Permutation-based Maximum Covariance Analysis (PMCA)

by Robyn L. Ball

1 Introduction

When presented with two different sets of measurements on the same samples, it is often of interest to identify similar patterns between the variables and thereby, map one set of variables onto the other set. Examples include measurements of gene expression and the proportion of a cell-type on the same biological samples and measurements of brain cell activity and behavior at particular time. In these examples, one is interested in knowing the genes that are specific to a cell type and which brain cells are specific to a behavior. However, one could consider the "sample" not as an actual physical sample but as the entity being measured. Examples of this type may include measurements of temperature or stress by two different methods or on two different levels. In these cases, the "samples" are the temperature and stress. In the case of stress measurements, the investigator could map the elements on the material level that have the same stress pattern as elements on the metal level.

A traditional method for analyzing these types of paired data is maximum covariance analysis (MCA). MCA was first developed in the meteorological sciences by John Prohaska in 1976 [1] and was popularized in the climatological sciences in the 1990s by Bretherton and Wallace [2, 3]. More recently, an MCA approach has been used in a bioinformatics context to clarify relationships between gene and protein expression [7]. While MCA is often an effective tool for detecting common signals in two sets of variables, Steve Cherry showed that it can be limited by a tendency to fit spurious patterns, especially when faced with increased sampling variation [4, 5]. Current methods [7, 6] employ a parametric smoothing model to mitigate this tendency but I wanted to take a nonparametric approach and thus, I developed a novel permutation-based maximum covariance analysis (PMCA) method that not only overcomes the spurious pattern identification liability but that does so without the need for any parametric assumptions.

2 Methodology

Consider two datasets, $\boldsymbol{X}_{(p\times n)}, \boldsymbol{Y}_{(q\times n)}$ that have different numbers of rows but the same number of columns (samples) and we want to know what elements of $\boldsymbol{Y}$ (rowterms of $\boldsymbol{Y}$) have the same pattern as elements of $\boldsymbol{X}$ (rowterms of $\boldsymbol{X}$).

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1n} \\ y_{21} & y_{22} & \dots & y_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ y_{q1} & y_{q2} & \dots & y_{qn} \end{bmatrix}$$

First, center the matrices by substracting the row averages. For $i = 1, 2, \dots p, k = 1, 2, \dots q, s = 1, 2, \dots n$.

$$\begin{aligned} \tilde{x}_{is} &= x_{is} - \frac{1}{n}\sum_{s=1}^{n} x_{is} \\ \tilde{y}_{ks} &= y_{ks} - \frac{1}{n}\sum_{s=1}^{n} y_{ks} \end{aligned} \tag{1}$$

so that

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \dots & \tilde{x}_{1n} \\ \tilde{x}_{21} & \tilde{x}_{22} & \dots & \tilde{x}_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{x}_{p1} & \tilde{x}_{p2} & \dots & \tilde{x}_{pn} \end{bmatrix}, \tilde{\boldsymbol{Y}} = \begin{bmatrix} \tilde{y}_{11} & \tilde{y}_{12} & \dots & \tilde{y}_{1n} \\ \tilde{y}_{21} & \tilde{y}_{22} & \dots & \tilde{y}_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{y}_{q1} & \tilde{y}_{q2} & \dots & \tilde{y}_{qn} \end{bmatrix}$$

Let $\boldsymbol{C}_{(p\times q)} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{Y}}^T/n$ be the covariance matrix of $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ and use Singular Value Decomposition (SVD) to decompose $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$.

Then, the column vectors of $\boldsymbol{U}$ correspond to the structures in $\boldsymbol{X}$ data that explain the covariance and the column vectors of $\boldsymbol{V}$ correspond to the structures in $\boldsymbol{Y}$ that explain the covariance. The singular values, $\text{diag}(\boldsymbol{\Sigma})$, are equivalent to the square root of the eigenvalues of $\boldsymbol{C}$.

We then use these results to compute the principal component matrices.

$$\begin{aligned} \boldsymbol{P}_x &= \boldsymbol{U}^T\tilde{\boldsymbol{X}} \text{ is } (p \times n) \\ \boldsymbol{P}_y &= \boldsymbol{V}^T\tilde{\boldsymbol{Y}} \text{ is } (q \times n) \end{aligned}$$

Up to this point, PMCA is consistent with traditional MCA but the following modifications are novel.

Because we are primarily interested in mapping rowterms of $\boldsymbol{Y}$ onto $\boldsymbol{X}$, we consider $\boldsymbol{P}_x$, the principal components of the covariance matrix that correspond to the elements of $\boldsymbol{X}$, and calculate the homogeneous and heterogeneous regressions using $\boldsymbol{P}_x$.

