Committee Meeting June 2016: Estimating Effect Sizes Applied to GTeX Data

Sarah Urbut

June 24, 2016

Contents

Introduction

Variation in gene expression is an important mechanism underlying susceptibility to complex disease. The simultaneous genome-wide assay of gene expression and genetic variation allows the mapping of the genetic factors that underpin individual differences in quantitative levels of expression (expression QTLs; eQTLs). The availability of this information provides immediate insight into a biological basis for disease associations identified through genome-wide association (GWA) studies, and can help to identify networks of genes involved in disease pathogenesis ([1,2]). Available methods are limited not only in their ability to *jointly analyze data on all tissues* to maximize power, but also in simultaneously *allowing for both qualitative and quantitative differences among eQTLs* present in each tissue.

Initial approaches to quantify the effect of a particular SNP on gene expression considered only one tissue at a time, and ignored the effect of the SNP on gene expression in other tissues. This failed to exploit the power of shared genetic variation in effects on expression - i.e. the information that the effect of the gene-snp pair in one tissue can provide about the effect in

another- and limited our understanding of multiple-tissue phenotypes. Furthermore, even past attempts at identifying eQTLs using the data across tissues jointly were limited in both the number of tissues considered, and also the level of heterogeneity considered. These joint attempts referred to the setting in which the gene-snp pair is active in all tissues as 'shared' and active in only one as 'tissue-specific' ([3, 4]). This method did not report an effect size, and centered on assigning the eQTL to a 'binary configuration' in which the eQTL was 'off' in some tissues and 'on' in others. However, a QTL may be 'active' in all or many tissues and with varying magnitude or sign; we refer to this as quantitative heterogeneity. Thus we sought to develop a method which could quantify the effect size across tissues in a multivariate manner, as well as acknowledge systematic 'patterns' of sharing of effect sizes and directions across tissue types.

Indeed, our initial motivation came from our group's past analysis of GTEx pilot data [5], in which we saw evidence that many (50%) QTLs are shared across all nine tissues. In this context, a QTL was called based on whether it demonstrates significant posterior probability of being active in a particular tissue. Applying our previous hierarchical model ('eQTL-BMA') to the dataset from Dimas *et al* [6] with 3 tissues, we found just 8% of eQTLs were specific to a single tissue, with an estimated 88% of eQTLs being common to fibroblasts, LCL cells and T-Cells [3]. Not all eQTLs are shared by all tissues; some tissues may share eQTLs more than others. To allow for this, our previous hierarchical model attempted to infer the extent of such sharing by estimating the proportion of eQTLs which were shared in various 'configurations' or patterns of binary activity in which a QTL was 'called' or 'absent' in each tissue. Such an approach did not quantify effect sizes across subgroups, and thus could not capture the quantitative variation present among tissues in which the SNP was considered active. Furthermore, even if such binary configurations were used to estimate effect sizes, the approach was not flexible enough to capture subtler patterns of systematic heterogeneity. Indeed, though most eQTL are active in all tissues, certain eQTL have consistently larger effects in some tissues than others, while other eQTL show strong effects in opposing tissues. In fact, as the number of tissues considered increases, perhaps the more interesting and biologically relevant question becomes one of quantitative heterogeneity - that is, how do the patterns of effect sizes and directions vary across tissues in which the SNP is called 'active'.

The novel approach we offer here allows groups of QTLs to be classified not only by their presence or absence in a particular tissue, but by their *relationships in continuous effects between tissues*, e.g., consistently larger effects in some tissues than other. Critically, quantifying the effect sizes of the gene-snp pair across tissues considering the evidence contained in all tissues jointly thus reveals new patterns of activity across tissues, which differ in their relationship in sign and magnitude within and between tissues. The novel framework described here allows that these effects be 'shared' but not necessarily 'consistent' across tissues, thus capturing continuous variation heretofore missed. The structure of this paper is as follows: we will describe our approach in brief for modeling and estimating these effect sizes across tissues, demonstrate the utility of such an approach on simulated data, and apply our method to data from the GTEX dataset, version 6.0, where we analyze effects size for 16,069 genes across 44 tissues. In so doing, we offer novel insight into the patterns of quantitative heterogeneity in genetic effects among a greater number of tissues than ever analyzed before.

In this project, we propose to address these issues as follows:

1 Aim 1: Method Development: Develop and apply statistical methods for efficiently mapping expression QTLs in large numbers of diverse cell-types and tissues from RNA-seq data.

We develop statistical methods capable of jointly mapping eQTLs across large numbers of conditions. While previous methods have been limited to tissue-by-tissue analyses or considering

joint effects across only a small number of conditions, our method will be tractable when applied to a much larger number (greater than 40) conditions. Critically, we hope to account for both the qualitative and quantitative heterogeneity among conditions, adding a novel feature beyond simply reporting reporting the probability of being 'on' or 'off'. This will capture similarities and differences among active QTLs across a variety of conditions.

1.1 Overview

We assume we have a matrix of estimates of many 'effects' in many conditions. For example, in the GTEx data considered here this matrix consists of estimated effects of many (potential) eQTLs in many tissues. Our goal is combine information across effects and conditions to produce improved estimates of the effects (and corresponding measures of significance) in each condition. In doing this we allow that some effects may be 'shared', being similar (though not identical) among conditions, while others may be 'specific' to only a subset of conditions. Our approach is sufficiently flexible to apply to many contexts where estimates can be naturally organized as a matrix.

Let b_{jr} $(j = 1, \ldots, J; r = 1, \ldots, R)$ denote the true value of effect j in condition r. Further let $\hat{b}_{jr}$ denote the (observed) estimate of this effect, and $\hat{s}_{jr}$ the standard error of this estimate (so $\hat{b}_{jr}/\hat{s}_{jr}$ is the usual z statistic for testing whether b_{jr} is zero). We use $\boldsymbol{b}_j$ to denote the vector of effects $(b_{j1}, \ldots, b_{jR})$, and $\hat{\boldsymbol{b}}_j$ for the corresponding vector of estimates.

Our method uses an Empirical Bayes (EB) approach. That is, we assume that the effects $\boldsymbol{b}_j$ come from some common ("prior") distribution g,

$$\boldsymbol{b}_j \sim g(\cdot), \tag{1}$$

and that the estimates $\hat{b}$ are normally distributed about the true effects:

$$\hat{b}_{jr} | b_{jr}, \hat{s}_{jr} \sim N(b_{jr}, \hat{s}_{jr}). \tag{2}$$

The EB approach first estimates g from the data, combining information across *all* effects, to yield an estimate $\hat{g}$ say. It then uses Bayes theorem to compute the posterior distribution for b_{jr} given the prior $\hat{g}$ and the data $(\hat{b}_{jr}, \hat{s}_{jr})$.

