Babraham Bioinformatics

Exercises: Sequencing QC

Version 2023-04

Licence

Software

The software which will be used in this session is listed below.

- FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- FastqScreen (http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/)
- MultiQC (https://multiqc.info/)

For the exercises today, we will just look at the output reports generated by these programmes. Example code for running these programmes in a Linux environment is given below:

```
fastqc path_to_fastq_file.fq.gz
fastscreen path_to_fastq_file.fq.gz
multiqc path_to_directory_containing_qc_reports
```

For more details/ options see the specific programme documentations

Data

The data in this practical comes from:

- Dataset 1: GSE115964
- Dataset 2: GSE176389
- Dataset 3: GSE135318
- Dataset 4: GSE68618
- Dataset 5: GSE52071
- Dataset 6: GSE81795

*Please note some datasets have been modified for demonstrative purposes

Exercise Background

Exercise Overview

- Review QC reports generated from processing public datasets of various sequencing type.
- Identify any QC issues
- Assess whether datasets can still be used

Important Feature of Different Library Types

In order to address whether there are any QC issues, remember FastQC expects a genomic library however our actual expectations for the dataset will depend on the library type. Below is a quick reminder of some key ways that the libraries we will analyse may differ from a genomic library.

RNA-Seq

- Library is prepared from the transcriptome rather than the genome, therefore will observe a reduced diversity of sequences relative to genomic analysis
- Library may have been prepared using total RNA or RNA depleted of ribosomal RNA
- Library preparation involves reverse transcription, this introduces a preference in the start site of the reads based on the random priming of the reverse transcriptase

ChIP-Seq

- Library are prepared by fragmented DNA:protein complexes of interested are isolated using antibodies, often these targets are associated with promoters which can have a more enriched GC content than the rest of the genome
- Often fragmented DNA not subject to immunoprecipitation will be included for comparison, termed input controls

ATAC-Seq

- Library preparation involves transposases to target accessible regions of DNA, this introduces a preference in the start site of the reads based on their binding

WGBS

- DNA is subject to bisulfite conversion prior to library generation, in this process unmethylated C's are converted to T's

Think about what impact these may have on certain QC metrics

There is also a library-dependent element to assessing the usability of datasets, depending on what we need the data to tell us.

For RNA-Seq, ChIP-Seq and ATAC-Seq:

- Typically just need to know aligned positions to a genome
- So confidence in individual base calls are less important

