RNA sequence data analysis via VectorBase Galaxy

Part II: View & interpret the results

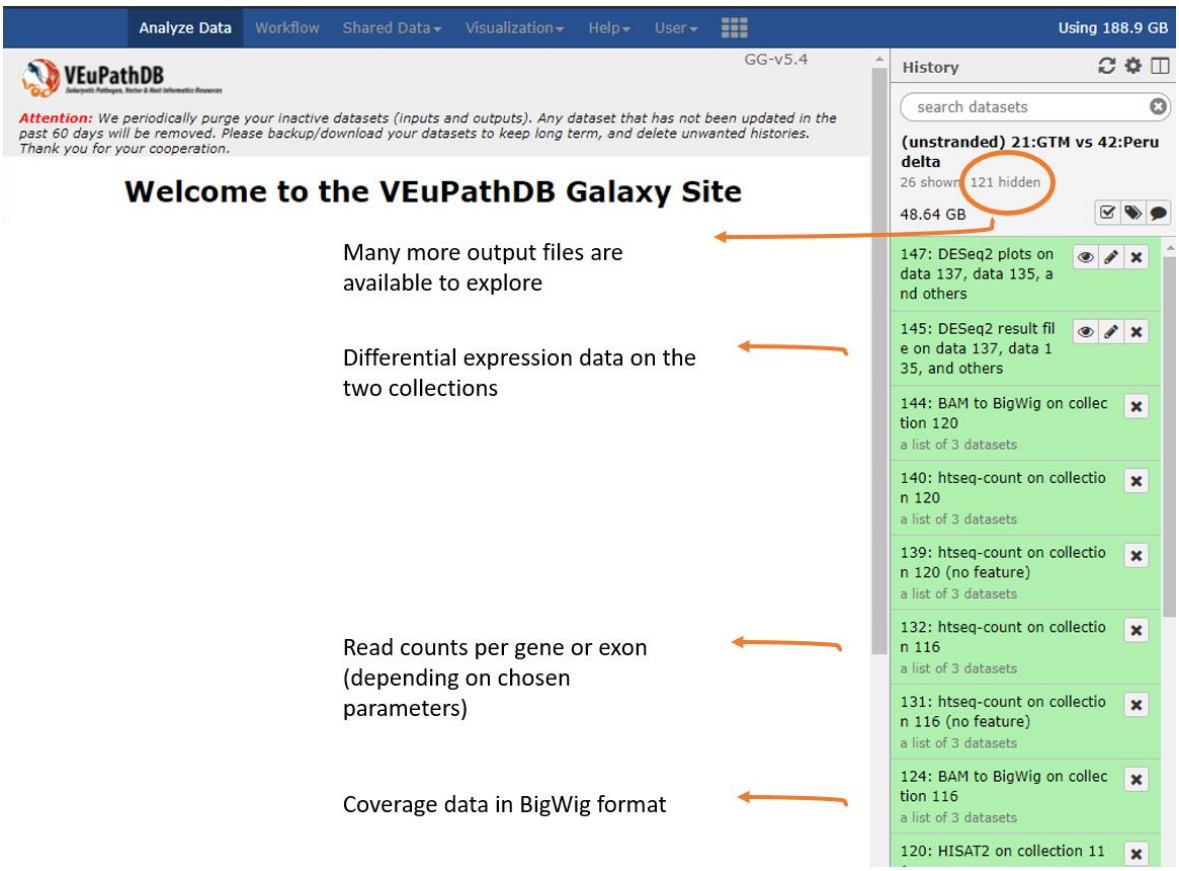
Learning objectives:

- Examine the results from the Galaxy RNA-Seq analysis workflow
- Import data from Galaxy to VectorBase My Workspace
- Analyze the results using VectorBase interface and tools

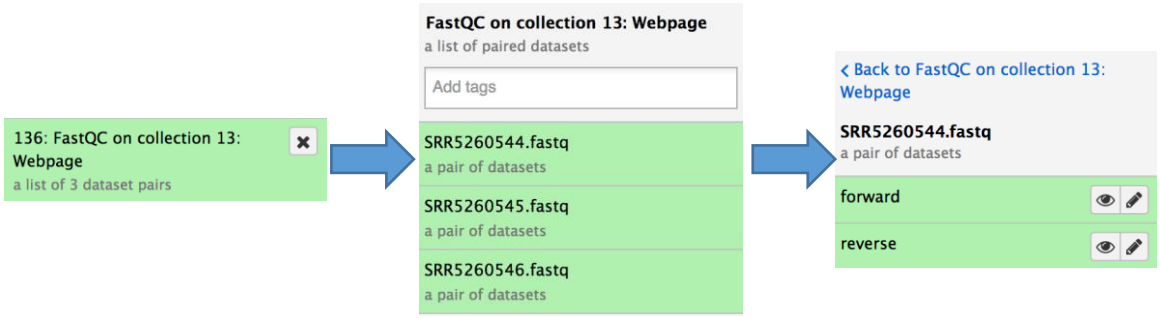
Additional resources:

- FastQC Result Interpretation: https://workshop.eupathdb.org/athens/2019/exercises/fastqc_results-2.pdf
- Beginner DESeq2 guide: https://bioc.ism.ac.jp/packages/2.14/bioc/vignettes/DESeq2/inst/doc/beginner.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

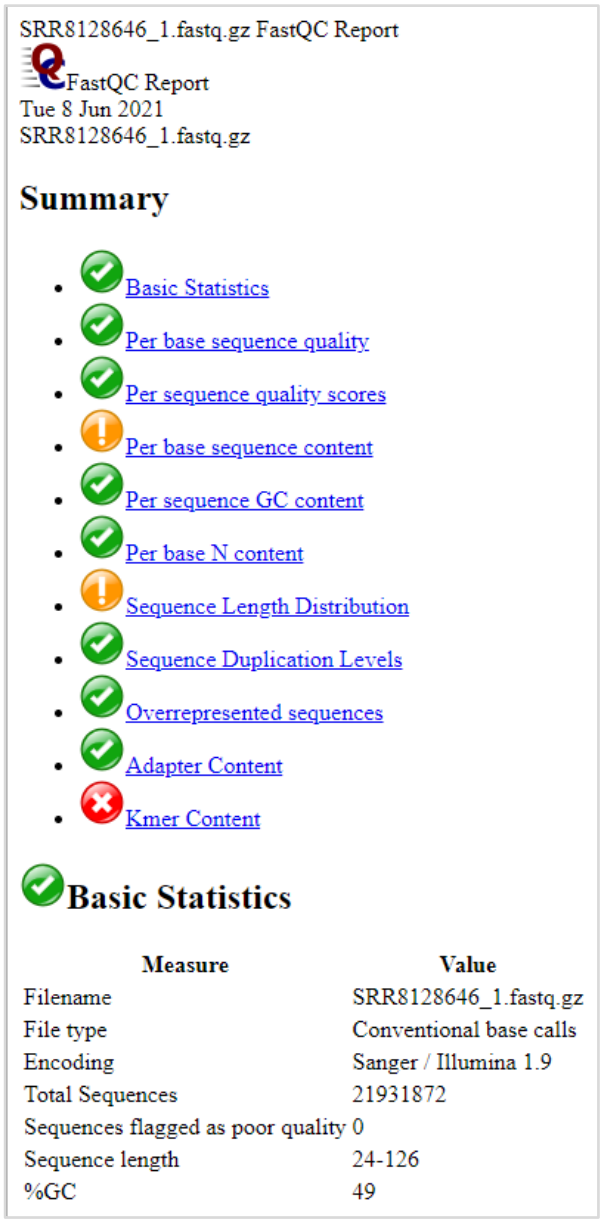
If everything worked out you should see a list of completed workflow steps (Green). The workflow generates many output files, however not all of the output files are visible. You can explore all the hidden files clicking on the word "hidden" (red circle) – this will reveal all hidden files.



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 first page of this tutorial or at the bottom of the FastQC results page.



SRR8128646_1.fastq.gz FastQC Report

FastQC Report
Tue 8 Jun 2021
SRR8128646_1.fastq.gz

Summary

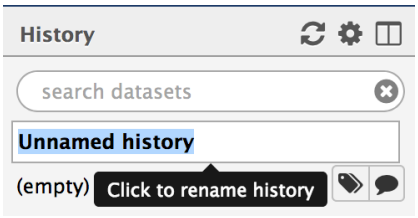
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

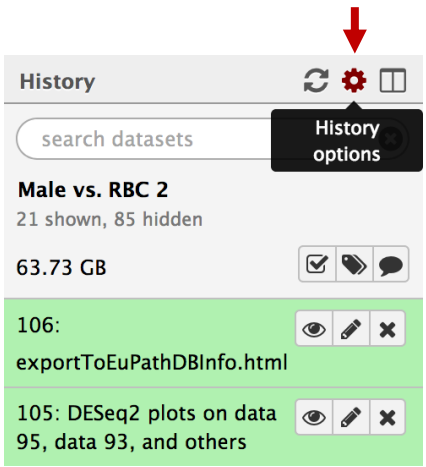
Measure	Value
Filename	SRR8128646_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	21931872
Sequences flagged as poor quality	0
Sequence length	24-126
%GC	49

Step 2: Sharing histories with others:

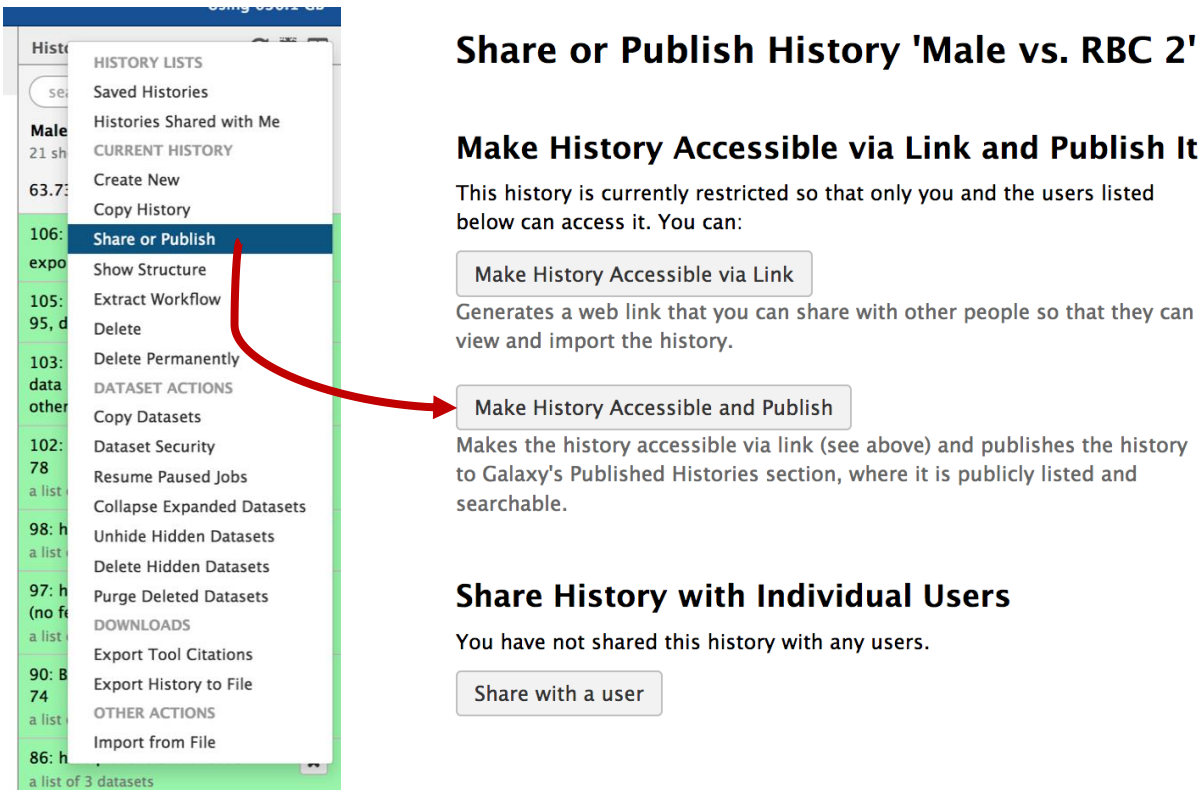
a. Make sure your history has a useful name – you can change the name by clicking on "unnamed history"



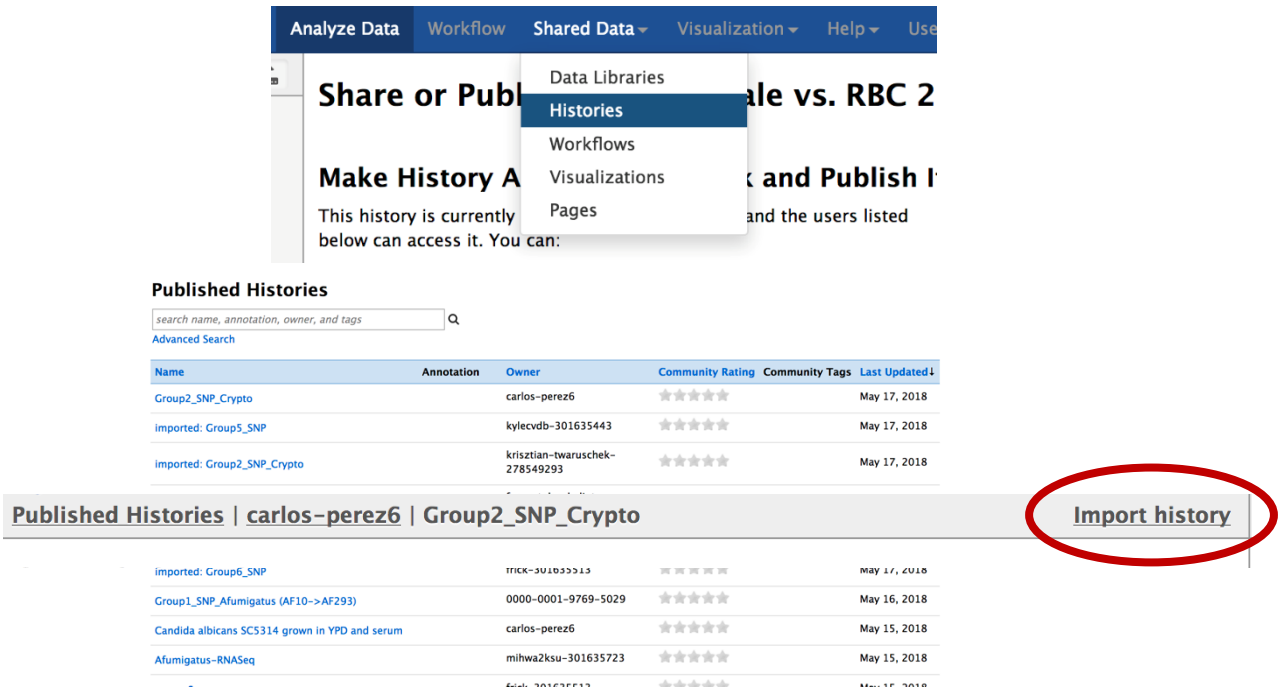
b. Click on the history options menu icon



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.



