Sequence Data Quality Assessment Exercises.

Starting Note: Please do not copy and paste the commands. Characters in this document may not be copied correctly. Please type the commands and use **<tab> complete** for commands, directories and long names.

Loading Modules:

First do	`module load bioinfo-tools`	and then
FastQC:	`module load FastQC/0.11.5`	
Seqtk:	`module load seqtk/1.0-r68e`	
Trimmomatic:	`module load trimmomatic/0.32`	

1. Use **`md5sum`** to calculate the checksum of all data files in **`/proj/g2016024/nobackup/QC_Data/`**.
   Redirect (**`>`** operator) the output into a file called **`checksum.txt`** in your workspace.

2. Make a copy of the data in your workspace (note the **`.`** at the end):
   ```
   cp -vr /proj/g2016024/nobackup/QC_Data/* .
   ```
   Use **`md5sum`** with the **`-c`** option and **`checksum.txt`** to check the files are complete.

3. Use **`file`** to test the files. In what format is the data compressed?

4. Use **`zcat`** and **`head`** to view the first 8 lines of **`Bacteria/bacteria_R1.fastq.gz`**.

5. From which sequencing technology is

   a. **Bacteria/bacteria_R{1,2}.fastq.gz**

   b. **Ecoli/E01_1_135x.fastq.gz**

6. What is each part of the FastQ header?
   `@HWI-ST486:212:D0C8BACXX:6:1101:2365:1998 1:N:0:ATTCCT`

7. What is each part of this FastQ header?
   `@m151121_235646_42237_c100926872550000001823210705121647_s1_p0/81/22917_25263`

8. What does each tool in this command do?

   zcat <fastq.gz> | seqtk seq -A - | grep -v “^>” | tr -dc “ACGTNacgtn” | wc -m

9. Use the command above to calculate how much data is in
   a. **Bacteria/bacteria_R{1,2}.fastq.gz**

   b. **Ecoli/E01_1_135x.fastq.gz**

10. How much data in **Ecoli/E01_1_135x.fastq.gz** are contained in reads 10kb or longer?

11. Run FastQC (**fastqc**) on the data files:
    fastqc -t 6 Bacteria/*.fastq.gz Ecoli/*.fastq.gz
    How many sequences are in each file (use either **fastqc** or **firefox** to open the html)?

12. What is the average GC% in each data set?

13. Which quality score encoding is used?

14. What does a quality score of 20 (Q20) mean?

15. What does a quality score of 40 (Q40) mean?

16. For **Bacteria/bacteria_R{1,2}.fastq.gz**, in the per base sequence plot, what percentage should the G and C lines be at, and why?

17. For **Bacteria/bacteria_R{1,2}.fastq.gz**, in the per base sequence plot, what percentage should the A and T lines be at, and why?

18. What distribution should the per base sequence plot follow?

19. What distribution should the per base GC plot follow?

20. What value should the per base GC distribution be centered on?

21. How much duplication is present in **Bacteria/bacteria_R{1,2}.fastq.gz**?

22. What is adapter read through?

23. After loading Trimmomatic look at **$TRIMMOMATIC_HOME/adapters** using
    ls $TRIMMOMATIC_HOME/adapters.
    This folder contains adapter sequence files from various library preparation kits.
    Trim **Bacteria/bacteria_R{1,2}.fastq.gz** using the TruSeq3-PE.fa file.

```
java -jar $TRIMMOMATIC_HOME/trimmomatic.jar PE
bacteria_R1.fastq.gz bacteria_R2.fastq.gz
bacteria_R1.trimmed.paired.fastq.gz
bacteria_R1.trimmed.unpaired.fastq.gz
```

```
bacteria_R2.trimmed.paired.fastq.gz
bacteria_R2.trimmed.unpaired.fastq.gz
ILLUMINACLIP:$TRIMMOMATIC_HOME/adapters/TruSeq3-
PE.fa:2:3:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15
MINLEN:36
```