BerryGenomics
贝瑞基因

scRNA-Seq (Part 1)

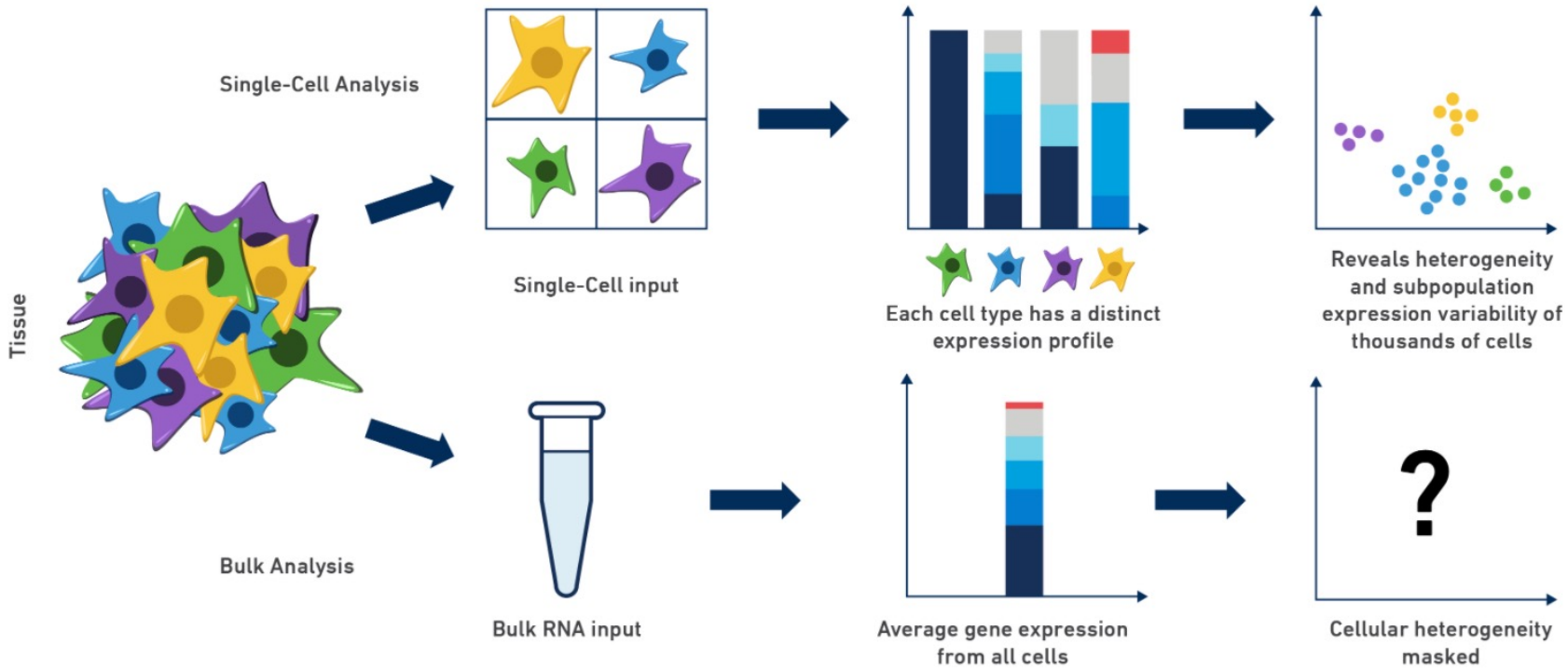
王　鹏

科技服务事业部

1

scRNA-Seq前期流程

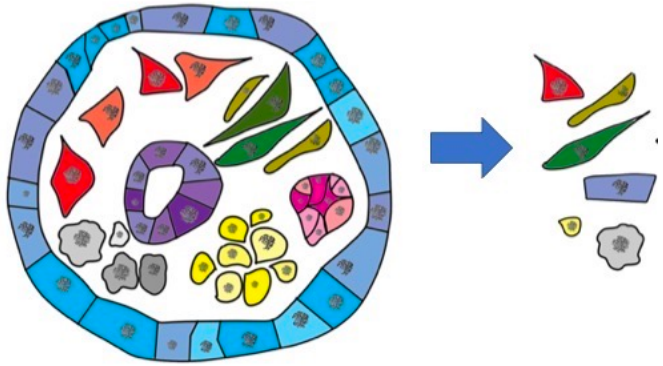
scRNA-Seq VS RNA-Seq



https://www.biocompare.com/Bench-Tips/345311-Single-Cell-Set-Up-Sample-Preparation-Tips/

1. Amplification bias; 2. Drop-out rates; 3. Background noise; 4. batch effects, 5. Bias due to cell-cycle, size, etc.

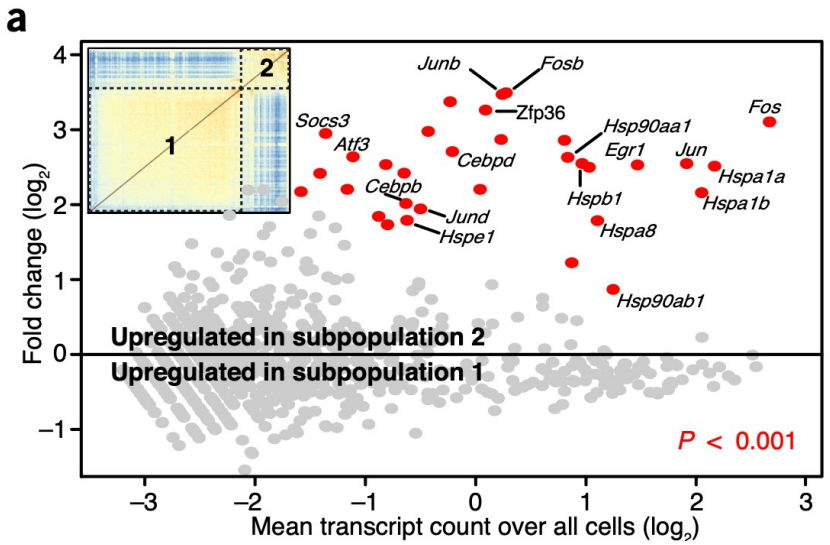
scRNA-Seq (Cell dissociation)



细胞活性和完整性

- 不完整解离导致细胞成团【多细胞的影响】
- 太重的解离会损伤细胞【细胞碎片/RNA泄露】

- 解离影响细胞类群的鉴定
- 解离会引入转录的变化



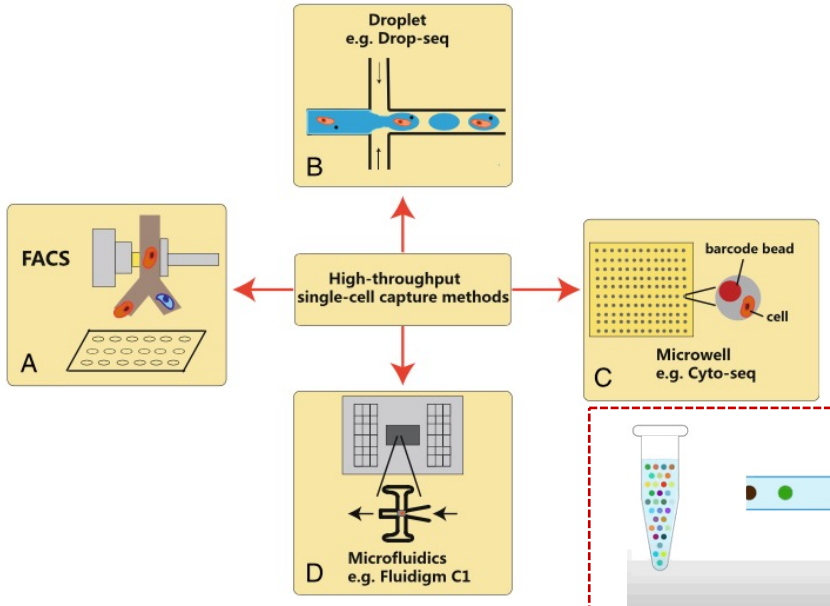
10.1038/nmeth.4437

scRNA-Seq (Cell capture)

