Decision on your article submitted to PLOS Genetics (PGENETICS-D-16-00755)

Dear Dr Im,

Thank you very much for submitting your Research Article entitled 'Survey of the Heritability and Sparse Architecture of Gene Expression Traits Across Human Tissues' to PLOS Genetics. Your manuscript was fully evaluated at the editorial level and by independent peer reviewers. Each reviewer appreciated the attention to an important problem, but all raised substantial concerns about the current manuscript. Based on the reviews, we will not be able to accept this version of the manuscript, but we would be willing to review again a much-revised version. We cannot, of course, promise publication at that time.

Should you decide to revise the manuscript for further consideration here, your revisions should address the specific points made by each reviewer. We will also require a detailed list of your responses to the review comments and a description of the changes you have made in the manuscript.

If you decide to revise the manuscript for further consideration at PLOS Genetics, please aim to resubmit within the next 60 days, unless it will take extra time to address the concerns of the reviewers, in which case we would appreciate an expected resubmission date by email to plosgenetics@plos.org.

If present, accompanying reviewer attachments are included with this email; please notify the journal office if any appear to be missing. They will also be available for download from the link below. You can use this link to log into the system when you are ready to submit a revised version having first consulted our Submission Checklist http://journals.plos.org/plosgenetics/s/submit-now#loc-submission-checklist.

PLOS offers a figure-checking tool, PACE (http://pace.apexcovantage.com/), to help authors to ensure all figures meet PLOS requirements so that the quality of published figures is as high as possible. Please use this tool to help you format your figures. PACE is a digital diagnostic and conversion tool for figure files. It will provide information about any failed check(s) and, if able, will automatically convert the figure file into an acceptable file that passes quality checks. PACE requires you to register for an account to ensure your figure files are processed securely.

Please note, PLOS has incorporated Similarity Check (http://www.crossref.org/crosscheck.html), powered by iThenticate, into its journal-wide submission system in order to screen submitted content for originality before publication. Each PLOS journal undertakes screening on a proportion of submitted articles. You will be contacted if needed following the screening process.

You should also be aware that as part of our data availability policy we do not permit the inclusion of phrases such as "data not shown" or "unpublished results" in manuscripts. All points should be backed up by data provided with the submission.

To resubmit, please use the link below and 'Revise Submission' in the 'Submissions Needing Revision' folder. \*\*\*\*\*\*\*\*

We are sorry that we cannot be more positive about your manuscript at this stage. Please do not hesitate to contact us if you have any concerns or questions.

http://www.editorialmanager.com/pgenetics/viewLetter.asp?id=891776&lsid={B7EAA7BE-E347-4BF1-9C83-562039F5EE20} Page 1 of 7

View Letter 5/26/16, 11:02 AM

Yours sincerely,

Stephen B Montgomery Guest Editor

PLOS Genetics

Gregory Barsh Editor-in-Chief PLOS Genetics

Reviewer's Responses to Questions

Comments to the Authors:

Please note here if the review is uploaded as an attachment.

**Reviewer #1:**

Wheeler et al. perform a systematic survey of genetic heritability in multiple human tissues and the DGN whole blood cohort. They concluded that local architecture of gene expression traits is sparse rather than polygenic, among other findings. The scientific problems addressed in this paper are interesting and important. The analyses described are well designed and performed by the state-of-the-art analytic tools. That said, I have some concerns in their interpretation of their statistical analysis, especially when drawing biological and statistical conclusions beyond the data sets analyzed.

I hope the following comments helpful for the authors to revise the manuscript:

1. The first subsection in the Results section concludes that local genetic variation can be \*well\* characterized for all tissues. It seems that the referred "well characterization" is only \*relative\* to the trans estimates in the paper, which I feel a bit unsatisfying. Claiming the local heritability is well characterized suggests the estimates are reliable, but how do they compare with the estimates from other existing studies? In particular, the local heritability estimates for DGN whole blood and GTEx whole blood have very large discrepancy (14.3% vs 3.3%, Table 1) and their 95% CIs are clearly separated. The authors should address this issue more carefully and explicitly. The authors also claims that h^2 in DGN is relatively well estimated, because more genes showing h^2 \ne 0 (line 86 page 6). Without knowing the biological truth I find this type of the reasoning problematic (there are other similar claims in the paper).

