转录组分析

25.5.28

邓之怡

Content

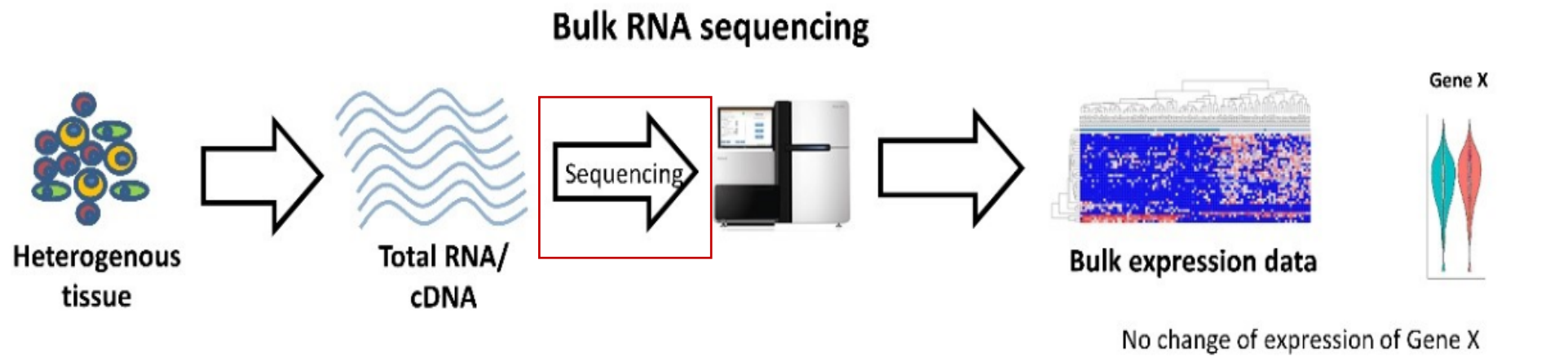
- 转录组定义
- 转录组测序技术
- 分析思路
- 实操

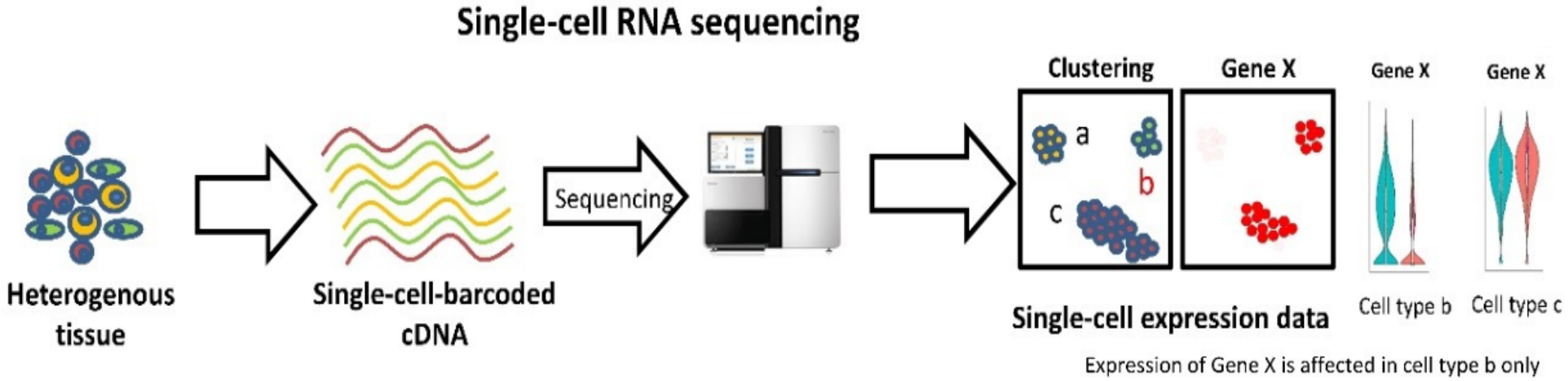
转录组定义

RNA-seq (RNA-sequencing): examine the quantity and sequences of RNA in a sample using next-generation sequencing (NGS).
It analyzes the transcriptome, indicating which of the genes encoded in our DNA are turned on or off and to what extent.

转录组测序方法

- 混合细胞转录组测序 bulk RNA-seq
- 单细胞转录组测序 single cell RNA-seq





转录组测序技术

- 短读长 short read length：

Illumina：

准确率高

- 长读长 long read length：

PacBio Iso-Seq：10-20kb

Oxford Nanopore：>100kb

准确率低

Next-generation sequencing

generations of tech

1st generation

Sanger sequencing, 1 read at time (1977-present)

2nd generation

“short-reads” sequencing, millions of reads at a time

Illumina (2006 - present)

3rd generation

“single molecule” sequencing, no amplification needed

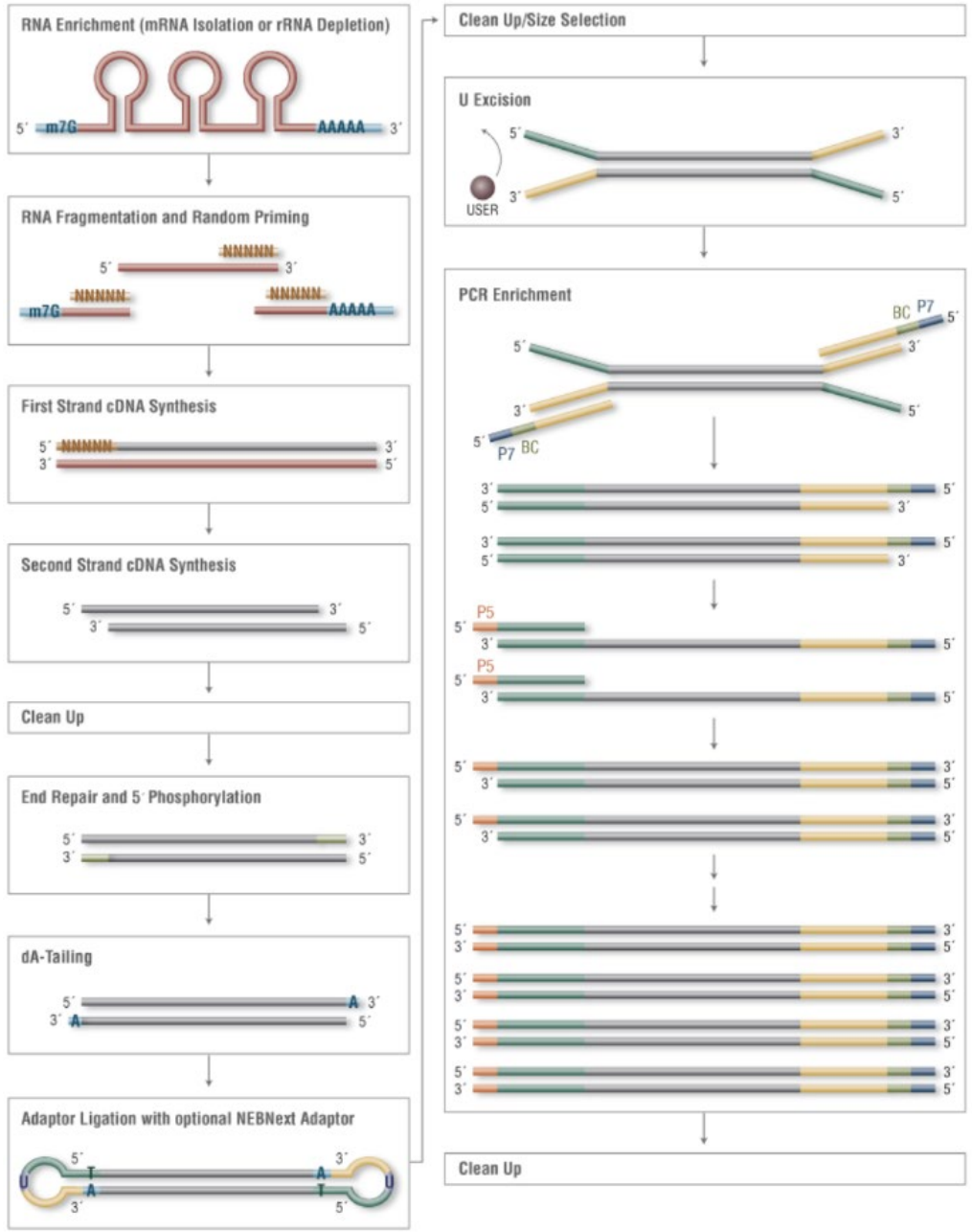
PacBio, SMRT (2010-present)

