

# RNA-seq analysis in R

## QC of Aligned Reads - exercise solutions

### 1. A quick look at the alignment metrics with samtools

#### Exercise 1

1. Use the `samtools flagstat` command to generate alignment metrics for the sorted bam file you created in the previous practical.

```
samtools flagstat bam/SRR7657883.sorted.bam
```

```
64434897 + 0 in total (QC-passed reads + QC-failed reads)
4193507 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
61244035 + 0 mapped (95.05% : N/A)
60241390 + 0 paired in sequencing
30120695 + 0 read1
30120695 + 0 read2
54707584 + 0 properly paired (90.81% : N/A)
55067498 + 0 with itself and mate mapped
1983030 + 0 singletons (3.29% : N/A)
161370 + 0 with mate mapped to a different chr
139571 + 0 with mate mapped to a different chr (mapQ>=5)
```

Q) What percentage of the reads have aligned to the genome?

95.05% of reads have been mapped to the reference genome.

### 2. More detailed metrics with Picard Tools

#### 2.1 Duplication metrics

##### Exercise 2.1

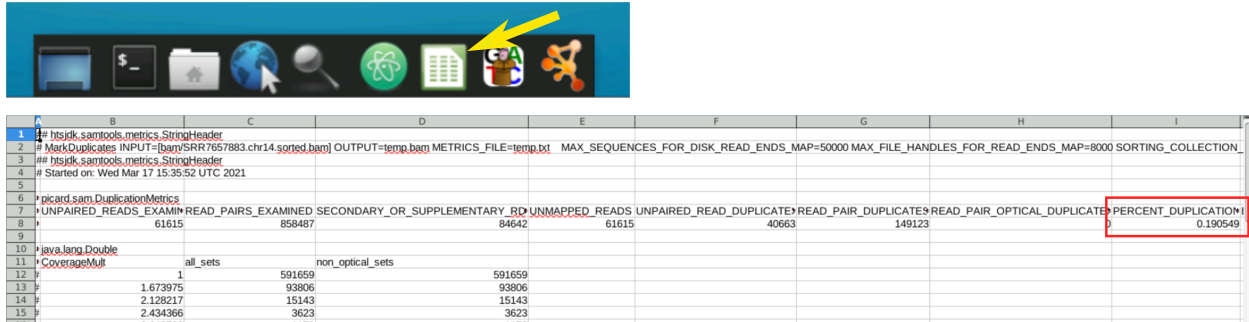
1. Run Picard's MarkDuplicates tool on the sorted bam file using the following command:

```
java -jar picard/picard.jar MarkDuplicates \
  INPUT=bam/SRR7657883.chr14.sorted.bam \
  OUTPUT=bam/SRR7657883.chr14.mkdup.bam \
  METRICS_FILE=bam/SRR7657883.chr14.mkdup_metrics.txt \
  CREATE_INDEX=true
```

⇒ *bam/SRR7657883.chr14.mkdup.bam* - The new bam file with duplicated marked  
⇒ *bam/SRR7657883.chr14.mkdup.bai* - The index for the new bam file  
⇒ *bam/SRR7657883.chr14.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 of the desktop.



~19%. 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.

## 2.2 Alignment metrics

### Exercise 2.2

- Run Picard's `CollectAlignmentSummaryMetrics` tool on the chr14 sorted bam providing the following options.
  - INPUT - The sorted chr 14 only bam file
  - OUTPUT - `bam/SRR7657883.chr14.alignment_metrics.txt`
  - REFERENCE\_SEQUENCE - `references/Mus_musculus.GRCm38.dna_sm.primary_assembly.fa`

```
java -jar picard/picard.jar CollectAlignmentSummaryMetrics \
INPUT=bam/SRR7657883.chr14.sorted.bam \
OUTPUT=bam/SRR7657883.chr14.alignment_metrics.txt \
REFERENCE_SEQUENCE=references/Mus_musculus.GRCm38.dna_sm.primary_assembly.fa
```

