Modeling Posterior Effects: A Continuous Journey ....

Sarah Urbut

June 18, 2015

Motivation

- Variation in gene expression is an important mechanism underlying susceptibility to complex disease.
- However, most studies to date have been conducted in a single immortalized peripheral cell type or single tissue framework
- The solution! GTEx! by 2016: 900 post-mortem donors, with approximately 30 tissues collected from each donor
- Our mission: *jointly analyze data on all tissues* to maximize power, and to identify and quantify the variability in effect sizes.

Specific Aims

- Develop statistical methods for estimating the effect size of QTLs in large numbers of diverse cell-types and tissues, thereby mapping QTLs.
- Compare among methods of estimation.
- Apply to GTEx Data

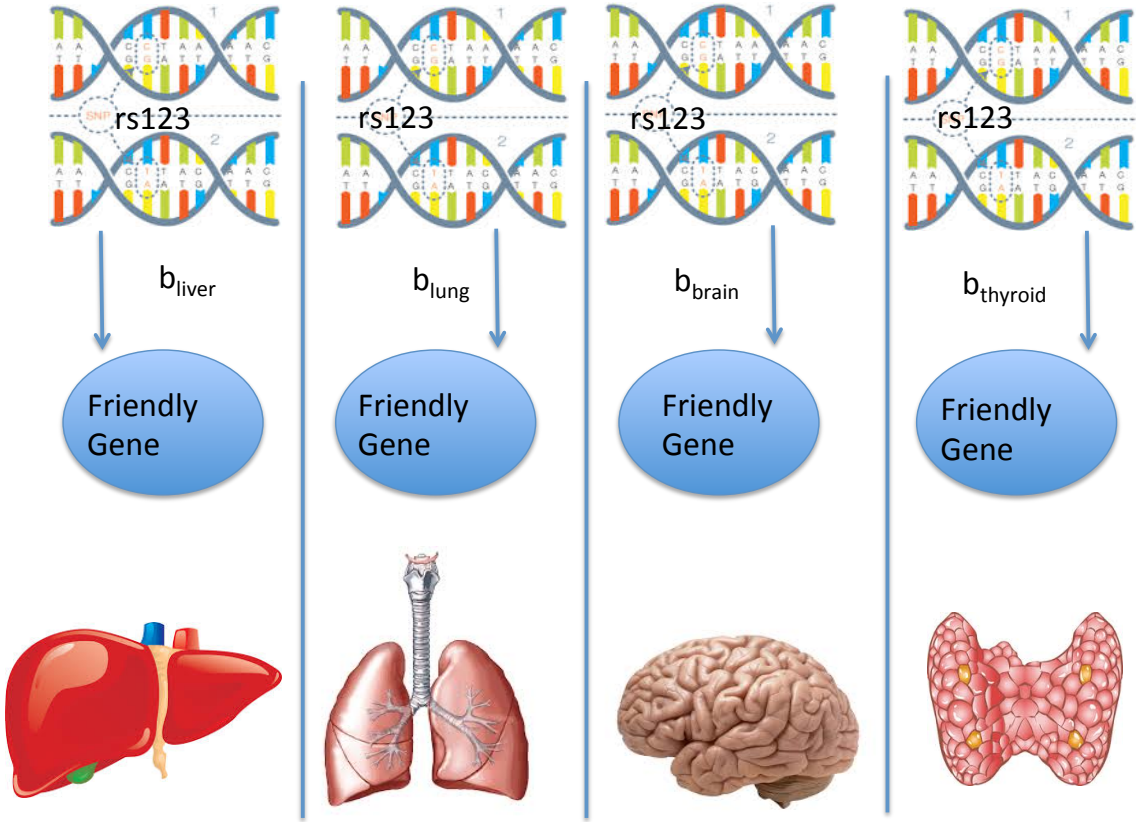
Aim 1

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'

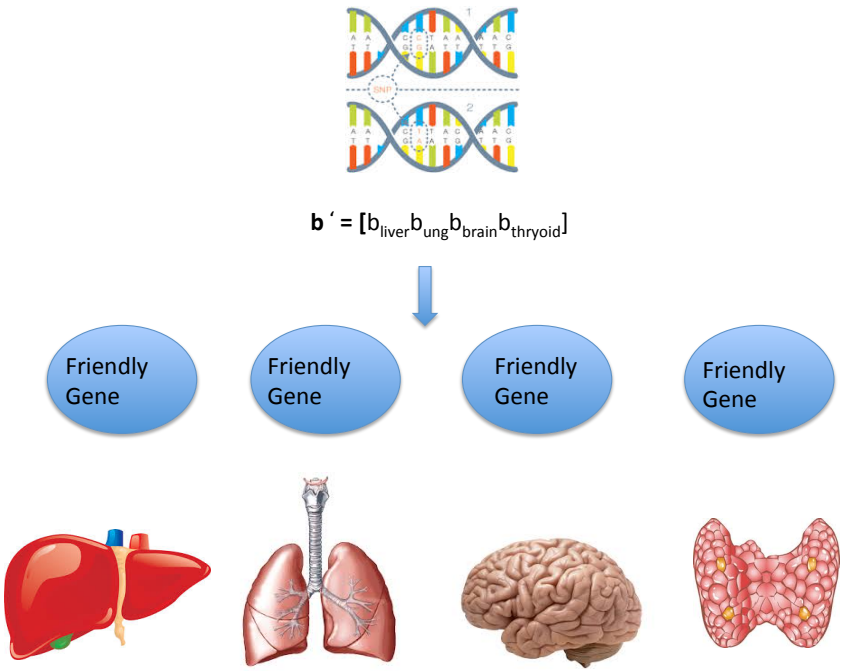
The Setting

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



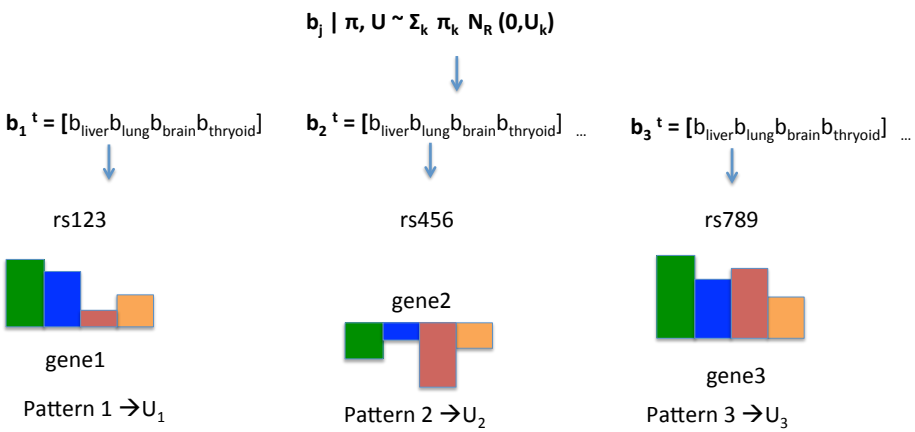
The Setting

- 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?



Considering ALL the evidence!

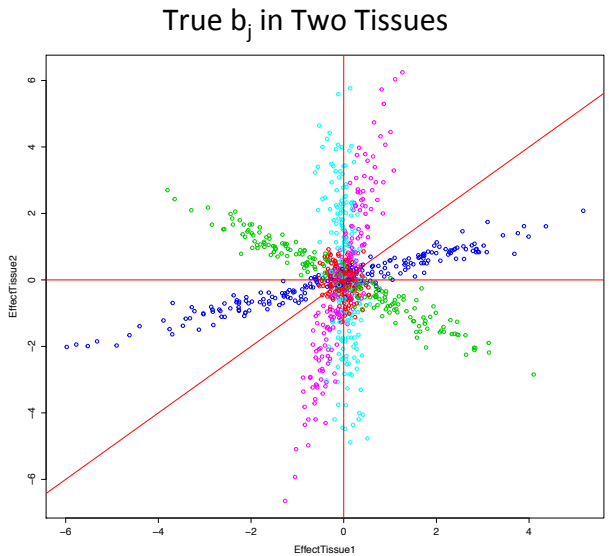
- Each eQTL may follow a particular pattern of activity
- Within these groups, the tissues exhibit characteristic patterns of sharing of effects



- 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

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 SNPs $\begin{pmatrix} Var(b_1) = 2.0 & Cov(b_1, b_2) = 0.56 \\ Cov(b_2, b_1) = 0.56 & Var(b_2) = 0.20 \end{pmatrix}$

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

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

Mixture Prior

For a given gene-snp pair, $\boldsymbol{b}$ represents the R vector of unknown standardized effect. We model the prior distribution from which $\boldsymbol{b}$ is drawn as a mixture of multivariate *Normals*.

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

- Choice of U_k determines the direction, while ω_l determines the 'stretch' or 'tails' of each distribution
- $\pi_{k,l}$ to represent the (unknown) prior weight on prior covariance matrix $U_{k,l}$
- Use the EM algorithm to estimate the optimal combination of weights: How often does this particular pattern of sharing occur in the data?

