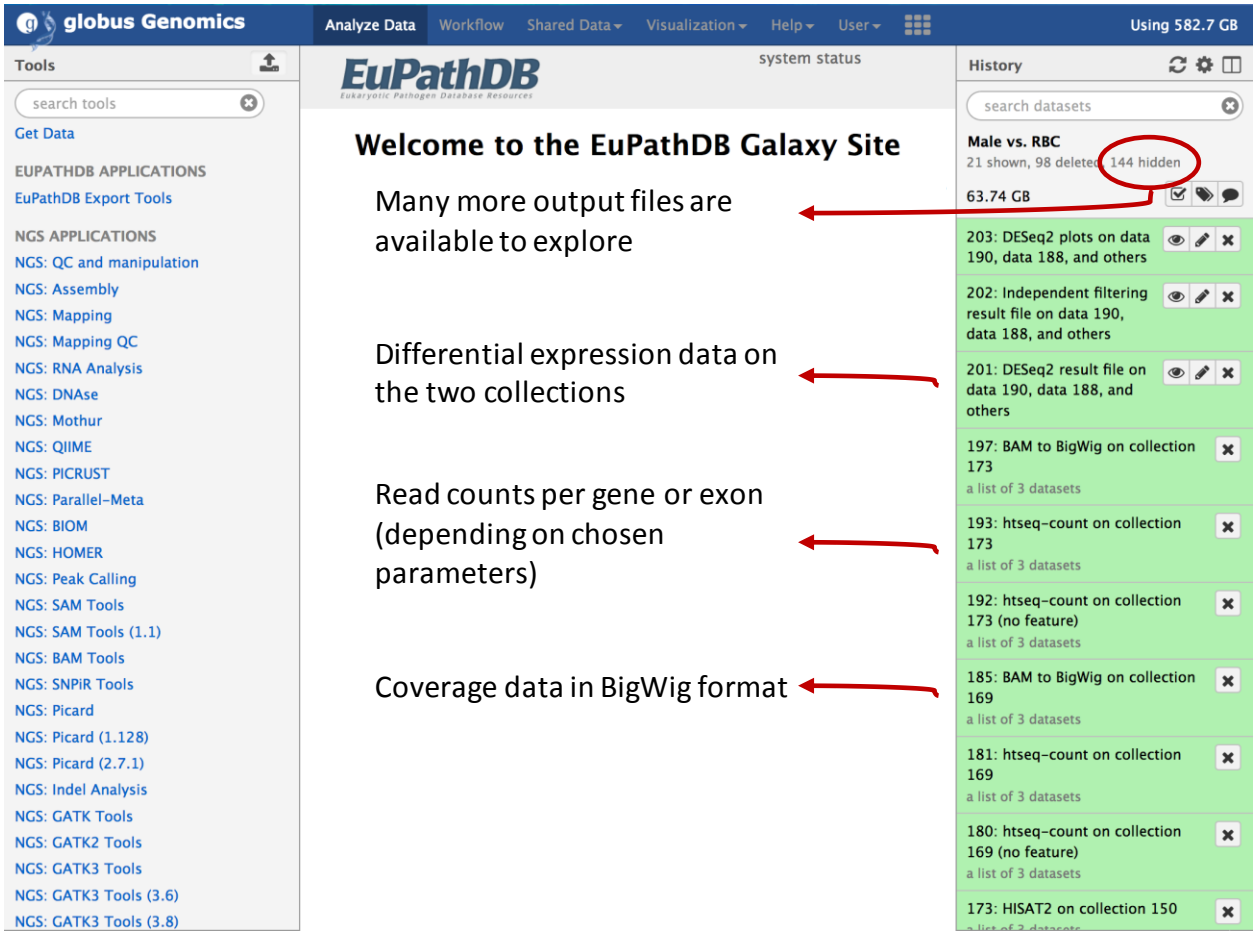
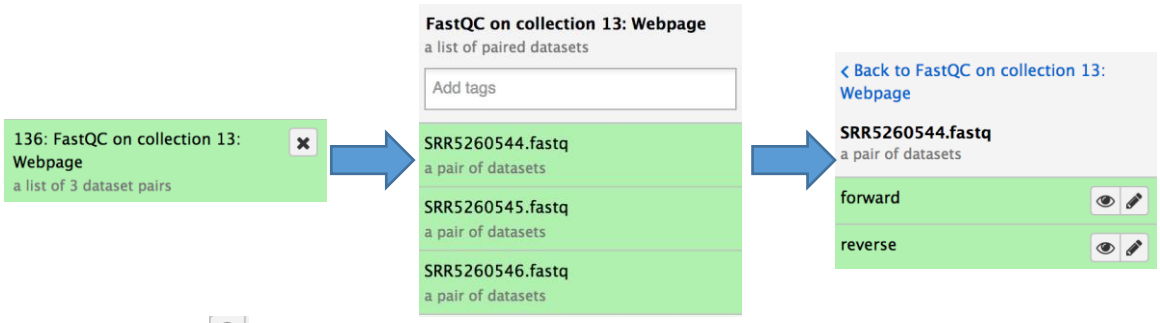
RNA sequence data analysis via Galaxy, Part II Uploading data and starting the workflow (Group Exercise)

