Functional annotation and analysis of de novo transcriptomes with TRAPID

Francois Bucchini,Michiel Van Bel,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 functional annotation of a transcriptome using the TRAPID platform. It also describes basic analyses that can be performed within TRAPID, after initial processig of the data.

Citation: Francois Bucchini,Michiel Van Bel,Klaas Vandepoele Functional annotation and analysis of de novo transcriptomes with TRAPID. **protocols.io**

dx.doi.org/10.17504/protocols.io.nhxdb7n

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

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

User Authentication

Step 1.
Setting up an account

Before you can use TRAPID, you need to create an account.

1. Go to the registration page. To go there from TRAPID's homepage, click on the "*Register*" link at the bottom of the page.
2. Fill in the required information (valid email address used as login, organizaion and country). Make sure to provide **a valid e-mail address** as your password will be sent immediately by e-mail.
3. Click on the "*Register*" button.

Log in

On TRAPID's homepage, click "*Login*" at the bottom of the page. This will take you to the login page. There, use the **e-mail address** used to register and the **password** sent by mail to login.

After you've logged in, the homepage will be replaced by the *Experiments overview*. Since you are a new user, this overview will be empty. This page is composed of 3 main sections:

1. *Current experiments* (experiments you uploaded and own),
2. *Shared experiments* (experiments uploaded by others you are allowed to view),
3. *Add new experiment* where a new experiment can be started.

NOTES

Klaas Vandepoele 27 Feb 2018

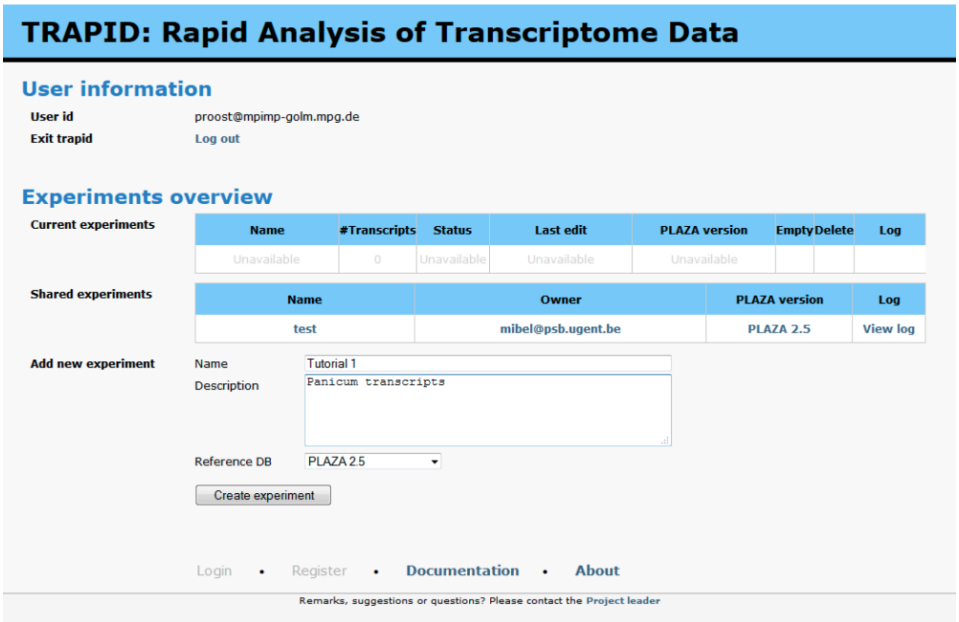
By requiring authentification, we ensure that no user has access to the data of any other user and that each experiment's data is kept private.

Creating experiments

Step 2.
Create a new experiment

From the *Experiments overview* page, **create a new experiment**. To do so, just fill in tour experiment's information in the *Add new experiment* section, as shown in figure 1. You can for example enter 'Tutorial 1' as a name and 'Panicum transcripts' as description. The reference DB should be left to its default setting (PLAZA 2.5).

We recommend using PLAZA as reference database when working with plants and algae. If data from other organisms is analyzed, select OrthoMCL-DB 5.0 as reference database.



