Kiley Huffman
QBIO-478
Spring 2025

QBIO 478 HW #5

1. (10 pts) Biotech assays often exploit naturally occurring enzymes. In the "ChIPseq-1" lecture we discussed numerous assays. List three enzymes from this lecture. In addition, for each enzyme state which assay it is used in and what it does in the assay.

Enzymes from the "ChIPseq-1" Lecture:

1) ***Tn5 Transposase***

   Assay Used In: ATAC-seq (Assay for Transposase-Accessible Chromatin)

   What it does: Tn5 is normally inactive in wild-type form, but a hyperactive mutant is used in ATAC-seq. In ATAC-seq, this mutant form of Tn5 simultaneously fragments and tags open chromatin regions with sequencing adapters, allowing for efficient mapping of accessible regions in the genome.

2) ***DNase I***

   Assay Used In: DNase-seq

   What it does: DNase I is an endonuclease that cuts accessible, non-nucleosome DNA. In DNase-seq, it is used to identify DNase hypersensitive sites, which correspond to regulatory regions such as promoters and enhancers.

3) ***Micrococcal Nuclease (MNase)***

   Assay Used In: MNase-seq

   What it does: MNase digests unprotected linker DNA between nucleosomes. In MNase-seq, it is used to map nucleosome positions by isolating DNA fragments protected by histones.

2. (10 pts) In the MACS algorithm for analyzing ChIP-seq experiments, in the second step, we model the shift size parameter "d." Assume we do single-end sequencing.

a. **Imagine two ChIP-seq experiments. Everything is identical in the two experiments, except we do the fragmentation differently. In the first experiment the DNA fragments are ~250 bp while in the second experiment the DNA fragments are ~500 bp. Let d_1 be the shift size parameter for the first experiment and let d_2 be the shift size parameter for the second experiment. Do you expect d_1 will be greater than d_2, d_1 will be less than d_2, d_1 will be approximately equal to d_2, or is there not enough information to know? Explain your answer.**

   250 bp fragments → $d_1 \approx 125$ bp

   500 bp fragments → $d_2 \approx 250$ bp

   Therefore, $d_1 \approx 125$ bp $< d_2 \approx 250$ bp.

   I expect that d_1 will be less than d_2. This is because the shift size parameter "d" models the distance between the forward and reverse strand peaks, which is approximately half of the average DNA fragment length. Thus, if the fragment size increases from ~250 bp to ~500 bp, the average shift size should also increase. Therefore, $d_1 < d_2$.

b. **Again imagine two ChIP-seq experiments. Everything is identical in the two experiments, except now we sequence different lengths. Unlike in part (a) the fragmentation is the same, for both experiments in part (b) the DNA fragments are ~500 bp. In the first experiment we sequence 100 bp of every DNA fragment while in the second experiment we sequence 200 bp of every fragment. Let d_1 be the shift size parameter for the first experiment and let d_2 be the shift size parameter for the second experiment. Do you expect d_1 will be greater than d_2, d_1 will be less than d_2, d_1 will be approximately equal to d_2, or is there not enough information to know? Explain your answer.**

   500 bp fragments → $d_1 \approx 250$ bp

   500 bp fragments → $d_2 \approx 250$ bp

   Therefore, $d_1 \approx 250$ bp $\approx$ $d_2 \approx 250$ bp

   I expect that d_1 will be approximately equal to d_2. This is because the shift size "d" is based on the fragment length, not the read length. Since both experiments have the same DNA fragment size (~500 bp), the expected distance between the start of forward and reverse reads will be roughly the same, regardless of the read length. Since read length does not significantly affect d, I expect $d_1 \approx d_2$.

3. (10 pts) In the MACS algorithm, in the third step, for each candidate peak we do the following Poisson test. For each candidate peak, we estimate a parameter $\lambda_{!"\#\$!}$ and calculate the probability that a Poisson random variable with parameter $\lambda_{!"\#\$!}$ will have as many reads or more reads than was observed. If this probability (the p-value) is less than 10^{-5} then the candidate peak makes the final list.