The goal of this exercise is to examine the results from the Galaxy RNAseq analysis workflow that ran overnight. If everything worked out you should see a list of completed workflow steps (Green). The workflow generates many output files, however not all of the output files are visible. You can explore all the hidden files clicking on the word "hidden" (red circle) – this will reveal all hidden files.



Step 1: Explore the FastQC results. To do this find the step called "FastQC on collection ##: Webpage". Click on the name this will open up the FastQ pairs, click on one of them then click



on view data icon () on either forward or reverse. Note that each FastQ file will have its own FastQC results. An explanation of each of the FastQC results is provided as a link on the main workshop website or at the bottom of the FastQC results page.

SRR5260544_1.fastq.gz FastQC Report

FastQC Report
Tue 12 Jun 2018
SRR5260544_1.fastq.gz

Summary

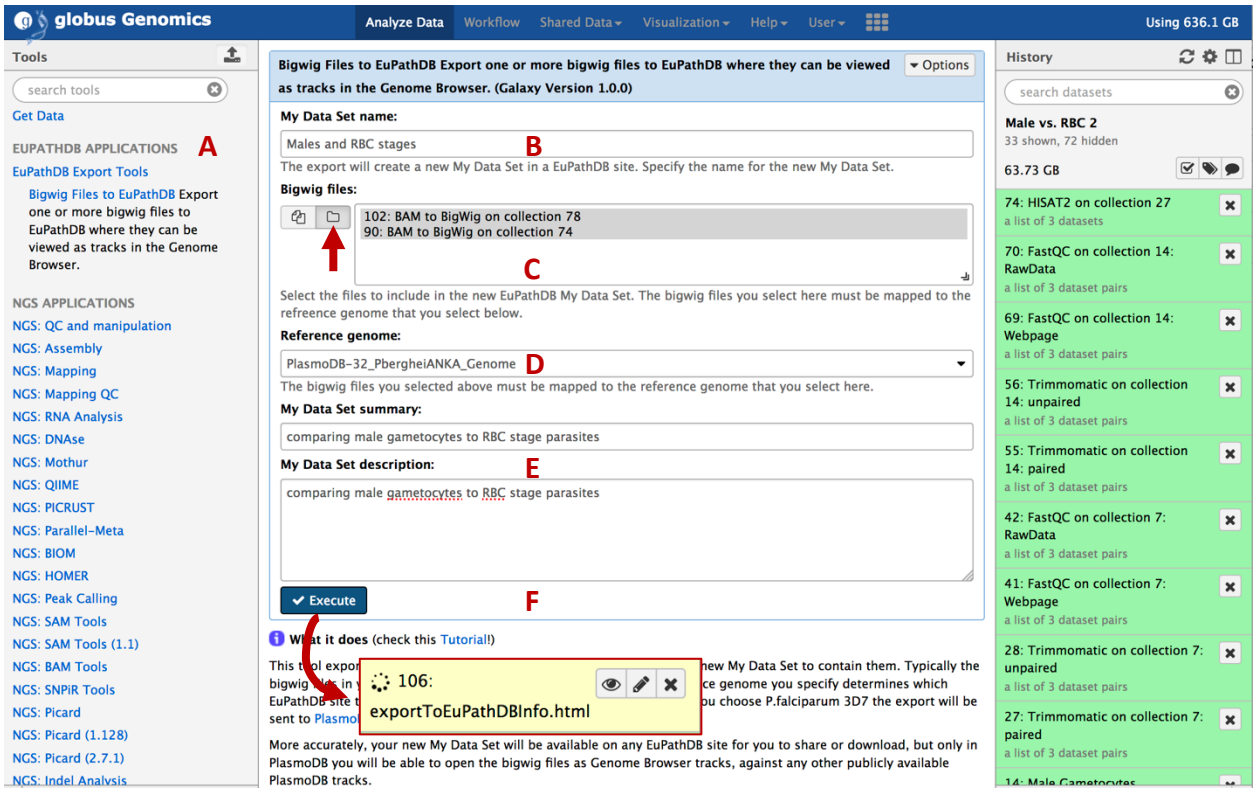
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	SRR5260544_1.fastq.gz
File type	Conventional base calls

