Survey of the Heritability and Sparse Architecture of Gene Expression Traits Across Human Tissues

Short title: Heritability and Sparse Architecture of Gene Expression

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 Center for Research Informatics, 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 50% of expressed genes showing significant h^2 in DGN and 8-19% 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 <1% in GTEx). Thus, we focus on local regulation. Bayesian Sparse Linear Mixed Model (BSLMM) analysis provide strong evidence that local architecture of gene expression traits is sparse rather than polygenic across DGN and all 40 GTEx tissues examined. This result is further confirmed by the sparsity of optimal performing gene expression predictors via elastic net modeling. To further explore 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   1

Regulatory variation plays a key role in the genetics of complex traits [1–3]. Methods   2
that partition the contribution of environment and genetic components are useful tools   3
to understand the biology underlying complex traits. Partitioning heritability into   4
different functional classes has been successful in quantifying the contribution of   5
different mechanisms that drive the etiology of diseases [3–5].   6

   Most human expression quantitative trait loci (eQTL) studies have focused on how   7
local genetic variation affects gene expression in order to reduce the multiple testing   8
burden that would be required for a global analysis [6,7]. Furthermore, when both local   9
and distal eQTLs are reported [8–10], effect sizes and replicability are much higher for   10
local eQTLs. Indeed, while the heritability of gene expression attributable to local   11
genetic variation has been estimated accurately, large standard errors have prevented   12
accurate estimation of the contribution of distal genetic variation to gene expression   13
variation [10,11].   14

   While many common diseases are likely polygenic [12–14], it is unclear whether gene   15
expression levels are also polygenic or instead have simpler genetic architectures. It is   16
also unclear how much these expression architectures vary across genes [6].   17

   The relative prediction performance of sparse and polygenic models can provide   18

useful information about the underlying distribution of effect sizes. For example, if the 19
true model of a trait is polygenic, it is natural to expect that polygenic models will 20
predict better than sparse ones. We assessed the ability of various models, with different 21
underlying assumptions, to predict gene expression in order to both understand the 22
underlying genetic architecture of gene expression and to further optimize predictors for 23
our gene-level association method, PrediXcan [15]. When we calibrated the prediction 24
model that was used in the PrediXcan paper, we showed that sparse models such as 25
LASSO performed better than a polygenic score model. We also showed that a model 26
that uses the top eQTL variant outperformed the polygenic score but did not do as well 27
as LASSO or elastic net [15], suggesting that for many genes, the genetic architecture is 28
sparse, but not regulated by a single SNP. 29

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   58

Local genetic variation can be well characterized for all tissues   59

We estimated the local and distal heritability of gene expression levels in 40 tissues from   60
the GTEx consortium and whole blood from the Depression Genes and Networks   61
(DGN) cohort. The sample size in GTEx varied from 72 to 361 depending on the tissue,   62
while 922 samples were available in DGN [22]. We used mixed-effects models (see   63
Methods) and calculated variances using restricted maximum likelihood as implemented   64
in GCTA [24].   65

For the local heritability component, we used variants within 1Mb of the   66
transcription start and end of each protein coding gene, whereas for the distal   67
component, we used variants outside of the chromosome where the gene was located.   68
Different approaches to pick the set of distal variants were explored, but results were   69
robust to different selections. See more details in Methods.   70

Table 1 summarizes the local heritability estimate results across all tissues. In order   71
to obtain an unbiased estimates of mean h^2, we allow the values to be negative when   72
fitting the REML (unconstrained), as done previously [10, 11]. This approach reduces   73
the standard error of the estimated mean of heritability, especially important for the   74
distal component. Even though each individual gene's distal heritability is noisy,   75
averaging across all genes reduces the error substantially. For the DGN dataset, we were   76
able to estimate the mean distal h^2, which was 0.034 (SE = 0.0024). However for the   77
GTEx samples, the sample size was too small and the REML algorithm became   78
unstable when allowing for negative values. This numeric instability would cause only a   79

small number of genes with large positive (and noisy) heritability values to converge 80 biasing the mean value. For this reason we do not show mean distal heritability 81 estimates for GTEx tissues. 82

The left column of Fig. 1 shows the estimated local and distal h^2 from DGN. Even 83 though many genes show relatively large point estimates of distal h^2, only the ones 84 colored in blue are significantly different from zero ($P < 0.05$). The local component of 85 h^2 is relatively well estimated in DGN with 50% of genes (6399 out of 12719) showing 86 $P < 0.05$. In contrast, the distal heritability is significant for only 7.3% (931 out of 87 12719) of the genes ($P < 0.05$). This is not much larger than the expected number at 88 this significance level (5%) and the genes with $P < 0.05$ and negative h^2 in Fig. 1 are 89 obvious false positives, within the type 1 error rate. 90

It has been shown that local-eQTLs are more likely to be distal-eQTLs of target 91 genes [25]. Thus, we tested whether restricting the distal genetic similarity computation 92 to QTLs (as determined in the Framingham mRNA dataset of over 5000 individuals [26] 93 independent of the DGN and GTEx cohorts) for other genes could improve distal 94 heritability precision by prioritizing functional variants. We exclude eQTLs on the same 95 chromosome as the tested gene to avoid contaminating distal h^2 with cis associations. 96

While using functional priors (known eQTLs) to define distal h^2 decreased the mean 97 standard error of the heritability estimates across genes from 0.24 to 0.14, the number 98 of significant genes did not change dramatically (Fig. 1). Also, using the subset of 99 known eQTLs (from an independent source) in other chromosomes for computing distal 100 heritability reduced the mean value from 0.027 to 0.015. Therefore, while we gain some 101 power to detect significant distal heritability by using cis eQTL priors as indicated by 102 the standard error reduction, a good portion of the distal regulation is lost when using 103 only the smaller subset of known cis-eQTL variants. We used functional priors to 104 estimate distal h^2 in the GTEx cohort, but less than 1% of genes had a $P < 0.05$ (S1 105 Fig). Given the limited sample size we will focus on local regulation for the remainder 106 of the paper. 107

Table 1. Estimates of unconstrained local h^2 across genes within tissues.

tissue	n	mean h^2 (SE)	% $P < 0.05$	num $P < 0.05$	num expressed
DGN Whole Blood	922	0.143 (0.0015)	50.3	6399	12719
Cross-tissue	450	0.062 (0.0008)	35.4	6003	16951
Adipose - Subcutaneous	298	0.038 (0.0008)	16.9	2402	14205
Adrenal Gland	126	0.043 (0.0011)	11.4	1610	14150
Artery - Aorta	198	0.042 (0.0010)	16.6	2294	13844
Artery - Coronary	119	0.037 (0.0010)	10.5	1477	14127
Artery - Tibial	285	0.042 (0.0009)	17.3	2333	13504
Brain - Anterior cingulate cortex (BA24)	72	0.028 (0.0015)	10.6	1532	14515
Brain - Caudate (basal ganglia)	100	0.037 (0.0012)	10.5	1540	14632
Brain - Cerebellar Hemisphere	89	0.049 (0.0014)	13.3	1901	14295
Brain - Cerebellum	103	0.050 (0.0013)	13.5	1955	14491
Brain - Cortex	96	0.045 (0.0012)	10.7	1575	14689
Brain - Frontal Cortex (BA9)	92	0.038 (0.0013)	10.9	1581	14554
Brain - Hippocampus	81	0.037 (0.0013)	9.8	1422	14513
Brain - Hypothalamus	81	0.017 (0.0013)	9.9	1460	14759
Brain - Nucleus accumbens (basal ganglia)	93	0.029 (0.0012)	10.2	1486	14601
Brain - Putamen (basal ganglia)	82	0.032 (0.0013)	10.1	1456	14404
Breast - Mammary Tissue	183	0.029 (0.0008)	11.7	1714	14700
Cells - EBV-transformed lymphocytes	115	0.058 (0.0013)	11.6	1448	12454
Cells - Transformed fibroblasts	272	0.051 (0.0010)	19.0	2420	12756
Colon - Sigmoid	124	0.033 (0.0011)	12.3	1760	14321
Colon - Transverse	170	0.036 (0.0009)	12.5	1832	14676
Esophagus - Gastroesophageal Junction	127	0.032 (0.0010)	11.6	1638	14125
Esophagus - Mucosa	242	0.042 (0.0009)	17.4	2476	14239
Esophagus - Muscularis	219	0.039 (0.0009)	16.8	2355	14047
Heart - Atrial Appendage	159	0.042 (0.0010)	12.9	1787	13892
Heart - Left Ventricle	190	0.034 (0.0009)	14.7	1960	13321
Liver	98	0.033 (0.0012)	10.0	1350	13553
Lung	279	0.032 (0.0007)	15.7	2315	14775
Muscle - Skeletal	361	0.033 (0.0007)	17.0	2180	12833
Nerve - Tibial	256	0.052 (0.0009)	18.8	2724	14510
Ovary	85	0.037 (0.0013)	8.5	1194	14094
Pancreas	150	0.047 (0.0011)	14.0	1954	13941
Pituitary	87	0.038 (0.0013)	10.7	1621	15183
Skin - Not Sun Exposed (Suprapubic)	196	0.041 (0.0009)	13.4	1966	14642
Skin - Sun Exposed (Lower leg)	303	0.039 (0.0008)	17.7	2589	14625
Small Intestine - Terminal Ileum	77	0.036 (0.0014)	9.0	1341	14860
Spleen	89	0.059 (0.0014)	10.4	1508	14449
Stomach	171	0.032 (0.0009)	12.0	1747	14531
Testis	157	0.054 (0.0010)	16.5	2792	16936
Thyroid	279	0.044 (0.0009)	18.2	2670	14642
Whole Blood	339	0.033 (0.0008)	16.1	1956	12160

