Matrix Ash

Sarah Urbut [1,2], Gao Wang 1, Matthew Stephens [1,3,‡], with the GTEX Consortium[¶]

1 Department of Human Genetics/ University of Chicago, Chicago, IL USA
2 Pritzker School of Medicine/Growth and Development Training Program/University of Chicago, Chicago, IL USA
3 Department of Statistics/ University of Chicago, Chicago, IL USA

‡These authors also contributed equally to this work.
¶Membership list can be found in the Acknowledgments section.
*** CorrespondingAuthor@institute.edu**

Abstract

Author Summary

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). By analyzing these effects across multiple tissues, we exploit the information that the effect of the gene-snp pair in one tissue can provide about its effect in alternative tissues. Furthermore, quantifying the effect size as opposed to simply calling QTLs present or absent reveal many patterns of sharing of effects among tissues which differ in both sign and magnitude. We provide a novel framework for estimating effect sizes across multiple subgroups, considering the evidence contained in all subgroups jointly, which provides a powerful and detailed insight into quantitative heterogeneity present in the genome.

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 ([**?**, **?**]). 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 fails 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 limits our understanding of multiple-tissue phenotypes. Furthermore, even past attempts at quantifying heterogeneity of eQTLs using the data across tissues jointly were limited in both the number of tissues considered, and also the level of heterogeneity considered. Qualitative heterogeneity refers to calling a snp 'active' or 'inactive' in a given tissue. For example, previous work has 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' ([**?**, **?**]). However, a QTL may be'active' in all or many tissues and with varying magnitude or sign; we refer to this as quantitative heterogeneity. Indeed, our initial motivation came from our analysis of GTEx pilot data, in which we see evidence that many (50%) QTLs are shared across all nine tissues. In this context, a QTL is called based on whether it demonstrates significant posterior probability of being active in a particular tissue. Applying our hierarchical model to the dataset from Dimas *etal* 1001[?] with 3 tissues, we found just 8% of eQTLs are specific to a single tissue, with an estimated 88% of eQTLs being common to fibroblasts, LCL cells and T-Cells [**?**]. 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 eQTL which were shared in various 'configurations' or patterns of binary activity. However, these binary configurations are still
limited in their ability to capture continuous variation in levels of activity among tissues in which the SNP is considered active. 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 vary across tissues in which the SNP is called 'active'. 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. Thus effects can be 'shared' but not 'consistent' across tissues. The structure of this paper is as follows: we will describe our approach for modeling and estimating these effect sizes across tissues,

0.1 Methods Overview: Approach

We aim to learning about patterns of sharing across tissues within a SNP and among SNPs, which join to help us better understand the global and snp-specific patterns of effects of genetics on gene expression. This allows us to make comparisons among tissues in which the QTL is called active, and among gene-snp pairs with a similar degree of activity in a given tissue. Thus as an additional level of combining information, we assume that each eQTL may follow a particular pattern of activity characterized by its effects across tissues. Within these groups, the tissues exhibit characteristic patterns of sharing, which can be captured by considering the covariance structure of the genetic effects among tissues. This lends itself to a mixture model, in which we assume all the gene-snp pairs arise from a mixture of a finite number of multivariate normal (MVN) distribution, each characterised by the covariance matrix from which the vector of effects is though to arise. For each of J gene-snp pairs, we observe an R dimensional vector of standardized effect sizes $\hat{b}$ and their standard error and assume that these effects descend from some true effect size $\boldsymbol{b}$.

$$\boldsymbol{b_j}|\boldsymbol{\pi}, \mathbf{U} \sim \sum_{\mathbf{k,l}} \pi_{\mathbf{k,l}} \; N_{\mathbf{R}}(\mathbf{0}, \omega_{\mathbf{l}}\mathbf{U_k}) \tag{1}$$

Here, the covariance matrix U_k captures the particular patterns of sharing - variation in effect sizes within and between tissues, while ω_l determines the scale of each pattern - the magnitude of the effect size. Thus we recognize that while two eQTL may obey a similar pattern or shape, the absolute scale may vary. For example, two eQTL may both have strong correlation between tissues 1 and 2 with consistently larger effects in 2, but the absolute size of the effects may vary between SNPs.

