LabBook_08_04_16

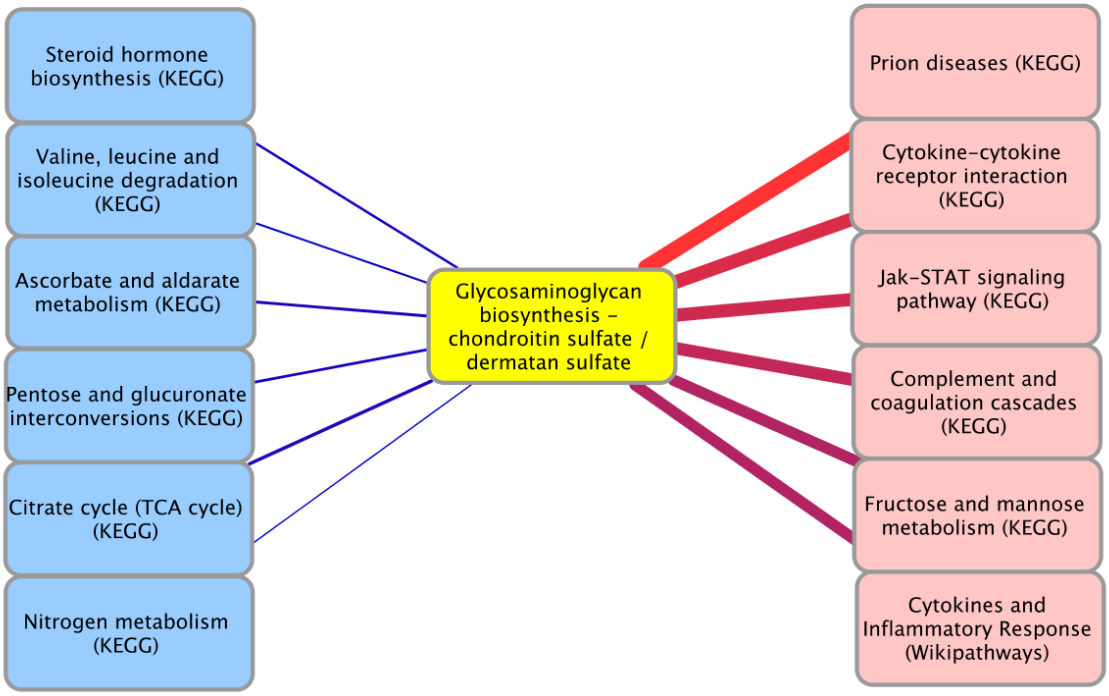
Claire Green

Monday

Today I tidied up a lot of the geneMANIA stuff I did last week. It seems that there are some genes that more consistently come up as coexpressed with the pathways than others. Namely, ACTN1, CST3, ANXA1, PPP2CA, PPP2CB, SPARC, BGN, DCN, and PLOD2. ANXA1 is typically associated with inflammatory responses, ACTN1 and SPARC with platelet activation, SPARC, DCN, CST3, BGN, ACTN1 & PLOD2 with extracellular matrix organisation, and PPP2CA and PPP2CB with fibroblast growth receptor activation (though whether this has anything to do with neurodegeneration I don't know).

Platelet activation has been recently implicated in neurodegeneration (Mazereeuw et al. 2013, "Platelet activating factors in depression and coronary artery disease: A potential biomarker related to inflammatory mechanisms and neurodegeneration")

I looked this afternoon at comparing the DEG enriched pathways and the pathprint pathways. This was not easy as GeneMANIA uses GO terms whereas pathprint uses KEGG, Wikipathways etc. I had to try and find equivalents of the GO pathways in the Pathprint database, which I think is super sketchy but I didn't have a choice. To be honest, I'm just trying to get somthing interesting for the presentation so it doesn't have to be groundbreaking. It was interesting though - I could only find a pathway related to the dermatan/chondroitin metabolism, but when I correlated that with the pathprint pathways I ended up with metabolic pathways being negatively correlated, and immunity/inflammation pathways positively correlated.