For a given gene-snp pair: Likelihood for $\boldsymbol{b}$

- $\hat{\boldsymbol{b}}$: R x 1 vector of standardized maximum likelihood estimates of effect sizes
- $\hat{V} \approx Var(\hat{\boldsymbol{b}})$: RxR diagonal 'accurate' approximation of standardized standard error of $\hat{\boldsymbol{b}}$.
- $\hat{\boldsymbol{b}}$ and $\hat{V}$ as *observations* (i.e. known)

For a given gene-snp pair, the Likelihood on $\boldsymbol{b}$:

$$\hat{\boldsymbol{b}}|\boldsymbol{b} \sim N_R(\boldsymbol{b}, \hat{V}) \tag{2}$$

For a given gene-snp pair: Posterior on $\boldsymbol{b}$

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:

- $\boldsymbol{\mu}_1 = U_1(\hat{V}^{-1}\hat{\boldsymbol{b}})$;
- $U_1 = (U_0^{-1} + \hat{V}^{-1})^{-1}$.

For a given gene-snp pair: Posterior on b

- Normal prior+ Normal likelihood = Normal posterior
- Mix normal prior + normal likelihood = Mix normal posterior

$$
\begin{aligned}
p(\boldsymbol{b}_{|}\hat{\boldsymbol{b}}, \hat{V}, \hat{\pi}) &= \sum_{k=1,l=1}^{K,L} \sim N_R(\boldsymbol{\mu}_{1kl}, U_{1kl}) p(z = k, l|\hat{\boldsymbol{b}}, \hat{V}, \hat{\pi}), \\
&= \sum_{k=1,l=1}^{K,L} \sim N_R(\boldsymbol{\mu}_{1kl}, U_{1kl}) \tilde{\pi}_{k,l}
\end{aligned}
\tag{3}
$$

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

The key: Mixture!

- But how do you choose the set of all **U** ?
- Goal: Capture all the patterns of sharing in the data
- If we knew the truth, choose a set that contains for SNPs of each class:

$$U_k = \begin{pmatrix} Var(b_1) & \cdots & Cov(b_1, b_r) \\ \vdots & \ddots & \vdots \\ Cov(b_r, b_1) & \cdots & Var(b_r) \end{pmatrix} \tag{4}$$

- Look to the data for hints ...

Choice of Covariance Matrices

For a given ω_l, we specify 5 'types' of RxR prior covariance matrices $U_{k,l}$.

- $U_{k=1,l} = \omega_l \ \mathbf{I}_R$
- $U_{k=2,l} = \omega_l \ \frac{1}{J} \ X_t^t \ X_t$ The (naively) estimated tissue covariance matrix as estimated from the column-centered J $\times R$ matrix of t statistics,
- $U_{k=3,l} = \omega_l \ \frac{1}{J} \ \mathrm{V}_{1\ldots p} \ d^2_{1\ldots p} \ V^t_{1..p}$
- $U_{k=4:4+Q-1,l} = \frac{1}{J}[(\Lambda\mathbf{F})^t \Lambda\mathbf{F}]_q$
- $U_{k=4+Q,l} = \frac{1}{J} \ (\Lambda\mathbf{F})^t \Lambda\mathbf{F}$
- Λ represent the J x Q matrix of loadings
- F is then the KxR matrix of factors.
- Sparse Prior on Λ such that each SNP can be a member of a minimal number of factor classes

Incorporating Tissue Specific and Tissue Consistent Effects

- $U_{k=5+Q:R+4+Q,I} = \frac{1}{J}$ ([100..]'[100...]
- $U_{k=R+5+Q,I} = \frac{1}{J}$ ([111...]'[111...]
- [1000...] or [111...] represent configurations such that given membership, $\boldsymbol{b_j}$ arise from the same prior variance.

$$U_0 \mid \omega = \begin{pmatrix} \omega^2 & \cdots & \omega^2 \\ \vdots & \ddots & \vdots \\ \omega^2 & \cdots & \omega^2 \end{pmatrix} \tag{5}$$

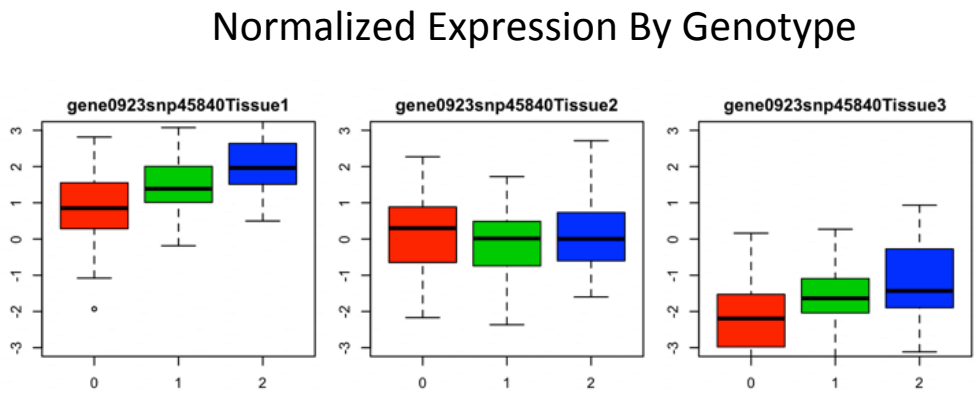
New Metrics: LFSR For a given gene-snp pair

- Compare the effect size among tissues in which the snp is called active, and among snps active in a particular tissues
- Accurate estimates of the true effect size $\boldsymbol{b}$ increase power
- Succinct comparison: Local False Sign Rate ($LFSR$)

$$\begin{aligned} LFSR\text{=}1 \text{ - } \max(\mathrm{P}(\mathrm{b}_r > 0|Data), P(b_r < 0|Data)) \\ = 1 - max[\sum_{kxl} p(b_r > 0|Data)\tilde{\pi}_{k,l}, \\ \sum_{kxl} p(b_r < 0|Data)\tilde{\pi}_{k,l}] \end{aligned} \tag{6}$$

- Posterior probability of incorrectly identifying the direction of effect.
- When the effect is large and the error is small ...
- $LFSR$ incorporates **tissue-specific quantitative** information through the tail-probability on the effect in a particular tissue.

A very simple example



True b	0.701	-0.007	0.610
LFSR	0.17e-15	0.47	1.36e-15

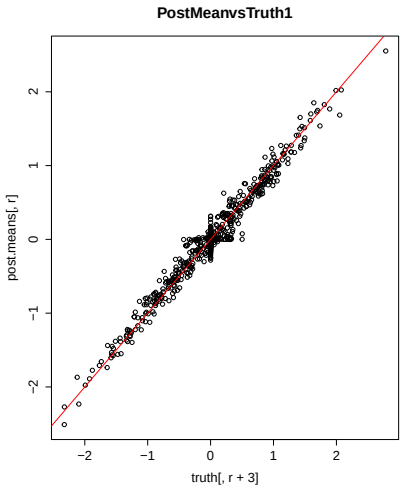
- Confident that a SNP acts in a tissue because our prior belief in the 'sharing' is high.
- Effect size might be relatively modest in tissue 2 for example.
- Reflected our relatively high probability of falsely assigning the direction of the effect
- This translates continuous outcome into a binary one.
- More resolution when compared with qualitative local *FDR*

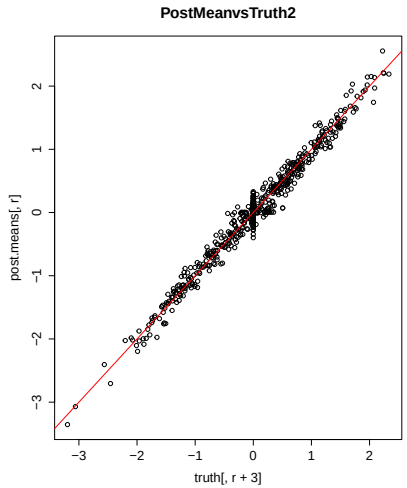
Aim 2: Comparing Among Methods

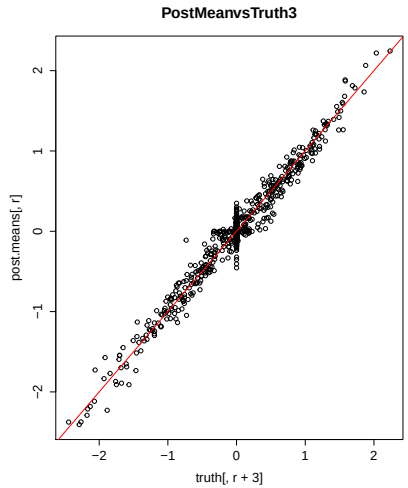
- Reduce Mean Squared Error in Simulations
- Maximize observed data likelihood in cross-validated data

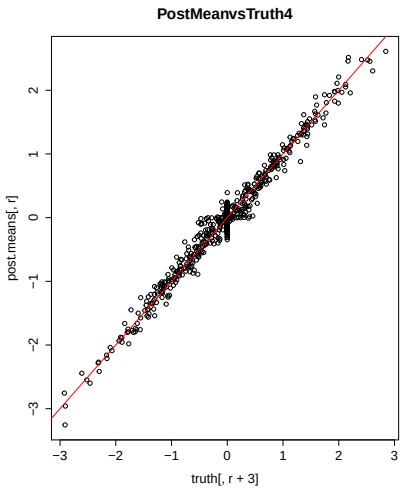
Simulations First ...

