Alexander Gilson

Genomics Spring 2021
Units 11 - 12 Graded Homework

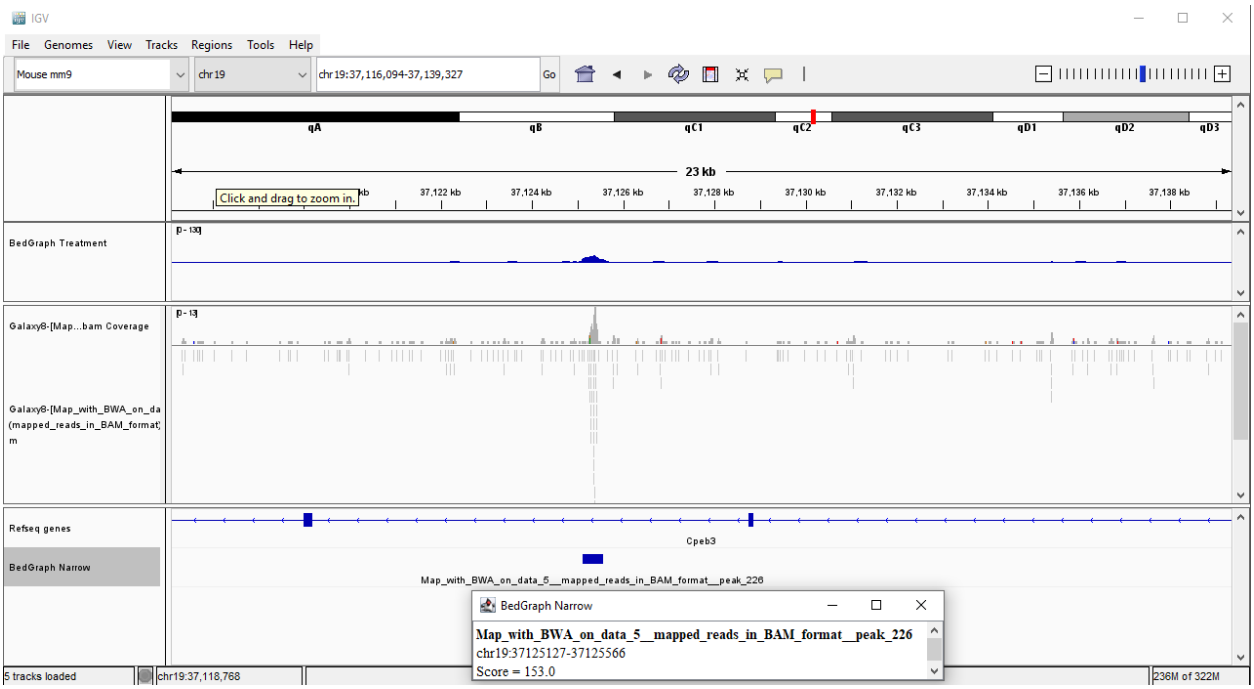
Part 1: ChIP-seq data analysis

Import the attached downsampled FASTQ file of ChIP-seq reads from mouse (mm9). The full dataset was downsampled to the subset reads from chr19. Follow the analysis protocol below:

Run FASTQC to determine the quality score encoding(**DONE**)
Run FASTQ Groomer to convert the file to Sanger/Illumina 1.9 phred encoding ONLY IF NEEDED (**was illumina 1.5, converted to 1.8+)**
Run Trimmomatic and set the minimum phred score in a 4 nt sliding window to 25(**used single-end)**
Re-run FASTQC to check the quality scores and encoding scheme (**Checked, the groomer did not convert the file, had to run again, selected wrong argument on initial groomer run.**)
Run Map with BWA with default settings (make sure to select for single-end reads), aligning against the mouse genome version mm9.
Run MACS2 callpeak on the BAM file, setting the Effective genome size to the mouse genome. Use default settings for the rest of the parameters and leave the control field blank. (**used M. musculus, 1.87e9**)

Part 1: Submission (2.25 pts)

(1 pts) Load the MACS2 Bedgraph Treatment file and narrow Peaks BED file from step 6 and aligned BAM file from step 5 to IGV or UCSC. Find a gene locus that has ChIP peaks nearby. Submit a screenshot image of the locus. Be sure the tracks are labeled so I know which is which.



