BIN508 Assignment 4

Due date: 7 May, Wednesday, 13:30 **Cut-off:** 7 May, Wednesday, 19:30

Late policy: After the due time, 5 points of deduction will be applied for each extra hour.

- **Use EU servers instead of USA servers for this assignment.**
- **You need to install IGV on your local machine.**
- **Unless you are very experienced in Bioconductor, Conda/Mamba, and R, you should use Galaxy, as installation of and parameter tuning for DESeq2 can be really tricky.**

- This tutorial (Reference-based RNA-Seq data analysis) will be completed to answer questions 2 and 3.

1) Briefly describe RPKM, FPKM, and TPM. Which metric is more appropriate for RNA-Seq analysis?

2) The first part of the tutorial: *Aligning the reads, and counting the reads*. Load the '`.fastqsanger`' files given to you in the tutorial: See "Hands-on: Data upload".

- Please add screenshots that help with your explanations. You don't have to show every step.

a) FastQC/MultiQC: Check the read quality and show what problems there are.

b) RNA STAR: Map your reads. Share details about the mapping results

c) IGV: Inspect your reads on IGV as shown in the tutorial. Compared to the previous assignment's visuals, now you will see lines connecting the reads. What do these lines represent?

d) IGV: Right-click on the name of your data and select "Sashimi plot". Share a screenshot and explain what the lines and the numbers represent.

e) Infer Experiment: Briefly explain what strandedness is and use "Infer Experiment" (you can skip "Estimate strandness with IGV for a paired-end library" and "Estimate strandness with STAR") on your data. Share and explain the output.

f) featureCounts: Count the assigned reads. What are your observations about read counts? How are the assigned reads distributed amongst the genome?

3) The second part of the tutorial: *Analysis of differential gene expression, functional enrichment analysis*. Create a new history and load the '`.counts`' files given to you in the tutorial (These are also called 'count matrices'): See "Hands-on: Import all count files".

a) DESeq2: Run DESeq2 for the Differential Gene Expression (DGE) analysis to estimate the biological variance between each condition. Submit your PCA plot, distance matrix, and MA plot, and briefly explain what they represent.

b) Table editing/math/heatmap2: What are heatmap and Z-Score? Explain how they can be used to understand the overall DEG profile in the experiment. Add a screenshot of your heatmap.

c) Table editing: Use FILTER to select significant DEGs (genes with an adjusted p-value below 0.05 and abs(log2FC)>1). Add a screenshot of the resulting table. What does this table contain?

d) goseq: Submit the GO and KEGG Analysis results, and explain what the plots represent. Which biological functions are enriched?

4) Briefly answer the following questions.

a) Why is normalisation important in RNASeq data analysis?

b) What are “differential expression analysis” and “functional analysis” in RNASeq data analysis? What data is given as the inputs and taken as the outputs of those steps?

c) Is it a good practice to keep the overrepresented sequences, and not remove them, in RNASeq? Why?

5) SNPnexus is a web-based variant annotation tool designed to simplify and assist in selecting and prioritising known and novel genomic alterations. Visit their website here. Check the video tutorial. Using the variant file given to you (`sample.vcf`), run your analysis to answer the following questions:

a) How many variants are listed in the .vcf from the Ensembl database?

b) How many exonic variants are present in the list?

c) Compare the number of deleterious/damaging variants annotated with SIFT vs. Polyphen.