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 the are similar, we have added the following paragraph to the first subsection of the results (lines 83-89):

“Gene expression local heritability estimates were consistent between whole blood tissue from the DGN and GTEx cohorts (Spearman's rho = 0.28 [95% CI: 0.27-0.29]). DGN estimates made here were also consistent with those made in independent blood cohorts from previous studies. Comparing DGN to Price et al. and Wright et al., rho = 0.31 [0.30-0.32] and 0.083 [0.070-0.095], respectively. Spearman's rho between Price et al. and Wright et al. gene expression heritability estimates was 0.12 [0.11-0.14].”

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.”

explain that h2 is dependent on number of PEER factors, overall level of noise, etc. So the values are not directly comparable. However, we expect the ranks to remain the same but not the absolute value.

We expand the discussion of this point in the discussion section to make it more clear.

Regarding the biological truth, we will compare Price and Wright estimates. Although they are not the "true" values, they should be good proxies of them. Unless the method has some systematic bias, which we don't believe is the case.

2. I find the conclusion "sparse local architecture implied by sparsity of best prediction models" is a bit weak. I have slightly different interpretation of the cross-validation results for the \alpha parameter in the elastic net. To me, the panel A of the Figure 2 indicates 1) the model completely excluding strong sparse effects seems inferior 2) the data cannot completely rule out the polygenic component in a predictive model. i.e., many curves are flat from \alpha \in (0.05,1), which indicates the predictive power does not significantly improve by focusing only on sparse component. I do understand the authors' point is to suggest approaches like PrediXscan can effectively use sparse model for predication purpose (with the current sample size) and I agree. However, I think extend it to the biological conclusion of general genetic architecture is a bit of stretch.

We thank the reviewer for raising a potential point of confusion. While our BSLMM results clearly indicate the sparse component dominates the total variance explained by genetic variants (Figure 3A), the elastic net results in Figure 2A 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. The flatness of the curve means the relative contribution of any polygenic component is minimal. To clarify what is presented in 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.”

In the next subsection, we build upon our elastic net results with BSLMM. We describe how BSLMM directly estimates the sparse component PGE, which was near 1 for most heritable genes (Fig. 3A). The title of this subsection has been changed from “Direct estimation of sparsity using BSLMM also points to sparse local architecture” to “Sparse local architecture revealed by direct estimation using BSLMM.”

The BSLMM method directly estimates the sparse component and it clearly indicates that the sparse component is dominant. For lower h2 genes, we cannot say much but we believe that this indicates lack of power to determine that. The flatness of the mixing parameter means the polygenic component is not adding much to the power to predict. Consistent with BSLMM results.

the polygenic component may be there but the relative contribution is minimal.

3. I like the authors' use of BSLMM to dissect PGE and PVE, but I think the reasoning to the conclusion that "BSLMM outperforms LMM in estimating h^2 for small samples" seems problematic. The evidence seems based on that "we found that many genes had measurably larger BSLMM-estimated PVE than GCTA-estimated h^2". This is again concluded without ground truth. But more importantly, it should be noted as a Bayesian approach BSLMM limits its h^2 estimate in (0,1), while GCTA allows negative estimates for unbiasedness. This may explain some of the over-estimates by BSLMM. I would also expect by different criteria, one may find completely different conclusion (e.g., GCTA is better to obtain the unbiased estimate).

We note that the GCTA results presented originally in Figures 3 and 4 used a REML model constrained between 0 and 1, just like the BSLMM results (see lines 374-375 of the original Methods section). This has been clarified in the new manuscript (lines 160-162). We have also updated Figures 3B and 4 such that the points are colored according to the elastic net (alpha=1) cross-validated R2 value for each gene. To Figure 4, we added correlations between the elastic net and BSLMM results and between the elastic net and GCTA results. The genes with high BSLMM PVE and near zero GCTA 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. Comparing elastic net R2 to the heritability estimates, the Pearson correlations with BSLMM are higher than those with GCTA for all tissues, consistent with a sparse genetic architecture of gene expression. 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 h\textsuperscript{2} for small samples” subsection starting at line 149 and the Figure 3-4 legends have been updated to reflect these changes.

The constrained GCTA is also underestimating many genes h2. (Fig 3 and 4 are using constrained REML h2--I verified min(h2)=1e-06)

Another evidence that GCTA is underestimating h2 can be seen in the plot R2 vs GCTA h2. When cross validated R2 is large, this is strong evidence that h2 should no be zero but we are finding many such genes (0 h2 and large R2)

4. I don't quite follow the OTD approach. The explanation in the Material and Methods section is over- simplified and may contain critical typos in notations. In the first linear mixed model, are the variance parameters \sigma\_{CT} and \sigma\_{TS} both constant across tissues and samples? If so, are they identifiable? 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? Does this imply all eQTLs share the same effects across tissues? 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.

