dnbc4tools

发行版本 *2.0*

lishuangshuang

2022 年 11 月 09 日

Contents

DNBelab C4 便携式单细胞系统

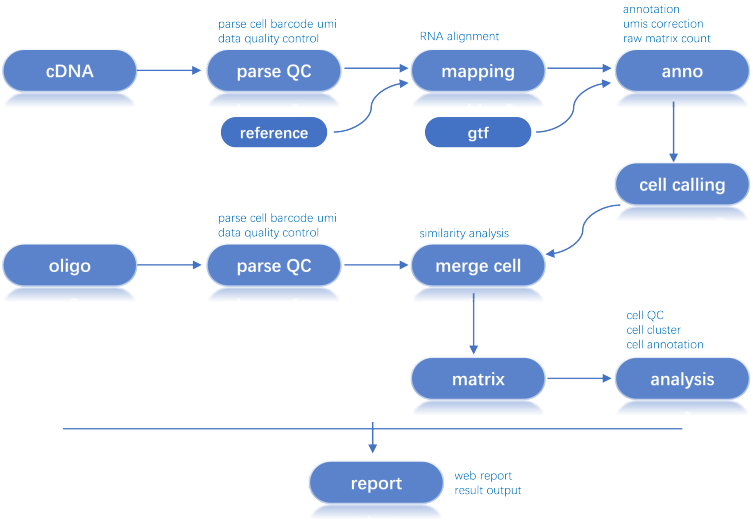
单细胞组学技术的迅猛发展，正不断更新我们对生物系统的理解和认知。DNBelab C 系列细胞组学整体产品组合，基于独特的 DNBelab C 系列单细胞文库制备技术和强大的 DNBSEQ 测序技术，由 DNBelab C4 便携式单细胞系统、DNBelab C 系列单细胞 RNA 文库制备套装，搭配华大智造 DNBSEQ 系列测序平台，以及相关的单细胞分析软件构成，可实现便携式、即时化、一站式单细胞组学研究全流程。

DNBelab_C_Series_HT_scRNA-analysis-software 单细胞 RNA 分析软件用于 DNBelab C4 便携式单细胞系统的 RNA 数据分析。

CHAPTER 1

软件说明

1.1 1. 流程分析原理概览



1.2 2. 软件文件结构说明

github 软件包含 5 个目录、2 个说明文件及 1 个 yaml 文件。

文件目录名称	描述
DNBC4tools	DNBelab C4 分析流程、脚本、软件以及配置文件存放目录。
doc	帮助使用文档存放目录。
example	DNBelab C4 分析示例文件存放目录。
scripts	DNBelab C4 常用分析脚本存放目录。
wdl	该目录包含 WDL 编写的主流程。
CHANGELOG.md	软件版本升级说明。
DNBC4tools.yaml	分析流程 conda 环境安装文件。
LICENSE	DNBelab C4 分析流程许可证。
README.md	DNBelab C4 简要使用说明。

1.3 3. 功能结构

软件整体可以划分为 4 个功能模块：

功能	详细描述
数据质控比对注释	提取 cell barcode 和 UMI 序列，并对下机数据进行质控与参考基因组比对使用注释文件注释，获取所有 beads 的原始表达量矩阵。
细胞获取表达量计算	获取真实液滴内 beads，合并同一个液滴内的多个 beads，计算细胞表达矩阵。
质控聚类	对细胞表达矩阵进行质控，过滤低质量的细胞，对过滤后的细胞进行聚类分析和标记基因筛选以及细胞群体注释。
报告生成	数据汇总和可视化网页报告生成。

CHAPTER 2

软件安装

DNBelab_C_Series_HT_scRNA-analysis-software 分析软件可基于 conda 环境、docker/singularity 容器。

2.1 设备需求

- Language Python3(>=3.8.*),R scripts.
- x86-64 compatible processors.
- require at least 50GB of RAM and 4 CPU.
- centos 7.x 64-bit operating system (Linux kernel 3.10.0, compatible with higher software and hardware configuration).

2.2 安装说明

2.2.1 1. Conda

1.1 Conda 安装

Conda 是一个开源的软件包管理系统和环境管理系统，可在 Windows，macOS 和 Linux 上运行。Conda 可快速安装、运行和更新包及其依赖项。Conda 可轻松创建、保存、加载本地计算机上的环境并在环境之间切换。它是为 Python 程序创建的，但它可以为任何语言打包和分发软件。

```
# 下载miniconda3安装包
wget https://repo.anaconda.com/miniconda/Miniconda3-py38_4.12.0-Linux-x86_64.sh -O␣
↪Miniconda3-latest-Linux-x86_64.sh
# 安装miniconda3
# Regular installation
bash Miniconda3-latest-Linux-x86_64.sh
# Installing in silent mode
bash Miniconda3-latest-Linux-x86_64.sh -b -p $HOME/miniconda
```

1.2 dnbc4tools 环境安装

1.2.1 Clone repo

github地址

```
git clone https://github.com/MGI-tech-bioinformatics/DNBelab_C_Series_HT_scRNA-
↪analysis-software.git
chmod 755 -R DNBelab_C_Series_HT_scRNA-analysis-software
cd DNBelab_C_Series_HT_scRNA-analysis-software
```

1.2.2 Create DNBC4tools environment

基于 conda 创建 dnbc4tools 环境

```
source /miniconda3/bin/activate
conda env create -f DNBC4tools.yaml -n DNBC4tools
conda activate DNBC4tools
Rscript -e "devtools::install_github(c('chris-mcginnis-ucsf/DoubletFinder','ggjlab/
↪scHCL','ggjlab/scMCA'),force = TRUE);"
```

如果使用 wdl 去运行流程需要下载 cromwell

```
wget https://github.com/broadinstitute/cromwell/releases/download/35/cromwell-35.jar
```

1.3 更新

在版本更新时，不需要对环境进行重新安装

```
# 如果只使用命令行模式，只需要更新dnbc4tools环境的dnbc4tools
source activate DNBC4tools
pip install --upgrade -i https://pypi.tuna.tsinghua.edu.cn/simple dnbc4tools
# 如果还需要使用wdl，则还需要重新更新repo
git clone https://github.com/MGI-tech-bioinformatics/DNBelab_C_Series_HT_scRNA-
↪analysis-software.git
chmod 755 -R DNBelab_C_Series_HT_scRNA-analysis-software
```

2.2.2 2. 基于容器技术

2.1 docker

Docker 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镜像中，然后发布到任何流行的 Linux 或 Windows 操作系统的机器上，也可以实现虚拟化。

```
# 下载docker镜像
docker pull lishuangshuang3/dnbc4tools
```

2.2 singularity

singularity 是一个容器平台。Singularity 旨在以简单、可移植和可重现的方式在 HPC 集群上运行复杂的应用程序。

```
# 创建sif文件
singularity build dnbc4tools.sif docker://lishuangshuang3/dnbc4tools
```

CHAPTER 3

使用说明

软件安装完成后可使用命令行或者 WDL 流程进行分析。

conda 环境和 docker、singularity 是基于命令行模式。

WDL 流程只有主流程分析功能，不能进行数据库构建等其他功能。

3.1 命令行 DNBC4tools

- conda 运行：

```
/miniconda3/envs/DNBC4tools/bin/DNBC4tools
```

- docker 运行：

```
docker run -P  -v $Database_LOCAL:/database -v $Rawdata_LOCAL:/data -v $Result_LOCAL:/
↪result lishuangshuang3/dnbc4tools DNBC4tools
# docker通过-v挂载目录到容器内
# $Database_LOCAL: 将基因组数据库绝对路径挂载到容器/database目录下
# $Rawdata_LOCAL: 将下机原始数据绝对路径挂载到容器/data目录下
# $Result_LOCAL: 将分析结果的绝对路径挂载到容器/result目录下
# 可以使用 --user $(id -u):$(id -g) 使生成文件为使用用户属主和属组信息
```

- singularity 运行：

```
export SINGULARITY_BIND=$cDNA_data,$oligo_data,$result,$database
singularity exec dnbc4tools.sif DNBC4tools
# 通过export SINGULARITY_BIND将目录挂载到容器内，可以挂载多个目录
# 也可以在singularity exec -B $data,-B参数挂载目录或多个目录
```

3.1.1 1. DNBC4tools mkref (构建数据库)

分析需要比对参考基因组和注释文件注释分析。在分析前需要创建对应物种的参考基因组数据库。需要准备两个文件，基因组的 DNA 序列文件（FASTA 格式）和基因的注释文件（GTF 格式）。常用的 Ensembl 和 GENECODE 数据库提供了这两种格式的文件。

1.1 查看注释文件的基因类型

在构建数据库之前可以对 gtf 文件的基因类型进行过滤，先查看 gtf 中的 gene 类型确定需要过滤哪些基因类型。

```
DNBC4tools mkref --action stat --ingtf gene.gtf --type gene_type --outstat gtf_type.
↪txt
```

- DNBC4tools mkref --action stat 输入项：

参数	类型	描述
--ingtf	File Path	输入需要查看 gene 类型统计的 gtf 文件。
--type	String	gtf 中 gene 类型的 tag。
--outstat	File Path	结果保存的文件，默认为 gtf_type.txt。

