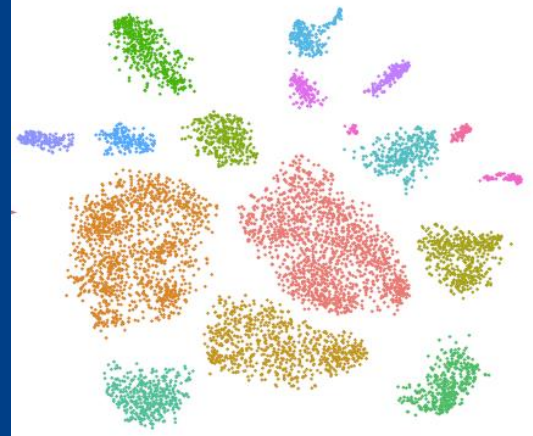
浙江大学
ZHEJIANG UNIVERSITY

单细胞转录组数据分析

——从原始数据到细胞类型注释



褚琴洁　qinjiechu@zju.edu.cn

2023年10月16日

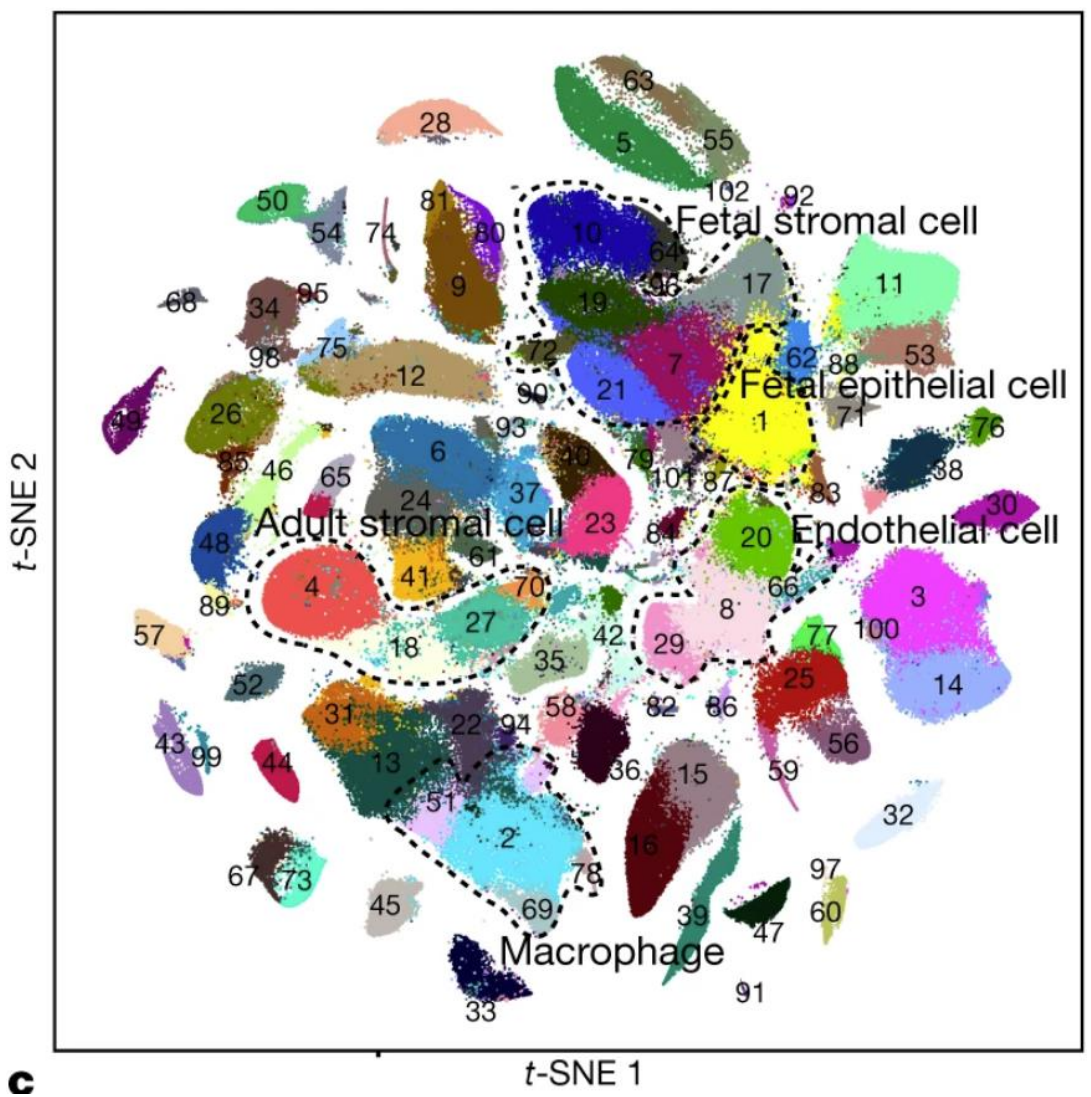
课件PPT

http://ibi.zju.edu.cn/bioinplant/courses/scomics/

人类细胞图谱

nature

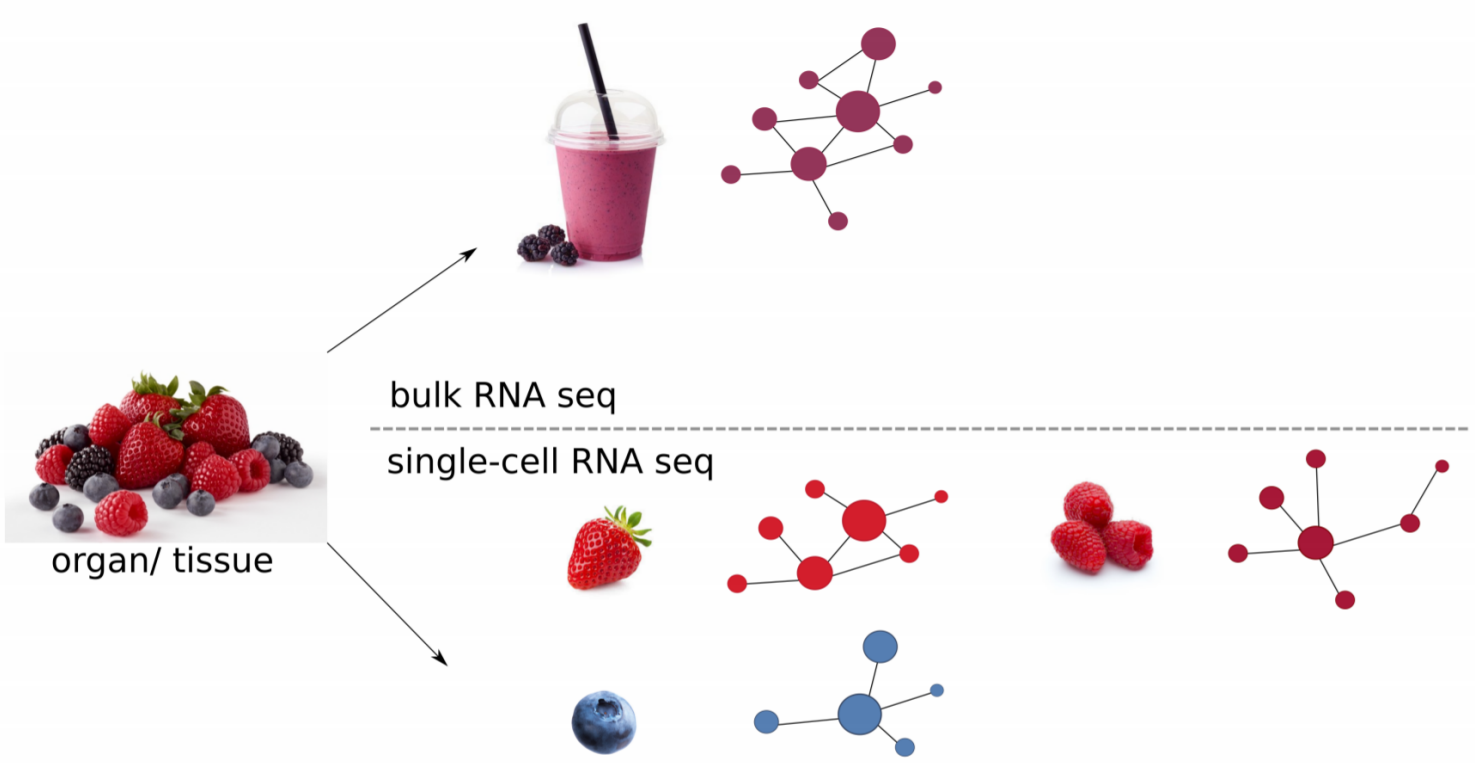
Construction of a human cell landscape at single-cell level



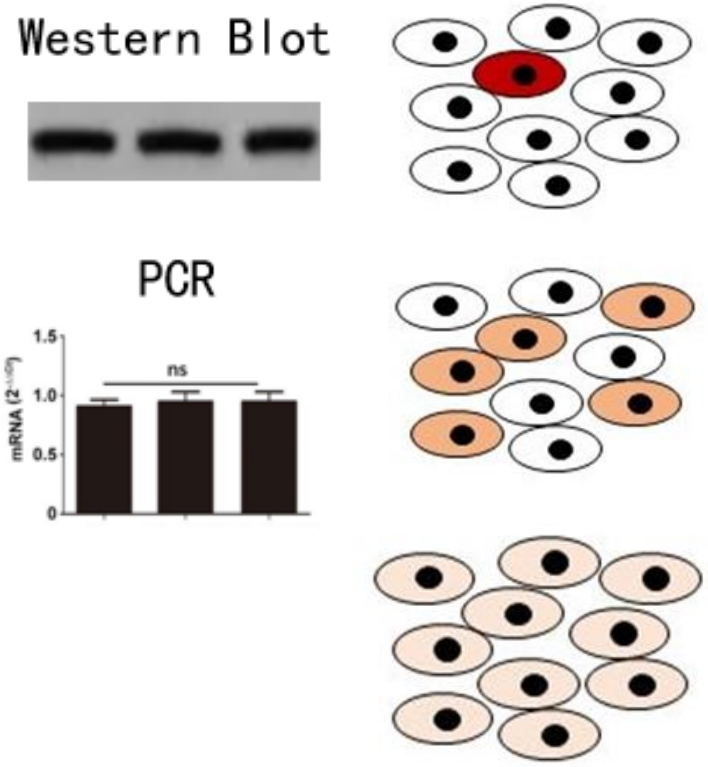
- Adult adipose
- Adult adrenal gland
- Adult artery
- Adult ascending colon
- Adult bladder
- Adult bone marrow
- Adult cerebellum
- Adult cervix
- Adult duodenum
- Adult epityphlon
- Adult oesophagus
- Adult fallopian tube
- Adult gall bladder
- Adult heart
- Adult ileum
- Adult jejunum
- Adult kidney
- Adult liver
- Adult lung
- Adult muscle
- Adult omentum
- Adult pancreas
- Adult peripheral blood
- Adult pleura
- Adult prostate
- Adult rectum
- Adult sigmoid colon
- Adult spleen
- Adult stomach
- Adult temporal lobe
- Adult thyroid gland
- Adult trachea
- Adult transverse colon
- Adult ureter
- Adult uterus
- Chorionic villus
- Cord blood
- Cord blood CD34P
- Fetal adrenal gland
- Fetal brain
- Fetal calvaria
- Fetal eyes
- Fetal female gonad
- Fetal heart
- Fetal intestine
- Fetal kidney
- Fetal liver
- Fetal lung
- Fetal male gonad
- Fetal muscle
- Fetal pancreas
- Fetal rib
- Fetal skin
- Fetal spinal cord
- Fetal stomach
- Fetal thymus
- Human ES cells
- Neonatal adrenal gland
- Placenta

为什么要进行单细胞研究?

- 没有完全相同的两片树叶



- 无法区分以下情况

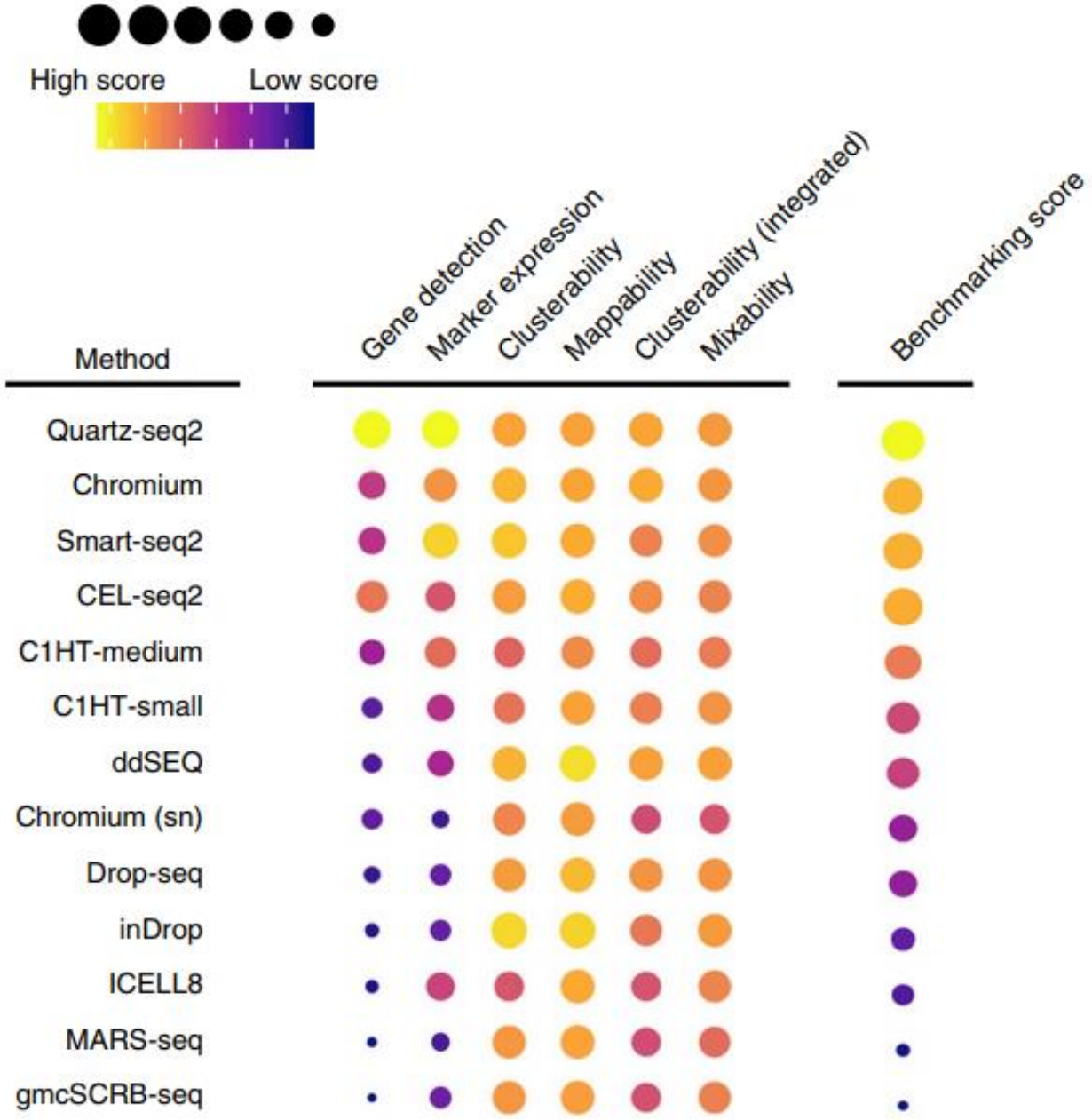


Steinheuer et al., bioRxiv, 2021
https://mp.weixin.qq.com/s/5laEGHM2iWSLRu0iaBbr7Q

单细胞转录组测序技术比较

- 综合比较结果

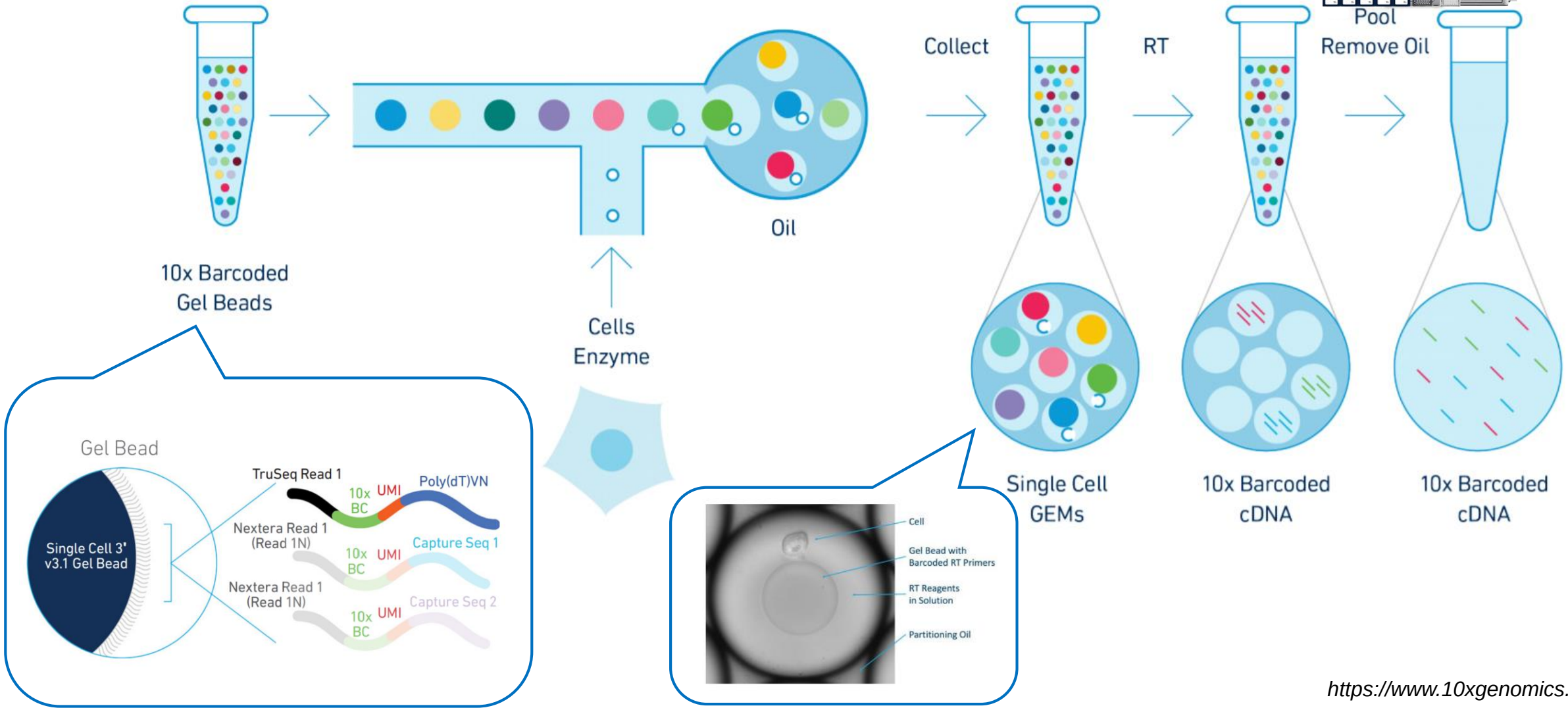
Methods	Transcript coverage	UMI possibility	Strand specific	References
Tang method	Nearly full-length	No	No	Tang et al., 2009
Quartz-Seq	Full-length	No	No	Sasagawa et al., 2013
SUPeR-seq	Full-length	No	No	Fan X. et al., 2015
Smart-seq	Full-length	No	No	Ramskold et al., 2012
Smart-seq2	Full-length	No	No	Picelli et al., 2013
MATQ-seq	Full-length	Yes	Yes	Sheng et al., 2017
STRT-seq and STRT/C1	5′-only	Yes	Yes	Islam et al., 2011, 2012
CEL-seq	3′-only	Yes	Yes	Hashimshony et al., 2012
CEL-seq2	3′-only	Yes	Yes	Hashimshony et al., 2016
MARS-seq	3′-only	Yes	Yes	Jaitin et al., 2014
CytoSeq	3′-only	Yes	Yes	Fan H.C. et al., 2015
Drop-seq	3′-only	Yes	Yes	Macosko et al., 2015
InDrop	3′-only	Yes	Yes	Klein et al., 2015
Chromium	3′-only	Yes	Yes	Zheng et al., 2017
SPLiT-seq	3′-only	Yes	Yes	Rosenberg et al., 2018
sci-RNA-seq	3′-only	Yes	Yes	Cao et al., 2017
Seq-Well	3′-only	Yes	Yes	Gierahn et al., 2017
DroNC-seq	3′-only	Yes	Yes	Habib et al., 2017
Quartz-Seq2	3′-only	Yes	Yes	Sasagawa et al., 2018



Mereu et al., Nature Biotechnology , 2020
Chen et al., Frontiers in Genetics. 2019

单细胞转录组测序

- 微流控 The Chromium Single Cell Gene Expression Solution



https://www.10xgenomics.com

单细胞转录组测序

- 微孔板 BD Rhapsody

Microwell technology
- No sample loss due to clogging of channels
- No electronics, portable
- 575 µL cells suspension loading volume

Visual workflow QC
Confidence with every experiment

Up to **80** % cartridge capture rate
for certain cell types

Low multiplet rate
2–3% @ 10,000 cell load
8–10% @ 40,000 cell load

Broad range of cell throughput
100–40,000 cells per cartridge

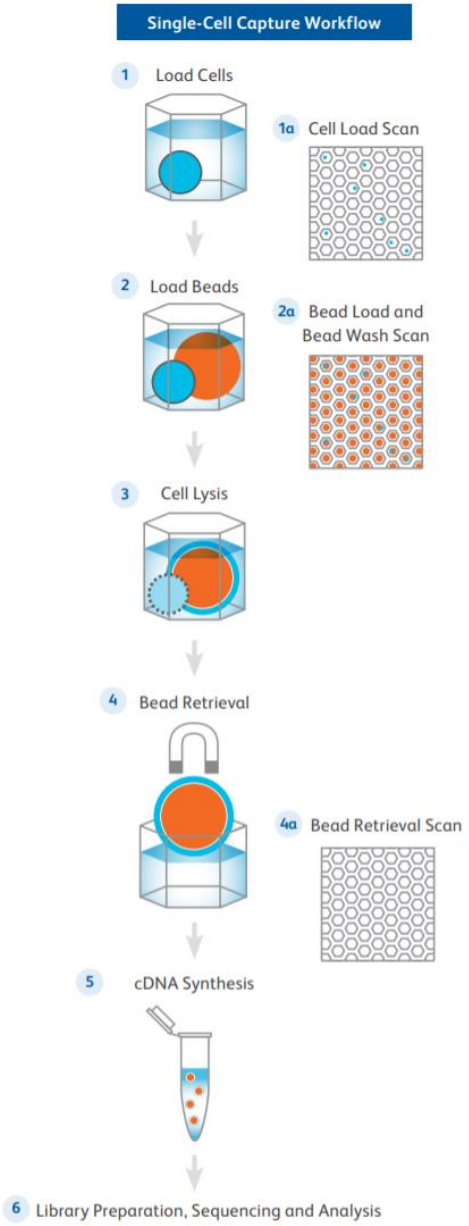
Capture and analyze fragile cells
Granulocytes, neutrophils, CAR-T cells, stem cells, tumor xenograft-derived cells, myeloma, T cells, NK cells and more

Minimal batch effects*
Consistent, reliable results with technical, biological, site-to-site and user-to-user replicates

High correlation with flow data
The same trusted BD antibodies for flow and single-cell multiomics

Subsample beads
Flexibility with experimental design, tool to measure reliability, work with collaborators

Archive beads
Equivalent data obtained from fresh beads and beads stored for several months



Single-Cell Capture Workflow

1 Load Cells
- Load 575 µL cell suspension
- Clogging of channels are not a concern as there are no microfluidic channels
- Gentle settling of cells by gravity allows capture of fragile cells
- The BD Rhapsody™ Cartridge contains >220,000 partitions for capture of up to 40,000 cells at a low multiplet rate

1a Cell Load Scan
- Estimate the number of viable cells captured and the cell multiplet rate

2 Load Beads
- Geometry and dimension of the microwell prevents bead multiplets

2a Bead Load and Bead Wash Scan
- Estimate the number of wells with a viable cell and a bead
- Measure cell retention rate to assess if cells have been lost

3 Cell Lysis
- Strong lysis buffers ensure complete lysis of cells

4 Bead Retrieval
- Easy, efficient magnetic retrieval of beads

4a Bead Retrieval Scan
- Confirm complete retrieval of beads

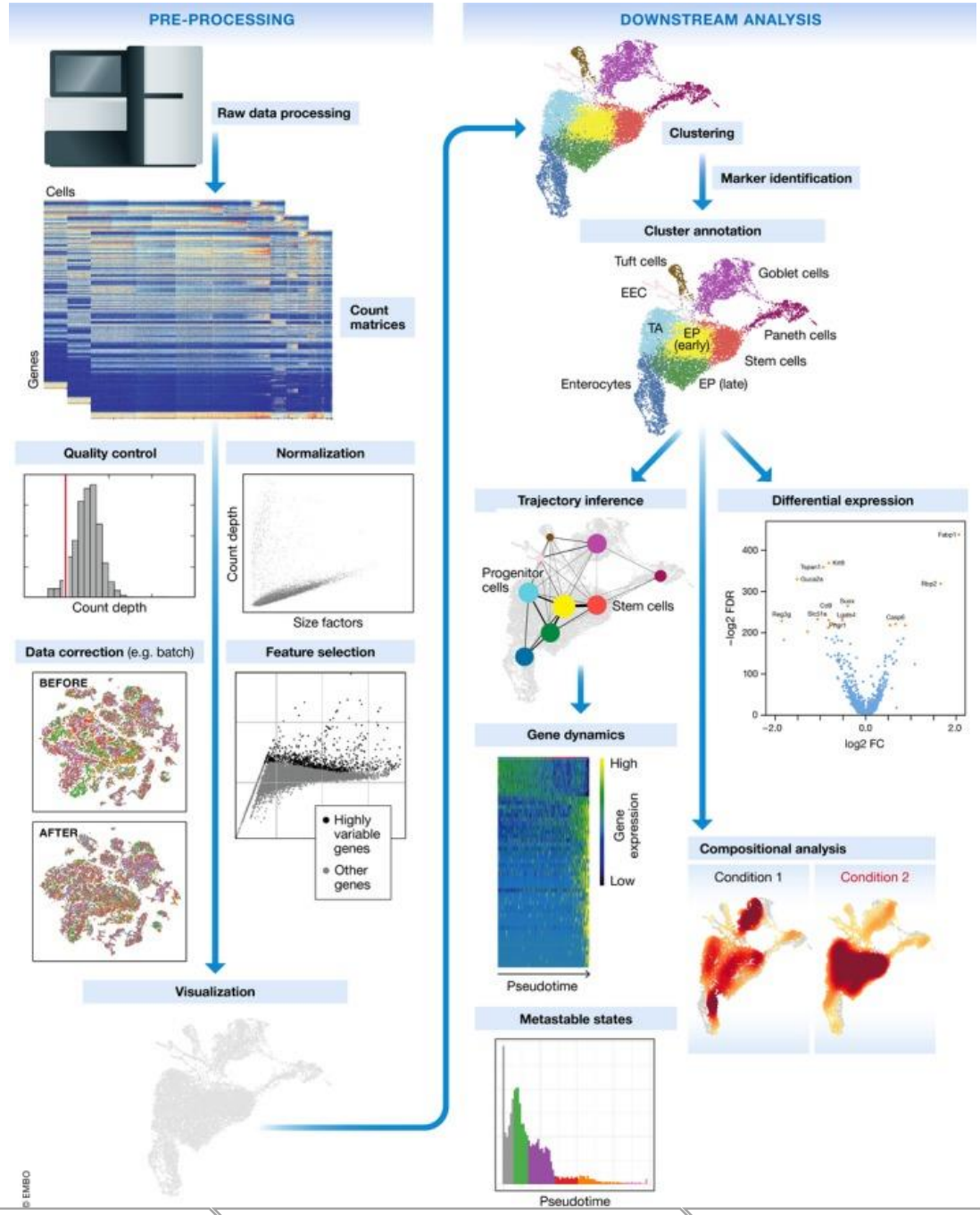
5 cDNA Synthesis
- Multiple bead washes remove contaminants and allow for more effective reverse transcription
- Beads can be archived or subsampled for more experimental flexibility

6 Library Preparation, Sequencing and Analysis
- Bioinformatics solutions including the BD Rhapsody™ Analysis Pipelines and SeqGeq™ Software provide a complete end-to-end single-cell solution

https://www.bdbiosciences.com/

单细胞转录组数据分析流程

- 前处理 Pre-processing
- 质控 Quality control
- 标准化 Normalization
- 特征基因选择 Feature selection
- 降维 Dimensionality reduction
- 聚类 Cluster analysis
- 细胞类型注释 Cell type annotation
- 数据整合 Data integration
- 拟时分析 Trajectory analysis
- 差异基因分析
- 细胞通讯
- ……

