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 anlaysis 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 og 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?

```r
library(IsoformSwitchAnalyzeR)

packageVersion("IsoformSwitchAnalyzeR")
```

```
## [1] '1.2.0'
```

```r
R.Version()
```

```
## $platform
## [1] "x86_64-apple-darwin15.6.0"
##
## $arch
## [1] "x86_64"
##
## $os
## [1] "darwin15.6.0"
##
## $system
## [1] "x86_64, darwin15.6.0"
##
## $status
## [1] ""
##
## $major
## [1] "3"
##
## $minor
## [1] "5.0"
##
## $year
## [1] "2018"
##
## $month
## [1] "04"
##
## $day
## [1] "23"
##
## $`svn rev`
## [1] "74626"
##
## $language
## [1] "R"
##
## $version.string
## [1] "R version 3.5.0 (2018-04-23)"
##
## $nickname
## [1] "Joy in Playing"
```

```r
?importIsoformExpression()

IsoformList1 <-  importIsoformExpression(parentDir = "./salmon_result_part2")
```

```
## Step 1 of 3: Identifying which algorithm was used...
```

```
##     The quantification algorithm used was: Salmon
```

```
## Step 2 of 3: Reading data...
```

```
## reading in files with read_tsv
```

```
## 1 2 3 4 5 6
## Step 2 of 3: Normalizing TxPM values via edgeR...
## Removing 2433 rows with all zero counts
## Done
```

```
head(data.frame(IsoformList1$counts))
```

```
##                      isoform_id      KO1      KO2 KO3          WT1       WT2
## TCONS_00000001 TCONS_00000001 1.003786 2.890696   0 4.400760e+00  0.000000
## TCONS_00000002 TCONS_00000002 0.000000 0.000000   0 0.000000e+00  0.000000
## TCONS_00000003 TCONS_00000003 0.000000 0.000000   0 1.069622e+01  2.199315
## TCONS_00003946 TCONS_00003946 0.000000 0.000000   0 0.000000e+00  0.000000
## TCONS_00003947 TCONS_00003947 0.000000 8.972704   0 1.398578e+02 81.796689
## TCONS_00003948 TCONS_00003948 0.000000 0.000000   0 1.196270e-06  0.000000
##                      WT3
## TCONS_00000001  0.000000
## TCONS_00000002  0.000000
## TCONS_00000003  1.378978
## TCONS_00003946 14.541497
## TCONS_00003947 32.194612
## TCONS_00003948  0.000000
```

```
repCount1 <- IsoformList1$counts[,2:7]
#colnames(repCount1) <- c('isoform_id','KO1','KO2','KO3','WT1','WT2','WT3')

#rownames(repCount1)

repCount1$isoform_id <- rownames(repCount1)


?data.matrix
#head(IsoformList1$counts,4)

OurDesignMatrix <- data.frame(sampleID = c('KO1','KO2','KO3','WT1','WT2','WT3'),condition=c('KO','KO','

colnames(OurDesignMatrix) <- c('sampleID','condition')


My.SwitchAnalyzerRList <-  importRdata(isoformCountMatrix = repCount1, addAnnotatedORFs=FALSE, isoformE
```

```
## Step 1 of 5: Obtaining annotation...
## importing GTF (this may take a while)
## Step 2 of 5: Calculating gene expression...
```

```
## 
  |                                                                       
  |                                                                 |   0%
  |                                                                       
  |=============                                                    |  20%
  |                                                                       
  |==========================                                       |  40%
  |                                                                       
  |=======================================                          |  60%
  |                                                                       
  |====================================================             |  80%
  |                                                                       
  |=================================================================| 100%
```

```
## Step 3 of 5: Merging gene and isoform expression...
```

```
##
  |
  |                                                                 |   0%
  |
  |================================                                 |  50%
  |
  |=================================================================| 100%

## Step 4 of 5: Making comparisons...

##
  |
  |                                                                 |   0%
  |
  |=================================================================| 100%

## Step 5 of 5: Making switchAnalyzeRlist object...
## Done
```

```r
head(My.SwitchAnalyzerRList$isoformFeatures,2)
```

```
##            iso_ref          gene_ref     isoform_id     gene_id
## 1 isoComp_00000001 geneComp_00000001 TCONS_00000001 XLOC_000001
## 2 isoComp_00000002 geneComp_00000002 TCONS_00000002 XLOC_000002
##   condition_1 condition_2 gene_name gene_overall_mean gene_value_1
## 1          KO          WT   Gm16088           1.44736      1.60474
## 2          KO          WT        U6           0.00000      0.00000
##   gene_value_2 gene_stderr_1 gene_stderr_2 gene_log2_fold_change
## 1     1.289981     0.9751635      1.289981            -0.3128111
## 2     0.000000     0.0000000      0.000000             0.0000000
##   gene_q_value iso_overall_mean iso_value_1 iso_value_2 iso_stderr_1
## 1           NA          1.44736     1.60474    1.289981    0.9751635
## 2           NA          0.00000     0.00000    0.000000    0.0000000
##   iso_stderr_2 iso_log2_fold_change iso_q_value IF_overall IF1 IF2 dIF
## 1     1.289981           -0.3128111          NA          1   1   1   0
## 2     0.000000            0.0000000          NA        NaN NaN NaN NaN
##   isoform_switch_q_value gene_switch_q_value
## 1                     NA                  NA
## 2                     NA                  NA
```

```r
#?importRdata

#cuffDB <- prepareCuffExample()

#designMatrix <- cummeRbund::replicates(IsoformList1$counts)[,c('rep_name','sample_name')]

#isoRepCount <- repCountMatrix(isoforms(cuffDB))
```

The addAnnotatedORFs will add annotated open reading frames, but since our data is not a quantification of known annotated transcripts, we can not use this option.

