# Introduction to Bulk RNAseq data analysis

QC of Aligned Reads - exercise solutions

# More detailed metrics with Picard Tools

# **Duplication metrics**

#### Exercise 1

1. Run Picard's MarkDuplicates tool on the sorted bam file using the following command:

 $\Rightarrow$  salmon\_qc\_demo/SRR7657883/SRR7657883.salmon.mkdup.bam - The new bam file with duplicated marked

- ⇒ salmon\_qc\_demo/SRR7657883/SRR7657883.salmon.mkdup.bai The index for the new bam file
- ⇒ salmon\_qc\_demo/SRR7657883/

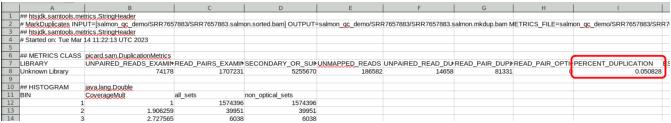
SRR7657883.salmon.mkdup metrics.txt - The duplication metrics

**Note**: The \ at the end of each line tells the terminal that when you press Enter, you have not yet finished typing the command. You can if you wish, type the whole command on a single line, omitting the \ - The command is written across multiple lines here just to make it easier to read.

Q. What is the duplication rate for this bam file? You'll need to look at the metrics file. The easiest way is to open in a spreadsheet. On the course machines we have LibreOffice Calc. You can find this in the launcher bar at the bottom or side of the desktop, e.g.:



You can find details about the contents of the metrics file in the Picard documentation (https://broadinstitute.github.io/picard/picard-metric-definitions.html#DuplicationMetrics).



The duplication rate reported  $\sim$ 5%.

Note that although the column headers for Picard say "PERCENT" or "PCT" the number is in fact the decimal fraction and need to be multiplied by 100 for percent. Just an odd quirk of Picard:

Note: Metrics labeled as percentages (with 'percent' in the full metric name or 'PCT' in the name given in the output file) are actually expressed as fractions. For example, 'PCT\_TARGET\_BASES\_20X = 0.85' should be interpreted as '85 percent of targeted bases are covered to 20X coverage or more!

# RNA alignment metrics

Generate the refFlat file

```
./scripts/create_refflat_from_sam_header.py \
-b salmon_qc_demo/SRR7657883/SRR7657883.salmon.sorted.bam \
-o references/GRCm38_transriptome_refFlat.txt
```

⇒ references/GRCm38\_transriptome\_refFlat.txt

#### Exercise 2

- Run Picard's CollectRnaSeqMetrics tool on the sorted bam file providing the following options:
  - o INPUT The sorted bam file
  - OUTPUT salmon\_qc\_demo/SRR7657883/ SRR7657883.salmon.RNA metrics.txt
  - REF\_FLAT the RefFlat reference file
  - STRAND NONE

```
java -jar picard/picard.jar CollectRnaSeqMetrics \
 INPUT=salmon_qc_demo/SRR7657883/SRR7657883.salmon.sorted.bam \
 OUTPUT=salmon_qc_demo/SRR7657883/SRR7657883.salmon.RNA_metrics.txt \
 REF_FLAT=references/GRCm38_transriptome_refFlat.txt \
 STRAND=NONE \
 VALIDATION_STRINGENCY=SILENT
```

⇒ bam/SRR7657883.chr14.RNA metrics.txt - The RNAseq metrics

The results of this analysis are best viewed graphically, we will do this in the next exercise.

# 3. Visualising QC results with MultiQC

#### Exercise 3

1. Run multiqc on the salmon\_qc\_demo directory:

```
multiqc \
  -n Salmon_QC_Report.html \
  -o salmon_qc_demo \
  salmon_qc_demo
```

- -n a name for the report
- -o the directory in which to place the report
- 2. Open the html report that was generated by multiqc and inspect the QC plots The easiest way to do this is type xdg-open salmon\_qc\_demo/Salmon\_QC\_Report.html, which will open the report in a web browser.

### Exercise 4

In the salmon directory you should find Salmon outputs, duplication metrics and RNAseq metrics for all of the samples from the study.

1. Run multiqc on the contents of the salmon directory.

```
multiqc -z -n Salmon_QC_Report.html -o salmon
```

```
⇒ salmon/Salmon_QC_Report.html (additional_scripts_and_materials/Salmon_QC_Report.html)
```

- Open the html report that was generated by multiqc and inspect the QC plots
- Q. Are there any samples that look problematic?

SRR7657893 has low alignment rate, an insert size profile that is skewed to left with a median at ~180 bp and a transcript coverage profile that shows a strong 3' bias. This suggests that the RNA in the this sample has been degraded. NOTE: This sample is not real - we have mocked up the metrics files for the purpose of illustrating a poor quality data set.