Notice that the last column of $\boldsymbol{P}_x = \boldsymbol{0}$.

Here, for notation purposes, we assume that $p < q$.

The amplitudes of $\boldsymbol{X}$ are

$$\begin{aligned} \boldsymbol{Z}_{x(p\times p)} &= \boldsymbol{X}\boldsymbol{P}_x^T \\ &= \begin{bmatrix} z_{x11} & z_{x12} & \dots & z_{x1p} \\ z_{x21} & z_{x22} & \dots & z_{x2p} \\ \vdots & \vdots & \vdots & \vdots \\ z_{xp1} & z_{xp2} & \dots & z_{xpp} \end{bmatrix} \end{aligned}$$

and the amplitudes of $\boldsymbol{Y}$ are

$$\begin{aligned} \boldsymbol{Z}_{y(q\times p)} &= \boldsymbol{Y}\boldsymbol{P}_x^T \\ &= \begin{bmatrix} z_{y11} & z_{y12} & \dots & z_{y1p} \\ z_{y21} & z_{y22} & \dots & z_{y2p} \\ \vdots & \vdots & \vdots & \vdots \\ z_{yq1} & z_{yq2} & \dots & z_{yqp} \end{bmatrix} \end{aligned}$$

To put $\boldsymbol{Z}_x$ and $\boldsymbol{Z}_y$ on the same scale, divide each row by its respective root mean square.

For $i = 1, 2, \ldots, p, j = 1, 2, \ldots, p$, and $k = 1, 2, \ldots q$

$$z_{xij} = z_{xij} / \sqrt{\sum_{j=1}^{p} z_{xij}^2/(p-1)} \quad (2)$$

$$z_{ykj} = z_{ykj} / \sqrt{\sum_{j=1}^{p} z_{ykj}^2/(p-1)} \quad (3)$$

A measure of similarity between the amplitudes in $\boldsymbol{Z}_x$ and the amplitudes in $\boldsymbol{Z}_y$ is the absolute difference, or score:

$$\text{score}_{ikj} = |z_{xij} - z_{ykj}| \quad (4)$$

Now, we can directly compare $\boldsymbol{Z}_x$ and $\boldsymbol{Z}_y$ using the scores: we say the pattern of ith row of $\boldsymbol{Z}_x$ is *similar* to the pattern in kth row of $\boldsymbol{Z}_y$ if the score_{ikj} is small across all $p-1$ components. It is not enough to calculate the total score because we need the pattern to match across all components.

We must specify how similar is similar across the $p-1$ components while controlling for false positives. To choose the optimal window widths for the $p-1$ components, we use an iterative permutation procedure.

2.1 Permutation procedure

To avoid the traditional liability of indentifying spurious patterns from the data, we estimate the false positive rate (FPR) by breaking the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

There are two methods to do this. If we are concerned with an overall FPR, we permute $\boldsymbol{X}$ and leave $\boldsymbol{Y}$ fixed. If however, we want a FPR for each rowterm of $\boldsymbol{X}$, we can permute $\boldsymbol{Y}$ and leave $\boldsymbol{X}$ fixed. I will illustrate both methods.

Method 1: Control the overall FPR
for b in 1 to B (B large) {

independently shuffle the columns of $\boldsymbol{X} \rightarrow \boldsymbol{X}_b^*$

calculate $\boldsymbol{Z}_x^*, \boldsymbol{Z}_y^*$ and $\text{score}^*_{(ikj)b}$ using $\boldsymbol{X}_b^*, \boldsymbol{Y}$

}

Method 2: Control the FPR for each of the rowterms of $\boldsymbol{X}$
for b in 1 to B (B large) {

independently shuffle the rows of $\boldsymbol{Y} \rightarrow \boldsymbol{Y}_b^*$

calculate $\boldsymbol{Z}_x^*, \boldsymbol{Z}_y^*$ and $\text{score}^*_{(ikj)b}$ using $\boldsymbol{X}, \boldsymbol{Y}_b^*$

}

Now that we have a set of scores indicative of no relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$, we use an iterative procedure to determine optimal window widths for the $j = 1, 2, \ldots, p-1$ components (columns of $\boldsymbol{Z}_x, \boldsymbol{Z}_y$).

2.2 Iterative procedure

To compare $\boldsymbol{Z}_x, \boldsymbol{Z}_y$, we have to determine the optimal window widths for the $p-1$ components while controlling the FPR; we use the scores* calculated in the permutation procedure.

