Lab_book_12_2_16

Claire Green

Monday

I have now completed my second attempt at a literature review. This will be uploaded to the repository. Now that this is finished my next step is to look at creating a statistically robust protocol for generating lists of DEGs.

Notes for meeting with Wenbin

Data: Affy HG Plus 2.0 RNA expression (except FTLD - HG U133A 2.0)

C9orf72 motor neurons: 8 patients, 3 controls CHMP2B motor neurons: 3 patients, 6 controls sALS motor neurons: 7 patients, 3 controls FTLD cortical neurons: 16 patients, 8 controls VCP muscle: 7 patients, 3 controls

R script

1. Run each data set through Wenbin's code individually
2. Take top X most DEGs –> Do I rank by fold change, pvalue or adjusted p value?
3. Find genes that are consensus across all data sets –> which genes are commonly dysregulated?
4. Validate number is more than you would expect by chance by random permutations test
    - say top 2000 from each data set yields 10 common genes
    - take 5 sets of 2000 random genes
    - calculate consensus value
    - repeat 10,000 times
    - number of times value is >10 / 10,000 should equal p value (less than 0.05 is acceptable)

Tuesday

I discussed with Wenbin about the optimal way of ordering DEGs, and he essentially said that different people have different methods. I decided then to compare my output of DEGs ordered by adjusted p value as compared to fold change. Expression was calculated by WenBin DE Gene.R and sorted using the following lines of code:

```
###Write results to CSV files for consensus analysis
setwd("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/DE Genes/TopGenes_2016-02-09/")
# dir.create(paste("TopGenes", Sys.Date(), sep = "_"))
#Take results, remove duplicate rows for genes, order by adjusted p value and take top X number of gene
uniqueresult <- result[!duplicated(result[,15]),]

#For ordering by adjusted p value

genesort <- uniqueresult[order(uniqueresult$adj.P.Val),]
topgene <- genesort[1:1000,]
write.csv(x = topgene, file = "VCP_ap_1000")
topgene <- genesort[1:2000,]
```

```
write.csv(x = topgene, file = "VCP_ap_2000")
topgene <- genesort[1:3000,]
write.csv(x = topgene, file = "VCP_ap_3000")
topgene <- genesort[1:4000,]
write.csv(x = topgene, file = "VCP_ap_4000")
topgene <- genesort[1:5000,]
write.csv(x = topgene, file = "VCP_ap_5000")

#For ordering by fold change
genesort <- uniqueresult[order(uniqueresult$`Fold Change`),]
topgene <- genesort[1:500,]

genesort <- uniqueresult[order(-uniqueresult$`Fold Change`),]
botgene <- genesort[1:500,]

topFC <- rbind(topgene,botgene)

write.csv(x = topgene, file = "VCP_fc_1000")
```

I evaluated consensus using the file DE consensus.R. It is as follows:

```
#####DIFFERENTIAL GENE EXPRESSION INTERSECT
#takes csv files of top X DE genes and identifies any consensus genes

setwd("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/GeneExpressionAnalysis/TopGenes_2016-02-09/")
A <- read.csv(file = "VCP_ap_1000")

B <- read.csv(file = "VCP_fc_1000")

# sALS<- read.csv(file = "sALS_anno_5000")
#
# FTLD<- read.csv(file = "FTLD_anno_5000")
#
# VCP<- read.csv(file = "VCP_anno_5000")


C9_DE<- A$Gene.Symbol
CHMP2B_DE <- B$Gene.Symbol
# sALS_DE <- sALS$Gene.Symbol
# FTLD_DE <- FTLD$Gene.Symbol
# VCP_DE <- VCP$Gene.Symbol
overlap <- Reduce(intersect, list(C9_DE, CHMP2B_DE))
print(overlap)
```

According to this example, 250 genes were shared by the two methods. This indicates there could be a big difference in results by ranking by adjusted p value rather than fold change.

My next task is to identify whether one method produces more consensus than the other. To do this I will generate top 1000 DEGs for all 5 data sets using the two methods, and compare number of consensus results. If 1000 is not enough, I will try 2000.

UPDATE When I tried to conduct the experiment again, I got completely different results. In december, I generated:

1000 Genes

0

2000 Genes

CSRP1 RNF130

3000 Genes

CSRP1 RNF13 TUBB3 PSAP DDX39A NKTR NAGA NUTF2 /// NUTF2P4 RPL6 CCT2 DCN RNF130 TARS PFDN1 HSD17B4

4000 Genes

UPF3A DYNLT1 CSRP1 EEF1A1 RNF13 MAP4K4 CDK5R1 CST3 ICMT TUBB3 BRD3 DMD PSAP DDX39A NKTR TFRC NAGA WBSCR22 NUTF2 /// NUTF2P4 PAICS RPL6 CCT2 DCN RPS6 GSTO1 TARDBP PRKD1 RNF130 TARS PFDN1 BPTF HSD17B4 PSMD1 GNPAT C14orf1 TMEM59 TRO ATP6V1G2-DDX39B /// DDX39B /// SNORD84 ABCD3 PTEN ZFYVE16 DDX5 /// MIR3064 /// MIR5047 GTF2I /// GTF2IP1 /// LOC100093631 MYL12A /// MYL12B LBR MTMR9 LSM5 WFS1 TANK PPP4R1 C18orf32 /// RPL17 /// RPL17-C18orf32 /// SNORD58A /// SNORD58B /// SNORD58C RAB40B

When I ran the same results (generated Dec 4th 2015) again using the DE consensus.R script, I ended up generating slightly different results:

1000

0

2000

CSRP1

3000

CSRP1 RNF13 FN1 TUBB3 PSAP RPL6 CCT2 NKTR NUTF2 /// NUTF2P4 NAGA PFDN1 TARS RNF130 HSD17B4 DDX5 /// MIR3064 /// MIR5047 GTF2I /// GTF2IP1 /// LOC100093631

4000

UPF3A DYNLT1 CSRP1 EEF1A1 ETS2 RNF13 FN1 CST3 MAP4K4 BRD3 TUBB3 PSAP RPL6 CCT2 DMD ICMT NKTR NUTF2 /// NUTF2P4 RPS6 CDK5R1 BPTF PRKD1 NAGA GSTO1 PFDN1 ATP6V1G2-DDX39B /// DDX39B /// SNORD84 TARDBP TARS PTEN PAICS RNF130 HSD17B4 TMEM59 TRO DDX5 /// MIR3064 /// MIR5047 GNPAT GTF2I /// GTF2IP1 /// LOC100093631 WBSCR22 C14orf1 MTMR9 TCF4 WDR78 LBR DDX39A C18orf32 /// RPL17 /// RPL17-C18orf32 /// SNORD58A /// SNORD58B /// SNORD58C DCN RPLP2 /// SNORA52 PSMD1 RECQL MPHOSPH9

This was generated using the following DE consensus.R script:

```
#####DIFFERENTIAL GENE EXPRESSION INTERSECT
#takes csv files of top X DE genes and identifies any consensus genes

setwd("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/DE Genes/TopGenes_2016-02-09/")
C9orf72 <- read.csv(file = "C9_ap_5000")

CHMP2B <- read.csv(file = "CH_ap_5000")

sALS<- read.csv(file = "sALS_ap_5000")

FTLD<- read.csv(file = "FTLD_ap_5000")

VCP<- read.csv(file = "VCP_ap_5000")


C9_DE<- C9orf72$Gene.Symbol
CHMP2B_DE <- CHMP2B$Gene.Symbol
sALS_DE <- sALS$Gene.Symbol
FTLD_DE <- FTLD$Gene.Symbol
VCP_DE <- VCP$Gene.Symbol
overlap <- Reduce(intersect, list(C9_DE, CHMP2B_DE, sALS_DE, FTLD_DE, VCP_DE))
print(overlap)
```

