scRNA-seq: Quality Control

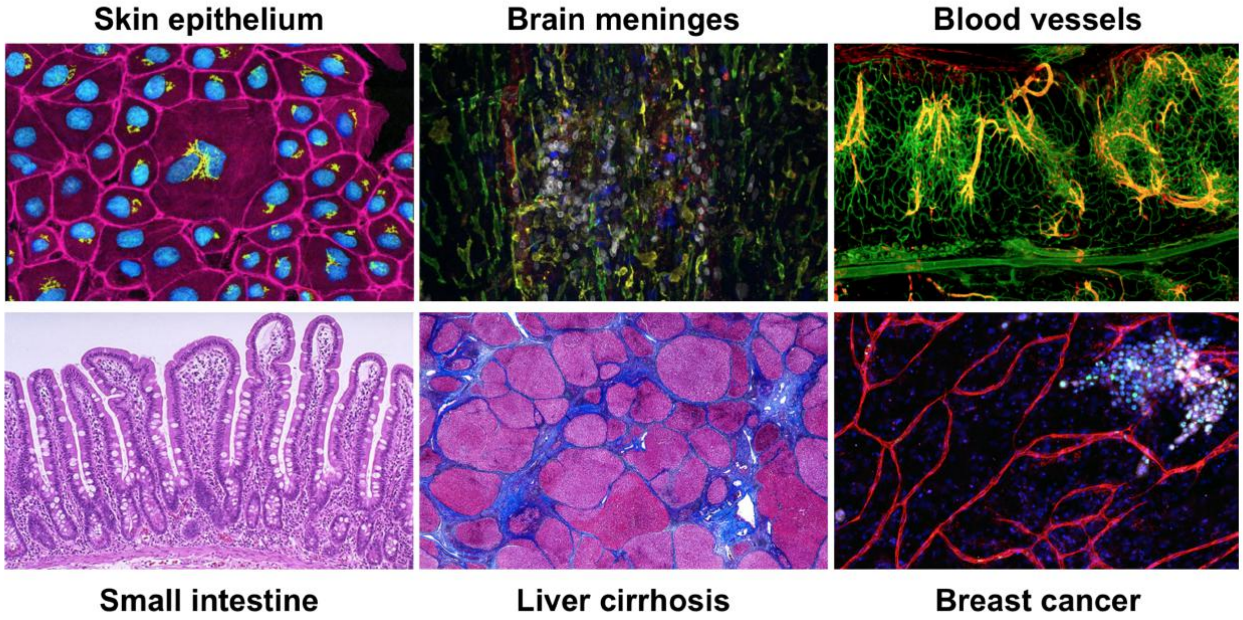
Meng Luo
A bioinformatician

Outline

- Intro Single-cell RNA-seq
- sc/nRNA-seq workflow
- Quality control:
  - Filtering low quality cells
  - Doublet cells
  - Filtering of genes
  - Removal of cell cycle effect
- PCA for quality control

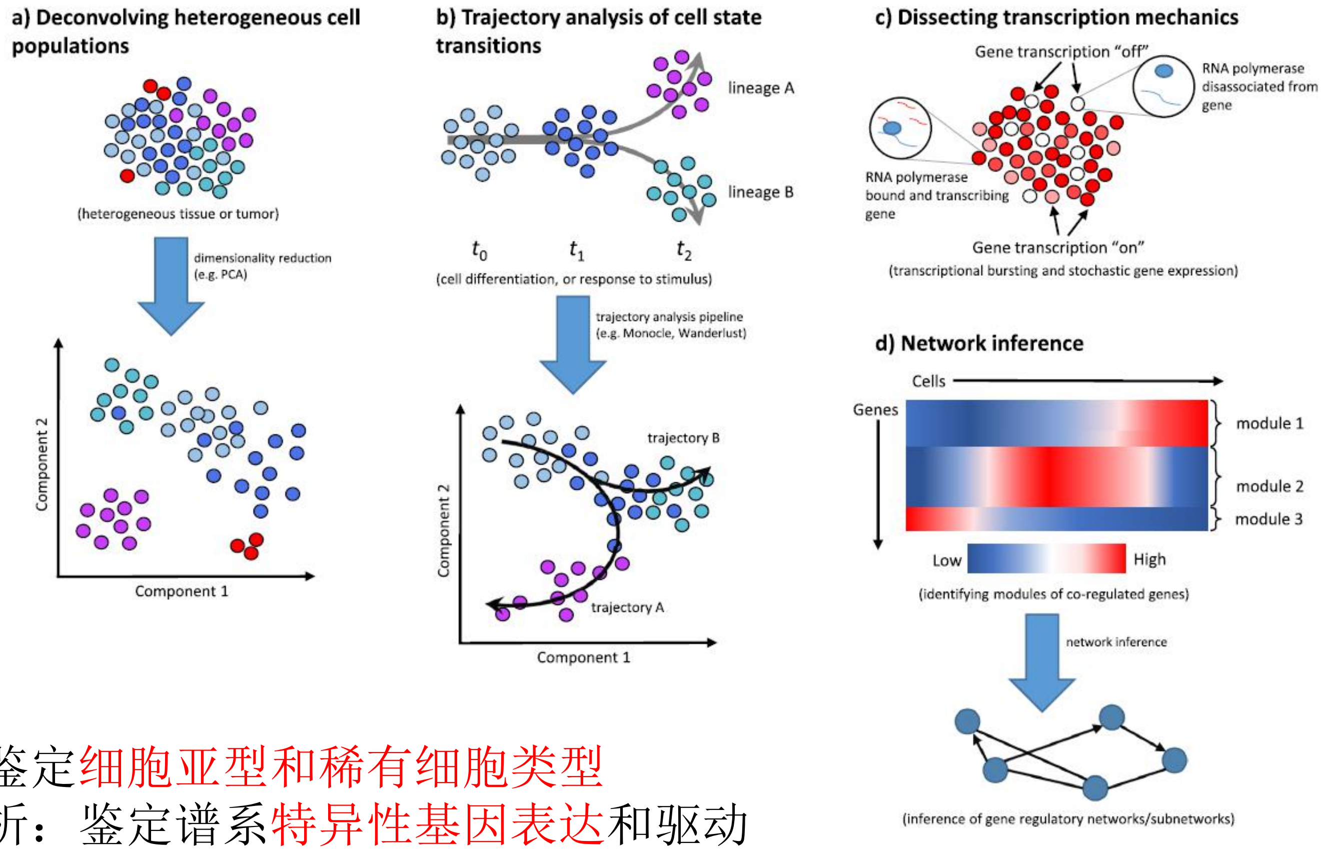
✓ Why Single-cell RNA-seq？

- ✓ 为了更好的了解组织和存在的细胞类型，需要更高分辨率的技术
- ✓ scRNA-seq提供了在单个细胞水平上表达哪些基因的信息
- ✓ 探索组织中存在哪些细胞类型
- ✓ 识别未知/稀有的细胞类型或状态
- ✓ 阐明分化过程中或跨时间或不同状态下的基因表达变化
- ✓ 识别在特定条件下（例如，治疗或疾病）在特定细胞类型中差异表达的基因

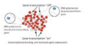


https://www.cell.com/pictureshow

✓ Common applications of single-cell RNA sequencing


a) Deconvolving heterogeneous cell populations
(heterogeneous tissue or tumor)
dimensionality reduction
(e.g. PCA)
Component 2
Component 1
b) Trajectory analysis of cell state transitions
lineage A
lineage B
t_0
t_1
t_2
(cell differentiation, or response to stimulus)
trajectory analysis pipeline
(e.g. Monocle, Wanderlust)
trajectory B
trajectory A
Component 2
Component 1
c) Dissecting transcription mechanics
Gene transcription "off"
RNA polymerase disassociated from gene
RNA polymerase bound and transcribing gene
Gene transcription "on"
(transcriptional bursting and stochastic gene expression)
d) Network inference
Cells
Genes
module 1
module 2
module 3
Low
High
(identifying modules of co-regulated genes)
network inference
(inference of gene regulatory networks/subnetworks)

- 细胞异质性研究：能够鉴定细胞亚型和稀有细胞类型
- 细胞状态转变的轨迹分析：鉴定谱系特异性基因表达和驱动分支的关键基因
- 解剖转录动力学：转录爆发，基因在每个细胞中的打开和关闭
- 网络推断：推断模块，共同调节的基因-推断基因调节网络



Single-cell transcriptome sequencing recent advances and remaining challenges

Serena Liu, Cole Trapnell

F1000 Research (2016)

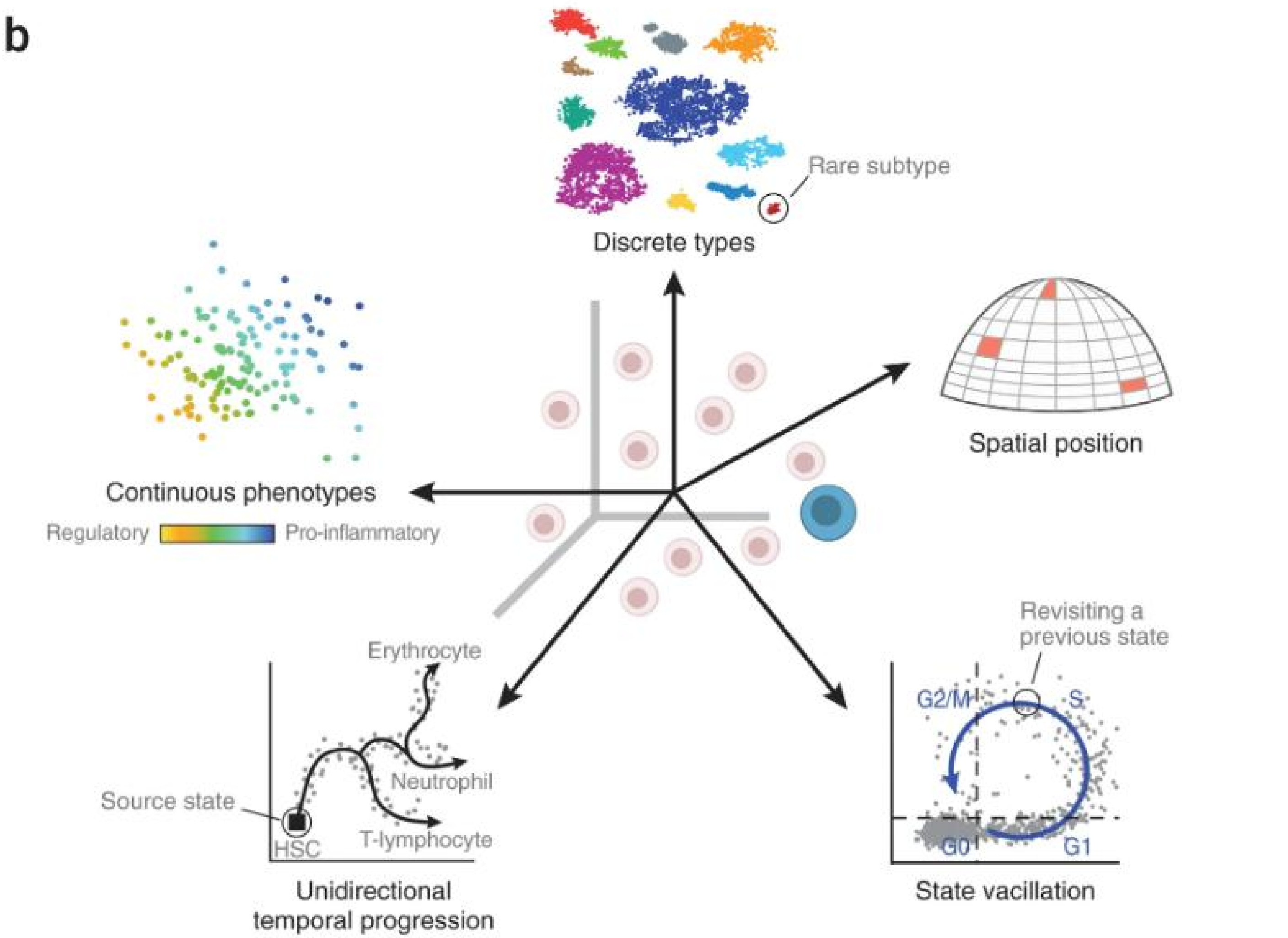
✓ Challenges(complexities) of scRNA-seq analysis

- Large volume of data (high dimension)
  - Expression data from scRNA-seq experiments represent ten or hundreds of thousands of reads for thousandsof cells. The data output is much larger, requiring higher amounts of memory to analyze, larger storage requirements, and more time to run the analyses

- Low depth of sequencing per cell
  - For the droplet-based methods of scRNA-seq, the depth of sequencing is shallow, often detecting only 10- 50% of the transcriptome per cell. This results in cells showing zero counts for many of the genes. However, in a particular cell, a zero count for a gene could either mean that the gene was not being expressed or the transcripts were just not detected. Across cells, genes with higher levels of expression tend to have fewer zeros. Due to this feature, many genes will not be detected in any cell and gene expression will be highly variable between cells.