Set $\tau = 1$ and let $\hat{\sigma}_j$ be the estimated standard deviation of column j of $\boldsymbol{Z}_x$. We optimize the window width by iteratively increasing τ (narrowing the width of the window).

Method 1: Control the overall FPR
Since $\boldsymbol{X}$ is permuted, choose $i = 1$ (i doesn't matter).
Specify α and $J \in 1, 2, \ldots, p-1$ such that the estimated FPR $\leq \alpha$ by component J.

Step 1:
$\boldsymbol{w} = [\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_{p-1}]/\tau$

for b in 1 to B {

$g_1 = \{$all k rows of $\boldsymbol{Z}^*_{yk1}$ for which score$^*_{(1k1b)} \leq w_1\}$

calculate $\hat{\text{FPR}}_{1b} = |g_1|/q$

for j in 2 to $p-1$ {
$g_j = g_{j-1} \cap \{$all k rows of $\boldsymbol{Z}^*_{ykj}$ for which score$^*_{(1kj)b} \leq w_j\}$
calculate $\hat{\text{FPR}}_{jb} = |g_j|/q$ }

}

Step 2:
for j in 1 to $p-1$ {

calculate the estimated FPR for component j : $\hat{\text{FPR}}_j = \sum_{b=1}^{B} \hat{\text{FPR}}_{jb}/B$

}
if $\hat{\text{FPR}}_J \leq \alpha$, STOP
else, $\tau = \tau + .1$. Go to Step 1.

The optimal window width is the final $\boldsymbol{w}_{\text{opt}}$. Use the values of $\boldsymbol{w}_{\text{opt}}$ to do the same procedure for the real data, $\boldsymbol{Z}_x, \boldsymbol{Z}_y$, and map rowterms of $\boldsymbol{Y}$ onto rowterms of $\boldsymbol{X}$ (see Section 2.3).

Method 2: Control the FPR for each of the rowterms of $\boldsymbol{X}$
This method is more computationally expensive than Method 1 because we have to do the procedure for all rows of $\boldsymbol{X}$ however, it may be desirable to know the FPR for each of the rowterms of $\boldsymbol{X}$.

As before, set $\tau = 1$.
This time, choose α, i, and J such that the estimated FPR of rowterm $i \leq \alpha$ by component J.
Or, one could specify that all rowterms of $\boldsymbol{X}$ must have a FPR $\leq \alpha$ by component J, depending on the project. For illustration purposes, I will specify a rowterm i but the algorithm could be generalized.

Step 1:
$\boldsymbol{w} = [\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_{p-1}]/\tau$

for b in 1 to B {

for i in 1 to p {

$g_{i1} = \{$all k rows of $\boldsymbol{Z}^*_{yk1}$ for which score$^*_{(ik1b)} \leq w_1\}$

calculate $\hat{\text{FPR}}_{i1b} = |g_{i1}|/q$

for j in 2 to $p-1$ {

$g_{ij} = g_{i,j-1} \cap \{$all k rows of $\boldsymbol{Z}^*_{ykj}$ for which score$^*_{(ikj)b} \leq w_j\}$

calculate $\hat{\text{FPR}}_{ijb} = |g_{ij}|/q$

}

}

}

Step 2:
for i in 1 to p {

for j in 1 to $p-1$ {

calculate the estimated FPR for rowterm i and component j : $\hat{\text{FPR}}_{ij} = \sum_{b=1}^{B} \hat{\text{FPR}}_{ijb}/B$

}

if $\hat{\text{FPR}}_{iJ} \leq \alpha$, STOP

}

else, $\tau = \tau + .1$. Go to Step 1.

The optimal window width is the final $\boldsymbol{w}_{\text{opt}}$. Use the values of $\boldsymbol{w}_{\text{opt}}$ to do the same procedure for the real data, $\boldsymbol{Z}_x, \boldsymbol{Z}_y$, and map rowterms of $\boldsymbol{Y}$ onto rowterms of $\boldsymbol{X}$ (see Section 2.3).

