3000788 Intro to Comp Molec Biol

Lecture 9: ChIP-seq and DNA motif discovery

Fall 2025

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today’s agenda

- Epigenetics
- Chromatin immunoprecipitation technique
- Analysis of ChIP-seq data
- DNA motif discovery

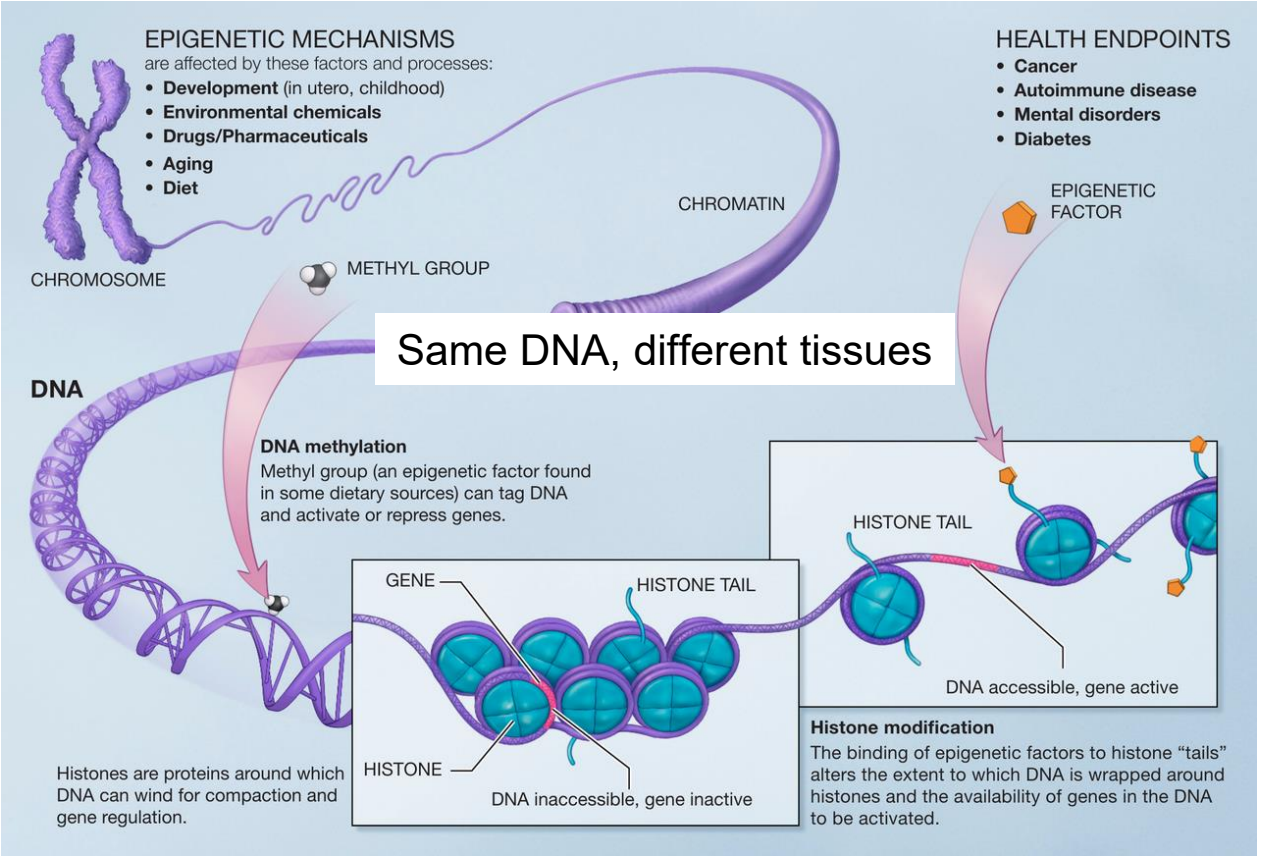
Epigenetic mechanisms

Epigenetics

- Regulation of gene expression **without changing the DNA sequence**

Major mechanisms

- DNA methylation
- Chromatin accessibility
- Histone modification



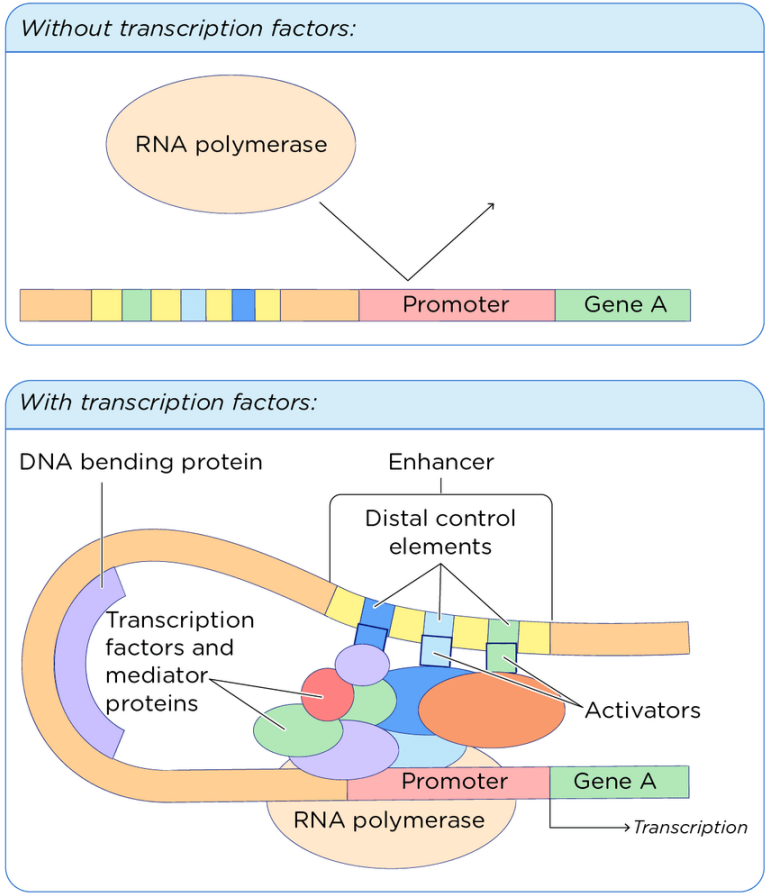
https://en.wikipedia.org/wiki/Epigenetics

Transcription factor (TF)

- TFs are proteins that binds to DNA segments called enhancer, repressor, or promoters

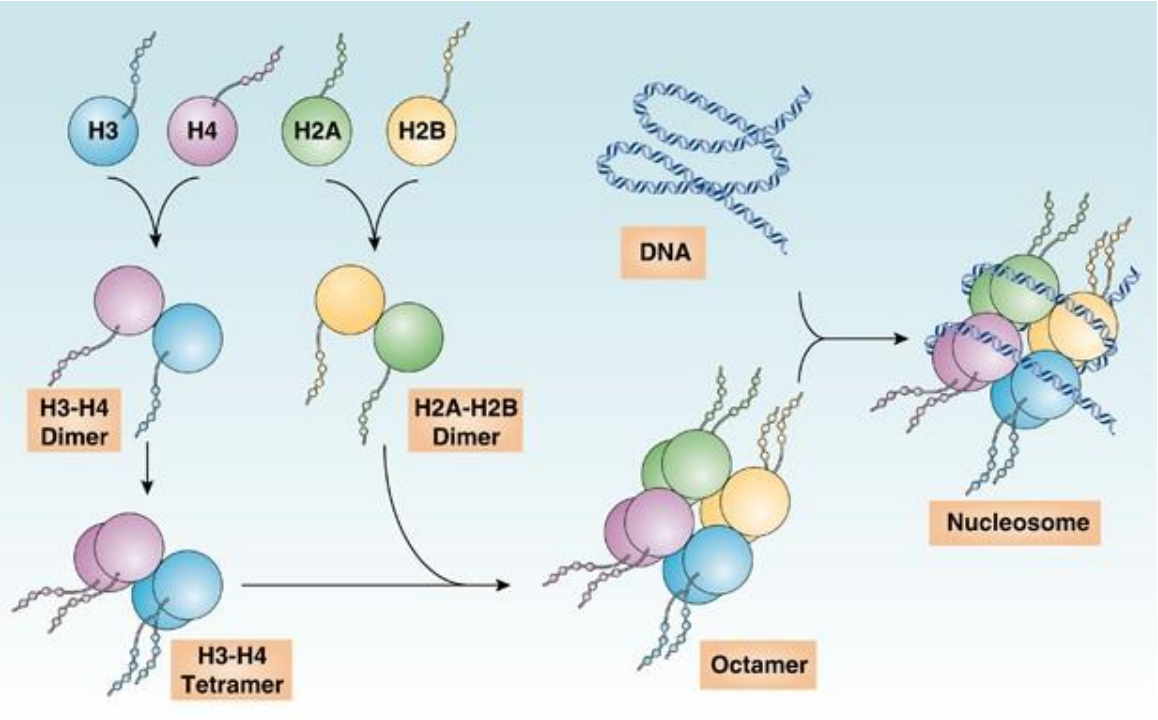
- [**Activator, Promoter**] TFs recruit and stabilize RNA polymerase

- [**Repressor**] TF-bound repressor blocks RNA polymerase from the promoter



https://jackwestin.com/resources/mcat-content/control-of-gene-expression-in-eukaryotes/dna-binding-proteins-transcription-factors