In this problem we will use R. If you do not have R on your computer, you can use R online at the following website: www.programiz.com/r/online-compiler/ The following R code will compute the p-value for the Poisson test for lam equal to $\lambda_{!"\#\$!}$ and k equal to the number of reads observed at the candidate peak. If you are using R online, type the code below in the left column, then press the blue "Run" box, and the answer will appear in the right column:

```
lam = 10

k = 27

x = 0:(k-1)

1 - sum(exp(-lam)*(lam^x)/factorial(x))
```

You should find that the p-value is 6.42×10^{-6}. If in the code above you decrease to k = 26 you should find the p-value is now 1.77×10^{-5}. So for $\lambda_{!"\#\$!} = 10$ the threshold for the number of observed reads at a candidate peak is 27, since fewer reads than that have a p-value > 10^{-5} and more reads than that have a p-value < 10^{-5}.

I want you to find the threshold number of reads for the case when $\lambda_{!"\#\$!} = 30$. So first change lam = 30, and then by trial-and-error try different k values until you find the value such that the p-value is less than 10^{-5} but if k is one less the p-value is greater than 10^{-5}:

Using the method described above, I kept increasing k until the result dropped below 10^{-5}. I found that:

- For k = 49, p-value is 1.22e-5 > 10^{-5}
- For k = 50, p-value is 6.45e-6 < 10^{-5}

Thus, the threshold number of reads when $\lambda = 30$ is k = 50. This is because when k = 50, the p-value is less than 10^{-5}, but if k is one less (k = 50-1 = 49), the p-value is greater than 10^{-5}.

4. (10 pts) The MACS algorithm was developed for analyzing ChIP-seq experiments, but it can be used to analyze ATAC-seq experiments too. List the four main steps of the MACS algorithm, and for each step discuss any modifications required when analyzing ATAC- seq experiments.

The MACS algorithm has four main steps:

1. ***Tag Shifting (Shift Size "d")***

   Modifications Required: When analyzing ATAC- seq experiments, you should skip the tag shifting or set it to 0 because ATAC-seq already captures the actual cut sites via the Tn5 transposase.

2. ***Background Modeling***

   Modifications Required: When analyzing ATAC- seq experiments, you should use the "--nolambda" option or rely solely on local background estimation because ATAC-seq experiments often don't include an input/control sample. As a result, the MACS's dynamic background modeling may not work well.

3. ***Peak Width and Shape***

   Modifications Required: When analyzing ATAC- seq experiments, you should adjust peak width parameters because ATAC-seq peaks are often wider and less sharp than ChIP-seq peaks, especially in regions like promoters or enhancers with open chromatin.

4. **Filtering and Thresholds**

   Modifications Required: When analyzing ATAC- seq experiments, you should consider relaxing p-value or q-value thresholds because ATAC-seq peaks may have lower signal-to-noise than strong TF ChIP-seq peaks. Thus, stricter thresholds might discard real open chromatin sites.

5. (10 pts)

a. First do an internet search on ACE2. Briefly report what you find:

I found that ACE2 stands for Angiotensin-Converting Enzyme 2, and that it is a membrane-bound enzyme. It plays a key role in the renin-angiotensin system (RAS) by converting angiotensin II to angiotensin-(1–7), which helps regulate blood pressure. I also learned that ACE2 is the entry receptor for SARS-CoV-2, the virus that causes COVID-19; the virus binds to ACE2 on the surface of human cells to gain entry. Moreover, ACE2 is expressed in many tissues, including the lungs, small intestine, heart, kidneys, and blood vessels.

b. Next, do an internet search on H3K4me3. Briefly report what you find:

I found that H3K4me3 stands for Histone 3 Lysine 4 Trimethylation, which is a histone modification. H3K4me3 has three methyl groups added to the 4th lysine residue on histone H3, and is strongly associated with active promoters. It is often found near the transcription start sites (TSS) of actively transcribed genes. Most significantly, it serves as an epigenetic marker of gene activation; it also plays a key role in chromatin remodeling and regulation of gene expression.

