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

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

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). The availability of systematically generated eQTL information could provide 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 [**?**]. However, most studies to date have been conducted in a single immortalized peripheral cell type, and it is unclear to what extent these findings will translate to human disease mapping across more varied cell types. Furthermore, even analyses performed on additional cell types are often performed in a single tissue framework [**?**, **?**] and fail to correlate the effect of genetics across multiple tissue types. The Genotype Tissue Expression Project, GTEx, Project will provide the data necessary to address this situation: by 2016, the resource is expected to enroll a total of approximately 900 post-mortem donors, with approximately 30 tissues collected from each donor, and the project will generate extensive genotype data and RNA-seq data on each individual. However, available methods are limited in their ability to *jointly analyze data on all tissues* to maximize power, while simultaneously *allowing for both qualitative and quantitative differences among eQTLs* presentin each tissue.

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 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

However, most studies to date have been conducted in a single immortalized peripheral cell type, and it is unclear to what extent these findings will translate to human disease mapping across more varied cell types.

Furthermore, even analyses performed on additional cell types are often performed in a single tissue framework and fail to correlate the effect of genetics across multiple tissue types.

The Genotype Tissue Expression Project, GTEx, Project will provide the data necessary to address this situation: by 2016, the resource is expected to enroll a total of approximately 900 post-mortem donors, with approximately 30 tissues collected from each donor, and the project will generate extensive genotype data and RNA-seq data on each individual. However, available methods are limited in their ability to *jointly analyze data on all tissues* to maximize power, while simultaneously *allowing for both qualitative and quantitative differences among eQTLs* present in each tissue.

0.1 Aim 1

Develop methods of estimating the posterior effect size across multiple subgroups, thereby mapping eQTLs

- Combine information across tissues
- Report an effect size
- Capture distinct variation in effect sizes within and between subgroups: 'patterns of sharing'

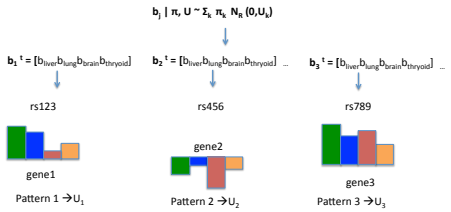
0.2 The Setting

- How do we quantify the effect of a particular SNP on gene expression among tissues?
- Approach 1: The Isolationist Approach

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.

0.3 The Setting

- How do we quantify the effect of a particular SNP on gene expression among tissues?



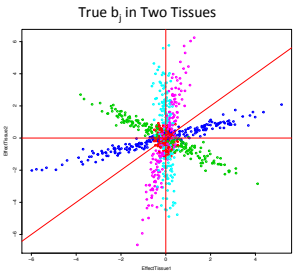
- Approach 2: The Joint Committee Approach the multivariate nature of this activity
- Many different patterns of sharing of effects among tissues.
- But how do we learn about the nature and frequency?

Here, we recognize the multivariate nature of the effects. Now, we aim to make an inference about the vector of true effects across tissue and thus model this effect b as an R dimensional vector composed of the true effect of the gene snp pair across multiple tissues. across tissue and thus model this effect across tissue and thus model this effect. Acknowledging the multivariate nature of this effect opens the door to illustrating many different patterns of sharing of effects among tissues. But how do we learn about the nature and frequency of these patterns?

0.4 Considering ALL the evidence!

- Each eQTL may follow a particular pattern of activity
- Within these groups, the tissues exhibit characteristic patterns of sharing of effects
- Captured by considering the covariance structure of the genetic effects among tissues.
- Natural mixture model: Each component of the mixture is defined by the prior covariance matrix U_k from which the vector of standardized effect sizes of this class is thought to be drawn.
- Learn relative frequencies from the data

Not only do we have many tissues, we also have an entire genome from which to 'learn' about these patterns of sharing. 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 $\boldsymbol{b}$. 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 natural mixture model, in which we assume all the gene-snp pairs arise from a mixture of a finite number of Gaussian distributions with unknown parameters. Because of the the multivariate nature of this activity, within a particular 'pattern of sharing' some tissues may be more active than others, but not completely on or off. Thus each component of the mixture is defined by the prior covariance matrix from which the vector of standardized effect sizes $\boldsymbol{b}$ is thought to be drawn. The covariance matrix of the true effects thus reflects a particular pattern of sharing, such that the diagonal elements of the component-specific covariance matrix then represents the variance of the effect size



within and and the off–diagonal between tissues . A large prior variance in one tissue and small in another means that effects in the first tissue tend to be large while effects in tissue two tend to be small. 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.

As a critical innovation on our previous method (1001[?, ?]) 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 variation in effect sizes within and between subgroups better than restricting effects to simply 'shared' or 'unshared' between subgroups.

Previous work from our lab considered only the idea that the covariance between two tissues was the same across tissues thought to contain a QTL in a given pattern, or 'configuration', and thus failed to incorporate the much richer covariance structure between tissues.