- 输出结果 `gtf_type.txt`

列名	描述
Type	gtf 中的基因类型。包括 protein_coding、lncRNA、pseudogene 等。
Count	每种基因类型的数目。

可查看 gtf 的类型选取需要的基因类型

```
$cat gtf_type.txt

Type Count
protein_coding       21884
processed_pseudogene 9999
lncRNA       9949
TEC  3237
```

(续下页)

(接上页)

```
unprocessed_pseudogene       2718
miRNA        2206
snoRNA       1507
snRNA        1381
misc_RNA     562
rRNA 354
transcribed_processed_pseudogene     300
transcribed_unprocessed_pseudogene   272
IG_V_gene    218
IG_V_pseudogene      158
TR_V_gene    144
```

Notice：--type 选择需要根据 gtf 类型的 tag 来选择，如 ensemble 是`--type gene_biotype`,genecode 是`--type gene_type`。

1.2 过滤注释文件

在构建数据库之前可以对 gtf 文件的基因类型进行过滤，使其中仅包含感兴趣的基因类别，过滤哪些基因取决于您的研究问题。

软件分析中，gtf 中存在 overlap 的基因将导致 reads 被舍弃。通过过滤 gtf 文件使其只有少量重叠的基因。

```
DNBC4tools mkref --action mkgtf --ingtf gene.gtf --outgtf gene.filter.gtf \
            --attribute gene_type:protein_coding \
                        gene_type:lncRNA \
                        gene_type:IG_C_gene \
                        gene_type:IG_D_gene \
                        gene_type:IG_J_gene \
                        gene_type:IG_LV_gene \
                        gene_type:IG_V_gene \
                        gene_type:IG_V_pseudogene \
                        gene_type:IG_J_pseudogene \
                        gene_type:IG_C_pseudogene \
                        gene_type:TR_C_gene \
                        gene_type:TR_D_gene \
                        gene_type:TR_J_gene \
                        gene_type:TR_V_gene \
                        gene_type:TR_V_pseudogene \
                        gene_type:TR_J_pseudogene
```

- DNBC4tools mkref --action mkgtf 输入项:

参数	类型	描述
--ingtf	File Path	输入需要进行过滤的 gtf 文件。
--outgtf	File Path	输出过滤后的 gtf 文件。
--attribute	File Path	通过 attribute 属性来筛选基因类型，每个组合使用 tag 对应 type 冒号连接，多个类型使用空格间隔。

Notice：--type 选择需要根据 gtf 类型的 tag 来选择，如 ensemble 是`--type gene_biotype`,genecode 是`--type gene_type`。

1.3 构建数据库

使用比对软件 scStar 进行数据库的构建。scStar 的 STAR 版本为 2.7.2b，基因组版本为 2.7.1a，相同基因组版本的 STAR 构建的数据库可通用，不同的基因组版本不可互用。数据库不向下兼容 v1 版本的数据库。

```
DNBC4tools mkref --action mkref --ingtf gene.filter.gtf \
            --fasta genome.fa \
            --genomeDir $star_dir \
            --thread $threads
```

- DNBC4tools mkref --action mkref 输入项：

参数	类型	描述
--ingtf	File Path	输入构建 star 数据库的 gtf 文件。
--fasta	File Path	输入与 gtf 文件配套的参考基因组。
--genomeDir	Directory	构建数据库的结果目录。
--thread	Integer	程序运行时所调用的进程数，默认为 4。
--limitGenomeGenerateRAM	Integer	程序运行时所调用的内存大小，默认为 125000000000。

1.4 构建数据库参考文件

Ref-202203

- **Human(GRCh38)**

```
http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_32/GRCh38.
↪primary_assembly.genome.fa.gz
http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_32/gencode.v32.
↪primary_assembly.annotation.gtf.gz
```

- **Mouse(GRCm38)**

```
http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M23/GRCm38.
↪primary_assembly.genome.fa.gz
http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M23/gencode.vM23.
↪primary_assembly.annotation.gtf.gz
```

3.1.2 2. DNBC4tools run (运行主程序)

run 命令为运行主程序

```
# 主程序示例
DNBC4tools run --cDNAfastq1 cDNA_R1.fastq.gz \
             --cDNAfastq2 cDNA_R2.fastq.gz \
             --oligofastq1 oligo1_1.fq.gz,oligo2_1.fq.gz \
             --oligofastq2 oligo1_2.fq.gz,oligo2_2.fq.gz \
             --genomeDir /database/Mouse/mm10/ --gtf /database/Mouse/mm10/genes.gtf \
             --name test --species Mus_musculus --thread 10
```

分析参数如下：

- 必选参数

参数	类型	描述
--name	String	样本名称。
--cDNAfastq1	File Path	cDNA 文库 fastq 格式的 R1 端序列，多个文件使用逗号隔开。
--cDNAfastq2	File Path	cDNA 文库 fastq 格式的 R2 端序列，多个文件使用逗号隔开，顺序与 cDNAfastq1 相同。
--oligofastq1	File Path	oligo 文库 fastq 格式的 R1 端序列，多个文件使用逗号隔开。
--oligofastq2	File Path	oligo 文库 fastq 格式的 R2 端序列，多个文件使用逗号隔开，顺序与 oligofastq1 相同。
--genomeDir	Directory	参考基因组构建数据库索引路径。
--gtf	File Path	参考基因组注释文件 gtf 路径。

- 可选参数

参数	类型	描述
--species	String	样本物种名称，默认为 undefined。只有物种名为 Homo_sapiens,Mus_musculu s,Human,Mouse 时可进行细胞注释分析。
--outdir	Directory	分析结果路径，默认为当前路径。
--thread	Integer	程序运行时调用的进程数，默认为 4。
--calling_method	String	默认：emptydrops。cell calling 鉴定有效液滴内 beads 的方法，可选 barcoderanks,emptydrops。
--expectcells	Integer	默认：3000。期望细胞数，仅当 calling method 为 emptydrops 时参数有效。期望细胞数建议按照投入活细胞数量的 50% 去设置（细胞捕获率 50% 计算）。
--forcecells	Integer	默认：0。截取 beads 数量进行分析。
--chemistry	String	默认：aut o。试剂版本，建议自动获取试剂版本。该参数需要和--darkreaction 一起使用。试剂版本包括 scRNAv1HT,scRNAv2HT。
--darkreaction	String	默认：auto。暗反应设置，建议自动获取测序是否使用暗反应。该参数需要和--che mistry 一起使用。参数格式为” cDNA,oli go “, 中间使用逗号分隔，比如”R1,R1R2 ” 代表 cDNA 的 R1 使用了暗反应，oligo 的 R 1 和 R2 都使用了暗反应。包括”R1,R1R2”, ”R1,R1”, ”unset,unset” 等等。
--customize	String	使用自定义的文库结构文件进行分析。文件为 json 格式，包含结构位置、汉明距离允许错配碱基数量和 cell barcode 白名单信息。参数格式为” cDNA,oligo “，比如” scRNA _beads_readStructure.json,scRNA_oli go_readStructure.json “。customize 的优先级高于 chemistry 和 darkreaction。
--process	String	默认：data,count,anlysis ,report。选择需要分析的步骤，可选择 data,count,anlysis,report 其中几项（该参数常用于已分析完成需要重新调整参数时使用，更改某一步骤参数后面的步骤也需要重新分析），用逗号分隔。
--mtgenes	String	默认：auto。线粒体基因列表文件，auto 表示选择基因名前缀为 mt 或 MT 的基因作为线粒体基因。

- flag 参数

参数	类型	描述
--no_introns	Flag	比对到 intronic 区域的 reads 不纳入进表达量矩阵计算。
--no_bam	Flag	添加该参数则不会将 02.count 中的 anno_decon_sorted.bam 和 anno_decon_sorted.b am.bai 移动到 output 目录中。后续使用 DNBC4tools clean 时则会删除该 bam 文件减少存储占用。
--dry	Flag	不进行流程分析。只打印分析步骤的 shell 文件。

对参数的详细说明：

- `--chemistry` 和`--darkreaction` 需要一起使用。建议使用自动检测的试剂版本和测序暗反应。当测序时 R1 没有进行暗反应才可检测到试剂版本，scRNAv1HT,scRNAv2HT 的 cDNA 和 oligo 的 R1 端在有暗反应的情况下结构是一样的。
- `--customize`，在使用 customize 参数时，chemistry 和 darkreaction 参数是无法起作用的。json 文件的格式内容可参考常见问题说明。
- `--callling_method`，在默认情况下会使用 emptydrops，如果对结果不满意也可以尝试 barcoderanks。两种 cell calling 方法的原理请参考常见问题说明。
- `--mtgenes`，默认为 auto，表示选择基因名前缀为 mt 或 MT 的基因作为线粒体基因。也可以使用自定义 mtgenes 的列表文件。文件内容如下：

```
mt-Nd1
mt-Nd2
mt-Co1
mt-Co2
mt-Atp8
mt-Atp6
mt-Co3
```

