

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

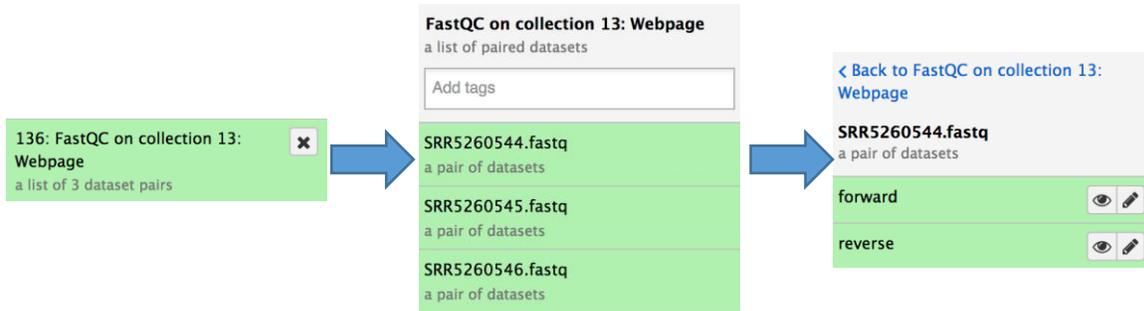
The goal of this exercise is to examine the results from the Galaxy RNAseq analysis workflow that ran overnight. If everything worked out you should see a list of completed workflow steps (Green). The workflow generates many output files, however not all of the output files are visible. You can explore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

The screenshot displays the EuPathDB Galaxy Site interface. The main content area shows a welcome message: "Welcome to the EuPathDB Galaxy Site" and "Many more output files are available to explore". Below this, it lists several types of data generated: "Differential expression data on the two collections", "Read counts per gene or exon (depending on chosen parameters)", and "Coverage data in BigWig format".

The right-hand panel, titled "History", shows a list of workflow steps. The top entry is "Male vs. RBC" with 21 shown, 98 deleted, and 144 hidden files. Below it are several DESeq2 plots and result files, and BAM to BigWig conversions. Red arrows point from the text in the main content area to the corresponding workflow steps in the history panel. A red circle highlights the "144 hidden" text in the "Male vs. RBC" entry.

Workflow Step ID	Description	Visibility
Male vs. RBC	21 shown, 98 deleted, 144 hidden	Hidden
203	DESeq2 plots on data 190, data 188, and others	Visible
202	Independent filtering result file on data 190, data 188, and others	Visible
201	DESeq2 result file on data 190, data 188, and others	Visible
197	BAM to BigWig on collection 173	Hidden
193	htseq-count on collection 173	Hidden
192	htseq-count on collection 173 (no feature)	Hidden
185	BAM to BigWig on collection 169	Hidden
181	htseq-count on collection 169	Hidden
180	htseq-count on collection 169 (no feature)	Hidden
173	HISAT2 on collection 150	Hidden

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.

The screenshot shows the 'globus Genomics' interface. The central panel is titled 'Bigwig Files to EuPathDB Export one or more bigwig files to EuPathDB where they can be viewed as tracks in the Genome Browser. (Galaxy Version 1.0.0)'. It contains several sections: 'My Data Set name:' with a text input field containing 'Males and RBC stages'; 'Bigwig files:' with a list of files including '102: BAM to BigWig on collection 78' and '90: BAM to BigWig on collection 74'; 'Reference genome:' with a dropdown menu set to 'PlasmoDB-32_PbergheiANKA_Genome'; 'My Data Set summary:' with a text area containing 'comparing male gametocytes to RBC stage parasites'; and 'My Data Set description:' with a text area containing 'comparing male gametocytes to RBC stage parasites'. An 'Execute' button is at the bottom. A tooltip for the 'Execute' button is visible, stating 'What it does (check this Tutorial!) This tool exports bigwig files in a new My Data Set to contain them. Typically the reference genome you specify determines which you choose P.falciparum 3D7 the export will be sent to PlasmoDB exportToEuPathDBInfo.html'. The right sidebar shows a 'History' panel with a list of datasets, including 'Male vs. RBC 2', '74: HISAT2 on collection 27', '70: FastQC on collection 14: RawData', '69: FastQC on collection 14: Webpage', '56: Trimmomatic on collection 14: unpaired', '55: Trimmomatic on collection 14: paired', '42: FastQC on collection 7: RawData', '41: FastQC on collection 7: Webpage', '28: Trimmomatic on collection 7: unpaired', '27: Trimmomatic on collection 7: paired', and '14: Male Gametocytes'. The top navigation bar includes 'Home', 'New Search', 'My Strategies', 'My Basket (1)', 'My Data Sets', 'Tools', 'Data Summary', 'Downloads', 'Community', 'Analyze My Experiment', and 'My Favorites'.

G. Once 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.

The screenshot shows the top navigation bar of the globus Genomics interface. The 'My Data Sets' tab is highlighted with a red box and a red '1/1' badge, indicating one dataset is present. Other tabs include 'Home', 'New Search', 'My Strategies', 'My Basket (1)', 'Tools', 'Data Summary', 'Downloads', 'Community', 'Analyze My Experiment', and 'My Favorites'.

