Differential Expression Analysis for Single-cell RNA-seq

The Data Lab

Why differential expression (DE) analysis with single-cell?

- Bulk tissue collection is a mixture of cell types, so gene expression differences between samples may be dependent on changes in cell type composition
- Additionally, changes in gene expression may not be due to the cell type we are interested in
- With single-cell, we can narrow in on a specific population of interest and identify differentially expressed genes in that specific population across a set of samples
- Focusing on a single population of cells minimizes the effect changes in cell type composition has on differentially expressed genes

The Do's and Don'ts of Differential Expression

Do:

- Compare expression of genes in specific cell types across sample groups (e.g., between two treatment types)
- Identify differentially expressed genes in a subpopulation of cells across sample groups

Don't:

- Try to perform differential expression on a subpopulation/cell type that is not found among all samples of interest
- Perform differential expression without replicates! You should have *at least* 3 samples for each group being compared.

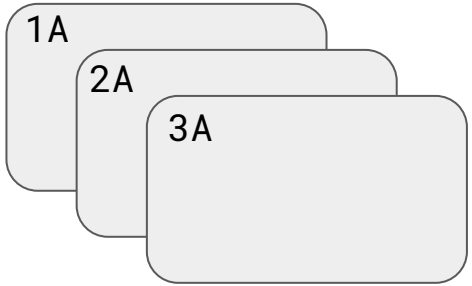
Differential expression analysis starts with good experimental design

- The same rules of bulk RNA-seq DE analysis apply – you need biological replicates!
- Cells are NOT replicates
  - Cells from one sample are more similar to each other than to cells from different samples
  - The goal is to highlight variation between samples not between cells in a sample
- Differential expression is not the same as identifying marker genes for clusters or cell types for a single sample
  - ***Marker genes*** - Identifying genes that are specific markers of a single group of cells within a single sample (e.g. CD4 is expressed in T-cells)
  - ***Differential expression*** - Identifying genes that are differentially expressed in a group of cells between two conditions (e.g. genes expressed in stimulated T-cells vs. unstimulated T-cells)

The goal of differential expression is to identify genes that are differentially expressed between two groups

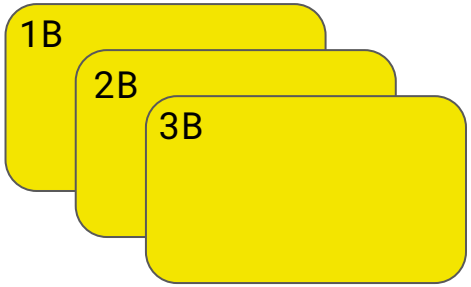
Single-cell counts matrices
≥3 samples per group

Group A - Control

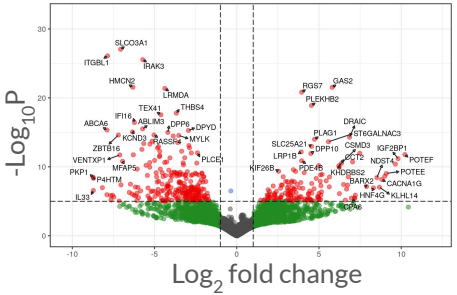


vs.

Group B - Treated



Use gene expression counts to identify DE genes



How do we perform DE analysis on single-cell data?

There are plenty of tools that exist for DE analysis in bulk RNA-seq (DESeq2, EdgeR, limma, etc.), so why can't we just use the same tools?

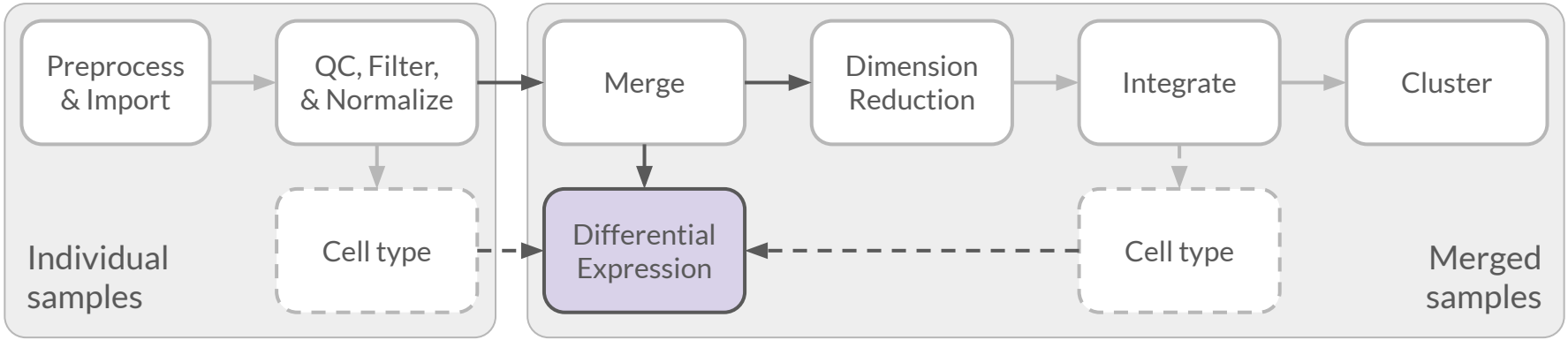
- High drop-out rate in single-cell, low gene expression counts
- Cells are treated independently, masking variation across the sample population
- Correlation of gene expression within cells from the same sample is unaccounted for

