NGS Bioinformatics

Module topic: NGS Data Processing and QC
Contact session title: Read Data QC
Trainer: Fatma Guerfali and Shaun Aron
Participant: <*write your name here*>
Date: <*write today's date here*>

FASTQ Data QC

Introduction

In the assignment, you will have an opportunity to view and run a set of FASTQ files through the FASTQC program and assess the html output reports.

- *View raw high throughput sequencing files (FASTQ).*
- *Check the quality of the dataset = run a standard QC pipeline.*
    - *This step is crucial to ensure that your dataset is of good quality before you continue further with the analysis.*
- *Remove adapters and low quality reads from the dataset.*
- *Compare a « good » and a « bad » quality dataset.*

Tools used in this session

FASTQC - https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trimmomatic - https://github.com/usadellab/Trimmomatic

Please note

- **Hand-in information** please upload your completed assignment to the Vula 'Assignments' tab. Take note of the final hand-in date for each assignment, which will be indicated on Vula.

Task 1: Quality Control of a FASTQ dataset

The datasets for this assignment can be found on your VM here: ~/course_data/data_formats/data
The dataset you will be working with first is paired end reads from an Escherichia coli K1 sample. The data has been produced on the Illumina MiSeq platform and is stored in the SRR957824.zip archive.

```
cd /course_data/data_formats/data
```

Unzip the SRR957824 archive and cd into the unzipped directory

```
gunzip SRR957824
```

```
cd SRR957824
```

View the files in the directory.

```
ls SRR957824
```

View the FASTQ files

```
zcat SRR957834_1.fastq.gz | less
```

Tip: You can use the *space bar* to scroll through the file and *Q* to quit. The suffix of _1 indicates that the file contains data for Read1.

```
@SRR957824.1 1/1
ACTATGAGCGAAACGGCATCCTGGCAGCCGAGCGCATCCATTCCTAACTTATTAAAACGCGCGGCGATTATGGCGGAGAACCCTCGTTTCTTTTCCCATCGTGGAGTGCTGGAGGTGGAGACGCCCTGTATGTGCCAGGCGACGGTAAC
+
?????BBBDDDDDDDBDFFFFFFFFFHHHICEEEEHBHHIIIIIIIFGHHIIIIHFGHHCEEHEEHBEDFEFEFD>B@##########################################################################
@SRR957824.2 2/1
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGCGCTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAAAAAAAATCATACTAATATGAGTTCGATATATAAACTGAAAGATCAAACATGAGTTCATCAGTGTACTATAT
+
????????DDDDDDDDGGGGGGIIIIIIIHIIHIIHHHHHHHHHHIFHHIIHHHHHHHIHGHHIIIFGHDHHEHHE##########################################################################
@SRR957824.3 3/1
TGGCTAACCCGGATACCACTTCCGGCGAAATGTGCGTATTATCCACAGATTCATCGTTTAACACGAATTTTCAAAACGGAACAGCTTATGAGTGCAATCGCGCCTGGAATGATCCTCATCGCGTACCTCTGCGGCTCCATTTCCAGTGC
+
???A?BBBDDDDDDDDGGGGGGIHHHHHHHHIIFHHHHHHHHEHHHIHGHHIIIIIIIHHIIIIIIHHHFHIIIHHFHHHE?DHHHHHHHHHFHFFGGGGGFF=BGGDECCEGGGGGGGGGGGC6<>DEEGG?E>D<>EEEEGGEGCGC:C
@SRR957824.4 4/1
AACGCCAGCATTCAGGGGACCGCAAACCTGGTGAGAACGCTCGGACATCTTAATGGCGACGGGCCGTTACTGGAGCGCCTTCGCTGGAAGAATCTCAATGTACCGCAGAAAATCGCGCTTGCAGCTGTGGGAGCGGGTTATCTGCTTGC
+
,????@?@DDDDDDDDFBEFEFHHHHHHHHHAFGHHFFHHHEHACECHHHHHGHHHHHHDHEEAHEHHHEFFFFFFEEEEEFFEEEEFFA;==EEFBEEEABEEEDD;>?AEC?EFDDDEFFFFEECCE?EFCA;;D;AAEA?:A?::?
@SRR957824.5 5/1
TGGCGATACGACGGAAGTGATAAGTACCCGTGGTCTCAGCCTGTGCTGTATCACATTTGCCTCGAGATGCGTTCAAAGGGGATTGAGCGCCAGATGACCGAAGGGGAATTAAAACGGCTTGCAGAACGGCAACTGACGAAATGGGCCAA
+
???AAAAADDEEEDDDGGGGGGIHIIIIIHHHGFGHIIIIIIIIIIIHHIIIIIIIIIIIIIIIHHEHIIHHHHHIIHHIHHHHHHHHHBHHHHHHGDFGG:@EGEGGGGGGGGGGEGGGGB'?EGC**8?4'48*?EGGE8C8:CE?*8?
@SRR957824.6 6/1
CTGGCGATTGCTGCTCTGGGCGGCGGATTACTATCATCATCCGATTGGCGATGTGCTGTTTCATGCCTTGCCGATTTTACTGCGCCAGGGGCGGCCTGCGGCGAACGCGCCGATGTGGTACTGGTTTGCCACTGAACAAGGCCATGCGC
+
```