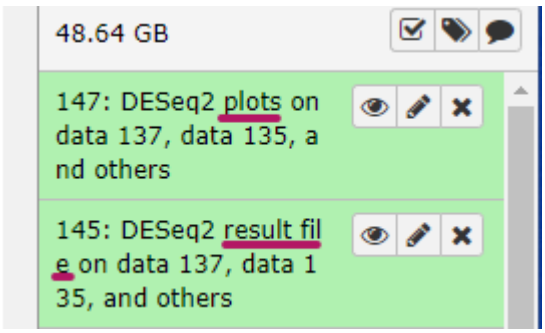
f. Click on the import link.

Step 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.

We will explore two output files:

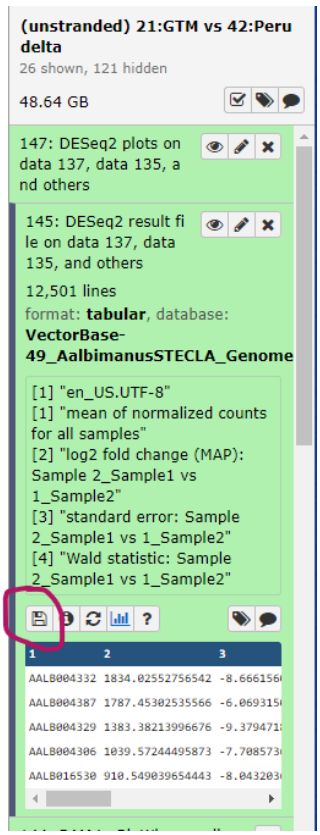
A. DESeq2 **Plots** – you can view these directly in galaxy by clicking on the view icon. These plots give you an idea about the quality of the experiment. This program user guide (link in the 1st page of this tutorial), includes a detailed description of the graphs.
B. DESeq2 **results file** – this 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 an spreadsheet program (e.g. Excel) so you can sort results and genes of interest.



The tabular file contains 7 columns:

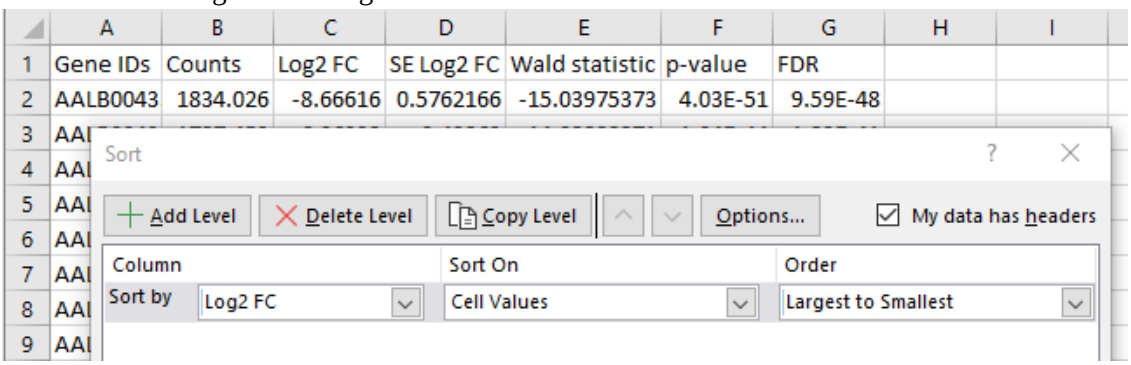
COLUMN	DESCRIPTION
1	Gene Identifiers
2	mean normalized counts, averaged over all samples from both conditions
3	the logarithm (to basis 2) of the fold change (See the note in inputs section)
4	standard error estimate for the log2 fold change estimate
5	Wald statistic
6	p value for the statistical significance 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.