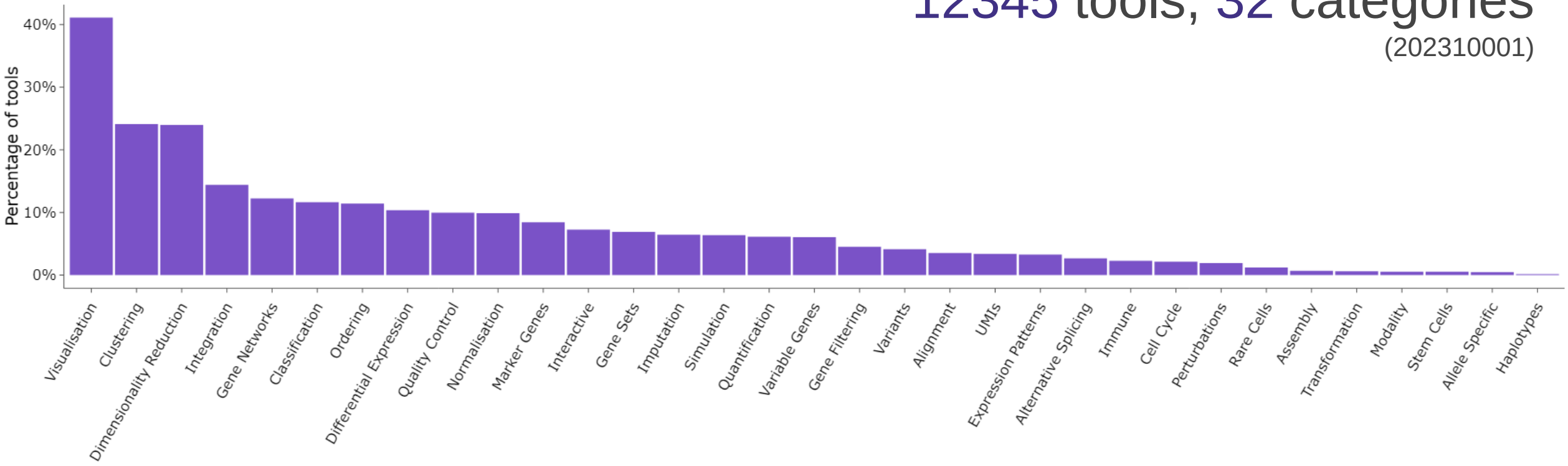


单细胞转录组数据分析方法

- 不同种类，涉及到不同的分析步骤

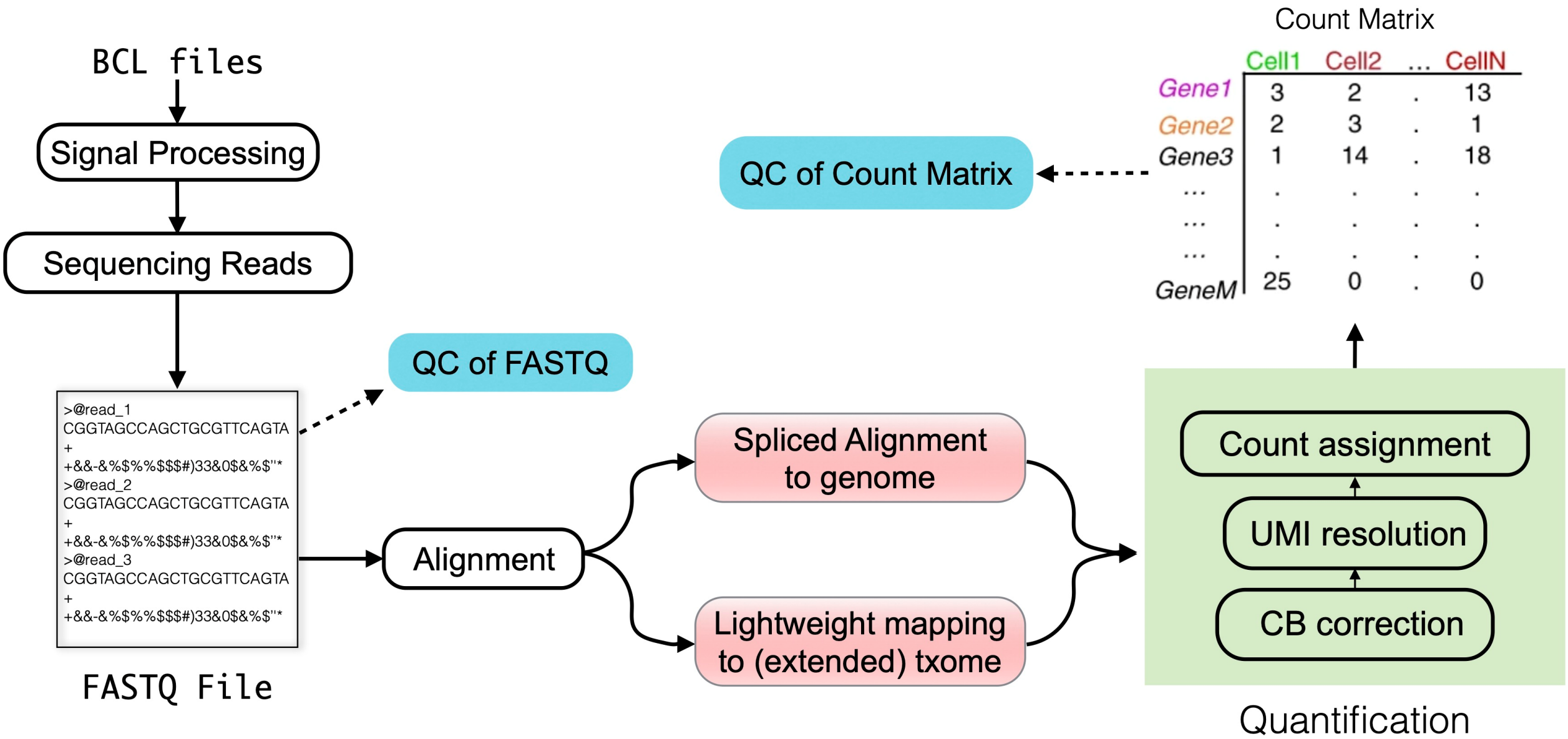
12345 tools, 32 categories

(202310001)



https://www.scrna-tools.org/

从FASTQ到表达矩阵



原始数据格式（FASTQ）

```
@ST-E00126:128:HJFLHCCXX:2:1101:7405:1133 1:N:0:CTTGTA
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

第1行主要储存序列测序时的坐标等信息；

第2行就是测序得到的序列信息，一般用ATCGN来表示，其中N用于荧光信号干扰无法判断到底是哪个碱基时的代表符号；

第3行以“+”开始，可以储存一些附加信息，但目前的测序fastq文件这一行一般是空的。

第4行储存的是质量信息，与第2行的碱基序列是一一对应的，其中的每一个符号对应的ASCII值是经过换算的phred值，可以简单理解为对应位置碱基的测序质量值，越大说明测序的质量越好。不同的版本对应的phred值范围不同。

原始数据文件形式

```
pbmc_1k_v3_fastqs
├── pbmc_1k_v3_S1_L001_I1_001.fastq.gz
├── pbmc_1k_v3_S1_L001_R1_001.fastq.gz
├── pbmc_1k_v3_S1_L001_R2_001.fastq.gz
├── pbmc_1k_v3_S1_L002_I1_001.fastq.gz
├── pbmc_1k_v3_S1_L002_R1_001.fastq.gz
└── pbmc_1k_v3_S1_L002_R2_001.fastq.gz
```

```
@A00228:279:HFWFVDMXX:1:1101:8486:1000 1:N:0:NCATTACT
NCATTACT
+
#FFFFFFF
@A00228:279:HFWFVDMXX:1:1101:10782:1000 1:N:0:NCATTACT
NCATTACT
+
#FFFFFFF
```

```
@A00228:279:HFWFVDMXX:1:1101:8486:1000 1:N:0:NCATTACT
NGTGATTAGCTGTACTCGTATGTAAGGT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFF
@A00228:279:HFWFVDMXX:1:1101:10782:1000 1:N:0:NCATTACT
NTCATGAAGTTTGGCTAGTTATGTTCAT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFF
```

```
@A00228:279:HFWFVDMXX:1:1101:8486:1000 2:N:0:NCATTACT
NACAAAGTCCCCCCCATAATACAGGGGGAGCCACTTGGGCAGGAGGCAGGGAGGGGTCCATTCCCCCTGGTGGGGCTGGTGGGGAGCTGTA
+
#FFFFFFFFFFFFFFF:FFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFF
@A00228:279:HFWFVDMXX:1:1101:10782:1000 2:N:0:NCATTACT
NTTGCAGCTGAACTGGTAAACTTGTCCCTAAAGAGACATAAGAATGGTCAACTGGAATGTGGATTCATCTGTAACATTACTCAGTGGGCCT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

原始数据处理——质控（FASTQC）

- 运行命令

```
fastqc -t 8 -o path/fastqc
sample1_R1.fq sample1_R2.fq
```

- 参数

-o --outdir:输出路径

--extract：结果文件解压缩

--noextract：结果文件压缩

-f --format:输入文件格式

-t --threads:线程数

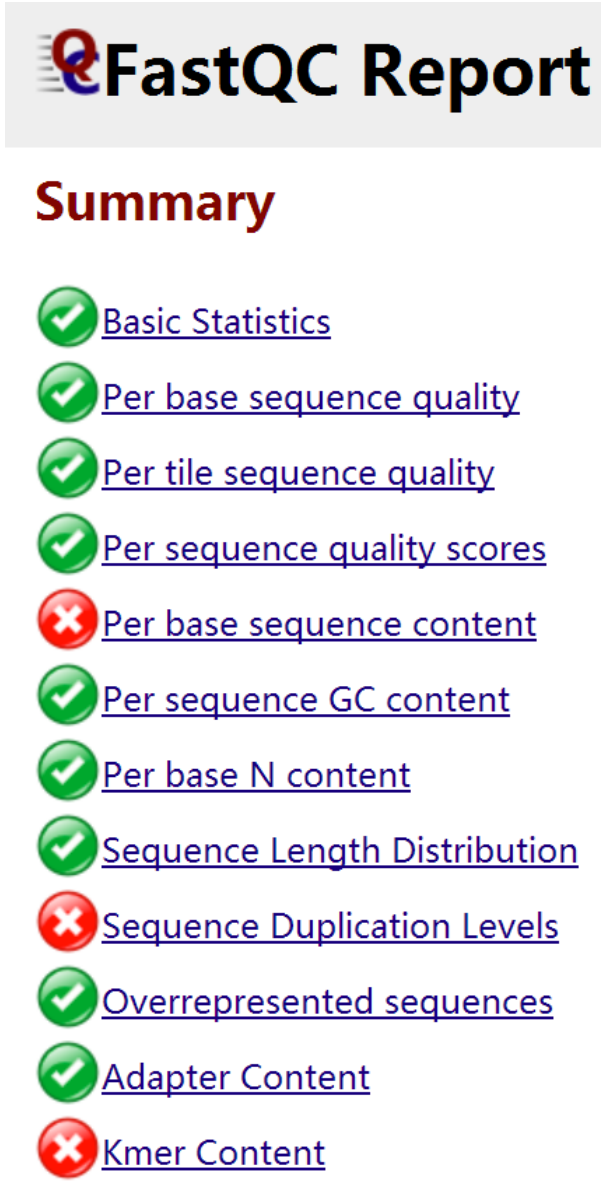
-c --contaminants：制定污染序列

-a --adapters：指定接头序列

-k --kmers：指定kmers长度（2-10bp,默认7bp）

-q --quiet： 安静模式

- 运行结果（html和zip）

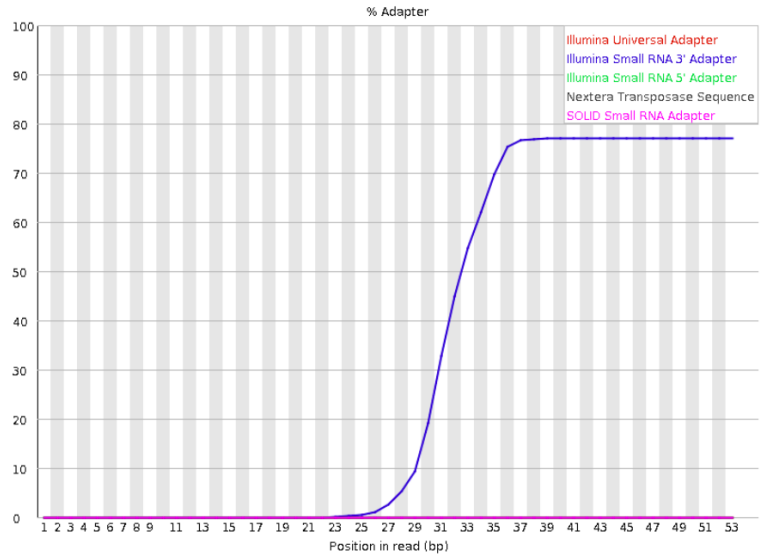


原始数据处理——FASTQC结果解读

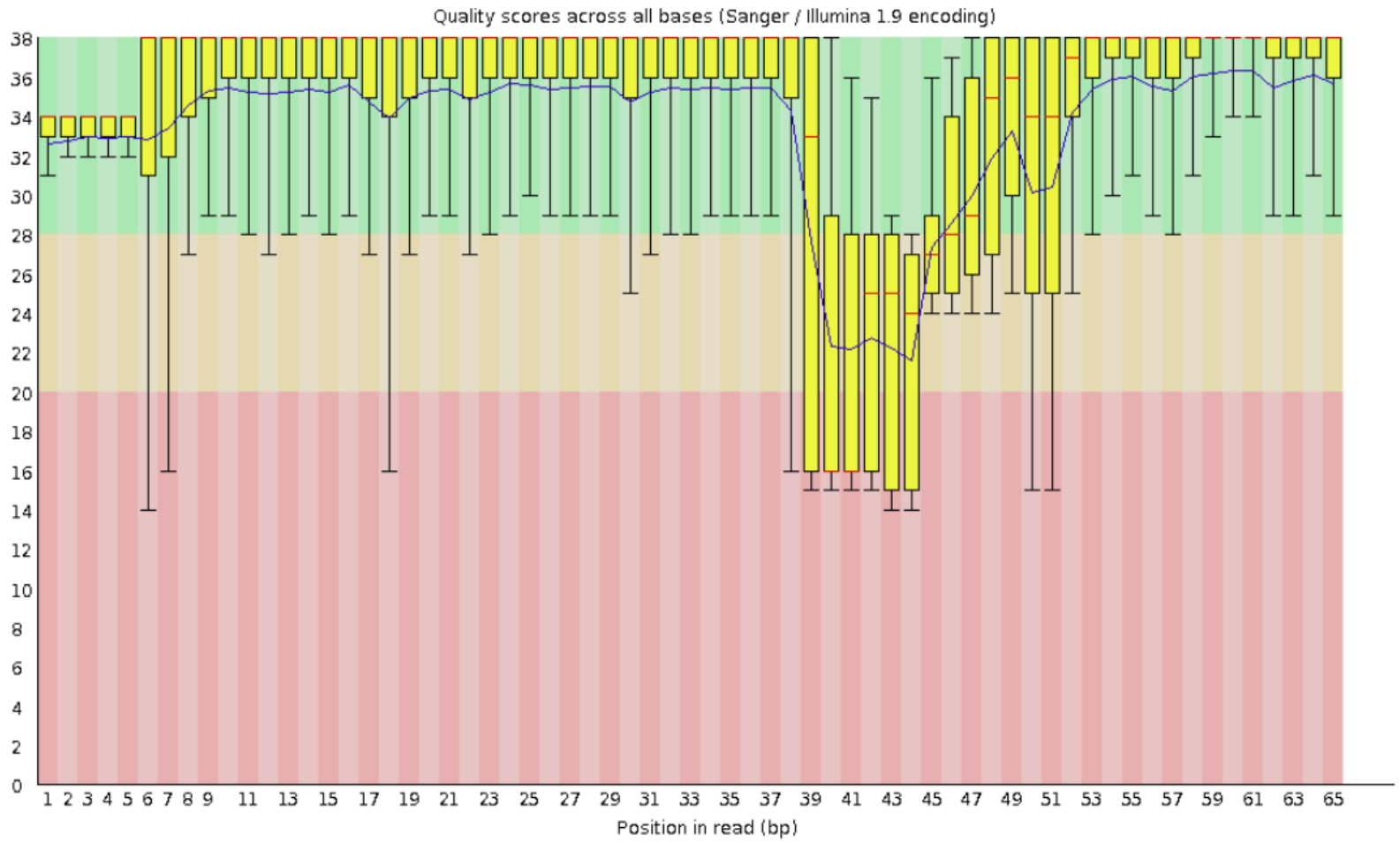
Basic Statistics

Measure	Value
Filename	SRR5345622.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	47613077
Sequences flagged as poor quality	0
Sequence length	65
%GC	64

Adapter Content

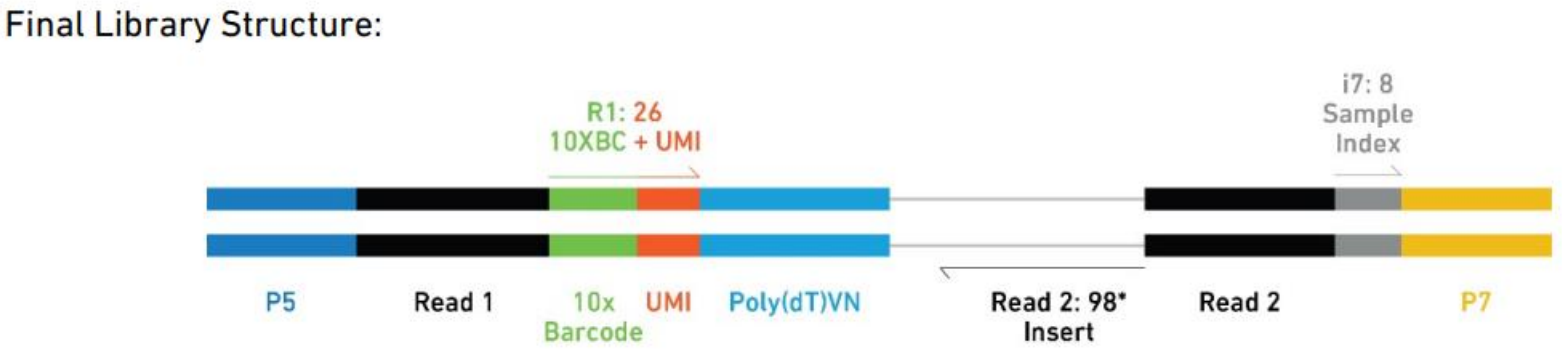


Per base sequence quality

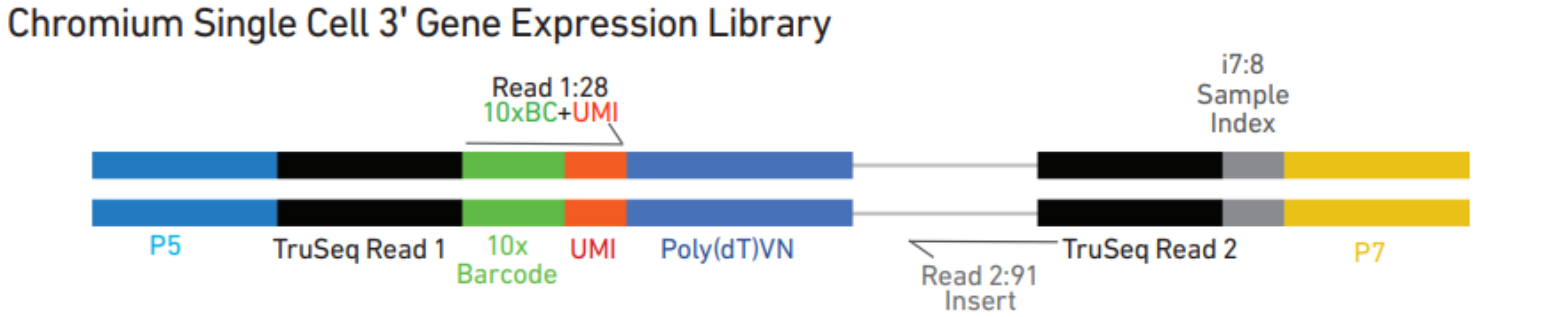


原始数据处理——不同测序平台的BC和UMI

10X V2



10X V3



BD Rhapsody

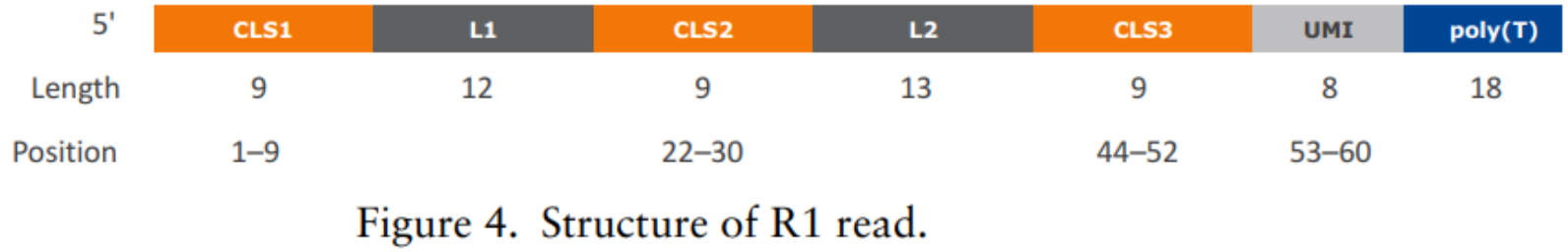
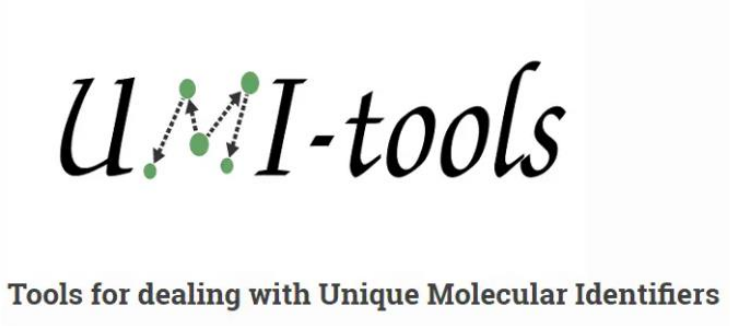


Figure 4. Structure of R1 read.

原始数据处理——分细胞（UMI-tools）



Step 1: get data

Step 2: Identify correct cell barcodes

Step 3: Extract barcdoes and UMIs and add to read names

Step 4: Map reads

Step 5: Assign reads to genes

Step 6: Count UMIs per gene per cell

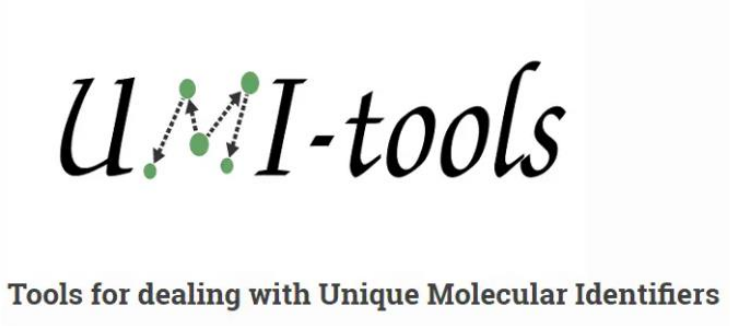
R1

```
@A00228:279:HFWFVDMXX:1:1101:8486:1000 1:N:0:NCATTACT
NGTGATTAGCTGTACTCGTATGTAAGGT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFF
@A00228:279:HFWFVDMXX:1:1101:10782:1000 1:N:0:NCATTACT
NTCATGAAGTTTGGCTAGTTATGTTCAT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFF
```

R2

```
@A00228:279:HFWFVDMXX:1:1101:8486:1000 2:N:0:NCATTACT
NACAAAGTCCCCCCCATAATACAGGGGGAGCCACTTGGGCAGGAGGCAGGGAGGGGTCCATTC
CCCCTGGTGGGGCTGGTGGGGAGCTGTA
+
#FFFFFFFFFFFFFFF:FFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFF
FFFFF:FFFFFFFFFFFFFFFFFFFF
@A00228:279:HFWFVDMXX:1:1101:10782:1000 2:N:0:NCATTACT
NTTGCAGCTGAACTGGTAAACTTGTCCCTAAAGAGACATAAGAATGGTCAACTGGAATGTGGA
TTCATCTGTAACATTACTCAGTGGGCCT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

原始数据处理——分细胞（UMI-tools）



R1

```
@A00228:279:HFWFVDMXX:1:1101:8486:1000 1:N:0:NCATTACT
NGTGATTAGCTGTACTCGTATGTAAGGT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFF
@A00228:279:HFWFVDMXX:1:1101:10782:1000 1:N:0:NCATTACT
NTCATGAAGTTTGGCTAGTTATGTTCAT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFF
```

Step 1: get data

Step 2: Identify correct cell barcodes

Step 3: Extract barcdoes and UMIs and add to read names

Step 4: Map reads

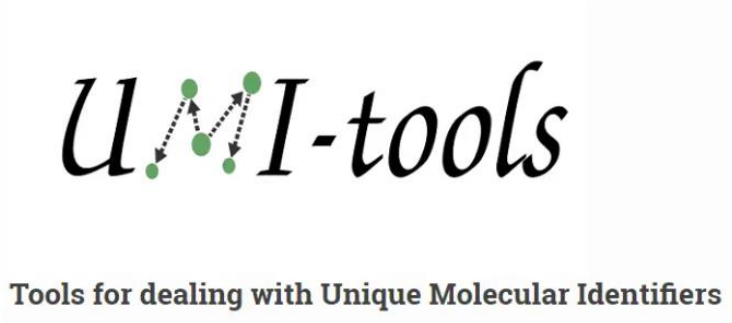
Step 5: Assign reads to genes

Step 6: Count UMIs per gene per cell

whitelist.txt

```
AAACCCAAGGAGAGTA      AAAACCAAGGAGAGTA,AAACACAAGGAGAGTA,…
           50531      51,67,…
AAACGCTTCAGCCCAG      AAAAGCTTCAGCCCAG,AAACACTTCAGCCCAG,…
           46071      47,11,…
AAAGAACAGACGACTG      AAACAACAGACGACTG,AAAGAAAAGACGACTG,…
           33466      1,28,…
AAAGAACCAATGGCAG      AAAAAACCAATGGCAG,AAAGAAACAATGGCAG,…
           21680      1,28,…
AAAGAACGTCTGCAAT      AAAAAACGTCTGCAAT,AAACAACGTCTGCAAT,…
           46538      2,1,…
AAAGGATAGTAGACAT      AAAAGATAGTAGACAT,AAACGATAGTAGACAT,…
           56493      2,1…
```

原始数据处理——分细胞（UMI-tools）



Step 1: get data

Step 2: Identify correct cell barcodes

Step 3: Extract barcdoes and UMIs and add to read names

Step 4: Map reads

Step 5: Assign reads to genes

Step 6: Count UMIs per gene per cell

R1_extracted

```
@A00228:279:HFWFVDMXX:1:1101:8486:1000_NGTGATTAGCTGTACT_CGTATGT
AAGGT 1:N:0:NCATTACT

+

@A00228:279:HFWFVDMXX:1:1101:10782:1000_NTCATGAAGTTTGGCT_AGTTAT
GTTCAT 1:N:0:NCATTACT

+
```

R2_extracted

```
@A00228:279:HFWFVDMXX:1:1101:8486:1000_NGTGATTAGCTGTACT_CGTATGT
AAGGT 2:N:0:NCATTACT
NACAAAGTCCCCCCCATAATACAGGGGGAGCCACTTGGGCAGGAGGCAGGGAGGGGTCCATTC
CCCCTGGTGGGGCTGGTGGGGAGCTGTA
+
#FFFFFFFFFFFFFFF:FFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFF
FFFFF:FFFFFFFFFFFFFFFFFFFFF
@A00228:279:HFWFVDMXX:1:1101:10782:1000_NTCATGAAGTTTGGCT_AGTTAT
GTTCAT 2:N:0:NCATTACT
NTTGCAGCTGAACTGGTAAACTTGTCCCTAAAGAGACATAAGAATGGTCAACTGGAATGTGGA
TTCATCTGTAACATTACTCAGTGGGCCT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

原始数据处理——比对（STAR）

Alexander Dobin
alexdobin

STAR: ultrafast universal RNA-seq aligner

Alexander Dobin[1,*], Carrie A. Davis[1], Felix Schlesinger[1], Jorg Drenkow[1], Chris Zaleski[1], Sonali Jha[1], Philippe Batut[1], Mark Chaisson[2] and Thomas R. Gingeras[1]
[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA and [2]Pacific Biosciences, Menlo Park, CA, USA
Associate Editor: Inanc Birol

- STAR比对结果

```
STAR_res/
├── Aligned.sortedByCoord.out.bam
├── Log.final.out
├── Log.out
├── Log.progress.out
├── SJ.out.tab
└── _STARtmp
    └── BAMsort
        ├── 0
        ├── 1
        ├── 2
        └── 3
```

Step 1: get data

Step 2: Identify correct cell barcodes

Step 3: Extract barcdoes and UMIs and add to read na

Step 4: Map reads

Step 5: Assign reads to genes

Step 6: Count UMIs per gene per cell

```
A00228:279:HFWFVDMXX:1:1102:1108:22388_CTCAAGAGTCAAAGAT_TTTTGTCAATAG    0       1       14473   255     85M583N6M       *       0       0       GGCTGGGTGGAGCCGTCCCC
CCATGGAGCACAGGCAGACAAAAGTCCCCGCCCCAGCTGTGTGGCCTCAAGCCAGCCTGCGCCACTGTGTT FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFN
H:i:1   HI:i:1  AS:i:79 nM:i:5
A00228:279:HFWFVDMXX:1:1109:28619:29731_CCTGCATAGTGGTGAC_CGGGGCCGACTC   16      1       14666   255     91M     *       0       0       GGTCTGGGGGGGAAGGTGTCATGGAGCC
CCCTACGATTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGA FFF,:FFFFFFFFFFFFF:FFFFFFF,FFFFFFFFF,F,FFFFFFF,FFFFF,,FFFFFFFFFFFFFFFFFFF,FFFF,FFFFFFFFFFF     NH:i
:1      HI:i:1  AS:i:89 nM:i:0
A00228:279:HFWFVDMXX:2:2311:8865:19852_GAGTTGTAGTGCGTCC_TTGCTGAGAAAC    16      1       14666   255     91M     *       0       0       GGTCGGGGGGGGAAGGTGTCATGGAGCC
CCCTACGATTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGA FFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFF     NH:i
:1      HI:i:1  AS:i:87 nM:i:1
A00228:279:HFWFVDMXX:2:2252:14064:2675_CCATCACCATTAAAGG_AGTATAGGGGCT    16      1       14666   255     91M     *       0       0       GGTCTGGGGGGGAAGGTGTCATGGAGCC
CCCTACGATTCCCTGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGA FFFFFF,FFF,F,FFFFFFFFFFFFFFFFF::FF,FFFFFF,FFFF:FFFFFFF:F:FFFFFFFFF,FFFF:FFFFFFFFFFFFFFFFF:     NH:i
:1      HI:i:1  AS:i:87 nM:i:1
A00228:279:HFWFVDMXX:2:1359:32226:23202_TGAGCGCCAGCATACT_GATACGTTTAGT   16      1       14666   255     91M     *       0       0       GGTCTGGGGGGGAAGGTGTCATGGAGCC
CCCTACGAGTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGATGCTGCGGTGGCGGCAGAGGA FFFF,FFFFFFF,FFF:,F:,FFF,FFFFFFF:FF:,FFFF:FF:FFF,FFFFFFFF,FF,F,FFF:FF,FF,FFFF:FFFF::FFFFFFF     NH:i
:1      HI:i:1  AS:i:85 nM:i:2
A00228:279:HFWFVDMXX:2:2402:17897:14669_TCATCATAGCACACAG_CACCTCTAACAG   16      1       14669   255     91M     *       0       0       CTGGGGGGGAAGGTGTCATGGAGCCCCC
TACGATTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGAGGG FFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF     NH:i
:1      HI:i:1  AS:i:89 nM:i:0
A00228:279:HFWFVDMXX:2:2402:18069:14810_TCATCATAGCACACAG_CACCTCTAACAG   16      1       14669   255     91M     *       0       0       CTGGGGGGGAAGGTGTCATGGAGCCCCC
TACGATTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGAGGG FFFFFFFFFF:FFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF     NH:i
:1      HI:i:1  AS:i:89 nM:i:0
A00228:279:HFWFVDMXX:1:1469:6587:29496_CGATGGCCACTCGATA_CGCTCACGCGAC    16      1       14670   255     91M     *       0       0       TGGGGGGGAAGGTGTCATGGAGCCCCCT
ACGATTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGAGGGA FFFFFFFFFFFFFFFFFFFFFF:FFF,FFFFFF,FFFFFFFFFFFFFFFFFFFFFFF:F:FFFFFFFFFFFFF,FFFFF:FFFFF,FF,FFF     NH:i
:1      HI:i:1  AS:i:89 nM:i:0
A00228:279:HFWFVDMXX:2:1282:21495:2300_AACCATGAGCTCGAAG_CTTTTACTGGGG    16      1       14671   255     91M     *       0       0       GGGGGGGAAGGTGTCATGGAGCCCCCTA
CGATTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGAGGGAT FFFFFFFFFFFFFFFFF:FFFFFFFFFFFFF,FFFFF:FFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF     NH:i
:1      HI:i:1  AS:i:89 nM:i:0
A00228:279:HFWFVDMXX:2:1158:28646:15843_GAAATGAAGCTACTGT_CGGATAAAAGGG   16      1       14672   255     91M     *       0       0       GGGGGGAAGGTGTCATGGAGCCCCCTAC
GATTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGAGGGATG FF:FFFFFFFFFF:FFFFFFF,FFF,FFFF:,FFF:FFFFF:FF,,:FF,F::FFF:FFF,F,FF:::,FFFFFF,F,FFF:FFFF:FF:F     NH:i
:1      HI:i:1  AS:i:89 nM:i:0
```

