003486

26. Hemani G, Yang J, Vinkhuyzen A, Powell JE, Willemsen G, Hottenga J-J, et al. Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. The American Journal of Human Genetics. Elsevier BV; 2013;93: 865–875. doi:10.1016/j.ajhg.2013.10.005

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

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

29. 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

30. 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/

31. 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/

32. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. Nature Publishing Group; 2012;44: 821–824. doi:10.1038/ng.2310

33. 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

34. 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/

35. Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4 [Internet]. 2015. Available: http://CRAN.R-project.org/package=lme4

36. Bates D, Maechler M, Bolker BM, Walker S. Fitting linear mixed-effects models using lme4 [Internet]. 2015. Available: http://arxiv.org/abs/1406.5823

37. 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. Nat Protoc. Nature Publishing Group; 2012;7: 500–507. doi:10.1038/nprot.2011.457