THINK ABOUT HOW TO ADDRESS THIS

Other minor comments:

page 5, line 76: "averaging across all genes reduces the error substantially" -> "reduces the error \*for mean estimate\* substantially"

We have made the requested wording change at line 76.

page 5, line 78: "REML algorithm became unstable when allowing for negative values" why this is only a problem for trans estimate? Would it cause problem for cis estimate in GTEx or IGN?

As can be seen in Fig. 1, only a few of the local h2 estimates are less than 0, while thousands of the distal h2 estimates are. 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 clarify, we have added the <10% to the following statement (lines 79-81):

“This numeric instability would cause only a small number of genes (<10%) with large positive (and noisy) heritability values to converge biasing the mean value.”

the problem only arises when the likelihood is flat (not enough information, i.e. sample size) but not when there is enough information such that the likelihood is very low for negative h2 or h2 greater than 1. We'll explain this better.

page 6, line 98: it is better to provide the complete 95% CIs here

This sentence has been changed, see line 103:

While using known eQTLs to define distal h2 decreased the 95% confidence interval mean across genes from [-0.45-0.50] to [-0.26-0.29], the number of significant genes did not change dramatically (Fig. 1).

page 9, line 123: "nearly identically" -> "nearly identical"

We have fixed the typo, this sentence is now at line 130.

**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.

It's not the higher heritability but the higher predictive power of sparse models. We also estimate a direct parameter that represents sparsity using BSLMM. We clarify this in the text...

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.

To clarify the second sentence in the Methods (line 308), we have changed “For our analyses,” to “For all analyses, we used the HCP (hidden covariates with prior) normalized gene-level expression data

used for the trans-eQTL analysis in Battle et al. and downloaded from the NIMH repository.”

We believe there is value in acknowledging the lack of power to address distal h2. By addressing other comments, including items 12 and 16-18 from Reviewer 2, we have clarified sources of confusion in this section.

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 results (lines 115-120):

“To determine if the local estimates were affected by linkage disequilibrium (LD), we also estimated the local heritability of gene expression in the DGN cohort using LDAK, a software for computing LD-adjusted kinships \cite{speed2012improved}. While the mean estimates were slightly higher using GCTA (0.143) compared to LDAK (0.118), the Pearson correlation between h2 estimates from GCTA and LDAK was high (R=0.96), indicating that the local component of heritability using either method is about the same (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.

I don't understand this point. Haky’s R2 vs pLI?

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 potentials areas of confusion have been addressed in our updates to the paper and our 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 \cite{Gusev\_2014,torres2014cross,davis2013partitioning}.”

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. 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.

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 26. In the second sentence of the results (line 64), we inserted the “(REML)” to clarify the definition:

“We used mixed-effects models (see Methods) and calculated variances using restricted maximum likelihood (REML) as implemented in GCTA \cite{Yang\_2011}.”

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 starting at line 45 to say:

“Corroborating this finding, our prediction model built in DGN whole blood showed robust prediction \cite{Gamazon\_2015} across the nine tissues with the largest sample size from the GTEx Pilot Project \cite{Ardlie\_2015}.”

We also added “(n > 70)” to this statement in the Methods (line 328): “we used the 40 tissues

with the largest sample sizes (n > 70) when quantifying tissue-specific effects (see Table 1).”

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: “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 83-89):

“Gene expression local heritability estimates were consistent between whole blood tissue from the DGN and GTEx cohorts (Spearman's rho = 0.28 [95\% CI: 0.27-0.29]). DGN estimates made here were also consistent with those made in independent blood cohorts from previous studies. Comparing DGN to Price et al. and Wright et al., rho = 0.31 [0.30-0.32] and 0.083 [0.070-0.095], respectively. Spearman's rho between Price et al. and Wright et al. gene expression heritability estimates was 0.12 [0.11-0.14].”

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. 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. This is something we need to keep in mind. Cite Fisher when he says that "...". This is a rather vague concept that we agree to use. Cell type heterogeneity effect will be to some degree captured by the PEER factors.

12.\* In Figure 1, do the order of the genes change from the left column to right one in the second row when consider different distal settings? What is the Wilcox rank test across tissues?