Previous work from our lab considered the idea of configuration - i.e., that a tissue was simply 'active' or 'inactive' in a particular tissues - and thus for R tissues, there were $2^R possible configurations. which becomes computationally infeasible as R grows. Furthermore, t$ $[?, ?]) these matrices contain distinct diagonal and off -$ $diagonal elements which reflect data -$ $specific patterns of variation within and covariance between subgroups (tissues). This captures the$

Because we can't know the 'true covariance matrix' for each gene-snp pair, we aim to assemble a list which sufficiently captures the various patterns, and then 'learn' the relative proportions of each pattern of sharing from the data. One can now model each vector of effect sizes $\boldsymbol{b}$ each as arising from a mixture that captures all the covariance patterns.

The primary novelty of this approach is *to estimate this multivariate posterior distribution on the effect size in a data-sensitive way* - i.e., using the mixture model to capture information about the covariance structure among subgroups (here, tissues). Thus we might identify a situation in which it is common to have large effects in some tissues and not others, and thus if a gene-snp pair demonstrates a small effect in one of the 'off issues', we might be inclined to conclude that it is indeed a member of this particular class and shrink the small effect in this tissue accordingly. However, if we see the same small effect in a setting in which 'similar tissues' have large effects, we might 'shrink' this effect size less, due to our high prior belief in the SNP's effectiveness garnered from adjacent tissues. Thus we deem this method 'adaptive Shrinkage' because the appropriate amount of shrinkage is learned from the data.

Because our prior belief in consistency is strong, we identify many more ?significant associations? in setting where perhaps the observed univariate statistic in one tissue is small but otherwise large in additional tissues, nudging these effects towards something more consistent. This is in contrast to a univariate shrinkage approach, in which all effects of the same size would be 'shrunk' equivalently, due to lack of information garnered from adjacent tissues.

In fact, Shrinkage towards 0 of small effects is a result, not a necessity - since the majority of the prior weight is on small ω components which emphasize components with small prior variance of the effect size b, many of the modest z statistics will be smoothed or shrunk towards the prior mean, 0. An additional novelty is that in learning something about the effect size in each tissue for a given gene-snp pair, we can make statements about the degree of heterogeneity - that is the proportion of the time we expect a SNP to have effects of different sign. We will be confident in our ability to identify the direction of the effect for A SNP with a large effect and relatvilely high precision, and thus we can use an estimate of the posterior mean in each component and the proportion to quantify the distribution of gene SNP pairs who have effects of opposite direction (or lack convincing evidence of effects in a consistent direction across tissues).

Materials and Methods

Let $\boldsymbol{b_j}$ represents the genetic effect of SNP-gene pair j across $R = 44$ tissues.

We assume the following mixture prior for the R dimensional vector of true effects,

$$\boldsymbol{b_j}|\boldsymbol{\pi}, \mathbf{U}, \boldsymbol{\omega} \sim \sum_{\mathbf{k,l}} \pi_{\mathbf{k,l}} \; N_{\mathbf{R}}(\mathbf{x}; \mathbf{0}, \omega_{\mathbf{l}} \mathbf{U_k}) \tag{2}$$

Where $N_R(.; \mathbf{0}, \omega_l U_k)$ denotes the density of a normal distribution with mean $\mathbf{0}$ and variance $\omega_l U_k$.

Each component of the mixture distribution is characterized by these prior covariance matrices, U_k which capture the pattern of effects across tissues. Critically, this prior distribution is the same for all J - hence the hierarchical incorporation of shared information.

0.2 Covariance Matrices

For a given ω_l, we specify 4 'types' of RxR prior covariance matrices $U_{k,l}$.

