Measuring APA from single cell RNA-Seq with precision weights

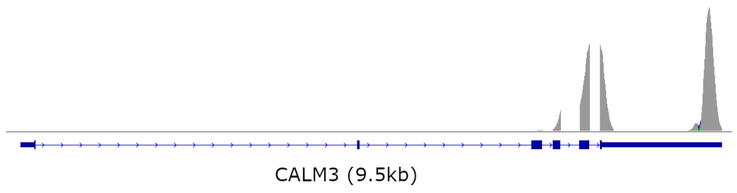
Paul Harrison[1] @paulfharrison, Sarah Williams[1], David Albrecht[2], David Powell[1], and Traude Beilharz[3]

[1]Monash Bioinformatics Platform, Monash University [2]Faculty of Information Technology, Monash University [3]Biomedicine Discovery Institute, Monash University

Bonjour! We're using this multi-assay data to explore multivariate techniques [1] which are new to us. Let's talk ideas!

Alternative Polyadenylation (APA)

Measuring APA provides novel information about cell state in addition to RNA expression level. Different lengths of 3' UTR for a transcript may contain different sets of regulatory elements.

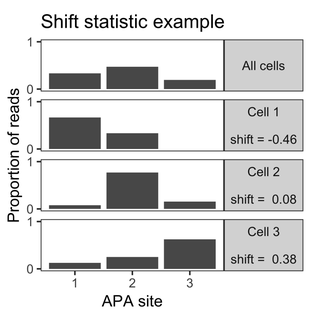


We use 10x Genomics single-cell RNA-Seq data, which amplifies RNA sequence immediately before the poly(A) tail, specifically the 10x supplied PBMC-8k dataset [2].

In some reads the sequence proceeds into the poly(A) tail itself, allowing APA sites to be located to base-level accuracy.

We found 339 genes with 2 or more suitable APA peaks.

Gene-level shift statistic



Define shift scores $y_{i,j}$ summarizing site usage for a single gene in a single cell as a single number between -1 and 1.

Estimated shift score is **unbiassed, but noisy when there are few reads**.

For each gene $i \in 1..n_{\text{gene}}$, and for each cell $j \in 1..n_{\text{cell}}$ cells j, and for each site $k 1..n_{,i}, observe the UMI count $u_{i,j,k}$. The proportion site usage is:

$$p_{i,j,k} = \frac{u_{i,j,k}}{\sum_{k'=1}^{n_{\text{site}\,i}} u_{i,j,k'}}$$

We compare this to the average over all cells, $\bar{p}_{i,k}$, omitting cells with zero for a particular gene. Define a shift score for an individual UMI based on the proportion of UMIs downstrand minus the proportion of UMIs upstrand

$$s_{i,k} = \sum_{k'=1}^{n_{\text{site}\,i}} \text{sign}(k'-k)\bar{p}_{i,k'}$$

Define the shift for a particular cell and gene as the mean over each UMI

$$y_{i,j} = \sum_{k=1}^{n_{\text{site}\,i}} p_{i,j,k} s_{i,k}$$

Precision weights

Similar to `voom` for gene expression [3], we estimate precision weights $w_{i,j}$ for each shift $y_{i,j}$ based on the number of reads.

A precision weight as used here is 1 over the variance. A weight of 0 indicates missing data (no reads).

An estimate of the variance each UMI contributes is

$$\hat{\sigma}_i^2 = \text{E}_i(s_{i,k}^2) = \sum_k \bar{p}_{i,k} s_{i,k}^2$$

The shift is an average over individual UMI scores, so we initially estimate the weight as

$$w_{i,j} = \frac{\sum_k u_{i,j,k}}{\hat{\sigma}_i^2}$$

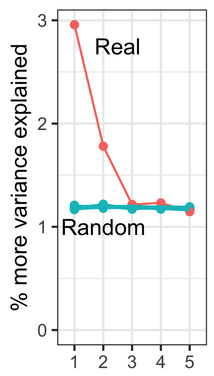
then use Maximum Likelihood with a Principal Components model to make adjustments:

- $\hat{\sigma}_i^2$ is an over-estimate, as some proportion of the variance can be modelled.
- Allow for biological variation by placing a soft maximum on the effective number of reads. However the PBMC-8k data showed little need for this.

