ChiP-Seq Analysis Pipeline

Lecture 4

Outline

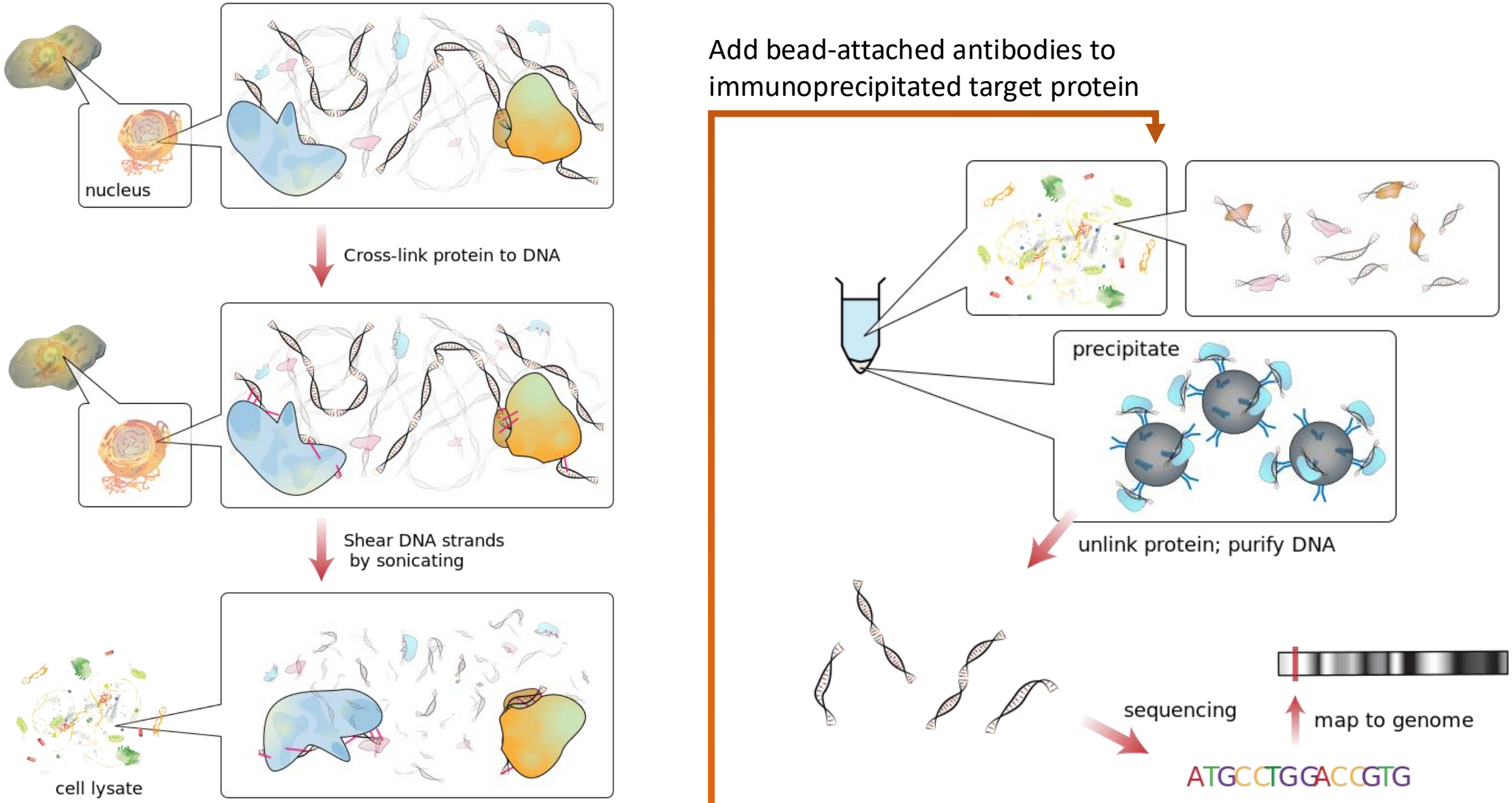
- ChIP-seq
- ChIP-seq Analysis Pipeline
- Transcription Factor Binding Site (TFBS) Motif Discovery

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

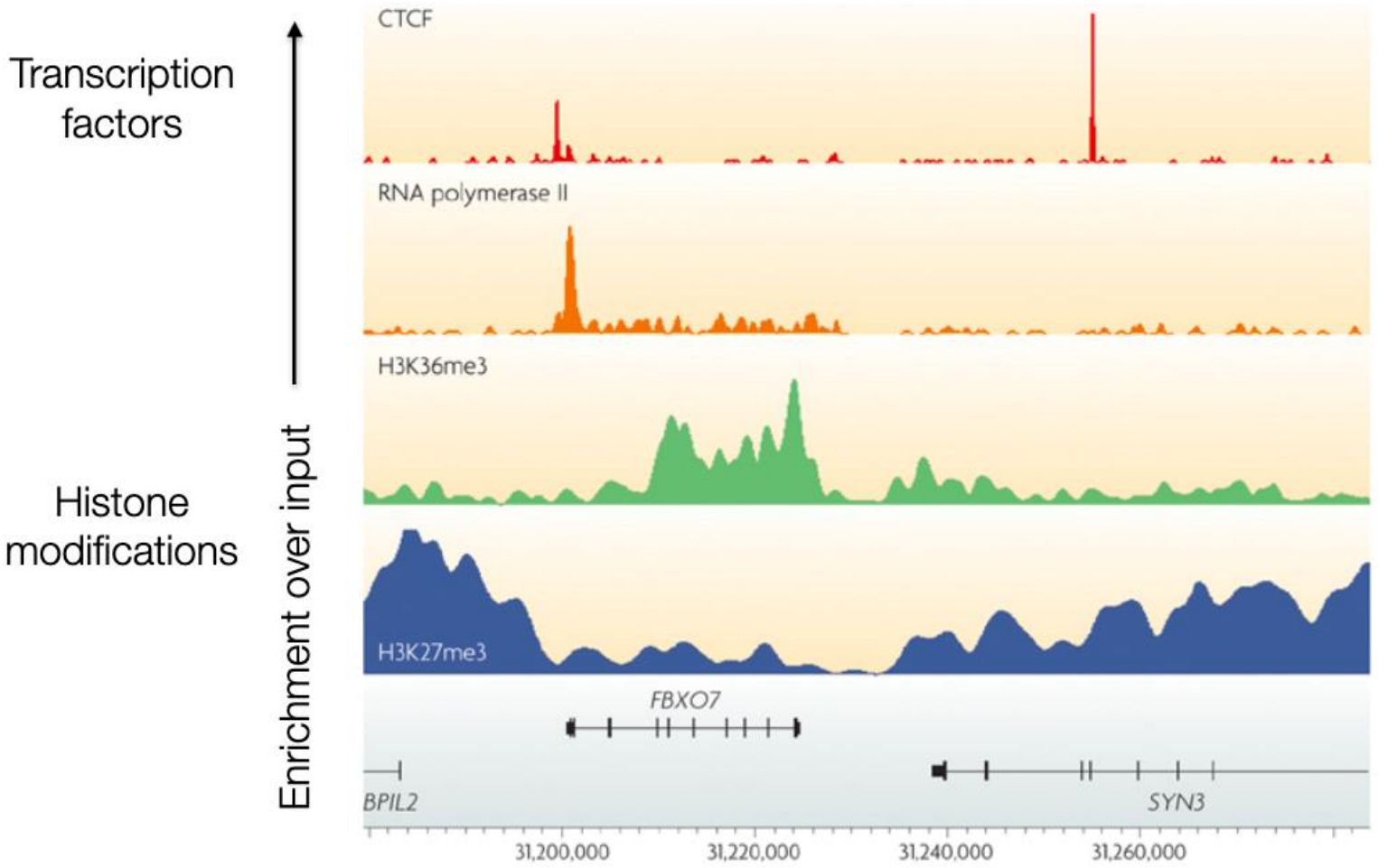
How Does ChIP-Seq Work?

- 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 DNA-protein complex.
- The DNA is extracted and sequenced, giving high-resolution sequences of the protein-binding sites.



ChIP-seq Workflow (wikipedia)

Types of Signals by ChIP-seq



Adapted from Park (2009). Nature Reviews Genetics.

Advantages of ChIP-Seq

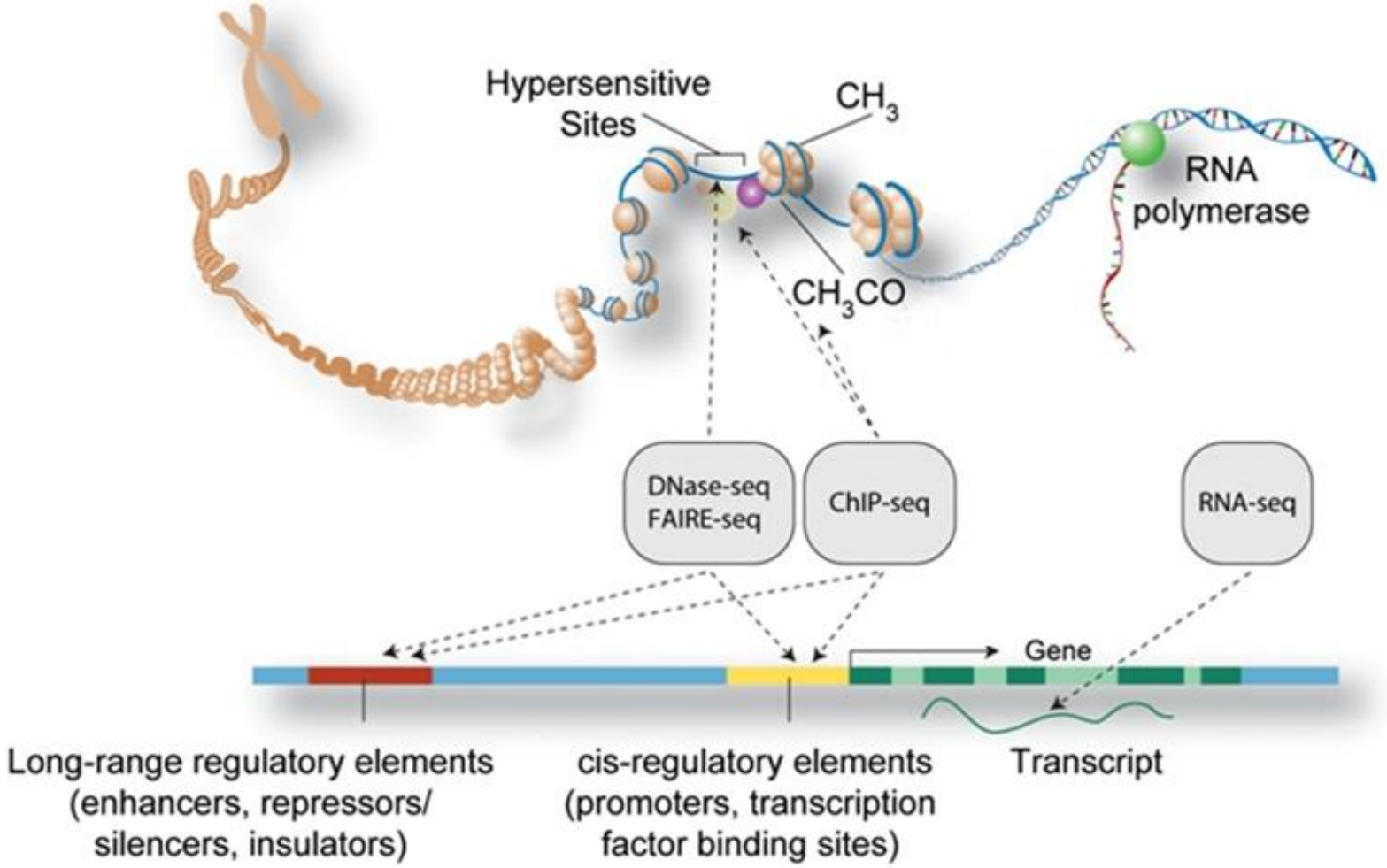
- ChIP-Seq delivers genome-wide profiling with massively parallel sequencing, generating millions of counts across multiple samples for cost-effective, precise, unbiased investigation of epigenetic patterns.

- Captures DNA targets for transcription factors, histone modifications, or nucleosomes 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



ENCODE Project Consortium 2011. *PLoS Biol* 9:e1001046.

