University of California, Irvine
Statistics Seminar

ClusterDE: A Post-clustering Differential Expression (DE) Method Robust to False-positive Inflation Caused by Double Dipping

Jessica Jingyu Li
Professor
Departments of Statistics and Human Genetics and Biomathematics
UCLA

4 p.m., Thursday, February 22, 2024
6011 DBH

In typical single-cell RNA-seq (scRNA-seq) data analysis, a clustering algorithm is applied to find discrete cell clusters as putative cell types, and then a statistical test is employed to identify the differentially expressed (DE) genes between the cell clusters. However, this common procedure suffers the ``double dipping" issue: the same data are used twice to find discrete cell clusters as putative cell types and DE genes as potential cell-type marker genes, leading to false-positive cell-type marker genes even when the cell clusters are spurious. To overcome this challenge, we propose ClusterDE, a post-clustering DE method for controlling the false discovery rate (FDR) of identified DE genes regardless of clustering quality, which can work as an add-on to popular pipelines such as Seurat. The core idea of ClusterDE is to generate real-data-based synthetic null data containing only one cell type, in contrast to the real data, for evaluating the whole procedure of clustering followed by a DE test. Using comprehensive simulation and real data analysis, we show that ClusterDE has solid FDR control and the ability to identify canonical cell-type marker genes as top DE genes, distinguishing them from common housekeeping genes. Notably, the DE genes identified by ClusterDE are informative markers for discrete cell types and can guide the merging of spurious clusters. ClusterDE is fast, transparent, and adaptive to a wide range of clustering algorithms and DE tests.