

# Ecole Doctorale Sciences de la Vie et de la Santé

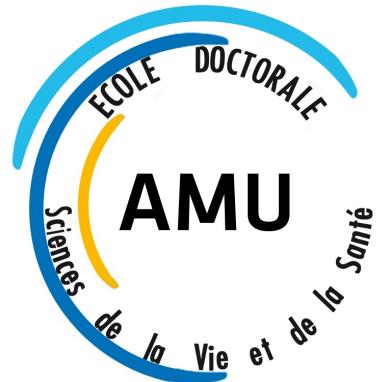
## ***Clustering Transcription Factor Binding Motifs***

CASTRO-MONDAGÓN Jaime Abraham

PhD student 2<sup>nd</sup> Year

Lab. Technological Advances for Genomics and Clinics (TAGC)

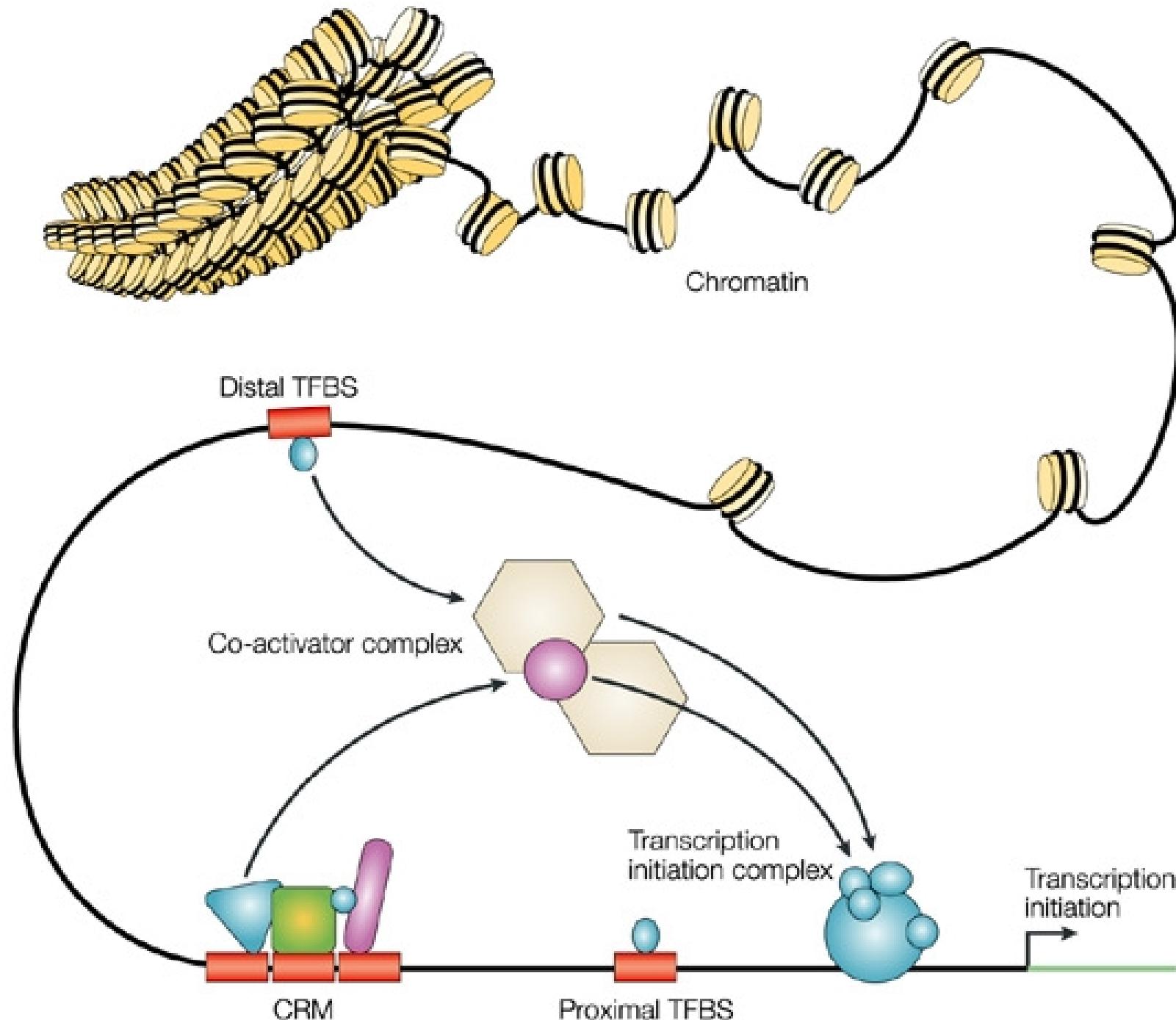
Supervisor: Jacques VAN HELDEN



Aix-Marseille Université

Marseille, France

# *Introduction – Transcriptional regulation*



# *Introduction – TFBS experimental detection*

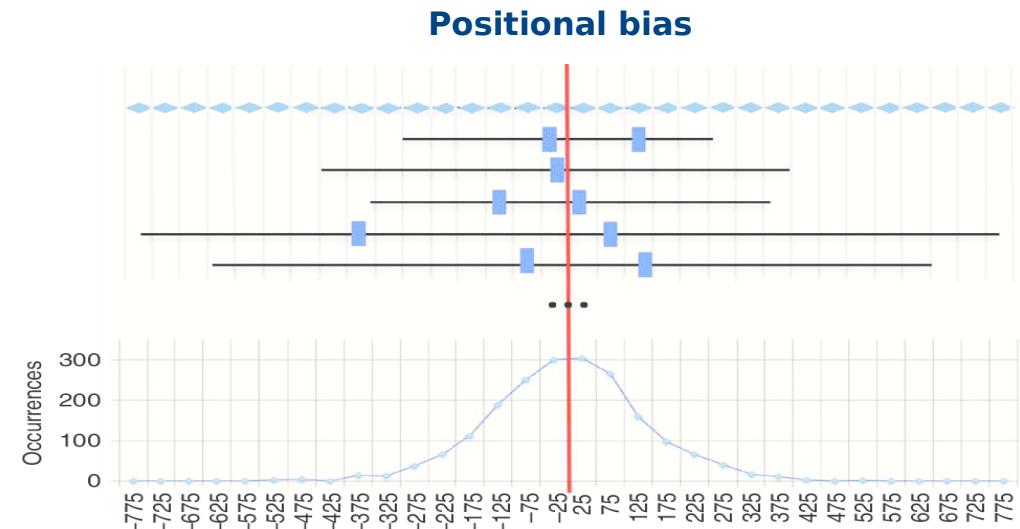
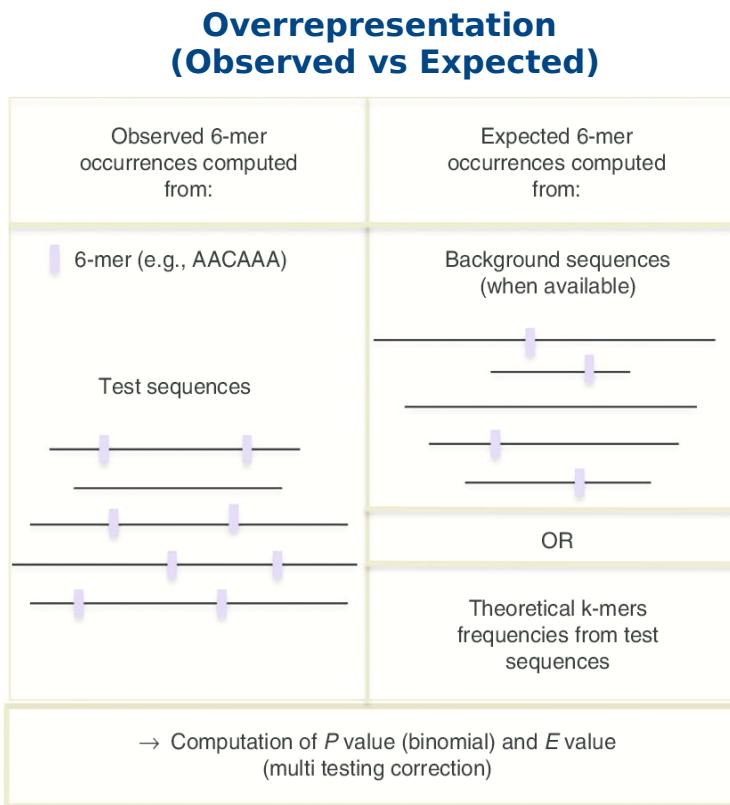
## Experimental Methods

**Table I:** Comparison of *in vitro*- and *in vivo*-based methods to characterize TF binding specificities

Method	Synonyms	Throughput (DNA sequence space)	Materials needed <sup>c</sup>	Data type	Resolution
<i>In vitro</i> approaches					
Selection of target	SELEX, CASTing	>200 000 sites	mg of P	+	Consensus site
Selection of target coupled to NGS	HT-SELEX, Bind-n-Seq	>200 000 sites	mg of P	++	PWM <sup>a</sup> , relative $K_D$
Protein binding microarray	PBM, CSI	up to 1 million sites	mg of P	++	PWM <sup>a</sup> , relative $K_D$
DNA immunoprecipitation	DIP-chip	all genomic sites	$\mu$ g of P	+	PWM <sup>a</sup>
Mechanical trapping	MITOMI	1000 to 100 sites	ng of P	+++(+)	Absolute $K_D$ ( $k_{on}$ , $k_{off}$ )
