DNBelab_C_Series_HT_scRNA-analysis-software 2.0流程、安装与使用

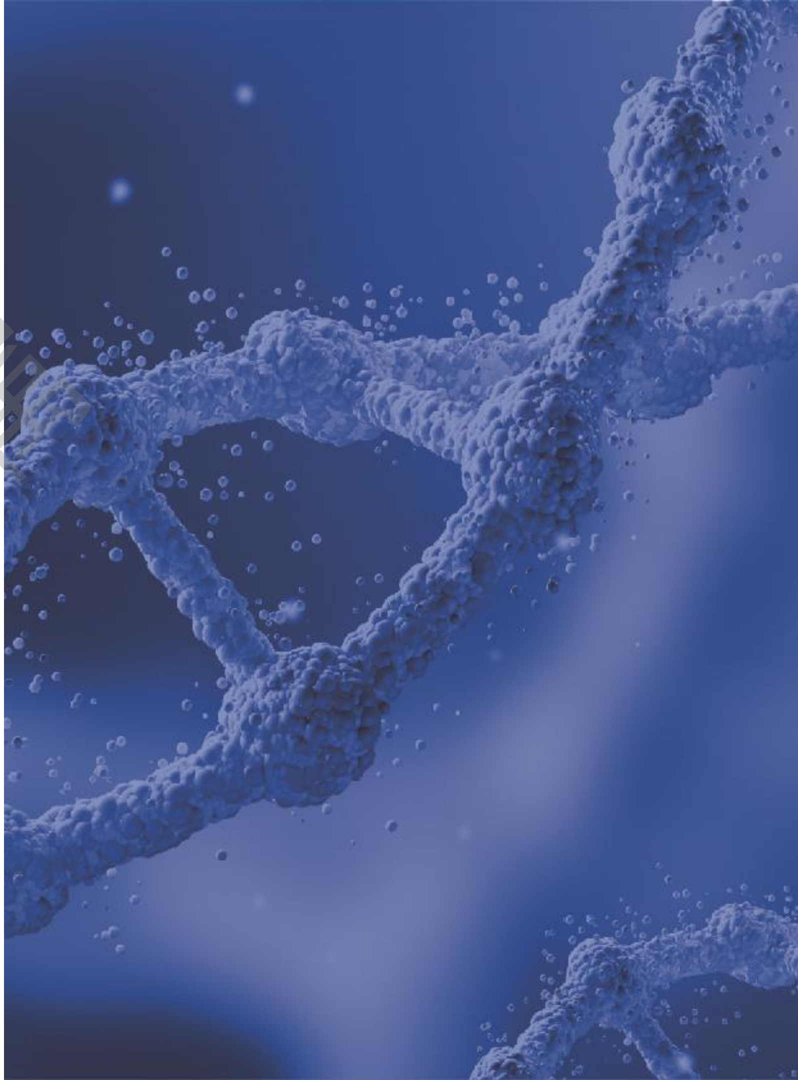
MGI
华大智造

CONTENTS 目录

01 下机数据结构&生信流程

02 软件安装与使用

03 报告解读

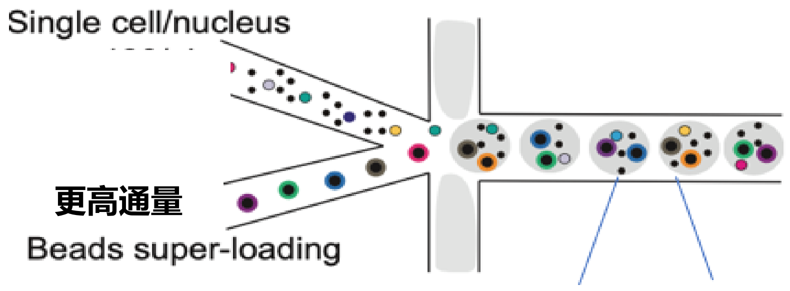


1

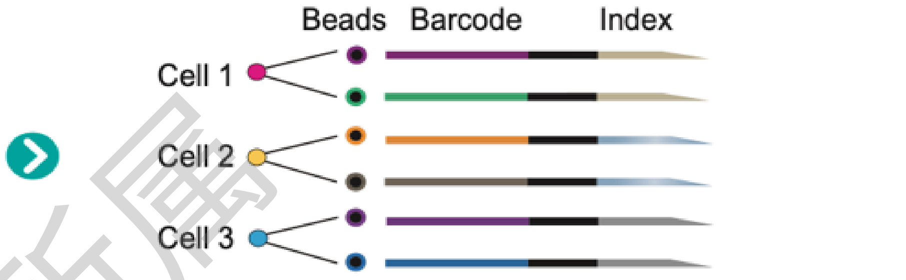
下机数据结构&生信流程

DNBelab_C4_scRNA_V2技术原理

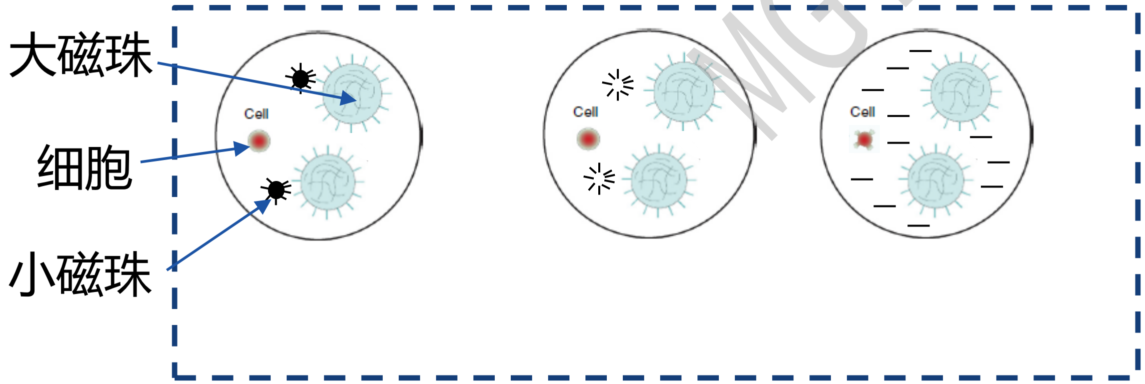
Droplet-based indexing

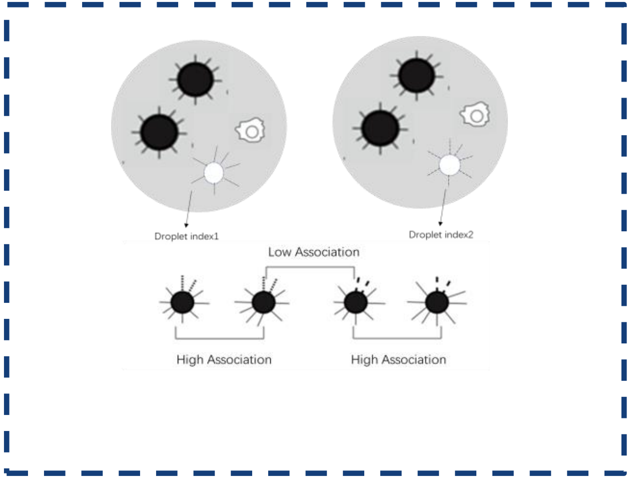


Deconvolution based on barcode and M280 Index

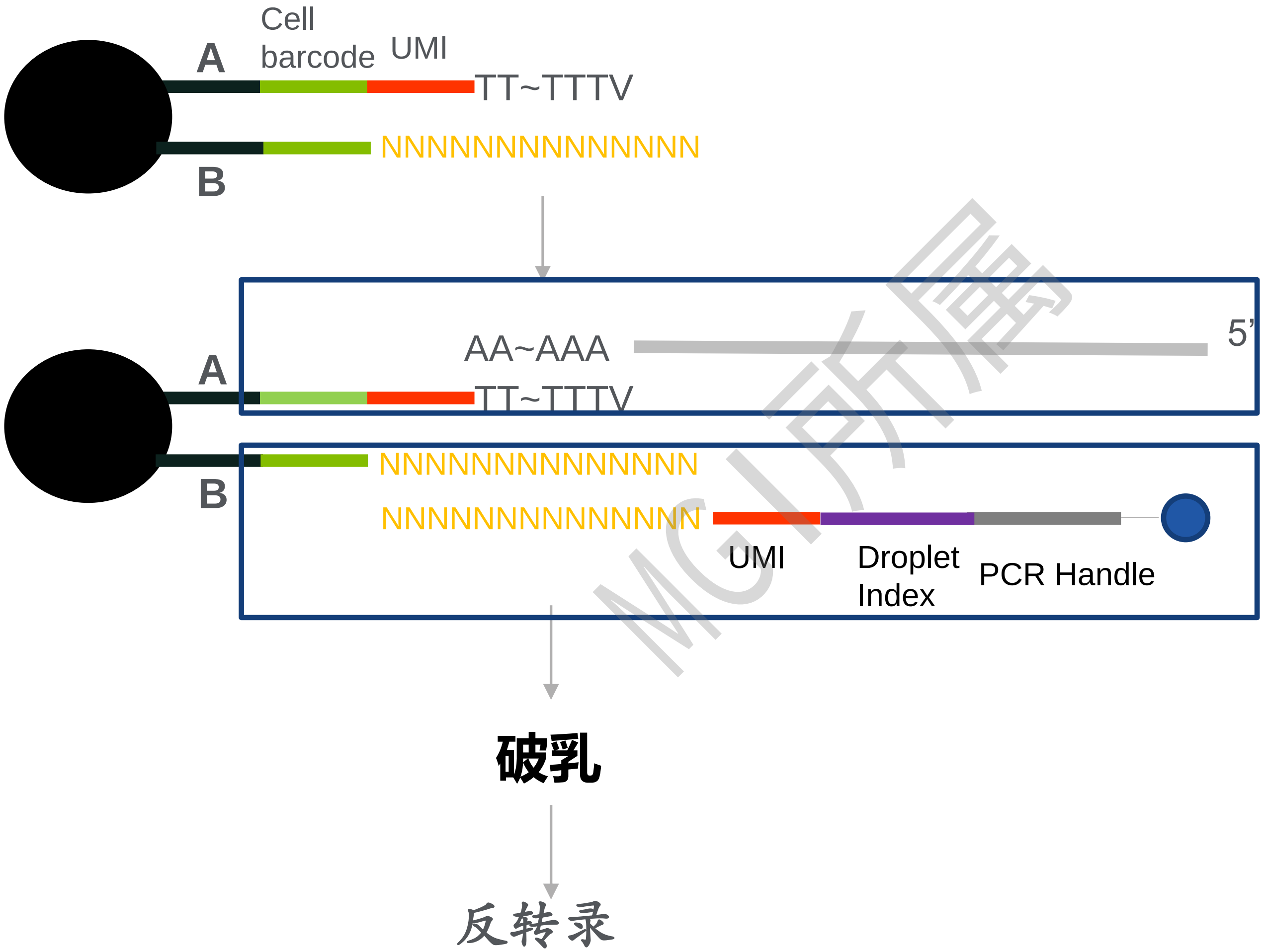


有效液滴：n个大Beads+n个小Beads+1个细胞（n>0）





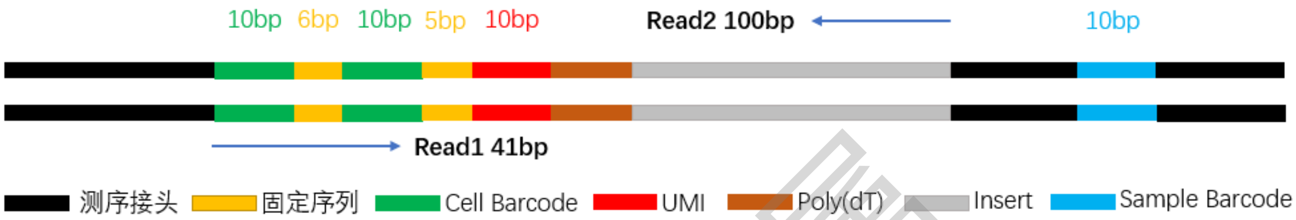
DNBelab_C4_scRNA_V2技术原理



A: mRNA （cDNA）

B: Index Carrier （Oligo）

DNBelab_C4_scRNA_V2文库结构&下机数据结构



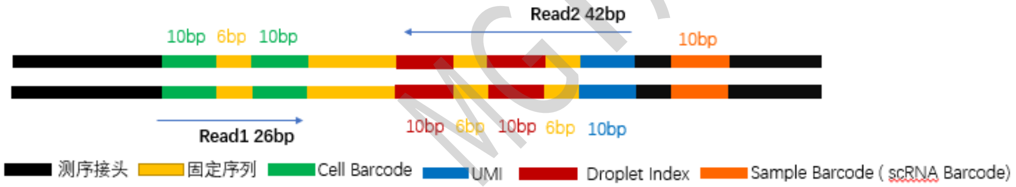
cDNA：

fastq1 1-10bp barcode1；11-20bp barcode 2；21-30 UMI.

fastq2 1-100bp insertion

r1设定：10bp+6bp暗反应+10bp+5bp暗反应+10bp

r2设定：100bp



Oligo（分开测序）：

fastq1 1-10bp barcode1; 11-20bp barcode 2;

fastq2 1-10bp Oligo UMI; 11-20bp Oligo barcode1; 21-30bp Oligo barcode2.

