实验 2.1 测序数据预处理

一、实验背景

DNA（脱氧核糖核酸）作为生物遗传信息的载体，分子构建当中含 4 种碱基，即腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶，分别用 4 个字母 A、G、C 和 T 表示。基因测序就是测定基因序列当中 A、G、C、T 这 4 种碱基的排布顺序。基因测序仪读取基因序列得到的 A、G、C、T 字符串序列片段称为 reads。接下来，我们就从实验 1.3 所得的下机 reads 开始，进入数据分析操作环节。

二、教学目标

序列是什么？ 哪些因素可能引起测序错误？哪些指标可用于评价测序质量？低质量的数据如何处理？

本节课程，我们将从了解下机数据的格式和数据质量概貌开始学习下机数据质量评估的基本思路和操作方法。

三、实验原理

下机数据概貌：

本次下机数据为 FASTQ 格式，用于保存测序仪的原始下机数据，文件后缀通常为.fastq 或.fq。以@开头的一行是一条 read 的名字；第二行，是测序仪下机数据中这条 read 的碱基序列；以“+”开头的行，旧版 FASTQ 格式中会在这一行重复第一行信息，现在为节约存储，一般是不再加重复信息；最后一行是测序 read 的碱基质量值，这一行信息至关重要，用于描述第二行测序数据中每一个测序碱基的可靠程度，用 ASCII 码表中的控制符号表示。

```
@FS2000L1C002R004000052
GGACAGTTCACCCCTCCTTAGGCAACCCGGTGGTCCCCTGCTCCTGGCAG
+
IIIIIII=IIFIIIIIIIIIIIIIIIIIIIIIIIIIIGIIIIIIIII
@FS2000L1C002R004000106
CATTAAACCCAGCACCTACCCTCAGAAATCGCCTCCCAAGCGTTTACATC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@FS2000L1C002R004000116
TTCAGCCCACACCTCTCCTCAGCCCATTACTGTGCAAAGTAGTTCCTAGA
+
IIIIIIIIIIIII?IIIIIIIIBIII@EII7I@I?IIIIBIII%IIIIII
@FS2000L1C002R004000133
CGAAAACCTTTTTCCAAGGACAAATCAGAGAAAAAGTCTTTAACTCCACC
+
IIIIIIIIII;9IIIIIIIIIIBIIIIIIIIIII<IIIIIII=IAIII
```

碱基质量体系：

在以上 FASTQ 文件中，有一行碱基质量值信息，每一个字符对应测序 read 上该位置碱基质量值，用来衡量碱基的测序准确度。其中碱基质量值计算公式：

Q = -10 * lg(p)

其中 p 为碱基被识别错误的概率。根据公式计算所得 Q 值越高，代表碱基被测错的概率越小。测序数据下机质量评价指标中，Q20，Q30，Q40 就分别代表测序错误率为 1%，0.1%，0.01%。