I re ran the original data again in WenBin DE Gene.R and got completely different results. This is the version of the script I used: ##Differential Expression of Genes##

```
setwd("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/DE Genes/FTLD/")
library(affy)
library(Biobase)
library(tkWidgets)

#run program to choose .CEL files from directory
celfiles <- fileBrowser(textToShow = "Choose CEL files", testFun = hasSuffix("[cC][eE][lL]"))
#celfiles<-basename(celfiles)
Data<-ReadAffy(filenames=celfiles) #read in files
rmaEset<-rma(Data) #normalise using RMA
#analysis.name<-"CHMP2B" #Label analysis
dataMatrixAll<-exprs(rmaEset) #takes expression from normalised expression set


#mas5call generates presence/absence calls for each probeset
mas5call<-mas5calls(Data)
callMatrixAll<-exprs(mas5call)
colnames(callMatrixAll)<-sub(".CEL", ".mas5-Detection", colnames(callMatrixAll),fixed=TRUE)
colnames(callMatrixAll)<-sub(".cel", ".mas5-Detection", colnames(callMatrixAll),fixed=TRUE)
callMatrixAll<-as.data.frame(callMatrixAll)
callMatrixAll$ProbeSetID<-rownames(callMatrixAll)
countPf<-function(x){
  sum(x=="P")
}

#count how many samples have presence calls
countPl<-apply(callMatrixAll, 1, countPf)
```

```
callMatrixAll$ProbeSetID<-rownames(callMatrixAll)
countPdf<-data.frame(ProbeSetID=names(countPl), countP=countPl)

#read annotation file

###USING BIOMART
# library (biomaRt)
# mart <- useMart("ENSEMBL_MART_ENSEMBL",dataset="hsapiens_gene_ensembl", host="www.ensembl.org")
# x <- rownames(dataMatrixAll) #create vector containing probe IDs
# mart_attribute <- listAttributes(mart)
# annotation <- getBM(attributes=c("affy_hg_u133a_2", "hgnc_symbol", "description"),
#                     filters = "affy_hg_u133a_2", values = x, mart = mart)
# annotation<-subset(annotation, subset=(hgnc_symbol !="")) #if no gene symbol, discount

#USING ANNOTATION FILE (if .csv, convert to .txt using excel)
#annotation.file<-"/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/HG-U133_Plus_2.na35.annot.csv/HG
annotation.file<-"/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/HG-U133A_2.na35.annot.csv/HG-U133A

annotation<-read.table(annotation.file, header = TRUE, row.names=NULL, sep="\t", skip=0, stringsAsFactor
dim(annotation)
nrow(annotation)
#[1] 39699
annotation<-subset( annotation, subset=(Gene.Symbol !="---")) #if no gene symbol, discount
nrow(annotation)

expressionMatrix<-exprs(rmaEset)
colnames(expressionMatrix)

#this is for matched samples
#tonsil<-factor(c("T101","T101","T102","T102","T103","T103"))
Treat<-factor(rep(c("Control", "Patient"),c(8,16)), levels=c("Control", "Patient"))
design<-model.matrix(~Treat)
rownames(design)<-colnames(expressionMatrix)
design

#Conduct statistical analysis of expression
library(limma)
fit<-lmFit(expressionMatrix, design) #linear model fit
fit<-eBayes(fit)
result<-topTable(fit, coef="TreatPatient", adjust="BH", number=nrow(expressionMatrix)) #"BH" adjust for
#toptable normally takes top number but this takes all


result$"ProbeSetID"<-rownames(result)
head(result$"ProbeSetID")
result$"Fold Change"<-2^result$logFC
result$"Fold Change"[result$"Fold Change"<1]<-(-1)/result$"Fold Change"[result$"Fold Change"<1] #conver
expressionLinear<-as.data.frame(2^expressionMatrix)
expressionLinear$ProbeSetID<-rownames(expressionLinear)
result<-merge(result, expressionLinear, by.x="ProbeSetID", by.y="ProbeSetID") #merge values into one ar
result<-merge(annotation, result, by.x="Probe.Set.ID", by.y="ProbeSetID")
#result<-merge(result, callMatrixAll, by.x="Probe_Set_ID", by.y="ProbeSetID")
result<-merge(result, countPdf, by.x="Probe.Set.ID", by.y="ProbeSetID")
```

```
# # write.table(result, file=paste(analysis.name, "anno RMA limma na35.txt", sep=""), sep="\t", row.nam

# result<-subset(result, Gene.Symbol!="")
# result<-subset(result, subset=(countP>2))
# nrow(result)
# foldchange<-1.5
# pvalue<-0.05
# #adj_P_Val<-0.05
# siggenes<-subset(result, subset=(P.Value < pvalue) & abs(logFC) > log2(foldchange))
# #siggenes<-subset(result, subset=(adj.P.Val < 0.05))
# nrow(siggenes)
# siggenesup<-subset(siggenes, subset= logFC > 0)
# siggenesdown<-subset(siggenes, subset=logFC < 0)
# colnames(siggenesup)
# nrow(siggenesup)
# nrow(siggenesdown)
# UpandDown<-intersect(siggenesup$"Gene.Symbol", siggenesdown$"Gene.Symbol")
# length(UpandDown)
#
# UporDown<-subset(siggenes, subset=(!siggenes$"Gene.Symbol"%in% UpandDown))
# upsiggenes<-subset(siggenesup, subset=(!siggenesup$"Gene.Symbol"%in% UpandDown))
# downsiggenes<-subset(siggenesdown, subset=(!siggenesdown$"Gene.Symbol"%in% UpandDown))
# length(unique(siggenes$"Gene.Symbol"))
# uniquesiggenes <- unique(siggenes$Gene.Symbol)
# length(unique(upsiggenes$"Gene.Symbol"))
# length(unique(downsiggenes$"Gene.Symbol"))
# length(unique(UporDown$"Gene.Symbol"))


###Write results to CSV files for consensus analysis
setwd("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/DE Genes/TopGenesb_2016-02-09")
#dir.create(paste("TopGenesb", Sys.Date(), sep = "_")) #create directory using the day's date
#Take results, remove duplicate rows for genes, order by adjusted p value and take top X number of gene
uniqueresult <- result[!duplicated(result[,15]),]

#For ordering by adjusted p value

genesort <- uniqueresult[order(uniqueresult$adj.P.Val),]
topgene <- genesort[1:1000,]
write.csv(x = topgene, file = "FTLD_ap_1000")
topgene <- genesort[1:2000,]
write.csv(x = topgene, file = "FTLD_ap_2000")
topgene <- genesort[1:3000,]
write.csv(x = topgene, file = "FTLD_ap_3000")
topgene <- genesort[1:4000,]
write.csv(x = topgene, file = "FTLD_ap_4000")
topgene <- genesort[1:5000,]
write.csv(x = topgene, file = "FTLD_ap_5000")

#For ordering by fold change
# genesort <- uniqueresult[order(uniqueresult$`Fold Change`),]
# topgene <- genesort[1:500,]
#
```

```
# genesort <- uniqueresult[order(-uniqueresult$`Fold Change`),]
# botgene <- genesort[1:500,]
#
# topFC <- rbind(topgene,botgene)
#
# write.csv(x = topgene, file = "VCP_fc_1000")

# dir.create(paste("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43 Data Sets/DE Genes/CHMP2B/unique"))
# setwd("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43 Data Sets/DE Genes/CHMP2B/unique")
#
# write.table(siggenes, file=paste(analysis.name," RMA limma siggenes biomart p_", pvalue, "_fold chang
# write.table(upsiggenes, file=paste(analysis.name," RMA limma upsiggenes biomart p_", pvalue, "_fold c
# write.table(downsiggenes, file=paste(analysis.name," RMA limma downsiggenes biomart p_", pvalue, "_fo
# write.table(uniquesiggenes, file=paste(analysis.name,"unique RMA limma siggenes biomart p_", pvalue,
```

