LabBook_19_02_2016

Claire Green

Monday

At the Friday lab meeting, Gabriel mentioned that the gene pool that I sample from for my random permutation test is not quite identical to my test sample because I was not removing the genes that did not have enough presence calls (at least 3). What I did was take the results table which was ranked to find my top X genes and sample from all HGNC symbols in that list. This means that both my random pool and experimental pool contain genes that are 1) not blank 2) not duplicates 3) not antisense matching 4) no less than 3 presence calls.

This essentially halved my sample pools, meaning that I was taking random genes from between 8 and 12 thousand genes. Consequently my p values became much larger, with my value of 16 from top 3000 reaching a value of 0.02 (still significant), and 50 from top 4000 0.07 (not significant). I feel confident that my list is still important because it's highly enriched for genes that are known to be either associated with neurological diseases or processes known to be dysfunctional in neurodegeneration. My next step is to take smaller increments between 3000 and 4000 to see where the genes lie and where it no longer becomes significant. I will take incremenets of 100 genes.

Change of Plan

I noticed that the section of code that I took from Wenbin to take out any genes with negative matching strand probes was not working (I had not seen the error) because the name for the column annotation notes was slightly different in my output. When I fixed this, it changed my results.

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

Interestingly, my output table now looks like this:

Top 1000	Top 2000	Top 3000	Top 4000
0	CSRP1	STOM	STOM
	RNF13	CSRP1	UPF3A
		RNF13	FBXO9
		TUBB3	DYNLT1
		PSAP	CSRP1
		RPL6	ETS2
		CCT2	RNF13
		NKTR	WASL
		MAP3K13	CST3
		NUTF2	MAP4K4
		RPS6	TUBB3
		NAGA	PSAP
		PFDN1	RPL6
		TARDBP	CCT2

Top 1000	Top 2000	Top 3000	Top 4000
		TARS	PCNA
		PTEN	SMPD4
		RNF130	DMD
		HSD17B4	ICMT
		DDX5	SUPT7L
		GTF2I	NKTR
			MAP3K13
			NUTF2
			RPS6
			MTR
			CREB1
			ACAT1
			CDK5R1
			BPTF
			PRKD1
			NAGA
			GSTO1
			PFDN1
			DDX39B
			TARDBP
			TARS
			PTEN
			USP11
			PAICS
			UNC119B
			RNF130
			HSD17B4
			TMEM59
			RTN1
			TRO
			DDX5
			GNPAT
			CDK16
			RSRC2
			GTF2I
			WBSCR22
			MARS
			GTF3C2
			C14orf1
			TAF5L
			TCF4
			WDR78
			LBR
			ZIC1
			ZFP36
			FBXL14
			DDX39A
			C18orf32
			DCN
			CAPN2
			RPLP2
			LDLRAD4

Top 1000	Top 2000	Top 3000	Top 4000
			PSMD1
			MPHOSPH9
			ITM2A
			MSL3
			TANK
			TNFAIP1
			LSM5

The genes look largely the same, there is only loss of the genes BRD3, EEF1A1, and RECQL. TARDBP is now commonly DE in the top 3000 genes which is promising.

Next I ran the random permutations test, again using the same table from the results (_uniqueresult.csv). This was a sample pool of 8050 for C9orf72, 10,065 for CHMP2B, 9506 for FTLD, 10405 for sALS, and 11935 for VCP.

For 73 genes from sampling top 4000, the p value was 0.0015 For 20 genes from sampling top 3000, the p value was 0.0158 For 2 genes from sampling top 2000, the p value was not significant at 0.2301

I find it surprising that the p values are getting more significant the more genes I include in the consensus. It could be that there are a large number of common genes that are less differentially expressed than the very top genes.

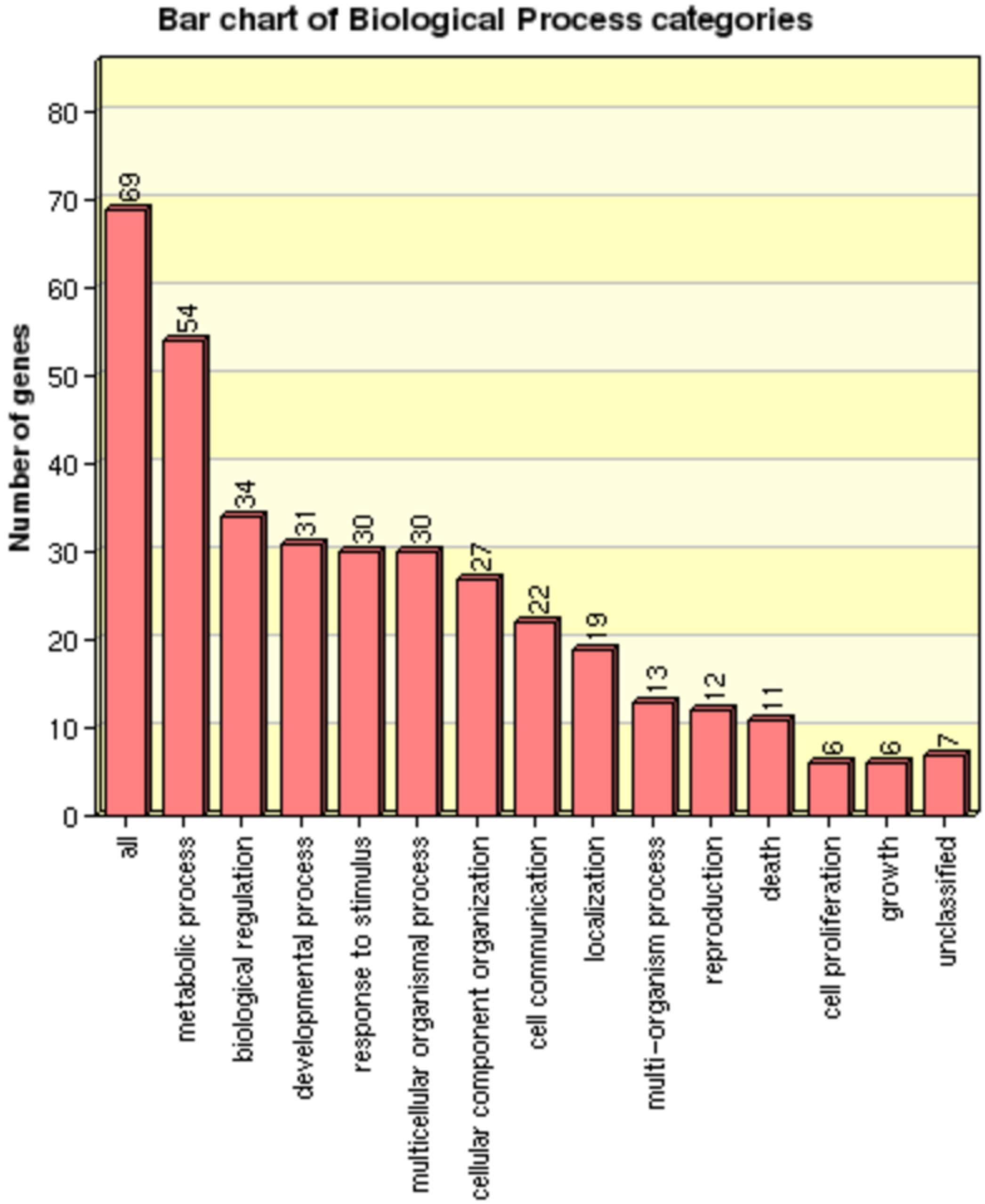
List of genes and names

Gene	Gene Names
ACAT1	acetyl-CoA acetyltransferase 1
BPTF	bromodomain PHD finger transcription factor
C14orf1	chromosome 14 open reading frame 1
C18orf32	chromosome 18 open reading frame 32
CAPN2	calpain 2, (m/II) large subunit
CCT2	chaperonin containing TCP1, subunit 2 (beta)
CDK16	cyclin-dependent kinase 16
CDK5R1	cyclin-dependent kinase 5, regulatory subunit 1 (p35)
CREB1	cAMP responsive element binding protein 1
CSRP1	cysteine and glycine-rich protein 1
CST3	cystatin C
DCN	decorin
DDX39A	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39A
DDX39B	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39B
DDX5	DEAD (Asp-Glu-Ala-Asp) box helicase 5
DMD	dystrophin
DYNLT1	dynein, light chain, Tctex-type 1
ETS2	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)
FBXL14	F-box and leucine-rich repeat protein 14
FBXO9	F-box protein 9
GNPAT	glyceronephosphate O-acyltransferase
GSTO1	glutathione S-transferase omega 1
GTF2I	general transcription factor IIi
GTF3C2	general transcription factor IIIC, polypeptide 2, beta 110kDa
HSD17B4	hydroxysteroid (17-beta) dehydrogenase 4
ICMT	isoprenylcysteine carboxyl methyltransferase