We have added the following sentence to lines 110-112:

“The two different methods to define distal h2 did not affect the local h2 estimates (Spearman’s rho = 0.86), demonstrating the robustness of local h2 component in the linear mixed model.”

13.\* The paragraph starting "Table 1 summarizes..." is hard to follow. What does "unconstrained" mean? Averaging what across genes? what is the mean distal h2 referring to, and what does the distribution look like? What does "unstable" mean?

We have rewritten this paragraph to explicitly answer these questions (lines 69-82):

“Table 1 summarizes the local heritability estimate results across all tissues. In order to obtain an unbiased estimates of mean h2 across genes, we do not constrain the model to only output h2 estimates between 0 and 1. Instead, as done previously \cite{Price\_2011,Wright\_2014}, 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, especially important for the distal component. Even though each individual gene's distal heritability is noisy (Fig. 1), averaging across all genes reduces the error for the mean estimate substantially. For the DGN dataset, we were able to estimate the mean distal h2, which was 0.034 (SE = 0.0024). However for the GTEx samples, the sample size was too small and the REML algorithm became unstable (did not converge) when allowing for negative values. This numeric instability would cause only a small number of genes (<10%) 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.”

14.\* While the idea of sparsity is central do the paper, an analysis of how sparsity relates to heritability is lacking. Is there a relationship between the effect sizes or sparsity patterns of genes significantly heritable within a tissue, and their patterns in the tissues where they are not significantly heritable?

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. 3A). If a gene is significantly heritable, we show the sparse component is large (i.e. near 1)--see Fig. 3A, Supplemental Figures S5, S6.

15.\* What does (P < 0.05) mean? Does this mean it is an uncorrected p-value? Why was it not Bonferroni corrected, or some other approach (e.g., q-values)? What is the FDR of that threshold? Could the null and alternative hypotheses be explicitly spelled out?

Yes, this is the uncorrected p-value, which is now explicitly stated on line 92. We want to compare tissues with each other and Bonferroni correction is too conservative to do so. We have added the FDR at P < 0.05 for each tissue to Table 1.

16.\* "from an independent source"?

17.\* what is the difference between "cis" and "local"?

18.\* "Functional priors" to describe "known eQTLs" is a strange way to describe that.

To clarify items 16-18, we have edited the paragraph describing our distal h2 analysis, now starting at line 98:

“It has been shown that local-eQTLs are more likely to be distal-eQTLs of target genes \cite{pierce2014mediation}. Thus, we tested whether restricting the distal h2 estimates to known eQTLs on non-gene chromosomes could improve distal h2 precision by prioritizing functional variants. We defined the "known eQTLs" by using the Framingham mRNA cohort of over 5000 individuals \cite{Zhang\_2015} (see Methods), which is independent from the DGN and GTEx cohorts. While using known eQTLs to define distal h2 decreased the mean standard error of the heritability estimates across genes from 0.24 to 0.14, the number of significant genes did not change dramatically (Fig. 1). Also, using the subset of known eQTLs reduced the mean distal h2 from 0.027 to 0.015. Therefore, while we gain some power to detect significant distal h2 by using eQTL priors as indicated by the standard error reduction, a good portion of the distal regulation is lost when using only the smaller subset of known eQTLs. When we used known eQTLs to estimate distal h2 in the GTEx cohort, less than 1% of genes had a P < 0.05 (Fig S1). Given the limited sample size we will focus on local regulation for the remainder of the paper.”

We removed any reference to “cis” and instead use “local” throughout the paper, since in humans, we usually don’t test phase (see Albert and Kruglyak 2015). We use “known eQTLs” rather than “functional priors” throughout.

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

Adaptive shrinkage, see <http://biorxiv.org/content/early/2016/06/08/038216.abstract> and <http://arxiv.org/abs/1605.07787> Say we will explore in future studies?

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 130:

“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 404-405:

“In order to compare prediction R$^2$ 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 the 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 component). As stated in lines 148-151:

“For example, all genes with PVE > 0.50 had PGE > 0.82 and their median PGE was 0.989 (Fig. 3A). The median PGE for genes with PVE > 0.1 was 0.949. Fittingly, for most (96.3%) of the genes with PVE estimates > 0.10, the number of SNPs included in the model was no more than 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 line . 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.

