Cellbender module

Leland Taylor

December 4, 2020

1 Example methods

We used CellRanger v3.0.2 to demultiplex reads, align reads to GRCh38 (GRCh38-3.0.0 reference file distributed by 10X Genomics, which corresponds to Ensembl version 93 transcript definitions), and generate feature-barcode matrices.

To identify droplets containing cells and adjust the raw counts matrix for background, ambient transcript contamination from the single cell experiment (i.e., the soup), we used CellBender v2.1 (Fleming et al., 2019). For training, CellBender requires rough estimate of the number of droplets containing cells (cell droplets) and the number of droplets without cells (empty droplets) derived from the UMI curve—the rank ordering droplet barcodes according to total UMI counts (x axis) by the total number of UMI counts per droplet (y axis). To estimate the cell droplet threshold, we calculated the UMI curve, selected droplets with a UMI count >1,000, and estimated the threshold using the "barcoderanks-inflection" procedure from DropletUtils v1.9.16 (Lun et al., 2019). To estimate the number of empty droplets, we calculated the UMI curve as before, selected droplets with a UMI count between 250 and 10, and estimated the threshold by performing both the "barcoderanks-inflection" and "barcoderanks-knee" procedure from DropletUtils—using 1/3rd of the distance between the two estimates as the final threshold. We ran CellBender with the default parameters apart from excluding droplets with <10 UMI counts from analysis (–low-count-threshold) and using 300 epochs with a learning rate of 0.0000001 to fit a model with 50 latent variable dimensions. We adjusted the final counts matrix for the ambient soup signature at a false positive rate (rate at which a true signal count is erroneously removed) of 0.01, the CellBender default.

2 Notes

(1) If you run into an error during training, decrease the learning rate by an order of magnitude. Continue to do so until the error goes away. If this does not work, try to decrease zdim to $\tilde{2}0$. (2) You will need to play around with the threshold identification to fine tune the estimators for your dataset.

References

Fleming, S. J., Marioni, J. C., and Babadi, M. (2019). CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. *bioRxiv*, page 791699. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Lun, A. T. L., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., Marioni, J. C., and participants in the 1st Human Cell Atlas Jamboree (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology*, 20(1):63.