Gel shift	EMSA	around 10 sites	mg of P	++++	Absolute $K_D$ , $k_{on}$ , $k_{off}$
Surface plasma resonance	BIAcore	up to 100 site	$\mu$ g of P	++++	Absolute $K_D$ , $k_{on}$ , $k_{off}$
<i>In vivo</i> approaches					
ChIP coupled to microarray	ChIP-chip	all genomic sites	ng of D	+	PWM <sup>a,b</sup>
ChIP coupled to NGS	ChIP-seq	all genomic sites	ng of D	+	PWM <sup>a,b</sup>
TF mediated DNA methylation profiling	DamID	all genomic sites	ng of D	+	PWM <sup>a</sup>
Reverse ChIP	PICh	one genomic site	*	–	
DNasel sensitivity profiling coupled to NGS	DNasel-seq	all genomic sites	ng of D	+	PWM <sup>a</sup>

# *Introduction – TFBS computational detection*

## Different approaches



## Different tools



# Introduction – TFBS representation

## Sox2 Binding sites

1	G	C	C	C	T	C	A	T	T	G	T	T	A	T	G	C
2	A	A	A	C	T	C	T	T	T	G	T	T	T	G	A	A
3	T	T	C	A	C	C	A	T	T	G	T	T	C	T	A	G
4	G	A	C	T	C	T	A	T	T	G	T	C	T	C	T	G
5	G	A	T	A	T	C	T	T	T	G	T	T	T	T	T	T
6	T	G	C	A	C	C	T	T	T	G	T	T	A	T	G	C
7	A	A	T	T	C	C	A	T	T	G	T	T	A	T	G	A
8	A	A	A	C	T	C	T	T	T	G	T	T	T	G	G	A
9	A	T	G	G	A	C	A	T	T	G	T	A	A	T	G	C
10	A	G	G	C	C	T	T	T	T	G	T	C	C	T	G	G
11	T	G	T	G	C	T	T	T	T	G	T	N	N	N	N	N
12	C	T	C	A	A	C	T	T	T	G	T	A	A	T	T	T
13	G	C	A	G	C	C	A	T	T	G	T	G	A	T	G	C
14	C	A	C	C	C	T	T	T	T	G	T	T	A	T	G	C
15	T	T	T	T	C	T	A	T	T	G	T	T	T	T	T	A
16	A	A	A	G	G	C	A	T	T	G	T	G	T	T	T	C

