


Interpreting RNAseq Mapping results (Part 2: Loading data from *RNA-Rocket* in the Genome Browser) Exercise 9

For this exercise we will be using:

<http://pathogenportal.org>

<http://plasmodb.org>





































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.

a. TopHat in RNA-Rocket generates five files:

- *Align_summary*: this includes a summary of how the alignment went (ie. the number of reads that were aligned).
- *Insertions*: reported insertions.
- *Deletions*: reported deletions.
- *Splice junctions*: reported junctions. Each junction consists of two connected BED blocks, where each block is as long as the maximal overhang of any read spanning the junction. The score is the number of alignments spanning the junction.
- *Accepted hits*: 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.

14: Tophat2 on data 2 and data 1: accepted hits (Genome Coverage BedGraph)			
13: Tophat2 on data 2 and data 1: accepted hits (- BigWig)			
12: Tophat2 on data 2 and data 1: accepted hits (+ BigWig)			
10: Cufflinks on data 7: assembled transcripts			
9: Cufflinks on data 7: transcript expression			
8: Cufflinks on data 7: gene expression			
7: Tophat2 on data 2 and data 1: accepted hits			
6: Tophat2 on data 2 and data 1: splice junctions			
5: Tophat2 on data 2 and data 1: deletions			
4: Tophat2 on data 2 and data 1: insertions			
3: Tophat2 on data 2 and data 1: align_summary			
2: EBI SRA: SRX129648 File: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR445/SRR445171/SRR445171.2.fast			

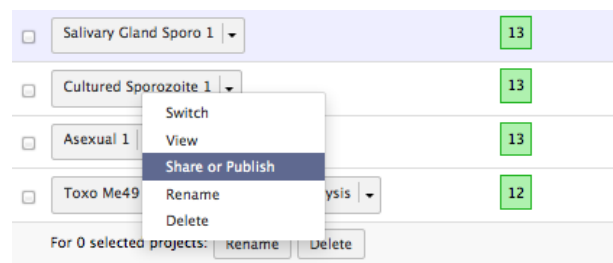
b. Cufflinks generates three files: *gene expression*, *transcript expression* and *assembled transcripts*. The gene expression and transcript expression files for our purposes should be identical since EuPathDB genomes do not have separate genes and transcripts. These files include the FPKM values (Fragments Per Kilobase of transcript per Million mapped reads) for each gene in the genome analyzed – in this case *Giardia* assemblages.

Additional files include files of the format BigWig and BedGraph. You can read more about these file formats here:

<http://genome.ucsc.edu/goldenPath/help/bigWig.html>

In a nutshell, these are file formats created from large binary files like BAM files and makes it possible to load these data in a genome browser.

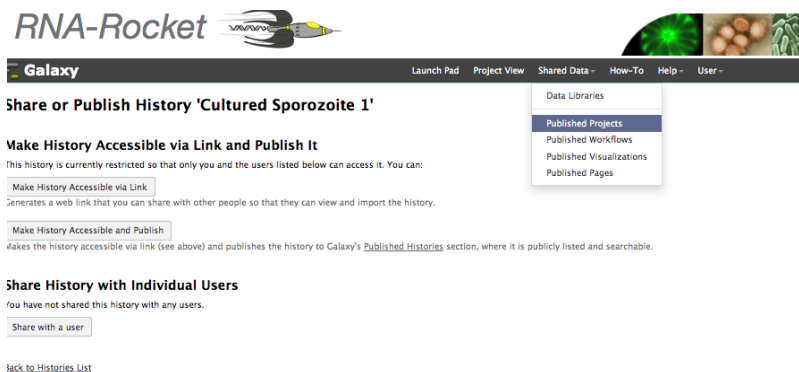
Note: to share your data with the rest of the workshop, select “*Share or Publish*” from the drop down menu on project you want to share. On the next page click



Project Results can be accessed via the “Project” panel to the right.

on “*Make History Accessible and Publish*”.

To import a history, select *Published Projects* from the Shared data menu item. Select the project you want to import then click on import history in the upper right hand side of the screen.



2. **Load your BAM data (accepted hits) into GBrowse.** Click on your “Tophat2 on data 2 and data 1: accepted_hits” in your project history panel. This will show you information about the file including a link to display data in PlasmoDB – click on the link.
3. **Load the assembled transcript data.** Essentially use a similar procedure as above. Wait a couple of minutes for GBrowse to load your data.

