

# 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](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](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)

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.

**Welcome to the VEuPathDB 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

History

search datasets

(unstranded) 21:GTM vs 42:Peru delta  
26 shown 121 hidden  
48.64 GB

147: DESeq2 plots on data 137, data 135, and others

145: DESeq2 result file on data 137, data 135, and others

144: BAM to BigWig on collection 120  
a list of 3 datasets

140: htseq-count on collection 120  
a list of 3 datasets

139: htseq-count on collection 120 (no feature)  
a list of 3 datasets

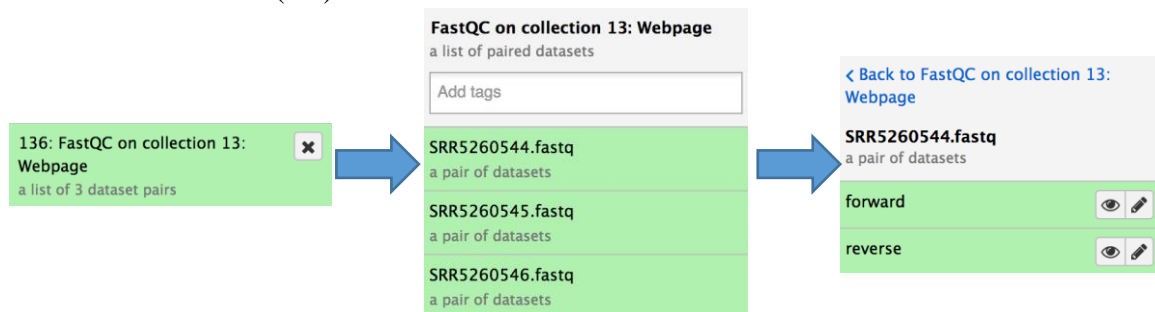
132: htseq-count on collection 116  
a list of 3 datasets

131: htseq-count on collection 116 (no feature)  
a list of 3 datasets

124: BAM to BigWig on collection 116  
a list of 3 datasets

120: HISAT2 on collection 11

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

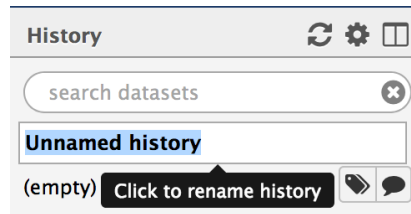
- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)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

### Basic Statistics

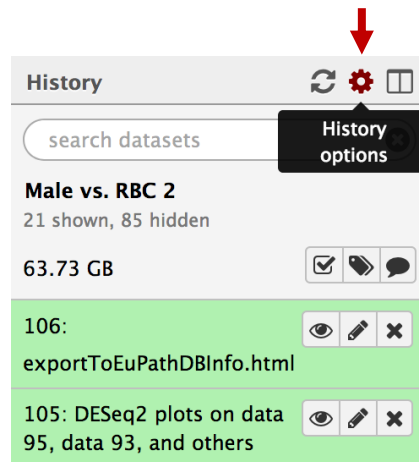
Measure	Value
Filename	SRR8128646_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	21931872
Sequences flagged as poor quality	0
Sequence length	24-126
%GC	49

## Step 2: Sharing histories with others:

- 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, showing options like 'Data Libraries', 'Histories', 'Workflows', 'Visualizations', and 'Pages'. Below this, the 'Published Histories' section is visible, featuring a search bar and a table of shared histories. The table has columns for Name, Annotation, Owner, Community Rating, Community Tags, and Last Updated. A red circle highlights the 'Import history' link in the top right corner of the table.

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		trick-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.

We will explore two output files:

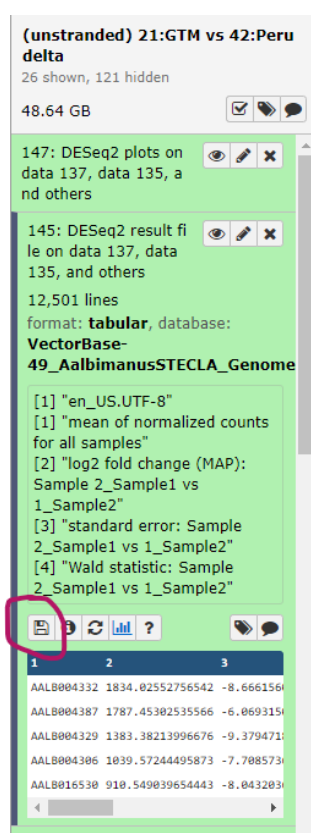
- A. DESeq2 **Plots** – you can view these directly in galaxy by clicking on the view icon. These plots give you an idea about the quality of the experiment. This program user guide (link in the 1<sup>st</sup> page of this tutorial), includes a detailed description of the graphs.
- B. DESeq2 **results 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 screenshot shows two output files in the Galaxy interface. The first file is '147: DESeq2 plots on data 137, data 135, and others' and the second is '145: DESeq2 result file on data 137, data 135, and others'. Both files have a green background and icons for viewing, editing, and deleting.

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.

	A	B	C	D	E	F	G	H	I
1	Gene IDs	Counts	Log2 FC	SE Log2 FC	Wald statistic	p-value	FDR		
2	AALB0043	1834.026	-8.66616	0.5762166	-15.03975373	4.03E-51	9.59E-48		
3	AAI								
4	AAI								
5	AAI								
6	AAI								
7	AAI								
8	AAI								
9	AAI								

- 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.

	A	B	C	D	E	F	G	H	I
1	Gene ID	Counts	Log2 FC	SE Log2	Wald statis	p-value	FDR		
3177	AALB0016	5.910172	6.481772	1.9507769	3.322661552	0.000892	0.008626		
3179	AAI								
3180	AAI								
3181	AAI								
3182	AAI								
3186	AAI								
3187	AAI								
3189	AAI								
3202	AAI								

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

Gene ID(s)

125 Genes

Step 1

+

Add a step

125 Genes (111 ortholog groups)

Revise this search

Gene Results

Genome View

Gene Ontology Enrichment

Analyze Results

Organism Filter

select all | clear all | expand all | collapse all

Hide zero counts

Search organisms...

Arthropoda

125

Mollusca

0

select all | clear all | expand all | collapse all

Hide zero counts

Hide Organism Filter

Gene Ontology Enrichment

Find Gene Ontology terms that are enriched in your gene result. [Read More](#)

Parameters

Organism

Anopheles albimanus STECLA

Ontology

Molecular Function

Cellular Component

Biological Process

Evidence

Computed

Curated

Limit to GO Slim terms

No

Yes

P-Value cutoff

0.05 (0 - 1)

Submit

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

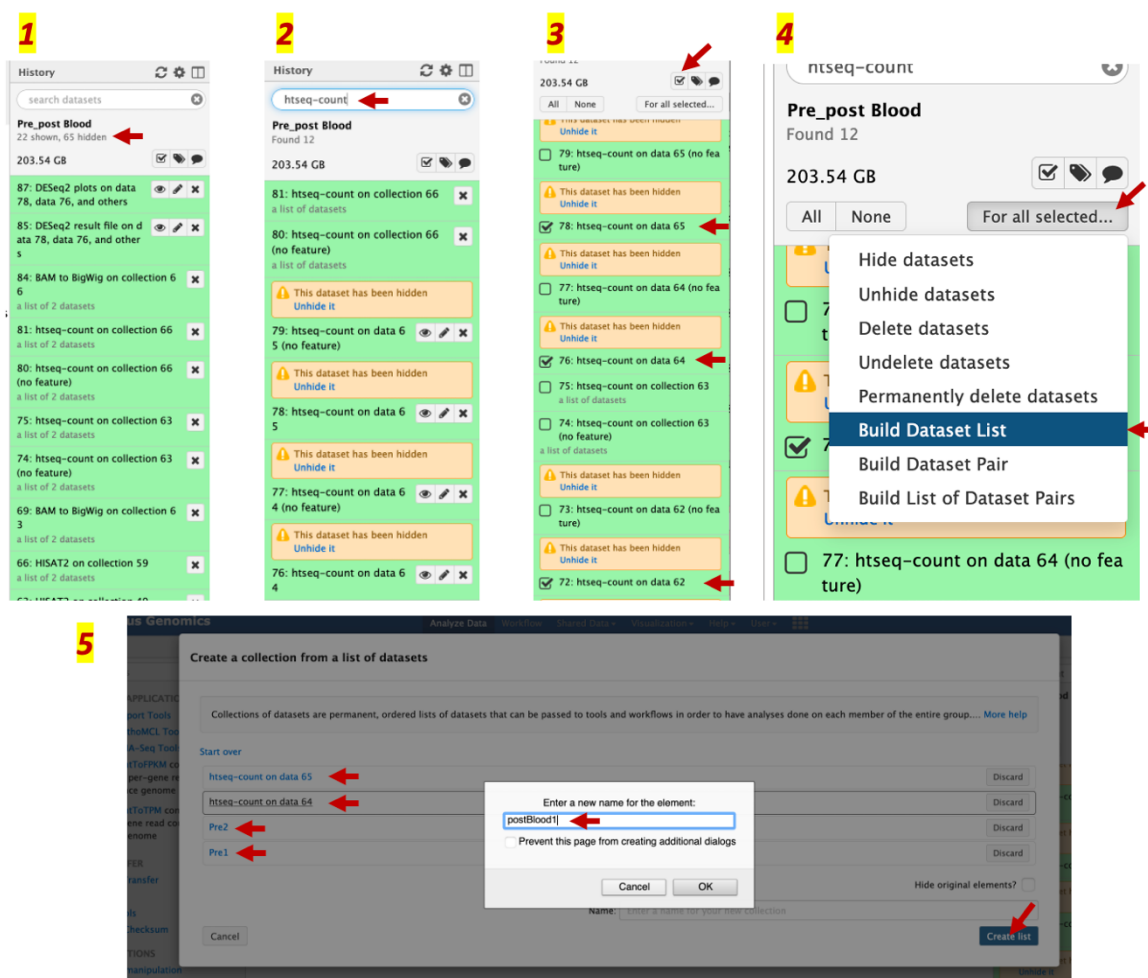
## Exporting data to VEuPathDB

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.

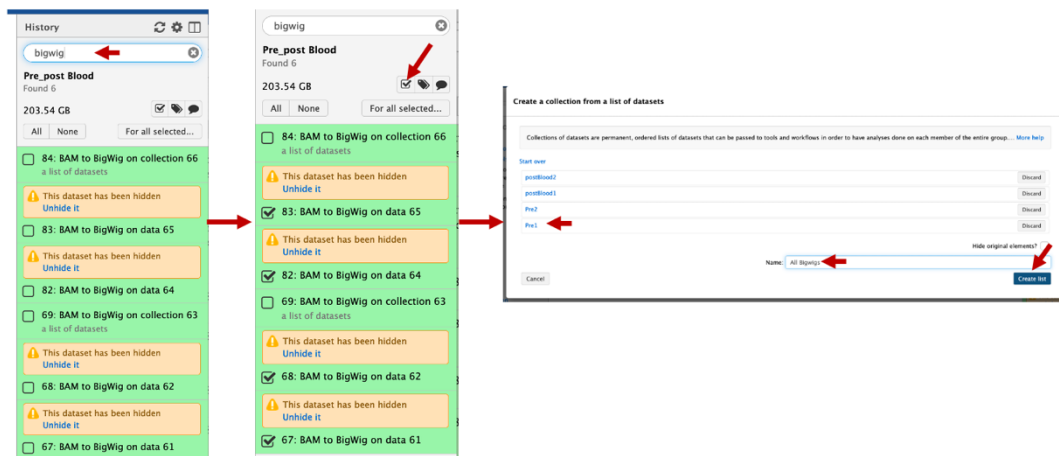
First let's organize the files (see matching screen shots below):

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.

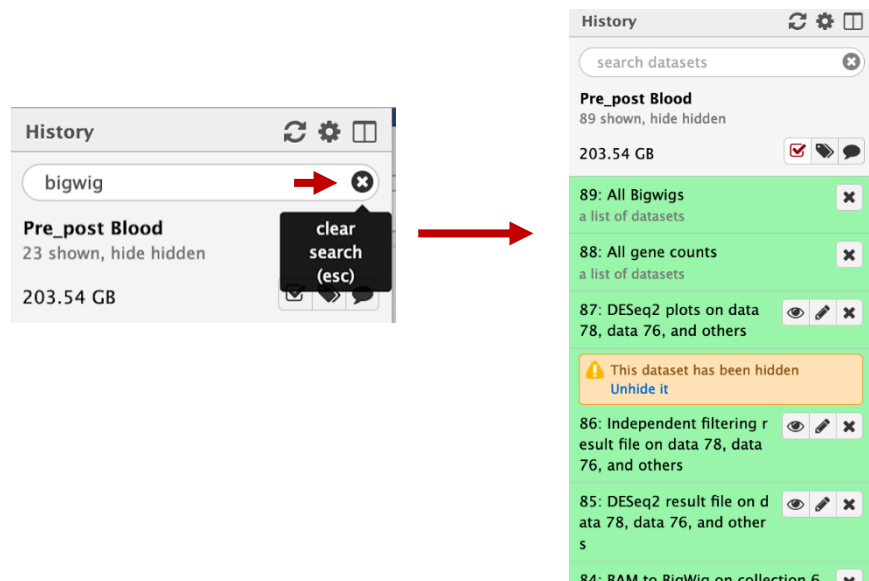




6. Repeat the same steps to create the list of BigWig files (See screen shots).



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



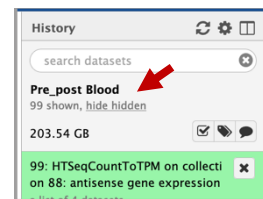
Now that your count and bigwig files are nice and organized, the next step is to convert the counts into TPMs. To do this follow these steps:

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”.

The screenshot shows the VectorBase website interface. The top navigation bar includes links for 'My Strategies', 'Searches', 'Tools', 'My Workspace', 'Data', 'About', 'Help', and 'Contact Us'. The 'My Workspace' link is highlighted with a red underline. A dropdown menu is open from 'My Workspace', showing options: 'Analyze my data (Galaxy)', 'My baskets', 'My BLAST <sup>beta</sup> jobs', 'My data sets' (highlighted with a red arrow), 'My favorites', and 'Public search strategies'. The 'My data sets' option is the target of the instruction.

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

The screenshot shows the 'My Data Sets' page in VectorBase. The page displays a table of datasets. The first dataset, 'Guatemala vs Peru DELTAMETHRIN An albimanus (4037421)', is highlighted with a red box and a red arrow. To the right, a detailed view of this dataset is shown. The 'Available Searches' section is circled in red, showing 'RNA-Seq user dataset (fold change)'. Below this, the 'Genome Browser Tracks' section is also circled in red, showing three tracks: 'Per\_delta3.bw', 'Per\_delta2.bw', and 'GTM\_delta1.bw', each with a 'View in Genome Browser' button.

Name / ID	Summary	Type
Guatemala vs Peru DELTAMETHRIN An albimanus (4037421)	Guatemala vs Peru: delta (An albimanus)	RNA-Seq (1.0)
eryth_vs_sporo (4036992)	eryth_vs_sporo	RNA-Seq (1.0)
Erythrocyte and cultured sporozoites (4036922)	eryth and cultures sporozoites	RNA-Seq (1.0)

**My Dataset: DELTAMETHRIN An albimanus**

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

Owner: Me

Description: Guatemala vs Peru: delta (An albimanus)

ID: 4037421

Data Type: RNA-Seq (RnaSeq 1.0)

Summary: Guatemala vs Peru: delta (An albimanus)

Created: 12 minutes ago

Dataset Size: 17.83 M

Quota Usage: 0.40% of 10.00 G

Available Searches: RNA-Seq user dataset (fold change)

**Genome Browser Tracks**

Filename	Genome Browser Link
Per_delta3.bw	<a href="#">View in Genome Browser</a>
Per_delta2.bw	<a href="#">View in Genome Browser</a>
GTM_delta1.bw	<a href="#">View in Genome Browser</a>

- 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**  $\geq 5$

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

☐ PostBlood2  
☐ postBlood1  
☒ Pre2  
☒ Pre1

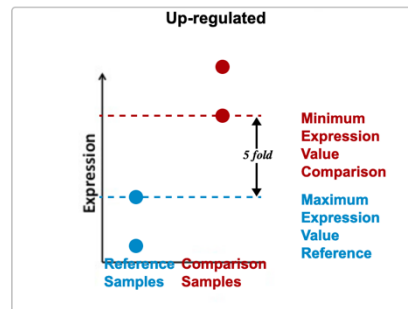
[select all](#) | [clear all](#)

and its **minimum** expression value  
 in the following **Comparison Samples**

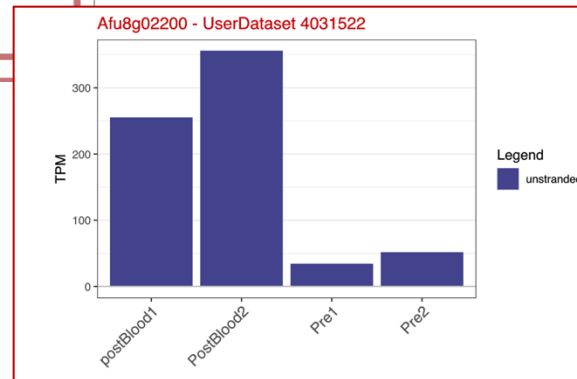
☒ PostBlood2  
☒ postBlood1  
☐ Pre2  
☐ Pre1

[select all](#) | [clear all](#)

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



For each gene, the search calculates:



- Explore the coverage plots in the genome browser.

