

Introduction to gene expression regulation by transcription factors and its computational analysis

Ieva Rauluseviciute & Jaime A Castro-Mondragon



Centre for Molecular Medicine Norway



UiO : Universitetet i Oslo

Who are we?



NCMM

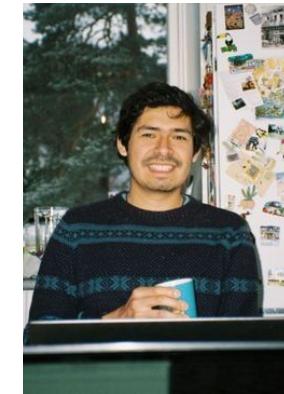


Ieva Rauluševičiūtė

Doctoral Research Fellow in Anthony Mathelier
Group at the Centre for Molecular Medicine
Norway



nykode
therapeutics



Jaime A. Castro Mondragon

Bioinformatician at the department of In Silico
analysis at Nykode Therapeutics

Who are you?



Wait a minute, who are you?



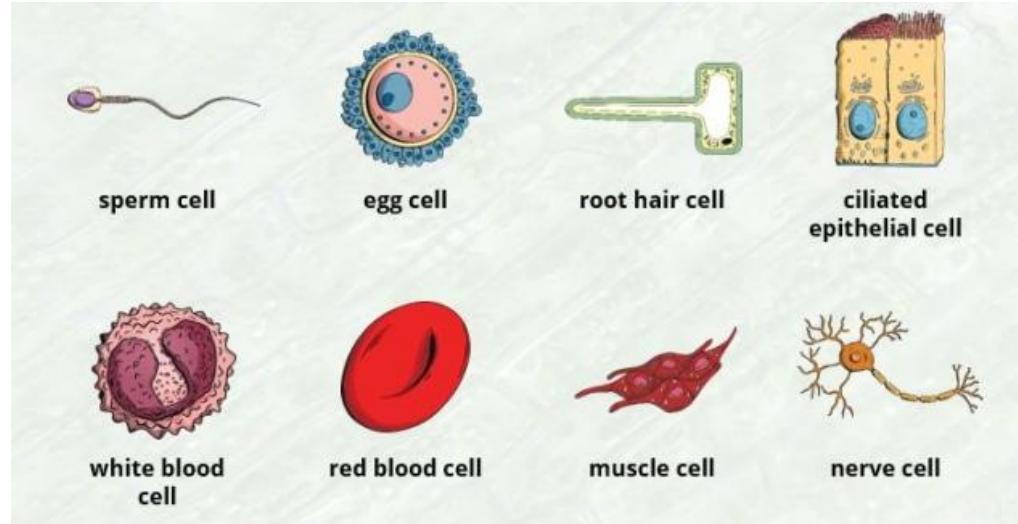
PART 1

Gene regulation by TF binding



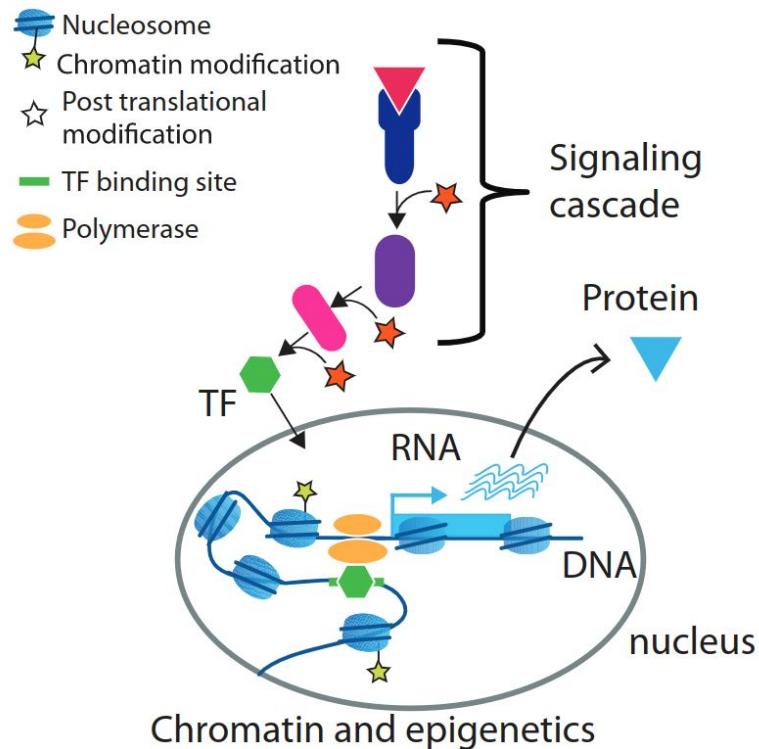
Why do we have multiple cell types?

- All cells in our bodies have the same DNA sequence but ...
- Why do we have hundreds of cell types with multiple morphology and functions ?
- Something ‘beyond’ the DNA itself must encode information of when, how and how much a gene should be correctly expressed.



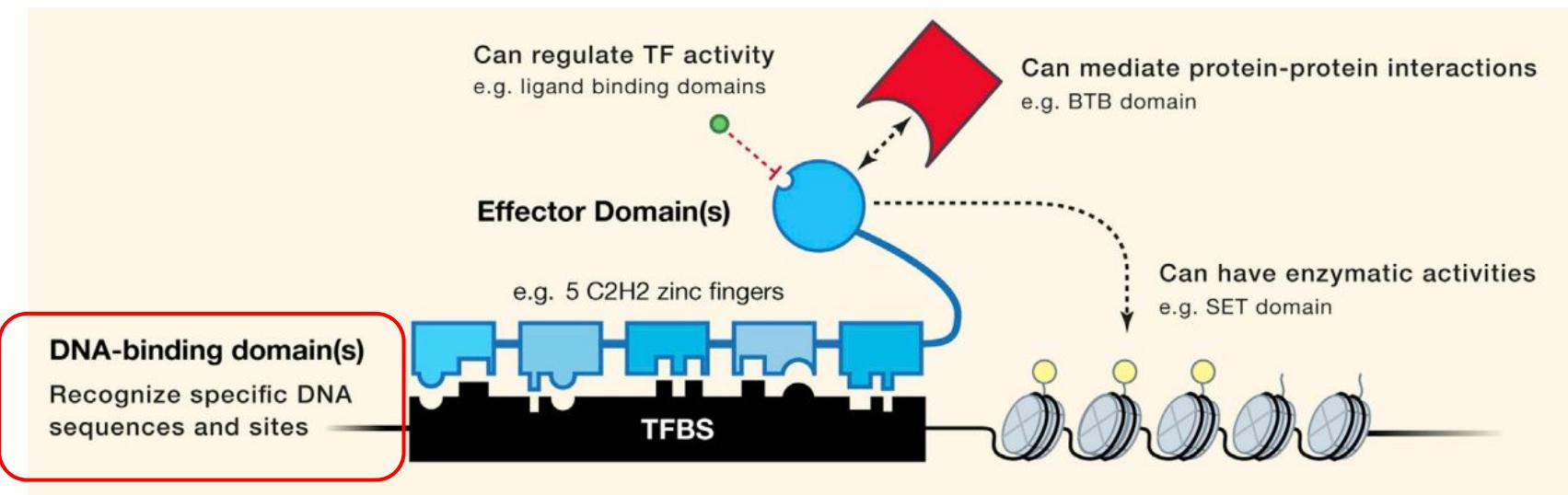
TFs link cell signaling with gene regulation

- Transcription Factors (TFs)
- Regulate key processes such as development and differentiation
- TFs activate or repress transcription.
- Combinatorics of TFs give rise to different cell types.

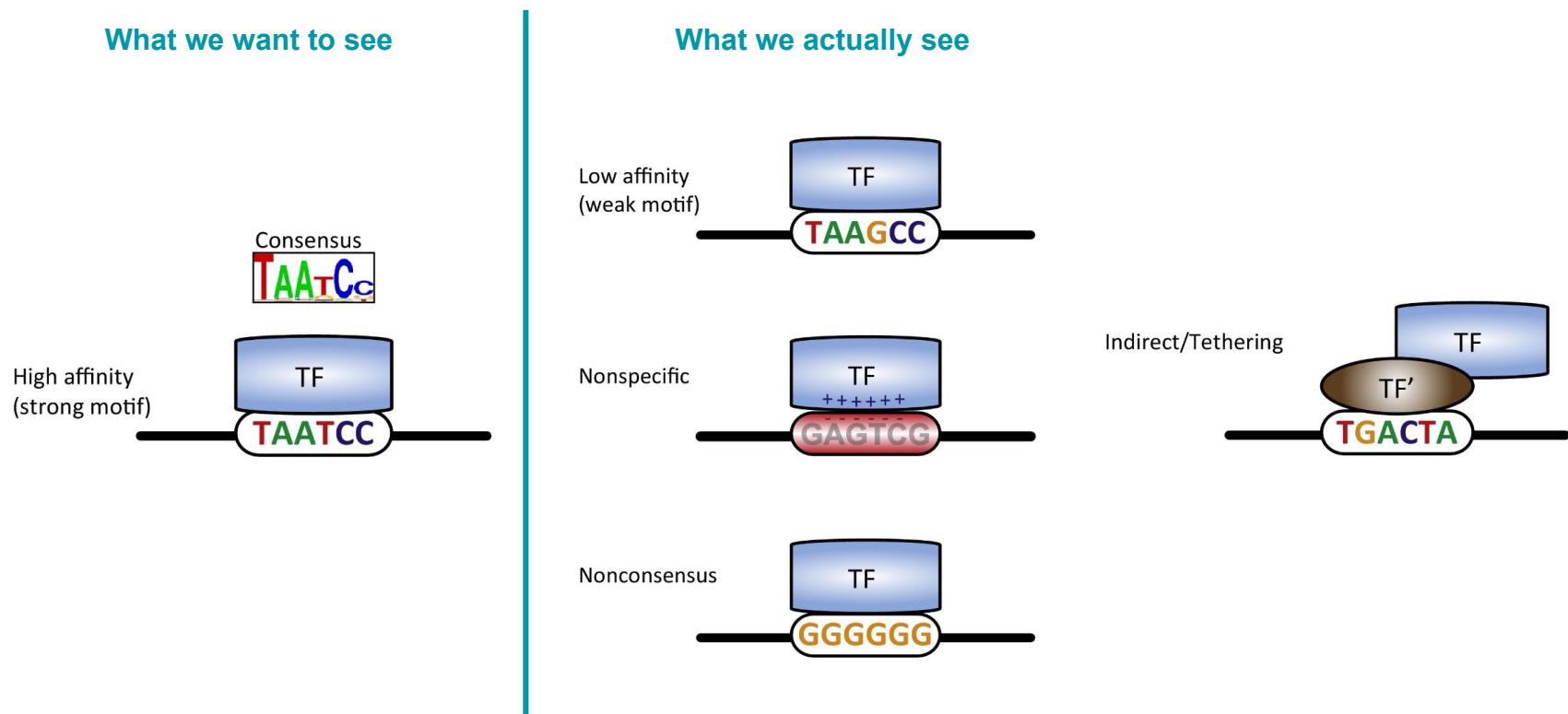


DNA-binding Transcription Factors

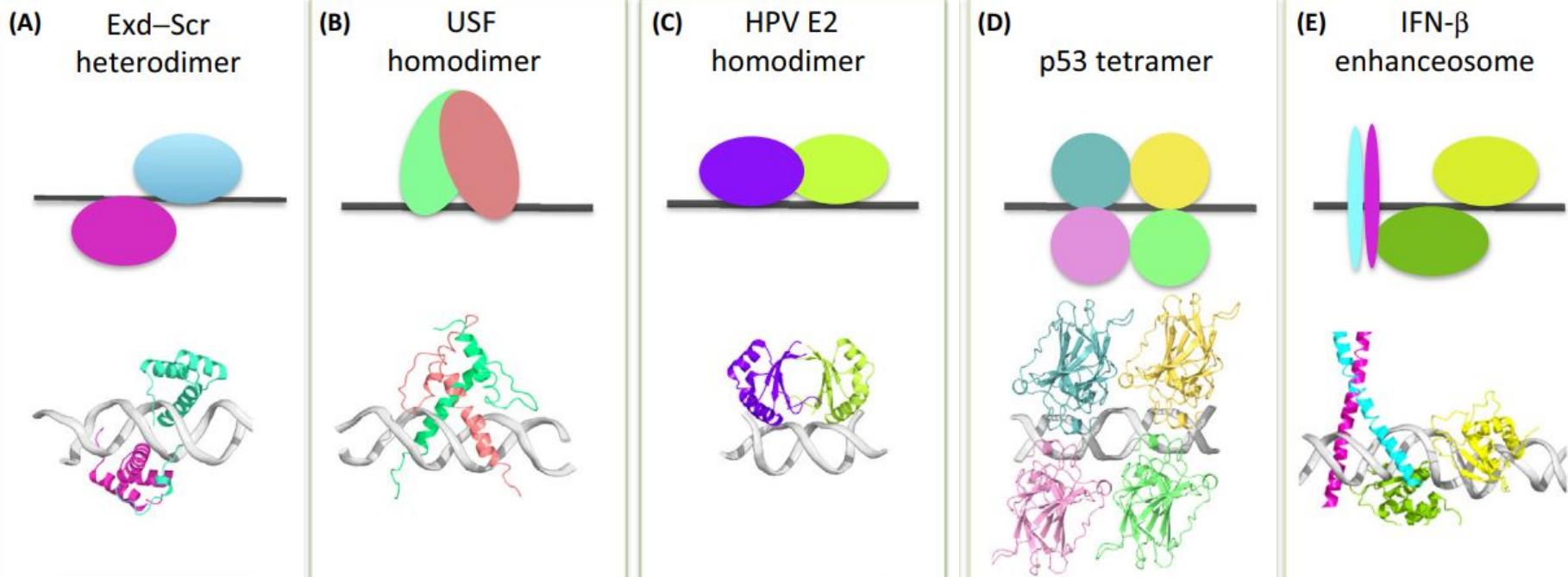
- Transcription Factors (TFs) bind DNA in a sequence-specific manner.
- TFs regulate transcription.



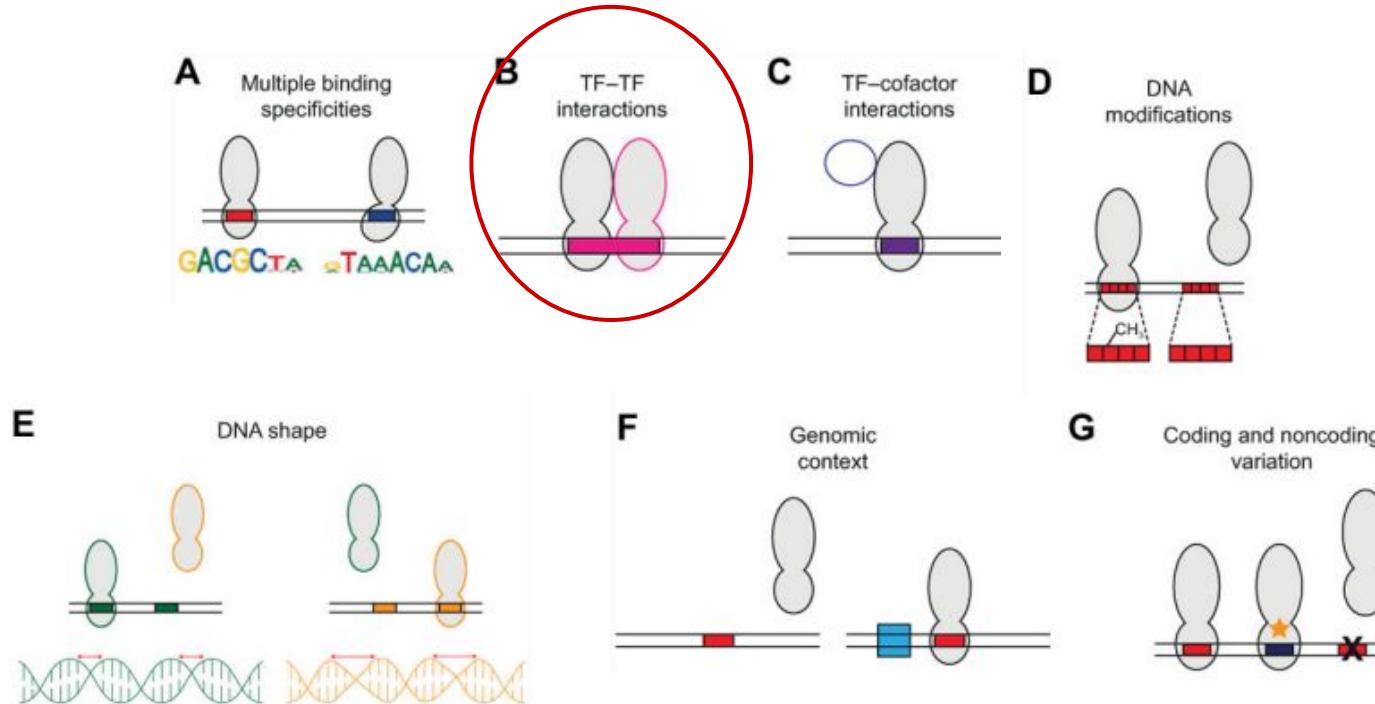
One binding mode to rule them all?



