# **ChiP-Seq Analysis Pipeline**

Lecture 4

# Outline

ChIP-seq

• ChIP-seq Analysis Pipeline

• Transcription Factor Binding Site (TFBS) Motif Discovery Chromatin ImmunoPrecipitation Followed by Sequencing (ChIP-seq)

- Combine chromatin immunoprecipitation assays with sequencing.
- DNA-bound protein is immunoprecipitated using a specific antibody.
- The bound DNA is then coprecipitated, purified, and sequenced.
- Mapping DNA-Protein Interactions: map DNA-binding proteins and histone modifications in a genome-wide manner at base-pair resolution

# How Does ChIP-Seq Work?

- ChIP-Seq identifies the binding sites of DNA-associated proteins and can be used to map global binding sites for a given protein.
- ChIP-Seq typically starts with crosslinking of DNA-protein complexes. Samples are then fragmented and treated with an exonuclease to trim unbound oligonucleotides.
- Protein-specific antibodies are used to immunoprecipitated the DNAprotein complex.
- The DNA is extracted and sequenced, giving high-resolution sequences of the protein-binding sites.



Add bead-attached antibodies to immunoprecipitated target protein



#### Types of Signals by ChIP-seq



Adapted from Park (2009). Nature Reviews Genetics.

# Advantages of ChIP-Seq

- ChIP-Seq delivers genome-wide profiling with massively parallel sequencing, generating millions of counts across multiple samples for cost-effective, precise, unbiased investigation of epigenetic patterns.
- Captures DNA targets for <u>transcription factors</u>, <u>histone modifications</u>, <u>or nucleosomes</u> across the entire genome of any organism
- Defines transcription factor binding sites
- Reveals gene regulatory networks in combination with RNA sequencing and methylation analysis

#### **Transcriptional Regulation is Complex**





#### Overview of ChIP– seq Analysis

Figure 4 | **Overview of ChIP–seq analysis.** The raw data for chromatin immunoprecipitation followed by sequencing (ChIP–seq) analysis are images from the next-generation sequencing platform (top left). A base caller converts the image data to sequence tags, which are then aligned to the genome. On some platforms, they are aligned with the aid of quality scores that indicate the reliability of each base call. Peak calling, using data from the ChIP profile and a control profile (which is usually created from input DNA), generates a list of enriched regions that are ordered by false discovery rate as a statistical measure. Subsequently, the profiles of enriched regions are viewed with a browser and various advanced analyses are performed.

#### ChIP-seq Analysis Workflow



Adapted from Bailey T. et al, PLOS Comp. Bio. 2013







#### Example Study Design



Kuehner J.N. et al, Cell Rep. 2021

#### **Evaluate Data Qualify**

 QC of samples with mapped data:
<u>Pearson correlations</u> with read counts at each genomic position.



Kuehner J.N. et al, Cell Rep. 2021

#### **Evaluate Data Qualify**

• QC of samples with mapped data: <u>Principal Components Analysis</u> (PCA) with read counts at each genomic position.



# Peak Calling: MACS

- Peak Calling: Predict the regions of the genome where the ChIPed protein is bound by finding regions with significant numbers of mapped reads (peaks).
- MACS (Model-based Analysis of ChIP-Seq, Zhang Y. et al. Genome Biology, 2008)
  - Captures the <u>influence of genome complexity</u> to evaluate the significance of enriched ChIP regions
  - Improves the spatial resolution of binding sites through combining the information of both <u>sequencing tag position and orientation</u>.
  - Input/Control sample: Sequence data of one sample without IP. Improves specificity.
  - Although it was developed for the detection of transcription factor binding sites it is also suited for larger regions.
- MACS3: <u>https://macs3-project.github.io/MACS/</u>

### MACS: Modeling the Shift Size

- MACS randomly samples 1,000 of these high-quality peaks, separates their positive and negative strand tags, and aligns them by the midpoint between their centers.
- The distance between the modes of the two peaks in the alignment is defined as 'd' and represents the estimated fragment length.
- MACS shifts all the tags by d/2 toward the 3' ends to the most likely protein-DNA interaction sites.