The primary novelty of this approach l 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.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

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).

0.5 So what are the Prior Covariance Matrices $U_k s$ specifiying?

Suppose we have just two tissues

- Direction defined by relative ratio in effect size between tissues, specified in prior covariance of $\boldsymbol{b}$

Generative U_k for blueSNPs $\begin{pmatrix} Var(b_1) = blue2.0 & Cov(b_1, b_2) = 0.56 \\ Cov(b_2, b_1) = 0.56 & Var(b_2) = blue0.20 \end{pmatrix}$

Additional novelty: Ratio between tissues is flexible (not simply shared or tissue-specific) and data sensitive (stay-tuned)

To illustrate the utility of using a variety of covariance matrices, consider that we have snps of 4 'types' here defined, by their effects in tow tissues. eQTL of the blue class tend to have large effects in tissue 1 and small in tissue 2, while snps in the purple class have very large effects in tissue 2 and small effects in tissue 2. These directions are thus specified in the prior covariance matrix which defines the direction - here simply ratio - in prior effect size between tissues,

The 45 degree angle and the lines would be simply using tissue specific or shared effects, while we can have a much richer understanding of the relationship between effect sizes using a set ofcovariance matrices that aims to recapitulate patterns found in the data

0.6 Why care about the effect size?

- 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.
- The addition of the quantitative comparison captures the continuous nature of biological phenomenon.
- How confident are we in the sign of the effect?
- Acknowledge the many patterns of sharing present in the data, wide array of prior covariance matrices allows our gene-snp pair to find it's true pattern of sharing

Given that an eQTL is called active in two tissues, we want to make more statements about our confidence in the sign and magnitude of the effect among tissues.

Similarly, if many SNPs are called active in a particular tissue, we can resolve differences among these gene=snp pairs and using this wide array of prior covariance matrices allows our gene snp pair to find it's true pattern of sharing.

Adaptive Shrinkage: So why does the likelihood increase at the 'right component'? Consider the univariate case. Here, think about x as $\hat{b_1}$, and σ as $U_{k[1,1]} + \hat{V}_{j[1,1]}$.

To compute the likelihood at each component:

$$f(x \mid \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x)^2}{2\sigma^2}} \tag{1}$$

We can see that this will be largest when σ^2 approaches the MLE, which where is simply $\hat{b}^2$. This is the intuition behind the bayes factor being the largest at the 'true component'.

Materials and Methods

We assume the following mixture prior for the R dimensional vector of true effects, $\boldsymbol{b_j}$ represents the genetic effect of SNP-gene pair j across $R = 44$ tissues:

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

As mentioned above, each component of the mixture distribution is characterised by these prior covariance matrices, U_k which capture the pattern of effects across tissues.

0.7 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. $U_{k=3,l} = \omega_l\ \frac{1}{J}\ V_{1\ldots p}\ d^2_{1\ldots p}\ V^t_{1..p}$ 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.
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. $U_{k=4+Q+1,l} = \frac{1}{J}\ (\Lambda\mathbf{F})^t\Lambda\mathbf{F}$ is the sparse factor representation of the tissue covariance matrix, estimated using all q factors.
7. $U_{k=5+Q:R+4+Q,l} = \frac{1}{J}\ ([100..]'[100...])$
8. $U_{k=R+5+Q,l} = \frac{1}{J}\ ([111...]'[111...])$
9. [1000...] or [111...] represent configurations such that given membership,$\boldsymbol{b_j}$ arise from the same prior variance.

0.8 Deconvolution

To retrieve a 'denoised' or 'deconvoluted' estimate of the non-single rank dimensional reduction matrices, I then perform deconvolution.em which initializes 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 component 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:

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

Where α_k represents $p(k)$ and q_{jk} is the latent identifier variable..

0.9 Generation of List of Covariance Matrices

I then use these three non single-rank covariance matrix in place of our original choice of the empirical covariance matrix, SFA and SVD approximations. Here, I also used the Identity (K=1), 5 single-rank SFA factors (K=4-9), and the 44+1 *eqtlbma.lite* configurations (K=10:54) in steps (7) and (8) to a assemble a full list of covariance matrices. Briefly, these *eqtlbma.lite* are an attempt to capture 'singleton' and 'fully shared' configurations in which the gene-snp pair is active in only one or all tissues. In the latter case, the variance of the distribution of underlying effect sizes is equal in all tissues. This is 54 matrices, and we then proceed to chooses an $'L'$ element grid according to the range of effect sizes present in the initial 16,069 x 44 matrix of strong Z statistics to create a KxL list of covariance matrices. In the GTeX data set we choose a grid with 22 omegas for a total of 1188 covariance matrices.

