

Using Galaxy for RNA-seq Analysis

1. Go to the main Penn State University galaxy server which can be accessed at <http://main.g2.bx.psu.edu/>
2. Rather than upload some illumina data, we will be using an online dataset. This will allow us to skip the step of waiting for everyone to upload their files. To retrieve the dataset follow the link below:

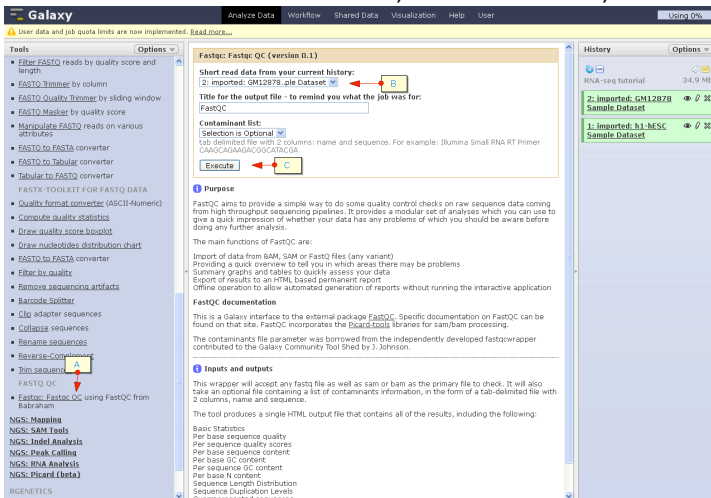
GM12878 cell line <http://main.g2.bx.psu.edu/u/jeremy/d/257ca40a619a8591>

h1-hESC cell line <http://main.g2.bx.psu.edu/u/jeremy/d/7f717288ba4277c6>

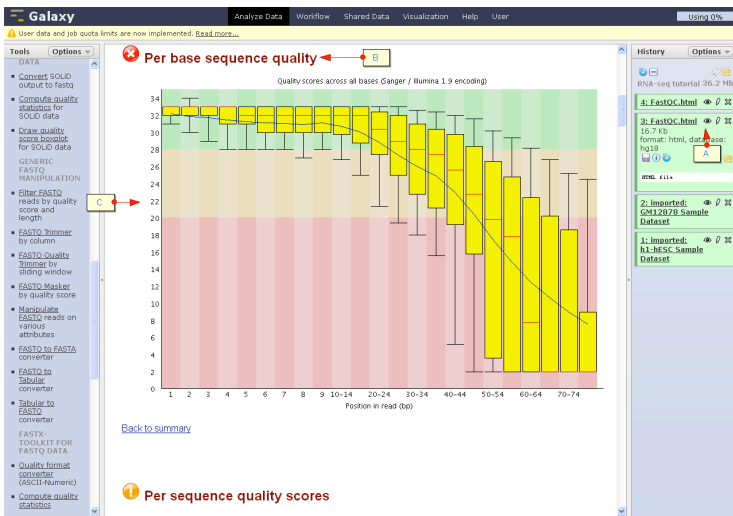
Click the **A. green +** near the top right corner to add the dataset to your history then click on **start using the dataset** to return to your history.

These subsequent steps will need to be run on each of the datasets we imported. I will be showing the screenshots for the steps related to the GM12878 cell line.

3. Click on **NGS: QC and manipulation**, scroll down to the FASTQ QC section, and choose the **A. Fastqc: Fastqc QC**. Be sure to run this on each **B. dataset**, leave the defaults, and hit **B. Execute**.



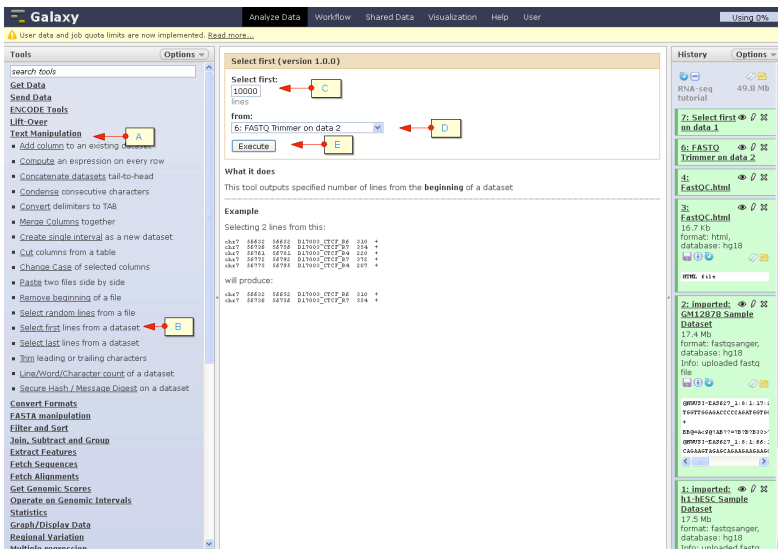
4. Next, we will want to trim the portion of the reads if the quality score drops below a certain value. Click on the **A. eye icon**, go to the **B. Per base sequence quality** boxplot results from the previous FastQC step, and **C. observe** where the quality score drops below 15.