- `--no_introns`，分析中默认会将比对到内含子的 reads 加入到表达量矩阵分析。虽然不推荐，但用户可以使用这个参数将内含子数据丢弃。
- `--species`，信息会展示在结果报告中，如果信息为 Homo_sapiens,Mus_musculus,Human,Mouse 会进行细胞群体注释分析。
- `--process`，默认：data,count,anlysis,report。选择需要分析的步骤，可选择 data,count,anlysis,report 其中几个步骤。

每个步骤可单独使用 DNBC4tools 去分析，具体见以下步骤。

DNBC4tools data

提取 barcode 和 UMI 序列，并对下机数据进行质控与参考基因组进行比对注释，获取所有 beads 的原始表达量矩阵。

参数保持与 DNBC4tools run 一致。

DNBC4tools count

确定有效液滴内 beads，合并同一个液滴内的多个 beads 计算细胞表达矩阵。

分析参数如下：

参数	类型	描述
--bam	File Path	必选，data 步骤生成的 final.bam 文件。
--raw_matrix	Directory	必选，da ta 步骤生成的 raw_matrix 矩阵目录。
--cDNAbarcodeCount	File Path	必选，data 步骤生成的 c DNA_barcode_counts_raw.txt 文件。
--Indexreads	File Path	必选，data 步骤生成的 Index_reads.fq.gz 文件。
--oligobarcodeCount	File Path	必选，data 步骤生成的 Ind ex_barcode_counts_raw.txt. 文件。
--minumi	Integer	可选，默认 1000。cell calling 中 emptydrops 方法中可获取的 beads 最小的 umi 数量。

DNBC4tools analysis

对细胞表达矩阵进行质控，过滤低质量的细胞根据表达矩阵进行细胞聚类分析和标记基因筛选。

分析参数如下：

参数	类型	描述
--matrix	Directory	必选，count 步骤生成的 filter_matrix 表达量矩阵目录。
--qcdim	String	可选，默认 20。DoubletFin der 的 PCs 参数显著的主成分的数量。
--clusterdim	Integer	可选，默认 20。用于 PCA 降维后的降维聚类使用的显著主成分的数量。
--doubletpercentage	Float	可选，默认：0.05。预测双胞比例。
--mitpercentage	Integer	可选，默认：15。过滤线粒体基因比例。
--minfeatures	Integer	可选，默认：2 00。细胞含有的基因数目的最小值。
--PCusage	Integer	可选，默认：50。用于 PCA 降维的主成分的数量。
--resolution	Integer	可选，默认：0.5。细胞聚类分辨率。该参数设置下游聚类的细胞群体数量，增加该值导致更多的分群。

DNBC4tools report

数据汇总和可视化网页报告生成。

参数保持与 DNBC4tools run 一致。

Notice：在 data,count,analysis,report 中有些参数在主程序 run 中没有。通常情况下这些参数使用默认值分析即可。如果需要修改这些参数，可使用 data,count,analysis,report 模块进行分析，再使用 run -process 参数将后续的结果分析。例如，使用 run 得到分析结果和报告后，对细胞分群的结果不满意，可使用 DNBC4tools analysis –resolution 调整分群的分辨率，分析完成后在使用 DNBC4tools run –process report 完成后续的 report 分析。

3.1.3 3. DNBC4tools multi (对多个样本生成 DNBC4tools run)

```
# 分析示例
DNBC4tools multi --list samplelist
             --genomeDir /database/Mouse/mm10/ --gtf /database/Mouse/mm10/genes.gtf \
             --thread 10
```

其中 samplelist 的格式如下：

```
test1 cDNA1_L01_1.fq.gz;cDNA1_L01_2.fq.gz    oligo1_L01_1.fq.gz,oligo1_L02_1.fq.gz;/
↪oligo1_L01_2.fq.gz,oligo1_L02_2.fq.gz Mouse
test2 cDNA2_L01_1.fq.gz,cDNA2_L02_1.fq.gz;cDNA1_L01_2.fq.gz,cDNA2_L02_2.fq.gz  ␣
↪oligo2_L01_1.fq.gz;/oligo2_L01_2.fq.gz  Mouse
test3 cDNA3_L01_1.fq.gz;cDNA3_L01_2.fq.gz    oligo3_L01_1.fq.gz,oligo3_L02_1.fq.gz;/
↪oligo3_L01_2.fq.gz,oligo3_L02_2.fq.gz Mouse
```

- 文件包含四列，使用水平制表符 (t) 进行分隔
- 不设置表头，第一列为样本名称，第二列为 cDNA 文库信息，第三列为 oligo 文库信息，第四列为物种名称。
- cDNA 文库和 oligo 文库，多个 fastq 以逗号分隔，R1 和 R2 以分号分隔。R1 和 R2 中的多个 fastq 顺序需保持一致。
- 分析样本物种名须保持一致，因为`--genomeDir` 和`--gtf` 只能分析一个物种。

3.1.4 4. DNBC4tools clean (分析完成后清理中间文件)

对分析中的存储较大的中间文件进行清除。确定不需要对结果重新分析时使用。

```
### 删除该目录下所有样本的中间大文件
DNBC4tools clean
### 删除该目录下样本sampleA的中间大文件
DNBC4tools clean --name sampleA
```

分析参数如下：

参数	类型	描述
--name	String	可选，默认该目录下的所有样本。需要进行中间文件清楚的样本名，多个样本使用逗号连接。
--outdir	Directory	可选，默认当前路径。分析结果的输出目录。
--combine	Flag	对选择的样本的统计文件 metrics_summary.xls 进行合并且将样本的网页报告拷贝到 result 目录中。

3.2 WDL 流程

工作流描述语言（Workflow Description Language），简称 WDL，是一门开源的、标准化的以及人类可读写的用于描述任务和工作流的编程语言。

3.2.1 1. 准备配置文件

config.json 文件包含一下内容，可拷贝/example/wdl/config.json 文件然后修改：

- 必选参数

参数	类型	描述
main.Outdir	Directory	输出结果目录的路径。
main.SampleName	String	样本名，不允许有空格。
main.cDNA_Fastq1	Fastq	cDNA 文库 fastq 格式的 R1 端序列，多个文件使用逗号隔开。
main.cDNA_Fastq2	Fastq	cDNA 文库 fas tq 格式的 R2 端序列，多个文件使用逗号隔开，顺序与 main.cDNA_Fastq1 相同。
main.Oligo_Fastq1	Fastq	oligo 文库 fastq 格式的 R1 端序列，多个文件使用逗号隔开。
main.Oligo_Fastq2	Fastq	oligo 文库 fas tq 格式的 R2 端序列，多个文件使用逗号隔开，顺序与 main.oligo_Fastq1 相同。
main.BeadsBarcode	JSON file	cDNA 文库结构文件路径，为 json 格式文件，包含结构位置、汉明距离允许错配碱基数量和 cell barcode 白名单信息。
main.OligoBarcode	JSON file	oligo 文库结构文件路径，为 json 格式文件，包含结构位置、汉明距离允许错配碱基数量和 cell barcode 白名单信息。
main.Root	Directory	DNBelab C4 分析流程路径。
main.Refdir	Directory	参考基因组构建数据库索引路径。
main.Gtf	File Path	参考基因组注释文件 gtf 路径。
main.Species	String	样本物种名。

- 可选参数

参数	类型	描述
main.expectCellNum	Integer	默认：3000。期望细胞数，仅当 calling_met hod 为 emptydrops 时参数有效。
main.calling_method	String	默认：emptydrops。cell calling 鉴定有效液滴内 beads 的方法，可选 b arcoderanks 和 emptydrops。
main.forceCellNum	Integer	默认：0。截取 beads 数量。
main.Intron	Boolean	默认：true。是否将比对到 intr onic 区域的 reads 纳入分析。
main.mtgenes	String	默认：auto 。线粒体基因列表文件，aut o 表示选择基因名前缀为 mt 或 MT 的基因作为线粒体基因。
main.Oligo_type8	File Path	默认：D NBC4tools/config/oligo_ty pe8.txt。oligo 文库 droplet index 白名单信息文件路径。
main.Adapter	File Path	默认：DNBC4tools/config/adapter .txt。Adapter 列表文件路径
main.clusterdim	Integer	默认：20 。用于 PCA 降维后的降维聚类使用的显著主成分的数量。
main.doublepercentage	Float	默认：0.05。预测双胞比例。
main.mitpercentage	Integer	默认：15。过滤线粒体基因比例。
main.minfeatures	Integer	默认：200。细胞含有的基因数目的最小值
main.PCusage	Integer	默认：50。用于 PCA 降维的主成分的数量。
main.resolution	Integer	默认：0.5。细胞聚类分辨率。该参数设置下游聚类的细胞群体数量，增加该值能得到更多的分群。

准备主分析脚本 run.sh：

```
# export环境变量，将所有路径替换成真实分析路径
export PATH=/miniconda3/envs/DNBC4tools/bin:$PATH
export LD_LIBRARY_PATH=/miniconda3/envs/DNBC4tools/lib:$LD_LIBRARY_PATH
java -jar /pipeline/wdl/cromwell-35.jar run -i config.json /pipeline/wdl/DNBC4_scRNA.
↪wdl
```

