Identifying Signature Pathways for TDP-43 Pathology

Claire Green

The purpose of this code is to use the tool Pathprint on TDP-43 pathology-associated RNA expression data from multiple sources to identify a signature set of pathways that are represenative of patients suffering ALS with TDP-43 pathology. Further investigation of this set of pathways could potentially reveal a mechanism behind the loss of nuclear TDP-43 and formation inclusions. This mechanism in turn could be a target for intervention, i.e. drug treatment.

Accumulating RNA Expression Data

The only criterion permitting inclusion of a potential data source is that TDP-43 pathology has been identified and reported. This means that we can be sure that the RNA expression data is representative of TDP-43 pathology. Certain variants of ALS are therefore automatically excluded, i.e. those with SOD1 or FUS mutations. Also, data is only acceptable if collected from motor neurons, astrocytes or muscle tissue, as more peripheral tissues do not present with TDP-43 pathology.

Currently, we have 5 data sets:
- LCM motor neurons from C9orf72 patients
- LCM motor neurons from CHMP2B patients
- LCM motor neurons from sporadic ALS patients
- Cortical neurons from FTLD patients
- Muscle tissue from VCP patients

The aim is to collect as many data sets as possible, to increase our confidence in the signature.

Pre-processing Array Data

Pre-processing array data involves taking the signal values for each set of gene probes and converting them into estimated expression levels for each gene. This is done using the Multi-chip modified gamma Model for Oligonucleotide Signal (mmgmos) function in the puma library. An example can be seen below:

First, you must read in the .CEL files into an AffyBatch using ReadAffy, which will also annotate the genes

```r
datadir <-("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43 Data Sets/CHMP2B")

CHMP2BFilenames <- c("LC 96-06.CEL","LC 299-99.CEL","LC 04-06.CEL",
                     "JK 85-07.CEL", "JK CC4.CEL","JK C 39-97.CEL","JK C 35-96.CEL",
                     "JK C 12-07.CEL","JK CC2.CEL", "JK C 05-07.CEL")

affybatch.CHMP2B <- ReadAffy( filenames=CHMP2BFilenames , celfile.path=datadir)
```

Next, phenotypic meta-data describing the experimental condition is assigned

```r
pData(affybatch.CHMP2B) <- data.frame("Subgroup"=c("CHMP2B_ALS","CHMP2B_ALS",
```

Finally, the mmgmos function is utilised to convert the signal values into expression values, and this is written into a .csv file which can be used later for Pathprint analysis

```
eset_CHMP2B_mmgmos <- mmgmos(affybatch.CHMP2B)
write.reslts(eset_CHMP2B_mmgmos, file="eset_CHMP2B_250615")
```

This process was repeated for each of the data sets listed above.

TDP43_Signature.R

TDP43_Signature is a piece of code employing the Pathprint library to identify common pathways across multiple data sets that are differentially expressed. As mentioned above, we are trying to identify pathways that are differentially expressed in all cases presenting with TDP-43 pathology. This is achieved by the following steps:

First, the pathprint library is called, and a pathway threshold is set

```
library (pathprint)
options(stringsAsFactors = FALSE)

thres <-100
```

The threshold will later define how many of the top differentially expressed paths in each list will be selected. Next, the working directory is set to the location of the first data set, and the .csv file created in the pre-processing stage (see above) is read in

```
Setwd ("/Users/clairegreen/Documents/PhD/TDP-43/Data Sets/CHMP2B")
exp_CHMP2B.LCM <- read.csv ("eset_CHMP2B_250615_exprs.csv", header=TRUE)
```

The first column is defined as containing the gene names, and the remaining columns are assigned to a variable which can be used for analysis.

```
row.names (exp_CHMP2B.LCM) <- exp_CHMP2B.LCM[,1]
exp_CHMP2B.LCM<- exp_CHMP2B.LCM[,2:11]
```

The experimental data is then processed by the function 'exprs2fingerprint'. This is a function of pathprint which uses the RNA expression data to assign a ternary score (-1,0,1) to a list of pre-definied pathways. The score represents whether that particular pathways is upregulated (1), downregulated (-1), or unchanged (0)

```
CHMP2B.LCM_pathprint <- exprs2fingerprint (exp_CHMP2B.LCM,
                                           platform = "GPL570",
                                           species="human",
                                           progressBar=T)
```

The platform refers to the GEO reference for the type of array used (in this example, the Affymetrix Human Genome U133 Plus 2.0 array), and species is defined as human.

Once this process is complete, the differential expressed between disease and control is calculated for each pathway. First, ternary scores for each pathway are averaged across disease conditions, and control conditions. The control scores are then deducted from the disease scores, indicating the difference in expression. These scores are ranked by decreasing value, and the top X number of pathways (indicated by 'thres') is selected and stored as a variable.

```
c <- apply (CHMP2B.LCM_pathprint[,1:7], 1,mean )
d <-  apply (CHMP2B.LCM_pathprint[,8:10], 1,mean )
t <- d-c
t1 <- t[order(abs(t), decreasing=T)]
CHMP2B.lcm <- (names(t1))[1:thres]
```

This process is completed for each of the data sets, resulting in a list of the top (thres) differentially expressed pathways for each. The next stage is to find the pathways that are common across all these data sets. This is completed via the following:

```
overlap <- Reduce(intersect, list(c9.lcm, CHMP2B.lcm, SALS.lcm, FTLD_FCx, VCP))
```

This line of code takes the intersecting pathways of each of the 5 datasets, and stores them in ‘overlap’. By printing overlap, we can see the pathways that are commonly differentially expressed across all TDP-43 pathology-containing data sets.