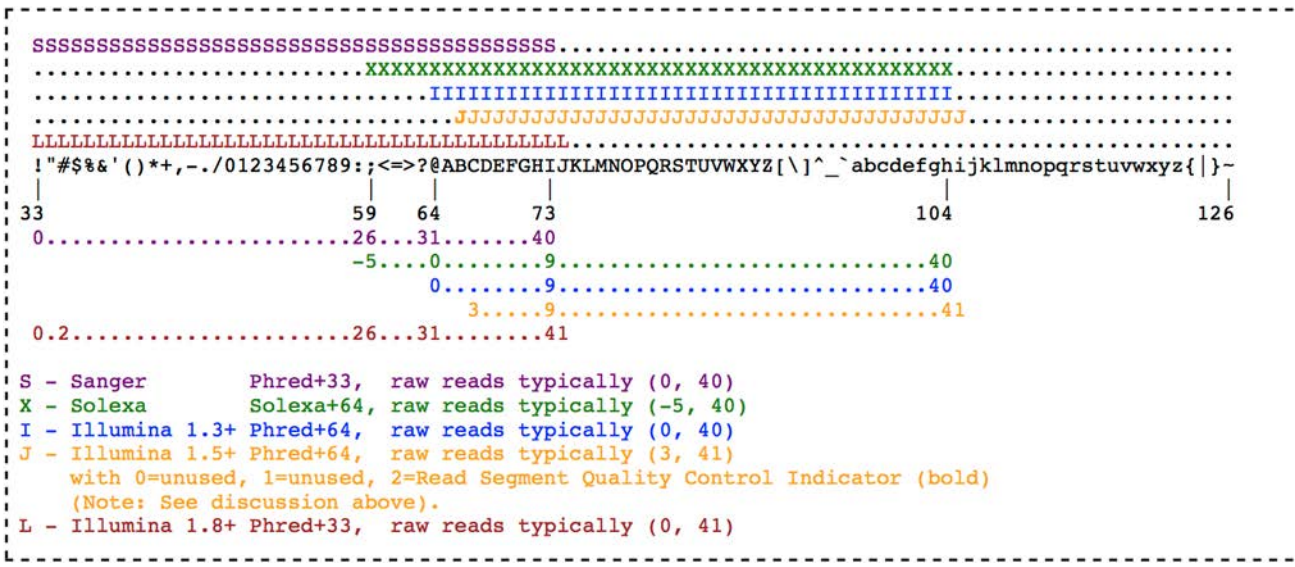
下机数据 Q30 >85%，代表所有下机数据当中测序错误率小于 0.1%的碱基占总碱基数的比例超过 85%。

为了对碱基质量值进行准确记录并节约存储，采用 ASCII 表对碱基质量值进行转换。对应转换方式比较常见的有 Phred33 和 Phred64 两种不同的质量体系。

Phred33：从 ASCII 码表可打印的字符开始，将质量值 0~40 对应转换为 ASCII 值 33 到 73 对应的控制字符“!”~“I”：（即用“实际碱基质量值+33”所得数值对应的 ASCII 码表中的控制字符代表该碱基质量值）

Phred64：用“实际碱基质量值+64”所得数值对应的 ASCII 码表中的控制字符代表该碱基质量值。

其他质量值体系如 Solexa+64 等的质量值与控制字符的对应关系见以下图示。



影响下机数据质量的常见因素与评价指标

1. 接头污染（测序接头残留）

2. 低质量（按照经验值，通常将 Q10 占比超过 10%的 reads 视为低质量 reads）

3. N rate （测序过程中未识别的碱基占总碱基的比例）

常见序列保存信息格式补充介绍：

分析流程产生的文件：

过滤➞clean.fq➞比对➞clean.bwa.sam➞数据格式转换➞

clean.bwa.bam➞索引结果排序➞clean.bwa.sorted.bam➞变异检测➞

clean.bwa.snp.vcf

基因序列信息保存还有一种常见格式为 FASTA，与 FASTQ 相比，不包含“+”行和碱基质量值信息，以“>”作为名字开头而非“@”，并且 fasta 文件只记录核酸序列信息或蛋白序列信息，后缀通常为.fasta 或.fa，后续分析过程中也会用到。

注：以“>”开头的行为序列头信息，记录该序列的名字，有时还包含序列长度、接头等其他信息；紧接的一行是序列信息，直到碰到下一个“>”开头的新序列或文件末尾为止。

我们采用 SOAPnuke filter 对下机数据各项质量指标进行统计，并对不符合质量要求的数据进行过滤。以下是 SOAPnuke filter 主要参数：

（运行 SOAPnuke filter –help 命令行，可查看 SOAPnuke 更多完整参数）

-l, --lowQual INT low quality threshold [default:5]

-q, --qualRate FLOAT low quality rate [default:0.5]

-n, --nRate FLOAT N rate threshold [default:0.05]

