**iMeta｜ScRNAPip单细胞分析流程使用教程**

2023年X月X日，iMeta期刊发表了题为 “ScRNAPip: a Systematic and Dynamic Pipeline for Single-Cell RNA Sequencing Analysis” 的论文，该论文介绍了一个单细胞分析工具，为用户提供了一种高效、全面的单细胞分析解决方案，访问地址是<https://github.com/OpenGene/scrnapip>。

ScRNAPip分析流程涵盖了从下机数据到定量、降维分析、细胞聚类、marker基因鉴定、细胞轨迹分析、基因组不稳定性分析、细胞干性分析、CNV、富集分析等个性化分析的全过程，并提供了相关的配置文件和可视化工具。这种综合性的设计可以帮助用户在一个流程中完成所有步骤，减少学习和分析的复杂性，提高效率。

ScRNAPip流程使用了Docker容器封装了该流程所需的所有软件和依赖项，减少了用户在本地安装和配置各种软件的麻烦。用户只需通过Docker容器即可访问流程中所需的工具和库，无需担心版本兼容性和依赖冲突。此外，为了适应不同实验目的和研究需求，配置文件还提供了多样化的配置选项，如遇到程序长时间运行后意外崩溃，还可以通过修改配置文件实现断点重启。

**下面是流程使用说明：**

本教程主要从安装、输入文件说明，配置文件说明、结果输出和其他五个方面介绍如何使用ScRNAPip进行分析。

# 环境配置

## （1） Docker简介

Docker官网：

Docker (<https://www.docker.com/>) 是一个开源的应用容器引擎，让开发者可以打包他们的应用以及依赖包到一个可移植的容器中，然后发布到其他的的[Linux](https://baike.baidu.com/item/Linux?fromModule=lemma_inlink" \t "_blank)/[Windows/mac操作系统](https://baike.baidu.com/item/Windows%E6%93%8D%E4%BD%9C%E7%B3%BB%E7%BB%9F/852149?fromModule=lemma_inlink" \t "_blank)的机器上，也可以实现虚拟化，容器是完全使用[沙箱机制](https://baike.baidu.com/item/%E6%B2%99%E7%AE%B1%E6%9C%BA%E5%88%B6/677601?fromModule=lemma_inlink" \t "_blank)，相互之间不会有任何接口。

Docker的常用命令如下：

1）查看镜像列表：docker images；

2）从镜像仓库中拉取镜像：docker pull 镜像名:tag；

3）删除镜像：docker rmi -f 镜像名/镜像ID；

4）查看所有容器：docker ps -a；

5）运行容器：docker run -it -name 容器名 -p 宿主机端口:容器内端口 镜像名 /bin/bash；

6）退出容器：exit；

7）重新运行关闭的容器：docker restart 容器名；docker attach 容器名；

## （2）拉取镜像

在拉取镜像前，请先确保服务器上已经安装了Docker软件，可以从Docker的官网查看安装方法 (https://docs.docker.com/engine/install/)，根据服务器的平台选择对应的安装方法。安装完成后使用Docker version查看是否成功。

之后，从Docker hub (https://hub.docker.com/r/zhangjing12/scrnapip)或者GitHub ( <https://github.com/OpenGene/scrnapip>) 上拉取该项目的镜像。

命令：docker pull zhangjing12/scrnapip

## （3）启动镜像

命令：docker run -d -p 1921:8787 -p 1882:3838 -e PASSWORD=yourpassword -e USERID=youruserid -e GROUPID=yourgroupid -v /yourdatapath:/dockerpath zhangjing12/scrnapip

注意：以下部分需要修改

yourpassword：密码信息；

youruserid,yourgroupid：通过命令”id“获取；

yourdatapath：linux服务器上的路径，将原始数据和需要的文件放入该目录下；

dockerpath：Docker容器中的路径，挂载在yourdatapath下；

这一部分的各个内容在下方会有详细介绍。

# 输入文件说明

输入文件包含两个部分：

第一个是配置文件 config file，详见3.配置文件说明。

第二个是参考基因组的配置，如果无参考基因组，则可通过10x官网的地址进行下载：

Human reference (GRCh38)：

wget <https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-GRCh38-2020-A.tar.gz>

Mouse reference：

wget <https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-mm10-2020-A.tar.gz>

这两个文件下载后最好放置在一个可以同时被读取到的目录下，在启动Docker时将yourdatapath设置为这个目录，便于分析和管理。

例如：

mkdir -p $yourdatapath/data $yourdatapath/reference

vi $yourdatapath/data/config\_Example.ini

wget <https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-GRCh38-2020-A.tar.gz> -P $yourdatapath/reference

tar -zxvf $yourdatapath/reference/refdata-gex-GRCh38-2020-A.tar.gz

# 配置文件说明

配置文件使用的是ini格式的文件，通过R进行读取。综括号内的部分为各个模块的标题，标题之后的内容就是该模块里提供的变量，也就是读取到程序中的参数。通过该配置文件，将后续分析中所有需要的参数统一进行修改。在修改它时，综括号中的标题不可更改。需要修改的变量和它们的实际含义如下：

