欧易 10x 单细胞多组学

（ATAC + 基因表达）测序项目

生物信息分析说明文档

oebiotech 欧易生物

目 录

I 生物信息分析流程

1 测序数据质量控制及定量

建库测序及数据分析部分由上海欧易生物医学科技有限公司完成。高通量测序中产生的原始数据（raw reads）为 fastq 格式序列，采用 10x genomics 官方软件 CellRanger ARC[1] 对原始数据进行数据质量统计以及比对于参考基因组，该软件通过识别序列中的区分细胞 Barcode 序列标记和每个细胞内不同 mRNA 分子的 UMI 标记对高通量单细胞转录组进行定量；通过识别序列中的区分细胞的 Barcode 序列标记和统计 Tn5 酶切位点获得每个细胞所对应的染色质开放区 peak。鉴定获得高质量细胞数、基因中位值、peak 相关等统计信息。

2 定量质控及数据预处理

使用 Seurat[2]软件包在 Cellranger 初步质控结果的基础上，对数据进行进一步质控及处理。 理论上大部分细胞表达的基因数量、UMI 数量和线粒体基因表达量、peak 中片段比例会集中分布在某一区域内，依据这个特征，我们首先通过拟合广义线性模型过滤离域细胞，然后根据 nUMI、nGene 、percent.mito、percent.fragment.peaks 三项指标的分布，过滤剔除双细胞、多细胞或死细胞等低质量的细胞。

3 转录组水平降维与聚类分析

利用基因表达量进行 PCA (主成分)线性降维分析，通过 tSNE 将 PCA 结果在二维空间进行可视化。如果样本间存在批次，则采用互享最近邻降维法 (mutual nearest neighbors)[3]矫正单细胞表达谱数据的批次效应。

4 转录组水平 MARKER 基因鉴定

使用 Seurat 包中的 FindAllMarkers 函数进行 marker 基因鉴定，即找到每种细胞分类相对于其他细胞群差异上调表达的基因，这些基因就是每种细胞分类潜在的 marker 基因。通过 VlnPlot 和 FeaturePlot 函数对鉴定得到的 marker 基因进行可视化。

5 转录组水平细胞类型鉴定

通过 SingleR[4] 包基于单细胞参考表达定量公共数据集，将待鉴定的细胞表达谱与参考数据集计算相关性，把参考数据集中相关性最高的细胞类型赋予待鉴定细胞，一定程度摒除了人为主观因素的干扰。鉴定原理为将样本中每一个细胞的表达谱与参考数据集中注释的每个细胞表达谱计算 spearman 相关性，选择数据集中与样本细胞表达相关性最大的细胞类型作为最终待鉴定的细胞类型。

6 染色质水平降维与聚类分析

利用基因表达量进行 LSI (潜在语义分析) 线性降维分析，通过 UMAP 将 LSI 结果在二维空间进行可视化。如果样本间存在批次，则采用 Harmony[5] 矫正批次效应。

7 染色质水平 MARKER PEAK 鉴定

使用 Seurat 包中的 FindAllMarkers 函数进行 marker peak 鉴定，即找到每种细胞分类相对于其他细胞群差异高度开放的 peak，这些 peak 就是每种细胞分类潜在的 marker peak。通过 VlnPlot 和 FeaturePlot 函数对鉴定得到的 marker peak 进行可视化。

8 染色质水平细胞类型映射

根据 Barcode 在染色质水平降维聚类结果中展示对应细胞类型（转录组水平鉴定出的细胞类型）。并计算每个细胞群中不同细胞类型的占比。

9 WNN 整合分析

根据表达量和 ATAC 数据的相似度的加权组合来计算每个细胞在数据集中的加权最近邻（weighted nearest neighbor, WNN）[6]。使用 Seurat 包中的利用 FindMultiModalNeighbors 函数计算细胞特异性模态权重和多模态近邻，该算法需要指定每个模态的维数。基于表达量和 ATAC 数据的加权组合创建数据的 UMAP 可视化结果。

根据 WNN 整合后细胞群鉴定 marker 基因和 marker peak ，方法分别参考上述 4、7。将细胞类型映射至 WNN 整合后 UMAP 结果，方法参考 8。

II 参考文献

[1] https://support.10xgenomics.com/single-cell-multiome-atac-gex/software/pipelines/latest/what-is-cell-ranger-arc.

[2] Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species[J]. Nature biotechnology, 2018, 36(5): 411-420

[3] Haghverdi L, Lun A T L, Morgan M D, et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors[J]. Nature biotechnology, 2018, 36(5): 421-427.

[4] Aran D, Looney A P, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage[J]. Nature immunology, 2019, 20(2): 163-172.

[5] Tran H T N, Ang K S, Chevrier M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data[J]. Genome biology, 2020, 21(1): 1-32.

[6] Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data[J]. bioRxiv, 2020.

申 明

本项目报告由上海欧易生物医学科技有限公司提供给项目相关客户。本公司承诺：未经客户同意，不向第三方泄露数据及数据分析内容，不将客户数据用于任何商业行为（遵循合同保密协议）。客户未经本公司同意，不得以任何目的向第三方出示项目报告。本报告的最终解释权归上海欧易生物医学科技有限公司。