Oxford nanopore, minION (2015 -present)

转录组测序 Workflow

Illumina

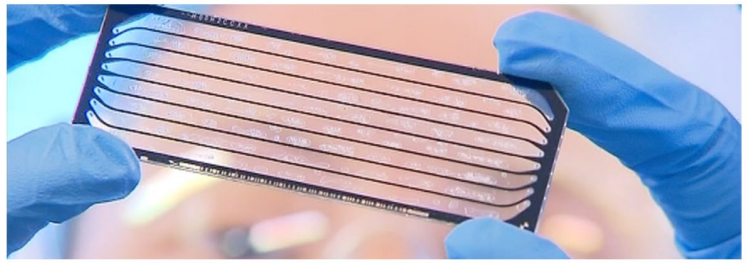
Library preparing 建库
为测序做准备

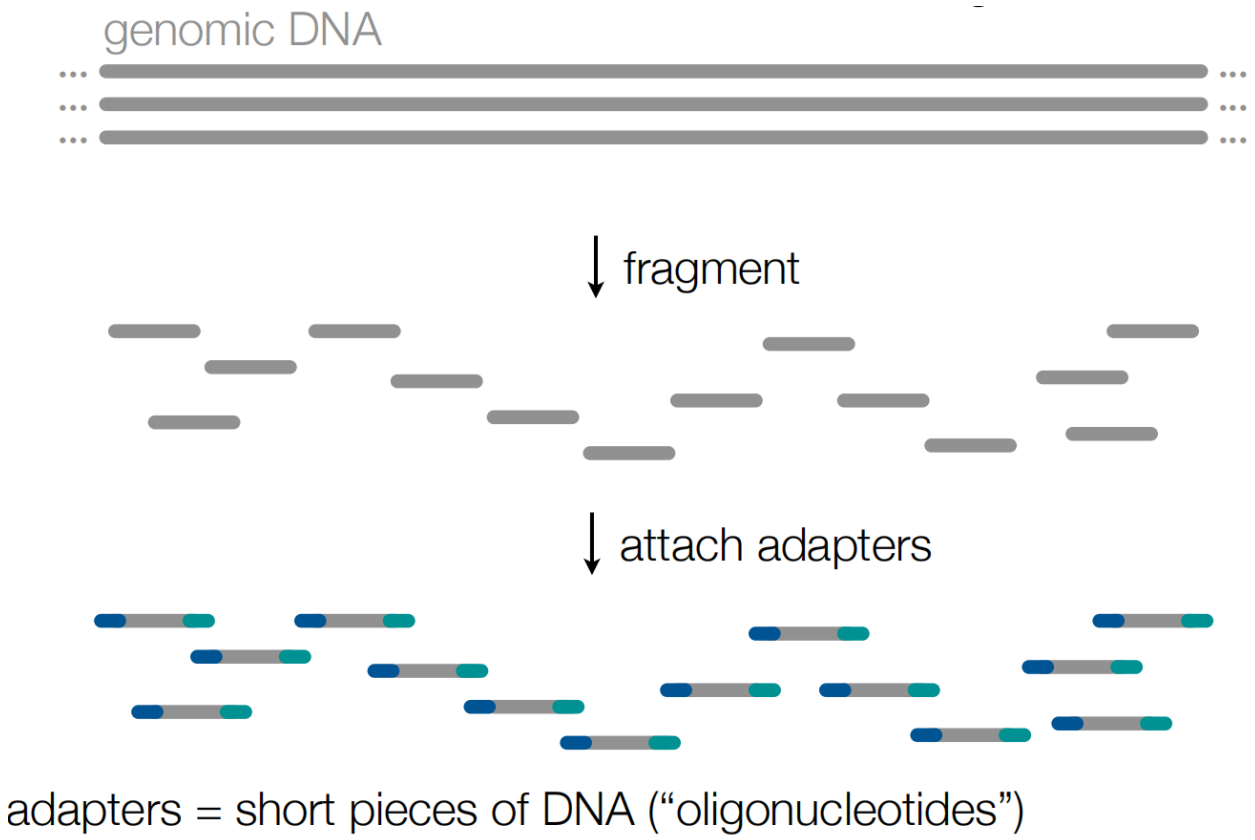


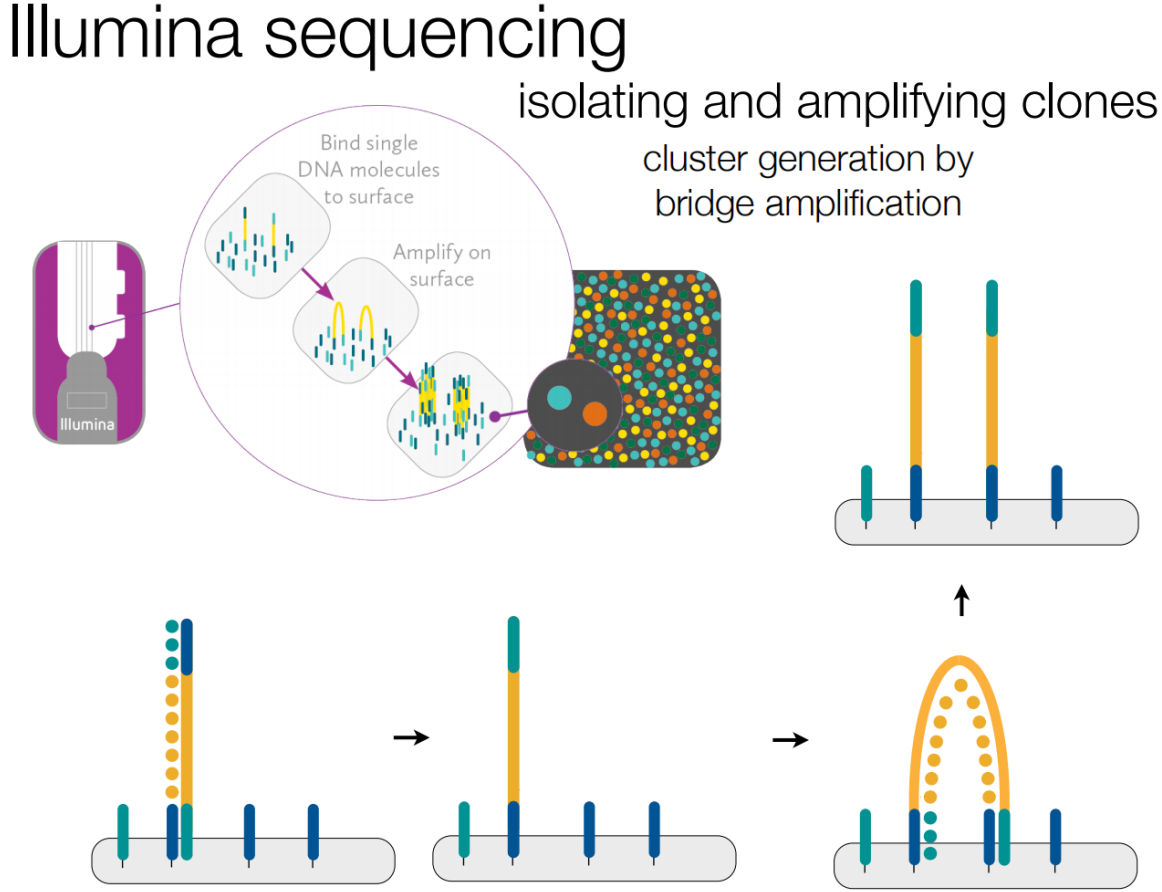
- 逆转录、建库——片段化——加修饰——分离——扩增成簇——合成测序

转录组测序 Workflow

Illumina







- 逆转录、建库——片段化——加修饰——分离——扩增成簇——合成测序

转录组测序 Workflow

