Function enrichment analysis

After RNA-seq data analysis, you will get a list of differentially expressed (DE) genes. In this exercise, you will use two different ways to test which function categories are enriched in your DE gene list.

There are three GO domains: cellular component (CC), biological process (BP), and molecular function (MF). During the exercise, you will only do function enrichment test for the BP domain. When you work on your real project, you will need to test all three.

1. Prepare working directory and software

1.1 Make a working directory and copy over data files for this exercise

```
mkdir -p  /workdir/$USER/exercise3

cd /workdir/$USER/exercise3

cp /shared_data/RNAseq-function/* ./

ls -l
```

1.2 Install GSEA on your laptop.

Go to http://software.broadinstitute.org/gsea/index.jsp and click "Downloads" to download and install the software.

(Optional) Install Cytoscape. Go to https://cytoscape.org/ and click "Download x.x.x" to download and install the software. After installation, you will need to install an app within Cytoscape named "EnrichmentMap". Start Cytoscape. Click "App"->"App Manager". Search for "EnrichmentMap" and install this app.

2. Over representation analysis (ORA)

The topGO package of Bioconductor will be used to do ORA. You will perform ORA on RNA-seq results of a yeast experiment.

2.1 Download GO annotation from Ensembl BioMart.

Ensembl (https://ensembl.org/) is a good resource to retrieve existing GO annotation (For plant genomes, go to https://plants.ensembl.org/)

- Use a web browser to navigate to Ensembl web site: https://ensembl.org/ , and click BioMart;
- From the pull-down menu "CHOOSE DATABASE" select "Ensembl Genes xxx";
- From the pull-down menu "CHOOSE DATASET" select "Sacchromyces cerevisiae";
- In the left panel, click "Attributes";
- In the right panel, expand "GENE", make sure "Gene stable ID" is checked, and Transcript stable ID" is unchecked.
- In the right panel, expand "EXTERNAL", check " GO term accession"
- At the top of the page, click "Results". You should see a table with two columns.
- Click "Go". You will be prompt to save the file as "mart_export.txt".

2.2 Convert the GO annotation file into a format compatible with topGO.
Use FileZilla to upload the "mart_export.txt" file to your assigned Linux server, keep it in the directory "/workdir/$USER/exercise3".

We provide you with a script that converts the BioMart file into topGO format. Run the commands:

```
cd /workdir/$USER/exercise3
/shared_data/RNAseq-function/toTopGO.pl mart_export.txt
```

After this, you would see a new file created: topgoAnnot.

Check the difference between the original file "mart_export.txt" and the converted file "topgoAnnot".

```
less mart_export.txt

less topgoAnnot
```

If a gene has multiple GO accessions, in the file "mart_export.txt", each of these GO accessions are in separate lines. In the the converted file "topgoAnnot", all GO accessions of the same gene are in a single line.

2.3 Run topGO

You are provided with two gene lists: refset and DElist_pos.

Check the content of these two files:

```
less refset

less DElist_pos
```

- refset: Genes in the reference set. Normally we use all genes that have none-zero expression in your RNA-seq experiments.
- DElist_pos: Up-regulated DE genes.

Run the R script topGO.r:

```
/programs/R-4.0.0/bin/Rscript /shared_data/RNAseq-function/topGO.r topgoAnnot
refset DElist_pos 0.1 BP myBP
```

- topgoAnnot, refset and DElist_pos: three input files;
- 0.1: cutoff p-value for reporting enriched GO categories.
- BP: test for biological processing GO. You can also test for MF (molecular function) and CC (cellular component).
- myBP: output file name prefix.

After this, you should find two new files "myBP".

- myBP: a text file.
- myBP_Topgo_weight01_134_all.pdf: a PDF file with tables and plots.

You can use FileZilla to download both files to your laptop to examine the contents.

The myBP file can be opened with a text editor. You can copy-paste each section in this file into an Excel spreadsheet (using text import wizard with "fixed width"). The enriched GO ids are sorted by p-values in the column "topgoFisher".

When you write "method" section of your manuscript, you can write that "gene set enrichment analysis was performed with BioConductor topGO package, using its default weight01 algorithm and topgo Fisher test, and p-value threshold is set at ...".

3. Gene Set Enrichment Analysis (GSEA)

We provide you with two data files (bp.gmt & genes.rnk) to run GSEA. When you work on your own project, here is how to prepare the two files.

- .gmt: The file format is explained in https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats . If you start with a GO annotation file like the one downloaded from Ensembl, use this procedure to convert to .gmt file (https://biohpc.cornell.edu/lab/userguide.aspx?a=software&i=719#c)
- .rnk: The file format is explained in the same web site as "".gmt". Briefly, it is a tab delimited text file with two columns: gene name & log2(FoldChange). The order of genes in this file does NOT matter, as GSEA will sort this file based on column 2. You can get the data from DEseq2, removing genes with expression level too low (filter by the DEseq2 output column baseMean)

You can run GSEA either on your laptop, or on the LINUX server.

3.1 Run GSEA on laptop.

Use FileZilla to download the two data files bp.gmt & genes.rnk from the directory /shared_data/RNAseq-function/.

- Start GSEA.
- Click "Load data", then click "Browse for files", and open the two files bp.gmt & genes.rnk
- Click "Run GSEA Preranked", and set the following:
    - Gene sets database: open-> click "Gene matrix(local gmx/gmt)" -> select "bp.gmt" -> "OK";
    - Number of permutations: Enter 1000;
    - Ranked List: select "genes";
    - Collapse: select "no collapse";
    - Chip platform: leave blank
    - Under "Basic fields" -> Enrichment statistics-> select "weighted" *
    - Set analysis name and output directory. OK to use default.
    - Under "Advanced fields"-> "Plot graphs for the top sets", you might want to increase to 40.
- Click "Run". It would take a few minutes. After it is done, you should see "Success".

**Enrichment statistics:* the high p value would put more weight to genes with high fold change when calculating enrichment scores (ES). Use the default "Weight" in most cases.

3.2 Examine results

Click "Success". If it does not work, click "Show results folder", which should open the directory. Open the directory of your analysis, double click the file "index.html".

On the page, you should see two block of "Enrichment in phenotype". One for gene sets enriched in up-regulated genes (na_pos) and one for gene sets enriched in down-regulated genes (na_neg). Click "Detailed enrichment results in html ", you would see enriched gene sets.

3.3 (Optional) Enrichment map visualization.

Interpretation of the map can be found at this page: https://enrichmentmap.readthedocs.io/en/docs-2.2/Tutorial_GSEAInterface.html

- Start the software Cytoscape;
- In GSEA, click the "Enrichment map visualization";
- In GSEA, select a GSEA result from the application cache;
- In GSEA, click "Build Enrichment Map";
- In Cytoscape, you should see the network map;
- In Cytoscape, it might look messy, you will need to manualy adjust it;
- In Cytoscape, "Style" -> "Label" -> delete (click the "trash can");
- In Cytoscape, "Style" -> "Label" -> "Select value" -> "Name", "Mapping type"->"Passthrough mapping";
- In Cytoscape, drag each dots to make it neat.

3.4 (Optional) Run GSEA command on LINUX

With GSEA open on your laptop, click "command" at the bottom of the GSEA window. Copy-paste the command to a text editor.

You need to change a few things for this command to run on BioHPC computer:

- "gsea-cli.bat" -> "/programs/GSEA_Linux_4.0.3/gsea-cli.sh"
- Fix the file paths for "-rnk" "-gmx" "-out".

Now run the command on Linux server. After done, you can examine the results in the "out" directory.