10.1186/s13045-017-0401-7

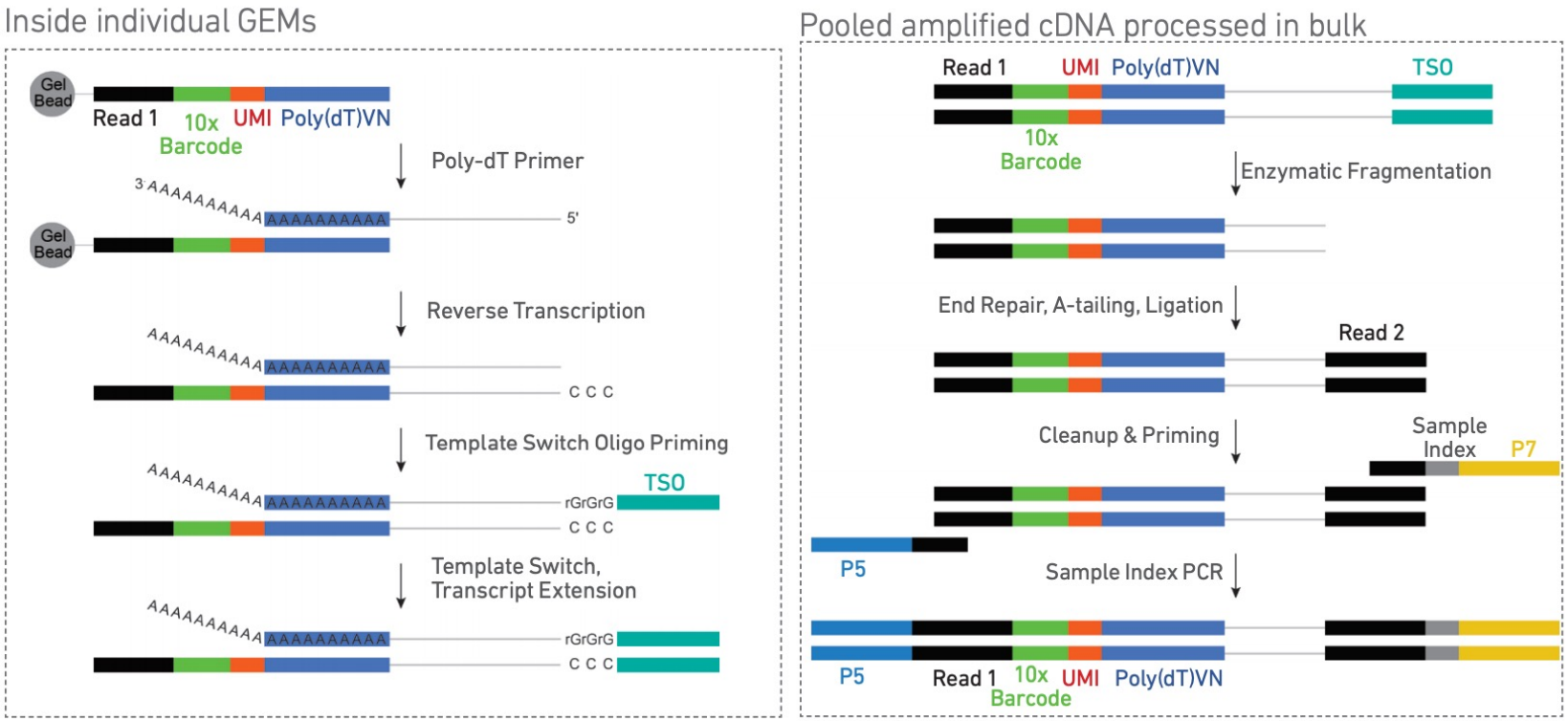
Table 2 The advances of single-cell capture methods

Methods	Advantage	Drawback	Application
Mouth pipetting	Low cost	Time consuming	Rare sample
Laser capture microdissection	Visualization	Time consuming	Specific target
Flow cytometry	Marker selection	Require sorting	MARS-seq
Microwell platform	High throughput	mRNA capture rate	Cyto-seq
Microdroplet platform	High throughput	mRNA capture rate	Drop-seq, inDrop
Fluidigm C1 platform	Automatic library prep	High cost	qPCR, mRNA-seq
DEPArray	Visualization	High cost	Specific target



- 空孔或空微滴（empty droplets）/双细胞或多细胞
- 对细胞大小的选择具有偏好【30μm, ~7-60 um】
- 对细胞类型的选择具有偏好性
- 长时间分选捕获可能会造成细胞损伤

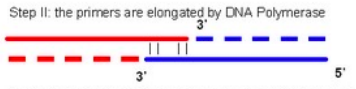
scRNA-Seq (Cell lysis and Library preparation)



裂解的影响：1）过度的裂解会影响建库的质量；2）裂解不足会导致核RNA释放不完整；3）不同细胞类型的裂解条件不一样

反转录效率： Drop-out rate

扩增的影响：扩增的偏好性和扩增的效率，以及引物二聚体的影响



Step1：**细胞裂解**，微滴捕获mRNA；Step2：**反转录**（MMLV：莫罗尼鼠白血病病毒）并加Cs；Step3：加TSO（template switching oligs）合成第二链（cDNA）；Step4：加入cDNA的前后引物，扩增出全长cDNA；Step5：使用片段酶将cDNA打断并尾部加A；Step6：加入Illumina Truseq adaptor进行连接；Step7：加入库**PCR引物**，扩增建库 https://teichlab.github.io/scg_lib_structs/methods_html/10xChromium3.html