Tuesday

I worked a little on the presentation today, still trying to cut it down as it's far too long. In the mean time John had sent me some different GWAS lists, which tightened the thresholds of significance (and therefore lost me some genes). I also went through GWAS catalog and looked to see if there were any on there, and I found that there were some in there that had been reported. So overall:

GWAS central (ALS) CSRP1 (0.0006729)
RAB40B (0.0005151)
SERBP1 (0.0008722)
TUG1 (0.0006935)

GWAS catalog PFDN1 (Alzheimer's Disease, 3x10-7)
ZFYVE26 (ALS, 6x10-7)

Note that TUG1 is a non-protein-coding gene so is often missed. For full details on the SNPs, go to the file ALLDEGs.xlsx

Wednesday

Today I looked at extending my gene list to see if there are any SNP-containing genes associated with my seed list. I generated a list of 95 by including 50 more genes by genemania (co-expression, genetic interaction, physical interaction, weighting based on biological processes)

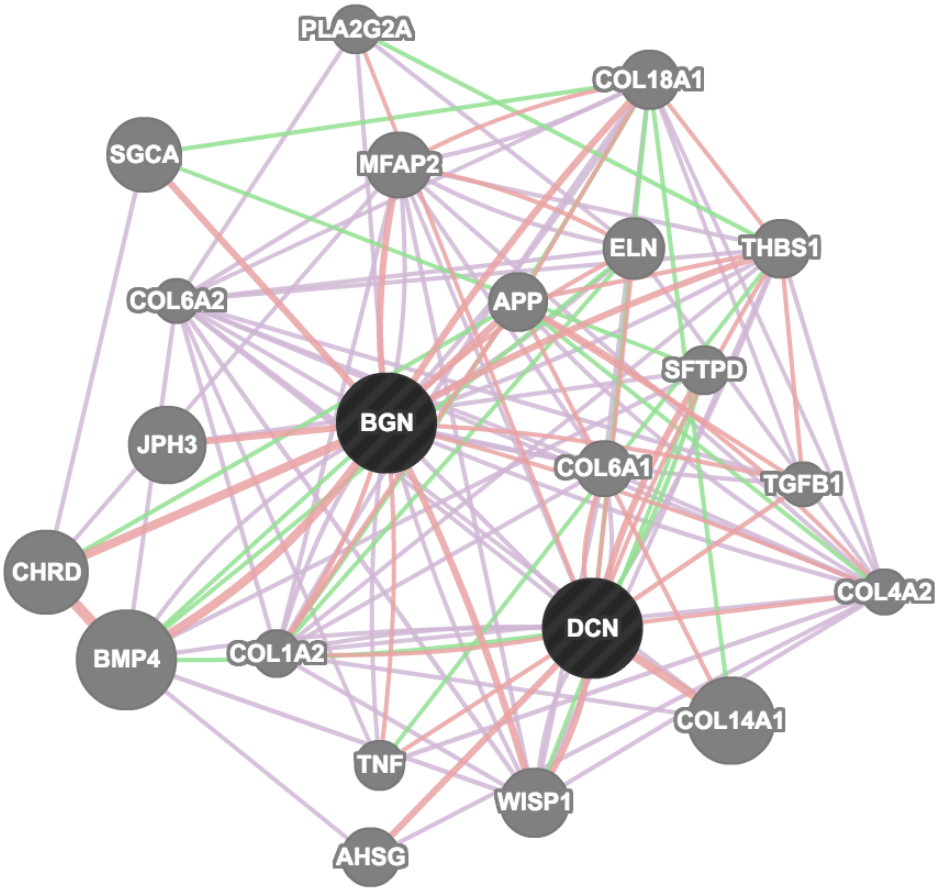
This was the output: TUG1 NDUFS5 PLEKHB1 BBIP1 DCN RPL35A RAB40B VPS13B JAG1 KPNA6 MPHOSPH9 NKTR ZNF518A ZFYVE26 CDH11 PTPN13 PPP1R7 STMN1 SERBP1 POGZ PFDN1 SCN1B CST3 BPTF CREG1 RPL37 PLOD2 OTUB1 BGN CSRP1 NUTF2 ANXA1 TCF4 MXI1 GBAS KCTD12 PPP2CB TUBB4B PRPF3 ACTN1 SPARC ZFP36 SF3B1 PPP2CA TARDBP SF3B14 COL3A1 NELFE MAZ PPP2R1A UBC MAX OTUD6B ATXN2 SH2D3A RAN AGO4 PPP1CA MACF1 NOTCH1 CALD1 RBM17 UHMK1 PLRG1 ACTN4 SCN2B DDX17 ANXA4 KPNB1 PFDN5 TMEM189-UBE2V1 C4A SMNDC1 TLX1 BMP4 MFAP1 TUBB COL14A1 THOC2 PCBP1 SET SF3A2 TUBA1B GNGT1 PPP2R1B S100A11 RNPS1 PAN2 PRKAR2A NEURL LRRC7 CTNNB1 CNOT1 FOS EIF4A3

