

10X genomics 单细胞转录组分析上机文档

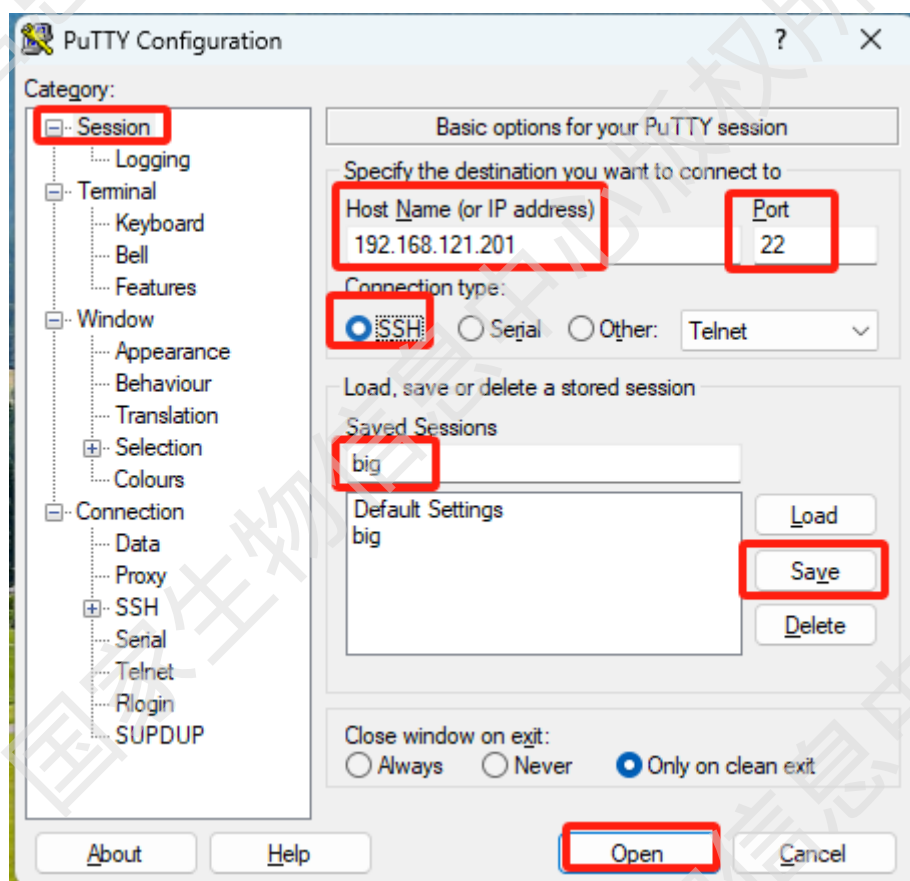
一、通过 putty 登录服务器

1. SSH 客户端 putty 安装及登录

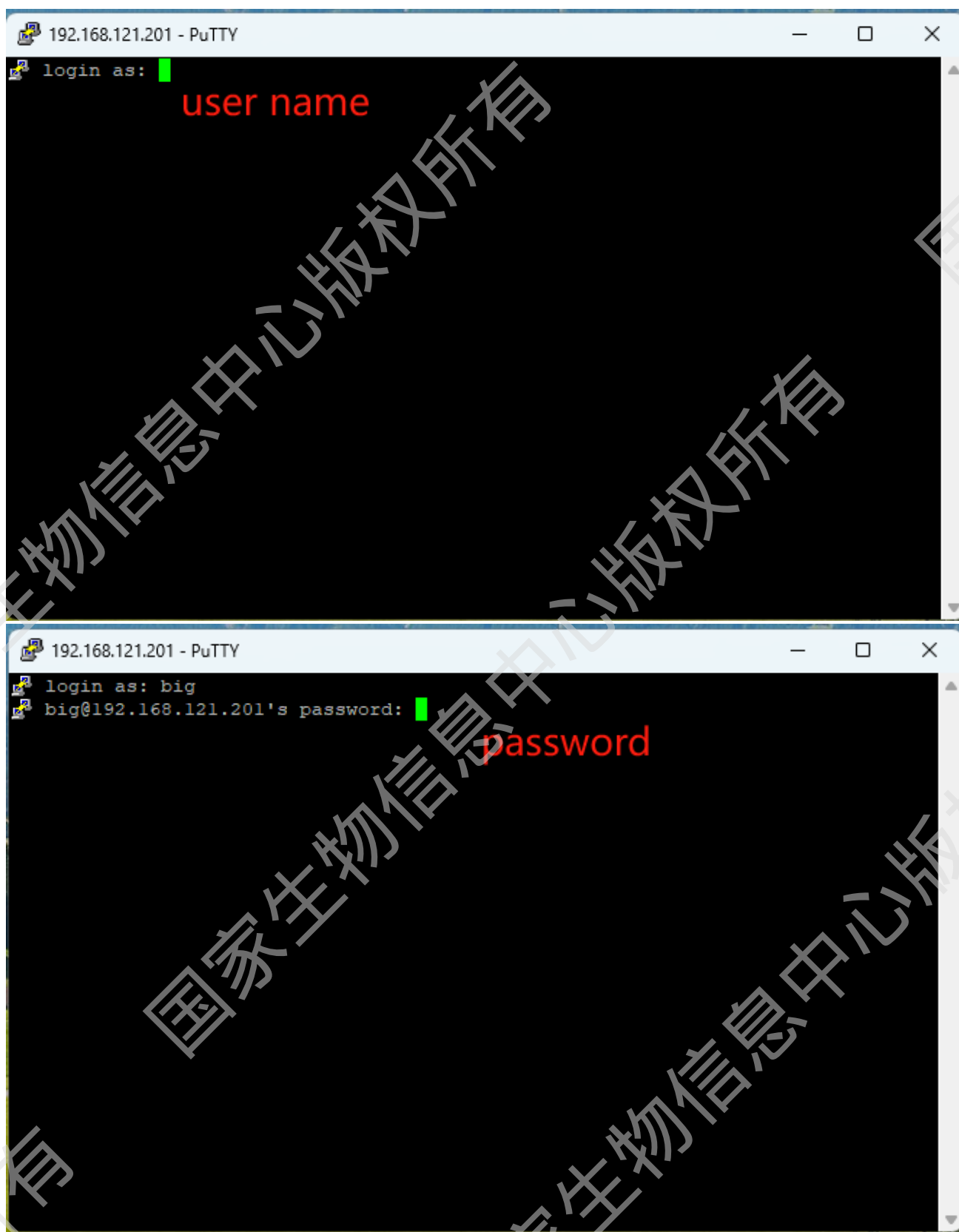
SSH 客户端 putty:

点击链接 <http://the.earth.li/~sgtatham/putty/latest/x86/putty.exe> 下载 putty.exe;

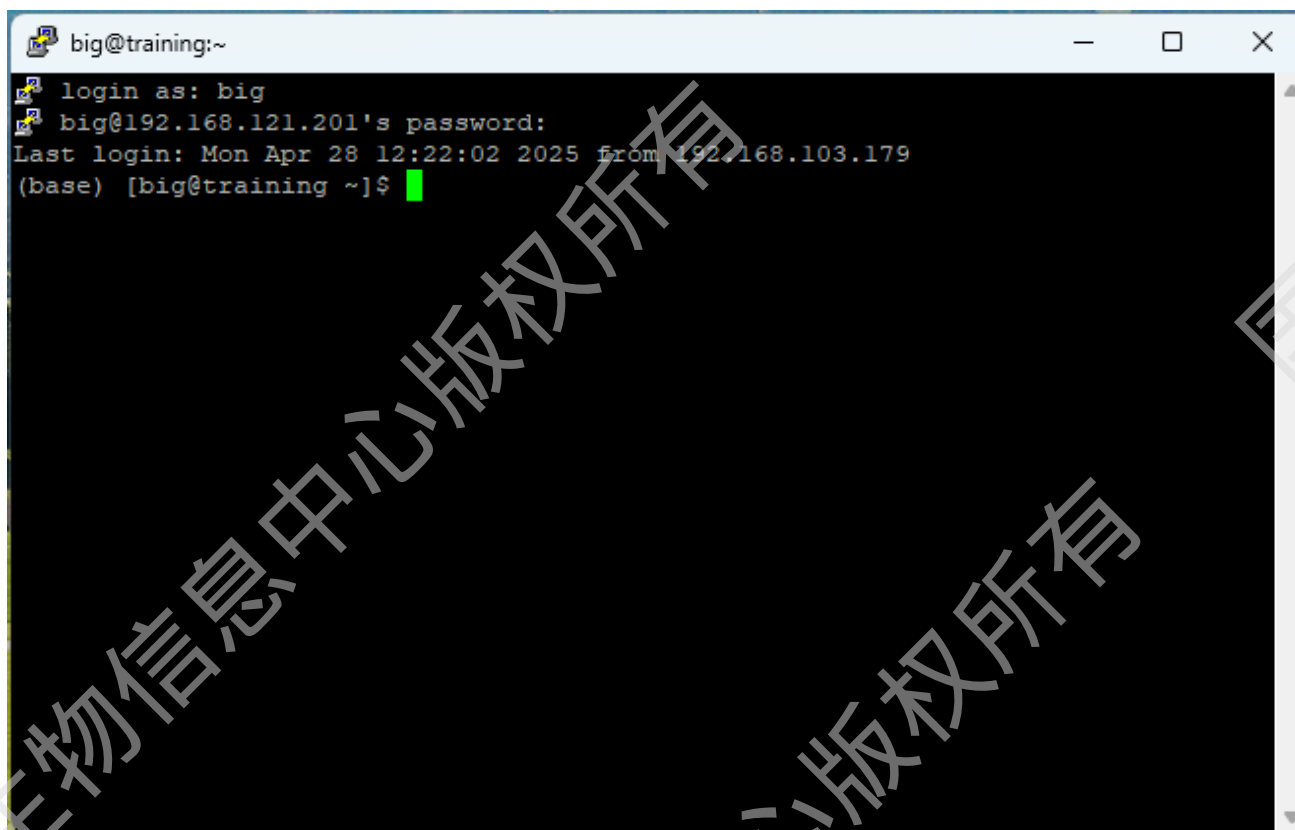
双击打开 putty，在 Session 栏的主机地址输入服务器地址 192.168.121.201；输入端口号：22；选择 ssh 链接方式；点击 Load；然后点击 open 打开。如果出现安全警告框，点击 YES。（见下图）



再输入用户名和密码登陆服务器。



远程登录服务器成功:



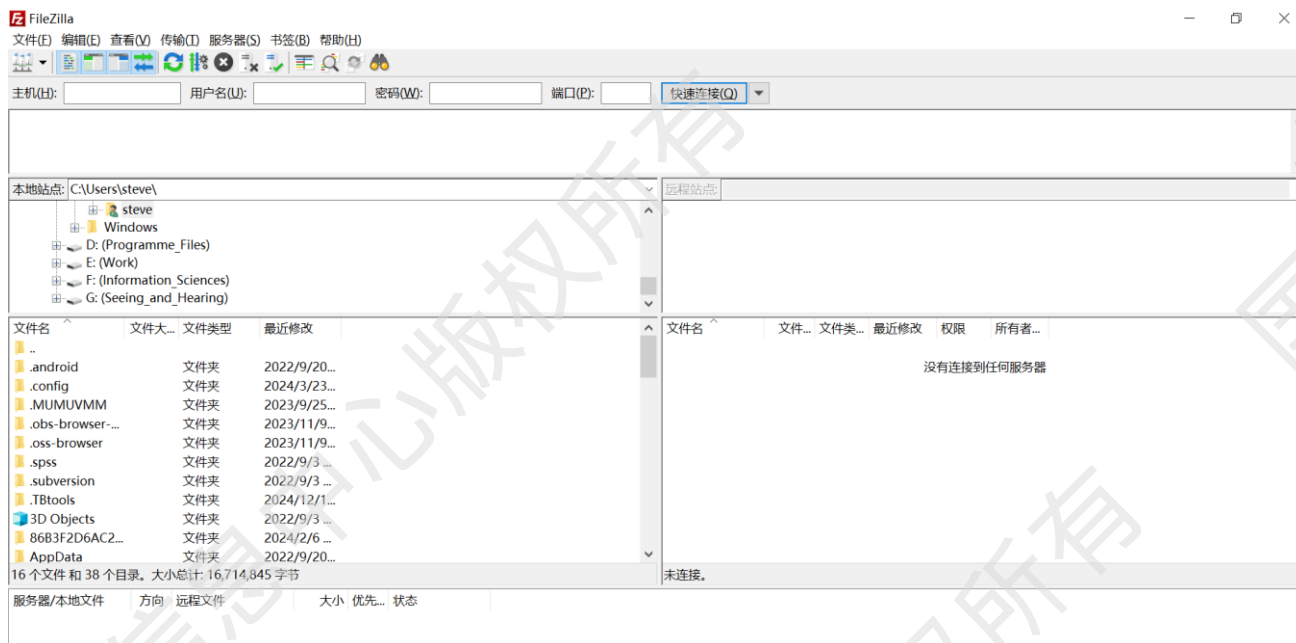
2. FileZilla 上传下载软件使用

FileZilla 是一个常用的上传下载软件，它可以远程登陆服务器，通过拖拽方式从服务器上下载资源到 pc 机，在有写入权限的情况下，也可以从本地上传文件到远程服务器。