From this I generated the following results

1000

0

2000

0

3000

0

4000

MAP4K4

5000

MRPL23 MAP4K4 METTL17

Wednesday

As a result, I went back to the closest backup of the file I had (Dec 16th 2015) and the only discernible difference in Wenbin DE Gene.R was that I had commented out two lines that may be important.

```
result<-subset(result, Gene.Symbol!="") #removes any probes for which there are no gene symbols
result<-subset(result, subset=(countP>2)) #only takes results that have at least 2 samples with a prese
```

There were no differences to the consensus script. I then re-calculated expression and consensus was identical to that generated in Febrary from the December results.

Talking to Wenbin

In my discussion with Wenbin, we talked first about me removing duplicated genes before taking the top X DEGs. He at first was sceptical about removing duplicates because you couldn't be sure that you weren't left with 'bad' probes, but I explained that the only probes accepted at this point were those with 3 or more samples with a presence call (which suggests good quality).

Next, we discussed again the difference between ranking by p value or fold change, but we came to the conclusion that it's up to the individual and the project. I'm going to continue my experiment if only out of interest.

When it came to validating the results of the consensus DEGs, Wenbin said that both random permutations test and Fisher's exact test could be ways to approach it. For the RPT, I wanted his opinion on what my population pool should be for selecting random genes. We settled on using the annotation file, removing blanks, removing duplicates and also removing any probes noted to bind to negative strand matching probes. NSMPs are expressed sequence tags that have been designed in the wrong direction to the gene and so are antisense.

The random permutation test including these changes are below. Note that I have used the annotation files from both the U133 plus 2.0 and U133A to accurately represent the fact we have 4 data sets from the former and 1 from the latter.

```
## Save annotation file locations to variable
annotation.U133plus2 <- "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/HG-U133_Plus_2.na35.annot.
annotation.U133A2 <- "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/HG-U133A_2.na35.annot.csv/HG-U

# Read in annotation files
annotationU133plus2 <- read.table(annotation.U133plus2, header = TRUE,
    row.names = NULL, sep = "\t", skip = 0, stringsAsFactors = F,
    quote = "", comment.char = "!", fill = TRUE)
annotationU133A2 <- read.table(annotation.U133A2, header = TRUE,
    row.names = NULL, sep = "\t", skip = 0, stringsAsFactors = F,
    quote = "", comment.char = "!", fill = TRUE)

# Remove rows in which gene symbol is absent
annotationU133plus2 <- subset(annotationU133plus2, subset = (Gene.Symbol !=
    "---"))  #if no gene symbol, discount
annotationU133A2 <- subset(annotationU133A2, subset = (Gene.Symbol !=
    "---"))  #if no gene symbol, discount

# Remove rows in which gene symbol is duplicated
annotationU133plus2 <- annotationU133plus2[!duplicated(annotationU133plus2[,
    15]), ]
annotationU133A2 <- annotationU133A2[!duplicated(annotationU133A2[,
    15]), ]


# Remove rows in which genes are noted to have negative
# strand matching probes
idxNegativeStrand <- grep("Negative Strand Matching Probes",
    annotationU133plus2$Annotation.Notes)
if (length(idxNegativeStrand) > 0) {
    annotationU133plus2 <- annotationU133plus2[-idxNegativeStrand,
        ]
}
nrow(annotationU133plus2)
```

```
idxNegativeStrand <- grep("Negative Strand Matching Probes",
    annotationU133A2$Annotation.Notes)
if (length(idxNegativeStrand) > 0) {
    annotationU133A2 <- annotationU133A2[-idxNegativeStrand,
        ]
}
nrow(annotationU133A2)


# indicate the number of overlapping genes identified by DE
# analysis
test <- 50

m = 10000  #number of repetitions
r <- c(1:m)  #store repetition numbers in vector 'r'

for (j in 1:m) {
    random1 <- sample(annotationU133plus2$Gene.Symbol, size = 4000,
        replace = F)
    random2 <- sample(annotationU133plus2$Gene.Symbol, size = 4000,
        replace = F)
    random3 <- sample(annotationU133plus2$Gene.Symbol, size = 4000,
        replace = F)
    random4 <- sample(annotationU133plus2$Gene.Symbol, size = 4000,
        replace = F)
    random5 <- sample(annotationU133A2$Gene.Symbol, size = 4000,
        replace = F)
    random <- Reduce(intersect, list(random1, random2, random3,
        random4, random5))
    r[j] <- length(random)
}

test1 <- which(r > test)  # count number of times r is larger than test value
result <- (length(test1)/m)  # calculate P value
```

For both 16 (consensus value for top 3000) and 50 (consensus value for top 4000), $p<0.01$

Next, I conducted the same analysis but using the most extreme fold change values. FC was sorted by number and the top and bottom genes were selected to create 1000, 2000, 3000, and 4000 gene sets. These were analysed for common genes.

1000

NOV

2000

MCTP1 TUBB4B NOV HN1 NAP1L2 TSPAN13 TMEM255A DMD RGS2 CDK5R1 HBA1 /// HBA2 ZIC1

3000

MCTP1 TUBB4B NOV SPP1 HN1 UPF3A NAP1L2 TSPAN13 TMEM255A DMD ESRRG RGS2 CDK5R1 ATP6V1A PVALB RTN1 NPTX2 COL4A1 HBA1 /// HBA2 ITM2A TJP2 RND3 ZIC1 MAN2A1 DCN ATP2A2 ADAMTS1 MAP1A

4000

MCTP1 EIF1AY TUBB3 TUBB4B NOV RNF13 SPP1 YWHAH WDR47 MLF2 NEK7 KIF3A HN1 NMRK1 UPF3A GSTO1 CCT2 NAP1L2 TMEM97 PHACTR2 TSPAN13 TMEM255A DMD ESRRG RGS2 CDK5R1 ATP6V1A STAT3 BTN3A2 /// BTN3A3 PVALB ZFPM2 P4HTM VDAC3 SLC25A12 RAB15 RTN1 NQO2 IL6ST PEG3 LSM5 UCHL1 KIFAP3 ZNF271 NUPL1 CEBPD NPTX2 IFIT3 COL4A1 HBA1 /// HBA2 RGCC DDIT4 PLEKHB1 ITM2A CNKSR2 ORAI2 ADARB1 MAN2A1 MAP1A CSTA DCN SLC14A1 ATP2A2 EFS TJP2 ADAMTS1 SCN1B RND3 KATNBL1 ETNK1 ZIC1 TRIM22 WSB2

The first thing to note is that when using fold change to rank genes, there are more common genes. The second thing to note is that the genes TUBB3, RNF13, UPF3A, GSTO1, CCT2, DMD, and CDK5R1 are common to both lists (top 3000 ap, top 4000 fc). However, one has to bear in mind that taking equal quantities of up and downregulated genes does not necessarily give the full picture. For example, a diseased condition may have proportionally more upregulation of genes than down, but by taking 500 from each end you are treating them as equal. By using adjusted p value, you are taking the most DE regardless of direction, which may be more useful.