Except for DGN Whole Blood, all tissues are from the GTEx Project. Cross-tissue uses derived expression levels from our orthogonal tissue decomposition (OTD) of GTEx data. Mean heritability (h^2) and the standard error of the mean (SE) are calculated across genes for each tissue. The percentage (%) and number (num) of genes with significant h^2 estimates ($P < 0.05$) and the number of genes expressed in each tissue are also reported.
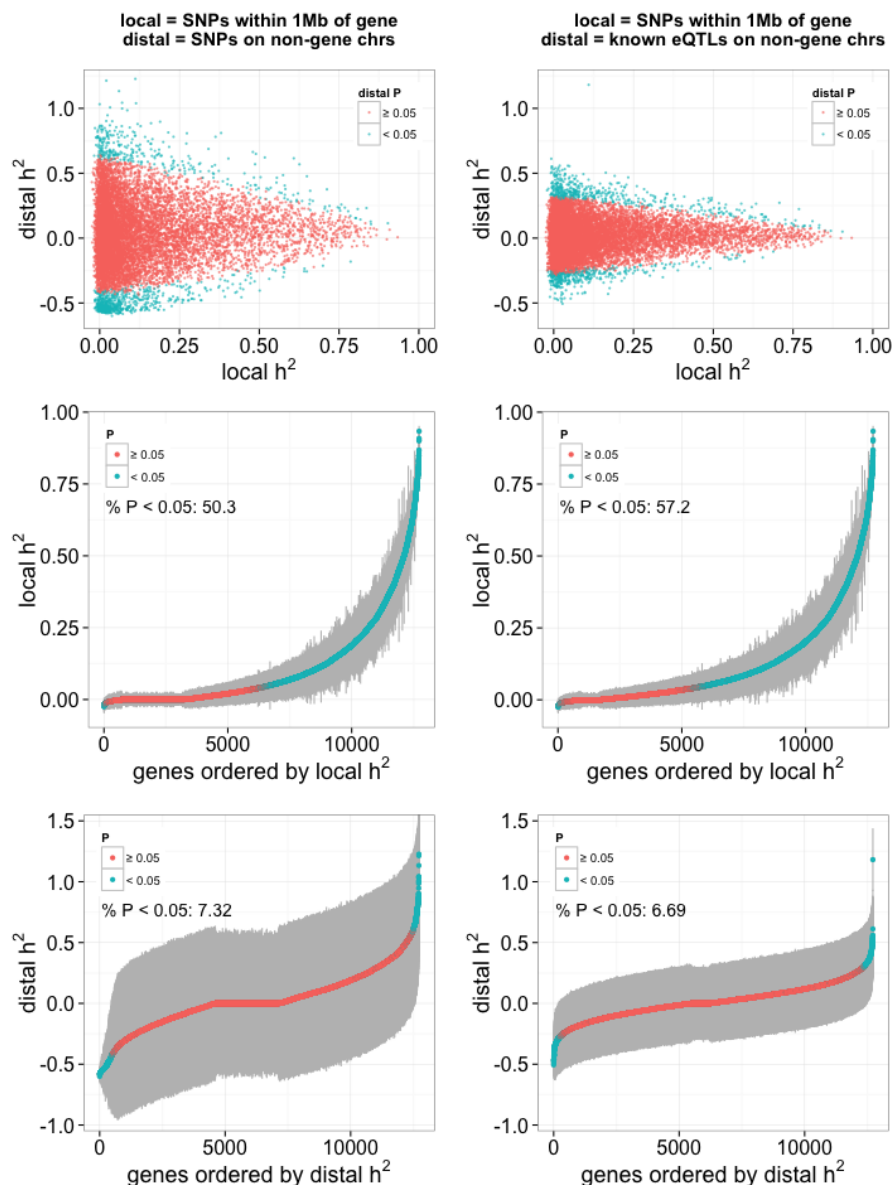
Figure 1. DGN whole blood expression local and distal heritability (h^2).
This figure shows the local and distal heritability of gene expression levels in whole
blood from the DGN RNA-seq dataset. In order to obtain an unbiased estimates of
mean h^2, we allow the values to be negative when fitting the REML (unconstrained).
Notice that only a few genes have distal heritability that is significantly different from 0
($P < 0.05$). Local was defined as 1Mb from each gene. For the left side panel, distal
heritability was computed using all SNPs outside of the gene's chromosome. On the the
right side, distal heritability was computing using SNPs that were cis-eQTLs in the
Framingham study. (**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. As a measure of uncertainty, we have added two times the standard
errors of each h^2 estimate in gray segments. Genes with significant h^2 ($P < 0.05$) are
shown in blue. To be conservative, we set $h^2 = 0$ when GCTA did not converge. Genes
in blue with negative h^2 are false positives, within the type 1 error (5%).

Sparse local architecture implied by sparsity of best prediction models

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

More specifically, we performed 10-fold cross-validation using the elastic net [27] 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. 2). An $\alpha = 0.05$ was also clearly suboptimal for gene expression prediction in the GTEx tissues, while models with $\alpha = 0.5$ 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.
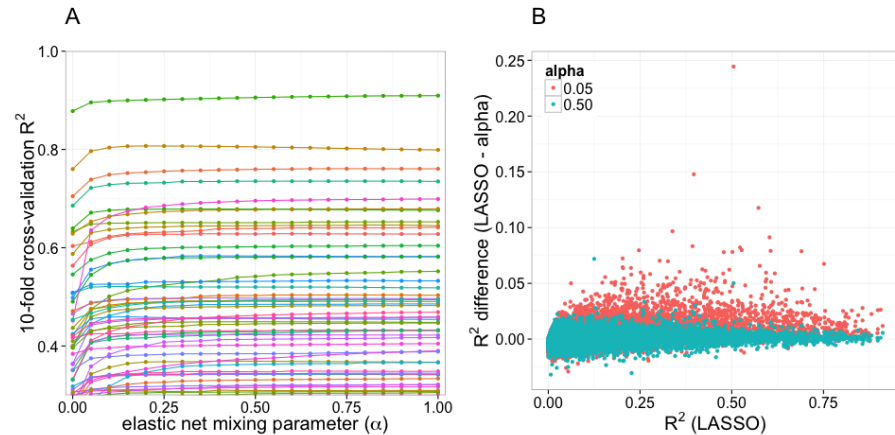
Figure 2. DGN cross-validated predictive performance across the elastic net. This figure shows the cross validated R^2 between observed and predicted expression levels using elastic net prediction models in DGN. (**A**) This panel shows the 10-fold cross validated R^2 for 51 genes with $R^2 > 0.3$ from chromosome 22 as a function of the elastic net mixing parameters (α). Smaller mixing parameters correspond to more polygenic models while larger ones correspond to more sparse models. Each line represents a gene. The performance is in general flat for most values of the mixing parameter except very close to zero where it shows a pronounced dip. Thus polygenic models perform more poorly than sparse models. (**B**) This panel shows the difference between the cross validated R^2 of the LASSO model and the elastic net model mixing parameters 0.05 and 0.5 for autosomal protein coding genes. Elastic net with $\alpha = 0.5$ values hover around zero, meaning that it has similar predictive performance to LASSO. The R^2 difference of the more polygenic model (elastic net with $\alpha = 0.05$) is mostly above the 0 line, indicating that this model performs worse than the LASSO model.