(0.5 pts) Submit the narrow Peaks BED file.

- File name: Galaxy9-[MACS2_callpeak_on_data_8_(narrow_Peaks)].bed

(0.25 pts) MACS2 produces a Bedgraph file, not a WIG file. How do those two file types differ? Can Bedgraph files be converted to WIG format, and vice versa?

- While both bedGraph and Wig can be used to display data in track format in programs such as IGV, bedGraph is able to export data to these tracks in their original state, while WIG must be changed to be viewed, though WIG is better suited for large datasets. BedGraph can be converted to WIG and vice versa on Galaxy through the Wig/BedGraph-to-bigWig converter found in the tools section of the site.

https://genome.ucsc.edu/goldenPath/help/bedgraph.html

(0.25 pts) MACS2 has the option of generating 'broad peaks'. What type of ChIP-seq data should be analyzed for 'broad peaks' instead of 'narrow peaks'? Why?

- Broad peaks are used for analysis of ChIP-seq data when attempting to identify regions in which histone modifications are covering entire genes, while narrow peaks(used for this assignment) are good for identifying regions at which transcription factors bind, therefore making the ChIP-seq data that analyzes histone modification patterns a method that should use broad peaks rather than narrow.

- https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

(0.25 pts) MACS2 has the option to remove duplicate reads before peak-calling. What are duplicate reads and why would one choose to remove them?

- These duplicate reads are any reads that occur at the same locus on the same strand, which can occasionally become troublesome when attempting to read what is occurring at each locus. One would remove these reads if the initial starting material is low in the sample, and therefore creates a region that is biased by PCR which can lead to many duplicates of the same region, leading to a high signal where one should not exist.

