LabBook_17_03_2017

Claire Green

```
## Warning: package 'knitr' was built under R version 3.3.2
```

Monday

Today I spent time doing a little reading and writing from home. I identified a paper that has also done a similar analysis to my own that was submitted to the journal Bioinformatics.

https://academic.oup.com/bioinformatics/article/25/7/875/210443/Meta-analysis-of-age-related-gene-expression

Tuesday

Today I worked on the threshold problem for my common DEGs. I set up a script that is able to take every threshold from 1-8000 top DEGs from each dataset, run hypergeometric testing of my benchmark lists and extract the adjusted p value for each. I hoped to find a threshold which best identified the highest enrichments for each dataset, but Win suggested why don't I just make one list with all the genes I think are important and do a single enrichment analysis on that. I will have to consider which genes are important to include.

Wednesday

Spent all day on this sodding med school research day abstract.

Thursday

On Thursday I got back to work (after the sofa fiasco) and I decided to start with writing out my idea of using the variance of a gene across datasets as a measure of proximity to phenotype. This can be found in my google drive. What I realised as I wrote it out was that my methodology had one giant assumption - that you know which DEGs are upstream and which are downstream of the phenotype. Since I don't yet have a way of discriminating that, I decided to put it aside for a later date. It's worth discussing theoretically though, this is what I wrote so far:

*Methodologies such as differential expression analysis of genes are proficient at providing lists of genes we suspect are related to a particular phenotype. What it cannot do, however, is provide a metric of the degree to which each gene is related to the phenotype. This concept has led to increased research in target prioritisation for drug development, in which a list of potential targets must be assessed for the likelihood that they contribute a real effect on the phenotype in questions. This process is laborious and expensive if you have to consider every DEG, so in silico prioritisation is key. Current methods, such as Varelect (http://varelect.genecards.org) and Endeavour (https://endeavour.esat.kuleuven.be/), draw upon existing literature and experimental data to generate an in-house scoring system. However, these systems are based upon information that is globally generated, rather than locally identified from the experimenter's own data.

When biological samples are all deemed to have a shared phenotype, we make an assumption that there is common biochemistry leading to that phenotype. However, in diseases where different mutations can cause the same phenotype, it is much harder to determine the point of convergence - where disparate mutations

meet to cause common phenotype. This convergence can be identified through identifying which DEGs are common to all samples. For example,

In a scenario where we have a list of genes we suspect to be upstream, it would be useful to be able to position each DEG along an axis starting at mutation, and ending in phenotype. In other words, which genes are *

It's obviously unfinished but you get the gist.