## （1）[fastp\_cellrange]

该项下输入原始数据和路径，基本格式为

S1.R1=["$dockerpath/data/SAMPLE1\_S1\_L001\_R1\_001.fastq.gz"]

S1.R2=["$dockerpath/data/SAMPLE1\_S1\_L001\_R2\_001.fastq.gz"]

其中，S1为样本名，R1/R2代表数据的R1/2端，[“”]内的路径为原始数据的路径。若一个样本有多个原始数据，则需要使用逗号分割开，如：

S1.R1=["$dockerpath/data/SAMPLE1.1\_S1\_L001\_R1\_001.fastq.gz"，"$dockerpath/data/SAMPLE1.2\_S1\_L001\_R1\_001.fastq.gz"]

S1.R2=["$dockerpath/data/SAMPLE1.1\_S1\_L001\_R2\_001.fastq.gz"，"$dockerpath/data/SAMPLE1.2\_S1\_L001\_R2\_001.fastq.gz"]

将所有的原始数据按以上格式配置。每个样本分为两行R1/R2，最终有2\*样本的行数。

## （2）[indata]数据路径

该项输入的为cellranger文件的表达矩阵结果路径，例如S1样本的路径为：

S1=“$dockerpath/workout/02.cellranger/S1/outs/filtered\_feature\_bc\_matrix”

每个样本都需要一个路径信息。

这里需要注意的有两点：

1）$dockerpath/workout路径为用户输入的[outpath]，请根据实际情况更改，例如[outpath]设置为outpath=”analysis“则此处就是$dockerpath/analysis；

2）路径中的S1为样本名，请与之前的样本名保持一致。

特殊情况：如果跳过了fastp和cellranger，直接从表达矩阵进行分析，请将[run]中的fastp和cellranger设置为false，并将此处的路径改为10X表达矩阵的路径。

## （3）[outpath]输出路径

此为输出结果的路径，即为结果创建一个新的文件夹outpath=“workout”。

## （4）[tempdata]rds路径

Rds文件路径

若跳过了fastp和cellranger直接从rds进行分析则需要此路径。如果没有跳过前面步骤则将其设置为空（程序优先会在默认目录下查找该文件，所以如果没有跳过步骤，这里的tempdata就不会起作用）。

tempdata=$rdsfilepath

tempdata=""

## （5）[run]分析步骤设置

将需要运行的模块设置为true，各步骤对应如下：

|  |  |
| --- | --- |
| fastp | 原始数据质控 |
| cellrangle | 比对和定量生成表达矩阵 |
| step1 | 生成seruat对象，过滤低质量细胞 |
| step2 | 样本合并 |
| step3 | 降维分析(PCA和UMAP)和细胞聚类 |
| step4 | 查找marker基因 |
| step5 | 细胞轨迹分析 |
| step6 | circos图和cerebro可视化 |
| step7 | 拷贝数变异分析 |
| step8 | 干性分析 |
| step9 | 基因组不稳定性分析 |
| step10 | 细胞通讯分析 |
| step11 | 通路富集分析 |

设置方法为：fastp=true

## （6）[fastp]数据质控

fastp是一款对fastq进行质控的软件，其速度快和占用资源少且一次完成多项质控操作的特点，被广泛用于原始数据的预处理中。通常该部分只需要使用默认参数。

fastppath="/usr/fastp" #fastp的路径，无需更改；

longr=28 #R1端保留的最长长度，10x单细胞R1端的Barcode+UMI为26/28bp；

ncode=5#允许的N碱基的数量；

## （7）[cellrangle]参考基因组比对

Cell Ranger是10X Genomics公司为单细胞RNA测序分析量身打造的数据分析软件。整合了比对，定量，细胞过滤，生成表达矩阵，初步分析等功能。

需要设置的参数如下：

ref="$dockerpath/reference/refdata-gex-GRCh38-2020-A" #设置参考基因组的路径；

expectcell=10000 #预期的细胞数量；

localcores=32 #分配的线程；

localmem=64 #分配的内存；

include\_introns="false" #是否在cellranger分析时纳入内含子；

## （8）[step1]细胞过滤

初步创建了seurat对象并根据基因和线粒体数过滤细胞，过滤的标准设置方法如下。

nFeature\_RNA=[200,8000] #保留基因数200到5000之内的细胞；

percent\_mt=[0,20] #保留线粒体占比在20%以内的细胞；

这里只提供了通常情况下的过滤数值，不同的组织样本有所不同，可以通过查看相关文献中给出的过滤标准。

## （9）[step2]样本合并

normethod="none" #样本合并参数。

none：merge将样本直接合并；

SCT：使用CCA算法进行矫正；

nFeature=3000 #选择高变基因进行后续分析，默认设置为3000；

不推荐seurat的CCA进行批次矫正，如果有相关需求请使用下面的算法【harmony MNN】。

## （10）[step3]降维分析、细胞聚类

使用PCA和UMAP对数据降维，随后进行聚类。

如果样本差异较大，例如来自于不同的平台，则需要进行批次矫正。

recluster=true #是否进行批次矫正，不进行矫正改为false；

mode="harmony" #有两种方法可选，harmony和mnn；

resolution=0.6 #聚类的分辨率，在cluster较多的时候可以通过调低该值减少cluster的数量；

algorithm=1 #聚类的算法：1 = original Louvain algorithm；2 = Louvain algorithm with multilevel refinement； 3 = SLM algorithm；