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   130
BSLMM [20] approach, which models the genetic contribution as the sum of a sparse   131
and a polygenic component. The parameter PGE in this model represents the   132
proportion of genetic variance explained by sparse effects. Another parameter, the total   133
variance explained (PVE) by additive genetic variants, is a more flexible Bayesian   134
equivalent of the chip heritability we have estimated using a linear mixed model (LMM)   135
as implemented in GCTA.   136

As anticipated, we find that for highly heritable genes, the sparse component is large.   137
For example, all genes with PVE > 0.50 had PGE > 0.82 and their median PGE was   138
0.989 (Fig. 3A). The median PGE for genes with PVE > 0.1 was 0.949. Fittingly, for   139
most (96.3%) of the genes with PVE estimates > 0.10, the median number of SNPs   140
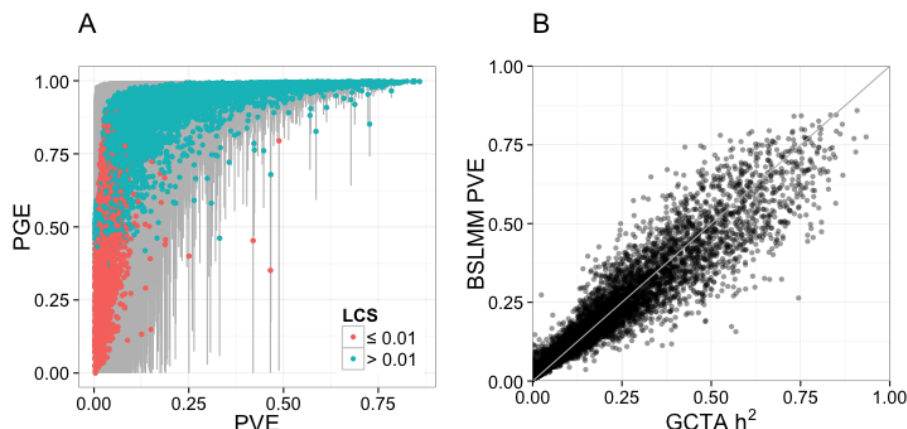
included in the model was no more than 10. ₁₄₁

Figure 3. Sparsity estimates using Bayesian Sparse Linear Mixed Models.
(**A**) This panel shows a measure of sparsity of the gene expression traits represented by
the PGE parameter from the BSLMM approach. PGE is the proportion of the sparse
component of the total variance explained by genetic variants, PVE (the BSLMM
equivalent of h^2). The median of the posterior samples of BSLMM output is used as
estimates of these parameters. Genes with a lower credible set (LCS) > 0.01 are shown
in blue and the rest in red. The 95% credible set of each estimate is shown in gray. For
highly heritable genes the sparse component is close to 1, thus for high heritability
genes the local architecture is sparse. For lower heritability genes, there is not enough
evidence to determine sparsity or polygenicity. (**B**) This panel shows the heritability
estimate from BSLMM (PVE) vs the estimates from GCTA, which are found to be
similar (R=0.96).

BSLMM outperforms LMM in estimating h^2 for small samples ₁₄₂

Also as expected, we find that when the sample size is large enough, such as in DGN, ₁₄₃

there is a strong correlation between BSLMM-estimated PVE and GCTA-estimated h^2 ₁₄₄

(Fig. 3B, R=0.96). In contrast, when we applied BSLMM to the GTEx data, we found ₁₄₅

that many genes had measurably larger BSLMM-estimated PVE than GCTA-estimated ₁₄₆

h^2 (Fig. 4). This is further confirmation of the local sparse architecture of gene ₁₄₇

expression traits: the underlying assumption in the GCTA (LMM) approach to estimate ₁₄₈

heritability is that the genetic effect sizes are normally distributed, i.e. most variants ₁₄₉

have small effect sizes. LMM is quite robust to departure from this assumption, but ₁₅₀

only when the sample size is rather large. For the relatively small sample sizes in GTEx ₁₅₁

($n \leq 361$), we found that a model that directly addresses the sparse component such as ₁₅₂

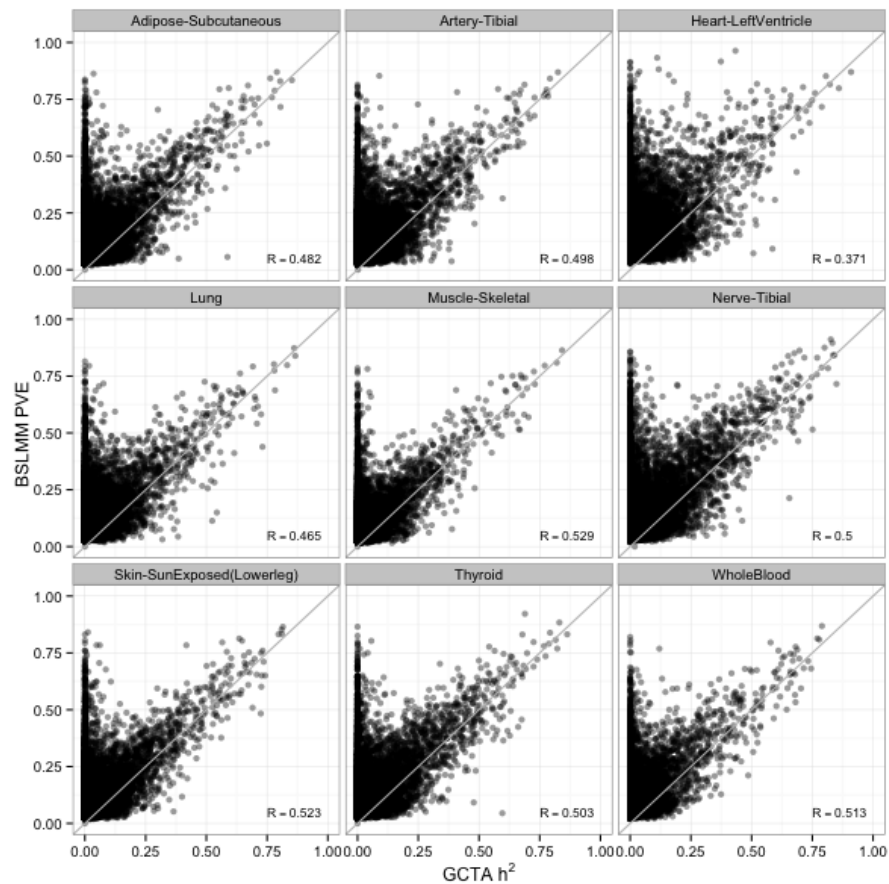BSLMM outperforms GCTA for estimating h^2. ₁₅₃

Figure 4. BSLMM vs LMM estimates of heritability in GTEx. This figure shows the comparison between estimates of heritability using BSLMM vs LMM for GTEx data. For most genes BSLMM estimates are larger than LMM estimates reflecting the fact that BSLMM yields better estimates of heritability because of its ability to account for the sparse component. R = Pearson correlation.

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 [23], we sought to decompose the expression levels into a component that is common across all tissues and tissue-specific components. For this we use a linear mixed effects model with a person-level random effect. See details in Methods. We use the posterior mean of this random effect as an estimate of the cross tissue component. We consider the residual component of this model as the tissue specific component. Below we describe the properties of these derived phenotypes.

We call this approach orthogonal tissue decomposition (OTD) because the

cross-tissue and tissue-specific components are assumed to be independent in the model. 164
The decomposition is applied at the expression trait level so that the downstream 165
genetic regulation analysis is performed separately for each derived trait, cross-tissue 166
and tissue-specific expression, which greatly reduces computational burden. For all the 167
derived phenotypes, one cross-tissue and 40 tissue-specific ones, we computed the local 168
heritability and generated prediction models. 169

Cross-tissue expression phenotype is less noisy and shows 170
higher predictive performance 171

Our estimates of h^2 for cross tissue expression traits are larger than the corresponding 172
estimates for each whole tissue expression traits (S3 Fig). This is due to the fact that 173
our OTD approach increases the ratio of genetically regulated component to noise by 174
averaging across multiple tissues. In addition to the increased h^2 we observe reduction 175
in standard errors of the estimated h^2. This is partly due to the increased h^2 – higher 176
h^2 are better estimated – but also due to the larger effective sample size for cross tissue 177
phenotypes. There were 450 samples for which cross tissue traits were available whereas 178
the maximum sample size for whole tissue phenotypes was 362. As consequence of this 179
increased h^2 and decreased standard errors, the percentage of cross h^2 estimates with 180
$P < 0.05$ was 35.4% whereas for whole tissue expression traits they ranged from 181
8.5-19.0% (Table 1). Similarly, cross-tissue BSLMM PVE estimates had lower error 182
than whole tissue PVE (S4 Fig, S5 Fig). 183

As for the tissue-specific components, the cross-tissue heritability estimates were also 184
larger and the standard errors were smaller reflecting the fact that a substantial portion 185
of regulation is common across tissues (S6 Fig). The percentage of GCTA h^2 estimates 186
with $P < 0.05$ was much larger for cross-tissue expression (35.4%) than the 187
tissue-specific expressions (7.6-17.7%, S1 Table). Similarly, the percentage of BSLMM 188
PVE estimates with a lower credible set greater than 0.01 was 49% for cross-tissue 189
expression, but ranged from 24-27% for tissue-specific expression (S5 Fig). 190

Cross-tissue predictive performance exceeded that of both tissue-specific and whole 191
tissue expression as indicated by higher cross-validated R^2 (S2 Fig, S7 Fig). Like whole 192
tissue expression, cross-tissue and tissue-specific expression showed better predictive 193

performance when using more sparse models. In other words elastic-net models with   194