Then I looked into identifying which genes were commonly upregulated and which commonly downregulated across all the datasets. This script is called updownregulatedgenes.R

```
setwd("/users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/noMedian/")

thresh <- 5996

C9 <- read.csv("C9_unique.csv")
C9 <- C9[order(C9$P.Value),]
C96500 <- C9[1:thresh,]
CH <- read.csv("CH_unique.csv")
CH <- CH[order(CH$P.Value),]
CH6500 <- CH[1:thresh,]
sals <- read.csv("sals_unique.csv")
sals <- sals[order(sals$P.Value),]
sals6500 <- sals[1:thresh,]
ftld <- read.csv("ftld_unique.csv")
ftld <- ftld[order(ftld$P.Value),]
ftld6500 <- ftld[1:thresh,]
vcp <- read.csv("vcp_unique.csv")
vcp <- vcp[order(vcp$P.Value),]
vcp6500 <- vcp[1:thresh,]

setwd("/users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/TDP-43_DEseq2/")
pet <- read.csv("PET_results_keepfiltering.csv")
pet <- pet[!duplicated(pet$hgnc_symbol),]
pet6500 <- pet[1:thresh,]
rav <- read.csv("RAV_results_keepfiltering.csv")
rav <- rav[!duplicated(rav$hgnc_symbol),]
rav6500 <- rav[1:thresh,]



C9up <-subset(C96500, subset=(logFC > 0))
C9upgene <- C9up$Gene.Symbol
C9down <-subset(C96500, subset=(logFC < 0))
C9downgene <- C9down$Gene.Symbol

CHup <-subset(CH6500, subset=(logFC > 0))
CHupgene <- CHup$Gene.Symbol
CHdown <-subset(CH6500, subset=(logFC < 0))
CHdowngene <- CHdown$Gene.Symbol

salsup <-subset(sals6500, subset=(logFC > 0))
salsupgene <- salsup$Gene.Symbol
salsdown <-subset(sals6500, subset=(logFC < 0))
salsdowngene <- salsdown$Gene.Symbol
```

```
ftldup <-subset(ftld6500, subset=(logFC > 0))
ftldupgene <- ftldup$Gene.Symbol
ftlddown <-subset(ftld6500, subset=(logFC < 0))
ftlddowngene <- ftlddown$Gene.Symbol

vcpup <-subset(vcp6500, subset=(logFC > 0))
vcpupgene <- vcpup$Gene.Symbol
vcpdown <-subset(vcp6500, subset=(logFC < 0))
vcpdowngene <- vcpdown$Gene.Symbol

petup <-subset(pet6500, subset=(log2FoldChange > 0))
petupgene <- petup$hgnc_symbol
petdown <-subset(pet6500, subset=(log2FoldChange < 0))
petdowngene <- petdown$hgnc_symbol

ravup <-subset(rav6500, subset=(log2FoldChange > 0))
ravupgene <- ravup$hgnc_symbol
ravdown <-subset(rav6500, subset=(log2FoldChange < 0))
ravdowngene <- ravdown$hgnc_symbol

intersect_up <- Reduce(intersect, list(C9upgene, CHupgene, salsupgene, ftldupgene, vcpupgene, petupgene
intersect_up
intersect_down <- Reduce(intersect, list(C9downgene, CHdowngene, salsdowngene, ftlddowngene, vcpdowngen
intersect_down

write.table(intersect_up, "intersect_up.txt", quote = F, col.names = F, row.names = F)
write.table(intersect_down, "intersect_down.txt", quote = F, col.names = F, row.names = F)
```

I will report the results in a moment as I need to explain how I got round to picking my final threshold, as this changes the results of a bunch of things.

Friday

I looked into making my ULTIMATE LIST. I decided to include the gene lists from Malacards ALS, AD and parkinsons, ALSOD, the two TDP-43 PPI lists and John's TDP-43 pathology tracking module. This provided a total list of 820 unique genes.