针对多个样本，使用 samplelist 样本列表文件，参考 *DNBC4tools multi*，修改 scripts 目录下的 `wdl.json`，替换成真实路径，使用如下命令：

```
/miniconda3/envs/DNBC4tools/bin/python creat_wdl_json.py --infile samplelist --outdir␣
↪outdir
```

3.2.2 2. 运行分析流程

```
### 运行分析流程
sh run.sh
```

- 执行 `sh run.sh` 后，会在当前目录下生成 DNBelab C4 分析的执行目录 `cromwell-executions` 该目录下为主进程 `main` 目录，在主进程 `main` 目录下包含 workflow id，每次运行 run.sh 都会自动生成一个 workflow id。每个 workflow id 下包含 4 个功能模块对应的任务运行脚本及运行日志。
- `symbol` 记录每个功能步骤是否完成的标志文件。重新分析时如果 symbol 目录中存在该步骤完成标志文件，则该分析步骤跳过。
  - `01.oligoparse_sigh.txt`，oligo 文库数据质控过滤。
  - `02.cDNAAnno_sigh.txt`，cDNA 文库数据质控过滤、比对和注释生成所有 beads 的表达量矩阵。
  - `03.M280UMI_stat_sigh.txt`，cell calling 获取有效液滴内 beads、合并同一液滴内 beads。
  - `04.count_matrix_sigh.txt`，生成细胞表达量矩阵。
  - `04.saturation_sigh.txt`，饱和度分析。
  - `05.Cluster_sigh.txt`，过滤后细胞降维聚类注释。
  - `05.QC_sigh.txt`，对细胞进行过滤。
  - `06.splice_matrix_sigh.txt`，exonic 区域的表达量矩阵、RNA velocity 分析的表达量矩阵。
  - `07.report_sigh.txt`，分析结果整理，生成网页报告。

CHAPTER 4

结果说明

DNBelab C4 流程分析顺利执行完后，指定的输出目录结构如下：

- **01.data** 提取 barcode 和 UMI 序列，并对下机数据进行质控，与参考基因组进行比对注释生成的 bam 文件，获取所有 beads 的原始表达量矩阵 raw_matrix。
- **02.count** 确定真实有效 beads 结果文件，合并同一个液滴内的多个 beads 的组合文件将细胞 tag 添加到 bam 文件。生成细胞表达矩阵的结果目录 filter_matrix。
- **03.analysis** 对细胞表达矩阵进行质控，过滤低质量的细胞根据表达矩阵进行细胞聚类分析和 marker 基因的结果文件。
- **04.report** 数据汇总和可视化网页报告的结果文件。
- **output** 输出分析结果文件目录。
- **log** 分析日志，分析使用的脚本和程序以及分析开始完成时间。

4.1 1. 运行步骤说明

4.1.1 1.1 数据质控比对注释

1.1.1 功能描述

- 使用 scStar 去除 cDNA 文库原始数据中低质量、cell barcode 含 N 的 reads；提取 reads 中的 cell barcode 和 UMI 序列，过滤掉带有无法识别或矫正后仍无法与 white list 比对上的 reads，并对质控结果进行统计

将合格的 reads 比对到参考基因组。使用 Anno 根据基因注释信息文件进行注释，进行 UMIs 校正，并对 beads 的原始 reads 数目，UMIs 数以及基因数目进行统计。使用 PISA count 生成所有 beads 的表达矩阵。

- 使用 parseFq 去除 oligo 文库原始数据中低质量、cell barcode 含 N 的 reads；提取 reads 中的 cell barcode 序列，过滤掉带有无法识别或矫正后仍无法与 white list 比对上的 reads，并对质控结果进行统计。

1.1.2 输入项

1）**scStar** 对 cDNA 文库数据进行质控并提取 cell barcode 和 umi 序列信息，将有效的数据与参考基因组进行比对生成 bam 文件。

输入参数如下:

参数	类型	描述
--outSAMattributes	String	默认值：single Cell。输出 BAM 文件的格式 ta g 包含 UR(UMI 序列) 和 CB(cell barcode 的序列)。
--outSAMtype	String	默认值：BAM Unsorted。输出结果格式，默认为 BAM 格式且不排序。
--genomeDir	Directory	参考基因组 scStar index 目录。
--runThreadN	Integer	程序运行时调用的进程数。
--limitOutSJcollapsed	Integer	默认值：10000000。最大检测到的剪接点数量。
--outFileNamePrefix	String	输出结果目录。
--stParaFile	File Path	scStar 质控的配置文件。
--limitIObufferSize	Integer	默认值：350000000 。每个线程的输入/输出的最大可用缓冲区大小 (字节)。
--outSAMmode	String	默认值：NoQC。输出 BAM 文件中不包含质量数据。

Notice：`--outSAMtype` 在分析时不能进行排序，因为后续针对 multi mapping 的 reads 会进行矫正舍去步骤需要同一条 reads 的比对在相邻位置。`--outSAMmode` 不包含质量值是为了降低 bam 存储的空间。

2）**Anno** 对 scStar 生成的 Aligned.out.bam 注释，对 umi 序列进行汉明距离为 1 的碱基错配纠正。

输入参数如下:

参数	类型	描述
-I	File Path	scStar 生成的 bam 文件路径。
-A	File Path	GTF 注释文件。
-L	File Path	scStar 生成的 cDNA_barcode_counts_raw.txt 文件路径。
-O	File Path	输出结果目录。
-C	Integer	程序运行时调用的进程数。
-M	String	默认值：chrM。线粒体染色体名称。
-B	File Path	cDNA 文库结构文件路径，为 json 格式文件，包含结构位置、汉明距离允许错配碱基数量和 cell barcode 白名单信息。
--intron	Flag	默认添加该参数。将比对到 intronic 区域的 reads 纳入分析。
--anno	Integer	数字 0 为使用 v1 版本的注释逻辑，数字 1 使用 v2 版本注释逻辑。

3）**PISA count** 对注释后的 final.bam 分析获取所有 beads 的表达矩阵。

输入参数如下：

参数	类型	描述
-@	Integer	解析 bam 文件时调用的进程数。
-cb	TAG	默认值：CB。定义 cell barcode 在 bam 记录中标签名称。
-anno_tag	TAG	默认值：GN。定义注释标签在 DNBelab C4 中为基因名。
-umi	TAG	默认值：UB。定义 UMI 标签，若同一个 anno_t ag 有超过一个记录有相同标签，则只计数一次。
-outdir	Director	表达矩阵输出目录。
-bam	File Path	输入的 bam 文件。

Notice：PISA 软件可参考PISA Wiki

1.1.3 输出项

- **cDNA_barcode_counts_raw.txt** cDNA 文库磁珠 cell barcode 对应 reads 数目文件，第一列为磁珠的 cell barcode 序列，第二列为带有该 barcode 的 reads 数。
- **Index_barcode_counts_raw.txt** oligo 文库磁珠 cell barcode 对应 reads 数目文件，第一列为磁珠的 cell barcode 序列，第二列为带有该 barcode 的 reads 数。
- **cDNA_sequencing_report.csv** cDNA 文库质控后统计文件。
- **Index_sequencing_report.csv** oligo 文库质控后统计文件。
- **Index_reads.fq.gz** oligo 文库通过质控并完成 cell barcode、droplet index 和 umi 提取的 fastq 格式的序列文件。
- **final.bam** cDNA 文库数据比对注释后的 bam 文件。
- **alignment_report.csv** 比对统计结果。
- **anno_report.csv** 功能区域注释统计结果文件。
- **beads_stat.txt** beads 统计结果。
- **Log.final.out** STAR 比对结束后比对统计信息。
- **Log.out** STAR 软件运行时的信息。
- **Log.progress.out** STAR 运行进程监控文件。
- **raw_matrix** 所有 beads 的表达矩阵文件目录。

部分结果内容展示：

1）`sequencing_report.csv` 内容如下：

- **Number of Fragments** 下机数据 reads 总数。
- **Fragments pass QC** 通过质控的 reads 数目。
- **Fragments Filtered on Low Qulity** cell barcode 含 N 或不满足质量值条件而被舍弃的 reads 数目。
- **Fragments with Failed Barcodes** 配对 cell barcode 白名单失败的 reads 数目。
- **Fragments too short after Adapter Trimming** 序列中存在接头序列并切除接头序列剩下区域过短的 reads 数目。
- **Fragments with Exactly Matched Barcodes** 完全匹配上不需要错配纠错的 cell barcode 的 reads 数目。
- **Fragments with Adapter** 序列中存在接头序列的 reads 数目占比。
- **Q30 bases in Cell Barcode** cell barcode 区域碱基质量值 > 30 的碱基个数占 cell barcode 区域碱基总数百分比。
- **Q30 bases in Sample Barcode** 样本 barcode 区域碱基质量值 > 30 的碱基个数占样本 barcode 区域碱基总数百分比。

- **Q30 bases in UMI** UMI 区域碱基质量值＞ 30 的碱基个数占 UMI 区域碱基总数百分比。
- **Q30 bases in Reads** 碱基质量值＞ 30 的碱基个数占总碱基总数百分比。

2）`alignment_report.csv` 内容如下:

- **Raw reads** bam 文件中所有的比对条目数目（为了降低 bam 文件的存储大小，bam 文件中不包含未比对上的 reads，所以 Raw reads 和 Mapped reads 的数目相同）。
- **Mapped reads** 成功比对上的 reads 百分比。
- **Plus strand** 比对上参考基因组正链的 reads 数目。
- **Minus strand** 比对上参考基因组负链的 reads 数目。
- **Mitochondria ratio** 比对上参考基因组中线粒体染色体的 reads 比例（默认线粒体染色体名称为 chrM）。
- **Mapping quality corrected reads** 比对到多个位置的 reads，将比对到外显子区域的条目设置成 primary hit 并将 MAPQ 调整成 255，统计调整质量值的 reads 数目。

3）`anno_report.csv` 内容如下:

- **Reads Mapped to Genome (Map Quality >= 0)** 比对上参考基因组的 reads 比例（为了降低 bam 文件的存储大小，bam 文件中不包含未比对上的 reads，所以该值为 0。
- **Reads Mapped Confidently to Exonic Regions** 比对上外显子区域的 reads 比例。
- **Reads Mapped Confidently to Intronic Regions** 比对上内含子区域的 reads 比例。
- **Reads Mapped to both Exonic and Intronic Regions** 同时比对上外显子和内含子的 reads 比例（在 v2 中由于注释逻辑的更改，该值为 0.0%）。
- **Reads Mapped Antisense to Gene** 比对上基因的 reads 中，比对上反义链的比例。
- **Reads Mapped to Intergenic Regions** 比对上基因间区的 reads 比例。
- **Reads Mapped to Gene but Failed to Interpret Type** 比对上基因但没有注释信息的 reads 比例（在 v2 中由于注释逻辑的更改，该值为 0.0%）。

4.1.2 1.2 细胞获取表达量计算

1.2.1 功能描述

分析 raw matrix 矩阵，区分有效液滴内和背景的 beads 使用 barcoderanks 或 emptydrops 方法进行 cell calling。计算 beads 之间的相似度，根据 beads 间的相似度对同一液滴内的 beads 合并，合并后的生成的 bam 计算细胞表达量矩阵。

1.2.2 输入项

1）**cell_calling.R** 对所有 beads 表达量矩阵计算区分有效液滴内 beads 和背景 beads。

输入参数如下：

参数	类型	描述
--matrix	Directory	所有 beads 表达量矩阵目录。
--outdir	Directory	分析输出结果目录。
--method	String	默认值：emptydrops。cell calling 使用方法，包含 barcoderanks 和 emptydrops。
--expectcells	Integer	默认值：3000。期望获取 beads 数。
--forcecells	Integer	默认值：0。截取 beads 数。
--minumi	Integer	使用 emptydrops 方法时，定义 beads 最小的 umi 数目，低于该数目的 beads 过滤舍弃。

2）**mergeBarcodes** oligo 数据的 cell barcode 和 droplet index 对应统计 counts 数目。

输入参数如下：

参数	类型	描述
-b	File Path	所有 beads 的 cell barcode 列表，用于过滤 olligo 的 cell barcode。
-f	File Path	质控并完成 cell barcode、droplet index 和 umi 提取的 fastq 格式的序列文件。
-n	String	样本名称。
-o	Directory	分析输出结果目录。

3）**s1.get.similarityOfBeads** 计算 beads 之间相似度（同一液滴内的 beads 具有较一致的 oligo droplet index）。

输入参数如下：

参数	类型	描述
Sample name	String	输入样本名称。
CB_UB_count.txt	File Path	cell barcode 和 droplet inde x 对应统计 counts 数目文件。
beads_barcodes.txt	File Path	cell calling 获取的有效液滴内 beads 列表。
oligo_type8.txt	File Path	oligo droplet index 白名单文件。
Similarity.all.csv	File Path	输出所有 cell barcode 之间存在的相似度统计结果文件。
Similarity.droplet.csv	File Path	输出初步过滤后的 cell barcode 相似度统计结果文件（根据有效液滴内 beads 过滤）。
Simila rity.droplet.filtered.csv	File Path	对 Similarity.droplet. csv 中存在的相同条目进行去除（第一列和第二列的 cell barcode 互换类型）。
-n	Inte- ger	程序运行所调用的进程数。

4）**combinedListOfBeads.py** 对相似度进行过滤并生成液滴内的 beads 对应信息。

输入参数如下:

参数	类型	描述
-- similarity_droplet	File Path	初步过滤后的 cell ba rcode 相似度统计结果文件（根据真实有效 beads 过滤）。
--beads_list	File Path	cell calling 获取的有效液滴内 beads 列表。
--combined_list	File Path	液滴内的 beads 对应信息列表。
--simi_threshold	Float	默认值：0.2。相似度过滤阈值。

5）**tagAdd** 将细胞 tag 信息存入 bam 文件中

输入参数如下:

参数	类型	描述
-bam	File Path	输入 final.bam 文件。
-file	File Path	有效液滴内 beads 对应信息列表。
-out	File Path	输出添加了细胞 tag 信息后的 bam 文件。
-tag_check	TAG	默认值：CB:Z:。beads 的 cell barcode 信息。
-tag_add	TAG	默认值：DB:Z:。添加细胞 tag 信息。
-n	Integer	程序运行所调用的进程数。

1.2.3 输出项

- **beads_barcodes.txt** 有效液滴内 beads 的 cell barcode 信息文件。
- **beads_barcodes_hex.txt** 有效液滴内 beads 的十六进制 cell barcode 信息文件。
- **cutoff.csv** 按照 umi 数量排序的 cell barcode 和是否为有效液滴内 beads。
- **beads_barcode_all.txt** 所有 beads 的 cell barcode 信息。
- **CB_UB_count.txt** oligo 的 cell barcode 和 droplet index 组合 count 统计表，统计每个磁珠 cell barcode 捕获到的 droplet index 序列的 UMIs 数目。第一列表示 droplet index UMI 数量，第二列是 droplet index 序列，第三列是磁珠 cell barcode 序列。
- **Similarity.all.csv** 所有 beads 之间存在的相似度统计结果文件。
- **Similarity.droplet.csv** 初步过滤后的 cell barcode 相似度统计结果文件（有效液滴内 beads 过滤）。
- **Similarity.droplet.filtered.csv** 对 Similarity.droplet.csv 中存在的相同条目进行去除（第一列和第二列的 cell barcode 互换类型）。
- **combined_list.txt** 在同一液滴中的 beads 的 cell barcode 组合的文件。第一列磁珠 cell barcode，第二列为 cell ID。
- **barcodeTranslate_hex.txt** barcodeTranslate.txt 中 beads 的 cell barcode 为十六进制。
- **barcodeTranslate.txt** 在同一液滴中的 beads 的 cell barcode 组合的文件。第一列磁珠 beads barcode，第二列为 cell ID。（v2 版本中与 combined_list.txt 相同）
- **cellNumber_merge.png** 每个 cell 含有 beads 数量统计结果条形图 png 格式图片。
- **cellNumber_merge.pdf** 每个 cell 含有 beads 数量统计结果条形图 pdf 格式图片。
- **filter_matrix** 细胞表达量矩阵目录。
- **cellCount_report.csv** 细胞统计信息文件。
- **anno_decon_sorted.bam** 排序后的 anno_decon.bam（将合并后的细胞 tag 信息存入 bam 文件）文件。
- **cell_count_detail.xls** 每个细胞中基因、umi 的组合测序 reads 的数量。
- **saturation.xls** 不同 fraction 饱和度分析结果文件。

部分结果内容展示：

1）`cellCount_report.csv` 内容如下:

- **Fraction Reads in Cells** 位于有效液滴内 beads 且比对上转录本的 reads 和所有比对上转录本的 reads 的比值。
- **Estimated Number of Cells** 鉴定的细胞数量。
- **Total Reads Number of Cells** 所有比对上细胞的 reads 数量。
- **Mean reads per cell** 每个细胞中平均的 reads 数量。

- **Mean UMI counts per cell** 每个细胞中平均的 umi 数量。
- **Median UMI Counts per Cell** 细胞中 umi 数量的中位数。
- **Total Genes Detected** 统计所有比对上的基因数量。
- **Mean Genes per Cell** 每个细胞中平均的基因数量。
- **Median Genes per Cell** 细胞中基因数量的中位数。

4.1.3 1.3 质控聚类

1.3.1 功能描述

对细胞表达矩阵进行质控过滤双胞，低质量的细胞。降维聚类区分不同细胞群体以及输出候选的各细胞群标记基因，对细胞群体注释。

1.3.2 输入项

1）**QC_analysis.R** 对细胞表达矩阵进行过滤。

输入参数如下:

参数	类型	描述
-I	Directory	细胞表达矩阵文件目录。
-D	Integer	默认值：20。Dou bletFinder 预测双胞的 PCs 参数显著的主成分的数量。
-P	Float	默认值：0.05。预测双胞比例。
-M	String	默认值：auto。线粒体基因列表文件，auto 表示选择基因名前缀为 mt 或 MT 的基因作为线粒体基因。
-MP	Integer	默认值：15。过滤线粒体基因比例。
-F	Integer	默认值：200。细胞含有的基因数目的最小值。
-B	String	样本名称。
-O	Director	输出文件路径。