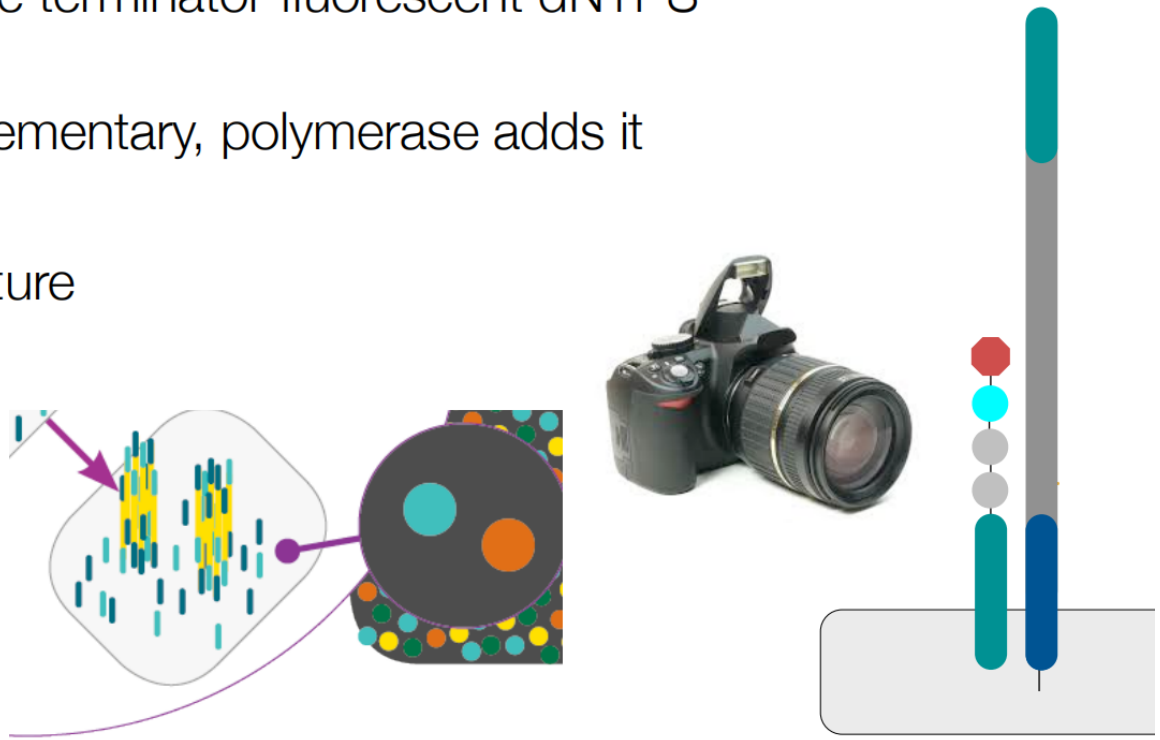
Illumina

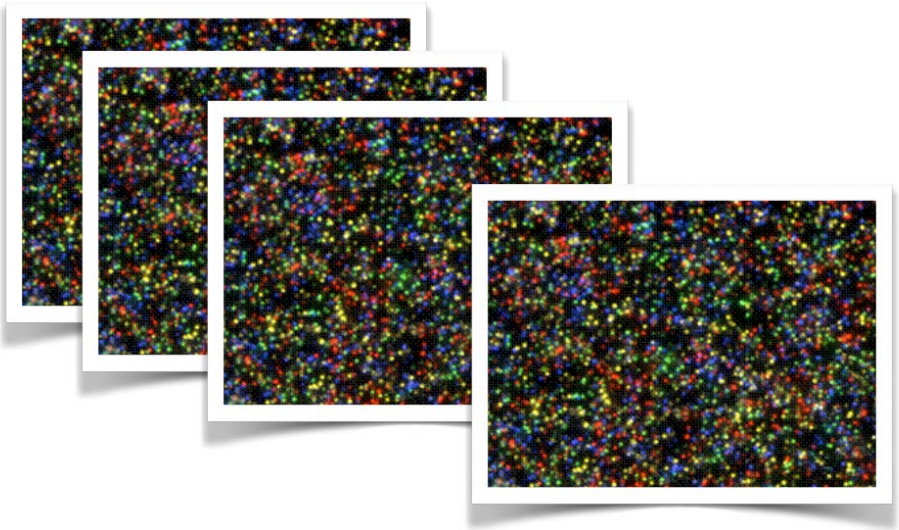
sequencing by synthesis

polymerase
reversible terminator fluorescent dNTPS

if complementary, polymerase adds it

take picture



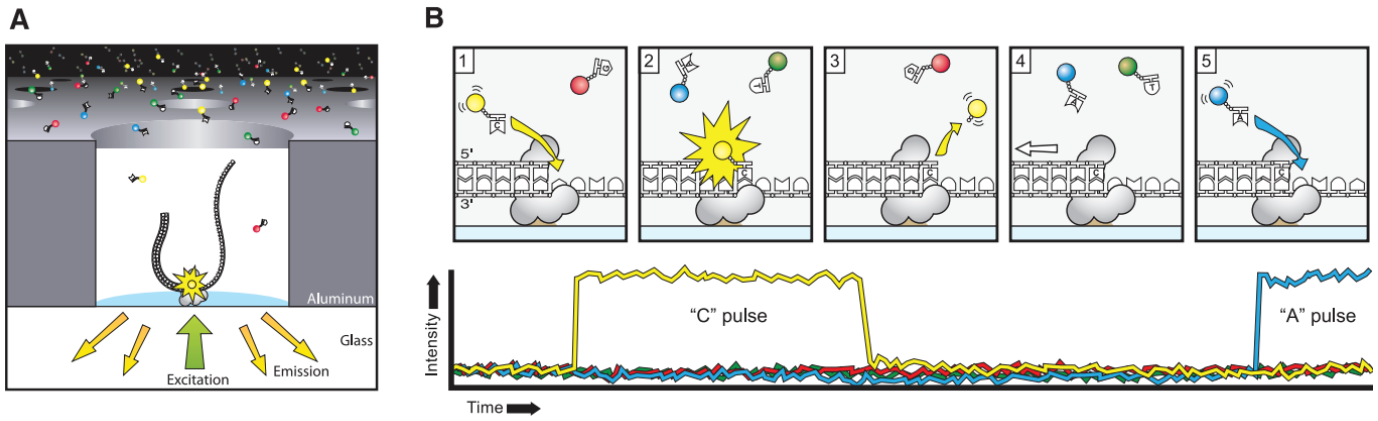


- 逆转录、建库——片段化——加修饰——分离——扩增成簇——合成测序

转录组测序 Workflow

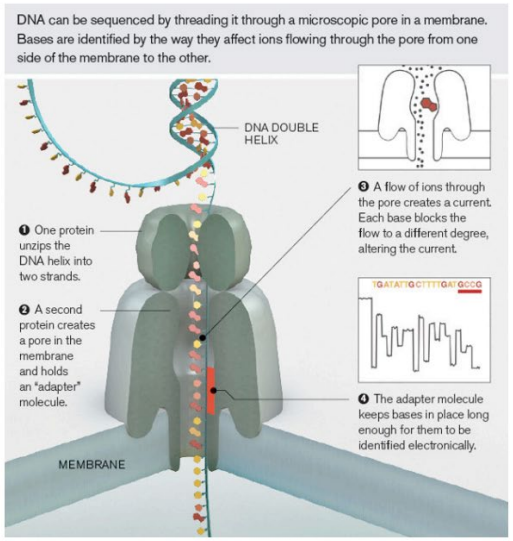