✓ Challenges(complexities) of scRNA-seq analysis

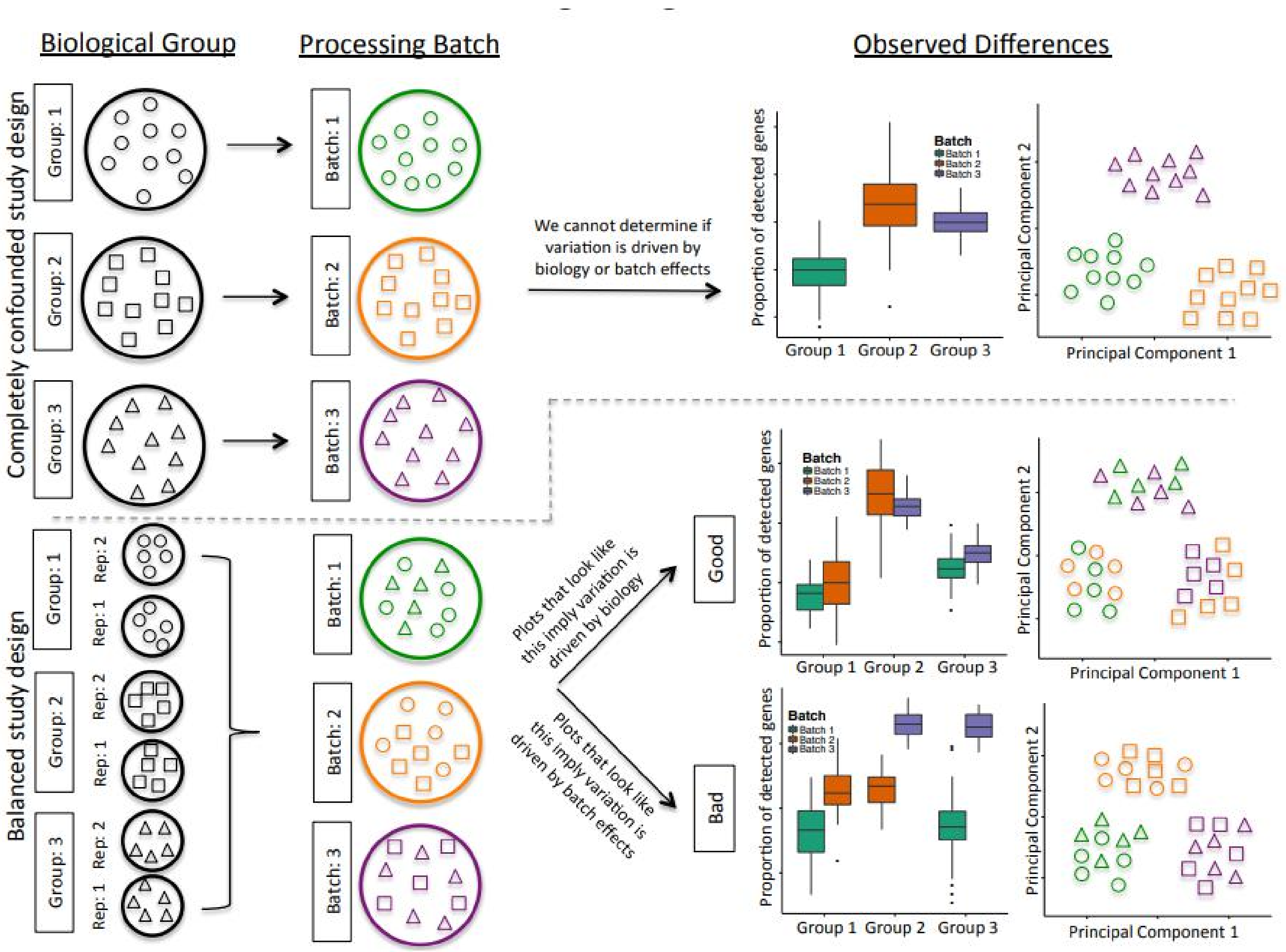
- Bio
- Tra ime for all g ch cell.
- Var ifferent rate
- Cor are by defi ıs from the
- Env nce the gen .
- Ten as cell cycle, can affect the gene expression profiles of individual cells.



✓ Challenges(complexities) of scRNA-seq analysis

- Technical variability across cells/samples
  - Cell-specific capture efficiency: Different cells will have differing numbers of transcripts captured resulting in differences in sequencing depth (e.g. 10-50% of transcriptome).
  - Library quality: Degraded RNA, low viability/dying cells, lots of free floating RNA, poorly dissociated cells, and inaccurate quantitation of cells can result in low quality metrics
  - Amplification bias: During the amplification step of library preparation, not all transcripts are amplified to the same level.
  - Batch effects: Batch effects are a significant issue for scRNA-Seq analyses, since you can see significant differences in expression due solely to the batch effect.

✓ Challenges of scRNA-seq analysis


Biological Group
Processing Batch
Observed Differences
Completely confounded study design
Group: 1
Group: 2
Group: 3
Batch: 1
Batch: 2
Batch: 3
We cannot determine if variation is driven by biology or batch effects
Proportion of detected genes
Batch
Batch 1
Batch 2
Batch 3
Group 1
Group 2
Group 3
Principal Component 2
Principal Component 1
Balanced study design
Rep: 1
Rep: 2
Plots that look like this imply variation is driven by biology
Plots that look like this imply variation is driven by batch effects
Good
Bad

Stephanie C Hicks, F William Townes, Mingxiang Teng, Rafael A Irizarry, Missing data and technical variability in single-cell RNA-sequencing experiments, Biostatistics,October 2018.

✓ Challenges(complexities) of scRNA-seq analysis

- ✓ While scRNA-seq is a powerful and insightful method for the analysis of gene expression with single-cell resolution, there are many challenges and sources of variation that can make the analysis of the data complex or limited.

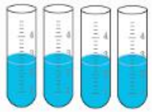
✓ Workflows

Samples

Human and mouse mixture

Rep1 Rep2 Rep3 Rep4

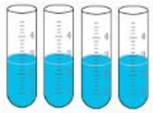
Human



Rep1 Rep2 Rep3 Rep4

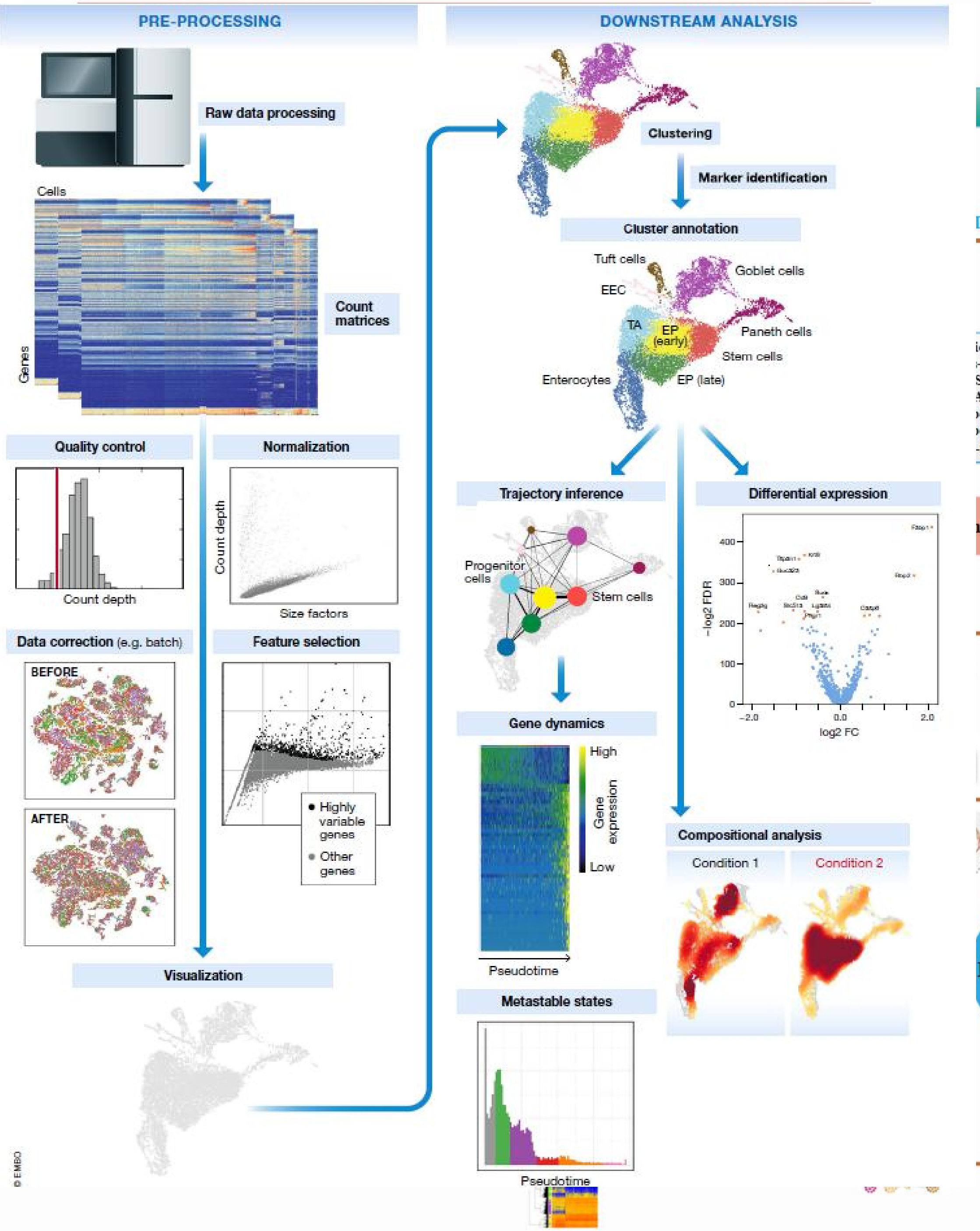
Resection, biopsy, ascites
Fresh, Frozen

Mouse



Rep1 Rep2 Rep3 Rep4

Copyrigth
Meng Luo



T2

SCUMI
Cumulus

Data analysis tools

Cellranger count

ion of exon mapping reads
–for full length methods
S2
A-mapping reads
er of UMIs/reads
er of detected genes
-in detection

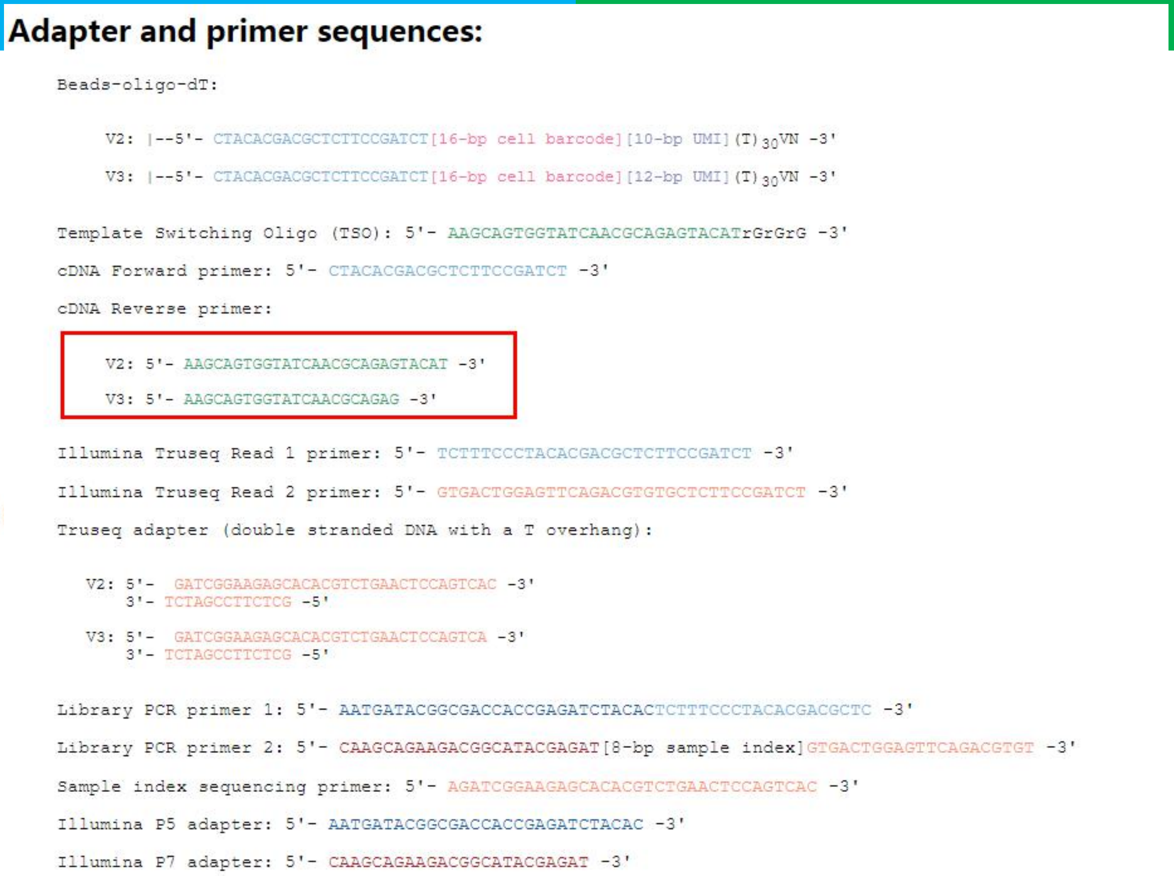
es

Hierarchical
k-Means
Graph-based
(scanpy, seurat...)