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.**

It is essential that the GTF file is the exact genomic annotation that was quantified with Salmon, since it otherwise won't be possible to match exactly which isoforms match the same genes.

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.

```
My.SwitchList <- readRDS("./hw3switchList.Rdata")


My.SwitchList.select <-  select(as.tibble(My.SwitchList$isoformFeatures), iso_ref, gene_ref, isoform_id
group_by(gene_name) %>%
summarise(ene_q_value = min(gene_q_value), gene_switch_q_value = min(gene_switch_q_value)) %>%
arrange(gene_switch_q_value)

knitr::kable(My.SwitchList.select[1:10,],digits=100, caption="2.3 \n Sorted genes of lowest gene switch
```

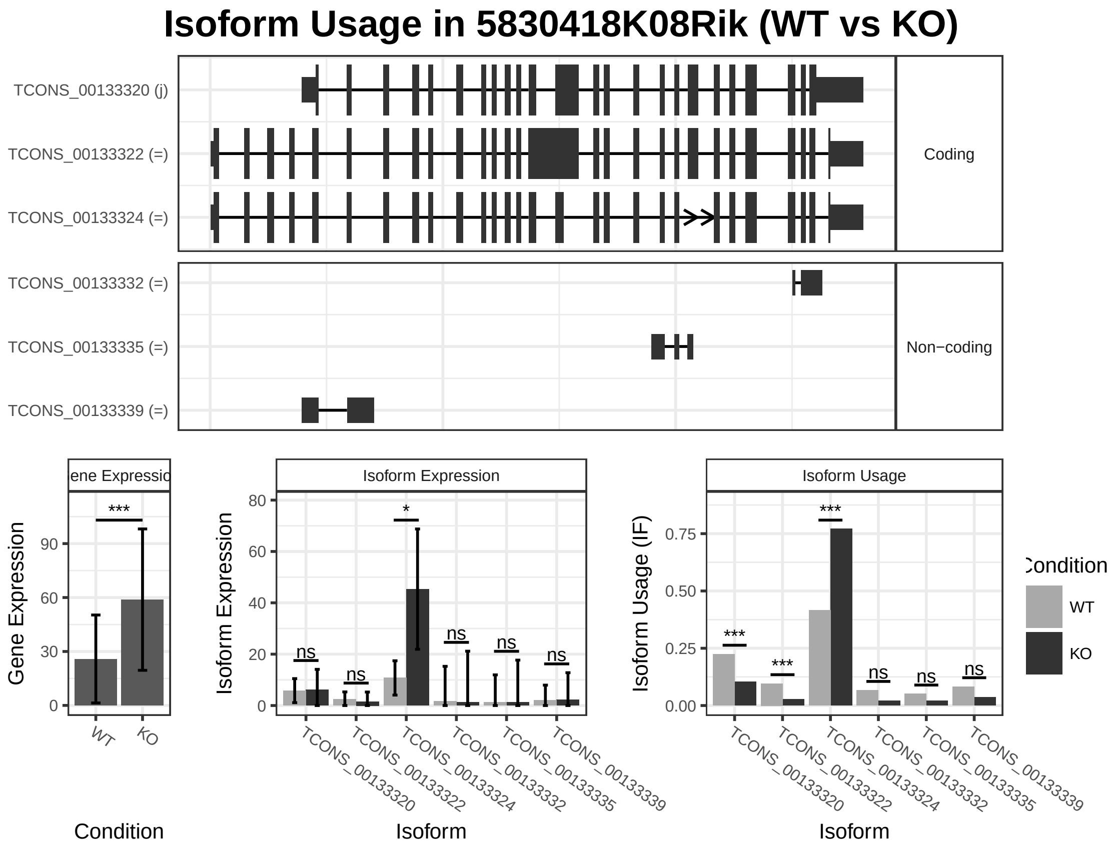
Table 1: 2.3 Sorted genes of lowest gene switch q-value

