In-depth functional and comparative analyses of transcriptomes with TRAPID

Francois Bucchini,Michiel Van Bal,Klaas Vandepoele

Abstract

TRAPID is an online tool for the fast, reliable and user-friendly analysis of *de novo* transcriptomes, developed and maintained by the CNB group at VIB-UGent Center for Plant Systems Biology.

This protocol explains how to perform advanced functional and comparative analyses after a transcriptome dataset was functionally annotated using TRAPID.

Citation: Francois Bucchini,Michiel Van Bal,Klaas Vandepoele In-depth functional and comparative analyses of transcriptomes with TRAPID. **protocols.io**
dx.doi.org/10.17504/protocols.io.nhzdb76
Published: 02 Mar 2018

Guidelines

In this protocol, we work with an example dataset consisting in 25,392 transcript sequences from *Panicum hallii* (Meyer et al. 2012). This dataset can be retrieved from TRAPID's FTP here.

Before start

In order to use TRAPID, make sure that:

- You are using a web browser with JavaScript and Adobe Flash enabled. Supported browsers: Safari, Chrome, Firefox (most recent versions).
- Java is installed on your machine (version JRE 6.0 or higher). This is required to launch the JalView alignment editor and the Archeopteryx tree viewer.

Protocol

Initial processing

Step 1.

The analyses presented in this protocol require to know how to perform initial processing of transcript sequences using TRAPID. For detailed guidelines regarding how to perform initial processing and a general overview of the TRAPID platform, please see our 'Functional annotation and analysis of de novo transcriptomes with TRAPID' protocol.

In this second protocol, we will work with the same example dataset as in the first one. It consists in 25,392 transcript sequences from *Panicum hallii* (Meyer et al. 2012). This dataset can be retrieved from TRAPID's FTP here.

Frameshift correction

Step 2.

For transcripts that were flagged as potentially containing frameshifts during the initial processing, we offer the possibility to run FrameDP to putatively correct the transcript sequence and identify the correct Open Reading Frame (ORF). FrameDP is a program that uses BLAST together with machine learning methods to build models used to test whether a sequence has a putative frameshift or not. The model is then used to correct the sequence (by inserting N-nucleotides at the necessary locations), which of course also impacts the associated ORF.

To perform frameshift correction of a transcript, select '*Correct framseshifts with FrameDP*' in the '*Toolbox*' on any transcript page.

Frameshift correction example (single *Panicum* transcript containing an indel)

1. Create a new experiment (named '*Tutorial 2*' for example) using the example *Panicum* dataset and PLAZA 2.5 as reference database.
2. Start the transcript processing using '*Eudicots*' as the phylogenetic clade.
3. Once the processing is finished, go to the *Experiment overview* page and select a transcript (e.g. *contig17160*).
4. The transcript page will show the line 'A putative frameshift was detected in this sequence'. To attempt to correct this frameshift, select *Correct framseshifts with FrameDP* in the toolbox.
5. The next page asks if you want to correct additional genes from the same family. This is not needed here, so you can click '*Perform frameshift correction*'.
6. A new job is started and after a while you will receive a notification via e-mail.
7. After completion, TRAPID will indicate if a frameshift was corrected or not and on the transcript page the corrected ORF can be obtained.

Phylogenetic analysis of a specific gene family

Step 3.

Multiple sequence alignments

Starting from any given transcript, you can generate amino acid multiple sequence alignment (MSA) within a gene family context. As such, you can create an MSA containing the transcripts within a gene family together with a selection of coding sequences from the reference database. Within TRAPID, MSAs are created using MUSCLE. In order to reduce the computation time, the maximum number of iterations in the MUSCLE algorithm is fixed at three, and all other settings are left at default values.

Phylogenetic trees

Similar to the MSA, you can also create a phylogenetic tree within a gene family context. In order to create phylogenetic trees which are less dependent on putative large gaps, the standard MSA is transformed to a *stripped* MSA. In this stripped MSA the alignment length is reduced by removing all positions for which a certain fraction (0.10 for stringent editing, 0.25 for relaxed editing) is a gap. As such, all gaps introduced by a small number of sequences will be removed. Note that in some cases the stringent editing yields a stripped MSA with zero or only a few conserved alignment positions. In this happens, please re-run the tree analysis using the relaxed editing option (which will yield more conserved alignment positions).

