LabBook_17_02_2017

Claire Green

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

Monday

After collecting all the information about the data in SITraN that may be relevent to our analyses, I decided to start looking at the count matrix that Sandeep sent me of Guillaume's data. There were 8 conditions, each with 3 samples

- GRASPS Translatome, Control
- GRASPS Translatome, Q133K
- GRASPS Translatome, GFP low
- GRASPS Translatome, GFP high
- Whole cell transcriptome, Control
- Whole cell transcriptome, Q133K
- Cytoplasmic transcriptome, Control
- Cytoplasmic transcriptome, Q133K

Using my limma script (DiffExpr_RNAseq.R) I analysed 8 comparisons:

- GRASPS TRL Q133K vs GRASPS TRL Control
- WCT Q133K vs WCT Control
- CYT Q133K vs WCT Control
- GRASPS TRL vs GRASPS GFP low
- GRASPS TRL vs GRASPS GFP high
- GRASPS TRL vs WCT
- GRASPS TRL vs CYT
- WCT vs CYT

GH-TDP-43_RNAEXPR.R

```
##RNA-Seq Gene Expression Analysis using Limma##

library(limma)
library(edgeR)
library(biomaRt)
library(plyr)

setwd(dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/Guillaume_HEK48
# Counts <- read.table(file = 'GSE67196_Petrucelli2015_ALS_genes.rawcount.txt', header = TRUE)
#
# write.csv(x = Counts, file = "counts_petrucelli.csv")

Counts <- read.csv(file = "combined.counts.csv", header = TRUE)

Counts[Counts == 0] <- NA
# Counts[Counts<30] <- NA
Counts <- na.omit(Counts)
rownames(Counts)<-Counts[,1]
Counts[,1] <- NULL

## Divide into different analyses
```

```
TRL <- Counts[,1:6]
WCT <- Counts[,13:18]
CYT <- Counts[,19:24]
GFP_L <- Counts[,c("GRASPS.TRL.Control.1","GRASPS.TRL.Control.2","GRASPS.TRL.Control.3","GRASPS.TRL.GFPL
                   "GRASPS.TRL.GFPLow.2","GRASPS.TRL.GFPLow.3")]
GFP_H <- Counts[,c("GRASPS.TRL.Control.1","GRASPS.TRL.Control.2","GRASPS.TRL.Control.3","GRASPS.TRL.GFPH
                   "GRASPS.TRL.GFPHigh.2","GRASPS.TRL.GFPHigh.3")]
GRASPS_VS_WCT <- Counts[,c("GRASPS.TRL.Control.1","GRASPS.TRL.Control.2","GRASPS.TRL.Control.3","WCT.Co
                           "WCT.Control.2","WCT.Control.3")]
GRASPS_VS_CYT <- Counts[,c("GRASPS.TRL.Control.1","GRASPS.TRL.Control.2","GRASPS.TRL.Control.3","CytTrc
                           "CytTrc.Control.2","CytTrc.Control.3")]
WCT_VS_CYT <- Counts[,c("WCT.Control.1","WCT.Control.2","WCT.Control.3","CytTrc.Control.1",
                           "CytTrc.Control.2","CytTrc.Control.3")]


#####
analysis.name<-"GH_HEK_48h_TRL" #Label analysis

Countnum <- TRL

#DGElist
dge <- DGEList(counts=Countnum)
dge <- calcNormFactors(dge)

#Design
Treat<-factor(rep(c("Control", "Patient"),c(3,3)), levels=c("Control", "Patient"))
design<-model.matrix(~Treat)
rownames(design)<-colnames(Countnum)
design

#Voom transformation
v <- voom(dge,design,plot=FALSE)

#Limma fitting
fit <- lmFit(v,design)
fit <- eBayes(fit)
result<-topTable(fit, coef="TreatPatient", adjust="BH", number=nrow(Countnum)) #"BH" adjust for multipl
result <- merge(result, Countnum, by="row.names", all=TRUE)
#result <- result[,1:7]

#Count tables from bcbio have ensembl gene IDs. This must be annotated with HGNC symbols

#Download the HGNC symbols and gene IDs using a vector containing the IDs from results

genes <- as.vector(result[,1])
mart <- useMart("ENSEMBL_MART_ENSEMBL",dataset="hsapiens_gene_ensembl", host="www.ensembl.org")
mart_back <- getBM(attributes =c("ensembl_gene_id", "hgnc_symbol"), filters="ensembl_gene_id", values=g

# library(org.Hs.eg.db)
# library(GeneNetworkBuilder)

#Merge the tables using ensembl ID
result <- merge(result, mart_back, by.x = "Row.names", by.y = "ensembl_gene_id")
```

```
# result[,1] <- NULL




#### Take median value for gene duplicates ############
result3 <- ddply(result,"hgnc_symbol", numcolwise(median, (result$adj.P.Val)))
#result3 <- aggregate(result, by=list("Gene.Symbol"), FUN=median)

genesort <- result3[order(result3$adj.P.Val),]

setwd(dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/Guillaume_HEK48
write.csv(genesort, file=paste(analysis.name, "_HGNCrankeduniqueresult.csv", sep=""), row.names=FALSE, 









#
# uniqueresult <- result[!duplicated(result$hgnc_symbol),]
# rownames(uniqueresult) <- uniqueresult$hgnc_symbol
# genesort <- uniqueresult[order(uniqueresult$adj.P.Val),]
#
# setwd(dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/DEG_Test2/")
# write.csv(genesort, file=paste(analysis.name, "EXPRSrankeduniqueresult.csv", sep=""), sep="\t", row.n

# topgene <- genesort[1:1000,]
# write.csv(x = topgene, file = paste(analysis.name,"_ap_1000.csv", sep = ""))
# topgene <- genesort[1:2000,]
# write.csv(x = topgene, file = paste(analysis.name,"_ap_2000.csv", sep = ""))
# topgene <- genesort[1:3000,]
# write.csv(x = topgene, file = paste(analysis.name,"_ap_3000.csv", sep = ""))
# topgene <- genesort[1:4000,]
# write.csv(x = topgene, file = paste(analysis.name,"_ap_4000.csv", sep = ""))
# topgene <- genesort[1:5000,]
# write.csv(x = topgene, file = paste(analysis.name,"_ap_5000.csv", sep = ""))
```

