Survey of Heritability and Sparsity of Gene Expression Traits across a Comprehensive Set of Human Tissues

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 2015-12-02 15:29:45

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. Thus, understanding of the genetic architecture of gene expression traits is key to elucidate the underlying mechanisms of complex traits. However, systematic survey of the heritability and the distribution of effect sizes across all representative tissues in the human body is not available.

Here we take advantage of the RNAseq data on a comprehensive set of tissue samples generated by the GTEx Consortium to fill this gap. We find that local h2 can be well characterized with xx% of significant h2. However, the current sample sizes of (<400?) only allow us to compute distant h2 for a handful of genes (xx% range). Thus we focus on local regulation. Bayesian Sparse Linear Mixed Model (BSLMM) analysis and the sparsity of optimal performing predictors provided compelling evidence that local architecture of gene expression traits is sparse rather than polygenic across all xx tissues examined.

To further delve into the tissue context specificity, we decompose the expression traits into cross tissue and tissue specific components. Heritability and sparsity estimates of these derived expression phenotypes

show similar characteristics to the original traits. Consistent properties relative to prior GTEx multi tissue analysis results suggests that these traits reflect the expected biology.

Finally, we apply this knowledge to develop prediction models of gene expression traits for all tissues. The prediction models, heritability, and prediction performance R^2 for original and (OTD-) derived phenotypes are made publicly available (https://github.com/hakyimlab/PrediXcan).

Introduction

Regulatory variation plays a key role in the genetics of complex traits [1–3]. Methods that partition the contribution of environment and geneitc components useful tools to understand the biology underlying complex traits. Partitioning heritability to different functional classes have been successful in quantifying the contribution of different mechanisms that drive the etiology of diseases.

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].

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].

The relative performance of sparse and polygenic models can provide useful information about the underlying distribution of effect sizes. For example, if the true model of a trait is polygenic, it is natural to expect that polygenic models will perform better than sparse ones. 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]. When we calibrated our prediction model that was used in our PrediXcan paper, we showed that sparse models such as LASSO performed better than a polygenic score model. We also showed that a model that uses the top eQTL variant outperformed the polygenic score but did not do as well as LASSO or elastic net.[13], suggesting that for many genes, the genetic architecture is sparse, but not regulated by a single SNP.

a bit redundant now 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]. Although more recently, studies with a few other tissues have been published, a comprehensive coverage of human tissues was not available until the launching of the Genotype-Tissue Expression (GTEx) Project. GTEx aims to examine the genetics of gene expression more comprehensively and has 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]. Here we use a much larger set of 8xxx samples corresponding to up to xxx individuals from the 2015 release. One of the findings of this comprehensive analysis was that a large portion of the local regulation of expression traits is shared across multiple tissues. This was corroborated by the fact that our prediction model based on whole blood showed good prediction across the 9 core GTEx tissues chosen by initial sample size [13].

This shared regulation implies that there is much to be learned from large sample studies of easily accessible tissues. Still a portion of the regulation seems to be tissue dependent. this may need rewording

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 use a mixed effects model called orthogonal tissue decom-

position (OTD) to decouple the cross-tissue and tissue-specific mechanisms 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

local and distal h2 across tissues

We estimated the local and distal heritability of gene expression levels in 40 tissues from the GTEx consortium and whole blood from the Depression Genes and Networks (DGN) cohort. The sample size in GTEx varied from xxx to xxx depending on the tissue while the DGN had 922 samples [20]. We used mixed-effects model (see Methods) and calculated variances using restricted maximum likelihood as implemented in GCTA [22].

For the local component, we used variants within 1Mb of the TSS and TSE of each protein coding gene whereas for the distal component we used variants outside of the chromosome where the gene was located. Different approaches to compute the distal genetic relatedness were explored but results did not change substantively. See more details in Methods.

The left column of Figure 1 shows the estimated local and distal h2 from DGN. Even though many genes show relatively large point estimates of distal h2, only the ones colored in blue are significantly different from zero. The local component of h2 is relatively well estimated in DGN with 52.6% of genes (5631 out of 10704) showing h2 values significantly different from 0. In contrast, the distal heritability is significantly different from 0 for only 3.2% (343 out of 10704) of the genes.

TODO: generate table 1

Table 1 summarizes the heritability estimate results across all tissues. It shows the mean local h2, its standard error, proportion of significant h2, number of genes with significant heritability, and number of expressed genes would it be too much work to generate this # of expressed genes? we could skip if so (definition? more than xx individuals with xx RPKM) - is this in Methods.

In order to obtain an unbiased estimates of mean h2, we allow the values to be negative when fitting the REML, as done in Price et al and Wright et al. This approach reduces the standard error of the estimated mean of heritability, especially important for the distal component. Even though invididual gene'sdistal heritability is noisy, the average across all genes reduces the error substantially. For the DGN dataset, we were able to estimate the mean distal h2 with enough accuracy. However for the GTEx samples, the sample size was too small and the REML algorithm became instable when allowing for negative values. This numeric instability would cause a small number of genes with large positive (and noisy) heritability values to converge biasing the mean value. For this reason we do not show mean distal heritability estimates for GTEx tissues.

In DGN (whole blood), the mean local h2 was 14.3% and the distal mean was 3.4% such that the local variation contribution is estimated as 14.3/(3.4+14.3) = 80%. This is much higher than the 37% reported by Price et al [] based on blood expression data for a cohort of Icelandic individuals. This potentially under-estimation of the distal component could be due to over-correction of confounders used in the preprocessing of the expression trait data we used. Indeed, PEER cite, SVA cite, and other types of hidden confounder corrections have been shown to increase local eqtl replicability but there are concerns about the detrimental effects on the distal eqtl identification. As larger sample sizes become available we will test this hypothesis in GTEx data by computing the distal h2 without PEER factor correction.

Since it has been shown that local-eQTLs are more likely to be distal-eQTL of target genes, we tested whether restricting the distal genetic similarity computation to cis-eQTLs (as determined in Framingham mRNA dataset ib over 5000 individuals independent of the DGN and GTEx cohort) for other genes could improve distal heritability precision by prioritizing functional variants. We exclude eQTLs in the same chr as the tested gene to avoid contaminating distal h2 with cis associations.

Using functional priors (known eQTLs) to define distal h^2 increased the percentage of genes with a positive CI from 3.2% (343 genes) to 4.2% (457) in whole blood (Fig 1). In GTEx tissues the increase ranged from xxx% (xx genes) to xxx% (xx genes). 125 genes have significant distal h2 by both approaches, i.e. all variants in other chr or only cis-eQTL variants in other chr.

However, using the subset of local eqtls (from an independent source) in other chromosomes for computing

distal heritability reduced the mean value from 3.4% to 1.6% indicating that even thought we gain some power to detect significant distal heritability but a good portion of the distal regulation is lost when using only the smaller subset potentially more functional variants.

Given the limited sample size we will focus on local regulation for the remainder of the paper.

Sparse local architecture implied by sparsity of best prediction models

Next, we sought to determine whether local genetic contribution to gene expression trait was polygenic or sparse. In other words, whether many variants of small effects or a small number of large effects were contributing to the trait variability. For this, we first looked at the prediction performance of a range of models with different degree of polygenicity, such as the elastic net model with different mixing parameter values that range from 0 (fully polygenic, ridge regression) to 1 (sparse, lasso).

