Yangyang Li
5674840
li002252@umn.edu

Homework Assignment 1
BICB 8510 2021-01-30

2021-01-30

1 Question 1

I get the current work directory by using command **pwd**:

/home/bicb8510/li002252

2 Question 2

2.1 a

There are **16** files in */home/bicb8510/public/Mills_HW_1* and the command:

```
ls -l /home/bicb8510/public/Mills_HW_1 |wc -l
```

Note: we need to remove the line of total size of this directory

2.2 b

8 unique samples exists in this directory in that each sample, which are pair-end sequenced, has two read files.

3 Question 3

3.1 a

I run the fastq file **/home/bicb8510/public/Mills_HW_1/163C_S2_R1_001.fastq**

3.2 b

There are **34299108** sequences totally in this file in terms of the report.

3.3 c

1. the length of the sequences is **51**.
2. the mean quality of all bases more than **30**, which indicates the sample has high enough quality that enables us to conduct the following analysis. However, the bases, which are at the head and tail of reads, have relatively larger variances for quality across all sequences than others. We can use some software to filter these bases with bad quality.

4 Question 4

After conducting quality control for the fastq files of all samples, and results locates in the directory */home/yang4414/li002252/fq*.

I find out that the quality of the sample *176C_S8_R2_001.fastq* is relatively worse than that of others, although there are some other samples whose quality are not good. What is more, the quality of the sample is not stable, especially for the bases of the head and tail.
Note:

1. I use the **multiqc** software to integrate all report of quality, and the integrated report locates in */home/yang4414/li002252/fq/multiqc_report.html*
2. the report of the sample locates in */home/yang4414/li002252/fq/176C_S8_R2_001_fastqc.html*