
TBro Demo Documentation

Release 0.9

Markus Ankenbrand

April 21, 2016

1	Cannabis sativa transcriptome	3
1.1	Installation	3
1.2	Introduction	4
1.3	Data	4
1.4	Bringing the data into TBro	5
1.5	Exploring the imported Data	14

This tutorial leads you through the process of installing and loading transcriptomic data into TBro as well as analyzing it. If you read this on readthedocs.org, please be aware that [Google Analytics](#) is used.

Cannabis sativa transcriptome

Original data published in:

Harm van Bakel et al. "The draft genome and transcriptome of Cannabis sativa." In: Genome biology 12.10 (Jan. 2011), R102. issn: 1465-6914. doi: 10.1186/gb-2011-12-10-r102. url: <http://genomebiology.com/2011/12/10/R102>.

1.1 Installation

Installation of TBro is straightforward and easy using preconfigured [docker](#) containers. See [docker documentation](#) on how to install docker on your machine.

After installation of docker execute the following commands to pull the docker images:

```
docker pull greatfireball/generic_postgresql_db
docker pull tbroteam/generic_chado_db_reload
docker pull tbroteam/tbro_worker_ftp
docker pull tbroteam/tbro_worker
docker pull tbroteam/tbro_apache
docker pull tbroteam/tbro_demo
```

Now start the Chado database container and install the schema:

```
docker run -d -e DB_NAME=chado -e DB_USER=tbro -e DB_PW=tbro --name "Chado_DB_4_TBro_official" greatfireball/generic_postgresql_db
sleep 60
docker run --rm -i -t --link Chado_DB_4_TBro_official:CHADO --name "Chado_DB_4_TBro_load_official" tbroteam/generic_chado_db_reload
```

Now start the database container for the BLAST worker:

```
docker run -d -e DB_NAME=worker -e DB_USER=worker -e DB_PW=worker --name "Worker_DB_4_TBro_official" tbroteam/generic_postgresql_db
```

Start an ftp server to host the BLAST databases:

```
docker run -d --name "Worker_FTP_4_TBro_official" -e FTP_USER="tbro" -e FTP_PW="ftp" tbroteam/tbro_worker_ftp
```

Start a worker to execute the BLAST jobs:

```
docker run -d --link Worker_DB_4_TBro_official:WORKER --link Worker_FTP_4_TBro_official:WORKERFTP --name "TBro_Worker_official" tbroteam/tbro_worker
docker exec -i -t TBro_Worker_official /home/tbro/worker_build_installation.sh 2> run_worker_build_installation.err
```

Finally start and install the main TBro container:

```
docker run -d --link Chado_DB_4_TBro_official:CHADO --link Worker_FTP_4_TBro_official:WORKERFTP --name "TBro_official" tbroteam/tbro_demo
docker exec -i -t TBro_official /home/tbro/build_installation.sh 2> tbro_build_installation.err
```

You can now access the TBro web interface by pointing your browser to <http://localhost> However there is no data loaded, yet. To load data you can either perform the automatic demo installation (see next section), follow the step-by-step tutorial (next chapter) or load your own data.

1.1.1 Preload demo data

Run the following command to fill your TBro instance with demo data from *Cannabis sativa*.

```
docker run --rm -i -t --link Worker_DB_4_TBro_official:WORKER --link Worker_FTP_4_TBro_official:WORKER
```

Congratulations, you have a full-featured TBro instance up and running.

1.2 Introduction

The aim of this manual is to teach you how to use the TBro. At first you will learn how to bring the data into the TBro. For this purpose you can follow the step by step guide using example data. When you work through this tutorial you have two options. You can either start from scratch, load the data from the original sources and perform the analyses yourself or you can just use the precomputed results in the example directory. In the manual it is assumed that you perform the analyses yourself so all commands that created the data are included. However the scope of this manual is not to teach you how to perform transcriptomic analyses and some of the methods may soon be out of date. The methods of importing and interpreting the data however will remain the same. If you choose to use the precomputed data just ignore all commands regarding their creation and use the accordingly named files in the example directory.

1.3 Data

The demo data consists of the published transcriptome of . And additional short read libraries of different samples.

1.3.1 Cannabis sativa

The raw data is available through NCBI. The transcriptome is described in @CSATIVA and has accession numbers (JP449145 - JP482359). Read data from different samples is available through the SRA at NCBI. The used samples are:

- Mature flower (SRR306868, SRR306869, SRR306870)
- Mature leaf (SRR306875, SRR306885, SRR306886)
- Entire root (SRR306861, SRR306862, SRR306863)

We analyzed the data in different ways and final results for each analysis are included in the example data. You can use this pregenerated results or generate it yourself, for this purpose each command for data creation along with program version is given. For an overview of the used programs and versions see table [tab:PROGRAMS].

[tab:PROGRAMS]

Task	Program	Version
Peptide prediction	Transdecoder	r2012-08-15
Repeat annotation	RepeatMasker	4.0.3
General annotations	InterproScan	5 RC5
GO/EC annotation	Blast2GO (B2G4Pipe)	2.5.0
General annotation	MapMan (Mercator)	2013/07/18
Count quantification	RSEM	1.2.5
Normalization / Differential expression	DESeq	1.12.1

Data creation and import into TBro is described as a workflow in the next section.

1.4 Bringing the data into TBro

For this tutorial it is assumed, that you have a fresh install of TBro as described in the manual @DIP_TBRO. So you start with a CHADO database with just a few slight modifications. So lets start populating the database with our first transcriptome.

1.4.1 The Cannabis sativa transcriptome

Preparation

There is one preparation necessary before we can start importing data into TBro. We have to create the organism as is not part of the default CHADO repertoire.

```
tbro-db organism insert --genus Cannabis --species sativa\
  --common_name Weed --abbreviation C.sativa

tbro-db organism list
```

The last command shows us all organisms known to TBro. As we see the ID of is 13. We will need this ID for later commands. You have probably noticed that it is possible to use autocompletion for commands, subcommands and parameters. In addition the very useful `-help` option is present for every command and gives information on usage, and available parameters.

The Transcripts

Create a directory `cannabis_sativa` with subdirectory `transcriptome`

```
mkdir -p cannabis_sativa/transcriptome
cd cannabis_sativa/transcriptome
```

Download the transcriptome from NCBI. To do so, search for 74271[BioProject] on <http://www.ncbi.nlm.nih.gov/nuccore> and download all hits as fasta. There should be sequences. Save those to the file `cannabis_sativa_transcriptome.fasta` in the newly created folder. As we have no isoform unigene relationship for those transcripts we treat all of them as separate isoforms. So the id file comes down to a simple list of all ids in the fasta.

```
grep ">" cannabis_sativa_transcriptome.fasta\
| perl -pe 's/>(\S+).*/$1/'\
>cannabis_sativa_transcriptome.ids
```

Now, it's time to import the sequence IDs into TBro. As we have no isoform - unigene relationship we import each transcript as a single isoform:

```
tbro-import sequence_ids --organism_id 13 --release 1.CasaPuKu\
  --file_type only_isoforms cannabis_sativa_transcriptome.ids
```

We had to pass the previously given organism-id and a release name.