1. $U_{k=1,l} = \omega_l \; \mathbf{I}_R$
2. $U_{k=2,l} = \omega_l \mathrm{X}_z$ The (naively) estimated tissue covariance matrix as estimated from the column-centered J $\times R$ matrix of Z statistics, Z_{center}: $\frac{1}{J} \; {Z_{center}}^t \; Z_{center}$
3. $U_{k=3,l} = \omega_l \; \frac{1}{J} \; V_{1...p} \; d^2_{1...p} \; V^t_{1..p}$ is the rank p eigenvector approximation of the tissue covariance matrices, i.e., the sum of the first p eigenvector approximations, where $_{1...p}$ represent the eigenvectors of the covariance matrix of tissues and $_{1...p}$ are the first p eigenvalues.
4. $U_{k=4:4+Q-1,l} = \frac{1}{J}((\Lambda\mathbf{F})^t \Lambda\mathbf{F})_q$ corresponding to the q_{th} sparse factor representation of the tissue covariance matrix
5. $U_{k=4+Q,l} = \frac{1}{J} \; (\Lambda\mathbf{F})^t \Lambda\mathbf{F}$ is the sparse factor representation of the tissue covariance matrix, estimated using all q factors.
6. $U_{k=5+Q:R+4+Q,l} = \frac{1}{J} \; ([100..]'[100...])$
7. $U_{k=R+5+Q,l} = \frac{1}{J} \; ([111...]'[111...])$
8. [1000...] or [111...] represent configurations such that given membership,$\boldsymbol{b_j}$ arise from the same prior variance.

0.3 Deconvolution

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 matrices specified in (2), (3) and (5). The final results 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 q_{ij}) (Bovy et al, 2014). This allows us to write down the 'full data' log likelihood as follows:

$$\phi = \sum_J \sum_K q_{jk} ln\alpha_k N(\hat{\boldsymbol{b_j}}|0,\, U_k + V_j)$$
$$\phi = \sum_J \sum_K q_{jk} ln\alpha_k N(\boldsymbol{b_j}|0,\, U_k) \tag{3}$$

Where α_k represents π_k and q_{jk} is the latent identifier variable.

0.4 Likelihood

By maximum likelihood in each tissue separately, we can easily obtain the observed estimates of the standardized genotype effect sizes, $\hat{\boldsymbol{b}}_j$, and their observed squared standard errors recorded on the diagonal of an $R \times R$ matrix noted $\hat{V}_j = (\hat{\boldsymbol{b}}_j)$. We assume that the matrix of standard errors of $\hat{\boldsymbol{b}}_j$, V_j as approximated by $\hat{V}_j$ is diagonal and that $\hat{V}_j$ is an accurate point estimate for the standard error and that these standard errors are independent between tissues.

If we now view $\hat{\boldsymbol{b}}_j$ and $\hat{V}_j$ as *observed data*, we can write a new "likelihood" using only the sufficient statistics, $\hat{\boldsymbol{b}}_j$ and $\hat{V}_j$:

$$\hat{\boldsymbol{b}}_j|\boldsymbol{b}_j \sim_R (\mathbf{x}; \mathbf{b_j}, \hat{\mathbf{V}}_\mathbf{j}) \tag{4}$$

0.5 Posterior Quantities

We aim to report posterior quantities for a given gene-snp pair **j**. We know that for a single multivariate *Normal* the posterior on $\boldsymbol{b}|U$ is simply:

$$\boldsymbol{b}|\hat{\boldsymbol{b}} \sim N_R(\tilde{\boldsymbol{\mu}}, \tilde{U})$$

where:

- $\tilde{\boldsymbol{\mu}} = \tilde{U}(\hat{V}^{-1}\hat{\boldsymbol{b}})$
- $\tilde{U} = (U^{-1} + \hat{V}^{-1})^{-1}$.

Let us concatenate the list of all KxL combinations of prior covariance matrices U_k and their scaling parameters ω_l into a KxL list and assign this length K for simplicity of notation.

