ChIP-seq & Motif Finding

Quantitative Biology 2022
10/7/22

Doctoral Program in
Cell, Molecular, Developmental
Biology, and Biophysics

JOHNS HOPKINS
UNIVERSITY

Transcription Regulation

- What are some ways that transcription is regulated?

- Transcription factor binding
- Histone modification
- DNA methylation
- Chromatin folding
- Histone positioning

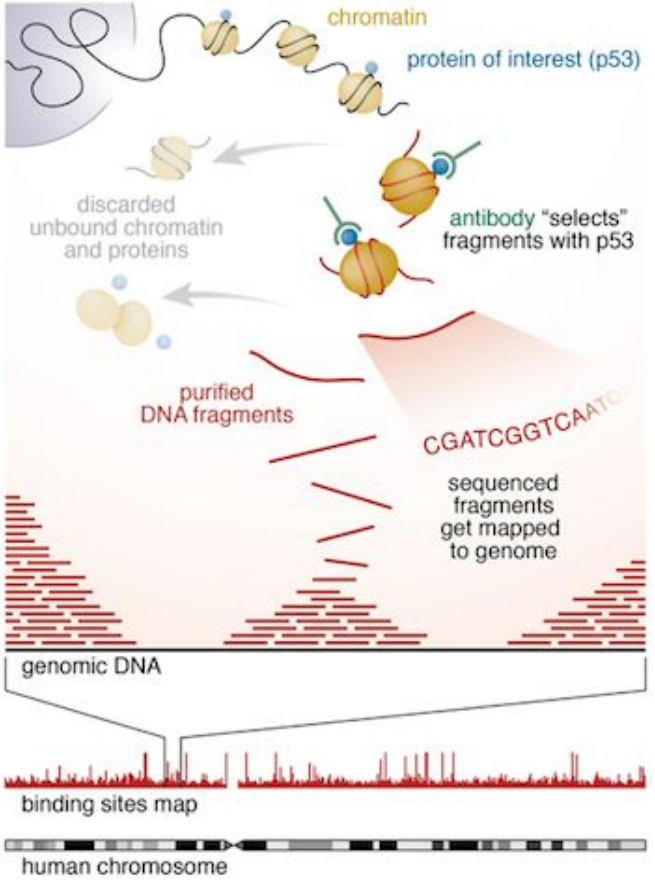
Strategies for assaying regulatory signals

- Antibody-based targeting of specific protein or chemical modification

- Detection of chromatin accessibility

Immunoprecipitation (IP) based assays

- Fragment DNA
- Bind target with antibodies
- Immunoprecipitate bound fragments
- Sequence purified fragments
- Map fragments to genome
- Find significant regions of enrichment

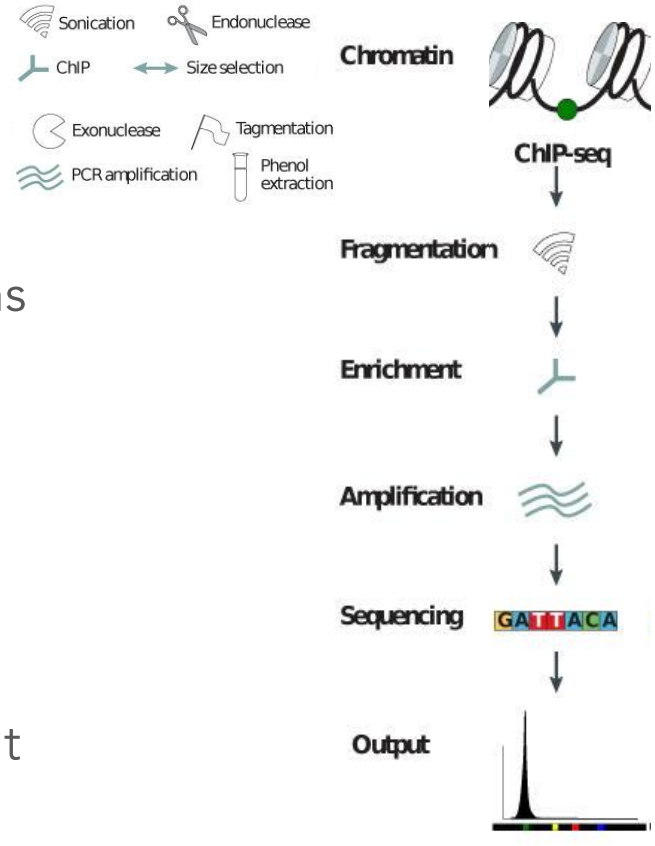


https://hbctraining.github.io/Intro-to-ChIPseq/lessons/01_Intro_chipseq_data_organization.html

ChIP-seq

Advantages

- Assaying a known target, highly specific
- Easy to perform
- Can be used to detect chemical modifications
    - histone modifications
    - DNA methylation



