欧易生物单细胞转录组生信分析方法

I 生物信息分析流程

1 测序数据质量控制及基因定量

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[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的 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，获得高质量细胞数、基因中位值、测序饱和度等质控统计信息。

2 基因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 和 percent.mito 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 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(非线性降维)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 Marker 基因鉴定

使用 Seurat[2]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 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 差异基因及富集分析

使用 Seurat[2]包中的 FindMarkers 函数进行差异基因筛选，根据 p 值小于 0.05 以及差异倍数大于 1.5 倍的条件筛选出差异显著基因，并通过超几何分布检验进行差异显著基因的 GO 和 KEGG 富集分析。

II 参考文献

[1] https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger
[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.