Speaker: Joshua Welch, University of Michigan 2/26/2020

Quantitative Definition of Cellular Identity by Single-Cell Multi-Omic Integration

Recent experimental advances have enabled high-throughput single-cell measurement of gene expression, chromatin accessibility and DNA methylation, providing an opportunity to identify cell types and states from molecular information in a principled, quantitative fashion. However, most existing technologies can measure only one modality per cell, making it challenging to link datasets that share neither identical features nor the same instances. An additional challenge is the growing scale of single-cell datasets: For instance, it is no longer feasible to use the entire available datasets as inputs to implement standard pipelines on a personal computer with limited memory capacity. Moreover, there is a need for an algorithm capable of iteratively refining the definition of cellular identity as efforts to create a comprehensive human cell atlas continually sequence new cells.

In this talk, I will describe how metagene factors inferred by integrative nonnegative matrix factorization (iNMF) provide quantitative definition of cellular identity and its variation across biological contexts, allowing robust and scalable integration of highly heterogeneous single-cell datasets. I will also present an online learning algorithm for integrating massive and continually arriving single-cell datasets. To derive an online iNMF algorithm, we extended previous online learning approaches for NMF to minimize the expected cost of a surrogate function. Our online approach accesses the training data as mini-batches, decoupling memory usage from dataset size and allowing on-the-fly incorporation of new data as it is generated. The online implementation of iNMF converges much more quickly using memory independent of dataset size, without sacrificing solution quality. Our new approach enables factorization of 939489 single cells from 9 regions of the mouse brain on a single core of a standard laptop in $\sim$ 30 minutes using less than 1 GB of RAM. Furthermore, we construct a multi-modal cell atlas of the mouse motor cortex by iteratively incorporating seven single-cell datasets from three different modalities generated by the BRAIN Initiative Cell Census Network over a period of two years.

Our approach obviates the need to recompute results each time additional cells are sequenced, dramatically increases convergence speed, and allows processing of datasets too large to fit in memory. Most importantly, it facilitates continual refinement of cell identity as new single-cell datasets from different biological contexts and data modalities are generated.