

ATGC Application User Manual

Table of Contents

1 Description.....	4
2 Install application.....	4
2.1 System Requirements.....	4
2.2 Complete instalation (without virtual machine).....	4
2.2.1 Software requirements.....	4
2.2.2 PostgeSQL configuration.....	6
2.3 Uncompress application file.....	6
2.4 Instalation using Virtual machine.....	6
2.4.1 VirtualBox install.....	6
2.4.2 Import VM Image.....	7
3 Runing application (Using Web2Py executable).....	7
3.1 Using screen command:	7
3.2 Using background (choose a password):.....	7
3.3 Using the graphical user interface (GUI):	7
3.4 Using web2py http server.....	7
3.5 Using Apache web server	7
4 Application access using the web browser.....	8
5 Application Usage.....	9
5.1 Database creation and load ontologies (Setup menu).....	9
5.1.1 Database creation.....	9
5.1.2 Database selection.....	10
5.1.3 Load Ontologies.....	10
5.1.4 Dump and Restore.....	10
5.2 Data loading	11
5.2.1 Create Organism.....	11
5.2.2 Load features from fasta.....	11
5.2.3 Features → CV associations (load cvterms for features)	12
5.2.4 Blast run results (XML Files).....	12
5.2.5 Genotypes (lines), markers and alleles.....	13
5.2.6 Load expression information.....	15
5.2.7 Feature relationships.....	17
5.3 Sequence search.....	18
5.3.1 Search features by name.....	18
5.3.2 Search features by list of names.....	19
5.3.3 Search features by ontology term name or accession.....	20
5.3.4 Search by Blast matches.....	21
5.4 Ontology exploration.....	23
5.5 Download.....	24
5.6 Software.....	25
5.6.1 Blast.....	25
5.7 Modify and delete.....	27
6 Share information.....	29
7 Testing the application.....	33

Table of Figures

Figure 1: First step in application.....	9
Figure 2: Database creation.....	9
Figure 3: Database selection.....	10
Figure 4: Load ontologies.....	10
Figure 5: Data loading after create an organism.....	11
Figure 6: Load features from fasta file.....	11
Figure 7: Feature → CV associations.....	12
Figure 8: Load Blast results.....	12
Figure 9: Create new line.....	13
Figure 10: Lines → CV associations.....	13
Figure 11: Lines → CV associations.....	13
Figure 12: Load markers.....	14
Figure 13: Markers → Line associations.....	14
Figure 14: Create experiment.....	15
Figure 15: Create library associated to a line and experiment.....	15
Figure 16: Experiments → CV associations.....	15
Figure 17: Experiments → CV associations (second part).....	16
Figure 18: Feature → Library associations.....	16
Figure 19: Load relationships between features.....	17
Figure 20: Load features relationships from gff3 file.....	17
Figure 21: Search by name.....	18
Figure 22: Search by name result.....	18
Figure 23: Search list of features by name.....	19
Figure 24: Search features by ontology annotation.....	20
Figure 25: Search by ontology, summary result.....	20
Figure 26: Search by ontology, feature list result.....	21
Figure 27: Search features by Blast matches.....	21
Figure 28: Search by Blast matches results.....	22
Figure 29: Pie chart: Biological process.....	23
Figure 30: Feature annotation tree (Dropdown menu).....	24
Figure 31: Download feature file.....	24
Figure 32: Download annotations.....	24
Figure 33: Create Blast database.....	25
Figure 34: Run Blast alignments.....	25
Figure 35: Example of Blast alignment.....	26
Figure 36: Example of Blast result.....	26
Figure 37: Search object to modify.....	27
Figure 38: Table to select object to modify.....	27
Figure 39: Update form to modify values.....	28
Figure 40: Web2py administration interface: write a password.....	29
Figure 41: Choosing "Edit" option.....	30
Figure 42: Click on "database administration" button.....	31
Figure 43: Click on db.navbar option.....	31
Figure 44: Hide the "Setup" part of the menu.....	32

Index of Tables

Table 1: Id values for main menu bar options in db.navbar table.....32

1 Description

ATGC is a web application that allows users to work with NGS transcriptomic data without a reference genome, using an ontology driven database schema to data store and management and provide interfaces to create schema structure and load ontologies and data, visualization, searching possibilities and data integration. Using this application is possible visualize, explore, analyze and share de novo transcriptomic data generated by NGS platforms using a Chado database to store the data. It is an open source and free available application, with support to store information in several Chado modules and uses different ontologies to classify data and then explore this data using a ontology structure with graphics, searches and description tables. ATGC is open source, so all we ask is that you cite our paper in any publications that use this application:

Cita

ATGC is a open source application, for more information, new releases, download source code and complete virtual machines (see below) or download the manual please visit the homepage: <http://atgcinta.sourceforge.net>

ATGC is available for installing in two ways, the first is a complete instalation mode (only tested for Linux machines) or the second is the use of virtual machines, downloading a VM image with the complete application (all requirements installed there). For detailed installation instructions please view the next points of this manual.

2 Install application

The application is dowloadble from: <http://atgcinta.sourceforge.net>, from this URL is possible to download source code and virtual machine (VM) images.

2.1 System Requirements

The software have been tested in the following Operative Systems:

- Distributions of Linux derived from Debian (Extensively tested in Debian7.0/5.0, Ubuntu12.04/14.04/15.04/16.04)
- Distributions of Linux derived from Red-Hat (Tested in Fedora20)
- For Windows or Mac (if you want also in any Linux distribution) please view the details below.

The application was tested (and run correctly) in a machine with:

- 1 CPU
- 1 Gb RAM (Using 500 Mb in average)

2.2 Complete instalation (without virtual machine)

2.2.1 Software requirements

The ATGC application require the following to run succesfully. In the absence of one or more of these packages, some ATGC parts may fail to run correctly. Listed in parenthesis are the versions used to test the application. These versions, or subsequent versions should assure the proper execution. These utilities must be accessible via the system path:

To install in Linux Debian (or related distributions), packages required:

- postgresql (9.3.5-0)
- ncbi-blast+ (2.2.28-2)
- emboss (6.6.0-1)
- python (2.7)
- python packages:
 - python-matplotlib (1.3.1-1)
 - python-pygraphviz (1.2-1)
 - python-biopython (1.63-1)
 - python-psycopg2 (2.4.5-1)
- perl (5.18.2)
- perl packages:
- perl-doc (3.19)
 - bioperl (1.6.923-1)
 - libgo-perl (0.15-1) (After install this package please change in the script: /usr/bin/go2chadoxml: go2fmt.pl to go2fmt)
 - libpg-perl (1:2.1.1-4)
 - libdata-stag-perl (0.13-1)
 - libdbix-dbstag-perl (0.12-1)
 - libsql-translator-perl (0.11018-1)

To install in Linux Red-Hat (or related distributions), packages required:

- postgresql-server (9.3.5-1)
- postgresql-contrib (9.3.5-1)
- postgresql-devel (9.3.5-1)
- ncbi-blast+ (2.2.29-2)
- EMBOSS (6.4.0-11)
- python27 (2.7)
- python packages:
 - python-matplotlib (1.3.1-3)
 - graphviz-python (2.34.0-8)
 - python-biopython (1.64-1)
 - python-psycopg2 (2.5.1-2)
- perl (5.18.2)
- perl packages:
 - perl-doc (3.20)
 - bioperl (1.6.924)
 - go-perl (0.15)
 - PgSQL (0.51)
 - Data-Stag (0.14)
 - DBIx-DBStag (0.12)
 - SQL-Translator (0.11020)

2.2.2 PostgreSQL configuration

The application need to create and enable the user with which are running in the postgresql package. For this, execute the following commands replacing <username> by the correct name.

```
$ su -  
$ su - postgres  
$ createuser <username>  
Enter to postgresql to enable the user to database creation:  
$ psql  
$ ALTER ROLE <username> CREATEDB;  
$ exit
```

2.3 Uncompress application file

To uncompress the application run the following command:

```
tar xzvf ATGC-1.0_source.tar.gz
```

This file contains a directoy called web2py and include all files to run the application.

For the specific files for ATGC application, you must go to <path to>/web2py/applications/ATGC

2.4 Instalation using Virtual machine

For the instalation using the virtual machine (VM) image and VirtualBox software you can download the VM image from <http://www.mediafire.com>

This VM has all packages installed and is avaiable in 32 and 64 bits.

The VM has the following features:

- LUbuntu16.04
- 1 CPU
- 1 Gb RAM
- 100 Gb Hard Drive
- username: atgc
- password: atgc
- root password: atgc

2.4.1 VirtualBox install

VirtualBox software can be downloaded and install from: <http://www.virtualbox.org/wiki/Downloads> Choose the correct package according to your operating system.

2.4.2 Import VM Image

To import VM image download and decompress **ATGC_XXb.ova** (XX = 32 or 64) file from:

http://www.mediafire.com/file/0bbocuj5uvcjc6o/ATGC_32b.ova

or

http://www.mediafire.com/file/dsx3evld8w5y1a6/ATGC_64b.ova

Open VirtualBox :

- **File -> Import Appliance (Use the .ova file to import)**

3 Runing application (Using Web2Py executable)

The password is needed when the user want to access to the administrative interface.

For the virtual machine mode installation replace <path to> by /home/atgc, for complete instalations replace by the correct path.

3.1 Using screen command:

```
$ screen -S web2py
```

```
$ <path to>/web2py.py --nogui
```

3.2 Using background (choose a password):

```
$ <path to>/web2py.py --nogui --password=password &
```

3.3 Using the graphical user interface (GUI):

```
$ <path to>/web2py.py
```

3.4 Using web2py http server

```
$ screen -S web2py
```

```
$ <path to>/web2py.py --nogui --interface='localhost:8000;<serverIP>:<port>'
```

In this way, other machines in the network can access to the application, writing the following URL in a web browser: **<http://<serverIP>:<port>/ATGC/>**. The administrative interface is only accessible from the server machine.

3.5 Using Apache web server

To use apache http server, you must install 2 packages and configure Web2py applications on apache configuration files with the follow commands:

```
$ apt-get install apache2
```

```
$ apt-get install libapache2-mod-wsgi
```

Download the apache configuration file from **<http://atgcinta.sourceforge.net>** called web2py.conf-apache.2.X (X= 2 or 4, according to apache version) and put this file in **/etc/apache2/sites-enabled**

directory with the **web2py.conf** name.

Then, open **/etc/apache2/sites-enabled/web2py.conf** file and replace:

<hostname> for the correct hostname defined in /etc/hosts file with the selected IP adress (atgc-VirtualBox in the virtual machine)

<username> for the correct username (atgc in the case of virtual machine)

<path to web2py home> for the complete path to web2py main directory (/home/atgc/web2py in the case of the virtual machine)

Put the follow two lines in **/etc/apache2/apache2.conf** file:

ServerName <hostname>

Listen 8000

Finally, you can acces using the web browser the follow ways:

<http://localhost:8000/>

<http://hostname/>

http://ip_adress/

4 Application access using the web browser

To access to the application, write this URL in the web browser: <http://localhost:8000> (other options above)

5 Application Usage

5.1 Database creation and load ontologies (Setup menu)

For the application usage, the first steps are: create a Chado database and load ontologies in the database. You must use the functions in the setup option of the navigation bar. When you start the application for the first time or when not exists a database, the application looks like this:

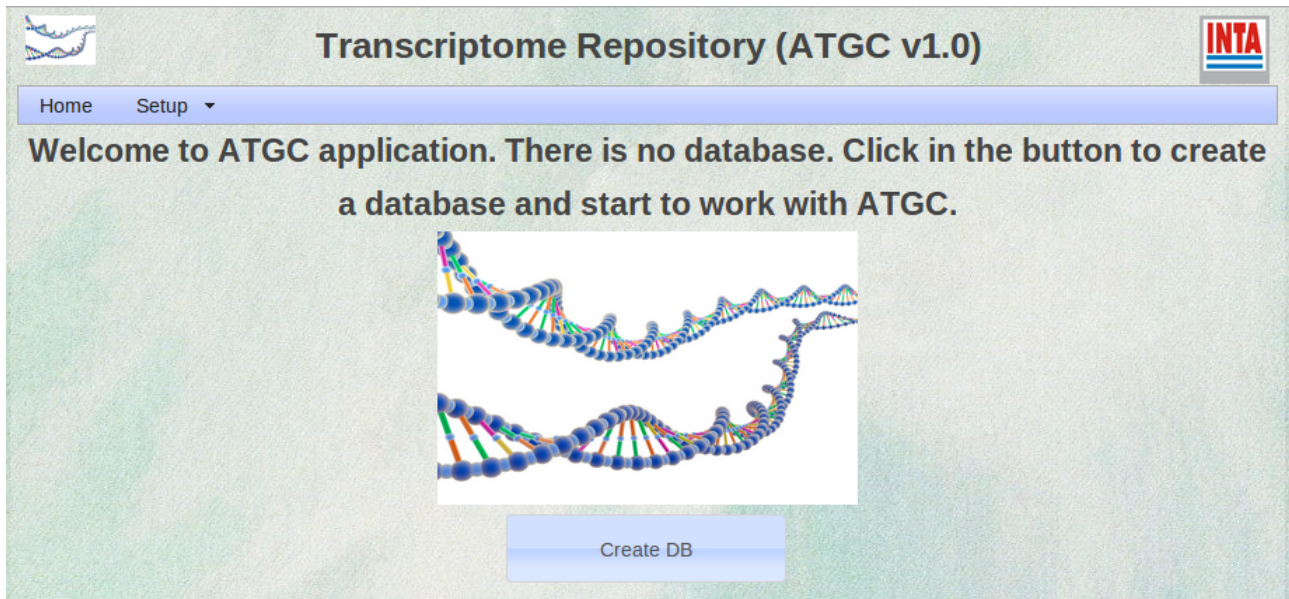


Figure 1: First step in application

5.1.1 Database creation

For database creation you only must choose a database name, and the application creates a database using a template of Chado schema and automatically load basic ontologies:

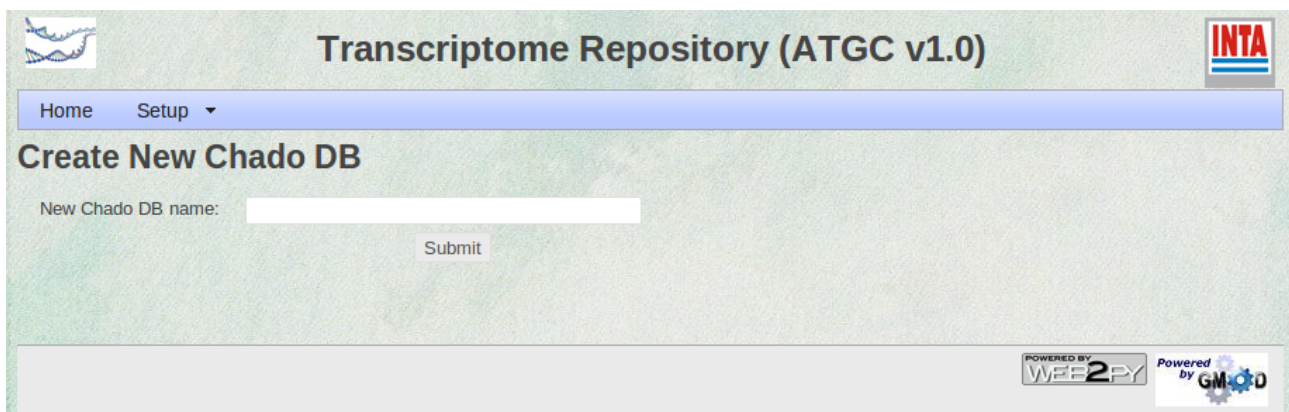


Figure 2: Database creation

5.1.2 Database selection

You can have several databases in the same instance of the application, but you must select the database with you want work.

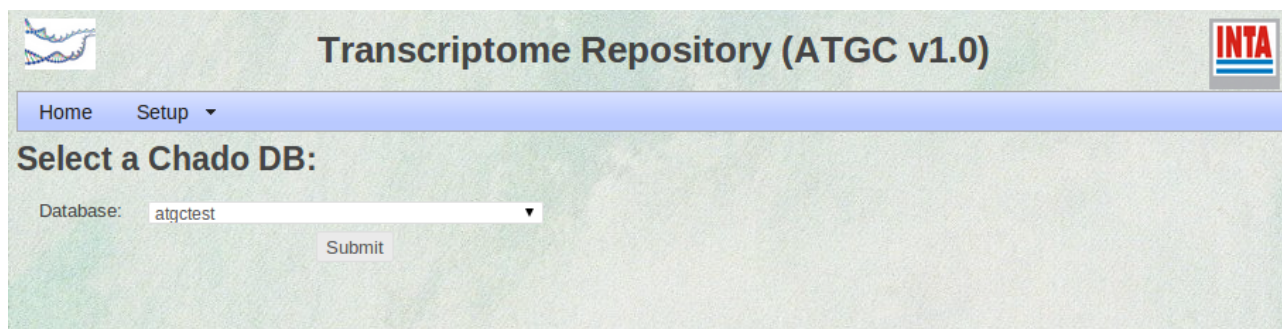


Figure 3: Database selection

5.1.3 Load Ontologies

After database creation, you must load the ontologies into the database, the main ontologies to follow with the data loading are Sequence Ontology (SO) and Gene Ontology (GO). You can load ontologies using “Load Ontology” option in Setup menu. When you load these two ontologies, you can follow loading data into the database (i.e. fasta files, GO annotations).

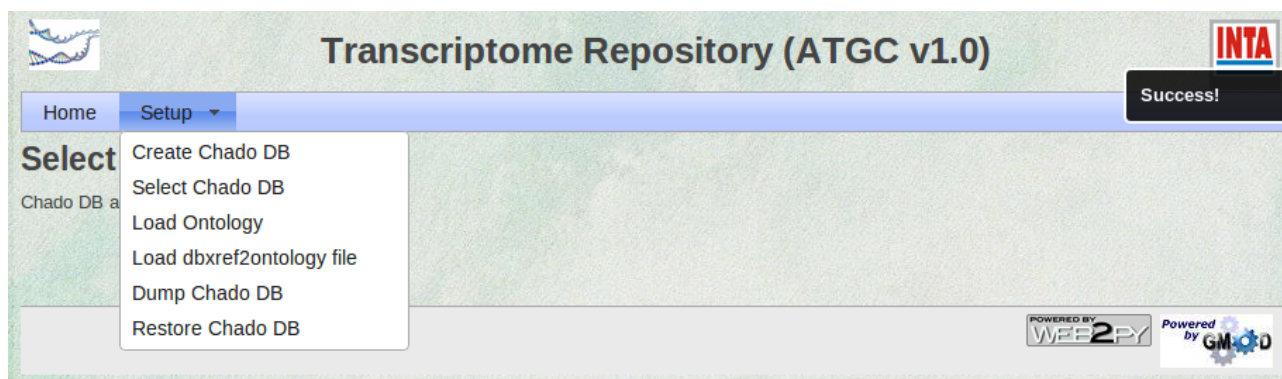


Figure 4: Load ontologies

If you use a software with a output you want load into the database using a GO ontology, you must load <software>2GO file using “Load dbxref2ontology file” option. One example of this, is InterProScan software, you can download interpro2go file from <http://geneontology.org/external2go/> and load this file to generate relationships between InterPro and GO terms into the database (after loading GO ontology).

5.1.4 Dump and Restore

The “Dump Chado DB” is used to do a backup a complete database in a “sql” compatible file format. Then is possible to use this file to restore a database to a previous state (before an error or unwanted changes) using the “Restore Chado DB” option.

5.2 Data loading

5.2.1 Create Organism

The first step then load ontologies is the creation of a organism for the database, along with this step, you must load a image to identify your organism, this image will be used in the application header (left image) and as image in the browser tab. Only is possible create one organism by database, if you need to use the application for several organisms you must create several databases (one for each organism). For this job, you don't need a new complete process of database creation and load ontologies, you can make a dump of the original database, create a new database and then, you can load the complete “dump” of the original database into the new database (basically creating a copy of the original database) with “Restore Chado DB”.

After create an organism, you can start with loading the data, creating experiments, libraries, lines and load features from fasta files and lists files (for features without sequence). The menu looks like this:

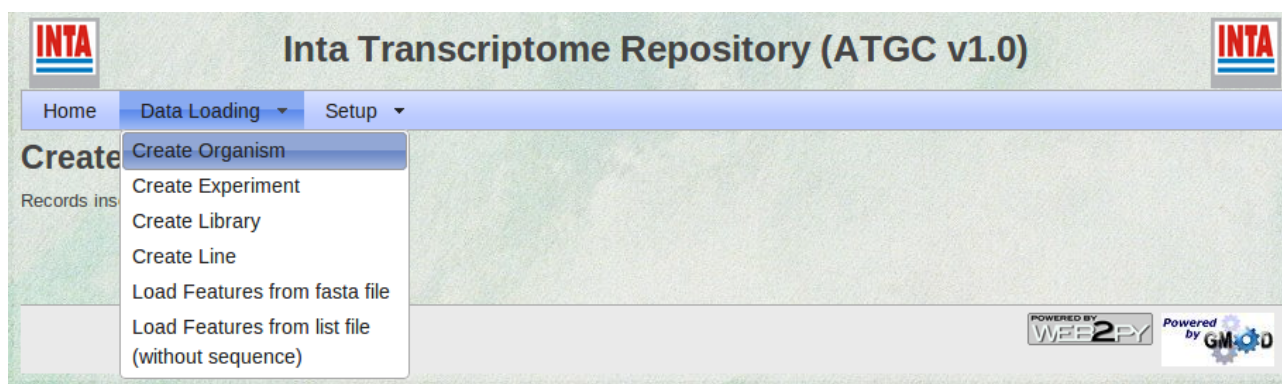


Figure 5: Data loading after create an organism

5.2.2 Load features from fasta

After creating the organism, you can load the sequences of transcripts or genes in “fasta” format, choosing the feature type for Sequence Ontology (contig for example), then all data can be loaded using this sequences as repository or reference.

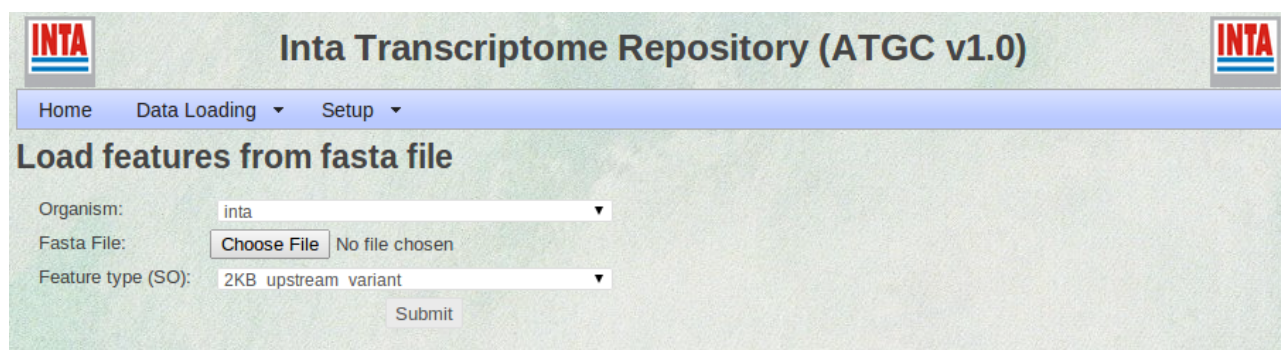


Figure 6: Load features from fasta file

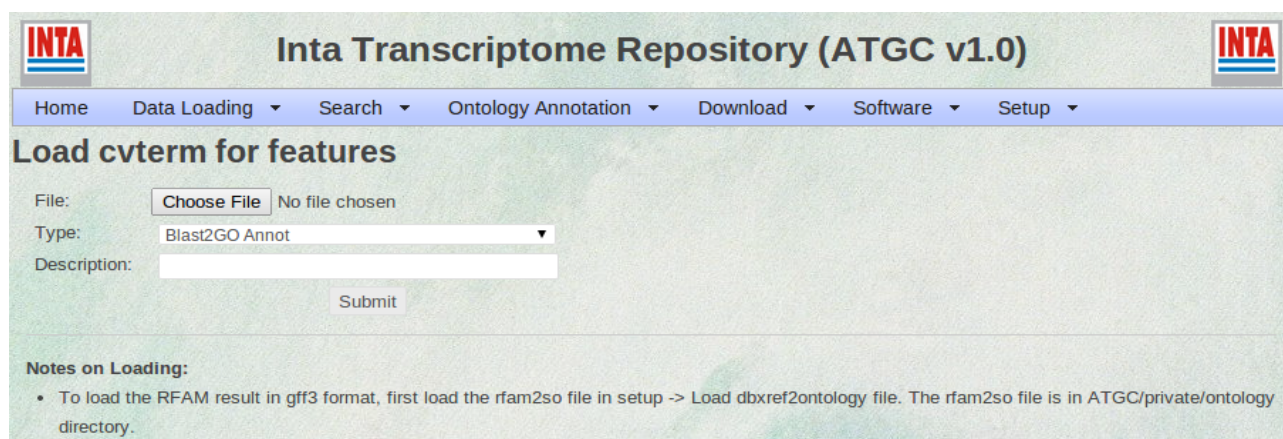
Loading features in fasta format allows to add all associated information, such as, functional annotation, blast results, expression levels, markers, alleles, genotypes and relationships with other features.

5.2.3 Features → CV associations (load cvterms for features)

Associations from features to controlled vocabularies can be made using several software results, such as, Blast2GO (annot file), Interproscan (raw file), RFAM (gff3 file) or tab files (in general). To load Interproscan results you must load “interpro2go” file from:

<http://geneontology.org/external2go/>, to load RFAM results you can load “rfam2go” file from the same place or “rfam2so” file developed by ATGC creators placed into:

web2py/applications/ATGC/private/ontology directory. The information placed on “Description” field will be used to identify the source of annotations in feature detail pages.