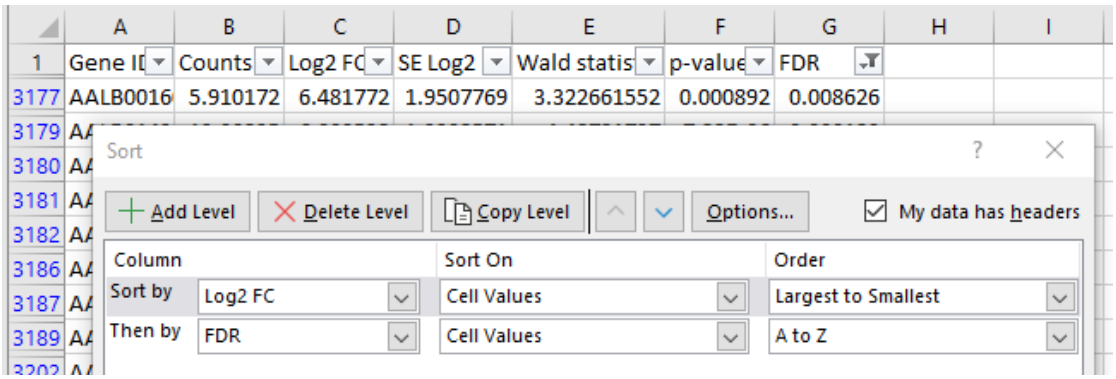


***** important: the file name ends with the extension .tabular – change this to .txt then open the file in Excel.**

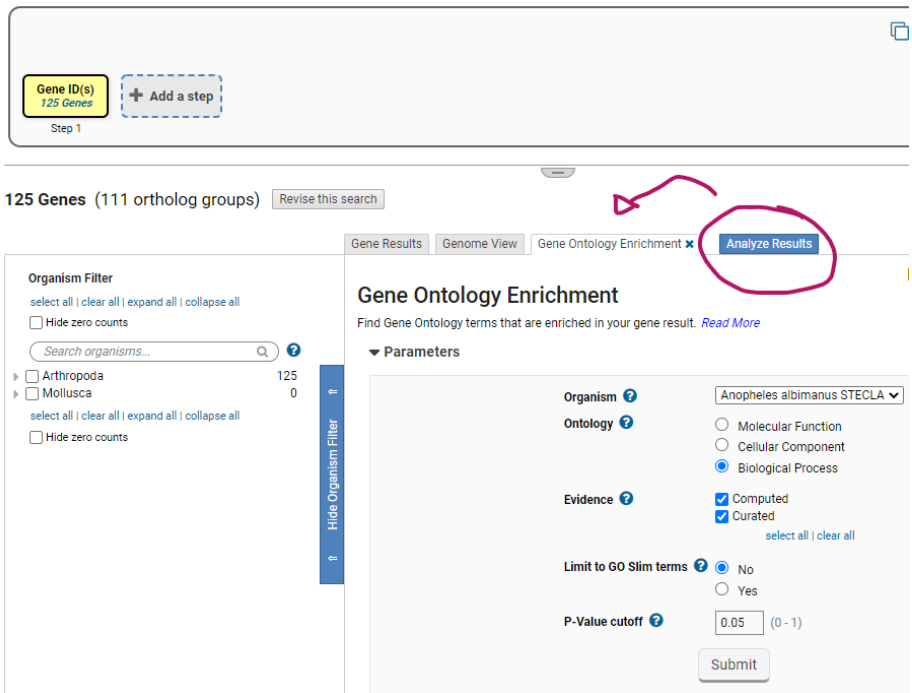
D. Type the column headers. Explore the results in Excel. For example, sort them based on the log2 fold change – column 3.



E. Pick a list of gene IDs from column 3 that are up-regulated with a good corrected P-value (column 7; Filter the NA values) and load then into VectorBase using the Gene by ID search.