TFs like to interact with other TFs

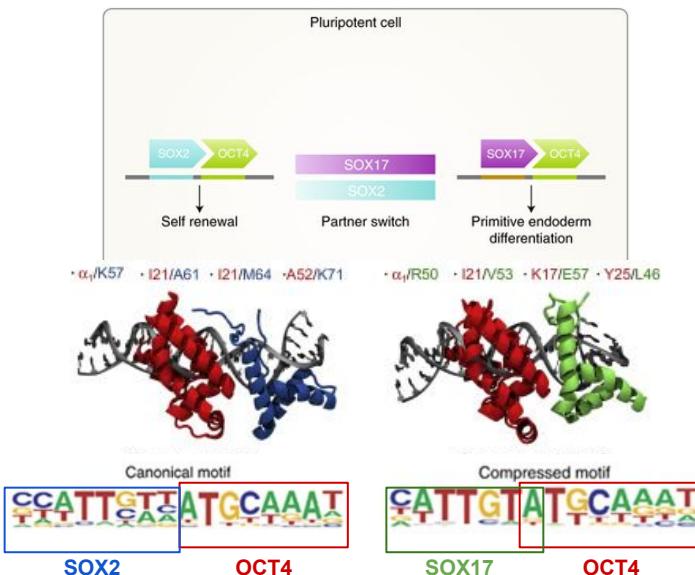


TF binding is also affected by other features



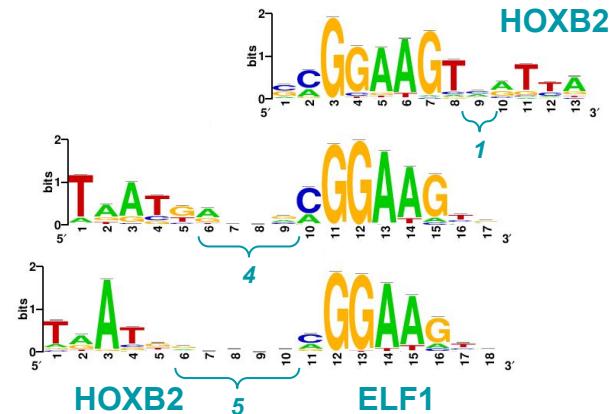
Cooperativity makes regulation very complex

TF cooperativity gives rise to TF binding combinations



Adapted from Aksoy et al., 2013
and Li & Belmonte, 2018

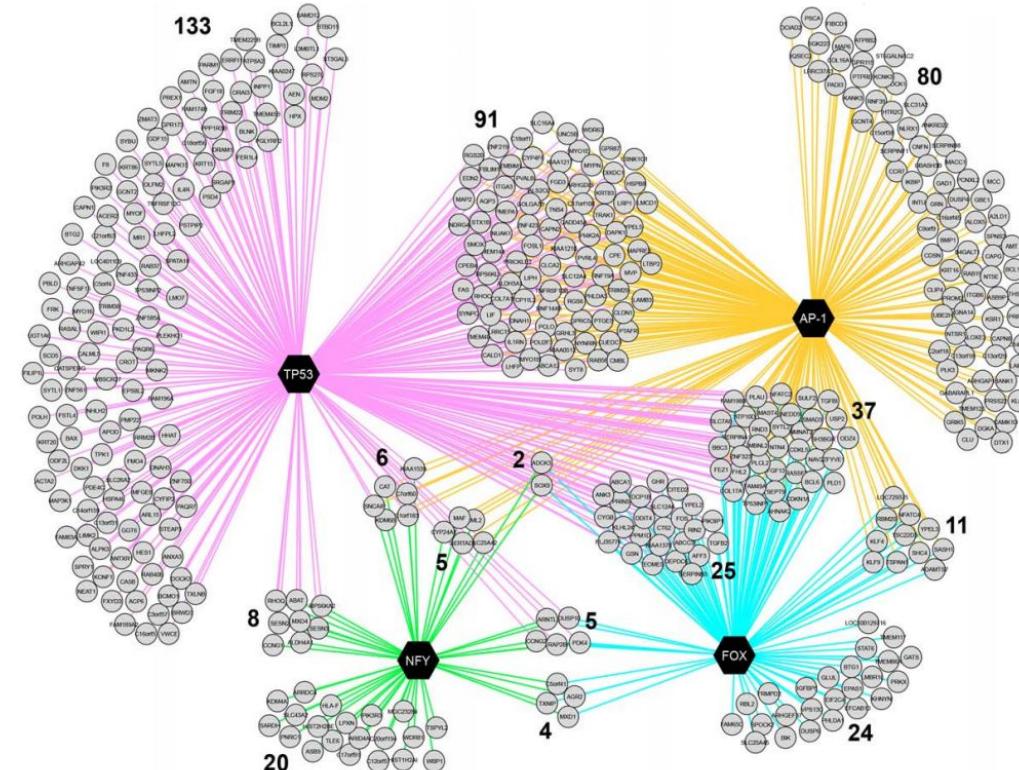
Different spacings and orientations of HOXB2 and ELF1¹



Adapted from Jolma et al., 2016

Transcription factors targets seen as networks

- One TF may regulate from ten to thousands of genes.
- One gene may be regulated by several TFs.



Summary

- TFs link signaling with gene expression.
- TFs bind DNA to activate or repress gene expression, but not all TF binding is functional.
- TFs may act alone or form complexes with other TFs.
- Co-factors, DNA modifications or DNA shape may alter TF binding.



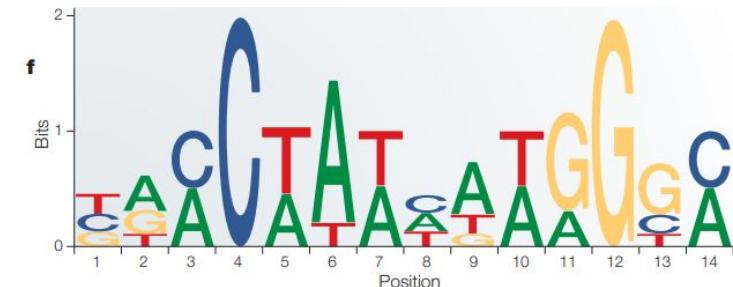
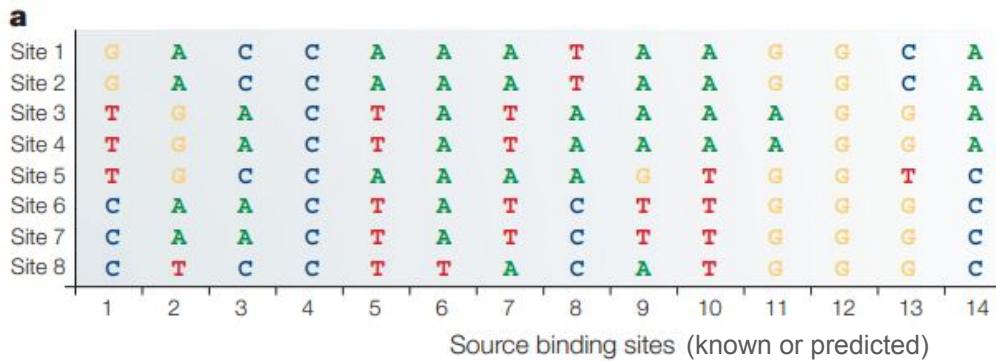
PART 2

Sequence-specific TF binding



DNA-binding Transcription Factors

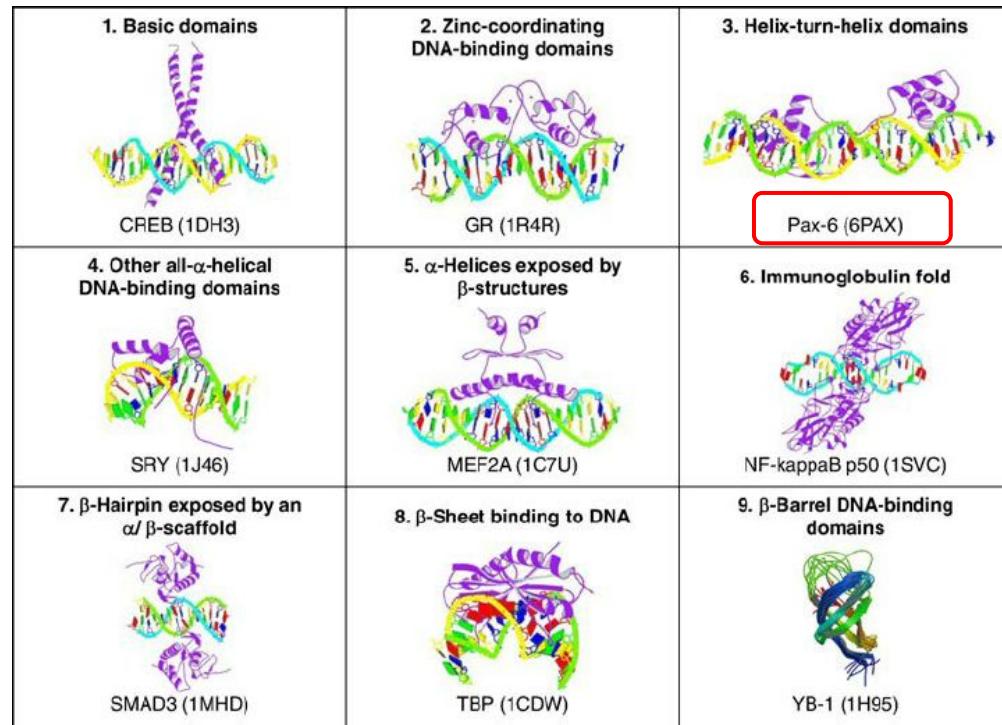
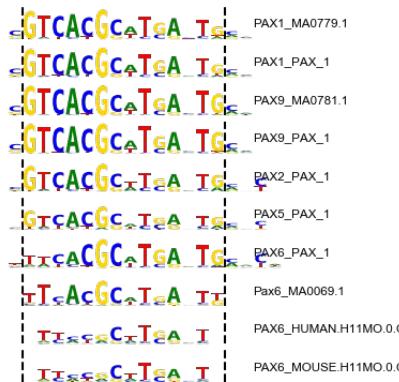
- One TF may bind thousands of different places in a genome: TF binding sites (**TFBSs**).
 - Each individual TFBS may be slightly different.
 - We can collect TFBSs to generate a model of the TF binding preference and find commonalities among the collected TFBSs.
-
- Position Frequency Matrix (PFM): simplest model, assumes nucleotide independency.



Example of a TF binding model

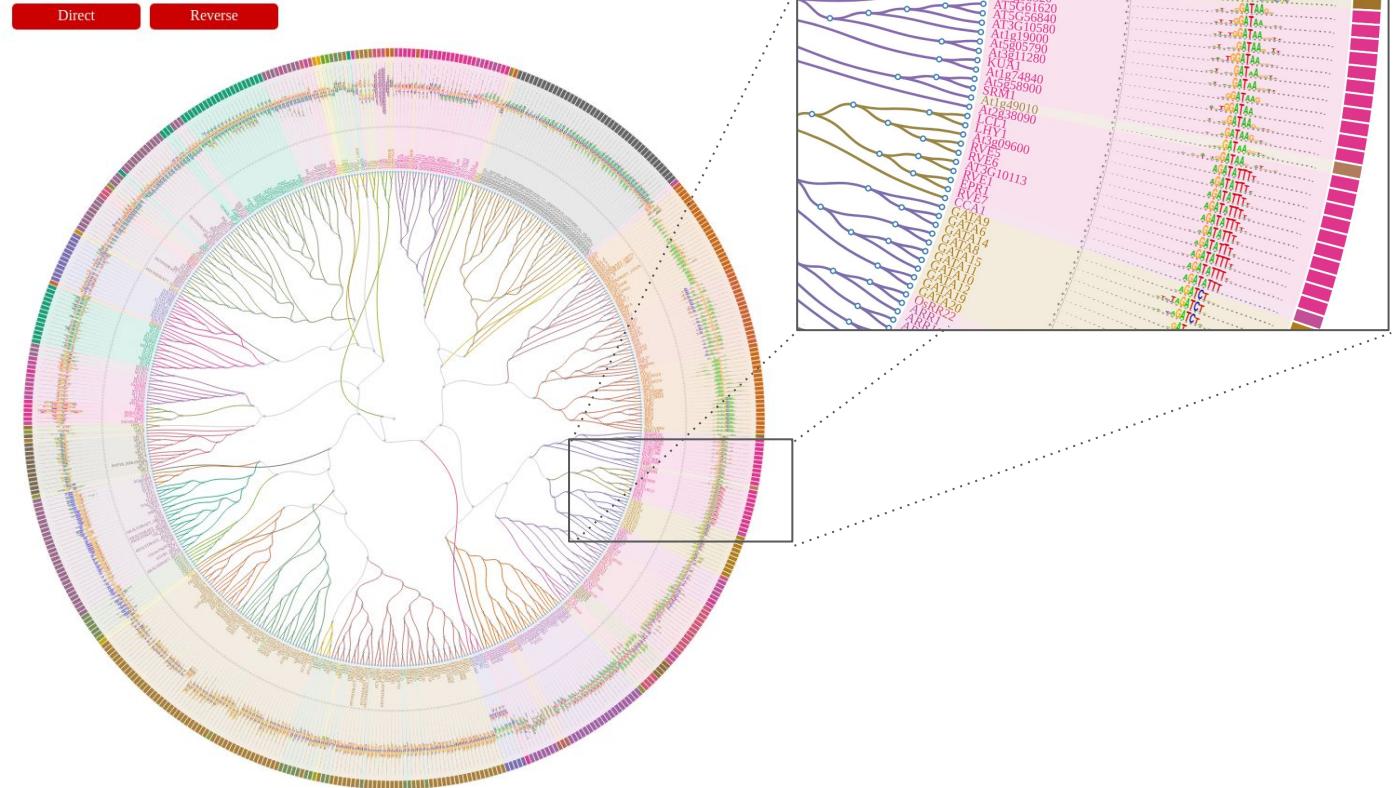
TFs are grouped by their DNA-binding domain sequence

- The mechanism of action is determined by the TF DNA-binding domain (DBD).
- TFs with similar (homologue) DBS tend to recognize similar sequences.



TFs with similar DBD have similar motifs

Color	TF Class
Green	Basic leucine zipper factors (bZIP)
Dark Green	B3 domain
Brown	AP2/ERF domain
Orange	C2H2 zinc finger factors
Dark Orange	MADS box factors
Red	A.T hook factors
Purple	Basic helix-loop-helix factors (bHLH)
Dark Purple	NAC/NAM
Light Purple	Homeo domain factors
Magenta	Tryptophan cluster factors
Dark Magenta	Helix-Turn-Helix
Pink	Other
Light Pink	Myb/SANT domain factors
Dark Grey	CG-1 domain
Grey	Beta-Hairpin-Ribbon
Light Grey	Winged Helix-Turn-Helix
Yellow	Zinc-coordinating
Light Yellow	SWIM-type zinc finger
Gold	Fork head / winged helix factors
Dark Gold	Other C4 zinc finger-type factors
Dark Brown	High-mobility group (HMG) domain factors
Medium Brown	Heat shock factors
Light Brown	SBP-type zinc finger
Black	WRKY

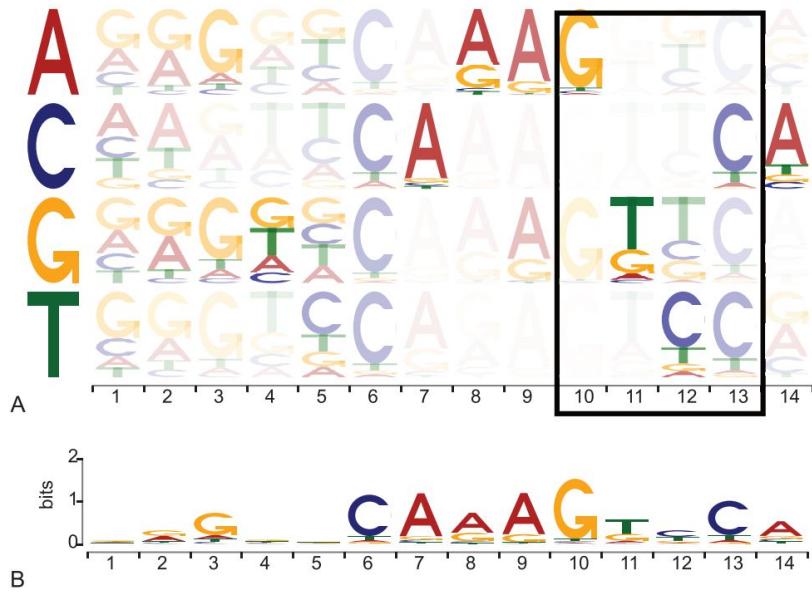


Complex representation of motifs

HMM

Mathelier 2013

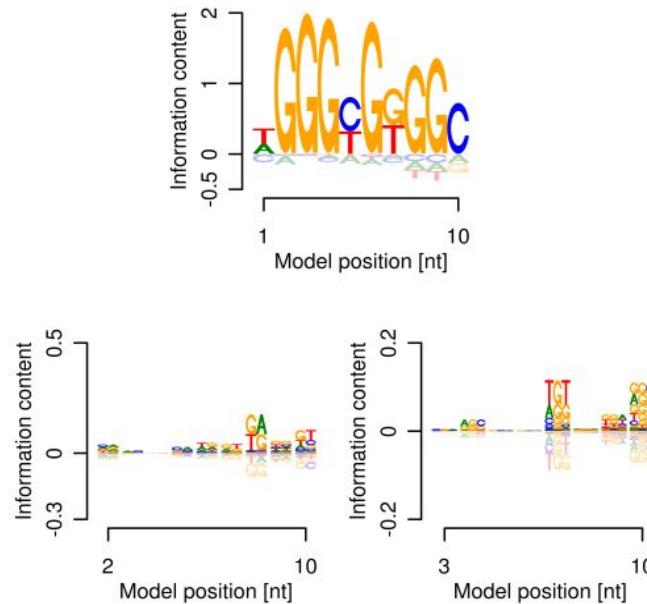
Nucleotide
dependencies
are modelled



Bayesian Markov Models

Siebert 2016

Supports long
k-mers



Complex representation of motifs

Dinucleotide PWMs

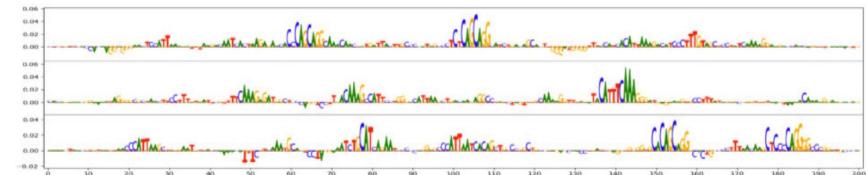
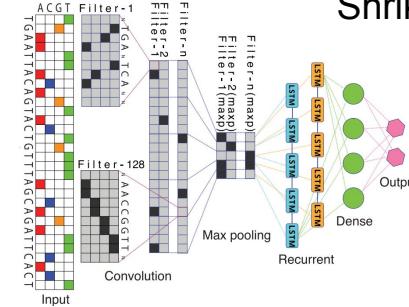
Kulakovskiy 2018



	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
01	38.0	11.0	44.0	0.0	51.0	15.0	10.0	3.0	86.0	15.0	43.0	3.0	70.0	22.0	53.0	3.0
02	0.0	0.0	0.0	245.0	0.0	0.0	0.0	63.0	0.0	0.0	0.0	150.0	0.0	0.0	0.0	9.0
03	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.0	450.0
04	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.0	0.0	1.0	8.0	46.0	0.0	354.0	50.0
05	0.0	54.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	353.0	1.0	0.0	0.0	58.0	0.0	0.0
06	0.0	0.0	1.0	0.0	423.0	5.0	34.0	3.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
07	10.0	5.0	4.0	404.0	0.0	0.0	0.0	6.0	0.0	2.0	0.0	33.0	0.0	0.0	0.0	3.0
08	1.0	9.0	0.0	0.0	0.0	7.0	0.0	0.0	0.0	4.0	0.0	0.0	9.0	436.0	0.0	1.0
09	9.0	0.0	1.0	0.0	454.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
10	7.0	59.0	186.0	212.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0
11	1.0	4.0	1.0	2.0	13.0	28.0	3.0	16.0	62.0	82.0	15.0	27.0	34.0	80.0	50.0	49.0
12	26.0	37.0	16.0	31.0	26.0	85.0	8.0	75.0	13.0	28.0	11.0	17.0	6.0	37.0	16.0	35.0
13	15.0	15.0	27.0	14.0	67.0	48.0	13.0	59.0	17.0	13.0	10.0	11.0	24.0	44.0	38.0	52.0

Motifs derived from deep learning models

Shrikumar 2019

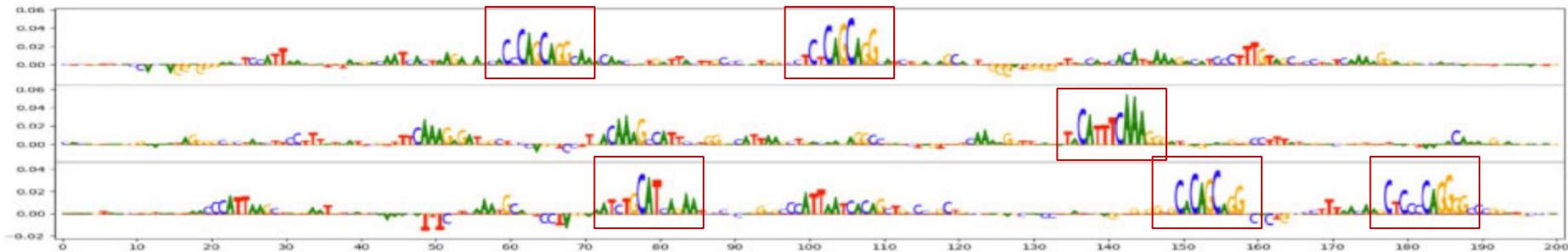


Finding nucleotides that discriminates between classes, e.g., bound/unbound

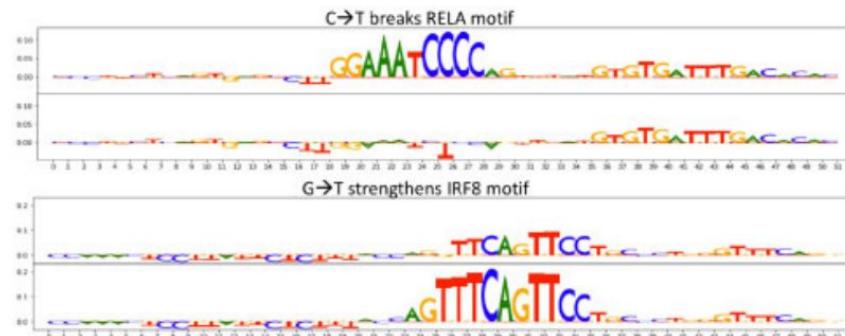
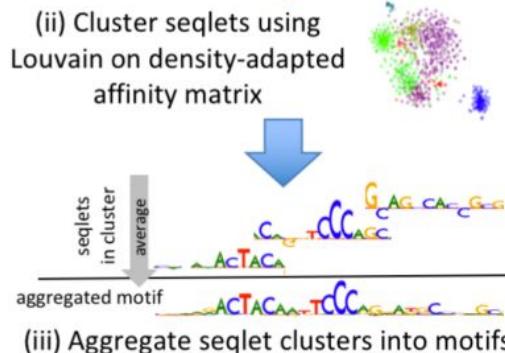
Complex representation of motifs

Motifs derived from deep learning or SVM or NMF models

Shrikumar 2019



Each nucleotide in the input sequences have a contribution score in the final layer of the NN.