- Summary of 55 TI methods
- Monocle, PAGA, Slingshot

Infer CNV

GO, KEGG
Enrichment
TopGO



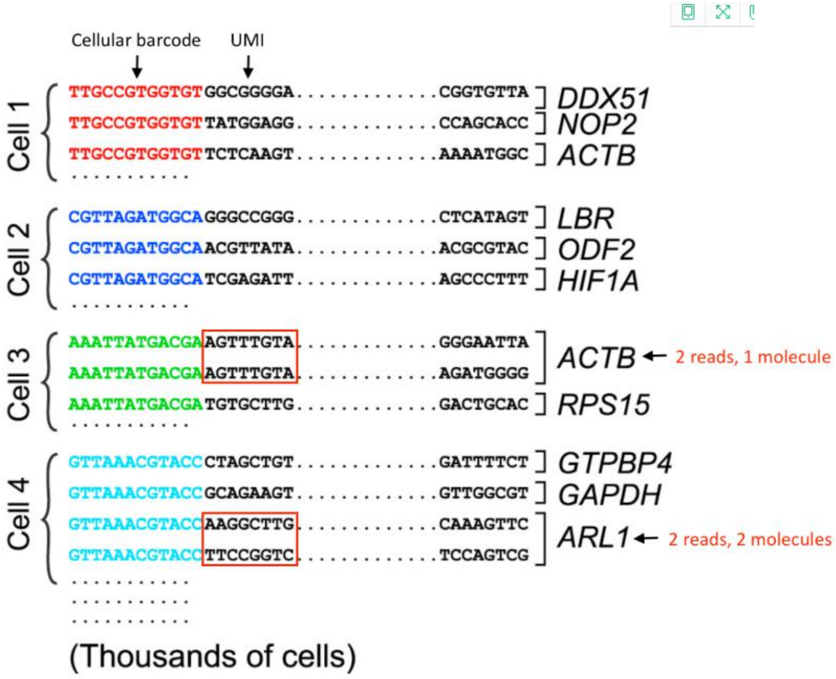
Adapter and primer sequences:

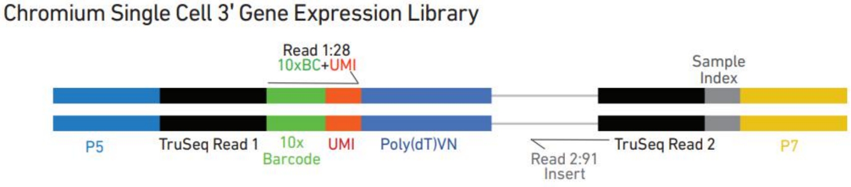
```
Beads-oligo-dT:

    V2: |--5'- CTACACGACGCTCTTCCGATCT[16-bp cell barcode][10-bp UMI](T)30VN -3'

    V3: |--5'- CTACACGACGCTCTTCCGATCT[16-bp cell barcode][12-bp UMI](T)30VN -3'

Template Switching Oligo (TSO): 5'- AAGCAGTGGTATCAACGCAGAGTACATrGrGrG -3'

cDNA Forward primer: 5'- CTACACGACGCTCTTCCGATCT -3'

cDNA Reverse primer:

    V2: 5'- AAGCAGTGGTATCAACGCAGAGTACAT -3'

    V3: 5'- AAGCAGTGGTATCAACGCAGAG -3'

Illumina Truseq Read 1 primer: 5'- TCTTTCCCTACACGACGCTCTTCCGATCT -3'

Illumina Truseq Read 2 primer: 5'- GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT -3'

Truseq adapter (double stranded DNA with a T overhang):

   V2: 5'-  GATCGGAAGAGCACACGTCTGAACTCCAGTCAC -3'
       3'- TCTAGCCTTCTCG -5'

   V3: 5'-  GATCGGAAGAGCACACGTCTGAACTCCAGTCA -3'
       3'- TCTAGCCTTCTCG -5'

Library PCR primer 1: 5'- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC -3'

Library PCR primer 2: 5'- CAAGCAGAAGACGGCATACGAGAT[8-bp sample index]GTGACTGGAGTTCAGACGTGT -3'

Sample index sequencing primer: 5'- AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -3'

Illumina P5 adapter: 5'- AATGATACGGCGACCACCGAGATCTACAC -3'

Illumina P7 adapter: 5'- CAAGCAGAAGACGGCATACGAGAT -3'
```

10X Chromium 单细胞转录组测序可分为3'端polyA 附近区域捕获和5'端转录起始位置附近捕获建库测序。3'端转录本测序适用于各种类型的细胞，对于10X 单细胞3'的V2 试剂盒处理的样本，Read 1 由16 bp 的10X 细胞Barcode 和10 bp 的UMI 序列组成；而V3 试剂盒处理的样本，Read 1 由16 bp 的10X 细胞Barcode 和12 bp 的UMI 序列组成。其中，10X chromium 的Barcode 用于标记单个细胞，存在于逆转录引物上的随机核苷酸序列上。Read 2 是151 bp 的cDNA 序列, 一般只将前98 bp 用于下游分析。

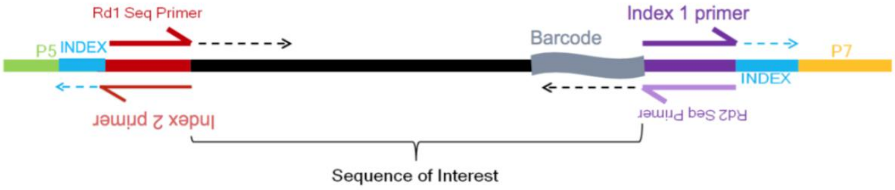
https://teichlab.github.io/scg_lib_structs/methods_html/10xChromium3.html
https://teichlab.github.io/scg_lib_structs/methods_html/10xChromium3fb.html

10x Genomics


Cellular barcode
UMI
Cell 1
TTGCCGTGGTGTGGCGGGGA............CGGTGTTA DDX51
TTGCCGTGGTGTTATGGAGG............CCAGCACC NOP2
TTGCCGTGGTGTTCTCAAGT............AAAATGGC ACTB
Cell 2
CGTTAGATGGCAGGGCCGGG............CTCATAGT LBR
CGTTAGATGGCAACGTTATA............ACGCGTAC ODF2
CGTTAGATGGCATCGAGATT............AGCCCTTT HIF1A
Cell 3
AAATTATGACGAAGTTTGTA............GGGAATTA
AAATTATGACGAAGTTTGTA............AGATGGGG ACTB ← 2 reads, 1 molecule
AAATTATGACGATGTGCTTG............GACTGCAC RPS15
Cell 4
GTTAAACGTACCCTAGCTGT............GATTTTCT GTPBP4
GTTAAACGTACCGCAGAAGT............GTTGGCGT GAPDH
GTTAAACGTACCAAGGCTTG............CAAAGTTC
GTTAAACGTACCTTCCGGTC............TCCAGTCG ARL1 ← 2 reads, 2 molecules
(Thousands of cells)



10X v3



inDrops v3

液滴方法:

- Sample index(样本索引)：确定read来自哪个样本(在库准备过程中添加—需要记录)
- Cellular barcode：确定read来自哪个细胞(每种库制备方法都有在库制备过程中使用的细胞条形码的库)
- UMI(唯一分子标识符)：确定read来自哪个转录分子
- Sequencing read1：Read1序列
- Sequencing read2：Read2序列

✓ Take home massage

nCount_RNA： 每个细胞的UMI数量

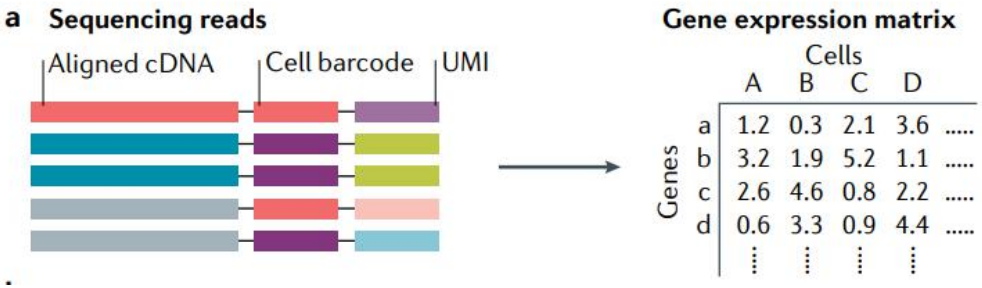
nFeature_RNA： 每个细胞检测到的基因数量

number of genes detected per UMI: 每个UMI检测到的基因(越多，我们的数据就越复杂)

mitochondrial ratio: 自线粒体基因的细胞读数的百分比

Ribosomal ratio:编码核糖体蛋白基因的UMI 序列比例

✓ Assessing the quality metrics

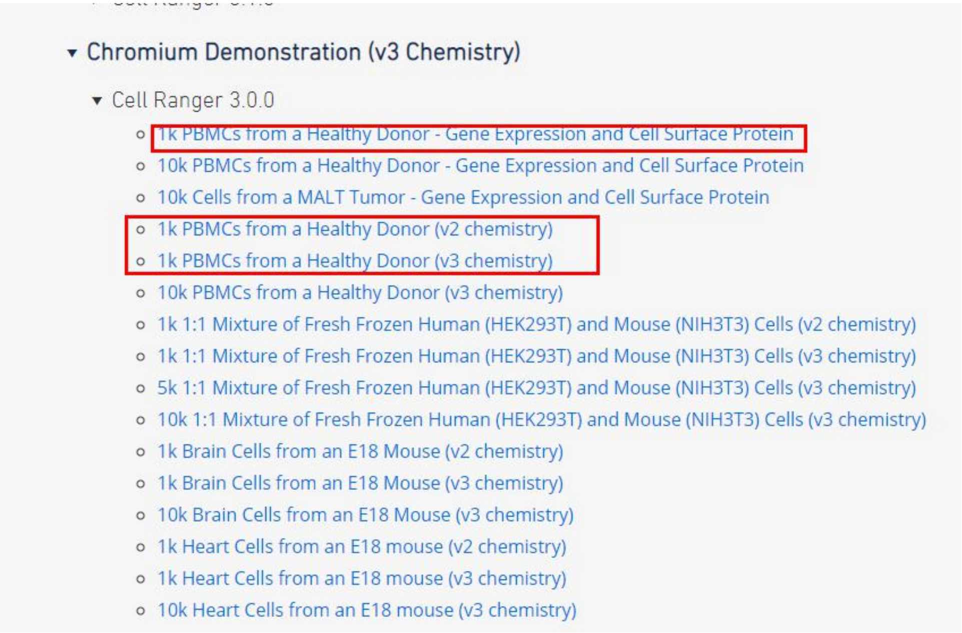


	A	B	C	D	
a	1.2	0.3	2.1	3.6	.....
b	3.2	1.9	5.2	1.1	.....
c	2.6	4.6	0.8	2.2	.....
d	0.6	3.3	0.9	4.4	.....
	⋮	⋮	⋮	⋮	

- Cell counts
- UMI counts per cell
- Genes detected per cell
- UMIs vs. genes detected
- Mitochondrial counts ratio
- doublets: doublets are generated from two cells. They typically arise due to errors in cell sorting or capture, especially in droplet-based protocols involving thousands of cells. Doublets are obviously undesirable when the aim is to characterize populations at the single-cell level.

✓ filtering in different ways and exploring variablility data

- 3 different PBMC datasets from the 10x Genomics
  - 1k PBMCs using 10x v2 chemistry
  - 1k PBMCs using 10x v3 chemistry
  - 1k PBMCs using 10x v3 chemistry in combination with cell surface proteins, but disregarding the protein data and only looking at gene expression.

