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

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

1 Department of Biology, 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 Division of Genetic Medicine, Vanderbilt University, Nashville, TN, USA

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

Abstract

For most complex traits, gene regulation is known to play a crucial mechanistic role as demonstrated by the consistent enrichment of expression quantitative trait loci (eQTLs) among trait-associated variants. Thus, understanding of the genetic architecture of gene expression traits is key to elucidate the underlying mechanisms of complex traits. However, systematic survey of the heritability and the distribution of effect sizes across all representative tissues in the human body is not available.

Here we take advantage of the RNAseq data on a comprehensive set of tissue samples generated by the GTEx Consortium to fill this gap. We find that local h^2 can be well characterized with xx% of significant h^2. However, the current sample sizes of (<400?) only allow us to compute distant h^2 for a handful of genes (xx% range). Thus we focus on local regulation. Bayesian Sparse Linear Mixed Model (BSLMM) analysis and the sparsity of optimal performing predictors provided compelling evidence that local architecture of gene expression traits is sparse rather than polygenic across all xx tissues examined.

To further delve into the tissue context specificity, we decompose the expression traits into cross tissue and tissue specific components. Heritability and sparsity estimates of these derived expression phenotypes show similar characteristics to the original traits. Consistent properties relative to prior GTEx multi tissue analysis results suggests that these traits reflect the expected biology.

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

Author Summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam.

Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

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 useful tools to understand the biology underlying complex traits. Partitioning heritability to different functional classes have been successful in quantifying the contribution of different mechanisms that drive the etiology of diseases.

Most human expression quantitative trait loci (eQTL) studies have focused on how local genetic variation affects gene expression in order to reduce the multiple testing burden that would be required for a global analysis [4, 5]. Furthermore, when both local and distal eQTLs are reported [6–8], effect sizes and replicability are much higher for local eQTLs. Indeed, while the heritability of gene expression attributable to local genetic variation has been estimated accurately, large standard errors have prevented accurate estimation of the contribution of distal genetic variation to gene expression variation [8, 9].

While many common diseases have are likely polygenic [10–12], 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 [4].

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

Thus, gene expression traits with sparse architecture should be better predicted with models such as LASSO (Least Absolute Shrinkage and Selection Operator), which prefers solutions with fewer parameters, each of large effect [14]. Conversely, highly polygenic traits should be better predicted with ridge regression or similarly polygenic models that prefer solutions with many parameters, each of small effect [15–17]. To obtain a more thorough understanding of gene expression achitecture, we used the hybrid approaches of the elastic net and BSLMM (Bayesian Sparse Linear Mixed Model) [18] to quantify sparse and polygenic effects.

Most previous human eQTL studies were performed in whole blood or lymphoblastoid cell lines due to ease of access or culturabilty [6, 19, 20]. Although more recently, studies with a few other tissues have been published, a comprehensive coverage of human tissues was not available until the launching of the Genotype-Tissue Expression (GTEx) Project. GTEx aims to examine the genetics of gene expression more comprehensively and has recently published a pilot analysis of eQTL data from 1641 samples across 43 tissues from 175 individuals confirming that eQTLs are highly shared across tissues [21]. Here we use a much larger set of 8555 samples across 53 tissues corresponding to 544 individuals from the 2014-06-13 release. One of the

findings of this comprehensive analysis was that a large portion of the local regulation of expression traits is shared across multiple tissues. This was corroborated by the fact that our prediction model based on whole blood showed good prediction across the 9 core GTEx tissues chosen by initial sample size [13].

This shared regulation implies that there is much to be learned from large sample studies of easily accessible tissues. Yet, a portion of gene regulation seems to be tissue dependent [21]. In order to harness this cross-tissue effect for prediction and to better understand the genetic architecture of tissue-specific and cross-tissue gene regulation, we use a mixed effects model called orthogonal tissue decomposition (OTD) to decouple the cross-tissue and tissue-specific mechanisms in the rich GTEx dataset. We modeled the underlying genetic architecture of the cross-tissue and tissue-specific gene expression components and developed predictors for use in PrediXcan [13].

