Survey of the Heritability and Sparsity of Gene Expression Traits Across Human Tissues

Heather E. Wheeler[1,2,*], Kaanan P. Shah[3], Jonathon Brenner[2], Tzintzuni Garcia[4], Keston Aquino-Michaels[3], . . ., GTEx Consortium, . . ., Nancy J. Cox[5], Dan L. Nicolae[3], Hae Kyung Im[3,*]

1 Department of Biology, Loyola University Chicago, Chicago, IL, USA

2 Department of Computer Science, Loyola University Chicago, Chicago, IL, USA

3 Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA

4 Bioinformatics Core, CRI <mark>check this</mark>, University of Chicago, Chicago, IL, USA

5 Division of Genetic Medicine, Vanderbilt University, Nashville, TN, USA

* hwheeler1@luc.edu, haky@uchicago.edu

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 the genetic architecture of gene expression traits is key to elucidating the underlying mechanisms of complex traits. However, a systematic survey of the heritability and the distribution of effect sizes across all representative tissues in the human body has not been reported.

Here we fill this gap through analyses of the RNA-seq data from a comprehensive set of tissue samples generated by the GTEx Project and the DGN whole blood cohort. We find that local h^2 can be relatively well characterized with 49% of expressed genes

showing significant h^2 in DGN and 2-10% in GTEx. However, the current sample sizes ($n < 362$ in GTEx) only allow us to compute distal h^2 for a handful of genes (3% in DGN and <0.8% in GTEx). Thus, we focus on local regulation. Bayesian Sparse Linear Mixed Model (BSLMM) analysis and the sparsity of optimal performing predictors provide compelling evidence that local architecture of gene expression traits is sparse rather than polygenic across DGN and all 40 GTEx 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 suggest 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 decomposed expression phenotypes are made publicly available (`https://github.com/hakyimlab/PrediXcan`).

Author Summary

Gene regulation is known to contribute to the underlying mechanisms of complex traits. 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. Regardless of tissue type, we found that local heritability can be well characterized with current sample sizes. Unless strong functional priors and large sample sizes 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 orthogonal tissue decomposition model recapitulate complex Bayesian multi-tissue analysis results. This knowledge was applied to develop prediction models of gene expression traits for

all tissues, which we make publicly available.

Introduction

Regulatory variation plays a key role in the genetics of complex traits [1–3]. Methods that partition the contribution of environment and genetic components are useful tools to understand the biology underlying complex traits. Partitioning heritability into different functional classes has been successful in quantifying the contribution of different mechanisms that drive the etiology of diseases [3–5].

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 [6,7]. Furthermore, when both local and distal eQTLs are reported [8–10], 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 [10, 11].

While many common diseases are likely polygenic [12–14], 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 [6].

The relative prediction 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 predict 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 gene-level association method, PrediXcan [15]. When we calibrated the prediction model that was used in the 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 [15], suggesting that for many genes, the genetic architecture is sparse, but not regulated by a single SNP.

Thus, gene expression traits with sparse architecture should be better predicted with   30
models such as LASSO (Least Absolute Shrinkage and Selection Operator), which   31
prefers solutions with fewer parameters, each of large effect [16]. Conversely, highly   32
polygenic traits should be better predicted with ridge regression or similarly polygenic   33
models that prefer solutions with many parameters, each of small effect [17–19]. To   34
obtain a more thorough understanding of gene expression architecture, we used the   35
hybrid approaches of the elastic net and BSLMM (Bayesian Sparse Linear Mixed   36
Model) [20] to quantify sparse and polygenic effects.   37

Most previous human eQTL studies were performed in whole blood or   38
lymphoblastoid cell lines due to ease of access or culturabilty [8, 21, 22]. Although   39
studies with a few other tissues have been published, comprehensive coverage of human   40
tissues was not available until the launching of the Genotype-Tissue Expression (GTEx)   41
Project. GTEx aims to examine the genetics of gene expression more comprehensively   42
and has recently published a pilot analysis of eQTL data from 1641 samples across 43   43
tissues from 175 individuals [23]. Here we use a much larger set of 8555 samples across   44
53 tissues corresponding to 544 individuals. One of the findings of this comprehensive   45
analysis was that a large portion of the local regulation of expression traits is shared   46
across multiple tissues. Corroborating this finding, our prediction model based on whole   47
blood showed robust prediction across the 9 core GTEx tissues chosen by initial sample   48
sizes [15].   49

This shared regulation implies that there is much to be learned from large sample   50
studies of easily accessible tissues. Yet, a portion of gene regulation seems to be tissue   51
dependent [23]. In order to harness this cross-tissue effect for prediction and to better   52
understand the genetic architecture of tissue-specific and cross-tissue gene regulation,   53
we use a mixed effects model called orthogonal tissue decomposition (OTD) to decouple   54
the cross-tissue and tissue-specific mechanisms in the rich GTEx dataset. We modeled   55
the underlying genetic architecture of the cross-tissue and tissue-specific gene expression   56
components and developed predictors for use in PrediXcan [15].   57

Results

Local genetic variation can be well characterized for all 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 72 to 361 depending on the tissue, while 922 samples were available in DGN [22]. We used mixed-effects models (see Methods) and calculated variances using restricted maximum likelihood as implemented in GCTA [24].

For the local heritability component, we used variants within 1Mb of the transcription start and end 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 pick the set of distal variants were explored, but results were robust to different selections. See more details in Methods.

Table 1 summarizes the unconstrained heritability estimate results across all tissues. In order to obtain an unbiased estimates of mean h^2, we allow the values to be negative when fitting the REML, as done previously [10,11]. This approach reduces the standard error of the estimated mean of heritability, especially important for the distal component. Even though each invididual gene's distal heritability is noisy, averaging across all genes reduces the error substantially. For the DGN dataset, we were able to estimate the mean distal h^2. However for the GTEx samples, the sample size was too small and the REML algorithm became unstable 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.

The left column of Fig. 1 shows the estimated local and distal h^2 from DGN, this time using REML constrained between 0 and 1 (GCTA default) [24]. Even though many genes show relatively large point estimates of distal h^2, only the ones colored in blue are significantly different from zero. The local component of h^2 is relatively well estimated in DGN with 48.6% of genes (6180 out of 12719) showing h^2 values significantly different from zero. In contrast, the distal heritability is significantly different from zero for only 2.7% (343 out of 12719) of the genes.

