LAB REPORT

Comparison of transcriptomic and proteomic data using R

Syuzanna Matevosyan

2022-05-27

Table of Contents

[Introduction 1](#_Toc105775015)

[Procedure 2](#_Toc105775016)

[Result 12](#_Toc105775017)

[Conclusion 14](#_Toc105775018)

## Introduction

The analysis is based on transcriptomic and proteomic data from healthy (control) and dystrophin-deficient (DMD) cardiomyocytes human(samples are from 2 patients). Applying data analysis methods to the transcriptomic and proteomic data of the same sample has the goal to check if the results from proteomics are consistent with the data from transcriptomic analysis. Having a list of genes (witch foldchange (Fch) and adjusted p-values) from the transcriptomic part where we compare disease vs. controls and a list of proteins (also characterized by foldchange and p-values) from the proteomic part where the same comparison was done. For this goal, using R scripts, the following questions were considered.

* how many records are common on both lists
* how many of them are consistent (up-regulated gene = higher protein level; down-regulated gene = lower protein level)
* how many of them are significant (adjusted p-value<0.05) botch on protein and gene/ only on protein or only on transcript level
* what is the correlation of fold-changes (for consistent genes-protein) e.g., if gene Fch=2 is protein level also 2-times higher
* what is the correlation for each gene between mRNA and protein levels across samples (i.e. spearman or pearson, get r^2 and pvalue, and see how is the distribution of correlations)

## Procedure

Following libraries were used during comparison.

library("readxl")  
library("VennDiagram")  
library("dplyr")  
library("ggrepel")  
library("tidyverse")

Data loaded from the file “rna\_seq\_protein\_data.xls” to a DataFrame objects referenced by df\_proteomic and df\_transcriptomic.

df\_proteomic = read\_excel("rna\_seq\_protein\_data.xls", sheet = 1)  
df\_transcriptomic = read\_excel("rna\_seq\_protein\_data.xls", sheet = 2)

Looking up for the structure of our DataFrame to make sure the data have been loaded correctly.

str(df\_proteomic)

## tibble [4,085 x 8] (S3: tbl\_df/tbl/data.frame)  
## $ Protein\_IDs : chr [1:4085] "P11532;E9PDN5;A0A075B6G3;A0A087WV90;A0A087WTU7;E7EQS5;F5GZY3;E7EQR9;Q4G0X0;Q14174;H0Y864;H0Y3E8;Q14172;A0A0B4J1W6;Q13474" "J3KTL2;Q07955;J3KSR8;J3QQV5;J3KSW7" "Q9Y5S2;H0YLY0;A0A0U1RRC3;H0Y6V3;H0Y7V8" "P17655" ...  
## $ Protein\_names : chr [1:4085] "Dystrophin" "Serine/arginine-rich splicing factor 1" "Serine/threonine-protein kinase MRCK beta" "Calpain-2 catalytic subunit" ...  
## $ Gene\_names : chr [1:4085] "DMD" "SRSF1" "CDC42BPB" "CAPN2" ...  
## $ Significant\_protein: chr [1:4085] "+" NA NA NA ...  
## $ -log10(p-value) : num [1:4085] 3.13 2.67 2.56 2.41 2.28 ...  
## $ P-value : num [1:4085] 0.000741 0.002142 0.002723 0.003917 0.005221 ...  
## $ log2(fold\_change) : num [1:4085] -6.205 -0.374 0.43 0.424 0.618 ...  
## $ Test statistic : num [1:4085] -7.41 -2.22 2.3 2.22 3.01 ...

str(df\_transcriptomic)

## tibble [20,812 x 7] (S3: tbl\_df/tbl/data.frame)  
## $ GeneID : chr [1:20812] "KCNH7" "PRSS23" "DUOX2" "NDUFA4L2" ...  
## $ Significant\_gene: chr [1:20812] "+" "+" "+" "+" ...  
## $ log2FoldChange : num [1:20812] -0.977 1.381 1.727 -1.391 1.008 ...  
## $ stat : num [1:20812] -5.51 5.02 4.9 -4.86 4.64 ...  
## $ pvalue : num [1:20812] 3.68e-08 5.18e-07 9.80e-07 1.15e-06 3.46e-06 ...  
## $ padj : num [1:20812] 0.000593 0.004168 0.004632 0.004632 0.011142 ...  
## $ Effect : chr [1:20812] "Down" "Up" "Up" "Down" ...