2.3 Map rowterms of $\boldsymbol{Y}$ onto rowterms of $\boldsymbol{X}$

Using the optimal window width, $\boldsymbol{w}_{\text{opt}}$, calculated above (through either method), map rowterms of $\boldsymbol{Y}$ onto rowterms of $\boldsymbol{X}$.

for i in 1 to p {

$g_{i1} = \{$all k rows of $\boldsymbol{Z}_{yk1}$ for which score$_{ik1} \leq w_{\text{opt}1}\}$

for j in 2 to $p-1$ {

$g_{ij} = g_{i,j-1} \cap \{$all k rows of $\boldsymbol{Z}_{ykj}$ for which score$_{ikj} \leq w_{\text{opt}j}\}$

}

}

It is straightforward to also determine those rowterms of $\boldsymbol{Y}$ that are anti-associated with rowterms of $\boldsymbol{X}$. Just use $-\boldsymbol{Z}_x, \boldsymbol{Z}_y$ in the above procedure. Anti-associated elements will have the exact opposite pattern; when the value in $\boldsymbol{X}$ is high, the value in $\boldsymbol{Y}$ will be low, and vice versa.

3 Application

Reproductive biologists desire to know what genes are specific to particular substages of meiosis. To define the meiotic substage-specific transcriptome, we collected five samples of mouse germ cells at six time points: 8, 10, 12, 14, 16, and 18 days post partum (dpp), for a total of 30 samples. We completed cytological analysis and gene expression analysis (via RNA-seq) on each biological sample, resulting in two coupled datasets (Figure 1).

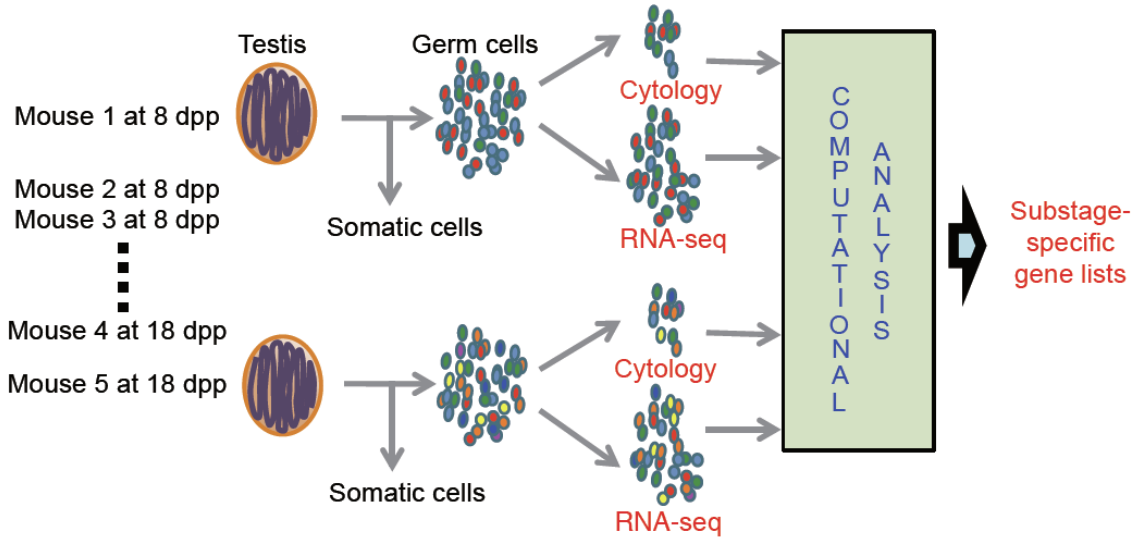

Figure 1: Experimental design.

The datasets show a high degree of concordance in the independent principal component analyses (PCAs). In Figure 2 the principal components were scaled to show them in the same figure.