Table 1. Estimates of local h² across whole tissues.

tissue	n	mean h²	mean se	prop CI > 0	num CI > 0	num expressed
DGN-WholeBlood	922	0.143	0.0292	0.486	6180	12719
Adipose-Subcutaneous	298	0.0385	0.0381	0.0735	1040	14205
AdrenalGland	126	0.0432	0.075	0.0425	601	14150
Artery-Aorta	198	0.0421	0.0561	0.0649	898	13844
Artery-Coronary	119	0.0371	0.0773	0.0337	476	14127
Artery-Tibial	285	0.0417	0.0402	0.0798	1080	13504
Brain-Anteriorcingulatecortex(BA24)	72	0.0275	0.133	0.031	450	14515
Brain-Caudate(basalganglia)	100	0.0367	0.091	0.0343	502	14632
Brain-CerebellarHemisphere	89	0.0492	0.11	0.0509	728	14295
Brain-Cerebellum	103	0.0504	0.094	0.0544	788	14491
Brain-Cortex	96	0.0451	0.0937	0.0393	578	14689
Brain-FrontalCortex(BA9)	92	0.0379	0.101	0.0341	496	14554
Brain-Hippocampus	81	0.0368	0.114	0.0285	414	14513
Brain-Hypothalamus	81	0.017	0.115	0.0235	347	14759
Brain-Nucleusaccumbens(basalganglia)	93	0.0293	0.0965	0.0292	426	14601
Brain-Putamen(basalganglia)	82	0.0324	0.108	0.0291	419	14404
Breast-MammaryTissue	183	0.0289	0.053	0.0365	537	14700
Cells-EBV-transformedlymphocytes	115	0.0578	0.0814	0.0497	619	12454
Cells-Transformedfibroblasts	272	0.0508	0.0424	0.0925	1180	12756
Colon-Sigmoid	124	0.0327	0.0807	0.0389	557	14321
Colon-Transverse	170	0.0358	0.059	0.0436	640	14676
Esophagus-GastroesophagealJunction	127	0.0318	0.0761	0.0332	469	14125
Esophagus-Mucosa	242	0.0416	0.0479	0.0764	1090	14239
Esophagus-Muscularis	219	0.0393	0.0505	0.069	969	14047
Heart-AtrialAppendage	159	0.0418	0.0638	0.0475	660	13892
Heart-LeftVentricle	190	0.0342	0.0558	0.0503	670	13321
Liver	98	0.033	0.0935	0.0299	405	13553
Lung	279	0.0315	0.0401	0.0577	853	14775
Muscle-Skeletal	361	0.0327	0.0324	0.0732	939	12833
Nerve-Tibial	256	0.0523	0.0445	0.095	1380	14510
Ovary	85	0.0369	0.0988	0.0258	364	14094
Pancreas	150	0.0469	0.0674	0.0557	777	13941
Pituitary	87	0.0379	0.107	0.0358	544	15183
Skin-NotSunExposed(Suprapubic)	196	0.0407	0.0532	0.0502	735	14642
Skin-SunExposed(Lowerleg)	303	0.0385	0.0385	0.0765	1120	14625
SmallIntestine-TerminalIleum	77	0.0365	0.112	0.0281	418	14860
Spleen	89	0.059	0.0998	0.0417	603	14449
Stomach	171	0.0315	0.0579	0.0367	533	14531
Testis	157	0.0539	0.0679	0.0727	1230	16936
Thyroid	279	0.0436	0.0413	0.0861	1260	14642
WholeBlood	339	0.0326	0.0328	0.0677	823	12160

Except for DGN-WholeBlood, all tissues are from the GTEx Project.

It has been shown that local-eQTLs are more likely to be distal-eQTLs of target genes [25]. Thus, we tested whether restricting the distal genetic similarity computation to QTLs (as determined in the Framingham mRNA dataset of over 5000 individuals [26] independent of the DGN and GTEx cohorts) for other genes could improve distal

heritability precision by prioritizing functional variants. We exclude eQTLs on the same 93
chromosome as the tested gene to avoid contaminating distal h^2 with cis associations. 94

Using functional priors (known eQTLs) to define distal h^2 increased the percentage 95
of genes with a positive CI from 2.7% (343 genes) to 3.6% (458) in whole blood (Fig. 1). 96
A total of 125 genes have significant distal h^2 by both approaches, i.e. all variants on 97
other chromosomes or only known eQTL variants on other chromosomes. 98

However, using the subset of known eQTLs (from an independent source) in other 99
chromosomes for computing distal heritability reduced the mean value from 3.4% to 100
1.6%. Therefore, while we gain some power to detect significant distal heritability by 101
using cis eQTL priors, a good portion of the distal regulation is lost when using only 102
the smaller subset of known cis-eQTL variants. We used functional priors to estimate 103
distal h^2 in the GTEx cohort, but less than 1% of genes had a positive CI (S1 Fig). 104

Given the limited sample size we will focus on local regulation for the remainder of 105
the paper. 106

Sparse local architecture implied by sparsity of best prediction models 107 108

Next, we sought to determine whether the local genetic contribution to gene expression 109
is polygenic or sparse. In other words, whether many variants with small effects or a 110
small number of large effects were contributing to expression trait variability. For this, 111
we first looked at the prediction performance of a range of models with different degrees 112
of polygenicity, such as the elastic net model with mixing parameter values ranging 113
from 0 (fully polygenic, ridge regression) to 1 (sparse, LASSO). 114

More specifically, we performed 10-fold cross-validation using the elastic net [27] to 115
test the predictive performance of local SNPs for gene expression across a range of 116
mixing parameters (α). The mixing parameter that yields the largest cross-validation 117
R^2 informs the degree of sparsity of each gene expression trait. That is, at one extreme, 118
if the optimal $\alpha = 0$ (equivalent to ridge regression), the gene expression trait is highly 119
polygenic, whereas if the optimal $\alpha = 1$ (equivalent to LASSO), the trait is highly 120
sparse. We found that for most gene expression traits, the cross-validated R^2 was 121
smaller for $\alpha = 0$ and $\alpha = 0.05$, but nearly identically for $\alpha = 0.5$ through $\alpha = 1$ in the 122

DGN cohort (Fig. 2). An $\alpha = 0.05$ was also clearly suboptimal for gene expression prediction in the GTEx tissues, while models with $\alpha = 0.5, 0.95$, or 1 had similar predictive power (S2 Fig). 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 also points to sparse local architecture

To further confirm the local sparsity of gene expression traits, we turned to the BSLMM [20] 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 of genetic variance explained by sparse effects. 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 (LMM) 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. 3A). 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. 3B).

