2024/05/13-2024/0519

2024/05/16

整理基于R语言的单细胞数据分析--preprocessing部分

1.实验目的

2.1熟悉基于R的seurat标准预处理分析流程

2.2理解seurat obeject数据结构

2.3为后续细胞注释分析工作作准备

3.实验材料：

3.1数据来源：

Single-cell analysis of human glioma and immune cells identifies S100A4 as an immunotherapy target(GSE182109)

Single Cell Portal数据，人，脑胶质瘤GBM，T cells（subset）

链接：https://singlecell.broadinstitute.org/single_cell/study/SCP1985/single-cell-analysis-of-human-glioma-and-immune-cells-identifies-s100a4-as-an-immunotherapy-target-gse182109

3.2软件与平台：

R（v4.3.3）；RStudio；Seurat（v5.0.3，https://github.com/satijalab/seurat）；Seurat tutorial（https://satijalab.org/seurat/）

4.实验步骤

简单介绍预处理流程

4.1Load raw data

1.read10X / readhd5 / readRDS

2.subset my interested cell type：T cells

4.2Quality control

1.add percent.mt

2.subset according to percent.mt , nFeature_RNA and nCount_RNA

4.3Standard work flow

1.Normalization

2.FindVariableFeatures

3.Scale

4.PCA

5.FindNeighbours

6.FindClusters (we can setup resolution e.g. res=0.1, 0.8)

7.UMAP

7.Finally, we will get a clustering map

Remember: if the data comes from muti datasets, do integration after PCA(e.g. CCA-Integration, RPCA-Integration)

5.注意事项

1.抽取数据时，尽量少用subset函数，此函数不能多次进行子集操作

2.做质量控制时，务必设定好筛选条件：nFeature_RNA, nCount_RNA, Percent.mt。可通过绘制Vlnplot可视化直观的了解各自的分布。

3.标准流程不必多说，再做之前，要了解数据来源，是否是多个数据集，考虑批次差异的影响，应进行Integrate Layers操作，而后方可进行PCA等后续降维聚类。

4.对于降维聚类，即FindClusters和UMAP，需要设定不同的分辨率（resolution），因为后续的注释是一个极其费时费力的工作，需要先以低倍看整体大群（例如res=0.1），而后以高倍看细分小群（例如res=0.8）

5.运行过程中，需要保持注释的好习惯，尽量用英文注释，并注意划分功能段落，增加可读性的同时也能锻炼自己的英语水平

6.对于过程中产生的临时文件，对于需要的数据，例如大文件（分析时间过长），中部关键数据，要随时保存，并写下读取代码。

7.变量的命名，在保证可读的基本要求下，尽量简洁。（例如：读取的原始单细胞counts矩阵可命名为：sc.raw.counts）