原始数据处理——比对（STAR）

- STAR比对结果文件（bam/sam）

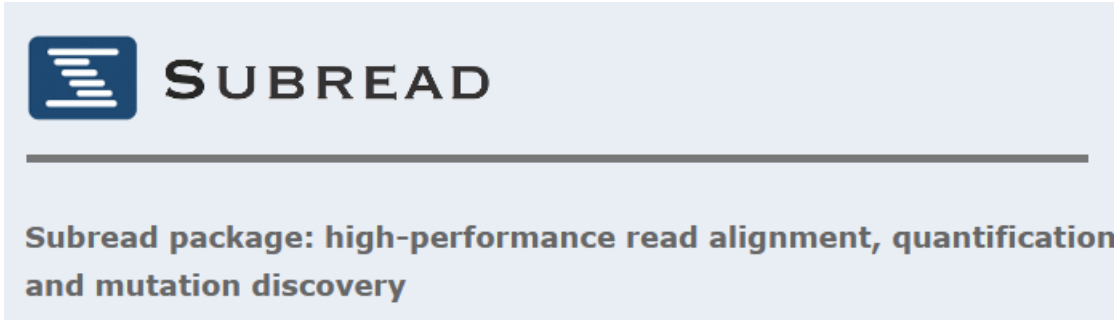
```
A00228:279:HFWFVDMXX:1:1102:1108:22388_CTCAAGAGTCAAAGAT_TTTTGTCAATAG
        0       1       14473   255     85M583N6M       *
        0       0
        GGCTGGGTGGAGCCGTCCCCCCATGGAGCACAGGCAGACAAAAGTCCCCGCCCCAGCTG
TGTGGCCTCAAGCCAGCCTGCGCCACTGTGTT
        FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,F
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFNH:i:1     HI:i:1     AS:i:79
        nM:i:5
```

Col	Field	Brief Description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	References sequence NAME
4	POS	1- based leftmost mapping Position
5	MAPQ	Mapping Quality
6	CIGAR	CIGAR String
7	MRNM/RNEXT	Ref. name of the mate/next read
8	MPOS/NEXT	Position of the mate/next read
9	ISIZE/TLEN	observed Template LENgth
10	SEQ	segment SEQuence
11	QUAL	ASCII of Phred-scaled b
12	TAGs	TAGs

```
                                 Started job on |       Oct 12 10:07:39
                             Started mapping on |       Oct 12 10:36:00
                                    Finished on |       Oct 12 12:21:40
       Mapping speed, Million of reads per hour |       37.82

                           Number of input reads |       66601887
                       Average input read length |       91
                                    UNIQUE READS:
                   Uniquely mapped reads number |       58334733
                        Uniquely mapped reads % |       87.59%
                          Average mapped length |       89.27
                       Number of splices: Total |       9210314
            Number of splices: Annotated (sjdb) |       9023795
                       Number of splices: GT/AG |       9116393
                       Number of splices: GC/AG |       27095
                       Number of splices: AT/AC |       1644
               Number of splices: Non-canonical |       65182
                      Mismatch rate per base, % |       0.54%
                         Deletion rate per base |       0.01%
                        Deletion average length |       1.55
                        Insertion rate per base |       0.02%
                       Insertion average length |       1.38
                            MULTI-MAPPING READS:
          Number of reads mapped to multiple loci |       0
               % of reads mapped to multiple loci |       0.00%
          Number of reads mapped to too many loci |       5401311
               % of reads mapped to too many loci |       8.11%
                                UNMAPPED READS:
Number of reads unmapped: too many mismatches |       0
     % of reads unmapped: too many mismatches |       0.00%
          Number of reads unmapped: too short |       2702824
               % of reads unmapped: too short |       4.06%
            Number of reads unmapped: other |       163019
                 % of reads unmapped: other |       0.24%
                              CHIMERIC READS:
                   Number of chimeric reads |       0
                        % of chimeric reads |       0.00%
```

原始数据处理——定量（featurecounts）

```
//========================== featureCounts setting ===========================\\
||                                                                            ||
||             Input files : 1 BAM file                                       ||
||                           o Aligned.sortedByCoord.out.bam                  ||
||                                                                            ||
||             Output file : gene_assigned                                    ||
||                 Summary : gene_assigned.summary                            ||
||              Annotation : Homo_sapiens.GRCh38.110.gtf (GTF)                ||
||      Dir for temp files : ./                                               ||
||      Assignment details : <input_file>.featureCounts.bam                   ||
||                           (Note that files are saved to the output directory)||
||                                                                            ||
```



Step 1: get data

Step 2: Identify correct cell barcodes

Step 3: Extract barcdoes and UMIs and add to read names

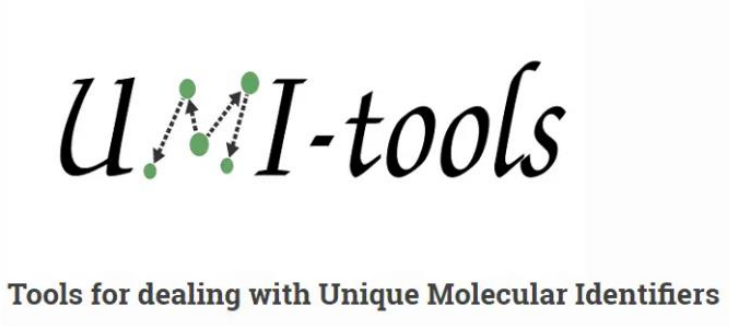
Step 4: Map reads

Step 5: Assign reads to genes

Step 6: Count UMIs per gene per cell

输出文件格式	举例
Geneid	ENSG00000269896
Chr	1;1
Start	2350414;2351644
End	2352820;2351857
Strand	-;-
Length	2407
Aligned.sortedByCoord.out.bam	8

原始数据处理——定量



Step 1: get data

Step 2: Identify correct cell barcodes

Step 3: Extract barcdoes and UMIs and add to read names

Step 4: Map reads

Step 5: Assign reads to genes

Step 6: Count UMIs per gene per cell

输出文件格式：

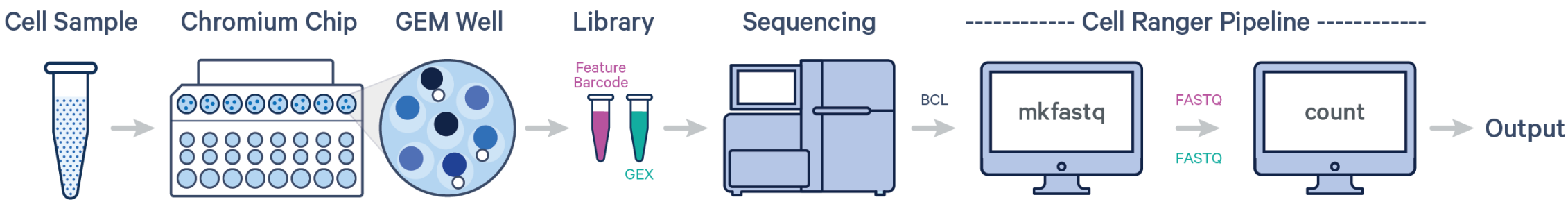
- the gene_id, the cell barcode and the count of deduplicated UMIs

gene	cell	count
ENSG00000000003	ACGTAACAGGTGCTTT	1
ENSG00000000003	CTGGCAGTCGTCTCAC	1
ENSG00000000003	GTCAAACCAAGCACCC	1
ENSG00000000003	TTACGTTTCTCGCTTG	1
ENSG00000000419	AAAGAACGTCTGCAAT	1
ENSG00000000419	AAAGGTACAACTGCTA	2
ENSG00000000419	AAAGTCCCACCAGCCA	1
ENSG00000000419	AAATGGAGTACCGCGT	1
ENSG00000000419	AACAAAGGTGATGAAT	1
ENSG00000000419	AACCTGACATCCTATT	1
ENSG00000000419	AAGAACAAGCCTCAGC	1
ENSG00000000419	AAGCGTTCACTGATTG	1
ENSG00000000419	AAGTACCAGCGCCTTG	1
ENSG00000000419	AAGTACCCAAAGAACT	2
ENSG00000000419	AAGTTCGAGGATACAT	1
ENSG00000000419	AAGTTCGGTCAACACT	1
ENSG00000000419	AATCACACACCCTGTT	1
ENSG00000000419	AATCGACAGTATGTAG	1
ENSG00000000419	AATCGACGTGAGACGT	1
ENSG00000000419	AATGACCGTGTCATCA	4
ENSG00000000419	ACAACCATCTGCCCTA	2
ENSG00000000419	ACACGCGGTGTTGCCG	1
ENSG00000000419	ACATCAGTCGGTGTCG	1
ENSG00000000419	ACGATCAGTCGTTCAA	2

原始数据处理——CellRanger获取表达矩阵

- mkfastq：可直接分析从测序之后扫描剪辑信号的BCL文件，生产每个样本的 FASTQ 文件用于后续分析
- count：获取 FASTQ 文件并执行比对、过滤、barcode和UMI计数，分群和表达差异分析。
- aggr：用于多样本的整合
- multi：用于混合样本的分析

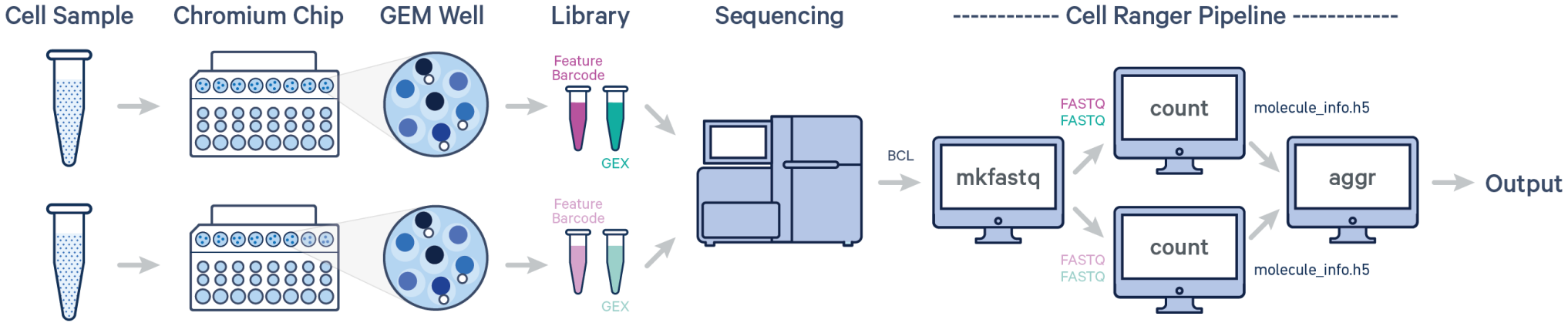
- 单个样本分析



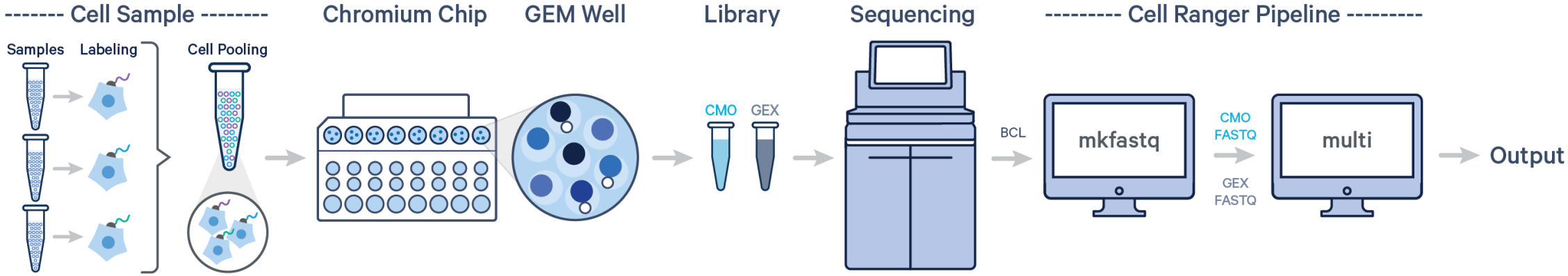
https://www.10xgenomics.com

原始数据处理——CellRanger获取表达矩阵

- CellRanger多样本分析：aggr用于多样本的整合



- multi 用于混合样本的分析



https://www.10xgenomics.com

原始数据处理——CellRanger具体使用（1）

（1）软件下载

Cell Ranger 7.2.0 (Sep 13, 2023) 直接解压即可使用

链接：https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest

```
/public/home/chuqj/software/cellranger-7.2.0
├── bin
│   ├── cellranger
│   ├── _cellranger_internal
│   ├── rna
│   ├── sc_rna
│   └── tenkit
├── builtwith.json
├── cellranger -> bin/cellranger
├── external
│   ├── anaconda
│   ├── cellranger_tiny_fastq
│   ├── cellranger_tiny_ref
│   ├── martian
│   └── tenx_feature_references
├── lib
│   ├── bin
│   ├── python
│   └── rust
├── LICENSE
├── mro
│   ├── rna
│   └── tenkit
├── probe_sets -> external/tenx_feature_references/targeted_panels
├── sourceme.bash
├── sourceme.csh
└── THIRD-PARTY-LICENSES.cellranger.txt
```

```
cellranger cellranger-7.2.0

Process 10x Genomics Gene Expression, Feature Barcode, and Immune Profiling data

Usage: cellranger <COMMAND>

Commands:
  count           Count gene expression and/or feature barcode reads from a single sample
                      and GEM well
  multi           Analyze multiplexed data or combined gene expression/immune
                      profiling/feature barcode data
  multi-template  Output a multi config CSV template
  vdj             Assembles single-cell VDJ receptor sequences from 10x Immune Profiling
                      libraries
  aggr            Aggregate data from multiple Cell Ranger runs
  reanalyze       Re-run secondary analysis (dimensionality reduction, clustering, etc)
  mkvdjref        Prepare a reference for use with CellRanger VDJ
  mkfastq         Run Illumina demultiplexer on sample sheets that contain 10x-specific
                      sample index sets
  testrun         Execute the 'count' pipeline on a small test dataset
  mat2csv         Convert a gene count matrix to CSV format
  mkref           Prepare a reference for use with 10x analysis software. Requires a GTF
                      and FASTA
  mkgtf           Filter a GTF file by attribute prior to creating a 10x reference
  upload          Upload analysis logs to 10x Genomics support
  sitecheck       Collect linux system configuration information
  help            Print this message or the help of the given subcommand(s)

Options:
  -h, --help     Print help
  -V, --version  Print version
```

原始数据处理——CellRanger具体使用（2）

（2）基因组建库

- 直接下载
  （https://www.10xgenomics.com/support/software/cell-ranger/downloads）

- 自己创建：下载基因组序列文件（FASTA）和注释文件（GTF）
  - http://ftp.ensembl.org/pub/release-110/
  - https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-57/

```
Homo_sapiens_GRCh38/
├── fasta
│   ├── genome.fa
│   └── genome.fa.fai
├── genes
│   └── genes.gtf.gz
├── reference.json
└── star
    ├── chrLength.txt
    ├── chrNameLength.txt
    ├── chrName.txt
    ├── chrStart.txt
    ├── exonGeTrInfo.tab
    ├── exonInfo.tab
    ├── geneInfo.tab
    ├── Genome
    ├── genomeParameters.txt
    ├── SA
    ├── SAindex
    ├── sjdbInfo.txt
    ├── sjdbList.fromGTF.out.tab
    ├── sjdbList.out.tab
    └── transcriptInfo.tab
```

原始数据处理——CellRanger具体使用（3）

（3）count获得表达矩阵

- 文件名的格式：[Sample Name]_S1_L00[Lane Number]_[Read Type]_001.fastq.gz

Read Type:
I1: Dual index i7 read (optional)
I2: Dual index i5 read (optional)
R1: Read 1
R2: Read 2

```
pbmc_1k_v3_S1_L001_I1_001.fastq.gz
pbmc_1k_v3_S1_L001_R1_001.fastq.gz
pbmc_1k_v3_S1_L001_R2_001.fastq.gz
pbmc_1k_v3_S1_L002_I1_001.fastq.gz
pbmc_1k_v3_S1_L002_R1_001.fastq.gz
pbmc_1k_v3_S1_L002_R2_001.fastq.gz
```

```
Usage: cellranger count [OPTIONS] --id <ID> --transcriptome <PATH>

Options:
      --id <ID>                                A unique run id and output folder name [a-zA-Z0-9_-]+
      --description <TEXT>                     Sample description to embed in output files [default: ]
      --transcriptome <PATH>                   Path of folder containing 10x-compatible transcriptome reference
      --fastqs <PATH>                          Path to input FASTQ data
      --project <TEXT>                         Name of the project folder within a mkfastq or bcl2fastq-generated folder from which to pick FASTQs
      --sample <PREFIX>                        Prefix of the filenames of FASTQs to select
      --lanes <NUMS>                           Only use FASTQs from selected lanes
      --libraries <CSV>                        CSV file declaring input library data sources
      --feature-ref <CSV>                      Feature reference CSV file, declaring Feature Barcode constructs and associated barcodes
      --expect-cells <NUM>                     Expected number of recovered cells, used as input to cell calling algorithm
      --force-cells <NUM>                      Force pipeline to use this number of cells, bypassing cell calling algorithm. [MINIMUM: 10]
      --no-bam                                 Set --no-bam to not generate the BAM file. This will reduce the total computation time for the pipestance and the size of the output directory. If unsure, we recommend not to use this option. BAM file
                                               could be useful for troubleshooting and downstream analysis
      --nosecondary                            Disable secondary analysis, e.g. clustering. Optional
      --r1-length <NUM>                        Hard trim the input Read 1 to this length before analysis
      --r2-length <NUM>                        Hard trim the input Read 2 to this length before analysis
      --include-introns <true|false>           Include intronic reads in count [default: true] [possible values: true, false]
      --chemistry <CHEM>                       Assay configuration. NOTE: by default the assay configuration is detected automatically, which is the recommended mode. You usually will not need to specify a chemistry. Options are: 'auto' for
                                               autodetection, 'threeprime' for Single Cell 3', 'fiveprime' for  Single Cell 5', 'SC3Pv1' or 'SC3Pv2' or 'SC3Pv3' for Single Cell 3' v1/v2/v3, 'SC3Pv3LT' for Single Cell 3' v3 LT, 'SC3Pv3HT' for Single
                                               Cell 3' v3 HT, 'SC5P-PE' or 'SC5P-R2' for Single Cell 5', paired-end/R2-only, 'SC-FB' for Single Cell Antibody-only 3' v2 or 5'. To analyze the GEX portion of multiome data, chemistry must be set to
                                               'ARC-v1'; 'ARC-v1' chemistry cannot be autodetected [default: auto]
      --no-libraries                           Proceed with processing using a --feature-ref but no Feature Barcode libraries specified with the 'libraries' flag
      --check-library-compatibility <true|false>  Whether to check for barcode compatibility between libraries. [default: true] [possible values: true, false]
      --dry                                    Do not execute the pipeline. Generate a pipeline invocation (.mro) file and stop
      --jobmode <MODE>                         Job manager to use. Valid options: local (default), sge, lsf, slurm or path to a .template file. Search for help on "Cluster Mode" at support.10xgenomics.com for more details on configuring the pipeline
                                               to use a compute cluster [default: local]
      --localcores <NUM>                       Set max cores the pipeline may request at one time. Only applies to local jobs
      --localmem <NUM>                         Set max GB the pipeline may request at one time. Only applies to local jobs
      --localvmem <NUM>                        Set max virtual address space in GB for the pipeline. Only applies to local jobs
      --mempercore <NUM>                       Reserve enough threads for each job to ensure enough memory will be available, assuming each core on your cluster has at least this much memory available. Only applies to cluster jobmodes
      --maxjobs <NUM>                          Set max jobs submitted to cluster at one time. Only applies to cluster jobmodes
      --jobinterval <NUM>                      Set delay between submitting jobs to cluster, in ms. Only applies to cluster jobmodes
      --overrides <PATH>                       The path to a JSON file that specifies stage-level overrides for cores and memory. Finer-grained than --localcores, --mempercore and --localmem. Consult https://support.10xgenomics.com/ for an example
                                               override file
      --output-dir <PATH>                      Output the results to this directory
      --uiport <PORT>                          Serve web UI at http://localhost:PORT
      --disable-ui                             Do not serve the web UI
      --noexit                                 Keep web UI running after pipestance completes or fails
      --nopreflight                            Skip preflight checks
  -h, --help                                   Print help
```

原始数据处理——CellRanger具体使用（3）

（3）count获得表达矩阵

```
├── _cmdline
├── _filelist
├── _finalstate
├── Homo_sapiens_GRCh38.mri.tgz
├── _invocation
├── _jobmode
├── _log
├── _mrosource
├── outs
│   ├── analysis
│   ├── cloupe.cloupe
│   ├── filtered_feature_bc_matrix
│   ├── filtered_feature_bc_matrix.h5
│   ├── metrics_summary.csv
│   ├── molecule_info.h5
│   ├── raw_feature_bc_matrix
│   ├── raw_feature_bc_matrix.h5
│   └── web_summary.html
├── _perf
├── SC_RNA_COUNTER_CS
│   ├── CELLRANGER_PREFLIGHT
│   ├── CELLRANGER_PREFLIGHT_LOCAL
│   ├── fork0
│   ├── FULL_COUNT_INPUTS
│   ├── GET_AGGREGATE_BARCODES_OUT
│   ├── SC_MULTI_CORE
│   ├── _STRUCTIFY
│   └── WRITE_GENE_INDEX
├── _sitecheck
├── _tags
├── _timestamp
├── _uuid
├── _vdrkill
└── _versions
```

barcodes.tsv.gz

```
AAACCCAAGGAGAGTA-1
AAACGCTTCAGCCCAG-1
AAAGAACAGACGACTG-1
AAAGAACCAATGGCAG-1
AAAGAACGTCTGCAAT-1
AAAGGATAGTAGACAT-1
AAAGGATCACCGGCTA-1
AAAGGATTCAGCTTGA-1
AAAGGATTCCGTTTCG-1
AAAGGGCTCATGCCCT-1
AAAGGGCTCCGTAGGC-1
AAAGGTACAACTGCTA-1
AAAGTCCAGCGGGTTA-1
AAAGTCCAGTCAACAA-1
AAAGTCCCACCAGCCA-1
AAAGTGATCGTACACA-1
AAATGGAAGCCGCTTG-1
AAATGGACAATGCTCA-1
AAATGGAGTACCGCGT-1
AAATGGATCCTATTTG-1
AACAAAGGTGATGAAT-1
AACAACCAGTAGTCCT-1
AACAACCCACGCTATA-1
AACAAGAGTTATAGAG-1
AACAGGGGTGGGAGAG-1
AACCAACAGCTTGTTG-1
AACCCAACAACTGATC-1
AACCCAAGTGGGCTTC-1
AACCCAATCTTACCGC-1
AACCTGACATCCTATT-1
```

features.tsv.gz

```
ENSG00000279928 DDX11L17        Gene Expression
ENSG00000228037 ENSG00000228037 Gene Expression
ENSG00000142611 PRDM16  Gene Expression
ENSG00000284616 ENSG00000284616 Gene Expression
ENSG00000157911 PEX10   Gene Expression
ENSG00000269896 ENSG00000269896 Gene Expression
ENSG00000228463 ENSG00000228463 Gene Expression
ENSG00000260972 ENSG00000260972 Gene Expression
ENSG00000224340 RPL21P21        Gene Expression
ENSG00000226374 LINC01345       Gene Expression
ENSG00000229280 EEF1DP6 Gene Expression
ENSG00000142655 PEX14   Gene Expression
ENSG00000232596 LINC01646       Gene Expression
ENSG00000235054 LINC01777       Gene Expression
ENSG00000231510 LINC02782       Gene Expression
ENSG00000149527 PLCH2   Gene Expression
ENSG00000284739 ENSG00000284739 Gene Expression
ENSG00000171621 SPSB1   Gene Expression
ENSG00000272235 ENSG00000272235 Gene Expression
ENSG00000284694 ENSG00000284694 Gene Expression
ENSG00000224387 ENSG00000224387 Gene Expression
ENSG00000142583 SLC2A5  Gene Expression
ENSG00000284674 LINC02781       Gene Expression
ENSG00000224338 MTCYBP45        Gene Expression
ENSG00000287727 ENSG00000287727 Gene Expression
ENSG00000286448 ENSG00000286448 Gene Expression
ENSG00000284703 ENSG00000284703 Gene Expression
ENSG00000226457 RPL22P3 Gene Expression
ENSG00000173614 NMNAT1  Gene Expression
ENSG00000215720 MFFP1   Gene Expression
```

matrix.tsv.gz

```
62754 1223 4016643
37 1 1
40 1 1
45 1 3
51 1 5
54 1 1
58 1 1
76 1 8
78 1 1
93 1 1
95 1 2
100 1 1
106 1 5
113 1 1
125 1 6
134 1 1
138 1 5
162 1 1
170 1 3
177 1 9
182 1 4
188 1 1
200 1 1
234 1 1
248 1 1
254 1 1
```