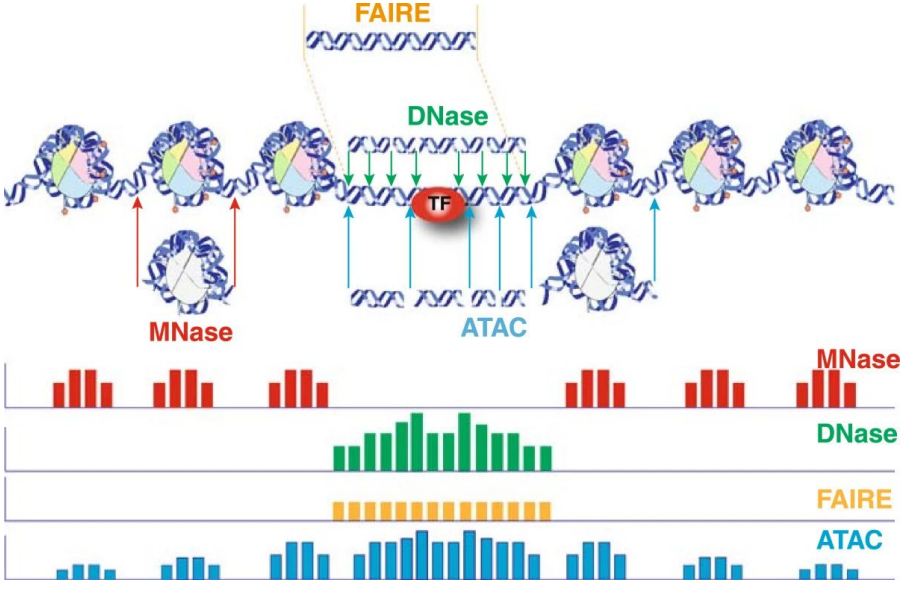
Disadvantages

- Needs a good antibody to exist
- Can only assay a single target per experiment
- Broad signal
- Requires input (although many people do not use one)
- Strand ambigous

Meyer, C. A., & Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews. Genetics*, *15*(11), 709–721.

Accessibility-based assays

- MNase-seq
  - Indirect assay of chromatin accessibility and nucleosome positioning
- DNAse-seq
  - Direct assay of chromatin accessibility
- FAIRE-seq (formaldehyde-assisted isolation of regulatory elements)
  - Maps open chromatin
- ATAC-seq (assay for transposase-accessible chromatin)
  - Maps open chromatin, transcription factor binding, and nucleosome occupancy



Tsompana, M., Buck, M.J. Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin* 7, 33 (2014).

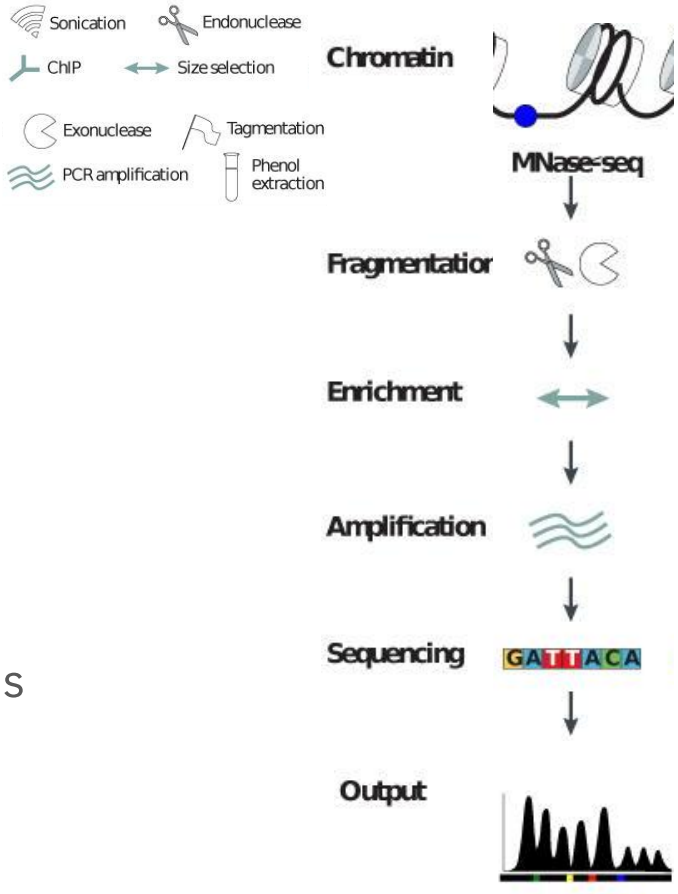
MNase-seq

Advantages

- Assays nucleosome positioning
- Can infer TF binding

Disadvantages

- Difficult to perform
- Requires a large number of cells (1-10M)
- Requires a large number of sequenced reads (150-200M)



Meyer, C. A., & Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews. Genetics*, *15*(11), 709–721.

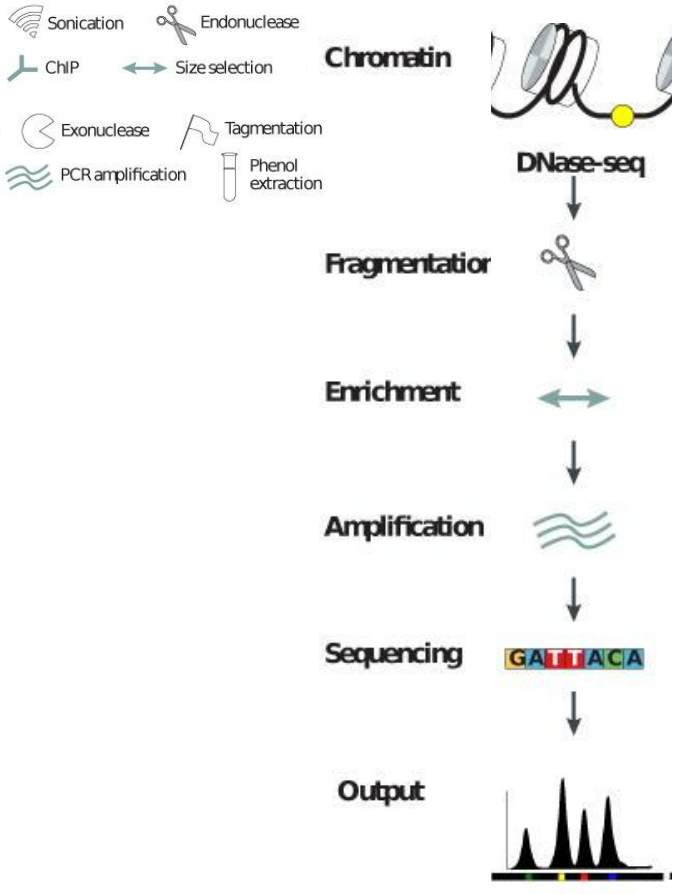
DNase-seq

Advantages

- Easier to perform
- General TF binding detection
- Requires fewer sequenced reads (20-50M)

