Exercise 2. Function enrichment analysis & gene function annotation

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

If you work on a non-model organism, chances are that you do not have Gene Ontology (GO) annotation of your reference genome. GO annotation is required for function enrichment analysis. We will use BLAST2GO software to generate GO annotation.

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

Part 1. Over representation analysis (ORA)

The topGO package of Bioconductor will be used to do ORA. We will perform ORA on RNA-seq results of a yeast experiment. The reference is https://www.ncbi.nlm.nih.gov/pubmed/16606683 .

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 xx";
   - 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. Convert the file you downloaded into a format compatible with topGO.
   Use FileZilla to upload the "mart_export.txt" file to your assigned Linux server, keep it in the /workdir/$USER.

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

```
cd /workdir/$USER
/shared_data/annotation2019_2/toTopGO.pl mart_export.txt
```

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

3. You are provided with two gene lists: refset and DElist_pos.
   Run the R script topGO.r:

```
   cp /shared_data/annotation2019_2/* ./
   Rscript /shared_data/annotation2019_2/topGO.r topgoAnnot refset DElist_pos
0.1 BP myBP
```

· topgoAnnot: the GO annotation file in the topGO required format;

· refset: a text file with list of reference set of genes with one gene per line (normally all genes that have none-zero expression in your experiments)

· DElist_pos: a text file with the list of up-regulated DE genes.

· 0.1: cutoff p-value for enriched 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_132_all.pdf: a PDF file with tables and plots.

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

In the myBP file, the enriched GO ids are listed sorted by p-values in the column "topgoFisher".

In your publication, you can say that "gene set enrichment analysis was performed with BioConductor topGO package, using its default method weight01, and p-value threshold is set at ...".

Part 2. 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 either GSEA on your laptop, or on the LINUX server.

1. Install and run GSEA on your laptop.

- Install GSEA. 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.

- Use FileZilla to download the two data files bp.gmt & genes.rnk from the directory /shared_data/annotation2019_2/;
- 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.

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. (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.
4. Run GSEA command on LINUX (optional).

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.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.

Part 3. Run BLAST2GO to create GO annotation. (Do it at home)

In part 1 of this exercise, you have copied all files in /shared_data/annotation2019_2/ to /workdir/$USER. You will run BLAST2GO on the fasta file: annot_exercise_aa.fasta

1. Run BLASTX against SWISSPROT database. BLASTX is a command line software.

```
  cd /workdir/$USER
  cp /shared_data/genome_db/BLAST_NCBI/swissprot* ./

  blastp -num_threads 4 -query annot_exercise_aa.fasta  -db swissprot \
    -out blastresults.xml  -max_target_seqs 20 -evalue 1e-5 -outfmt 5 \
    -culling_limit 10 >& logfile &
```

2. After it is done, copy the result file into your home directory.

```
cp blastresults.xml   ~/
```

3. Run BLAST2GO on cbsumm10.

   As we only have one license of BLAST2GO. It is installed on the computer cbsumm10.biohpc.cornell.edu.

   Login to the computer cbsumm10.

```
mkdir /workdir/$USER
cd /workdir/$USER

cp /shared_data/blast2go/* ./
cp ~/blastresults.xml  ./

/usr/local/blast2go/blast2go_cli.run -properties annotation.prop -useobo go.obo
\
     -loadblast blastresults.xml -mapping -annotation -annex -statistics all \
     -saveb2g myresult -saveannot myresult -savereport myresult -tempfolder ./ \
     >& annotatelogfile &
```

After this step, you will get three result files:

- myresult.annot. It is a text file with GO annotation. You can open it in a text editor or Excel.
- myresult.b2g; It is a project file that you can open in the free version of BLAST2GO GUI software.
- myresult.pdf. A report file with statistics of your data set.