TRAPID: Rapid Analysis of Transcriptome Data

User information

User id proost@mpimp-golm.mpg.de

Exit trapid Log out

Experiments overview

Current experiments

Name	#Transcripts	Status	Last edit	PLAZA version	Empty	Delete	Log
Unavailable	0	Unavailable	Unavailable	Unavailable			

Shared experiments

Name	Owner	PLAZA version	Log
test	mibel@psb.ugent.be	PLAZA 2.5	View log

Add new experiment

Name: Tutorial 1

Description: Panicum transcripts

Reference DB: PLAZA 2.5

Create experiment
Login • Register • Documentation • About
Remarks, suggestions or questions? Please contact the Project leader

Figure 1: experiments overview. At the bottom a name and description for the new experiment needs to be filled in for the new experiment. Experiment are created by clicking the *Create experiment* button.

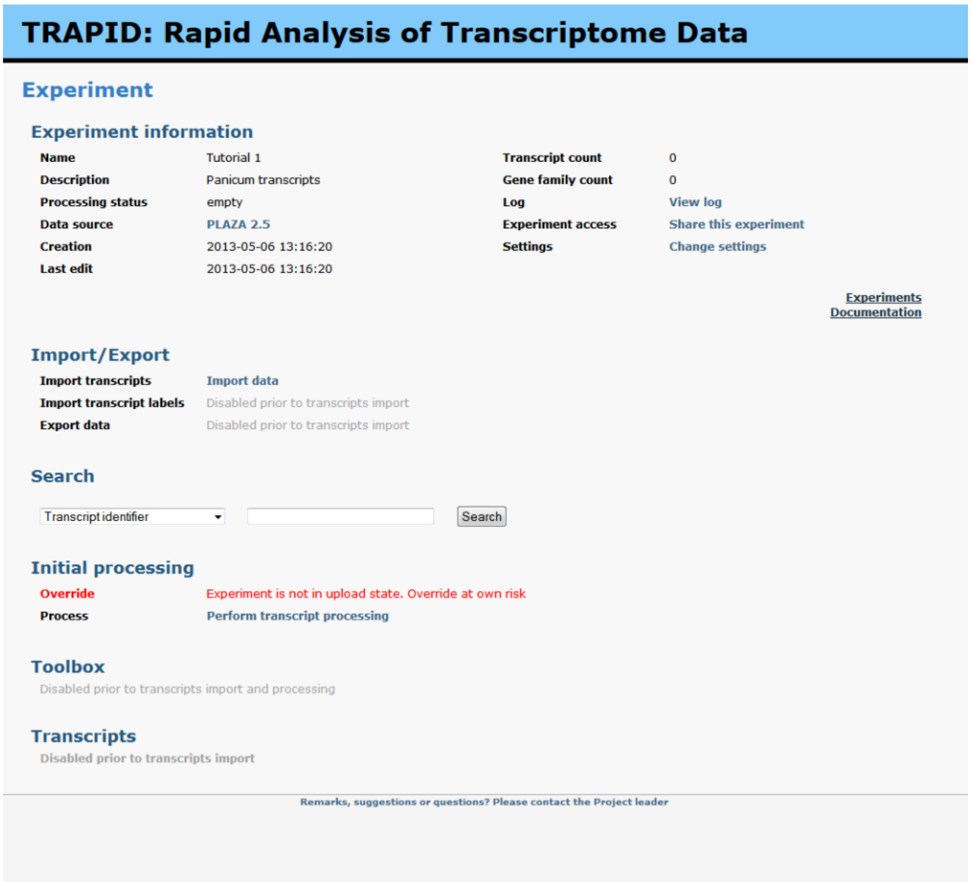
Add the experiment by clicking *Create experiment*. The new experiment will now appear in the current experiments list. Note that each user can have a **maximum of 10 experiments**.

Accessing your experiment and experiment overview

Click on the experiment's name in the *Current experiments* table to continue. This will take you to the experiment page.

From there, general statistics are shown (*Experiment information*), sequences can be imported and exported (*Import/Export*) and a search function is available to find specific sequences (*Search*). There also are options to start processing transcripts (*Initial processing*, more details in section 4). Once data is imported, a detailed overview of the transcripts and a toolbox will also become available.

