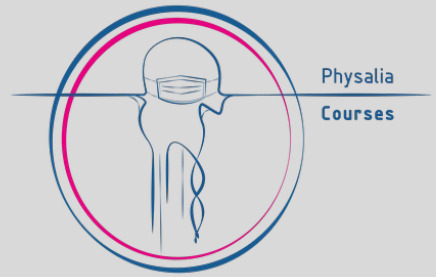
Motif identification from ChIP-seq

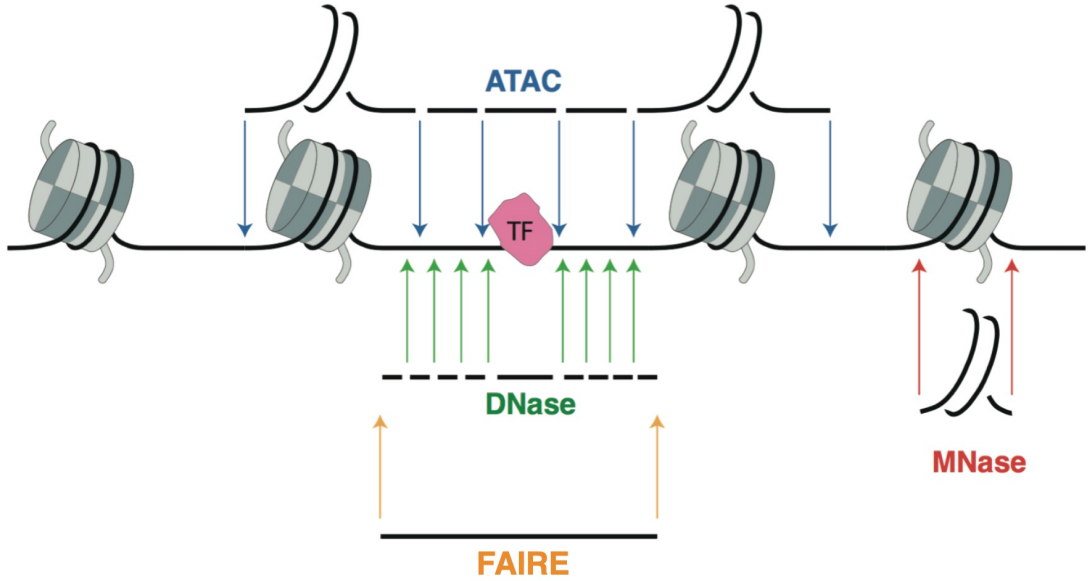
Epigenomics Data Analysis

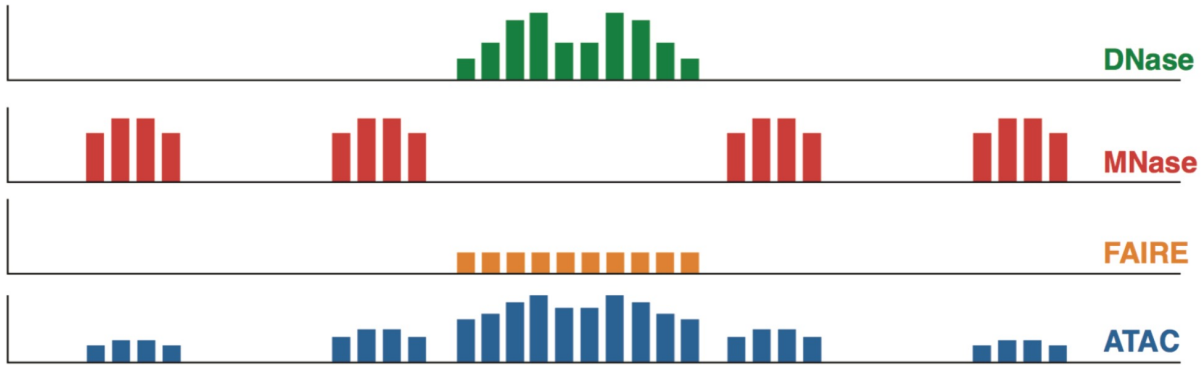
Jacques Serizay

Physalia 2025

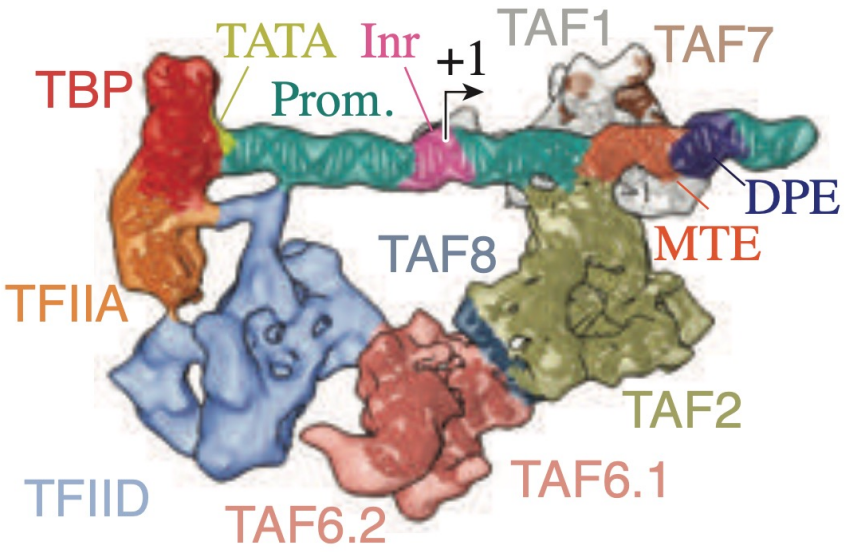