I used the following for loop to conduct fisher's exact test on each threshold up to 8000 genes:

```
library(hgu133plus2.db)

setwd("/users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/noMedian/")

C9 <- read.csv("C9_unique.csv")
C9 <- C9[order(C9$P.Value),]
CH <- read.csv("CH_unique.csv")
CH <- CH[order(CH$P.Value),]
sals <- read.csv("sals_unique.csv")
sals <- sals[order(sals$P.Value),]
ftld <- read.csv("ftld_unique.csv")
ftld <- ftld[order(ftld$P.Value),]
vcp <- read.csv("vcp_unique.csv")
vcp <- vcp[order(vcp$P.Value),]
```

```r
setwd("/users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/TDP-43_DEseq2/")

pet <- read.csv("PET_results_keepfiltering.csv")
pet <- pet[!duplicated(pet$hgnc_symbol),]
rav <- read.csv("RAV_results_keepfiltering.csv")
rav <- rav[!duplicated(rav$hgnc_symbol),]

## extract gene lists
c9_gene <- C9$Gene.Symbol
ch_gene <- CH$Gene.Symbol
sals_gene <- sals$Gene.Symbol
ftld_gene <- ftld$Gene.Symbol
vcp_gene <- vcp$Gene.Symbol
pet_gene <- pet$hgnc_symbol
rav_gene <- rav$hgnc_symbol

# num_overlap <- matrix(data=NA)
List <- list()

for (i in 1:8000){
  C9_int <- c9_gene[1:i]
  CH_int <- ch_gene[1:i]
  sals_int <- sals_gene[1:i]
  ftld_int <- ftld_gene[1:i]
  vcp_int <- vcp_gene[1:i]
  pet_int <- pet_gene[1:i]
  rav_int <- rav_gene[1:i]
  List[[i]] <- Reduce(intersect, list(C9_int, CH_int, sals_int, ftld_int, vcp_int, pet_int, rav_int))
}

#Load file with all genes
sym <- hgu133plus2SYMBOL
sym1 <- mappedkeys(sym)
sym2 <- as.list (sym[c(sym1)])
sym3 <- data.frame (sym2)
sym.probes <- names (sym2)
sym.genes <- sym3[1,]
sym.genes <- t(sym.genes)
allgenes <- sym.genes[!duplicated(sym.genes),]

#Remove list elements with less than 5 genes (to aid calculations)
List_5 <- List[lengths(List) > 4]
#Leaves final 5087 elements (elements 1:2913 removed)

#Create new empty list
enrich_result <- list()

setwd(dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/")
S <- read.table(file = "OneBenchmarkList.txt")
s <- S$V1


for (i in 1:length(List_5)){
```

```
  write.table(List_5[i], "benchmark_genelist.txt", quote = FALSE, row.names = FALSE, col.names = FALSE)
  bgene <- read.table("benchmark_genelist.txt")
  bgene <- bgene$V1

  ur.list <- bgene
  int.list <- s

  #How many test geneset genes contain snps
  x.in <- length (which(ur.list %in% int.list))

  #how many do not
  x.out <- length(ur.list) - x.in

  #total number of snp genes
  tot.in <- length(int.list)

  #total number of all genes
  tot.out <- length(allgenes)-length(tot.in)


  #create count matrix
  counts <- matrix (nrow=2, ncol=2)
  counts [1,] <- c(x.in, tot.in)
  counts [2,] <- c(x.out, tot.out)


  #Conduct fisher's exact test for count data
  a5 <-fisher.test (counts)

  enrich_result[[i]] <- a5$p.value
}

enrich_result_df <- t(as.data.frame(enrich_result))
rownames(enrich_result_df) <- (1:nrow(enrich_result_df))+2913
```

You have to add 2913 at the end so that you get the correct rowname for the value, as 2913 rows were removed before the analysis.

I removed the rows with less than 5 genes because it wouldn't affect the results and having zero or one genes was messing with the calculations.

In the end, results showed that the threshold with the highest enrichment value was 5996 (113 genes), with an enrichment of the ULTIMATE LIST of 1.979756e-10. After random permutation tests, this threshold with this number of genes was significant way beyond chance ($p < .0001$). The p value was was actually zero because the simulation never even got near 113, as the range was 27-79, with a mean value of 52.2726.

Enrichment results from EnrichR are as follows:

GO Biological Process

Term	Overlap	P-value	Adjusted P	Z-score	Combined Score	Genes
RNA splicing (GO:0008380)	12/313	2.11E-07	3.12E-04	-2.340	18.890	RBM39 PNN SFPQ RBFOX2 SCAF11 SREK1 TARDBP SF3B1 SRSF11 RBM5 SKIV2L2 SNRPB
mRNA processing (GO:0006397)	12/397	2.58E-06	1.91E-03	-2.385	14.936	RBM39 PNN SFPQ RBFOX2 SCAF11 SREK1 TARDBP SF3B1 SRSF11 RBM5 SKIV2L2 SNRPB
RNA splicing, via transesterification reactions (GO:0000375)	8/184	9.90E-06	4.88E-03	-2.228	11.856	PNN SFPQ SCAF11 SF3B1 SRSF11 RBM5 SKIV2L2 SNRPB
RNA splicing, via transesterification reactions with bulged adenosine as	7/177	6.63E-05	1.96E-02	-2.215	8.710	PNN SFPQ SF3B1 SRSF11 SKIV2L2 RBM5 SNRPB
mRNA splicing, via spliceosome (GO:0000398)	7/177	6.63E-05	1.96E-02	-2.213	8.700	PNN SFPQ SF3B1 SRSF11 SKIV2L2 RBM5 SNRPB
viral transcription (GO:0019083)	5/84	1.14E-04	2.81E-02	-2.093	7.475	RPS16 RPS29 RPL12 RPL22 RPL35A
placenta blood vessel development (GO:0060674)	3/30	6.38E-04	4.73E-02	-2.407	7.344	RBM15 NR2F2 RBPJ

