Ben Anderson
Project Outline
3/25/2022

Effects of different normalization methods for spatial transcriptomics using the Spatial LIBD dataset

Problem: Normalization of transcriptomics data is a critical aspect of the RNA-seq preprocessing workflow, however, normalization of spatial transcriptomics is currently understudied. This study aims to assess the effect of at least 4 normalization methods by measuring the difference between the cluster assignment found in the paper versus the same data processed with a different normalization method.

Methods: I plan to assess the 4 suggested normalization methods in R (TPM, scran, scTransform, and Dino), assuming that they are amenable to these spatial data. I may also extend analysis to methods available in Python. Finally, I may also assess whether the normalization significantly changes the conclusions of their differential expression for certain genes.

Deliverables: For assessing the quality of normalization, I will assume that the Layer 1-6/white matter assignment given to each spot via their clustering method is a generally reliable ground truth. If a normalization method leads to a significant deviation of cluster assignment for a spot that is in the middle of a layer, then this will be a red flag. Overall, I would expect to see at least 90% of all spots to have the same cluster assignment. I hypothesize that spots at boundaries—i.e. boundaries between layers, or a boundary from background to tissue—will be more prone to different layer assignment depending on normalization. I will attempt to specially monitor those spots for changes in cluster assignment or cluster probability (if available).

Challenges: I have run into problems using the data from the SpatialLIBD package in R, mostly due to unfamiliarity with the language. To avoid the problem of R language unfamiliarity inhibiting learning about RNA-seq, I would like to extend the analysis to Python as much as possible. To that end, I downloaded and wrangled the data into a convenient form in Python: sparse matrix of counts and multi-indexes for the row and column metadata. I'm in the process of identifying any Python packages with RNA-seq normalization methods.

I'm working on identifying how best to index into certain spots, specifically to identify the spot indices that correspond to being on a layer boundary. Unless I'm missing some metadata, I may have to generate indices of boundary spots with custom code. I would develop a metric of "number of spots away from a layer boundary" with 0=spot on boundary, 1=one spot away, etc.

I'm unsure how computationally intensive the normalization and clustering methods will be on these data, so my plan for number of replicates is uncertain. I may do 12, or only 3 depending on time constraints. The same consideration applies to the clustering methods. The authors performed PCA, UMAP and a nearest-neighbors approach, and I assume that PCA is fastest and UMAP and NN will be slower. Whether I repeat the UMAP and NN clustering on normalized data depends on computational feasibility on my local computer.