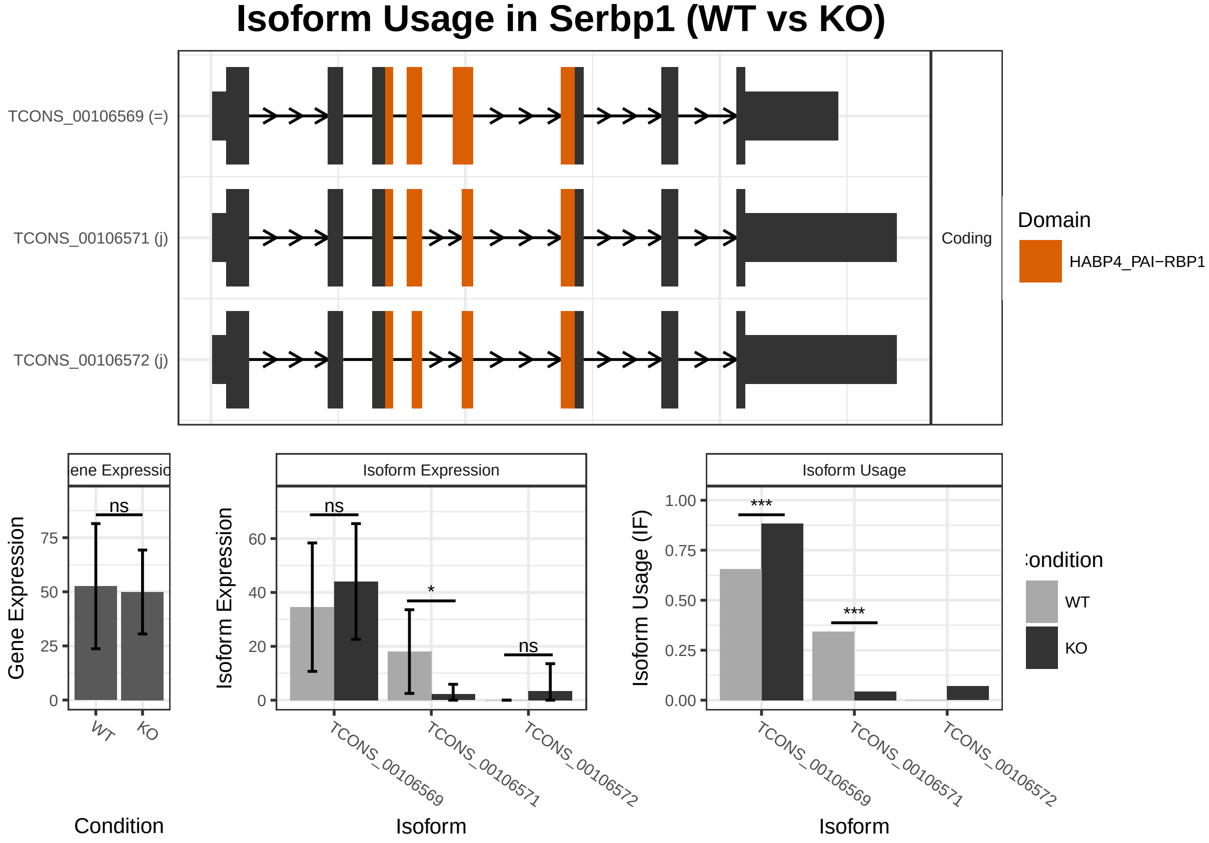
gene_name	ene_q_value	gene_switch_q_value
5830418K08Rik	0.000662452	3.175544e-64
Serbp1	0.905557000	1.478945e-19
Ablim1	0.035215400	1.155042e-15
Tef	0.039535100	4.686282e-15
Map4k4	0.000662452	4.770904e-15
Postn	0.005988670	9.818809e-14
Myo9a	0.896290000	8.903610e-13
Xrcc6	0.786097000	9.951012e-13
Snx14	1.000000000	4.031854e-12
Slmap	0.000662452	6.992658e-11

Question 2.4

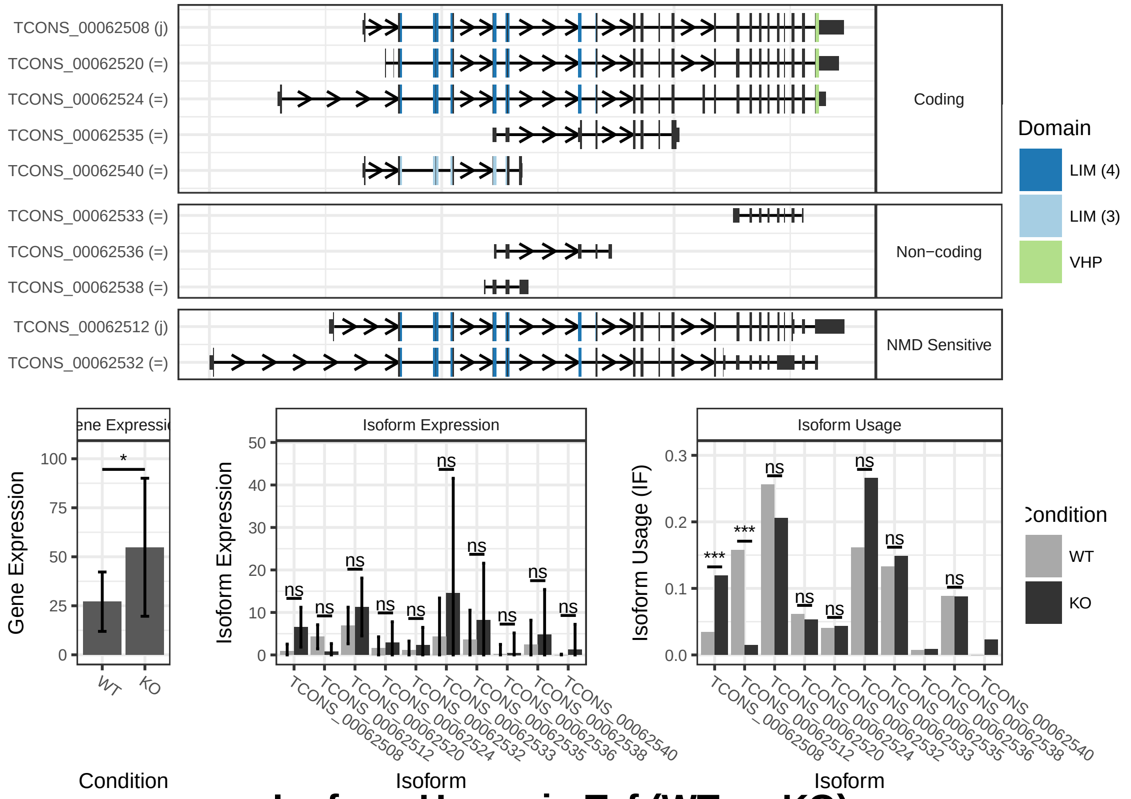
Show code for how to produce switchPlot for these 10 genes and save them to your own computer. The plots should not be included in the report (only the code for how to produce it)!

```
gene.names <- pull(My.SwitchList.select[1:10,], gene_name)

#switchPlot(switchAnalyzeRlist = My.SwitchList, gene='Tef')

plotter <- function(name){
  switchPlot(switchAnalyzeRlist = My.SwitchList, gene=name)
}

sapply(gene.names, FUN = plotter)
```