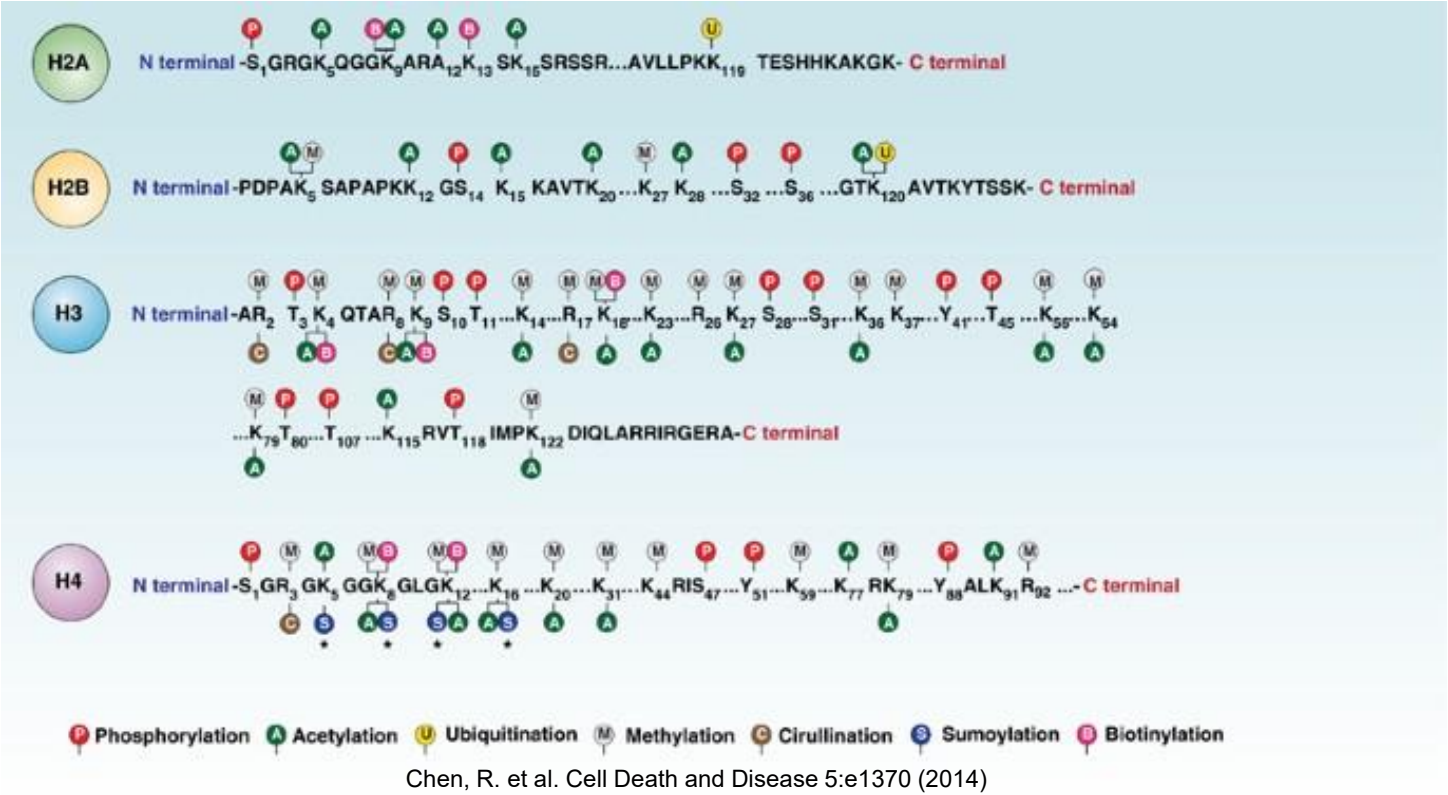
Histone and nucleosome



- Histone is a family of proteins that together form an octamer
- DNA wraps around histones for packaging
- A unit of DNA-histone is called nucleosome (~150bp)

Chen, R. et al. Cell Death and Disease 5:e1370 (2014)

Modification of histone tails



Chen, R. et al. Cell Death and Disease 5:e1370 (2014)

Regulatory roles of histone tail modification

Recognition by RNA pol II complex

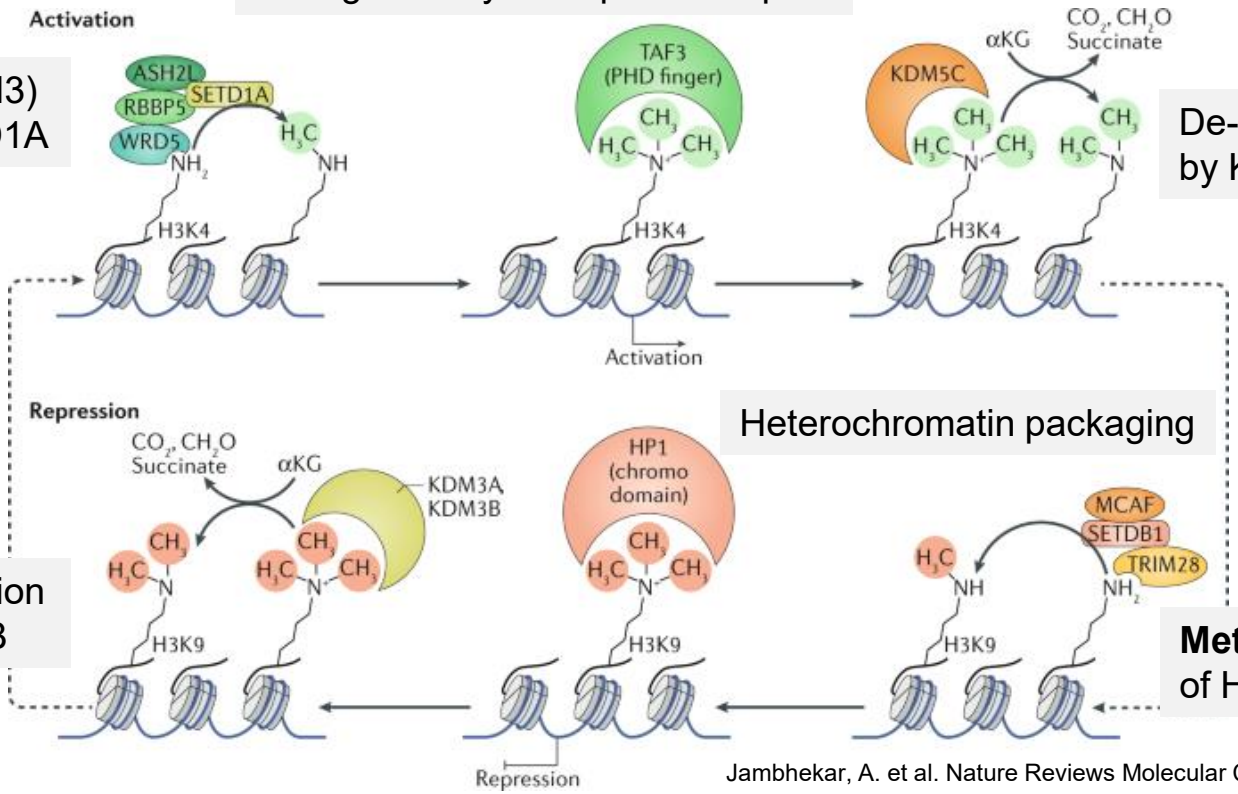
Methylation (-CH3) of H3K4 by SETD1A

De-methylation by KDM5C

Heterochromatin packaging

De-methylation by KDM3A/B

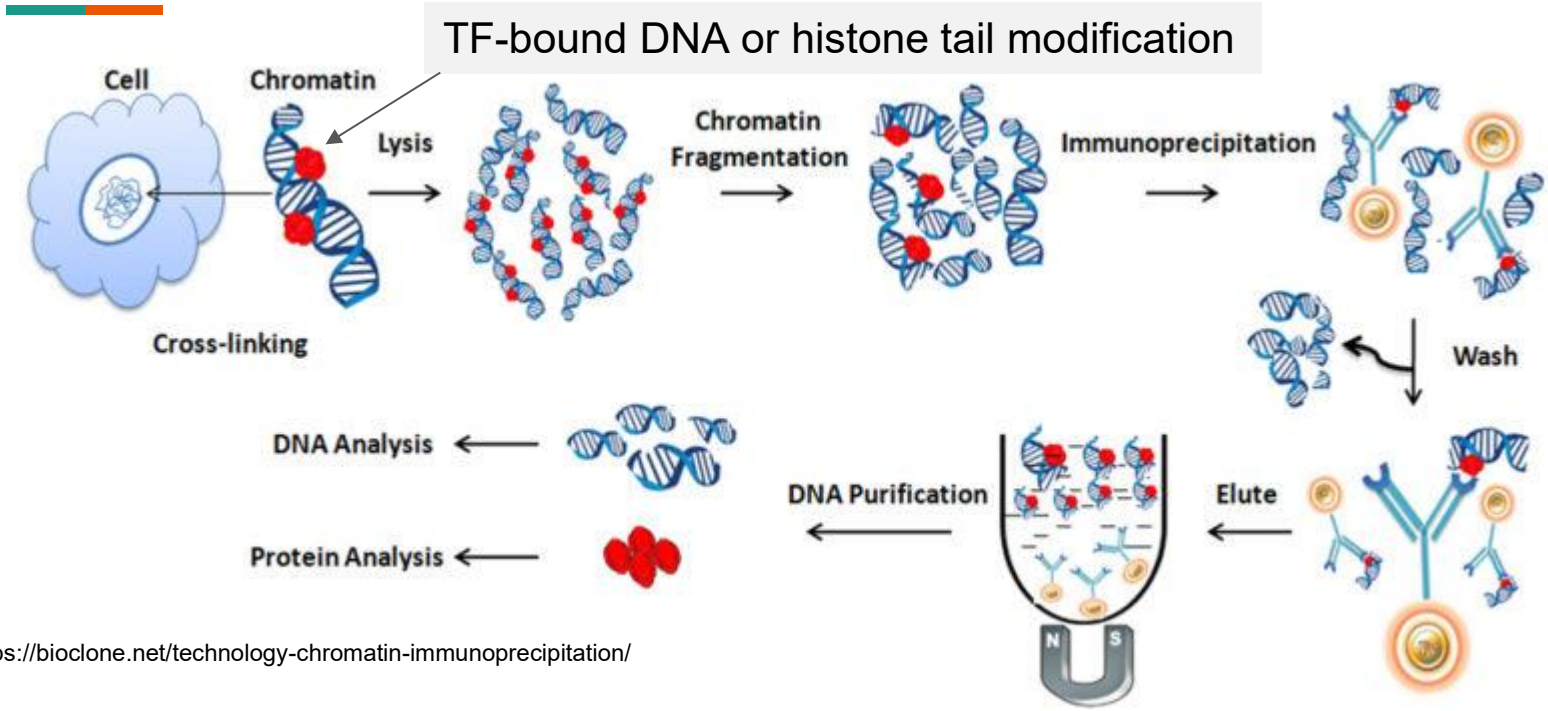
Methylation (-CH3) of H3K9 by SETDB1



Jambhekar, A. et al. Nature Reviews Molecular Cell Biology 20:625-641 (2019)