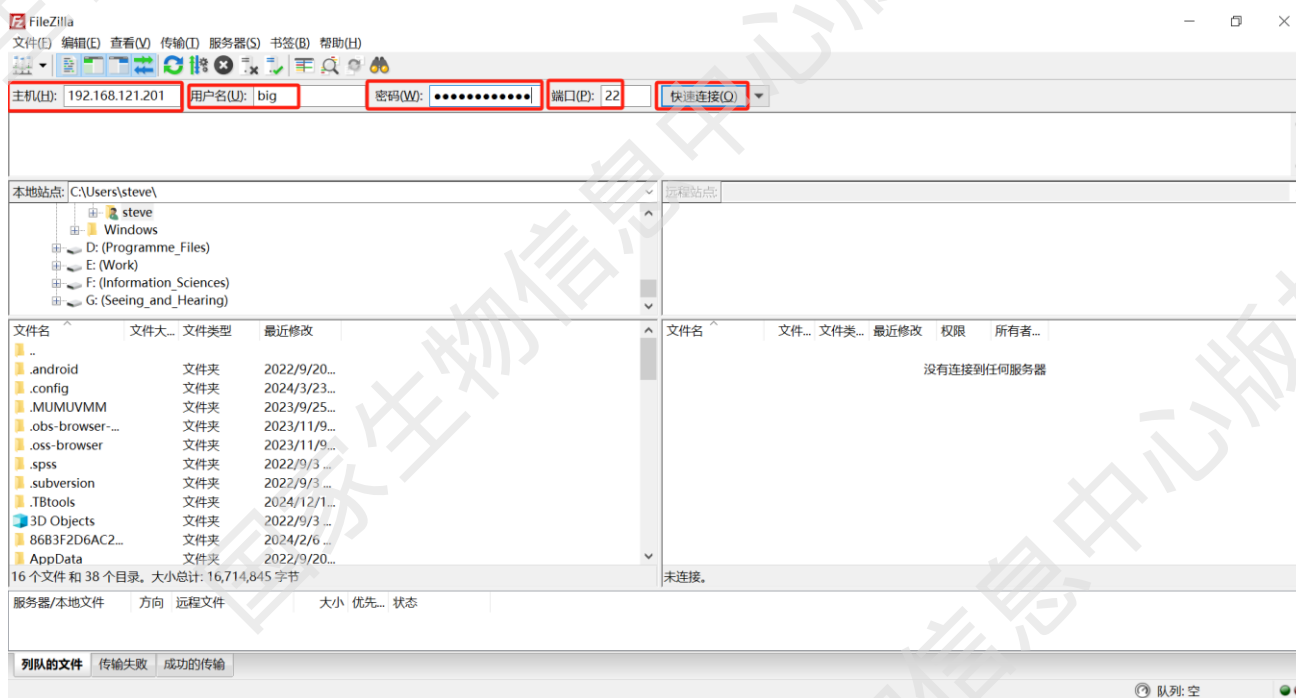
FileZilla 下载链接为 <https://www.filezilla.cn/download/client>，下载并安装。