The release name can be selected freely and the release is automatically created upon first usage. The file-type was set to `only_isoforms` as we have no unigenes. Other possible values are `only_unigenes` and `map`. The last thing we pass is the path to the file containing the sequence-ids which we have just created. | TBro now knows about the sequence ids so lets feed it with the associated sequences:

```
tbro-import sequences_fasta --organism_id 13\  
--release 1.CasaPuKu cannabis_sativa_transcriptome.fasta
```

Now it's time to start up your browser and visit your TBro instance. Use the quick search field in the upper right corner to find `gi|351628922|gb|JP481805.1|`. You will see the isoform page with the basic information about this transcript. By now there is just the general info (date of import, organism, release) and the sequence together with a visualization as a horizontal bar. You can check back to this page after every successful import to watch how the new features are presentet. Of course you can choose any other isoform that is of interest to you.

Predicted Peptides

After we have the nucleotide sequences, the next step is to predict peptides and load this info into TBro. There are many tools available to predict peptides, we chose `Transdecoder` but the TBro does not restrict you to a certain tool.

```
mkdir -p ../peptids  
cd ../peptids  
  
transcripts_to_best_scoring_ORFs.pl -t \  
../transcriptome/cannabis_sativa_transcriptome.fasta\  
-m 30 -v --CPU 4 >log >error.log
```

Note that we have set the minimum protein length to 30 and number of threads to 4, you can adjust those parameters to your own requirements. Unfortunately the output format for predicted peptides is not standardized. To make the peptide import generic and not rely on the output format of a special program the import into TBro is split into two steps. First a list of peptides is imported. This list has to be in tab delimited format and contain the following columns:

1. peptide id
2. isoform id
3. start position
4. end position
5. strand (+/-)

This file can easily be created from the output of every peptide prediction program. TBro contains a tool to get the table from the `Transdecoder` output so lets use that:

```
tbro-tools transToProt -o predicted_peptides.tbl\  
best_candidates.eclipsed_orfs_removed.pep
```

Lets have a look to see that the table has the desired format.

```
head -n5 predicted_peptides.tbl  
m.243266      gi|351590686|gb|JP449145.1|      165      893      +  
m.243259      gi|351590687|gb|JP449146.1|      1751     1894     +  
m.243253      gi|351590687|gb|JP449146.1|      2        1684     +  
m.243237      gi|351590688|gb|JP449147.1|      1        1986     +  
m.243247      gi|351590688|gb|JP449147.1|      2173     2295     +
```

Now to import this peptide table issue the following command:

```
tbro-import peptide_ids --organism_id 13\
--release 1.CasaPuKu predicted_peptides.tbl
```

TBro now knows about the predicted peptides and their locations. What's missing is the sequences. They are added the same way as the nucleotide sequences of the transcripts before. It is important, that the fasta IDs exactly match the IDs in the first column of the peptide table.

```
tbro-import sequences_fasta --organism_id 13\
--release 1.CasaPuKu best_candidates.eclipsed_orfs_removed.pep
```

You might want to check back to the web interface to see our newly imported peptides.

RepeatMasker

Another basic type of annotation are repeats. So we create repeat annotations using RepeatMasker and import them.

```
mkdir -p ../analyses/repeats
cd ../analyses/repeats

RepeatMasker -pa 4 -dir . -xm -species viridiplantae\
../../transcriptome/cannabis_sativa_transcriptome.fasta
```

All we need to do now is tell TBro to import the generated file as RepeatMasker annotations:

```
tbro-import annotation_repeatmasker --organism_id 13\
--release 1.CasaPuKu cannabis_sativa_transcriptome.fasta.out.xml
```

Interpro

Interpro is a usefull tool to annotate protein sequences with information from different databases. There exists a command line version of this tool called InterproScan. We use this tool to generate the interpro annotations for our transcriptome:

```
mkdir -p ../interpro
cd ../interpro

interproscan.sh --pa --iprlookup --goterms --fasta\
../../peptids/best_candidates.eclipsed_orfs_removed.pep\
--output-file-base interpro >interpro.log
```

The results in the tsv format can be importet into TBro with the following command:

```
tbro-import annotation_interpro --organism_id 13\
--release 1.CasaPuKu -i interproscan-5-RC5 interpro.tsv
```

Note that it is important to know the InterproScan version used as each version uses different versions of the underlying databases. Interpretation of the results requires knowledge of this versions so the `-i` switch taking the version is required for this import.

Blast2GO

Blast2GO uses BLAST to find sequence similarities to annotated sequences. The hits are then used to assign GO terms and EC numbers to the input sequences.

```
mkdir -p ../blast2go
cd ../blast2go

blastx -query ../../transcriptome/cannabis_sativa_transcriptome.fasta\
  -db /path/to/databases/NCBI/nr -evalue 1e-3 -outfmt 5\
  -num_alignments 250 -num_descriptions 250\
  -out cannabis_sativa_transcriptome.nr.xml\
  2> cannabis_sativa_transcriptome.nr.log

java -Xmx20G -cp *:ext/*: es.blast2go.prog.B2GAnnotPipe\
  -in cannabis_sativa_transcriptome.nr.xml\
  -out cannabis_sativa_transcriptome.blast2go.annot\
  -prop b2gPipe.properties -v -annot -dat -img\
  > cannabis_sativa_transcriptome.blast2go.log
```

First all sequences are blasted against a local copy of nr. The output format is set to 5 (xml output). The e-value cutoff was set to 10^{-3} . Afterwards the BLAST output is passed to the Blast2GO annotation pipeline. We can extract three different kinds of annotations from the Blast2GO output:

GO

Gene Ontology

```
grep "GO:" cannabis_sativa_transcriptome.blast2go.annot\
  >cannabis_sativa_transcriptome.blast2go.annot.go

tbro-import annotation_go --organism_id 13\
  --release 1.CasaPuKu cannabis_sativa_transcriptome.blast2go.annot.go
```

The lines containing “GO:” are selected and imported into TBro as `annotation_go`

EC

Enzyme Commission

```
grep "EC:" cannabis_sativa_transcriptome.blast2go.annot\
  >cannabis_sativa_transcriptome.blast2go.annot.ec

tbro-import annotation_ec --organism_id 13\
  --release 1.CasaPuKu cannabis_sativa_transcriptome.blast2go.annot.ec
```

The lines containing “EC:” are selected and imported into TBro as `annotation_ec`

Description

```
perl -ane 'print if(@F>2)'\
  cannabis_sativa_transcriptome.blast2go.annot.go\
  >cannabis_sativa_transcriptome.blast2go.annot.go.description

tbro-import annotation_description --organism_id 13 --release 1.CasaPuKu\
  cannabis_sativa_transcriptome.blast2go.annot.go.description
```

Descriptions are arbitrary text that describes a transcript. Some GO Terms contain a meaningful description so we import the lines containing such a description into TBro. However this is just an example, the source of the description

does not matter. The format is a tab delimited format with the feature ID in the first column and the description in the second.