Overview of ChIP–seq Analysis

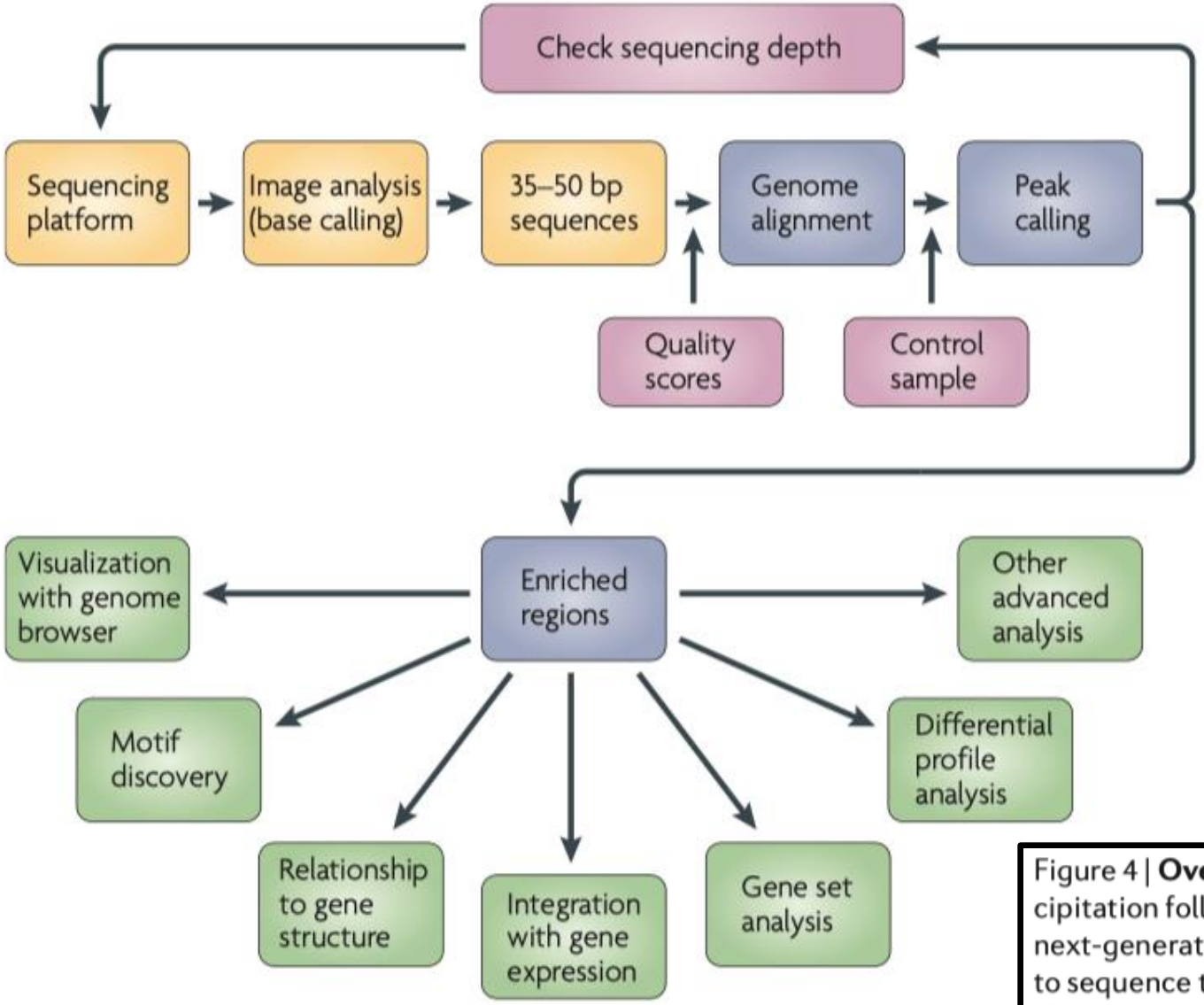
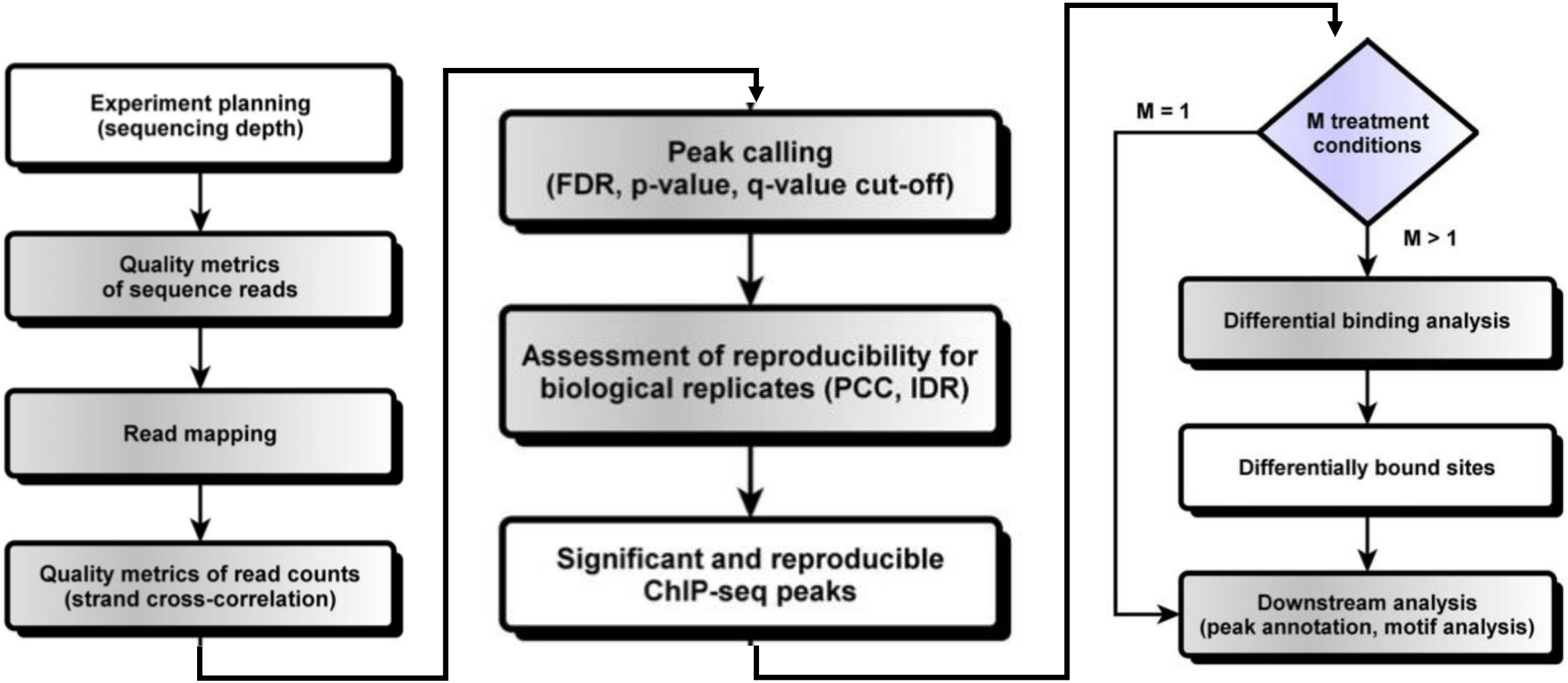


Figure 4 | **Overview of ChIP–seq analysis.** The raw data for chromatin immunoprecipitation followed by sequencing (ChIP–seq) analysis are images from the next-generation sequencing platform (top left). A base caller converts the image data to sequence tags, which are then aligned to the genome. On 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.

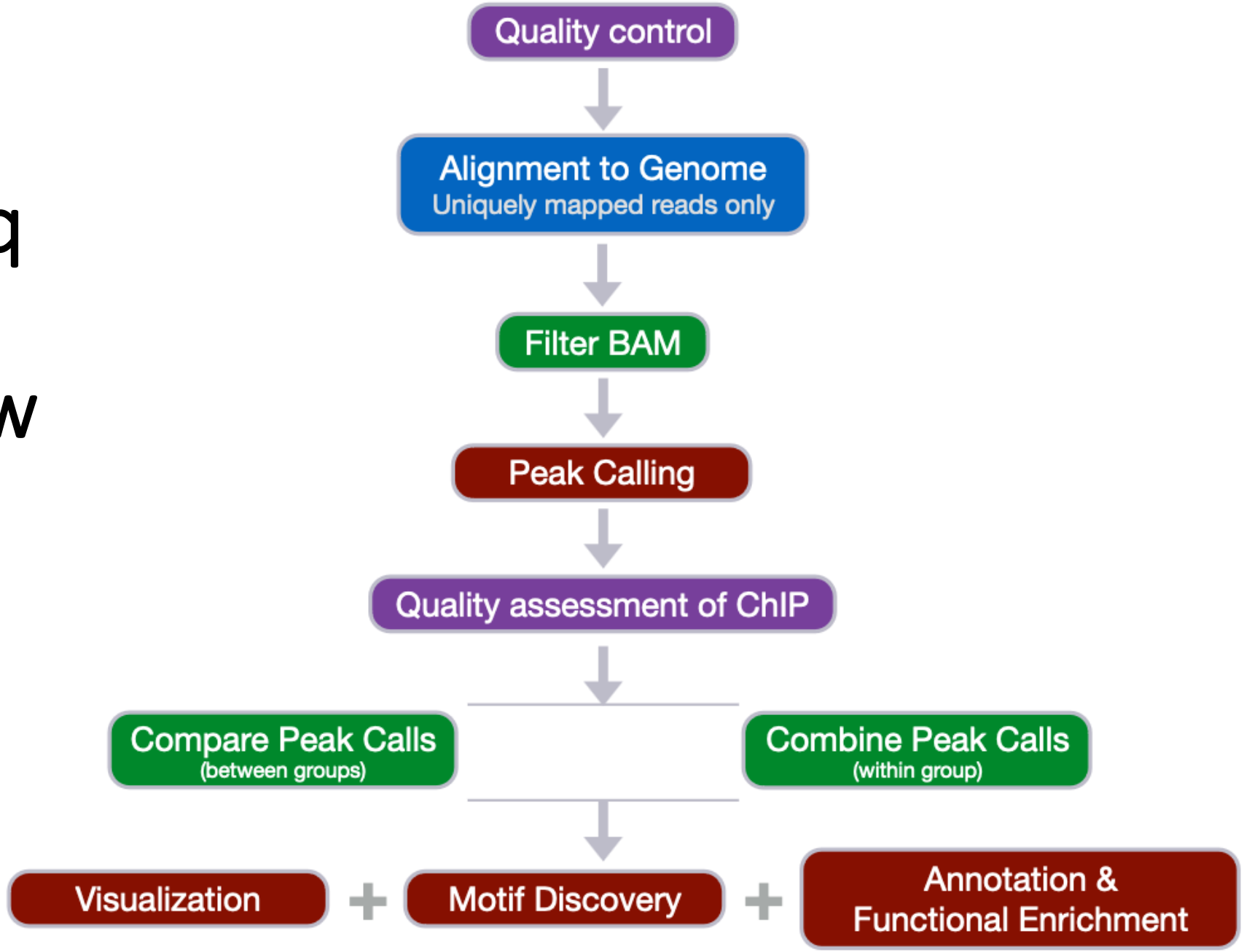
Park J.P. Nat. Reviews 2009.

