Comparing multiple conditions with the same reference level using MASH

Yuxin Zou, Sarah Margaret Urbut, Matthew Stephens

November 9, 2018

Abstract

When we estimate and compare the gene expression in multiple conditions withe the same reference level, we want to estimate them jointly using mash to gain more power. But failure to deal with the induced inherent correlation among deviations causes many false discoveries. Here, we describe a method to include the inherent correlations and thus reduce the number of false discoveries dramatically.

1 Introduction

The multivariate adaptive shrinkage, mash, method could estimate and compare many effects across multiple conditions jointly (Urbut et al., 2018). This flexible approach can be used to identify "significant" non-zero effects and compare effects, identify differences in effect among conditions. For example, in eQTL (expression Quantitative Trait Locus) studies, researchers are often interested in identifying tissue-specific effects, in the belief that they may have particular biological relevance.

However, in some genomic studies, there might be no obvious "effect" in each condition. We want to estimate the change in some quantity computed in multiple conditions over a common reference level. Such analyses are common in practice in case and control literature. Scientists are often interested in identifying patterns of differential gene regulation under many conditions, compared to a condition in which there exists no external stimuli. In these cases, differential expression in any condition is defined as a difference in expression over a common control condition. In other cases, there might be no control group in a study. The differential expression in any condition can be defined as a difference in expression over sample mean or median. In both cases, the differential expression is computed over a common reference level.

The typical approaches to solve this problem (e.g Robinson et al., 2010; López-Kleine and González-Prieto, 2016) focusing on analyzing differential expression over the reference level in each condition separately, and thus vastly underestimate sharing. The `mash` method estimates and compares effects across multiple conditions jointly. It can capture both shared and condition-specific effects and it is adaptive. It adapts to the patterns present in the data set being analyzed. Although we wish to take advantage of correlation and thus boost power as in `mash`, we now must consider the additional burden of comparing all subsequent conditions to the same reference group. Specifically, comparing all condition estimates to the same reference condition without accounting for the correlation in errors artificially induces many false positives, and thus any successful joint method needs to account for this correlation in the error structure.

Suppose we observe gene expression in R-1 treatments, and we want to compare the expression with a common control group. On the one hand, simply estimating the expression by subtracting the mean expression in the control condition from every subsequent condition might create false positives, because of the inherent correlation in errors. On the other hand, analyzing each condition separately fails to exploit the power of a joint analysis.

In this paper, we propose a method to apply the `mash` framework in these situations. We call this method `mash commonbaseline`, because there exists no baseline within each condition but each condition is compared to the same baseline. Our goal is to reduce the number of false positives due to the inherent correlation induced when comparing all conditions to a common baseline. A quick review for `mash` is in Section 2. The method for `mash commonbaseline` is described in detail in Section 3. When there is a control group in the study, we estimate the deviation over the control condition. When there is no control group in the study, we estimate the deviation over the mean. Section 4 shows the improvement of the `mash commonbaseline` method through simulations. There is a drawback when we estimate the deviation over the mean. We discuss it in detail in Section 5, and propose methods to estimate the deviation over the median. Section 6 shows the improvement in the real application.

Notations:

We denote matrices by boldface uppercase letters ($\mathbf{A}$), vectors are denoted by boldface lowercase letters ($\mathbf{a}$), and scalars are denoted by non-boldface letters (a or A). All vectors are column-vectors. Lowercase letters may represent elements of a vector or matrix if they have subscripts. For example, a_{ij} is the (i, j)th element of $\mathbf{A}$, a_i is the ith element of $\mathbf{a}$, and $\mathbf{a}_i$ is either the ith row or ith column of $\mathbf{A}$. For indexing, we will generally use capital non-boldface letters to denote the total number of elements and their lowercase non-boldface versions to denote the index. For example, $i = 1, \dots, I$. We denote the matrix transpose by $\mathbf{A}^\intercal$, the matrix inverse by $\mathbf{A}^{-1}$, and the matrix determinant by $\det(\mathbf{A})$.

2 Background

Let b_{jr} ($j = 1, \cdot, J$; r $= 1, \cdots, R$) denote the true value of effect j in condition r. Further let $\hat{b}_{jr}$ denote the (observed) estimate of this effect, and $\hat{s}_{jr}$ denote the standard error of this estimate. Let $\hat{\boldsymbol{B}}$, $\boldsymbol{B}$ and $\hat{\boldsymbol{S}}$ denote the corresponding $J \times R$ matrices, and let $\boldsymbol{b}_j$, $\hat{\boldsymbol{b}}_j$ denote the j-th row of $\boldsymbol{B}$

and $\hat{\boldsymbol{B}}$.

The `mash` model assumes the vector $\hat{\boldsymbol{b}}_j$ is normally distributed about the true effects $\boldsymbol{b}_j$, with variance-covariance matrix $\hat{\boldsymbol{S}}_j \boldsymbol{V} \hat{\boldsymbol{S}}_j$, and that the true effects follow a mixture of multivariate normal. That is,

$$\hat{\boldsymbol{b}}_j | \boldsymbol{b}_j, \hat{\boldsymbol{S}}_j \sim N_R(\boldsymbol{b}_j, \hat{\boldsymbol{S}}_j \boldsymbol{V} \hat{\boldsymbol{S}}_j) \tag{2.1}$$

$$\boldsymbol{b}_j | \boldsymbol{\pi} \sim \sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{kl} N_R(\boldsymbol{0}, \omega_l \boldsymbol{U}_k) \tag{2.2}$$

where $N_R(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of the R-dimensional multivariate normal (MVN) distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and the scaling parameters $\omega_1, \cdots, \omega_L$ are fixed on a dense grid. $\boldsymbol{V}$ is a full rank correlation matrix that accounts for correlations among the measurements in the R conditions, and $\hat{\boldsymbol{S}}$ is the $R \times R$ diagonal matrix with diagonal elements $(\hat{s}_{j1}, \cdots, \hat{s}_{jR})$.

If $\boldsymbol{V}$ is known, we can use it to estimate $\boldsymbol{\pi}$; otherwise, we estimate $\boldsymbol{V}$ and $\boldsymbol{\pi}$ iteratively. In the method we introduced below (Section 3), we cannot estimate $\boldsymbol{V}$. We assume it is known.

The steps of `mash` are:

1. Create a list of both data-driven and canonical covariance matrices, $\boldsymbol{U}$.

   The data-driven covariance matrices are estimated based on the rows of $\hat{\boldsymbol{B}}$ that likely have an effect in at least one condition. One way to find these "strongest effects" is to run univariate adaptive shrinkage `ash` (Stephens, 2016) on the effects in each condition r separately, and computed $lfsr_{jr}$ for each gene j (see below for details about lfsr). We then choose the effects j for which at least one of the conditions showed a significant effect in this univariate analyses ($\min_r lfsr_{jr} < 0.05$). Let $\tilde{J}$ denote the number of selected "strongest effects", and let $\tilde{\boldsymbol{Z}}$ denote the column-centered $\tilde{J} \times R$ matrix of $\boldsymbol{Z}$ scores for these "strong effects". We apply Principal Component Analysis (through Singular Value Decomposition, SVD) and Sparse Factor Analysis (FLASH Wang and Stephens (2018)) on $\tilde{\boldsymbol{Z}}$. SVD yields a set of eigenvalues and eigenvectors of $\tilde{\boldsymbol{Z}}$. Let λ_p, $\boldsymbol{v}_p$ denote the pth eigenvalue and corresponding (right) eigenvector. FLASH yields

$$\tilde{\boldsymbol{Z}} = \boldsymbol{L}\boldsymbol{F}^T + \boldsymbol{E} \tag{2.3}$$

   where $\boldsymbol{L}$ is a sparse $\tilde{J} \times Q$ matrix of loadings, and $\boldsymbol{F}$ is a $R \times Q$ matrix of factors. We construct the following data-driven covariance matrices:

   (a) $\tilde{\boldsymbol{U}}_1 = \frac{1}{\tilde{J}} \tilde{\boldsymbol{Z}}^T \tilde{\boldsymbol{Z}}$, the empirical covariance matrix of $\tilde{\boldsymbol{Z}}$.