## （11）[step4]marker基因鉴定

上一步已经对细胞进行聚类得到了全部cluster，这一步是进行差异分析。

如果需要进行marker基因的分析，需要将clustermarkers设置为true。注意，marker基因分析默认只输出各cluster中上调的基因。

clustermarkers=true #是否进行marker基因分析；

min\_pct=0.25 #基因所占细胞数的最低比例，默认0.25。这个值越小，包含的基因数量越多，但占比较小的基因通常对marker分析用处不大。

findmarkers\_testuse="wilcox" #marker基因分析使用的算法：1）wilcox：Wilcoxon秩和检验（默认）；2） bimod：似然比测试；roc：ROC分析识别；3） t：学生t检验； negbinom：负二项式广义线性模型；4）poisson：泊松广义线性模型。

通过设置不同细胞类型的基因，可以输出这些基因的umap图结果。格式为：custer.XXX=["testgene"]；

以T细胞为例：custer.T\_cell=["CD3E","CD3D"]。

此外，可以自定义cluster或样本进行差异表达基因的分析。

例：cluster之间差异分析

difcluster.test.a=[0,1] #设置case组为cluster0和cluster1；

difcluster.test.b=[5,6] #设置control组为cluster5和cluster6；

difcluster.test.testuse="wilcox" #差异分析算法为wilcox。

例：样本之间差异分析

difident.tVSn.a=["sample1"] #设置case组为样本sample1；

difident.tVSn.b=["sample2"] #设置control组为样本sample2；

difident.tVSn.testuse="bimod" #差异分析算法为bimod。

## （12）[step5]细胞轨迹分析

拟时序分析可以便于我们发现细胞状态和功能的转变。我们使用被广泛应用的Monocle进行拟时序分析的细胞轨迹构建，并进行节点的beam分析来查看经过该位点的不同branch的差异基因。

genenum=50 #差异分析热图使用的基因数量；

numclusters=4 #差异分析热图聚类的cluster数量；

pointid=1 #Beam分析选择的节点，默认为1；

BEAMgn=50 #Beam分析热图基因数量；

BEAMnumclusters=4 #Beam分析热图的cluster数量；

BEAMgenelist=["S100A12", "ALOX5AP", "PAD14", "NRG1", "MCEMP1", "THBS1"] #BEAM分析后自定义部分用于绘图的基因。除此之外还会绘制显著差异top10的基因。

## （13）[step11]富集分析

使用GO，KEGG，reactome等数据库进行通路富集分析，得到每个cluster可能的功能特征。

设置如下：

ClusterProfiler=["true","Rscript","/home/bin/clusterProfiler.R","-a true -s org.Hs.eg.db,hsa,human -g 6 -t SYMBOL -d KEGG, BioCyc, PID, PANTHER, BIOCARTA -C 0.05"] ；

# -a：是否使用全部背景基因；-s：物种； -g：输入文件的基因列；-t：基因名类型 (SYMBOL, ENTREZID)；-d：选择进行分析的数据库名称。

# 各步骤及结果文件说明

## （1）fastp数据质控

── 01.fastp/

└── <SampleName>/

├── <SampleName>\_fastp.html

├── <SampleName>\_fastp.json

├── <SampleName>\_S1\_L001\_R1\_001.fastq.gz

└── <SampleName>\_S1\_L001\_R2\_001.fastq.gz

使用fastp软件对原始数据进行质控，并将R1端多余的reads切除。

结果中的两个fastq.gz文件为过滤后的R1/R2端数据，且名称已转换成cellranger要求的格式。

若跳过了fastp步骤，需要自行修改该名称。其余两个html文件和json文件为fastp软件给出的两种记录统计信息的格式，便于查看数据质量。

## （2）cellranger参考基因组比对

── 02.Cellranger/

└── <SampleName>/

└── outs/

├── analysis/

├── raw\_feature\_bc\_matrix/

├── raw\_feature\_bc\_matrix.h5

├── filtered\_feature\_bc\_matrix/

├── filtered\_feature\_bc\_matrix.h5

├── molecule\_info.h5

├── metrics\_summary.xls

├── possorted\_genome\_bam.bam

├── possorted\_genome\_bam.bam.bai

└── web\_summary.html

使用cellranger软件进行比对和定量分析，生成10x的标准matrix文件，具体路径在02.Cellranger/<SampleName>/outs/filtered\_feature\_bc\_matrix

cellranger分析的结果报告主要查看web\_summary.html结果。

需要关注的几个数据是：

Estimated number of cells：细胞数量，正常细胞数量在5000到10000之间。过低的细胞数量不利于分析，多高的细胞数量则会导致双细胞率的上升；