-f, --adapter1 STR adapter sequence or list file of read1

-r, --adapter2 STR adapter sequence or list file of read2 (if PE)

-1, --fq1 FILE fq1 file(required), .gz or normal text format are both supported(required)

-2, --fq2 FILE fq2 file(used when process PE data), format should be same as fq1 file, both are gz or both are normal text

-C, --cleanFq1 STR reads which passed QC from fq1 file would output to this file

-D, --cleanFq2 STR reads which passed QC from fq2 file would output to this file

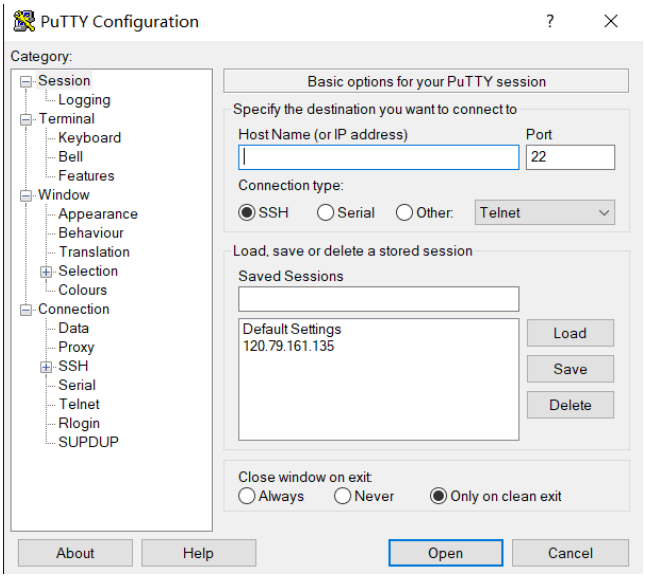
-o, --outDir STR Output directory. Processed fq files and statistical results would be output to here

〖教学重点〗在本实验环节中，学生需掌握下机数据的基本格式，并了解碱基质量值的含义与计算方法，理解不同碱基质量体系的核心差异，做到可以举一反三地理解其他采用不同碱基质量体系的常见下机数据基本格式。掌握下机数据质量的评价指标与评估方法。

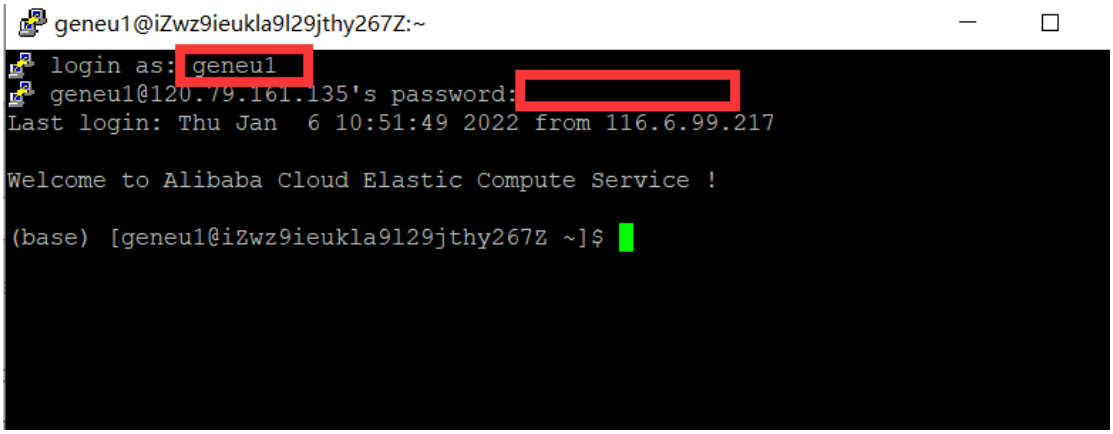
四、软件安装

✧ 集群服务器登录：

Windows 系统：根据自己的处理器型号下载并安装相应版本的 putty
https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html