Now each U_k imparts information about both *scale* and *direction* . Furthermore, a mixture-multivariate normal prior and a normal likelihood yields a mixture multivariate posterior, where the final posterior distribution is simply a weighted combination of multivariate normal distributions, where for each gene-snp pair **j** is now characterized by it's posterior mean $\tilde{\boldsymbol{\mu}}_{jk}$ and covariance $\tilde{U}_{jk} = (U_k^{-1} + \hat{V}_j^{-1})^{-1}$.

$$\boldsymbol{b}_j|\hat{\boldsymbol{b}_j}, \hat{V}_j, \hat{\boldsymbol{\pi}} \sim \sum_k^K \tilde{\pi}_{jk} N_R(\mathbf{x}; \tilde{\boldsymbol{\mu}}_\mathbf{jk}, \tilde{\mathbf{U}}_\mathbf{jk}) \tag{6}$$

Where again, $N_R(.; \mathbf{0}, \omega_l U_k)$ denotes the density of a normal distribution with mean $\tilde{\boldsymbol{\mu}}_k$ and variance $\tilde{U}_k$ and the posterior mixture weight $\tilde{\pi}_k$ is simply

$$\tilde{\pi}_{jk} = \frac{p(\hat{\boldsymbol{b}}_j|\hat{V}_j, z_j = k)\hat{\pi}_k}{\sum_{k=1}^K p(\hat{\boldsymbol{b}}_j|\hat{V}_j, z_j = k)\hat{\pi}_k} \tag{7}$$

Where $z_j = k$ is the latent variable indicator of the component identity and each $\hat{\pi}_k$ represents the Maximum Likelihood Estimate of the prior mixture weights assigned to each component.

0.6 Reported Quantities

For every gene-snp pair 'j', we aim to report the effect size as the posterior mean, defined as:

$$E(\boldsymbol{b}_j|\hat{\boldsymbol{b}_j}, \hat{V}_j, \hat{\boldsymbol{\pi}}) = \sum_k^K \tilde{\pi}_k \tilde{\boldsymbol{\mu}}_k \tag{8}$$

And the local false sign rate, or posterior probability of incorrectly identifying the sign of the effect for a given tissues 'r' as :

$$P(b_{jr}) = 1 - max[\sum_k p(b_{j,r} > 0|\hat{\boldsymbol{b}}_j, \hat{V}_j, z_j = k)\tilde{\pi}_k, \sum_k p(b_{j,r} < 0|\hat{\boldsymbol{b}}_j, \hat{V}_j, z_j = k)\tilde{\pi}_{jk}] \tag{9}$$

1 Results

1.1 Demonstrating Features of the Method

To get a sense of the accuracy of our novel approach to estimating multivariate effects, we simulated two types of data.

In the first set, in which we expect our method to be superior to both univariate methods and methods in which the configuration approach is utilized, we simulate 50,000 gene-snp pairs, with only 400 representing true signal. This represents roughly 500 genes with 100 snps in cis, 80% of which contain one active QTL. Thus naturally, if the gene contains such a QTL, it is the same QTL among all tissues in which the tissue is active. This puts a dual burden on both features of the method: The small number of true associations present in these simulations tests whether the method accurately encourages small observed effects toward zero while preserving the true signal when it exists. Furthermore, the multivariate nature of these simulated effects when they exist tests the ability of the method to accurately infer patterns of sharing from the dataset. These true effects are thus simulated from the 'learned' covariance matrices representing U_k 2-9, and thus aim to emulate the patterns of sharing present in real biological data. We compare with univariate 'shrinkage' method Ash (Stephens et al, unpublished) as well as the eqtlBMA-lite (Flutre et al, 2013) which uses the singleton and fully consistent (i.e., active in only one tissue, or active with the same effect size in all tissues) configurations to estimate these effects jointly. We call this the 'sharing' (S) scenario.