## Position Specific Scoring Matrix

<b>A</b>	6	7	4	4	2	0	<b>8</b>	0	0	0	0	0	2	7	0	2	4
<b>C</b>	2	2	6	5	<b>9</b>	<b>12</b>	0	0	0	0	0	0	2	2	2	0	6
<b>G</b>	4	3	2	4	1	0	0	0	0	<b>16</b>	0	2	0	2	<b>9</b>	3	
<b>T</b>	4	4	4	3	4	4	8	<b>16</b>	<b>16</b>	0	<b>16</b>	9	6	<b>11</b>	5	2	

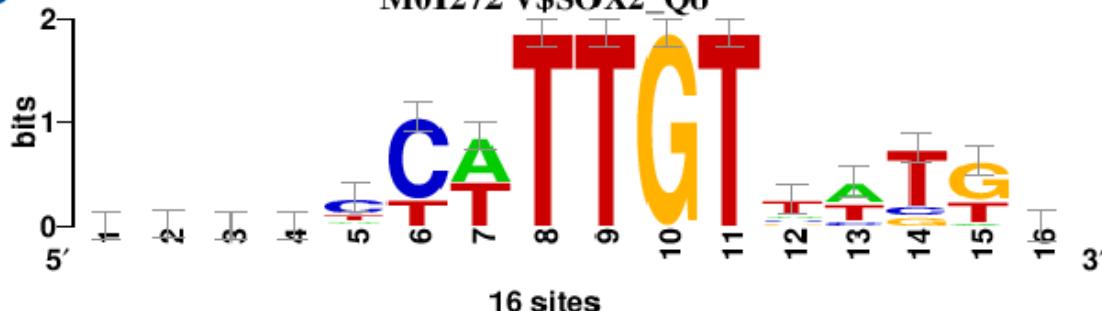
## Consensus

**Y = C or T**  
**W = A or T**

d w h v y Y w T T G T t w T k m

(Logo)

M01272 V\$SOX2\_Q6



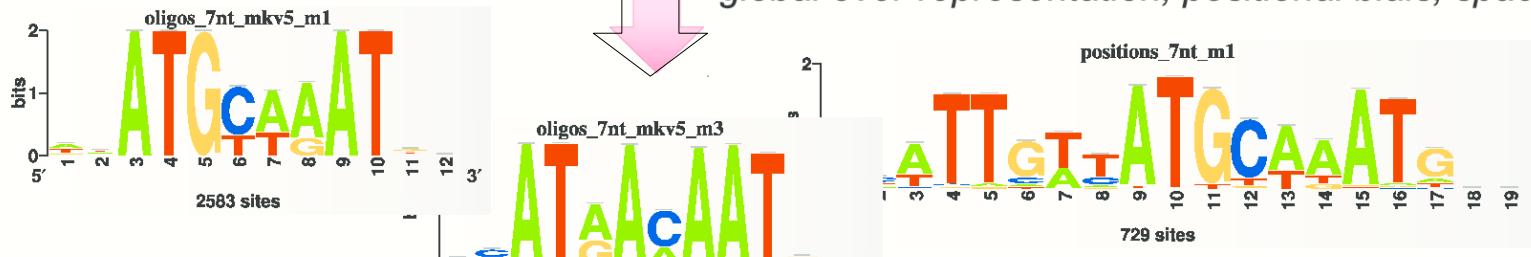
# Introduction – In silico motif discovery

```
>mm9_chr1_3473041_3473370+
ctgtctcttatcttgcttaataaaaggat
ctctttgtattggaaattgggtgttggg
tatatcctgtgcctaattgcatatgga
```