#### **MACS:** Peak Detection

- Slides across the genome using a window size of 2d to find candidate peaks.
- The tag distribution along the genome can be modeled by a Poisson distribution.
- The Poisson is a one parameter model, where the parameter λ is the expected number of reads in that window.

$$P_{\lambda}(X=k) = \frac{\lambda^{k}}{k! * e^{-\lambda}}$$

- $\lambda$  = mean = expectated value = variance
- $\lambda = \underline{\text{total number of events (k)}}$ number of units (n) in the data
  - = <u>Read length (nt) \* Total read number</u> Effective genome length (nt)



#### **MACS:** Peak Detection



- MACS uses a dynamic parameter,  $\lambda_{local}$ , defined for each candidate peak. The  $\lambda_{local}$  parameter is estimated from the <u>control sample (NULL)</u> and is given by the maximum value across various window sizes:  $\lambda_{local} = \max(\lambda_{BG}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k})$ .
- In this way  $\lambda$ \_local captures the influence of local biases, and is **robust against** occasional low tag counts at small local regions.
- A region is considered to have a significant tag enrichment if the Poisson distribution p-value < 10E-5 or FDR < 5%.</li>
- Overlapping enriched peaks are merged, and each tag position is extended 'd' bases from its center. <u>The location in the peak with the highest fragment pileup</u>, <u>hereafter referred to as the summit</u>, is predicted as the precise binding location.
- The ratio between the ChIP-seq tag count and  $\lambda\_local$  is reported as the fold enrichment.  $^{\rm 20}$





Peak Visualization



are combined into one BAM file.

# Differential Binding Analysis

- Test the difference of read counts in cases vs. controls, for each genomic region. DESeq2 (R library)
- <u>https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst</u> /doc/DESeq2.html
- Identify genomic regions with differential binding between two groups.
- Genes linked with these differential methylated regions can be used for follow-up gene ontology analysis.

#### Visualizing Differential Binding Regions



(A) Number of established and disappeared 5hmC peaks at D56, D84, and D112.

(B–D) Heatmaps of developmental-stage-specific differentially hydroxymethylated regions (**DhMRs**), where the color scale represents normalized 5hmC read counts.

#### Peak Annotation: ChIPseeker

Associate the ChIPseq peaks with functionally relevant genomic regions, such as gene promoters, transcription start sites, intergenic regions, etc.



http://bioconductor.org/packages/release/bioc/vignettes/ChIPseeker/inst/doc/ChIPseeker.html

# Gene Ontology Analysis

- Determine if the ChIPed protein is involved in particular biological processes.
- Get the list of genes whose transcription starting site is annotated as the nearest ones to the peak regions of interest.
- Tools:
  - DAVID: https://david.ncifcrf.gov/home.jsp
  - GREAT:
    - http://great.stanford.edu/public/html/
  - **GSEA**: <u>https://www.gsea-</u> msigdb.org/gsea/index.jsp

D GO: AD Enriched DhMRs and Increasing Gene Expression



GO: AD Depleted DhMRs and Decreasing Gene Expression



Kuehner J.N. et al, Cell Rep. 2021

# What Are Motifs?

- A <u>motif</u> in molecular biology is a relatively short sequence of nucleotides or amino acids that <u>changes little during evolution</u> and, at least presumably, has a definite biological function.
- A motif is sometimes meant not a specific sequence, but <u>a spectrum of</u> <u>sequences</u> described in some way, each of which is capable of <u>performing a</u> <u>certain biological function</u> of a given motif.
- Motifs are <u>ubiquitous in living organisms</u> and perform many vital functions, such as regulation of transcription and translation (in the case of nucleotide motifs), post-translational modification and cellular localization of proteins, and partially determine their functional properties (leucine zipper).
- They are widely used in bioinformatics for <u>predicting the functions of genes</u> and proteins, constructing regulation maps, and are important for many problems in genetic engineering and molecular biology in general.