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.

My Data Sets [?](#)

Share Datasets Remove

Search Datasets Showing 7 of 7 data sets Only show data sets related to **PlasmoDB** 621.74 M (0.06%) of 10.00 G used

<input type="checkbox"/>	Name / ID	Summary	Type	EuPathDB Websites	Status	Owner	Created	File Count	Size	Quota Usage
	Males and RBC stages (4010547)	comparing male gametocytes to RBC stage parasites	Bigwig (1.0)	PlasmoDB		Me	5 minutes ago	6	211.73 M	2.22%
<input type="checkbox"/>	RBC vs Sporozoites (4010506)	RBC vs. Sporozoites	Bigwig (1.0)	PlasmoDB		Me	2 days ago	4	137.73 M	1.44%
<input type="checkbox"/>	test (4010428)	test	Bigwig (1.0)	FungiDB		Me	a month ago	1	42.99 M	0.45%
<input type="checkbox"/>	test (4010335)	test	Bigwig (1.0)	PlasmoDB		Me	a month ago	1	5.73 M	0.06%
<input type="checkbox"/>	Male Gametocytes (4010288)	Male Gametocytes	Bigwig (1.0)	PlasmoDB		Me	a month ago	3	106.48 M	1.11%

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

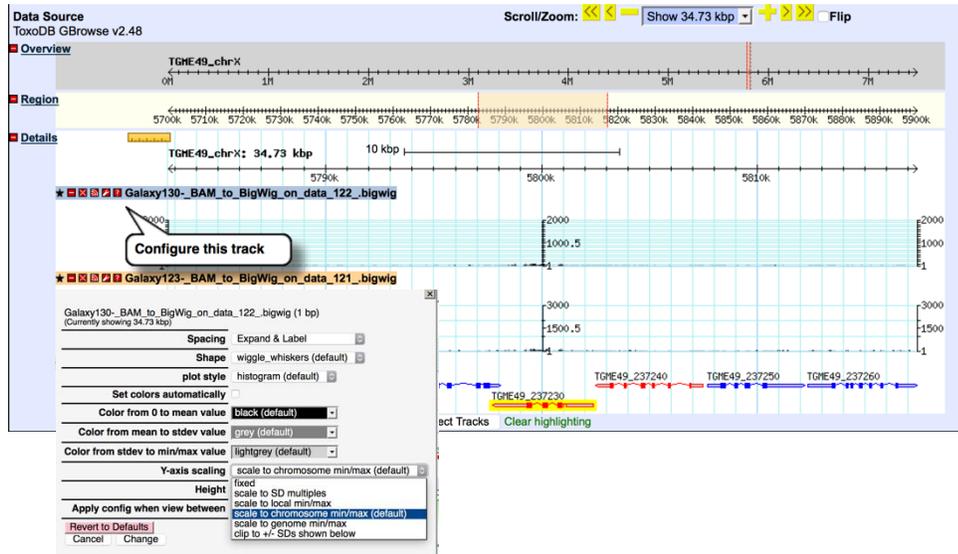
GBrowse Tracks

Filename	GBrowse Status	
BAM_to_BigWig_on_data_71	<input type="radio"/> This file has not been added to GBrowse .	
BAM_to_BigWig_on_data_72	<input type="radio"/> This file has not been added to GBrowse .	
BAM_to_BigWig_on_data_77	<input type="radio"/> This file has not been added to GBrowse .	
BAM_to_BigWig_on_data_73	<input type="radio"/> This file has not been added to GBrowse .	
BAM_to_BigWig_on_data_75	<input type="radio"/> This file has not been added to GBrowse .	
BAM_to_BigWig_on_data_76	<input type="radio"/> This file has not been added to GBrowse .	

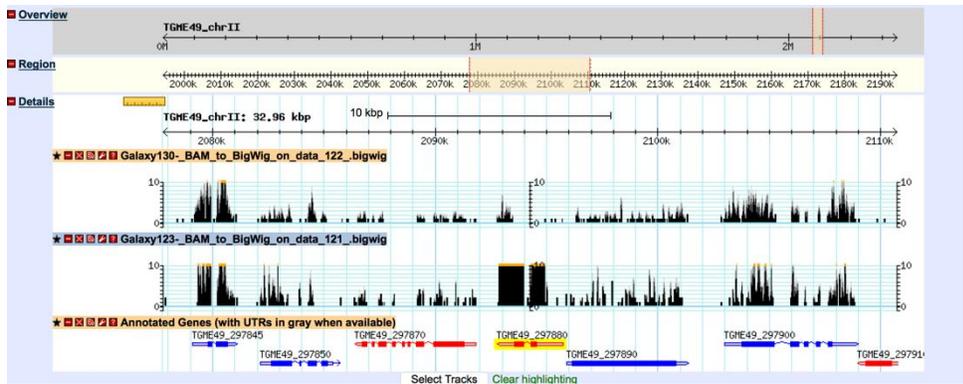
- K. The send to GBrowse button will change to “View in GBrowse”. Click on these buttons to view data. This may take a while so you can move on to step 3 called “Sharing histories with others” after clicking the buttons.