Mean Reads per cell：平均每个细胞的reads，通常认为在50k左右是比较理想的数值；

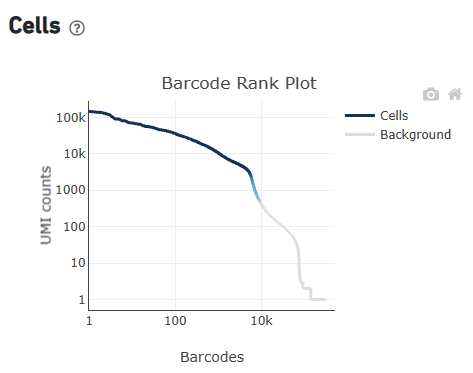
Median Genes per cell：该数值主要看组织的来源和数据量大小，干细胞或者肿瘤组织表达的基因较多，这类细胞偏多的样本可能会达到3k到4k或更高。如果是免疫浸润强的组织一般会在1k-2k之间；

Valid Barcodes：被识别到的barcode，该值越大越好；

Sequencing Saturation：测序的饱和度Q30 Base：一些质量指标；

Reads Mapped to Genome=比对到基因组的序列的数量。比对率在 80%以上都是可以接受的值。

曲线：



纵轴是一个 UMI 的数量，横轴是 Barcode 的数量。该图根据 UMI的数量将细胞从大到小往下排。随着barcode 的增加，UMI Counts 会逐渐下降，随后会在曲线上有一个突然的下降，之后的细胞被认为是背景或死亡的细胞。

其他结果文件：

raw\_feature\_bc\_matrix：原始表达矩阵；

filtered\_feature\_bc\_matrix：过滤后的表达矩阵，也就是过滤了上方曲线中灰色部分的背景，通常后续分析使用过滤后的矩阵；

上面两个matrix对应的两个h5文件：feature-barcode HDF5矩阵；

molecule\_info.h5：该文件用于进行cellranger aggr的样本合并；

metrics\_summary.xls：数据统计表格，与web\_summary.html中的数据一致，但用表格的形式存储，便于将多个样本的信息整合在一起方便比较；

passorted\_genome\_bam.bam/bai：使用STAR软件得到的序列比对结果。

## （3）Step2样本合并

── 03.CellFilter/

├── <SampleName>/

│ ├── <SampleName>\_countVfeature.png/pdf

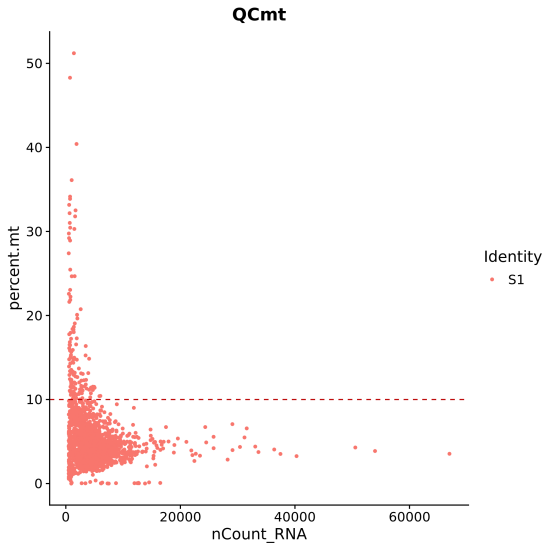
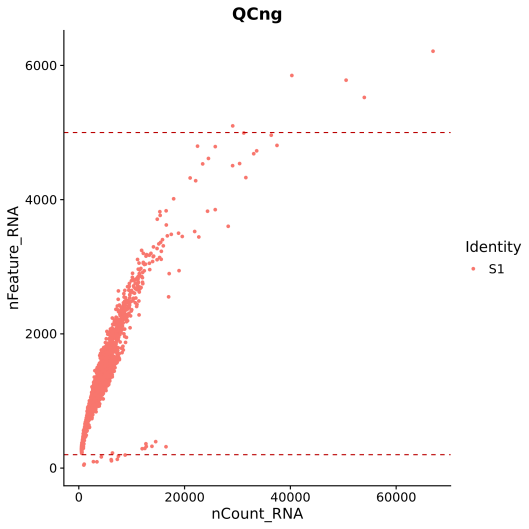
│ ├── <SampleName>\_countVmt.png/pdf