Mercator

Mercator is a tool to classify sequences into MapMan functional plant categories.

```
mkdir -p ../mercator
cd ../mercator
```

To perform the Mercator classification start up your browser and go to

<http://mapman.gabipd.org/web/guest/mercator>. In the web interface you can upload the `cannabis_sativa_transcriptome.fasta`. Unfortunately, there is a restriction on the input file size. This limit is exceeded by our transcriptome. So you can either contact the people at MapMan to allow you the submission of a larger dataset or just split the input file into two parts. We split the file by sequence length but you can just as well open the file in a text editor and split it. Then run Mercator on each chunk and download the results afterwards. It is no problem to import the two reports, one after the other:

```
tbro-import annotation_mapman --organism_id 13\
--release 1.CasaPuKu mercator.results_max1499.txt

tbro-import annotation_mapman --organism_id 13\
--release 1.CasaPuKu mercator.results_min1500.txt
```

Expression Counts

Now we have all kinds of annotation for each transcript in the TBro so we can start with the fun part. Expression data and differential expression data in particular are the main prospects why we perform RNASeq experiments. So go ahead and download the SRA files listed above.

```
mkdir -p ../../samples
cd ../../samples

/path/to/sratoolkit/bin/fastq-dump *
```

In the samples directory we now have a `.fq` file for each downloaded `.sra` file. The SRA files are no longer required so you can delete them to save some space. The next step is the quantification by mapping the reads onto the transcripts. This quantification is done separately for each sample in the for loop:

```
rsem-prepare-reference cannabis_sativa_transcriptome.fasta\
cannabis_sativa_transcriptome

for SAMPLE in *.fq
do
BASE=$(basename $SAMPLE .fq)
rsem-calculate-expression -p 4 $SAMPLE cannabis_sativa_transcriptome\
$BASE >$BASE.log 2>$BASE.err
done
```

The results for each sample are aggregated into a single large table with the perl script `aggregator_Count.pl`.

```
perl aggregator_CountMat.pl --in_RSEM\
SRR306868.isoforms.results SRR306869.isoforms.results\
SRR306870.isoforms.results SRR306875.isoforms.results\
SRR306885.isoforms.results SRR306886.isoforms.results\
SRR306861.isoforms.results SRR306862.isoforms.results\
SRR306863.isoforms.results\
--labels_RSEM Flower.mature_L1 Flower.mature_L2 Flower.mature_L3\
Leaf.mature_L1 Leaf.mature_L2 Leaf.mature_L3\
Root.entire_L1 Root.entire_L2 Root.entire_L3\
--out rsem_aggregated.mat
```

The resulting table could be imported into TBro as it is. However the data is not normalized yet. You should always(!) normalize your expression data. One way to do that is using the DESeq R-package provided by Bioconductor. So fire up R and install DESeq if you don't already have it. As we use DESeq also to create the differential expression data we will already do that and use the results in the next section.

```
# installing and loading DESeq
source("http://bioconductor.org/biocLite.R")
biocLite("DESeq")
library(DESeq)
# loading the expression data
cmat <- read.table(file="rsem_aggregated.mat", row.names=1, header=T)
cond <- sub("_L.*", "", colnames(cmat))
# TMM normalization
cds <- newCountDataSet(round(cmat), cond)
cds <- estimateSizeFactors(cds)
write.table(file="rsem_aggregated_TMM.mat", counts(cds, normalized=T),
  quote=F, sep="\t")

# differential expressions
cds <- estimateDispersions(cds)
res.FvsR <- na.omit(nbinomTest(cds, "Flower", "Root"))
res.FvsL <- na.omit(nbinomTest(cds, "Flower", "Leaf"))
res.RvsL <- na.omit(nbinomTest(cds, "Root", "Leaf"))
write.csv(res.FvsL, file="rsem_aggregated_TMM_diff_FvsL.mat", quote=F)
write.csv(res.FvsR, file="rsem_aggregated_TMM_diff_FvsR.mat", quote=F)
write.csv(res.RvsL, file="rsem_aggregated_TMM_diff_RvsL.mat", quote=F)
```

So now we have the expression counts in the file `rsem_aggregated_TMM.mat`. This file just lacks the header for the first column so we add it with the following command:

```
sed -i '1{s/^/ID\t/}' rsem_aggregated_TMM.mat
```

Before we can go ahead and import the data into TBro we have to make some preparations. Normally RNASeq experiments are performed on biomaterials in different conditions. To differentiate between biological signals and random noise it is mandatory to have replicates for each condition. Each replicate is called a sample. This hierarchical structure of biomaterial → condition → sample is also represented in the TBro. So let's tell TBro about our samples:

```
# Prepare database for Expression Data Import
# Add missing biomaterials (Flower and Root are already present)
tbro-db biomaterial insert --name Flower
tbro-db biomaterial insert --name Leaf
tbro-db biomaterial insert --name Root

# Add conditions
tbro-db biomaterial add_condition --name Flower.mature\
  --parent_biomaterial_name Flower
tbro-db biomaterial add_condition --name Leaf.mature\
  --parent_biomaterial_name Leaf
```

```

tbro-db biomaterial add_condition --name Root.entire\
  --parent_biomaterial_name Root

# Add samples
tbro-db biomaterial add_condition_sample --name Flower.mature_L1\
  --parent_condition_name Flower.mature
tbro-db biomaterial add_condition_sample --name Flower.mature_L2\
  --parent_condition_name Flower.mature
tbro-db biomaterial add_condition_sample --name Flower.mature_L3\
  --parent_condition_name Flower.mature
tbro-db biomaterial add_condition_sample --name Leaf.mature_L1\
  --parent_condition_name Leaf.mature
tbro-db biomaterial add_condition_sample --name Leaf.mature_L2\
  --parent_condition_name Leaf.mature
tbro-db biomaterial add_condition_sample --name Leaf.mature_L3\
  --parent_condition_name Leaf.mature
tbro-db biomaterial add_condition_sample --name Root.entire_L1\
  --parent_condition_name Root.entire
tbro-db biomaterial add_condition_sample --name Root.entire_L2\
  --parent_condition_name Root.entire
tbro-db biomaterial add_condition_sample --name Root.entire_L3\
  --parent_condition_name Root.entire

```

Also the experiments and analyses should be traceable. So we also have to include information about the experiment and the different steps in the analysis. Also the person who performed the analyses has to be specified:

```

# Add contact
tbro-db contact insert --name TBroDemo --description 'TBro Demo User'
# New item ID is 5.

#Add experiments
tbro-db assay insert --name SRX082027 --operator_id 4
# New item ID is 1.

# Add acquisitions (corresponding to experiments)
tbro-db acquisition insert --name SRX082027 --assay_id 1
# New item ID is 1.

# Add analyses
tbro-db analysis insert --name RSEM_TMM --program RSEM\
  --programversion 1.2.5 --sourcename Mapping\
  --description 'RSEM quantification with subsequent TMM normalization'
# New item ID is 50.
tbro-db analysis insert --name DESeq_isoform --program DESeq\
  --programversion 1.12.1 --sourcename Mapping_isoform
# New item ID is 51.

# Add quantifications
tbro-db quantification insert --name RSEM_SRX082027\
  --acquisition_id 1 --analysis_id 50
# New item ID is 1.

