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:

1. **Run FASTQC to determine the quality score encoding**

   Quality was encoding Illumina 1.5.

2. **Run FASTQ Groomer to convert the file to Sanger/Illumina 1.9 phred encoding ONLY IF NEEDED**

   Ran and confirmed with FASTQC to be Illumina 1.9

3. **Run Trimmomatic and set the minimum phred score in a 4 nt sliding window to 25**

   Done

4. **Re-run FASTQC to check the quality scores and encoding scheme**

   Done, quality improved.

5. **Run Map with BWA with default settings (make sure to select for single-end reads), aligning against the mouse genome version mm9**

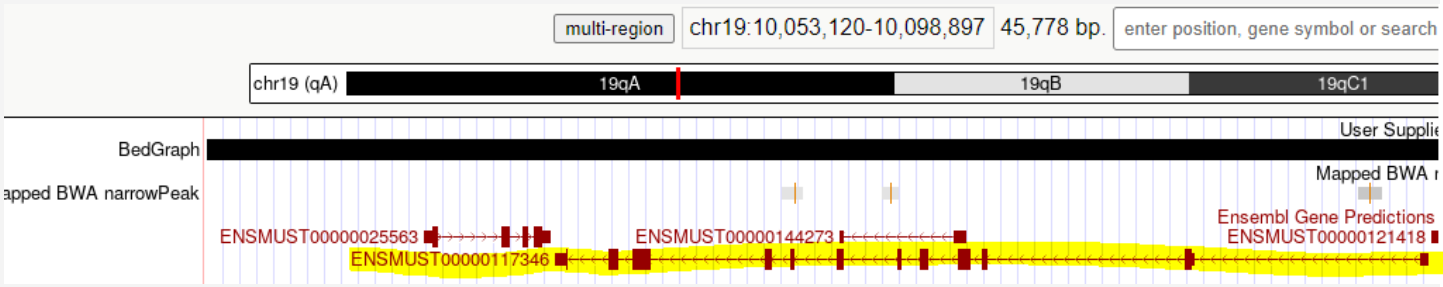
   Done

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.**

   Done

Part 1: Submission

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.**



*Full image also attached

B. Submit the narrow Peaks BED file.

*File is attached

C. 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?

Unlike WIG files, BedGraph files stay in their original state and are not compressed, taking lots of memory. BedGraphs can be converted to each other, but converting BedGraphs to WIG format is more challenging.

Source: http://genome.ucsc.edu/goldenPath/help/bedgraph.html

D. 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 is good for histone modifications, while narrow peaks is good for transcription factors. Histone modifications spans large broad areas, while transcription factors span a much smaller area.

Source: Starmer, J., Magnuson, T. Detecting broad domains and narrow peaks in ChIP-seq data with *hiddenDomains* . *BMC Bioinformatics* 17, 144 (2016). https://doi.org/10.1186/s12859-016-0991-z

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

Part 2: RNA-seq data analysis

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

When you upload files, use the following links to avoid errors.

file type: fastqsanger, genome:mm10

https://zenodo.org/record/583140/files/G1E_rep1_forward_read_%28SRR549355_1%29
https://zenodo.org/record/583140/files/G1E_rep1_reverse_read_%28SRR549355_2%29
https://zenodo.org/record/583140/files/G1E_rep2_forward_read_%28SRR549356_1%29
https://zenodo.org/record/583140/files/G1E_rep2_reverse_read_%28SRR549356_2%29
https://zenodo.org/record/583140/files/Megakaryocyte_rep1_forward_read_%28SRR549357_1%29
https://zenodo.org/record/583140/files/Megakaryocyte_rep1_reverse_read_%28SRR549357_2%29
https://zenodo.org/record/583140/files/Megakaryocyte_rep2_forward_read_%28SRR549358_1%29
https://zenodo.org/record/583140/files/Megakaryocyte_rep2_reverse_read_%28SRR549358_2%29

file type: gtf, genome:mm10

https://zenodo.org/record/583140/files/RefSeq_reference_GTF_%28DSv2%29

Part 2: Submission

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.

For the most part, transcripts from G1E and Megakaryocyte cell lines agree, but there are some with differential expression.

A. 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).

The RNA-seq duplicates generally agree.

B. 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?

51 transcripts had significant change in expression to a p-value of <.01. 30 transcripts were upregulated, whereas 21 genes were downregulated.

C. 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.

The transcript MSTRG.28.1 has the highest differential expression with log2 value of over 8. MSTRG.28.1 is overexpressed substantially in the G1E cell line compared than compared to the Megakaryocyte expression. This transcript corresponds to a gene that is part of the septin family in GTP binding.

Source: https://www.genecards.org/cgi-bin/carddisp.pl?gene=SEPTIN14

Mouse_ChIP-Seq_Example_Experimental_Data_chr19_mm9.fastq.gz