0.10 Mixture Weights

We now need to compute the mixture weights $\boldsymbol{pi_{kl}}$ hierarchically - that is, using all of the data to determine the optimal mixture of covariance matrices. I use a randomly chosen set of 20,000 gene-snp pairs to estimate these mixture proportions. This set does not contain the strongest gene-snp pairs, and thus will allow for substantial shrinkage, as a majority of these gene-snp pairs will have their likelihood maximized at low ω components. To compute the likelihood at each component: For a given gene-snp pair, the Likelihood on $\boldsymbol{b}$:

$$\hat{\boldsymbol{b}}|\boldsymbol{b} \sim N_R(\boldsymbol{b}, \hat{V}) \tag{4}$$

Furthermore, we take advantage of the fact that

$$N_R(\hat{\mathbf{b}}_j; \mathbf{0}, U_k + \hat{V}_j) \tag{5}$$

Thus we can now treat the matrices U_k as fixed and computed the JxK matrix of likelihoods to the combinations of weights which maximizes the probability of observing the data. Concatenating the list of JxK matrices to a list of K components, use the EM algorithm to estimate $\hat{\pi_k}$ as the set of all weights which maximizes the following likelihood:

$$L(\boldsymbol{\pi}; \hat{\boldsymbol{b}}) = \prod_{j=1}^{J} \sum_{k}^{K} \pi_k P(\hat{\boldsymbol{b_j}}|z_j = k) \tag{6}$$

0.11 Posterior Quantities

Now that I have the estimated these prior mixture weights stored in the vector . I proceed to the inference step, where I compute the posterior weights and corresponding posterior quantities across all original 16,069 gene-snp pairs. In brief, the posterior mean, post covariance matrix and tissue specific tail probabilities are computed across all K components for each gene snp pair, and then weighted according to the posterior weights. This is performed in the **weightedquants** step.

We know that for a single multivariate *Normal* the posterior on $\boldsymbol{b}|U_0$ is simply:

$$\boldsymbol{b}|\hat{\boldsymbol{b}} \sim N_R(\boldsymbol{\mu}_1, U_1)$$

where:

- $\boldsymbol{\mu}_1 = U_1(\hat{V}^{-1}\hat{\boldsymbol{b}})$;

- $U_1 = (U_0^{-1} + \hat{V}^{-1})^{-1}$.

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, each now characterized by it's posterior mean $\boldsymbol{\mu}_{1k}$ and covariance $U_1 = (U_0^{-1} + \hat{V}^{-1})^{-1}$.

$$\begin{aligned} p(\boldsymbol{b}|\hat{\boldsymbol{b}}, \hat{V}, \hat{\boldsymbol{\pi}}) &= \sum_{k=1,l=1}^{K,L} \sim N_R(\boldsymbol{\mu}_{1kl}, U_{1kl}) p(z = k, l|\hat{\boldsymbol{b}}, \hat{V}, \hat{\boldsymbol{\pi}}), \\ &= \sum_{k=1,l=1}^{K,L} \sim N_R(\boldsymbol{\mu}_{1kl}, U_{1kl}) \tilde{\pi}_{k,l} \end{aligned} \tag{8}$$

Furthermore, the posterior weights or responsibilities combine hierarchical and snp-specific information, as they are proportional to the product of the original mixture weights (estimated hierarchically) and the likelihood for a particular gene-snp pair, which allows the gene-snp pair to 'find its' true match. A large likelihood in a particular component will mean that the majority of the posterior weight is at this component, and correspondingly, the majority of the posterior mean will be comprised of the posterior mean at this component which will specify a large effect at this component. If a SNP shows a relatively 'flat likelihood' at all components, than the posterior weights will appear similar to the prior weights, and correspondingly, the posterior mean will look a lot like the null case (since the prior weights are computed from 'mostly null data' and thus the prior weights will heavily weight the components with small posterior means (as determined by small prior variance in U_k).

- $\tilde{\pi_{k,l}} = P(Component|Data) \propto P(Data|Comp.) \times P(Comp.)$
- Combine hierarchical and snp-specific information
- Allows pair to find its true match!

Here, the posterior weight $\tilde{\pi}_{k,l}$ is simply

$$\tilde{\pi}_{k,l} = \frac{p(\hat{\boldsymbol{b}}_j|\hat{V}_j, z_j = k, l)\hat{\pi}_{kl}}{\sum_{k=1,l=1}^{K,L} p(\hat{\boldsymbol{b}}_j|\hat{V}_j, z_j = k, l)\hat{\pi}_{kl}} \tag{9}$$

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 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.

2 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:

$$L(\boldsymbol{\pi}; \hat{\boldsymbol{b}}) = \prod_{j=1}^{J} \sum_{k}^{K} \pi_k P(\hat{\boldsymbol{b_j}} | z_j = k) \tag{10}$$

- To estimate the hierarchical prior weights π_k: compute the likelihood at each each gene snp pair j by evaluating the probability of observing $\hat{\boldsymbol{b}}_{\boldsymbol{j}}$ 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.

Results

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.