An example of experiment page is given in figure 2.



TRAPID: Rapid Analysis of Transcriptome Data

Experiment

Experiment information

Name	Tutorial 1	Transcript count	0
Description	Panicum transcripts	Gene family count	0
Processing status	empty	Log	View log
Data source	PLAZA 2.5	Experiment access	Share this experiment
Creation	2013-05-06 13:16:20	Settings	Change settings
Last edit	2013-05-06 13:16:20		

Experiments
Documentation

Import/Export

Import transcripts	Import data
Import transcript labels	Disabled prior to transcripts import
Export data	Disabled prior to transcripts import

Search

Initial processing

Override	Experiment is not in upload state. Override at own risk
Process	Perform transcript processing

Toolbox

Disabled prior to transcripts import and processing

Transcripts

Disabled prior to transcripts import

Remarks, suggestions or questions? Please contact the Project leader

Figure 2: experiment page. The experiment page of our currently empty experiment. Click on '*Import data*' to import transcript sequences.

NOTES

Klaas Vandepoele 27 Feb 2018

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.

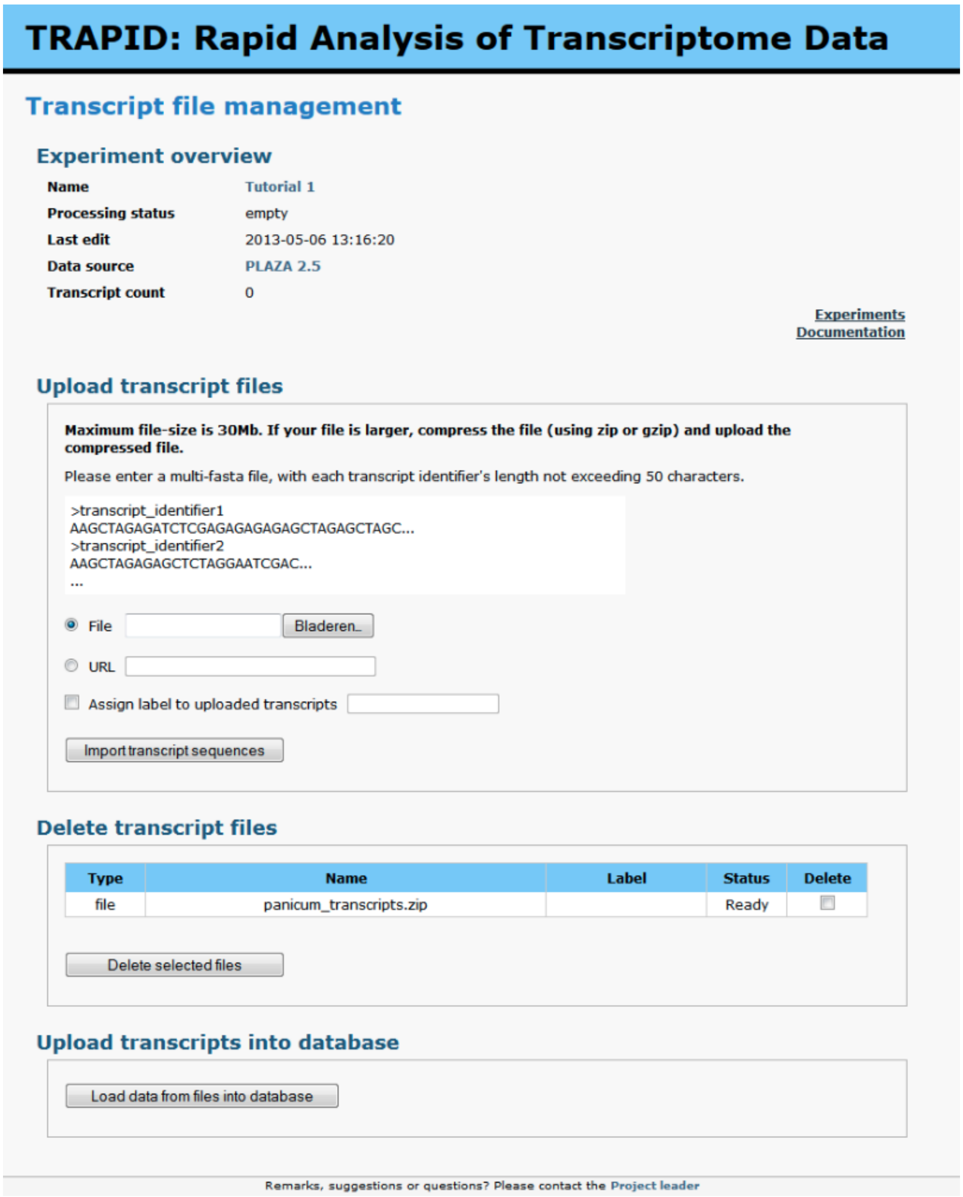
Uploading transcript sequences & job management

