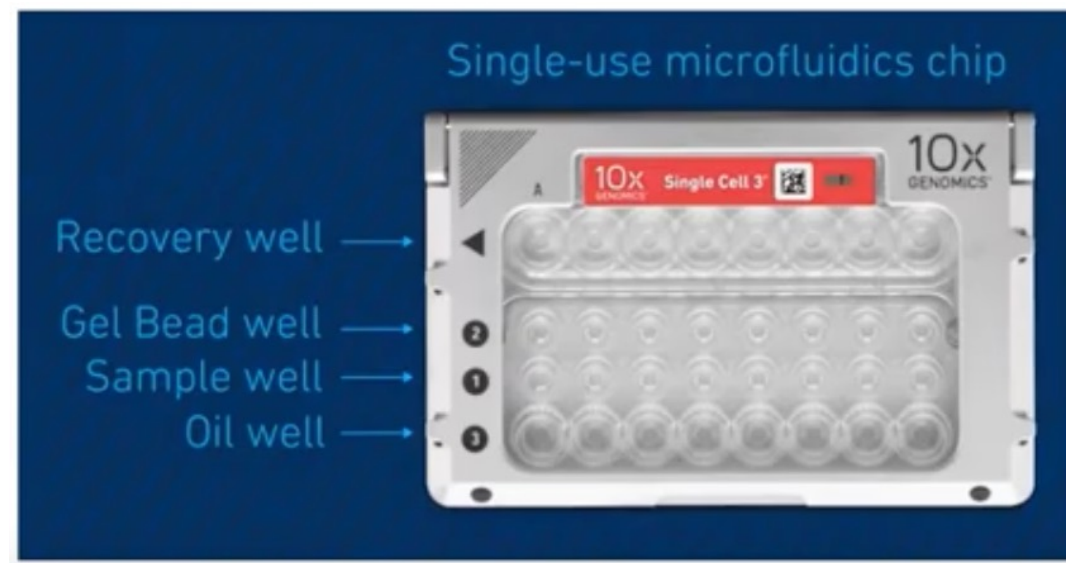


10X单细胞测序数据分析流程培训

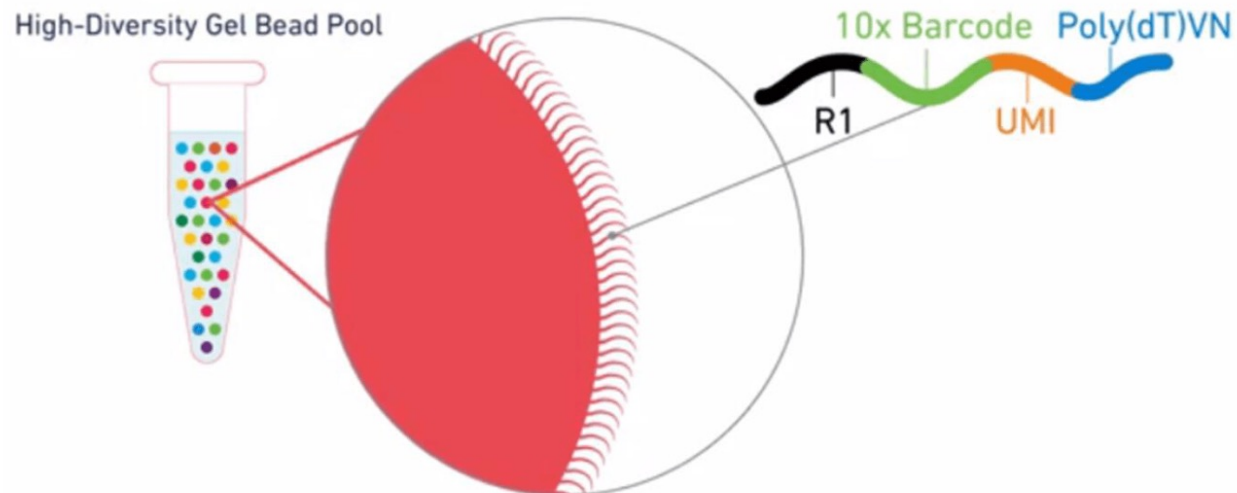
生信培训
2022年4月

10X Genomics 全基因组解码系统



2
1
3

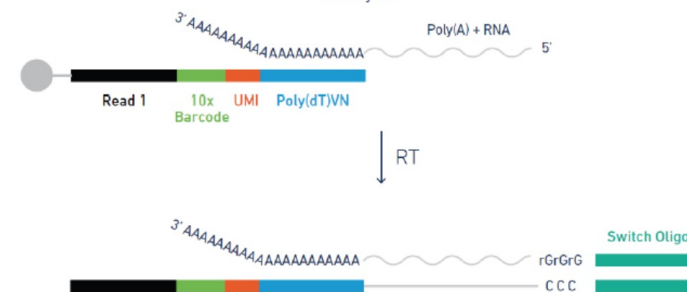
预制凝胶微珠



The Single Cell 3' v2 Assay Scheme

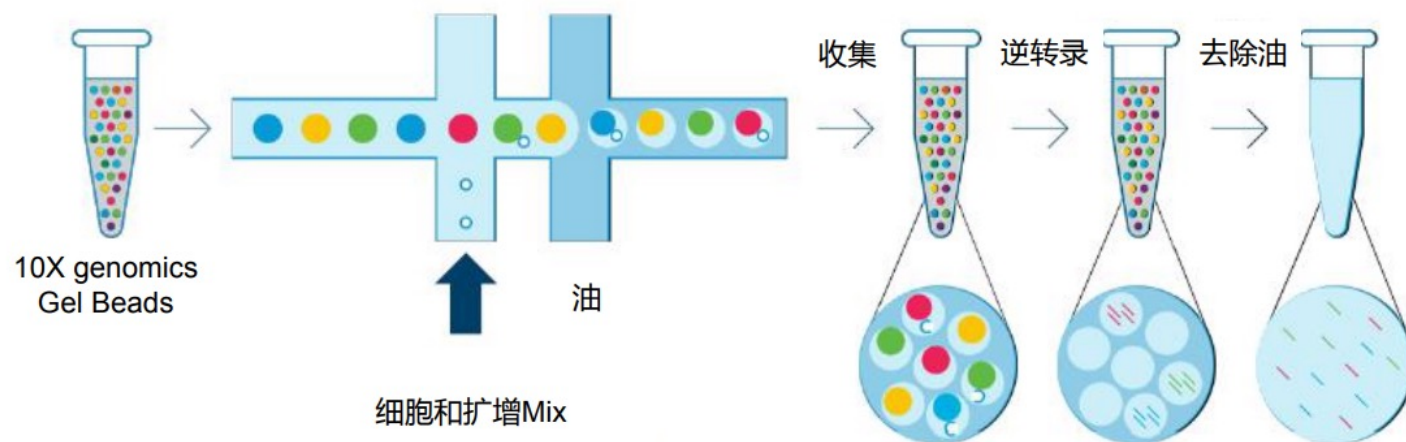
GEMs

Cell Lysis



微流控技术

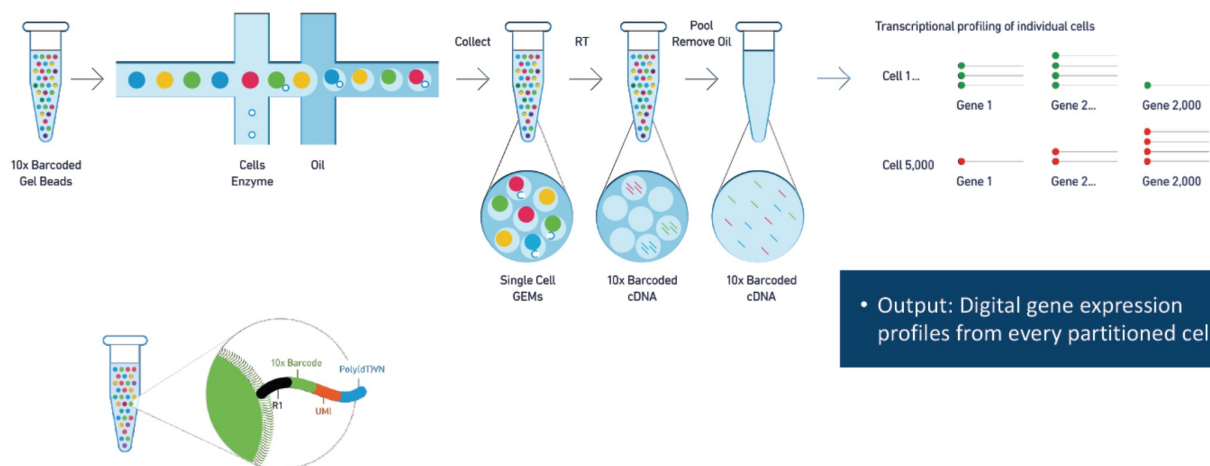
真实生成微滴过程



简而言之：一个油滴 = 一个细胞（单细胞） = 一个RNA_Seq

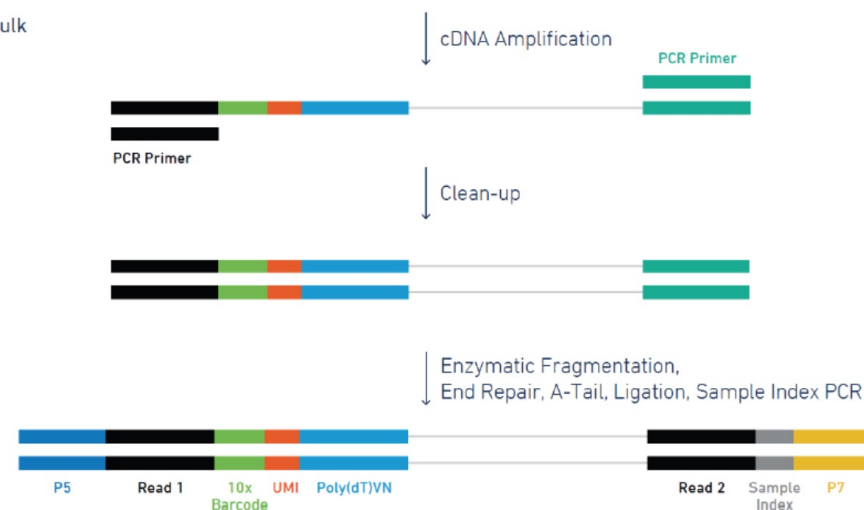
测序

Single Cell 3' Digital Gene Expression

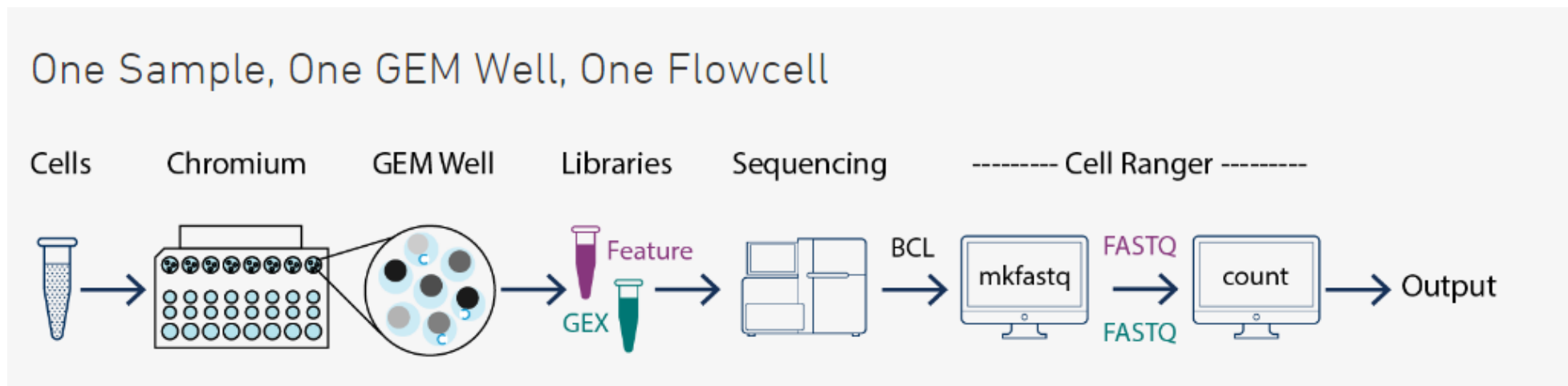