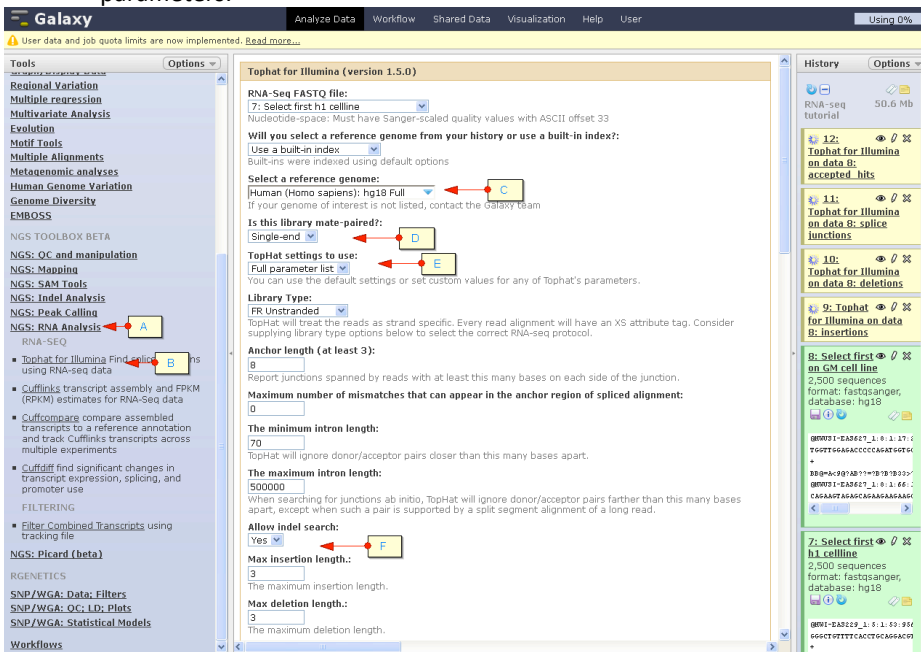
- Is trimming needed for the dataset? If so, which base pairs should be trimmed?

- Click on **A. NGS: QC and manipulation**, scroll down to the Generic FASTQ Manipulation section, and choose the **B. FASTQ Trimmer**. Choose your previous dataset, be sure to use **C. absolute values**, and **D. take the difference of your total sequence length with the number of good quality bases**. When you put this value into the correct text box, hit **E. execute**.

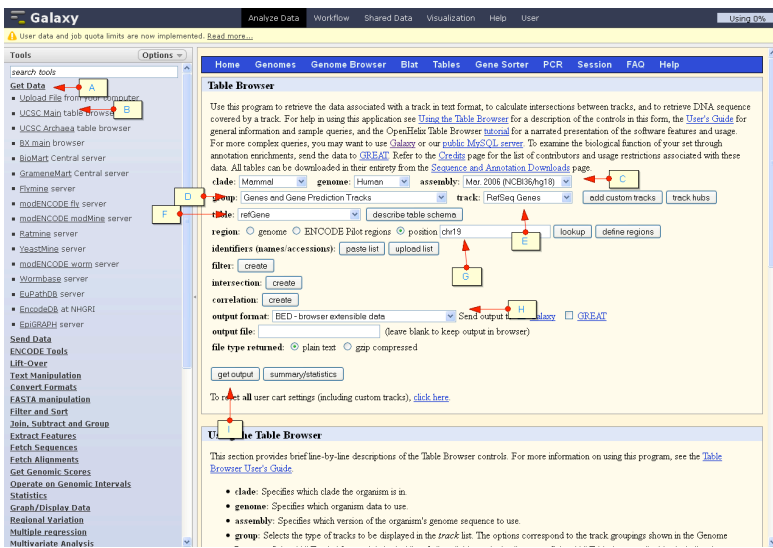
- Since these datasets will take a while to run, we are going to take a subset of sequences so that we can complete the tutorial within the lab session. To do this, choose **A. Text Manipulation** and click on **B. Select first**. We will be working with the first 2,500 sequences, which would be the first **C. 10000** lines, be sure to do this to **D. both datasets**, and hit **E. execute**.