Thursday

Now I have a list of 50 DEGs associated with TDP-43 pathology, I wanted to find out a little bit more about them. I first used WebGestalt to get some information on the genes themselves.

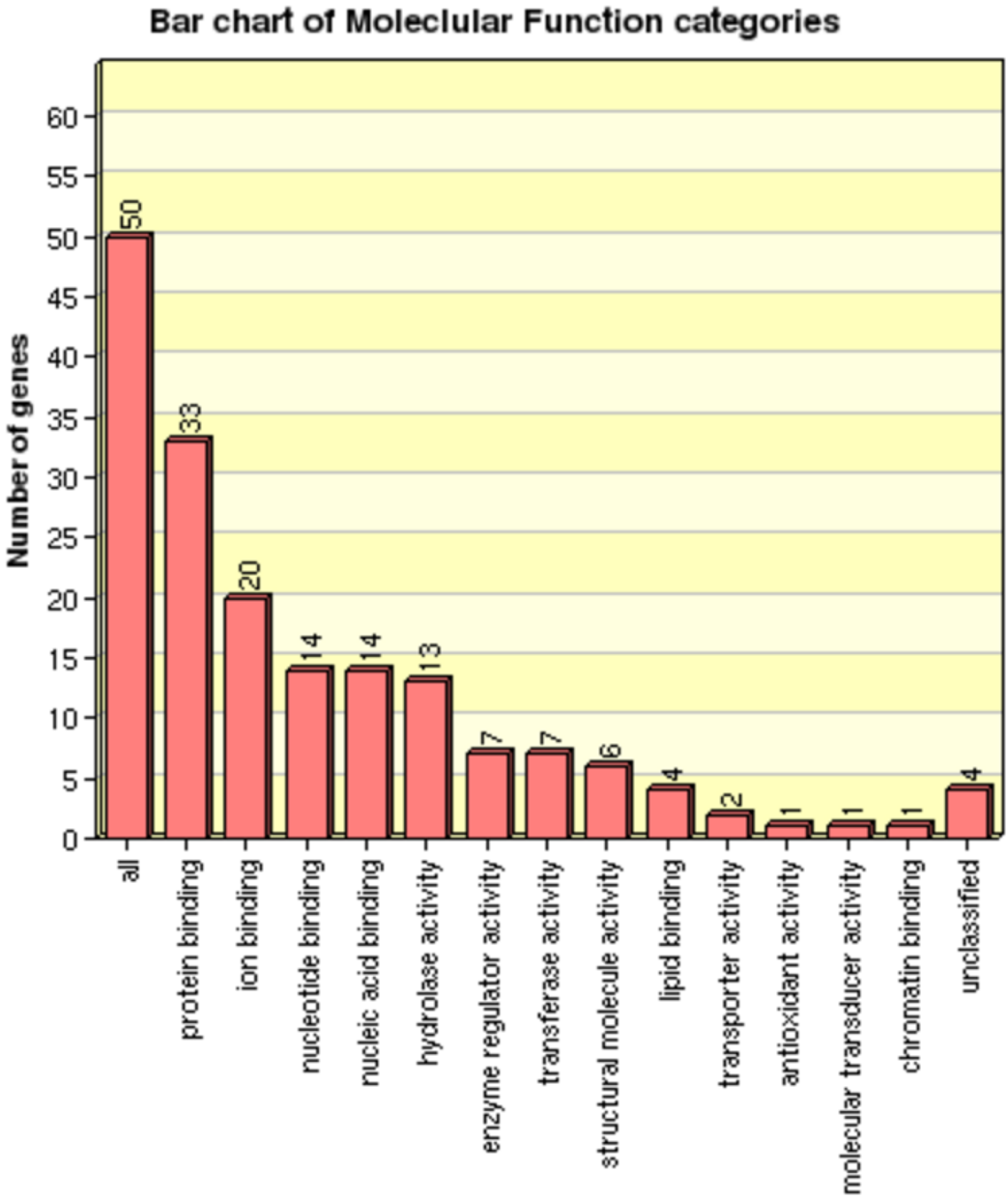
4000	Full Name
DDX39B	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39B
BPTF	bromodomain PHD finger transcription factor
BRD3	bromodomain containing 3
C14orf1	chromosome 14 open reading frame 1
C18orf32	chromosome 18 open reading frame 32
CCT2	chaperonin containing TCP1, subunit 2 (beta)
CDK5R1	cyclin-dependent kinase 5, regulatory subunit 1 (p35)
CSRP1	cysteine and glycine-rich protein 1
CST3	cystatin C
DCN	decorin
DDX39A	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39A
DDX5	DEAD (Asp-Glu-Ala-Asp) box helicase 5
DMD	dystrophin
DYNLT1	dynein, light chain, Tctex-type 1
EEF1A1	eukaryotic translation elongation factor 1 alpha 1
ETS2	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)
FN1	fibronectin 1
GNPAT	glyceronephosphate O-acyltransferase
GSTO1	glutathione S-transferase omega 1
GTF2I	general transcription factor IIi
HSD17B4	hydroxysteroid (17-beta) dehydrogenase 4
ICMT	isoprenylcysteine carboxyl methyltransferase
LBR	lamin B receptor