2）**Cluster_analysis.R** 降维聚类区分不同细胞群体以及输出候选的各细胞群标记基因。

输入参数如下:

参数	类型	描述
-I	Directory	QC 分析结果目录。
-D	Integer	默认值：20。用于 PCA 降维后的降维聚类使用的显著主成分的数量。
-PC	Float	默认值：50。用于 PCA 降维的主成分的数量。
-RES	Float	默认值：0.5。细胞聚类分辨率。该参数设置下游聚类的细胞群体数量，增加该值能得到更多的分群。
-O	Director	输出文件路径。
-SP	String	输入样本物种名称。只有 Homo_sapiens,Mus _musculus,Human,Mouse 可以进行细胞群体注释分析。

1.3.2 输出项

1）过滤输出文件位于输出目录下的 **QC** 目录内。

- **raw_QCplot.png** 所有细胞的基因、UMIs 数目和线粒体比例小提琴图。如果未识别到线粒体基因将没有线粒体比例小提琴图。
- **filter_QCplot.png** 过滤后细胞的基因、UMIs 数目和线粒体比例小提琴图。如果未识别到线粒体基因将没有线粒体比例小提琴图。
- **doublets_info.txt** 双胞统计结果文件。第一列为细胞名称，最后一列为是否鉴定为双胞。
- **QCobject.RDS** rds 格式的文件用于存储 QC 的结果用于后续降维聚类分析。

2）降维聚类输出文件位于输出目录下的 **Clustering** 目录内。

- **clustering_plot.png** 细胞聚类结果的 UMAP 展示图片。
- **cluster.csv** 记录每个细胞 meta 数据的表格文件（包括群体、umap 坐标、umi 数量、基因数量和预测的细胞类型）。
- **cluster_cell.stat** 细胞聚类的结果及每个类群的细胞数目统计。
- **marker.csv** 所有 marker 基因的表格文件第一列为基因名，第二列为群体、第三列矫正后的 p_value，第四列为 p_value，第五列为该基因在该群体与其他群体之间的差异倍数，第六列 pct.1 为在当前 cluster 细胞中检测到该基因表达的细胞比例，第七列 pct.2 为在其它 cluster 细胞中检测到该基因表达的细胞比例。
- **cell_report.csv** 用于降维聚类的细胞个数统计。
- **cluster_annotation.png** 细胞聚类且注释后的 UMAP 展示图片。
- **clustering_annotation_object.RDS** rds 格式的文件用于存储降维聚类注释的结果用于后续复现该分析结果。

4.1.4 1.4 报告生成

1.4.1 功能描述

对前三个步骤的分析结果进行整理整合，生成 html 格式的分析报告。

1.4.2 输出项

- **scRNA_report.html** 网页分析报告。
- **anno_decon_sorted.bam** 排序后的 anno_decon.bam 文件，可用于后续分析。
- **anno_decon_sorted.bam.bai** 排序后的 anno_decon.bam 文件的 bai 文件。
- **attachment** 目录内包括细胞过滤的结果 QC 和降维聚类的结果 Clustering。如果分析包含 intronic reads，目录内会增加 exonic 区域的 reads 的表达量矩阵 splice_matrix 以及用于 RNA velocity 分析的表达量矩阵 RNAvelocity_matrix。
- **filter_feature.h5ad** h5ad 格式的细胞表达量矩阵。
- **filter_matrix** 细胞表达量矩阵。
- **metrics_summary.xls** 部分参数的结果统计。
- **raw_matrix** 所有 beads 的表达量矩阵。

4.2 2. 结果报告说明

网页报告由 **SUMMARY** 和 **ANALYSIS** 组成，可点击切换。

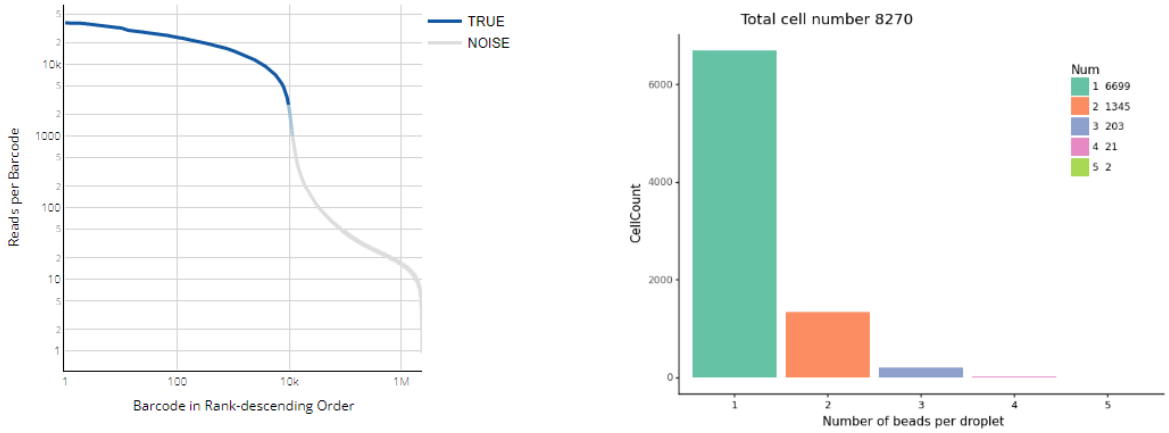
SUMMARY 包括 **Sample information**、**Beads to cells**、**Summary**、**Sequencing** 和 **Mapping & Annotation** 五部分。

4.2.1 2.1 Sample information

- **Estimated number of cell** 细胞数目。
- **Median UMI counts per cell** 细胞 UMIs 中位数。
- **Median genes per cell** 细胞基因中位数。
- **Mean reads per cell** 细胞平均 reads 数。

4.2.2 2.2 Beads to cells

Beads to cells



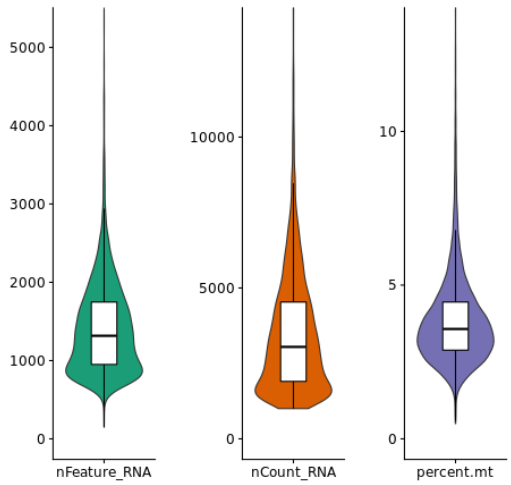
- 左图展示了 beads 的 UMIs 数目分布，并推测出存在细胞的液滴内的磁珠深蓝色区域）、低 UMIs 和背景磁珠混合区域（浅蓝色渐变区域）、背景磁珠（位于空液滴的磁珠，灰色区灰色区域）。来自同一细胞的不同 mRNA 会带有相同的磁珠条形码序列和随机的 UMI 序列，但由于建库过程中存在的凋亡损伤细胞所释放到背景环境中的 mRNA 会混入反应体系中，所以空液滴内磁珠也会捕获到环境中的 mRNA。
- 右图展示了每个有效液滴中包含的磁珠数目统计。

4.2.3 2.3 Summary

- **Sample name** 样本名称。
- **Species** 样本物种名称。
- **Estimated number of cell** 鉴定到细胞数目。
- **Mean reads per cell** 细胞平均 reads 数目。
- **Mean UMI count per cell** 细胞平均 UMI 数目。
- **Median UMI counts per cell** 细胞 UMI 中位数。
- **Total genes detected** 检测到的总基因种类数目。
- **Mean genes per cell** 细胞平均基因数目。
- **Median genes per cell** 细胞基因中位数。
- **Fraction Reads in cells** 比对到转录本上的 reads 位于有效液滴内 beads 的比例。
- **Sequencing saturation** 测序饱和度。
- **Number of cells used for clustering** 质控后用于聚类分析的细胞数目。

Summary

Sample name	demo
Species	Human
Estimated number of cells	7,564
Mean reads per cell	46,507
Mean UMI count per cell	3,557
Median UMI counts per cell	3,047
Total genes detected	33,151
Mean genes per cell	1,420
Median genes per cell	1,317
Fraction Reads in cells	75.8%
Sequencing saturation	91.18%
Number of cells used for clustering	6,802



4.2.4 2.4 Sequencing

- **Number of reads** 下机数据 reads 总数。
- **Reads pass QC** 通过质控的 reads 数目。
- **Reads with exactly matched barcodes** 完全匹配上不需要错配纠错的 cell barcode 的 reads 数目。
- **Reads with failed barcodes** 配对 cell barcode 白名单失败的 reads 数目。
- **Reads filtered on low quality** cell barcode 含 N 或不满足质量值条件而被舍弃的 reads 数目。
- **Q30 bases in Cell Barcode** cell barcode 区域碱基质量值 > 30 的碱基个数占 cell barcode 区域碱基总数百分比。
- **Q30 bases in UMI** UMI 区域碱基质量值 > 30 的碱基个数占 UMI 区域碱基总数百分比。
- **Q30 bases in reads** 序列碱基质量值 > 30 的碱基个数占总碱基总数百分比。

