LabBook_15_04_16

Claire Green

11 April 2016

Monday

I started with calculating the enrichment of my DEGs in the pathprint pathway gene list. I did so using this script:

```
pathprint <- read.table(file = "Pathprintgenes.txt")
pathprint <- pathprint$V1

pathprintunique <- pathprint[!duplicated(pathprint)]

overlap <- Reduce(intersect, list(x, pathprint))
print(overlap)


x <- read.table(file = "DEGs.txt")
x <- x$V1

library(hgu133plus2.db)
sym <- hgu133plus2SYMBOL
sym1 <- mappedkeys(sym)
sym2 <- as.list (sym[c(sym1)])
sym3 <- data.frame (sym2)
sym.probes <- names (sym2)
sym.genes <- sym3[1,]

x.in <- length (which(x %in% pathprintunique))
x.out <- length(x) - x.in
tot.in <- length (pathprintunique)
tot.out <- length (sym.genes)

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

a5 <-fisher.test (counts)
enrich <- a5$p
```

Overlapping genes were "KPNA6" "NUTF2" "PLOD2" "PPP2CA" "PPP2CB"

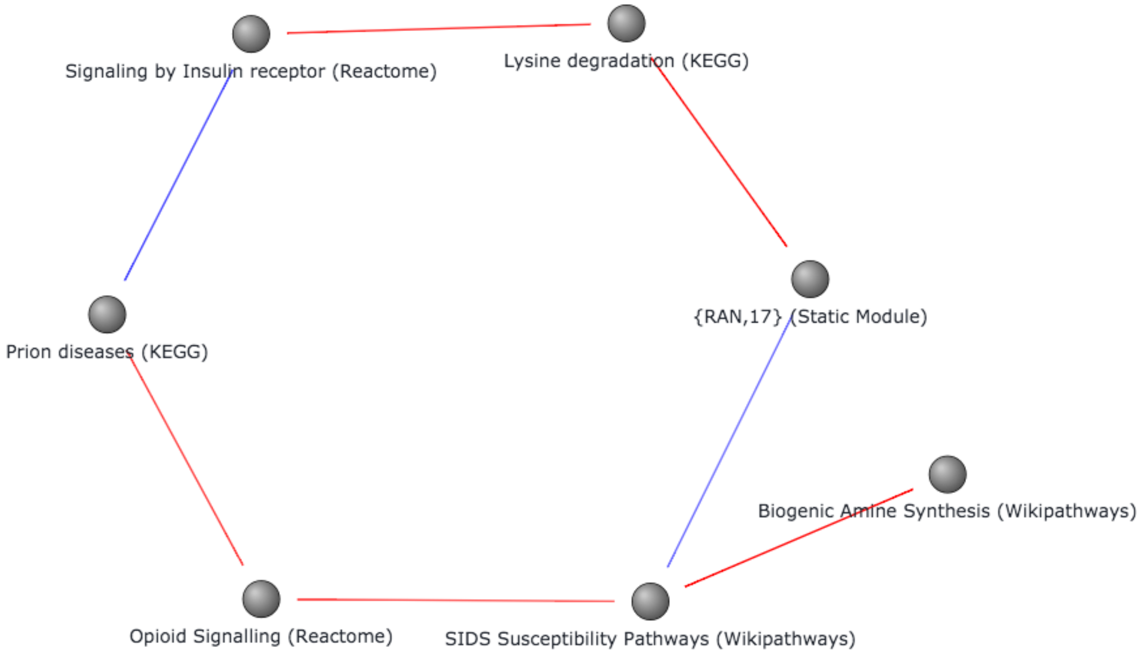
The result was that when the duplicates were not removed, significance was 0.004. With duplicates removed, significance was p = 0.002. This means that my 5 genes are significantly enriched in the pp gene list as compared to the proportion of all genes represented by the list.

I was interested to see if the Pathprint genes were enriched with SNPs. I used the no-duplicates pathprint list and the GWAS central list of SNPs where p<.0001 ("signif.snp.GWAScentral.p0.0001.1.txt"). There were 4 genes containing SNPS ("KCNQ1" "PPARGC1A" "GNG7" "ITPR2") but no significant enrichment.