Step 2: Displaying coverage results in the EuPathDB genome browser:

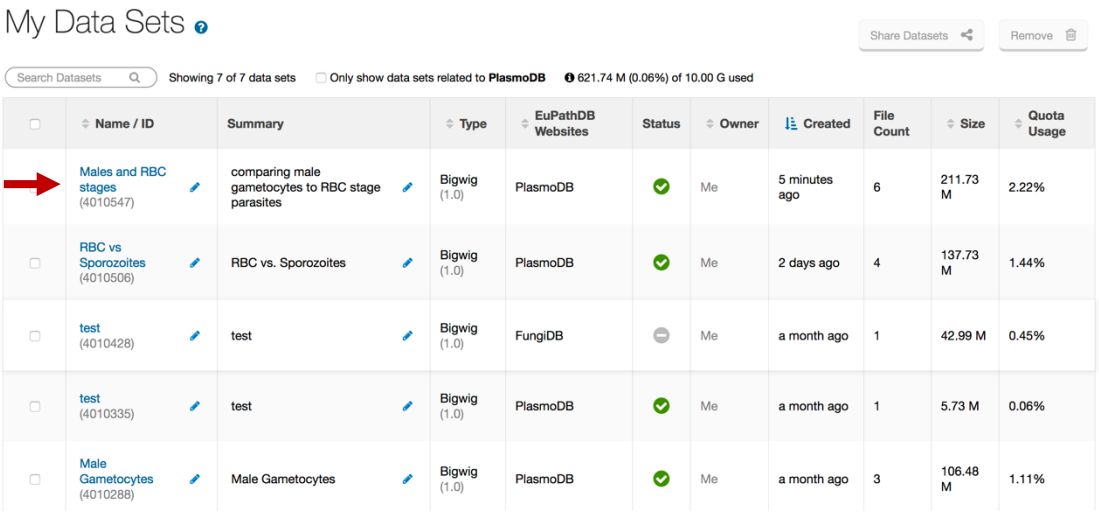
A. Click on EuPathDB Export Tools, then click on BigWig Files to EuPathDB (left Tools panel). The export tool will appear in the central portion.
B. Give your dataset a name.
C. Select "Dataset Collections" (icon looks like a folder). Then select all the BigWig collections that appear (Shift click).
D. Select the reference genome for your experiment.
E. Provide a short summary and dataset description – these could be the same for the purpose of this exercise.
F. Click on the Execute button. This will initiate a new step in your history which will indicate the transfer progress.