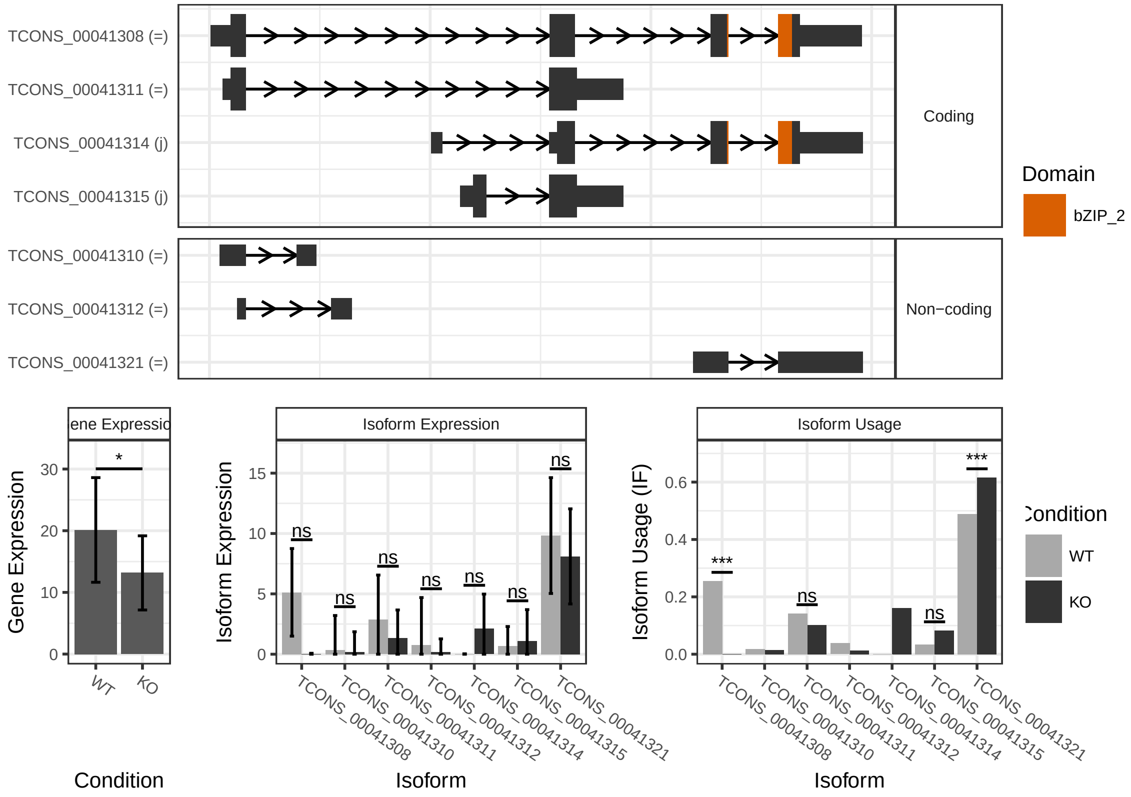

Isoform Usage in 5830418K08Rik (WT vs KO)
TCONS_00133320 (j)
TCONS_00133322 (=)
TCONS_00133324 (=)
Coding
TCONS_00133332 (=)
TCONS_00133335 (=)
TCONS_00133339 (=)
Non−coding
ene Expressio
Gene Expression
***
0
30
60
90
WT
KO
Condition
Isoform Expression
Isoform Expression
0
20
40
60
80
ns
ns
*
ns
ns
ns
TCONS_00133320
TCONS_00133322
TCONS_00133324
TCONS_00133332
TCONS_00133335
TCONS_00133339
Isoform
Isoform Usage
Isoform Usage (IF)
0.00
0.25
0.50
0.75
***
***
***
ns
ns
ns
TCONS_00133320
TCONS_00133322
TCONS_00133324
TCONS_00133332
TCONS_00133335
TCONS_00133339
Isoform
:ondition
WT
KO