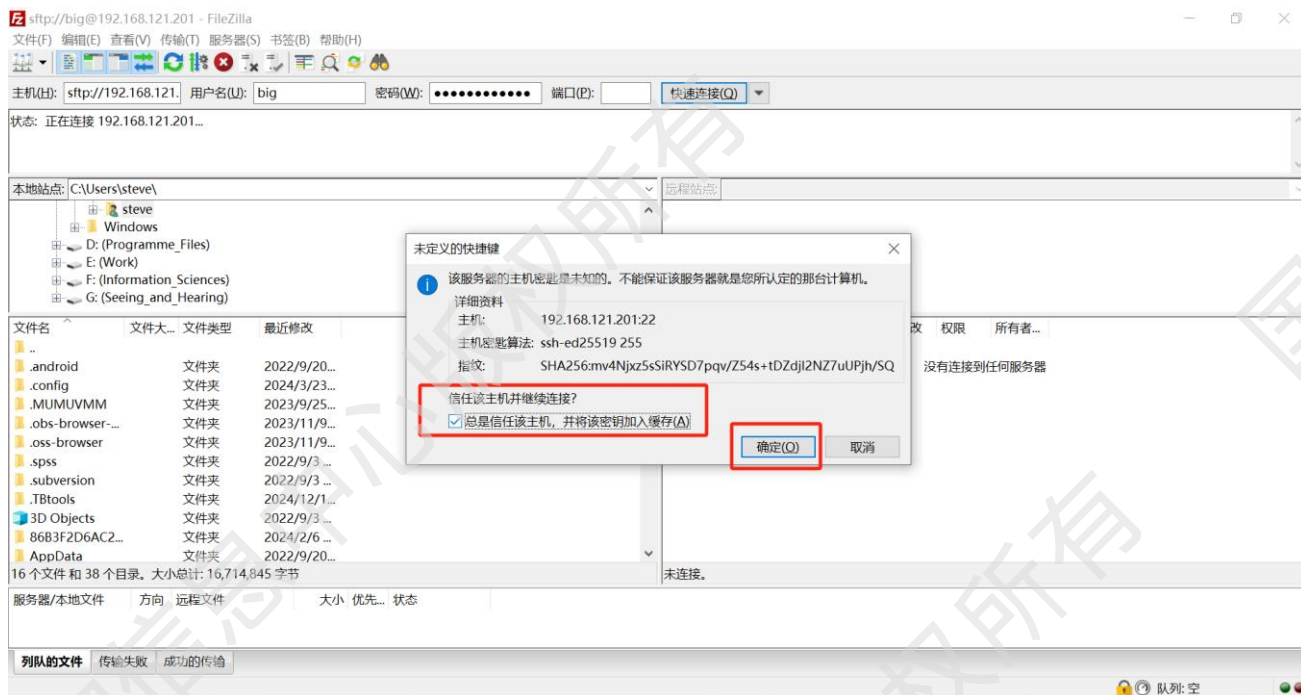
打开 FileZilla，进行如下操作链接我们的服务器：

- 1、打开 FileZilla，如下图：

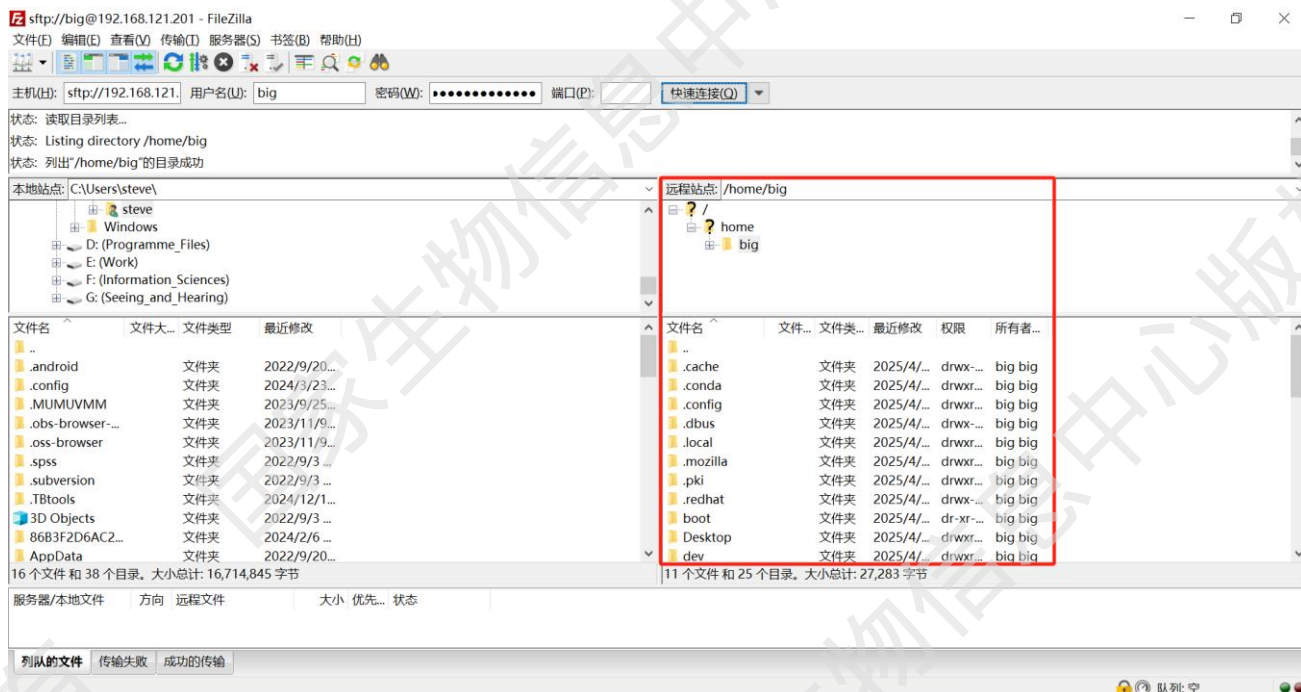


2、点击新建站点，依次填入 “主机 ID: 192.168.121.201”，“端口: 22”，“用户名”，“密码”，点击 “快速连接”，即可以建立本地 pc 机与远程服务器的连接。





3、在左侧选择本地站点的路径(即 pc 文件要存放的路径), 用于存放将要下载的数据。在右侧输入服务器路径, 比如 “/home/user1”, 注意 user1 换成自己的 ID。就可以看到 WGS 文件夹, 双击文件夹进入, 选择自己需要下载的文件左右拖拽就可以开始上传下载。



3. R studio Server

课程全程操作均可以在 R studio server 上完成，Mac 和 Windows 操作系统均可在浏览器访问：

192.168.121.201:8787

之后请输入您服务器登录时的用户名和密码

例如：user01，password

二、 软件和数据

1. 分析用到的软件或工具包

单细胞转录组数据的基因组比对用到的软件是：

cellranger[官网：<https://www.10xgenomics.com/support/software/cell-ranger/latest>]

Linux 服务器上已经安装 cellranger 在路径 “/data/biosoft/big/cellranger-9.0.1/cellranger”。

FastQC 官网：<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Linux 服务器上已经安装 FastQC，整个文件夹（FastQC-0.12.1）也可以拖到 windows 本地，鼠标双击 run_fastqc.bat 打开使用，输出结果没有区别：/data/biosoft/big/FastQC-0.12.1/fastqc

单细胞转录组数据的基因表达分析用到的软件是：

Seurat [官网：<https://satijalab.org/seurat/install.html>]

常用的学习新软件包的网址：

新 R 包：Bioconductor[官网：<https://www.bioconductor.org/>]

新 R 包/新软件：GitHub[官网：<https://github.com/>]

国内平替版（可能不全）Gitee[官网：<https://gitee.com/>]

分析软件均为开源软件，可以下载最新版本，进行安装。

2. 参考基因组数据准备

使用的参考基因组路径为 “/data/ref/refdata-gex-GRCh38-2020-A”，大家可以直接使用。

Hg38 参考基因组及基因注释文件：<https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-GRCh38-2020-A.tar.gz>

3. 准备分析数据及代码

在自己的工作目录下建立 **scRNA** 文件夹并进入，然后把 “/data/biosoft/big/scRNA” 路径下的文件拷贝到当前目录。同时也可以拖到个人电脑本地。

命令行：

```
cd /home/user01
```

```
cp -r /data/biosoft/big/scRNA ./
```

```
cd scRNA/
```

```
ls -la
```

```
(base) [big@training scRNA]$ ls -la
total 24264
drwxrwxr-x. 3 big big 4096 Jun 26 22:26 .
drwxrwxr-x. 10 big big 4096 Jun 26 15:03 ..
-rwxrwxr-x. 1 big big 154 Jun 26 15:38 1_fastqc.sh
-rwxrwxr-x. 1 big big 270 Jun 26 15:30 2_cellranger.sh
-rwxrwxr-x. 1 big big 570 Jun 26 16:25 3_scrublet.py
-rwxrwxr-x. 1 big big 8001 Jun 26 22:05 4_Seurat.R
drwxrwxr-x. 2 big big 61 Jun 26 16:24 filtered_gene_bc_matrices
-rw-rw-r--. 1 big big 649797 Jun 26 22:01 genes_exp_in_each_cell_type.txt
-rw-rw-r--. 1 big big 142410 Jun 26 21:47 ggplot2.barplot.celltype_ratio.png
-rw-rw-r--. 1 big big 1065446 Jun 26 22:00 ggplot2.jjVolcano.DEG_in_each_cell_type.png
-rw-rw-r--. 1 big big 2172085 Jun 26 21:59 ggplot2.jjVolcano.cell_type_maker_gene.png
-rw-rw-r--. 1 big big 359527 Jun 26 21:25 ggplot2.violin.gene.png
-rw-rw-r--. 1 big big 19264114 Jun 26 22:07 pbmc3k.rds
-rw-rw-r--. 1 big big 667185 Jun 26 22:07 pbmc3k_umap.jpg
-rw-rw-r--. 1 big big 116334 Jun 26 16:11 scrublet_result.txt
-rwxrwxr-x. 1 big big 199779 Jun 26 15:14 test_S1_L001_R1_001.fastq.gz
-rwxrwxr-x. 1 big big 161526 Jun 26 15:14 test_S1_L001_R2_001.fastq.gz
```