TF binding motifs are represented in different formats

MEME version 5.4.0

MEME

ALPHABET= ACGT

strands: + -

Background letter frequencies (from unknown source):

A 0.290 C 0.210 G 0.210 T 0.290

MOTIF TACTGTATATAHAHMCAG MEME-1

letter-probability matrix: alength= 4 w= 18 nsites= 14 E= 3.7e-033

0.142857 0.000000 0.000000 0.857143

0.857143 0.000000 0.071429 0.071429

0.000000 1.000000 0.000000 0.000000

0.000000 0.000000 0.000000 1.000000

0.000000 0.000000 1.000000 0.000000

0.000000 0.000000 0.071429 0.928571

1.000000 0.000000 0.000000 0.000000

0.000000 0.071429 0.000000 0.928571

0.928571 0.000000 0.071429 0.000000

0.214286 0.000000 0.000000 0.785714

0.642857 0.071429 0.214286 0.071429

0.357143 0.285714 0.000000 0.357143

1.000000 0.000000 0.000000 0.000000

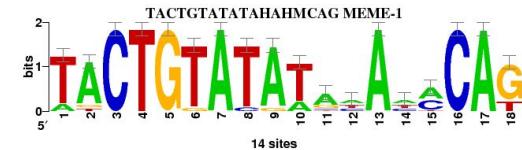
0.357143 0.285714 0.000000 0.357143

0.500000 0.428571 0.000000 0.071429

0.000000 1.000000 0.000000 0.000000

1.000000 0.000000 0.000000 0.000000

0.000000 0.000000 0.785714 0.214286



Many formats can be visualized as a logo

JASPAR

>Motif_1 USF1

A	[2	12	0	0	0	0	14	0	13	3	9	5	14	5	7	0	14	0]
C	[0	0	14	0	0	0	0	0	1	0	0	1	4	0	4	6	14	0	0
G	[0	1	0	0	14	1	0	0	1	0	3	0	0	0	0	0	0	11]
T	[12	1	0	14	0	13	0	13	0	11	1	5	0	5	1	0	0	3]

A	2.0	12.0	0.0	0.0	0.0	14.0	0.0	13.0	3.0	9.0	5.0	14.0	5.0	7.0	0.0	14.0	0.0	//
C	0.0	0.0	14.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	4.0	0.0	4.0	6.0	14.0	0.0	0.0	
G	0.0	1.0	0.0	0.0	14.0	1.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	
T	12.0	1.0	0.0	14.0	0.0	13.0	0.0	13.0	0.0	11.0	1.0	5.0	0.0	5.0	1.0	0.0	0.0	3.0

TRANSFAC

```

AC Motif_1
XX
ID USF1
XX
DE TACTGTATATAHAHMCAG
P0   A   C   G   T
1    2.0  0.0  0.0  12.0
2    12.0 0.0  1.0  1.0
3    0.0  14.0 0.0  0.0
4    0.0  0.0  0.0  14.0
5    0.0  0.0  0.0  0.0
6    0.0  0.0  1.0  13.0
7    14.0 0.0  0.0  0.0
8    0.0  1.0  0.0  13.0
9    13.0 0.0  1.0  0.0
10   3.0  0.0  0.0  11.0
11   9.0  1.0  3.0  1.0
12   5.0  4.0  0.0  5.0
13   14.0 0.0  0.0  0.0
14   5.0  4.0  0.0  5.0
15   7.0  6.0  0.0  1.0
16   0.0  14.0 0.0  0.0
17   14.0 0.0  0.0  0.0
18   0.0  0.0  11.0 3.0
XX
CC program: meme
CC matrix_nb: 1
CC id: TACTGTATATAHAHMCAG
CC ac: MEME_1
CC name: TACTGTATATAHAHMCAG
CC residue_type:
CC sites: 14
CC meme_E-value: 3.7e-033
CC matrix_nb: 1
CC consensus.strict: TACTGTATATAAccCAG
CC consensus.strict_rc: CTGGGTATATACAGTA
CC consensus.IUPAC: TACTGTATATAhAhCAG
CC consensus.IUPAC_rc: CTGKDTDYATATACAGTA
CC consensus-regexp: TACTGTATAT[ag][act][ac]CAG
CC consensus-regexp_rc: CTG[GT][AGT]T[AGT][CT]ATATACAGTA
CC consensus.//: TAB
XX
//
```

TAB

JASPAR

- Public database of **curated** (human-verified) TF binding motifs.
- Motifs for 6 taxonomic groups.
- <https://jaspar.genereg.net/>

Detailed information of matrix profile **MA0143.4**

Profile summary Add

Name: SOX2
Matrix ID: MA0143.4
Class: High-mobility group (HMG) domain factors
Family: SOX-related factors
Collection: CORE
Taxon: Vertebrates
Species: Homo sapiens
Data Type: ChIP-seq
Validation: 15863505
Uniprot ID: P48431
Source: ReMap
Comment:

Sequence logo Download SVG

Frequency matrix

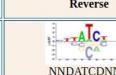
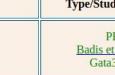
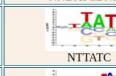
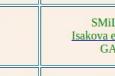
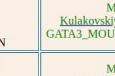
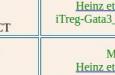
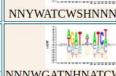
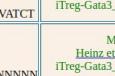
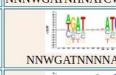
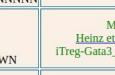
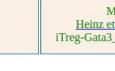
	JASPAR	TRANSFAC	MEME	RAW PFM	Reverse comp.							
A[23321	21971	10304	1656	79277	1185	2412	4848	1385	23164	24174]
C[15998	17765	54414	73372	930	2002	550	1243	418	17157	21434]
G[17475	17158	6264	800	732	1742	1205	78356	436	15154	11041]
T[28960	28860	14772	9926	4815	80825	81587	1307	83515	30279	29105]

Browse JASPAR CORE for 6 different taxonomic groups

Fungi
Insecta
Nematoda
Plantae
Urochordata
Vertebrata

CIS-BP

- The largest public database of TF binding motifs.
- Motifs for >700 organisms: in total >390,000 TFs.
- Most motifs are inferred.
- <http://cisbp.ccbr.utoronto.ca/>

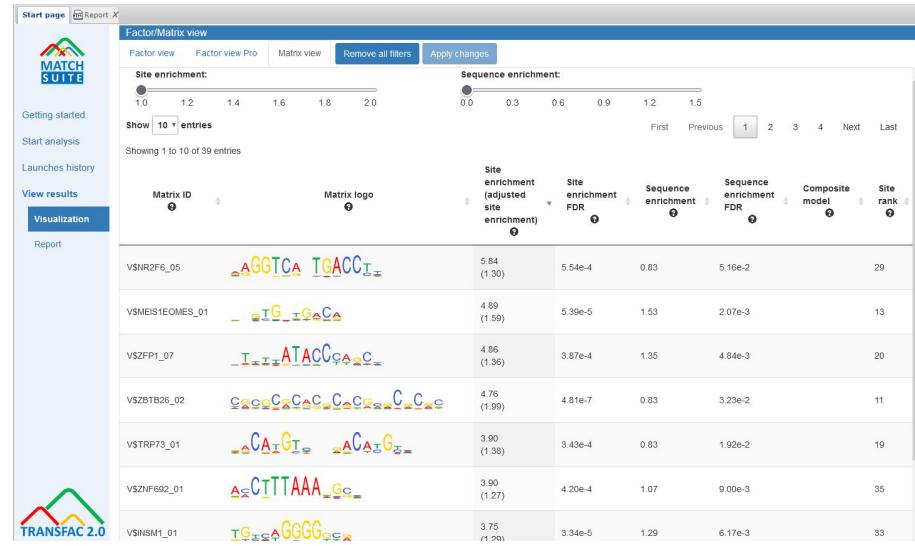
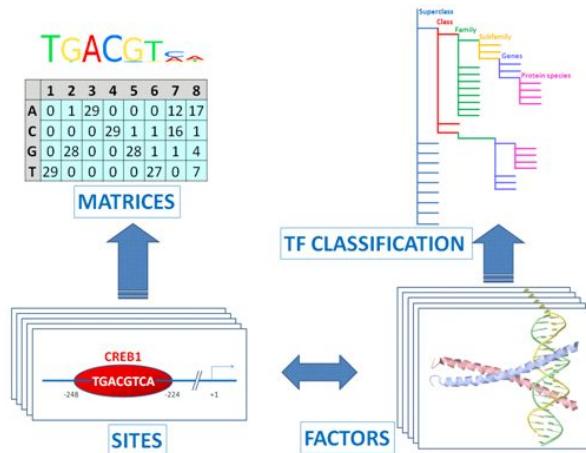
Name/Motif ID	Species	Forward	Reverse	Type/Study/Study ID	SR Score	DBD Identity
Gata3 M00166_2.00	<i>Mus musculus</i>			PBM Badis et al.(2009) Gata3_1024	(Direct)	(Direct)
Gata3 M05867_2.00	<i>Mus musculus</i>			SMILE-seq Isakova et al.(2017) GATA3	(Direct)	(Direct)
Gata3 M09122_2.00	<i>Mus musculus</i>			Misc Kulakovskiy et al.(2013) GATA3_MOUSE.H1MO.0.A	(Direct)	(Direct)
Gata3 M09551_2.00	<i>Mus musculus</i>			Misc Heinz et al.(2010) iTreg-Gata3_GSE20898_1	(Direct)	(Direct)
Gata3 M09552_2.00	<i>Mus musculus</i>			Misc Heinz et al.(2010) iTreg-Gata3_GSE20898_2	(Direct)	(Direct)
Gata3 M09553_2.00	<i>Mus musculus</i>			Misc Heinz et al.(2010) iTreg-Gata3_GSE20898_3	(Direct)	(Direct)
Gata3 M09554_2.00	<i>Mus musculus</i>			Misc Heinz et al.(2010) iTreg-Gata3_GSE20898_4	(Direct)	(Direct)
Gata3 M09555_2.00	<i>Mus musculus</i>			Misc Heinz et al.(2010) iTreg-Gata3_GSE20898_5	(Direct)	(Direct)

Name	Species	Gene ID	Family	Motif Evidence
<u>GATA3</u>	<i>Arabidopsis thaliana</i>	AT4G34680	GATA	Direct
<u>GATA3</u>	<i>Homo sapiens</i>	ENSG00000107485	GATA	Direct
<u>Gata3</u>	<i>Mus musculus</i>	ENSMUSG00000015619	GATA	Direct
<u>GATA3</u>	<i>Anolis carolinensis</i>	ENSACAG00000009109	GATA	Inferred
<u>GATA3</u>	<i>Arabidopsis lyrata</i>	fgenesh2_kg.7_620_	GATA	Inferred
<u>GATA3</u>	<i>Bos taurus</i>	ENSBTAG00000017243	GATA	Inferred
<u>GATA3</u>	<i>Canis familiaris</i>	ENSCAFG00000005019	GATA	Inferred
<u>GATA3</u>	<i>Cavia porcellus</i>	ENSCPOG00000008595	GATA	Inferred
<u>GATA3</u>	<i>Chloepus hoffmanni</i>	ENSCHOG00000005894	GATA	Inferred
<u>gata3</u>	<i>Danio rerio</i>	ENSDARG00000016526	GATA	Inferred
<u>GATA3</u>	<i>Dasypus novemcinctus</i>	ENSDNOG00000042141	GATA	Inferred

Inferred motifs are predicted based on DNA-binding domains similarity.

Transfac database

- The oldest database of TF binding motifs.
- Private database.
- The motifs and the TFBS are manually curated
- ~9000 motifs
- <https://genexplain.com/transfac-2-0/>



Summary

- TF DNA binding motifs can be represented using various models.
- It is important to know what your model represents in order to properly interpret the TF binding.
- TF DNA binding motifs can be represented in different formats.
- Multiple resources are storing TF DNA binding motifs, but it is important to know where that data is coming from and how it was generated.

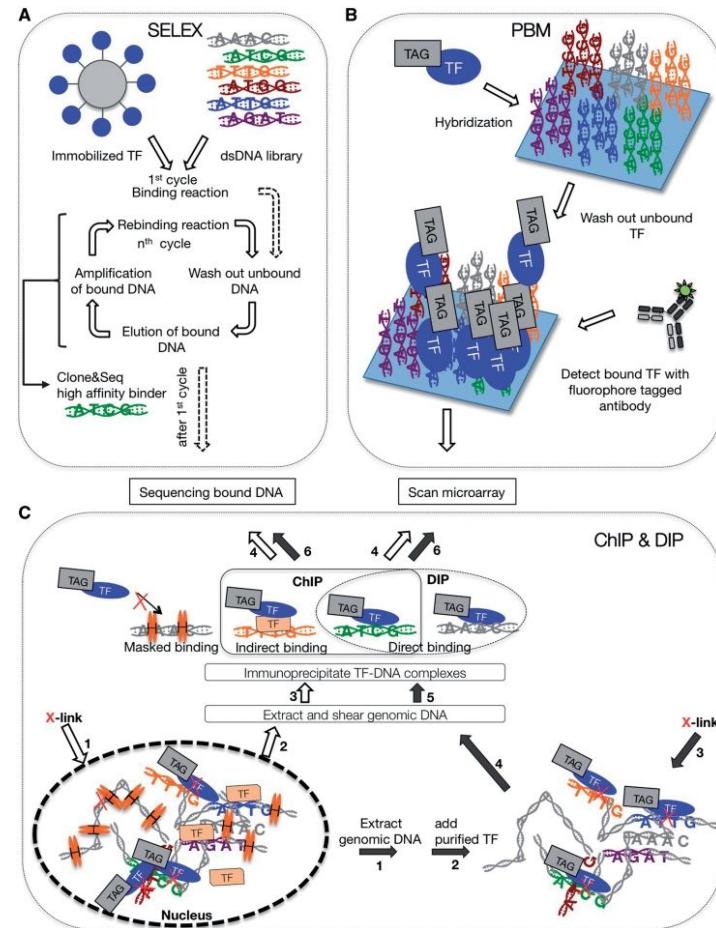


PART 3

Capturing TF binding

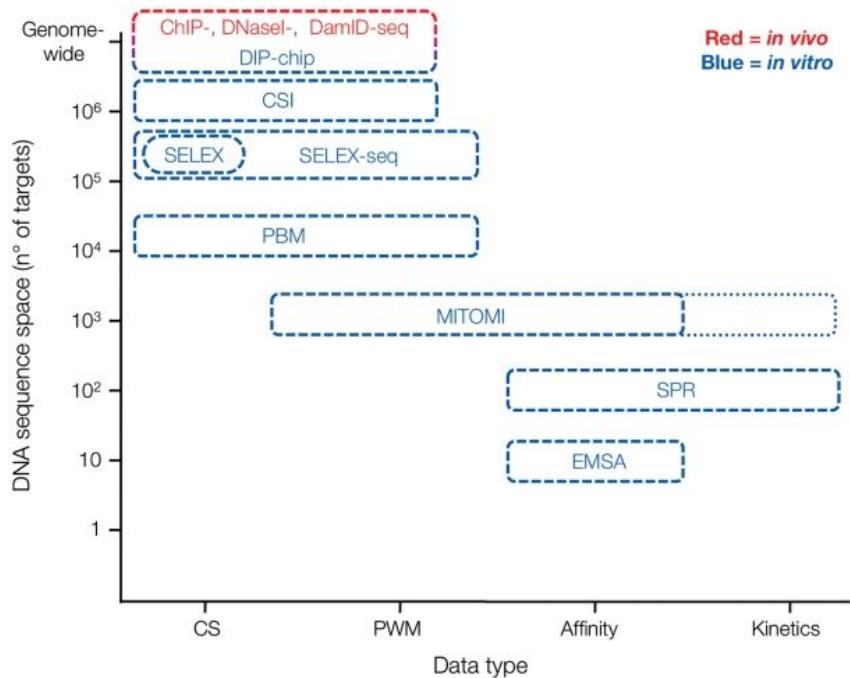
Why a bioinformatician should care about experimental techniques and their nuances?