   (b) $\tilde{\boldsymbol{U}}_2 = \frac{1}{\tilde{J}} \sum_{p=1}^{P} \lambda_p^2 \boldsymbol{v}_p \boldsymbol{v}_p^T$, which is a rank P approximation of the covariance matrix of $\tilde{Z}$.

   (c) The rank 1 matrices that reflect the effects captured by the p-th eigenvector in the PCA,

   (d) $\tilde{\boldsymbol{U}}_3 = \frac{1}{\tilde{J}} (\boldsymbol{L}\boldsymbol{F}^T)^T (\boldsymbol{L}\boldsymbol{F}^T)$, which is a rank Q approximation of the covariance matrix of $\tilde{\boldsymbol{Z}}$.

   (e) The rank 1 matrices that reflect the effects captured by the q-th factor in the FLASH analysis

We perform extreme deconvolution for the above $\boldsymbol{U}_k$.

2. Estimate unknown weights $\hat{\boldsymbol{\pi}}$ for covariance matrices (and correlation $\hat{\boldsymbol{V}}$).
3. Compute, for each j, the posterior distribution $p(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \boldsymbol{U}, \hat{\boldsymbol{\pi}}, \boldsymbol{V})$.

To measure "significance" of a effect b_{jr}, we use the local false sign rate (lfsr), which is defined as

$$lfsr_{jr} = \min\{p(b_{jr} \geq 0|\hat{\boldsymbol{b}}_j, \hat{\boldsymbol{S}}, \hat{\boldsymbol{\pi}}, \boldsymbol{U}), p(b_{jr} \leq 0|\hat{\boldsymbol{b}}_j, \hat{\boldsymbol{S}}, \hat{\boldsymbol{\pi}}, \boldsymbol{U})\} \tag{2.4}$$

lfsr is the probability that we would get the sign of effect incorrect if we were to use our best guess of the sign. Therefore, a small lfsr indicates high confidence in determining the sign of an effect. Notice that lfsr is more conservative than the local false discovery rate (lfdr), since we can infer lfsr $\geq$ lfdr from the definition.

3 Method

The traditional `mash` model can be used when there is obvious "effect" in each condition. When there is no obvious "effect" in each condition, we estimate the change in some quantity computed in multiple conditions over a reference level. In this section, we discuss the details of our method.

We observe a vector of uncertered noisy mean feature expression $\hat{\boldsymbol{c}}_j$ across R conditions, for each gene j,

$$\hat{\boldsymbol{c}}_j|\boldsymbol{c}_j \sim N_R(\boldsymbol{c}_j, \hat{\boldsymbol{S}}_j\boldsymbol{V}\hat{\boldsymbol{S}}_j) \tag{3.1}$$

where $\boldsymbol{c}_j$ represents the "true" means across R conditions, $\hat{\boldsymbol{S}}_j$ is the $R \times R$ diagonal matrix with diagonal elements $\left(\hat{s}_{j1}, \cdots, \hat{s}_{jR}\right)$, the standard error of the observations. $\boldsymbol{V}$ is a full rank correlation matrix that accounts for correlations among the measurements in the R conditions. In settings where measurements in the R conditions are independent one would set $\boldsymbol{V} = I_R$, the $R \times R$ identity matrix. We assume the correlation $\boldsymbol{V}$ is known and full rank.

We can express the "true" $\boldsymbol{c}_j$ as a mixture of multivariate Normals which centered at an underlying mean $\mu_j\mathbf{1}_R$, each covariance matrix $\boldsymbol{U}_k$ represents the underlying covariance matrix from which the "true" expression $\boldsymbol{c}_j$ are thought to arise. The scaling parameters $\omega_1, \cdots, \omega_L$ are fixed on a dense grid.

$$\boldsymbol{c}_j|\boldsymbol{\pi} \sim \mu_j\mathbf{1}_R + \sum_{k,l} \pi_{kl} N_R(\mathbf{0}, w_l U_k) \tag{3.2}$$

Our goal is to fit a model for the distribution of a R-1 dimensional deviations, $\hat{\boldsymbol{\delta}}_j$, using a set of J observational data points that measure the observed deviations over the reference level. To get the deviations, we use contrast matrix.

Let $\boldsymbol{L}$ denotes the $R - 1 \times R$ matrix of contrasts which removes expression in the reference level from each subsequent condition. When there is a control condition, suppose it is the last condition in the data, the contrast matrix takes the form:

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 & 0 & \cdots & -1 \\ 0 & 1 & 0 & & -1 \\ \vdots & & \ddots & & \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix}_{R-1 \times R} \tag{3.3}$$

When there is no common control condition in the study, we estimate the deviation as differences with the mean. The contrast matrix takes the form:

$$\boldsymbol{L} = \begin{pmatrix} \frac{R-1}{R} & -\frac{1}{R} & -\frac{1}{R} & \cdots & -\frac{1}{R} \\ -\frac{1}{R} & \frac{R-1}{R} & -\frac{1}{R} & & -\frac{1}{R} \\ \vdots & & \ddots & & \\ -\frac{1}{R} & -\frac{1}{R} & \cdots & \frac{R-1}{R} & -\frac{1}{R} \end{pmatrix}_{R-1 \times R} \tag{3.4}$$

Using the contrast matrix $\boldsymbol{L}$, the deviations can be obtained,

$$\hat{\boldsymbol{\delta}}_j = \boldsymbol{L}\hat{\boldsymbol{c}}_j \sim N_{R-1}(\boldsymbol{L}\boldsymbol{c}_j, \boldsymbol{L}\hat{\boldsymbol{S}}_j\boldsymbol{V}\hat{\boldsymbol{S}}_j\boldsymbol{L}^T) \tag{3.5}$$

The “true” deviations have expression

$$\boldsymbol{\delta}_j = \boldsymbol{L}\boldsymbol{c}_j|\boldsymbol{\pi} \sim \mu_j\boldsymbol{L}\mathbf{1}_R + \sum_{k,l}\pi_{kl}N_{R-1}(\mathbf{0}, w_l\boldsymbol{L}\boldsymbol{U}_k\boldsymbol{L}^T)$$

Denote $\boldsymbol{L}\boldsymbol{U}_k\boldsymbol{L}^T$ using $\boldsymbol{U}'_k$

$$\boldsymbol{\delta}_j|\boldsymbol{\pi} \sim \sum_{k,l}\pi_{kl}N_{R-1}(\mathbf{0}, w_l\boldsymbol{U}'_k) \tag{3.6}$$

The distribution 3.5, 3.6 fit in `mash` framework.

