第一次上机实验任务：数据探索

一、数据介绍

单细胞 RNA-seq 测序数据介绍

1. **Visium_Human_Breast_Cancer_filtered_feature_bc_matrix.tar.gz**

   下载自 10× Genomics 的浸润性导管癌乳腺组织的 RNA-seq 数据。Barcodes.tsv.gz 主要包含细胞信息。features.tsv.gz 包含基因信息。matrix.mtx.gz 为基因表达矩阵，行表示细胞，列表示基因，其中的值为特定基因在特定细胞中的表达值。

   Scanpy 中的 read_10x_mtx( )函数可以直接将数据集读取为 AnnData 对象，以便后续分析。

2. **GSE212890 数据集**

   来自 47 名癌症患者包括肿瘤和正常组织的肿瘤浸润 NK 细胞，包括'THCA', 'FTC', 'RC', 'BRCA', 'PACA', 'OV', 'UCEC', 'ESCA'八种癌症；包含 11963 个细胞， 28991 个基因（6044 个正常细胞，5019 个肿瘤细胞）。

   NK_rawcount.barcodes 数据包含两列，第一列是 index，第二列为细胞的标识（cellID），用来标记不同的细胞。

   NK_rawcount.genes：数据包含两列，第一列是 index，第二列为基因的别名（gene_alias，与 gene symbol 的区别自行搜索，此处不过多赘述），用来标记不同的基因。

   NK_rawcount.mtx：细胞中基因表达量的稀疏矩阵。第一行三个数值分别表示：细胞总数，基因总数和基因表达总量。包含三列，其中第一列数值 i 表示第 i 个细胞，第二列数值 j 表示第 j 个基因，第三列数值 m 表示第 i 个细胞中，基因 j 的表达量为 m。

   注意：在 NK_rawcount.barcodes 与 NK_rawcount.genes 中 index 是从 0 开始的，而在 NK_rawcount.mtx 中均从 1 开始，需要注意对应关系。

   GSE212890_NK_metadata.csv：该文件记录数据的一些其他信息，其中 cellID 与 NK_rawcount.barcodes 文件中的 cellID 对应；meta_histology 为肿瘤类型；meta_tissue 标记该细胞是肿瘤细胞还是正常细胞；Majortype 为该细胞的主要类型。

   以上数据解压后可在 PyCharm 上直接进行预览或者用 excel 预览（除 csv 文件外不推荐 excel），数据读取和处理可参考 Scanpy 文档。

二、任务分析

1、 数据的汇总统计分析

针对单个细胞（样本），基因表达的取值范围

数据矩阵中"0"的出现及其含义，给出统计分析结果，并据此给出处理的方案

给出均值、方差等的统计分析结果

2、 数据的可视化

针对上述统计分析给出其可视化结果

针对以上统计分析及可视化结果给出分析和总结

3、 构建共表达网络并可视化

通过计算任意两个基因对之间的皮尔森相关稀疏，构建基因之间的共表达网络，并进行分析；

通过计算两个细胞之间的相似性，构建细胞之间的相似性网络，并进行分析；

注：

网络可视化可以采用 cytoscape，但是不限于此工具。

数据可视化可以采用任何熟悉的编程工具或者软件

三、实验报告撰写

1、针对该数据分析任务，撰写实验报告，包含任务描述、数据描述、详细分析步骤实结合验结果及可视化，但不限于此。

2、以自己的数据探索结果为主要内容，给出丰富的分析结果，并探讨以上结果对于后续数据挖掘的影响和启示。

四、验收

最后一次上机验收实验代码和结果。

最终提交材料为实验报告+实验代码、实验结果。

执笔人：郭杏莉

2023-10-10