More specifically, 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 mixing parameter that yields the largest cross-validation R^2 informs the degree of 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 smaller for $\alpha = 0$ and $\alpha = 0.05$, but nearly identically 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 all? 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.

Direct estimation of sparsity using BSLMM

To further confirm the local sparsity of gene expression traits, we turned to the BSLMM [18] approach, which models the genetic contribution as the sum of a sparse and a polygenic component. The parameter PGE in this model represents the proportion sparse to polygenic component using. Another parameter, the total

variance explained (PVE) by additive genetic variants, is a more flexible Bayesian equivalent of the chip heritability we have estimated using a linear mixed model as implemented in GCTA.

As anticipated, we find that for highly heritable genes, the sparse component is large. 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).

Also as expected, we find that when the sample size is large enough, such as in DGN, there is a strong correlation between BSLMM-estimated PVE and GCTA-estimated h^2 (Fig 6A, R=0.96). In constrast, 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 is further confirmation of the local sparse architecture of gene expression traits: the underlying assumption in LMM approaches to estimate heritability is that the genetic effect sizes are normally distributed, i.e. most variants have small effect sizes. LMM is quite robust to departure from this assumption but only when the sample size is rather large. For the relatively small sample sizes in GTEx ($n \leq 361$), we are finding that a model that directly addresses the sparse component such as BSLMM outforms GCTA for estimating h2.

Orthogonal decomposition of cross-tissue and tissue specific expression traits

Since a substantial portion of local regulation was shown to be common across multiple tissues cite gtex-pilot, we sought to decompose the expression levels into a component that is common across all tissues and a tissue specific components. For this we use a linear mixed effects model as described in the Methods section. We call this approach orthogonal tissue decomposition (OTD) because the cross tissue and tissue specific components are assumed to be independent in the model. The decomposition is applied at the expression trait level so that the downstream genetic regulation analysis is performed separately for each derived trait, cross tissue and tissue specific expression, which greatly reduces computational burden.

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

An clear benefit of OTD for the cross tissue trait is that the effective sample size of the trait is 450 even though each of the tissues had less than 360 individuals. This is reflected in more precise estimates of h2 as shown below.

For all the derived phenotypes, one cross tissue and 40 tissue specific ones, we computed the local heritability and generated prediction models.

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 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 did you download the top eQTL's prob of being eqtl in the

tissue? From what I talked with Matthew Stephens I thought it was the probability at the gene level, of the gene being an egene, i.e. being regulated in the tissue by some eqtl., we defined an entropy statistic (see Methods) that combines the nine posterior probabilities into one value such that higher entropy values mean the gene regulation is more uniform 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. Figure 15 shows the correlation between the heritability of tissue specific gene expression traits and the posterior probability of a gene being an e-gene in the tissue is thir correct?. 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

Motivated by the key role that regulatory variation plays in the genetic control of complex traits [1–3], we performed a survey of the heritability and patterns of effect sizes of gene expression traits across a comprehensive set of human tissues. We quantified the local and distal heritability of gene expression in DGN and 40 different tissues from the GTEx consortium. For the DGN dataset, we estimate the relative proportion of mean local and distal genetic contribution to gene expression traits. For GTEx samples it was not possible to estimate the mean distal heritability because of the limited sample size. As the number of GTEx samples grows nearing the 1000 individuals we expect to be able to estimate these values.

The proportion of local contribution in whole blood from DGN was 80%, which is more than twice values reported previously citeprice et al. This could be due to the over-correction of distal effects that may occur when applying hidden factor adjustments and thus leading to underestimates of the distal component. These corrections have been shown to increase the number and reproducibility of identified local eQTLs but their consequence on distal regulation is not well understood.

we need to check whether Battle et al computed the proportion of local vs distal for DGN

We showed that restricting distal variants to known functional variants such as eQTL data from independent studies improves the precision of distal heritability estimates but also reduces mean distal heritability estimates by half. For GTEx tissues, we ended up computing the distal heritability using only the subset of functional (cis eqtl) variants to reduce the errors in the estimates at the expense of missing variants that contribute to the traits.

Using results implied by the improved predictive performance of sparse models and by directly estimating sparsity using BSLMM (Bayesian Sparse Linear Mixed Model) we show evidence that for highly heritable genes local regulation is sparse across all the tissues analyzed here. For genes with moderate and low heritability the evidence is not as strong but results are consistent with a sparse local architecture. Better methods to correct for hidden confounders that does not dilute distal signals and larger sample sizes will be needed to determine the properties of distal regulation.

Given that a substantial portion of local regulation is shared across tissues, we propose here to decompose the expression traits into cross tissue and tissue specific components. This approach, called orthogonal tissue decomposition, aims to decouple the shared regulation from the tissue specific regulation. We examined the genetic architecture of these derived traits and find that they follow similar patterns to the original whole tissue expression traits. The cross tissue component benefits from an effectively larger sample size than any individual tissue trait that is reflected in more accurate heritability estimates and consistently better prediction performance. Encouragingly, heritabiilty estimates of cross tissue traits correlate well with a measure of uniformity of regulation across tissues defined as the entropy of the vector of probability for a gene to be regulated in a given tissue. Higher entropy genes will show more uniform regulation across tissues. As for the tissue specific expression traits, we found that they recapitulate correlation with the vector of probability of tissue specific regulation. The main application for which these traits were devised is to be used as prediction models in PrediXcan. We expect tesults from the cross tissue models to relate to mechanisms that are shared across multiple tissues whereas results from the tissue specific models will inform us about the context specific mechanisms. Further tests need to be performed to assess their usefulness.

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 PredictDB 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

Motivated by the observed differences in regulatory effect sizes of variants located in the vicinity of the genes and distal to the 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 $w_{k,g} \sim N(0, \sigma_w^2)$ and $\epsilon \sim N(0, \sigma_\epsilon^2 I_n)$, 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.

Approximate confidence intervales were computed as the point estimate + or - 2 times the estimated standard error. The intervals were also forced to be ≥ 0 or ≤ 1. Genes were considered to have heritability significantly different from 0 if the confidence interval did not include 0.

By default we restricted the heritability estimates to be in the 0 to 1 interval. However, for the purpose of estimating the mean heritability we performed separate runs allowing the heritability estimates to take negative values with the –reml-no-constrain option in GCTA. Despite the lack of obvious biological interpretation of a negative heritability, it is an accepted procedure used in order to avoid bias in the estimated mean. cite alkes, wright

Determining polygenicity versus sparsity using the elastic net

We used the glmnet package to fit an elastic net model where the tuning parameter is chosen via 10 fold cross validation to maximize prediction performance measure by Pearson's R2 [30,31].

