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.

**1.** First we need to add the human genome reference to your HOMER

**perl /home/ucsd-train36/software/homer\_new/configureHomer.pl -install hg19**

**2**. 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?

**Answer:**

Command I ran –

*for f in <path>/\*bam; do fname=`basename $f -chr7.bam`; makeTagDirectory <path>tagDir/$fname -genome hg19 -checkGC $f; done*

1. PHF8: averageFragmentGCcontent=0.481
2. H3K27me3: averageFragmentGCcontent=0.449
3. H3K4me3: averageFragmentGCcontent=0.482
4. RNF2: averageFragmentGCcontent=0.394
5. CTCF: averageFragmentGCcontent=0.449
6. Input: averageFragmentGCcontent=0.415

**3**. Next visualize the ChIP-seq experiments by creating bedGraph files from the tag directories and (Using IGV or 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?

Answer:

Command I ran –

*for dir in <path>/tagDir/hES-\*; do makeUCSCfile $dir -o auto; done*

H3K27me3 and RNF2 show high similarity in their binding (reassuring as RNF2 is part of the polycomb complex that is involved in the deposition of H3K27me3). The co-occurance of H3K4me3 on this same region allows us to say these are putative bivalent genes (to make sure they are indeed bivalent, one should do sequential ChIP). Interestingly, even though PHF8 is known to be binding to H3K4me3, it devoid of these regions. This is an observation that we (my group) is studying now.

**4**. Find peaks for the ChIP-seq experiments. RNF2 and PHF8 are chromatin regulators. Would you think should be used -style histone or factor? Try both. 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?

**Answer:**

Command –

*for dir in <path>/tagDir/\*; do findPeaks $dir -i <path>/tagDir/hES-control/ -style factor -o auto; done*

and

*for dir in <path>/tagDir/\*; do findPeaks $dir -i <path>/tagDir/hES-control/ -style histone -o auto; done*

Then:

*mkdir <path>bedfiles*

*for dir in <path>/tagDir/\* ; do dirname=${dir##\*/}; pos2bed.pl $dir/regions.txt > <path>/bedfiles/$dirname-regions.bed; done*

and

*for dir in <path>/tagDir/\* ; do dirname=${dir##\*/}; pos2bed.pl $dir/peaks.txt > <path>/bedfiles/$dirname-peaks.bed; done*

- style factor:

* Control : 0
* CTCF: 3184
* H3K27me3: 461
* H3K4me3: 1305
* PHF8: 1199
* RNF2: 6
* MYC: 964 (in case they also used this one)

-style histone:

* Control : 0
* CTCF: 3604
* H3K27me3: 1212
* H3K4me3: 1289
* PHF8: 1285
* RNF2: 406
* MYC: 4935 (in case they also used this one)

**5**. 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/>.

**Answer:**

*mkdir <path>annotations*

*for dir in ~/1-homework-data/tagDir/\* ; do dirname=${dir##\*/}; annotatePeaks.pl $dir/peaks.txt hg19 > ~/1-homework-data/annotations/$dirname-peaks.txt; done*

and

*for dir in ~/1-homework-data/tagDir/\* ; do dirname=${dir##\*/}; annotatePeaks.pl $dir/regions.txt hg19 > ~/1-homework-data/annotations/$dirname-regions.txt; done*

**PHF8:**

Regions (these make more sense for PHF8 signal structure):

promoter-TSS (NR\_024586) KMT2E-AS1;

Peaks:

promoter-TSS (NM\_001270643) LUC7L2, but the results of this analysis (using -style factor) did not make sense according to the IGV.

**CTCF:**

Peaks:

Intron (NR\_026999, intron 1 of 11); LINC00265, but the results of this analysis (using -style histone) did not make sense according to the IGV

Regions:

Intron (NR\_026999, intron 1 of 11); LINC00265 (same area).

**RNF2:**

Regions:

Intron (NM\_153631, intron 2 of 3); HOXA2

Peaks:

The results of this analysis (using -style factor) did not make sense at all, both according to the IGV and according to the number of peaks identified (6!).

**6.** Use findMotifsGenome.pl to analyze the peaks (peaks or 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).

**Answer:**

Command:

*for dir in <path>/tagDir/\*; do dirname=${dir##\*/}; findMotifsGenome.pl $dir/peaks.txt hg19r <path>/motifs-peaks/motifs-$dirname/ -size 100 -p 10; done*

and

*for dir in <path>/tagDir/\*; do dirname=${dir##\*/}; findMotifsGenome.pl $dir/regions.txt hg19r <path>/motifs-regions/motifs-$dirname/ -size 100 -p 10; done*

Note the difficulty in finding the motifs for histone modifications and chromatin regulators. Since TFs bind specific DNA sequences, their motifs are much more specific.

CTCF, peaks:



Pval: 1e-1358

Name: BORIS(Zf)/K562-CTCFL-ChIP-Seq(GSE32465)/Homer(0.923)

PHF8, regions:



Pval: 1e-4 **(\* note the difference between the CTCF and this one!)**

Name: Bach2(bZIP)/OCILy7-Bach2-ChIP-Seq(GSE44420)/Homer

RNF2 (regions)



Pval: 1e-10 (Note the false positive star)

Name: ZNF528(Zf)/HEK293-ZNF528.GFP-ChIP-Seq(GSE58341)/Homer(0.705)

H3K4me3 (regions):



Pval: 1e-15

Name: CREB1/MA0018.2/Jaspar(0.567)

H3K27me3 (regions):



Pval: 1e-12

Name: NFkB-p65-Rel(RHD)/ThioMac-LPS-Expression(GSE23622)/Homer(0.642)