安装 putty 后打开 putty 集群登录工具，在 Host Name 栏输入 IP 地址，Port 端口设置为22，Connection type 类型选 SSH，Saved Sessions 处再次输入 IP 地址，设置完成后，双击右侧 Load，即可进入集群登录界面。在此界面上 login as 处输入集群登录账号，回车后再输入登录密码，即可开始数据分析操作。



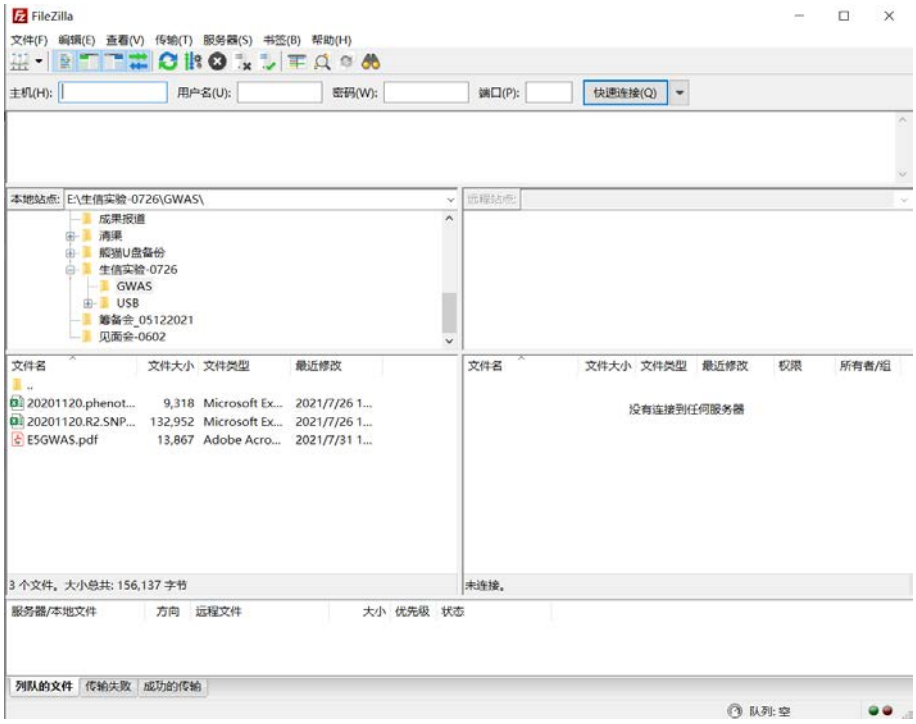
MacOS 系统：打开 Terminal（终端），键入：（以10.184.150.35为例，此 IP 地址为测试用，上课用 IP 地址以教学指导老师课堂提供的为准）