F. You can then analyze these results by GO enrichment for example. Do the same for down-regulated genes.



G. Can you find genes are that are uniquely up or down regulated in the conditions tested?

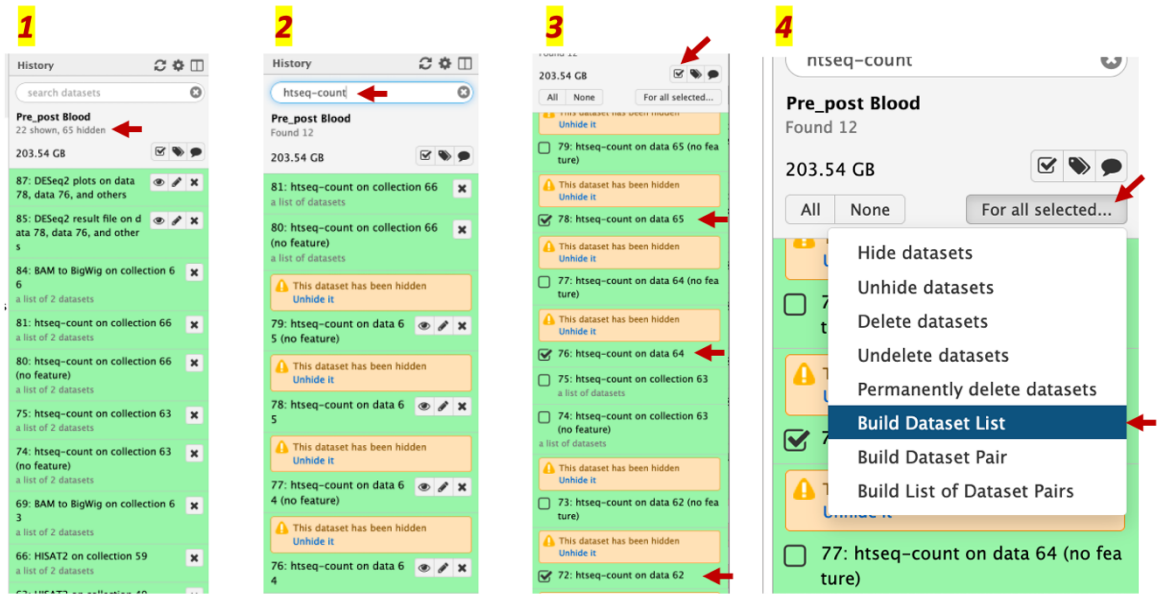
Exporting data to VEuPathDB

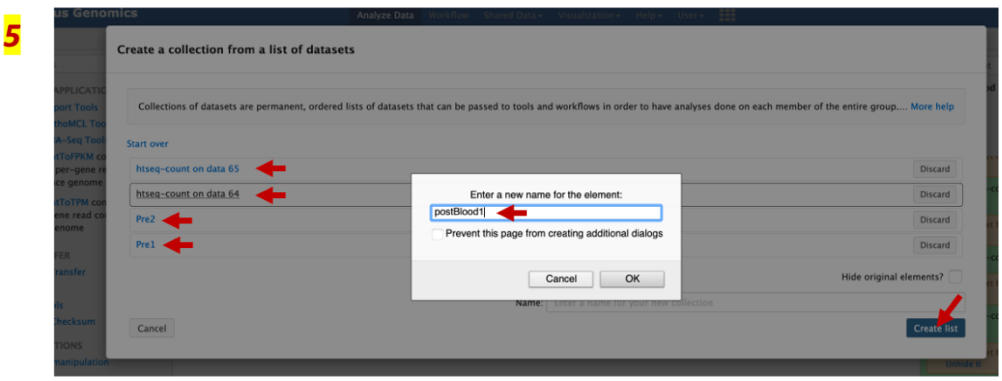
The VEuPathDB RNAseq export tool provides a mechanism to export your RNAseq results (TPM values) and BigWig RNAseq coverage files. The advantage of doing this is that it allows you to search the TPM data using the RNAseq search in VEuPathDB and view the BigWig files in the genome browser.

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.

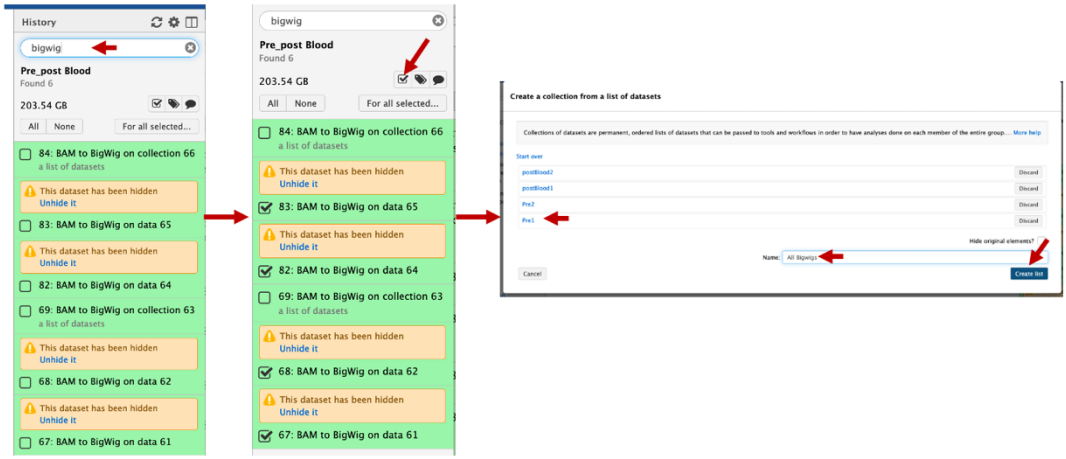
First let's organize the files (see matching screen shots below):

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".
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 all 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.

