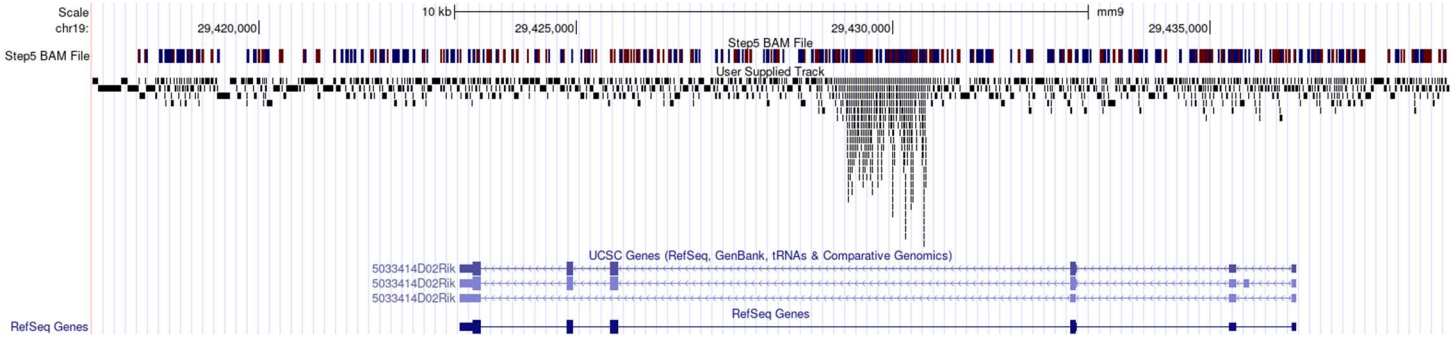
CHIP-SEQ DATA ANALYSIS

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:

1. Run FASTQC to determine the quality score encoding
2. Run FASTQ Groomer to convert the file to Sanger/Illumina 1.9 phred encoding ONLY IF NEEDED
3. Run Trimmomatic and set the minimum phred score in a 4 nt sliding window to 25
4. Re-run FASTQC to check the quality scores and encoding scheme
5. Run Map with BWA with default settings (make sure to select for single-end reads), aligning against the mouse genome version mm9
6. 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.

a. 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.



Top = STEP 5 BAM File
User Supplied Track = STEP 6 BED File

b. 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?

A wig file can be converted to a Bedgraph file, and it is advisable to do so when there are distances of over 100bp between variable step points. Bed graph is zero-based and WIG is 1-based. Bedgraph cannot be converted to wig.
http://genome.ucsc.edu/goldenPath/help/wiggle.html

c. 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?

When MACS2 generates 'broad peaks', it tries to make a BED file (BED6 + 3 format) with broad regions, but 'narrow peaks' is a BED6 + 4 format that has the peak summit, pvalue, and qvalue. The broad peaks file doesn't have a 10th column containing the peak summits. Broad peaks are better for large regions such as histones, but narrow regions is better for shorter regions such as transcription factors so you can see the peak summit.

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

Duplicate reads could mean the DNA is bound to a histone, PCR bias, or noise due to some type of error. You would remove these if you want to be sure your signals are all high-quality, and there is no noise.