- https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

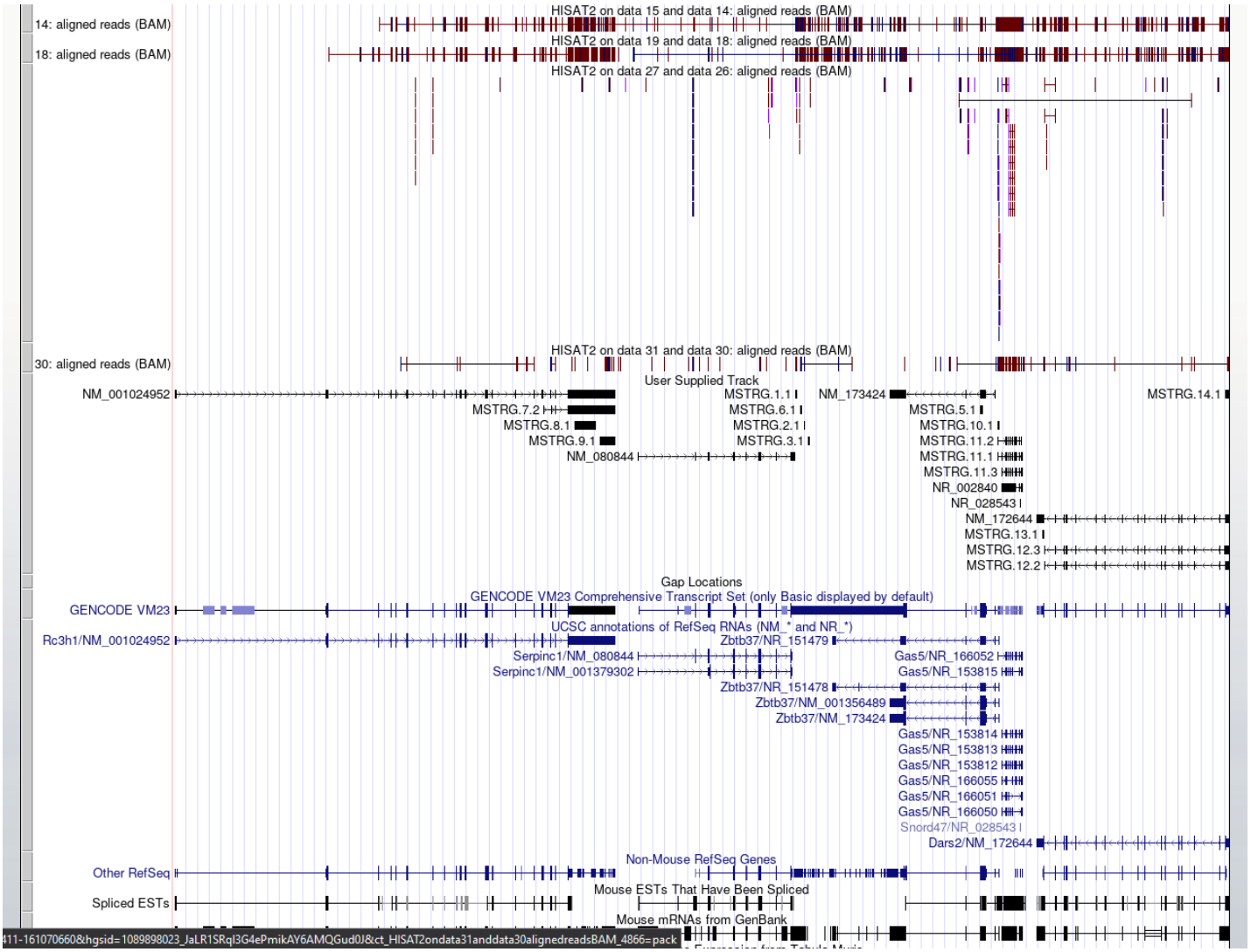
Part 2: RNA-seq data analysis

Follow this Galaxy RNA-seq tutorial on Galaxy Main (usegalaxy.org), including the Vizualization section. It will take some time to run various steps, so don't wait until the last minute!

Part 2: Submission (2.75 pts)

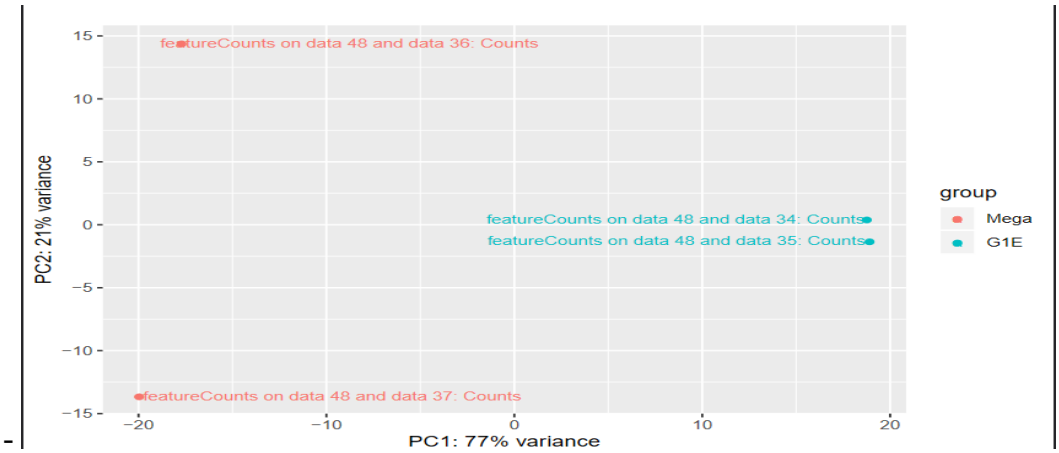
(1 pt) Use any genome browser, e.g., IGV, UCSC Genome Browser, or Trackster, to the coordinates: chr11:96193539-96206376. Describe what you see in terms of known and novel transcripts, from both the G1E and Megakaryocyte cell lines.