And then I did any enrichment analysis using hyperPathway_GH_TDP-43.R

```
library(pathprint)
library(hgu133plus2.db)

setwd(dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Data/GSEA/PCxN Example/probesets/")

### Gene background ###

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),]

### DEG Thresholds ###
setwd (dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/DEG_Test2/")
Y <- read.csv(file = "AllgenesNO.csv", na.strings = c("", "NA)"))
Y <- as.list(Y)
Y <- lapply(Y, function(x) x[!is.na(x)])

setwd (dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/")
W <- read.csv(file = "BenchmarkGenes.csv", na.strings = c("", "NA)"))
W <- as.list(W)
W<- lapply(W, function(x) x[!is.na(x)])

setwd (dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/Pathprint/Pathprint 25.04.16/
Z <- read.csv(file = "pathprintgenes.csv", na.strings = c("", "NA)"))
Z <- as.list(Z)
Z <- lapply(Z, function(x) x[!is.na(x)])

setwd (dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/Guillaume_HEK4
TRL_all <- read.csv(file = "TRL_Comparison.csv", na.strings = c("", "NA)"))
TRL_all <- as.list(Z)
 <- lapply(Z, function(x) x[!is.na(x)])



### GH Genes ###

setwd (dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/Guillaume_HEK4

X <- read.table(file = "GH_TRL_1000.txt", header = FALSE)
X <- X$V1

### Enrichment analysis ###
pathwayEnrichment <- hyperPathway(
  genelist = LC_CYT_gene,
  geneset = W,
  Nchip = length(allgenes)
)
setwd (dir = "/Users/clairegreen/Desktop/")
write.csv(pathwayEnrichment, file = "VCPpathways_enrich.csv")




setwd (dir = "/Users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/Guillaume_HEK4
```

```r
SA_TRL <- read.table("tdp43_top_TRL_sigDEG.genes.txt")
SA_CYT <- read.table("tdp43_top_CYT_sigDEG.genes.txt")
SA_WCT <- read.table("tdp43_top_WCT_sigDEG.genes.txt")

SA_TRL_gene <- SA_TRL$GeneSymbol
SA_CYT_gene <- SA_CYT$GeneSymbol
SA_WCT_gene <- SA_WCT$GeneSymbol


CG_TRL <- read.csv("GH_HEK_48hHGNC_TRL_rankeduniqueresult.csv")
CG_TRL_sig <- subset(CG_TRL, subset=(P.Value < 0.05))
CG_CYT <- read.csv("GH_HEK_48h_CYT_HGNCrankeduniqueresult.csv")
CG_CYT_sig <- subset(CG_CYT, subset=(adj.P.Val < 0.05))
CG_WCT <- read.csv("GH_HEK_48h_WCT_HGNCrankeduniqueresult.csv")
CG_WCT_sig <- subset(CG_WCT, subset=(adj.P.Val < 0.05))

CG_TRL_gene <- CG_TRL_sig$hgnc_symbol
CG_CYT_gene <- CG_CYT_sig$hgnc_symbol
CG_WCT_gene <- CG_WCT_sig$hgnc_symbol

LC_TRL <- read.table("Q331K GRASPS_BitSeq_EdgeR_DE transcripts.txt")
LC_TRL_sig <- subset(LC_TRL, subset=(FDR < 0.05))
LC_CYT <- read.table("Q331K cytoplasmic transcriptomes_BitSeq_EdgeR_DE transcripts.txt")
LC_CYT_sig <- subset(LC_CYT, subset=(FDR < 0.05))
LC_WCT <- read.table("Q331K whole cell transcriptomes_BitSeq_EdgeR_DE transcripts.txt")
LC_WCT_sig <- subset(LC_WCT, subset=(FDR < 0.05))

LC_TRL_gene <- LC_TRL_sig$V3
LC_CYT_gene <- LC_CYT_sig$V3
LC_WCT_gene <- LC_WCT_sig$V3

# write.table(x = CG_TRL_gene, file = "CG_TRL_gene.txt", quote = FALSE, row.names = FALSE)
# write.table(x = SA_TRL_gene, file = "SA_TRL_gene.txt", quote = FALSE, row.names = FALSE)
# write.table(x = LC_TRL_gene, file = "LC_TRL_gene.txt", quote = FALSE, row.names = FALSE)
#
#
library(VennDiagram)
results <- read.csv("TRL_Comparison.csv", na.strings = c("", "NA)"))
results <- as.list(results)
# results <- lapply(results, function(x) x[!duplicated(x)])
results <- lapply(results, function(x) x[!is.na(x)])

venn <- calculate.overlap(
  x = list(
    "SA" = SA_TRL_gene,
    "LC" = LC_TRL_gene,
    "CG" = CG_TRL_gene))

venn <- calculate.overlap(x = results)

grid.newpage()
draw.triple.venn(area1 = 697, area2 = 2528, area3 = 2390, n12 = , n23 = , n13 = 35,
                 n123 = 0, category = c("Manifesting vs Control", "Non-manifesting vs Control", "Manifes
```

```
                fill = c("skyblue", "violet", "coral"), cat.dist = -0.1)

vennDiagram(results)

### Intersect ###
overlap <- Reduce(intersect, list(LC_TRL_gene, CG_TRL_gene))
print(overlap)
```

