This homework follows the basic framework of the practical session we do on the class using bam files (same as sam files, but binary, hence smaller). The data was downloaded from <https://www.encodeproject.org/> and includes ChIP-seq for the insulator CTCF, two histone modifications we mentioned (H3K4me3 and H3K27me3) two chromatin regulators RNF2 and PHF8. RNF2 is part of the PcG (polycomb group) that “writes” H3K27me3, and PHF8 “reads” H3K4me3. Read here some info about these enzyme [http://www.genecards.org/](http://www.genecards.org/cgi-bin/carddisp.pl?gene=RNF2). It is a good practice to read a bit about each enzyme you encounter.

The bam files for this assignment are located in:

/oasis/tscc/scratch/biom200/cmm262/ChIP-seq-Homework/

**Answer Format:**

1. **Please either directly add your code to the word document or supply a separate file containing all steps you ran with a comment indicating what step what command belongs to.**
2. **Submit your answers to Patrick by Tuesday, 02/27 (**[**pfiaux@ucsd.edu**](mailto:pfiaux@ucsd.edu)**)**
3. **Send you file in the format: Lastname\_First\_CMM262\_Homework3**
4. **Send your answers as a word document.**

**1**. Create tag directories for each of the bam files using a for loop.

**Q1:** What is the average GC% of the ChIP-fragments found in each of the datasets?

2. Next visualize the ChIP-seq experiments by creating bedGraph files from the tag directories and use the UCSC genome browser upload and view the tracks. Go to the hox gene cluster (e.g. go to HOXA1, and zoom out to chr7:26,987,134-27,482,887).

**Q2:** Within this range, what is the level of co-localization between H3K4me3 and H3K27me3? Between RNF2 and H3K27me3? Given the nature of H3K27me3 and H3K4me3, how would you define these genes?

**3**. Find peaks for the three ChIP-seq experiments. RNF2 and PHF8 are chromatin regulators. What do you think should be used as ‘-style’: histone or factor? Report the peaks for both running either on the entire data set. Don’t forget the Input experiment as well! Convert the regions.txt and peaks.txt to bed files using pos2bed.

**Q3:** How many peaks were found for each dataset?

**4**. Use annotatePeaks.pl to provide annotation for the PHF8, CTCF and RNF2 peaks from the previous step. Look at the entry for the top/best peak in the file (top row after the header).

**Q4:** Which gene is the top peak nearest to for each dataset?  Is it located in an exon, intron, intergenic, or promoter region? How does it compare between datasets? Study the functions of these genes using <http://www.genecards.org/>.

**5.** Use findMotifsGenome.pl to analyze the peaks (regions.txt) for the enriched DNA motifs found associated with them.

**Q5:** What was the consensus sequence for the top *de novo* motif discovered by the motif finding program? What was the p-value? Which known motif was the best match to this motif? (NOTE: since background sequences are random, the exact results may vary a little bit, particularly the exact p-value.  Also, the motif may appear in the reverse opposite direction, i.e. AACCGG vs. CCGGTT).