A novel computational method integrates single cell RNAseq and bulk RNAseq analyses

by Robyn L. Ball & Xulong Wang (co-1st authors), Harriet Jackson, Gareth Howell, and Gregory Carter

To do list:

- PMCA R package (Robyn)
- immunological data as a well-characterized control (Xulong)
- experimental methods (Harriet and Gareth)
- More details on results (all)

1 Obtaining signals from single cell RNAseq (scRNAseq)

A total of 47 molecularly distinct cell classes were identified by single cell RNA-sequencing of 3005 cortical cells and BackSPIN clustering (Zeisel et al., 2015). To identify marker genes for each of the 47 cell types, we modeled each gene's expression profile across the 3005 cells with a generalized linear model, in which gene expression values were modeled with a negative binomial model, and the mean of the binomial model was modeled with two predictors: a 47 levels factor describing cell types and a basal variable describing the basal expression levels. Model parameters were estimated with Stan in the Bayesian framework.

Let y_{nm} represent the number of unique molecular identifiers (UMIs), or molecule counts, for gene m ($m = 1, 2, \ldots, M$) and cell n ($n = 1, 2, \ldots, 3005$). Also, let T_k be the set of cells that were clustered as cell type k for $k = 1, 2, \ldots, t$. Then,

$$y_{nm} = \mu_n + \sum_{k=1}^{47} \alpha_{mk} \mathbb{1}(n \in T_k) + e_{nm}$$

where $\boldsymbol{y}_m = [y_{1m}, y_{2m}, \ldots, y_{3005m}]^T$ and $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_{3005}]^T$ and $\boldsymbol{y}_m \sim \text{NB}\left[\boldsymbol{\mu} + \sum_{k=1}^{47} \alpha_{mk} \mathbb{1}(n \in T_k), \phi\right]$.

Bayesian methodology was used to estimate α_{mk} such that the estimate a_{mk} is the mode of the distribution.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1M} \\ a_{21} & a_{22} & \ldots & a_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ a_{t1} & a_{t2} & \ldots & a_{tM} \end{bmatrix}$$

(We need to specify the prior and posterior of α_{mk} that was used in the analysis.)

$\phi, \alpha_{mk} \sim$ Cauchy(0,10)

2 Map scRNAseq onto bulk RNAseq (bRNAseq) using permutation-based maximum covariance analysis (PMCA)

After obtaining gene expression signals from both scRNAseq and bRNAseq, we map celltypes in scRNAseq onto important groups in bRNAseq analyses. For example, groups could be ages, time points, genotypes, treatment groups, etc., for which there are multiple samples. For notation purposes, suppose there are g groups of interest, $G_1, G_2, \ldots, G_g$ with $n_1, n_2, \ldots, n_g$ replicates.

In bRNAseq experiments, differential expression analysis is conducted to determine the genes that are differentially expressed among the g groups. These M differentially expressed genes are used for the subsequent analysis. Define the bRNAseq signal for gene m of group g as the median divided by the median absolute deviation (MAD) of the group's replicates.:

$$x_{gm} = \text{median}(x_{gm,1}, x_{gm,2}, \ldots x_{gm,n_g})/(1.4826 * \text{MAD}) \tag{1}$$

Then, the bRNAseq signals for the g groups are:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1M} \\ x_{21} & x_{22} & \ldots & x_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ x_{g1} & x_{g2} & \ldots & x_{gM} \end{bmatrix} \tag{2}$$

First, center the matrices by subtracting the row averages. For $i = 1, 2, \ldots g, k = 1, 2, \ldots t, m = 1, 2, \ldots M$.

$$\begin{aligned} \tilde{x}_{im} &= x_{im} - \frac{1}{M}\sum_{m=1}^{M} x_{im} \\ \tilde{a}_{km} &= a_{km} - \frac{1}{M}\sum_{m=1}^{M} a_{km} \end{aligned} \tag{3}$$

so that

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \ldots & \tilde{x}_{1M} \\ \tilde{x}_{21} & \tilde{x}_{22} & \ldots & \tilde{x}_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{x}_{g1} & \tilde{x}_{g2} & \ldots & \tilde{x}_{gM} \end{bmatrix}, \tilde{\boldsymbol{A}} = \begin{bmatrix} \tilde{a}_{11} & \tilde{a}_{12} & \ldots & \tilde{a}_{1M} \\ \tilde{a}_{21} & \tilde{a}_{22} & \ldots & \tilde{a}_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{a}_{T1} & \tilde{a}_{T2} & \ldots & \tilde{a}_{TM} \end{bmatrix}$$

Let $\boldsymbol{C}_{(g\times t)} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{A}}^T/M$ be the covariance matrix of $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{A}}$ 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 the bRNAseq groups, $\boldsymbol{X}$, that explain the covariance and the column vectors of $\boldsymbol{V}$ correspond to the structures in the scRNAseq cell types, $\boldsymbol{A}$, that explain the covariance.