Step 3.

From the experiment page, click on *Import data* to go to the data upload page (referred as *Transcript file management* page in TRAPID). Two options are available:

1. in case you downloaded the sample *Panicum* dataset, you can upload the file (use *Browse* to locate the file on your system),
2. otherwise, you can simply provide the URL of the file.

Loading the data into TRAPID is then done in two steps. First, you need to upload the dataset (or datasets in case you are using multiple input files) by clicking *Import transcript sequences*. Then, click *Load data from files into database* to get the sequences into our database. **Both steps are essential before the data can be processed.** A screenshot of the transcript file management page is shown in figure 3.



TRAPID: Rapid Analysis of Transcriptome Data

Transcript file management

Experiment overview

Name	Tutorial 1
Processing status	empty
Last edit	2013-05-06 13:16:20
Data source	PLAZA 2.5
Transcript count	0

Upload transcript files

Maximum file-size is 30Mb. If your file is larger, compress the file (using zip or gzip) and upload the compressed file.

Please enter a multi-fasta file, with each transcript identifier's length not exceeding 50 characters.

```
>transcript_identifier1
AAGCTAGAGATCTCGAGAGAGAGCTAGAGCTAGC...
>transcript_identifier2
AAGCTAGAGAGCTCTAGGAATCGAC...
...
```

● File

○ URL

□ Assign label to uploaded transcripts

Delete transcript files

Type	Name	Label	Status	Delete
file	panicum_transcripts.zip		Ready	□

Upload transcripts into database

Remarks, suggestions or questions? Please contact the Project leader

Experiments
Documentation
Bladeren..
Import transcript sequences
Delete selected files
Load data from files into database

Figure 3: transcript file management page. From this page, transcriptome dataset(s) can be uploaded from your machine or a URL. After the data file(s) were added to our system, click the *Load data from files into database* button.

Once your data was successfully uploaded to TRAPID, an email will be sent to your login email address.

During all TRAPID processing steps (upload, transcript processing, running frameshift correction or computing alignment/phylogenetic tree), you can check the experiment status to see if your job is queued, running or in error status. In case you want to cancel or stop your job, go to the *Experiment Status* page (accessible form the experiments overview) and modify the status to *Finished*.

NOTES

Klaas Vandepoele 27 Feb 2018

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.

Klaas Vandepoele 27 Feb 2018

Requirements for input data:

* All data needs to be provided in FASTA format.

* Each sequence has a unique identifier (max. 50 characters) and no empty sequences should be present.

* Single files cannot be bigger than 32Mb (the limit is 300Mb if the file is provided as URL). For large datasets it is possible to use (g)zipped files.

Process transcript sequences (functional annotation)

Step 4.

After the sequences have been loaded to the TRAPID database (you received the completion email), you can now proceed with the initial processin of your transcripts. To do so, go to the *initial processing* page by clicking the *Perform Initial Processing* link on the experiment page. During this step, you should consider the options carefully, as they may seriously impact the subsequent analyses.

On the initial processing page, you have to specify how transcripts should be assigned to gene families and how should functional annotation be performed. For this protocol, you can select the settings shown in figure 4. Finish by clicking the *Start transcriptome pipeline* button. While an experiment is being processed, it cannot be accessed: the experiment will be available again once the initial processing finished.