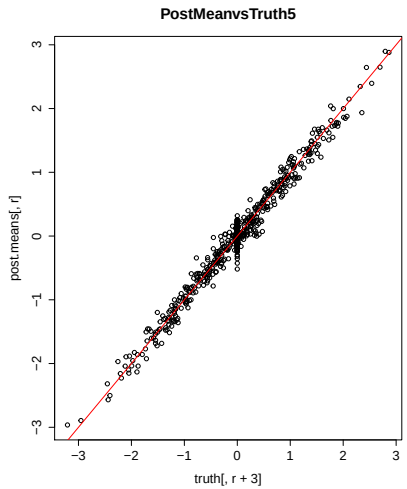
$$\text{MSE} = \frac{1}{J}\sum_{i=1}^{n}(\boldsymbol{b}_{jtruth} - E(\boldsymbol{b}|\hat{\boldsymbol{b}}_j)^2) = 4.57\text{e-}4$$



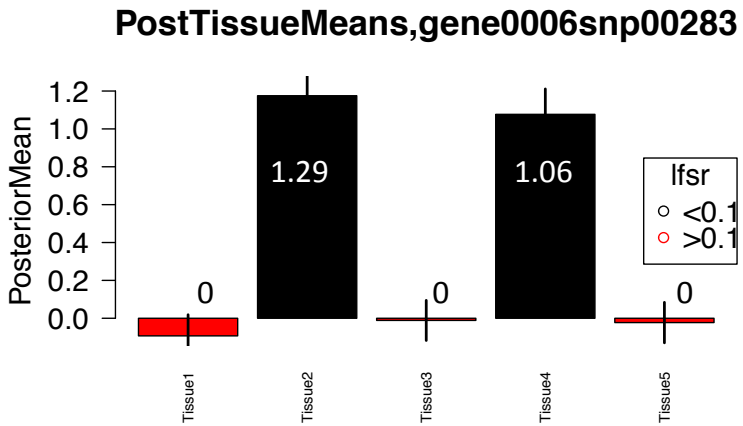






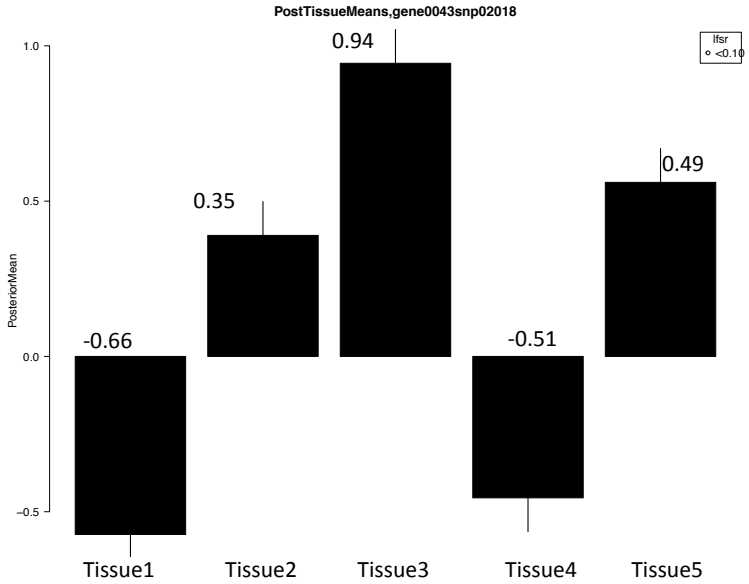


The *LFSR* at work: Nimbleness in the Making



- Detect associations that are truly active in only real tissue and accurately assign an elevated local false sign rate to the alternative off tissues
- Use of an *lfsr* threshold detects the null effect in tissues 1,3 and 5

Can we actually accurately predict differences in sign?



- Confidently capture differences in the direction of effect with ample power.
- Subtle differences in direction captured by covariance structure

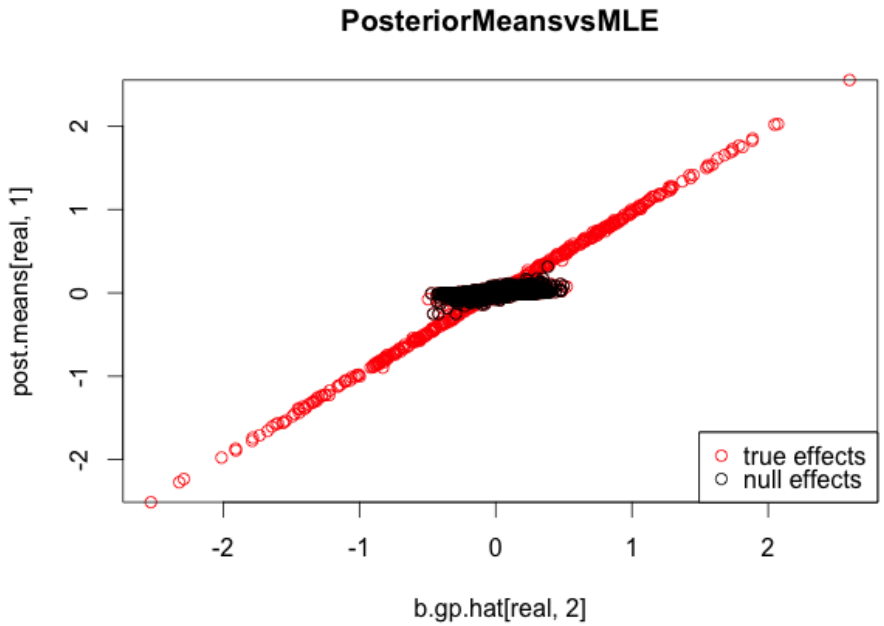


Figure: **Simulated Gene-SNP Pair**

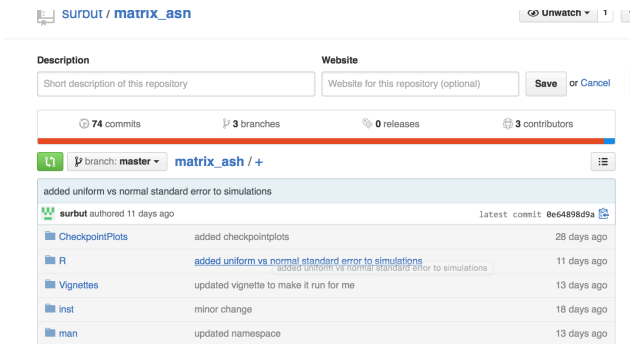
For gene-snp pairs we know to be inactive, we plot the posterior mean that results from a 336 component mixture prior that included 12 SFA approximations on $\boldsymbol{b}_j$ and see that it appears to serve as a shrunken approximation of the MLE for null SNPs and an accurate approximation for real SNPs.

Evaluating A Method: But how do we compare?

- We can use the π_k as computed from a training data set of 1000 gene SNP pairs
- Compute log of the marginal likelihood for each 'test' gene-snp pair
- We then sum the log likelihood for test set to compare the log likelihoods between two methods - a method which captures the data and thus reflects accurate hierarchical weights well will maximize the likelihood of the test set.
- With simulated data, we can estimate the likelihood for the 'true prior' using true mixture proportions

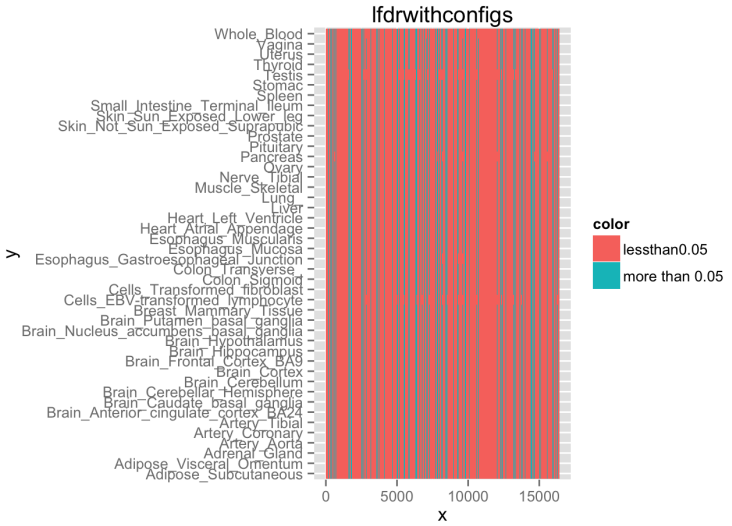
$$\ln \mathcal{L}(\pi\,;, \hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_j) = \sum_{j=1}^{J} \ln \sum_{k=1}^{K} \mathbf{p}(\hat{\mathbf{b}}_{j.\mathbf{test}} | \pi, \mathbf{z} = \mathbf{k}) \mathbf{p}_{\mathbf{train}}(\mathbf{z} = \mathbf{k}) \qquad (7)$$

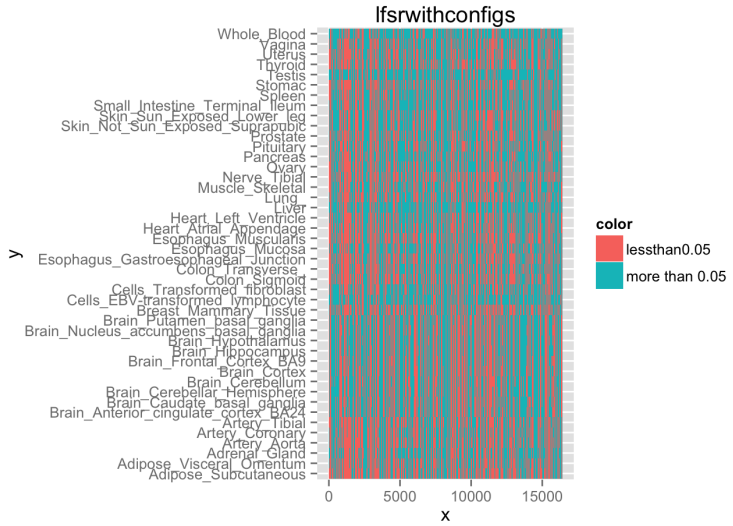
Evaluating A Method: But how do we compare?