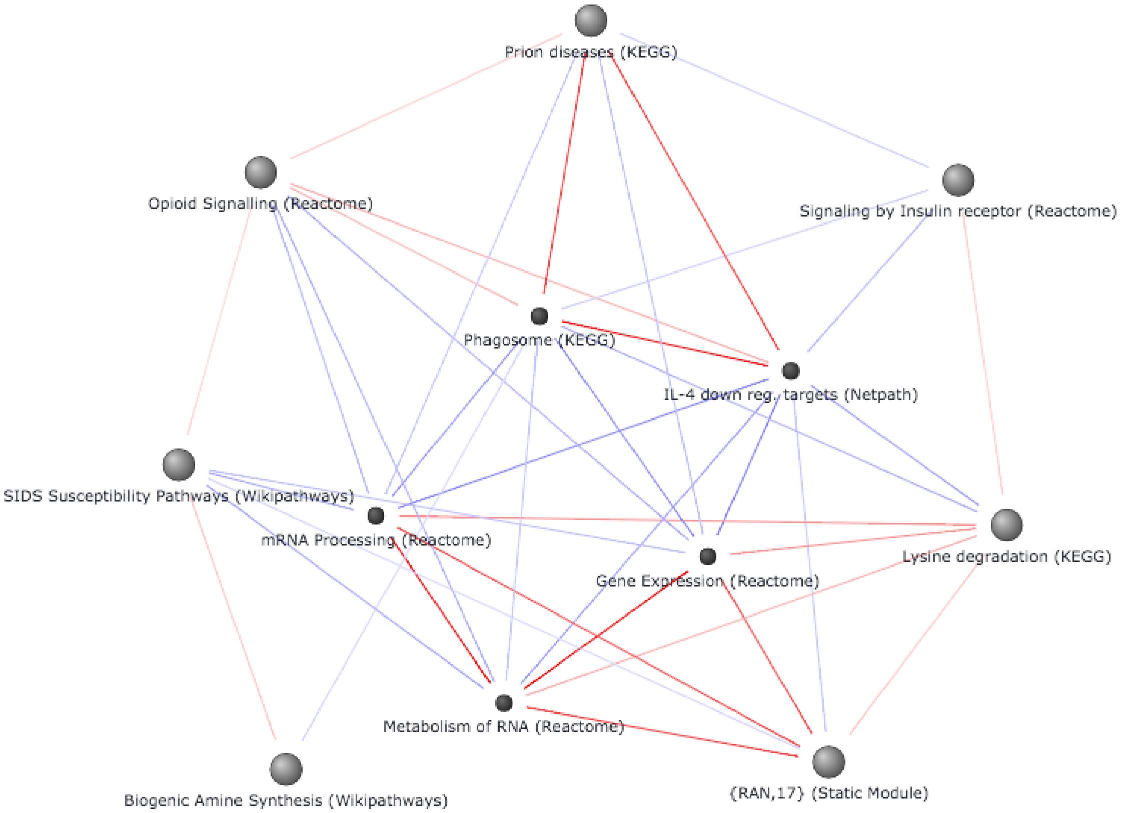
Next, I looked at the NeuroX list. There were 7 genes overlapping these lists (“NOTCH1” “SOD1” “COMT” “CHRNA4” “FGFR3” “STK11” “TSC2”) and this enrichment was significant (p = 0.0002)

When I look at the pathways in which these are enriched, Prion diseases (KEGG) contains 2 NeuroX genes (SOD1, NOTCH1), Biogenic Amine Synthesis (KEGG) contains 1 (COMT), SIDS susceptibility pathway (Wikipathways) contains 1 neuroX (CHRNA4) and two GWAS central (KCNQ1, PPARGC1A), Signalling by insulin receptor (Reactome) contains 3 neuroX genes (FGFR3, STK11, TSC2), Opioid signalling (Reactome) contains 2 GWAS central genes (GNG7, ITPR2) and 2 DEGS (PPP2CA, PPP2CB).

The first thing I did was put these pathways in PCxN, including RAN,17 which contains 2 DEGs.

