Abstract: Statistical and computational approaches for identifying cell types and classifier genes from massive single-cell RNA-sequencing datasets, Y1

Understanding how neural circuits mediate animal behaviors is a fundamental problem in neuroscience, one that requires an inventory of the cell types comprising these biological circuits and experimental access to specific cell types to elucidate their roles in neural circuit function. With the advent of high-throughput single cell transcriptome profiling, the past few years have witnessed an explosion of new information on the complexity of cell types in the nervous system based on genes expressed by individual cells [reviewed in ref. 1]. In addition to providing a principled basis with which to create a taxonomy of cell types in the brain, knowledge of the genes expressed by specific neurons also provides information to target genetically encoded reporters and actuators to these newly discovered cell types [2]. An avalanche of data is expected from research consortia -- including the BRAIN Initiative Cell Census Network -- whose goal is to create a comprehensive census of cell types in the mouse brain using single-cell RNA-sequencing (scRNA-seq) approaches. New statistical methods will be needed to analyze these large datasets -- containing upwards of millions of cells -- to classify cells based on their transcriptomes and identify biomarker genes that can be used to identify and interrogate them experimentally. We will develop new statistical approaches for data normalization, clustering and biomarker identification that can be scaled for analyzing large scRNA-seq datasets. As a related case, we will be working with single-cell RNA-sequencing data from the olfactory epithelium. We have previously used scRNA-seq to understand how olfactory stem cells contribute to the remarkable regenerative capacity of this neurogenic tissue. We are currently revisiting these experiments at higher cellular resolution, to better distinguish different cell types in the regenerating tissue. These will provide an additional dataset for the development of normalization and clustering methods.