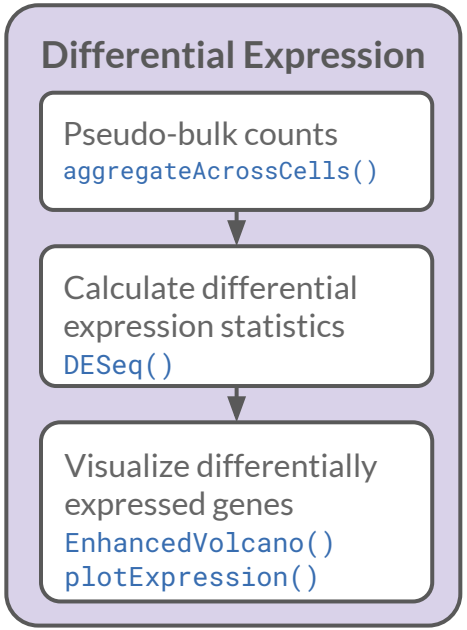
How do we perform DE analysis on single-cell data?

Potential solutions:

- **Calculate pseudo-bulk counts prior to using DE methods developed for bulk RNA-seq**
- Fitting a mixed-effects model to consider both drop-out and correlation between cells from the same sample (computationally intensive)
- Test for differences between expression distributions of a gene across a group of cells rather than between mean gene expression values

Differential Gene Expression Analysis in scRNA-seq Data between Conditions with Biological Replicates - 10x Genomics





- Pseudo-bulk
  - Collapse gene expression counts by sample and cell type rather than individual cells
  - Treats each sample rather than each cell as a replicate
- Differential expression statistics
  - Set up differential expression between two groups of samples
  - Use tools used for bulk RNA-seq analysis, like DESeq2, to identify differentially expressed genes between groups
- Visualize
  - Look at all differentially expressed genes using a volcano plot
  - Plot expression of individual genes across samples and subpopulations to validate results

Creating a pseudo-bulk SCE object

colData

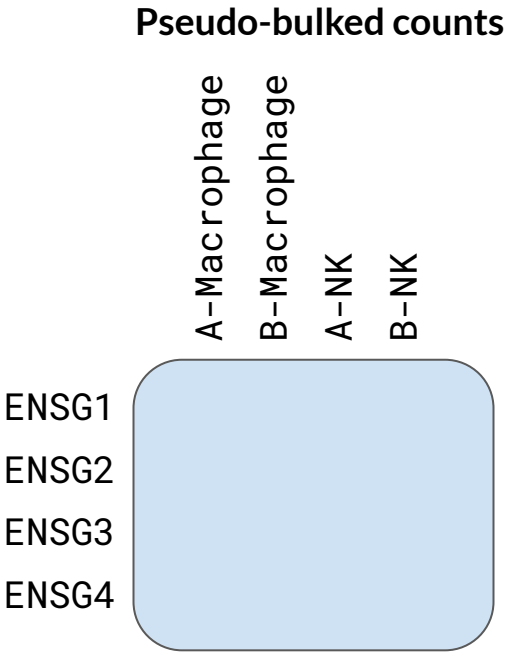
	sample	celltype
A-cell_1	A	Macrophage
A-cell_2	A	Macrophage
A-cell_3	A	NK
A-cell_4	A	NK
B-cell_5	B	Macrophage
B-cell_6	B	Macrophage
B-cell_7	B	NK
B-cell_8	B	NK

counts

	A-cell_1	A-cell_2	A-cell_3	A-cell_4	B-cell_5	B-cell_6	B-cell_7	B-cell_8
ENSG1	merged counts							
ENSG2								
ENSG3								
ENSG4								

Sum counts for each gene across all cells in desired group to use for differential expression (e.g., sample and celltype)

Creating a pseudo-bulk SCE object



- The resulting object will have one column for each group of cells and one row for each gene
- Similar to bulk RNA-seq we now have a `sample` by `gene` counts matrix rather than `cell` by `gene` matrix

Why do we pseudo-bulk?