Because we are primarily interested in mapping the cell types of $\boldsymbol{A}$ onto the groups of $\boldsymbol{X}$, we consider $\boldsymbol{P}_{x(g\times M)} = \boldsymbol{U}^T\tilde{\boldsymbol{X}}$, the principal components of the covariance matrix that correspond to the groups of $\boldsymbol{X}$, and calculate the homogeneous and heterogeneous regressions using $\boldsymbol{P}_x$.

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

The group amplitudes, or "group patterns", of $\boldsymbol{X}$ are

$$\begin{aligned} \boldsymbol{Z}_{x(g\times g)} &= \boldsymbol{X}\boldsymbol{P}_x^T \\ &= \begin{bmatrix} z_{x11} & z_{x12} & \ldots & z_{x1g} \\ z_{x21} & z_{x22} & \ldots & z_{x2g} \\ \vdots & \vdots & \vdots & \vdots \\ z_{xg1} & z_{xg2} & \ldots & z_{xgg} \end{bmatrix} \end{aligned}$$

and the cell type amplitudes, or "cell type patterns", of $\boldsymbol{A}$ are

$$\begin{aligned} \boldsymbol{Z}_{A(t\times g)} &= \boldsymbol{A}\boldsymbol{P}_x^T \\ &= \begin{bmatrix} z_{a11} & z_{a12} & \cdots & z_{a1g} \\ z_{a21} & z_{a22} & \cdots & z_{a2g} \\ \vdots & \vdots & \vdots & \vdots \\ z_{at1} & z_{at2} & \cdots & z_{atg} \end{bmatrix} \end{aligned}$$

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

For $i = 1, 2, \ldots, g, j = 1, 2, \ldots, g$, and $k = 1, 2, \ldots t$

$$z_{xij} = z_{xij} / \sqrt{\sum_{j=1}^{g} z_{xij}^2/(g-1)} \tag{4}$$

$$z_{akj} = z_{akj} / \sqrt{\sum_{j=1}^{g} z_{akj}^2/(g-1)} \tag{5}$$

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

$$\text{score}_{ikj} = |z_{xij} - z_{akj}| \tag{6}$$

Now, we can directly compare $\boldsymbol{Z}_x$ and $\boldsymbol{Z}_A$ using the scores: we say the group pattern of ith group of $\boldsymbol{Z}_x$ is *similar* to the cell type pattern in kth cell type of $\boldsymbol{Z}_A$ if the score_{ikj} is small across all $g-1$ 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 $g-1$ components while controlling for false positives. To choose the optimal window widths for the $g-1$ components, we use an iterative permutation procedure.

2.1 Permutation procedure

To avoid the traditional liability of indentifyng spurious patterns from the data, we estimate the false positive rate (FPR) for each group by breaking the relationship between $\boldsymbol{X}$ and $\boldsymbol{A}$. (Note: we chose $B = 1000$.)

for b in 1 to B (B large) {

independently shuffle each row of $\boldsymbol{A} \rightarrow \boldsymbol{A}_b^*$

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

}

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

2.2 Iterative procedure

Set $\tau = \tau_0$ 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).

Choose α and J such that the estimated FPR of every group is $\leq \alpha$ by component J.

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

for b in 1 to B {

for i in 1 to g {

$H_{i1} = \{$all k rows of $\boldsymbol{Z}^*_{Ak1}$ for which score$^*_{(ik1b)} \leq w_1\}$

calculate $\hat{\text{FPR}}_{i1b} = |H_{i1}|/t$

for j in 2 to $g-1$ {

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

calculate $\hat{\text{FPR}}_{ijb} = |H_{ij}|/t$

}

}

}

Step 2:
for i in 1 to g {

for j in 1 to $g-1$ {

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

}

if $\forall \hat{\text{FPR}}_{iJ} \leq \alpha$, 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, $\boldsymbol{Z}_x, \boldsymbol{Z}_A$, and map cell types of $\boldsymbol{A}$ onto the groups of $\boldsymbol{X}$ (see subsection 2.3). Note that by taking the mean, the estimated FPR is a conservative estimate.

2.3 Map cellt ypes of $\boldsymbol{A}$ onto groups of $\boldsymbol{X}$

Using the optimal window width, $\boldsymbol{w}_{\text{opt}}$, calculated above, map cell types of $\boldsymbol{A}$ onto groups of $\boldsymbol{X}$.

for i in 1 to g {

$H_{i1} = \{$all k rows of $\boldsymbol{Z}_{ak1}$ for which score$_{ik1} \leq w_{\text{opt}1}\}$

for j in 2 to $g-1$ {

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

}

}

It is straightforward, and sometimes quite informative, to also determine those cell types of $\boldsymbol{A}$ that are anti-associated with groups of $\boldsymbol{X}$. Just use $-\boldsymbol{Z}_x, \boldsymbol{Z}_A$ in the above procedure. Anti-associated elements will have the exact opposite pattern; when the gene signal of the group is high, the gene signal of the cell type is low, and vice versa.