/data/biosoft/big/scRNA 目录中有一对 fastq 压缩文件，是 10X genomics 平台双端测序产生的原始数据。

```
-rw-rw-r--. 1 big big 199779 May 1 20:36 test_R1_001.fastq.gz
-rw-rw-r--. 1 big big 161526 May 1 20:36 test_R2_001.fastq.gz
```



三、 分析数据

1. 对数据进行质控：FastQC

注：Linux 服务器上已经安装 FastQC，整个文件夹（FastQC-0.12.1）也可以拖到 windows 本地，鼠标双击 run_fastqc.bat 打开使用，输出结果没有区别

查看 1_fastqc.sh 中的代码：cat 1_fastqc.sh

```
(base) [big@training scRNA]$ cat 1_fastqc.sh
#!/bin/bash

mkdir fastqc_reports
/data/biosoft/big/FastQC-0.12.1/fastqc -t 8 -o fastqc_reports test_S1_L001_R1_001.fastq.gz test_S1_L001_R2_001.fastq.gz
```

运行脚本：sh 1_fastqc.sh

将整个 fastqc_reports 目录拖到本地，双击打开 html 文件查看质控结果

2. 生成细胞-基因表达矩阵

Cell Ranger 将前面产生的 fastq 测序数据比对到参考基因组上，然后进行基因表达定量，生成细胞-基因 表达矩阵

查看 2_cellranger.sh 中的代码：cat 2_cellranger.sh

```
(base) [big@training scRNA]$ cat 2_cellranger.sh
#!/bin/bash

/data/biosoft/big/cellranger-9.0.1/cellranger count \
  --id test \
  --transcriptome /data/ref/refdata-gex-GRCh38-2020-A \
  --fastqs /data/biosoft/big/scRNA \
  --create-bam true \
  --localmem 30 \
  --localcores 10
```

说明：一般情况下，--chemistry 无须明文指定，cellranger 能自动检测数据来源于哪种建库方式，并将 chemistry 设置成对应类型。

```
--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' or 'SC3Pv4' for Single Cell 3' v1/v2/v3/v4,
'SC3Pv3LT' for Single Cell 3' v3 LT, 'SC3Pv3HT' for Single Cell 3' v3 HT, 'SC5P-PE' or
'SC5P-PE-v3' or 'SC5P-R2' or 'SC5P-R2-v3' 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' [default: auto]
```

运行: sh ./2_cellranger.sh
运行完成之后, cellranger 会明确说明 “successfully”, 如果没有, 那就是出 bug 了, 需要检查问题后重跑。

```
Outputs:
- Run summary HTML: /data/biosoft/big/scRNA/test/outs/web_summary.html
- Run summary CSV: /data/biosoft/big/scRNA/test/outs/metrics_summary.csv
- BAM: /data/biosoft/big/scRNA/test/outs/possorted_genome_bam.bam
- BAM BAI index: /data/biosoft/big/scRNA/test/outs/possorted_genome_bam.bam.bai
- BAM CSI index: null
- Filtered feature-barcode matrices MEX: /data/biosoft/big/scRNA/test/outs/filtered_feature_bc_matrix
- Filtered feature-barcode matrices HDF5: /data/biosoft/big/scRNA/test/outs/filtered_feature_bc_matrix.h5
- Unfiltered feature-barcode matrices MEX: /data/biosoft/big/scRNA/test/outs/raw_feature_bc_matrix
- Unfiltered feature-barcode matrices HDF5: /data/biosoft/big/scRNA/test/outs/raw_feature_bc_matrix.h5
- Secondary analysis output CSV: /data/biosoft/big/scRNA/test/outs/analysis
- Per-molecule read information: /data/biosoft/big/scRNA/test/outs/molecule_info.h5
- CRISPR-specific analysis: null
- Antibody aggregate barcodes: null
- Loupe Browser file: /data/biosoft/big/scRNA/test/outs/cLoupe.cLoupe
- Feature Reference: null
- Target Panel File: null
- Probe Set File: null
- cell_types: null
- Symlink web_summary_cell_types.html: null

Waiting 6 seconds for UI to do final refresh.
Pipestance completed successfully!

2025-06-26 23:02:27 Shutting down.
Saving pipestance info to "test/test.mri.tgz"
```

主体结果都在./test/outs 里, 查看数据质量及基因组比对情况 ./test/outs/web_summary.html

10x GENOMICS

Cell Ranger • count

test

Summary

Gene Expression

1,263

Estimated Number of Cells

8

Mean Reads per Cell

3

Median Genes per Cell

Run Summary

Sample ID	test
Sample Description	
Chemistry	Single Cell 5' PE
Include introns	True
Transcriptome	GRCh38-2020-A
Pipeline Version	cellranger-9.0.1

Sequencing

Cells

Barcode Rank Plot

Estimated Number of Cells	1,263
Fraction Reads in Cells	87.7%
Mean Reads per Cell	8
Median UMI Counts per Cell	3
Median Genes per Cell	3
Total Genes Detected	2,712

Sequencing [?]

Number of Reads	10,000
Number of Short Reads Skipped	0
Valid Barcodes	87.2%
Valid UMIs	100.0%
Sequencing Saturation	2.1%
Q30 Bases in Barcode	93.8%
Q30 Bases in RNA Read	92.9%
Q30 Bases in RNA Read 2	88.9%
Q30 Bases in UMI	93.5%

Total Genes Detected

2,712

Mapping [?]