- Data processing depends on the used protocol and experimental setting
- Interpretation of the results highly depends on experimental setting
- Sometimes devil is in the details!
- Ideal scenario - collaboration between wet and dry labs **to plan the experiment**

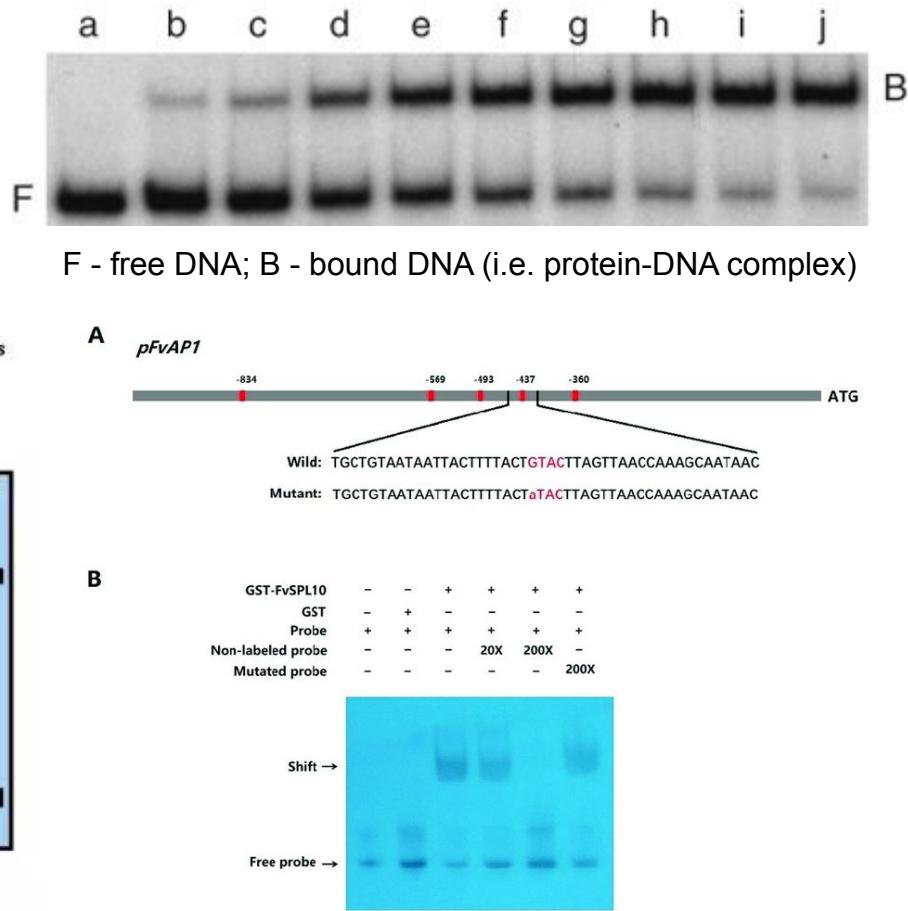
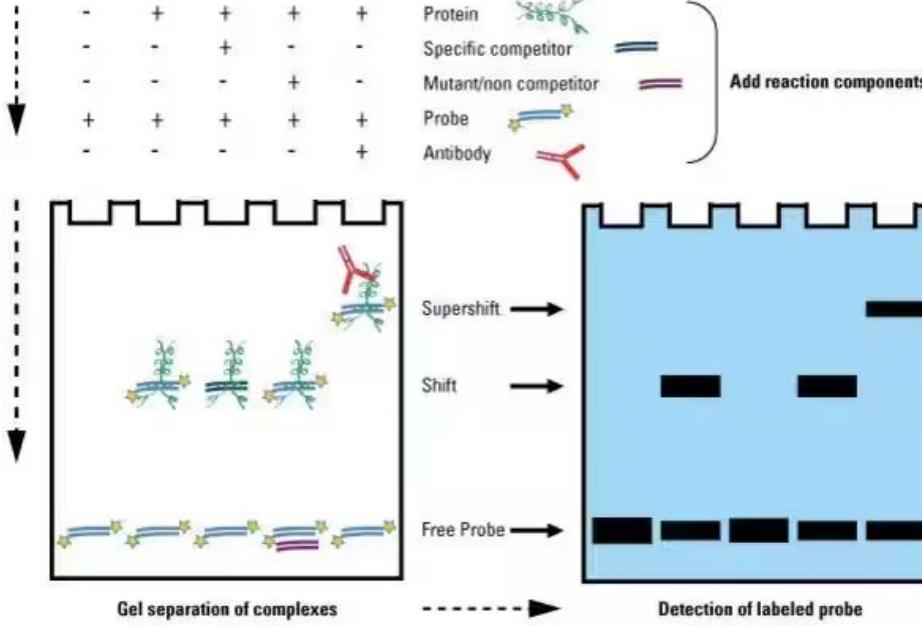


In vitro and *In vivo* techniques

- *In vitro*:
 - Low-throughput:
 - EMSA
 - Mid-throughput:
 - DAP-seq
 - High-throughput:
 - PBM
 - SELEX-based methods
- *In vivo*:
 - ChIP-based techniques
 - All techniques are low-throughput



EMSA: electrophoretic mobility shift assay

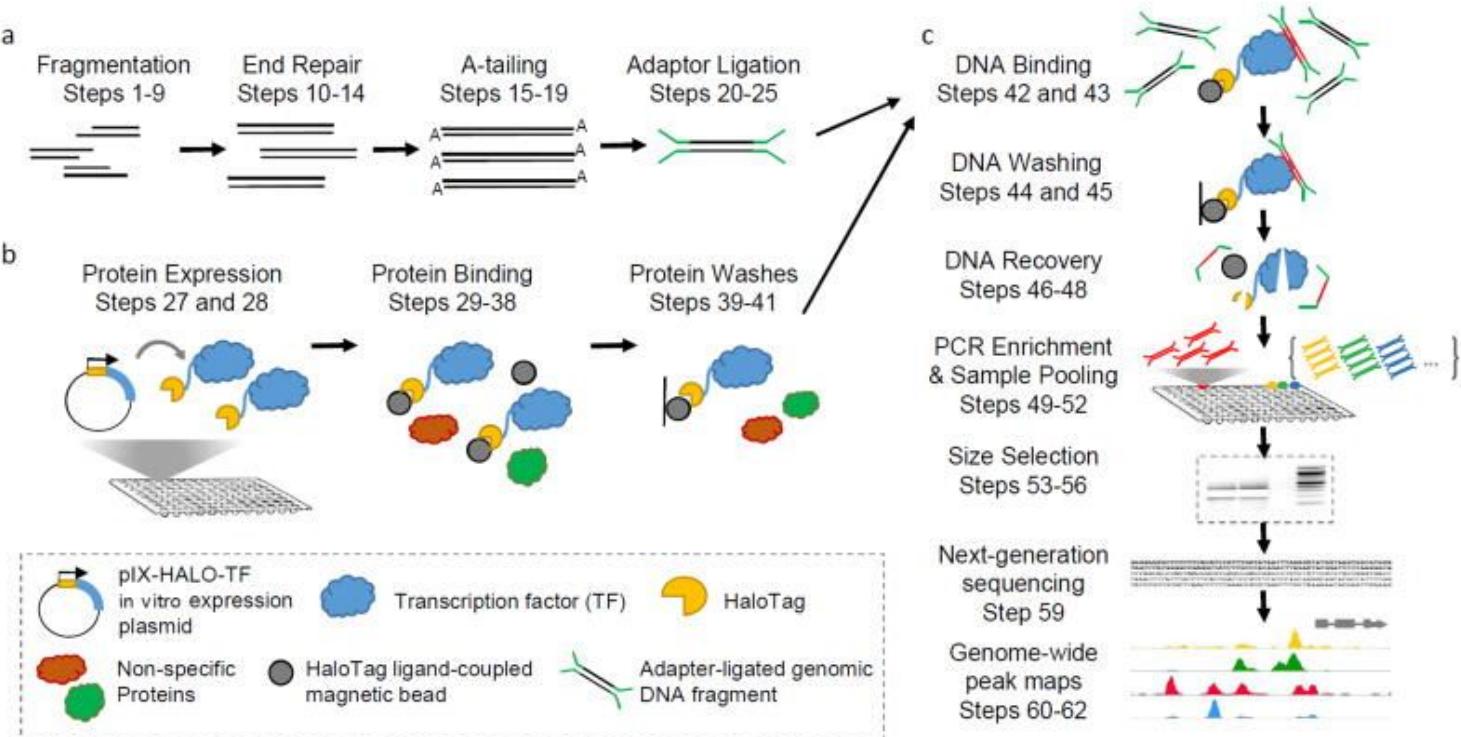


Xiong et al. 2019. Molecular cloning and characterization of SQUAMOSA-promoter binding-like gene FvSPL10 from Woodland Strawberry (*Fragaria vesca*)
 Hellman LM and Fried MG 2007. Electrophoretic mobility assay (EMSA) for detecting protein-nucleic acid interactions

<https://www.thermofisher.com/no/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/gel-shift-assays-emsa.html>

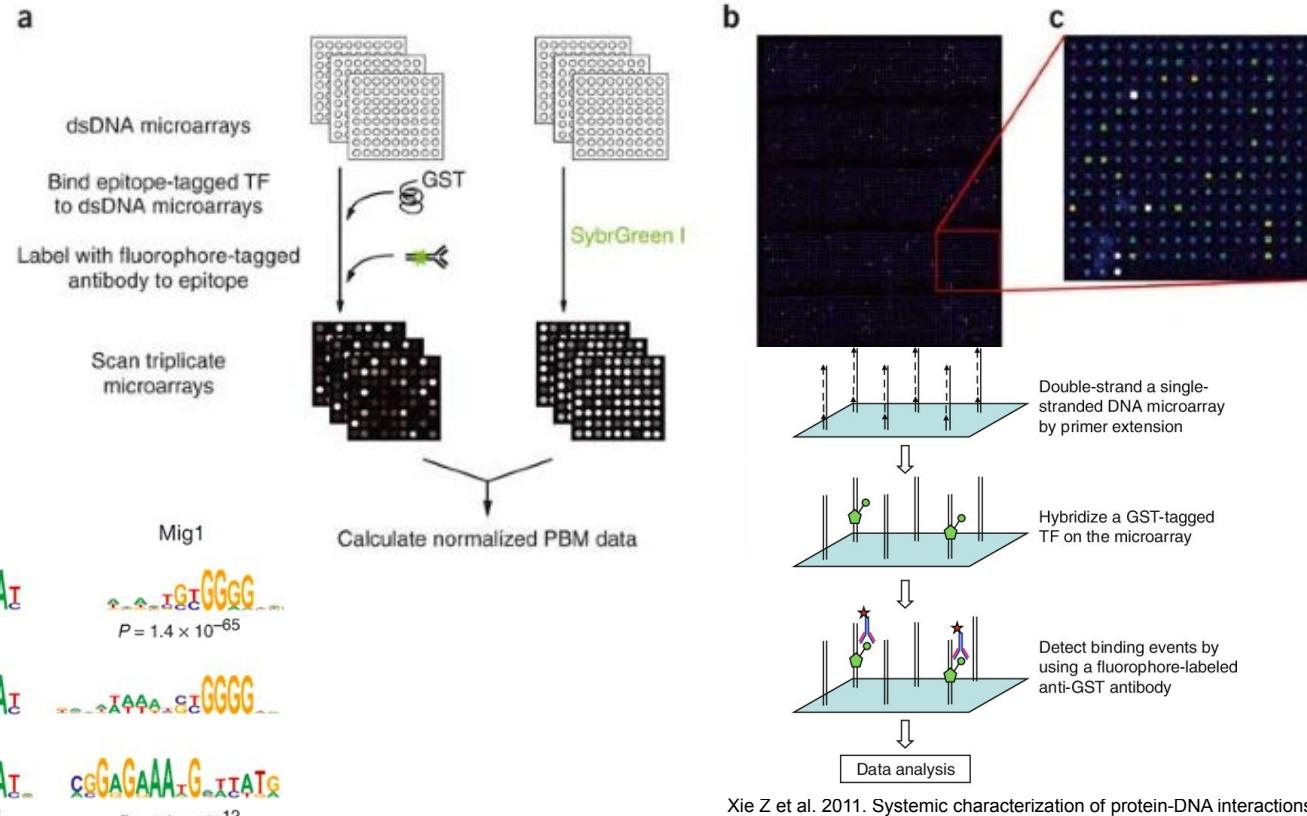
DAP-seq: DNA affinity purification sequencing

- Scalable
- Cheap
- *In vitro*



PBM: protein binding microarray

- High-throughput
- *In vitro*
- Quick



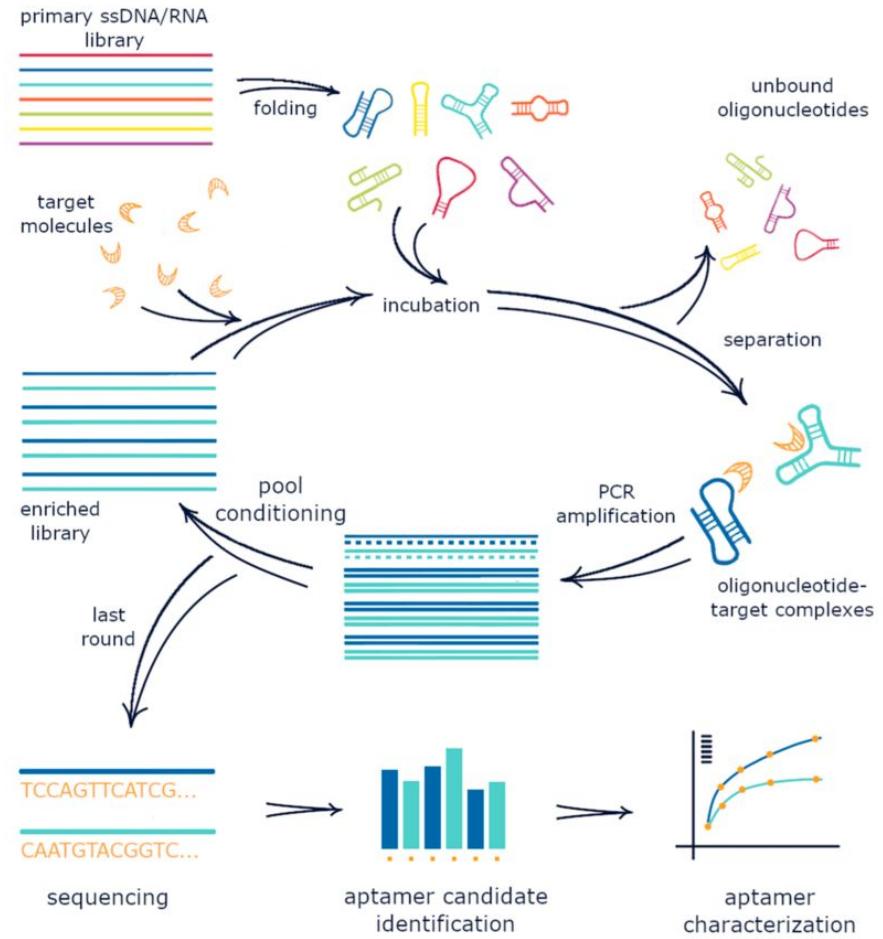
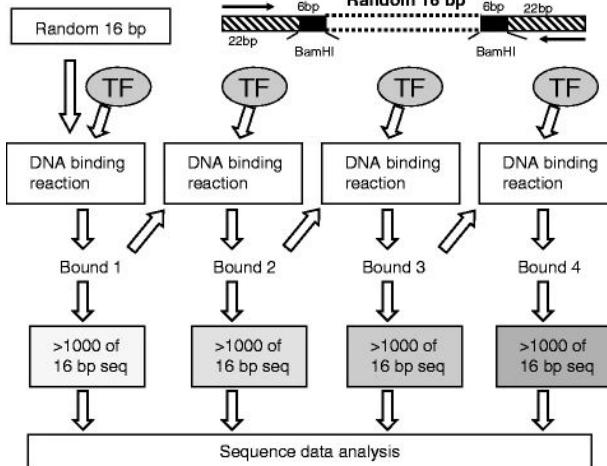
PBM

	Rap1	Abf1	Mig1
PBM			
TRANSFAC			
ChIP-chip*			

Xie Z et al. 2011. Systemic characterization of protein-DNA interactions
Mukherjee S et al. 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays

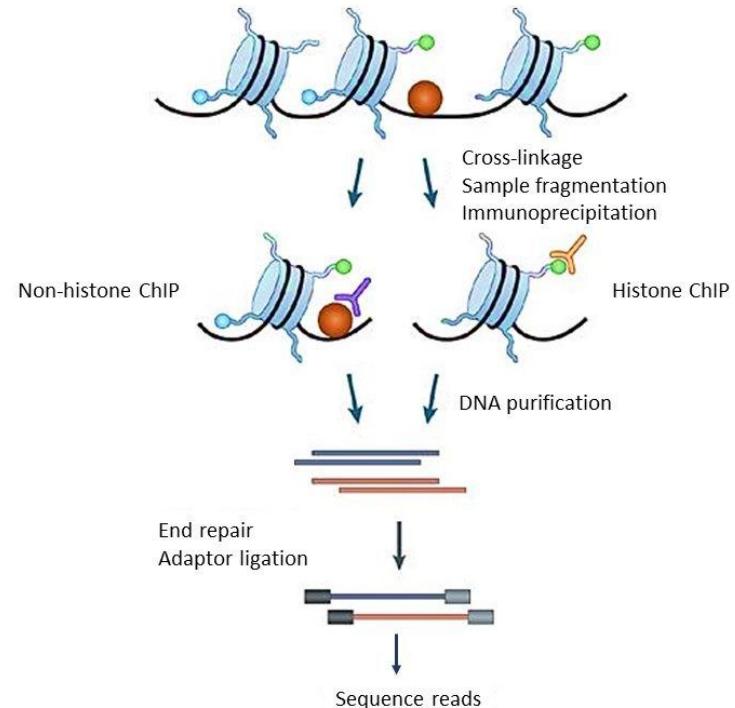
SELEX: systemic evolution of ligands by exponential enrichment

- *In vitro*
- Could be applied for:
 - cooperative binding predictions
 - differential TF binding
- HT-SELEX: high throughput SELEX



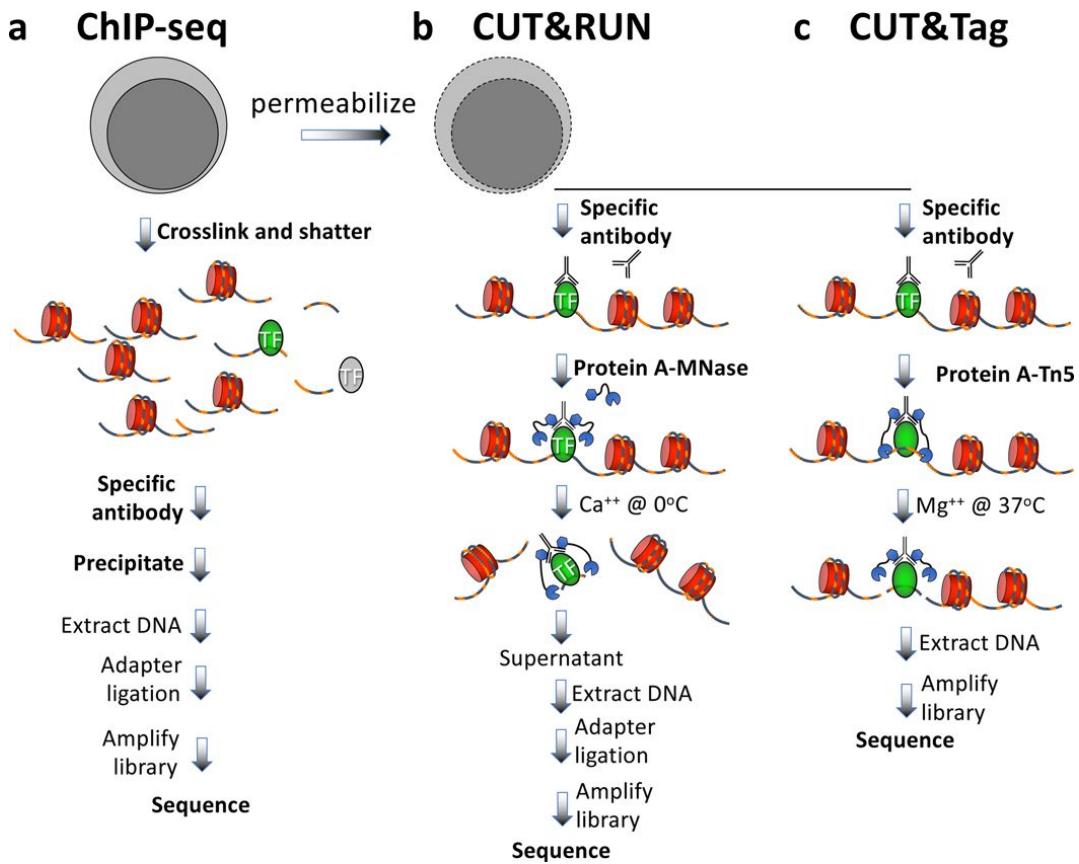
ChIP-seq: chromatin immunoprecipitation followed by sequencing

