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

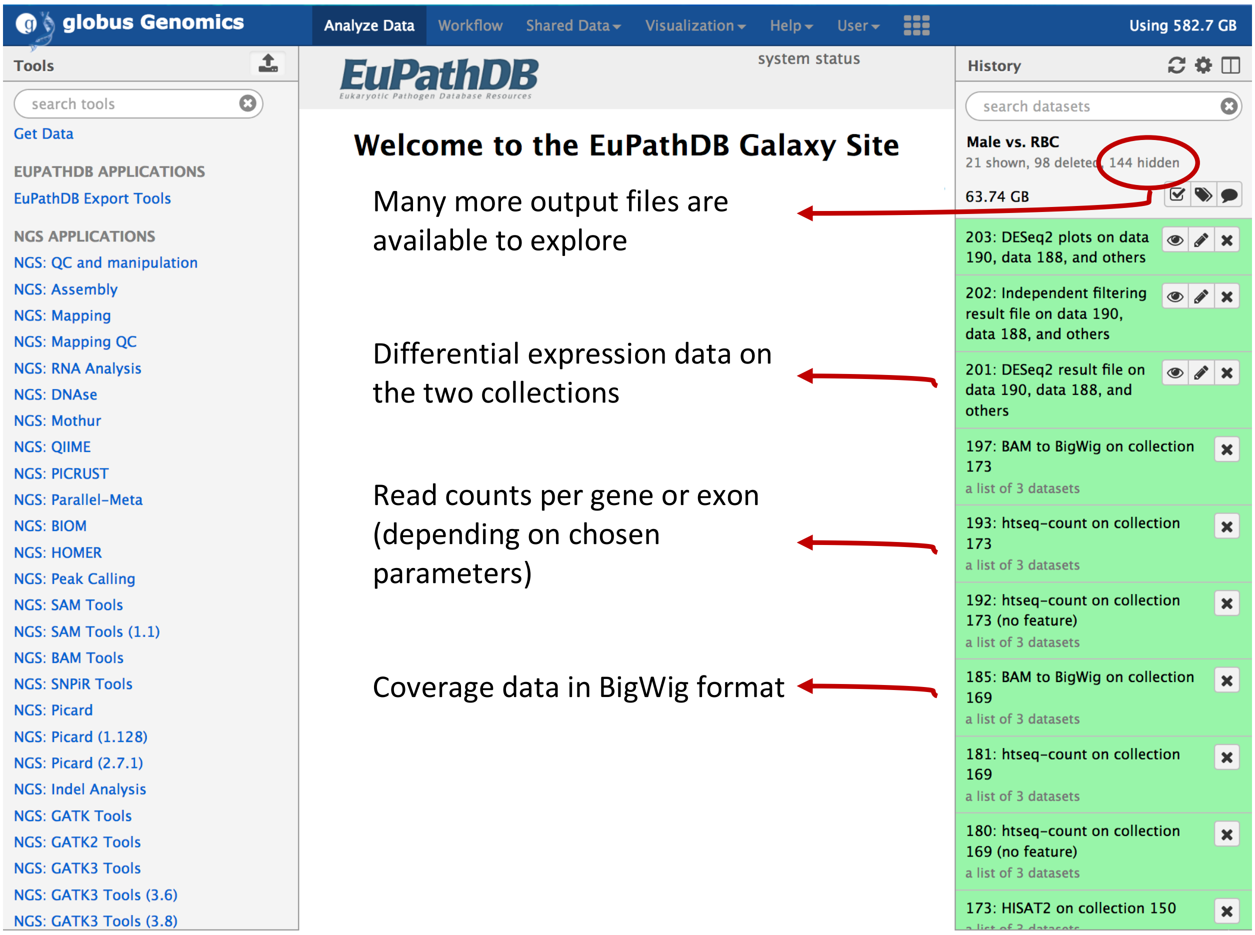
Learning objectives:

- examine the results from the Galaxy RNA-Seq analysis workflow
- Import data from Galaxy to FungiDB My Workspace
- Analyse the results using FungiDB interface and tools
-

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)
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)



Welcome to the EuPathDB Galaxy Site

Many more output files are available to explore

Differential expression data on the two collections

Read counts per gene or exon (depending on chosen parameters)

Coverage data in BigWig format

Male vs. RBC
21 shown, 98 deleted, 144 hidden
63.74 GB

203: DESeq2 plots on data 190, data 188, and others

202: Independent filtering result file on data 190, data 188, and others

201: DESeq2 result file on data 190, data 188, and others

197: BAM to BigWig on collection 173
a list of 3 datasets

193: htseq-count on collection 173
a list of 3 datasets

192: htseq-count on collection 173 (no feature)
a list of 3 datasets

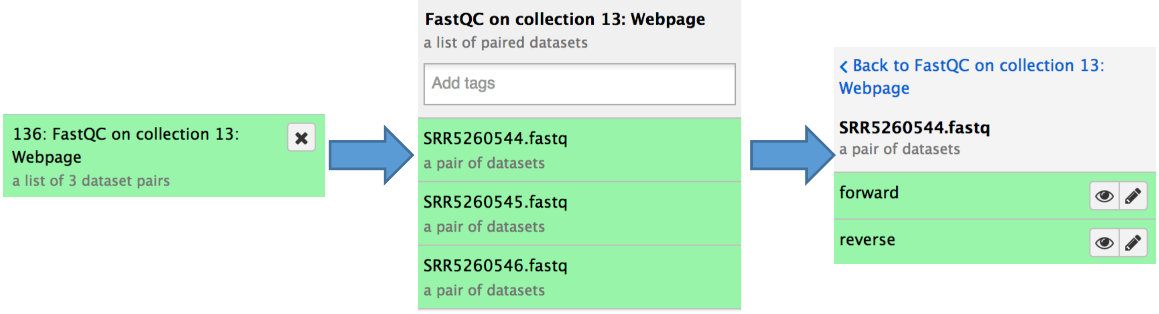
185: BAM to BigWig on collection 169
a list of 3 datasets

