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

Ieva Rauluseviciute & Jaime A Castro-Mondragon

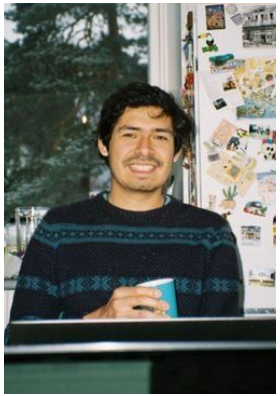
NCMM
Centre for Molecular Medicine Norway

UiO : Universitetet i Oslo

Who are we?

Ieva Rauluševičiūtė

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



Jaime A. Castro Mondragon

Bioinformatician at the department of In Silico analysis at Nykode Therapeutics

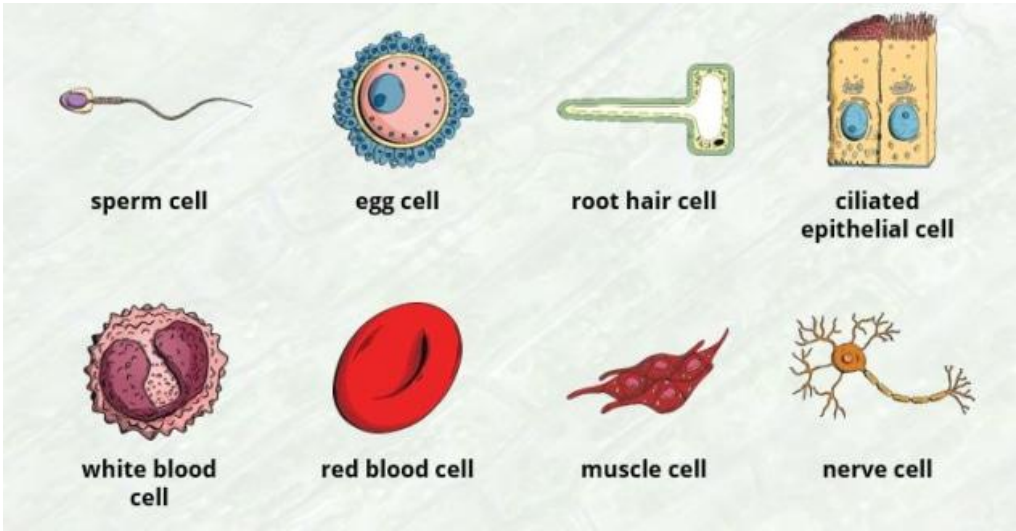
Who are you?

PART 1

Gene regulation by TF binding

Why do we have multiple cell types?

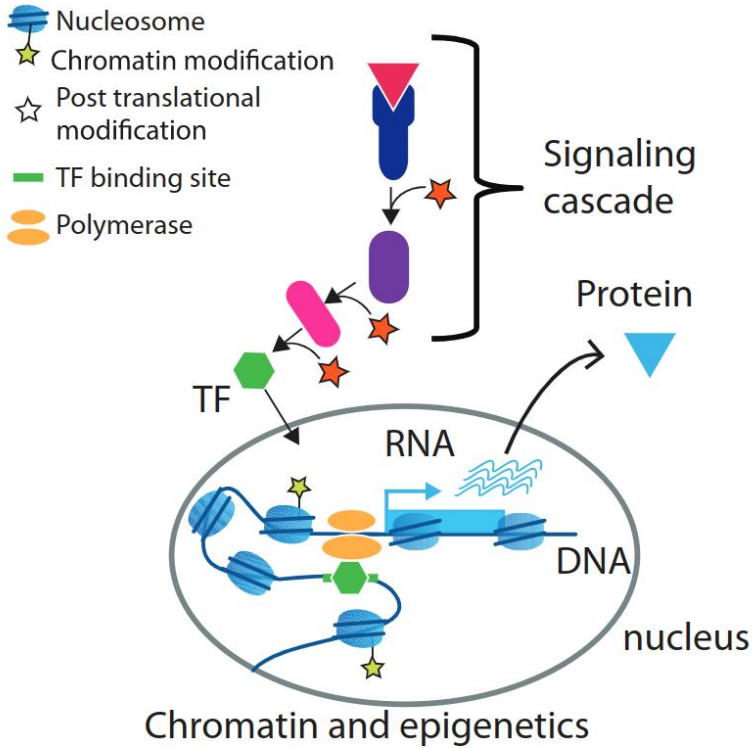
- All cells in our bodies have the same DNA sequence but …
- Why do we have hundreds of cell types with multiple morphology and functions ?
- Something ‘beyond’ the DNA itself must encode information of when, how and how much a gene should be correctly expressed.



https://www.twinkl.no/teaching-wiki/cells-in-the-human-body

TFs link cell signaling with gene regulation

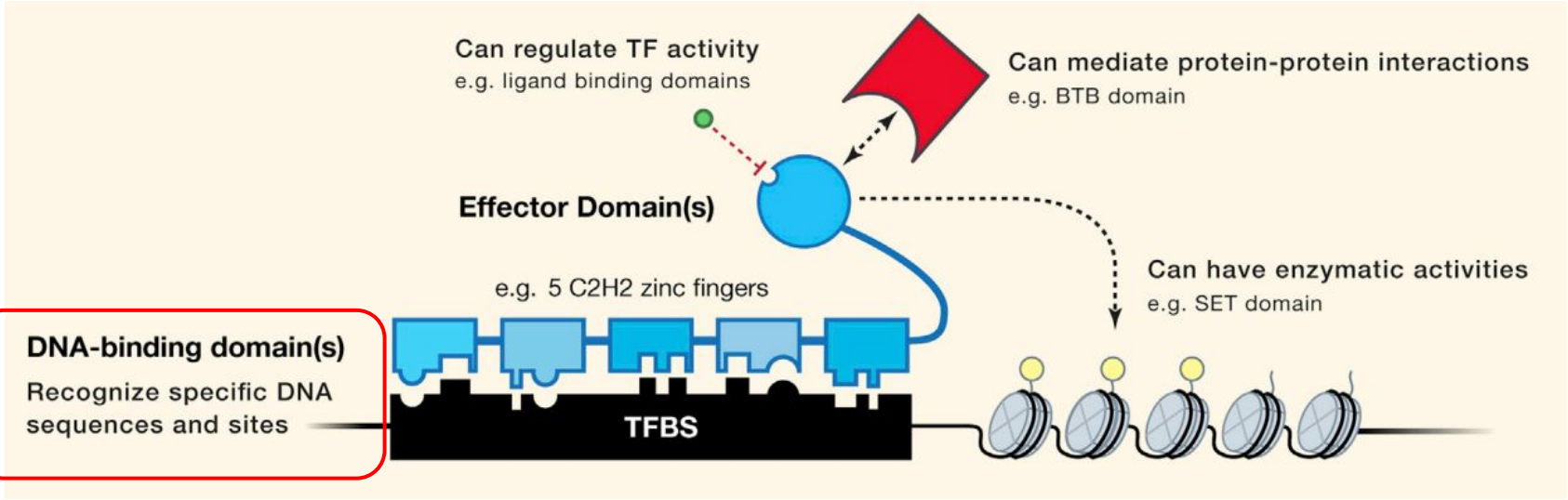
- Transcription Factors (TFs)
- Regulate key processes such as development and differentiation
- TFs activate or repress transcription.
- Combinatorics of TFs give rise to different cell types.



Weidemüller et al. 2021. Transcription factors: Bridge between cell signaling and gene regulation.

DNA-binding Transcription Factors

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



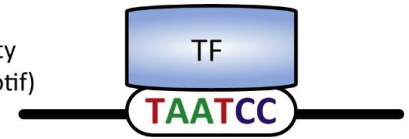
Lambert SA et al. 2018. The Human Transcription Factors

One binding mode to rule them all?

What we want to see

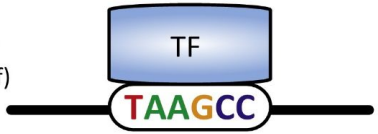
Consensus

High affinity
(strong motif)

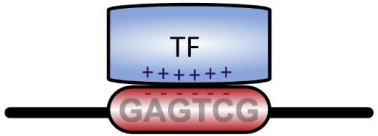


What we actually see

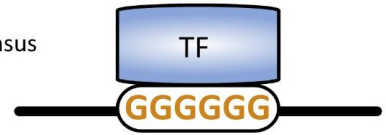
Low affinity
(weak motif)



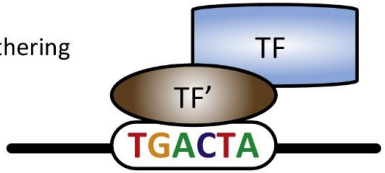
Nonspecific



Nonconsensus

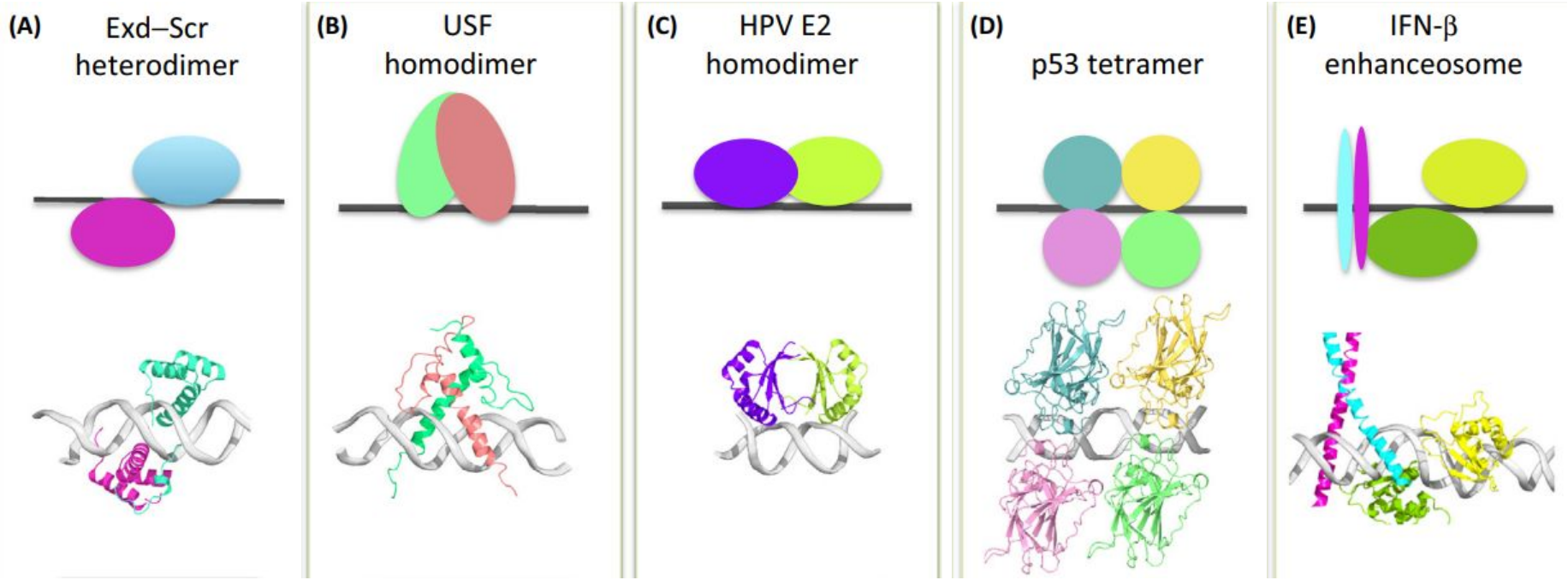


Indirect/Tethering



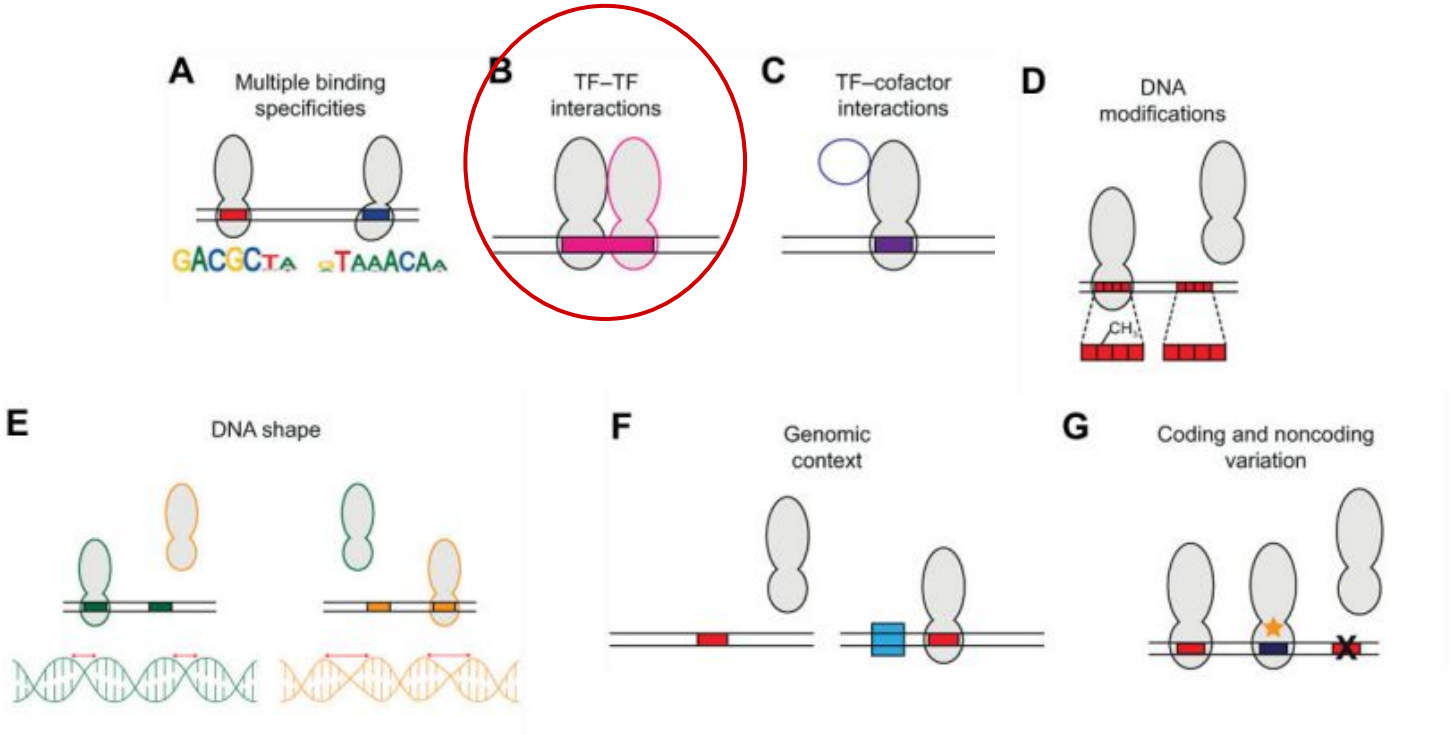
Adapted from Slattery et al., 2014. Absence of a simple code: how transcription factors read the genome.

TFs like to interact with other TFs



Adapted from Slattery et al., 2014. Absence of a simple code: how transcription factors read the genome.

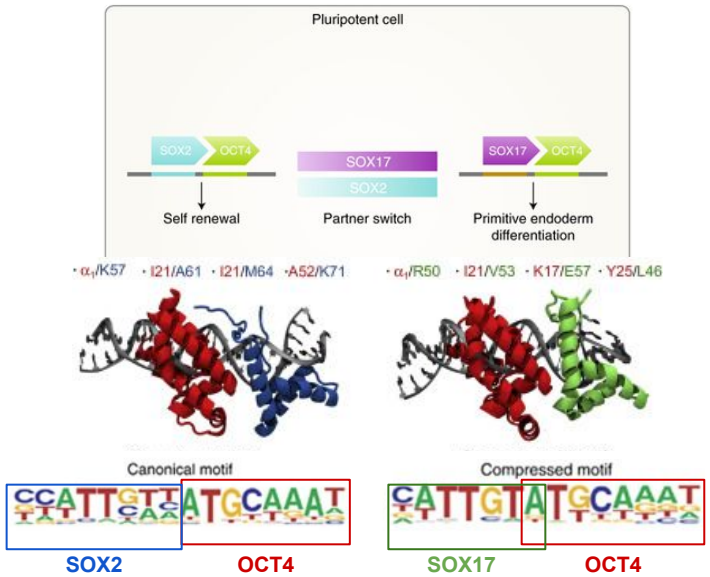
TF binding is also affected by other features



Adapted from Inukai et al., 2017. Transcription factor–DNA binding: beyond binding site motifs.

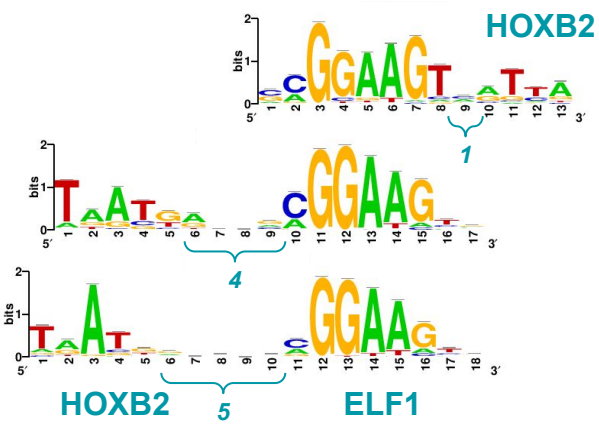
Cooperativity makes regulation very complex

TF **cooperativity** gives rise to TF binding **combinations**



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

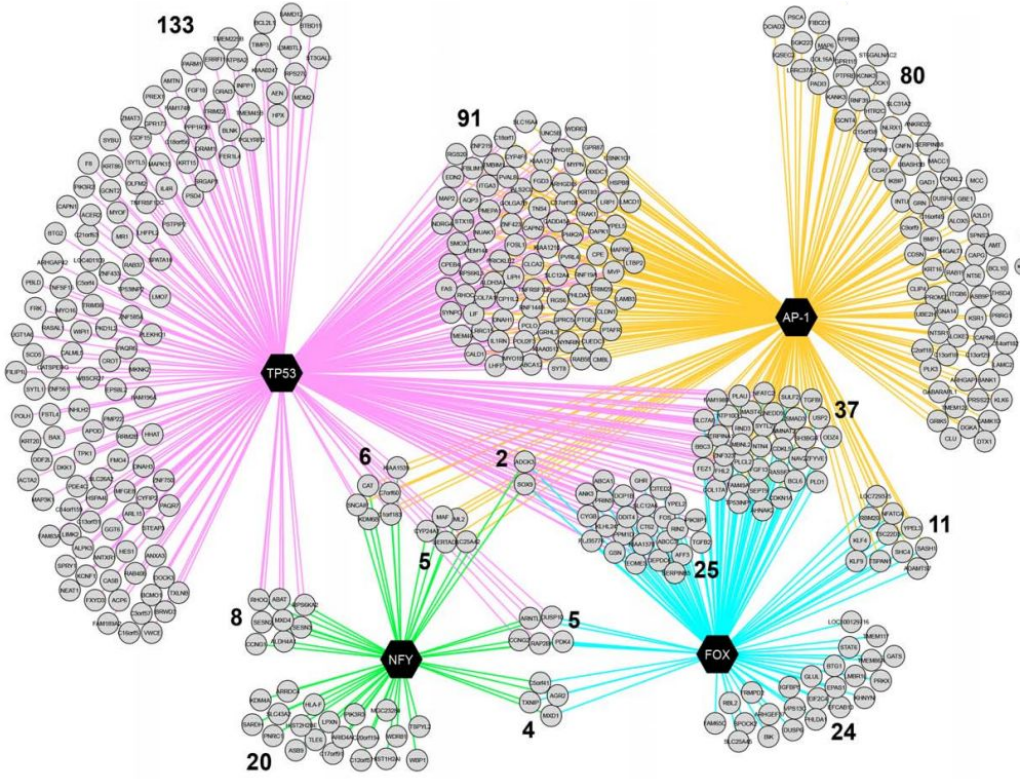
Different spacings and orientations of HOXB2 and ELF1 [1]



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.