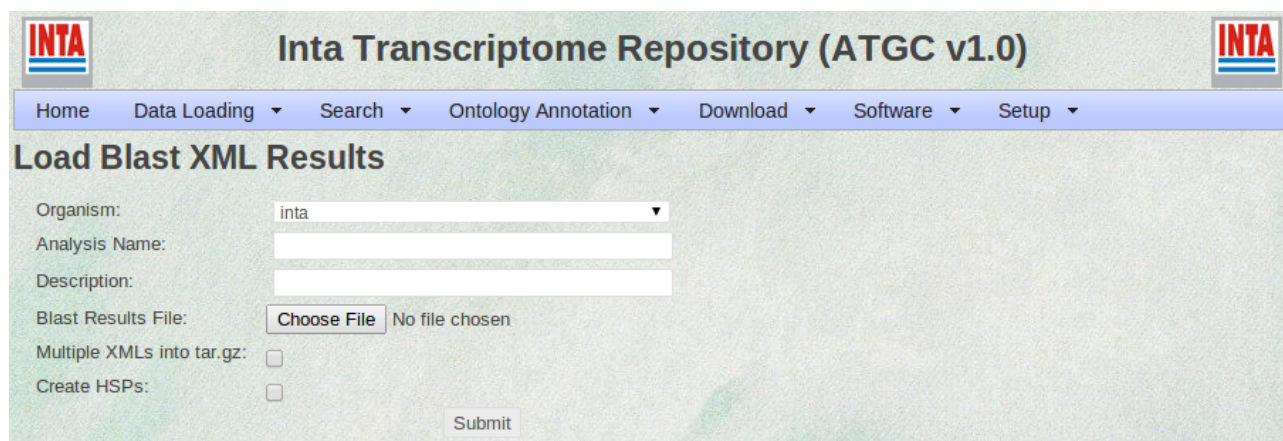


The screenshot shows the 'Inta Transcriptome Repository (ATGC v1.0)' web interface. At the top is a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. Below this is the section 'Load cvterm for features'. It contains a form with the following fields: 'File:' with a 'Choose File' button and 'No file chosen' text; 'Type:' with a dropdown menu currently set to 'Blast2GO Annot'; and 'Description:' with a text input field. A 'Submit' button is located below the description field. At the bottom, there is a 'Notes on Loading:' section with a bullet point: 'To load the RFAM result in gff3 format, first load the rfam2so file in setup -> Load dbxref2ontology file. The rfam2so file is in ATGC/private/ontology directory.'

Figure 7: Feature → CV associations

5.2.4 Blast run results (XML Files)

In the same way that CV association, is possible to load the results of Blast alignments against a database using xml as output type.

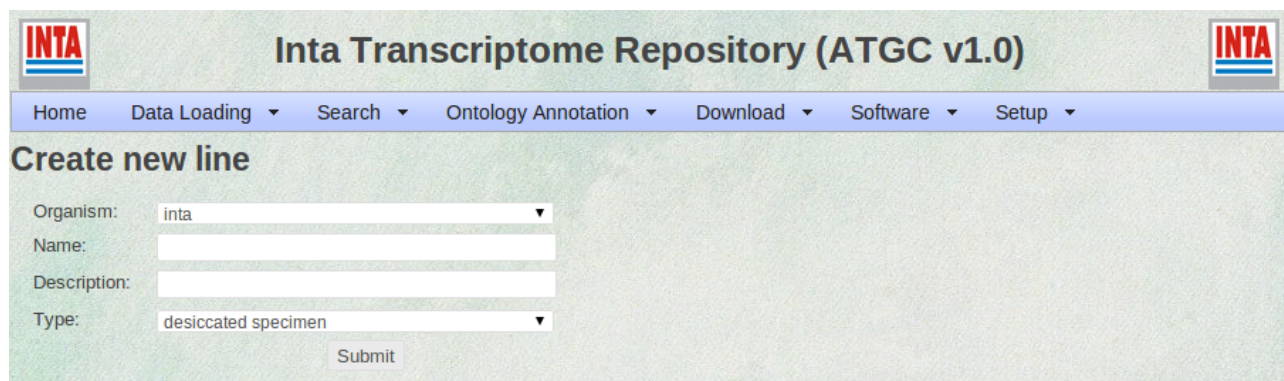


The screenshot shows the 'Inta Transcriptome Repository (ATGC v1.0)' web interface. At the top is a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. Below this is the section 'Load Blast XML Results'. It contains a form with the following fields: 'Organism:' with a dropdown menu currently set to 'inta'; 'Analysis Name:' with a text input field; 'Description:' with a text input field; 'Blast Results File:' with a 'Choose File' button and 'No file chosen' text; 'Multiple XMLs into tar.gz:' with an unchecked checkbox; and 'Create HSPs:' with an unchecked checkbox. A 'Submit' button is located at the bottom right of the form.

Figure 8: Load Blast results

5.2.5 Genotypes (lines), markers and alleles

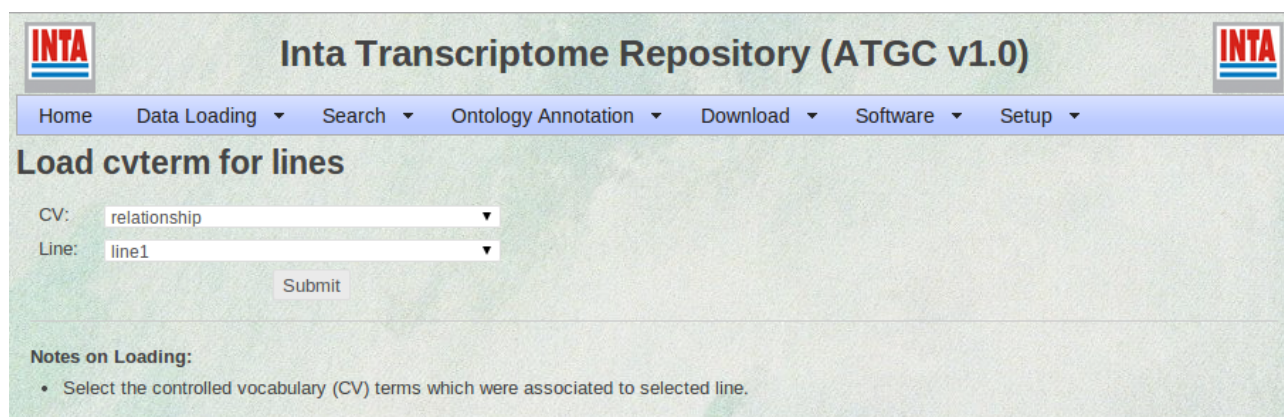
You can load a set of markers and genotypes defined by alleles of this markers, to make this, you must load first a genotypes and then the markers and alleles.



The screenshot shows the 'Create new line' form in the Inta Transcriptome Repository (ATGC v1.0). The form includes a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The form fields are: Organism (dropdown menu with 'inta' selected), Name (text input), Description (text input), and Type (dropdown menu with 'desiccated specimen' selected). A 'Submit' button is located at the bottom right of the form.

Figure 9: Create new line

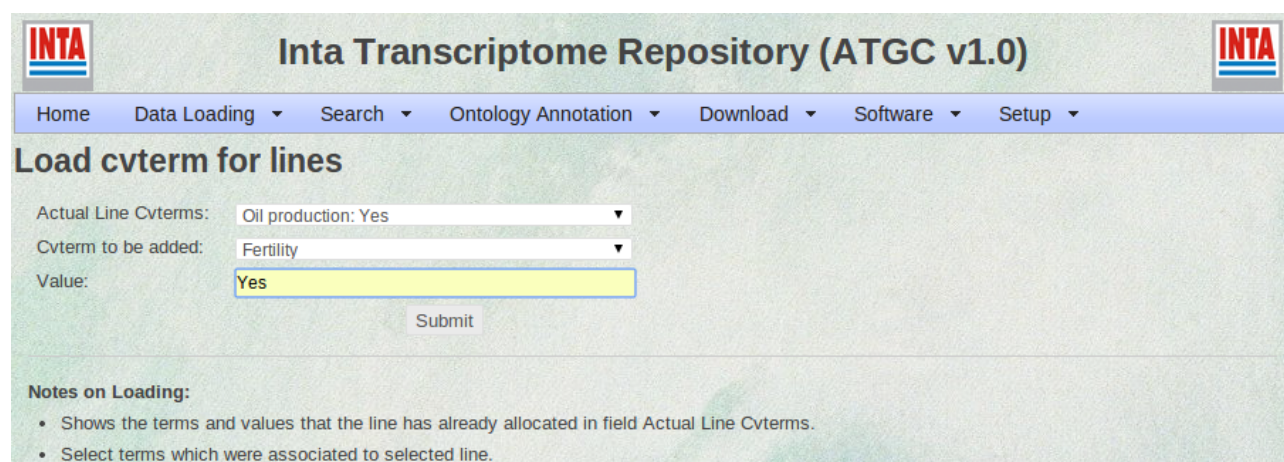
Is possible to describe the characteristics of the genotypes using terms of controlled vocabularies, in the tab of Lines → CV associations, first choosing the CV (for example INTA_CV or any ontology previously loaded) and then adding a characteristic and value for this line.



The screenshot shows the 'Load cvterm for lines' form in the Inta Transcriptome Repository (ATGC v1.0). The form includes a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The form fields are: CV (dropdown menu with 'relationship' selected) and Line (dropdown menu with 'line1' selected). A 'Submit' button is located at the bottom right of the form. Below the form, there is a section titled 'Notes on Loading:' with a bullet point: 'Select the controlled vocabulary (CV) terms which were associated to selected line.'

Figure 10: Lines → CV associations

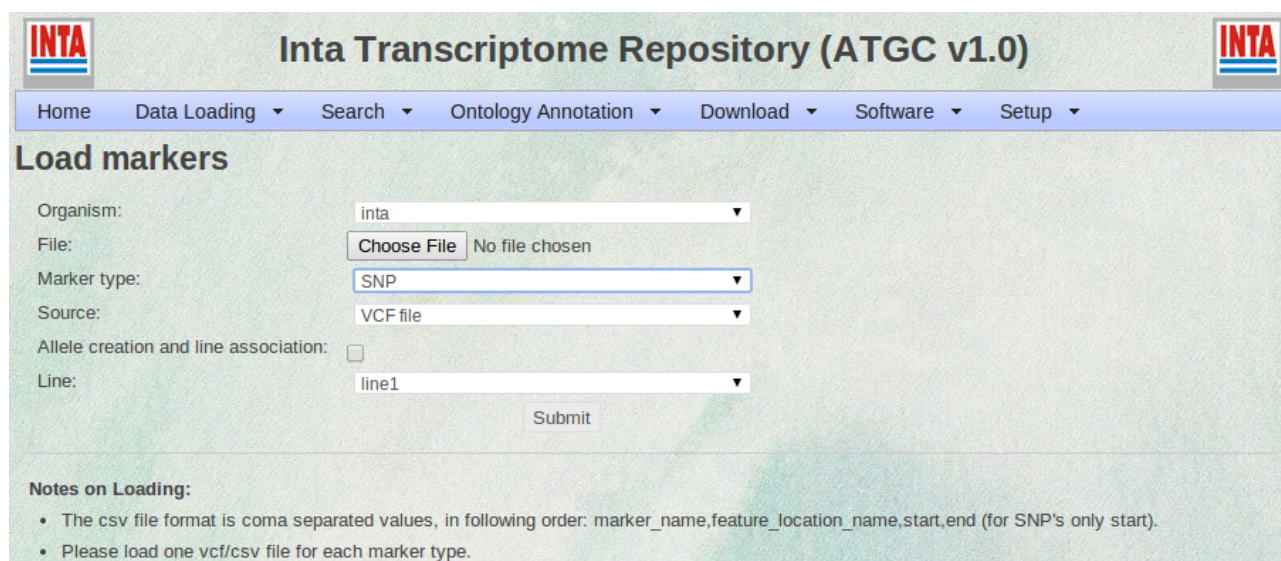
For example, setting the characteristic “Fertility” on “Yes”



The screenshot shows the 'Load cvterm for lines' form in the Inta Transcriptome Repository (ATGC v1.0). The form includes a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The form fields are: Actual Line Cvterms (dropdown menu with 'Oil production: Yes' selected), Cvterm to be added (dropdown menu with 'Fertility' selected), and Value (text input with 'Yes' entered). A 'Submit' button is located at the bottom right of the form. Below the form, there is a section titled 'Notes on Loading:' with two bullet points: 'Shows the terms and values that the line has already allocated in field Actual Line Cvterms.' and 'Select terms which were associated to selected line.'

Figure 11: Lines → CV associations

To load markers you can use Load Markers section, given the markers information on several file types, such as: VCF, CSV or specific formats of some softwares. If you want to create a relation between markers and genotype, you must to select the option “Allele creation and line association” and select the genotype name.