│ ├── <SampleName>\_libraryVmt.png/pdf

│ └── <SampleName>\_voilin.png/pdf

└── summary.txt/

这里主要查看每个样本下的<SampleName>\_countVmt.png和<SampleName>\_countVfeature.png



根据组织和样本的实际情况调整feautre和线粒体的阈值，对应配置文件中的nFeature\_RNA=[200,5000]、percent\_mt=[0,20]。根据样本的实际情况确定阈值，线粒体需要观察图中的峰值，将阈值设置在峰值以上，通常设置为10%或20%。若为肿瘤组织可以设置的更高。超过这个阈值的细胞都会被过滤掉，即认为线粒体占比高的细胞质量差，有可能为死细胞。而feature的阈值可以稍微宽松，由于有中性粒这种基因数本身偏低的细胞类型存在，将基因下限设为200比较合适，在细分中如果单独分析某类细胞可以另外进行卡值。

## （4）Step3降维分析、细胞聚类

── 04.PCA\_UMAP/

├── pca/

│ ├── bowplot.png/pdf

│ ├── pca.png/pdf

│ └── pcaheatmap1.png/pdf

└── umap/

├── barplot.png/pdf

├── celltype.png/pdf

├── cluster0.png/pdf

├── plotall\_ident.png/pdf

├── plotby\_cluster.png/pdf

├── plotby\_ident.png/pdf

├── plotby\_nCount.png/pdf

├── singleR\_celltype.xls

└── summary.xls

该部分分析内容为降维，分为PCA和UMAP降维两步。

PCA文件夹下的结果中关注bowplot，根据拐点选择使用的pc，pcaheatmap中可以查看每个pc中的关键基因。

Umap文件夹下则是在umap降维并进行了聚类之后得到的结果，包含cluster，样本，count的分布图以及singleR进行细胞类型鉴定的结果。

由于singleR对不同样本的分析差别准确率差别较大，这里的结果仅作为细胞类型判定的参考，实际需要依据marker基因和已有的文献。

选择使用的pc对应配置文件中的[step3] dims，reduction的方法可以选择tsne或umap对应[step3] reduction，调整聚类的分辨率对应[step3] resolution，若样本间的批次效应过大，可以选择[step3]recluster=true来进行批次矫正，并通过[step3]mode来调整批次矫正的方法，目前支持mnn和harmony两种。通常推荐使用harmony进行矫正。

## （5）Step4 marker 基因鉴定

── 05.MarkerGene/

├── cluster\* /

│ ├── cluster\_genelist.xls

│ ├── cluster\_genereduction.png/pdf

│ ├── cluster\_genevlnplot.png/pdf

│ └── cluster\_padj0.05\_logFC0.5genelist.xls

├── custer/

│ └── Celltype

│ ├── \*\_reduction.png/pdf

│ └── \*\_vlnplot.png/pdf

├── cluster\_top10geneheatmap.png/pdf

├── clusterall\_adj0.05\_logFC0.5genelist.xls

└── clusterall\_top10genelist.xls

这一步主要是分析每个cluster的marker基因并绘制各cluster中marker基因的umap图和violin图等。

第一层文件夹中有所有cluster的显著差异的marker基因表格和top10的marker基因以及热图。此外每个cluster命名的文件夹下则是该cluster单独的结果。

根据各cluster的marker基因，可以进行细胞类型鉴定，也可以进行通路富集分析等。

## （6）step5细胞轨迹分析

── 06.Pseudotime/

├── pseudotime /

│ ├── geneplot.png/pdf

│ ├── pseudotime\_byclusters.png/pdf

│ ├── pseudotime\_byPseudotime.png/pdf

│ ├── pseudotime\_bystate.png/pdf

│ ├── pseudotime\_splitbyclusters.png/pdf

│ ├── pseudotime\_splitbyorig.ident.png/pdf

│ └── cell\_state\_clusters.xls

├── BEAM /

│ ├── custergene.png/pdf

│ ├── top10gene.png/pdf

│ ├── topgeneheatmap.pdf

│ └── diffgenelist.xls

├── diffgenes /

│ ├── topgeneheatmap.pdf

│ └── cell\_diffgene.xls

└── QC.png/pdf

这一步使用全部细胞进行拟时序分析，通常我们需要在分析了细胞类型后选择几种细胞类型进行分析，可以看到它们之间的转化关系，细胞类型太杂会影响轨迹的构建和对结果的理解。

所有的结果中需要关注的有：

1）pseudotime/pseudotime\_byclusters.png：最终的轨迹图，颜色代表的是不同的cluster。原始数据表在cell\_state\_clusters.csv中。

2）pseudotime/pseudotime\_byPseudotime.png：查看monocle判定的轨迹起始点。

BEAM/topgeneheatmap.png：基因在经过节点的两个branch中的表达变化结果。这张图的起始点在中间，最上方的CellType显示灰色的部分为Pre-branch，红色和蓝色分别为两种细胞状态。根据配置文件中的设置[step5]BEAMnumclusters=4将热图分为四类。

