

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

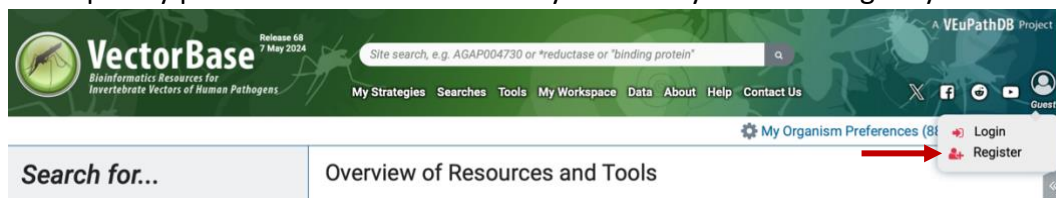
Learning objectives:

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

Part 1

Use this site: <https://veupathdb1.globusgenomics.org>

You will need a VEuPathDB account to access this galaxy instance. To create an account, go to <https://vectorbase.org> then select the register link from the upper right hand side login icon. A temporary password will be emailed to you which you can change if you would like.



Once you have an account you will be able to access the VEuPathDB galaxy instance.

1 Log in to use veupathdb1

Use your existing organizational login

e.g., university, national lab, facility, project

VEuPathDB

By selecting Continue, you agree to Globus [terms of service](#) and [privacy policy](#).

Continue

2



Please log in

Username or Email:

Password:

Login

Cancel

[Forgot Password?](#)

[Register/Subscribe](#)

Visit our partner Bioinformatics Resource Center, [BV-BRC](#)



Follow the instructions on the next pages to access the VEuPathDB galaxy instance.

This exercise is designed to start with a completed galaxy RNAseq analysis history. You will be divided into groups and each person may import their group's history:

Group 1 will compare *Anopheles stephensi* mosquitos fed with blood infected with Wild type or CSP mutant strains of *Plasmodium*.

<https://veupathdb1.globusgenomics.org/u/oharb.391/h/csp-wt-vs-csp-mutant-results>

The raw data is available in the sequence repositories:

<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA734759>

Groups 2 & 3 will compare samples from ticks infected with *B. burgdorferi* during adiponectin receptor silencing.

Group 2 (Uninfected Wt vs Silenced):

<https://veupathdb1.globusgenomics.org/u/oharb.391/h/b-burgdorferi-uninfected-wt-vs-silenced-results>

Group 3 (Infected Wt vs Silenced):

<https://veupathdb1.globusgenomics.org/u/oharb.391/h/infected-wt-vs-infected-silenced-results>

The raw data is available in the sequence repositories:

<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA716187>

Groups 4 & 5 will examine insecticide resistant and susceptible strains of *Anopheles coluzzii*. The resistant strains are either original field isolates or reselected resistance after loss of resistance (susceptible).

Group 4 (Susceptible vs Original):

<https://veupathdb1.globusgenomics.org/u/oharb.391/h/susceptible-vs-original-results>

Group 5 (Susceptible vs Reselected):

<https://veupathdb1.globusgenomics.org/u/oharb.391/h/reselected-vs-original-results>

The raw data is available in the sequence repositories:

<https://www.ncbi.nlm.nih.gov/bioproject/750256>

Group 6 will be examining data from a study called "RNAseq from adult male and female *Anopheles stephensi*" We will compare the male to the female samples.

<https://veupathdb1.globusgenomics.org/u/oharb.391/h/female-vs-male-results>

The raw data is available in the sequence repositories:

<https://www.ncbi.nlm.nih.gov/bioproject/277477>

To import a history, click on the link for your group then click on the import ‘+’ icon in the upper right-hand side of the window.

If everything worked out, you should see a list of completed workflow steps (Green). The workflow generates many output files, however not all 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)