1. Produces larger and less sparse counts so we can use standard methods for normalization and DE
2. Collapses gene expression counts by samples so samples rather than cells represent replicates
3. Masks variance within a sample to emphasize variance across samples

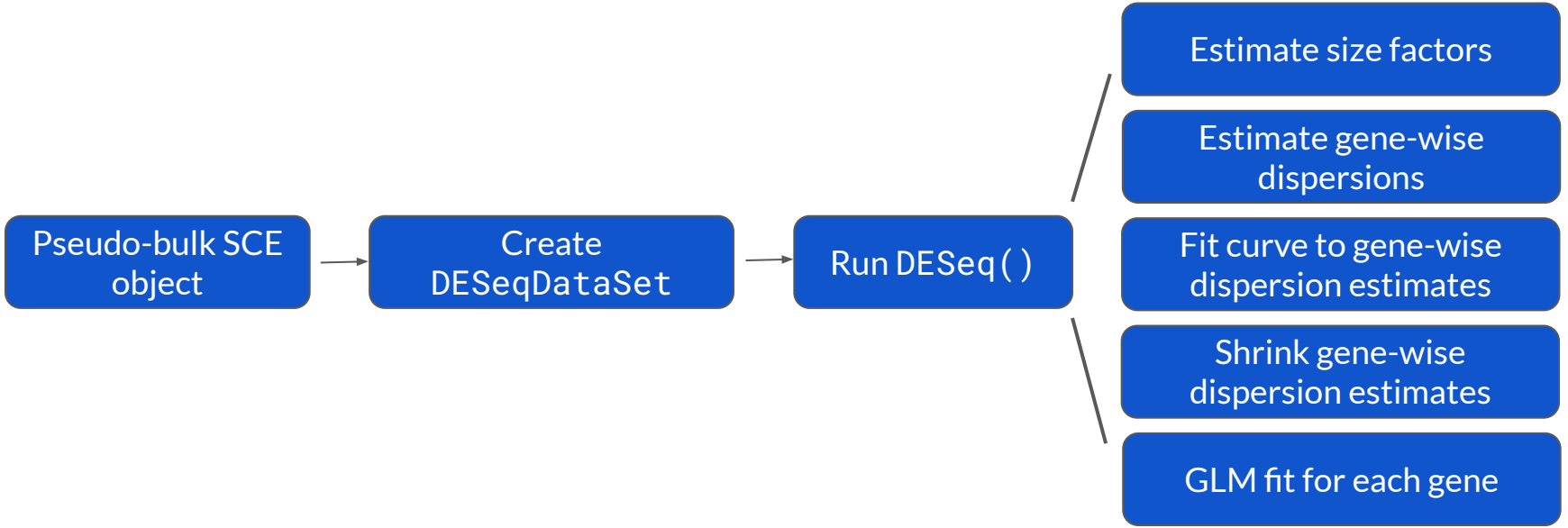
Using DESeq2 for differential expression analysis

Input: Pseudo-bulk counts matrix based on the DE experiment design

- Create the pseudo-bulked counts matrix using the raw, uncorrected, unnormalized counts
- Looking at a specific cell type between treatment groups? Group by sample, cell type, and treatment

Output: Table of genes with associated log2-fold change, p-value, and adjusted p-value

Using DESeq2 for differential expression analysis



Modified from HBC Training Intro to Single Cell RNA-Seq

Some caveats of single-cell differential expression:

- Differential expression comes with some light circularity
  - When we pick cell types or groups of cells to perform DE, we typically have picked those groups based on expression of a subset of genes
  - This may mean we miss differences between samples, especially if those differences are large enough to change a cell label

- Use raw counts, not corrected gene expression data!
  - Correction/integration will transform the data so that between sample variation is not preserved, sometimes resulting in negative gene expression values
  - `DESeq2` has been optimized for count data such that normalization and correction will affect the distribution in ways that may not be compatible with the model