Isoform Usage in Serbp1 (WT vs KO)
TCONS_00106569 (=)
TCONS_00106571 (j)
TCONS_00106572 (j)
Coding
Domain
HABP4_PAI−RBP1
ene Expressio
Gene Expression
ns
0
25
50
75
WT
KO
Condition
Isoform Expression
Isoform Expression
0
20
40
60
ns
*
ns
TCONS_00106569
TCONS_00106571
TCONS_00106572
Isoform
Isoform Usage
Isoform Usage (IF)
0.00
0.25
0.50
0.75
1.00
***
***
TCONS_00106569
TCONS_00106571
TCONS_00106572
Isoform
:ondition
WT
KO

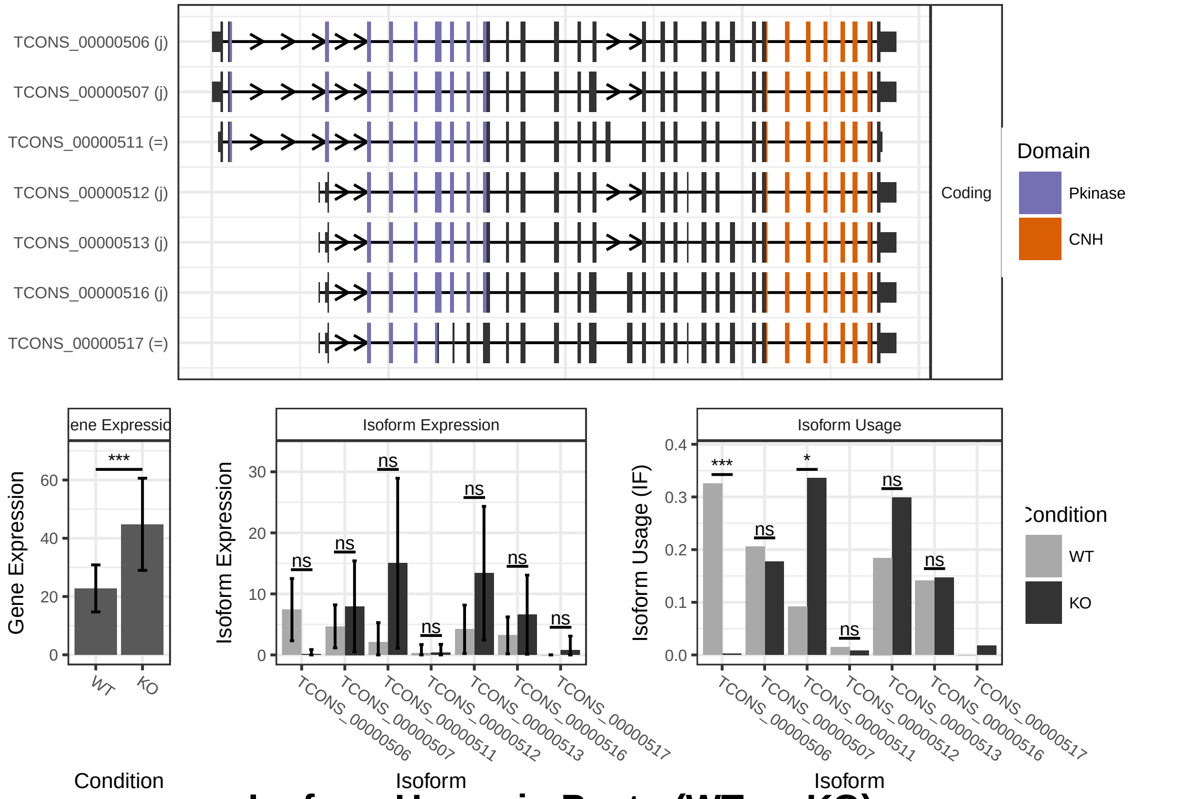
Isoform Usage in Ablim1 (WT vs KO)