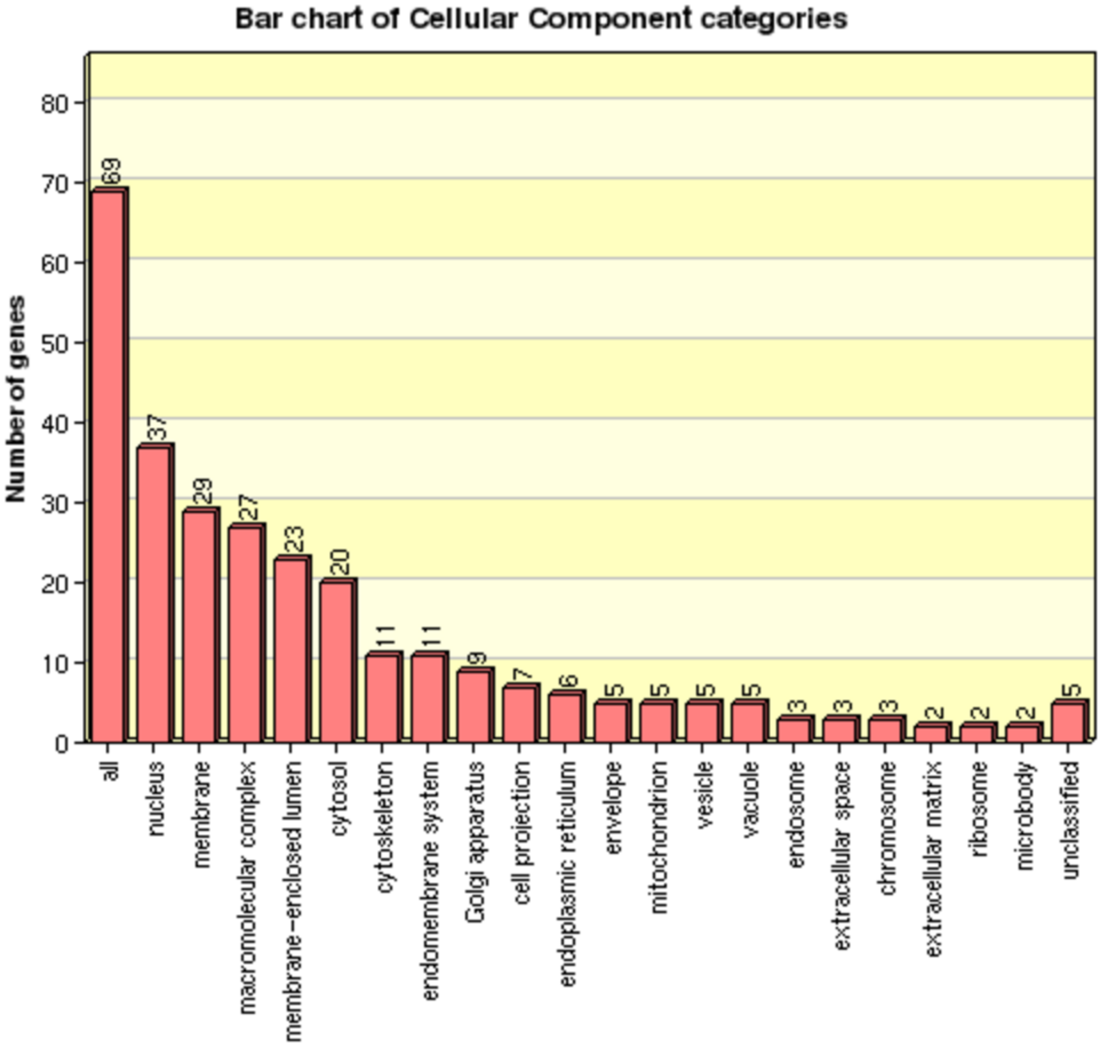
Gene	Gene Names
ITM2A	integral membrane protein 2A
LBR	lamin B receptor
LDLRAD4	Low Density Lipoprotein Receptor Class A Domain Containing 4
LSM5	LSM5 homolog, U6 small nuclear RNA associated (S. cerevisiae)
MAP3K13	mitogen-activated protein kinase kinase kinase 13
MAP4K4	mitogen-activated protein kinase kinase kinase kinase 4
MARS	Methionyl-TRNA Synthetase
MPHOSPH9	M-phase phosphoprotein 9
MSL3	male-specific lethal 3 homolog (Drosophila)
MTR	5-methyltetrahydrofolate-homocysteine methyltransferase
NAGA	N-acetylgalactosaminidase, alpha-
NKTR	natural killer-tumor recognition sequence
NUTF2	nuclear transport factor 2
PAICS	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazole succinocarboxamide synthetase
PCNA	proliferating cell nuclear antigen
PFDN1	prefoldin subunit 1
PRKD1	protein kinase D1
PSAP	prosaposin
PSMD1	proteasome (prosome, macropain) 26S subunit, non-ATPase, 1
PTEN	phosphatase and tensin homolog
RNF13	ring finger protein 13
RNF130	ring finger protein 130
RPL6	ribosomal protein L6
RPLP2	ribosomal protein, large, P2
RPS6	ribosomal protein S6
RSRC2	arginine/serine-rich coiled-coil 2
RTN1	reticulon 1
SMPD4	sphingomyelin phosphodiesterase 4, neutral membrane (neutral sphingomyelinase-3)
STOM	stomatin
SUPT7L	suppressor of Ty 7 (S. cerevisiae)-like
TAF5L	TAF5-like RNA polymerase II, p300/CBP-associated factor (PCAF)-associated factor, 65kDa
TANK	TRAF family member-associated NFKB activator
TARDBP	TAR DNA binding protein
TARS	threonyl-tRNA synthetase
TCF4	transcription factor 4
TMEM59	transmembrane protein 59
TNFAIP1	tumor necrosis factor, alpha-induced protein 1 (endothelial)
TRO	trophinin
TUBB3	tubulin, beta 3 class III
UNC119B	unc-119 homolog B (C. elegans)
UPF3A	UPF3 regulator of nonsense transcripts homolog A (yeast)
USP11	ubiquitin specific peptidase 11
WASL	Wiskott-Aldrich syndrome-like
WBSCR22	Williams Beuren syndrome chromosome region 22
WDR78	WD repeat domain 78
ZFP36	zinc finger protein 36, C3H type, homolog (mouse)
ZIC1	Zic family member 1

Like before, I inputted the gene list into WebGestalt to identify associated GO terms and diseases