Chromatin immunoprecipitation (ChIP)

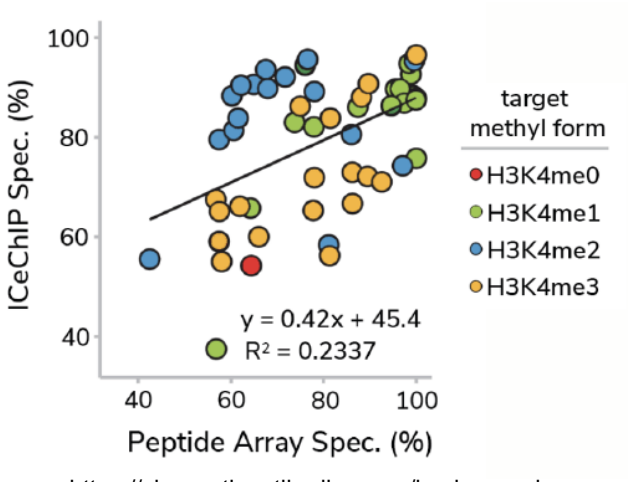
Chromatin immunoprecipitation



https://bioclone.net/technology-chromatin-immunoprecipitation/

- Selective enrichment of DNA segments via antibody pull-down

Antibodies for histone modifications



https://chromatinantibodies.com/background

- Specificity is important
- Study literature to find reliable Ab

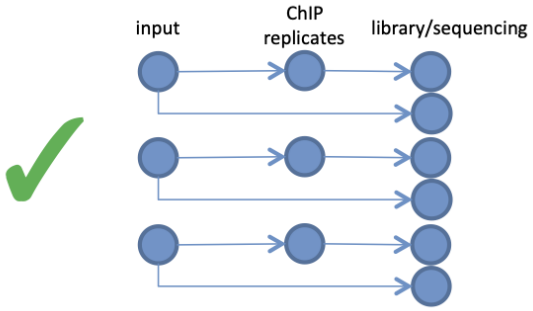
Monoclonal histone modification antibodies

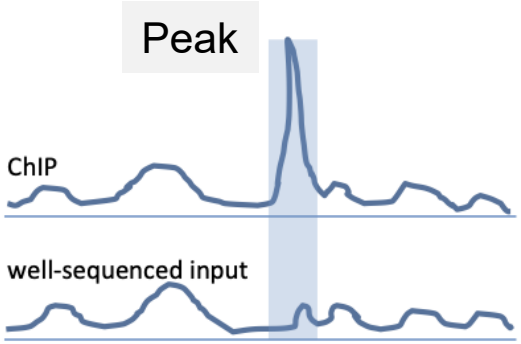
Abcam catalog

Modification	Function	Rabbit monoclonal antibody	Mouse monoclonal antibody
H3K4me1	Activation	ab176877	-
H3K4me3	Activation	ab213224	-
H3K9ac	Activation	ab32129	-
H3K27ac	Activation	ab177178	-
H3K9me2	Repression	ab32521	ab1220
H3K9me3	Repression	ab176916	-
H3K27me3	Repression	ab192985	-
γH2A.X	DNA damage	ab81299	ab26350
H3S10p	DNA replication	ab177218	ab14955

ChIP-seq experiment setup

- Need matched control to model the baseline read count per genomic segment
- **Input control** = no immunoprecipitation
- **Peak calling** = detection of high local read count in ChIP sample relative to control
- Single-end sequencing is good enough!



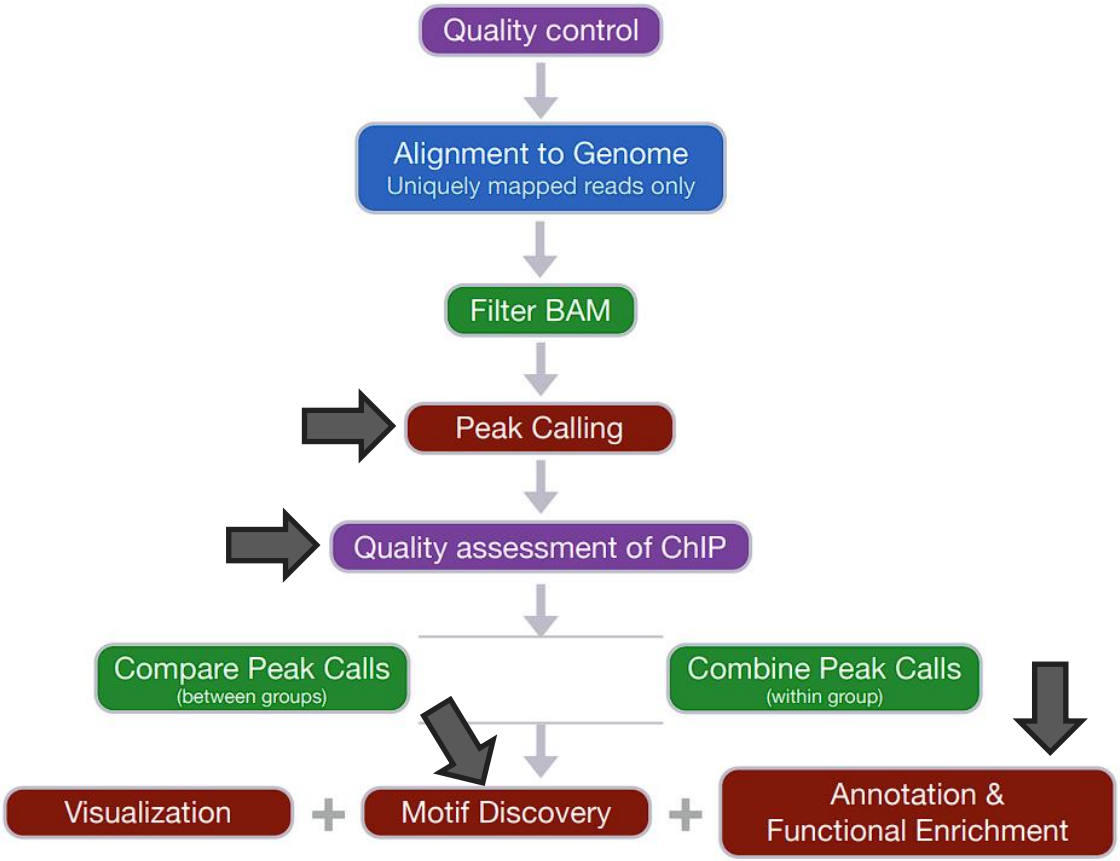


https://nbis-workshop-epigenomics.readthedocs.io/en/latest/content/tutorials/expdesign/expdes-ChIPseq.html

Analysis of ChIP-seq data

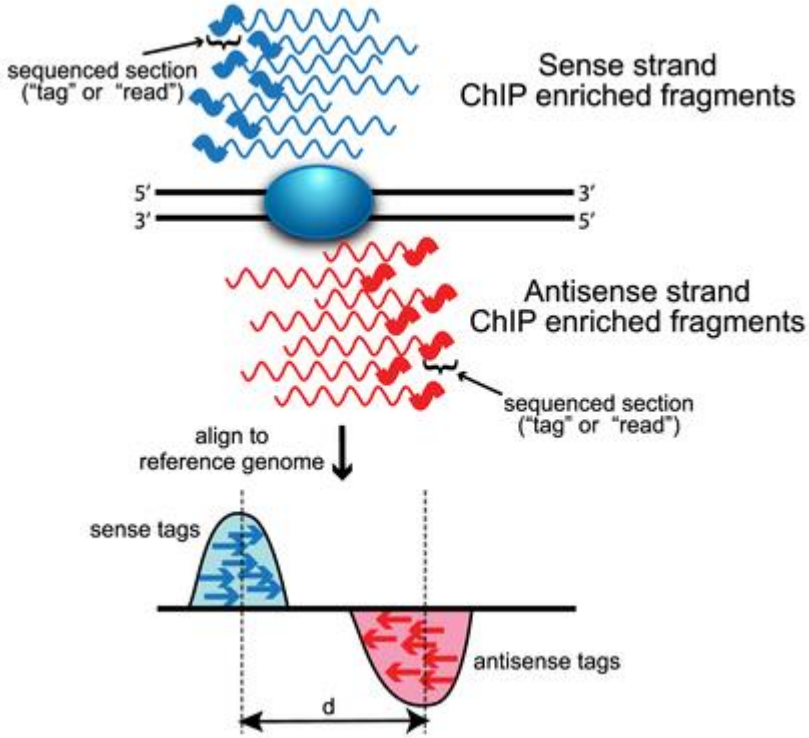
Overview

- What's new:
  - Peak calling
  - Quality check
  - Peak annotation
  - Functional enrichment
  - Motif discovery



https://github.com/hbctraining/Intro-to-ChIPseq

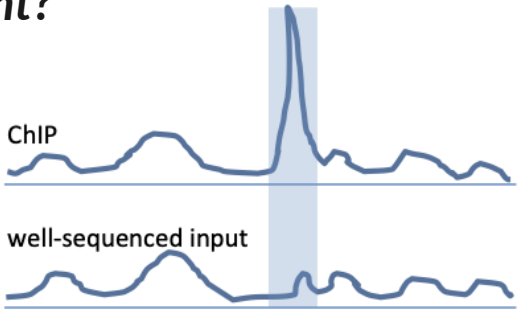
Peak calling (for TF)



Wilbanks, E.G. and Facciotti, M.T. PLoS ONE 5:e11471 (2010)