$\alpha \geq 0.5$ predicted better than the ones with $\alpha = 0.05$ (S7 Fig).   195

Cross Tissue expression phenotype recapitulates published   196
multi-tissue eQTL results   197

To verify that the cross tissue phenotype has the properties we expect, we compared our   198
OTD results to those from a joint multi-tissue eQTL analysis [28], which was previously   199
performed on a subset of the GTEx data [23] covering 9 tissues. In particular, we used   200
the posterior probability of a gene being actively regulated (PPA) in a tissue. These   201
analysis results are available on the GTEx portal.   202

First, we reasoned that genes with high cross tissue h^2 would be actively regulated   203
in most tissues so that the PPA of a gene would be roughly uniform across tissues. By   204
contrast, a gene with tissue specific regulation would have concentrated posterior   205
probability in one or a few tissues. Thus we decided to define a measure of uniformity of   206
the posterior probability vector across the 9 tissues using the concept of entropy. More   207
specifically, for each gene we normalized the vector of posterior probabilities so that the   208
sum equaled 1. Then we applied the usual entropy definition (negative of the sum of the   209
log of the posterior probabilities weighted by the same probabilities, see Methods). In   210
other words, we defined a uniformity statistic that combines the nine posterior   211
probabilities into one value such that higher values mean the gene regulation is more   212
uniform across all nine tissues, rather than in just a small subset of the nine.   213

Thus we expected that genes with high cross tissue heritability, i.e. large cross tissue   214
regulation would show high probability of being active in multiple tissues, thus high   215
uniformity measure. Reassuringly, this is exactly what we find. Figure 5 shows that   216
genes with high cross tissue heritability concentrate on the higher end of the uniformity   217
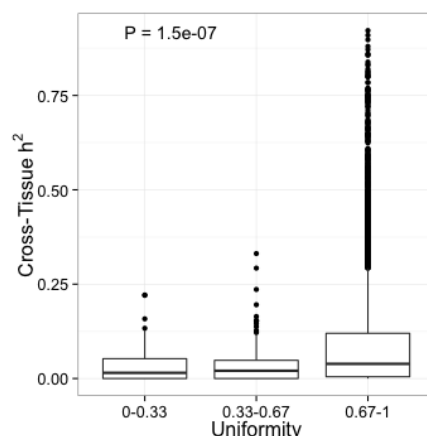measure.   218

Figure 5. Measure of uniformity of the posterior probability of active regulation vs. cross-tissue heritability. This figure shows the distribution of heritability of the cross-tissue component vs. a measure of uniformity of genetic regulation across tissues. The measure of uniformity was computed using the posterior probability of a gene being actively regulated in a tissue, PPA, from the Flutre et al. [28] multi-tissue eQTL analysis. Genes with PPA concentrated in one tissue were assigned small values of the uniformity measure whereas genes with PPA uniformly distributed across tissues were assigned high value of uniformity measure. See Methods for the entropy-based definition of uniformity.

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. This is confirmed in Figure 6A where PPA in each tissue is correlated with the BSLMM PVE of the expression in that tissue. In the off diagonal elements we observe high correlation between tissues, which was expected given that large portion of the regulation has been shown to be common across tissues. Whole blood has the lowest correlation consistent with whole blood clustering aways from other tissues [23]. In contrast, panel B of Figure 6 shows that the tissue specific expression PVE correlates well with matching tissue PPA but the off diagonal correlations are substantially reduced consistent with these phenotypes representing tissue specific components. Again whole blood shows a negative correlation which could be indicative of some over correction of the cross tissue component. Overall these results indicate that the cross tissue and tissue-specific phenotypes have properties that are consistent with the intended decomposition.
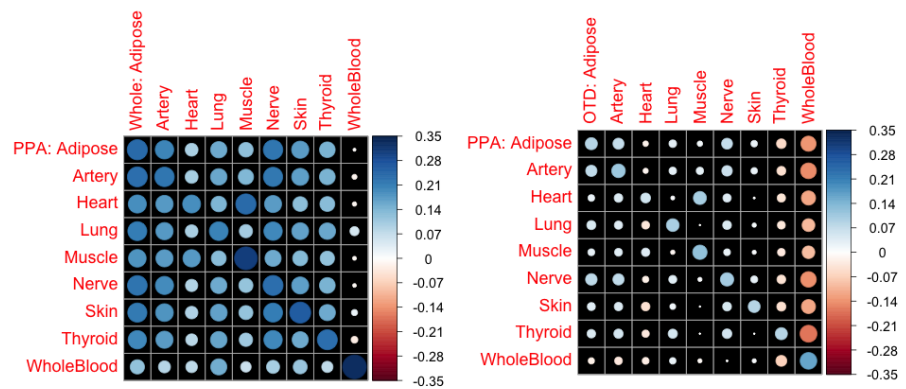
Figure 6. Comparison of heritability of whole tissue or tissue specific components vs. PPA. Panel (**A**) of this figure shows the Pearson correlation (R) between the BSLMM PVE of the original (we are calling whole here) tissue expression levels vs. the probability of the tissue being actively regulated in a given tissue (PPA). Matching tissues show, in general, the largest correlation values but most of the off diagonal correlations are also relatively high consistent with the shared regulation across tissues. Panel (**B**) shows the Pearson correlation between the PVE of the tissue-specific component of expression via orthogonal tissue decomposition (OTD) vs. PPA. Correlations are in general lower but matching tissues show the largest correlation. Off diagonal correlations are reduced substantially consistent with properties that are specific to each tissue. Area of each circle is proportional to the absolute value of R.

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 $\quad$ 247
preprocessing of the expression trait data we used. Indeed, PEER [29], SVA [30], and $\quad$ 248
other types of hidden confounder corrections have been shown to increase local eQTL $\quad$ 249
replicability, but their consequences on distal regulation is not well understood. As $\quad$ 250
larger sample sizes become available, we will test this hypothesis in GTEx data by $\quad$ 251
computing the distal h^2 without PEER factor correction. $\quad$ 252

We showed that restricting distal variants to known functional variants such as $\quad$ 253
eQTL data from independent studies improves the precision of distal heritability $\quad$ 254
estimates, but also reduces mean distal heritability estimates by half. $\quad$ 255

Using results implied by the improved predictive performance of sparse models and $\quad$ 256
by directly estimating sparsity using BSLMM (Bayesian Sparse Linear Mixed Model), $\quad$ 257
we show evidence that for highly heritable genes, local regulation is sparse across all the $\quad$ 258
tissues analyzed here. For genes with moderate and low heritability the evidence is not $\quad$ 259
as strong, but results are consistent with a sparse local architecture. Better methods to $\quad$ 260
correct for hidden confounders that do not dilute distal signals and larger sample sizes $\quad$ 261
will be needed to determine the properties of distal regulation. $\quad$ 262

Given that a substantial portion of local regulation is shared across tissues, we $\quad$ 263
propose here to decompose the expression traits into cross-tissue and tissue-specific $\quad$ 264
components. This approach, called orthogonal tissue decomposition, aims to decouple $\quad$ 265
the shared regulation from the tissue-specific regulation. We examined the genetic $\quad$ 266
architecture of these derived traits and find that they follow similar patterns to the $\quad$ 267
original whole tissue expression traits. The cross-tissue component benefits from an $\quad$ 268
effectively larger sample size than any individual tissue trait, which is reflected in more $\quad$ 269
accurate heritability estimates and consistently better prediction performance. $\quad$ 270
Encouragingly, we find that genes with high cross tissue heritability tend to be regulated $\quad$ 271
more uniformly across tissues. As for the tissue-specific expression traits, we found that $\quad$ 272
they recapitulate correlation with the vector of probability of tissue-specific regulation. $\quad$ 273

Prediction models of these decoupled expression traits will be useful to interpret the $\quad$ 274
association results from PrediXcan [15]. We expect results from the cross-tissue models $\quad$ 275
to relate to mechanisms that are shared across multiple tissues whereas results from the $\quad$ 276
tissue-specific models will inform us about the context specific mechanisms. $\quad$ 277

In this paper, we quantitate the genetic architecture of gene expression and develop $\quad$ 278

predictors across tissues. We show that local heritability can be accurately estimated   279
across tissues, but distal heritability cannot be reliably estimated at current sample   280
sizes. Using two different approaches, the elastic net and BSLMM, we show that for   281
local gene regulation, the genetic architecture is mostly sparse rather than polygenic.   282
Using new expression phenotypes generated in our OTD model, we show that   283
cross-tissue predictive performance exceeded that of both tissue-specific and whole   284
tissue expression as indicated by higher elastic net cross-validated R^2. Predictors,   285
heritability estimates and cross-validation statistics generated in this study of gene   286
expression architecture have been added to our PredictDB database   287
(`https://github.com/hakyimlab/PrediXcan`) for use in future studies of complex   288
trait genetics.   289