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


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

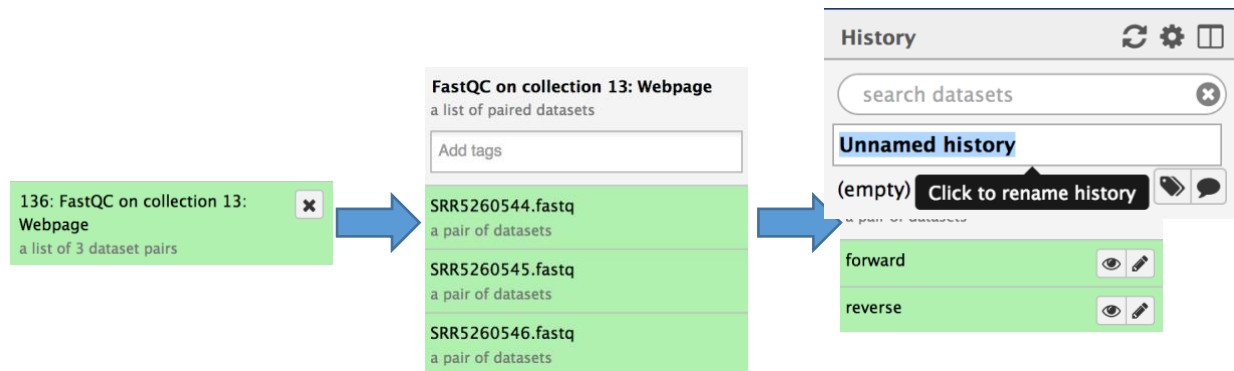
[SNP Eff manual](#) (http://snpeff.sourceforge.net/SnpEff_manual.html)

[Trimmomatic Manual](#)


(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 ##: 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 3: Explore the differential expression results:

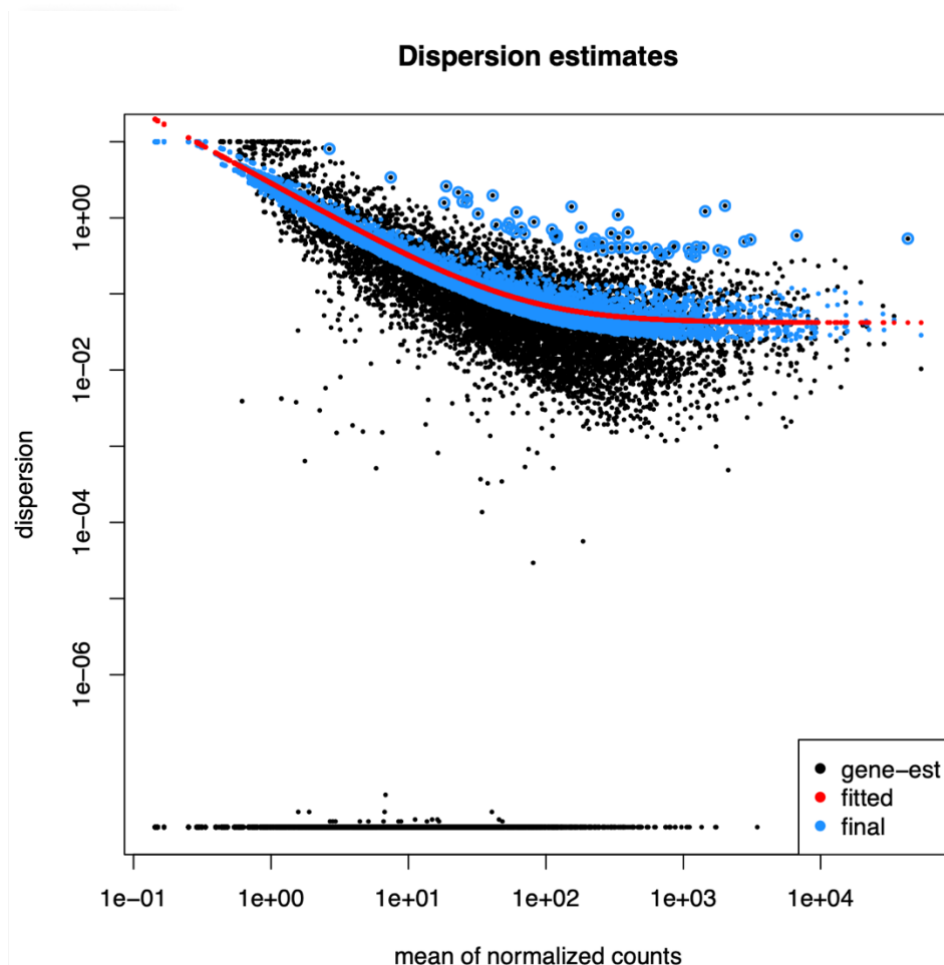
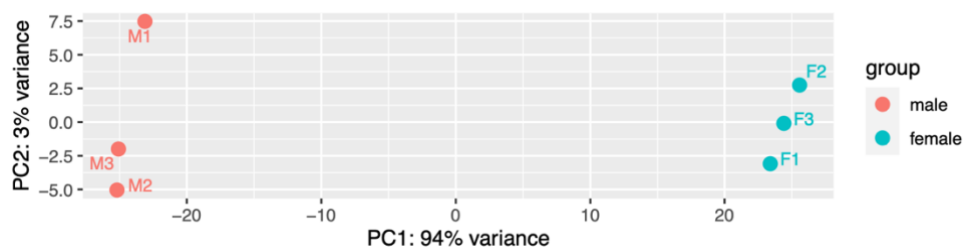
DESeq2 is a package with essential estimates expression values and calculates differential expression. DESeq2 requires counts as input files. You can explore details of DESeq2 here:

<https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>

https://youtu.be/0b24mpzM_5M?si=TwF7OBrbhe4tzlg2

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.