#### Motifs in Nucleic Acids

- In the case of DNA, most often motifs are short sequences that are binding sites for proteins, such as nucleases and transcription factors.
- Or are involved in important regulatory processes already at the RNA level, such as ribosome entry, mRNA processing, and transcription termination.

#### **Transcription over-simplified**

- 1. **TF** binds to DNA at **TF** binding site
- 2. TF recruits RNA polymerase II
- 3. RNA polymerase II produces RNA



#### **Anatomy of transcriptional regulation**

WARNING: Terms vary widely in meaning between scientists



- Core promoter Sufficient for initiation of transcription; orientation dependent
  - TSS transcription start site
    - Often really a transcription start region
- TFBS single transcription factor binding site
- Regulatory regions
  - Proximal/distal vague reference to distance from TSS
  - May be positive (enhancing) or negative (repressing)
  - Orientation independent (generally)
  - Modules Sets of TFBS within a region that function together
- Transcriptional unit
  - DNA sequence transcribed as a single polycistronic mRNA

#### **Motif Discovery Problem**

Given sequences



Find motif

IGRGGFGEVY at position 515 LGEGCFGQVV at position 430 VGSGGFGQVY at position 682



#### Motif Discovery Problem

- Given:
  - a sequence or family of sequences.
- Find:
  - the number of motifs
  - the width of each motif
  - the locations of motif occurrences



#### Why is this hard?

- Input sequences are long (thousands or millions of residues).
- Motif may be *subtle* 
  - Instances are short.
  - Instances are only slightly similar.





#### **Representing binding sites for a TF**

- Single site
  - AAGTTAATGATTAAC
- Set of sites, represented as a consensus
  - VDRTWRWWSHDWVDH (IUPAC degenerate DNA)
- Set of sites, represented as a position frequency matrix (PFM)





Set of binding sites AAGTTAATGATTAAC CAGTTAATAAATAAC GAGTTAAACACTAAA CAGTTAATTAGTAAC GAGTTAATAAATAAC CAGTTATTCAGTAAC GAGTTAATAAATCAT CAGTTAATCAGTAAT AGATTAAAGAATAAT AAGTTAACGATTAAC AGGTTAACGATACAC ATGTTGATGATAAAC AAGTTAATGATAAAT AAGTTAACGATAAAC AAATTAATGATTCAC GAGTTAATGATTAAA AAGTTAATCATTGAC AAGTTGATGATTAAG AAATTAATGATTGAC ATGTTAATGATTAAC AAGTAAATGATTAAA AAGTTAATGATTGCC AAGTTAATGATTGAC AAATTAATGATTGAC AAGTTAATGATTAGG AAGTTAATGATTAAT AAGTTAATGATTAGC AAGTTAATGATTAAT

#### Challenges

- PWMs can accurately reflect *in vitro* binding properties of DNA-binding proteins
- Suitable binding sites occur at a rate far too frequent to reflect *in vivo* function
- In vivo presence of a DNA-binding protein often occurs without a strong motif
- Bioinformatics methods that use PWMs for binding site studies must incorporate additional information to enhance specificity
  - Unfiltered predictions are too noisy for most applications
  - Organisms with short regulatory sequences are less problematic (such as yeast and *E. coli*)

**PWM:** Position Weight Matrix

TFBSshape: an expanded motif database for DNA shape features of transcription factor binding sites (Chiu T. et al. NAR, 2020)

- TF–DNA binding preferences are commonly described as <u>consensus</u> <u>sequence</u> represented by a <u>position weight matrix (PWM)</u> and visualized as <u>motif logo</u>.
- Traditional PWM-based methods assume that each nucleotide independently contributes to TF–DNA binding.
- An alternative representation of interdependencies between base pairs is the <u>three-dimensional (3D) DNA structure</u>.
- Thus, it becomes essential to understand the structural readout mechanisms underlying the recognition through <u>DNA shape changes</u> due to CpG methylation.

