

RNA sequence data analysis via Galaxy, Part II Analyzing your results (Group Exercise)

Learning objectives:

- examine the results from the Galaxy RNA-Seq analysis workflow
- Import data from Galaxy to the VectorBase “My Workspace”
- Analyze the results using the VectorBase interface and tools
- Analyzing DEseq2 results

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.

Resources:

[FastQC Result Interpretation](https://workshop.eupathdb.org/athens/2019/exercises/fastqc_results-2.pdf) (https://workshop.eupathdb.org/athens/2019/exercises/fastqc_results-2.pdf)

[Beginner DESeq2 guide](https://workshop.eupathdb.org/athens/2019/exercises/beginner_DeSeq2.pdf) (https://workshop.eupathdb.org/athens/2019/exercises/beginner_DeSeq2.pdf)

[FastQC output](https://workshop.eupathdb.org/athens/2019/exercises/fastqc_output.pdf) (https://workshop.eupathdb.org/athens/2019/exercises/fastqc_output.pdf)

[SNP Eff manual](http://snpeff.sourceforge.net/SnpEff_manual.html) (http://snpeff.sourceforge.net/SnpEff_manual.html)

[Trimmomatic Manual](http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf)

(http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf)

Step 1: Explore the FastQC results. To do this find the step called “FastQC on collection ##:

The screenshot shows the Globus Genomics interface with a list of workflow steps on the right. Red arrows point from text annotations to specific items in the list:

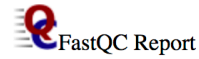
- Annotation: "Many more output files are available to explore" points to the "Uninfected hoxes dsGFP vs dsGFP" step.
- Annotation: "Differential expression data on the two collections" points to the "222: DESeq2 plots on data a 215, data 213, and others" step.
- Annotation: "Coverage data in BigWig format" points to the "197: BAM to BigWig on collection n 189" step.
- Annotation: "FastQC results – there will be one for each FastQ file" points to the "14: dsISARL1_uninfected" step.

The list of steps includes:

- Uninfected hoxes dsGFP vs dsGFP (8 shown, 104 deleted, 198 hidden, 483.43 GB)
- 222: DESeq2 plots on data a 215, data 213, and others
- 221: DESeq2 result file on data 215, data 213, and others
- 217: BAM to BigWig on collection n 193 (a list with 3 items)
- 197: BAM to BigWig on collection n 189 (a list with 3 items)
- 147: FastQC on collection 14: W ebpage (a list of pairs with 3 items)
- 126: FastQC on collection 13: W ebpage (a list of pairs with 3 items)
- 14: dsISARL1_uninfected (a list of pairs with 3 items)
- 13: dsGFP_uninfected

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

136: FastQC on collection 13:
Webpage
a list of 3 dataset pairs

ection 13:



Summary

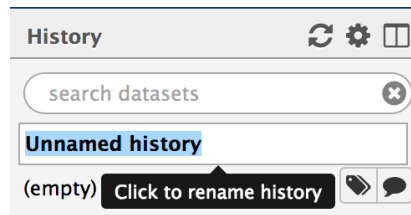
- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented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

Basic Statistics

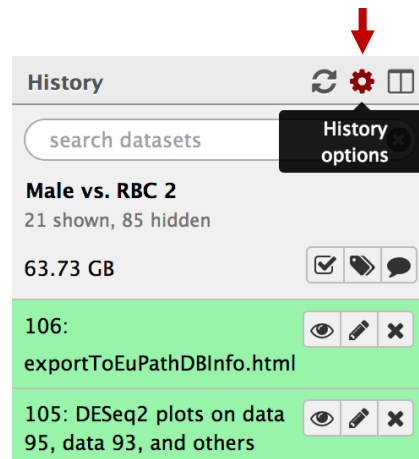
	Measure	Value
Filename		SRR5260544_1.fastq.gz
File type		Conventional base calls

Step 2: Sharing histories with others:

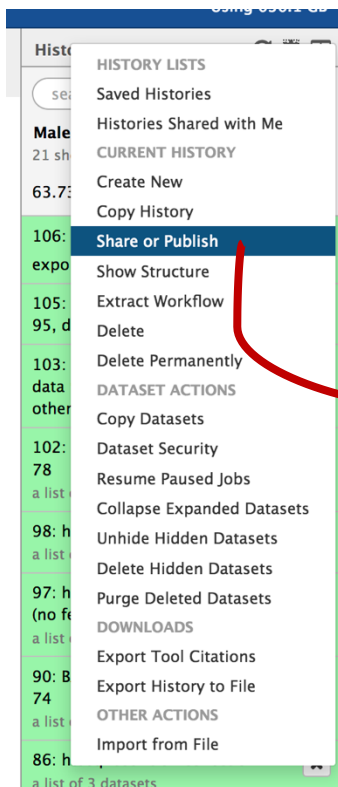
- Make sure your history has a useful name – you can change the name by clicking on “unnamed history”



- Click on the history options menu icon



- 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.