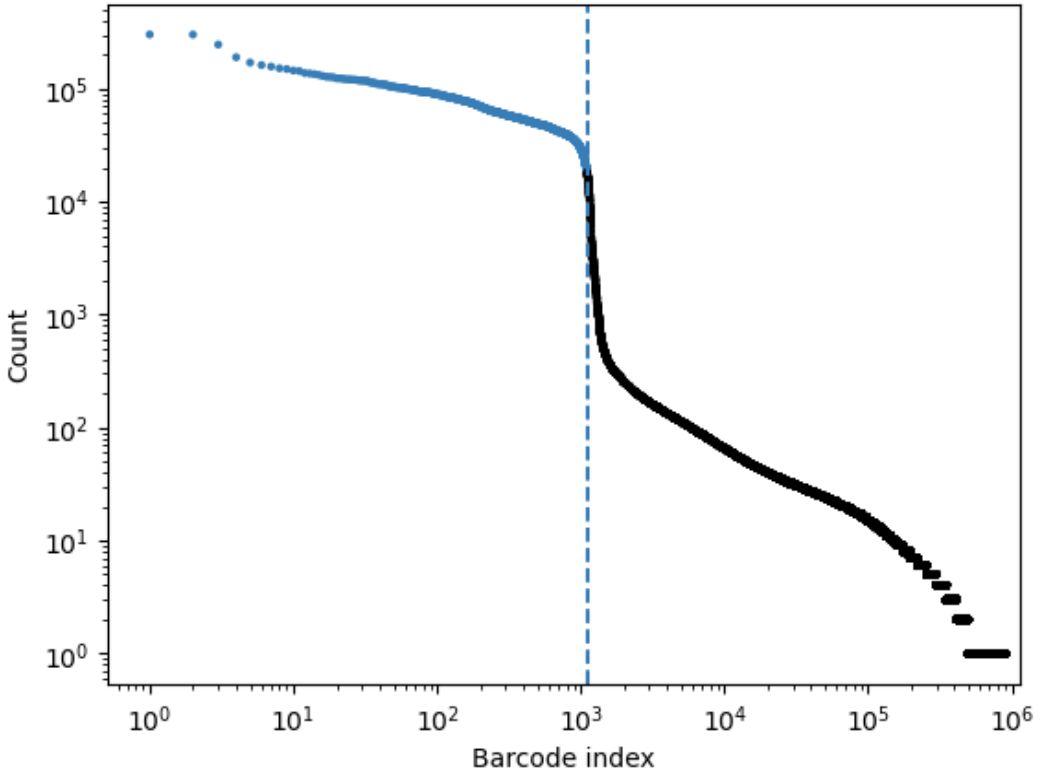
barcodes

features

UMI-tools+STAR+featurecounts VS CellRanger

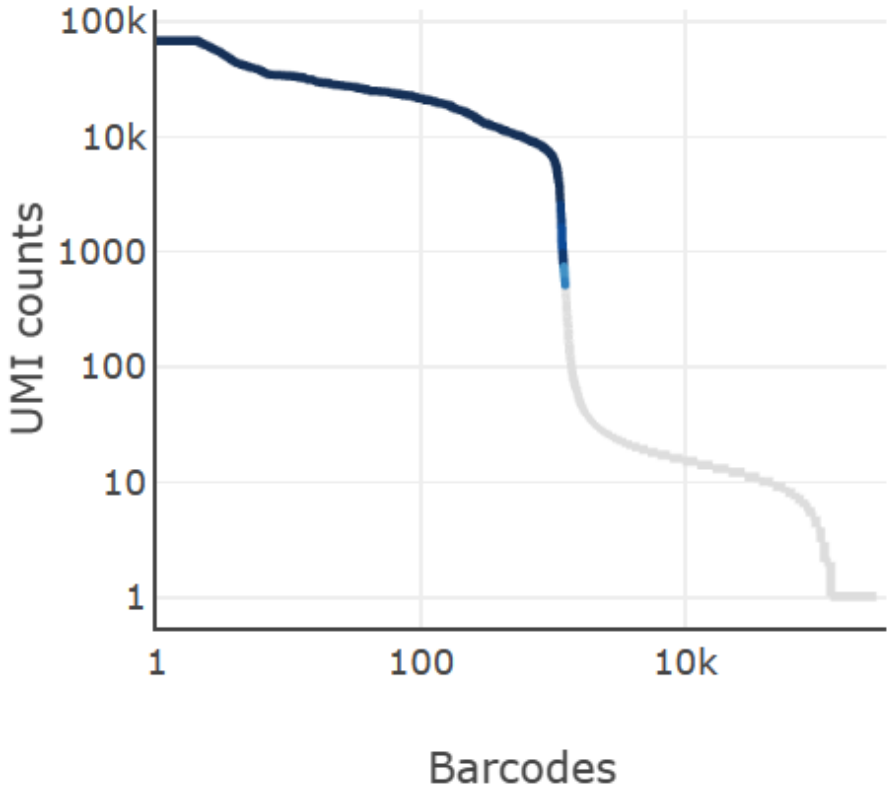
UMI-tools

- 1108 cells



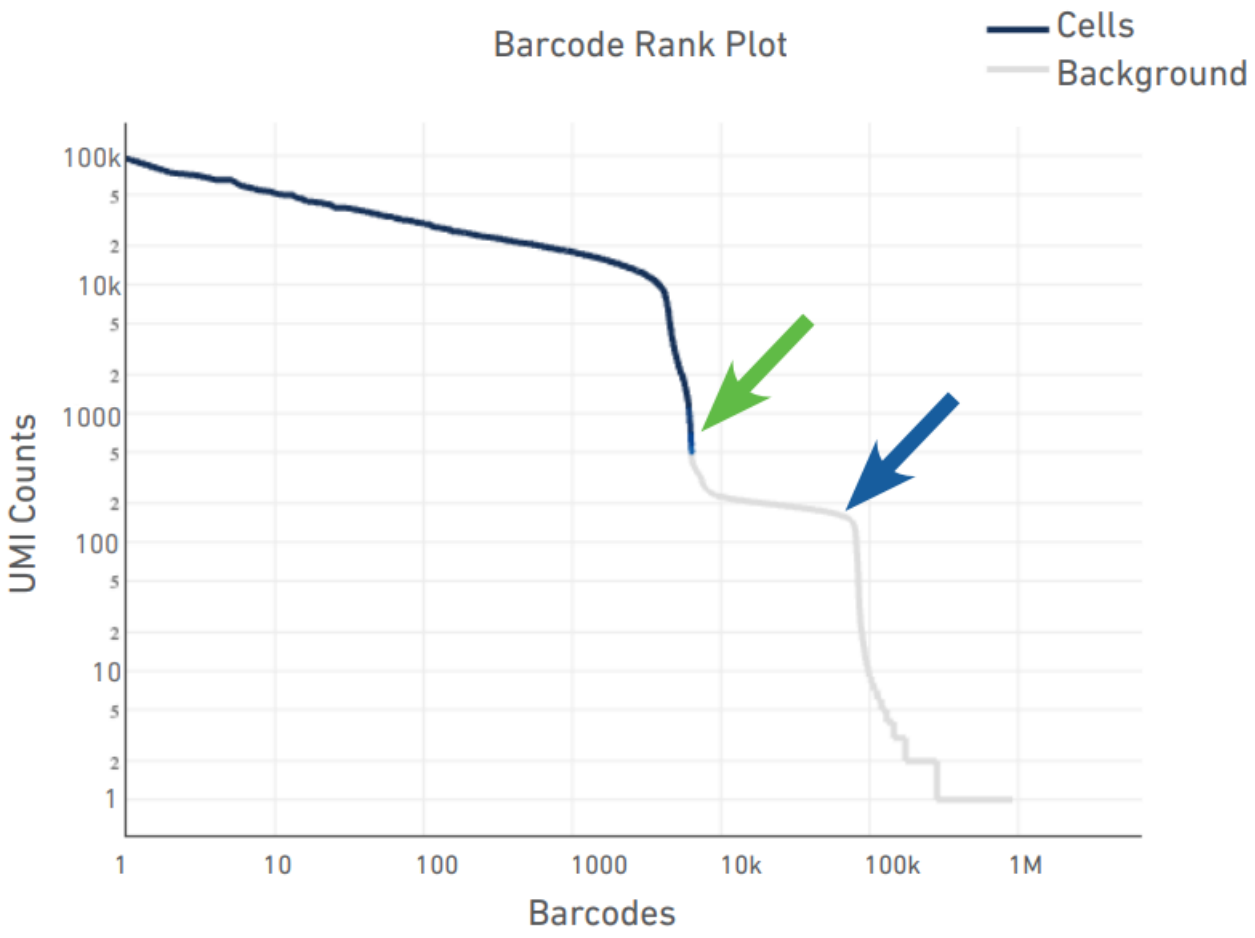
CellRanger

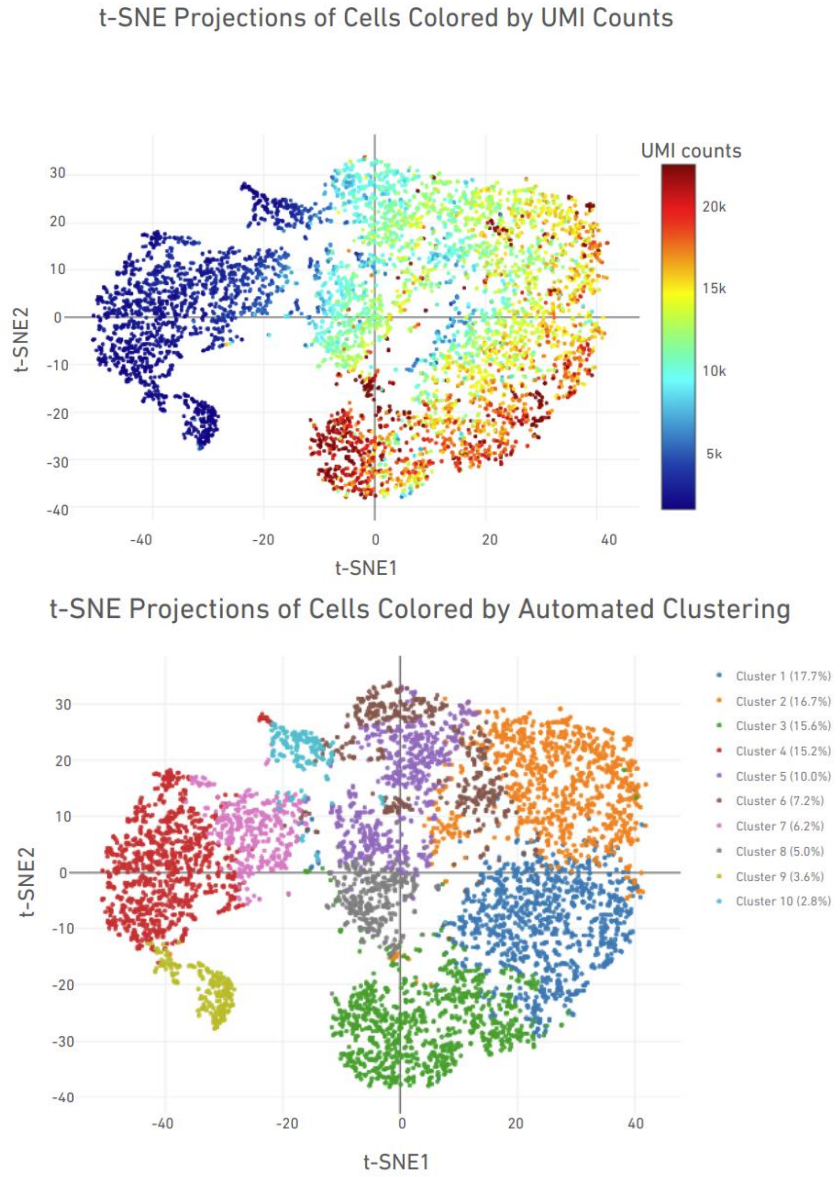
- 1223 cells



基因-细胞表达矩阵（案例1）

- UMI-barcode图判断测序质量和细胞数：典型案例
- 急剧下降表明细胞相关条形码和与空 GEM 相关的条形码之间的良好分离。

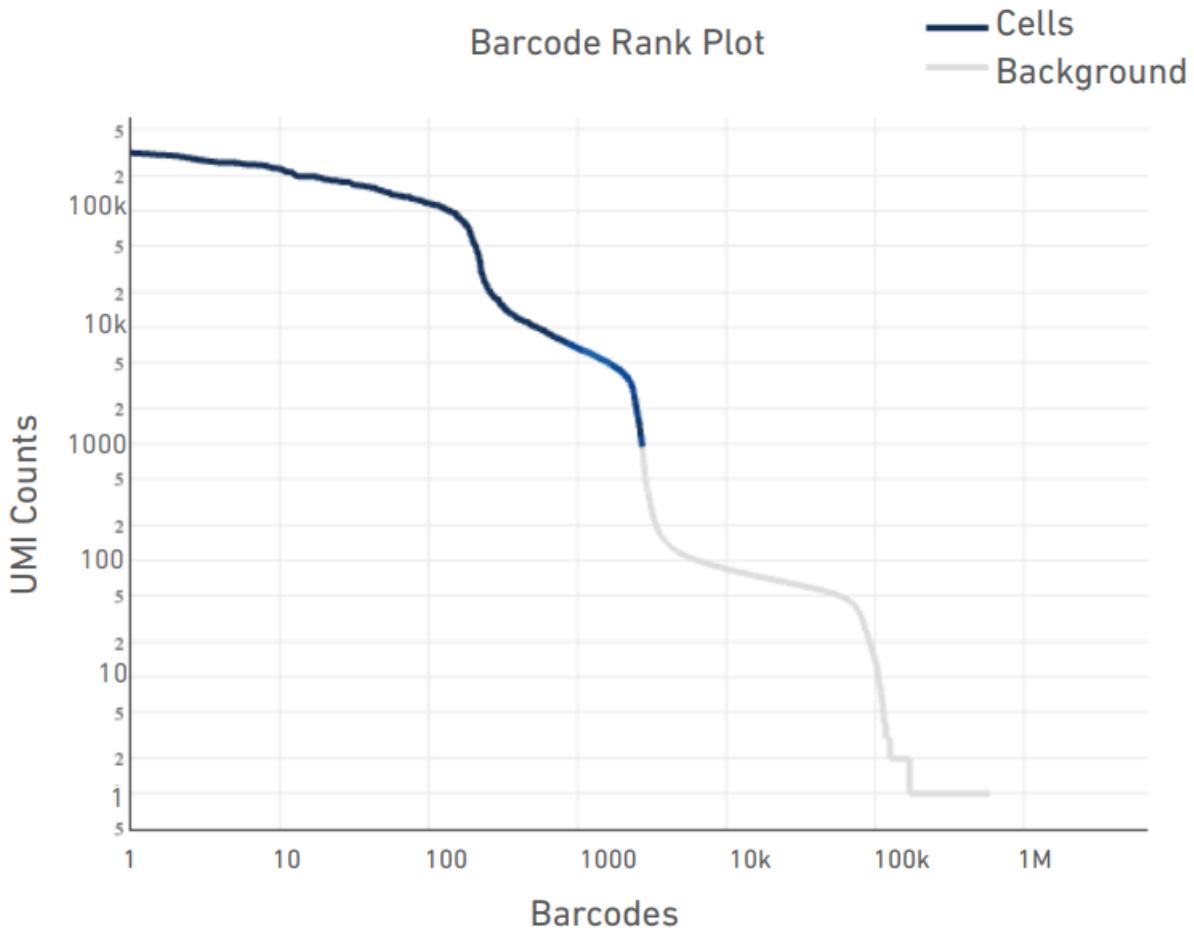


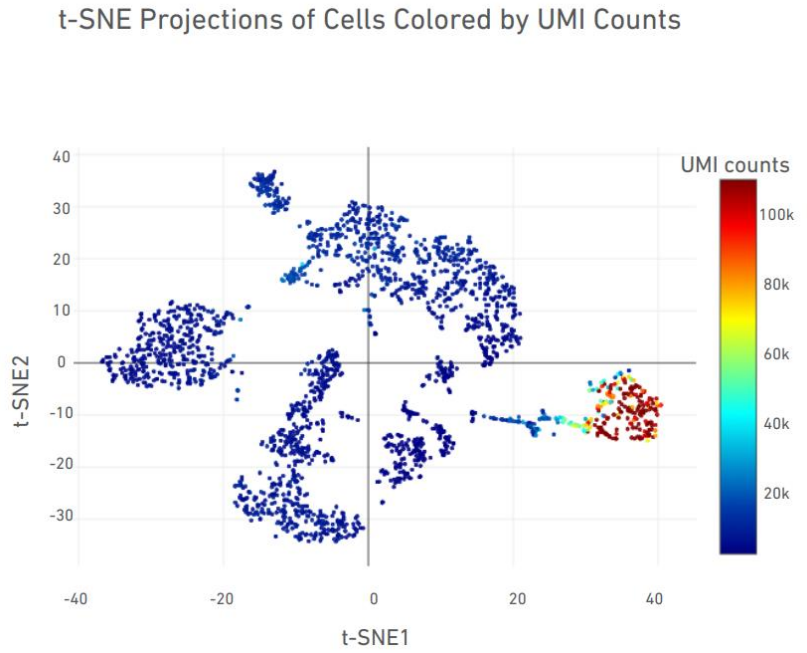


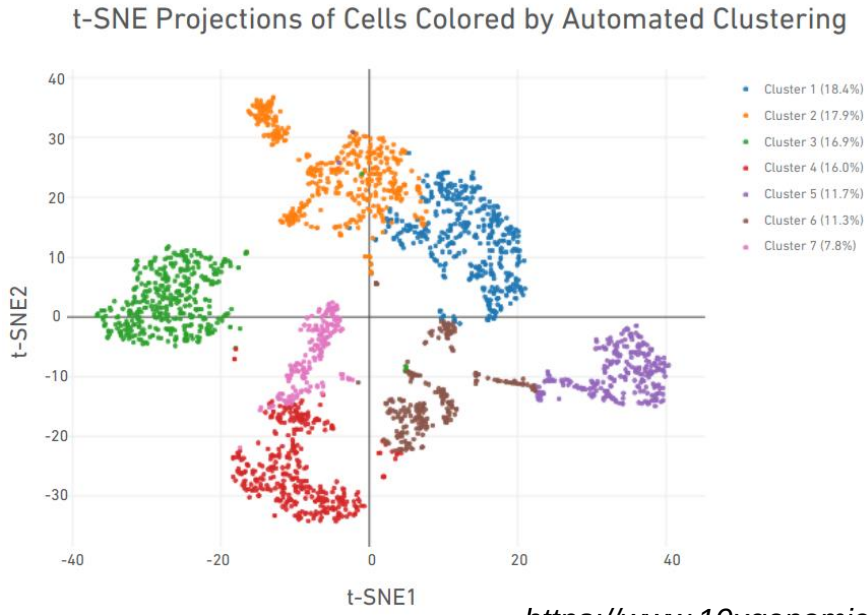
https://www.10xgenomics.com

基因-细胞表达矩阵（案例2）

- UMI-barcode图判断测序质量和细胞数：异质样本案例
- 样本中可能存在异质细胞群，这可能会导致双峰图。



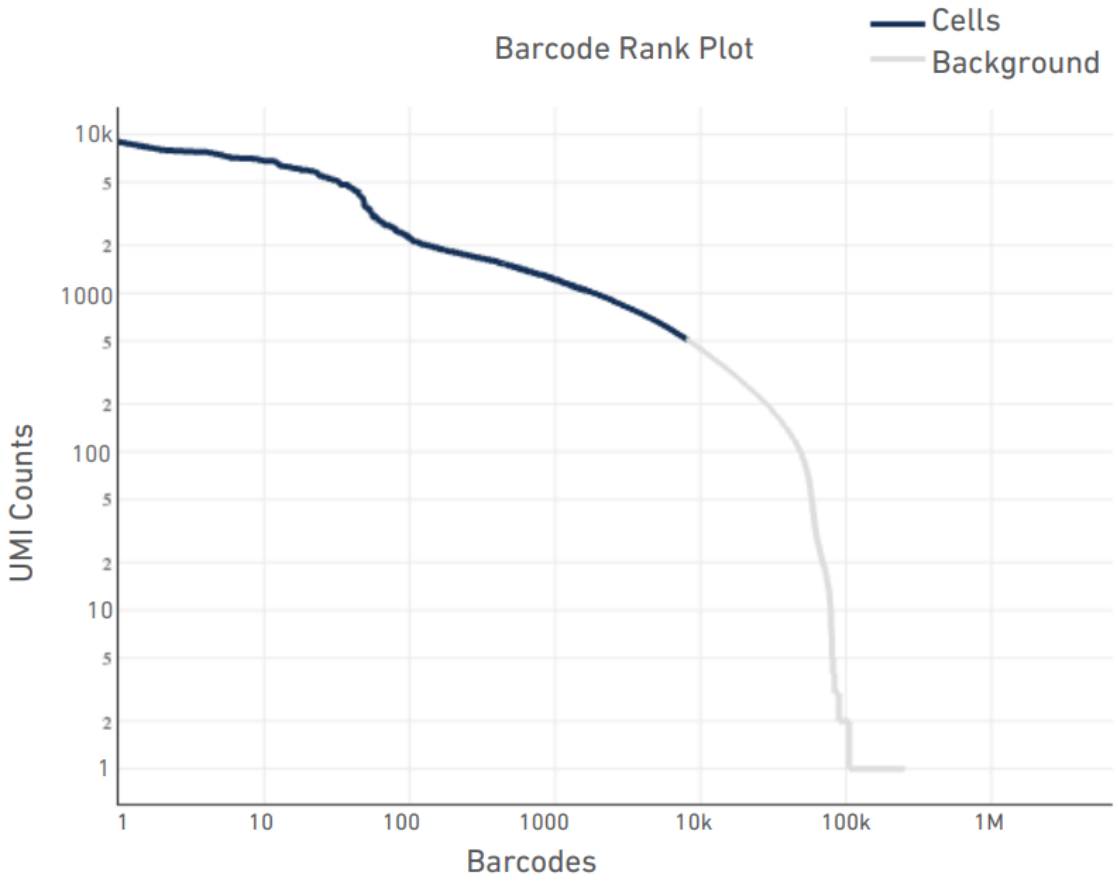


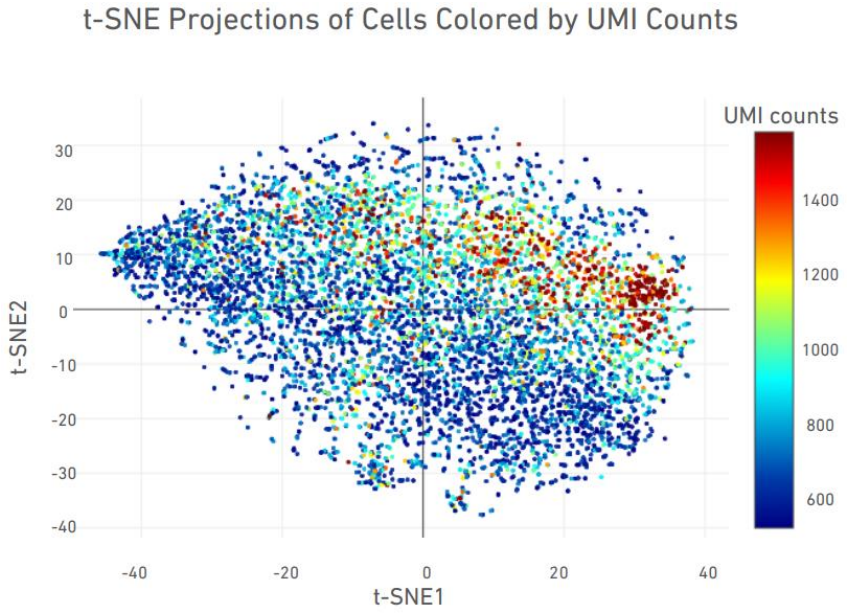


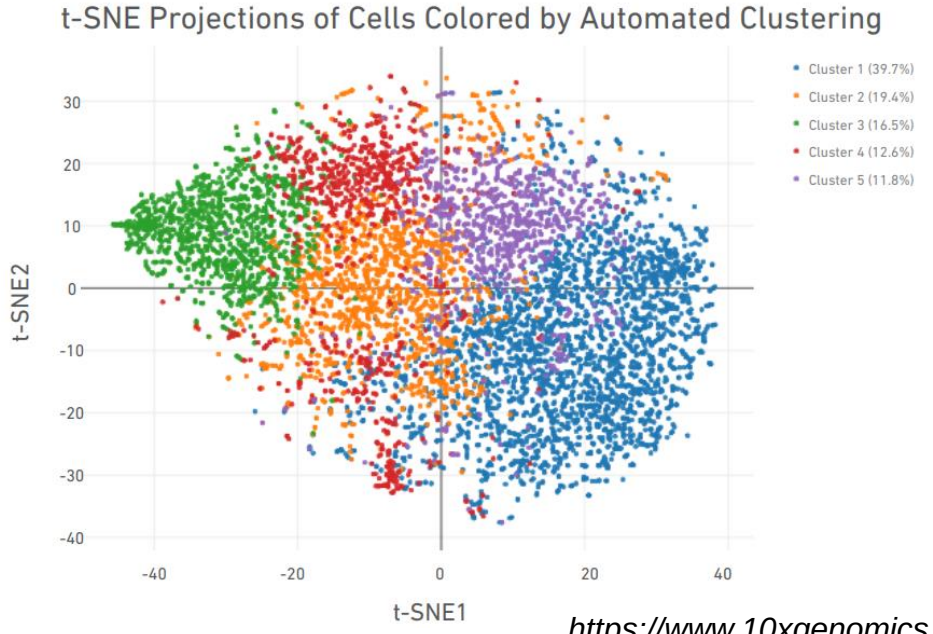
https://www.10xgenomics.com

基因-细胞表达矩阵（案例3）

- UMI-barcode图判断测序质量和细胞数：受损样本案例1
- 圆形曲线和没有陡峭的悬崖可能表明样品质量低或单细胞行为丧失。
- 这可能是由于润湿失败、细胞过早裂解或细胞活力低



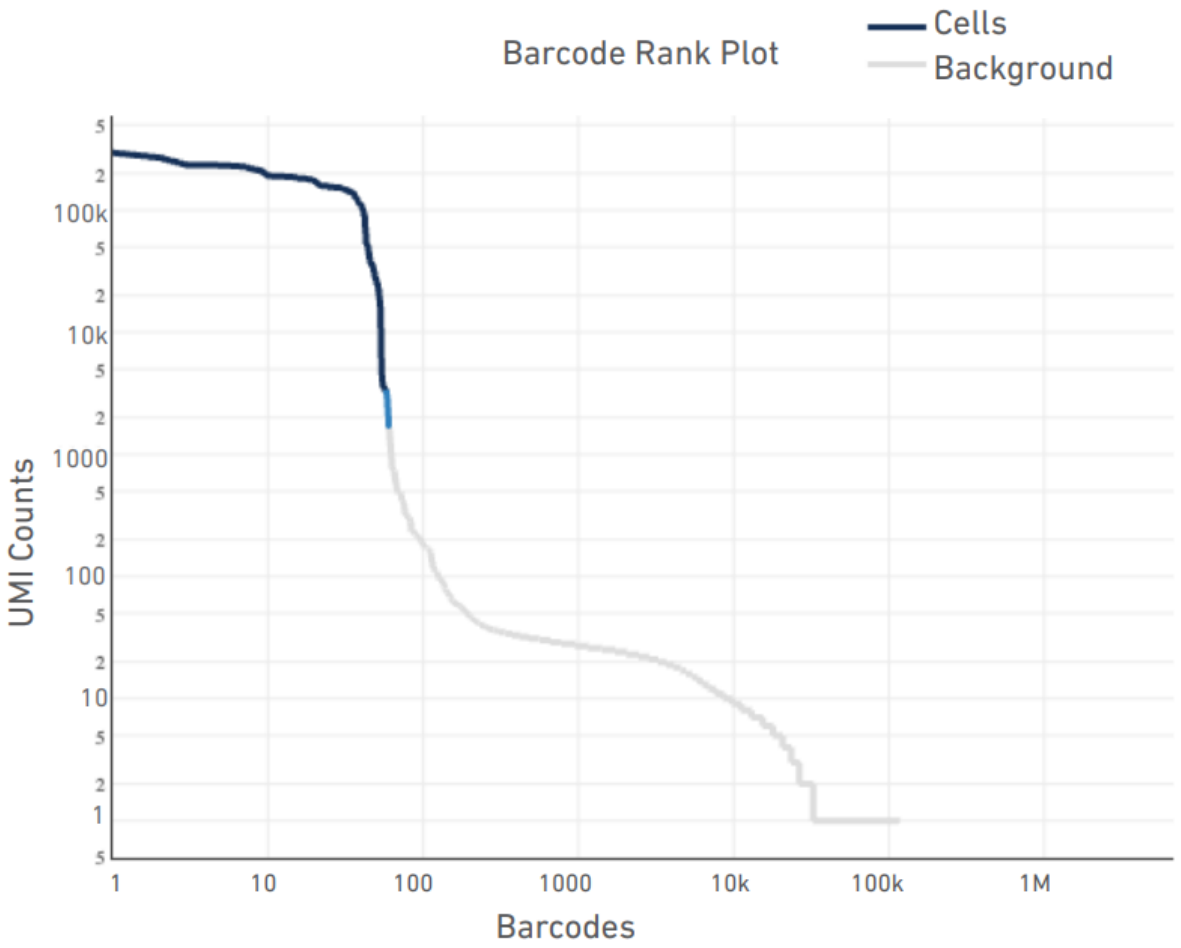


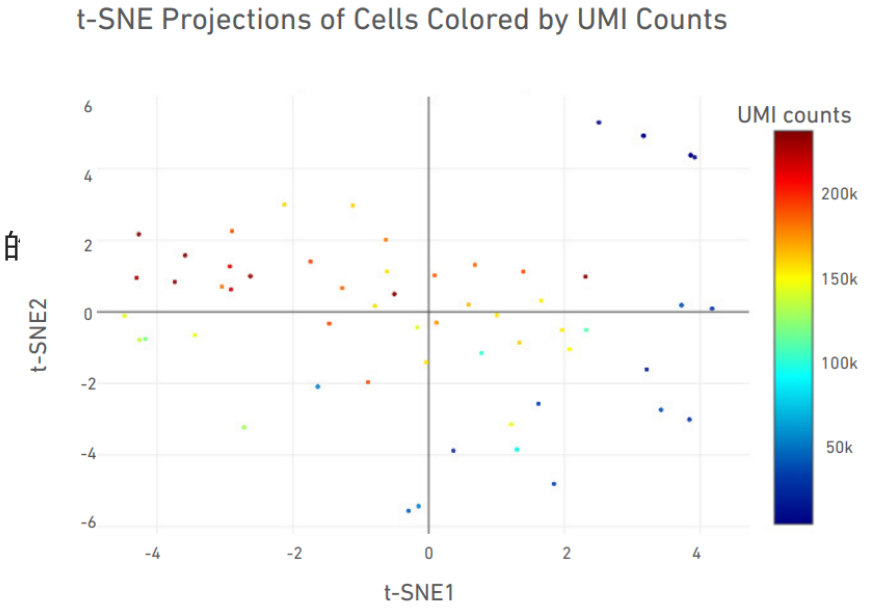


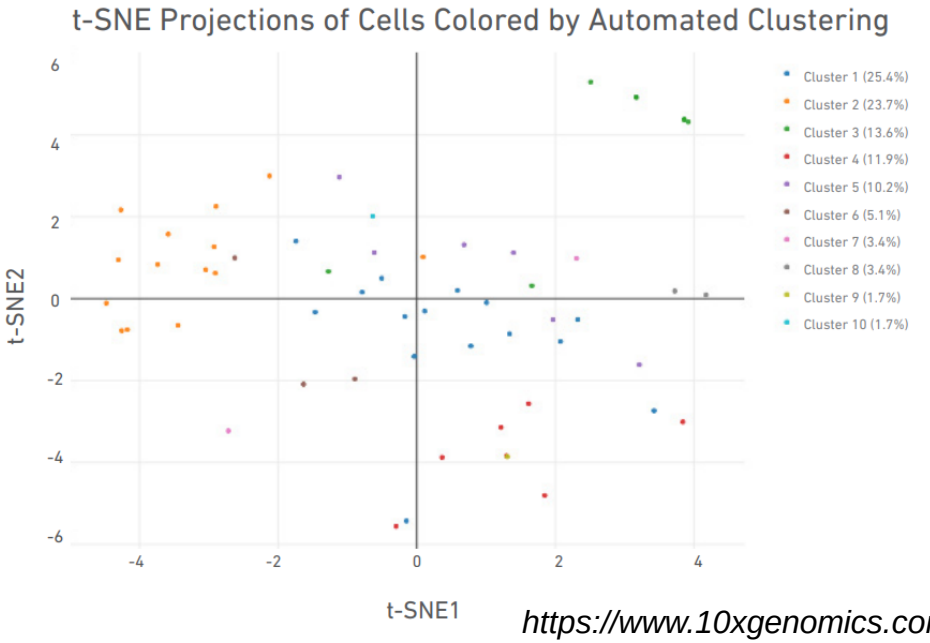
https://www.10xgenomics.com

基因-细胞表达矩阵（案例4）

- UMI-barcode图判断测序质量和细胞数：受损样本案例2
- 检测到的条码总数可能低于预期。这可能是由样品堵塞或细胞计数不准确引起的

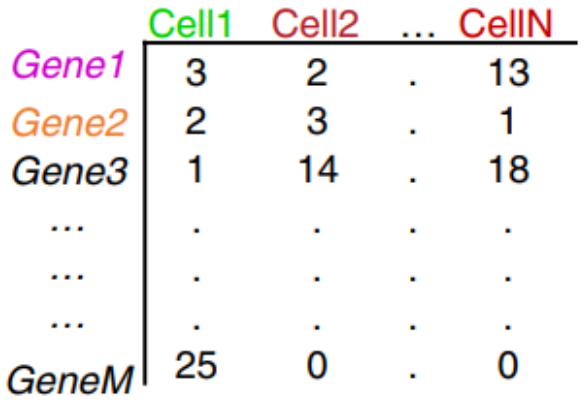






https://www.10xgenomics.com

表达矩阵



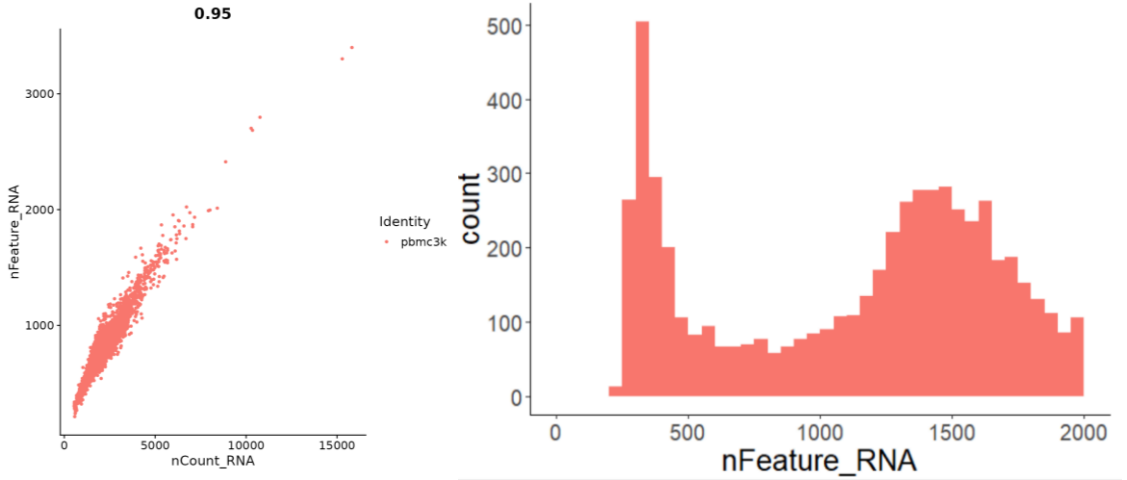
	Cell1	Cell2	…	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
…	.	.	.	.
…	.	.	.	.
…	.	.	.	.
GeneM	25	0	.	0

