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 (48), however only 11 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.

Step 1: Explore the FastQC results. You can do this by clicking on the view data icon (👁). An explanation of each of the FastQC results is provided as a link on the main workshop website.

Step 2: Display coverage results in GBrowse on the EuPathDB site associated with the reference genome you used to analyze your data. Take the following steps to load the BigWig file into Gbrowse:

  A. Click on the name of the file "BAM to BigWig on data …". Then click on the "Download" icon (red circle):

B. Repeat step A, with the second BigWig file.
C. Go to the EuPathDB database that contains the reference genomes that you used in your analysis. Visit a gene page in the reference organism that you used in your analysis and click on the "View in Genome Browser" link. This will take you to the genome browser with displaying your reference genes and genome.
D. In Gbrowse, click on the "Custom Tracks" tab.

E. Click on the "From File" link at the bottom of the page.

F. Click on the Browse… button and select one of the BigWig files you downloaded. Next click on the Upload button.
G. Repeat the steps E and F for the second BigWig file.

H. Once both files are done uploading, click on the Browser tab to view the coverage plots.

I. Adjust the Y axis by clicking on the "configure this track icon". Adjust to a fixed Y axis and set the maximum to a value that makes sense for the results you are looking at.
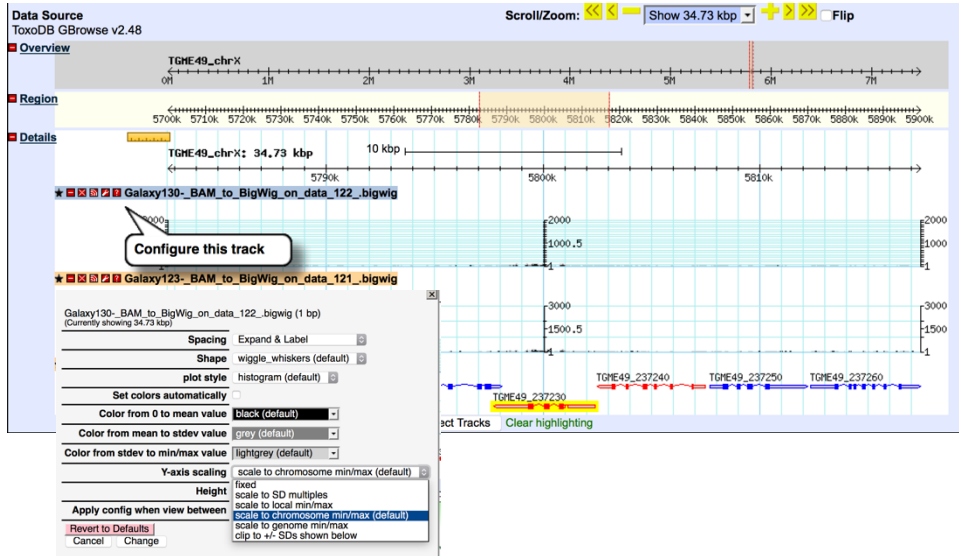
J. Explore your results in Gbrowse – zoom in or zoom out. Find any regions of interest. For example, one of your samples came from a knock out strain, go to that gene and see if you can find the difference.
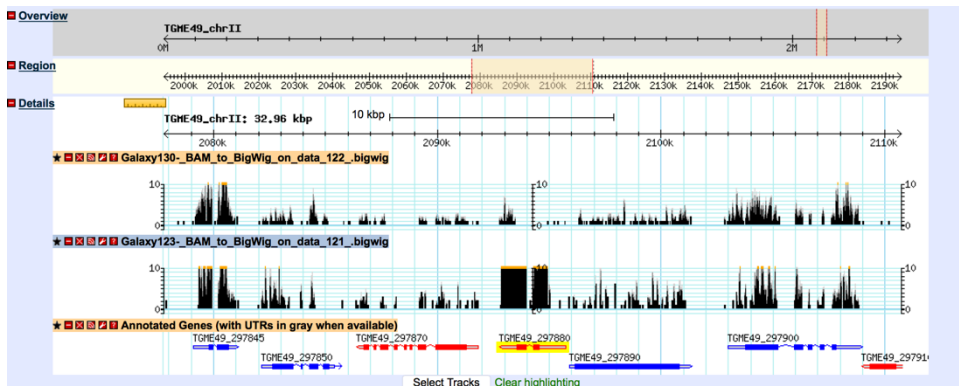
Step 3: Explore the CuffDiff results –

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. Download the CuffDiff output the same way you downloaded the BigWig files above.

*** important: the file name ends with the extension .tabular – change this to .txt then open the file in Excel.

B. Explore the results in Excel. For example, sort them based on the log2 fold change. Or by the the columns called "value1" or "value2"

C. Pick a list of 100-200 gene IDs from column C and run an ID search in EuPathDB. How would you analyze these genes in EuPathDB? One possibility is to run a GO enrichment analysis on these genes. Or you can simply look at the product descriptions, do the genes in this list make sense based on the samples you compared?