Julie Garcia
Genomics Spring 2018
Units 11 - 12 Graded Homework

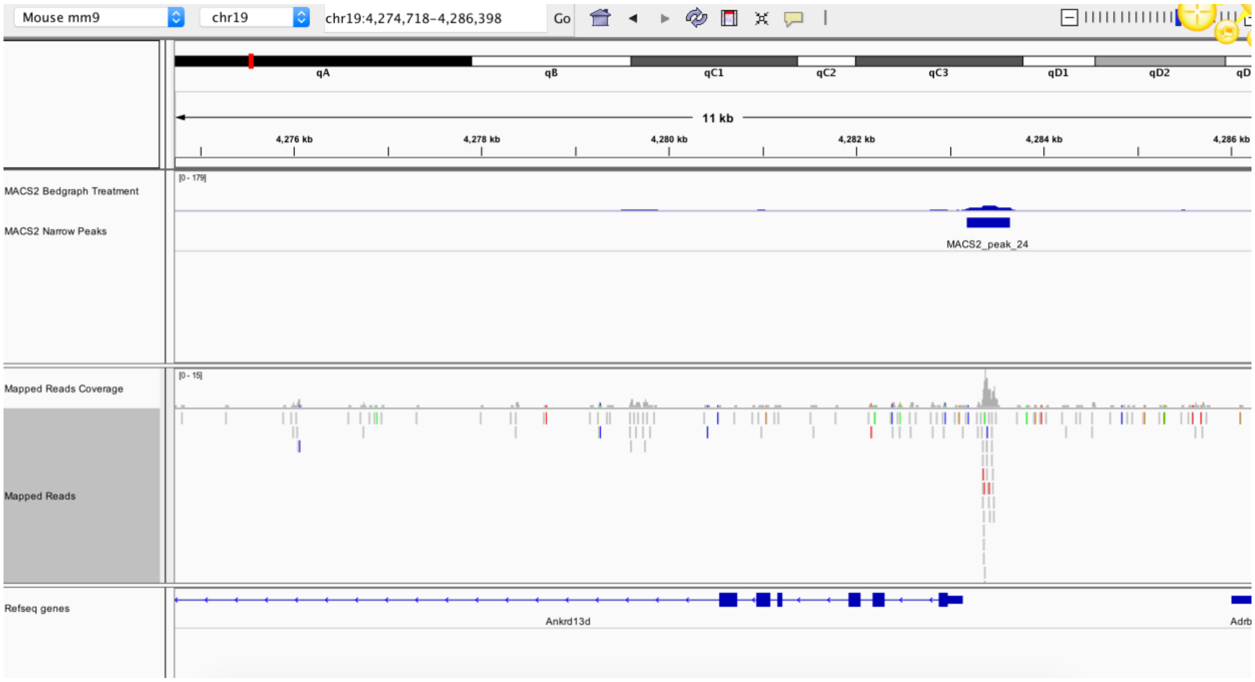
Part 1: ChIP-seq data analysis

I imported the attached downsampled FASTQ file of ChIP-seq reads from mouse (mm9) and ran the following protocol.

1. I ran **FASTQC** to determine the quality score encoding.
2. I ran **FASTQ Groomer** and converted the file from Illumina 1.5 to Sanger/Illumina 1.9 phred encoding.
3. I ran **Trimmomatic** with minimum phred score in a 4 nt sliding window to 25.
4. I re-ran **FASTQC** to check the quality scores and encoding scheme.
5. I ran **Map with BWA** and aligned the fastq data to the mouse reference genome (mm9).
6. I ran **MACS2 callpeak** on the BAM file, setting the Effective genome size to the mouse genome using the default settings.