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 is a linear model with a polygenic component (small effects) and a sparse component (large effects) enforced by 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

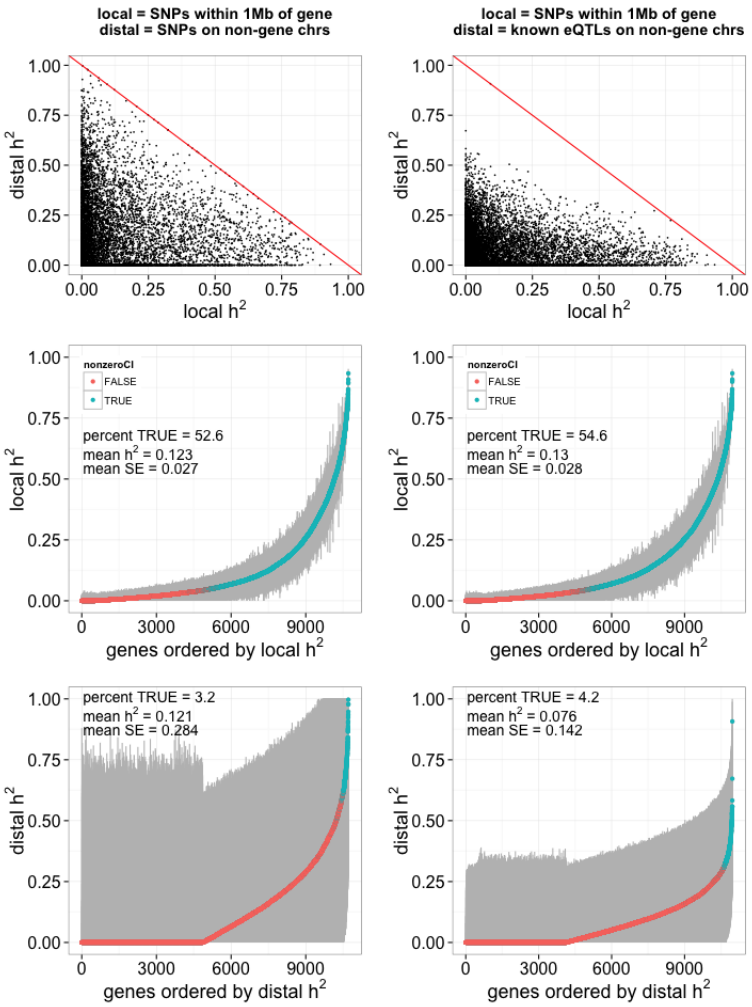

Figure 1: DGN whole blood expression joint heritability (h^2). Local (SNPs within 1 Mb of each gene) and distal (Left: SNPs on non-gene chromsomes. Right: SNPs that are eQTLs in the Framingham Heart Study on other chromosomes [FDR < 0.05]) h^2 for gene expression were jointly estimated. (**Top**) Distal 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.

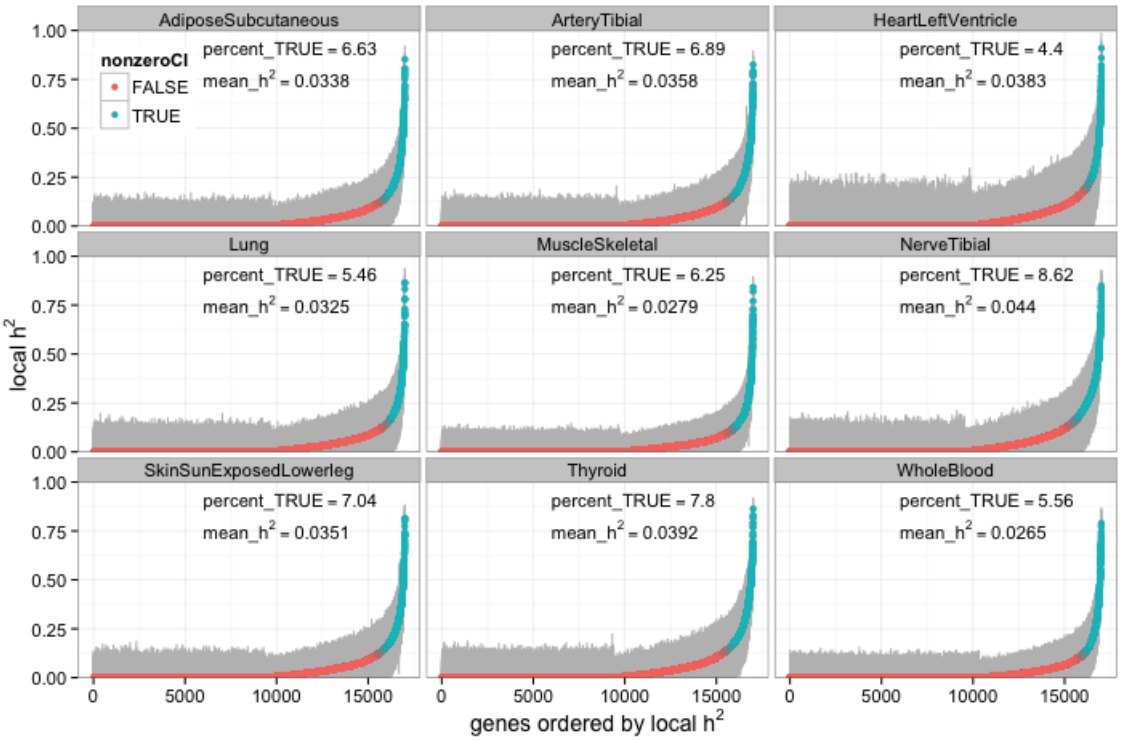

Figure 2: GTEx whole tissue local heritability (h^2) estimation. Local (SNPs within 1 Mb of each gene) gene expression h^2 estimates ordered by increasing h^2. The 95% confidence interval (CI) of each h^2 estimate is in gray and genes with a lower bound greater than zero are in blue

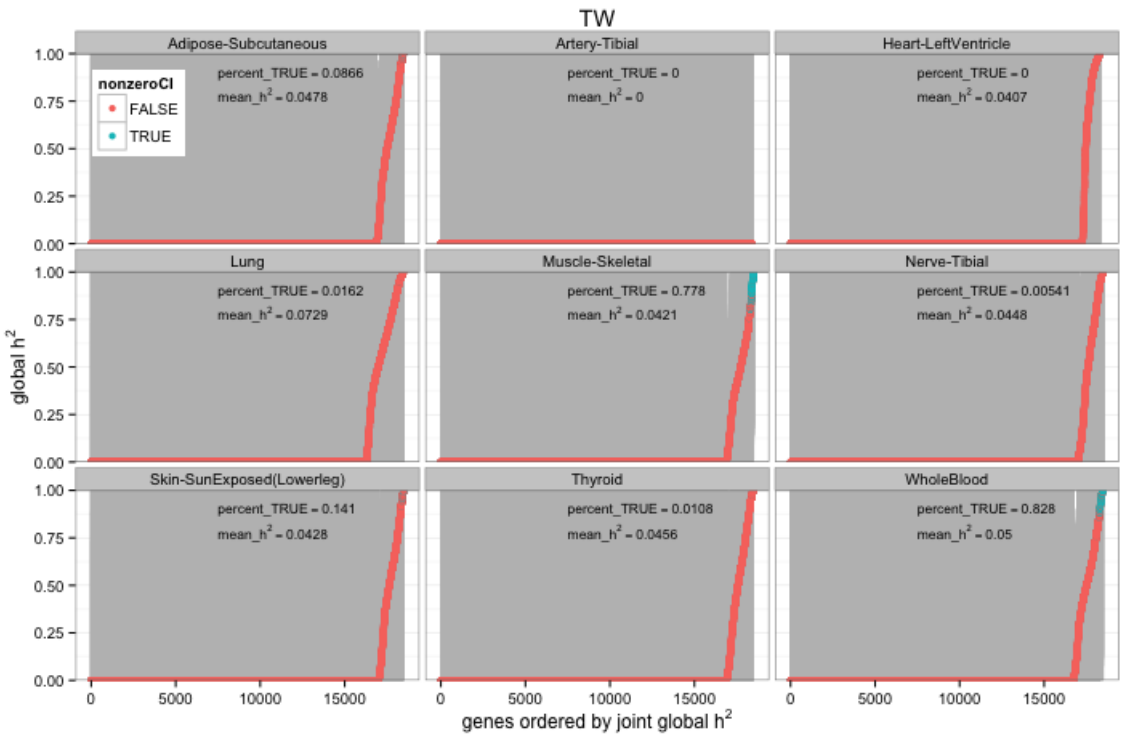

Figure 3: GTEx whole tissue distal heritability (h^2) estimation. Distal (SNPs that are eQTLs in the Framingham Heart Study on other chromosomes [FDR < 0.05]) gene expression h^2 estimates from a joint model are ordered by increasing h^2. The 95% confidence interval (CI) of each h^2 estimate is in gray and genes with a lower bound greater than zero are in blue