A key point is that the distribution g captures the similarity, or otherwise, of effects among conditions. Thus, in estimating g from the data, the EB approach adapts to the specific patterns of effects in the data at hand. If effects tend to be similar among conditions then the estimated $\hat{g}$ will reflect this, and the resulting posterior distributions will be "shrunk" (or "nudged") towards a common value, producing estimates that tend to be similar among conditions. Conversely, if effects tend to be specific to one or a few conditions then the estimated $\hat{g}$ and resulting estimates will reflect this.

To flexibly model the effect distribution g we use a mixture of multivariate normal (MVN) distributions:

$$g(\cdot; \boldsymbol{\pi}, \boldsymbol{U}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{k,l} \; N_R(\cdot; \mathbf{0}, \omega_l U_k). \tag{3}$$

Here each covariance matrix U_k corresponds to a different "class" of effect, which allow for different patterns of sharing across subgroups of effects, whereas each scaling coefficient ω_l corresponds to a different "size" of effect. The mixture proportions $\pi_{k,l}$ determine the relative frequency of each size-class combination. We allow the scaling coefficients ω_l to vary on a fixed dense grid of values spanning "very small" to "very large", to capture the full range of effects that could reasonably occur (as in [7]). Estimating g then involves estimating the parameters π, U, as described in Detailed Methods (not included here). Note that every component in 3 is centered on $\mathbf{0}$, and so the overall distribution of effects will be centered on $\mathbf{0}$. This will be appropriate in many applications, specifically those where effects can be both positive or negative and many will be close to 0.

The use of a mixture of MVNs for g effectively generalizes several previous approaches to this problem. For example, each "configuration" in eQTL-BMA [3] corresponds to a multivariate normal distribution. A key difference with previous work is that we estimate the matrices U_k from the data. In contrast previous work either used a large number of fixed U_k 'configurations' (e.g. eQTL-BMA [3]) making them computationally prohibitive for moderate R; or a small number of fixed U_k (e.g. eQTL-BMAlite) making them less flexible. By instead estimating U_k from the data our approach is both flexible, and computationally tractable for moderately large R (here, $R = 44$).

In our application to the GTEX data $R = 44$, so each U_k is a 44 by 44 covariance matrix, and each component of the mixture (3) is a distribution in 44 dimensions. Visualizing such a distribution is challenging, but we can get some insight from the first eigenvector of U_k, v_k say, which captures the principal direction of the effects in component k. If U_k is dominated by this principal direction (which is the case for many k in our application) then we can think of effects from that component as being of the form λv_k for some scalar λ. For example, if the elements of the vector v_k are approximately equal then component k captures effects that are approximately equal in all conditions. Or, if v_k has one large element, with other elements close to 0, then component k corresponds to an effect that is strong in only one condition. Though not evident in the gene expression case but possible in the more general case, these effects may be inversely correlated among subgroups, and our model also allows for this by estimating these patterns from the data. See Figure 1 for illustration.

1.2 Updates

- In the past year, I have added a deconvolution framework to the estimation of these covariance matrices. To retrieve a 'denoised' or 'deconvoluted' estimate of the non-single rank dimensional reduction matrices, we then perform deconvolution after initializing the EM algorithm with the non-single rank matrices (2), (3) and (5) specified in 2.2. The final result of this iterative procedure preserves the rank of the initialization matrix, and allows us to use the 'true' effect at each component component $\boldsymbol{b_j}$ as missing data in deconvoluting the prior covariance matrices. In brief, this algorithm works by treating not only the component identity but also the true effect $\boldsymbol{b_j}$ as unobserved data, and maximizing the likelihood over the expectation of the complete data likelihood, considering the values $\boldsymbol{b_j}$ as extra missing data (in addition to the indicator variables of component identity) [8]. We found that this improved the likelihood of the model on cross validation (see Aim 2).

- I also added a component with an effect at $\mathbf{0}$ to understand the estimation of the null effect. On simulated data, I find that putting a likelihood penalty of approximately $2R$ allows for accurate estimation of π_0, but further illustrating the difficulty in estimating the mixture components and only further validating the need for methods which estimate effects rather than focus on assigning items to a particular category.

- I have added extensive simulations to demonstrate the ability of MASH to accurately and powerfully estimate effects shared effects, and improve on existing methods ([3]) when structure among shared effects exist. Furthermore, `mash` is superior to univariate shrinkage method `ash` on joint data in both its ability to estimate 'true' effects and shrink noise.

1.3 Simulations

We conducted simulations under three very different scenarios:

1. "Shared, structured effects": data were simulated using the model (3), based on the fit of this model to the GTEX data (see Methods for details). In this scenario effects tend to be shared among many conditions, and furthermore these shared effects are highly

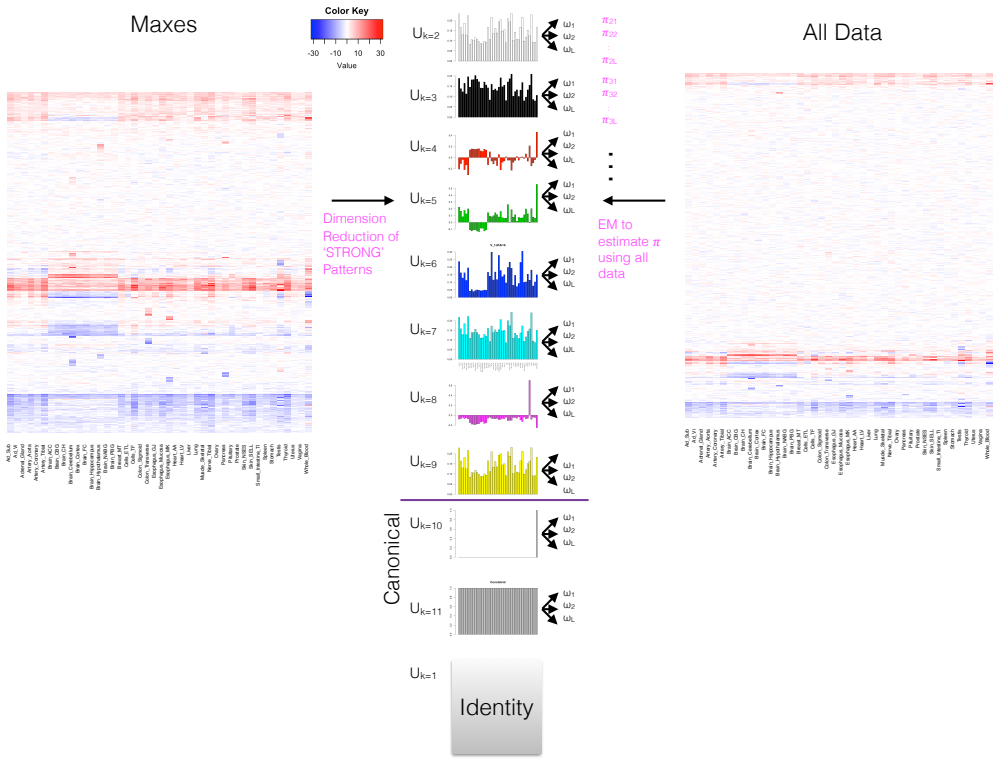

Figure 1. Overview of Modeling Approach. A: In our pipeline, we estimate patterns of sharing from the strongest Z statistics across 44 tissues, using a variety of rank approximations. We hope to capture the 'true' patterns in doing so. We then learn the range of effect sizes (ω) and estimate the proportional representation of each pattern from the larger data set, using the EM algorithm in typical EB fashion. This allows us to succinctly capture the pattern of sharing of effects across tissues.