The screenshot shows the 'Load markers' section of the Inta Transcriptome Repository (ATGC v1.0) web interface. The header includes the INTA logo and a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The main form contains the following fields:

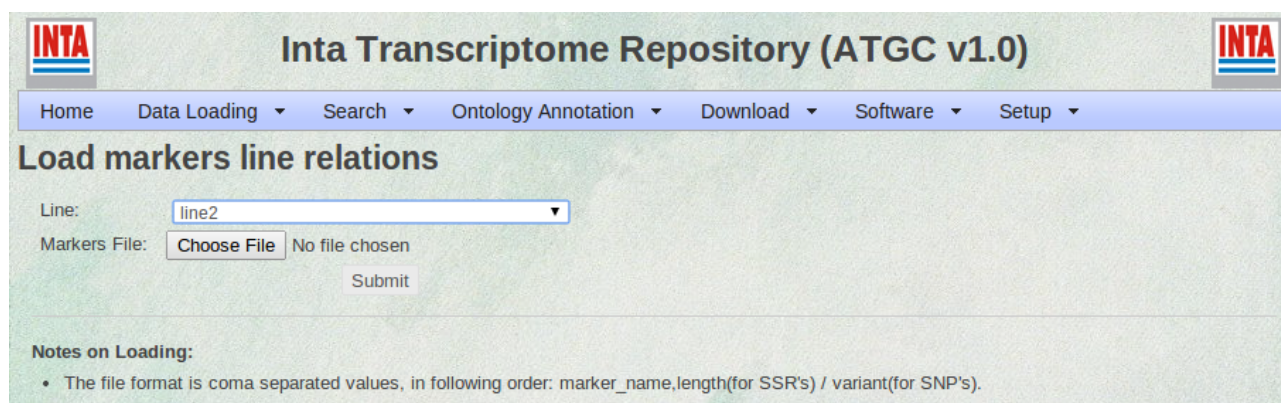
- Organism: A dropdown menu with 'inta' selected.
- File: A 'Choose File' button and the text 'No file chosen'.
- Marker type: A dropdown menu with 'SNP' selected.
- Source: A dropdown menu with 'VCF file' selected.
- Allele creation and line association: An unchecked checkbox.
- Line: A dropdown menu with 'line1' selected.
- A 'Submit' button.

Below the form, the 'Notes on Loading:' section contains two bullet points:

- The csv file format is coma separated values, in following order: marker_name,feature_location_name,start,end (for SNP's only start).
- Please load one vcf/csv file for each marker type.

Figure 12: Load markers

In the same way, if you want create an association of marker with other line, you can use the window: Markers → Line associations, from a csv file to describe the allele of each marker in the new line.



The screenshot shows the 'Load markers line relations' section of the Inta Transcriptome Repository (ATGC v1.0) web interface. The header is identical to Figure 12. The main form contains the following fields:

- Line: A dropdown menu with 'line2' selected.
- Markers File: A 'Choose File' button and the text 'No file chosen'.
- A 'Submit' button.

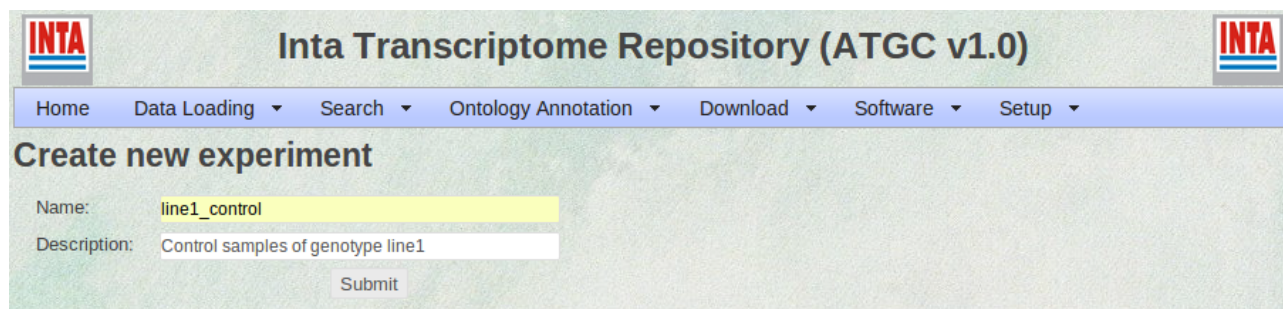
Below the form, the 'Notes on Loading:' section contains one bullet point:

- The file format is coma separated values, in following order: marker_name,length(for SSR's) / variant(for SNP's).

Figure 13: Markers → Line associations

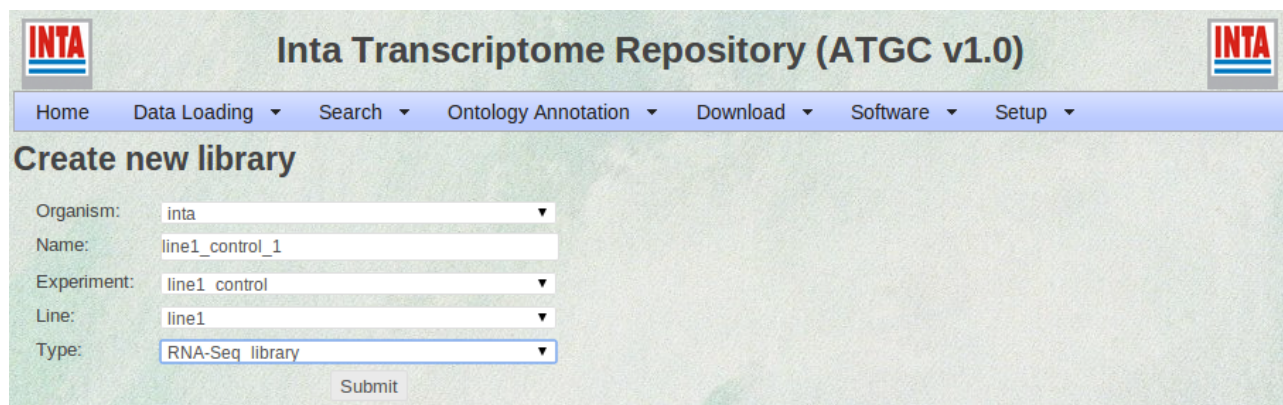
5.2.6 Load expression information

To load expression information in the database, you first create a structure of experiments (with your characteristics, any treatment in the assay is a different experiment, for example, if you have two treatments, control and treated, you must to create two experiments) and libraries (like biological or technical replicates) and finally you can use the “Feature → Library associations” window to load expression or other measure variables to any feature (contigs for example).



The screenshot shows the 'Create new experiment' form in the Inta Transcriptome Repository (ATGC v1.0). The form has a header with the INTA logo and a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The form itself has a title 'Create new experiment' and two input fields: 'Name' with the value 'line1_control' and 'Description' with the value 'Control samples of genotype line1'. A 'Submit' button is located at the bottom right of the form.

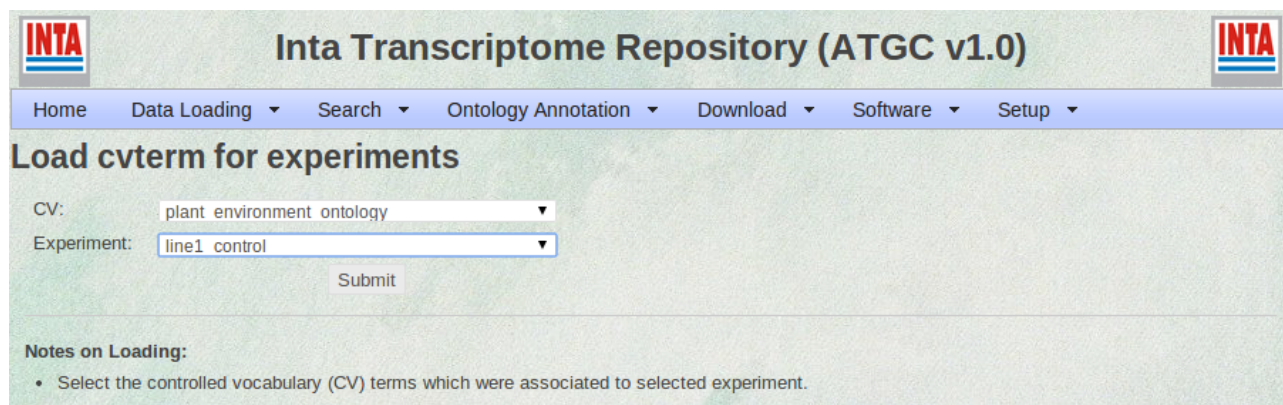
Figure 14: Create experiment



The screenshot shows the 'Create new library' form in the Inta Transcriptome Repository (ATGC v1.0). The form has a header with the INTA logo and a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The form itself has a title 'Create new library' and five input fields: 'Organism' with the value 'inta', 'Name' with the value 'line1_control_1', 'Experiment' with the value 'line1_control', 'Line' with the value 'line1', and 'Type' with the value 'RNA-Seq library'. A 'Submit' button is located at the bottom right of the form.

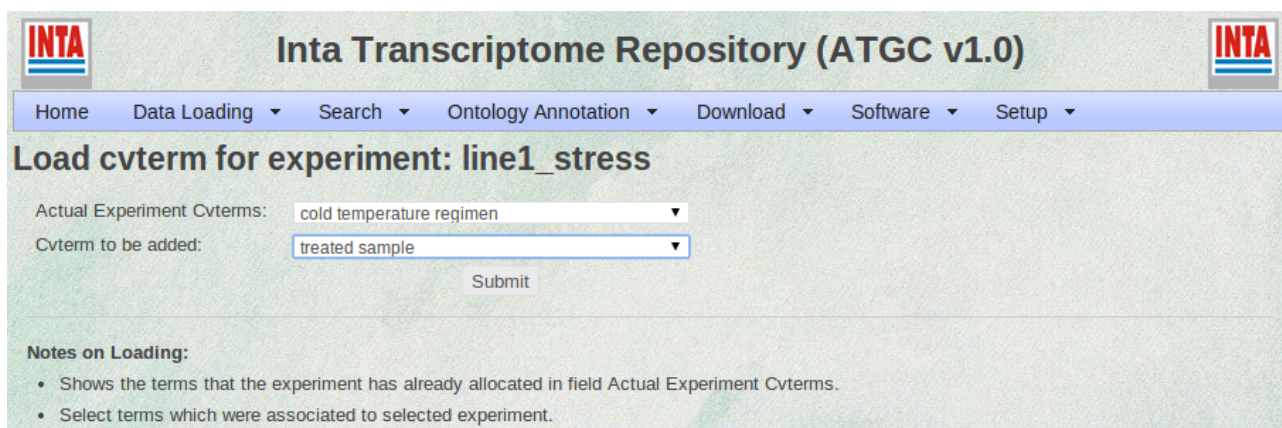
Figure 15: Create library associated to a line and experiment

Is very important to add information to fully describe the experiments, then this information will be used to create dynamic expression graphics in feature detail pages.



The screenshot shows the 'Load cvterm for experiments' form in the Inta Transcriptome Repository (ATGC v1.0). The form has a header with the INTA logo and a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The form itself has a title 'Load cvterm for experiments' and two input fields: 'CV' with the value 'plant environment ontology' and 'Experiment' with the value 'line1_control'. A 'Submit' button is located at the bottom right of the form. Below the form, there is a section titled 'Notes on Loading:' with a bullet point: 'Select the controlled vocabulary (CV) terms which were associated to selected experiment.'

Figure 16: Experiments → CV associations



Inta Transcriptome Repository (ATGC v1.0)

Home Data Loading Search Ontology Annotation Download Software Setup

Load cvterm for experiment: line1_stress

Actual Experiment Cvterms: cold temperature regimen

Cvterm to be added: treated sample

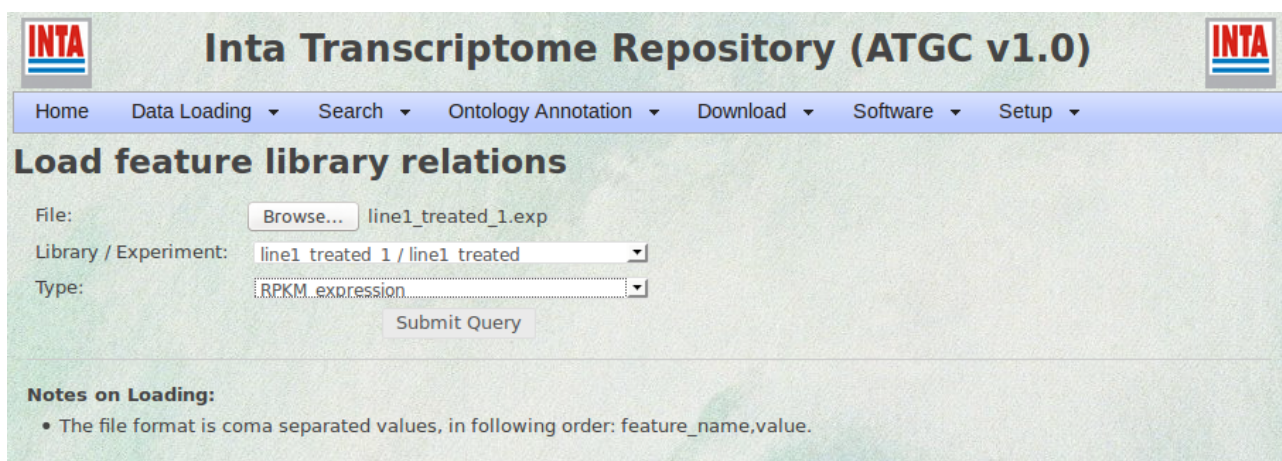
Submit

Notes on Loading:

- Shows the terms that the experiment has already allocated in field Actual Experiment Cvterms.
- Select terms which were associated to selected experiment.