- SFA modest improvement
- Grid: Go big or go home - a wide array of effect sizes to capture many 'tails' or 'stretches'
- Covariance Matrices on real (or strong) estimated effects
- Add singleton and 'ALL' configurations
- Ultimately: 5 SFA factors, 44 + 1 configs and 23 element grid
- Iterative shrinkage ...

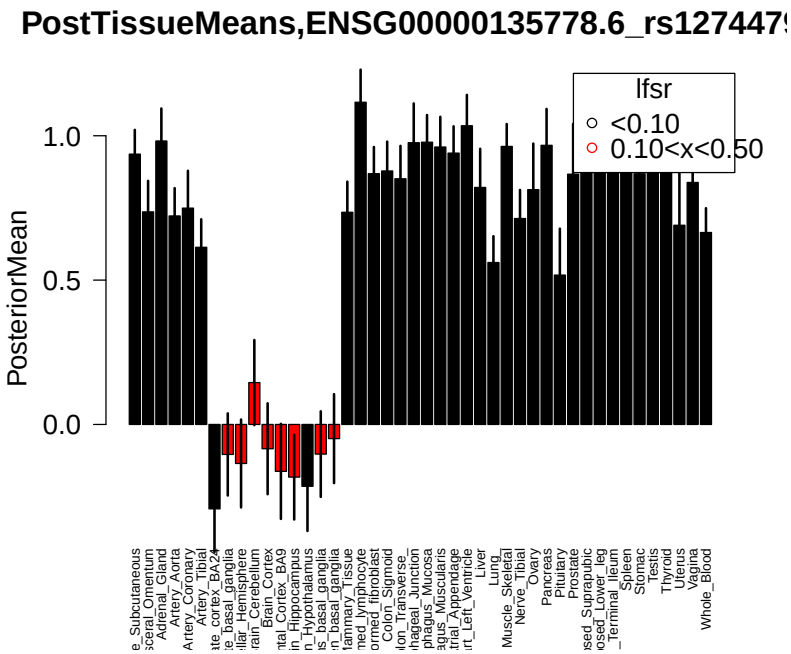
Aim 3: Getting to the real data!

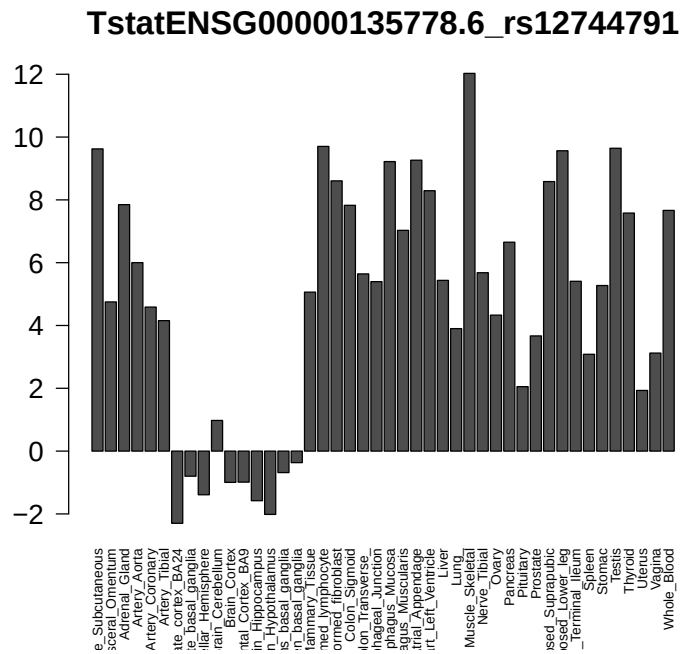




- Probability of being 'on' no longer of interest
- How confident are we in the sign of the effect?
- Power: at a given LFSR ...
- Best method yet: 5 SFA factors and 23 element grid

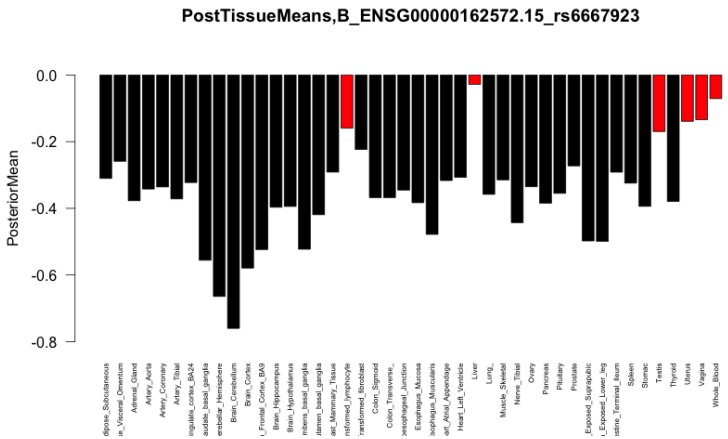
Getting to the real data

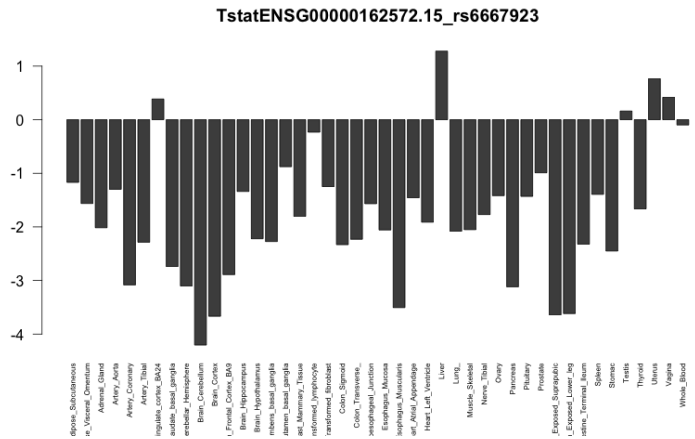




- Highly suspect 'off direction' caught redhanded
- Accurately recapitulates strong signals

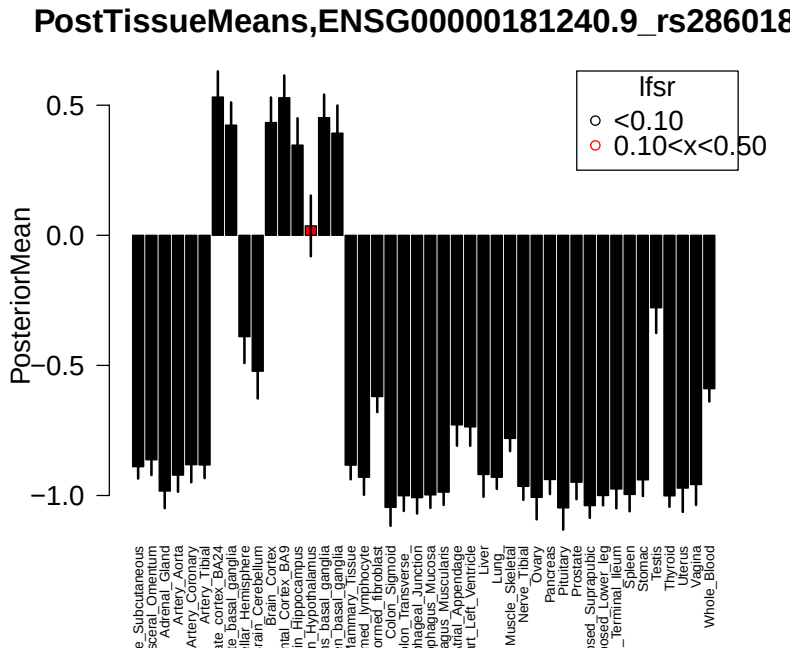
Getting to the real data?





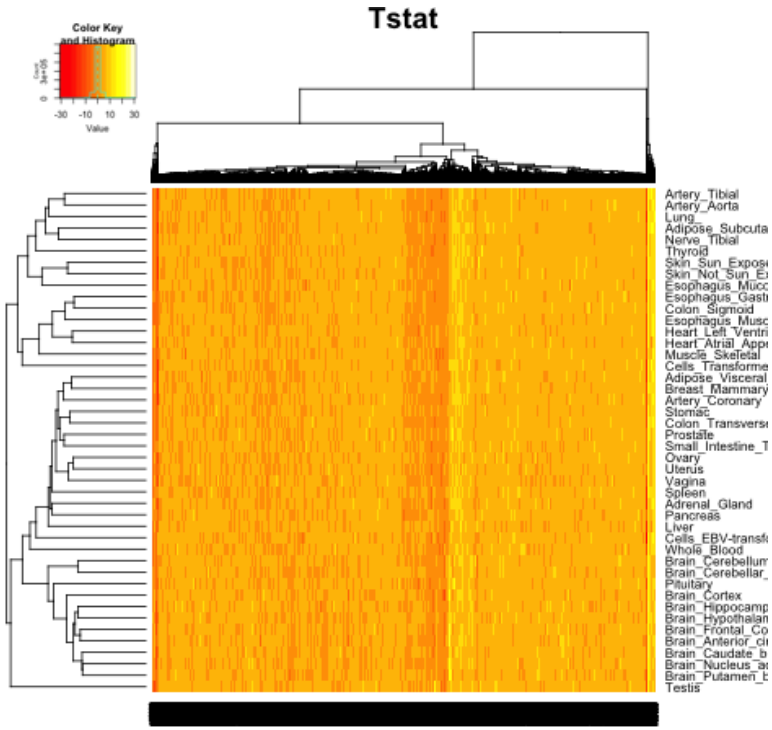
- Fills in for potentially errant tibial artery and liver
- Identifies signals in which we lack confidence in direction or magnitude