Results

Local genetic variation can be well characterized for all tissues

We estimated the local and distal heritability of gene expression levels in 40 tissues from the GTEx consortium and whole blood from the Depression Genes and Networks (DGN) cohort. The sample size in GTEx varied from 72 to 361 depending on the tissue while the DGN had 922 samples [20]. We used mixed-effects model (see Methods) and calculated variances using restricted maximum likelihood as implemented in GCTA [22].

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

Table 1 summarizes the unconstrained heritability estimate results across all tissues. In order to obtain an unbiased estimates of mean h^2, we allow the values to be negative when fitting the REML, as done previously [8, 9]. This approach reduces the standard error of the estimated mean of heritability, especially important for the distal component. Even though invididual gene's distal heritability is noisy, the average across all genes reduces the error substantially. For the DGN dataset, we were able to estimate the mean distal h^2 with enough accuracy. However for the GTEx samples, the sample size was too small and the REML algorithm became unstable when allowing for negative values. This numeric instability would cause a small number of genes with large positive (and noisy) heritability values to converge biasing the mean value. For this reason we do not show mean distal heritability estimates for GTEx tissues.

In DGN (whole blood), the mean local h^2 was 14.3% and the mean distal h^2 was 3.4% such that the local variation contribution is estimated as 14.3/(3.4+14.3) = 81%. move to discussion??? This is much higher than the 37% reported by Price et al [9] based on blood expression data for a cohort of Icelandic individuals. This potentially underestimation of the distal component could be due to over-correction of confounders used in the preprocessing of the expression trait data we used. Indeed, PEER [23], SVA [24], and other types of hidden confounder corrections have been shown to increase local eQTL replicability but there are concerns about the detrimental effects on the distal eQTL identification. As larger sample sizes become available we will test this hypothesis in GTEx data by computing the distal h^2 without PEER factor correction.

The left column of Fig. 1 shows the estimated local and distal h^2 from DGN, this time using REML constrained between 0 and 1 (GCTA default) [22]. Even though many genes show relatively large point estimates of distal h^2, only the ones colored in blue are significantly different from zero. The local component of h^2 is relatively well

Table 1. Estimates of local h^2 across whole tissues.

tissue	n	mean h^2	mean se	prop CI > 0	num CI > 0	num expressed
DGN-WholeBlood	922	0.143	0.0292	0.486	6180	12719
Adipose-Subcutaneous	298	0.0385	0.0381	0.0735	1040	14205
AdrenalGland	126	0.0432	0.075	0.0425	601	14150
Artery-Aorta	198	0.0421	0.0561	0.0649	898	13844
Artery-Coronary	119	0.0371	0.0773	0.0337	476	14127
Artery-Tibial	285	0.0417	0.0402	0.0798	1080	13504
Brain-Anteriorcingulatecortex(BA24)	72	0.0275	0.133	0.031	450	14515
Brain-Caudate(basalganglia)	100	0.0367	0.091	0.0343	502	14632
Brain-CerebellarHemisphere	89	0.0492	0.11	0.0509	728	14295
Brain-Cerebellum	103	0.0504	0.094	0.0544	788	14491
Brain-Cortex	96	0.0451	0.0937	0.0393	578	14689
Brain-FrontalCortex(BA9)	92	0.0379	0.101	0.0341	496	14554
Brain-Hippocampus	81	0.0368	0.114	0.0285	414	14513
Brain-Hypothalamus	81	0.017	0.115	0.0235	347	14759
Brain-Nucleusaccumbens(basalganglia)	93	0.0293	0.0965	0.0292	426	14601
Brain-Putamen(basalganglia)	82	0.0324	0.108	0.0291	419	14404
Breast-MammaryTissue	183	0.0289	0.053	0.0365	537	14700
Cells-EBV-transformedlymphocytes	115	0.0578	0.0814	0.0497	619	12454
Cells-Transformedfibroblasts	272	0.0508	0.0424	0.0925	1180	12756
Colon-Sigmoid	124	0.0327	0.0807	0.0389	557	14321
Colon-Transverse	170	0.0358	0.059	0.0436	640	14676
Esophagus-GastroesophagealJunction	127	0.0318	0.0761	0.0332	469	14125
Esophagus-Mucosa	242	0.0416	0.0479	0.0764	1090	14239
Esophagus-Muscularis	219	0.0393	0.0505	0.069	969	14047
Heart-AtrialAppendage	159	0.0418	0.0638	0.0475	660	13892
Heart-LeftVentricle	190	0.0342	0.0558	0.0503	670	13321
Liver	98	0.033	0.0935	0.0299	405	13553
Lung	279	0.0315	0.0401	0.0577	853	14775
Muscle-Skeletal	361	0.0327	0.0324	0.0732	939	12833
Nerve-Tibial	256	0.0523	0.0445	0.095	1380	14510
Ovary	85	0.0369	0.0988	0.0258	364	14094
Pancreas	150	0.0469	0.0674	0.0557	777	13941
Pituitary	87	0.0379	0.107	0.0358	544	15183
Skin-NotSunExposed(Suprapubic)	196	0.0407	0.0532	0.0502	735	14642
Skin-SunExposed(Lowerleg)	303	0.0385	0.0385	0.0765	1120	14625
SmallIntestine-TerminalIleum	77	0.0365	0.112	0.0281	418	14860
Spleen	89	0.059	0.0998	0.0417	603	14449
Stomach	171	0.0315	0.0579	0.0367	533	14531
Testis	157	0.0539	0.0679	0.0727	1230	16936
Thyroid	279	0.0436	0.0413	0.0861	1260	14642
WholeBlood	339	0.0326	0.0328	0.0677	823	12160

