

# Introduction to Bulk RNAseq data analysis

## Quantification of Gene Expression with Salmon

### Contents

Exercise 1 - Create Salmon index . . . . .	1
Exercise 2 - Quantify with Salmon . . . . .	2

### Exercise 1 - Create Salmon index

1. Create concatenated transcriptome/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:

<code>-v [ --version ]</code>	print version string
<code>-h [ --help ]</code>	produce help message
<code>-t [ --transcripts ] arg</code>	Transcript fasta file.
<code>-k [ --kmerLen ] arg (=31)</code>	The size of k-mers that should be used for the quasi index.
<code>-i [ --index ] arg</code>	salmon index.
<code>--gencode</code>	This flag will expect the input transcript
<code>...</code>	
<code>...</code>	
<code>...</code>	
<code>-d [ --decoys ] arg</code>	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. There should already be a directory called `salmon_output` in the `Course_materials` directory. If not, create it.

```
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 Automatically infer this from the data.
  - **File containing the #1 mates** - *fastq/SRR7657883.subset\_2M.sra\_1.fastq.gz*
  - **File containing the #2 mates** - *fastq/SRR7657883.subset\_2M.sra\_2.fastq.gz*
  - **Output quantification directory** - *salmon\_output/SRR7657883*
  - **--writeMappingssalmon\_output/SRR7657883.salmon.sam** - Instructs Salmon to output the read alignments in SAM format to the file *salmon\_output/SRR7657883.salmon.sam*.
  - **--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 \  
-i references/salmon_index \  
-l A \  
-1 fastq/SRR7657883.subset_2M.sra_1.fastq.gz \  
-2 fastq/SRR7657883.subset_2M.sra_2.fastq.gz \  
-o salmon_output/SRR7657883 \  
--writeMappings=salmon_output/SRR7657883/SRR7657883.salmon.sam \  
--gcBias \  
-p 7
```

## Exercise 3

1. Sort and transform your aligned SAM file into a BAM file called `SRR7657883.salmon.sorted.bam`. Use the option `-@ 7` to use 7 cores, this vastly speeds up the compression.

```
samtools sort \  
-@ 7 \  
-O BAM \  
-o salmon_output/SRR7657883/SRR7657883.salmon.sorted.bam \  
salmon_output/SRR7657883/SRR7657883.salmon.sam
```

⇒ *salmon\_output/SRR7657883/SRR7657883.salmon.sorted.bam*

2. Check your bam file

```
samtools view salmon_output/SRR7657883/SRR7657883.salmon.sorted.bam | more
```