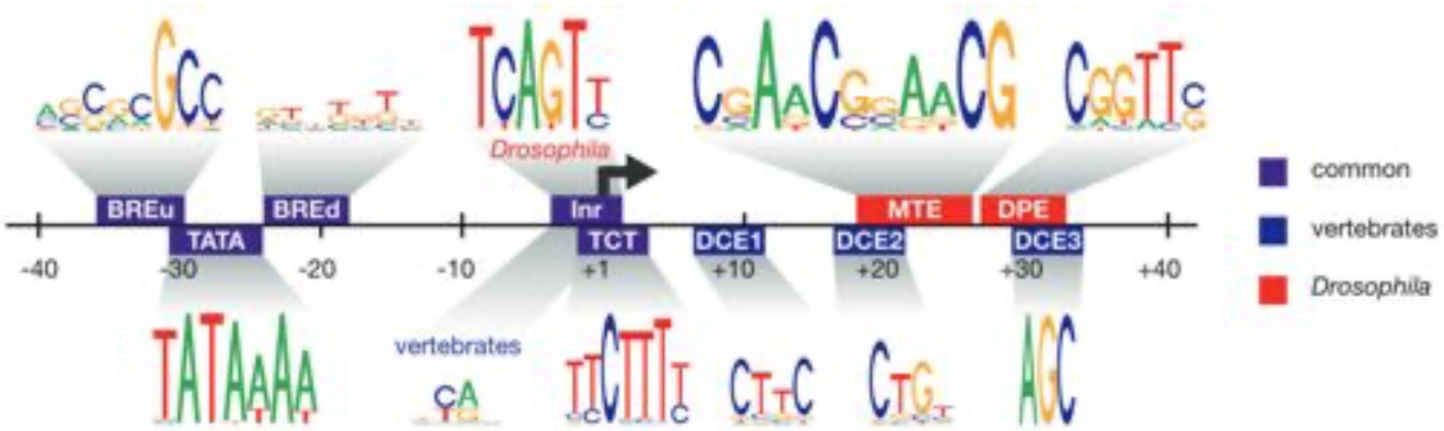
Accessible sites are often regulatory elements



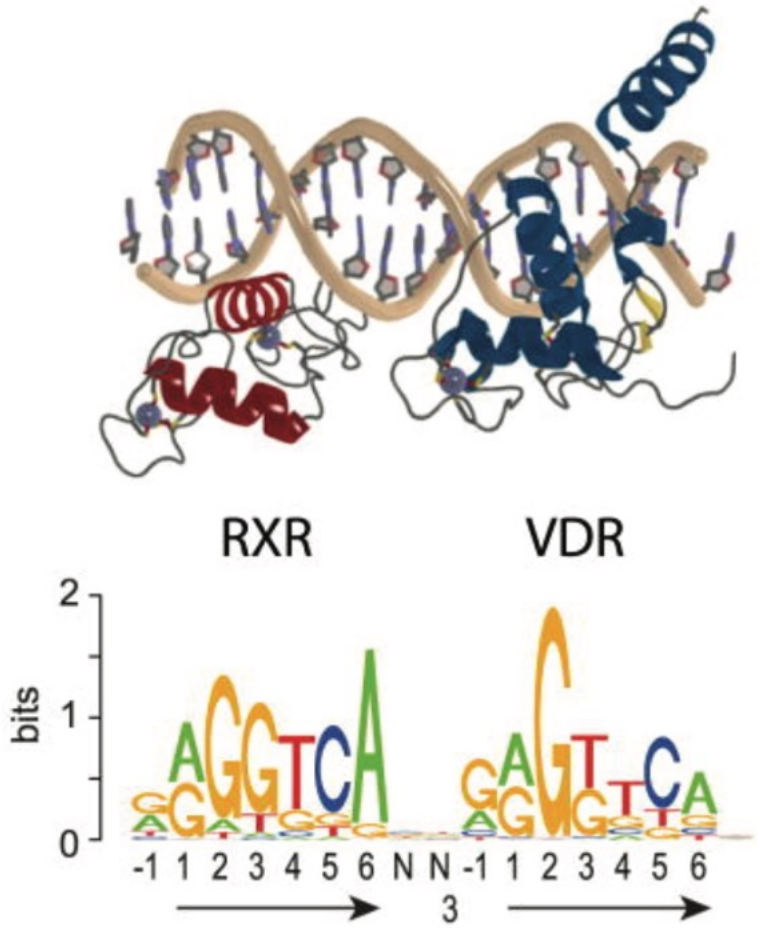


General transcription machinery is assembled around promoters using sequence affinity