Figure 1:

GO Cellular Component

Term	Overlap	P-value	Adjusted P	Z-score	Combined Score	Genes
nucleolus (GO:0005730)	26/1653	1.3031E-06	0.00024	-2.212	18.441	CEP57 ZBTB20 RBPJ MTDH FXR1 PNN PSMB2 IFI16 SNX27 POGZ MCMBP SF3B1 RBM5 SRSF11 RBM39 RBM15 RBFOX2 SCAF11 TBCB NAV2 HCFC2 FOSL2 ETV6 CIAPIN1 SREK1 TARDBP
cell-substrate junction (GO:0030055)	10/362	3.9648E-05	0.003648	-2.384	13.382	PPFIBP1 ANXA1 RPS16 RPS29 RPL12 PRKAR2A G3BP1 RPL22 DMD LPP
focal adhesion (GO:0005925)	9/352	0.00017513	0.008024	-2.331	11.247	PPFIBP1 ANXA1 RPS16 RPS29 PRKAR2A RPL12 RPL22 G3BP1 LPP
cell-substrate adherens junction (GO:0005924)	9/358	0.0001986	0.008024	-2.327	11.229	PPFIBP1 ANXA1 RPS16 RPS29 PRKAR2A RPL12 G3BP1 RPL22 LPP
spliceosomal complex (GO:0005681)	6/151	0.00021804	0.008024	-2.125	10.256	PNN SREK1 SF3B1 SKIV2L2 RBM5 SNRPB
adherens junction (GO:0005912)	9/405	0.00048963	0.01287	-2.250	9.795	PPFIBP1 ANXA1 RPS16 RPS29 RPL12 PRKAR2A RPL22 G3BP1 LPP
anchoring junction (GO:0070161)	9/419	0.00062461	0.014366	-2.237	9.491	PPFIBP1 ANXA1 RPS16 RPS29 RPL12 PRKAR2A RPL22 G3BP1 LPP
extracellular vesicular exosome (GO:0070062)	29/2717	0.0004351	0.01287	-2.148	9.348	TMED10 RPL12 RAP1GDS1 GNAI3 TNFAIP3 SRI GLG1 UBE2L3 CST3 RPS16 PSMB2 PRKAR2A ANXA1 RPL22 MYOF BGN RPL35A PTPN13 TUBB4B MOB1A COL1A2 NACA RPS29 SERBP1 MMRN2 CRYL1 KCTD12 SNRPB PPIC
catalytic step 2 spliceosome (GO:0071013)	4/80	0.00109813	0.020206	-1.978	7.718	PNN SF3B1 SKIV2L2 SNRPB
ribosomal subunit (GO:0044391)	5/135	0.00101926	0.020206	-1.955	7.626	RPS16 RPS29 RPL12 RPL22 RPL35A
cytosol (GO:0005829)	26/2529	0.0016125	0.026973	-1.911	6.903	CEP57 RPL12 RAP1GDS1 TNFAIP3 SRI RPS16 PSMB2 IFI16 CALD1 PRKAR2A DMD ARIH1 MAP2K5 MCL1 RANBP2 OSBPL3 RPL22 RPL35A TUBB4B PLA2G16 MOB1A RPS29 ALMS1 CRYL1 CYCS SNRPB
cytosolic large ribosomal subunit (GO:0022625)	3/52	0.00317337	0.048658	-1.887	5.705	RPL12 RPL22 RPL35A