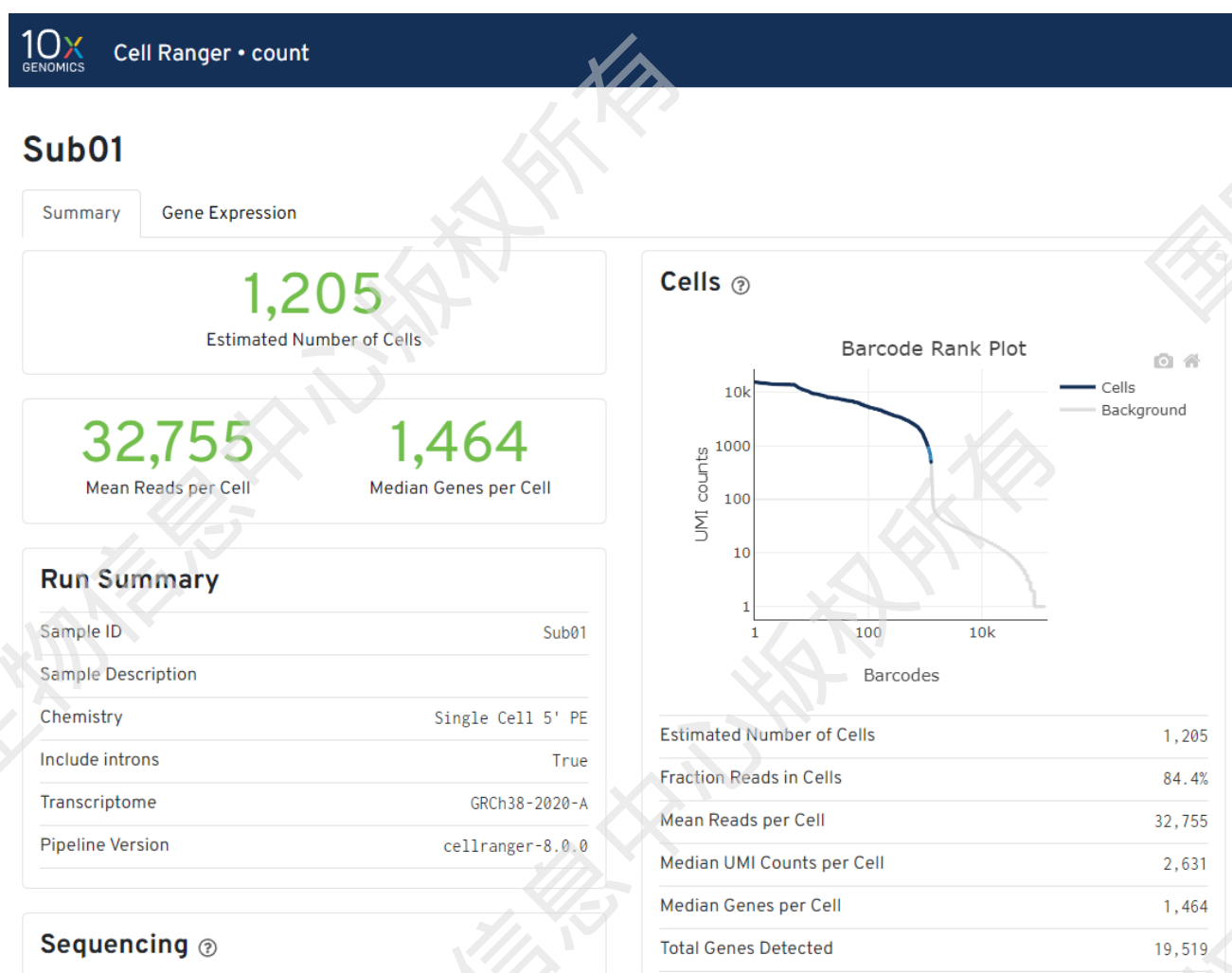
Reads Mapped to Genome	89.4%
Reads Mapped Confidently to Genome	82.0%
Reads Mapped Confidently to Intergenic Regions	4.1%
Reads Mapped Confidently to Intronic Regions	48.2%
Reads Mapped Confidently to Exonic Regions	29.6%
Reads Mapped Confidently to Transcriptome	73.2%
Reads Mapped Antisense to Gene	3.5%

Command Line Arguments

Hide

```
/data/biosoft/big/cellranger-9.0.1/cellranger count --id test --transcriptome /data/ref/refdata-gex-GRCh38-2020-A --fastqs /data/biosoft/big/scRNA --create-bam true --localmem 30 --localcores 10
```

以上是测试数据集的 web summary， 以下是一个完整数据集的 web summary:



Sequencing [?]

Number of Reads	39,470,341
Number of Short Reads Skipped	0
Valid Barcodes	89.8%
Valid UMIs	99.9%
Sequencing Saturation	82.6%
Q30 Bases in Barcode	96.8%
Q30 Bases in RNA Read	95.6%
Q30 Bases in RNA Read 2	91.6%
Q30 Bases in UMI	96.5%

Total Genes Detected

19,519

Mapping [?]

Reads Mapped to Genome	90.0%
Reads Mapped Confidently to Genome	82.3%
Reads Mapped Confidently to Intergenic Regions	4.0%
Reads Mapped Confidently to Intronic Regions	48.6%
Reads Mapped Confidently to Exonic Regions	29.7%
Reads Mapped Confidently to Transcriptome	73.8%
Reads Mapped Antisense to Gene	3.4%

Command Line Arguments

Hide

```
/xtdisk/jiangl_group/yuchw/software/cellranger-8.0.0/bin/cellranger count --id=Sub01 --transcriptome=/xtdisk/
```

以下是整个 outs 目录结构和内容说明:

```
# 输出结果
#   .outs
#   ├── analysis 【数据分析文件夹】
#   │   ├── clustering 【聚类，图聚类 and k-means 聚类】
#   │   ├── diffexp 【差异分析】
#   │   ├── pca 【主成分分析线性降维】
#   │   └── tsne 【非线性降维信息】
#   ├── cloupe.cloupe 【Loupe Cell Browser 输入文件】
#   ├── filtered_feature_bc_matrix 【过滤掉的barcode信息】
#   │   ├── barcodes.tsv.gz
#   │   ├── features.tsv.gz
#   │   └── matrix.mtx.gz
#   ├── filtered_feature_bc_matrix.h5 【过滤掉的barcode信息HDF5 format】
#   ├── metrics_summary.csv 【CSV format数据摘要】
#   ├── molecule_info.h5 【UMI信息，aggregate的时候会用到的文件】
#   ├── raw_feature_bc_matrix 【原始barcode信息】
#   │   ├── barcodes.tsv.gz
#   │   ├── features.tsv.gz
#   │   └── matrix.mtx.gz
#   ├── possorted_genome_bam.bam 【比对文件】
#   ├── possorted_genome_bam.bam.bai 【索引文件】
#   ├── raw_feature_bc_matrix.h5 【原始barcode信息HDF5 format】
#   ├── web_summary.html 【网页简版报告以及可视化】
#   └── *_gene_bar.csv_temp 【过程文件】
```