- Still is probably the main technique to map TF binding
- Major limitation - one TF/histone modification per sample
- Genome-wide
- Unlike arrays does not require prior knowledge



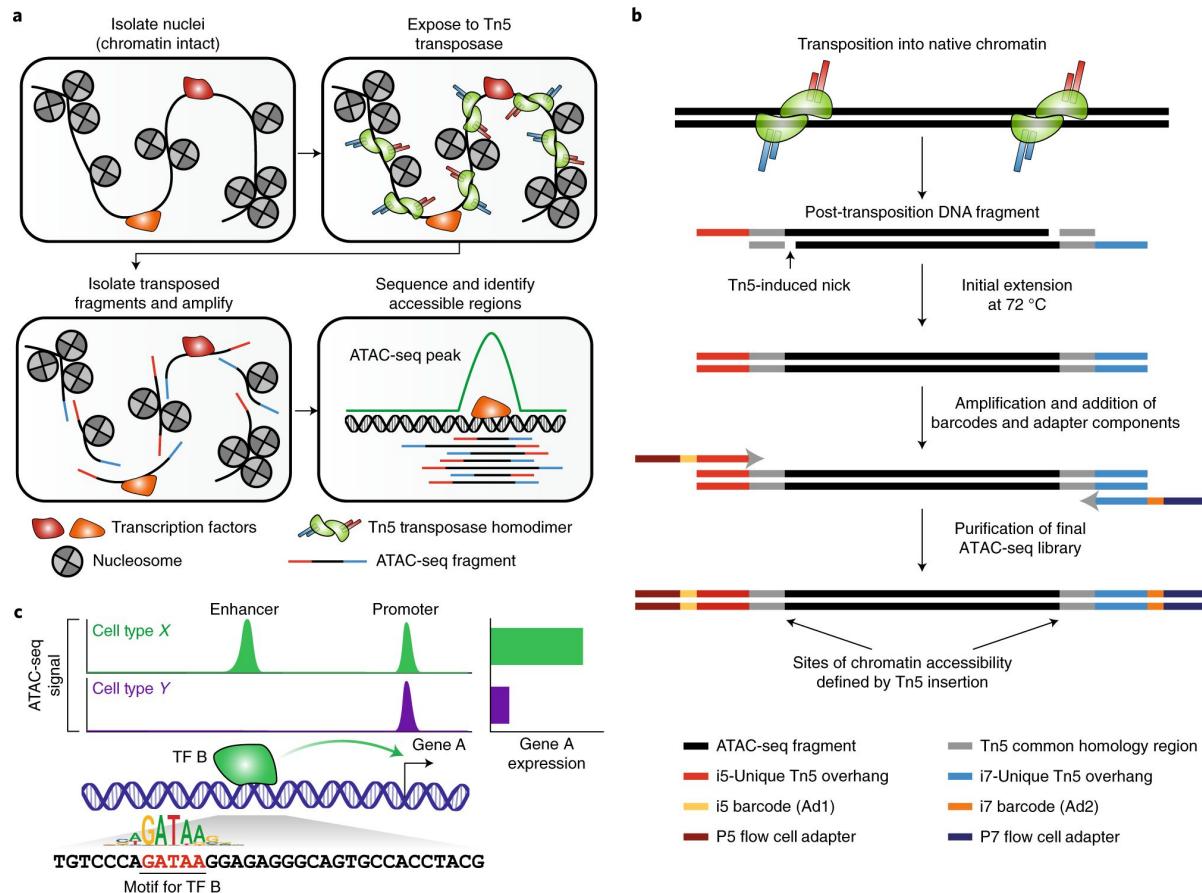
CUT&Tag/CUT&RUN: cleavage under targets and tagmentation/release under nuclease

- For small samples
- Possibility to map in single-cells
- Quite needy protocol and data analysis



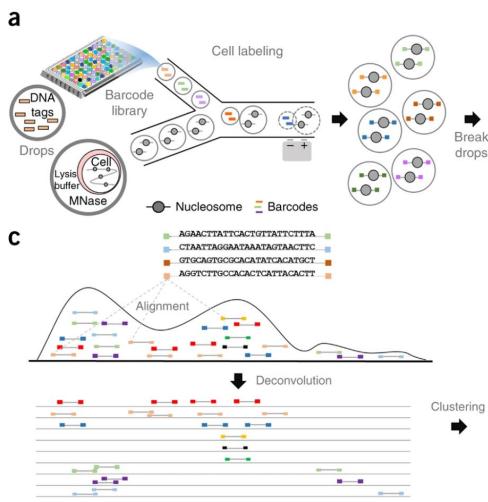
ATAC-seq: Assay for Transposase-Accessible Chromatin using sequencing

- Mapping open chromatin
- Inferring the binding of multiple TFs: what are advantages and disadvantages of this?

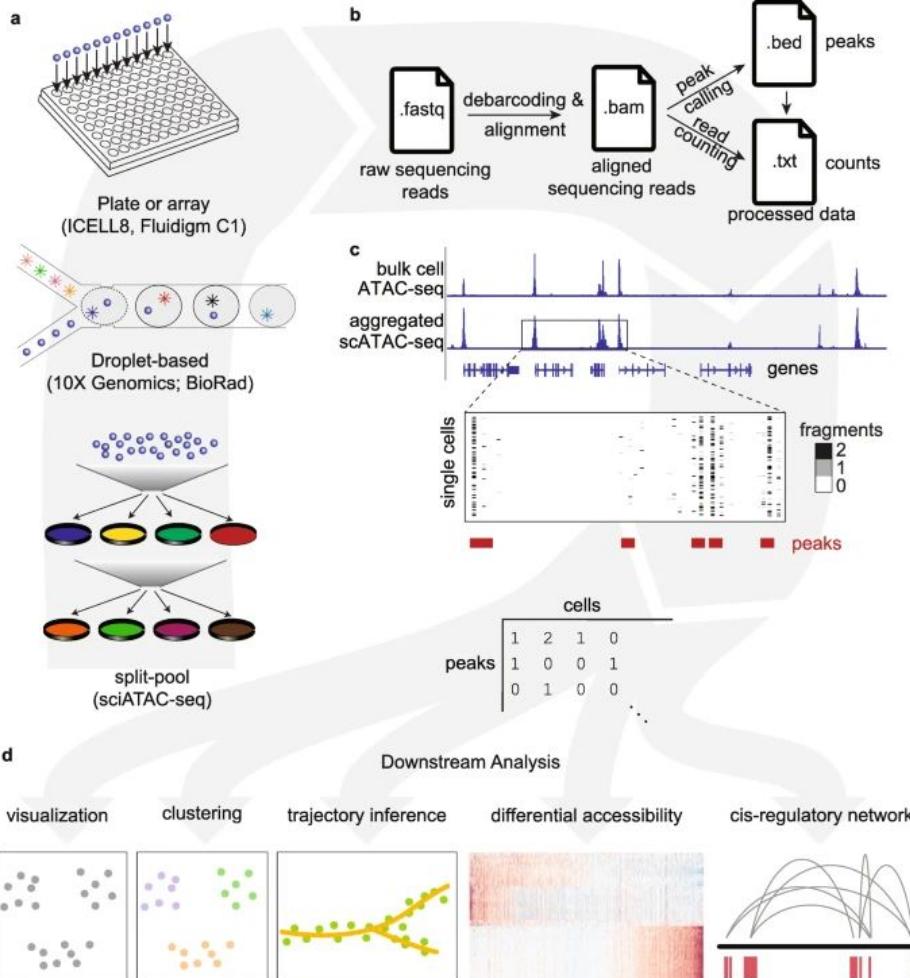


Single-cell techniques

scATAC-seq



Drop-ChIP/scChIP-Seq



Grosselin K et al. 2019. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer
Chen H et al. 2019. Assessment of computational methods for the analysis of single-cell ATAC-seq data

Summary. Which method to choose?

- What do you want to study? What is your study design and hypothesis?
- What is your budget?
- Sometimes... how fancy do you want to be?

- Do you need a new experiment at all? Have you explored available data?

- It is probably more important to have a proper experimental design



PART 4

TF binding data analysis

TF binding data analysis: general principles

- If not produced with array or gel-based technique, most of the resulting raw data today is **sequencing data**.
- Data analysis starts with processing raw reads in FASTA format (aka sequence alignment to the reference genome) and performing quality assessments.
- Sequencing facilities most likely will do the quality assessment, some might perform the alignment and/or some of the downstream analysis.
- After alignment there are several things one can do...

ChIP-seq data analysis

- Raw data analysis:
 - Quality control
 - Adapter trimming
 - Sequence alignment
 - Duplicate removal
- Peak calling
- Downstream analysis:
 - Visualization

Visualizing the data in Genome Browser



[NR3C1 ChIP-seq experiments in ENCODE](#)

Kaya-Okur et al., 2020. CUT&Tag for efficient epigenomic profiling of small samples and single cells

ChIP-seq data analysis

- Raw data analysis:
 - Quality control
 - Adapter trimming
 - Sequence alignment
 - Duplicate removal
 - Peak calling
 - Downstream analysis:
 - Visualization
-  Raw signals and peaks make sense, the quality of ChIP-seq experiment seems good! **What's next?**

ChIP-seq data analysis

- Raw data analysis:
 - Quality control
 - Adapter trimming
 - Sequence alignment
 - Duplicate removal
- Peak calling
- Downstream analysis:
 - Visualization
 - Peak analysis
 - Motif enrichment/discovery
 - Annotations
 - Functional enrichment analysis

What to do when you **already know** that ChIP'ed protein is a TF and motif is known?

Let's say you are analyzing
GATA1 TF

⇒ you have ChIP-seq peaks where
GATA1 TF was targeted

What's next?

1. What is the expected motif?



2. Do you find the expected motif?

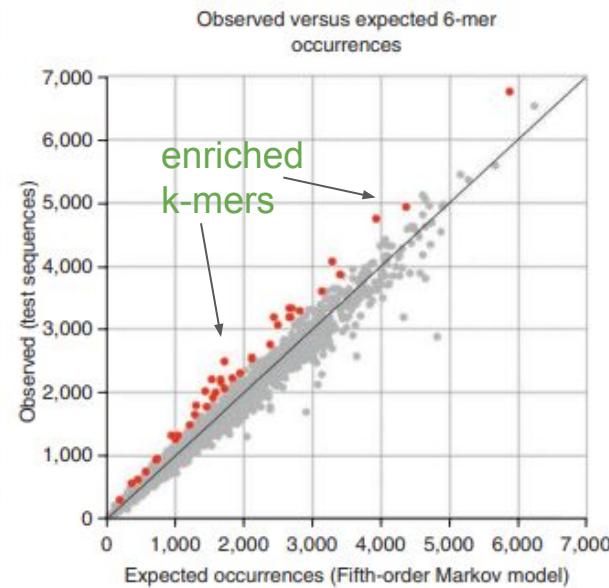
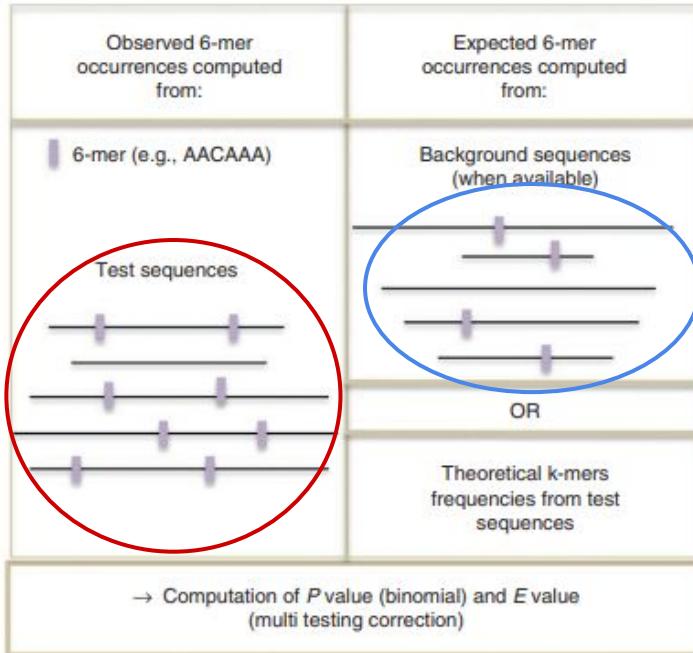
CAREFUL! ⇒ Do *de novo* motif discovery to check what is in your peaks

3. Which genomic regions have the expected motif?

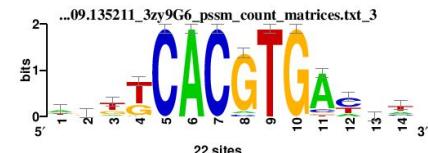
⇒ Motif scanning

De novo motif discovery

Step 1: detection of over-represented k-mers
(**observed** vs **expected**.)



assembly #	seed:	score
1	aliant	2.22
	rev_cpl	2.09
	...acgtgac	0.29
	...cacgtac	10.78
cgtgac	4.85
cacgtg..	10.78
acgtga.	9.00
tcacgt..	2.09
cacgtg..	4.85
cacgtgac	2.22
acgtga.	0.29
acgtgac	10.78
	gtcacgtgac	best consensus



ChIPMunk

- Web application and command line
- <https://opera.autosome.org/chipmunk/discovery/mono/new>

MACRO-APE ▾ PERFECTOS-APE ▾ ChIPMunk ▾ Databases ▾ Contacts

Ticket number:

Toy examples for the [simple](#), [weighted](#), [peak](#) data sets.

Mode:

Paste a list of sequences in a textarea or select a file with sequences to scan

Sequence set in extended multi-fasta format:

```
> 1 example data: 48 Drosophila bicoid footprints
CCGATCGAAGGGATTATAATTCC
> 2
GTACTAATTAAAGCATGGCTCAAG
> 3
TGGGCGGATTTGCATATGG
> 4
GCGTCGGATTTGCAGCCAT
> 5
TGGCCAAGGATTGTCCGTGGG
> 6
```

or load a file with sequences: No file chosen

Advanced options: [\[+\]](#)

 De novo motif discovery in nucleotide sequences.

Click a field in the form on the left to get a hint

MEME

- Only takes the first 1000 regions!
- Web application and command line
- <https://meme-suite.org/meme/tools/meme>

The screenshot shows the MEME Suite 5.5.0 web application. At the top right, there's a magnifying glass icon and the text "MEME Multiple Em for Motif Elicitation Version 5.5.0". To the right of this, a sidebar contains the following navigation links:

- Motif Discovery
 - MEME
 - STREME
 - XSTREME
 - MEME-ChIP
 - GLAM2
 - MoMo
 - DREME (deprecated)
- Motif Enrichment
- Motif Scanning
- Motif Comparison
- Gene Regulation
- Utilities
- Manual
- Guides & Tutorials
- Sample Outputs
- File Format Reference
- Databases
- Download & Install
- Help
- Alternate Servers

Authors & Citing

Recent Jobs

Clear All

« Previous version 5.4.1

On the right side, the main content area is titled "Data Submission Form". It includes sections for "Select the motif discovery mode" (with "Classic mode" selected), "Select the sequence alphabet" (with "DNA, RNA or Protein" selected), "Input the primary sequences" (with a file input field showing "No file chosen" and a "Depressed" status indicator), "Select the site distribution" (with "Zero or One Occurrence Per Sequence (zoops)" selected), "Select the number of motifs" (with a value of "3" entered), "Input job details" (with fields for email address and job description), and "Advanced options" (with a note about file size limits). At the bottom, there are "Start Search" and "Clear Input" buttons, along with copyright and citation information.

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. See this Manual for more information.

MEME Suite 5.5.0

Motif Discovery

Motif Enrichment

Motif Scanning

Motif Comparison

Gene Regulation

Utilities

Manual

Guides & Tutorials

Sample Outputs

File Format Reference

Databases

Download & Install

Help

Alternate Servers

Authors & Citing

Recent Jobs

Clear All

« Previous version 5.4.1

Version 5.5.0

Please send comments and questions to: meme-suite@uw.edu

Powered by [Opal](#)

Data Submission Form

Perform motif discovery on DNA, RNA, protein or custom alphabet datasets.

Select the motif discovery mode [?](#)

Classic mode Discriminative mode Differential Enrichment mode

Select the sequence alphabet

Use sequences with a standard alphabet or specify a custom alphabet. [?](#)

DNA, RNA or Protein Custom [Choose File](#) No file chosen

Input the primary sequences

Enter sequences in which you want to find motifs. [?](#)

[Upload sequences](#) [Choose File](#) No file chosen

Select the site distribution

How do you expect motif sites to be distributed in sequences? [?](#)

[Zero or One Occurrence Per Sequence \(zoops\)](#)

Select the number of motifs

How many motifs should MEME find? [?](#)

3

Input job details

(Optional) Enter your email address. [?](#)

(Optional) Enter a job description. [?](#)

Advanced options

Note: if the combined form inputs exceed 80MB the job will be rejected.

Start Search Clear Input

STREME

- Web application and command line
- <https://meme-suite.org/meme/tools/streme>

MEME Suite 5.5.0

▼ Motif Discovery

- MEME
- STREME
- XSTREME
- MEME-ChIP
- GLAM2
- MoMo
- DREME (deprecated)

► Motif Enrichment

► Motif Scanning

► Motif Comparison

► Gene Regulation

► Utilities

► Manual

► Guides & Tutorials

► Sample Outputs

► File Format Reference

► Databases

► Download & Install

► Help

► Alternate Servers

▼ Authors & Citing

- Authors
- Citing the MEME Suite

▼ Recent Jobs

- Clear All

↳ Previous version 5.4.1

 **STREME**
Sensitive, Thorough, Rapid, Enriched Motif Elicitation

Version 5.5.0

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). See this [Manual](#) or this [Tutorial](#) for more information.

Data Submission Form

Perform discriminative motif discovery in sequence datasets (including in very **large** datasets). The sequences may be in the DNA, RNA or protein alphabet, or in a custom alphabet.