Data was cleaned. We are interested in columns that contain Gene names, Fold changes and p-values

df\_proteomic = df\_proteomic[c(3, 7, 6)]

df\_transcriptomic = df\_transcriptomic[c(1, 3, 6)]

DataFrame was filtered from values to retain only rows for which foldchane and pvalue are specified.

df\_proteomic <-df\_proteomic[!is.na(df\_proteomic$`log2(fold\_change)`),]  
df\_proteomic <- df\_proteomic[!is.na(df\_proteomic$`P-value`),]  
  
df\_transcriptomic <- df\_transcriptomic[!is.na(df\_transcriptomic$log2FoldChange),]  
df\_transcriptomic <- df\_transcriptomic[!is.na(df\_transcriptomic$padj),]

Some gene\_names were duplicated. DataFrames was filtered further to retain only one best (of lowest P-value) match for each gene. It made by sorting DataFrames by values of the column P-value, and then by removing duplicates (which by default leaves the first row).

df\_proteomic <- df\_proteomic[order(df\_proteomic$`P-value`),]  
df\_proteomic <- df\_proteomic[!duplicated(df\_proteomic$Gene\_names),]  
  
df\_transcriptomic <- df\_transcriptomic[order(df\_transcriptomic$padj),]  
df\_transcriptomic <- df\_transcriptomic[!duplicated(df\_transcriptomic$GeneID),]

Looking to data again

head(df\_proteomic)

## # A tibble: 6 x 3  
## Gene\_names `log2(fold\_change)` `P-value`  
## <chr> <dbl> <dbl>  
## 1 DMD -6.21 0.000741  
## 2 SRSF1 -0.374 0.00214   
## 3 CDC42BPB 0.430 0.00272   
## 4 CAPN2 0.424 0.00392   
## 5 PTRH2 0.618 0.00522   
## 6 PVR 0.854 0.00616

head(df\_transcriptomic)

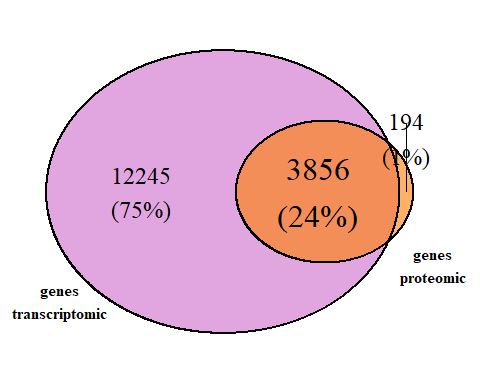
## # A tibble: 6 x 3  
## GeneID log2FoldChange padj  
## <chr> <dbl> <dbl>  
## 1 KCNH7 -0.977 0.000593  
## 2 PRSS23 1.38 0.00417   
## 3 DUOX2 1.73 0.00463   
## 4 NDUFA4L2 -1.39 0.00463   
## 5 PHACTR1 1.01 0.0111   
## 6 GABRE 1.67 0.0377

The list of Gene names from protein analysis compared with a list of geneID from the transcriptomic analysis for identification common genes.

common\_genes <- intersect(df\_proteomic$Gene\_names, df\_transcriptomic$GeneID)

For visualization “VennDiagramm” library were used

v <- venn.diagram(list(`genes\nproteomic`=df\_proteomic$Gene\_names,  
 `genes\ntranscriptomic`=df\_transcriptomic$GeneID),  
 col = "black",  
 fill = c("darkorange1", "orchid3"),  
 alpha = c(0.6, 0.6),   
 cat.cex = 1,   
 cat.fontface = "bold",   
 cat.dist = c(0.06, 0.04),  
 cex = c(1.5, 2, 1.5),  
 margin = 0.05,  
 filename=NULL)  
  
  
v[[5]]$label <- "12245\n(75%)"  
v[[6]]$label <- "\n194\n(1%)"  
v[[8]]$label <- "3856\n(24%)"  
  
grid.newpage()  
grid.draw(v)



3856 records are common from proteomic and transcriptomic data. Common records are in object referenced by common\_genes

str(common\_genes)