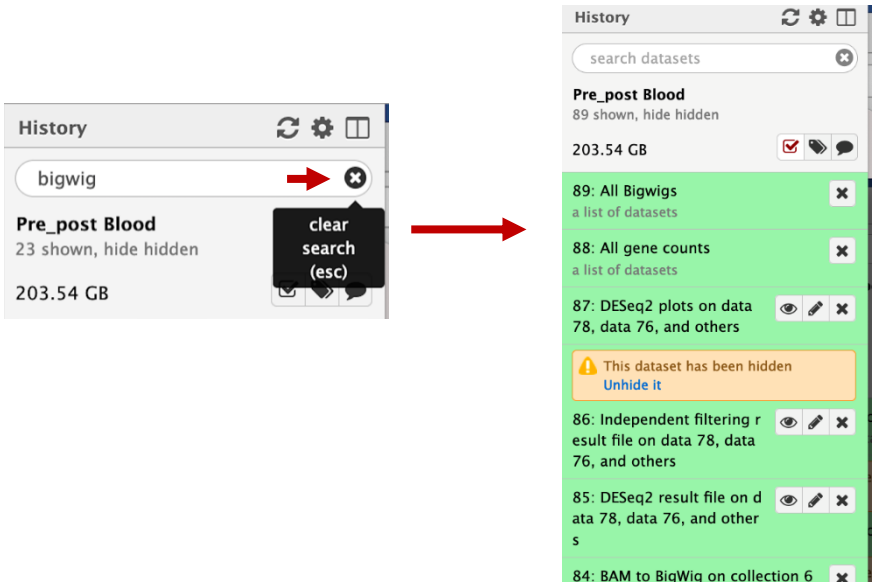




6. Repeat the same steps to create the list of BigWig files (See screen shots).



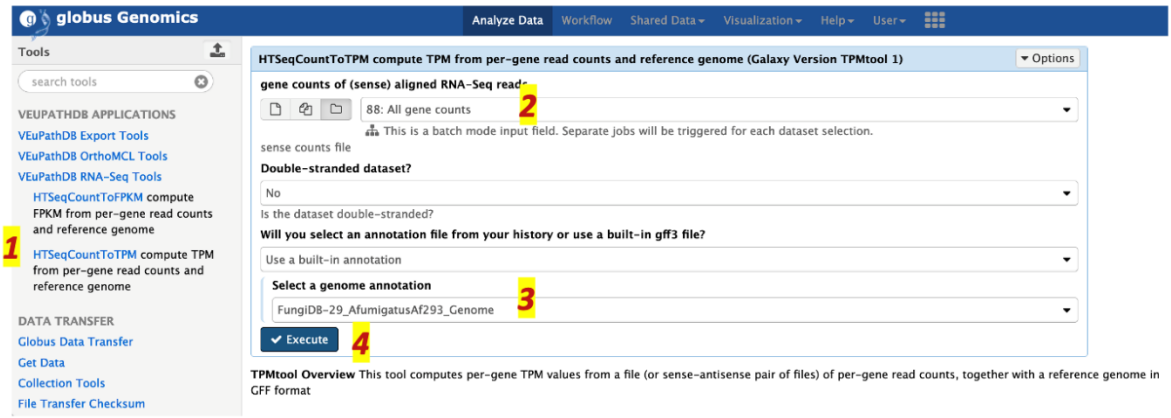
7. Click on clear search to see all results in your history.



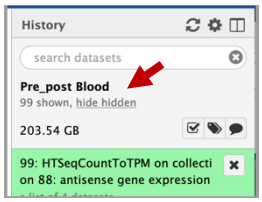
Now that your count and bigwig files are nice and organized, the next step is to convert the counts into TPMs. To do this follow these steps:

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.

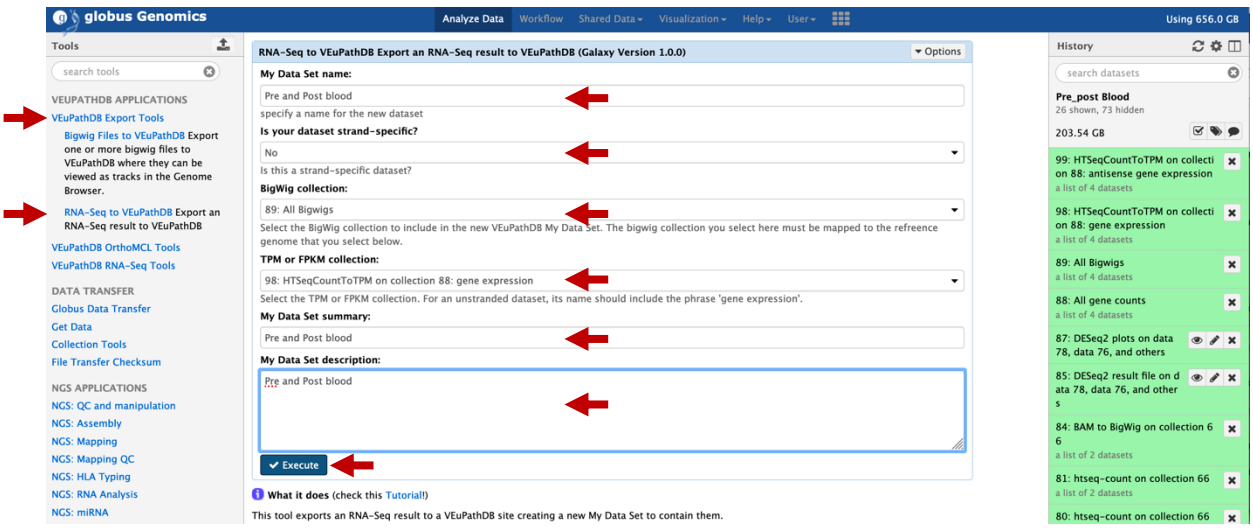


Optional: Click on "hide hidden" to clean up your history a bit.