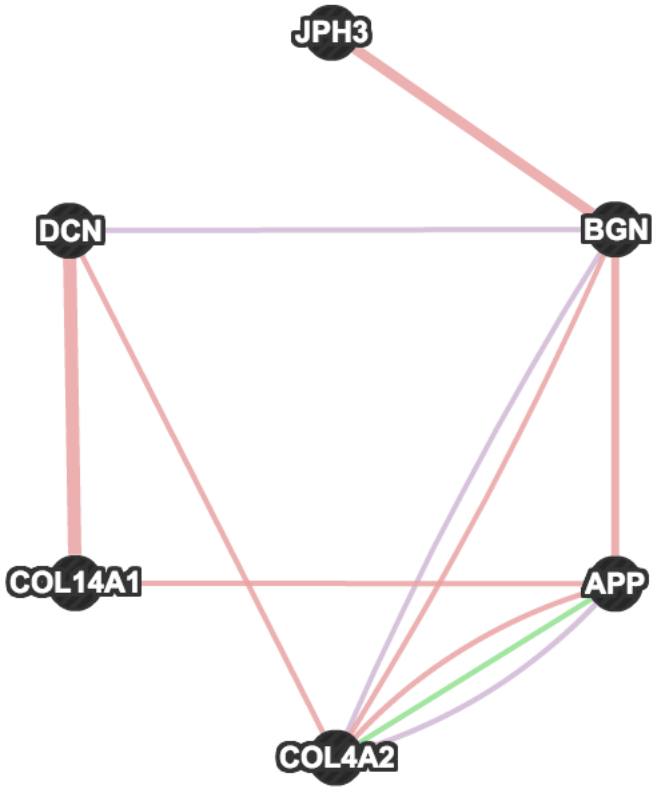
From this I identified 4 more SNPS

COL14A1 (ALS) COL3A1 (AD and Lewy body dementia) CTNNB1 (ALS) LRCC7 (ALS)

I tried using GWAS central myself but it's horrible to navigate. John says that they don't allow large amounts of information to be downloaded, and when he tried to contact them about AD SNPs they basically said no because of collaboration/authorship problems.

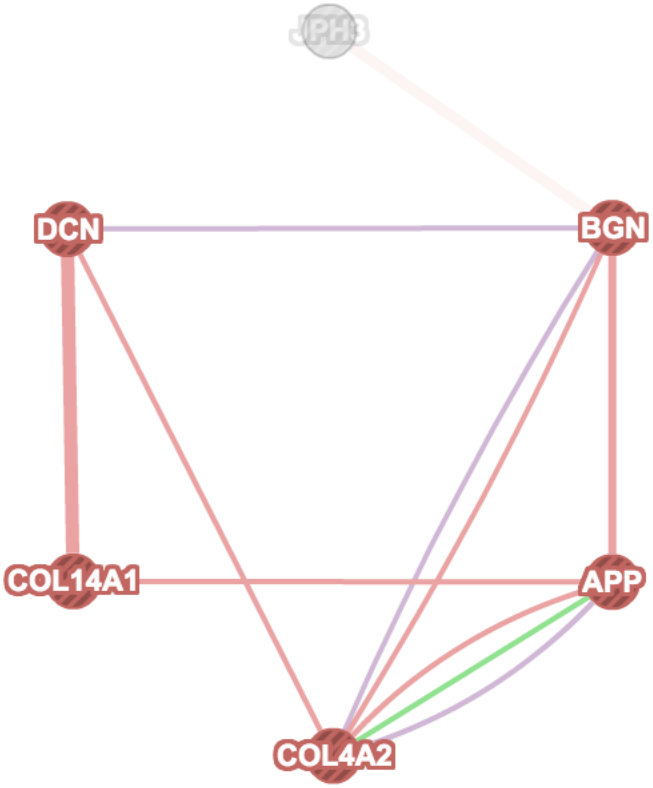
I did a little investigation of the DCN/BGN story as I had been emphasising it in my presentation. I decided to input just DCN and BGN in genemania and ask for the 20 most related genes. This contained 4 GWAScentral SNPs, including APP.