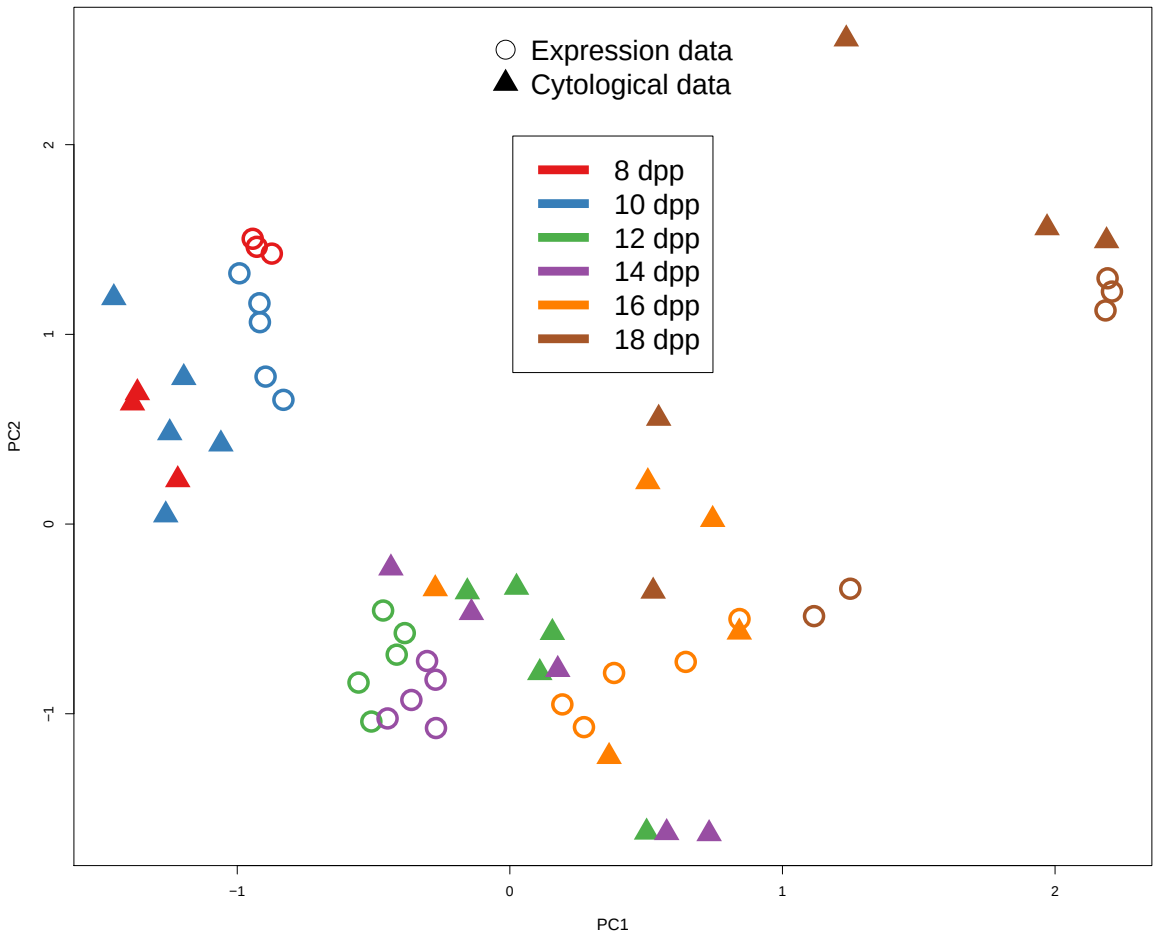

Figure 2: Plot of the first two principal components.

We subsequently discovered there was an inadequate amount of material for accurate gene expression analysis in two of the 8 day samples. These samples were excluded from further analysis, resulting in 28 samples.

Let $\boldsymbol{X}$ represent the 6 cytological proportions of interest and $\boldsymbol{Y}$ represent the expression levels of the 20,368 genes.

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1,28} \\ x_{21} & x_{22} & \dots & x_{2,28} \\ \vdots & \vdots & \vdots & \vdots \\ x_{61} & x_{62} & \dots & x_{6,28} \end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1,28} \\ y_{21} & y_{22} & \dots & y_{2,28} \\ \vdots & \vdots & \vdots & \vdots \\ y_{20368,1} & y_{20368,2} & \dots & y_{20368,28} \end{bmatrix}$$

Then, let $\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}$ be centered on the row average (Equation 1) and calculate the covariance matrix

$$\begin{aligned} \boldsymbol{C}_{6\times 20368} &= \tilde{\boldsymbol{X}}\tilde{\boldsymbol{Y}}^T/28 \\ &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T. \end{aligned}$$

Then, $\boldsymbol{P}_{x(6\times 20368)} = \boldsymbol{U}^T\tilde{\boldsymbol{X}}$ and

$$\begin{aligned}
\boldsymbol{Z}_{x(6\times 6)} &= \boldsymbol{X}\boldsymbol{P}_x^T \\
&= \begin{bmatrix} z_{x11} & z_{x12} & \cdots & z_{x16} \\ z_{x21} & z_{x22} & \cdots & z_{x26} \\ \vdots & \vdots & \vdots & \vdots \\ z_{x61} & z_{x62} & \cdots & z_{x66} \end{bmatrix} \\
\boldsymbol{Z}_{y(20368\times 6)} &= \boldsymbol{Y}\boldsymbol{P}_x^T \\
&= \begin{bmatrix} z_{y11} & z_{y12} & \cdots & z_{y16} \\ z_{y21} & z_{y22} & \cdots & z_{y26} \\ \vdots & \vdots & \vdots & \vdots \\ z_{y20368,1} & z_{y20368,2} & \cdots & z_{y20368,6} \end{bmatrix}
\end{aligned}$$

Calculate scores using Equations 2 and 4. We estimate the overall FPR according Method 1, and specify that $\hat{\text{FPR}} \leq .05$ by $j = 3$. The histogram of the $B = 1000$ estimated FPRs is shown in Figure 3. Note that using the average as the estimator results in a conservative estimate.