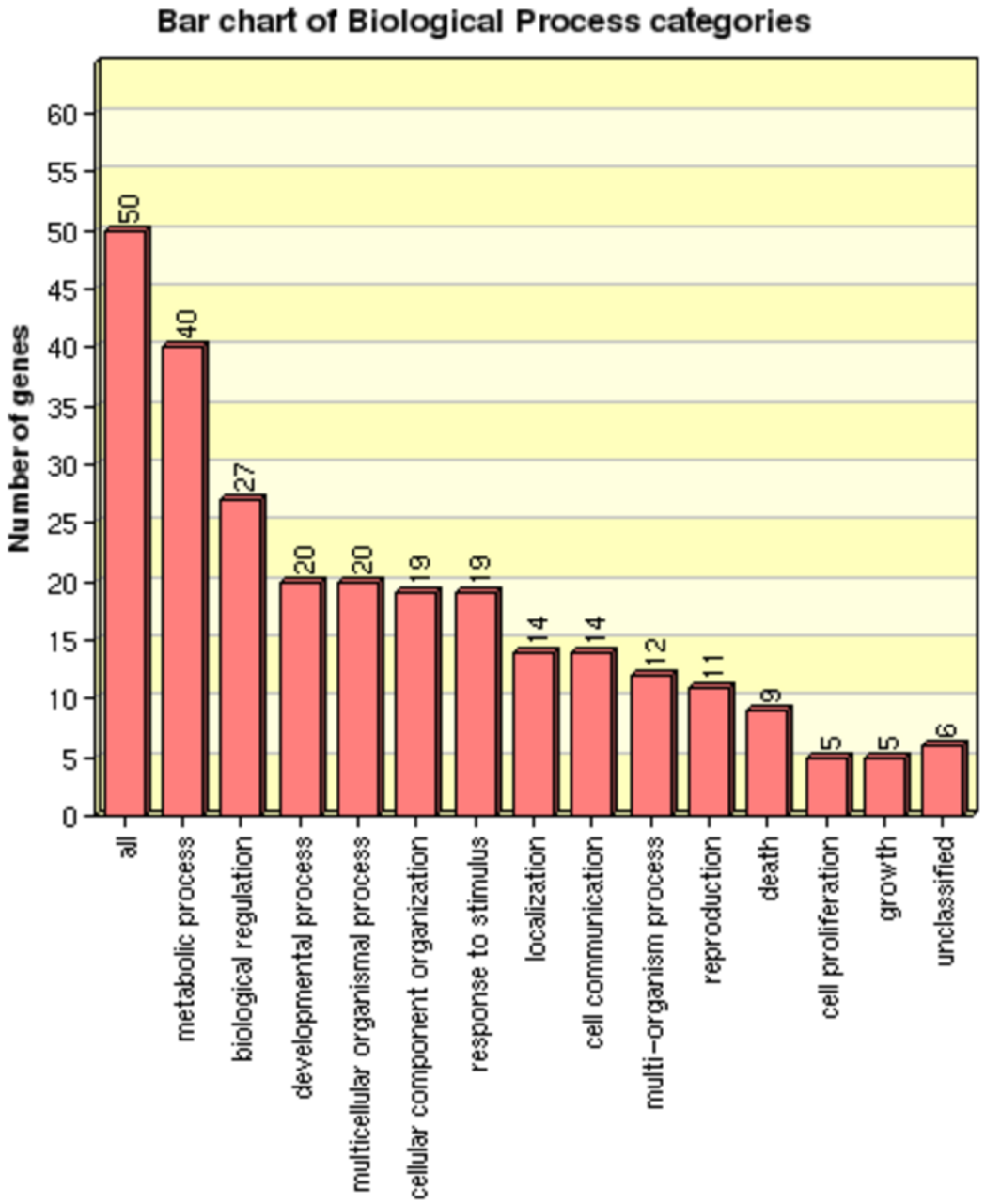
4000	Full Name
MAP4K4	mitogen-activated protein kinase kinase kinase kinase 4
MPHOSPH9	M-phase phosphoprotein 9
MTMR9	myotubularin related protein 9
NAGA	N-acetylgalactosaminidase, alpha-
NKTR	natural killer-tumor recognition sequence
NUTF2	nuclear transport factor 2
PAICS	phosphoribosylaminoimidazole carboxylase
PFDN1	prefoldin subunit 1
PRKD1	protein kinase D1
PSAP	prosaposin
PSMD1	proteasome (prosome, macropain) 26S subunit, non-ATPase, 1
PTEN	phosphatase and tensin homolog
RECQL	RecQ protein-like (DNA helicase Q1-like)
RNF13	ring finger protein 13
RNF130	ring finger protein 130
RPL6	ribosomal protein L6
RPLP2	ribosomal protein, large, P2
RPS6	ribosomal protein S6
TARDBP	TAR DNA binding protein
TARS	threonyl-tRNA synthetase
TCF4	transcription factor 4
TMEM59	transmembrane protein 59
TRO	trophinin
TUBB3	tubulin, beta 3 class III
UPF3A	UPF3 regulator of nonsense transcripts homolog A
WBSCR22	Williams Beuren syndrome chromosome region 22
WDR78	WD repeat domain 78

4000	Associated Diseases (WebGestalt)	Associated Drugs (WebGestalt)
DDX39B		Dactinomycin, Adenosine
BPTF	Dementia, Alzheimer's Disease	Adenosine
BRD3		
C14orf1		
C18orf32		
CCT2		
CDK5R1	Dementia, Alzheimer's Disease, Brain Diseases, Tauopathies	
CSRP1		
CST3	Dementia, Alzheimer's Disease, Brain Diseases, Tauopathies	
DCN		
DDX39A		Dactinomycin
DDX5		Dactinomycin, Adenosine
DMD	Mental Retardation, Monosomy	
DYNLT1		
EEF1A1		
ETS2	Chordoma, Mental Retardation	
FN1		
GNPAT	Mental Retardation	
GSTO1	Dementia, Alzheimer's Disease, Brain Diseases, Tauopathies	
GTF2I	Breast/Ovarian Cancer Syndrome, Mental Retardation, Monosomy	
HSD17B4	Brain Diseases	
ICMT		Adenosine

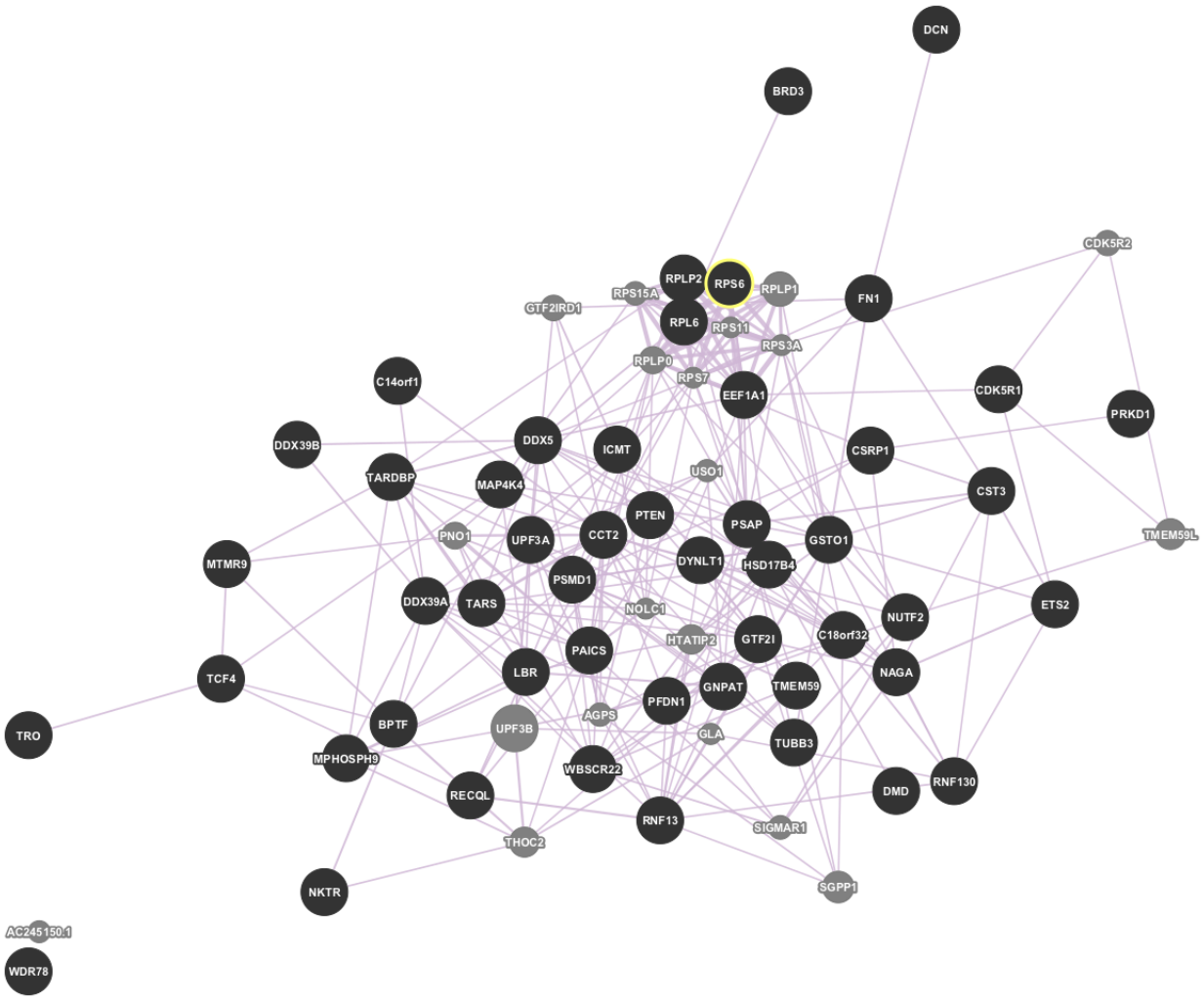
4000	Associated Diseases (WebGestalt)	Associated Drugs (WebGestalt)
LBR		
MAP4K4		
MPHOSPH9		
MTMR9		
NAGA	Brain Diseases, Sandhoff Disease	
NKTR		
NUTF2		
PAICS		
PFDN1		
PRKD1		
PSAP	Brain Diseases, Sandhoff Disease	
PSMD1		Adenosine
PTEN	Chordoma, Breast/Ovarian Cancer Syndrome	
RECQL		Adenosine
RNF13		
RNF130		
RPL6		Dactinomycin
RPLP2		
RPS6		
TARDBP	Dementia, Alzheimer's Disease, Brain Diseases, Tauopathies	
TARS		
TCF4	Mental Retardation, Monosomy	
TMEM59		
TRO		
TUBB3		
UPF3A		
WBSCR22	Mental Retardation, Monosomy	
WDR78		