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

We have edited this sentence at lines 162-165:

“...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 Methods at line 413:

“BSLMM assumes the genotypic effects come from a mixture of two normal distributions and thus is flexible to both polygenic and sparse genetic architectures.”

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?

Those are great ideas. Our proof-of-concept results in this manuscript are encouraging, therefore, we will develop the method more thoroughly when additional brain data (larger subtissue samples sizes) are released by the GTEx Project.

27.\* Assumptions of OTD need to be spelled out clearly.

We added a summary of the assumptions in OTD, which are described in the Methods section, to the xxx section.

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 (line 19):

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

29.\* It would be helpful to provide background on what the PGA parameter is and on why one should expect statistically for PGE to be large

when PVE is large. It is also unclear whether the data in Figure 3 correspond to a particular tissue from GTEx, to DGN, or to a

cumulative set.

PGE is defined in the Methods lines 416-419:

“The 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. 3 to read:

“Sparsity estimates using Bayesian Sparse Linear Mixed Models in DGN whole blood.”

30.\* How do you justify the entropy metric you use? It seem that by normalizing the vector of posterior probabilities, you are removing the essential "total magnitude" parameter that will differentiate distributions that look non-uniform (but are really low magnitude and cannot be differentiated from uniform) from uniform.

We have added a plot with a few examples that shows that entropy is a good measure of uniformity. TODO: explain plot here

31.\* It is mentioned that when the sample size is large enough, there is a strong correlation between PVA and $h^2$. However, there is no figure showing how across the tissues in GTEx this relationship

fades for the tissues with less samples.

We have added a supplemental figure (S4 Fig), which shows a strong correlation between sample size and the PVE:h2 correlation.

32.\* In figure 4 it is not clear why the overestimates given by BSLMM are more reliable than the estimates of LMM given that the true heritability values are not known.

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 GCTA results. The genes with high BSLMM PVE and near zero GCTA 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 GCTA for all tissues, consistent with a sparse genetic architecture of gene expression. The “BSLMM outperforms LMM in estimating h\textsuperscript{2} for small samples” subsection starting at line 149 and the Figure 4 legend have been updated to reflect these changes.

33.\* More care in the write-up of the equation would be helpful to the reader. Do we assume Y\_g to be n dimensional vectors? In the orthogonal tissue decomposition section, is Y\_{i,t} in fact Y\_{i,t,g}? What is the reduced model? What is the cross - tissue dependence in the equation Y\_{i,t} = Y\_i^{CT} + Z\_{i} \beta + \epsilon\_{i,t}. What is parameter W, mentioned in the text but not in the equations? All of the dimensions must be stated clearly, and the distributions of each of the random effects.

clarify dimension of vectors. Verify equations are right.

34.\* More detail regarding the relationship between the posterior probability of active regulation (PPA) and heritability could be provided starting from a primer on how PPA is computed, and why such a computation is justified.

We have added a description of PPA to the Methods at lines 467-473:

“PPA can be interpreted as the probability a SNP is an eQTL in tissue *t* given the data. PPA is computed from a joint analysis of all tissues and takes account of sharing of eQTLs among tissues \cite{Flutre\_2013}. 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 \cite{Flutre\_2013}.”

35.\* Figure 5 is misleading. put the y-axis on log scale and x-axis on a continuous scale? What is the p-value here representing?

We added the following to the Fig. 5 legend to describe the p-value: “The Kruskal-Wallis rank sum test revealed a significant difference in the cross-tissue h$^2$ of uniformity groups ($\chi^2 = 31.4$, $P = 1.5e-07$).”

We also added Supplementary Fig. S10 to show the continuous scale relationship.

36.\* Going beyond heritability with predictability in mind, are the more heritable genes good markers for tissue or cell type?

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.

37.\* quite a number of typos in Methods.

TODO: check this

38.\* why did you use only chr 22 instead of doing, e.g., 10 fold cross validation?

As shown in Fig. 2, 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 chr22. We perform 10-fold cross validation at 3 mixing parameters (α = 0.05, 0.5, 1) for all genes.

**Reviewer #4:**

The manuscript by Wheeler et al presents analyses of multiple tissue gene expression data sets primarily from the GTEx consortium. The manuscript investigates the genetic architecture of gene expression through the lens of heritability (variance explained) and prediction and claims a sparsity in the genetic architecture of gene expression. Whereas using heritability/prediction is clearly useful if the goal is expression prediction (see Gamazon et al Nat Genet 2015, Gusev et al Nat Genet 2016, Wang et al AJHG 2016), it is unclear what conclusions can be drawn about the distribution of effect sizes at the hidden unknown causal variants (i.e. the true underlying genetic architecture). For example, since the authors focus only on common variants (MAF>5%), risk prediction (or variance explained) from multiple common variants could be due to tagging of a single hidden causal variant; conversely, a single common variant explaining a big proportion of the variance in expression could be a tag for one or multiple untyped variants.