Tuesday

After analysis, the results were somewhat confusing. As Sandeep had done his own DEG analysis, as had Luisa, I decided to do a comparison.

	SA	CG	Overlap
TRL	696	2930 (rawp)	109
CYT	26	3116	18
WCT	9	1463	5

	LC	CG	Overlap
TRL	3400	2930 (rawp)	641
CYT	3116	3116	1654
WCT	2652	1463	195

	LC	SA	Overlap
TRL	3400	696	6
CYT	3116	26	20
WCT	2652	9	6

LC TRL Enrichment

ListName	ID	P-value	BHadjP-value	nGenes	nPathway
Exac	6	1.02E-17	1.12E-16	601	2680
Pasterkamp	10	8.41E-08	4.63E-07	44	124
Taylor	11	2.52E-07	9.25E-07	75	261
Carulli	7	0.010507983	0.028896954	11	36
NeuroX.GWS	3	0.088503998	0.194708796	12	53
GeneCards.AD	8	0.169585307	0.310906396	36	191
NeuroX.FDR..05	2	0.332412926	0.406282465	20	114
GWAS.AD	4	0.304388895	0.406282465	11	61
GeneCards.ALS	9	0.285859326	0.406282465	27	151
Subnetwork.28	5	0.841251305	0.925376436	9	77
GWAS.ALS	1	1	1	154	1817

SA TRL Enrichment

ListName	ID	P-value	BHadjP-value	nGenes	nPathway
Taylor	11	1.04E-26	1.14E-25	59	261
Pasterkamp	10	3.16E-18	1.74E-17	33	124
Exac	6	7.91E-08	2.90E-07	172	2680
Carulli	7	2.30E-06	6.34E-06	9	36
GeneCards.ALS	9	0.002733299	0.006013257	14	151
NeuroX.GWS	3	0.08260092	0.15143502	4	53
NeuroX.FDR..05	2	0.129434867	0.203397648	7	114
GeneCards.AD	8	0.22083241	0.303644564	10	191
GWAS.AD	4	0.755512359	0.923403994	1	61
Subnetwork.28	5	0.858159454	0.9439754	1	77
GWAS.ALS	1	1	1	29	1817

CG TRL Enrichment

ListName	ID	P-value	BHadjP-value	nGenes	nPathway
Exac	6	5.45E-10	5.99E-09	409	2680
NeuroX.FDR..05	2	0.067854807	0.373201436	18	114
NeuroX.GWS	3	0.157674379	0.578139391	8	53
Carulli	7	0.236457772	0.650258874	5	36
GeneCards.ALS	9	0.303481905	0.667660192	19	151
Subnetwork.28	5	0.407085274	0.746323003	9	77
GeneCards.AD	8	0.554741243	0.871736238	21	191
GWAS.AD	4	0.728393208	0.890258365	5	61
Pasterkamp	10	0.697408505	0.890258365	12	124
GWAS.ALS	1	0.890478313	0.912857831	195	1817
Taylor	11	0.912857831	0.912857831	23	261