BSLMM outperforms LMM in estimating h^2 for small samples

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. 3C, R=0.96). In contrast, when we applied BSLMM to the GTEx data, we found that many genes had measurably larger BSLMM-estimated PVE than GCTA-estimated h^2 (Fig. 4). This is further confirmation of the local sparse architecture of gene expression traits: the underlying assumption in the GCTA (LMM) approach 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 found that a model that directly addresses the sparse component such as

BSLMM outperforms GCTA for estimating h^2. 152

Orthogonal decomposition of cross-tissue and tissue-specific expression traits 153 154

Since a substantial portion of local regulation was shown to be common across multiple 155
tissues [23], we sought to decompose the expression levels into a component that is 156
common across all tissues and tissue-specific components. For this we use a linear mixed 157
effects model with a person-level random effect. See details in Methods. We use the 158
posterior mean of this random effect as an estimate of the cross tissue component. We 159
consider the residual component of this model as the tissue specific component. Below 160
we describe the properties of these derived phenotypes. 161

We call this approach orthogonal tissue decomposition (OTD) because the 162
cross-tissue and tissue-specific components are assumed to be independent in the model. 163
The decomposition is applied at the expression trait level so that the downstream 164
genetic regulation analysis is performed separately for each derived trait, cross-tissue 165
and tissue-specific expression, which greatly reduces computational burden. For all the 166
derived phenotypes, one cross-tissue and 40 tissue-specific ones, we computed the local 167
heritability and generated prediction models. 168

Cross-tissue expression phenotype is less noisy and more predictable 169 170

Our estimates of h^2 for cross tissue expression traits are larger than the corresponding 171
estimates for each whole tissue expression traits (S3 Fig). This means that our OTD 172
approach increases the ratio of genetically regulated component to noise by averaging 173
across multiple tissues. In addition to the increased h^2 we observe reduction in standard 174
errors of the estimated h^2. This is partly due to the increased h^2 – higher h^2 are better 175
estimated– but also due to the larger effective sample size for cross tissue phenotypes. 176
There were 450 samples for which cross tissue traits were avialable whereas the 177
maximum sample size for whole tissue phenotypes was 362. As consequence of this 178
increase h^2 and decreased standard errors, the percentage of cross h^2 estimates with 179
positive CIs was 19.8% whereas for whole tissue expression traits they ranged from 180

2.4-9.5%. Similarly, cross-tissue BSLMM PVE estimates had lower error than whole tissue PVE (S4 Fig, S5 Fig).

As for the tissue specific components, the cross-tissue heritability estimates were also larger and the standard errors were smaller reflecting the fact that a substantial portion of regulation is common across tissues (S6 Fig). The percentage of GCTA h^2 estimates with positive CIs was much larger for cross-tissue expression (19.8%) than the tissue-specific expressions (all less than 4%, S1 Table). 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 (S5 Fig).

Cross-tissue predictive performance exceeded that of both tissue-specific and whole tissue expression as indicated by higher cross-validated R^2 (S2 Fig, S7 Fig). Like whole tissue expression, cross-tissue and tissue-specific expression showed better predictive performance when using more sparse models. In other words elastic-net models with $\alpha \geq 0.5$ predicted better than the ones with $\alpha = 0.05$ (S7 Fig).

——

Cross Tissue expresion phenotype recapitulates published multi-tissue eQTL results

We compared our OTD results to those from a joint multi-tissue eQTL analysis method [28], which was previously performed on a subset of the GTEx data [23] covering 9 tissues. In particular, we used the posterior probability of a gene being actively regulated (PPA) in a tissue. These analysis results are available on the GTEx portal.

First, we reasoned that genes with high cross tissue h^2 would be actively regulated in most tissues so that the PPA of a gene would be roughly uniform across tissues. By contrast, a gene with tissue specific regulation would have concentrated posterior probability in one or a few tissues. Thus we decided to define a measure of uniformity of the posterior probability vector across the 9 tissues using the concept of entropy. More specifically, for each gene we normalized the vector of posterior probabilities so that the sum equaled 1. Then we applied the usual entropy definition (negative of the sum of the log of the posterior probabilities weighted by the same probabilities, see Methods). In other words, we defined an entropy statistic 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 small subset of the nine.

Reassuringly, we found that the cross tissue phenotype heritability estimates were well correlated with the entropy statistics indicating that our derived cross tissue phenotype is capturing the common regulatory component. Figure 5 shows the comparison between this entropy statistics and cross-tissue expression heritability (R=0.082) and PVE (R = 0.12).

For the original whole tissue, we expected the whole tissue expression heritability to correlate with the posterior probability of a gene being actively regulated in a tissue as shown in Figure 6A. Strikingly, the tissue specific phenotype heritability showed highest correlation with the PPA of the matching tissue and lower correlations with other tissues, confirming that OTD also captures tissue-specific components of gene expression (Fig. 6B). However, for whole blood there is a consistently negative correlation with other tissues. This issue needs to be further investigated.

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 to near 1000 individuals, we expect to be able to estimate these values.