Materials and Methods   290

Genomic and Transcriptomic Data   291

DGN Dataset. We obtained whole blood RNA-seq and genome-wide genotype data   292
for 922 individuals from the Depression Genes and Networks (DGN) cohort [22], all of   293
European ancestry. For our analyses, we used the HCP (hidden covariates with prior)   294
normalized gene-level expression data used for the *trans*-eQTL analysis in Battle et   295
al. [22] and downloaded from the NIMH repository. The 922 individuals were unrelated   296
(all pairwise $\hat{\pi} < 0.05$) and thus all included in downstream analyses. Imputation of   297
approximately 650K input SNPs (minor allele frequency [MAF] > 0.05,   298
Hardy-Weinberg Equilibrium [P > 0.05], non-ambiguous strand [no A/T or C/G SNPs])   299
was performed on the Michigan Imputation Server   300
(`https://imputationserver.sph.umich.edu/start.html`) [31,32] with the following   301
parameters: 1000G Phase 1 v3 ShapeIt2 (no singletons) reference panel, SHAPEIT   302
phasing, and EUR population. Approximately 1.9M non-ambiguous strand SNPs with   303
MAF > 0.05, imputation $R^2 > 0.8$ and, to reduce computational burden, inclusion in   304
HapMap Phase II were retained for subsequent analyses.   305

GTEx Dataset. We obtained RNA-seq gene expression levels from 8555 tissue   306
samples (53 unique tissue types) from 544 unique subjects in the GTEx Project [23]   307

data release on 2014-06-13. Of the individuals with gene expression data, genome-wide 308
genotypes (imputed with 1000 Genomes) were available for 450 individuals. While all 309
8555 tissue samples were used in the OTD model (described below) to generate 310
cross-tissue and tissue-specific components of gene expression, we used the 40 tissues 311
with the largest sample sizes when quantifying tissue-specific effects (see Table 1). 312
Approximately 2.6M non-ambiguous strand SNPs included in HapMap Phase II were 313
retained for subsequent analyses. 314

Framingham Expression Dataset. We obtained exon array expression and 315
genotype array data from 5257 individuals from the Framingham Heart Study [26]. The 316
final sample size after QC was 4286. We used the Affymetrix power tools (APT) suite 317
to perform the preprocessing and normalization steps. First the robust multi-array 318
analysis (RMA) protocol was applied which consists of three steps: background 319
correction, quantile normalization, and summarization [33]. The background correction 320
step uses antigenomic probes that do not match known genome sequences to adjust the 321
baseline for detection, and is applied separately to each array. Next, the normalization 322
step utilizes a 'sketch' quantile normalization technique instead of a memory-intensive 323
full quantile normalization. The benefit is a much lower memory requirement with little 324
accuracy trade-off for large sample sets such as this one. Finally, the adjusted probe 325
values were summarized (by the median polish method) into log-transformed expression 326
values such that one value is derived per exon or gene. Additionally an analysis of the 327
detection of probes above the background noise (DABG) was carried out. It provides 328
further diagnostic information which can be used to filter out poorly performing probes 329
and weakly expressed genes. The summarized expression values were then annotated 330
more fully using the annotation databases contained in the huex10stprobeset.db 331
(exon-level annotations) and huex10sttranscriptcluster.db (gene-level annotations) R 332
packages available from Bioconductor [34, 35]. In both cases gene annotations were 333
provided for each feature. 334

Plink [36] was used for data wrangling and cleaning steps. The data wrangling steps 335
included updating probe IDs, unifying data to the positive strand, and updating 336
locations to GRCh37. The data cleaning steps included a step to filter for variant and 337
subject missingness and minor alleles, one to filter variants with Hardy-Weinberg exact 338

test, and a step to remove unusual heterozygosity. Additionally, we used the 339

HRC-check-bin tool in order to carry out data wrangling steps required to make our 340

data compatible with the Haplotype Reference Consortium (HRC) panel 341

(`http://www.well.ox.ac.uk/~wrayner/tools/`). Having been prepared thusly, the 342

data were split by chromosome and pre-phased with SHAPEIT [37] using the 1000 343

Genomes phase 3 panel and converted to vcf format. These files were then submitted to 344

the Michigan Imputation Server 345

(`https://imputationserver.sph.umich.edu/start.html`) [31, 32] for imputation 346

with the HRC version 1 panel [38]. We applied Matrix eQTL [39] to the normalized 347

expression and imputed genotype data to generate prior eQTLs for our heritability 348

analysis. 349

Partitioning local and distal heritability of gene expression 350

Motivated by the observed differences in regulatory effect sizes of variants located in the 351

vicinity of the genes and distal to the gene, we partitioned the proportion of gene 352

expression variance explained by SNPs in the DGN cohort into two components: local 353

(SNPs within 1Mb of the gene) and distal (eQTLs on non-gene chromosomes) as defined 354

by the GENCODE [40] version 12 gene annotation. We calculated the proportion of the 355

variance (narrow-sense heritability) explained by each component using the following 356

mixed-effects model: 357

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

where Y_g represents the expression of gene g, X_k is the allelic dosage for SNP k, local 358

refers to the set of SNPs located within 1Mb of the gene's transcription start and end, 359

distal refers to SNPs in other chromosomes, and ϵ is the error term representing 360

environmental and other unknown factors. We assume that the local and distal 361

components are independent of each other as well as independent of the error term. We 362

assume random effects for $w_{k,g}^{local} \sim N(0, \sigma_{w,\text{local}}^2)$, $w_{k,g}^{distal} \sim N(0, \sigma_{w,\text{distal}}^2)$, and 363

$\epsilon \sim N(0, \sigma_\epsilon^2 I_n)$, where I_n is the identity matrix. We calculated the total variability 364

explained by local and distal components using restricted maximum likelihood (REML) 365

as implemented in the GCTA software [24]. 366

For the purpose of estimating the mean heritability (see Table 1, Figure 1 and S1 Table), we allowed the heritability estimates to take negative values (unconstrained model). 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 [10,11]. Heritabilities are plotted as point estimates with bars that extend 2 times the estimated standard error up and down. Genes were considered to have heritability significantly different from 0 if the p value from GCTA was less than 0.05.

For comparing to BSLMM PVE, we restricted the GCTA heritability estimates to be within the [0,1] interval (constrained model, see Figures 3, 4 and 5).

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 measured by Pearson's R^2 [41,42].

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

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 $\alpha = 0.05, 0.5, 1$.

Quantifying sparsity with Bayesian Sparse Linear Mixed Models (BSLMM)

We used BSLMM [20] 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 [20]. We used the

software GEMMA [43] to implement BSLMM for each gene with 100K sampling steps

per gene. The BSLMM estimates the PVE (the proportion of variance in phenotype

explained by the additive genetic model, analogous to the chip heritability in GCTA)

and PGE (the proportion of genetic variance explained by the sparse effects terms

where 0 means that genetic effect is purely polygenic and 1 means that the effect is

purely sparse). 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

We use a mixed effects model to decompose the expression level of a gene into a subject

specific and subject by tissue specific components. The expression of a gene for

individual i in tissue t, $Y_{i,t}$, is modeled as

$$Y_{i,t} = Y_i^{\mathrm{CT}} + Y_{i,t}^{\mathrm{TS}} + \mathbf{Z}_i \boldsymbol{\beta} + \epsilon_{i,t}$$

where Y_i^{CT} is the random subject level intercept, $Y_{i,t}^{\mathrm{TS}}$ is the random subject by tissue

intercept, $\mathbf{Z}_i$ represents covariates (for overall intercept, tissue intercept, gender, and

PEER factors), and $\epsilon_{i,t}$ is the error term. We assume $Y_i^{\mathrm{CT}} \sim N(0, \sigma_{\mathrm{CT}}^2)$,

$Y_{i,t}^{\mathrm{TS}} \sim N(0, \sigma_{\mathrm{TS}}^2)$, $\epsilon \sim N(0, \sigma_\epsilon^2)$, and all three independent of each other.

For the cross tissue component to be identifiable, multiple replicates of expression is

needed for each subject. In the same vein, for the tissue specific component to be

identifiable multiple replicates of expression is needed for a given tissue/subject pair.