Select the type of control sequences to use

Shuffled input sequences User-provided sequences [?](#)

Select the sequence alphabet

Use sequences with a standard alphabet or specify a custom alphabet. [?](#)

DNA, RNA or Protein Custom [Choose File](#) No file chosen

Input the sequences

Enter the sequences in which you want to find motifs. [?](#)

[Upload sequences](#) [Choose File](#) No file chosen [?](#) 

Input job details

(Optional) Enter your email address. [?](#)

(Optional) Enter a job description. [?](#)

Advanced options

Note: if the combined form inputs exceed 80MB the job will be rejected.

[Start Search](#)

[Clear Input](#)

Version 5.5.0

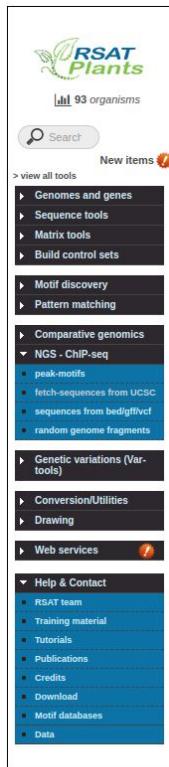
Please send comments and questions to: meme-suite@uw.edu

Powered by [Opal](#)

[Home](#) [Documentation](#) [Downloads](#) [Authors](#) [Citing](#)

RSAT peak-motif

- Web application and command line
- http://rsat.eead.csic.es/plants/peak-motifs_form.cgi



RSAT - peak-motifs

Discover exceptional motifs (over-represented, positionally biased) in a collection of ChIP-seq peaks.

References

- Thomas-Chollier, M., Hermann, C., Defrance, M., Sand, O., Thieffry, D. and van Helden, J. (2011). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets Nucleic Acids Research doi:10.1093/nar/gkr1104, 9. [Open access]
- Thomas-Chollier M, Darbo E, Hermann C, Defrance M, Thieffry D, van Helden J. (2012). A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nat Protoc 7(8): 1551-1568. [PMID 22836136]

► Information on the methods used in peak-motifs

Peak Sequences

Title (mandatory)

Peak sequences (mandatory) Paste your sequence (fasta format)
Or select a file to upload (.gz compressed files supported)
Choose File No file chosen
URL of a sequence file available on a Web server (e.g. Galaxy).

Optional: control dataset for differential analysis (test vs control)

Control sequences Paste your sequence (fasta format)
Or select a file to upload (.gz compressed files supported)
Choose File No file chosen
URL of a sequence file available on a Web server (e.g. Galaxy).

Mask

(I only have coordinates in a BED file, how to get sequences ?)

► Reduce peak sequences

► Motif discovery parameters

► Compare discovered motifs with databases (e.g. against Jaspar) or custom reference motifs

► Locate motifs and export predicted sites as custom UCSC tracks

► Reporting options

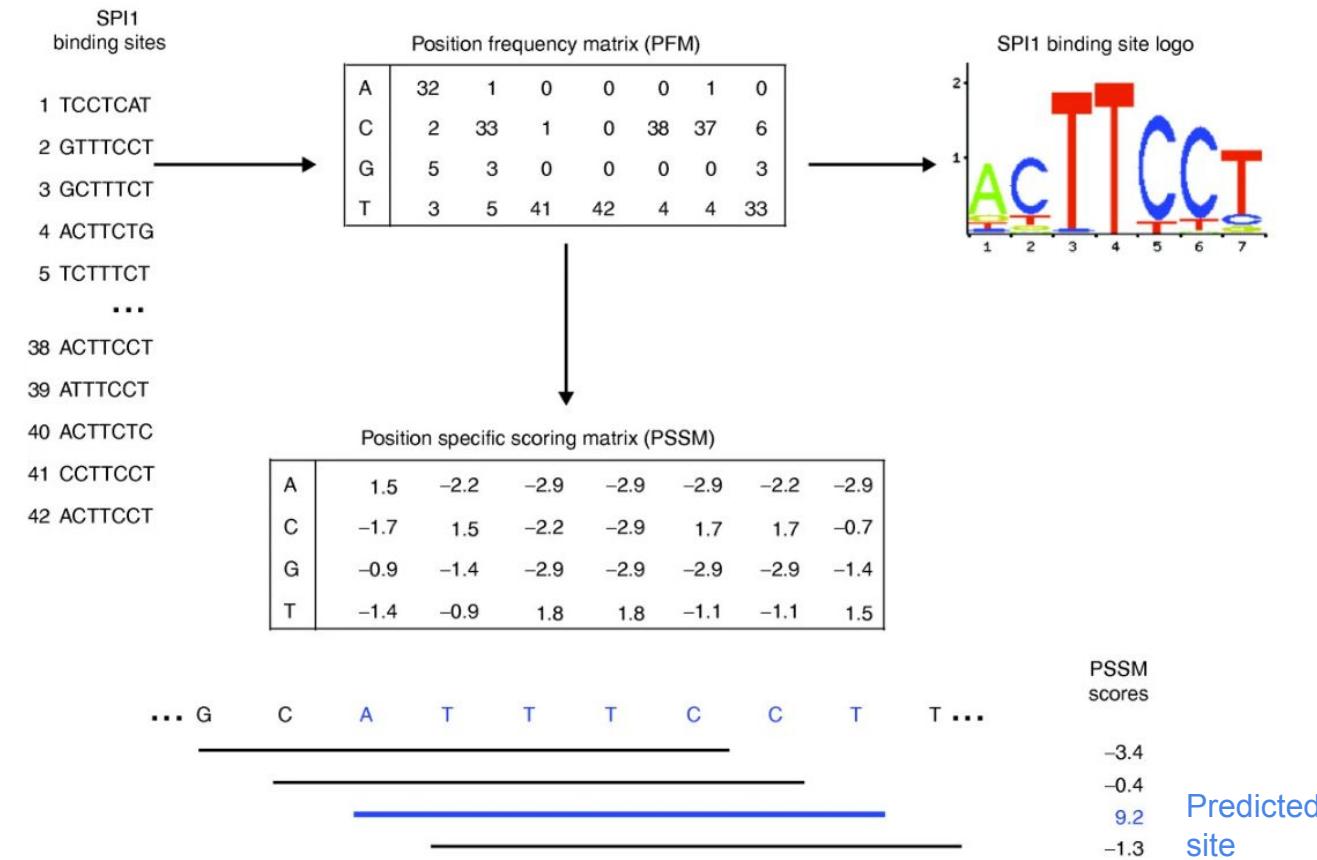
Output display email

Note: email output is preferred for very large datasets or many comparisons with motifs collections

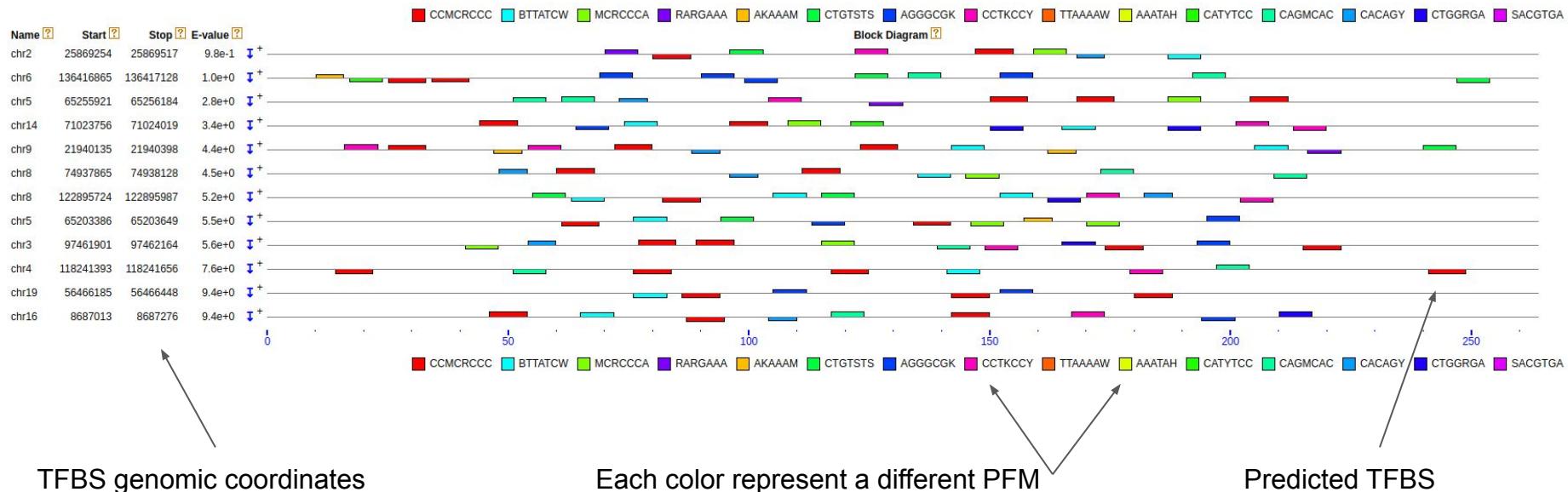
[\[MANUAL\]](#) [\[TUTORIAL\]](#)

Motif scanning

- Every sequence is evaluated.
- We select those positions with a score above a threshold.
- Repeat this process for each PFM.



Motif scanning

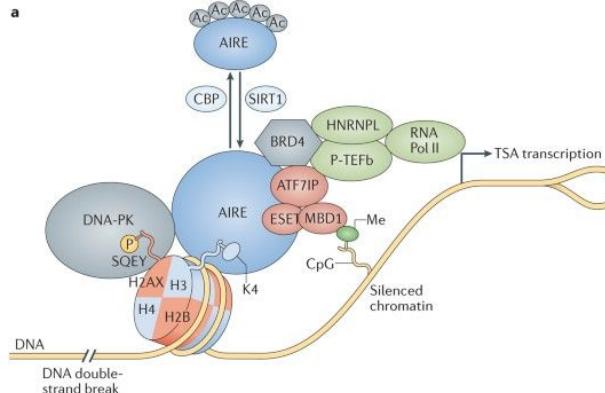


What to do when you do **not know** the motif of a ChIP'ed protein? Or rather when you want to find a regulatory role of a certain protein?

Let's say you are analyzing AIRE

⇒ you have ChIP-seq peaks where AIRE TF was targeted

What's next?



1. What are the motifs enriched in peaks?

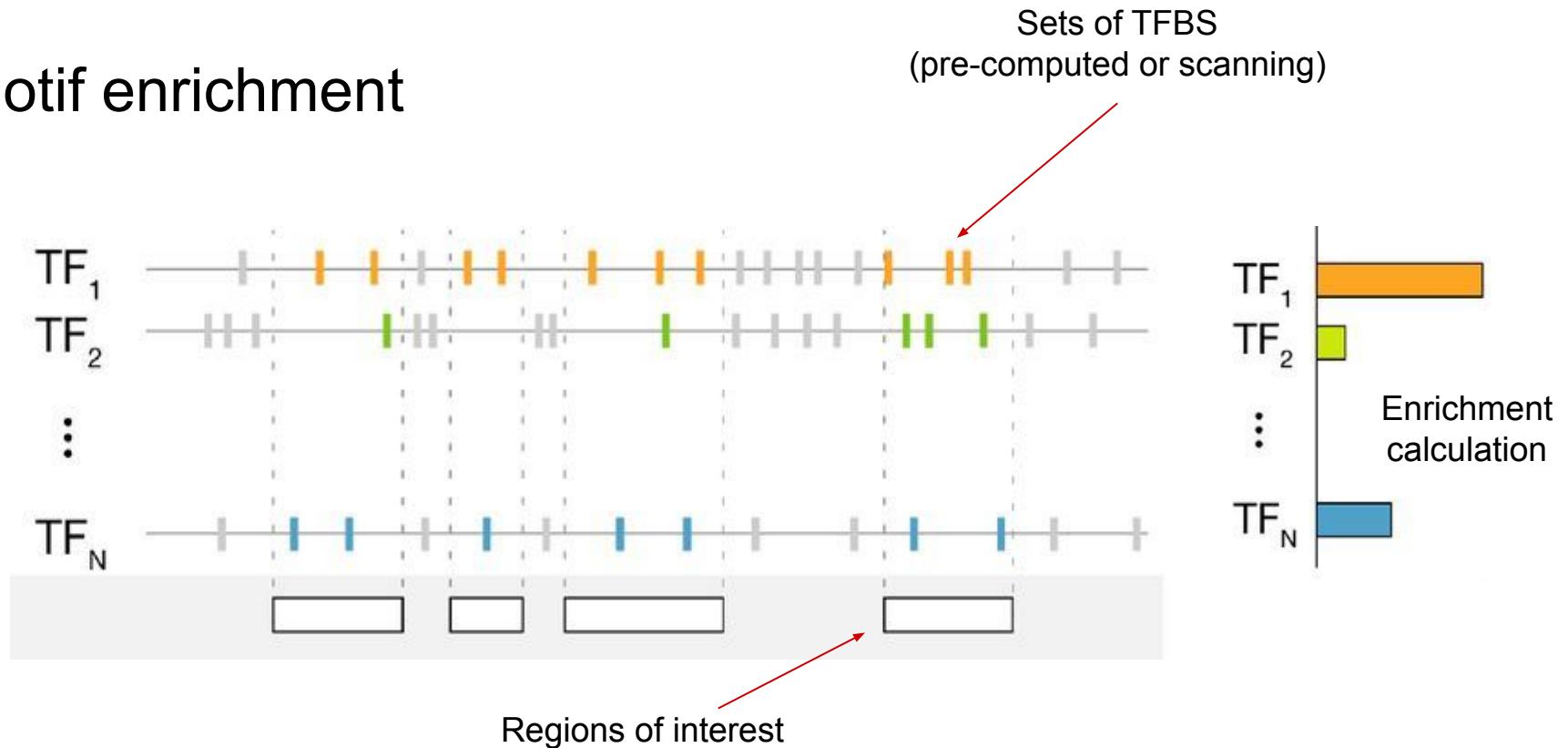
CAREFUL! ⇒ Should you take the top motif?

2. What do you know about AIRE binding?

3. Which genomic regions have the motifs of interest?

⇒ Motif scanning

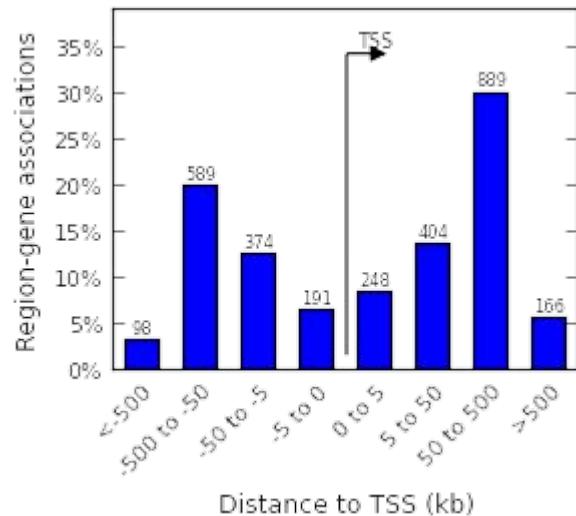
Motif enrichment



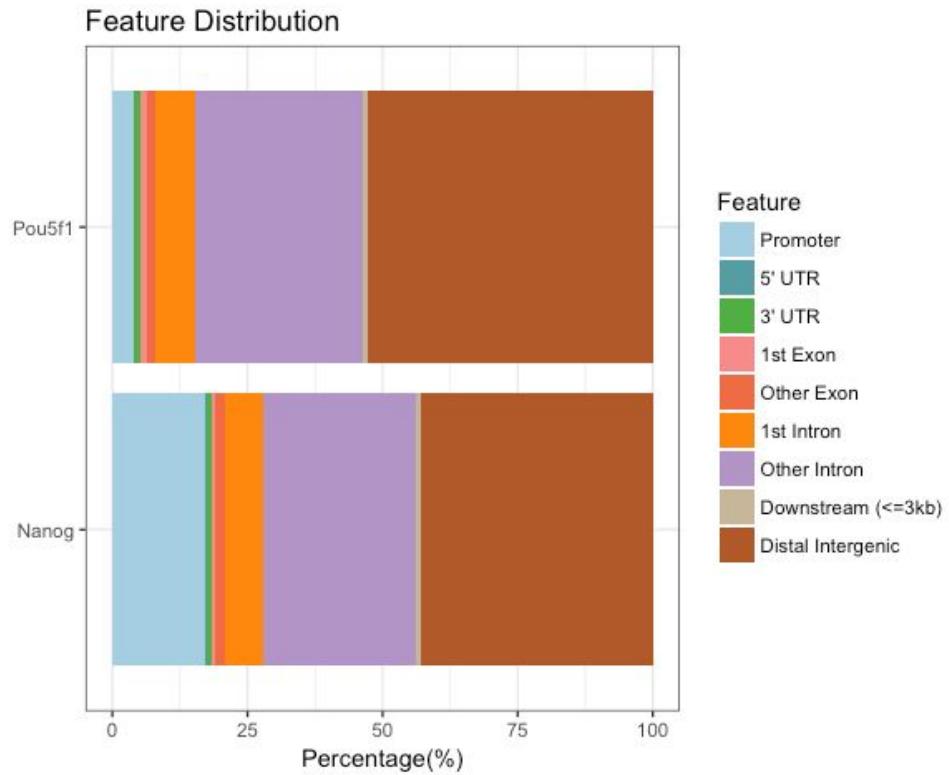
What could have gone wrong if you do not see the expected motif?

- Biology:
 - ChIP'ed TF does not bind DNA in sequence-specific manner.
 - TF binds through other proteins - **this might be interesting!**
- Technical reasons:
 - Your antibody was not specific enough or too specific.
 - Discovery algorithm is not able to detect the motif:
 - there is too much noise
 - inappropriate k-mer size

Peak annotation



Distance to TSS with GREAT



ChIPseeker R package

Functional enrichment analysis

GREAT Overview News Use GREAT Demo Video How to Cite Help Forum Bejerano Lab, Stanford University

GREAT version 4.0.4 current (08/19/2019 to now)

Job Description

Region-Gene Association Graphs

What do these graphs illustrate?

Number of associated genes per region
Download as PDF.

Genomic regions associated with one or more genes
Genomic regions not associated with any genes

Number of associated genes per region	Count
0	12
1	441
2	1259

Binned by orientation and distance to TSS
Download as PDF.

Region-gene associations
Distance to TSS (kb)

Distance to TSS (kb)	Count
<500	46
500 to 500	589
50 to 500	374
5 to 500	391
0 to 5	248
5 to 50	405
50 to 500	889
>500	165

Binned by absolute distance to TSS
Download as PDF.

Region-gene associations
Absolute distance to TSS (kb)

Absolute distance to TSS (kb)	Count
0 to 5	43
5 to 50	778
50 to 500	1478
>500	264

Global Controls Global Export Which data is exported by each option?