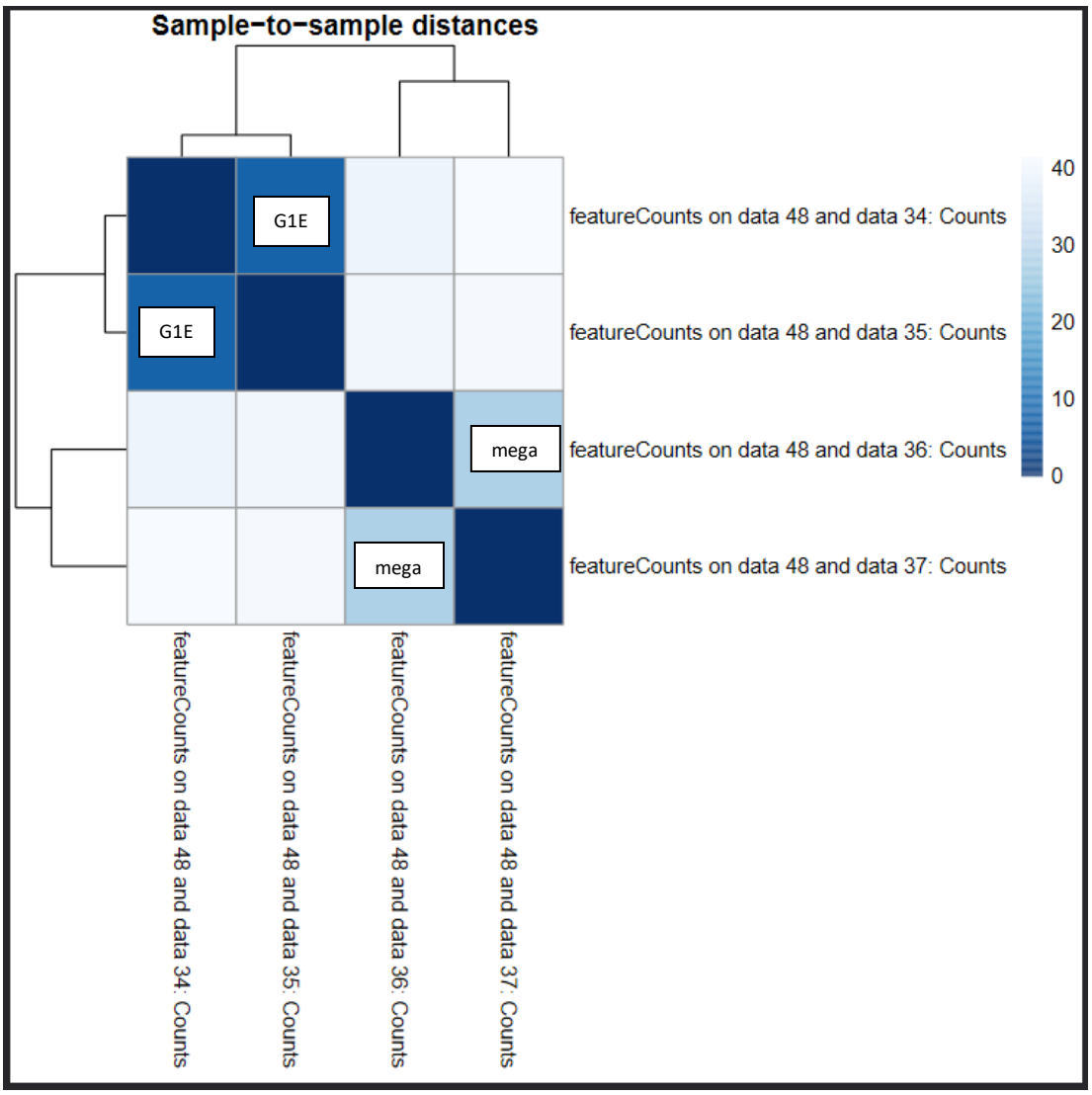
- By examining the StringTie Merge file, I was able to see each transcript displayed as a track on UCSC table browser. This not only showed the 4 different original transcripts from the original files used at the start of the Galaxy tutorial, but included their trimmed versions yielding more significant data points than all of the low quality data points included with the original files. These low quality data points occur over stretches of longer reads, where they become less informative as the sequencing goes on. The StringTie merge file also includes these sequences after they have been aligned to a reference sequence, making comparison to known transcripts easier. When compared to the track of transcripts from UCSC, differences between the two cell lines become a bit more clear and easier to understand. The G1E reads (HISAT2 on 15/14 and 19/18) are shown in the UCSC genome browser to match up well with gaps seen in the GENCODE VM23 track shown in the Gap Locations section. Meanwhile, the Mega cell line gaps are a lot less informative than the G1E reads. In terms of how the different cell lines match up with the sequences shown in the UCSC browser, the G1E cell line has areas more dense in regions that match well with NM_001024952, as well as the NM_173424 sequence shown. It is interesting to note that there are regions of the G1E transcript shown that do not match wit a transcript. The Mega cell line on the other hand does not match well with most of the transcripts, but is found to match to multiple with those areas not in the gaps having some relation to the known transcripts in that region.