PLA2G2A
COL18A1
SGCA
MFAP2
ELN
THBS1
COL6A2
APP
SFTPD
BGN
JPH3
COL6A1
TGFB1
CHRD
COL4A2
DCN
BMP4
COL1A2
COL14A1
TNF
WISP1
AHSG



Just SNP genes

As it turned out, of these 6 genes, 5 were enriched for extracellular matrix organisation, but not dermatan/chondroitin metabolism.



Thursday

Sandeep mentioned a function of pathprint called diffPathways, which appears to more succintly conduct the analysis that John wrote in TDP-43_Signature.R. What you do is take the fingerprint output, define which columns are sample and disease, and it calculated DE for you based on a threshold you define.

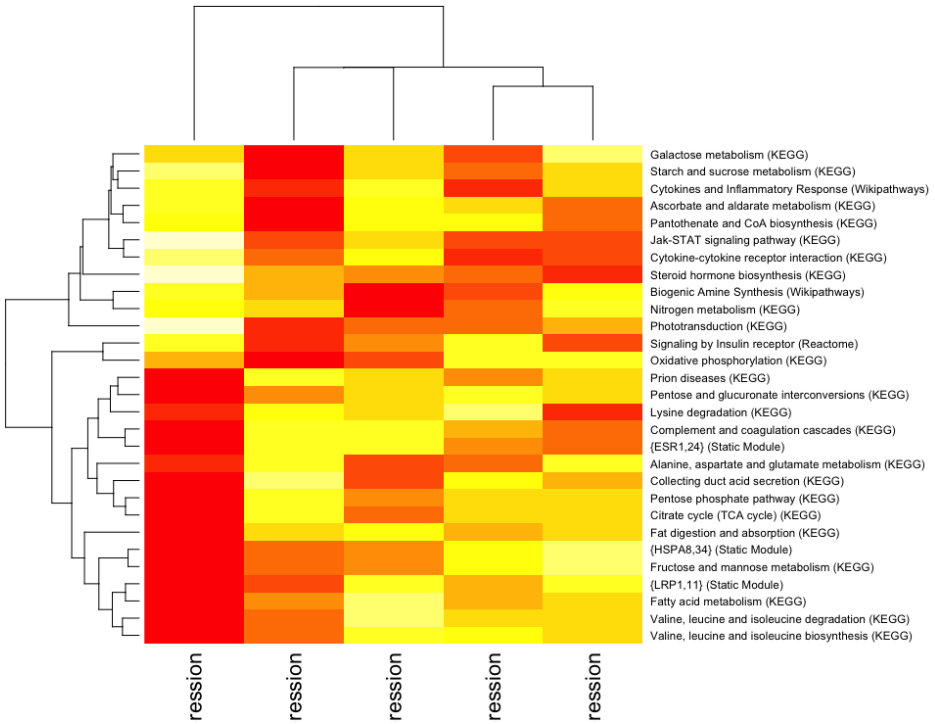
```
#Using diffPathways
C9fac <- c(1,1,1,1,1,1,1,1,0,0,0) #create vector assigning columns to disease or control
C9DP <- diffPathways(C9.LCM_pathprint, C9fac, 0.1)

CHfac <- c(1,1,1,0,0,0,0,0,0)
CHDP <- diffPathways(CHMP2B.LCM_pathprint, CHfac, 0.1)

sALSfac <- c(0,0,0,1,1,1,1,1,1,1)
sALSDP <- diffPathways(SALS.LCM_pathprint, sALSfac, 0.1)

FTLDfac <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0)
FTLDDP <- diffPathways(FTLD_pathprint, FTLDfac, 0.1)
```

```
VCPfac <- c(0,0,0,1,1,1,1,1,1,1)
VCPDP <- diffPathways(VCP_pathprint, VCPfac, 0.1)


#Intersect
overlap <- Reduce(intersect, list(C9DP, CHDP, sALSDP, FTLDDP, VCPDP)) #selects pathways that are presen
print(overlap)

setwd ("/Users/clairegreen/Desktop/")

write.csv(overlap, file = "overlap.csv")

#Heatmap

overlap <- as.data.frame(overlap) #convert to data frame

#After running pathprint
C9t1 <- as.data.frame(C9t1)
CHt1 <- as.data.frame(CHt1)
sALSt1 <- as.data.frame(sALSt1)
FTLDt1 <- as.data.frame(FTLDt1)
VCPt1 <- as.data.frame(VCPt1)

colnames(C9t1) <- "C9Expression"
colnames(CHt1) <- "CHExpression"
colnames(sALSt1) <- "sALSExpression"
colnames(FTLDt1) <- "FTLDExpression"
colnames(VCPt1) <- "VCPExpression"

C9t1[,2] <- rownames(C9t1)
CHt1[,2] <- rownames(CHt1)
sALSt1[,2] <- rownames(sALSt1)
FTLDt1[,2] <- rownames(FTLDt1)
VCPt1[,2] <- rownames(VCPt1)

all.fingerprints <- merge(x = C9t1, y = CHt1, by.x= "V2", by.y="V2")
all.fingerprints <- merge(x = all.fingerprints, y = sALSt1, by.x= "V2", by.y="V2")
all.fingerprints <- merge(x = all.fingerprints, y = FTLDt1, by.x= "V2", by.y="V2")
all.fingerprints <- merge(x = all.fingerprints, y = VCPt1, by.x= "V2", by.y="V2")

#sig.names <- read.delim(file = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/Pathprint/
sig.fingerprints <- merge(overlap, all.fingerprints, by.x = "overlap", by.y = "V2" ) #take only pathway

rownames(sig.fingerprints) <- sig.fingerprints[,1]
sig.fingerprints[,1] <- NULL

sig.fingerprints <- as.matrix(sig.fingerprints) #must be converted to numeric data frame for heatmap to

heatmap(sig.fingerprints)
```