Disadvantages

- Requires a large number of cells (1-10M)
- Cutting bias can give false signal



Meyer, C. A., & Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews. Genetics*, *15*(11), 709–721.

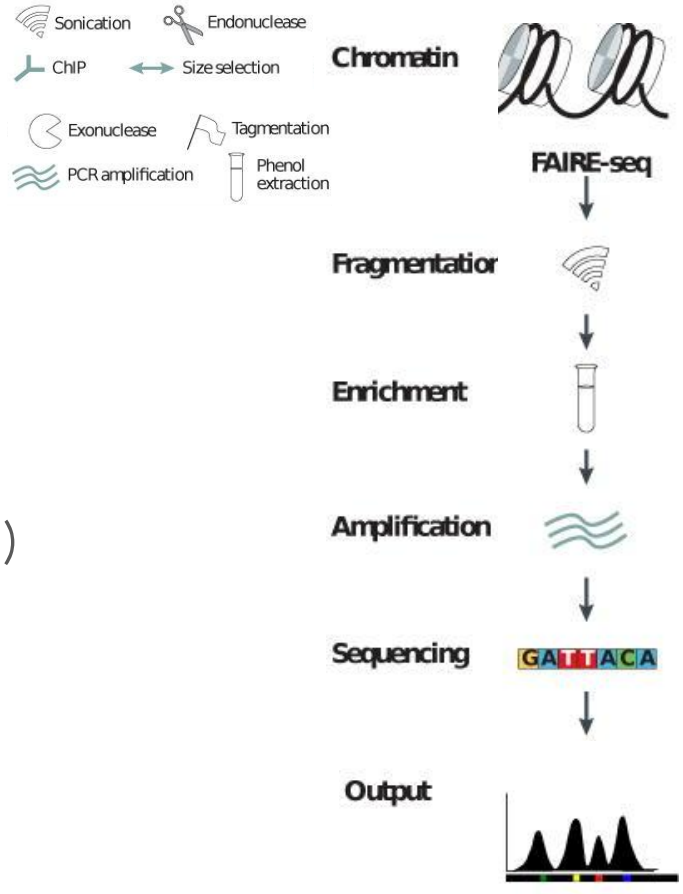
FAIRE-seq

Advantages

- Very easy to perform
- Requires fewer sequenced reads (20-50M)

Disadvantages

- Requires a large number of cells (100K-10M)
- Has a low signal to noise ratio
- Is sensitive to the fixation efficiency



Meyer, C. A., & Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews. Genetics*, *15*(11), 709–721.

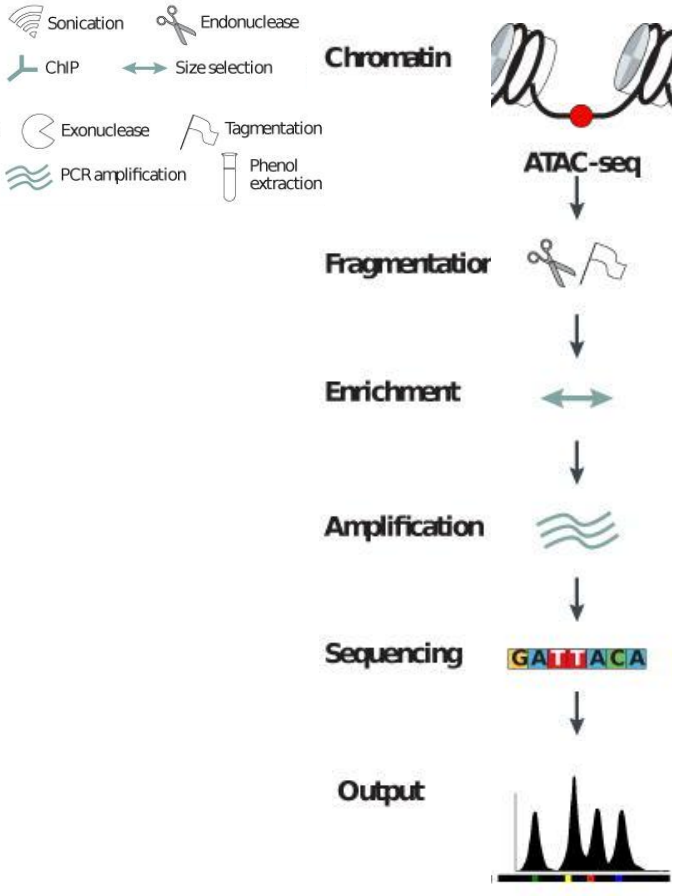
ATAC-seq

Advantages

- Easier to perform
- Requires few cells (500-50K)
- Maps open chromatin, TF binding, and nucleosome occupancy

