Drop-Seq Analysis Pipeline Overview

Guilherme de Sena Brandine

March 29, 2017

Drop-Seq is a high throughput single cell sequencing methods that has been gaining popularity and being continuously optimized in the last 2 years. The reads are divided in two Fastq files: The first end read is 20bp and composed of a unique cell barcode appended to a unique molecular identifier, the second end read is from a cDNA reverse-transcribed by one of the mRNA molecules in the cell.

1 Upstream Analysis

The upstream analysis is done differently whether the data comes from the traditional drop-seq protocol [1] [2] or from 10X genomics [3].

1.1 Standard Protocol

1.1.1 Library Quality Control

In the traditional protocol we use FastQC [4] on the second end read to assess the Illumina read quality and overall sequence biases. If necessary we use Trim Galore [5] to trim out possible adapter contamination.

1.1.2 Mapping

We further use the software package developed by the McCarroll lab [1] to create BAM files that, for every mapped read, attaches the cell/molecule barcode as metadata and further uses this to create a count table with barcodes as columns and genes as rows. Reads with equal UMIs and mapped to the same gene are counted as one and cell barcodes with less than 1000 reads are discarded.

1.2 10X Data

The CellRanger software from 10X proceeds in a similar fashin as described above, but their barcode standards are different. They have adapters between cell and molecular barcodes that overcome potential ambiguities in sequencing errors and they have a list of valid barcodes as reference. Each sequenced cell barcode is associated with a whitelist barcode based on sequence similarity.

2 Downstream Analysis

Most of the methods described below (eccept for the mapped reads quality control and ontology analysis) are implemented in the Seurat package [6].

2.1 Mapped Reads Quality Control

To estimate the percentage of the gene expression that is present on each cell we apply the Good-Turing estimate [7] to find the expected value of the percentage of genes that are expressed but not present in the library. A density plot of these values gives us a qualitative indication of saturation for each cell. Good libraries have most of the cells in the 60-90% saturation range.

2.2 Normalization and PCA

We use the $\log(1 + tpm)$ normalization to stabilize the mean-variance dependence across genes. To overcome technical noise, we do PCA on the mean-centered data and use the JackStraw [8] method to assess the statistical significance of each PC. There is often a clear cutoff between signal and noise PCs based on the JackStraw test p-value, so we reject PCs that have $p > 10^{-3}$.

2.3 Clustering and Subpopulation Discovery

We use the aforementioned PCs to cluster cells using the Louvain-Jaccard method proposed by Shekhar et al [9]. The method requires the construction of a k nearest neighbors graph using the Euclidean Distance between cells. The value of k is chosen as to minimize the gap statistic. If necessary, we further validate the clusters using t-SNE [10] and visualize known marker genes associated with cell phenotypes expected to be present in our dataset.

2.4 Differential Expression

For each cluster i, we use a likelihood ratio test proposed by McDavid et al [11] to test the hypothesis that the average expression on cluster i is greater than that of all cells except cluster i.

2.5 Ontology Analysis

To infer potential cluster phenotypes, we use the TopGO [12] package to assess ontologies enriched in the top differentially expressed genes for each cluster using the entire set of DE genes from all clusters as background. We use Fisher's Exact Test and select nodes spanning from the Biological Process (BP) root. Leaf nodes with small p-values are a strong indication of the underlying phenotypes. If biological processes that aren't of interest (e.g. cell cycle, metabolism, protein translation) are driving the clusters, the analysis may be redone using linear correction on genes associated with such processes.

References

Adrian Alexa and Jorg Rahnenfuhrer. topgo: enrichment analysis for gene ontology. *R package version*, 2(0), 2010.

Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

Neo Christopher Chung and John D Storey. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, 31(4):545–554, 2015.

Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, pages 237–264, 1953.

Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

F Krueger. Trim galore!: A wrapper tool around cutadapt and fastqc to consistently apply quality and adapter trimming to fastq files, 2015.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Andrew McDavid, Greg Finak, Pratip K Chattopadyay, Maria Dominguez, Laurie Lamoreaux, Steven S Ma, Mario Roederer, and Raphael Gottardo. Data exploration, quality control and testing in single-cell qpcr-based gene expression experiments. *Bioinformatics*, 29(4):461–467, 2013.

Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.

Karthik Shekhar, Sylvain W Lapan, Irene E Whitney, Nicholas M Tran, Evan Z Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z Levin, James Nemesh, Melissa Goldman, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5):1308–1323, 2016.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *bioRxiv*, page 065912, 2016.