In DGN (whole blood), the mean local h^2 was 14.3% and the mean distal h^2 was 3.4% such that the local variation contribution is estimated as $14.3/(3.4+14.3) = 81\%$. This is much higher than the 37% reported by Price et al. [11] based on blood expression data from a cohort of Icelandic individuals. This potentially underestimation 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 [29], SVA [30], and

other types of hidden confounder corrections have been shown to increase local eQTL $_{241}$ replicability, but their consequences on distal regulation is not well understood. As $_{242}$ larger sample sizes become available, we will test this hypothesis in GTEx data by $_{243}$ computing the distal h^2 without PEER factor correction. $_{244}$

We showed that restricting distal variants to known functional variants such as $_{245}$ eQTL data from independent studies improves the precision of distal heritability $_{246}$ estimates, but also reduces mean distal heritability estimates by half. For GTEx tissues, $_{247}$ we computed the distal heritability using only the subset of functional (known eQTLs) $_{248}$ variants to reduce the errors in the estimates at the expense of missing variants that $_{249}$ contribute to the traits. $_{250}$

Using results implied by the improved predictive performance of sparse models and $_{251}$ by directly estimating sparsity using BSLMM (Bayesian Sparse Linear Mixed Model), $_{252}$ we show evidence that for highly heritable genes, local regulation is sparse across all the $_{253}$ tissues analyzed here. For genes with moderate and low heritability the evidence is not $_{254}$ as strong, but results are consistent with a sparse local architecture. Better methods to $_{255}$ correct for hidden confounders that do not dilute distal signals and larger sample sizes $_{256}$ will be needed to determine the properties of distal regulation. $_{257}$

Given that a substantial portion of local regulation is shared across tissues, we $_{258}$ propose here to decompose the expression traits into cross-tissue and tissue-specific $_{259}$ components. This approach, called orthogonal tissue decomposition, aims to decouple $_{260}$ the shared regulation from the tissue-specific regulation. We examined the genetic $_{261}$ architecture of these derived traits and find that they follow similar patterns to the $_{262}$ original whole tissue expression traits. The cross-tissue component benefits from an $_{263}$ effectively larger sample size than any individual tissue trait, which is reflected in more $_{264}$ accurate heritability estimates and consistently better prediction performance. $_{265}$ Encouragingly, heritability estimates of cross-tissue traits correlate well with a measure $_{266}$ of uniformity of regulation across tissues defined as the entropy of the vector of $_{267}$ probability for a gene to be regulated in a given tissue. Higher entropy genes will show $_{268}$ more uniform regulation across tissues. As for the tissue-specific expression traits, we $_{269}$ found that they recapitulate correlation with the vector of probability of tissue-specific $_{270}$ regulation. The main application for which these traits were devised is to be used as $_{271}$ prediction models in PrediXcan [15]. We expect results from the cross-tissue models to $_{272}$

relate to mechanisms that are shared across multiple tissues whereas results from the tissue-specific models will inform us about the context specific mechanisms.

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`) for use in future studies of complex trait genetics.

Materials and 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 [22], 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. [22] 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 Michigan Imputation Server (`https://imputationserver.sph.umich.edu/start.html`) [31, 32] 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   302
samples (53 unique tissue types) from 544 unique subjects in the GTEx Project [23]   303
data release on 2014-06-13. Of the individuals with gene expression data, genome-wide 304
genotypes (imputed with 1000 Genomes) were available for 450 individuals. While all   305
8555 tissue samples were used in the OTD model (described below) to generate   306
cross-tissue and tissue-specific components of gene expression, we used the 40 tissues 307
with the largest sample sizes when quantifying tissue-specific effects (see Table 1). 308
Approximately 2.6M non-ambiguous strand SNPs included in HapMap Phase II were   309
retained for subsequent analyses.   310

Framingham Expression Dataset. We obtained exon array expression and   311
genotype array data from 5257 [Tzuni, what is your n?] individuals from the   312
Framingham Heart Study [26]. We used the Affymetrix power tools (APT) suite to   313
perform the preprocessing and normalization steps. First the robust multi-array   314
analysis (RMA) protocol was applied which consists of three steps: background   315
correction, quantile normalization, and summarization [33]. The background correction 316
step uses antigenomic probes that do not match known genome sequences to adjust the 317
baseline for detection, and is applied separately to each array. Next, the normalization 318
step utilizes a 'sketch' quantile normalization technique instead of a memory-intensive 319
full quantile normalization. The benefit is a much lower memory requirement with little 320
accuracy trade-off for large sample sets such as this one. Finally, the adjusted probe 321
values were summarized (by the median polish method) into log-transformed expression 322
values such that one value is derived per exon or gene. Additionally an analysis of the 323
detection of probes above the background noise (DABG) was carried out. It provides   324
further diagnostic information which can be used to filter out poorly performing probes 325
and weakly expressed genes. The summarized expression values were then annotated   326
more fully using the annotation databases contained in the huex10stprobeset.db   327
(exon-level annotations) and huex10sttranscriptcluster.db (gene-level annotations) R   328
packages available from Bioconductor [34, 35]. In both cases gene annotations were   329
provided for each feature.   330

Plink [36] was used for data wrangling and cleaning steps. The data wrangling steps 331
included updating probe IDs, unifying data to the positive strand, and updating   332

locations to GRCh37. The data cleaning steps included a step to filter for variant and   333

subject missingness and minor alleles, one to filter variants with Hardy-Weinberg exact   334

test, and a step to remove unusual heterozygosity. Additionally, we used the   335

HRC-check-bin tool in order to carry out data wrangling steps required to make our   336

data compatible with the Haplotype Reference Consortium (HRC) panel   337

(`http://www.well.ox.ac.uk/~wrayner/tools/`). Having been prepared thusly, the   338

data were split by chromosome and pre-phased with SHAPEIT [37] using the 1000   339

Genomes phase 3 panel and converted to vcf format. These files were then submitted to   340

the Michigan Imputation Server   341

(`https://imputationserver.sph.umich.edu/start.html`) [31,32] for imputation   342

with the HRC version 1 panel [38]. We applied Matrix eQTL [39] to the normalized   343

expression and imputed genotype data to generate prior eQTLs for our heritability   344

analysis.   345

Partitioning local and distal heritability of gene expression   346

Motivated by the observed differences in regulatory effect sizes of variants located in the   347

vicinity of the genes and distal to the gene, we partitioned the proportion of gene   348

expression variance explained by SNPs in the DGN cohort into two components: local   349

(SNPs within 1Mb of the gene) and distal (eQTLs on non-gene chromosomes) as defined   350

by the GENCODE [40] version 12 gene annotation. We calculated the proportion of the   351

variance (narrow-sense heritability) explained by each component using the following   352

mixed-effects model: 353

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

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   354

identity matrix, we calculated the total variability explained by local and distal   355

components by estimating σ_w^2 with restricted maximum likelihood (REML) using   356

GCTA software [24]. For heritability analyses in the GTEx cohort, we removed the   357

distal term from the model and only estimated marginal *local* h^2 due to the smaller   358

sample sizes of both cross-tissue and tissue-specific expression levels compared to DGN.   359