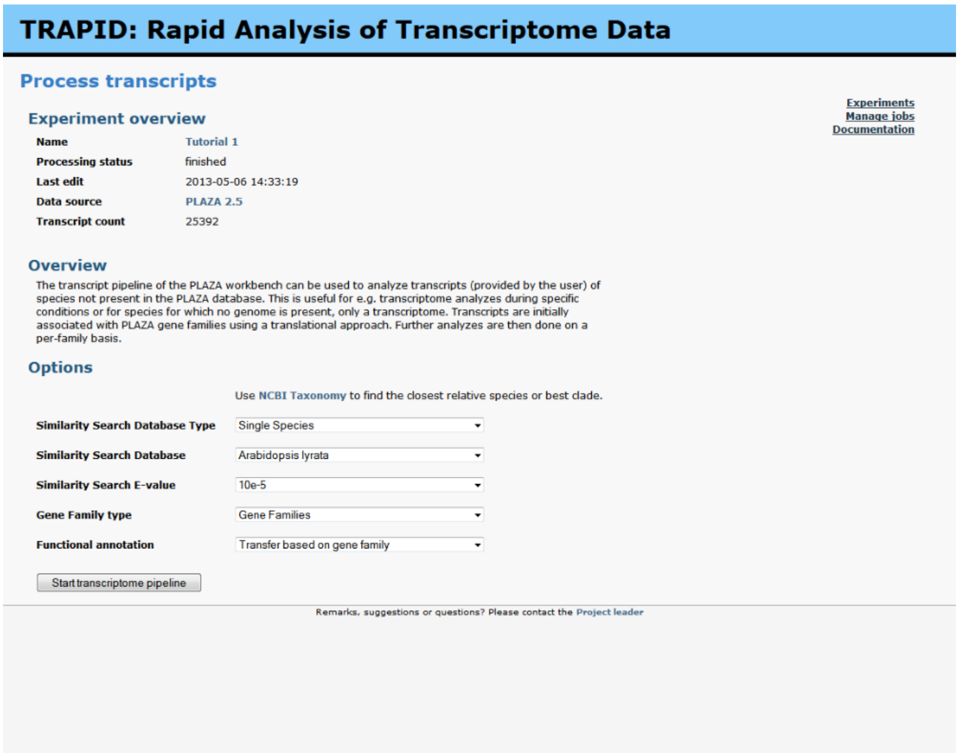
Initial processing options

Several settings can be adjusted for initial processing:

1. **Similarity search database type:** whether either a single species, a phylogenetic clade or the gene family representatives will be used for the similarity search (first step of TRAPID's initial processing). A single species is a good choice if in the reference database a close relative of the transcriptome species is present. If a good encompassing phylogenetic clade is available,

then this is also a solid choice. If none of the above, then the gene family representatives will provide a good sample distribution of the gene content within each reference database.

2. **Similarity search database:** depending on the choice made for the similarity search database type, this list offers you choices of database to select.
3. **Similarity search e-value:** set an E-value cutoff for the RapSearch2 similarity search.
4. **Gene family type:** define the Gene Family Type. For PLAZA 2.5 this is Gene families (TribeMCL clusters) or Integrative Orthology (only available if a single species was selected as database type), for OrthoMCL-DB this is Gene families (OrthoMCL clusters).
5. **Functional annotation:** define how the functional annotation should be transferred from the family to the transcript level. In general, *transfer based on gene family* is the most conservative approach while *transfer based on best hit* is yielding a larger number of functionally annotated transcripts. Logically, combining both methods using *transfer from both GF and best hit* yields the largest fraction of annotated transcripts.



TRAPID: Rapid Analysis of Transcriptome Data

Process transcripts

Experiment overview

Name	Tutorial 1
Processing status	finished
Last edit	2013-05-06 14:33:19
Data source	PLAZA 2.5
Transcript count	25392

Overview

The transcript pipeline of the PLAZA workbench can be used to analyze transcripts (provided by the user) of species not present in the PLAZA database. This is useful for e.g. transcriptome analyzes during specific conditions or for species for which no genome is present, only a transcriptome. Transcripts are initially associated with PLAZA gene families using a translational approach. Further analyzes are then done on a per-family basis.