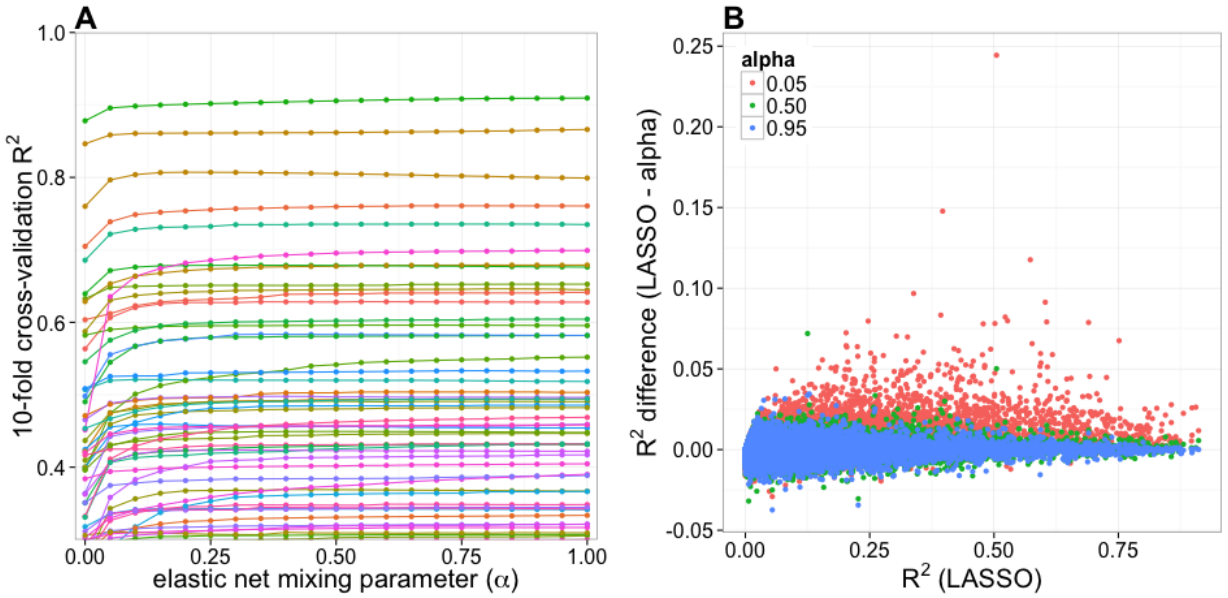

Figure 4: DGN cross-validated predictive performance across the elastic net. (**A**) 10-fold cross-validated R^2 of predicted vs. observed expression in DGN whole blood compared to a range of elastic net mixing parameters (α) for genes on chromosome 22 with $R^2 > 0.3$. (**B**) Predictive R^2 difference between LASSO ($\alpha = 1$) and several other values of α compared to LASSO predictive R^2 for 13171 autosomal genes.

Supplemental Figures

Acknowledegments

We thank Nicholas Knoblauch and Jason Torres for initial pipeline development and planning.

Grants

We acknowledge the following US National Institutes of Health grants: R01MH107666 (H.K.I.), K12 CA139160 (H.K.I.), T32 MH020065 (K.P.S.), R01 MH101820 (GTEx), P30 DK20595 and P60 DK20595 (Diabetes Research and Training Center), P50 DA037844 (Rat Genomics), P50 MH094267 (Conte). H.E.W. was supported in part by start-up funds from Loyola University Chicago.

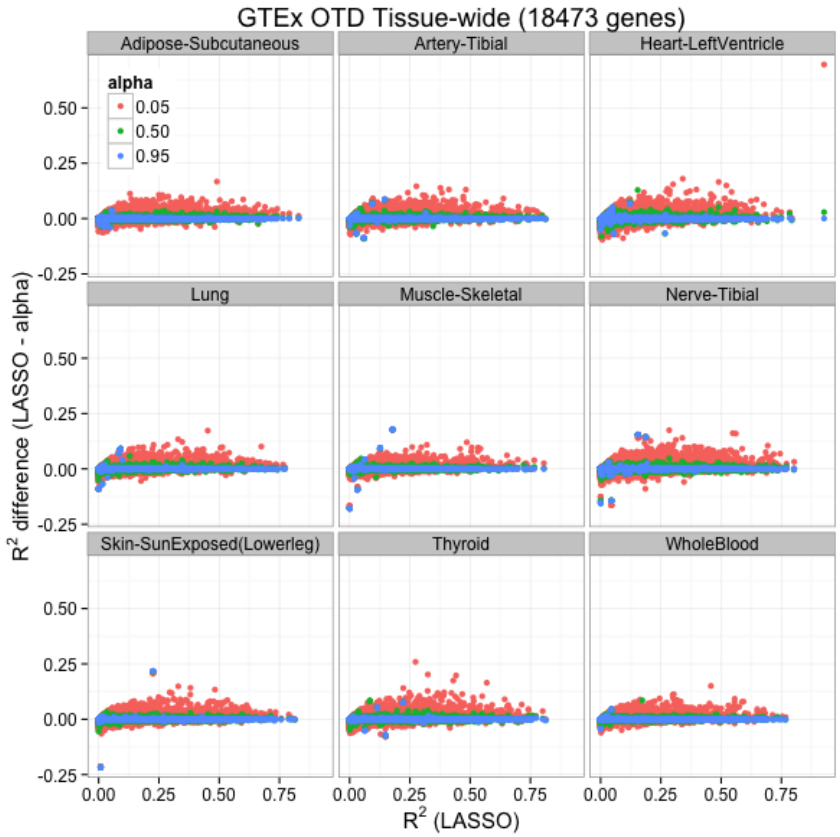

Figure 5: GTEx whole tissue cross-validated predictive performance across the elastic net. Predictive R^2 difference between LASSO ($\alpha = 1$) and several other values of α compared to LASSO predictive R^2 for 18473 autosomal genes per tissue.