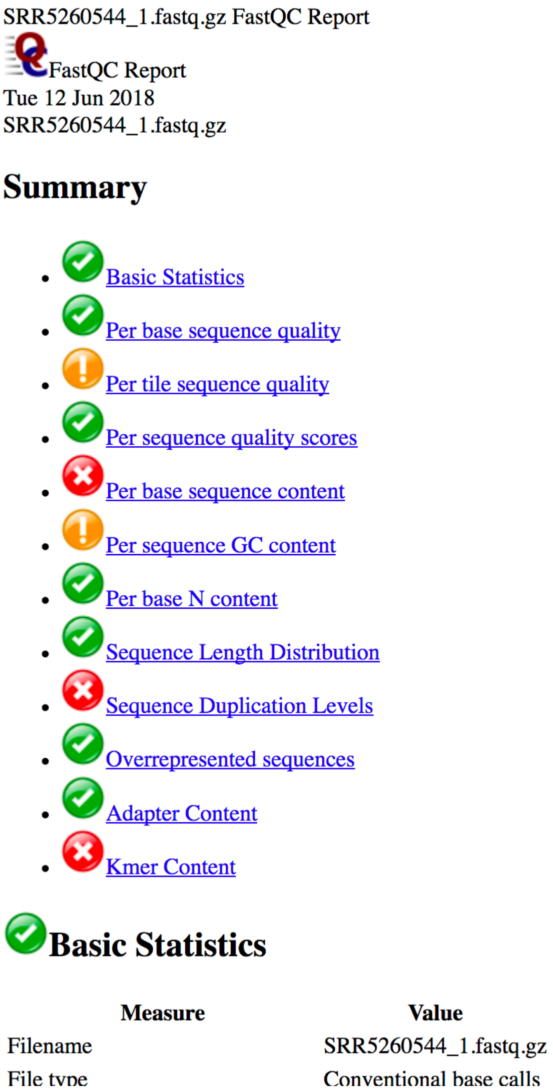
181: htseq-count on collection 169
a list of 3 datasets

180: htseq-count on collection 169 (no feature)
a list of 3 datasets

173: HISAT2 on collection 150

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.


136: FastQC on collection 13: Webpage
a list of 3 dataset pairs
✖
FastQC on collection 13: Webpage
a list of paired datasets
Add tags
SRR5260544.fastq
a pair of datasets
SRR5260545.fastq
a pair of datasets
SRR5260546.fastq
a pair of datasets
< Back to FastQC on collection 13: Webpage
SRR5260544.fastq
a pair of datasets
forward
reverse



SRR5260544_1.fastq.gz FastQC Report

FastQC Report
Tue 12 Jun 2018
SRR5260544_1.fastq.gz

Summary

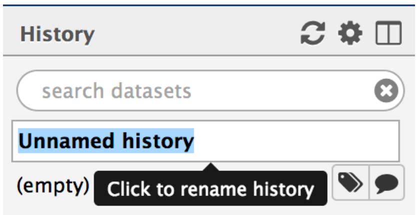
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ! Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter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

✓ Basic Statistics

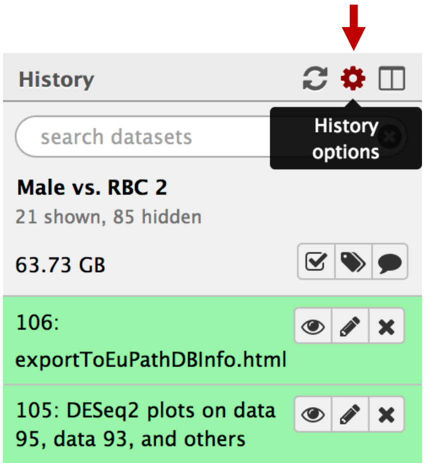
Measure	Value
Filename	SRR5260544_1.fastq.gz
File type	Conventional base calls

Step 2: Sharing histories with others:

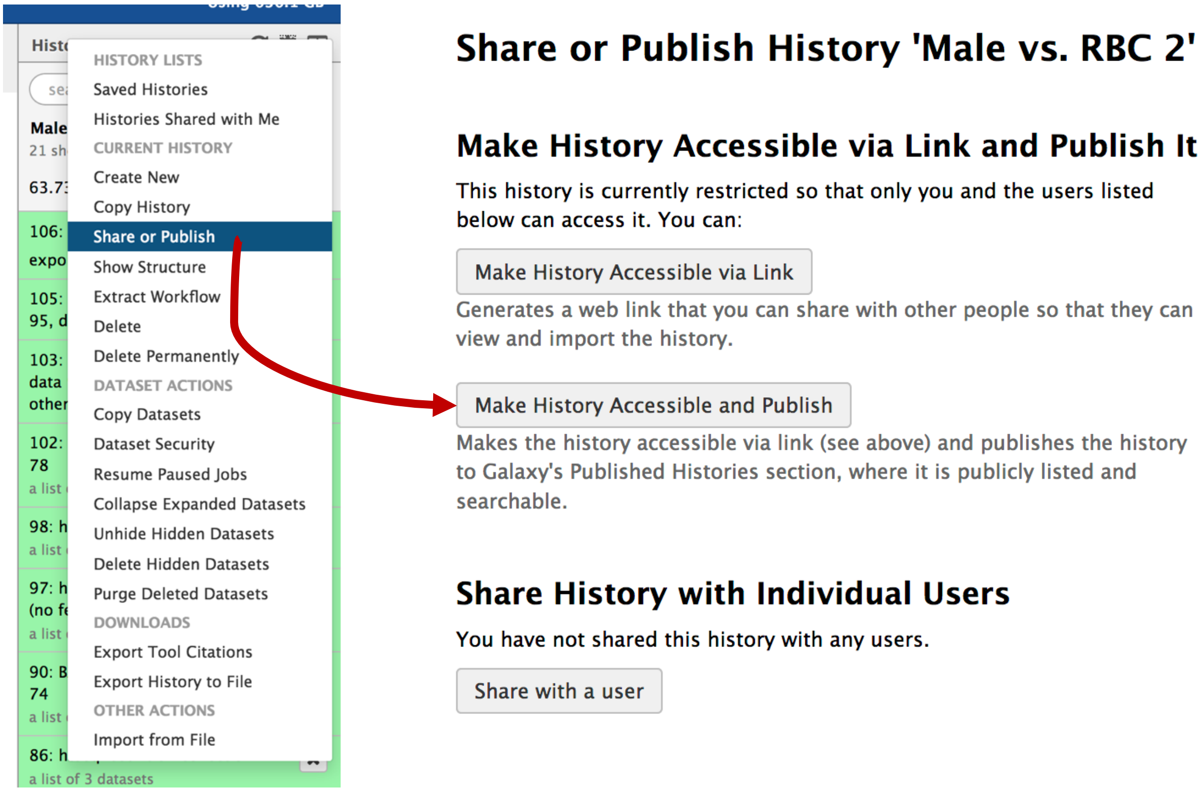
a. Make sure your history has a useful name – you can change the name by clicking on “unnamed history”



