Manual Galaxy Pipeline

Access to the Galaxy pipeline

The Galaxy instance can be accessed via the following link: 10.42.1.212:8080.

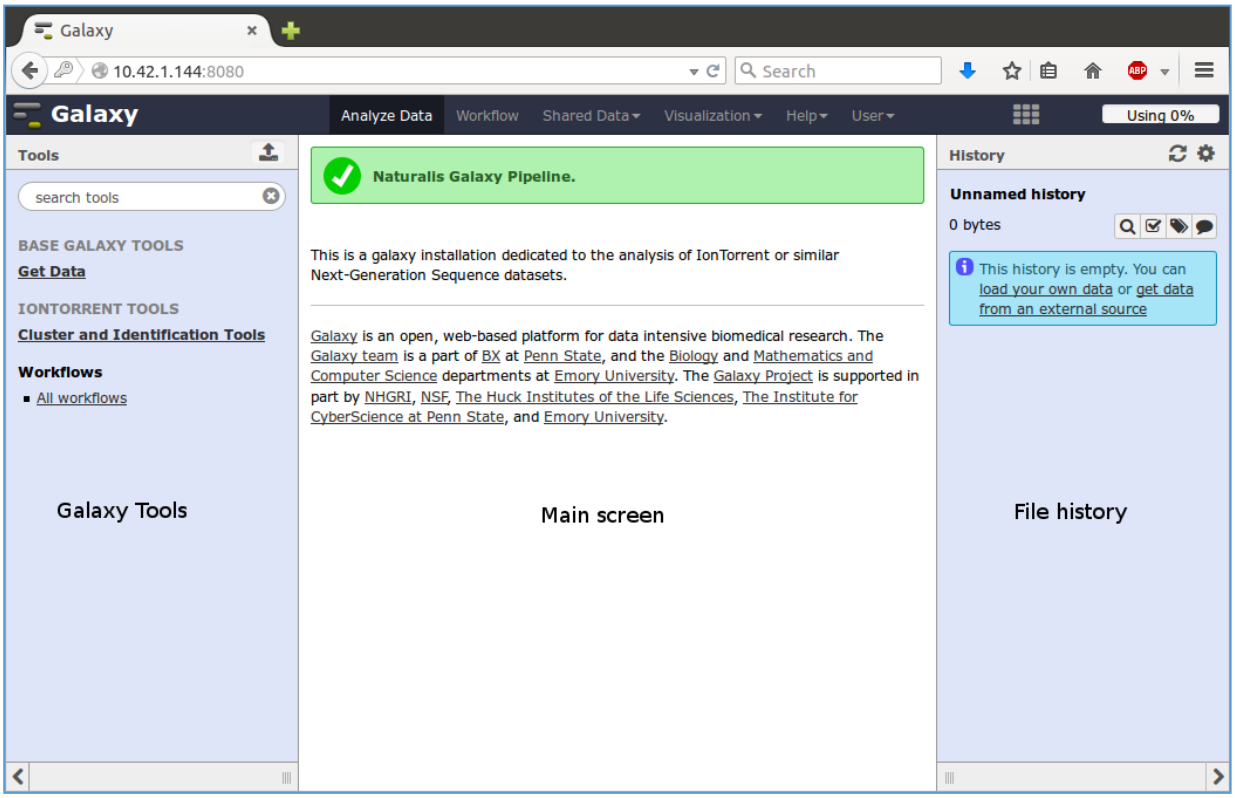
Note: Currently the Galaxy instance only accessible from within the Naturalis network.

Creating an account

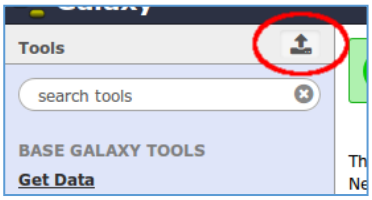
The first thing that needs to be obtained is an account for the Galaxy server. To start the account creation click on: “you may create one” on the login page. This link will send you to a form where you can fill in your e-mail, display name and password. After submitting the account will be created and you will be automatically logged in and send to the Galaxy main menu.

Layout

The layout of Galaxy is divided into three areas. On the left you'll find the available tools, on the right the user history that will contain all the files for a certain project and the centre will display general messages, tools forms or displayed data.