GTEx [23] data consisted of measurement of expression for multiple tissues for each

subject, thus multiple replications per subject. However, there were very few replicated

measurement for a given tissue/subject pair. Thus, we fit the reduced model and use

the estimates of the residual as the tissue specific component.

$$Y_{i,t} = Y_i^{\mathrm{CT}} + \mathbf{Z}_i \boldsymbol{\beta} + \epsilon_{i,t}$$

The mixed effects model parameters were estimated using the `lme4` package [44] in

R [45]. Batch effects and unmeasured confounders were accounted for using 15 `PEER` 421
factors computed with the `PEER` [29] package in R. Posterior modes of the subject level 422
random intercepts were used as estimates of the cross tissue components whereas the 423
residuals of the models were used as tissue specific components. 424

The model included whole tissue gene expression levels in 8555 GTEx tissue samples 425
from 544 unique subjects. A total of 17,647 Protein-coding genes (defined by 426
GENCODE [40] version 18) with a mean gene expression level across tissues greater 427
than 0.1 RPKM (reads per kilobase of transcript per million reads mapped) and RPKM 428
> 0 in at least 3 individuals were included in the model. 429

Comparison of OTD trait heritability with multi-tissue eQTL 430
results 431

To verify that the newly derived cross tissue and tissue specific traits were capturing the 432
expected properties we used the results of the multi-tissue eQTL analysis performed by 433
Flutre et al. [28] on nine tissues from the pilot phase of the GTEx project [23]. In 434
particular, we used the posterior probability of a gene being actively regulated (PPA) in 435
a tissue downloaded from the GTEx portal at 436
`http://www.gtexportal.org/static/datasets/gtex_analysis_pilot_v3/multi_` 437
`tissue_eqtls/Multi_tissue_eQTL_GTEx_Pilot_Phase_datasets.tar`. 438

We reasoned that genes with large cross tissue component (i.e. high cross-tissue h^2) 439
would have more uniform PPA across tissues. Thus we defined for each gene a measure 440
of uniformity, U_g, across tissues based on the nine dimensional vector of PPAs using the 441
entropy formula. More specifically, we divided each vector of PPA by their sum across 442
tissues and computed the measure of uniformity as follows: 443

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

where $p_{t,g}$ is the normalized PPA for gene g and tissue t. 444

Acknowledgments

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 L, Vattikuti S, Purcell SM, Lee JJ.
Second-generation PLINK: rising to the challenge of larger and richer datasets.
Gigascience. 2015;4(7).

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. 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`.

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

46. Grossman RL, Greenway M, Heath AP, Powell R, Suarez RD, Wells W, et al.
    The design of a community science cloud: The open science data cloud
    perspective. In: High Performance Computing, Networking, Storage and Analysis
    (SCC), 2012 SC Companion:. IEEE; 2012. p. 1051–1057.

47. Heath AP, Greenway M, Powell R, Spring J, Suarez R, Hanley D, et al.
    Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets.
   Journal of the American Medical Informatics Association. 2014;21(6):969–975.

Supporting Information

S1 Fig
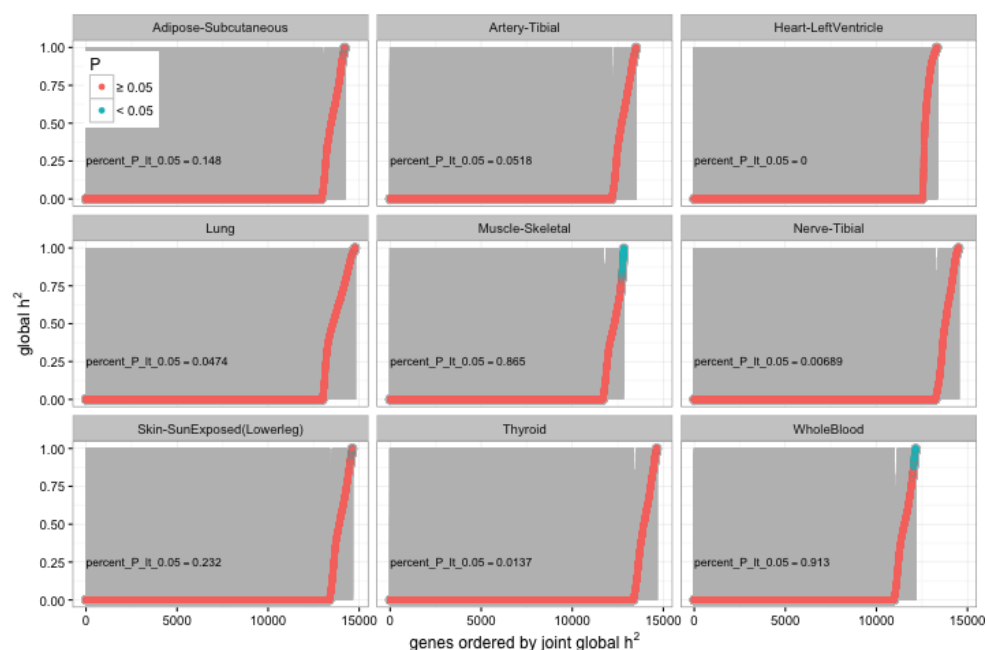
GTEx whole tissue distal heritability (h^2) estimation. Distal (SNPs that are eQTLs in the Framingham Heart Study on other chromosomes [FDR < 0.05]) gene expression h^2 estimates from a joint model in the nine GTEx tissues with the largest sample sizes are ordered by increasing h^2. The 95% confidence interval (CI) of each h^2 estimate is in gray and significant genes ($P < 0.05$) are in blue. Less than 1% of genes show significant distal h^2.

S2 Fig
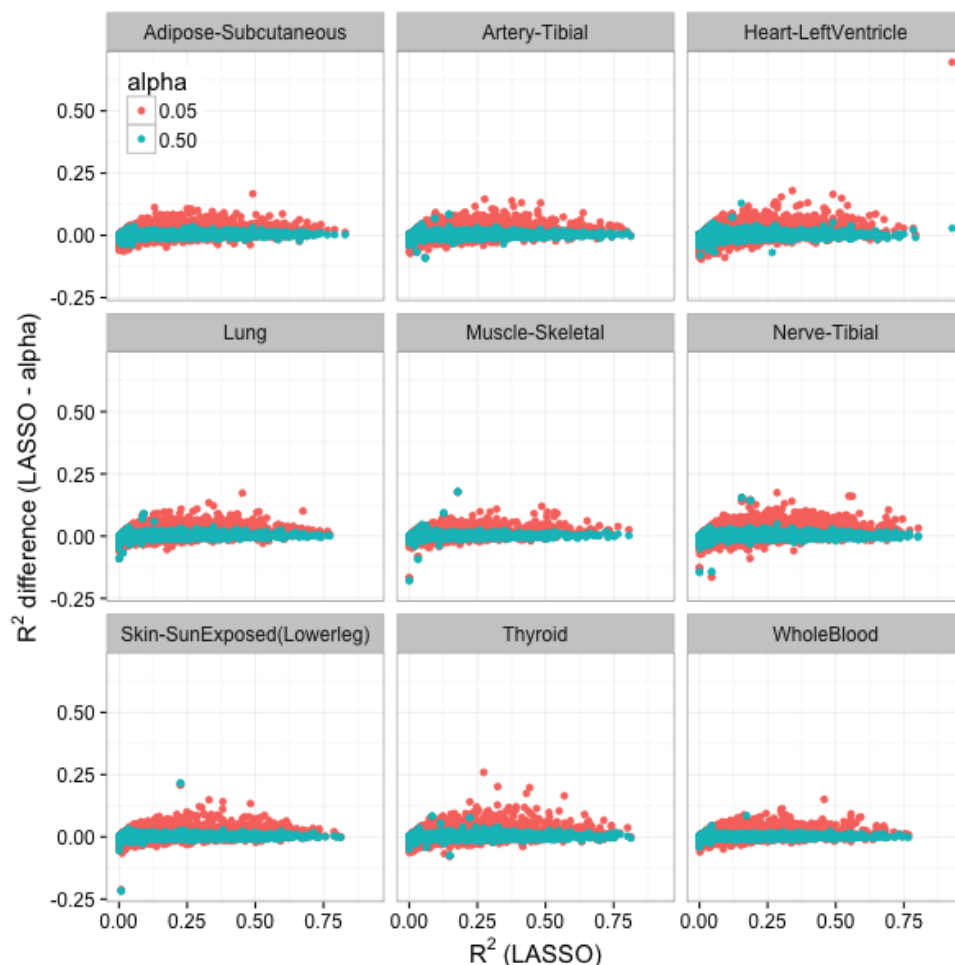
GTEx whole tissue cross-validated predictive performance across the elastic net. The difference between the cross validated R^2 of the LASSO model and the elastic net model mixing parameters 0.05 and 0.5 for autosomal protein coding genes per tissue. Elastic net with $\alpha = 0.5$ values hover around zero, meaning that it has similar predictive performance to LASSO. The R^2 difference of the more polygenic model (elastic net with $\alpha = 0.05$) is mostly above the 0 line, indicating that this model performs worse than the LASSO model across tissues.