I also generated graphs showing GO biological processes and molecular functions


Bar chart of Molecular Function categories
Number of genes
0
5
10
15
20
25
30
35
40
45
50
55
60
50
33
20
14
14
13
7
7
6
4
2
1
1
1
4
all
protein binding
ion binding
nucleotide binding
nucleic acid binding
hydrolase activity
enzyme regulator activity
transferase activity
structural molecule activity
lipid binding
transporter activity
antioxidant activity
molecular transducer activity
chromatin binding
unclassified



The next thing I did was input the genes into GeneMANIA to look at the co-expression network.



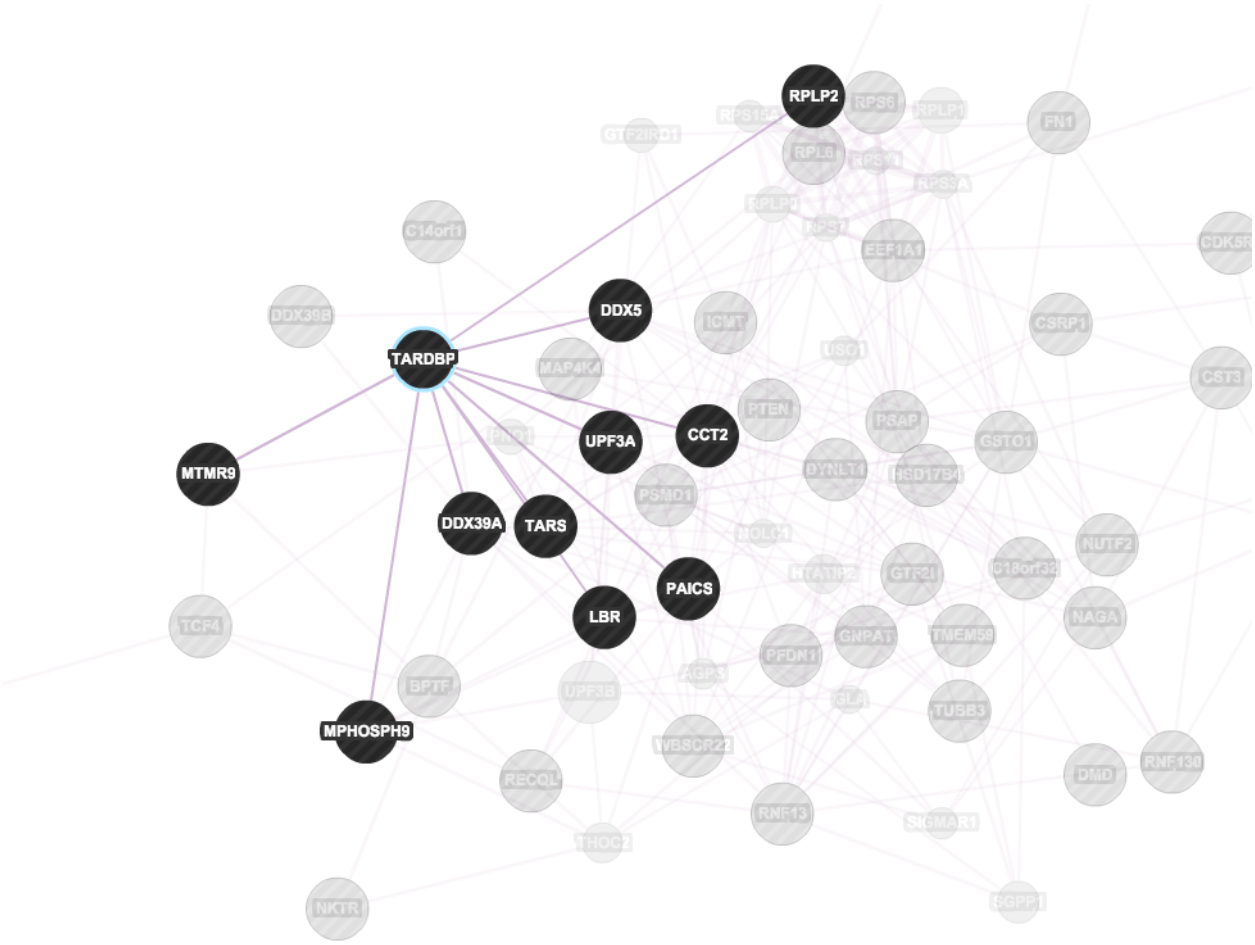
There is a group of 9 genes (RPS15A, RPLP2, RPS6, RPLP1, RPL6, RPS11, RPS3A, RPLP0 and RPS7) that is very tightly connected. These genes are found in the top 29 functions listed in the table below. In one of these categories, viral transcription, these RPLP2 is significantly co-expressed with TARDBP.

Feature	FDR	Genes in network	Genes in genome
viral life cycle	4.32E-09	12	159
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	4.32E-09	11	116
translational elongation	2.10E-08	10	102
cytosolic ribosome	1.37E-07	9	90
translational termination	1.37E-07	9	89
nuclear-transcribed mRNA catabolic process	1.37E-07	11	179
viral gene expression	1.45E-07	9	94
mRNA catabolic process	1.96E-07	11	190
multi-organism metabolic process	2.84E-07	9	104
SRP-dependent cotranslational protein targeting to membrane	2.97E-07	9	107
protein targeting to membrane	2.97E-07	10	157
protein targeting to ER	2.97E-07	9	108
cotranslational protein targeting to membrane	2.97E-07	9	108
establishment of protein localization to endoplasmic reticulum	2.97E-07	9	109
single-organism cellular localization	2.97E-07	12	272
single-organism localization	2.97E-07	12	272
viral transcription	3.16E-07	10	159
RNA catabolic process	3.28E-07	11	215
protein localization to endoplasmic reticulum	4.88E-07	9	120
ribosomal subunit	5.00E-07	9	121
cellular protein complex disassembly	1.27E-06	9	135
establishment of protein localization to membrane	1.50E-06	11	253
translational initiation	2.94E-06	9	150
cytosolic part	2.99E-06	9	151
ribosome	3.61E-06	9	155
protein complex disassembly	3.67E-06	9	156
macromolecular complex disassembly	5.20E-06	9	163
protein localization to membrane	6.08E-06	10	227
protein targeting	7.89E-05	10	299
cytosolic small ribosomal subunit	1.20E-04	5	36
structural constituent of ribosome	8.64E-04	6	97
small ribosomal subunit	9.85E-04	5	55
cytosolic large ribosomal subunit	1.97E-02	4	51
sphingolipid metabolic process	1.97E-02	5	102
nuclear export	3.09E-02	5	113
nucleocytoplasmic transport	3.51E-02	7	278
nuclear transport	3.66E-02	7	281
large ribosomal subunit	5.18E-02	4	67
mRNA export from nucleus	5.35E-02	4	68
mRNA transport	5.52E-02	4	69
membrane lipid metabolic process	7.27E-02	5	140
RNA export from nucleus	7.66E-02	4	76
regulation of neurogenesis	7.75E-02	6	227
multi-organism intracellular transport	8.61E-02	3	32
multi-organism transport	8.61E-02	3	32
intracellular transport of virus	8.61E-02	3	32
transport of virus	8.61E-02	3	32
microbody lumen	9.06E-02	3	33
peroxisomal matrix	9.06E-02	3	33
establishment of RNA localization	9.35E-02	4	85
nucleic acid transport	9.35E-02	4	85
RNA transport	9.35E-02	4	85
RNA localization	9.35E-02	4	84