Janky R et al 2014.

Summary

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

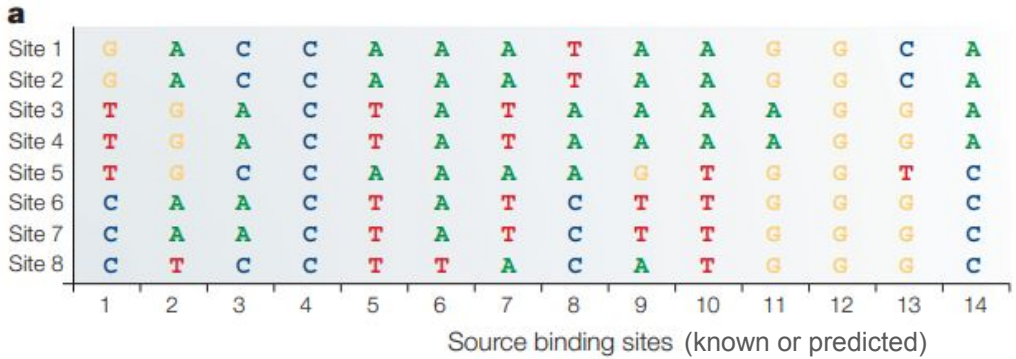
PART 2

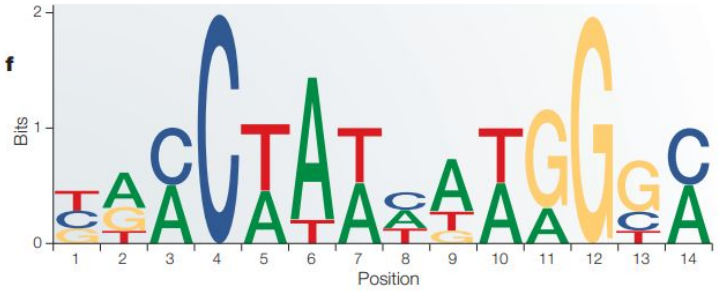
Sequence-specific TF binding

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.



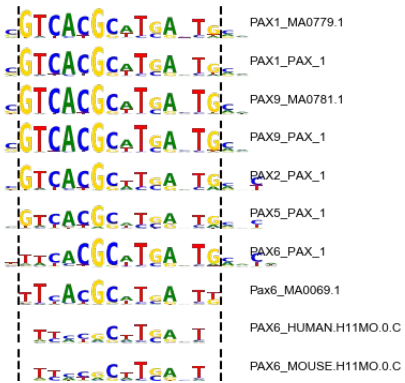


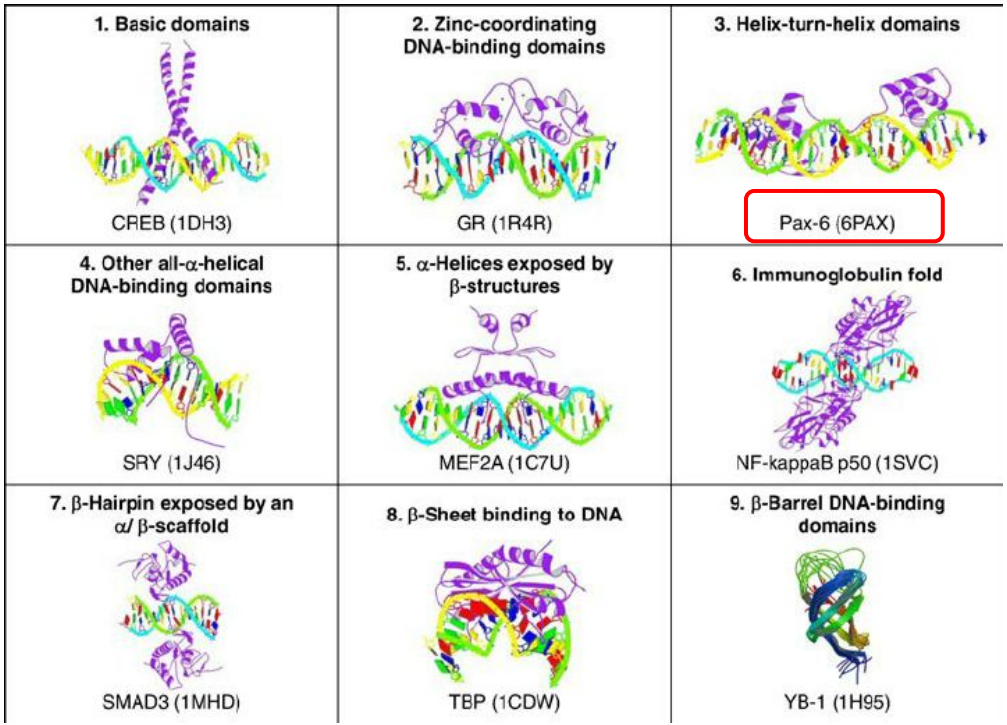
Example of a TF binding model

Wassermand & Sandelin 2004. Applied bioinformatics for the identification of regulatory elements.

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.

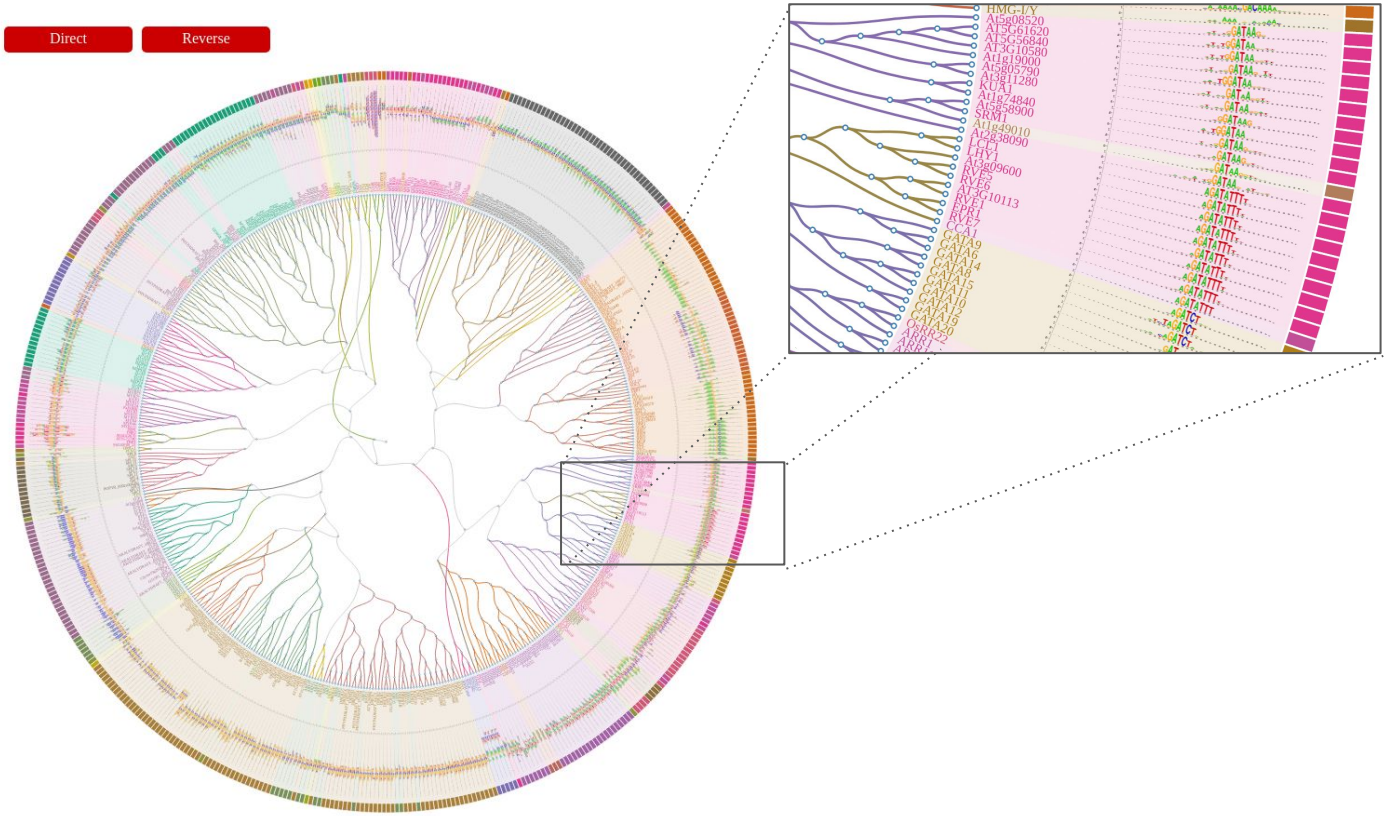




Wingender 2011. Criteria for an updated classification of human transcription factor DNA-binding domains.
https://resources.altius.org/~jvierstra/projects/motif-clustering/releases/v1.0/cluster_viz.html

TFs with similar DBD have similar motifs

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



- Click on the TF name to open its corresponding link in JASPAR
- Click on the *Direct*/*Reverse* buttons to change the motif orientation
- Ring colors: TF classes
- Branch colors: Clusters assigned by RSAT matrix-clustering
  See summary

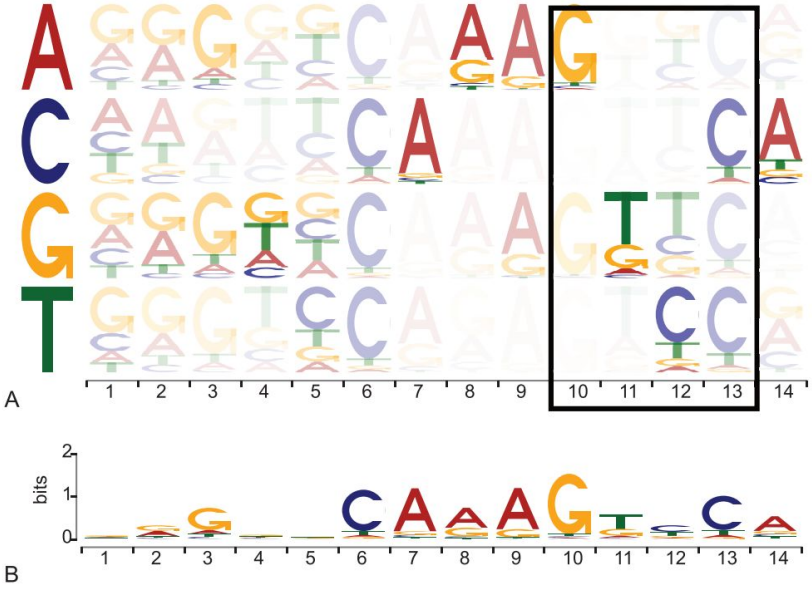
https://jaspar.genereg.net/matrix-clusters/plants/

Complex representation of motifs

HMM

Mathelier 2013

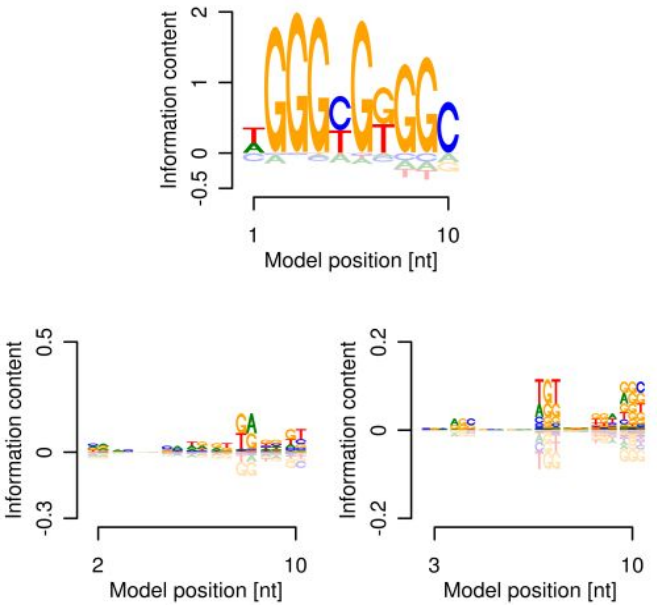
Nucleotide dependencies are modelled



Bayesian Markov Models

Siebert 2016

Supports long k-mers



Complex representation of motifs

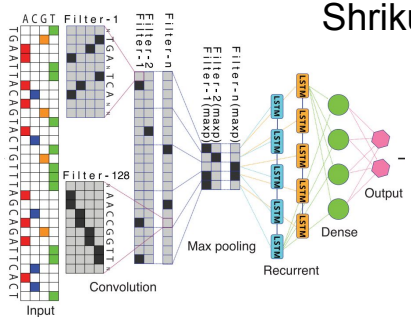
Dinucleotide PWMs

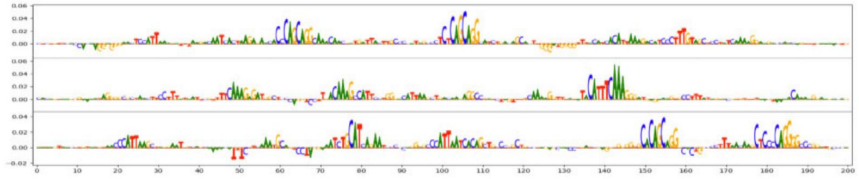
Kulakovskiy 2018

	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



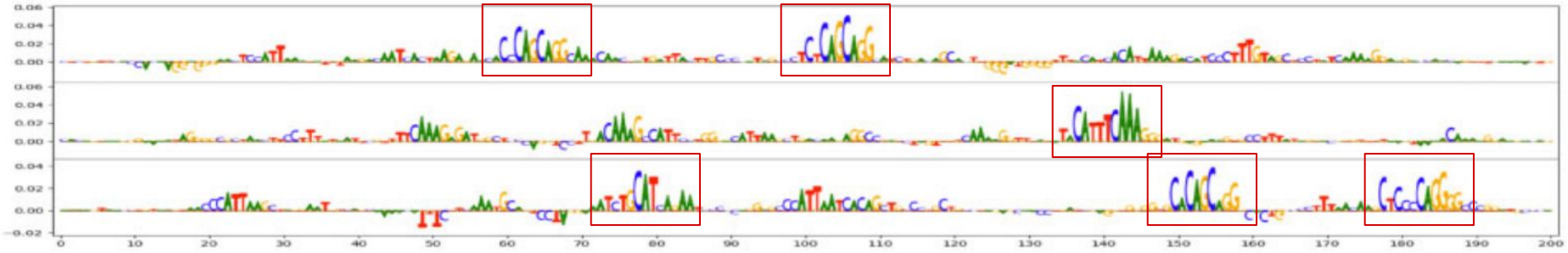


Finding nucleotides that discriminates between classes, e.g., bound/unbound

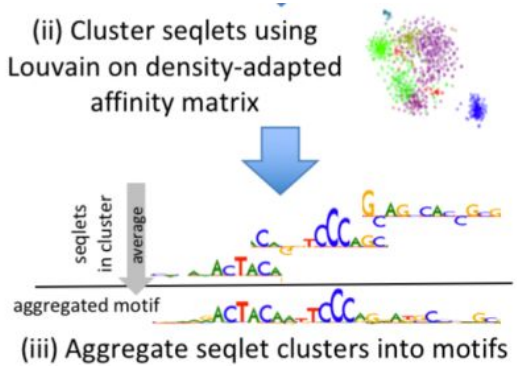
Complex representation of motifs

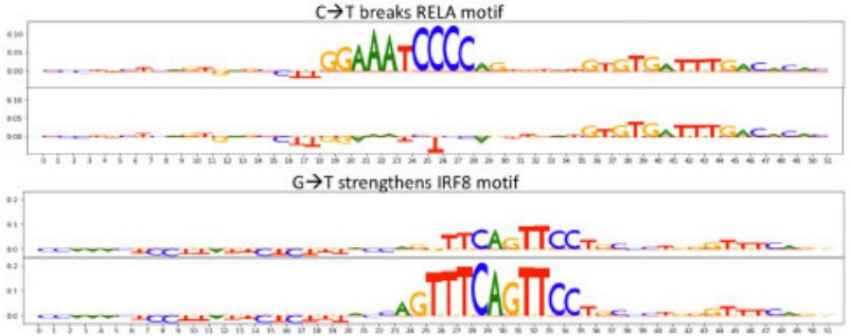
Motifs derived from deep learning or SVM or NMF models

Shrikumar 2019



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

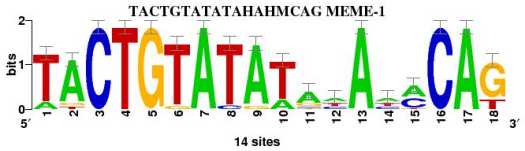




TF binding motifs are represented in different formats

MEME

```
MEME version 5.4.0

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  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 ]
```

