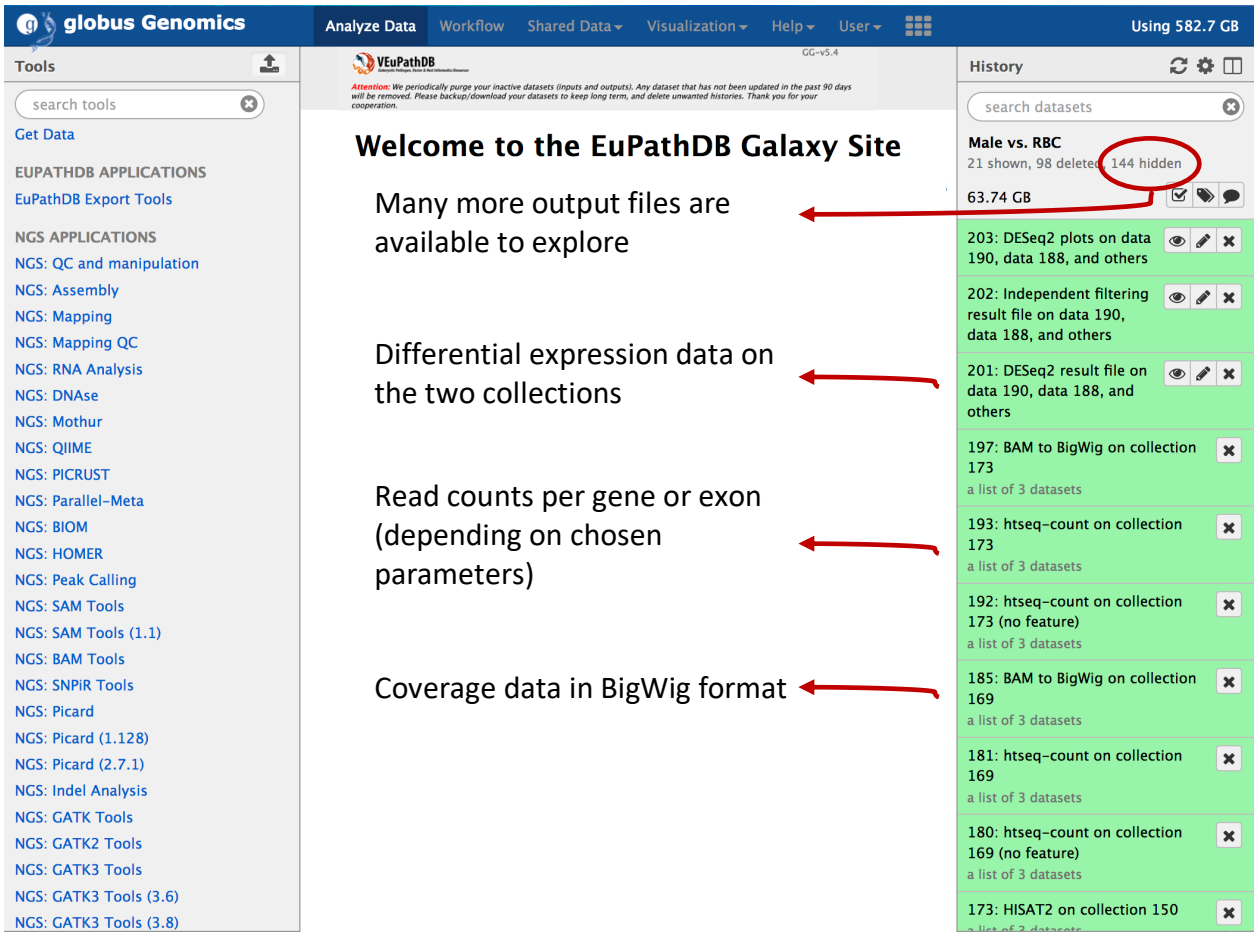
# RNA sequence data analysis via Galaxy, Part II Uploading data and starting the workflow (Group Exercise)

**Learning objectives:**

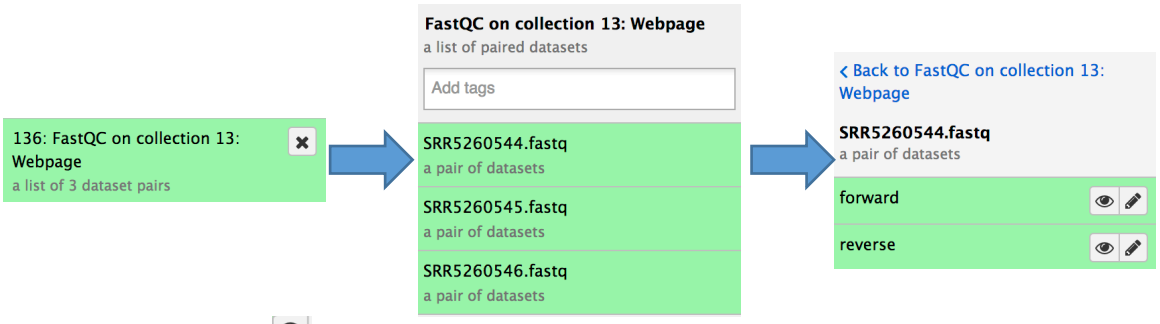
- Examine Galaxy output results
- Analyze output results
- Export TPM data to the My Datasets section of the VEuPathDB workspace

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.

Resources:
FastQC Result Interpretation
(https://workshop.VEuPathDB.org/athens/2019/exercises/fastqc_results-2.pdf)
Beginner DESeq2 guide (https://workshop.VEuPathDB.org/athens/2019/exercises/beginner_DeSeq2.pdf)
FastQC output
(https://workshop.VEuPathDB.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

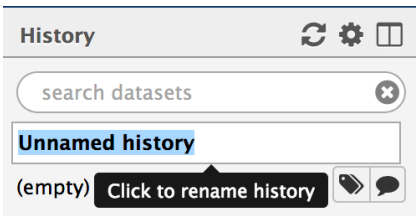
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

## Basic Statistics

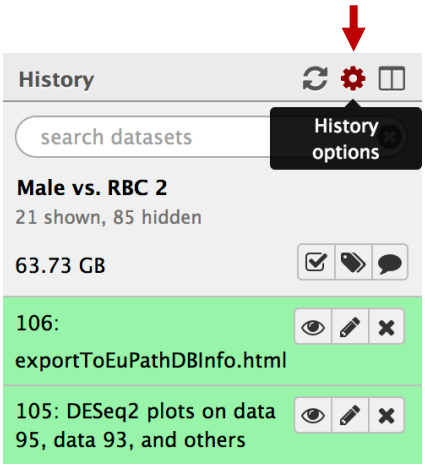
| Measure | Value |
|---|---|
| Filename | SRR5260544_1.fastq.gz |
| File type | Conventional base calls |

Step 2: Sharing histories with others:

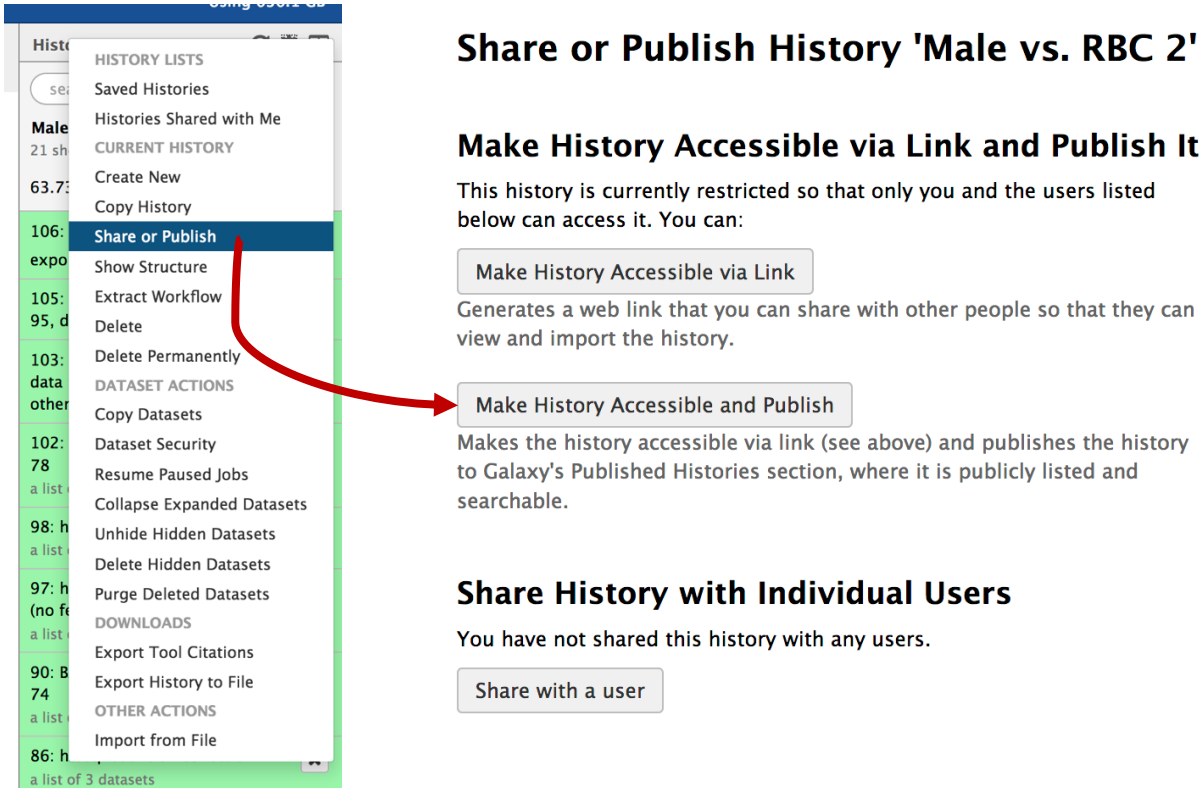
a. Make sure your history has a useful name – you can change the name by clicking on “unnamed history”



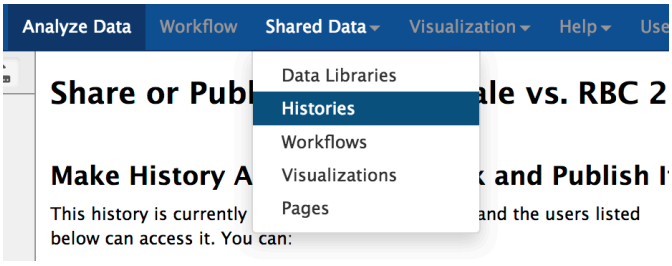

b. Click on the history options menu icon



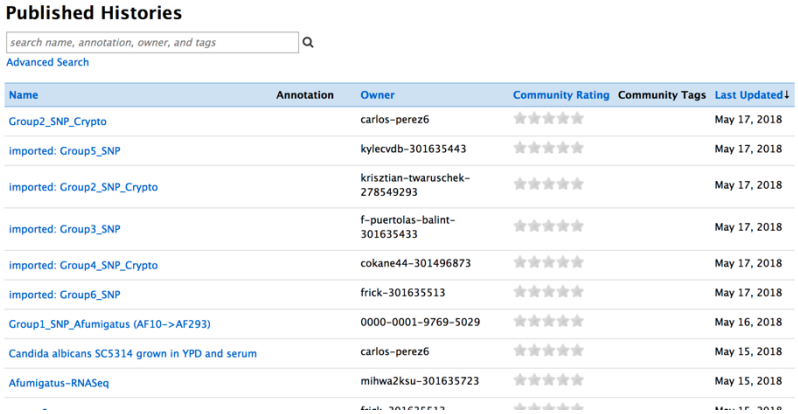

c. Select the “Share or Publish” option, the click on the “Make History Accessible and Publish” button in the center section.



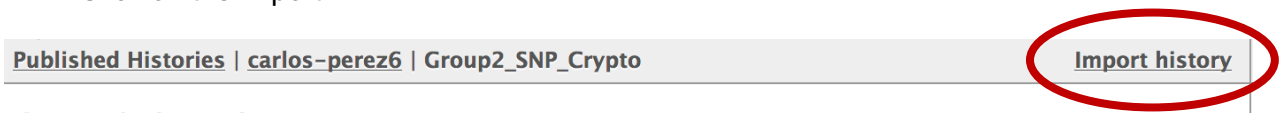
d. To import a shared history, go to the “histories” section (under the shared data menu item).




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. You can explore details of DESeq2 here: https://bioc.ism.ac.jp/packages/2.14/bioc/vignettes/DESeq2/inst/doc/beginner.pdf

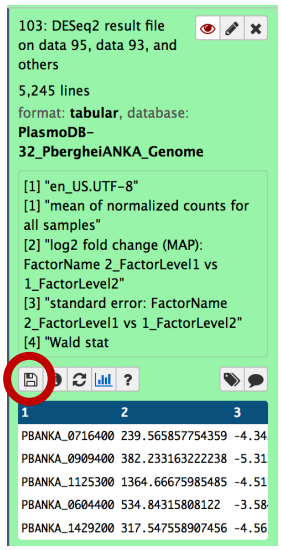
We will explore two output files:

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 query your RNAseq results (TPM values) using VEuPathDB search tools.

However, to use this feature you need to do two things:

1. Generate TPM values for genes in your datasets from your htseq-count output files.
2. Put all the TPM files you want to query in the same search in VEuPathDB into a single collection.

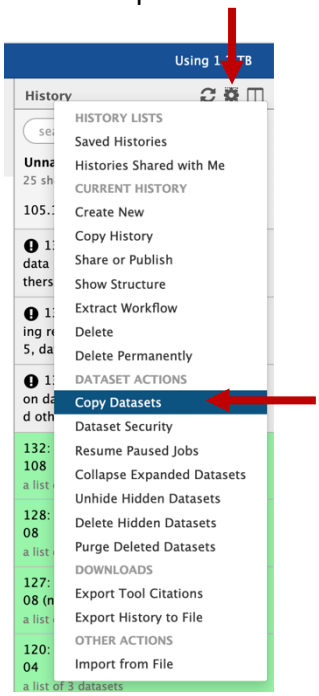
Follow these steps to generate TPM values:

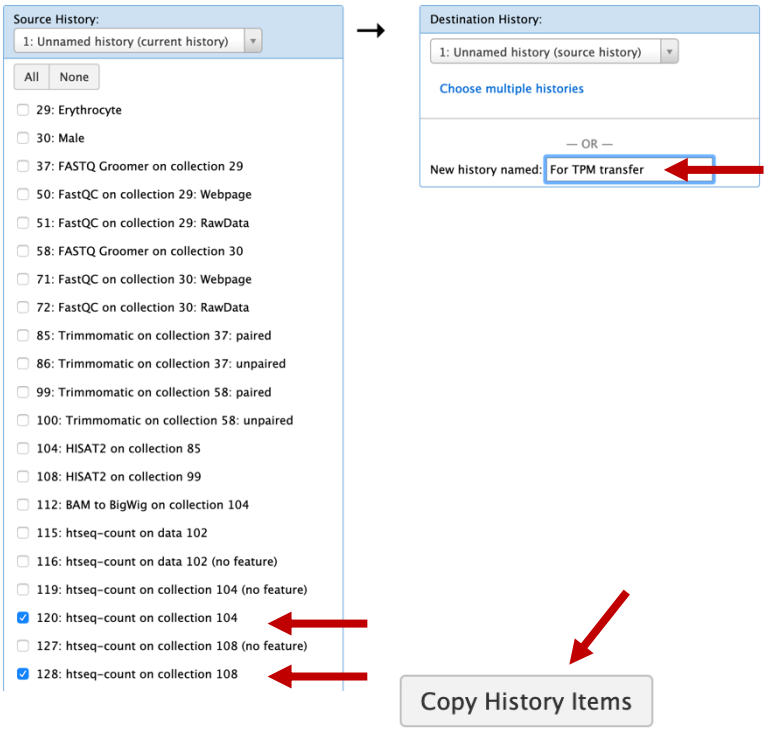
1. Copy the htseq-count collections called "htseq-count on collection ###" to a new history. Here the ### refers to the output file in your history that htseq-count was performed on. For this exercise you should have two collections that you are copying E.g.
   - htseq-count on collection 104
   - htseq-count on collection 108

   Follow these steps to copy:

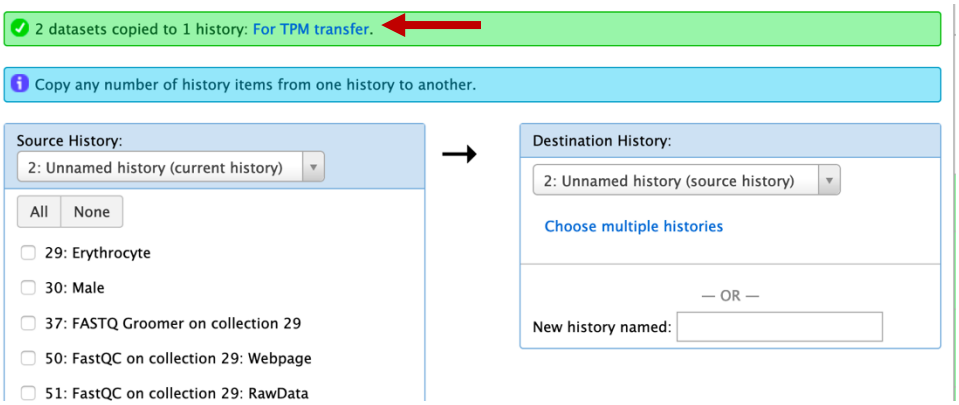
   a. select history options and click on the Copy Datasets option.

   b. On the next page select the collections you want to copy and if you want, give the new history a name (see figure below).

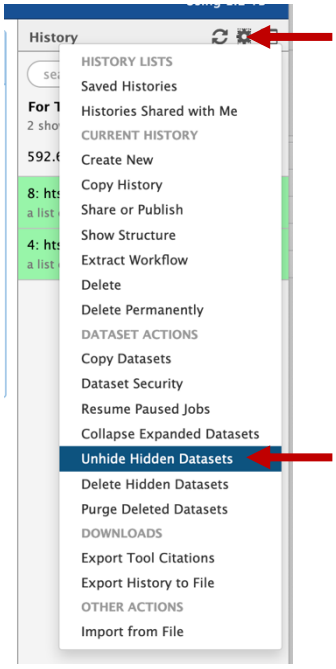


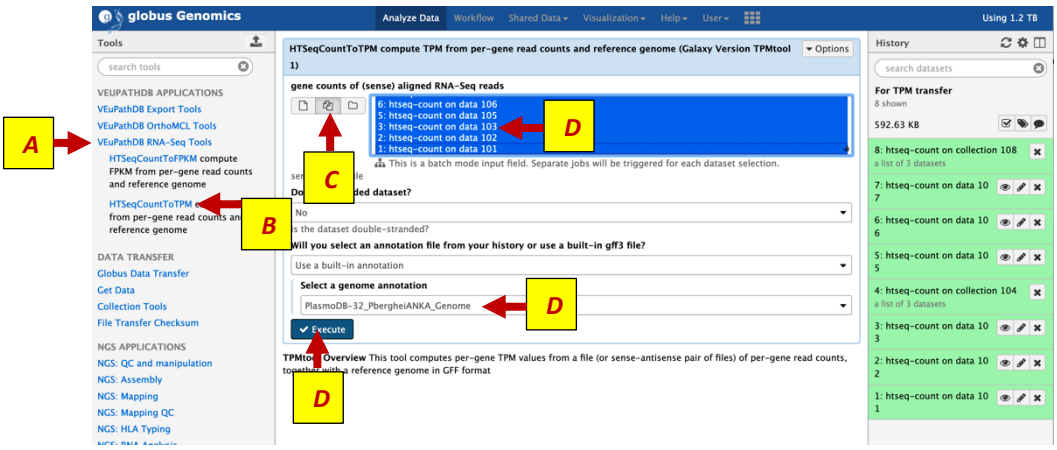
2. Go to the new history you just created. You can do this by clicking on “dataset copied” link at the top of the page.




3. Select the history options menu item and click on “unhide hidden datasets” then click ok. You should see several datasets appear in your history. These are the count files, there should be one for each sample/condition. So if you had 3 samples in each collection and two collections, you should end up with an additional six datasets appearing in your history.

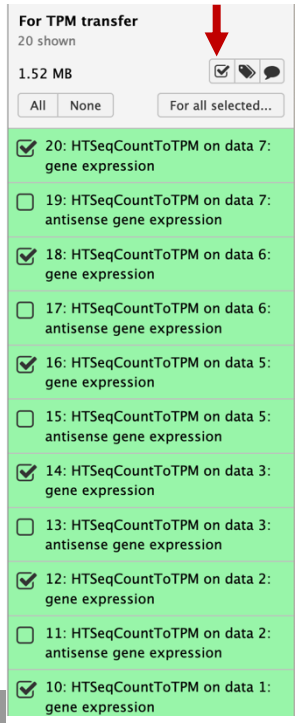


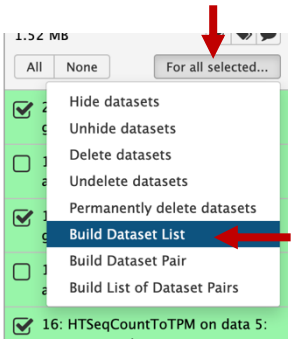

4. Now follow these steps (See figure below):
    - Click on the left-hand menu item called “VEuPathDB RNA-Seq tools” (A).
    - Select the tool called ***HTSeqCountToTPM*** (B).
    - In the middle window select the multiple datasets option (C).
    - Select all the datasets that appear in the box (D).
    - Select the correct reference annotation (E).
    - Click on execute (F).



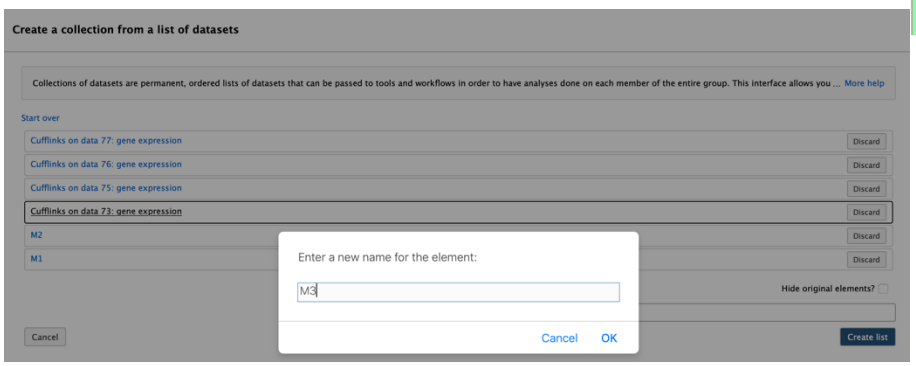

- Calculating TPM values will take a few minutes. Once this is done add the TPM files to a single collection.

- Click on the check box to perform an operation on multiple datasets (arrow in image).
- Select all files containing the words “gene expression”. Do not select the antisense gene expression files.
- Click on the “For all selected” button and select “Build dataset list”.




- Rename each of the datasets in the list and give this collection a meaningful name.

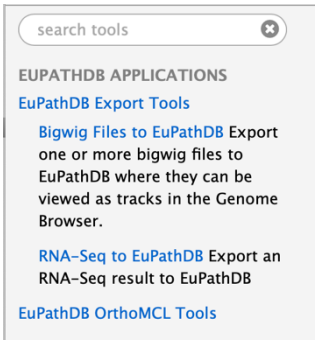


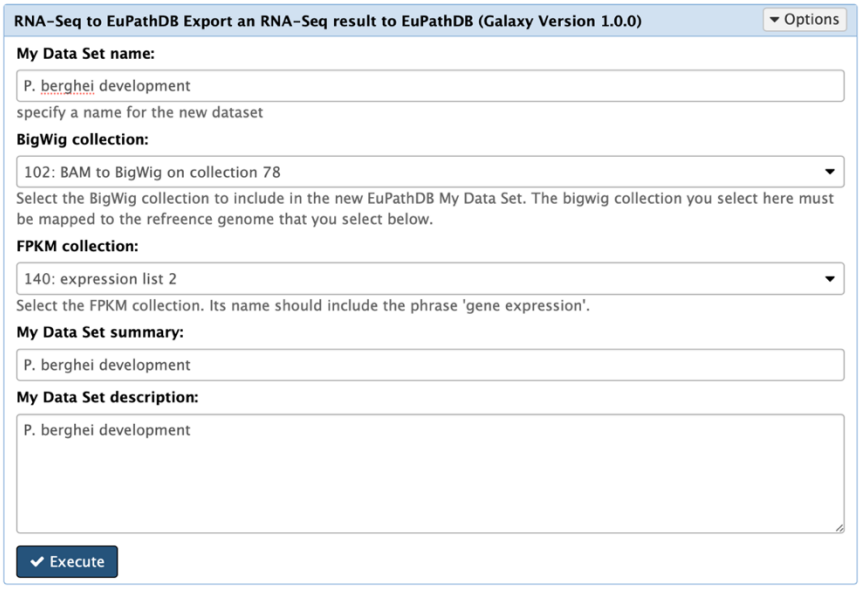

Step 4: Copy a bigwig collection from your original history.

1. Go back to you previous history which contains your entire DEseq analysis.
2. Copy one of the bigwig collections to the history that contains your TPM files.

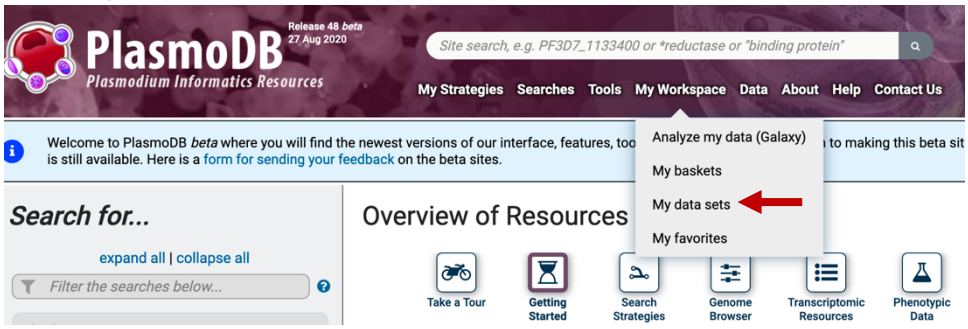
Step 4: Export Expression files to VEuPathDB

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.
4. Click on the “My Datasets” link in the grey menu bar. You should see the dataset you exported from galaxy in this list. Click on it and explore the dataset page.
5. Click on Execute and wait for the export step to complete.
6. When export is complete, go to the VEuPathDB website with the genomes for this data, e.g PlasmoDB.
7. Go to the My datasets section.




8. Click on the available search and explore this page. Can you run a search to identify genes differentially expressed between the two conditions you analyzed in Galaxy? How do these compare to the results you got from DEseq2?