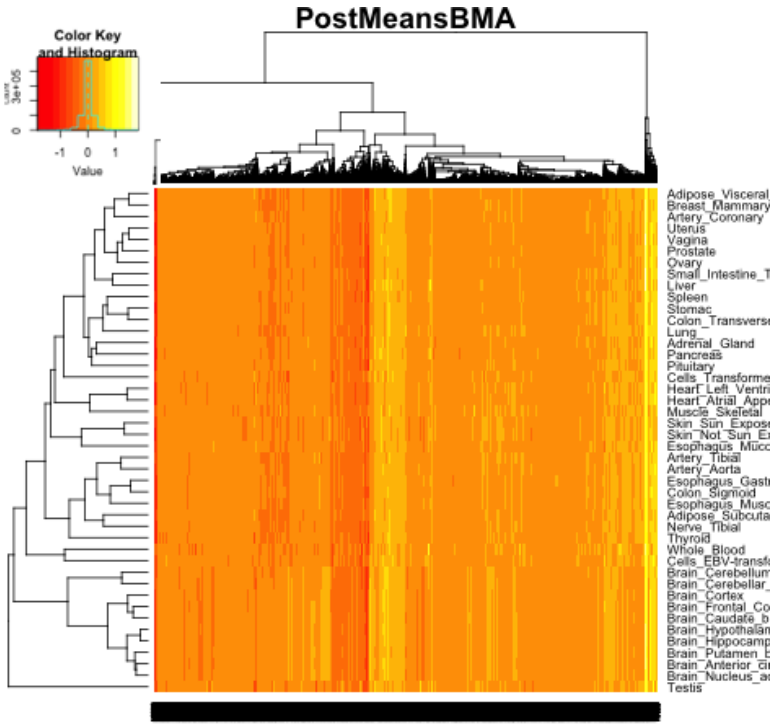
But what about effects in different directions?



- Two SNPs show a significant effect in both tissues, but with effects in the opposite direction
- SNP by SNP analysis, looks like opposite effects

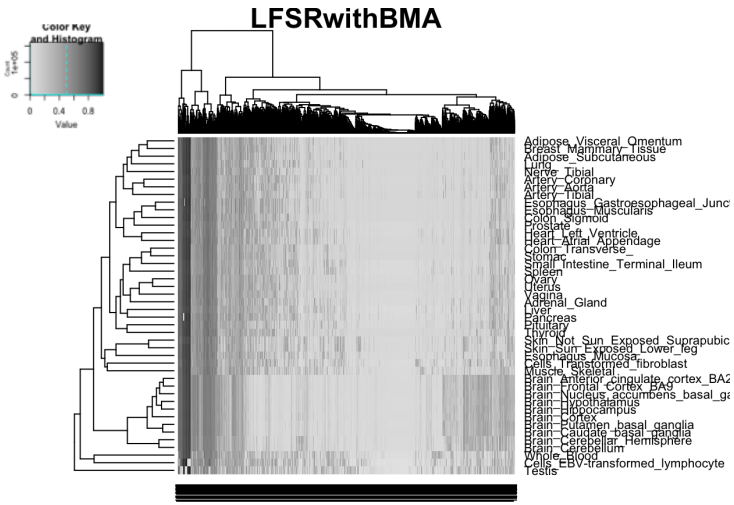
On a grand scale?

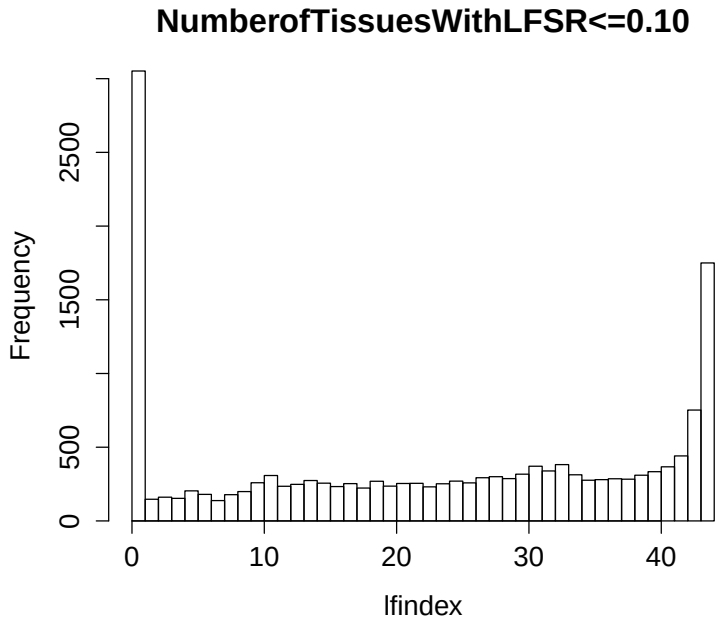




- 'Shrinks' the noise and emphasizes the strong signals

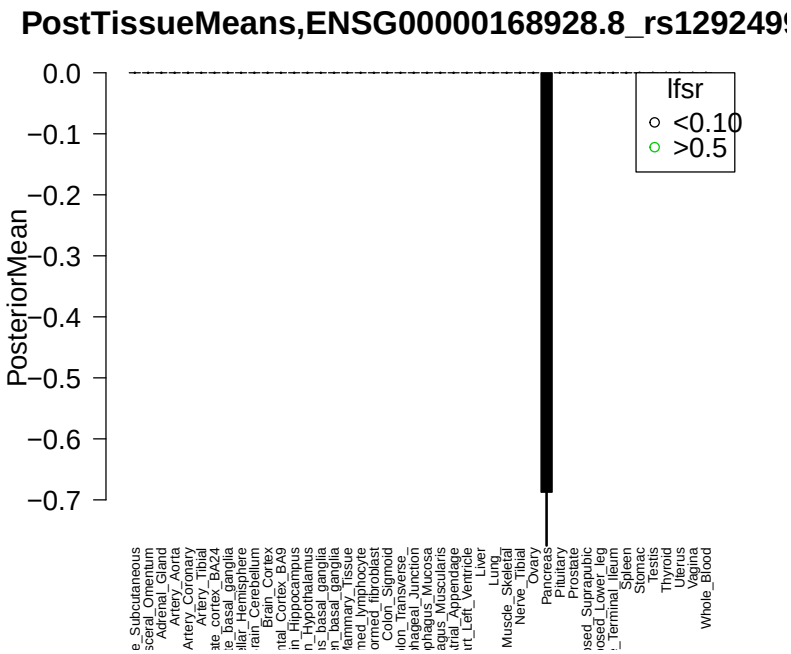
So what's the biology behind tissues with similar lfsr?

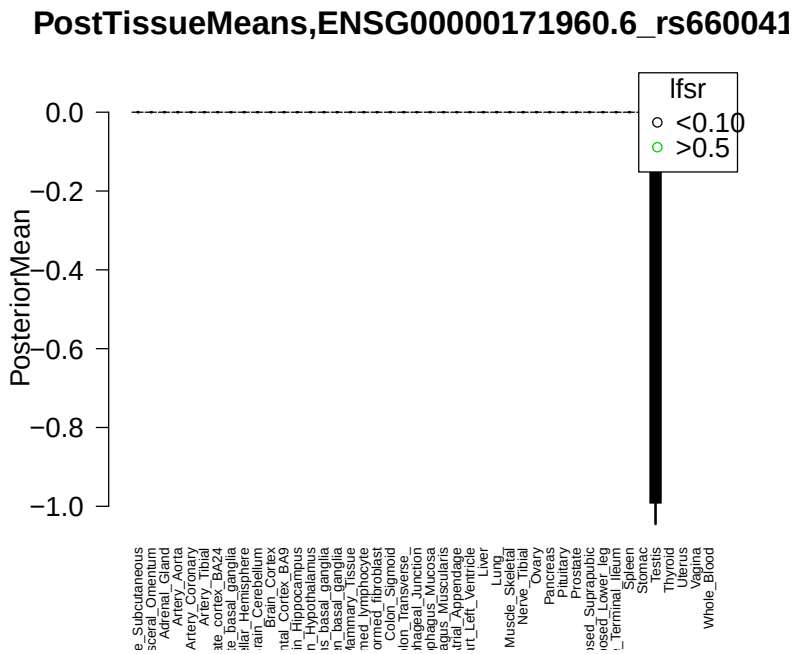




- Brains and Guts cluster together
- True distribution represents a mixture of tissue specific, shared and heterogenous effects