```

So we have created a contact, assay, acquisition, quantification and two analyses. Warning: It is important to use the right IDs. Those may differ in your case so carefully watch the output of each command and note the ID given. If you forget an ID you can always have a list of all available entries by issuing:

```
tbro-db <subcommand> list
```

After a lot of groundwork we are finally there. Import the expression counts with this command:

```
tbro-import expressions -o 13 -r 1.CasaPuKu -q 1 -a 50\  
rsem_aggregated_TMM.mat
```

Differential Expression

Differential expression is the comparison of the expressions in two different conditions. When calculating differential expressions statistical methods are applied to correct for the multiple testing problem. We have already performed this analysis in the previous section. So if you have skipped the Expression section you have to use the R snippet there. We also already created the biomaterials, conditions, analyses, etc. Therefore we can go ahead and import the differential expression results:

```
tbro-import differential_expressions -o 13 -r 1.CasaPuKu --analysis_id\  
51 -A Flower.mature -B Leaf.mature rsem_aggregated_TMM_diff_FvsL.mat  
tbro-import differential_expressions -o 13 -r 1.CasaPuKu --analysis_id\  
51 -A Flower.mature -B Root.entire rsem_aggregated_TMM_diff_FvsR.mat  
tbro-import differential_expressions -o 13 -r 1.CasaPuKu --analysis_id\  
51 -A Root.entire -B Leaf.mature rsem_aggregated_TMM_diff_RvsL.mat
```

Blast DB

To search the transcriptome by homology. Let's add a blast database. To do so we create a nucleotide database and a protein database and zip them:

```
makeblastdb -in cannabis_sativa_transcriptome.fasta -dbtype nucl  
makeblastdb -in cannabis_sativa_predpep.fasta -dbtype prot  
zip cannabis_sativa_transcriptome.zip cannabis_sativa_transcriptome.fasta*  
zip cannabis_sativa_predpep.zip cannabis_sativa_predpep.fasta*  
md5sum *.zip  
# b2ab466c7bfb7d41c27a89cf40837fb4 cannabis_sativa_predpep.zip  
# 1f87bbeee5a623e6d2f8cab8f68c9726 cannabis_sativa_transcriptome.zip
```

This zip files should now be moved in a location where it can be reached from the worker machines. To tell TBro about the BLAST databases you should issue the following command in your main TBro directory:

```
phing queue-install-db
```

This will create a file called `queue_config.example.sql` in the current directory. Rename it to `queue_config.sql` and adjust the appropriate sections like this:

```
...  
-- database files available. name is the name it will be referenced by, md5 is the zip file's sum, do  
INSERT INTO database_files  
  (name, md5, download_uri) VALUES  
  ('cannabis_sativa_transcriptome.fasta', '1f87bbeee5a623e6d2f8cab8f68c9726',  
   'http://yourdomain/location/cannabis_sativa_transcriptome.zip'),  
  ('cannabis_sativa_predpep.fasta', 'b2ab466c7bfb7d41c27a89cf40837fb4',  
   'http://yourdomain/location/cannabis_sativa_predpep.zip');  
  
-- contains information which program is available for which program.  
-- additionally, 'availability_filter' can be used to e.g. restrict use for a organism-release combin  
INSERT INTO program_database_relationships  
  (programname, database_name, availability_filter) VALUES  
  ('blastn', 'cannabis_sativa_transcriptome.fasta', '13_1.CasaPuKu'),  
  ('blastp', 'cannabis_sativa_predpep.fasta', '13_1.CasaPuKu'),  
  ('blastx', 'cannabis_sativa_predpep.fasta', '13_1.CasaPuKu'),
```



```
('tblastn','cannabis_sativa_transcriptome.fasta', '13_1.CasaPuKu'),
('tblastx','cannabis_sativa_transcriptome.fasta', '13_1.CasaPuKu');
...
```

You have to specify a location that can be reached by your worker

machine. If you just want to have a single worker on the same machine as the server you can specify the location in the local file system starting with `file://`. To perform the changes run the `queue_config.sql` commands in your queue database. Now TBro knows about the database and shows it in the web interface. To perform BLAST searches we need a worker to execute them. So create one with this command:

```
phing queue-build-worker
unzip unix-worker.zip
```

Modify the `config.php` to your needs. Most values should be preconfigured through your `build.properties`. After that you can start the worker (preferably in a screen):

```
screen -S blastworker
./worker.php config.php
```

Have fun blasting.

Synonym / Publication

Synonyms and publications can be added using the API key and internal name of your bibsonomy account. The structure of such a command is as follows:

```
tbro-db feature add_synonym -f 555 --synonym 'InterestingTranscript'\
-b '[[publication/1adaa3fb03xxxxxxxxxxxxxxxxaec4cef920/bibsonomy_username]]'
-u 'bibsonomy_username' -t symbol -k 34a2149d8xxxxxxxxxxxxxxxxbebd342aa
```

Pathways

To use TBros pathway feature we have to connect the imported data to pathways. As of now this connection is made via EC numbers and KEGG pathways. We have to import two tables containing descriptions for EC and KEGG identifiers in the simple formatL:

```
<Identifier><TAB><Description>
```

Additionally a mapping of which EC occurs in each KEGG pathway is required in the following format:

```
<EC number><TAB><KEGG ID>
```

We collected EC and KEGG information from ENZYME, Interpro and priam to get the descriptions and mapping. The resulting tables may not be complete and up to date so you might wish to create your own tables and mapping. For a quick start you find the three files `ec_info.tab`, `kegg_info.tab` and `ec_kegg_map.tab`

```
tbro-tools addECInformationToDB ec_info.tab
tbro-tools addPathwayInformationToDB kegg_info.tab
tbro-tools addEC2PathwayMapping ec_kegg_map.tab
```

1.4.2 Any other transcriptome

With the description above it should be easy for you to import any transcriptome that interests you. The only thing that could differ significantly from the description above is if you have predicted unigenes for your transcriptome. This is common practice and if you use a de novo transcriptome assembler like *Trinity* you will get unigenes with corresponding isoforms. In this case the main difference is in the first step importing ids. Instead of importing a plain list of sequence IDs you import a map of the following format:

```
<Unigene ID><TAB><Isoform ID>
```

With a separate line for each isoform. The import command would then be:

```
tbro-import sequence_ids --organism_id 14 --release 1.0\
--file_type map my_new_transcriptome.map
```

Of course you have to adjust the `organism_id` and `release` parameters. The use of unigenes brings a number of advantages. You can easily find isoforms that belong to the same unigene as each isoform contains a connection to its parent and on the unigene page you have a list of all corresponding isoforms. In addition you can now load expression and differential expression results on unigene level as well as on isoform level. Many programs like RSEM can readily handle that.

