**Testing if specific histone marks are enriched in true v/s spurious peaks**

Aim: To check if true peaks are enriched for specific histone marks in comparison to spurious peaks.

Marks to test:

1. Active transcription: H3K4me3, H3K27ac
2. Repressive: H3K27me3, H3K9me3
3. Bivalency: H3K4me3 + H3K27me3 (peaks for both must be present)

Histone data:

1. For the histone files, run a multi intersect for all replicates (i.e. datasets for the same cell line and histone mark) and then sub select peaks that are found in ALL of the samples (I think you have already done this part?). Create a folder called “*Histones*” and save these files as “*Histones/{CellLine}\_{Histone mark}\_consensus\_all.bed*”. You should have a total of 5 files in this folder.

Note: For bivalency, run multi-inter with files for replicates of H3K4me3 and H3K27me3 in a given cell line.

ChIP-seq data:

1. First, you will need to access the data. I would suggest copying the files locally or into your own Quest folder. For each cell line and TF, the file corresponding to peaks that are observed across replicates and are accessible (based on ATAC-seq) in the genome.  
     
   Location*:/projects/b1042/AmaralLab/Maalavika/CellLineSpecificNetworks/TFbindingModel/TF\_ChIPfitv2/{CellLine}/ChIP-seq/{TF}\_consensus\_accessible\_peaks.bed*

*projects/b1042/AmaralLab/Maalavika/CellLineSpecificNetworks/TFbindingModel/TF\_ChIPfitv2/A549/ChIP-seq/H3K4me3\_consensus\_accessible\_peaks.bed*

There are files for multiple TFs (around 15). Run the analysis below for each TF in the 5 cell lines. I would prefer automating the code using a for loop for the tasks below rather than individually running them since that script would be way too long and unreadable.

1. To get the accurate edge list from this file you need to follow a couple of steps:
   1. Select peaks that are accessible and present in at least one sample: The column labelled *“./ATAC-seq/consensuspeaks\_all\_targets\_10kb\_sorted.bed*” should be equal to 1 and at least one of the other sample columns should be 1.
   2. Once you select the correct rows – run “[merge](https://daler.github.io/pybedtools/autodocs/pybedtools.bedtool.BedTool.merge.html)” to remove any overlapping edges.
   3. You can also save these files in a separate folder if you like. Suggested name: “*processed\_ChIP/{CellLine}\_{TF}\_accessible\_processed.bed”*
2. Now, run an intersection of this file with each of the marks labelled above (*Histones/{CellLine}\_ {Histone mark}\_consensus\_all.bed* files). Don’t run a multi intersect, it will break each peak into multiple small regions.   
     
   So for each */{CellLine} /ChIP-seq/{TF}\_consensus\_accessible\_peaks.bed* file you will have 5 intersection files (with H3K4me3,H3K27ac, H3K27me3,H3K9me3 and H3K4me3 + H3K27me3) .   
     
   Make sure the intersection is run for the respective cell line (for e.g. /A549/ChIP-seq/CTCF\_consensus\_accessible\_peaks.bed must be intersected with histone marks from the A549 cell line).

Save the output in a separate folder called “intersections” with the following format:   
“*intersections/{CellLine}\_{TF}\_{Histone mark}\_intersection.bed*”. You should have 375 files (5 cell lines x 15 TFs x 5 Histone marks)

Notes:

1. It might help to have 2 separate scripts to keep them readable. You can have one script to process the histone data (point 1) and a separate one to process the TF data (point 2-3) and the intersection (point 4). Or you can make point 4 a separate script of its own.
2. Feel free to use any of the scripts in the *projects/b1042/AmaralLab/Maalavika/CellLineSpecificNetworks/TFbindingModel/TF\_ChIPfitv2/scripts* folder as a reference.
3. Be sure to comment your script so that it is readable, and we can go over it!

Quest Directory: /projects/b1042/AmaralLab/Rohan/

Local Directory: /Users/rohankrishnamurthi/Downloads/Maalavika\_Pillai\_Research/

ORGANIZATION:

Folders

* histone\_data 🡪 raw histone data for each cell line and histone mark, as bed files (Part 1)
* histone\_processed\_data 🡪 results of histone\_overlap\_1.1.ipynb, the multi intersect of raw histone data (Part 1)
  + Relevant output files end in “*consensus\_all.bed”*
* tf\_ChIP\_data 🡪 raw transcription factor ChIP data organized by cell line, from Maalavika’s folder, as “consensus\_accessible\_peaks.bed” (Part 2)
* tf\_ChIP\_data/filtered\_file 🡪 filtered TF data as now bed files, following selection of peaks and merging (Part 3)
* tf\_ChIP\_data/processed\_data 🡪 filtered TF data as bed files, following selection of peaks and merging (Part 3)
  + End in “accessible\_processed.bed”
* intersections 🡪 results of intersection of TF and histone files for each cell line (Part 4)

Python Scripts

* histone\_overlap\_1.1.ipynb 🡪 multi intersect for histone data from replicate experiments, including bivalency case (Part 1)
* tf\_data\_1.1.ipynb 🡪 selection of peaks and merging of raw TF files (Part 3)
* intersection\_1.1.ipynb 🡪 intersection of raw TF files with processed histone files, as bed files (Part 4)

Other stuff regarding packages

* brew install Bedtools
  + Version: bedtools v2.31.1
* Which Bedtools
  + /opt/homebrew/bin/Bedtools
* nano ~/.bashrc
* export PATH=$PATH:/path/to/Bedtools
* source ~/.bashrc
* D