Note that the contrast matrix $\boldsymbol{L}$ has only $R-1$ rows, instead of R. This requirement is necessary, since `mash` framework requires the full rank correlation among observed deviations. When there is a control condition, the deviation for the control condition is always zero. When we compare the expression with mean, any deviation can be expressed using the rest deviations, i.e. $\hat{\delta}_{j,i} = \hat{c}_{j,i} - \bar{\hat{c}}_j = -\sum_{r=1,r\neq i}^{R}(\hat{c}_{j,r} - \bar{\hat{c}}_j) = -\sum_{r=1,r\neq i}^{R}\hat{\delta}_{j,r}$. We must discard the deviation in one condition to perform the analysis.

The contrast matrix $\boldsymbol{L}$ defined in (3.4) discards the deviation in the last condition. The deviations are $\hat{c}_{j,1} - \bar{\hat{c}}_j, \hat{c}_{j,2} - \bar{\hat{c}}_j, \cdots, \hat{c}_{j,R-1} - \bar{\hat{c}}_j$. However, the contrast matrix $\boldsymbol{L}$ can discard any condition from $\hat{c}_{j,1} - \bar{\hat{c}}_j, \cdots, \hat{c}_{j,R} - \bar{\hat{c}}_j$. The choice of the discarded condition has no influence on the result.

We can fit the `mash` model as described in Section 2 using 3.5, 3.6. The quantity of interest now is $\boldsymbol{\delta}_j$, which represents the true deviations from the reference level. The $\boldsymbol{\delta}_j$ can be treated as the effects in `mash`. With the model, we get the posterior distribution of true deviations $\boldsymbol{\delta}_j$. The estimation and inference can therefore obtained.

The critical step above is the covariance of the observed deviations. Even if the original covariance matrix is diagonal ($\hat{\boldsymbol{S}}_j \boldsymbol{V} \hat{\boldsymbol{S}}_j$ is diagonal), and thus the observed noisy mean expression measurements in each condition are independent, $\boldsymbol{L}\hat{\boldsymbol{S}}_j \boldsymbol{V} \hat{\boldsymbol{S}}_j \boldsymbol{L}^T$ is not diagonal and thus accounts for the induced correlation in errors.

For deviations over means, the posteriors from the `mash` model are only for the first R-1 conditions. Using a linear transformation of the posteriors, we can obtain the posteriors for all R conditions. The linear transformation corresponding to contrast matrix 3.4 is

$$\boldsymbol{A} = \left(\begin{array}{ccc} & \boldsymbol{I}_{R-1} & \\ \hline -1 & \cdots & -1 \end{array} \right)_{R\times(R-1)} \tag{3.7}$$

$$\boldsymbol{A}\boldsymbol{\delta}_j = \delta_{j,1}, \cdots, \delta_{j,R-1}, -\sum_{r=1}^{R-1} \delta_{j,r} \tag{3.8}$$

All procedures are implemented in `mash` framework. The R package is available at https://github.com/stephenslab/mashr.

4 Simulations

In this section, we demonstrate that failing to account for the inherent correlations induced by subtracting the same noisy observed reference measurement from each subsequent condition inflates our identification of true deviations. We compare the results from our new method `mash commonbaseline` with the one from `mash` that ignoring the induced correlations, we call this `independent mash` model. We show the simulation with the control group and without the control group.

4.1 Study with a control condition

We first show the improvement for study with a control condition. Suppose the last condition in the simulation study below is the control condition.

4.1.1 Without Deviation

There is no true deviation exists in this simulation.

$$\boldsymbol{c}_j = \mu_j \mathbf{1}_{10} \tag{4.1}$$

$$\hat{\boldsymbol{c}}_j \sim N_{10}(\mu_j \mathbf{1}_{10}, \frac{1}{2}\boldsymbol{I}) \tag{4.2}$$

Let L be the contrast matrix as defined in (3.3). The `mash commonbaseline` uses the model

$$\hat{\boldsymbol{\delta}}_j \sim N_9(\mathbf{0}, \frac{1}{2}\boldsymbol{L}\boldsymbol{L}^T) \tag{4.3}$$

However, one might subtract the expression in the control condition from every subsequent condition, and ignore the induced correlations, which leads to the `independent mash` model.

$$\hat{\boldsymbol{\delta}}_j \sim N_9(\mathbf{0}, \boldsymbol{I}) \tag{4.4}$$

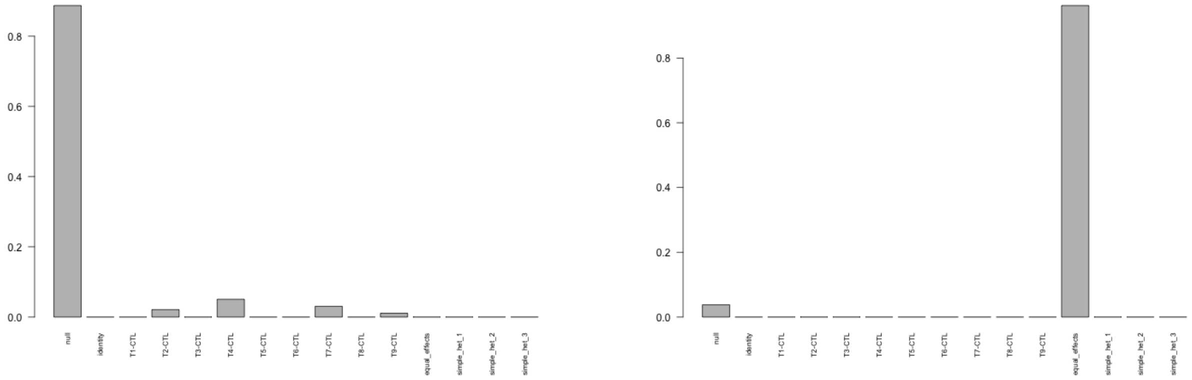
The variance of $\hat{\delta}_{jr}$ is calculated by

$$\text{Var}(\hat{c}_{jr} - \hat{c}_{jR} | c_{jr}, c_{jR}) = \text{Var}(\hat{c}_{jr}|c_{jr}) + \text{Var}(\hat{c}_{jR}|c_{jR}) = \frac{1}{2} + \frac{1}{2} = 1 \quad r = 1, \cdots, R-1 \tag{4.5}$$

The correlation between $\hat{\delta}_{jr}$ and $\hat{\delta}_{js'}$, $r \neq r'$, is ignored.

The `mash commonbaseline` method yields a much higher log-likelihood. Including the induced correlations, there are no discoveries which is as we expected. Because the true deviations, $\boldsymbol{\delta}_j$, are zero for all samples. However, the `independent mash` model produces around 30% discoveries, which are all false discoveries.

Moreover, the `independent mash` model produces wrong weights on the covariance structures. The estimated weights for covariance matrices are plotted in Figure 1. The `mash commonbaseline` method produces the correct weights on covariance matrices, the majority of the weights is on the null matrix. In contrast, the `independent mash` model puts the majority of weights on the equal effect matrix. This is caused by the ignorance of correlations in errors. From this simulation, we see the improvement of false discoveries by `mash commonbaseline` clearly.