1.5 Exploring the imported Data

1.5.1 Feature annotations

gi|351628922|gb|JP481805.1|

Add to Cart ▼

Imported into TBro	2013-07-12 12:16:31.93198
Release	1.CasaPuku
Organism	Cannabis sativa

Description

abc transporter g family member 11-like

Sequence

GAGACGGCCTAGTTACGGGGGACCACATTCTCTTTTCTCATCAGTCCCTCTCTTGATTCTCAGCCATTAACCAAATATGGCTTCTGTTACTCC
CTCAACTACTATTGACCAAGACCATTATCATAACAAAAAGTTACCATTTACGAACTACTAGTACTACAAGTAGTAATAATAACTAATCATGGA
AACCTAGATAAAGAGTGGTAGAGATTGAAACAGCTAATAATCGCAATGCGAGTTATCGTAGTTGATTTTCGGGGCTGATCATGAAGTTAGTTT
GCCAGTAGTGGAAGATCATCAGGGAGTTTTCTTGACATGGCAGGATTTGTGGGTGACAGTTTCTAGCGGAAAGAACAGTAACAGACCAATTCTT
CGGGGTCTGACTGGCTATGCTAAGCCAGGGGAGCTGTTGGCTATAATGGGTCTTCTGGTTGTGGCAAATCTACTCTCCTTGATGCATTGGCA
GGCAGATTGAATGCAACACAAGACAAAGTGGGATATATTGATAAATGGAATAAACAGACTCTAGCATATGGAACATCGGCATACGTAACACA

Fig. 1.1: Basic information

[fig:basic]

[fig:annotation]

[fig:predpep]

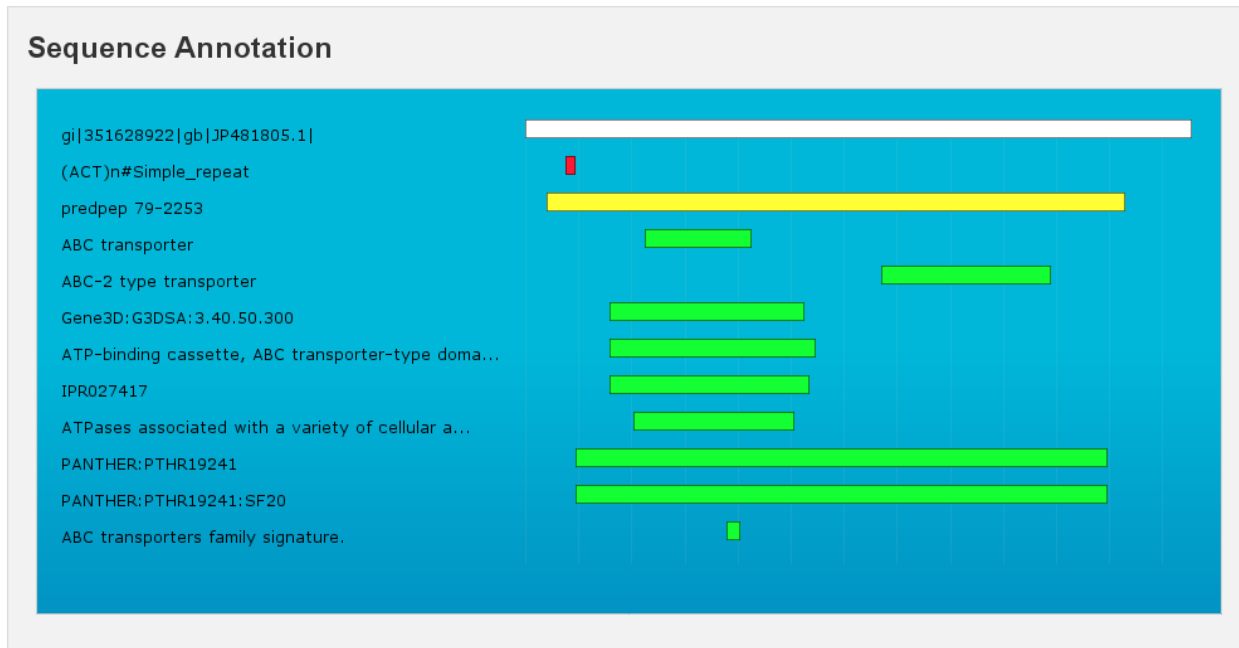


Fig. 1.2: Sequence Annotation

[fig:go]

[fig:ec]

[fig:mapman]

[fig:repeatmasker]

1.5.2 Expressions

[fig:barplot:sub:iso]

[fig:barplot:sub:cart]

1.5.3 Differential Expressions

[fig:diffexp]

[fig:diffexp:sub:results]

1.5.4 Carts

[fig:carts]

1.5.5 Pathways

[fig:pathway]

