Chapter 5. Overview

5.2 Experimental Design

* Choice of technology
  + Droplet-based
    - De facto, low cost per cell
  + Plate-based
    - Can capture other phenotypic information (like morphology)
    - More amenable to customization
  + Plate-based with reads
    - Whole-transcript coverage
* How many cells should be captured and to what depth
  + Droplet-based has a trade-off between throughput and doublet rate that affects the true efficiency of sequencing
* There should be multiple biological replicates for each condition and conditions should not be confounded with batch

5.3 Obtaining a count matrix

* A count matrix containing the number of UMIs or reads mapped to each gene in each cell
* First do quantification
* Import the count matrix into R and create a SingleCellExperiment object

5.4 Data Processing and downstream analysis

* Workflow
  + Compute quality control metrics to remove low quality cells that would interfere with downstream analyses
  + Convert the counts into normalized expression values to eliminate cell-specific biases (in capture efficiency)
  + Perform feature selection to pick a subset of interesting features for downstream analysis
  + Dimensionality reduction
  + Cluster cells into groups

Chapter 6. Quality Control

6.1 Motivation

* Low quality
  + Form their own distinct clusters
  + Distort the characterization of population heterogeneity during variance estimation or PCA
  + Contain genes that appear to be strongly upregulated (the process of increasing the response to a stimulus specifically)

6.2 Choice of QC metrics

* Library size
  + Defined as the total sum of counts across all relevant features for each cell
  + Cells with small library sizes are of low quality as the RNA has been lost at some point during library preparation
* The proportion of reads mapped to spike-in transcripts is calculated relative to the total count across all features (including spike-ins) for each cell.
  + High proportions are indicative of poor-quality cells where endogeneous RNA has been lost
* In the absence of spike-in transcripts, the proportion of reads mapped to genes in the mitochondrial genome can be used.
* Function to use: perCellQCMetrics()
* QC metrics are independent of the biological state of each cell

6.3 Identifying low-quality cells

6.3.1 With fixed thresholds

* Apply thresholds on the QC metrics
* Thresholds can vary from run ot run

6.3.2 With adaptive thresholds

* Identifying outliers
  + Based on the median absolute deviation (MAD) from the median value of each metric across all cells
  + More than 3 MADs from the median
  + Can apply log transformation
  + Is Outlier(), quckPerCellQC()
* Assumptions of outlier detection
  + Most cells are of acceptable quality
  + QC metrics are independent of the biological state of each cell
* Considering experimental factors
  + The adaptive strategy should be applied to each batch separately
  + Parameter batch=
    - Assuming each batch are of high quality
  + Identify problematic batches: find batches with QC thresholds that are themselves outliers compared to the thresholds of other batches

6.4 Checking diagnostic plots

* Distributions of QC metrics
* Plotting proportion of mitochondrial counts against some of the other QC metrics
* Comparison of the ERCC and mitochondrial percentages

6.5 Cell calling for droplet data

6.5.1 Background

* Droplet-based data
  + no prior knowledge about whether a particular library corresponds to cell-containing or empty droplets

6.5.2 Testing for empty droplets

* function: emptyDrops()
  + To test whether the expression profile for each cell barcode is significantly different from the ambient RNA pool
  + Uses Monte Carlo simulations to compute p-values for the multinomial sampling transcripts from the ambient pool.
  + Assumes that barcodes with low total UMI counts are empty droplets

6.5.3 Relationship with other QC metrics

* emptyDrops(): Possible for droplets to contain damaged or dying cells

6.6 Removing low-quality cells

* can either remove or mark the low-quality cells
* the true technical quality of a cell may also be correlated with its type

6.7 Marking low-quality cells

* aim: allow clusters of low-quality cells to form
* give opportunity to decide whether a cluster of such cells represents a genuine biological state.