Lab Book 15_1_16

Claire Green

18 January 2016

Monday

I started by having a look at the random permutation test that I had been conducting before Christmas. The concept is that when conducting differential gene expression, and finding a consensus, I needed to prove that the liklihood of me discovering the genes I did was higher than chance. Subsequently I had to create a script that generated a random set of genes the same length as the samples from the 5 data sets (2000, 3000, 4000, 5000), and look for consensus between those sets. My aim is to prove that there are more consensus genes for my 5 data sets than would be expected by chance.

The script is as follows:

```
annotation <- read.table(annotation.file, header = TRUE, row.names =
NULL,
    sep = "\t", skip = 0, stringsAsFactors = F, quote = "",
comment.char = "!",
    fill = TRUE)  #import list of all genes
annotation <- subset(annotation, subset = (Gene.Symbol != "---"))  #if
no gene symbol, discount

# indicate the number of overlapping genes identified by DE
# analysis
test <- 181

m = 10000  #number of repetitions
r <- c(1:m)  #store repetition numbers in vector 'r'

for (j in 1:m) {
    random1 <- sample(annotation$Gene.Symbol, size = 5000, replace = F)
#size = size of sample for consensus
    random2 <- sample(annotation$Gene.Symbol, size = 5000, replace = F)
    random3 <- sample(annotation$Gene.Symbol, size = 5000, replace = F)
    random4 <- sample(annotation$Gene.Symbol, size = 5000, replace = F)
    random5 <- sample(annotation$Gene.Symbol, size = 5000, replace = F)
    random <- Reduce(intersect, list(random1, random2, random3,
        random4, random5))
    r[j] <- length(random)
}

test1 <- which(r > test)  #how many replicates had larger consensus
```

```
lists than the experimental condition
result <- (length(test1)/m)  #calculate p value
```

I need to run this by Wenbin to check the statistical robustness of this technique.

Tuesday, Wednesday, Thursday

I decided that now I know more about what I am doing, and the techniques that I am using, I would write my own experimental plan for the forseeable future. This was developed over the week, with comments from Win

```
What pathways are commonly dysregulated in diseases exhibiting TDP-43
pathology?
Do these pathways teach us anything new about the nature of TDP-43
pathology?
Can these pathways act as a predictive signature for disease?
Are these pathways suitable targets for drug intervention?
Are there common pathways in dysregulation that provide insight into
the mechanism of TDP-43 pathology?

--------------------------------------------------------------------------
------------------------------
RNA Expression Data

To fit the criteria to be accepted, data must be:

. RNA expression data from a platform covered by Pathprint
. Samples must be from patients known to have TDP-43 pathology
. Samples must have controls

ALS

*All Affymetrix HG-U133 Plus 2.0 Array*

8 C9orf72 LCM motor neuron samples + 3 controls
3 CHMP2B LCM motor neuron samples + 6 controls
7 sALS LCM motor neuron samples + 3 controls

FTLD

*All Affymetrix HG-U133A 2.0 Array*

6 GRN LCM cortical neuron samples
10 sFTLD LCM cortical neuron samples
+ 8 controls

VCP Myopathy

*Affymetrix HG-U133 Plus 2.0 Array*
```

7 VCP muscle cell samples + 3 controls

Quality Control of Data

Quality control allows the identification of samples that are deemed to act as noise in the data. These are outliers, and must be evaluated in terms of their inclusion in experimentation.

- Principle Component Analysis - plots the points according to variables that show the most variance thereby highlighting outliers.
- Box plots - plots the variance of data within each sample. Analysis of the spread of components indicates if any particular samples are outliers.
- Hierarchical Clustering - All samples are clustered using a clustering algorithm. It's expected that similar samples will cluster together. If any samples are separated from the group or clustered in an unlikely group, they are highlighted as outliers.

For a sample to be labelled an outlier, it has to be implicated in all three tests. Samples are only removed if absolutely necessary.