Sequence dataset

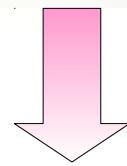


**de novo motif discovery**  
global over-representation, positional bias, spaced motifs



positions\_7nt\_m1

729 sites



**Comparison with collections of motifs**  
various metrics to calculate motif similarity

## DataBases with redundant motifs



## Combine several approaches + tools to discover motifs.

**Pros:** complementarity (motifs found by one program but not by the others).

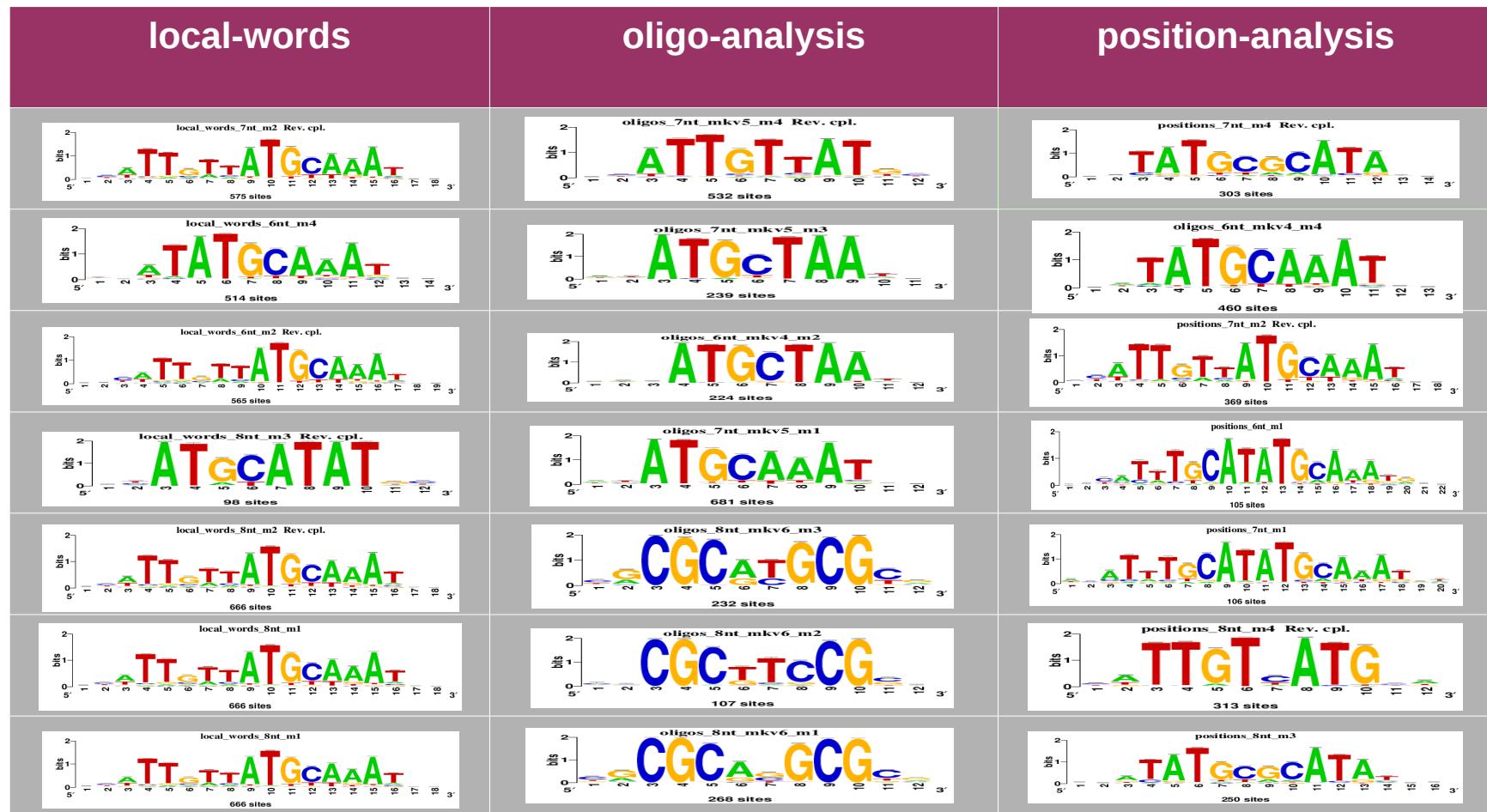
**Cons:** redundancy (same motifs found by several programs).



# Introduction – Where does redundancy come from?

## In Motif discovery algorithms:

- Complementarity between different algorithms.
- One algorithm can be set with different parameters  
(e.g. different k-mer size, different background model, etc).



# Introduction – Where does redundancy come from?

## In Motif Databases:

- 1) TFBMs come from different sources (e.g. Footprint Assays, Selex, ChIP-seq, ...)
- 2) TFBMs can come from data of several species.
- 3) TFs from the same Family use to have conserved structure of DBD.
- 4) Difficulties in annotation.

3				
MA0142.1	Pou5f1::Sox2	10090	Helix-Turn-Helix	Homeo
MA0143.1	Sox2	10090	Other Alpha-Helix	High Mobility Group box (HMG)
MA0143.2	Sox2	10090	Other Alpha-Helix	High Mobility Group box (HMG-box)
MA0143.3	Sox2	10090	Other Alpha-Helix	High Mobility Group box (HMG)
MA0442.1	SOX10	10090,10116,9606	Other Alpha-Helix	High Mobility Group box (HMG)