One might expect that our method would prove superior only in the setting in which true effects are shared among all tissues, and thus fail in the setting of tissue specificity. Thus, building on the situation above, we add a simulation in which 35% of the true effects are active in only one tissue, according to 5 different patterns of tissue specificity. We call this the 'tissue-specific' (TS) scenario.

We see that in terms of both power and accuracy, Matrix Ash is superior in both setting to both univariate methods, and to joint analysis under the setting of orthogonal configurations.

$$RMSE\sqrt{(\sum_{jr}(b_{jr} - E(b_{jr}|Data)^2))} \quad (10)$$

Table 1. Accuracy Comparison: RMSE

Inference Method	MASH	ASH	eqtlBMA-lite
RMSE_S	0.010	0.030	0.047
RMSE_{TS}	0.008	0.025	0.043
cor.with.truth_S	0.99	0.94	0.84
$\text{cor.with.truth}_{TS}$	.99	0.94	0.82

Table 2. Accuracy Analysis Here, we compare the ability of matrix ash ('MASH') to capture the true effect size estimates. We compare with univariate-shrinkage method 'ASH' and configuration-specific joint approach 'eqtl-BMA-lite'. We report the Root Mean Squared Error (RMSE) and the correlation with the truth.

To demonstrate the ability of Matrix Ash to powerfully capture these accurately estimated effect sizes, we compare the proportion of true associations called significant at a given significance threshold among the three methods. Indeed, Matrix Ash proves superior to both methods under each condition (i.e., sharing or tissue-specific).

INSERT POWER VS. ACCURACY FIGURE HERE

In introducing a method to quantify the heterogeneity of effect sizes, we have developed a 'heterogeneity index' which attempts to capture the heterogeneity in magnitude among tissues in which the gene-snp pair is active. For each gene-snp pair **j**, we normalize its vector of effects across tissues by the effect which has the maximum absolute value; thus for a fully 'consistent' gene-snp pair in which all the effects are equal in magnitude, the new vector of normalized effects would consist of all ones, and R=44 tissues would be greater than 50% of the maximum effect. By contrast, for a tissue-specific gene-snp pair, the vast majority of effects would be small fraction of the maximum effect and thus the number of tissues greater than 50% of the maximum effect would be 1 (the effect used to normalize). We can apply this heterogeneity index, here deemed 'HI' to the real data, but first wanted to demonstrate the superiority of Matrix Ash in estimating these quantities on simulated data. To quantity the ability of each method to accurately ascertain the heterogeneity, we can compute the heterogeneity index of the real data, and the inferred quantities, and use a modified RMSE:

$$RMSE_{HI}\sqrt{(\sum(true_{HI} - estimated_{HI})^2)}) \tag{11}$$

Table 3. Accuracy Comparison: RMSE

Inference Method	MASH	ASH	eqtlBMA-lite
HI_S	39.38	40.87	39.78
HI_{TS}	39.98	40.77	39.51

1.2 Adaptive Shrinkage: The Multivariate Approach

To demonstrate the utility of shrinking effect size estimates jointly, we consider the estimated effect sizes vs their observed input summary statistics using our joint (Matrix Ash) and comparing to a univariate shrinkge method (ASH). On simulated data, we can also then plot the estimated effect sizes against the true values, again comparing among methods. Here, we show the results under the setting of tissue specificity, to analyze the behavior of eQTL of each class.

INSERT TSPEC SIMULATED SCATTERPLOTS HERE

Comparing estimated Z statistics vs the observe 'raw' input valuesIn both Matrix Ash and univariate methods, values with large standard errors will be shrunk more harshly. Furthermore, in this simulated data, where there is an abundance of small effects, tend to shrink small z statistics towards prior mean at 0 as their likelihoods will be maximized by componenent with small . However, now considering z statistics of the same size (so now accounting for standard erro), not all small values are shrunk to the same extent using Matrix Ash, due to the power of joint analysis to consider the effects across tissues in inferring the final vector of effect sizes/ Acknowledging consistency, small effects in one tissue will be augmented in the presence of larger effects present in other tissues, resulting in dramatic power increases.