Weighted Principal Components

$$\underset{\text{genes}\times\text{cells}}{Y} = \underset{\text{genes}\times\text{components}}{A} \quad \underset{\text{components}\times\text{cells}}{B^\top} \quad + \underset{\text{genes}\times\text{cells}}{\varepsilon}$$

$$\underset{\text{original}}{Y} = \underset{\text{loadings}}{A} \quad \underset{\text{scores}}{B^\top} \quad + \underset{\text{noise}}{\varepsilon}$$



Maximum Likelihood solution sought by Criss-cross Weighted Least Squares [5].

Parallel Analysis indicated each component will account for ~1.2% of variance by chance. 2 components clearly exceed this.

Varimax rotation seeks sparse loadings, ideally separating distinct biological processes.

- In addition to columns for components, we include a column of all 1s in B. The corresponding column in A is the (weighted) mean for each gene.
- This approach is similar to matrix factorization used in recommender systems [4], which work with big data and with a high proportion of missing data.
- Parallel Analysis requires randomized versions of the data to compare against. We leave the weights unchanged, and draw random values

$$y_{i,j} \sim \mathcal{N}\left(0, \frac{1}{w_{i,j}}\right)$$

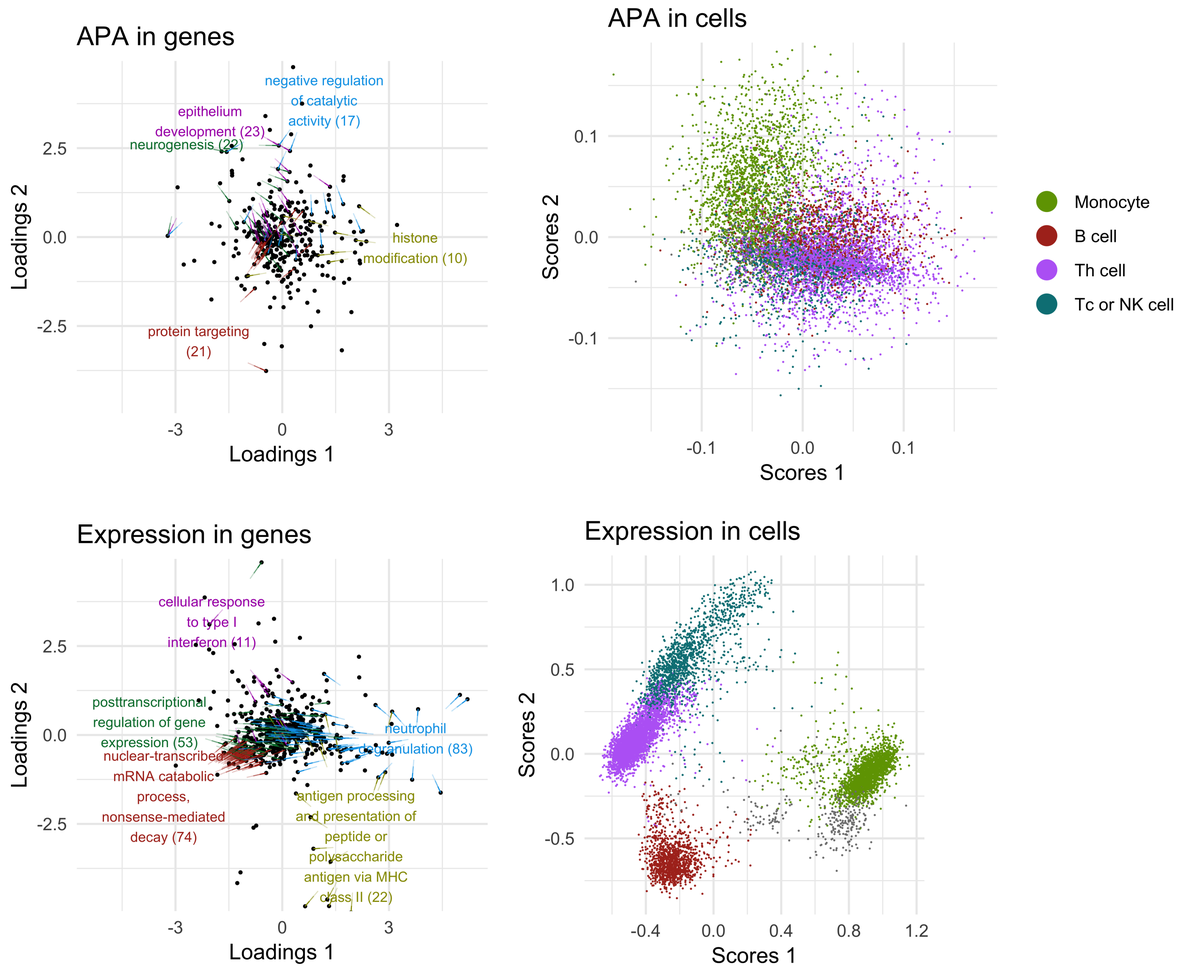
Weighted $\log_2$ gene expression

Parallel Analysis supports 9 components for gene expression.

Filtered for genes with more than $n_{\text{cell}}/2$ total UMIs. There were 644 such genes. Transformed and weighted with `limma::vooma(edgeR::cpm(counts, log=TRUE, prior.count=0.25), design=B)`.

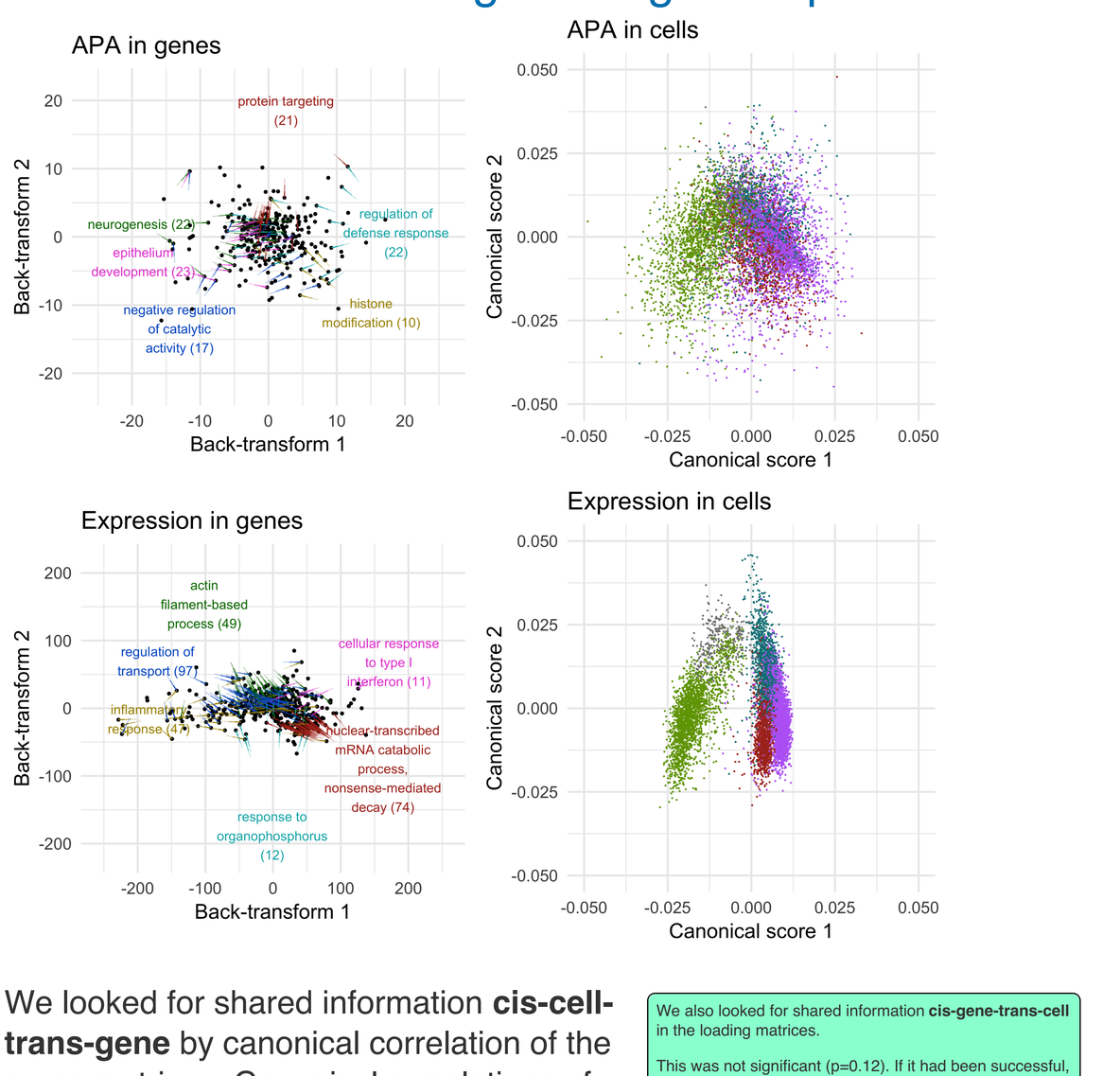
Use of a prior count necessarily introduces bias. In general, issues around weighting here seem more complex than for APA!