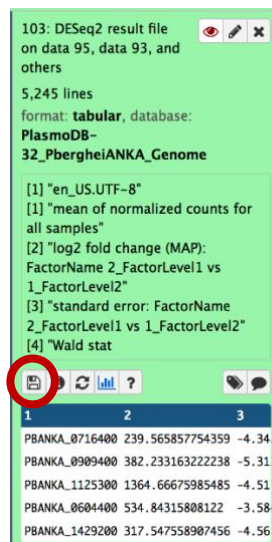
Part 2

- 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 PlasmaDB 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?

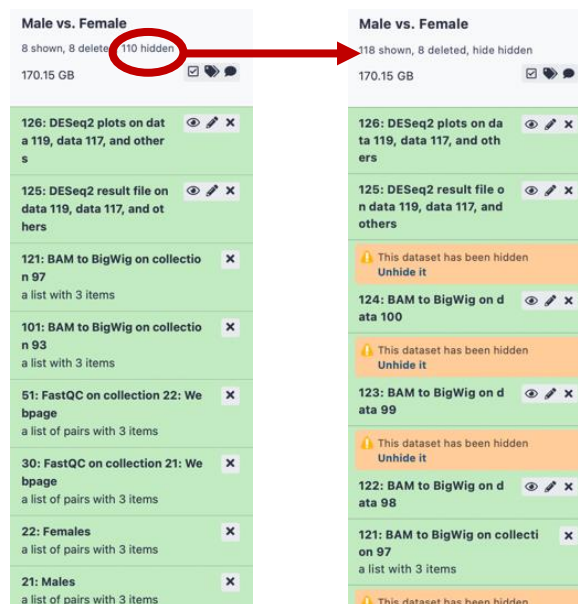
Exporting data to VEuPathDB

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.

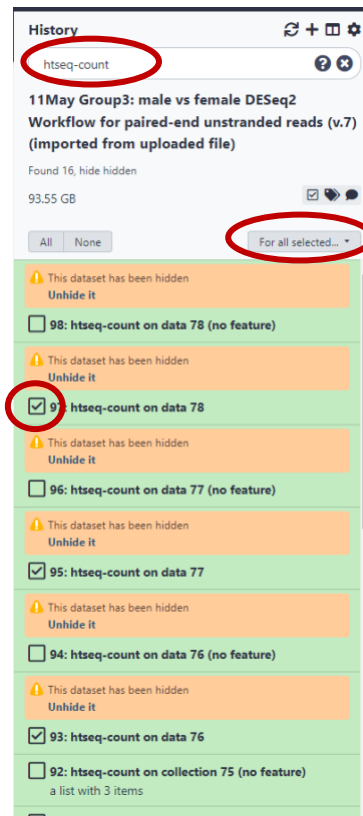
However, to use this feature you need to generate TPM values for genes in your datasets and organize your results into two collections, one for the TPMs and one for the BigWigs.