BAM_to_BigWig_on_data_71	Sent to GBrowse.	
--------------------------	------------------	--

- 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.

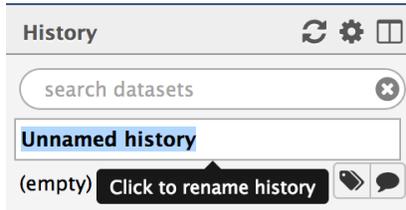


- M. Explore your results in Gbrowse – zoom in or zoom out. Find any regions of interest. For example, one of your samples came from a knock out strain, go to that gene and see if you can find the difference.

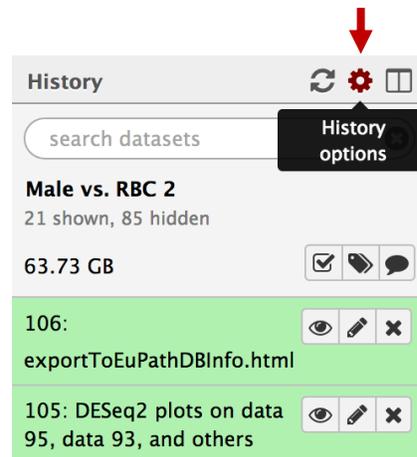


Step 3: Sharing histories with others:

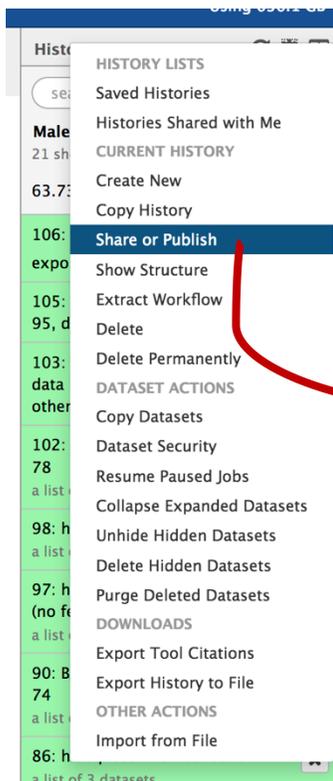
- a. Make sure your history has a useful name – you can change the name by clicking on “unnamed history”



- b. Click on the history options menu icon



- c. Select the “Share or Publish” option, then click on the “Make History Accessible and Publish” button in the center section.



Share or Publish History 'Male vs. RBC 2'

Make History Accessible via Link and Publish It

This history is currently restricted so that only you and the users listed below can access it. You can:

[Make History Accessible via Link](#)

Generates a web link that you can share with other people so that they can view and import the history.

[Make History Accessible and Publish](#)

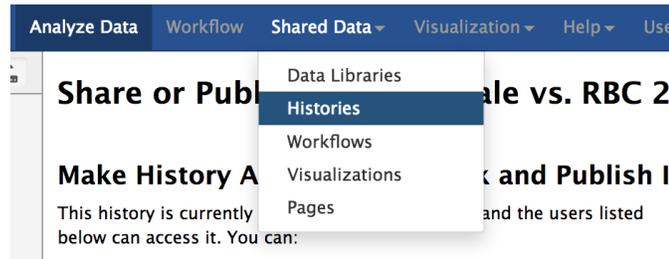
Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, where it is publicly listed and searchable.

Share History with Individual Users

You have not shared this history with any users.

[Share with a user](#)

- 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.

Published Histories

search name, annotation, owner, and tags Q

Advanced Search

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
Group2_SNP_Crypto		carlos-perez6	★★★★★		May 17, 2018
imported: Group5_SNP		kylecvdb-301635443	★★★★★		May 17, 2018
imported: Group2_SNP_Crypto		krisztian-twaruscek-278549293	★★★★★		May 17, 2018
imported: Group3_SNP		f-puertolas-ballint-301635433	★★★★★		May 17, 2018
imported: Group4_SNP_Crypto		cokane44-301496873	★★★★★		May 17, 2018
imported: Group6_SNP		frick-301635513	★★★★★		May 17, 2018
Group1_SNP_Afumigatus (AF10->AF293)		0000-0001-9769-5029	★★★★★		May 16, 2018
Candida albicans SCS314 grown in YPD and serum		carlos-perez6	★★★★★		May 15, 2018
Afumigatus-RNASeq		mihwa2ksu-301635723	★★★★★		May 15, 2018

- 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>

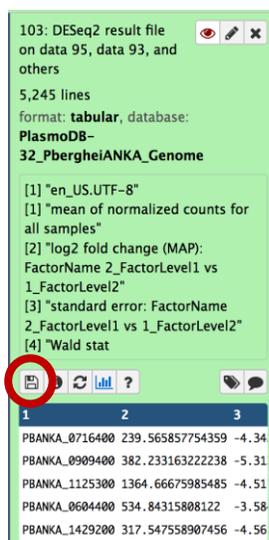
We will explore two output files:

- 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.
- 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 that are uniquely up or down regulated in the conditions tested?