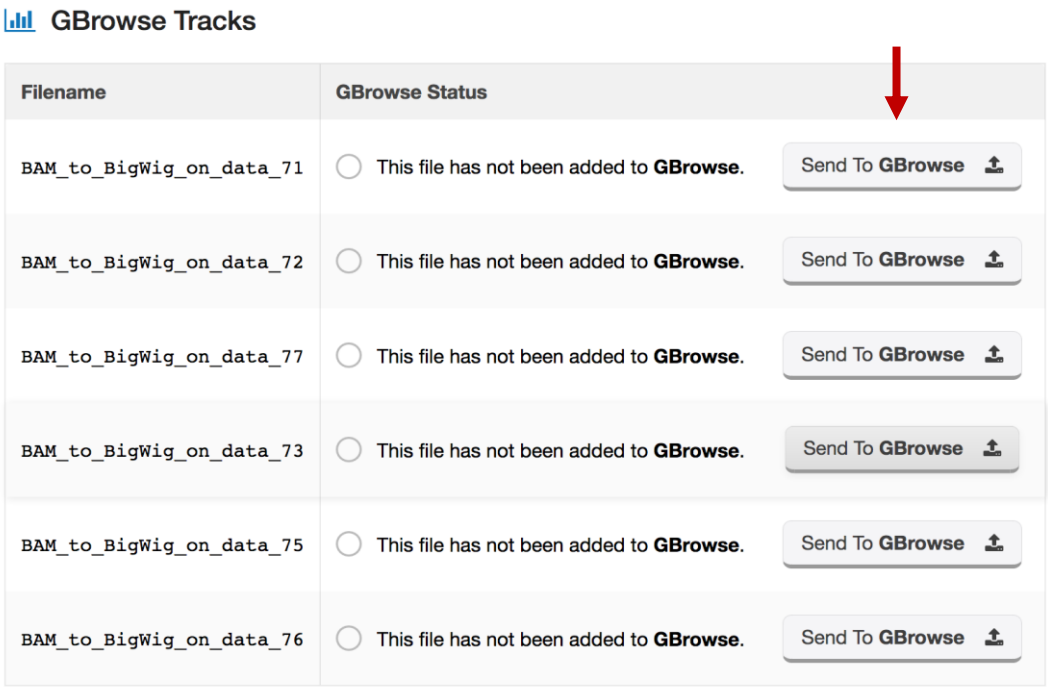
G. One the export is completed go to the database of the reference genome used. In this case PlasmoDB. Make sure you are logged in then click on the "My Datasets" tab in the grey menu bar.

H. You should see your dataset in the list. If this is the first dataset you transfer to EuPathDB then you will only see one. If it is not then the most recently transferred dataset will be at the top.

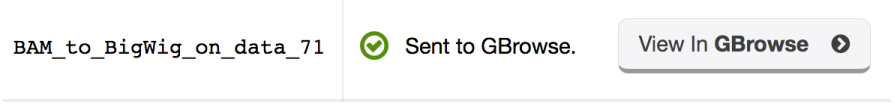
I. Click on the name of the dataset to view and interact with the dataset details.