• Output: Digital gene expression profiles from every partitioned cell

Bulk



10X Genomics的单细胞测序基本流程



- 控制原始的细胞数，减少混合pool的形成
- 细胞被读到30万条reads之后，能得到的基因数量随reads数的增长变少，即被读到基因数量的增加进入平台期

测序结果

名称	修改日期	类型	大小
 MD5_MT1-ACCACGAA_BKDL190804409-1a-AK4358	2019/4/16 16:11	文本文档	1 KB
 MT1-ACCACGAA_BKDL190804409-1a-AK4358_1.clean.fq	2019/4/17 10:11	WinRAR archive	4,531,092 KB
 MT1-ACCACGAA_BKDL190804409-1a-AK4358_2.clean.fq	2019/4/17 10:08	WinRAR archive	3,756,652 KB

有两个文件 1和2 1代表了Read 1文件 2代表了Read 2文件

有两个文件，1和2，1代表了Read 1文件，2代表了Read 2文件

[illegible]

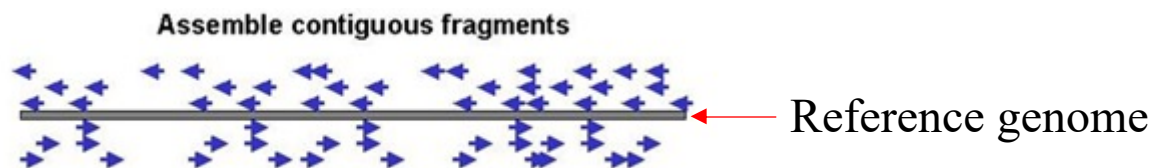
- Poly(dT) primer sequence

Read 1 file

Barcode and UMI

Cell Ranger Count的使用方式

- Cell Ranger已在集群上安装，需要开通账户；
- 首先需要有一个参考基因组，不同物种的参考基因组不同；



- 输入的FASTQ文件要符合Cell Ranger的格式要求；
- 详细声明输出的文件夹和输入的样本名。

Cell Ranger 安装

- 下载
 - `wget -O cellranger-6.1.2.tar.gz`
- 进行解包、解压缩
 - `tar -zxvf cellranger-6.1.2.tar.gz`
- 添加Cell Ranger到\$PATH，或者添加到.bashrc，下次登陆就不用再次export
 - `export PATH=~/.download/cellranger-6.1.2:$PATH`
- 检验是否添加到\$PATH
 - `which cellranger`
- Cellranger试运行
 - `cellranger testrun --id=tiny`