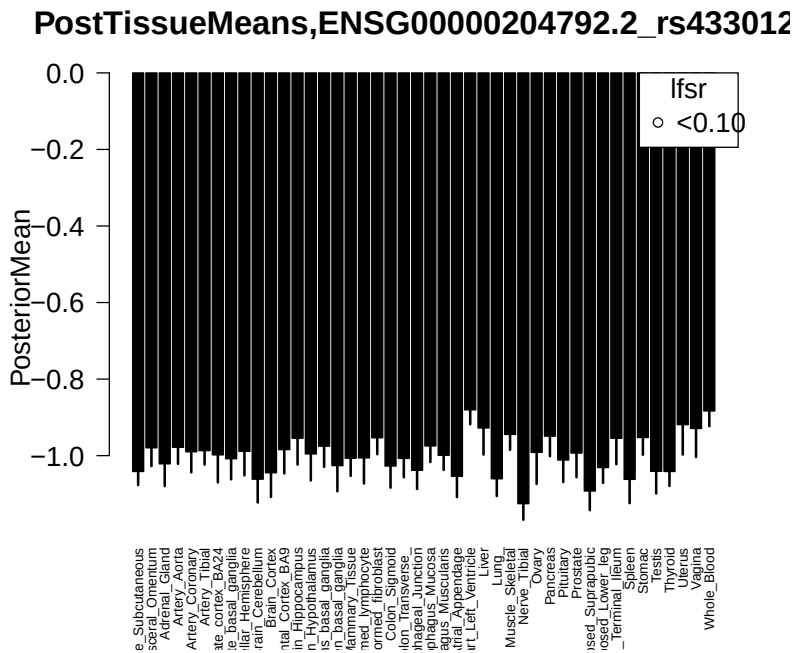
Tissue Specific Effects





What about Consistent effects

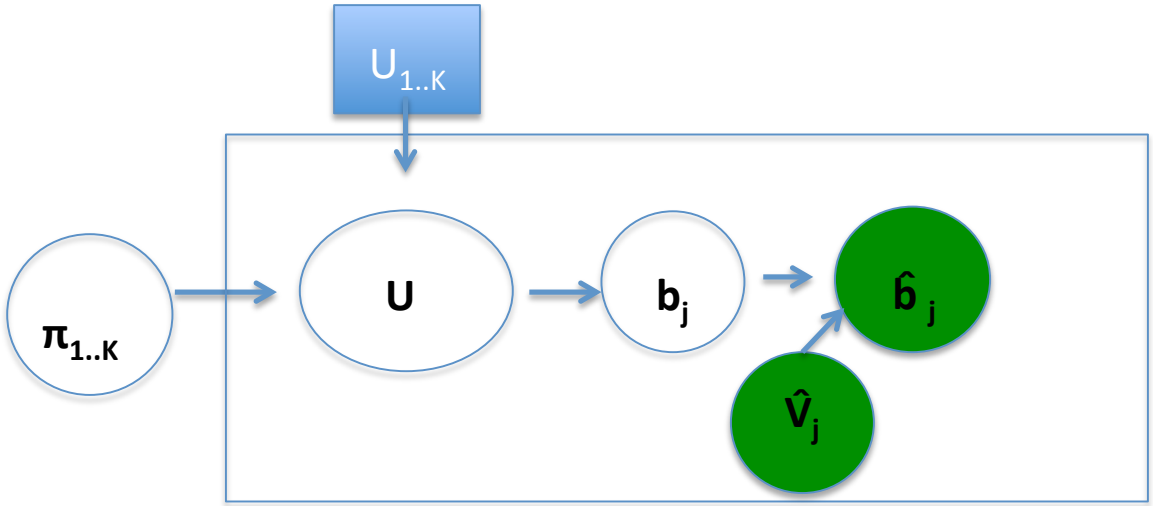
- Pilot Data - 50% of SNPs 'active in all tissues'
- Tends to push SNPs to similar effect size in all tissues



Thank you so much for listening!

Appendix

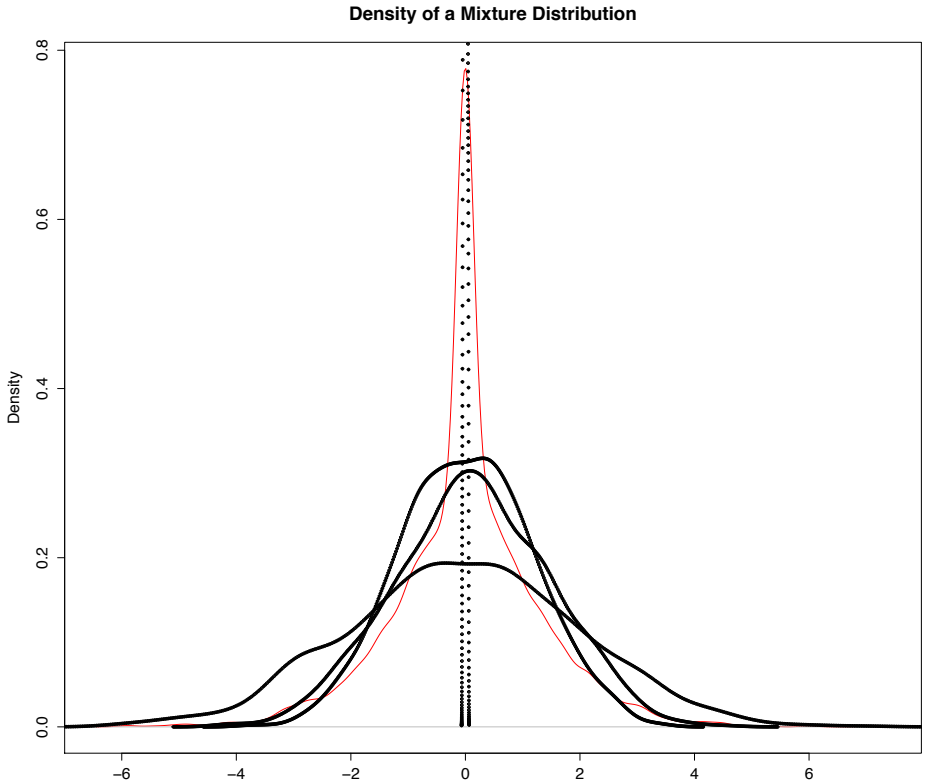
If we knew the truth ...



We assume that each observed $\hat{\boldsymbol{b}}_j$ arises from some $\boldsymbol{b}_j$ which in turn arises from a specific multivariate distribution defined by it's prior covariance matrix U_k. The proportion of gene-snps pairs in each class, π_k is estimated hierarchically.

Because we can't know the true component, we model $\boldsymbol{b}_j$ as arising from a mixture that captures all of the covariance matrices.

For one tissue ...



Estimating Hierarchical Weights

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

Incomplete Data likelihood:

$$L(\boldsymbol{\pi}; \hat{\boldsymbol{b}}) = \prod_{j=1}^{J} \sum_{k}^{K} \pi_k P(\hat{\boldsymbol{b}}_{\boldsymbol{j}} | z_j = k) \tag{8}$$

- 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 $\boldsymbol{b}_{\boldsymbol{j}}$ 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?

Potential

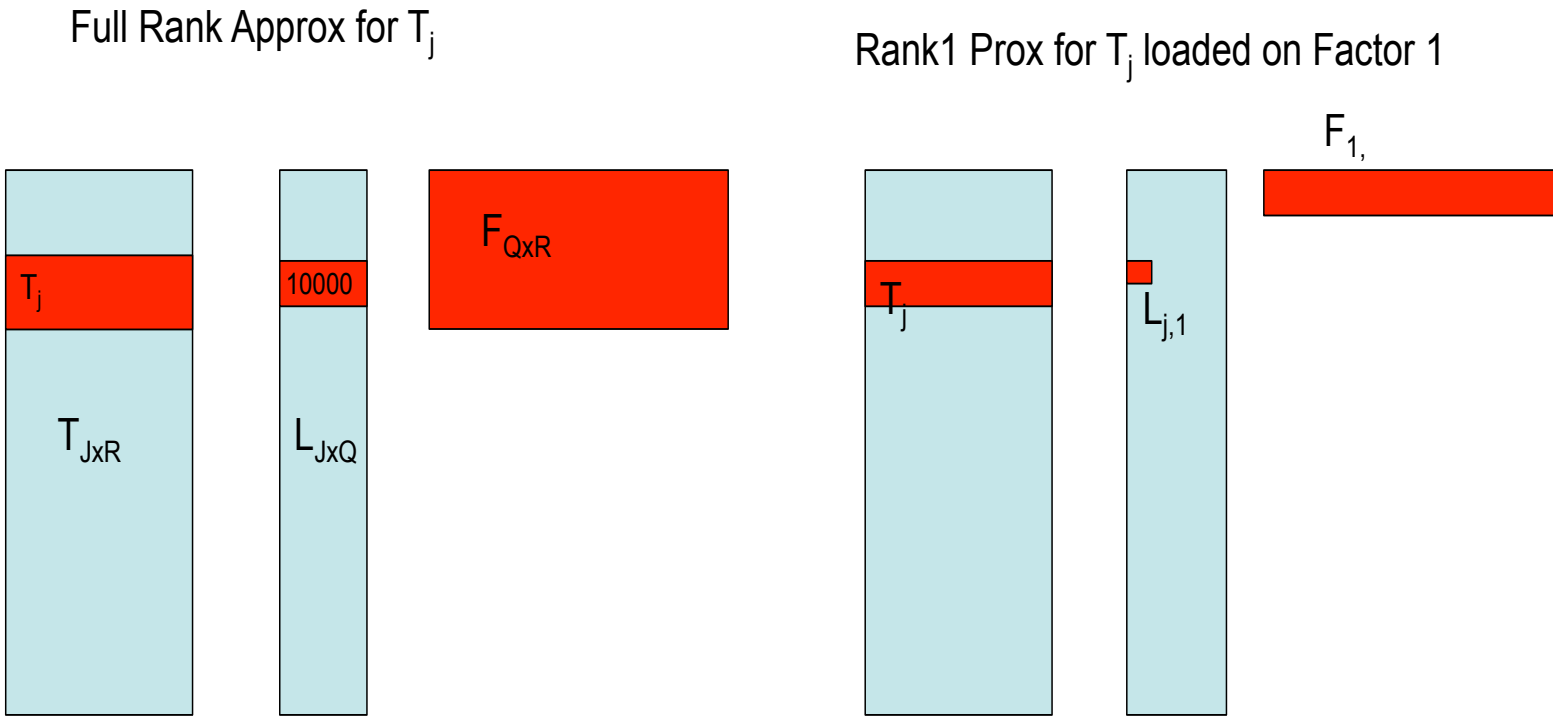
By contrast, the local false discovery rate or *LFDR* simply averages the marginal likelihood over all components in which a tissue is inactive, thus representing the posterior probability that a SNP is inactive given its level of significance.

$$
\begin{aligned}
LFDR{=}\mathrm{P}(\mathrm{b}_r = 0|data) & \\
& = 1 - \sum_{kxl} p(b_{jr} \neq 0|Data)\tilde{\pi}_{k,l} \\
& = 1 - \sum_{kxl, U_{r,r} \neq 0} \tilde{\pi}_{k,l}
\end{aligned}
\tag{9}
$$

- but reconsider $\tilde{\pi}_{k,l}$
- Large emphasis on prior weights so if the prior weight on any 'shared' pattern is high, the *lfdr* will be low
- $\tilde{\pi}_{k,l}$ augments modest evidence in tissue r with information from all tissues

Understanding SFA

A brief simulatory interlude ....