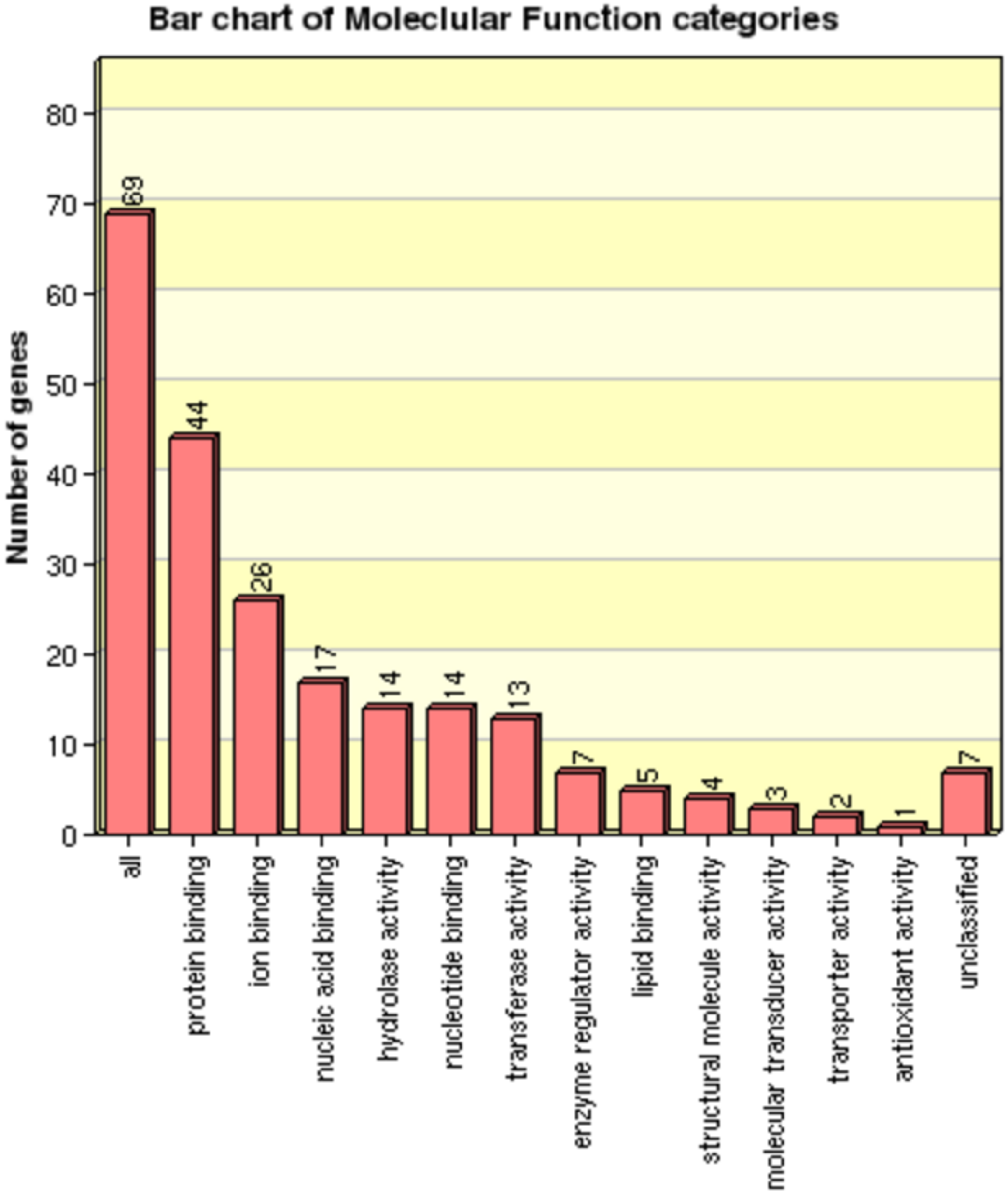
Biological Process



Cellular Component



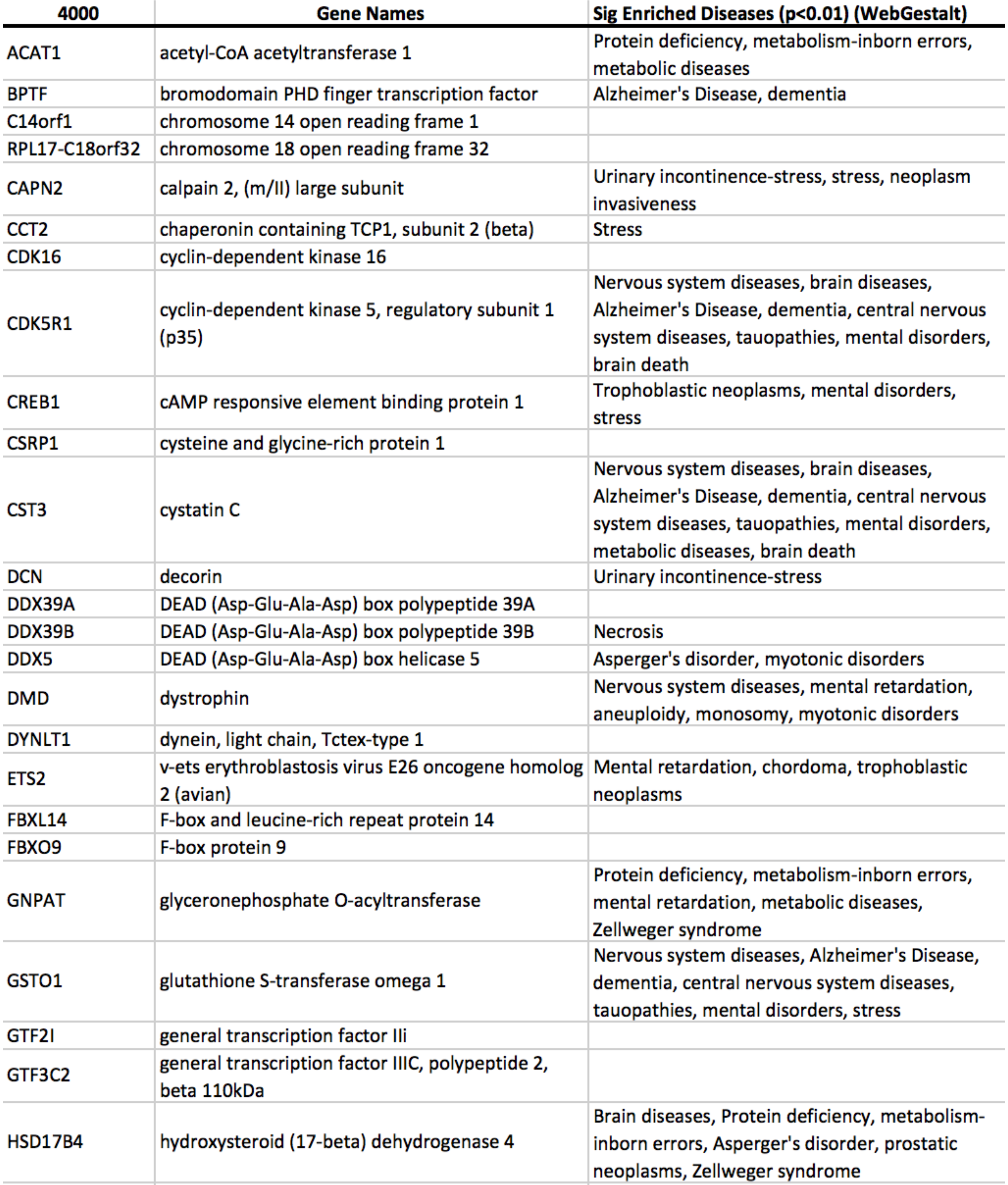
Molecular Function



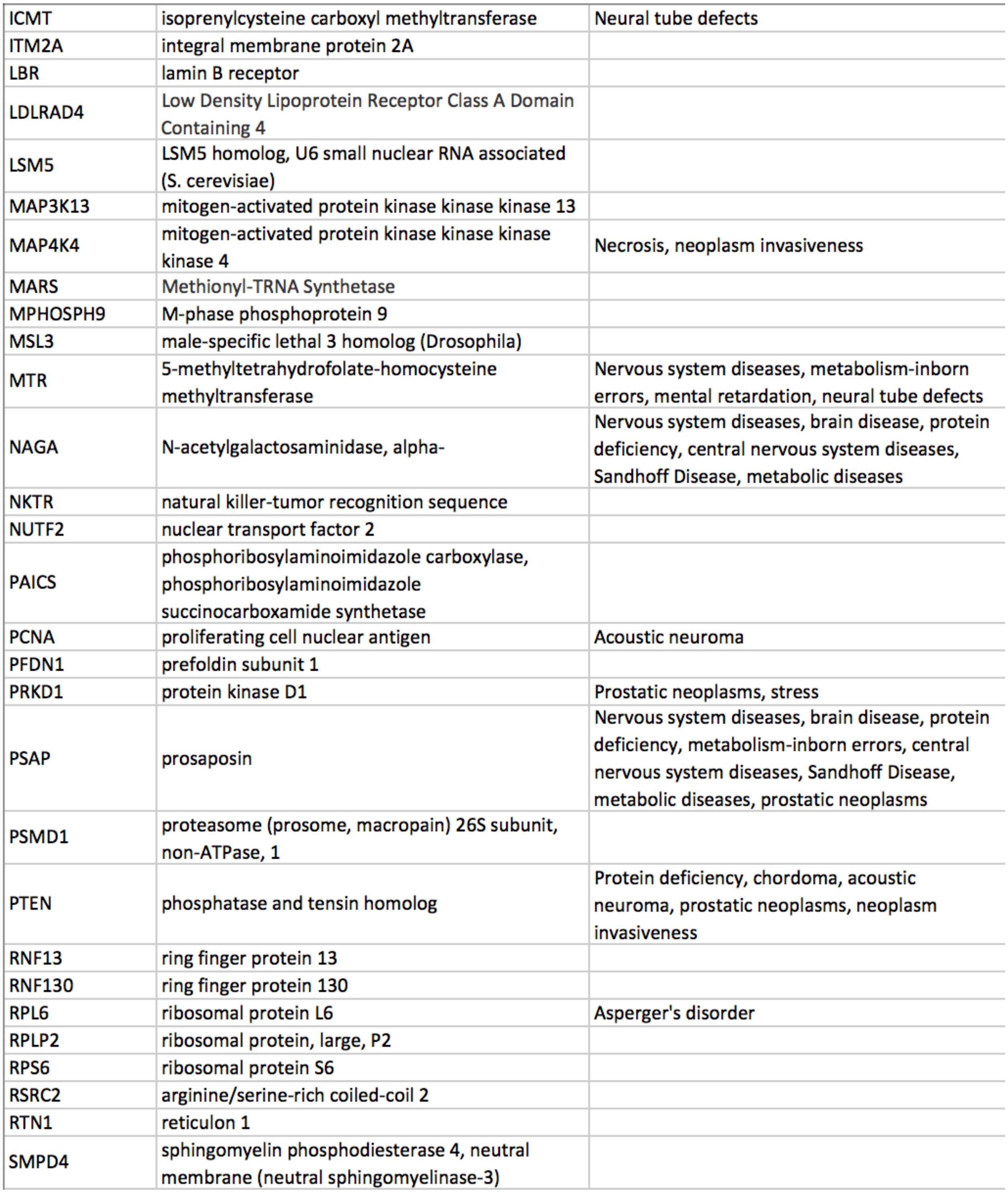
Tuesday

Next, I ran the gene list through WebGestalt and CTTV to see which genes have been associated with TDP-43 diseases.