Figure 2:

GO Molecular Function

Term	Overlap	P-value	Adjusted P-v	Z-score	Combined Sc	Genes
mRNA binding (GO:0003729)	6/112	4.1861E-05	0.011	-2.357	10.555	FXR1 RBFOX2 SERBP1 G3BP1 TARDBP RBM5
cholesterol binding (GO:0015485)	3/34	0.00092497	0.046	-2.651	8.148	OSBPL8 SOAT1 OSBPL3
mRNA 3'-UTR binding (GO:0003730)	3/35	0.00100754	0.046	-2.535	7.793	FXR1 SERBP1 TARDBP
sterol binding (GO:0032934)	3/38	0.00128277	0.046	-2.415	7.421	OSBPL8 SOAT1 OSBPL3
steroid binding (GO:0005496)	4/82	0.00120397	0.046	-2.389	7.343	OSBPL8 AR SOAT1 OSBPL3
double-stranded DNA binding (GO:0003690)	5/109	0.0003857	0.046	-2.352	7.229	IFI16 RBMS1 ZNF148 TARDBP FOSL2
transcription factor binding (GO:0008134)	9/464	0.00127875	0.046	-2.346	7.211	AR RBFOX2 IFI16 NACA JMJD1C TCF4 SRI RBPJ MTDH
sequence-specific DNA binding RNA polymerase II transcription factor activity (GO:0000981)	8/397	0.00187775	0.046	-2.317	7.122	AR ZFHX3 ZBTB20 TCF4 NR2F2 TARDBP RBPJ ETV6
peptidyl-prolyl cis-trans isomerase activity (GO:0003755)	3/41	0.00160114	0.046	-2.252	6.922	RANBP2 NKTR PPIC
structural constituent of ribosome (GO:0003735)	5/160	0.00215796	0.049	-2.225	6.721	RPS16 RPS29 RPL12 RPL22 RPL35A
cis-trans isomerase activity (GO:0016859)	3/43	0.0018385	0.046	-2.173	6.678	RANBP2 NKTR PPIC
RNA polymerase II transcription factor binding (GO:0001085)	4/90	0.00169916	0.046	-2.126	6.535	AR TCF4 RBPJ MTDH

Figure 3:

KEGG 2016

Term	Overlap	P-value	Adjusted P-v	Z-score	Combined Score	Genes
Gap junction_Homo sapiens_hsa04540	4/88	0.00156411	0.085	-1.772	4.363	GNAI3 LPAR1 TUBB4B MAP2K5
Ribosome_Homo sapiens_hsa03010	5/137	0.0010885	0.085	-1.746	4.299	RPS16 RPS29 RPL12 RPL22 RPL35A
Apoptosis_Homo sapiens_hsa04210	3/140	0.04504868	0.691	-1.833	0.678	CYCS PTPN13 MCL1
Regulation of lipolysis in adipocytes_Homo sapiens_hsa04923	2/56	0.03997967	0.691	-1.831	0.677	PLA2G16 GNAI3
Viral myocarditis_Homo sapiens_hsa05416	2/59	0.04393948	0.691	-1.767	0.654	CYCS DMD
Notch signaling pathway_Homo sapiens_hsa04330	2/48	0.03014593	0.691	-1.680	0.622	JAG1 RBPJ
RNA transport_Homo sapiens_hsa03013	3/172	0.07386865	0.691	-1.673	0.619	FXR1 RANBP2 PNN
TNF signaling pathway_Homo sapiens_hsa04668	2/110	0.12831503	0.691	-1.645	0.609	JAG1 TNFAIP3
Parkinson's disease_Homo sapiens_hsa05012	3/142	0.0466479	0.691	-1.625	0.601	GNAI3 CYCS UBE2L3
Pathways in cancer_Homo sapiens_hsa05200	4/397	0.18720527	0.691	-1.603	0.593	AR LPAR1 GNAI3 CYCS
Toxoplasmosis_Homo sapiens_hsa05145	2/118	0.14364561	0.691	-1.569	0.580	GNAI3 CYCS
Platelet activation_Homo sapiens_hsa04611	2/122	0.15145061	0.691	-1.547	0.572	COL1A2 GNAI3