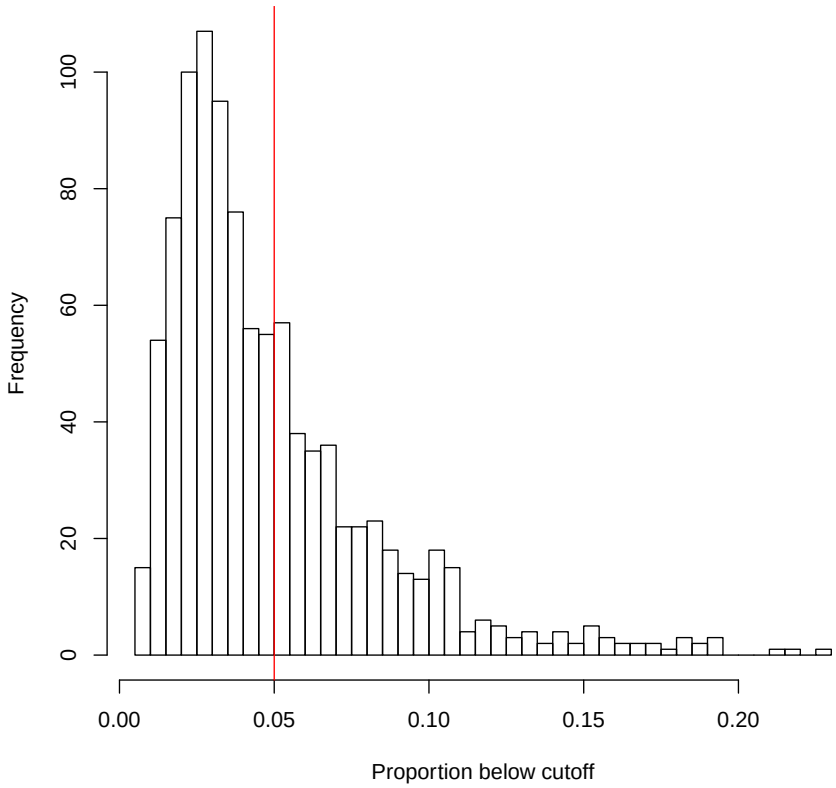

Figure 3: Histogram of the estimated FPR for $j = 3$. The red line is $\alpha = .05$.

The resulting estimated FPRs for all 5 components are given in Table 1.

Using the most specific lists of $j = 4, 5(\hat{\text{FPR}} \leq 0.005)$, the number of genes that are substage specific and the number of anti-associated genes for each of the substages is given in Table 2.

component (j)	1	2	3	4	5
$\hat{\text{FPR}}$	0.482	0.164	0.049	0.005	<0.001
$\boldsymbol{w}_{\text{opt}}$	1.435	0.694	0.544	0.168	0.063

Table 1: Estimated FPRs and optimal window widths.

Substage	Substage-specific	Anti-associated
Spermatogonia	1234	1694
Preleptotene	55	12
Early Leptotene	92	49
Late Leptotene+Zygotene	742	556
Eerly Pachytene	922	1068
Late Pachytene+Diplotene	4004	4366

Table 2: Substage-specific and anti-associated gene lists with $\hat{\text{FPR}} \leq 0.005$

We can compare the patterns of the cytological substage with the patterns of gene expression of the PMCA substage-specific and anti-associated lists. In Figures 4, 5, and 6, we examine the most striking pattern – the late pachytene and diplotene substage. Notice in Figure 4 that there are no late pachytene cells until 16 dpp, resulting in a flat pattern and then sharp increase at 16 and 18 dpp. The gene expression values of the PMCA late pachytene gene list (Figure 5) have the same pattern; the expression values do not change until 16 and 18 dpp, when they have a sharp increase in expression. Contrast the substage specific late pachytene gene expression with the anti-associated late pachytene gene expression in Figure 6. The expression levels of the anti-associated genes have the exact opposite pattern; they have relatively flat expression until day 16 and 18, when they sharply decrease. Not only do we know what genes are turning on in the substage but we also know what genes are turning off, or are silenced.

