Sequence Data Quality Assessment Exercises and Solutions.

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.

   ```
   $ cd /proj/g2016024/nobackup/QC_Data
   $ md5sum */* > ~/checksum.txt
   ```

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.

   ```
   $ md5sum -c checksum.txt
   ```
   (If the checksum wasn't generated in the QC_Data folder, the paths in checksum.txt need to be edited to reflect the new locations they are checking)

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

   `$ file */*`

   gzip compressed data

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

   `$ zcat Bacteria/bacteria_R1.fastq.gz | head -n8`

5. From which sequencing technology is
   a. **`Bacteria/bacteria_R{1,2}.fastq.gz`**

      Illumina

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

      `$ zcat E01/E01_1_135x.fastq.gz | head`

      Pacific Biosciences

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

   Machine number:
   Run number:
   Flowcell ID:
   Lane:
   Tile:
   X-coord:
   Y-coord

   First/Second in pair:
   Failed Illumina QC (Chastity):
   Control bit:

Barcode index sequence/ID

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

   Movie name consisting of the date, time, instrument, and smrt cell
   barcode
   set number
   part number
   ZMW number
   subread start_subread end

8. What does each tool in this command do?

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

   zcat – concatenates compressed files to one output stream.
   seqtk – toolkit for manipulating sequence data (seqtk seq –A converts the file to a fasta output).
   grep – searches through files for lines containing the given string (-v excludes lines containing the given string).
   tr – translates characters from one set to another (-dc deletes any character not in the given character set).
   wc – word count (-m counts characters).

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

      $ zcat Bacteria/bacteria_R{1,2}.fastq.gz | seqtk seq -A - | grep -v "^>" | tr -dc "ACGTNacgtn" | wc -m

225890464 (nucleotides)

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

$ zcat Ecoli/E01_1_135x.fastq.gz | seqtk seq -A - | grep -v "^>" | tr -dc "ACGTNacgtn" | wc -m
748508257 (nucleotides)

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

$ zcat Ecoli/E01_1_135x.fastq.gz | seqtk seq -A -L10000 - | grep -v "^>" | tr -dc "ACGTNacgtn" | wc -m
510546313 (nucleotides)

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)?

766616 (Bacteria/bacteria_R{1,2}.fastq.gz)
87217 (Ecoli/E01_1_135x.fastq.gz)

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

40 (Bacteria/bacteria_R{1,2}.fastq.gz)
49 (Ecoli/E01_1_135x.fastq.gz)

13. Which quality score encoding is used?

Sanger / Illumina 1.9

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

    An expectation of 1 error in 100bp

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

    An expectation of 1 error in 10000bp

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?

    20, because the mean GC is 40% and G and C should be in equal proportions and therefore half of the mean GC%.

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?

    30, because the mean AT is 60% and A and T should be in equal proportions and therefore half of the mean AT%.

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

    A Uniform distribution

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

    A Gaussian/Normal distribution

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

    Average GC content

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

    24% (R1) and 15% (R2)

22. What is adapter read through?

    When the sequence reads past the insert into the adapter sequence on the other end.

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
    ```