单细胞免疫组数据分析上机文档

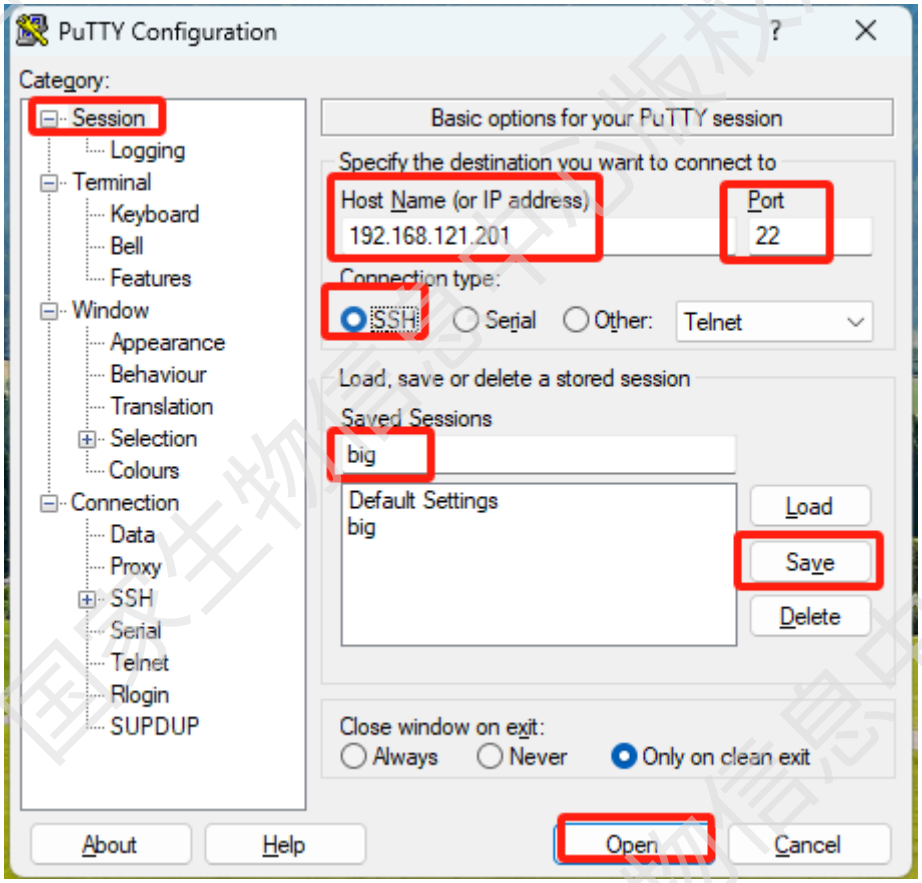
一、 通过 putty 登录服务器

1. SSH 客户端 putty 安装及登录

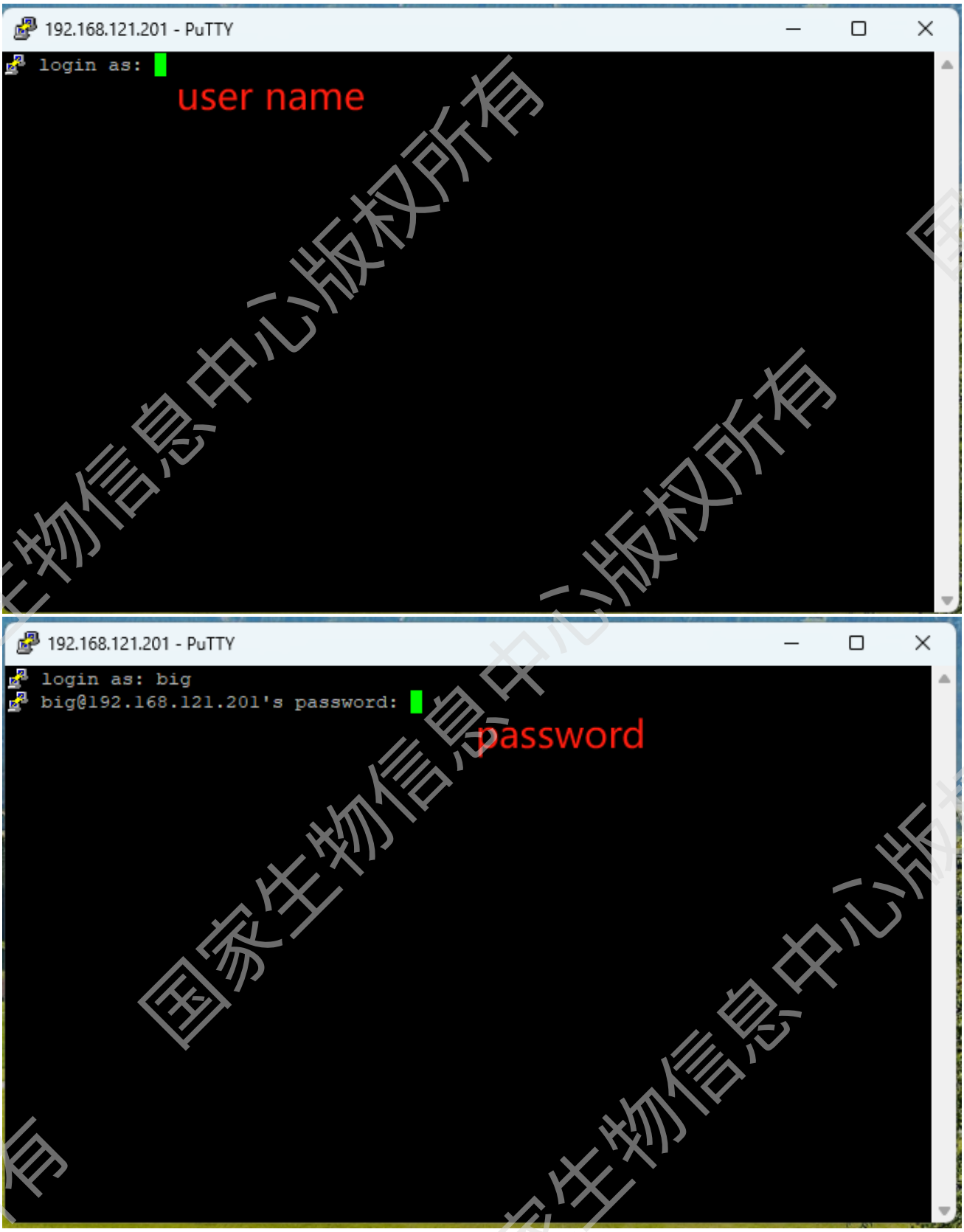
SSH 客户端 putty：

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

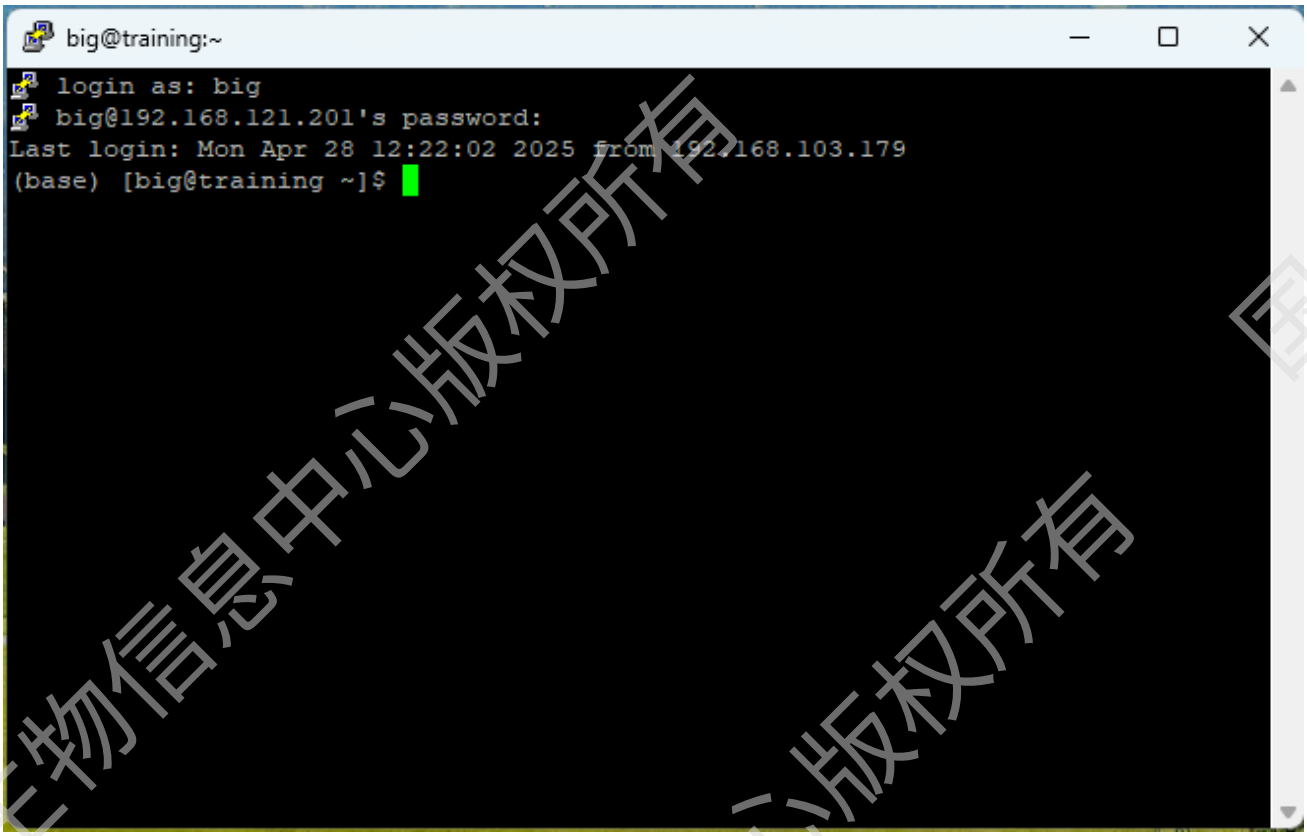
双击打开 putty，在 Session 栏的主机地址输入服务器地址 192.168.121.201；输入端口号： 22；选择 ssh 链接方式；点击 Load，然后点击 open 打开。如果出现安全警告框，点击 YES。（见下图）



再输入用户名和密码登陆服务器。



远程登录服务器成功：



二、 软件和数据

1) 分析用到的软件

免疫组的序列比对部分用到的软件是：

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

免疫组的分析软件

scRepertoire [官网：https://www.borch.dev/uploads/screpertoire/]

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) 软件准备

Linux 系统里使用的软件 cellranger 已经安装在路径“/data/biosoft/big/cellranger-9.0.1/cellranger”，使用的参考基因组路径为“ /data/biosoft/big/scTCR_scBCR/refdata-cellranger-vdj-GRCh38-alts-ensembl-7.1.0/”，大家可以直接使用。