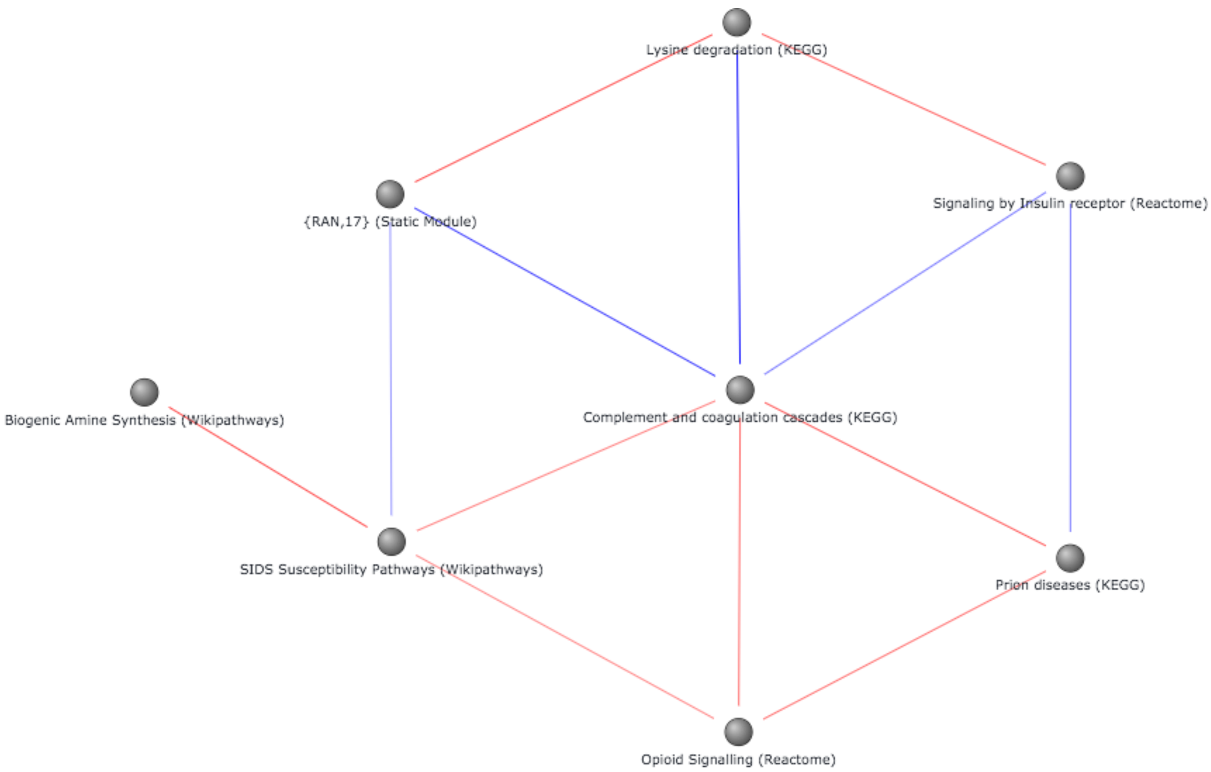

Signaling by Insulin receptor (Reactome)
Lysine degradation (KEGG)
{RAN,17} (Static Module)
Prion diseases (KEGG)
Biogenic Amine Synthesis (Wikipathways)
Opioid Signalling (Reactome)
SIDS Susceptibility Pathways (Wikipathways)

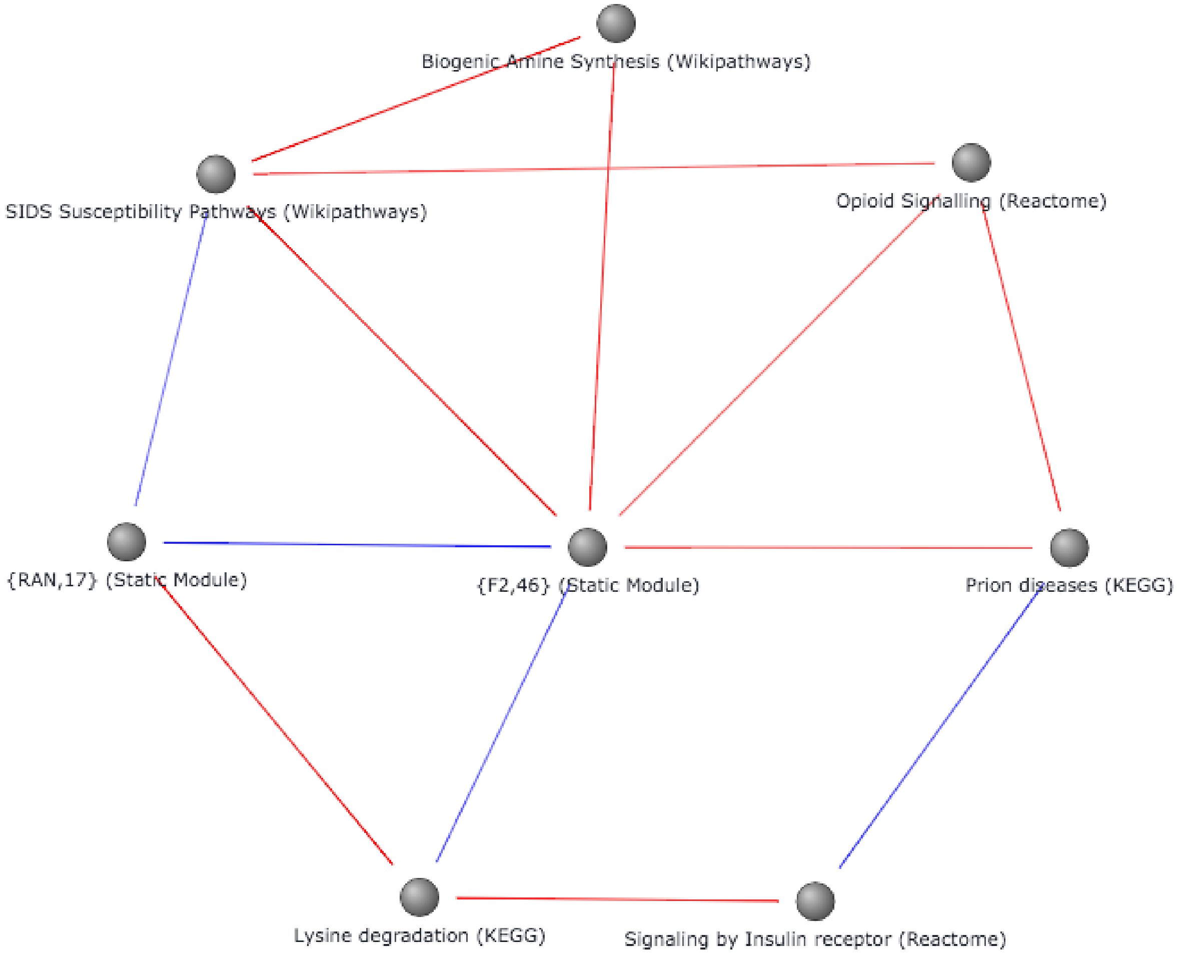
It’s not particularly connected, so I tried adding the 5 most correlated gene sets