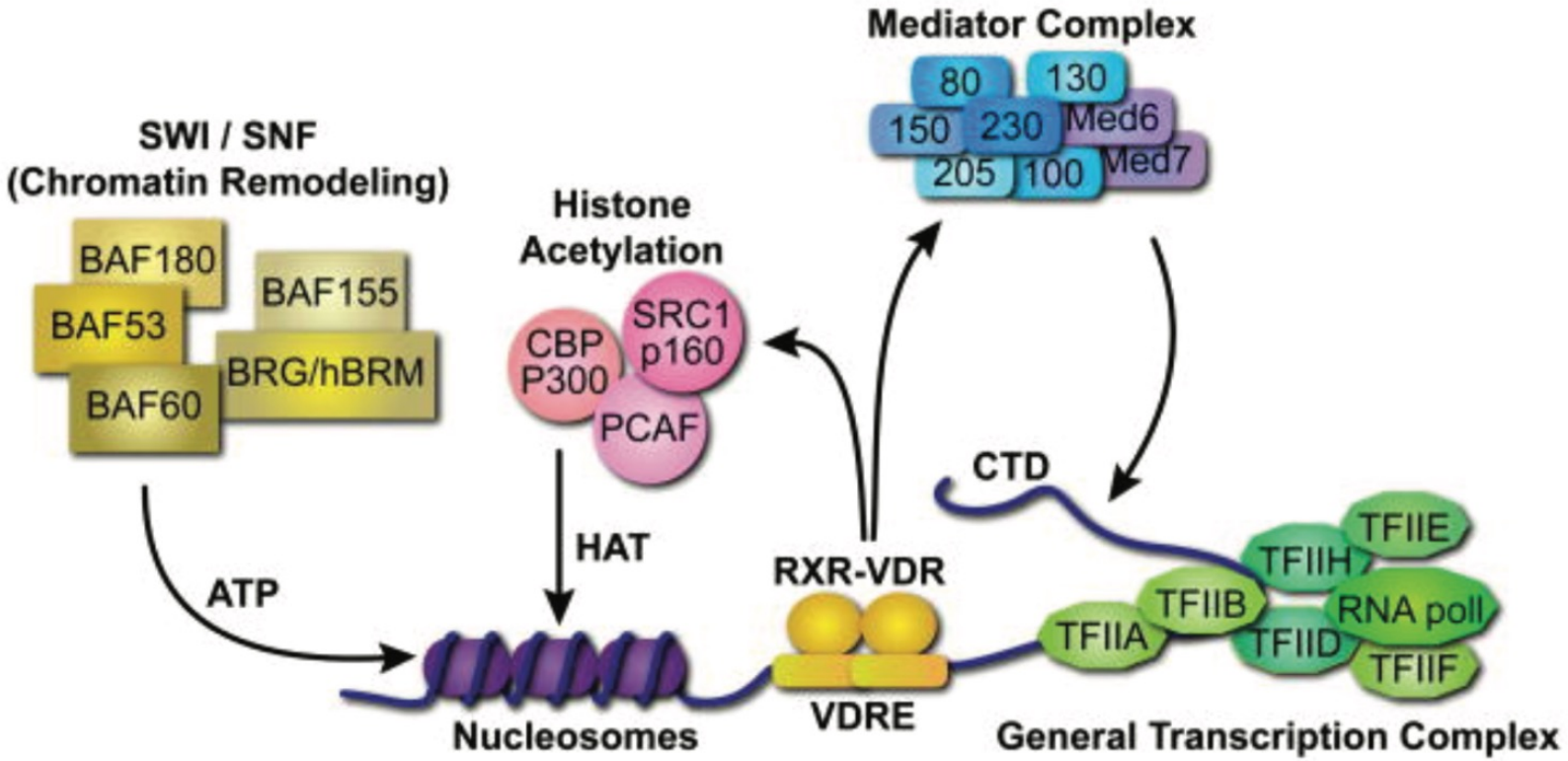




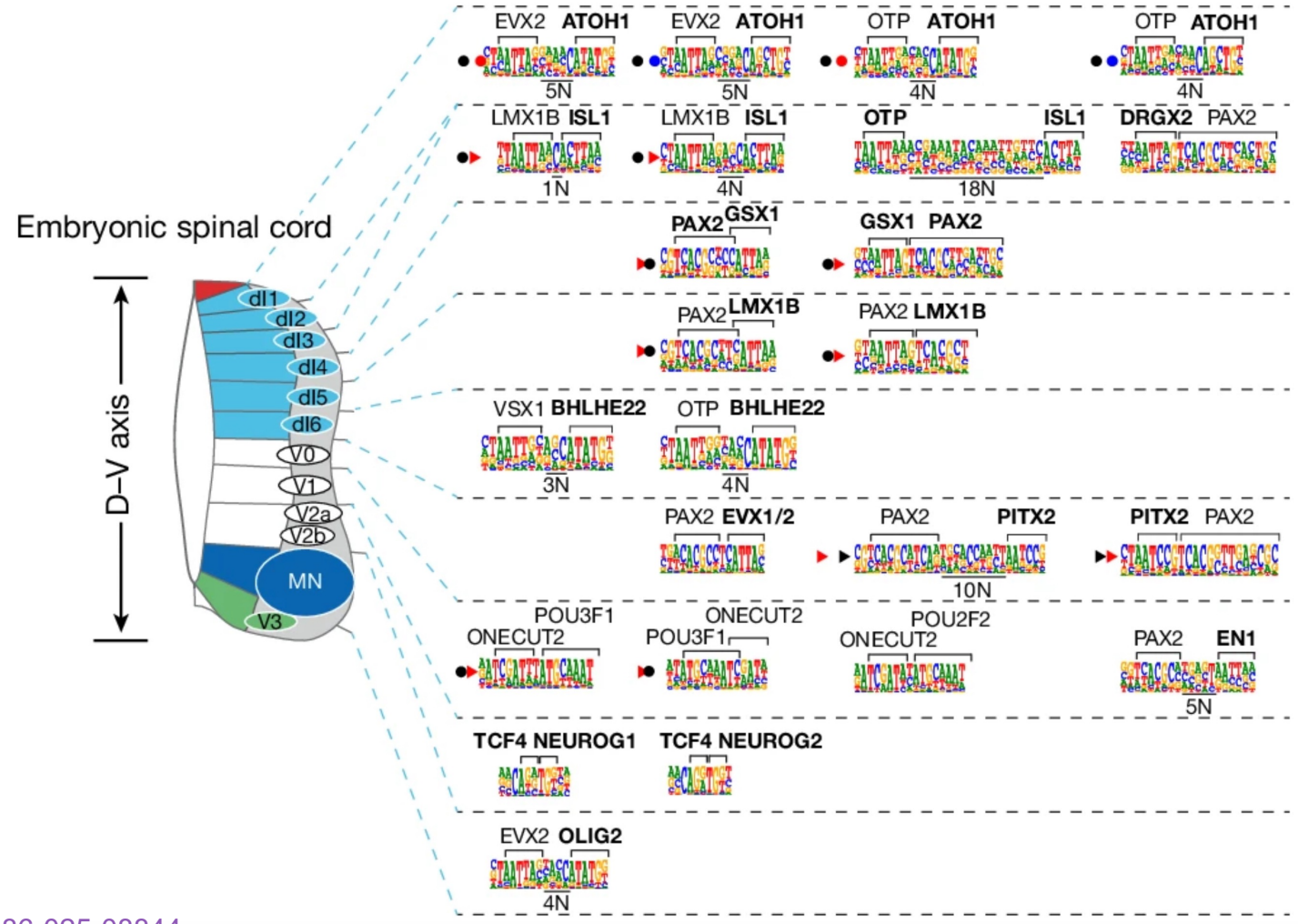
Transcription factors can be recruited directly to DNA motifs



TFs can also mediate INDIRECT