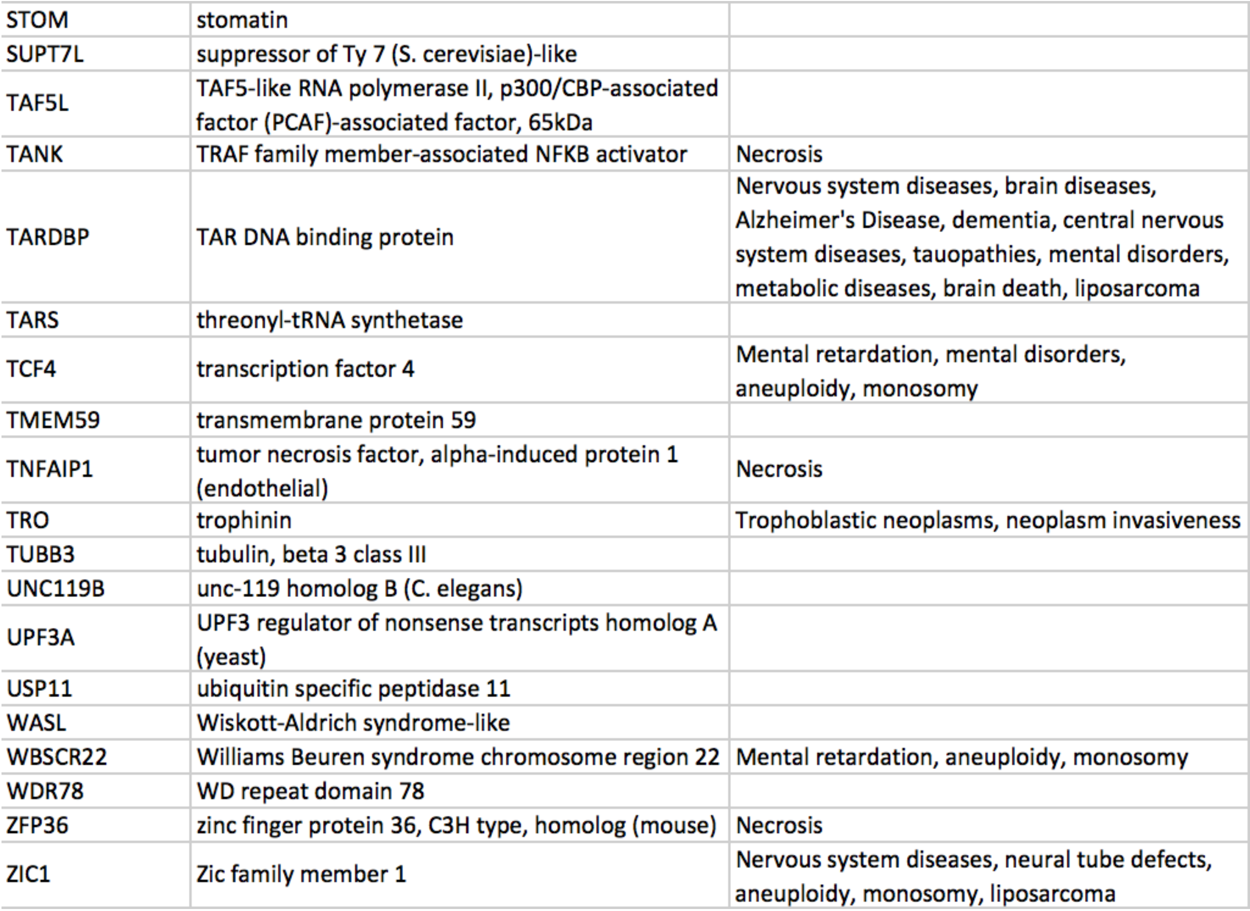
WebGestalt Associated Diseases



4000	Gene Names	Sig Enriched Diseases (p<0.01) (WebGestalt)
ACAT1	acetyl-CoA acetyltransferase 1	Protein deficiency, metabolism-inborn errors, metabolic diseases
BPTF	bromodomain PHD finger transcription factor	Alzheimer's Disease, dementia
C14orf1	chromosome 14 open reading frame 1	
RPL17-C18orf32	chromosome 18 open reading frame 32	
CAPN2	calpain 2, (m/II) large subunit	Urinary incontinence-stress, stress, neoplasm invasiveness
CCT2	chaperonin containing TCP1, subunit 2 (beta)	Stress
CDK16	cyclin-dependent kinase 16	
CDK5R1	cyclin-dependent kinase 5, regulatory subunit 1 (p35)	Nervous system diseases, brain diseases, Alzheimer's Disease, dementia, central nervous system diseases, tauopathies, mental disorders, brain death
CREB1	cAMP responsive element binding protein 1	Trophoblastic neoplasms, mental disorders, stress
CSRP1	cysteine and glycine-rich protein 1	
CST3	cystatin C	Nervous system diseases, brain diseases, Alzheimer's Disease, dementia, central nervous system diseases, tauopathies, mental disorders, metabolic diseases, brain death
DCN	decorin	Urinary incontinence-stress
DDX39A	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39A	
DDX39B	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39B	Necrosis
DDX5	DEAD (Asp-Glu-Ala-Asp) box helicase 5	Asperger's disorder, myotonic disorders
DMD	dystrophin	Nervous system diseases, mental retardation, aneuploidy, monosomy, myotonic disorders
DYNLT1	dynein, light chain, Tctex-type 1	
ETS2	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)	Mental retardation, chordoma, trophoblastic neoplasms
FBXL14	F-box and leucine-rich repeat protein 14	
FBXO9	F-box protein 9	
GNPAT	glyceronephosphate O-acyltransferase	Protein deficiency, metabolism-inborn errors, mental retardation, metabolic diseases, Zellweger syndrome
GSTO1	glutathione S-transferase omega 1	Nervous system diseases, Alzheimer's Disease, dementia, central nervous system diseases, tauopathies, mental disorders, stress
GTF2I	general transcription factor IIi	
GTF3C2	general transcription factor IIIC, polypeptide 2, beta 110kDa	
HSD17B4	hydroxysteroid (17-beta) dehydrogenase 4	Brain diseases, Protein deficiency, metabolism-inborn errors, Asperger's disorder, prostatic neoplasms, Zellweger syndrome