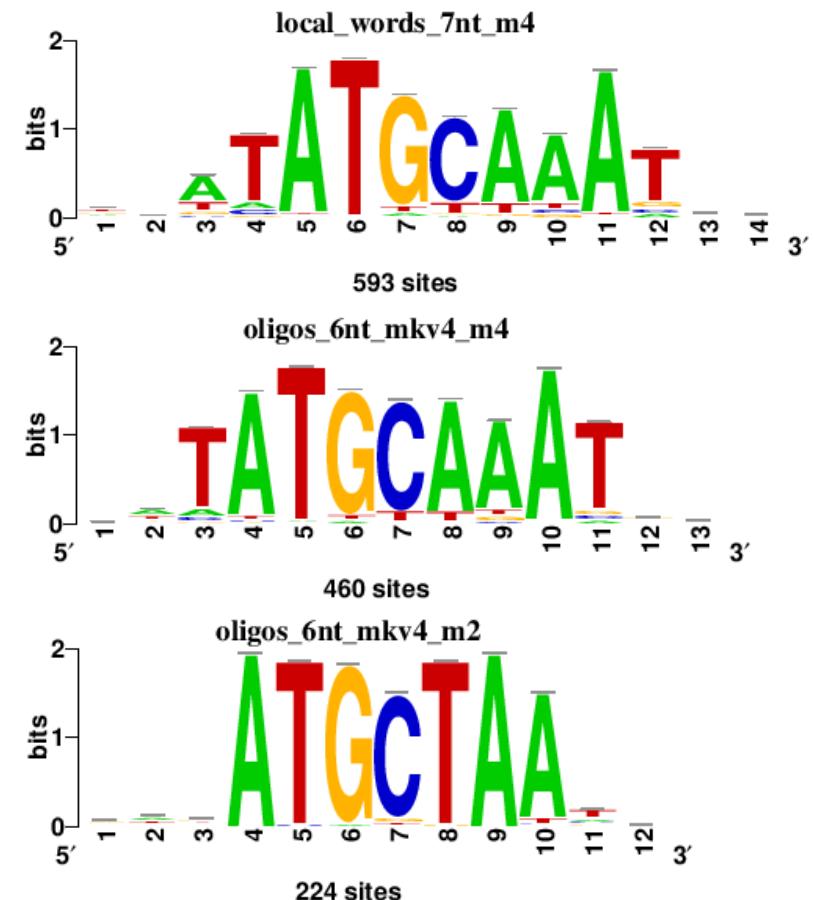
SOX2_DB3_2 (HumanTF 1.0)	SOX2	aTGAATGkyATTCAk	Homo sapiens	Site type: dimeric; SELEX cycle: 4; Possible binding to single +
SOX2_DB3_3 (HumanTF 1.0)	SOX2	rATCAATGkyATTGAt	Homo sapiens	Site type: dimeric; SELEX cycle: 4; Possible binding to single +
SOX2_full_1 (HumanTF 1.0)	SOX2	sAACAAATAaCATTGTTc	Homo sapiens	Site type: dimeric; SELEX cycle: 4; Possible binding to single +
SOX2_full_2 (HumanTF 1.0)	SOX2	hATCAATAmCATTGATm	Homo sapiens	Site type: dimeric; SELEX cycle: 4; Possible binding to single +
SOX2_full_3 (HumanTF 1.0)	SOX2	aTGAATAmCATTCAat	Homo sapiens	Site type: dimeric; SELEX cycle: 4; Possible binding to single +
SOX2_f1 (HOCOMOCO v9)	SOX2	TttgcAtrACAAwRg	Homo sapiens Mus musculus	2
MA0142 (JASPAR 2014)	Pou5f1::Sox2	ywTTswyATGCaaAt	Mus musculus	ChIP-seq
MA0143 (JASPAR 2014)	Sox2	CCwTTGTy	Mus musculus	ChIP-seq; ChIP-seq; ChIP-seq

# Introduction – Motif comparison

Similarity Metric	Formula
Pearson correlation coefficient (PCC)	$PCC(X, Y) = \frac{\sum_{b=A}^T (f_X(b) - \bar{f}_X) \cdot (f_Y(b) - \bar{f}_Y)}{\sqrt{\sum_{b=A}^T (f_X(b) - \bar{f}_X)^2 \cdot \sum_{b=A}^T (f_Y(b) - \bar{f}_Y)^2}}$
Chi-square (pCS) ( $1-p$ -value of)	$\chi^2_3(X, Y) = \sum_{K=\{X,Y\}} \sum_{b=A}^T \frac{(n_K(b) - n_K^e(b))^2}{n_K^e(b)}$
Average Kullback–Leibler (AKL)	$AKL(X, Y) = 10 - \frac{\sum_{b=A}^T f_X(b) \cdot \log \frac{f_Y(b)}{f_X(b)} + \sum_{b=A}^T f_Y(b) \cdot \log \frac{f_Y(b)}{f_X(b)}}{2}$
Sum of squared distances (SSD)	$SSD(X, Y) = 2 - \sum_{b=A}^T (f_X(b) - f_Y(b))^2$
Average log-likelihood ratio (ALLR)	$ALLR(X, Y) = \frac{\sum_{b=A}^T n_X(b) \cdot \log \frac{f_Y(b)}{p_{ref}(b)} + \sum_{b=A}^T n_Y(b) \cdot \log \frac{f_X(b)}{p_{ref}(b)}}{\sum_{b=A}^T (n_X(b) + n_Y(b))}$
ALLR with lower limit (ALLR_LL)	Same as above, but a lower limit of $-2$ is imposed on the score (see text)

## How similar are these motifs?

- Different width
- Different orientation
- Different number of sites
- Differences in nucleotide frequencies.
- Different information content



# **Introduction – Motif comparison : a battlefield**

- 25 papers published (since 2004) with new methods to measure motif similarity.
- Some metrics works well with monomers, dimer, spaced motifs, high information content columns, etc.
- Each publication says that its method outperforms the others!
- **Conclusion:** No a standard metric to measure motif similarity. A combination of them could help to measure correctly the similarity.

## **FISim: A new similarity measure between transcription factor binding sites based on the fuzzy integral**

Fernando Garcia\*, Francisco J Lopez, Carlos Cano and Armando Blanco

## **A Novel Alignment-Free Method for Comparing Transcription Factor Binding Site Motifs**

Minli Xu\*, Zhengchang Su\*

Jaccard index based similarity measure to compare transcription factor binding site models

Ilya E Vorontsov<sup>2,3†</sup>, Ivan V Kulakovskiy<sup>1,2\*†</sup> and Vsevolod J Makeev<sup>1,2,4</sup>

## **Quantifying similarity between motifs**

Shobhit Gupta\*, John A Stamatoyannopoulos\*, Timothy L Bailey<sup>†</sup> and William Stafford Noble<sup>\*‡</sup>

## **A Discriminative Approach for Unsupervised Clustering of DNA Sequence Motifs**

Philip Stegmaier<sup>1,2\*</sup>, Alexander Kel<sup>2</sup>, Edgar Wingender<sup>2,3</sup>, Jürgen Borlak<sup>4</sup>

## **Natural similarity measures between position frequency matrices with an application to clustering**

Utz J. Pape<sup>1,2,\*</sup>, Sven Rahmann<sup>3,4</sup> and Martin Vingron<sup>1</sup>

**SPIC: A novel similarity metric for comparing transcription factor binding site motifs based on information contents**

Shaoqiang Zhang<sup>1\*</sup>, Xiguo Zhou<sup>1</sup>, Chuanbin Du<sup>2</sup>, Zhengchang Su<sup>2\*</sup>