Part 1: Submission (2.25 pts)

a. I loaded the Mouse (mm9) genome into IGV and loaded the MACS2 Bedgraph Treatment file, narrow peaks BED file and BAM alignment file. I zoomed into the gene ANKRD13D that had a CHIP-seq MACS peak near the promoter.



b. I submitted the narrow peaks bed file.
c. Bedgraph files are in BED format with zero based chromosome start locations and have four columns (Chr, chrstart, chrend, datavalue). WIG files have two columns (chrStart datavalue) and can be 0-based or 1-based depending on how they were created. There

is a conversion tool to go from Bedgraph to BigWig and WIG to BigWig in Galaxy. BigWig files can also be converted to WIG. To go from WIG to Bedgraph, there is a tool Wiggle-to-interval in Galaxy.

d. Narrow peaks are generally used for finding transcription factor binding sites and broad peaks are generally used to find larger areas of modification like histone modifications. This is because the broad peaks show larger areas of modification, while narrow peaks show small areas where the TF binding sites are located.

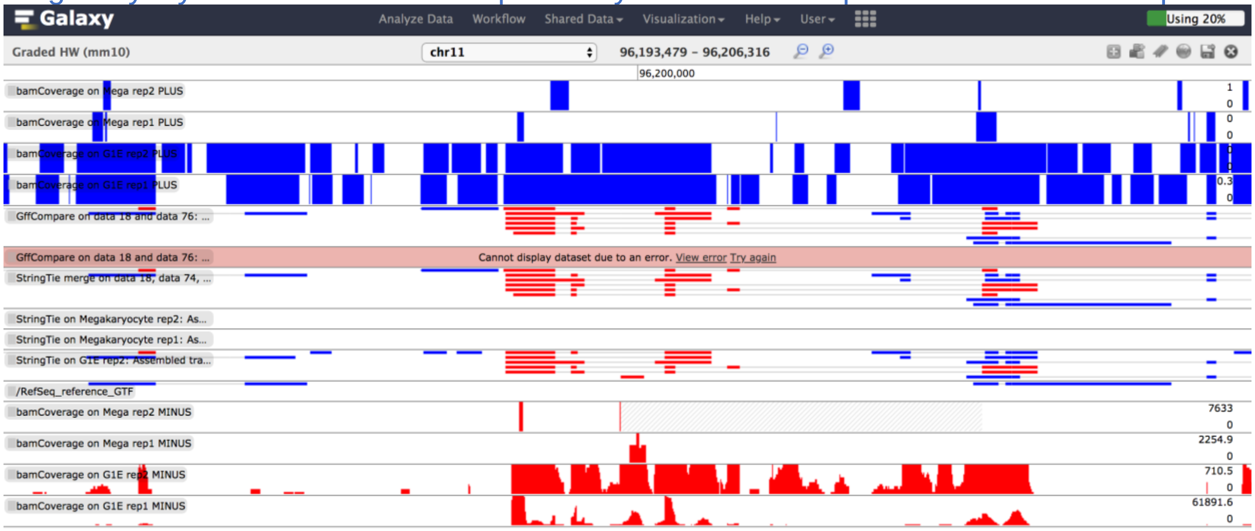
e. Duplicate reads are copies of the same read, and can represent experimental artifacts and can be due to PCR amplification bias, or they can be legitimate duplicates that repepresent highly enriched areas. They are often flagged and removed when doing Chip-Seq analysis and you would want to revmoe them in order to normalize the data before analysis, although there seems to be some debate about whether or not this step is necessary.

Part 2: RNA-seq data analysis

I uploaded the fastqsanger and gtf files to Galaxy and ran the tutorial.

