Lee TDP-43 Analysis

Claire Green

17/05/2017

Rationale

The purpose of this analysis was twofold: firstly to provide some validation of the findings of the Lee lab, and secondly to see if there is any concordance with the results I have obtained from my TDP-43 consensus analysis. After running the DESeq protocol, the Lee lab will be able to make their own comparisons with their results, and I will make my own comparison with my results.

Data

A pre-processed count matrix was provided with the expression values for 14 paired samples. These samples were procured from the frontal cortex of ALS patients with C9orf72 mutations. Two samples were collected from each patient, one containing the nuclei which stained positively for presence of TDP-43 protein, and one containing nuclei that stained negatively. Further information about the patients and the nuclei extraction protocol can be found on Basecamp.

The count matrix was analysed using the R package DESeq2. Below is the script:

```
#### Analysis of EL's TDP-43 data ####

options(scipen=999)

library(DESeq2)
library(biomaRt)

#Generate a column of Ensembl IDs and a column of gene names
setwd("/users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/EL_TDP-43/")
load("Count_Matrix_Genes.rda")

EnsNames <- rownames(counts.g)
EnsNames <- EnsNames[grepl("^ENS", EnsNames)]

counts.g.ens <- subset(counts.g, rownames(counts.g) %in% EnsNames)
counts.g.hgnc <- subset(counts.g, !(rownames(counts.g) %in% EnsNames))


ens_genes <- row.names(counts.g.ens)
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=en


hgnc_genes <- row.names(counts.g.hgnc)
mart <- useMart("ENSEMBL_MART_ENSEMBL",dataset="hsapiens_gene_ensembl", host="www.ensembl.org")
mart_back_hgnc <- getBM(attributes =c("ensembl_gene_id", "hgnc_symbol"), filters="hgnc_symbol", values=h

#Rearrange columns to match and merge
```

```
counts_ens <- merge(counts.g, mart_back, by.x = 0, by.y = "ensembl_gene_id")
counts_hgnc <- merge(counts.g, mart_back_hgnc, by.x = 0, by.y = "hgnc_symbol")


counts_ens <- counts_ens[,c(16,1,3,5,7,9,11,13,15,2,4,6,8,10,12,14)]
counts_hgnc <- counts_hgnc[,c(1,16,3,5,7,9,11,13,15,2,4,6,8,10,12,14)]

colnames(counts_ens)[1] <- "hgnc_symbol"
colnames(counts_hgnc)[1] <- "hgnc_symbol"
colnames(counts_ens)[2] <- "Ensembl_ID"
colnames(counts_hgnc)[2] <- "Ensembl_ID"

counts_all <- rbind(counts_ens, counts_hgnc)
counts_all<- subset(counts_all, subset=(hgnc_symbol !="")) #if no gene symbol, discount


exp_info <- data.frame(condition = factor(c(rep("1", 7), rep("2", 7))),
                       patientID = factor(c(1,2,3,4,5,6,7,1,2,3,4,5,6,7)),
                       Sex = factor(c("F","M","F","F","M","M","M")),
                       Disease = factor(c("FTLD","FTLD-ALS","FTLD","FTLD","FTLD","FTLD-ALS","FTLD-ALS"))


#Rows must have at least 3 samples with scores of 10 or higher
keep <- rowSums(counts_all>=10) >= 3
counts_all<-counts_all[keep,]

counts_all_data <- counts_all
rownames(counts_all_data) <- counts_all$Ensembl_ID
counts_all_data <- counts_all_data[,3:16]

#Create a coldata frame
coldata <- data.frame(row.names=colnames(counts_all_data), exp_info)
coldata

#Create a DESeqDataSet object
dds <- DESeqDataSetFromMatrix(countData=counts_all_data, colData=coldata, design=~patientID + conditions
dds

#Run DEseq2 pipleline
dds <- DESeq(dds)

res <- results(dds)
table(res$padj<0.05)
## Order by p-value
res <- res[order(res$pvalue), ]

# ## Merge with raw count data
# resdata_raw <- merge(as.data.frame(res), as.data.frame(counts(dds, normalized=FALSE)), by="row.names"

## Merge with normalised count data
resdata_norm <- merge(as.data.frame(res), as.data.frame(counts(dds, normalized=TRUE)), by="row.names", s
# resdata_norm_data <- resdata_norm[,8:21]
genenames <- counts_all[,1:2]
```

```
result <- merge(resdata_norm, genenames, by.x = "Row.names", by.y = "Ensembl_ID")
result <- result[,c(1,22,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21)]
colnames(result)[1] <- "Ensembl_ID"

genesort <- result[order(result$pvalue),]
genesort <- genesort[!duplicated(genesort$hgnc_symbol),]

setwd("/users/clairegreen/Documents/PhD/TDP-43/TDP-43_Code/Results/GeneExpression/EL_TDP-43/2017_05_31/
write.csv(genesort, "EL_results_31052017", row.names = F)

Sig.padj <- subset(genesort, subset=(padj < 0.05))
Sig.padj.gene <- Sig.padj$hgnc_symbol
Sig.padj.gene <- Sig.padj.gene[!duplicated(Sig.padj.gene)]

Sig.padj.up <- subset(Sig.padj, subset=(log2FoldChange > 0))
Sig.padj.up.gene <- Sig.padj.up$hgnc_symbol
Sig.padj.up.gene <- Sig.padj.up.gene[!duplicated(Sig.padj.up.gene)]

Sig.padj.down <- subset(Sig.padj, subset=(log2FoldChange < 0))
Sig.padj.down.gene <- Sig.padj.down$hgnc_symbol
Sig.padj.down.gene <- Sig.padj.down.gene[!duplicated(Sig.padj.down.gene)]

Sig.padj.both <- c(Sig.padj.down.gene, Sig.padj.up.gene)

write.table(Sig.padj.both, "sig_padj_genenames.txt", quote = FALSE, row.names = FALSE, col.names = FALSE
write.table(Sig.padj.up.gene, "sig_padj_up_genenames.txt", quote = FALSE, row.names = FALSE, col.names =
write.table(Sig.padj.down.gene, "sig_padj_down_genenames.txt", quote = FALSE, row.names = FALSE, col.nam
```