ssh 用户名@10.184.150.35

然后输入密码进入集群系统。

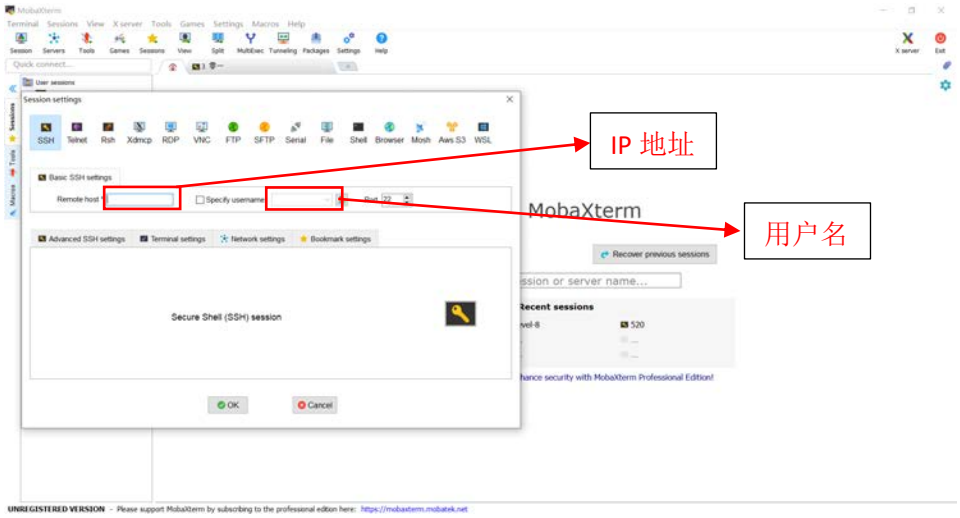
✧ FileZilla 下载及安装

打开链接，选择下载适用自己系统的数据传输工具 FileZilla Client，并安装。安装完成后打开 FileZilla 数据传输工具，在主机口填入 IP 地址，用户名、密码与登录集群的账号、密码相同，端口设置 22，点击快速连接，即可将本地计算机与集群服务器相连。



五、实验步骤

第一部分　创建文件夹及拷贝数据



1、在 MobaXterm 上通过指定账号、密码以及 IP 地址远程登录集群，查看当前所在数据路径：

pwd

```
(base) [geneu1@iZwz9ieukla9l29jthy267Z Tom]$ pwd
/home/geneu1/Tom
```

例，本测试路径使用 pwd 查看时，显示当前路径为：/home/geneu1/Tom

2、创建工作目录

mkdir 00_ref 01_clean 02_align 03_SNPCalling 04_annotation raw_data

3、查看目录是否创建成功
ls

```
(base) [geneu1@iZwz9ieukla9l29jthy267Z Tom]$ mkdir 00_ref 01_clean 02_align 03_SNPCalling 04_annotation raw_data
(base) [geneu1@iZwz9ieukla9l29jthy267Z Tom]$ ls
00_ref  01_clean  02_align  03_SNPCalling  04_annotation  raw_data
```

4、将当前目录下的测序数据，挪到 raw_data 目录中
mv ~/*.fq.gz ~/raw_data

```
LY@ecoli-ThinkCentre-M930q-N000:~$ mv *.fq.gz ./raw_data
LY@ecoli-ThinkCentre-M930q-N000:~$ ll
```

5、将以下路径目录中所有文件（/home/00_ref）拷贝至自己个人的新建目录 00_ref 下：
cd
cd 00_ref
cp /home/00_ref/*.* ./

```
LY@ecoli-ThinkCentre-M930q-N000:~/00_ref$ cd
LY@ecoli-ThinkCentre-M930q-N000:~$ cd 00_ref
LY@ecoli-ThinkCentre-M930q-N000:~/00_ref$ ll
total 8
drwxr-xr-x  2 LY student 4096 10月  24 16:17 ./
drwxr-xr-x 13 LY student 4096 10月  24 16:17 ../
LY@ecoli-ThinkCentre-M930q-N000:~/00_ref$ cp /home/00_ref/*.* ./
LY@ecoli-ThinkCentre-M930q-N000:~/00_ref$ ll
total 12804
drwxr-xr-x  2 LY student    4096 10月  24 16:28 ./
drwxr-xr-x 13 LY student    4096 10月  24 16:17 ../
-rw-r--r--  1 LY student    2331 10月  24 16:28 alleles_all.vcf.gz
-rw-r--r--  1 LY student    3143 10月  24 16:28 alleles_all.vcf.gz.tbi
-rw-r--r--  1 LY student  448559 10月  24 16:28 MGI358.SNP.fa
-rw-r--r--  1 LY student  891846 10月  24 16:28 MGI358.SNP.fa.0123
-rw-r--r--  1 LY student      13 10月  24 16:28 MGI358.SNP.fa.amb
-rw-r--r--  1 LY student    7087 10月  24 16:28 MGI358.SNP.fa.ann
-rw-r--r--  1 LY student 5351187 10月  24 16:28 MGI358.SNP.fa.bwt.2bit.64
-rw-r--r--  1 LY student 6243027 10月  24 16:28 MGI358.SNP.fa.bwt.8bit.32
-rw-r--r--  1 LY student    6861 10月  24 16:28 MGI358.SNP.fa.fai
-rw-r--r--  1 LY student  111482 10月  24 16:28 MGI358.SNP.fa.pac
-rw-r--r--  1 LY student    4661 10月  24 16:28 target.358.SE50.subSNP.bed
-rw-r--r--  1 LY student    4957 10月  24 16:28 vcf2geno_free.pl
```