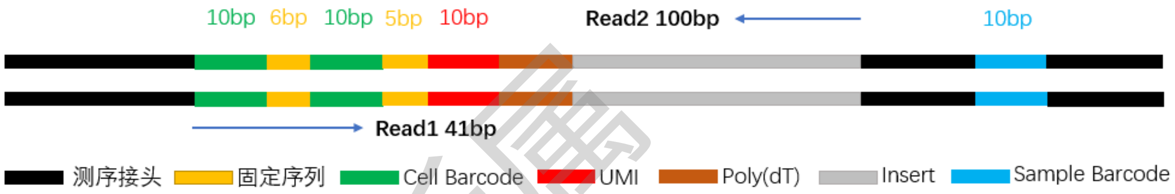
cDNA+oligo混测新流程下机数据结构

双文库

A: cDNA 文库

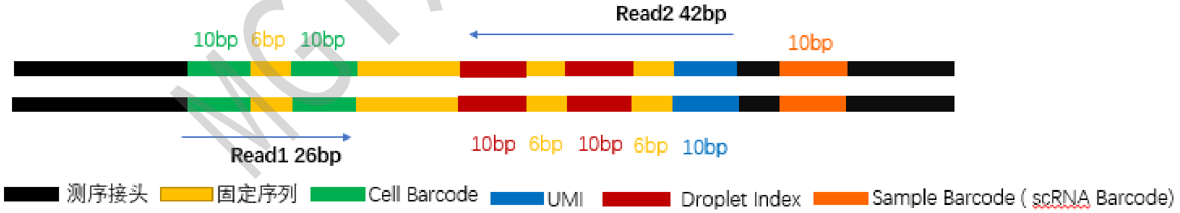
300~500bp

cDNA文库不变



B: Droplet Index文库

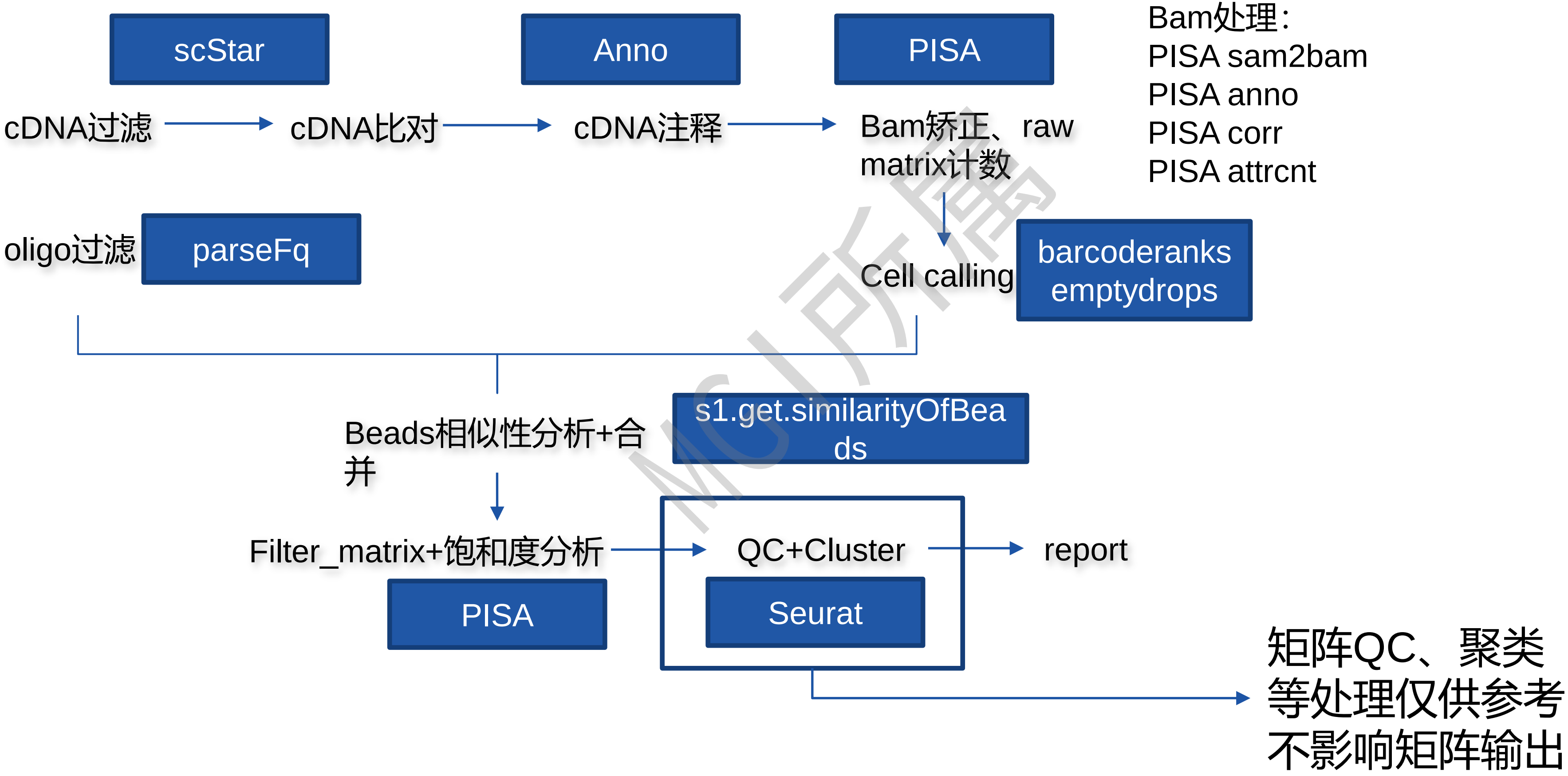
~170bp



Oligo（混测）：
fastq1：20bp barcode + 10bp固定序列
fastq2：10bp umi + 6bp固定 + 10bp index + 6bp固定 + 10bp index + 58bp固定序列
*混测的oligo fastq2需要处理

DNBelab_C4_scRNA_V2生信分析流程概览

过滤和比对使用到的软件：
scRNA_parse
Star
Bam处理：
PISA sam2bam
PISA anno
PISA corr
PISA attrcnt



scStar

Anno

PISA

cDNA过滤

cDNA比对

cDNA注释

Bam矫正、raw matrix计数

oligo过滤

parseFq

Cell calling

barcoderanks
emptydrops

Beads相似性分析+合并

s1.get.similarityOfBeads

Filter_matrix+饱和度分析

PISA

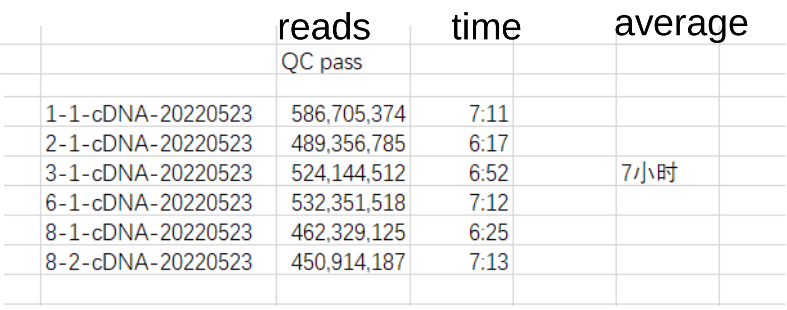
QC+Cluster

Seurat

report

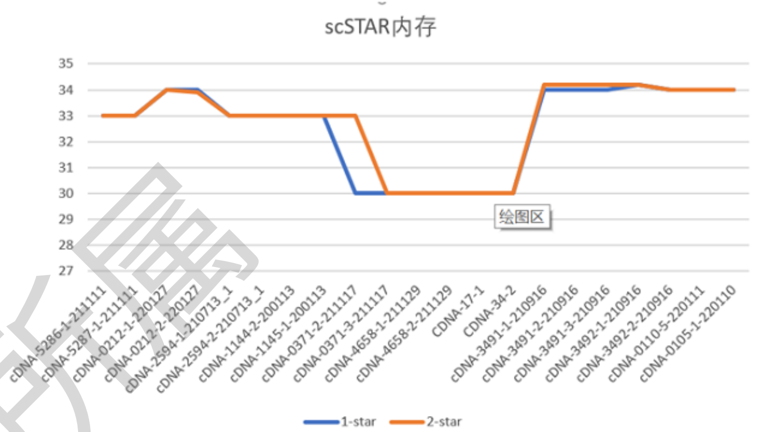
矩阵QC、聚类
等处理仅供参考
不影响矩阵输出

cDNA比对分析时间和内存（限速步骤）

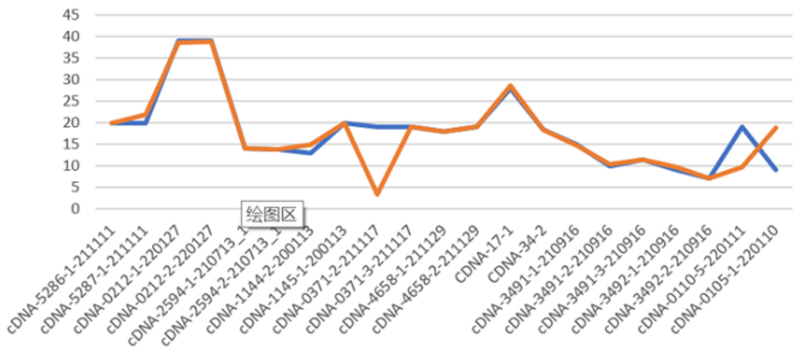


	reads QC pass	time	average
1-1-cDNA-20220523	586,705,374	7:11	
2-1-cDNA-20220523	489,356,785	6:17	
3-1-cDNA-20220523	524,144,512	6:52	7小时
6-1-cDNA-20220523	532,351,518	7:12	
8-1-cDNA-20220523	462,329,125	6:25	
8-2-cDNA-20220523	450,914,187	7:13	

scSTAR内存



Anno内存

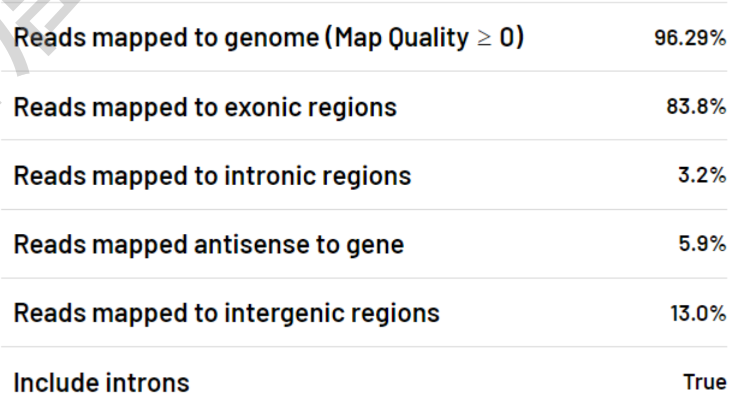


测试范围：200M – 1200M reads

cDNA注释逻辑

通过gtf文件对cDNA进行注释，并划分reads归属（exon/intron/intergenic/antisense区域）

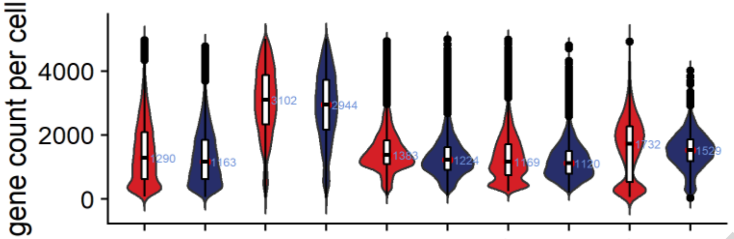
	单细胞V1	10X	单细胞V2
Exon判断标准	100%与exon区域相交	大于50%与exon区域相交	大于50%与exon区域相交
Intron判断标准	reads与intron区域相交	reads与intron区域相交	reads与intron区域相交
Antisense统计标准	比对上的基因任意一个链反向	注释为exon或intron，且链反向	注释为exon或intron，且链反向
Intergenic统计标准		1-exon%-intron%	1-exon%-intron%



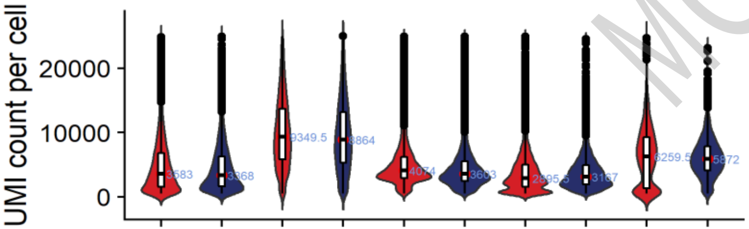
Reads mapped to genome (Map Quality ≥ 0)	96.29%
Reads mapped to exonic regions	83.8%
Reads mapped to intronic regions	3.2%
Reads mapped antisense to gene	5.9%
Reads mapped to intergenic regions	13.0%
Include introns	True

Exon+Intron 用于后续分析

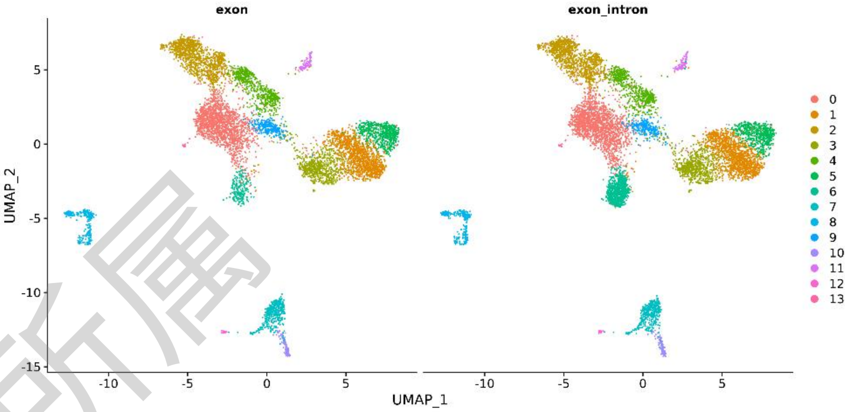
Including intronic reads, for both cellular and nuclei samples, could lead to higher sensitivity (higher UMI counts, more genes per cell) and less wasted sequencing.