Sequencing

mRNA	
Number of reads	605,713,832
Reads pass QC	92.64%
Reads with exactly matched barcodes	80.53%
Reads with failed barcodes	6.7%
Reads filtered on low quality	0.66%
Q30 bases in cell barcode	91.25%
Q30 bases in UMI	90.37%
Q30 bases in reads	92.65%

Droplet index	
Number of Reads	227,294,205
Reads pass QC	96.75%
Reads with exactly matched barcodes	91.25%
Reads with failed barcodes	2.84%
Reads filtered on low quality	0.41%
Q30 bases in cell barcode	95.88%
Q30 bases in reads	97.25%

4.2.5 2.5 Mapping & Annotation

Mapping & Annotation

Reads pass QC	561,134,393
Mapped reads	537,174,377
Plus strand	51.12%
Minus strand	48.88%
Mitochondria ratio	3.33%
Mapping quality corrected reads	4.45%

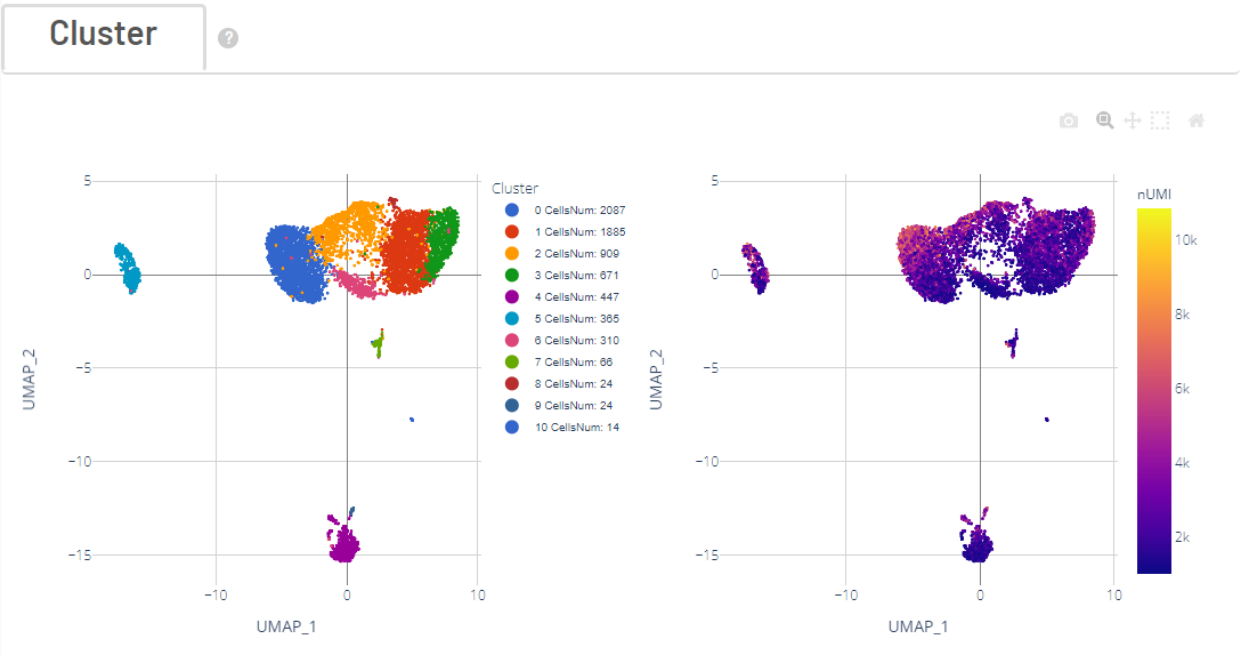
Reads mapped to genome (Map Quality ≥ 0)	95.73%
Reads mapped to exonic regions	65.4%
Reads mapped to intronic regions	12.8%
Reads mapped antisense to gene	10.6%
Reads mapped to intergenic regions	21.8%
Include introns	True

- **Reads pass QC** 通过质控的 reads 数目。
- **Mapped reads** 比对上参考基因组的 reads 数目。
- **Plus strand** 比对上参考基因组正链的 reads 数目。
- **Minus strand** 比对上参考基因组负链的 reads 数目。
- **Mitochondria ratio** 比对上参考基因组中线粒体染色体的 reads 比例（默认线粒体染色体名称为 chrM）。
- **Mapping quality corrected reads** 比对到多个位置的 reads，将比对到外显子区域的条目设置成 primary hit 并将 MAPQ 调整成 255，统计调整质量值的 reads 数目。
- **Reads Mapped to Genome (Map Quality >= 0)** 比对上参考基因组的 reads 比例。
- **Reads mapped to exonic regions** 比对上外显子区域的 reads 比例。

- **Reads mapped to intronic regions** 比对上内含子区域的 reads 比例。
- **Reads mapped antisense to gene** 比对上基因的 reads 中，比对上反义链的比例。
- **Reads mapped to intergenic regions** 比对上基因间区的 reads 比例。
- **Include introns** 分析中是否包含比对到内含子区域的 reads 用于表达量计算。

ANALYSIS 包括 Cluster、Marker、Cell Annotation 和 Saturation。

4.2.6 2.6 Cluster



- 左边 UMAP 图展示的是通过 lovain 算法对每个细胞进行聚类，聚为同一类的细胞具有相似的表达谱。每个点代表一个细胞，并按照不同的细胞类别予以着色。
- 右边 UMAP 图展示的是每个细胞的中 UMI 数分布。利用 UMAP 算法处理得到二维横纵坐标，每个点代表一个细胞，并按照 UMI 数不同予以着色。

4.2.7 2.7 Marker

显示了每个细胞类别中差异表达基因。每个基因在每个簇与其余样品之间进行差异表达测试。P-val 值是表达差异的统计显著性的量度，P-val 值越小，与理论相似程度越高。p_val_adj 是基于 bonferroni 校正，使用数据集中的所有基因进行调整后的 p 值。avg_log2FC 是指一个簇中某基因表达与其他细胞中平均表达比例的对数值。pct.1 在当前 cluster 细胞中检测到该基因表达的细胞比例 pct.2 是在其它 cluster 细胞中检测到该基因表达的细胞比例。

Marker

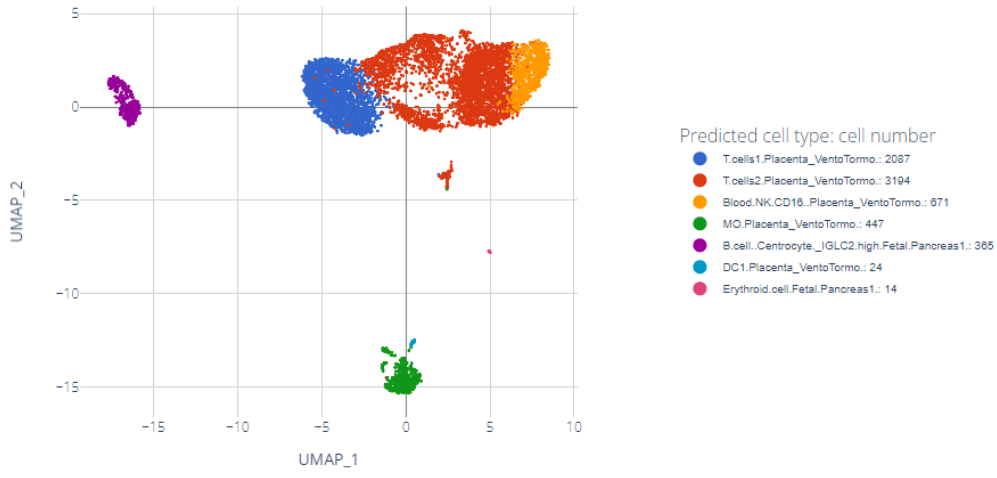
Show 10 entries Search:

gene	cluster	p_val_adj	p_val	avg_log2FC	pct.1	pct.2
AAK1.1	5	2.70801637601182e-16	8.16873209258188e-21	-1.11206601401795	0.079	0.307
AAK1.2	6	7.28720579408643e-11	2.19818581463197e-15	1.02967861772067	0.429	0.288
AAMP	8	1	0.00275144184743092	0.66656859660234	0.25	0.083
ABCA1	4	6.03389579419634e-156	1.82012482102994e-160	1.32431039667817	0.233	0.014
ABCA2	8	0.00678592267061935	2.0469737475851e-07	0.954749318466195	0.333	0.069
ABCB4	5	3.9909266702198e-104	1.20386313240017e-108	0.631660734099145	0.132	0.005
ABCC1	4	1	0.00250405124353833	0.272397665072224	0.136	0.096
ABCC1.1	6	9.80073487051254e-07	2.95639192498342e-11	0.869569290056188	0.197	0.093