ChIP–seq Analysis Workflow

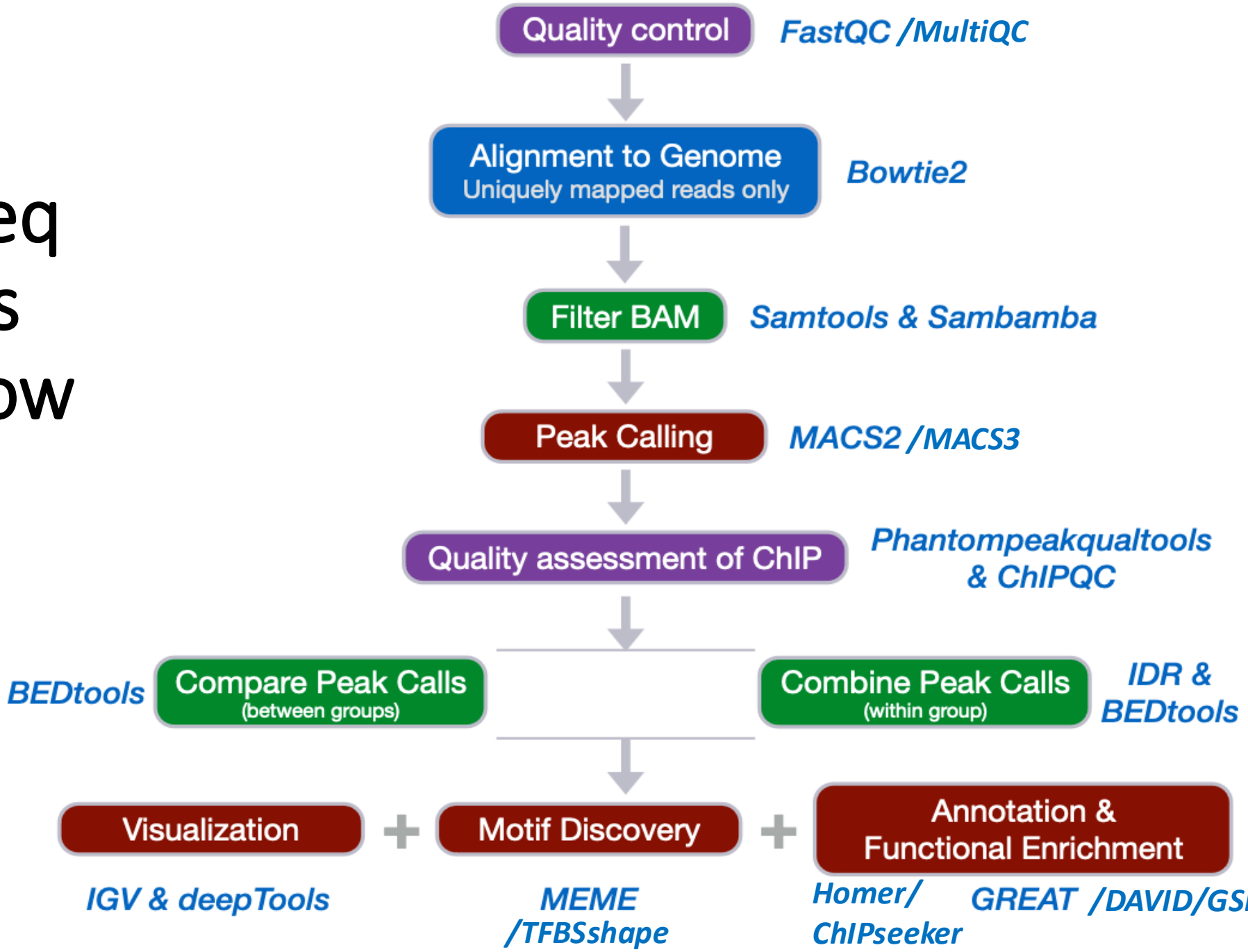


Adapted from Bailey T. et al, PLOS Comp. Bio. 2013

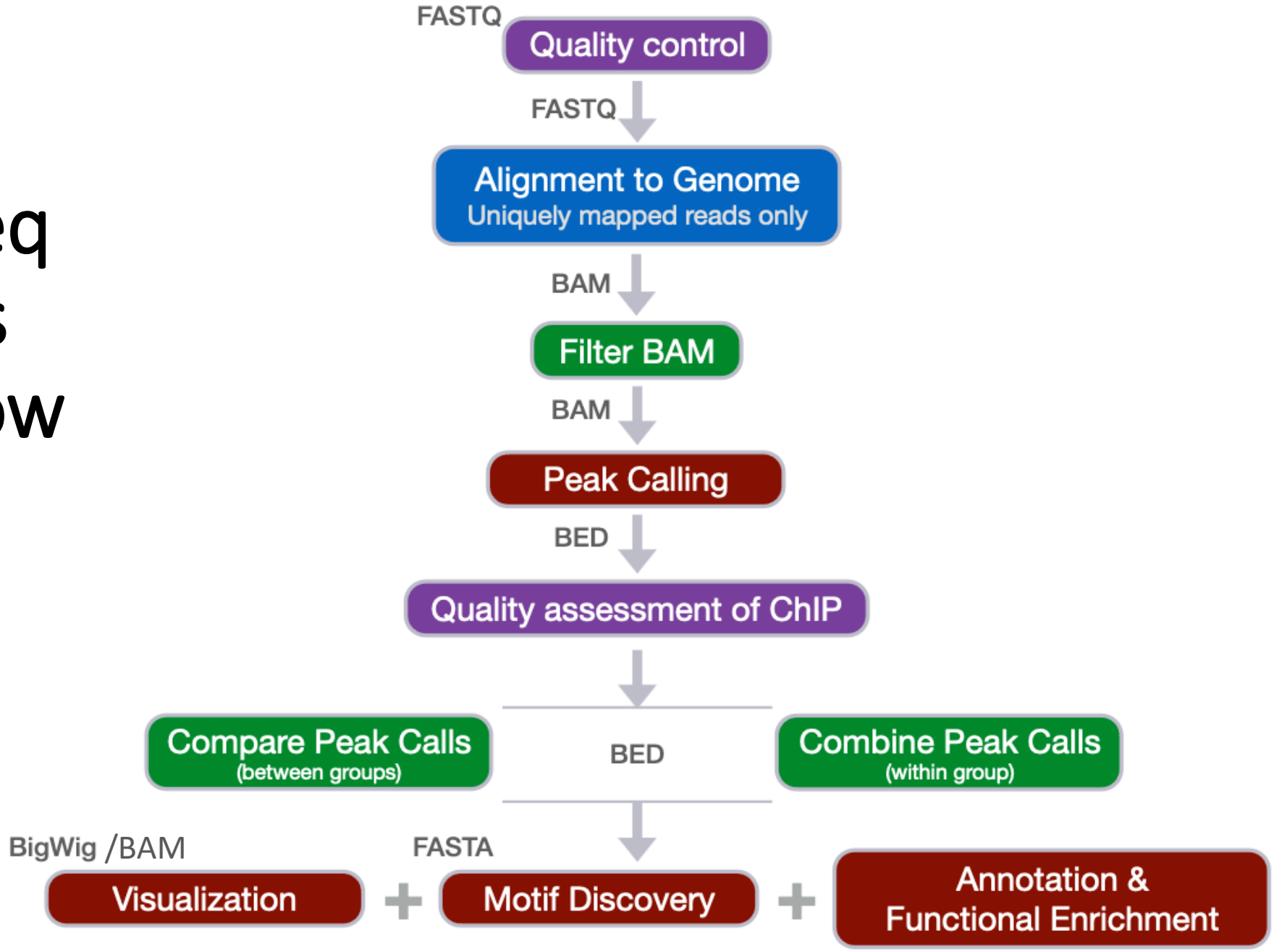
ChIP–seq Analysis Workflow



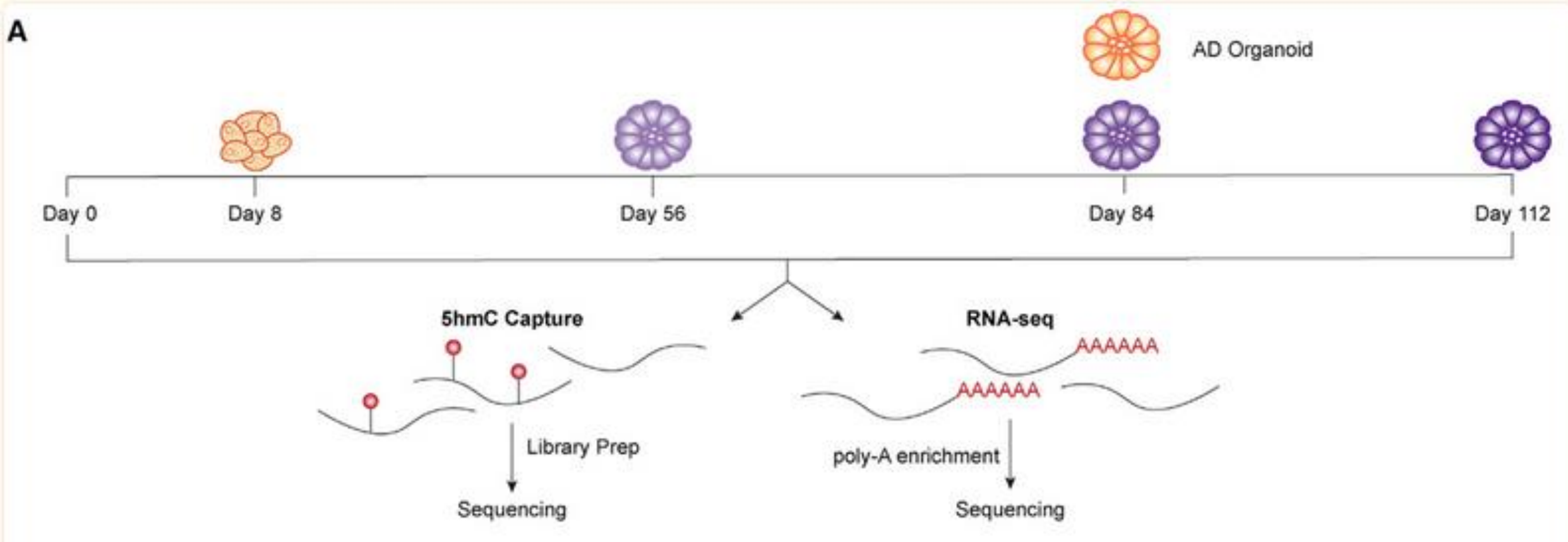
ChIP–seq Analysis Workflow



ChIP–seq Analysis Workflow



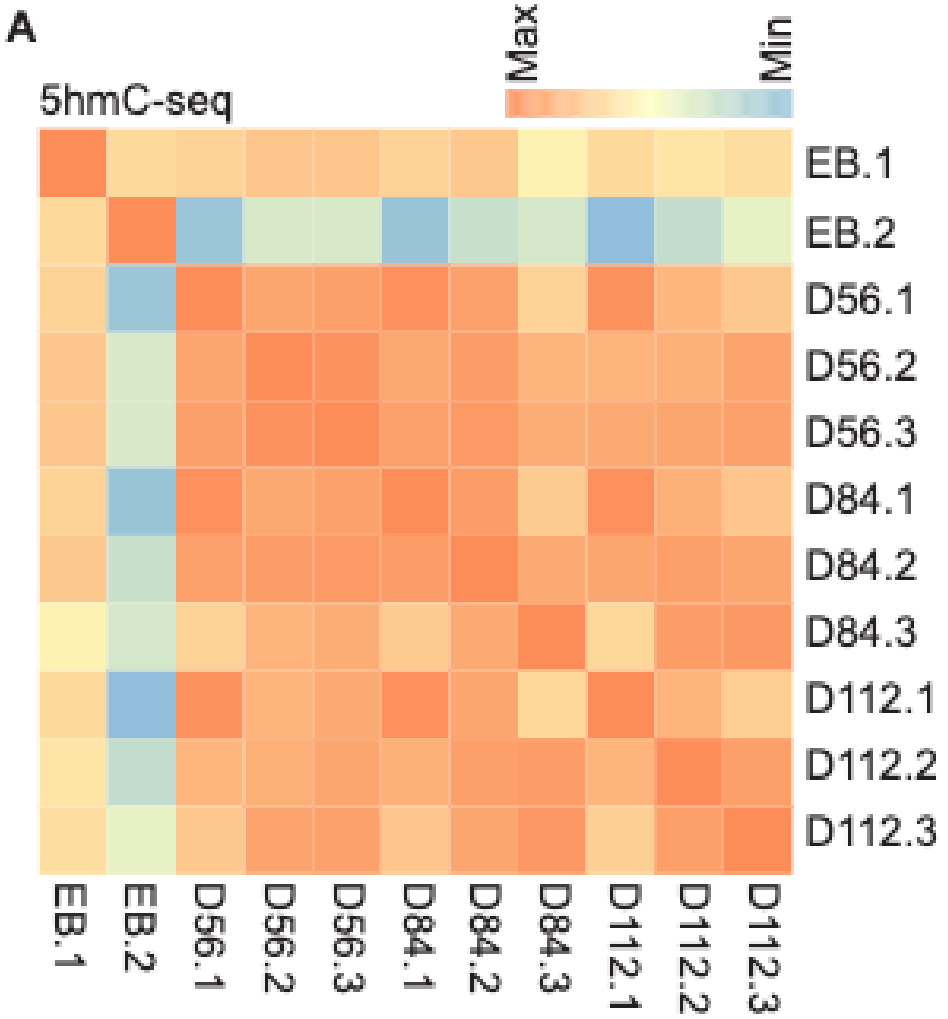
Example Study Design



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

Evaluate Data Qualify

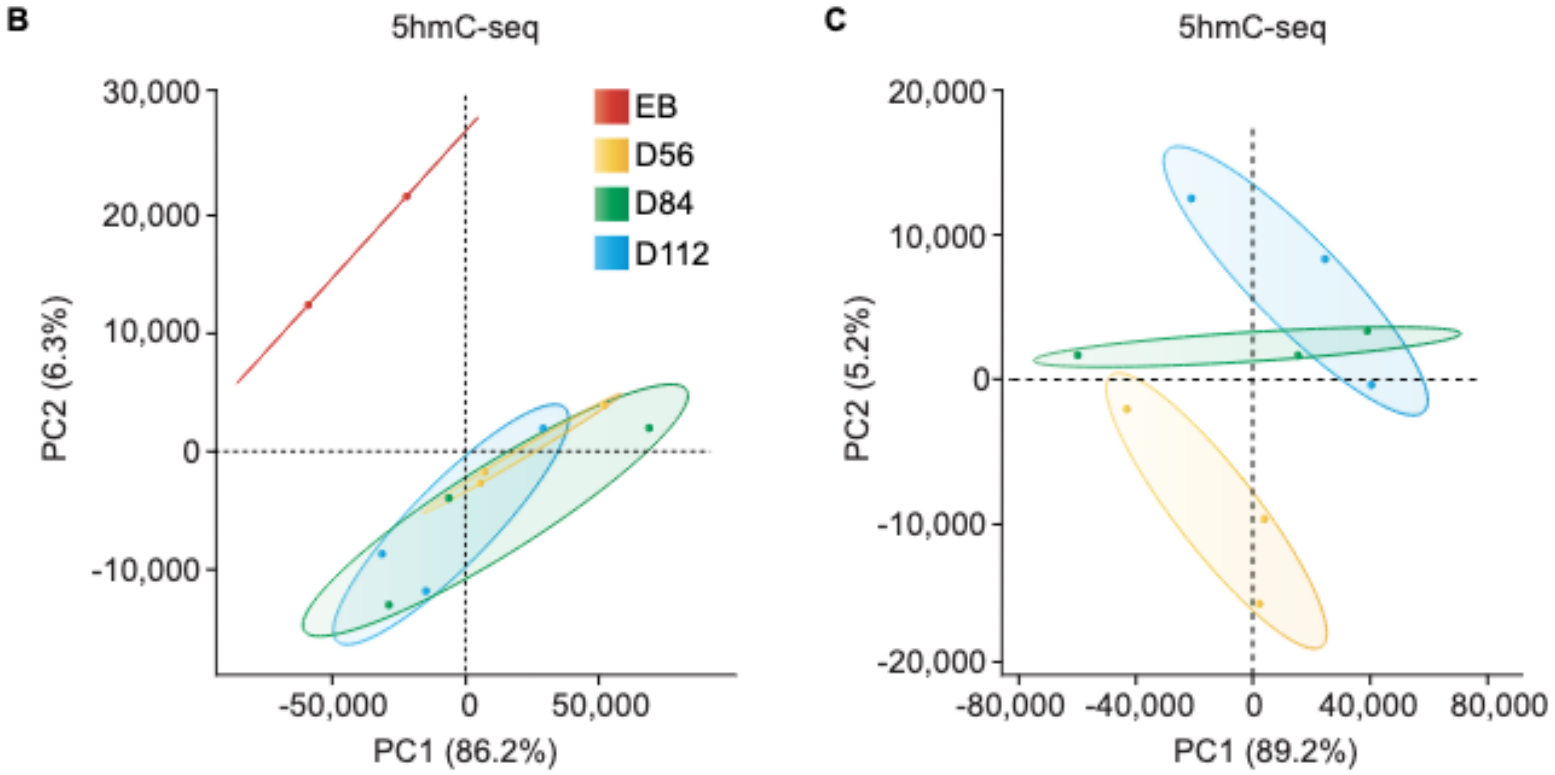
- **QC of samples with mapped data**: Pearson correlations with read counts at each genomic position.



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

Evaluate Data Qualify

- **QC of samples with mapped data**: Principal Components Analysis (**PCA**) with read counts at each genomic position.



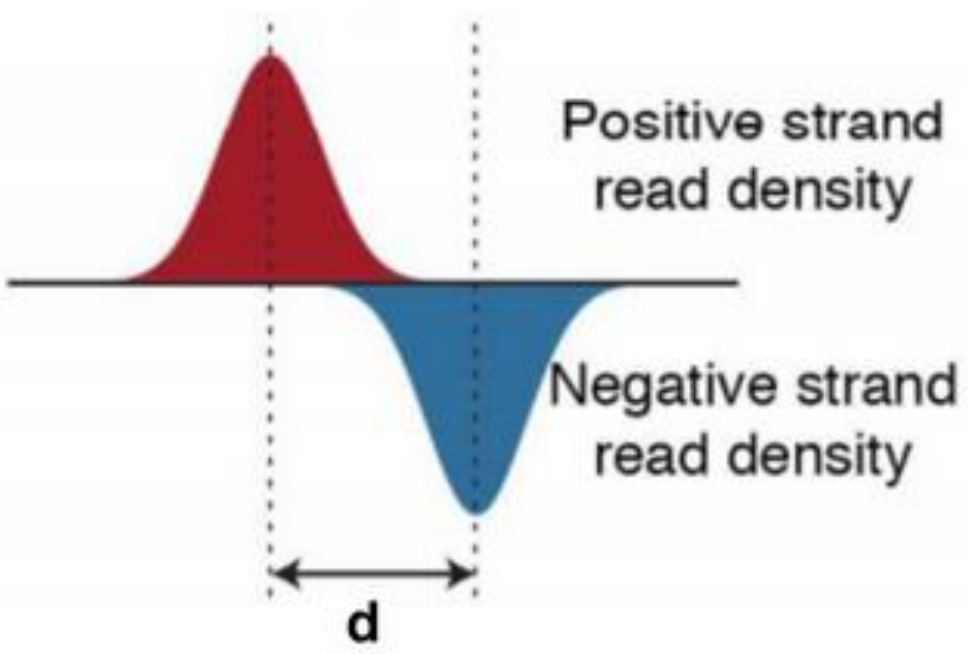
Kuehner J.N. et al, Cell Rep. 2021

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 influence of genome complexity to evaluate the significance of enriched ChIP regions
  - Improves the spatial resolution of binding sites through combining the information of both sequencing tag position and orientation.
  - 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**: https://macs3-project.github.io/MACS/

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

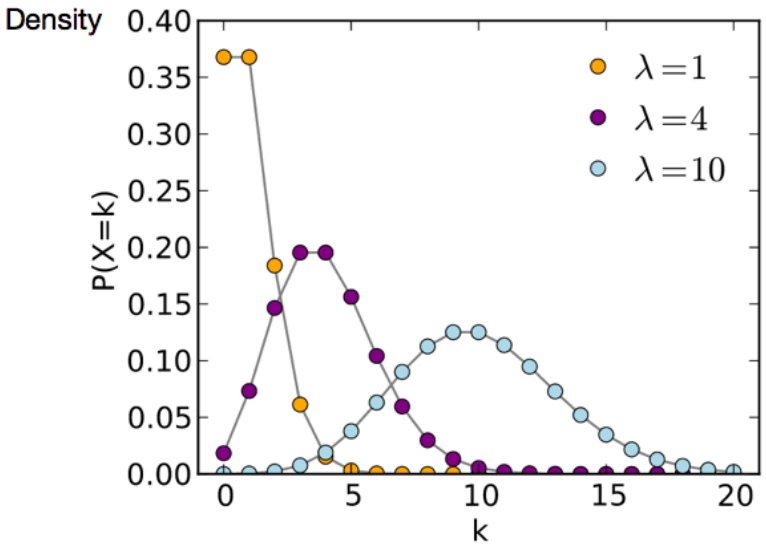
- Slides across the genome using a window size of *2d* to find candidate peaks.
- The tag distribution along the genome can be modeled by a Poisson distribution.
- The Poisson is a one parameter model, where the parameter **λ is the expected number of reads in that window**.