6. (10 pts) Go to the ENCODE website: https://www.encodeproject.org/ Enter ACE2 in the window near the top. Click on "ChIP-seq experiments" near the bottom. Click on "Homo sapiens," "Histone," and "tissue." Then click on "H3K4me3" and "small intestine" (note if you don't see "small intestine" you can scroll left and right to see more tissues).

a. You should see 3 experiments. For each experiment, list the age and gender for the sample.

1. *Experiment ENCSR792IJA*

   Age, Gender: 34 years, Male

2. *Experiment ENCSR944QSH*

   Age, Gender: 30 years, Female

3. *Experiment ENCSR237QFJ*

   Age, Gender: Embryo (108 days), Male

b. Click on "Visualize" (if a window pops up giving options for visualize, always select "GRCh38" and "UCSC"). You should see 6 tracks, and above the tracks you should see blocks for peaks. If you double-click on the blocks, they should expand. Click on Peak_11425 (for me it's colored dark black, and longer than the other peaks, the number 11425 is on the left side of the window). Which sample is this peak from? What is the p-value (they report -log10 of the p-value, so if they report x, the p-value is 10^{-x})?

Which sample is this peak from?

This peak is from the male embryo (108 days) sample in Experiment ENCSR237QFJ. Here is more information about where this sample is from:

- Male embryo (108 days)
- Biosample: ENCBS054KUO
- Experiment: ENCSR237QFJ
- Title: "Homo sapiens small intestine tissue male embryo (108 days)"

What is the p-value?

They report the p-value as 121.356 in -log10. Thus, the p value is:

$$p = 10^{-131.356} = 2.8 \times 10^{-122}$$

c. Click on Peak_51891 (this is near the same genomic location as Peak_11425 but much narrower). Which sample is this peak from? What is the p-value?

Which sample is this peak from?

This peak is from the 34-year old male sample in Experiment ENCSR792IJA. Here is more information about the sample:

- 34-year old Male
- Biosample: ENCBS853LFM
- Experiment:ENCSR792IJA
- Title: "Homo sapiens small intestine tissue male adult (34 years) ENCBS853LFM"

What is the p-value?

They report the p-value as 3.069 in -log10. Thus, the p value is:

$$p = 10^{-3.069} = 8.54 \times 10^{-4}$$

d. Go back to ENCODE. Now select "H3K4me3" and "lung". How many experiments? Click on "Visualize". Double-click on the blocks to expand the peaks. How many of the lung samples have peaks that overlap Peak_11425?

How many experiments: 4

Number of Lung Samples that have peaks that overlap Peak_11425: 4

7. (10 pts) Go back to the first ENCODE website page: https://www.encodeproject.org/ In the window near the top enter whatever gene, protein, genomic region, etc. that interests you. Click on "Reference epigenome" near the bottom. You can now select ATAC-seq or many other assays (including DNase-seq). Note: you can also click the arrow to reveal more tissues. Explore. Report on something interesting you find with ATAC-seq.

While exploring the ENCODE database, I focused on the FOXP3 gene, known for its critical role in the development and function of regulatory T cells. Using the "Reference Epigenome" section and selecting ATAC-seq data, I examined chromatin accessibility across various human tissues. One interesting thing I learned was that there is high chromatin accessibility at the FOXP3 promoter region in CD4-positive, alpha-beta regulatory T cells, as indicated by prominent ATAC-seq peaks. This suggests that the FOXP3 gene is actively transcribed in these cells. I also saw that other cell types, such as CD8-positive T cells or B cells, showed minimal ATAC-seq signal at the FOXP3 promoter, indicating lower accessibility.