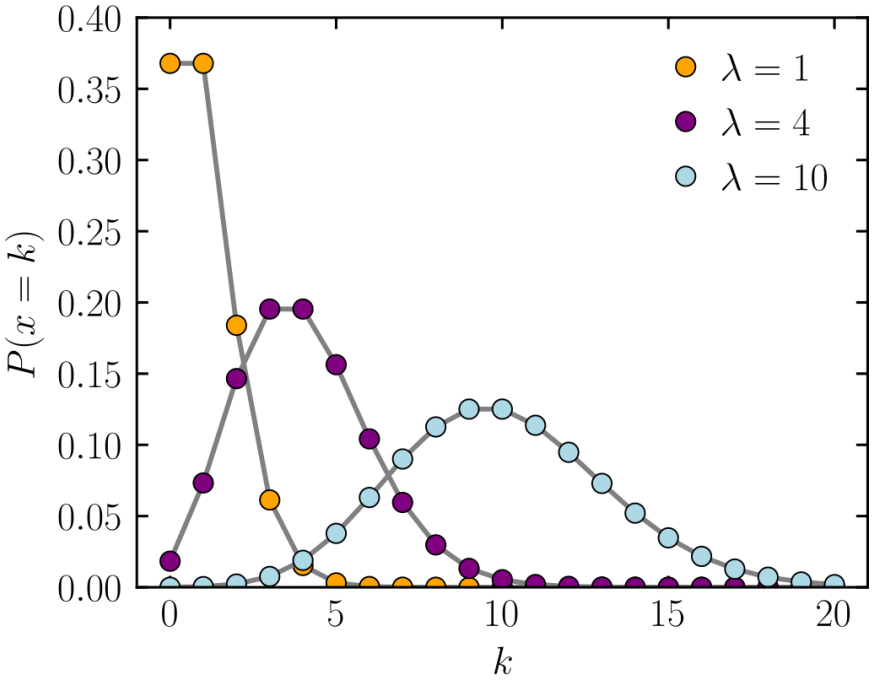
- Clusters of forward and reverse reads surrounding the binding sites
- *d* = DNA fragment size
- Peak height = read counts
  - ***Is it significant?***

Which statistical model to use?



https://nbis-workshop-epigenomics.readthedocs.io/en/latest/content/tutorials/expdesign/expdes-ChIPseq.html

Poisson distribution



https://en.wikipedia.org/wiki/Poisson_distribution

- The probability that an event will occur *k* times within a certain time or space (with expectation = λ)
- P(k) = $\frac{\lambda^k e^{-\lambda}}{k!}$
- **For ChIP-seq**: The probability that we observe *k* reads a DNA segment (with expectation = λ, number of reads in the input control)

Poisson model for peak calling

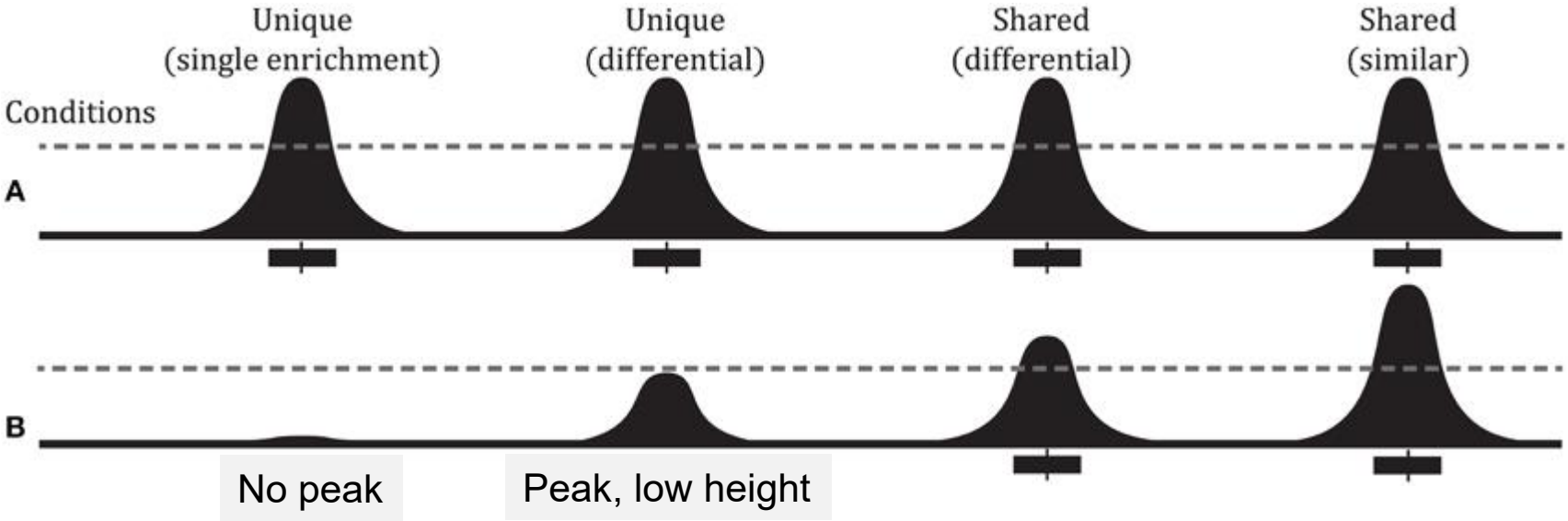
- **Null Hypothesis**: There is no peak. We expect the number of reads in ChIP sample to be the same as the input control (λ reads).

- P-value = P(observe ≥k reads in ChIP | expected λ reads) = $\sum_{x=k}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!}$

- **Low p-value**: Unlikely to observe k reads in ChIP sample by chance under the Null Hypothesis → **There is a peak in ChIP sample**

- This is why we need the input control, with sufficient sequencing depth to estimate λ

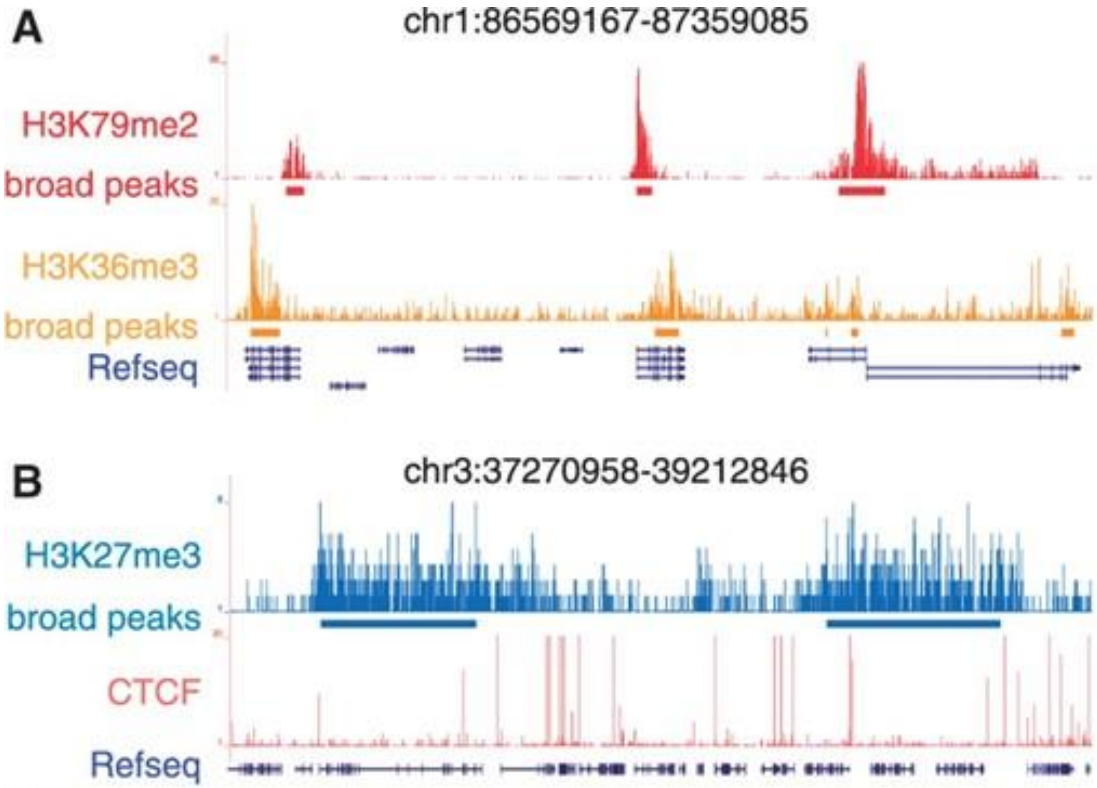
Differential peak calling



Wu, D.-Y. et al. Frontiers in Genetics 6:fgene.2015.00169 (2015)

- Comparing across experimental conditions, not with input control
- **Two-stage**: Peak calling → Compare height

Narrow and broad peaks

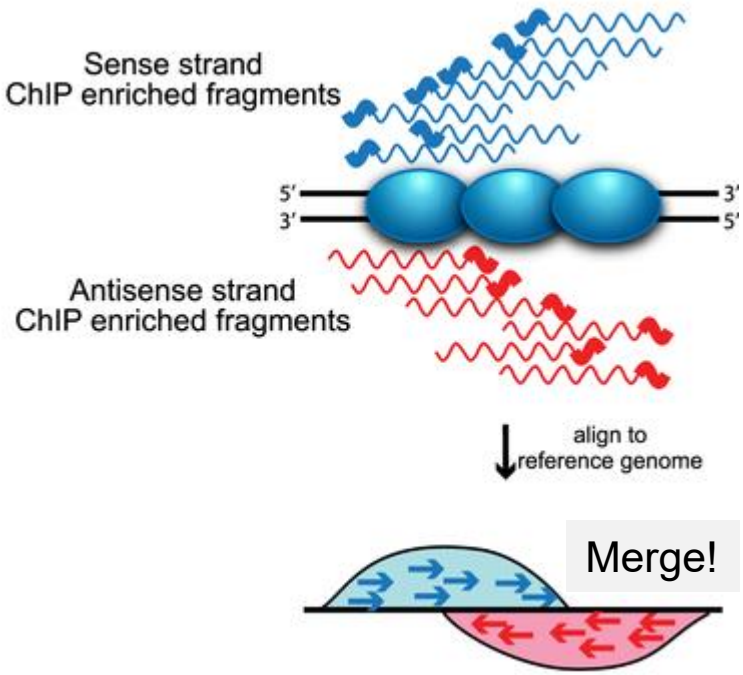


- [**Narrow**] TFs bind at precise locations
- [**Broad**] Histone modifications span a long DNA segment

DNA-binding protein

Wang, J. et al. Bioinformatics 29:492-493 (2013)