A Closer Look

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

$$\tilde{\pi}_{k,l} = \frac{p(\hat{\boldsymbol{b}}_j | \hat{V}_j, z_j = k, l)\hat{\pi}_{kl}}{\sum_{k=1,l=1}^{K,L} p(\hat{\boldsymbol{b}}_j | \hat{V}_j, z_j = k, l)\hat{\pi}_{kl}} \tag{10}$$

- $\hat{\pi}_{kl}$ represents the prior weights which are estimated hierarchically
- But $p(\hat{\boldsymbol{b}}_j | \hat{V}_j, z_j = k, l)$ considers the marginal probability combining evidence over all tissues!
- Cheat! $N_R(\hat{\mathbf{b}}_j; \mathbf{0}, U_k + \hat{V}_j)$

A Closer Look

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:

$$f(x \mid \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \, e^{-\frac{(x)^2}{2\sigma^2}} \tag{11}$$

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'.

Furthermore, now we know that the posterior mean is $\boldsymbol{\mu}_1 = U_1(\hat{V}^{-1}\hat{\boldsymbol{b}})$ where $U_1 = (U_0^{-1} + \hat{V}^{-1})^{-1}$ and so quite obviously, the posterior mean for a particular tissue in the components with a large prior variance will also be large because they are 'very roughly' $\propto U_{0k}(\hat{V}^{-1}\hat{\boldsymbol{b}})$.

A Closer Look

Putting these things together, we see that 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).

Breaking It Down

But where did that come from Sarah?

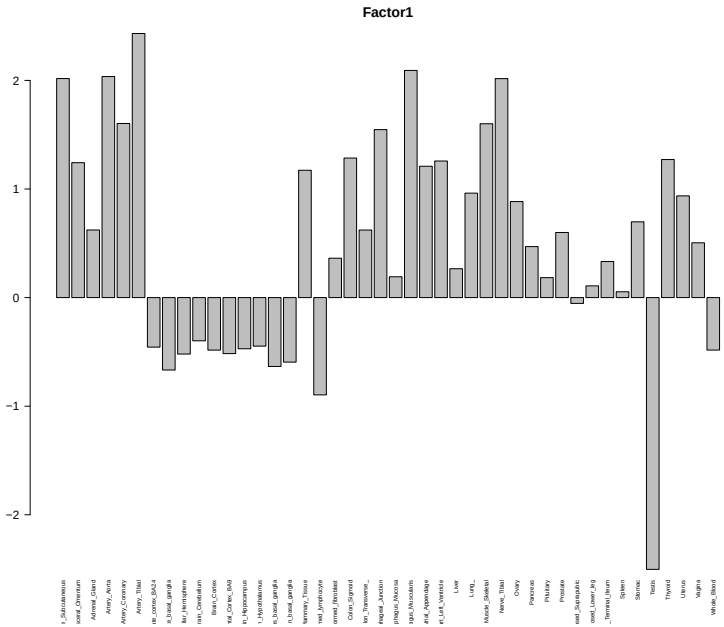
For a given gene snp pair j,

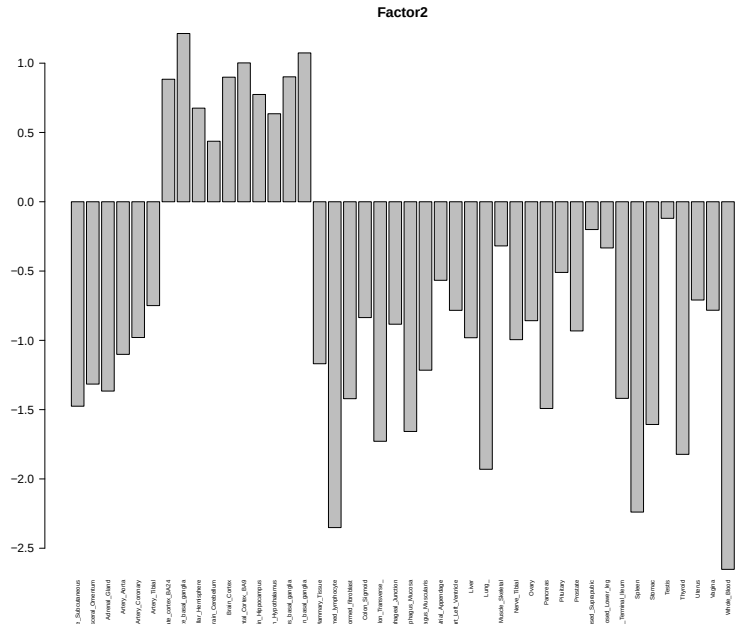
$$E(\mathbf{T_j}|\lambda_\mathbf{j}) = \lambda_\mathbf{j}\mathbf{F}$$

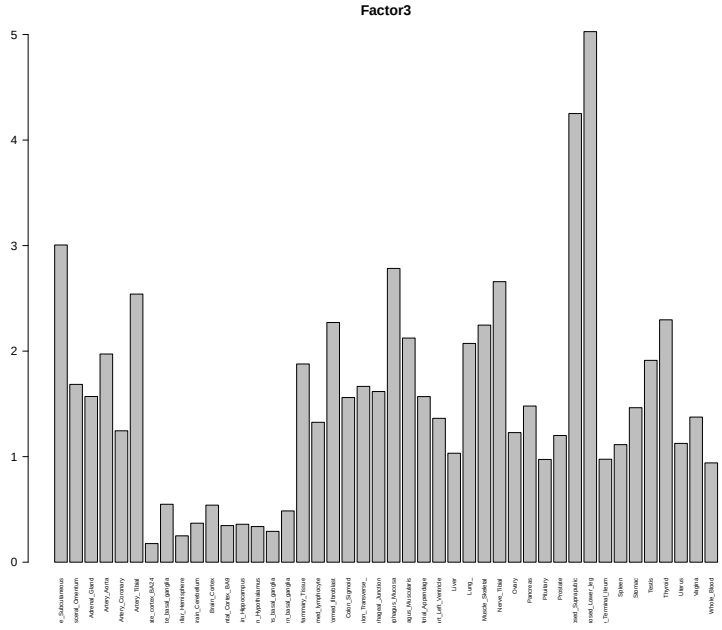
$$\begin{aligned} \lambda_j &\sim N(0, \Sigma_j) \\ \mathbf{T}_j &\sim N(0, F'\Sigma_j F) \end{aligned} \tag{12}$$

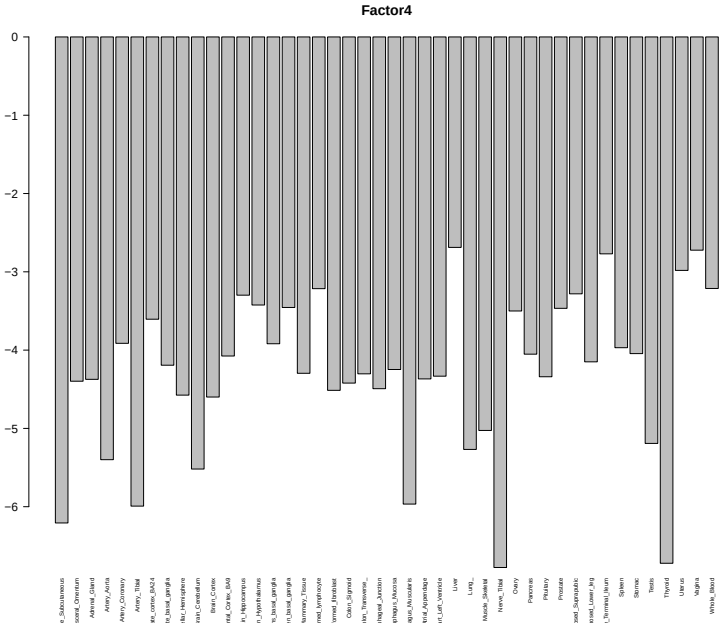
- Where $\mathbf{T_j}$ represents an R vector of T statistics for a given individual and λ represents a K-vector of loadings for a particular SNP
- Think of $L'L$ as an approximation for the 'next' Σ_j.