recruitment of factors to chromatin



Motifs are important in shaping all biological processes



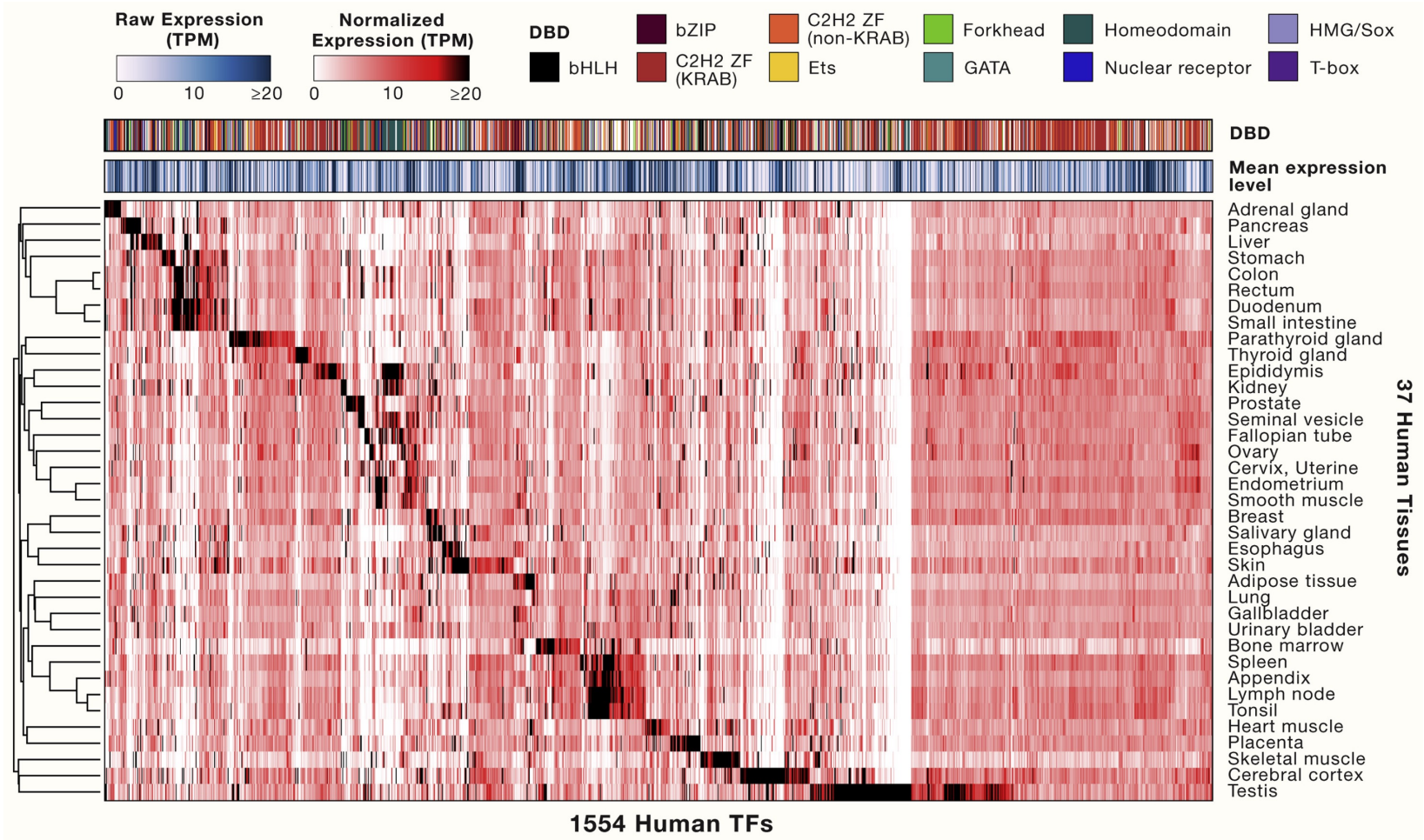
https://www.nature.com/articles/s41586-025-08844-z