Loading data

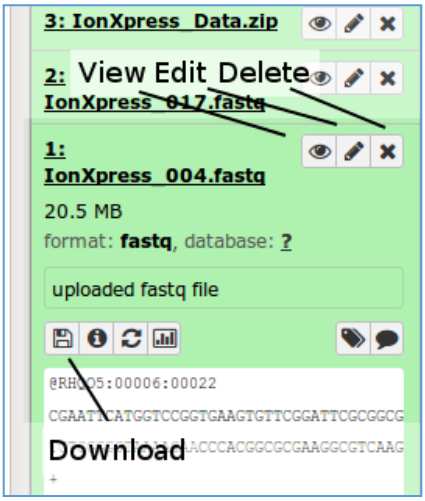


The quickest way to load your data into Galaxy is via the “up arrow” on the tools section of the Galaxy layout. You can either drag and drop files into the upload screen, or use the file browser to add files for uploading.

Note: The Galaxy upload will automatically try to determine the file type of the uploaded files, however this is not working correctly for comma separated value files (CSV) and ZIP archives, you can click on “type” right of the file name and manually set it to the correct format.

After selecting a number of files you can click on “start” on the bottom of the screen. After uploading the files will appear in you history on the right side of the screen.

Data management



Each file in the history has its own info “box”. By clicking on a file name the box opens and shows the file type, file size, a preview of the data (if available) and downloading the file back to the computer (the floppy disk symbol).

The three symbols right of the file name can be used to display the data in the centre screen (eye symbol), edit file attributes (such as the name) (pencil symbol) or delete the file from the history (cross). Renaming files can be helpful when running multiple analysis on the same file or when dealing with obfuscated file names (Such as the default IonTorrent output).

ZIP archives

Most tools for analysing IonTorrent data accept both single files and ZIP archives. ZIP files can contain multiple IonTorrent input files. If a ZIP archive is used as input for a tool, the tool will be ran for each of the files in the archive, with the output returned to the user as a new ZIP archive. The advantage of using ZIP archives is that multiple files can be analysed simultaneously while keeping the history free from heaps of output files.

ZIP archives can be managed with the “Manage ZIP” tool in the “Cluster and identification tools” section of the tool menu. Several action for ZIP management are available:

- Display ZIP content. List the files in the ZIP (default), a new file is created in the history that contains a list of the files in the ZIP.
- Create ZIP / Add items to ZIP. Multiple files from the history can either be added to an existing ZIP or a new ZIP.
- Unpack items from ZIP. Extract either specific files or all files from a zip to the history.
- Delete items in ZIP. Delete specified files.
- Rename items in ZIP. Rename files in the ZIP.
- Create subset from ZIP. Create a new ZIP containing a subset of the files from an existing ZIP.

Analysing a dataset

Demultiplexing

For demultiplexing a sequence file use the “Split on Primer” tool which can be found in the tools menu under the section: “Cluster and Identification Tools”.

Set the input type to the file type of your input file (either FASTA, FASTQ or ZIP), select the sequence file, the primer file type (primers are either provided in a CSV file or via a text form), the primers, the maximum number of mismatches allowed between the sequences and the and if the primers need to be trimmed from the sequences after demultiplexing.

After the tools has been ran several files are created in the history: a log file which should be empty and can be deleted and multiple sequence output files (one output file for each input primer, plus a file for unmatched sequences).

Sequence trimming and filtering

2 tools are available for trimming and filtering of sequence data: the “Sequence analyser” tool and the “Sequence trimmer” tool. The analyser tool can be used to get insight in the quality and length of the sequence files, the trimmer tools does the actual trimming and filtering of the sequences.

To analyse a sequence file select the “Sequence analyser” tool. Select the type of file you want to analyse (Single sequence file or ZIP archive) and the format of the data (FASTA or FASTQ), start the tools by clicking on “Execute”. After the tool is finished a HTML file will appear in the history containing the analyse output graphs, the data can be displayed by clicking on the “eye” symbol. The output graphs can be used to determine the appropriate filter and trim settings.

Trimming the data happens via the “Sequence trimmer” tools. The file type and sequence format needs to be selected for the input file and the output sequence format needs to be set.

Note: the cluster and BLAST tools downstream of the pipeline require input files in the FASTA format (the default output type of the Sequence trimmer tool). Output in FASTQ can be set if the desired for other tools outside the Galaxy pipeline.

A range of filter and trim settings is available, by default all these settings are ignored (when the value is set to zero) but can be adjusted adjusted to the user requirements.

Note: Not every setting has to be set, some values can be left to the default settings while other are changed.

Note: The Sequence trimmer tool filters and trims the sequences in the same order as the settings are listed in the form, this means for example that the sequences are first trimmed before filtered based on minimum or maximum length.

The output of the trim tool (in the selected output format) is listed in the history.

Clustering and filtering

Clustering of the data can happen via two "different" methods, separate clustering and filtering or combined clustering and filtering. The "Cluster and filter" tool clusters the input data and immediately filters the output based on minimum cluster size, the separate "Cluster" and "Cluster filter" tools allow the user to check the output clusters before deciding if they require filtering.