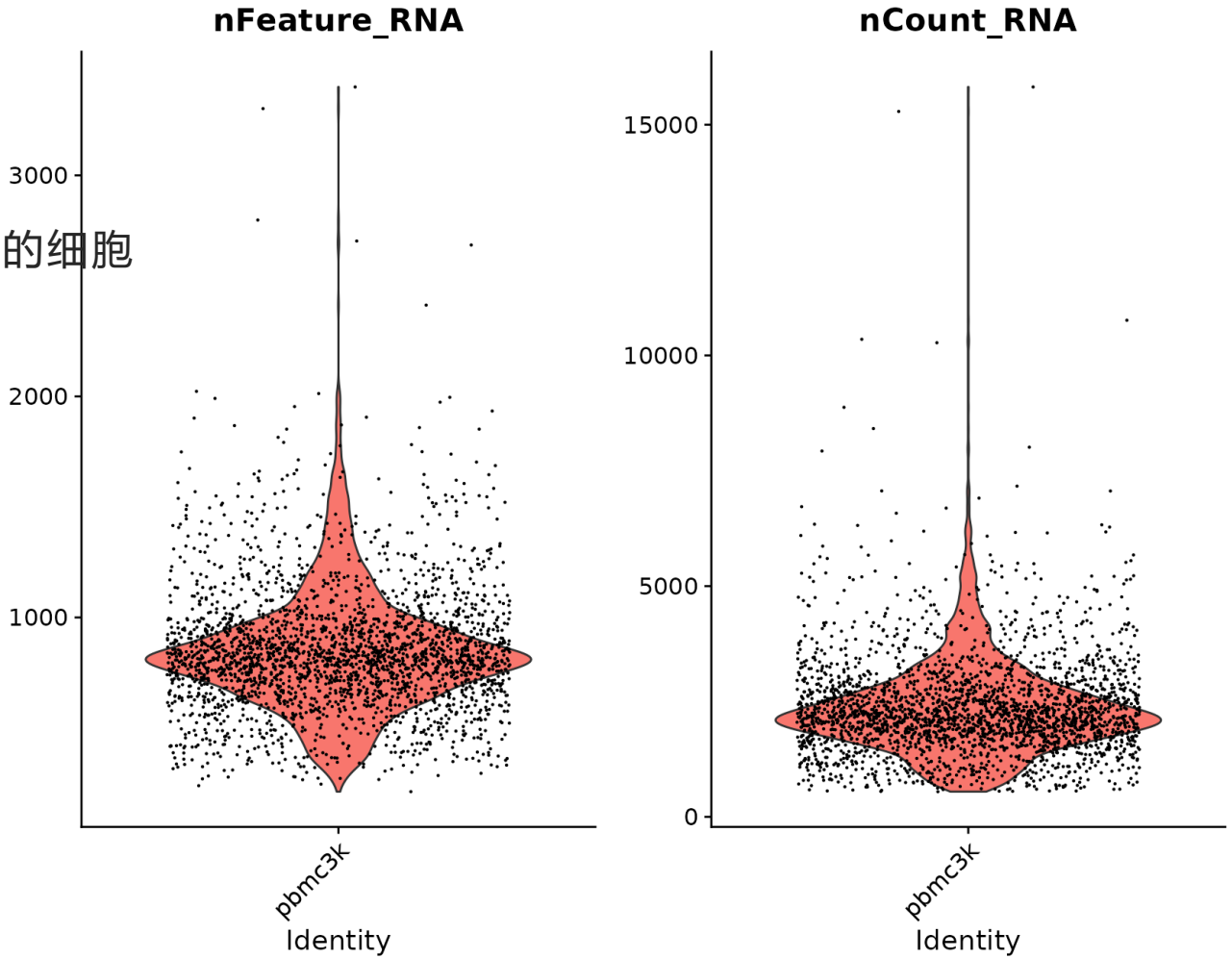
- 质控 Quality control
  - 基因数和UMI数、线粒体比例
  - 双细胞判断、去除空液滴、去除环境RNA、细胞周期判断（optional）
- 标准化 Normalization
- 特征基因选择 Feature selection
- 中心化 Scaling
- 降维 Dimensionality reduction
- 聚类 Cluster analysis
- 细胞类型注释 Cell type annotation

表达矩阵质控——基因数和UMI数

- 低质量的细胞或空滴通常只有很少的基因
- 双细胞或多细胞可能表现出异常高的基因计数

- 一般情况下：删去基因数大于 2500 或少于 200 的细胞
- 基因数与UMI数成正比
- 可通过画直方图确定最小表达的基因数

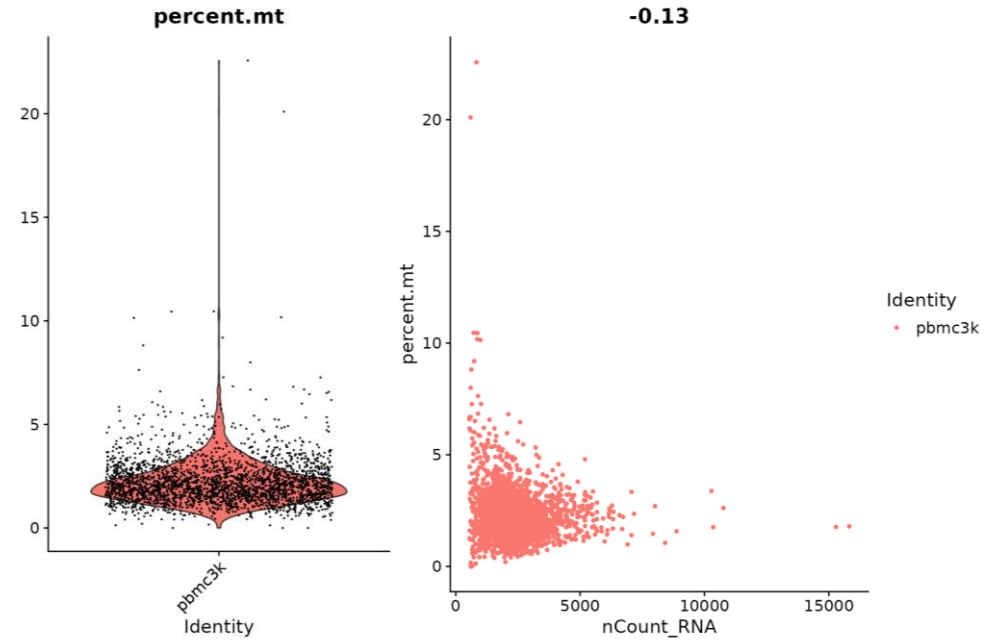


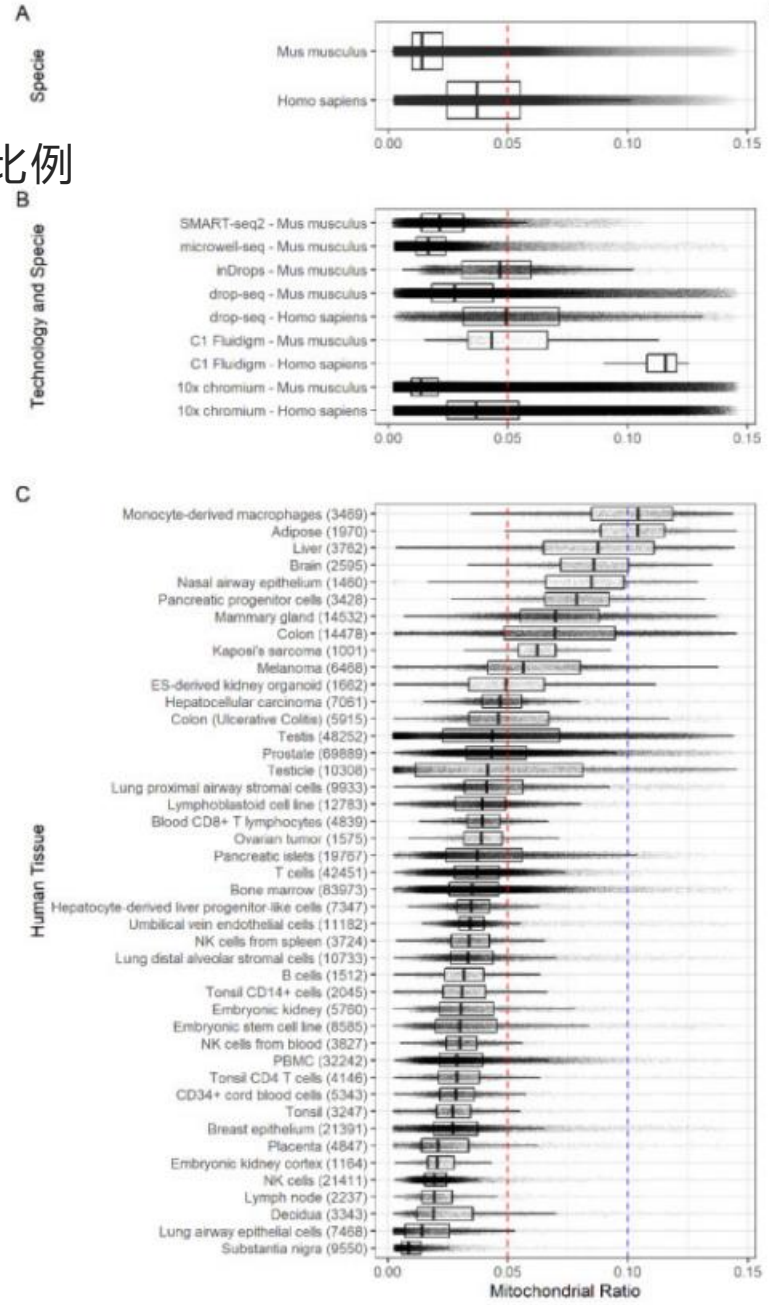


https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

表达矩阵质控——去除死细胞

- 低质量/垂死细胞通常表现出较高的线粒体比例
- 一般情况下：
- 删去 线粒体基因计数 > 5% 的细胞





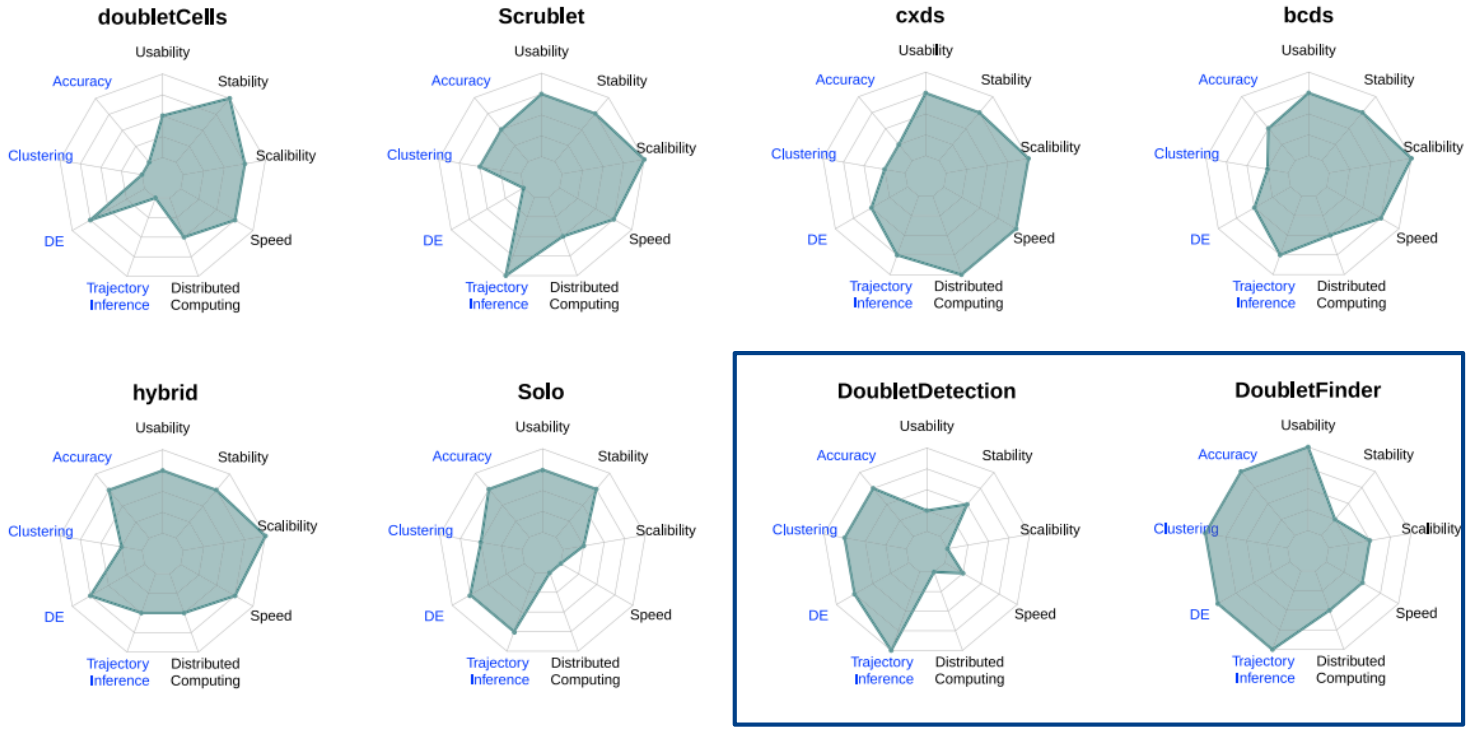
Osorio and Cai. Bioinformatics. 2021

表达矩阵质控——双细胞判断（optional）

- 单细胞或者多细胞的比例随细胞浓度的增加而增加，推荐上机浓度 700-1,200 cells/μl （10X）
- 不同预测结果之间有差异，自行判断是否去除（真正的双细胞单独成簇、有多种细胞marker；不是重要研究对象可以不去除）

Low multiplet rate
2–3% @ 10,000 cell load
8–10% @ 40,000 cell load

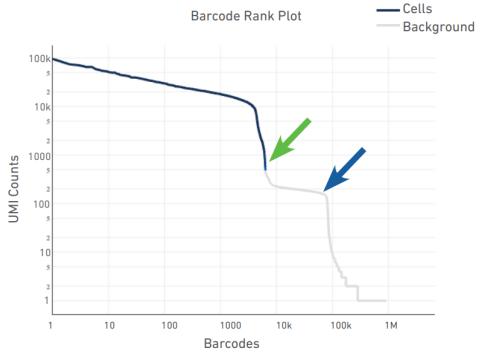
Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~800	~500
~0.8%	~1,600	~1,000
~1.6%	~3,200	~2,000
~2.3%	~4,800	~3,000
~3.1%	~6,400	~4,000
~3.9%	~8,000	~5,000
~4.6%	~9,600	~6,000
~5.4%	~11,200	~7,000
~6.1%	~12,800	~8,000
~6.9%	~14,400	~9,000
~7.6%	~16,000	~10,000

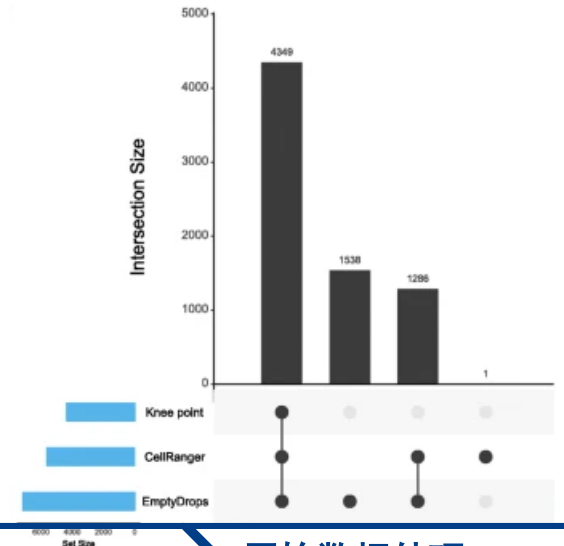


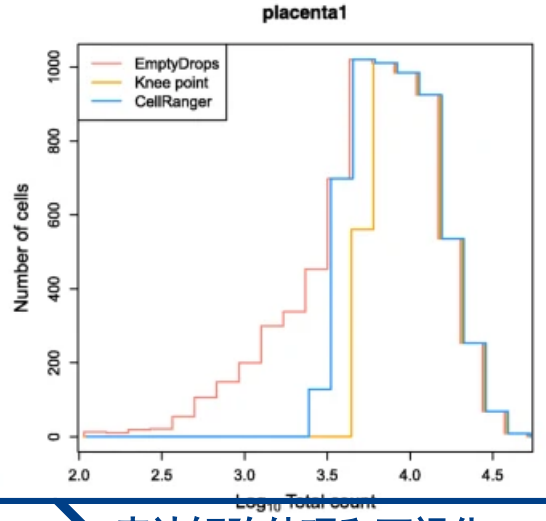
https://www.10xgenomics.com
Xi and Li. Cell Systems. 2021

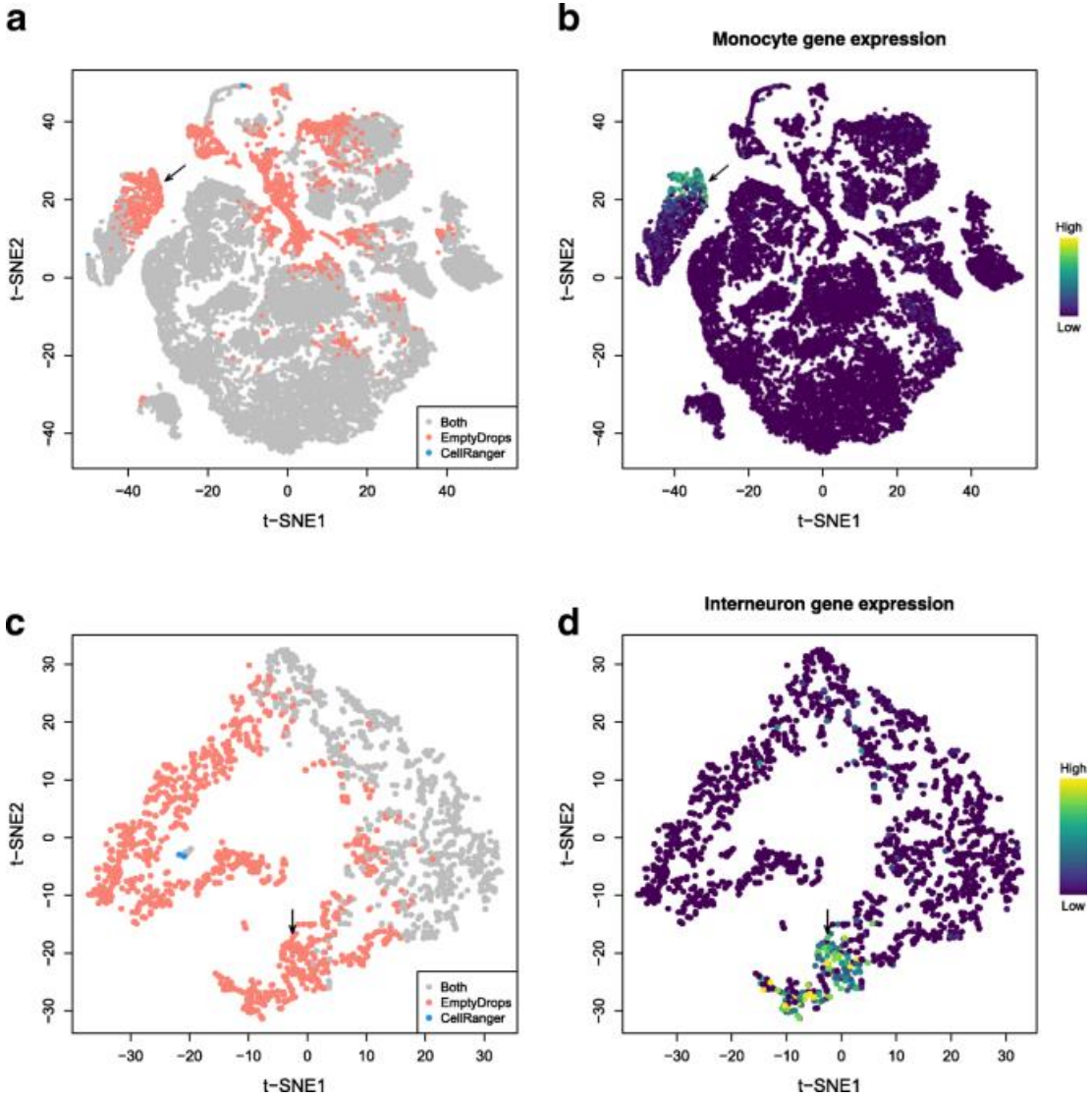
表达矩阵质控——EmptyDrops识别空液滴（optional）

- EmptyDrops
- 将真实细胞与空液滴区分开来
- 根据观察到的每个液滴的表达谱与周围溶液的表达谱来区分空液滴和含细胞的液滴。





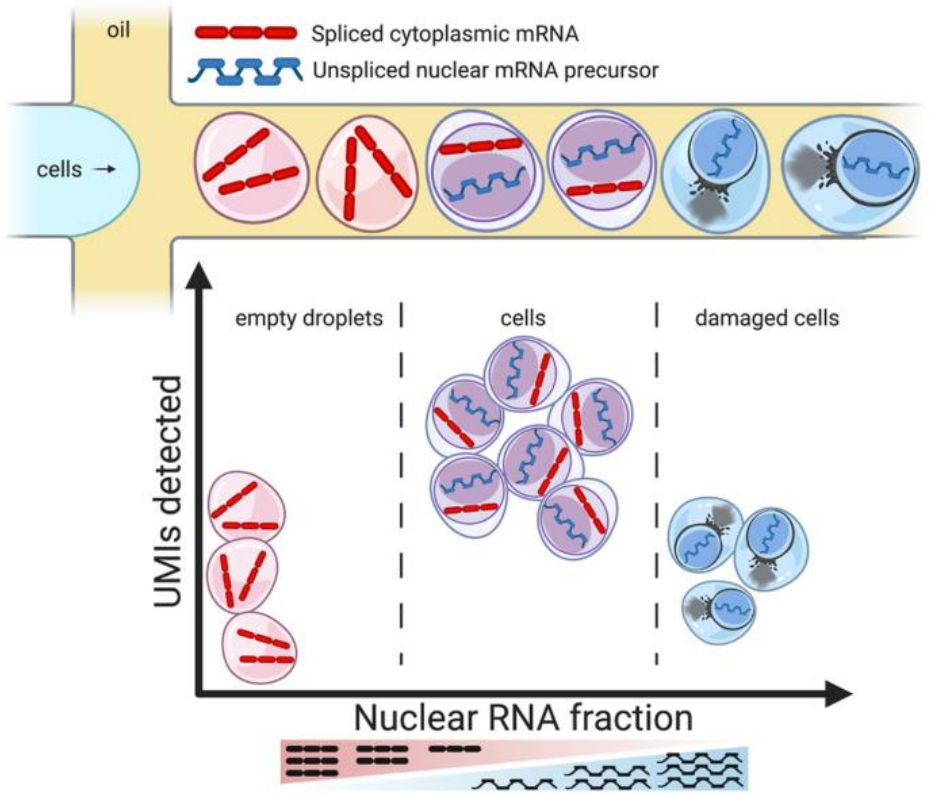


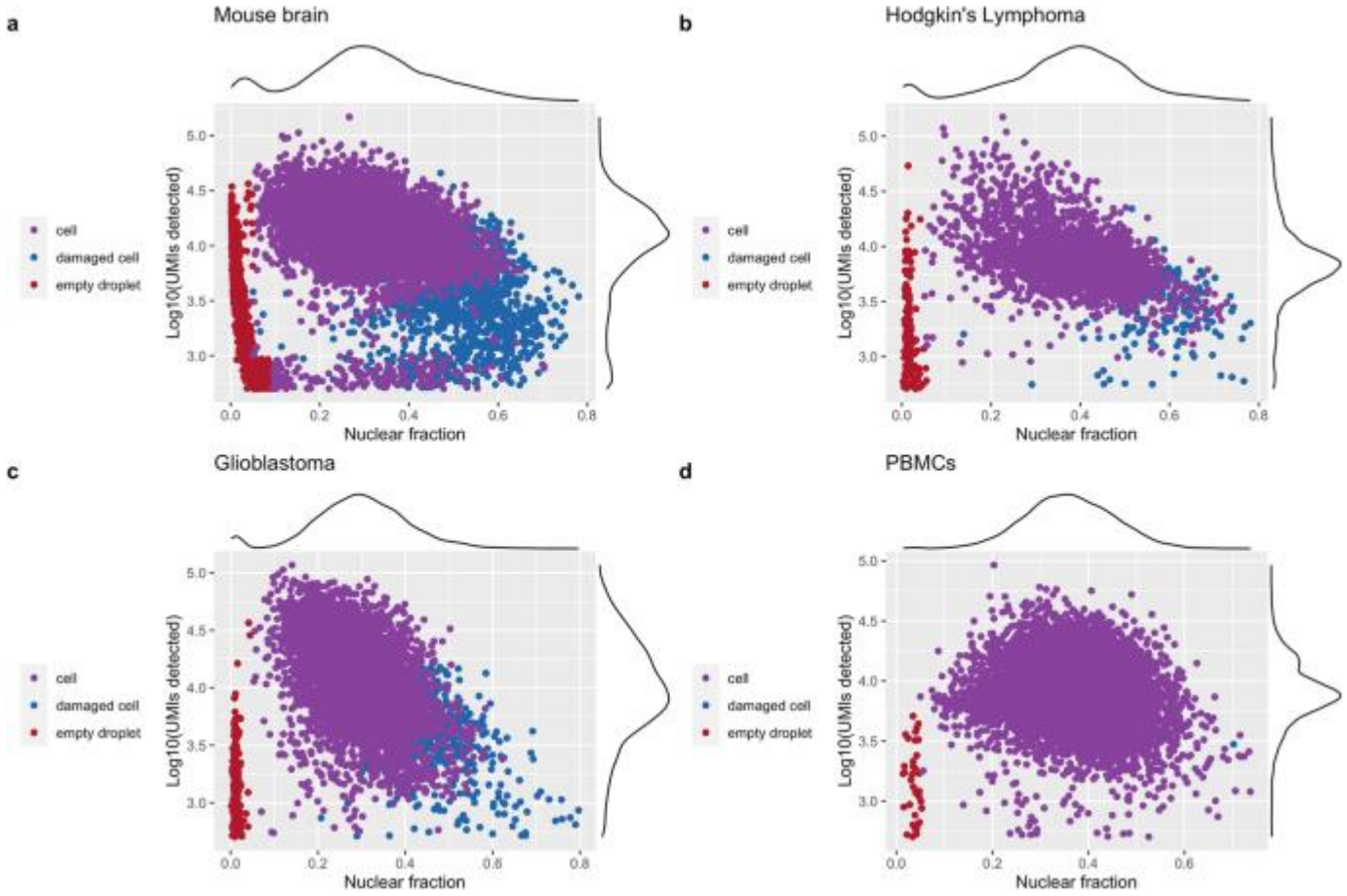


Lun et al., Genome Biology 2019

表达矩阵质控——DropletQC识别空液滴和受损的细胞（optional）

- DropletQC
- 量化了每个液滴中源自未剪接的核内前体 mRNA 的占比

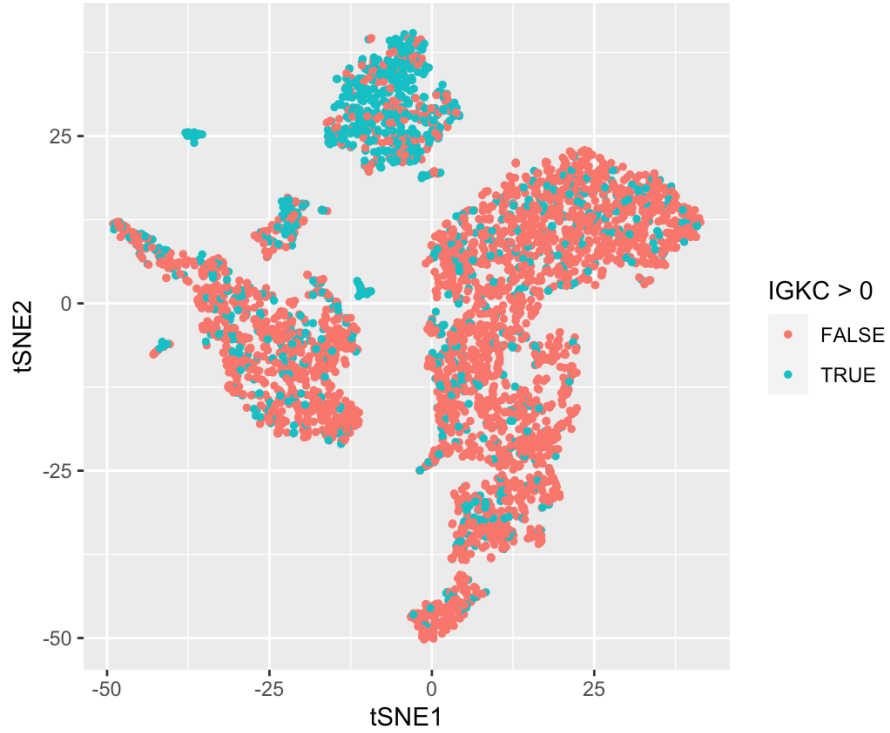




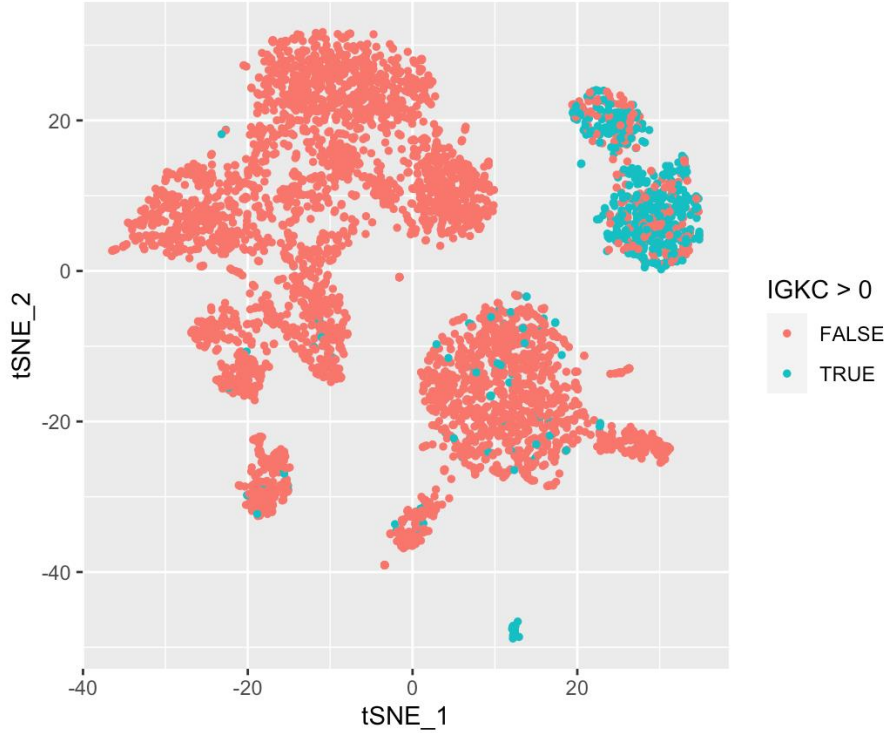
Muskovic and Powell. Genome Biology 2021

表达矩阵质控——SoupX去除环境RNA（optional）

- 由于实验原因导致某些基因的转录本扩散到大多数细胞，使得部分基因在大多数细胞中均出现高表达的现象
- 外周血中（PBMC）常见的是血细胞的污染，一般是HBB之类的基因
- 在大脑中因为神经元比较多，可能会有兴奋性神经元或抑制性神经元的污染，可能是SLC17A7或GAD1等基因的污染
- SoupX：对环境基因表达做估量并从表达基因表达矩阵中去除其影响