Options

Use NCBI Taxonomy to find the closest relative species or best clade.

Similarity Search Database Type	Single Species
Similarity Search Database	Arabidopsis lyrata
Similarity Search E-value	10e-5
Gene Family type	Gene Families
Functional annotation	Transfer based on gene family

Remarks, suggestions or questions? Please contact the Project leader

Figure 4: initial processing page. On this page, you can select multiple options regarding how the sequences should be added to gene families and how to perform funtional annotation.

Once the initial processing is done, all sequences will be included in gene families and will be functionally annotated, if possible. Additionally, as transcript data often includes truncated sequences or sequences with indels, problematic sequences are flagged, for example if they contain a putative frameshift.

You will a receive an email when the processing is finished.

NOTES

Klaas Vandepoele 27 Feb 2018

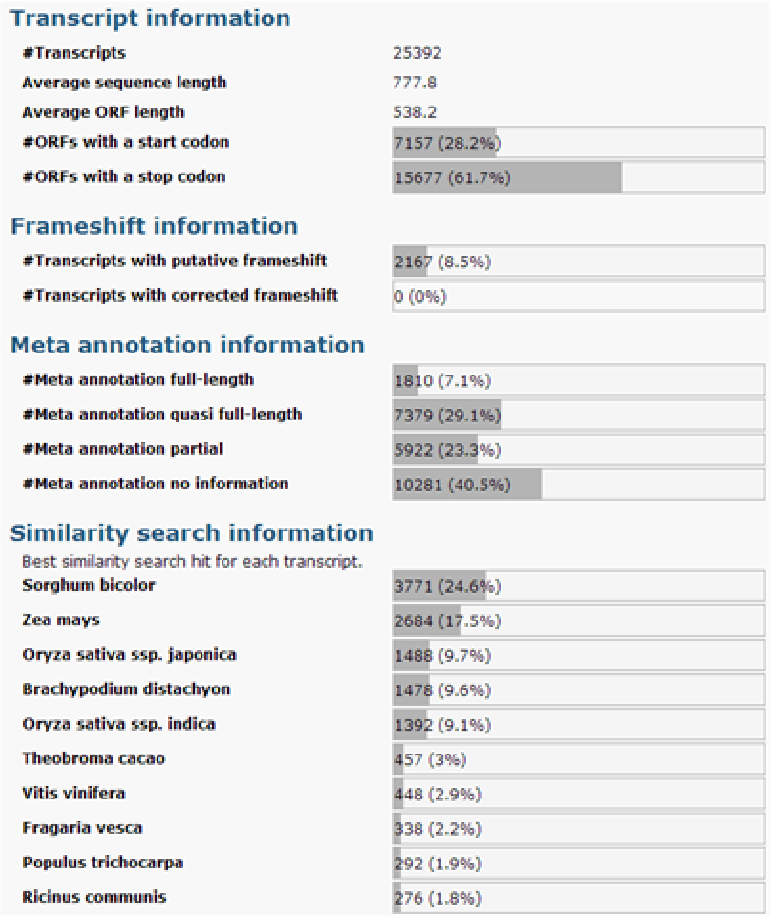
Initial processing is rather fast. A test data set containing 90,000 transcripts can be processed in less than 3 hours, with approximately 28% of these transcripts assigned to gene families. For our example *Panicum* data set (25,392 transcripts), the complete processing, including upload and transcript processing (using *Monocots* clade) takes around 1 hour (and yields 60% of transcripts assigned to gene families). Note that the fraction of very short or very long transcripts will impact the total processing time during this initial phase.

Exploring TRAPID output: basic analyses

Step 5.

General statistics

From the experiment page, use the *General statistics* link in the toolbox to access the general statistics page (example shown in figure 5). This page offers a complete overview of ORF finding, gene family assignments, similarity search species information, meta-annotation and functional information.