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

estimated in DGN with 44.3% of genes (5633 out of 12719) showing h^2 values significantly different from zero. In contrast, the distal heritability is significantly different from zero for only 2.7% (343 out of 12719) of the genes.

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

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

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

However, using the subset of local eQTLs (from an independent source) in other chromosomes for computing distal heritability reduced the mean value from 0.102 to 0.065. Therefore, while we gain some power to detect significant distal heritability by using priors, a good portion of the distal regulation is lost when using only the smaller subset of potentially more functional variants.

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

Sparse local architecture implied by sparsity of best prediction models

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

More specifically, we performed 10-fold cross-validation using the elastic net [24] to test the predictive performance of local SNPs for gene expression across a range of mixing parameters (α). The mixing parameter that yields the largest cross-validation R^2 informs the degree of sparsity of each gene expression trait. That is, at one extreme, if the optimal $\alpha = 0$ (equivalent to ridge regression), the gene expression trait is highly polygenic, whereas if the optimal $\alpha = 1$ (equivalent to LASSO), the trait is highly sparse. We found that for most gene expression traits, the cross-validated R^2 was smaller for $\alpha = 0$ and $\alpha = 0.05$, but nearly identically for $\alpha = 0.5$ through $\alpha = 1$ in the DGN cohort (Fig 4). An $\alpha = 0.05$ was also clearly suboptimal for gene expression prediction in the all? nine GTEx tissues, while models with $\alpha = 0.5, 0.95$, or 1 had similar predictive power (Fig 5-[**SUP**]). This suggests that for most genes, the effect of local genetic variation on gene expression is sparse rather than polygenic.

Direct estimation of sparsity using BSLMM

To further confirm the local sparsity of gene expression traits, we turned to the BSLMM [18] approach, which models the genetic contribution as the sum of a sparse and a polygenic component. The parameter PGE in this model represents the proportion sparse to polygenic component using. Another parameter, the total variance explained (PVE) by additive genetic variants, is a more flexible Bayesian equivalent of the chip heritability we have estimated using a linear mixed model as implemented in GCTA.

As anticipated, we find that for highly heritable genes, the sparse component is large. For example, all genes with PVE > 0.50 had PGE > 0.82 and their median PGE was 0.989 (Fig 6B). The median PGE for genes with PVE > 0.1 was 0.949. Fittingly, for most (96.3%) of the genes with PVE estimates > 0.10, the median number of SNPs included in the model was no more than 10 (Fig 6C).

Also as expected, we find that when the sample size is large enough, such as in DGN, there is a strong correlation between BSLMM-estimated PVE and GCTA-estimated h^2 (Fig 6A, R=0.96). In constrast, when we applied BSLMM to the GTEx data, we found

that many genes had strikingly larger BSLMM-estimated PVE than GCTA-estimated h^2 (Fig 7). This is further confirmation of the local sparse architecture of gene expression traits: the underlying assumption in LMM approaches to estimate heritability is that the genetic effect sizes are normally distributed, i.e. most variants have small effect sizes. LMM is quite robust to departure from this assumption but only when the sample size is rather large. For the relatively small sample sizes in GTEx ($n \leq 361$), we are finding that a model that directly addresses the sparse component such as BSLMM outperforms GCTA for estimating h^2.