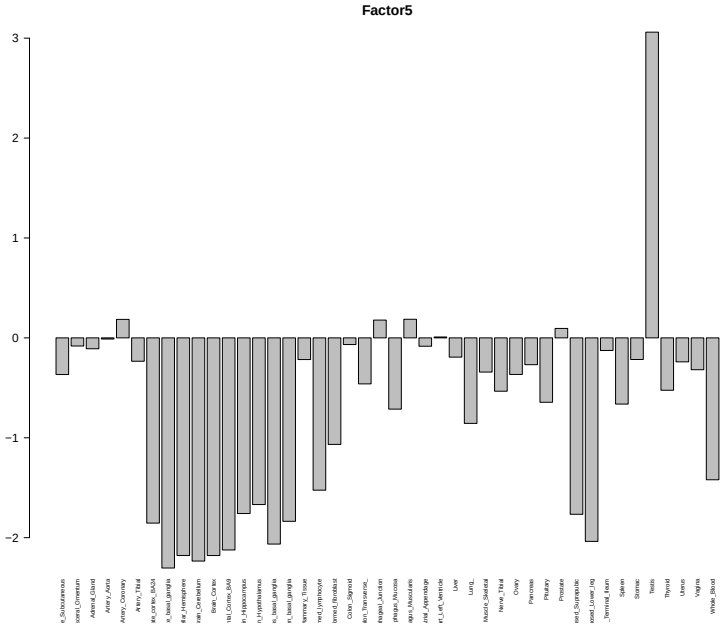
Understanding SFA: Recapitulating the patterns in the T statistics



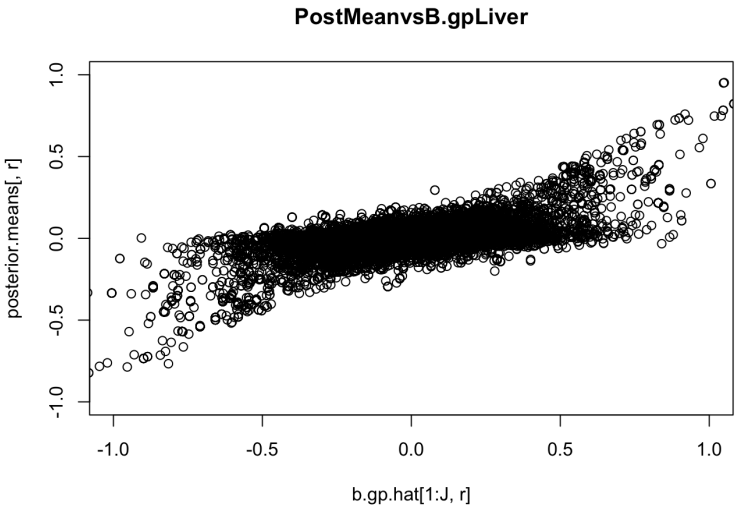


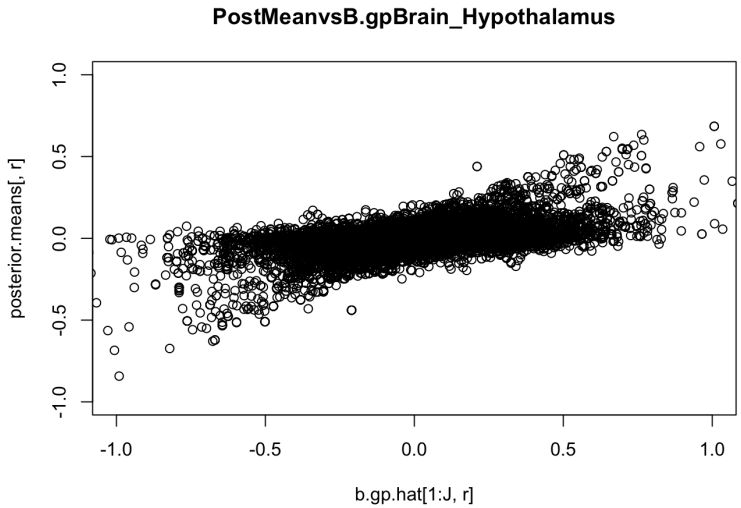

Factor3
5
4
3
2
1
0
e_Subcutaneous
scend_Omentum
Adrenal_Gland
Artery_Aorta
Artery_Coronary
Artery_Tibial
ex_corex_BA24
te_basal_ganglia
ital_Hemisphere
rain_Cerebellum
Brain_Cortex
ntal_Cortex_BA9
n_Hippocampus
n_Hypothalamus
te_basal_ganglia
n_basal_ganglia
Mammary_Tissue
med_Lymphocyte
ormed_Fibroblast
Colon_Sigmoid
ion_Transverse_
Vaginal_Junction
ginglis_Mucosa
agus_Muscularis
ntal_Appendage
rt_Left_Ventricle
Liver
Lung_
Muscle_Skeletal
Nerve_Tibial
Ovary
Pancreas
Pituitary
Prostate
ased_Suprapubic
ernal_Lower_Leg
l_Terminal_Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
White_Blood


Factor4
0
−1
−2
−3
−4
−5
−6
e_Subcutaneous
scend_Omentum
Adrenal_Gland
Artery_Aorta
Artery_Coronary
Artery_Tibial
ex_corex_BA24
te_basal_ganglia
ital_Hemisphere
rain_Cerebellum
Brain_Cortex
ntal_Cortex_BA9
n_Hippocampus
n_Hypothalamus
te_basal_ganglia
n_basal_ganglia
Mammary_Tissue
med_Lymphocyte
ormed_Fibroblast
Colon_Sigmoid
ion_Transverse_
Vaginal_Junction
ginglis_Mucosa
agus_Muscularis
ntal_Appendage
rt_Left_Ventricle
Liver
Lung_
Muscle_Skeletal
Nerve_Tibial
Ovary
Pancreas
Pituitary
Prostate
ased_Suprapubic
ernal_Lower_Leg
l_Terminal_Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
White_Blood


Factor5
3
2
1
0
−1
−2
e_Subcutaneous
scend_Omentum
Adrenal_Gland
Artery_Aorta
Artery_Coronary
Artery_Tibial
ex_corex_BA24
te_basal_ganglia
ital_Hemisphere
rain_Cerebellum
Brain_Cortex
ntal_Cortex_BA9
n_Hippocampus
n_Hypothalamus
te_basal_ganglia
n_basal_ganglia
Mammary_Tissue
med_Lymphocyte
ormed_Fibroblast
Colon_Sigmoid
ion_Transverse_
Vaginal_Junction
ginglis_Mucosa
agus_Muscularis
ntal_Appendage
rt_Left_Ventricle
Liver
Lung_
Muscle_Skeletal
Nerve_Tibial
Ovary
Pancreas
Pituitary
Prostate
ased_Suprapubic
ernal_Lower_Leg
l_Terminal_Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
White_Blood

Yeah .... but do you shrink?





- Significant shrinkage of MLE - expected with random sampling of gene-snp pairs
- Avoids shrinking strongest t statistics

The MLE for π can be obtained by using the following EM algorithm. Let $D_i = (\hat{\boldsymbol{b}}, \hat{V})$ and $D = (D_1, \ldots, D_n)$. Consider unobserved latent variables $Z = (Z_1, \ldots, Z_n)$, where $Z_i \in \{1, \ldots, M\}$ and $\Pr(Z_i = m) = \pi_m$. Then, a complete data likelihood can be written

$$\begin{aligned}
\Pr(D, Z \mid \pi) &= \prod_{i=1}^{n} \Pr(D_i, Z_i \mid \pi) && (13)\\
&= \prod_{i=1}^{n} \prod_{m=1}^{M} \Pr(D_i, Z_i = m \mid \pi)^{I(Z_i = m)} && (14)\\
&= \prod_{i=1}^{n} \prod_{m=1}^{M} [\Pr(D_i \mid Z_i = m, \pi)\pi_m]^{I(Z_i = m)}, && (15)
\end{aligned}$$

yielding a log likelihood

$$\log \Pr(D, Z \mid \pi) = \sum_{i=1}^{n} \sum_{m=1}^{M} I(Z_i = m) \Big[\log \Pr(D_i \mid Z_i = m) + \log \pi_m \Big] \quad (16)$$

We denote by π^l the vector of probabilities at step l of the EM algorithm. In each step, we update the vector, i.e. compute in the l-th step π^{l+1} from π^l and the data.

E-step: For each i and m,

$$A_{im}^l \equiv \Pr(Z_i = m \mid D_i, \pi^l) = \frac{\Pr(Z_i = m, D_i \mid \pi^l)}{\sum_{n=1}^M \Pr(Z_i = n, D_i \mid \pi^l)} \quad (17)$$

$$= \frac{\pi_m^l \Pr(D_i \mid Z_i = m)}{\sum_{n=1}^M \pi_n^l \Pr(D_i \mid Z_i = n)} \quad (18)$$

$Pr(D_i \mid Z_i = m) = N_R(\hat{\mathbf{b}}_j; \mathbf{0}, U_k + \hat{V}_j)$

M-step: Find the parameters π which maximizes $\mathrm{E}_{Z|D,\pi^l}[\log \Pr(D, Z \mid \pi)]$.

$$\pi^{l+1} = \underset{\pi}{\operatorname{argmax}}\, \mathrm{E}_{Z|D,\pi^l}[\log \Pr(D, Z \mid \pi)] \quad (19)$$

and for each $m = 1, \ldots, M-1$,

$$\pi_m = \frac{\sum_{i=1}^n A_{im}^l}{n}. \quad (20)$$