Figure 17: Experiments → CV associations (second part)

With the structure of experiments and libraries created, you can load to the database the information of features related with each library, for example, in the Figure 18 the file contains values of RPKM counts of expression.



Inta Transcriptome Repository (ATGC v1.0)

Home Data Loading Search Ontology Annotation Download Software Setup

Load feature library relations

File: Browse... line1_treated_1.exp

Library / Experiment: line1 treated 1 / line1 treated

Type: RPKM expression

Submit Query

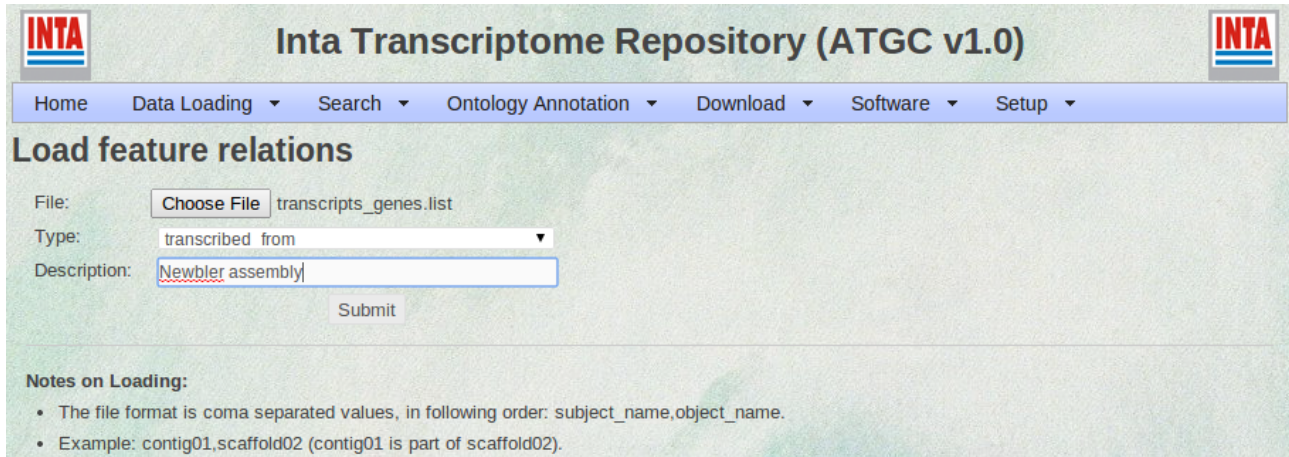
Notes on Loading:

- The file format is coma separated values, in following order: feature_name,value.

Figure 18: Feature → Library associations

5.2.7 Feature relationships

You can load relationships between features, for example, if you predict clusters of contigs that possible come from the same gene, you can load this information using the option of Figure 19, you must define the type of the relationship using ontology terms and load previously contigs and genes as features.



The screenshot shows the 'Load feature relations' form in the Inta Transcriptome Repository (ATGC v1.0). The form includes a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The main form area has the following fields:

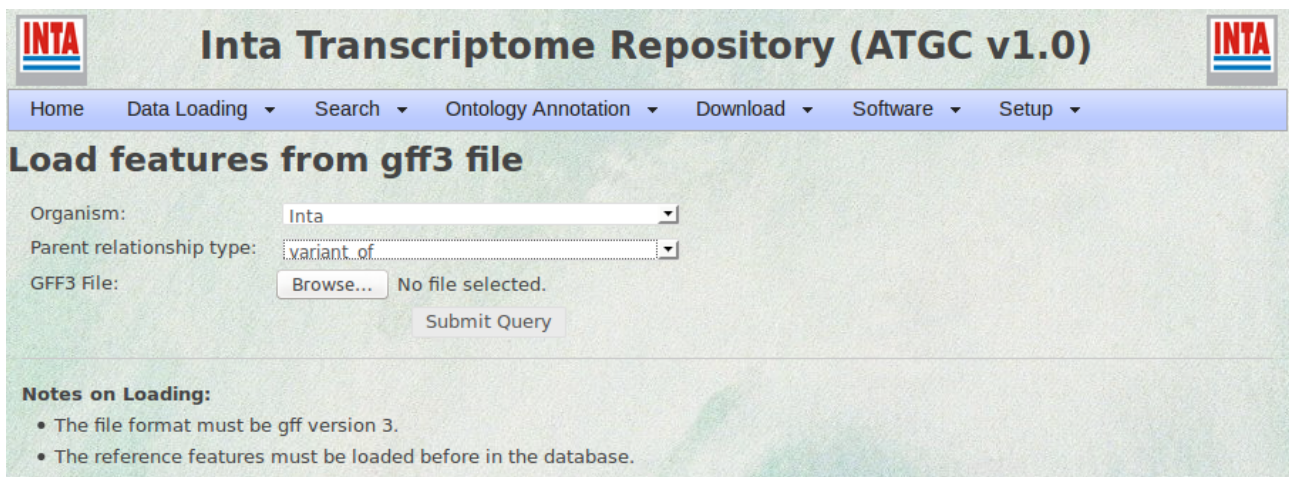
- File:** A text input field containing 'transcripts_genes.list' and a 'Choose File' button.
- Type:** A dropdown menu with 'transcribed from' selected.
- Description:** A text input field containing 'Newbler assembly'.
- Submit:** A button to submit the form.

Below the form, there is a section titled 'Notes on Loading:' with the following bullet points:

- The file format is coma separated values, in following order: subject_name,object_name.
- Example: contig01,scaffold02 (contig01 is part of scaffold02).

Figure 19: Load relationships between features

Other possibility to load feature relationships is using gff3 files, and you can choose the type of parental relationship between features (for example, relation between exon and transcript), in this case, is only necessary have loaded the references of the gff3 file as features in the database.



The screenshot shows the 'Load features from gff3 file' form in the Inta Transcriptome Repository (ATGC v1.0). The form includes a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The main form area has the following fields:

- Organism:** A dropdown menu with 'Inta' selected.
- Parent relationship type:** A dropdown menu with 'variant_of' selected.
- GFF3 File:** A text input field containing 'No file selected.' and a 'Browse...' button.
- Submit Query:** A button to submit the query.

Below the form, there is a section titled 'Notes on Loading:' with the following bullet points:

- The file format must be gff version 3.
- The reference features must be loaded before in the database.

Figure 20: Load features relationships from gff3 file

5.3 Sequence search

5.3.1 Search features by name

Using the application, you can explore all information related to the features from different ways, one of this ways, is search features from the name of the feature, or an expression related with the name (using % character as wildcard)

Figure 21: Search by name

Feature Name	Go Term Annotation	SO Term Annotation
contig00004	endonuclease activity	contig
contig00006		
contig00012	ATP binding;abscisic acid transport;membrane;response to ozone;plasma membrane;ATP catabolic process;organic phosphonate transmembrane-transporting ATPase activity;lead ion transport	contig
contig00033		
contig00037	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor;oxidation-reduction process;regulation of proton transport;photosynthesis, light reaction;chloroplast thylakoid membrane;plasma membrane;protein autophosphorylation;NADH dehydrogenase complex (plastoquinone) assembly;NAD(P)H dehydrogenase complex (plastoquinone)	contig
isotig00001	chlorophyll binding;metal ion binding;oxidation-reduction process;photosystem I;membrane;electron carrier activity;oxidoreductase activity;chloroplast thylakoid membrane;photosystem II;photosynthesis, light harvesting;protein-chromophore linkage;integral component of membrane	contig
isotig00002	chlorophyll binding;metal ion binding;oxidation-reduction process;photosystem I;membrane;electron carrier activity;oxidoreductase activity;chloroplast thylakoid membrane;photosystem II;photosynthesis, light harvesting;protein-chromophore linkage;integral component of membrane	contig

Figure 22: Search by name result

5.3.2 Search features by list of names

If you have a list of features of interest for any reason, you can search information for these features using this option



The screenshot shows the web interface of the Inta Transcriptome Repository (ATGC v1.0). The header includes the INTA logo and the title "Inta Transcriptome Repository (ATGC v1.0)". Below the header is a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The main content area is titled ">Search List of Features by name". It contains a "Type:" dropdown menu set to "contig" and a "Feature List:" text box. Below the text box is a "Search" button. At the bottom, there is a section titled "Notes on Searching:" with two bullet points: "To search insert a list of sequence names in text box. Items may be separated by new lines, spaces, commas (,) and/or semicolons (;)." and "The search is case-insensitivity, so 'a' will match both 'a' and 'A'."

INTA Inta Transcriptome Repository (ATGC v1.0) INTA

Home Data Loading Search Ontology Annotation Download Software Setup

>Search List of Features by name

Type: contig

Feature List:

Search

Notes on Searching:

- To search insert a list of sequence names in text box. Items may be separated by new lines, spaces, commas (,) and/or semicolons (;).
- The search is case-insensitivity, so "a" will match both "a" and "A".

Figure 23: Search list of features by name

5.3.3 Search features by ontology term name or accession

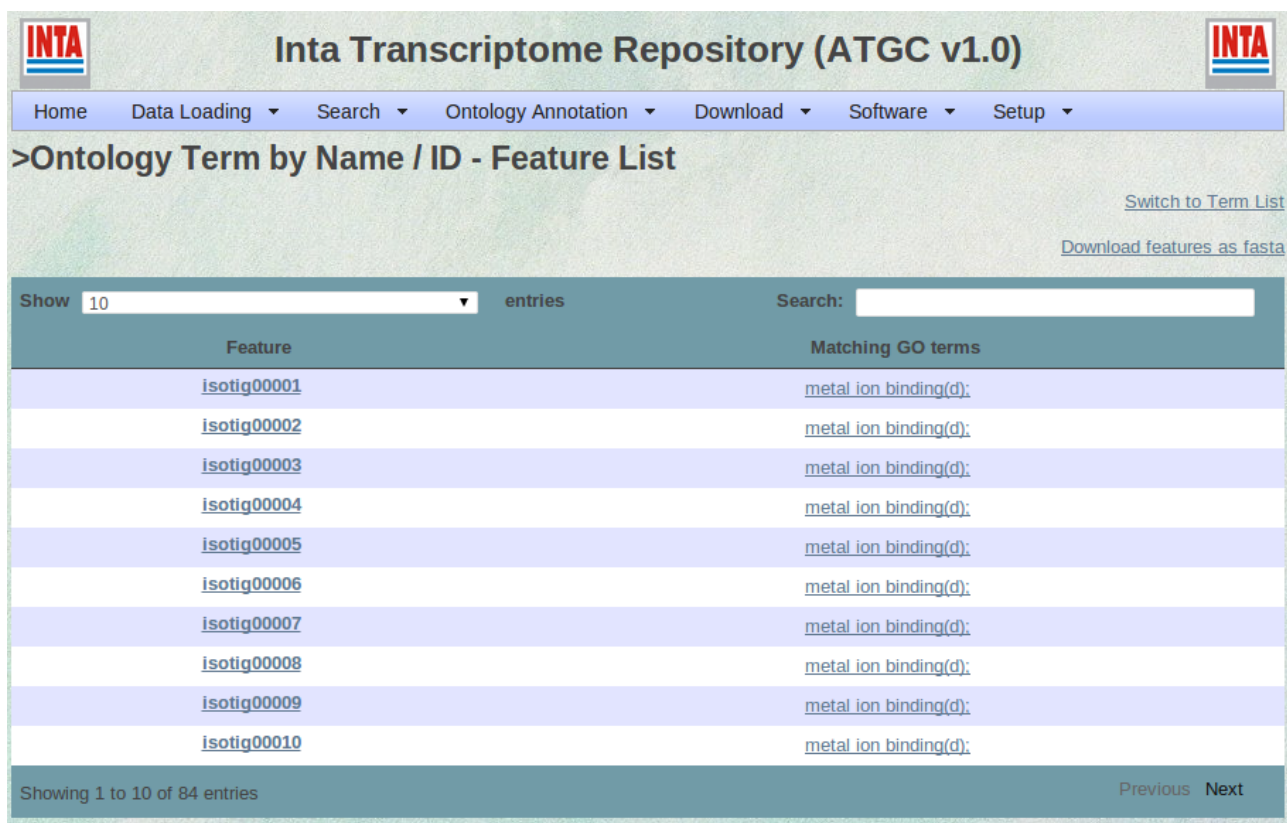
In the other way, you can search features by functional annotation, using ontology term name or accession, and then obtain the list of features annotated with this ontology term.