Orthogonal decomposition of cross-tissue and tissue specific expression traits

Since a substantial portion of local regulation was shown to be common across multiple tissues [21], we sought to decompose the expression levels into a component that is common across all tissues and a tissue specific components. For this we use a linear mixed effects model as described in the Methods section. We call this approach orthogonal tissue decomposition (OTD) because the cross tissue and tissue specific components are assumed to be independent in the model. The decomposition is applied at the expression trait level so that the downstream genetic regulation analysis is performed separately for each derived trait, cross tissue and tissue specific expression, which greatly reduces computational burden.

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

An clear benefit of OTD for the cross tissue trait is that the effective sample size of the trait is 450 even though each of the tissues had less than 360 individuals. This is reflected in more precise estimates of h^2 as shown below.

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

The cross-tissue heritabilities were larger and the standard errors were smaller than the tissue-specific estimates (Fig 9-[**SUP**]). The percentage of GCTA h^2 estimates with positive CIs was much larger for cross-tissue expression (17.3%) than the tissue-specific expressions (all less than 3%, Fig 10). Similarly, the percentage of BSLMM PVE estimates with a lower credible set greater than 0.01 was 49% for cross-tissue expression, but ranged from 24-27% for tissue-specific expression (Fig 11).

We also compared the cross-tissue h^2 from the OTD to h^2 estimates from the pre-OTD measures of gene expression in each of the nine tissues, which we term whole tissue expression. Again, the cross-tissue heritabilities were larger and the standard errors were smaller than the whole tissue estimates (Fig 12-[**SUP**]), though less striking than the tissue-specific comparison. The percentage of whole tissue h^2 estimates with positive CIs ranged from 4.4-8.6% and thus were all larger than the tissue-specific postive CI percentages, but smaller than the cross-tissue percentage (Fig 2). Cross-tissue BSLMM PVE estimates had lower error than whole tissue PVE (Fig 8, Fig 11). Like whole tissue expression, cross-tissue and tissue-specific expression showed better predictive performance using the elastic-net when $\alpha \geq 0.5$ than when $\alpha = 0.05$ (Fig 13-[**SUP**]). Cross-tissue predictive performance exceeded that of both tissue-specific and whole tissue expression as indicated by higher cross-validated R^2 (Fig 5-[**SUP**], Fig 13-[**SUP**]).

We compared our OTD results to those from a joint multi-tissue eQTL analysis method [25], which was previously performed on a subset of the GTEx data [21]. The results of this analysis include eQTL posterior probabilities for nine tissues, which can

be interpreted as the probability a SNP is an eQTL in tissue x given the data. Using the top eQTL for each gene did you download the top eQTL's prob of being eqtl in the tissue? From what I talked with Matthew Stephens I thought it was the probability at the gene level, of the gene being an egene, i.e. being regulated in the tissue by some eqtl., we defined an entropy statistic (see Methods) that combines the nine posterior probabilities into one value such that higher entropy values mean the gene regulation is more uniform across all nine tissues, rather than in just a subset of the nine. We observed a strong correlation between entropy and both the cross-tissue expression heritability (R = 0.082, Fig 14A) and PVE (R = 0.12, Fig 14B) estimates, using the cross-tissue expression derived from the OTD. Thus, genes with high cross-tissue heritability are more likely to have cross-tissue eQTLs, confirming that OTD is capturing the cross-tissue component of gene expression. Figure 15 shows the correlation between the heritability of tissue specific gene expression traits and the posterior probability of a gene being an e-gene in the tissue is thir correct?. Also, the correlation between tissue-specific OTD gene expression PVE and the posterior probability that the gene has an eQTL in that tissue is strongest in each respective tissue, confirming that OTD also captures tissue-specific components of gene expression (Fig 15).

Nulla mi mi, venenatis sed ipsum varius, Table 2 volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Table 2. Table caption Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam.