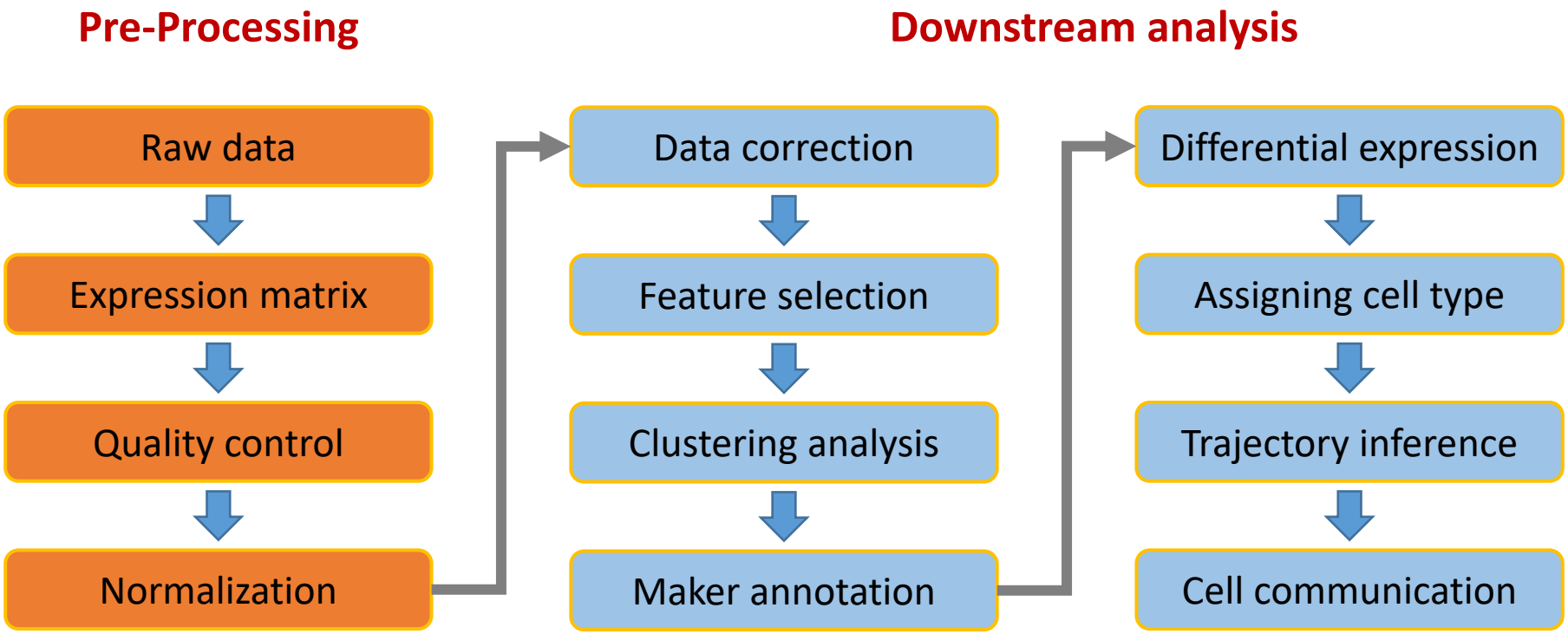
2

scRNA-Seq分析流程

scRNA-Seq pipeline

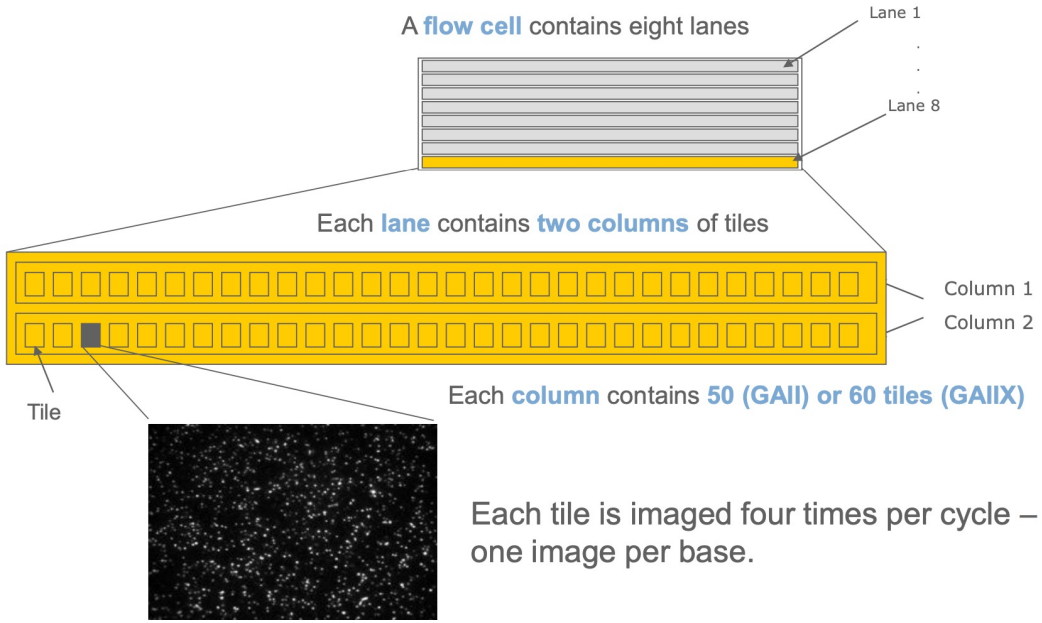
Current best practices in single-cell RNA-seq analysis: a tutorial

Malte D Luecken[1] & Fabian J Theis[1,2,*]

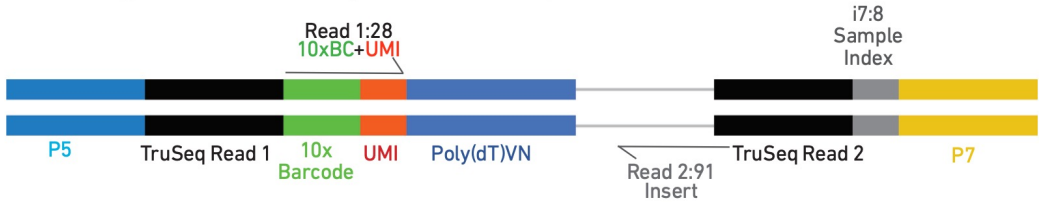


Rawdata (bcl to fastq)

Data\Intensities\BaseCalls\L001\C1.1\s_1_3.bcl , **Base calls for lane 1, cycle 1, tile 3**

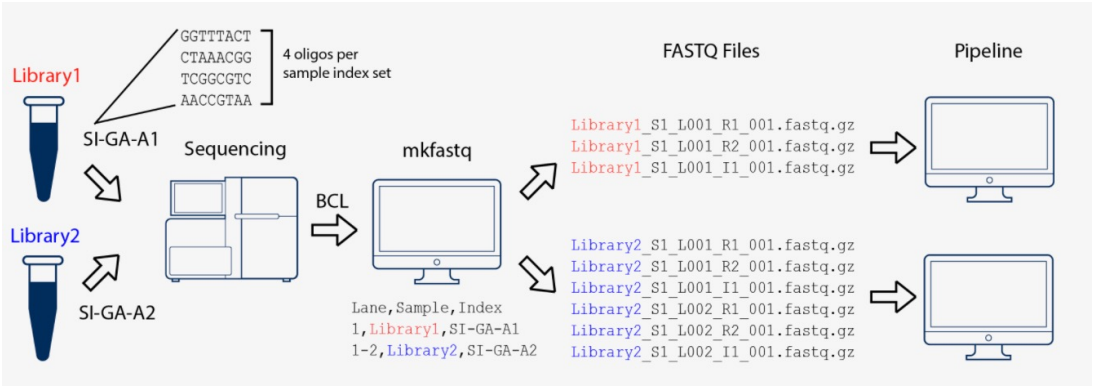


https://www.broadinstitute.org/files/shared/illuminavids/sequencingSlides.pdf



Each sample index set is base-balanced to avoid monochromatic signal issues when it is the sole sample loaded on an Illumina sequencer.

`cellranger mkfastq`



https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/mkfastq

bcl (binary base calling) => fastq

I1: sample index (8bp)

R1: barcode+UMI(28bp)

R2:insert (91bp)

https://divingintogeneticsandgenomics.rbind.io/post/understand-10x-scrnaseq-and-scatac-fastqs/

Reads aligning and statistics (1)

- Number of Reads
- Valid Barcodes (737000 barcodes + mismatch)
- Valid UMIs (不含Ns，不是均聚物)
- Sequencing Saturation（如何计算）

Sequencing Saturation = 1 - (n_deduped_reads / n_reads)

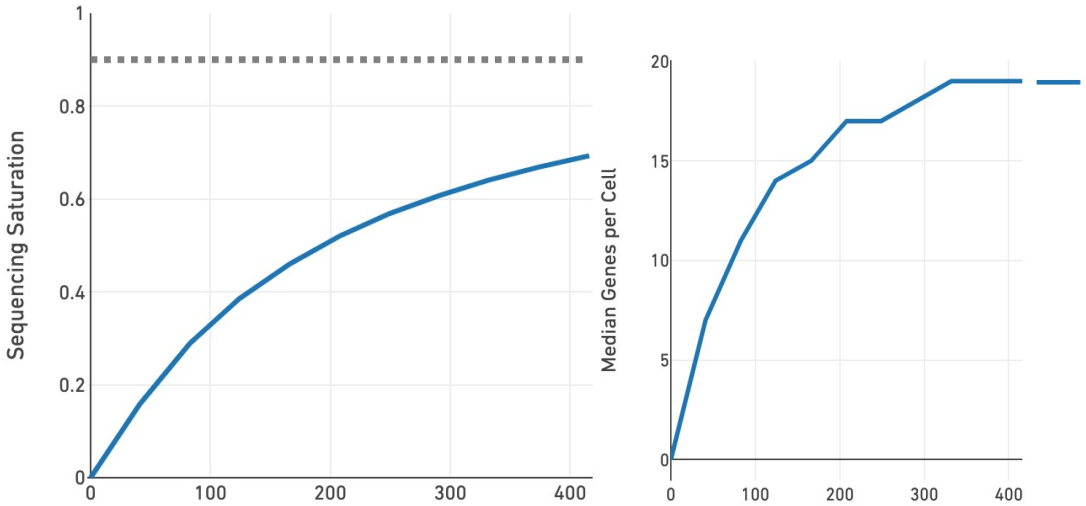
n_deduped_reads = Number of unique (valid cell-barcode, valid UMI, gene) combinations among confidently mapped reads（unique，具有有效barcode和有效UMI且可信比对的唯一reads）