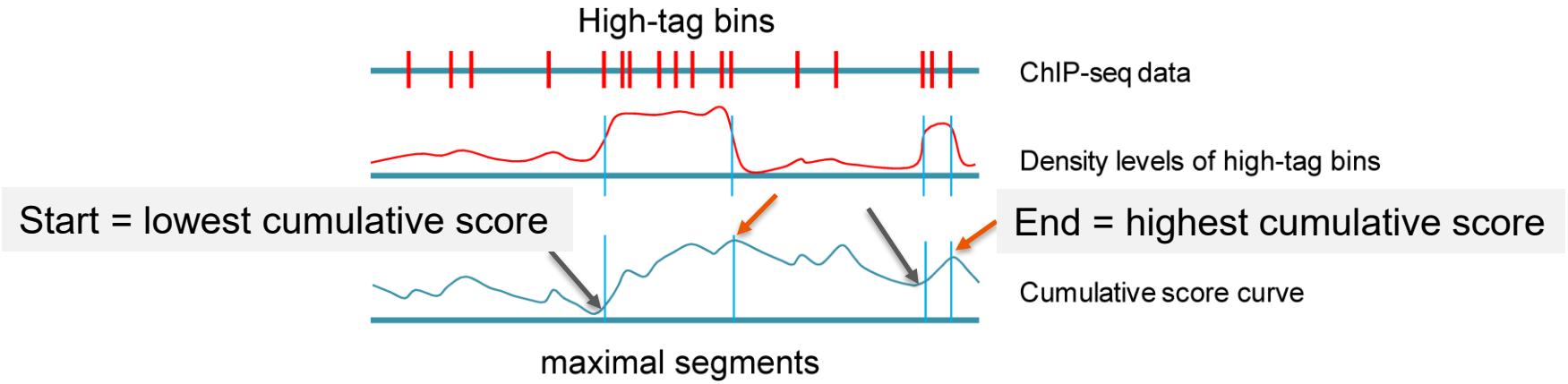
Calling of broad peaks



- Call individual peaks
- Merge adjacent peaks into broader areas
- What is the range (in bp) for merging adjacent peaks?
  - Optimized by known data
  - Visual inspection

Wilbanks, E.G. and Facciotti, M.T. PLoS ONE 5:e11471 (2010)

Calling of broad peaks



Wang, J. et al. Bioinformatics 29:492-493 (2013)

- Call individual peaks
- Score each DNA segment with the number of observed peaks
- Identify the **start** and **end** of high-density segments

BED file format

track type=n	Peak name="Somite narrowPeak"								
chr14	93429597	93429897	.	971	.	6.14099	-1	0.1492	150
chr2	217436588	217436888	.	1000	.	6.14825	-1	0.14907	150
chr2	63964529	63964829	.	1000	.	6.14955	-1	0.14903	150
chr9	115258329	115258629	.	954	.	6.17257	-1	0.14984	150
chr9	20692737	20693037	.	1000	.	6.18178	-1	0.14996	150
chr10	3828442	3828742	.	1000	.	6.20276	-1	0.15	150
chr3	4763989	4764289	.	732	.	6.28842	-1	0.15822	150
chr6	143037411	143037711	.	887	.	6.32704	-1	0.16192	150
chrX	55138332	55138632	.	1000	.	6.35559	-1	0.16467	150
chr8	126231677	126231977	.	1000	.	6.36141	-1	0.16485	150
chr2	120245492	120245792	.	1000	.	6.3983	-1	0.16748	150

- Tabular file with 3 requires columns:
  - Sequence (chromosome/scaffold/contig) name
  - Start position
  - End position
- Designate genomic regions (ChIP peaks, exomes, etc.)

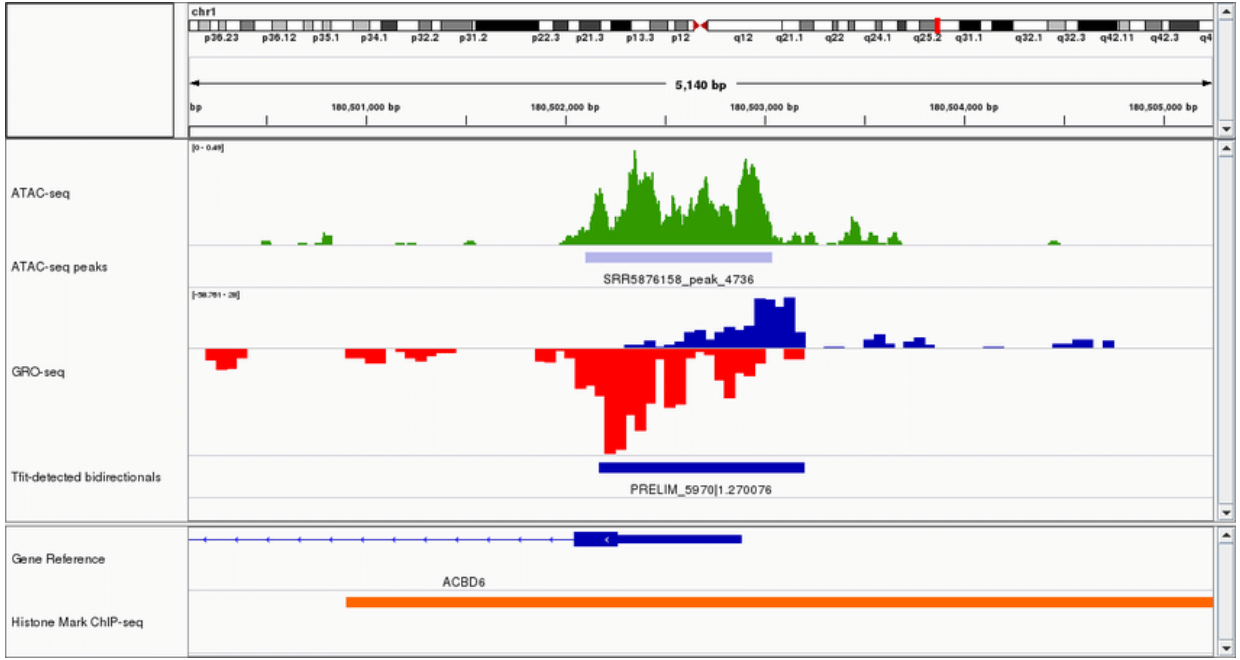
Quality check of called peaks: ChIPQC

Table 1. Summary of ChIP-seq filtering and quality metrics.

ID	Tissue	Factor	Condition	Replicate	Reads	Dup%	ReadL	FragL	RelCC	SSD	RiP%	RiBL%
Nanog.Rep1		Nanog		1	969186	0	36	121	1.4	2	4.1	2.9
Nanog.Rep2		Nanog		2	2283248	0	36	136	2.7	1.9	6.5	1.5
Pou5f1.Rep1		Pou5f1		1	1085316	0	36	164	2.3	2.6	3.9	3.2
Pou5f1.Rep2		Pou5f1		2	1995385	0	36	151	5.8	3	0.92	2.6
Nanog-Input1	NA	NA	NA	NA	4080970	0	36	73	0.2	4.3	NA	5.2
Nanog-Input2	NA	NA	NA	NA	1817134	0	50	104	0.64	0.91	NA	0.56

- **RiP%:** Percentage of reads in peak
- **SSD**: Variance of coverage across the genome
- **RiBL%:** Percentage of reads mapped to regions known to have artificially high read counts (microsatellite, mobile element, repeats, ribosomal DNA)
- **RelCC**: Consistency of the DNA fragment sizes

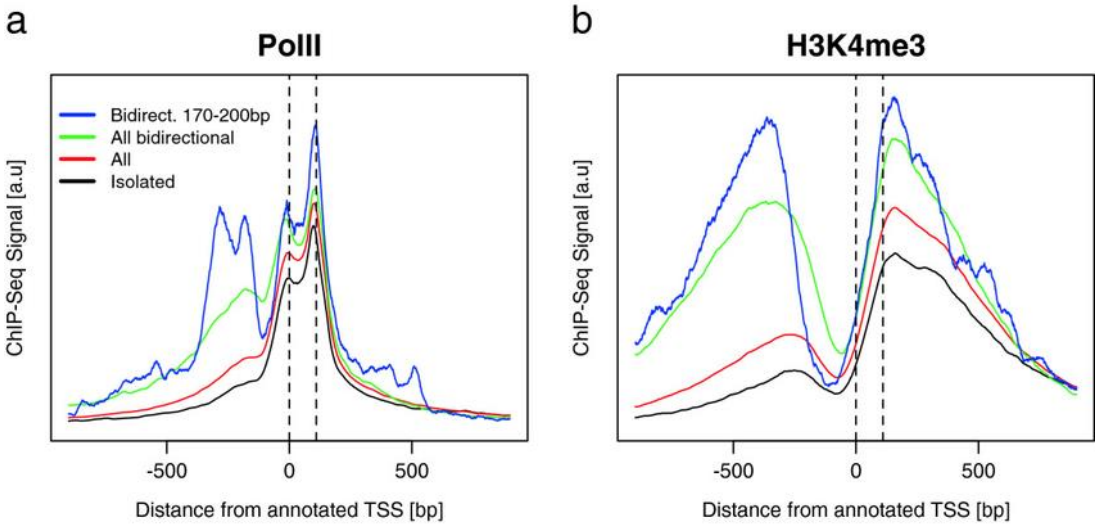
Visualization with Integrative Genomic Viewer