Using a similar approach you can view the contents of the Read2 file.

```
zcat SRR957834_2.fastq.gz
```

```
@SRR957824.1 1/2
AAACGTGTCTCAAACGGGACCAAATGAATATCGGTTACCGTCGCCTGGCTCATACAAGGCGTCTCCACCTCCAGCACTCCACGATCGGCAAAGAAACGACGGATCTCCGCCATAATCGCCGCGCGTTTTACTAAGTTAGGAATGGATGC
+
<????B?BBBBBBBBEEEECEHF>CFHHHHGHHBHHHHHEHHHHHHHHHHGHHBGHDGHHHHHHHHHHHHHHFHHHHHHHHHHCFHHDEHEDDFFDF@DE=@EBBCEEAEE6?CECE)?E;????28;?:*:::**0:*0?*:A::AEE
@SRR957824.2 2/2
ATCGGGAGGGGGGCGGGGTGGGGAAGAGGGGAGAATTCGGGGGGGGGCGGGTTATTTAAAAAAAAAAAAAAAAAAAAGAAAATAGAACGTAAAACAAGATGAGATAGACTAAAGTAATCAAGGCTGAGCATATTCTATCTTAGTATAGT
+
##########################################################################################################################################################
@SRR957824.3 3/2
CCACACAGGCGGCAAACCAGAATGGCACTGGAAATGGAGCCGCAGAGGTACGCGATGAGGATCATTCCAGGCGCGATTGCACTCATAAGCTGTTCCGTTTTGAAAATTCGTGTTAAACGATGAATCTGTGGATAATACGCACATTTCGC
+
???????DDDDDDEDFFFFFFFHHFFHHHHIHHGFGHCFHEHHHCCACCGGHDHCHDCGFFFHHGHHHHDFFEHEFFFFFFFFDEEE;DDEFBEDEE=CAEE*=C3BECCB?CBCB:CAA88*:::?E**::AAECCC*?)4;ECAEC?
@SRR957824.4 4/2
ACACCGATTTCATAATGACGAGGTTACGCTCTTTTACATGAGACGGCACCTTGTCATACCAGTTAACCCCGTCATCATCCTCCCCCGCAACACTGCGGTTAAGCGAGCCAAGCAGATAACCCGCTCCCACAGCTGCAAGCGCGATTTTC
+
5????@@DBDDDBDDFFFFFEFEH>FFF>C@EFHHFFHGHHHHGHHHHHHHHEFDFFBFHGFFHG?FFCGHHBHH<CFGGFADGHHHCEHHHED=FFFEEEEE,6=:B@<:==CEEEEC=B0:?>82.'8?E?CA?:CE?##########
@SRR957824.5 5/2
CCGACGGGCGTTTGGGGGCCGCCAGTTGTCGCCGGACTGGCGGAACGCTCAGGCCGTTACTAACATGCTTTGCCCATTTCGTCAGTTGCCGTTCTGCAAGCCGTTTTAATTCCCCTTCGGTCCTCTGGCGCTCAATCCCCTTTGAACGC
+
????BBADDEDDDEEEFFFFDHHHHIIHIHHIHHHEHHHHHHHHFDEEDE)@DDFEEEEE;BE?CEEF?EFCCEECCEFECCE*.1???ADAAEFFEE?*8A>EEEECCEFAACE:?E*8AE'*:?AEEDE884ACCE:E:C?######
@SRR957824.6 6/2
GGCTGTTCAGATCCACCGCATGGCCTTGTTCAGTGGCAAACCAGTACCACATCGGCGCGTTCGCCGCAGGCCGCCCCTGGCGCAGTAAAATCGGCAAGGCATGAAACAGCACATCGCCAATCGGATGATGATAGAAATCCGCCGCCCAC
+
```

Question 1: Why are there two files for the single sample dataset?

The dataset is from a paired-end Illumina experiment. The file with the _1 suffix is for read 1 and the file with the _2 suffix is for read 2 for the sample.

Task 2: Run FASTQC on your dataset

The next step is to access the quality of our raw reads. We will do this by running FASTQC on our dataset.

```
fastqc SRR957824_1.fastq.gz
```

You may also run the analysis in parallel if you have multiple FASTQ files in a single directory.

```
fastqc SRR957824_1.fastq.gz SRR957824_2.fastq.gz
```

or using a wildcard

```
fastqc *.fastq.gz
```

Tip: A tool called MultiQC (https://multiqc.info/) can be used to combine the results from several QC outputs into a single report. This may be useful when working with a large number of samples.