14: [Tophat2 on data 1 and data 2: accepted_hits \(Genome Coverage BedGraph\)](#)

13: [Tophat2 on data 1 and data 2: accepted_hits \(- BigWig\)](#)

12: [Tophat2 on data 1 and data 2: accepted_hits \(+ BigWig\)](#)

10: [Cufflinks on data 7: assembled transcripts](#)

9: [Cufflinks on data 7: transcript expression](#)

8: [Cufflinks on data 7: gene expression](#)

7: [Tophat2 on data 1 and data 2: accepted_hits](#)
1.5 GB
format: bam, database: pfal3D7
Log: tool progress Log: tool progress
[2014-06-14 08:55:03] Beginning TopHat run (v2.0.10)

----- [2014-06-14 08:55:03] Checking for Bowtie Bowtie version: 2.1.0.0 [2014-06-14 08:55:03] Checking for Samtools

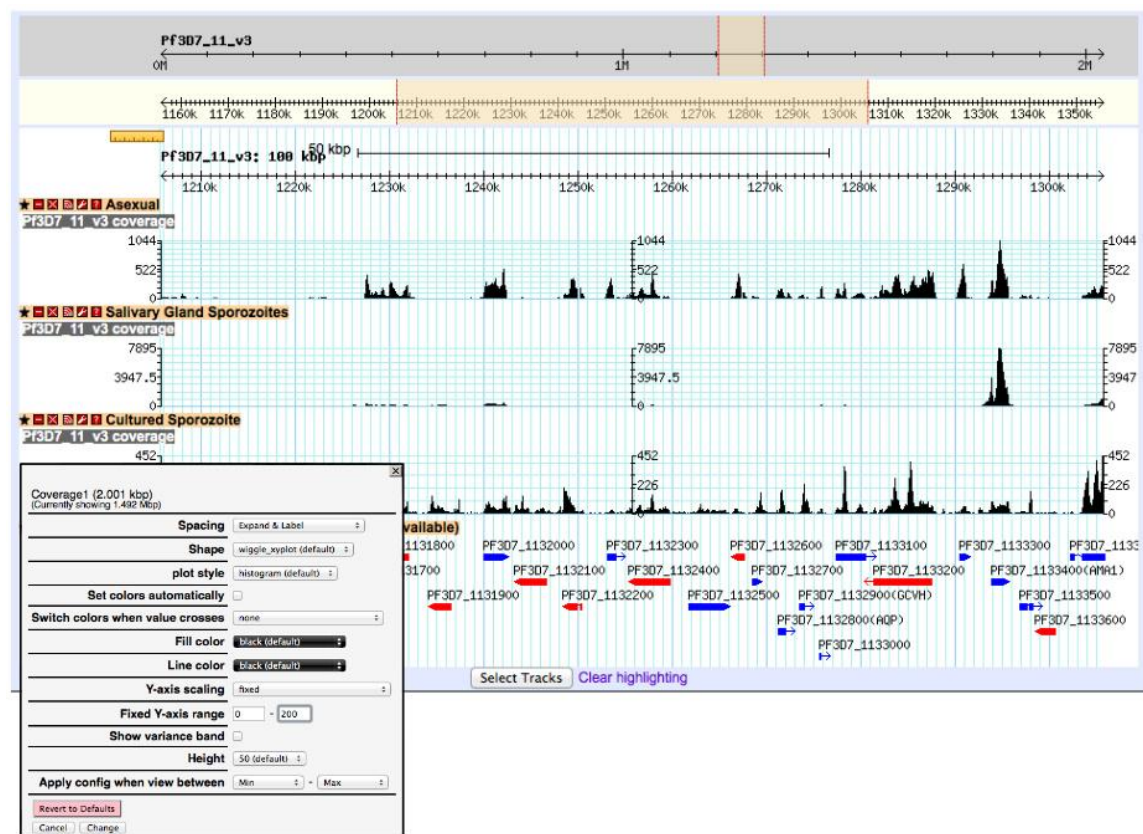
display at EupathDB [plasmodb](#)
Binary bam alignments file

6: [Tophat2 on data 1 and data 2: splice junctions](#)

5: [Tophat2 on data 1 and](#)



Once data has been loaded, you can configure the track display settings. For example, you can adjust the Y-axis scaling to a fixed axis.

