PLOS Genetics

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

Manuscript Number:	PGENETICS-D-16-00755R1
Full Title:	Survey of the Heritability and Sparse Architecture of Gene Expression Traits Across Human Tissues
Short Title:	Heritability and Sparse Architecture of Gene Expression
Article Type:	Research Article
Section/Category:	General
Keywords:	gene expression, gene regulation, eQTL, genetic architecture, polygenic, tissue specificity
Corresponding Author:	Hae Kyung Im University of Chicago Chicago, IL UNITED STATES
Corresponding Author's Institution:	University of Chicago
First Author:	Heather E Wheeler
Order of Authors:	Heather E Wheeler
	Kaanan P Shah
	Jonathon Brenner
	Tzintzuni Garcia
	Keston Aquino-Michaels
	Nancy J Cox
	Dan L Nicolae
	Hae Kyung Im

Abstract:	Understanding the genetic architecture of gene expression traits is key to elucidating the underlying mechanisms of complex traits. Here, for the first time, we perform a systematic survey of the heritability and the distribution of effect sizes across all representative tissues in the human body. We find that local h2 can be relatively well characterized with 59\% of expressed genes showing significant h2 (FDR<0.1) in the DGN whole blood cohort. However, current sample sizes (n<=922) do not allow us to compute distal h2. Bayesian Sparse Linear Mixed Model (BSLMM) analysis provides strong evidence that the genetic contribution to local expression traits is dominated by a handful of genetic variants rather than by the collective contribution of a large number of variants each of modest size. In other words, the local architecture of gene expression traits is sparse rather than polygenic across all 40 tissues (from DGN and GTEx) examined. This result is 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 using a novel Orthogonal Tissue Decomposition (OTD) approach. Through a series of simulations we show that the cross-tissue and tissue-specific components are identifiable via OTD. 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 R2 for original and decomposed expression phenotypes are made publicly available (https://github.com/hakyimlab/PrediXcan)

Suggested Reviewers:	Hongyu Zhao Yale University

	hongyu.zhao@yale.edu
	Amanda Myers University of Miami AMyers@med.miami.edu
	Xiang Zhou University of Michigan xiang.zhou.email@gmail.com
	Yontao Guan Baylor College of Medicine yongtaog@bcm.edu
	Marylyn Ritchie Pennsylvania State University marylyn.ritchie@psu.edu
Opposed Reviewers:	
Additional Information:	
Question	**Response**

| **Competing Interests**

You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work, | None |

acknowledging all financial support and any other relevant financial or non-financial competing interests. Do any authors of this manuscript have competing interests (as described in the [PLOS Policy on Declaration and Evaluation of Competing Interests](#))? **If yes**, please provide details about any and all competing interests in the box below. Your response should begin with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests:* **If no** authors have any competing interests to declare, please enter this statement in the box: "*The authors have declared that no competing interests exist.*" * typeset	
Data Availability PLOS journals require authors to make all data underlying the findings described in their manuscript fully available, without restriction and from the time of publication, with only rare exceptions to address legal and ethical concerns (see the [PLOS Data Policy](#) and [FAQ](#) for further details). When submitting a manuscript, authors must provide a Data Availability Statement that describes where the data underlying their manuscript can be found. Your answers to the following constitute your statement about data availability and will be included with the article in the event of publication. **Please note that simply stating 'data available on request from the author' is not acceptable.** *If, however, your data are only available upon request from the author(s), you must answer "No" to the first question below, and explain your exceptional situation in the text box provided.* Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	Yes - all data are fully available without restriction
Please describe where your data may be	https://github.com/hakyimlab/PrediXcan

found, writing in full sentences. **Your answers should be entered into the box below and will be published in the form you provide them, if your manuscript is accepted.** If you are copying our sample text below, please ensure you replace any instances of XXX with the appropriate details.

If your data are all contained within the paper and/or Supporting Information files, please state this in your answer below. For example, "All relevant data are within the paper and its Supporting Information files."

If your data are held or will be held in a public repository, include URLs, accession numbers or DOIs. For example, "All XXX files are available from the XXX database (accession number(s) XXX, XXX)." If this information will only be available after acceptance, please indicate this by ticking the box below.

If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so in the box below. For example:

"Data are available from the XXX Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data."

"Data are from the XXX study whose authors may be contacted at XXX."

* typeset

Additional data availability information:

THE UNIVERSITY OF
CHICAGO
BIOLOGICAL
SCIENCES

Department of Medicine
Section of Genetic Medicine
Hae Kyung Im, PhD
Assistant Professor
900 E 57th St, KCBD 3220, Chicago, IL 60637
T 773.702.3898 | F 773.702.2567 | @hakyim
haky@uchicago.edu | imlab.uchicago.edu

August 18, 2016

Dear Editor

We thank you and the reviewers for the thorough examination of our manuscript and extensive feedback. We appreciate the opportunity to address ~~them~~ here.

The main concerns of the reviewers can be summarized as follows.

1. Difference in magnitude with heritability estimates with previous reports

2. Interpretation of sparsity parameters

3. Lack of ground truth for heritability values

4. Details of the novel Orthogonal Tissue Decomposition (OTD) method and need for simulations

5. Effect of LD in heritability estimation

Below is a brief summary of how we addressed them. Detailed response to each comment can be found in the response to reviewers document.

1. We clarify that the magnitude of h2 is dependent on the total variance of the expression traits. This denominator will depend on overall noise level, batch effects, number of covariates used, number of PEER factors or other latent variables used for correction, normalizing and variance stabilizing transformation approaches, etc. Thus we emphasize that only ranking of h2 can be expected to be preserved across studies not the magnitude. Given this limitations, we show concordance of our estimate with previous reports using correlations.

2. We are finding through several different approaches that the local contribution to expression traits is predominantly explained by a handful of variants with rather large effects than by the collective contribution of most variants. We have several independent lines of evidence that lead to our conclusion. In the manuscript, we explain the reasoning

behind our conclusions more clearly and make our argument more convincing.

3. Even though we do not have ground truth for heritability values, we do have independent estimates, from other authors and from the different approaches we use (GCTA, BSLMM, cross-validated prediction R2). For example, heritability cannot be lower (on average, since we are dealing with estimates not true values) than the cross-validated R2. Thus, for the substantial number of genes with near 0 GCTA heritability, but large cross validated R2 and BSLMM heritability, we can be quite confident that it is GCTA that is underestimating the true value.

4. We provide more detailed explanation of the OTD and show the results of multiple simulations where we demonstrate that the method can identify the cross-tissue and tissue-specific components. We also added a figure describing the procedure.

5. We account for LD using LDAK and show good concordance of results.

We have dedicated substantial effort to address all the comments raised by the reviewers and we hope that you will find this work acceptable for publication in PLoS Genetics.

Sincerely yours,

Hae Kyung Im on behalf of the coauthors

Response to Reviwers

We thank the reviewers for the thorough examination of our manuscript and extensive feedback. We appreciate the opportunity to address them here.

The main concerns of the reviewers can be summarized as follows.

1. Difference in magnitude with heritability estimates with previous reports

2. Interpretation of sparsity parameters

3. Lack of ground truth for heritability values

4. Details of the novel Orthogonal Tissue Decomposition (OTD) method and need for simulations

5. Effect of LD in heritability estimation

Below is a brief summary of how we addressed them.

1. We clarify that the magnitude of h2 is dependent on the total variance of the expression traits. This denominator will depend on overall noise level, batch effects, number of covariates used, number of PEER factors or other latent variables used for correction, normalizing and variance stabilizing transformation approaches, etc. Thus we emphasize that only ranking of h2 can be expected to be preserved across studies not the magnitude. Given this limitations, we show concordance of our estimate with previous reports using correlations.

2. We are finding through several different approaches that the local contribution to expression traits is predominantly explained by a handful of variants with rather large effects than by the collective contribution of most variants. We have several independent lines of evidence that lead to our conclusion. In the manuscript, we explain the reasoning behind our conclusions more clearly and make our argument more convincing.

3. Even though we do not have ground truth for heritability values, we do have independent estimates, from other authors and from the different approaches we use (GCTA, BSLMM, cross-validated prediction R2). For example, heritability cannot be lower (on average, since we are dealing with estimates not true values) than the cross-validated R2. Thus, for the substantial number of genes with near 0 GCTA heritability but large cross validated R2 and BSLMM heritability, we can be quite confident that it is GCTA that is underestimating the true value.

4. We provide more detailed explanation of the OTD and show the results of multiple simulations where we demonstrate that the method can identify the cross-tissue and tissue-specific components. We also added a figure describing the procedure.

5. We account for LD using LDAK and show good concordance of results.

Next, we respond to each item and list the action we have taken to address them. and our responses are in black.

Reviewer #1:

Heritability estimates are dependent on sample size, the number of PEER factors used to correct for confounding in gene expression levels, the overall noise level, etc. Thus, when such variables differ, the estimates are not directly comparable between cohorts. However, if the heritability estimates are capturing the genetic architecture of gene expression, we do expect the rankings between cohorts to be similar. To illustrate that they are similar, we have added the following paragraph to the first subsection of the results (lines 96-102):

"Gene expression local heritability estimates were consistent between whole blood tissue from the DGN and GTEx cohorts (Spearman's ρ = 0.27 [95% confidence interval (CI): 0.25-0.30]). DGN estimates made here were also consistent with those made in independent blood cohorts from previous studies. Comparing DGN to Price et al. [29] and Wright et al. [10], ρ = 0.32 [0.30-0.35] and 0.073 [0.047-0.10], respectively. Spearman's ρ between Price et al. [29] and Wright et al. [10] gene expression heritability estimates was 0.11 [0.080-0.13]."

While we do not know the biological truth, significant rank correlations with the prior estimates of Price et al. and Wright et al. indicate that estimates of local heritability capture a replicable component of the genetic architecture of gene expression. To reduce confusion, we have changed the title of this subsection from "Local genetic variation can be well characterized for all tissues" to "Significant local heritability of gene expression in all tissues."

We thank the reviewer for raising a potential point of confusion. Our BSLMM results clearly indicate the sparse component dominates the total variance explained by genetic variants (Figure 3A) and thus we have moved this subsection (now lines 107-124) to precede our elastic net results.

The elastic net results in Figure 4A show a predominantly flat curve across mixing parameters except at the fully polygenic end (alpha=0), which shows inferior predictive power (cross-validated R2) across genes. As the reviewer says, this mostly rules out a fully polygenic model.

To clarify what is presented in Figure 4 (formerly Figure 2), we have changed the title of this subsection from "Sparse local architecture implied by sparsity of best prediction models" to "Prediction models with a sparse component outperform polygenic models."

We note that the GCTA results presented originally in this section used a REML model constrained between 0 and 1, just like the BSLMM results. This has been clarified in the new manuscript (lines 154-156). We have also updated Figures 3B and 5 such that the points are colored according to the elastic net (alpha=1) cross-validated R2 value for each gene. To Figure 5, we added correlations between the elastic net and BSLMM results and between the elastic net and LMM (GCTA) results. The genes with high BSLMM PVE and near zero LMM h2 have high elastic net prediction R2, providing strong evidence that heritability is not zero and that the underlying genetic architecture is sparse for these genes.

Also elastic net R2 estimates correlate better with BSLMM than GCTA for al tissues, consistent with a sparse genetic architecture of gene expression.

We further validated this behavior by showing Price et al. identity-by-descent (IBD) heritability estimates correlate more strongly with BSLMM than LMM estimates in the added S4 Fig. These results indicate that it is GCTA that is underestimating h2 and not that BSLMM is over estimating h2. The "BSLMM outperforms LMM in estimating h2 for small samples" subsection starting at line 143 and the Figure 3 and 5 legends have been updated to reflect these changes.

We have reworded the description of the method (lines 169-179 and Methods starting at line 447), added Figure 6, and added results of simulations (lines 180-196) to address this point.

For each gene, the cross-tissue and tissue-specific components are assumed to be normally distributed with one variance parameter for the cross-tissue component and one for the tissue-specific component. Differences between tissues will be reflected in the posterior means of the tissue by individual effect. Because of the lack of multiple replications for a given tissue per individual, in practice we use a reduced model and consider the residuals as the tissue-specific effect (see lines 431-461).

We show that the cross-tissue and tissue-specific components are identifiable using simulated expression traits for which true values are known. To preserve the correlation between tissues, we performed OTD to the observed data and used the OTD-estimates of cross-tissue and tissue-specific components as the ground truth. We added a noise term with variance equal to the total variance (and multiples of the total variance). We computed the OTD estimates of cross-tissue and tissue-specific components and showed that they correlated significantly with the true values (see lines 180-196, 474-485, S6 Fig).

In the reduced model (with only Y_i^{CT} term), the expression level of an individual only depends on the tissue through the residual error term?

Yes. The CT term is the one that is common across all tissues. Whatever is tissue-specific is represented by the remainder, i.e. the residual term.

Does this imply all eQTLs share the same effects across tissues?

When estimating tissue-specific eQTL or heritability we use only the residuals for the tissue and exclude all other tissues. Thus eQTLs will be specific to each tissue. The component that is common across all tissue is represented by the cross-tissue component, the individual specific random effect.

Without clear understanding of the model, it is difficult for me to interpret the results. More importantly, if OTD is a new method, the authors need to provide a thorough evaluation of the proposed approach (simulation study and comparison to existing approaches are warranted). Otherwise, necessary literature should be cited.

We have improved the description of the method, clarified assumptions, and added simulations that demonstrate that our method is able to identify the cross-tissue and tissue-specific components. We added a supplemental figure and a description of this comparison at lines 186-196:

"We simulated expression traits using different magnitudes of the noise level and then applied our OTD approach. Simulated values were constructed using noise variance equal in magnitude to the original variance of the expression levels of each gene. This is a rather conservative choice since the variance of the noise component is unlikely to be larger than the total variance of the expression trait. In S6 Fig, we show for one simulated example the correlation between true and OTD-estimated components for one representative gene, SLTM, and the distribution of the correlation between true and OTD-estimates for all genes. The median Pearson correlation between true and OTD-estimted cross-tissue components was 0.54 [95% CI: 0.33-0.77] and for tissue-specific components was 0.75 [0.46-0.88], demonstrating the robustness of our model (S6 Fig)."

Other minor comments:

This sentence has been changed (now at line 80) to: "This approach reduces the standard error of the estimated mean of heritability (law of large numbers)."

To avoid confusion we have dropped the detailed description of the distal heritability estimates from the manuscript. See explanation below.

The instability problem arises only for trans estimation because the likelihood is very flat so that the optimization algorithm searches in parameter values that produce negative definite covariance matrices. This does not occur with cis estimations in our case because we have enough information, i.e. the likelihood has a clear peak within the 0 to 1 range. An example of this can be seen in Figure S1 where even though we did not constrain heritability to be between 0 and 1, most local estimates fall within that range.

Only a few of the local h2 estimates were less than 0, while we found thousands of the distal h2 estimates were negative. We were able to obtain local h2 estimates for most genes in GTEx (Table 1), but when we attempted a joint local and distal model, >90% of genes did not converge. To avoid confusion we have dropped the detailed description of the distal heritability estimates from the manuscript.

To avoid confusion we have dropped the detailed description of the distal heritability estimates from the manuscript.

We have fixed the typo, this sentence is now at line 137.