Tripodi, I. et al. Preprint DOI:10.1101/531517

- BAM or BED file from ChIP-seq analysis can be uploaded into IGV

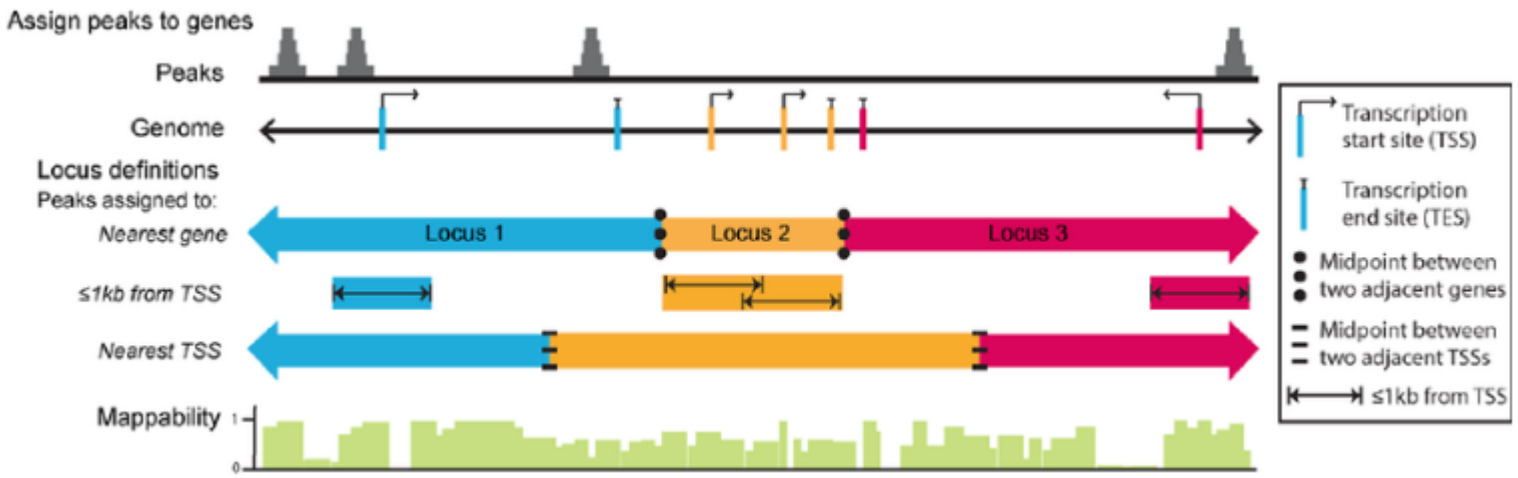
Relative visualization



- Summarization of peak location relative to the **transcription start site (TSS)** of the nearest gene

Quinodoz, M. et al. Genome Biology 15:R85 (2014)

Annotation of ChIP peaks

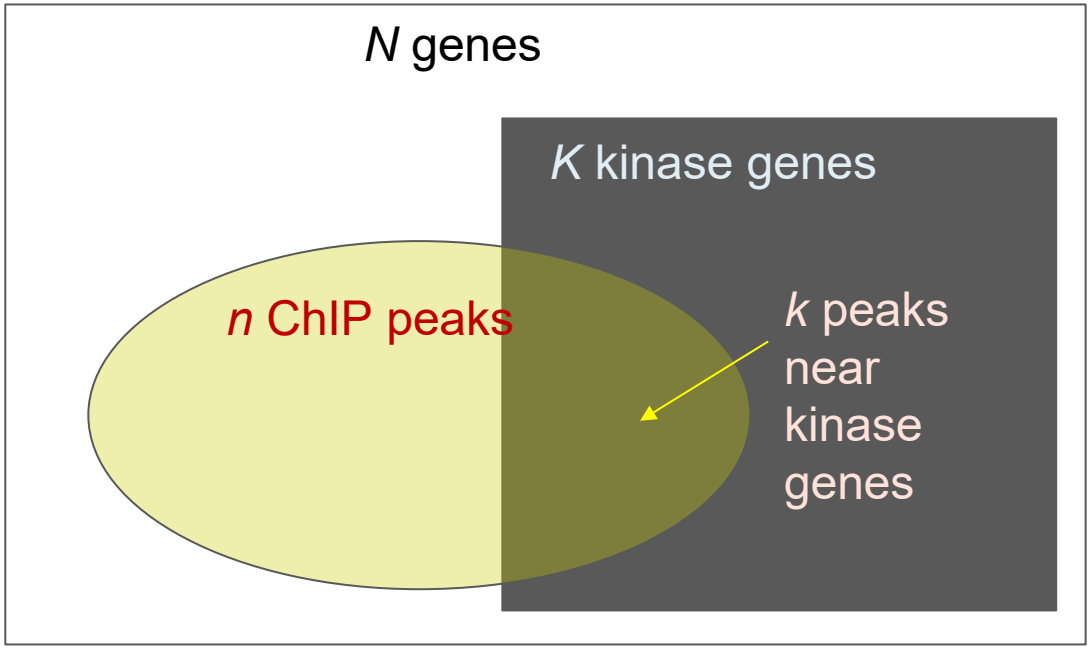


https://github.com/hbctraining/Intro-to-ChIPseq

- With no other evidence, peaks are mapped to the nearest genes and transcription start sites (TSS)
- Functional annotation of the genes are transferred to ChIP peaks

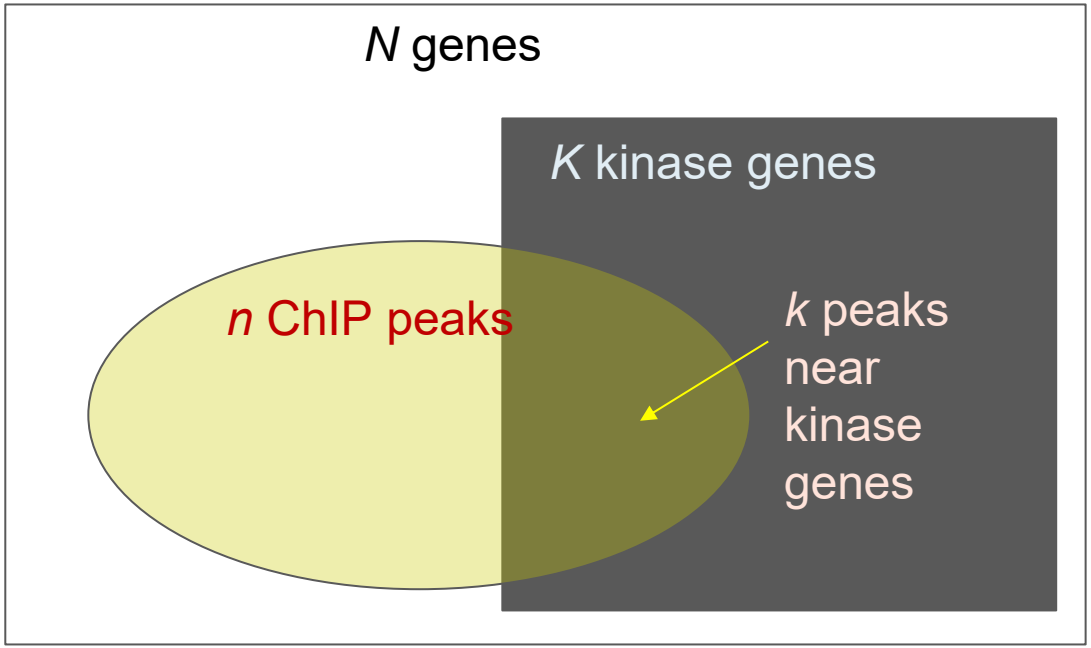
Functional enrichment analysis

Hypergeometric distribution



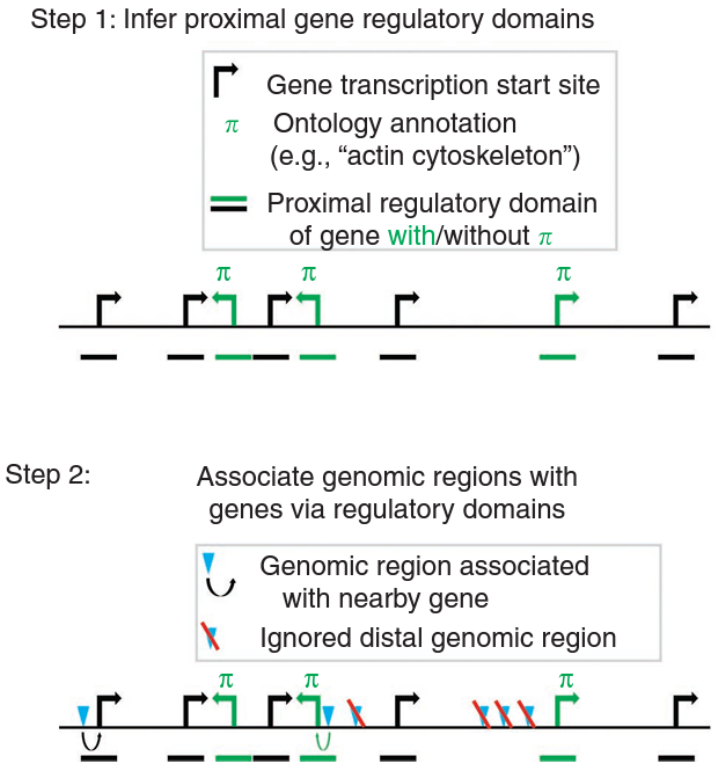
- **Null Hypothesis**: No association between ChIP peaks and kinases
- What is the probability of observing *k* out of *n* ChIP peaks being near kinase genes?
- Given total *N* genes, *K* of which are kinases

Hypergeometric distribution



- $P(N, K, n, k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$
- P-value = sum of $P(N, K, n, x)$ for $x \geq k$
- **Low p-value**: Reject Null Hypothesis, there is an association between ChIP peaks and kinases

Functional enrichment analysis for ChIP peaks



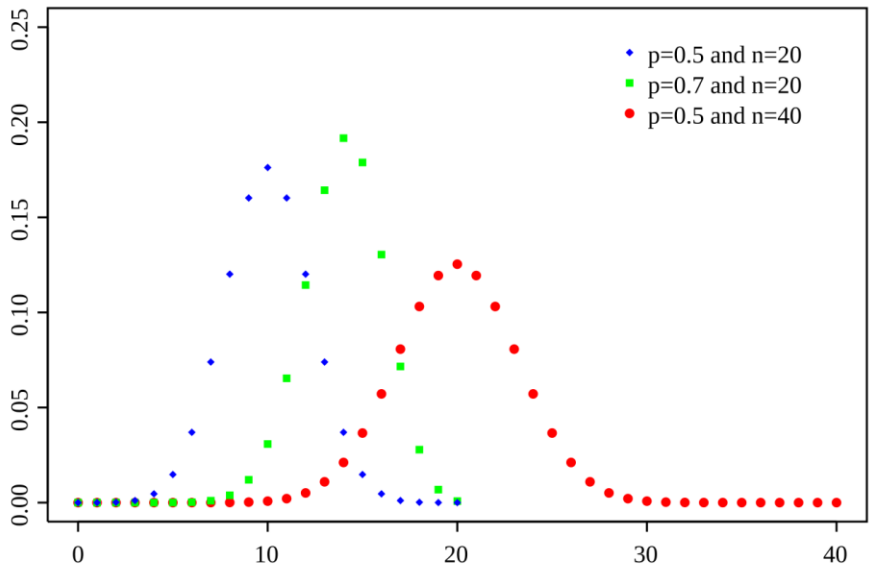
- 20,000 total genes
- 200 genes linked to cholesterol metabolism

- 1,000 total peaks identified
- 700 peaks are within 1kb of some genes
- 100 are near genes linked to cholesterol metabolism

- **Expectation**: 200 x 700 / 20,000 = 7 peaks linked to cholesterol metabolism
- 100 / 7 = 14-fold enrichment!

