

# HW3: Transcriptome Analysis

## Homework 3 - to be done as groups

Names:

Group:

For deadlines etc., see Absalon.

You have to supply both the answer (whatever it is: numbers, a table, plots or combinations thereof), as well as the R or Linux code you used to make the plots. This should be done using this R markdown template: we want both the R markdown file and a resulting PDF. For PDF output, you may have to install some extra programs - RStudio will tell you.

Note that:

1. If the R code gives different results than your results, you will get severe point reductions or even 0 points for the exercise
2. Some questions may request you to use R options we have not covered explicitly in the course: this is part of the challenge
3. While this is a group work, we expect that everyone in the group will have understood the group solution: similar or harder question might show up in the individual homework. So, if something is hard, it means you need to spend more time on it
4. The results should be presented on a level of detail that someone else could replicate the analysis.

For statistical tests, you have to:

- 1) Motivate the choice of test
- 2) State exactly what the null hypothesis is (depends on test!)
- 3) Comment the outcome: do you reject the null hypothesis or not, and what does this mean for the actual question we wanted to answer (interpretation)?

## Intro

As you already know how to quantify RNAseq data (see the quantification exercise from RNAseq lecture) this homework is about the downstream analysis of such quantifications.

Please use `knitr::kable()` to produce nicely formatted tables when you are asked to provide a table.

## Part1: Data analysis and clustering

Use the supplied subset of Salmon quantifications stored in the “salmon\_result\_part1.zip” folder. These files contain the Salmon quantifications of 6 samples – 3 biological replicates of non-treated cells (WT) and 3 biological replicates of cells treated with a cancer promoting drug called TPA (WTTPA). Salmon was run with the `-seqBias` option.

### Question 1.1

Read the “quant.sf” file from the Salmon result folder for WT1 into R with `read_tsv()`. Plot the isoform lengths versus the effective lengths as a scatter plot, add a smoothed line and a dashed line along the diagonal.

Scale both axis using log10-scaling in ggplot2. Comment on the comparison of the differences between the trend line and the diagonal line with respect to what is expected. **Use max 100 words.**

### Question 1.2

Analyze and comment on the strange outliers in the plot from Question 1.1. **Use max 100 words.**

### Question 1.3

Use IsoformSwitchAnalyzeR's `importIsoformExpression()` to import all the data into R. Convert the abundancies imported by `importIsoformExpression()` into a log2 transformed abundance matrix (using a pseudocount of 1) where columns are samples and isoform ids are stored as rownames. Report the first 4 rows as a table and discuss the advantage of a pseudocount of 1. **Use max 100 words.**

### Question 1.4

Use tidyverse to extract the 100 most variable isoforms (aka those with highest variance) from the log2-transformed expression matrix. Provide a table with top five most variable isoforms.

### Question 1.5

Use the pheatmap package to make one visually appealing heatmap of the isoforms from 1.4 and comment on the result. Columns should be samples and rows isoforms. Furthermore, discuss pros and cons of the argument `scale = "row"` vs `scale = "none"`. **Use max 100 words.**

## Part2: Isoform switch analysis with IsoformSwitchAnalyzeR

Use the supplied Salmon quantification subsets stored in the “salmon\_result\_part2.zip” folder (Different than the one you used in part 1!). These files contain the Salmon quantifications of 6 samples – 3 biological replicates of wildtype (WT) and 3 biological replicates of a knock out (KO) of a suspected splice factor – let us call it *factor X* for the sake of drama. Salmon was run with the `-seqBias` option.

Please note that you need IsoformSwitchAnalyzeR version > 1.1.10. You might need to update it first.

### Question 2.1

Use the `importIsoformExpression` and `importRdata(addAnnotatedORFs=FALSE)` functions to create a `switchAnalyzeRList` object from the Salmon output supplied in the “salmon\_result\_part2.zip” folder. Use the GTF file also included in the zip file. Report the summary statistics of the resulting `switchAnalyzeRList`. What does the `addAnnotatedORFs=FALSE` argument do and why do you think it is enabled here?

### Question 2.2

Why is it essential the annotation stored in the GTF file is the exact annotation quantified with Salmon (in the context of IsoformSwitchAnalyzeR functionalities)? **Use max 100 words.**

### Question 2.3

Load the supplied “switchList.Rdata” object into R with the `readRDS()` function. This is the result of running the whole IsoformSwitchAnalyzeR workflow on the full dataset. Make a table with the Top 10 switching genes with predicted consequences when sorting on q-values.

### Question 2.4

Show code for how to produce switchPlot for these 10 genes and save them to your own computer. The plots should not be included in the report (only the code for how to produce it)!

### Question 2.5

Which of the top 10 genes with switches do you think is the most important? Include/produce the switchPlot for that particular gene in the report and provide references if necessary. **Use max 100 words.**

### Question 2.6

Plot the global enrichment of switch consequences and alternative splicing and comment on it. What are the general patterns? **Use max 100 words.**