### DNA Shape Features of TFBSs

- Minor Groove Width (MGW)
- Roll
- Propeller Twist (ProT)
- Helix Twist (HelT)

IA shape features at Transcription Factor Binding Sit Using data from JASPAR2014, the Rohs' lab developed the TFBSshape database storing DNA shape features of TFBSs. Considered DNA shape features are :

#### **Including Shape Properties**



#### TFBSshape: Expanding to 18 DNA Features

- 14 Features for unmethylated DNA
  - Six intra-base pair features: Buckle, Opening, ProT (propeller twist), Shear, Stagger, and Stretch
  - Six inter-base pair features: HelT (helix twist), Rise, Roll, Shift, Slide, and Tilt,
  - MGW (minor groove width)
  - EP (minor groove electrostatic potential)
- 4 features for methylated DNA
  - HelT, MGW, ProT, Roll
- Quantified by DNAshapeR (R/Bioconductor)

Fig. 1. Schematic overview of the architecture and key functionality of TFBSshape: A motif database

https://tfbsshape. usc.edu/

Chiu T. et al. NAR, 2020



#### Fig. 2. Schematic illustration of the pentamer model for highthroughput prediction of DNA shape.





А

Pentamer query table entries:

Pentamer	MGW [Å]	Roll1 [°]	Roll2 [°]
TTTGA	5.60	-2.67	5.24
TTGAC	5.91	6.83	-0.77
TGACT	4.89	-1.77	-3.69
GACTT	3.74	-4.01	-4.07
ACTTC	4.31	-3.34	-4.20
CTTCA	5.15	-3.01	0.22

C Input sequence: TTTGACTTCA

Sliding window for intra-base pair parameter:



Chiu T. et al. NAR, 2020

#### Fig. 3. Search Interface in the TFBSshape Webtool

С

Quick Search TFs	Name	Species	Source ID	Domain	Publication
JASPAR	AMAG_00796	Allomyces macrogynus	UP00515	Fork_head	Nakagawa et al., PNAS 2013
UniPROBE	AMAG_02766	Allomyces macrogynus	UP00514	Fork_head	Nakagawa et al., PNAS 2013
pecies:	<u>HLHmgamma</u>	Drosophila melanogaster	UP01527	Myc-type, bHLH	Shokri et al., Cell Reports 2018
Choose Species	Max	Drosophila melanogaster	UP01549	Myc-type, bHLH	Shokri et al., Cell Reports 2018
omain: Choose Domain	MAML1	Homo sapiens	UP00485		Del Bianco et al., PLoS ONE 2010
per: Choose Paper 🗘	Mafb	Mus musculus	UP00045	BRLZ	Badis et al., Science 2009
ne: F Name	Mafg	Mus musculus	UP01375	BRLZ	Mariani et al., Cell Systems 2017
earch	Mafk	Mus musculus	UP00044	BRLZ	Badis et al., Science 2009
	Max	Mus musculus	UP00060	HLH	Badis et al., Science 2009
ipiSELEX-seq	Smad3	Mus musculus	UP00000	DWA	Badis et al., Science 2009
MeDReaders	MAL8P1.153	Plasmodium falciparum	UP00429	AP2	Campbell et al., PLoS Pathog 2010

Name	Max				
Species	Mus musculus				
Domain	НГН				
Reference	Badis et al., Science 2009				
PWM	O Max_3863_015681.bml.pwm				
DeBruijn	Max_3863.1_v1_deBruijn.txt				
Source	UP00060				
Motif Distribution	TFBS distribution       10000     20000     30000     40000       0 17BS     1 17BS     2×17BS     Ranked probes (high -> low signal intensity)				
Download	Unmethylated   Methylated				
Update					
Minor Groove Width					
Roll					

#### Fig. 3. Example search output by the TFBSshape



Fig. 4. Example TF compare interface in the TFBSshape Webtool



Fig. 4. Example Output of TF compare by TFBSshape

Name	MAX		Name	Мус		
Species	Homo sapiens		Species	Mus musculus		
Database	JASPAR		Database	JASPAR		
Download	Unmethylated   Methylated		Download	Unmethylated   Methylated		
Sequence Logo			Sequence Logo			
Reference Posi	tion 1	R	eference Positior 0	n 2		
Align by Refe	Align by Reference					