The screenshot shows the Galaxy web interface. At the top, the 'Shared Data' menu is open, and 'Histories' is selected. Below this, a table titled 'Published Histories' is displayed. The table has columns for Name, Annotation, Owner, Community Rating, Community Tags, and Last Updated. The 'Import history' button is circled in red.

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-twarushek-278549293	★★★★★		May 17, 2018
imported: Group6_SNP		ITRICK-301635513	★★★★★		May 17, 2018
Group1_SNP_Afumigatus (AF10->AF293)		0000-0001-9769-5029	★★★★★		May 16, 2018
Candida albicans SC5314 grown in YPD and serum		carlos-perez6	★★★★★		May 15, 2018
Afumigatus-RNASeq		mihwa2ksu-301635723	★★★★★		May 15, 2018

- f. Click on the import link.

Step 3: 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>

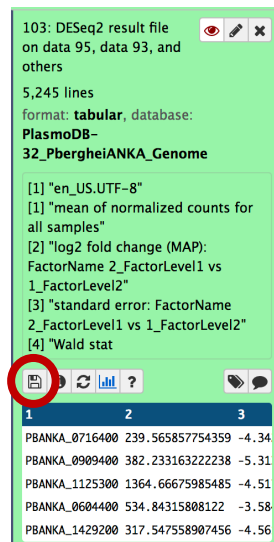
We will explore two output files:

- 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?

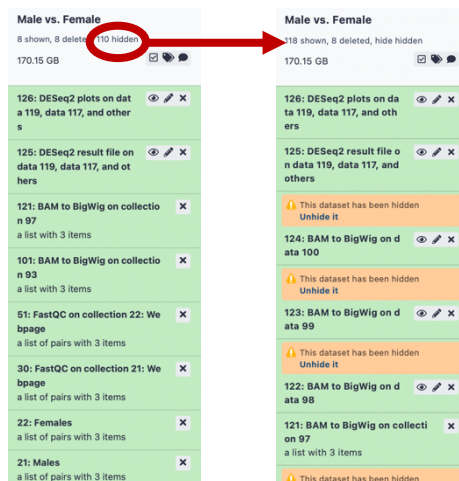
Exporting data to VEuPathDB

The VEuPathDB RNAseq export tool provides a mechanism to export your RNAseq results (TPM values) and BigWig RNAseq coverage files. The advantage of doing this is that it allows you to search the TPM data using the RNAseq search in VEuPathDB and view the BigWig files in the genome browser.

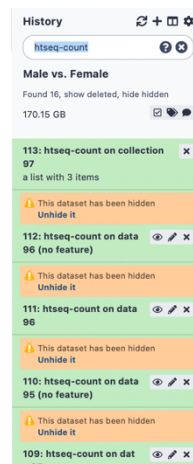
However, to use this feature you need to generate TPM values for genes in your datasets and organize your results into two collections, one for the TPMs and one for the BigWigs.

First let's organize the files (see matching screen shots below):

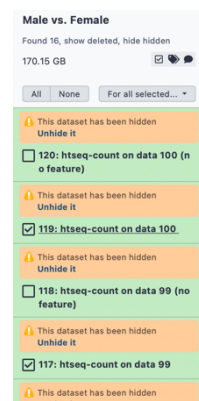
1. Click on the link at the top of your history that says “## hidden”. This will show all hidden files.