Disadvantages

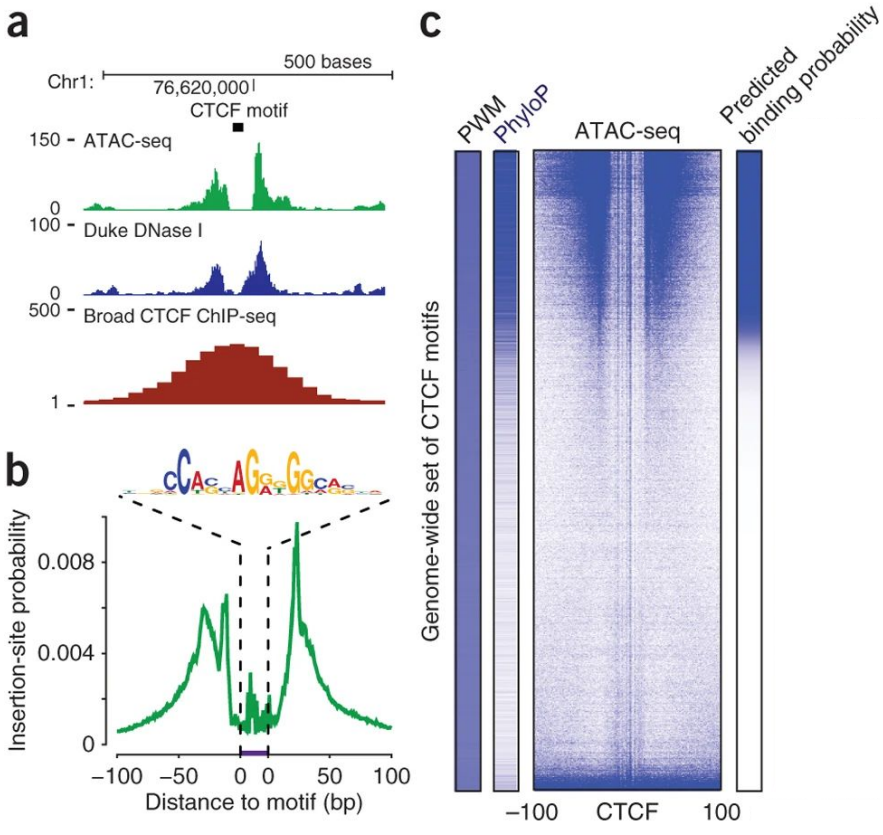
- Requires many sequenced reads (60-100M)
- Has issues with mitochondrial DNA contamination



Meyer, C. A., & Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews. Genetics*, *15*(11), 709–721.

ATAC-seq data

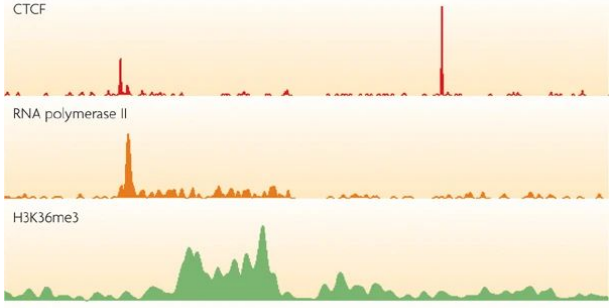
- Transcription factor shape-specific peaks
- Strand-specific orientation of peak shape
- Local restriction of nucleosome positioning
- Narrow footprint allows direct inference of binding sequence



Buenrostro, J., Giresi, P., Zaba, L. *et al.* Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10, 1213–1218 (2013).