TRANSFAC

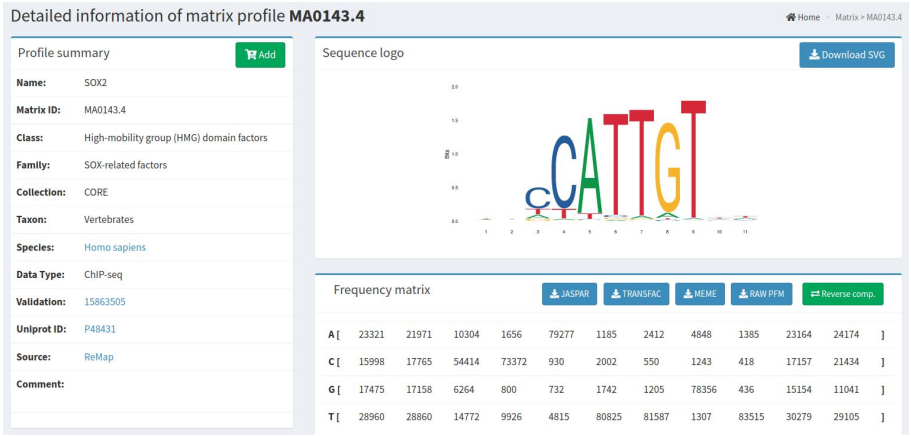
```
AC  Motif_1
XX
ID  USF1
XX
DE  TACTGTATAtrhAhmCAG
P0          A          C          G          T
1         2.0        0.0        0.0       12.0
2        12.0        0.0        1.0        1.0
3         0.0       14.0        0.0        0.0
4         0.0        0.0        0.0       14.0
5         0.0        0.0       14.0        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: TACTGTATAtacAccCAG
CC  consensus.strict.rc: CTGGGTGTATATACAGTA
CC  consensus.IUPAC: TACTGTATAtrhAhmCAG
CC  consensus.IUPAC.rc: CTGKDTDYATATACAGTA
CC  consensus.regexp: TACTGTATAt[ag][act]A[act][ac]CAG
CC  consensus.regexp.rc: CTG[GT][AGT]T[AGT][CT]ATATACAGTA
XX
//
```

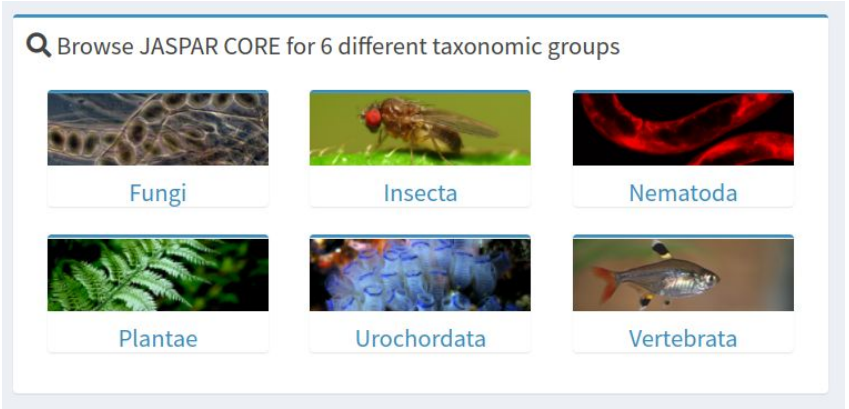
TAB

```
A       2.0     12.0    0.0     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     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     0.0     14.0    1.0     0.0     0.0     1.0     0.0     3.0     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
//
```

JASPAR

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





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/

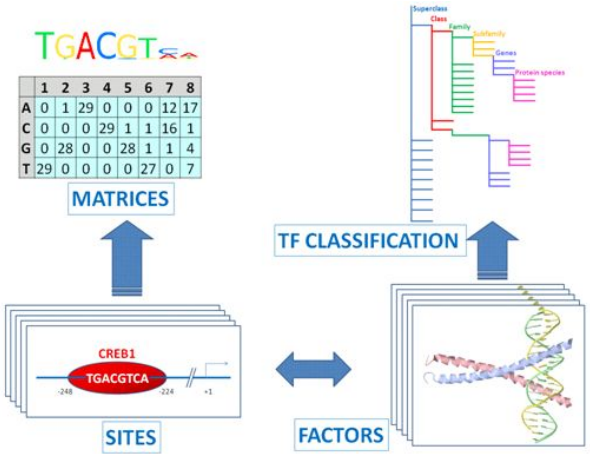
Name/Motif ID	Species	Forward	Reverse	Type/Study/Study ID	SR Score	DBD Identity
Gata3 M00166_2.00	*Mus musculus*	 NNNHGATHNN	 NNDATCDNNN	PBM Badis et al.(2009) Gata3_1024	(Direct)	(Direct)
Gata3 M05867_2.00	*Mus musculus*	 GATAAN	 NTTATC	SMiLE-seq Isakova et al.(2017) GATA3	(Direct)	(Direct)
Gata3 M09122_2.00	*Mus musculus*	 NVWGATAAVVDBH	 DVHBBTTATCWBN	Misc Kulakovskiy et al.(2013) GATA3_MOUSE.H11MO.0.A	(Direct)	(Direct)
Gata3 M09551_2.00	*Mus musculus*	 AGATBNNNAGATARV	 BYTATCTNNNVATCT	Misc Heinz et al.(2010) iTreg-Gata3_GSE20898_1	(Direct)	(Direct)
Gata3 M09552_2.00	*Mus musculus*	 AGATBDNNNNDSWGATWRNN	 NNYWATCWSHNNNNHVATCT	Misc Heinz et al.(2010) iTreg-Gata3_GSE20898_2	(Direct)	(Direct)
Gata3 M09553_2.00	*Mus musculus*	 NNNNNNWGATNDNATCWNNN	 NNNWGATNHNATCWNNNNNN	Misc Heinz et al.(2010) iTreg-Gata3_GSE20898_3	(Direct)	(Direct)
Gata3 M09554_2.00	*Mus musculus*	 NWGATNNNNATCWNN	 NNWGATNNNNATCWN	Misc Heinz et al.(2010) iTreg-Gata3_GSE20898_4	(Direct)	(Direct)
Gata3 M09555_2.00	*Mus musculus*	 WGATAAVV	 BBTTATCW	Misc Heinz et al.(2010) iTreg-Gata3_GSE20898_5	(Direct)	(Direct)

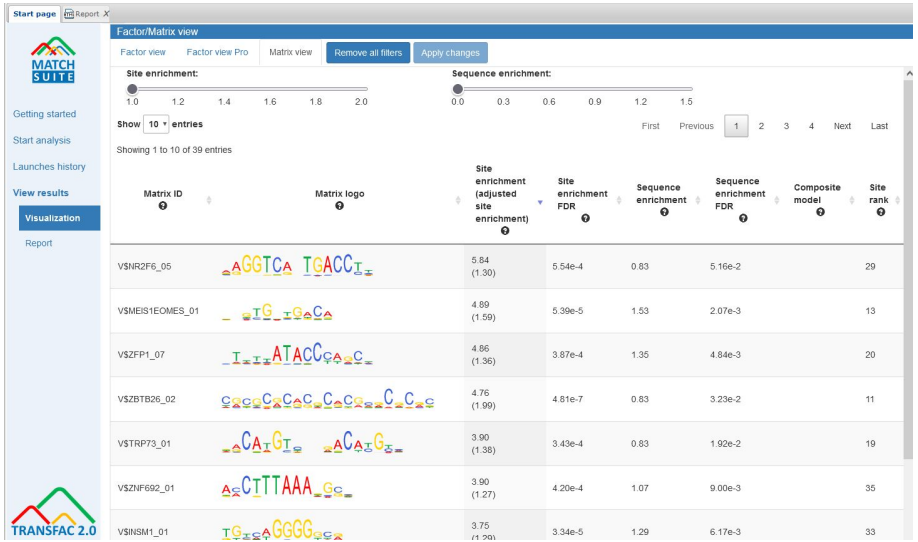
Name	Species	Gene ID	Family	Motif Evidence
GATA3	*Arabidopsis_thaliana*	AT4G34680	GATA	Direct
GATA3	*Homo_sapiens*	ENSG00000107485	GATA	Direct
Gata3	*Mus_musculus*	ENSMUSG00000015619	GATA	Direct
GATA3	*Anolis_carolinensis*	ENSACAG00000009109	GATA	Inferred
GATA3	*Arabidopsis_lyrata*	fgenesh2_kg.7__620__	GATA	Inferred
GATA3	*Bos_taurus*	ENSBTAG00000017243	GATA	Inferred
GATA3	*Canis_familiaris*	ENSCAFG00000005019	GATA	Inferred
GATA3	*Cavia_porcellus*	ENSCPOG00000008595	GATA	Inferred
GATA3	*Choloepus_hoffmanni*	ENSCHOG00000005894	GATA	Inferred
gata3	*Danio_rerio*	ENSDARG00000016526	GATA	Inferred
GATA3	*Dasypus_novemcinctus*	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

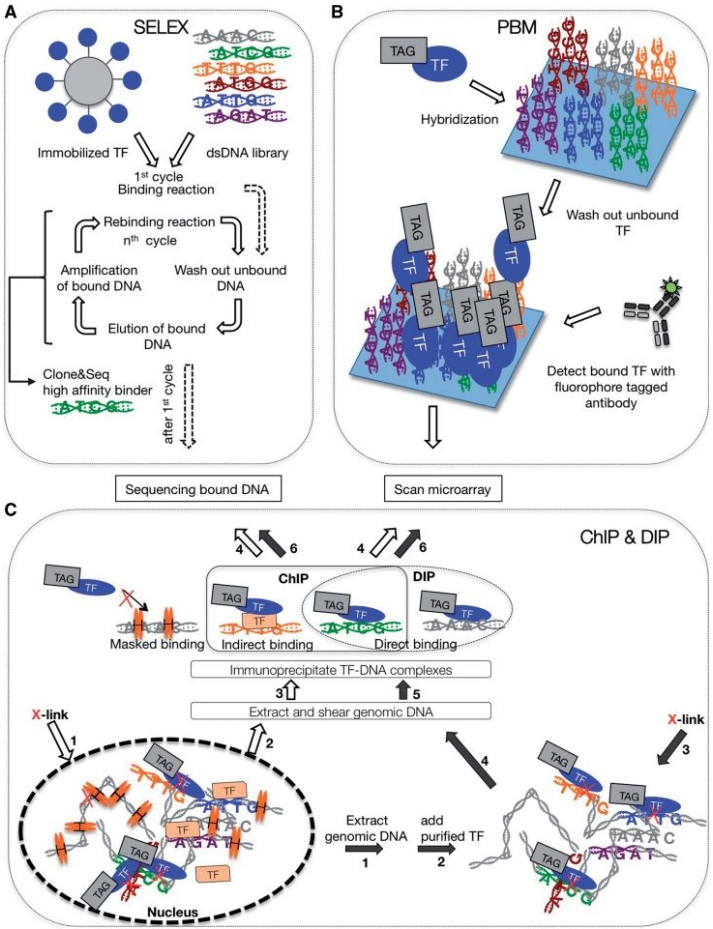
- TF DNA binding motifs can be represented using various models.
- It is important to know what your model represents in order to properly interpret the TF binding.
- TF DNA binding motifs can be represented in different formats.
- Multiple resources are storing TF DNA binding motifs, but it is important 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?

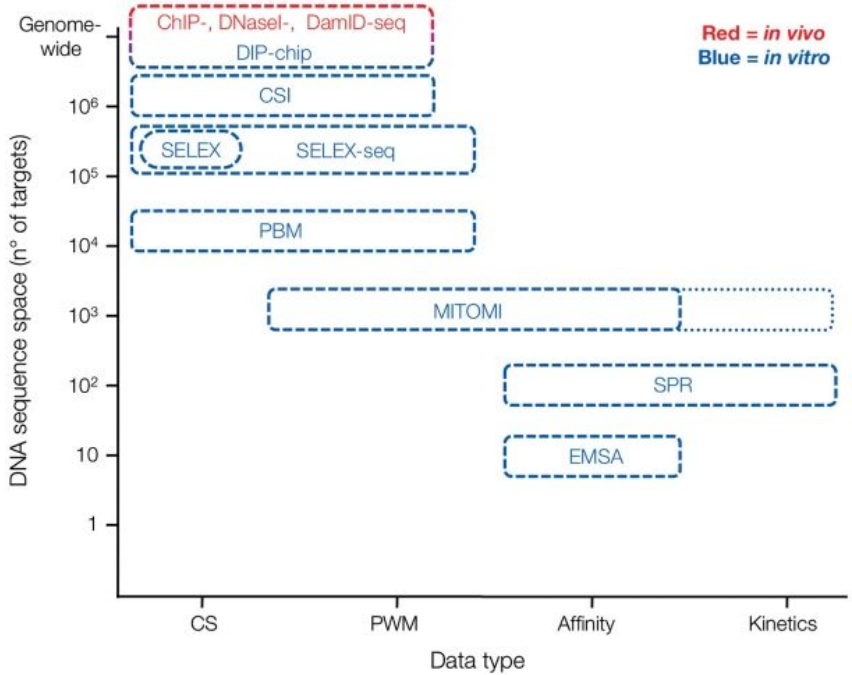
- Data processing depends on the used protocol and experimental setting
- Interpretation of the results highly depends on experimental setting
- Sometimes devil is in the details!
- Ideal scenario - collaboration between wet and dry labs **to plan** the experiment



Geertz M and Maerki SJ 2010. Experimental strategies for studying transcription factor-DNA binding specificities

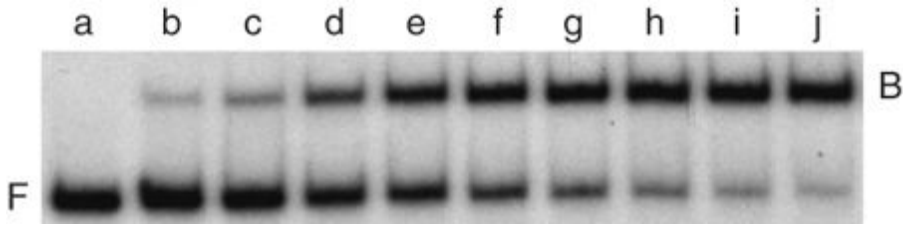
In vitro and *In vivo* techniques

- *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

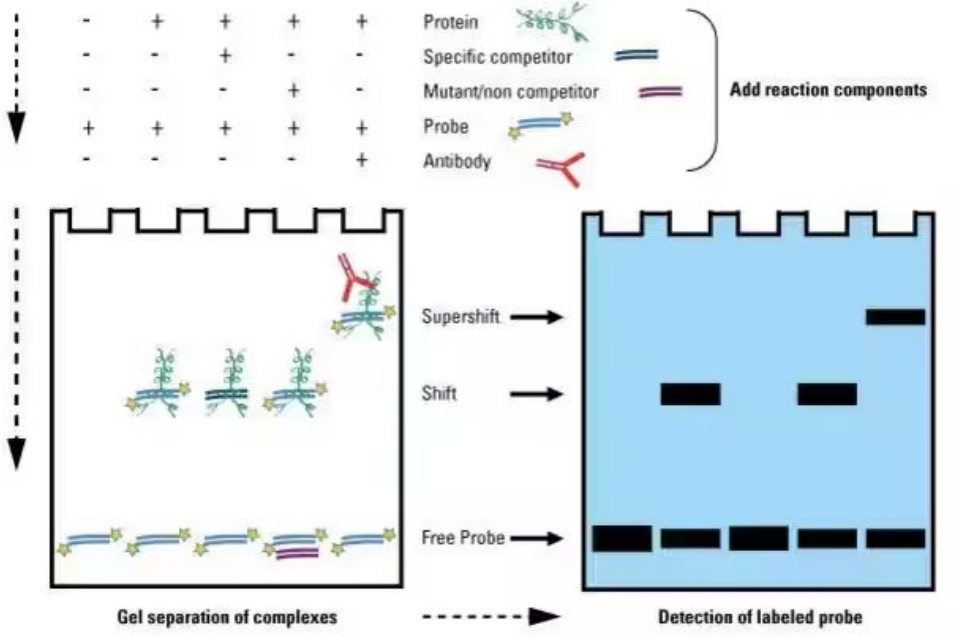


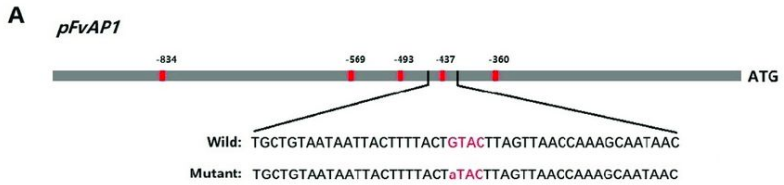
Geertz M and Maerki SJ 2010. Experimental strategies for studying transcription factor-DNA binding specificities

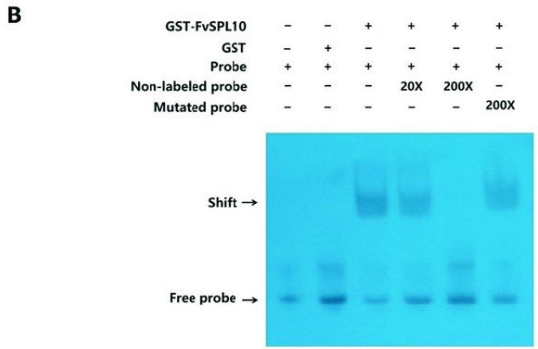
EMSA: electrophoretic mobility shift assay



F - free DNA; B - bound DNA (i.e. protein-DNA complex)