Heading1				**Heading2**			
*cell*1*row*1	cell2 row 1	cell3 row 1	cell4 row 1	cell5 row 1	cell6 row 1	cell7 row 1	cell8 row 1
*cell*1*row*2	cell2 row 2	cell3 row 2	cell4 row 2	cell5 row 2	cell6 row 2	cell7 row 2	cell8 row 2
*cell*1*row*3	cell2 row 3	cell3 row 3	cell4 row 3	cell5 row 3	cell6 row 3	cell7 row 3	cell8 row 3

Table notes Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed.

Nulla mi mi, Fig. 2 venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, S1 Video vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Discussion

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

The proportion of local contribution in whole blood from DGN was 80%, which is more than twice values reported previously citeprice et al. This could be due to the

over-correction of distal effects that may occur when applying hidden factor adjustments and thus leading to underestimates of the distal component. These corrections have been shown to increase the number and reproducibility of identified local eQTLs but their consequence on distal regulation is not well understood.

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

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

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

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

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

Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat. Ut neque ipsum, luctus id lacus ut, laoreet scelerisque urna. Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl.

Ut ornare mauris tellus, vel dapibus arcu suscipit sed. Nam condimentum sem eget mollis euismod. Nullam dui urna, gravida venenatis dui et, tincidunt sodales ex. Nunc est dui, sodales sed mauris nec, auctor sagittis leo. Aliquam tincidunt, ex in facilisis elementum, libero lectus luctus est, non vulputate nisl augue at dolor. For more information, see S1 Text.

Materials and Methods

Genomic and Transcriptomic Data

DGN Dataset We obtained whole blood RNA-Seq and genome-wide genotype data for 922 individuals from the Depression Genes and Networks (DGN) cohort [20], all of European ancestry. For our analyses, we used the HCP (hidden covariates with prior) normalized gene-level expression data used for the *trans*-eQTL analysis in Battle et al. [20] and downloaded from the NIMH repository. The 922 individuals were unrelated (all pairwise $\hat{\pi} < 0.05$) and thus all included in downstream analyses. Imputation of approximately 650K input SNPs (minor allele frequency [MAF] > 0.05, Hardy-Weinberg Equilibrium [P > 0.05], non-ambiguous strand [no A/T or C/G SNPs]) was performed on the University of Michigan Imputation-Server (`https://imputationserver.sph.umich.edu/start.html`) [27,28] with the following parameters: 1000G Phase 1 v3 ShapeIt2 (no singletons) reference panel, SHAPEIT phasing, and EUR population. Approximately 1.9M non-ambiguous strand SNPs with MAF > 0.05, imputation $R^2 > 0.8$ and, to reduce computational burden, inclusion in HapMap Phase II were retained for subsequent analyses.

GTEx Dataset We obtained RNA-Seq gene expression levels from 8555 tissue samples (53 unique tissue types) from 544 unique subjects in the GTEx Project [21] data release on 2014-06-13. Of the individuals with gene expression data, genome-wide genotypes (imputed with 1000 Genomes) were available for 450 individuals. While all 8555 tissue samples were used in the OTD model (described below) to generate cross-tissue and tissue-specific components of gene expression, we used the nine tissues with the largest sample sizes when quantifying tissue-specific effects. Tissues and sample sizes (both RNA-seq and genotypes available) included cross-tissue ($n = 450$), skeletal muscle ($n = 361$), whole blood ($n = 339$), skin from the sun-exposed portion of the lower leg ($n = 303$), subcutaneous adipose ($n = 298$), tibial artery ($n = 285$), lung ($n = 279$), thyroid ($n = 279$), tibial nerve ($n = 256$) and left ventricle heart ($n = 190$). Approximately 2.6M non-ambiguous strand SNPs included in HapMap Phase II were retained for subsequent analyses.

Partitioning local and distal heritability of gene expression

Motivated by the observed differences in regulatory effect sizes of variants located in the vicinity of the genes and distal to the gene, we partitioned the proportion of gene expression variance explained by SNPs in the DGN cohort into two components: local (SNPs within 1Mb of the gene) and distal (eQTLs on non-gene chromosomes) as defined by the GENCODE [29] version 12 gene annotation. We calculated the proportion of the variance (narrow-sense heritability) explained by each component using the following mixed-effects model:

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

Assuming a random effects for $w_{k,g} \sim N(0, \sigma_w^2)$ and $\epsilon \sim N(0, \sigma_\epsilon^2 I_n)$, where I_n is the identity matrix, we calculated the total variability explained by local and distal components by estimating σ_w^2 with restricted maximum likelihood (REML) using GCTA software [22]. For heritability analyses in the GTEx cohort, we removed the *distal* term from the model and only estimated marginal *local* h^2 due to the smaller sample sizes of both cross-tissue and tissue-specific expression levels compared to DGN.