Approximate confidence intervals were computed as the point estimate $\pm$ 2 times the   360

estimated standard error. The intervals were also forced to be ≥ 0 or ≤ 1. Genes were   361
considered to have heritability significantly different from 0 if the confidence interval did   362
not include 0.   363

By default we restricted the heritability estimates to be in the 0 to 1 interval.   364
However, for the purpose of estimating the mean heritability (see Table 1 and S1 Table),   365
we performed separate runs allowing the heritability estimates to take negative values   366
with the `--reml-no-constrain` option in GCTA. Despite the lack of obvious biological   367
interpretation of a negative heritability, it is an accepted procedure used in order to   368
avoid bias in the estimated mean [10, 11].   369

Determining polygenicity versus sparsity using the elastic net   370

We used the glmnet package to fit an elastic net model where the tuning parameter is   371
chosen via 10 fold cross validation to maximize prediction performance measure by   372
Pearson's R^2 [41, 42].   373

The elastic net penalty is controlled by mixing parameter α, which spans LASSO   374
($\alpha = 1$, the default) [16] at one extreme and ridge regression ($\alpha = 0$) [17] at the other.   375
The ridge penalty shrinks the coefficients of correlated SNPs towards each other, while   376
the LASSO tends to pick one of the correlated SNPs and discard the others. Thus, an   377
optimal prediction R^2 for $\alpha = 0$ means the gene expression trait is highly polygenic,   378
while an optimal prediction R^2 for $\alpha = 1$ means the trait is highly sparse.   379

In the DGN cohort, we tested 21 values of the mixing parameter   380
($\alpha = 0, 0.05, 0.1, ..., 0.90, 0.95, 1$) for optimal prediction of gene expression of the 341   381
genes on chromosome 22. For the rest of the autosomes in DGN and for whole tissue,   382
cross-tissue, and tissue-specific expression in the GTEx cohort, we tested   383
$\alpha = 0.05, 0.5, 0.95, 1$.   384

Quantifying sparsity with Bayesian Sparse Linear Mixed Models (BSLMM)   385, 386

We used BSLMM [20] to model the effect of local genetic variation (SNPs within 1 Mb   387
of gene) on the genetic architecture of gene expression. The BSLMM is a linear model   388
with a polygenic component (small effects) and a sparse component (large effects)   389

enforced by sparsity inducing priors on the regression coefficients [20]. We used the   390

software GEMMA [43] to implement BSLMM for each gene using the following   391

parameters: 392

```
gemma -g [genoFile] -p [expFile] -a [snpFile] -bslmm 1 -s 100000 -o [out]
```

The `-bslmm 1` option specifies a linear BSLMM and the `-s 100000` option specifies   393

the number of sampling steps per gene. The BSLMM estimates the PVE (the total   394

proportion of variance in phenotype explained by the sparse effects and random effects   395

terms together) and PGE (the proportion of genetic variance explained by the sparse   396

effects terms). From the second half of the sampling iterations for each gene, we report   397

the median and the 95% credible sets of the PVE, PGE, and the $|\gamma|$ parameter (the   398

number of SNPs with non-zero coefficients).   399

Orthogonal tissue decomposition   400

To better understand the context specificity of gene expression regulation, we developed   401

a method called orthogonal tissue decomposition (OTD). This approach is an extension   402

of our method to develop an intrinsic growth phenotype [44]. We applied OTD to   403

GTEx Project [23] data and decomposed the expression of each gene into cross-tissue   404

and tissue-specific components. The tissue availability is unbalanced across individuals   405

because of the difficulties of sample collection and the uneven quality of the tissues.   406

OTD decomposes the expression traits into orthogonal components as represented by   407

the following model: 408

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

Specifically, to generate cross-tissue and tissue-specific expression levels, we used the   409