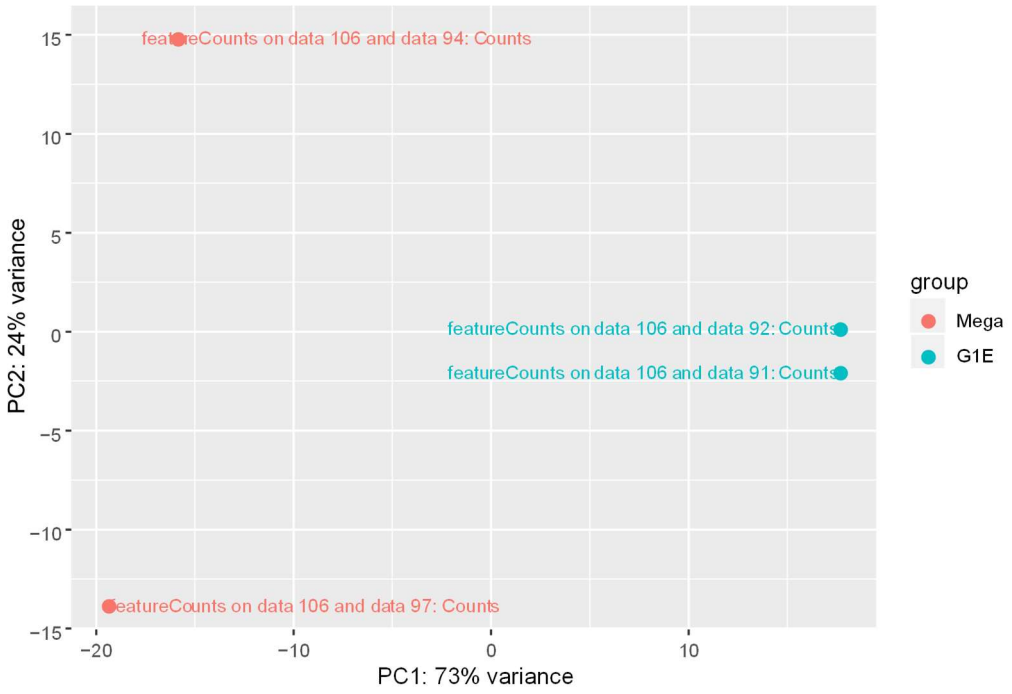
RNA-SEQ DATA ANALYSIS

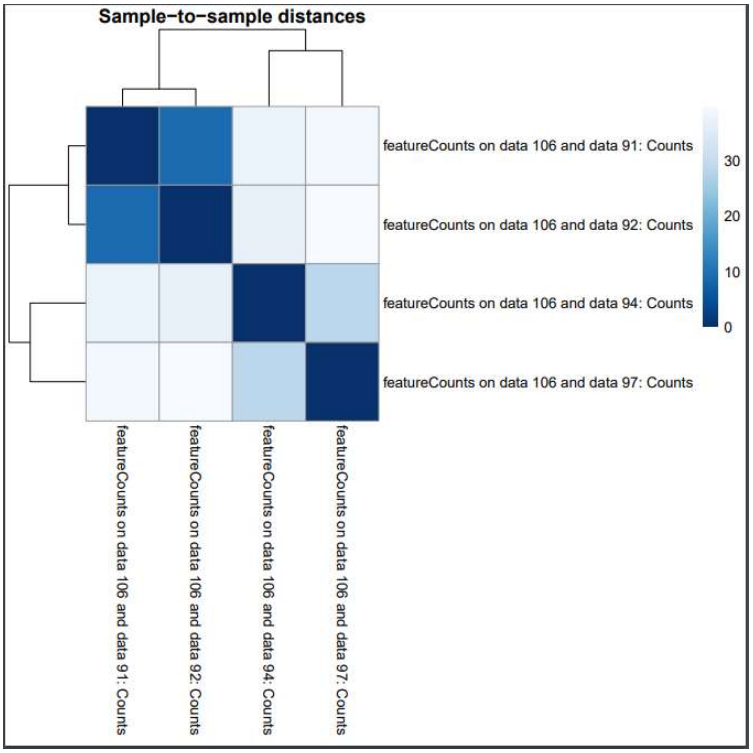
1. Use any genome browser, e.g., IGV, UCSC Genome Browser, or Trackster, to the coordinates: chr11:96193539-96206376. Describe what you see in terms of known and novel transcripts, from both the G1E and Megakaryocyte cell lines.

   There is a known transcript on the forward strand, and one on the reverse strand of the reference genome. There are genes in the GFF Compare track indicating they are novel

2. 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).

   G1E and Megakaryocyte RNA-seq replicates do not agree well. They do not line up in the visualization, the PCA plot shows a large distance between the two groups, and the heatmap of sample-to-sample distances shows low correlation between the G1E samples and the Megakaryocyte samples.





3. 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?

   There are 51 transcripts that have significant changes in expression. 21 of these genes are down regulated and 30 are upregulated.