b. Click on the history options menu icon


History
search datasets
History options
Male vs. RBC 2
21 shown, 85 hidden
63.73 GB
106: exportToEuPathDBInfo.html
105: DESeq2 plots on data 95, data 93, and others

c. Select the “Share or Publish” option, the 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:

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

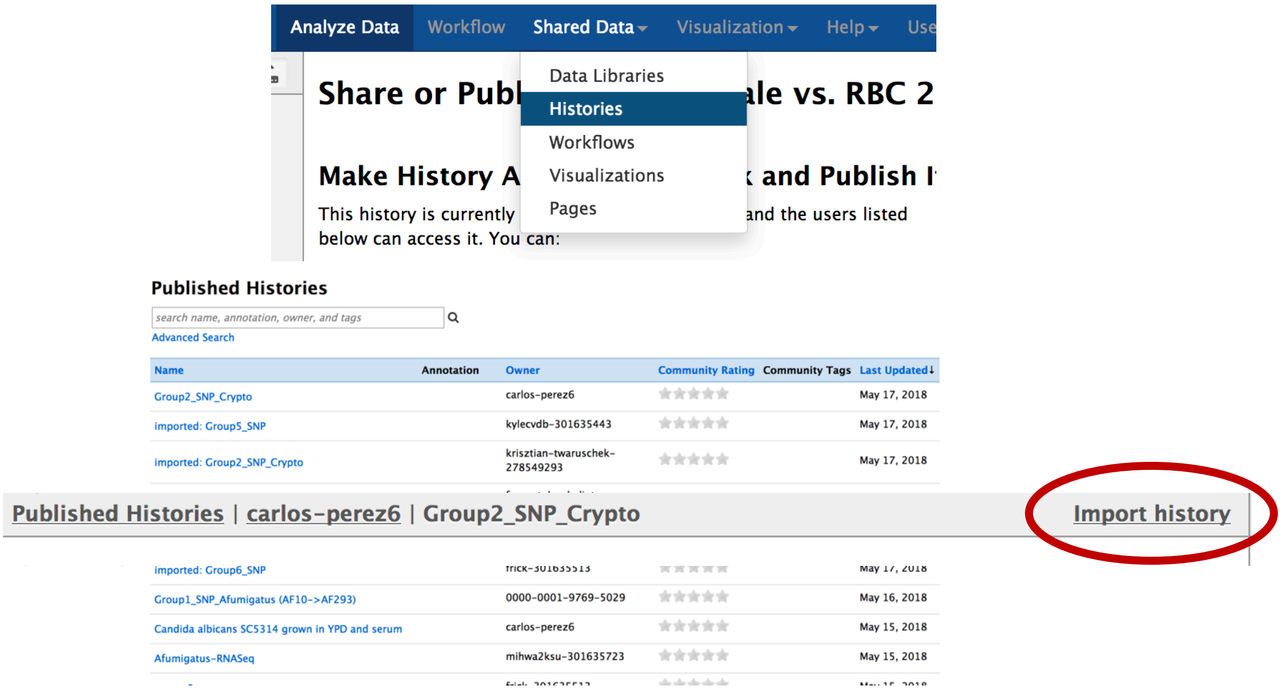
Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, where it is publicly listed and searchable.

Share History with Individual Users

You have not shared this history with any users.

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.



Share or Publ... ale vs. RBC 2

Make History A... k and Publish I

This history is currently ... and the users listed below can access it. You can:

Published Histories

search name, annotation, owner, and tags

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-twaruschek-278549293	☆☆☆☆☆		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 SC5314 grown in YPD and serum		carlos-perez6	☆☆☆☆☆		May 15, 2018
Afumigatus-RNASeq		mihwa2ksu-301635723	☆☆☆☆☆		May 15, 2018
		frick-301635513	☆☆☆☆☆		May 15, 2018

Published Histories | carlos-perez6 | Group2_SNP_Crypto

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:

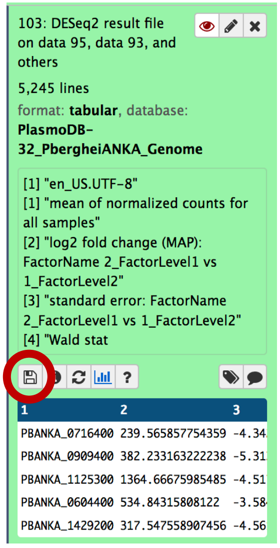
A. DESeq2 Plots – you can view these directly in galaxy by clicking on the view icon. These plots give you an idea about the quality of the experiment. The link above includes a detailed description of the graphs.

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.



103: DESeq2 result file on data 95, data 93, and others

5,245 lines

format: **tabular**, database: **PlasmoDB-32_PbergheiANKA_Genome**

```
[1] "en_US.UTF-8"
[1] "mean of normalized counts for all samples"
[2] "log2 fold change (MAP): FactorName 2_FactorLevel1 vs 1_FactorLevel2"
[3] "standard error: FactorName 2_FactorLevel1 vs 1_FactorLevel2"
[4] "Wald stat
```

1	2	3
PBANKA_0716400	239.565857754359	-4.34
PBANKA_0909400	382.233163222238	-5.31
PBANKA_1125300	1364.66675985485	-4.51
PBANKA_0604400	534.84315808122	-3.58
PBANKA_1429200	317.547558907456	-4.56

***** 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

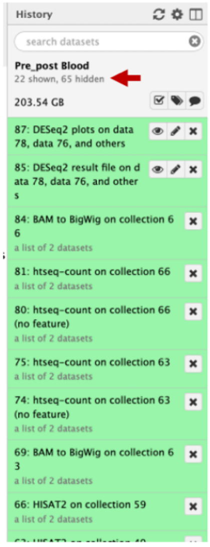
The VEuPathDB RNAseq export tool provides a mechanism to export your RNAseq results (TPM values) and BigWig RNAseq coverage files. The advantage of doing this is that it allows you to search the TPM data using the RNAseq search in VEuPathDB and view the BigWig files in the genome browser.