(0.5 pts) How well do the G1E and Megakaryocyte RNA-seq replicates agree? What is your evidence? Submit and describe two figures that support your conclusion. (HINT: Look at DESeq2 output).

- Fortunately, running DESeq2 via Galaxy provides not only an informative file that can be filtered to examine differential expression or normalization of the data, but also provides a visualization PDF file that includes plots of varying means to examine the data. Two plots in particular, the principal component analysis and distance analysis plots, can help determine how well the replicates of these cell lines agree. From the principal component analysis, one can see that the 2 G1E cell line replicates (light blue) agree rather well, in both PC1 and PC2 since they are relatively close together on the plot. The Mega cell lines(red) seem not to agree as much, seen spread rather far away from each other on the Y(PC2) axis of the plot. For the sample-to-sample distance plot, it can be seen that the data for G1E distance(top left) is much closer than those seen for the Mega replicates, showing a more dark blue indicating closer distance/relation of the datapoints.





(0.5 pts) How many transcripts have a significant (adjusted p-value < 0.01) change in expression between these conditions? How many transcripts are up-regulated in G1E? How many transcripts are down-regulated in G1E?

- According to another filter run of c7 <=0.01 shows that 70 of the 587 transcripts found with DESeq2 have a significant change of expression between these two conditions. Determining which of the G1E transcripts were up-regulated was accomplished with the filter c3 >=0.0, showing that 37 transcripts are up-regulated. The same was done for down-regulation with c3 < 0.0, showing 33 transcripts are down-regulated.

(0.75 pts) Choose a transcript that is differentially expressed from part c and has a log2 fold change of at least 2 or -2. What is the transcript? What is the biological function of the gene corresponding to this transcript? In which cell type is this transcript more highly expressed, and by how much? Make a conjecture about how the difference in expression of this gene might explain or be a result of the cell types examined.

-

GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj
NM_001003947	57.6194639393993	-9.932868739288848	1.695218697635542	-5.859343548501 06	4.64700450809862e-09	6.72448887642506e-08

I chose the transcript with the GeneID of NM_001003947. This is shown to be heavily down-regulated in the G1E cell line with a log2(FC) score of -9.93. Its NCBI page (https://www.ncbi.nlm.nih.gov/nuccore/NM_001003947) states that this is related to cytochrome P450, specifically the CYP4x1 mRNA. Cytochrome P450 is a super family of enzymes that are implicated in hormone synthesis, breakdown, and drug metabolism(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4093435/). The CYP4x1 mRNA is specifically associated with the mouse brain, which could be implicated with many different hormonal changes, so its down-regulation in the G1E cell line could indicate a serious up-regulation of a hormonal substrate, by not having enough of the CYP4x1 enzyme readily available. It could also affect drug metabolism in the brain of mice from the G1E cell line, and therefore could impact animal model studies for drugs that could be impacted by this transcript's down-regulation.