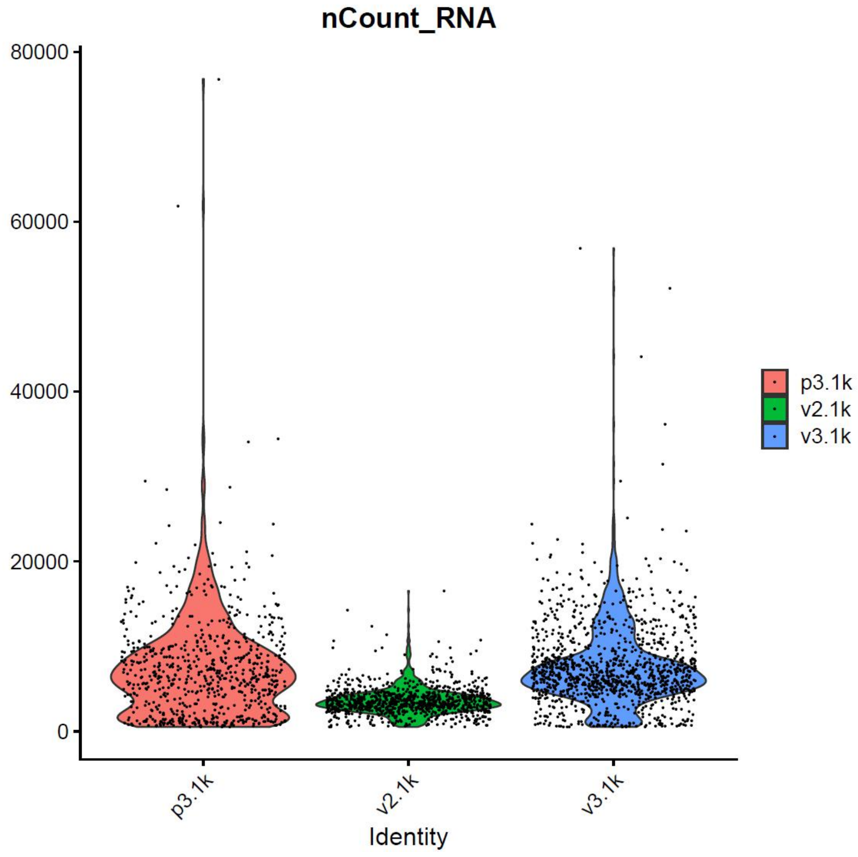


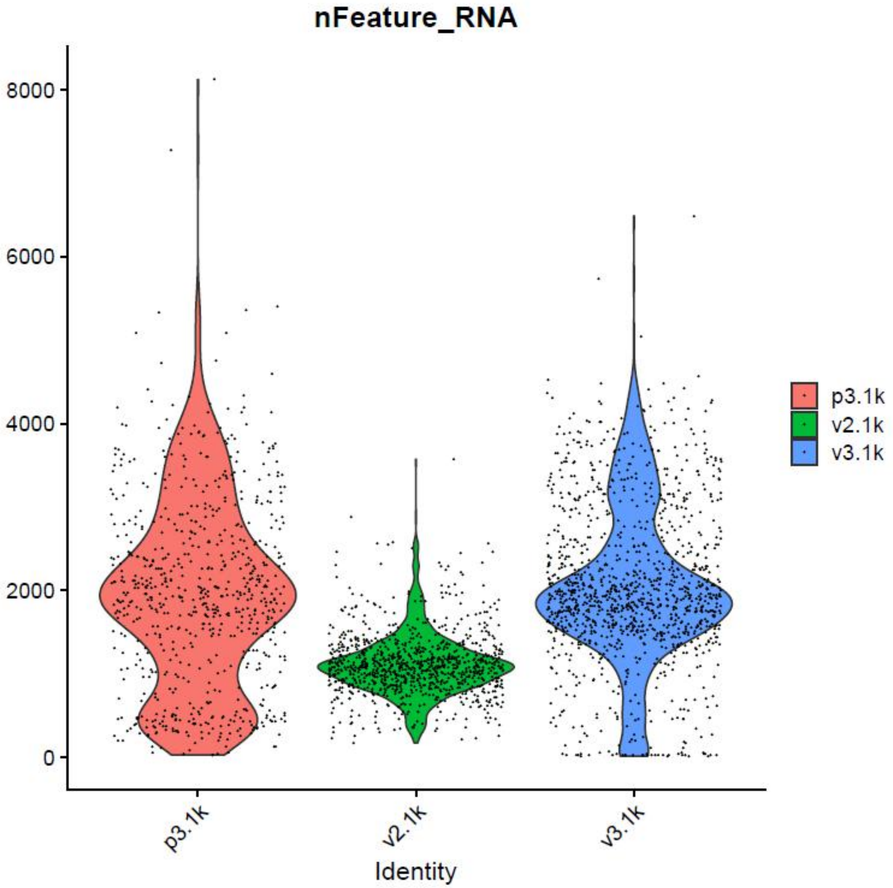
- Chromium Demonstration (v3 Chemistry)
  - Cell Ranger 3.0.0
    - 1k PBMCs from a Healthy Donor - Gene Expression and Cell Surface Protein
    - 10k PBMCs from a Healthy Donor - Gene Expression and Cell Surface Protein
    - 10k Cells from a MALT Tumor - Gene Expression and Cell Surface Protein
    - 1k PBMCs from a Healthy Donor (v2 chemistry)
    - 1k PBMCs from a Healthy Donor (v3 chemistry)
    - 10k PBMCs from a Healthy Donor (v3 chemistry)
    - 1k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells (v2 chemistry)
    - 1k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells (v3 chemistry)
    - 5k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells (v3 chemistry)
    - 10k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells (v3 chemistry)
    - 1k Brain Cells from an E18 Mouse (v2 chemistry)
    - 1k Brain Cells from an E18 Mouse (v3 chemistry)
    - 10k Brain Cells from an E18 Mouse (v3 chemistry)
    - 1k Heart Cells from an E18 mouse (v2 chemistry)
    - 1k Heart Cells from an E18 mouse (v3 chemistry)
    - 10k Heart Cells from an E18 mouse (v3 chemistry)

