Examining cell quality in association with specific cell types for droplet-based single cell RNA-Sequencing

Shila Ghazanfar[1,2], Jean Y.H. Yang[2,1], Ellis Patrick[2]
[1]*Judith and David Coffey LifeLab, Charles Perkins Centre, The University of Sydney*
[2]*School of Mathematics and Statistics, The University of Sydney*

@shazanfar
@TheEllisPatrick
@sydneybioinfo

Droplet scRNA-Seq data contains more information than you think

Motivating data

Bach *et al* (2017) present data on epithelial cells in the mouse mammary gland, across **four developmental stages** in adulthood, resulting in 8 runs of 10x Genomics scRNA-Seq data. They identified a set of 23,184 high quality cells and labelled them into **eight main cell types**:

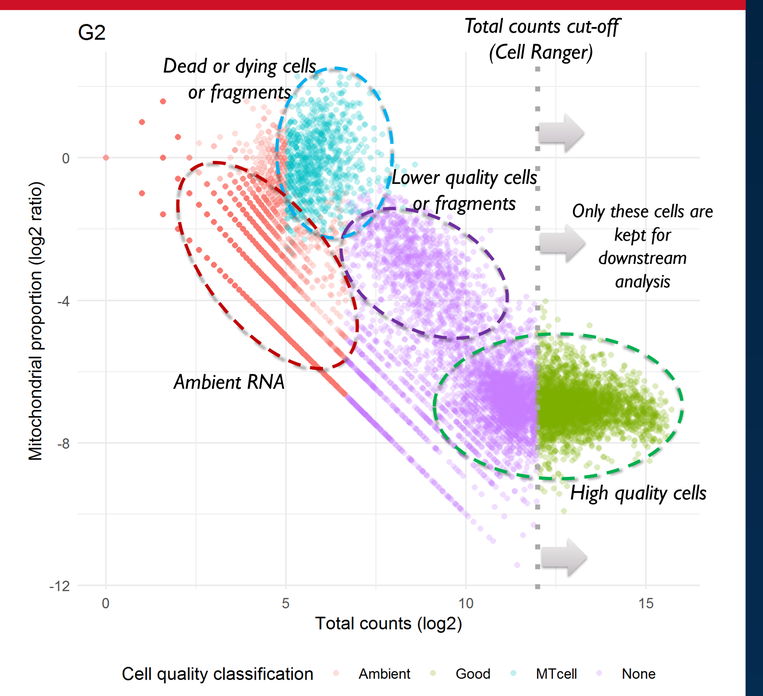
Avd – Alveolar differentiated cells
Avp – Alveolar progenitor cells
Bsl – Basal cells
Hsd – Hormone sensing differentiated
Hsp – Hormone sensing progenitors
Lp – Luminal progenitor
Myo – Myoepithelial cells
Prc – Procr+ basal cells

Summary and significance

We will demonstrate that cell quality selection in droplet scRNA-Seq can have a drastic effect on the make-up of the cell types per sample.
Significance: This can have big consequences for downstream analysis like differential expression and differences in cell type proportions between comparison groups.

Breakdown of cell barcodes

For a given droplet scRNA-Seq experiment, we can typically identify sets of high and lower quality cells, cell barcodes with dead or dying cells, as well as cell barcodes containing ambient RNA.
We suggest initial selection of cell barcodes using emptyDrops (Lun *et al*, 2018) followed by removal of cells with very high mitochondrial gene proportion. We identify these cells by clustering on total counts, mitochondrial proportion, and ribosomal proportion, and removing the cluster with the lowest counts and highest mitochondrial proportion.