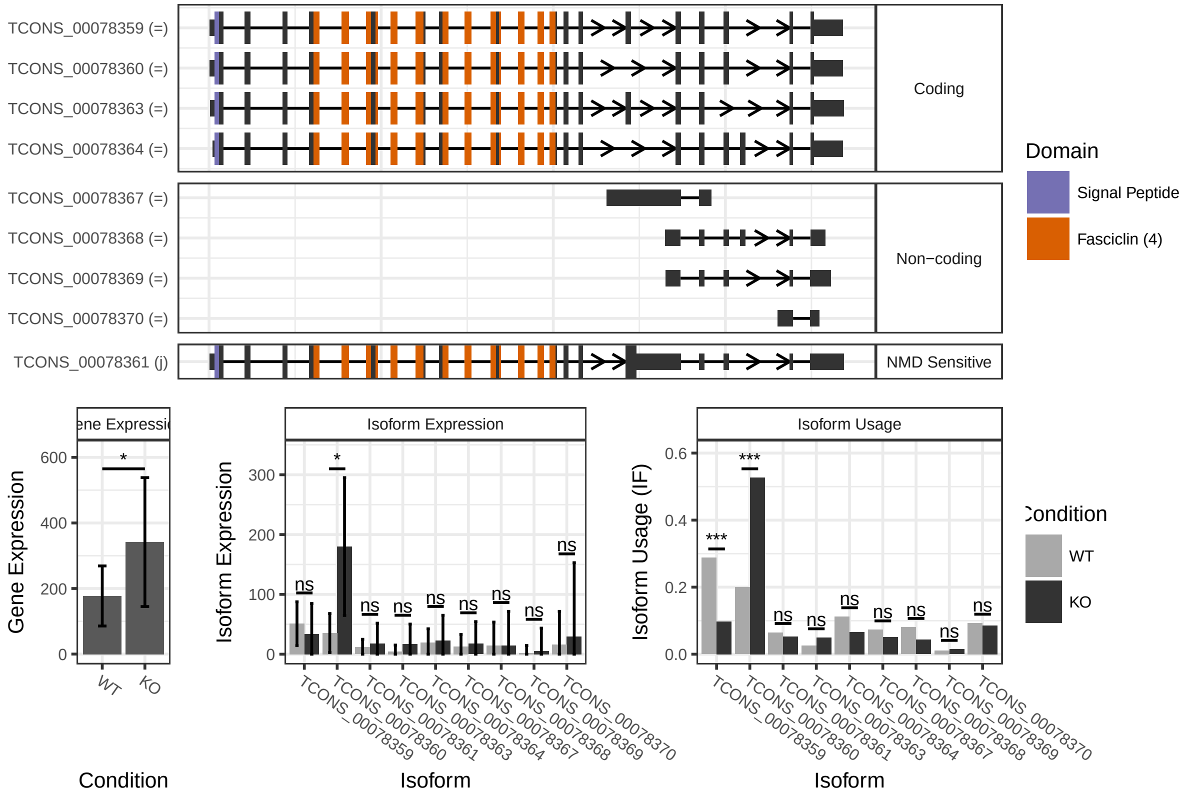
Isoform Usage in Tef (WT vs KO)