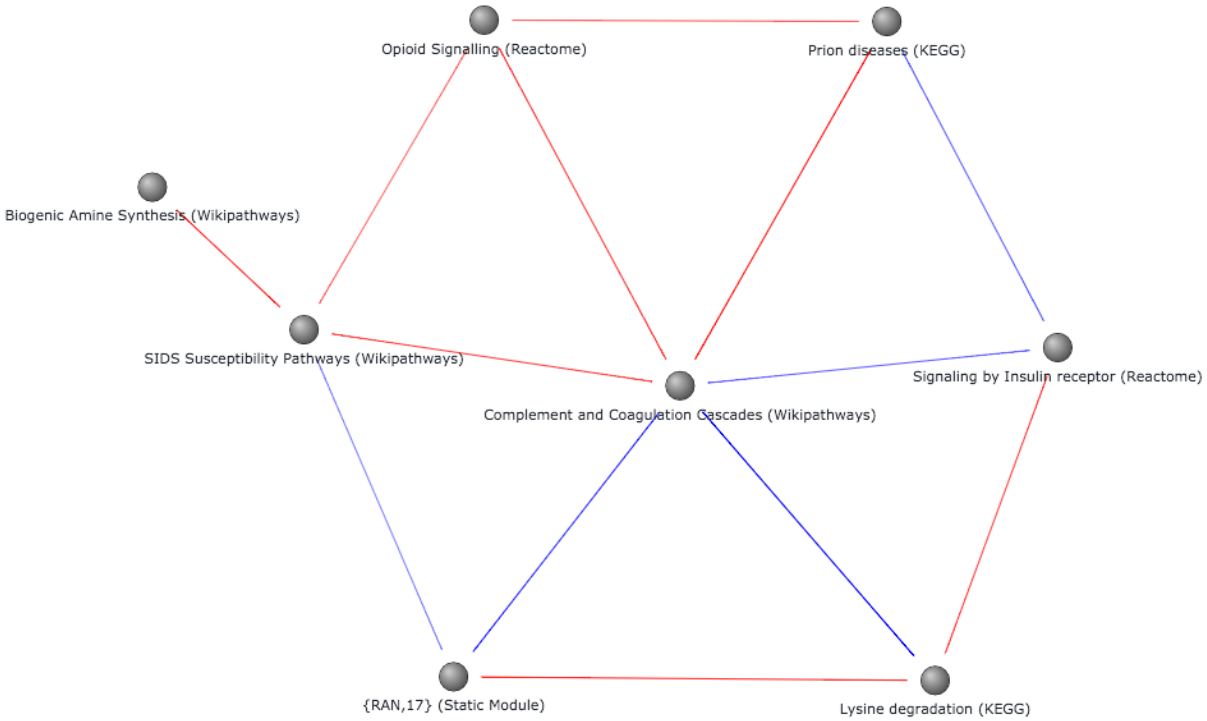

Prion diseases (KEGG)
Opioid Signalling (Reactome)
Signaling by Insulin receptor (Reactome)
Phagosome (KEGG)
IL-4 down reg. targets (Netpath)
SIDS Susceptibility Pathways (Wikipathways)
mRNA Processing (Reactome)
Gene Expression (Reactome)
Lysine degradation (KEGG)
Metabolism of RNA (Reactome)
Biogenic Amine Synthesis (Wikipathways)
{RAN,17} (Static Module)

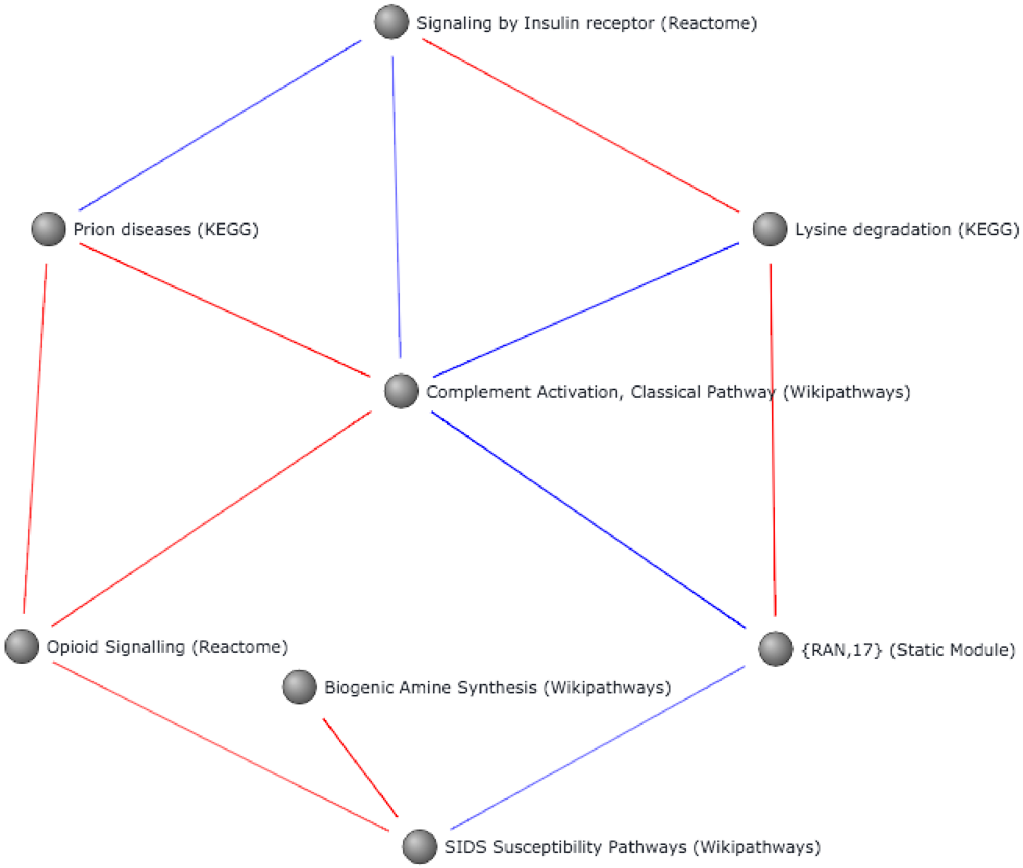
What we get out is Phagosome, IL-4 down reg. targets, mRNA processing, metabolism of RNA and gene expression. The last three can particularly be linked to TDP-43 activity.

