Presenter: Lan Zhang

Authors: Lan Zhang, Andrew McDavid

Title: MAKING OUT-OF-SAMPLE PREDICTIONS FROM UNSUPERVISED CLUSTERING OF SINGLE CELL RNASEQ

Abstract:
Background: A common goal in single cell RNA sequencing is to categorize subtypes of cells (observations) using unsupervised clustering on thousands of gene expression features. Each input cell is assigned a discrete label, interpreted as a cellular subpopulation. However, it has been challenging to characterize the robustness of the clustering, because most of the steps do not directly provide out-of-sample predictions. Methods: We introduce extensions to the steps in a common clustering workflow (i.e feature selection of highly variable genes, dimension reduction using principal component analysis, Louvain community detection) that allow out-of-sample prediction. These are implemented as wrappers around the R packages SingleCellExperiment and scran. The data is partitioned into a training set, where the workflow parameters are learned, and a test set where parameters are fixed and predictions are made. We compare the clustering of a set of observations in training vs test using the Adjusted Rand Index (ARI), which is a measure of the similarity between two data clusterings that ranges from 0 and 1. Result: We illustrate the approach using cells from the mouse brain originally published in Zeisel et al. 2015. We compare the impact on clustering concordance when splitting the cells into test/train subsets either a) uniformly at random or b) stratified by biological replicates (mice). Although we found agreement of clustering (approx. 0.80 ARI), the number of identified subpopulations was less stable. The ARI was further reduced (approx. 0.68) when our held out-data consisted of independent biological replicates. Conclusion: Typical clustering workflows contain steps that only implicitly learn various parameters. Formalizing the estimation of these implicit parameters allows quantification of the sensitivity of the clustering to changes in the input data, and can interrogate the generalizability of cell population discoveries made using single cell RNA-seq data.