## chr [1:3856] "DMD" "SRSF1" "CDC42BPB" "CAPN2" "PTRH2" "PVR" "SEC22B" ...

For further analysis data was filtered to remain only data belonging to common records.

df\_proteomic <- df\_proteomic %>% filter(Gene\_names %in% common\_genes)

df\_transcriptomic <- df\_transcriptomic %>% filter(GeneID %in% common\_genes)

New dataframe was created collecting all common records. Before that, rows of df\_proteomic and df\_transcriptomic reordered by using common\_genes order.

df\_proteomic <- df\_proteomic %>% arrange(factor(Gene\_names, levels = common\_genes))  
df\_transcriptomic <- df\_transcriptomic %>% arrange(factor(GeneID, levels = common\_genes))

proteomic\_pvalue <- df\_proteomic$`P-value`  
transcriptomic\_pvalue <- df\_transcriptomic$padj  
proteomic\_foldchange <- df\_proteomic$`log2(fold\_change)`  
transcriptomic\_foldchange <- df\_transcriptomic$log2FoldChange  
  
df\_common <- data.frame(common\_genes, proteomic\_pvalue, transcriptomic\_pvalue, proteomic\_foldchange, transcriptomic\_foldchange)

New data\_frame is

head(df\_common)

## common\_genes proteomic\_pvalue transcriptomic\_pvalue proteomic\_foldchange  
## 1 DMD 0.0007405255 0.9986817 -6.205250  
## 2 SRSF1 0.0021416081 0.9986817 -0.374339  
## 3 CDC42BPB 0.0027234537 0.9986817 0.429546  
## 4 CAPN2 0.0039173286 0.8609095 0.423963  
## 5 PTRH2 0.0052205950 0.9986817 0.618248  
## 6 PVR 0.0061563032 0.9986817 0.853700  
## transcriptomic\_foldchange  
## 1 -0.11286383  
## 2 -0.07996891  
## 3 -0.17343846  
## 4 0.68476046  
## 5 -0.75685377  
## 6 0.19533893

Using fold\_change values, data was seperated into 3 groups. 1: (High regulated genes) Fold change is positive number in both transcriptomic and proteomic data. 2: (Low regulated genes) Fold change is negative number in both transcriptomic and proteomic data 3: (Different regulation) Fold change is positive in one data and negative in another.

positive <- (df\_common$proteomic\_foldchange > 0 & df\_common$transcriptomic\_foldchange > 0)  
negative <- (df\_common$proteomic\_foldchange < 0 & df\_common$transcriptomic\_foldchange < 0)  
different <- ((df\_common$proteomic\_foldchange < 0 & df\_common$transcriptomic\_foldchange > 0) | (df\_common$proteomic\_foldchange > 0 & df\_common$transcriptomic\_foldchange < 0))

3th group seperetaed into 2 subgroups 3.1: Negative fold change in proteomic and Positive fold change in transcriptomic 3.2: Positive fold change in proteomic and Negative fold change in transcriptomic

negative\_proteomic\_positive\_transcriptomic <- (df\_common$proteomic\_foldchange < 0 & df\_common$transcriptomic\_foldchange > 0)  
positive\_proteomic\_negative\_transcriptomic <- (df\_common$proteomic\_foldchange > 0 & df\_common$transcriptomic\_foldchange < 0)

Record numbers of each group was stored on variables.

nump = length(which(positive == TRUE))  
numn = length(which(negative == TRUE))  
numd = length(which(different == TRUE))  
numnppt = length(which(negative\_proteomic\_positive\_transcriptomic == TRUE))  
numppnt = length(which(positive\_proteomic\_negative\_transcriptomic == TRUE))

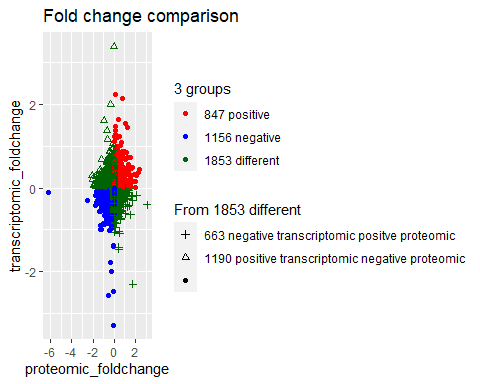
Status column added to the df\_common, which represents the regulation of genes.