3) 准备分析数据

在自己的工作目录下建立文件夹 scTCR_scBCR 并进入，然后把“/data/biosoft/big/scTCR_scBCR”路径下的数据拷贝到当前目录。

命令行：

cd

cp -r /data/biosoft/big/scTCR_scBCR .

cd scTCR_scBCR/

ll

```
mkdir scTCR_scBCR                  #建立文件夹
cd   scTCR_scBCR                     #进入文件夹
cp /home/big/scTCR_scBCR/* ./          #复制数据
ll   ./                  #查看文件夹内容
```

打开文件夹进行查看，

```
(base) [big@training scTCR_scBCR]$ ll
total 12
-rwxrwxr-x. 1 big big  268 Apr 21 11:53 1_cellranger_Sub01.sh
drwxr-xr-x. 3 big big   41 Nov 22  2022 refdata-cellranger-vdj-GRCh38-alts-ensembl-7.1.0
-rw-rw-r--. 1 big big 3350 Apr 21 11:45 Sub01_S1_L001_R1_001.fastq.gz
-rw-rw-r--. 1 big big 3350 Apr 21 11:45 Sub01_S1_L001_R2_001.fastq.gz
```

文件说明：

refdata-cellranger-vdj-GRCh38-alts-ensembl-7.1.0 目录是用来做比对的参考基因组，包含基因组的 fasta 序列以及 cellranger 需要用到的其他文件。