Reviewer #2:

The paper surveys heritability estimates of gene expression across samples from various tissues in the context of two datasets, namely the GTEx Project and the DGN whole blood cohort. The authors used a Bayesian linear mixed model and an elastic net sparsity inducing penalty to estimate that 50% of the genes in the DGN dataset and 8 -19% of the genes in the GTEx dataset showed significant local narrow sense heritability h2. The authors further contrast the estimates from a Bayesian sparse linear mixed model (BSLMM) to a simple linear mixed model (LMM). The authors claim that higher heritability predicted by the models tending towards higher levels of sparsity (i.e., elastic net) suggest a sparse regulatory architecture as opposed to a highly polygenic model in aggregate across genes.

Major comments

1.* Too much attention is paid to the distal model. There are a number of reasons why the distal model appears to be ineffective in the context of these two datasets. I'd be inclined to avoid this
discussion entirely. A major source of potential problems, as stated in the discussion, is that there are different QC approaches necessary for local and global effects, and, from the Methods section, it is not clear whether the evaluation of local and global effects were derived from the same post QC data or from data QCed according to the downstream purpose. This will obviously impact the analysis where only local-effects are used to identify distant heritability.

Following the reviewer's advice, we have dropped the detailed description of the distal heritability estimates from the manuscript.

2.* I do not believe one of the main conclusions of the paper, that the superior fit of sparse models "shows that" genetic effects of gene expression are sparse (what exactly does this mean?) rather than polygenic. There are a number of reasons that this superior predictive fit does not lead to this conclusion, including low effect sizes, small sample sizes, (ironically) highly polygenic QTLs (with opposite directions), and high LD. The high LD in particular will confound these conclusions.

See our response to item 3 from Reviewer 1. Our analyses indicate that the cumulative effect of any polygenic component is so small that it does not add to the heritability of gene expression that we detect with sparse models (i.e. BSLMM and LASSO). We will not be able to detect very small effect sizes, but would detect opposite directions of effect with current techniques. As for LD, as suggested in item 6 by Reviewer 4, we have added the following LDAK analysis and S2 Fig (lines 90-94):

"We examined the sensitivity of our local h2 estimates to uneven linkage disequilibrium (LD) across the genome using LDAK [31], a method proposed to account for LD. Overall, we find good concordance between GCTA and LDAK estimates, which tended to be slightly lower (S2 Fig)."

3.* There is no functional analyses. Throughout the paper I was hoping to understand what it means for the heritability of certain subsets of genes to be higher (or lower), or subsets to have greater (or less) sparsity of signals. For every analysis performed, I think a (simple) analysis of the functional features of the genes or the local regions identified should accompany them.

Heritability is a measure of the proportion of the variability accounted for by genetic variation. As such it is a useful parameter to understand the sources of variation of gene expression traits.

We also found that more heritable genes tend to be more tolerant to loss of function mutations and added Fig. 1 to show this result. Presumably, such genes are more tolerant to both loss of function mutations and mutations that alter gene expression regulation. Our result is consistent with the finding from the Exome Aggregation Consortium that more eQTLs are found in more tolerant genes [32].

Our sparsity finding, i.e. a handful of genetic variants seem to contribute to the variability in gene expression traits, has important implications for the strategies used in future investigations of gene regulation and downstream effects on complex traits. For example, most fine mapping methods that attempt to find the causal SNPs that drive a GWAS locus focus on a limited number of variants. This only makes sense in cases where the underlying genetic architecture is sparse, i.e. that a handful of causal variants are determining the variability of the traits. We have added this to the discussion section starting at line 270.

4.* Throughout the paper, there are insufficient details given in the text to understand the motivation behind the analyses, how the methods support the conclusions, and what exactly was performed to get to the results.

We believe potential areas of confusion have been addressed in our updates to the paper and in our thorough responses to the reviewers outlined in this document.

Minor points

5.* what is a "functional class" of heritability?

To clarify, we have added the examples in parentheses to the following sentence in the Introduction, see lines 4-7:

"Partitioning heritability into different functional classes (e.g. promoters, coding regions, DNase I hypersensitivity sites) has been successful in quantifying the contribution of different mechanisms that drive the etiology of diseases [3–5]."

6.* How do you know that "the heritability of gene expression attributable to local genetic variation has been estimated accurately"?

We have removed this sentence from the Introduction and compare our estimates to those from other studies in the Results section (lines 96-102). In our paper, we show that we are able to accurately estimate the heritability for genes with a large sparse component.

7.* There have been a number of studies describing the polygenicity (allelic heterogeneity) of gene expression levels. Those should be reviewed, and specific exampled examined.

We have added the following to the Introduction at lines 34-37:

"These results suggest that for many genes, the genetic architecture is sparse, but not regulated by a single SNP consistent with previous work describing the allelic heterogeneity of gene expression [16–18]."

8.* LASSO defined after it was used. REML never defined.

We have moved the LASSO definition up to where it is first mentioned, now line 31. In the second sentence of the results (line 71), we inserted "(REML)" to clarify the definition:

"We used mixed-effects models (see Methods) and calculated variances using restricted maximum likelihood (REML) as implemented in GCTA [28]."

9.* what are the "9 core GTEx tissues" and the "40 tissues" (they are in the table, but how were they selected?) from the current version of the data? what was the sample size threshold?

To clarify, we removed the "9 core GTEx tissues" portion and edited the sentence at lines 54-56:

"Corroborating this finding, our prediction model built in DGN whole blood showed robust prediction [15] across the nine tissues with the largest sample size from the GTEx Pilot Project [27]."

We also added "(n > 70)" to this statement in the Methods (lines 333-335): "we used the 40 tissues with the largest sample sizes (n > 70) when quantifying tissue-specific effects (see S1 Table)."

10.* what is the role of the DGN data in this paper? The cell type is never described, the explicit comparison and replication with matched tissue in GTEx is not performed, and the implications of the different collection mechanisms not discussed. I recommend incorporating it or removing it.

The first sentence of the results stated/states (now lines 67-69): "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."

DGN has more than twice the sample size of GTEx so we feel it is important to keep it in the manuscript. We compare the estimates of whole blood from both cohorts. As described in our answer to item 1 from Reviewer 1, we have added the following paragraph to the first subsection of the results (lines 96-102):

"Gene expression local heritability estimates were consistent between whole blood tissue from the DGN and GTEx cohorts (Spearman's ρ = 0.27 [95% confidence interval (CI): 0.25-0.30]). DGN estimates made here were also consistent with those made in independent blood cohorts from previous studies. Comparing DGN to Price et al. [29] and Wright et al. [10], ρ = 0.32 [0.30-0.35] and 0.073 [0.047-0.10], respectively. Spearman's ρ between Price et al. [29] and Wright et al. [10] gene expression heritability estimates was 0.11 [0.080-0.13]."

11.* Related question: how are the tissue "covariates" (e.g., cell type heterogeneity, sample size, etc.) explored here in relation to estimates of h2 and "sparsity" of signal? presumably each of these covariates impact heritability.

The use of covariates will change the estimated h2 because it will change the total variance of the trait, i.e. the denominator in the definition of h2 (genetic variance/ total variance). For example, if we don't adjust for PEER factors, the estimates of h2 is decreased because of the additional variance added by the batch effects and other unwanted variation. Cell type heterogeneity effect will be, to some degree, captured by the PEER factors. Covariates may affect the precision with which we can determine the genetic architecture of the traits but they do not change the underlying distribution of the genetic effects.

Given the uncertainty in the distal h2 results, we have dropped the detailed description of distal heritability results.

We have rewritten this paragraph to explicitly answer these questions (lines 76-85):

"Table 1 summarizes the local heritability estimate results across all tissues. In order to obtain an unbiased estimate of mean h2 across genes, we do not constrain the model to only output h2 estimates between 0 and 1. Instead, as done previously [10,29], we allow the h2 estimates to be negative when fitting the model and thus refer to it as the unconstrained REML. This approach reduces the standard error of the estimated mean of heritability (law of large numbers). For distal heritability however, the errors in the individual heritability estimates were still too large to render a significant mean distal heritability, even in DGN whole blood, the tissue with the largest sample size (S1 Fig). The local component of h2 is relatively well estimated in DGN whole blood with 59% of genes (7474 out of 12719) showing FDR < 0.1 (Table 1)."

If gene expression is not significantly heritable, we cannot say anything about the sparse component of heritability (see large PGE error for low PVE in Fig. 2A). If a gene is significantly heritable, we show the sparse component is large (i.e. near 1)--see Fig. 2A, S8 Fig, S9 Fig.

Yes, this was the uncorrected p-value, which we have replaced with FDR < 0.1 (Benjamini-Hochberg) throughout the manuscript. To calculate the FDR for each gene, we had to recalculate GCTA two-sided p-values to range between 0 and 1, because GCTA only reports one-sided p-values between 0 and 0.5.

Given the uncertainty in the distal h2 results, we have dropped the detailed description of distal heritability results.

We removed any reference to "cis" and instead use "local" throughout the paper, since here, these local eQTLs are defined by position rather than mechanism, which could be cis-acting (allele-specific) or trans-acting (see definitions in Albert and Kruglyak 2015). We use "known eQTLs" rather than "functional priors" in lines 86-88:

"It has been shown that local-eQTLs are more likely to be distal-eQTLs of target genes [30]. However, restricting the distal h2 estimates to known eQTLs on non-gene chromosomes discovered in a separate cohort (see Methods) did not improve distal h2 estimates."

19.* What about methods like ASH (Stephens)? How would that compare with LASSO regression?

ASH is designed to deal with multiple testing issues assuming unimodal distribution of effect sizes, which induces shrinkage. However, it does not seem obvious that this could be used to train multivariate prediction models as LASSO or elastic net does. Also, the ASH model development assumes independence of markers which is not appropriate given the LD between them. The authors acknowledge that lifting this assumption would be difficult.

20.* "The mixing parameter that yields the largest cross-validation..." This is only true under a series of assumptions that do not hold here.

To clarify, the following clause was added to the beginning of this sentence at line 132:

"Given common folds and seeds, the mixing parameter that yields the largest cross-validation R2 informs the degree of sparsity of each gene expression trait."

We also added the following sentence to the Methods at lines 426-427:

"In order to compare prediction R2 values across alpha values, we used common folds and seeds for each run."

21.* what is "chip heritability"?

This is the proportion of variance in the gene expression level explained by an additive model of the SNPs included in the analysis. Since throughout the manuscript we usually use "heritability," the two places where we used "chip heritability" have been simplified to "heritability" to reduce any confusion.

22.* Biologically, what does it mean when for "highly heritable genes, the sparse component is large"?

For genes with high heritability (high PVE), we have clear evidence that most of the genetic component can explained by a small number of variants (large PGE--sparse componsent). As stated in lines 120-124:

"For example, the median PGE was 0.99 [95% CI: 0.94-1] for genes with PVE > 0.50 (Fig. 2A). The median PGE was 0.95 [95% CI: 0.72-0.99] for genes with PVE > 0.1. Fittingly, for most (96.3%) of the genes with heritability > 0.10, the number of SNPs included in the model was below 10."

For less heritable genes, we do not have the power to determine the number of variants that explain the variance in gene expression.

23.* it is not clear what sample size is "large enough" -- DGN is larger than GTEx, but certainly not large enough to get a good estimate of global h2.

We have changed "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 h2" to:

"In DGN, there is a strong correlation between BSLMM-estimated PVE and GCTA-estimated h2" at lines 144-145. Then we explain how this correlation is reduced in the smaller GTEx samples and that the BSLMM estimates are more correlated with the elastic net cross-validated R2 results than the GCTA estimates, as described above in our response to item 3 from Reviewer 1.

24.* what does "addresses the sparse component" mean? what are the assumptions of BSLMM?

We have edited this sentence at lines 152-154:

"...we found that directly modeling the sparse component with BSLMM (polygenic + sparse components) outperforms LMM (single polygenic component) for estimating h2."

We have added the following about BSLMM assumptions to the Introduction at lines 16-23:

"BSLMM models complex traits as a mixture of sparse and polygenic contributions. The sparse component consists of a handful of variants of small effect sizes whereas the polygenic component allows for most variants to contribute to the trait albeit with small effect sizes. BSLMM assumes the genotypic effects come from a mixture of two normal distributions and thus is flexible to both polygenic and sparse genetic architectures as well as everything in-between [14]. The model is enforced by sparsity inducing priors on the regression coefficients. BSLMM allows us to directly estimate the sparse and polygenic components of a trait."

25.* I liked how you used the posterior mean of the random effect as the estimate of the cross tissue component.

Thank you for the positive feedback.

26.* the OTD assumes either tissue specificity or across-tissues. What about all of the intermediate situations where it is a subset of tissues? Does this affect particular tissues (e.g., ones that are very close a priori, like the skin or brain?) in a decrease in expected tissue-specific heritability?

This is a great idea for future extension of the method. We will extend our method to account for known similarities between tissues. We will use a hierarchical approach with subsets such as endoderm, mesoderm, and exoderm to improve the deconvolution of expression levels.

27.* Assumptions of OTD need to be spelled out clearly.
We reworded the Methods section starting at line 430 to make assumptions of the OTD model more explicit.

28.* "estimated better" and "predicted better" continue to be mentioned. what do they mean and how do you quantify them?

The first time we mention "predict better" in the Introduction, we edit to (lines 26-28):

"For example, if the true genetic architecture of a trait is polygenic, it is natural to expect that polygenic models will predict better (higher predicted vs. observed R2) than sparse ones."

PGE is defined in the Methods lines 407-410:

"BSLMM estimates the PVE (the proportion of variance in phenotype explained by the additive genetic model, analogous to the heritability estimated 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)."

We do not expect PGE to be large every time PVE is large. We expect PGE to be large when the architecture of a trait is predominantly sparse (e.g. gene expression) and PGE to be small when the architecture is mainly polygenic. If the PVE estimate is low, PGE cannot be determined.

We have updated the title of Fig. 2 to read:
"Sparsity estimates using Bayesian Sparse Linear Mixed Models in DGN whole blood."

We have added a plot (Fig. 6A) with a few examples that shows that entropy is a good measure of uniformity. The figure shows the posterior probability of a gene being actively regulated in a tissue (PPA) and entropy (Uniformity) values for a number of representative genes with different combinations of PPA values. Entropy was normalized so that it would be in the 0 to 1 range. The top row corresponds to entropy whereas the remaining values correspond to PPAs. Each column corresponds to a gene. Among low entropy genes (toward the left of the plot) we find that there is a single tissue in which the gene was actively regulated (high PPA). Toward the right (high entropy) we find that genes have high PPA or mid levels of PPA across all or most tissues.

We have added a supplemental figure (S4 Fig), which shows a strong correlation between sample size and the PVE-h2 correlation, see lines 150-151.

As described in our answer to item 3 from Reviewer 1, we have updated Figure 4 such that the points are colored according to the elastic net (alpha=1) cross-validated R2 value for each gene. We also added correlations between the elastic net and BSLMM results and between the elastic net and LMM results. The genes with high BSLMM PVE and near zero LMM h2 have high elastic net R2, providing strong evidence that heritability is not zero and that the underlying genetic architecture is sparse for these genes. Comparing elastic net R2 to the heritability estimates, the Pearson correlations with BSLMM are higher than those with LMM for all tissues, consistent with a sparse genetic architecture of gene expression. We further validated this behavior

by showing Price et al. [29] identity-by-descent (IBD) heritability estimates correlate more strongly with BSLMM than LMM estimates (see the added S5 Fig). For all tissues, correlation is higher between IBD estimates and BSLMM than between IBD and LMM. The "BSLMM outperforms LMM in estimating h2 for small samples" subsection starting at line 154 and the Figure 4 legend have been updated to reflect these changes.

