欧易生物单细胞转录组基础分析方法

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 BD 官方软件 BD Rhapsody WTA Analysis Pipeline(version 2.0)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区分细胞的 Barcode 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数据进行定量，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。

批注 [A1]: 人：
https://bd-rhapsody-public.s3.amazonaws.com/Rhapsody-WTA/Pipeline-version2.x_WTA_references/RhapRef_Human_WTA_2023-02.tar.gz
小鼠：
https://bd-rhapsody-public.s3.amazonaws.com/Rhapsody-WTA/Pipeline-version2.x_WTA_references/RhapRef_Mouse_WTA_2023-02.tar.gz

2 基因定量质控及数据预处理

在 BD 初步质控的基础上，使用 Seurat[1](version 4.0.0）软件包对数据做进一步质控处理。理论上，大部分细胞表达的基因数量、UMI 数量和线粒体转录本表达占比会集中分布在某一区域内，因此，我们根据 nUMI、nGene 和 percent.mito 等指标的分布，过滤低质量细胞，具体质控方案为：剔除保留基因数小于 200 、UMI 数小于 1000、log10GenesPerUMI 小于 0.7、线粒体 UMI 占比高于 5%、血红细胞基因占比高于 5% 的细胞作为高质量细胞，同时使用 DoubletFinder[2] (version 2.0.3)软件去除双细胞。质控完成后，使用 Seurat 包中的 NormalizeData 函数对数据进行标准化处理。

批注 [A2]: 默认质控过滤标准，如有调整可自行修改。

3 降维与聚类分析

使用 Seurat 包中的 FindVariableGenes 函数(mean.function = FastExpMean, dispersion.function = FastLogVMR)筛选 Top 2000 个高变基因（HVGs, highly variable genes），利用高变基因的表达谱进行 PCA (主成分）降维分析，通过 UMAP(非线性降维)将结果在二维空间上进行可视化。

批注 [A3]: 如果样本间存在批次，则将 PCA 降维替换为 batchelor(version 1.6.3)包中的互享最近邻降维法 (mutual nearest neighbors)，矫正单细胞表达谱数据的批次效应。

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 Marker 基因鉴定

使用 Seurat 包中的 FindAllMarkers 函数(test.use = presto)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[3] (version 1.4.1)包基于公共参考数据集，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 包中的 FindMarkers 函数(test.use = presto)进行差异基因筛选，根据 pvalue 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。

如果您的研究课题使用了欧易的测序和分析服务，我们期望您在文章发表时，在文章方法部分或致谢部分引用或提及欧易生物：The sequencing and bioinformatics analysis were provided by OE Biotech Co., Ltd. (Shanghai, China).

II 参考文献

[1] Hao, Yuhan, et al. "Integrated analysis of multimodal single-cell data." Cell 184.13 (2021): 3573-3587.

[2] McGinnis, Christopher S., Lyndsay M. Murrow, and Zev J. Gartner. "DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors." Cell systems 8.4 (2019): 329-337.

[3] Aran, Dvir, et al. "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage." Nature immunology 20.2 (2019): 163-172.

批注 [A4]: 细胞类型鉴定部分使用的参考数据集，可根据实际情况自行添加修改。（如果是发 SCI，本部分基本不会用到，舍弃，如是中文核心或者毕业论文，可以选择性纳入）

人的可选参考数据集：
1、HPCA（Human Primary Cell Atlas）：Mabbott N A, Baillie J K, Brown H, et al. An expression atlas of human primary cells: inference of gene function from coexpression networks[J]. BMC genomics, 2013, 14(1): 632.
2、blueprint+encode：
Blueprint:
Martens, J. H. A, Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes[J]. Haematologica, 98(10):1487-1489.
Encode：
Bernstein B E, Birney E, Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome[J]. Nature, 2012, 489(7414): 57-74.
3、schcl： http://bis.zju.edu.cn/HCL/index.html

小鼠可选参考数据集：
1、 immgen：Heng T S P , Painter M W , Elpek K , et al. The Immunological Genome Project: networks of gene expression in immune cells[J]. Nature Immunology, 2008, 9(10):1091-1094.
2、scmca：Xiaoping Han, Renying Wang, Yincong Zhou,et al. Mapping the Mouse Cell Atlas by Microwell-Seq[J]. Cell, 2018, 172(5):1091-1107.e17.
3、mouse.rnaseq：Benayoun B A, Pollina E A, Singh P P, et al. Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses[J]. Genome research, 2019, 29(4): 697-709.

批注 [A5]: 或 1.2 倍，根据实际结果修改。