Figure 24: Search features by ontology annotation

The results of these searches can be seen on two formats, the list of features annotated (direct and indirect) called term list view or a summary called term list view, with only the amount of features of each condition.

Term name	GO ID	Direct Annotated Features	Indirect Annotated Features
metal ion binding	GO:0046872	84	0

Figure 25: Search by ontology, summary result

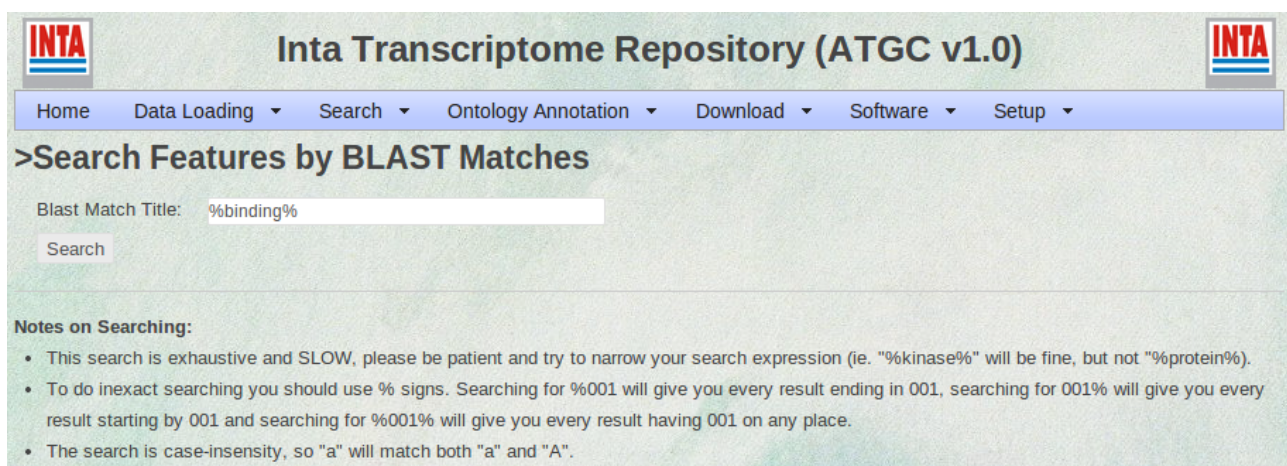


Feature	Matching GO terms
isotig00001	metal ion binding(d)
isotig00002	metal ion binding(d)
isotig00003	metal ion binding(d)
isotig00004	metal ion binding(d)
isotig00005	metal ion binding(d)
isotig00006	metal ion binding(d)
isotig00007	metal ion binding(d)
isotig00008	metal ion binding(d)
isotig00009	metal ion binding(d)
isotig00010	metal ion binding(d)

Figure 26: Search by ontology, feature list result

5.3.4 Search by Blast matches

The last way, is search through the headers or descriptions of blast results for each transcript, such as a more general and nonspecific form to relate features with functional annotation or similar sequences.



Notes on Searching:

- This search is exhaustive and SLOW, please be patient and try to narrow your search expression (ie. "%kinase%" will be fine, but not "%protein%").
- To do inexact searching you should use % signs. Searching for %001 will give you every result ending in 001, searching for 001% will give you every result starting by 001 and searching for %001% will give you every result having 001 on any place.
- The search is case-insensivity, so "a" will match both "a" and "A".

Figure 27: Search features by Blast matches

Inta Transcriptome Repository (ATGC v1.0)		
Home Data Loading Search Ontology Annotation Download Software Setup		
>Search Features by BLAST Matches		
Download features as fasta		
Show 10 entries	Search:	
Feature	Blast Matches	
	Score	Description
contig00012	676.00	ATP-binding cassette transporter, putative [Ricinus communis] >gi 223547739 gb EEF49231.1 ATP-binding cassette transporter, putative [Ricinus communis]
isotig00001	816.00	PREDICTED: chlorophyll a-b binding protein 3C, chloroplastic-like isoform 1 [Brachypodium distachyon] >gi 357132520 ref XP_003567878.1 PREDICTED: chlorophyll a-b binding protein 3C, chloroplastic-like isoform 2 [Brachypodium distachyon] >gi 357132522 ref
	816.00	PREDICTED: chlorophyll a-b binding protein 3C, chloroplastic-like isoform 4 [Brachypodium distachyon]
	816.00	PREDICTED: chlorophyll a-b binding protein 3C, chloroplastic-like isoform 1 [Brachypodium distachyon] >gi 357118766 ref XP_003561120.1 PREDICTED: chlorophyll a-b binding protein 3C, chloroplastic-like isoform 2 [Brachypodium distachyon] >gi 357118768 ref
	816.00	light harvesting chlorophyll a/b-binding protein Lhcb1 [Hordeum vulgare subsp. vulgare]
	816.00	putative chloroplast chlorophyll a/b binding protein [Solanum nigrum]
	816.00	chlorophyll a-b binding protein 3C-like [Solanum tuberosum]
	816.00	PREDICTED: chlorophyll a-b binding protein 3C, chloroplastic-like isoform 1 [Brachypodium distachyon] >gi 357132520 ref XP_003567878.1 PREDICTED: chlorophyll a-b binding protein 3C, chloroplastic-like isoform 2 [Brachypodium distachyon] >gi 357132522 ref
isotig00002	816.00	

Figure 28: Search by Blast matches results

5.4 Ontology exploration

You can explore the data across the ontology using the graph structure of terms and connections, showing a pie chart with the distribution of annotated features of each ontology term and a dropdown menu to move along the graph.

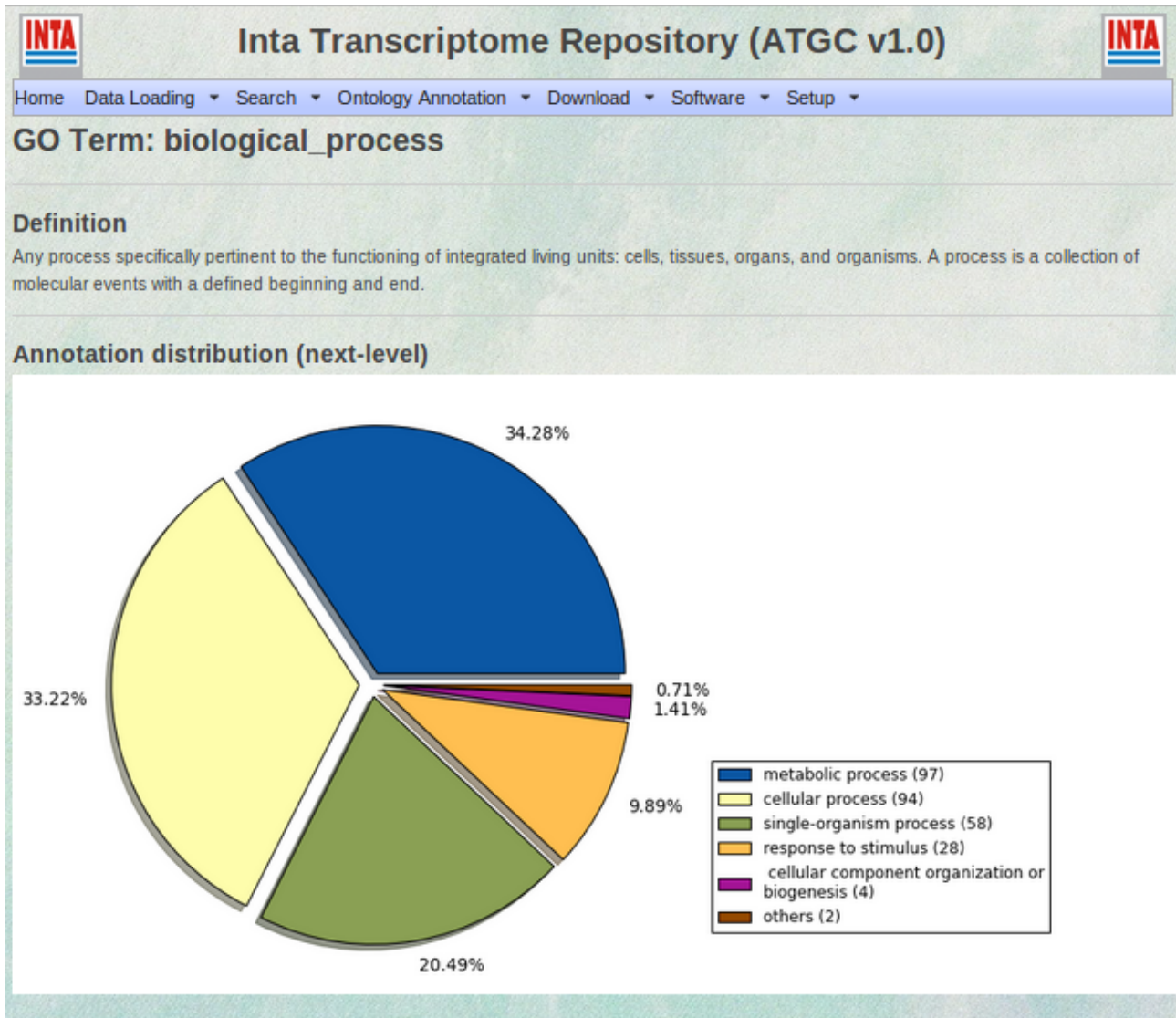


Figure 29: Pie chart: Biological process

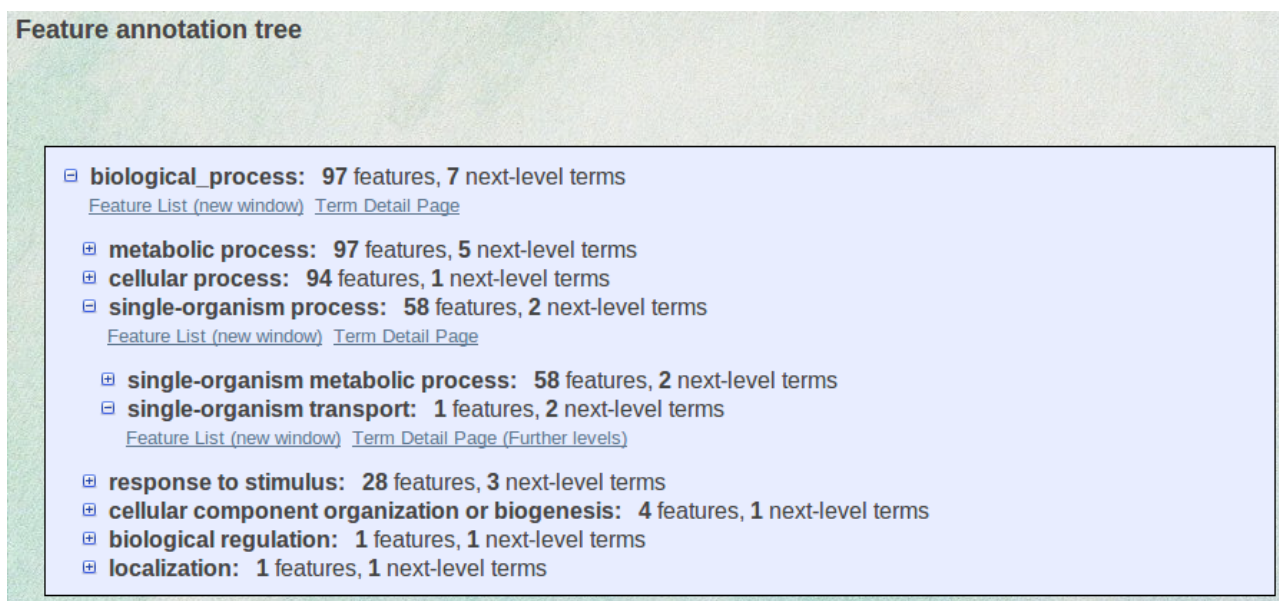


Figure 30: Feature annotation tree (Dropdown menu)

5.5 Download

Using the Download section, is possible to obtain sequences and annotations from database to text files to make other analysis, such as, functional enrichment. You can select the sequence type and download directly the file in fasta format or download the complete set of functional annotations in tabular format.