df\_common <- df\_common %>%  
 mutate(Status = case\_when(  
 (positive == TRUE) ~ nump,  
 (negative == TRUE) ~ numn,  
 (different == TRUE) ~ numd  
 ))  
  
df\_common <- df\_common %>%  
 mutate(Status1 = case\_when(  
 (negative\_proteomic\_positive\_transcriptomic == TRUE) ~ numnppt,  
 (positive\_proteomic\_negative\_transcriptomic == TRUE) ~ numppnt,  
 (different == FALSE) ~ nump+numn  
 ))  
  
  
head(df\_common)

## common\_genes proteomic\_pvalue transcriptomic\_pvalue proteomic\_foldchange  
## 1 DMD 0.0007405255 0.9986817 -6.205250  
## 2 SRSF1 0.0021416081 0.9986817 -0.374339  
## 3 CDC42BPB 0.0027234537 0.9986817 0.429546  
## 4 CAPN2 0.0039173286 0.8609095 0.423963  
## 5 PTRH2 0.0052205950 0.9986817 0.618248  
## 6 PVR 0.0061563032 0.9986817 0.853700  
## transcriptomic\_foldchange Status Status1  
## 1 -0.11286383 1156 2003  
## 2 -0.07996891 1156 2003  
## 3 -0.17343846 1853 663  
## 4 0.68476046 847 2003  
## 5 -0.75685377 1853 663  
## 6 0.19533893 847 2003

Data was ploted using ggplot function. For axis Fold change used from transcriptomic and proteomic.

p <- ggplot(df\_common, aes(x = proteomic\_foldchange, y = transcriptomic\_foldchange)) + geom\_point(aes(color =   
as.factor(Status), shape = as.factor(Status1)))  
p <- p + labs(color = "3 groups", shape = "From 1853 different", title = "Fold change comparison")  
p <- p + guides(color = guide\_legend(order = 1), shape = guide\_legend(order = 2))  
cols = c("847" = "red", "1156" = "blue", "1853" = "darkgreen")  
p <- p + scale\_color\_manual(values = cols, labels = c("847 positive", "1156 negative", "1853 different"))  
p <- p + scale\_shape\_manual(values=c(3, 2, 16), labels = c("663 negative transcriptomic positve proteomic", "1190 positive transcriptomic negative proteomic", ""))  
p



Result of 3 groups. 1: 847 (High regulated genes) Fold change is positive number in both transcriptomic and proteomic data 2: 1156 (Low regulated genes) Fold change is negative number in both transcriptomic and proteomic data 3: 1853 (Different regulation) Fold change is positive in one data and negative in another.

3.1: 1190 genes have negative fold change in proteomic and Positive fold change in transcriptomic 3.2: 663 genes have positive fold change in proteomic and Negative fold change in transcriptomic

For further steps data sepereated into 3 groups based on p-value. 1: (Significant genes) p-value is smaller than 0.05 in both transcriptomic and proteomic data 2: (Non sginificant genes) p-value is higher than 0.05 in both transcriptomic and proteomic data 3: (Different) p-value is smaller than 0.05 in one data and higher in another

sign <- (df\_common$proteomic\_pvalue < 0.05 & df\_common$transcriptomic\_pvalue < 0.05)  
notsign <- (df\_common$proteomic\_pvalue > 0.05 & df\_common$transcriptomic\_pvalue > 0.05)  
differents <- ((df\_common$proteomic\_pvalue < 0.05 & df\_common$transcriptomic\_pvalue > 0.05) | (df\_common$proteomic\_pvalue > 0.05 & df\_common$transcriptomic\_pvalue < 0.05))

Number of each group was calculated

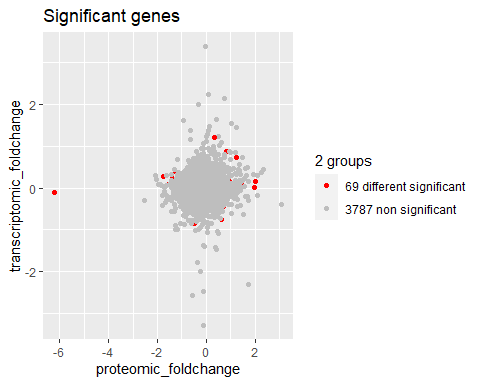
nums = length(which(sign == TRUE))  
numns = length(which(notsign == TRUE))  
numds = length(which(differents == TRUE))