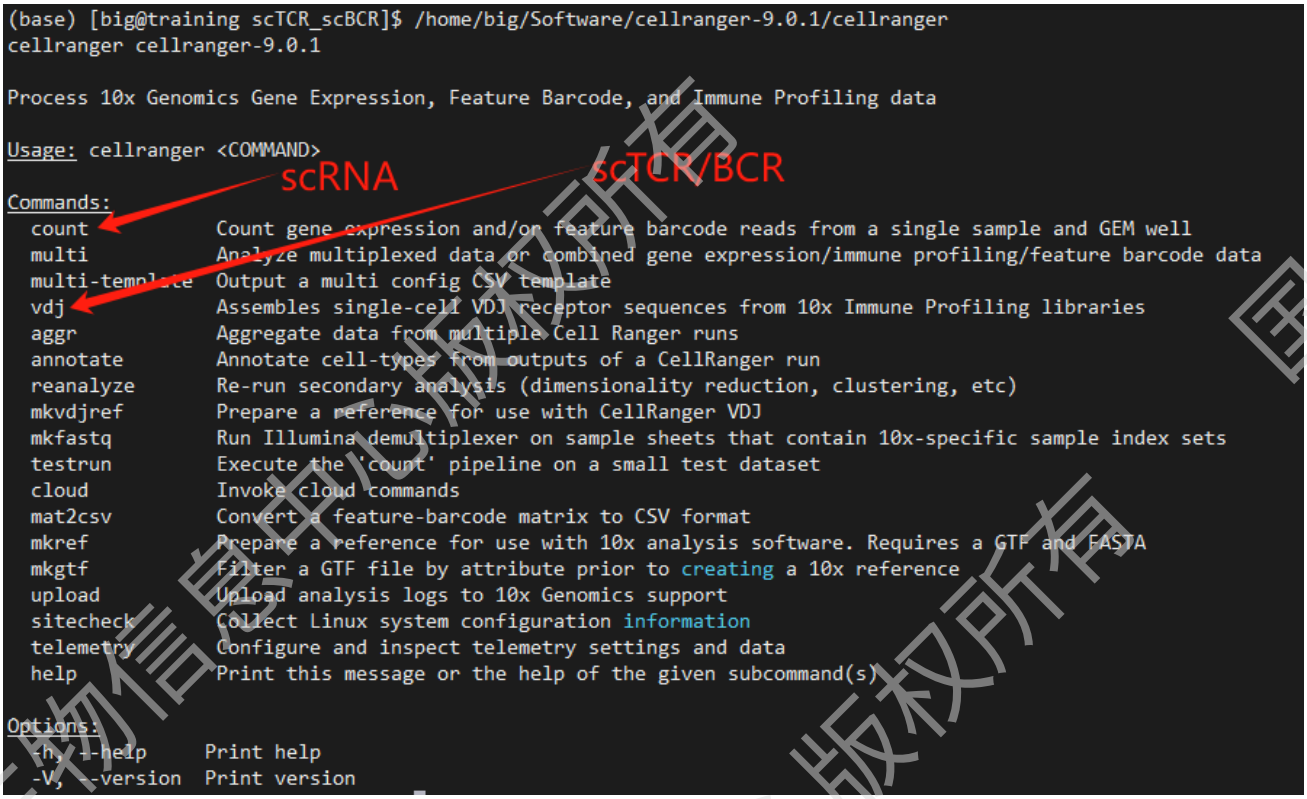
*.fastq.gz 文件为测序数据，这里有一个样本的 scTCR 双端测序数据，Sub01 表示测序样本 ID; L001 表示测序重复，R1 和 R2 表示双末端 pair-end 数据。

三、 分析流程

1. 基因组比对(约 7min)

1.1 认识 cellranger

基因组比对是以参考基因组为模版，把测序的短片段 reads 比对到参考基因组的具体位置上。10X Genomics 的 Universal 5' Gene Expression 建库方式，会检测同一个细胞的 5'scRNA-seq 和 scTCR/BCR 这两个模态。常用的 scRNA-seq 比对软件为 cellranger count 或 STARsolo，常用的 scTCR/BCR 比对软件为 cellranger vdj。



1.2 输入/data/biosoft/big/cellranger-9.0.1/cellranger vdj --help 可以查看软件的帮助信息 (下图)

```
(base) [big@training scTCR_scBCR]$ /home/big/Software/cellranger-9.0.1/cellranger vdj --help
Assembles single-cell VDJ receptor sequences from 10x Immune Profiling libraries

Usage: cellranger vdj [OPTIONS] --id <ID> --fastqs <PATH>

Options:
      --id <ID>                          A unique run id and output folder name [a-zA-Z0-9_-]+
      --description <TEXT>               Sample description to embed in output files [default: ]
      --reference <PATH>                 Path of folder containing 10x-compatible VDJ reference. Optional if '--denovo' is
                                         specified
      --fastqs <PATH>                    Path to input FASTQ data
      --project <TEXT>                   Name of the project folder within a mkfastq or bcl2fastq-generated folder to pick
                                         FASTQs from
      --sample <PREFIX>                  Prefix of the filenames of FASTQs to select
      --lanes <NUMS>                     Only use FASTQs from selected lanes
      --denovo                           Run in reference-free mode (do not use annotations)
      --chain <CHAIN_SPEC>               Chain type to display metrics for: 'TR' for T cell receptors, 'IG' for B cell
                                         receptors, or 'auto' to autodetect [default: auto]
      --inner-enrichment-primers <PATH>  If inner enrichment primers other than those provided in the 10x kits are used,
                                         they need to be specified here as a textfile with one primer per line. Disable
                                         secondary analysis, e.g. clustering
      --dry                              Do not execute the pipeline. Generate a pipeline invocation (.mro) file and stop
      --jobmode <MODE>                   Job manager to use. Valid options: local (default), sge, lsf, slurm or path to a
                                         .template file. Search for help on "Cluster Mode" at support.10xgenomics.com for
                                         more details on configuring the pipeline to use a compute cluster
      --localcores <NUM>                 Set max cores the pipeline may request at one time. Only applies to local jobs
      --localmem <NUM>                   Set max GB the pipeline may request at one time. Only applies to local jobs
      --localvmem <NUM>                  Set max virtual address space in GB for the pipeline. Only applies to local jobs
      --mempercore <NUM>                 Reserve enough threads for each job to ensure enough memory will be available,
                                         assuming each core on your cluster has at least this much memory available. Only
                                         applies to cluster jobmodes
      --maxjobs <NUM>                    Set max jobs submitted to cluster at one time. Only applies to cluster jobmodes
      --jobinterval <NUM>                Set delay between submitting jobs to cluster, in ms. Only applies to cluster
                                         jobmodes
      --overrides <PATH>                 The path to a JSON file that specifies stage-level overrides for cores and memory.
                                         Finer-grained than --localcores, --mempercore and --localmem. Consult
                                         https://10xgen.com/resource-override for an example override file
      --output-dir <PATH>                Output the results to this directory
      --uiport <PORT>                    Serve web UI at http://localhost:PORT
      --disable-ui                       Do not serve the web UI
      --noexit                           Keep web UI running after pipestance completes or fails
      --nopreflight                      Skip preflight checks
  -h, --help                             Print help
```

1.3 比对（mapping）

命令行：

/data/biosoft/big/cellranger-9.0.1/cellranger vdj --id=Sub01 --sample=Sub01 --reference=./refdata-cellranger-vdj-GRCh38-alts-ensembl-7.1.0/ --fastqs=./ --localmem=30 --localcores=10

（不要运行，仅为示例）

```
cd /home/user01/scTCR_scBCR        #回到工作目录，注释 user01 换成自己的 ID
/home/big/Software/cellranger-9.0.1/cellranger vdj --id=Sub01 --sample=Sub01 --reference=./refdata-cellranger-vdj-GRCh38-alts-ensembl-7.1.0 --fastqs=./ --localmem=30 --localcores=10
```