PBMC数据集，IGKC作为B细胞的特异表达基因，不仅在B细胞中表达，还在其他细胞中表达

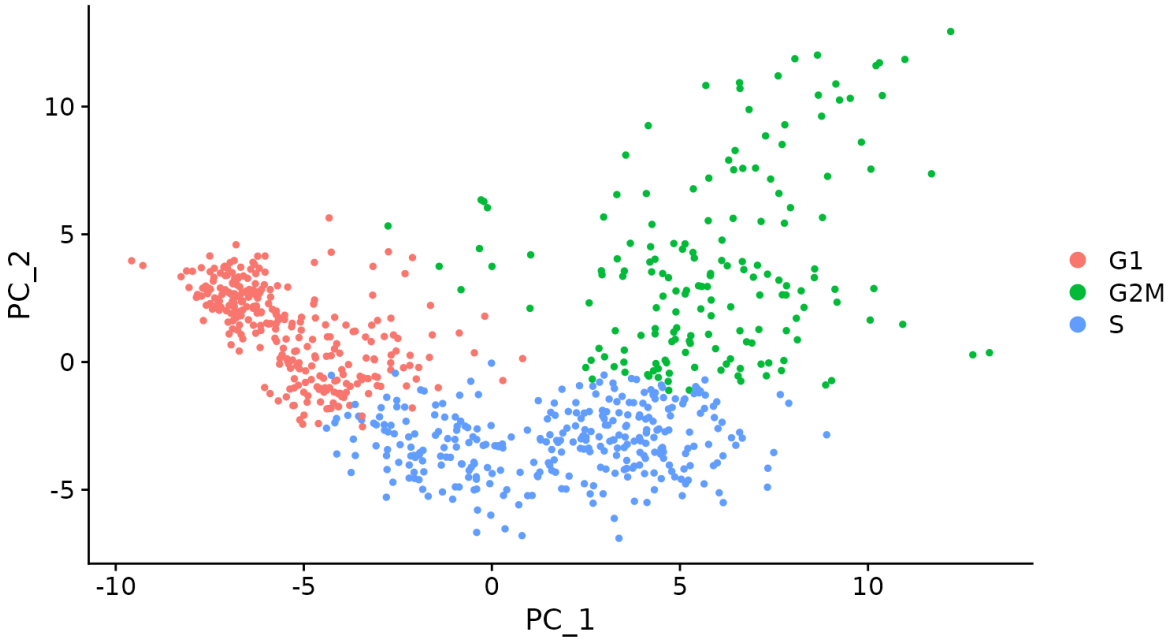


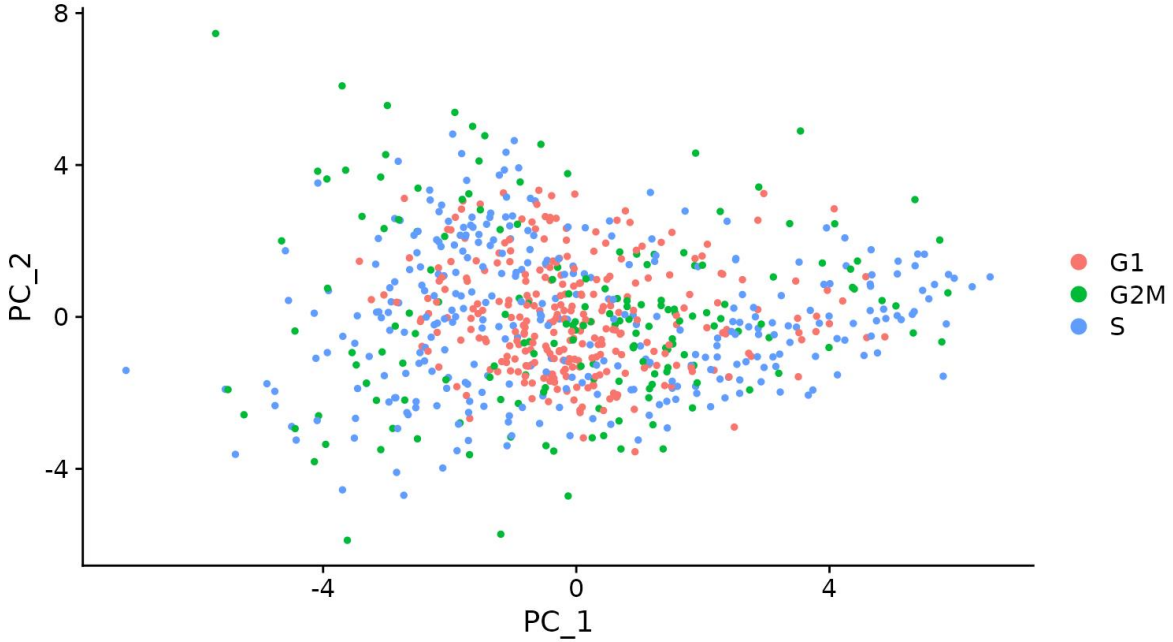
除了B细胞，其他表达IGKC的细胞数量大幅度减少了

SoupX: Young and Behjati. GigaScience 2020

表达矩阵质控——细胞周期判断（optional）

- 同一类型的细胞来自不同的细胞周期阶段，可能会对下游的聚类、细胞类型注释产生影响。
- 根据细胞周期相关基因进行打分
- 在数据中心化处理时回归细胞周期分数






```
> s.genes
 [1] "MCM5"     "PCNA"     "TYMS"     "FEN1"     "MCM2"     "MCM4"     "RRM1"     "UNG"      "GINS2"    "MCM6"     "CDCA7"    "DTL"      "PRIM1"    "UHRF1"    "MLF1IP"
[16] "HELLS"    "RFC2"     "RPA2"     "NASP"     "RAD51AP1" "GMNN"     "WDR76"    "SLBP"     "CCNE2"    "UBR7"     "POLD3"    "MSH2"     "ATAD2"    "RAD51"    "RRM2"
[31] "CDC45"    "CDC6"     "EXO1"     "TIPIN"    "DSCC1"    "BLM"      "CASP8AP2" "USP1"     "CLSPN"    "POLA1"    "CHAF1B"   "BRIP1"    "E2F8"
> g2m.genes
 [1] "HMGB2"    "CDK1"     "NUSAP1"   "UBE2C"    "BIRC5"    "TPX2"     "TOP2A"    "NDC80"    "CKS2"     "NUF2"     "CKS1B"    "MKI67"    "TMPO"     "CENPF"    "TACC3"    "FAM64A"
[17] "SMC4"     "CCNB2"    "CKAP2L"   "CKAP2"    "AURKB"    "BUB1"     "KIF11"    "ANP32E"   "TUBB4B"   "GTSE1"    "KIF20B"   "HJURP"    "CDCA3"    "HN1"      "CDC20"    "TTK"
[33] "CDC25C"   "KIF2C"    "RANGAP1"  "NCAPD2"   "DLGAP5"   "CDCA2"    "CDCA8"    "ECT2"     "KIF23"    "HMMR"     "AURKA"    "PSRC1"    "ANLN"     "LBR"      "CKAP5"    "CENPE"
[49] "CTCF"     "NEK2"     "G2E3"     "GAS2L3"   "CBX5"     "CENPA"
```

https://satijalab.org/seurat/articles/cell_cycle_vignette.html

质控前后的表达矩阵比较

```
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k",
min.cells = 3, min.features = 200)
```

```
> pbmc
An object of class Seurat
13714 features across 2700 samples within 1 assay
Active assay: RNA (13714 features, 0 variable features)
```

```
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500
& percent.mt < 5)
```

```
> pbmc
An object of class Seurat
13714 features across 2638 samples within 1 assay
Active assay: RNA (13714 features, 0 variable features)
```

表达矩阵标准化（Normalizing）

- 消除文库大小的差异，从而使不同细胞之间的表达谱的能够比较
- 确保了在不同细胞群体中观察到的异质性或差异表达都是由生物学而不是技术差异引起的

```
pbmc <- NormalizeData(pbmc, normalization.method = "LogNormalize")
```

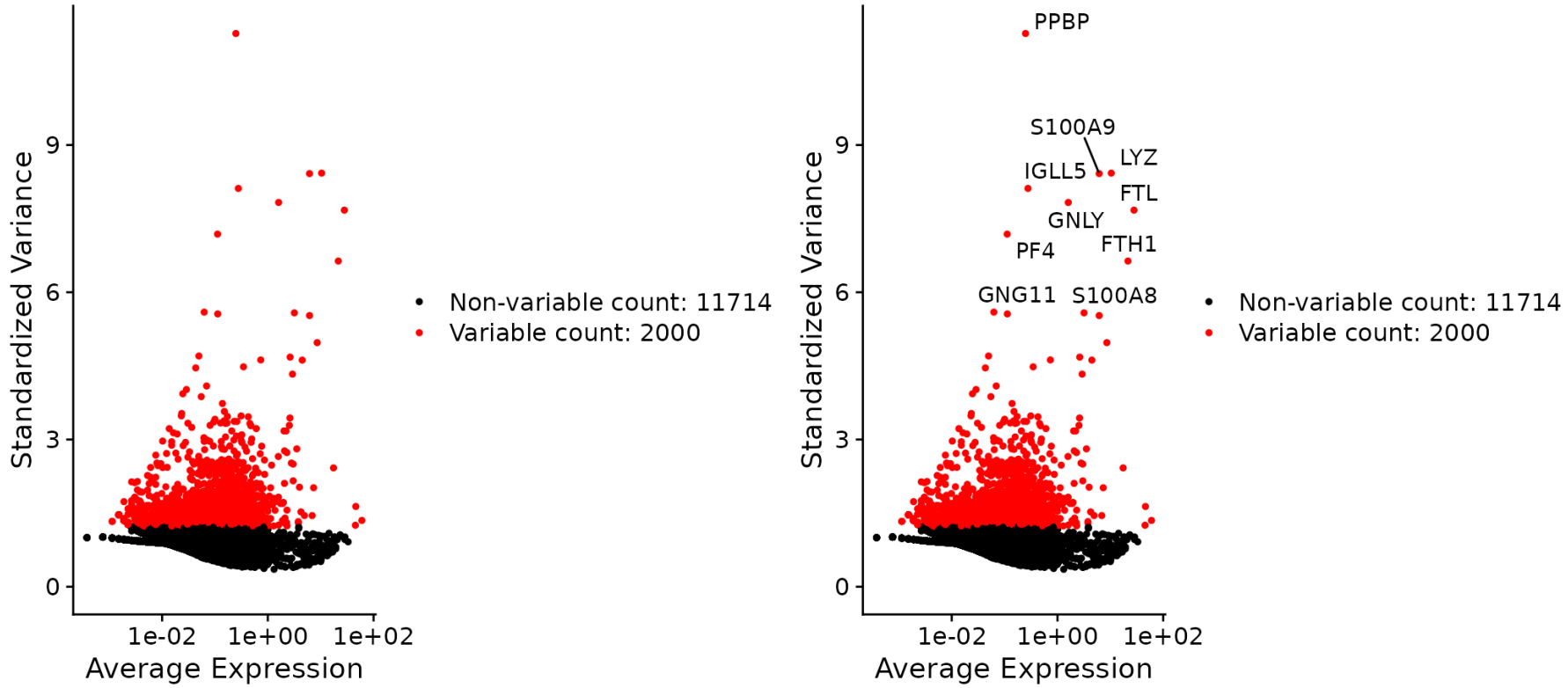
```
> pbmc@assays[["RNA"]]@counts
13714 x 2638 sparse Matrix of clas
   [[ suppressing 56 column names
   [[ suppressing 56 column names

AL627309.1    . . . . . .
AP006222.2    . . . . . .
RP11-206L10.2 . . . . . .
RP11-206L10.9 . . . . . .
LINC00115     . . . . . .
NOC2L         . . . . . .
KLHL17        . . . . . .
PLEKHN1       . . . . . .
RP11-54O7.17  . . . . . .

 ..............................
 ........suppressing 2582 columns
 ..............................
```

```
> pbmc@assays[["RNA"]]@data
13714 x 2638 sparse Matrix of class "dgCMatrix"
   [[ suppressing 56 column names 'AAACATACAACCAC-1', 'AAACATTGAGCTAC-1', 'AAACATTGATCAGC-1' ... ]]
   [[ suppressing 56 column names 'AAACATACAACCAC-1', 'AAACATTGAGCTAC-1', 'AAACATTGATCAGC-1' ... ]]

AL627309.1    . . . . . .
AP006222.2    . . . . . .
RP11-206L10.2 . . . . . .
RP11-206L10.9 . . . . . .
LINC00115     . . . . . .
NOC2L         . . . . . 1.646272 . . . . . 1.398186 . . . . . 1.89939 . . . . . 1.36907
KLHL17        . . . . . .
PLEKHN1       . . . . . .
RP11-54O7.17  . . . . . 1.015884 . . . . . .

AL627309.1    .       . . . . . .       .....
AP006222.2    .       . . . . . .       .....
RP11-206L10.2 .       . . . . . .       .....
RP11-206L10.9 .       . . . . . .       .....
LINC00115     .       . . . . . .       .....
NOC2L         1.721224 . . . . . 1.568489 .....
KLHL17        .       . . . . . .       .....
PLEKHN1       .       . . . . . .       .....
RP11-54O7.17  .       . . . . . .       .....

 ..............................
 ........suppressing 2582 columns and 13697 rows in show(); maybe adjust 'options(max.print= *, width = *)'
 ..............................
```

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

特征基因选择

- 高变特征基因有助于突出单细胞数据集中的生物信号
- 特征基因将用于下游分析，如PCA



https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

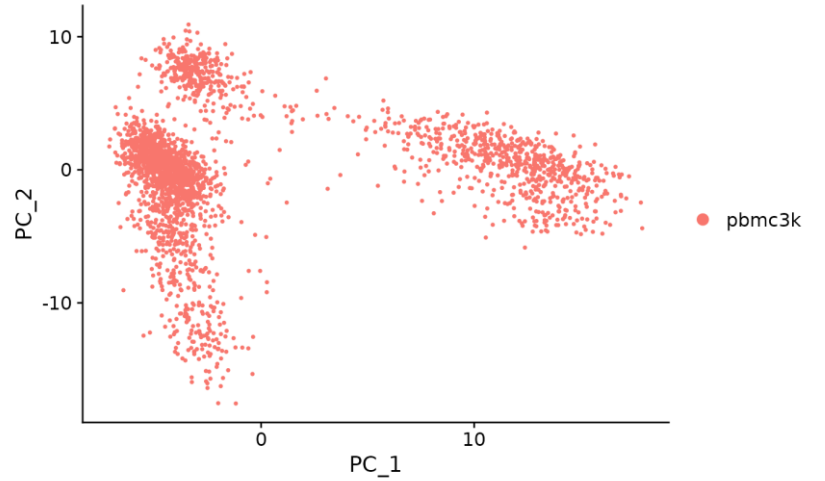
表达矩阵中心化（Scaling）

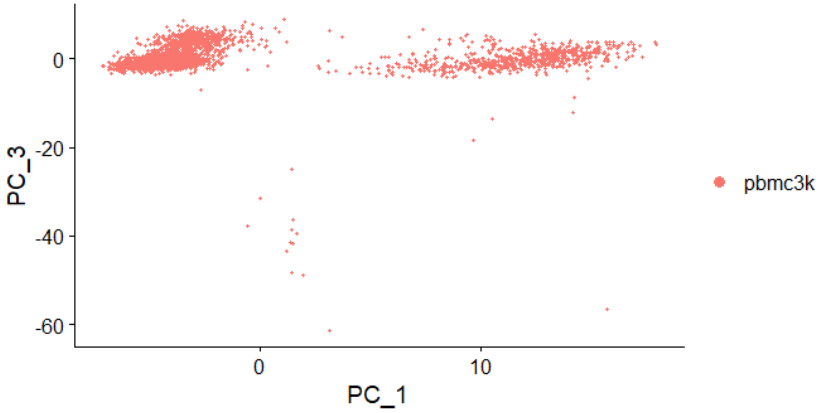
- 对基因表达量进行处理，目的是消除不同样本中基因表达的平均水平和偏离度的影响，为了后续的分析不受基因表达的极值影响（改变每个基因的表达，使细胞中的平均表达为0；缩放每个基因的表达，使细胞间的差异为1）
- 默认基于选择的高变基因进行中心化，这样得到的 scale.data 矩阵只有2000个基因

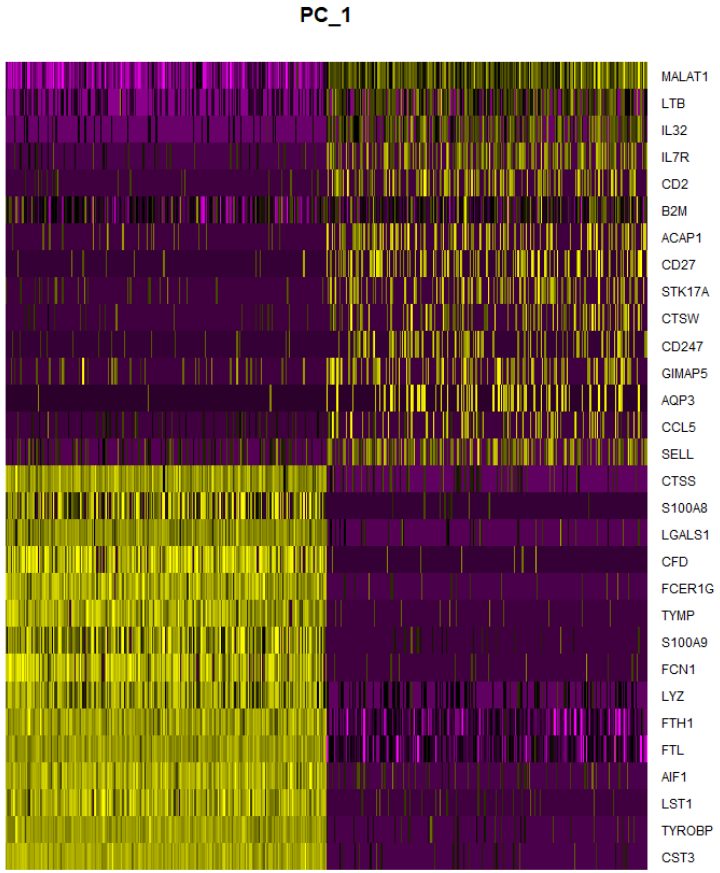
```
> head(pbmc@assays[["RNA"]]@counts)[,1:5]
6 x 5 sparse Matrix of class "dgCMatrix"
              AAACATACAACCAC-1 AAACATTGAGCTAC-1 AAACATTGATCAGC-1 AAACCGTGCTTCCG-1 AAACCGTGTATGCG-1
AL627309.1                  .                .                .                .                .
AP006222.2                  .                .                .                .                .
RP11-206L10.2               .                .                .                .                .
RP11-206L10.9               .                .                .                .                .
LINC00115                   .                .                .                .                .
NOC2L                       .                .                .                .                .
> head(pbmc@assays[["RNA"]]@data)[,1:5]
6 x 5 sparse Matrix of class "dgCMatrix"
              AAACATACAACCAC-1 AAACATTGAGCTAC-1 AAACATTGATCAGC-1 AAACCGTGCTTCCG-1 AAACCGTGTATGCG-1
AL627309.1                  .                .                .                .                .
AP006222.2                  .                .                .                .                .
RP11-206L10.2               .                .                .                .                .
RP11-206L10.9               .                .                .                .                .
LINC00115                   .                .                .                .                .
NOC2L                       .                .                .                .                .
> head(pbmc@assays[["RNA"]]@scale.data)[,1:5]
              AAACATACAACCAC-1 AAACATTGAGCTAC-1 AAACATTGATCAGC-1 AAACCGTGCTTCCG-1 AAACCGTGTATGCG-1
AL627309.1        -0.05812316      -0.05812316      -0.05812316      -0.05812316      -0.05812316
AP006222.2        -0.03357571      -0.03357571      -0.03357571      -0.03357571      -0.03357571
RP11-206L10.2     -0.04166819      -0.04166819      -0.04166819      -0.04166819      -0.04166819
RP11-206L10.9     -0.03364562      -0.03364562      -0.03364562      -0.03364562      -0.03364562
LINC00115         -0.08223981      -0.08223981      -0.08223981      -0.08223981      -0.08223981
NOC2L             -0.31717081      -0.31717081      -0.31717081      -0.31717081      -0.31717081
```

表达矩阵降维（PCA）

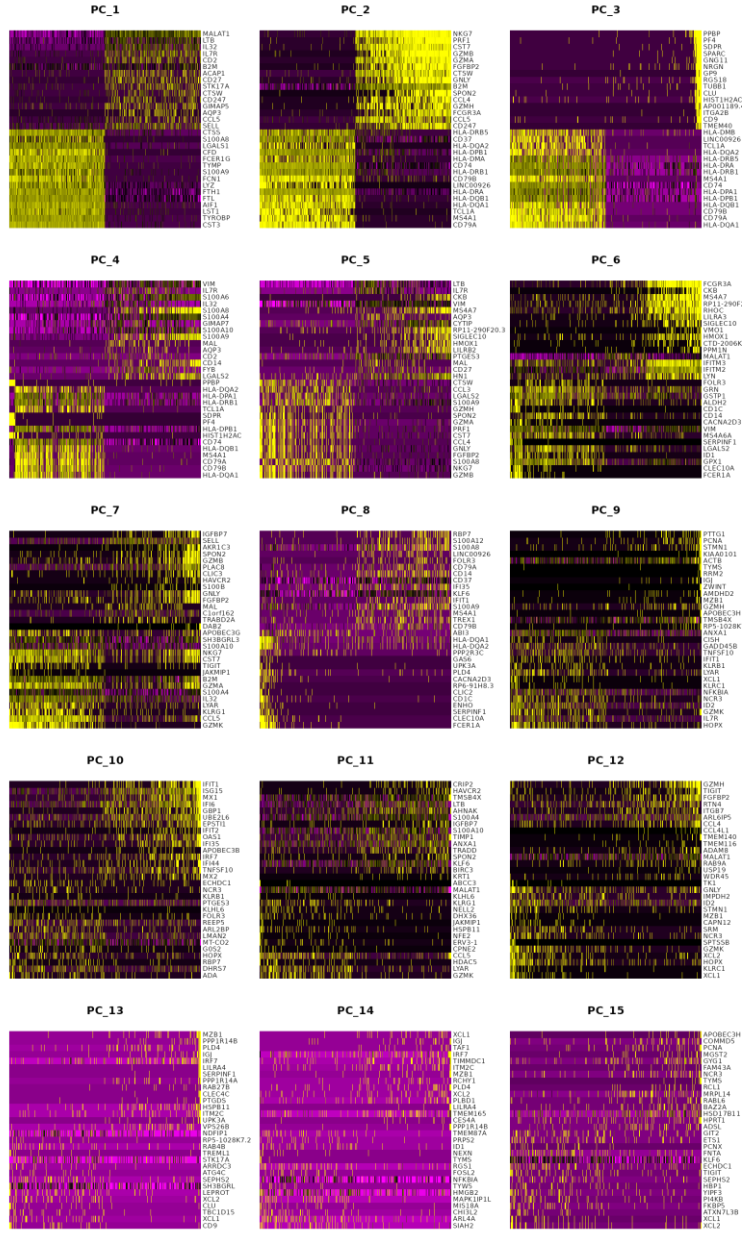
- 对 scale.data 矩阵进行线性降维，PCA（Principle Component Analysis）
- 前几个主成分能够解释绝大部分数据的方差，体现最多的差异





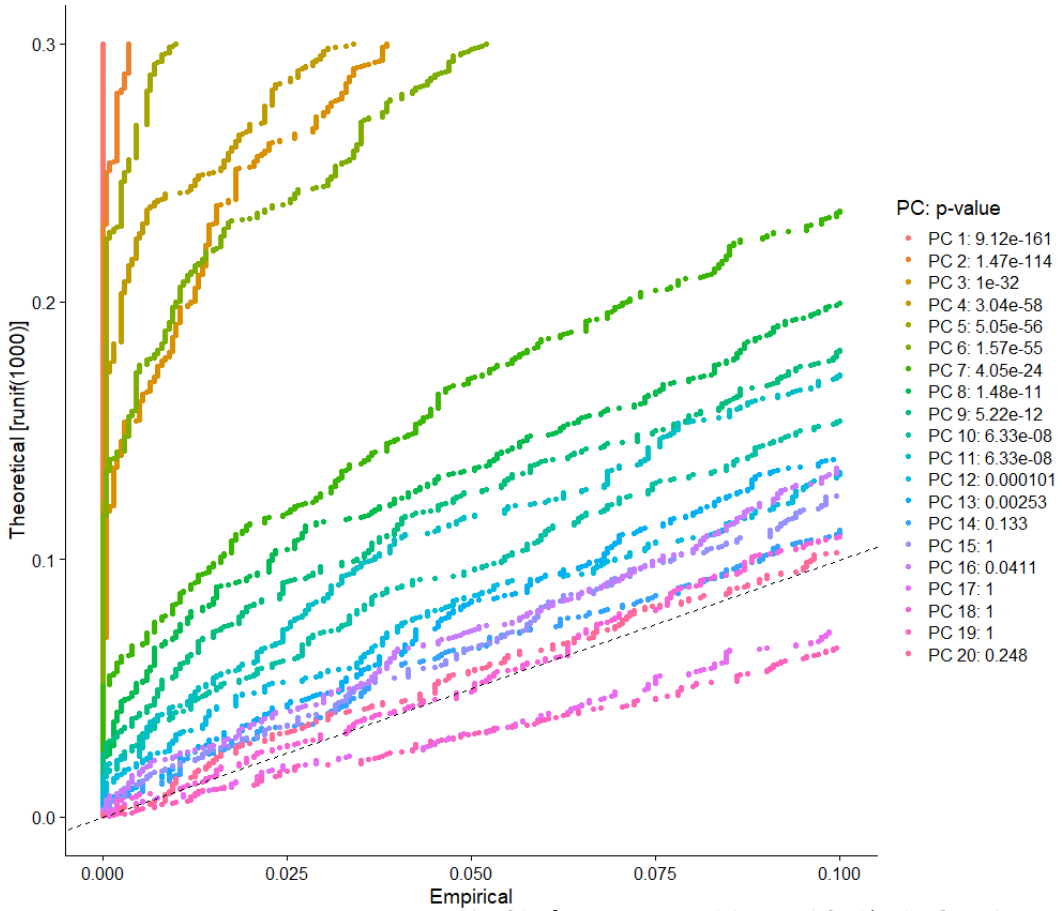


- 该PC的变异最大的基因



https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

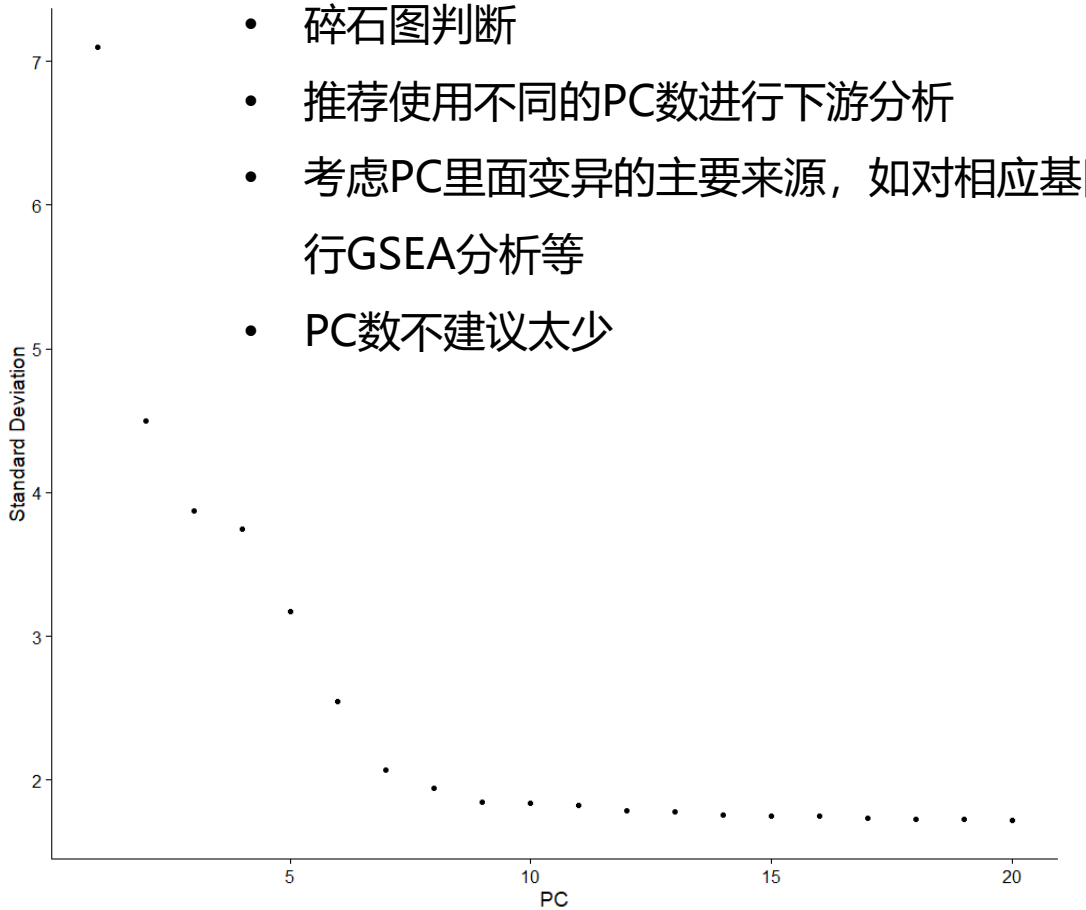
选择合适的主成分数量



- JackStrawPlot() 函数 将每个 PC 的 p 值分布与均匀分布（虚线）进行比较。
- 在前 10-12 个 PC 之后，重要性急剧下降。

如何选择合适的主成分数量：

- 显著性P值
- 碎石图判断
- 推荐使用不同的PC数进行下游分析
- 考虑PC里面变异的主要来源，如对相应基因进行GSEA分析等
- PC数不建议太少



- ElbowPlot() 函数 根据每个主成分解释的方差百分比对主成分进行排名。
- 大部分真实信号是在前 10 个 PC 中捕获。

聚类

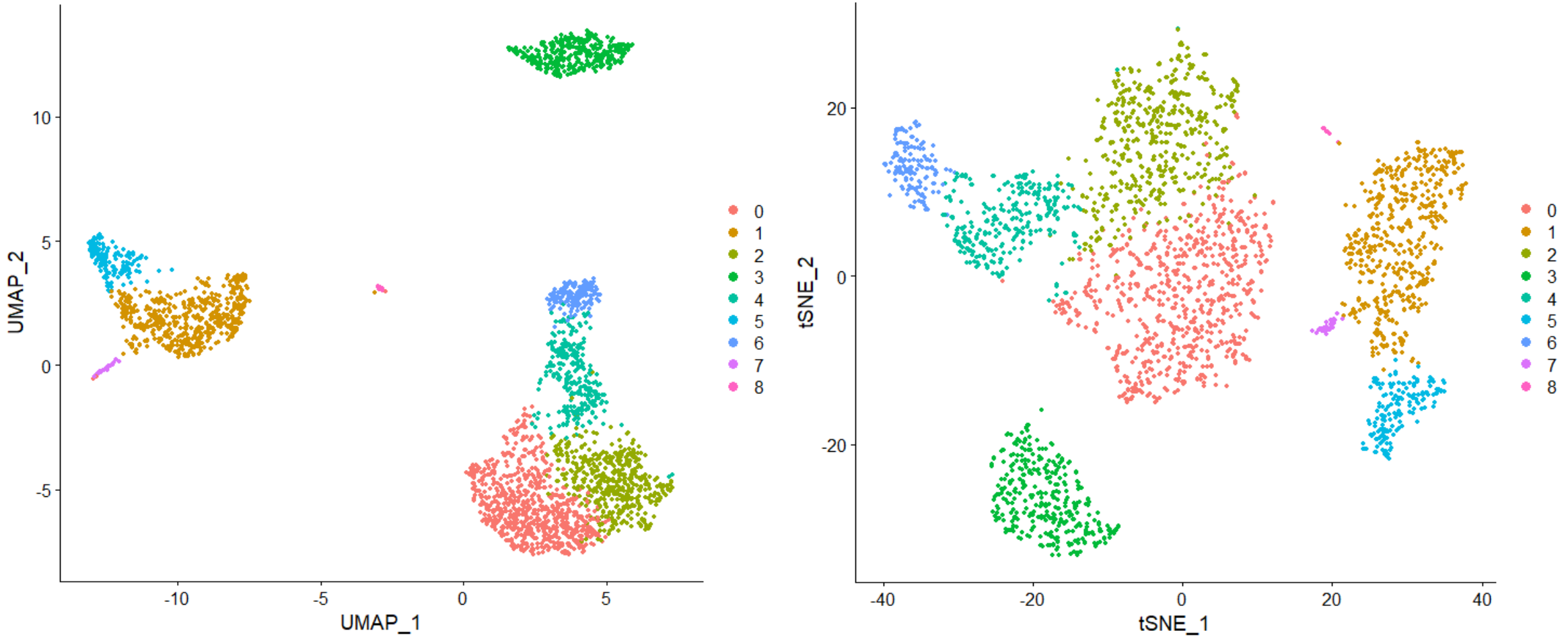
- **FindNeighbors()**: construct a KNN graph based on the euclidean distance in PCA space, and refine the edge weights between any two cells based on the shared overlap in their local neighborhoods
  FindClusters(): apply modularity optimization techniques to iteratively group cells together, with the goal of optimizing the standard modularity function

```
> pbmc <- FindNeighbors(pbmc, dims = 1:10)
Computing nearest neighbor graph
Computing SNN
> pbmc <- FindClusters(pbmc, resolution = 0.5)
Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan

Number of nodes: 2638
Number of edges: 95965

Running Louvain algorithm...
0%   10   20   30   40   50   60   70   80   90   100%
[----|----|----|----|----|----|----|----|----|----|
**************************************************|
Maximum modularity in 10 random starts: 0.8723
Number of communities: 9
Elapsed time: 0 seconds
```

