

RNA sequence data analysis via Galaxy, Part 2 (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) (https://workshop.eupathdb.org/athens/2019/exercises/fastqc_results-2.pdf)

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

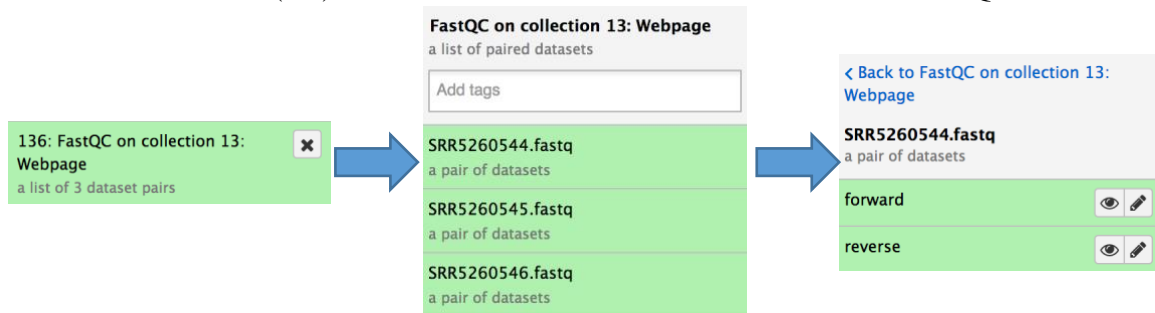
(http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf)

The screenshot shows the EuPathDB Galaxy Site interface. The main content area displays a welcome message and a list of datasets. Annotations with red arrows point to specific parts of the interface:

- Many more output files are available to explore**: Points to the "144 hidden" text in the "Male vs. RBC" dataset entry.
- Differential expression data on the two collections**: Points to the "201: DESeq2 result file on data 190, data 188, and others" entry.
- Read counts per gene or exon (depending on chosen parameters)**: Points to the "193: htseq-count on collection 173" entry.
- Coverage data in BigWig format**: Points to the "185: BAM to BigWig on collection 169" entry.


The interface includes a left sidebar with navigation links, a top navigation bar, and a right sidebar with a history of datasets.

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

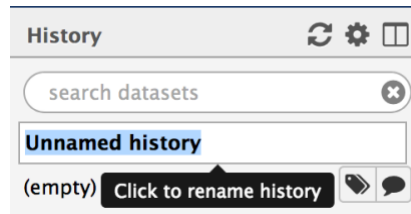
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented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

Basic Statistics

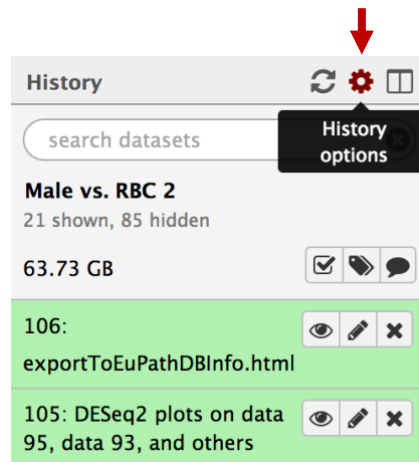
| Measure | Value |
|-----------|-------------------------|
| Filename | SRR5260544_1.fastq.gz |
| File type | Conventional base calls |

Step 2: Sharing histories with others:

- 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. The 'Import history' link is circled in red.

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

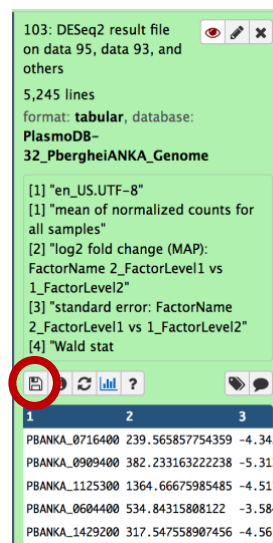
We will explore two output files:

- 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 that are uniquely up or down regulated in the conditions tested?