查看输出文件: `./test/outs/filtered_feature_bc_matrix`

```
(base) [big@training scRNA]$ ll ./test/outs/filtered_feature_bc_matrix
total 344
-rw-rw-r--. 1 big big 7793 Jun 26 23:01 barcodes.tsv.gz
-rw-rw-r--. 1 big big 333437 Jun 26 23:01 features.tsv.gz
-rw-rw-r--. 1 big big 7845 Jun 26 23:01 matrix.mtx.gz
```

```
(base) [big@training test]$ cd outs/filtered_feature_bc_matrix/
(base) [big@training filtered_feature_bc_matrix]$ ls
barcodes.tsv.gz features.tsv.gz matrix.mtx.gz
(base) [big@training filtered_feature_bc_matrix]$ zcat barcodes.tsv.gz | head
AAACGAACAAGCTACT-1
AAACGCTAGGAGATAG-1
AAAGAACAGAATCGAT-1
AAAGGATCAATGGGTG-1
AAAGGATCAGATCCTA-1
AAAGGATTCCTAGCGG-1
AAAGGGCGTTTGAAAG-1
AAAGGTAGTCTTACTT-1
AAAGTCCCAACAGTGG-1
AAAGTCCCACAACCGC-1
(base) [big@training filtered_feature_bc_matrix]$ zcat features.tsv.gz | head
ENSG00000243485 MIR1302-2HG Gene Expression
ENSG00000237613 FAM138A Gene Expression
ENSG00000186092 OR4F5 Gene Expression
ENSG00000238009 AL627309.1 Gene Expression
ENSG00000239945 AL627309.3 Gene Expression
ENSG00000239906 AL627309.2 Gene Expression
ENSG00000241860 AL627309.5 Gene Expression
ENSG00000241599 AL627309.4 Gene Expression
ENSG00000286448 AP006222.2 Gene Expression
ENSG00000236601 AL732372.1 Gene Expression
(base) [big@training filtered_feature_bc_matrix]$ zcat matrix.mtx.gz | head
%%MatrixMarket matrix coordinate integer general
%%metadata_json: {"software_version": "cellranger-9.0.1", "format_version": 2}
36601 1359 1487
34658 1 2
31576 2 1
1540 3 1
36569 4 1
35379 5 1
5887 6 1
25618 7 1
```

3. 推断 doublets, 从该步骤开始, 演示数据使用 Seurat 软件包提供的 PBMC

3K 数据: /data/biosoft/big/scRNA/filtered_gene_bc_matrices

查看 3_scrublet.py 中的代码: cat 3_scrublet.py

```
(base) [big@training scRNA]$ cat 3_scrublet.py
#!/home/big/miniconda3/envs/scrublet/bin/python
# -*- coding: utf-8 -*-
__metaclass__ = type

import os
import scanpy as sc
import scrublet as scr
import numpy as np
import pandas as pd

a = sc.read_10x_mtx("./filtered_gene_bc_matrices")
sc.pp.filter_genes(a, min_cells=3)
b = scr.Scrublet(a.X)
doublet_scores, predicted_doublets = b.scrub_doublets(n_prin_comps = 3)
c = { "BARCODE" : np.array(a.obs_names),
      "DBL" : np.array(predicted_doublets),
      "DBL_SCORE" : np.array(doublet_scores) }
d = pd.DataFrame(c)
d.to_csv("scrublet result.txt", sep="\t", index=0)
```

查看输出结果

```
(base) [big@training scRNA]$ head scrublet_result.txt
BARCODE DBL      DBL_SCORE
AAACATACAACCAC-1      False    0.11475409836065571
AAACATTGAGCTAC-1      False    0.09833024118738404
AAACATTGATCAGC-1      False    0.06990881458966564
AAACCGTGCTTCCG-1      False    0.03062302006335797
AAACCGTGTATGCG-1      False    0.01066447908121411
AAACGCACTGGTAC-1      False    0.09352517985611511
AAACGCTGACCAGT-1      False    0.14285714285714285
AAACGCTGGTTCTT-1      False    0.07692307692307691
AAACGCTGTAGCCA-1      False    0.1707317073170731
```

使用 linux 命令组合统计 DBL 这一列为 True 和 False 的细胞个数，True 即为 doublet:

```
cat scrublet_result.txt | tail -n+2 | cut -f2 | sort | uniq -c
```

代码解释：打印 scrublet_result.txt 中的全部内容到屏幕 | 从第二行开始打印 | 只打印第二列 | 排序 | 去重复的同时输出重复个数

```
(base) [big@training scRNA]$ cat scrublet_result.txt | tail -n+2 | cut -f2 | sort | uniq -c
2649 False
51 True
```

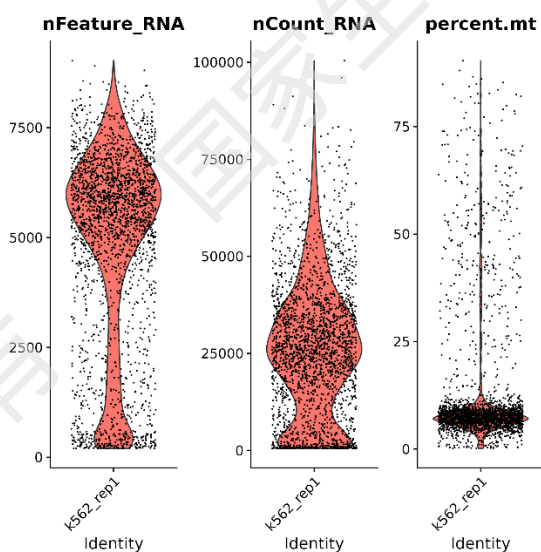
4. Seurat 读取、分析、可视化单细胞转录组数据，详细教程参考官方网站

安装：https://satijalab.org/seurat/articles/install_v5

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

打开 Rstudio，打开 4_Seurat.R，逐行执行代码：

1) scRNA-seq 数据质控



根据基因表达数、转录本的表达数以及线粒体基因的表达比例，过滤低质量的细胞 (不同组织/细胞类型)

不一样，阈值要根据情况调整/参考文献)