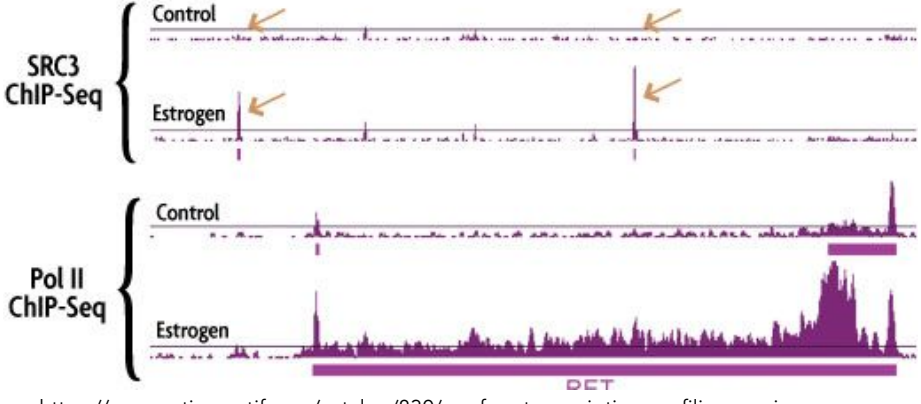
ChIP-seq data

- Broader peaks (10s-100s bp wide)
- TF vs. histone modification peaks

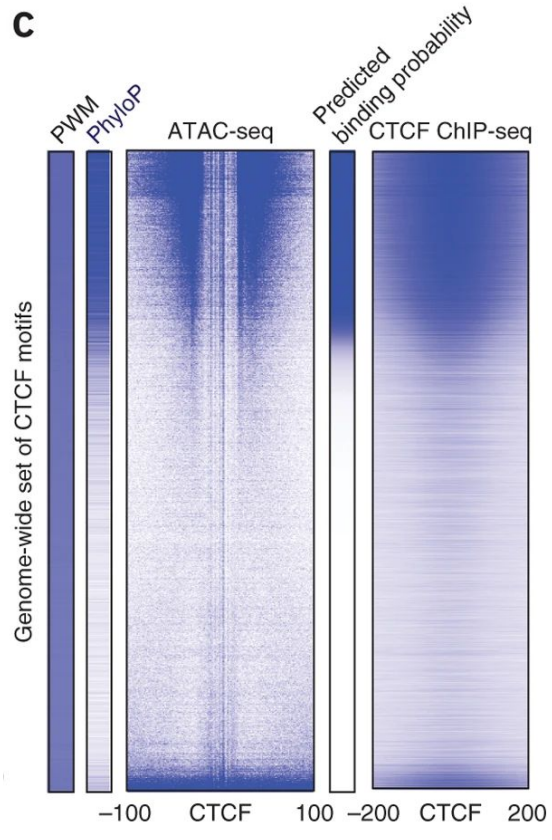


Park, P. ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 10, 669–680 (2009).

- Why does ChIP-seq need controls?

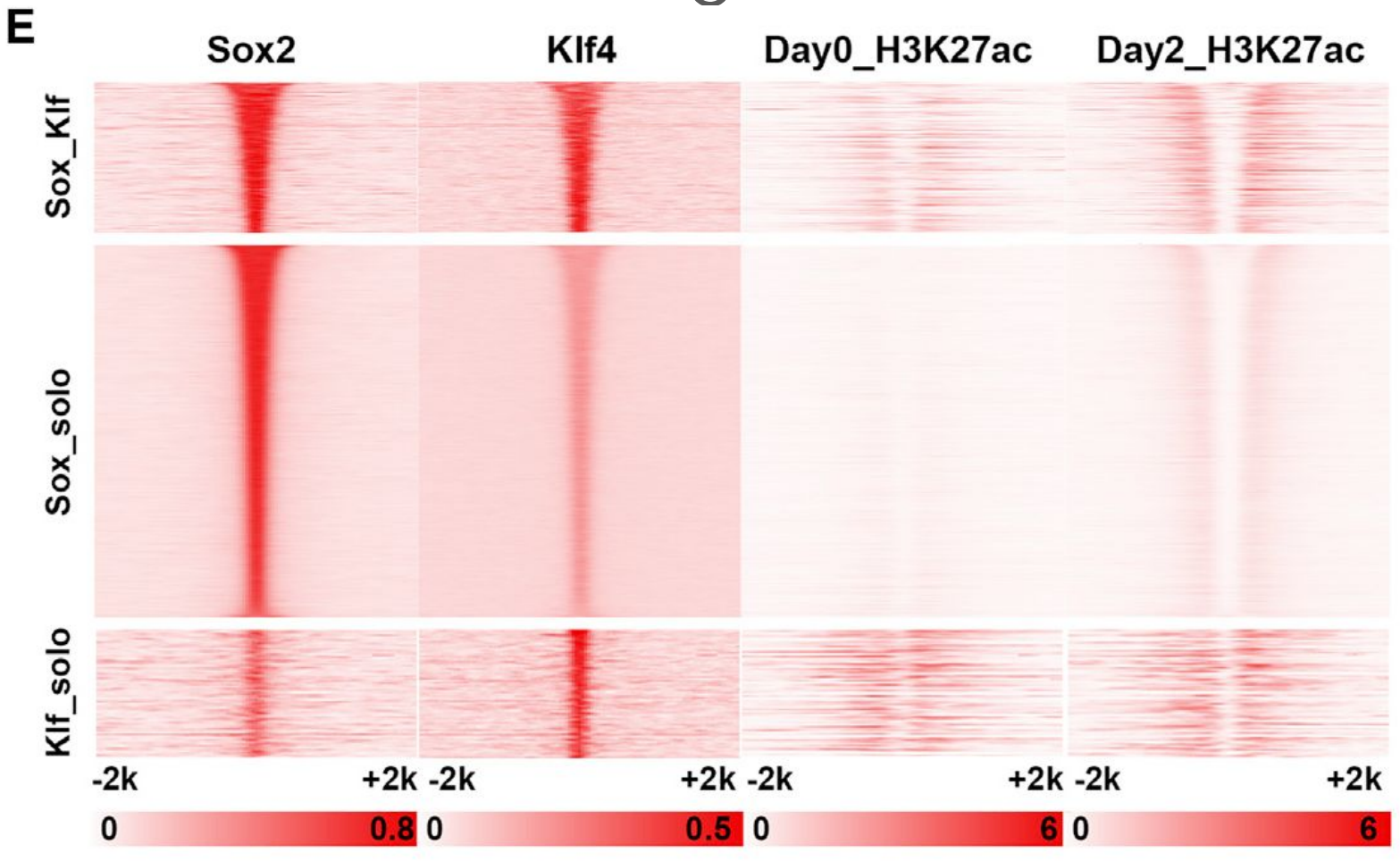


https://www.activemotif.com/catalog/830/rna-free-transcription-profiling-services



Buenrostro, J., Giresi, P., Zaba, L. *et al*. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10, 1213–1218 (2013).