https://support.10xgenomics.com/single-cell-gene-expression/datasets

✓ filtering in different ways and exploring variablility data

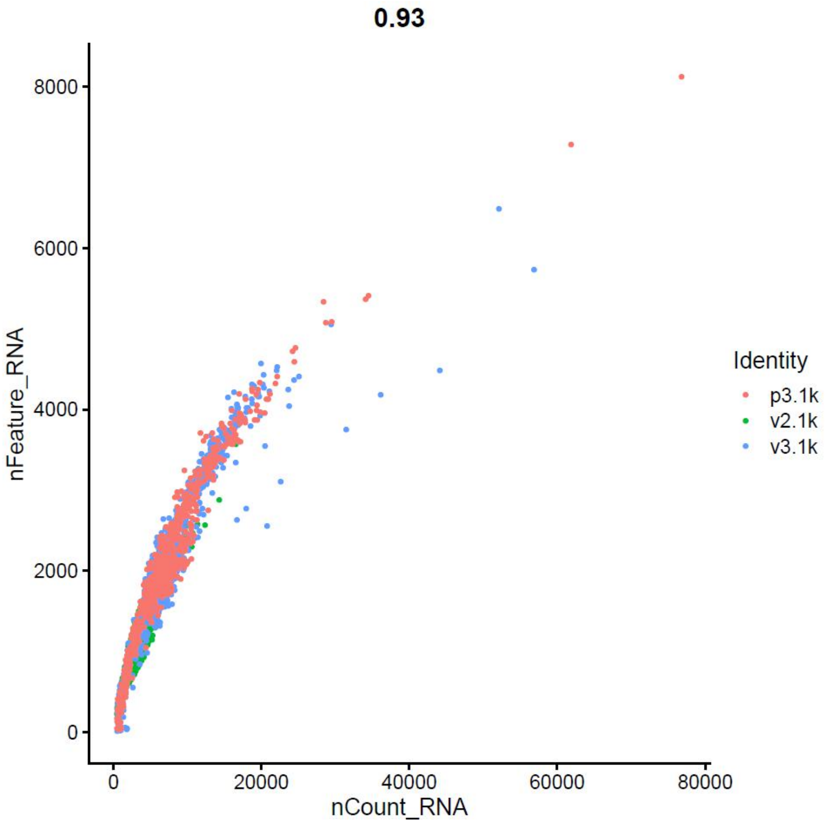
➢ QC-features :nFeature_RNA, nCount_RNA

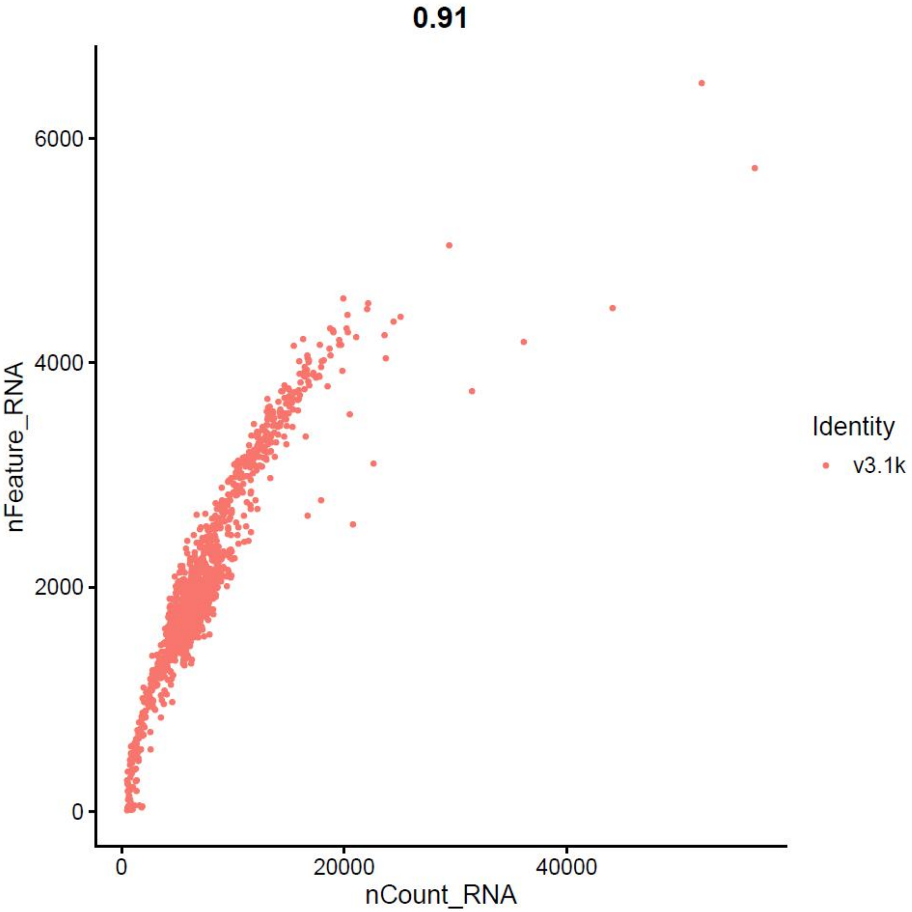




✓ filtering in different ways and exploring variablility data

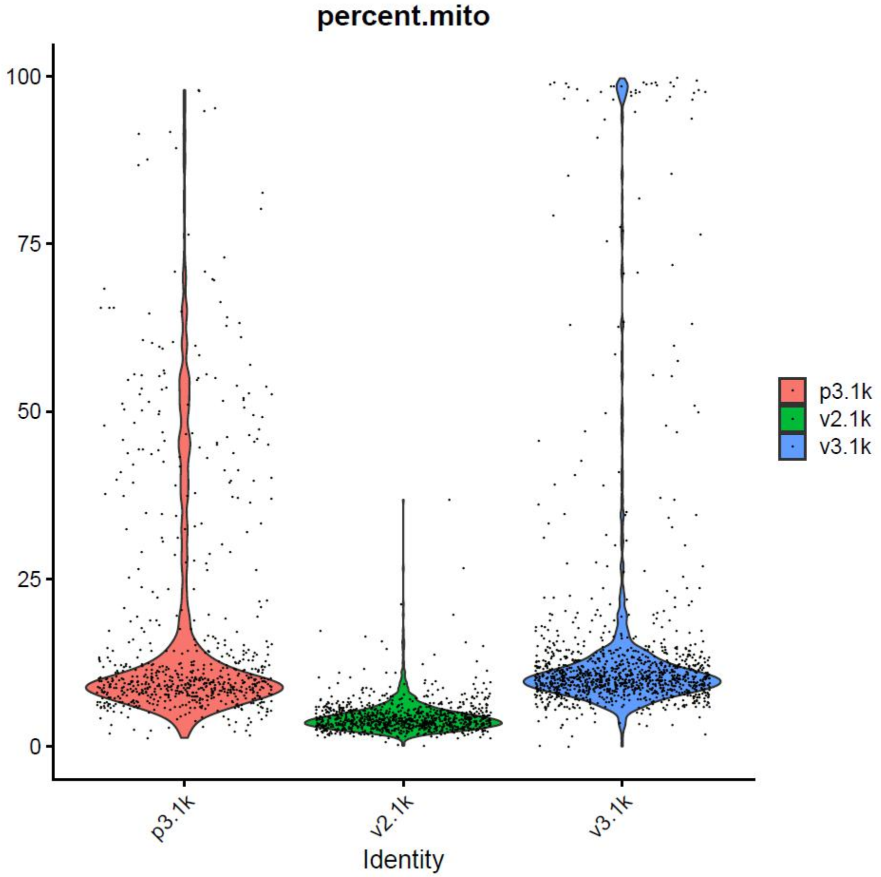
➢ correlation

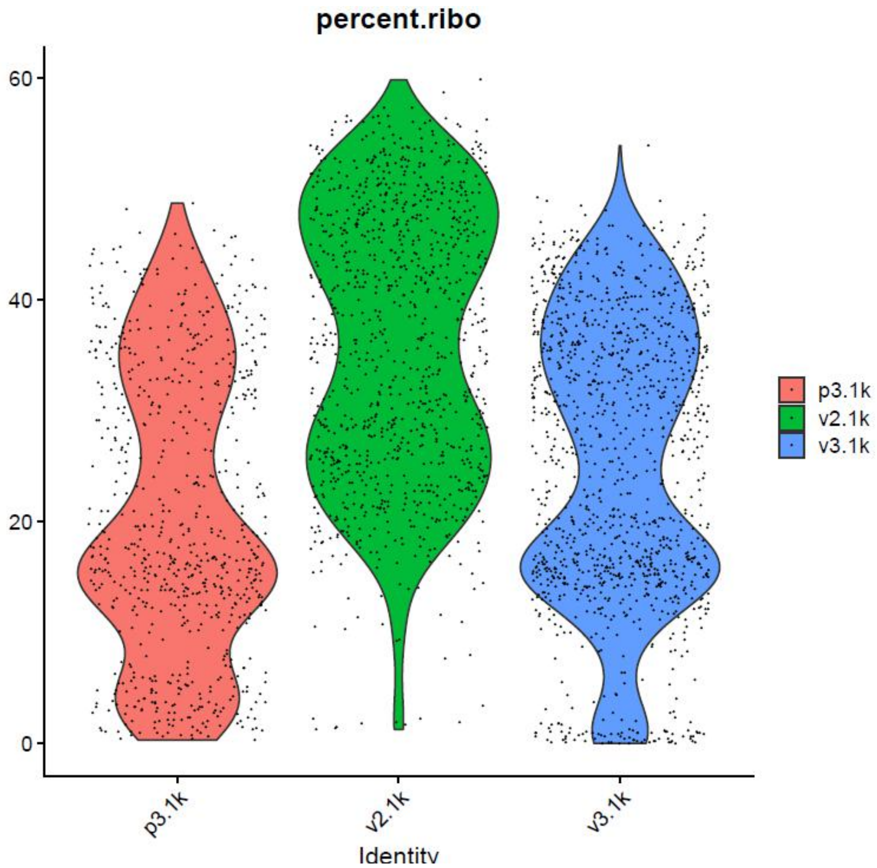




✓ filtering in different ways and exploring variablility data

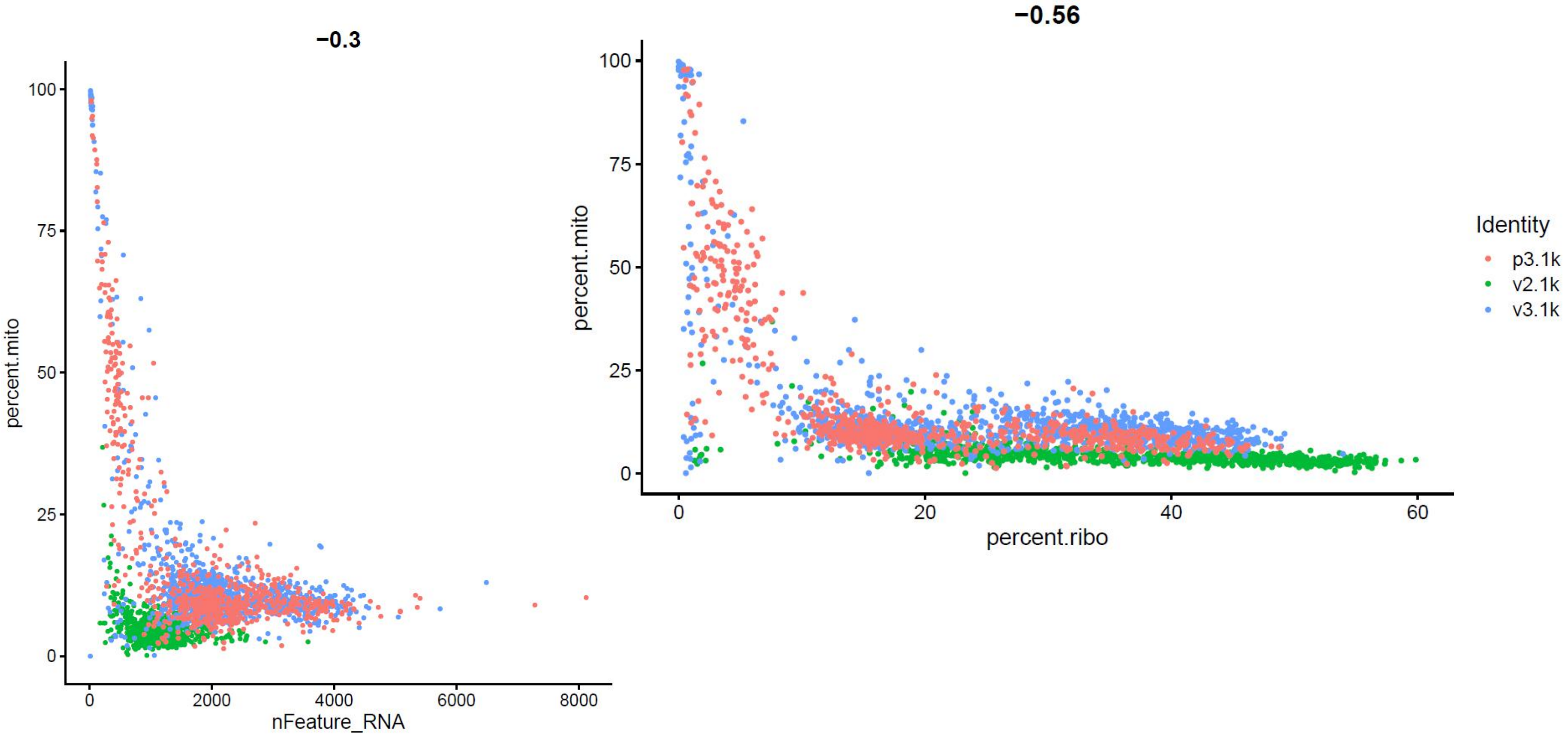
➢ Calculate mitochondrial, ribosomal proportion





✓ Doing filtering in different ways and exploring variablility data

➢ Calculate mitochondrial, ribosomal proportion



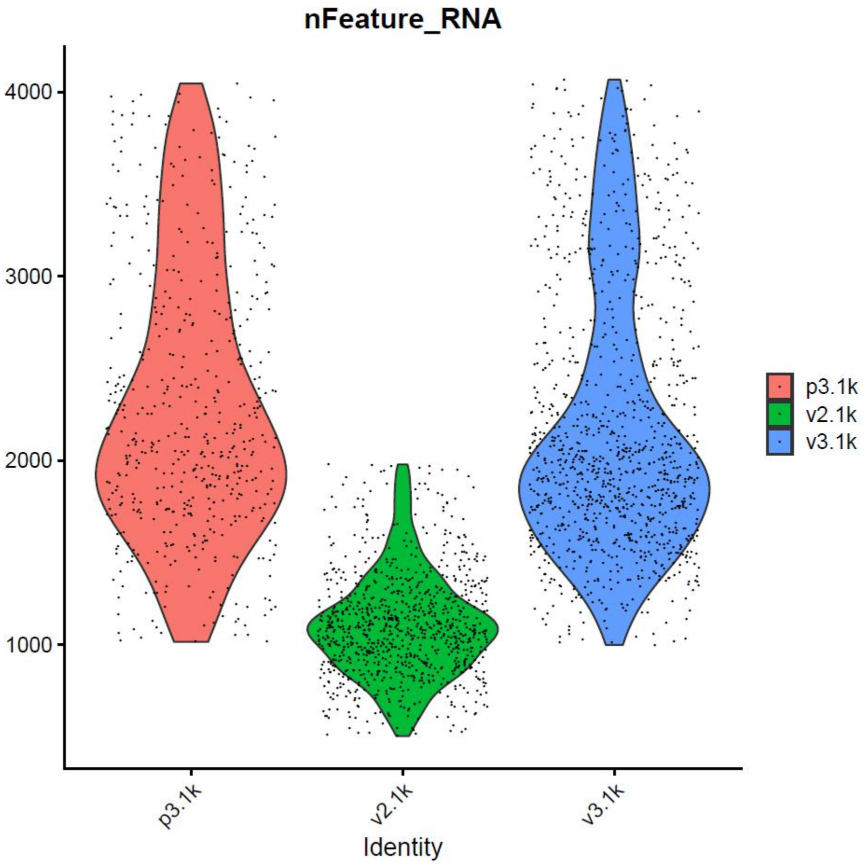
➢ Gene detection filtering

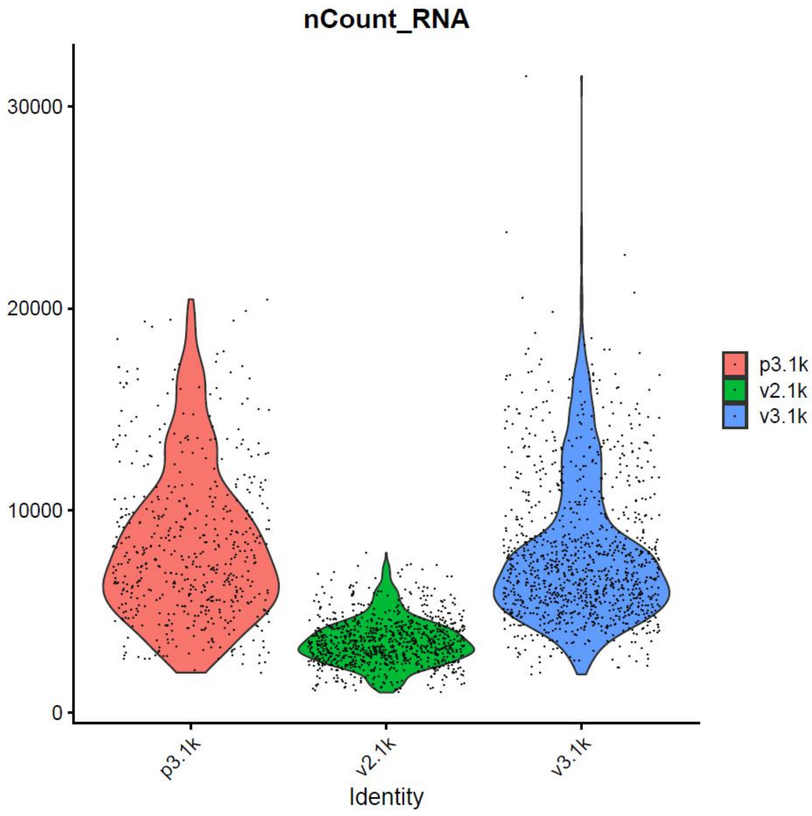
- 被检测基因数量极高的可能表明doublets
- 基因检测方面，v2和v3也有明显的差异，数据的过滤不能用都采用相同的界限
- 有蛋白质分析数据中，有许多细胞几乎没有检测到的基因，但呈双峰分布。这种类型的分布在其他两个数据集中没有看到。考虑到它们都是PBMC数据集，把这个分布看作低质量的库是有意义的
- 过滤高基因检测的细胞(假定doublets)，v3的cutoff为4100，v2的cutoff为2000

➢ Mitochondrial filtering

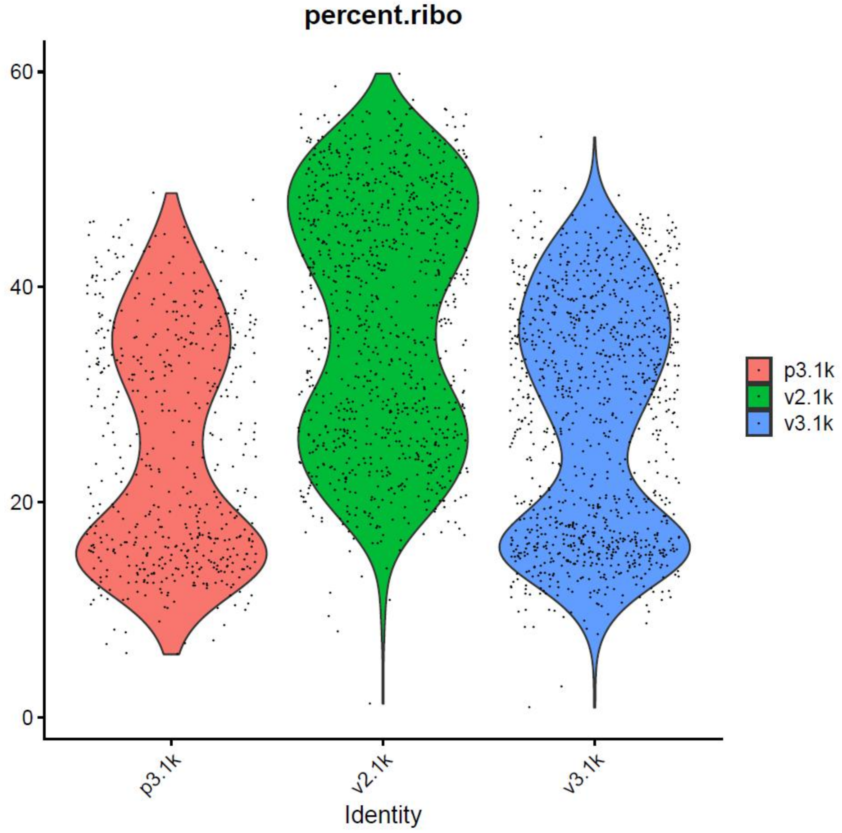
- 有相当多的细胞具有高比例的线粒体reads。如果过滤后我们还有足够的细胞，最好去除这些细胞。另一种是从数据集中移除所有线粒体，剩余的基因仍然有足够的信号
- 以上数据分析中，有高达99.7%的线粒体存在细胞中，所以不太可能有很多细胞类型的标记留在这些细胞中
- 看图作出合理的决定，在哪里划出界限。看以上的数据中，大部分细胞的线粒体读数低于25%

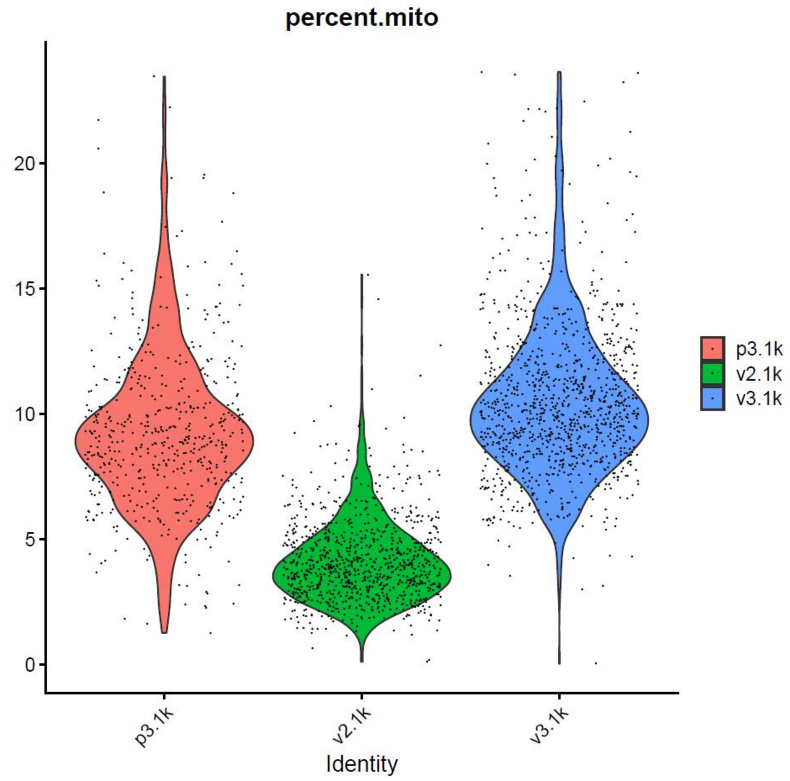
➢ QC-Filterling :nFeature_RNA,nCount_RNA



