Stefan Wyder
URPP Evolution
University of Zurich

May 2014

Exercises NGS Tutorial – Part 4 – Gene list Annotation

Exercise 4: Pathway Over-Representation Analysis using DAVID

Next we will do a Pathway Enrichment Analysis using the DAVID webtool (http://david.abcc.ncifcrf.gov/). Pathway analysis often gives better biological insight than Gene Ontology. DAVID is one of most widely used tools to annotate gene lists in model species. Its interface is not very intuitive but its pathway visualization is quite good.

First we have to prepare the input file (1 gene name per line). We take the 734 genes overexpressed in Adults relative to Newborns (Exercise 1) with a FDR < 5%.

```
awk 'BEGIN {FS="\t"} NR>1 && $5>1 && $8<=0.05 {print $1}' diffExp_N-A.txt | sed 's/"//g' > diffExp_N-A.genes_FDR5perc_greater1.txt
```

Open the site http://david.abcc.ncifcrf.gov in your web browser.

1. Press <Start Analysis>
2. Click on the tab <Upload> and choose the file containing the gene list we created above
3. Under Step2 select identifier: <ENSEMBL_GENE_ID>
4. Under step 3 select <Gene List>
5. Submit the List, then In the appearing Analysis Wizard choose: <Functional Annotation Chart>
6. Click on <Pathways> in red and then click on <Chart> for <KEGG pathways>.
7. A new window will open containing sorted according to their enrichment statistics.

Choose some pathways and open them by clicking on the underlines Pathway name. Blinking asterisks denote genes present in our input list. The pathway maps are **clickable,** if you click on a gene a new windows opens with the gene's description.

DAVID uses from the KEGG database. Check the scope of available pathways in KEGG on http://www.genome.jp/kegg/pathway.html

Exercise 5: Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a more sensitive and robust alternative to Over-Representation Analysis. Next we want to find out using GSEA which pathways are enriched in the Newborn - Adult comparison from Exercise 1. Here we will be using the original GSEA software that has a nice Graphical User Interface providing its output as simple webpages. However, many derivatives of the GSEA approach are now available in Bioconductor with more precise statistics (check http://www.bioconductor.org/packages/release/BiocViews.html#___Pathways).

Download the GSEA software and exercise files by:

```
wget https://dl.dropboxusercontent.com/u/3435091/GSEA_files.zip
```

First we need to provide a ranked list of genes. Here we take the probabilities for differential expression between the 2 groups, Newborn vs Adults, and we separate the 2 possible directions by multiplying FDR with -1 for genes which are underexpressed in Newborns relative to Adults (we can take either p-value or FDR which give the same result).

```
awk '$5<1 {factor=-1} {print $1"\t"$8*factor;factor=1}' diffExp_N-A.txt
| sed 's/"//g' > diffExp_N-A.4GSEA.rnk
```

We also have to provide gene sets. The GSEA gene set files are available for human only therefore I have prepared a file c2.cp.v2.5.symbols_MmProjection_ensembl.gmt which projects the human pathway annotation to mouse genes via orthology.

Now we are ready to run GSEA. Launch GSEA in the virtual machine:

```
java -jar /software/GSEA/gsea2-2.0.14.jar
```

Next we load the data. Cllck on the on the left <Load data> and <Browse for files…> then select the 2 required files.

In the menu select <Tools | GseaPreranked>. Fill out the required fields (Gene matrix local gmx/gmt) Run with <Normal> CPU Usage.

Once finished, click on the green <Success> in the lower left corner of GSEA. A webpage will then open in your browser.

Explore the most significant pathways.

 Are they all plausible?

 Go to <Detailed enrichment results in html format> and on the following page go to the <Details> for some pathways. Try to understand the different columns.

Documentation of the GSEA method (User guide, tutorials, file formats etc) are available under http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Main_Page

You can also access the GSEA output directly from the web browser (the folder containing the date might be called differently - use the <Tab> autocompletion!)

```
firefox
/home/student/gsea_home/output/may26/my_analysis.GseaPreranked.1401102143051/index.html
```