(a) `mashcommonbaseline`
Estimated Weights

(b) `independent mash`
Estimated Weights

Figure 1: Without deviation data: Estimated weights for covariance matrices

4.1.2 With Deviation

We add signal to a number of "non-null" simulations such that for any subgroup $r = 1, \cdots, R-1$, an expression different from the control exists:

$$\boldsymbol{c}_{j1\ldots(R-1)} = c_{jR}\mathbf{1}_{R-1} + \boldsymbol{\delta}_j \tag{4.6}$$

$$\hat{\boldsymbol{c}}_j | \boldsymbol{c}_j \sim N_R(\boldsymbol{c}_j, \frac{1}{2}\boldsymbol{I}) \tag{4.7}$$

We simulated data with 10 conditions, and four different types of deviations $\boldsymbol{\delta}_j$: null ($\boldsymbol{\delta}_j = 0$), independent among conditions, condition-specific in condition 1 ($\delta_{j1} \neq 0$), and shared (equal deviations in all sub-conditions, $\delta_{j1\cdots 9} = \boldsymbol{f}\mathbf{1}_9$). The data contains 10% non-null deviations

$$\boldsymbol{\delta}_j \sim \frac{9}{10}N_9(\mathbf{0}, \mathbf{0}) + \frac{1}{30}N_9(\mathbf{0}, \boldsymbol{I}) + \frac{1}{30}N_9(\mathbf{0}, \mathbf{e}_1\mathbf{e}_1^T) + \frac{1}{30}N_9(\mathbf{0}, \mathbf{1}\mathbf{1}^T) \tag{4.8}$$

Let L be the contrast matrix as in (3.3). Therefore,

$$\hat{\boldsymbol{\delta}}_j | \boldsymbol{\delta}_j = \boldsymbol{L}\hat{\boldsymbol{c}}_j | \boldsymbol{c}_j \sim N_9(\boldsymbol{\delta}_j, \frac{1}{2}\boldsymbol{L}\boldsymbol{L}^T) \tag{4.9}$$

Figure 2a compares the accuracy of deviation size estimates, as measured by the relative root mean squared error (RRMSE), which is the RMSE of the estimates divided by the RMSE achieved by simply using the original observed estimates $\hat{\boldsymbol{\delta}}_j$ for the deviations. It can be written as

$$RRMSE = \sqrt{\frac{\mathbb{E}((\delta_{jr} - \hat{\hat{\delta}}_{jr})^2)}{\mathbb{E}((\delta_{jr} - \hat{\delta}_{jr})^2)}} \tag{4.10}$$

Both methods have $RRMSE < 1$, indicating a substantial improvement in accuracy compared with the original observed effects $\hat{\boldsymbol{\delta}}_j$. As expected, the `mash commonbaseline` outperforms the `independent mash` model. The accuracy of the estimated deviations improved.

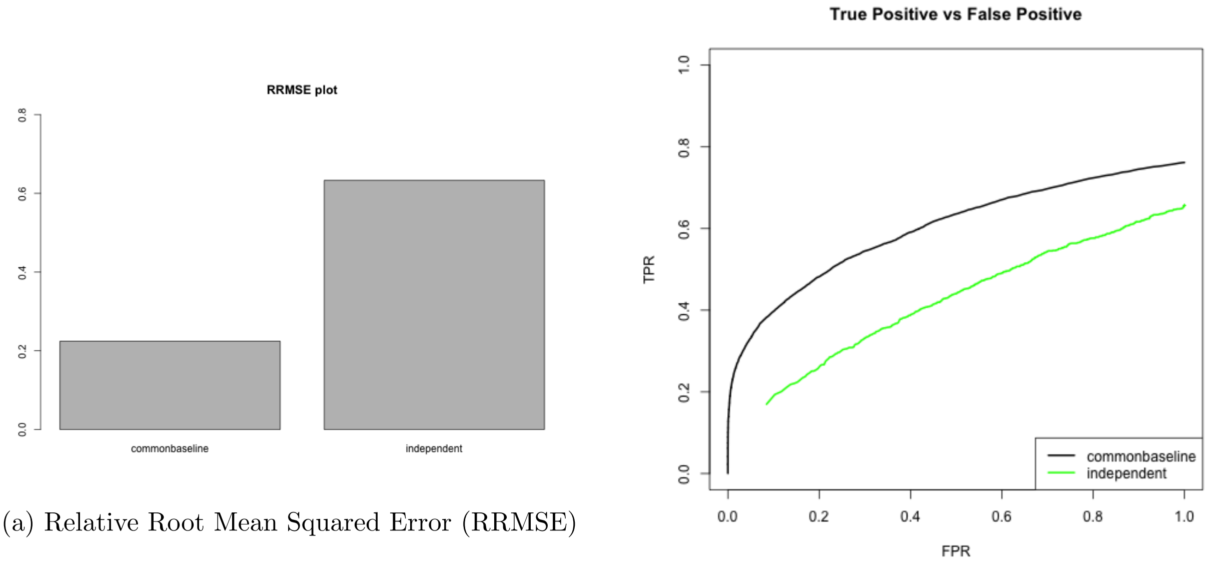
For the ROC curves in Figure 2b, the True Positive Rate and False Positive Rate are computed at any given threshold t as

$$\text{True Positive Rate} = \frac{|CS \cap S|}{|T|} \quad \text{False Positive Rate} = \frac{|N \cap S|}{|N|} \tag{4.11}$$

where S is the set of significant results at threshold t, CS is the set of correctly-signed results, T is the set of true (non-zero) deviations and N is the set of null deviations:

$$S = \{j, r : lfsr_{jr} \leq t\} \tag{4.12}$$
$$CS = \{j, r : E(\delta_{jr}|\hat{\Delta}) \times \delta_{jr} > 0\} \tag{4.13}$$
$$N = \{j, r : \delta_{jr} = 0\} \tag{4.14}$$
$$T = \{j, r : \delta_{jr} \neq 0\} \tag{4.15}$$

We require the estimated sign (+/-) of each significant deviation to be correct to be considered a "true positive". Our `mash commonbaseline` method outperforms the `independent mash` method.



(a) Relative Root Mean Squared Error (RRMSE)

(b) ROC curve

Figure 2: With deviation data: Relative Root Mean Square Error (RRMSE) and the ROC curve

4.2 Study without control condition

We show the improvement for study without control condition here. We estimate the change in expression computed in R conditions over their mean.

4.2.1 Without Deviation

The simulation scheme is similar as in Section 4.1.1.

$$\hat{\boldsymbol{c}}_j|\boldsymbol{c}_j \sim N_{10}(\boldsymbol{c}_j, \boldsymbol{I}) \tag{4.16}$$

$$\boldsymbol{c}_j = \mu_j \mathbf{1}_{10} \tag{4.17}$$

Let L be the contrast matrix as defined in (3.4). The `mash commonbaseline` has the model

$$\hat{\boldsymbol{\delta}}_j|\boldsymbol{\delta}_j \sim N_9(\boldsymbol{\delta}_j, \boldsymbol{L}\boldsymbol{L}^T) \tag{4.18}$$

$$\boldsymbol{\delta}_j = \mathbf{0} \tag{4.19}$$

In the `independent mash` model, the correlation among $\hat{\delta}_{jr}$ and $\hat{\delta}_{jr'}$ is ignored.

$$\hat{\boldsymbol{\delta}}_j|\boldsymbol{\delta}_j \sim N_9(\boldsymbol{\delta}_j, \boldsymbol{S}) \tag{4.20}$$

where $\boldsymbol{S}$ is a diagonal matrix with diagonal elements

$$\text{Var}(\hat{\boldsymbol{c}}_{j,r} - \bar{\hat{c}}_j|\boldsymbol{c}_j) = \frac{R-1}{R} \tag{4.21}$$

