

# **Exercises: Sequencing QC**

## Software

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

- FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)
- FastQ Screen ([http://www.bioinformatics.babraham.ac.uk/projects/fastq\\_screen/](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
fastq_screen 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:

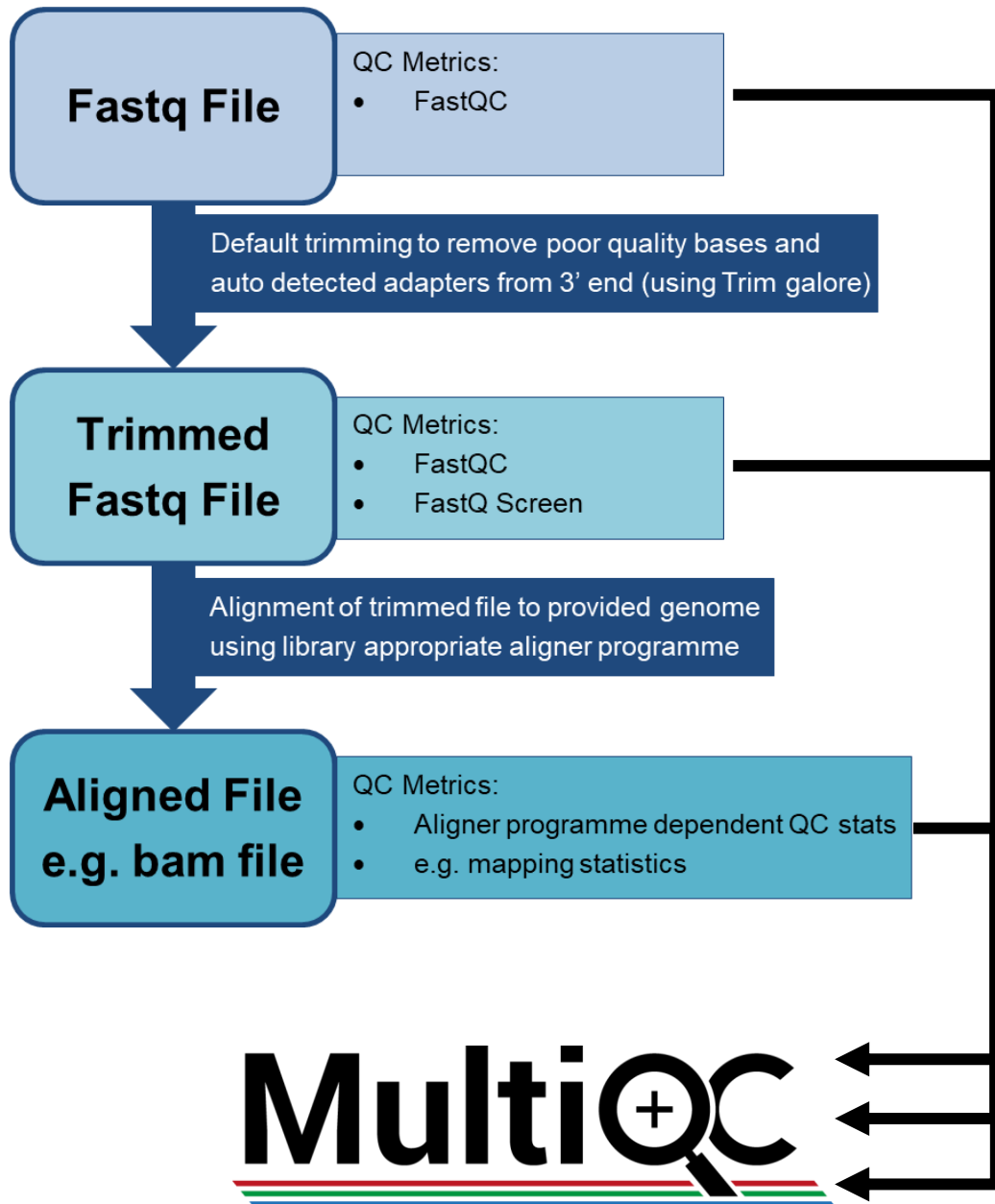
- Dataset 1: GSE176389
- Dataset 2: GSE68618
- Dataset 3: GSE81795
- Dataset 4: GSE52071
- Dataset 5: GSE135318
  
- Demo Dataset: GSE115964

\*Please note some datasets have been modified for demonstrative purposes

## Background

### Summarising Quality Control in Bioinformatics Processing

All the datasets provided, 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:



We will look at the individual and collated QC reports from these pipelines to assess data quality in the following exercises.

## Part 1: Assessing Universal Metrics

### Exercise Overview:

- Review FastQC and FastQ Screen Reports for selected datasets
- For FastQC reports you should **focus only on the following sections**:
  - Per base sequence quality
  - Per tile sequence quality\*
  - Per sequence quality scores
  - Adapter Content

\*Note these datasets are public data and as such do not all have detail on flow-cell tile positions. This means the “per tile sequence quality” section of the FastQC report is not always available. This is the case for Dataset 1.