## **Alignment-free clustering of transcription factor binding motifs using a genetic-k-medoids approach**

Pilib Ó Broin<sup>1,2</sup>, Terry J Smith<sup>1</sup> and Aaron AJ Golden<sup>1,3\*</sup>

# *Motivation – From comparison to clustering*

- The discovered motifs are compared against several motif databases.
- The motif comparison results can be used to cluster the similar motifs.
- The clustering of motifs would allow to:
  - a) Filter out the redundant motifs.
  - b) Classify the motifs (TF families)
- Few programs are available to cluster motifs (stamp<sup>1</sup>, m2match<sup>2</sup>, matlign<sup>3</sup>, GMACS<sup>4</sup>)

Limitations:

- Low number of input motifs.
- One (or few) metrics can be used.
- No visualization of the clustered TFBMs.
- Only one dataset of TFBMs can be grouped.

1.- Mahony S and Benos P. (2007) **STAMP: a web tool for exploring DNA-binding motif similarities.** NAR.

2.- Kankainen M and Löytynoja A. (2007) **MATLIGN: a motif clustering, comparison and matching tool.** BMC Bioinformatics.

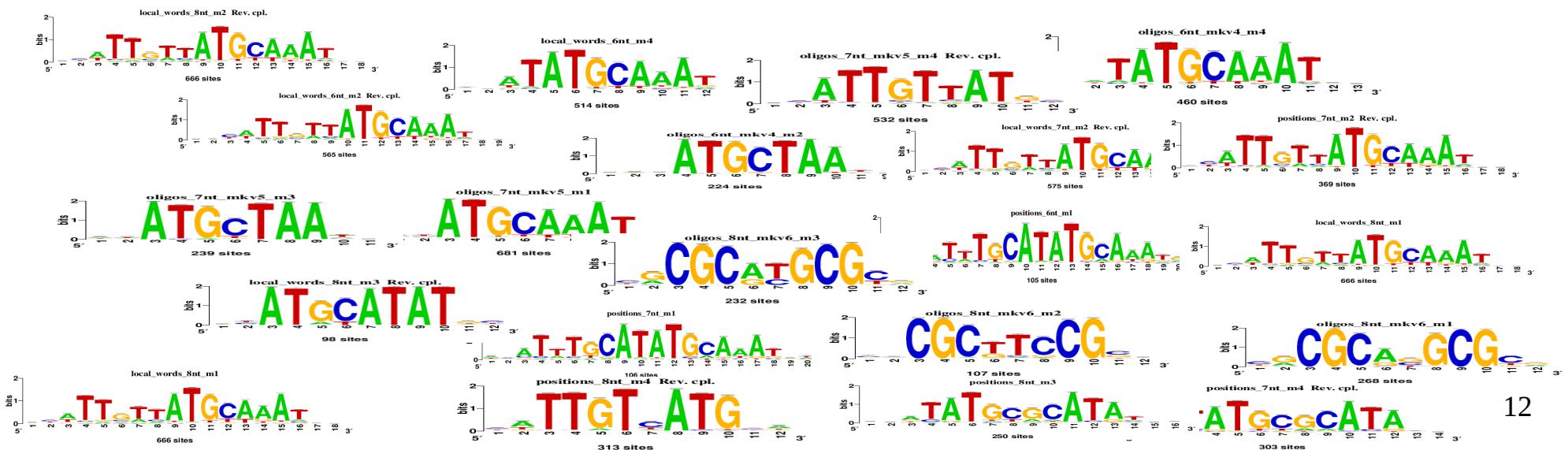
3.- Stegmaier P et al. (2013) **A Discriminative Approach for Unsupervised Clustering of DNA Sequence Motifs.** PloS Computational Biology.

4.- BroinP, SmithJY and Golden AJ. (2015) **Alignment-free clustering of transcription factor binding motifs using a genetic-k-medoids approach.** BMC Bioinformatics.

# Motivation – Clustering of Motifs

## Questions:

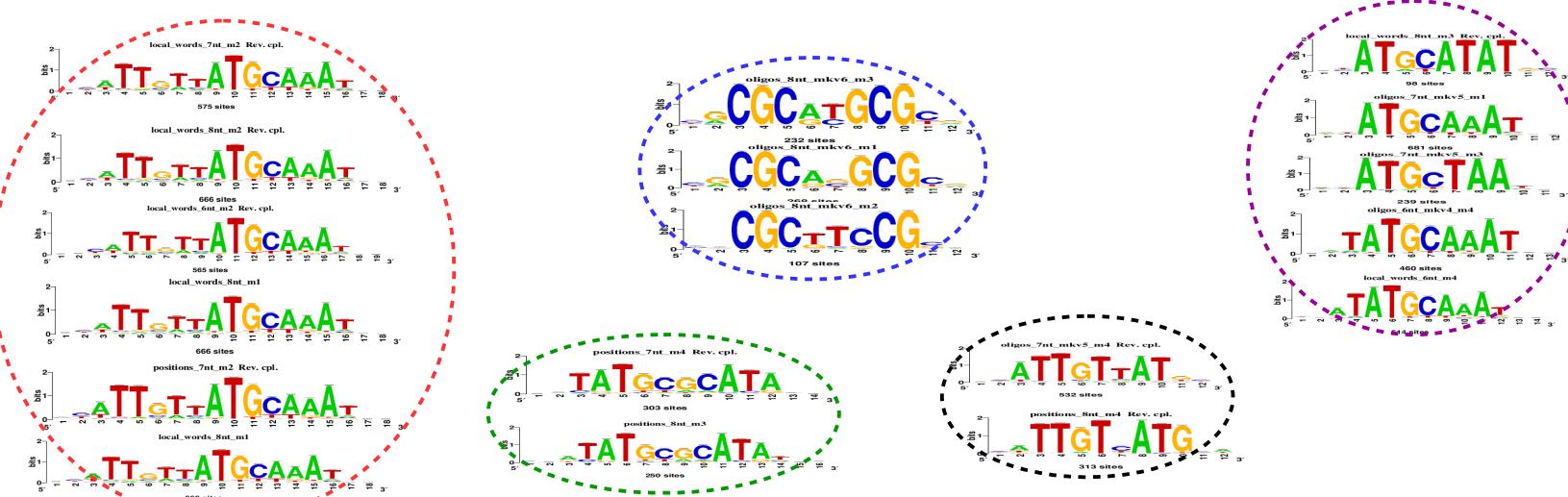
- Which metric use to compare/cluster the motifs ?
- Can we group the similar motifs ?
- Does the clustering reflect binding properties of TFs ?
- Does the grouped TFBMs correspond to the same Family ?
- Is it possible to create a non-redundant collection of motifs ?
- Does the public collections of motifs encompass the private ones ?