`lmer` function in the R [45] package `lme4` [46] to fit the following mixed-effects model: 410

```
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 [40] version 18) with a mean gene expression level across tissues greater than 0.1 RPKM (reads per kilobase of transcript per million reads mapped) and RPKM > 0 in at least 3 individuals 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` [29] 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 [28] performed with a subset of the GTEx data (maximum n=175 in the nine tissues of the pilot phase, see [23]), 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.

Acknowledgments

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.,

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. PLoS Genetics. 2010;6(4):e1000888. Available from:
   `http://dx.doi.org/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. PLoS Genetics. 2010;6(4):e1000895.
   Available from: `http://dx.doi.org/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.
   2014;95(5):535–552. Available from:
   `http://dx.doi.org/10.1016/j.ajhg.2014.10.004`.

4. Torres JM, Gamazon ER, Parra EJ, Below JE, Valladares-Salgado A, Wacher N,
   et al. Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a
   complex trait. The American Journal of Human Genetics. 2014;95(5):521–534.

5. Davis LK, Yu D, Keenan CL, Gamazon ER, Konkashbaev AI, Derks EM, et al.
   Partitioning the heritability of Tourette syndrome and obsessive compulsive
   disorder reveals differences in genetic architecture. PLoS Genet. 2013;.

6. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and
   disease. Nat Rev Genet. 2015;16(4):197–212. Available from:
   `http://dx.doi.org/10.1038/nrg3891`.

7. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al.
   Patterns of Cis Regulatory Variation in Diverse Human Populations. PLoS

Genetics. 2012;8(4):e1002639. Available from:
`http://dx.doi.org/10.1371/journal.pgen.1002639`.

8. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al.
Population genomics of human gene expression. Nature Genetics.
2007;39(10):1217–1224. Available from: `http://dx.doi.org/10.1038/ng2142`.

9. 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. PLoS Genetics.
2011;7(5):e1002078. Available from:
`http://dx.doi.org/10.1371/journal.pgen.1002078`.

10. 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. 2014
apr;46(5):430–437. Available from: `http://dx.doi.org/10.1038/ng.2951`.

11. 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. PLoS Genetics.
2011;7(2):e1001317. Available from:
`http://dx.doi.org/10.1371/journal.pgen.1001317`.

12. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF,
et al. Common polygenic variation contributes to risk of schizophrenia and
bipolar disorder. Nature. 2009;Available from:
`http://dx.doi.org/10.1038/nature08185`.

13. 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. 2012;44(5):483–489. Available from:
`http://dx.doi.org/10.1038/ng.2232`.

14. 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.
2012;44(9):981–990. Available from: `http://dx.doi.org/10.1038/ng.2383`.

15. 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. 2015;47(9):1091–1098. Available from: http://dx.doi.org/10.1038/ng.3367.

16. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological). 2015;58(1):267–288. Available from: http://www.jstor.org/stable/2346178.

17. Hoerl AE, Kennard RW. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics. 1970;12(1):69–82. Available from: http://dx.doi.org/10.1080/00401706.1970.10488635.

18. de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010;11(12):880–886. Available from: http://dx.doi.org/10.1038/nrg2898.

