Introduction to Bulk RNAseq data analysis

QC of Aligned Reads - exercise solutions

Contents

Duplication metrics

Exercise 1

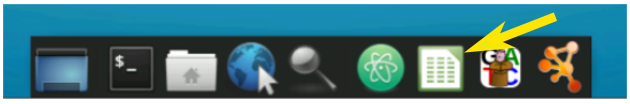
1. Run Picard's MarkDuplicates tool on the sorted bam file using the following command:

```
java -jar picard/picard.jar MarkDuplicates \
        INPUT=salmon_qc_demo/SRR7657883/SRR7657883.salmon.sorted.bam \
        OUTPUT=salmon_qc_demo/SRR7657883/SRR7657883.salmon.mkdup.bam \
        METRICS_FILE=salmon_qc_demo/SRR7657883/SRR7657883.mkdup_metrics.txt \
        CREATE_INDEX=true \
        VALIDATION_STRINGENCY=SILENT
```

⇒ *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.

	A	B	C	D	E	F	G	H	I
1	## htsjdk.samtools.metrics.StringHeader								
2	# MarkDuplicates INPUT=[salmon_qc_demo/SRR7657883/SRR7657883.salmon.sorted.bam] OUTPUT=salmon_qc_demo/SRR7657883/SRR7657883.salmon.mkdup.bam METRICS_FILE=salmon_qc_demo/SRR7657883/SRR7								
3	## htsjdk.samtools.metrics.StringHeader								
4	# Started on: Tue Mar 14 11:22:13 UTC 2023								
5									
6	## METRICS CLASS	picard.sam.DuplicationMetrics							
7	LIBRARY	UNPAIRED_READS_EXAMIN	READ_PAIRS_EXAMIN	SECONDARY_OR_SUP	UNMAPPED_READS	UNPAIRED_READ_DU	READ_PAIR_DUPI	READ_PAIR_OPTI	PERCENT_DUPLICATION
8	Unknown Library	74178	1707231	5255670	186582	14658	81331	0	0.050828
9									
10	## HISTOGRAM	java.lang.Double							
11	BIN	CoverageMult	all_sets	non_optical_sets					
12	1	1	1574396	1574396					
13	2	1.906259	39951	39951					
14	3	2.727565	6038	6038					

The duplication rate reported ~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

1. Run Picard's `CollectRnaSeqMetrics` tool on the sorted bam file providing the following options:
   - 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
```

⇒ *salmon_qc_demo/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/Salmon_QC_Report.html

2. 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.