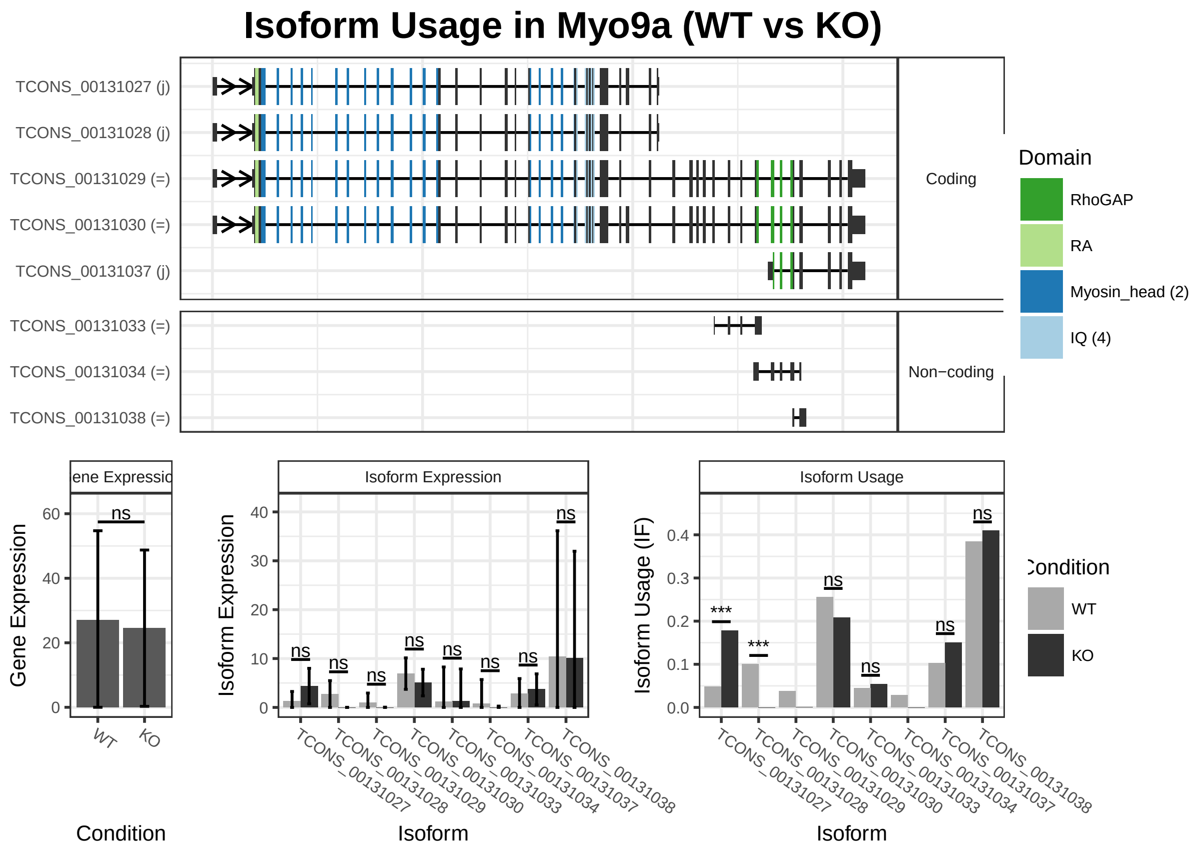


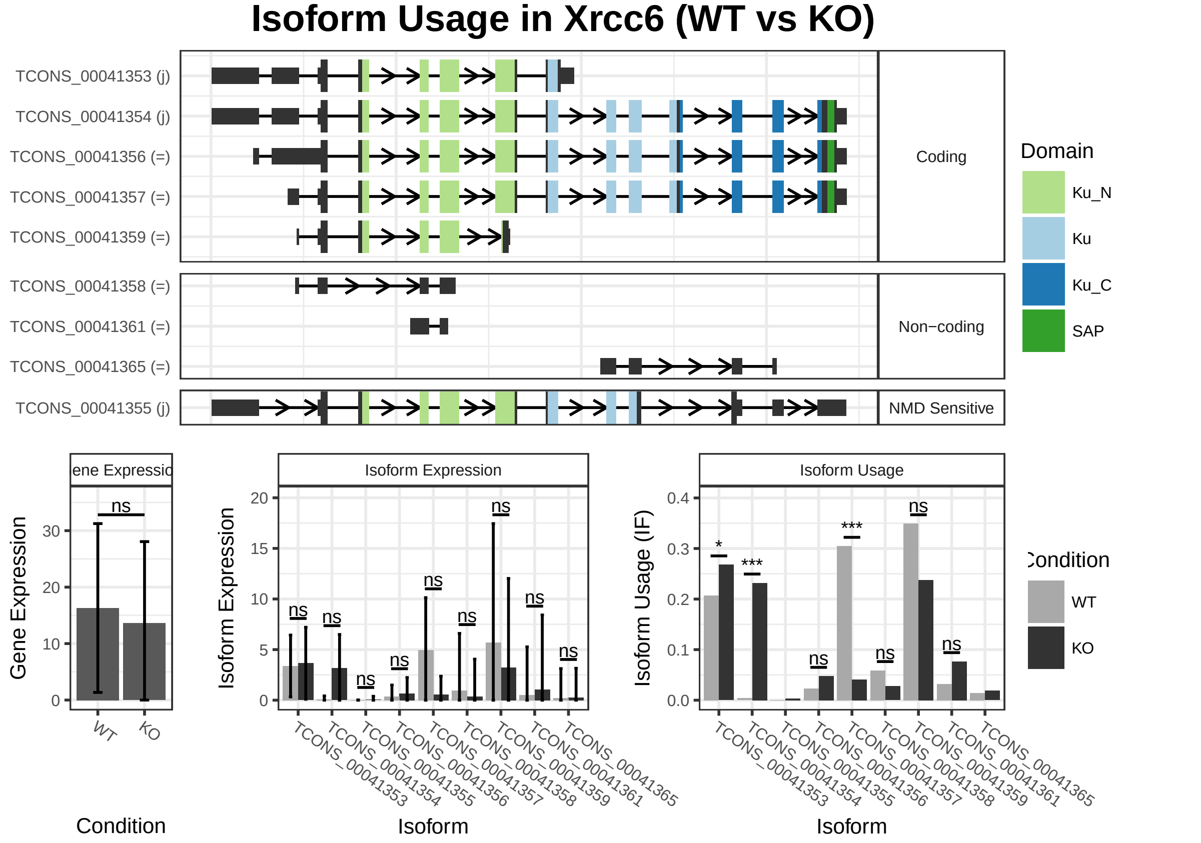
Isoform Usage in Map4k4 (WT vs KO)



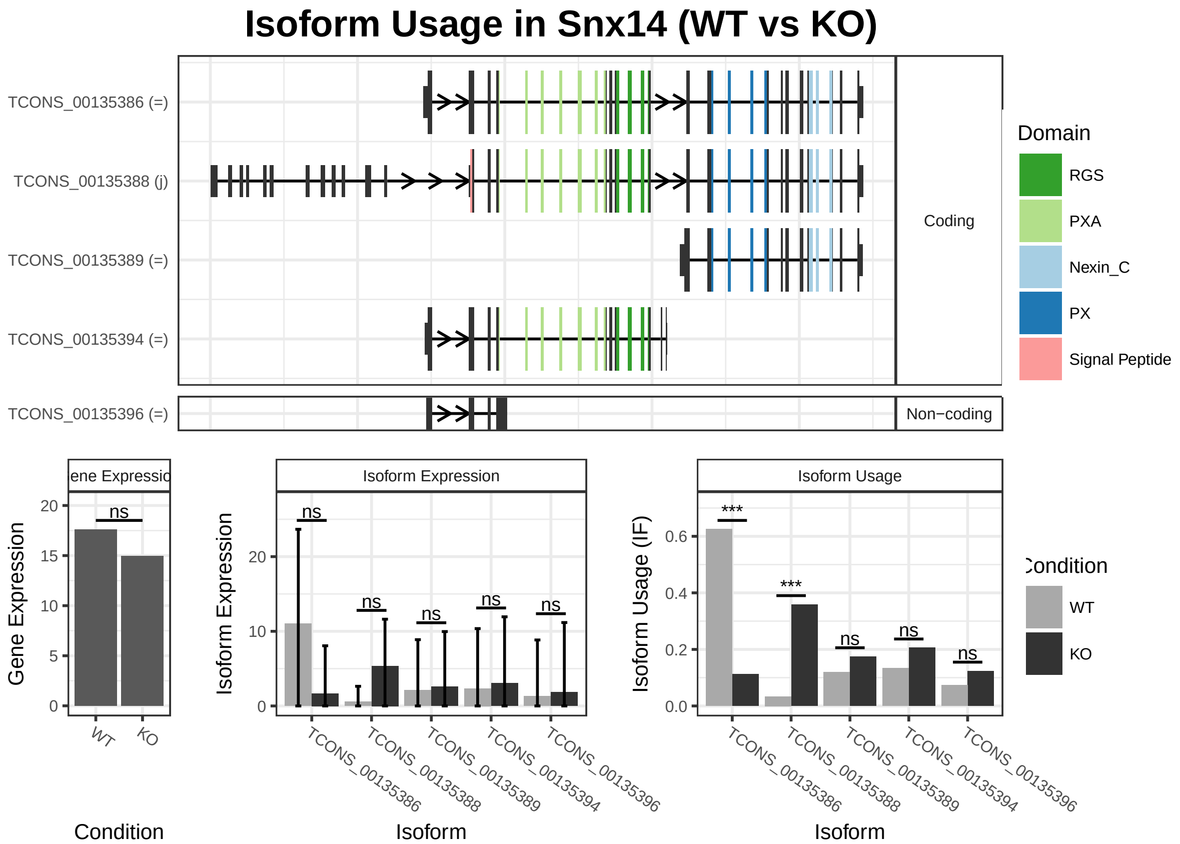
Isoform Usage in Postn (WT vs KO)