19. Wheeler HE, Aquino-Michaels K, Gamazon ER, Trubetskoy VV, Dolan ME, Huang RS, et al. Poly-Omic Prediction of Complex Traits: OmicKriging. Genetic Epidemiology. 2014;38(5):402–415. Available from: http://dx.doi.org/10.1002/gepi.21808.

20. Zhou X, Carbonetto P, Stephens M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genetics. 2013;9(2):e1003264. Available from: http://dx.doi.org/10.1371/journal.pgen.1003264.

21. 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. 2005;437(7063):1365–1369. Available from: http://dx.doi.org/10.1038/nature04244.

22. 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. 2013;24(1):14–24. Available from: http://dx.doi.org/10.1101/gr.155192.113.

23. 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. 2015;348(6235):648–660. Available from: http://dx.doi.org/10.1126/science.1262110.

24. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics. 2011;88(1):76–82. Available from: http://dx.doi.org/10.1016/j.ajhg.2010.11.011.

25. Pierce BL, Tong L, Chen LS, Rahaman R, Argos M, Jasmine F, et al. Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians. PLoS genetics. 2014;10(12):e1004818.

26. 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. 2015;47(4):345–352. Available from: http://dx.doi.org/10.1038/ng.3220.

27. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;67(2):301–320. Available from: http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.

28. Flutre T, Wen X, Pritchard J, Stephens M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. PLoS Genetics. 2013;9(5):e1003486. Available from: http://dx.doi.org/10.1371/journal.pgen.1003486.

29. 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. 2012;7(3):500–507. Available from: http://dx.doi.org/10.1038/nprot.2011.457.

30. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–1735.

31. 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. 2012;44(8):955–959. Available from: http://dx.doi.org/10.1038/ng.2354.

32. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics. 2014;31(5):782–784. Available from: http://dx.doi.org/10.1093/bioinformatics/btu704.

33. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic acids research. 2003;31(4):e15–e15.

34. MacDonald JW. huex10stprobeset.db: Affymetrix huex10 annotation data (chip huex10stprobeset). R package;version 8.3.1.

35. MacDonald JW. huex10sttranscriptcluster.db: Affymetrix huex10 annotation data (chip huex10sttranscriptcluster). R package;version 8.3.1.

36. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. ArXiv e-prints. 2014;.

37. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nature methods. 2012;9(2):179–181.

38. McCarthy S, Das S, Kretzschmar W, Durbin R, Abecasis G, Marchini J. A reference panel of 64,976 haplotypes for genotype imputation. bioRxiv. 2015;p. 035170.

39. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353–1358.

40. 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. 2012;22(9):1760–1774. Available from: http://dx.doi.org/10.1101/gr.135350.111.

41. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):1–22. Available from: `http://www.jstatsoft.org/v33/i01/`.

42. 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(5):1–13. Available from: `http://www.jstatsoft.org/v39/i05/`.

43. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. 2012;44(7):821–824. Available from: `http://dx.doi.org/10.1038/ng.2310`.

44. 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. PLoS Genetics. 2012;8(2):e1002525. Available from: `http://dx.doi.org/10.1371/journal.pgen.1002525`.

45. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015. Available from: `http://www.R-project.org/`.

46. Bates D, Mächler M, Bolker BM, Walker S. Fitting Linear Mixed-Effects Models using lme4. Journal of Statistical Software. 2015;67(1):1–48. Available from: `http://dx.doi.org/10.18637/jss.v067.i01`.