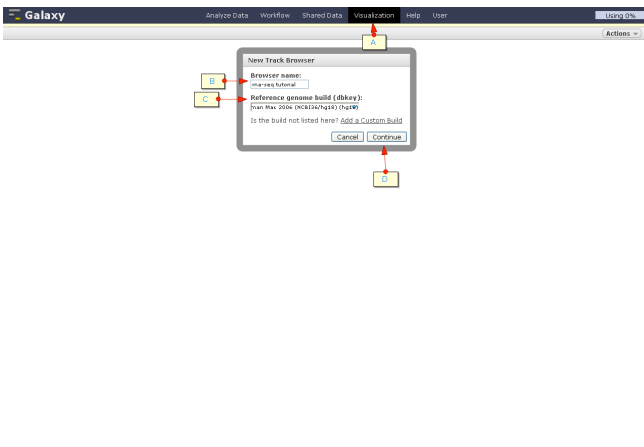
- Click on **A. NGS: RNA Analysis** and choose **B. Tophat for Illumina**. Choose your previous dataset, set your reference genome as **C. hg18 Full**, keep the default **D. single-end** setting, and choose **E. Full Parameter list** for TopHat settings. With these new options, change **F. Allow indel search** to **yes** and press the execute button when you are done setting these parameters.



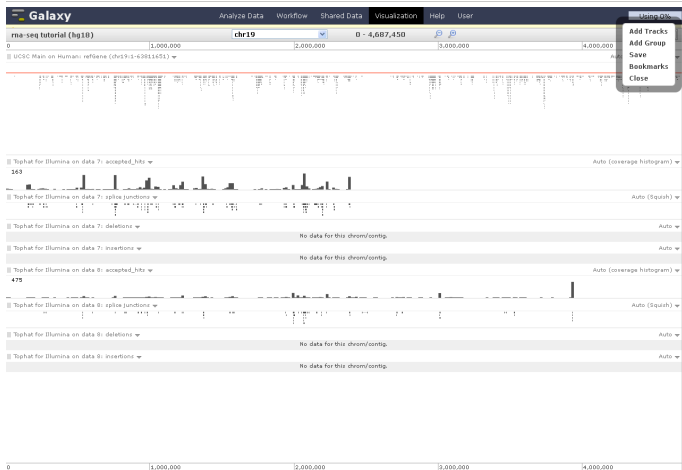
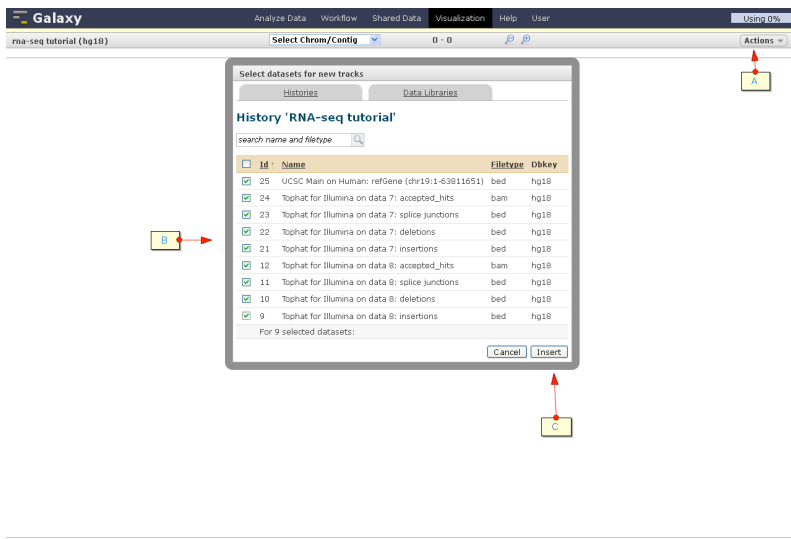
- While TopHat is running, we are going to import some gene information from the UCSC Genome Browser for visualization. Click on **A. Get Data** then **B. UCSC Main**. From this screen, make sure to use **C. Mar. 2006 hg18**, **D. Genes and Gene Prediction Tracks**, **E. RefSeq Genes**, and the table **F. refGene**. Set **G. chr19** for the position, make sure **H. BED** and send output to Galaxy is selected, and **I. press get output**. On the next page, be sure to Create one BED record per: **Coding Exons** and then press the **Send query to Galaxy** button.