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".
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.

1



Pre_post Blood
22 shown, 65 hidden
203.54 GB

87: DESeq2 plots on data 78, data 76, and others

85: DESeq2 result file on data 78, data 76, and others

84: BAM to BigWig on collection 66
a list of 2 datasets

81: htseq-count on collection 66
a list of 2 datasets

80: htseq-count on collection 66 (no feature)
a list of 2 datasets

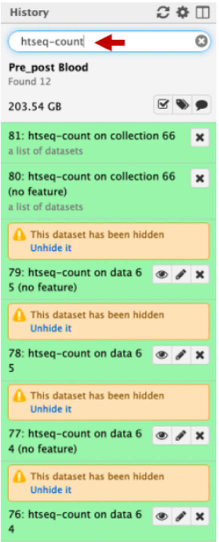
75: htseq-count on collection 63
a list of 2 datasets

74: htseq-count on collection 63 (no feature)
a list of 2 datasets

69: BAM to BigWig on collection 63
a list of 2 datasets

66: HISAT2 on collection 59
a list of 2 datasets

2



Pre_post Blood
Found 12
203.54 GB

81: htseq-count on collection 66
a list of datasets

80: htseq-count on collection 66 (no feature)
a list of datasets

This dataset has been hidden
Unhide it

79: htseq-count on data 65 (no feature)

This dataset has been hidden
Unhide it

78: htseq-count on data 65

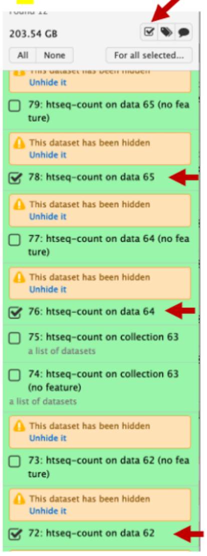
This dataset has been hidden
Unhide it

77: htseq-count on data 64 (no feature)

This dataset has been hidden
Unhide it

76: htseq-count on data 64

3



203.54 GB

Unhide it

□ 79: htseq-count on data 65 (no feature)

This dataset has been hidden
Unhide it

☑ 78: htseq-count on data 65

This dataset has been hidden
Unhide it

□ 77: htseq-count on data 64 (no feature)

This dataset has been hidden
Unhide it

☑ 76: htseq-count on data 64

□ 75: htseq-count on collection 63
a list of datasets

□ 74: htseq-count on collection 63 (no feature)
a list of datasets

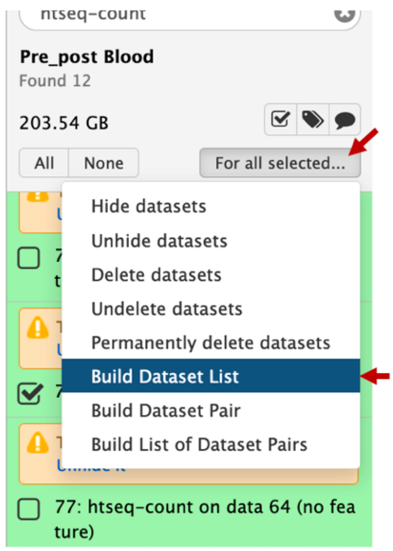
This dataset has been hidden
Unhide it

□ 73: htseq-count on data 62 (no feature)

This dataset has been hidden
Unhide it

☑ 72: htseq-count on data 62

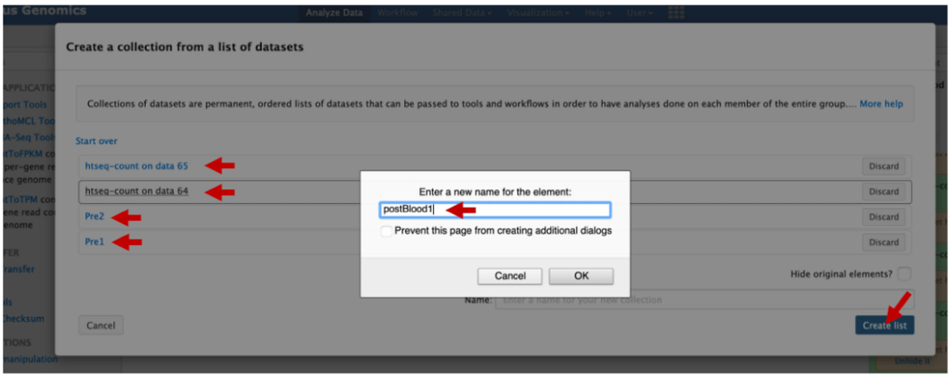
4



Pre_post Blood
Found 12
203.54 GB

□ 77: htseq-count on data 64 (no feature)

5



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.... More help

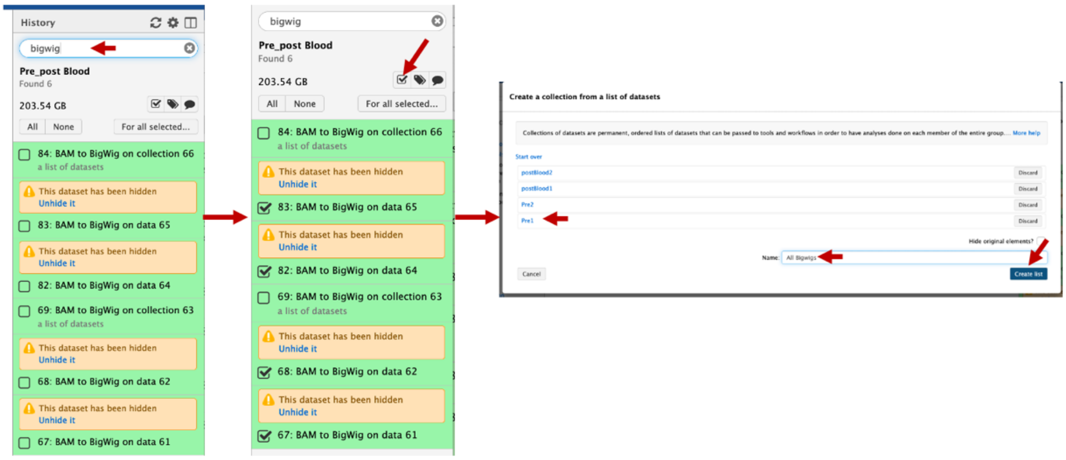
htseq-count on data 65
htseq-count on data 64
Pre2
Pre1