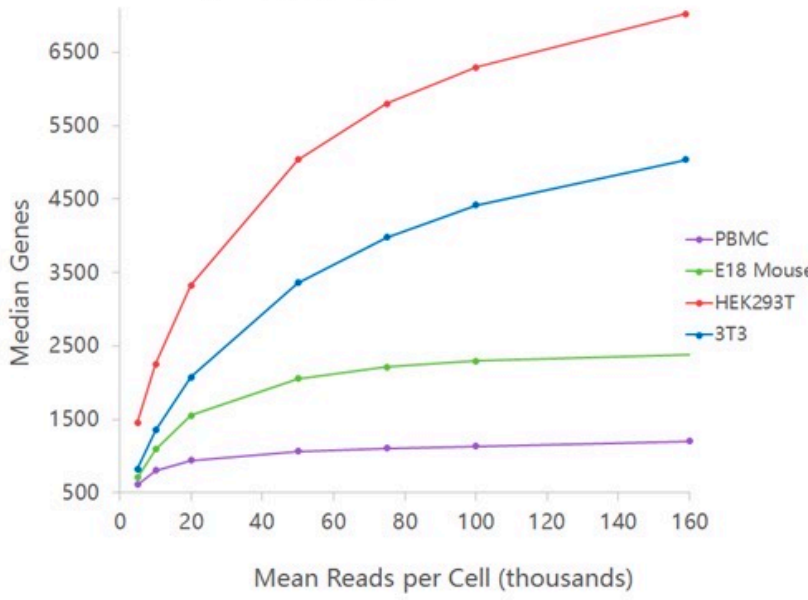
n_reads = Total number of confidently mapped, valid cell-barcode, valid UMI reads（non-unique，具有可信比对、有效barcode和有效UMI的reads）

50%：2条新reads会产生1个新的UMI数（1-1/2）

90%：10条新reads会产生1个新UMI数（1-1/10）

不同的细胞类型所需要的数据饱和度有很大差异





https://kb.10xgenomics.com/hc/en-us/articles/115005062366-What-is-sequencing-saturation-

Reads aligning and statistics (2)

- Fraction of genome/exon mapping reads
- mRNA-mapping reads
- Spike-in detection and spike-in ratio
- Number of detected genes
- Mitochondrial read fraction
- Ribosomal RNA read fraction

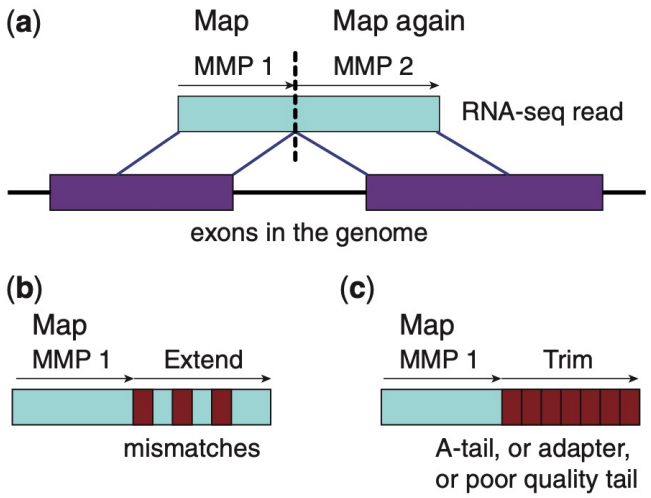
低mapping比例的原因：样本质量、建库质量、静态细胞群（quiescent cell populations）

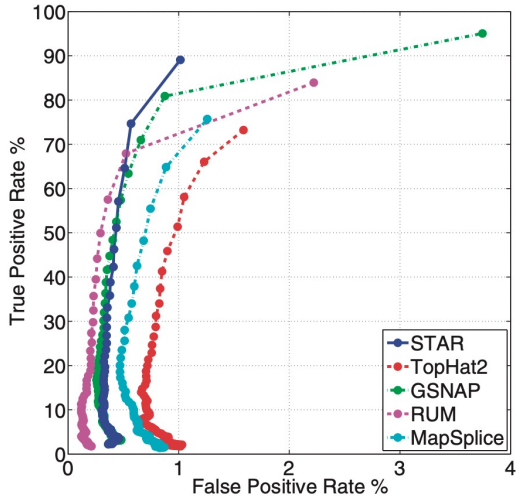
此值异常：建库失败、低mRNA（小细胞/RNA泄露）

基因数异常：高，双/多细胞、较大细胞；低，同上

高线粒体基因：膜破裂，胞质RNA丢失、呼吸活动

核糖体RNA比例高：RNA的降解（RNA聚合酶I介导）





Spliced Transcripts Alignment to a Reference (STAR)：1）能够比对到非连续区域 ；2）允许reads的错配比对；3）过滤掉A-尾/adapter/低质量尾序列

Kallisto

10.15252/msb.20188746
10.1016/j.cell.2009.01.019
10.1093/bioinformatics/bts635

Expression matrix

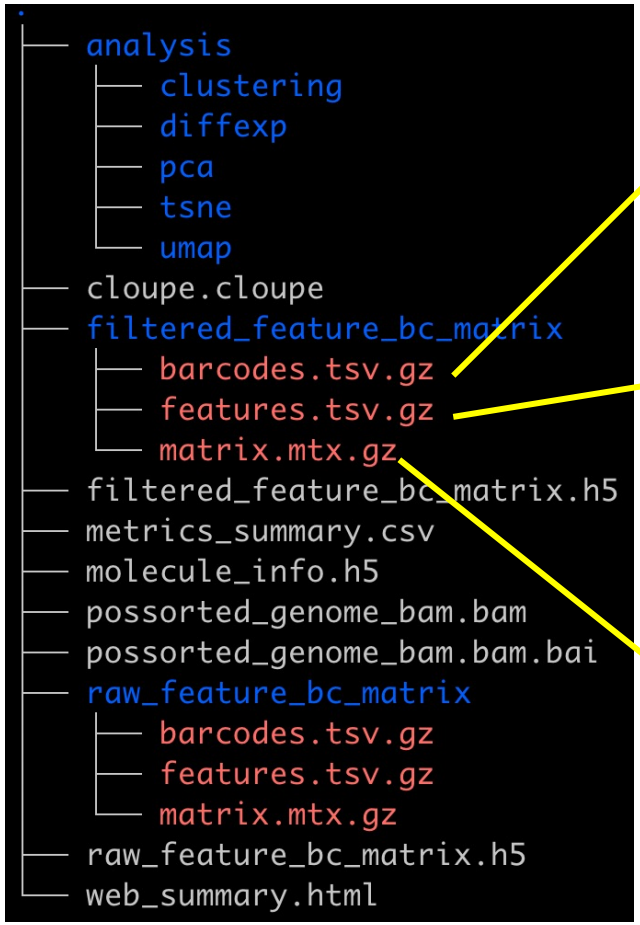
cellranger count

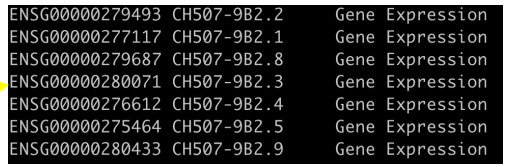
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