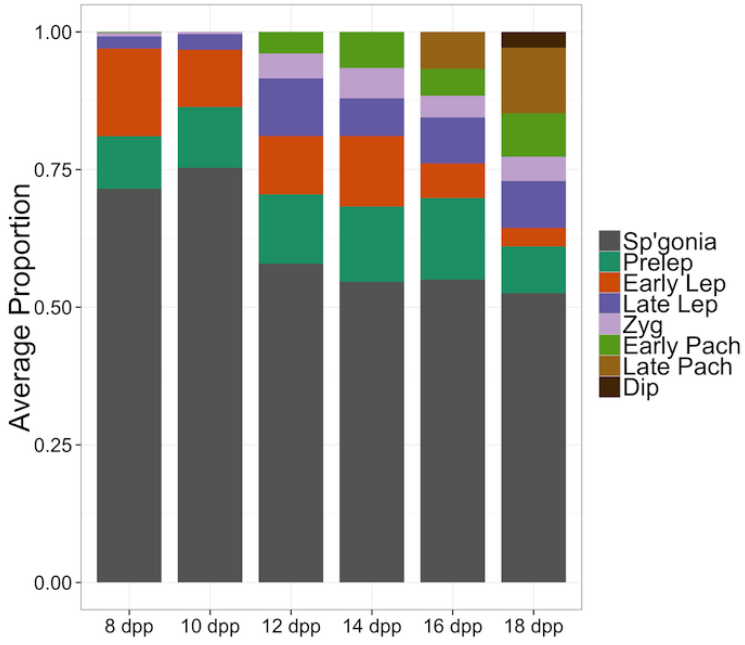

Figure 4: Average cytological proportions of the substages. The late pachytene and diplotenet substage proportions are in light and dark brown.

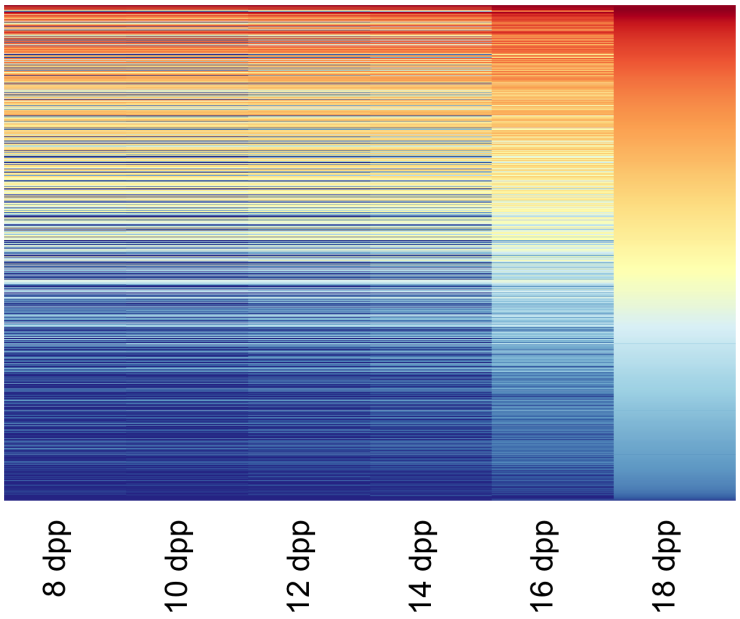

Figure 5: Gene expression ($\log_2$(TPM+1)) of the PMCA late pachytene+diplotene gene list.

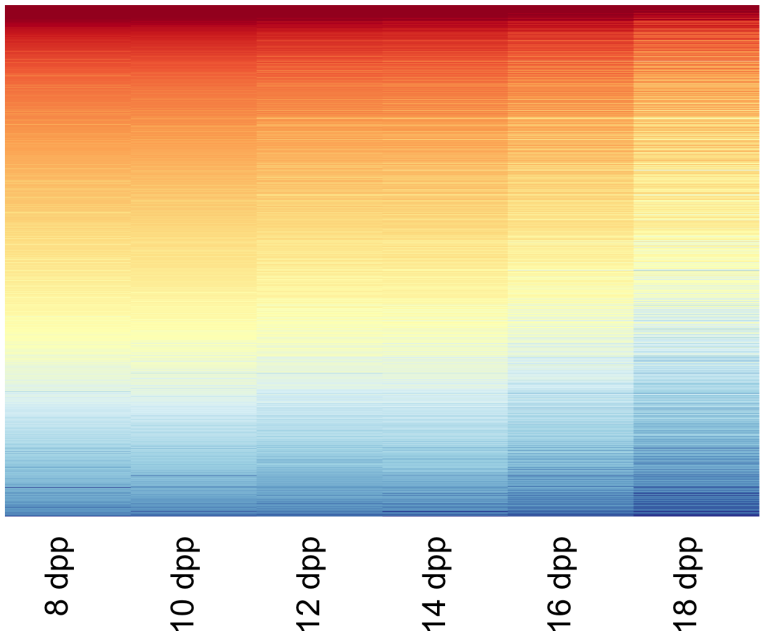

Figure 6: Gene expression ($\log_2$(TPM+1)) of the PMCA anti-associated late pachytene+diplotene gene list.