Transcript information

#Transcripts	25392
Average sequence length	777.8
Average ORF length	538.2
#ORFs with a start codon	7157 (28.2%)
#ORFs with a stop codon	15677 (61.7%)

Frameshift information

#Transcripts with putative frameshift	2167 (8.5%)
#Transcripts with corrected frameshift	0 (0%)

Meta annotation information

#Meta annotation full-length	1810 (7.1%)
#Meta annotation quasi full-length	7379 (29.1%)
#Meta annotation partial	5922 (23.3%)
#Meta annotation no information	10281 (40.5%)

Similarity search information

Best similarity search hit for each transcript.

Sorghum bicolor	3771 (24.6%)
Zea mays	2684 (17.5%)
Oryza sativa ssp. japonica	1488 (9.7%)
Brachypodium distachyon	1478 (9.6%)
Oryza sativa ssp. indica	1392 (9.1%)
Theobroma cacao	457 (3%)
Vitis vinifera	448 (2.9%)
Fragaria vesca	338 (2.2%)
Populus trichocarpa	292 (1.9%)
Ricinus communis	276 (1.8%)

Figure 5: general statistics page.

Exploring TRAPID output: basic analyses

Step 6.

Subsets and labels

If the dataset is comprised of transcriptome data from different sources (corresponding for example to different tissues of a multicellular organism, or different stress conditions), then it is possible to assign labels to the subsets.

By doing so, it becomes easy to browse the function of a subset of transcripts, and several new analyses become available: comparison of functional annotation, and functional enrichment analysis. The latter is covered with more details in our other protocol (see section 10).

To define a subset, click the *Import transcript labels/Import data* link in the *import/export data* section of the experiment page. You then have to provide, as file, a list of transcripts that should be form the new transcript subset. Note that a transcipt can be in multiple subsets.

You could for example define a new subset using a set of Cell Cycle genes for our *Panicum* dataset. To do so, hit B*rowse* to select the file, enter a label name (e.g. *Cell_cycle*) and click *Import labels*.

Exploring TRAPID output: basic analyses

Step 7.

Searching for data

From the experiment page, use the *Search* box to search for a number of possible data types within the current experiment.

Relevant families can be found using the search function, for example by looking for a protein domain / GO term of interest and browsing the list of associated sequences.

> NOTES
>
> **Klaas Vandepoele** 27 Feb 2018
>
> Functional annotation can be searched for both through direct term identifiers (e.g. 'GO:0005509') or through the descriptions (e.g. 'Calcium ion binding').

Exploring TRAPID output: basic analyses

Step 8.

Exporting data

TRAPID enables users to export both the original data and the annotated/processed data of an experiment. To access the export page, click the *Export data* link in the *import/export data* section of the experiment page.

It is possible to export:

- Structural data: ORF information, transcript/ORF/protein sequences
- Gene family data: transcripts with their associated gene family, gene families with their associated transcript, gene content of the reference gene families
- Functional data: GO & InterPro information (annotation and meta-data).
- Transcript subsets, in case you defined any (list of identifiers)

Data export is performed in two steps: first, click on the button corresponding to the data type you want to export (for example *transcripts with GF*), then click on the generated download link (*download file*). This two-step implementation is due to the fact that files can take a while to be generated.

NOTES

Klaas Vandepoele 27 Feb 2018

The export of the functional GO information has an extra column 'is_hidden', indicating whether a GO term is flagged as hidden, due to the presence of more informative GO terms in the GO graph for the given transcript.

Exploring TRAPID output: basic analyses

Step 9.

The toolbox

On most pages (experiment/transcript/gene family/GO/protein domain), a toolbox is available which contains the most common analyses to be performed on the given data object.

For example, click on a gene family in the transcript table of the experiment page. From there, you would be able to create multiple sequence alignments or phylogenetic trees, in the gene family context, for transcripts associated to this gene family, using the options available in the toolbox.

Going beyond

Step 10.

Although this protocol provides a good introduction to TRAPID and its basic functionality, it is possible to conduct more advanced functional and comparative analyses on the platform. Please read our other protocol, 'In-depth functional and comparative analyses of transcriptomes with TRAPID', for more details.

For a comprehensive overview of TRAPID capabilities, please have a look at TRAPID's documentation.