Cell type prediction for low count data

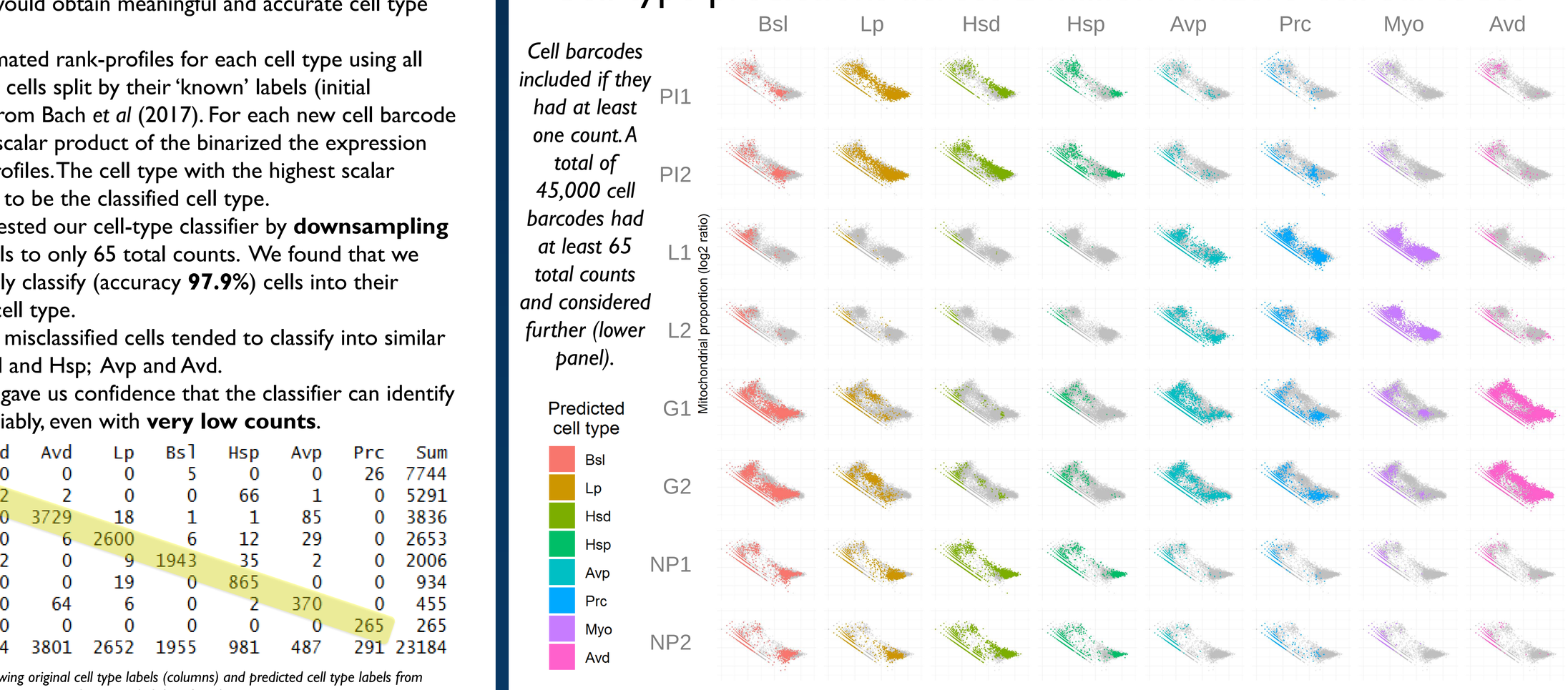
Aim: To predict the cell type for all cell barcodes, but first need to establish that we would obtain meaningful and accurate cell type predictions.
Method: We estimated rank-profiles for each cell type using all 23,184 high quality cells split by their 'known' labels (initial clustering results from Bach *et al* (2017). For each new cell barcode we then took the scalar product of the binarized the expression vector and rank-profiles. The cell type with the highest scalar product was taken to be the classified cell type.
Evaluation: We tested our cell-type classifier by **downsampling** the high quality cells to only 65 total counts. We found that we could still accurately classify (accuracy **97.9%**) cells into their originally labelled cell type.
We found that the misclassified cells tended to classify into similar categories, e.g. Hsd and Hsp; Avp and Avd.
Conclusion: This gave us confidence that the classifier can identify these cell types reliably, even with **very low counts**.

	Myo	Hsd	Avd	Lp	Bsl	Hsp	Avp	Prc	Sum
Myo	7713	0	0	0	5	0	0	26	7744
Hsd	0	5222	2	0	0	66	1	0	5291
Avd	2	0	3729	18	1	1	85	0	3836
Lp	0	0	6	2600	6	12	29	0	2653
Bsl	15	2	0	9	1943	35	2	0	2006
Hsp	0	50	0	19	0	865	0	0	934
Avp	13	0	64	6	0	2	370	0	455
Prc	0	0	0	0	0	0	0	265	265
Sum	7743	5274	3801	2652	1955	981	487	291	23184