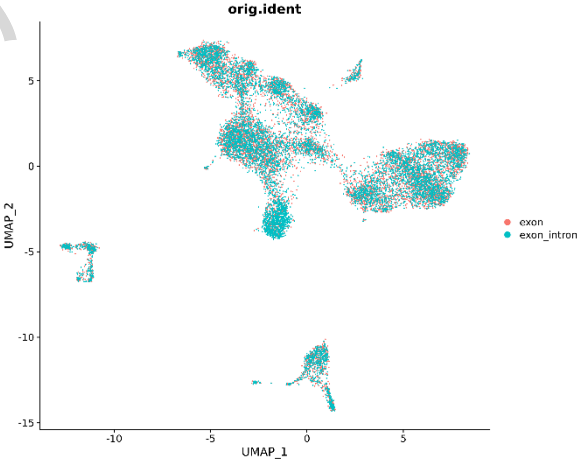


Median	1290	1163	3102	2944	1383	1224	1169	1120	1732	1529
	样本1		样本2		样本3		样本4		样本5	

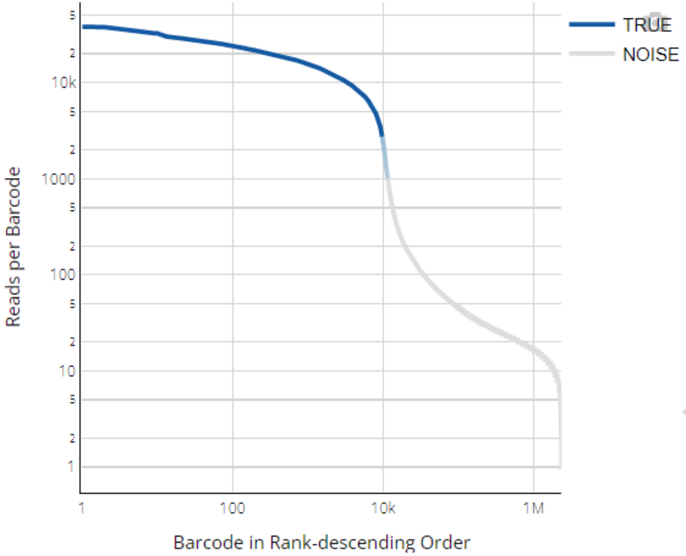


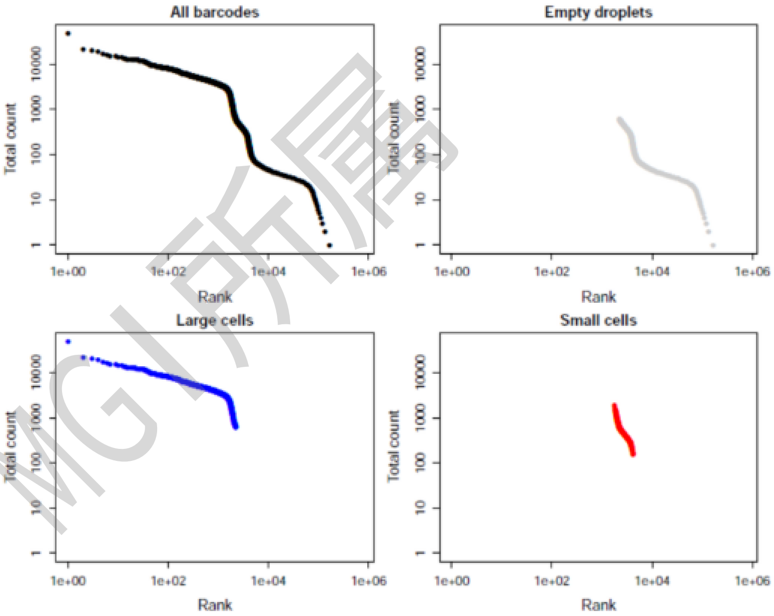
Median	3583	3368	9350	8864	4074	3603	2896	3167	6260	5872
	样本1		样本2		样本3		样本4		样本5	





Cell calling





Cell calling算法

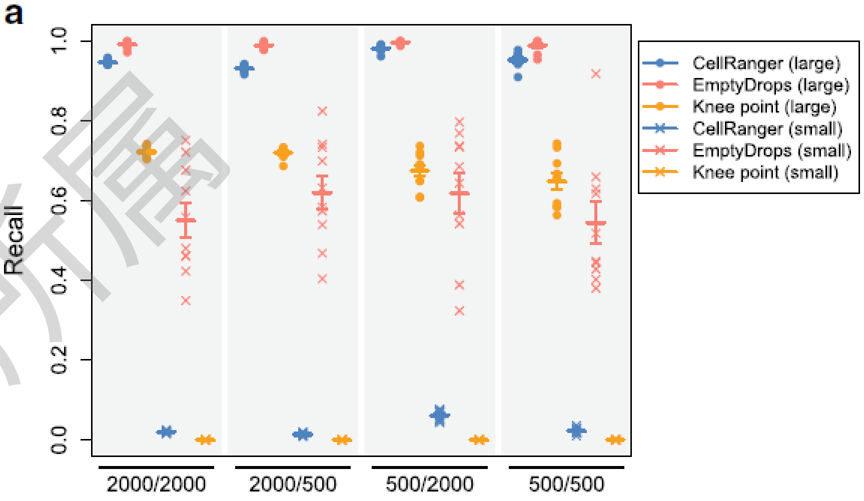
高UMI阈值法：这种方法即是通过UMI数值高低进行判断的方法。如果预期捕获N个细胞，则按照每个Barcode对应的UMI数进行排序，在UMI数最高的N个Barcode中，取第99分位Barcode对应的UMI数目除以10，作为cut-off。所有Barcode中对应的UMI数目高于该cut-off即为细胞，否则为背景。

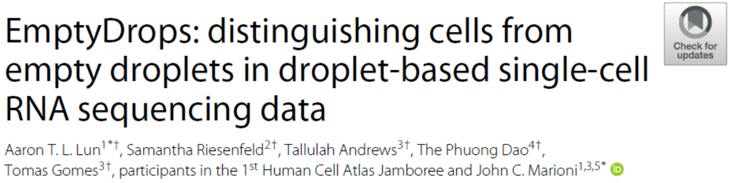
EmptyDrops：这种方法解决了低UMI细胞与背景数据的区分，首先，对ambient RNA的集合进行估计，然后使用Dirichlet-multinomia模型，将其与每个Barcode对应的UMI count进行差异显著性检验，差异显著即为细胞，否则为背景。

Barcoderanks方法：这种方法使用UMI数值变化的“拐点”作为细胞判断cut-off的方法。将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数目为拐点，即cut-off。所有Barcode对应的UMI数目高于该cut-off为细胞，否则为背景。

Barcoderanks法对于高质量数据截取效果较好，下游分析时数据质量高

Cellranger方法：高阈值+emptydrops（默认）



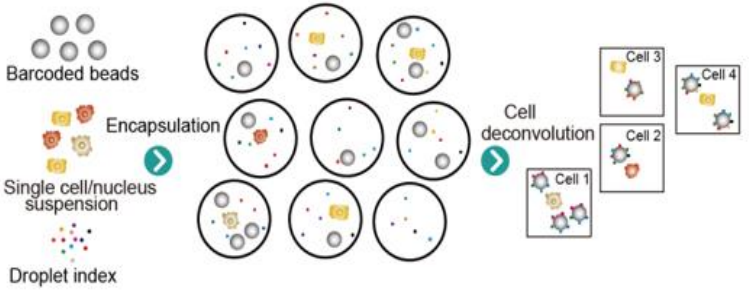


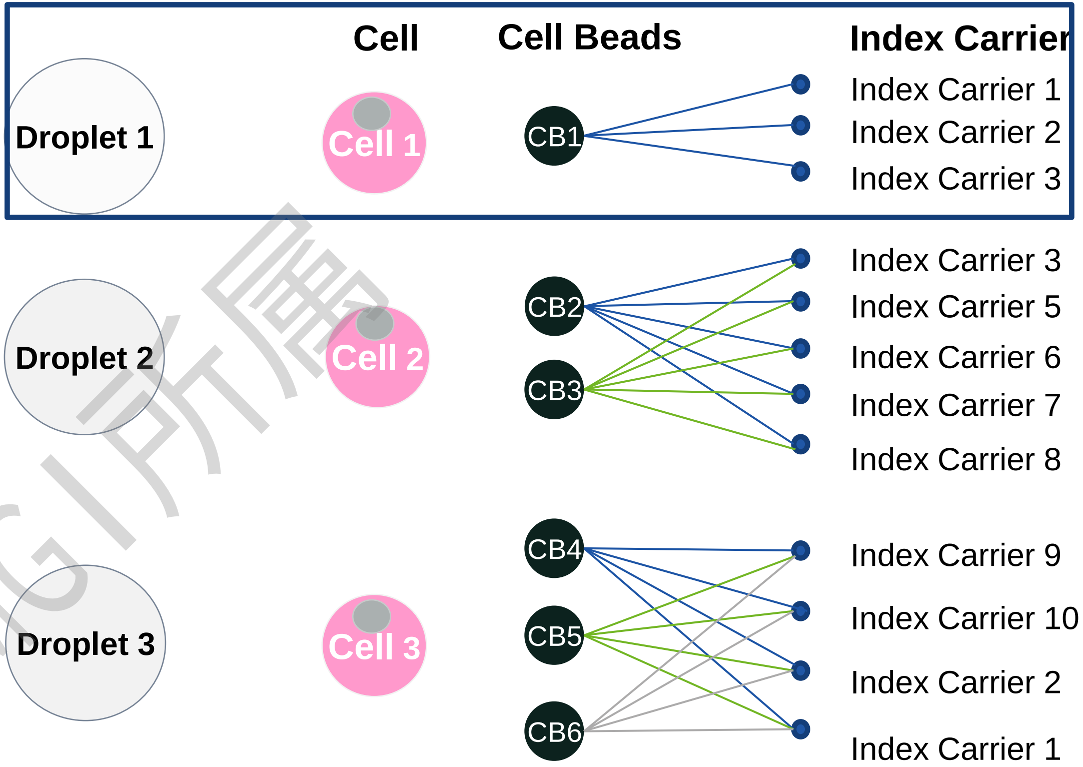
EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data

Check for updates

Aaron T. L. Lun[1*†], Samantha Riesenfeld[2†], Tallulah Andrews[3†], The Phuong Dao[4†], Tomas Gomes[3†], participants in the 1st Human Cell Atlas Jamboree and John C. Marioni[1,3,5*]

Beads相似性计算与合并



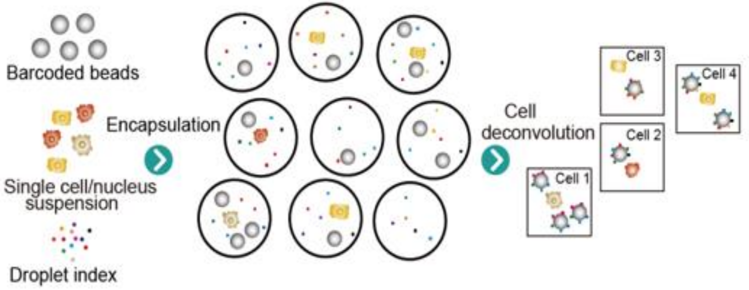


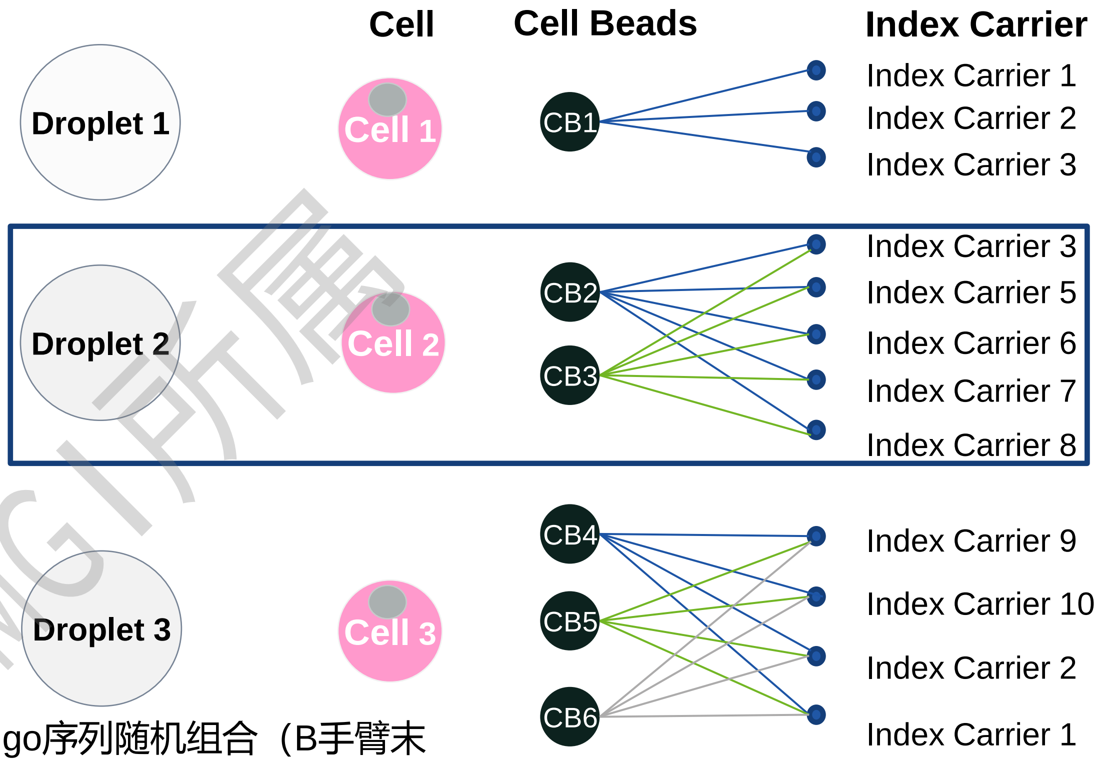
磁珠种类丰富度为10^5-10^6

单个液滴中捕获的大小磁珠数量遵循泊松分布，油滴内含有n个大磁珠（一般情况下1<n<7，70%以上为1），m个小磁珠（m>=3),1个细胞（双胞、多胞的情况可以通过异常的表达量离群值去除）

在该阶段之前，每个cell barcode对应1个bead，计数以beads为单位

Beads相似性计算与合并



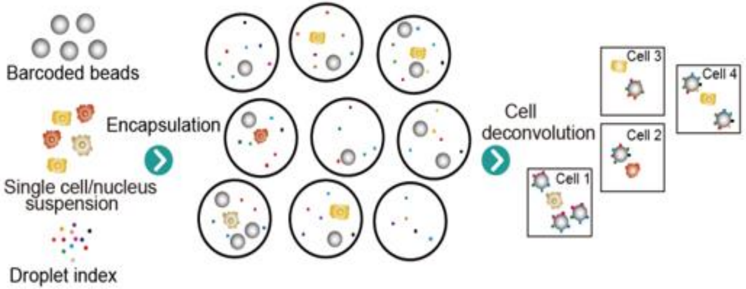


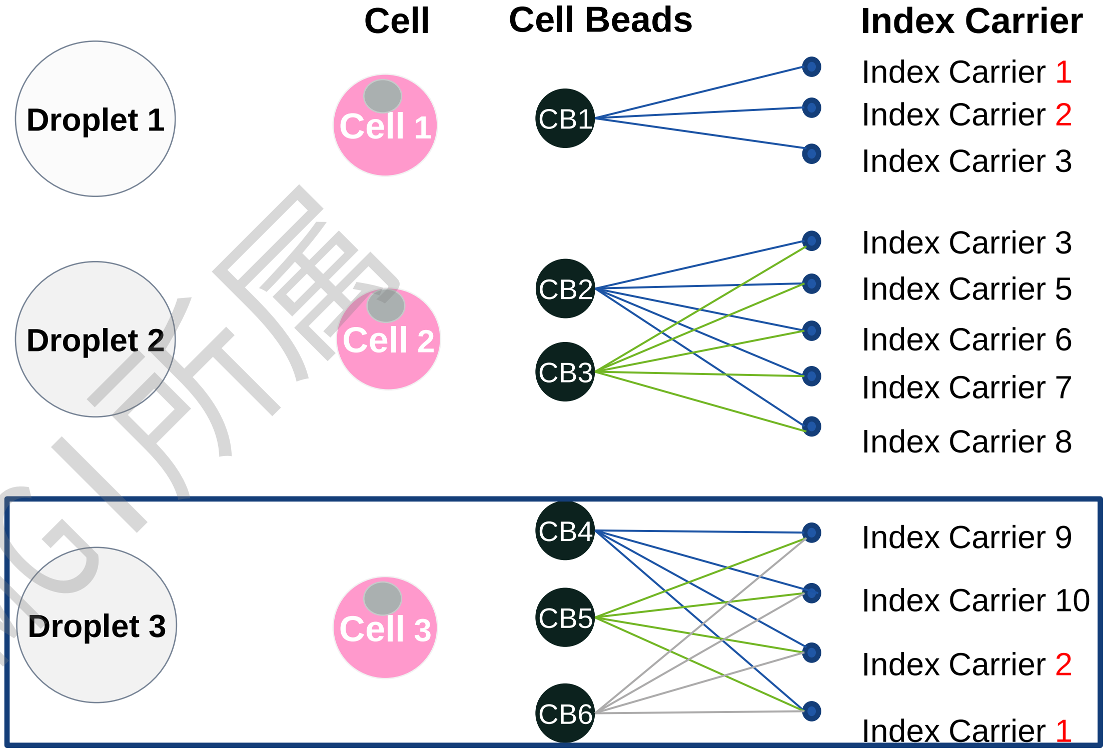
磁珠种类丰富度为10^5-10^6

同一个液滴中，大磁珠的B手臂与小磁珠释放的oligo序列随机组合（B手臂末端与小磁珠固定序列部分互补配对），组合信息则通过B手臂上的cell barcode-oligo barcode对应关系，保存在oligo测序文件中