6、在集群分析路径下使用以下命令检查数据文件拷贝是否齐全。
cd
ls raw_data
ls 00_ref/

```
LY@ecoli-ThinkCentre-M930q-N000:~$ cd
LY@ecoli-ThinkCentre-M930q-N000:~$ ls raw_data
test.fq.gz
LY@ecoli-ThinkCentre-M930q-N000:~$ ls 00_ref
alleles_all.vcf.gz      MGI358.SNP.fa       MGI358.SNP.fa.amb  MGI358.SNP.fa.bwt.2bit.64  MGI358.SNP.fa.fai  target.358.SE50.subSNP.bed
alleles_all.vcf.gz.tbi  MGI358.SNP.fa.0123  MGI358.SNP.fa.ann  MGI358.SNP.fa.bwt.8bit.32  MGI358.SNP.fa.pac  vcf2geno_free.pl
```

7、使用以下命令查看下机数据基本格式：
less ~/raw_data/test.fq.gz
###(查看完毕，单击 q 即可退出)

```
@FS2000L1C002R004000021
CGAAAACCTTTTTCCAAGGACAAATCAGAGAAAAAGTCTTTAACTCCACC
+
IIII=CII<IIIIIIIIIIIIIIIHIIIIIII0<IHIIIIII12IIIID
@FS2000L1C002R004000025
ACTCTATTAGTCAACGTTCTCCAGAGAAACAGAACCAATAGCATAGATAG
+
IIIIII;IIIDIIIIGIIIIIDIHIII>IIIIIIIIIIII'IIIHIIIC
@FS2000L1C002R004000032
TTATGTTCAGCCCTTGGATTGTCTCAGGATGTTGCAGGCGCAACTTTCAT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

8、使用 SOAPnuke 生成数据质量评估表并按指定标准对低质量数据进行过滤：
/Pipeline/FIS.Traits/tools/SOAPnuke filter -1 ~/raw_data/test.fq.gz -l 10 -q 0.5 -n 0.01 -T 1 -o ~/01_clean/ -C test_clean.fq.gz
（注意，此命令为连续一行，需手动将此 word 文档中第一行的行尾尾换行符改为一个空格符）