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

By default we restricted the heritability estimates to be in the 0 to 1 interval. However, for the purpose of estimating the mean heritability we performed separate runs allowing the heritability estimates to take negative values with the `--reml-no-constrain` option in GCTA. Despite the lack of obvious biological interpretation of a negative heritability, it is an accepted procedure used in order to avoid bias in the estimated mean [8, 9].

Determining polygenicity versus sparsity using the elastic net

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

The elastic net penalty is controlled by mixing parameter α, which spans LASSO ($\alpha = 1$, the default) [14] at one extreme and ridge regression ($\alpha = 0$) [15] at the other. The ridge penalty shrinks the coefficients of correlated SNPs towards each other, while the LASSO tends to pick one of the correlated SNPs and discard the others. Thus, an optimal prediction R^2 for $\alpha = 0$ means the gene expression trait is highly polygenic, while an optimal prediction R^2 for $\alpha = 1$ means the trait is highly sparse. An optimal prediction R^2 in between (e.g. $\alpha = 0.5$) means the trait has a mixed genetic architecture.

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

Quantifying sparsity with Bayesian Sparse Linear Mixed Models (BSLMM)

We used BSLMM [18] to model the effect of local genetic variation (SNPs within 1 Mb of gene) on the genetic architecture of gene expression. The BSLMM is a linear model with a polygenic component (small effects) and a sparse component (large effects) enforced by sparsity inducing priors on the regression coefficients [18]. We used the software GEMMA [32] to implement BSLMM for each gene using the following parameters:

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

The `-bslmm 1` option specifies a linear BSLMM and the `-s 100000` option specifies the number of sampling steps per gene. The BSLMM estimates the PVE (the total proportion of variance in phenotype explained by the sparse effects and random effects terms together) and PGE (the proportion of genetic variance explained by the sparse

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

Orthogonal tissue decomposition

To better understand the context specificity of gene expression regulation, we developed a method called orthogonal tissue decomposition (OTD). This approach is an extension of our method to develop an intrinsic growth phenotype [33]. We applied OTD to GTEx Project [21] data and decomposed the expression of each gene into cross-tissue and tissue-specific components. The tissue availability is unbalanced across individuals because of the difficulties of sample collection and the uneven quality of the tissues. OTD decomposes the expression traits into orthogonal components as represented by the following model:

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

Specifically, to generate cross-tissue and tissue-specific expression levels, we used the `lmer` function in the R [34] package `lme4` [35,36] to fit the following mixed-effects model:

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

The model included whole tissue gene expression levels in 8555 GTEx tissue samples from 544 unique subjects. A total of 17,647 Protein-coding genes (defined by GENCODE [29] version 18) with a mean gene expression level across tissues greater than 0.1 RPKM (reads per kilobase of transcript per million reads mapped) and RPKM ¿ 0 in at least 3 individuals were included in the model. `SUBJID` was a random effect and the covariates `TISSUE`, `GENDER`, and `PEERs` were fixed effects used to predict whole tissue expression levels (`expression` in the model). `PEERs` included the top 15 PEER factors estimated across all tissues using the R package `PEER` [37] to control for batch effects and experimental confounders. Cross-tissue expression was defined as the random effects from the model (`ranef(fit)`) and tissue-specific expression as the residuals (`resid(fit)`).

Comparison of OTD PVE to multi-tissue eQTL results

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

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

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

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

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. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics. 2015;47(9):1091–1098. Available from: http://dx.doi.org/10.1038/ng.3367.

14. 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.

15. 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.

16. 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.

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

18. 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.

19. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature. 2005;437(7063):1365–1369. Available from: http://dx.doi.org/10.1038/nature04244.

20. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research. 2013;24(1):14–24. Available from: http://dx.doi.org/10.1101/gr.155192.113.

21. Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348(6235):648–660. Available from: http://dx.doi.org/10.1126/science.1262110.

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

23. 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.

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

25. 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.