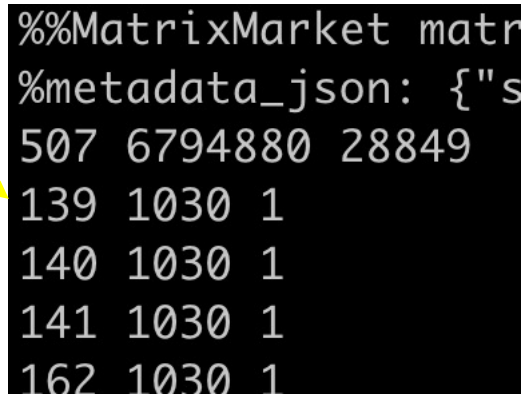
raw_feature_bc_matrix: all feature barcode matrix

filtered_feature_bc_matrix: filtered cell barcode matrix

bcl to fastq => seq of QC => barcode and UMI correction => reads aligning => expression matrix => ……







matrix.mtx说明：第三行开始才是有效信息

第三行：基因数、细胞数、总行数

第四行之后：基因索引、barcode索引、UMI数

Filtering cells: (1) empty dropltes

空液滴：不含细胞但包含ambient RNA，在某一个barcode中检测到非零的UMI数（假细胞/背景噪音）

正常的细胞应当含有较多的mRNA，导致大量的UMI数检出

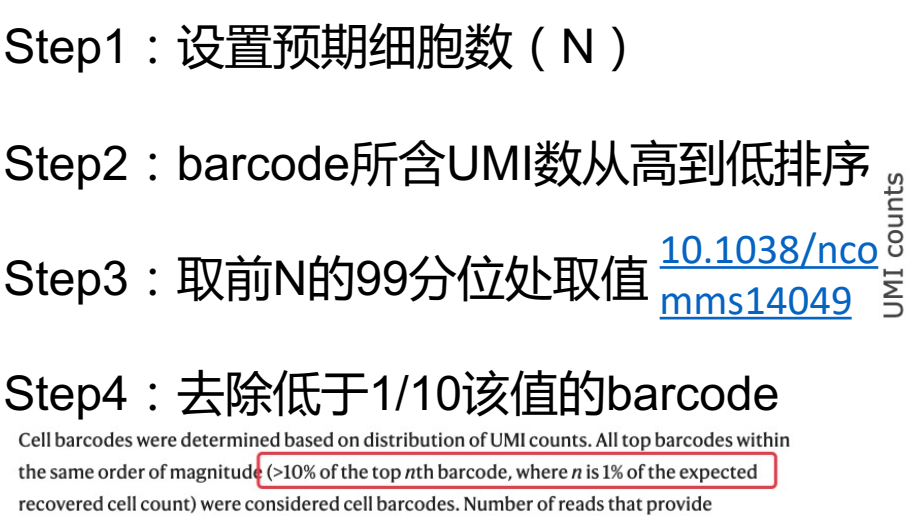
方法一

Step1：设置预期细胞数（N）

Step2：barcode所含UMI数从高到低排序

Step3：取前N的99分位处取值 10.1038/ncomms14049

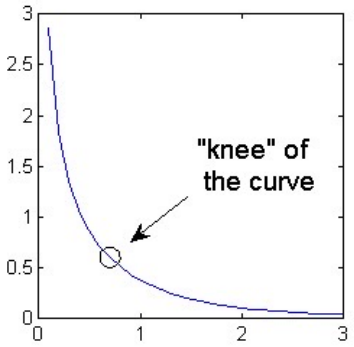
Step4：去除低于1/10该值的barcode



该方法的弊端就是小细胞容易被去掉（低RNA含量的小细胞和空液滴难以区分）

Cell barcodes were determined based on distribution of UMI counts. All top barcodes within the same order of magnitude (>10% of the top *n*th barcode, where *n* is 1% of the expected recovered cell count) were considered cell barcodes. Number of reads that provide

方法二



寻找拐点（knee point），拐点下即为空液滴

10.1038/ncomms14049

方法三

(scRNA-seq) data. We estimate the profile of the ambient RNA pool and test each barcode for deviations from this profile using a Dirichlet-multinomial model of UMI count sampling. Barcodes with significant deviations are

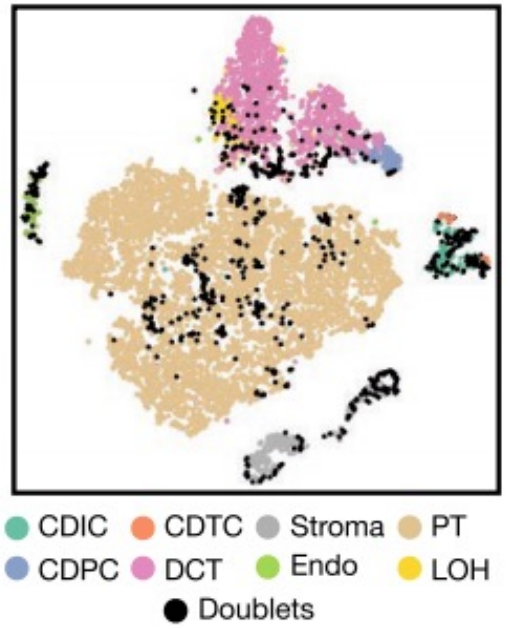
DropletUtils 10.1186/s13059-019-1662-y

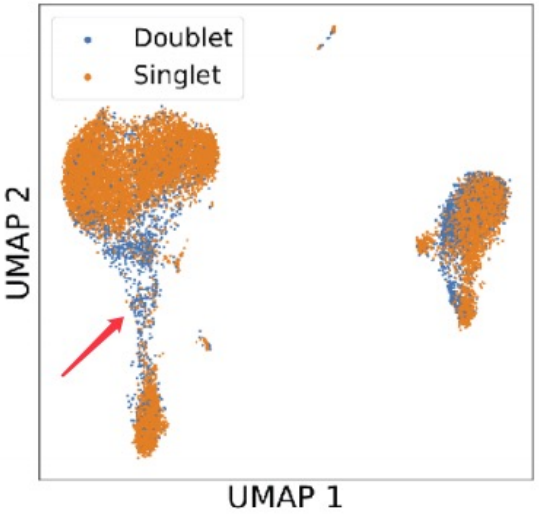
Filtering cells: (2) doublets（双/多细胞）

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

CG000204_ChromiumNextGEMSingleCell3_v3.1_Rev_D.pdf

- 被归为不同的类群（与主群有明显的界限）
- 不同类群之间的峰（类群之间的连接）
- 散落在主群之间，图上无明显特征





10.1016/j.cels.2019.03.003

https://www.biorxiv.org/content/10.1101/841981v1

DoubletFinder: https://github.com/search?q=DoubletFinder

DoubletDetection:

https://github.com/JonathanShor/DoubletDetection

DoubletDecon: https://github.com/EDePasquale/DoubletDecon

Solo: https://github.com/calico/solo

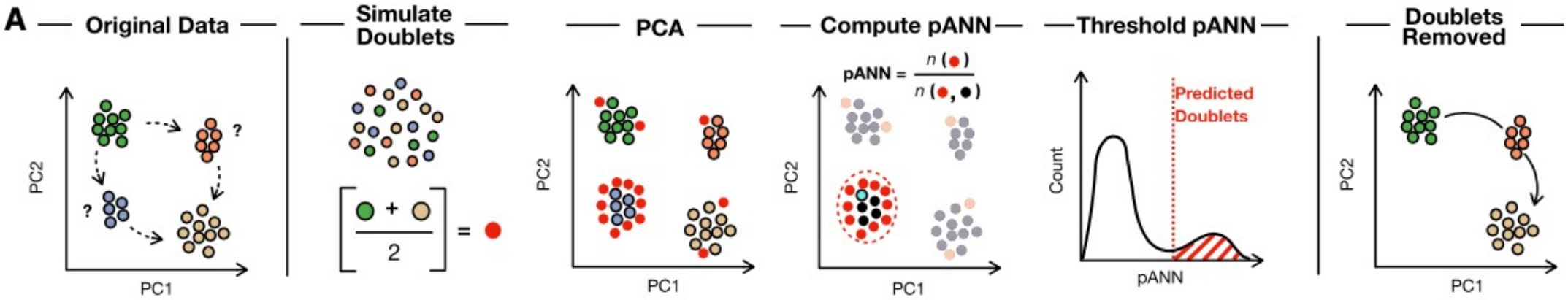
Filtering cells: (2) DoubletFinder软件

特征

- 检测到的基因数（UMIs）较高
- 一个细胞内具有>=2个的基因Marker

注意：较大的细胞也可能具有较多的基因数；

去除双细胞也应当根据平台的经验值确定



Step1：原始数据标准化（μ=0,σ=1）之后PCA聚类；**Step2**：两种类型的细胞模拟doublets（~25%）；

Step3：合并真实数据和模拟数据，并重新进行聚类；**Step4**：计算pANN值（proportion of artificial nearest neighbors，pN/pK）；**Step5**：判定>pANN阈值的cell为doublets

10.1016/j.cels.2019.03.003

由已知确定阈值来推断未知（物以类聚）

缺点：同类型或表达相似的doubltes敏感性低

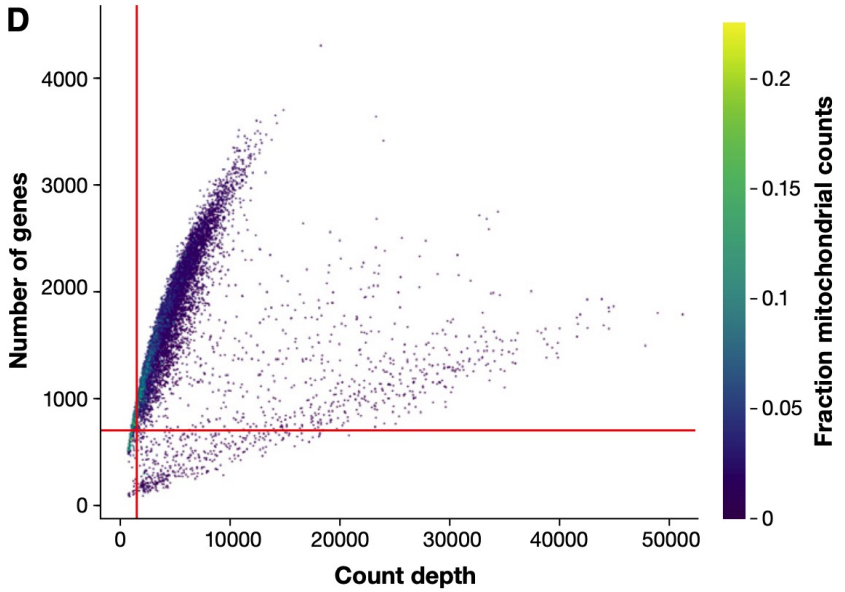
Filtering cells: (3) other standards

1 nUMI + nGene + log10GenesPerUMI + mitoRatio

How to determine MT ration thresholds?

parameter is highly dependent on the tissue type and the questions being investigated. For example, 30% of total mRNA in the heart is mitochondrial due to high energy needs of cardiomyocytes, compared with 5% or less in tissues with low energy demands.42 For instance, 30% mitochondrial mRNA is representative of a healthy heart muscle cell, but would represent a stressed lymphocyte.

https://www.biostars.org/p/366809/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6072887/

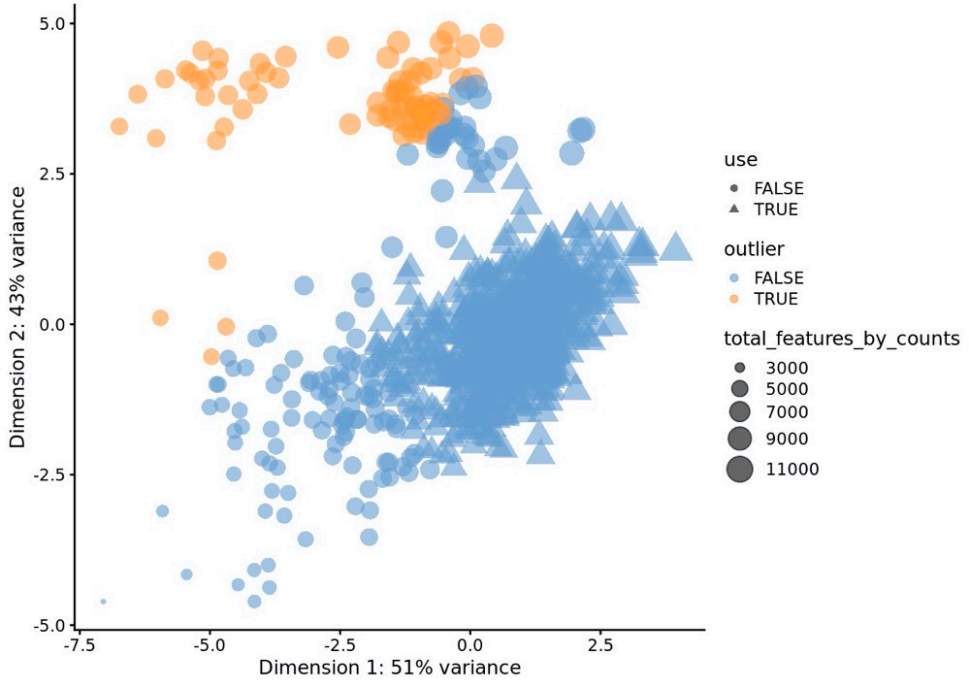


2. Clustering analysis (PCA) to identify outliers

MAD: median absolute deviation

1, 1, 2, 2, 4, 6, 9 => 1, 1, 0, 0, 2, 4, 7; MAD=1

偏离MAD的数据定义为outliers



Filtering genes

1. Mitochondrial/ribosomal genes
2. Low expressed genes
3. Low dispersal genes
4. Technical noise (dropout: correction)

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

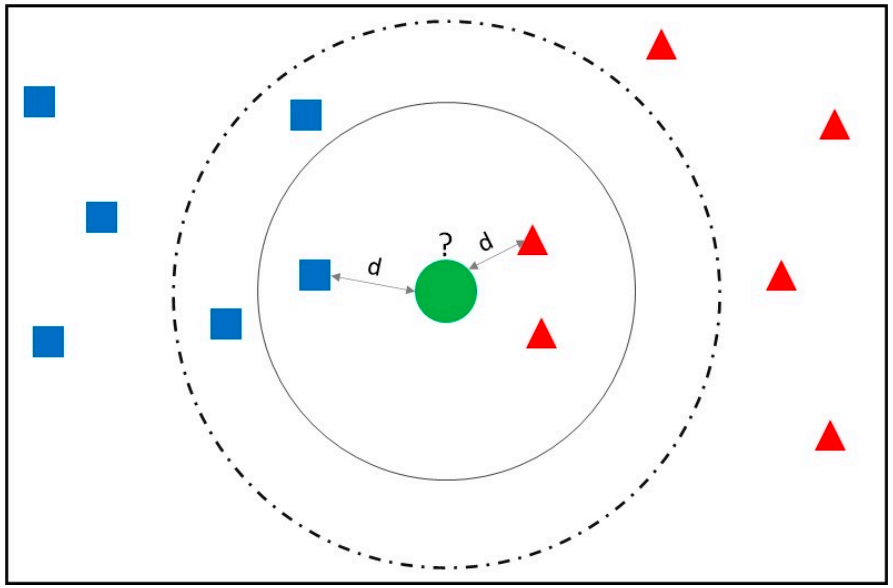
mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