3）BEAM/custergene.png：使用[step5]BEAMgenelist中选择的基因绘制的beam基因散点图。图中不同的颜色代表state，对应pseudotime/pseudotime\_bystate.png。横坐标从左到右是monocle判定对细胞的时间排序，对应pseudotime/pseudotime\_byPseudotime.png的结果。虚线和实线是两条branch的细胞拟合得到的曲线。从这个图可以看到，有些基因的两个branch在初始state中趋势保持一致，而到了不同的state后会发生明显的变化，某一个会显著上调或下调，这种基因就是在这个节点中显著变化的基因。

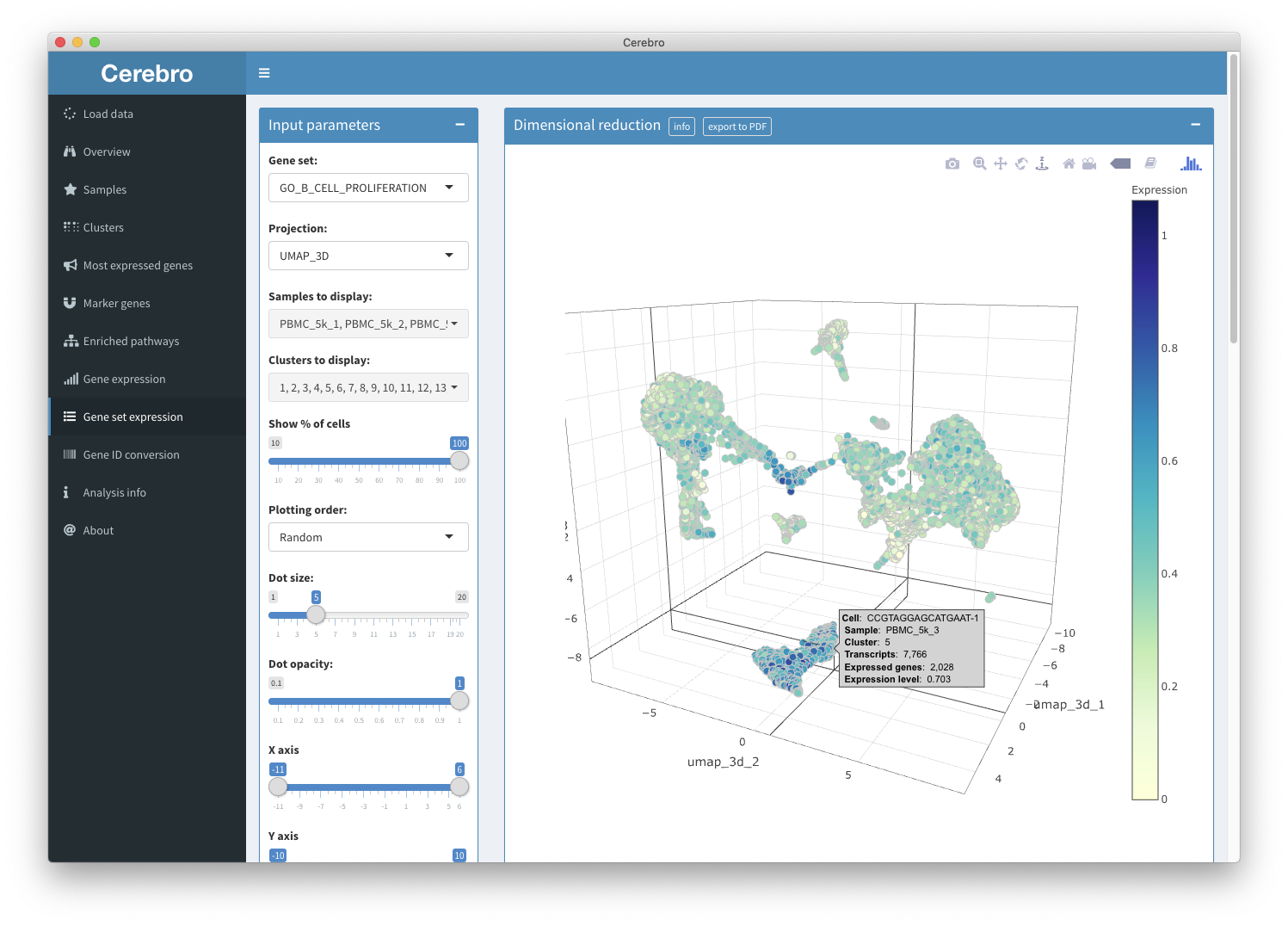
## （7）step6 cerebro可视化、circos图

为了更好的进行结果的展示，便于用户对结果进行查找，整合了单细胞可视化工具cerebro。

常见的单细胞转录组图形基本都包含，用起来也很方便，很适合不擅长生信分析的人来做数据探索，对于做分析的人来讲，也降低了和非分析合作者的沟通成本。

利用生成的输入文件导入到cerebro的loaddata中，即可查看seurat分析的所有结果，用户可以自行选择想要查看的样本，cluster以及改变展示的图片类型。

circos文件夹下提供了使用marker基因绘制的circos图。一共分为三层，气泡图，热图和散点图。气泡图的大小代表的基因表达的细胞在该cluster中所占比例，颜色区分不同样本，透明度表示p值大小。热图展示的则是不同cluster，不同样本的基因表达量。散点图则显示了所有marker基因的差异倍数高低，越接近中心的基因差异倍数越大。