第三代测序

PacBio Single Molecule Real-Time Sequencing (SMRT)



single molecule; 50,000 bp

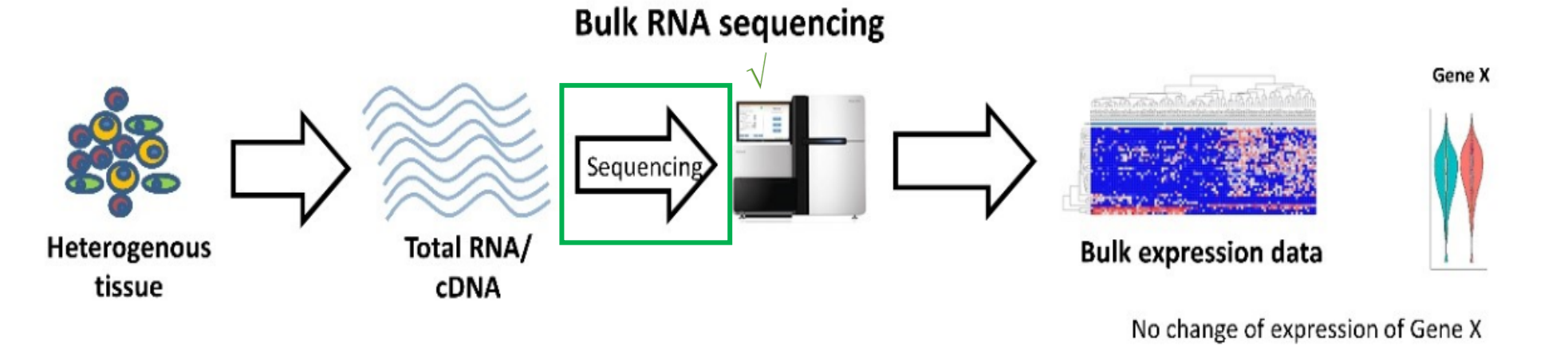
Oxford nanopore sequencing (minION)

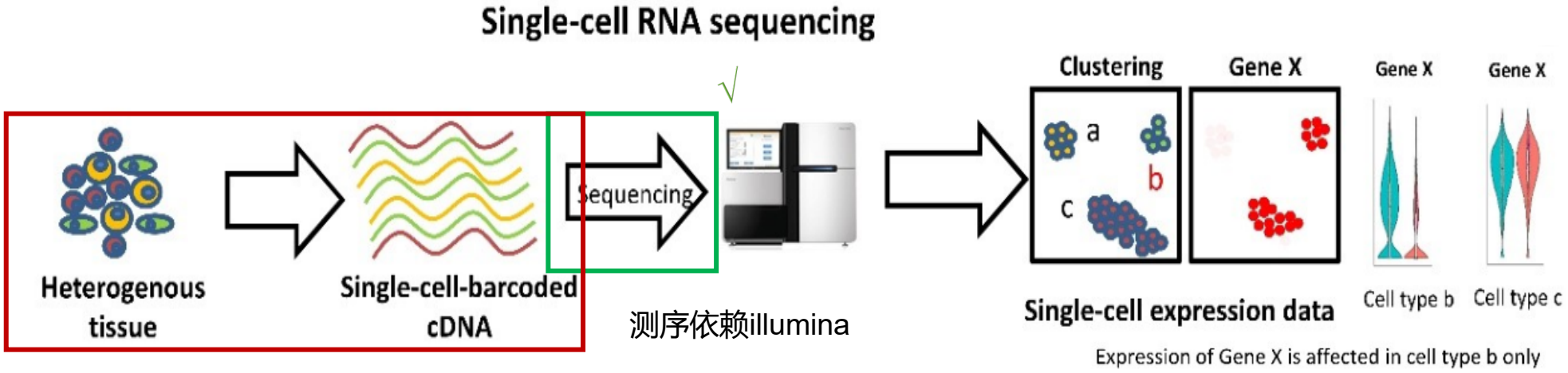


single molecule, 100,000 bp

转录组测序方法

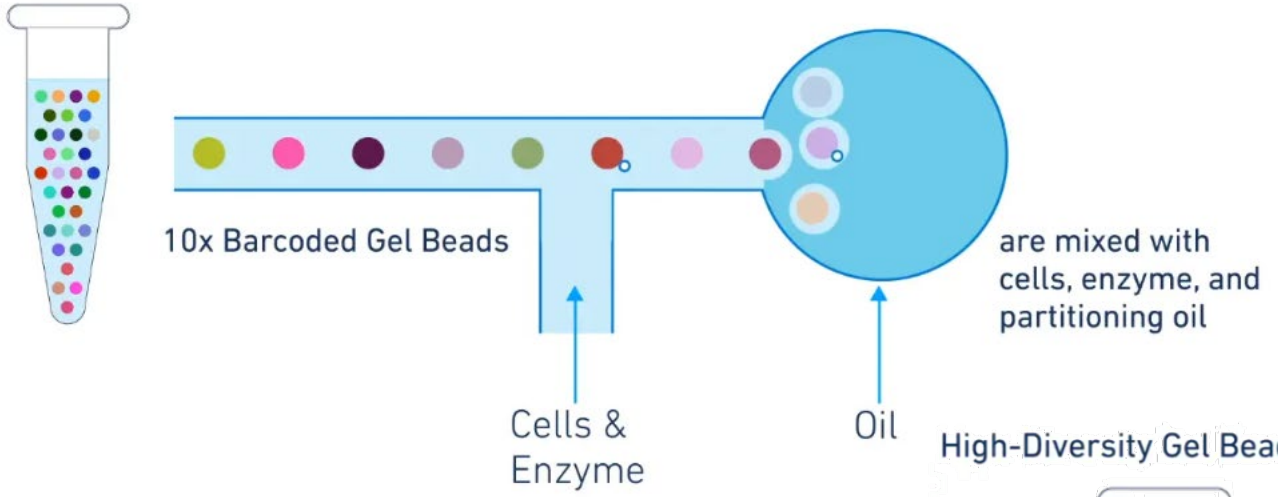
- 混合细胞转录组测序 bulk RNA-seq
- 单细胞转录组测序 single cell RNA-seq