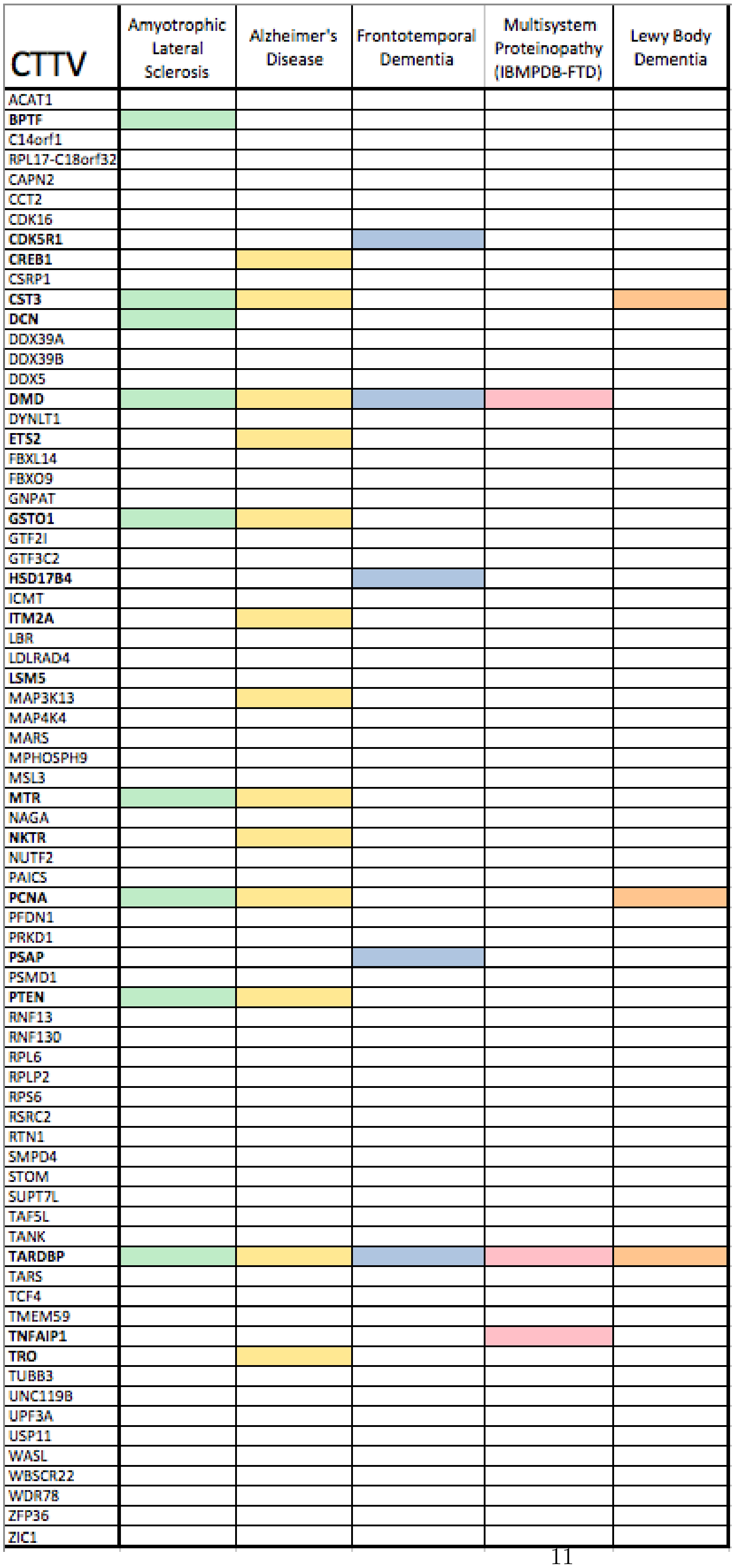


ICMT	isoprenylcysteine carboxyl methyltransferase	Neural tube defects
ITM2A	integral membrane protein 2A	
LBR	lamin B receptor	
LDLRAD4	Low Density Lipoprotein Receptor Class A Domain Containing 4	
LSM5	LSM5 homolog, U6 small nuclear RNA associated (S. cerevisiae)	
MAP3K13	mitogen-activated protein kinase kinase kinase 13	
MAP4K4	mitogen-activated protein kinase kinase kinase kinase 4	Necrosis, neoplasm invasiveness
MARS	Methionyl-TRNA Synthetase	
MPHOSPH9	M-phase phosphoprotein 9	
MSL3	male-specific lethal 3 homolog (Drosophila)	
MTR	5-methyltetrahydrofolate-homocysteine methyltransferase	Nervous system diseases, metabolism-inborn errors, mental retardation, neural tube defects
NAGA	N-acetylgalactosaminidase, alpha-	Nervous system diseases, brain disease, protein deficiency, central nervous system diseases, Sandhoff Disease, metabolic diseases
NKTR	natural killer-tumor recognition sequence	
NUTF2	nuclear transport factor 2	
PAICS	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazole succinocarboxamide synthetase	
PCNA	proliferating cell nuclear antigen	Acoustic neuroma
PFDN1	prefoldin subunit 1	
PRKD1	protein kinase D1	Prostatic neoplasms, stress
PSAP	prosaposin	Nervous system diseases, brain disease, protein deficiency, metabolism-inborn errors, central nervous system diseases, Sandhoff Disease, metabolic diseases, prostatic neoplasms
PSMD1	proteasome (prosome, macropain) 26S subunit, non-ATPase, 1	
PTEN	phosphatase and tensin homolog	Protein deficiency, chordoma, acoustic neuroma, prostatic neoplasms, neoplasm invasiveness
RNF13	ring finger protein 13	
RNF130	ring finger protein 130	
RPL6	ribosomal protein L6	Asperger's disorder
RPLP2	ribosomal protein, large, P2	
RPS6	ribosomal protein S6	
RSRC2	arginine/serine-rich coiled-coil 2	
RTN1	reticulon 1	
SMPD4	sphingomyelin phosphodiesterase 4, neutral membrane (neutral sphingomyelinase-3)	



STOM	stomatin	
SUPT7L	suppressor of Ty 7 (S. cerevisiae)-like	
TAF5L	TAF5-like RNA polymerase II, p300/CBP-associated factor (PCAF)-associated factor, 65kDa	
TANK	TRAF family member-associated NFKB activator	Necrosis
TARDBP	TAR DNA binding protein	Nervous system diseases, brain diseases, Alzheimer's Disease, dementia, central nervous system diseases, tauopathies, mental disorders, metabolic diseases, brain death, liposarcoma
TARS	threonyl-tRNA synthetase	
TCF4	transcription factor 4	Mental retardation, mental disorders, aneuploidy, monosomy
TMEM59	transmembrane protein 59	
TNFAIP1	tumor necrosis factor, alpha-induced protein 1 (endothelial)	Necrosis
TRO	trophinin	Trophoblastic neoplasms, neoplasm invasiveness
TUBB3	tubulin, beta 3 class III	
UNC119B	unc-119 homolog B (C. elegans)	
UPF3A	UPF3 regulator of nonsense transcripts homolog A (yeast)	
USP11	ubiquitin specific peptidase 11	
WASL	Wiskott-Aldrich syndrome-like	
WBSCR22	Williams Beuren syndrome chromosome region 22	Mental retardation, aneuploidy, monosomy
WDR78	WD repeat domain 78	
ZFP36	zinc finger protein 36, C3H type, homolog (mouse)	Necrosis
ZIC1	Zic family member 1	Nervous system diseases, neural tube defects, aneuploidy, monosomy, liposarcoma

CTTV gene-disease associations



CTTV	Amyotrophic Lateral Sclerosis	Alzheimer's Disease	Frontotemporal Dementia	Multisystem Proteinopathy (IBMPDB-FTD)	Lewy Body Dementia
ACAT1					
BPTF	🟩				
C14orf1					
RPL17-C18orf32					
CAPN2					
CCT2					
CDK16					
CDK5R1			🟦		
CREB1		🟨			
CSRP1					
CST3	🟩	🟨			🟧
DCN	🟩				
DDX39A					
DDX39B					
DDX5					
DMD	🟩	🟨	🟦	🟥	
DYNLT1					
ETS2		🟨			
FBXL14					
FBXO9					
GNPAT					
GSTO1	🟩	🟨			
GTF2I					
GTF3C2					
HSD17B4			🟦		
ICMT					
ITM2A		🟨			
LBR					
LDLRAD4					
LSM5					
MAP3K13		🟨			
MAP4K4					
MARS					
MPHOSPH9					
MSL3					
MTR	🟩	🟨			
NAGA					
NKTR		🟨			
NUTF2					
PAICS					
PCNA	🟩	🟨			🟧
PFDN1					
PRKD1					
PSAP			🟦		
PSMD1					
PTEN	🟩	🟨			
RNF13					
RNF130					
RPL6					
RPLP2					
RPS6					
RSRC2					
RTN1					
SMPD4					
STOM					
SUPT7L					
TAF5L					
TANK					
TARDBP	🟩	🟨	🟦	🟥	🟧
TARS					
TCF4					
TMEM59					
TNFAIP1				🟥	
TRO		🟨			
TUBB3					
UNC119B					
UPF3A					
USP11					
WASL					
WBSCR22					
WDR78					
ZFP36					
ZIC1					

Clearly a large number of genes within my gene list have been associated with TDP-43 diseases, mostly through RNA expression experiments. For some of the genes, such as CAPN2 and many of the ribosomal proteins, very similar genes have been associated with these diseases. For example, both CAPN1 and CAPN3 were implicated by CTTV. So even if not explicitly associated, the processes they contibute to are likely the same.

Now that I have a list of genes that looks, at a first glance, , there are two large steps I need to take. Firstly, although I have validated the *number* of genes, I also need to validate the particular selection itself. One way I can do this is to perform the same analysis on another dataset, preferably RNA seq. RNA seq would be good for two reasons; firstly by showing that a similar list is produced on a completely different technology, this provides extremely good validation (although it is extremely rare to get even a remotely similar list). Secondly, it would allow me to learn how to perform expression analysis on RNA seq data which I have not done yet.