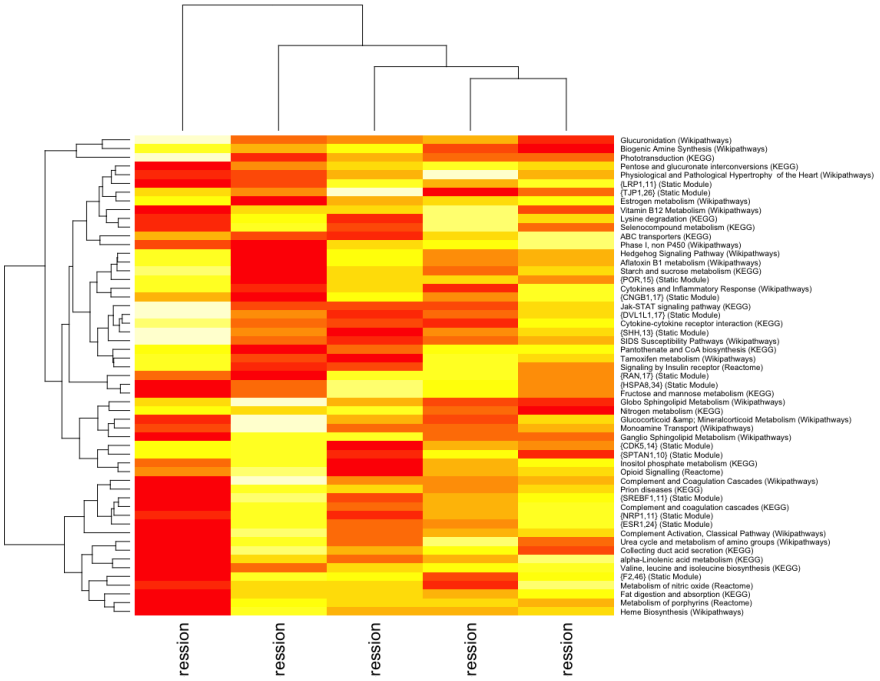
The output of this is mostly similar to what I already have. The two lists are:

Taking Mean	Using diffPathways
Fructose and mannose metabolism (KEGG)	Pentose and glucuronate interconversions (KEGG)
Ascorbate and aldarate metabolism (KEGG)	Fructose and mannose metabolism (KEGG)
Starch and sucrose metabolism (KEGG)	Lysine degradation (KEGG)
Complement and coagulation cascades (KEGG)	Starch and sucrose metabolism (KEGG)
{ESR1,24} (Static Module)	Pantothenate and CoA biosynthesis (KEGG)
{HSPA8,34} (Static Module)	Nitrogen metabolism (KEGG)
Biogenic Amine Synthesis (Wikipathways)	ABC transporters (KEGG)
Signaling by Insulin receptor (Reactome)	Complement and coagulation cascades (KEGG)
Phototransduction (KEGG)	Jak-STAT signaling pathway (KEGG)
Cytokines and Inflammatory Response (Wikipathways)	Phototransduction (KEGG)
Jak-STAT signaling pathway (KEGG)	Prion diseases (KEGG)
Pentose and glucuronate interconversions (KEGG)	Phase I, non P450 (Wikipathways)
Nitrogen metabolism (KEGG)	Ganglio Sphingolipid Metabolism (Wikipathways)
Oxidative phosphorylation (KEGG)	Urea cycle and metabolism of amino groups (Wikipathways)
Alanine, aspartate and glutamate metabolism (KEGG)	Complement Activation, Classical Pathway (Wikipathways)
Prion diseases (KEGG)	Biogenic Amine Synthesis (Wikipathways)
Collecting duct acid secretion (KEGG)	Complement and Coagulation Cascades (Wikipathways)
Pantothenate and CoA biosynthesis (KEGG)	Glucuronidation (Wikipathways)
Valine, leucine and isoleucine biosynthesis (KEGG)	SIDS Susceptibility Pathways (Wikipathways)
Valine, leucine and isoleucine degradation (KEGG)	Signaling by Insulin receptor (Reactome)
Lysine degradation (KEGG)	Opioid Signalling (Reactome)
Steroid hormone biosynthesis (KEGG)	{ESR1,24} (Static Module)
Fat digestion and absorption (KEGG)	{F2,46} (Static Module)
{LRP1,11} (Static Module)	{HSPA8,34} (Static Module)
Citrate cycle (TCA cycle) (KEGG)	{NRP1,11} (Static Module)
Pentose phosphate pathway (KEGG)	{POR,15} (Static Module)
Galactose metabolism (KEGG)	{RAN,17} (Static Module)
Cytokine-cytokine receptor interaction (KEGG)	{SPTAN1,10} (Static Module)
Fatty acid metabolism (KEGG)	{SREBF1,11} (Static Module)

About 50% of the pathways are shared. The heatmaps however look no more consistent in one than the other



Taking Mean



Using diffPathways

So in that respect I don't think I can bank on getting dysregulation occurring in the same direction. What that means for the whole model, I don't know, but I can't really see any way of getting around it. Maybe it's a data problem - too low signals, samples skewing results, I don't know. It's not like leaving one out is going to help as they all seem to contradict oneanother.

Sandeep showed me a function called getEntropy, using the package 'entropy'

```
getEntropy <- function(mat, index){
  if (index > 2 | index < 1)
```

```
    stop("Indicate 1 for rows or 2 for columns")
  d <- apply(as.matrix(mat), index, function(x){discretize(x, numBins = 3, r=c(-1,1))})
  entropy.vec <- apply(d, 2, entropy)
  return(entropy.vec)
}
```

What it does is remove any pathways where the variance between samples is thresholded to remove pathways which aren't that differnent.

```
library(entropy)

#Run TDP43_signature first

x=getEntropy(C9.LCM_pathprint,1)
x.selectC9=C9.LCM_pathprint[names(x)[which(x>0.5)],]
heatmap(x.selectC9)

x=getEntropy(CHMP2B.LCM_pathprint,1)
x.selectCH=CHMP2B.LCM_pathprint[names(x)[which(x>0.5)],]
heatmap(x.selectCH)

x=getEntropy(SALS.LCM_pathprint,1)
x.selectsals=SALS.LCM_pathprint[names(x)[which(x>0.5)],]
heatmap(x.selectsals)

x=getEntropy(FTLD_pathprint,1)
x.selectFTLD=FTLD_pathprint[names(x)[which(x>0.5)],]
heatmap(x.selectFTLD)

x=getEntropy(VCP_pathprint,1)
x.selectVCP=VCP_pathprint[names(x)[which(x>0.5)],]
heatmap(x.selectVCP)

#Using diffPathways

DEthresh <- 0

C9fac <- c(1,1,1,1,1,1,1,1,0,0,0) #create vector assigning columns to disease or control
C9DP <- diffPathways(x.selectC9, C9fac, DEthresh)

CHfac <- c(1,1,1,0,0,0,0,0,0)
CHDP <- diffPathways(x.selectCH, CHfac, DEthresh)

sALSfac <- c(0,0,0,1,1,1,1,1,1,1)
sALSDP <- diffPathways(x.selectsals, sALSfac, DEthresh)

FTLDfac <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0)
FTLDDP <- diffPathways(x.selectFTLD, FTLDfac, DEthresh)

VCPfac <- c(0,0,0,1,1,1,1,1,1,1)
VCPDP <- diffPathways(x.selectVCP, VCPfac, DEthresh)
```

```
#Intersect
overlap <- Reduce(intersect, list(C9DP, CHDP, sALSDP, FTLDDP, VCPDP)) #selects pathways that are presen
print(overlap)
```