真0（未表达）；假0（表达未捕获）；未知0（随机丢失）

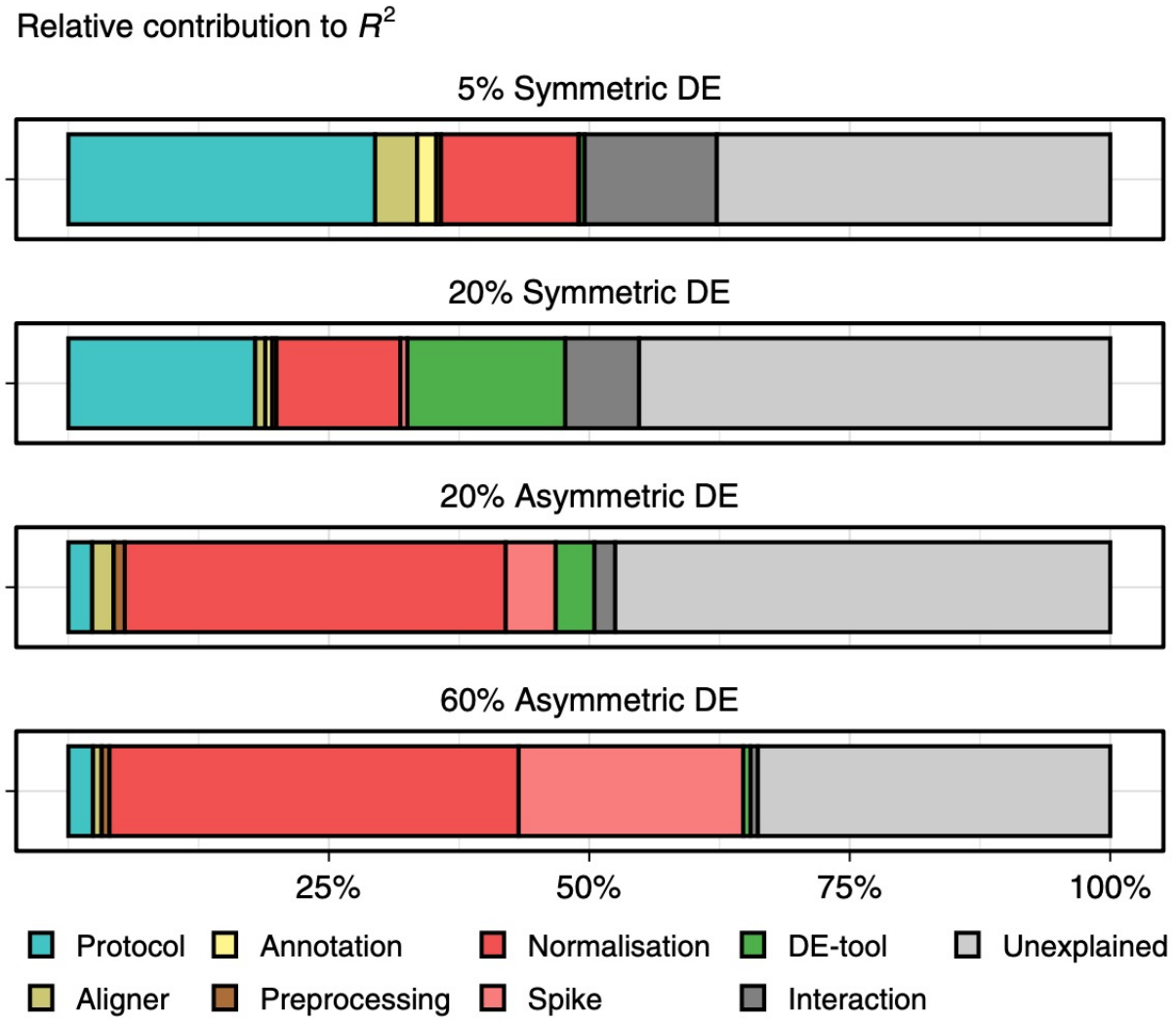
Dropout: 0非真0的现象，属于一种缺失数据。通常由低RNA含量和基因表达的随机性引起，最终会影响细胞亚群的鉴定和谱系发育分析【10.1038/nmeth.2967】

缺失值数据填充方法：1）平均值法；2）KNN（最近邻近法估算，KNNImpute）；3）基于模型（高斯混合聚类方法，GMCimpute）；4）多种方法结合（多种回归方法，seqimpute）【10.1186/s12859-018-2226-y】



https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779

Normalization



Normalization

- 标准化的基因表达不受测序深度的影响
- 标准化的基因差异表达反映生物学差异

	Cell1	Cell2	Cell3	Cell4
Gene1	4	2	7	5
Gene2	10	4	3	9
Gene3	5	10	10	5
Gene4	8	7	4	2
Gene5	5	10	7	7
Gene6	8	2	10	8
Gene7	9	7	5	7
Gene8	6	2	3	3
Gene9	8	9	9	4
Total	63	53	58	50
Median	55.5			

$R=n_i/Total$

	Cell1	Cell2	Cell3	Cell4
Gene1	0.063	0.038	0.121	0.100
Gene2	0.159	0.075	0.052	0.180
Gene3	0.079	0.189	0.172	0.100
Gene4	0.127	0.132	0.069	0.040
Gene5	0.079	0.189	0.121	0.140
Gene6	0.127	0.038	0.172	0.160
Gene7	0.143	0.132	0.086	0.140
Gene8	0.095	0.038	0.052	0.060
Gene9	0.127	0.170	0.155	0.080

$R'=R*median$

	Cell1	Cell2	Cell3	Cell4
Gene1	3.524	2.094	6.698	5.550
Gene2	8.810	4.189	2.871	9.990
Gene3	4.405	10.472	9.569	5.550
Gene4	7.048	7.330	3.828	2.220
Gene5	4.405	10.472	6.698	7.770
Gene6	7.048	2.094	9.569	8.880
Gene7	7.929	7.330	4.784	7.770
Gene8	5.286	2.094	2.871	3.330
Gene9	7.048	9.425	8.612	4.440

$FR=\ln(R')$

	Cell1	Cell2	Cell3	Cell4
Gene1	1.260	0.739	1.902	1.714
Gene2	2.176	1.432	1.055	2.302
Gene3	1.483	2.349	2.259	1.714
Gene4	1.953	1.992	1.342	0.798
Gene5	1.483	2.349	1.902	2.050
Gene6	1.953	0.739	2.259	2.184
Gene7	2.070	1.992	1.565	2.050
Gene8	1.665	0.739	1.055	1.203
Gene9	1.953	2.243	2.153	1.491