Using ChIP-seq to see relationships



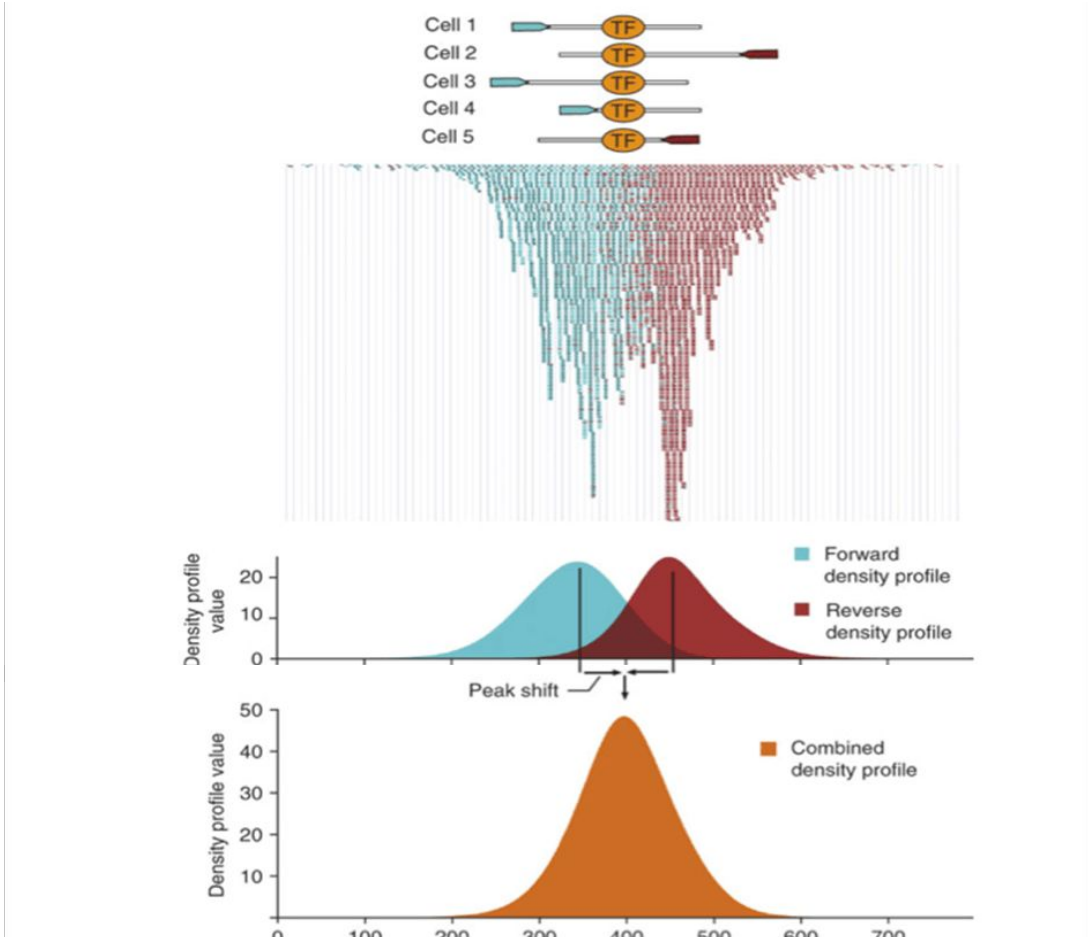
An Z, Liu P, Zheng J, Si C et al. Sox2 and Klf4 as the Functional Core in Pluripotency Induction without Exogenous Oct4. Cell Rep 2019 Nov 12;29(7):1986-2000.e8.

MACS (model-based analysis of ChIP-seq)

- In order to analyze ChIP-seq data, it's crucial to identify significant peaks

- MACS accomplishes this in a 2-part process
  1. Model the sequenced fragment size and shift reads accordingly (single-end only)
  2. With a sliding window, estimate the probability of the level of enrichment within each window

MACS - fragment size modeling

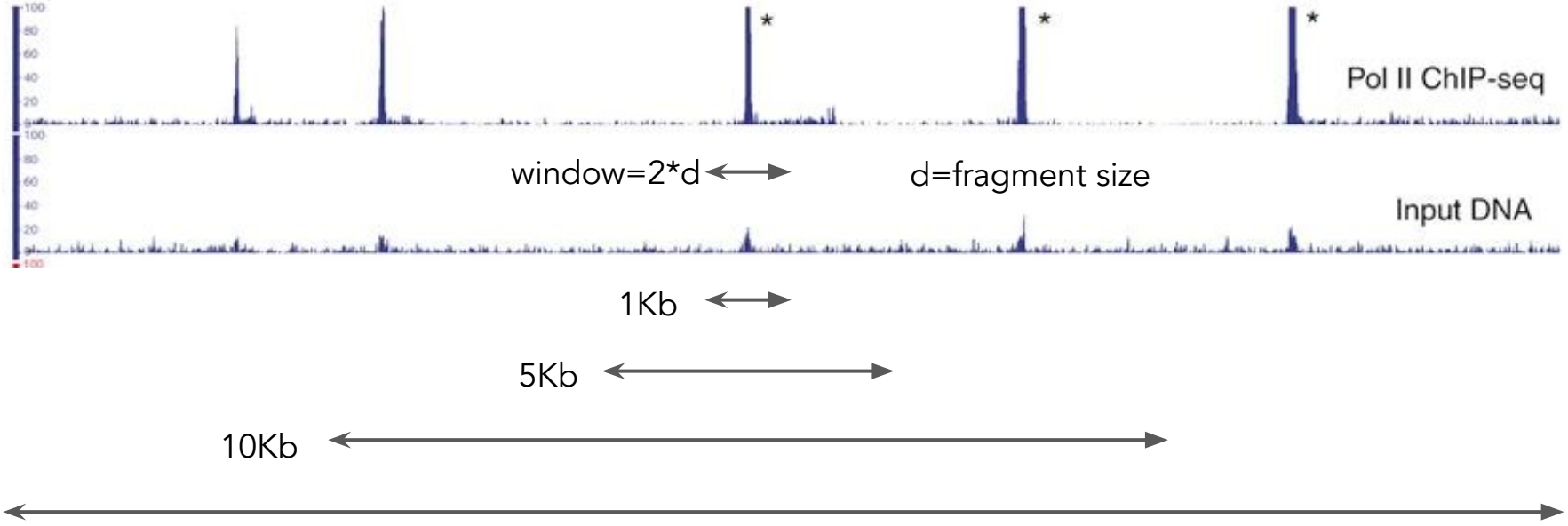


https://www.rasalsi.com/courses/next-generation-sequencing/new-developments-in-ngs-data-analysis-chip-seq/