### Datasets to Review:

#### Dataset 1

- **Library strategy:** RNA-Seq
- **Organism:** mouse

#### Dataset 2

- **Library strategy:** WGBS
- **Organism:** mouse

#### Dataset 3

- **Library strategy:** CHIP-Seq
- **Organism:** human

### Files:

- You have been given the QC reports for 1 sample for each data set – these include:
  - A FastQ Screen report (one per sample)
  - A trimmed FastQC report for each fastq file:
    - There will be 1 FastQC reports if the data is Single End
    - There will be 2 FastQC reports if the data is Paired End

### Look at the reports to answer the following questions:

- Can you see any problems with the datasets?
- Which modules do you see an issue in and what does this suggest about the run?
  - e.g. Technical issues with the sequencer, contamination of the library ect.

### Extra points if you have time:

- Do you think there would be a way to improve any of these sequencing datasets?
- What might this involve?
  - e.g. removing poor quality reads by location

## Part 2: Assessing Library Dependent Metrics

### Exercise Overview:

- Review FastQC Reports for selected datasets
- For FastQC reports you should **focus on the following sections**:
  - Per base sequence content
  - Per sequence GC content
  - Sequence Duplication Levels
  - Overrepresented sequences

### Datasets to Review:

#### Dataset 4

- **Library strategy**: RNA-Seq
- **Organism**: mouse

### Files:

- You have been given the trimmed FastQC report for 2 samples from dataset 4:
  - E14\_D0
  - E14\_D4

### Question:

**One of these samples is normal for an RNA-Seq library, the other is not.  
Which is which?**

- Look at the reports – remember to focus this time on modules related to the library type
  - Though feel free to look at the modules we've already discussed.
- You can also refer to the data expectations and helpful pointers below, to help get you started.

### Dataset Expectations for an RNA-Seq Library:

- Library is prepared from the transcriptome rather than the genome, therefore we 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

### Pointers:

Below are some questions for you to consider while looking at the reports:

- Thinking about the above expectations what modules do you think could flag as problems in FastQC that are related to the nature of the library?
- Can you spot any issues flagged by FastQC, are they in-keeping with your expectations?
- How similar do the samples look – are there any differences?

## Part 3: Putting it All Together with MultiQC

### Exercise Overview:

- Review MultiQC Reports for selected datasets

### Datasets to Review:

#### Dataset 5

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

### Files:

- You have been given the MultiQC report which contains a combined summarised view of the different QC reports for each replicate
  - See Background section (p3) for an overview of the QC/ processing pipeline.

### Dataset Expectations

- Key expectations for an ATAC library are:
  - 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
- Our expectations for mapping are:
  - All samples should map to the mouse genome
- Our need from this data:
  - Typically for ATAC data we are just interested in where our sequences map to in a reference genome – so here individual base calls are less important.

### Look at the MultiQC to answer the following questions:

#### General Impressions

- Can you see the different sections for QC statistics collated from different programmes e.g. Bowtie vs FastQC vs FastQ Screen?
- Can you tell which statistics refer to which sample?
- Can you differentiate statistics from the trimmed and untrimmed fastq files?
  - Hint: trimmed files in this case end “val\_1 or val\_2

#### Reviewing Data-Set Quality

- Can you identify any issues with the data-set?
  - Is the alignment of the samples as expected?
  - Does the data look as expected for an ATAC library?
- Can you suggest the possible cause of any problems with this dataset?

**With all of the above in mind, do you think all (or some) of this library is still usable?**

## Part 3: Extended Exercises (If you have time)

### Additional Exercise Overview:

- Pick one (or more!) of the complete MultiQC reports for the datasets we looked at in Part 1 and Part 2 to review

### Questions to Consider:

- Can you see evidence of the issues we identified in the earlier exercises?
  - If there were problems with only certain samples, can you identify which ones from this view?
  - Are there any cases where you think it would be helpful to refer back the original FastQC or FastQ Screen reports?

### Dataset Details:

If you are unsure of what the different library strategies involve, please check out the “Extended MultiQC Background” section on p8 for a few hints on the libraries involved and some general expectations.

#### Dataset 1

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

#### Dataset 2\*

- **Library strategy:** WGBS
- **Organism:** mouse
- **Biological Replicates:** 1 Old and 1 Young

#### Dataset 3

- **Library strategy:** ChIP-Seq
- **Organism:** drosophila (accidentally mapped to human)
- **Biological Replicates:** 2 H3K4me1 ChIP and 1 input control

#### Dataset 4

- **Library strategy:** RNA-Seq
- **Organism:** mouse
- **Biological Replicates:** 1 D0 and 1 D4

\* Note dataset 2 has been mapped with bismark, this generates:

- Mapping statistics which include detail on the strand alignment
- Additional bismark specific QC metrics
- They are left in so you can check them out, but are beyond the scope of this course

## ***Extended MultiQC Background:***

### **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**

- Typically ChIP-seq libraries are randomly fragmented by sonication – sometimes will still see a little bit of bias at the 5'end due to ligation point of adapter
- 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**