Ensembl Genes (no terms) Global controls

GO Biological Process (20+ terms) Global controls

Table controls: Export Shown top rows in this table: [20] Set Term annotation count: Min: [1] Max: [Inf] Set Visualize this table: [select one]

Term Name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
chemical synaptic transmission	1	9.7715e-20	1.2845e-15	2.0606	183	10.69%	2	1.5578e-22	2.7200	123	400	5.87%
trans-synaptic signaling	2	2.2912e-19	1.5059e-15	2.0395	185	10.81%	3	1.2894e-22	2.7016	124	406	5.91%
neurotransmitter transport	5	4.8925e-12	1.2862e-8	2.5482	70	4.09%	68	6.8270e-8	2.7364	43	139	2.05%
signal release	8	1.1633e-11	1.9114e-8	2.3440	79	4.61%	62	2.8706e-8	2.6110	49	166	2.34%

GREAT (web application and R package)

Enrichr Login | Register 53,477,000 sets analyzed 427,081 terms 203 libraries

Analyze What's new? Libraries Gene search Term search About Help

Input data

Expand a gene, a term, or a variant into a gene set:
e.g. STAT3, breast cancer, or rs28897756

Try an example STAT3, breast cancer, rs28897756

Include the top 100 most relevant genes

Paste a set of valid Entrez gene symbols on each row in the text-box below. Try a gene set example.

Paste a set of valid Entrez gene symbols (e.g. STAT3) on each row in the text-box

0 gene(s) entered

In order to enable others to search your set please enter a brief description of it.

Contribute your set so it can be searched by others

Submit

Please acknowledge Enrichr in your publications by citing the following references:
Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A.
Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013; 12(14).

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gunderson GW, Ma'ayan A.
Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*. 2016; gkw377.

Xie Z, Bailey A, Kuleshov MV, Clarke DJ., Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, Jeon M, & Ma'ayan A.
Gene set knowledge discovery with Enrichr. *Current Protocols*, 1, e90. 2021. doi: 10.1002/cpzi.90

modEnrichr A suite of gene set enrichment analysis tools

FlyEnrichr YeastEnrichr WormEnrichr FishEnrichr

Click here to raise an issue on GitHub

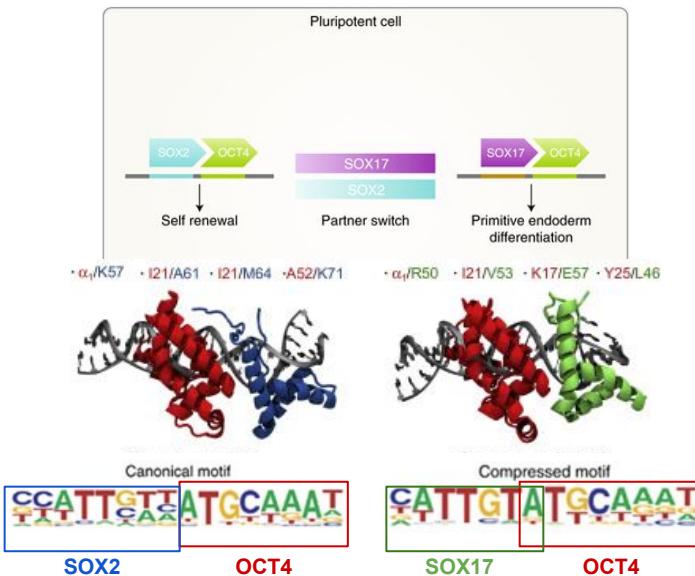
Enrichr (web application, API, R package)

So far we looked into individual TFs, but how about cooperativity?

- Nowadays deep learning techniques are powerful way to detect TF binding syntax across genome
 - ⇒ computationally challenging, might require a lot of data
- Simpler methods use different ChIP-seq datasets to infer possible co-binding
 - ⇒ a lot of combinations to go through!
- We can look into co-binding through DNA patterns

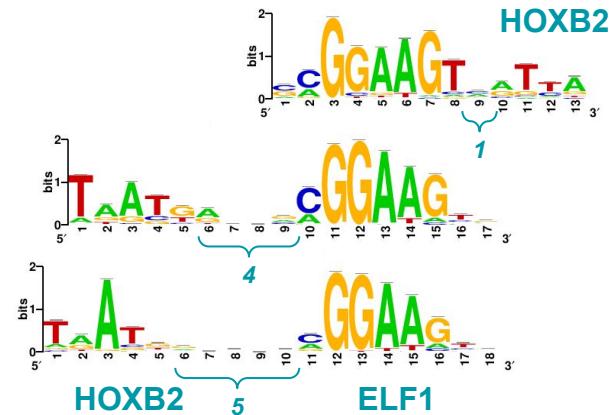
Cooperativity makes regulation very complex

TF cooperativity gives rise to TF binding combinations



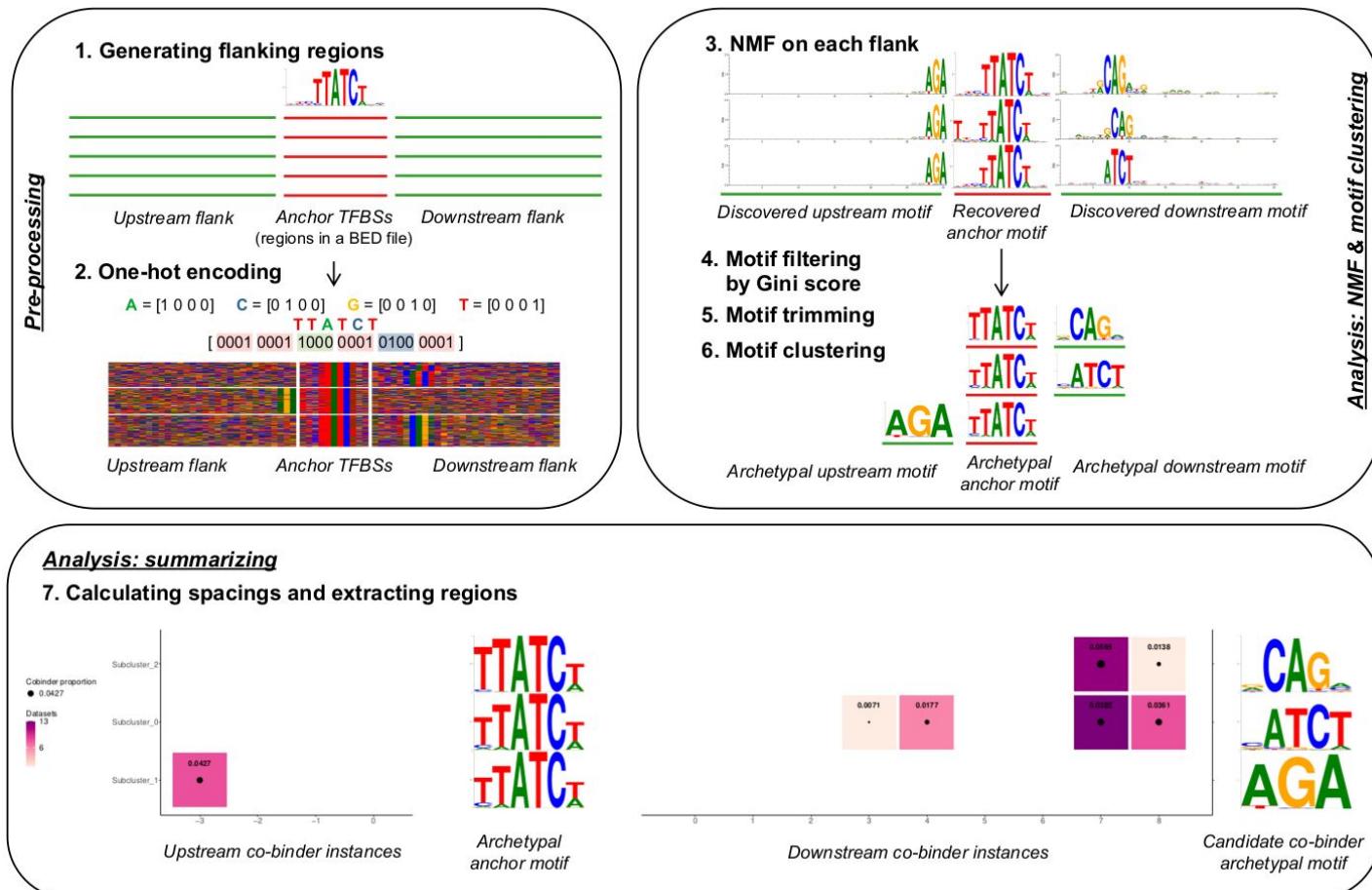
Adapted from Aksoy et al., 2013
and Li & Belmonte, 2018

Different spacings and orientations of HOXB2 and ELF1¹

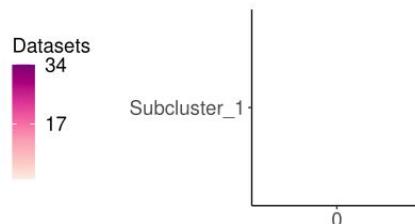


Adapted from Jolma et al., 2016

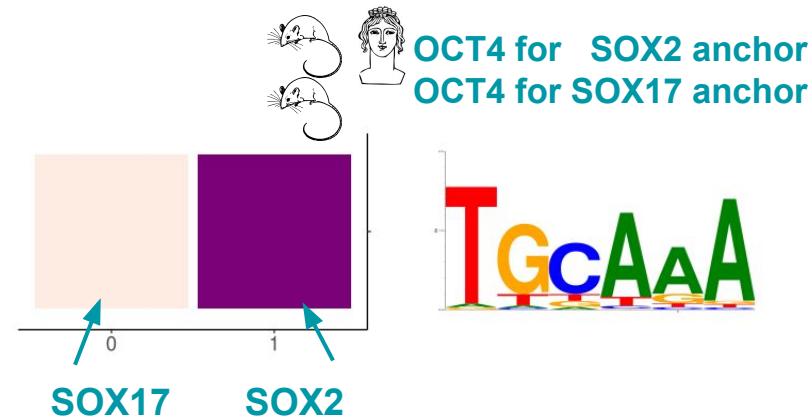
COBIND



Same co-binders are found within TF families: SOX-related factors



Anchors of different family members



Article | 8 March 2013 | FREE ACCESS

Oct4 switches partnering from Sox2 to Sox17 to reinterpret the enhancer code and specify endoderm

Irene Aksoy, Ralf Jauch, Jiaxuan Chen, Mateusz Dyla, Ushashree Divakar, Gireesh K Bogu, Roy Teo, Calista Keow Leng Ng, Wishva Herath, Sun Lili, Andrew P Hutchins, Paul Robson, Prasanna R Kolatkar, Lawrence W Stanton

Author Information

EMBO J (2013) 32: 938-953 | <https://doi.org/10.1038/emboj.2013.31>



Summary

- You can start the analysis from raw or processed data. If you start from processed - make sure you understand how it was processed.
- It is important to understand the experimental setup.
- There are some standard steps that could help guide you into more specific analysis.
 - Visualize your data in the genome browser - this helps you understand the quality.
 - Do *de novo* motif discovery even if you know what motif should be in your peaks.

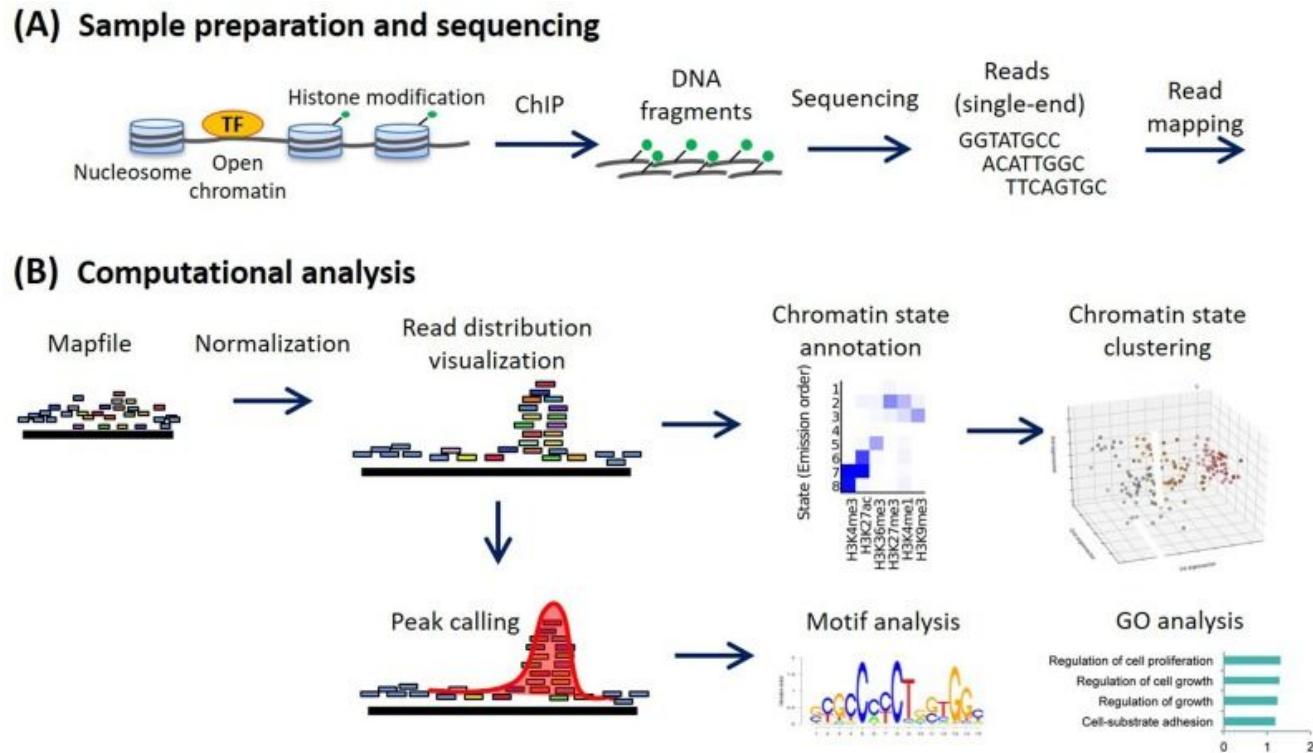


PART 5

TF binding data resources

What data can we get?

- Raw
- Pre-processed
- Processed



Where to get publicly available data?

- Individual studies that are relevant to your research (PUBMED, GEO).
- Consortiums:
 - TCGA
 - GTRD
- Individual databases:
 - ReMap
 - UniBind
 - Factorbook

GEO database

 NCBI  Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO | NCBi > GEO > Accession Display | Reviewer access | Sign Out

Scope: Self Format: HTML Amount: Quick GEO accession: GSE211852 | GO

Series GSE211852 Query DataSets for GSE211852

Status Public on Oct 05, 2022

Title Transcriptional constraint of EWS/FLI by an ETS transcription factor promotes Ewing sarcoma growth [CUT&RUN]

Organism *Homo sapiens*

Experiment type Genome binding/occupancy profiling by high throughput sequencing

Summary Pediatric cancers frequently harbor sentinel mutations involving transcription factors (TFs) that dysregulate normal development. A recurrent mechanism involves the ability of mutant TFs to co-opt cell lineage-specific, activating TFs to promote cancer growth. Ewing sarcoma, the second most common pediatric bone cancer, is defined by the presence of a 11;22 chromosomal translocation fusing the N-terminal of the EWS protein with the C-terminal DNA binding domain of an ETS (E26 Transformation Specific) family member, most commonly (85-90% of cases), FLI1. FLI1 gain-of-function exhibits the neoplastic ability to pioneer de novo enhancers at repeating 5'-GGAA-3' motifs in the cell-of-origin, which has not been identified. To date, efforts to elucidate the key mechanisms by which EWS/FLI promotes oncogenesis have prioritized identifying the genes that are profoundly activated by EWS/FLI and highly expressed in Ewing sarcoma compared to other cancers, with particular focus on transcription factors capable of altering cell state. However, it is not known whether, globally, these genes constitute the most critical drivers of Ewing sarcoma cell growth. Here, we describe the results of an unbiased deletion screen revealing that the wild-type repressive ETS family TF, ETV6 (ETS Variant 6, or TEL), is a novel and most critical TF dependency specific to Ewing sarcoma. We demonstrate that the repressive activity of ETV6 constrains EWS/FLI gene activation at GGAA repeat enhancers to promote Ewing sarcoma cell growth.

Overall design CUT&RUN (Skene and Henikoff, elife, 2017) was used to evaluate FLI1 and H3K4me3 binding in Ewing sarcoma using the newly derived CCLF-PEDS_0009_T (PEDS0009) cell line (Guenther et al., Clinical Cancer Research, 2010) in cells transduced with lentivirally-packaged CRISPR/Cas9 constructs. All sgRNA sequences used in the Broad Institute AVANA CRISPR/Cas9 screen are available for download at the DepMap Portal (<https://depmap.org>). sgRNA sequences used to target ETV6 were taken from this screen. The sgETV6-1 forward sequence was 5'-GCAGCCAATTACTGGAGCC-3'. The sgETV6-2 forward sequence was 5'-GCAGGGATGACGTAGCCCCAG-3'. The sgETV6-3 forward sequence was 5'-GTGTGTTGATAGTTTCCA-3'. The sgETV6-4 forward sequence was 5'-GTTATGGTCACATTATCCA-3'. For control sgRNA, sgChr2.2 was used as a cutting control and targets a gene desert on Chromosome 2, 5'-GGTGTGCCTATGAAGCAGTG-3'. For ligation into the LentCRISPRv2 plasmid, additional bases were added: 5'-CACCG-3' was added to the beginning of the forward sequence, 5'-AAC-3' and 5'-C-3' were added at the beginning and end of the reverse sequence, respectively. Paired-end experiments were used to evaluate FLI1 and H3K4me3 binding. All CUT&RUN samples have matching IgG reference sample from the same batch and treatment condition. The samples were processed with a pipeline based upon the bulk-level method outlined in CUT&RUNTools 2.0 (Yu, F., Sankaran, V.G., Yuan, G. CUT&RUN Tools 2.0: a pipeline for single-cell and bulk-level CUT&RUN and CUT&Tag analysis. *Bioinformatics*, 38(1), 2022, 252-254. <https://github.com/f-yu/CUT-RUNTools-2.0/blob/master/docs/INSTALL.md>).

Platforms (1)

GPL24676 Illumina NovaSeq 6000 (Homo sapiens)

Samples (6)

[Less...](#)

GSM6503406 PEDS0009, sgChr2, FLI1

GSM6503407 PEDS0009, sgChr2, IgG

GSM6503408 PEDS0009, sgETV6, FLI1

GSM6503409 PEDS0009, sgETV6, IgG

GSM6503410 PEDS0009, sgChr2, H3K4me3

GSM6503411 PEDS0009, sgETV6, H3K4me3

This SubSeries is part of SuperSeries:

GSE181554 Transcriptional constraint of EWS/FLI by an ETS transcription factor promotes Ewing sarcoma growth.

Relations

BioProject PRJNA872360

Download family

SOFT formatted family file(s)

MINIML formatted family file(s)

Series Matrix File(s)

Format

SOFT [?](#)MINIML [?](#)TXT [?](#)

Supplementary file	Size	Download	File type/resource
GSE211852_RAW.tar	1.3 Gb	(http)(custom)	TAR (of BW, NARROWPEAK)
SRA Run Selector ?			