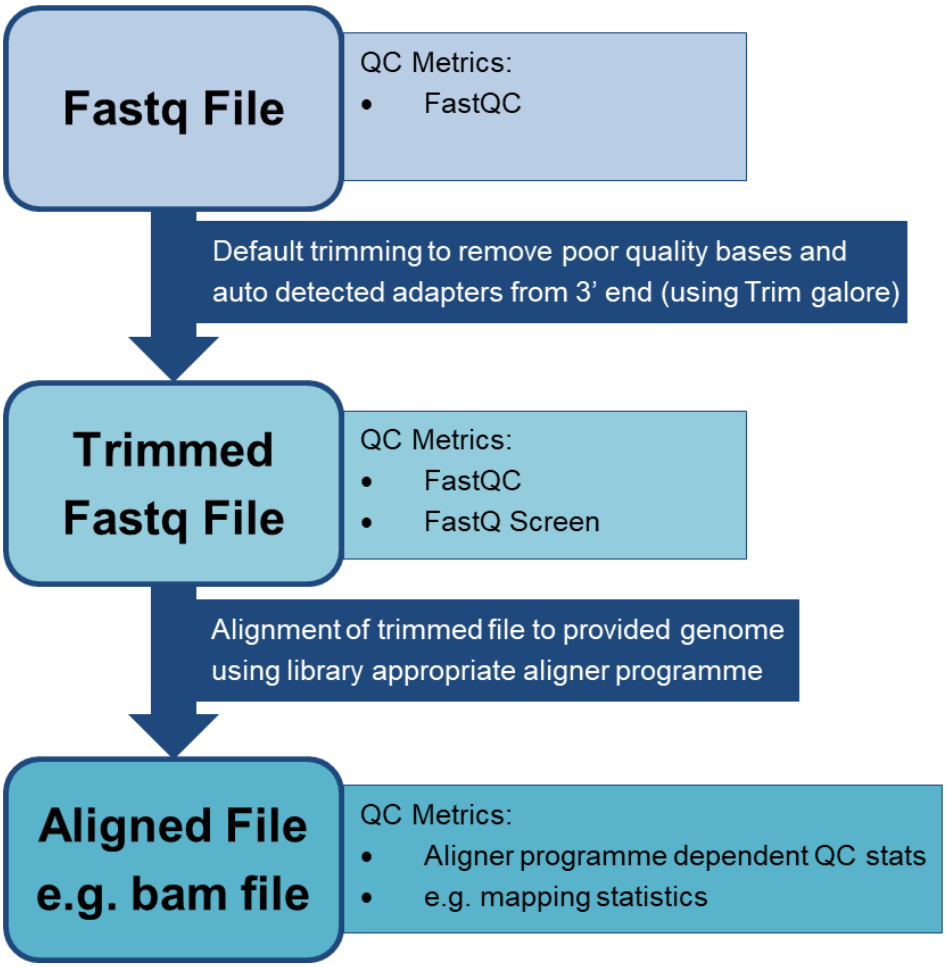
For WGBS

- Normally interested in the proportion of methylated and unmethylated C's
- So confidence in individual base calls is more important

Keep this in mind when considering whether a dataset with QC issues might still be usable

Quality Control in Bioinformatics Processing

You have been provided with a selection of QC reports for different datasets. All the datasets have been processed using a standard pipeline, the exact details of which vary depending on the library type. While the exact processes involved vary depending on the pipeline/ library type the processing can be broadly summarised as follows:



Individual QC files are generated at the stages indicated and the metrics from these are aggregated together in the multiQC report. You are provided with the multiQC report along with the individual FastQC and FastQ Screen html reports for both the trimmed (which will either include "trimmed" or "val_1/2" in their name) and untrimmed Fastq files, should you wish to refer back to these.

Some Analysis Tips:

- It is helpful to start with the multQC report for an overall summary of the data
- You can refer back to individual QC reports for a more focused/ detailed view
    - For instance, not all QC stats can be summarised visually in the multiQC e.g. the "per Tile Sequence QC" is only given in a summary status format.
- The exact QC stats included will vary slightly depending on how the data has been processed and what format it was in, notably:
    - Bisulfite data must be mapped with bismark and associated mapping statistics include detail on the strand alignment. This is beyond the scope of this course so can be ignored. In addition it should also be noted that for simplicity, other bismark specific QC metrics have been excluded from the final multi-QC report.
    - As this is public data and not all datasets have detail on flowcell tile positions therefore the "per Tile Sequence QC" section of the FastQC report is not always available.

Main Exercise

Dataset Details:

Dataset 1

Library strategy: ATAC-Seq
Organism: mouse
Biological replicates: 2 ES-WT and 2 ES-Tet2-KO

Dataset 2

Library strategy: RNA-Seq
Organism: mouse
Biological Replicates: 3 Tet1-WT and 3 Tet1-KO

Dataset 3

Library strategy: ATAC-Seq
Organism: mouse
Biological Replicates: 2 NPC and 3 mESC

Dataset 4

Library strategy: WGBS
Organism: mouse
Biological Replicates: 1 Oldbeta and 1 YoungBeta

Dataset 5

Library strategy: RNA-Seq
Organism: mouse
Biological Replicates: 3 day 0 ESC and 3 day 4 ESC

Dataset 6

Library strategy: ChIP-Seq
Organism: human, drosophila
Biological Replicates: 2 H3K4me1 ChIP and 1 input control per species

For each dataset consider the following:

Given the background of the dataset, is there anything that concerns you?

- Are there any issues with universal QC metrics?
- Are there any issues with library-dependent QC metrics?
    - Is this expected from the library type?
- Do the samples look consistent?

If there is an issue can you diagnose a likely cause from its impact?

- If there are multiple issues which ones are the most fundamental
    - How might the different QC metrics be interlinked
- Broadly, do you think this issue could be due to:
    - Technical error – something wrong with the sequencing run
    - Experimental error – something wrong with the samples that the library was made from
    - Human error – a superficial mistake was made e.g. sample swaps

Finally, would you be happy using this data?

- Consider what needs to understood from the sequencing data and if, even with issues, this might be possible
- If not, could you apply additional processing to perhaps improve it?
    - What might this involve e.g. removing poor quality reads by location?