• Pearson correlation coefficient

$$r(\mathbf{p}, \mathbf{q}) = rac{\sum_{i=1}^{n} (p_i - ar{p})(q_i - ar{q})}{\sqrt{\sum_{i=1}^{n} (p_i - ar{p})^2} \sqrt{\sum_{i=1}^{n} (q_i - ar{q})^2}}$$

where:

- $ar{p}$  is the mean of vector  ${f p}:ar{p}=rac{1}{n}\sum_{i=1}^n p_i$
- $ar{q}$  is the mean of vector  ${f q}$ :  $ar{q}=rac{1}{n}\sum_{i=1}^n q_i$
- Euclidian distance

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \dots + (p_n-q_n)^2}$$

where:

• 
$$\mathbf{p} = (p_1, p_2, \dots, p_n)$$
  
•  $\mathbf{q} = (q_1, q_2, \dots, q_n)$ 

Fig5. Example Mutation Design Interface in the TFBSshape Webtool

$\bigcirc$	JASPAR
------------	--------

UniPROBE

TF Name:	TF Name
Search	



MAX | Homo sapiens

○ MAX | Homo sapiens

○ MAX::MYC | Homo sapiens



Select file(s) for TF 1MA0058.1

**Mutation Design** 



#### Fig5. Example Mutation Design Interface in the TFBSshape Webtool

Name	MAX	
Species	Homo sapiens	
Database	JASPAR	
File	TF ID: MA0058.1	
Input Sequence for Query (i)	GTGAgCACGTGgTT Wild sequence	e
Target	<ul><li>Keep sequence, change shape</li><li>Keep shape, change sequence</li></ul>	
Sequence Threshold (i)	60	
Shape Threshold (i)	80	
Maximum Number of Mutations Allowed (i)	2	
Number of Outputs	10	
Shape Feature(s) for Mutation Design:	☑ Minor Groove Width 🗹 Roll 🗹 Propeller Twist ☑ Helix Twist	

	Sequence	ls in the Pool	Sequence Distance	Shape Distance
Fig5.	GATAAAATGAGACCAgtgcatTATTAGTTGTACACC	No	0	0.0
Example	GATAAAATGAGACCAtcgcatTATTAGTTGTACACC	No	2	0.446928134427
Mutation	GATAAAATGAGACCAgtgcgtTATTAGTTGTACACC	No	1	0.476333023662
Design	GATAAAATGAGACCAgtgtatTATTAGTTGTACACC	No	1	0.499588222314
Output by	GATAAAATGAGACCAttgcatTATTAGTTGTACACC	No	1	0.507074315589
TFBSshape	GATAAAATGAGACCAgtacatTATTAGTTGTACACC	No	1	0.516044767294
	GATAAAATGAGACCAccgcatTATTAGTTGTACACC	No	2	0.518749448563

#### **Motif Analysis**

 The MEME Suite: Motif-based Sequence Analysis Tools: <u>https://meme-</u> <u>suite.org/meme/</u>

#### Motif-based sequence analysis tools **MEME Suite** 5.5.5 Mouse-over for information or Discovered Your DNA, RNA or Motif Discovery Sequence each software tool or resource. motifs protein sequences MEME databases Click to submit a job to the too (de novo) Jobs running: 4 STREME or to view database details. XSTREME Jobs waiting to MEME-ChIP Motif Scanning Annotated sequences run: 0 GLAM2 MoMo FIMO MAST Motif Discovery Motif Enrichment MCAST Enriched motifs SEA Motif MEME GLAM2SCAN CentriMo databases 3 -STREME AME Aligned motifs **Motif Comparison** SpaMo XSTREME Tomtom GOMo Your DNA, RNA or protein motifs MEME-ChIP Annotated motifs Your BED file of Regulatory gene GO function GLAM2 GO genomic loci targets GO compartm T-Gene databases GO process MoMo AA DREME (deprecated) MEME Multiple Em for Motif **O SEA** 0 ▶ Motif Simple Enrichment Eind Individual Motif Enrichment Elici Motif Scanning STREME AME analysis of Motif AST tive Thorough, Actif Alignment & ▼Motif Comparison TREME CentriMo Tomtom otif Cluster Alignment Discovery and Gene GLAM2Scan Scanning with Gapped SpaMo Spaced Motif Analysis MEME-ChP Regulation Utilities GLAM2 Gapped Local Alignment GOMO Gene Ontology for Motifs Manual Predicting Target Genes Guides & GT-Scan dentifying Unique MoMo Modification Motifs Tomtom Motif Comparison Tool Tutorials ► Sample EBED2FASTA Outputs DREME Discriminative Regular File Format Reference