Status2 column added to the df\_common, which represents the significant of genes.

df\_common <- df\_common %>%  
 mutate(Status2 = case\_when(  
 (sign == TRUE) ~ nums,  
 (notsign == TRUE) ~ numns,  
 (differents == TRUE) ~ numds  
 ))

Visualization

p <- ggplot(df\_common, aes(x = proteomic\_foldchange, y = transcriptomic\_foldchange)) + geom\_point(aes(color =   
as.factor(Status2)))  
p <- p + scale\_color\_manual(values = c("red", "grey"), labels = c("69 different significant", "3787 non significant"))  
p <- p + labs(color = "2 groups", title = "Significant genes")  
p

 Result of 3 groups. 1: 0 (Significant genes) p-value is smaller than 0.05 in both transcriptomic and proteomic data 2: 3787 (Non significant genes) p-value is higher than 0.05 in both transcriptomic and proteomic data 3: 69 (Different) p-value is smaller than 0.05 in one data and higher in another

Data was filtered and only genes that have the same regulation and are significant in one data are left.

df\_common\_fch\_sign <- df\_common %>% filter(different == FALSE & differents == TRUE)

Data was devided into 2 groups 3.1: p-value is smaller than 0.05 in proteomic data 3.2: p-value is smaller than 0.05 in transcriptomic data

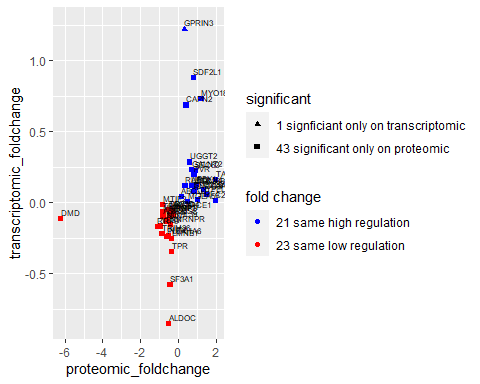
sign\_onlyprot <- df\_common\_fch\_sign$proteomic\_pvalue < 0.05  
sign\_onlytrans <- df\_common\_fch\_sign$transcriptomic\_pvalue < 0.05

Number of each group was calculated

numsop <- length(which(sign\_onlyprot == TRUE))  
numsot <- length(which(sign\_onlytrans == TRUE))

df\_common\_fch\_sign <- df\_common\_fch\_sign %>%  
 mutate(Status4 = case\_when(  
 (sign\_onlyprot == TRUE) ~ numsop,  
 (sign\_onlytrans == TRUE) ~ numsot  
 ))  
  
positive <- (df\_common\_fch\_sign$proteomic\_foldchange > 0 & df\_common\_fch\_sign$transcriptomic\_foldchange > 0)  
negative <- (df\_common\_fch\_sign$proteomic\_foldchange < 0 & df\_common\_fch\_sign$transcriptomic\_foldchange < 0)  
nump = length(which(positive == TRUE))  
numn = length(which(negative == TRUE))  
df\_common\_fch\_sign <- df\_common\_fch\_sign %>%  
 mutate(Status5 = case\_when(  
 (positive == TRUE) ~ nump,  
 (negative == TRUE) ~ numn  
 ))

p <- ggplot(df\_common\_fch\_sign, aes(x = proteomic\_foldchange, y = transcriptomic\_foldchange)) + geom\_point(aes(color =   
as.factor(Status5), shape = as.factor(Status4))) + geom\_text(size = 2, hjust = 0, nudge\_y = 0.05, aes(label = common\_genes))  
p <- p + labs(color = "fold change", shape = "significant")  
p <- p + scale\_color\_manual(values = c("blue", "red"), labels = c("21 same high regulation", "23 same low regulation"))  
p <- p + scale\_shape\_manual(values = c(17, 15), labels = c("1 signficiant only on transcriptomic", "43 significant only on proteomic"))  
p



## Result

To see the correlation for each gene between mRNA and protein levels across samples, r^2 and pvalue counted.