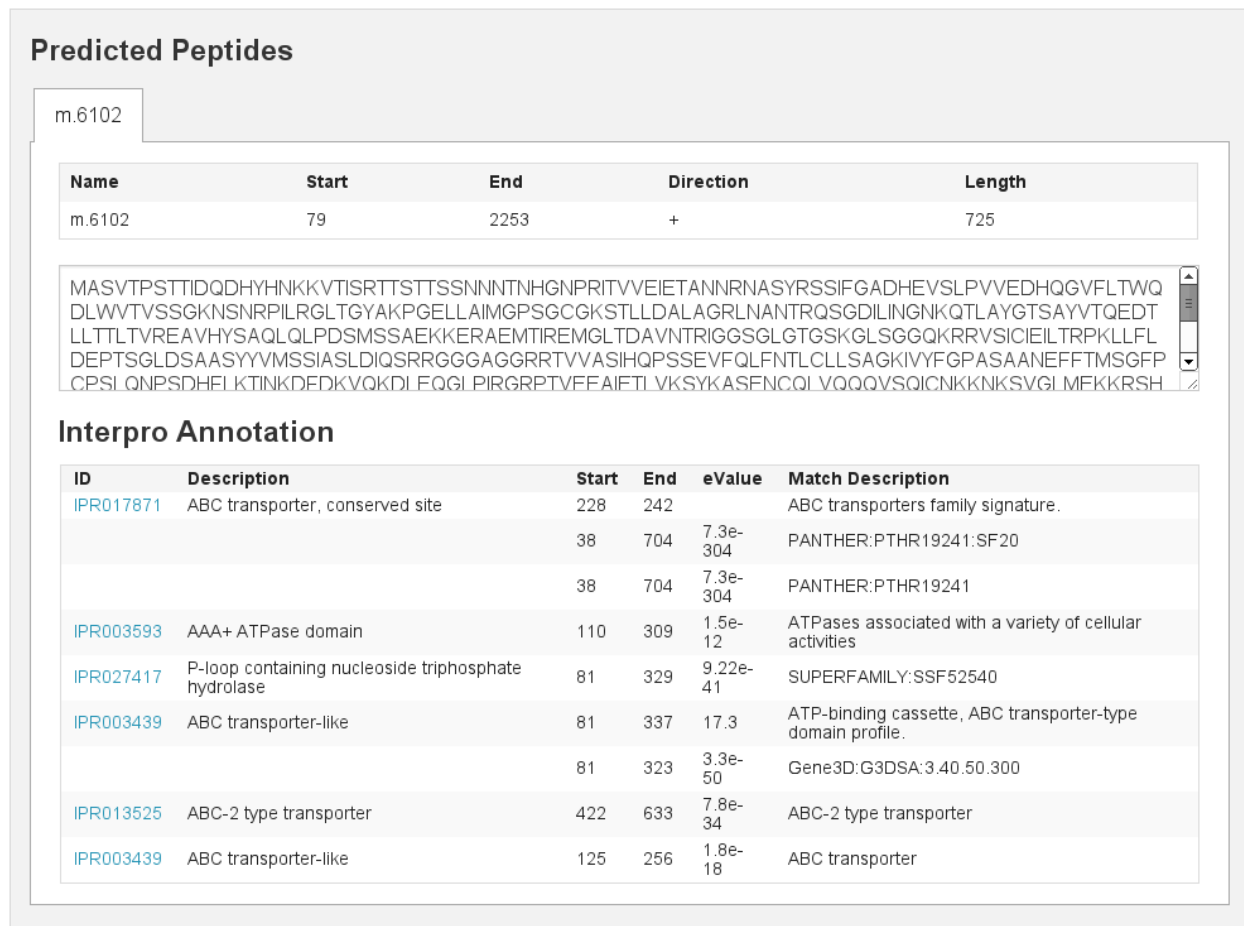


Fig. 1.3: Predicted peptides



Fig. 1.4: GO

Enzyme Commission

EC:3.6.1.3 Adenosinetriphosphatase.

Associated Pathways

00230Purine metabolism

Fig. 1.5: EC

MapMan Annotation

Annotation: highly similar to (630) AT1G17840 | Symbols: WBC11, ABCG11, DSO, COF1, ATWBC11 | WBC11 (WHITE-BROWN COMPLEX HOMOLOG PROTEIN 11); ATPase, coupled to transmembrane movement of substances / fatty acid transporter | chr1:6142870-6145894 FORWARDweakly similar to (163) PDR4_ORYSA Pleiotropic drug resistance protein 4 - Oryza sativa (Rice) TSA: Cannabis sativa PK09255.1_1.CasaPuKu mRNA sequence