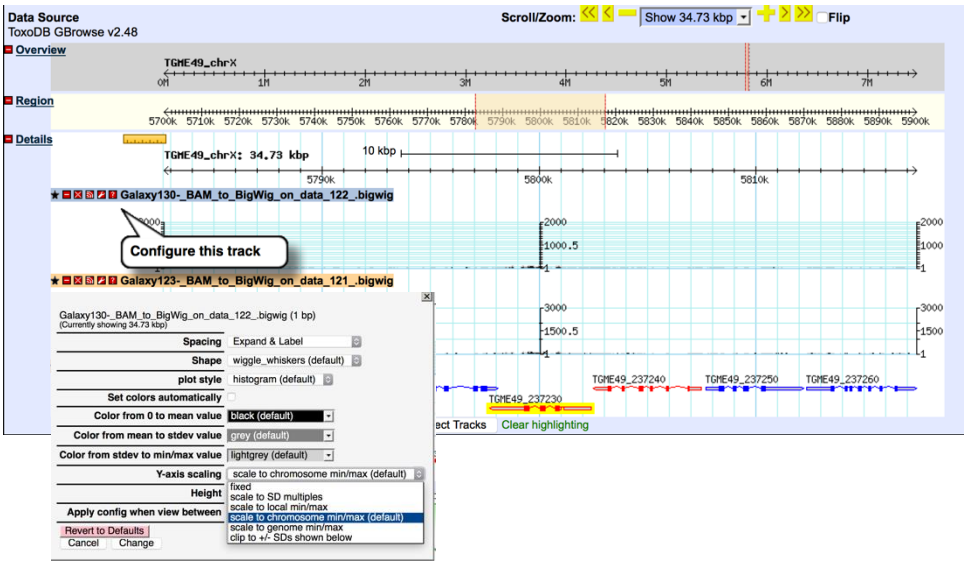
J. Scroll down to the GBrowse tracks section and click on the "Send to GBrowse" buttons for each of the files in the list.



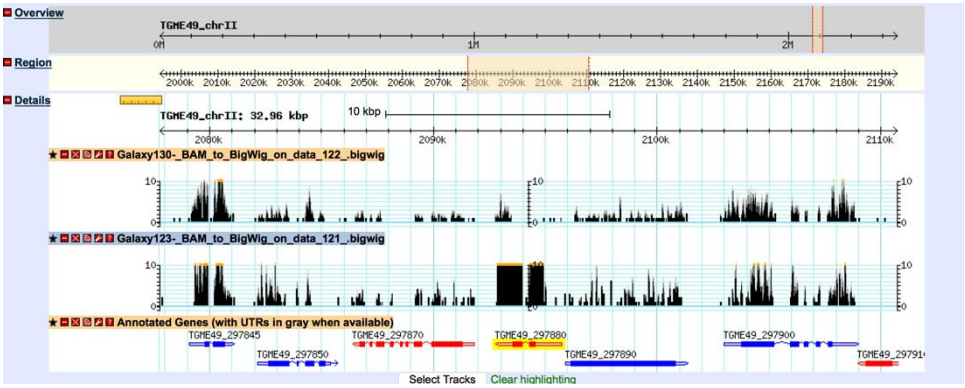
K. The send to GBrowse button will change to "View in GBrowse". Click on these buttons to view data. This may take a while so you can move on to step 3 called "Sharing histories with others" after clicking the buttons.



L. Adjust the Y axis by clicking on the “configure this track icon”. Adjust to a fixed Y axis and set the maximum to a value that makes sense for the results you are looking at.

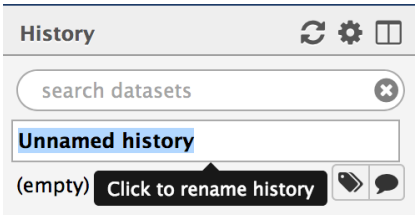


M. Explore your results in Gbrowse – zoom in or zoom out. Find any regions of interest. For example, one of your samples came from a knock out strain, go to that gene and see if you can find the difference.

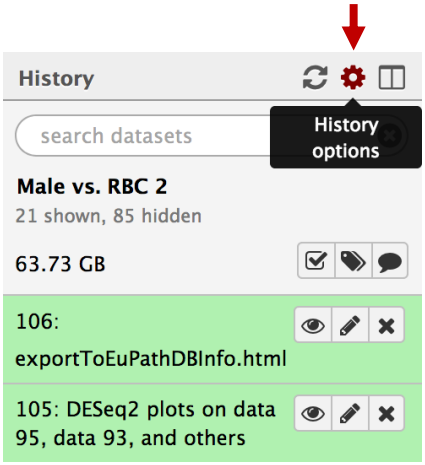


Step 3: Sharing histories with others:

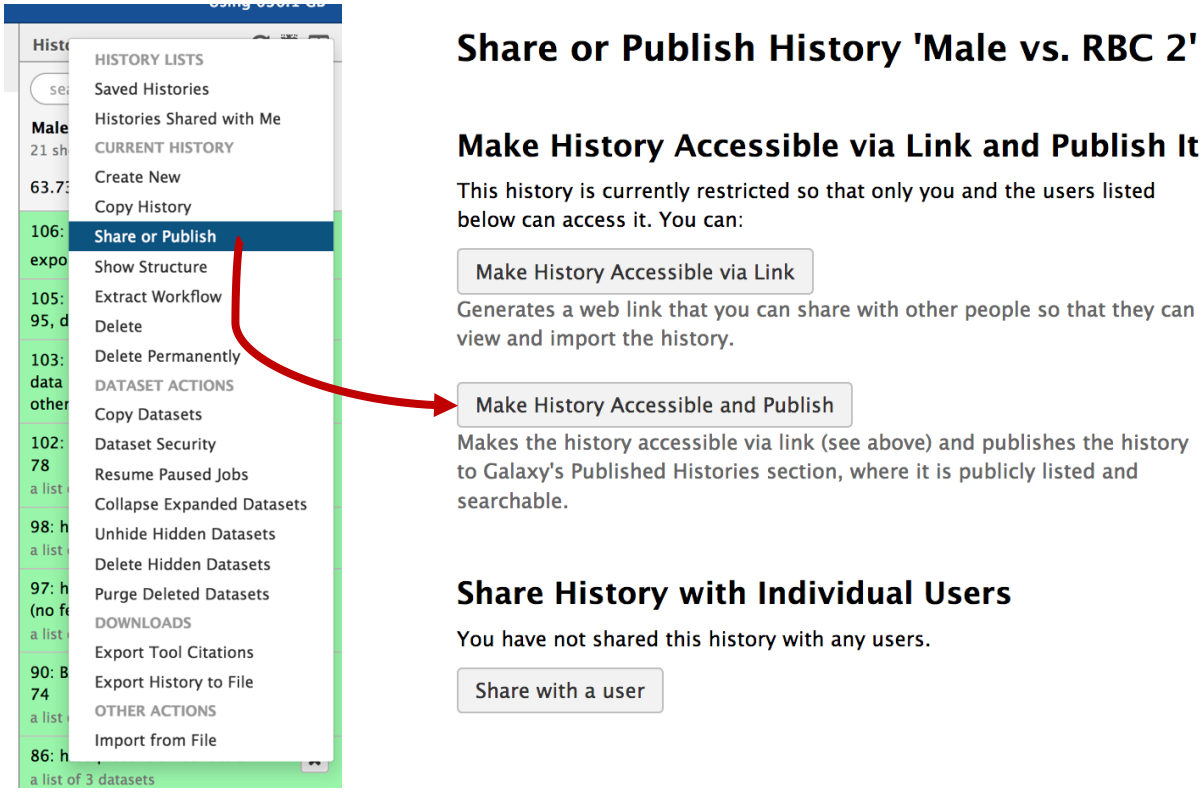
a. Make sure your history has a useful name – you can change the name by clicking on "unnamed history"



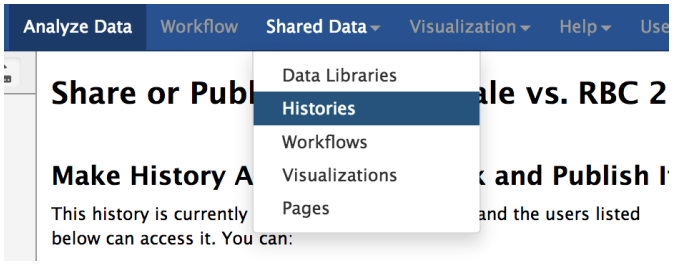
b. Click on the history options menu icon