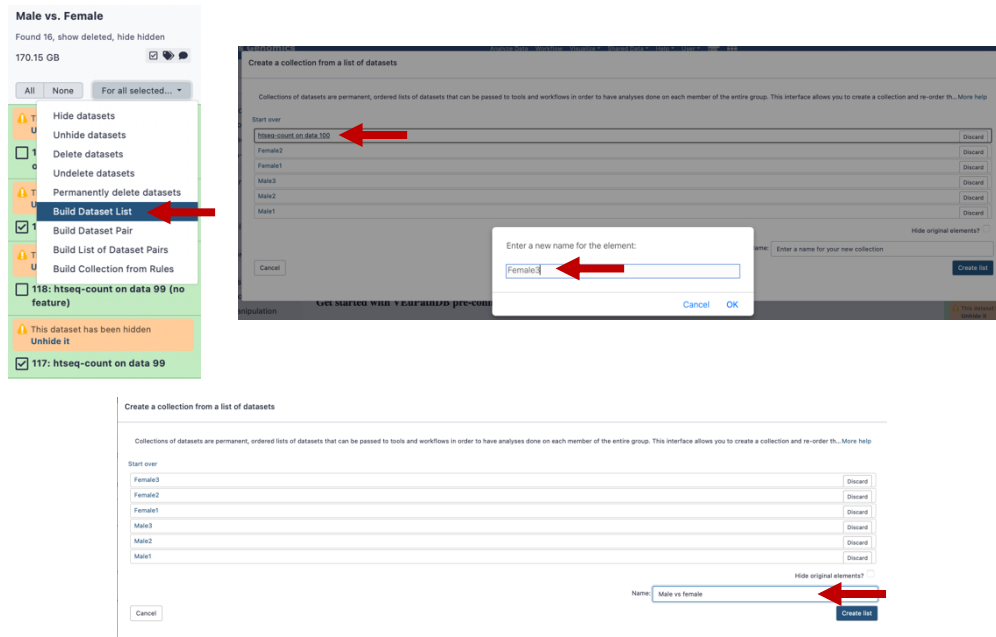
2. Use the search datasets box at the top of your history to find any file in your history with the work “htseq-count”. Ignore the ones that include (no feature) in their names or that are a collection.



3. Click on the “operation on multiple datasets” tool and select the individual htseq-count files. These should look something like this: htseq-count on data 65. *Note if you are comparing two conditions each done in triplicate then you should have selected 6 files.*
4. Click on the “for selected button” and choose the “Build dataset list” option.



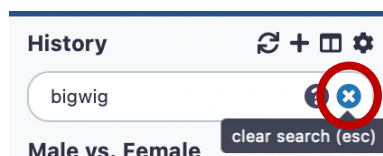
5. In the popup, rename each of the samples and give the collection a name, then click on the Create List button.



6. Repeat the same steps to create the list of BigWig files.

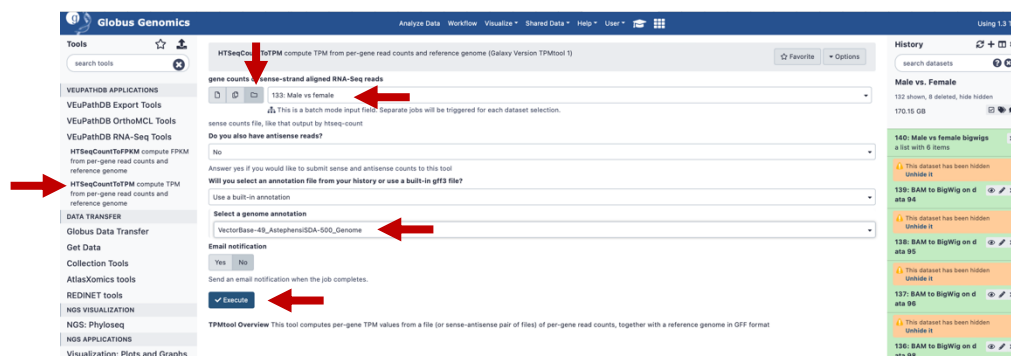


7. Click on clear search to see all results in your history.



Now that your count and bigwig files are nice and organized, the next step is to convert the counts into TPMs. To do this follow these steps:

1. Select the HTSeqCountToTPM tool (under the VEuPathDB RNAseq tools in the left menu).
2. Make sure the list of count files is selected.
3. Select the reference organism.



4. Click on Execute.

Optional: Click on “hide hidden” to clean up your history a bit.

Export data to VEuPathDB. To export the TPM and BigWig files follow these steps:

1. Click on “VEuPathDB Export Tools” in the left-hand panel.
2. Click on the tool called “RNA-Seq to VEuPathDB”
3. Fill up the export tool and select the correct files to export (see screen shot).