1. cellranger 数据标准化 10.1038/ncomms14049

2. counts-per-million (CPM) ; 3. size-factor CPM

$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

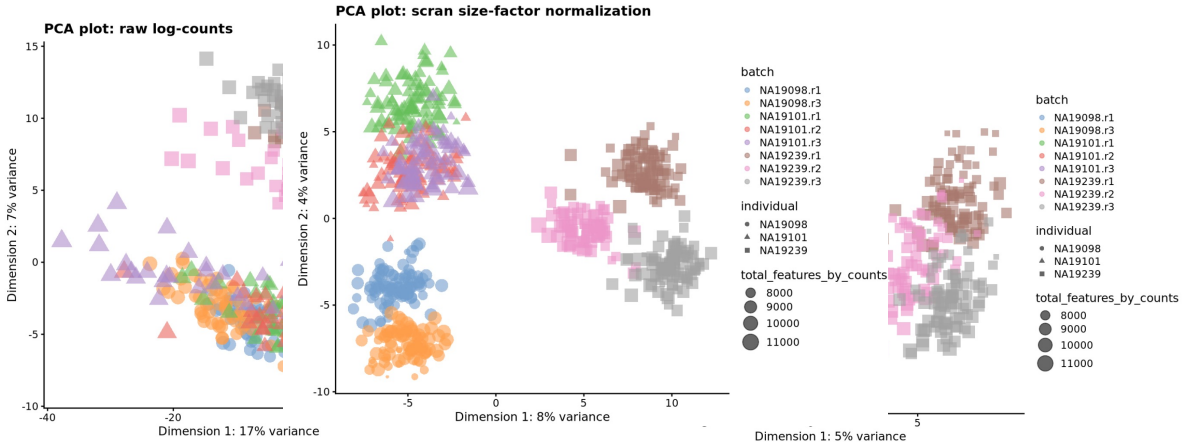
为什么进行log转化？

① log fold change是常用的基因表达变化方式

② 缓和均值-方差的关系，使数据更加集中

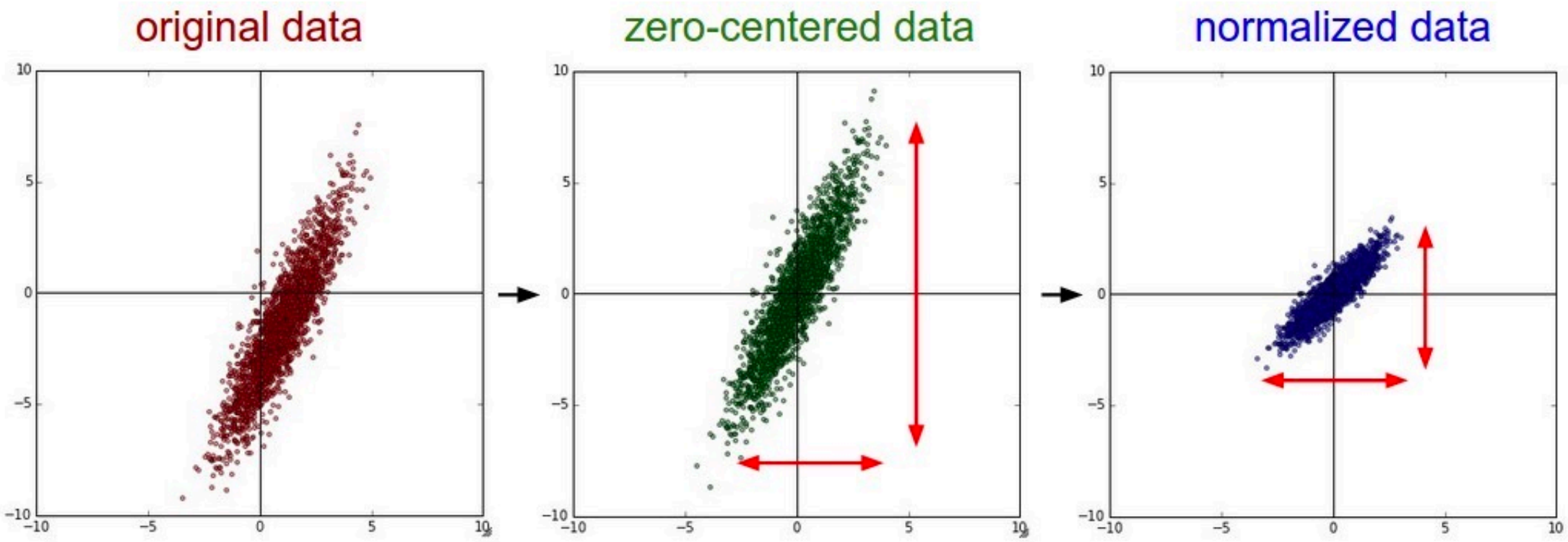
③ 降低数据的偏斜，使之更符合正态分布

缺点：由于数据集中导致差异表达分析假阳性



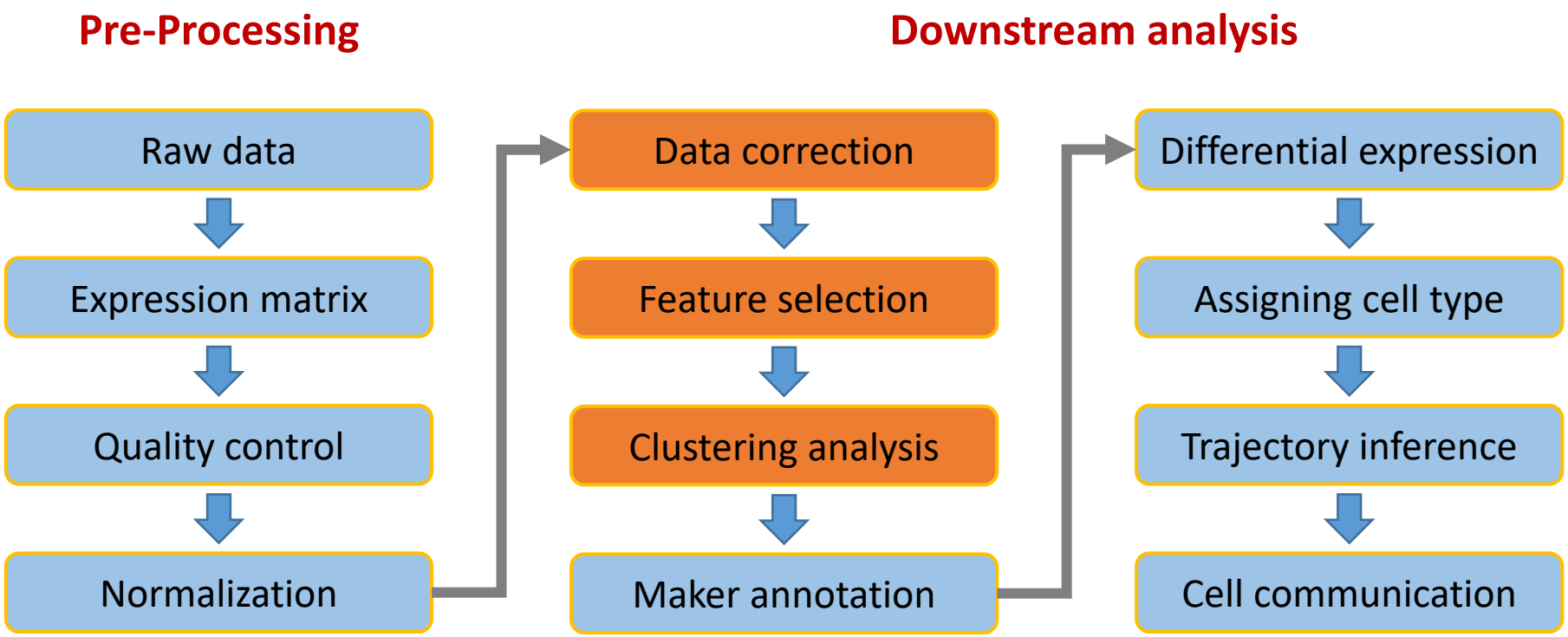
https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/public/normalization-confounders-and-batch-correction.html

Scale : Z-score



$xi' = (xi - \mu) / \sigma$ 使数据平均值=0，标准差=1，近似符合正态分布，方便下游分析

Next plan


Pre-Processing
Downstream analysis
Raw data
Expression matrix
Quality control
Normalization
Data correction
Feature selection
Clustering analysis
Maker annotation
Differential expression
Assigning cell type
Trajectory inference
Cell communication

BerryGenomics
贝瑞基因

Thank You!

官 方 网 站

官 方 微 信

www.berrygenomics.com