Results

Through this analysis I identified 4803 significantly differentially expressed genes at an adjusted p value of 0.05. This list of genes can be found in the file "sig_padj_genenames.txt" and have also been split into two separate files based on upregulation and downregulation.

Functional enrichment

I conducted functional enrichment using the webtool EnrichR. This provides a static link for a variety of databases in which enrichment has been calculated.

Upregulated genes http://amp.pharm.mssm.edu/Enrichr/enrich?dataset=1hnmt

Downregualted genes http://amp.pharm.mssm.edu/Enrichr/enrich?dataset=1hnn0

As you can see for upregulated genes there seems to be a trend towards synaptic transmission and mRNA processing. This can easily be related back to TDP-43 disease as we know there is a huge dysregulation in synaptic function, particularly glutamate toxicity from overexcitation in ALS. Additionally, as TDP-43 is involved in a large number of mRNA-related processes it makes sense to see a dysregulation in this area.

Interestingly, you also see *downregulation* in the glutamatergic synapse. These results together perhaps represent a battle between dysregulatory and compensatory mechanisms that are trying to bring synaptic activity back to homeostasis. There is also enrichment of downregulated genes in the GABAergic synapse, suggesting a downregulation of mechanisms that are normally inhibitory. Aside from synaptic transmission,

there are implications of downregulation of cytoskeletal genes, which paired with the suggestion of cancer and cell development may point towards mechanisms of the cell cycle and apoptosis.

Comparison with my results

There are quite a few comparisons that I can draw between these results and my own. Through intersecting differentially expressed genes from 6 TDP-43+ datasets I was able to extract 285 common genes (full list in file "filtered_upanddown.txt". Running a Fisher's Exact Test of enrichment of my genes in the 4803 DEGs from the Lee data, I found a significant enrichment p value of 3.4e-11. This translated to an overlap of 102/285 genes in my DEG list present in the Lee DEG list. These genes are:

RPH3A GNB5 WASF1 PPP3CA KCNAB1 NDEL1 CDH22 SEPT6 UCHL1 TMOD2 PDE4A LPCAT4 EEF1A1 ATF7IP FAM129A PLOD2 ADAM17 CTBP2 CALD1 LPL CCNL1 ZFYVE26 ZBTB10 BTG1 CNOT1 NEK4 PABPC1 VCAN AMD1 N4BP2L2 KCTD12 PNISR HIPK1 ZCCHC6 PNN TIMP2 RPS6KA2 SEC24A PALLD CPSF7 RAP1A EFEMP1 TRIT1 CMTM6 TARDBP ZFC3H1 MYOF CLK4 THUMPD2 ANKRD10 ABI1 PTPRA MAP4K3 RAP1B IFRD1 RBM5 ALMS1 YY1AP1 ANKRD28 JMJD1C CRLF3 ZNF83 PROSER1 ITGA6 SETD2 CREG1 ZCCHC8 SORBS1 NBPF1 NEK1 SOCS2 PLSCR1 U2SURP MAP4K5 EIF3E CLIC4 USP3 KIAA0141 CPOX ANXA1 WSB1 FHL1 SH2B3 HEXB TCF12 TPP1 OGT CHD1 PHKB NAA50 STX2 CBLB DENND4A SENP7 BCLAF1 POT1 LRCH3 ATP8B1 HNMT SRSF10 TRIM5 CPSF6

There are also functional similarities with results I have generated, both with my 285 DEGs and subsequently with the 3566-node PPI network I recently generated from the first-degree neighbours of my 285 DEGs. You can find the results from various enrichments here:

Enrichment of my 285 DEGs http://amp.pharm.mssm.edu/Enrichr/enrich?dataset=1hous

Enrichment of 102 overlap DEGs http://amp.pharm.mssm.edu/Enrichr/enrich?dataset=1hovu

Enrichment of my expanded PPI network (3566 genes) http://amp.pharm.mssm.edu/Enrichr/enrich?dataset=1hovv

Out of interest, I decided to look at the overlap of my expanded PPI network with the Lee lab 4803 DEGS. I found a significant enrichment of 3e-18 with an overlap of 904 genes. This gene list can be found in the file "CG_PPI_EL_DEG_overlap.txt"

Enrichment of this overlap can be found here: http://amp.pharm.mssm.edu/Enrichr/enrich?dataset=1hox1

Interpretation

Overall it appears that some consistent signals are coming up in both datasets. This includes cell signalling, neurotransmission, cell cycle and apoptosis, processing of genetic information (particularly RNA), the immune system etc. In the context of neurodegeneration these are certainly processes that we hope and expect to see. It's extremely positive to see a significant overlap between the genesets I have produced from each dataset, and hopefully we can both use that information to narrow down the search field for interesting targets.