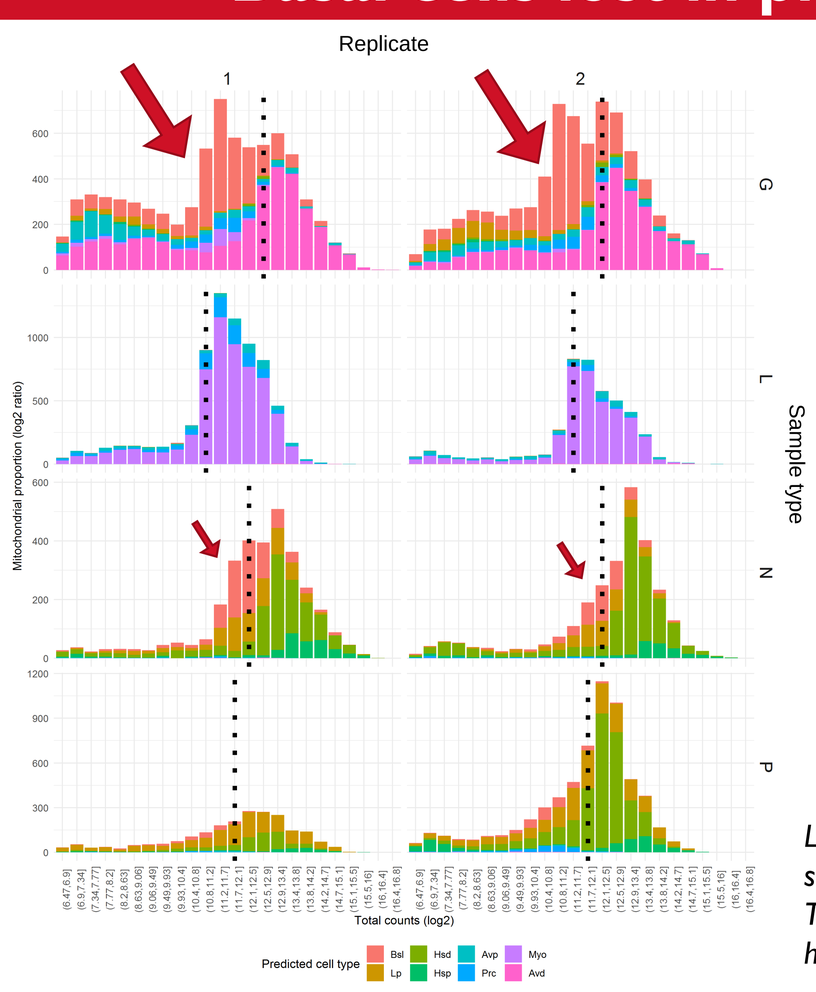
Confusion matrix showing original cell type labels (columns) and predicted cell type labels from downsampled data (rows)

Cell type predictions across 2 million nonzero cell barcodes

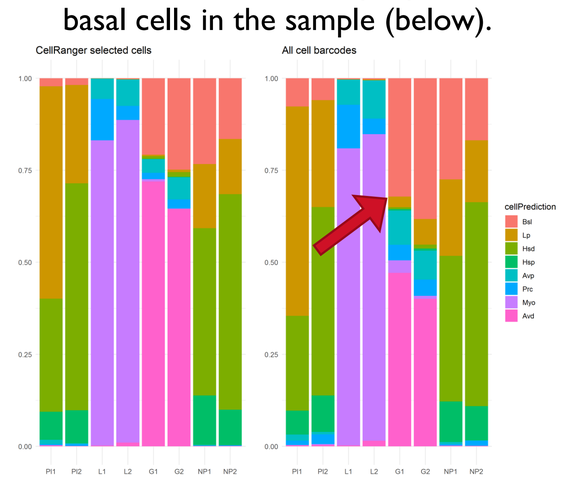
Cell barcodes included if they had at least one count. A total of 45,000 cell barcodes had at least 65 total counts and considered further (lower panel).



Basal cells lost in preprocessing



With all cell barcodes labelled with putative cell types, we observe that basal cells exhibit much lower total counts than other cell types like alveolar differentiated cells. In the gestational time point we observe many basal cells are lost in preprocessing (left of dotted line). This affects the estimated proportion of basal cells in the sample (below).



Left: Histogram of classified cells, with dotted line showing the cut-off employed for Cell Ranger Top: Barplots showing cell type proportions for high-quality cells only and for all cell barcodes

Future work

- Incorporate a **confidence score** for cell type classification: Evaluate if there is enough information in the damaged/low quality cells to be able to accurately classify cell type?
- Downstream impact: bias correction for **cell proportion** testing
- Experimental **quality control**: When have key cell types been compromised in the experimental set up?

Conclusion

This work immediately points to the potential for discovering more information from droplet scRNA-Seq data than simply selecting based on total counts. Utilising all cell barcodes could result in more accurate estimates of cell type proportions among samples, especially important when aiming to incorporate estimates or perform inference of cell type proportion or total cell type expression across various comparison groups.

References

Bach, K. *et al.* Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat. Commun.* **8,** (2017).
Lun, A. *et al.* Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *bioRxiv* 234872 (2018). doi:10.1101/234872

We sincerely thank John Marioni, Joseph Powell, and Karsten Bach for their support and comments on this project.
Contact: shila.ghazanfar@sydney.edu.au jean.yang@sydney.edu.au ellis.patrick@sydney.edu.au