We fit the mixed effects model one gene at a time. We have clarified this in the text. Also to simplify notation we prefer to use an implicit indexing of gene rather than an explicit one. Thus $Y_{i,t}$ is in fact $Y_{i,t,g}$ and is a scalar (not a vector).

In the reduced model we do not explicitly estimate a subject-by-tissue component because GTEx data does not have replicates for these. Thus this would not be identifiable. Instead we estimate the cross-tissue component Y_i^{CT} and the remainder is considered to be the tissue specific component.

We consider the expression level of a gene at a given tissue to be composed of a cross=tissue component which we represent as Y_i^{CT} and a tissue specific component, which given the lack of replications we estimate as the difference between the expression level and the cross tissue components (after adjusting for covariates).

We have revised the wording of the description and added the details requested to address this point (see lines 431-461.

We have added a description of PPA to the Methods at lines 494-501:

"PPA can be interpreted as the probability a gene is regulated by an eQTL in tissue *t* given the data. In the GTEx pilot, the most significant eQTL per gene was used to compute PPA [27]. PPA is computed from a joint analysis of all tissues and takes account of sharing of eQTLs among tissues [33]. For example, consider a SNP showing modest association with expression in tissue *t*. If this SNP also shows strong association in the other tissues, then it will be assigned a higher probability of being an active eQTL in tissue *t* than if it showed no association in the other tissues [33]."

The previous Fig. 5 is now Fig. 6B, which has been changed to a density plot rather than a boxplot. We added the following to the legend to describe the p-value: "The Kruskal-Wallis rank sum test revealed a significant difference in the cross-tissue h2 of uniformity groups ($\chi 2 = 31.4$, $P < 10^{-6}$)."

We also added S12 Fig to show the continuous scale relationship.

Thank you for the suggestion. We did not seek to identify markers for tissue type here, but that is a potentially useful application that could be explored in detail in a subsequent paper.

Typos have been fixed.

As shown in Fig. 3, we limit our 10-fold cross validation of 21 mixing parameters (α = 0, 0.05, 0.1, ..., 0.90, 0.95, 1) to genes on chromosome 22. We perform 10-fold cross validation at 3 mixing parameters (α = 0.05, 0.5, 1) for all genes.

Reviewer #4:

We clarify that our analysis only considers the effect of common variants, see our response to item 1 below. If common variants are tagging rare variants, we would expect higher predictive R2 with a more polygenic model. The predominant sparse component we found through BSLMM and elastic net suggests that our method is not picking up untyped rare causal variants. We note that any rare variants not tagged by common variants are not included in our heritability estimates and prediction models. For common variants, our method accounts for the heritability of untyped variants through linkage disequilibrium.

Throughout our manuscript, we change "local" variation to "local common" variation. We add the MAF info to the first mention of local h2 in the results at lines 72-73: "For the local heritability component, we used common (minor allele frequency > 0.05) variants within 1Mb of the transcription start and end…"

Given the uncertainty in the distal h2 results, we have dropped the detailed description of distal heritability results and our comparison to Price et al.

We added the following caveat to the Discussion at lines 276-278:

"We note that any rare variants not tagged by common variants are not included in our heritability estimates and prediction models."

multiple traits? Third, same authors proposed methods for expression prediction across tissues in (Wang et al AJHG 2016); what is the connection between these two works?

As suggested by the reviewer, we have performed the simulations and show that OTD-estimates of cross-tissue and tissue-specific components are concordant with the true values used in the simulations. More details can be found in the manuscript lines 180-196 and S6 Fig.

The goal of Bivariate REML and OTD are different. Bivariate REML does not try to decouple expression levels into cross tissue and tissue specific components. The same is true for the method proposed by Wang et al, which consists of imputing missing expression data to improve downstream analysis. Our goal is to decouple the expression levels into a component that is common across tissues and a tissue-specific component. Also our mixed effects models accounts for missingness implicitly without imputation.

We have reworded the description of the OTD model (see Methods starting at line 430) to make the goals of OTD more clear.

3. Indeed, several recent works (some cited here some not) have shown that a sparse predictor can accurately predict gene expression; however it is unclear what conclusions can be drawn on the effect size distribution at the unknown (or untyped) causal variants due to tagging from linkage disequilibrium (see main comment above).

Common variants that tag untyped causal variants will not change our results. As explained in item 1, we note that any rare variants not tagged by common variants are not included in our heritability estimates and prediction models.

We add citations to several studies that have developed predictors of gene expression at lines 280-282:

"Sparse models capture the most variance in gene expression at current sample sizes and have been used successfully in gene expression prediction methods [15,36–38].

4. Introduction section, 3nd paragraph. There is an extensive literature (entire field of research) in prediction/regression for sparse models; I find it strange that the authors claim that LASSO should be the preferred prediction approach for all sparse architectures.

We add the following sentence to the Introduction at lines 42-44:

"Elastic net [23] is a good multi-purpose model that encompasses both LASSO and ridge regression at its extremes and has been shown to predict well across many complex traits with diverse genetic architectures [24]."

We also note that in addition to elastic net, we used BSLMM to model sparse effects. However, the runtime of BSLMM prevented us from running 10-fold cross-validation for each gene, which was easily implemented in elastic net.

5. Certain paragraphs of the manuscript (e.g., end of introduction, discussion etc) give the impression that this work is an appendix of existing work from the same authors (i.e. PrediXcan, Gamazon et al Nat Genet 2015). The presentation could be significantly improved if the findings are presented to be useful for the entire community interested in understanding the genetic architecture of expression. For example, just restricting to

We have changed the wording of the manuscript to make it clear that the usefulness of the results are not restricted to PrediXcan application. Specifically, we add the following to the Discussion at lines 296-300:

"Many groups have proposed integrating genotype and expression data to understand complex traits \cite{Gamazon_2015,giambartolomei2014bayesian,gusev2016integrative,he2013sherlock,Wheeler_2014,zhu2016integration}. Through integration of our OTD expression traits with studies of complex diseases, we expect results from the cross-tissue models to relate to mechanisms that are shared across multiple tissues, whereas results from the tissue-specific models will inform us about the context specific mechanisms."

Based on the reviewer's suggestion, we estimated the local heritability of gene expression using LDAK in DGN and compared the estimates to our GCTA results. We have added the following paragraph and supplemental figure to the results (lines 90-94):

"We examined the sensitivity of our local h2 estimates to uneven linkage disequilibrium (LD) across the genome using LDAK [31], a method proposed to account for LD. Overall, we find good concordance between GCTA and LDAK estimates, with slightly lower LDAK estimates (S2 Fig)."

As in Price et al. [29], the standard error (SE) is estimated from the h2 estimates obtained when the expression sample labels are shuffled 100 times. The standard deviation of the mean of the permuted h2 estimates is our estimate of SE.

The known eQTL set is fewer SNPs than the all non-chromosome SNP set, thus less of the genome is tagged. More samples will be required for accurate distal h2 estimates.

Fig. 1 is now S1 Fig and was re-designed to present an overview of each gene's local and distal h2 and to show the large errors of the distal estimates. The mean local heritabilities are reported in the text and Table 1.

The number of predictors in our model are not used as a basis to estimate sparsity of the models. As explained by the authors of BSLMM in Zhou et al. 2013: "we define PVE (the total proportion of variance in phenotype explained by the sparse effects and random effects terms together) and PGE (the proportion of genetic variance explained by the sparse effects terms)".

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

Understanding the genetic architecture of gene expression traits is key to elucidating the underlying mechanisms of complex traits. Here, for the first time, we perform a systematic survey of the heritability and the distribution of effect sizes across all representative tissues in the human body. We find that local h^2 can be relatively well characterized with 59% of expressed genes showing significant h^2 (FDR < 0.1) in the DGN whole blood cohort. However, current sample sizes ($n \leq 922$) do not allow us to compute distal h^2. Bayesian Sparse Linear Mixed Model (BSLMM) analysis provides strong evidence that the genetic contribution to local expression traits is dominated by a handful of genetic variants rather than by the collective contribution of a large number of variants each of modest size. In other words, the local architecture of gene expression traits is sparse rather than polygenic across all 40 tissues (from DGN and GTEx) examined. This result is 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 using a novel Orthogonal Tissue Decomposition (OTD) approach. Through a series of simulations we show that the cross-tissue and tissue-specific components are identifiable via OTD. 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, but not distal heritability, can be well characterized with current sample sizes. We found that the distribution of effect sizes is more consistent with a sparse local architecture in 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 (e.g. promoters, coding regions, DNase I hypersensitivity sites) 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. While many common diseases are likely polygenic [11–13], 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].

Bayesian Sparse Linear Mixed Modeling (BSLMM) models complex traits as a mixture of sparse and polygenic contributions. The sparse component consists of a handful of variants of large effect sizes whereas the polygenic component allows for most variants to contribute to the trait albeit with small effect sizes. BSLMM assumes the genotypic effects come from a mixture of two normal distributions and thus is flexible to

both polygenic and sparse genetic architectures as well as everything in-between [14]. $_{21}$
The model is enforced by sparsity inducing priors on the regression coefficients. $_{22}$
BSLMM allows us to directly estimate the sparse and polygenic components of a trait. $_{23}$

As a somewhat independent approach to determine the sparsity and polygenicity of $_{24}$
gene expression traits traits, we can look at the relative prediction performance of $_{25}$
sparse and polygenic models. For example, if the true genetic architecture of a trait is $_{26}$
polygenic, it is natural to expect that polygenic models will predict better (higher $_{27}$
predicted vs. observed R^2) than sparse ones. We assessed the ability of various models, $_{28}$
with different underlying assumptions, to predict gene expression in order to understand $_{29}$
the underlying genetic architecture of gene expression. For gene expression prediction, $_{30}$
we have shown that sparse models such as LASSO (Least Absolute Shrinkage and $_{31}$
Selection Operator) perform better than a polygenic score model and that a model that $_{32}$
uses the top eQTL variant outperformed the polygenic score but did not do as well as $_{33}$
LASSO or elastic net (mixing parameter $\alpha = 0.5$) [15]. These results suggest that for $_{34}$
many genes, the genetic architecture is sparse, but not regulated by a single SNP, which $_{35}$
is consistent with previous work describing the allelic heterogeneity of gene $_{36}$
expression [16–18]. $_{37}$

Thus, gene expression traits with sparse architecture should be better predicted with $_{38}$
models such as LASSO, which prefers solutions with fewer parameters, each of large $_{39}$
effect [19]. Conversely, highly polygenic traits should be better predicted with ridge $_{40}$
regression or similarly polygenic models that prefer solutions with many parameters, $_{41}$
each of small effect [20–22]. Elastic net [23] is a good multi-purpose model that $_{42}$
encompasses both LASSO and ridge regression at its extremes and has been shown to $_{43}$
predict well across many complex traits with diverse genetic architectures [24]. $_{44}$

Most previous human eQTL studies were performed in whole blood or $_{45}$
lymphoblastoid cell lines due to ease of access or culturabilty [8, 25, 26]. Although $_{46}$
studies with a few other tissues have been published, comprehensive coverage of human $_{47}$
tissues was not available until the launching of the Genotype-Tissue Expression (GTEx) $_{48}$
Project. GTEx aims to examine the genetics of gene expression more comprehensively $_{49}$
and has recently published a pilot analysis of eQTL data from 1641 samples across 43 $_{50}$
tissues from 175 individuals [27]. Here we use a much larger set of 8555 samples across $_{51}$
53 tissues corresponding to 544 individuals. One of the findings of the pilot analysis was $_{52}$

that a large portion of the local regulation of expression traits is shared across multiple   53
tissues. Corroborating this finding, our prediction model built in DGN whole blood   54
showed robust prediction [15] across the nine tissues with the largest sample size from   55
the GTEx Pilot Project [27].   56

This shared regulation across tissues implies that there is much to be learned from   57
large sample studies of easily accessible tissues. Yet, a portion of gene regulation seems   58
to be tissue dependent [27]. To further investigate the genetic architecture that is   59
common across tissues or specific to each tissue, we use a mixed effects model-based   60
approach termed orthogonal tissue decomposition (OTD) and sought to decouple the   61
cross-tissue and tissue-specific mechanisms in the rich GTEx dataset. We show that the   62
cross-tissue and tissue-specific expression phenotypes constructed with our OTD model   63
reflect the expected biology.   64

Results   65

Significant local heritability of gene expression in all tissues   66

We estimated the local and distal heritability of gene expression levels in 40 tissues from   67
the GTEx consortium and whole blood from the Depression Genes and Networks   68
(DGN) cohort. The sample size in GTEx varied from 72 to 361 depending on the tissue,   69
while 922 samples were available in DGN [26]. We used linear mixed models (see   70
Methods) and calculated variances using restricted maximum likelihood (REML) as   71
implemented in GCTA [28]. For the local heritability component, we used common   72
(minor allele frequency > 0.05) variants within 1Mb of the transcription start and end   73
of each protein coding gene, whereas for the distal component, we used common   74
variants outside of the chromosome where the gene was located.   75

Table 1 summarizes the local heritability estimate results across all tissues. In order   76
to obtain an unbiased estimate of mean h^2 across genes, we do not constrain the model   77
to only output h^2 estimates between 0 and 1. Instead, as done previously [10, 29], we   78
allow the h^2 estimates to be negative when fitting the model and thus refer to it as the   79
unconstrained REML. This approach reduces the standard error of the estimated mean   80
of heritability (law of large numbers). For distal heritability, the errors in the individual   81

heritability estimates were still too large to render a significant mean distal heritability, even in DGN whole blood, the tissue with the largest sample size (S1 Fig). The local component of h^2 is relatively well estimated in DGN whole blood with 59% of genes (7474 out of 12719) showing FDR < 0.1 (Table 1).

It has been shown that local-eQTLs are more likely to be distal-eQTLs of target genes [30]. However, restricting the distal h^2 estimates to known eQTLs on non-gene chromosomes discovered in a separate cohort (see Methods) did not improve distal h^2 estimates.

We examined the sensitivity of our local h^2 estimates to uneven linkage disequilibrium (LD) across the genome using LDAK [31], a method proposed to account for LD. Overall, we find good concordance between GCTA and LDAK estimates, with slightly lower LDAK estimates (S2 Fig). Given the limited sample size we will focus on local regulation for the remainder of the paper.

Across all tissues 15-59% of genes had significant estimates (FDR < 0.1) as shown in Table 1. Gene expression local heritability estimates were consistent between whole blood tissue from the DGN and GTEx cohorts (Spearman's $\rho = 0.27$ [95% confidence interval (CI): 0.25-0.30]). DGN estimates made here were also consistent with those made in independent blood cohorts from previous studies. Comparing DGN to Price et al. [29] and Wright et al. [10], $\rho = 0.32$ [0.30-0.35] and 0.073 [0.047-0.10], respectively. Spearman's ρ between Price et al. [29] and Wright et al. [10] gene expression heritability estimates was 0.11 [0.080-0.13].

We also found that more heritable genes tend to be more tolerant to loss of function mutations (Fig. 1). Presumably, such genes are more tolerant to both loss of function mutations and mutations that alter gene expression regulation. This result is consistent with the finding that more eQTLs are found in more tolerant genes [32].