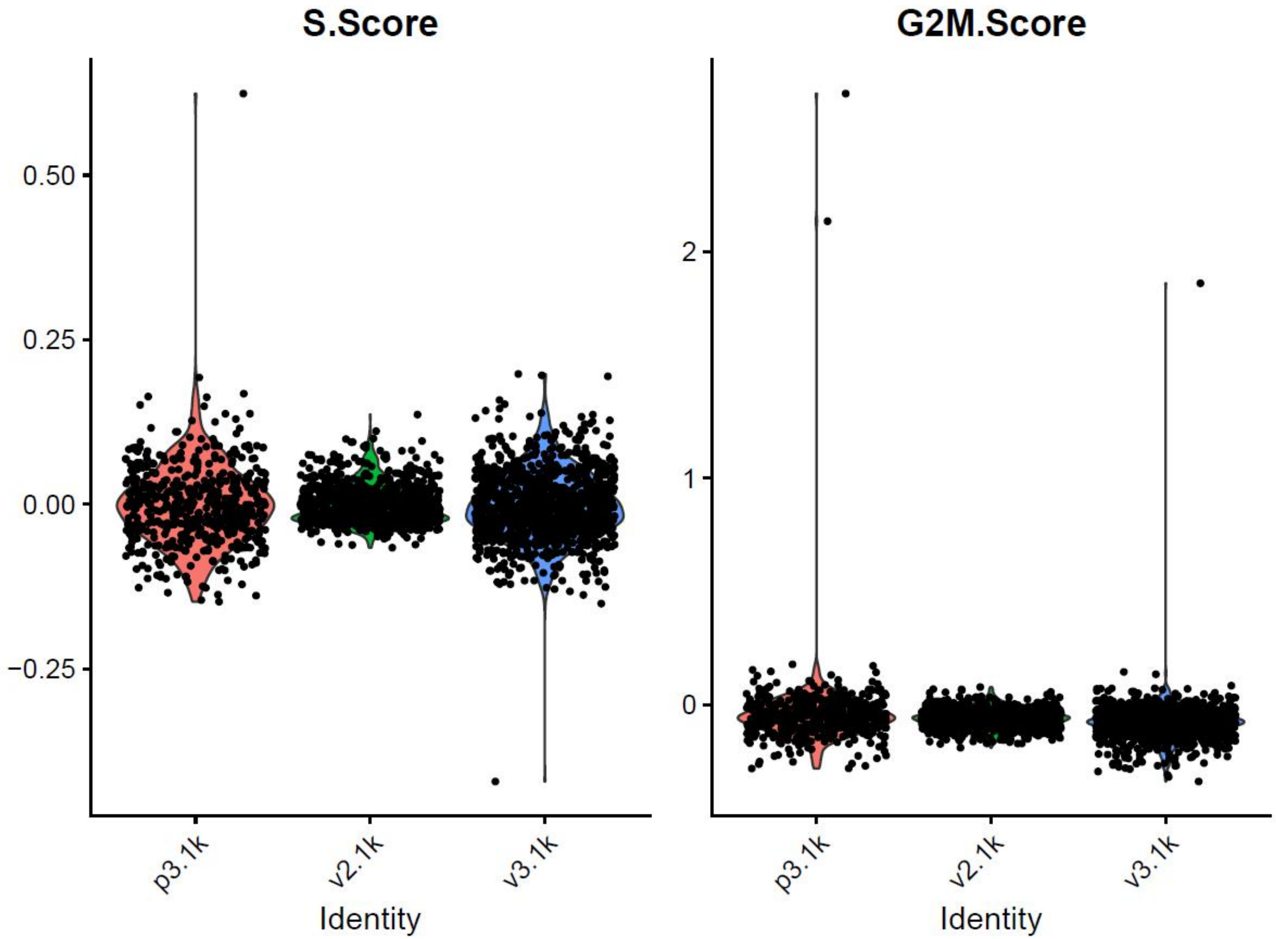


➢ QC-Filterling :mitochondrial, ribosomal proportion





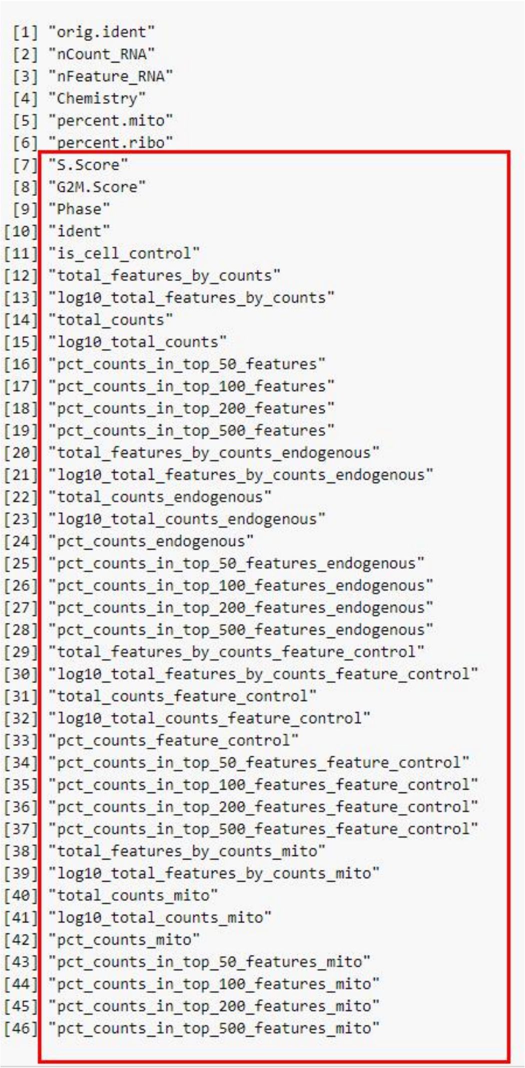
➢ Removal of cell cycle effect



- Calculating cell cycle scores based on a list of know S-phase and G2/M-phase genes.

- **SCATER: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R**

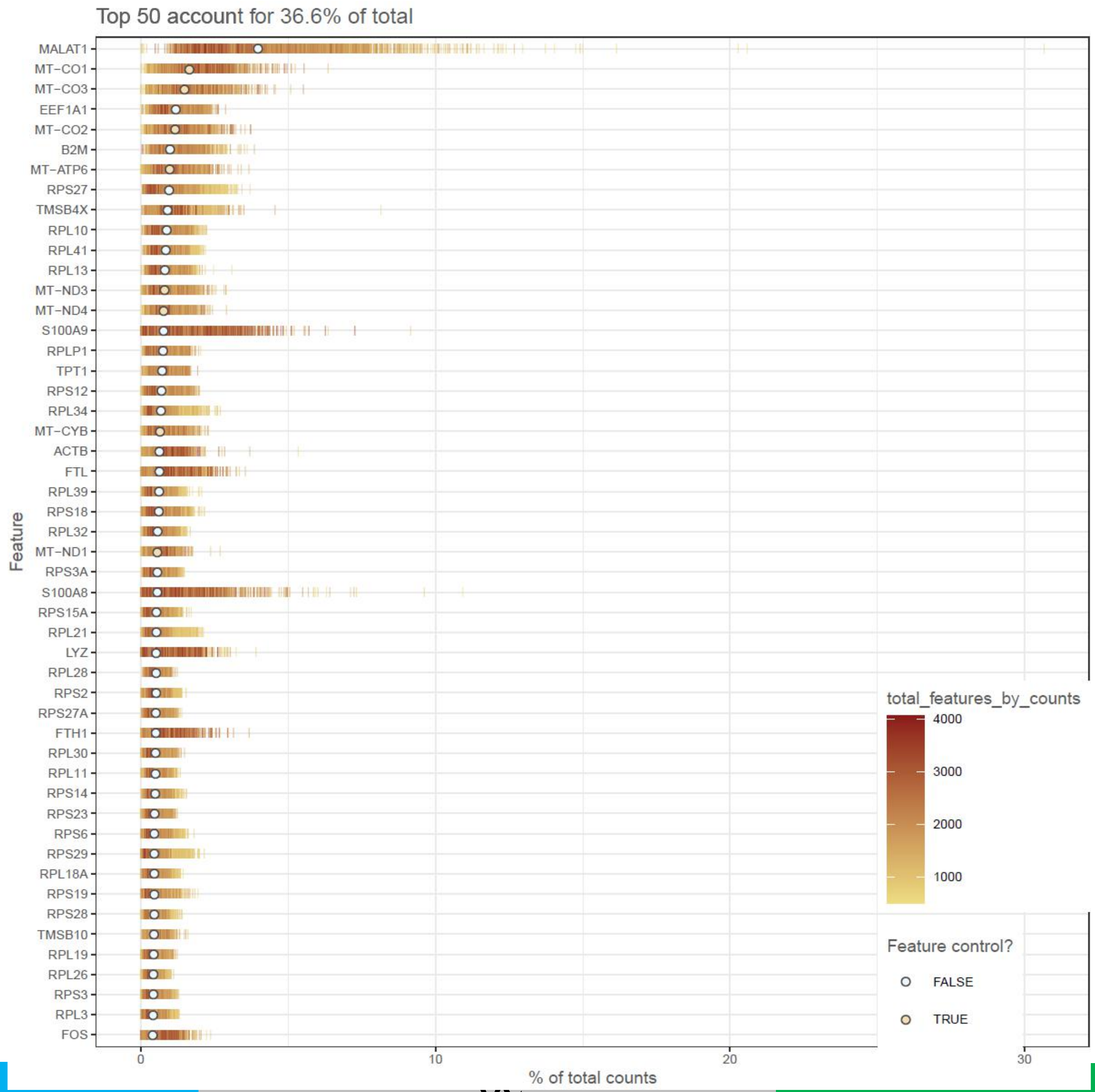
- Most expressed features



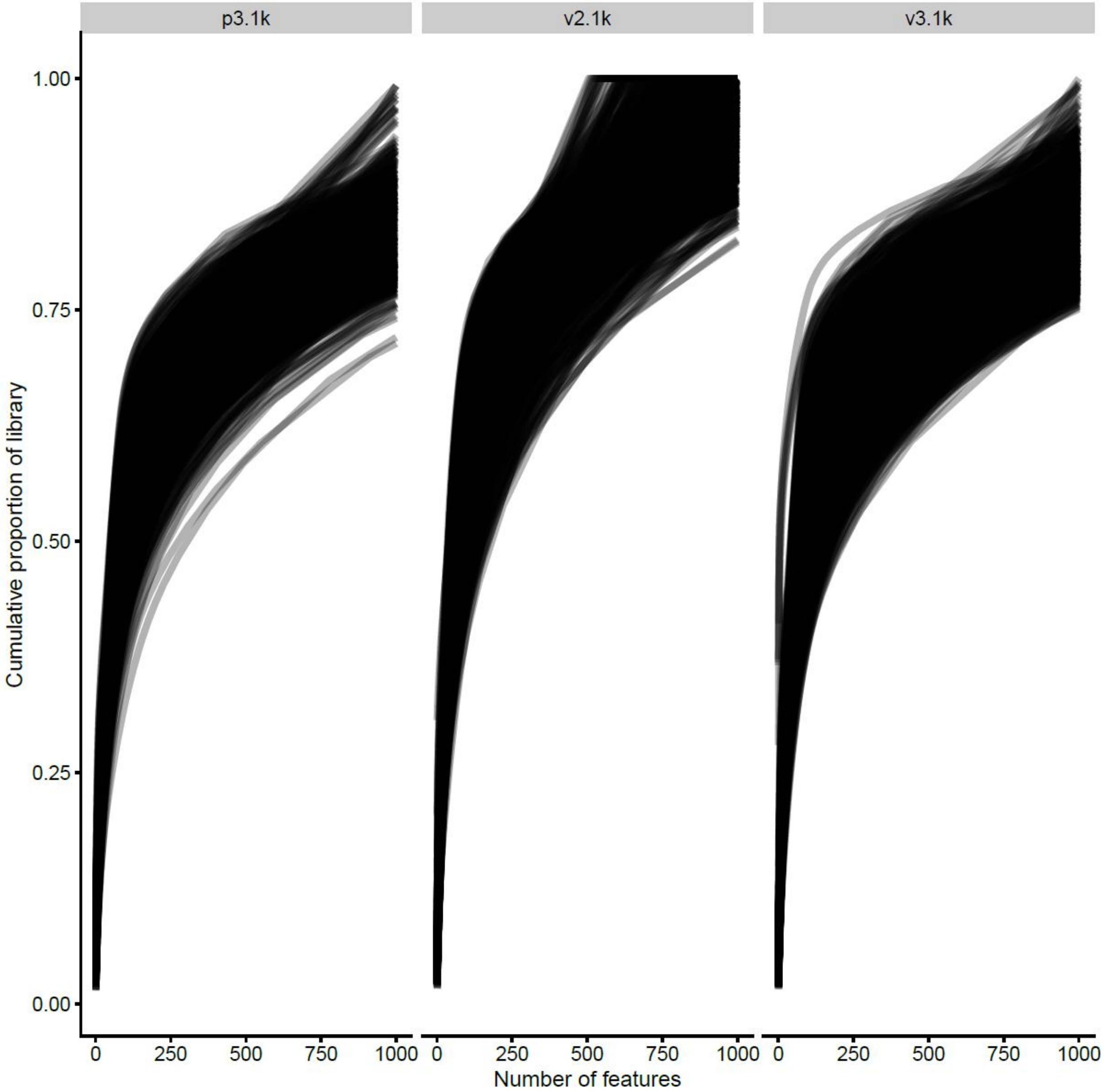
```
 [1] "orig.ident"
 [2] "nCount_RNA"
 [3] "nFeature_RNA"
 [4] "Chemistry"
 [5] "percent.mito"
 [6] "percent.ribo"
 [7] "S.Score"
 [8] "G2M.Score"
 [9] "Phase"
[10] "ident"
[11] "is_cell_control"
[12] "total_features_by_counts"
[13] "log10_total_features_by_counts"
[14] "total_counts"
[15] "log10_total_counts"
[16] "pct_counts_in_top_50_features"
[17] "pct_counts_in_top_100_features"
[18] "pct_counts_in_top_200_features"
[19] "pct_counts_in_top_500_features"
[20] "total_features_by_counts_endogenous"
[21] "log10_total_features_by_counts_endogenous"
[22] "total_counts_endogenous"
[23] "log10_total_counts_endogenous"
[24] "pct_counts_endogenous"
[25] "pct_counts_in_top_50_features_endogenous"
[26] "pct_counts_in_top_100_features_endogenous"
[27] "pct_counts_in_top_200_features_endogenous"
[28] "pct_counts_in_top_500_features_endogenous"
[29] "total_features_by_counts_feature_control"
[30] "log10_total_features_by_counts_feature_control"
[31] "total_counts_feature_control"
[32] "log10_total_counts_feature_control"
[33] "pct_counts_feature_control"
[34] "pct_counts_in_top_50_features_feature_control"
[35] "pct_counts_in_top_100_features_feature_control"
[36] "pct_counts_in_top_200_features_feature_control"
[37] "pct_counts_in_top_500_features_feature_control"
[38] "total_features_by_counts_mito"
[39] "log10_total_features_by_counts_mito"
[40] "total_counts_mito"
[41] "log10_total_counts_mito"
[42] "pct_counts_mito"
[43] "pct_counts_in_top_50_features_mito"
[44] "pct_counts_in_top_100_features_mito"
[45] "pct_counts_in_top_200_features_mito"
[46] "pct_counts_in_top_500_features_mito"
```