如：CB2和CB3都会随机捕获到5-8号小磁珠，而CB1、CB4/5/6则没有5-8号小磁珠。因此显然CB2与CB3来自同一液滴，即同一细胞。

Beads相似性计算与合并

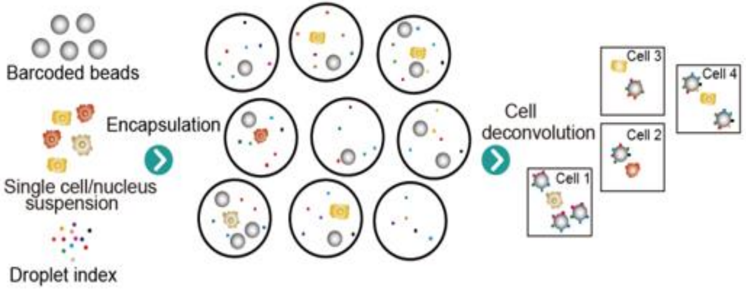


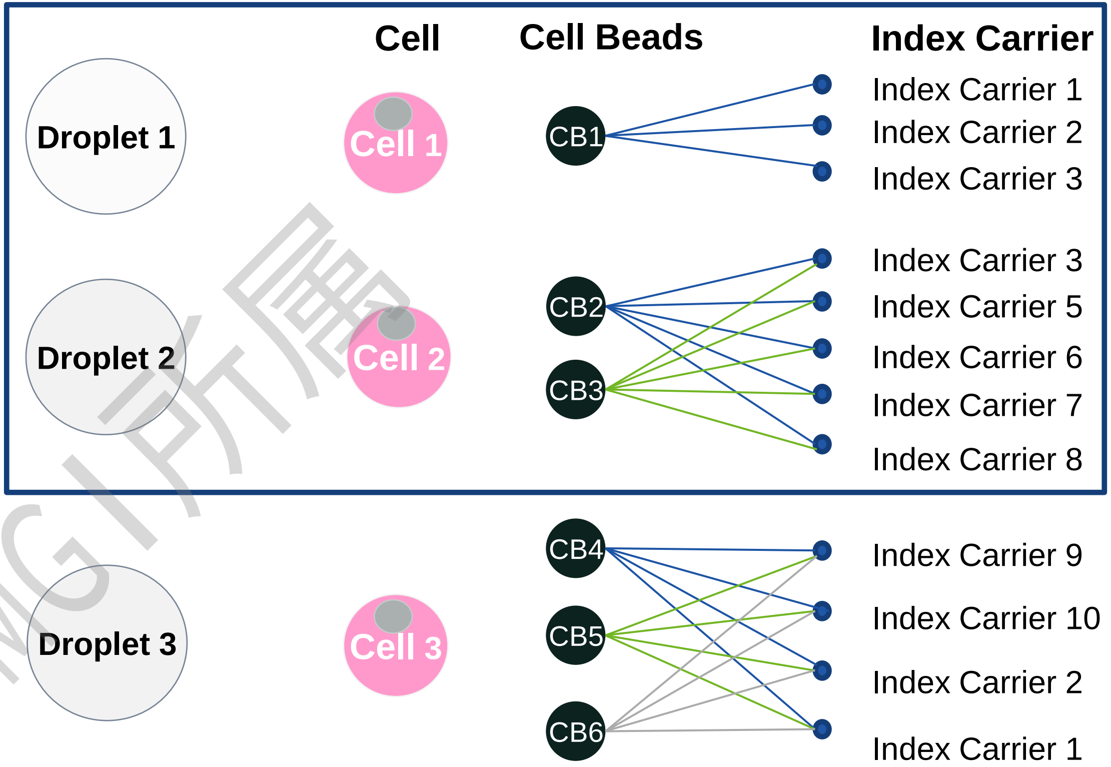


磁珠种类丰富度为10^5-10^6

大磁珠不会重复，小磁珠则会过量
不同液滴内捕获到的小磁珠barcode在概率学上会有随机重复
因此，当Cell barcode对应的Oligo barcode有overlap时，除了它们来自于同一个油滴的情况以外，也有可能是不同油滴内随机出现重复。
如：CB1和CB4/5/6对应的oligo barcode都有1/2

Beads相似性计算与合并



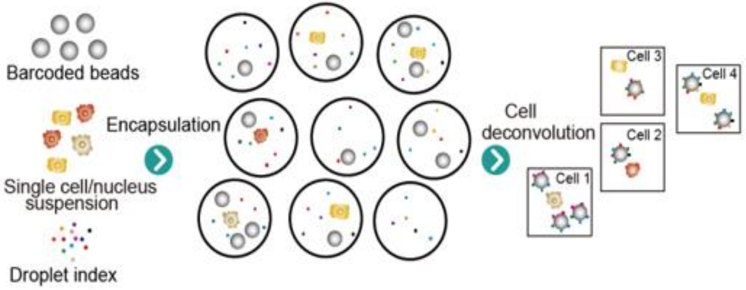


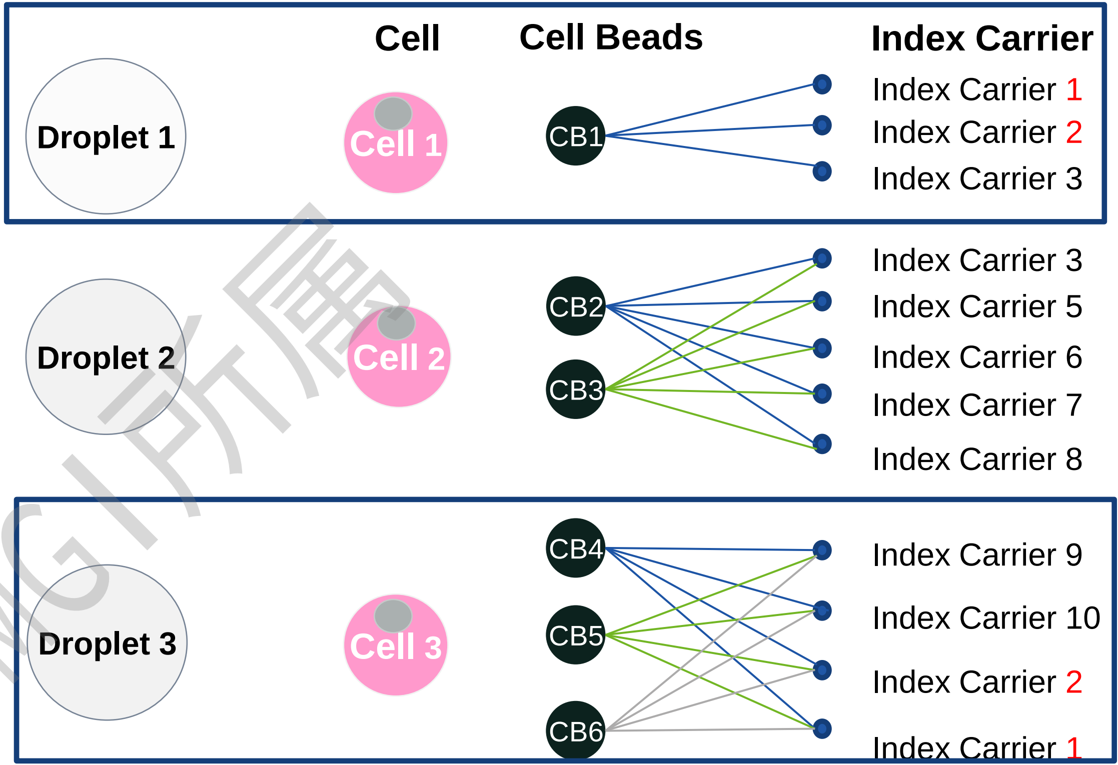
磁珠种类丰富度为10^5-10^6

但由于小磁珠种类的较高丰富度，不同液滴内的大磁珠间的overlap（小磁珠barcode）在大多数情况下为0/1，这些情况可以判定为其不在同一个油滴内

如：CB1与CB2/CB3之间只有1个overlap

Beads相似性计算与合并





如：CB4/5/6两两之间余弦相似度值>0.2
CB1与CB4/5/6之间的余弦相似度值<0.2
因此CB4/5/6会进行合并

磁珠种类丰富度为10^5-10^6

对于有大于2个overlap以上的小磁珠的beads单位，以每个beads为单位（即图中的CB）构建轶和加权余弦相似度模型，通过计算判定哪些beads单位需要合并（认为余弦相似度大于0.2的beads单位来自于同一个油滴，这些beads捕获到的reads则应来自同一个细胞，因此合并整合为1个cell单位）

PISA count

one-hit提取：基因之间可能存在overlap，比对到overlap的reads无法判断其归属，因此不会被计算进基因表达矩阵matrix中
*在构建数据库时，有效地过滤gtf能排除较多overlap。

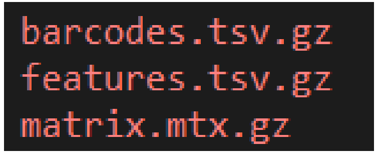
提供3类矩阵输出以供各种分析需求

#### all have GN tag
PISA count -one-hit -cb DB -anno-tag GN -umi UB

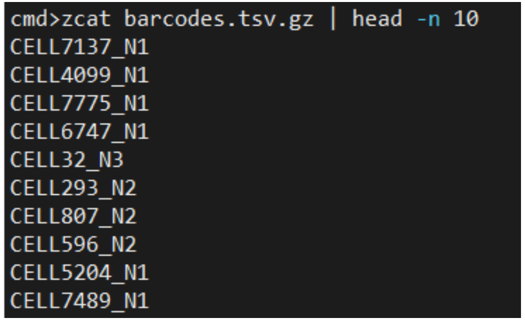
#### Exon
PISA count -one-hit -cb DB -ttype E -anno-tag GN -umi UB

##### RNA velocity
PISA count -one-hit -cb DB -velo -anno-tag GN -umi UB

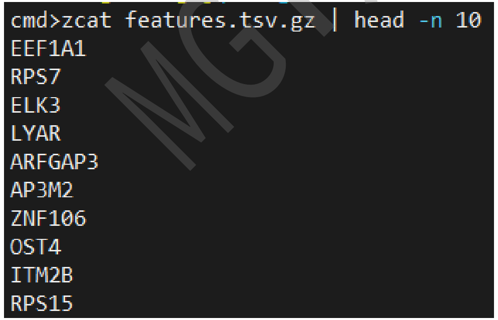
Only introns Included



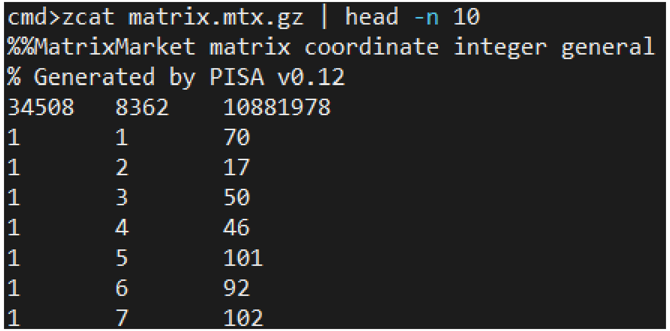
```
barcodes.tsv.gz
features.tsv.gz
matrix.mtx.gz
```



```
cmd>zcat barcodes.tsv.gz | head -n 10
CELL7137_N1
CELL4099_N1
CELL7775_N1
CELL6747_N1
CELL32_N3
CELL293_N2
CELL807_N2
CELL596_N2
CELL5204_N1
CELL7489_N1
```



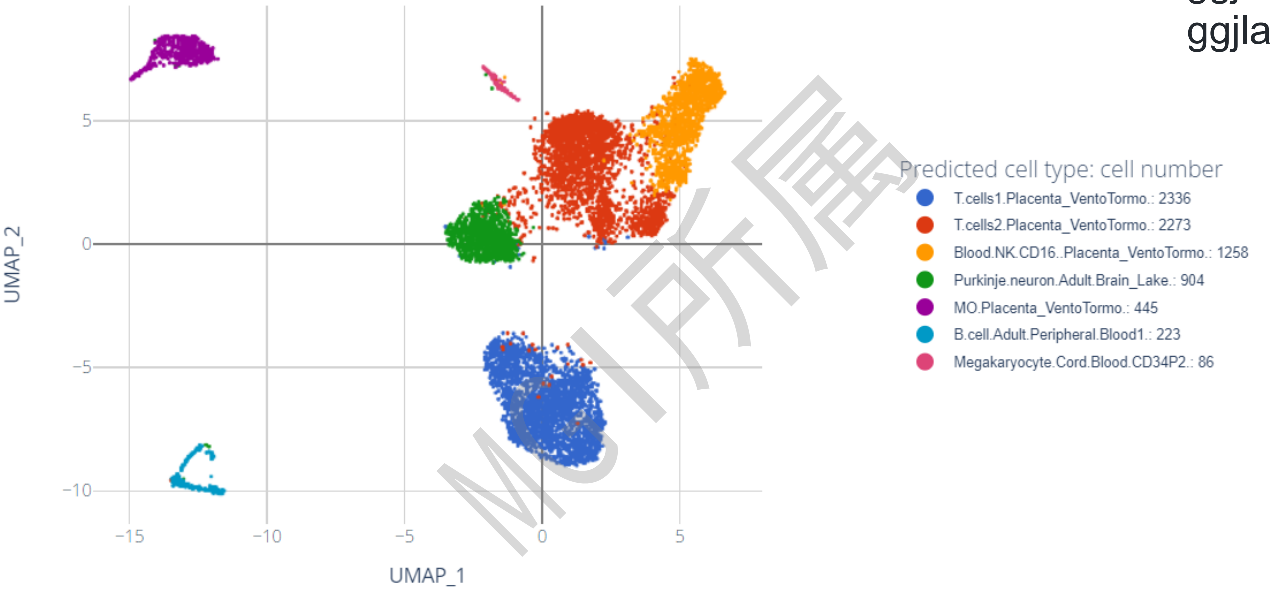
```
cmd>zcat features.tsv.gz | head -n 10
EEF1A1
RPS7
ELK3
LYAR
ARFGAP3
AP3M2
ZNF106
OST4
ITM2B
RPS15
```



```
cmd>zcat matrix.mtx.gz | head -n 10
%%MatrixMarket matrix coordinate integer general
% Generated by PISA v0.12
34508   8362    10881978
1       1       70
1       2       17
1       3       50
1       4       46
1       5       101
1       6       92
1       7       102
```

