Bayesian Non-Parametric Models for Analyzing Single-Cell RNA Sequencing Data

Skanda Koppula (skoppula@mit.edu), Karren Yang (karren@mit.edu)

6.882 Project Pre-Proposal, March 11, 2017

Cancer cells injected in mice grow into tumors composed of distinct sub-populations of cells, which exhibit patterned levels of gene expression. In order to characterize sub-populations of tumor cells and identify useful gene modules, we propose applying Bayesian nonparametrics to model single-cell RNA-sequencing (scRNA-seq) data from mouse tumors. Parametric topic models based on matrix factorization have shown promise in identifying meaningful clusters of cells and genes in heterogeneous cell populations; however, the need to pre-determine the number of clusters is a major limitation, particularly when extending these models to time-series data, as sub-populations of cells may appear, disappear, or diverge in expression over time. We propose exploring the utility of nonparametric models based on the Hierarchical Dirichlet Process (HDP) and the Indian Buffet Process (IDP) for modeling scRNA-seq data from single time points. We plan to compare their performance to that of existing methods for scRNA-seq analysis, as well as that of the top parametric topic model, Latent Dirichlet Allocation (LDA). Time permitting, we would also like to explore the extension of these models to time-series data, in order to understand how sub-populations of cells in the tumor microenvironment evolve over time.