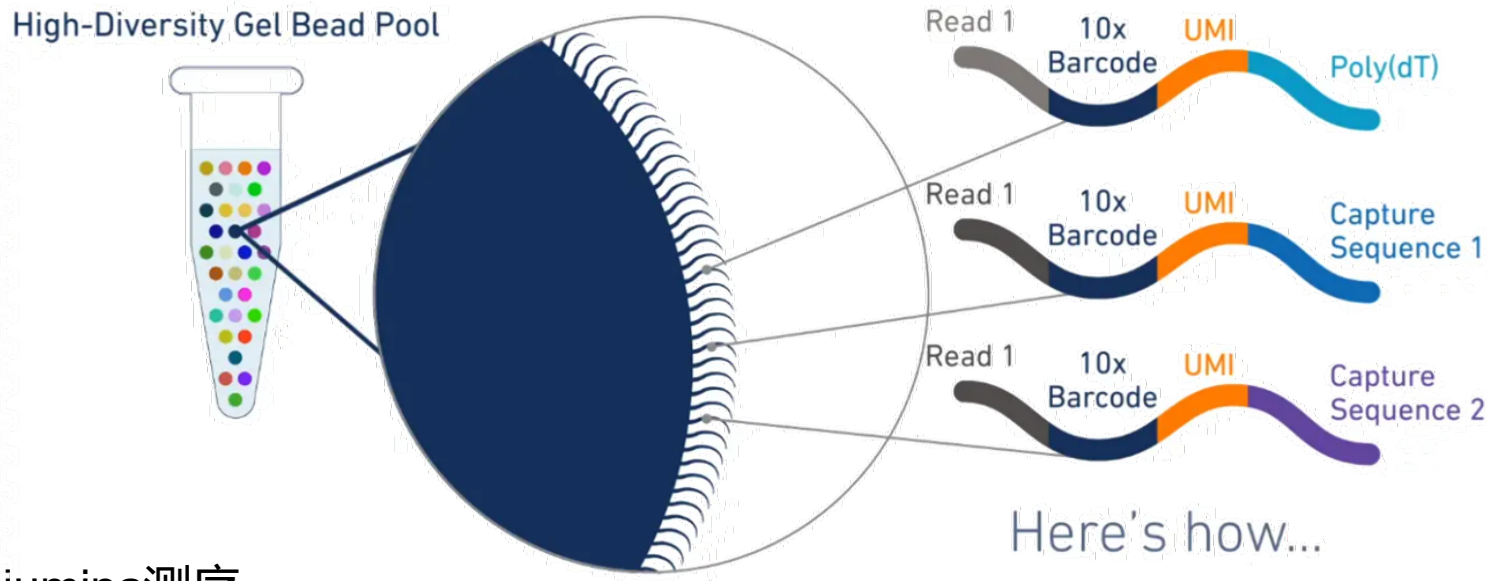




单细胞转录组测序 Workflow

10X genomics

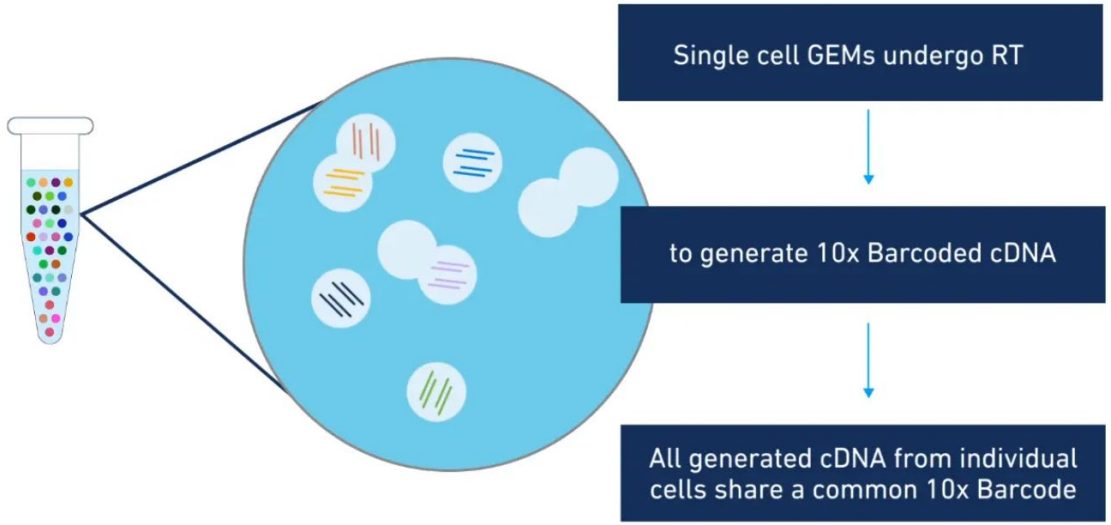


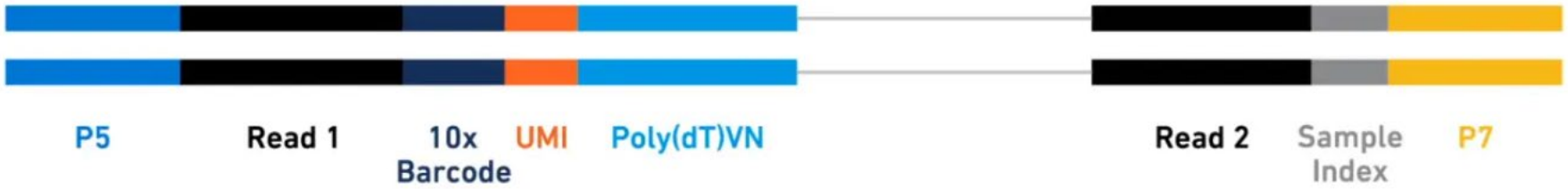


- 细胞分选——加条码标记——建库——Iliumina测序

单细胞转录组测序 Workflow

10X genomics





- 细胞分选——加条码标记——建库——Illumina测序

单细胞转录组测序数据结构

3 matrix：
barcodes
genes
matrix

```
jmzengdeMacBook-Pro:SRR7722939 jmzeng$ head barcodes.tsv
AAACCTGAGCGAAGGG-1
AAACCTGAGGTCATCT-1
AAACCTGAGTCCTCCT-1
AAACCTGCACCAGCAC-1
AAACCTGGTAACGTTC-1
AAACCTGGTAAGGATT-1
AAACCTGGTTGTCGCG-1
AAACCTGTCCTGCCAT-1
AAACGGGAGTCATCCA-1
AAACGGGCATGGATGG-1
```

```
jmzengdeMacBook-Pro:SRR7722939 jmzeng$ head genes.tsv
hg38_ENSG00000243485    hg38_RP11-34P13.3
hg38_ENSG00000237613    hg38_FAM138A
hg38_ENSG00000186092    hg38_OR4F5
hg38_ENSG00000238009    hg38_RP11-34P13.7
hg38_ENSG00000239945    hg38_RP11-34P13.8
hg38_ENSG00000239906    hg38_RP11-34P13.14
hg38_ENSG00000241599    hg38_RP11-34P13.9
hg38_ENSG00000279928    hg38_FO538757.3
hg38_ENSG00000279457    hg38_FO538757.2
hg38_ENSG00000228463    hg38_AP006222.2
```