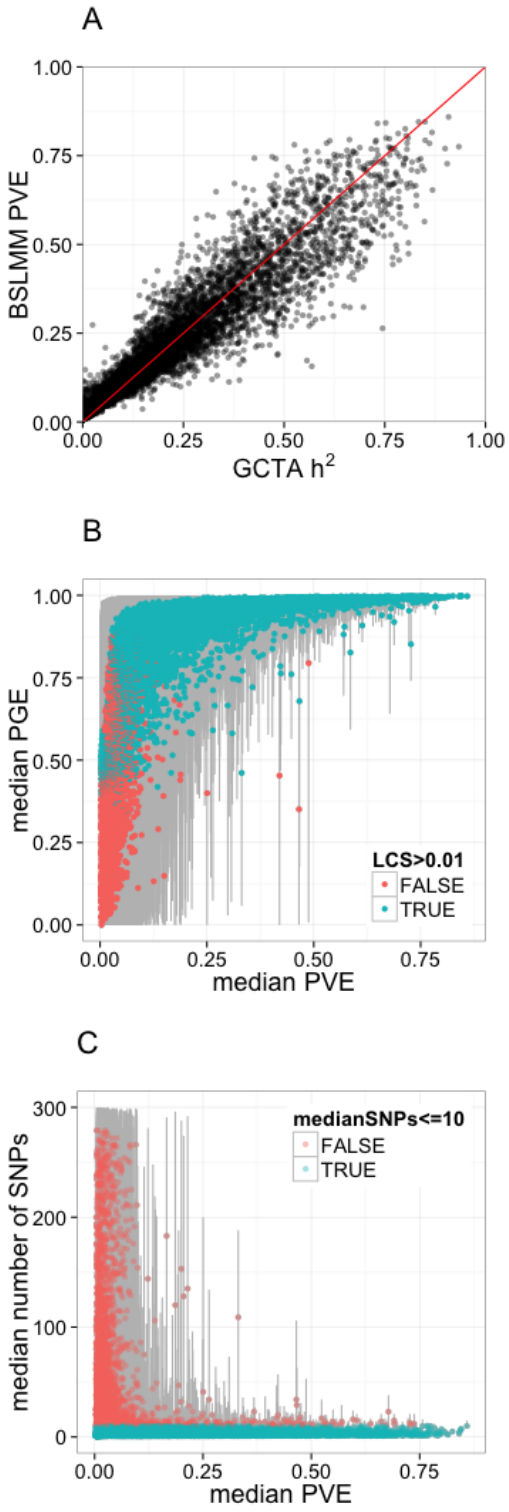

Figure 6: Bayesian Sparse Linear Mixed Models reveal the sparsity of gene expression architecture. (**A**) BSLMM-estimated PVE (total proportion of variance explained) compared to GCTA-estimated heritability per gene (R=0.96) (**B**) Comparison of median PGE (proportion of PVE explained by sparse effects) 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.

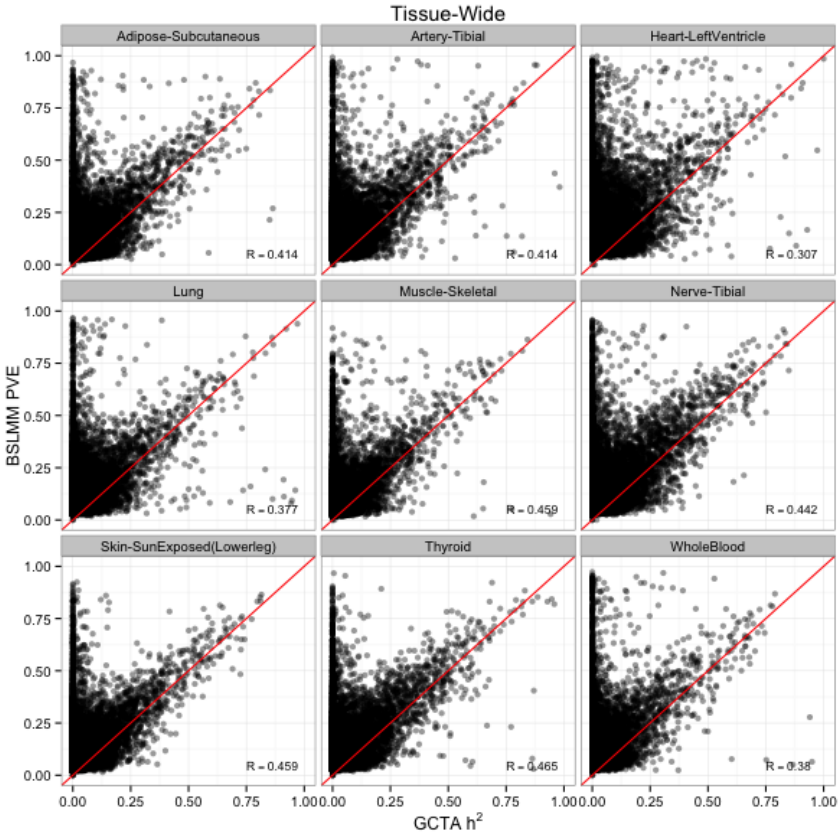

Figure 7: GTEx whole tissue expression BSLMM-estimated PVE (total proportion of variance explained) compared to GCTA-estimated heritability per gene. R = Pearson correlation.

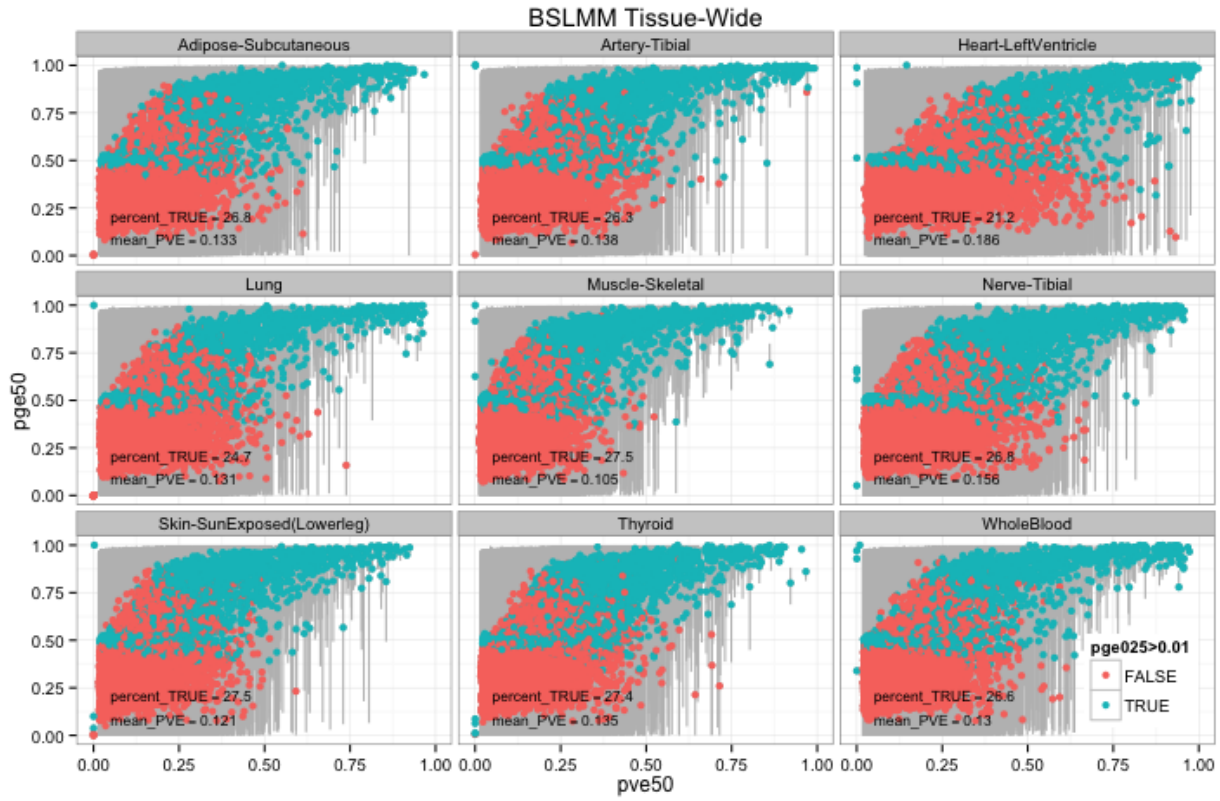

Figure 8: GTEx whole tissue expression comparison of median PGE (proportion of PVE explained by sparse effects) to median PVE (total proportion of variance explained) for expression of each gene. The 95% credible set of each PGE estimate is in gray and genes with a lower credible set (LCS) greater than 0.01 are in blue.