Enter a new name for the element:
postBlood1
□ Prevent this page from creating additional dialogs

Hide original elements? □

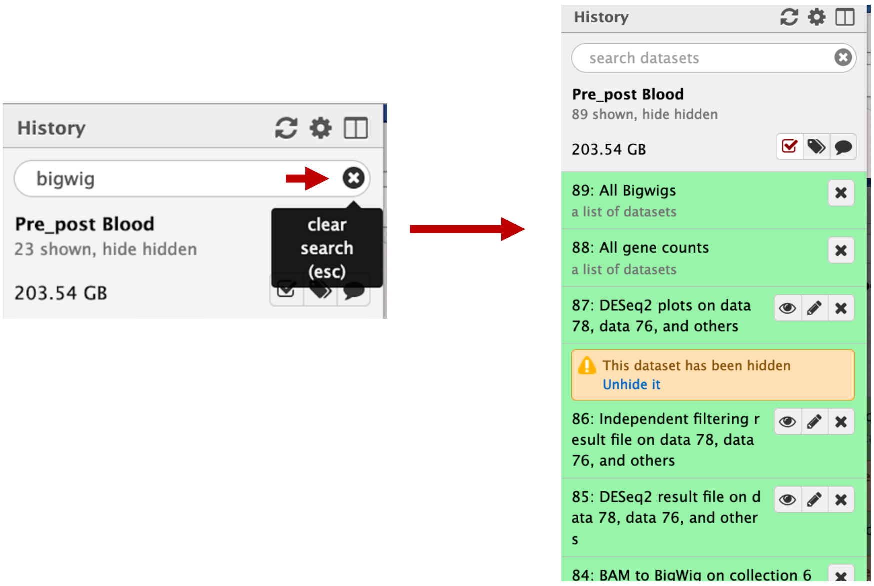
Name: Enter a name for your new collection

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


History
bigwig
Pre_post Blood
Found 6
203.54 GB
All
None
For all selected...
84: BAM to BigWig on collection 66
a list of datasets
This dataset has been hidden
Unhide it
83: BAM to BigWig on data 65
82: BAM to BigWig on data 64
69: BAM to BigWig on collection 63
68: BAM to BigWig on data 62
67: BAM to BigWig on data 61
Create a collection from a list of datasets
Start over
postBlood2
postBlood1
Pre2
Pre1
Discard
Hide original elements?
Name:
All Bigwigs
Cancel
Create list

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.... More help

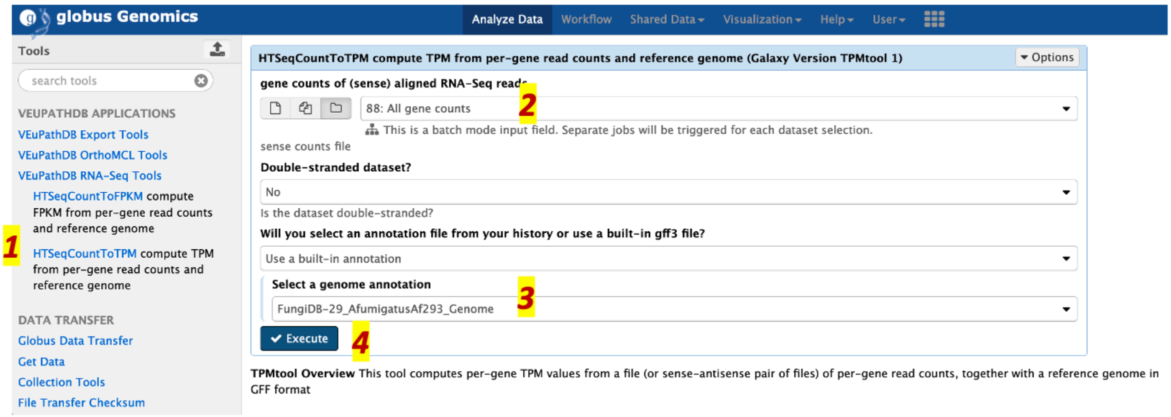
7. Click on clear search to see all results in your history.


History
bigwig
clear search (esc)
Pre_post Blood
23 shown, hide hidden
203.54 GB
search datasets
89 shown, hide hidden
89: All Bigwigs
a list of datasets
88: All gene counts
a list of datasets
87: DESeq2 plots on data 78, data 76, and others
This dataset has been hidden
Unhide it
86: Independent filtering result file on data 78, data 76, and others
85: DESeq2 result file on data 78, data 76, and others
84: BAM to BigWig on collection 6

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.



VEUPATHDB APPLICATIONS
VEuPathDB Export Tools
VEuPathDB OrthoMCL Tools
VEuPathDB RNA-Seq Tools
HTSeqCountToFPKM compute FPKM from per-gene read counts and reference genome
HTSeqCountToTPM compute TPM from per-gene read counts and reference genome

DATA TRANSFER
Globus Data Transfer
Get Data
Collection Tools
File Transfer Checksum

HTSeqCountToTPM compute TPM from per-gene read counts and reference genome (Galaxy Version TPMtool 1) ▾ Options

gene counts of (sense) aligned RNA-Seq reads
88: All gene counts
This is a batch mode input field. Separate jobs will be triggered for each dataset selection.
sense counts file

Double-stranded dataset?
No
Is the dataset double-stranded?