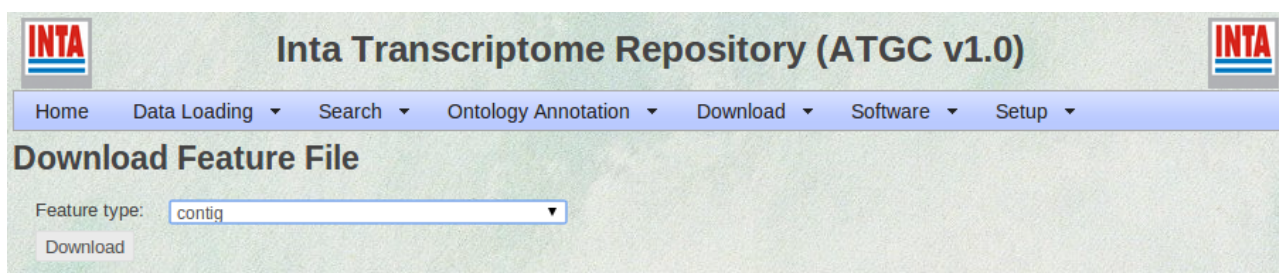


Figure 31: Download feature file

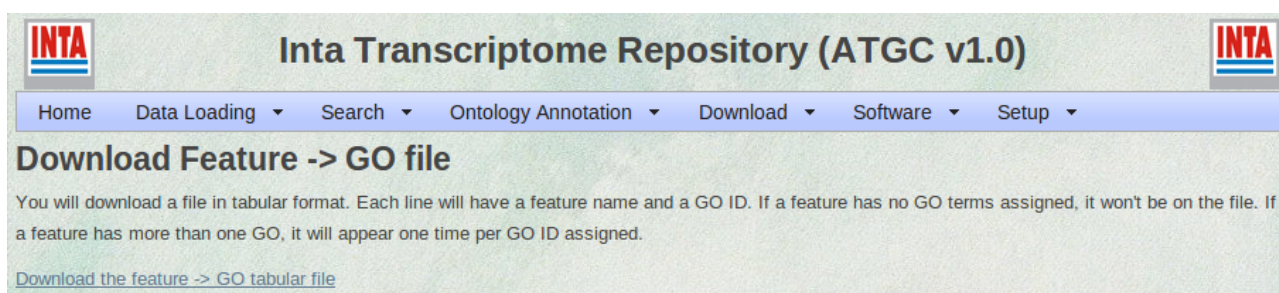


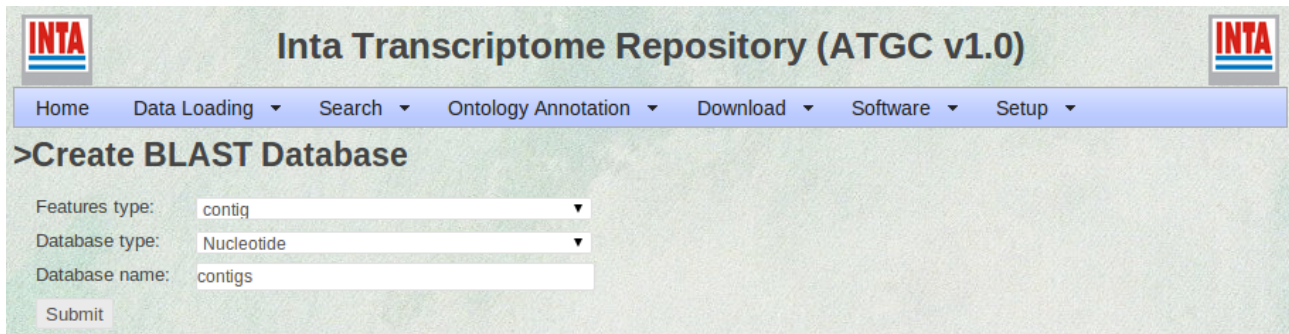
Figure 32: Download annotations

5.6 Software

Using the application is possible to make Blast alignments from sequences stored in the database as subject and external sequences as query.

5.6.1 Blast

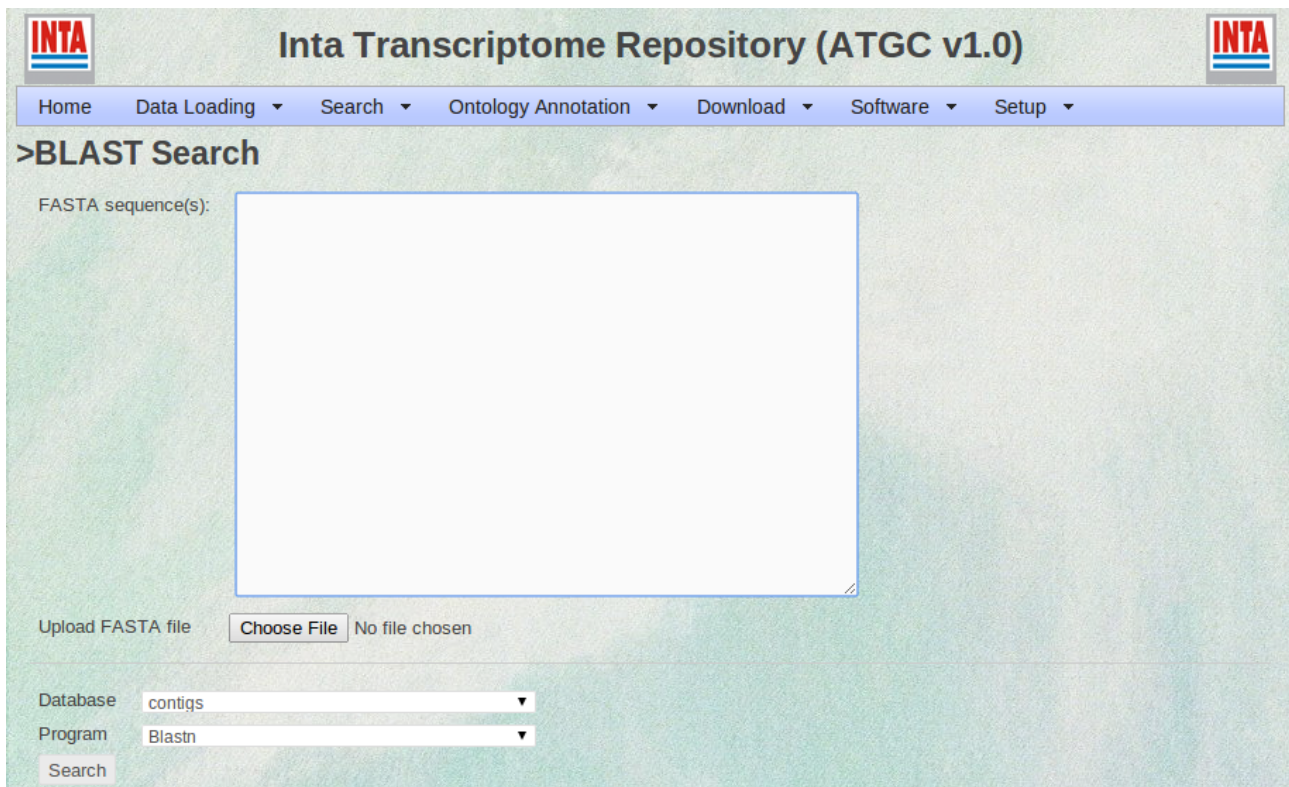
You can make blast alignments of query sequences using the sequences stored in the database as subject. The first step to make this alignments is to create a database choosing the sequences of interest from the database. For example in the Figure 33 the user is creating a Blast database called contigs that contain the sequences of all contigs stored in the database.



The screenshot shows the 'Create BLAST Database' form in the Inta Transcriptome Repository (ATGC v1.0) interface. The form includes a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The main heading is '>Create BLAST Database'. Below this, there are three input fields: 'Features type:' with a dropdown menu set to 'contig', 'Database type:' with a dropdown menu set to 'Nucleotide', and 'Database name:' with a text input field containing 'contigs'. A 'Submit' button is located at the bottom left of the form.

Figure 33: Create Blast database

With the database created, you can make the alignment using the option Software → BLAST → Run BLAST, in this place is possible to choose the type of Blast alignment among the options: BlastN, TblastN or TblastX. The query sequence or sequences can be inserted directly in the text box or using a text file.



The screenshot shows the 'BLAST Search' form in the Inta Transcriptome Repository (ATGC v1.0) interface. The form includes a navigation bar with links: Home, Data Loading, Search, Ontology Annotation, Download, Software, and Setup. The main heading is '>BLAST Search'. Below this, there is a large text area for 'FASTA sequence(s):'. Below the text area, there is a section for 'Upload FASTA file' with a 'Choose File' button and the text 'No file chosen'. At the bottom, there are two dropdown menus: 'Database' set to 'contigs' and 'Program' set to 'Blastn'. A 'Search' button is located at the bottom left of the form.

Figure 34: Run Blast alignments



Inta Transcriptome Repository (ATGC v1.0)



[Home](#)
[Data Loading](#)
[Search](#)
[Ontology Annotation](#)
[Download](#)
[Software](#)
[Setup](#)

>BLAST Search

FASTA sequence(s):

```
>sequence1
GGCCTCAACTACACCCTCTGGGCCAACCAACCACACCTTCTACGTCGGCGACTTC|
ATCTCGTTTAGGTACCAGAAAACGCAGTACAACGTGTCGAGGTGAACcAGACCGGCTAC
GACAACTGCACGACGGAGGGAGcgACCGGAACTGGAGCAGCGGCAAGGATTTTCATCCTC
CTCGACAAGGCCAAGCGgTACTtTTCATCTGCGGTAATGGCGGCTGCTTCAGCGGCATG
AAGGTTTCGGTGGTCGTCCaIGCGCTGCCGCGcGCCGCCAAGTCGTCGGTAGCCGTCAAG
gaCTCTTCGCTGTCGCCGcAgCGGCGCGTGGCTGGCACGGGGTAGCCgTGGGGgTTTTG
GgCGCGTTGGCcaccgcTTGGCTGTGACCGGTGCGGTGGGCCCAAGCGT
```

Upload FASTA file Choose File No file chosen

Database contigs

Program Blastn

Search

Figure 35: Example of Blast alignment



Inta Transcriptome Repository (ATGC v1.0)



[Home](#)
[Data Loading](#)
[Search](#)
[Ontology Annotation](#)
[Download](#)
[Software](#)
[Setup](#)

>BLAST Search

[Download BLAST pairwise output file](#)

Query Name	Subject Name	Identity	Alignment Length	E-value	Query Frame	Subject Frame
sequence1	isotig00001	100.00	403	0.0	1	1
sequence1	isotig00002	100.00	403	0.0	1	1
sequence1	isotig00003	100.00	403	0.0	1	1
sequence1	isotig00004	100.00	403	0.0	1	1
sequence1	isotig00005	100.00	403	0.0	1	1
sequence1	isotig00006	100.00	403	0.0	1	1
sequence1	isotig00007	100.00	403	0.0	1	1
sequence1	isotig00008	100.00	403	0.0	1	1
sequence1	isotig00009	100.00	403	0.0	1	1
sequence1	isotig00010	100.00	403	0.0	1	1

Showing 1 to 10 of 24 entries
Previous Next

Figure 36: Example of Blast result

5.7 Modify and delete

You can modify and remove data stored in the database using the section “Modify and Delete”, having two options to select the work entry: Using the entry name or selecting the entry from a complete list of features.

Figure 37: Search object to modify

feature_id	dbxref_id	organism_id	name	uniquename	residues
1	None	Inta	contig00004	contig00004	TACCTGGTTGATCCTGCCAGTAGTCATATG...
10	None	Inta	isotig00005	isotig00005	CTCCTCCTCTCTCCcGCCGCCGCCACC...
100	None	Inta	isotig00095	isotig00095	CGAAAGGTTTATAATATTAATACTTtGG...
1000	None	Inta	gi 357497757 ref XP_003619167....	blast_match: isotig00029 to gi...	None
1001	None	Inta	gi 238858956 dbj BAH70299.1 c...	blast_match: isotig00029 to gi...	None
1002	None	Inta	gi 385298939 gb AFI60242.1 li...	blast_match: isotig00029 to gi...	None
1003	None	Inta	gi 242088861 ref XP_002440263....	blast_match: isotig00029	None

Figure 38: Table to select object to modify

INTA

Inta Transcriptome Repository (ATGC v1.0)

INTA

Home

Data Loading ▾

Search ▾

Ontology Annotation ▾

Download ▾

Software ▾

Modify and Delete ▾

Setup ▾

Update Data

library_id	Actual value:	11	
organism_id	Actual value:	Inta	New value:
name	Actual value:	line2_control_1	New value:
unique_name	Actual value:	line2_control_1	New value:
stock_id	Actual value:	line2	New value:
type_id	Actual value:	RNA-Seq_library	New value: line1 line2
is_obsolete	Actual value:	None	New value:
timeaccessioned	Actual value:	2016-03-31 13:42:51.197663	New value:
timelastmodified	Actual value:	2016-03-31 13:42:51.197663	New value:

Update Data

Figure 39: Update form to modify values

6 Share information

To share information, you can use a Web2py internal web server or install and configure apache web server, in any way you can choose that part of menu bar is visible to other people as follows:

Go to the follow direction in the web browser and enter the password choosen when you run Web2py (Web2py administrative interface, only available whit the Web2py web server):

localhost:8000/admin



The screenshot shows a web browser window with the address bar set to `localhost:8000/admin`. The page title is "web2py™ administrative interface". The main heading is "web2py™ Web Framework", followed by the sub-heading "Login to the Administrative Interface". Below this is a login form with the label "Administrator Password:" and a text input field. A "Login" button is positioned below the input field. At the bottom of the page, a footer line reads "Powered by web2py™ created by Massimo Di Pierro ©2007-2016 - Admin language" followed by a dropdown menu currently showing "English (US)".