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

λ = mean = expectated value = variance

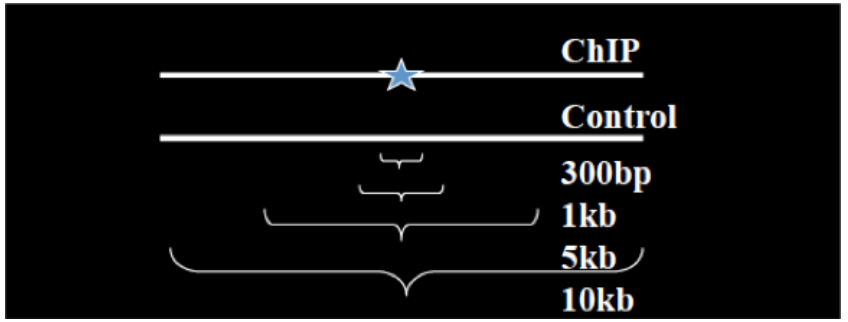
$$\lambda = \frac{\text{total number of events (k)}}{\text{number of units (n) in the data}}$$

$$= \frac{\text{Read length (nt)} * \text{Total read number}}{\text{Effective genome length (nt)}}$$



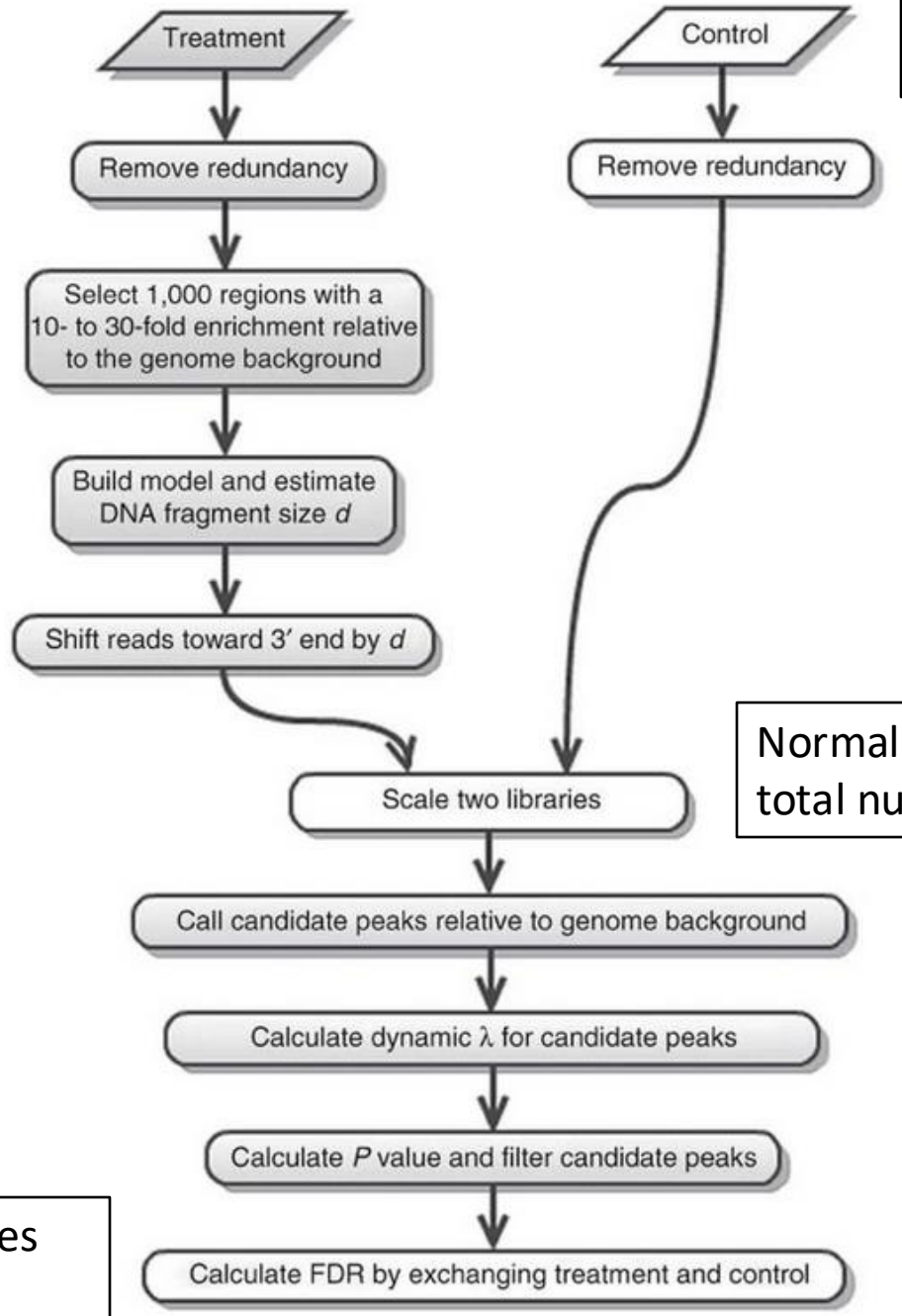
http://en.wikipedia.org/wiki/Poisson_distribution

MACS: Peak Detection



- MACS uses a dynamic parameter, λ_local, defined for each candidate peak. The λ_local parameter is estimated from the control sample (NULL) 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 λ_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%.
- Overlapping enriched peaks are merged, and each tag position is extended 'd' bases from its center. The location in the peak with the highest fragment pileup, hereafter referred to as the summit, is predicted as the precise binding location.
- The ratio between the ChIP-seq tag count and λ_local is reported as the fold enrichment.

MACS

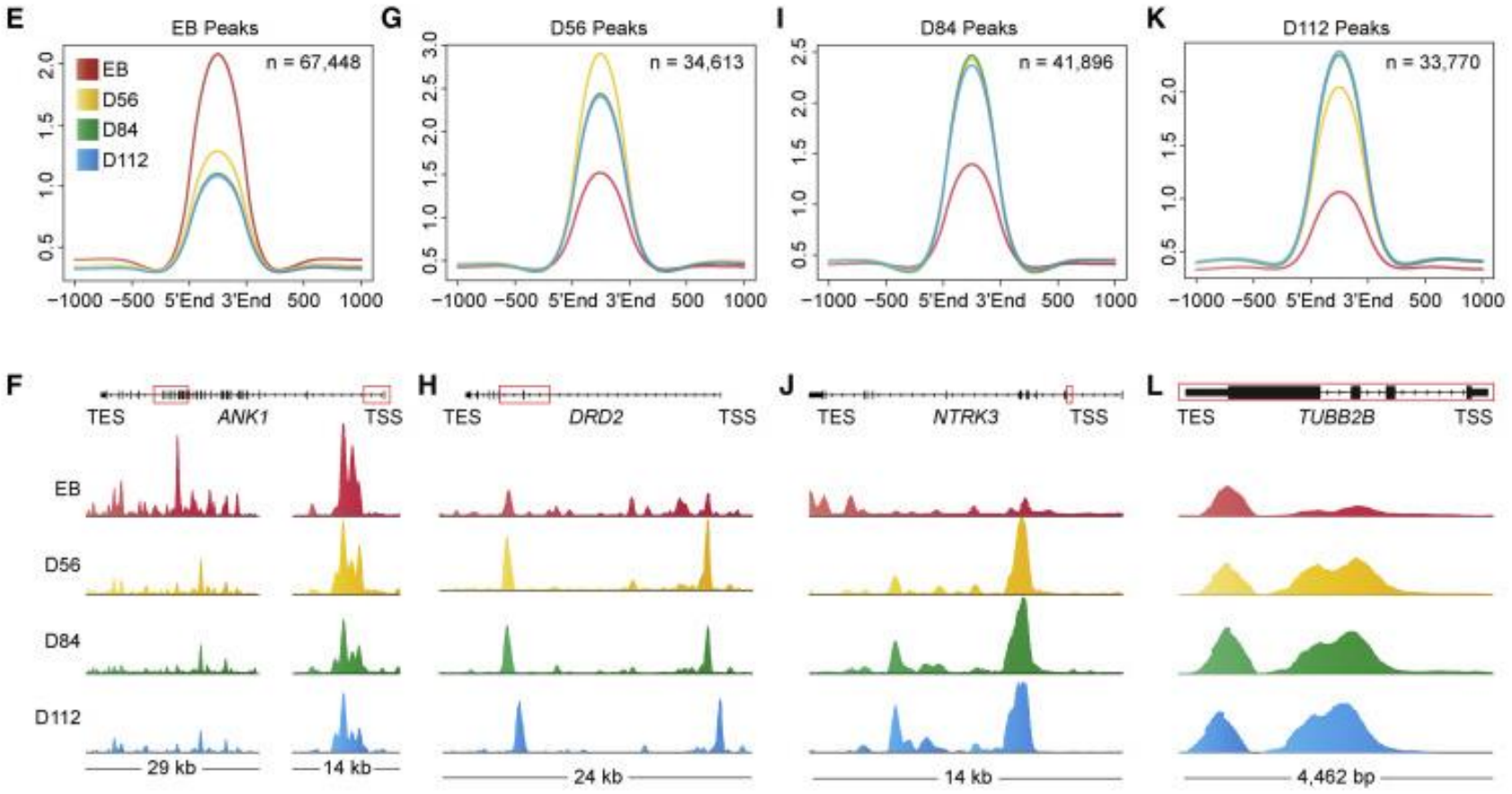


Control/Input Sample:
Without ImmunoPrecipitation

Normalize read counts to the same total number of reads per sample.

MACS3 obtains FDR from p-values across all called peaks.

Peak Visualization



Average 5hmC read counts were plotted globally by ngsplot: https://github.com/shenlab-sinai/ngsplot

Visualize Peaks by IGV: https://igv.org/

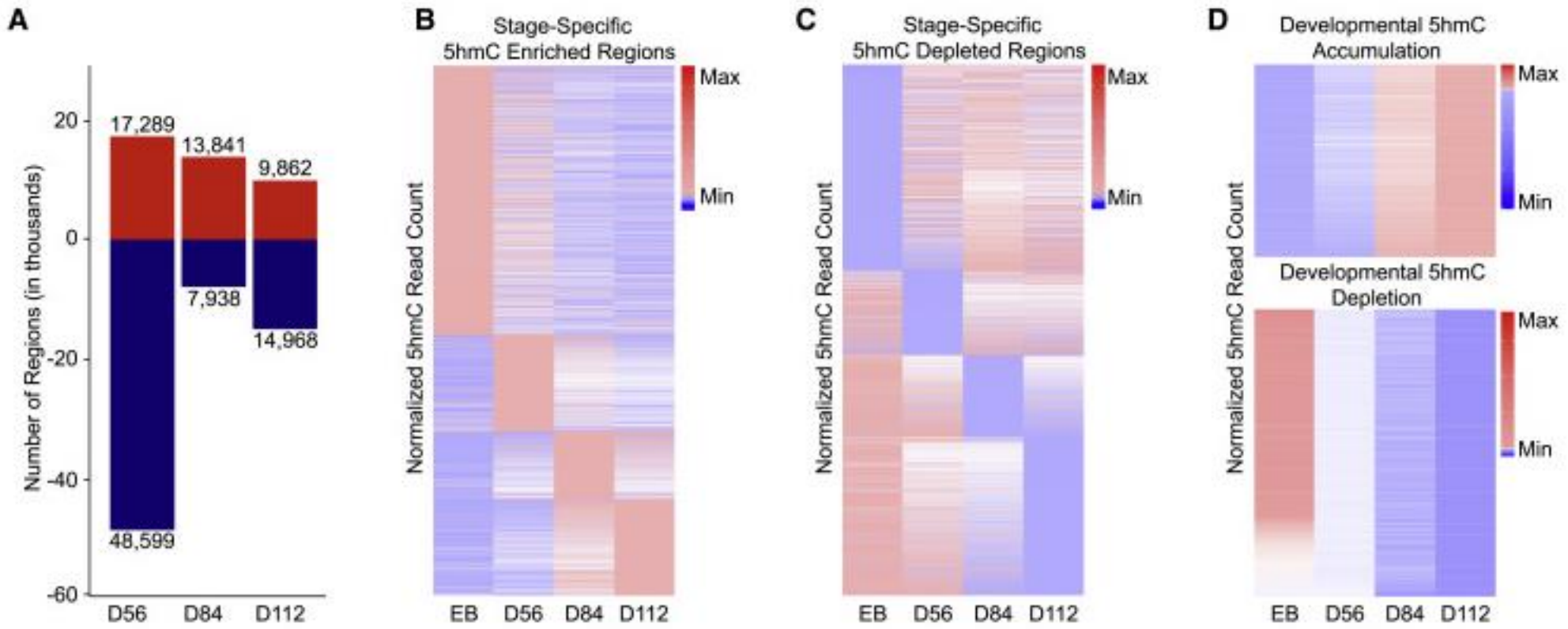
Kuehner J.N. et al, Cell Rep. 2021

Reads of multiple replicates per condition 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)
- https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/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