9. Now that TopHat has finished running, let's visualize our results. Do this by selecting **A. Visualization > New Track Browser** from the main Galaxy menu at the top. Give it a **B. name**, be sure to select **C. hg18** as the reference genome and press **D. continue**.

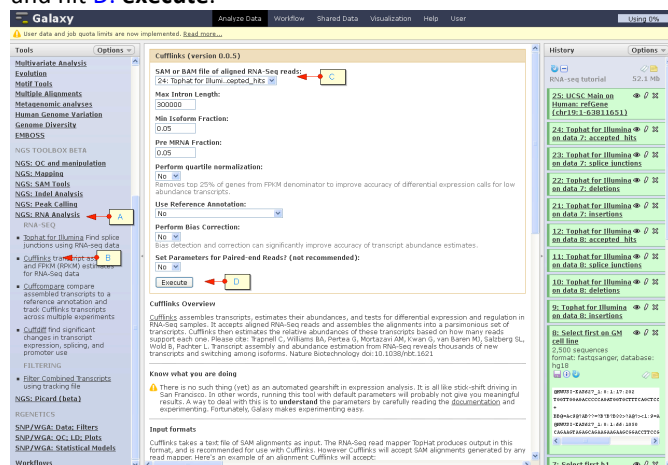


10. In the right hand corner, from the **A. Actions** dropdown menu, select **Add Tracks**. You will now have to select the rna-seq history we are working on to import the dataset. Check mark your **B. UCSC Track, TopHat accepted hits, TopHat splice junctions, TopHat insertions, and TopHat deletions**. Hit **C. continue** and these tracks will now be indexed by Galaxy. In the middle portion of the screen, from the Select Chrom/Contig dropdown menu, **select chr19**.



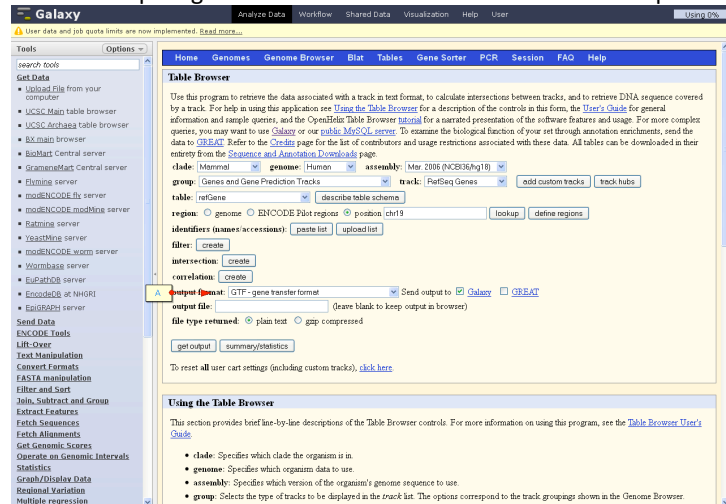
- Find an example of a splice junction between 2 known exons, and find an example where a splice junction should be found but is not.

11. Click on **A. NGS: RNA Analysis** and choose **B. Cufflinks**. Be sure to run this on **C. both datasets**, keep the default parameters and hit **D. execute**.