软件参数说明：

--id 用户自定义一个唯一的 ID，cellranger vdj 的全部输出结果会保存在名为该 ID 的目录下

--sample 给 cellranger 读取的 fastq 的前缀名（即“Sub01_S1_L001_R1_001.fastq.gz”中的“Sub01”）

--reference 参考基因组的目录，下载网址和教程：https://www.10xgenomics.com/support/software/cell-ranger/downloads#reference-downloads

--fastqs 原始 fastq 数据所在路径

--localmem 所需内存，以 GB 为单位

--localcores 所需线程数

注：从方便核对结果的角度，id 和 sample 可以是相同的字符串；cellranger vdj 一般需要 30-50GB 内存，线程数给多少都可以，给多了跑得更快，一般给 40GB 内存+10 线程，可以在 1 小时以内跑完 cellranger 的比对。

如果运行了上述代码，当前结果由于 fastq 序列为示例所以显示如下

```
[error] Pipestance failed. Error log at:
Sub01/SC_VDJ_ASSEMBLER_CS/SC_MULTI_CORE/MULTI_CHEMISTRY_DETECTOR/VDJ_CHEMISTRY_DETECTOR/DETECT_VDJ_RECEPTOR/fork0/chnk0-u851e0f0431/_errors

Log message:
V(D)J Chain detection failed for Sample Sub01 in "/home/user50/scTCR_scBCR".

Total Reads          = 10
Reads mapped to TR   = 3
Reads mapped to IG   = 0

In order to distinguish between the TR and the IG chain the following conditions need to be satisfied:
- A minimum of 10000 total reads
- A minimum of 5.0% of the total reads needs to map to TR or IG
- The number of reads mapped to TR should be at least 3.0x compared to the number of reads mapped to IG or vice versa
Please check the input data and/or specify the chain via the --chain argument.


Waiting 6 seconds for UI to do final refresh.
Pipestance failed. Use --noexit option to keep UI running after failure.

2025-04-28 12:29:43 Shutting down.
Saving pipestance info to "Sub01/Sub01.mri.tgz"
For assistance, upload this file to 10x Genomics by running:

cellranger upload <your_email> "Sub01/Sub01.mri.tgz"
```

请返回家目录重新拷贝数据

cd

cp -r /data/biosoft/big/scTCR_scBCR/ .

cd scTCR_scBCR/

ll

1.4 运行完后，在当前目录[scTCR_scBCR 目录]下进行查看，可看到生成以下结果目录：

```
(base) [big@training scTCR_scBCR]$ ll
total 16
-rwxrwxr-x. 1 big big  250 Apr 21 11:57 1_cellranger.Sub01.sh
drwxr-xr-x. 3 big big   41 Nov 22  2022 refdata-cellranger-vdj-GRCh38-alts-ensembl-7.1.0
drwxrwxr-x. 5 big big 4096 Apr 21 11:52 Sub01
-rw-rw-r--. 1 big big 3350 Apr 21 11:45 Sub01_S1_L001_R1_001.fastq.gz
-rw-rw-r--. 1 big big 3350 Apr 21 11:45 Sub01_S1_L001_R2_001.fastq.gz
```

目录里的 outs 文件夹包含供下游分析使用的输出结果

cd Sub01/outs

ll

```
(base) [big@training scTCR_scBCR]$ ll Sub01/outs
total 724256
-rw-rw-r--. 1 big big      2704 Apr 21 11:52 airr_rearrangement.tsv
-rw-rw-r--. 1 big big     61380 Apr 21 11:52 all_contig_annotations.bed
-rw-rw-r--. 1 big big     63902 Apr 21 11:52 all_contig_annotations.csv
-rw-rw-r--. 1 big big   1605431 Apr 21 11:52 all_contig_annotations.json
-rw-rw-r--. 1 big big 732444785 Apr 21 11:52 all_contig.bam
-rw-rw-r--. 1 big big     29296 Apr 21 11:52 all_contig.bam.bai
-rw-rw-r--. 1 big big    175750 Apr 21 11:52 all_contig.fasta
-rw-rw-r--. 1 big big     16969 Apr 21 11:52 all_contig.fasta.fai
-rw-rw-r--. 1 big big    341618 Apr 21 11:52 all_contig.fastq
-rw-rw-r--. 1 big big        28 Apr 21 11:52 cell_barcodes.json
-rw-rw-r--. 1 big big       165 Apr 21 11:52 clonotypes.csv
-rw-rw-r--. 1 big big       735 Apr 21 11:52 concat_ref.bam
-rw-rw-r--. 1 big big        96 Apr 21 11:52 concat_ref.bam.bai
-rw-rw-r--. 1 big big      1022 Apr 21 11:52 concat_ref.fasta
-rw-rw-r--. 1 big big        39 Apr 21 11:52 concat_ref.fasta.fai
-rw-rw-r--. 1 big big      1013 Apr 21 11:52 consensus_annotations.csv
-rw-rw-r--. 1 big big       677 Apr 21 11:52 consensus.bam
-rw-rw-r--. 1 big big        96 Apr 21 11:52 consensus.bam.bai
-rw-rw-r--. 1 big big       504 Apr 21 11:52 consensus.fasta
-rw-rw-r--. 1 big big        38 Apr 21 11:52 consensus.fasta.fai
-rw-rw-r--. 1 big big         0 Apr 21 11:52 donor_regions.fa
-rw-rw-r--. 1 big big       891 Apr 21 11:52 filtered_contig_annotations.csv
-rw-rw-r--. 1 big big       509 Apr 21 11:52 filtered_contig.fasta
-rw-rw-r--. 1 big big       991 Apr 21 11:52 filtered_contig.fastq
-rw-rw-r--. 1 big big       931 Apr 21 11:52 metrics_summary.csv
-rw-rw-r--. 1 big big   2476194 Apr 21 11:52 vdj_contig_info.pb
drwxrwxr-x. 3 big big        41 Apr 21 11:52 vdj_reference
-rw-rw-r--. 1 big big     27365 Apr 21 11:52 vloupe.vloupe
-rw-rw-r--. 1 big big   4312114 Apr 21 11:52 web_summary.html
```