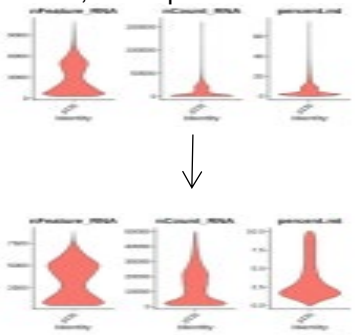
```
jmzengdeMacBook-Pro:SRR7722939 jmzeng$ head matrix.mtx
%%MatrixMarket matrix coordinate integer general
%
33694 2049 1878957
28 1 1
55 1 2
59 1 1
60 1 1
62 1 1
78 1 2
111 1 1
```

```
2049 barcodes.tsv
  33694 genes.tsv
1878960 matrix.mtx
```

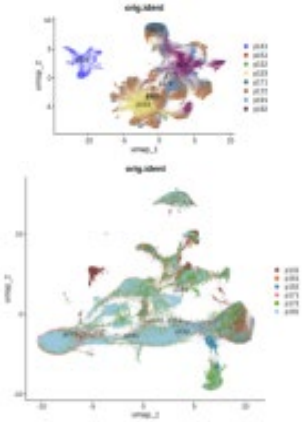
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

单细胞转录组数据分析流程

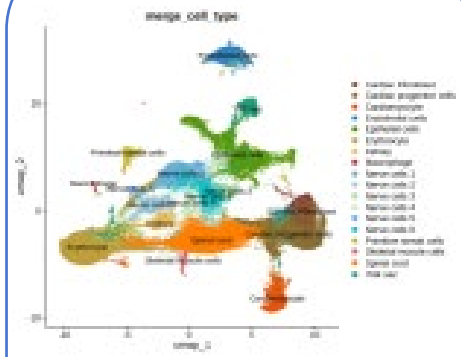
10X, 6 samples



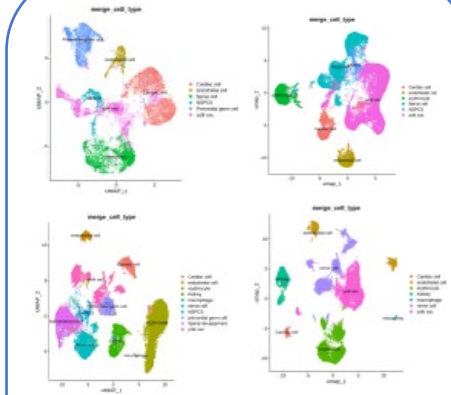
Quality Control



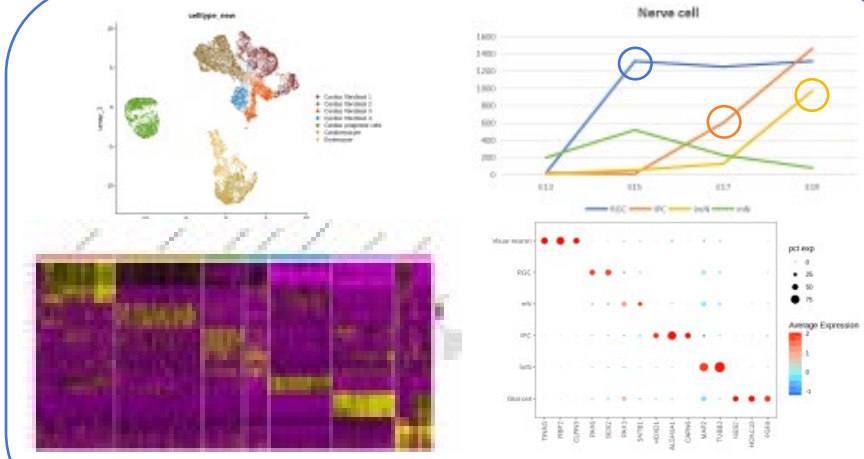
Harmony remove batch effect



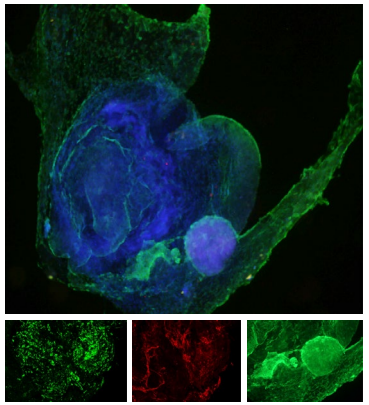
Cell type annotation



Verification



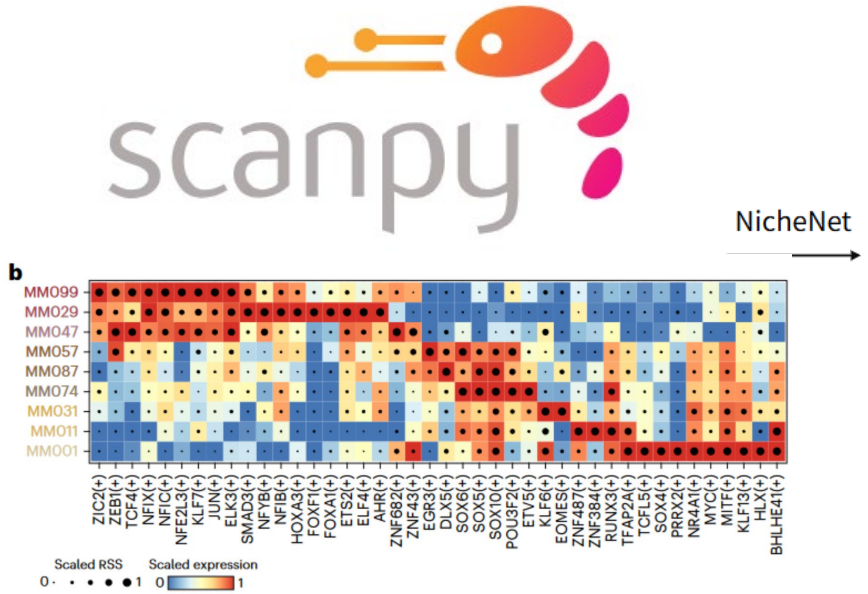
Detailed subtype analysis

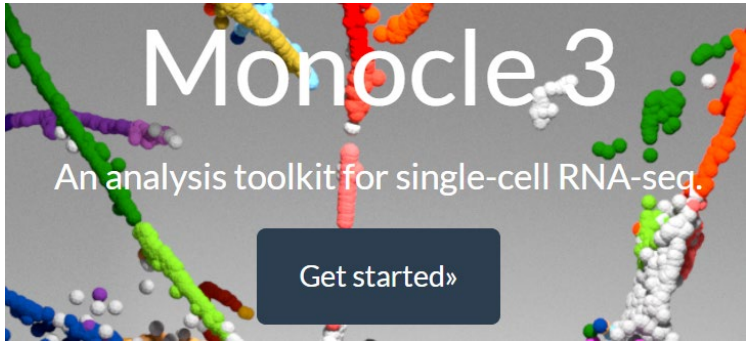


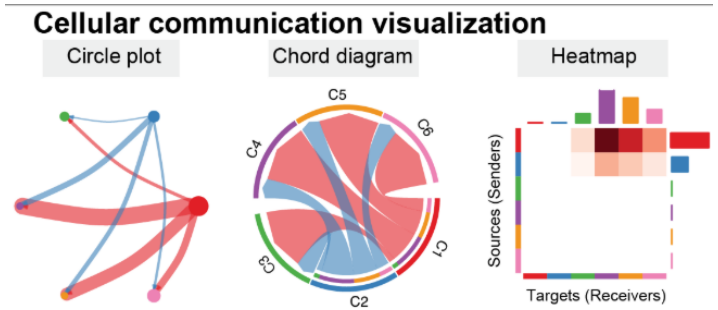
immunofluorescent staining
If possible

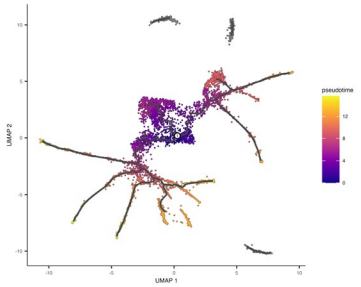
单细胞转录组数据分析流程

常用包：
python：
scanpy： https://scanpy.readthedocs.io/en/stable/
Nichenet： https://github.com/saeyslab/nichenetr
Scenic： https://github.com/aertslab/SCENIC
R：
seurat:https://satijalab.org/seurat/
monocle： https://cole-trapnell-lab.github.io/monocle3/
CellChat:https://github.com/sqjin/CellChat



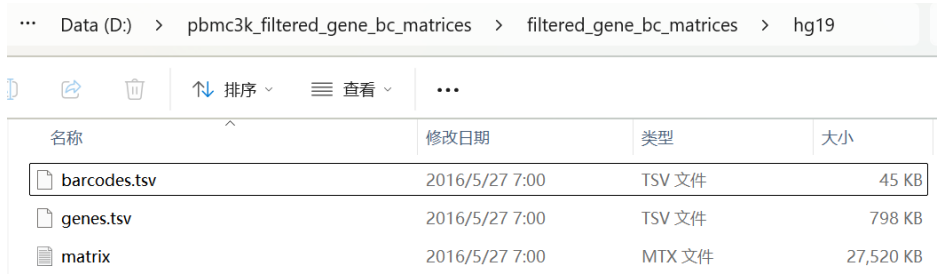