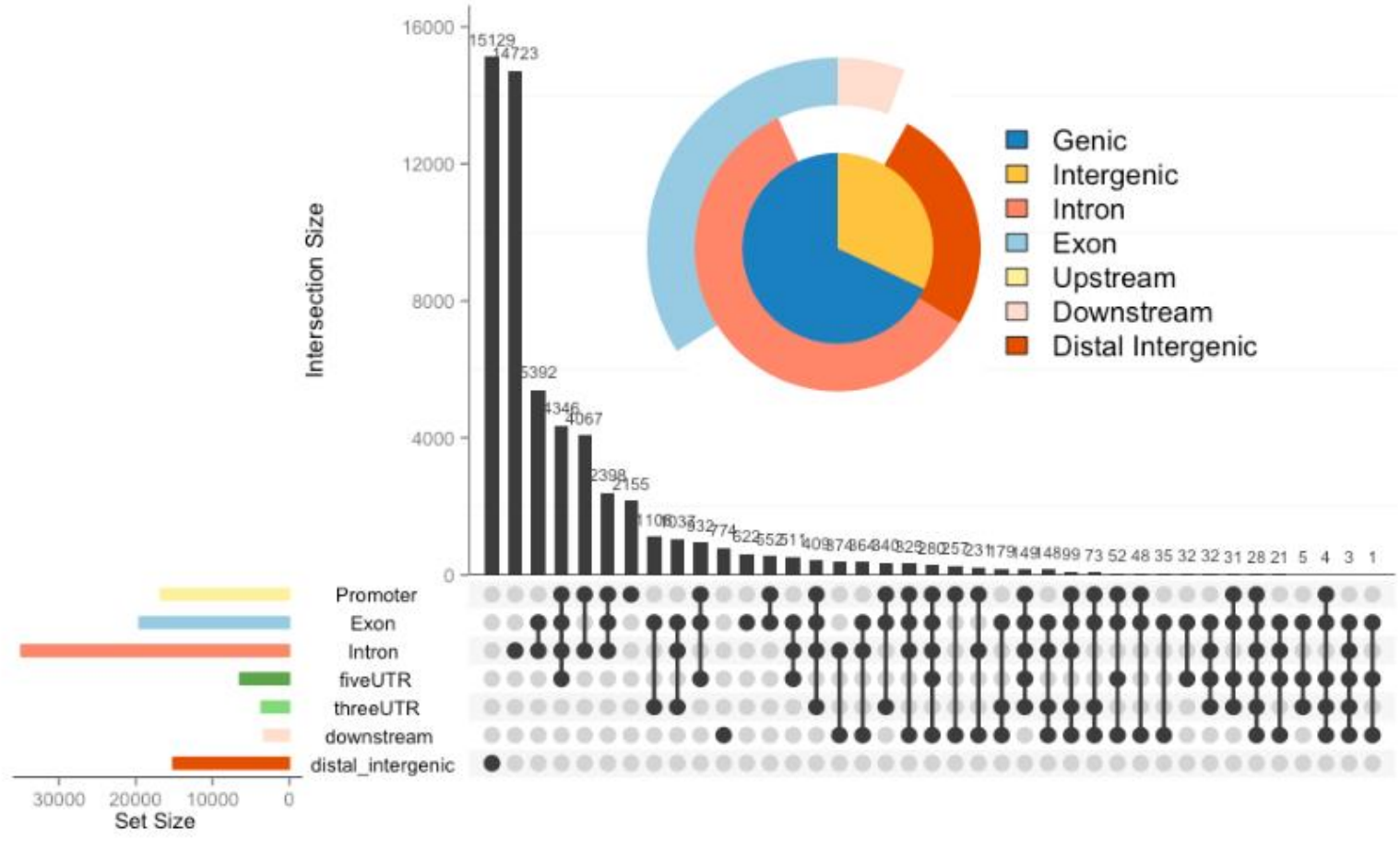


Kuehner J.N. et al, Cell Rep. 2021

(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

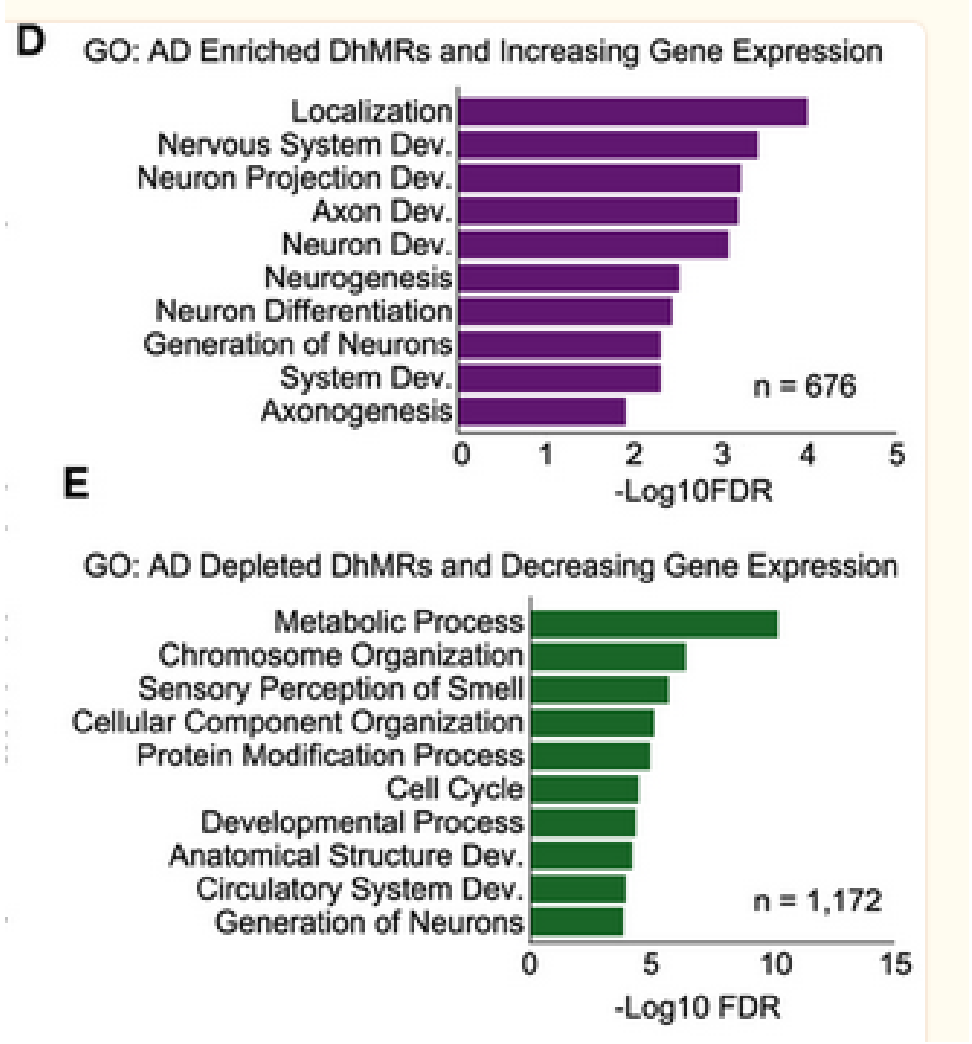
Associate the ChIP-seq peaks with functionally relevant genomic regions, such as gene promoters, transcription start sites, intergenic regions, etc.



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**: https://www.gsea-msigdb.org/gsea/index.jsp



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

What Are Motifs?

- A motif in molecular biology is a relatively short sequence of nucleotides or amino acids that changes little during evolution and, at least presumably, has a definite biological function.
- A motif is sometimes meant not a specific sequence, but a spectrum of sequences described in some way, each of which is capable of performing a certain biological function of a given motif.
- Motifs are ubiquitous in living organisms 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 predicting the functions of genes and proteins, constructing regulation maps, and are important for many problems in genetic engineering and molecular biology in general.

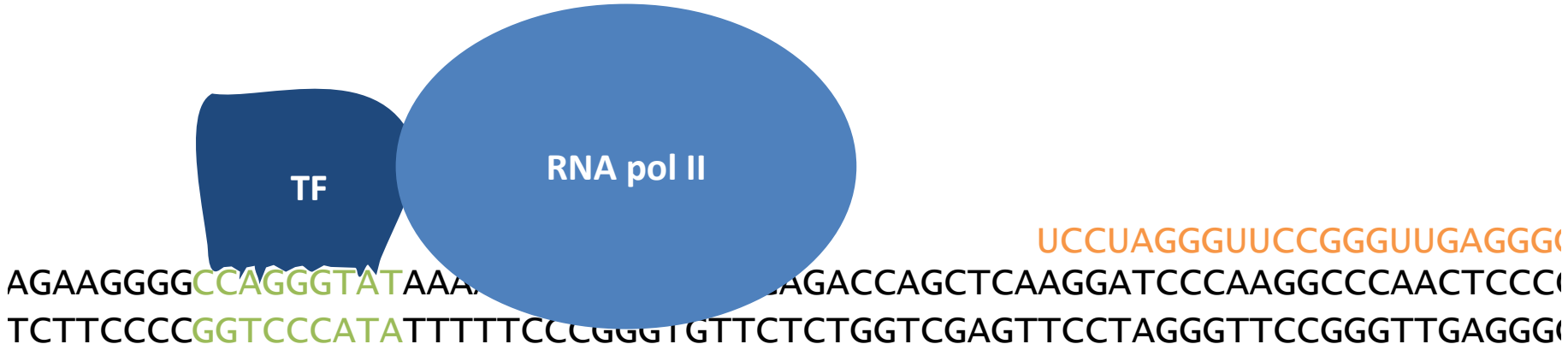
Motifs in Nucleic Acids

- 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.

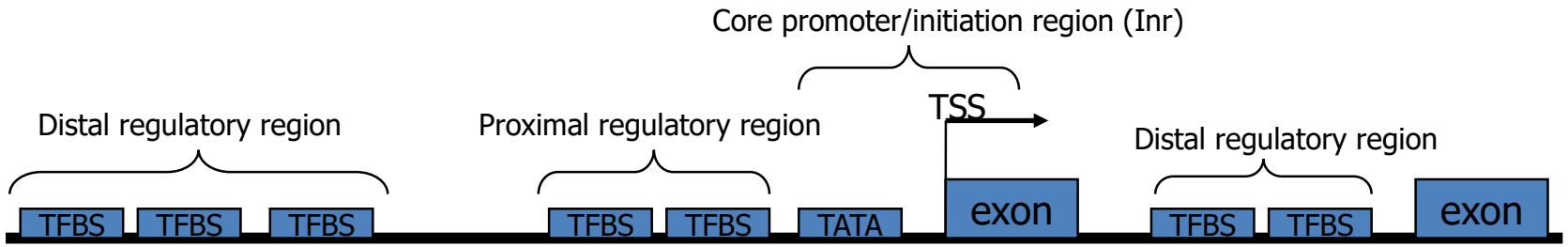
Transcription over-simplified

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



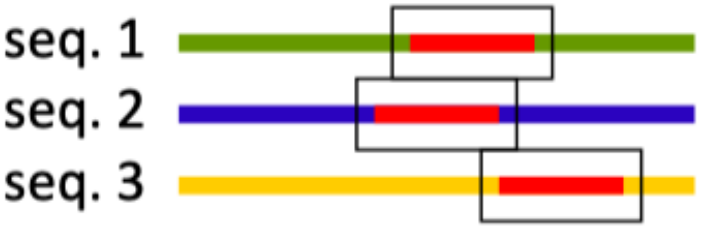
- Core promoter – Sufficient for initiation of transcription; orientation dependent
  - TSS – transcription start site
    - Often really a transcription start region
- TFBS – single transcription factor binding site
- Regulatory regions
  - Proximal/distal – vague reference to distance from TSS
  - 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

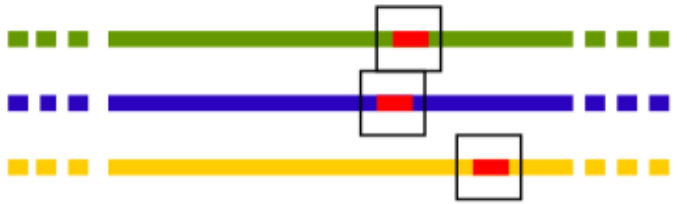


Motif Discovery Problem

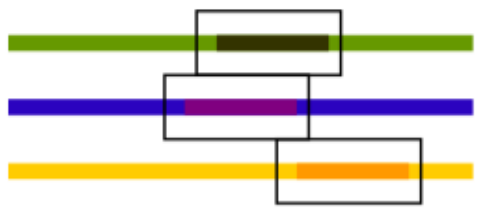
- Given:
  - a sequence or family of sequences.
- Find:
  - the number of motifs
  - the width of each motif
  - the locations of motif occurrences

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.



?

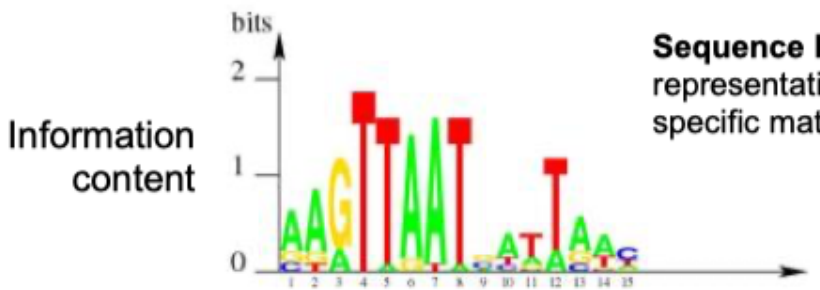


?

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)

A	14	16	4	0	1	19	20	1	4	13	4	4	13	12	3
C	3	0	0	0	0	0	0	0	7	3	1	0	3	1	12
G	4	3	17	0	0	2	0	0	9	1	3	0	5	2	2
T	0	2	0	21	20	0	1	20	1	4	13	17	0	6	4



Sequence logo: graphical representation of position-specific matrix.

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

DNA Shape Features of TFBSs

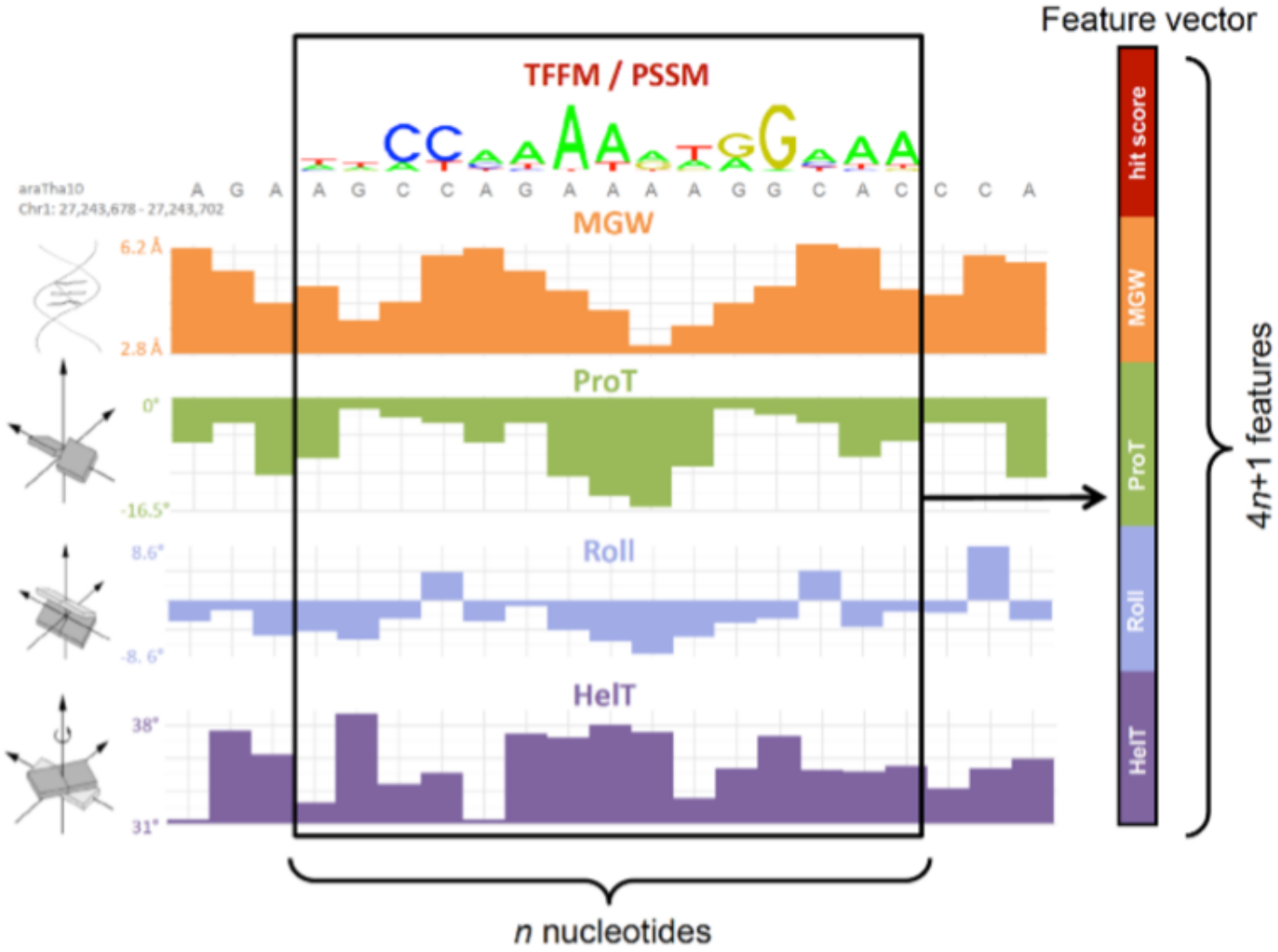
- Minor Groove Width (MGW)
- Roll
- Propeller Twist (ProT)
- Helix Twist (HelT)

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)