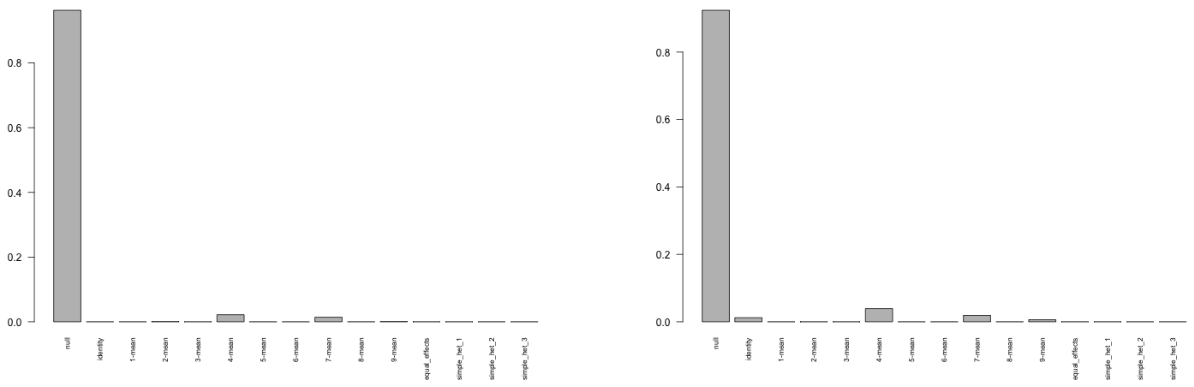
The `mash commonbaseline` method yields higher log-likelihood. There are no discoveries from both models, since the true deviations are all zero. Both methods yields large weights on the null matrix (See Figure 3). When there are no signals, subtracting the mean directly from the data performs as well as the `mash commonbaseline` model.

4.2.2 With Deviation

We simulate data with 10 conditions, half of the samples have equal expression among conditions. In the rest samples, half have higher and equal expression in the first 2 conditions, half have higher expression in the last condition.

$$\boldsymbol{c}_j \sim \mu_j \mathbf{1}_{10} + \frac{1}{2}N_{10}(\mathbf{0}, \mathbf{0}) + \frac{1}{4}N_{10}(\mathbf{0}, 9\begin{pmatrix} \mathbf{1}_2\mathbf{1}_2^T & \mathbf{0}_{2\times 8} \\ \mathbf{0}_{8\times 2} & \mathbf{0}_{8\times 8} \end{pmatrix}) + \frac{1}{4}N_{10}(\mathbf{0}, 9\begin{pmatrix} \mathbf{0}_{9\times 9} & \mathbf{0}_9 \\ \mathbf{0}_9^T & 1 \end{pmatrix}) \tag{4.22}$$

$$\hat{\boldsymbol{c}}_j|\boldsymbol{c}_j \sim N_{10}(\boldsymbol{c}_j, \boldsymbol{I}) \tag{4.23}$$



(a) `mashcommonbaseline`
Estimated Weights

(b) `independent mash`
Estimated Weights

Figure 3: Without deviation data: Estimated weights for covariance matrices

Let L be the contrast matrix in (3.4) that subtract the mean from each sample.

$$\hat{\boldsymbol{\delta}}_j|\boldsymbol{\delta}_j \sim N_9(\boldsymbol{\delta}_j, \boldsymbol{L}\boldsymbol{L}^T) \tag{4.24}$$

Half of the true deviations are zero, quarter of the deviations $\boldsymbol{\delta}_j$ have correlation that the first two conditions are negatively correlated with the rest conditions. For the rest quarter of the deviations $\boldsymbol{\delta}_j$, the first 9 conditions are negatively correlated with the last condition.

There are two ways to fit the wrong model. One is the `independent mash` model, which ignores the induced correlation (4.20). The other one is the `constant mash` model, which ignores the induced correlation and mis-calculates the standard error, i.e treat the subtracted mean as constant:

$$\hat{\boldsymbol{\delta}}_j|\boldsymbol{\delta}_j \sim N_9(\boldsymbol{\delta}_j, \boldsymbol{I}) \tag{4.25}$$

We apply six models using the simulated data.

1. mash commonbaseline, discard the 10-th condition (`m.10`)
2. mash commonbaseline, discard the 9-th condition (`m.9`)
3. `independent mash` model, discard the 10-th condition (`Indep.10`)
4. `independent mash` model, discard the 9-th condition (`Indep.9`)
5. `constant mash` model, discard the 10-th condition (`Const.10`)
6. `constant mash` model, discard the 9-th condition (`Const.9`)

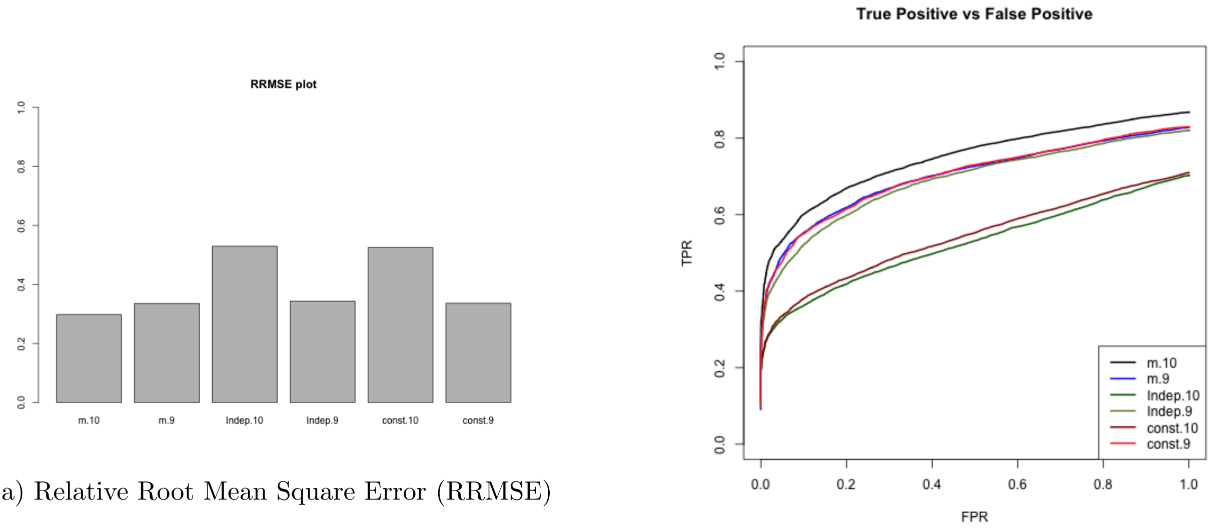
To better capture the covariance structures, we use `mash` framework with data driven covariance matrices.

From Figure 4a, 4b, we can see the results from `m.9` and `m.10` have similar accuracy and power, which confirms our statement in Section 3. `mash commonbaseline` is robust to the choice of the discarded condition. In contrast, the performance of `independent mash` and `constant mash` model depend heavily on the choice of discarded condition.

Based on our simulation scheme, quarter of the samples have higher expression in the 10-th condition. When we subtract the mean and discard the 10-th condition, the deviations in the first 9 conditions are equal with small magnitude. The `independent mash` and `constant mash` model fail to recognize this pattern, and shrink all deviations to zero, whereas the `mash commonbaseline` model captures the pattern successfully. Ignoring the induced correlation in `Indep.10` and `Const.10` causes the low accuracy of the estimated deviation and low power.

When we discard the 9-th condition, the `mash commonbaseline`, `independent mash` and `constant mash` models perform similarly. The reason is that the deviations in the remaining conditions are obvious and the model can capture the pattern even we ignore the induced correlation. There is not much benefit by using `mash commonbaseline` model in this scenario.

The `independent mash` and `constant mash` model perform very similarly. The variance of the deviation is $\frac{R-1}{R}$ in `independent mash` model, 1 in `constant mash` model. When the number of conditions is large, the difference is negligible.