Furthermore, when we plot the estimated effects vs the truth and segregate these effects by class, we see that the correlation among the truth and estimated effect sizes is much tighter using our multivariate approach. Similarly, truly null effects are shrunk more tightly, due to the fact that in the presence of consistency, small effects across subgroups will lead us to have a high prior belief that an additional small observed effect in that eQTL is also likely to be close to 0. Importantly, tissue specific SNPs are still captured using our joint approach, demonstrating that have tissue-specific patterns exist in the data, our prior belief will capture this phenomenon and accordingly our posterior estimates will reflect the underlying tissue specific nature of these activities.

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 heterogeneity of effect sizes among tissues.

2 Real Data

Now, we consider the results of our analysis, when applied to the GTEX data set. After estimating the covariance matrices from the strongest Z statistics in the data, thus demonstrating the strong underlying 'true patterns' of sharing in the data and adding the qualitatively specific configurations, we then inferred the relative frequency of each pattern of sharing and corresponding effect sizes from a large sample of 40,000 gene snp pairs. 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 effect size in absolute value across tissues. As described above and demonstrated in simulations, in the setting of an abundance of small effects in data set, we tends to shrink small z statistics towards the prior mean at 0. It should be noted that this is a result specific to a particular data set, and in that sense 'adaptive' - indeed, if small effects were rare and large effects abundant, such shrinkage would not occur.

But perhaps more importantly, the striking increase in power when compared to univariate methods is noted. 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 an $lfsr$ threshold of 0.05, we identify 393,414 significant snp-gene-tissue effects. Using the naive univariate summary statistics and using an FDR threshold of 0.05, we identify only 202,087, meaning that using univariate methods we would only call an effect non-zero in less than a third (28%) of cases, while using our posterior effect estimation we would say that we can confidently argue the SNP has a non-zero effect for a gene in a particular tissue over half (55%) of the time. We would expect this increase in power, because small Z statistics in a tissue will be increased for a given gene in the present of larger z statistics in other tissues.

SHOW REAL DATA SCATTERPLOT

Table 4. Power Comparison

Metric	$\text{LFSR}_{MatrixAsh}$	LFSR_{ASH}	
Significant $\mathbf{b_{jr}} \leq 0.05$	393414	91755	401552
$\mathbf{b_{jr}}$ significant in other not in MASH	NA	1447	
$\mathbf{b_{jr}}$ significant in MASH not in other	NA	303106	

Table 5. Power Restricting our analysis to thresholding by local false sign rates, we can quantify the number of associations we identify at a given local false sign rate threshold using the original summary statistics and posterior means computed using multivariate Matrix Ash and Univariate Ash. We can see that Matrix Ash calls nearly twice and 4 times as many associations significant, and misses very few associations identified in other methods. These are rare examples in which modestly large effects in one tissues are 'nudged' towards small effects as observed in the additional tissues for that SNP.

While the number of associations catpure is slightly greater using the BMA lite approach, we note that the likelihood of the data set under this model is much much worse (1298672 vs -1267997.5). Indeed, eQTLBMA would put the vast majority of the prior weight on the fully 'consistent' configuration, as SNPs demonstrating activity

acorss all tissues, regardless of how heterogeneous among subgroups, are forced into this configuration.

SHOW Matrix Ash PRIOR WEIGHT VS BMA

Considering a two tissue example, a configuration type approach recongizes only patterns constrained to lie along the x and y axis or along the x-y line. Matrix ash allows for patterns which show consistenty larger effects in one tissue over another, with varying amounts of correlation among tissues. In these example from real data, we can see that while SNPS of the green, blue and yellow class appear consistently active in both tissues plotted (Brain and Muscle), the blue and yellow effects have consistently larger effects in Muslce then brain, while SNPs of the green class show the reverse. Similarly, SNPS of the yellow close show only loose correlation among the pair of tissues, while SNPs of the green class show strong prediction of activity in one from activity in the other. Similarly, we can see that SNPS of the pink class tend to show tisu-specificity in Whole Blood relative to testis, while SNPs of the black class show tissue specificity in testis relative to whole blood.