结果文件说明：

web_summary.html 是网页版的比对结果（可以直接用浏览器打开），包括 mapping 率，饱和度等一系列结果。

metrics_summary.csv 是表格版的比对结果（可以直接用 excel 打开），内容与网页版一致。

all_contig_annotations.csv 是供下游 scRepotire 包分析 scTCR/BCR 用的结果，文件很小，只有 100KB 左右。

```
sample,barcode,is_cell,contig_id,high_confidence,length,chain,v_gene,d_gene,j_gene,c_gene,full_length,productive,fwr1,fwr1_nt,cdr1,cdr1_nt,fwr
2,fwr2_nt,cdr2,cdr2_nt,fwr3,fwr3_nt,cdr3,cdr3_nt,fwr4,fwr4_nt,reads,umis,raw_clonotype_id,raw_consensus_id,exact_subclonotype_id
Sub01,AAACCCATCAGCATTG-1,false,AAACCCATCAGCATTG-1_contig_1,false,348,TRB,TRBV27,,TRBJ1-5,TRBC1,false,false,,,,,,,,,,CASSFDLVGPQHF,TGTGCCAGCAG
TTTCGATTTAGTAGGACCCCAGCATTTT,,,76,1,,,
Sub01,AAACCCATCATTATCT-1,false,AAACCCATCATTATCT-1_contig_1,false,511,TRB,TRBV12-3,,TRBJ2-1,TRBC2,true,true,,,,,,,,,,CASSLISGVNNEQFF,TGTGCCAGC
AGTTTAATTAGCGGGGTGAACAATGAGCAGTTCTTC,,,9340,1,,,
Sub01,AAACCGGAGATTCAGT-1,false,AAACCGGAGATTCAGT-1_contig_1,false,449,TRB,TRBV25-1,,TRBJ2-1,TRBC2,true,true,,,,,,,,,,CASSEPGLEQFF,TGTGCCAGCAGT
GAACCGGGGCTTGAGCAGTTCTTC,,,78,1,,,
Sub01,AAACGCTGTGTTCGCT-1,false,AAACGCTGTGTTCGCT-1_contig_1,false,512,TRB,TRBV7-9,,TRBJ2-3,TRBC2,true,true,,,,,,,,,,CASRHPIRTDTQYF,TGTGCCAGCCG
GCACCCGATCAGGACAGATACGCAGTATTTT,,,10300,1,,,
Sub01,AAAGCCTGTAGTCGCA-1,false,AAAGCCTGTAGTCGCA-1_contig_1,false,362,TRB,TRBV7-3,TRBD1,TRBJ2-1,TRBC2,true,false,,,,,,,,,,CASSLGGRGEQFF,TGTGCC
AGCAGCTTAGGCGGGAGAGGGGAGCAGTTCTTC,,,100,2,,,
```

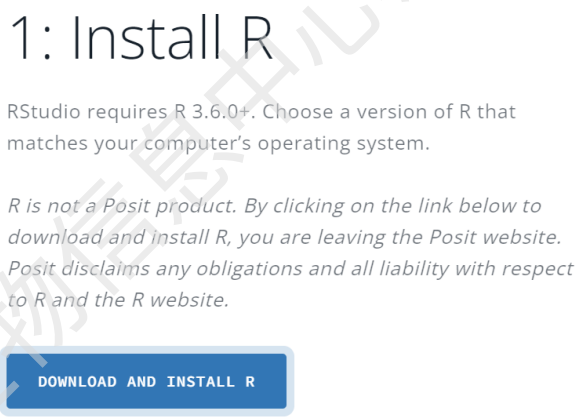
2. scRepertoire 包解析 scTCR/BCR 数据 (运行非常快，预计这部分的讲课用时 25min)

该部分可以在 windows 上进行，操作更友好，适合新手入门

2.1 下载并安装 R 和 Rstudio：

R 下载地址：https://cloud.r-project.org/

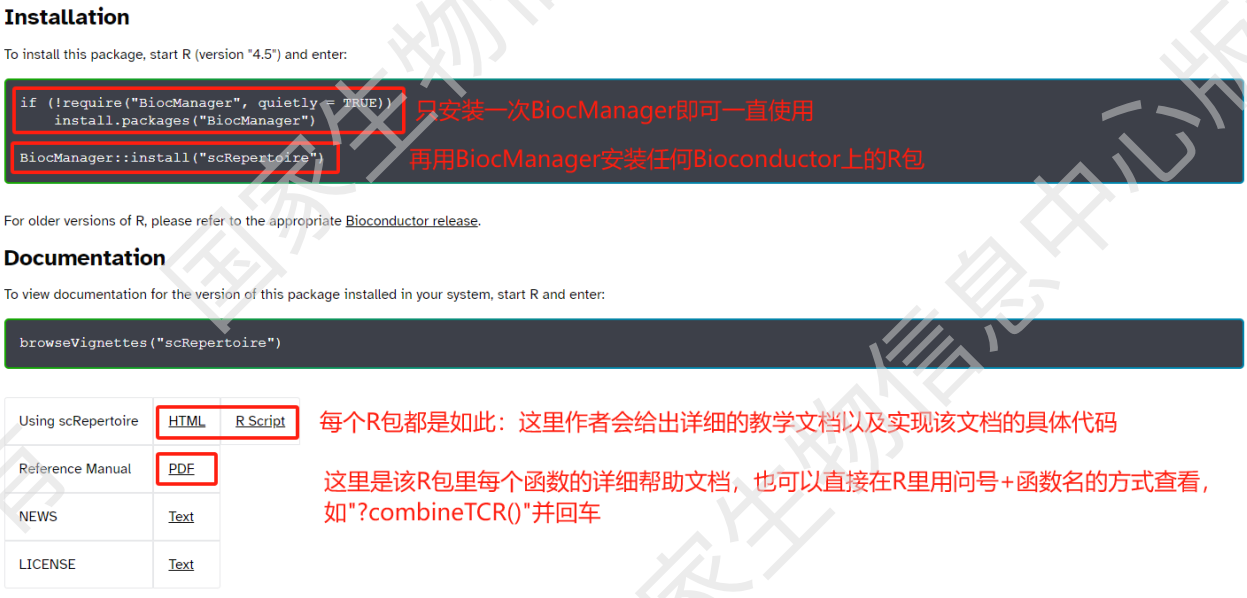
RStudio 下载地址：https://posit.co/download/rstudio-desktop/