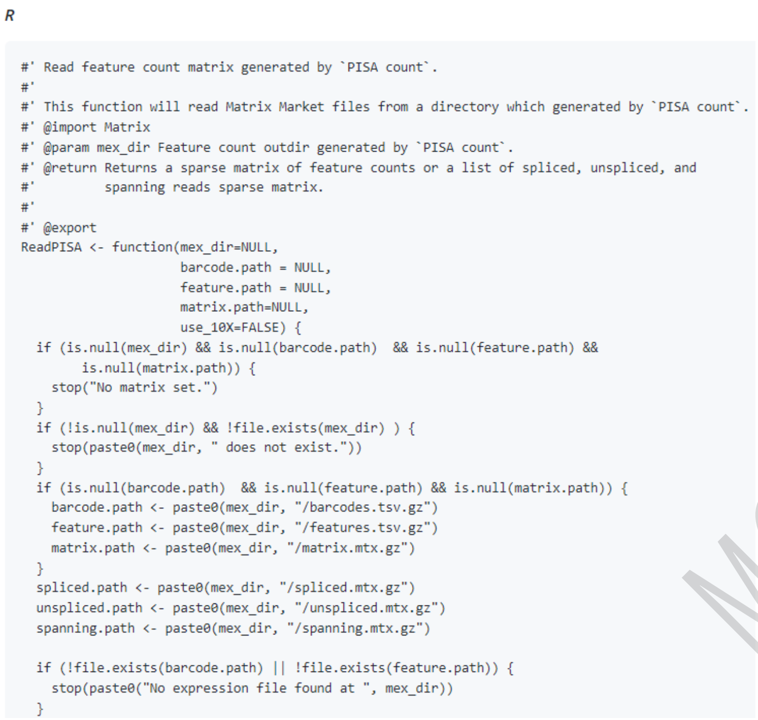
细胞注释

ggjlab/scHCL
ggjlab/scMCA



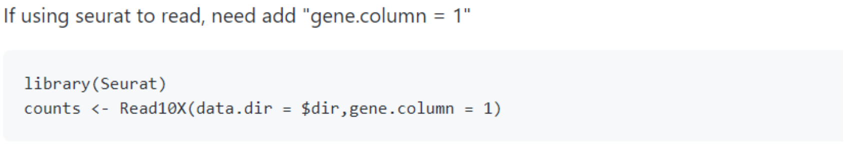
目前软件包提供了基于ggjlab R包的人类和小鼠的自动细胞注释，可以通过Species参数输入以下词条进行判定：Mouse、mouse、Human、human、hg19、hg38、mm10、Human nucleus等

下游分析软件对接



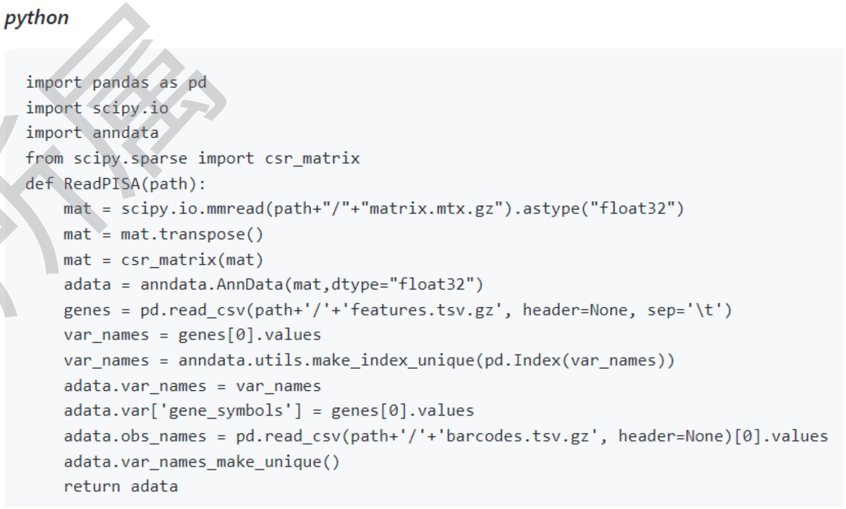
R

```r
#' Read feature count matrix generated by `PISA count`.
#'
#' This function will read Matrix Market files from a directory which generated by `PISA count`.
#' @import Matrix
#' @param mex_dir Feature count outdir generated by `PISA count`.
#' @return Returns a sparse matrix of feature counts or a list of spliced, unspliced, and
#'         spanning reads sparse matrix.
#'
#' @export
ReadPISA <- function(mex_dir=NULL,
                     barcode.path = NULL,
                     feature.path = NULL,
                     matrix.path=NULL,
                     use_10X=FALSE) {
  if (is.null(mex_dir) && is.null(barcode.path)  && is.null(feature.path) &&
      is.null(matrix.path)) {
    stop("No matrix set.")
  }
  if (!is.null(mex_dir) && !file.exists(mex_dir) ) {
    stop(paste0(mex_dir, " does not exist."))
  }
  if (is.null(barcode.path)  && is.null(feature.path) && is.null(matrix.path)) {
    barcode.path <- paste0(mex_dir, "/barcodes.tsv.gz")
    feature.path <- paste0(mex_dir, "/features.tsv.gz")
    matrix.path <- paste0(mex_dir, "/matrix.mtx.gz")
  }
  spliced.path <- paste0(mex_dir, "/spliced.mtx.gz")
  unspliced.path <- paste0(mex_dir, "/unspliced.mtx.gz")
  spanning.path <- paste0(mex_dir, "/spanning.mtx.gz")

  if (!file.exists(barcode.path) || !file.exists(feature.path)) {
    stop(paste0("No expression file found at ", mex_dir))
  }
```



If using seurat to read, need add "gene.column = 1"

```r
library(Seurat)
counts <- Read10X(data.dir = $dir,gene.column = 1)
```



python

```python
import pandas as pd
import scipy.io
import anndata
from scipy.sparse import csr_matrix
def ReadPISA(path):
    mat = scipy.io.mmread(path+"/"+"matrix.mtx.gz").astype("float32")
    mat = mat.transpose()
    mat = csr_matrix(mat)
    adata = anndata.AnnData(mat,dtype="float32")
    genes = pd.read_csv(path+'/'+'features.tsv.gz', header=None, sep='\t')
    var_names = genes[0].values
    var_names = anndata.utils.make_index_unique(pd.Index(var_names))
    adata.var_names = var_names
    adata.var['gene_symbols'] = genes[0].values
    adata.obs_names = pd.read_csv(path+'/'+'barcodes.tsv.gz', header=None)[0].values
    adata.var_names_make_unique()
    return adata
```

https://github.com/MGI-tech-bioinformatics/DNBelab_C_Series_HT_scRNA-analysis-software/blob/version2.0/doc/Downstream_Analysis.md

提供seurat、scanpy等常用的下游分析软件的对接帮助文档

2

软件安装与使用

DNBelab_C4_scRNA_V2流程安装

GitHub官网：

https://github.com/MGI-tech-bioinformatics/DNBelab_C_Series_HT_scRNA-analysis-software

硬件要求：

- 节点需要x86-64 架构，centos 7.x 64位操作系统(Linux内核版本3.10.x)
- 流程运行要求50G RAM ;4 CORE CPU ;4TB以上存储
- conda环境部署到本地时需要连接公网
- *提供离线docker版本

DNBelab_C4_scRNA_V2流程安装

- 安装：
- 支持环境部署（miniconda）
- 一键安装DNBC4tools软件
- 安装3个R包
- 下载cromwell-35.jar
- 构建index（与cellranger 6.0之后的index版本通用）

运行：

wdl模式：

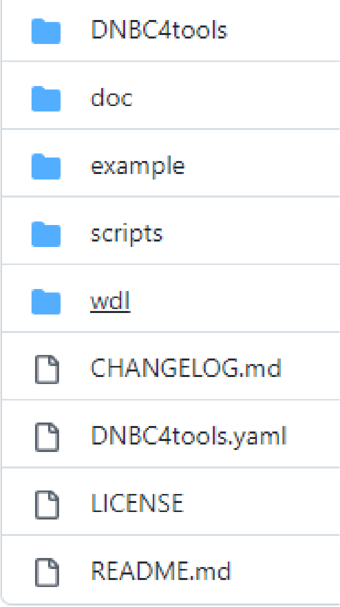
- 填写config-cDNA.json
- 运行主程序run.sh

命令行模式：

- 一键运行全流程
- 调整参数，单独运行部分模块

DNBelab_C4_scRNA_V2流程安装

流程下载：
运行 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更改权限



- DNBC4tools
- doc
- example
- scripts
- wdl
- CHANGELOG.md
- DNBC4tools.yaml
- LICENSE
- README.md

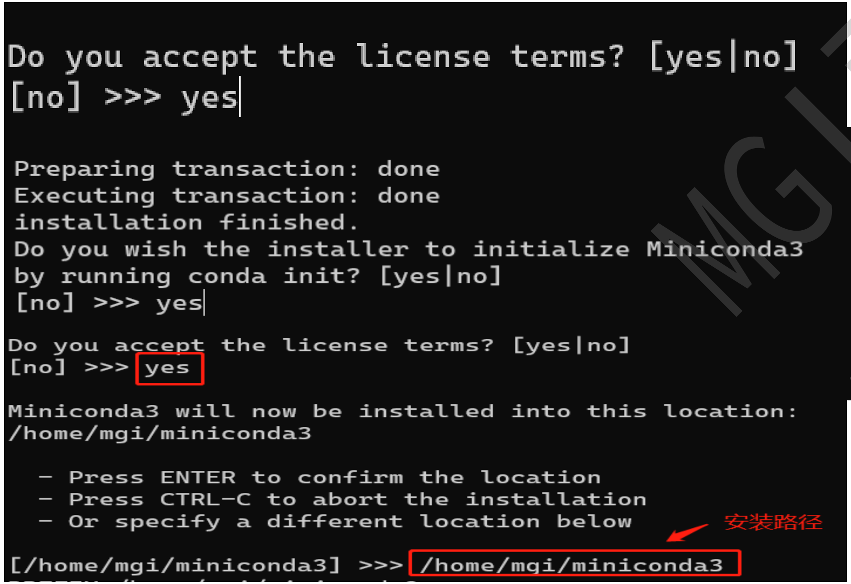
子环境配置文件(DNBC4tools.yaml)
DNBC4tools：包含运行所需的软件和脚本（其中有config文件，包括磁珠结构配置文件）
Doc：包括安装说明文档等各种帮助文档
Example：包含主程序和运行配置文件
Script：包含运行的脚本
Wdl：基于wdl流程化运行方式的流程文件（wdl流程可根据需要修改）

DNBelab_C4_scRNA_V2流程安装

Conda安装：

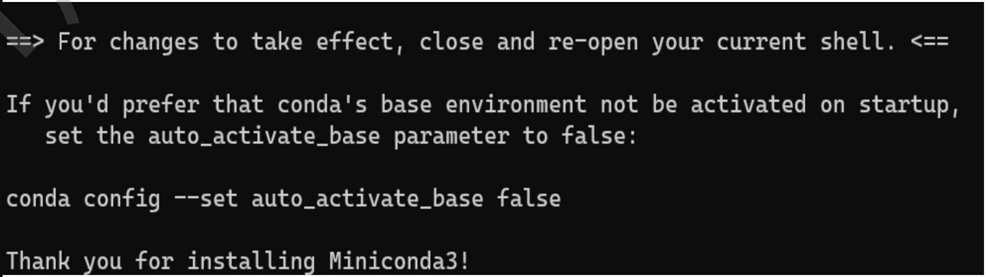
运行 wget -c https://mirrors.bfsu.edu.cn/anaconda/miniconda/Miniconda3-py37_4.9.2-Linux-x86_64.sh

运行 sh Miniconda3-py37_4.9.2-Linux-x86_64.sh



```
Do you accept the license terms? [yes|no]
[no] >>> yes

 Preparing transaction: done
 Executing transaction: done
 installation finished.
 Do you wish the installer to initialize Miniconda3
 by running conda init? [yes|no]
 [no] >>> yes

Do you accept the license terms? [yes|no]
[no] >>> yes

Miniconda3 will now be installed into this location:
/home/mgi/miniconda3

  - Press ENTER to confirm the location
  - Press CTRL-C to abort the installation
  - Or specify a different location below

[/home/mgi/miniconda3] >>> /home/mgi/miniconda3
```

安装路径



```
==> For changes to take effect, close and re-open your current shell. <==

If you'd prefer that conda's base environment not be activated on startup,
   set the auto_activate_base parameter to false:

conda config --set auto_activate_base false

Thank you for installing Miniconda3!
```

DNBelab_C4_scRNA_V2流程安装

DNBC4tools子环境部署：

首先进入conda的base环境source local_path_to/miniconda3/bin/activate

然后在软件包目录运行

```
conda env create -f DNBC4tools.yaml -n DNBC4tools
```

一键化部署DNBC4tools

安装3个R包：

```
conda activate DNBC4tools
Rscript -e “devtools::install_github(c(‘chris-mcginnis-ucsf/DoubletFinder’,‘ggjlab/scHCL’,‘ggjlab/scMCA’),force = TRUE);”
```

下载cromwell.jar并保存至wdl/文件夹下

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

比对数据库构建

1.过滤假基因

```
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
```

- 确认gtf中是gene_type还是gene_biotype
- 如果是biotype，需要将以下命令中的type改为biotype（如果是type则不需要改动）

比对数据库构建

2.构建STAR比对数据库：

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

*DNBC4tools构建数据库过程参考cell ranger，如果有cell ranger6.0之后的版本对应的数据库，可以直接使用。

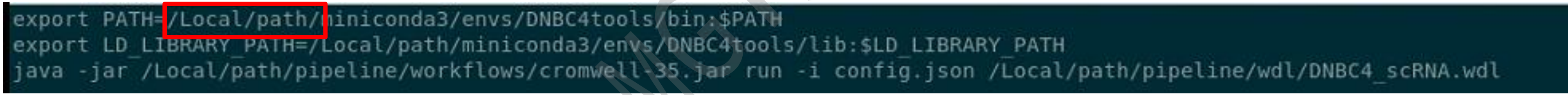
运行方法

1.wdl模式运行

（所需文件在example/wdl中）

- 修改run.sh路径，并填写config.json
- 运行run.sh

run.sh:



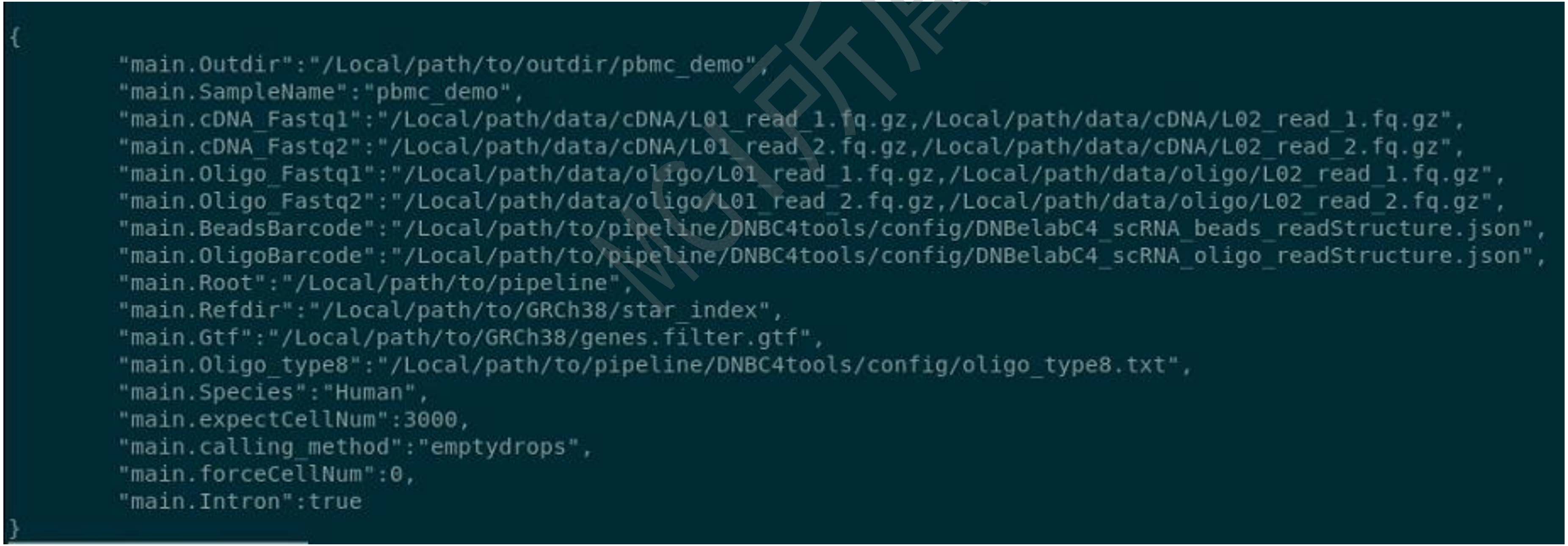
```
export PATH=/Local/path/miniconda3/envs/DNBC4tools/bin:$PATH
export LD_LIBRARY_PATH=/Local/path/miniconda3/envs/DNBC4tools/lib:$LD_LIBRARY_PATH
java -jar /Local/path/pipeline/workflows/cromwell-35.jar run -i config.json /Local/path/pipeline/wdl/DNBC4_scRNA.wdl
```

Local/path处需要替换为对应的路径，确保补充后路径正确（ls + 路径检查）。

运行方法

1.wdl模式运行

config.json

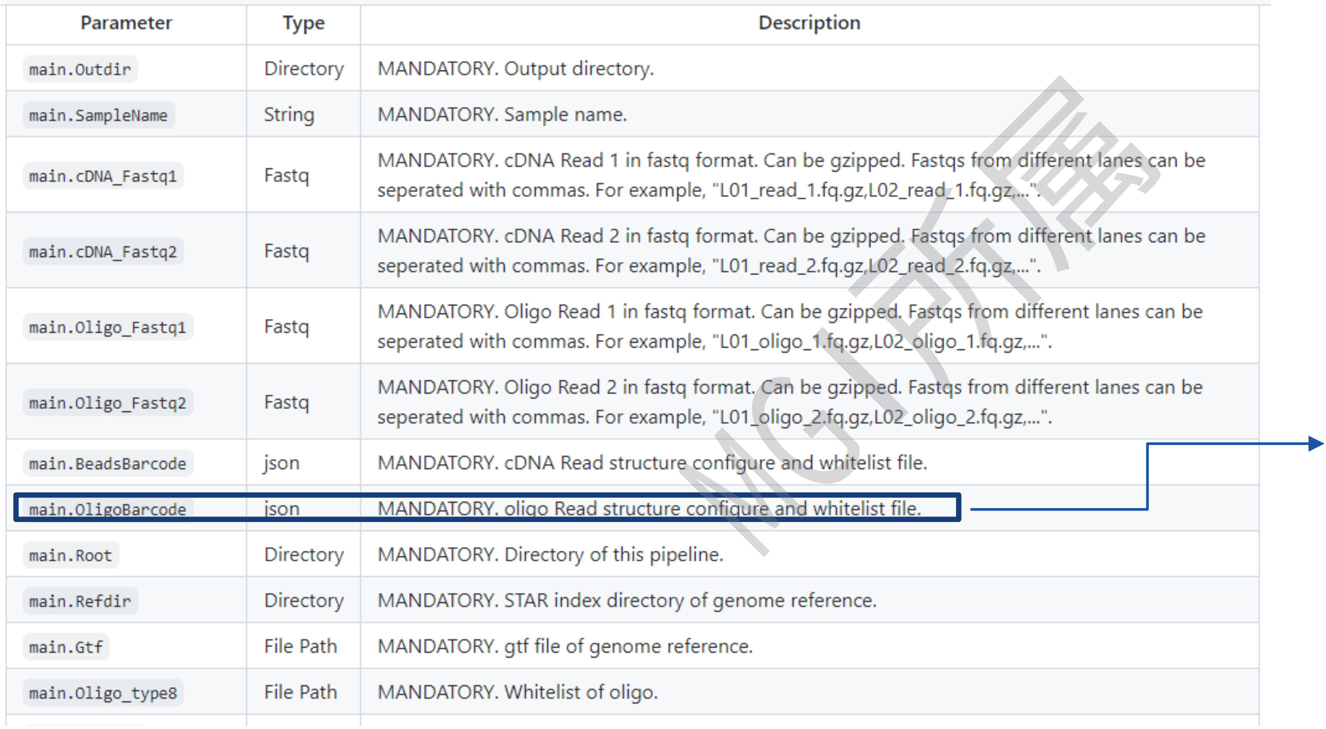


```
{
        "main.Outdir":"/Local/path/to/outdir/pbmc_demo",
        "main.SampleName":"pbmc_demo",
        "main.cDNA_Fastq1":"/Local/path/data/cDNA/L01_read_1.fq.gz,/Local/path/data/cDNA/L02_read_1.fq.gz",
        "main.cDNA_Fastq2":"/Local/path/data/cDNA/L01_read_2.fq.gz,/Local/path/data/cDNA/L02_read_2.fq.gz",
        "main.Oligo_Fastq1":"/Local/path/data/oligo/L01_read_1.fq.gz,/Local/path/data/oligo/L02_read_1.fq.gz",
        "main.Oligo_Fastq2":"/Local/path/data/oligo/L01_read_2.fq.gz,/Local/path/data/oligo/L02_read_2.fq.gz",
        "main.BeadsBarcode":"/Local/path/to/pipeline/DNBC4tools/config/DNBelabC4_scRNA_beads_readStructure.json",
        "main.OligoBarcode":"/Local/path/to/pipeline/DNBC4tools/config/DNBelabC4_scRNA_oligo_readStructure.json",
        "main.Root":"/Local/path/to/pipeline",
        "main.Refdir":"/Local/path/to/GRCh38/star_index",
        "main.Gtf":"/Local/path/to/GRCh38/genes.filter.gtf",
        "main.Oligo_type8":"/Local/path/to/pipeline/DNBC4tools/config/oligo_type8.txt",
        "main.Species":"Human",
        "main.expectCellNum":3000,
        "main.calling_method":"emptydrops",
        "main.forceCellNum":0,
        "main.Intron":true
}
```

运行方法

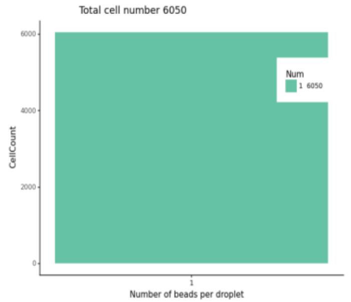
1.wdl模式运行

注：所有路径使用绝对路径，不能使用”~”



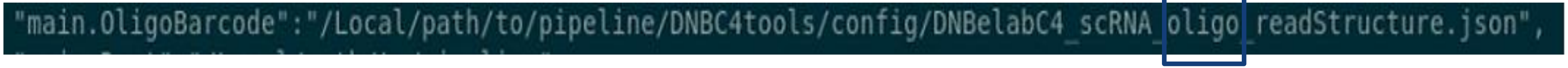
Parameter	Type	Description
main.Outdir	Directory	MANDATORY. Output directory.
main.SampleName	String	MANDATORY. Sample name.
main.cDNA_Fastq1	Fastq	MANDATORY. cDNA Read 1 in fastq format. Can be gzipped. Fastqs from different lanes can be seperated with commas. For example, "L01_read_1.fq.gz,L02_read_1.fq.gz,...".
main.cDNA_Fastq2	Fastq	MANDATORY. cDNA Read 2 in fastq format. Can be gzipped. Fastqs from different lanes can be seperated with commas. For example, "L01_read_2.fq.gz,L02_read_2.fq.gz,...".
main.Oligo_Fastq1	Fastq	MANDATORY. Oligo Read 1 in fastq format. Can be gzipped. Fastqs from different lanes can be seperated with commas. For example, "L01_oligo_1.fq.gz,L02_oligo_1.fq.gz,...".
main.Oligo_Fastq2	Fastq	MANDATORY. Oligo Read 2 in fastq format. Can be gzipped. Fastqs from different lanes can be seperated with commas. For example, "L01_oligo_2.fq.gz,L02_oligo_2.fq.gz,...".
main.BeadsBarcode	json	MANDATORY. cDNA Read structure configure and whitelist file.
main.OligoBarcode	json	MANDATORY. oligo Read structure configure and whitelist file.
main.Root	Directory	MANDATORY. Directory of this pipeline.
main.Refdir	Directory	MANDATORY. STAR index directory of genome reference.
main.Gtf	File Path	MANDATORY. gtf file of genome reference.
main.Oligo_type8	File Path	MANDATORY. Whitelist of oligo.

Beads未合并



如果为同一张芯片混测，OligoBarcode需要更换为oligomix

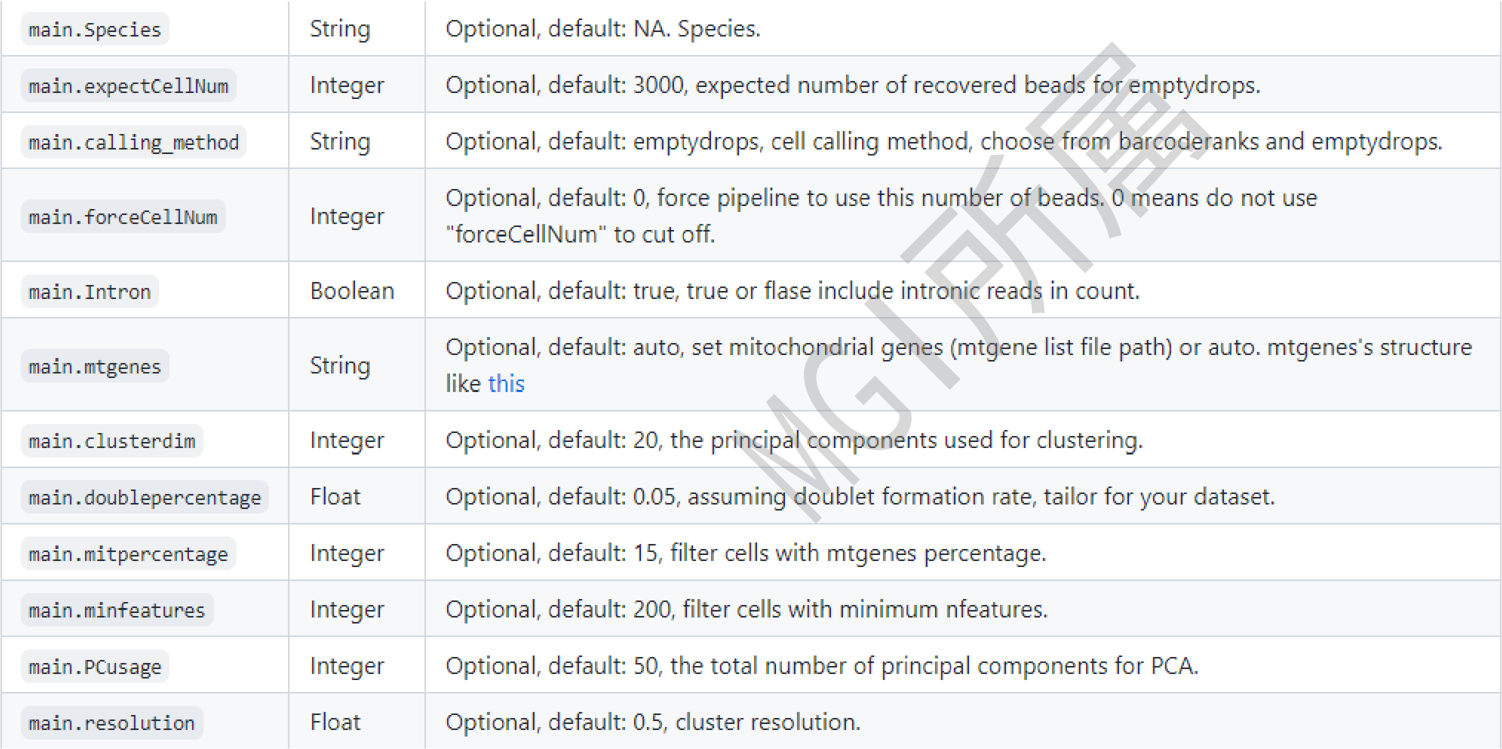
改为oligomix



```
"main.OligoBarcode":"/Local/path/to/pipeline/DNBC4tools/config/DNBelabC4_scRNA_oligo_readStructure.json",
```