SHOW COLORED SCATTERPLOT

2.1 A qualitative description of heterogeneity in the GTEX data

Indeed, from the prior weight assigned to the 'learned matrices' above coupled with the simulation results in the previous sections, we can see that Matrix ASH is able to accurately parse shared configurations. Focusing on the two predominant patterns, we see that Here we examine several examples with a high posterior probability of arising from this particular component while the learned matrix $U_k = 3$ seems to capture gene-snp pairs with large, correlated effects in brain, matrix $U_k = 2$ captures SNPs with small effects in brain and larger effects in thyroid and transformed cell-types (e.g., fibroblasts, lymphocytes). The lower which rank high in importance (e.g., $U_k 5 and \mathrm{U}_k = 9$) show somewhat tissue specific (i.e., high prior variance in only one tissue-type) effects in testes and whole blood, consistent with our conclusions that whole blood and testes indeed demonstrate an abundance of tissue-specific gene-snp pairs.

2.2 Examples of strong loading on Uk3, Uk9 and eQTL-BMA lite

Uk3: Captures correlation in sign, quantitative heterogeneity in magnitude along diagonal emphasizes the utility of continuous approach In this particular example, strong effects in brain matching an underlying pattern of shared effects among brain tissue is well-captured by the data and thus allows this gene-snp pair to find its true match. brain effect sizes thus borrow strength from one another, and the posterior estimates tend to nudge the brains towards a consistent, shared effect. Similarly, an overall tendency towards consistency tends to 'flip' erratic off directions towards the prevailing positive direction.

Uk 9: Quantitative specificity in magnitude in testes/Whole Blood In this example, though this paritculr pattern captures correlation in sign among all tissues, the quantitative heterogeneity is reflected in the intnesity of the badning along the diagonal, and thus intorduces the idea of quantitative specificity - e.g., that a SNP can be modestly 'active' in all tissues though to dramatically different degrees. here, though this matrix was learned (and not forced, as in eQTLBMAlite) from the data, the pattern of quantitative tissue specificity in testes and whole blood is evident. Again, erratic, off-directions are flipped in sign.

Lastly, the inclusion of the eQTLBMA lite configurations (in which the SNP has a non-zero effect in only one tissue) coupled with the learned patterns of tissue specificit evident in matrices $U_k 5-9$ serve to allow the preservation of qualitatively specific effecs. Here, we show a gene-snp pair demonstrating high loading on one of the eqtlbma lite configuration matrices - indeed, we reject the significance of the effect size estimates in all tissues but testes, a pattern consistent with the presence of tissue-specificity described below.

2.3 Tissue Specificity

One of the criticisims 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 recongize such specificity both quantitatively (through learned patterns of sharing which specifiy consistently larger effects in one tissue over others) and qualitatively (through forced prior effect size mass on 0). For each tissue, we can ask how many gene-snp pairs meet a given tsignficiance threshold in that tissue alone. **NUMBER OF QTL PER TISSUE PLOT**

Furthermore, tissue speicific eQTL demonstrate the smoothing feature of this joint shrinkage approach: for gene SNP pairs which demonstrate strong effects in only one tissue, the weaker errativ tissue are shrunk towards the prior mean at 0, resulting in a tissue specific smoothing.

TISSUESPECIFIC SMOOTHING PLOT

2.4 Quantifying Heterogeneity

Armed with a vector of effect size estimates across 44 tissues, 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. From a biological standpoint, we might consider think that effects of a different sign are rare. Similar to the heterogeneity index described in the simulation framework above which attempted to describe heterogeniety in magnitude, we can plot the number of tissues in which the sign is differ than the effect with maximum absolute value. Considering this 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) and removing brains from our analysis tends to push us towards consistency, suggesting that brain appears to behave as a large tissue-specific entity.