MACS - finding peak significance

- What should we expect if there is no biological signal (i.e. the control)?
  - Reads randomly spread through the genome
- What probability distribution would you expect the read density to conform to?
  - Poisson, of course!
- Can we compare the treatment and control samples if they have different sequencing depth?
  - No, not without up- or downsampling to equalize them
- Do you think the chances of finding reads are equal across the whole genome? Why or why not?
  - No. Differential accessibility, repeat elements

MACS - finding peak significance



$$\lambda = \max(\lambda_{1K}, \lambda_{5K}, \lambda_{10K}, \lambda_{wg})$$

Why do we need to consider multiple lambdas?

Motif Finding

Transcription factor target binding sequence

- How might we go about finding the sequence preference of a DNA-binding protein?
  - Collect many DNA fragments that we know the protein likely binds to and look for a common sequence
- Can a DNA-binding protein only bind to a single sequence?
  - No, the binding sequence is degenerate, or flexible. More so at some positions than others
- How can we find an enriched sequence if it's not the same every time?
  - Motif-finding algorithms like Meme

The Position Weight Matrix (PWM)

	A	C	T	G
Position 1	0.0000	1.0000	0.0000	0.0000
Position 2	0.0149	0.9851	0.0000	0.0000
Position 3	0.5223	0.4776	0.0000	0.0000
Position 4	0.0000	1.0000	0.0000	0.0000
Position 5	0.4029	0.0000	0.1642	0.4328
Position 6	0.0000	1.0000	0.0000	0.0000
Position 7	0.0000	1.0000	0.0000	0.0000

Meme - Multiple Expectation Maximization Enumerator

Meme uses two rounds of expectation maximization to first determine a good starting PWM and then to learn the best weights for the matrix

But what is expectation maximization?

Problem:
Given a set of observations X, some latent variable Z, and unknown parameters Θ along with a likelihood function, find the optimal values of Θ

Expectation Maximization

Problem:
Given a set of observations X, some latent variable Z, and unknown parameters $\boldsymbol{\theta}$ along with a likelihood function, find the optimal values of $\boldsymbol{\theta}$

$$L(\boldsymbol{\theta}; \mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\theta}) = \int p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})\, d\mathbf{Z} = \int p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}) p(\mathbf{Z} \mid \boldsymbol{\theta})\, d\mathbf{Z}$$

Because we don't know Z, we can represent it as a probability distribution across potential value. Thus, we can get the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ by finding the marginal likelihood of the data

Unfortunately, this is rarely doable as Z is unknown as is its distribution

Expectation Maximization

Solution:

Break the problem into two parts and iterate between them

1. Find the distribution of Z
2. Find the MLE of $\boldsymbol{\theta}$

Expectation step:

Define $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ as the expected value of the log likelihood function, where $\boldsymbol{\theta}^{(t)}$ is the estimate of parameters at the current step.

This is the weighted average of the log likelihood across Z's distribution.

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \underbrace{\mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}}[\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})]$$

This is value we are estimating in this step, the distribution of Z values

Expectation Maximization

Solution:

Break the problem into two parts and iterate between them

1. Find the distribution of Z
2. Find the MLE of $\boldsymbol{\theta}$

Maximization step:

Find the value of $\boldsymbol{\theta}$ that maximizes the log likelihood function. This uses the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$.

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}\left[\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})\right]$$

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$$

This is our MLE function to get an updated estimate of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(t+1)}$

Motif finding EM

In the motif-finding problem the values correspond to the following:

X = The set of sequences corresponding to ChIP-seq peaks

Z = The starting position of the motif in each sequence

$\boldsymbol{\theta}$ = The position weight matrix of our motif

Some assumptions:

- Each sequence contains exactly one occurrence of the motif
- The length of the motif is specified and fixed
- The probability of each base appearing in a given position is equal

Motif finding EM

Starting conditions - randomly assign values to the PWM for motif of width k

Step 1: Expectation

For each sequence, find the probability of each k-mer in the sequence using the PWM and then scale values to sum to 1.