Figure 40: Web2py administration interface: write a password

In the first page, you must click on “Manage” and choose “Edit” option

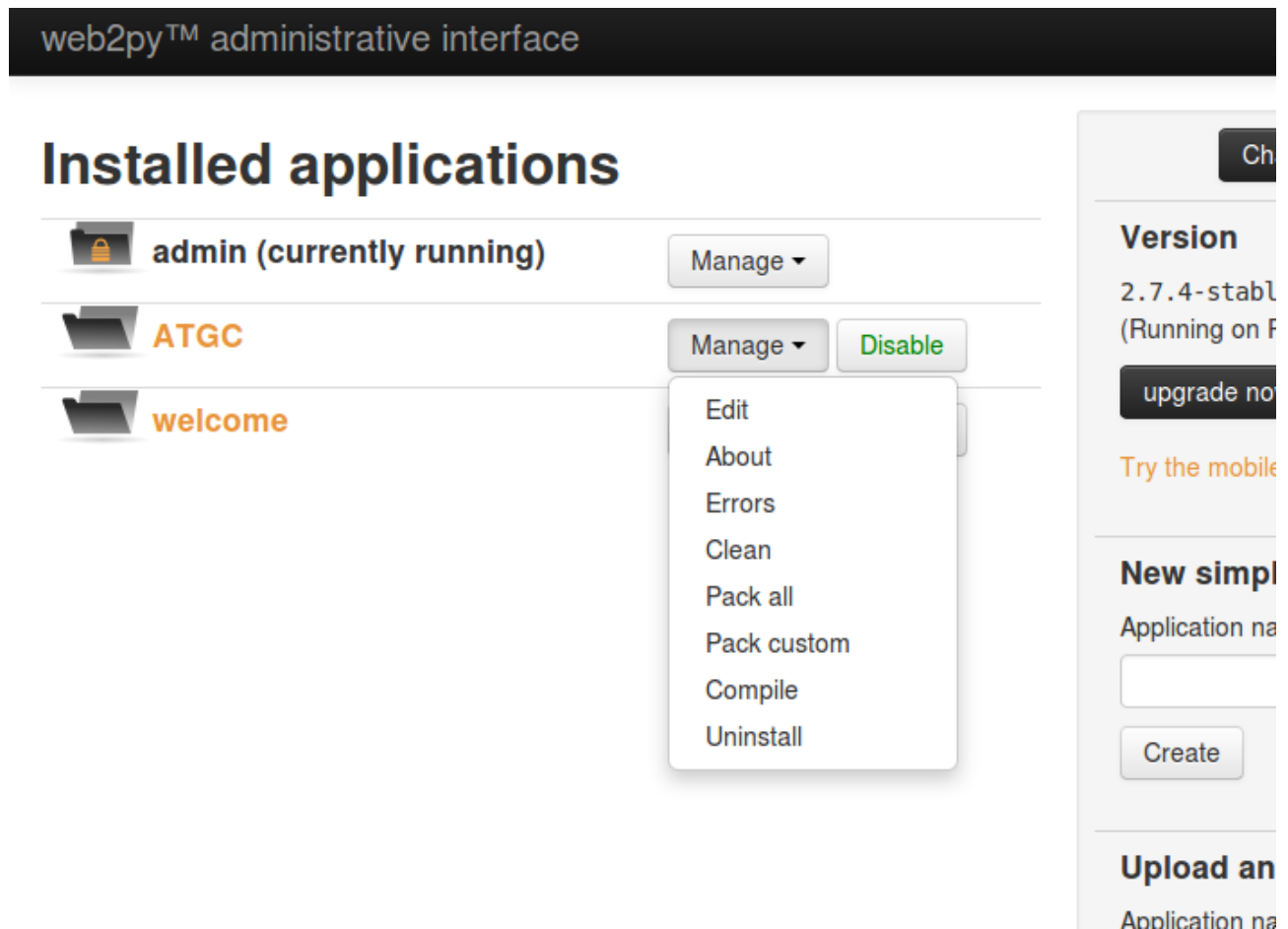


Figure 41: Choosing "Edit" option

Then, click in “database administration” button below of “Models” title

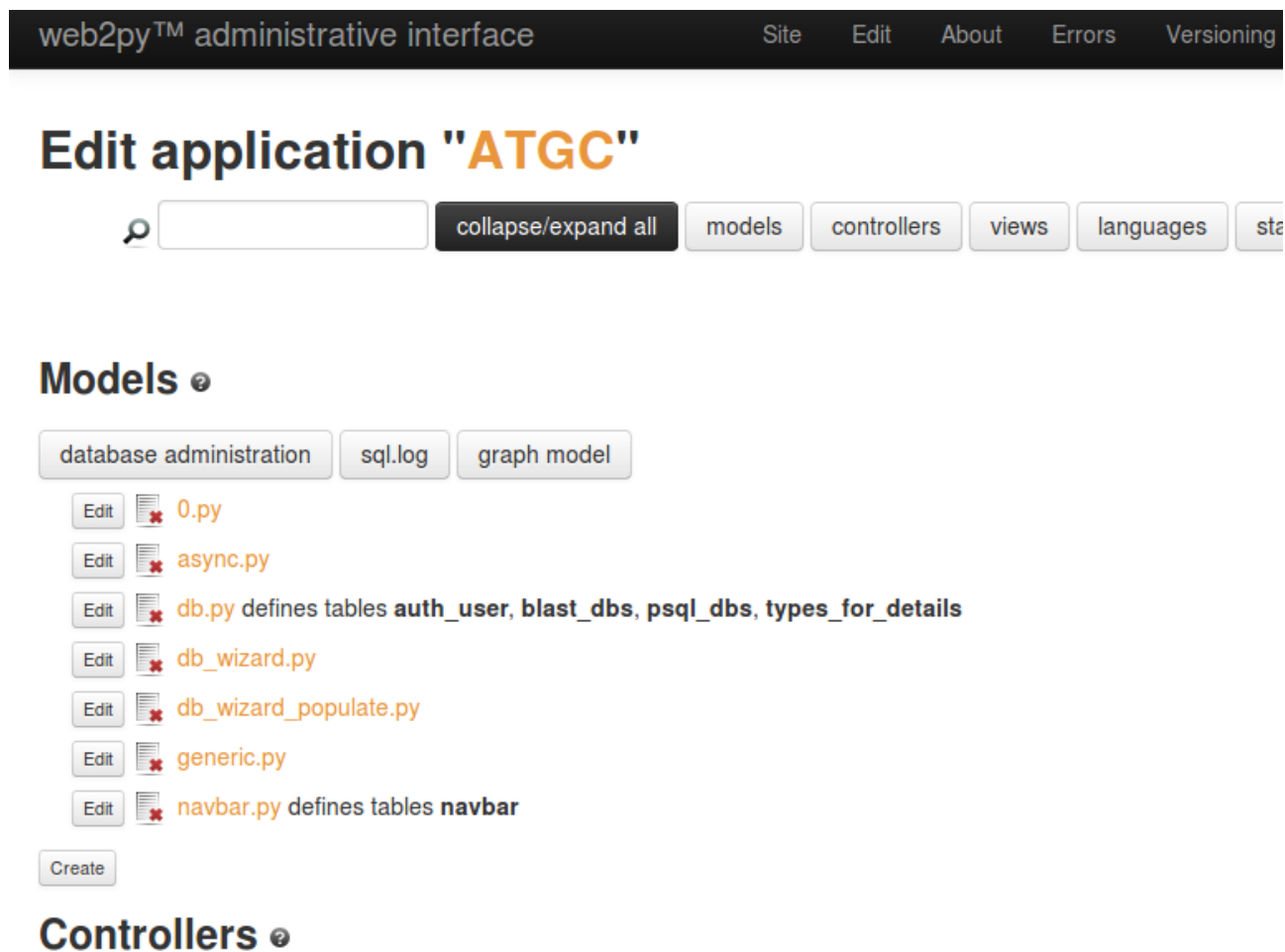


Figure 42: Click on "database administration" button

Choose the last option, db.navbar:

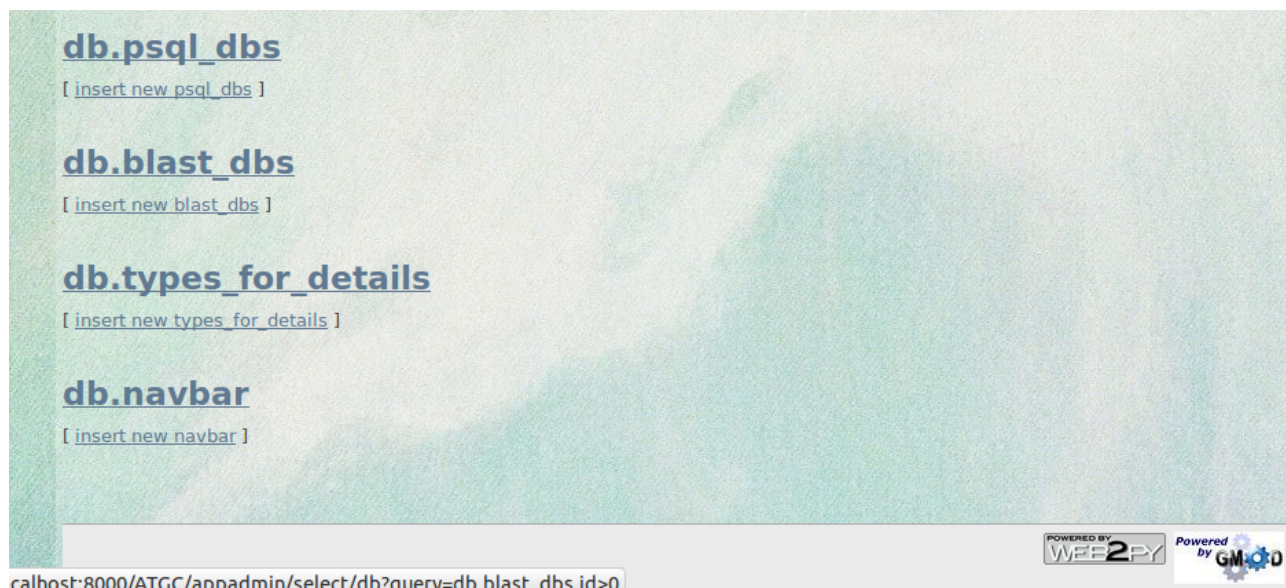


Figure 43: Click on db.navbar option

In the next page, you can see the table to set up the menu bar, at this point, you can choose which part of menu is visible or not, setting on “True” or “False” the field called “is_visible” for the correct table entry. For example, to hide the “Setup” menu you must complete the fields as the next figure:

database db select

[[insert new navbar](#)]

Rows in table

Query:

Update: ☒

Delete: ☐

The "query" is a condition like "db.table1.field1=='value'". Something like "db.table1.field1==db.table2.field2". Use (...)&(…) for AND, (...)|(…) for OR, and ~(…) for NOT to build more complex queries.

"update" is an optional expression like "field1='newvalue'". You cannot update or delete the results of a JOIN

Figure 44: Hide the "Setup" part of the menu

For others main parts of the menu bar, the next table has a id value, you can hide any option of the menu, such as, complete part (for example, all Setup menu) or a single part (for example, only):

Menu option	Id (db.navbar)
Data Loading	2
Search	18
Ontology Annotation	23
Download	33
Software	36
Modify and Delete	42
Setup	63

Table 1: Id values for main menu bar options in db.navbar table

7 Testing the application

Inside the 'web2py/applications/ATGC/private/test' directory you have a test dataset to try the application, this dataset is composed by the following files:

- go.obo: Reduced gene ontology (obo format)
- so.obo: Complete sequence ontology (obo format)
- interpro2go: File with relationships between gene ontology and InterProScan terms
- test.png: Image to load when you create an organism
- transcripts.fasta: Multifasta file with sequences of 100 test transcripts
- transcripts.raw: Functional annotation results of InterProScan software
- transcripts.xml: Blast results
- transcripts.annot: Functional annotation results of Blast2GO software
- transcripts.gff3: Structural annotation of transcripts (gff3 format)
- genes.list: List of gene names to load with “Load features from list file (without sequence)”
- transcripts_gene.csv: csv file with relationships between genes and transcripts to load with “Feature relationships”
- transcripts_SNP.vcf: SNPs in transcripts
- transcripts_snps_line2.csv: Alleles of SNPs in line2
- transcripts_SSR.vcf: SSR in transcripts
- transcripts_ssrs_line2.csv: Alleles of SSRs in line2
- experiment_design.txt: Experiment design schema to load expression data (view 5.2.6 section in this tutorial) and create assay structure in the database
- lineX_control_Y.exp (X=1,2 and Y=1,2): RPKM expression values for control samples
- lineX_treated_X.exp (X=1,2 and Y=1,2): RPKM expression values for treated samples