"structured", in that they are often similar in size and direction. A key feature of `mash` is that it attempts to learn this structure. If this learning is successful then `mash` should show improved performance over simpler multivariate methods.

2. "Shared, unstructured effects": in this scenario effects are shared among all conditions (i.e. either every condition shows an effect, or no condition shows an effect), but the effect sizes and directions are independent across conditions. This represents a relatively "unstructured" setting, where the attempts of `mash` to learn structure should have no advantage. However, we hope to demonstrate that `mash` performs no worse in this setting than simpler approaches.

3. "Independent effects": in this scenario effects are entirely independent across conditions, with no greater sharing then expected by chance. In this setting univariate methods, which analyze each condition separately, are expected to perform best.

In each case we simulate a matrix of data containing 20,000 estimated effect sizes in each of 44 conditions (and their associated standard errors). In the first two scenarios 400 of the 20,000 effects are non-null. In the last scenario 400 effects are non-null in each condition.

We analyzed each scenario using three methods

1. `mash`, the method we describe here.

2. A simpler multivariate method, `bmalite`, which is similar to the BMAlite method from [3] extended to output effect size estimates. `bmalite` allows for condition-specific effects (i.e. effects that occur in only one condition) and shared effects (i.e. effects that occur in all conditions), and allows for correlation of effects among conditions. However, the modeling assumptions underlying `bmalite` are less flexible than `mash`. For example, `bmalite` assumes all pairs of conditions are equally correlated in their effects, whereas `mash` can learn that some pairs are more correlated than others.

3. `ash` ([7]), which can be thought of as a univariate version of `mash`. Results from `ash` are obtained by applying it separately to each condition, and so represent what can be achieved a typical simple "condition-by-condition" analysis.

Each method produces a matrix of estimated effect sizes, and a matrix of the "local false sign rate" (*lfsr*) for each effect, which is an assessment of the probability that the estimated effect size has the incorrect sign. The *lfsr* is a measure of significance, analogous to the local false discovery rate [9], but more stringent in that it insists that effects be correctly signed to be considered "true discoveries."

We compared methods both in the accuracy of their estimates and power to detect true effects.

To compare accuracy of estimated effect sizes we use the relative root mean squared error (RRMSE) (4).

$$\text{RRMSE} = \frac{\sqrt{E((b_{jr} - E(b_{jr}|\text{Data}))^2)}}{\sqrt{E((b_{jr} - \hat{b}_{jr})^2)}}. \tag{4}$$

The RRMSE is the root mean squared error of the estimates, divided by the root mean squared error achieved by simply using the original observed estimates for the effects. Thus an RRMSE < 1 indicates that the method produces estimates that are more accurate than the original observations . Such improvements in accuracy come from two sources: i) the methods can shrink estimated effects towards zero. This improves average accuracy because most effects are indeed null; ii) the methods can share information across conditions to improve accuracy. For example, if a particular effect is shared, and similar in size, across a subset of conditions then averaging the observed effects in those conditions will improve estimation accuracy. The

Table 1. Power and Accuracy Comparison

Method	Simulation Framework	RRMSE^{ALL}	RRMSE^{TRUE}	RRMSE^{Noise}
`mash`	GTeX	0.06	0.44	0.015
`ash`	GTeX	0.21	1.34	0.076
`bmalite`	GTeX	0.11	0.78	0.018
`mash`	ωI	0.14	0.997	0.014
`ash`	ωI	0.21	1.37	0.078
`bmalite`	ωI	0.15	1.03	0.014
`mash`	Independent	0.28	1.82	0.112
`ash`	Independent	0.21	1.37	0.076
`bmalite`	Independent	0.28	1.82	0.118

Comparison of RRMSE on true vs noise effects. Here we compare the RMSE on the true ($\beta \neq 0$ vs noisy estimates).

challenge that the methods face is deciding how much to shrink, and how and when to share information.

To compare power to detect effects we use the "Sign Power", which is an analogue of power, but requires effects to be not only significant, but also correctly signed, to be considered "detected". We compare (1) the ability of Matrix Ash ('`mash`') to capture the true effect-size estimates across all 20,000 simulated pairs. We compare with univariate-shrinkage method '`ash`' with our effect size addition to the existing joint approach for analyzing joint data, `bmalite`.

The results (Table 1) demonstrate the benefits of `mash` for estimating and detecting shared, structured effects. Indeed, in the "shared, structured" scenario `mash` substantially outperforms `bmalite` in both accuracy and power (e.g. halving the RRMSE) which in turn substantially outperforms the univariate method `ash`. In the case of shared, unstructured effects both multivariate approaches perform similarly, and outperform the univariate method. This illustrates the fact that, although `mash` is designed to identify structure in the effects when they exist, it is capable of capturing unstructured patterns of sharing. Finally, in the independent effects scenario, the univariate approach performs best as expected. However, the differences between methods are smaller here than in the other scenarios, illustrating the relative robustness of the multivariate methods.

We might also want to parse out the ability of `mash` to estimate the 'real associations' and to simultaneously shrink noise. We can see that it is best at estimating real associations when sharing is present, no worse than the existing joint approach and superior to univariate methods in data that is shared but independent, and equal to existing joint methods on univariate data. Also, `mash` is superior in shrinking noise to univariate approaches to noise shrinkage on data that is simulated jointly (GTeX and Shared but Independent) and no worse than existing approaches (`bmalite`) which allow for a less flexible grid selection process.

1.4 Consistency Index

I also introduce the idea of a consistency index:

There is often considerable interest in assessing the extent to which effects are shared among conditions. For example, here we apply `mash` to estimate the effects of many eQTLs in many tissues, and want to assess whether eQTLs are shared broadly among tissues, or specific to one or a small number of tissues. Previous work in this context (e.g. [3, 5]) has focussed on assessing the number of tissues in which an eQTL is "active". However, an eQTL could be active in all tissues while varying greatly in the *strength* of activity among tissues. To capture this idea we introduce the notions of "normalized effect" and "consistency index".

We define the normalized effect $\tilde{b}$ in each condition as the ratio of its effect in that condition

to the largest effect across all conditions:

$$\tilde{b}_{jr} = \frac{b_{jr}}{b_{jr_0}} \tag{5}$$

where

$$r_0 = \arg\max_r |b_{jr}| \tag{6}$$

For example, in our eQTL context, a normalized effect $\tilde{b}_{jr} = 0.5$ means that the effect of eQTL j in tissue r is half that of its effect in the strongest tissue.