Chiu T. et al. NAR, 2020

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

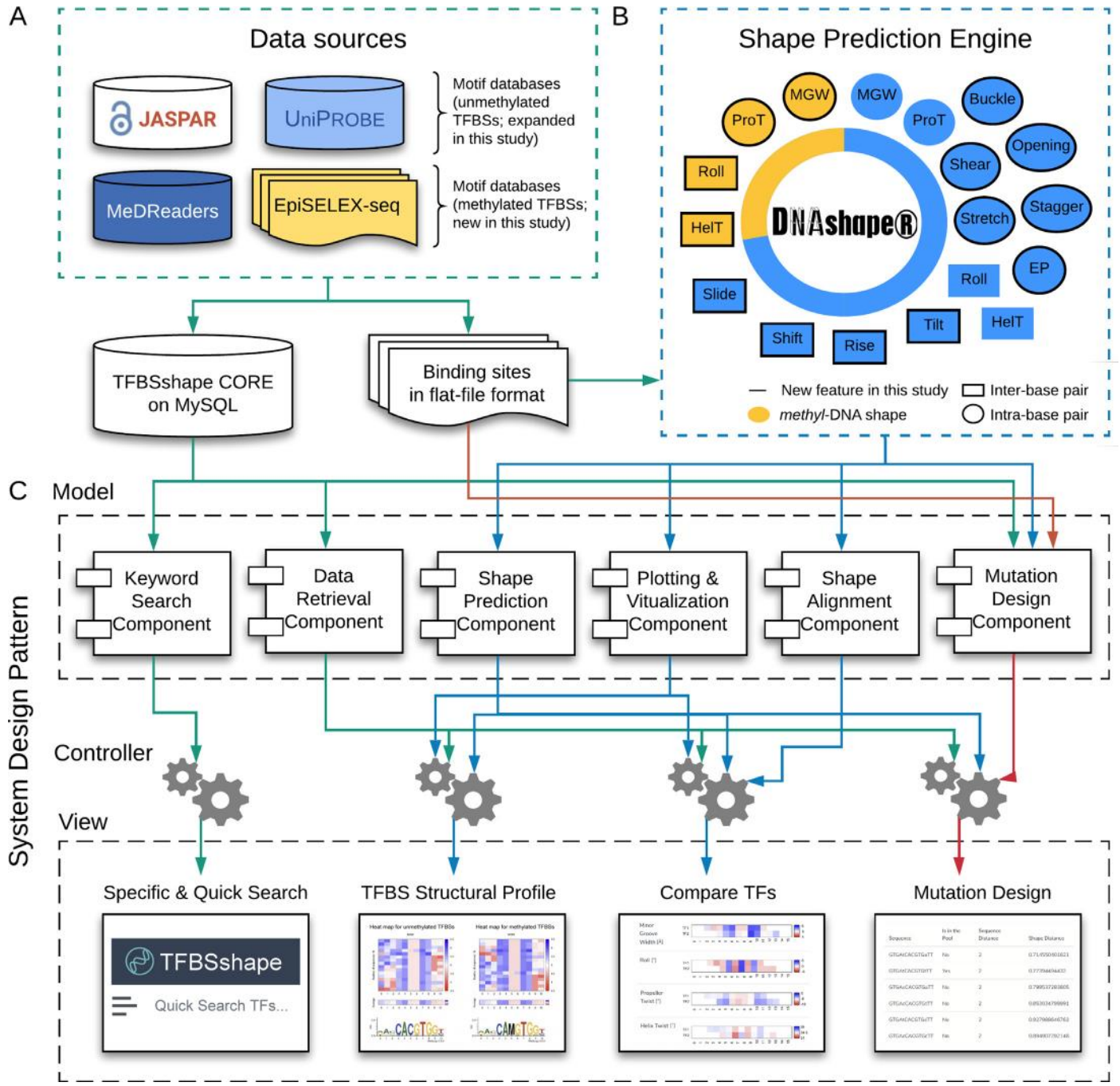
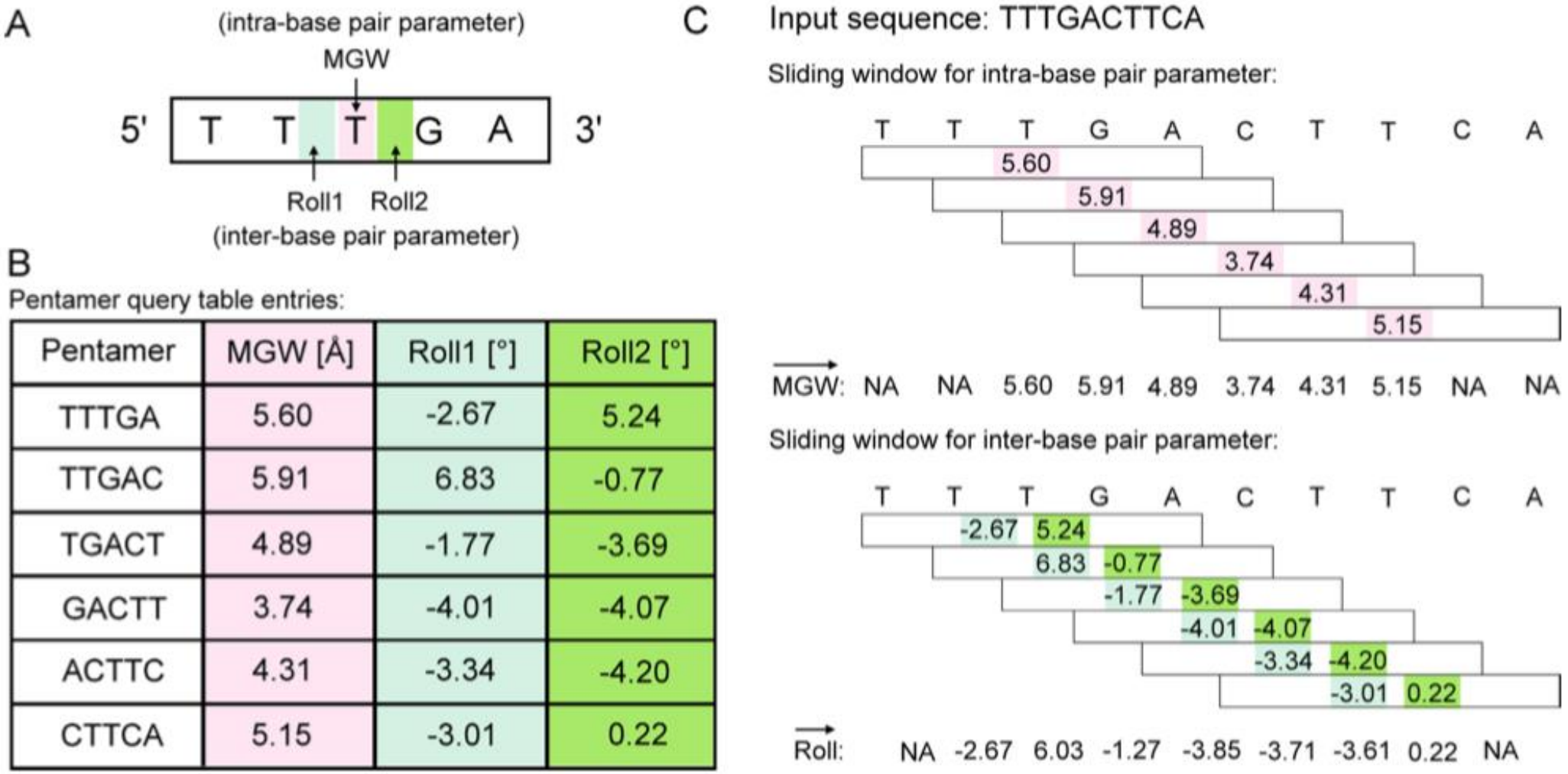


Fig. 2. Schematic illustration of the pentamer model for high-throughput prediction of DNA shape.



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

Chiu T. et al. NAR, 2020

Fig. 3. Search Interface in the TFBSshape Webtool

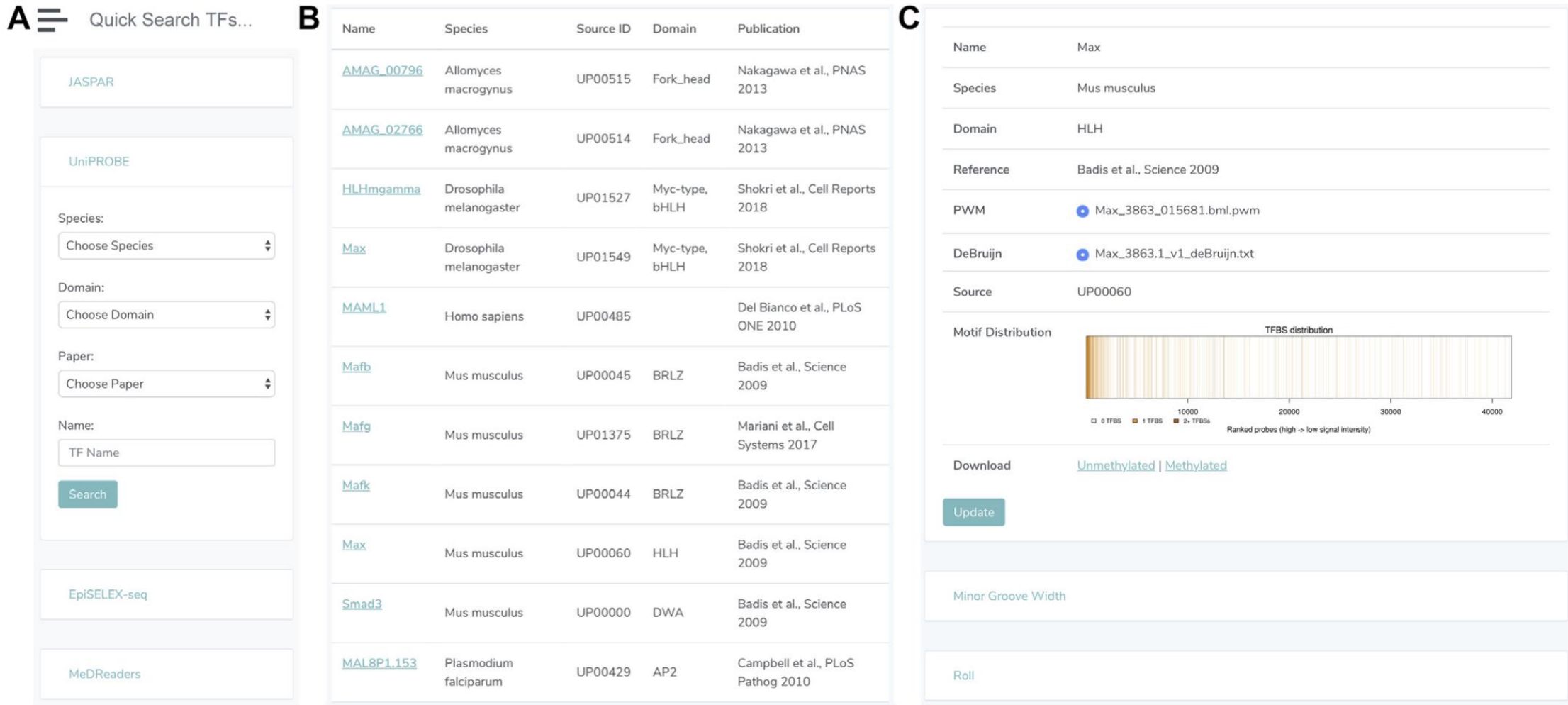


Fig. 3. Example search output by the TFBSshape

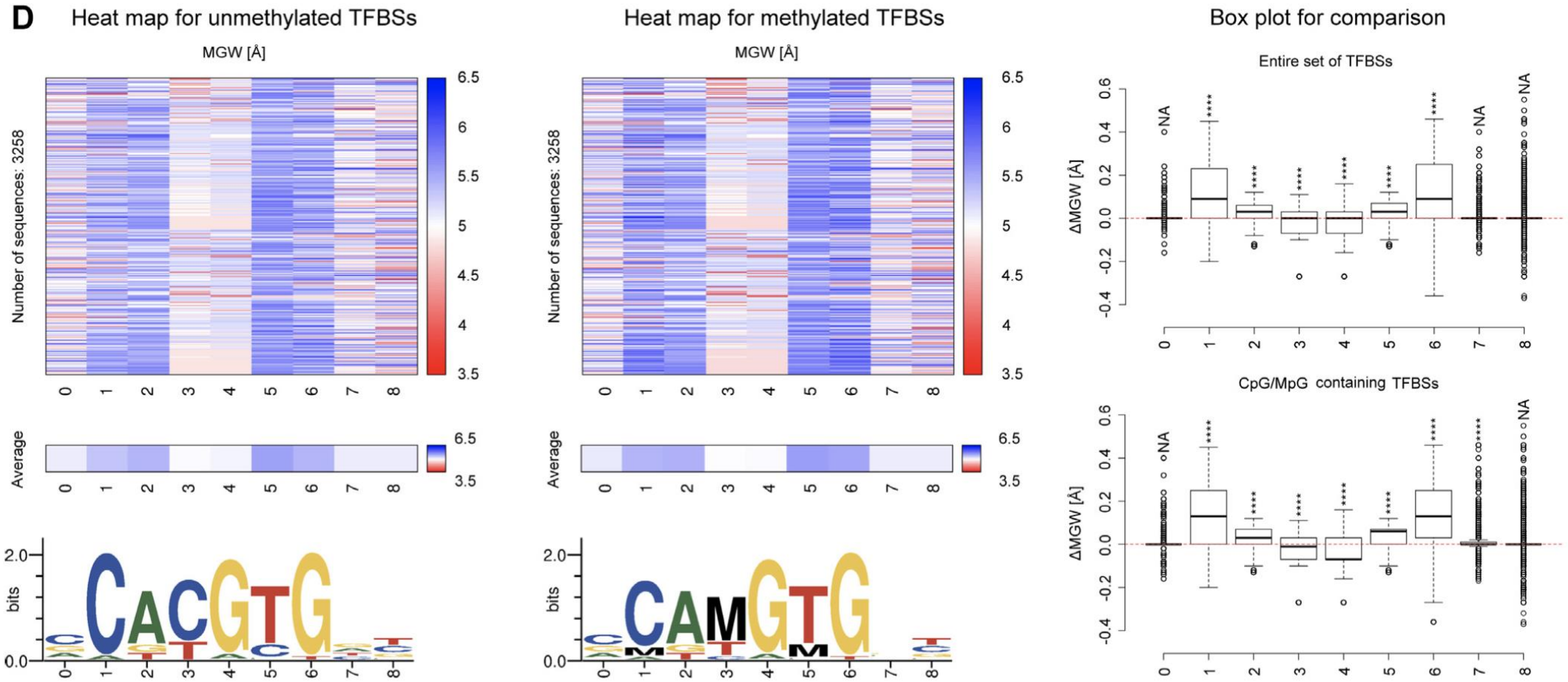


Fig. 4. Example TF compare interface in the TFBSshape Webtool

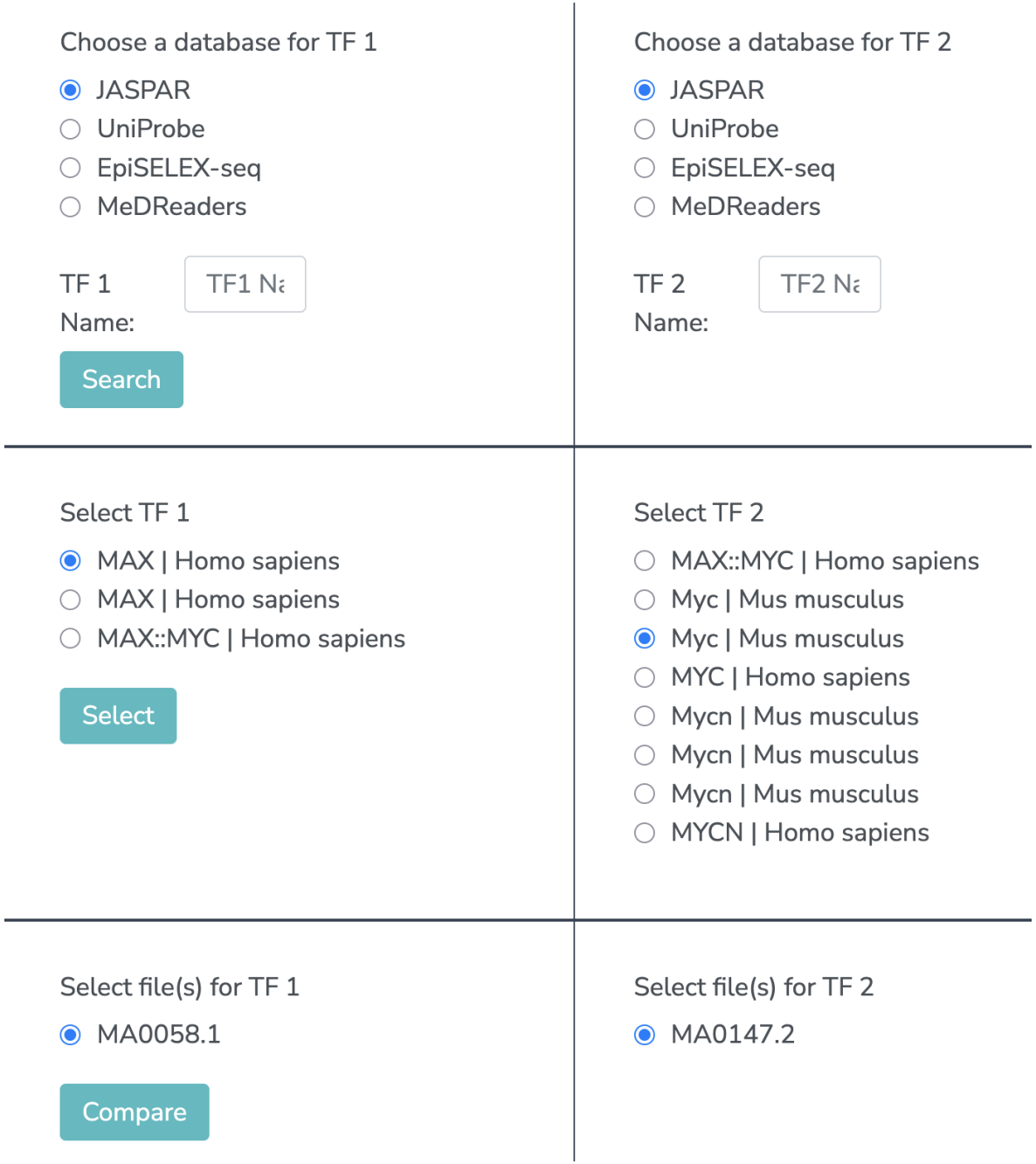


Fig. 4. Example Output of TF compare by TFBSshape

Name	MAX
Species	Homo sapiens
Database	JASPAR
Download	Unmethylated | Methylated
Sequence Logo	