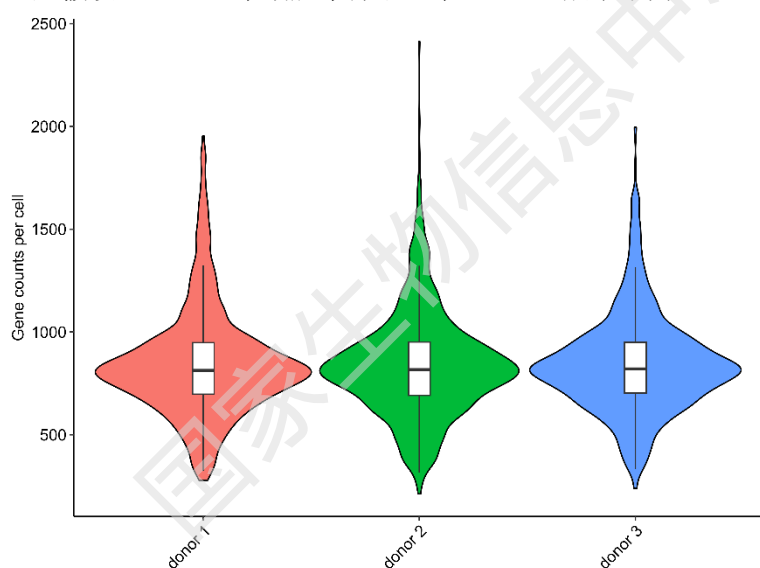
分开考虑这三个 QC 变量中的任何一个都可能导致对细胞状态的错误解读。例如，线粒体计数相对较高的细胞，可能死亡细胞比较多。同样，其他 QC 变量也具有相应的生物学意义。总分子数和/或基因数低的细胞可能是处于静息状态的细胞群体；总分子数和/或基因数高的细胞也可能是细胞自身体积较大。实际上，细胞之间的总分子数自身也可能有很大差异（具体见 Github 上的案例研究）。因此，在做出是否过滤的单阈值决策时，应联合考虑 3 个 QC 变量，并且应将这些阈值设置的尽可能宽松，以避免无意间滤除有效的细胞群。

5. 单细胞表达数据的标准化、鉴定 highly variable gene、ScaleData 和 PCA 降维、批次校正、无监督聚类、UMAP 降维、marker 基因鉴定、细胞类型注释

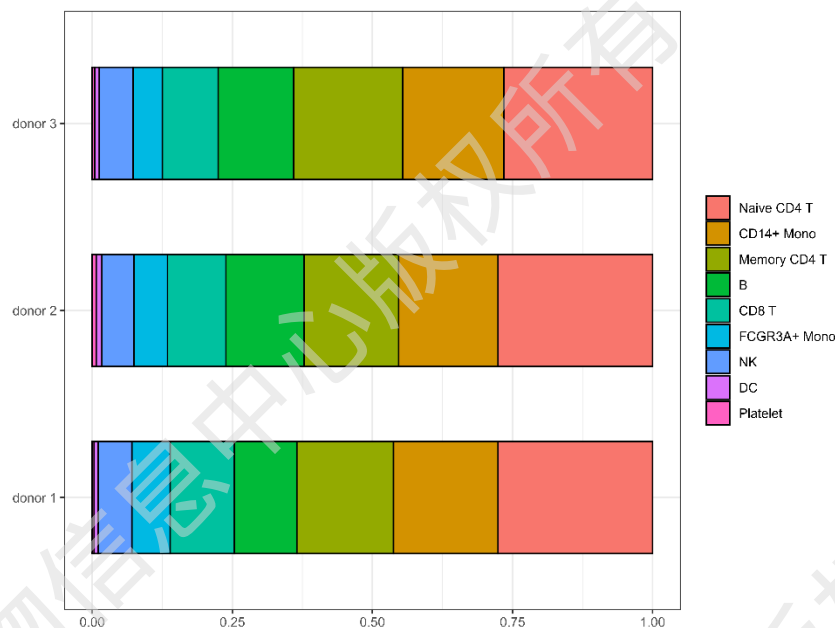
详见 4_Seurat.R 中对每步代码的注释

6. 单细胞转录组数据可视化：violin plot、barplot

此处假设 PBMC 3K 个细胞来源于 3 个 donor，展示不同 donor 各自的 gene counts per cell distribution

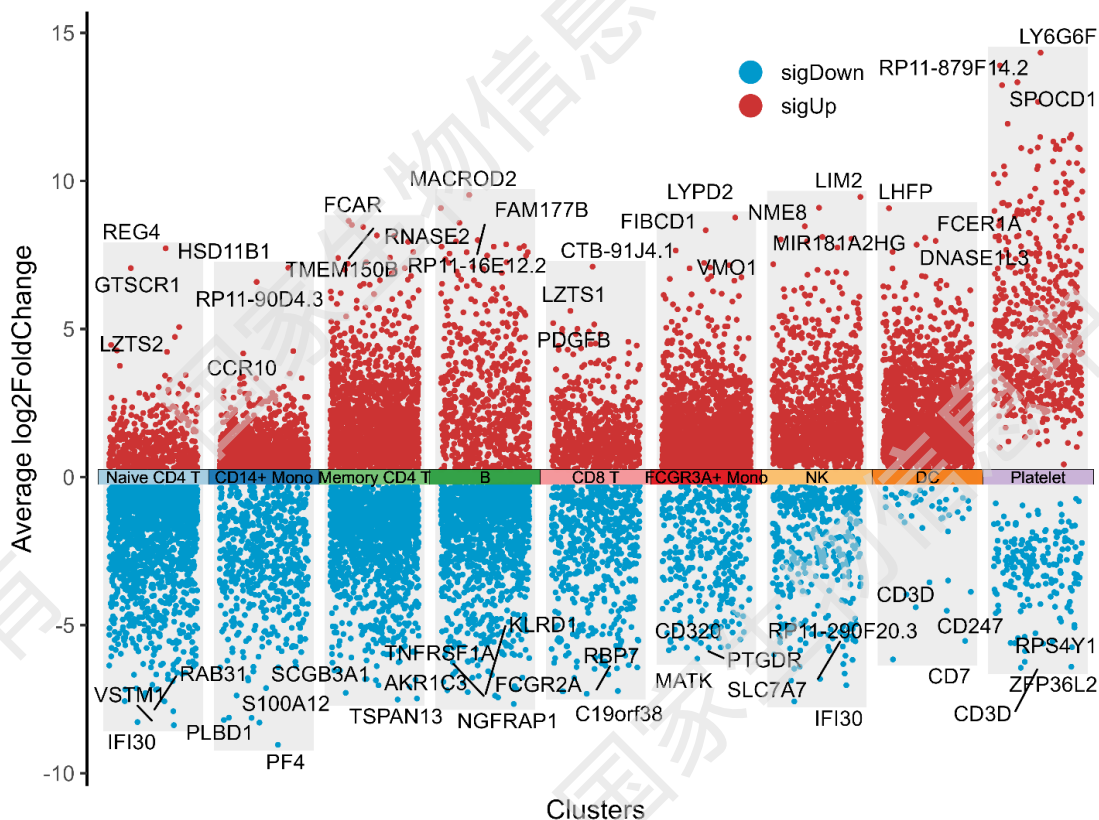


常用图：对比不同 donor（或不同 group）的 cell type ratio 变化情况



7. 按细胞类型鉴定 case vs control 组间差异表达基因

使用 FindAllMarkers 得到每个 cell type 的 marker gene，并使用 jjVolcano 进行可视化



灵活运用 FindMarkers 函数，得到每个 cell type 中的 case vs control 差异表达基因，并使用 jjVolcano 进行可视化