2.2 Bioconductor 是 R 语言中专门为生物信息 R 包建立的管理网站，可以在 Bioconductor 官网上搜索任何你想要学习的生信相关 R 包

Bioconductor 网站：https://www.bioconductor.org/

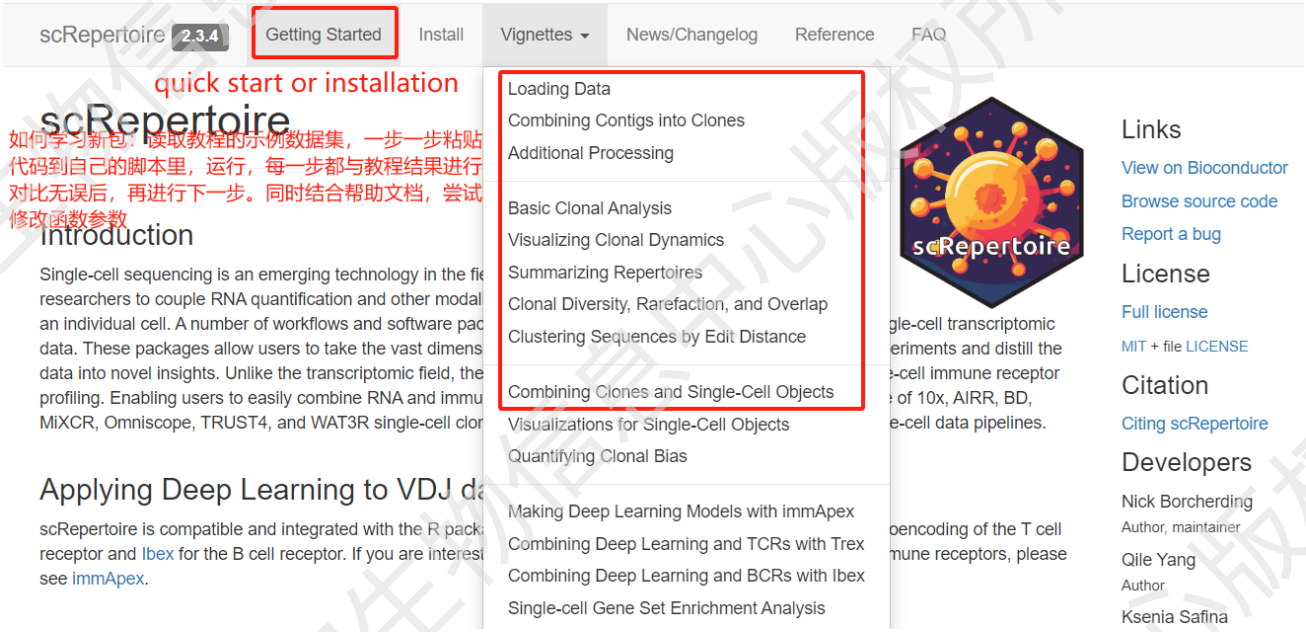
搜索 scRepertoire 后的结果：https://www.bioconductor.org/packages/release/bioc/html/scRepertoire.html



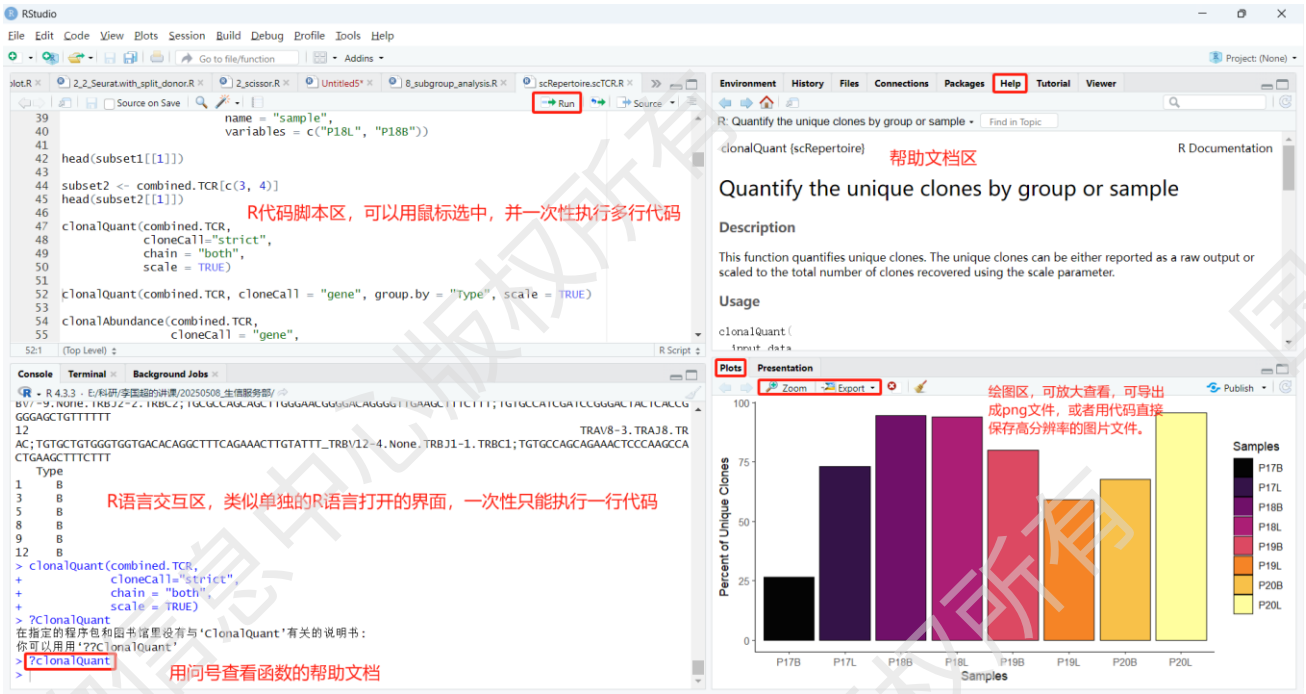
Details

biocViews	Annotation, Classification, ImmunoOncology, Sequencing, SingleCell, Software
Version	2.4.0
In Bioconductor since	BioC 3.12 (R-4.0) (4.5 years)
License	MIT + file LICENSE
Depends	ggplot2, R (>= 4.0)
Imports	assertthat, cubature, dplyr, evmix, ggalluvial, ggdendro, ggraph, grDevices, igraph, iNEXT, plyr, quantreg, Rcpp, reshape2, rjson, rlang, S4Vectors, SeuratObject, SingleCellExperiment, stringr, stringdist, SummarizedExperiment, tidygraph, truncdist, VGAM, purrr, lifecycle, methods
System Requirements	
URL	https://www.borch.dev/uploads/scRepertoire/
Bug Reports	https://github.com/BorchLab/scRepertoire/issues

