Functional transcriptome analysis in non-model species Hands-on 2

F. Bucchini – K. Vandepoele

The aim of this hands-on session is to learn how to extract functional information out of a *de novo* assembled RNA-Seq dataset processed using TRAPID2.0.

Tools

- TRAPID2.0 http://bioinformatics.psb.ugent.be/testix/trapid_frbuc/

Data sets

- TRAPID FTP ftp://ftp.psb.ugent.be/pub/trapid/workshop/datasets/

EXERCISE 1 – General statistics

Starting from the MMETSP transcriptome you processed in hands-on 1, answer the following questions.

a. How many sequences are present in this experiment? For how many could an ORF be found? [Overview – Experiment information and Statistics – General statistics]
b. How many transcripts have been assigned to a gene family and how many gene families have at least one transcript? [Statistics – General statistics]
c. How many transcripts have GO annotation. How many have Protein Domain information?

EXERCISE 2 – Taxonomic binning and core GF completeness

a. Under [Taxonomic binning], what fraction of transcripts are assigned to Unclassified, *Eukaryota* and *Bacteria*? For these taxa, generate 3 subsets using the CTRL button in combination with 'Define new subsets'. For the largest Phylum [Bar chart], also generate a new subset (so create 4 subsets in total).
b. In preparation of exercise 4, Under [Overview] section 'Functional enrichment preprocessing', start 'Perform functional enrichment preprocessing for … subsets' and select 'GO' as data type.
c. Under [Explore subsets], select the subset with the smallest number of transcripts, and perform in the Toolbox 'Compare Label Gene Family intersection'. Which HOM gene family is present in all/most subsets? Can you identify how many genes are present from each subset (use the search bar, select Gene family)?
d. Perform a core GF completeness analysis. Run your analysis using a phylogenetic clade which is well represented based on the results of step d. (e.g. *Eukaryota*, *Chlorophyta*, *Stramenopiles*). What fraction of core GFs is represented in your complete dataset? Repeat this step by processing a subset of the transcripts.

EXERCISE 3 – Gene family and gene function analysis*

a. Using the search bar, search for 'transporter' using the Protein Domain description option. Which IPR protein domain contains the largest number of transcripts?
b. Select this IPR domain and determine how many gene families are associated to it [Toolbox]. Which HOM gene family contains the largest number of transcripts? [Data – Transcript selection]
c. For one of these transporter-related HOM gene families, select the gene family and generate a phylogenetic tree [Gene family Toolbox – Create phylogenetic tree]. Also inspect the multiple sequence alignment (MSA); do all MMETSP transcripts in this family have a full-length ORFs?

d. Which GO terms are linked to this gene family? [Gene family Toolbox – Functional data – View associated functional annotation]
e. Select the GO category 'transport', how many transcripts are assigned to this GO term?
f. Which more specific GO terms can be found for these transcripts? [GO term Toolbox – Explore the child GO terms]

* if 'transporter' fails, try this exercise with 'transcription factor'.

EXERCISE 4 – GO enrichment analysis

a. Under [Subset enrichment], select 'GO term enrichment'. Search in one the defined subsets for enriched GO terms (tip: try your smallest subsets first).
b. Explore the results in the GO enrichment bar charts, the enrichment graph, and the GO enrichment data table. By comparing the two latter items, can you explain why some enriched GO terms are labeled as 'Shown: X = false' in the table?
c. For a given GO enrichment, how many transcripts are annotated with that term [GO enrichment data table – Subset ratio]?
d. Under [Overview] section 'Functional enrichment preprocessing', start 'Rerun functional enrichment preprocessing for … subsets' and select 'Protein domain' as data type.
e. Under [Subset enrichment], select 'Protein domain enrichment'. For the enriched IPR Protein domain with the largest subset ratio, how many transcripts and gene families are present?

EXERCISE 5 – Advanced functional analysis incl. data export and new subset creation

a. Determine how many transcripts are present in your dataset encoding for transcription factors*. [Search bar, GO description, 'transcription factor']
b. Create a new subset containing the transcripts annotated with '"transcription factor activity' (GO:0003700). [Export Data – tab Functional data – Transcripts with GO].
    1. Download the ZIP file to your PC/laptop and unzip this tab-delimited text file.
    2. Using Excel or another spreadsheet application, open this file and filter for GO:0003700.
    3. Copy-paste the column 'transcript_id' in a plain .txt document (NOT Word!)
    4. Import this set of transcript as a new subset [Import data – Transcript subsets/labels], name it 'TranscriptionFactors'
c. Under [Overview] section 'Functional enrichment preprocessing', start 'Rerun functional enrichment preprocessing for … subsets' and select 'Protein domain' as data type.
d. Which type of transcription factors is most prevalent in this subset? [Perform Protein domain enrichment analysis]
e. To which gene families have these transcription factors been assigned? [Sankey diagrams – Label→ Enriched IPR→GF].
f. What fraction of transcription factors are classified as *Bacteria* / *Eukaryota* / Unclassified? [Explore subsets – Venn diagram]
g. Which protein domains are present in the 'TranscriptionFactors' subset and classified as Bacteria**? [Explore subsets – select subset TranscriptionFactors – Toolbox Label Interpro intersection]

* as an alternative, use 'generation of precursor metabolites and energy' (GO:0006091)

** if '*Bacteria*' fails, overlap with another subset defined based on the taxonomic binning output.