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.

Comments:

1. The authors should clearly define the heritability they estimate through GCTA as that has implications on the interpretability of their results. I believe the authors estimate the proportion of variance explained by typed variants which (in general) does not include the contribution of rare variation and is a lower bound on the narrow-sense heritability. This could also explain the difference between results reported here and Price et al PlosGen 2011 (see Discussion) which used local IBD to include typed and untyped variants in the estimation.

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 69-70: “For the local heritability component, we used common (minor allele frequency > 0.05) variants within 1Mb of the transcription start and end…”

We edited the sentence describing the Price et al results at lines 266-268:

“This is much higher than the 37% reported by Price et al. [26] based on blood expression data from a cohort of Icelandic individuals using identity by descent, which captures both common and any rare variation.”

We added the following caveat to the Discussion at lines 281-282:

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

2. The manuscript proposes a potentially interesting approach to decompose expression into cross-tissue and tissue-specific components. However, they do not provide any simulations to understand how the empirical results should be interpreted. Second, what is the relationship between this approach and a standard bi- variate REML analysis (e.g. as implemented in GCTA)? Does it have a similar interpretability but extended to 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?

LET'S THINK ABOUT WHAT SIMULATIONS WE CAN RUN TO MAKE THESE REVIEWERS HAPPY.

Wang et al and this paper have different purposes. Wang et al is used to impute unobserved gene expression levels for eQTL analysis or other types.

COMPARE TO BIVARIATE REML? <http://cnsgenomics.com/software/gcta/reml_bivar.html>

Run --reml-bivar CT TS and compare to --reml TW ???

See if V(G)/Vp\_CT + V(G)/Vp\_TS correlates with V(G)/Vp ???

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 284-286:

“Sparse models capture the most variance in gene expression at current sample sizes and have been used successfully in gene expression prediction methods \cite{manor2013robust,Gamazon\_2015,gusev2016integrative,zhu2016integration}.

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 34-36:

“Elastic net \cite{Zou\_2005} 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 \cite{abraham2013performance}.”

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 the goals of PrediXcan (i.e., integration of expression and GWAS), lots of approaches could be informed by these findings (e.g. Giambartolomei et al PlosGen 2014, He et al AJHG 2013, and many more).

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 300-304:

“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.”

6. The authors use use 2 components to split variance into local and distal. Since the local component is likely to be impacted by LD, it would be interesting to quantify the impact of LD on the local component using methods such as LDAK.

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 115-120):

“To determine if the local estimates were affected by linkage disequilibrium (LD), we also estimated the local heritability of gene expression in the DGN cohort using LDAK, a software for computing LD-adjusted kinships \cite{speed2012improved}. While the mean estimates were slightly higher using GCTA (0.143) compared to LDAK (0.118), the Pearson correlation between h2 estimates from GCTA and LDAK was high (R=0.96), indicating that the local component of heritability using either method is about the same (S2 Fig).”

7. How were the standard errors computed in the average heritabilities reported in Table 1? Is there an assumption that every gene is independent of each other (i.e. SE=sd/sqrt(n))? If so, the SE may be biased downwards as many genes will have correlated expression/genetics.

We calculated the correlation between genes and used a chi2 test to estimate significance of the mean h2 estimate.

8. The authors perform an interesting analysis aiming to check whether known-QTL genes contribute more to the distal heritability of a given expression but do not find an enrichment. I wonder whether this is due to LD tagging in GCTA (i.e. the more known-QTLs are included the more of the genome will be tagged).

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.

On a similar note, it is hard to see the mean heritabilities reported in Figure 1.

Fig. 1 is designed to present an overview of each gene’s h2. The mean heritabilities are reported in the text and Table 1.

9. In general, the manuscript uses very sophisticated methodology (i.e. variance components, prediction through elastic net) which although powerful comes with assumptions. I suspect that a simple iterative conditioning strategy (i.e. take top QTL variant, condition it from expression, and iterate) will yield rather similar results in figuring out the number of predictors needed.

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)”.