3.1 Validation

The validity of the PMCA procedure was queried first by analysis of genes expressed by a highly enriched substage cell population, pachytene spermatocytes from adult mice retrieved by unit gravity sedimentation, and second, by analysis of a known pattern of gene expression (X-chromosome silencing during meiotic prophase).

To further validate the substage transcriptomes revealed by RNA-seq and PMCA, we compared the substage-specific gene lists to genes identified by RNA-seq as expressed in four samples of highly enriched adult pachytene spermatocytes, obtained by sedimentation at unit gravity. Each sample was from germ cells pooled from testes of 6 mice at 9 weeks of age and the he purity of late pachytene/diplotene spermatocytes was about 90%. There was a highly significant enrichment (hypergeometric test) of pachytene spermatocyte-expressed genes among the gene lists from later meiotic substages, but there was no enrichment in the gene lists for spermatogonia or the early substages of meiosis (Table 3). In fact, 99% of the late pachytene/diplotene genes were also found in the enriched adult pachytene spermatocyte samples, thereby validating the PMCA procedure for identification of meiotic substage-specific gene expression.

Substage	Number of overlap	p-value
Spgonia	614 of 1235	1
Prelep	17 of 55	1
Early Lep	28 of 92	1
Late Lep+Zyg	550 of 742	1.2 x 10-09
Early Pach	809 of 922	2.0 x 10-62
Late Pach/Dip	3955 of 4004	<1.0 x 10-62

Table 3: Results of hypergeometric enrichment analysis.

3.2 In progress

Steve Cherry illustrated that traditional MCA can result in the identification of spurious patterns. He used 7 simulated datasets and I will use these same datasets to show that PMCA overcomes this liability. I am writing an R package so it will be easier for investigators to apply this technique.

We have used PMCA for other datasets and these results are forthcoming.

4 Summary

PMCA is a novel extension of MCA that overcomes the chief liability of traditional MCA, fitting spurious patterns, without the need for parametric assumptions. Furthermore, an acceptable FPR is specified and the investigator can choose to compute an overall FPR or to compute a FPR for each of the rowterms of $\boldsymbol{X}$, which may be of particular interest.

PMCA can be used in a number of ways in a variety of settings, depending on how the investigator frames the question. In this application, the "sample" was a biological sample and two measurements were taken on the same sample. However, there are other situations in which the "sample" need not be a physical sample but could be the element of interest, such as temperature or stress. Here the data could be two different ways of measuring the temperature or stress and PMCA can map one of the measurement methods onto the other method. This could be particularly useful when, for example, the measurements are taken on both the material and metal levels and the investigator may want to know what elements on the material level and metal level have the same stress pattern.

References

[1] Prohaska, J. (1976) A technique for analyzing the linear relationships between two meteorological fields. *Mon. Wea. Rev.*, **104**, 1345–1353

[2] Bretherton, C.S., Smith, C., and Wallace, J.M. (1992) An intercomparison of methods for finding coupled patterns in climate data sets. *J. Climate*, **5**, 541–560.

[3] Wallace, J.M., Smith, C., and Bretherton, C.S. (1992) Singular value decomposition of wintertime sea-surface-temperature and 500 mb height anomalies. *J. Climate*, **5**, 561–576.

[4] Cherry, S. (1996) Singular value decomposition and canonical correlation analysis. *J. Climate*, **9**, 2003–2009.

[5] Cherry, S. (1997) Some comments on singular value decomposition analysis. *J. Climate* **10** 1759–1761.

[6] Salim, A., Pawitan, Y., and Bond, K. (2005) Modelling association between two irregularly observed spatiotemporal processes by using maximum covariance analysis. *Appl. Statist.*, **54**(3), 555–573.

[7] Tan, C.S., Salim, A., Ploner, A., Lehtio, J., Chia, K.S., and Pawitan, Y. (2009) Correlating gene and protein expression data using correlated factor analysis. *Bioinformatics*, **10**:272.