This is the probability distribution of Z for that sequence

Step 2: Maximization

For each sequence, sum to the occurrences of bases in each position of each k-mer, weighted by that sequence's Z. Convert counts into probabilities for each position in the PWM

Motif finding EM

Step 1: Expectation

Current PWM

A	C	G	T
0.7	0.1	0.1	0.1
0.1	0.6	0.2	0.1
0.2	0.1	0.2	0.5
0.5	0.1	0.1	0.3

Current sequence

A	C	G	G	T	A	G	T

Motif finding EM

Step 1: Expectation

A	0.7	0.1	0.2	0.5
C	0.1	0.6	0.1	0.1
G	0.1	0.2	0.2	0.1
T	0.1	0.1	0.5	0.3

	A	C	G	G	T	A	G	T

Z Distribution

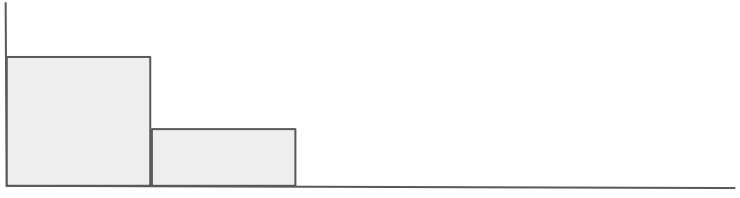
Motif finding EM

Step 1: Expectation

A	0.7	0.1	0.2	0.5
C	0.1	0.6	0.1	0.1
G	0.1	0.2	0.2	0.1
T	0.1	0.1	0.5	0.3

A	C	G	G	T	A	G	T

Z Distribution

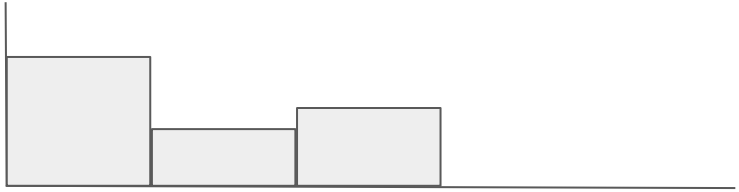


Motif finding EM

Step 1: Expectation

	A	0.7	0.1	0.2	0.5		
	C	0.1	0.6	0.1	0.1		
	G	0.1	0.2	0.2	0.1		
	T	0.1	0.1	0.5	0.3		
A	C	G	G	T	A	G	T

Z Distribution

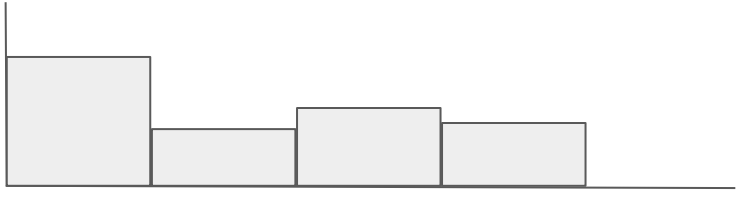


Motif finding EM

Step 1: Expectation

		A	0.7	0.1	0.2	0.5	
		C	0.1	0.6	0.1	0.1	
		G	0.1	0.2	0.2	0.1	
		T	0.1	0.1	0.5	0.3	
A	C	G	G	T	A	G	T

Z Distribution



Motif finding EM

Step 1: Expectation

				A	0.7	0.1	0.2	0.5
				C	0.1	0.6	0.1	0.1
				G	0.1	0.2	0.2	0.1
				T	0.1	0.1	0.5	0.3
A	C	G	G	T	A	G	T	
0.35	0.1	0.15	0.1	0.3				

Z Distribution

Motif finding EM

Step 2: Maximization

A	C	G	G	T	A	G	T

Uninitialized PWM

A	0.0	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0
G	0.0	0.0	0.0	0.0
T	0.0	0.0	0.0	0.0

Z

0.35
0.1
0.15
0.1
0.3

Motif finding EM

Step 2: Maximization

A	C	G	G	T	A	G	T

	A	C	G	G
A	0.35	0.0	0.0	0.0
C	0.0	0.35	0.0	0.0
G	0.0	0.0	0.35	0.35
T	0.0	0.0	0.0	0.0

0.35
0.1
0.15
0.1
0.3

Motif finding EM

Step 2: Maximization

A	C	G	G	T	A	G	T

	C	G	G	T
A	0.35	0.0	0.0	0.0
C	0.1	0.45	0.1	0.0
G	0.0	0.0	0.35	0.35
T	0.0	0.0	0.0	0.1

0.35
0.1
0.15
0.1
0.3

Motif finding EM

Step 2: Maximization

A	C	G	G	T	A	G	T

	G	G	T	A
A	0.35	0.0	0.0	0.15
C	0.1	0.45	0.25	0.0
G	0.15	0.15	0.35	0.35
T	0.0	0.0	0.0	0.1

0.35
0.1
0.15
0.1
0.3

Motif finding EM

Step 2: Maximization

A	C	G	G	T	A	G	T

0.35
0.1
0.15
0.1
0.3

	G	T	A	G
A	0.35	0.0	0.1	0.15
C	0.1	0.45	0.25	0.0
G	0.25	0.15	0.35	0.45
T	0.0	0.1	0.0	0.1

Motif finding EM

Step 2: Maximization

A	C	G	G	T	A	G	T

0.35
0.1
0.15
0.1
0.3

	T	A	G	T
A	0.35	0.3	0.1	0.15
C	0.1	0.45	0.25	0.0
G	0.25	0.15	0.65	0.45
T	0.3	0.1	0.0	0.4

Motif finding EM

Step 2: Maximization

A	C	G	G	T	A	G	T

0.35
0.1
0.15
0.1
0.3

Our updated PWM

A	0.35	0.3	0.1	0.15
C	0.1	0.45	0.25	0.0
G	0.25	0.15	0.65	0.45
T	0.3	0.1	0.0	0.4

MEME

Meme uses two sets of EM, one to initialize the PWM, and one to find the starting points and PWM.

Phase 1:

1. Using each possible k-mer in the sequences as the initial PWM, find the Z distributions
2. Calculate the weighted sum of log likelihoods given the PWM and Z for each k-mer
3. Select the k-mer with the highest weighted log likelihood

Phase 2:

1. Perform the standard motif EM to optimize the PWM