单细胞转录组数据分析流程

环境配置 & 数据下载
Rstudio version：4.3.3 （R>=4.0即可）
Seurat version：5.2.0



安装包：
install.packages("dplyr")
install.packages("patchwork")
install.packages('Seurat')

调用包：
library(dplyr)
library(Seurat)
library(patchwork)

数据下载：https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz
数据读取：

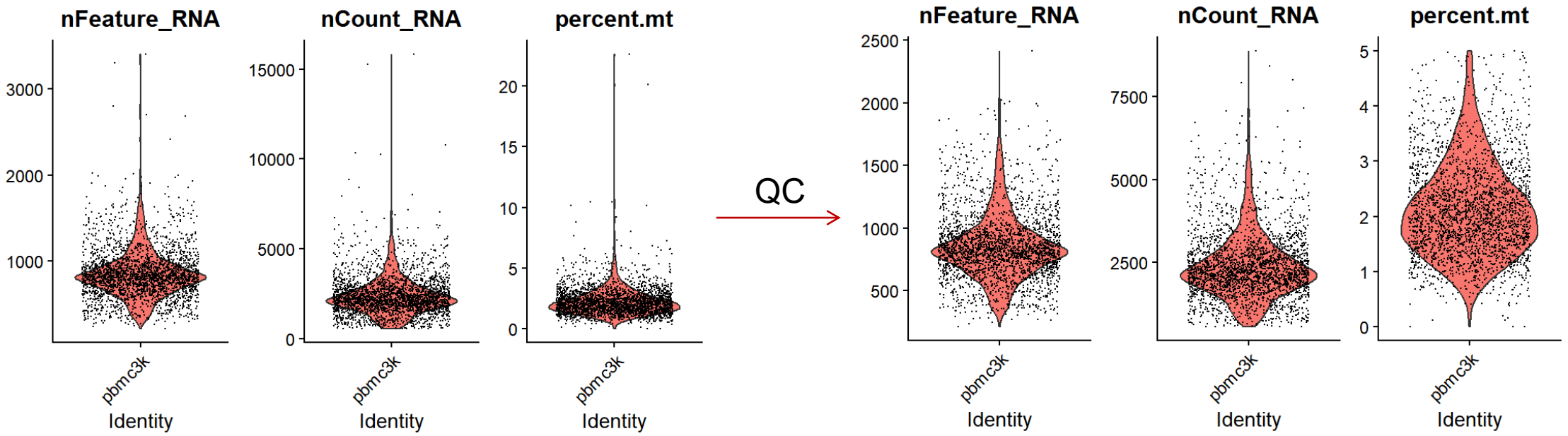
```
pbmc.data <- Read10X(data.dir = "D:/pbmc3k_filtered_gene_bc_matrices/filtered_gene_bc_matrices/hg19/")
#自己的3个matrix存放的位置
# Initialize the Seurat object with the raw (non-normalized data).
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k", min.cells = 3, min.features = 200)
pbmc
```

```
> library(dplyr)
> library(patchwork)
> # Load the PBMC dataset
> pbmc.data <- Read10X(data.dir = "D:/pbmc3k_filtered_gene_bc_matrices/filtered_gene_bc_matrices/hg19/")
> # Initialize the Seurat object with the raw (non-normalized data).
> pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k", min.cells = 3, min.features = 200)
Warning: Feature names cannot have underscores ('_'), replacing with dashes ('-')
> pbmc
An object of class Seurat
13714 features across 2700 samples within 1 assay
Active assay: RNA (13714 features, 0 variable features)
 1 layer present: counts
```