Two different tree inference algorithms can be use within TRAPID: FastTree and PhyMl. Although FastTree is used by default (because of its very short execution time), you can choose which algorithm to use, and -if desired- how many bootstrap runs will be applied. If you defined subsets within your experiment, these subsets can also be displayed on the phylogenetic tree, making subsequent analyses much easier. By default, the meta-annotation is displayed as domains next to the phylogenetic tree.

From the phylogenetic tree page, the tree can be exported in PhyloXML and Newick formats.

How to generate MSAs and trees

1. Select a transcript, either through the search function, or through any link in the platform
2. On the transcript page, select the associated gene family
3. On the gene family page, select the *create multiple sequence alignment* or *create phylogenetic tree* from the toolbox
4. Select the reference species from the reference database you want to include in the tree

MSA and phylogenetic tree creation example

Using the processed *Panicum* dataset (as shown in our first protocol):

1. Search for the GO term *leaf senescence* and look at the genes assigned this label in the dataset.
2. Look at **contig04501**, this transcript is annotated as quasi full length. Now click the gene family ID next to the transcript identifier: this will take you to the corresponding gene family page.
3. Build the multiple sequence alignment for this gene family by clicking *Create multiple sequence alignment* in the toolbox. After the alignment is generated, a link will appear to start JalView to view the sequence alignment, or download the alignment.
4. Click *View full multiple sequence alignment* to start JalView. From this alignment can be seen the transcript has indeed a good alignment with most of the other members in the family, though at the N-terminal end there likely is a portion missing (and hence is indeed quasi full length). The other transcript (contig20276) is the opposite, here the C-terminal end appears to be missing. Potentially both contigs represent a single, split transcript.
5. Go back and look for **contig01069**. Find the gene family page of this transcript and in the toolbox select *Create phylogenetic tree*.
6. On the next page, tree creation settings can be adjusted. Switch the Editing from *Stringent editing* to *Relaxed editing* (optionally fewer species can be selected). Click *Create phylogenetic tree*.
7. Once the tree is built and you received the confirmation email (usually within a few minutes), go back to the gene family page and click View *phylogenetic tree* to view the created tree. Now an ATV applet will be included on the page which displays the tree, as shown in figure 1.