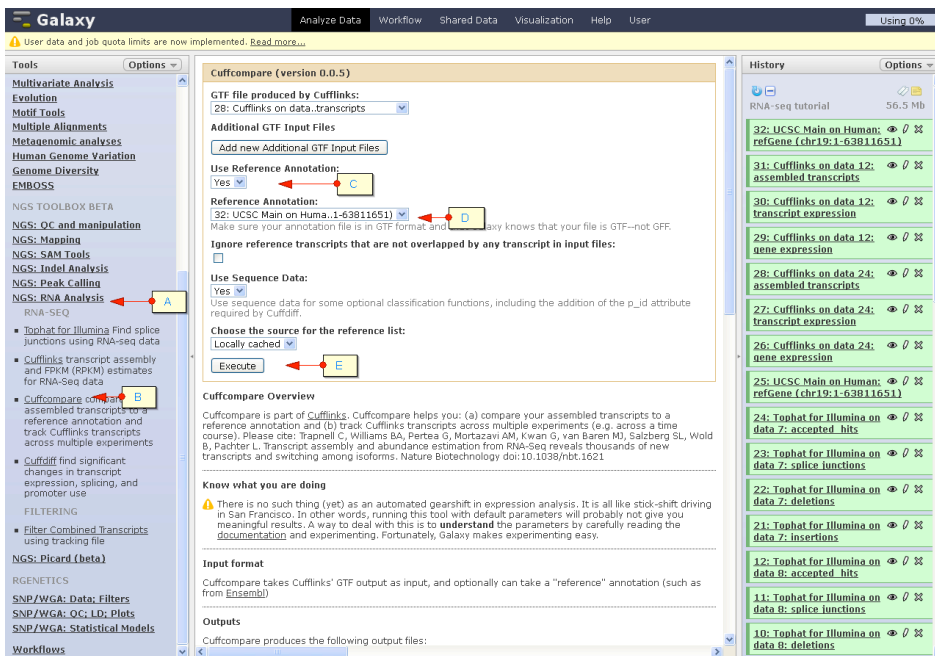


12. The next step is to run Cuffcompare on the assembled transcripts to UCSC RefSeq Genes dataset. Since Cuffcompare

requires the reference annotation to be in GTF format, we will need to get the chr19 dataset again except using GTF format. Perform step 8 again but be sure to select **A. GTF** as the output format.



13. Click on **A. NGS: RNA Analysis** and choose **B. Cuffcompare**. Choose your **cufflinks dataset**, change Use Reference Annotation to **C. Yes**, select the **D. UCSC GTF dataset** for the Reference Annotation, and hit **E. execute**.



- Find some transcripts that appear in both samples and have FPKM confidence bands that do not overlap. You can find this information by looking at the 'transcript tracking' dataset produced by Cuffcompare and [reading the Cuffcompare documentation to understand the data in this dataset](#).

14. Now, add the Cufflinks' assembled transcripts datasets to the visualization you created earlier in order to view the transcripts alongside the mapped reads, junctions, and reference genes. Refer to step 10 if you are unsure how to do this.

- Can you find examples where Cufflinks/Cuffcompare assembled a complete or almost complete transcript?

15. Our last step is to run Cuffdiff on the combined transcripts produced by cuffcompare and TopHat's accepted hits datasets

for each cell line.

Cuffdiff produces quite a few output datasets; [you'll want to browse the Cuffdiff documentation to get a sense of what they do.](#)

- Look at isoform expression dataset -- are there any significant isoform expression differences between the two samples?
- Look at the isoform FPKM tracking dataset -- find an entry for a novel isoform and an entry for an isoform that matches a reference isoform.
- What is the nearest gene and transcription start site for each entry? (Hint: [you'll need to understand the class codes, which are explained in the Cuffcompare documentation](#)).

Using Galaxy for CHIP-seq Analysis

1. Go to the main Penn State University galaxy server which can be accessed at <http://main.g2.bx.psu.edu/>
2. Rather than upload some illumina data, we will be using an online dataset. This will allow us to skip the step of waiting for everyone to upload their files. To retrieve the dataset follow the link below:
[G1E_ER4 CTCF \(chr9\) t](#)
Click the **A. green +** near the top right corner to add the dataset to your history then click on **start using the dataset** to return to your history.
3. First, for quality control, we will compute summary statistics on this dataset. Run the tool **A. NGS: QC and Manipulation** > **B. FASTQ Summary Statistics** on your dataset and hit **C. execute**. When the job completes, inspect the results.

6. Once MACS completes it will produce two datasets. One is a report on the peak calling process. The other contains the positions of the peaks.
 - How many peaks were found? Click the link to "Display at UCSC main" and you will be able to see the positions of the peaks on the genome.

Next, we will incorporate an input DNA control, import the following dataset into your history:

7. Rather than upload some illumina data, we will be using an online dataset. To retrieve the dataset follow the link below: G1E_ER4 input (chr19) <http://main.g2.bx.psu.edu/u/james/d/2979743ceed4f520>
Click the **A. green +** near the top right corner to add the dataset to your history then click on **start using the dataset** to return to your history.
8. Next, we will map these control reads to a reference genome. Use the **A. NGS: Mapping > B. Map with Bowtie for Illumina** tool. You will need to change the reference genome build you are mapping against to **C. mm9**. Otherwise you can leave the default mapping options.
9. Use the **A. NGS: Peak Calling > B. MACS** tool again. Select your **C. previous CTCF dataset** for Chip-seq tag file, but now select the **D. mapped input DNA** for "Chip-seq control file". Also, be sure to set the same **E. tag size** as you used previously.