运行方法

1.wdl模式运行



`main.Species`	String	Optional, default: NA. Species.
`main.expectCellNum`	Integer	Optional, default: 3000, expected number of recovered beads for emptydrops.
`main.calling_method`	String	Optional, default: emptydrops, cell calling method, choose from barcoderanks and emptydrops.
`main.forceCellNum`	Integer	Optional, default: 0, force pipeline to use this number of beads. 0 means do not use "forceCellNum" to cut off.
`main.Intron`	Boolean	Optional, default: true, true or flase include intronic reads in count.
`main.mtgenes`	String	Optional, default: auto, set mitochondrial genes (mtgene list file path) or auto. mtgenes's structure like this
`main.clusterdim`	Integer	Optional, default: 20, the principal components used for clustering.
`main.doublepercentage`	Float	Optional, default: 0.05, assuming doublet formation rate, tailor for your dataset.
`main.mitpercentage`	Integer	Optional, default: 15, filter cells with mtgenes percentage.
`main.minfeatures`	Integer	Optional, default: 200, filter cells with minimum nfeatures.
`main.PCusage`	Integer	Optional, default: 50, the total number of principal components for PCA.
`main.resolution`	Float	Optional, default: 0.5, cluster resolution.

calling_method可以切换barcoderank/emptydrop两种cell-calling方法

Mtgenes可以指定线粒体文件，格式在github有说明

Intron可调整是否使用intron作为表达量计数

运行方法

2.命令行模式运行

- 将以下内容添加入~/.bashrc，并且source ~/.bashrc

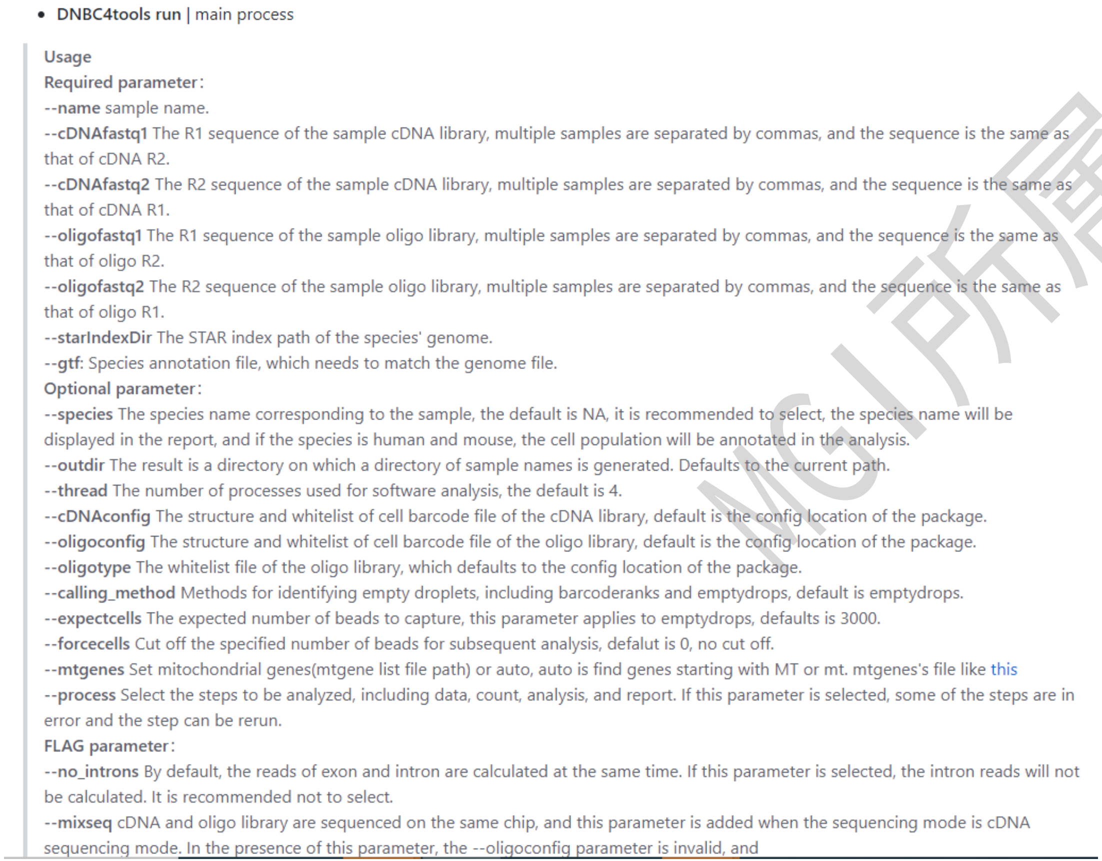
```
alias DNBC4tools='/MGI/miniconda3/envs/DNBC4tools/bin/DNBC4tools’
```

- 进入DNBC4tools环境，或者运行run.sh中的前两行，以载入需要的环境变量
- 运行示例：

```
DNBC4tools run --cDNAfastq1 /test/data/test_cDNA_R1.fastq.gz --cDNAfastq2
/test/data/test_cDNA_R2.fastq.gz
--oligofastq1 /test/data/test_oligo1_1.fq.gz,/test/data/test_oligo2_1.fq.gz --
oligofastq2 /test/data/test_oligo1_2.fq.gz,/test/data/test_oligo2_2.fq.gz
--starIndexDir /database/Mouse/mm10/ --gtf /database/Mouse/mm10/genes.gtf --
name test --species Mouse --thread 10
```

运行方法

2.命令行模式运行



- **DNBC4tools run** | main process

Usage

Required parameter:

--name sample name.

--cDNAfastq1 The R1 sequence of the sample cDNA library, multiple samples are separated by commas, and the sequence is the same as that of cDNA R2.

--cDNAfastq2 The R2 sequence of the sample cDNA library, multiple samples are separated by commas, and the sequence is the same as that of cDNA R1.

--oligofastq1 The R1 sequence of the sample oligo library, multiple samples are separated by commas, and the sequence is the same as that of oligo R2.

--oligofastq2 The R2 sequence of the sample oligo library, multiple samples are separated by commas, and the sequence is the same as that of oligo R1.

--starIndexDir The STAR index path of the species' genome.

--gtf: Species annotation file, which needs to match the genome file.

Optional parameter:

--species The species name corresponding to the sample, the default is NA, it is recommended to select, the species name will be displayed in the report, and if the species is human and mouse, the cell population will be annotated in the analysis.

--outdir The result is a directory on which a directory of sample names is generated. Defaults to the current path.

--thread The number of processes used for software analysis, the default is 4.

--cDNAconfig The structure and whitelist of cell barcode file of the cDNA library, default is the config location of the package.

--oligoconfig The structure and whitelist of cell barcode file of the oligo library, default is the config location of the package.

--oligotype The whitelist file of the oligo library, which defaults to the config location of the package.

--calling_method Methods for identifying empty droplets, including barcoderanks and emptydrops, default is emptydrops.

--expectcells The expected number of beads to capture, this parameter applies to emptydrops, defaults is 3000.

--forcecells Cut off the specified number of beads for subsequent analysis, defalut is 0, no cut off.

--mtgenes Set mitochondrial genes(mtgene list file path) or auto, auto is find genes starting with MT or mt. mtgenes's file like this

--process Select the steps to be analyzed, including data, count, analysis, and report. If this parameter is selected, some of the steps are in error and the step can be rerun.

FLAG parameter:

--no_introns By default, the reads of exon and intron are calculated at the same time. If this parameter is selected, the intron reads will not be calculated. It is recommended not to select.

--mixseq cDNA and oligo library are sequenced on the same chip, and this parameter is added when the sequencing mode is cDNA sequencing mode. In the presence of this parameter, the --oligoconfig parameter is invalid, and

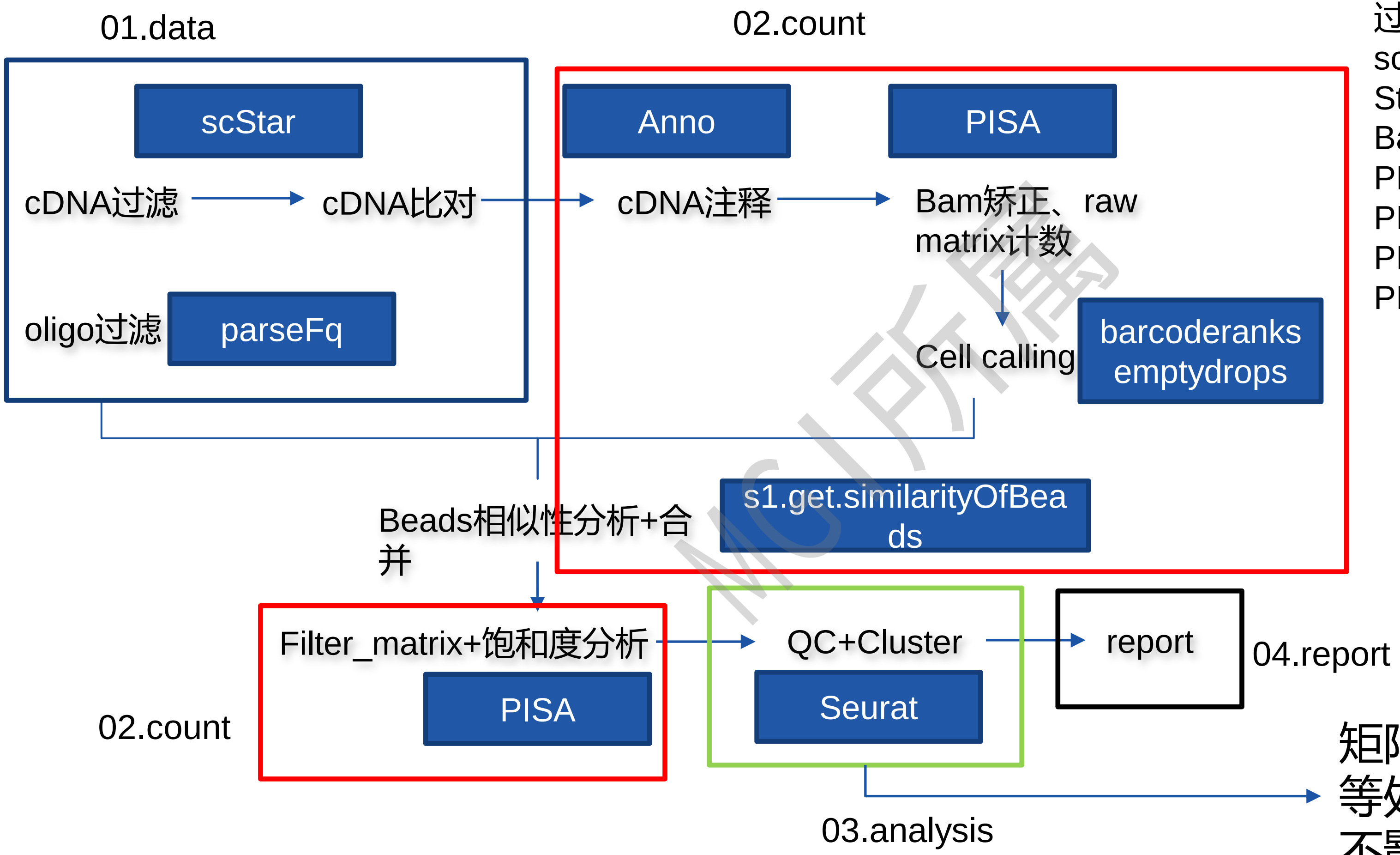
- 如果为同一张芯片混测，需要加--mixseq

- DNBC4tools multi可支持多个样本一键生成多个脚本，样本list格式见github

- --process + 对应的模块（data、count、analysis、report）可以运行部分模块

详细参数说明与范例见github

DNBelab_C4_scRNA_V2生信分析流程概览

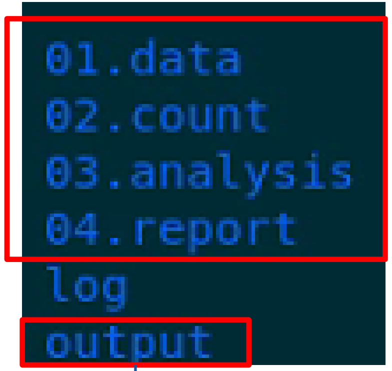


过滤和比对使用到的软件：
scRNA_parse
Star
Bam处理：
PISA sam2bam
PISA anno
PISA corr
PISA attrcnt

矩阵QC、聚类等处理仅供参考不影响矩阵输出

软件输出结果

中间文件



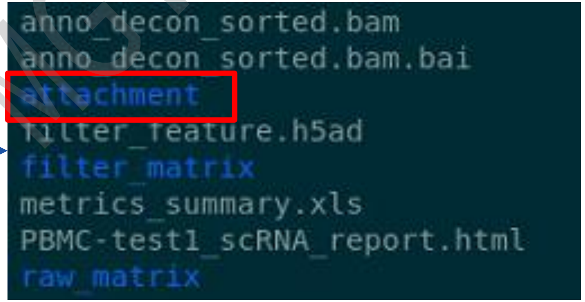
```
01.data
02.count
03.analysis
04.report
log
output
```

Parsefq 过滤后fastq、比对bam等文件

scSTAR+anno+PISA 处理后bam、相似度计算等文件

Seurat QC+cluster+细胞注释相关文件

网页Report、各种图表



```
anno_decon_sorted.bam
anno_decon_sorted.bam.bai
attachment
filter_feature.h5ad
filter_matrix
metrics_summary.xls
PBMC-test1_scRNA_report.html
raw_matrix
```

最终输出

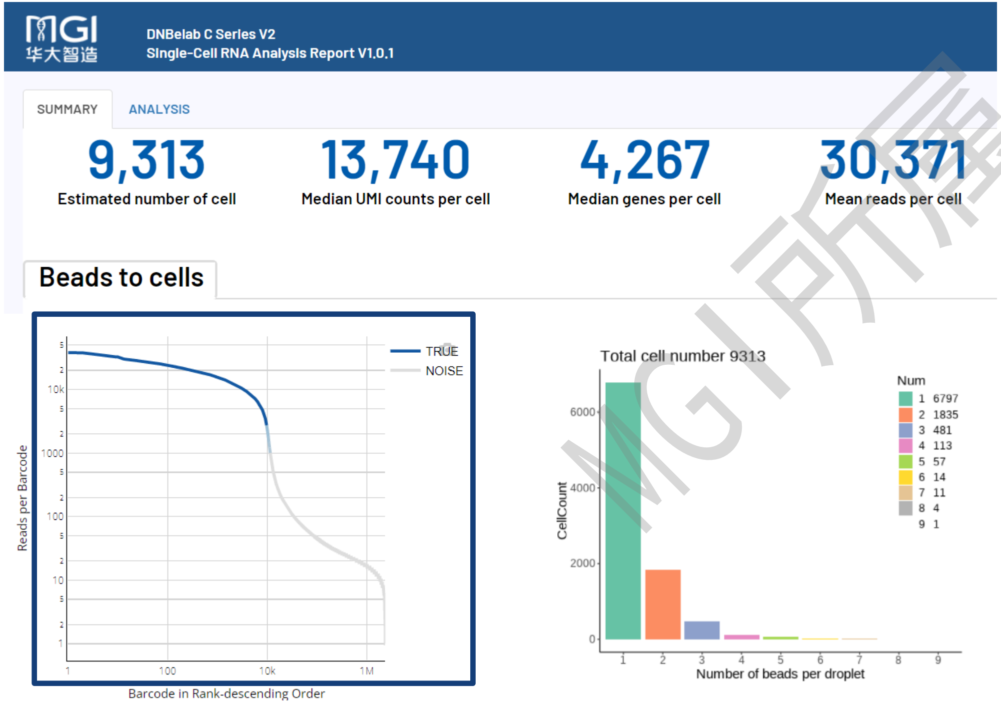
附件：
分析可能用到的各种中间文件
包括QC、cluster、RNA速率分析矩阵、仅exon矩阵