Supporting Information

S1 Fig

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 in the nine GTEx tissues with the largest sample sizes 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.

S2 Fig

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 all autosomal genes per tissue.

S3 Fig

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.

S4 Fig

GTEx whole tissue expression Bayesian Sparse Linear Mixed Model. 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.

S5 Fig

GTEx orthogonal tissue decomposition cross-tissue and tissue-specific expression Bayesian Sparse Linear Mixed Model. 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.

S6 Fig

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.

S7 Fig

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 all autosomal genes per tissue.

S1 Table

Estimates of cross-tissue and tissue-specific local h^2. Expression levels derived by Orthogonal Tissue Decomposition and h^2 estimated using the `--reml-no-constrain` method.
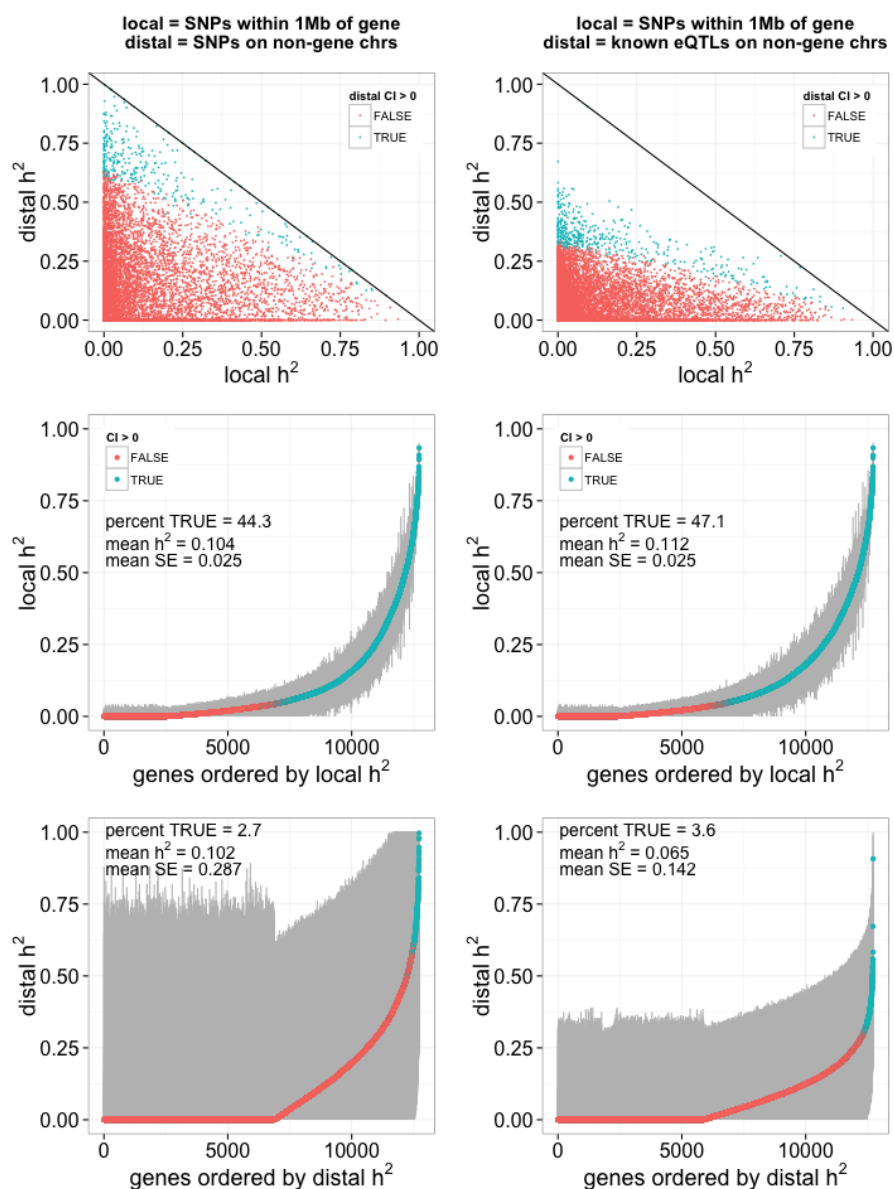
Figure 1. DGN whole blood expression joint heritability (h^2). Local (SNPs within 1 Mb of each gene) and distal (Left: SNPs on non-gene chromsomes. Right: SNPs that are eQTLs in the Framingham Heart Study on other chromosomes [FDR < 0.05]) h^2 for gene expression were jointly estimated. (**Top**) Distal 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.
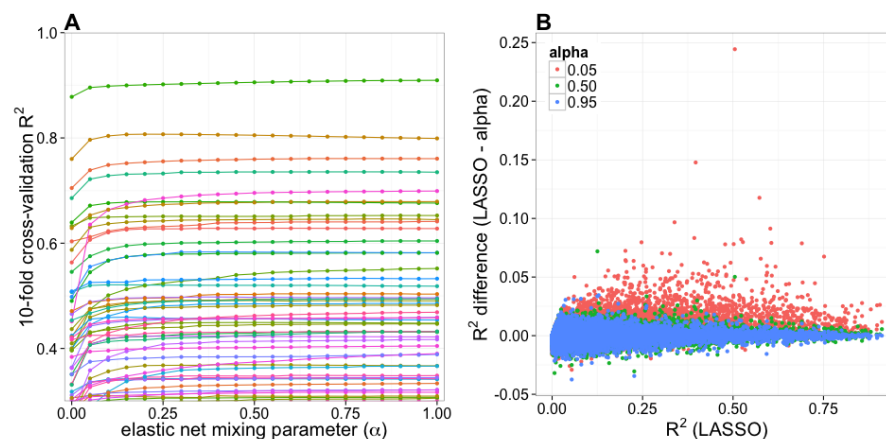
Figure 2. DGN cross-validated predictive performance across the elastic net. (**A**) 10-fold cross-validated R^2 of predicted vs. observed expression in DGN whole blood compared to a range of elastic net mixing parameters (α) for genes on chromosome 22 with $R^2 > 0.3$. (**B**) Predictive R^2 difference between LASSO ($\alpha = 1$) and several other values of α compared to LASSO predictive R^2 for all autosomal genes.
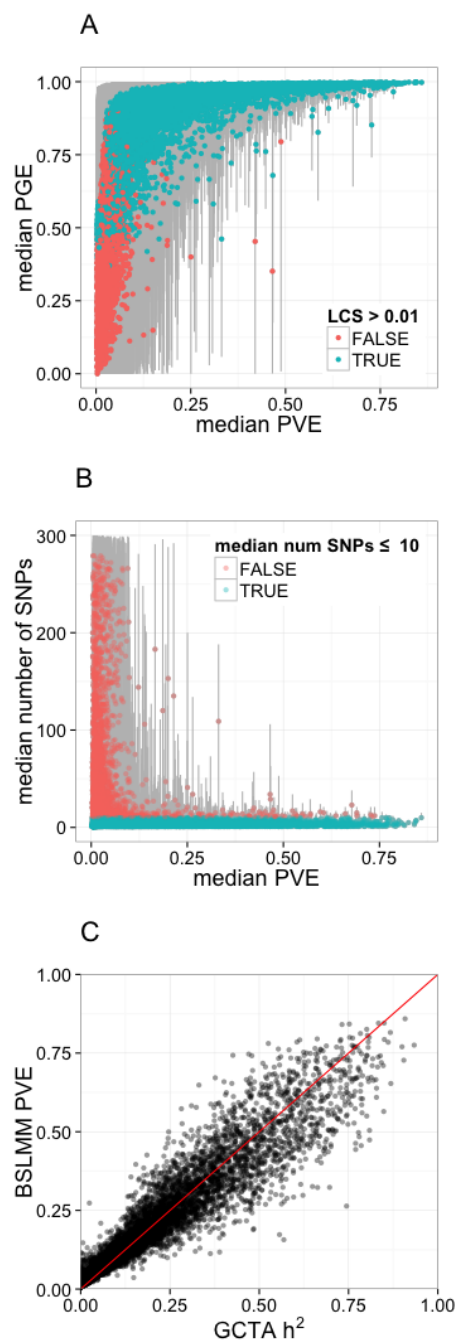
Figure 3. Bayesian Sparse Linear Mixed Models reveal the sparsity of gene expression architecture. (**A**) Comparison of median PGE (proportion of PVE explained by sparse effects) to median PVE (total proportion of variance explained) for expression of each gene. The 95% credible set of each PGE estimate is in gray and genes with a lower credible set (LCS) greater than 0.01 are in blue. (**B**) 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.(**C**) BSLMM-estimated PVE compared to GCTA-estimated heritability per gene (R=0.96).
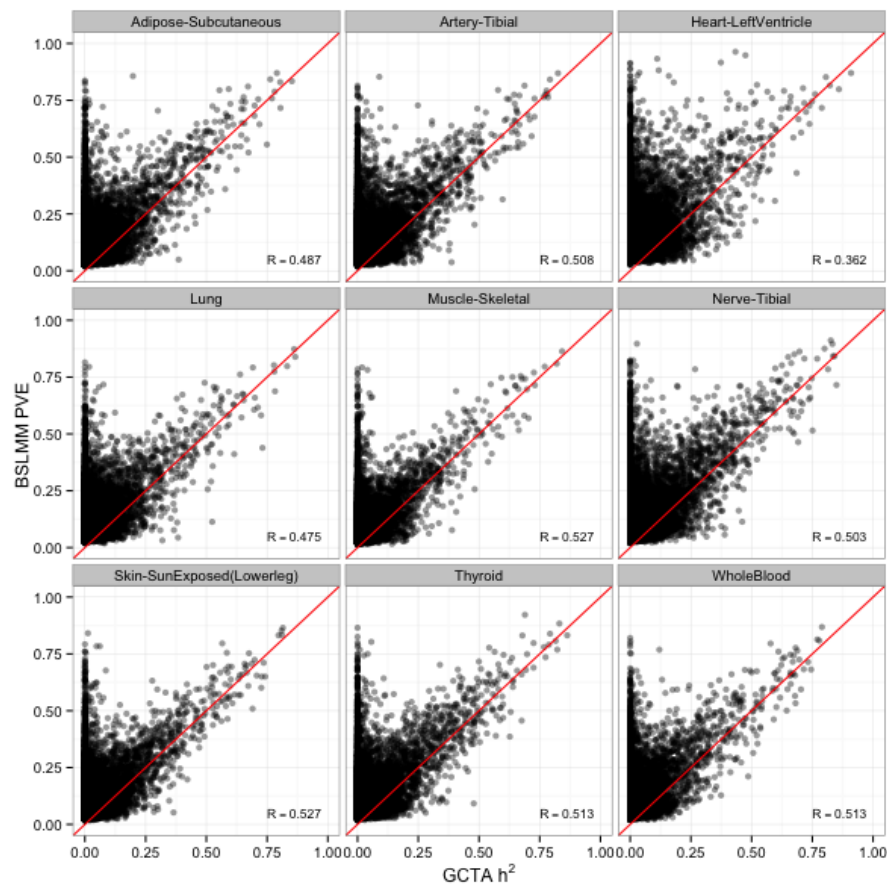
Figure 4. GTEx PVE vs. h^2. GTEx whole tissue expression BSLMM-estimated PVE (total proportion of variance explained) compared to GCTA-estimated heritability per gene. R = Pearson correlation.
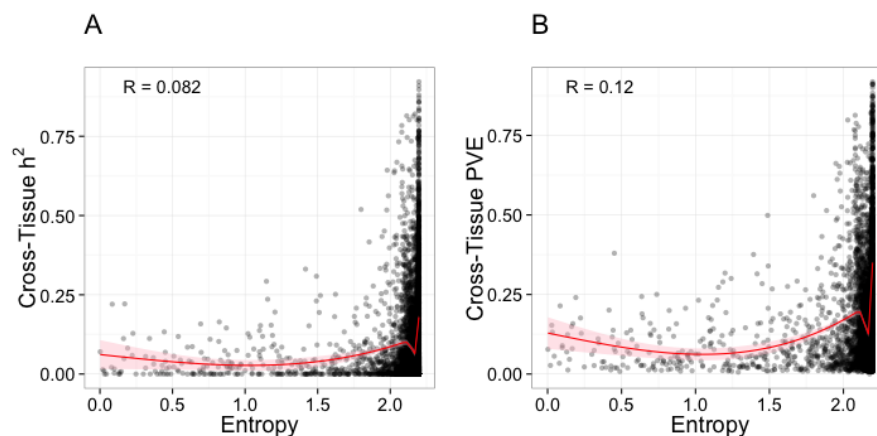

Figure 5. Cross-tissue expression from the orthogonal tissue decomposition (OTD) compared to multi-tissue eQTL results. Pearson correlation (R) between the entropy of the posterior probabilities from the Flutre et al. multi-tissue eQTL method and the estimates of (A) heritability and (B) PVE of cross-tissue gene expression derived from the OTD. The generalized additive model smoothing line is in red.
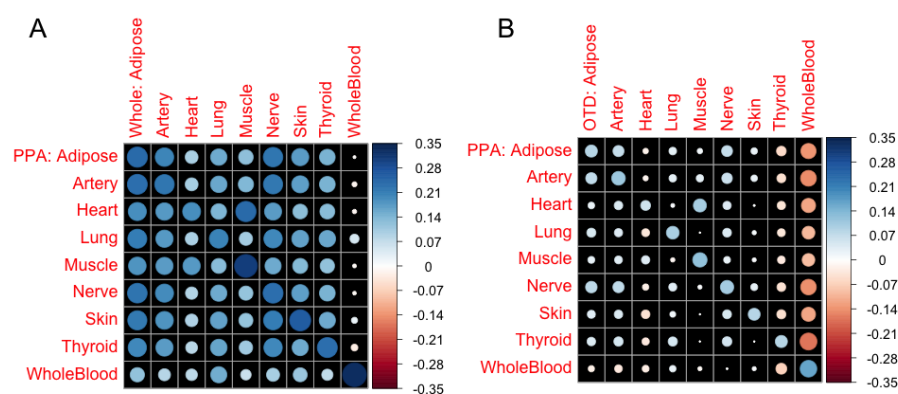
Figure 6. Comparison of whole tissue or tissue-specific expression PVE to the posterior probability of a gene being actively regulated in a tissue. Pearson correlation (R) between the posterior probability the top multi-tissue eQTL regulates its gene in a given tissue (PPA, Flutre et al. method) and the PVE of (A) whole tissue gene expression or (B) tissue-specific gene expression from the orthogonal tissue decomposition (OTD). Area of each circle is proportional to the absolute value of R.