Table 1. Estimates of unconstrained local h² across genes within tissues.

tissue	n	mean h^2 (SE)	% FDR<0.1	num FDR<0.1	num expressed
DGN Whole Blood	922	0.149 (0.0039)	58.8	7474	12719
Cross-tissue	450	0.062 (0.017)	20.2	2995	14861
Adipose - Subcutaneous	298	0.038 (0.028)	18.8	2666	14205
Adrenal Gland	126	0.043 (0.018)	17.5	2479	14150
Artery - Aorta	198	0.042 (0.024)	17.2	2385	13844
Artery - Coronary	119	0.037 (0.048)	16.5	2326	14127
Artery - Tibial	285	0.042 (0.02)	18.4	2489	13504
Brain - Anterior cingulate cortex (BA24)	72	0.028 (0.036)	15.4	2235	14515
Brain - Caudate (basal ganglia)	100	0.037 (0.047)	17.2	2521	14632
Brain - Cerebellar Hemisphere	89	0.049 (0.068)	16.0	2294	14295
Brain - Cerebellum	103	0.050 (0.06)	18.3	2646	14491
Brain - Cortex	96	0.045 (0.057)	17.7	2593	14689
Brain - Frontal Cortex (BA9)	92	0.038 (0.077)	16.6	2416	14554
Brain - Hippocampus	81	0.037 (0.05)	16.4	2376	14513
Brain - Hypothalamus	81	0.017 (0.043)	15.2	2238	14759
Brain - Nucleus accumbens (basal ganglia)	93	0.029 (0.04)	17.1	2498	14601
Brain - Putamen (basal ganglia)	82	0.032 (0.064)	17.3	2495	14404
Breast - Mammary Tissue	183	0.029 (0.037)	17.0	2496	14700
Cells - EBV-transformed lymphocytes	115	0.058 (0.1)	16.6	2066	12454
Cells - Transformed fibroblasts	272	0.051 (0.031)	20.0	2552	12756
Colon - Sigmoid	124	0.033 (0.019)	16.0	2298	14321
Colon - Transverse	170	0.036 (0.058)	16.3	2398	14676
Esophagus - Gastroesophageal Junction	127	0.032 (0.039)	16.1	2270	14125
Esophagus - Mucosa	242	0.042 (0.03)	17.1	2432	14239
Esophagus - Muscularis	219	0.039 (0.015)	17.7	2493	14047
Heart - Atrial Appendage	159	0.042 (0.03)	16.8	2327	13892
Heart - Left Ventricle	190	0.034 (0.034)	16.5	2203	13321
Liver	98	0.033 (0.062)	16.5	2230	13553
Lung	279	0.032 (0.034)	17.0	2509	14775
Muscle - Skeletal	361	0.033 (0.03)	17.9	2301	12833
Nerve - Tibial	256	0.052 (0.087)	20.6	2992	14510
Ovary	85	0.037 (0.094)	17.2	2418	14094
Pancreas	150	0.047 (0.024)	17.2	2398	13941
Pituitary	87	0.038 (0.055)	16.6	2527	15183
Skin - Not Sun Exposed (Suprapubic)	196	0.041 (0.034)	17.0	2490	14642
Skin - Sun Exposed (Lower leg)	303	0.039 (0.016)	18.4	2687	14625
Small Intestine - Terminal Ileum	77	0.036 (0.053)	17.4	2591	14860
Spleen	89	0.059 (0.061)	17.0	2451	14449
Stomach	171	0.032 (0.025)	16.1	2338	14531
Testis	157	0.054 (0.044)	17.8	3009	16936
Thyroid	279	0.044 (0.066)	19.4	2838	14642
Whole Blood	339	0.033 (0.023)	18.6	2260	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) is calculated across genes for each tissue. The standard error (SE) is estimated from the h^2 estimates obtained when the expression sample labels are shuffled 100 times. The percentage (%) and number (num) of genes with significant h^2 estimates (FDR < 0.1) in each tissue are reported. The number of genes in each tissue with mean RPKM > 0.1 (num expressed) is also reported.
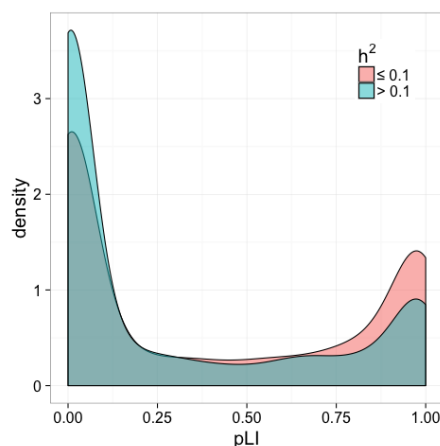
Figure 1. Genes with heritable expression in DGN whole blood are more tolerant to loss of function mutations. The distribution of the probability of being loss-of-function intolerant (pLI) for each gene (from the Exome Aggregation Consortium [32]) dichotomized by local heritability estimates. The Kruskal-Wallis rank sum test revealed a significant difference in the pLI of heritability groups ($\chi^2 = 234$, $P < 10^{-52}$). More heritable genes ($h^2 > 0.1$ in blue) have lower pLI metrics and are thus more tolerant to mutation than genes with lower h^2.

Sparse local architecture revealed by direct estimation using Bayesian Sparse Linear Mixed Modeling (BSLMM)

Next, we sought to determine whether the local common 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. To do this, we used the BSLMM [14] approach, which models the genetic contribution as the sum of a sparse component and a highly 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 heritability we have estimated using a linear mixed model (LMM) as implemented in GCTA.

For genes with low heritability, our power to discern between sparse and polygenic architecture is limited. In contrast, for highly heritable genes, the credible sets are tighter and we are able to determine that the sparse component is large. For example, the median PGE was 0.99 [95% CI: 0.94-1] for genes with PVE > 0.50 (Fig. 2A). The median PGE was 0.95 [95% CI: 0.72-0.99] for genes with PVE > 0.1. Fittingly, for most (96.3%) of the genes with heritability > 0.10, the number of SNPs included in the model
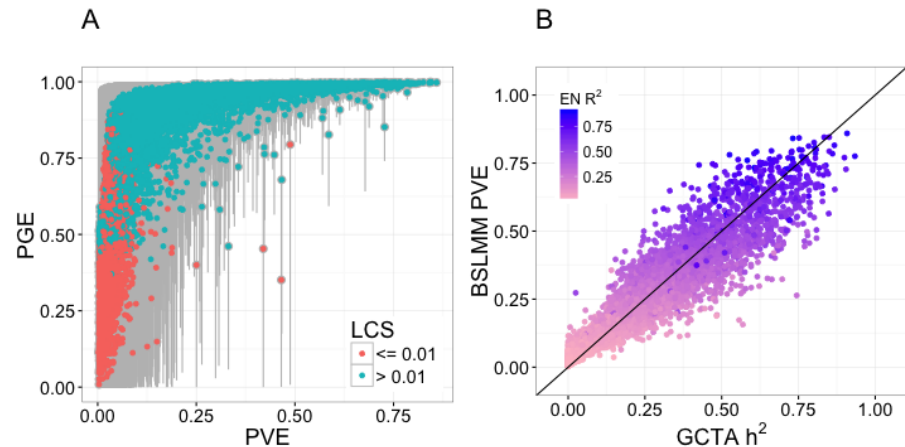
Figure 2. Sparsity estimates using Bayesian Sparse Linear Mixed Models in DGN whole blood. (**A**) This panel shows a measure of sparsity of the gene expression traits represented by the PGE parameter from the BSLMM approach. PGE is the proportion of the sparse component of the total variance explained 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). Here, the estimates are constrained to be between 0 and 1 in both models. Each point is colored according to that gene's elastic net $\alpha = 1$ cross-validated prediction correlation squared (EN R^2). Note genes with high heritability have high prediction R^2, as expected.

Prediction models with a sparse component outperform polygenic models

To further confirm the local sparsity of gene expression traits, we 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). We performed 10-fold cross-validation using the elastic net [23] to test the predictive performance of local SNPs for gene expression across a range of mixing parameters (α). Given common folds and seeds, 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 identical for $\alpha = 0.5$ through $\alpha = 1$ in the DGN cohort (Fig. 3). 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 (S3 Fig). Together with the BSLMM results, this suggests that for most genes, the effect of local common genetic variation on gene expression is sparse rather than polygenic.
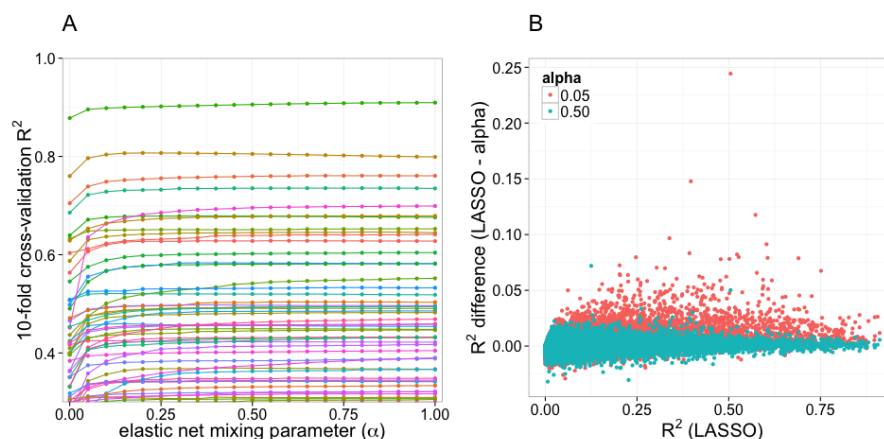
Figure 3. DGN cross-validated predictive performance across the elastic net. Elastic net prediction models were built in the DGN whole blood and performance was quantified by the cross-validated R^2 between observed and predicted expression levels. (**A**) This panel shows the 10-fold cross validated 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.

BSLMM outperforms LMM in estimating h^2 for small samples

In DGN, there is a strong correlation between BSLMM-estimated PVE and GCTA-estimated h^2 (Fig. 2B, 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 LMM-estimated h^2 (Fig. 4). This is further confirmation of the predominantly

:::PLOS | SUBMISSION

sparse local architecture of gene expression traits: the underlying assumption in the   148
LMM approach to estimate heritability is that the genetic effect sizes are normally   149
distributed, i.e. most variants have small effect sizes. LMM is quite robust to departure   150
from this assumption, but only when the sample size is rather large (S4 Fig). For the   151
relatively small sample sizes in GTEx ($n \leq 361$), we found that directly modeling the   152
sparse component with BSLMM (polygenic + sparse components) outperforms LMM   153
(single polygenic component) for estimating h^2. Here, unlike in previous sections (see   154
Table 1 and S1 Fig), both the BSLMM and LMM model estimates are constrained to be   155
between 0 and 1.   156

To ensure that the high values of heritability estimated by BSLMM are due to   157
actually higher h^2 and not to limitations or bias of the BSLMM, we use the fact that   158
cross-validated prediction R^2 (observed vs predicted squared correlation when   159
prediction parameters are estimated without including the observations to be predicted)   160
is bounded by the heritable component. In other words, genetically based prediction   161
models can only explain or predict the portion that is heritable. As shown in Figure 4,   162
most of the genes with high BSLMM heritability but low LMM h^2 have large values of   163
prediction R^2. We further validated this behavior by showing Price et al. [29]   164
identity-by-descent (IBD) heritability estimates correlate more strongly with BSLMM   165
than LMM estimates (S5 Fig). For all tissues, correlation is higher between IBD   166
estimates and BSLMM than between IBD and LMM. These findings provide strong   167
evidence that LMM is underestimating h^2 for these genes.   168

PLOS   11/47
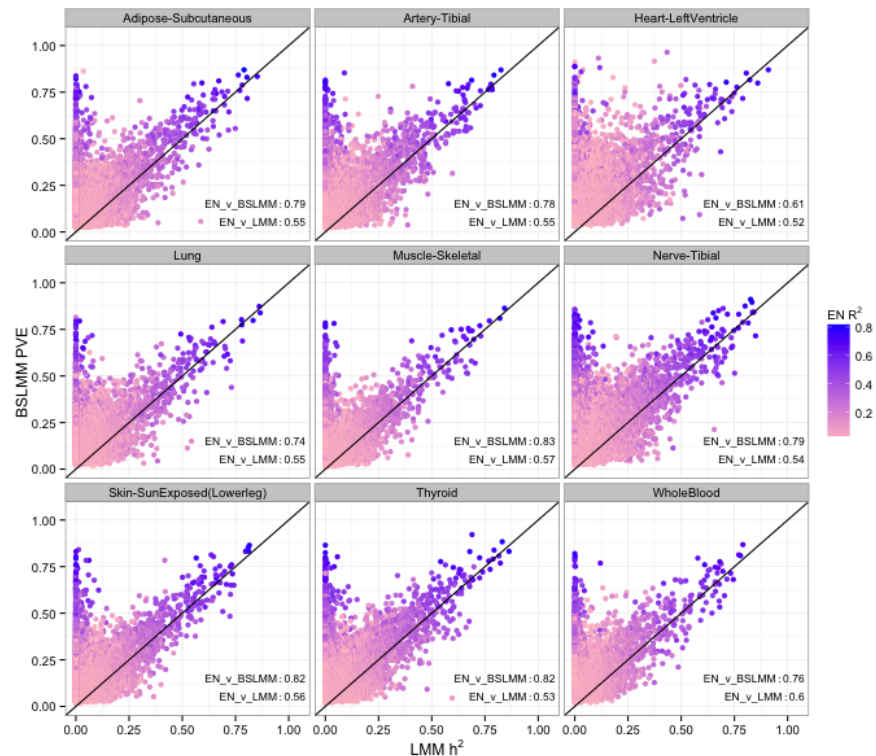
Figure 4. BSLMM vs LMM estimates of heritability in GTEx. This figure
shows the comparison between estimates of heritability using BSLMM vs. LMM
(GCTA) for GTEx data. Here, in both models the estimates are constrained to be
between 0 and 1. For most genes BSLMM estimates are larger than LMM estimates
reflecting the fact that BSLMM yields better estimates of heritability because of its
ability to account for the sparse component. Each point is colored according to that
gene's prediction R2 (correlation squared between cross-validated elastic net prediction
vs observed expression denoted EN R^2). At the bottom left of each panel, we show the
correlation between BSLMM (EN_v_BSLMM) and LMM (EN_v_LMM). BSLMM is
consistently more correlated with the elastic net correlation. This provides further
indication that the local architecture is predominantly sparse.