Cell Ranger Count 运行范例

```
#!/bin/bash
```

```
#SBATCH -J mouse_cheng
```

```
#SBATCH -n 20
```

```
#SBATCH -o run_%J
```

```
#SBATCH -e errput_%J
```

```
#SBATCH --mem=244800
```

```
#SBATCH -p earth
```

```
module load cellranger/3.0.2
```

```
cellranger count --id=sample1 \.    # 输出目录
```

```
--transcriptome=/data/DB/refdata-gex-mm10-2020-A \ #参考基因组及其所在的目录
```

```
--fastqs=/home/fastq_path \ #输入的FASTQ文件所在的目录
```

```
--sample=sample1_BKDL220007671 \ #样本名
```

**输入文件的名字必须严格
按照Cell Ranger的要求**

Cell Ranger Count 输入文件名范例

MT1-ACCACGAA_BKDL190804409-1a-AK4358			
搜索 MT1-ACCACGAA_BKDL190804409-1a-AK4358			
共享 刻录 新建文件夹			
名称	修改日期	类型	大小
MD5_MT1-ACCACGAA_BKDL190804409-1a-AK4358	2019/4/16 16:11	文本文档	1 KB
MT1-ACCACGAA_BKDL190804409-1a-AK4358_1.clean.fq	2019/4/17 10:11	WinRAR archive	4,531,092 KB
MT1-ACCACGAA_BKDL190804409-1a-AK4358_2.clean.fq	2019/4/17 10:08	WinRAR archive	3,756,652 KB

拿到手的测序结果，fq结尾的压缩文件

MT1-ACCACGAA_BKDL190804409-1a-AK4358_1.clean.fq
MT1-ACCACGAA_BKDL190804409-1a-AK4358_2.clean.fq

必须更名为
右边这个格
式的文件名

MT1_S1_L001_R1_001.fastq.gz
MT1_S1_L001_R2_001.fastq.gz

MT1为样本名

S1 代表了来自同以一个sample

L001 代表了来自同一个lane

R1和R2 代表了Read1和Read2

001.fastq.gz 是必须的fastq文件后缀名（.fq文件后缀，Cell Ranger不认）

中间用下划线隔开（MT1_S1_L001_R1_001.fastq.gz）

多lane处理

```
k1th-k1th_BKDL220007669-1a-AK4712_1.clean.fq.gz  
k1th-k1th_BKDL220007669-1a-AK4712_2.clean.fq.gz  
MD5.txt
```

```
k1th-k1th_BKDL220007669-1a-AK4713_1.clean.fq.gz  
k1th-k1th_BKDL220007669-1a-AK4713_2.clean.fq.gz  
MD5.txt
```

```
k1th-k1th_BKDL220007669-1a-AK4714_1.clean.fq.gz  
k1th-k1th_BKDL220007669-1a-AK4714_2.clean.fq.gz  
MD5.txt
```

```
k1th-k1th_BKDL220007669-1a-AK4715_1.clean.fq.gz  
k1th-k1th_BKDL220007669-1a-AK4715_2.clean.fq.gz  
MD5.txt
```

k1th_BKDL220007669	S1	L001	R1_001.fastq.gz
k1th_BKDL220007669	S1	L001	R2_001.fastq.gz
k1th_BKDL220007669	S1	L002	R1_001.fastq.gz
k1th_BKDL220007669	S1	L002	R2_001.fastq.gz
k1th_BKDL220007669	S1	L003	R1_001.fastq.gz
k1th_BKDL220007669	S1	L003	R2_001.fastq.gz
k1th_BKDL220007669	S1	L004	R1_001.fastq.gz
k1th_BKDL220007669	S1	L004	R2_001.fastq.gz

样本名
--sample

说明
是同
一个
样本

不同
lane

R1和
R2 代
表了
Read1
和
Read2

001.fastq.gz
是必须的
fastq文件后
缀名

输出文件

```
_cmdline
_filelist
_finalstate
_invocation
_jobmode
k1th_BKDL220007669.mri.tgz
_log
_mrosource
outs
_perf
SC_RNA_COUNTER_CS
_sitecheck
_tags
_timestamp
_uuid
_vdrkill
_versions
```



```
analysis
cloupe.cloupe
filtered_feature_bc_matrix
filtered_feature_bc_matrix.h5
metrics_summary.csv
molecule_info.h5
possorted_genome_bam.bam
possorted_genome_bam.bam.bai
raw_feature_bc_matrix
raw_feature_bc_matrix.h5
web_summary.html
```