https://bioconductor.org/packages/release/bioc/html/scater.html

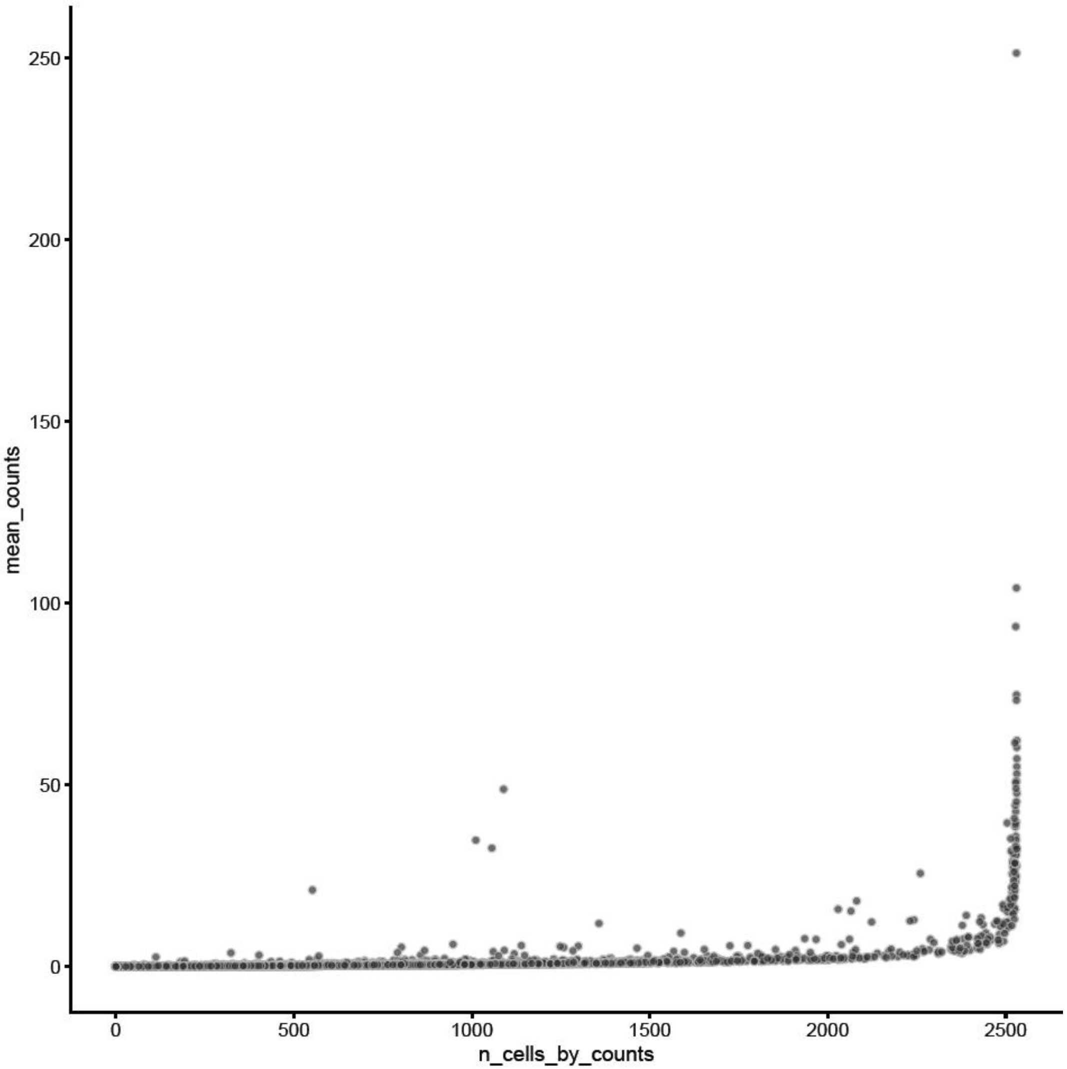
➢ Most expressed features



➢ Cumulative expression

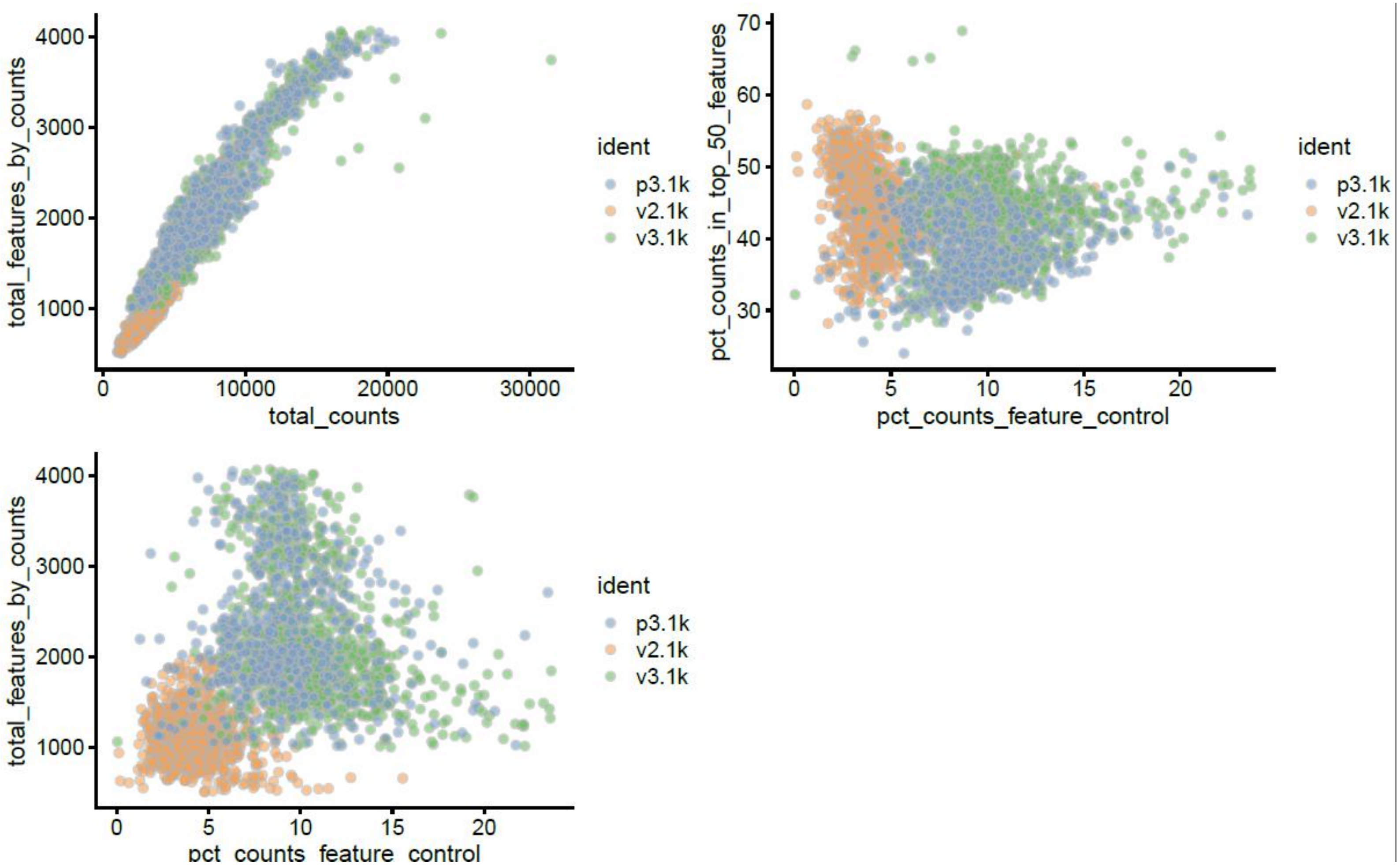


- gene stats



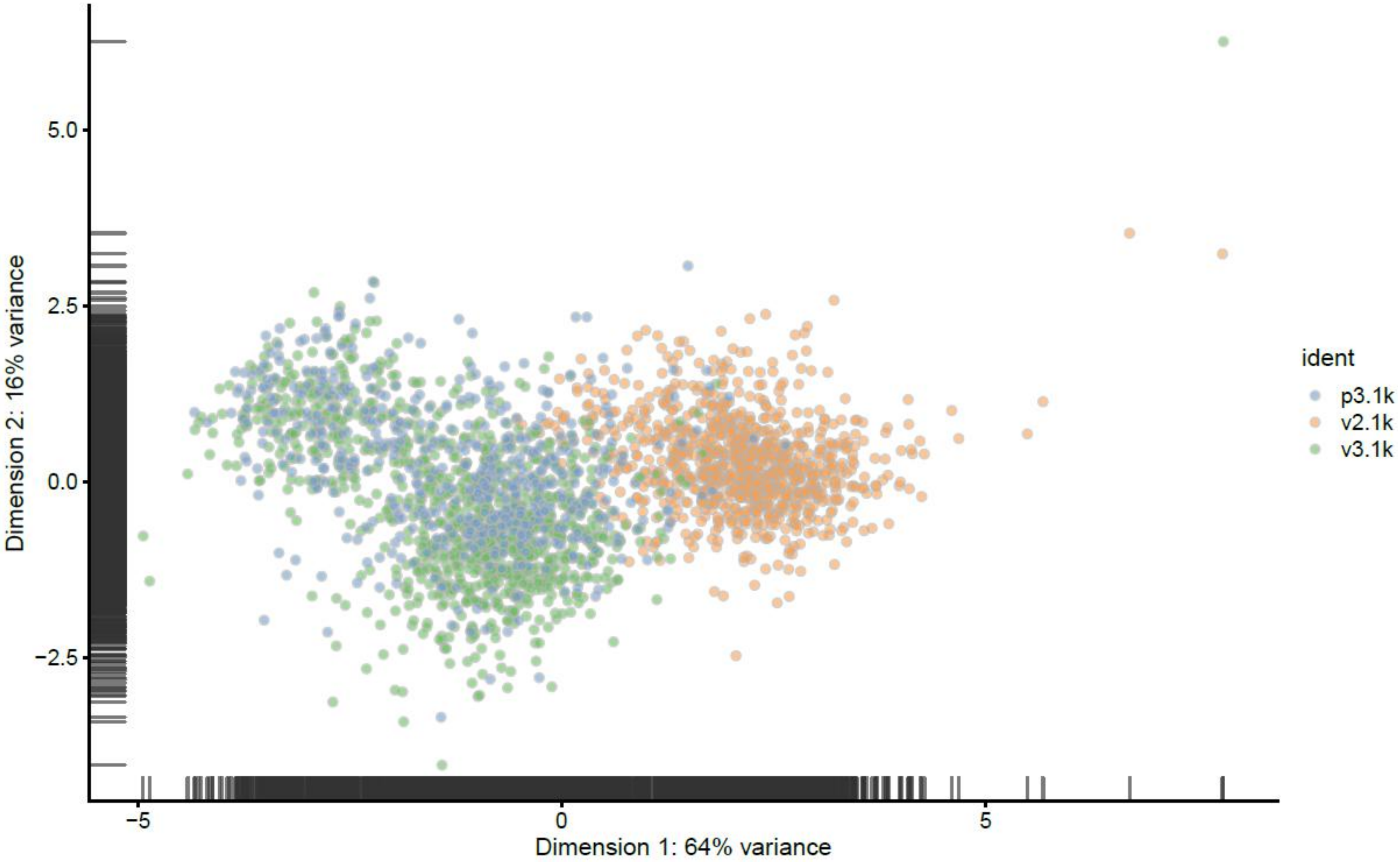
https://bioconductor.org/packages/release/bioc/html/scater.html