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

Since a substantial portion of local regulation was shown to be common across multiple   171

tissues [27], we sought to decompose the expression levels into a component that is   172

common across all tissues and tissue-specific components. Figure 5 shows a diagram   173

describing the decomposition for which we use a linear mixed effects model with a   174

person-level random effect (see Methods). We use the posterior mean of this random   175

effect as an estimate of the cross-tissue component. We consider the residual component   176

of this model as the tissue-specific component. We call this approach orthogonal tissue   177

decomposition (OTD) because the cross-tissue and tissue-specific components are   178

assumed to be independent in the model.   179

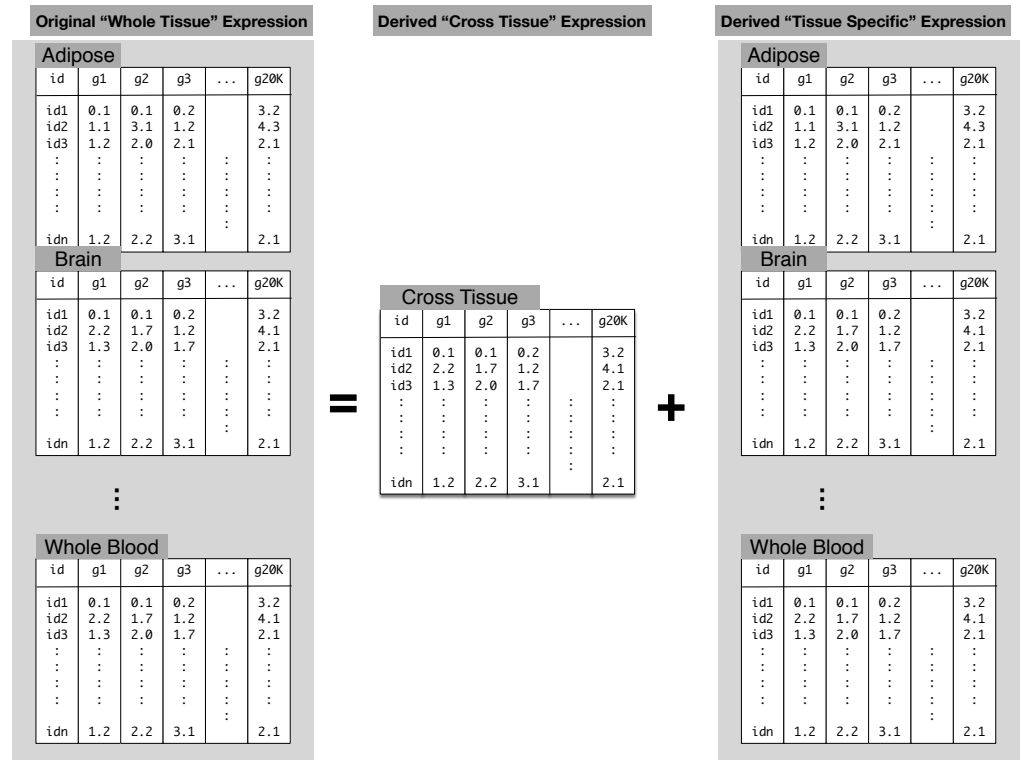**Figure 5. Orthogonal Tissue Decomposition of gene expression traits.** For a given gene, the expression level is decoupled into a component that is specific to the individual and another component that is specific to the individual and tissue. The left side of the equation in the figure corresponds to the original "whole tissue" expression levels. The right side has the component specific for the individual, independent of the tissue and the tissue-specific component. Given the lack of multiple replications for a given tissue/individual we use a mixed effects model with a random effect that is specific to the individual. The cross-tissue component is estimated as the posterior mean of the subject-specific random effect. The tissue-specific component is estimated as the residual of the model fit, i.e. the difference between the "whole tissue" expression and the cross-tissue component. The rationale is that once we remove the component that is common across tissues, the remaining will be specific to the tissue. Models are fit one gene at a time. Covariates are not shown to simplify the presentation.

First, we sought to demonstrate that OTD is able to identify the cross-tissue and   180

tissue-specific components via simulations. We generated simulated traits so that the   181

true cross-tissue and tissue-specific components are known. To preserve the correlation   182

structure between genes and tissues, we considered the estimated cross-tissue and   183

tissue-specific components as the true values. Then, we simulated expression traits as   184

the sum of the cross-tissue and tissue-specific components and an error term (expression   185

trait = cross-tissue + tissue-specific + noise). We simulated expression traits using 186
different magnitudes of the noise level and then applied our OTD approach. Simulated 187
values were constructed using noise variance equal in magnitude to the original variance 188
of the expression levels of each gene. This is a rather conservative choice since the 189
variance of the noise component is unlikely to be larger than the total variance of the 190
expression trait. In S6 Fig, we show for one simulated example the correlation between 191
true and OTD-estimated components for one representative gene, *SLTM*, and the 192
distribution of the correlation between true and OTD-estimates for all genes. The 193
median Pearson correlation between true and OTD-estimted cross-tissue components 194
was 0.54 [95% CI: 0.33-0.77] and for tissue-specific components was 0.75 [0.46-0.88], 195
demonstrating the robustness of our model (S6 Fig). 196

For the OTD derived cross-tissue and 40 tissue-specific expression phenotypes, we 197
computed the local heritability and generated prediction models. The decomposition is 198
applied at the expression trait level so that the downstream genetic regulation analysis 199
is performed separately for each derived trait, cross-tissue and tissue-specific expression, 200
which greatly reduces computational burden. 201

Cross-tissue expression phenotype is less noisy and shows 202 higher predictive performance 203

Our estimates of h^2 for cross-tissue expression traits are larger than the corresponding 204
estimates for each whole tissue (S7 Fig) because our OTD approach increases the ratio 205
of the genetically regulated component to noise by averaging across multiple tissues. In 206
addition to the increased h^2, we observe reduction in standard errors of the estimated 207
cross-tissue h^2. This is due, in part, to the larger effective sample size for cross-tissue 208
phenotypes. There were 450 samples for which cross-tissue traits were available whereas 209
the maximum sample size for whole tissue phenotypes was 362. Similarly, cross-tissue 210
BSLMM PVE estimates had lower error than whole tissue PVE (S8 Fig, S9 Fig). 211

As for the tissue-specific components, the cross-tissue heritability estimates were also 212
larger and the standard errors were smaller reflecting the fact that a substantial portion 213
of regulation is common across tissues (S10 Fig). The percentage of GCTA h^2 estimates 214
with FDR <0.1 was much larger for cross-tissue expression (20%) than the 215

tissue-specific expressions (6-13%, 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 (S9 Fig).

Cross-tissue predictive performance exceeded that of both tissue-specific and whole tissue expression as indicated by higher cross-validated R^2 (S11 Fig). Like whole tissue expression, cross-tissue and tissue-specific expression showed higher predictive performance when using more sparse models. In other words elastic-net models with $\alpha \geq 0.5$ predicted better than the ones with $\alpha = 0.05$ (S11 Fig).

Cross-tissue expression phenotype recapitulates published multi-tissue eQTL results

To verify that the cross-tissue phenotype has the properties we expect, we compared our OTD results to those from a joint multi-tissue eQTL analysis [33], which was previously performed on a subset of the GTEx data [27] 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 (see Methods).

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 a uniformity statistic that combines the nine posterior probabilities into one value such that higher values mean the gene regulation is more uniform across all nine tissues, rather than in just a small subset of the nine.

Thus, we expected that genes with high cross-tissue heritability would show high probability of being active in multiple tissues and have high uniformity measure. Reassuringly, this is exactly what we find. Genes with high cross-tissue heritability concentrate on the higher end of the uniformity measure (Fig. 6, S12 Fig).
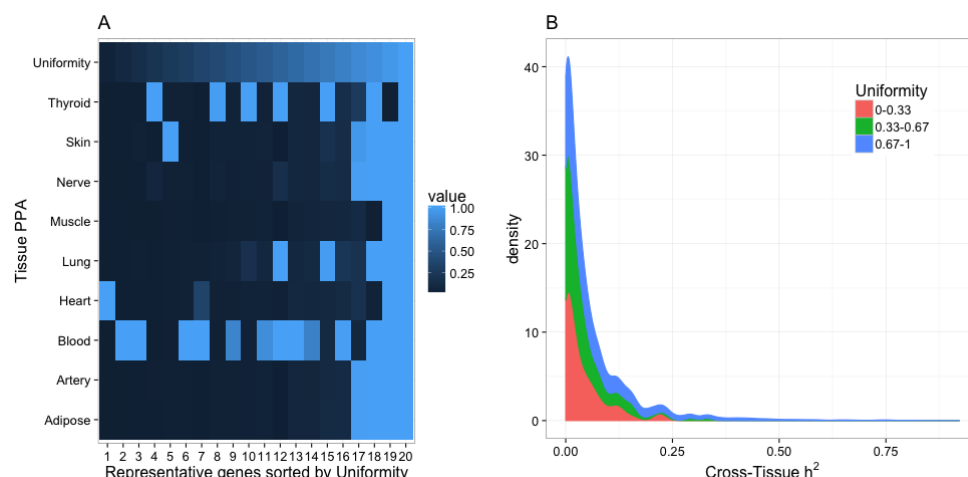
Figure 6. Measure of uniformity of the posterior probability of active regulation vs. cross-tissue heritability. Uniformity was computed using the posterior probability of a gene being actively regulated in a tissue, PPA, from the Flutre et al. [33] multi-tissue eQTL analysis. **(A)** Representative examples showing that 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. **(B)** This panel shows the distribution of heritability of the cross-tissue component vs. a measure of uniformity of genetic regulation across tissues. The Kruskal-Wallis rank sum test revealed a significant difference in the cross-tissue h^2 of uniformity groups ($\chi^2 = 31.4$, $P < 10^{-6}$).

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 7A 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 [27]. In contrast, Figure 7B 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.
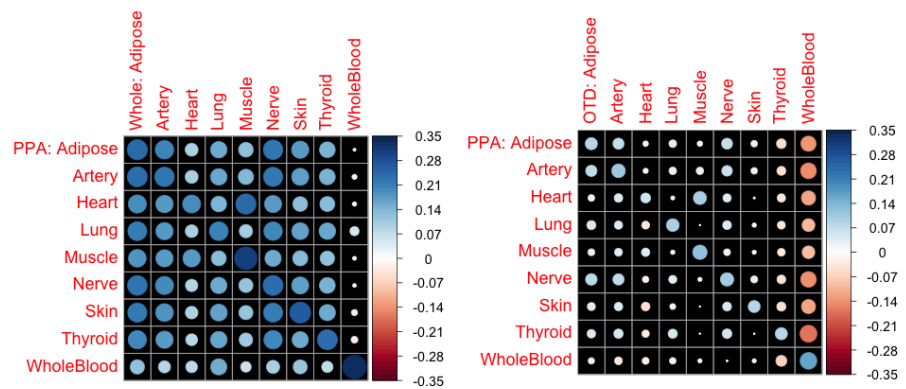
Figure 7. Comparison of heritability of whole tissue or tissue-specific components vs. PPA. Panel (**A**) of this figure shows the Pearson correlation (R) between the BSLMM PVE of the original (we are calling whole here) tissue expression levels vs. the probability of the tissue being actively regulated 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. Distal components estimates were too noisy to draw meaningful conclusions.

Using results implied by the improved predictive performance of sparse models and by directly estimating sparsity using BSLMM (Bayesian Sparse Linear Mixed Model), we show evidence that local common variant regulation is sparse across all the tissues analyzed here. Our finding that a handful of genetic variants seem to contribute to the variability in gene expression traits has important implications for the strategies used in future investigations of gene regulation and downstream effects on complex traits. For example, most fine mapping methods that attempt to find the causal SNPs that drive a

GWAS locus focus on a limited number of variants. This only makes sense in cases where the underlying genetic architecture is sparse, that is, when a handful of causal variants are determining the variability of the traits. We note that any rare variants not tagged by common variants are not included in our gene expression heritability estimates and prediction models. For genes with moderate and low heritability the evidence is not as strong, due to reduced power, but results are consistent with a sparse local architecture. Sparse models capture the most variance in gene expression at current sample sizes and have been used successfully in gene expression prediction methods [15, 36–38]. Better methods to correct for hidden confounders that do not dilute distal signals and larger sample sizes will be needed to determine the properties of distal regulation.

Given that a substantial portion of local regulation is shared across tissues, we proposed here to decompose the expression traits into cross-tissue and tissue-specific components. This approach, called orthogonal tissue decomposition (OTD), aims to decouple the shared regulation from the tissue-specific regulation. We examined the genetic architecture of these derived traits and find that they follow similar patterns to the original whole tissue expression traits. The cross-tissue component benefits from an effectively larger sample size than any individual tissue trait, which is reflected in more accurate heritability estimates and consistently higher prediction performance. Encouragingly, we find that genes with high cross-tissue heritability tend to be regulated more uniformly across tissues. As for the tissue-specific expression traits, we found that they recapitulate correlation with the vector of probability of tissue-specific regulation. Many groups have proposed integrating genotype and expression data to understand complex traits [15, 22, 37–40]. Through integration of our OTD expression traits with studies of complex diseases, we expect results from the cross-tissue models to relate to mechanisms that are shared across multiple tissues, whereas results from the tissue-specific models will inform us about the context specific mechanisms.

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, BSLMM and the elastic net, we show that for common local gene regulation, the genetic architecture is mostly sparse rather than

polygenic. Using new expression phenotypes generated in our OTD model, we show that 306
cross-tissue predictive performance exceeded that of both tissue-specific and whole tissue 307
expression as indicated by higher elastic net cross-validated R^2. Predictors, heritability 308
estimates and cross-validation statistics generated in this study of gene expression 309
architecture are freely available (`https://github.com/hakyimlab/PrediXcan`) for use 310
in future studies of complex trait genetics. 311

Methods 312

Genomic and Transcriptomic Data 313

DGN Dataset. We obtained whole blood RNA-seq and genome-wide genotype data 314
for 922 individuals from the Depression Genes and Networks (DGN) cohort [26], all of 315
European ancestry. For all analyses, we used the HCP (hidden covariates with prior) 316
normalized gene-level expression data used for the *trans*-eQTL analysis in Battle et 317
al. [26] and downloaded from the NIMH repository. The 922 individuals were unrelated 318
(all pairwise $\hat{\pi} < 0.05$) and thus all included in downstream analyses. Imputation of 319
approximately 650K input SNPs (minor allele frequency [MAF] > 0.05, 320
Hardy-Weinberg Equilibrium [P > 0.05], non-ambiguous strand [no A/T or C/G SNPs]) 321
was performed on the Michigan Imputation Server 322
(`https://imputationserver.sph.umich.edu/start.html`) [41, 42] with the following 323
parameters: 1000G Phase 1 v3 ShapeIt2 (no singletons) reference panel, SHAPEIT 324
phasing, and EUR population. Approximately 1.9M non-ambiguous strand SNPs with 325
MAF > 0.05, imputation $R^2 > 0.8$ and, to reduce computational burden, inclusion in 326
HapMap Phase II were retained for subsequent analyses. 327

GTEx Dataset. We obtained RNA-seq gene expression levels from 8555 tissue 328
samples (53 unique tissue types) from 544 unique subjects in the GTEx Project [27] 329
data release on 2014-06-13. Of the individuals with gene expression data, genome-wide 330
genotypes (imputed with 1000 Genomes) were available for 450 individuals. While all 331
8555 tissue samples were used in the OTD model (described below) to generate 332
cross-tissue and tissue-specific components of gene expression, we used the 40 tissues 333
with the largest sample sizes ($n > 70$) when quantifying tissue-specific effects (see S1 334

Table). Approximately 2.6M non-ambiguous strand SNPs included in HapMap Phase II ₃₃₅
were retained for subsequent analyses. ₃₃₆

Framingham Expression Dataset.   We obtained exon array expression and ₃₃₇
genotype array data from 5257 individuals from the Framingham Heart Study [43]. The ₃₃₈
final sample size after quality control was 4286. We used the Affymetrix power tools ₃₃₉
(APT) suite to perform the preprocessing and normalization steps. First the robust ₃₄₀
multi-array analysis (RMA) protocol was applied which consists of three steps: ₃₄₁
background correction, quantile normalization, and summarization [44]. The ₃₄₂
background correction step uses antigenomic probes that do not match known genome ₃₄₃
sequences to adjust the baseline for detection, and is applied separately to each array. ₃₄₄
Next, the normalization step utilizes a 'sketch' quantile normalization technique instead ₃₄₅
of a memory-intensive full quantile normalization. The benefit is a much lower memory ₃₄₆
requirement with little accuracy trade-off for large sample sets such as this one. Next, ₃₄₇
the adjusted probe values were summarized (by the median polish method) into ₃₄₈
log-transformed expression values such that one value is derived per exon or gene. The ₃₄₉
summarized expression values were then annotated more fully using the annotation ₃₅₀
databases contained in the huex10stprobeset.db (exon-level annotations) and ₃₅₁
huex10sttranscriptcluster.db (gene-level annotations) R packages available from ₃₅₂
Bioconductor [45, 46]. In both cases gene annotations were provided for each feature. ₃₅₃