3

报告解读

报告参数解读

报告概览 & Beads 筛选 + Beads 合并



DNBelab C Series V2
Single-Cell RNA Analysis Report V1.0.1

SUMMARY ANALYSIS

9,313 Estimated number of cell

13,740 Median UMI counts per cell

4,267 Median genes per cell

30,371 Mean reads per cell

Beads to cells

UMI

细胞数

默认筛选规则：

Umi阈值+emptydrop

可选：

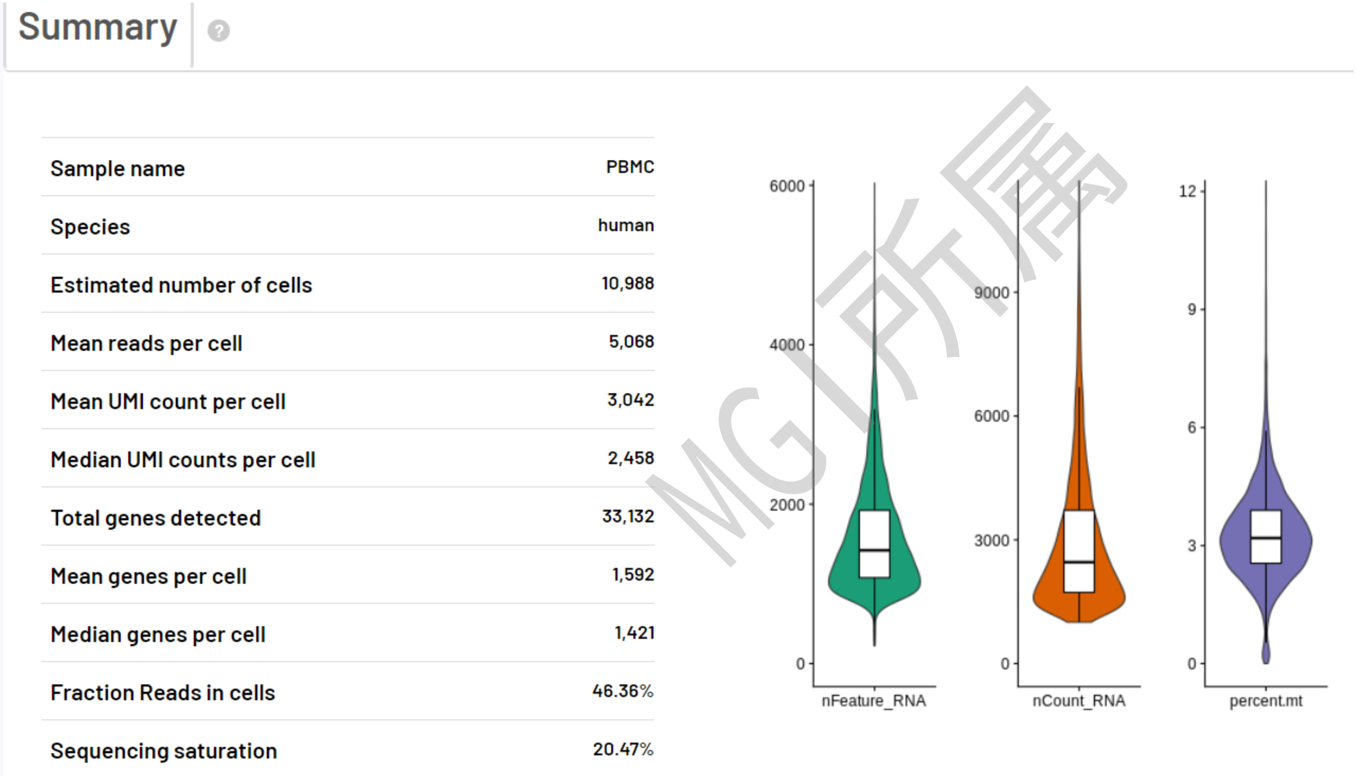
Barcoderank（斜率法）

标准曲线形状：
第一阶段平缓下降
第二阶段骤降并进入次平缓阶段，即形成cliff and knee
第三阶段降低至纵坐标为1

*若有杂质或者细胞状态不佳，则曲线的骤降将会不明显

报告参数解读

summary&质控图



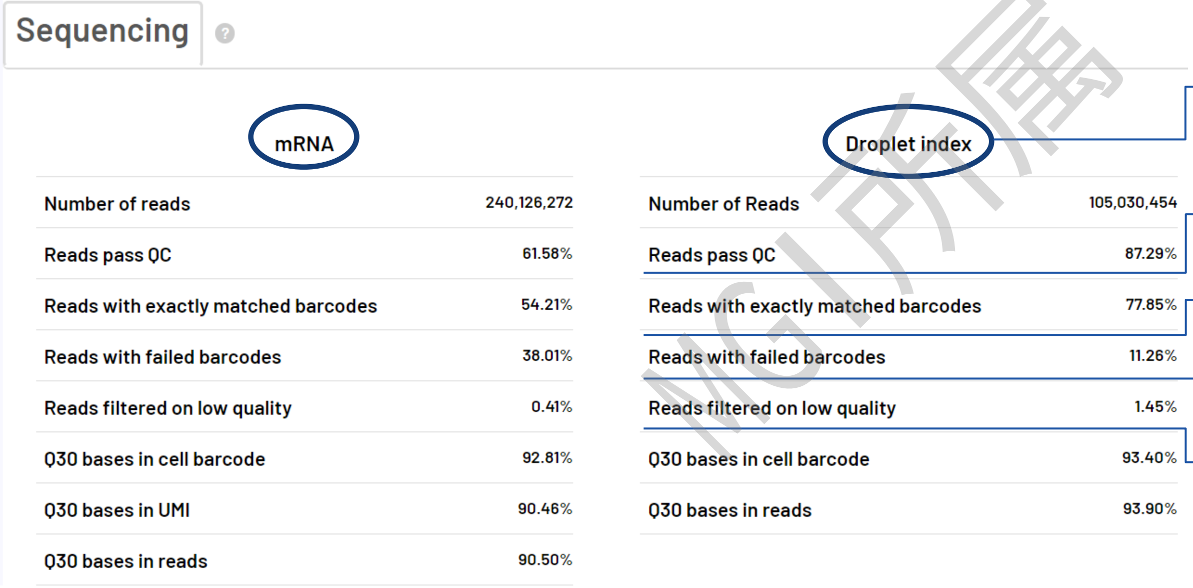
Summary

Sample name	PBMC
Species	human
Estimated number of cells	10,988
Mean reads per cell	5,068
Mean UMI count per cell	3,042
Median UMI counts per cell	2,458
Total genes detected	33,132
Mean genes per cell	1,592
Median genes per cell	1,421
Fraction Reads in cells	46.36%
Sequencing saturation	20.47%

报告参数解读

数据过滤和映射情况

Low quality+failed barcode=total reads（100%）-readspassQC



Sequencing

mRNA	
Number of reads	240,126,272
Reads pass QC	61.58%
Reads with exactly matched barcodes	54.21%
Reads with failed barcodes	38.01%
Reads filtered on low quality	0.41%
Q30 bases in cell barcode	92.81%
Q30 bases in UMI	90.46%
Q30 bases in reads	90.50%

Droplet index	
Number of Reads	105,030,454
Reads pass QC	87.29%
Reads with exactly matched barcodes	77.85%
Reads with failed barcodes	11.26%
Reads filtered on low quality	1.45%
Q30 bases in cell barcode	93.40%
Q30 bases in reads	93.90%

cDNA/oligo的数据
推荐300M/100M reads

白名单映射（默认容错1）& 质量值QC

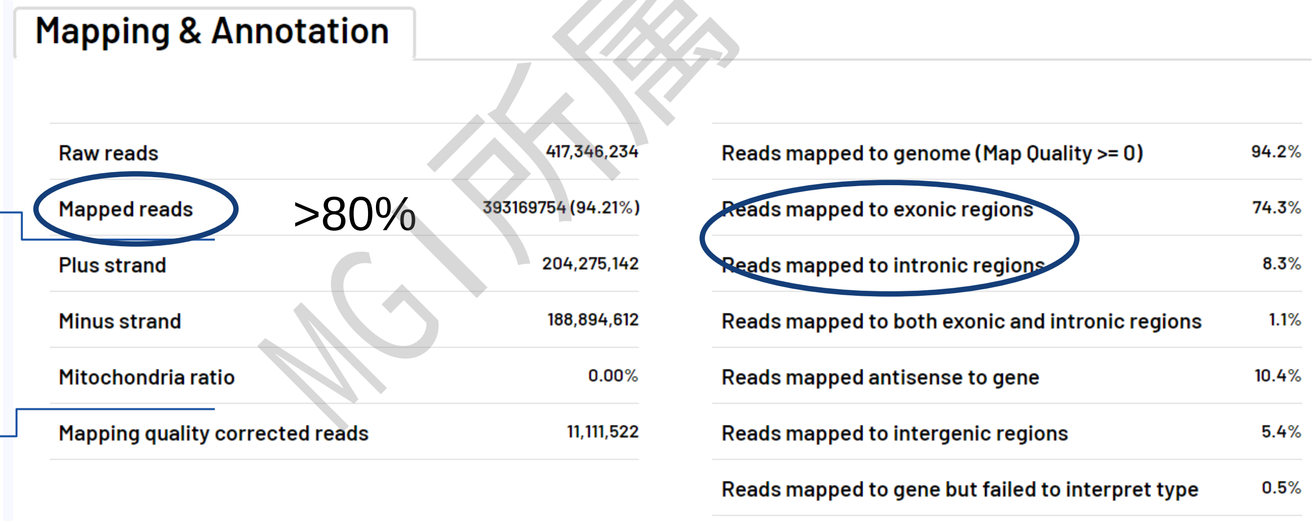
无容错映射结果

不在白名单内的reads

质量值过低reads

报告参数解读

比对和注释



Mapping & Annotation

Raw reads	417,346,234
Mapped reads	393169754 (94.21%)
Plus strand	204,275,142
Minus strand	188,894,612
Mitochondria ratio	0.00%
Mapping quality corrected reads	11,111,522

Reads mapped to genome (Map Quality >= 0)	94.2%
Reads mapped to exonic regions	74.3%
Reads mapped to intronic regions	8.3%
Reads mapped to both exonic and intronic regions	1.1%
Reads mapped antisense to gene	10.4%
Reads mapped to intergenic regions	5.4%
Reads mapped to gene but failed to interpret type	0.5%

参考库比对率

>80%

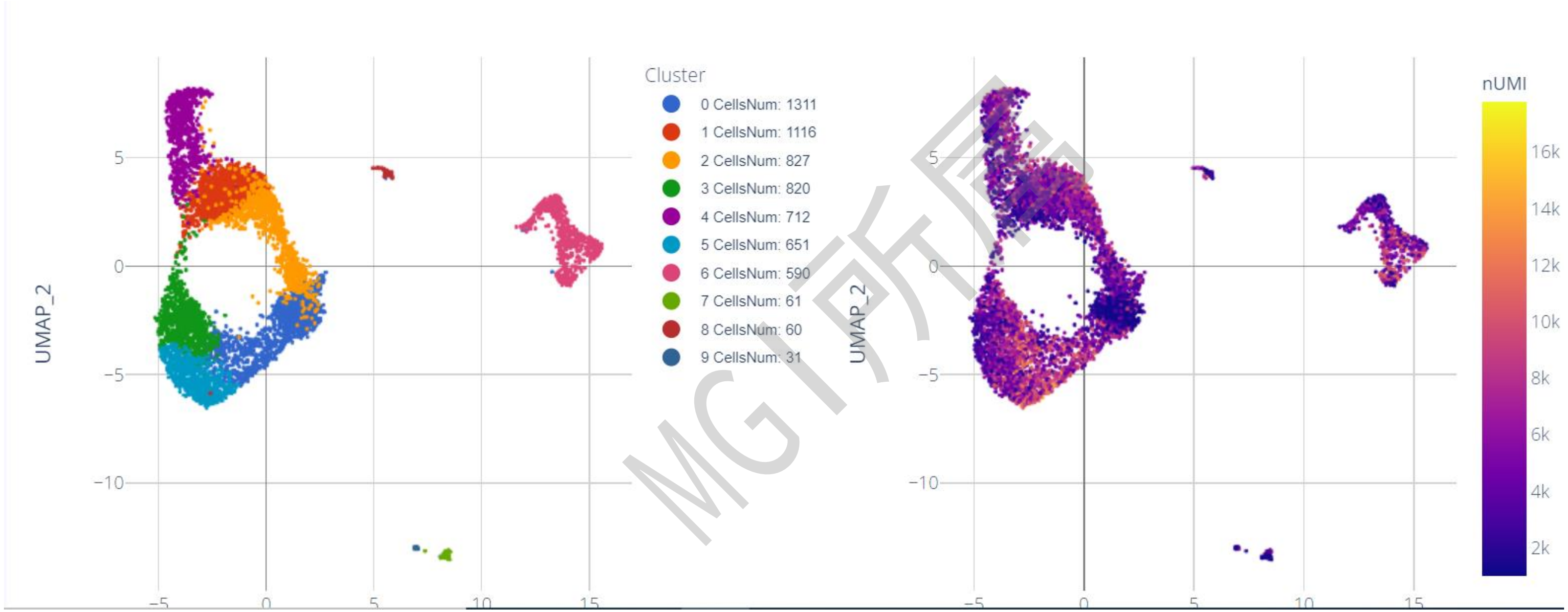
线粒体比例

(若为0则需要手动输入chrM文件)

Exon+intron>50%

报告参数解读

UMAP聚类图 & UMI丰度热图



高亮部分 = 聚类更可信
高亮平均分布 = 聚类整体情况更可信

Thanks for your time!

感谢你的聆听！

MGI 华大智造

深圳市盐田区北山工业区11栋
Building No.11, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China
MGI-service@mgi-tech.com | www.mgi-tech.com