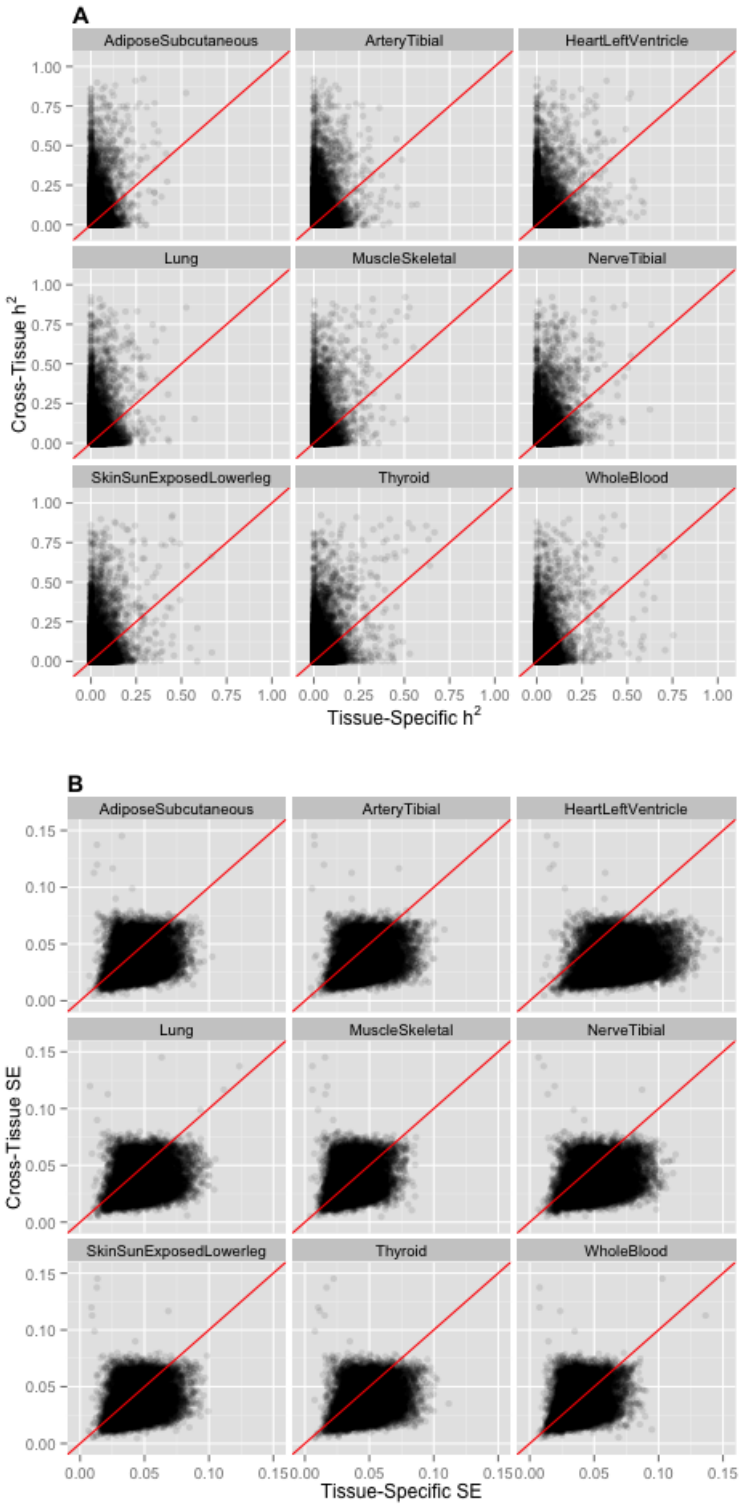

Figure 9: Cross-tissue and tissue-specific comparison of heritability (h^2, **A**) and standard error (SE, **B**) estimation. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-specifc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

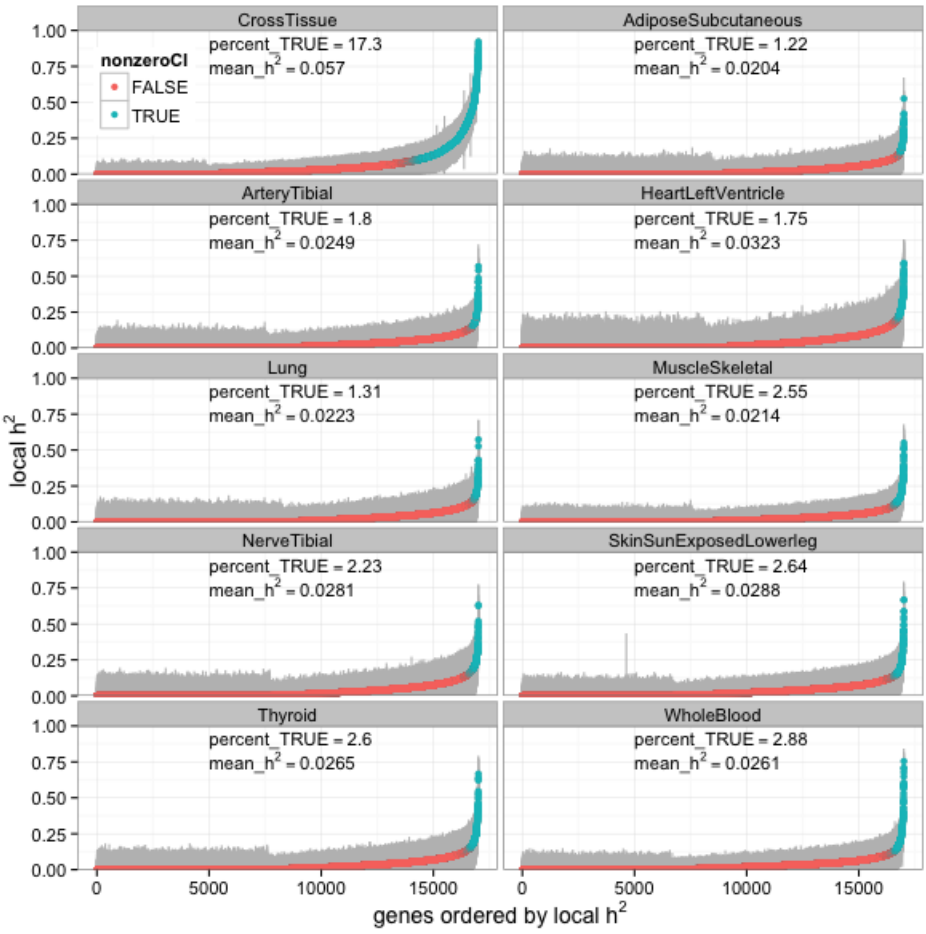

Figure 10: Cross-tissue heritability (h^2) compared to tissue-specific h^2. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-specifc local h^2 is estimated using the tissue-specific component (residuals) of the mixed effects model for gene expression for each respective tissue and SNPs within 1 Mb of each gene.