. Investigation at the Gene Level

Differential expression analysis

Determine consistent differentially expressed genes:
- Between healthy and dysregulated tissue --> MN, CN, Muscle
- By mutation --> C9orf72, CHMP2B, GRN, VCP
- Overall consensus DE genes across all data sets

2. Take top portion of each list of genes (most differentially expressed? Highest fold change? Variance?) and validate a functional relationship
see edgeRun: an R package for sensitive, functionally relevant differential expression discovery using an unconditional exact test Emmanuel Dimont, Jiantao Shi, Rory Kirchner and Winston Hide)

1. Validation - does my network have more 'connectedness' than is expected by chance? <>is this a validation?<>
<> what is the purpose of asking that question?

Create networks of random genes and generate p Value<> so this is the

first step - this step establishes whether you have found functionally valid DEG, its not a validation step, its a check to see if you have realistic results that may be viable for further analysis.<>

2. Further investigation for functional roles (also see Dimont paper) - build a co-expression network using background edges from the Genemania package

<>What is the purpose of enlarging the network ? Could you get what you need from simple straightforward network association of existing genes?<>

- look for any genes/modules that consistently appear by enrichment, By gene overlap?<>
- Maybe try another platform such as DAVID or GOrilla to see if those functional groups are consistent

>Why would you not just standardise upon GSEA categories and a single version of GSEA so that you have consistently comparable results?<>Also, please look at the endeavour and other papers I have shared with you on comparing enrichment categories<>

3. For the same lists (or those viewed as appropriate), look for co-regulation

- CORD (The CO-Regulation Database), Database of Gene Co-Regulation (dGCR), GeneFriends (read papers on these)
- Try to identify clusters of co-regulation
- Compare to co-expression

<>Although this sounds attractive, regulatory analyses are a huge ask.<> what question are you trying to address here? Are you trying to determine if these genes are all co-=regulated? They are likely to be so but by a number of different mechanisms and regulators, all interacting in some unknown manner....

Validation - Are my genes more co-regulated than is expected by chance? Test co-regulation with random sets of genes. Generate p Value. <> good to see that you are thinking along these lines><this is how co-regulation tests are built<>

4. Construct a PCA plot of gene sets to visualise clusters of genes <> (between and with studies)<> the purpose of the PCA analysis is to see if there are co-clustering sets of genes - from different experiments, mutations etc, that may be of interest in terms of investigating the commonality between mutant reagents<>

Validation - Are my genes more clustered than is expected by chance?

Test clustering with random sets of genes. Generate p Value.

By this point, analysis will hopefully have identified sets of DE genes that have both an internal functional relationship, but perhaps a functional relationship between sets as well. These relationships will be both biologically and statistically robust. The purpose of this stage of investigation is to highlight genes or gene sets that could be used to support results generated at the gene set level.

Investigations at the gene set level

Analysing dysregulation of pathways

Pathprint - Activity-based analysis >> pathway activity defined by differential expression of constituent genes as compared to a database
GSEA - Enrichment-based analysis >> pathway activity defined by enrichment of constituent genes at the extreme ends of a gene list

For each method, look for consensus dysregulated pathways across data sets. Validation - are there more common dysregulated pathways than you would expect by chance?<> is this a validation or ?<>
Attempt to identify commonalities between the two methods - relate back to stage two >> are there any patterns emerging?
In the same way that co-regulation of genes was investigated, are there any pathways that are co-regulated? Validation - are there more co-regulated pathways than you would expect by chance?

By the end of stage three, the aim is to have a pathway or set of pathways that is consistently dysregulated in association with TDP-43 pathology. The validity of this signature will have been evaluated both biologically and statistically, to ensure robustness.

STAGE FOUR: Target prioritisation <> this is a big area to explore<>

If we are able to generate a signature for TDP-43 pathology, this signature will likely contain many genes. It is within our interest therefore to identify which genes are:
Most connected (hub genes) and therefore most influential over the whole system
Realistic targets in terms of available drugs

Gene prioritisation software is available e.g. Endeavour, but another method is to see which genes are enriched for known SNPs.<> why would you consider this here, and not earlier in your analyses in order to determine if the DEGs you have discovered are in fact harbouring genes that are functionally and genetically linked to the phenotype?<>