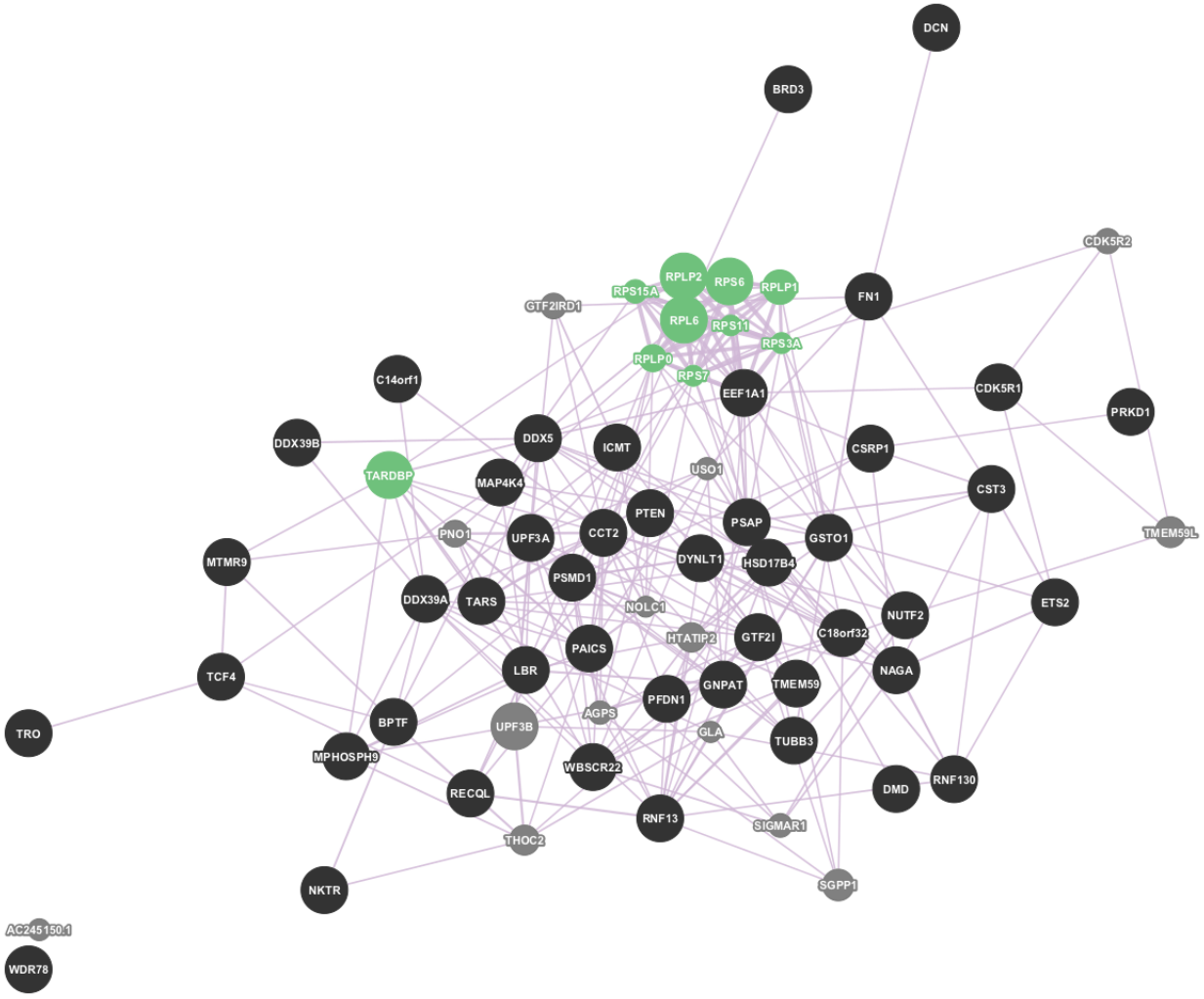
List of functions

4000	Gene Mania Functions
BPTF	
BRD3	
C14orf1	
C18orf32	
CCT2	
CDK5R1	Regulation of neurogenesis
CSRP1	
CST3	
DCN	
DDX39A	Nuclear export, nucleocytoplasmic transport, nuclear transport, mRNA transport from nucleus, mRNA transport, RNA transport from nucleus, establishment of RNA localization, nucleic acid transport, RNA transport, RNA localization
DDX39B	Viral life cycle, nuclear export, nucleocytoplasmic transport, nuclear transport, mRNA transport from nucleus, mRNA transport, multi-organism intracellular transport, multi-organism transport, intracellular transport of virus, transport of virus, establishment of RNA localisation, nucleic acid transport, RNA transport, RNA localisation
DDX5	
DMD	Regulation of neurogenesis
DYNLT1	Viral life cycle, single-organism cellular localization, single-organism localization, regulation of neurogenesis, multi-organism intracellular transport, multi-organism transport, intracellular transport of virus, transport of virus
EEF1A1	Transaltional elongation
ETS2	
FN1	
GNPAT	Microbody lumen, peroxisomal matrix
GSTO1	
GTF2I	
HSD17B4	Microbody lumen, peroxisomal matrix
ICMT	Protein targeting to membrane, single-organism cellular localization, single-organism localization, establishment of protein localization to membrane, protein localization to membrane, protein targeting
LBR	
MAP4K4	
MPHOSPH9	
MTMR9	
NAGA	Sphingoloid metabolic process, membrane lipid metabolic process
NKTR	
NUTF2	Nuclear export, nucleocytoplasmic transport, nuclear transport
PAICS	
PFDN1	
PRKD1	Sphingoloid metabolic process, membrane lipid metabolic process, regulation of neurogenesis
PSAP	Sphingoloid metabolic process, membrane lipid metabolic process
PSMD1	
PTEN	Single-organism cellular localization, single-organism localization, regulation of neurogenesis
RECQL	
RNF13	
RNF130	
RPL6	Viral life cycle, nuclear-transcribed mRNA catabolic process-nonsense-mediated delay, translational elongation, cytosolic ribosome, translational termination, nuclear transcribed mRNA catabolic process, viral gene expression, mRNA catabolic process, multi-organism metabolic process, SRP-dependent cotranslational protein targeting to membrane, protein targeting to membrane, protein targeting to ER, cotranslational protein targeting to membrane, establishment of protein localization to endoplasmic reticulum, single-organism cellular localization, single organism localisation, viral transcription, RNA catabolic process, protein localisation to endoplasmic reticulum, ribosomal subunit, establishment of protein localisation to membrane, translational initiation, cytosolic part, ribosome, protein complex disassembly, macromolecular complex disassembly, protein localization to membrane, protein targeting, **structural constituent of ribosome, cytosolic large ribosomal subunit, large ribosomal subunit**
RPLP2	Viral life cycle, nuclear-transcribed mRNA catabolic process-nonsense-mediated delay, translational elongation, cytosolic ribosome, translational termination, nuclear transcribed mRNA catabolic process, viral gene expression, mRNA catabolic process, multi-organism metabolic process, SRP-dependent cotranslational protein targeting to membrane, protein targeting to membrane, protein targeting to ER, cotranslational protein targeting to membrane, establishment of protein localization to endoplasmic reticulum, single-organism cellular localization, single organism localisation, viral transcription, RNA catabolic process, protein localisation to endoplasmic reticulum, ribosomal subunit, establishment of protein localisation to membrane, translational initiation, cytosolic part, ribosome, protein complex disassembly, macromolecular complex disassembly, protein localization to membrane, protein targeting, **structural constituent of ribosome, cytosolic large ribosomal subunit, large ribosomal subunit**
RPS6	Viral life cycle, nuclear-transcribed mRNA catabolic process-nonsense-mediated delay, translational elongation, cytosolic ribosome, translational termination, nuclear transcribed mRNA catabolic process, viral gene expression, mRNA catabolic process, multi-organism metabolic process, SRP-dependent cotranslational protein targeting to membrane, protein targeting to membrane, protein targeting to ER, cotranslational protein targeting to membrane, establishment of protein localization to endoplasmic reticulum, single-organism cellular localization, single organism localisation, viral transcription, RNA catabolic process, protein localisation to endoplasmic reticulum, ribosomal subunit, establishment of protein localisation to membrane, translational initiation, cytosolic part, ribosome, protein complex disassembly, macromolecular complex disassembly, protein localization to membrane, protein targeting, **cytosolic small ribosomal subunit, small ribosomal subunit**
TARDBP	Viral transcription
TARS	
TCF4	Regulation of neurogenesis
TMEM59	Establishment of protein localization to membrane
TRO	
TUBB3	
UPF3A	nuclear-transcribed mRNA catabolic process-nonsense-mediated delay, nuclear transcribed mRNA catabolic process, mRNA catabolic process, RNA catabolic process, nucleocytoplasmic transport, nuclear transport.
WBSCR22	
WDR78	