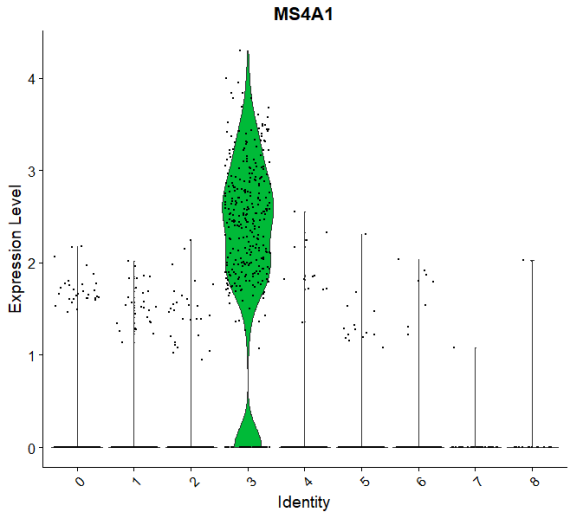
```
> head(pbmc@meta.data, n = 40)
                orig.ident nCount_RNA nFeature_RNA percent.mt RNA_snn_res.0.5 seurat_clusters
AAACATACAACCAC-1      pbmc3k       2419          779  3.0177759               2               2
AAACATTGAGCTAC-1      pbmc3k       4903         1352  3.7935958               3               3
AAACATTGATCAGC-1      pbmc3k       3147         1129  0.8897363               2               2
AAACCGTGCTTCCG-1      pbmc3k       2639          960  1.7430845               1               1
AAACCGTGTATGCG-1      pbmc3k        980          521  1.2244898               6               6
AAACGCACTGGTAC-1      pbmc3k       2163          781  1.6643551               2               2
AAACGCTGACCAGT-1      pbmc3k       2175          782  3.8160920               4               4
AAACGCTGGTTCTT-1      pbmc3k       2260          790  3.0973451               4               4
AAACGCTGTAGCCA-1      pbmc3k       1275          532  1.1764706               0               0
AAACGCTGTTTCTG-1      pbmc3k       1103          550  2.9011786               5               5
AAACTTGAAAAACG-1      pbmc3k       3914         1112  2.6315789               3               3
AAACTTGATCCAGA-1      pbmc3k       2388          747  1.0887772               0               0
AAAGAGACGAGATA-1      pbmc3k       2410          864  1.0788382               0               0
AAAGAGACGCGAGA-1      pbmc3k       3033         1058  1.4177382               1               1
AAAGAGACGGACTT-1      pbmc3k       1151          457  2.3457863               0               0
AAAGAGACGGCATT-1      pbmc3k        792          335  2.3989899               0               0
AAAGCAGATATCGG-1      pbmc3k       4584         1422  1.3961606               1               1
AAAGCCTGTATGCG-1      pbmc3k       2928         1013  1.7076503               2               2
AAAGGCCTGTCTAG-1      pbmc3k       4973         1445  1.5282526               3               3
AAAGTTTGATCACG-1      pbmc3k       1268          444  3.4700315               3               3
AAAGTTTGGGGTGA-1      pbmc3k       3281         1015  2.5906736               3               3
AAAGTTTGTAGAGA-1      pbmc3k       1102          417  1.5426497               0               0
AAAGTTTGTAGCGT-1      pbmc3k       2683          877  2.4972046               1               1
AAATCAACAATGCC-1      pbmc3k       2319          787  1.1642950               3               3
AAATCAACACCAGT-1      pbmc3k       1412          508  1.9830028               0               0
AAATCAACCAGGAG-1      pbmc3k       2800          823  2.2500000               0               0
AAATCAACCCTATT-1      pbmc3k       5676         1541  2.4312896               5               5
AAATCAACGGAAGC-1      pbmc3k       3473          996  1.7564066               0               0
AAATCAACTCGCAA-1      pbmc3k       2811          936  1.8498755               2               2
AAATCATGACCACA-1      pbmc3k       4128         1368  4.5784884               5               5
AAATCCCTCCACAA-1      pbmc3k        955          427  1.9895288               0               0
AAATCCCTGCTATG-1      pbmc3k        822          406  1.7031630               3               3
AAATGTTGAACGAA-1      pbmc3k       3208         1017  1.9326683               1               1
AAATGTTGCCACAA-1      pbmc3k       1760          785  0.6250000               2               2
AAATGTTGTGGCAT-1      pbmc3k       2761         1017  1.9195943               1               1
AAATTCGAAGGTTC-1      pbmc3k       2740          749  2.2262774               0               0
AAATTCGAATCACG-1      pbmc3k       2567          822  1.9867550               1               1
AAATTCGAGCTGAT-1      pbmc3k       2969          980  2.1892893               5               5
AAATTCGAGGAGTG-1      pbmc3k       2978          873  2.4848892               0               0
AAATTCGATTCTCA-1      pbmc3k       2641          928  1.4388489               4               4
```

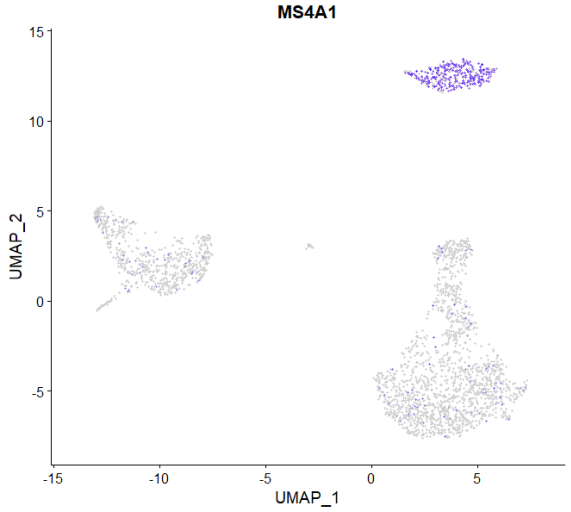
聚类可视化

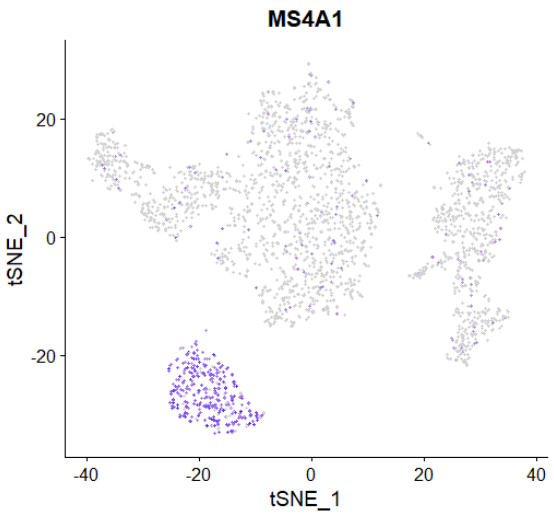
- t分布随机邻居嵌入（t-SNE）和均匀流形逼近与投影（UMAP）：将高维空间中具有相似局部邻域的细胞放置在低维空间中

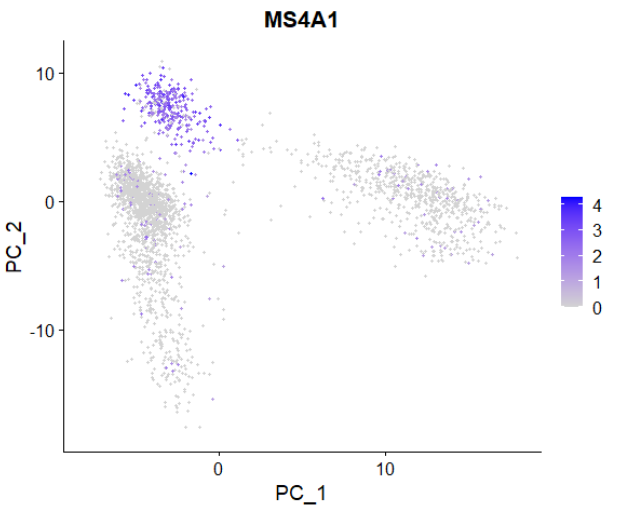


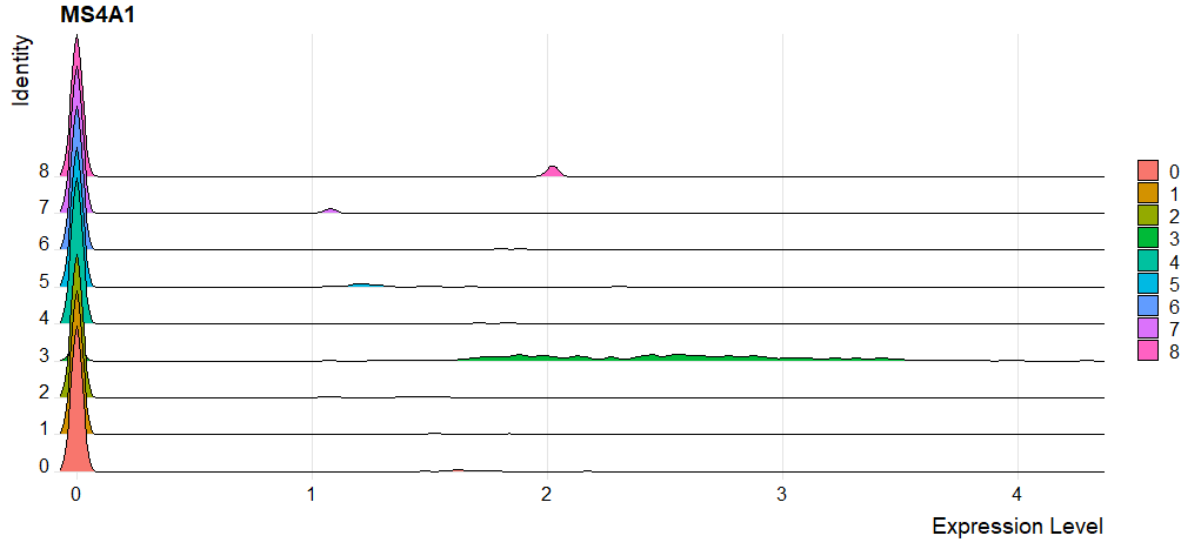
任意基因的可视化

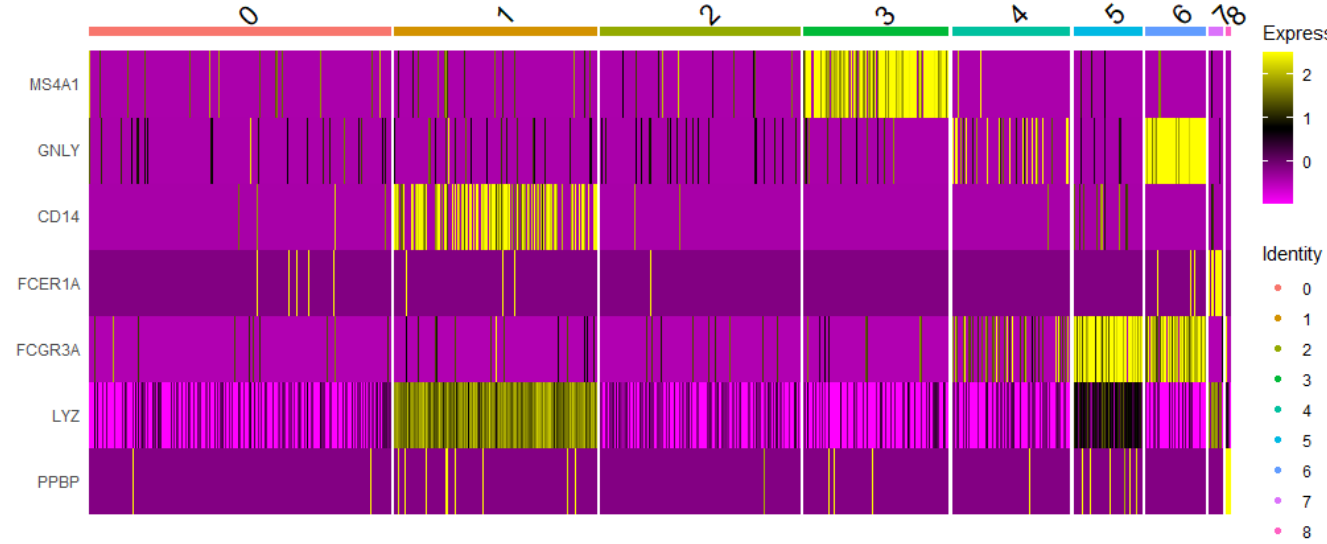

MS4A1
Expression Level
Identity


MS4A1
UMAP_2
UMAP_1


MS4A1
tSNE_2
tSNE_1

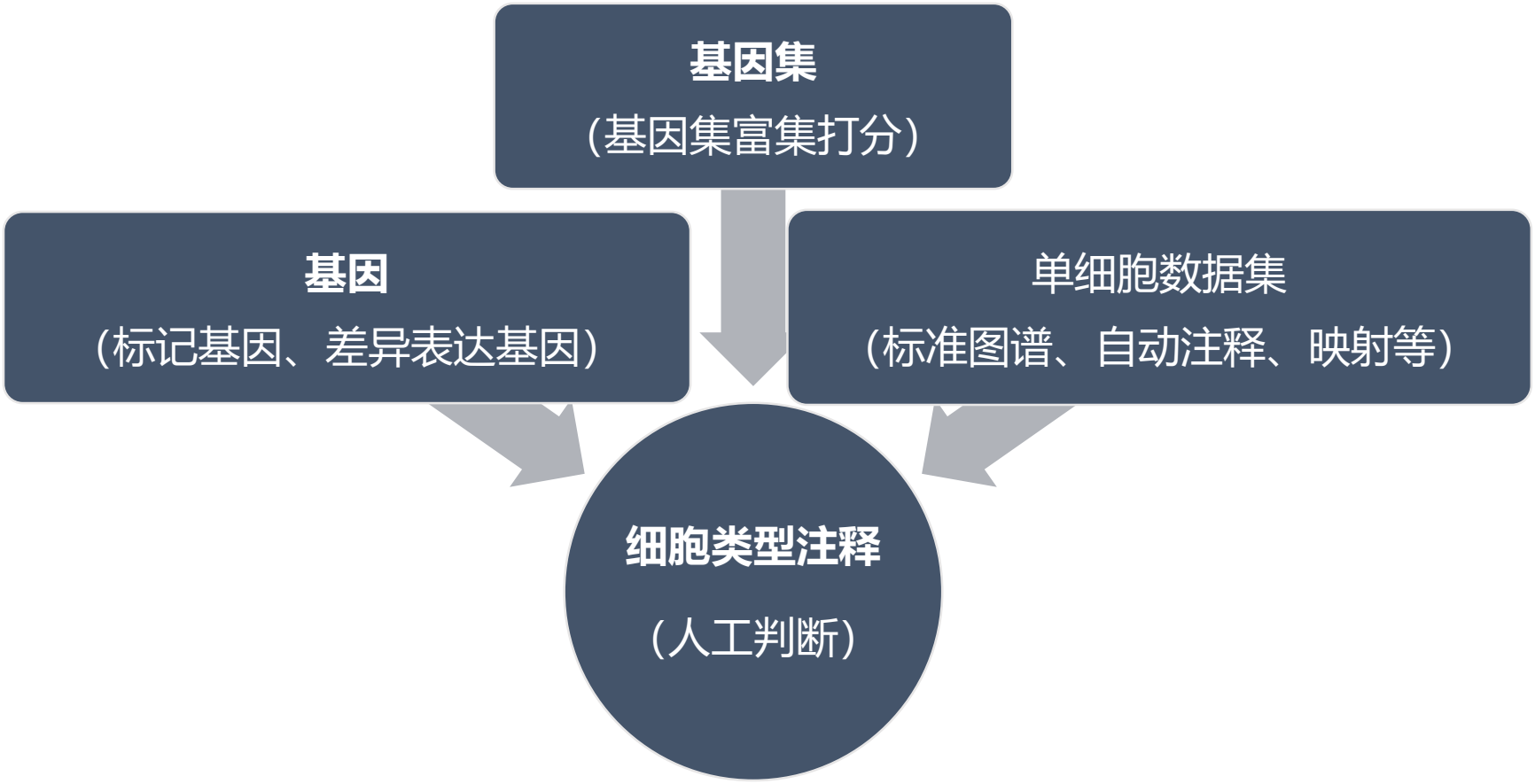

MS4A1
PC_2
PC_1


MS4A1
Identity
Expression Level


MS4A1
GNLY
CD14
FCER1A
FCGR3A
LYZ
PPBP
Identity

细胞类型注释

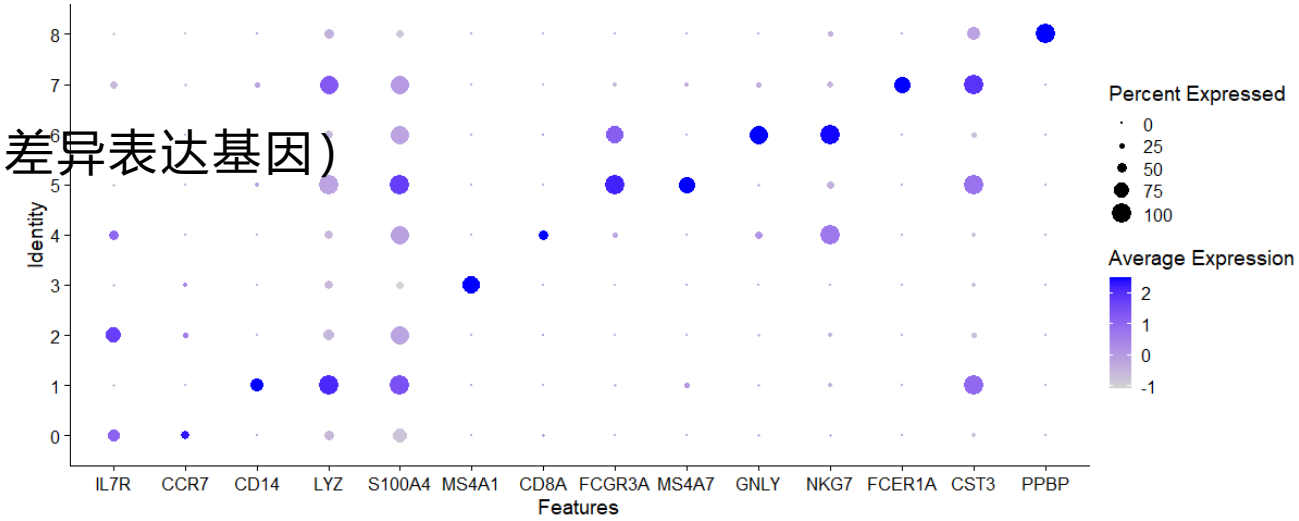
- 可能的细胞类型注释方法

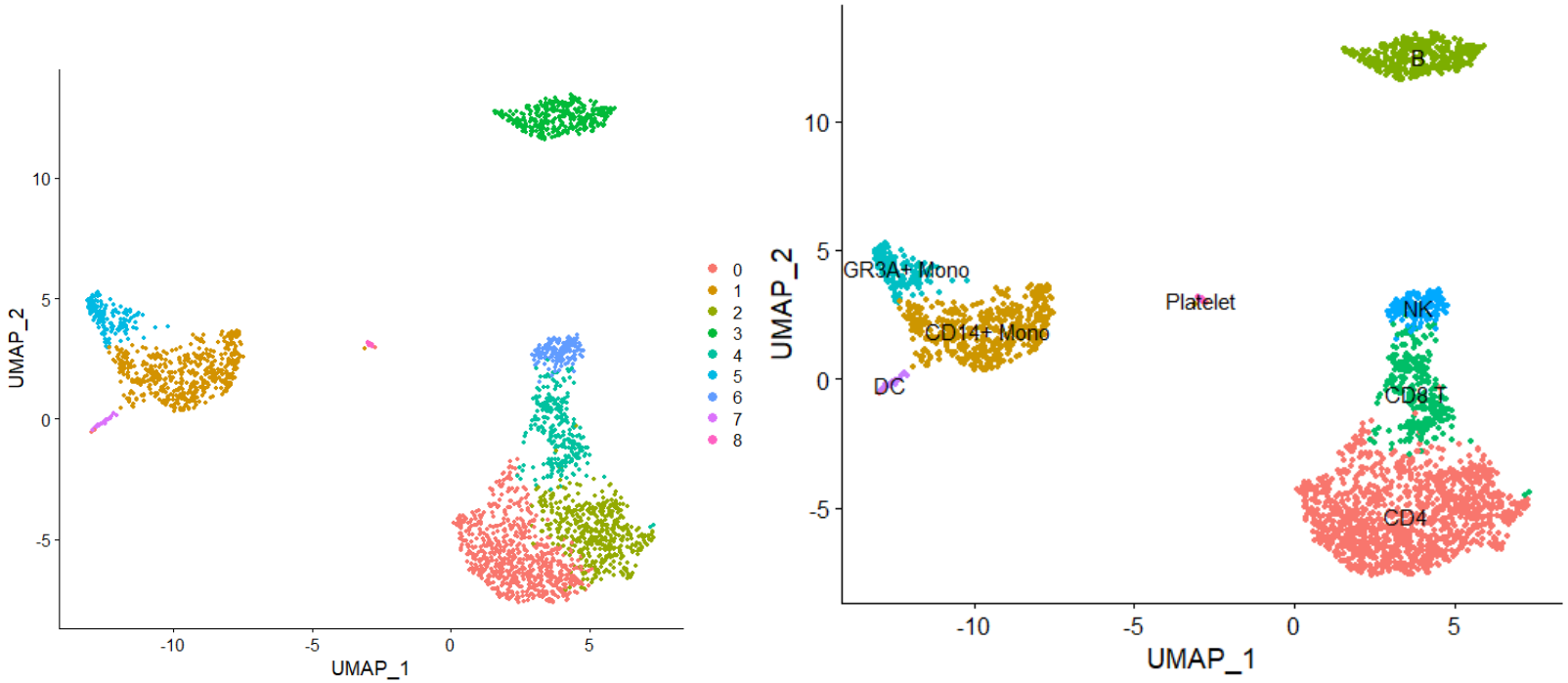


细胞类型注释

- （1）基于基因：利用标记基因（文献、数据库、差异表达基因）

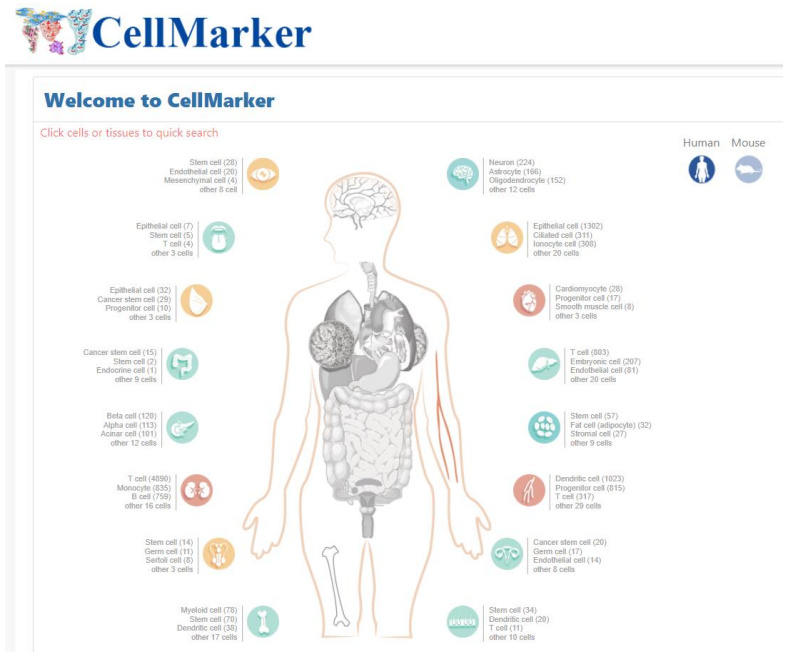
Markers	Cell Type
IL7R, CCR7	Naive CD4+ T
CD14, LYZ	CD14+ Mono
IL7R, S100A4	Memory CD4+
MS4A1	B
CD8A	CD8+ T
FCGR3A, MS4A7	FCGR3A+ Mono
GNLY, NKG7	NK
FCER1A, CST3	DC
PPBP	Platelet



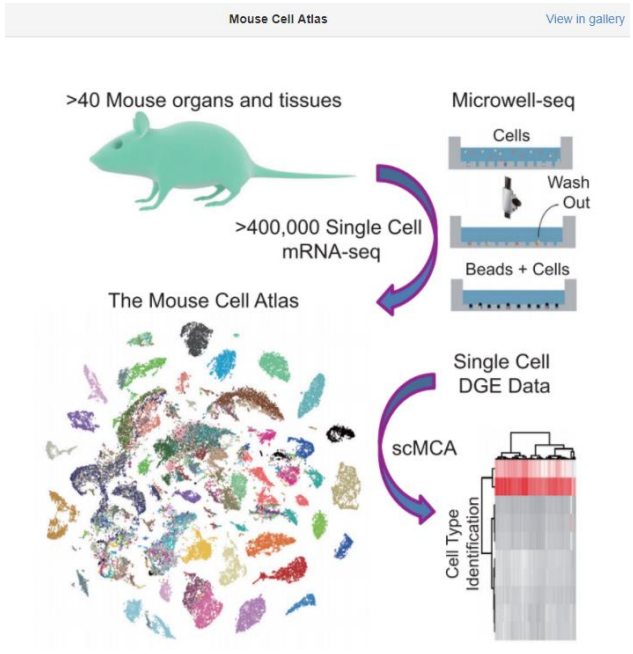


细胞类型注释

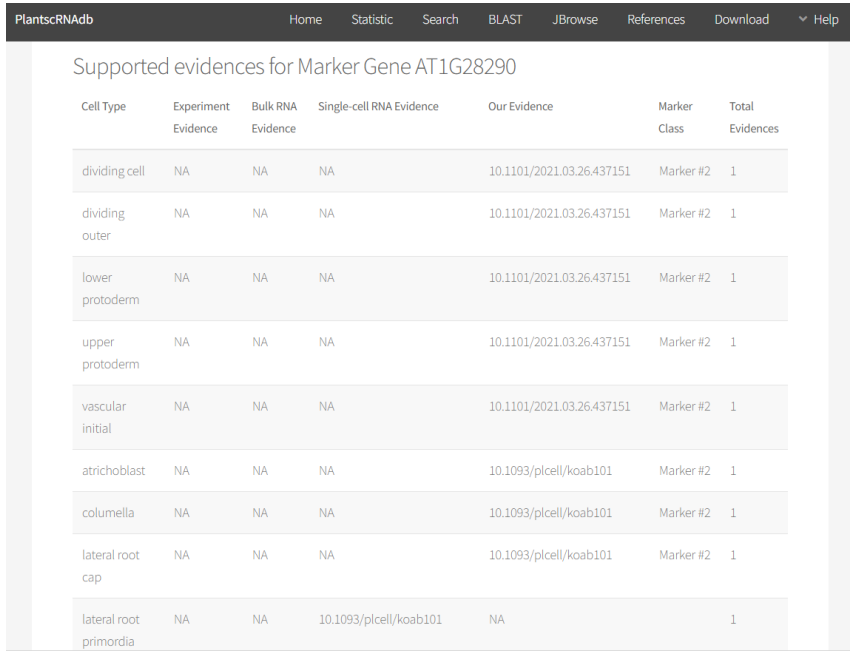
- （1）基于基因：利用标记基因（文献、数据库、差异表达基因）



通过梳理100,000+发表的文献，梳理出人的158个组织的467个细胞类型的13605个Marker基因，和鼠的81个组织的389个细胞类型的9148个Marker基因



第一版收录了小鼠近50种组织、40,0000细胞的基因表达数据；目前已经更新到第三版，>1,130,000个细胞



Cell Type	Experiment Evidence	Bulk RNA Evidence	Single-cell RNA Evidence	Our Evidence	Marker Class	Total Evidences
dividing cell	NA	NA	NA	10.1101/2021.03.26.437151	Marker #2	1
dividing outer	NA	NA	NA	10.1101/2021.03.26.437151	Marker #2	1
lower protoderm	NA	NA	NA	10.1101/2021.03.26.437151	Marker #2	1
upper protoderm	NA	NA	NA	10.1101/2021.03.26.437151	Marker #2	1
vascular initial	NA	NA	NA	10.1101/2021.03.26.437151	Marker #2	1
atrichoblast	NA	NA	NA	10.1093/plcell/koab101	Marker #2	1
columella	NA	NA	NA	10.1093/plcell/koab101	Marker #2	1
lateral root cap	NA	NA	NA	10.1093/plcell/koab101	Marker #2	1
lateral root primordia	NA	NA	10.1093/plcell/koab101	NA		1