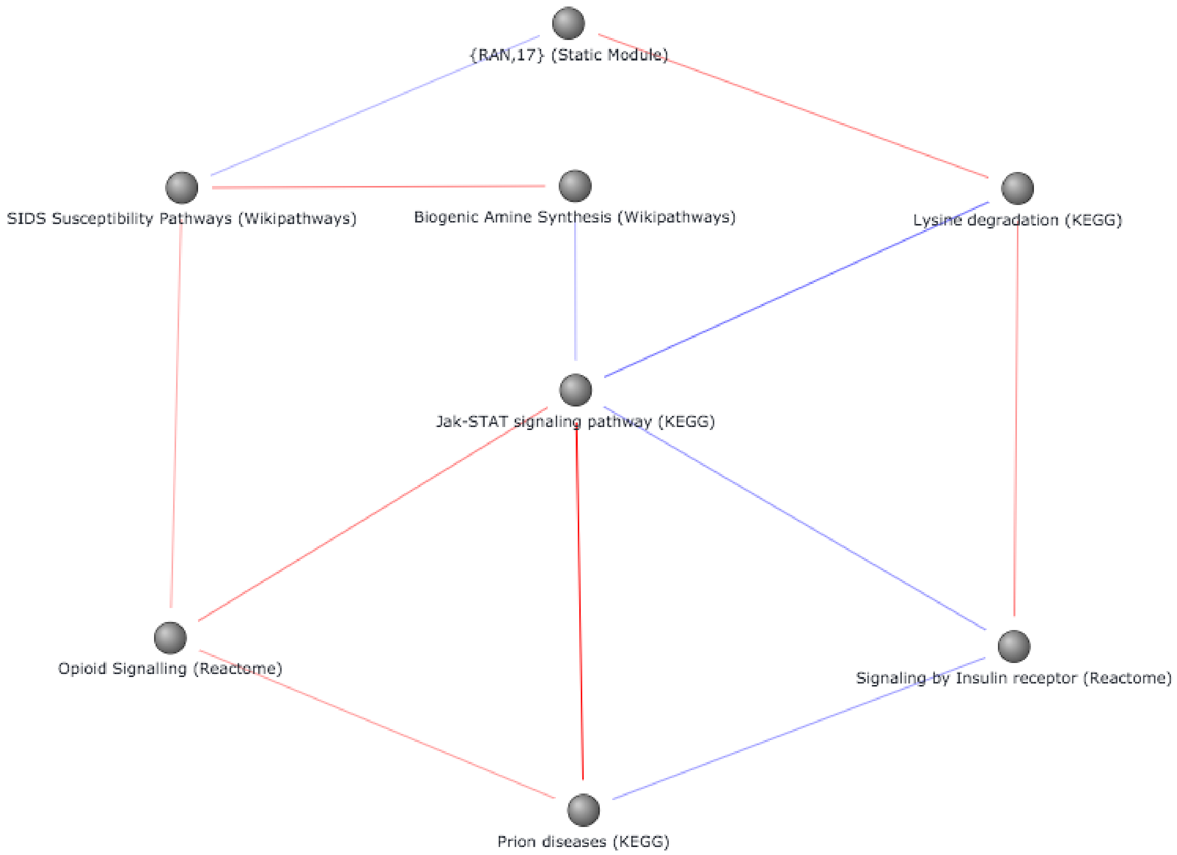
What I then tried to do is discover if any of the other pathprint pathways were able to connect the enriched pathways better. I identified 5 pathways:


Lysine degradation (KEGG)
{RAN,17} (Static Module)
Signaling by Insulin receptor (Reactome)
Biogenic Amine Synthesis (Wikipathways)
Complement and coagulation cascades (KEGG)
SIDS Susceptibility Pathways (Wikipathways)
Prion diseases (KEGG)
Opioid Signalling (Reactome)