Supporting Information

S1 Video

Bold the first sentence. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Text

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Fig

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S2 Fig

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

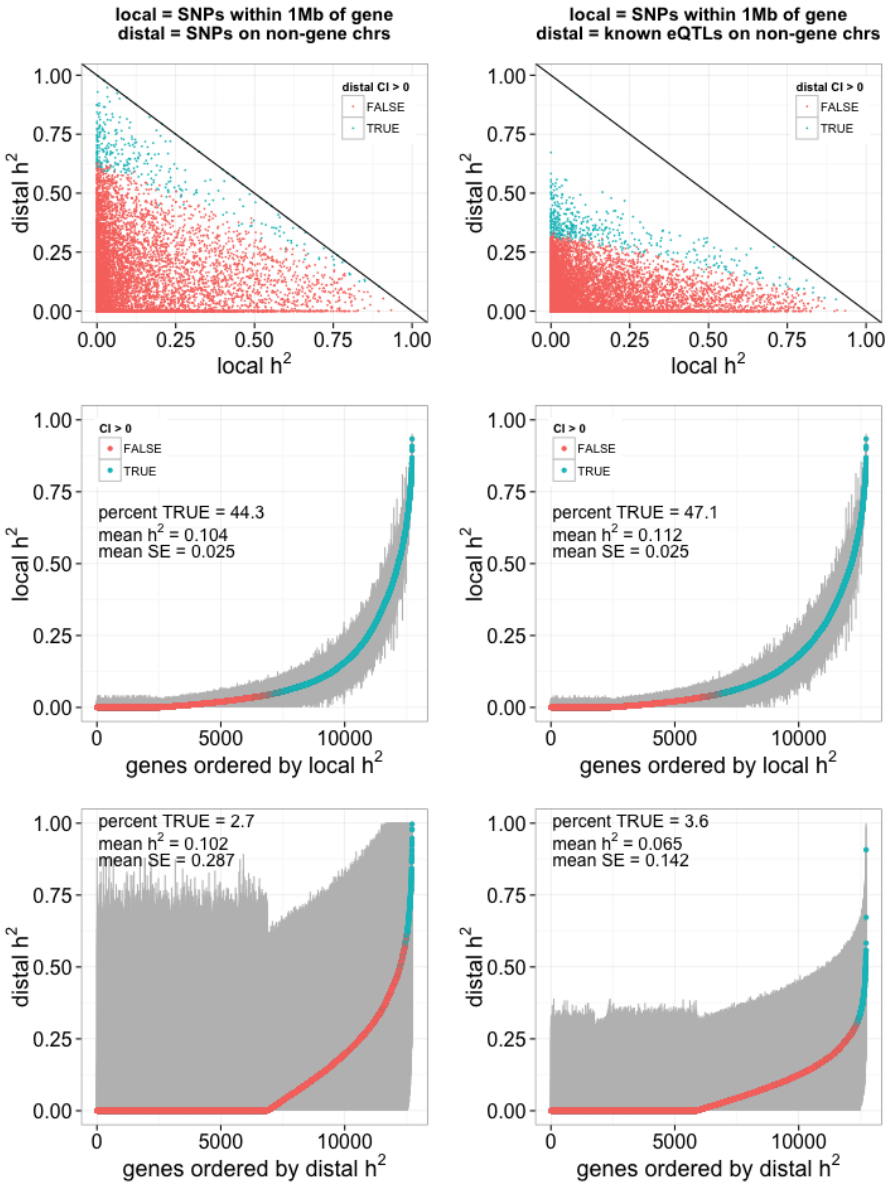

Figure 1. DGN whole blood expression joint heritability (h^2). Local (SNPs within 1 Mb of each gene) and distal (Left: SNPs on non-gene chromsomes. Right: SNPs that are eQTLs in the Framingham Heart Study on other chromosomes [FDR < 0.05]) h^2 for gene expression were jointly estimated. (**Top**) Distal h^2 compared to local h^2 per gene in each model. (**Middle**) Local and (**Bottom**) distal gene expression h^2 estimates ordered by increasing h^2. The 95% confidence interval (CI) of each h^2 estimate is in gray and genes with a lower bound greater than zero are in blue.

Figure 2. Figure Title first bold sentence Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Figure Caption Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. A: Lorem ipsum dolor sit amet. B: Consectetur adipiscing elit.