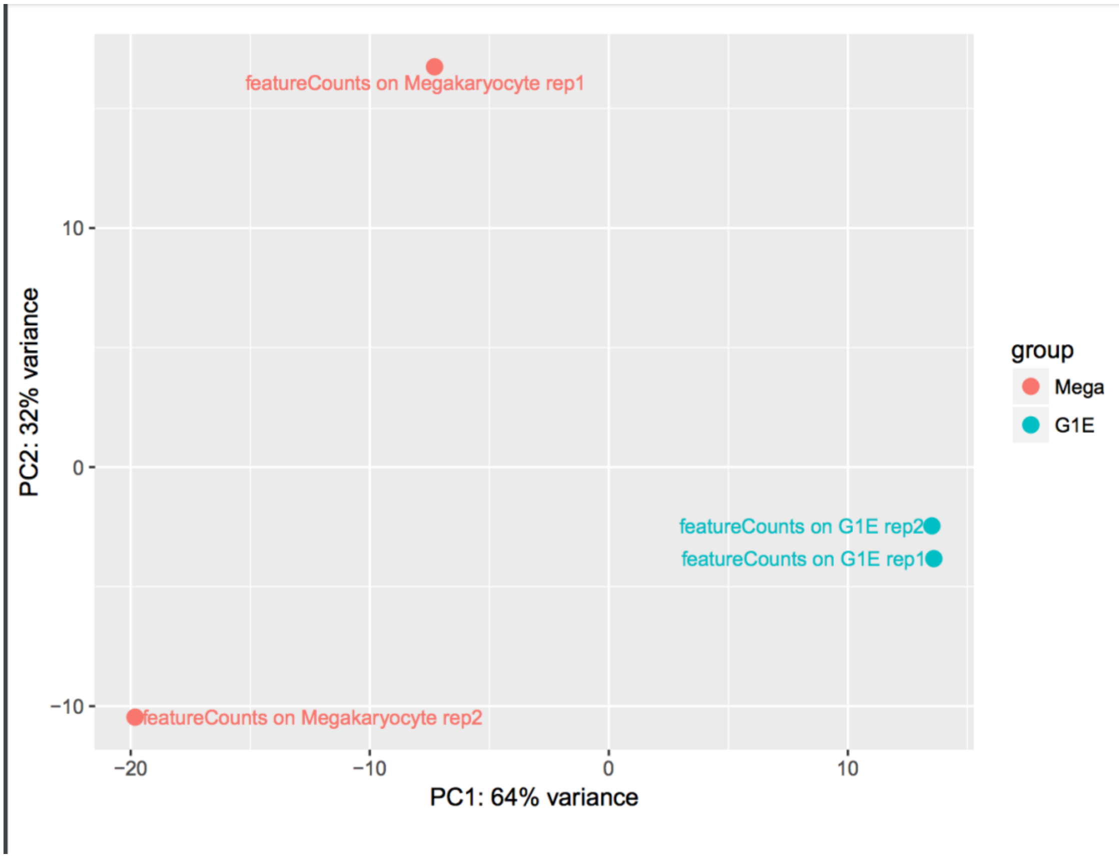
Part 2: Submission (2.75 pts)