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

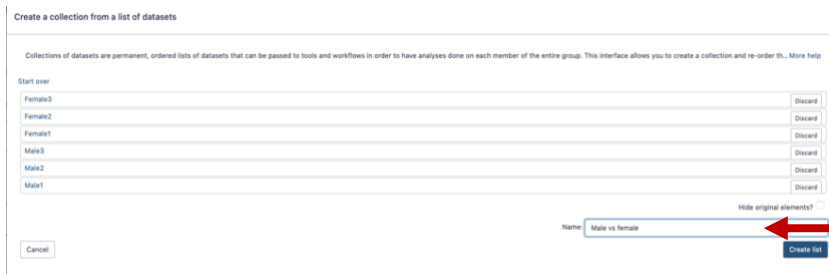
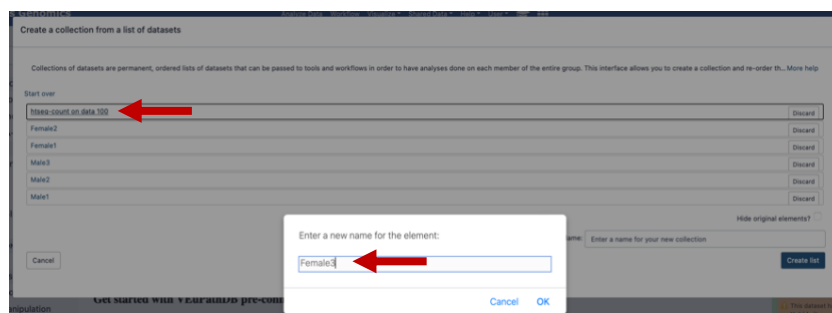
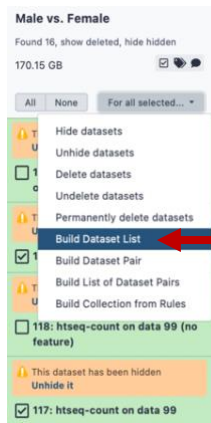
1. Click on the link at the top of your history that says “## hidden” to show hidden files.



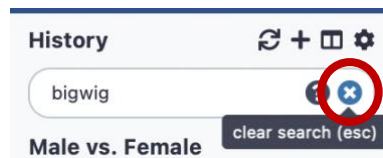
2. Use the search datasets box at the top of your history to find any file in your history with the work “htseq-count”. Ignore the ones that include (no feature) in their names or that are a collection.



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.



Create a collection from a list of datasets

Collections of datasets are permanent, ordered lists of datasets that can be passed to tools and workflows in order to have analyses done on each member of the entire group. This interface allows you to create a collection and re-order th...More help

Start over Clear selected

female3	<input type="button" value="Discard"/>
female2	<input type="button" value="Discard"/>
female1	<input type="button" value="Discard"/>
male3	<input type="button" value="Discard"/>
male2	<input type="button" value="Discard"/>
male1	<input type="button" value="Discard"/>

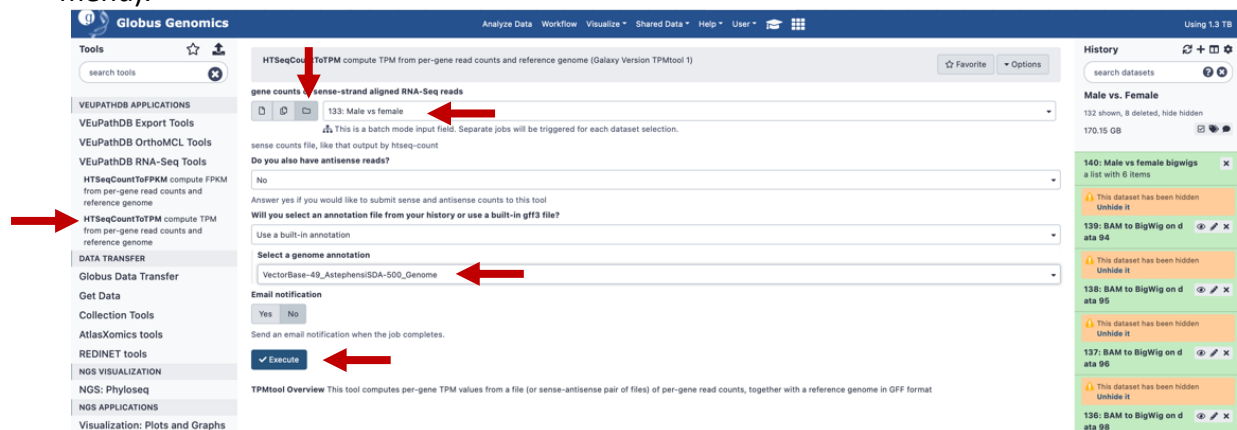
Hide original elements? ☐

Name:

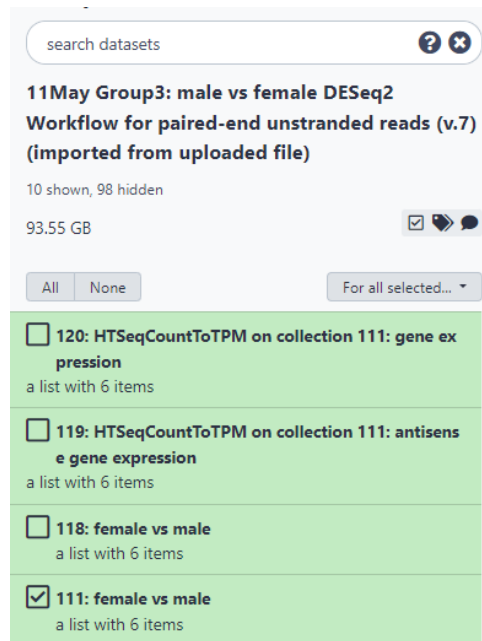
7. Click on clear search to see all results in your history.

