

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

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.

The screenshot displays the EuPathDB Galaxy Site interface. The top navigation bar includes 'globus Genomics', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', 'User', and 'Using 582.7 GB'. The left sidebar lists various tools and applications, including 'EuPathDB APPLICATIONS', 'EuPathDB Export Tools', and 'NGS APPLICATIONS'. The main content area features a 'Welcome to the EuPathDB Galaxy Site' message and a list of workflow steps. A red circle highlights the word 'hidden' in the 'Male vs. RBC' dataset entry, indicating that 144 files are hidden. Red arrows point from the text annotations to the corresponding workflow steps in the list.

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

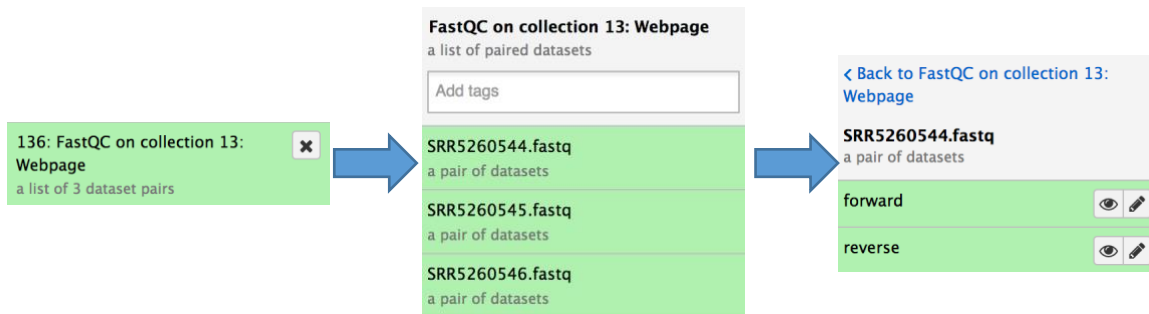
History


search datasets

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
a list of 3 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

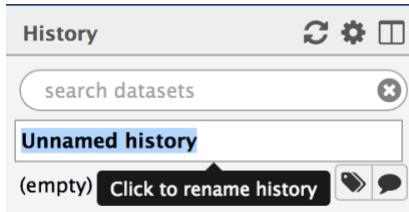
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented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

Basic Statistics

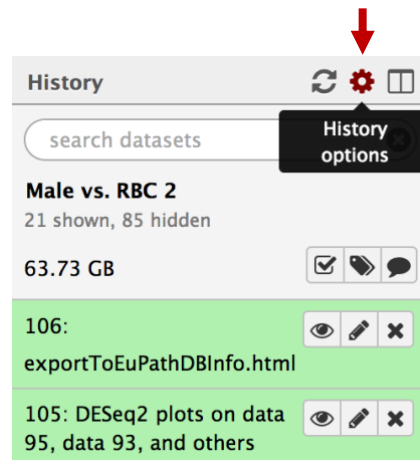
| Measure | Value |
|-----------|-------------------------|
| Filename | SRR5260544_1.fastq.gz |
| File type | Conventional base calls |

Step 2: Sharing histories with others:

- 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, 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

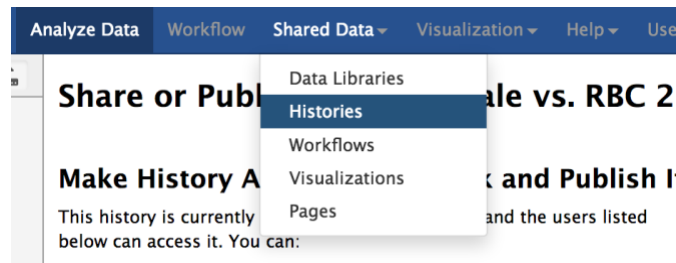
Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, where it is publicly listed and searchable.

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.

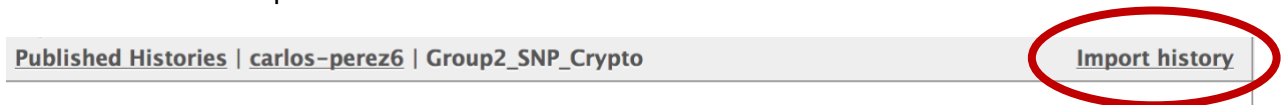
Published Histories

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: Group3_SNP | | f-puertolas-balint-301635433 | ★★★★★ | | May 17, 2018 |
| imported: Group4_SNP_Crypto | | cokane44-301496873 | ★★★★★ | | 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 |

- 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>

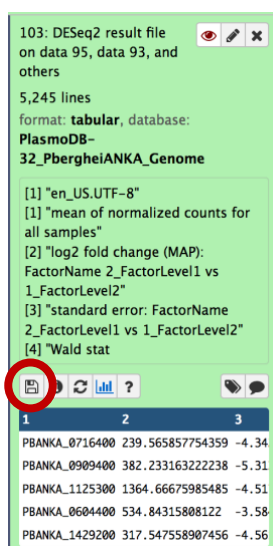
We will explore two output files:

- 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.
- 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 Expression files to EuPathDB

To take advantage of the EuPathDB RNAseq export tool which provides a mechanism to query your RNAseq results (FPKM values) using EuPathDB search tools follow these instructions:

To generate FPKM values you need a tool called Cufflinks and read alignment files – BAM files. Our workflow from yesterday generated BAM alignment files from a tool called HISAT2.

Follow these steps to generate FPKM values:

1. Find the tool called Cufflinks by typing the word cufflinks in the tool search box on the left-hand side.
2. Click on the tool to access its parameters.
3. Modify the cufflinks parameters
 - Change the input file to collection and select one of the HISAT2 collections
 - Change the Use Reference Annotation from “No” to “use reference annotation”
 - Select the appropriate reference genome from the drop down list
 - Click on execute.

Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data (Galaxy Version CUFFLINKS: 2.1.1) Versions Options

SAM or BAM file of aligned RNA-Seq reads

a

Dataset collection this is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Max Intron Length

Min Isoform Fraction

Pre MRNA Fraction

Perform quartile normalization

Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Use Reference Annotation **b**

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

Select a genome annotation

c

Perform Bias Correction

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Use multi-read correct

Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.

d

Tools

cufflinks

NCS: RNA Analysis

CUFFLINKS PACKAGE

Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data

Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments

Cuffmerge merge together several Cufflinks assemblies

Cuffdiff find significant changes in transcript expression, splicing, and promoter use

CUFFLINKS2 PACKAGE

Cuffquant Precompute gene expression levels

Cuffnorm Create normalized expression levels

StringTie transcript assembly and quantification

FILTERING

Filter Combined Transcripts using tracking file

Ballgown Flexible, isoform-level differential expression analysis

VISUALIZATION

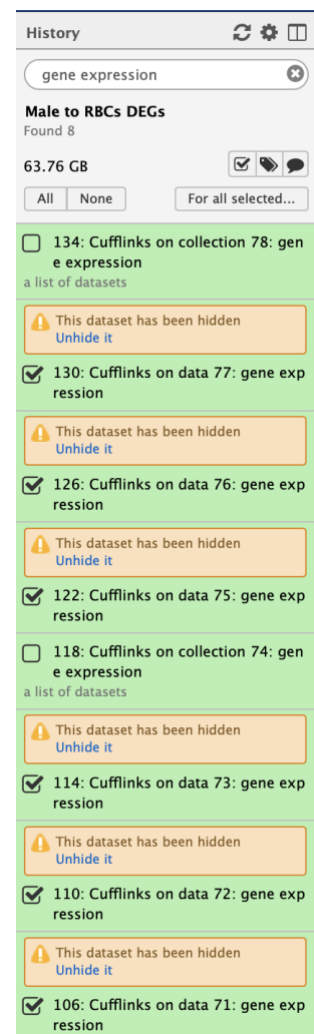
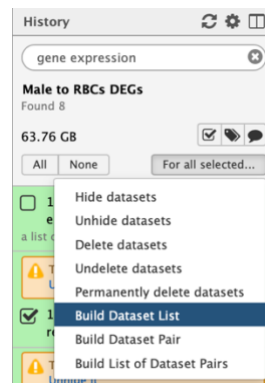
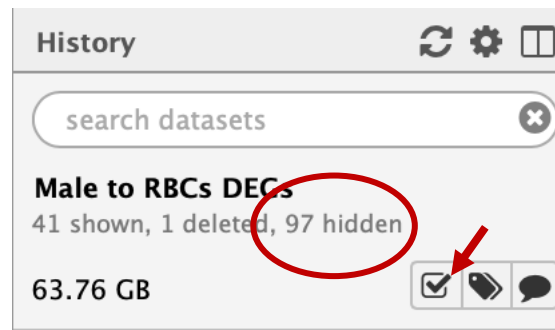
cummeRbund R package designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output

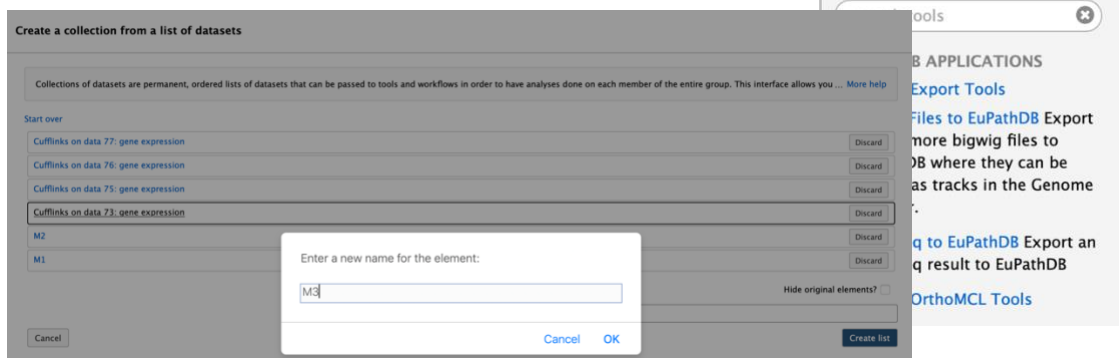
NCS: HOMER

findPeaks performs all of the peak calling and transcript identification analysis

After Cufflinks is done running, the next step is to take the FPKM output files from the collection outputs and put them into a single collection. Notice that cufflinks generates three types of FPKM files (or collections in this case): 1. Gene expression 2. Transcript expression 3. Assembled transcripts. We will only worry about the gene expression files for this section.

- Since we have collections of output files we will need to show all hidden files so we can generate the single collection. To do this, click on the word hidden in the upper right-hand side of the screen
- This will expose all hidden files.
- Click on the check box to perform an operation on multiple datasets (arrow in above image)
- Find all files containing the words “gene expression” and select all the cufflinks files (**NOT** the collections)
- Build a dataset list by clicking 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.





1. Click on “EuPathDB Export Tools” in the left-hand panel.
2. Click on the tool called “RNA-Seq to EuPathDB”
3. Fill up the export tool and select the correct files to export.

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

My Data Set name:

 specify a name for the new dataset

BigWig collection:

 Select the BigWig collection to include in the new EuPathDB My Data Set. The bigwig collection you select here must be mapped to the reference genome that you select below.

FPKM collection:

 Select the FPKM collection. Its name should include the phrase 'gene expression'.

My Data Set summary:

My Data Set description:

4. Click on Execute and wait for the export step to complete.

- When export is complete, go to the EuPathDB website with the genomes for this data, e.g PlasmoDB.
- 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.



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

Owner: Me

Description: testing manual cufflinks

ID: 4019810

Data Type: **RNASeq** (RnaSeq 1.0)

Summary: test25

Created: 2 days ago

Dataset Size: 108.12 M

Quota Usage: 1.13% of 10.00 G

Available Searches: • [genes by RNA-Seq user dataset \(fold change\)](#)

My Data Sets

Search Datasets Showing 4 of 17 data sets Only show data sets related to PlasmoDB 1.68 G (0.16% of 10.00 G us

| <input type="checkbox"/> | Name / ID | Summary | Type | EuPathDB Websites | Status | |
|--------------------------|---|-------------------------------------|--------------|-------------------|--------|----|
| <input type="checkbox"/> | Test25 (4019810) | test25 | RNASeq (1.0) | PlasmoDB | | Me |
| <input type="checkbox"/> | Differentiation 1 (4013803) | Differentiation 1 | RNASeq (1.0) | PlasmoDB | | Me |
| <input type="checkbox"/> | RBC vs Sporozoites (4010508) | RBC vs. Sporozoites | Bigwig (1.0) | PlasmoDB | | Me |
| <input type="checkbox"/> | berghel bigwig (4010222) | bigwig berghel | Bigwig (1.0) | PlasmoDB | | Me |

Use This Dataset in PlasmoDB

Compatibility Information

| EuPathDB Website | Required Resource | Required Resource Release | Installed Resource Release |
|------------------|---------------------|---------------------------|----------------------------|
| PlasmoDB | PberghelANKA Genome | 32 | 32 |

- 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?