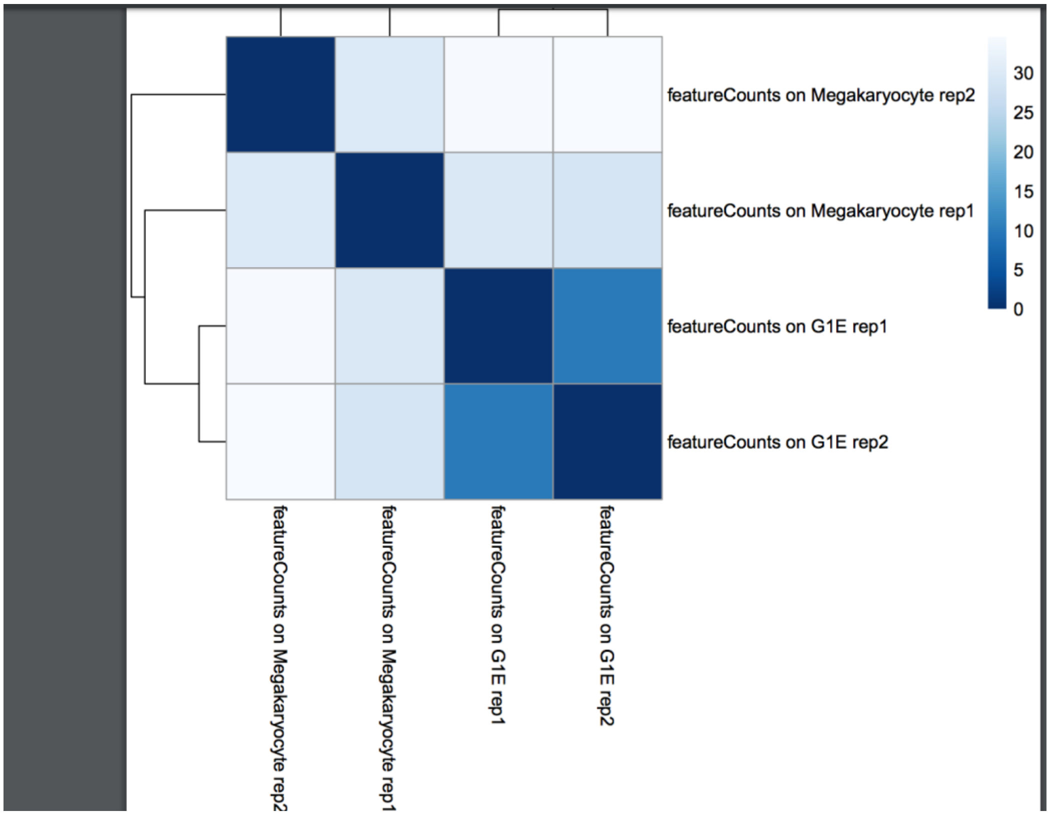
a. As you can see below the reference transcript shows only two gene transcripts on the plus strand. On G1E, has novel transcripts on the plus and minus strands and the Megakaryocyte cell line shows the possiblity of a few small possible novel transcripts.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 20%
Graded HW (mm10)
chr11
96,193,479 – 96,206,316
96,200,000
bamCoverage on Mega rep2 PLUS
bamCoverage on Mega rep1 PLUS
bamCoverage on G1E rep2 PLUS
bamCoverage on G1E rep1 PLUS
GffCompare on data 18 and data 76: ...
GffCompare on data 18 and data 76: ...
StringTie merge on data 18, data 74, ...
StringTie on Megakaryocyte rep2: As...
StringTie on Megakaryocyte rep1: As...
StringTie on G1E rep2: Assembled tra...
/RefSeq_reference_GTF
bamCoverage on Mega rep2 MINUS
bamCoverage on Mega rep1 MINUS
bamCoverage on G1E rep2 MINUS
bamCoverage on G1E rep1 MINUS
1
0
0
0
0.3
0
7633
0
2254.9
0
710.5
0
61891.6
0

Cannot display dataset due to an error. View error Try again

b. The replicates for G1E and Megakaryocyte do not match up well. The two plots below show the association between the G1E replicates and the Megakaryocyte cell replicates and you can clearly see that the replicates do not agree between the two cell lines.


featureCounts on Megakaryocyte rep1
featureCounts on G1E rep2
featureCounts on G1E rep1
featureCounts on Megakaryocyte rep2
10
0
−10
−20
−10
0
10
PC2: 32% variance
PC1: 64% variance
group
● Mega
● G1E


featureCounts on Megakaryocyte rep2
featureCounts on Megakaryocyte rep1
featureCounts on G1E rep1
featureCounts on G1E rep2
30
25
20
15
10
5
0
featureCounts on Megakaryocyte rep2
featureCounts on Megakaryocyte rep1
featureCounts on G1E rep1
featureCounts on G1E rep2

c. I filtered the DeSeq data for adjusted p-values < 0.01 and came up with 39 transcripts. Further filtering by the log2(FC) shows 20 underexpressed transcripts and 19 overexpressed transcripts in G1E.

d. NM_008267 has a log2 fold change of 12, so this transcript is overexpressed. This is a transcript of the gene Hoxb13, that produces homeobox protein HoxB-13. This protein is a transcription factor that plays a role in the regulation of positional identities in the anterior-posterior axis

References:

1. https://www.ncbi.nlm.nih.gov/protein/6680247
2. https://www.ncbi.nlm.nih.gov/protein/2495325