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

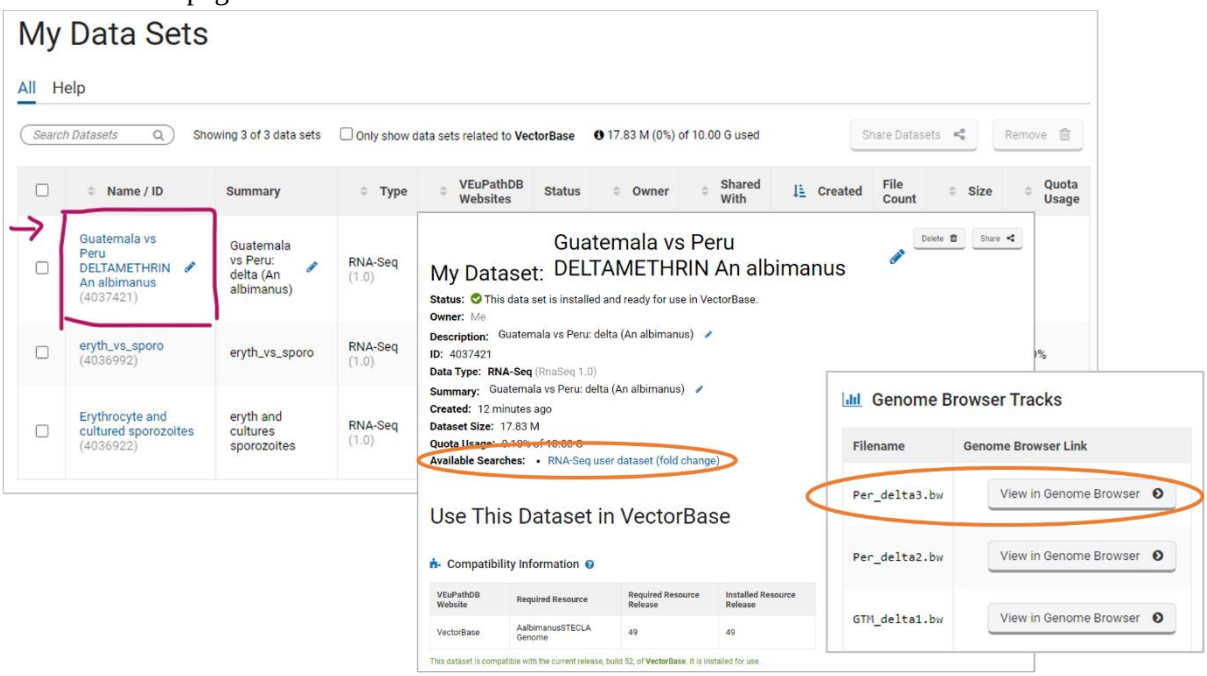
1. Click on "VEuPathDB Export Tools" in the left-hand panel.
2. Click on the tool called "RNA-Seq to VEuPathDB"
3. Fill up the export tool and select the correct files to export (see screen shot).



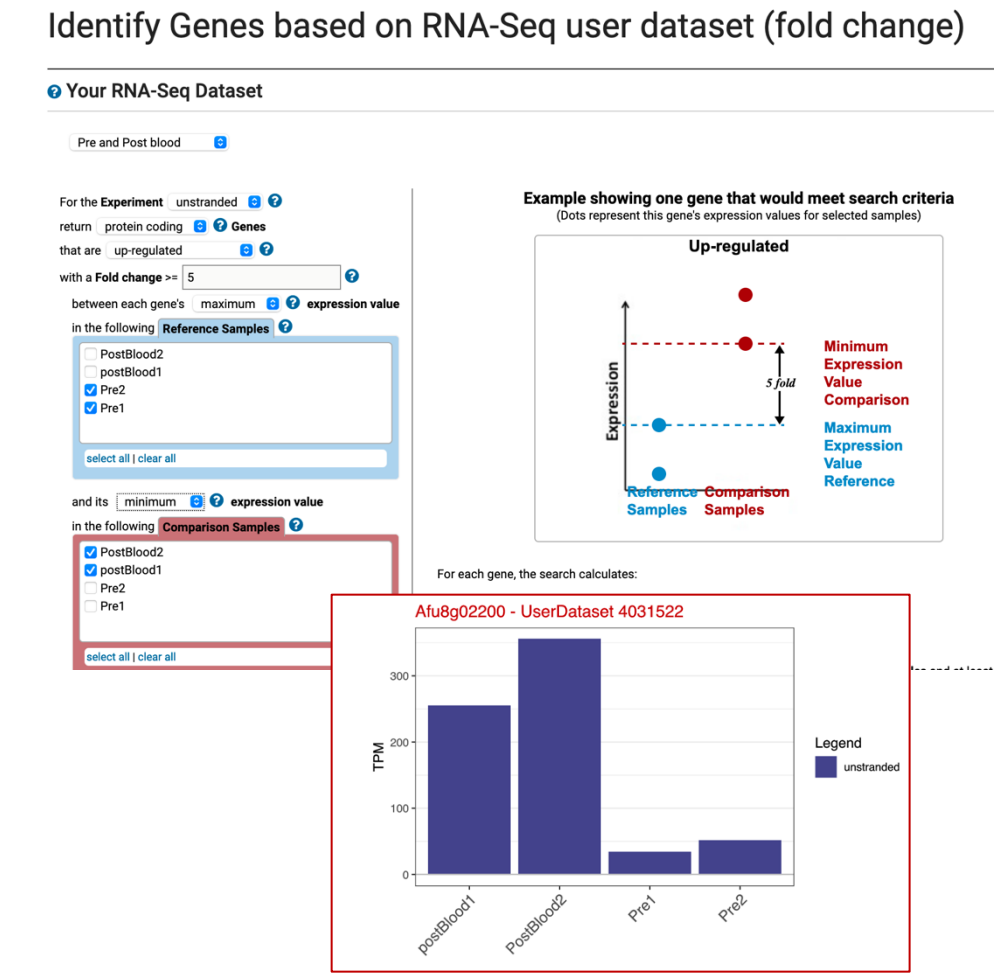
Explore your data in VEuPathDB: Go to the VEuPathDB database that your data belongs to (e.g. VectorBase).

1. Click on “My Workspace” > “My datasets”.

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



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.