Name	Myc
Species	Mus musculus
Database	JASPAR
Download	Unmethylated | Methylated
Sequence Logo	

Reference Position 1

0

Reference Position 2

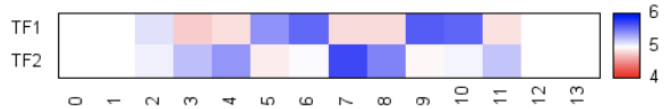
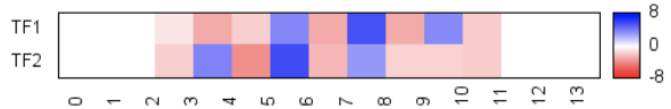
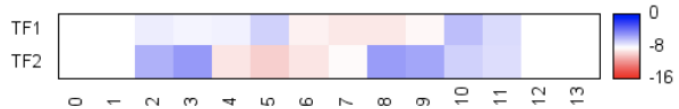
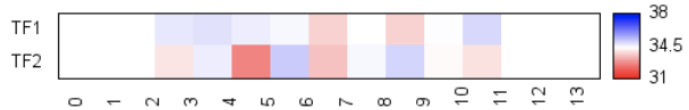
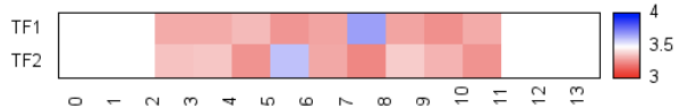
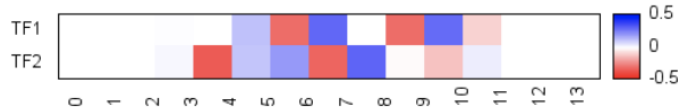
0

Align by Reference

Fig. 4. Example Output of TF compare by TFBSshape.

PCC: Pearson correlation coefficient

ED: Euclidian distance

Feature	Alignment	PCC	ED
Minor Groove Width		-0.782	1.81
Roll		0.496	9.51
Propeller Twist		-0.174	7.01
Helix Twist		-0.141	2.75
Rise		-0.458	0.575
Shift		-0.463	1.08

- Pearson correlation coefficient

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

where:

- $\bar{p}$ is the mean of vector $\mathbf{p}$: $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$
- $\bar{q}$ is the mean of vector $\mathbf{q}$: $\bar{q} = \frac{1}{n}\sum_{i=1}^{n} q_i$

- Euclidian distance

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

where:

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

Fig5. Example Mutation Design Interface in the TFBSshape Webtool

Fig5. Example Mutation Design Interface in the TFBSshape Webtool

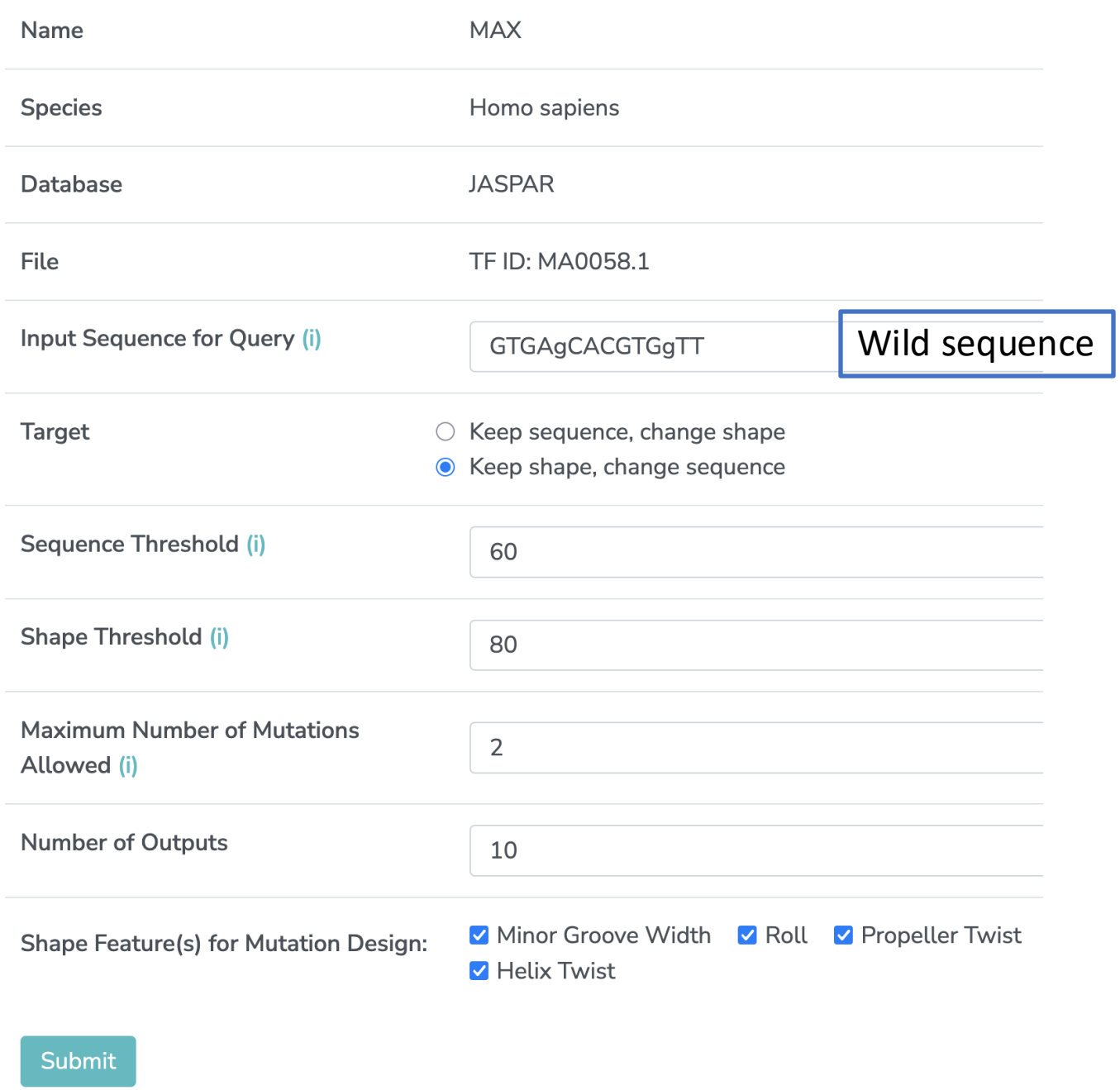
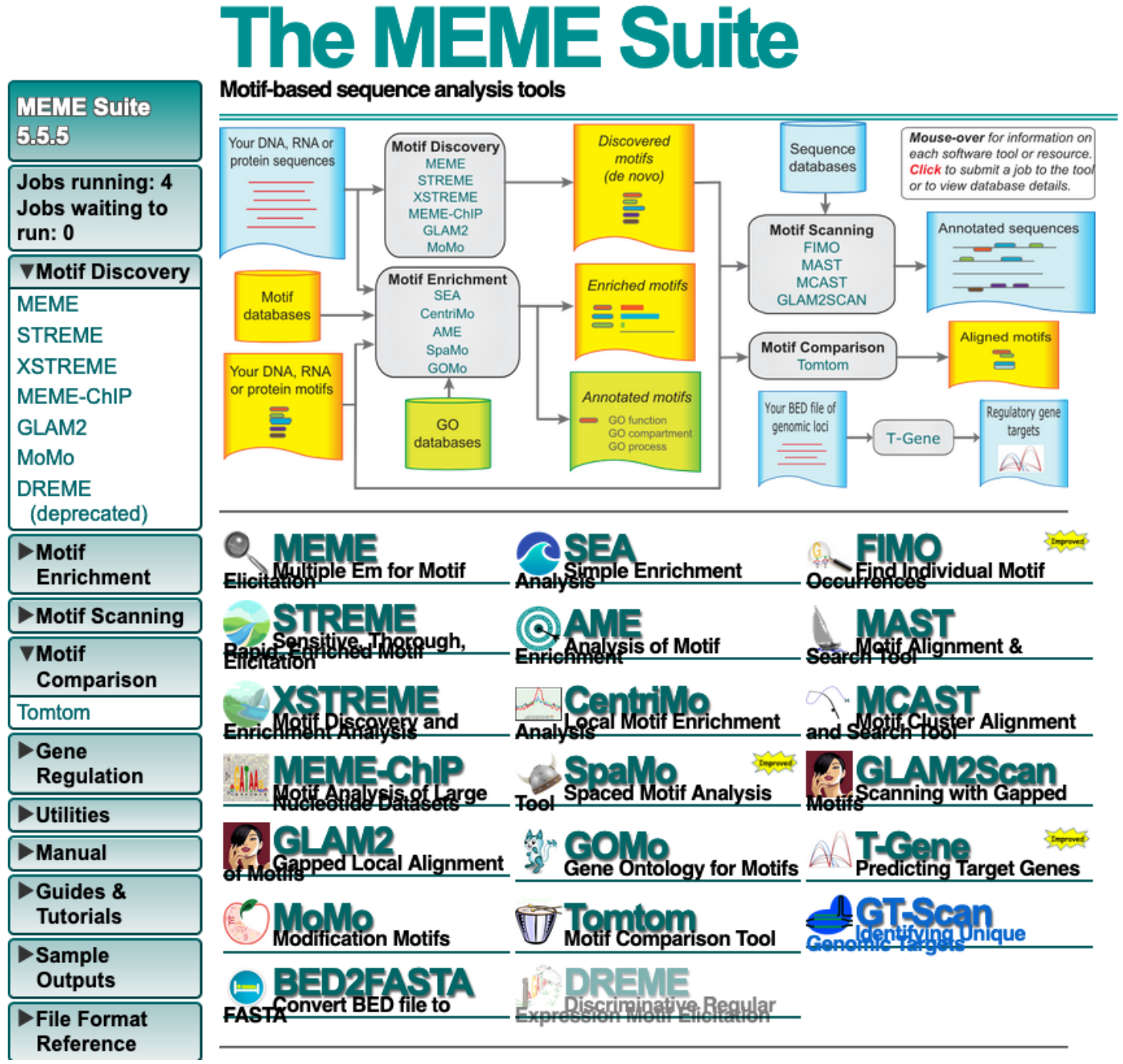


Fig5. Example Mutation Design Output by TFBSshape

Sequence	Is in the Pool	Sequence Distance	Shape Distance
GATAAAATGAGACCAgtgcatTATTAGTTGTACACC	No	0	0.0
GATAAAATGAGACCAtcgcatTATTAGTTGTACACC	No	2	0.446928134427
GATAAAATGAGACCAgtgcgtTATTAGTTGTACACC	No	1	0.476333023662
GATAAAATGAGACCAgtgtatTATTAGTTGTACACC	No	1	0.499588222314
GATAAAATGAGACCAttgcatTATTAGTTGTACACC	No	1	0.507074315589
GATAAAATGAGACCAgtacatTATTAGTTGTACACC	No	1	0.516044767294
GATAAAATGAGACCAccgcatTATTAGTTGTACACC	No	2	0.518749448563

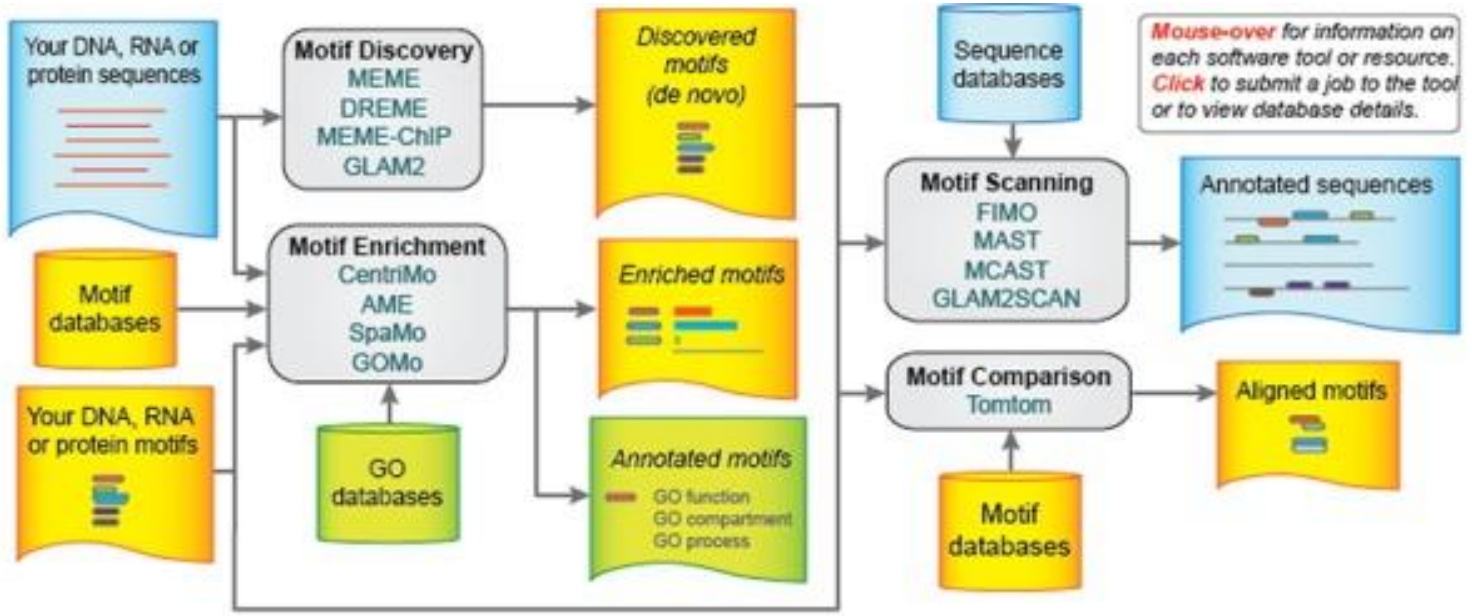
Motif Analysis

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



MEME Suit

(Bailey T.L., NAR, 2015)