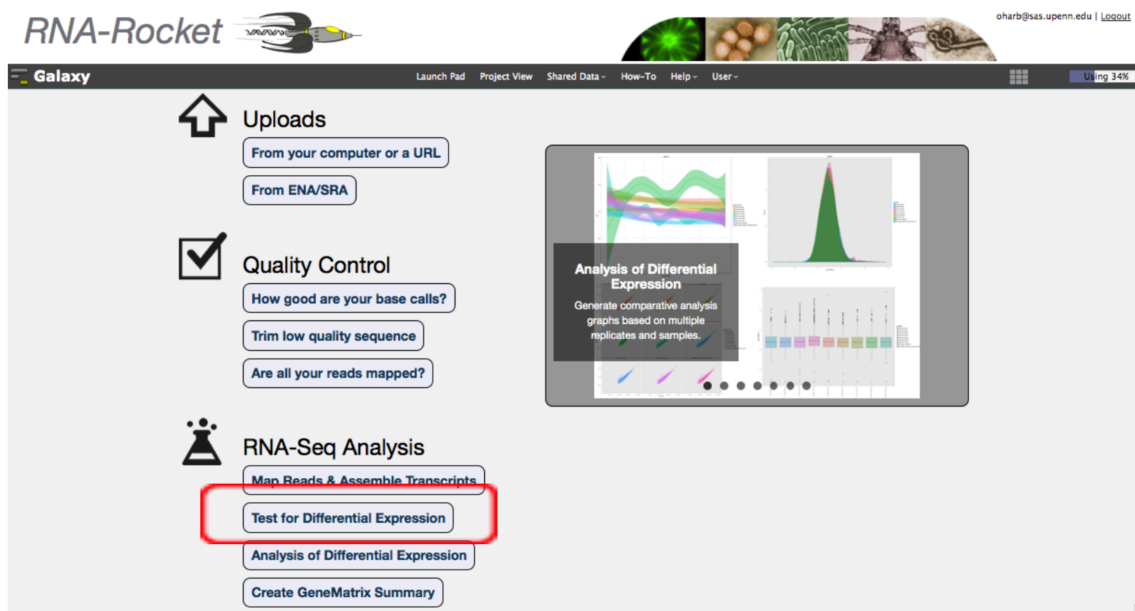


4. Find genes with significant differences in expression based on two of the samples you analyzed.

Cufflinks.cuffdiff finds significant changes in transcript expression, splicing, and promoter use. The Cufflinks.cuffdiff module takes a GTF file of transcripts as input, along with two or more SAM or BAM files containing the fragment alignments for two or more samples.

Cufflinks.cuffdiff produces a number of output files that contain test results for changes in expression at the level of transcripts, primary transcripts, and genes. It also tracks changes in the relative abundance of transcripts sharing a common transcription start site, and in the relative abundances of the primary transcripts of each gene. Tracking the former shows changes in splicing, and the latter shows changes in relative promoter use within a gene.

- a. Go back to the launch pad and select the “Test for Differential Expression” option.



- b. On the next page create a new project - you can call it whatever you want (diff. expression, for example). Type the name in the box then click on “Create Project”. Next, select the project you just created from the drop down menu called “Target Project: Select existing project”.
- c. The next step is to copy over two BAM files from your TopHat output. Select the BAM file from each of your two projects and

Differential Expression Analysis

Purpose:
Determine if assembled transcripts are significantly differentially expressed between RNA-Seq Samples.

Required Input:
BAM files from mapping reads to the Pathogen Portal provided genome

Output:
Tab delimited files tracking which genes are differentially expressed between sample pairs.

Select an existing Project or create a new Project to be used during this analysis and populate the Project with the necessary files. Output from this analysis will be saved in the selected Project.

Currently Selected Project: **None Selected**

Target Project:
Select existing project — OR — Create project

Copy-Source:
Select source

- ☐ Cufflinks (Eukaryotic on data 29: total map mass
- ☐ SRX1041266_1.fastq.gz
- ☐ SRX1041266_2.fastq.gz
- ☐ SRX1041267_1.fastq.gz
- ☐ SRX1041267_2.fastq.gz
- ☐ SRX1041268_1.fastq.gz
- ☐ SRX1041268_2.fastq.gz

Copy

click on “copy” to copy them into the “diff. expression” project.

Differential Expression Analysis

Purpose:
Determine if assembled transcripts are significantly differentially expressed between RNA-Seq Samples.

Required Input:
BAM files from mapping reads to the Pathogen Portal provided genome

Output:
Tab delimited files tracking which genes are differentially expressed between sample pairs.

