Functional transcriptome analysis in non-model species Hands-on 1

F. Bucchini – K. Vandepoele

The aim of this hands-on session is to learn i) the basic steps needed to perform functional sequence analysis for a *de novo* assembled RNA-Seq dataset and ii) how to perform the basic processing of a complete transcriptome using TRAPID2.0.

Tools

- NCBI BLASTX https://blast.ncbi.nlm.nih.gov/Blast.cgi
- TRANSLATE tool https://web.expasy.org/translate/
- 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 – BLAST-based ORF finding, taxonomic classification and functional analysis

Starting from the MMETSP0140 FASTA file in the TRAPID FTP folder, use BLASTX and the TRANSLATE tool to identify the open-reading frame (ORF; at which frame is there coding potential? Is this a full-length or partial ORF?), to perform taxonomic classification (to which kingdom, phylum and genus can this sequence be assigned to?) and try to assign a function to the sequence. Do this for these 3 transcripts:

MMETSP0140.Transcript_1002
MMETSP0140.Transcript_1023
MMETSP0140.Transcript_256

EXERCISE 2 – Processing full transcriptome using TRAPID2.0

2.1 Starting from the shared dataset MMETSP0140 processed in TRAPID2.0, answer the following questions.

a. Under [Menu bar – Log] how long took the taxonomic binning [tax_binning kaiju_mem]? How long took the complete initial processing [initial_processing start / stop]?
b. How many sequences are present in this experiment? [Overview – Experiment information]
c. Which reference database was used to process this transcriptome?
d. How many transcripts have a full-length ORF? [Statistics – General Statistics – Meta annotation information]
e. What can you tell about the length of the transcripts lacking ORF information (called 'non-information')? [Statistics – Length distribution sequences – Select 'Display 'partial' data separately' + select 'Display 'non information' data separately' – Select 'Graph type: stacked'].
f. Based on the plot from e., adjust the plot by plotting 10 bins. What is the fraction of partial sequences in the first and the second bin?

2.2 Analyze the taxonomic binning results for this experiment.

g. What fraction of transcripts could be taxonomically binned? [Taxonomic binning - Krona]
h. How many transcripts were taxonomically classified as *Bacteria* and *Eukaryota*? [Taxonomic binning – Tree]
i. Which phylum has the largest number of transcript based on the taxonomic binning? [Taxonomic binning - Bar]
j. What fraction of Eukaryotic transcripts is assigned to the *Stramenopiles*? [Explore subsets]

EXERCISE 3 – Submit your own transcriptome

Upload and start processing an MMETSP transcriptome present in the TRAPID FTP folder.

a. Go to the experiments overview page [TRAPID Home - Experiments]
b. Select 'Add new experiment'
c. Add as name the MMETSP identifier, select reference database 'Pico PLAZA 2.0' and click 'Create Experiment'
d. In the experiment table, select your experiment, and select 'Import data – Transcripts'
e. In case you downloaded the transcriptome FASTA file from the TRAPID FTP folder, upload the file. Otherwise, copy-paste the FTP URL to the FASTA file in the URL box. Select 'UPLOAD FILE/DEFINE URL'
f. Select 'LOAD DATA INTO DATABASE' (this should take 2-5 minutes; you will receive an e-mail when ready)
g. Select your experiment, and on the overview page select 'Perform transcript processing'
h. In 'Similarity search database – Phylogenetic clade', choose Eukaryotes
i. In 'Gene families and annotation options', for functional annotation, select 'Both'
j. Click 'RUN INITIAL PROCESSING ' to start processing your transcriptome

EXERCISE 4 – Gene space completeness using Core Gene Families (core GFs)

Starting from the shared dataset MMETSP0140 processed in TRAPID2.0, go to [Core GF completeness] tab and compare the results from 'Previous analyses'. Specifically:

a. When performing the core GF completeness using *Eukaryota* core GFs and default parameters, how many core GFs were defined and how many are present/missing in this transcriptome?
b. Give an example of missing core GF and its function. [Missing GFs table – follow linkout GF identifier]
c. How many core GFs were defined at the *Bacillariophyta* level, and how many are present in this experiment?
d. Can you explain why the number of (missing) core GFs is much larger for *Bacillariophyta* than for *Eukaryota*?