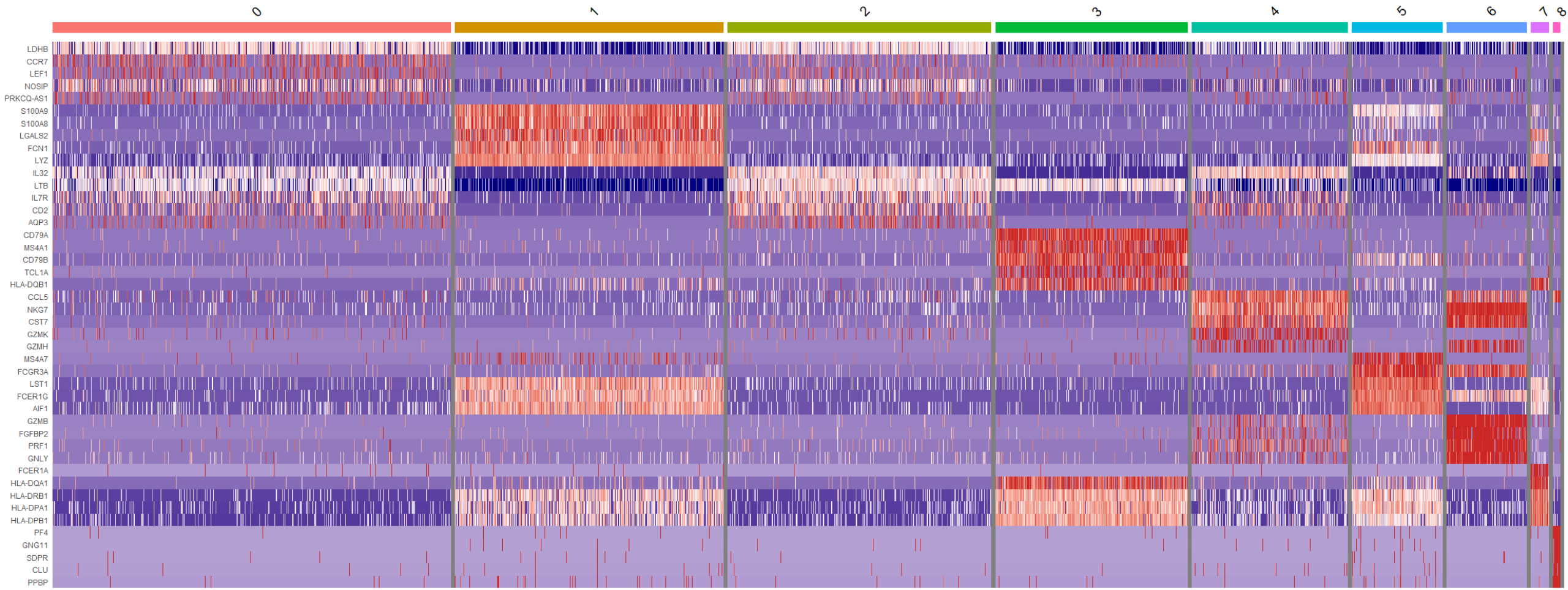
植物单细胞标记基因数据库，包括8个物种、近7万个标记基因

http://xteam.xbio.top/CellMarker/
http://bis.zju.edu.cn/MCA/
https://data.humancellatlas.org/
http://ibi.zju.edu.cn/plantscrnadb/

细胞类型注释

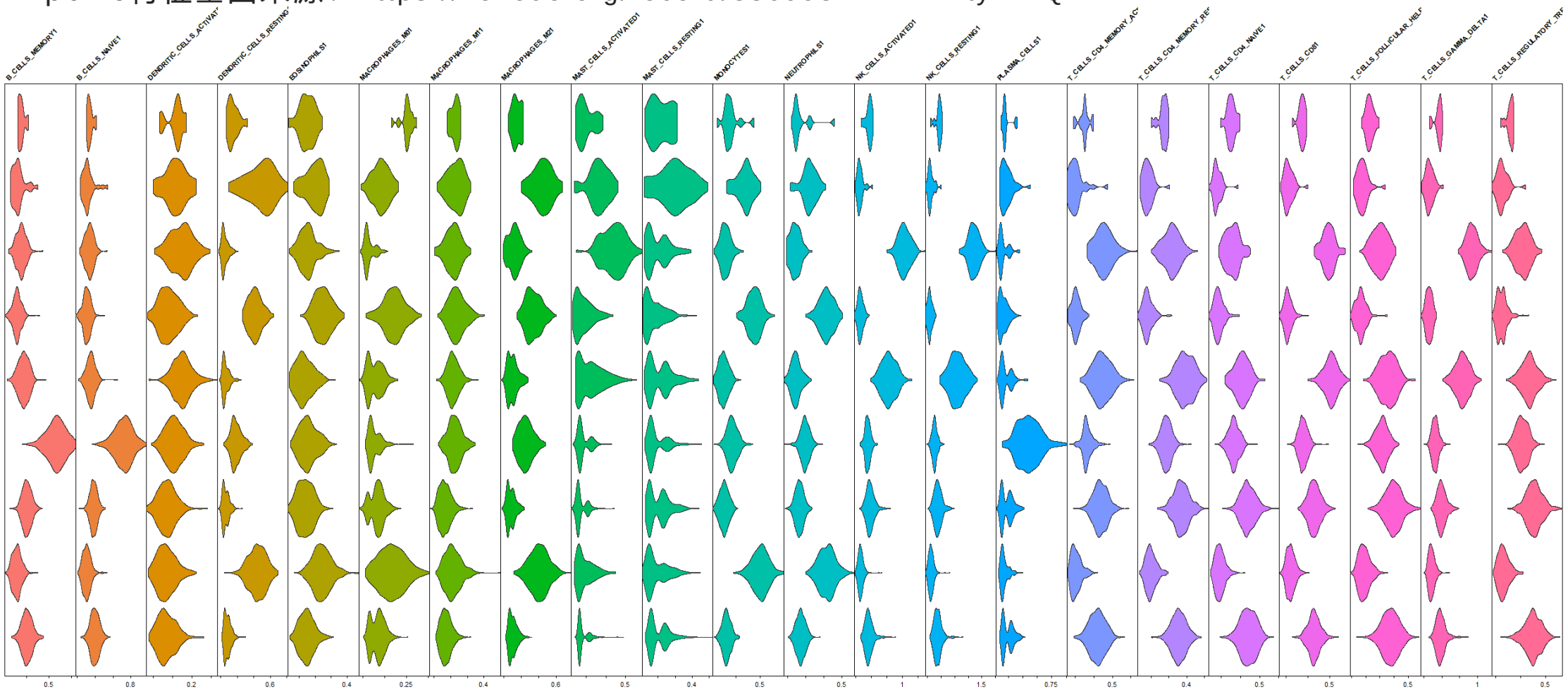
- （1）基于基因：利用类群差异基因注释

```
pbmc.markers <- FindAllMarkers(pbmc, only.pos = TRUE, min.pct = 0.25, logfc.threshold = 0.25)
```



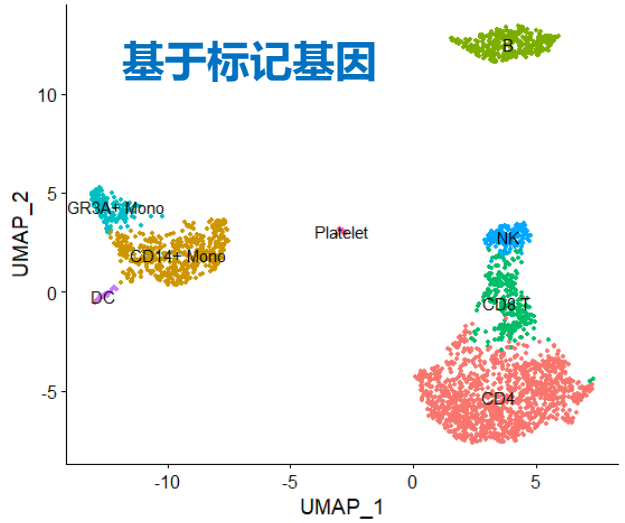
细胞类型注释

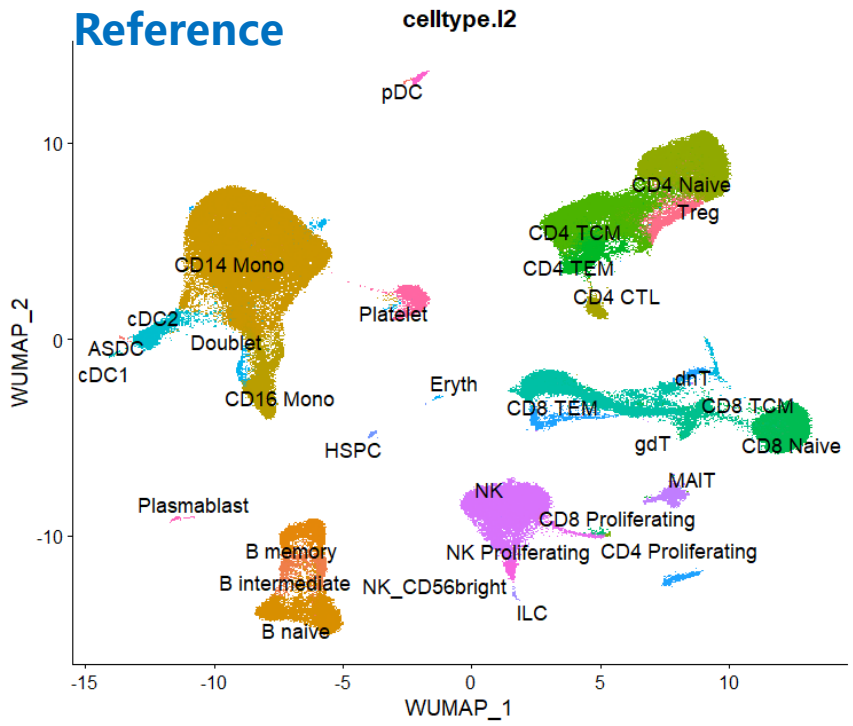
- （2）基于基因集：利用基因集富集分析打分
- pbmc特征基因来源：https://zenodo.org/record/3369934#.X2PWty2z1QI

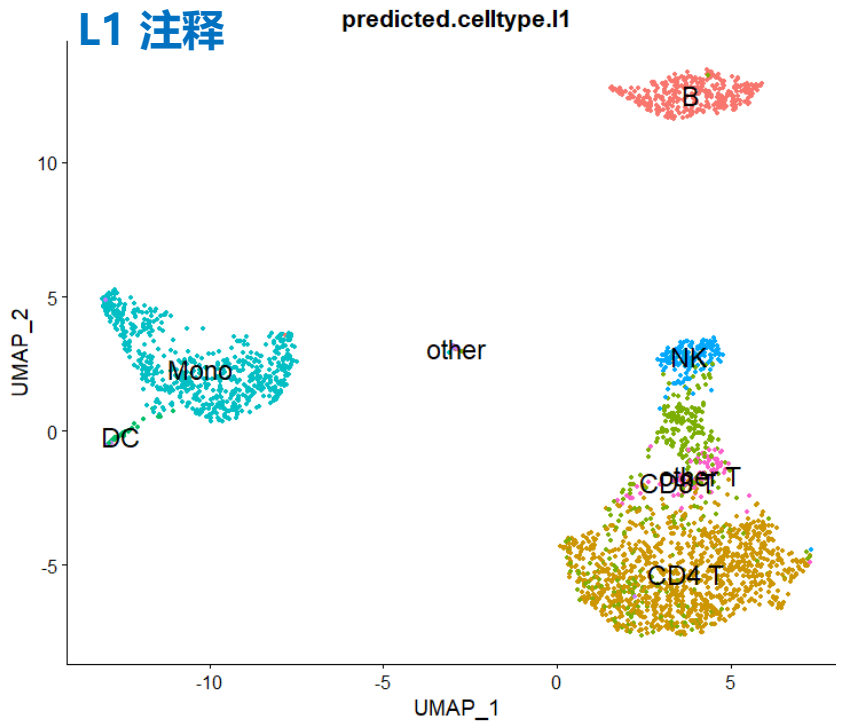


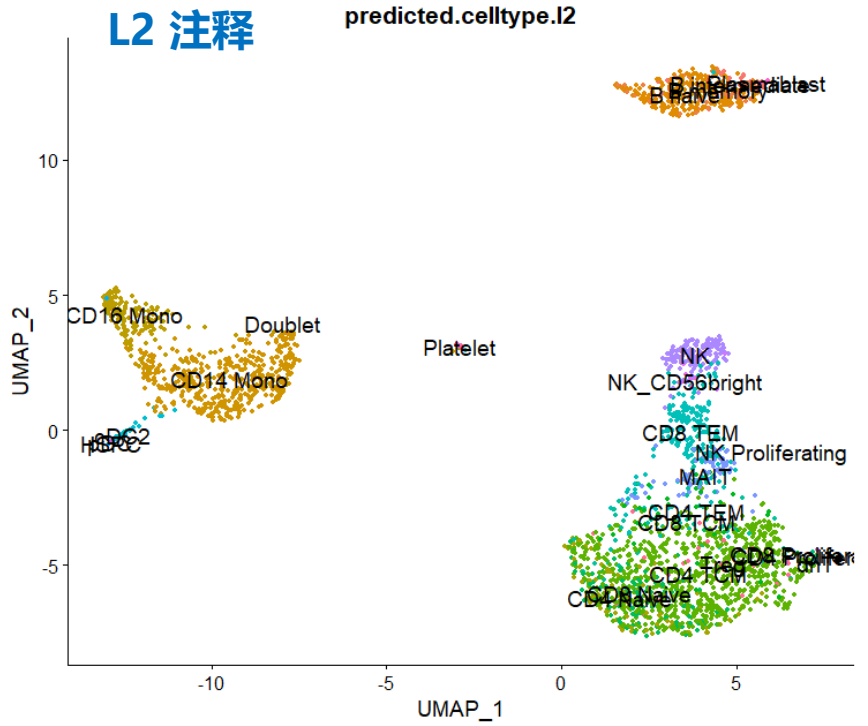
细胞类型注释

- （3）基于单细胞数据集：利用已知高质量数据集注释
- 利用Seurat中将查询数据集（query ）映射到参考数据集（references ）。参考数据集为包含228个抗体的162,000个PBMC的CITE-seq数据集







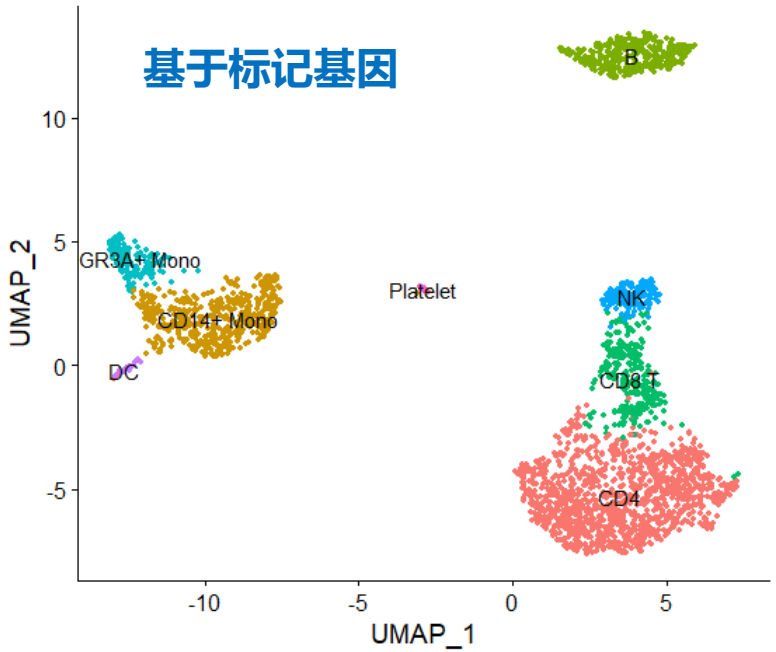


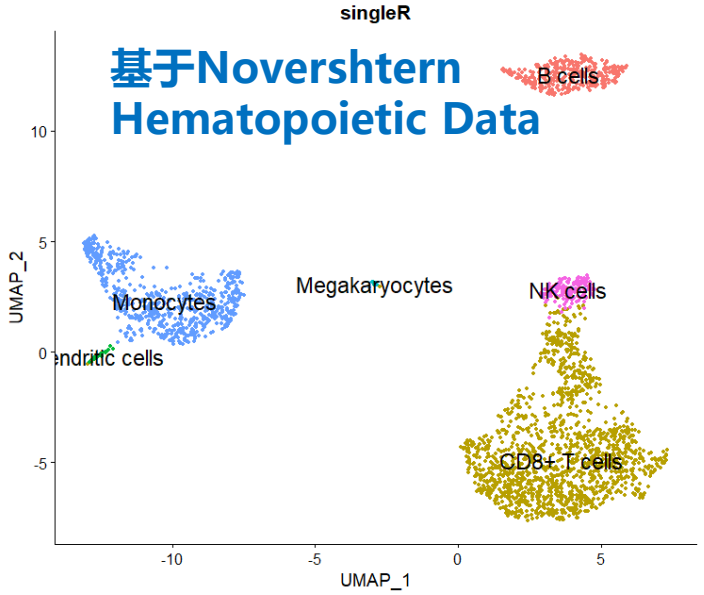
细胞类型注释

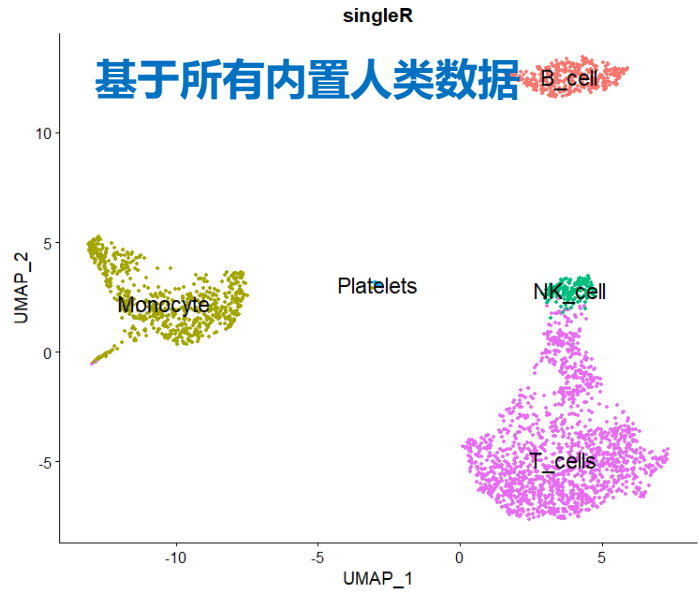
- （3）基于单细胞数据集：借助SingleR、 CellTypist等自动注释软件

SingleR在线版本：https://comphealth.ucsf.edu/app/singler

```
# human
hpca.se <- HumanPrimaryCellAtlasData()
bpe.se <- BlueprintEncodeData()
DICE <- DatabaseImmuneCellExpressionData()
NHD <- NovershternHematopoieticData()
MID <- MonacoImmuneData()
# mouse
MRD <- MouseRNAseqData()
IGD <- ImmGenData()
```

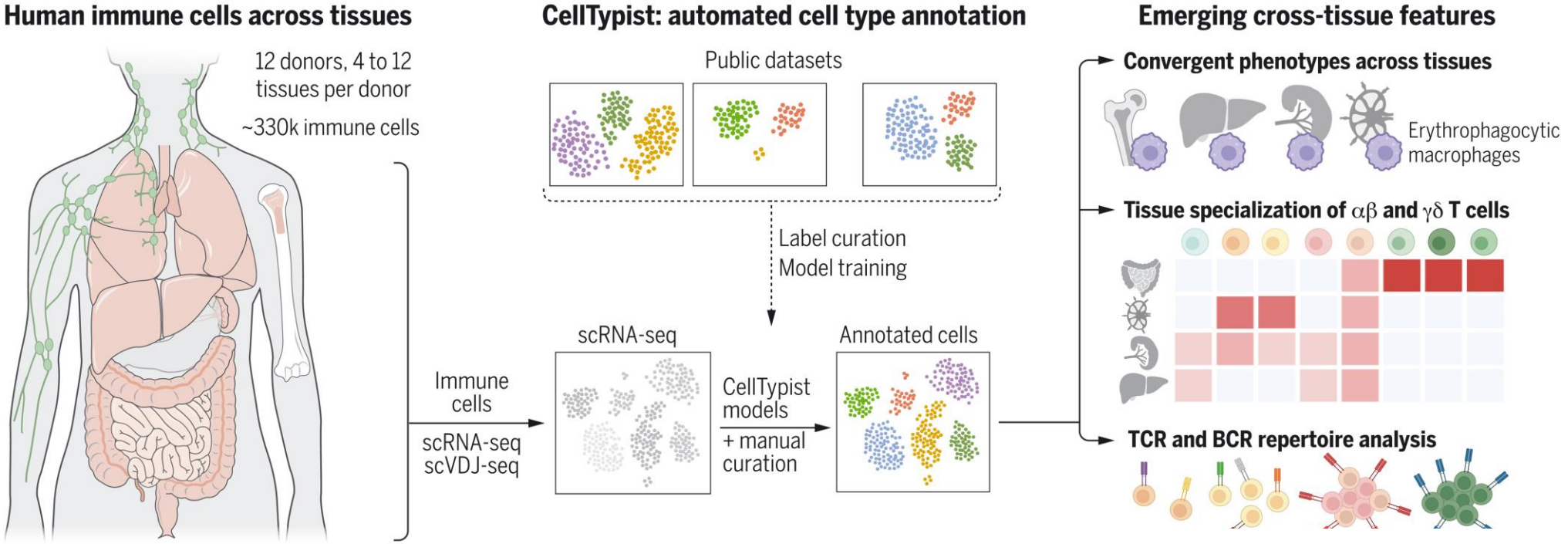






细胞类型注释

- （3）基于单细胞数据集：借助SingleR、 CellTypist等自动注释软件



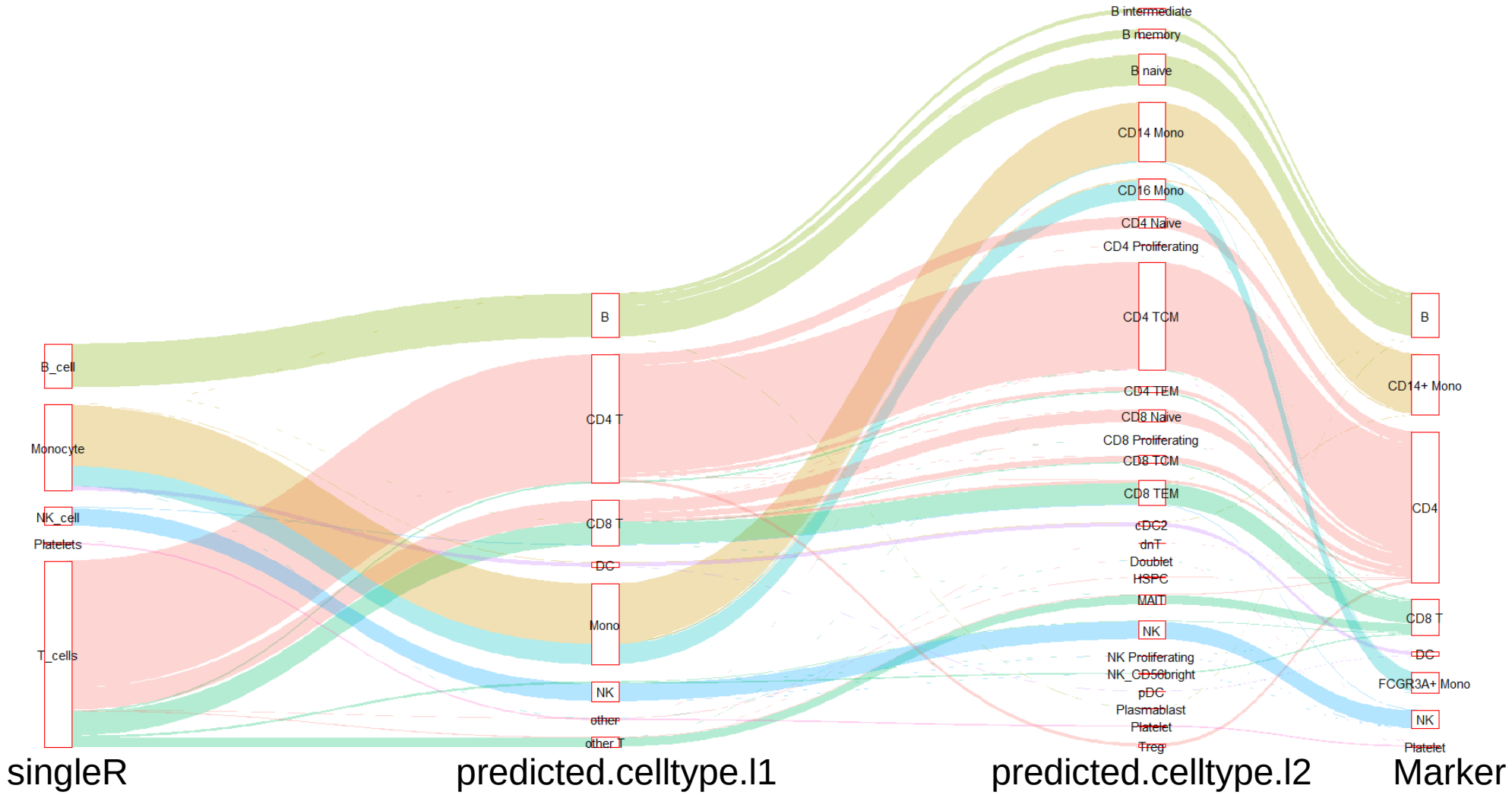

```
Skipping [1/14]: Immune_All_Low.pkl (file exists)
Skipping [2/14]: Immune_All_High.pkl (file exists)
Skipping [3/14]: Immune_All_PIP.pkl (file exists)
Skipping [4/14]: Immune_All_AddPIP.pkl (file exists)
Skipping [5/14]: Adult_Mouse_Gut.pkl (file exists)
Skipping [6/14]: COVID19_Immune_Landscape.pkl (file exists
Skipping [7/14]: Cells_Fetal_Lung.pkl (file exists)
Skipping [8/14]: Cells_Intestinal_Tract.pkl (file exists)
Skipping [9/14]: Cells_Lung_Airway.pkl (file exists)
Skipping [10/14]: Developing_Mouse_Brain.pkl (file exists)
Skipping [11/14]: Healthy_COVID19_PBMC.pkl (file exists)
Skipping [12/14]: Human_Lung_Atlas.pkl (file exists)
Skipping [13/14]: Nuclei_Lung_Airway.pkl (file exists)
Skipping [14/14]: Pan_Fetal_Human.pkl (file exists)
```

- 1. Use in the Python environment
- 2. Use as the command line
- 3. Use in the R environment
- 4. Use as Docker/Singularity container

CellTypist

细胞类型注释

- 不同注释方法得到的结果之间的差异比较



实例分析

- pbmc1k，尝试用不同的参数看结果的差异

1,223
Estimated Number of Cells

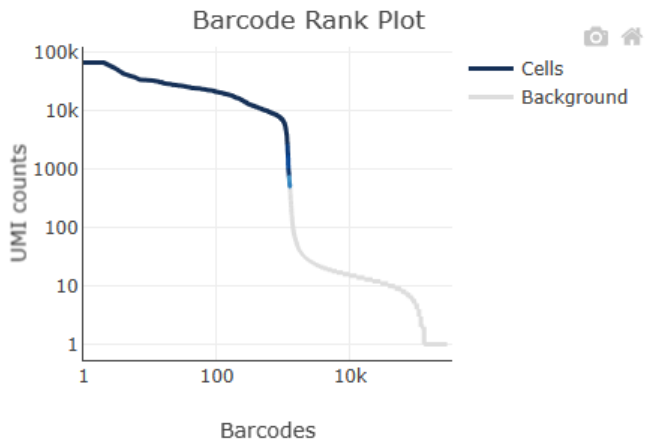
54,458
Mean Reads per Cell

3,219
Median Genes per Cell

Run Summary

Sample ID	Homo_sapiens_GRCh38
Sample Description	
Chemistry	Single Cell 3' v3
Include introns	True
Reference Path	…mo_sapiens_GRCh38
Transcriptome	Homo_sapiens_GRCh38-
Pipeline Version	cellranger-7.2.0

Cells



Estimated Number of Cells	1,223
Fraction Reads in Cells	95.5%
Mean Reads per Cell	54,458
Median UMI Counts per Cell	9,726
Median Genes per Cell	3,219
Total Genes Detected	30,001

Sequencing

Number of Reads	66,601,887
Number of Short Reads Skipped	0
Valid Barcodes	97.4%
Valid UMIs	99.9%
Sequencing Saturation	70.0%
Q30 Bases in Barcode	94.1%
Q30 Bases in RNA Read	90.2%
Q30 Bases in UMI	92.7%

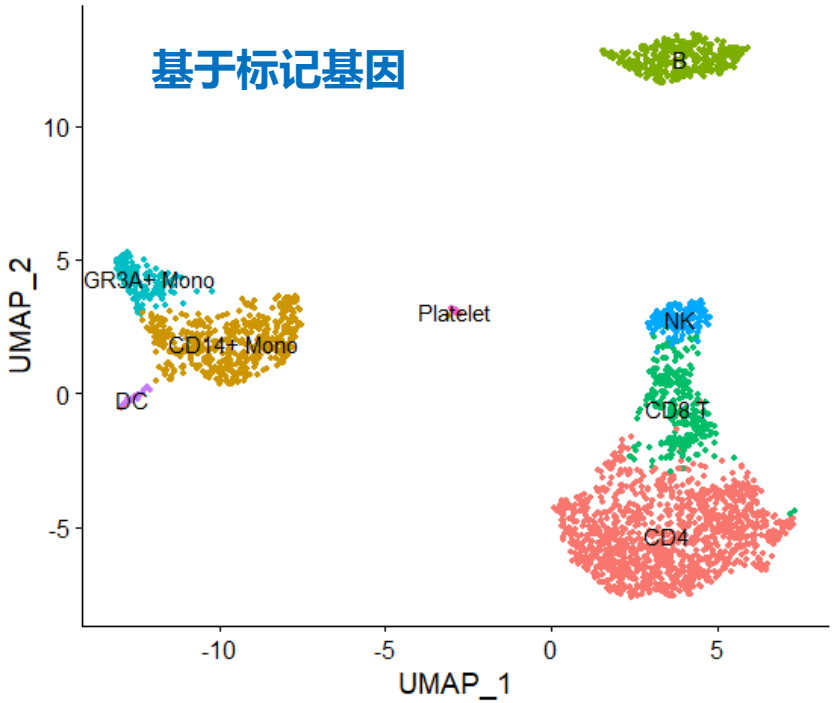
Mapping

Reads Mapped to Genome	96.1%
Reads Mapped Confidently to Genome	90.7%
Reads Mapped Confidently to Intergenic Regions	2.5%
Reads Mapped Confidently to Intronic Regions	31.0%
Reads Mapped Confidently to Exonic Regions	57.1%
Reads Mapped Confidently to Transcriptome	77.3%
Reads Mapped Antisense to Gene	7.9%

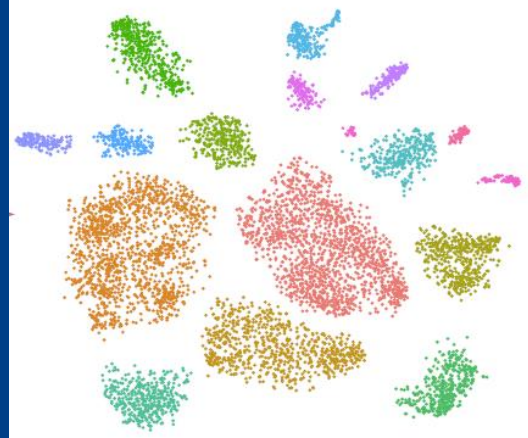
小结

	Cell1	Cell2	…	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
…	.	.	.	.
…	.	.	.	.
…	.	.	.	.
GeneM	25	0	.	0

- 质控 Quality control
  - 基因数和UMI数、线粒体比例
  - 双细胞判断、去除空液滴、去除环境RNA、细胞周期判断（optional）
- 标准化 Normalization
- 特征基因选择 Feature selection
- 中心化 Scaling
- 降维 Dimensionality reduction
- 聚类 Cluster analysis
- 细胞类型注释 Cell type annotation



浙江大学
ZHEJIANG UNIVERSITY



单细胞转录组数据分析

——从原始数据到细胞类型注释

褚琴洁　qinjiechu@zju.edu.cn

2023年10月16日