51

The MEME Suite



#### MEME Suit

#### (Bailey T.L., NAR, 2015)

	North LO	ment	ang ang	pairson		
Que	450	Scar	Cor	Tool	Ref.	Description
~				MEME	(2)	Discovers novel, ungapped motifs (recurring, fixed-length patterns) in nucleotide or protein sequences; MEME splits variable-length patterns into two or more separate motifs
~				DREME	(3)	Discovers short, ungapped motifs (recurring, fixed-length patterns) that are relatively enriched in your nucleotide sequences compared with shuffled sequences or your control sequences
~	1		~	MEME-ChIP	(4)	Performs comprehensive motif analysis (including motif discovery) on large (50MB maximum) sets of nucleotide sequences such as those identified by ChIP-seq or CLIP-seq experiments
1				GLAM2	(5)	Discovers novel, gapped motifs (recurring, variable-length patterns) in DNA or protein sequences
	~			CentriMo	(6)	Identifies known or user-provided motifs that show a significant preference for particular locations in nucleotide sequences; CentriMo can also show if the local enrichment is significant relative to control sequences
	~			AME	(7)	Identifies known or user-provided motifs that are relatively enriched in nucleotide sequences compared with shuffled sequences or control sequences; AME treats motif occurrences the same, regardless of their locations within the sequences
	1			SpaMo	(8)	Identifies significantly enriched spacings in a set of sequences between a primary motif and each motif in a set of secondary motifs; typically, the input sequences are centered on ChIP-seq peaks
	~			GOMo	(9)	Scans all promoters using nucleotide motifs you provide to determine if any motif is significantly associated with genes linked to one or more Genome Ontology (GO) terms; the significant GO terms can suggest the biological roles of the motifs
		~		FIMO	(10)	Scans a nucleotide or protein sequence database for individual matches to each of the motifs you provide
		× .		MAST	(11)	Searches sequences for matches to a set of nucleotide or protein motifs and sorts the sequences by the best combined match to all motifs
		1		MCAST	(12)	Searches sequences for clusters of matches to one or more nucleotide motifs
		1		GLAM2Scan	(5)	Searches sequences for matches to gapped DNA or protein GLAM2 motifs
			~	Tomtom	(13)	Compares one or more nucleotide motifs against a database of known motifs such as JASPAR (14); Tomtom will rank the motifs in the database and produce an alignment for each significant match

# Motif Discovery by DREME:

To identify overrepresented motifs

Input: A set of unaligned DNA, RNA, or protein sequences, e.g., ChIPseq peak regions.

DREME (deprecated; please consider usina MEME Suite instead) STREME iscriminative Regular Expression Motif Elicitation short. 5.5.5 discovers ungapped motifs Version 5.5.5 fixed-Jobs running: 7 (recurring. length patterns) that are relatively Jobs waiting to enriched in your sequences run: 0 compared with shuffled sequences Motif Discovery or your control sequences (sample output from sequences). See this ► Motif Manual or this Tutorial for more Enrichment information. Motif Scanning Data Submission Form ► Motif Perform motif discovery on DNA or RNA datasets for short regular Comparison expression motifs. ► Gene Regulation Select the type of control sequences to use Utilities Shuffled input sequences ○ User-provided sequences ? Manual Select the sequence alphabet Guides & Use sequences with a standard alphabet or specify a custom Tutorials alphabet. ? Sample • DNA, RNA or Protein O Custom Browse... No file selected. Outputs ► File Format Input the sequences Reference Enter sequences in which you want to find motifs ? Databases Upload sequences V Browse... No file selected. Download & Install Input job details ► Help (Optional) Enter your email address. 2 ► Alternate Servers (Optional) Enter a job description. 2 Authors & Citing Recent Jobs Advanced options ← Previous Note: if the combined form inputs exceed 80MB the job will be version 5.5.4 rejected. Start Search Clear Input