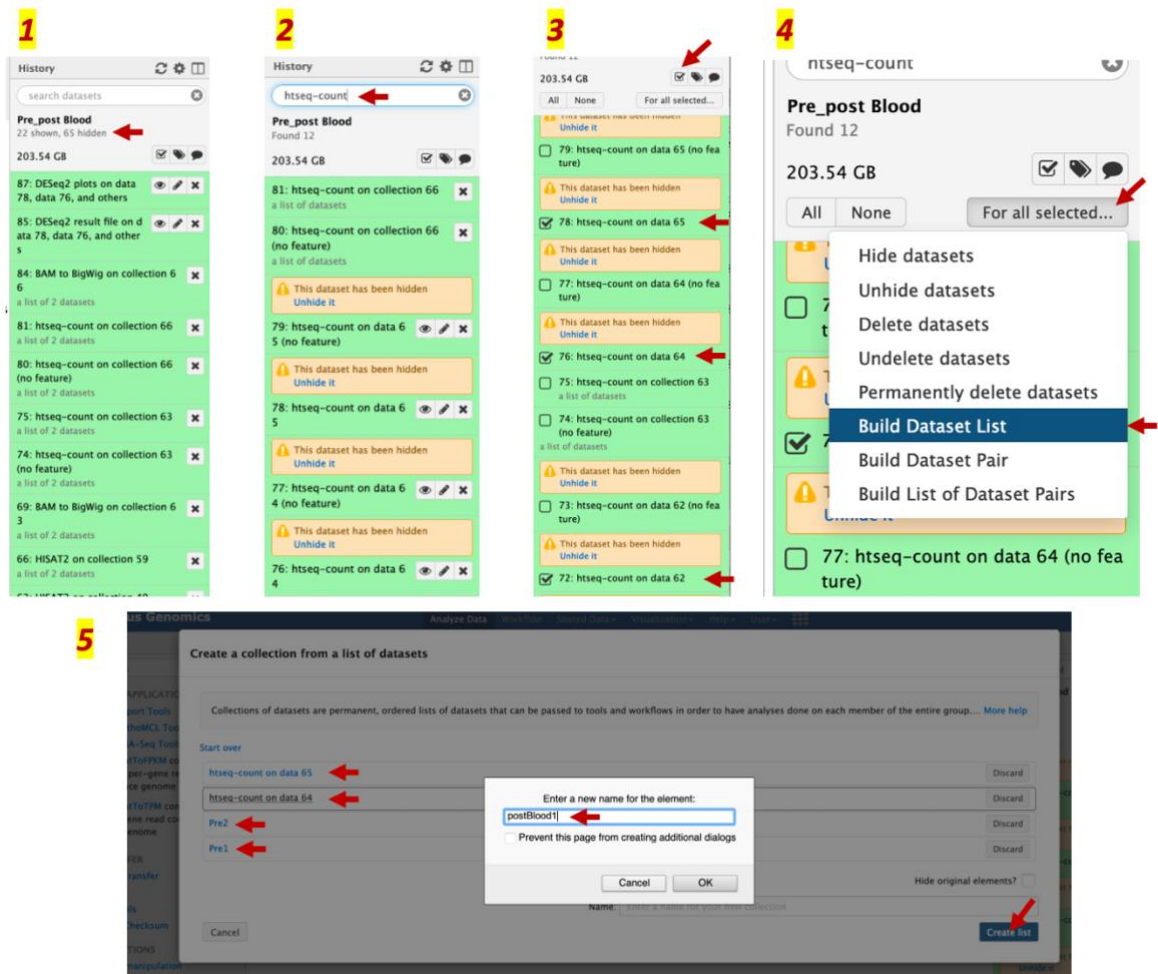
Exporting data to VEuPathDB

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.

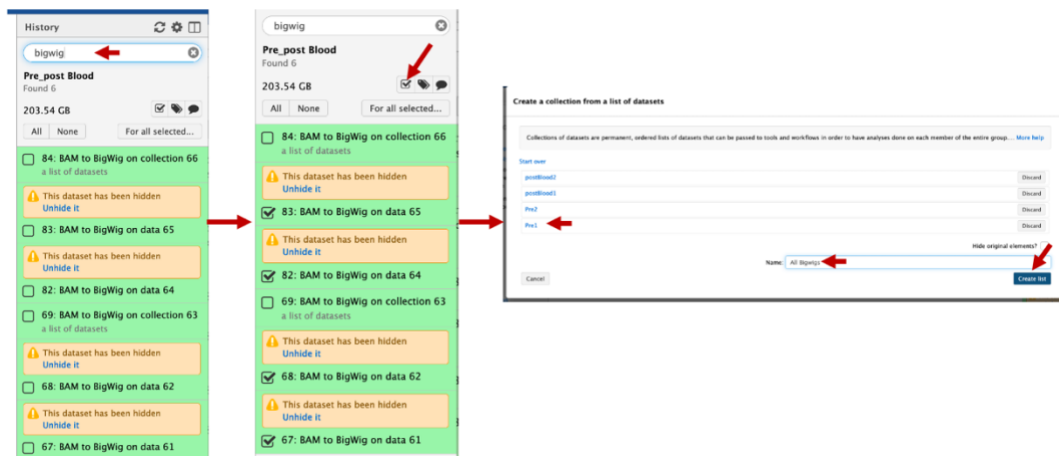
However, to use this feature you need to generate TPM values for genes in your datasets and organize your results into two collections, one for the TPMs and one for the BigWigs.

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

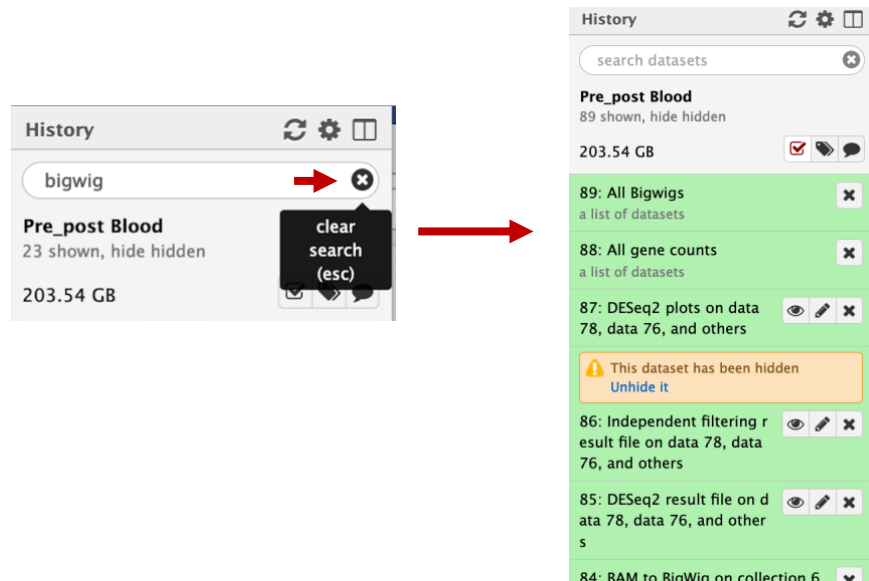
1. Click on the link at the top of your history that says “## hidden”. This will show all hidden files.
2. Use the search datasets box at the top of your history to find any file in your history with the work “htseq-count”.
3. Click on the “operation on multiple datasets” tool and select the individual 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.



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.

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

gene counts of (sense) aligned RNA-Seq read: 88: All gene counts

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?

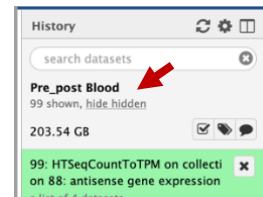
Use a built-in annotation

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.



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).

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

My Data Set name: Pre and Post blood

Is your dataset strand-specific? No

BigWig collection: 89: All BigWigs

TPM or FPKM collection: 98: HTSeqCountToTPM on collection 88: gene expression

My Data Set summary: Pre and Post blood

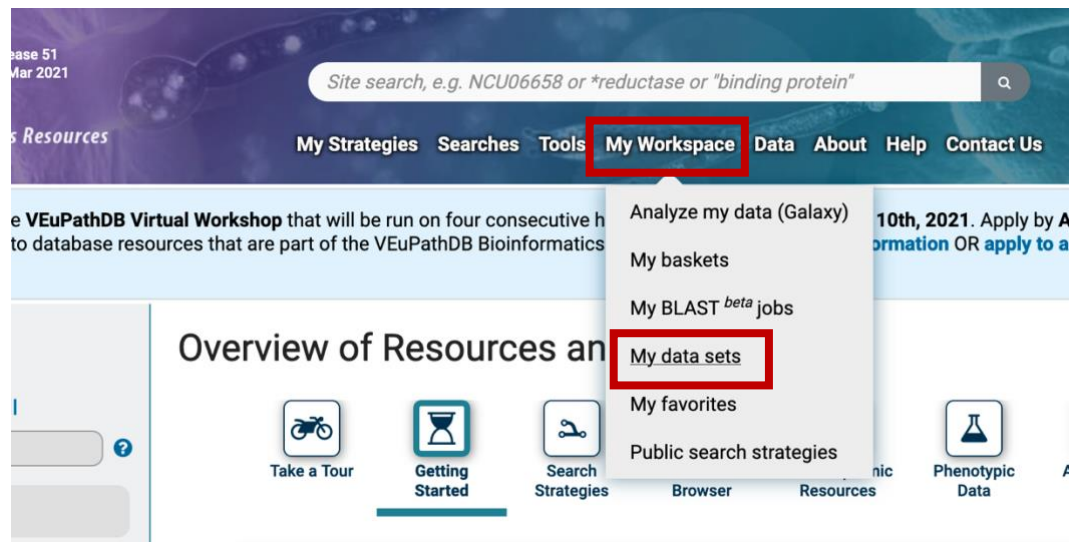
My Data Set description: Pre and Post blood

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.

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.



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](#)

| <input type="checkbox"/> | Name / ID | Summary | Type | VEuPathDB Website | Status | Owner | Shared With | Created | File Count | Size | Quota Usage |
|--------------------------|--|--------------------|---------------|-------------------|----------------|-------|-------------|-------------|------------|----------|-------------|
| <input type="checkbox"/> | 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 FungiDB

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 in

| 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 |

- 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](#)

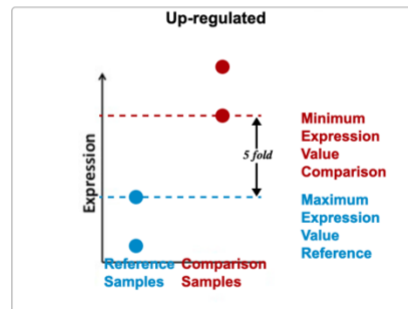
and its **minimum** expression value

in the following **Comparison Samples**

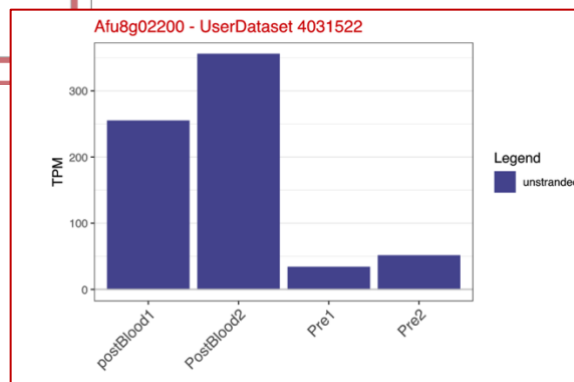
- ☒ PostBlood2
- ☒ postBlood1
- ☐ Pre2
- ☐ Pre1

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

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.