- analysis: 数据分析目录, 下面又包含聚类clustering (有graph-based & k-means)、差异分析diffexp、主成分线性降维分析pca、非线性降维tsne

- cloupe.cloupe: 官方可视化工具Loupe Cell Browser 输入文件

- filtered_gene_bc_matrices: 是重要的一个目录, 下面又包含了 barcodes.tsv.gz、features.tsv.gz、matrix.mtx.gz, 是下游Seurat、Scater、Monocle等分析的输入文件

- filtered_feature_bc_matrix.h5: 过滤掉的barcode信息HDF5 format

- metrics_summary.csv: CSV格式数据摘要

- molecule_info.h5: 下面进行aggregate使用的文件

- possorted_genome_bam.bam: 比对文件

- possorted_genome_bam.bam.bai: 索引文件

- raw_feature_bc_matrix: 原始barcode信息

- raw_feature_bc_matrix.h5: 原始barcode信息HDF5 format

- web_summary.html: 官方说明 summary HTML file

web_summary.html

该网页的结果分成了summary和analysis两部分， summary部分包含如下结果：

1. 有问题，会有黄色提醒：

The analysis detected some issues. [Details »](#)

Estimated Number of Cells

6,852

Mean Reads per Cell

37,914

Median Genes per Cell

2,304

Sequencing

Number of Reads	259,789,692
Valid Barcodes	97.6%
Sequencing Saturation	51.5%
Q30 Bases in Barcode	96.1%
Q30 Bases in RNA Read	91.6%
Q30 Bases in UMI	93.4%

Mapping

Reads Mapped to Genome	95.2%
Reads Mapped Confidently to Genome	92.9%
Reads Mapped Confidently to Intergenic Regions	6.0%
Reads Mapped Confidently to Intronic Regions	29.0%
Reads Mapped Confidently to Exonic Regions	57.9%
Reads Mapped Confidently to Transcriptome	54.4%
Reads Mapped Antisense to Gene	1.6%

Cells

Estimated Number of Cells	6,852
Fraction Reads in Cells	90.4%
Mean Reads per Cell	37,914
Median Genes per Cell	2,304
Total Genes Detected	21,651
Median UMI Counts per Cell	5,950

Sample

Name	w1th_BKDL220007668
Description	
Transcriptome	mm10
Chemistry	Single Cell 3' v3
Cell Ranger Version	3.0.2

Cell ranger 结果解读1

■ 细胞和基因数目的评估：

- *Estimated number of cells* - 样本测到的细胞数
- *Mean reads per cell* - 每个细胞测到的平均reads
- *Median genes per cell* - 每个细胞基因数的中位数

Estimated Number of Cells

6,852

Mean Reads per Cell

37,914

Median Genes per Cell

2,304

■ Sequencing中：

- *Number of reads* - 测到的总read数目
- *Valid barcodes* - UMI校正后匹配的UMI数量
- *Sequencing saturation*：测序饱和度
- *Q30 bases in barcode* - 基于barcode的分数，大于30的比率
- *Q30 bases in RNA read* - 基于RNA read的分数，大于30的比率
- *Q30 bases in UMI* - 基于UMI的分数，大于30的比率

Sequencing

Number of Reads	259,789,692
Valid Barcodes	97.6%
Sequencing Saturation	51.5%
Q30 Bases in Barcode	96.1%
Q30 Bases in RNA Read	91.6%
Q30 Bases in UMI	93.4%

Cell ranger 结果解读2

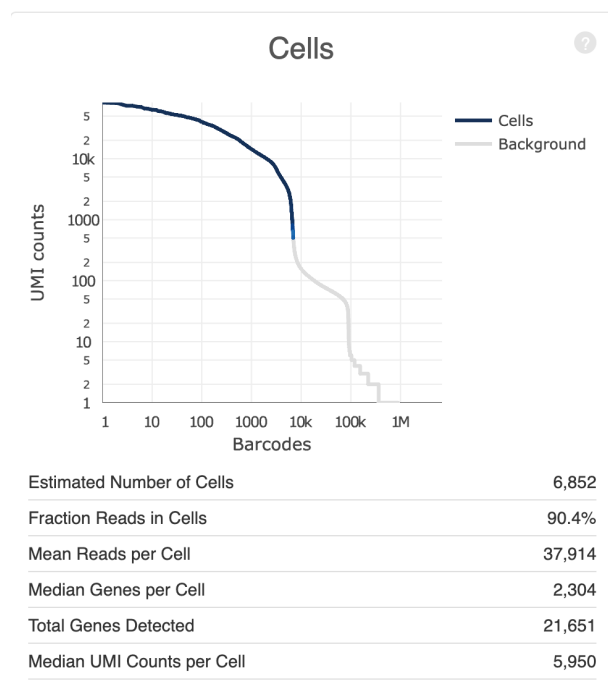
- 比对比例统计，统计reads的比对比例，同时给出比对到基因间区，外显子，内含子的比例