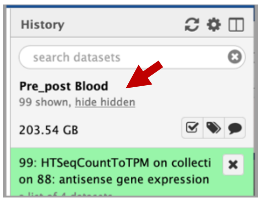
Will you select an annotation file from your history or use a built-in gff3 file?
Use a built-in annotation

Select a genome annotation
FungiDB-29_AfumigatusAf293_Genome

✔ Execute

TPMtool Overview This tool computes per-gene TPM values from a file (or sense-antisense pair of files) of per-gene read counts, together with a reference genome in GFF format

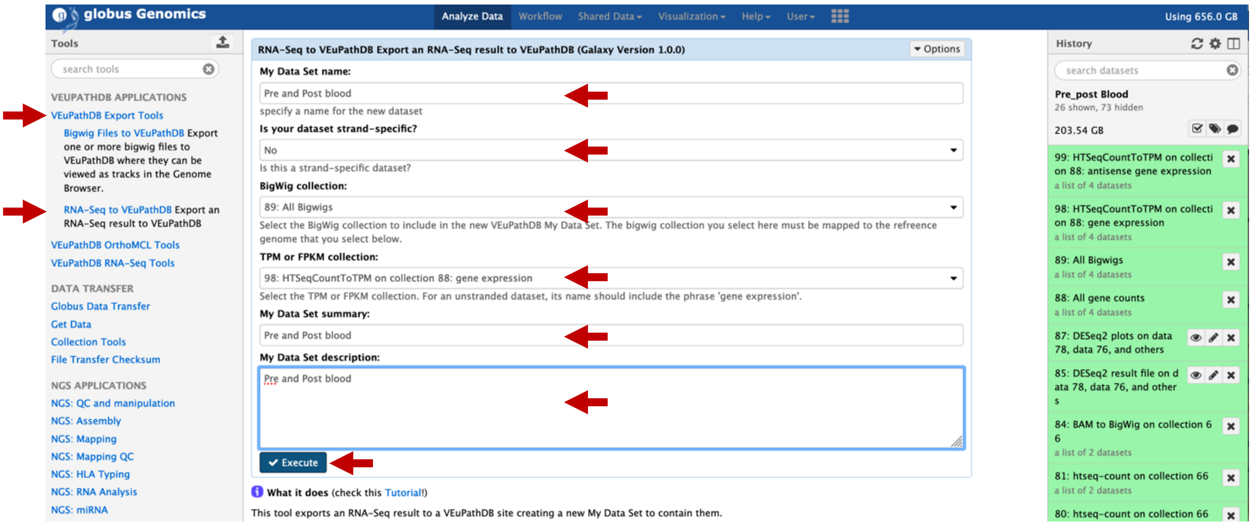
Optional: Click on "hide hidden" to clean up your history a bit.



Pre_post Blood
99 shown, hide hidden
203.54 GB
99: HTSeqCountToTPM on collection on 88: antisense gene expression

Export data to VEuPathDB. To export the TPM and BigWig files follw 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).



VEUPATHDB APPLICATIONS
VEuPathDB Export Tools
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
File Transfer Checksum

NGS APPLICATIONS
NGS: QC and manipulation
NGS: Assembly
NGS: Mapping
NGS: Mapping QC
NGS: HLA Typing
NGS: RNA Analysis
NGS: miRNA

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

My Data Set name:
Pre and Post blood
specify a name for the new dataset

Is your dataset strand-specific?
No
Is this a strand-specific dataset?

BigWig collection:
89: All Bigwigs
Select the BigWig collection to include in the new VEuPathDB My Data Set. The bigwig collection you select here must be mapped to the refreence genome that you select below.

TPM or FPKM collection:
98: HTSeqCountToTPM on collection 88: gene expression
Select the TPM or FPKM collection. For an unstranded dataset, its name should include the phrase 'gene expression'.

My Data Set summary:
Pre and Post blood

My Data Set description:
Pre and Post blood

✔ Execute

What it does (check this Tutorial!)

This tool exports an RNA-Seq result to a VEuPathDB site creating a new My Data Set to contain them.

Pre_post Blood
26 shown, 73 hidden
203.54 GB

99: HTSeqCountToTPM on collection on 88: antisense gene expression
a list of 4 datasets

98: HTSeqCountToTPM on collection on 88: gene expression
a list of 4 datasets

89: All Bigwigs
a list of 4 datasets

88: All gene counts
a list of 4 datasets

87: DESeq2 plots on data 78, data 76, and others

85: DESeq2 result file on data 78, data 76, and others

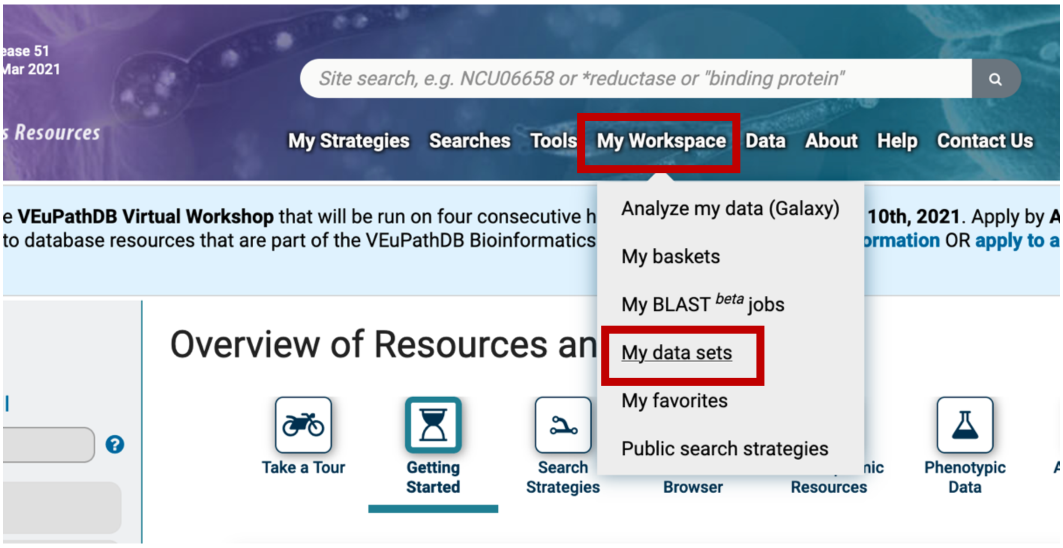
84: BAM to BigWig on collection 66
a list of 2 datasets

81: htseq-count on collection 66
a list of 2 datasets

80: htseq-count on collection 66

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

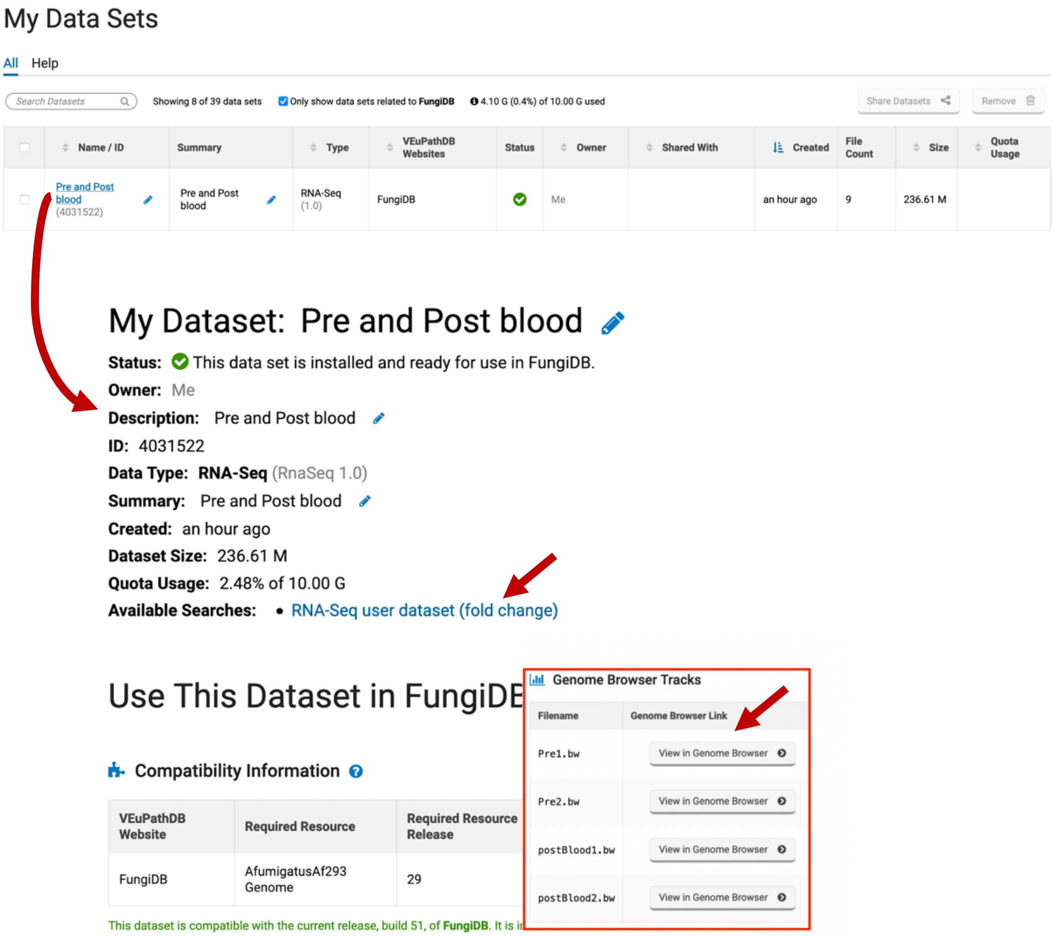
1. Click on the “My Workspace” link in the grey menu bar. Then select “My datasets” from the list.



e **VEuPathDB Virtual Workshop** that will be run on four consecutive h **10th, 2021**. Apply by **A**
to database resources that are part of the VEuPathDB Bioinformatics **ormation** OR **apply to a**

Overview of Resources an

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



My Data Sets

All Help

Search Datasets | Showing 8 of 39 data sets | ☑ Only show data sets related to **FungiDB** | 4.10 G (0.4%) of 10.00 G used | Share Datasets | Remove

□	Name / ID	Summary	Type	VEuPathDB Websites	Status	Owner	Shared With	Created	File Count	Size	Quota Usage
□	Pre and Post blood (4031522)	Pre and Post blood	RNA-Seq (1.0)	FungiDB	✓	Me		an hour ago	9	236.61 M	

My Dataset: Pre and Post blood

Status: ✓ This data set is installed and ready for use in FungiDB.
Owner: Me
Description: Pre and Post blood
ID: 4031522
Data Type: **RNA-Seq** (RnaSeq 1.0)
Summary: Pre and Post blood
Created: an hour ago
Dataset Size: 236.61 M
Quota Usage: 2.48% of 10.00 G
Available Searches: • RNA-Seq user dataset (fold change)

Use This Dataset in FungiDE

Compatibility Information

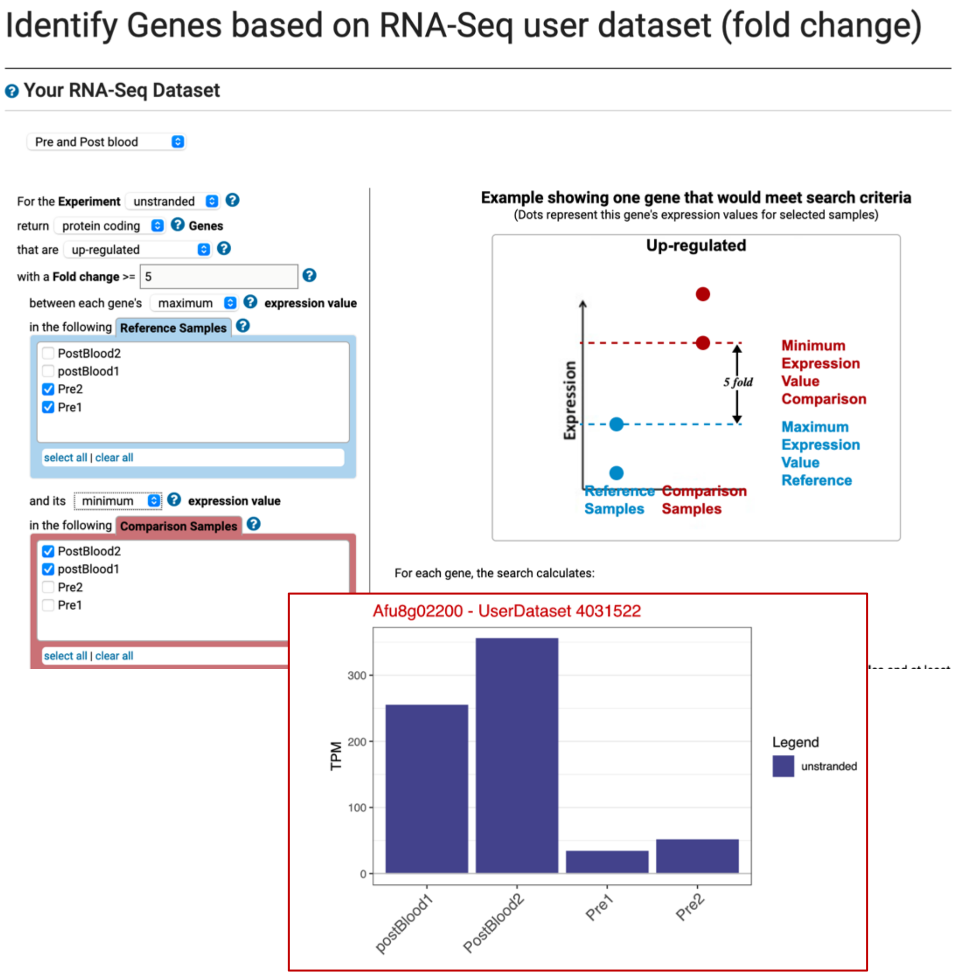
VEuPathDB Website	Required Resource	Required Resource Release
FungiDB	AfumigatusAf293 Genome	29

This dataset is compatible with the current release, build 51, of **FungiDB**. It is i

Genome Browser Tracks

Filename	Genome Browser Link
Pre1.bw	View in Genome Browser
Pre2.bw	View in Genome Browser
postBlood1.bw	View in Genome Browser
postBlood2.bw	View in Genome Browser

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!



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

Your RNA-Seq Dataset

Pre and Post blood

For the **Experiment** unstranded

return protein coding **Genes**

that are up-regulated

with a **Fold change** >= 5

between each gene's maximum **expression value**

in the following **Reference Samples**

- □ PostBlood2
- □ postBlood1
- ☑ Pre2
- ☑ Pre1

select all | clear all

and its minimum **expression value**

in the following **Comparison Samples**

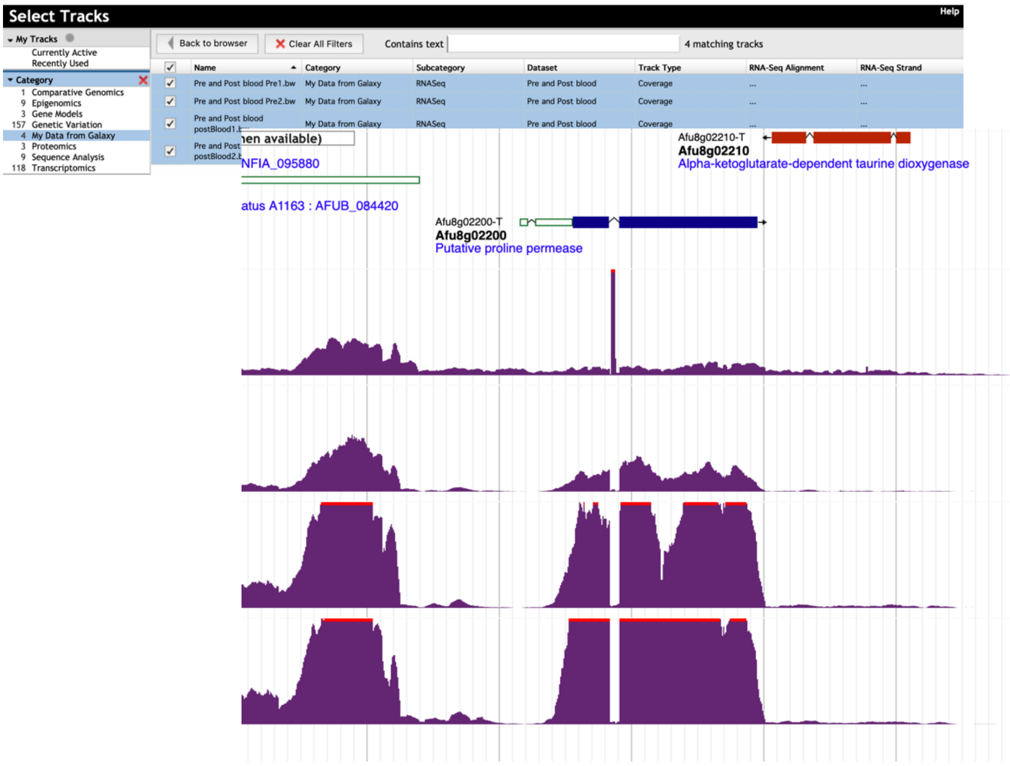
- ☑ PostBlood2
- ☑ postBlood1
- □ Pre2
- □ Pre1

select all | clear all

For each gene, the search calculates:

Example showing one gene that would meet search criteria
(Dots represent this gene's expression values for selected samples)
Up-regulated
Expression
5 fold
Minimum Expression Value Comparison
Maximum Expression Value Reference
Reference Samples
Comparison Samples
Afu8g02200 - UserDataset 4031522
TPM
300
200
100
0
postBlood1
PostBlood2
Pre1
Pre2
Legend
unstranded

4. Explore the coverage plots in the genome browser.



Select Tracks

☑	Name	Category	Subcategory	Dataset	Track Type	RNA-Seq Alignment	RNA-Seq Strand
☑	Pre and Post blood Pre1.bw	My Data from Galaxy	RNASeq	Pre and Post blood	Coverage	...	...
☑	Pre and Post blood Pre2.bw	My Data from Galaxy	RNASeq	Pre and Post blood	Coverage	...	...
☑	Pre and Post blood postBlood1.b...	My Data from Galaxy	RNASeq	Pre and Post blood	Coverage	...	...
☑	Pre and Post postBlood2.b						