With genes

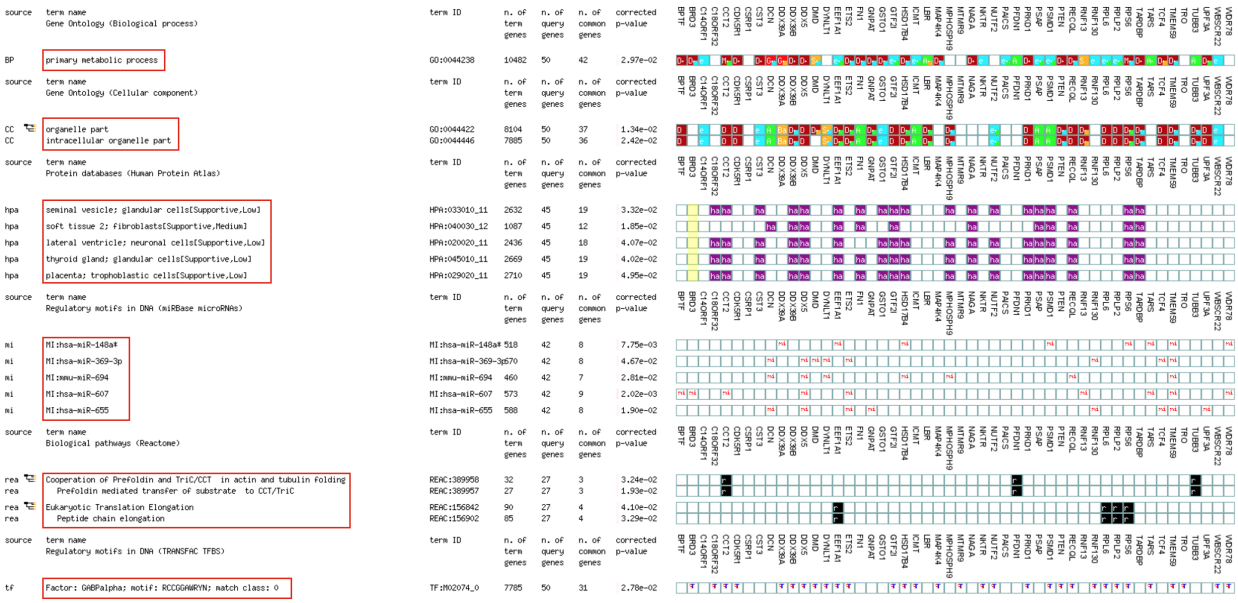


TARDBP coexp



Viral transcription

Next I looked at G:profiler. This produced a protein-protein interaction network using BioGRID. The high quality PDF can be found on basecamp or at /Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/GeneExpressionAnalysis/G.Profiler output. G:Profiler also produced results from multiple places including GO biological process and cellular component, as well as Human protein atlas and Reactome biological pathways.


source
term name
Gene Ontology (Biological process)
term ID
n. of term genes
n. of query genes
n. of common genes
corrected p-value
BP
primary metabolic process
GO:0044238
10482
50
42
2.97e-02
source
term name
Gene Ontology (Cellular component)
CC
organelle part
GO:0044422
8104
50
37
1.34e-02
CC
intracellular organelle part
GO:0044446
7885
50
36
2.42e-02
source
term name
Protein databases (Human Protein Atlas)
hpa
seminal vesicle; glandular cells[Supportive,Low]
HPA:033010_11
2632
45
19
3.32e-02
hpa
soft tissue 2; fibroblasts[Supportive,Medium]
HPA:040030_12
1087
45
12
1.85e-02
hpa
lateral ventricle; neuronal cells[Supportive,Low]
HPA:020020_11
2436
45
18
4.07e-02
hpa
thyroid gland; glandular cells[Supportive,Low]
HPA:045010_11
2669
45
19
4.02e-02
hpa
placenta; trophoblastic cells[Supportive,Low]
HPA:029020_11
2710
45
19
4.95e-02
source
term name
Regulatory motifs in DNA (miRBase microRNAs)
mi
MI:hsa-miR-148a*
MI:hsa-miR-148a*
518
42
8
7.75e-03
mi
MI:hsa-miR-369-3p
MI:hsa-miR-369-3p
570
42
8
4.67e-02
mi
MI:mmu-miR-694
MI:mmu-miR-694
460
42
7
2.81e-02
mi
MI:hsa-miR-607
MI:hsa-miR-607
573
42
9
2.02e-03
mi
MI:hsa-miR-655
MI:hsa-miR-655
588
42
8
1.90e-02
source
term name
Biological pathways (Reactome)
rea
Cooperation of Prefoldin and TriC/CCT in actin and tubulin Folding
REAC:389958
32
27
3
3.24e-02
rea
Prefoldin mediated transfer of substrate to CCT/TriC
REAC:389957
27
27
3
1.93e-02
rea
Eukaryotic Translation Elongation
REAC:156842
90
27
4
4.10e-02
rea
Peptide chain elongation
REAC:156902
85
27
4
3.29e-02
source
term name
Regulatory motifs in DNA (TRANSFAC TFBS)
tf
Factor: GABPalpha; motif: RCCGGAWRYN; match class: 0
TF:M02074_0
7785
50
31
2.78e-02