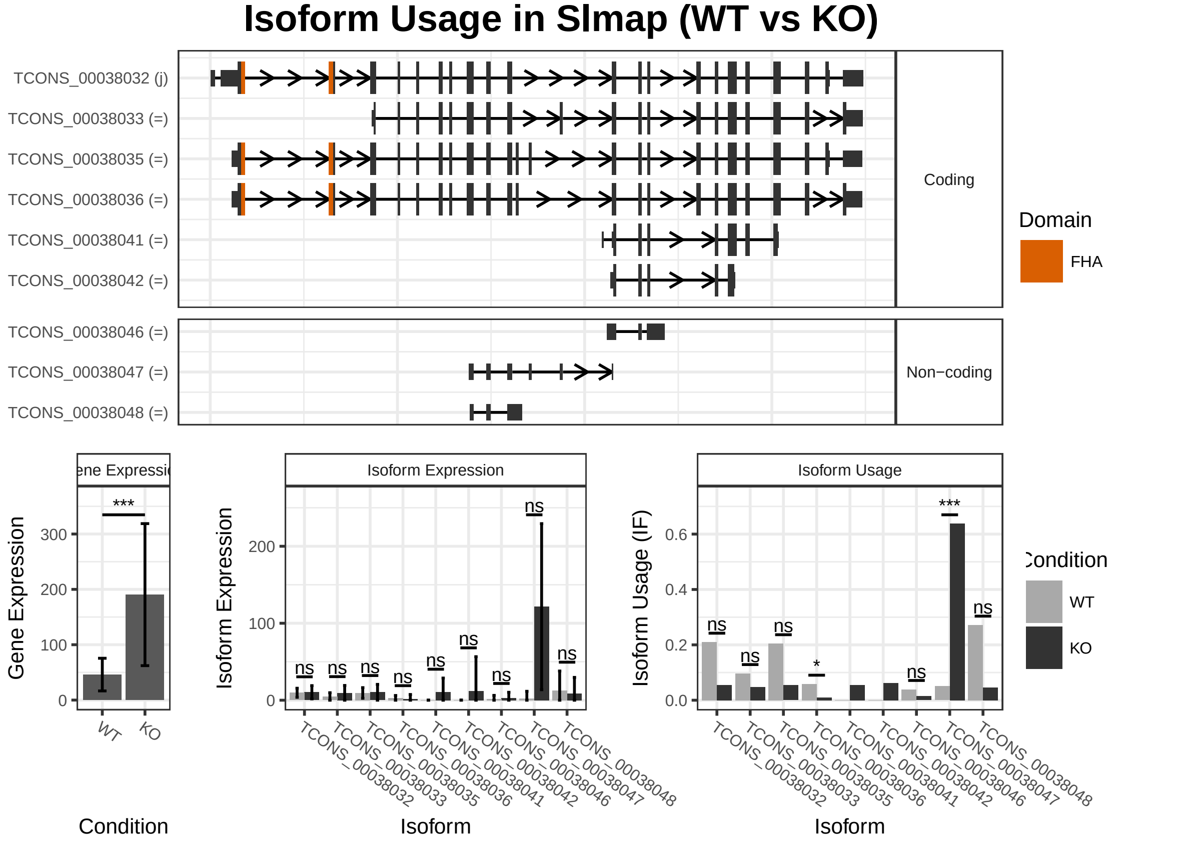






```
## Warning: Removed 2 rows containing missing values (geom_errorbar).
```





```
##        5830418K08Rik Serbp1 Ablim1 Tef     Map4k4 Postn   Myo9a   Xrcc6
## data   List,4        List,4 List,4 List,4 List,4 List,4 List,4 List,4
## layout ?             ?      ?      ?      ?      ?      ?      ?
## plot   List,9        List,9 List,9 List,9 List,9 List,9 List,9 List,9
##        Snx14  Slmap
## data   List,4 List,4
## layout ?      ?
## plot   List,9 List,9
```

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.**