FASTQC can also be run via a graphical user interface, via the terminal or in R Studio (fastqcr package).

Once the analysis is complete, you will notice that new files with a .html extension will be generated. This file contained the information report and images produced by the FASTQC program. We will open the file in the Firefox browser.

```
firefox SRR957824_1_fastqc.html
```

This command should open the html report for read1 in the Firefox browser

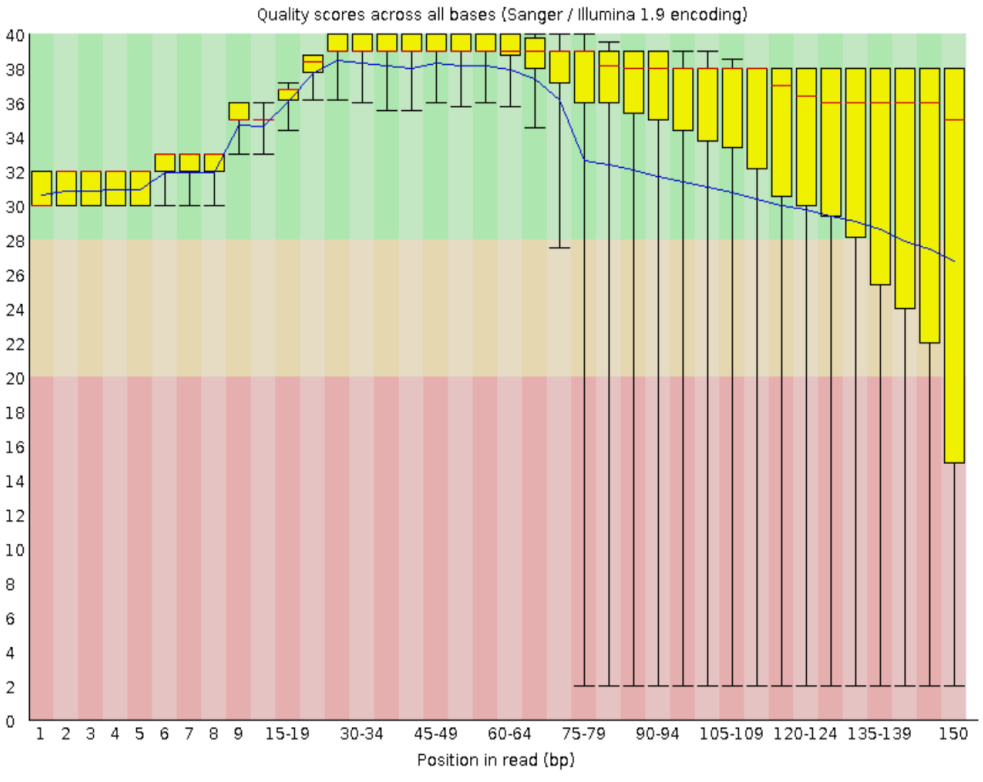
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	EBI SRA_ SRR957824 File_ ftp___ftp_sra_ebi_ac_uk_vol1_fastq_SRR957_SRR957824_SRR957824_1_fastq_gz.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1792335
Sequences flagged as poor quality	0
Sequence length	150
%GC	49

Per base sequence quality



Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

In a similar way open the quality control report for Read2.

```
firefox SRR957824_2_fastqc.html
```