Figure 4:

Reactome 2016

Term	Overlap	P-value	Adjusted P-	Z-score	Combined Score	Genes
Influenza Viral RNA Transcription and Replication_Homo sapiens_R-HSA-168273	6/128	8.8212E-05	0.007	-2.065	10.280	RANBP2 RPS16 RPS29 RPL12 RPL22 RPL35A
Influenza Life Cycle_Homo sapiens_R-HSA-168255	6/136	0.00012323	0.007	-2.050	10.207	RANBP2 RPS16 RPS29 RPL12 RPL22 RPL35A
Major pathway of rRNA processing in the nucleolus_Homo sapiens_R-HSA-6791226	7/166	4.4148E-05	0.007	-2.045	10.182	RPS16 RPS29 RIOK3 RPL12 RPL22 RPL35A SKIV2L2
rRNA processing_Homo sapiens_R-HSA-72312	7/180	7.3678E-05	0.007	-2.020	10.057	RPS16 RPS29 RIOK3 RPL12 RPL22 RPL35A SKIV2L2
Eukaryotic Translation Elongation_Homo sapiens_R-HSA-156842	5/89	0.00014986	0.007	-1.946	9.692	RPS16 RPS29 RPL12 RPL22 RPL35A
Peptide chain elongation_Homo sapiens_R-HSA-156902	5/84	0.00011398	0.007	-1.929	9.607	RPS16 RPS29 RPL12 RPL22 RPL35A
Viral mRNA Translation_Homo sapiens_R-HSA-192823	5/84	0.00011398	0.007	-1.892	9.420	RPS16 RPS29 RPL12 RPL22 RPL35A
Selenocysteine synthesis_Homo sapiens_R-HSA-2408557	5/87	0.0001346	0.007	-1.888	9.400	RPS16 RPS29 RPL12 RPL22 RPL35A
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)_Homo sapiens_R-HSA-975956	5/89	0.00014986	0.007	-1.871	9.318	RPS16 RPS29 RPL12 RPL22 RPL35A
Eukaryotic Translation Termination_Homo sapiens_R-HSA-72764	5/87	0.0001346	0.007	-1.863	9.275	RPS16 RPS29 RPL12 RPL22 RPL35A
Influenza Infection_Homo sapiens_R-HSA-168254	6/147	0.00018849	0.008	-2.001	9.696	RANBP2 RPS16 RPS29 RPL12 RPL22 RPL35A
Formation of a pool of free 40S subunits_Homo sapiens_R-HSA-72689	5/96	0.00021387	0.008	-1.865	8.965	RPS16 RPS29 RPL12 RPL22 RPL35A
GTP hydrolysis and joining of the 60S ribosomal subunit_Homo sapiens_R-HSA-72706	5/107	0.00035412	0.009	-1.878	8.838	RPS16 RPS29 RPL12 RPL22 RPL35A
SRP-dependent cotranslational protein targeting to membrane_Homo sapiens_R-HSA-1799339	5/107	0.00035412	0.009	-1.870	8.801	RPS16 RPS29 RPL12 RPL22 RPL35A
3' -UTR-mediated translational regulation_Homo sapiens_R-HSA-157279	5/106	0.00033909	0.009	-1.863	8.768	RPS16 RPS29 RPL12 RPL22 RPL35A
L13a-mediated translational silencing of Ceruloplasmin expression_Homo sapiens_R-HSA-156827	5/106	0.00033909	0.009	-1.838	8.654	RPS16 RPS29 RPL12 RPL22 RPL35A
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)_Homo sapiens_R-HSA-975957	5/106	0.00033909	0.009	-1.795	8.451	RPS16 RPS29 RPL12 RPL22 RPL35A
Nonsense-Mediated Decay (NMD)_Homo sapiens_R-HSA-927802	5/106	0.00033909	0.009	-1.788	8.418	RPS16 RPS29 RPL12 RPL22 RPL35A
Selenoamino acid metabolism_Homo sapiens_R-HSA-2408522	5/111	0.00041937	0.010	-1.796	8.249	RPS16 RPS29 RPL12 RPL22 RPL35A
Cap-dependent Translation Initiation_Homo sapiens_R-HSA-72737	5/114	0.00047396	0.010	-1.836	8.388	RPS16 RPS29 RPL12 RPL22 RPL35A
Eukaryotic Translation Initiation_Homo sapiens_R-HSA-72613	5/114	0.00047396	0.010	-1.827	8.349	RPS16 RPS29 RPL12 RPL22 RPL35A
Defective CHSY1 causes TPBS_Homo sapiens_R-HSA-3595177	2/7	0.00065228	0.012	-0.248	1.087	BGN DCN
Defective CHST14 causes EDS, musculocontractural type_Homo sapiens_R-HSA-3595174	2/7	0.00065228	0.012	-0.176	0.770	BGN DCN
Defective CHST3 causes SEDCJD_Homo sapiens_R-HSA-3595172	2/7	0.00065228	0.012	-0.133	0.582	BGN DCN
Disease_Homo sapiens_R-HSA-1643685	12/725	0.00081811	0.015	-2.179	9.146	RANBP2 JAG1 RPS16 PSMB2 ADAMTS1 RPS29 RPL12 RPL22 BGN RPL35A RBPJ DCN
Translation_Homo sapiens_R-HSA-72766	5/151	0.00167515	0.029	-1.839	6.536	RPS16 RPS29 RPL12 RPL22 RPL35A
Dermatan sulfate biosynthesis_Homo sapiens_R-HSA-2022923	2/11	0.00168328	0.029	-0.724	2.573	BGN DCN
Protein folding_Homo sapiens_R-HSA-391251	4/101	0.00258935	0.042	-1.774	5.604	GNAI3 PFDN1 TBCB TUBB4B
CS/DS degradation_Homo sapiens_R-HSA-2024101	2/14	0.00275439	0.044	-1.140	3.570	BGN DCN