Showing 1 to 010 of 12,529 entries Previous 1 2 3 4 5 ... 1253 Next

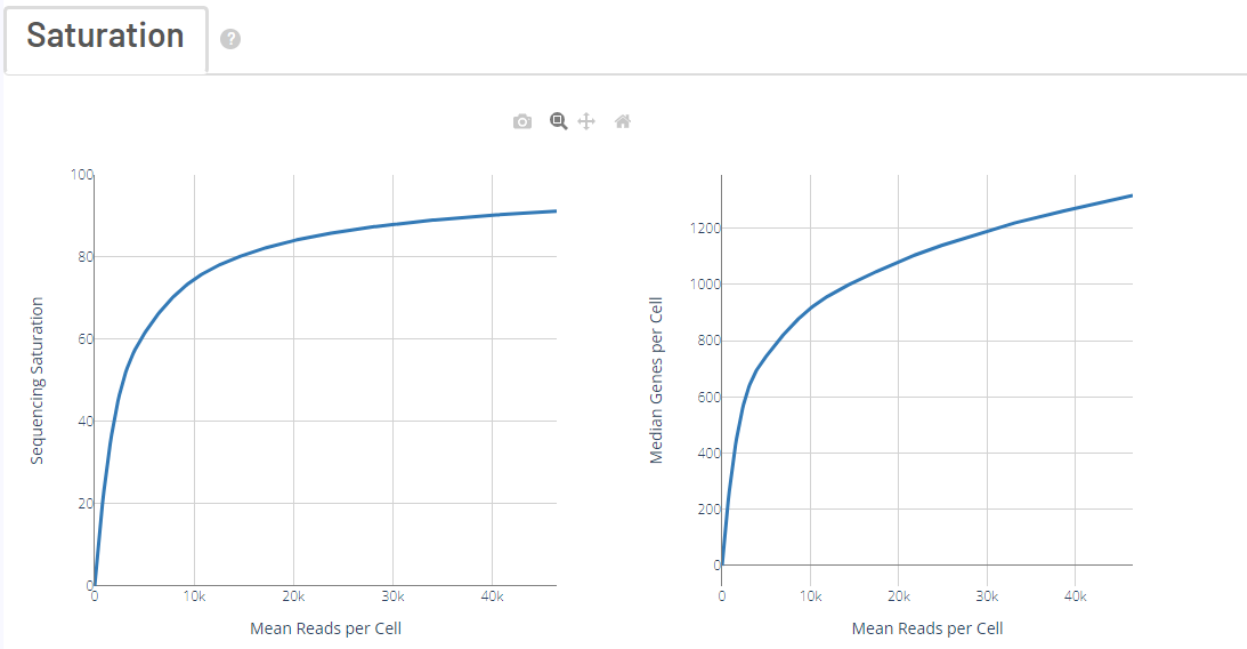
4.2.8 2.8 Cell Annotation

Cell Annotation



基于 R 包 **scHCL** (注释物种为人) 和 **scMCA** (注释物种为小鼠) 的自动注释结果。只有当物种为 **Human** 和 **Mouse** 时会得到该注释结果，其他物种时报告显示 *There is no such species reference for annnotation.*。

4.2.9 2.9 Saturation



- 左边曲线图展示了不同比例采样测序深度的测序饱和度指标。测序饱和度受测序深度和文库复杂性的影响，当所有 mRNA 转录本都已测序时，它接近 1.0(100%)。曲线末端接近平滑状态说明测序达到饱和，因为继续增加测序量，检测到的转录本也不会有特别大的变化。
- 右边曲线图展示了不同比例采样测序深度的每个细胞的基因中位数。曲线末端接近平滑状态说明测序达到饱和，因为继续增加测序量，每个细胞检测到的基因数也不会有特别大的变化。

CHAPTER 5

常见问题

5.1 1. 不同测序策略模式和试剂如何设置参数获取文库结构信息，软件如何适配分析？

我们设置了多个参数来读取文库的结构，包括--chemistry、--darkreaction 和--customize。

在使用默认参数时，软件会自动识别是否进行了暗反应以及试剂的版本。当然我们也可以使用--chemistry、--darkreaction 来定义它。目前--chemistry 包括 scRNAv1HT 和 scRNAv2HT，--darkreaction 可以分别对 cDNA 的 R1 和 oligo 的 R1R2 进行设置。比如 cDNA 的 R1 设置暗反应，oligo 的 R1 设置暗反应，R2 不设置暗反应，那么我们可以使用--darkreaction R1,R1。如果--chemistry、--darkreaction 依然无法来读取文库结构，我们可以使用--customize 来自定义文库结构。

`scRNAv1HT` 试剂的文库结构

- cDNA：

- oligo：

分析中使用 json 文件来识别 cell barcode、umi、read 等序列信息。

json 文件格式如下：

```
{
    "cell barcode tag":"CB",
    "cell barcode":[
     {
         "location":"R1:1-10",
            "distance":"1",
            "white list":[
                "TAACAGCCAA",
                "CTAAGAGTCC",
                ...
                "GTCTTCGGCT"
            ]
     },
     {
         "location":"R1:11-20"
            "distance":"1",
            "white list":[
                "TAACAGCCAA",
                "CTAAGAGTCC",
                ...
                "GTCTTCGGCT"
            ]
     },
    ],
    "UMI tag":"UR",
    "UMI":{
     "location":"R1:21-30",
    },
    "read 1":{
     "location":"R2:1-100",
    }
}
```

json 文件 key 对应的 tag 信息

key	comment
cell barcode tag	SAM tag for cell barcode, after corrected. ”CB” is suggested.
cell barcode	JSON array for cell barcode segments
cell barcode raw tag	SAM tag for raw cell barcode; ”CR” is suggested.
cell barcode raw qual tag	SAM tag for cell barcode sequence quality; ”CY” is suggested.
distance	minimal Hamming distance
white list	white list for cell barcodes
location	location of sequence in read 1 or 2
sample barcode tag	SAM tag for sample barcode
sample barcode	SAM tag for sample barcode sequence quality
UMI tag	SAM tag for UMI; ”UR” is suggested for raw UMI; ”UB” is suggested for corrected UMI
UMI qual tag	SAM tag for UMI sequence quality
UMI	location value for the UMI
read 1	read 1 location
read 2	read 2 location

cDNA 的 R1 和 oligo 的 R1R2 都进行了暗反应时位置信息

```
cDNA
cell barcode:R1:1-10、R1:11-20
umi:R1:21-30
read 1:R2:1-100
oligo
cell barcode:R1:1-10、R1:11-20
read 1:R2:1-30
```

cDNA 的 R1 和 oligo 的 R1 都进行了暗反应,oligo 的 R2 没有进行暗反应时位置信息

```
cDNA
cell barcode:R1:1-10、R1:11-20
umi:R1:21-30
read 1:R2:1-100
oligo
cell barcode:R1:1-10、R1:11-20
read 1:R2:1-10,R2:17-26,R2:33-42
```

其他测序策略可自定义 json 文件，根据位置信息填写 location。

5.2 2.cell_calling 应该选择哪个参数？

默认的 cell calling 方法是 emptydrops。

- emptydrops:

  先判定有效液滴 beads 先采用高 umi 阈值法预期捕获 N 个 beads，则按照每个 Barcode 对应的 UMI 数进行排序，在 UMI 数最高的 N 个 cell barcode 中，取第 99 分位 Barcode 对应的 UMI 数目除以 10，作为 cut-off。所有 cell barcode 中对应的 UMI 数目高于该 cut-off 即为细胞，否则为背景），然后使用 emptydrops 对低 umi 的 beads 与背景 beads 进行区分（确定背景空液滴集合，使用 Dirichlet-multinomial 模型将其与每个 beads 对应的 UMI count 进行差异显著性检验，差异显著即为有效液滴内 beads 否则为背景 beads）。

- barcoderanks：

  将 cell barcode 按照 UMI 数目从高到低排列，并拟合曲线，曲线斜率变化大的点对应的 UMI 数目即为 cut-off 所有 cell barcode 对应的 UMI 数目高于该 cut-off 为有效液滴内 beads，否则为背景 beads。

如果对获取的细胞结果不满意，可更换 cell calling 方法重新进行计算或者使用 forcecells 确定使用 umi 数量排序前 N 个 beads 用于分析。

5.3 3. 对某些参数不满意，重新分析？

DNBelab C4 分析流程支持跳过已完成的步骤。例如 02.count 步骤中合并多 beads 分析报错，则不需要重新分析 01.data 步骤步骤。DNBC4tools 只需要在原分析中添加参数 `--process count,analysis,report` 可以跳过 data 的分析步骤的分析步骤。分析结果不满意需要重新分析时，需要判定需要调整的分析参数位于哪个阶段，然后选择分析接下来的步骤。

在 DNBC4tools data,count,analysis,report 中有些参数在主程序 run 中没有。通常情况下这些参数使用默认值分析即可。如果确实需要修改这些参数，可使用 data,count,analysis,report 模块进行分析，再使用 run -process 参数将后续的结果分析。例如，使用 run 得到分析结果和报告后，对细胞分群的结果不满意，可使用 DNBC4tools analysis –resolution 调整分群的分辨率，分析完成后在使用 DNBC4tools run –process report 完成后续的 report 分析。

CHAPTER 6

LICENSE