The other step I need to take, probably after validation, is to extract as much meaning from my list as possible. By examining the relationships between the genes in my gene set, as well as with other unseeded genes, I can perhaps decipher 1) the hub genes that might be of more interest 2) certain biological processes that are enriched in this list. This is easier said than done, and I think I'm going to need some advice on the best way of doing this. I know that WGCNA is a good idea, as building a coexpression network and looking for enrichment of GO terms in the modules may be useful- but I don't know what my input should be. The whole expression table? Just the genes left after filtering? Or only the genes in my list? I'm not sure. I'm trying to get WGCNA to work in the meantime, and this is what I have so far:

```
#### Creating Co-expression Network using WGCNA ####

library(WGCNA)
### C9orf72 ###
# Display the current working directory
setwd ("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/GeneExpressionAnalysis/TopGenes_2016-02-15/

#Read in desired genes
C9Results <- read.csv ("C9rankeduniqueresult.csv", header=TRUE) #Taking only the genes we deemed accept
#gene expression analysis to find criteria

C9ID <- cbind(C9Results$Probe.Set.ID)

#Read in raw expression values
setwd ("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/C9orf72_LCM/")
C9RawExp <- read.csv("eset_NineP_150612_exprs.csv")

C9Exp <-merge(C9ID, C9RawExp, by.x="V1", by.y="X") #merge raw expression data with accepted genes
rownames(C9Exp) <- C9Exp[,1] #make probeset IDs row names
colnames(C9Exp) <- colnames(C9RawExp) #make file names column names
C9Exp <- cbind(C9Exp[,2:12]) #remove ID column

C9Exp <- t(C9Exp) #transpose for WGCNA analysis

###Choosing soft threshold
# Choose a set of soft-thresholding powers
powers = c(c(1:10), seq(from = 12, to=20, by=2))
# Call the network topology analysis function
sft = pickSoftThreshold(C9Exp, powerVector = powers, verbose = 5)
# Plot the results:
sizeGrWindow(9, 5)
par(mfrow = c(1,2));
```

```
cex1 = 0.9;
# Scale-free topology fit index as a function of the soft-thresholding power
plot(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2],
     xlab="Soft Threshold (power)",ylab="Scale Free Topology Model Fit,signed R^2",type="n",
     main = paste("Scale independence"));
text(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2],
     labels=powers,cex=cex1,col="red");
# this line corresponds to using an R^2 cut-off of h
abline(h=0.70,col="red")
# Mean connectivity as a function of the soft-thresholding power
plot(sft$fitIndices[,1], sft$fitIndices[,5],
     xlab="Soft Threshold (power)",ylab="Mean Connectivity", type="n",
     main = paste("Mean connectivity"))
text(sft$fitIndices[,1], sft$fitIndices[,5], labels=powers, cex=cex1,col="red")

##SOFT THRESHOLD VALUE OF 6 SELECTED##

C9Exp <- data.matrix(C9Exp) #csv files contain character matrices, the following code requires numeric

##One-step network construction and module detection
setwd ("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/WGCNA/C9orf72/")
net = blockwiseModules(C9Exp, power = 6,
                       TOMType = "unsigned", minModuleSize = 30,
                       reassignThreshold = 0, mergeCutHeight = 0.25,
                       numericLabels = TRUE, pamRespectsDendro = FALSE,
                       saveTOMs = TRUE,
                       saveTOMFileBase = "C9TOM",
                       verbose = 3)
table(net$colors)

# open a graphics window
sizeGrWindow(12, 9)
# Convert labels to colors for plotting
mergedColors = labels2colors(net$colors)
# Plot the dendrogram and the module colors underneath
plotDendroAndColors(net$dendrograms[[1]], mergedColors[net$blockGenes[[1]]],
                    "Module colors",
                    dendroLabels = FALSE, hang = 0.03,
                    addGuide = TRUE, guideHang = 0.05)

moduleLabels = net$colors
moduleColors = labels2colors(net$colors)
MEs = net$MEs;
geneTree = net$dendrograms[[1]];
save(MEs, moduleLabels, moduleColors, geneTree

##Looking for enrichment of GO terms in modules
library(S4Vectors)
library(IRanges)
library(AnnotationDbi)
library(GO.db)
library(org.Hs.eg.db)
```

```
EntrezIds <- cbind(C9Results$Entrez.Gene)

GOenr = GOenrichmentAnalysis(moduleColors, EntrezIds, organism = "human", nBestP = 10);
```

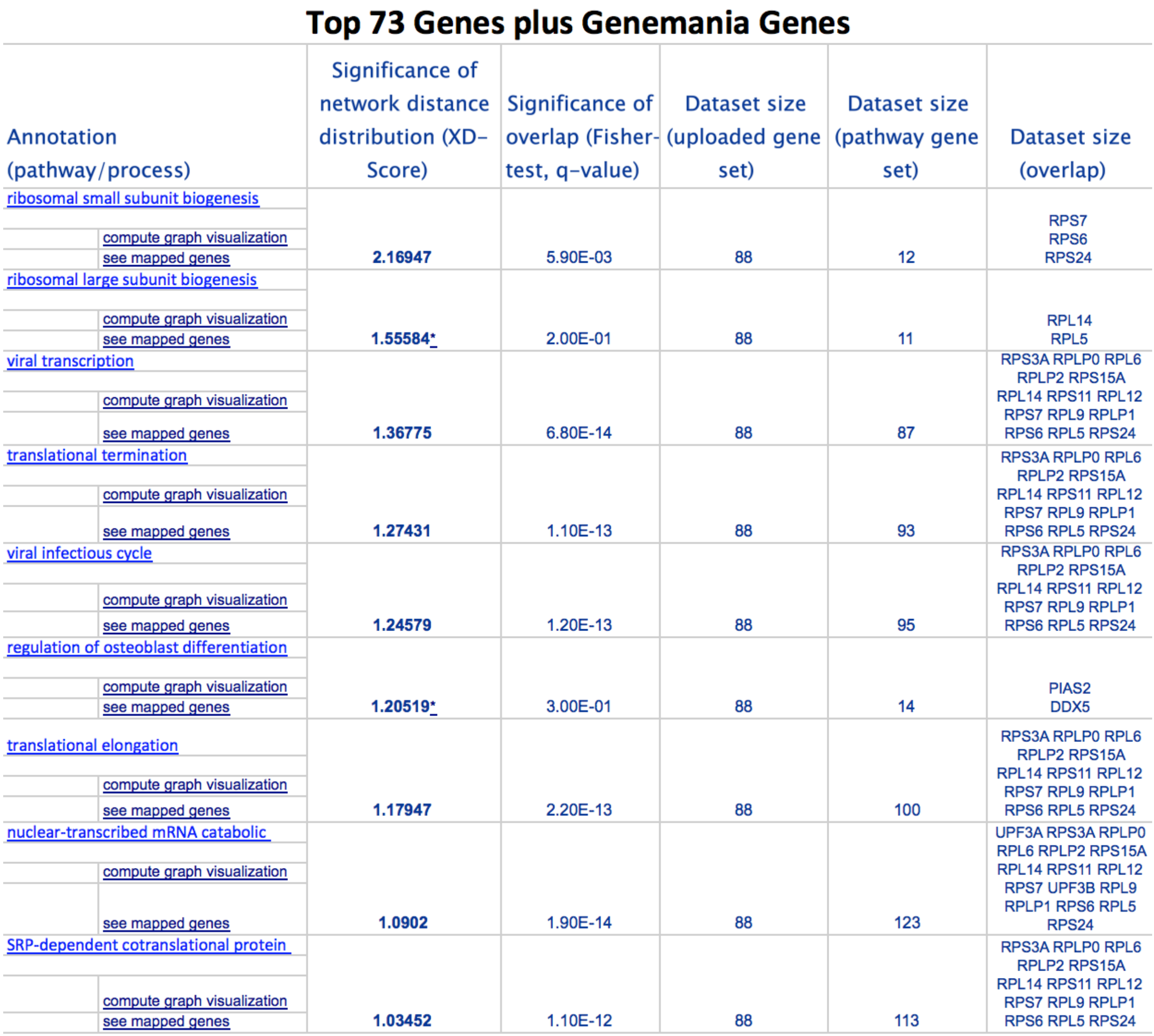
Thursday

When I generate this table it's a little confusing. In fact the WGCNA website itself says not to use these results as published enrichment analysis. So I'm not really sure what to do. I took my list of genes and looked for online tools for enrichment. I found a resource called EnrichNet and put my 73 genes through GO molecular function enrichment analysis, however none of the results were significant.

I decided to look at some other enrichment software. I looked at EnrichNet, using four different lists:

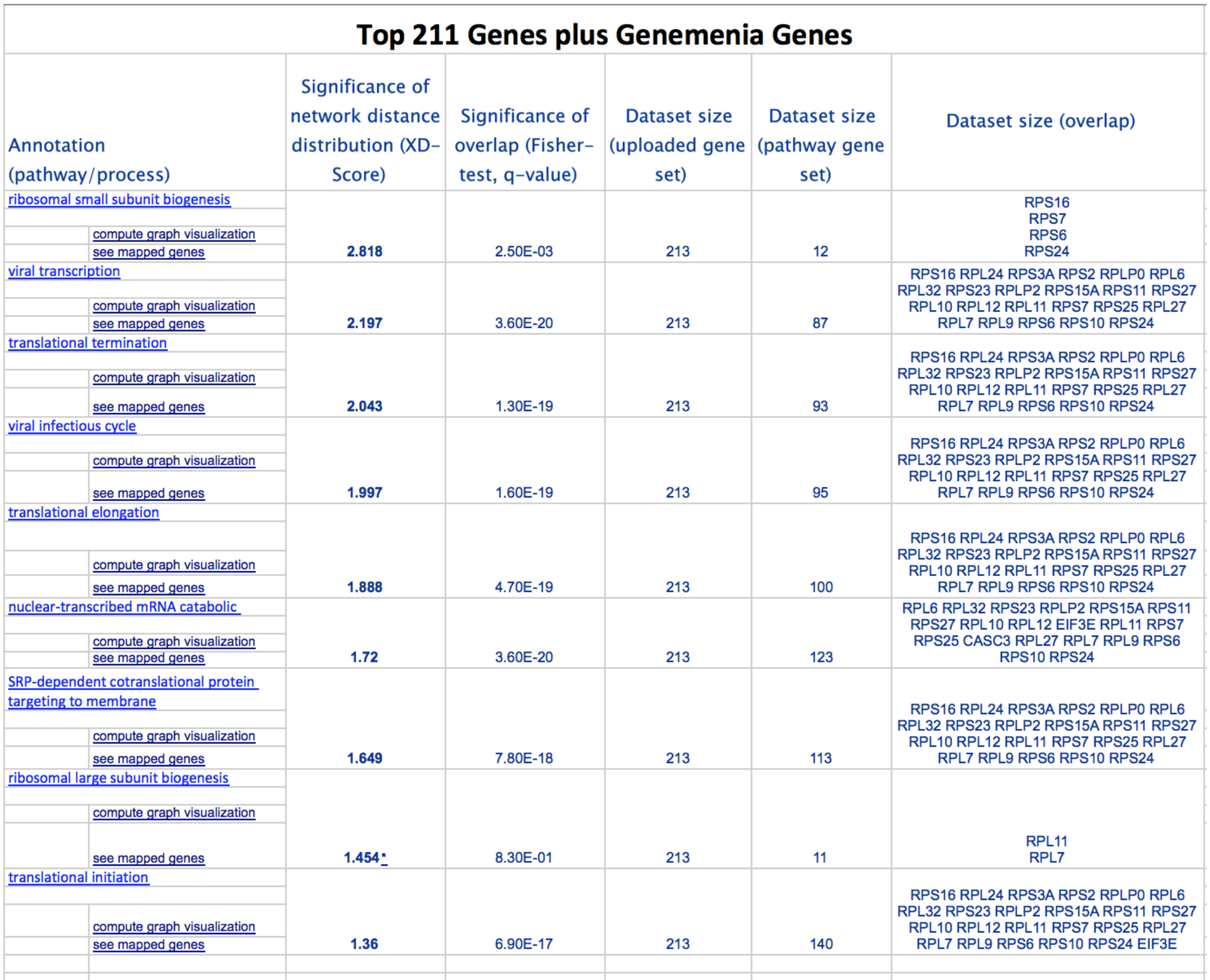
Conditions
73 Genes generated by top 4000
73 Genes plus GeneMANIA additions
211 Genes generated by top 5000
211 Genes plus GeneMANIA additions

I looked at Enrichment for GO Biological Process terms. Results were only significant if I used the additional genes contributed by GeneMANIA.



Top 73 Genes plus Genemania Genes

Annotation (pathway/process)	Significance of network distance distribution (XD–Score)	Significance of overlap (Fisher-test, q–value)	Dataset size (uploaded gene set)	Dataset size (pathway gene set)	Dataset size (overlap)
ribosomal small subunit biogenesis compute graph visualization see mapped genes	**2.16947**	5.90E-03	88	12	RPS7 RPS6 RPS24
ribosomal large subunit biogenesis compute graph visualization see mapped genes	**1.55584***	2.00E-01	88	11	RPL14 RPL5
viral transcription compute graph visualization see mapped genes	**1.36775**	6.80E-14	88	87	RPS3A RPLP0 RPL6 RPLP2 RPS15A RPL14 RPS11 RPL12 RPS7 RPL9 RPLP1 RPS6 RPL5 RPS24
translational termination compute graph visualization see mapped genes	**1.27431**	1.10E-13	88	93	RPS3A RPLP0 RPL6 RPLP2 RPS15A RPL14 RPS11 RPL12 RPS7 RPL9 RPLP1 RPS6 RPL5 RPS24
viral infectious cycle compute graph visualization see mapped genes	**1.24579**	1.20E-13	88	95	RPS3A RPLP0 RPL6 RPLP2 RPS15A RPL14 RPS11 RPL12 RPS7 RPL9 RPLP1 RPS6 RPL5 RPS24
regulation of osteoblast differentiation compute graph visualization see mapped genes	**1.20519***	3.00E-01	88	14	PIAS2 DDX5
translational elongation compute graph visualization see mapped genes	**1.17947**	2.20E-13	88	100	RPS3A RPLP0 RPL6 RPLP2 RPS15A RPL14 RPS11 RPL12 RPS7 RPL9 RPLP1 RPS6 RPL5 RPS24
nuclear-transcribed mRNA catabolic compute graph visualization see mapped genes	**1.0902**	1.90E-14	88	123	UPF3A RPS3A RPLP0 RPL6 RPLP2 RPS15A RPL14 RPS11 RPL12 RPS7 UPF3B RPL9 RPLP1 RPS6 RPL5 RPS24
SRP-dependent cotranslational protein compute graph visualization see mapped genes	**1.03452**	1.10E-12	88	113	RPS3A RPLP0 RPL6 RPLP2 RPS15A RPL14 RPS11 RPL12 RPS7 RPL9 RPLP1 RPS6 RPL5 RPS24



Top 211 Genes plus Genemenia Genes

Annotation (pathway/process)	Significance of network distance distribution (XD-Score)	Significance of overlap (Fisher-test, q-value)	Dataset size (uploaded gene set)	Dataset size (pathway gene set)	Dataset size (overlap)
ribosomal small subunit biogenesis compute graph visualization see mapped genes	2.818	2.50E-03	213	12	RPS16 RPS7 RPS6 RPS24
viral transcription compute graph visualization see mapped genes	2.197	3.60E-20	213	87	RPS16 RPL24 RPS3A RPS2 RPLP0 RPL6 RPL32 RPS23 RPLP2 RPS15A RPS11 RPS27 RPL10 RPL12 RPL11 RPS7 RPS25 RPL27 RPL7 RPL9 RPS6 RPS10 RPS24
translational termination compute graph visualization see mapped genes	2.043	1.30E-19	213	93	RPS16 RPL24 RPS3A RPS2 RPLP0 RPL6 RPL32 RPS23 RPLP2 RPS15A RPS11 RPS27 RPL10 RPL12 RPL11 RPS7 RPS25 RPL27 RPL7 RPL9 RPS6 RPS10 RPS24
viral infectious cycle compute graph visualization see mapped genes	1.997	1.60E-19	213	95	RPS16 RPL24 RPS3A RPS2 RPLP0 RPL6 RPL32 RPS23 RPLP2 RPS15A RPS11 RPS27 RPL10 RPL12 RPL11 RPS7 RPS25 RPL27 RPL7 RPL9 RPS6 RPS10 RPS24
translational elongation compute graph visualization see mapped genes	1.888	4.70E-19	213	100	RPS16 RPL24 RPS3A RPS2 RPLP0 RPL6 RPL32 RPS23 RPLP2 RPS15A RPS11 RPS27 RPL10 RPL12 RPL11 RPS7 RPS25 RPL27 RPL7 RPL9 RPS6 RPS10 RPS24
nuclear-transcribed mRNA catabolic compute graph visualization see mapped genes	1.72	3.60E-20	213	123	RPL6 RPL32 RPS23 RPLP2 RPS15A RPS11 RPS27 RPL10 RPL12 EIF3E RPL11 RPS7 RPS25 CASC3 RPL27 RPL7 RPL9 RPS6 RPS10 RPS24
SRP-dependent cotranslational protein targeting to membrane compute graph visualization see mapped genes	1.649	7.80E-18	213	113	RPS16 RPL24 RPS3A RPS2 RPLP0 RPL6 RPL32 RPS23 RPLP2 RPS15A RPS11 RPS27 RPL10 RPL12 RPL11 RPS7 RPS25 RPL27 RPL7 RPL9 RPS6 RPS10 RPS24
ribosomal large subunit biogenesis compute graph visualization see mapped genes	1.454*	8.30E-01	213	11	RPL11 RPL7
translational initiation compute graph visualization see mapped genes	1.36	6.90E-17	213	140	RPS16 RPL24 RPS3A RPS2 RPLP0 RPL6 RPL32 RPS23 RPLP2 RPS15A RPS11 RPS27 RPL10 RPL12 RPL11 RPS7 RPS25 RPL27 RPL7 RPL9 RPS6 RPS10 RPS24 EIF3E

From this it is clear again that the ribosome-associated genes seem to have the highest functional enrichment. Both transcription and translation are represented, as well as the ribosomal structures themselves.

I then had a go at inputting the 211 genes+GM into PATHER. Below is the table of significant results (p<.05) . . . .