有些优秀的R包会单独设置自己的教学网站，链接都放在这个位置，这个教程往往比上一页中的教程更好更新



用 RStudio 执行从教程网站粘贴来的代码：



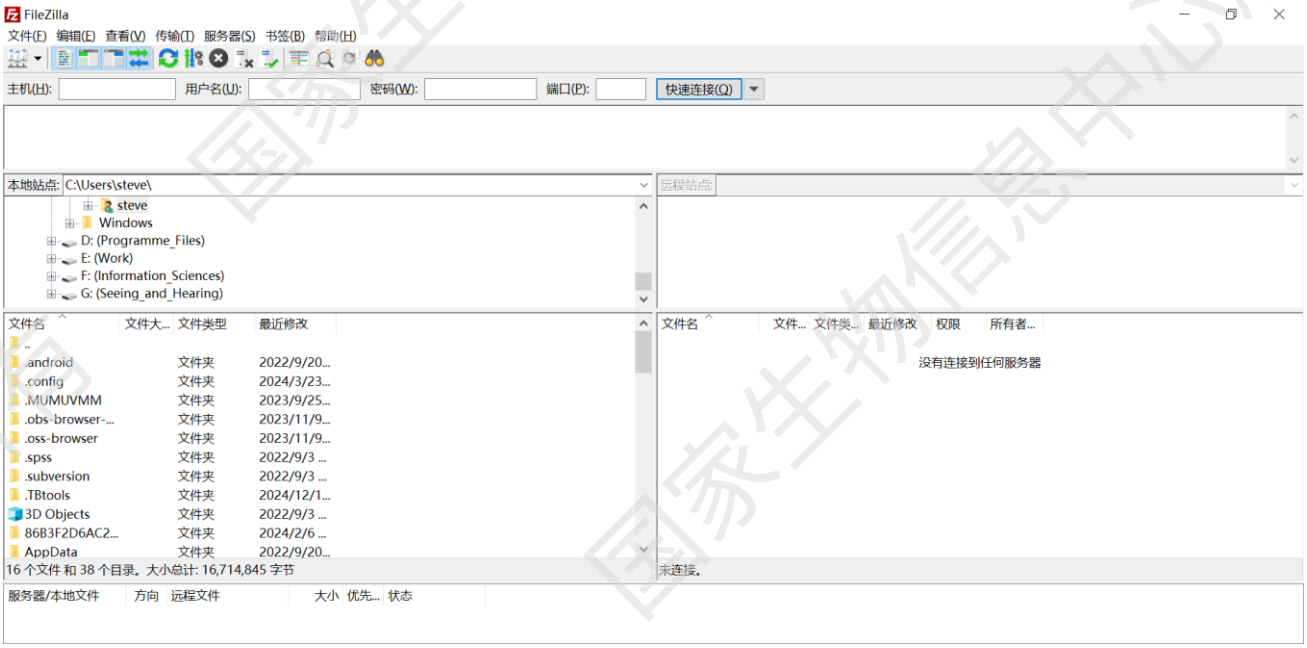
四、 补充：FileZilla 上传下载软件使用

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。

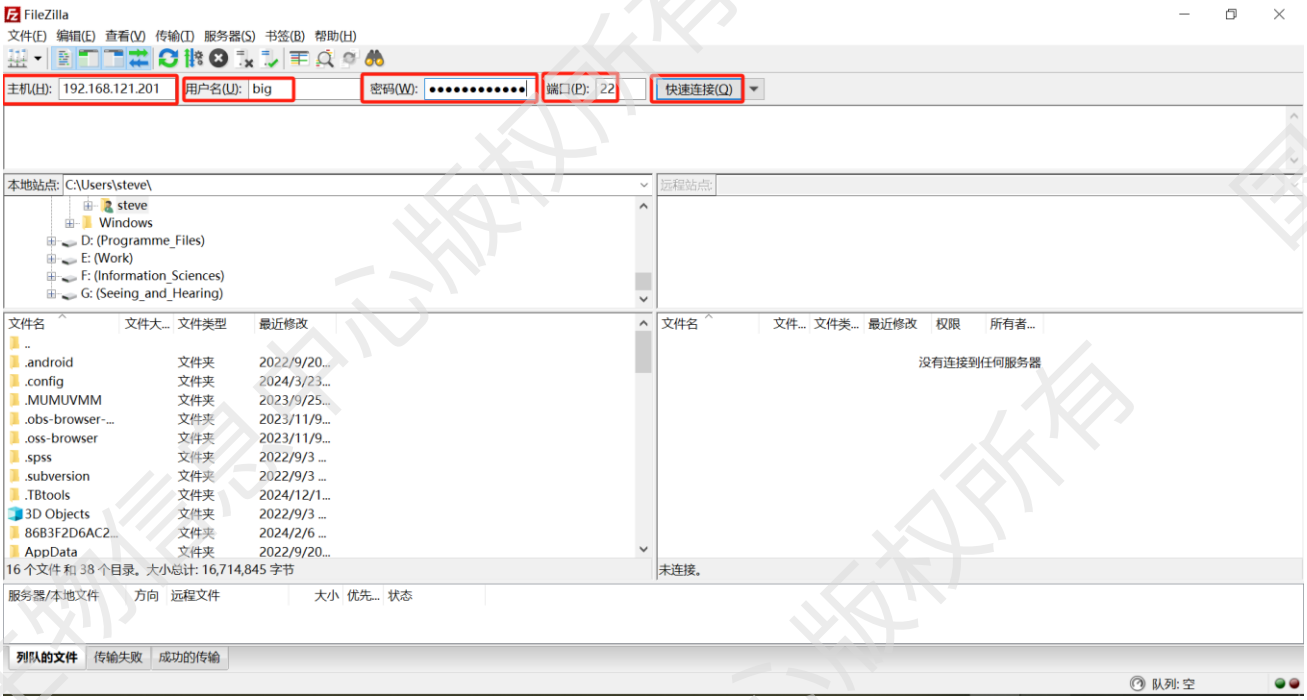
FileZilla 下载链接为 https://www.filezilla.cn/download/client，下载并安装。