(a) Relative Root Mean Square Error (RRMSE)

(b) ROC curve

Figure 4: Comparison of methods on data with true deviations: (a) shows the accuracy of the estimated deviations. (b) shows the ROC curve.

5 Compare with Median

From simulations above, it is clear that including the inherent correlation induced when comparing all conditions to a common control, the number of false positives reduce dramatically. When the comparison is made with the mean, we need to discard the deviation in one condition to get rid of the perfect dependency among deviations. If the discarded condition does not have large deviation, the improvement from `mash commonbaseline` is not obvious. The reason is that the quantity subtracted from each condition is a summary statistics, which contains information from every condition. The correlation between the deviations becomes weaker, and it is negligible. However, when we compare the expression with a common control condition, the correlation is not negligible.

We illustrate with a simple example that the standard error is common among different conditions and the measurements in different conditions are independent.

$$\mathrm{Var}(\boldsymbol{c}_j|\boldsymbol{c}_j) = s^2\boldsymbol{I}_R \tag{5.1}$$

If the last condition is the control group, the deviation is $c_{jr} - c_{jR}$, the correlation between two deviations is $\frac{1}{2}$.

$$\mathrm{Var}(\hat{c}_{jr} - \hat{c}_{jR}|\boldsymbol{c}_j) = 2s^2 \quad Cov(\hat{c}_{jr} - \hat{c}_{jR}, \hat{c}_{jr'} - \hat{c}_{jR}|\boldsymbol{c}_j) = s^2 \quad r \neq r'$$

$$Cor(\hat{c}_{jr} - \hat{c}_{jR}, \hat{c}_{jr'} - \hat{c}_{jR}|\boldsymbol{c}_j) = \frac{1}{2}$$

In contrast, if the deviation is computed over the mean, the correlation between two deviations is $\frac{1}{1-R}$.

$$\mathrm{Var}(\hat{c}_{jr} - \bar{\hat{\boldsymbol{c}}}_j|\boldsymbol{c}_j) = \frac{R}{R-1}s^2 \quad Cov(\hat{c}_{jr} - \bar{\hat{\boldsymbol{c}}}_j, \hat{c}_{jr'} - \bar{\hat{\boldsymbol{c}}}_j|\boldsymbol{c}_j) = -\frac{s^2}{R} \quad r \neq r'$$

$$Cor(\hat{c}_{jr} - \bar{\hat{\boldsymbol{c}}}_j, \hat{c}_{jr'} - \bar{\hat{\boldsymbol{c}}}_j|c_j) = \frac{1}{1-R}$$

Ignoring the correlation 0.5 will cause the false discovery. However, the correlation $\frac{1}{1-R}$ is close to zero if the number of conditions R is large, so it can be ignored.

There is one drawback when we subtract the mean from the quantity in each condition. When some conditions have large positive deviations over the mean, the other conditions must have negative deviations. For instance, we have result from a gene expression experiments. The first two conditions among R conditions have high gene expression level, the other conditions have nearly zero expression level. Subtracting mean from each condition leads to high positive deviations for the first two conditions, negative deviations for the other conditions. Using our `mash commonbaseline` framework, we conclude that all conditions have deviations, but the deviations in the first two conditions have opposite sign than others. However, it is more parsimonious to conclude that the first two conditions are different from others. It is better to report "condition specific" effect than a shared effect at all but one condition. To achieve this parsimonious statement, we could estimate the change in the quantity computed in R conditions over their median.

If we compare the gene expression with the median among R conditions in the above example, we can identify that the first two conditions have high gene expression level comparing with other conditions. Since subtracting median would not cause the rank deficiency problem in the correlation matrix $\boldsymbol{V}$, we do not need to discard deviations in any condition. However, there is no contrast matrix exists to get the deviations over median. Therefore, we cannot use the `mash commonbaseline` method directly. There are several ways to analysis the deviation over median using `mash`.

1. Subtract median directly:

   Like mean, median is also a summary statistics. Using the idea discussed above, we could ignore the correlation among deviations. The variance of median in a multivariate normal distribution is hard to compute. To the best of our knowledge, there is no result about the explicit variance of the median exists in literature. But there is no need to worry about it. We see the `independent mash` and `constant mash` model perform similarly from Section 4. This allows us to subtract the median directly from the observations without considering the variance and the correlation, and identify deviations using `mash`.

2. Estimate from posterior samples:

   We can first get the posterior samples of deviations over mean using `mash commonbaseline`. Then estimate the posterior for

   $$\hat{\boldsymbol{\delta}}_j - median(\hat{\boldsymbol{\delta}}_j) = (\hat{\boldsymbol{c}}_j - \bar{\hat{\boldsymbol{c}}}_j) - median(\hat{\boldsymbol{c}}_j - \bar{\hat{\boldsymbol{c}}}_j) = \hat{\boldsymbol{c}}_j - median(\hat{\boldsymbol{c}}_j)$$

   which is same as we perform `mash commonbaseline` on deviation over median.

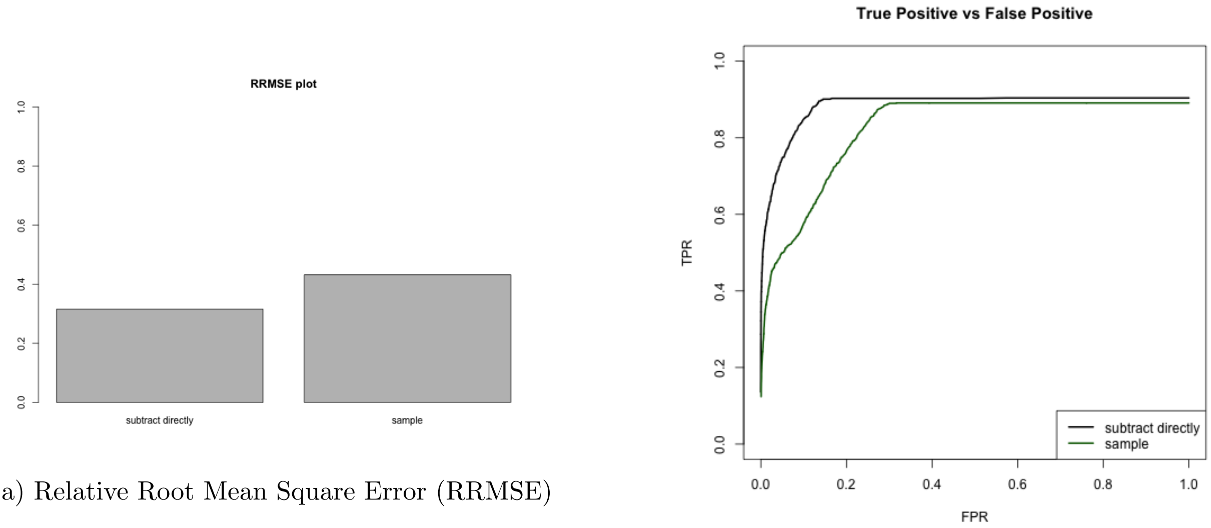
   We can summarize the posterior information based on the samples.

We apply the `mash median` model on the simulated data in Section 4. When there is no deviation in the data, the simulation scheme is same as 4.16, 4.17. There is no false discovery using both methods. When there are true deviations, the simulation scheme is same as 4.22, 4.23. Figure 5a, 5b shows the RRMSE and the ROC curve. Both methods have $RRMSE < 1$, indicating a substantial improvement in accuracy compared with the original observed effects $\hat{\boldsymbol{\delta}}_j$. The method 1, subtracting median directly from the data, performs slightly better.

6 Application

To illustrate the power of our `mash commonbaseline` approach, we applied it to data in which gene expression across multiple conditions had indeed been compared to a control. Blischak et al. (2015) analyzed gene expression in cells infected with 8 strains of Tuberculin bacteria and sought to understand changes in gene expression in comparison to uninfected controls.

We consider the gene expression patterns identified 18 hours post-infection in our analysis. The gene expression readings (see Blischak et al., 2015, for details) represent batch-corrected $\log_2$ counts



(a) Relative Root Mean Square Error (RRMSE)

(b) ROC curve

Figure 5: Comparison of methods on data with true deviations: (a) shows the accuracy of the estimated deviations. (b) shows the ROC curve.

per million for the 12,728 Ensemble genes, each has 156 samples gene expression data across the 8 conditions and control. To obtain summary statistics for each gene-condition pair, we used the Empirical Bayes linear model method Limma (Smyth, 2004) to estimate $\hat{\boldsymbol{c}}_j$ of mean gene expression and corresponding standard errors. The $[j, r]$ entry of the matrix $\hat{\boldsymbol{C}}$ represents the mean gene expression of gene j for individuals with infected condition r. There are 9 conditions in total including the uninfected control.

Denote the matrix of observed deviations, $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{C}}\boldsymbol{L}^T$, $\boldsymbol{L}$ is defined as in 3.3. We run univariate adaptive shrinkage `ash` (Stephens, 2016) on the deviations in each bacteria infection r separately, and computed $lfsr_{jr}$ for each gene j. We then choose the genes j for which at least one of the bacteria showed a significant deviation in this univariate analyses ($\min_r lfsr_{jr} < 0.05$).

We use these strongest genes to estimate the data-driven covariance matrices as described in Section 2 and use the larger 12,728 gene data set to estimate weights and compute posteriors for all 12,728 genes.

Figure 6a shows the patterns `cormotif` identified by Blischak et al. (2015). The four differential expressed patterns are collapsed into the primary patterns of sharing in `mash commonbaseline`. In `mash commonbaseline`, we see that the majority of the hierarchical weight falls on the pattern that reflect broad sharing of both sign and magnitude across conditions. `Yersinia` and `Salmonella` seem to share effects more closely (see Figure 6b, 6c).

Controlling the local false sign rate at 0.05, there are 3,009 genes differentially expressed for at least one bacteria infection. Blischak et al. (2015) discovers much more genes than our `mash` commonbaseline method, which may contains large amount of false discoveries.

We assess the quantitative similarity of deviations by sign and magnitude. Here we define similar in magnitude to mean both the same sign and within a factor of 2 of one another. Figure 6d and 6e show the heatmap for the proportion of differentially expressed genes have similar magnitude and sign in each pair of bacteria infection. Deviations tend to be shared by sign much more commonly than they are shared by magnitude, reflecting the fact that to be shared by magnitude must be shared by sign. Almost all genes share the deviation direction among bacteria. The magnitude of deviations in Rv+ and BCG are not similar with Yersinia and Salmonella. The observation agrees the primary pattern from the mash commonbaseline model.

Compare with median

We treat the expression level in uninfected control as the baseline in the analysis above. Here, we change the baseline to median. We compare the gene expression in the uninfected control and 8 bacteria infections with their median.

The observed gene expression matrix $\hat{\boldsymbol{C}}_{12,728\times 9}$ is same as above. The row of observed deviations $\hat{\boldsymbol{\Delta}}$ is $\hat{\boldsymbol{c}}_j - median(\hat{\boldsymbol{c}}_j)$. We run mash as described in Section 2. We see that the majority of the hierarchical weight from mash falls on the pattern that Yersinia and Salmonella are negatively correlated with the uninfected control (Figure 7a, 7b). Yersinia and Salmonella share deviations, and have opposite direction than deviation in uninfected condition.

Controlling the local false sign rate at 0.05, there are 3,068 genes differentially expressed in at least one condition, which is similar as we compare the expression with the uninfected control. We also assess the quantitative similarity of deviations by sign and magnitude. Figure 7c and 7d show the heatmap for the proportion of differentially expressed genes have similar magnitude and sign in each pair of infect condition. As we expected, Yersinia and Salmonella have similar deviations. Rv+ and BCG have similar deviations, and the deviations have opposite sign with Yersinia and Salmonella. The observations agree with the primary pattern from the mash model.

7 Discussion

The number of false positives reduce dramatically, when we include the inherent correlation induced when comparing all conditions to a common control. When there is no control condition in the study, we can compare the quantity with their mean and we need to discard the deviations in one condition to get rid of the perfect dependency among deviations. If the discarded condition does not have large deviation, the improvement from mash commonbaseline is not obvious. If the discarded condition has some high deviations, the results from mash commonbaseline has higher accuracy and larger power. In the real life, we don't know whether the discarded condition has any large deviations. In this case, it is better to use mash commonbaseline model when the comparison is made with the mean.

When there is no control condition in the study, we can also compare the quantity with their median, which provides more parsimonious conclusion. There are two ways to estimate the deviations over median, 1. subtract median from all conditions and use `mash`, 2. estimate from posteriors of deviations over mean. These two methods perform similarly. We recommend the first one, because it is simpler and faster than the sampling method.

References

Blischak, J. D., L. Tailleux, A. Mitrano, L. B. Barreiro, and Y. Gilad (2015, 12). Mycobacterial infection induces a specific human innate immune response. *Scientific Reports 5*(1), 16882.

López-Kleine, L. and C. González-Prieto (2016, 7). Challenges Analyzing RNA-Seq Gene Expression Data. *Open Journal of Statistics 06*(04), 628–636.