- cell stats



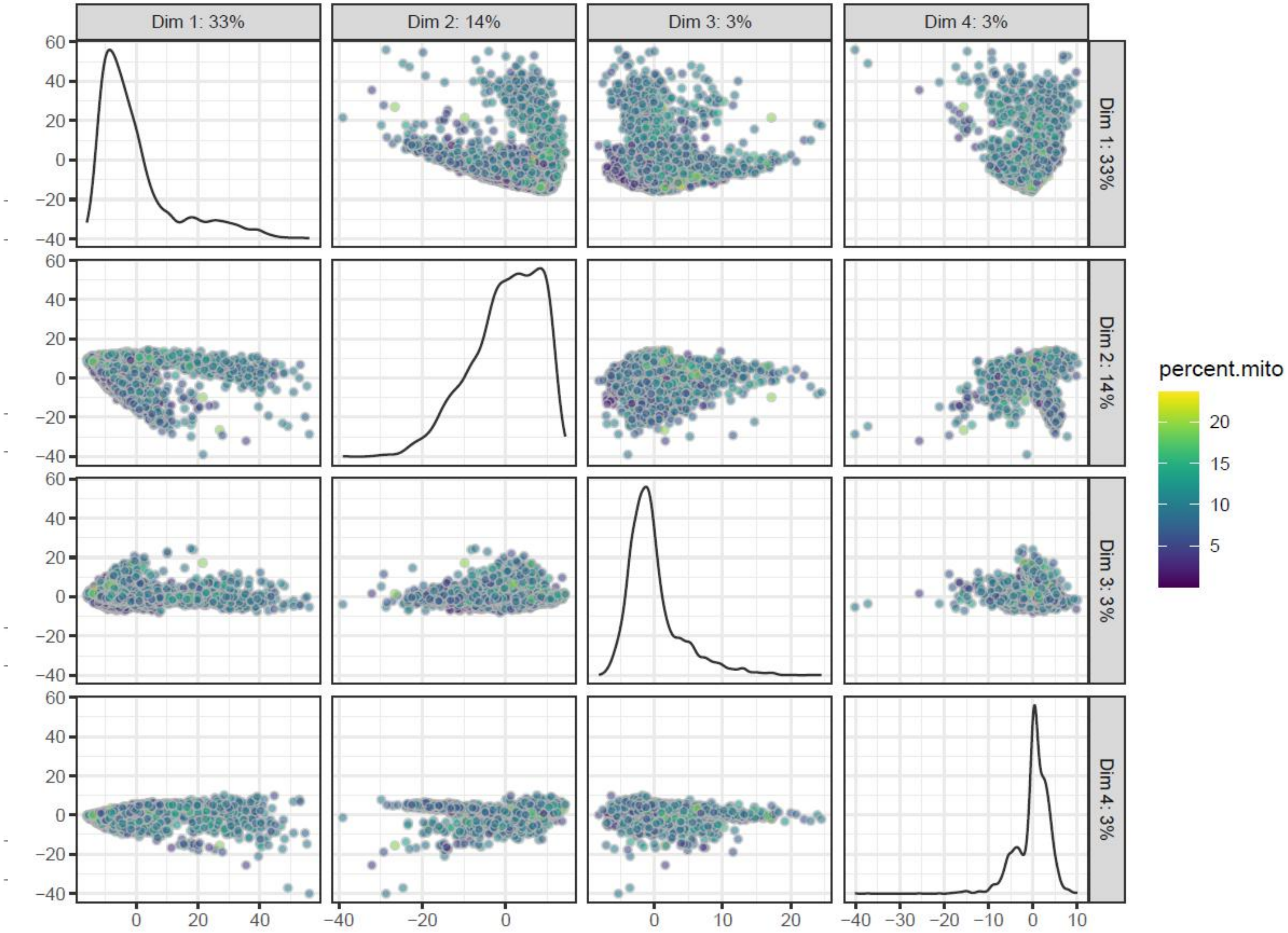
https://bioconductor.org/packages/release/bioc/html/scater.html

➢ PCA for quality control：Identify outliers in QC-stats



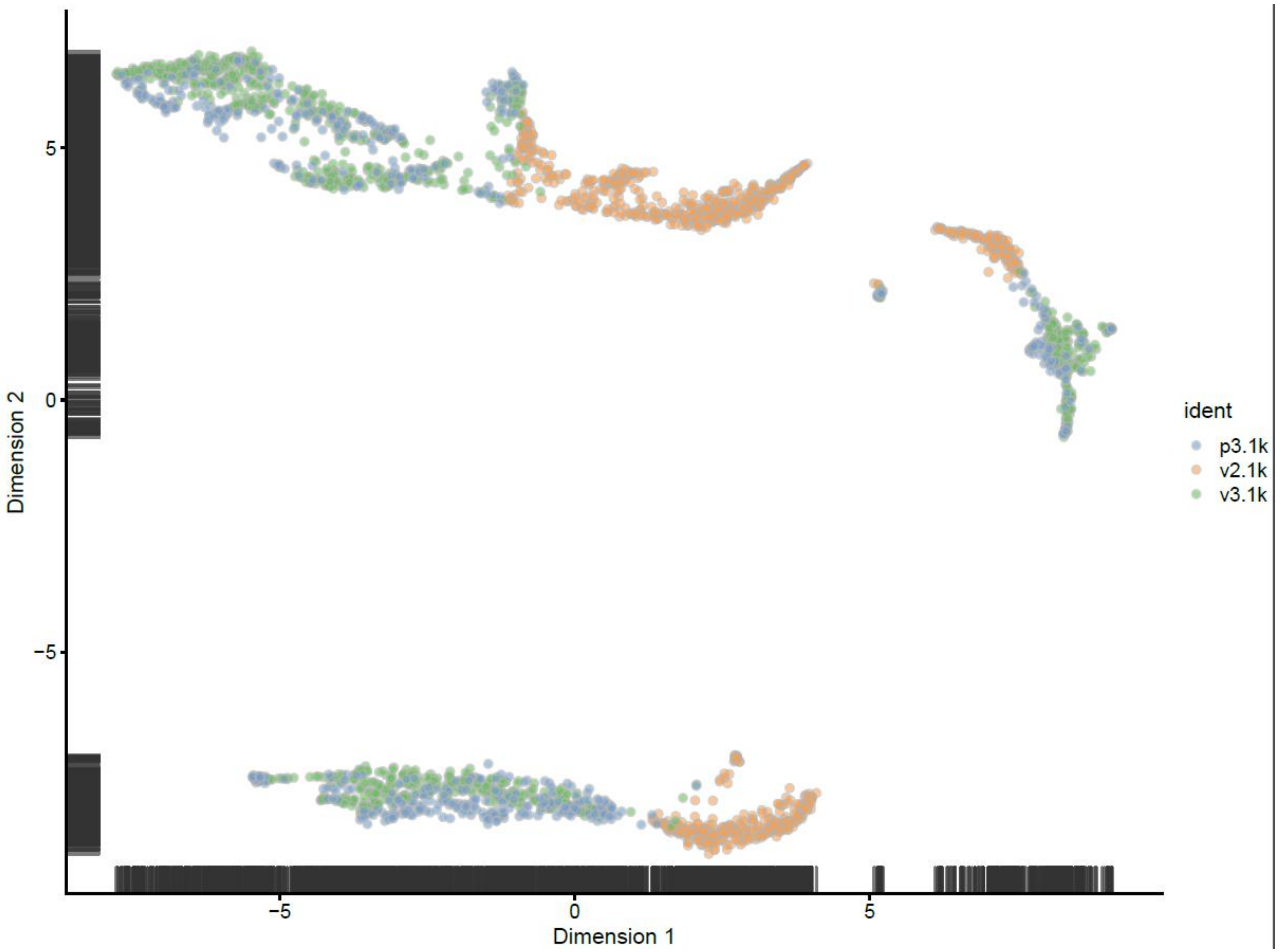
On method of identifying low quality cells is to run PCA on all the qc-stats and then identify outliers in PCA space.

➢ PCA for quality control：Identify outliers in QC-stats

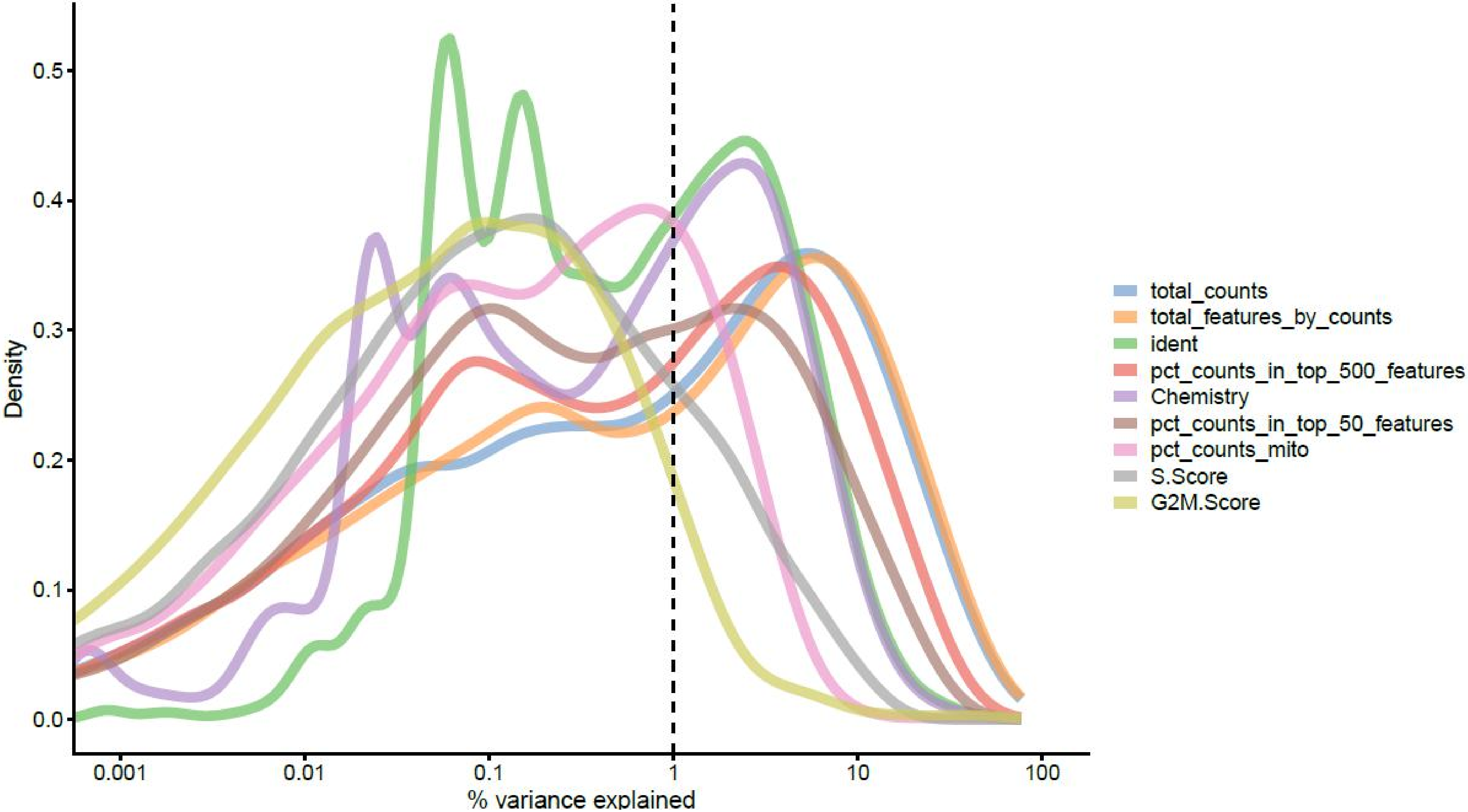


On method of identifying low quality cells is to run PCA on all the qc-stats and then identify outliers in PCA space.

➢ UMAP



➢ Explanatory factors



❖ Reference

- Ding, J., et al. (2020). "Systematic comparison of single-cell and single-nucleus RNA-sequencing methods." Nature Biotechnology 38(6): 737-746.
- Hicks, S. C., et al. (2018). "Missing data and technical variability in single-cell RNA-sequencing experiments." Biostatistics 19(4): 562-578.
- Kang, H. M., et al. (2018). "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation." Nat Biotechnol 36(1): 89-94.
- Klein, A. M., et al. (2015). "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells." Cell 161(5): 1187-1201.
- Liu, S. and C. Trapnell (2016). "Single-cell transcriptome sequencing: recent advances and remaining challenges." F1000Res 5.
- McCarthy, D. J., et al. (2017). "Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R." Bioinformatics 33(8): 1179-1186.
- Mereu, E., et al. (2020). "Benchmarking single-cell RNA-sequencing protocols for cell atlas projects." Nat Biotechnol 38(6): 747-755.
- Slyper, M., et al. (2020). "A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors." Nat Med 26(5): 792-802.
- Stuart, T. and R. Satija (2019). "Integrative single-cell analysis." Nat Rev Genet 20(5): 257-272.
- Wagner, A., et al. (2016). "Revealing the vectors of cellular identity with single-cell genomics." Nature Biotechnology 34(11): 1145-1160.
- Wagner, D. E. and A. M. Klein (2020). "Lineage tracing meets single-cell omics: opportunities and challenges." Nat Rev Genet.
- Ziegenhain, C., et al. (2017). "Comparative Analysis of Single-Cell RNA Sequencing Methods." Mol Cell 65(4): 631-643.e634.
- Wu, Y., Zhang, K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. Nat Rev Nephrol (2020).
- Lafzi, A., Moutinho, C., Picelli, S. et al. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. Nat Protoc 13, 2742–2757 (2018).
- Malte D Luecken;Fabian J Theis.Current best practices in single-cell RNA-seq analysis: a tutorial.Mol Syst Biol. (2019)

➢ 下期课程

- ❖降维与聚类（不同降维算法原理）
- ❖细胞亚群间表达差异分析
- ❖拟时序分析
- ❖细胞亚群注释
- ❖样本间表达量差异分析

ANY QUESTIONS！

谢谢！