Plink [47] was used for data wrangling and cleaning steps. The data wrangling steps ₃₅₄
included updating probe IDs, unifying data to the positive strand, and updating ₃₅₅
locations to GRCh37. The data cleaning steps included a step to filter for variant and ₃₅₆
subject missingness and minor alleles, one to filter variants with Hardy-Weinberg exact ₃₅₇
test, and a step to remove unusual heterozygosity. To make our data compatible with ₃₅₈
the Haplotype Reference Consortium (HRC) panel ₃₅₉
(`http://www.well.ox.ac.uk/~wrayner/tools/`), we used the HRC-check-bin tool in ₃₆₀
order to carry out data wrangling steps required. The data were then split by ₃₆₁
chromosome and pre-phased with SHAPEIT [48] using the 1000 Genomes phase 3 panel ₃₆₂
and converted to vcf format. These files were then submitted to the Michigan ₃₆₃
Imputation Server (`https://imputationserver.sph.umich.edu/start.html`) [41, 42] ₃₆₄
for imputation with the HRC version 1 panel [49]. We applied Matrix eQTL [50] to the ₃₆₅

normalized expression and imputed genotype data to generate prior eQTLs for our   366
heritability analysis.   367

Partitioning local and distal heritability of gene expression   368

Motivated by the observed differences in regulatory effect sizes of variants located in the   369
vicinity of the genes and distal to the gene, we partitioned the proportion of gene   370
expression variance explained by SNPs in the DGN cohort into two components: local   371
(SNPs within 1Mb of the gene) and distal (eQTLs on non-gene chromosomes) as defined   372
by the GENCODE [51] version 12 gene annotation. We calculated the proportion of the   373
variance (narrow-sense heritability) explained by each component using the following   374
mixed-effects model: 375

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

where Y_g represents the expression of gene g, X_k is the allelic dosage for SNP k, local   376
refers to the set of SNPs located within 1Mb of the gene's transcription start and end,   377
distal refers to SNPs in other chromosomes, and ϵ is the error term representing   378
environmental and other unknown factors. We assume that the local and distal   379
components are independent of each other as well as independent of the error term. We   380
assume random effects for $w_{k,g}^{\text{local}} \sim N(0, \sigma_{w,\text{local}}^2)$, $w_{k,g}^{\text{distal}} \sim N(0, \sigma_{w,\text{distal}}^2)$, and   381
$\epsilon \sim N(0, \sigma_\epsilon^2 I_n)$, where I_n is the identity matrix. We calculated the total variability   382
explained by local and distal components using restricted maximum likelihood (REML)   383
as implemented in the GCTA software [28].   384

In an effort to determine if distal heritability estimates could be improved, we also   385
tested a mixed-effect model restricting the distal component to known eQTLs on   386
non-gene chromosomes discovered in the Framingham cohort at FDR < 0.05. When we   387
found the distal estimate could not be improved, we focused on estimating local   388
heritability without the distal component in the equation above.   389

For the purpose of estimating the mean heritability (see Table 1, ~~Figure ??~~ and S1   390
Table), we allowed the heritability estimates to take negative values (unconstrained   391
model). Despite the lack of obvious biological interpretation of a negative heritability, it   392
is an accepted procedure used in order to avoid bias in the estimated mean [10, 29].   393

Genes with FDR < 0.1 (derived from the two-sided GCTA P-value) were considered to have significant heritability.

For comparing to BSLMM PVE, we restricted the GCTA heritability estimates to be within the [0,1] interval (constrained model, see Figures 2, 4 and 6).

Quantifying sparsity with Bayesian Sparse Linear Mixed Modeling (BSLMM)

We used BSLMM [14] to model the effect of local genetic variation (common SNPs within 1 Mb of gene) on the genetic architecture of gene expression. BSLMM uses a linear model with a polygenic component (small effects) and a sparse component (large effects) enforced by sparsity inducing priors on the regression coefficients [14]. BSLMM assumes the genotypic effects come from a mixture of two normal distributions and thus is flexible to both polygenic and sparse genetic architectures [14]. We used the software GEMMA [54] to implement BSLMM for each gene with 100K sampling steps per gene. BSLMM estimates the PVE (the proportion of variance in phenotype explained by the additive genetic model, analogous to the heritability estimated 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).

Determining polygenicity versus sparsity using the elastic net

We used the *glmnet* R 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 [52, 53].

The elastic net penalty is controlled by mixing parameter α, which spans LASSO ($\alpha = 1$, the default) [19] at one extreme and ridge regression ($\alpha = 0$) [20] 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. In order to compare prediction R^2 values across α values, we used common folds and seeds for each run. For the rest of the autosomes in DGN and for whole tissue, cross-tissue, and tissue-specific expression in the GTEx cohort, we tested $\alpha = 0.05, 0.5, 1$.

Orthogonal Tissue Decomposition (OTD)

We use a mixed effects model to decompose the expression level of a gene into a subject-specific component and a subject-by-tissue-specific component. We fit the model one gene at a time and to simplify notation we assume the gene index, g, is implicit and drop it from the equations below. The expression Y of a gene g for individual i in tissue t is modeled as

$$Y_{i,t,g} = Y_{i,t} = Y_i^{\text{CT}} + Y_{i,t}^{\text{TS}} + \mathbf{Z}_i'\boldsymbol{\beta} + \epsilon_{i,t}$$

where Y_i^{CT} is the random subject level intercept, $Y_{i,t}^{\text{TS}}$ 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 $Y_i^{\text{CT}} \sim N(0, \sigma_{\text{CT}}^2)$, $Y_{i,t}^{\text{TS}} \sim N(0, \sigma_{\text{TS}}^2)$, $\epsilon \sim N(0, \sigma_\epsilon^2)$, and that all three terms are independent of each other. All variances are estimated by restricted maximum likelihood (REML). $Y_{i,t,g} = Y_{i,t}$ is a scalar. $\mathbf{Z}_i$ is a vector of length p, which represents the number of covariates. $\boldsymbol{\beta}$ is a vector of length p and represents the effects of covariates on the expression level of the gene. The prime in $\mathbf{Z}_i'\boldsymbol{\beta}$ represents the inner product between the two vectors. All variances are estimated by restricted maximum likelihood (REML). These variances will be different for each gene. The tissue-specific component's variance is common across tissues. Differences between tissues will be reflected in the posterior mean of each tissue/individual random effects.

For the cross-tissue component (Y_i^{CT}) to be identifiable, multiple replicates of expression are needed for each subject. In the same vein, for the tissue-specific component ($Y_{i,t}^{\text{TS}}$) to be identifiable, multiple replicates of expression are needed for a

given tissue/subject pair. GTEx [27] data consisted of expression measurement in 451

multiple tissues for each subject, thus multiple replicates per subject were available. 452

However, there were very few replicated measurement for a given tissue/subject pair. 453

Thus, we fit the reduced model and use the estimates of the residual as the 454

tissue-specific component. 455

$$Y_{i,t} = Y_i^{\mathrm{CT}} + \mathbf{Z}_i' \boldsymbol{\beta} + \epsilon_{i,t}$$

We consider the expression level of a gene at a given tissue for individual i to be 456

composed of a cross-tissue component which we represent as Y_i^{CT} and a tissue-specific 457

component, which given the lack of replicates, we estimate as the difference between the 458

expression level and the cross-tissue components (after adjusting for covariates). The 459

assumptions are the same as the full model, i.e., $Y_i^{\mathrm{CT}} \sim N(0, \sigma_{\mathrm{CT}}^2)$, $\epsilon \sim N(0, \sigma_\epsilon^2)$, and 460

that both terms are independent of each other. 461

The mixed effects model parameters were estimated using the `lme4` package [55] in 462

R. Batch effects and unmeasured confounders were accounted for using 15 `PEER` factors 463

computed with the `PEER` [34] package in R. We also included gender as a covariate. 464

Posterior modes of the subject level random intercepts were used as estimates of the 465

cross-tissue components whereas the residuals of the models were used as tissue-specific 466

components. 467

The model included whole tissue gene expression levels in 8555 GTEx tissue samples 468

from 544 unique subjects. A total of 17,647 protein-coding genes (defined by 469

GENCODE [51] version 18) with a mean gene expression level across tissues greater 470

than 0.1 RPKM (reads per kilobase of transcript per million reads mapped) and RPKM 471

> 0 in at least 3 individuals were included in the model. 472

OTD simulation 473

To test the robustness of our method, we generated simulated expression levels so that 474

we could compare our estimates with the ground truth. To preserve correlation 475

structure between genes and tissues, we used the cross-tissue and tissue-specific 476

estimates of OTD based on the observed data as a basis for the simulations. Thus, we 477

simulated expression phenotypes as the sum of the observed OTD cross-tissue and 478

tissue-specific components plus an error term derived by randomly sampling a normal 479

distribution with a standard deviation equal to 1, 2, or 10 times the variance of each   480
gene's component sum. We performed OTD on the simulated phenotypes and compared   481
the OTD-estimated cross-tissue and tissue-specific expression levels for each gene to the   482
true levels using Pearson correlation. As expected, higher levels of noise reduced the   483
correlation between true and simulated values, but the correlation remained significant   484
for all simulated phenotypes.   485

Comparison of OTD trait heritability with multi-tissue eQTL   486
results   487

To verify that the newly derived cross-tissue and tissue-specific traits were capturing the   488
expected properties, we used the results of the multi-tissue eQTL analysis developed by   489
Flutre et al. [33] and performed on nine tissues from the pilot phase of the GTEx   490
project [27]. In particular, we downloaded the posterior probabilities of a gene being   491
actively regulated in a tissue (PPA) from the GTEx portal at   492
`http://www.gtexportal.org/static/datasets/gtex_analysis_pilot_v3/multi_`   493
`tissue_eqtls/Multi_tissue_eQTL_GTEx_Pilot_Phase_datasets.tar`. PPA can be   494
interpreted as the probability a gene is regulated by an eQTL in tissue t given the data.   495
In the GTEx pilot, the most significant eQTL per gene was used to compute PPA [27].   496
PPA is computed from a joint analysis of all tissues and takes account of sharing of   497
eQTLs among tissues [33]. For example, consider a SNP showing modest association   498
with expression in tissue t. If this SNP also shows strong association in the other   499
tissues, then it will be assigned a higher probability of being an active eQTL in tissue t 500
than if it showed no association in the other tissues [33].   501

We reasoned that genes with large cross-tissue component (i.e. high cross-tissue h^2)   502
would have more uniform PPA across tissues. Thus we defined for each gene a measure   503
of uniformity, U_g, across tissues based on the nine-dimensional vector of PPAs using the   504
entropy formula. More specifically, we divided each vector of PPA by their sum across   505
tissues and computed the measure of uniformity as follows: 506

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

where $p_{t,g}$ is the normalized PPA for gene g and tissue t.   507

Acknowledgments

Computing resources.

538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564

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. 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`.

12. 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`.

13. 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`.

14. 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`.

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. Brown CD, Mangravite LM, Engelhardt BE. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. PLoS Genet. 2013;9(8):e1003649.

17. King EG, Sanderson BJ, McNeil CL, Long AD, Macdonald SJ. Genetic dissection of the Drosophila melanogaster female head transcriptome reveals widespread allelic heterogeneity. PLoS Genet. 2014;10(5):e1004322.

18. Zhang X, Cal AJ, Borevitz JO. Genetic architecture of regulatory variation in Arabidopsis thaliana. Genome research. 2011;21(5):725–733.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. Abraham G, Kowalczyk A, Zobel J, Inouye M. Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. Genetic Epidemiology. 2013;37(2):184–195.

25. 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`.

26. 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`.

27. 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`.

28. 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`.

29. 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`.

30. 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.

31. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. The American Journal of Human Genetics. 2012;91(6):1011–1021.

32. Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. BioRxiv. 2016;p. 030338.

33. 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`.

34. 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`.

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

36. Manor O, Segal E. Robust prediction of expression differences among human individuals using only genotype information. PLoS Genet. 2013;9(3):e1003396.

37. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics. 2016;.

38. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature genetics. 2016;.

39. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383.

40. He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, et al. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. The American Journal of Human Genetics. 2013;92(5):667–680.

41. 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`.

42. 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.

43. 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.

44. 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.

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

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

47. 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).

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

49. 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.

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

51. 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.

52. 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/`.

53. 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/`.

54. 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`.

55. 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`.

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

57. 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.

58. 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
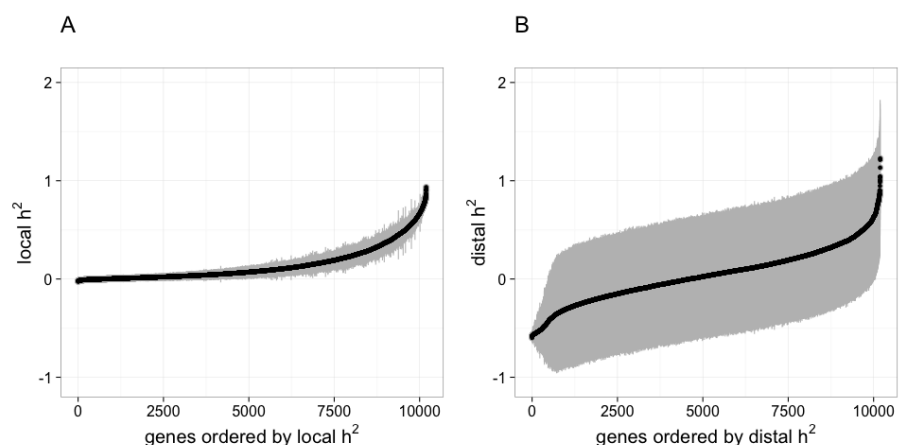
Local and distal heritability (h^2) of gene expression levels in whole blood from the DGN RNA-seq dataset. In order to obtain an unbiased estimate of mean h^2, we allow the values to be negative when fitting the restricted maximum likelihood (unconstrained REML). (**A**) Local (SNPs within 1Mb of each gene) and (**B**) distal (all SNPs on non-gene chromosomes) gene expression h^2 estimates ordered by increasing h^2. Gray segments show two times the standard errors of each h^2 estimate. Notice the larger range and errors of the distal estimates.

S2 Fig
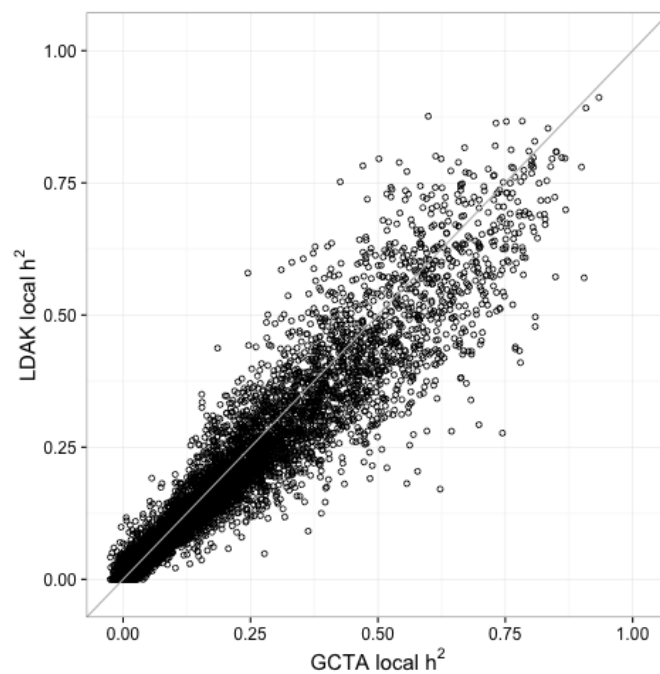
DGN local heritability estimation comparison between LDAK and GCTA.
The Pearson correlation between GCTA and the LD-adjusted kinship local h^2 estimates
of gene expression is 0.96.