- How many peaks are called this time?
- What is the effect of using the input control?

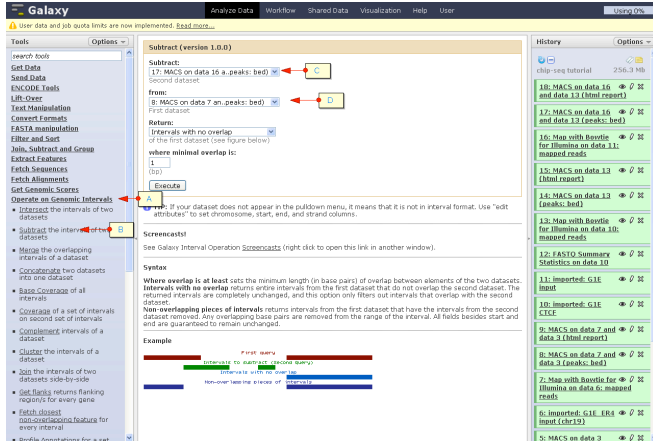
Now that we have ran through this pipeline with one sample, we are going to run through this process again with another G1E CTCF input file and another G1E Control input file.

G1E CTCF <http://main.g2.bx.psu.edu/u/james/d/d78ba454458040fd>
 G1E Control input <http://main.g2.bx.psu.edu/u/james/d/5aa51f1c13683c33>

10. Repeat steps 3-9 with these new datasets

G1E is a model for erythropoiesis, the G1E line is a GATA1 null derived line which can be induced to differentiate by estradiol treatment (thus G1E-ER4). Here we will use Galaxy to identify sites that have differential binding across the two developmental stages.

11. Select the **A. Operate on Genomic Intervals > B. Subtract** tool. For the first input ("Subtract") select your **C. second set of peaks** (Peaks from G1E), for the second input ("from") select your **D. first set of peaks** (Peaks from G1E-ER4) and run the tool. The resulting dataset contains peaks that are only present in the differentiated line.



- How many are there?

12. Perform the subtract operation, switching the input datasets, to find peaks that are unique to the undifferentiated line.

13. Finally, load the **A. Graph / Display Data > A. Build Custom Track** tool. Add each of your three tracks (Intersect, ER4, and G1E) by clicking **Add Track** and give them descriptive names. Run the tool, and inspect the resulting dataset when complete. Click "Display at UCSC main" and all three tracks will be displayed in the UCSC browser. You can now inspect differential CTCF binding sites between two differentiation time points.

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 0%

User data and job quota limits are now implemented. [Read more...](#)

Tools Options

ENCODE Tools

Lift-Over

Text Manipulation

Convert Formats

FASTA manipulation

Filter and Sort

Join, Subtract and Group

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Graph/Display Data A

- Histogram of a numeric column
- Scatterplot of two numeric columns
- Plotting tool for multiple series and graph types
- Boxplot of quality statistics
- GMAI Multiple Alignment Viewer
- Build custom track for [UCSC genome browser](#) B
- VCF to MAF Custom Track for display at UCSC
- Mutation Visualization

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

Human Genome Variation

Genome Diversity

EMBOSS

Build custom track (version 1.0.0)

Tracks

Track 1

Dataset: 20: Subtract on data 8 and data 17 C

name: ER4

description: ER4

Color: Red

Visibility: Dense

Remove Track 1

Track 2

Dataset: 19: Subtract on data 17 and data 8 D

name: G1E

description: G1E

Color: Orange

Visibility: Dense

Remove Track 2

Add new Track

Execute

! This tool allows you to build custom tracks using datasets in your history for the UCSC genome browser. You can view these custom tracks on the UCSC genome browser by clicking on [display at UCSC main/test](#) link in the history panel of the output dataset.

History Options

20: Subtract on data 8 and data 17

62 regions, 1 comments
format: bed, database: mm9

display at UCSC main
view in GeneTrack
display at Ensembl Current

19: Subtract on data 17 and data 8

238 regions, 1 comments
format: bed, database: mm9

display at UCSC main
view in GeneTrack
display at Ensembl Current

18: MACS on data 16 and data 13 (html report)

17: MACS on data 16 and data 13 (peaks: bed)

16: Map with Bowtie for Illumina on data 11: mapped reads