We define the consistency index for effect j, CI_j^α, as the number of conditions with normalized effect size at least α:

$$\mathrm{CI}_j^\alpha = \#\{r : \tilde{b}_{jr} > \alpha\}. \tag{7}$$

Here α can be chosen to demand a more or less stringent definition of consistency; we focus on $\alpha = 0.5$ here.

Both the normalized effect and the consistency index are only meaningful for units (eQTLs) j that have a substantive effect in at least one condition, and so when reporting these quantities we typically focus on units j that are "significant" in at least one condition ($lfsr_{jt} < 0.05$ for some t), recognizing that this will not be an unbiased representation of all units.

Table 2. Consistency Index Accuracy: Raw (Unstandardized)

Method	RMSE_{HI}
MASH	5.20
ASH	10.26
eQTL-BMAlite	7.62

Consistency Index Accuracy For the data simulated according to the GTEX patterns of sharing, we compute the RMSE of the consistency index for each method (unstandardized).

Together, these results demonstrate the tremendous power increase of using a multivariate method and the accuracy of estimating patterns of sharing from the data rather than imposing forced configurations which fail to capture the Consistency of effect-sizes among tissues.

2 Aim 2: Compare Among Statistical methods

While in simulated data, we can compare among methods in terms of accuracy and power (see Sec. 1.3), we introduce a novel 'testing-training' framework to find the model of best fit when varying the number and nature of components.

2.1 Testing and Training

In order to determine the optimal number and rank of the covariance matrices, we divide our data set into a training and test data set, each containing 8000 genes.

In the training set, we proceed as described in 1: choosing the top SNP for each of the 8000 genes, creating a list of covariance matrices through deconvolution and grid selection of these top 'training gene-snp' pairs.

Then, within the training data, we similarly choose a random set of gene-snp pairs (restricting our analysis to genes contained in the training set). Specifically, we choose 20,000 random-gene snp pairs and use the EM algorithm to learn the mixture proportions $\hat{\boldsymbol{\pi}}$ from this data set as in (8).

The likelihood is given by

$$\begin{aligned} L(\pi; \hat{B}, V) &= \prod_{j=1}^{J} p(\hat{\boldsymbol{b}}_j | \pi, V) \\ &= \prod_{j=1}^{J} \sum_{p}^{P} \pi_p N_R(\hat{\boldsymbol{b}}_j; \mathbf{0}, \Sigma_p + V_j) \end{aligned} \tag{8}$$

We then use the KxL vector of π from the training set to estimate the log likelihood of each data point in the test data set. If our model is biased (or 'overfit') to the training data set, then a larger number of covariance matrices may actually decrease the test log-likelihood. We compare with eQTL-BMA lite as an existing joint approach, as well as to a variety of U_k where the number and rank of each of the U_k varied. I also compared with univariate Ash to show that our method is more powerful.

In this particular data set, we found a combination of 9 'learned' matrices representing the Identity matrix, the empirical covariance matrix of the strong Z statistics, the rank 3 SVD approximation, 5 single rank approximations from Sparse Factor Analysis ([10]) and a rank 5 SFA approximation. Thus in the framework below, Q=5 and P= 3.

2.2 Heuristics for generating candidate covariance matrices U_k

We begin by identifying the rows of the matrix, $\hat{\boldsymbol{b}}_j$ that likely have strong effects in at least one condition. Specifically we compute for each row the maximum Z score across conditions,

$$Z_j^{\max} := \max_r \hat{\boldsymbol{b}}_{jr} / \hat{s}_{jr}, \tag{9}$$

and order the rows by the value of $Z_j^{\max}$.

Let $\tilde{J}$ denote the number of rows selected in this way, and let $\tilde{Z}$ denote the column-centered $\tilde{J} \times R$ matrix of Z scores for these "strong effects".

- $U_1 = \mathbf{I}_R$. This represents the situation where the effects in different conditions are independent, which may be unlikely in some applications (like the GTEx application here), but seems a useful case to include if only to exclude it .
- $U_2 = \frac{1}{\tilde{J}} \tilde{Z}' \tilde{Z}$, the empirical covariance matrix of the strong Z scores.
- $U_3 = \frac{1}{J}\ V_{1\ldots p}\ d^2_{1\ldots p}\ V^t_{1..p}$ is the rank p eigenvector approximation of the tissue covariance matrices, i.e., the sum of the first p eigenvector approximations, where ${}_{1\ldots p}$ represent the eigenvectors of the covariance matrix of tissues and ${}_{1\ldots p}$ are the first p eigenvalues.
- $U_{4:4+Q-1} = \frac{1}{J}((\Lambda\mathbf{F})^t \Lambda\mathbf{F})_q$ corresponding to the q_{th} sparse factor representation of the tissue covariance matrix
- $U_{k=4+Q} = \frac{1}{J}\ (\Lambda\mathbf{F})^t \Lambda\mathbf{F}$ is the sparse factor representation of the tissue covariance matrix, estimated using all q factors.
- $U_{k=5+Q:R+4+Q} = \frac{1}{J}\ ([100..]'[100...])$
- $U_{k=R+5+Q} = \frac{1}{J}\ ([111...]'[111...])$

We divide each of the U_k by the maximum element along the diagonal, so that the maximum element along the diagonal is now one and the matrices are equivalent in 'scale' but not direction.

3 Aim 3: Apply this novel approach to a data-sensitive modeling of the posterior effect-size estimate to the GTEx data set.

We provide the user with a comprehensive list of eQTLs across all conditions, intractable with earlier approaches, as well as effect size estimates for each variant. More broadly, our method applies to any assessment of genetic effect across multiple conditions, providing for an interpretation of the continuous relationship among conditions.

3.1 GTEX analysis

We applied `mash` to a matrix of summary statistics $\hat{\beta}$, together with their corresponding standard errors $\hat{s_j}$, estimating the effect of a SNP on gene-expression. These genes were chosen subject to the constraint that a univariate summary statistic could be obtained across all 44 tissues, and left us with 16,069 genes. We estimate the covariance matrices from the Z statistics of strongest pair for each gene, as estimated by the maximum absolute Z statistic across tissues. This allows us to capture the underlying 'true patterns' of sharing in the data. We then add the qualitatively specific canonical configurations '`bmalite`' configurations. Finally, we inferred the relative frequency of each pattern of sharing and corresponding scales from a larger sample of 40,000 randomly chosen gene-snp pairs derived from the same genes. Here, we report the analysis on the top SNP for each of 16,069 genes, where the 'top' snp is defined as the SNP with the largest observed univariate Z-statistic in absolute value across tissues. We applied the same procedure the univariate shrinkage procedure `ash`, fitting the model to a larger set of 40,000 gene snp pairs and then reporting posterior means and lfsr on the 'top' SNP per gene, thus using hierarchical shrinkage reflective of the overall scale of the data set and not just the maxes.