Biogenic Amine Synthesis (Wikipathways)
SIDS Susceptibility Pathways (Wikipathways)
Opioid Signalling (Reactome)
{RAN,17} (Static Module)
{F2,46} (Static Module)
Prion diseases (KEGG)
Lysine degradation (KEGG)
Signaling by Insulin receptor (Reactome)


Opioid Signalling (Reactome)
Prion diseases (KEGG)
Biogenic Amine Synthesis (Wikipathways)
SIDS Susceptibility Pathways (Wikipathways)
Signaling by Insulin receptor (Reactome)
Complement and Coagulation Cascades (Wikipathways)
{RAN,17} (Static Module)
Lysine degradation (KEGG)


Signaling by Insulin receptor (Reactome)
Prion diseases (KEGG)
Lysine degradation (KEGG)
Complement Activation, Classical Pathway (Wikipathways)
Opioid Signalling (Reactome)
{RAN,17} (Static Module)
Biogenic Amine Synthesis (Wikipathways)
SIDS Susceptibility Pathways (Wikipathways)

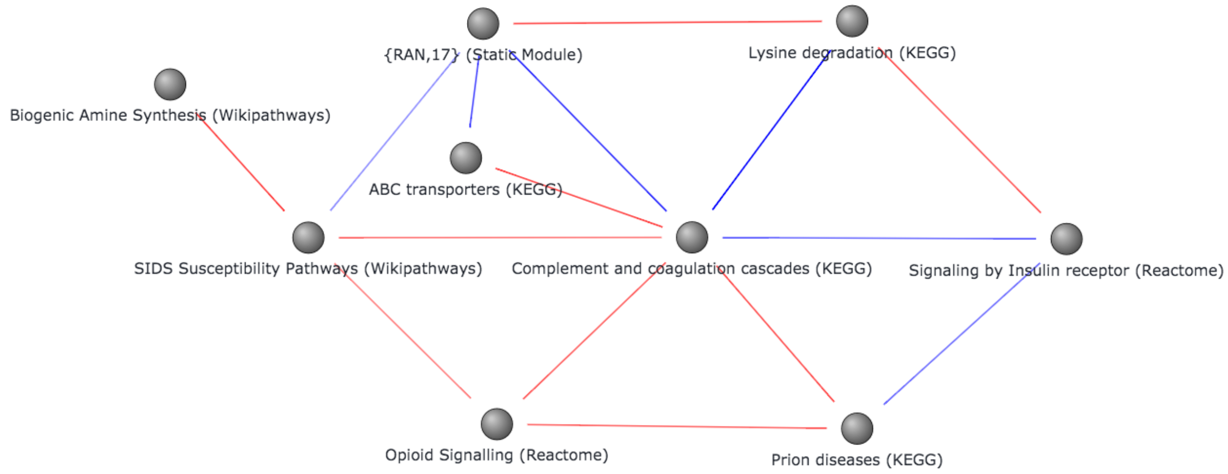

{RAN,17} (Static Module)
SIDS Susceptibility Pathways (Wikipathways)
Biogenic Amine Synthesis (Wikipathways)
Lysine degradation (KEGG)
Jak-STAT signaling pathway (KEGG)
Opioid Signalling (Reactome)
Signaling by Insulin receptor (Reactome)
Prion diseases (KEGG)

It appears that complemet and coagulation cascades seems to be the pathway that links these enriched pathways together.

Next, I downloaded all the SNPs identified by GWAS catalog and added them in. Now the pathways are this:

	GWAS		GWAS	
Pathways	Central	NeuroX	catalog	DEGS
Lysine degradation (KEGG)	x	x	HADH	PLOD2
ABC transporters (KEGG)	x	x	ABCG1 ABCC12	x
Complement and coagulation cascades (KEGG)	x	x	MASP1	x
Prion diseases (KEGG)	x	NOTCH1 SOD1	x	x
Biogenic Amine Synthesis (Wikipathways)	x	COMT	x	x
SIDS Susceptibility Pathways (Wikipathways)	KCNQ1 PPARGC1A	CHRNA4	x	x
Signaling by Insulin receptor (Reactome)	x	FGFR3 STK11 TSC2	x	x
Opioid Signalling (Reactome)	GNG7 ITPR2			PPP2CA PPP2CB
{RAN,17} (Static Module)	x	x	x	NUTF2 KPNA6

Now we can confirm the inclusion of complement and coagulation cascades.


{RAN,17} (Static Module)
Lysine degradation (KEGG)
Biogenic Amine Synthesis (Wikipathways)
ABC transporters (KEGG)
SIDS Susceptibility Pathways (Wikipathways)
Complement and coagulation cascades (KEGG)
Signaling by Insulin receptor (Reactome)
Opioid Signalling (Reactome)
Prion diseases (KEGG)

Tuesday

Since I was able to download all the SNPs from GWAS catalog for ALS, I thought I would try to do the same for Alzheimer's disease. What I did was downlaod the table of all associations, extract the column marked MAPPED GENES, and separate each gene into a cell after which I remove any duplicates. This meant that some genes were duplicated due to different names, but it allowed me to identify genes in my pathways that perhaps had alternative names. I wouldn't do stats analysis on this, it's just for identification purposes. I intersected the two lists and identified 23 genes containing AD SNPS.