S3 Fig
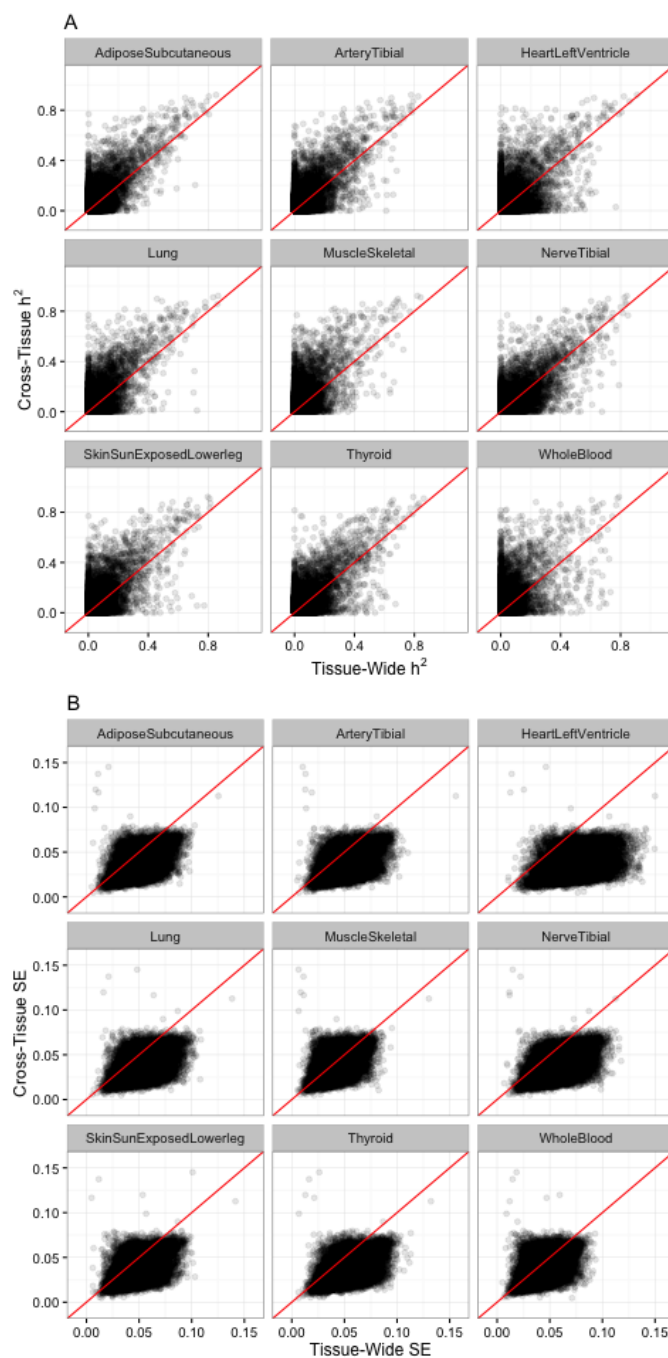
Cross-tissue and whole tissue comparison of heritability (h^2, A) and standard error (SE, B). Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Whole tissue local h^2 is estimated using the measured gene expression for each respective tissue and SNPs within 1 Mb of each gene. Estimates of h^2 for cross-tissue expression traits are larger and their standard errors are smaller than the corresponding estimates for each whole tissue expression trait.
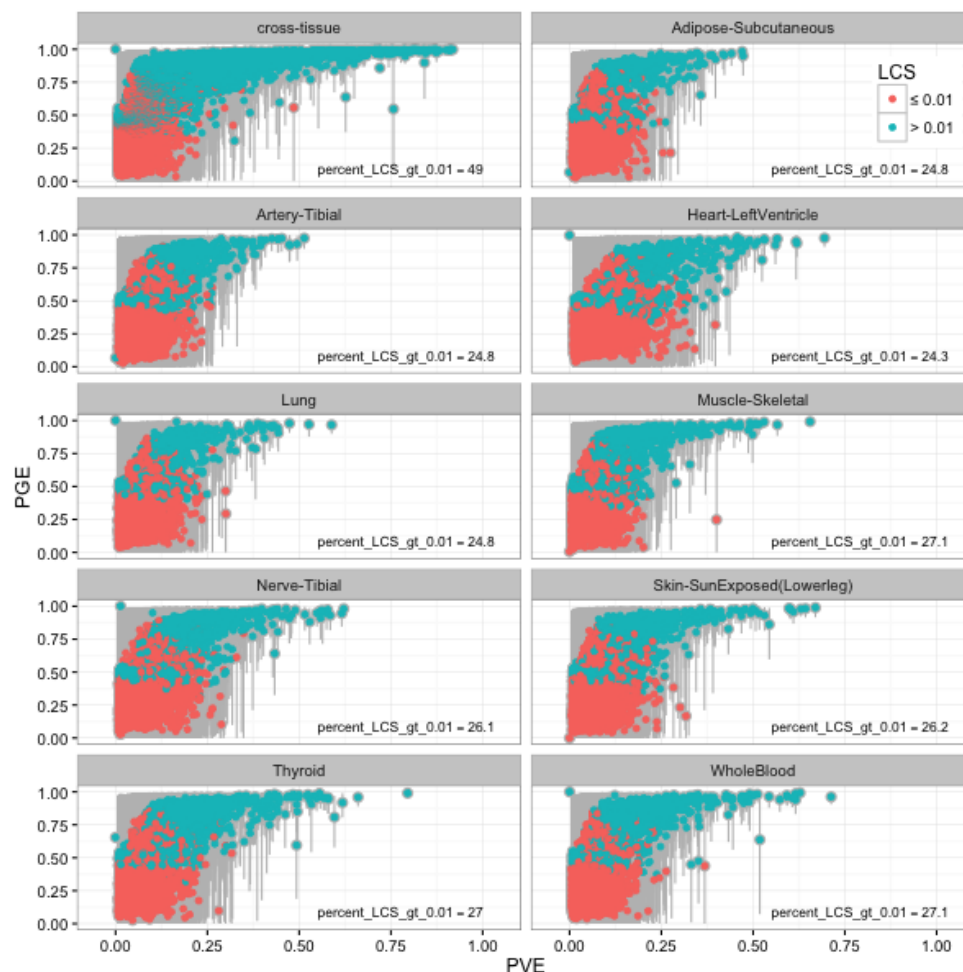
S4 Fig
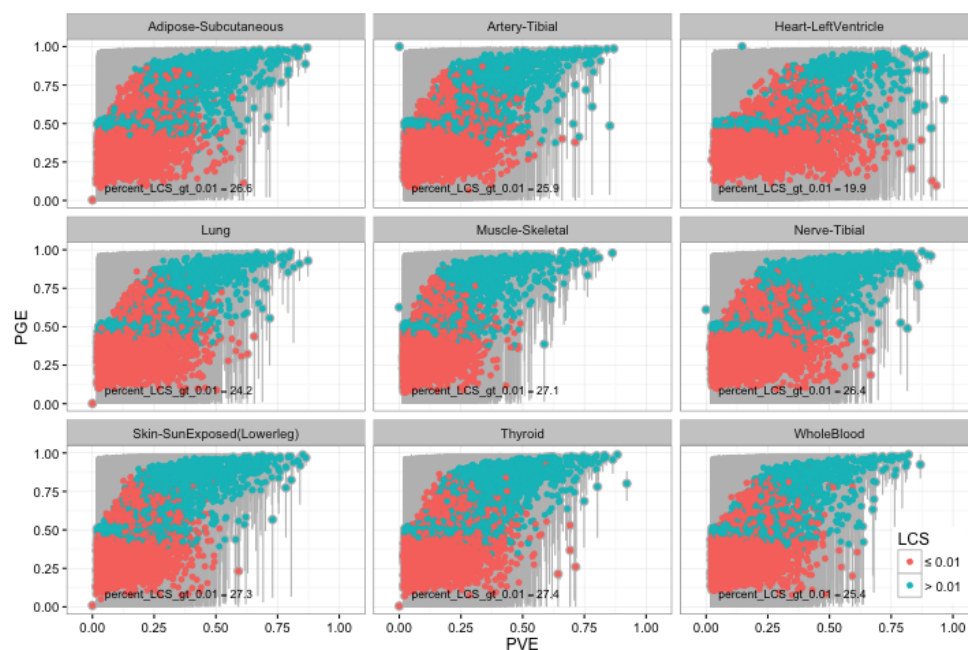
GTEx whole tissue expression Bayesian Sparse Linear Mixed Model.
Comparison of median PGE (proportion of PVE explained by sparse effects) to median PVE (total proportion of variance explained, the BSLMM equivalent of h^2) for expression of each gene. The 95% credible set of each PGE estimate is in gray and genes with a lower credible set (LCS) greater than 0.01 are in blue. For highly heritable genes the sparse component is close to 1, thus for high heritability genes the local architecture is sparse across tissues. For lower heritability genes, there is not enough evidence to determine sparsity or polygenicity.

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, the BSLMM equivalent of h^2) 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. For highly heritable genes the sparse component is close to 1, thus for high heritability genes the local architecture is sparse across tissues. About twice as many cross-tissue expression traits have significant PGE (LCS > 0.01) compared to the tissue-specific expression traits.