`mash` produces a striking increase in power (Table 3) when compared to the tissue-by-tissue analysis using `ash`. There are a total of 44 tissues x 16,069 gene-snp pair associations considered, or 707,036 total tissue-level effect-size coefficients. At $lfsr<0.05$, we identify 393,414 (55%) snp-gene-tissue effects can be confidently signed. Using `ash` only 91,755 (13%) are confidently signed.

This tremendous increase in power arises from the fact that in a data set with a great deal of 'sharing' of information as learned by the hierarchical model, small effects in on subgroup that exist in a gene containing large effects in alternative tissues will be augmented to reflect such consistency (see Figure 2), thus increasing our confidence in its size and direction. Indeed, examining the proportion of QTL that are shared between each pair of tissues at a given significance threshold, we see that many (65%) pairs of tissues share greater than 60% of the significant QTL contained in both tissues (see Figure 2 at left). There are several tissues - Whole Blood, Testes, and Transformed Cell Types - which exhibit tissue-specific behavior (see Section 3.5 for more details) but these patterns are also captured by our flexible model.

Furthermore, our method identifies more associations than existing joint approaches (BMA-lite). We have established the accuracy of our approach on simulations 1 and in cross-validation on real data (`mash` yields a log likelihood improvement of 15,215, see 'Testing and Training' procedure, for details).Critically, '`bmalite`' would put the vast majority of the prior weight on the fully shared and 'consistent' configuration.SNPs demonstrating activity across all tissues, are forced into this configuration without the ability to capture systematic 'sharing' between tissues.

3.2 Sharing: From Significance to Effect Size

We see that there is a tremendous amount of sharing of significance across tissues (see Figure 2), resulting in the power increase of the joint method cited in Table 3 . This has been noted in previous analyses ([3], [5]) and is reinforced by a model which is able to incorporate additional tissues and estimate the subtle patterns of sharing across tissues. Here, we define pairwise

Table 3. Number of SNP-gene pairs called significant by each method (at lfsr<0.05).

	b_{jr} called Significant
Method	
`mash`	391,104
`ash`	91,755
`bmalite`	314,771

Power. We quantify the number of associations we identify at a local false sign rate (LFSR) of 0.05. `mash` calls over four times as many associations significant when compared to univariate approach, and yields greater power when compared to a joint approach (bma-lite) with a lower likelihood (15,215 log units worse).

sharing of significance between tissues 'i' and 'j' as the Jaccard coefficient (10) and visualize in the heatmap in 2 at top left.

$$P_{shared} = \frac{QTL_i \cap QTL_j}{QTL_i \cup QTL_j} \tag{10}$$

We call associations 'significant' if they satisfy a given LFSR (Eqn. 12) threshold, indicating that we are confident of the sign of their effect. This is a novelty of our approach, as previous methods have simply asked if the effect is non-zero (see section 'LFSR vs LFDR'). When we then consider the genes sorted by similarity patterns of significance as measured by LFSR across tissues (Figure 2, top right), we see a variety of patterns of sharing of significance of effects, motivating a deeper exploration of effect sizes across tissues.

While a QTL can be significant across all tissues considered, there may be variation in the sign and magnitude across subgroups. In previous analyses we have considered the distribution of number of tissues in which a QTL is considered 'significant'. Simply counting the number of tissues in which a gene-snp is significant reveals a bimodal distribution of significance (Fig 2, Top Center), suggesting that many effects are significant in all tissues. However, the analogous plot considering the number of tissues in which the effects are 'similar' (Fig 2, bottom Center) reveals a different story - namely, that many effects are significant in all tissues yet not consistent in magnitude. Thus armed with a vector of effect-size estimates across 44 tissues, $\boldsymbol{b_j}$, we can move beyond asking in how many tissues is a given gene-snp pair significant, and ask about the relationship in effect-size and direction among tissues in which the gene-snp pair is active.

Table 4. Heterogeneity Analysis.

Data	All Tissues	No Brains	Brain Only
Consistent in Sign: $E(b_{jrnorm}\|D) > 0$	0.85	0.86(0.88)	0.96(0.98)
Consistent Magnitude: $\mathrm{E}(\mathrm{b}_{jrnorm}\|D) > 0.5$	0.37	0.42 (0.44)	0.78 (0.85)
Consistent Significance: LFSR ≤ 0.05	0.63	0.65(0.67)	0.84(0.93)

Heterogeneity Analysis. Top: After normalizing each gene-snp-effect-size coefficient by the effect-size with maximal value at that gene (b_{jrnorm}), we ask how many of these gene-snp effect coefficients are positive. **Center**: To evaluate consistency in magnitude, we can ask how many gene-snp-tissue effects are greater than 50% of the maximal effect across tissues for the pair, given that the gene contains a QTL in at least one tissue. **Bottom**: To evaluate consistency in significance, we can ask how many gene-snp-tissue effects are significant (at LFSR of 0.05), given that the gene contains a QTL in at least one tissue. For robustness, we also performed this analysis separately on each sub-selection of tissues and show these results in parenthesis. In all analyses, we restrict our consideration to gene-snp pairs that are significant in at least one tissue.

.

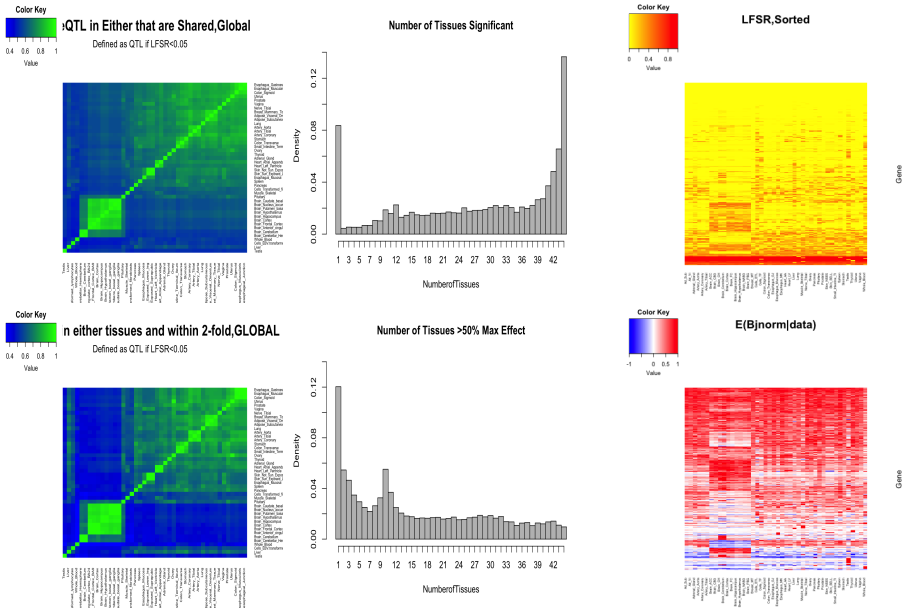

Figure 2. Sharing. We contrast the conclusions drawn by sharing of significance (**top**) with those drawn by sharing of effects (**bottom**). **Top Left**: For each pair of tissues, we consider the proportion of significant QTL contained by either tissues that are significant in both. **Top Center**: For each gene, we count the number of tissues in which a gene is significant. **Top Right**: We plot a heatmap of genes as sorted by their pattern of significance. We contrast this with the conclusions drawn by considering sharing of effects. Armed with new information about effect-sizes across tissues, we can ask additional questions about heterogeneity. **Bottom Left**: For each pair of tissues, we consider the proportion of significant QTL contained by either tissue that are within 2-fold of one another. **Bottom Center**: For each gene, we considering the number of tissues in which the effect is at least 50% of the maximal effect for the gene. **Bottom Right**: We plot a heatmap of genes as sorted by their normalized effect across tissues.