"SORD" "FUK" "CPS1" "ABCA1" "ABCC9" "ABCA7" "CR1" "F13A1" "CREBBP" "IL6R" "JAK2" "IL19" "IL21" "NCAM1" "KCNQ1" "RORA" "FGF1" "IRS1" "STK11" "ADCY5" "ADCY8" "CAMK4" "PDE1A"

And I did the same for the DEG list:

"PFDN1"

I'm not sure the second list (of one) will be significant, but the first could.

PCxN analysis

Win asked me to conduct GSEA on an ALS data set, take the enriched pathways, run them through PCxN, add pathways using PCxN, and then look for enrichment of ALS disease-associated loci.

GSEA

I talked to Gabriel and we decided that using the GSEA java platform is probably the best to use instead of the R script, as most people are going to use this method and the R script isn't supported any more.

I used the following settings:

Expression Dataset - C9 [35647x11, chip, na] Gene Sets Database - gseaftp.broadinstitute.org://pub/gsea/gene_sets/c5.bp.v5.1.s Number of permutations - 1000 Phenotype labels - /Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/GSEA/C9orf72 with outlier/C9_pheno.cls#PAT_versus_CON Collapse dataset to gene symbols - true Permutation type - phenotype Chip platform(s) - gseaftp.broadinstitute.org://pub/gsea/annotations/HG_U133_Plus_2.ch

Basic and Advanced fields were left unchanged.

I also conducted an identical analysis using a permutation type of permytating gene sets, as it was recommended that for small data sets this is a better way of calculating significance.

Pheno-perm results

80 / 274 gene sets are upregulated in phenotype PAT 0 gene sets are significant at FDR < 25% 1 gene sets are significantly enriched at nominal pvalue < 1% 5 gene sets are significantly enriched at nominal pvalue < 5%

Those 5 gene sets are:

Protein import Protein targeting Nucleocytoplasmic Transport Protein import into nucleus Nuclear transport

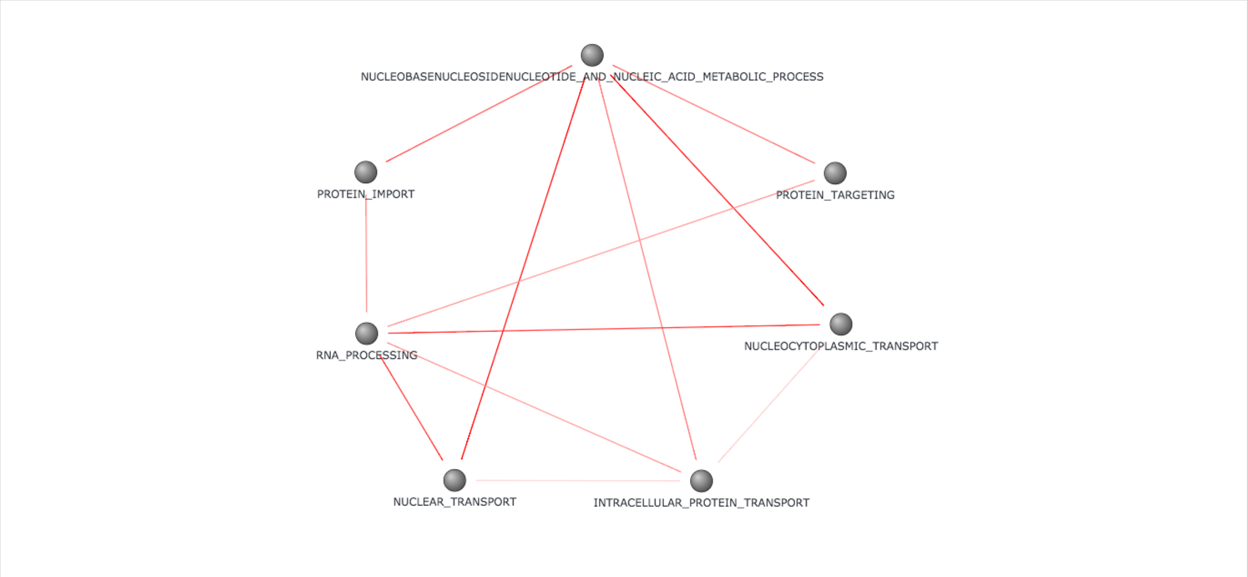
Gene-perm results

80 / 274 gene sets are upregulated in phenotype PAT 0 gene sets are significant at FDR < 25% 1 gene sets are significantly enriched at nominal pvalue < 1% 7 gene sets are significantly enriched at nominal pvalue < 5%

