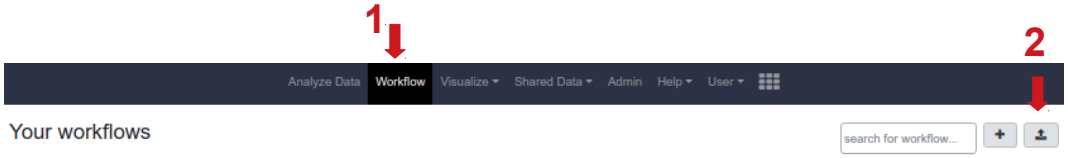
Galaxy Tutorial

This tutorial introduces the main functions of the Galaxy-distribution for RNA-analysis and modification calling. This includes a short manual for the general usage of the Galaxy-environment and the available workflows.

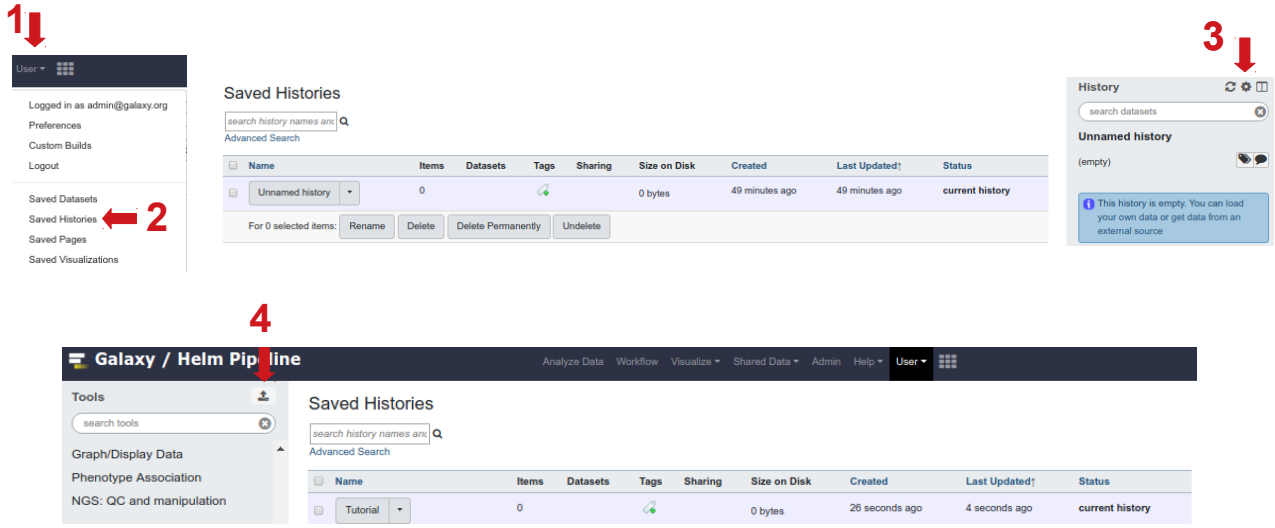
1. Upload workflows

After starting the Galaxy environment, click on the „Workflow“-tab (1) at the top of the window. Then click on the „Upload or import workflows“ (2) button. You can then browse for the workflows which are located in the workflows-folder that came with the download and upload all workflows. By selecting the „Show in tools panel“-option, the workflows can directly selected from the bottom of the tools panel (left side of your window). This allows to start the workflows a little bit faster.



2. Creating a new history and uploading files

The Galaxy environment allows you to organize different projects within so-called „histories“. An overview of all existing histories can be found by clicking on the „User“-tab (1) and selecting „Saved Histories“ (2). To create a new history, click on the „History options“-button (3) and selecting „Create new“. You will automatically switch into the new unnamed history. To upload files into your current history, click on the Download/Upload button (4) and select „Choose local files“. For this tutorial, use the test data which was included in the download (make sure the files are unzipped). Upload „total_tRNA_yeast_untreated_R1.fastq“ (forward read), „total_tRNA_yeast_untreated_R2.fastq“ (reverse read) and the yeast tRNA reference in fasta-format by selecting them and pressing „Start“. The uploaded files will show up on the panel on the right.



3. Analysis – Standard Workflow

To start analysis first click on the „Analyze Data“-tab, then select the „RNA_Seq_Standard_Workflow“ from the workflow-tab or from the tools-panel, this will open a window showing all steps of the workflow. Here, you have to select the forward read, reverse read and the reference file, respectively. You can also change multiple options for each individual step. The default options are optimized for the test data. After selecting the input files, click on „Run workflow“ to start the analysis. The progress of the workflow is shown in the History-panel on the right. The standard-workflow is the basis for further analysis and concludes with the creation of a „Profile“ (shown in the History-panel as „Pileup2ProfileV5“). The Standard Workflow serves as the basis for all other workflows and functionalities as they require a file in Profile-format.

4. Machine Learning and prediction

Upload the „Known_m1As_yeast“-file from the test data. Galaxy will automatically label the file as .txt. Next, select „Machine_Learning“ workflow and select the uploaded file and the profile-file, respectively. If the uploaded file cannot be selected, the file format has to be changed. This can be done by selecting the file and clicking on the „Edit attributes“-button (1). In the next window, under the „Datatypes“-tab (2), „New type“ (3) the file format „csv“ can be selected, which is required by the workflow. Press „Change datatype“ (4) to start the conversion. After this, run the workflow. In the default settings, the machine learning algorithm uses all available features except for the A-mismatch, which is irrelevant for detection of adenosine-based modifications.

The output contains the trained machine learning model in pkl-format. For the prediction process, it is necessary to select the exact same features. Otherwise, the results will be inaccurate. The best way to ensure that the correct features are selected in the prediction process is to save the information in the file name. This can be done via the “Edit attributes”-button. The prediction workflow requires a pkl-file and a profile-file. The two output files contain the predictions, separated into positive and negative class (e.g. m^1A/non-m^1A).

5. Visualization

The “Visualize_V3” workflow allows to plot modification-patterns in sequence context. This workflow allows the user to enter the sequence of interest. The name of the reference sequence can be found within the profile-file, where it is the first entry in each line. The starting and ending interval (e.g. leftmost and rightmost position respectively) also have to be entered. If one of these positions is not present in the sequence (for example selecting an ending interval of 1000 for a sequence containing only 70 bases), the workflow will produce an error. The remaining attributes can be used to adjust the plots configuration. To view the resulting plot, click the download button and open the plot in your download-directory.

6. Creating workflows and installing other tools

Besides the available workflows introduced here, Galaxy allows for the creation of additional workflows. This functionality can be accessed through the workflow-tab. If an analysis requires the use of additional bioinformatic tools, the Galaxy tool shed offers a great variety of software. The tool shed is accessible through the “Admin”-tab by clicking on the “Install new tools” option (Note: This action requires the addition of your nameserver to the resolv.conf file, as described in the installation instructions.).

7. Data management

The Galaxy-environment comes with an option to store files in Data-libraries which can be managed by the user. This option is especially useful for storage of heavily-used files (for example: reference genomes, trained machine learning models) and allows for fast adding of these files into new histories. Data-libraries can be accessed through “Shared data” → “Data libraries”.