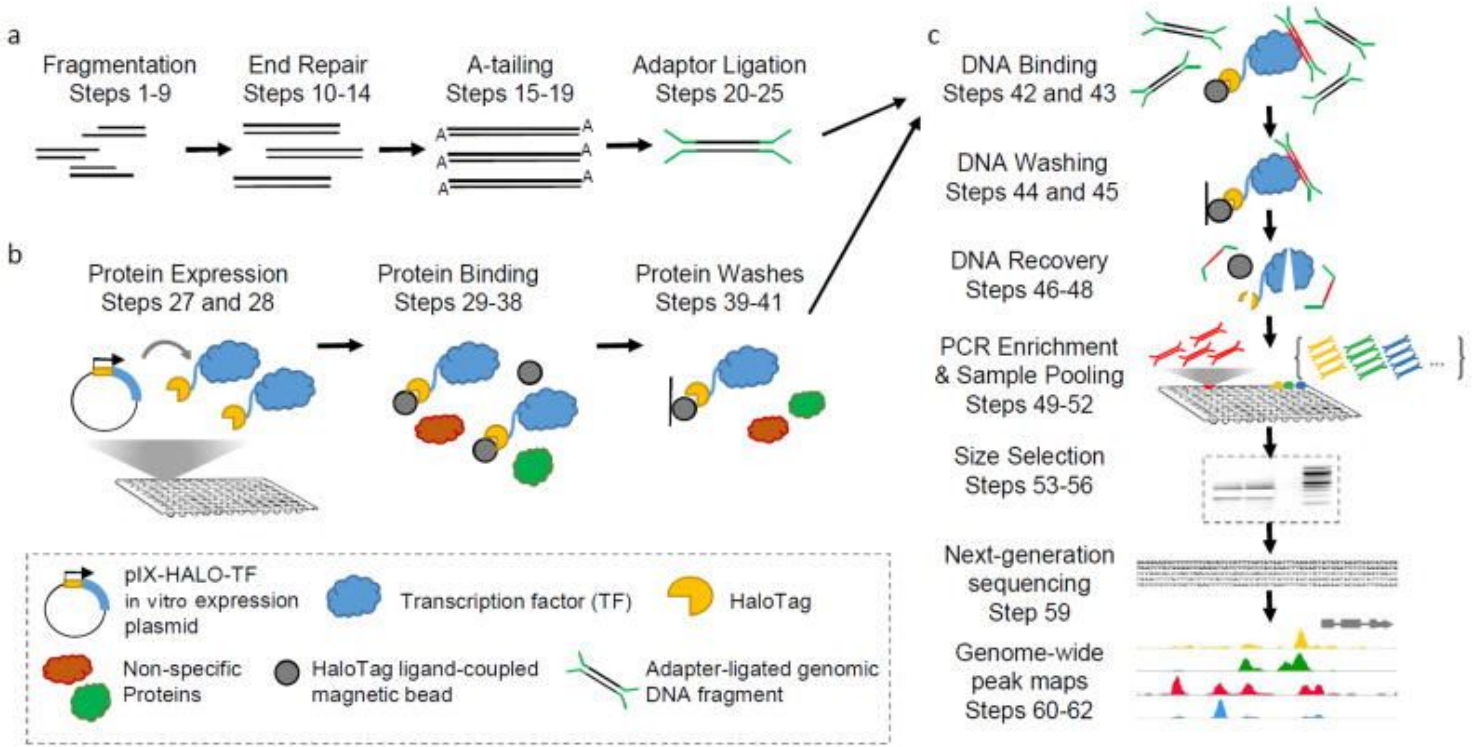




Xiong et al. 2019. Molecular cloning and characterization of SQUAMOSA-promoter binding-like gene FvSPL10 from Woodland Strawberry (Fragaria vesca)
Hellman LM and Fried MG 2007. Electrophoretic mobility assay (EMSA) for detecting protein-nucleic acid interactions
https://www.thermofisher.com/no/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/gel-shift-assays-emsa.html

DAP-seq: DNA affinity purification sequencing

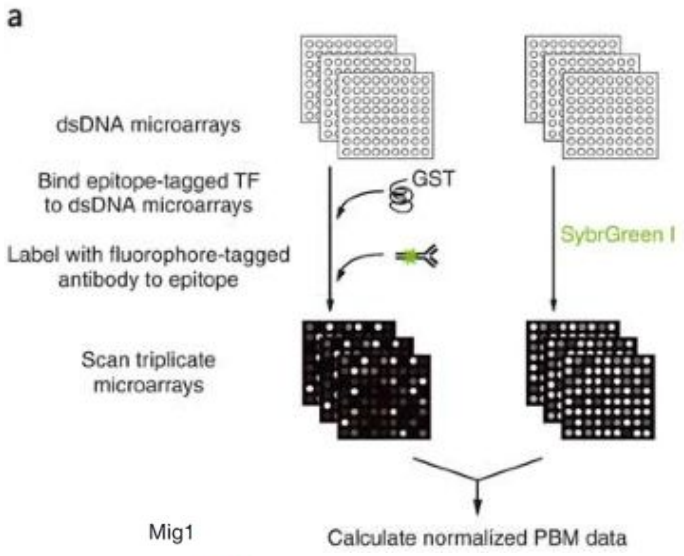
- Scalable
- Cheap
- *In vitro*

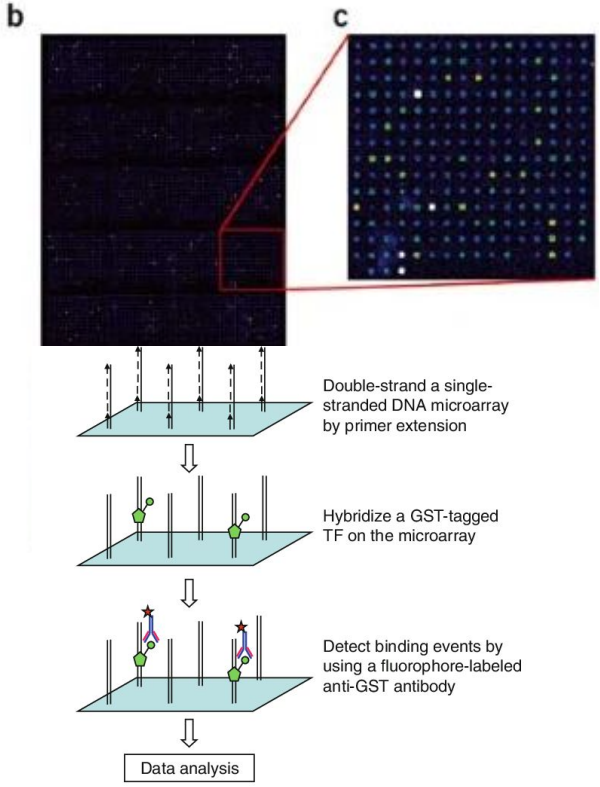


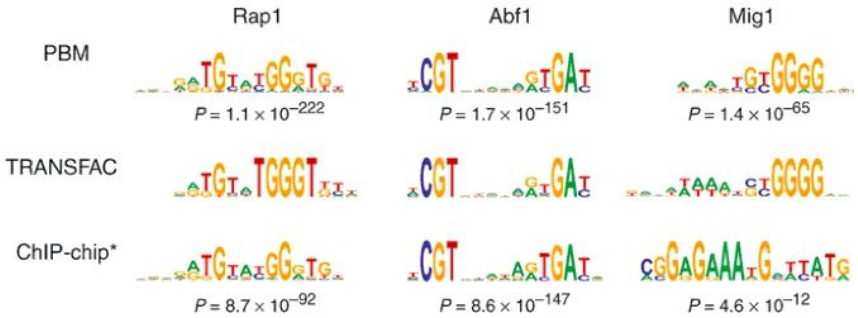
Bartlett A et al. 2018. Mapping genome-wide transcription factor binding using DAP-seq

PBM: protein binding microarray

- High-throughput
- *In vitro*
- Quick





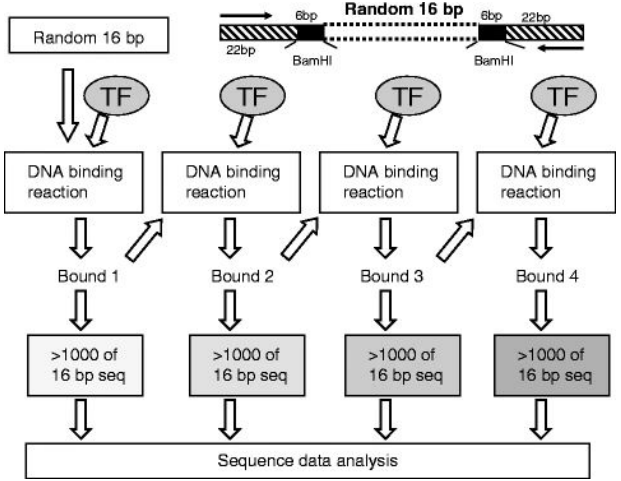


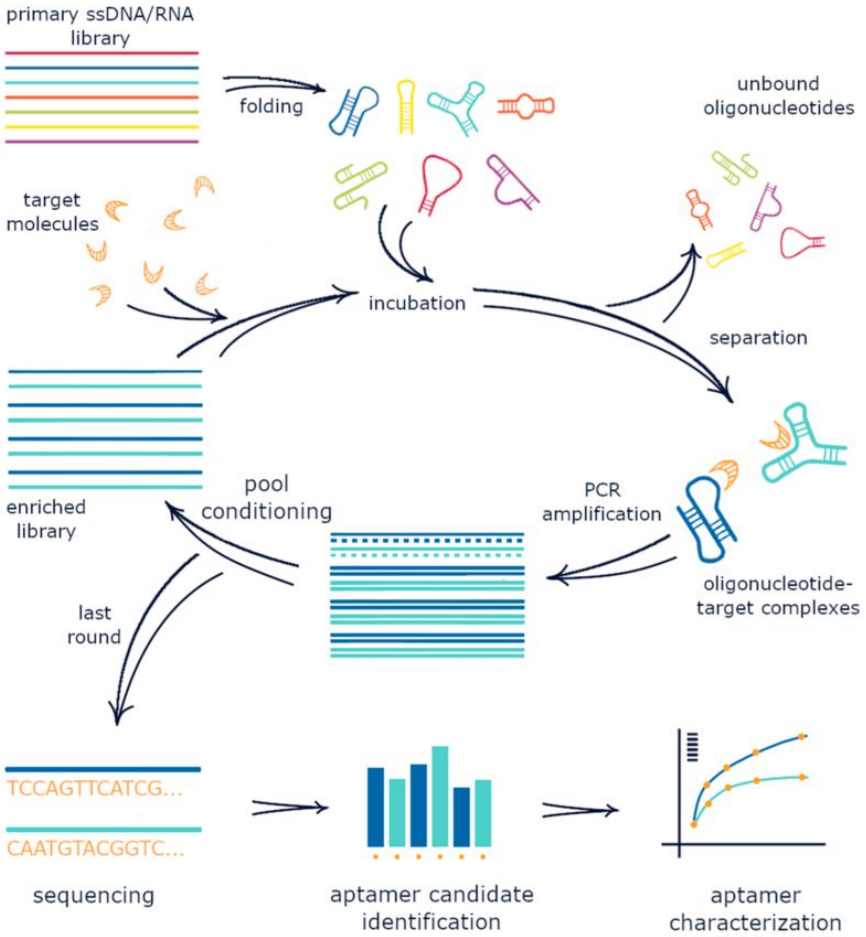
Xie Z et al. 2011. Systemic characterization of protein-DNA interactions

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

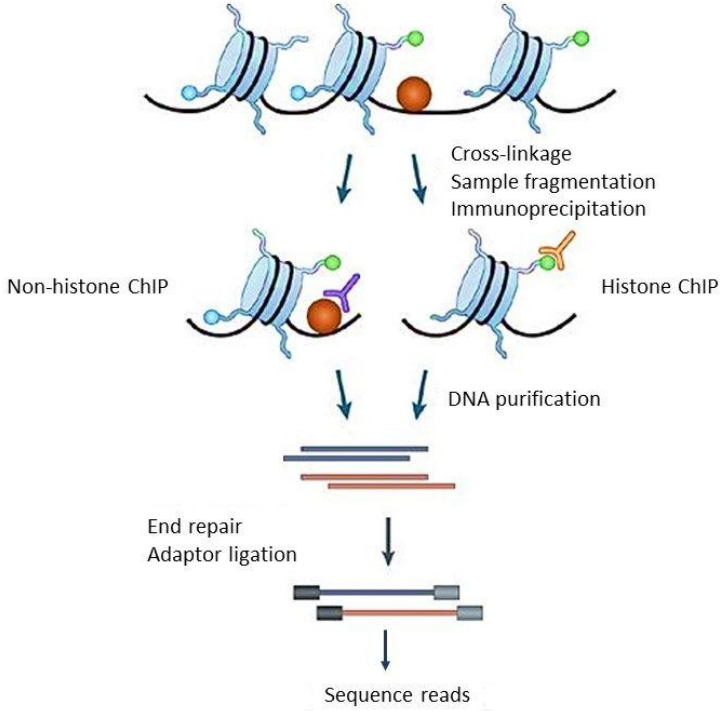




Komarova N and Kuznetsov A. 2019. Inside the black box: what makes SELEX better?
Ogawa N and Biggin MD. 2011. High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro

ChIP-seq: chromatin immunoprecipitation followed by sequencing

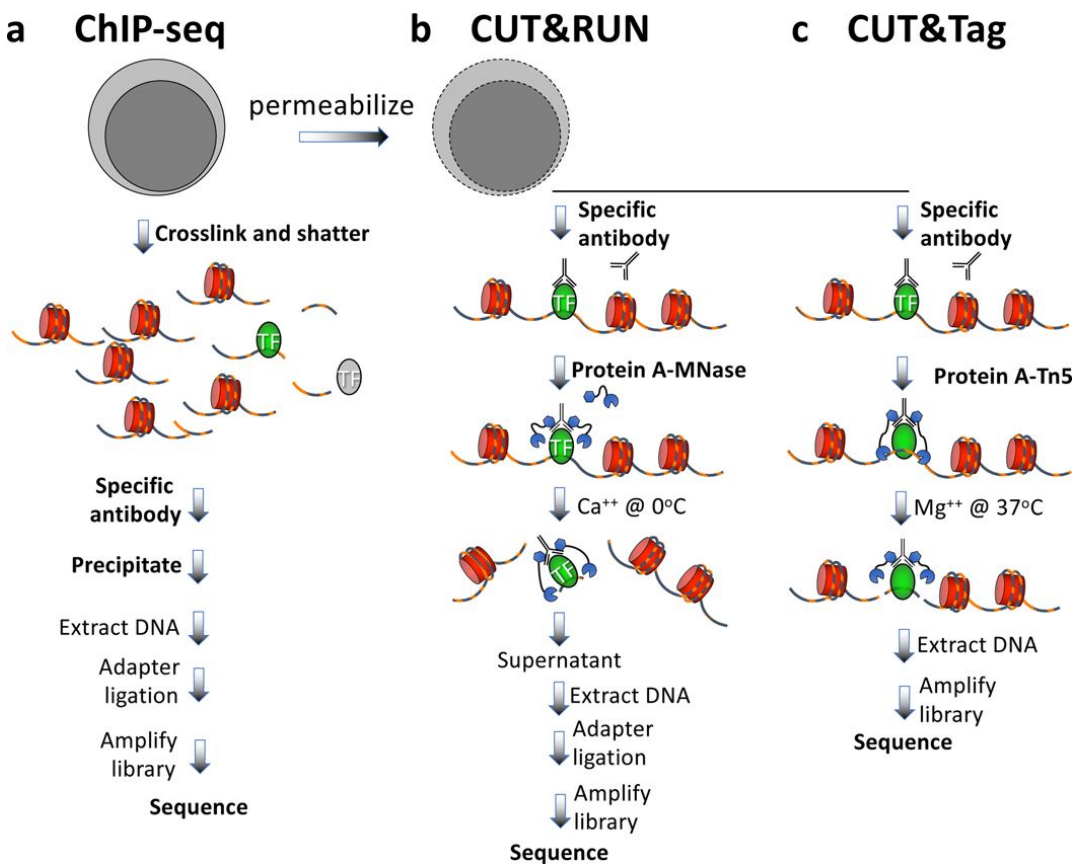
- Still is probably the main technique to map TF binding
- Major limitation - one TF/histone modification per sample
- Genome-wide
- Unlike arrays does not require prior knowledge



Park JP. 2009. ChIP-seq: advantages and challenges of maturing technology

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

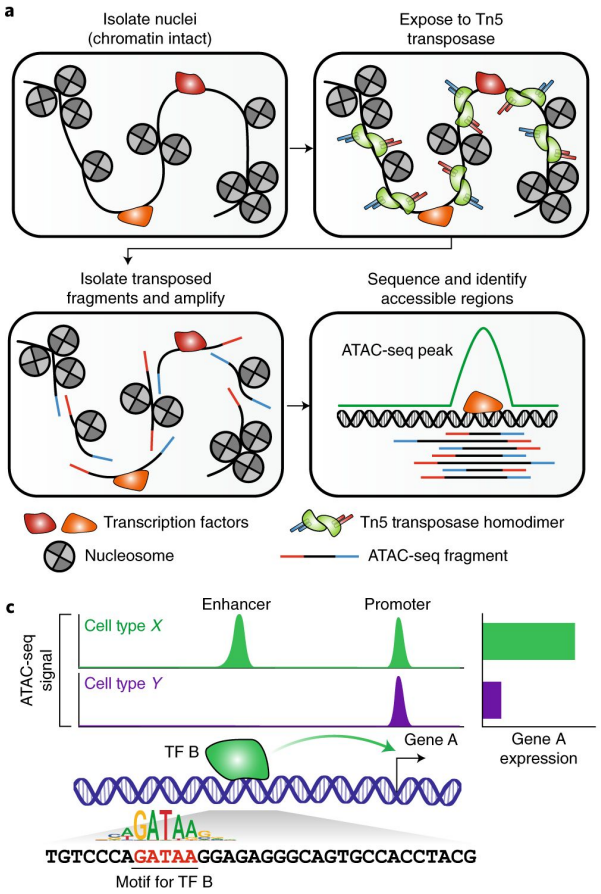
- For small samples
- Possibility to map in single-cells
- Quite needy protocol and data analysis

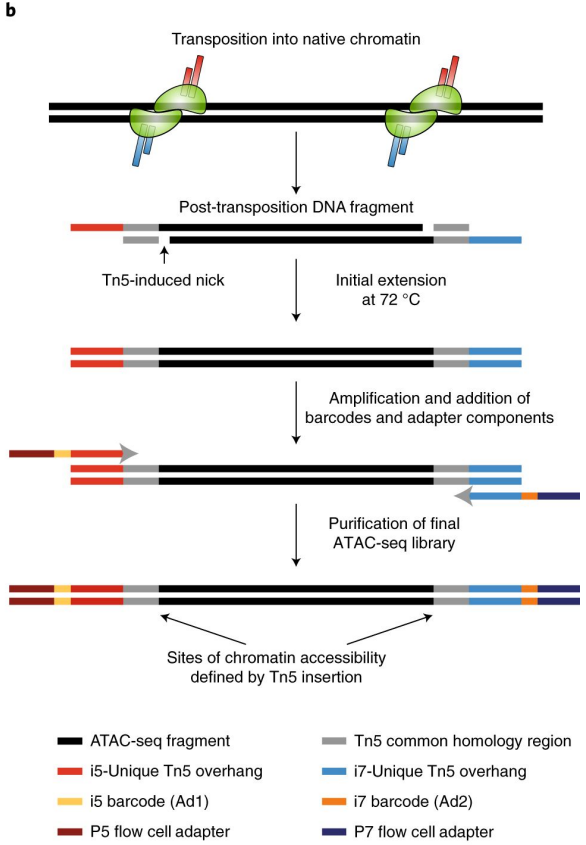


Kaya-Okur et al., 2020. Efficient low-cost chromatin profiling with CUT&Tag

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?