```
(base) [geneu1@iZwz9ieukla9l29jthy267Z Tom]$ SOAPnuke filter -1 ./raw_data/test.fq.gz -l 10 -q 0.5 -n 0.01 -T 1
(base) [geneu1@iZwz9ieukla9l29jthy267Z Tom]$ ls 01_clean/
Base_distributions_by_read_position_1.txt             Distribution_of_Q20_Q30_bases_by_read_position_1.txt  St
Base_quality_value_distribution_by_read_position_1.txt  log                                                 te
Basic_Statistics_of_Sequencing_Quality.txt            Statistics_of_Filtered_Reads.txt
(base) [geneu1@iZwz9ieukla9l29jthy267Z Tom]$
```

过滤前后数据质量对比统计：
9、可在01_clean 路径下使用 less 命令逐一查看测序数据质量评估表，我们简单展示过滤前后数据总体情况如下：
less ~/01_clean/Basic_Statistics_of_Sequencing_Quality.txt
###(查看完毕，单击 q 即可退出)

Item	raw reads(fq1)	clean reads(fq1)
Read length	50.0	50.0
Total number of reads	473994 (100.00%)	473598 (100.00%)
Number of filtered reads	396 (0.08%)	-
Total number of bases	23699700 (100.00%)	23679900 (100.00%
Number of filtered bases	19800 (0.08%)	-
Number of base A	6753420 (28.50%)	6747827 (28.50%)
Number of base C	5643559 (23.81%)	5638836 (23.81%)
Number of base G	4954525 (20.91%)	4950542 (20.91%)
Number of base T	6347790 (26.78%)	6342695 (26.79%)
Number of base N	406 (0.00%)	0 (0.00%)
Q20 number	23171219 (97.77%)	23152679 (97.77%)
Q30 number	22639917 (95.53%)	22622162 (95.53%)

〖教学重点〗本实验要求学生熟练掌握集群登录方法，linux 系统常见命令和数据过滤工具的使用，最终达到可轻松访问指定数据路径，根据数据分析需要创建工作目录，拷贝、查看数据等数据操作基本命令，逐步熟悉 linux 系统环境。同时需了解数据的基本格式、质量评估的主要参数和低质量数据过滤的方法，获取可用于下一步数据分析的有效数据。

六、实验后处理和预期结果

为帮助学生深入理解数据质量评估及处理的意义，希望同学进一步学习使用 FASTQC 工具对数据质量情况进行绘图，通过图形化展示，更加清晰直观地了解本环节对数据进行处理的必要性和效果。使用方法参考链接：https://www.jianshu.com/p/4d388cb26596

0、学习在自己电脑上进行 FASTQC 工具安装：

conda install -c bioconda fastqc

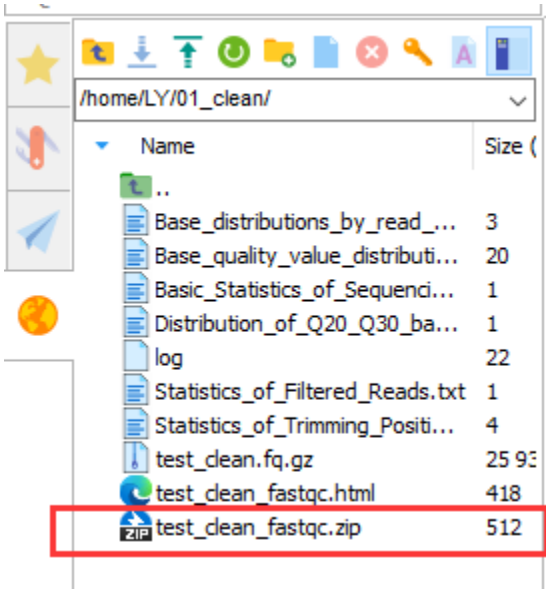
（服务器上已安装）

1、运行 FastQC：

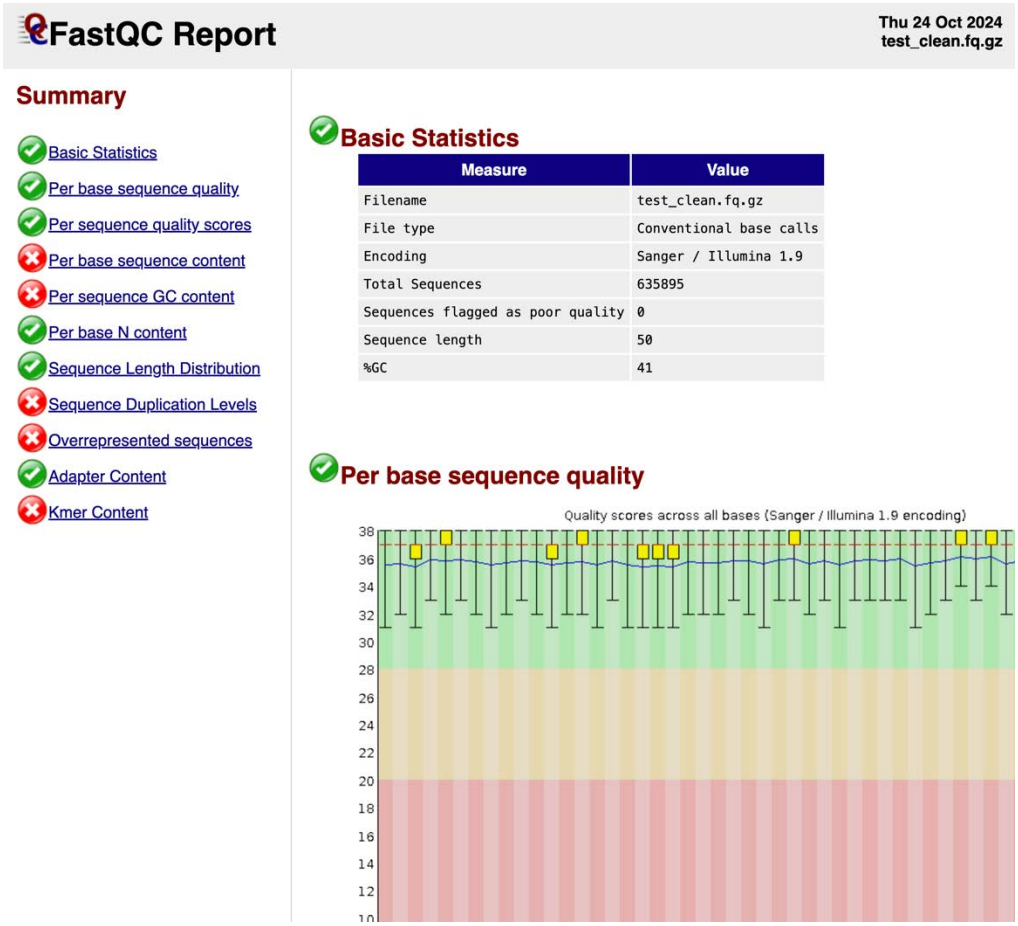
/Pipeline/FIS.Traits/tools/anaconda/bin/fastqc ~/01_clean/test_clean.fq.gz