## （8）step7 拷贝数分析

── 09.Copykat/

├── <SampleName1> /

│ ├── <SampleName1>\_copykat\_clustering\_results.rds

│ ├── <SampleName1>\_copykat\_CNA\_raw\_results\_gene\_by\_cell.txt

│ ├── <SampleName1>\_copykat\_CNA\_results.txt

│ ├── <SampleName1>\_copykat\_heatmap.jpeg

│ ├── <SampleName1>\_copykat.pdf

│ ├── <SampleName1>\_copykat\_prediction.txt

│ ├── <SampleName1>\_copykat\_with\_genes\_heatmap.pdf

│ ├── <SampleName1>\_rawdata.txt

│ ├── <SampleName1>.tumor\_subtype.pdf

│ └── <SampleName1>.tumor\_subtype.txt

├── <SampleName2> /

├── ...

│

├── copykat.counts.clusterbarplot.pdf/png

├── copykat.counts.samplebarplot.pdf/png

├── copykat.prop.clusterbarplot.pdf/png

├── copykat.prop.samplebarplot.pdf/png

├── copykat.umap.pdf/png

├── copykat.umap.splitsample.pdf/png

└── Summary\_copykat\_prediction.txt

肿瘤单细胞RNA测序的一个主要挑战是区分癌细胞和非恶性细胞类型和区分肿瘤亚克隆。CopyKAT是一种拷贝数变异的计算工具，使用综合性的贝叶斯方法在单个细胞中以5MB分辨率识别全基因组阳极，将肿瘤细胞与正常细胞分离。

虽然copykat支持直接输入表达矩阵，但提供正常细胞可以极大提高预测的准确性。

每个样本名文件夹下都会有该样本copyakat分析结果，其中<SampleName>\_copykat\_heatmap.jpeg为CNV结果的热图，横轴为按染色体顺序排列的基因，上方由灰色和黑色组成的bar区分了由1-22以及xy染色体。纵轴则是细胞，根据拷贝数变异对细胞进行了聚类，橙色部分为肿瘤细胞，绿色部分为正常细胞。

此外，Summary\_copykat\_prediction.txt为统计所有样本的肿瘤细胞预测情况，共有三种状态，aneuploid为肿瘤，diploid为正常细胞，not.defined为不符合软件预设的质量要求的细胞。

其他的柱状图展示的是这三种类型在不同cluster，不同样本中的占比。在最终确定肿瘤细胞时需要根据实际情况进行调整，如果在运行copykat时没有指定正常细胞，则正常的免疫细胞中会有少量被判定为肿瘤，这些假阳性的值需要将其清理掉。

## （9）step8干性分析

── 10.CytoTRACE/

├── <prefix>.CytoTRACE.boxplot\_raw.pdf/png

├── <prefix>.CytoTRACE.boxplot\_type.pdf/png

├── <prefix>.cytovalue.FeaturePlot.pdf/png

├── <prefix>CytoGenes.pdf

└── <prefix>.CytoTRACE.table.txt

CytoTRACE (Cellular (Cyto) Trajectory Reconstruction Analysis using gene Counts and Expression) 是一种从单细胞rna测序数据预测细胞分化状态的计算方法。研究发现，细胞中可检测到的表达基因的数量越多，它的发育潜力就会越大。CytoTRACE利用这一点对每个细胞进行打分，最终体现出细胞的分化水平。

结果中CytoTRACE.CytoTRACE.table.txt为每个细胞的打分表，根据该表绘制了箱线图和染色图。箱线图按照每个cluster的得分中位数进行排列，从左到右的cluster发育潜力越来越低。染色图显示了CytoTRACE得分在umap或tsne上的分布。

利用CytoTRACE的结果，可以帮助我们定义细胞类型，了解新发现细胞类型的发育状态。将其与拟时序结果结合还能帮助我们确定起始位点。

## （10）step9基因组不稳定性分析

── 11.genomicInstability/

├── <prefix>\_genomicinstability.pdf

└── <prefix>\_gis.result.txt

使用R包genomicInstability进行基因组不稳定性分析。该软件包包含从scRNA-Seq数据运行基因组不稳定性分析 (GIA) 的功能。GIA估计基因表达和编码基因的基因组位置之间的关联。