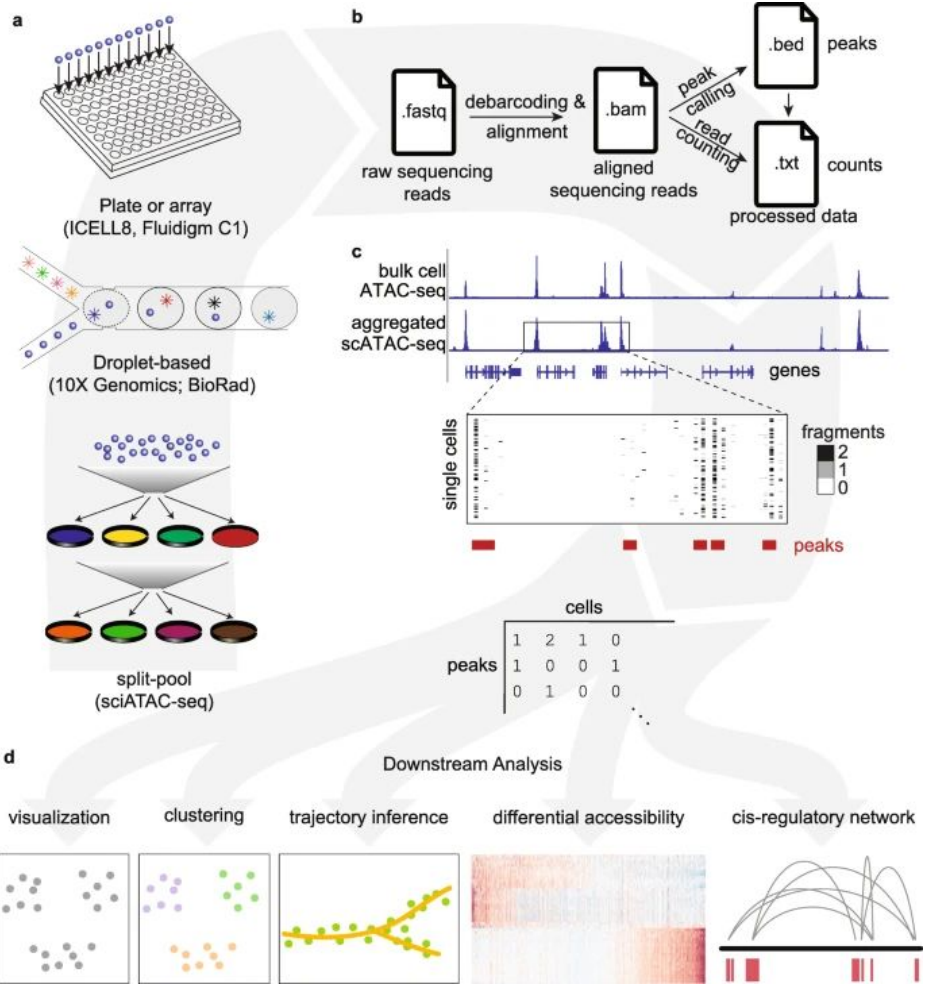


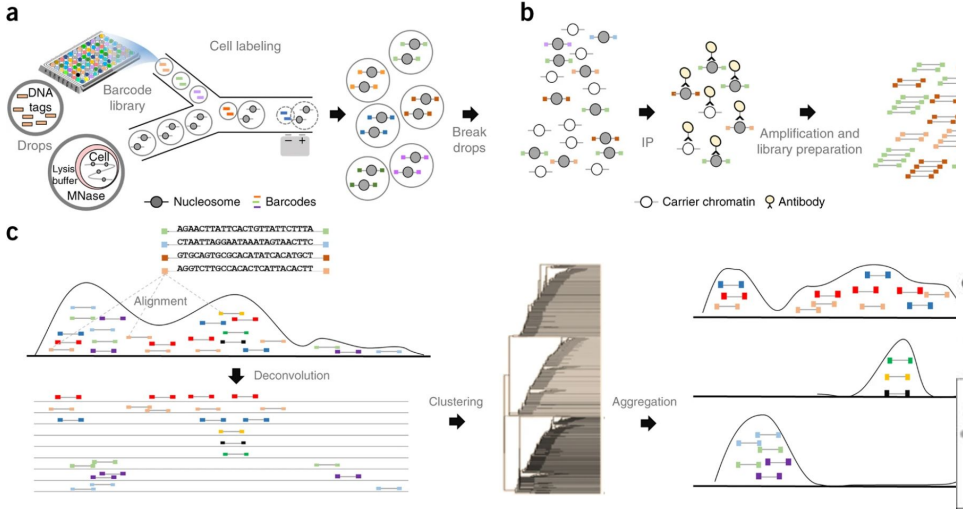


Grandi FC et al. 2022. Chromatin accessibility profiling by ATAC-seq

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?

Check motif databases

2. Do your find the expected motif?

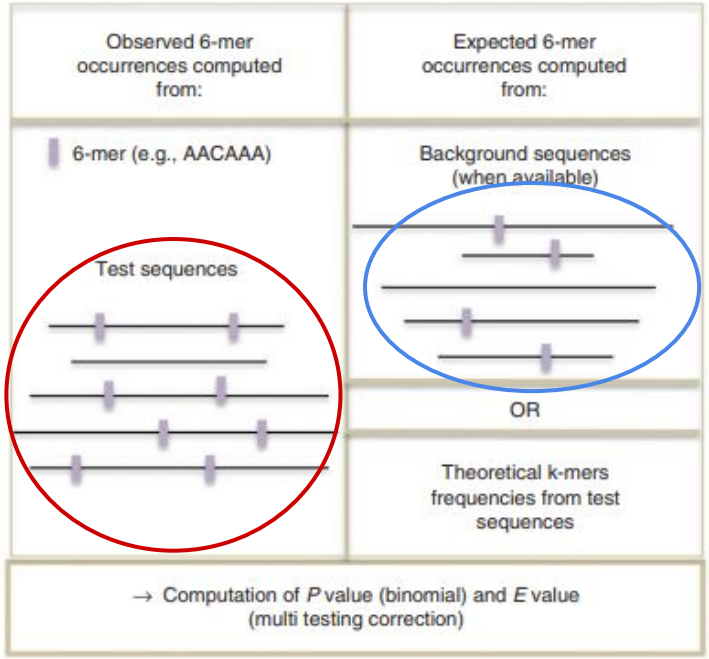
CAREFUL! ⇒ Do *de novo* motif discovery to check what is in your peaks

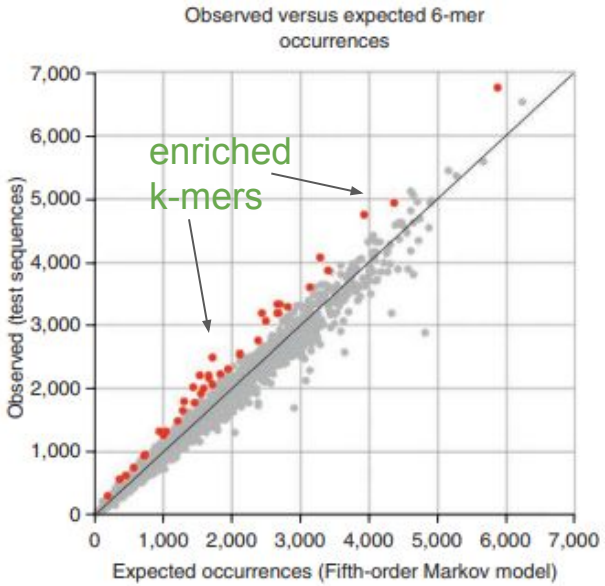
3. Which genomic regions have the expected motif?

⇒ Motif scanning

De novo motif discovery

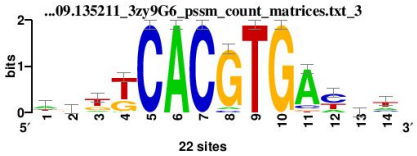
Step 1: detection of over-represented k-mers (observed vs expected.)





Step 2: alignment of enriched k-mers

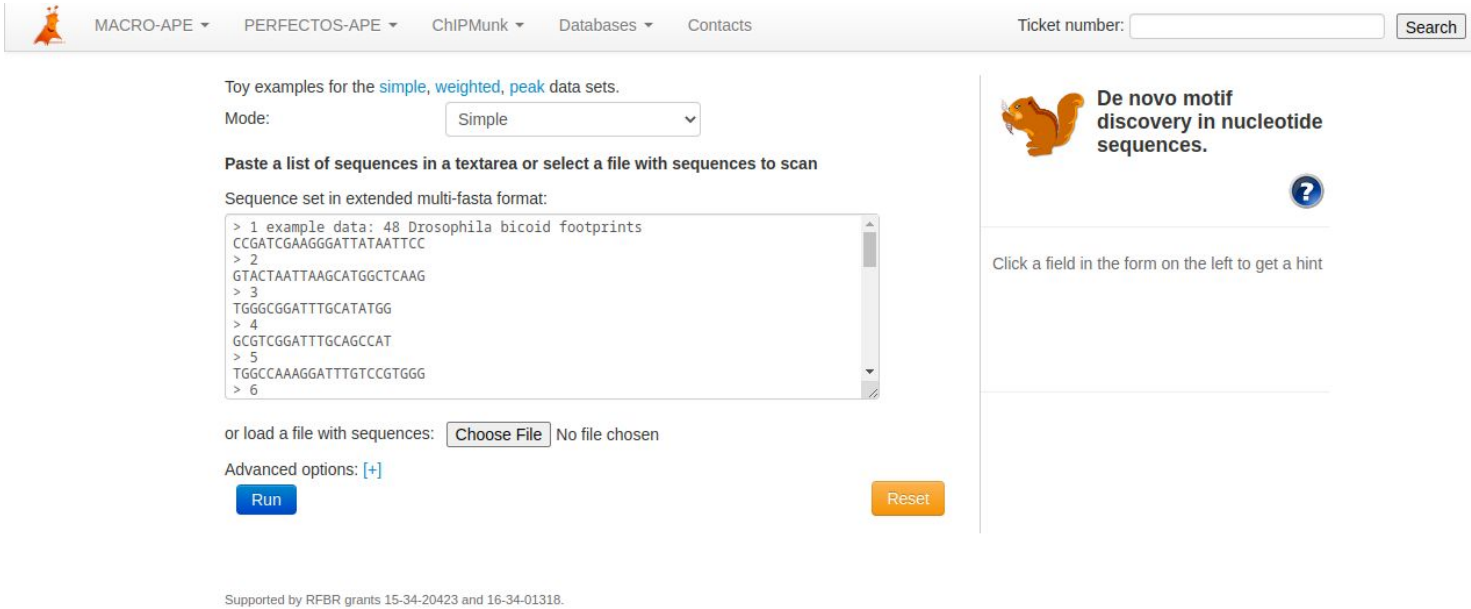
```
; assembly # 1  seed: cacgtga    11 assembled patterns
;    aligmt          rev_cpl      score
gtcacgt...       ...acgtgac      2.22
gtcacgtg..       ..cacgtgac      2.09
gtcacg....       ....cgtgac      0.29
.tcacgtg..       ..cacgtga.      10.78
.tcacgt...       ...acgtga.      4.85
..cacgtga.       .tcacgtg..      10.78
..cacgtg..       ..cacgtg..      9.00
..cacgtgac       gtcacgtg..      2.09
...acgtga.       .tcacgt...      4.85
...acgtgac       gtcacgt...      2.22
....cgtgac       gtcacg....      0.29
gtcacgtgac       gtcacgtgac      10.78   best consensus
```



Adapted from Thomas-Chollier et al 2012.

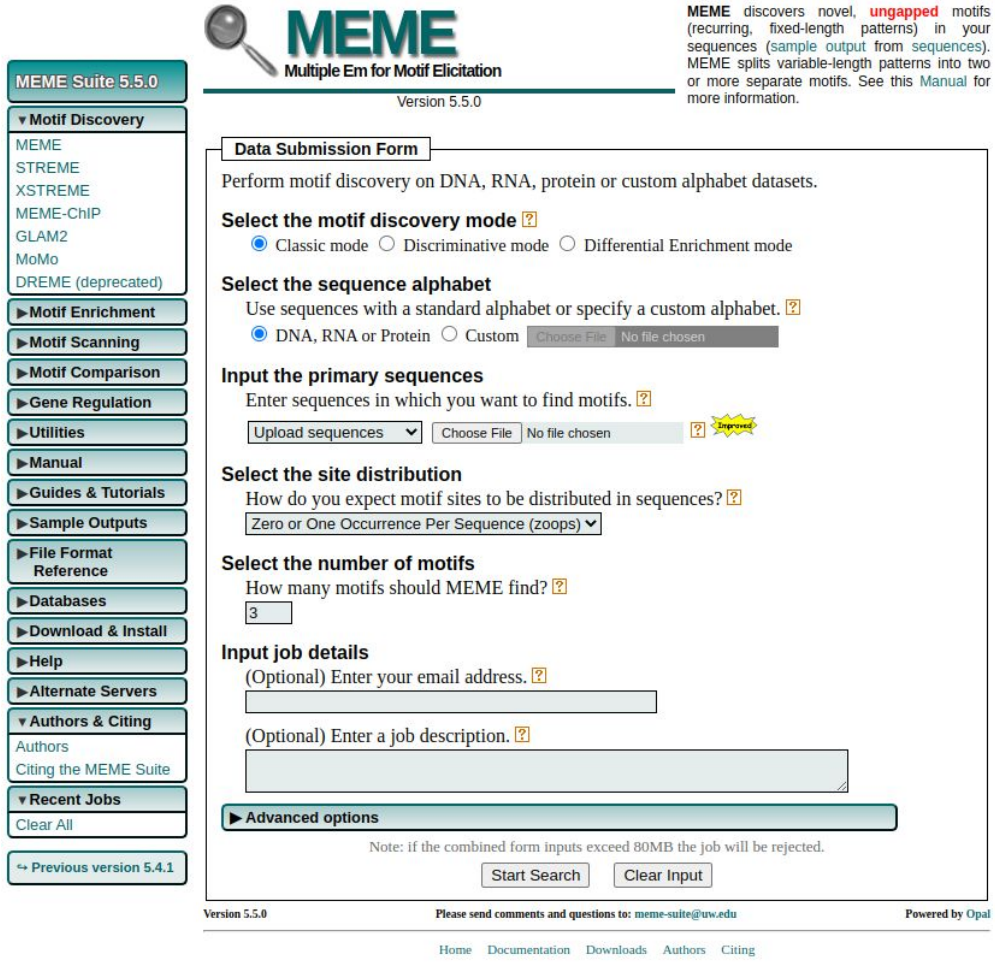
ChIPMunk

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



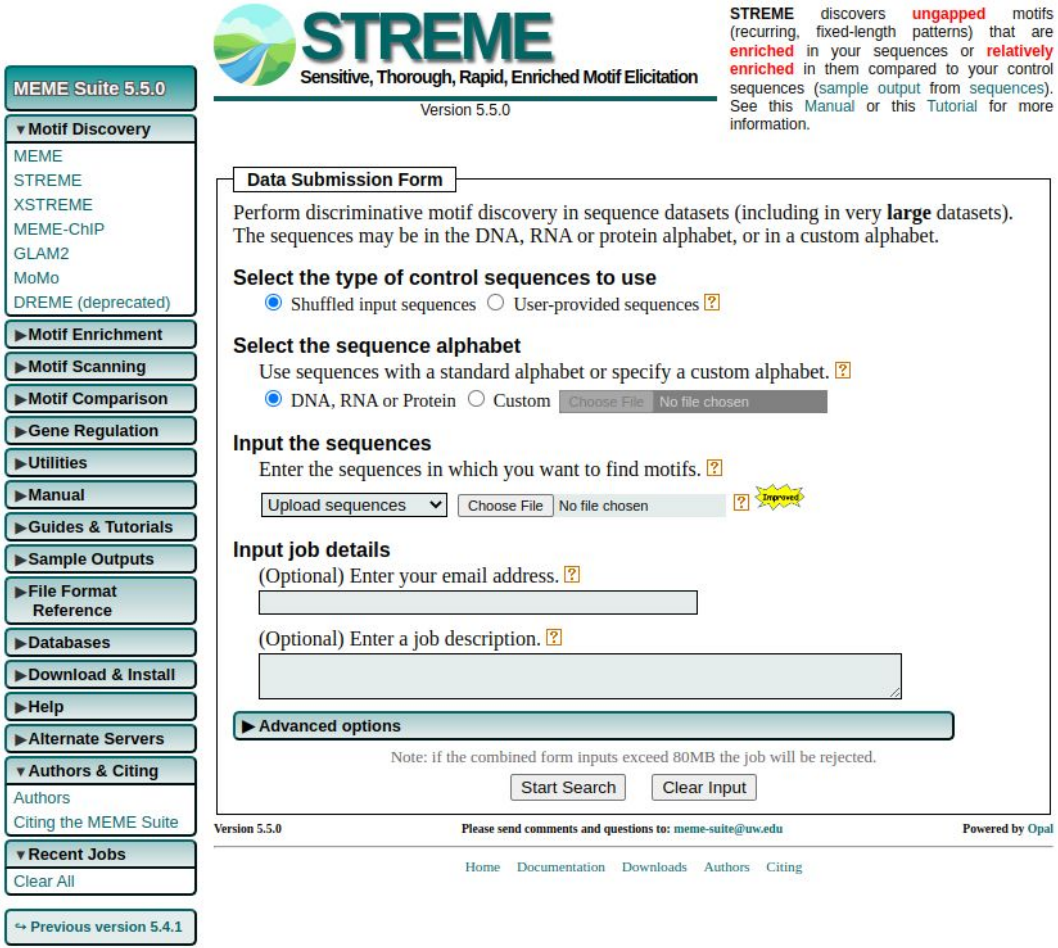
MEME

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



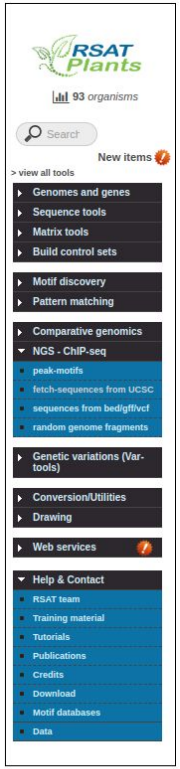
STREME

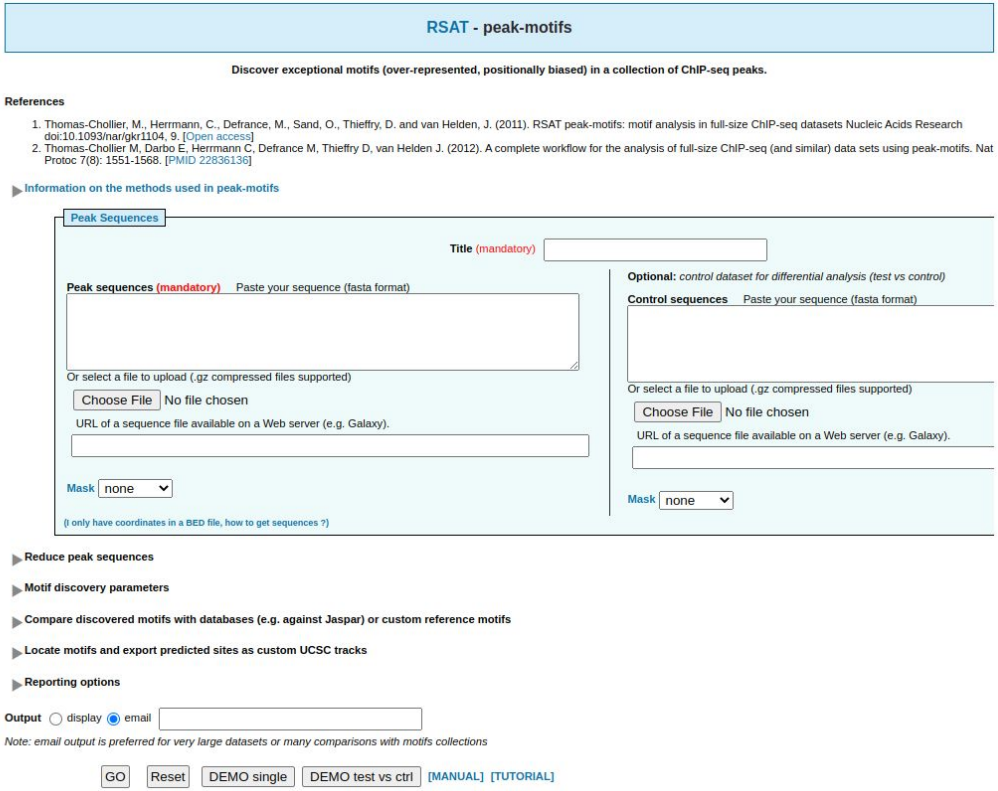
- Web application and command line
- https://meme-suite.org/meme/tools/streme