3.2.1 Quantifying Heterogeneity

As the results above illustrate, there may be sharing of significance in the presence considerable heterogeneity in terms of magnitude. Effects may be 'shared' in the sense of significance, but not necessarily consistent in size. In fact, we find consistency in significance is more common than consistency in magnitude (Table 4). To understand this sharing at the tissue level, in addition to our previous plot of pairwise sharing of significant effects (Figure 2, top left), we can now consider the proportion of genes which have effects that are 'close' (within 2-fold) to each other between a pair of tissues, again restricting our analysis to those effects which have demonstrated significance in at least one tissue (Figure 2, bottom left). In the heatmap in the bottom left, we consider pairwise sharing of effects between tissues 'i' and 'j' as 11:

$$P_{sharedeffects} = \frac{0.5b_i \leq b_j \leq 2b_i}{QTL_i \cup QTL_j} \tag{11}$$

We recognize familiar, organ level patterns of sharing of effect-sizes, as we can see that brain-tissue effects are often similar to one another, as are vascular effects (tibial and coronary arteries), and gut tissues (esophagus and colon, as well as the terminal ilium of the small intestine and colon). However, considering the sharing determined by fold-change in effect size clarifies this separation. Effects that are significant in either tissue tend to be larger, and thus closer in magnitude to one another. This picture eliminates 'modestly significant-effects' that are small but can still differ by orders of magnitude. This emphasizes the utility of a joint approach which can make use of subtler patterns of effect-size sharing among tissues, but also illustrates the problem of thresholding.

We can now quantify the consistency index (7) in magnitude and ask, for each gene, in how many tissues is the effect greater than or equal to a significant fraction, here 50% of the maximum effect (Fig 2, lower center). Genes binned in the left of this distribution are quantitatively specific because the effect is close to the maximal effect in few tissues, while homogenous genes will be featured towards the right of the distribution (maximal value at 44) as the majority of their effects across tissues are similar in magnitude.

Critically, while we can see that while there exists a bimodal distribution of the number of tissues in which a gene is significant, most genes have 'similar' effects in only a few tissues, as evidenced by the peak at one. Excluding brain from the analysis tends to nudge us towards a belief in consistency (Table 4).

The conclusions drawn about 'sharing' are different when we ask in how many tissues is the gene considered significant and turn instead to considering consistency in effect size. For instance, while 63% of effects are significant given that the QTL is significant in at least one tissue, only 37% of effects are greater than 50% of the maximum effect for the gene (4).

From a biological standpoint, we might predict that effects of a different sign are rare. Considering these results with and without the inclusion of the brain tissues, which appear to behave as a strongly correlated group, we observe several phenomenon. The majority of gene-snp pairs are consistent in sign (indeed, only about 20% of genes show two significant effects of a different sign when including brain, and even fewer (14.8%) when excluding brains, see Table 4. Removing brains from our analysis tends to push the tendency towards consistency, suggesting that brain appears to behave as a large tissue-specific entity. After normalizing each gene-snp effect-size coefficient b_{jr} by the effect-size with the maximum value for the gene, we can also ask what proportion of these are positive. We again recognize homogeneity with 83% (all tissues) and 86% (excluding brain) demonstrating positive normalized effects, respectively. Taken together, these results suggest the presence of consistency in sign in our data set, and a bimodal distribution of heterogeneity in magnitude, with certain groups of tissues exhibiting sharing of effect-sizes more frequently.

Of note, the proportion of effects that are significant given that the gene within which they exist contains an eQTL is higher in the separate analyses only because the total number of significant genes is fewer in this analysis (see 5).

3.2.2 Pulling out a subgroup of tissues

To contrast effects which are both consistently significant and homogenous in effect size with those that are consistently significant but have heterogenous effects, we compare the brain tissues with the remaining tissues.

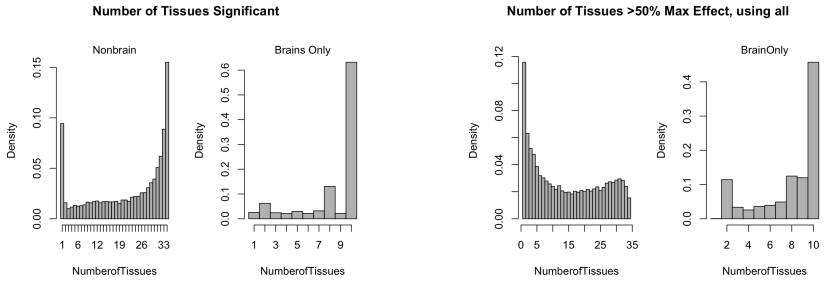

Figure 3. Non-Brain and Brain Alone. Here, we consider the distribution of the number of tissues in which the eQTL is significant top) and similar in magnitude to the max effect in non-brain and brain-only tissues. We see that brain tissues tend to behave 'consistently' in both significance and magnitude, as the majority of eQTL are significant in sign and magnitude in all tissues.

We emphasize two points: this subgroup of tissues behaves as one tissue, in that the most significant effects have an effect of similar magnitude, unlike the all tissue analysis in which effects can be 'significant' in many tissues and yet heterogeneous in magnitude. Secondly, a joint approach improves the total number of associations and genes detected as containing a QTL (5). We note that an analysis which considers more tissues ('Global') and then retrieves the subgroup of interest is more powerful than restricting the analysis to the subgroup alone 5 the results restricted to a subset of tissues versus performing the analyses separately results in different number of identified associations in both per gene-snp-tissue and per gene.

Table 5. Power for a Subset of Tissues

Number of Associations Significant		
	Tissue Type	
Analysis Type	**Non-brain**	**Brain Only**
Global	306,820	86,594
Subgroup	303,809	82,120
Global	13,945	10,353
Subgroup	13,277	8,817

Power for a Subset of Tissues. As described in Flutre et al, we recognize that including more tissue types in our analysis is a more powerful analysis. We compare he number of significant effects in a subset of tissues retrieved from a global analysis with those achieved by performing the analyses on the respective subgroup of data. **Bottom Rows:** We can also ask for the number of genes showing a QTL in at least one of the subsetted tissues, again using both subgroup and global methods.