Figure 5:

Wikipathways 2016

Term	Overlap	P-value	Adjusted P	Z-score	Combined Score	Genes
mRNA processing_Mus musculus_WP310	17/398	9.4828E-11	0.000	-2.136	39.259	RBM39 TIA1 RBFOX2 TMED10 RPL12 RPL22 FXR1 SFPQ RPS29 G3BP1 RBMS1 SREK1 TARDBP SF3B1 SRSF11 RBM5 SNRPB
mRNA Processing_Homo sapiens_WP411	7/127	7.7939E-06	0.000	-1.941	15.053	RBM39 SFPQ TMED10 SREK1 SF3B1 RBM5 SNRPB
Cytoplasmic Ribosomal Proteins_Mus musculus_WP163	5/70	4.7579E-05	0.002	-1.972	12.524	RPS16 RPS29 RPL12 RPL22 RPL35A
Cytoplasmic Ribosomal Proteins_Homo sapiens_WP477	5/89	0.00014986	0.004	-1.924	10.566	RPS16 RPS29 RPL12 RPL22 RPL35A

Figure 6:

Virus MINT

Term	Overlap	P-value	Adjusted P	Z-score	Combined Score	Genes
Epstein-Barr virus (strain GD1)	7/133	1.0553E-05	7.387E-05	-1.190	11.322	DDX17 SFPQ RPS16 RPL12 SERBP1 RPL22 SF3B1

Figure 7: