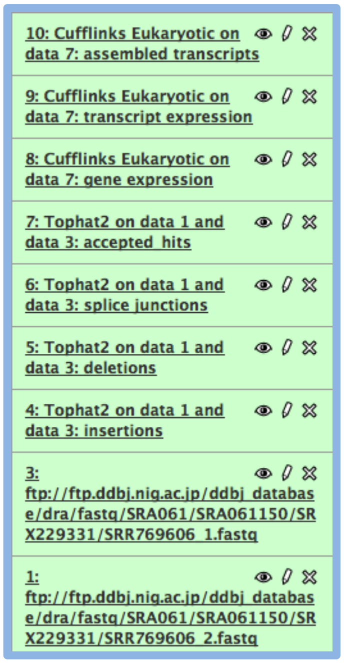
RNA sequence data analysis
(Part 2: Loading data generated by the pathogen portal's RNAseq pipeline in the Genome Browser)
Exercise 10

For this exercise we will be using:
http://pathogenportal.org
http://microsporidiadb.org



10: Cufflinks Eukaryotic on data 7: assembled transcripts 👁 ✎ ✕

9: Cufflinks Eukaryotic on data 7: transcript expression 👁 ✎ ✕

8: Cufflinks Eukaryotic on data 7: gene expression 👁 ✎ ✕

7: Tophat2 on data 1 and data 3: accepted_hits 👁 ✎ ✕

6: Tophat2 on data 1 and data 3: splice junctions 👁 ✎ ✕

5: Tophat2 on data 1 and data 3: deletions 👁 ✎ ✕

4: Tophat2 on data 1 and data 3: insertions 👁 ✎ ✕

3: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_1.fastq 👁 ✎ ✕

1: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_2.fastq 👁 ✎ ✕

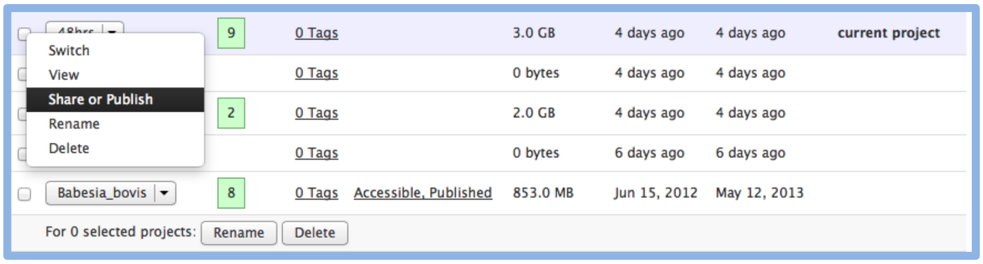
1. Explore the results of the RNA-sequence pipeline. What files were generated? To view contents of any of the results, click on the eye icon (👁) next to the file name.

!!! important note – do not click on the icon next to the file called "Tophat2 on data 1 and data 3: accepted_hits" – this file is huge and will not display but rather will download the contents to your computer.

TopHat generates four files:
insertions, deletions, splice junctions and accepted hits. The accepted hits file is the BAM file (binary alignment map). Note that many alignment programs will generate a file called a SAM file (sequence alignment map) which is a table including text of the alignment and mapping. However, for viewing results in a sequence browser like GBrowse, the file needs to be converted into the binary formatted (BAM) – you do not have to worry about this for this exercise.

Cufflinks generates three files:
gene expression, transcript expression and assembled transcripts. The gene expression and transcript expression files for our purposes should be identical since EuPathDB genomes do not have separate genes and transcripts. These files include the FPKM values for each gene in the genome analyzed – in this case *Encephalitozoon cuniculi* ECII.

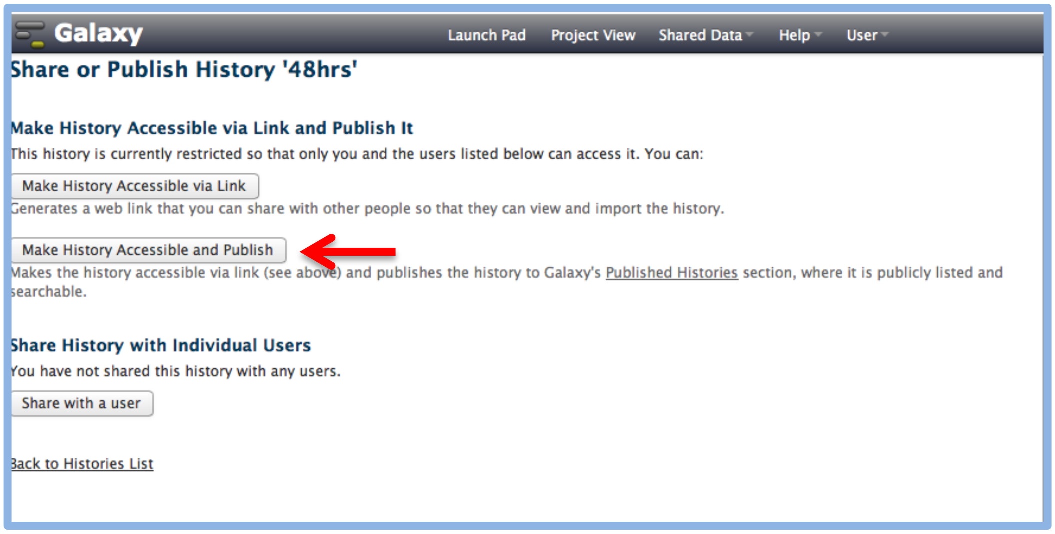
2. Share your accepted hits files. Click on the drop down menu for your project and select the option "share or publish".



□ 48hrs ▾	9	0 Tags		3.0 GB	4 days ago	4 days ago	current project
□		0 Tags		0 bytes	4 days ago	4 days ago	
□	2	0 Tags		2.0 GB	4 days ago	4 days ago	
□		0 Tags		0 bytes	6 days ago	6 days ago	
□ Babesia_bovis ▾	8	0 Tags	Accessible, Published	853.0 MB	Jun 15, 2012	May 12, 2013	

For 0 selected projects: Rename Delete

On the next page select the option: Make History Accessible and Publish



Share or Publish History '48hrs'

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.

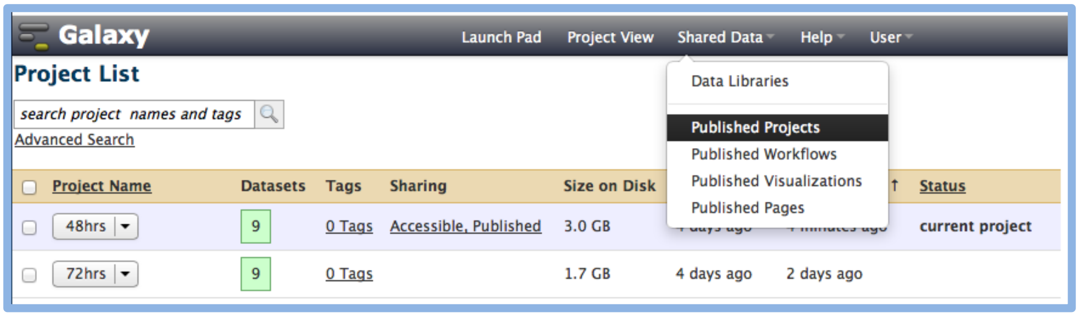
Share History with Individual Users

You have not shared this history with any users.

Share with a user

Back to Histories List

Once your project is published other people can access it by going to “Published Projects” section under the Shared data menu option in the Galaxy menu bar.

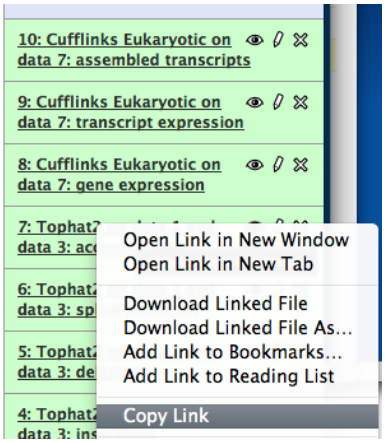


Project List

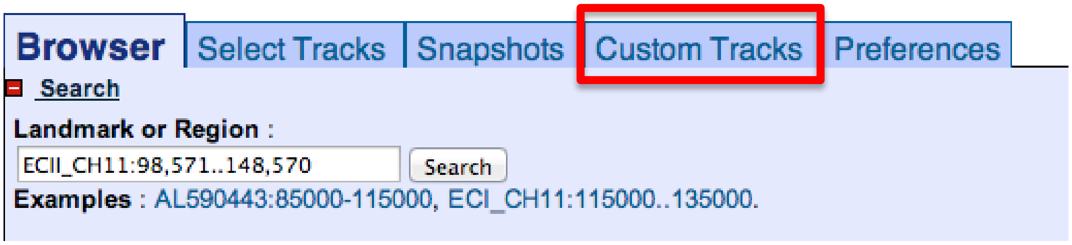
□	Project Name	Datasets	Tags	Sharing	Size on Disk		↑	Status
□	48hrs	9	0 Tags	Accessible, Published	3.0 GB	4 days ago	4 minutes ago	current project
□	72hrs	9	0 Tags		1.7 GB	4 days ago	2 days ago	

3. Load your BAM data into GBrowse. Navigate to the genome browser in MicrosporidiaDB and choose a landmark for *Encephalitozoon cuniculi* ECII you can just cut and paste the following into the “landmark or region” box: ECII_CH11:98,571..148,570

Next, do the following to copy the link to the tophat accepted hits in pathogenportal to GBrowse:



a. Control click (same as right click on a windows machine) on the eye icon for the tophat accepted hits.
b. In GBrowse click on the “Custom Tracks” tab.

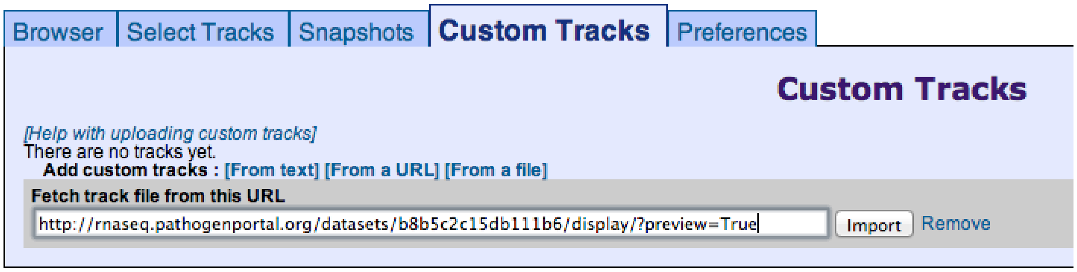


Landmark or Region :

ECII_CH11:98,571..148,570

Examples : AL590443:85000-115000, ECI_CH11:115000..135000.

c. Click on the "From a URL" link and paste the link you copied from pathogenportal.



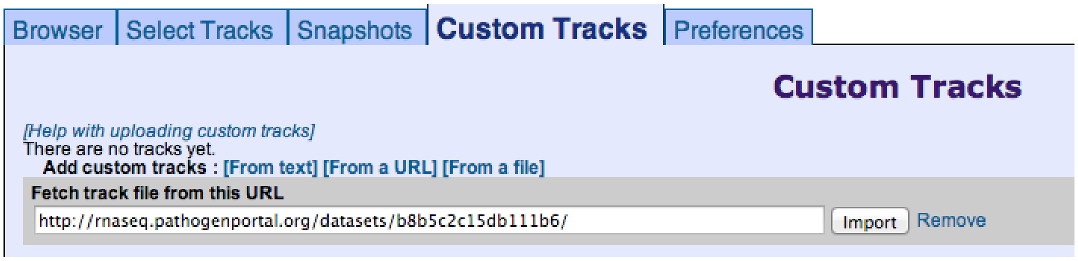
Custom Tracks

[Help with uploading custom tracks]
There are no tracks yet.
Add custom tracks : [From text] [From a URL] [From a file]

Fetch track file from this URL

http://rnaseq.pathogenportal.org/datasets/b8b5c2c15db111b6/display/?preview=True

d. Delete the last portion of the URL: display/?preview=True



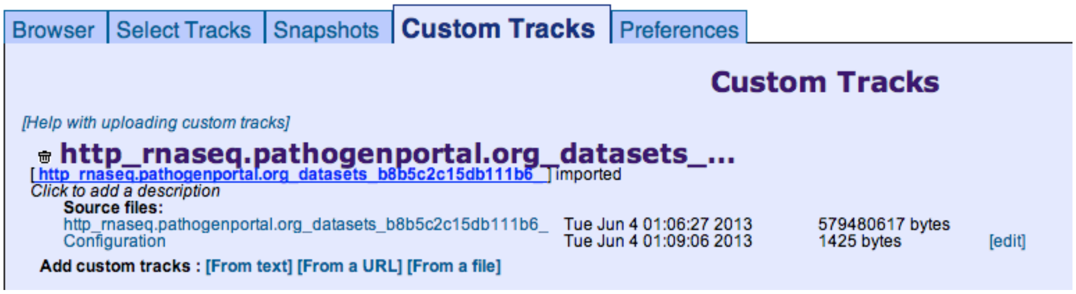
Custom Tracks

[Help with uploading custom tracks]
There are no tracks yet.
Add custom tracks : [From text] [From a URL] [From a file]

Fetch track file from this URL

http://rnaseq.pathogenportal.org/datasets/b8b5c2c15db111b6/

e. Click on import…..and be patient.



Custom Tracks

[Help with uploading custom tracks]

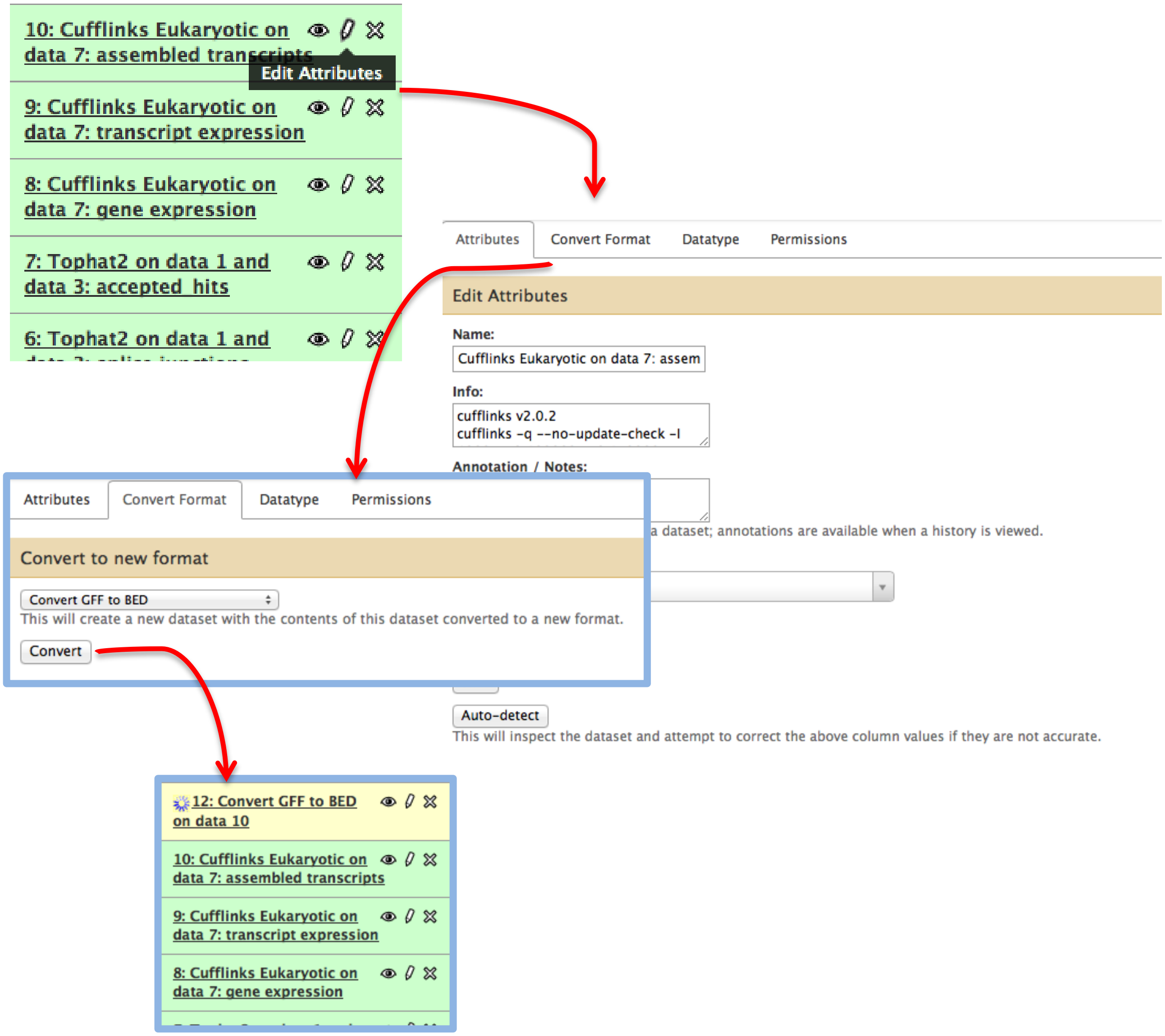
http_rnaseq.pathogenportal.org_datasets_...
[http_rnaseq.pathogenportal.org_datasets_b8b5c2c15db111b6_] imported
Click to add a description
Source files:

http_rnaseq.pathogenportal.org_datasets_b8b5c2c15db111b6_	Tue Jun 4 01:06:27 2013	579480617 bytes	
Configuration	Tue Jun 4 01:09:06 2013	1425 bytes	[edit]

Add custom tracks : [From text] [From a URL] [From a file]

f. Once the data has loaded click on the Browser tab to view your data.

4. Load the assembled transcript data. Cufflinks generates this file in a format called GFF. This format is not accepted by GBrowse so you have to convert it to another format called BED. To do this click on the pencil icon next to the file. Click on “Covert Format” then click on convert. A new file will be generated in BED format. You can not copy the link to the file and load it into GBrowse the same way you loaded the BAM file.



10: Cufflinks Eukaryotic on data 7: assembled transcripts

9: Cufflinks Eukaryotic on data 7: transcript expression

8: Cufflinks Eukaryotic on data 7: gene expression

7: Tophat2 on data 1 and data 3: accepted_hits

6: Tophat2 on data 1 and data 3: splice junctions

Name: Cufflinks Eukaryotic on data 7: assem

Info: cufflinks v2.0.2
cufflinks -q --no-update-check -I

Annotation / Notes:

a dataset; annotations are available when a history is viewed.

Convert to new format

Convert GFF to BED

This will create a new dataset with the contents of this dataset converted to a new format.

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

12: Convert GFF to BED on data 10

10: Cufflinks Eukaryotic on data 7: assembled transcripts

9: Cufflinks Eukaryotic on data 7: transcript expression

8: Cufflinks Eukaryotic on data 7: gene expression