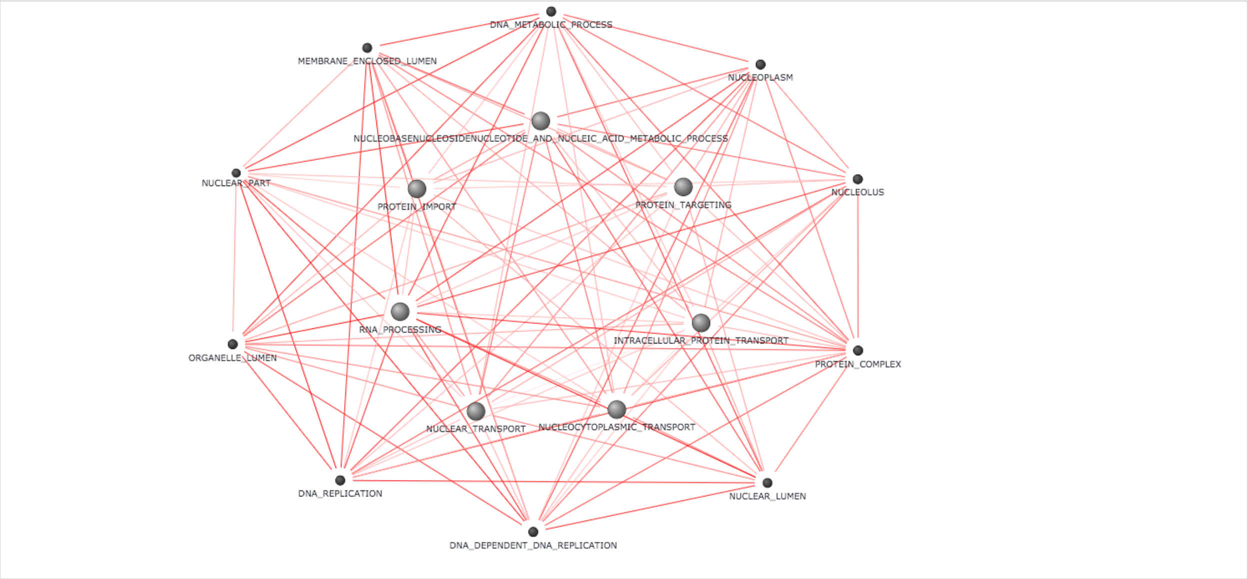
Those 7 gene sets are: Protein targeting Nucleocytoplasmic Transport Intracellular protein transport Protein import Nuclear transport Nuclear import Protein import into nucleus

PCxN

Here are the 7 gene sets in PCxN


NUCLEOBASENUCLEOSIDENUCLEOTIDE_AND_NUCLEIC_ACID_METABOLIC_PROCESS
PROTEIN_IMPORT
PROTEIN_TARGETING
RNA_PROCESSING
NUCLEOCYTOPLASMIC_TRANSPORT
NUCLEAR_TRANSPORT
INTRACELLULAR_PROTEIN_TRANSPORT

And adding the 10 most correlated gene sets



It concerns me a little that the added gene sets are so generic. Looking at the sizes of them as well - some of them are enormous. I'm not sure exactly how this is going to affect the p values later down the line in terms of enrichment.

GWAS enrichment

For this I used 4 different ALS SNP Databases. GWAS central (all) included about 4000 genes. GWAS central (p<0.0001) contained around 2700 genes. NeuroX contained 122 genes, and NeuroX (p<5x10-8) contained 54. I used the following script to calculate the intersect and enrichment significance.

```
#Load database of associations
setwd(dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/GWAS/")

a <- read.table(file = "signif.snp.GWAScentral.txt")
a <- a$V1

b <- read.table(file = "signif.snp.GWAScentral.p0.0001.1.txt")
b <- b$V1

c <- read.table(file = "signif.snp.NeuroX.txt")
c <- c$V1

d <- read.table(file = "signif.snp.NeuroX.p5E08.txt")
d <- d$V1

#load test file

setwd (dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/GSEA/PCxN Example/")
y <- read.table(file = "GSEAgenes.txt")
y <- y$V1

#remove any duplicates
y <- y[!duplicated(y)]
```

```
#Intersect
overlap <- Reduce(intersect, list(y, a))
print(overlap)

#Load file with all genes
library(hgu133plus2.db)
sym <- hgu133plus2SYMBOL
sym1 <- mappedkeys(sym)
sym2 <- as.list (sym[c(sym1)])
sym3 <- data.frame (sym2)
sym.probes <- names (sym2)
sym.genes <- sym3[1,]

#calculate counts

x.in <- length (which(y %in% a))
x.out <- length(y) - x.in
tot.in <- length (a)
tot.out <- length (sym.genes)

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 <- a5$p
print(enrich)
```

I compared results depending on whether I removed duplicates or not, as I wasn't sure which was statistically the right representation. Often pathways contain the same genes but it's not always clear that this is because the pathways are too overlapping or the gene just has multiple roles.

GSEA only	Dup.Gene	Dup.pvalue	noDup.Gene	noDup.pvalue
GWAS Central	89	0.316	64	0.387
GWAS Cental (p<0.001)	17	0.6	13	0.537
NeuroX	11	0.01	11	9.5x10-4
NeuroX (p<5x10-8)	4	0.166	4	0.07

GSEA +10	Dup.Gene	Dup.pvalue	noDup.Gene	noDup.pvalue
GWAS Central	225	0.0037	96	0.19
GWAS Cental (p<0.001)	57	0.03	19	0.6
NeuroX	47	9.85x10-12	17	5x10-5
NeuroX (p<5x10-8)	23	4.55x10-7	9	9x10-4