Bincode: 34.16: transport.ABC transporters and multidrug resistance systems

Fig. 1.6: Mapman Mercator

Repeatmasker Annotation

Name	Class	Family	Start	End	Direction	Length
(ACT)n	Simple_repeat		151	184	+	34

Fig. 1.7: RepeatMasker

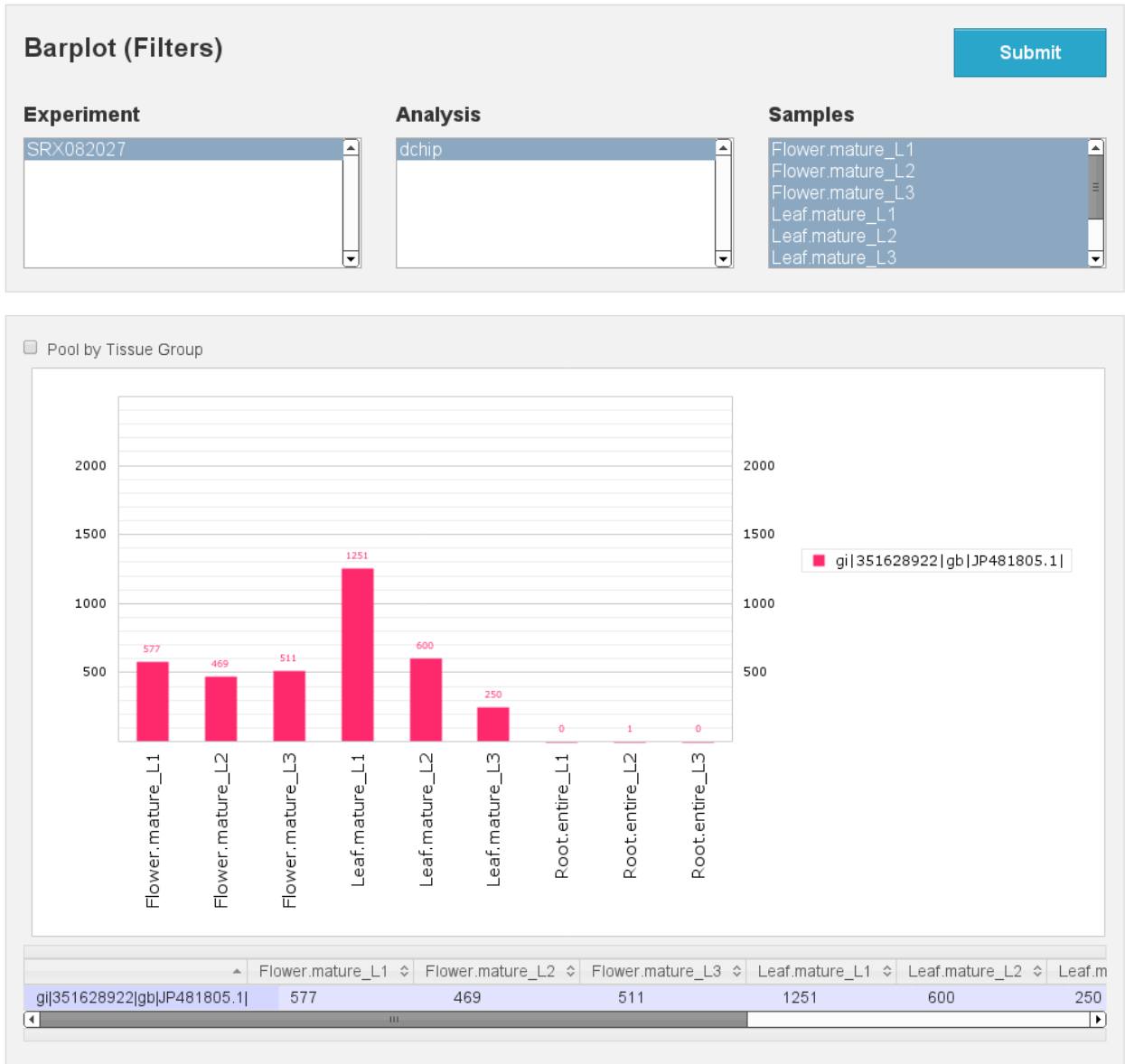


Fig. 1.8: Expression Barplot for a single Isoform

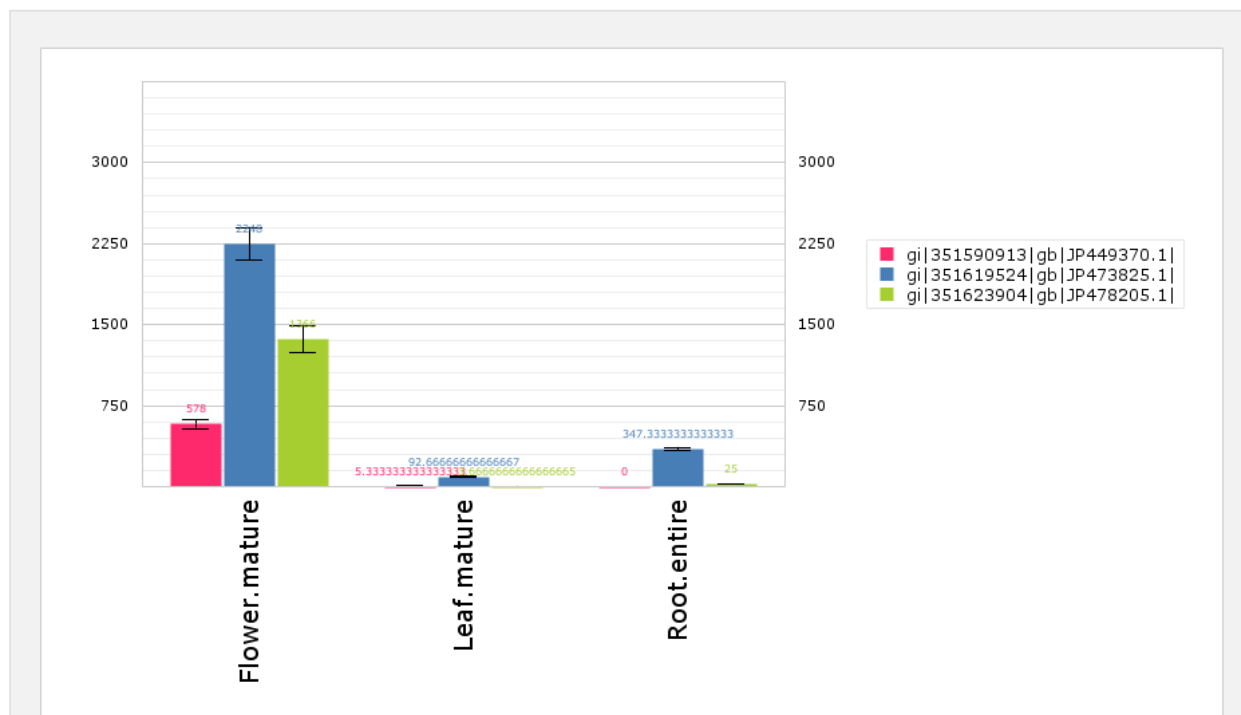


Fig. 1.9: Expression Barplot for all isoforms in a cart

1.5.6 Searches

[fig:combisearch]

[fig:multisearch]

1.5.7 Blast

[fig:blast]

[fig:blast:sub:results₁]

[fig:blast:sub:results₂]

{#section .unnumbered}

Experiment

SRX082027

Condition A

Flower.mature

Root.entire

Leaf.mature

Condition B

Flower.mature

Root.entire

Analysis

RSEM_TMM

Filters

baseMean

<

baseMeanA

<

baseMeanB

<

foldChange

<

log2foldChange

<

pval

<

pvaladj

<

Apply

Fig. 1.10: Differential expression

Filters

baseMean	<	
baseMeanA	<	
baseMeanB	<	
foldChange	<	
log2foldChange	<	
pval	<	
pvaladj	<	1e-30

Apply

Results Overview

Condition A	Flower.mature
Condition B	Leaf.mature
Analysis	RSEM_TMM
Organism	Cannabis sativa
Release	1.CasaPuku
Hits	21

Detailed Results

Show Entries	Show Columns	Select	Store	Export	
Sequence ID	Alias	baseMean	foldChange	log2foldChange	pvaladj
gl 351598622 gb JP453137.1		1.08550e+3	6.06301e-4	-1.06877e+1	4.19681e-72
gl 351616445 gb JP470746.1		7.72357e+2	4.87666e-3	-7.67989e+0	2.47351e-60
gl 351622338 gb JP476639.1		6.41593e+2	5.35890e-3	-7.54385e+0	9.12359e-50
gl 351629444 gb JP482327.1		8.68614e+2	1.01758e-2	-6.61872e+0	4.90196e-47
gl 351603594 gb JP457951.1		3.15386e+2	0	-Inf	5.20999e-47
gl 351609064 gb JP463365.1		1.02134e+3	5.34973e+1	5.74140e+0	1.06435e-44
gl 351609028 gb JP463329.1		2.45737e+2	9.84630e+1	6.62151e+0	1.97420e-41
gl 351606046 gb JP460383.1		4.04961e+2	9.29569e-3	-6.74922e+0	1.68781e-39
gl 351597377 gb JP451924.1		2.19363e+2	0	-Inf	2.54831e-39
gl 351609763 gb JP464064.1		3.72631e+3	3.21577e+1	5.00709e+0	3.21013e-38

Showing 1 to 10 of 21 entries

First Previous 1 2 3 Next Last




Fig. 1.11: Differential expression results Flower vs Leaf




Login with Google




Carts




NewImport




▼ Flower_vs_Leaf (21)Actions




gi|351597377|gb|JP451924.1|




gi|351603594|gb|JP457951.1|




gi|351598622|gb|JP453137.1|




gi|351623904|gb|JP478205.1|




gi|351616445|gb|JP470746.1|




gi|351622338|gb|JP476639.1|




gi|351606046|gb|JP460383.1|




gi|351590913|gb|JP449370.1|




gi|351629444|gb|JP482327.1|




gi|351591379|gb|JP449826.1|

gi|351602165|gb|JP456596.1|

gi|351619524|gb|JP473825.1|

gi|351621130|gb|JP475431.1|

gi|351598314|gb|JP452839.1|

gi|351614393|gb|JP468694.1|

There are 6 more items (go to cart)