单细胞转录组数据分析流程

质控：

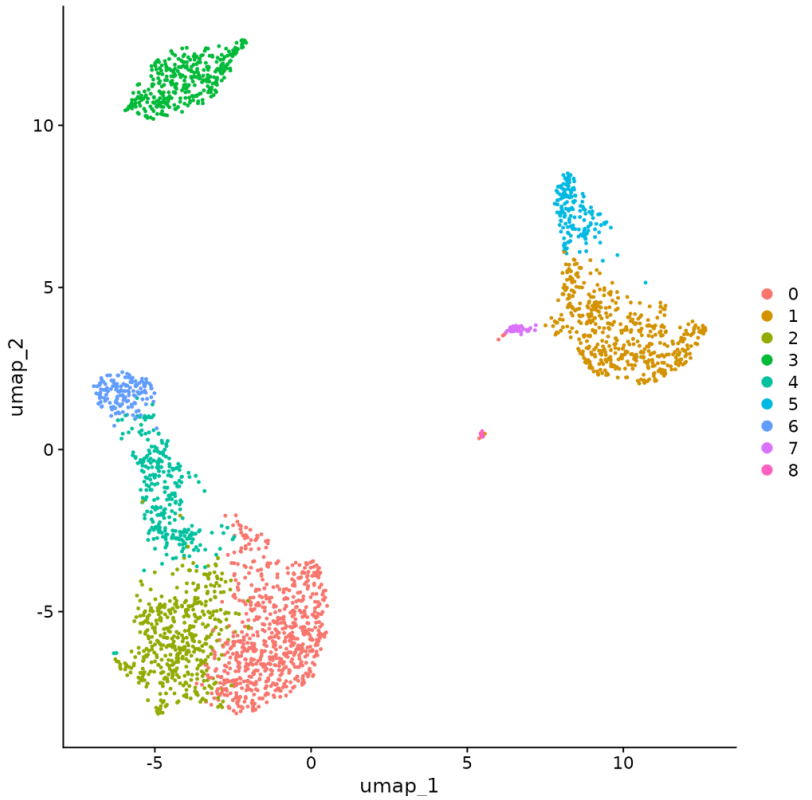
```
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")
VlnPlot(pbmc, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt < 5)
```



单细胞转录组数据分析流程

分群：

```
all.genes <- rownames(pbmc)
pbmc <- ScaleData(pbmc, features = all.genes)
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))
print(pbmc[["pca"]], dims = 1:5, nfeatures = 5)
VizDimLoadings(pbmc, dims = 1:2, reduction = "pca")
pbmc <- FindNeighbors(pbmc, dims = 1:10)
pbmc <- FindClusters(pbmc, resolution = 0.5)
pbmc <- RunUMAP(pbmc, dims = 1:10)
DimPlot(pbmc, reduction = "umap")
```

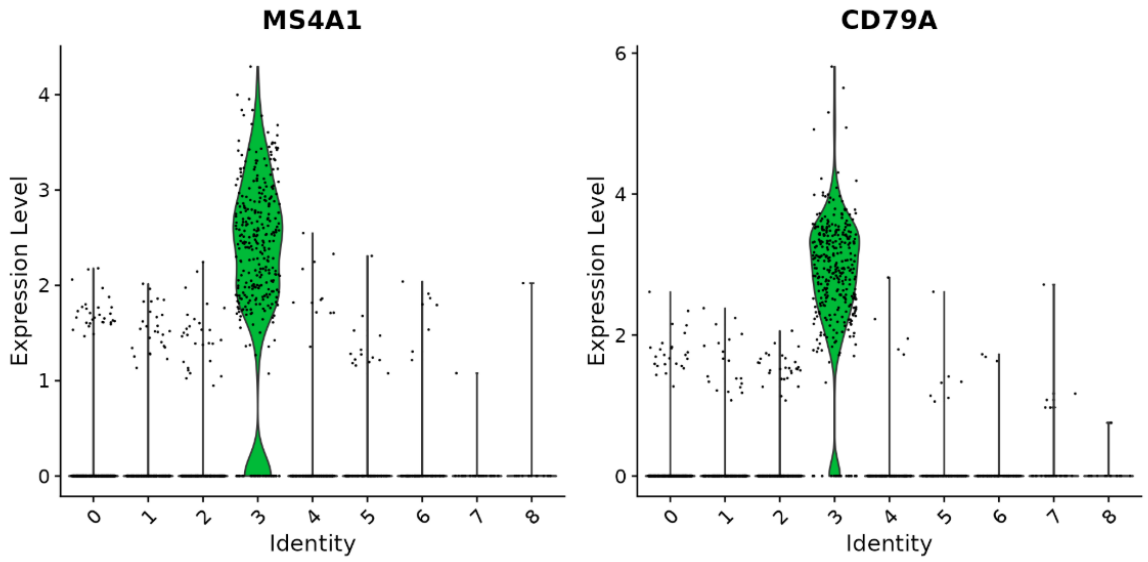


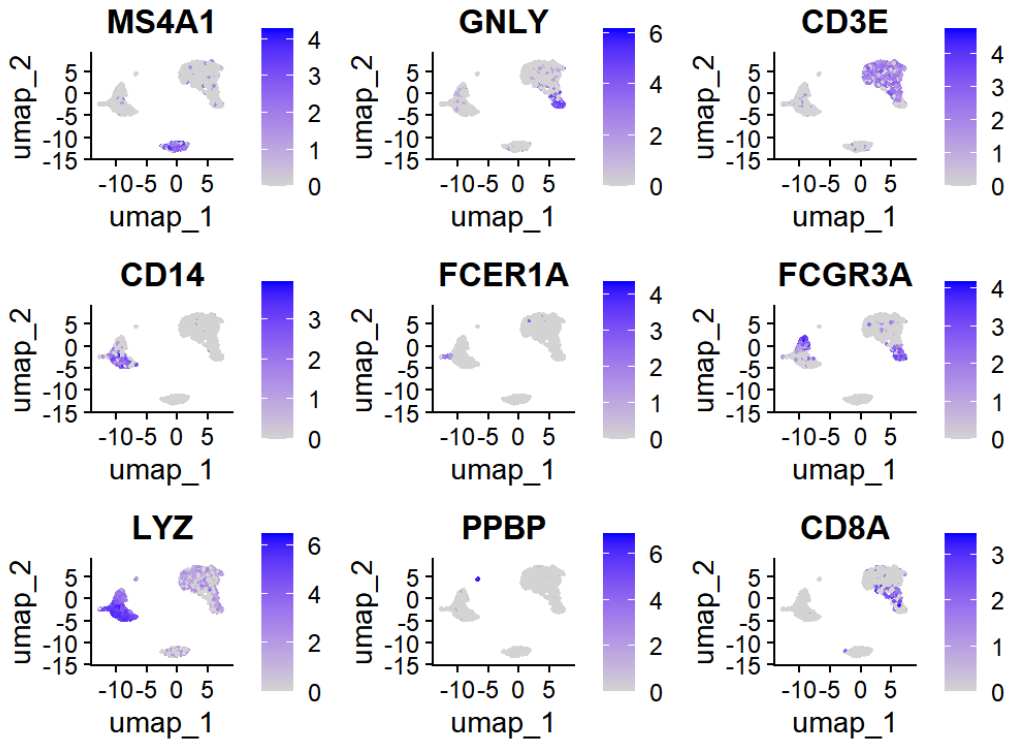
单细胞转录组数据分析流程

基因表达可视化：

#找基因
cluster2.markers <- FindMarkers(pbmc, ident.1 = 2)
head(cluster2.markers, n = 5)
#可视化基因表达
VlnPlot(pbmc, features = c("MS4A1", "CD79A"))
FeaturePlot(pbmc, features = c("MS4A1", "GNLY", "CD3E", "CD14", "FCER1A", "FCGR3A", "LYZ", "PPBP","CD8A"))

```
VlnPlot(pbmc, features = c("MS4A1", "CD79A"))
```





单细胞转录组数据分析流程

定群：

```
new.cluster.ids <- c("Naive CD4 T", "CD14+ Mono", "Memory CD4 T", "B", "CD8 T", "FCGR3A+
Mono","NK", "DC", "Platelet")
names(new.cluster.ids) <- levels(pbmc)
pbmc <- RenameIdents(pbmc, new.cluster.ids)
DimPlot(pbmc, reduction = "umap", label = TRUE, pt.size = 0.5) + NoLegend()
```