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).

Tools

search tools

VEUPATHDB APPLICATIONS

VEuPathDB Export Tools

Gene List to VEuPathDB Export a gene list to VEuPathDB

Bigwig Files to VEuPathDB Export one or more bigwig files to VEuPathDB where they can be viewed as tracks in the Genome Browser.

RNA-Seq to VEuPathDB Export an RNA-Seq result to VEuPathDB

VEuPathDB OrthoMCL Tools

VEuPathDB RNA-Seq Tools

DATA TRANSFER

Globus Data Transfer

Get Data

Collection Tools

AtlasXomics tools

REDINET tools

NGS VISUALIZATION

NGS: Phyloseq

NGS APPLICATIONS

Visualization: Plots and Graphs

NGS: QC and manipulation

NGS: Assembly

Display a menu

RNA-Seq to VEuPathDB Export an RNA-Seq result to VEuPathDB (Galaxy Version 1.0.0)

Favorite Options

My Data Set name:

uninfected WT vs Silenced

specify a name for the new dataset

Are you exporting sense and antisense TPM/FPKM datasets?

No

Select yes if your experiment is strand-specific and you are including sense and antisense datasets in this export.

BigWig collection:

252: bigwig_collection all

Select the BigWig collection to include in the new VEuPathDB My Data Set. The BigWig collection you select here must be mapped to the reference genome that you select below.

TPM or FPKM collection:

233: HTSeqCountToTPM on collection 231: gene expression

Select the TPM or FPKM collection. For an unstranded dataset, its name should include the phrase 'gene expression'.

My Data Set summary:

uninfected WT vs. Silenced

My Data Set description:

uninfected WT vs. Silenced

Email notification

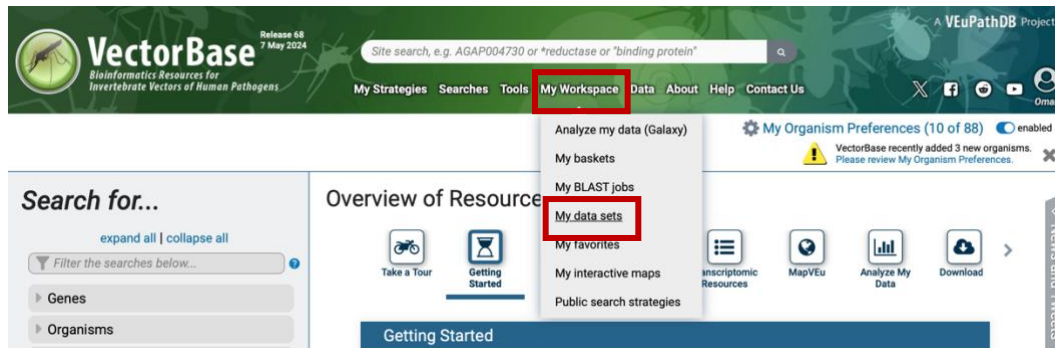
Yes No

Send an email notification when the job completes.

Execute

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

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	Required Resource	Required Resource Release	Installed Resource Release
VectorBase	IscapularisWikel Genome	49	49

Display a menu

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

Male vs Females

For the Experiment unstranded
return protein coding
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.

Product Description	Fold Change	Chosen Ref	Chosen Comp	Profiles
unspecified product	1963.7	2.57	5055.81	<p>ASTE008615 - UserDataset 4057359</p>
female reproductive tract protease GLEANR_896 [Source:Projected from Anopheles gambiae (AGAP00886...]	863.2	1.38	1187.11	<p>ASTE009598 - UserDataset 4057359</p>

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

Transcripts (UTRs in White when available)

ASTE008615
ASTE009598
ASTE008615-RA
ASTE009598-RA
unspecified product

gamma [Source:Projected from Anopheles gambiae...]

product

ASTE008615-RA
ASTE009598-RA
unspecified product

female1 female1.bw

female2 female2.bw

male1 male1.bw

male2 male2.bw

male3 male3.bw