Motifs are important in shaping all biological processes

		Method	Description	Features			
				Capability of de novo **motif discovery** (approx. length in base pairs with high information content)	Identifies **genomic binding locations** of a TF	Can measure effect of **CpG Methylation**	Can measure **cooperative binding and/or multimers**
High-throughput	*In Vitro Methods*	Protein Binding Microarray (PBM)	A GST-tagged TF is bound to a glass slide that has ~41,000 spots of short immobilized DNA sequences. Fluorescence-based detection of bound spots and k-mer enrichment analysis yields motifs.	✔ (< 12 bp)	✘	✔ Methyl-PBM	✔
		Bacterial one-hybrid	TF binding sites are selected in bacterial cells from a randomized library that is cloned in front of selectable marker genes. Can be reversed to select proteins able to bind a constant DNA sequence using a library of variant protein sequences.	✔ (< 14 bp)	✘	✘	✘
		SELEX-based methods	Systematic evolution of ligands through exponential enrichment (SELEX) involves adding TFs to a DNA pool containing many randomized sequences and selecting for binding in multiple rounds. Related methods include HT-SELEX, SELEX-seq, and Bind-n-Seq. Selection can be performed using affinity tags, or molecular trapping on a microfluidic platform (SMiLE-seq).	✔ (< 25 bp)	✘	✔ Methyl-HT-SELEX	✔ CAP-SELEX ✔ SMiLE-seq
Mid-throughput		DAP-seq	Single step SELEX using a library of fragmented genomic sequences. Sequence diversity is less than HT-SELEX, but genomic sequences that have co-evolved with the TF are included.	✔ Limited by skewed distribution of genomic sequences	✔ Peaks are not necessarily indicative of *in vivo* binding	✔ AmpDAP-seq	✘
		HiTS-FLIP	Uses an Illumina sequencer's flowcell as a PBM chip to measure binding to orders of magnitude more DNA sequences.	✔ (< 17 bp)	✘	✘	✘
		Spec-seq	Single step SELEX using a synthesized library of degenerate sequences of interest. The lower complexity library is useful for quantitatively measuring effects of binding site mutations using sequencing.	✔ Limited by number of sequences assayed	✘	✔ Methyl-Spec-seq	✘
		MITOMI	A microfluidic device is used to isolate DNA-protein complexes from free DNA instantaneously to accurately measure the relative binding affinities of TFs to ~10,000 individual sites.		✘	✘	✘
Low-throughput		EMSA	Tests if a DNA sequence is bound by a protein by observing a shift in the electrophoretic migration of DNA.	✘ Useful for validating known binding sites	✘	✔	✔ EMSA-FRET
		DNA footprinting	DNA is incubated with a TF and then degraded using DNase-I, resulting in cuts in all positions except those that were protected by the bound TF.		✘	✔	✔
		ITC, SPR, MSTP	Isothermal titration calorimetry (ITC), Surface plasmon resonance (SPR) and Microscale thermophoresis (MSTP) measure the binding affinity of TF-DNA interactions.		✘	✔	✔
	In Vivo Methods	ChIP-based assays	Proteins are crosslinked to DNA using formaldehyde and precipitated with an antibody. Bound DNA is detected with qPCR, microarray (ChIP-chip), or sequencing (ChIP-seq). The ChIP-Exo variant incorporates exonuclease treatment to enhance resolution.	✔ Limited by skewed distribution of genomic sequences, and inability to distinguish direct from indirect binding	✔	✔ ChIP + Bisulfite-sequencing	✔ Re-ChIP
		DamID-seq	A TF is expressed in mammalian cells as a fusion to bacterial Dam-methylase. The enzyme methylates a consensus sequence in close proximity to the TF's binding sites, which can be mapped using restriction enzymes and high-throughput sequencing.		✔	✘	✔ Split DamID-seq

