我院教授张景肖及学生余小康就scRNA-seq数据集整合问题在《Nature Communications》发文

时间：2023-02-24

我院教授张景肖及学生余小康在《Nature Communications》发表论文，其中中国人民大学统计学院博士生余小康和中央财经大学统计与数学学院讲师许欣怡为该文章共同第一作者，昌平实验室李向杰副研究员和本院张景肖教授为通讯作者。该研究主要针对scRNA-seq数据中广泛存在的批次效应问题，提出了一种新的整合方法scDML。以往的研究中，绝大多数的单细胞整合算法流程都被设计成先消除批次效应，后进行数据集的聚类分群，这种做法可能会导致整合过程中稀有细胞类型的丢失。为此，我们提出了一种基于度量学习的深度学习模型(Deep Metric Learning)来整合单细胞数据集，在该方法中，scDML首先对预处理后的单细胞数据集进行高分辨率的聚类初始化，然后通过计算数据集内部和数据集之间的邻居信息来度量类间的相似度，并且针对该类相度矩阵设计了一种基于分层聚类的合并算法，最终通过训练深度学习模型来消除批次效应。在多个模拟和实际数据集的测试结果表明，在消除批次效应的同时，scDML能准确识别真实和稀有的细胞类型，提高了聚类效果，同时能应用到多样本和大数据集上。

论文题目

Batch alignment of single-cell transcriptomics data using deep metric learning

论文摘要

scRNA-seq has uncovered previously unappreciated levels of heterogeneity. With the increasing scale of scRNA-seq studies, the major challenge is correcting batch effect and accurately detecting the number of cell types, which is inevitable in human studies. The majority of scRNA-seq algorithms have been specifically designed to remove batch effect firstly and then conduct clustering, which may miss some rare cell types. Here we develop scDML, a deep metric learning model to remove batch effect in scRNA-seq data, guided by the initial clusters and the nearest neighbor information intra and inter batches. Comprehensive evaluations spanning different species and tissues demonstrated that scDML can remove batch effect, improve clustering performance, accurately recover true cell types and consistently outperform popular methods such as Seurat 3, scVI, Scanorama, BBKNN, Harmony et al. Most importantly, scDML preserves subtle cell types in raw data and enables discovery of new cell subtypes that are hard to extract by analyzing each batch individually. We also show that scDML is scalable to large datasets with lower peak memory usage, and we believe that scDML offers a valuable tool to study complex cellular heterogeneity.

作者介绍

余小康，中国人民大学统计学院在读博士生，主要研究方向为单细胞转录组学，深度学习。

张景肖，中国人民大学统计学院教授、应用统计科学研究中心研究员，主要研究方向为高维统计，函数型数据，生物、医学数据分析。

论文发表截图