3 Application

Abnormal clusters of protein clusters in the brain, or plaques, are characteristic of Alzheimer's Disease patients. Mutations in the APP gene, thought to be responsible for brain plaques, can cause early-onset Alzheimer's Disease so a study was undertaken at the Jackson Laboratory by the Howell Lab to study the genetics of APP mutant mice as they age. A total of 59 mice, both normal wild-type mice (WT) and APP mutant mice were used in the study. Mice were sacrificed at 2, 4, 5, and 6 months (m) of age and whole brain samples were used for RNAseq analysis (Table 1). (Need more info from Gareth)

genotype	2 month	4 month	5 month	6 month
WT	5	9	13	7
APP	5	5	10	5

Table 1: Number of biological replicates in each group.

Differential gene expression analysis found sets of genes that were differential to genotype, age, and both age and genotype. From this group of differential genes, we comprised a set of 108 genes that were also expressed in the scRNAseq data for the 47 brain cell types. As described in Equations 1 and 2, we computed the gene signals for the following 10 groups: WT (all ages), WT2m, WT4m, WT5m, WT6m, APP (all ages), APP2m, APP4m, APP5m, and APP6m. Since there are 47 brain cell types in the scRNAseq data, a false positive rate (FPR) of 0.05 translates to 2.35 cell types. However, it is important to note that the estimated FPR, $\hat{\text{FPR}}$ is a conservative estimate so while care should be taken if the number of cell types falls below the estimated FPR, these cell types still warrant investigation.

The bRNAseq group data ($\boldsymbol{X}$) and the scRNAseq cell type data ($\boldsymbol{A}$) are illustrated in Equations 7 and 8.

Group data

$$\boldsymbol{X}_{10\times 108} = \begin{bmatrix} WT(Actn1) & WT(Akap5) & \dots & WT(Zdhhc9) \\ APP(Actn1) & APP(Akap5) & \dots & APP(Zdhhc9) \\ WT2m(Actn1) & WT2m(Akap5) & \dots & WT2m(Zdhhc9) \\ \vdots & \vdots & \vdots & \vdots \\ APP6m(Actn1) & APP6m(Akap5) & \dots & APP6m(Zdhhc9) \end{bmatrix} \tag{7}$$

Cell type data

$$Y_{47\times 108} = \begin{bmatrix} Astro1(Actn1) & Astro1(Akap5) & \dots & Astro1(Zdhhc9) \\ Astro2(Actn1) & Astro2(Akap5) & \dots & Astro2(Zdhhc9) \\ CA1Pyr1(Actn1) & CA1Pyr1(Akap5) & \dots & CA1Pyr1(Zdhhc9) \\ \vdots & \vdots & \vdots & \vdots \\ Vsmc(Actn1) & Vsmc(Akap5) & \dots & Vsmc(Zdhhc9) \end{bmatrix} \quad (8)$$

4 Results

We used PMCA as outlined in the Methods Section to map brain cell types onto groups of mice. The most compelling results are given in Table 2.

Group	$\hat{\text{FPR}}$	$\hat{\text{FPR}} * 47$	brain cell types
WT	0.001	0.047	**Int12,Epend**,Int5,Int7, CA1Pyr1,CA1PyrInt,CA2Pyr2
APP	0.001	0.047	**Int2,Int8,Int13**
WT2m	0.012	0.564	**ClauPyr,S1PyrL5,S1PyrL23,S1PyrL6**
WT4m	0.0004	0.02	**SubPyr,Pvm1,Astro2,Astro1**
WT6m	0.001	0.047	CA1Pyr2,CA2Pyr2,CA1Pyr1
APP4m	0.031	1.457	**Oligo2,Oligo6**
APP5m	0.006	0.282	CA1Pyr2,CA2Pyr2

Table 2: PMCA results. Brain cell types in **bold** are unique to the group. Brain cell types that are underlined are unique to older groups.

Interestingly, oligodendrocytes, subtypes 2 and 6, are specific to the APP 4 month mice. These mice do not yet have the characteristic brain plaques but we know they will have the plaques by 6 months of age. Oligodendrocytes are generally active early but are not thought to be active in normal mice past 2 months of age. Perhaps the enhanced activity of oligodendrocytes in the APP4m mice is an early precursor to the development of brain plaques. (More from Gareth).

Kegg pathway analysis on differential genes in subtype 2 and 6 of the oligodendrocytes shows particular enrichment of genes involved in cell adhesion. Cell type CA1Pyr2 is specific to older mice (WT6m and APP5m). Pathway analysis of differential genes in cell type CA1Pyr2 shows enrichment for Alzheimer's Disease, Parkinson's Disease, and Huntington's Disease. Perhaps this cell type is more specific to aging in general. (More analysis of other cell types needed).

Possible hypothesis that we need to investigate include:

- Oligo2 and Oligo6 are specific to APP4m (precursor to phenotype?)
- CA1Pyr2 is involved in aging in both WT and APP mice
- Int2, Int8, and Int13 are specific to APP mice (all ages)
- Int12 and Epend are specific to WT (is there something important that APP mice lack?)
- SubPyr,Pvm1,Astro2,and Astro1 are specific to WT4m (are these critical for health compared to APP?)