https://doi.org/10.1016/j.cell.2018.01.029

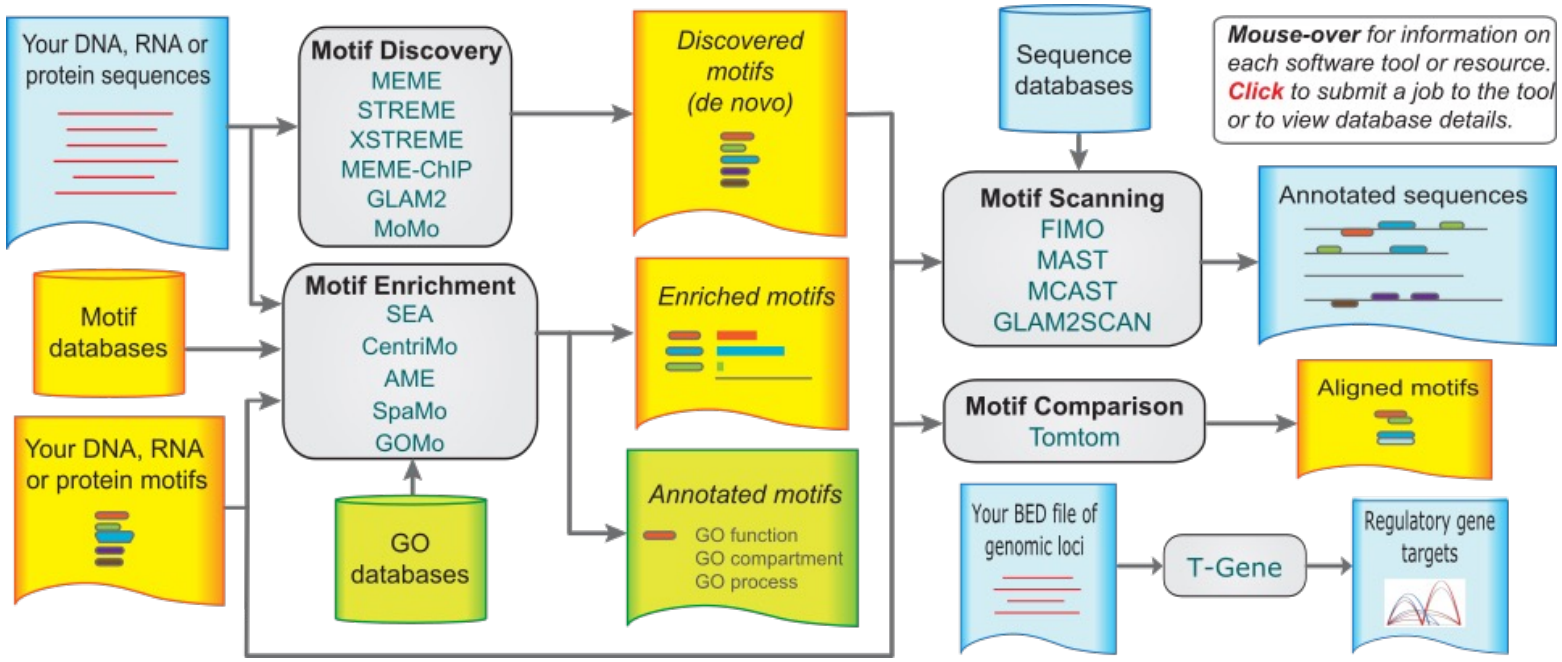
TF gene expression: an orthogonal approach to estimate TF's activity



https://doi.org/10.1016/j.cell.2018.01.029

Finding motifs enriched in sequences extracted from ChIP-seq peaks

MEME is a versatile suite of softwares largely used to identify known or *de novo* motifs enriched in a given set of sequences of interest.