Mapping	
Reads Mapped to Genome	95.2%
Reads Mapped Confidently to Genome	92.9%
Reads Mapped Confidently to Intergenic Regions	6.0%
Reads Mapped Confidently to Intronic Regions	29.0%
Reads Mapped Confidently to Exonic Regions	57.9%
Reads Mapped Confidently to Transcriptome	54.4%
Reads Mapped Antisense to Gene	1.6%

- Reads mapped to genome - 比对到选定基因组的reads
- Reads mapped confidently to genome - 仅仅比对到基因组的reads，如果一条reads既可以比对到外显子区又可以比对到非外显子区，那么算比对到了其中一个外显子区
- Reads mapped confidently to intergenic regions - 比对到基因组的基因间区域
- Reads mapped confidently to intronic regions - 比对到内含子区域
- Reads mapped confidently to exonic regions - 比对到外显子区域
- Reads mapped confidently to transcriptome - 比对到转录组的reads，这些读数可以用来UMI的计数
- Reads mapped antisense to gene - 比对到基因的相反的reads

Cell ranger 结果解读3

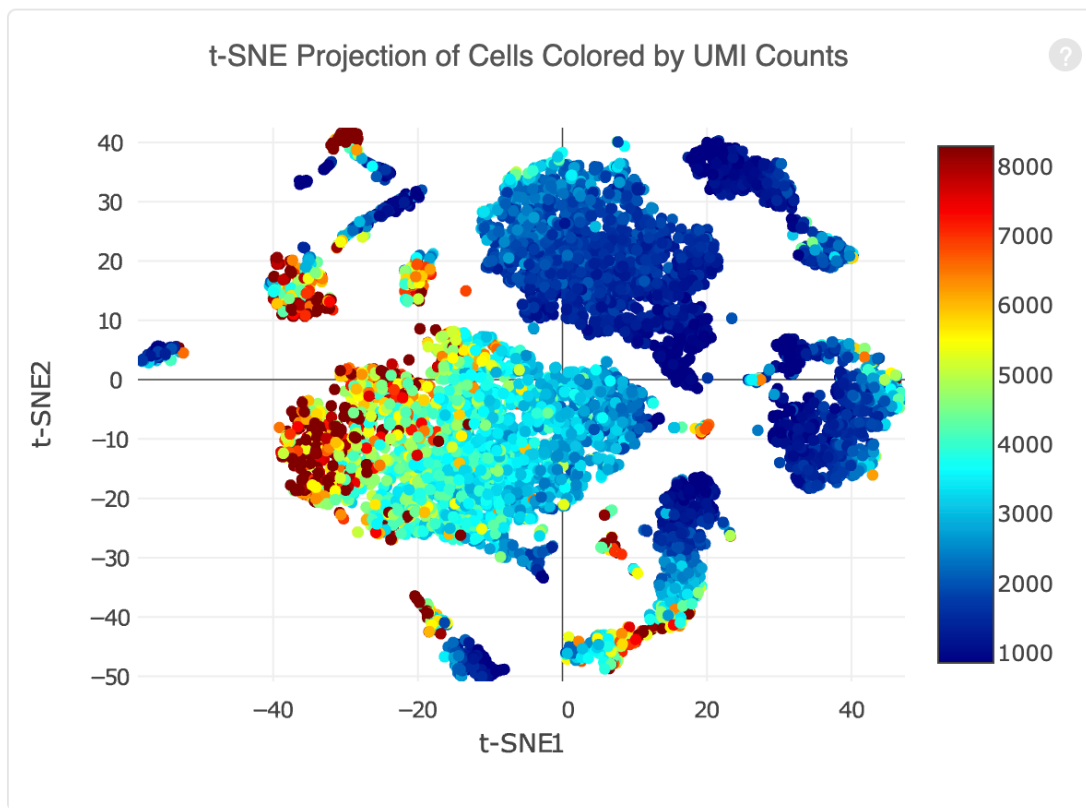
■ 细胞数目评估详细信息



- *Estimated number of cells* - 样本测到的细胞数
- *Fraction reads in cells* - valid-UMI的质量分数，代表与细胞相关的UMI可靠地比对到基因组，一般要在70%及以上，否则数据质量就不好
- *Mean reads per cell* - 每个细胞测到的平均reads
- *Median genes per cell* - 每个细胞的中间基因数
- *Total genes detected* - 测到的总基因数,至少有一条UMI
- *Median UMI counts per cell* - 细胞UMI数量的中间值