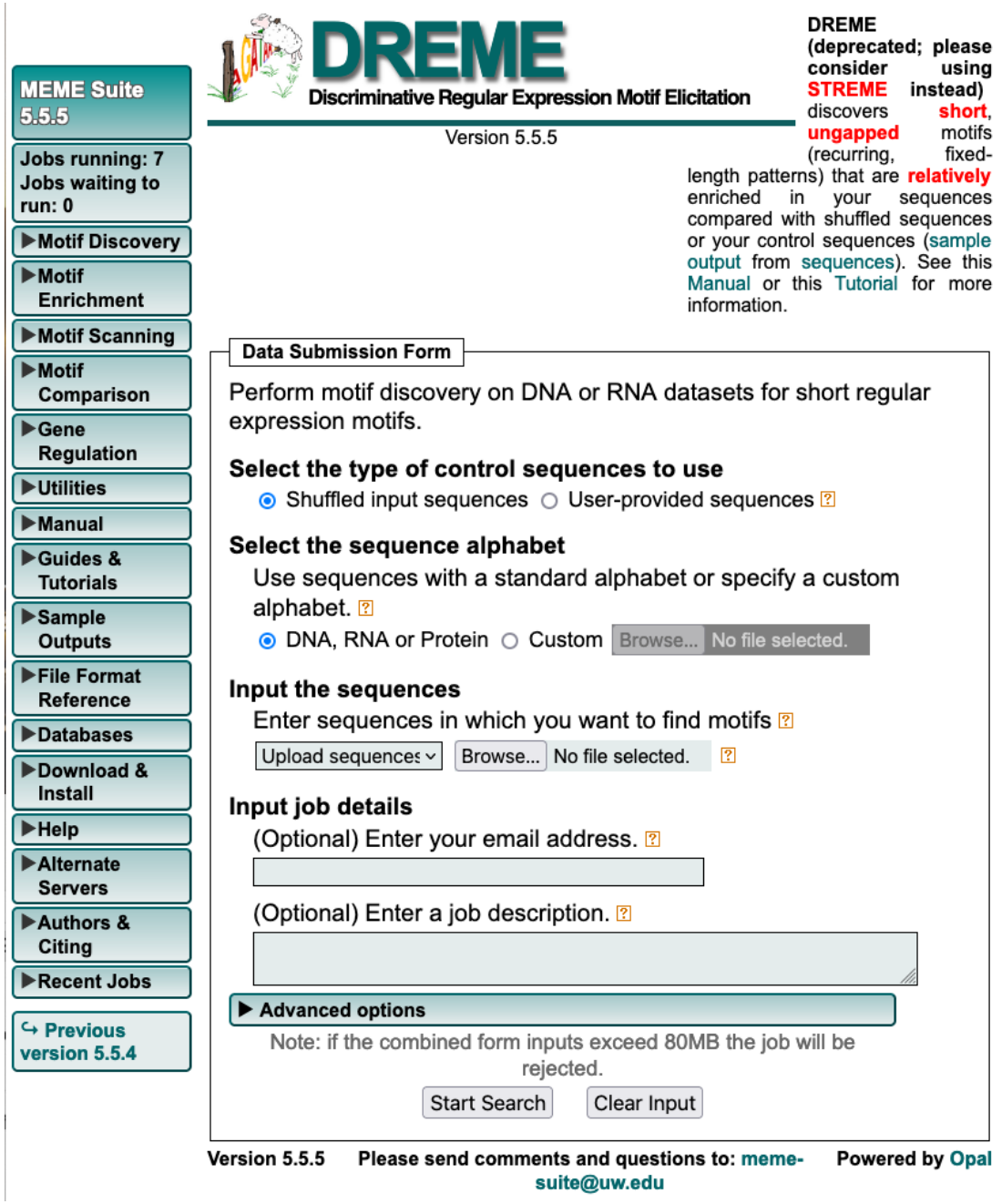


Discovery	Enrichment	Scanning	Comparison	Tool	Ref.	Description
✓				MEME	(2)	Discovers novel, ungapped motifs (recurring, fixed-length patterns) in nucleotide or protein sequences; MEME splits variable-length patterns into two or more separate motifs
✓				DREME	(3)	Discovers short, ungapped motifs (recurring, fixed-length patterns) that are relatively enriched in your nucleotide sequences compared with shuffled sequences or your control sequences
✓	✓		✓	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
✓				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
	✓			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
		✓		MCAST	(12)	Searches sequences for clusters of matches to one or more nucleotide motifs
		✓		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 over-represented motifs

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



DREME

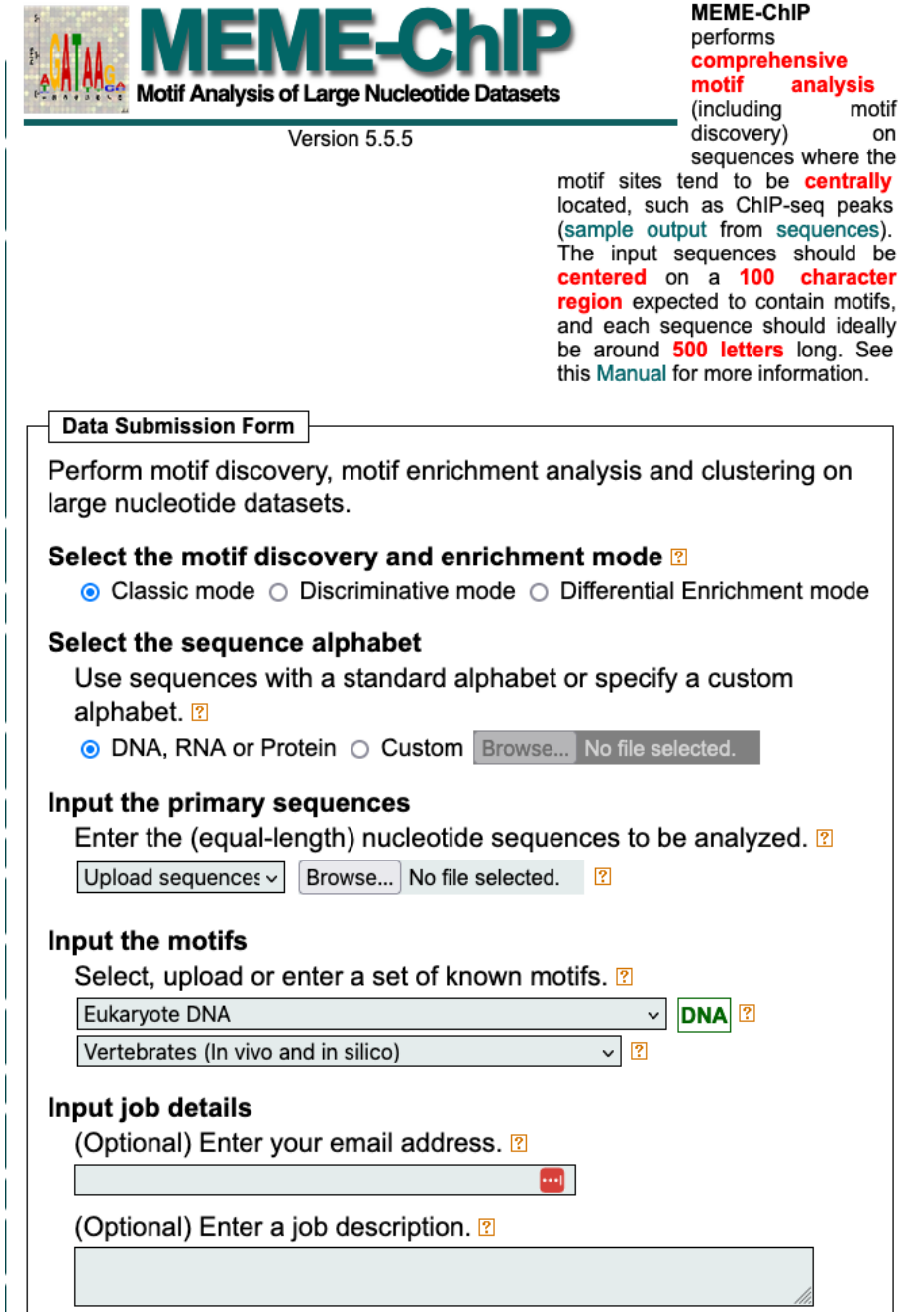
- DREME (Discriminative Regular Expression Motif Elicitation): https://meme-suite.org/meme/tools/dreme
  - A motif discovery algorithm designed to find short, core DNA-binding motifs of eukaryotic transcription factors and is optimized to handle large ChIP-seq data sets.
  - Tailored to eukaryotic data 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).

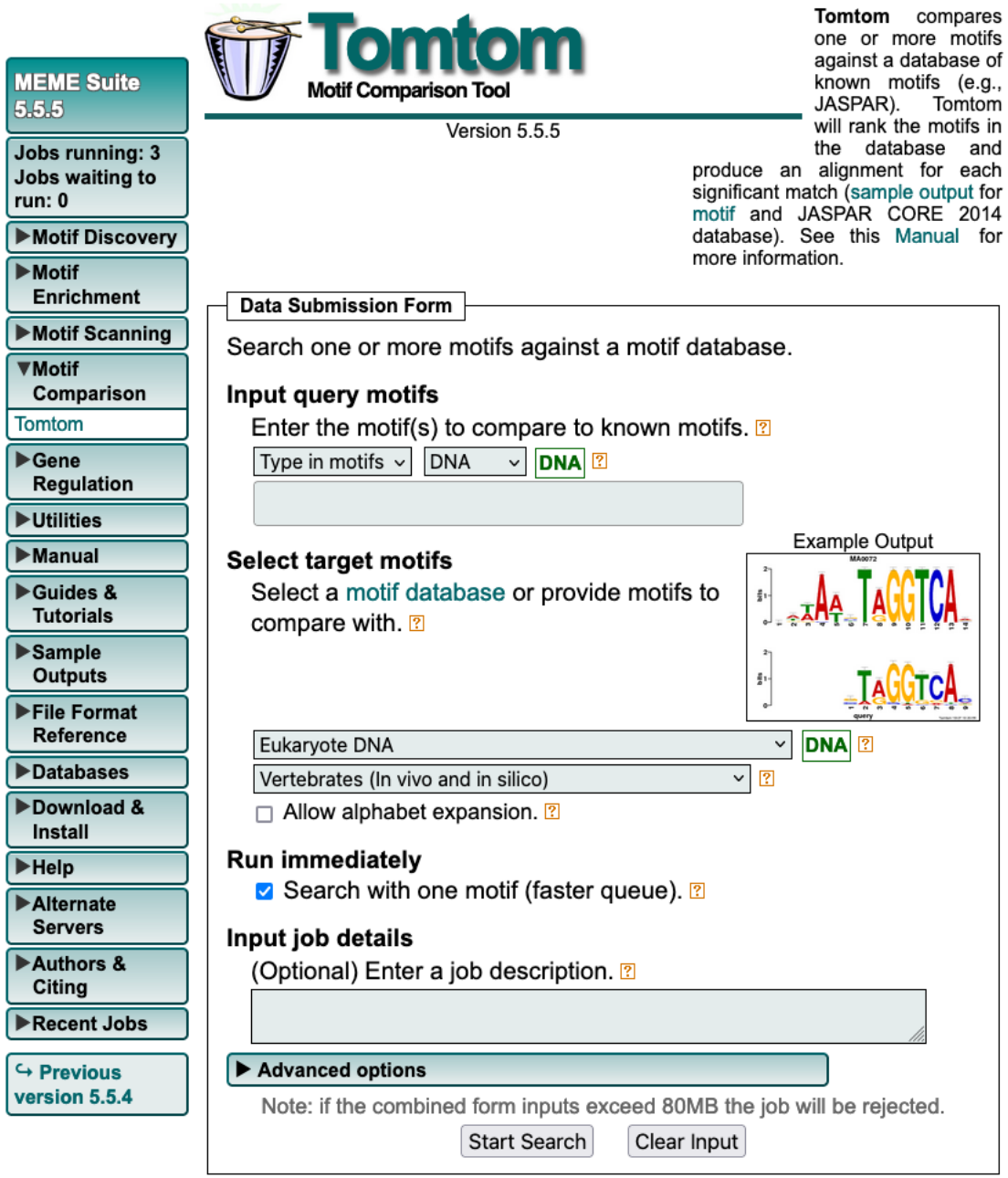
MEME-ChIP

- Part of the MEME Suite that is specifically designed for ChIP-seq analyses.
- 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.

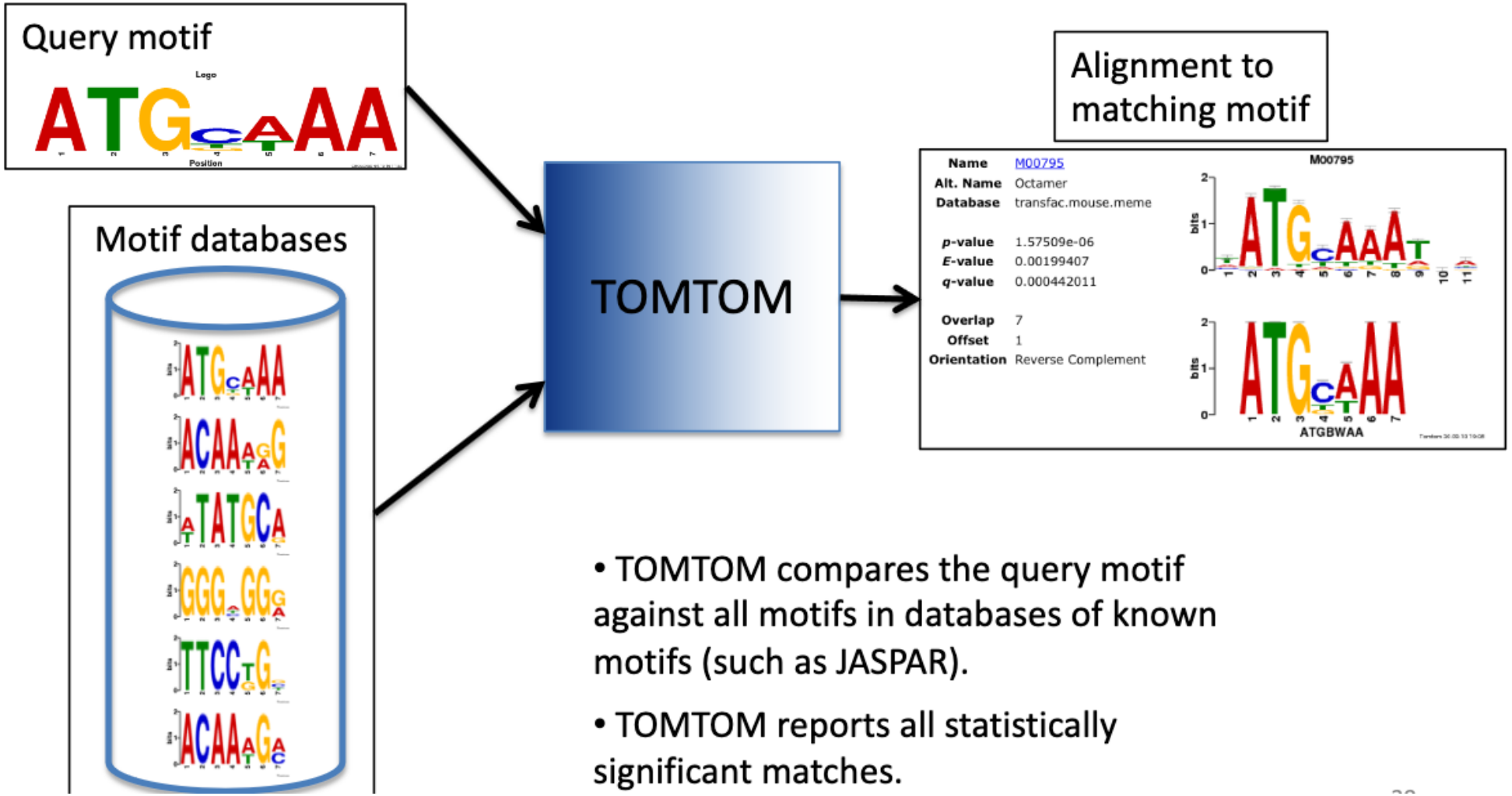


Tomtom

- Tomtom: https://meme-suite.org/meme/tools/tomtom
- 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 motif-motif similarity.



TOMTOM:
predict which proteins may bind a DNA motif



- TOMTOM compares the query motif against all motifs in databases of known motifs (such as JASPAR).
- TOMTOM reports all statistically significant matches.

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:**
  - https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/integrating_rna-seq_and_chip-seq.html
- TFBSshape:
  - https://tfbsshape.usc.edu/Home
- MEME Suite:
  - https://meme-suite.org/meme/