McLean *et al*. Nat Biotech, 28(5):495-501 (2010)

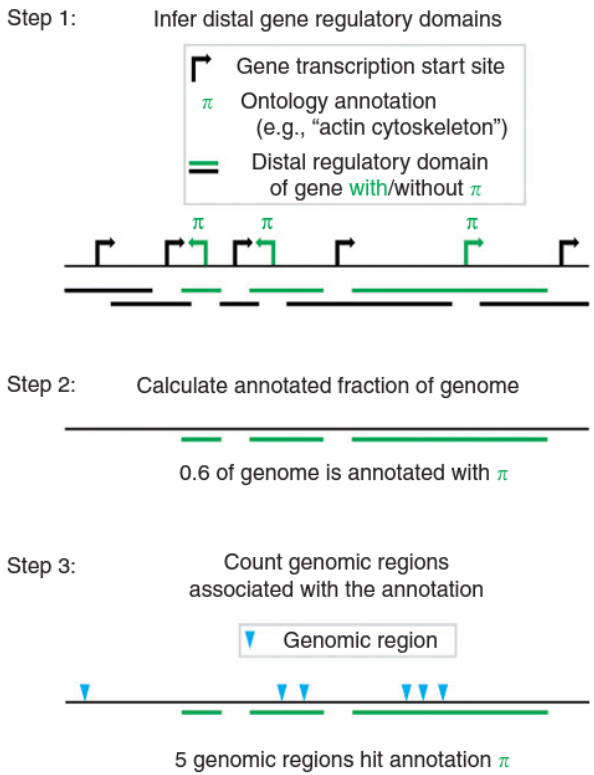
Binomial distribution



- The probability that an event with probability *p* will happen *k* times out of *n* trials
- P(k) = $\binom{n}{k} p^k (1-p)^{n-k}$
- P-value = sum of P(*x*) for *x* ≥ *k*
- How can we use this for ChIP analysis?

https://en.wikipedia.org/wiki/Binomial_distribution

Binomial model for functional enrichment



- Annotate genome segments according to the functions of nearby genes
- **0.0001** of the genome linked to cholesterol metabolism
- Out of **1,000** ChIP peaks, **50** fall in the regions linked to cholesterol metabolism
- **Expectation**: 0.0001 x 1,000 < 1 peak

Limitation of ChIP peak analysis

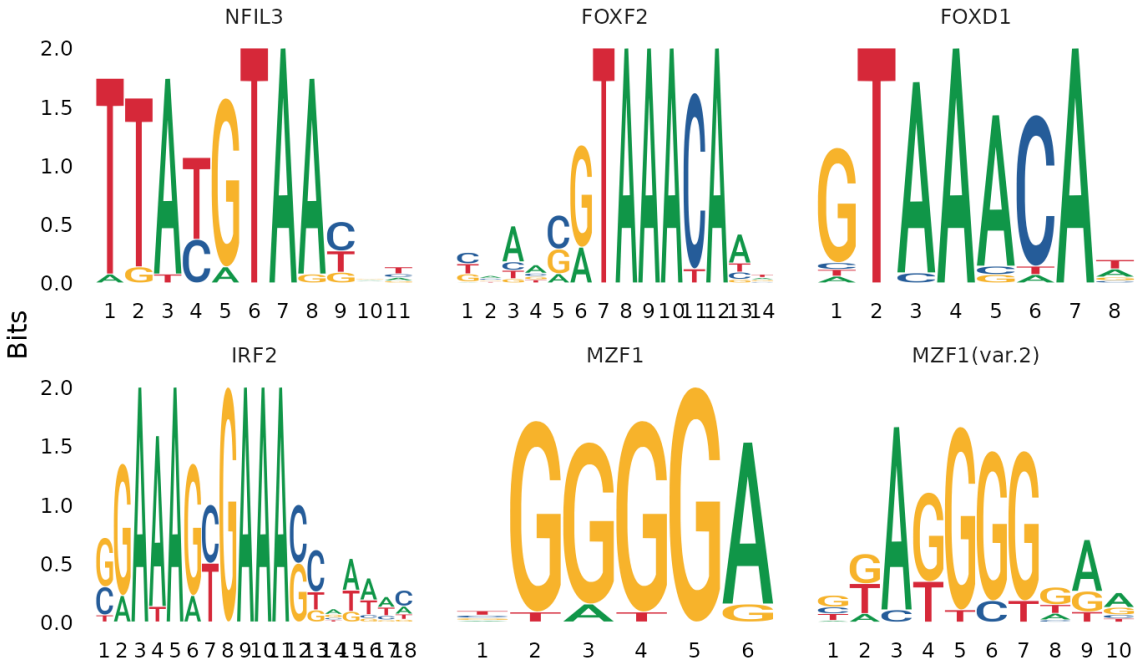
- Epigenetics can affect genes very far away from ChIP peak locations
  - Depend on distance in 3D, not distance on the genome sequence

- Epigenetics does not necessarily affect the nearest gene

- Interpret together with other omics data
  - **With transcriptomics**: Does increase in gene expression coincide with more TF binding or activating histone markers?
  - **With genomics**: Does disease-specific mutations disrupt epigenetics?

DNA motif

DNA motifs

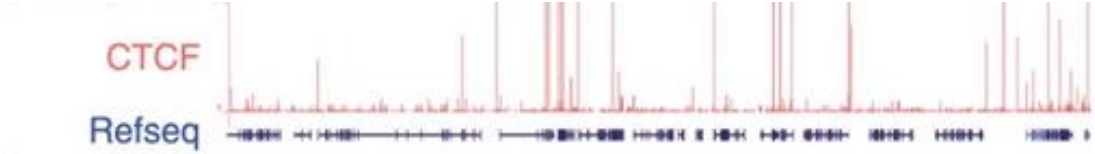


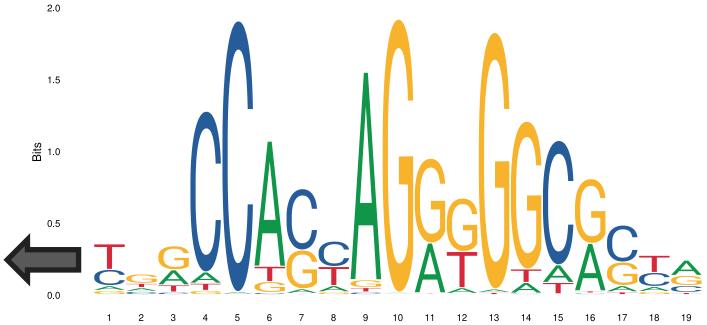
- Patterns of DNA sequence recognized by proteins or other molecules
- DNA binding motifs
- Similar concept to **Position-Specific Scoring Matrix** (PSSM)

https://stuartlab.org/signac/1.2.0/reference/motifplot

Discovery of DNA binding motifs from ChIP data

Wang, J. et al. Bioinformatics 29:492-493 (2013)



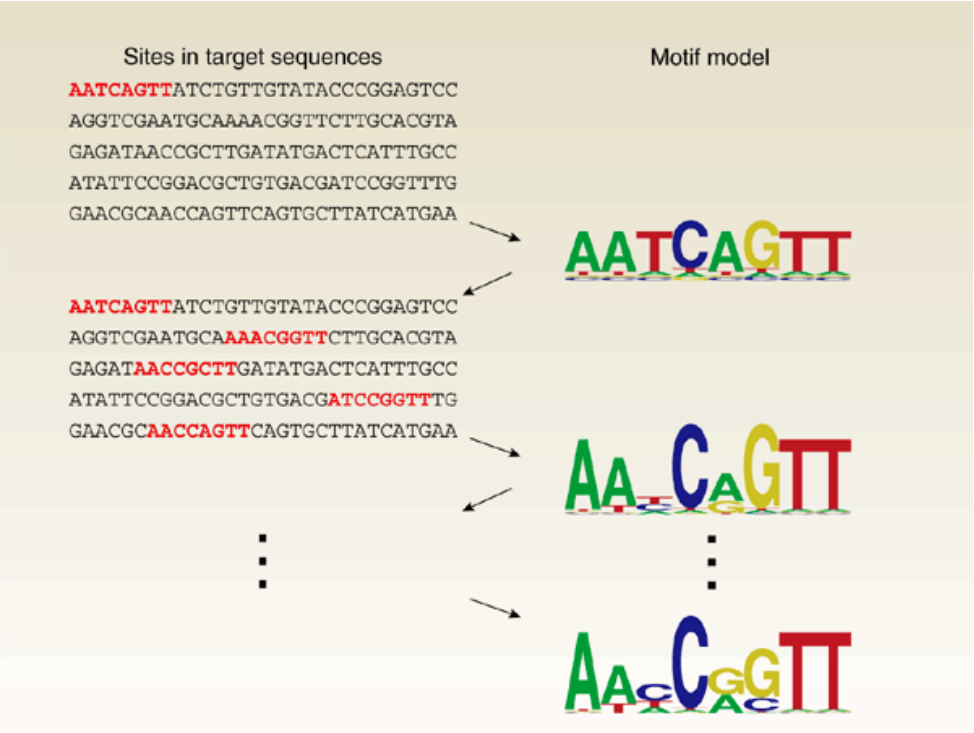


https://jaspar2018.genereg.net/matrix/MA0139.1/

- If TF recognize specific DNA motifs, every ChIP peak should contain at least 1 occurrence of those patterns!

- Is there an algorithm to find common DNA patterns shared by a collection of DNA segments? How to test for statistical significance?