⇒ `bam/SRR7657883.chr14.alignment_metrics.txt` - The alignment metrics

## 2.3 Insert Size metrics

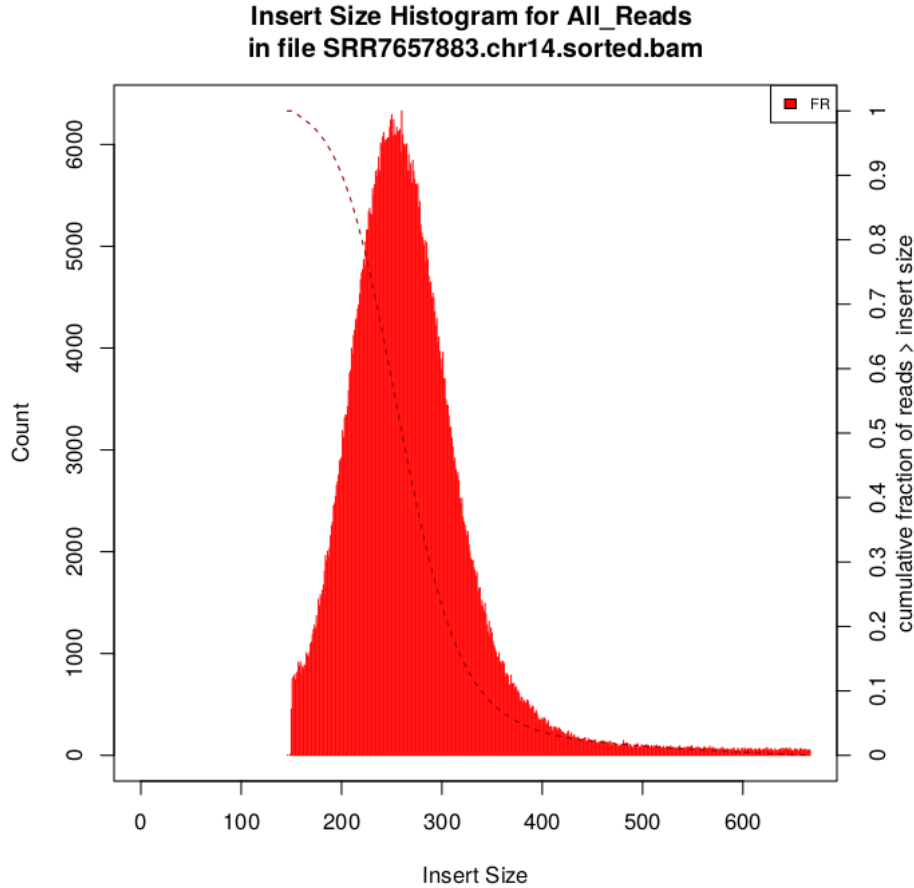
### Exercise 2.3

- Run Picard's `CollectInsertSizeMetrics` tool on the chr1 sorted bam providing the following options.
  - INPUT - The sorted chr 1 only bam file
  - OUTPUT - `bam/SRR7657883.chr1.insert_size.txt`
  - HISTOGRAM\_FILE - `bam/SRR7657883.chr1.insert_size.pdf`

```
java -jar picard/picard.jar CollectInsertSizeMetrics \
INPUT=bam/SRR7657883.chr14.sorted.bam \
OUTPUT=bam/SRR7657883.chr14.insert_size.txt \
HISTOGRAM_FILE=bam/SRR7657883.chr14.insert_size.pdf
```

- ⇒ *bam/SRR7657883.chr14.insert\_size.txt* - The insert size metrics
- ⇒ *bam/SRR7657883.chr14.insert\_size.pdf* - The PDF with a plot showing the insert size distribution

Open the PDF and look at the distribution fragment lengths (insert sizes) in the library.



Q. Considering this data is from paired 150 base reads, what are the implications of the fragment length distributions.

As we have PE 150 reads, the total amount of sequencing from each fragment is 300 bases. Looking at the distribution only ~25% of the fragments have lengths greater than 300 bases. This means that for about 80% of the fragments the reads are overlapping. From the perspective of our gene expression analysis, this doesn't matter, however, from a design perspective it means that we have unnecessarily sequenced a lot of bases twice and the more sequencing we carry out, the more expensive the study is. It would have been optimal to use a shorter read length.

## 2.4 RNA alignment metrics

### Exercise 2.4

1. Run Picard's `CollectRnaSeqMetrics` tool on the sorted bam file providing the following options:
  - INPUT - The sorted bam file
  - OUTPUT - `bam/SRR7657883.chr14.RNA_metrics.txt`
  - REF\_FLAT - the RefFlat reference file

- STRAND - NONE

```
java -jar picard/picard.jar CollectRnaSeqMetrics \  
INPUT=bam/SRR7657883.chr14.sorted.bam \  
REF_FLAT=references/Mus_musculus.GRCm38.102.txt \  
OUTPUT=bam/SRR7657883.chr14.RNA_metrics.txt \  
STRAND=NONE
```

⇒ *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

1. Run multiqc on the bam directory:

```
multiqc -n Alignment_QC_Report.html -o bam bam
```

- `-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 multiqc_report.html`, which will open the report in a web browser.

#### Exercise 3.2

In the `metrics` directory you should find Picard metrics for all of the samples.

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

```
multiqc -z -n Alignment_QC_Report.html -o metrics metrics
```

⇒ *metrics/Alignment\_QC\_Report.html*

2. Open the html report that was generated by multiqc and inspect the QC plots

Q. Are there any bam files 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.