RSAT peak-motif

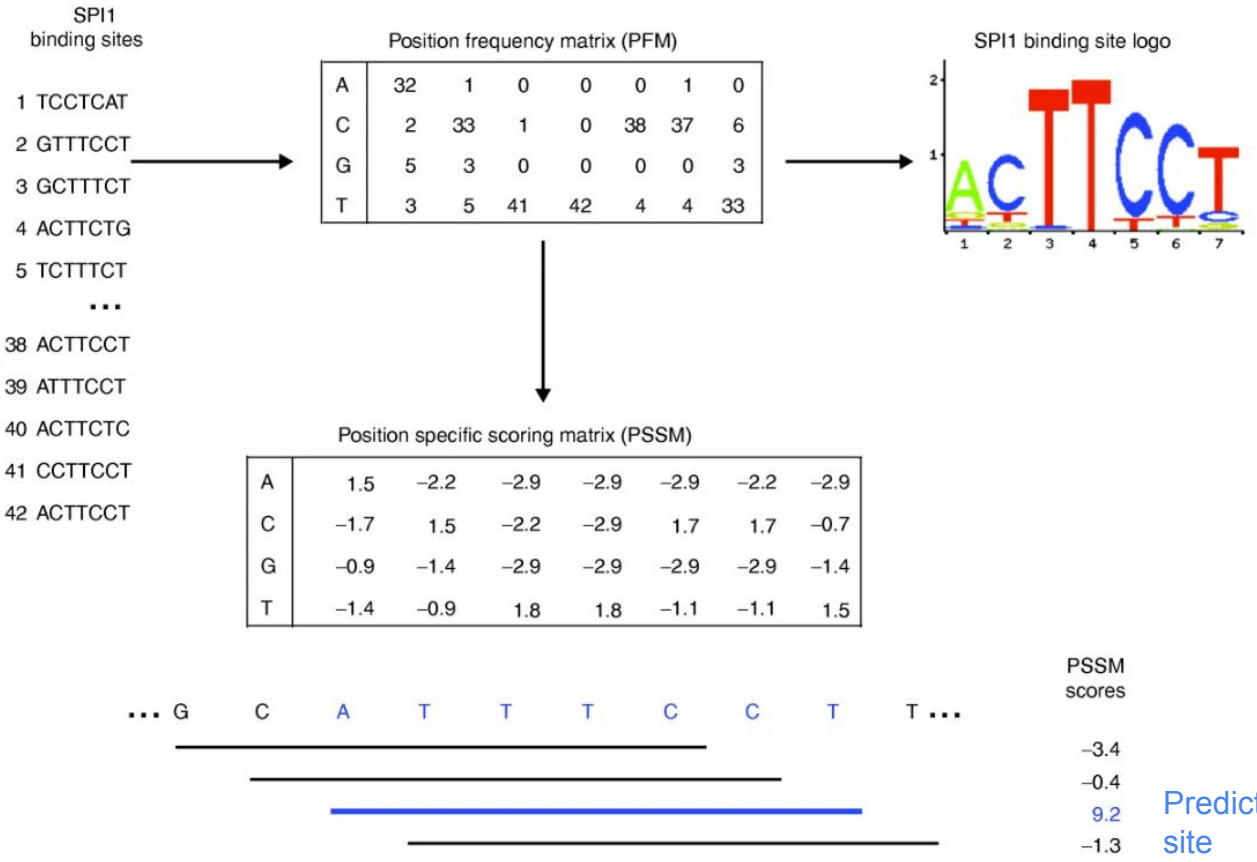
- Web application and command line
- http://rsat.eead.csic.es/plants/peak-motifs_form.cgi





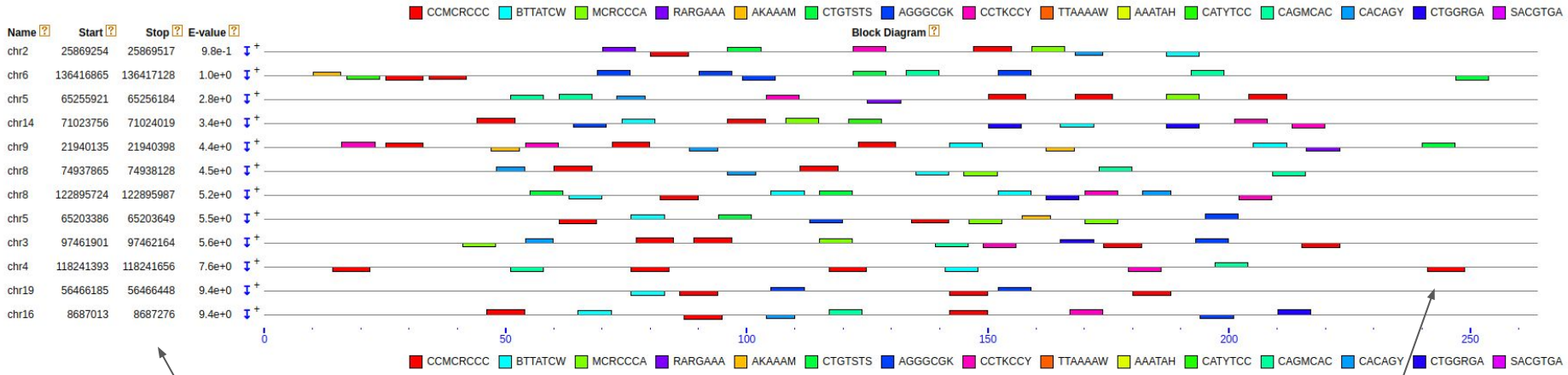
Motif scanning

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



Worsley-Hunt et al 2011.

Motif scanning



TFBS genomic coordinates

Each color represent a different PFM

Predicted TFBS

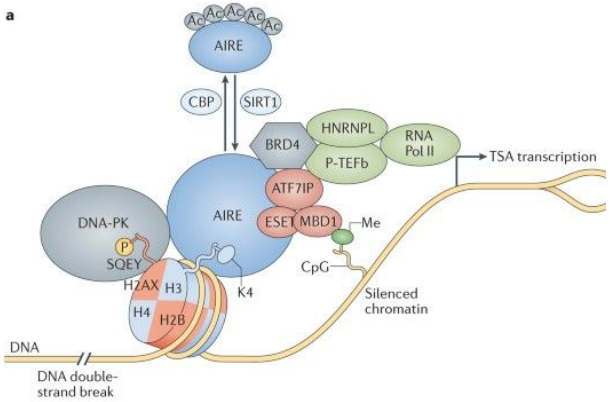
https://meme-suite.org/meme/doc/examples/mcast_example_output_files/mcast.html?man_type=web

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?

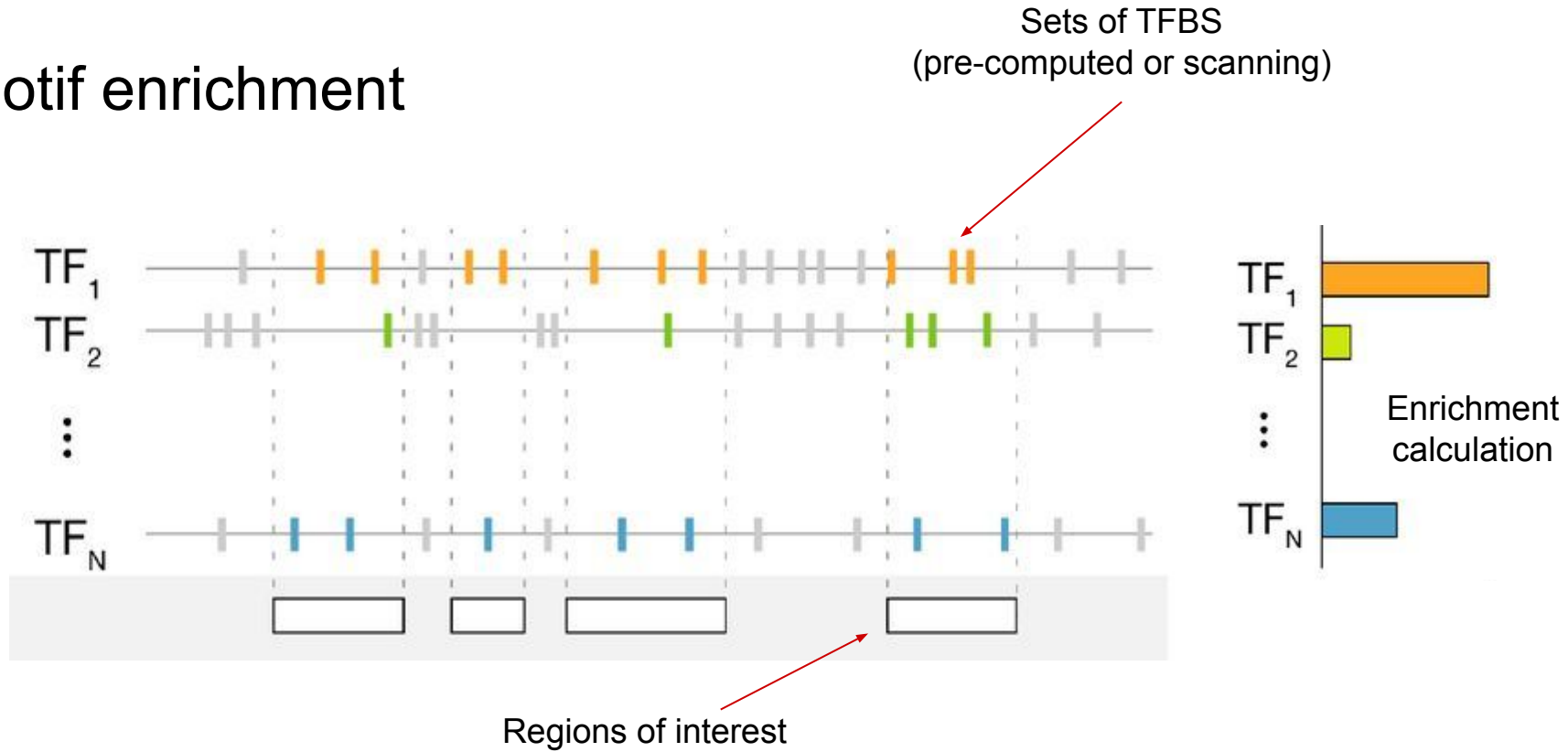
CAREFUL! ⇒ Should you take the top motif?

2. What do you know about AIRE binding?

3. Which genomic regions have the motifs of interest?

⇒ Motif scanning

Motif enrichment

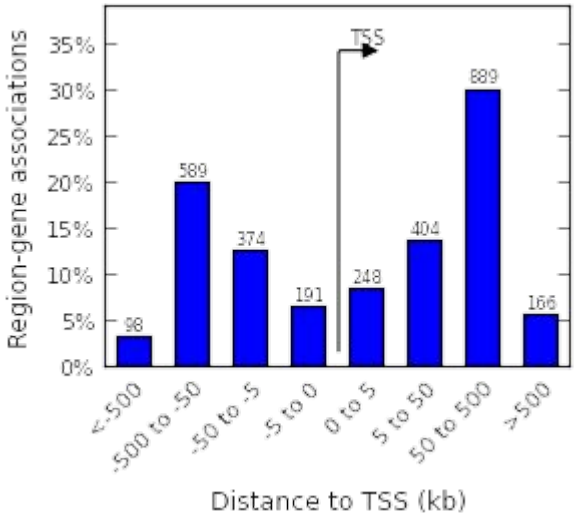


Adapted from Taniwaga et al 2022.

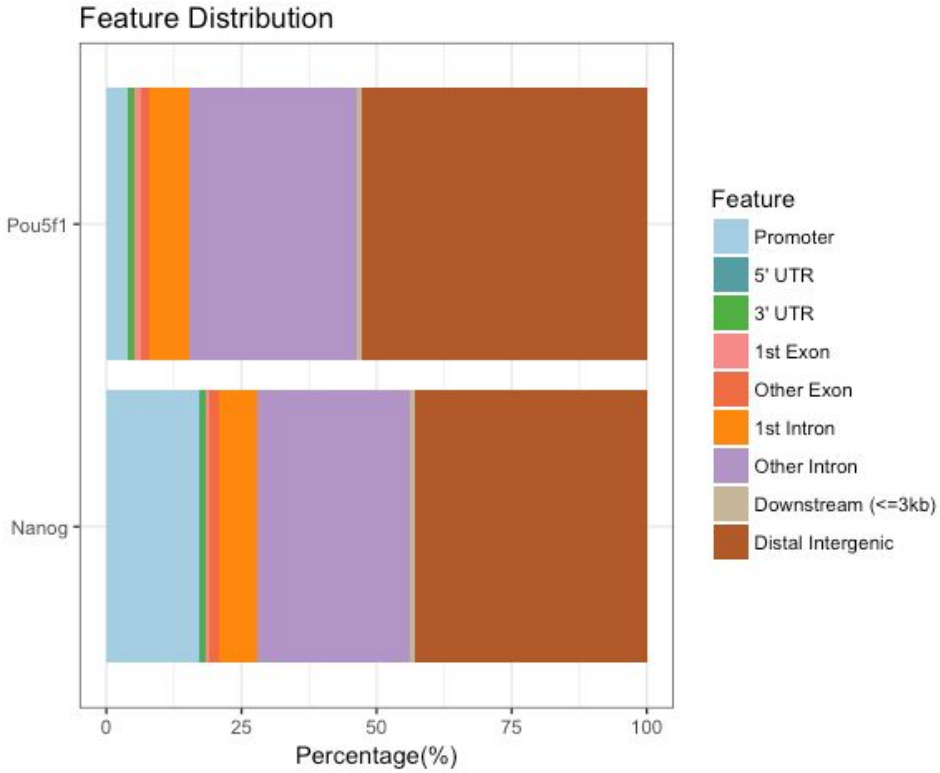
What could have gone wrong if you do not see the expected motif?