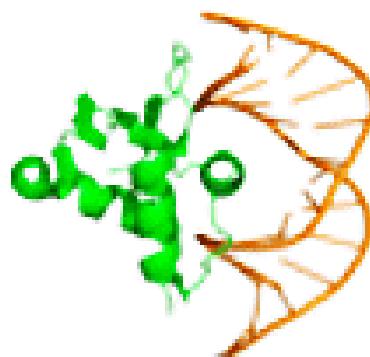
# Motivation – Clustering of Motifs

## Questions:

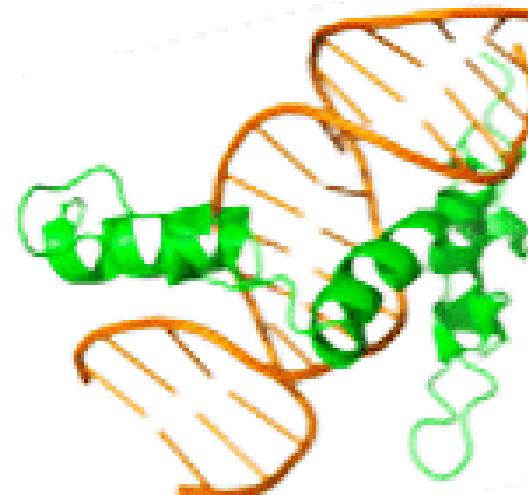
- Which metric use to compare/cluster the motifs ?
- Can we group the similar motifs ?
- Does the clustering reflect binding properties of TFs ?
- Does the grouped TFBMs correspond to the same Family ?
- Is it possible to create a non-redundant collection of motifs ?
- Does the public collections of motifs encompass the private ones ?



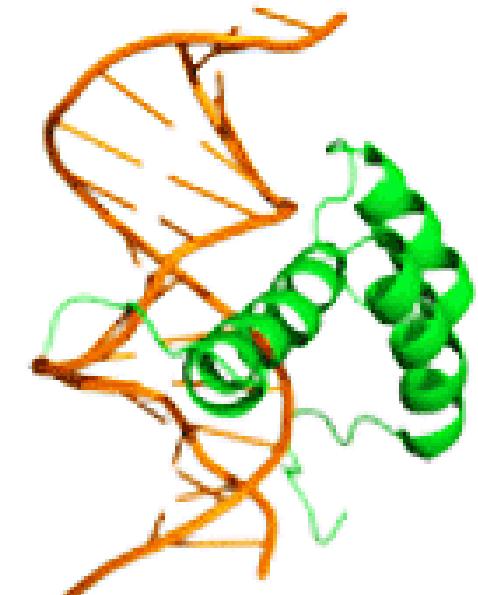
# *Introduction – Identification of DBD from TFBMs*



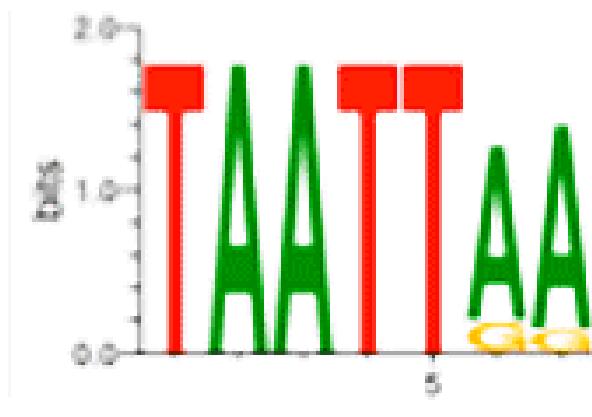
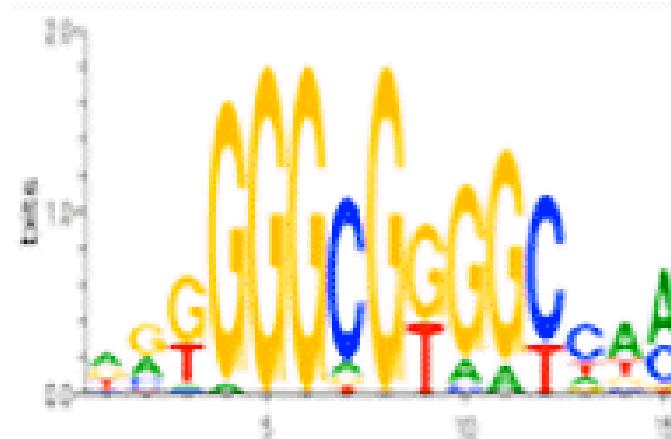
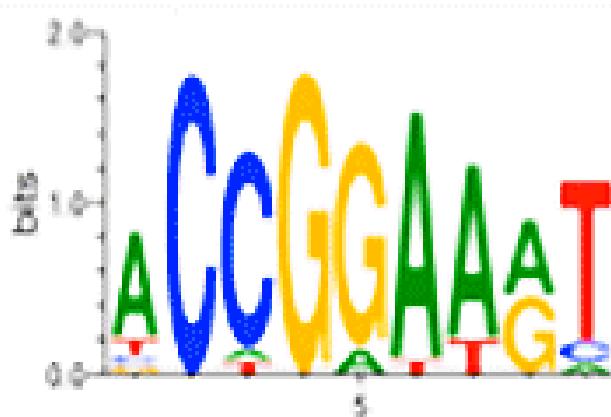
ETS domain



Zinc Fingers of TFIIIA



Homeodomain



# *Objectives*

- Create a tool to identify and visualize groups of similarities among a set of input TFBMs.