Biplots



With this many variables, we put the axes (genes, left) and points (cells, right) in separate plots. Cell types were identified using our `celaref` Bioconductor package [6], using the supplied "graph" clustering and cell type labels from [7] as reference. Gene Ontology terms sampled by an ad-hoc method.

Canonical correlation gives aligned biplots



We looked for shared information **cis-cell-trans-gene** by canonical correlation of the score matrices. Canonical correlations of 0.76 and 0.41 are supported (Wilks' Λ with Rao's F approximation, p<<0.001 for both).

We also looked for shared information **cis-gene-trans-cell** in the loading matrices.

This was not significant (p=0.12). If it had been successful, it would show a set of genes regulated by two distinct mechanisms.

The first canonical scores distinguish monocytes from other cells, and the second canonical scores highlight an interesting commonality between a subpopulation of monocytes and T_C and NK cells. 2 directions in the space of variation in gene expression most associated with variation in APA are also found, pointing to linked biological processes between APA and expression.

References

[1] S. Holmes, "Multivariate data analysis: The French way," in *Probability and Statistics: Essays in Honor of David A. Freedman*, Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2008, pp. 219–233 [Online]. Available: http://projecteuclid.org/euclid.imsc/1207580085
[2] 10x Genomics, "8k PBMCs from a Healthy Donor." [Online]. Available: https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc8k. [Accessed: 30-Jan-2019]
[3] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "voom: precision weights unlock linear model analysis tools for RNA-seq read counts," *Genome Biology*, vol. 15, no. 2, p. R29, Feb. 2014.
[4] Y. Koren, R. Bell, and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.
[5] K. R. Gabriel and S. Zamir, "Lower Rank Approximation of Matrices by Least Squares With Any Choice of Weights," *Technometrics*, vol. 21, no. 4, pp. 489–498, Nov. 1979.
[6] S. Williams, *celaref: Single-cell RNAseq cell cluster labelling by reference*. 2019 [Online]. Available: https://bioconductor.org/packages/release/bioc/html/celaref.html
[7] G. X. Y. Zheng *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, p. 14049, Jan. 2017.