Introduction to Bulk RNAseq data analysis

Quantification of Gene Expression with Salmon

Exercise 1 - Create Salmon index

1. Create concatenated trancriptome/genome reference file

```
cat references/Mus_musculus.GRCm38.cdna.chr14.fa.gz \
    references/Mus_musculus.GRCm38.dna_sm.chr14.fa.gz \
    > references/gentrome.chr14.fa.gz
```

2. Create decoy sequence list from the genomic fasta

```
echo "14" > references/decoys.txt
```

3. Use `salmon index` to create the index. You will need to provide three pieces of information:
    - the **Transcript fasta file** - *references/gentrome.chr14.fa.gz*
    - the **decoys** - *references/decoys.txt*
    - the **salmon index** - a directory to write the index to, use **references/salmon_index_chr14**

Also add `-p 7` to the command to instruct salmon to use 7 threads/cores. To find the flags for the other three pieces of information use:

```
salmon index --help
```

```
Version Info: This is the most recent version of salmon.

Index
==========
Creates a salmon index.

Command Line Options:
  -v [ --version ]              print version string
  -h [ --help ]                 produce help message
  -t [ --transcripts ] arg  Transcript fasta file.
  -k [ --kmerLen ] arg (=31)    The size of k-mers that should be used for the
                                quasi index.
  -i [ --index ] arg        salmon index.
  --gencode                     This flag will expect the input transcript
  ...
  ...
  ...
  -d [ --decoys ] arg       Treat these sequences ids from the reference as
                                the decoys that may have sequence homologous to
                                some known transcript. for example in case of
                                the genome, provide a list of chromosome name
                                --- one per line
```

```
salmon index \
    -t references/gentrome.chr14.fa.gz \
    -d references/decoys.txt \
    -p 7 \
    -i references/salmon_index_chr14
```

Exercise 2 - Quantify with Salmon

1. Make directory called `salmon_output`

```
mkdir salmon_output
```

2. Use `salmon quant` to quantify the gene expression from the raw fastq. To see all the options run `salmon quant --help-reads`. There are lot of possible parameters, we will need to provide the following:
    - **salmon index** - *references/salmon_index*
    - `-l A` - Salmon needs to use some information about the library preparation, we could explicitly give this, but it is easy enough for Salmon to **A**utomatically infer this from the data.
    - **File containing the #1 mates** - *fastq/SRR7657883.sra_1.fastq.gz*
    - **File containing the #2 mates** - *fastq/SRR7657883.sra_2.fastq.gz*
    - **Output quantification directory** - *salmon_output/SRR7657883*
    - `--gcBias` - salmon can optionally correct for GC content bias, it is recommended to always use this
    - **The number of threads to use** - *7*

```
salmon quant \
    -p 7 \
    -i references/salmon_index \
    --gcBias \
    -l A \
    -1 fastq/SRR7657883.sra_1.fastq.gz \
    -2 fastq/SRR7657883.sra_2.fastq.gz \
    -o salmon_output/SRR7657883
```

Exercise 3 - Run multiqc

1. Run `multiqc` on the `salmon_output` directory and create a report called `Salmon_quantification_report` in the `salmon_output` directory. Look back to the previous session if you need a reminder on how to do this.

```
multiqc -z -n Salmon_quantification_report -o salmon_output salmon_output
```

⇒ salmon_output/Salmon_quantification_report.html

2. Open the report and determine what percentage of the reads have been *aligned* to the transcriptome. Compare this to the QC report from the previous session.

With Salmon we have ~85% of reads mapped to the transcriptome.

If you look at the RNAseq metrics section in the multiqc report we generated in the previous section, you should see that for SRR7657883 we had ~85% reads aligned to Coding + UTR.