S3 Fig
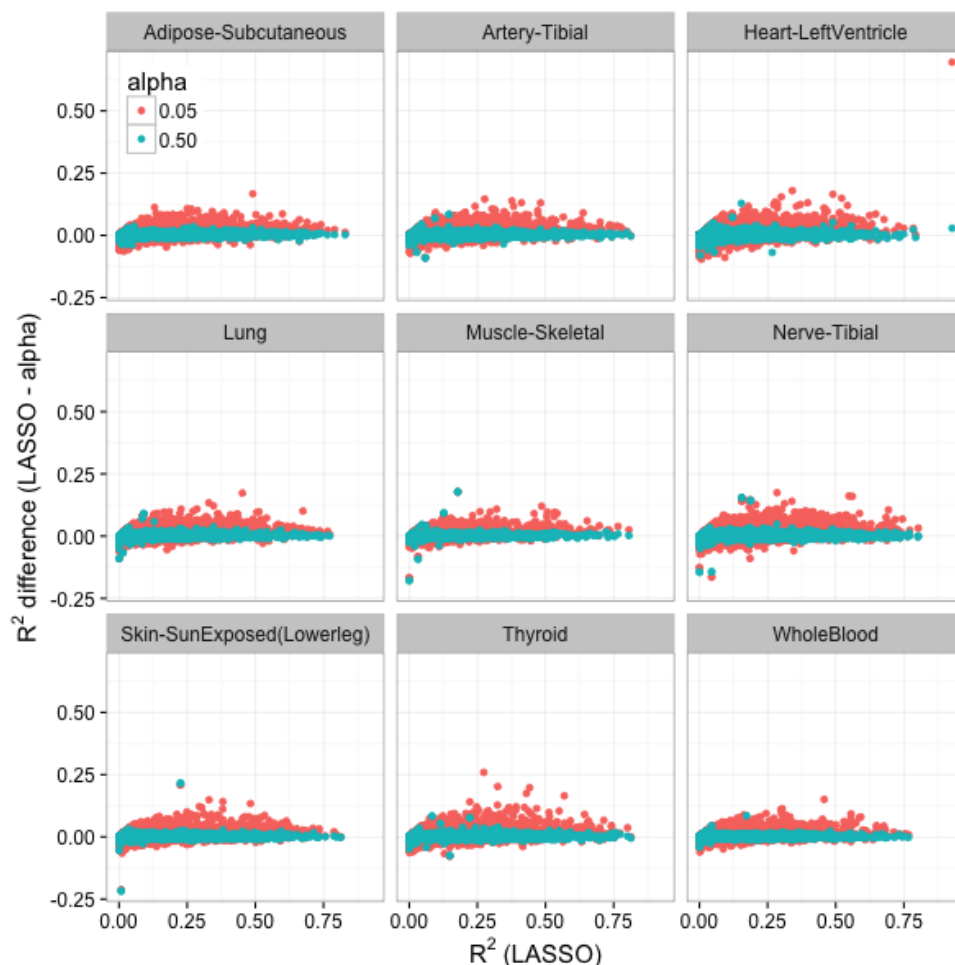
GTEx whole tissue cross-validated predictive performance across the elastic net. The difference between the cross validated R^2 of the LASSO model and the elastic net model mixing parameters 0.05 and 0.5 for autosomal protein coding genes per tissue. Elastic net with $\alpha = 0.5$ values hover around 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.

S4 Fig
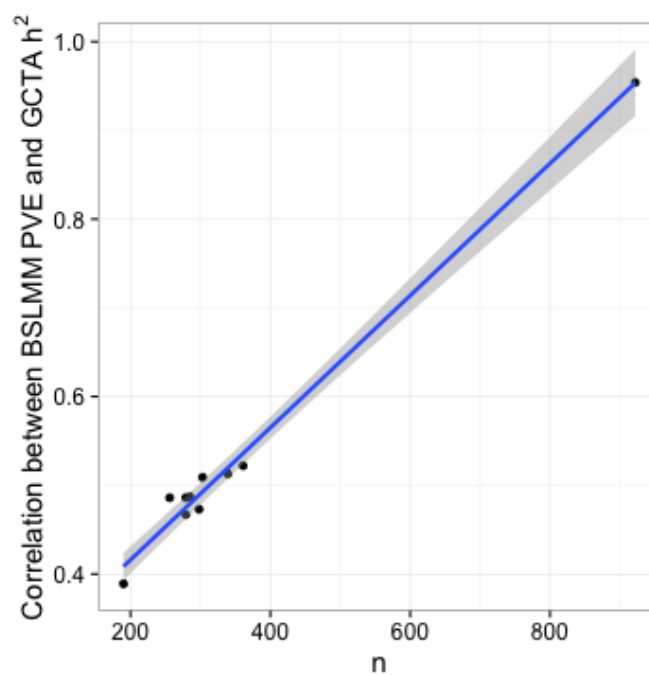
Correlation between heritability models compared to sample size. The Pearson correlation between estimates of heritability using BSLMM vs. GCTA is dependent on sample size (R $= 0.99$, $P = 3x10^{-9}$).

S5 Fig
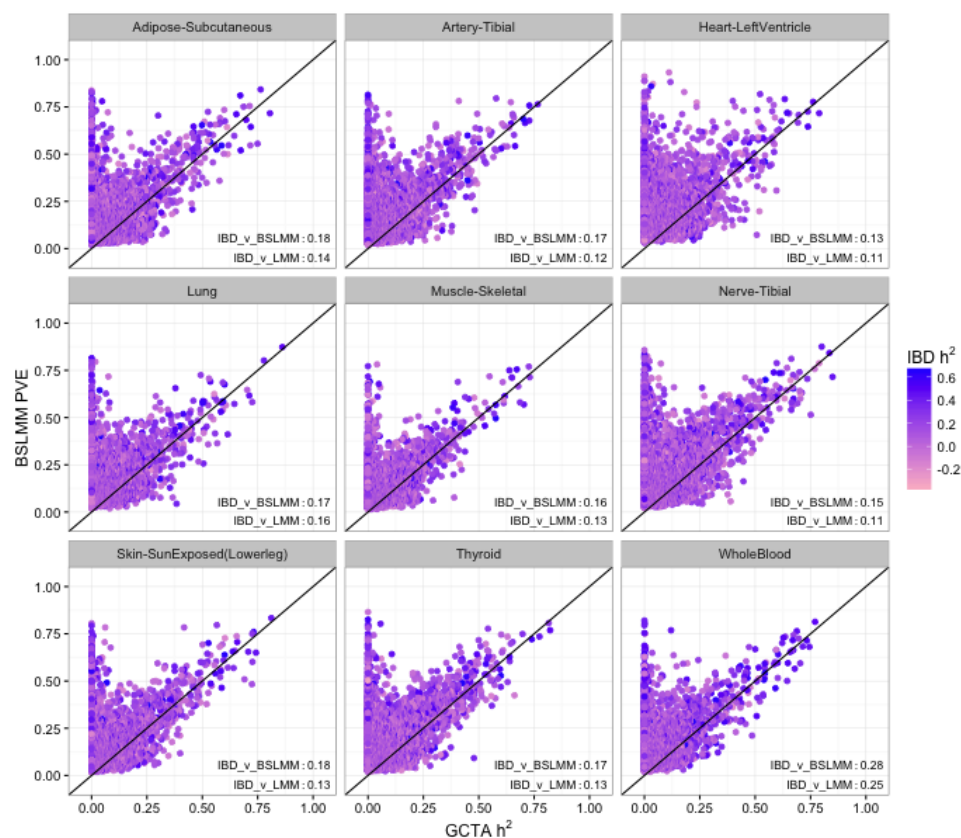
BSLMM and LMM estimates of heritability in GTEx compared to IBD estimates This figure shows the comparison between estimates of heritability using BSLMM vs. LMM (GCTA) for GTEx data. Here, in both models the estimates are constrained to be between 0 and 1. For most genes BSLMM estimates are larger than LMM estimates reflecting the fact that BSLMM yields better estimates of heritability because of its ability to account for the sparse component. Each point is colored according to that gene's estimate of h^2 by shared identity by descent (IBD) in Price et al. [29]. At the bottom left of each panel, we show the IBD correlation with BSLMM (IBD_v_BSLMM) and LMM (IBD_v_LMM). BSLMM is consistently more correlated with the IBD estimate. This provides further evidence that LMM is underestimating h^2 for these genes.

S6 Fig
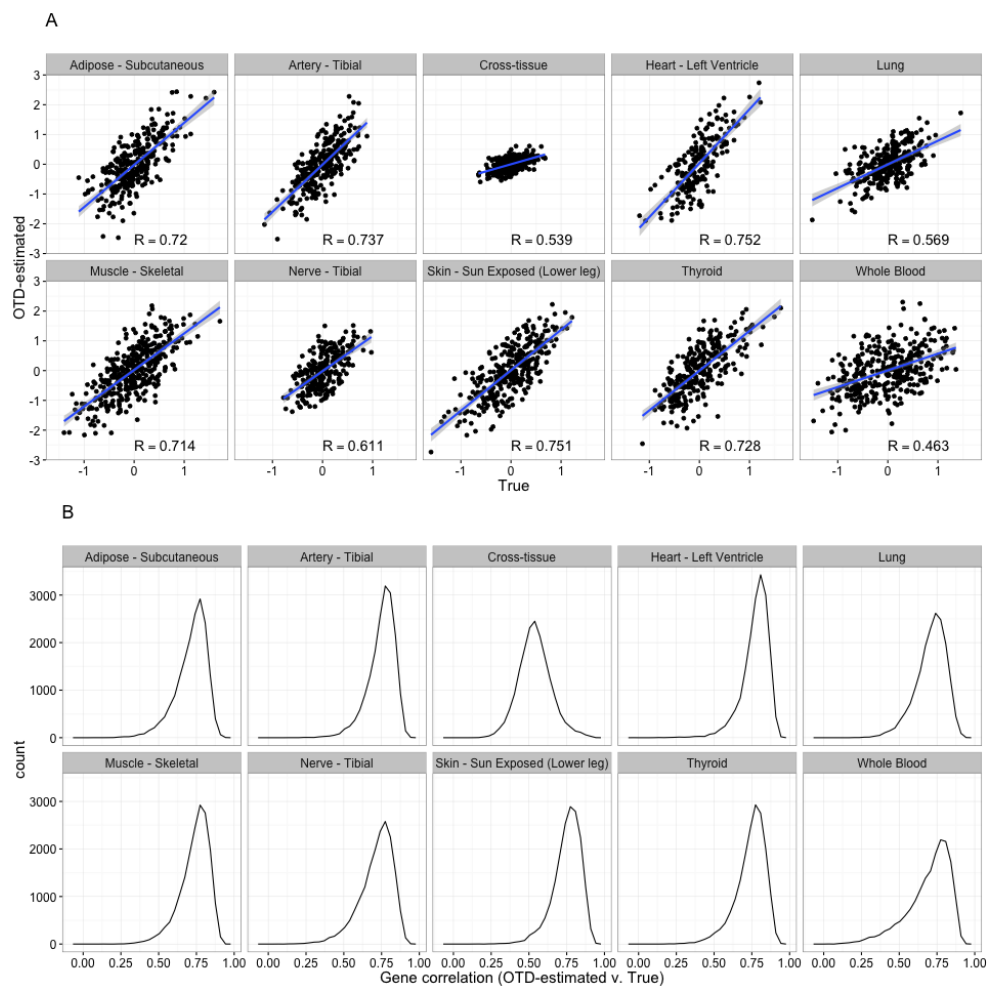
Identifiability of cross-tissue and tissue-specific components shown in simulated data. The simulated phenotype is the sum of the observed cross-tissue and tissue-specific components (estimated via OTD in the observed data) plus an error term derived by randomly sampling a normal distribution with variance equal to the variance of each gene. **(A)** Correlation between true and OTD-estimated cross-tissue and tissue-specific components for one gene, *SLTM*. **(B)** Histogram of the correlation between true and OTD-estimated components across all genes.

S7 Fig
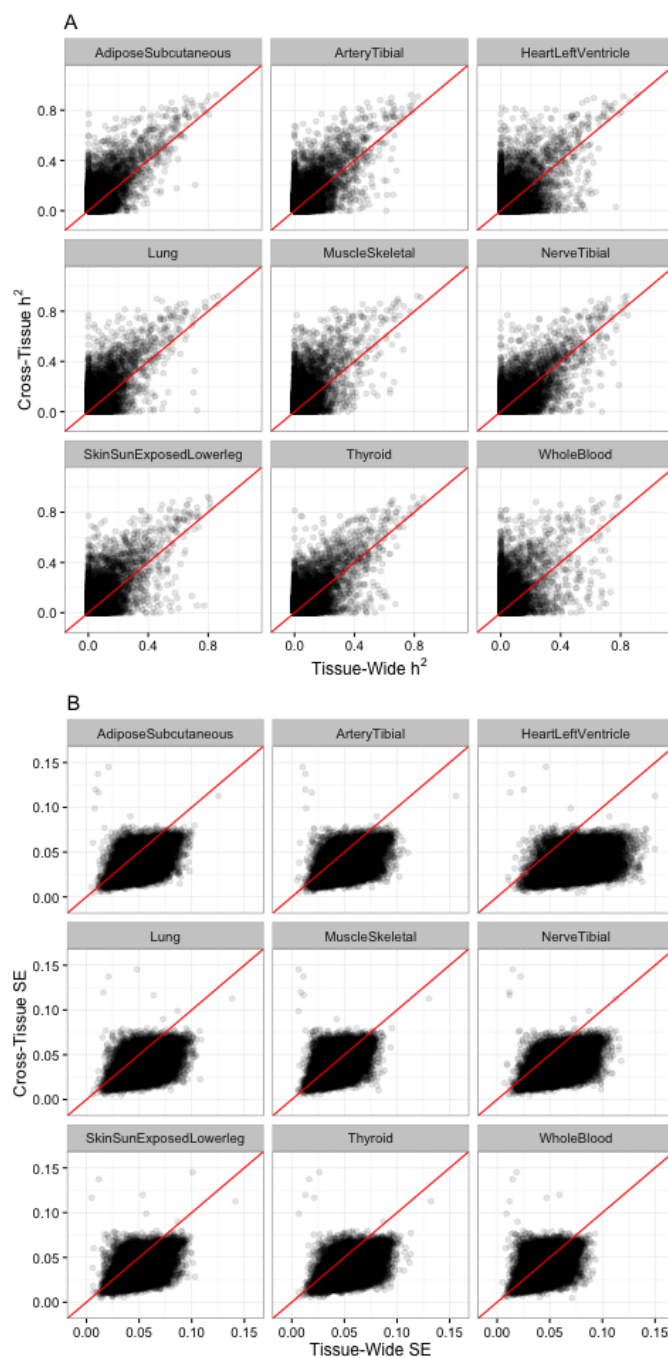
Cross-tissue and whole tissue comparison of heritability (h^2, A) and standard error (SE, B). Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Whole tissue local h^2 is estimated using 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.