► Example (1)Actions

Overview	Expression Analysis	Diff. Expression Analysis	Pathway Analysis																																																																
<table> <tr> <th>Pathway</th><th>Map</th><th>Components</th><th>Details</th></tr> <tr> <td>Metabolic pathways</td><td>01100</td><td>4</td><td>Details</td></tr> <tr> <td>Starch and sucrose metabolism</td><td>00500</td><td>2</td><td>Details</td></tr> <tr> <td>Alanine, aspartate and glutamate metabolism</td><td>00250</td><td>1</td><td>Details</td></tr> <tr> <td>Ascorbate and aldarate metabolism</td><td>00053</td><td>1</td><td>Details</td></tr> <tr> <td>Biosynthesis of phenylpropanoids</td><td>01061</td><td>1</td><td>Details</td></tr> <tr> <td>Biosynthesis of secondary metabolites</td><td>01110</td><td>1</td><td>Details</td></tr> <tr> <td>Glyoxylate and dicarboxylate metabolism</td><td>00630</td><td>1</td><td>Details</td></tr> <tr> <td>Methane metabolism</td><td>00680</td><td>1</td><td>Details</td></tr> <tr> <td>Microbial metabolism in diverse environments</td><td>01120</td><td>1</td><td>Details</td></tr> <tr> <td>Pentose and glucuronate interconversions</td><td>00040</td><td>1</td><td>Details</td></tr> <tr> <td>Phenylalanine metabolism</td><td>00360</td><td>1</td><td>Details</td></tr> <tr> <td>Phenylpropanoid biosynthesis</td><td>00940</td><td>1</td><td>Details</td></tr> <tr> <td>Pyrimidine metabolism</td><td>00240</td><td>1</td><td>Details</td></tr> <tr> <td>Pyruvate metabolism</td><td>00620</td><td>1</td><td>Details</td></tr> <tr> <td>T cell receptor signaling pathway</td><td>04660</td><td>1</td><td>Details</td></tr> </table>				Pathway	Map	Components	Details	Metabolic pathways	01100	4	Details	Starch and sucrose metabolism	00500	2	Details	Alanine, aspartate and glutamate metabolism	00250	1	Details	Ascorbate and aldarate metabolism	00053	1	Details	Biosynthesis of phenylpropanoids	01061	1	Details	Biosynthesis of secondary metabolites	01110	1	Details	Glyoxylate and dicarboxylate metabolism	00630	1	Details	Methane metabolism	00680	1	Details	Microbial metabolism in diverse environments	01120	1	Details	Pentose and glucuronate interconversions	00040	1	Details	Phenylalanine metabolism	00360	1	Details	Phenylpropanoid biosynthesis	00940	1	Details	Pyrimidine metabolism	00240	1	Details	Pyruvate metabolism	00620	1	Details	T cell receptor signaling pathway	04660	1	Details
Pathway	Map	Components	Details																																																																
Metabolic pathways	01100	4	Details																																																																
Starch and sucrose metabolism	00500	2	Details																																																																
Alanine, aspartate and glutamate metabolism	00250	1	Details																																																																
Ascorbate and aldarate metabolism	00053	1	Details																																																																
Biosynthesis of phenylpropanoids	01061	1	Details																																																																
Biosynthesis of secondary metabolites	01110	1	Details																																																																
Glyoxylate and dicarboxylate metabolism	00630	1	Details																																																																
Methane metabolism	00680	1	Details																																																																
Microbial metabolism in diverse environments	01120	1	Details																																																																
Pentose and glucuronate interconversions	00040	1	Details																																																																
Phenylalanine metabolism	00360	1	Details																																																																
Phenylpropanoid biosynthesis	00940	1	Details																																																																
Pyrimidine metabolism	00240	1	Details																																																																
Pyruvate metabolism	00620	1	Details																																																																
T cell receptor signaling pathway	04660	1	Details																																																																
<p>Details of Starch and sucrose metabolism (00500)</p> <p>Cellulase. EC:3.2.1.4</p> <ul style="list-style-type: none"> gi 351622889 gb JP477190.1 <p>Pectinesterase. EC:3.1.1.11</p> <ul style="list-style-type: none"> gi 351597529 gb JP452074.1 																																																																			

Fig. 1.13: List of Pathways in the cart

Annotation Search

Here you can search for annotations. All subsearches will be AND'ed.

Search for

Description contains

MapMan description contains

MapMan has Bincode

Has GO

Has GO or Children (slower!)

In Pathway (KEGG ID)

In Pathway (definition contains)

Interpro ID

Interpro Match Description

Has mapman Bincode.

Start Search

Fig. 1.14: Combisearch

Advanced Search

This field allows you to search for as many unigenes or isoforms as you want at once. (Up to a reasonable limit)
For every found isoform, corresponding unigene will be shown.
For each found unigene, all isoforms will be shown.
This search does not allow wildcards.

This List contains IDs and other text
gi|351603594|gb|JP457951.1|
gi|351598622|gb|JP453137.1| however the IDs are found
gi|351623904|gb|JP478205.1|
gi|351616445|gb|JP470746.1| and extracted
gi|351622338|gb|JP476639.1|
gi|351606046|gb|JP460383.1|
gi|351590913|gb|JP449370.1|
gi|351629444|gb|JP482327.1|

search

☒ Strict Search

Fig. 1.15: Multisearch

24

Chapter 1. Cannabis sativa transcriptome

Blast

Blast type:

Blast against Database:

FASTA sequence(s):

```
>test
TCTGTAGCTAGTTGTGTTTTATAATATCATCACTAAAAGTAAAAATGGCAGCCACCAGCAATGCTTTGAAGGGTAAGATTGATATTGATGTT
GATATCAAGGCATCTCCTACAAAAGTTCTACCACATGTTCCGGAAAACTGCTCACATCGTTCCCTCACTGCGCCGGTAGCCATATTCAAGGAG
TTGATGTCCATGAAGGTGACTGGGAGAGCCATGGCTCTATCAAGTTTTGGAAATATACAGTTGACGGAAAAAGAGGAGACGTTCAAGGAGA
AGGTGGAATTAGATGATGAGAACAAATAAGGTGAGTCTGATTGGAATTGAAGGTGATGTGTTCAAGCACTACAAGTGCTTCAATGTTATATA
CCAGGCGGTTCCCAATCCCAAGGGCGAGGGTTCCTTGGCCAACTCAGCATCGAGTATGAGAAGCTCGACCCTTCTGTTCCAACTCCAAA
CAAGTACCTCAATCTCATGATTAAACATCACCAAAGATATCGATTCTCACTTTGAGAAGGCTGCCCCCTCAAACTAAGACCGTTGTTTGGTGAA
AAATGAATTGATGAAAAACAAATAAAAGGGTCCATATATATCGTTTATATTATATGTGTGTGAAGAAATTAATAAGGATAATTTATTTTATCC
ATAATCAGTCACTCGAGCTTCCGCTGTCTCTACTACTTGCATGTACTACTATACAACAAGC
```

Max Target Sequences: Maximum E-value: Task: Wordsize:

Fig. 1.16: Blast interface



Fig. 1.17: Blast results



Fig. 1.18: Blast results