Sign Heterogeneity Hist

Furthermore, we can now quantify the heterogeneity index in magnitude described in the simulation framework above, and ask, for each gene, in how many tissues is the effect a certain proportion, suppose 50%, of the maximum effect? Again, homogenous genes will tend to be featured at the right of the distribution, with the majority of their tissues effect sizes similar in magnitude, while heterogneous genes will be featured towards the left, with tissue-specific genes at the extreme left. Again, excluding brain form the analysis tends to nudge us towards a belief in consistency. We can also consider how many gene-snp-tissue (i.e., b_{jr}) effects are greater than 50% of the maximal value for the gene (gene 'j'): such a contrast is evident here as

Taken together, these results suggest the presence of consistency in sign in our data set, and a bimodal distribution of heterogeneity in magnitude.

Magnitude Heterogeneity Index

Table 6. Heterogeneity Comparison

Data	All Tissues	No Brains
E(DifferentSignPosteriorMean| LFSR≤0.05)	0.802	0.852
E(At least 50% max value)	0.354	0.449

Table 7. Heterogeneity Analysis At a given significance threshold, we can ask how many gene-pairs contain effects of different signs across tissues. At an arbitrary LFSR threshold of 0.05 for instance, we note that 80% of genes are homogenous in sign when all tissues are considered. Excluding brains from our analysis, this rises to 85%. 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. Again, we see that excluding brains from our analysis tends to push this towards consistency.

Attempting to understand which genes tend to behave the most homogenously or heterogeneously, we can plot the value used to normalize each gene, e.g., the 'maximum' effect size across tissue of the gene, against the normalized values. We can see that if a large effect is present, it tends to be in the presence of homogenous effects across the board, while small normalizing effects tend to be in the presence of effects that are more variable in sign and magnitude. Furthermore, aggregating the gene-snp pairs at a given heterogeneity index and classifying them by the effect used to normalize (e.g., the 'max effect') we can see that gene-snp pairs with greater Heterogeneity indices tend to have larger effects.

Insert BIPLOT Insert Median MAX EFFECT by HI Index Insert

3 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 above: 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. Again, we choose 20,000 random-gene snp pairs and use the EM algorithm to learn the mixture proportions π from this data set.

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 'overfit' to the training data set, than a larger number of covariance matrices may actually decrease the test log-likelihood.

I found that the K=1188 set of covariance matrices containing the Identity, the denoised empirical covariance matrix, rank 5 SFA approximation and rank 3 SVD approximation as well as 5 single-rank SFA factors and the 45 *eqtl.bma.lite* configurations maximized this likelihood.

4 Training and Testing Procedure: Estimating Hierarchical Weights

We wish to choose the model which best maximizes the probability of observing the data set.

Incomplete Data likelihood:

$$L(\boldsymbol{\pi}; \hat{\boldsymbol{b}}) = \prod_{j=1}^{J} \sum_{k}^{K} \pi_k P(\hat{\boldsymbol{b_j}} | z_j = k) \tag{12}$$

- To estimate the hierarchical prior weights π_k: compute the likelihood at each each gene snp pair j by evaluating the probability of observing $\hat{\boldsymbol{b_j}}$ given that we know the true $\boldsymbol{b_j}$ arises from component k
- Use the EM algorithm to estimate the optimal combination of weights: How often does this particular covariance matrix occur in the data?

We then use these weights to estimate the test set log likelihood.

Discussion

Supporting Information

Acknowledgments

References

1. Devaraju P, Gulati R, Antony PT, Mithun CB, Negi VS. Susceptibility to SLE in South Indian Tamils may be influenced by genetic selection pressure on TLR2 and TLR9 genes. Mol Immunol. 2014 Nov 22. pii: S0161-5890(14)00313-7. doi: 10.1016/j.molimm.2014.11.005

2. Huynen MMTE, Martens P, Hilderlink HBM. The health impacts of globalisation: a conceptual framework. Global Health. 2005;1: 14. Available: http://www.globalizationandhealth.com/content/1/1/14.