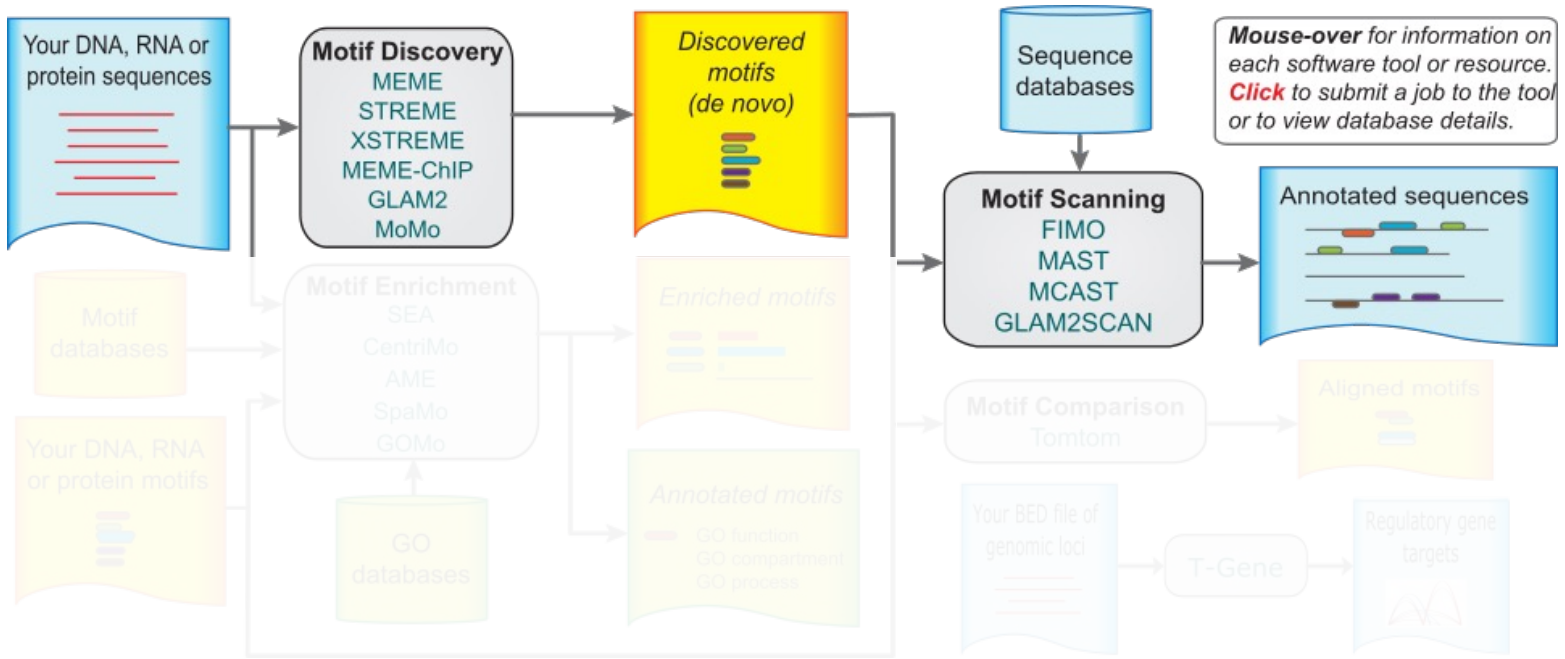


Timothy L. Bailey

Adjunct Research Professor
Department of Pharmacology
School of Medicine
University of Nevada, Reno

https://meme-suite.org/meme/

DE NOVO identification vs COMPARISON to known motifs



https://meme-suite.org/meme/

DE NOVO identification vs COMPARISON to known motifs

STREME

Sensitive, Thorough, Rapid, Enriched Motif Elicitation

View Command-Line Version

Description

STREME discovers **ungapped** motifs (recurring, fixed-length patterns) that are **enriched** in your sequences or **relatively enriched** in them compared to your control sequences (sample output from sequences).

STREME (Simple, Thorough, Rapid, Enriched Motif Elicitation) finds motifs (up to 30 columns wide) in large sequence datasets. The input to STREME is one or two sets of sequences. The control sequences should have approximately the same length distribution as the primary sequences. If you do not provide a control set, the program shuffles the primary set to create a control set. The program uses Fisher's Exact Test or the Binomial test to determine significance of each motif found in the positive set as compared with its representation in the control set, using a significance threshold that may be set on the command line.

STREME achieves high speed via an optimization algorithm that uses a suffix tree, and provides **accurate estimates** of the statistical significance of the motifs it discovers.

MEME

Multiple EM for Motif Elicitation

View Command-Line Version

Description

MEME discovers novel, **ungapped** motifs (recurring, fixed-length patterns) in your sequences (sample output from sequences). MEME splits variable-length patterns into two or more separate motifs.

A motif is an approximate sequence pattern that occurs repeatedly in a group of related sequences. MEME represents motifs as position-dependent letter-probability matrices that describe the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs.

MEME takes as input a group of sequences and outputs as many motifs as requested. MEME uses statistical modeling techniques to automatically choose the best width, number of occurrences, and description for each motif.

https://meme-suite.org/meme/

DE NOVO identification vs COMPARISON to known motifs

View Command-Line Version

Description

XSTREME performs **comprehensive motif analysis** (including motif discovery) on sequences where the motif sites can be **anywhere** in the sequences (sample output from sequences). The input sequences may be of **any length**, and their lengths may **vary**.

XSTREME will:

1. Discover novel motifs in the input sequences (with STREME and MEME).
2. Determine which motifs are most enriched (with SEA).
3. Analyze them for similarity to known motifs (with Tomtom).
4. Group significant motifs by similarity.
5. Create a GFF file for viewing each motif's predicted sites in a genome browser.

It is worth noting that XSTREME is **not** a motif scanner, but the motifs discovered by it can be used by FIMO, MAST or MCAST to scan for motif sites.