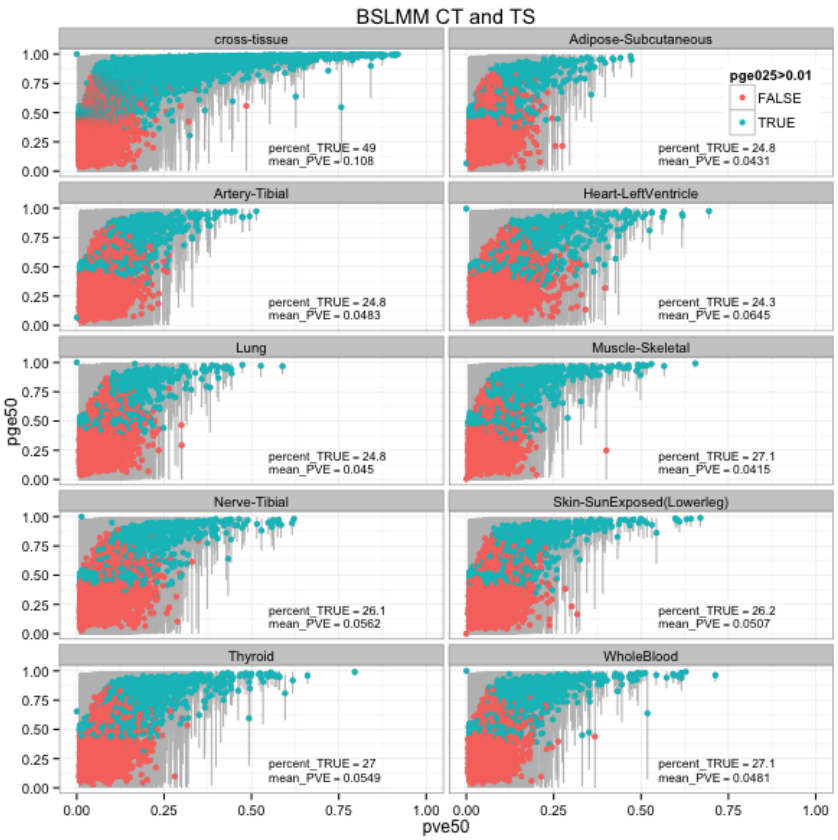

Figure 11: GTEx orthogonal tissue decomposition cross-tissue and tissue-specific expression comparison of median PGE (proportion of PVE explained by sparse effects) to median PVE (total proportion of variance explained) for expression of each gene. The 95% credible set of each PGE estimate is in gray and genes with a lower credible set (LCS) greater than 0.01 are in blue.

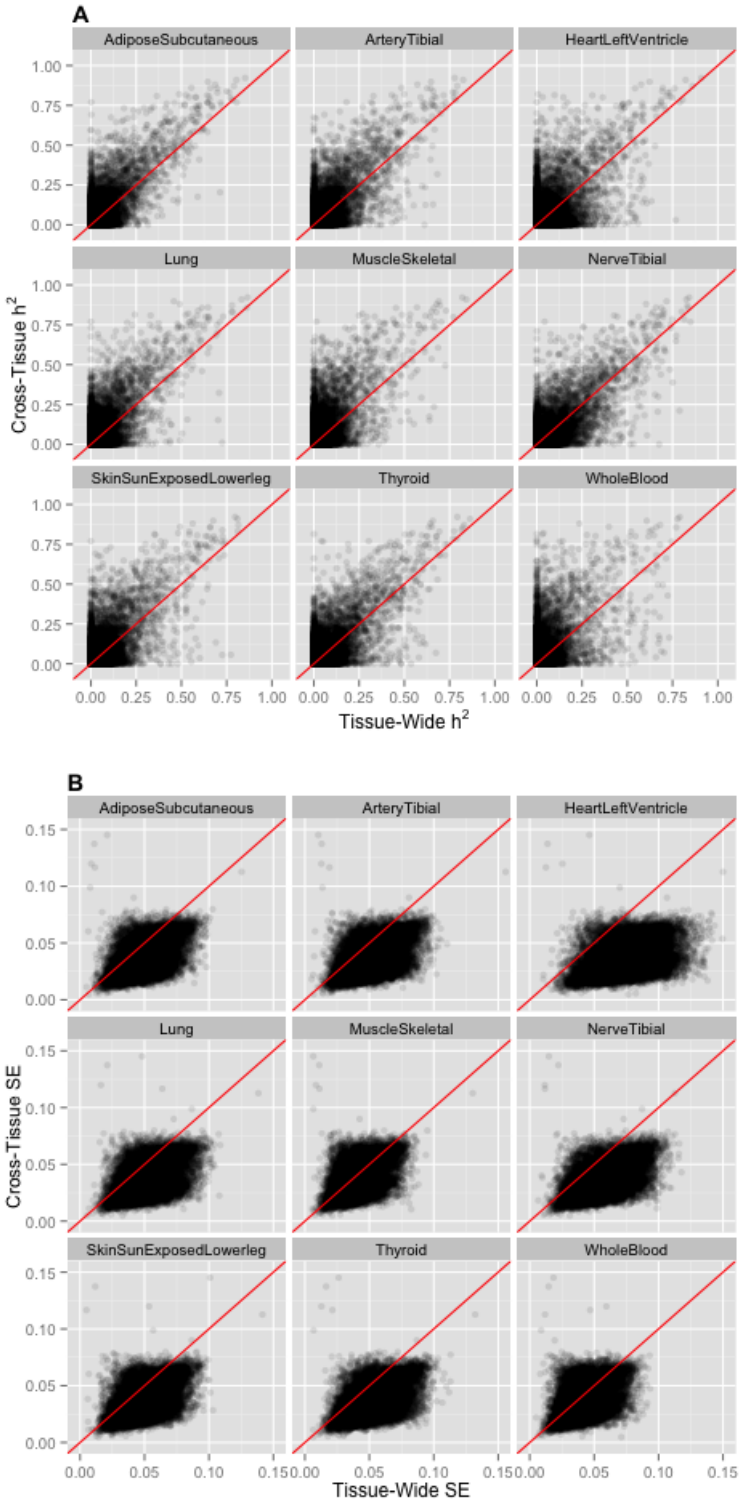

Figure 12: Cross-tissue and whole tissue comparison of heritability (h^2, **A**) and standard error (SE, **B**). Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Whole tissue local h^2 is estimated using the measured gene expression for each respective tissue and SNPs within 1 Mb of each gene.

