Chapter 9. Dimensionality reduction

9.1 Overview

* scRNA-seq analysis: involve comparing cells based on their expression values across multiple genes
  + scRNA: a small conditional RNA is a small RNA molecule or complex engineered to interact and change conformation conditionally in response to cognate molecular inputs
  + Two genes, 2D plot where each axis represents the expression of one gene and each pt in the plot represents a cell
  + Thousands of genes, each cell’s expression profile defines its location in the high-dimensional expression space
* Dimensionality reduction: reduce the number of separate dimensions in the data
  + Possible because different genes are correlated if they are affected by the same biological process

9.2 Principal components (?) analysis (PCA)

* Discovers axes in high-dimensional space that capture the largest amount of variation
  + Imagine each axis as a line, move all cells in our data set onto this line by the shortest path (?), the variance captured by this axis is defined as the variance across cells along that line
  + The first axis (principal component) is chosen such that it captures the greatest variance across cells.
  + Next PC orthogonal to the first.
* Top PCs capture the dominant factors of heterogeneity in the data set.
  + Then perform dimensionality reduction by restricting downstream analyses to the top PCs.
* Assumption: biological processes affect multiple genes in a coordinated manner.

9.3 Choosing the number of PCs

9.3.2 Using the elbow point

* Identify the elbow point in the percentage of variance explained by successive PCs
* Sharp drop

9.3.3 Using the technical noise

* Retain all PCs until the percentage of total variation explained reaches some threshold T.
* Perform best when the mean-variance trend reflects the actual technical noise

9.3.4 Based on population structure

* Number of subpopulations (?) in the data
  + Each subpopulation differs from the others along a different axis in the high-dimensional space, set d to be the number of unique subpopulations minus 1, which guarantees separation of all subpopulations while retaining as few dimensions (and noise) as possible. (?)
* Normally don’t know subpopulation number

9.3.5 Putting it Together

* After choosing d, use reducedDim() to reassign the subsetted matrix back into the SingleCellExperiment object.
* See demo

9.4 Dimensionality reduction for visualization

9.4.2 Visualizing with PCA

* Plot the top 2 PCs
  + Each point is a cell
* Problem: a linear technique
  + Fail to resolve some subpopulations
  + Soln: pairwise plots

9.4.3 t-stochastic neighbor embedding (t-SNE)

* Find a low-dimensional representation of the data that preserves the distances between each point and its neighbors in the high-dimensional space.
* Repeat the visualization several times to ensure that the results are representative
* Set the seed to ensure that the chosen results are reproducible
* Perplexity
  + Low perplexity: favor resolution of finer structure
  + Test different perplexity values

9.4.4 Uniform manifold approximation and projection (UMAP)

* Non-linear
* Tries to find a low-dimensional representation that preserves relationships between neighbors in high-dimensional space
* Difference between UMAP and t-SNE
  + Tends to have more compact visual clusters with more empty space
  + Preserve more of the global structure than t-SNE
  + Faster than t-SNE
* Parameters
  + Number of neighbors
  + Minimum distance between embedded points

9.4.5 Interpreting the plots

* t-SNE plot can be used for a diagnostic inspection of the clustering output
* plot most useful for checking whether two clusters are actually neighboring subclusters or whether a cluster can be split into further subclusters