S8 Fig
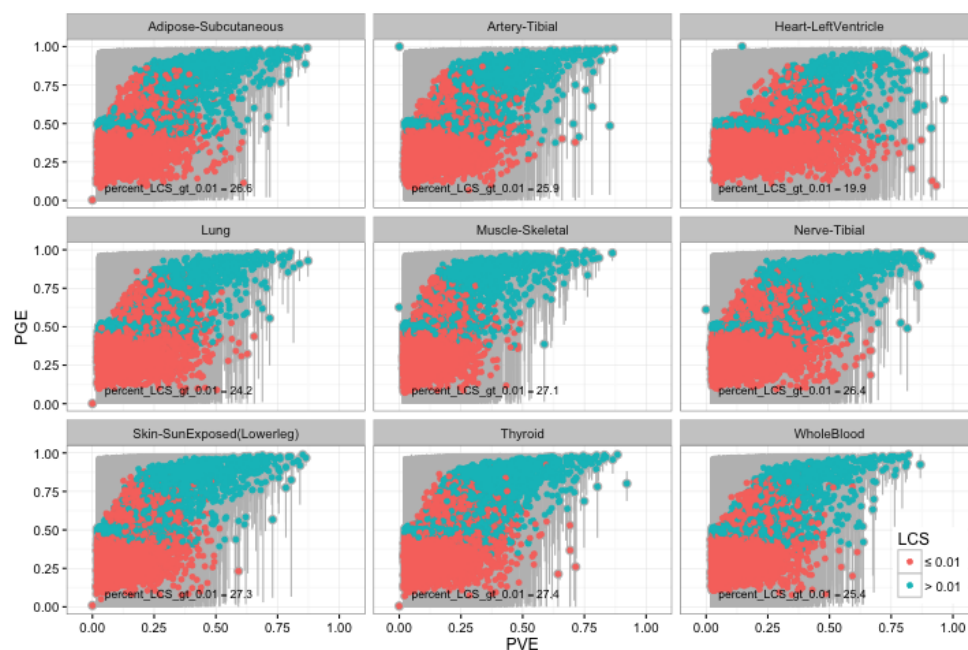
GTEx whole tissue expression Bayesian Sparse Linear Mixed Model.
Comparison of median PGE (proportion of PVE explained by sparse effects) to median PVE (total proportion of variance explained, the BSLMM equivalent of h^2) for expression of each gene. The 95% credible set of each PGE estimate is in gray 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.

S9 Fig

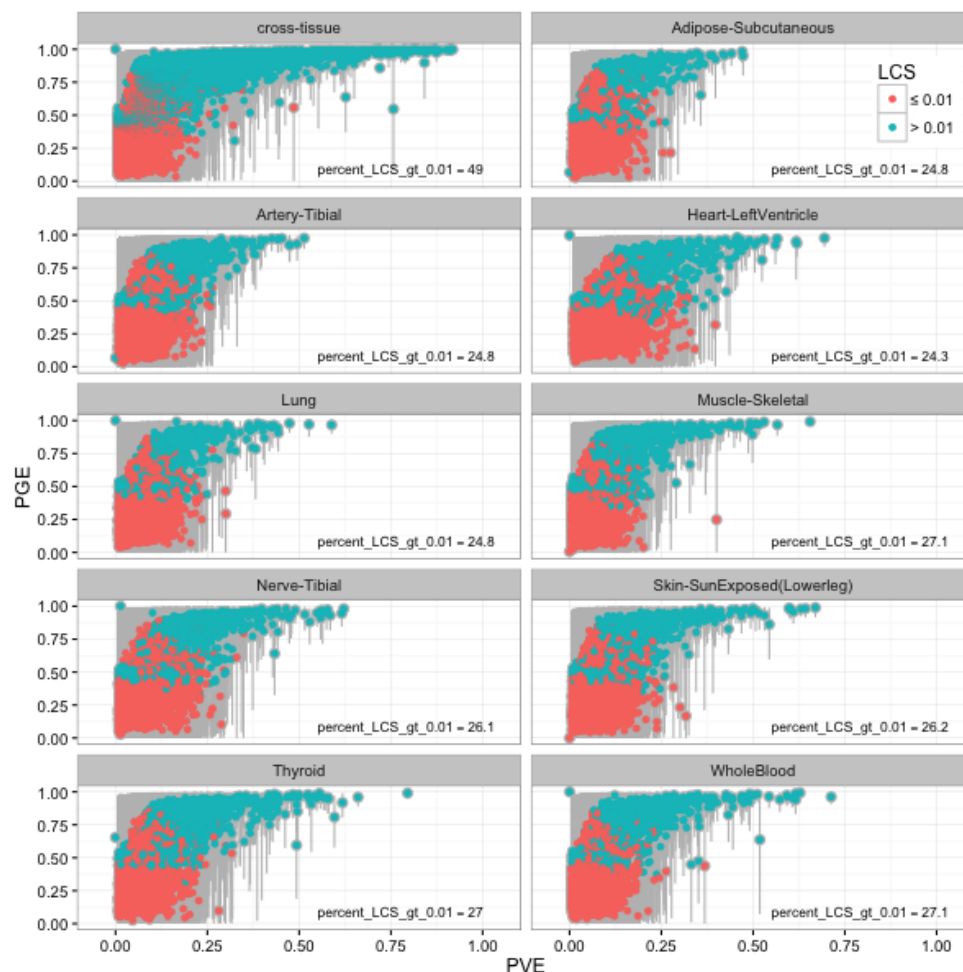**GTEx orthogonal tissue decomposition cross-tissue and tissue-specific expression Bayesian Sparse Linear Mixed Model.** Comparison of median PGE (proportion of PVE explained by sparse effects) to median PVE (total proportion of variance explained, the BSLMM equivalent of h^2) for expression of each 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.

S10 Fig
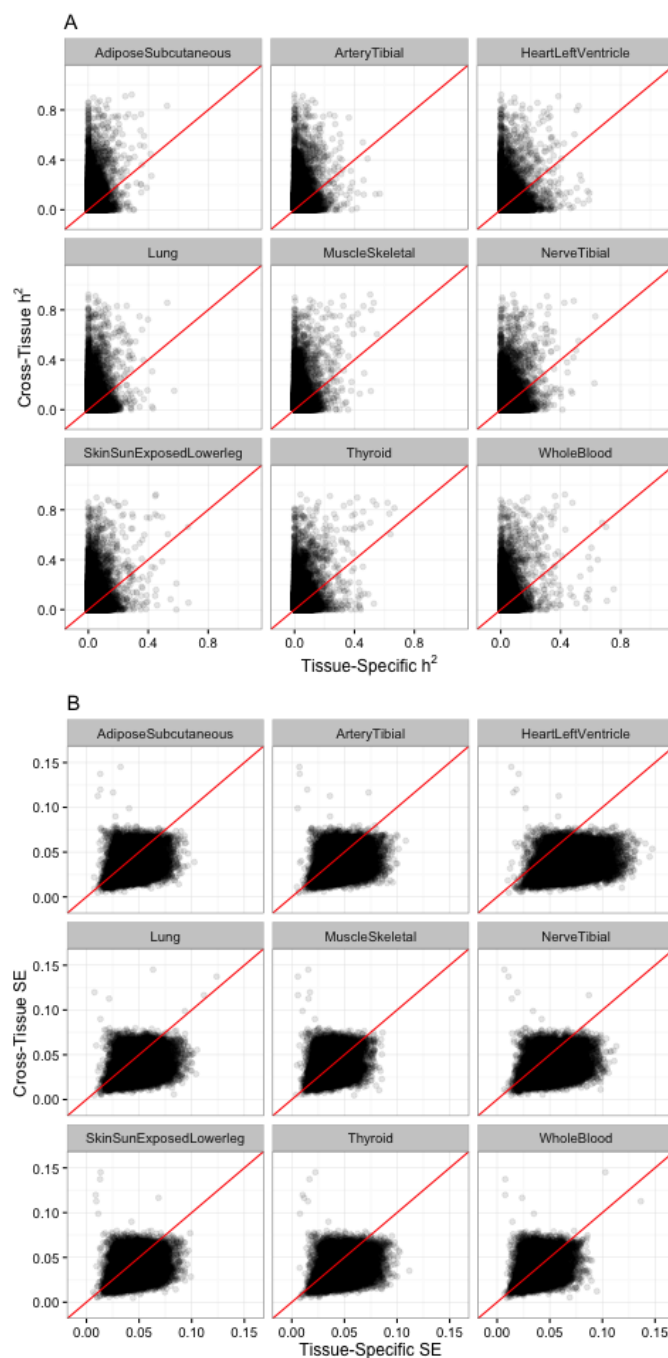
Cross-tissue and tissue-specific comparison of heritability (h^2, A) and standard error (SE, B) estimation. Cross-tissue local h^2 is estimated using the cross-tissue component (random effects) of the mixed effects model for gene expression and SNPs within 1 Mb of each gene. Tissue-specifc local h^2 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.

S11 Fig
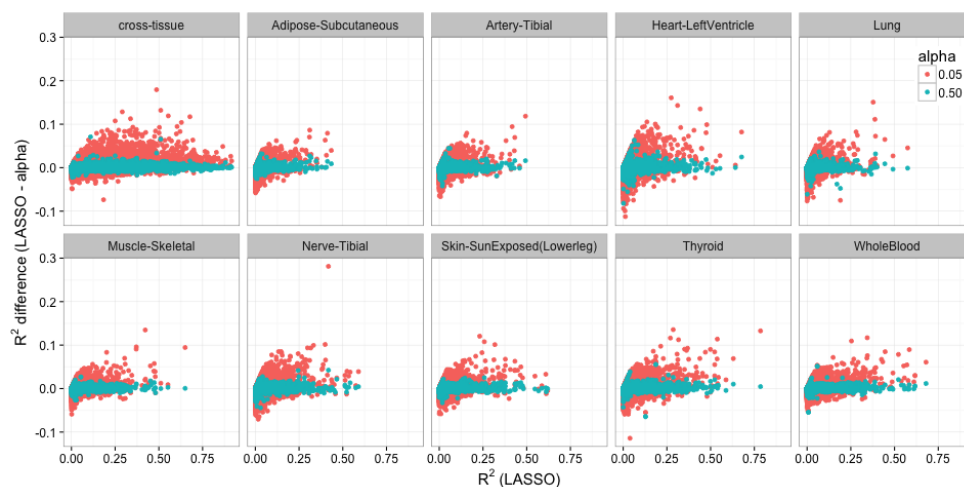
GTEx orthogonal tissue decomposition cross-tissue and tissue-specific expression cross-validated predictive performance across the elastic net.
The difference between the cross validated R^2 of the LASSO model and the elastic net model mixing parameters 0.05 and 0.5 for autosomal protein coding genes 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.

S12 Fig
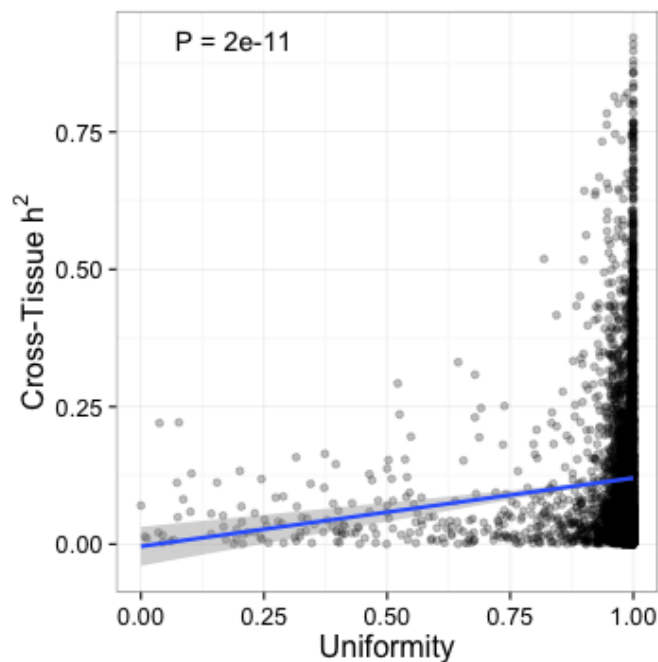
Continuous distribution of heritability of the cross-tissue component vs. a measure of uniformity of genetic regulation across tissues. The measure of uniformity was computed using the posterior probability of a gene being actively regulated in a tissue, PPA, from the Flutre et al. [33] multi-tissue 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. The blue line is the best fit linear regression and shows a significant positive correlation between h^2 and uniformity ($\beta = 0.12$, $P = 2e - 11$).

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	% FDR<0.1	num FDR<0.1	num expressed
Cross-tissue	450	0.062	20.2	2995	14861
Adipose - Subcutaneous	298	0.017	9.5	1408	14861
Adrenal Gland	126	0.037	8.9	1329	14861
Artery - Aorta	198	0.022	9.8	1449	14861
Artery - Coronary	119	0.048	8.1	1211	14861
Artery - Tibial	285	0.022	8.5	1262	14861
Brain - Anterior cingulate cortex (BA24)	72	0.037	10.9	1627	14861
Brain - Caudate (basal ganglia)	100	0.042	9.3	1388	14861
Brain - Cerebellar Hemisphere	89	0.046	10.9	1627	14861
Brain - Cerebellum	103	0.041	10.9	1618	14861
Brain - Cortex	96	0.053	10.0	1480	14861
Brain - Frontal Cortex (BA9)	92	0.037	10.6	1569	14861
Brain - Hippocampus	81	0.033	10.9	1618	14861
Brain - Hypothalamus	81	0.020	12.4	1839	14861
Brain - Nucleus accumbens (basal ganglia)	93	0.040	10.7	1585	14861
Brain - Putamen (basal ganglia)	82	0.033	10.6	1581	14861
Breast - Mammary Tissue	183	0.020	9.2	1367	14861
Cells - EBV-transformed lymphocytes	115	0.045	8.9	1323	14861
Cells - Transformed fibroblasts	272	0.019	9.0	1334	14861
Colon - Sigmoid	124	0.024	10.6	1578	14861
Colon - Transverse	170	0.022	10.6	1573	14861
Esophagus - Gastroesophageal Junction	127	0.029	9.1	1358	14861
Esophagus - Mucosa	242	0.026	8.2	1220	14861
Esophagus - Muscularis	219	0.024	9.0	1344	14861
Heart - Atrial Appendage	159	0.031	9.7	1437	14861
Heart - Left Ventricle	190	0.025	8.9	1325	14861
Liver	98	0.036	10.2	1519	14861
Lung	279	0.018	8.8	1308	14861
Muscle - Skeletal	361	0.020	9.0	1342	14861
Nerve - Tibial	256	0.026	8.3	1235	14861
Ovary	85	0.043	10.2	1523	14861
Pancreas	150	0.036	9.4	1396	14861
Pituitary	87	0.044	10.0	1481	14861
Skin - Not Sun Exposed (Suprapubic)	196	0.041	6.3	938	14861
Skin - Sun Exposed (Lower leg)	303	0.027	6.9	1020	14861
Small Intestine - Terminal Ileum	77	0.046	11.7	1733	14861
Spleen	89	0.062	9.7	1448	14861
Stomach	171	0.020	10.3	1532	14861
Testis	157	0.042	10.5	1561	14861
Thyroid	279	0.024	8.7	1293	14861
Whole Blood	339	0.025	7.7	1150	14861

All tissues are from the GTEx Project. Mean heritability (h^2) is calculated across genes for each tissue. The percentage (%) and number (num) of genes with significant h^2 estimates (FDR < 0.1) in each tissue are reported.