Basic quality control with FastQC

March 2023

Differential Gene Expression Analysis Workflow

Fastq file format

Fastq file format - Headers

Fastq file format - Sequences

Fastq file format - Third line

Fastq file format - Quality Scores

(Phred) Quality Scores

Sequence quality scores are transformed and translated p-values

- Sequence bases are called after image processing (base calling)
  - Each base in a sequence has a *p-value* associated with it
  - p-values range from 0-1 (e.g.: 0.05, 0.01, 1e-30)
  - p-value of 0.01 inferred as 1 in 100 chance that called base is wrong

(Phred) Quality Scores . . .

How do we assign p-values to bases in the fastq file?

- P-vales can be many characters long (e.g.:0.000005)
- Transform to Phred quality scores Q
- $Q = -10(log_{10}P)$ (e.g.: 0.01 = Q value of 20, 0.001 = Q value of 30)
- Translate Q values to ASCII characters (adding 33) (Q value of 30 = ?, Q value of 40 = I)

QC is important

Check for any problems before we put time and effort into analysing potentially bad data

- Start with FastQC
    - Quick
    - Outputs an easy to read html report

We run fastQC from the terminal with the command

```
fastqc <fastq>
```

but there are lots of other parameters which you can find to tailor your QC by typing

```
fastqc -h
```

Per base sequence quality

Good Data

Bad Data

Per base sequence content

Good Data

Bad Data

Per sequence GC content

Good Data

Bad Data

Adaptor content

Good Data

Bad Data

And now onto the exercise. . .

- Go to: https://ushers.bio.cam.ac.uk/guacamole2
- Log on with YOUR credentials that were emailed to you

A quick intro to the environment

- The terminal is just a text based version of the operating system
- We will look at an example with side by side GUI and text file system. . .
- You use commands instead of mouse clicks - commands are case-senstitve and can be followed by arguments with spaces
    - cd
    - pwd
    - ls
    - flags - e.g. ls -a
    - the directory structure is like a tree, you can go back with cd ..
    - Up arrows to get through history
    - tab complete to avoid errors
    - More to look at the files and q to exit
    - ctrl-c