它使用aREA算法来量化基因表达谱上连续基因（位点块）的富集程度，并估计每个分析细胞的基因组不稳定性评分 (GIS)。

结果中<prefix>\_gis.result.txt为基因组不稳定性评分GIS和似然值的表格。<prefix>\_genomicinstability.pdf展示了GIS的密度分布。将relative likelihood为0.5的位置所对应的GIS作为阈值，可以将所有细胞分为小于该阈值的基因组稳定细胞和大于该阈值的基因组不稳定细胞。

## （11）step10细胞通讯分析

── 13.CellChat/

├── bubble/

│ └── <CellType>\_bubble.pdf

├── pathway/

│ ├── <GeneName>\_circle.pdf

│ └── <GeneName>\_hier.pdf

├── <CellType>\_circle.pdf

└── topgeneheatmap.pdf

CellChat是一种能够从单细胞rna测序 (scRNA-seq) 数据中定量推断和分析细胞间通信网络的工具。CellChat预测细胞的主要信号输入和输出，以及这些细胞和信号如何使用网络分析和模式识别方法协调功能。通过多种学习和定量对比，CellChat对信号通路进行分类，并在不同的数据集上描绘出保守的和特定于环境的通路。

结果中cluster0\_circle.pdf中用连线的宽度显示了cluster0与其他所有cluster的通讯次数。topgeneheatmap.pdf则将所有cluster的结果进行了整合，该pdf文件中的两张图分别显示了所有cluster之间的通讯次数和通讯强度。bubble/cluster0\_bubble.pdf气泡图展示了cluster0作为配体，其他cluster作为受体的配受体基因对以及它们的调控强度和p值。pathway/genename.circle.pdf则是展示了某一个基因在所有cluster间的调控关系。而genename\_hier.pdf是分为了两张图，第一张是左侧的cluster作为target的调控关系，第二张是右侧的cluster作为target的调控关系。

## （12）step11富集分析

── 12.clusterProfiler/

├── <prefix>.BIOCARTA\_Enrich.xls

├── <prefix>.BioCyc.png/pdf

├── <prefix>.BioCyc\_Enrich.xls

├── <prefix>.BP.DAG.svg

├── <prefix>.CC.DAG.svg

├── <prefix>.MF.DAG.svg

├── <prefix>.go.pdf/png

├── <prefix>.GO\_Enrich.xls

├── <prefix>.KEGG.png/pdf

├── <prefix>.KEGG\_Enrich.xls

├── <prefix>.PANTHER\_Enrich.xls

├── <prefix>.PID\_Enrich.xls

├── <prefix>.reactome.png/pdf

└── <prefix>.Reactome\_Enrich.xls

ClusterProfiler集成了GO，KEGG分析，同时整合了基因ID的转换，富集分析和可视化如气泡图柱状图等。我们使用clusterprofiler对step4的结果中每个cluster的marker基因进行富集分析，可以帮助我们进一步了解每个cluster的主要功能。

结果中通常包含多个数据库的富集分析结果，如KEGG，BioCyc，PID，PANTHER，BIOCARTA。每个数据库的结果表格都以Enrich.xls为后缀，该表格中ID为每个数据库中通路或term的编号，Description为详细的描述信息或通路名称，GeneRatio为输入的基因在该通路中的数量比上输入基因的总数，BgRatio为该通路的基因总数比上数据库中的基因数量。*p* value，p.adjust和*q* value显示了该通路的显著水平，geneID为输入基因中在该通路中的基因的Entrezid，Count为基因的数量。

此外，还使用柱状图和气泡图对以上表格中的信息进行可视化。DAG.svg为后缀的网络图中分别显示了GO的三个模块，生物学过程BP，分子功能MF和细胞组分CC，它们中显著的term与其他term之间的上下游关系。

# 其他

## （1）其他物种分析配置

本流程目前物种仅支持人和小鼠。如果想分析其他物种，请按以下步骤创建参考基因组并在config中指定其位置。

1）从NCBI或Ensembl下载参考基因组的fa和gtf文件。

NCBI：https://www.ncbi.nlm.nih.gov/genome/?term=

Ensembl：1）http://asia.ensembl.org/info/about/species.html；2）http://ensemblgenomes.org/

注意，从NCBI下载的注释文件通常为gff格式，需要通过gffread将其转换成gtf格式。首先conda安装gffread：conda install gffread，随后进行转换：gffread genome.gff3 -T -o genome.gtf。

2）使用cellranger的mkref命令构建参考基因组。

如果本地没有安装cellranger，也可以进入docker使用。命令为：

cellrangermkref --genome=GenomeName --fasta=genome.fa --genes=genome.gtf

以上构建的参考基因组可用于cellranger分析生成表达矩阵，后续高级分析中step1到step6以及step8干性分析无需更改可直接运行，step7拷贝数变异分析，step9基因组不稳定性分析和step10细胞通讯分析仅支持人和小鼠。step11通路富集分析需要更改数据库。

## （2）断点重启

如果遇到突发情况，流程断了。可以设置断点重启，此时需要修改config.ini文件，需要将[run]中需要跳过的模块修改为false，重新投递即可。