mod1 = lm(transcriptomic\_foldchange~proteomic\_foldchange, data = df\_common\_fch\_sign)  
modsum = summary(mod1)  
  
r2 = modsum$adj.r.squared  
r2

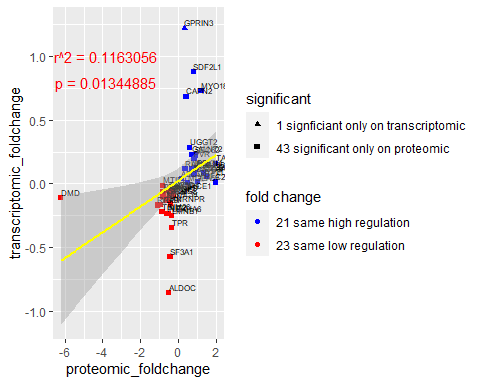
## [1] 0.1163056

my.p = modsum$coefficients[2,4]  
my.p

## [1] 0.01344885

Visualization of the distribution of correlations

p <- ggplot(df\_common\_fch\_sign, aes(x = proteomic\_foldchange, y = transcriptomic\_foldchange)) + geom\_point(aes(color =   
as.factor(Status5), shape = as.factor(Status4))) + geom\_text(size = 2, hjust = 0, nudge\_y = 0.05, aes(label = common\_genes))  
p <- p + labs(color = "fold change", shape = "significant")  
p <- p + scale\_color\_manual(values = c("blue", "red"), labels = c("21 same high regulation", "23 same low regulation"))  
p <- p + scale\_shape\_manual(values = c(17, 15), labels = c("1 signficiant only on transcriptomic", "43 significant only on proteomic"))  
p <- p + stat\_smooth(method = "lm", col = "yellow")  
p <- p + annotate(geom= "text", x = -3.8, y= 1, label = "r^2 = 0.1163056", color = "red")  
p <- p + annotate(geom = "text", x= -3.8, y = 0.8, label = "p = 0.01344885", color = "red")  
p



If we remove DMD from data, and make corralation again we will have the following results.

df\_common\_fch\_sign <- df\_common\_fch\_sign[-1,]

Make visualization

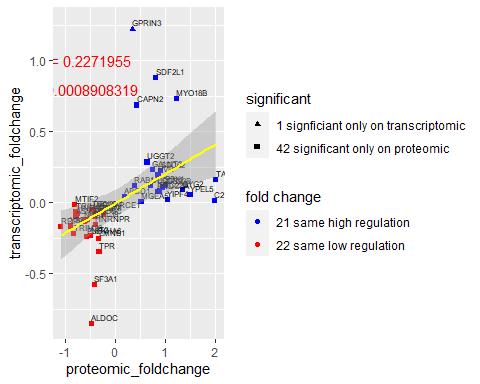
mod1 = lm(transcriptomic\_foldchange~proteomic\_foldchange, data = df\_common\_fch\_sign)  
modsum = summary(mod1)  
  
r2 = modsum$adj.r.squared  
r2

## [1] 0.2271955

my.p = modsum$coefficients[2,4]  
my.p

## [1] 0.0007272979

p <- ggplot(df\_common\_fch\_sign, aes(x = proteomic\_foldchange, y = transcriptomic\_foldchange)) + geom\_point(aes(color =   
as.factor(Status5), shape = as.factor(Status4))) + geom\_text(size = 2, hjust = 0, nudge\_y = 0.05, aes(label = common\_genes))  
p <- p + labs(color = "fold change", shape = "significant")  
p <- p + scale\_color\_manual(values = c("blue", "red"), labels = c("21 same high regulation", "22 same low regulation"))  
p <- p + scale\_shape\_manual(values = c(17, 15), labels = c("1 signficiant only on transcriptomic", "42 significant only on proteomic"))  
p <- p + stat\_smooth(method = "lm", col = "yellow")  
p <- p + annotate(geom= "text", x = -0.7, y= 1, label = "r^2 = 0.2271955", color = "red")  
p <- p + annotate(geom = "text", x= -0.7, y = 0.8, label = "p = 0.0008908319", color = "red")  
p



## Conclusion

Analysis has shown that there exists a poor correlation for each gene between mRNA and protein levels across samples. Results showed low R-square and low p-value (p-value <= 0.05), which means that this model doesn’t explain much of the variation of the data but is significant.