3.3 LFSR vs LFDR

Indeed, a novelty of our approach is its ability to consider how considering an effect size estimate as opposed to simply calling associations on or off helps us to understand the subtle patterns of

heterogeneity among SNPS that are called active in a particular tissue. In Figure 2, note how we would classify genes differently based on their patterns of 'calling' by LFDR vs their patterns of effect sizes - e.g., effects can be 'significant' in all tissues such that $p(b_{jr} = 0|Data)$ is small, but the actually size, as reflected our confidence in our ability to demonstrate their sign (12) can vary (4).

$$P(b_{jr}) = 1 - max[\sum_p p(b_{j,r} > 0|\hat{\boldsymbol{b}}_j, \hat{V}_j, z_j = p)\tilde{\pi}_p, \sum_p p(b_{j,r} < 0|\hat{\boldsymbol{b}}_j, \hat{V}_j, z_j = p)\tilde{\pi}_{jp}] \quad (12)$$

For instance, using LFDR, we find that 96% of genes contain a QTL in at least one tissue, while using LFSR, we find that 88% of genes contain a QTL in at least one tissue.

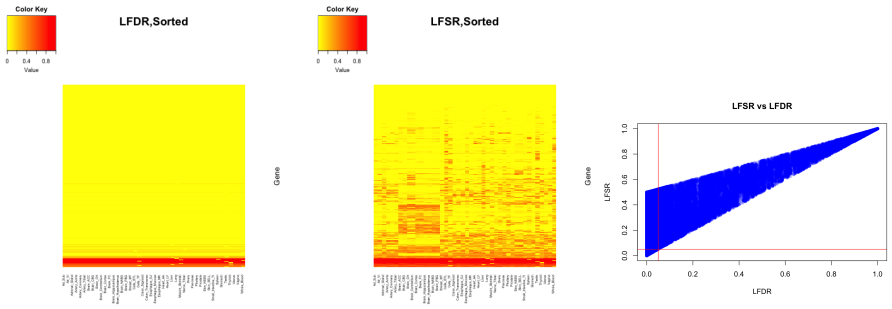

Figure 4. LFSR reveals more differences than LFDR. Left: Here, we characterize each effect by its lFDR: the posterior probability that the effect is 0. **Center**: We consider the lFSR for the same gene snp pairs across tissues. **Far right**: For each effect, we plotting its LFSR against the LFDR. Overall, we see that considering the LFSR reveals a wide array of variation among effects called 'non-zero' at a given LFDR threshold.

3.4 A qualitative description of heterogeneity in the GTEX data

We now want to understand the global patterns of heterogeneity present in the data set. Indeed, from the prior weight assigned to the 'learned matrices' coupled with the simulation results in the previous sections, we can see that `mash` is able to accurately parse shared configurations, thus resolving the relationship among tissues in which the QTL is active.

To emphasize the contrast between our approach and existing joint methods on this data-set, we compare our results to a configuration approach which recognizes only patterns constrained to lie along the x and y axis or along the $x - y$ line (Figure 1). `mash` allows for patterns which show consistently larger effects in one tissue over another, with varying amounts of correlation among tissues. We compare the first principal component of each of the the most important covariance matrices, as evaluated by their hierarchical weighting, reflecting the patterns 'learned' from the data and used as U_k to model the effect-size for each gene-snp pair $\boldsymbol{b}_j$ in Equation 3. Intuitively, this provides a single-rank 'summary' of the relationship in effect-sizes among tissues (Figure 5) captured by each pattern.

Each of the U_k thus reflects diverse relationships in effect-sizes among tissues: for instance, comparing the black and red pattern, we see that gene-snp pairs with high posterior probability of arising from the black class ($U_k = 2$) demonstrate consistently smaller and shared effects in brain than other tissues, while gene-snp pairs of the red class ($U_k = 3$) demonstrate strong effects in brain as compared to alternative tissues, for example. Indeed, matrix $U_k = 3$ captures gene-snp pairs with large, correlated effects in brain, and is the most prevalent pattern of sharing in the larger data set, as reflected by it's prior weight summed across grid-values. Matrix $U_k = 2$ captures SNPs with small effects in brain and larger effects in thyroid and transformed

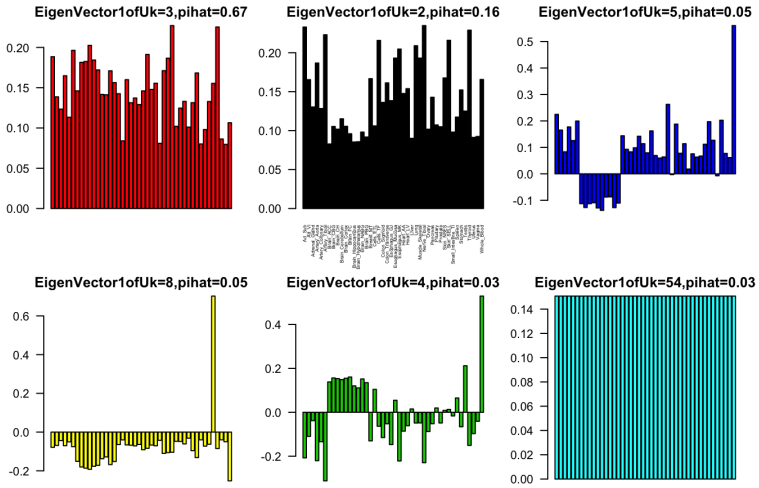

Figure 5. Relationships Among Tissues Captured. Here, we demonstrate the first principal direction of each pattern of effects across tissues, by simply taking the first eigenvector of each of the covariance matrices which receive more than 1% of the prior weighting, as well as the enforced `bmalite` configuration. Intuitively, these provide a rank 1 summary of the relationship in effect-sizes and directions among tissues captured by each pattern. They can be contrasted with the 'consistent configuration' which assumes the same effect-size for all tissues in which the tissue is active. We also consider the behavior of a gene-snp pair that had strong loading on one of the enforced `bmalite` configurations. See Text for details and possible interpretations, and Supplement for guide to tissue abbreviations.

cell-types (e.g., fibroblasts, lymphocytes). Similarly, we see that the patterns learned in U_k =4, 5 and 8 (blue and yellow) demonstrate a degree of 'quantitative' specificity: that is, consistently stronger effects in one tissue without restricting the effect-sizes to zero in alternative tissues. Because these patterns and their relative abundance are 'learned' from the data and allow for a superior model fit to methods which restrict effects as 'active' or 'qualitatively specific', we can use these to understand gross patterns present in the data as well as gene-snp specific effects. Examining the barplot of the relative importance of each of these patterns as learned from the data through empirical Bayes Methods and comparing it with a method which simply calls effects 'active or inactive' (`bmalite`) we see that such patterns are able to effectively 'dissect' the activity quantified in a fully 'consistent' configuration.