However, the results I get out are similar but not as tightly linked by function and coexpression

[1] “Pentose and glucuronate interconversions (KEGG)”
[2] “Phototransduction (KEGG)”
[3] “Collecting duct acid secretion (KEGG)”
[4] “Fat digestion and absorption (KEGG)”
[5] “Globo Sphingolipid Metabolism (Wikipathways)”
[6] “Physiological and Pathological Hypertrophy of the Heart (Wikipathways)”
[7] “Vitamin B12 Metabolism (Wikipathways)”
[8] “Complement Activation, Classical Pathway (Wikipathways)”
[9] “Biogenic Amine Synthesis (Wikipathways)”
[10] “Aflatoxin B1 metabolism (Wikipathways)”
[11] “Signaling by Insulin receptor (Reactome)”
[12] “{CNGB1,17} (Static Module)”
[13] “{F2,46} (Static Module)”
[14] “{LRP1,11} (Static Module)”

I talked to John and he said that he has tried ranking by variance before but the problem is that sometimes the most biologically relevent pathways are not the most varied. This method might be a little too stringent meaning that the delicate balance is lost. I don’t know, it’s difficult to tell.

We also discussed the methodology of using the averaging method rather than the diffPathways method. He said maybe try a different summary statistic, so first I tried just summing the columns and looking at the difference, however the results were more than weird:

[1] “Proximal tubule bicarbonate reclamation (KEGG)” “Pyrimidine metabolism (KEGG)”
[3] “Taurine and hypotaurine metabolism (KEGG)” “N-Glycan biosynthesis (KEGG)”
[5] “Other glycan degradation (KEGG)” “Linoleic acid metabolism (KEGG)”
[7] “Ribosome biogenesis in eukaryotes (KEGG)” “Ribosome (KEGG)”
[9] “RNA degradation (KEGG)” “Spliceosome (KEGG)”
[11] “Base excision repair (KEGG)” “Non-homologous end-joining (KEGG)”
[13] “Protein processing in endoplasmic reticulum (KEGG)” “Glutamatergic synapse (KEGG)”
[15] “Shigellosis (KEGG)” “Pentose Phosphate Pathway (Wikipathways)”
[17] “Proteasome Degradation (Wikipathways)” “IL-1 Signaling Pathway (Wikipathways)”
[19] “SREBP signalling (Wikipathways)” “IL-7 Signaling Pathway (Wikipathways)”
[21] “IL-9 Signaling Pathway (Wikipathways)” “TNF-alpha/NF-kB Signaling Pathway (Wikipathways)”
[23] “EBV LMP1 signaling (Wikipathways)” “FAS pathway and Stress induction of HSP regulation (Wikipathways)” [25] “Apoptosis Modulation by HSP70 (Wikipathways)” “Nucleotide Metabolism (Wikipathways)”
[27] “Diurnally regulated genes with circadian orthologs (Wikipathways)” “mRNA processing (Wikipathways)”
[29] “Non-homologous end joining (Wikipathways)” “Cytoplasmic Ribosomal Proteins (Wikipathways)”
[31] “EPO Receptor Signaling (Wikipathways)” “Cell Cycle Checkpoints (Reactome)”
[33] “Metabolism of proteins (Reactome)” “DNA Replication (Reactome)”

This list is completely different to my original list, but it does have some interesting points. For example TDP-43 is involved in mRNA processing and TNF-alpha/nfkb pathway is hugely implicated in ALS. Glutamate excitotoxicity is one of the biggest contributions to neuronal death. However, I think I trust the diffPathways function more than this, so I’m not entirely sure it’s the right direction.

Friday

I took the pathway list from diffPathways and collected all the gene lists. There are 29 pathways and 1042 genes in total. I find 5 of my genes in that list, however I don't know if that is significant.

I presented my oral presentation at lab meeting, and there were only a few corrections. I have been asked to do some work for PCxN - when you develop a method you need to show that it produces biologically relevent results, and hopefully something no one has discovered before. Win asked if I could do an example where I took some ALS data sets, conducted GSEA analysis, put those enriched functions in PCxN and hopefully find enrichment of GWAS SNPs in the genes of those modules.