GO biological process complete	Fold Enrichment	P value
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (GO:0000184)	> 5	1.12E-21
nuclear-transcribed mRNA catabolic process (GO:0000956)	> 5	8.26E-21
mRNA catabolic process (GO:0006402)	> 5	4.76E-20
SRP-dependent cotranslational protein targeting to membrane (GO:0006614)	> 5	1.64E-18
RNA catabolic process (GO:0006401)	> 5	2.11E-18
viral transcription (GO:0019083)	> 5	2.45E-18
cotranslational protein targeting to membrane (GO:0006613)	> 5	2.98E-18
protein targeting to ER (GO:0045047)	> 5	3.62E-18
establishment of protein localization to endoplasmic reticulum (GO:0072599)	> 5	7.76E-18
viral gene expression (GO:0019080)	> 5	1.62E-17
multi-organism metabolic process (GO:0044033)	> 5	4.66E-17
protein localization to endoplasmic reticulum (GO:0070972)	> 5	2.07E-16
organonitrogen compound biosynthetic process (GO:1901566)	4.66	2.16E-16

GO biological process complete	Fold Enrichment	P value
interspecies interaction between organisms (GO:0044419)	> 5	5.54E-16
symbiosis, encompassing mutualism through parasitism (GO:0044403)	> 5	5.54E-16
mRNA metabolic process (GO:0016071)	> 5	2.93E-15
viral process (GO:0016032)	> 5	3.56E-15
multi-organism cellular process (GO:0044764)	> 5	4.38E-15
protein targeting to membrane (GO:0006612)	> 5	1.32E-14
translational termination (GO:0006415)	> 5	2.89E-14
viral life cycle (GO:0019058)	> 5	3.35E-14
translational elongation (GO:0006414)	> 5	6.60E-14
nucleobase-containing compound catabolic process (GO:0034655)	> 5	2.24E-13
establishment of protein localization to membrane (GO:0090150)	> 5	6.36E-13
cellular protein complex disassembly (GO:0043624)	> 5	8.07E-13
aromatic compound catabolic process (GO:0019439)	> 5	1.04E-12
macromolecular complex disassembly (GO:0032984)	> 5	1.30E-12
translational initiation (GO:0006413)	> 5	1.72E-12
amide biosynthetic process (GO:0043604)	> 5	3.25E-12
peptide metabolic process (GO:0006518)	> 5	3.71E-12
cellular nitrogen compound catabolic process (GO:0044270)	> 5	4.29E-12
heterocycle catabolic process (GO:0046700)	> 5	4.29E-12
peptide biosynthetic process (GO:0043043)	> 5	4.87E-12
protein complex disassembly (GO:0043241)	> 5	5.99E-12
cellular metabolic process (GO:0044237)	1.64	6.27E-12
translation (GO:0006412)	> 5	8.75E-12
cellular macromolecule catabolic process (GO:0044265)	4.7	8.94E-12
macromolecule catabolic process (GO:0009057)	4.3	1.26E-11
establishment of protein localization to organelle (GO:0072594)	> 5	2.18E-11
cellular amide metabolic process (GO:0043603)	4.98	2.34E-11
cytoplasmic transport (GO:0016482)	4.78	3.21E-11
organic cyclic compound catabolic process (GO:1901361)	> 5	5.48E-11
cellular catabolic process (GO:0044248)	3.31	1.28E-10
organonitrogen compound metabolic process (GO:1901564)	3.04	1.82E-10
metabolic process (GO:0008152)	1.52	2.15E-10
protein localization to membrane (GO:0072657)	> 5	2.83E-10
protein targeting (GO:0006605)	> 5	3.07E-10
organic substance metabolic process (GO:0071704)	1.57	3.25E-10
establishment of localization in cell (GO:0051649)	2.95	3.72E-10
primary metabolic process (GO:0044238)	1.59	6.31E-10
biological_process (GO:0008150)	1.21	7.54E-10
macromolecular complex subunit organization (GO:0043933)	2.75	8.02E-10
intracellular protein transport (GO:0006886)	4.43	1.65E-09
cellular localization (GO:0051641)	2.67	1.90E-09
single-organism intracellular transport (GO:1902582)	3.47	4.10E-09
cellular macromolecule metabolic process (GO:0044260)	1.7	9.83E-09
intracellular transport (GO:0046907)	3.2	9.93E-09
protein localization to organelle (GO:0033365)	> 5	1.01E-08
cellular component disassembly (GO:0022411)	> 5	1.17E-08
cellular protein metabolic process (GO:0044267)	2.14	1.39E-08
catabolic process (GO:0009056)	2.82	2.01E-08
single-organism cellular localization (GO:1902580)	3.92	2.40E-08
nitrogen compound metabolic process (GO:0006807)	1.82	3.04E-08
organic substance catabolic process (GO:1901575)	2.99	3.36E-08

GO biological process complete	Fold Enrichment	P value
cellular macromolecule localization (GO:0070727)	3.38	3.65E-08
protein complex subunit organization (GO:0071822)	3.08	3.76E-08
cellular process (GO:0009987)	1.29	4.66E-08
multi-organism process (GO:0051704)	2.55	5.95E-08
protein transport (GO:0015031)	3.24	7.56E-08
organic substance biosynthetic process (GO:1901576)	1.91	1.12E-07
cellular protein localization (GO:0034613)	3.32	1.16E-07
single-organism membrane organization (GO:0044802)	4.12	1.28E-07
biosynthetic process (GO:0009058)	1.89	1.29E-07
membrane organization (GO:0061024)	3.68	2.95E-07
organic substance transport (GO:0071702)	2.62	3.29E-07
macromolecule metabolic process (GO:0043170)	1.6	4.14E-07
establishment of protein localization (GO:0045184)	3.03	5.99E-07
cellular nitrogen compound metabolic process (GO:0034641)	1.8	6.67E-07
macromolecule localization (GO:0033036)	2.52	9.23E-07
cellular biosynthetic process (GO:0044249)	1.87	1.44E-06
cellular component organization or biogenesis (GO:0071840)	1.78	1.60E-06
protein localization (GO:0008104)	2.67	1.68E-06
single-organism localization (GO:1902578)	2.07	2.02E-06
cellular component organization (GO:0016043)	1.78	2.33E-06
protein metabolic process (GO:0019538)	1.89	2.50E-06
ribonucleoprotein complex biogenesis (GO:0022613)	> 5	2.62E-06
single-organism process (GO:0044699)	1.31	2.92E-06
cellular nitrogen compound biosynthetic process (GO:0044271)	2.03	3.58E-06
localization (GO:0051179)	1.81	5.61E-06
single-organism transport (GO:0044765)	2.08	6.45E-06
transport (GO:0006810)	1.94	7.06E-06
gene expression (GO:0010467)	1.93	8.12E-06
cellular aromatic compound metabolic process (GO:0006725)	1.81	9.25E-06
organic cyclic compound biosynthetic process (GO:1901362)	2.03	1.66E-05
organic cyclic compound metabolic process (GO:1901360)	1.76	2.55E-05
ribosome biogenesis (GO:0042254)	> 5	2.64E-05
establishment of localization (GO:0051234)	1.89	2.72E-05
nucleobase-containing compound metabolic process (GO:0006139)	1.81	2.84E-05
RNA metabolic process (GO:0016070)	1.97	3.65E-05
aromatic compound biosynthetic process (GO:0019438)	2.03	4.73E-05
heterocycle metabolic process (GO:0046483)	1.77	4.96E-05
nucleic acid metabolic process (GO:0090304)	1.85	8.50E-05
nucleobase-containing compound biosynthetic process (GO:0034654)	2.02	9.91E-05
cellular component biogenesis (GO:0044085)	2.33	1.04E-04
heterocycle biosynthetic process (GO:0018130)	2	1.16E-04
macromolecular complex assembly (GO:0065003)	2.74	5.41E-04
single-organism cellular process (GO:0044763)	1.32	5.45E-04
cellular macromolecular complex assembly (GO:0034622)	3.64	5.47E-04
ribonucleoprotein complex subunit organization (GO:0071826)	> 5	1.99E-03
RNA biosynthetic process (GO:0032774)	1.96	3.07E-03
ribonucleoprotein complex assembly (GO:0022618)	> 5	7.63E-03
macromolecule biosynthetic process (GO:0009059)	1.74	1.02E-02
ribosome assembly (GO:0042255)	> 5	1.40E-02
negative regulation of biological process (GO:0048519)	1.64	1.45E-02
negative regulation of cellular process (GO:0048523)	1.67	2.04E-02

GO biological process complete	Fold Enrichment	P value
cellular macromolecule biosynthetic process (GO:0034645)	1.73	2.12E-02
ribosomal large subunit assembly (GO:0000027)	> 5	3.17E-02
ribosomal large subunit biogenesis (GO:0042273)	> 5	4.10E-02
cellular component assembly (GO:0022607)	2.09	4.35E-02

Friday - Lab Meeting

At the lab meeting, I explained that I now have a slightly different list of genes, but my numbers are significant. I said that there seems to be particular emphasis on the ribosomal proteins, but Win said to be wary because it could just be a literature bias. I explained that I'm not really sure what to do next, but one thing I definitely thought I should do is try and do some GEA on the RNA seq data I have found. Win said this was a good idea, but he also wants both me and John to talk to Jiantao about how the package EdgeRun could help us build co-expression networks that correlate with the gene lists we have generated.