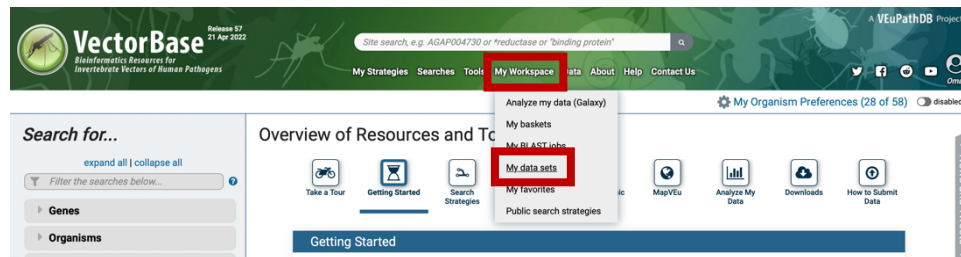
The screenshot shows the 'RNA-Seq to VEuPathDB' tool interface. On the left, a sidebar lists various tools under 'VEUPATHDB APPLICATIONS'. The main panel contains several input fields and dropdown menus. Red arrows point to the following elements:

- VEuPathDB Export Tools** in the left sidebar.
- RNA-Seq to VEuPathDB** in the left sidebar.
- My Data Set name:** A text input field containing 'uninfected WT vs Silenced'.
- BigWig collection:** A dropdown menu showing '252: bigwig_collection all'.
- TPM or FPKM collection:** A dropdown menu showing '233: HTSeqCountToTPM on collection 231: gene expression'.
- My Data Set summary:** A text input field containing 'uninfected WT vs. Silenced'.
- My Data Set description:** A text input field containing 'uninfected WT vs. Silenced'.
- Execute** button at the bottom.

The tool title is 'RNA-Seq to VEuPathDB Export an RNA-Seq result to VEuPathDB (Galaxy Version 1.0.0)'. It includes a 'Favorite' button and an 'Options' dropdown. The 'Are you exporting sense and antisense TPM/FPKM datasets?' dropdown is set to 'No'. The 'Email notification' section has 'Yes' and 'No' buttons, with a note 'Send an email notification when the job completes.'

Explore your data in VEuPathDB: Go to the VEuPathDB database that your data belongs to (e.g. FungiDB).

1. Click on the “My Workspace” link in the grey menu bar. Then select “My datasets” from the list.



2. You should see the dataset you exported from galaxy in this list. Click on it and explore the dataset page.

[All My Data Sets](#)

My Dataset: uninfected WT vs Silenced

Status: ✔ This data set is installed and ready for use in VectorBase.

Owner: Me

Description: uninfected WT vs. Silenced

ID: 4057319

Data Type: RNA-Seq (RnaSeq 1.0)

Summary: uninfected WT vs. Silenced

Created: an hour ago

Dataset Size: 475.19 M

Quota Usage: 4.98% of 10.00 G

Available Searches:

- RNA-Seq user dataset (fold change)

Use This Dataset in VectorBase

Compatibility Information

VEuPathDB Website

VectorBase

Display a menu

Identify Genes based on RNA-Seq user dataset (fold change)

Male vs Females

For the Experiment: unstranded

return: protein coding Genes

that are: up-regulated

with a Fold change >= 4

between each gene's average expression value in the following Reference Samples

Female3
Female2
Female1
Male3
Male2
Male1

select all | clear all

and its average expression value in the following Comparison Samples

Female3
Female2
Female1
Male3
Male2
Male1

select all | clear all

Example showing one gene that would meet search criteria
(Dots represent this gene's expression values for selected samples)

Up-regulated

For each gene, the search calculates:

$$\text{fold change} = \frac{\text{average expression value in comparison}}{\text{average expression value in reference}}$$

and returns genes when fold change >= 4.

You are searching for genes that are up-regulated between at least two reference samples and at least two comparison samples.

To narrow the window, use the maximum reference value, or minimum comparison value. To broaden the window,

3. Explore the available search to identify genes with expression differences. Note that a custom graph is generated for your data in the results and on gene pages!
4. Explore the coverage plots in the genome browser.



5.

Select Tracks

My Tracks

Currently Active

Recently Used

Category

- 1 Comparative Genomics
- 3 Gene Models
- 5 Genetic Variation
- 6 My Data from Galaxy
- 8 Sequence Analysis
- 135 Transcriptomics

Subcategory

- 6 RNASeq

Dataset

- 6 Male vs Females

Track Type

- 6 Coverage

RNA-Seq Alignment

- 6 (no data)

RNA-Seq Strand

- 6 (no data)

Back to browser

Clear All Filters

Name	Category
<input type="checkbox"/> Male vs Females female1.bw	My Data from
<input checked="" type="checkbox"/> Male vs Females female2.bw	My Data from
<input type="checkbox"/> Male vs Females female3.bw	My Data from
<input type="checkbox"/> Male vs Females male1.bw	My Data from
<input type="checkbox"/> Male vs Females male2.bw	My Data from
<input type="checkbox"/> Male vs Females male3.bw	My Data from