Select an existing Project or create a new Project to be used during this analysis and populate the Project with the necessary files. Output from this analysis will be saved in the selected Project.

Currently Selected Project: **diff. expression**

Target Project:
Select existing project — OR — Create project

TopHat2 on data 103 and data 102: accepted_hits
TopHat2 on data 2 and data 1: accepted_hits

Copy-Source:
Select source

- ☐ SRX1041267_1.fastq.gz
- ☐ SRX1041267_2.fastq.gz
- ☐ TopHat2 on data 2 and data 1: align_summary
- ☐ TopHat2 on data 2 and data 1: insertions
- ☐ TopHat2 on data 2 and data 1: deletions
- ☐ TopHat2 on data 2 and data 1: splice junctions
- ☐ TopHat2 on data 2 and data 1: accepted_hits
- ☐ Cufflinks on data 7: gene expression
- ☐ Cufflinks on data 7: transcript expression
- ☐ Cufflinks on data 7: assembled transcripts
- ☐ Cufflinks on data 7: total map mass
- ☐ TopHat2 on data 2 and data 1: accepted_hits (+ BigWig)
- ☐ TopHat2 on data 2 and data 1: accepted_hits (- BigWig)

Copy

Hint: the BAM file is the one that ends in “accepted_hits”.

- d. Click on continue and configure the Cuffdiff parameters on the next page.

Cuffdiff with cummeRbund support (version 0.0.7)

Select a reference annotation:
Plasmodium falciparum 3D7
If your annotation of interest is not listed, contact Pathogen Portal team.

Conditions

Condition 1

Name:
Sporozoite

Replicates

Replicate 1

Add replicate:
2: Tophat2 on data 2 and data 1: accepted_hits

Add new Replicate

Condition 2

Name:
Asexual

Replicates

Replicate 1

Add replicate:
1: Tophat2 on data 103 and data 102: accepted_hits

Add new Replicate

Add new Condition

Current Project History

diff. expression
2.9 GB

2: Tophat2 on data 2 and data 1: accepted_hits

1: Tophat2 on data 103 and data 102: accepted_hits

- i. Select the reference annotation, in the case *Plasmodium falciparum* 3D7.
- ii. There are two conditions that we are analyzing - for example, the asexual sample and the salivary gland sporozoites. Provide a useful name for each condition and select the replicate from the drop down menu (these are the BAM files that you copied over).
- iii. We will keep the rest of the parameters the same for the purposes of this exercise.
- iv. Click on execute.

Note that Cuffdiff will generate ~15 output files. In our case we are only going to be concerned with the file that contains differential expression analysis at the gene level. This file is called: “Cuffdiff with cummeRbund support on data 1 and data 2: gene differential expression testing”. This is a tabular file that includes many columns such as gene IDs,

14: Cuffdiff with cummeRbund support on data 1 and data 2: gene differential expression testing 5,778 lines
format: tabular, database: pfal3D7
Log: tool progress Log: tool progress [13:49:25] Loading reference annotation. Warning: No conditions are replicated, switching to 'blind' dispersion method [13:49:27] Inspecting maps and determining fragment length distributions. [13:53:43] Modeling fra

1 2 3 4

test_id	gene_id	gene	locus
PF3D7_0100100	PF3D7_0100100	PF3D7_0100100	PF3D7_01.v3:29509-371
PF3D7_0100200	PF3D7_0100200	PF3D7_0100200	PF3D7_01.v3:38981-402
PF3D7_0100300	PF3D7_0100300	PF3D7_0100300	PF3D7_01.v3:42366-465
PF3D7_0100400	PF3D7_0100400	PF3D7_0100400	PF3D7_01.v3:50362-516
PF3D7_0100500	PF3D7_0100500	PF3D7_0100500	PF3D7_01.v3:53168-532

expression values for each of the samples, fold-change and significance. You can click on the 'eye' icon to view the results. Alternatively you can click on the "visualize icon to graph a scatter plot of your results.

Scatterplot of 'Cuffdiff with cummeRbund support on data 1 and data 2: gene differential expression testing'

Log: tool progress Log: tool progress [13:49:25] Loading reference annotation. Warning: No conditions are replicated, switching to 'blind' dispersion method [13:49:27] Inspecting maps and determining fragment length distributions. [13:53:43] Modeling fra

This tab will display the chart

[Data Controls](#)

[Chart Controls](#)

[Statistics](#)

[Chart](#)