```
LY@ecoli-ThinkCentre-M930q-N000:~$ /Pipeline/FIS.Traits/tools/anaconda/bin/fastqc ~/01_clean/test_clean.fq.gz
Started analysis of test_clean.fq.gz
Approx 5% complete for test_clean.fq.gz
Approx 10% complete for test_clean.fq.gz
Approx 15% complete for test_clean.fq.gz
Approx 20% complete for test_clean.fq.gz
Approx 25% complete for test_clean.fq.gz
Approx 30% complete for test_clean.fq.gz
Approx 35% complete for test_clean.fq.gz
Approx 40% complete for test_clean.fq.gz
Approx 45% complete for test_clean.fq.gz
Approx 50% complete for test_clean.fq.gz
Approx 55% complete for test_clean.fq.gz
Approx 60% complete for test_clean.fq.gz
Approx 65% complete for test_clean.fq.gz
Approx 70% complete for test_clean.fq.gz
Approx 75% complete for test_clean.fq.gz
Approx 80% complete for test_clean.fq.gz
Approx 85% complete for test_clean.fq.gz
Approx 90% complete for test_clean.fq.gz
Approx 95% complete for test_clean.fq.gz
Analysis complete for test_clean.fq.gz
```

2、下载 FastQC 质控文件至自己本地电脑：



3、解压缩 FastQC 质控文件，用浏览器打开目录中的 fastqc_report.html 文件：



经 FASTQC 工具绘图后，我们将得到可视化的碱基质量值分布图。以上图为例，横轴为 50bp 碱基序列上不同的碱基位置，纵轴为碱基质量值。