Note: The "Cluster" and "Cluster filter" tools can only be used on single input files and not on ZIP archives.

To separately cluster and filter the clusters first the "Cluster" tool needs to be selected. A FASTA file has to be selected for clustering together with a cluster threshold (if the sequence similarity between two sequences equals or is larger than the threshold, the sentences are placed in the same cluster).

After the tool is finished 3 output files are produced in the history: A cluster "Stats" file (contains information on the cluster size and origin of the sequences), the "Clustered" file (contains the representative sequences for each cluster) and the "Histogram" file (contains a histogram with the number of clusters with per cluster size). The histogram tool can be used to select a minimum cluster size for the "Cluster filter" tools (if desired).

The "Cluster filter" tool can be used to remove small cluster (for example singletons). This tools requires the cluster "Stats" and "Clustered" output generated by the "Cluster" tool, together with a minimum cluster size set by the user. The tools will generate a "Clustered Filtered" file containing the cluster representative sequences for clusters above the minimum size.

The "Cluster and filter" tools is similar to the "Cluster" tools but besides a FASTA file and a cluster threshold a minimum cluster size needs to be provided before clustering. The output "Clustered" file only contains the clusters above the minimum size. This tool is convenient for bulk processing (ZIP archives) and if you have a general idea on the cluster output desired (no singletons for example).

Identification

BLAST

The “BLAST” tool can be used to BLAST a set of FASTA sequences against either the online NCBI GenBank databases or a collection of local database (see table).

Local Database	Online Database
GenBank Nucleotide (nt)	GenBank Nucleotide (nt)
Filtered BoLD (fully annotated sequences)	GenBank mRNA set
Unfiltered BoLD (all BoLD sequences)	GenBank Genome database
UNITE Fungal ITS database	GenBank EST database
SILVA 16S and 18S database	GenBank Environmental database
SILVA 23S and 28S database	

For local BLASTing blastn is automatically used, for online BLASTing different blast algorithms are available (blastn, blastp, blastx, tblastt, tblastx). To run the BLAST tools a input FASTA file or ZIP archive needs to be selected, followed by a selection of either online or offline reference databases (multiple offline database can be selected for a single run) and algorithms and a selection of BLAST thresholds (minimum sequence similarity for a match, minimum length, maximum e-value and the maximum number of results).

Note: Always try to run a local BLAST rather than an online BLAST. Online BLASTing has been capped to a maximum of a 100 sequences per run, with a maximum of 20 hits per sequence.

The BLAST output is a tab separated value file (TSV) that that lists the blast hits for each input sequence. The file can be displayed in the browser (via the eye symbol) but should preferably be downloaded and opened in spreadsheet software such as Excel.

HTS barcode checker

The “HTS barcode checker” tool is similar to the BLAST tools. A set of sequences is BLASTed against the GenBank nucleotide database (either online or offline) however the BLAST results are checked for the presence of CITES protected species.

Similar to the BLAST tools a input FASTA file has to be selected, together with either the BLAST reference database (local or online) and a set of BLAST thresholds (minimum sequence similarity, minimum match length, maximum e-value and maximum number of hits). Different is the option to provide a CITES database (lists the CITES protected species), Blacklist database (known erroneous GenBank sequences) and additional CITES database (contains species not listed on CITES but are required to be flagged regardless).

Note: The Galaxy pipeline has its own copy of the CITES appendix, if none is provided is the tools automatically falls back to its own copy and returns this together with the BLAST results to the user.

The tools produces a TSV file similar to the BLAST tools. The TSV contains the default BLAST information, but has three additional columns if the BLAST hit is from a CITES protected species containing: the CITES appendix, CITES species and additional CITES information (if available).

Expand BLAST

The TSV output of a BLAST search based on clustered sequences can be expanded with the "Expand BLAST" tool. This tools adds an additional column to the output of the BLAST or HTS barcode checker tools that contains the number of sequences present in the cluster.

The user needs to provide the BLAST output TSV file and the cluster "Stats" file from the clustering. The tools will produce a new TSV file containing both the BLAST results and the cluster size.

Histories

When working on multiple projects it is easy to lose track of the different data files and outputs, a convenient way to organise the data is by using multiple histories. The default history is named "Unnamed history" (see upper right corner), this can be changed by clicking on it and replacing it with a new name. A new history can be created by clicking on the "cog" symbol in the upper right corner, this will prompt a menu with the option: "Create new", which will create a new empty history if selected. You can swap between different histories on your account by clicking on the cog and selecting: "Saved histories".