Version 5.5.5 Please send comments and questions to: meme- Powered by Opal suite@uw.edu

#### DREME

- DREME (Discriminative Regular Expression Motif Elicitation): <u>https://meme-suite.org/meme/tools/dreme</u>
  - A motif discovery algorithm designed to find short, core DNA-binding motifs of eukaryotic transcription factors and is optimized to handle large ChIP-seq data sets.
  - Tailored to eukaryotic data by focusing on short motifs (4 to 8 nucleotides) encompassing the DNA-binding region of most eukaryotic monomeric transcription factors.
  - Therefore it may miss wider motifs due to binding by large transcription factor complexes.

#### Motif Discovery Algorithms

- meme is a general purpose motif discovery algorithm for both nucleotide and peptide motifs, but is less sensitive than DREME for finding short nucleotide motifs.
- Neither meme nor **DREME** allows insertions or deletions in the motifs they find, but glam2 does.
- meme-chip is adapted to very large datasets that cannot be handled by meme, and it actually performs motif discovery, motif enrichment and motif comparison on its input sequences, producing a fully integrated report. A comprehensive protocol for using meme-chip has recently been published (Ma W. et al, Naat. Protoc. 2014).

# MEME-ChIP

- Part of the MEME Suite that is specifically designed for <u>ChIP-seq</u> <u>analyses.</u>
- Performs DREME and Tomtom analysis
- Assess which motifs are most centrally enriched (motifs should be centered in the peaks)
- Combine related motifs into similarity clusters.
- It is able to identify longer motifs < 30bp, but takes much longer to run.</li>



### Tomtom

- Tomtom: <u>https://meme-</u> <u>suite.org/meme/tools/tomtom</u>
- Determine if the identified motifs resemble the binding motifs of known transcription factors.
- Tomtom searches a database of known motifs to find potential matches and provides a statistical measure of motifmotif similarity.

Tomtom compares omtom one or more motifs against a database of **MEME Suite** known motifs (e.g., JASPAR). Tomtom 5.5.5 will rank the motifs in Version 5.5.5 the database and Jobs running: 3 produce an alignment for each Jobs waiting to significant match (sample output for run: 0 motif and JASPAR CORE 2014 database). See this Manual for Motif Discovery more information. Motif Enrichment Data Submission Form Motif Scanning Search one or more motifs against a motif database. ▼Motif Comparison Input query motifs Tomtom Enter the motif(s) to compare to known motifs. 2 Type in motifs V DNA V DNA ? Gene Regulation Utilities Example Output Manual Select target motifs Select a motif database or provide motifs to Guides & Tutorials compare with. ? Sample Outputs File Format Reference DNA ? Eukaryote DNA  $\sim$ Databases Vertebrates (In vivo and in silico) × ? Download & Allow alphabet expansion. 2 Install Run immediately Help Search with one motif (faster queue). 2 Alternate Servers Input job details Authors & (Optional) Enter a job description. 2 Citing Recent Jobs Advanced options → Previous version 5.5.4 Note: if the combined form inputs exceed 80MB the job will be rejected. Start Search Clear Input

#### TOMTOM: predict which proteins may bind a DNA motif



58

#### Web Resources

- Intro to ChIPseq using HPC by Harvard Chan Bioinformatics Core
  - https://hbctraining.github.io/Intro-to-ChIPseq/
- Interactive Analysis of RNA-seq and ChIP-seq:
  - <u>https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/integrating\_rna-seq\_and\_chip-seq.html</u>
- TFBSshape:
  - <u>https://tfbsshape.usc.edu/Home</u>
- MEME Suite:
  - https://meme-suite.org/meme/