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

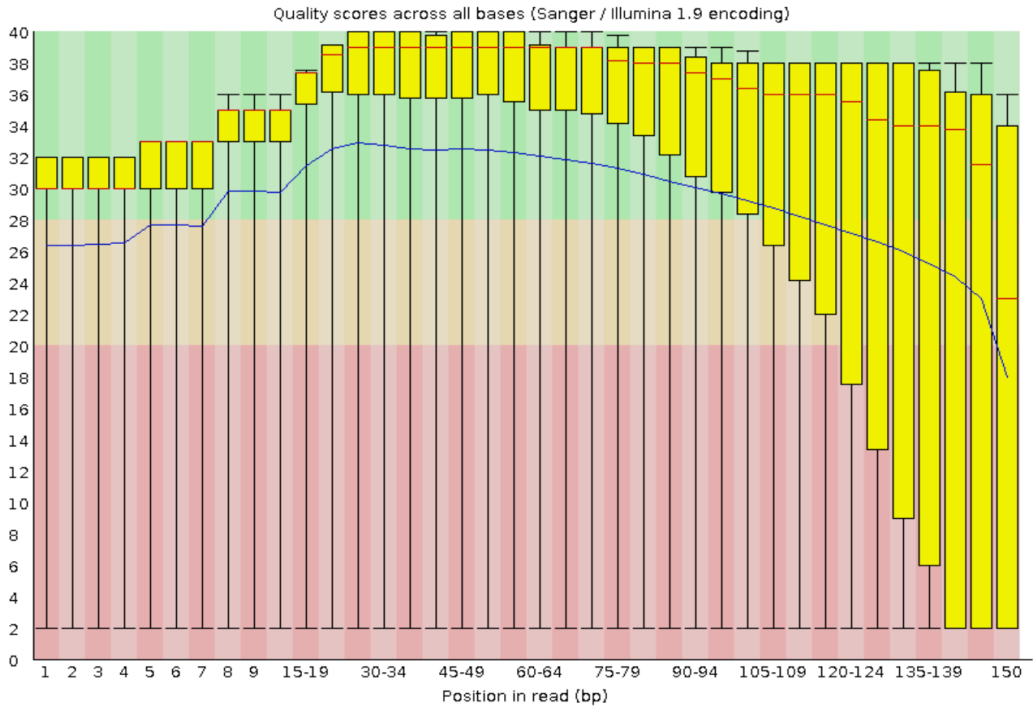
Basic Statistics

Measure	Value
Filename	EBI SRA_ SRR957824 File_ ftp___ftp_sra_ebi_ac_uk_vol1_fastq_SRR957_SRR957824_SRR957824_2_fastq_gz.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1792335
Sequences flagged as poor quality	0
Sequence length	150
%GC	50

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Per base sequence quality



Task 3: Interpretation of QC report

Compare the quality control reports for the two FASTQ files.

Question 2: Comment on the quality of the sequenced reads in file 1 and file 2.

Reads with quite poor quality (<20) towards the end. Reads as outliers with very poor quality for nearly the final third of the read length.
Reads length: reads with poor quality towards the end
Reads number: reads with poor quality

Read through the other reports in the FASTQC output html file and ensure that you understand the information in each report.

Question 3: What can be done to improve the quality of the reads?

The reads can be trimmed using trimming tools such as Trimmomatic or Cutadapt. Only retain reads with high quality base calls.

Trimming

Trimming can be done using a tool called Trimmomatic (https://github.com/usadellab/Trimmomatic)

The current trimming steps are:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
- TOPHRED33: Convert quality scores to Phred-33
- TOPHRED64: Convert quality scores to Phred-64

It works with FASTQ (using phred + 33 or phred + 64 quality scores, depending on the Illumina pipeline used), either uncompressed or gzipp'ed FASTQ. Use of gzip format is determined based on the .gz extension.

Below is an example for using Trimmomatic. Trimmomatic will produce both paired and unpaired datasets from a paired-end set of FASTQ files. The file ILLUMINACLIP:TrueSeq3-PE.fa contains the adapter sequences associated with that sequencing platform.

```
java -jar trimmomatic-0.35.jar PE -phred33 input_forward.fq.gz input_reverse.fq.gz output_forward_paired.fq.gz output_forward_unpaired.fq.gz output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

This will perform the following:

- Remove adapters (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10)
- Remove leading low quality or N bases (below quality 3) (LEADING:3)
- Remove trailing low quality or N bases (below quality 3) (TRAILING:3)
- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15)
- Drop reads below the 36 bases long (MINLEN:36)

Run the command below in order to trim your fastq files. Ensure that you understand what each of the options in the command means.

```
trimmomatic PE SRR957824_1.fastq SRR957824_2.fastq
SRR957824_1_paired.fq.gz SRR957824_1_unpaired.fq.gz
SRR957824_2_paired.fq.gz SRR957824_2_unpaired.fq.gz LEADING:15
TRAILING:15 SLIDINGWINDOW:4:20 MINLEN:100
```

Question 4: Explain what each of the options in the trimmomatic command above specifying. i.e. LEADING, TRAILING, SLIDINGWINDOW, MINLEN.

LEADING:15 – Trim off the leading N bases below Q=15, TRAILING:15 – Trim off the leading N bases below Q=15, SLIDINGWINDOW:4:20 – Using a sliding window of 4 bases, scan the read and trim whenever the average base quality drops below 20, MINLEN:100 – Drop any reads that less than 100 bases long.

If you list your files again you will notice that you now have some new unpaired and paired .fq.gz files. These are your qc'd files that have been trimmed using Trimmomatic in the last step. To see the effect of your trimming on your new files, run FASTQC on the paired output files as you have done previously. Compare the output of the first FASTQC run and the second FASTQC run.

Question 5: Comment on the quality of your two trimmed read files. Have you improved the quality of your data? How many reads have you lost for each file by trimming the data? Please note any other improvements in the dataset.

The quality of both files has been improved. For both FASTQ files we have removed the low-quality base calls towards the end of the reads. For the read2 file we have removed reads with overall low quality. For both files we went from 1 792 335 to 1 283 094 reads.