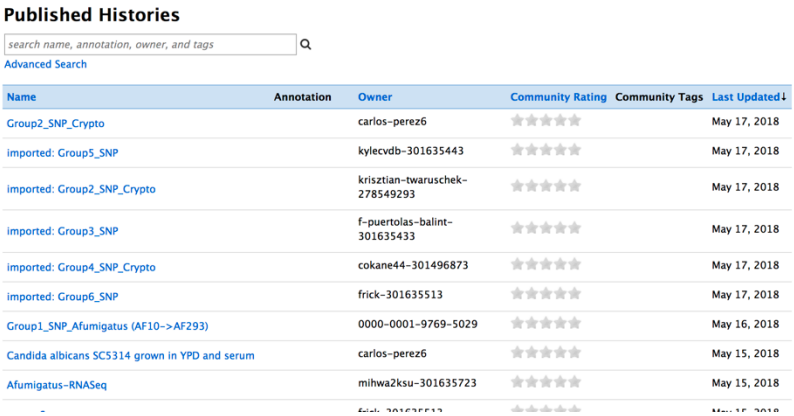
c. Select the "Share or Publish" option, the click on the "Make History Accessible and Publish" button in the center section.



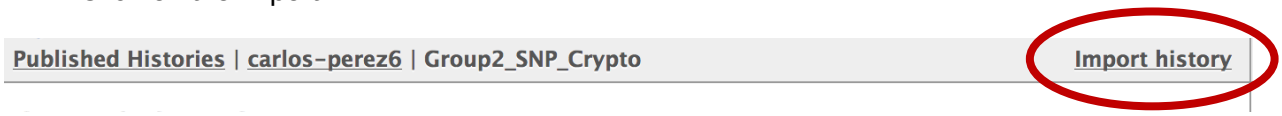
d. To import a shared history, go to the "histories" section (under the shared data menu item).



e. Find the history you would like to import and click on it.



f. Click on the import link.



Step 4: Explore the differential expression results:

DESeq2 is a package with essential estimates expression values and calculates differential expression. DESeq2 requires counts as input files. You can explore details of DESeq2 here: https://bioc.ism.ac.jp/packages/2.14/bioc/vignettes/DESeq2/inst/doc/beginner.pdf

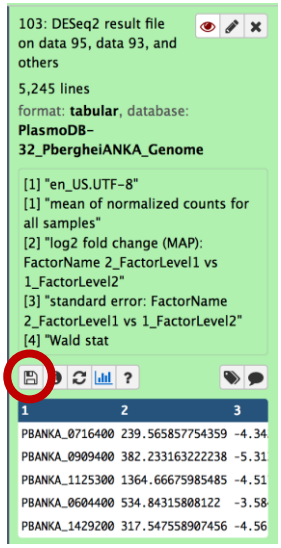
We will explore two output files:

A. DESeq2 Plots – you can view these directly in galaxy by clicking on the view icon. These plots give you an idea about the quality of the experiment. The link above includes a detailed description of the graphs.
B. DESeq2 results file – this is a table which contains the actual differential expression results. These can be viewed within galaxy but it will be more useful to download this table and open in Excel so you can sort results and big genes of interest.

The tabular file contains 7 columns:

COLUMN	DESCRIPTION
1	Gene Identifiers
2	mean normalized counts, averaged over all samples from both conditions
3	the logarithm (to basis 2) of the fold change (See the note in inputs section)
4	standard error estimate for the log2 fold change estimate
5	Wald statistic
6	p value for the statistical significance of this change
7	p value adjusted for multiple testing with the Benjamini-Hochberg procedure which controls false discovery rate (FDR)

C. To download the table, click on the step then click on the save icon.



*** **important: the file name ends with the extension .tabular – change this to .txt then open the file in Excel.**

D. Explore the results in Excel. For example, sort them based on the log2 fold change – column 3.

E. Pick a list of gene IDs from column 3 that are up-regulated with a good corrected P value (column 7) and load then into PlasmoDB using the Gene by ID search. You can then analyze these results by GO enrichment for example. Do the same for down-regulated genes.

F. Compare results from the other groups. Can you find genes are that are uniquely up or down regulated in the conditions tested?