- 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

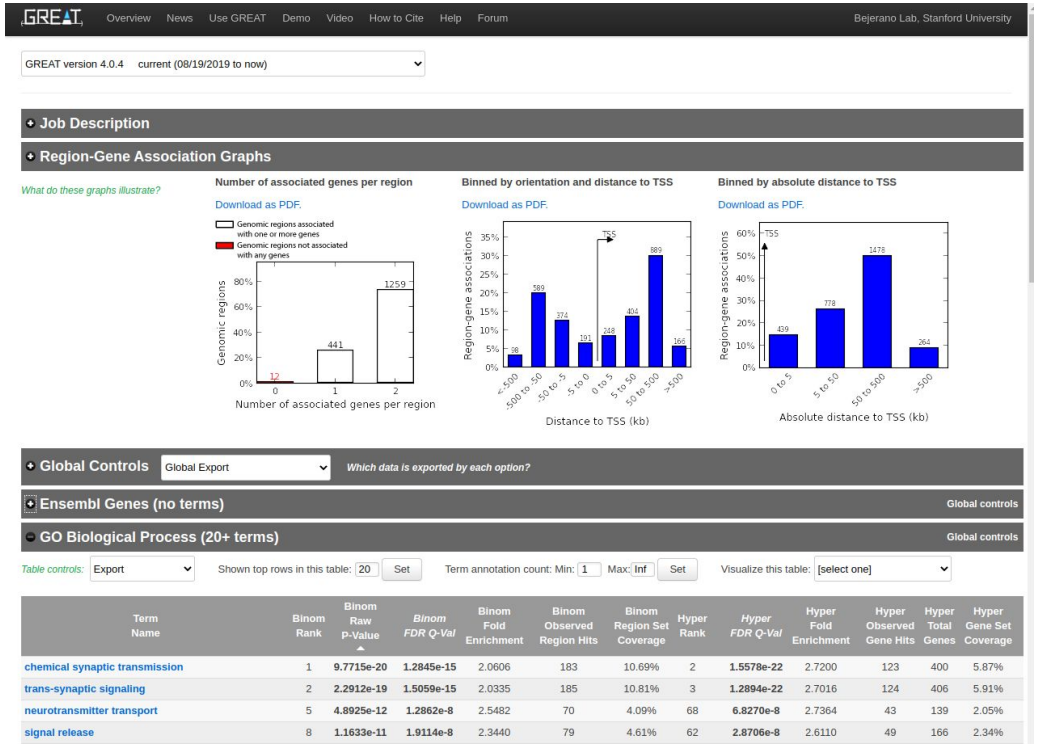


Distance to TSS with GREAT

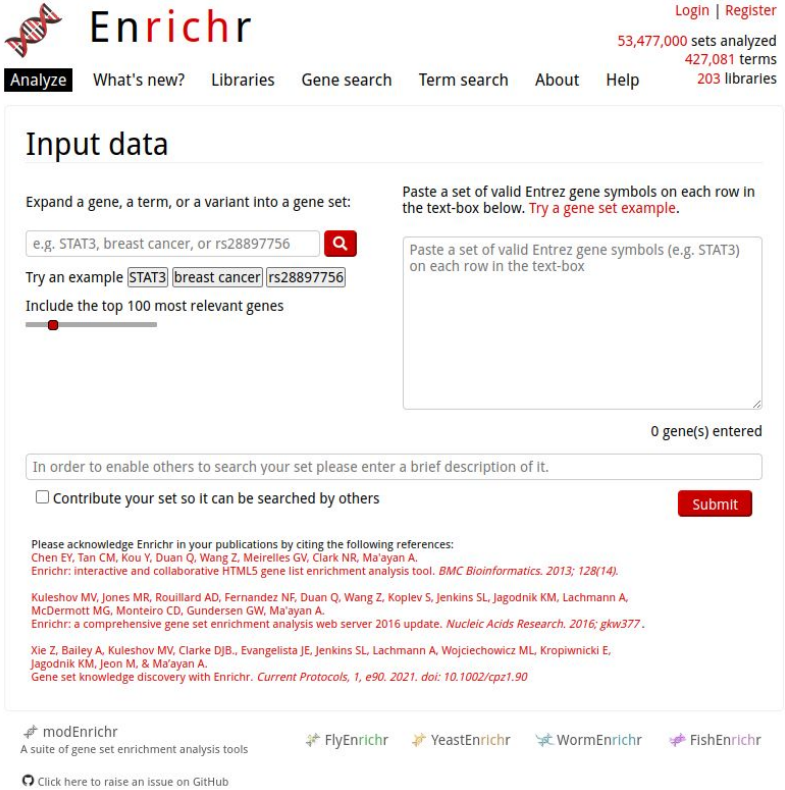


ChIPseeker R package

Functional enrichment analysis



GREAT (web application and R package)



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

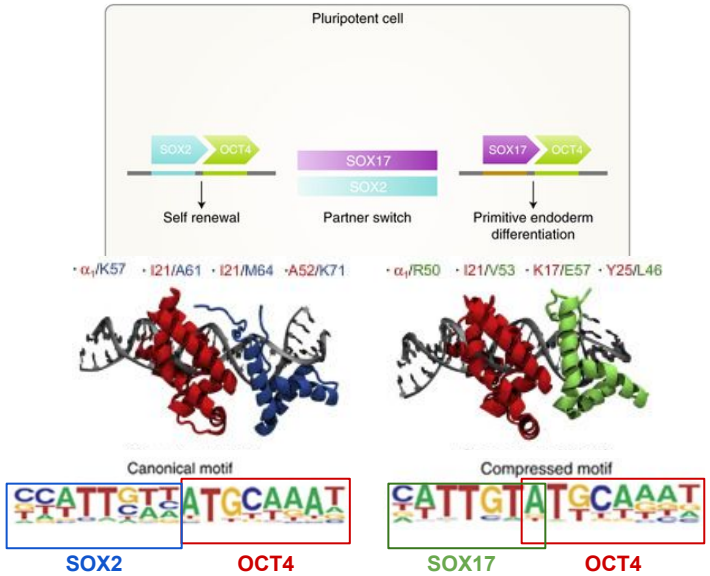
- Simpler methods use different ChIP-seq datasets to infer possible co-binding

  ⇒ a lot of combinations to go through!

- We can look into co-binding through DNA patterns

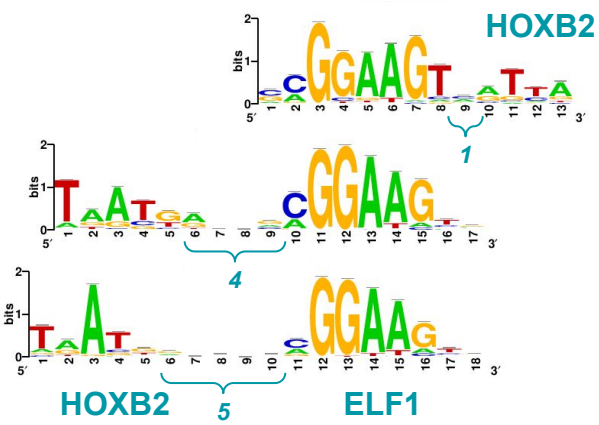
Cooperativity makes regulation very complex

TF **cooperativity** gives rise to TF binding **combinations**



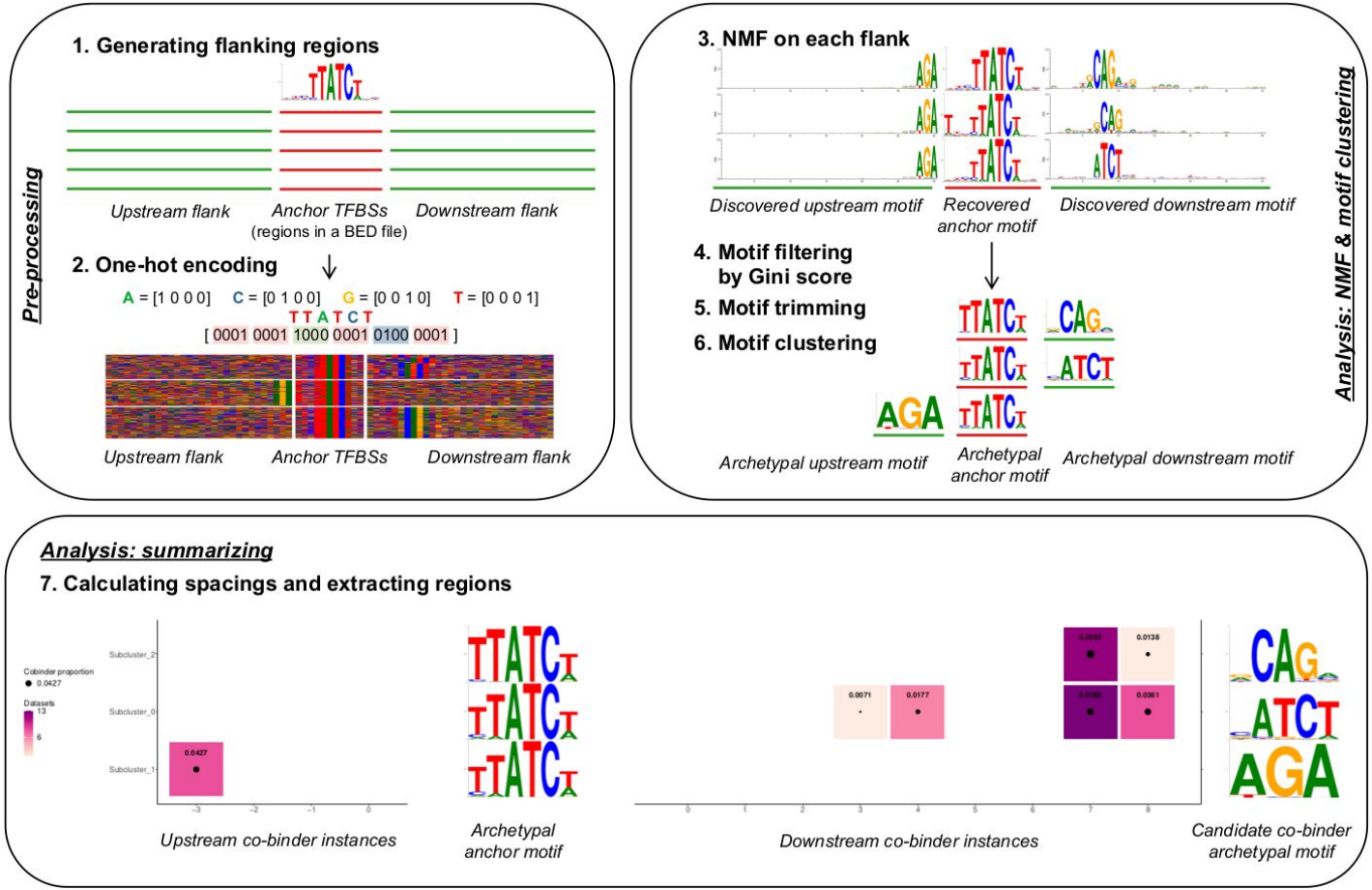
Adapted from Aksoy et al., 2013
and Li & Belmonte, 2018

Different spacings and orientations of HOXB2 and ELF1 [1]

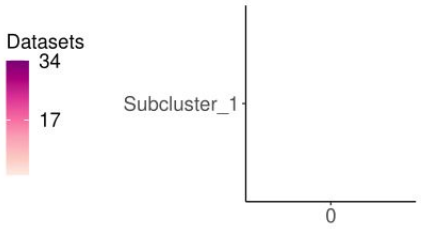


Adapted from Jolma et al., 2016

COBIND

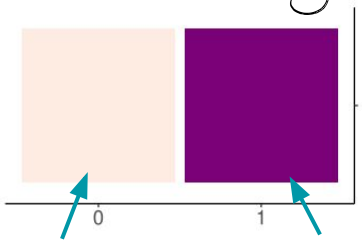

Pre-processing
1. Generating flanking regions
Upstream flank
Anchor TFBSs
(regions in a BED file)
Downstream flank
2. One-hot encoding
A = [1 0 0 0] C = [0 1 0 0] G = [0 0 1 0] T = [0 0 0 1]
T T A T C T
[0001 0001 1000 0001 0100 0001]
Upstream flank
Anchor TFBSs
Downstream flank
Analysis: NMF & motif clustering
3. NMF on each flank
Discovered upstream motif
Recovered anchor motif
Discovered downstream motif
4. Motif filtering by Gini score
5. Motif trimming
6. Motif clustering
Archetypal upstream motif
Archetypal anchor motif
Archetypal downstream motif
Analysis: summarizing
7. Calculating spacings and extracting regions
Cobinder proportion
0.0427
Datasets
13
6
Subcluster_2
Subcluster_0
Subcluster_1
0.0427
Upstream co-binder instances
Archetypal anchor motif
0.0071
0.0177
0.0138
0.0361
Downstream co-binder instances
Candidate co-binder archetypal motif

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



Anchors of different family members

OCT4 for SOX2 anchor
OCT4 for SOX17 anchor



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.

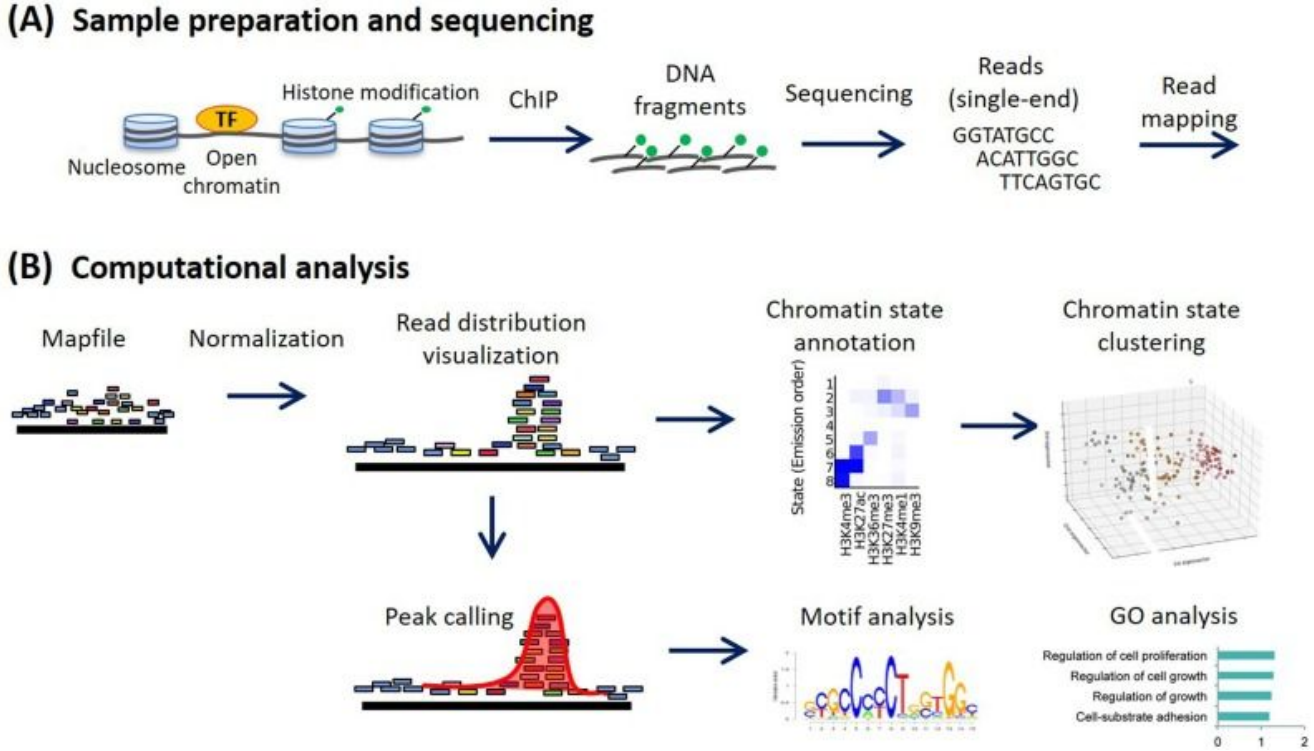
- There are some standard steps that could help guide you into more specific analysis.
  - Visualize your data in the genome browser - this helps you understand the quality.
  - Do *de novo* motif discovery even if you know what motif should be in your peaks.

PART 5

TF binding data resources

What data can we get?

- Raw
- Pre-processed
- Processed



Nakato R and Sakata T. 2021.Methods for ChIP-seq analysis: a practical workflow and advanced applications

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-terminus of the EWS protein with the C-terminal DNA binding domain of an ETS (E26 Transformation Specific) TF family member, most commonly (85-90% of cases), FLI1. The EWS/FLI fusion exhibits the neomorphic 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, 2019) 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'-GCAGCCAATTTACTGGAGCA-3'. The sgETV6-2 forward sequence was 5'-GCAGGGATGACGTAGCCCAG-3'. The sgETV6-3 forward sequence was 5'-GTGTGTGTATAGAGTTTCCA-3'. The sgETV6-4 forward sequence was 5'-GTTATGGTGCACATTATCCA-3'. For control sgRNA, sgChr2.2 was used as a cutting control and targets a gene desert on Chromosome 2, 5'-GGTGTGCGTATGAAGCAGTG-3'. For ligation into the LentiCRISPRv2 plasmid, additional bases were added. 5'-CACCG-3' was added to the beginning of the forward sequence. 5'-AAAC-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., and 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/fl-yu/CUT-RUNTools-2.0/blob/master/docs/INSTALL.md).

Platforms (1) GPL24676 Illumina NovaSeq 6000 (Homo sapiens)

Samples (6) GSM6503406 PEDS0009, sgChr2, FLI1
Less... 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	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

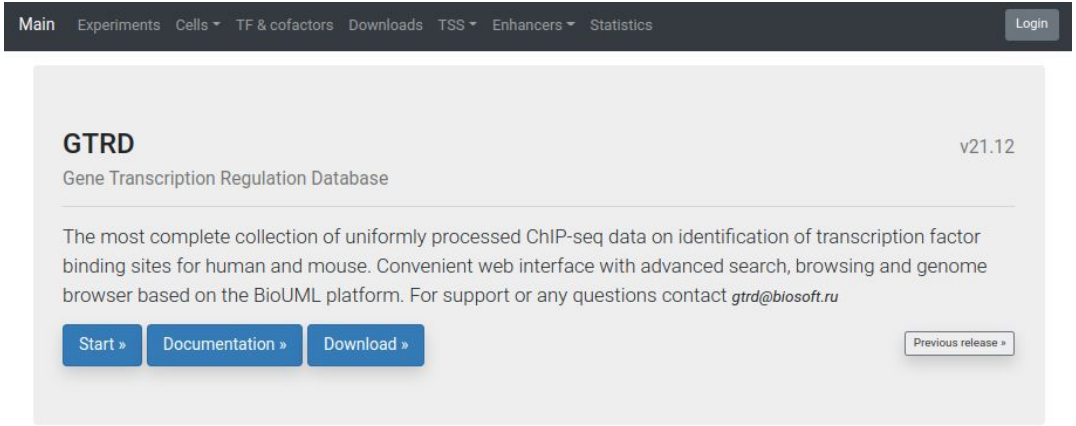
Supplementary file	Size	Download	File type/resource
GSE211852_RAW.tar	1.3 Gb	(http)(custom)	TAR (of BW, NARROWPEAK)