S6 Fig
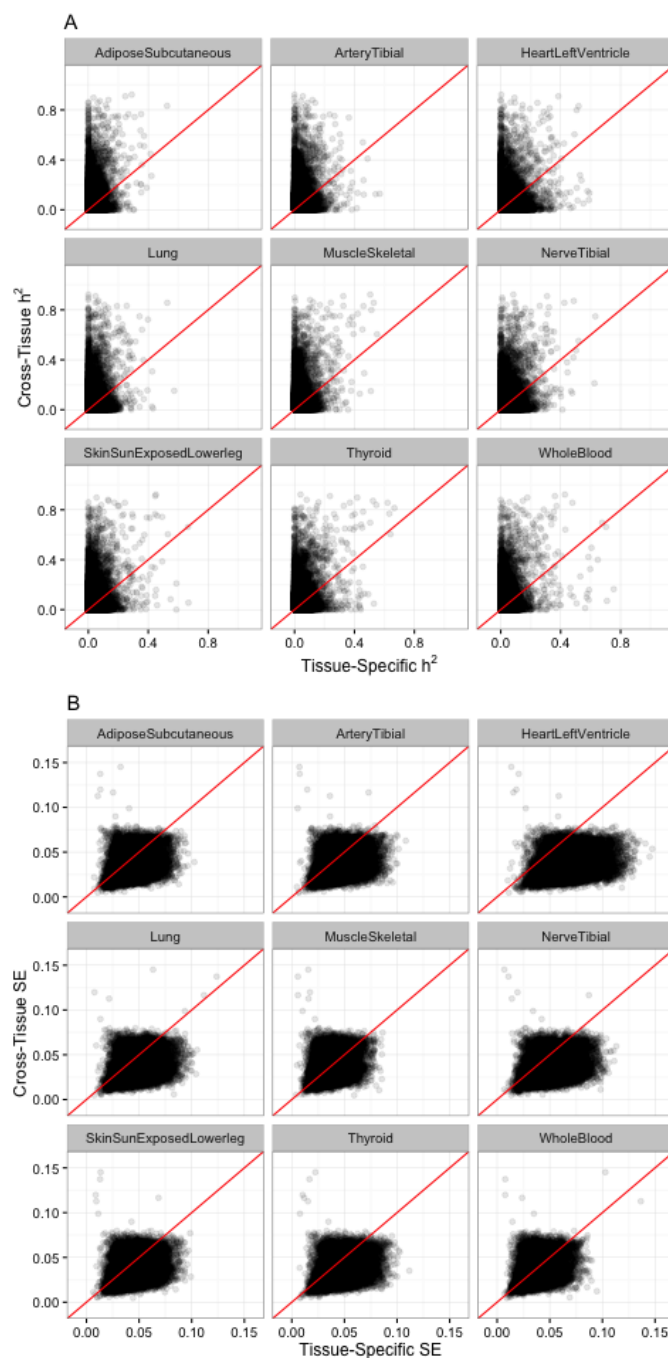
Cross-tissue and tissue-specific comparison of heritability (h^2, A) and standard error (SE, B) estimation. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-specifc local h^2 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. Estimates of h^2 for cross-tissue expression traits are larger and their standard errors are smaller than the corresponding estimates for each tissue-specific expression trait.

S7 Fig
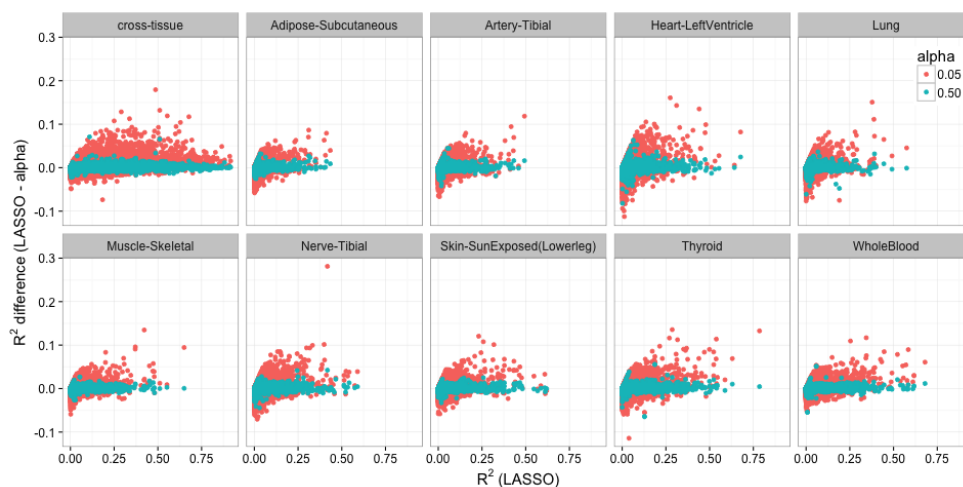
GTEx orthogonal tissue decomposition cross-tissue and tissue-specific expression cross-validated predictive performance across the elastic net. The difference between the cross validated R^2 of the LASSO model and the elastic net model mixing parameters 0.05 and 0.5 for autosomal protein coding genes per cross-tissue and tissue-specific gene expression traits. Elastic net with $\alpha = 0.5$ values hover around zero, meaning that it has similar predictive performance to LASSO. The R^2 difference of the more polygenic model (elastic net with $\alpha = 0.05$) is mostly above the 0 line, indicating that this model performs worse than the LASSO model across decomposed tissues.

S1 Table

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.

tissue	n	mean h^2 (SE)	% $P < 0.05$	num $P < 0.05$	num expressed
Cross-tissue	450	0.062 (0.0008)	35.4	6003	16951
Adipose - Subcutaneous	298	0.017 (0.0004)	10.3	1747	16951
Adrenal Gland	126	0.037 (0.0007)	10.2	1731	16951
Artery - Aorta	198	0.022 (0.0005)	10.2	1733	16951
Artery - Coronary	119	0.048 (0.0008)	12.2	2066	16951
Artery - Tibial	285	0.022 (0.0004)	13.3	2262	16951
Brain - Anterior cingulate cortex (BA24)	72	0.037 (0.0012)	8.6	1451	16951
Brain - Caudate (basal ganglia)	100	0.042 (0.0009)	9.8	1659	16951
Brain - Cerebellar Hemisphere	89	0.046 (0.0011)	10.2	1730	16951
Brain - Cerebellum	103	0.041 (0.0009)	10.8	1825	16951
Brain - Cortex	96	0.053 (0.0010)	10.6	1803	16951
Brain - Frontal Cortex (BA9)	92	0.037 (0.0009)	8.4	1418	16951
Brain - Hippocampus	81	0.033 (0.0010)	7.6	1292	16951
Brain - Hypothalamus	81	0.020 (0.0011)	7.8	1321	16951
Brain - Nucleus accumbens (basal ganglia)	93	0.040 (0.0010)	8.9	1511	16951
Brain - Putamen (basal ganglia)	82	0.033 (0.0010)	7.8	1323	16951
Breast - Mammary Tissue	183	0.020 (0.0005)	7.7	1307	16951
Cells - EBV-transformed lymphocytes	115	0.045 (0.0008)	10.8	1833	16951
Cells - Transformed fibroblasts	272	0.019 (0.0004)	10.6	1805	16951
Colon - Sigmoid	124	0.024 (0.0007)	9.0	1530	16951
Colon - Transverse	170	0.022 (0.0005)	8.6	1462	16951
Esophagus - Gastroesophageal Junction	127	0.029 (0.0007)	8.9	1512	16951
Esophagus - Mucosa	242	0.026 (0.0005)	13.7	2319	16951
Esophagus - Muscularis	219	0.024 (0.0005)	11.0	1871	16951
Heart - Atrial Appendage	159	0.031 (0.0006)	9.4	1601	16951
Heart - Left Ventricle	190	0.025 (0.0005)	11.2	1894	16951
Liver	98	0.036 (0.0009)	8.9	1506	16951
Lung	279	0.018 (0.0004)	10.7	1819	16951
Muscle - Skeletal	361	0.020 (0.0003)	13.1	2223	16951
Nerve - Tibial	256	0.026 (0.0004)	13.3	2250	16951
Ovary	85	0.043 (0.0009)	7.8	1323	16951
Pancreas	150	0.036 (0.0007)	11.6	1969	16951
Pituitary	87	0.044 (0.0010)	9.4	1590	16951
Skin - Not Sun Exposed (Suprapubic)	196	0.041 (0.0005)	19.2	3249	16951
Skin - Sun Exposed (Lower leg)	303	0.027 (0.0004)	17.3	2939	16951
Small Intestine - Terminal Ileum	77	0.046 (0.0011)	7.9	1340	16951
Spleen	89	0.062 (0.0010)	10.7	1820	16951
Stomach	171	0.020 (0.0005)	8.3	1406	16951
Testis	157	0.042 (0.0007)	14.0	2378	16951
Thyroid	279	0.024 (0.0004)	13.3	2250	16951
Whole Blood	339	0.025 (0.0004)	17.7	3004	16951

All tissues are from the GTEx Project. Mean heritability (h^2) and the standard error of the mean (SE) are calculated across genes for each tissue. The percentage (%) and number (num) of genes with significant h^2 estimates ($P < 0.05$) and the number of genes expressed in each tissue are also reported.