Motif discovery algorithm sketch



D'haeseleer, P. Nature Biotechnology 24:959-961 (2006)

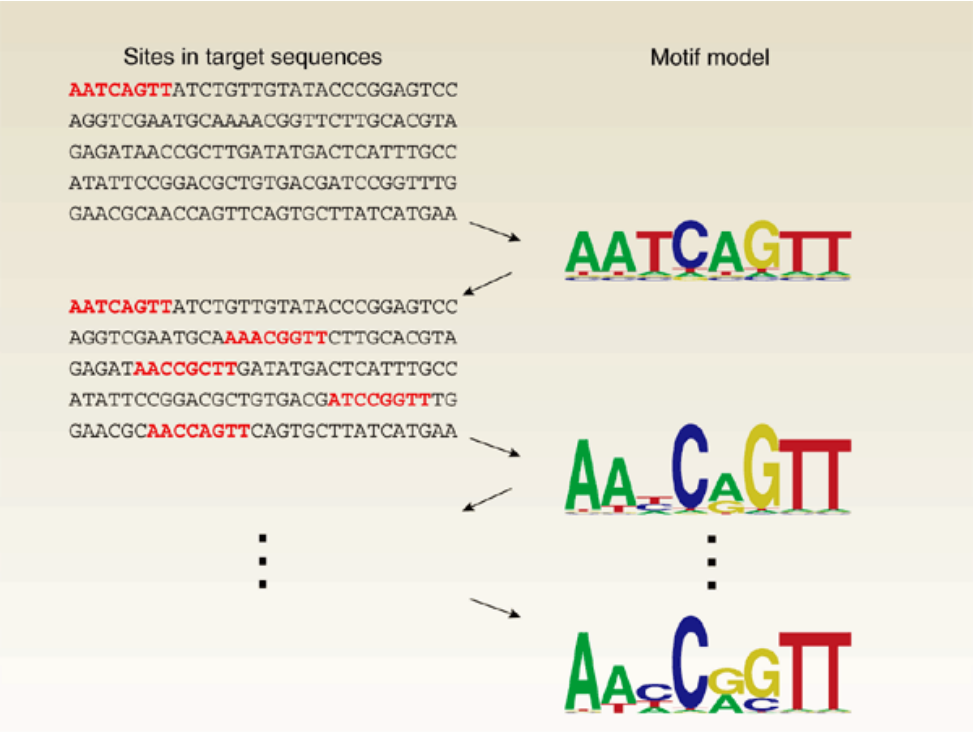
- Guess a motif (fixed length)
- Find the best match in each sequence (ChIP peak)

- Update motif's PSSM
- Find (possibly better) match in each sequence

- Repeat the two steps

- **Same idea as PSI-BLAST**

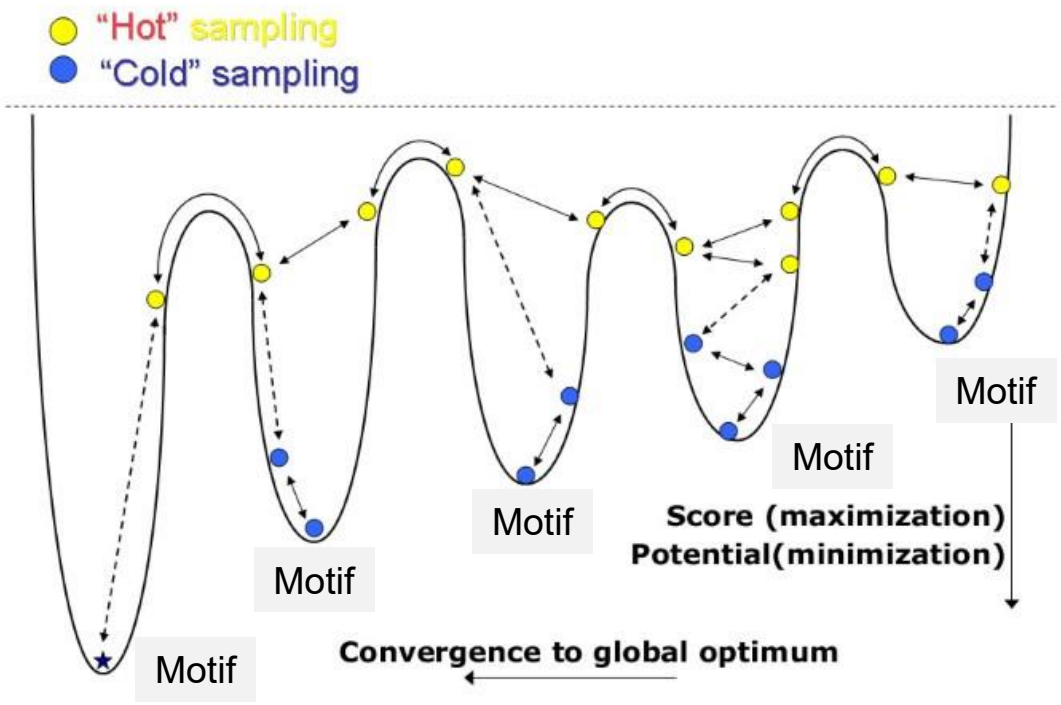
Issue of sampling algorithm



D'haeseleer, P. Nature Biotechnology 24:959-961 (2006)

- Final answer depend on the initial guess!
- Smart guess:
  - Compare sequences beforehand and identify matching DNA patterns
- Brute force:
  - Try multiple guesses
  - Select the best final motifs

An example of sampling strategy

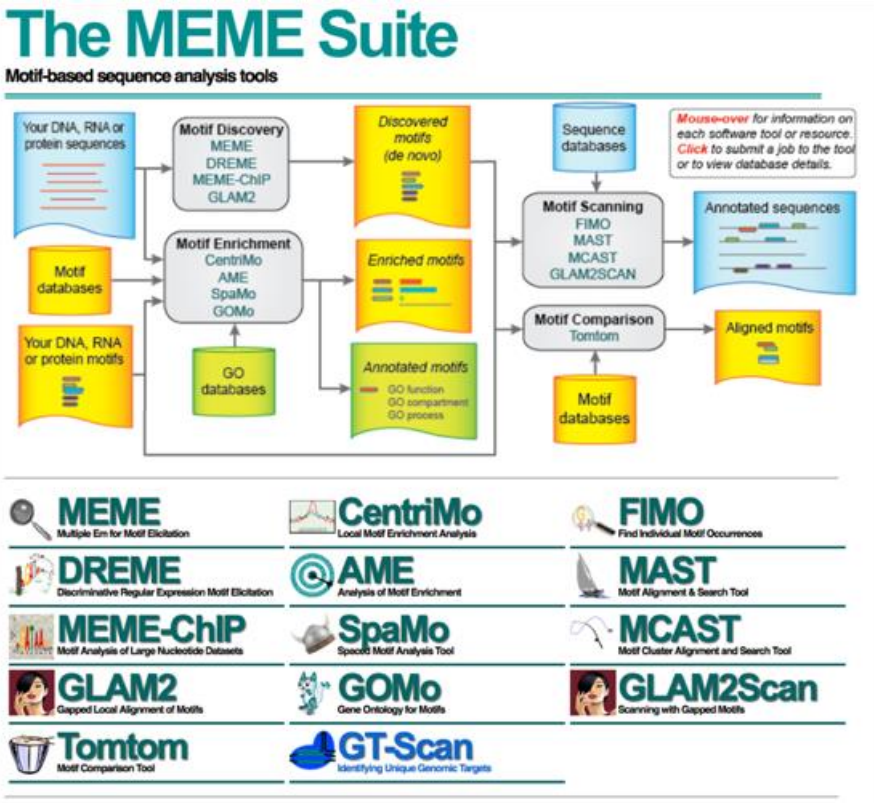


Shida, K. BMC Bioinformatics 7:486 (2006)

- Generate diverse guesses (hot sampling)
- For each guess, let the algorithm converge to a local best motif (cold sampling)
- **A general approach in AI and physics simulation**

One-stop service for DNA motif analysis

- Motif discovery from your DNA sequences
- Search for known motifs in your DNA sequences
- **Test if your DNA sequences contain some motifs more frequently than in the genomic background**



Scoring of motifs

$$\begin{pmatrix} A & 15 & 4 & 47 & 0 & 0 & 0 & 3 & 0 & 14 \\ C & 15 & 20 & 0 & 47 & 0 & 0 & 1 & 1 & 20 \\ G & 8 & 3 & 0 & 0 & 47 & 0 & 43 & 18 & 5 \\ T & 9 & 20 & 0 & 0 & 0 & 47 & 0 & 28 & 8 \end{pmatrix}$$

...CC**ACGT**AGC...

...CC**ACGTGT**C...

https://cs.rice.edu/~ogilvie/comp571/pssm/

- Given a PSSM, a DNA sequence can be scored according to the probability
- P(CC**ACGT**AGC| PSSM) $= \frac{15}{47} \cdot \frac{20}{47} \cdot \frac{47}{47} \cdot \frac{47}{47} \cdot \frac{47}{47} \cdot \frac{47}{47} \cdot \frac{3}{47} \cdot \frac{18}{47} \cdot \frac{20}{47} = 0.001412$
  - Log-odd = $\text{Log}\left(\frac{p}{1-p}\right) = -2.85$
- P(CC**ACGTGT**C| PSSM) $= \frac{15}{47} \cdot \frac{20}{47} \cdot \frac{47}{47} \cdot \frac{47}{47} \cdot \frac{47}{47} \cdot \frac{47}{47} \cdot \frac{43}{47} \cdot \frac{28}{47} \cdot \frac{20}{47} = 0.031498$
  - Log-odd = $\text{Log}\left(\frac{p}{1-p}\right) = -1.49$

Enrichment of a motif

Your DNA sequences
(ChIP-seq peaks)

# of motif occurrence
# of sequence with motif
Distribution of log-odd scores

Background DNA sequences
(Random genomic segments)

# of motif occurrence
# of sequence with motif
Distribution of log-odd scores

- **Null Hypothesis**: Your DNA sequences are not associated with the motif. Expect the same occurrence as random genomic segments

- **Caution**: The difference in DNA *k*-mers can bias motif occurrence!
  - Select from the same genome, or generated

Any question?

- See you next time