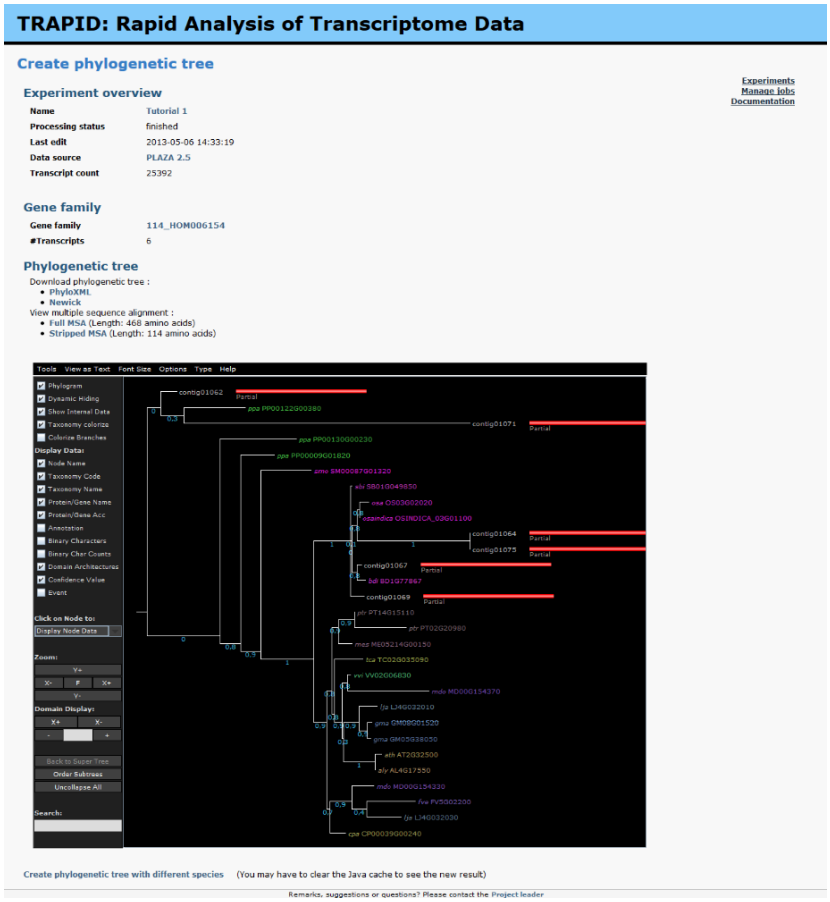

Figure 1: phylogenetic tree ATV applet displaying the phylogenetic tree of contig01069 and its orthologs.

NOTES

Klaas Vandepoele 28 Feb 2018

Settings used for phylogenetic tree creation: for FastTree we use the following non-default settings: '--wag gamma', which indicate that the algorithm uses the WAG+CAT model, and rescales the branch lengths. For PhyML we use the following non-default settings: '-m WAG -f e -c 4 -a e', which indicate that the algorithm uses the WAG+CAT model, that empirical amino acid frequencies are used, that 4 relative substitution rate categories are used, and that the parameter for the gamma distribution shape is based on the maximum likelihood estimate.

Within-transcriptome functional analysis using experiment subsets

Step 4.

Apart from the functional annotation of individual transcripts, and the analyses in the gene family context, TRAPID also supports the quantitative analysis of experiment subsets using GO and protein domain enrichment statistics. For more details regarding subset creation, refer to our other protocol.

Specific comparative analyses than can be performed using subsets are:

- GO terms or InterPro protein domains enrichment (subset versus all; hypergeometric distribution,
- GO terms / InterPro protein domains ratios between subsets, including subset-specific GO annotations,
- Subsets examination. It is possible to examine transcripts specific to a certain subset or a combination of subsets.

Functional enrichment analysis example: cell cycle genes

In this example, we want to see if there are any GO terms or protein domains that are significantly over-represented in a set of cell cycle transcripts compared to the whole *Panicum* transcriptome.

1. Download the list of cell cycle transcripts, available here (TRAPID FTP).
2. Add the labels to the dataset. From the experiment overview page, click *Import data* next to *Import transcript labels*. On the next page, hit browse to select the file, enter a label (e.g. *Cell_cycle*) and click *Import labels*.
3. Go back to the experiment overview page, in the *Toolbox* a number of new options are enabled. To check if this set is enriched for specific GO terms or protein domains, click *GO enrichment from a subset compared to background* or *Protein domain enrichment from a set compared to background*. For both GO and Protein domains, select the subset (here *Cell_cycle*) and the desired p-value and click *Compute enrichment*. The enrichment should be computed within a few minutes, at most.

Figure 10 shows the enrichment resultspage with for each of the enriched InterPro domains the fold enrichment, significance and a short description. Note that the InterPro domain codes are hyperlinks to pages with more detailed information.

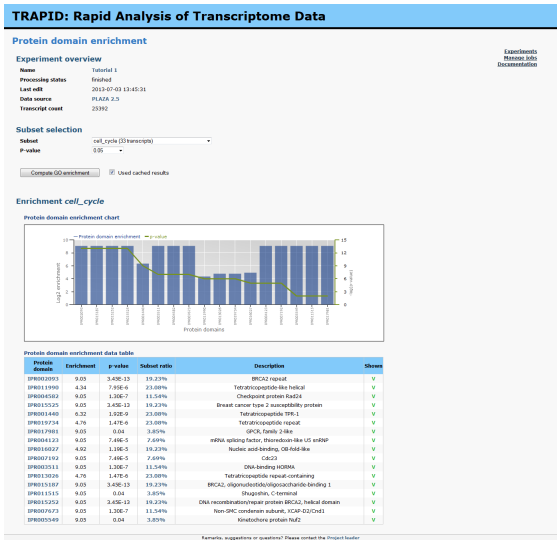

FIGURE 11 ENRICHED INTERPRO DOMAINS Overview of InterPro domain enrichted within the 'Cell Cycle' subset