3.5 Tissue Specificity

One of the criticisms of a joint approach might be its loss of tissue-specificity. That is, by considering effects across subgroups in estimating the effect-size, one might lose sight of tissue-specific activity when it exists. Here, we demonstrate our ability to recognize such specificity both quantitatively, as described above through learned patterns of sharing which specify consistently larger effects in one tissue over others, and qualitatively through forced prior effect-size mass on $\mathbf{0}$. For each tissue, we can ask how many gene-snp pairs meet a given significance threshold in that tissue alone. Furthermore, tissue specific eQTL demonstrate the smoothing feature of this joint shrinkage approach: for gene-SNP pairs which demonstrate strong effects in only one tissue, the weaker erratic tissue are shrunk towards the prior mean at $\mathbf{0}$, resulting in a tissue specific smoothing (Figure 6, at right). We recognize an enrichment of tissue-specific effects in the transformed cell types, testes, whole blood, and thyroid.

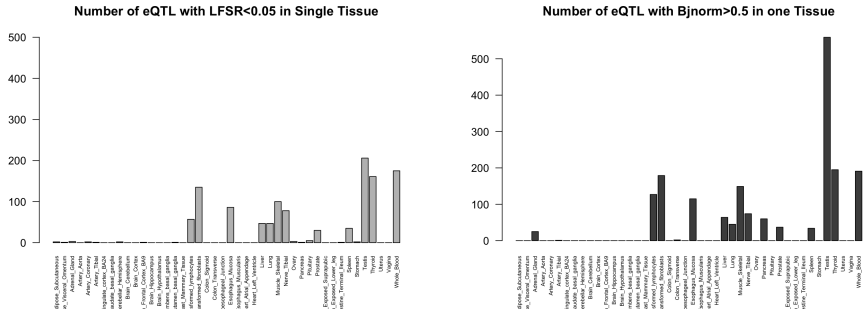

Figure 6. Tissue Specificity. Left: At an LFSR threshold of 0.05, we can ask how many QTL are specifically significant in only the given tissue. **Right:** We can also ask for the number of QTL that have a quantitatively specific effect in a given tissue, as defined by the number of b_{jrnorm}>0.5 in only one tissue.

3.6 Effective Sample Size Analysis

To try and understand which tissues might benefit the most from a joint calculation, we want to get an idea of the 'effective' sample size of a tissue.

We can consider the effect sample size for each tissue 'r' as

$$n_{jreffective} := n_{jroriginal} * \frac{\hat{s}^2_{jr}}{\tilde{s}^2_{jr}} \tag{13}$$

where $\hat{s}^2_{jr}$ is the standard error and $\tilde{s}^2_{jr}$ is the posterior variance for a particular tissue. We can also consider the increase in sample size incurred through our estimation - tissues which tend to 'share' more benefit from the joint analysis, and thus the ratio of $\frac{n_{jreffective}}{n_{jroriginal}}$ will be larger.

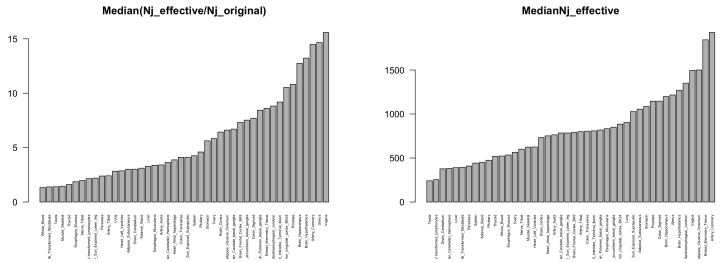

Figure 7. Posterior Marginal Variance Here, we report the increase in $n_{jreffective}$ and $n_{jreffective}$ (13).

We can see that the tissues with the greatest increase are part of 'groups' which intuitively share information: gender (uterus, vagina, ovary) and brain. This is particularly noticeable in brain, where sample sizes are often small.

I have also been applying MASH to additional data sets.

- I analyzed the association of genetic variants with 4 phenotypes in the global lipid data set [11] using input summary statistics for HDL, LDL, Total Cholesterol and Triglycerides.
- I analyzed the association of genetic variants a set of 5 kidney datasets providing input summary data for association with 5 kidney-related phenotypes [12].
- I applied Mash to Alan Hutchison's physics time series data. I also provided joint shrinkage estimates for Kushal Dey's gene-expression data across 5 tissues.

Presentations and Publications

- I presented a poster at both the Cold Spring Harbor Conference on Probabilistic Modeling in Genomics October 14-17 2015 and the Tukey Conference at Princeton University Sept 18 2015:

  `Matrix adaptive shrinkage: Modeling genetic effects across multiple subgroups`
  `Sarah M. Urbut, Gao Wang, Matthew Stephens`

- I presented a talk entitled *"Matrix Ash: Modeling Effect Sizes Across Tissues"* at the December 2 GTEx Analysis Working Group Jamboree meeting in Bethesda, Maryland.

- I have been working on a draft of the paper *"Matrix adaptive shrinkage: a general approach for estimating effects among multiple datasets*" since January.

Other Activities

- I served as a teaching assistant for Professor John Novembre at the Quantitative Bootcamp for incoming BSD students at the Woods Hold Marine Biological Laboratory.

- I completed my final required teaching assistant position for Introduction to Human Genetics in Fall 2015.

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 Apr;6(4). Available from: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2848547/`.

2. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, et al. High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. PLoS Genet. 2008 Oct;4(10):e1000214. Available from: `http://dx.doi.org/10.1371/journal.pgen.1000214`.

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

4. Wen X, Stephens M. Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene-environment interactions. The Annals of Applied Statistics. 2014 Mar;8(1):176–203. ArXiv:1111.1210 [stat]. Available from: `http://arxiv.org/abs/1111.1210`.

5. Consortium TG. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015 May;348(6235):648–660. Available from: `http://science.sciencemag.org/content/348/6235/648`.

6. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science (New York, NY). 2009 Sep;325(5945):1246–1250.

7. Stephens M. False Discovery Rates: A New Deal. bioRxiv. 2016 Jan;p. 038216. Available from: `http://biorxiv.org/content/early/2016/01/29/038216`.

8. Bovy J, Hogg DW, Roweis ST. Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations. The Annals of Applied Statistics. 2011 Jun;5(2B):1657–1677. Available from: http://projecteuclid.org/euclid.aoas/1310562737.

9. Efron B. Local False Discovery Rates; 2005.

10. Engelhardt BE, Stephens M. Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis. PLoS Genet. 2010 Sep;6(9):e1001117. Available from: http://dx.doi.org/10.1371/journal.pgen.1001117.

11. Global Lipids Genetics Consortium, Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, et al. Discovery and refinement of loci associated with lipid levels. Nature Genetics. 2013 Nov;45(11):1274–1283.

12. Pattaro C, Teumer A, Gorski M, Chu AY, Li M, Mijatovic V, et al. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. Nature Communications. 2016 Jan;7:10023. Available from: http://www.nature.com/ncomms/2016/160121/ncomms10023/full/ncomms10023.html.