https://meme-suite.org/meme/

DE NOVO identification vs COMPARISON to known motifs

View Command-Line Version

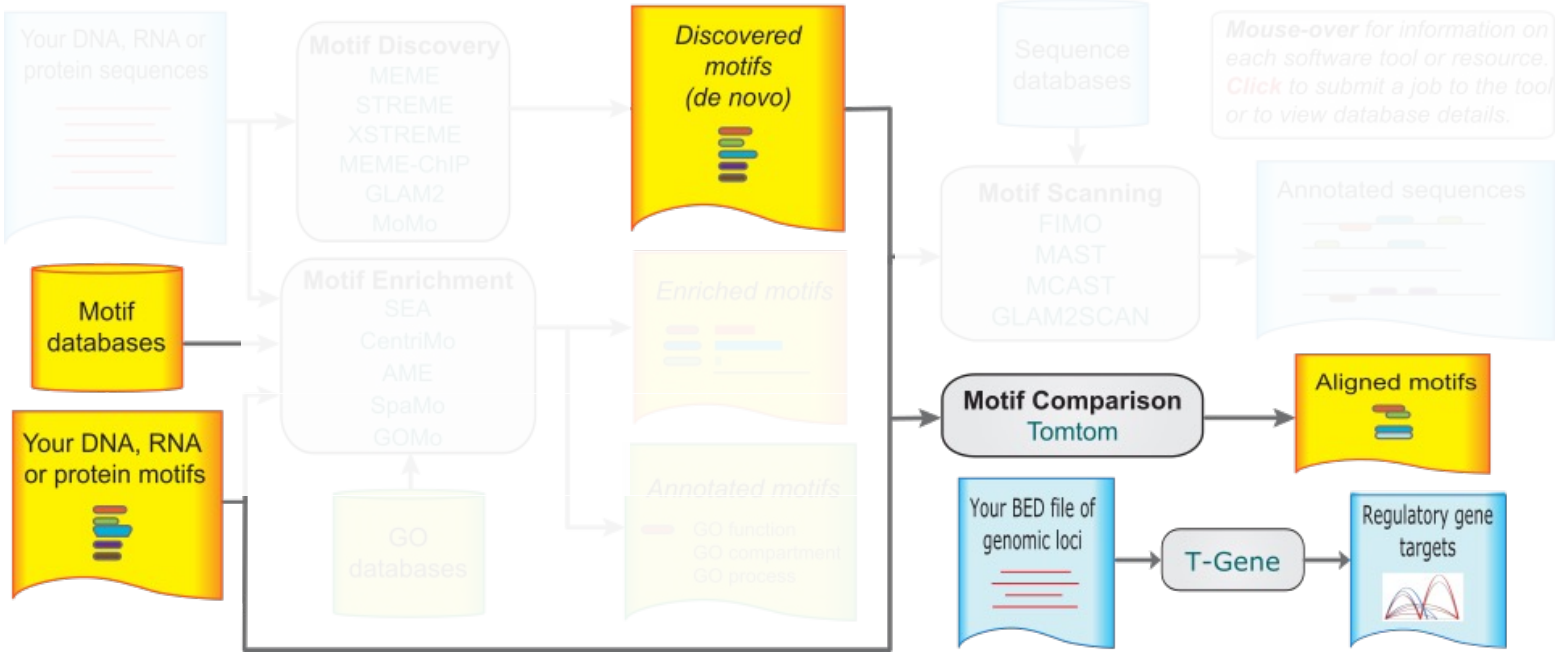
Description

FIMO scans a set of sequences for **individual matches** to each of the motifs you provide (sample output for motifs and sequences).

The name FIMO stands for 'Find Individual Motif Occurrences'. The program searches a set of sequences for occurrences of known motifs, treating each motif independently. Motifs must be in MEME Motif Format. The web version of FIMO also allows you to type in motifs in additional formats.

https://meme-suite.org/meme/

DE NOVO identification vs COMPARISON to known motifs



https://meme-suite.org/meme/

DE NOVO identification vs COMPARISON to known motifs

View Command-Line Version

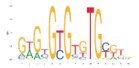
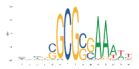
Description

Tomtom **compares** one or more **motifs** against a database of known motifs (e.g., JASPAR). Tomtom will rank the motifs in the database and produce an alignment for each significant match (sample output for motif and JASPAR CORE 2014 database).

Tomtom searches one or more query motifs against one or more databases of target motifs (and their reverse complements when applicable), and reports for each query a list of target motifs, ranked by *p*-value. The *E*-value and the *q*-value of each match is also reported. The *q*-value is the minimal false discovery rate at which the observed similarity would be deemed significant. The output contains results for each query, in the order that the queries appear in the input file.

https://meme-suite.org/meme/

Public motif databases provide extensive knowledge of TFBS

ID	Name	Species	Class	Family	Logo
MA0260.1	che-1	Caenorhabditis elegans	C2H2 zinc finger factors	More than 3 adjacent zinc fingers	
MA0261.1	lin-14	Caenorhabditis elegans			
MA0262.1	mab-3	Caenorhabditis elegans	DM-type intertwined zinc finger factors	DMRT factors	
MA0263.1	ceh-10::ttx-3	Caenorhabditis elegans	Homeo domain factors	Pair rela fact LIM	
MA0264.1	ceh-22	Caenorhabditis elegans	Homeo domain factors	NK	
MA0264.2	ceh-22	Caenorhabditis elegans	Homeo domain factors	NK	
MA0537.1	blmp-1	Caenorhabditis elegans	C2H2 zinc finger factors	More than 3 adjacent zinc fingers	
MA0538.1	daf-12	Caenorhabditis elegans	Other C4 zinc finger-type factors	C4-GATA-related	
MA0538.2	daf-12	Caenorhabditis elegans	Other C4 zinc finger-type factors	C4-GATA-related	
MA0541.1	efl-1	Caenorhabditis elegans	Fork head/winged helix factors	E2F	

https://jaspar.elixir.no/