Gene expression imputation and clustering with batch effect removal in single-cell RNA-seq analysis by deep learning

Mingyao Li

University of Pennsylvania
E-mail: ahuwsj@126.com

Abstract: A primary challenge in single-cell RNA-seq (scRNA-seq) analysis is the ever increasing number of cells, which can be thousands to millions in large projects such as the Human Cell Atlas. Identifying cell populations becomes challenging in these data, as many existing scRNA-seq clustering methods cannot be scaled up to handle such large datasets. For large data, it is desirable to learn cluster-specific gene expression signatures from the data itself. Another challenge in large-scale scRNA-seq analysis is batch effect, which refers to systematic gene expression difference from one batch to another. Failure to remove batch effect can obscure downstream analysis and interpretation of results. In this talk, I will present a method for scRNA-seq analysis that enables gene expression imputation and clustering simultaneously through the use a deep learning algorithm. We further extend this method to incorporate known cell type information from a well-labeled source dataset through the use of transfer learning, a machine learning method that transfers knowledge gained from one problem to a different but related problem. Through comprehensive evaluations across many datasets generated in different tissues, species and protocols, we show that our methods can significantly improve clustering accuracy as compared to existing methods, and is capable of removing complex batch effects while maintaining true biological variations. We expect that, with the increasing growth of single-cell studies, our methods will offer a useful set of tools for clustering of these data.