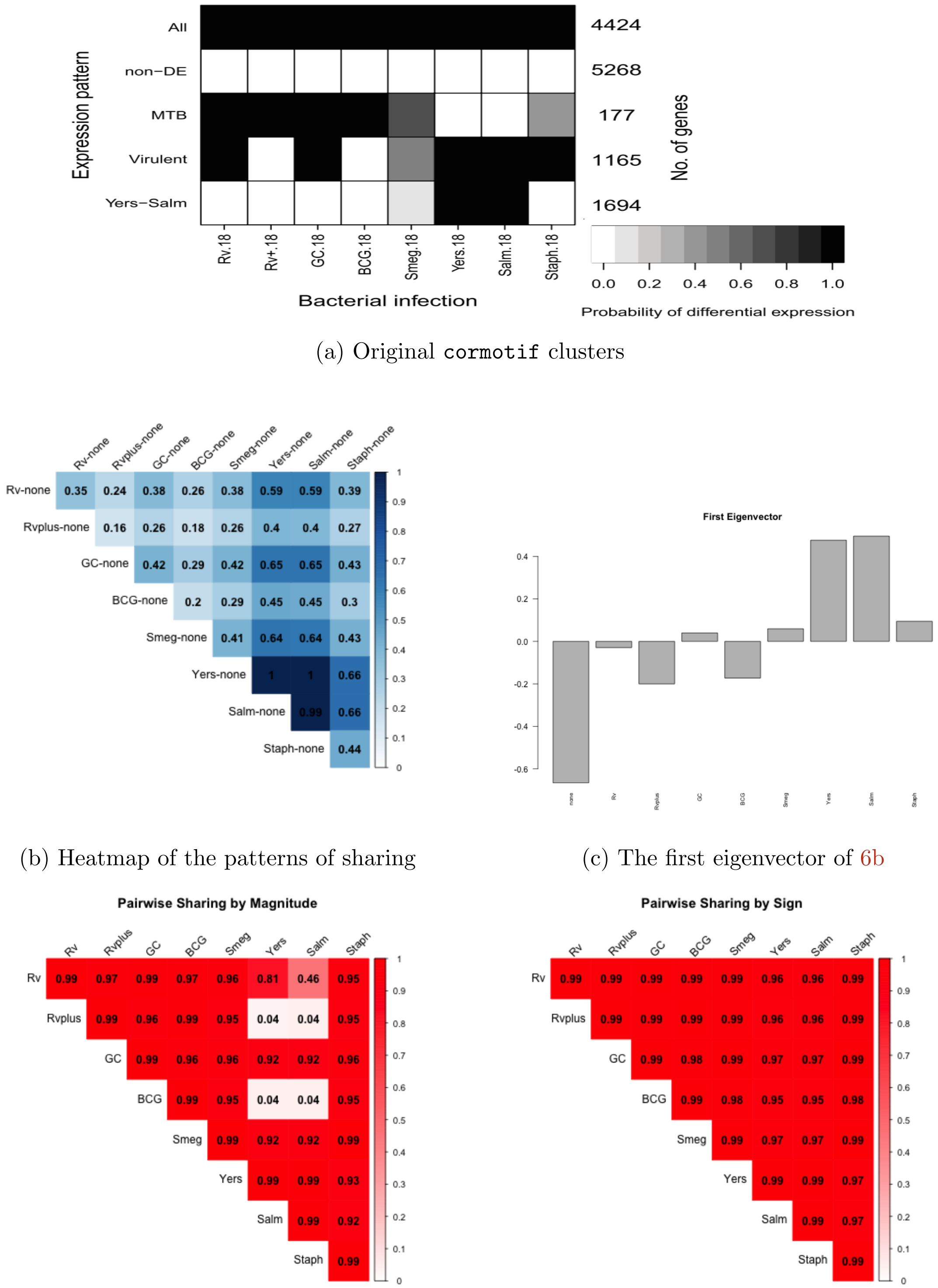
Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010, 1). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics 26*(1), 139–140.

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology 3*(1), 1–25.

Stephens, M. (2016, 10). False discovery rates: a new deal. *Biostatistics 18*(2), kxw041.

Urbut, S. M., G. Wang, P. Carbonetto, and M. Stephens (2018, 9). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *bioRxiv*, 096552.

Wang, W. and M. Stephens (2018, 2). Empirical Bayes Matrix Factorization.



(a) Original cormotif clusters

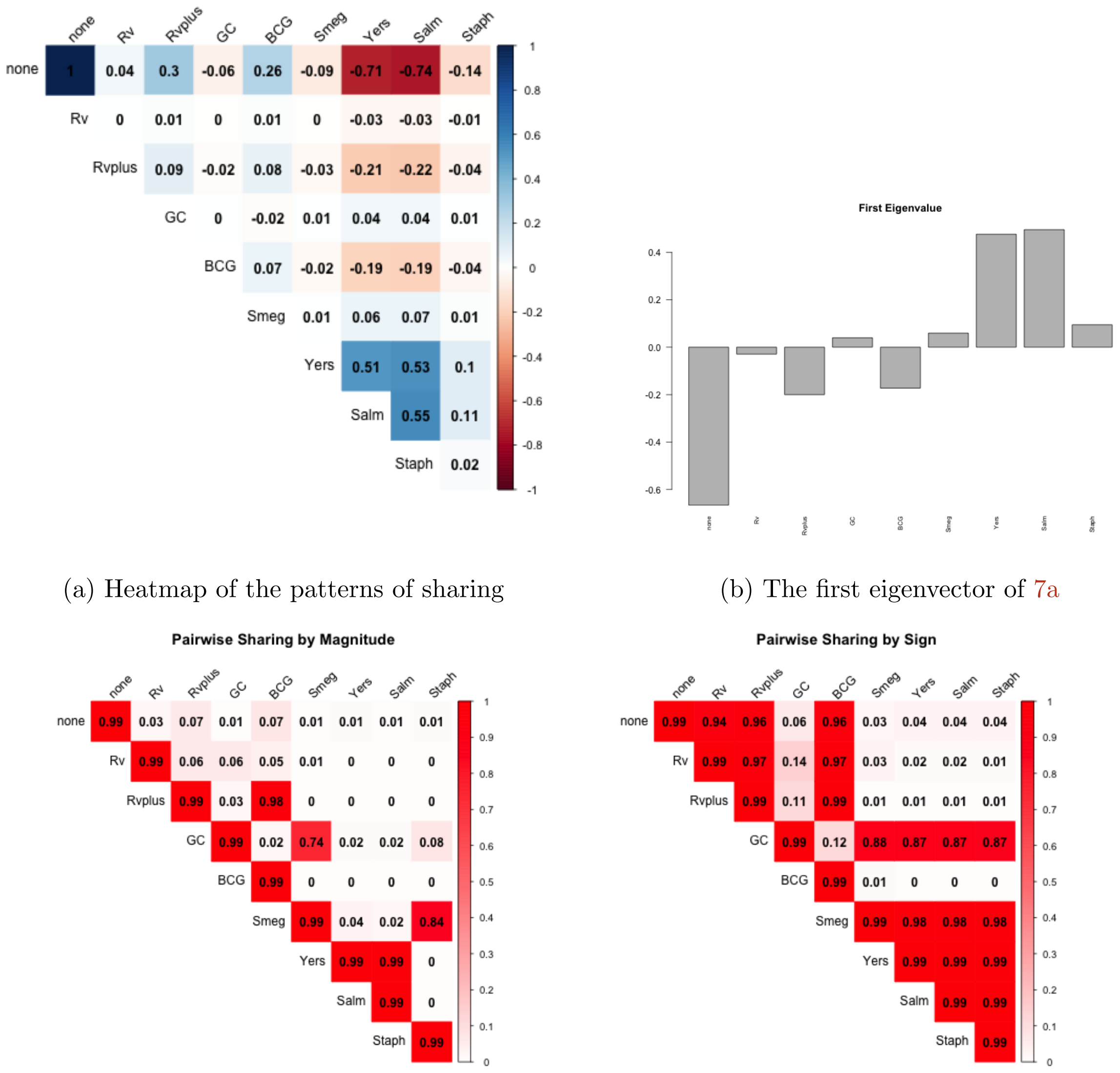
(b) Heatmap of the patterns of sharing

(c) The first eigenvector of 6b

(d) Pairwise sharing by Magnitude

(e) Pairwise sharing by Sign

Figure 6: The top plot 6a shows the five patterns cormotif identified. (Blischak et al., 2015). The plot 6b and 6c show the pattern of sharing with the most weight in mash commonbaseline. The left plot is the heatmap of the scaled covariance matrix. It's first eigenvector is in the right plot. The bottom plots show the similarity of deviations by sign and magnitude. The left plot shows the proportion of significant genes that are "shared in magnitude". The right plot shows the proportion of significant genes that are "shared in sign".



(a) Heatmap of the patterns of sharing

(b) The first eigenvector of 7a

(c) Pairwise sharing by Magnitude

(d) Pairwise sharing by Sign

Figure 7: The top figures 7a and 7b show the pattern of sharing with the most weight in `mash`. The left plot is the heatmap of the scaled covariance matrix. It's first eigenvector is in the right plot. The bottom plots show the similarity of deviations by sign and magnitude. The left plot shows the proportion of significant genes that are "shared in magnitude". The right plot shows the proportion of significant genes that are "shared in sign"