Analysis部分

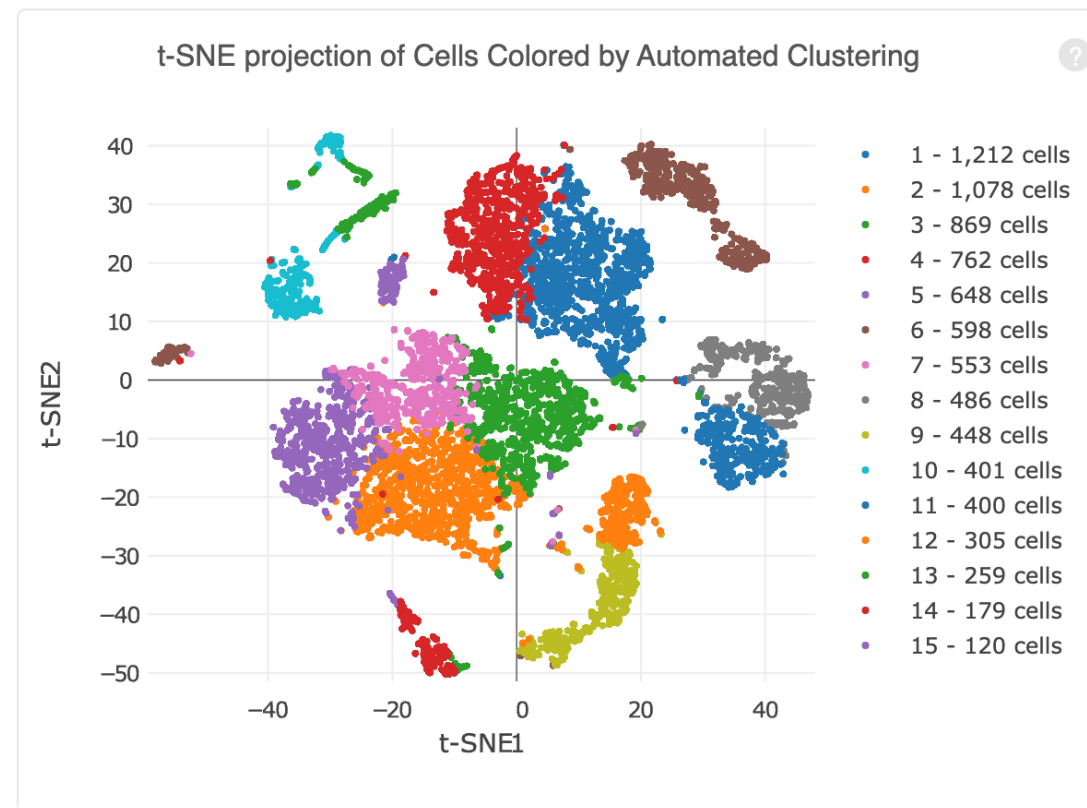
细胞表达量分布的t-SNE图



细胞亚型

Clustering Type:

Graph-based



Analysis部分

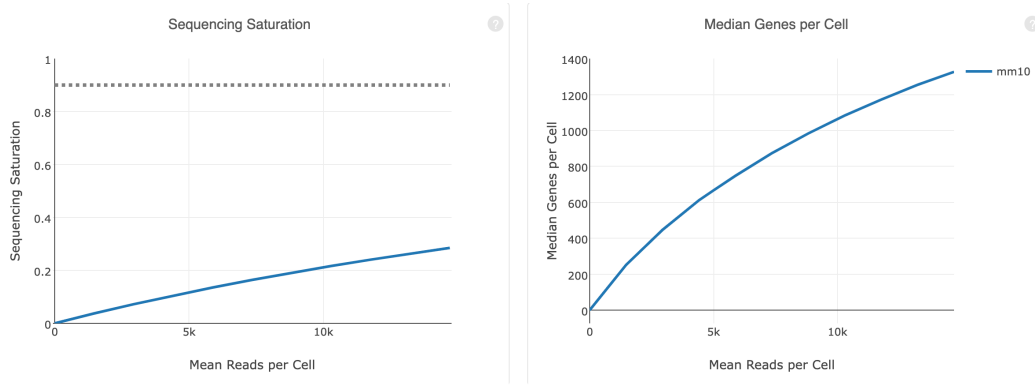
■ 基因差异表达分析

- 对cluster下的基因进行差异分析，将细胞分成了该cluster和其他cluster两类，然后进行差异分析

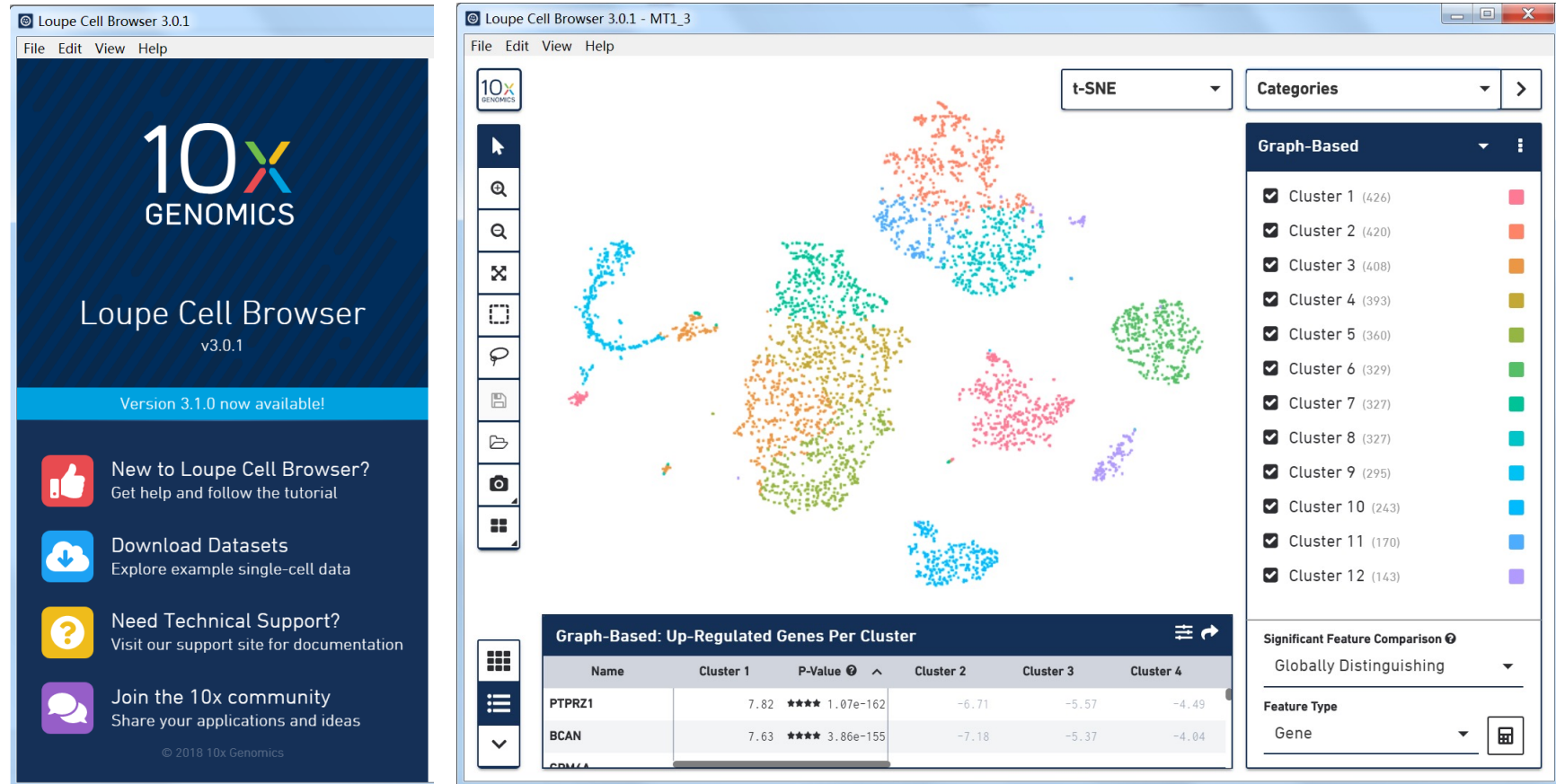
Top Genes By Cluster (Log2 fold-change, p-value)																	
Gene ID	Gene name	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		Cluster 7		Cluster 8	
		L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value
ENSMUSG00000022132	Cldn10	3.92	2e-31	-4.13	3e-16	-3.94	7e-09	3.20	5e-18	-4.94	4e-24	-2.52	1e-01	-4.42	2e-11	-3.38	8e-03
ENSMUSG00000027574	Nkain4	3.89	2e-30	-4.23	5e-16	-3.53	1e-07	3.07	4e-16	-4.65	9e-22	-2.95	7e-02	-3.98	6e-10	-3.29	1e-03
ENSMUSG00000031980	Agt	3.84	2e-30	-4.58	2e-19	-4.38	2e-10	3.31	9e-20	-5.26	1e-27	-3.15	4e-02	-4.42	3e-12	-2.92	2e-03
ENSMUSG00000000202	Btbd17	3.80	1e-28	-4.30	4e-16	-3.48	2e-07	3.21	6e-18	-5.30	3e-24	-2.43	2e-01	-4.15	3e-10	-2.73	4e-03
ENSMUSG00000048001	Hes5	3.78	7e-26	-4.89	5e-16	-4.31	3e-08	3.24	7e-17	-4.78	3e-19	-3.60	4e-02	-4.84	2e-10	-2.80	5e-03

■ 饱和度评估

- 对reads抽样，观察不同抽样条件下检测到的转录本数量占检测到的所有转录本的比例，并绘制如下曲线
- 对reads抽样，观察不同测序数据量情况下检测到的基因数目的分布，并绘制如下曲线



10X genomics 的 Loupe Cell Browser



图形化界面易于操作，使用cloupe.cloupe文件来展示。

<https://support.10xgenomics.com/single-cell-vdj/software/vdj-and-gene-expression/latest/lcb>