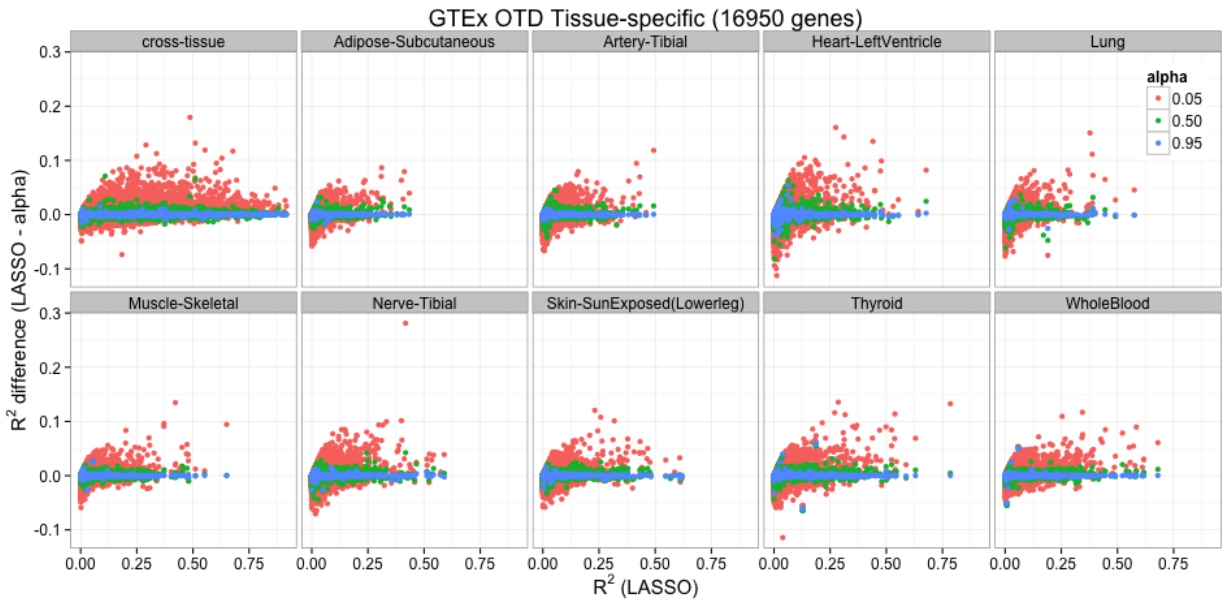

Figure 13: GTEx orthogonal tissue decomposition cross-tissue and tissue-specific expression cross-validated predictive performance across the elastic net. Predictive R^2 difference between LASSO ($\alpha = 1$) and several other values of α compared to LASSO predictive R^2 for 18473 autosomal genes per tissue.

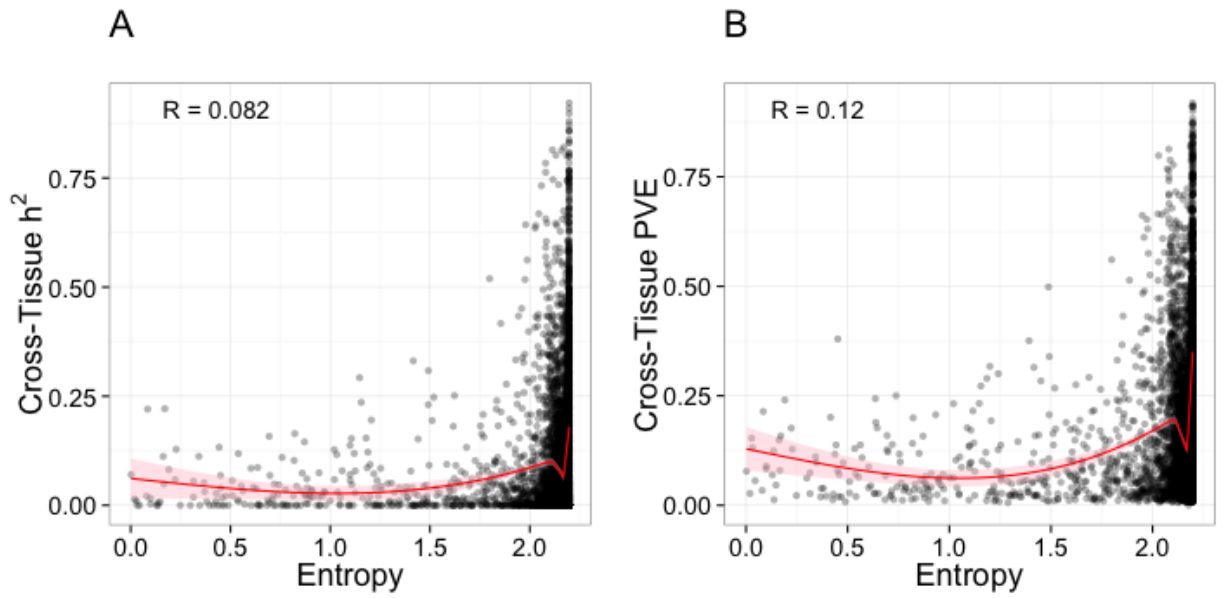

Figure 14: Entropy of the posterior probabilities from the Flutre et al. multi-tissue eQTL method compared to the estimates of (**A**) heritability and (**B**) PVE of cross-tissue gene expression derived from the orthogonal tissue decomposition. The generalized additive model smoothing line is in red.

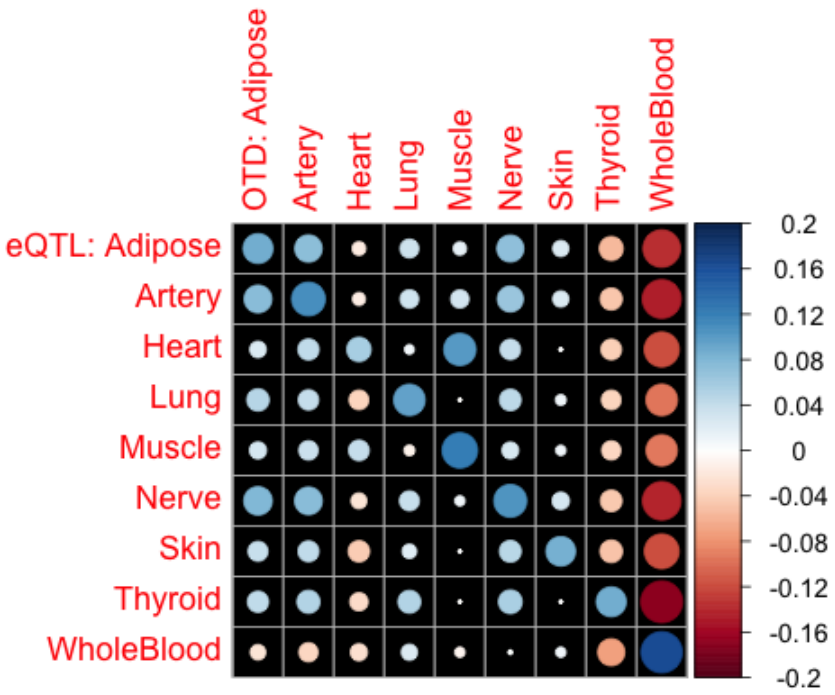

Figure 15: Pearson correlation (R) between the posterior probability the top multi-tissue eQTL regulates its gene in a given tissue (eQTL, Flutre et al. method) and the PVE of tissue-specific gene expression from the orthogonal tissue decomposition (OTD). Area of each circle is proportional to the absolute value of R.

GTEx data

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, 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. 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 https://www.opensciencedatacloud.org/ Grossman RL, Greenway M, Heath AP, Powell R, Suarez R, Wells 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; 2015. 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