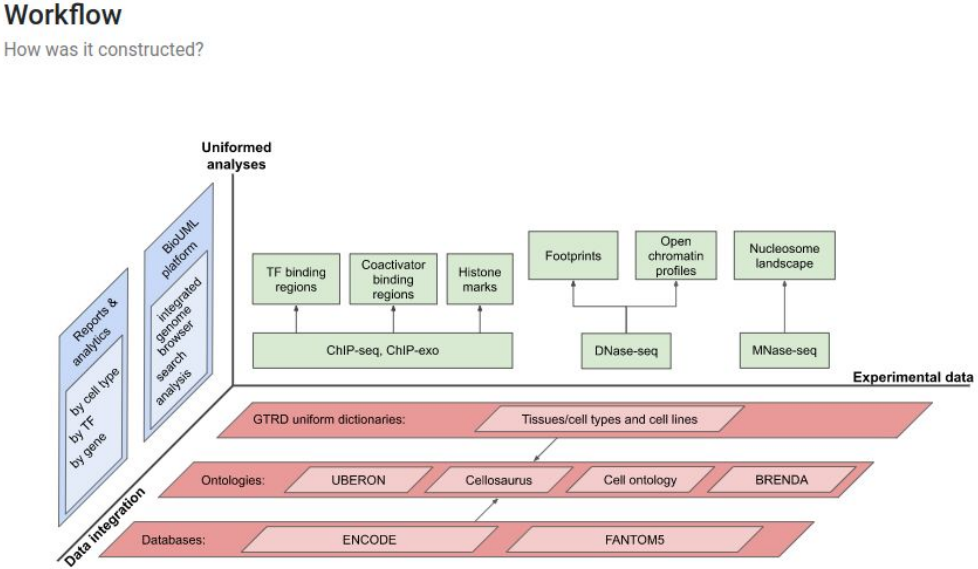
SRA Run Selector

Raw data are available in SRA

Processed data provided as supplementary file

GTRD: gene transcription regulation database





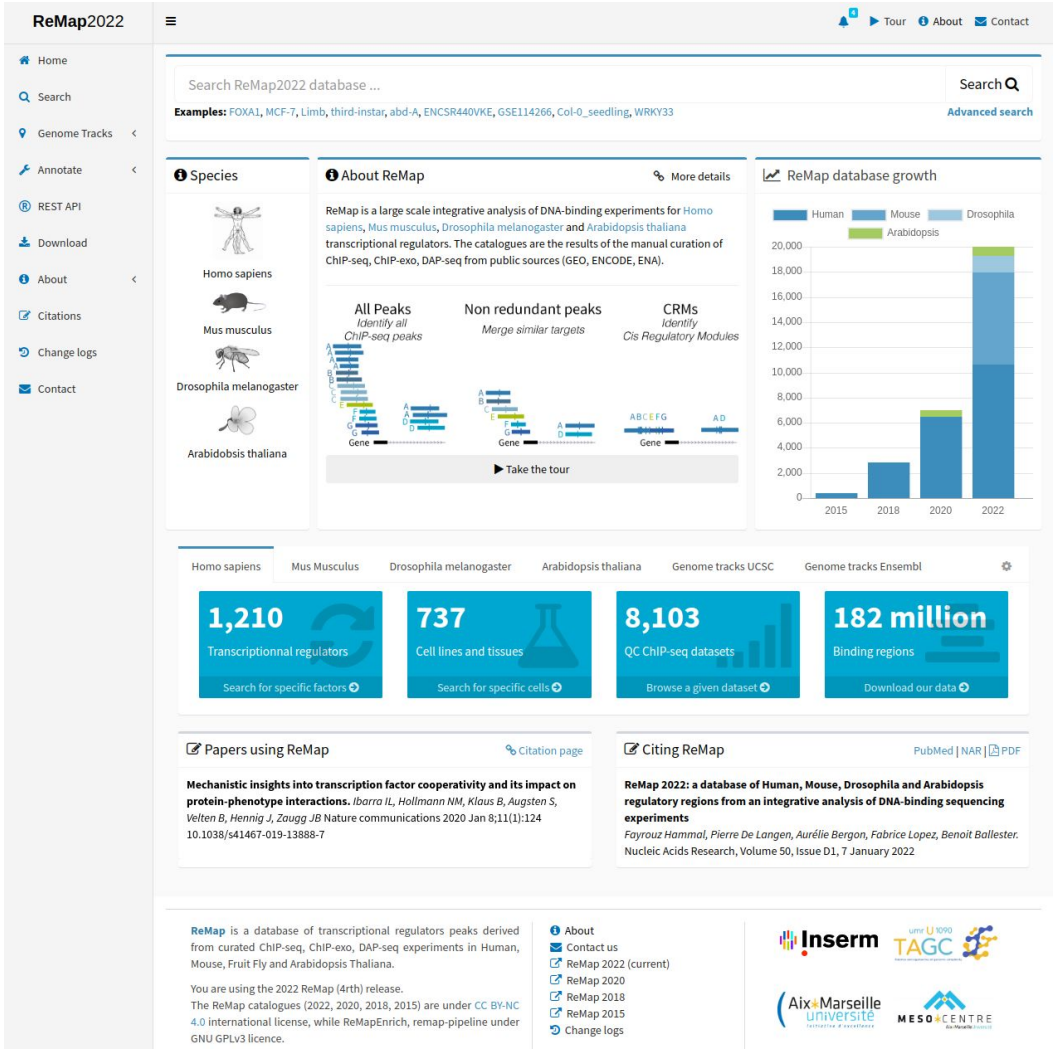
Uniformly processed data

ReMap

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

- Different species

Manually curated ChIP-seq and ChIP-exo data

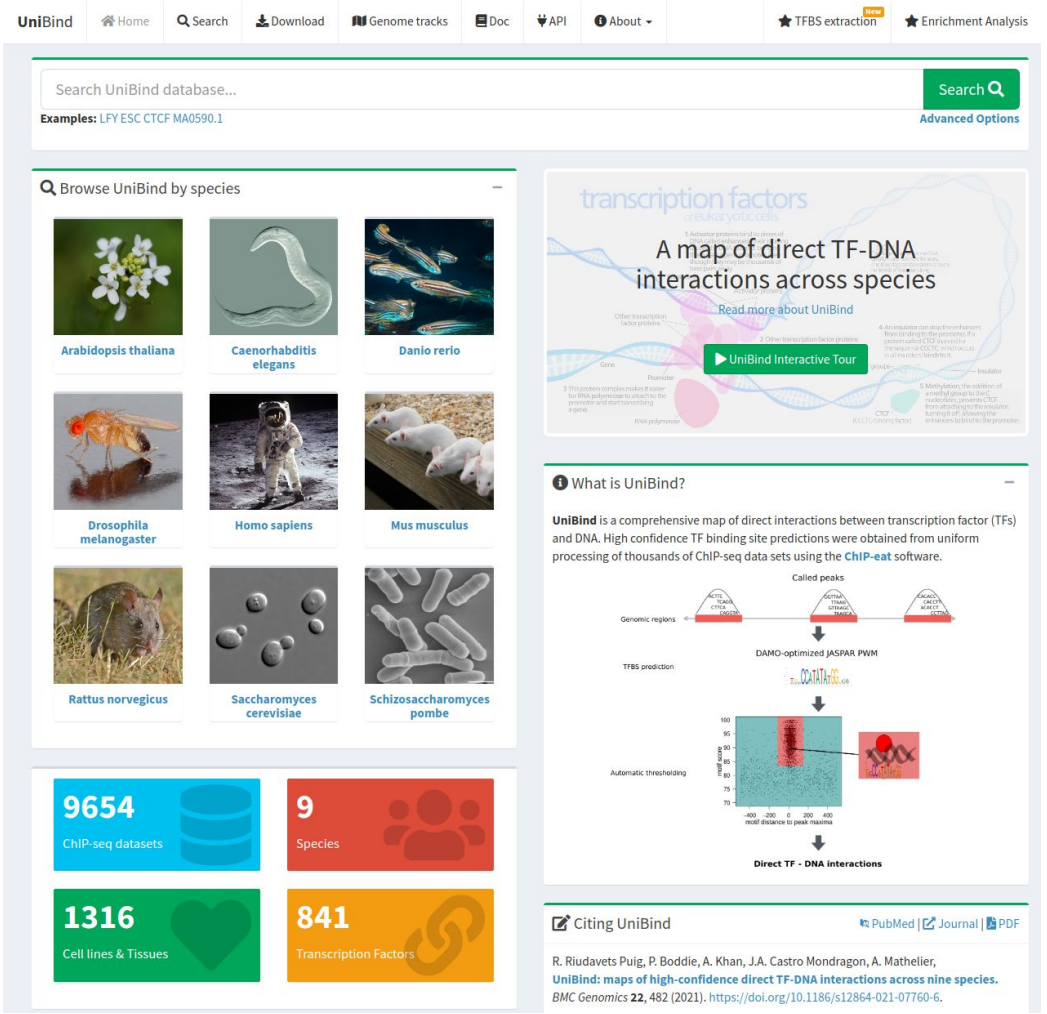


UniBind: database of direct TFBSs

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

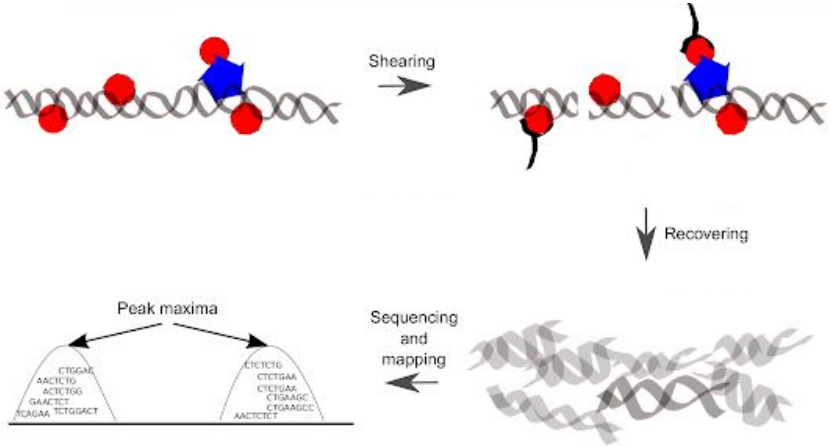
Processed data

Rafael Riudavets Puig



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)

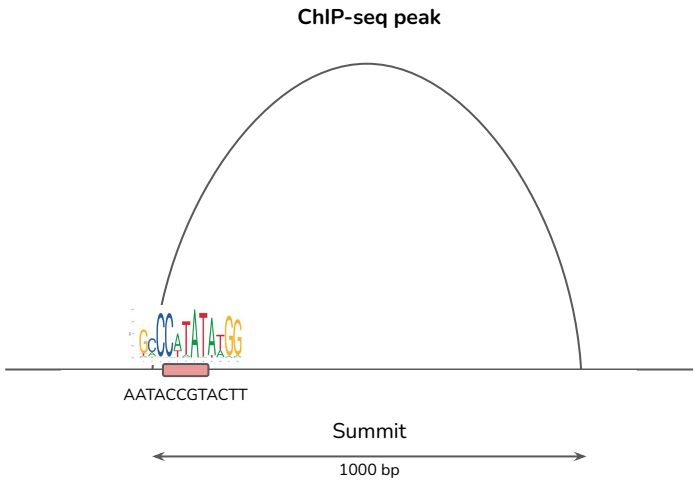
ChIP-eat

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.



TF motif similarity: 20

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

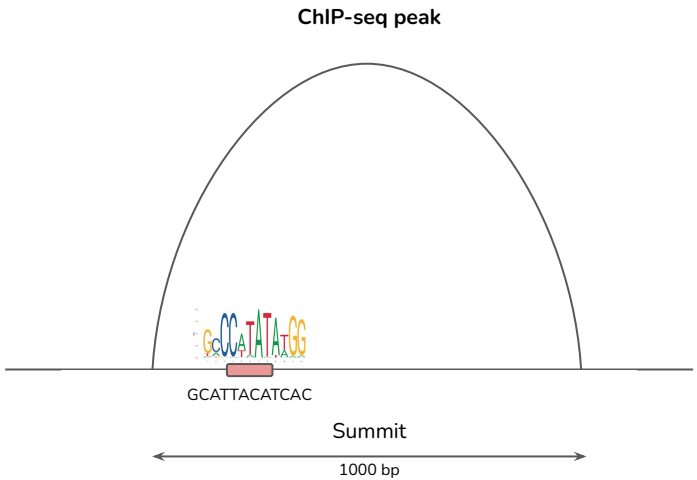
ChIP-eat

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/

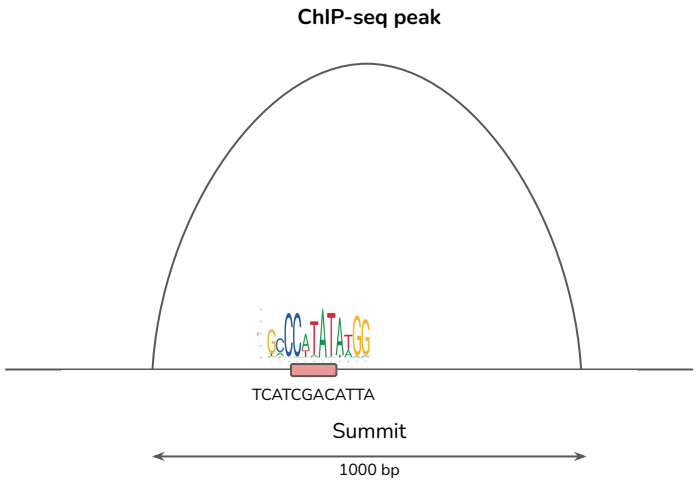
ChIP-eat

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.



TF motif similarity: 37

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

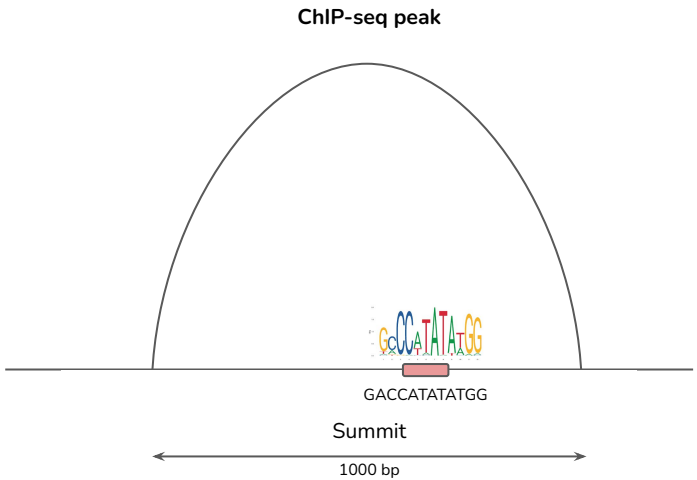
ChIP-eat

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: 99

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

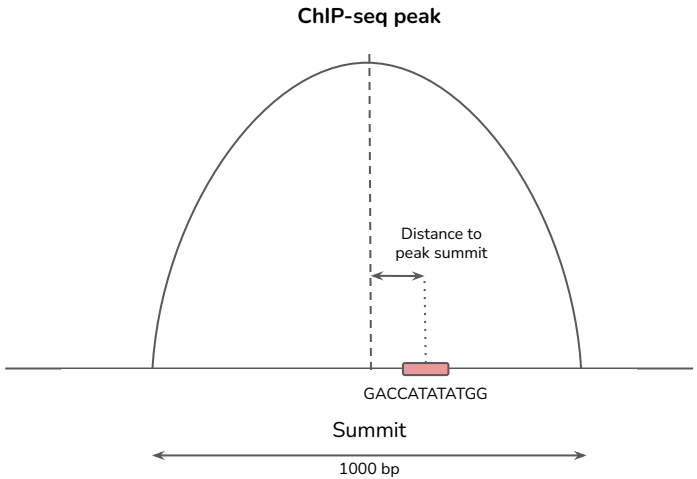
ChIP-eat

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: 99

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

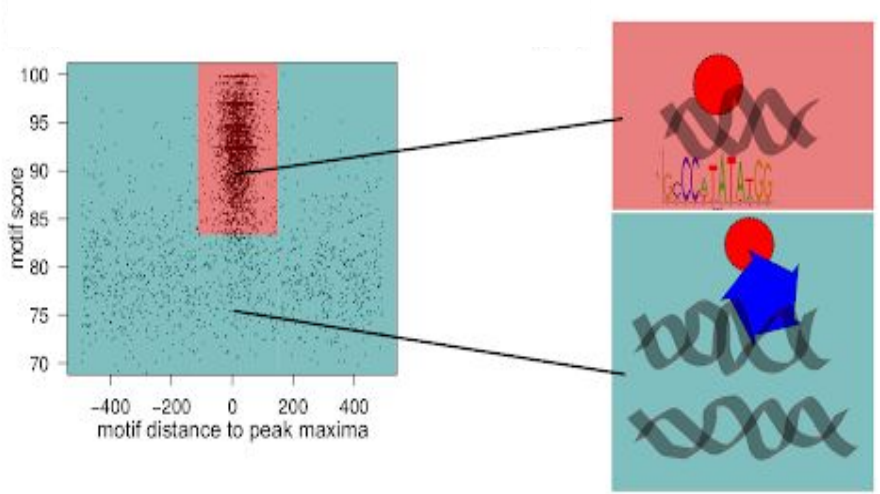
ChIP-eat

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.



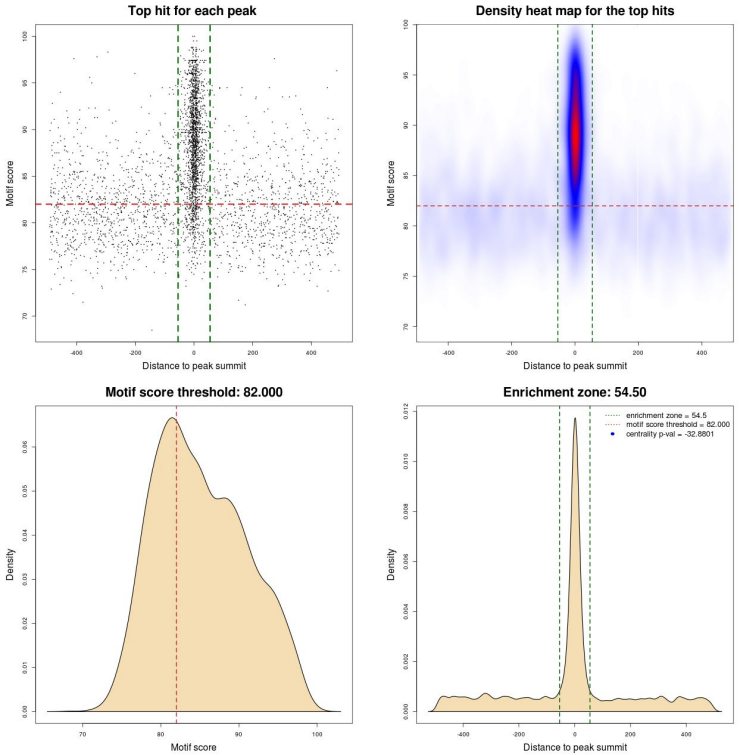
Modified from Mathelier, Shi, and Wasserman (2015)

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:

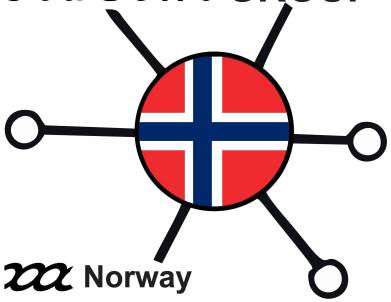
https://forms.gle/XpRV4oDTRPoUCnzM8

The End

NCMM

NORDIC COMPUTATIONAL BIOLOGY

ISCB REGIONAL Student GROUP



rsg-norway.iscbsc.org

twitter.com/RSGNorway

rsg-norway@iscbsc.org

slack

Get Involved: **bit.ly/370yAU0**