Raw data are available in SRA

Processed data provided as supplementary file

GTRD: gene transcription regulation database

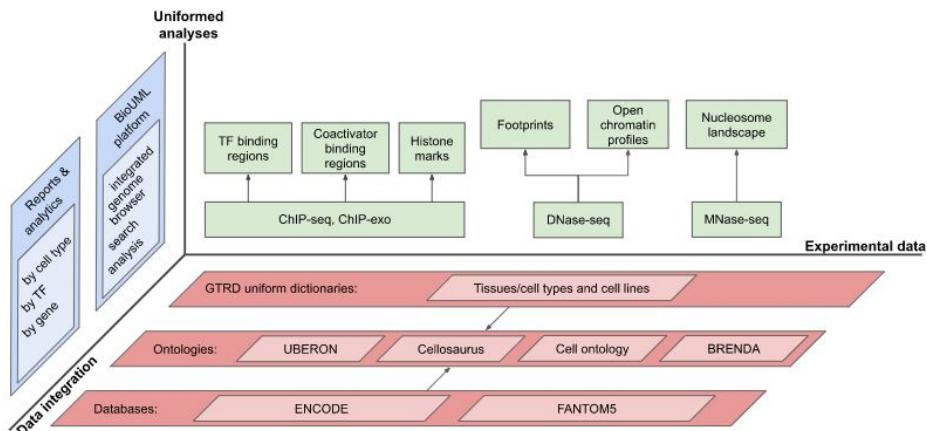
GTRD
Gene Transcription Regulation Database
v21.12

The most complete collection of uniformly processed ChIP-seq data on identification of transcription factor binding sites for human and mouse. Convenient web interface with advanced search, browsing and genome browser based on the BioUML platform. For support or any questions contact gtrd@biosoft.ru

[Start »](#) [Documentation »](#) [Download »](#) [Previous release »](#)

Workflow

How was it constructed?



Uniformly processed data

ReMap

- All peaks
- Non-redundant peaks
- Cis-Regulatory modules

- Different species

Manually curated ChIP-seq
and ChIP-exo data

ReMap2022

Home Search Genome Tracks Annotate REST API Download About Citations Change logs Contact

Search ReMap2022 database ...
Examples: FOXA1, MCF-7, Limb, third-instar, abd-A, ENCSR440VKE, GSE114266, Col-0_seedling, WRKY33 Advanced search

Species

Homo sapiens (Human) Mus musculus (Mouse) Drosophila melanogaster (Drosophila) Arabidopsis thaliana (Arabidopsis)

All Peaks (Identify all ChIP-seq peaks) Non redundant peaks (Merge similar targets) CRMs (Identify Cis Regulatory Modules)

Take the tour

ReMap is a large scale integrative analysis of DNA-binding experiments for *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Arabidopsis thaliana* transcriptional regulators. The catalogues are the results of the manual curation of ChIP-seq, ChIP-exo, DAP-seq from public sources (GEO, ENCODE, ENA).

ReMap database growth

Human Mouse Drosophila Arabidopsis

2015 2018 2020 2022

Key statistics

Homo sapiens: 1,210 Transcriptional regulators | Search for specific factors

Mus Musculus: 737 Cell lines and tissues | Search for specific cells

Drosophila melanogaster: 8,103 QC ChIP-seq datasets | Browse a given dataset

Arabidopsis thaliana: 182 million Binding regions | Download our data

Papers using ReMap Citation page

Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. Ibarra IL, Hollmann NM, Klaus B, Augsten S, Velten B, Hennig J, Zaug J. Nature communications 2020 Jan 8;11(1):124. 10.1038/s41467-019-13888-7

Citing ReMap

PubMed | NAR | PDF

ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Fayrouz Hammal, Pierre De Langen, Aurélie Bergon, Fabrice Lopez, Benoit Ballester. Nucleic Acids Research, Volume 50, Issue D1, 1 January 2022

About

Contact us ReMap 2022 (current) ReMap 2020 ReMap 2018 ReMap 2015 Change logs

Inserm TAGC

Aix-Marseille université MESO CENTRE

urc U1090

UniBind: database of direct TFBSS

- Processing of peaks in a specific way that addresses TF binding specificity

Processed data



Rafael Riudavets Puig

Search UniBind database...
Examples: LFY ESC CTCF MA0590.1 [Search](#) [Advanced Options](#)

[Browse UniBind by species](#)

Arabidopsis thaliana	Caenorhabditis elegans	Danio rerio
Drosophila melanogaster	Homo sapiens	Mus musculus
Rattus norvegicus	Saccharomyces cerevisiae	Schizosaccharomyces pombe

9654 ChIP-seq datasets

9 Species

1316 Cell lines & Tissues

841 Transcription Factors

transcription factors
A map of direct TF-DNA interactions across species
Read more about UniBind

UniBind is a comprehensive map of direct interactions between transcription factor (TFs) and DNA. High confidence TF binding site predictions were obtained from uniform processing of thousands of ChIP-seq data sets using the ChIP-eat software.

Called peaks

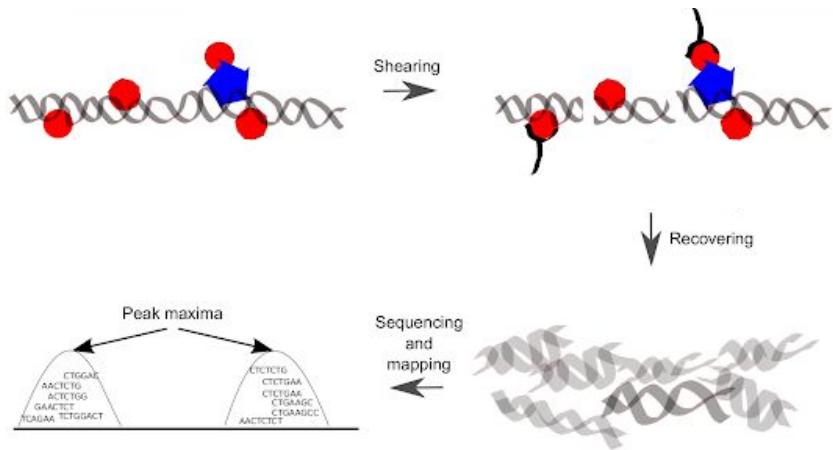
Genomic regions → DAMO-optimized JASPAR PWM → Automatic thresholding → Direct TF - DNA interactions

Citing UniBind [PubMed](#) [Journal](#) [PDF](#)

R. Riudavets Puig, P. Boddie, A. Khan, J.A. Castro Mondragon, A. Mathelier, **UniBind: maps of high-confidence direct TF-DNA interactions across nine species**. *BMC Genomics* **22**, 482 (2021). <https://doi.org/10.1186/s12864-021-07760-6>.

ChIP-seq

- Direct (red) vs indirect (blue + red) interactions.
- High quality direct interactions should:
 - Be more likely to be in the vicinity of the peak summit.
 - Contain the sequence recognised by the TF.
- The **ChIP-eat** pipeline was developed to obtain TFBSs corresponding to high quality direct TF - DNA interactions.

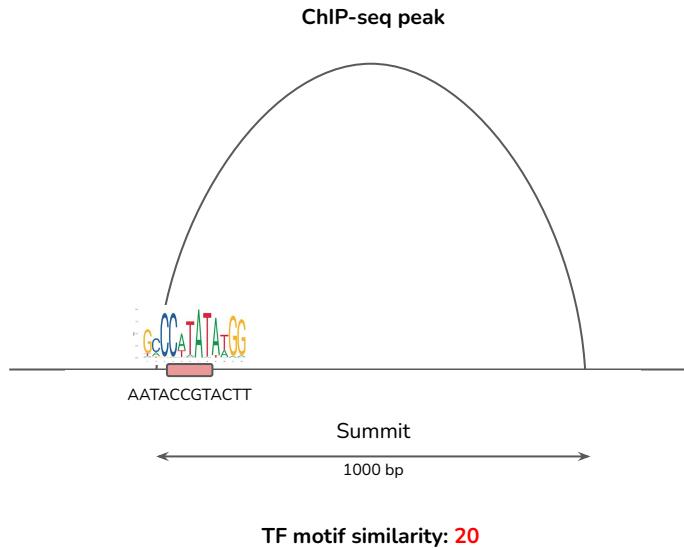


Modified from Mathelier, Shi, and Wasserman (2015)

ChIP-eat

Input: ChIP-seq called peaks (e.g. MACS2).

- 1) Scan 1000 bp windows centered at its summit with optimized TF JASPAR matrices.
- 2) Get best scoring subsequence and record its distance to the peak summit.
- 3) Entropy-based algorithm to determine a threshold in a parameter-free way.

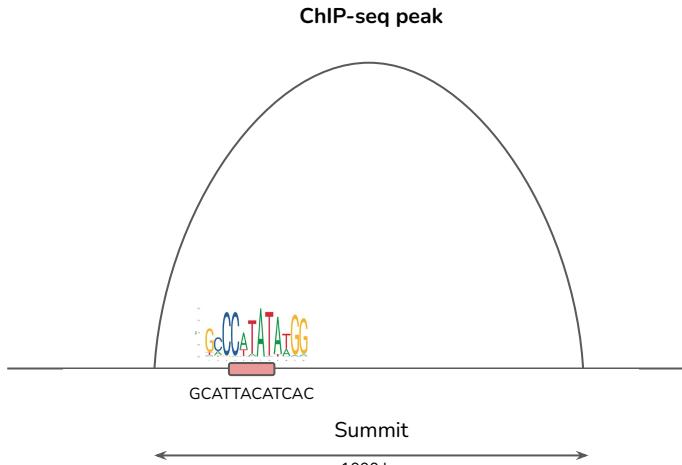


<https://bitbucket.org/CBGR/chip-eat/>

ChIP-eat

Input: ChIP-seq called peaks (e.g. MACS2).

- 1) Scan 1000 bp windows centered at its summit with optimized TF JASPAR matrices.
- 2) Get best scoring subsequence and record its distance to the peak summit.
- 3) Entropy-based algorithm to determine a threshold in a parameter-free way.



TF motif similarity: **27**

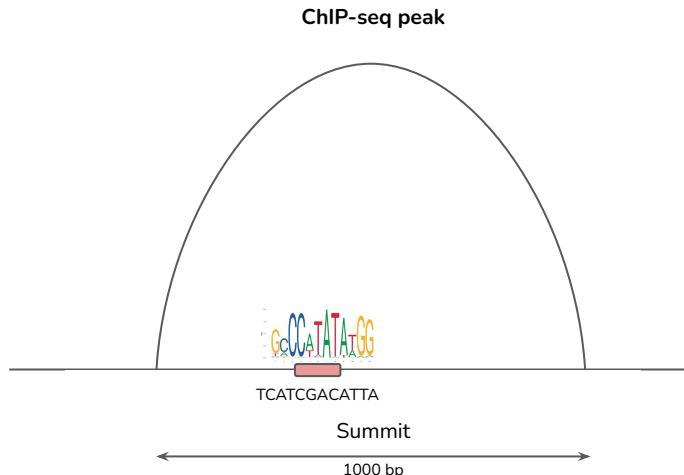


<https://bitbucket.org/CBGR/chip-eat/>

ChIP-eat

Input: ChIP-seq called peaks (e.g. MACS2).

- 1) Scan 1000 bp windows centered at its summit with optimized TF JASPAR matrices.
- 2) Get best scoring subsequence and record its distance to the peak summit.
- 3) Entropy-based algorithm to determine a threshold in a parameter-free way.

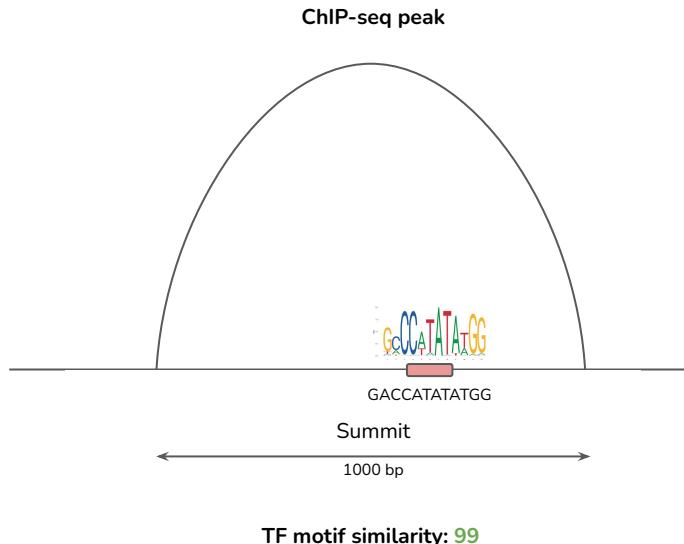


<https://bitbucket.org/CBGR/chip-eat/>

ChIP-eat

Input: ChIP-seq called peaks (e.g. MACS2).

- 1) Scan 1000 bp windows centered at its summit with optimized TF JASPAR matrices.
- 2) Get best scoring subsequence and record its distance to the peak summit.
- 3) Entropy-based algorithm to determine a threshold in a parameter-free way.

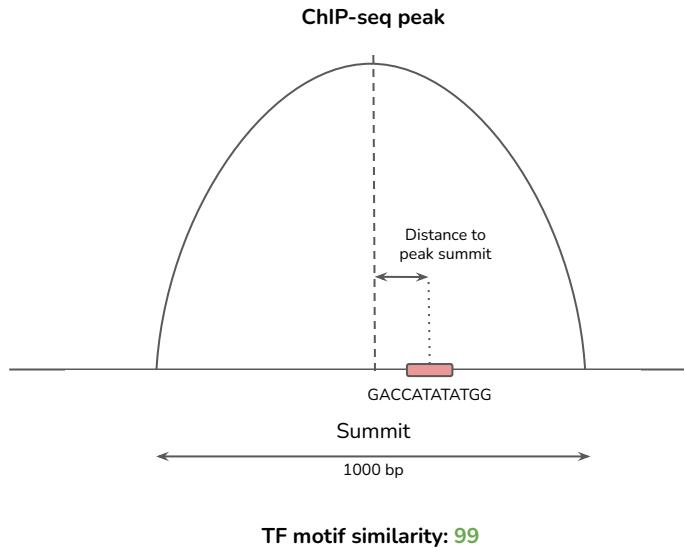


<https://bitbucket.org/CBGR/chip-eat/>

ChIP-eat

Input: ChIP-seq called peaks (e.g. MACS2).

- 1) Scan 1000 bp windows centered at its summit with optimized TF JASPAR matrices.
- 2) Get best scoring subsequence and record its distance to the peak summit.
- 3) Entropy-based algorithm to determine a threshold in a parameter-free way.

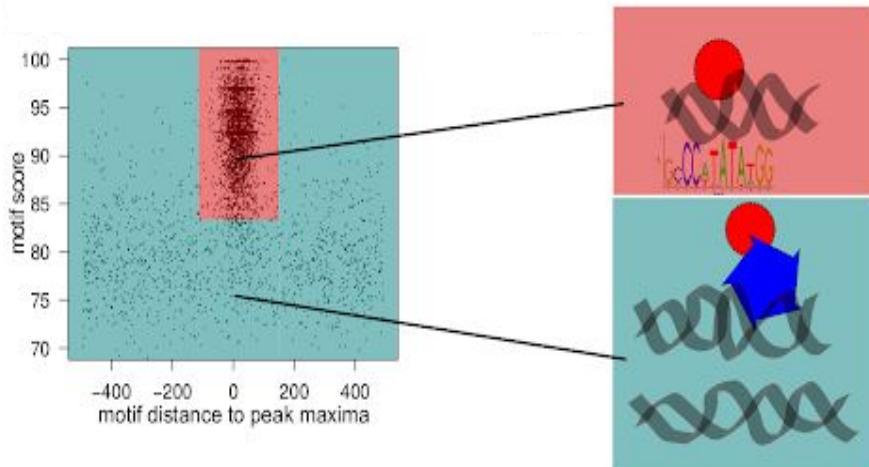


<https://bitbucket.org/CBGR/chip-eat/>

ChIP-eat

Input: ChIP-seq called peaks (e.g. MACS2).

- 1) Scan 1000 bp windows centered at its summit with optimized TF JASPAR matrices.
- 2) Get best scoring subsequence and record its distance to the peak summit.
- 3) Entropy-based algorithm to determine a threshold in a parameter-free way.



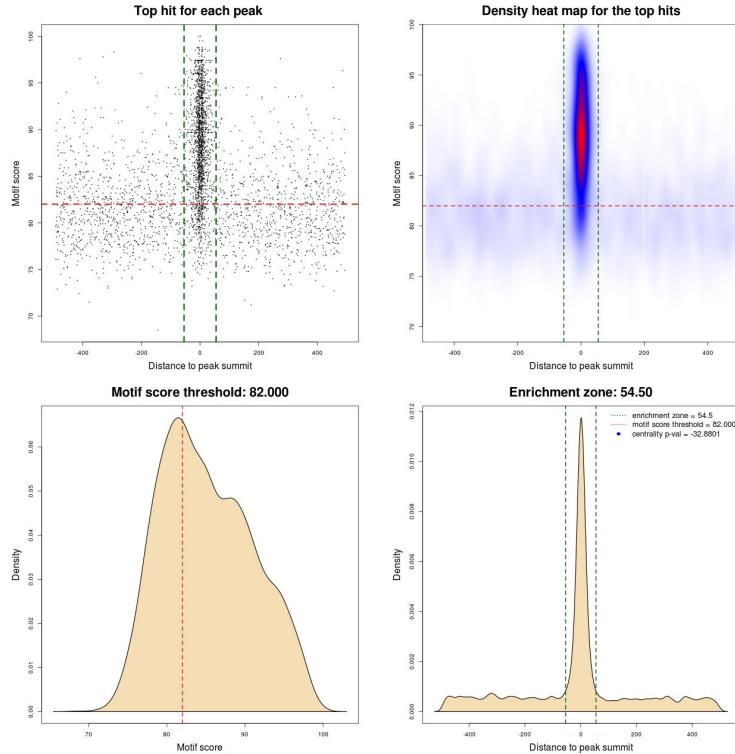
Modified from Mathelier, Shi, and Wasserman (2015)



<https://bitbucket.org/CBGR/chip-eat/>

Parameter-free thresholding

- Automatically separate signal (high quality direct TFBSs) from noise (indirect TFBSs, unspecific binding, etc).
- Similar problem as separating noise from signal in a black and white image.



Summary

- Explore already available data - there is a huge number of resources.
- Again: if using processed data - know how it was processed.
- Integrate different data from various data sources.
- Your research can be discovering biology or processing data.



PART 6

Exercise



Data

- Clone the repository:

```
git clone https://github.com/ievarau/intro\_to\_regulation\_workshop.git
```

- Use .fasta files in fasta_peak_sets directory

Objective

Using online tools investigate DNA binding of
GATA2, GATA4 and NR3C1 (GR) transcription factors.



Please take a few minutes to give some feedback:

<https://forms.gle/XpRV4oDTRPoUCnzM8>

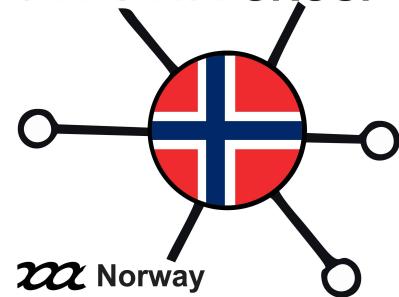
The End



NCMM

NORDIC
COMPUTATIONAL
BIOLOGY

*isCB REGIONAL
Student GROUP*



rsg-norway.iscbsc.org



twitter.com/RSGNorway



rsg-norway@iscbsc.org

Get Involved: bit.ly/370yAU0