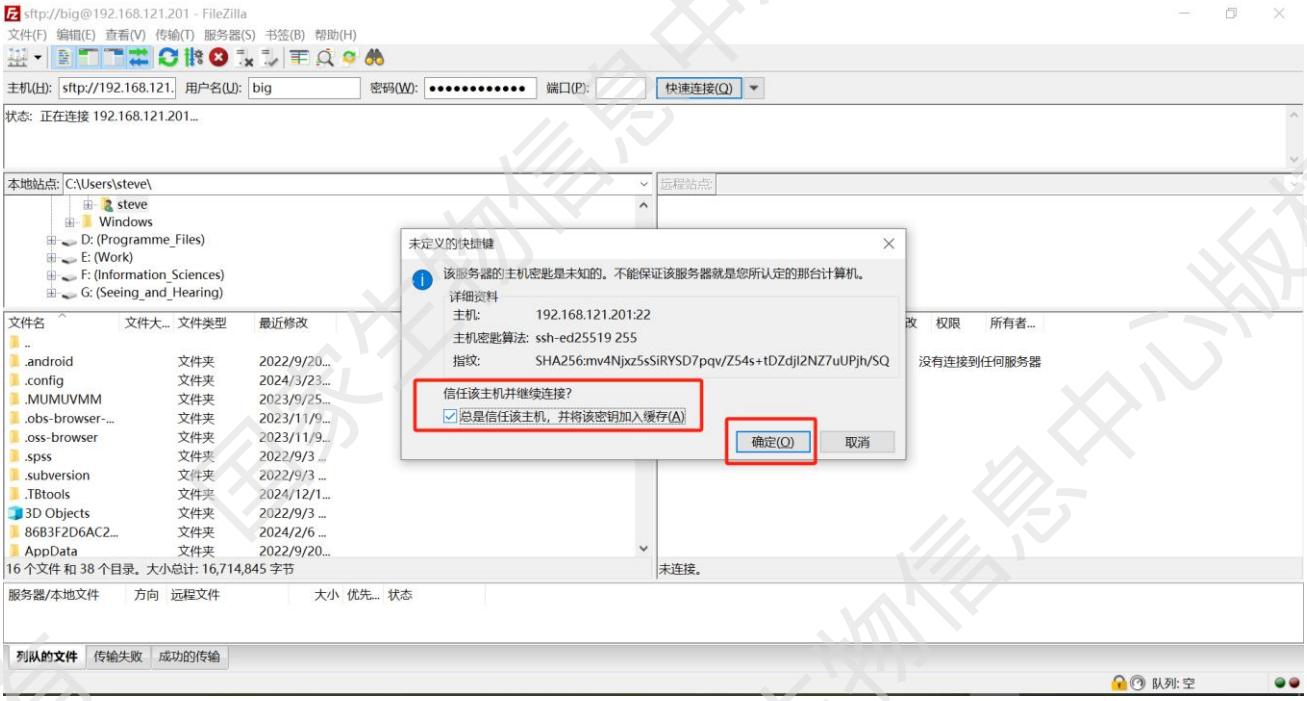
打开 FileZilla，进行如下操作链接我们的服务器：

1、打开 FileZilla，如下图：

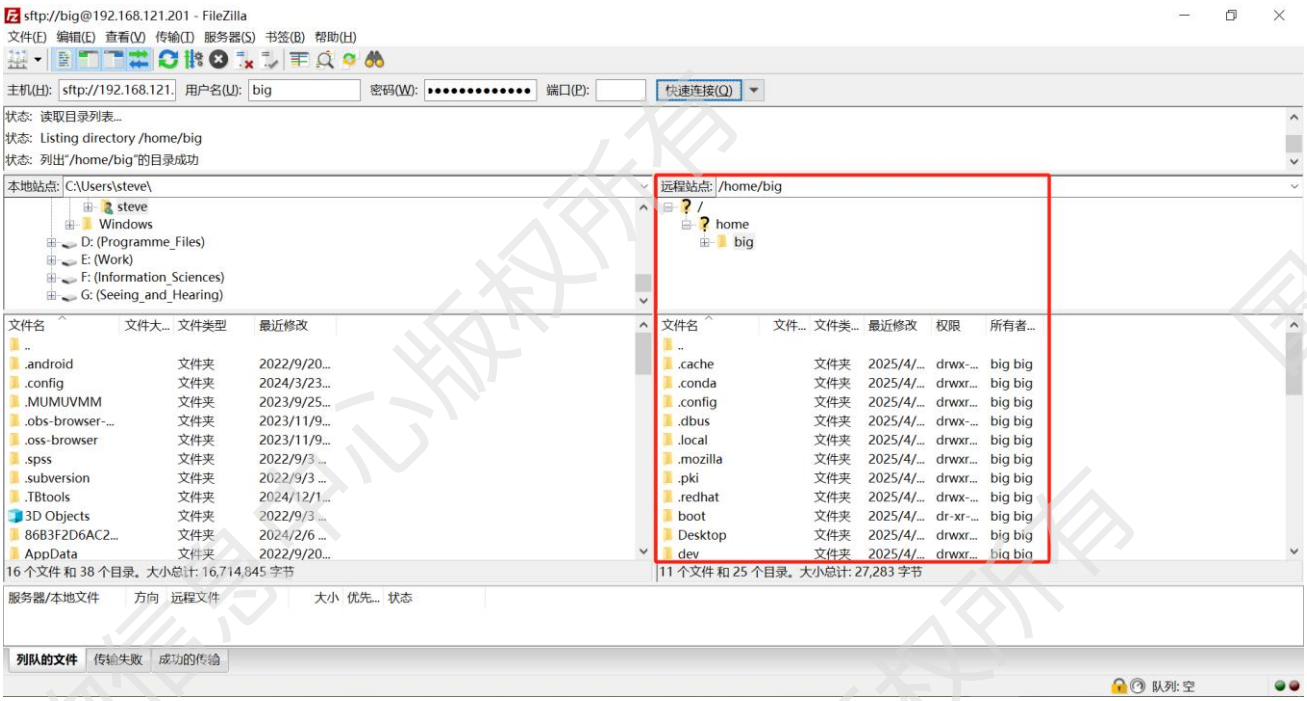


2、点击新建站点，依次填入 “主机 ID：192.168.121.201”，“端口：22”， “用户名”，“密码”，点击“快速连接”，即可以建立本地 pc 机与远程服务器的连接。





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。


sftp://big@192.168.121.201 - FileZilla
文件(F) 编辑(E) 查看(V) 传输(T) 服务器(S) 书签(B) 帮助(H)
主机(H): sftp://192.168.121. 用户名(U): big 密码(W): •••••••••••••• 端口(P): 快速连接(Q)
状态: 读取目录列表...
状态: Listing directory /home/big
状态: 列出"/home/big"的目录成功
本地站点: C:\Users\steve\
steve
Windows
D: (Programme_Files)
E: (Work)
F: (Information_Sciences)
G: (Seeing_and_Hearing)
远程站点: /home/big
/
home
big
文件名 文件大... 文件类型 最近修改
.. 文件夹
.android 文件夹 2022/9/20...
.config 文件夹 2024/3/23...
.MUMUVMM 文件夹 2023/9/25...
.obs-browser-... 文件夹 2023/11/9...
.oss-browser 文件夹 2023/11/9...
.spss 文件夹 2022/9/3 ...
.subversion 文件夹 2022/9/3 ...
.TBtools 文件夹 2024/12/1...
3D Objects 文件夹 2022/9/3 ...
86B3F2D6AC2... 文件夹 2024/2/6 ...
AppData 文件夹 2022/9/20...
16 个文件 和 38 个目录。大小总计: 16,714,845 字节
文件名 文件... 文件类... 最近修改 权限 所有者...
.. 文件夹
.cache 文件夹 2025/4/... drwx-... big big
.conda 文件夹 2025/4/... drwxr... big big
.config 文件夹 2025/4/... drwxr... big big
.dbus 文件夹 2025/4/... drwx-... big big
.local 文件夹 2025/4/... drwxr... big big
.mozilla 文件夹 2025/4/... drwxr... big big
.pki 文件夹 2025/4/... drwxr... big big
.redhat 文件夹 2025/4/... drwx-... big big
boot 文件夹 2025/4/... dr-xr-... big big
Desktop 文件夹 2025/4/... drwxr... big big
dev 文件夹 2025/4/... drwxr... big big
11 个文件 和 25 个目录。大小总计: 27,283 字节
服务器/本地文件 方向 远程文件 大小 优先... 状态
列队的文件 传输失败 成功的传输
队列: 空