## **Specific goals:**

- (1)** Partitioning the input set of TFBMs into distinct clusters.
- (2)** Different TFBMs format should be used as input.
- (3)** Cluster more than one collection of motifs.
- (4)** Align the TFBMs to highlight common/different positions.
- (5)** Provide intuitive and dynamic visualization of motifs and their relationships.

## **Applications:**

- (1)** Simplify the interpretation of motif discovery results.
- (2)** Identify groups of similar motifs (TF families).
- (3)** Create non-redundant collections of TFBMs.

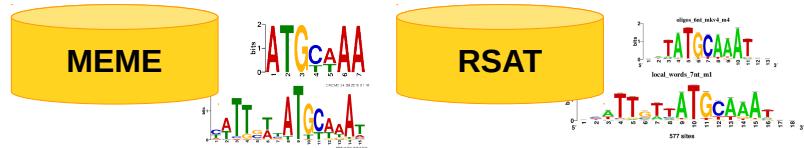
# *matrix-clustering*

- *matrix-clustering* : dynamic visualization of clusters of TFBMs in RSAT<sup>1</sup>.
  - 1) Segmentation of the input set of TFBMs into separated trees (forest) rather than a single tree.
  - 2) Displays the cluster of motifs as logo alignments.
  - 3) Generate Branch-wise motifs (also known as Familial Binding Profiles)
  - 4) Supports for a large series of metrics to measure TFBM (dis)similarities.
  - 5) Possibility to set a custom combination of these metrics to compute an integrative threshold to separate the TFBMs.
  - 6) Multiple and dynamic representation of clusters (logo alignment, heatmaps).
  - 7) Possibility to cluster two or more collection of TFBMs.
- Available in RSAT website (<http://www.rsat.eu/>)

# *Matrix-clustering - algorithm*

## Collection of Motifs

Motifs from several motif discovery tools



Motifs from different experiments/conditions



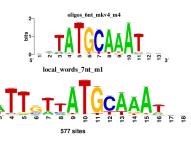
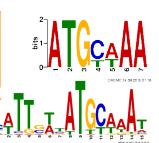
Motif Databases

(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

# *Matrix-clustering - algorithm*

## Collection of Motifs

Motifs from several motif discovery tools



Motifs from different experiments/conditions

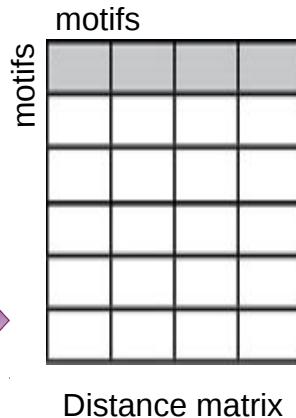


Motif Databases

(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

## Comparison of all motifs

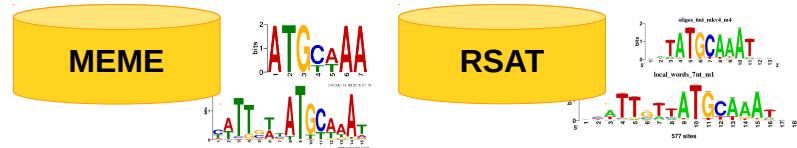
Various metrics to calculate motif similarity



# Matrix-clustering - algorithm

## Collection of Motifs

Motifs from several motif discovery tools



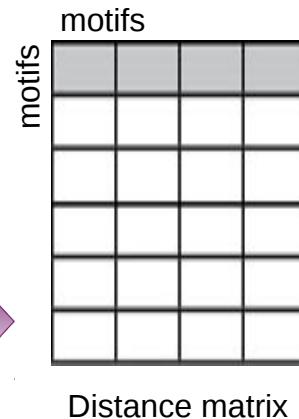
Motifs from different experiments/conditions



Motif Databases  
(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

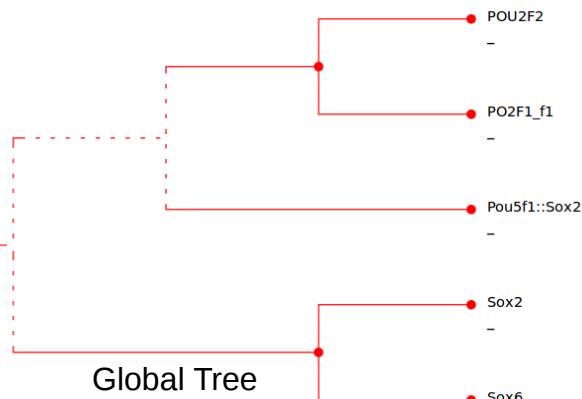
## Comparison of all motifs

Various metrics to calculate motif similarity



## Hierarchical Clustering

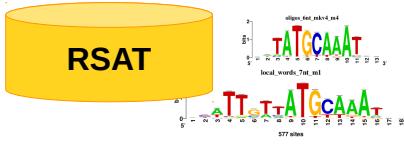
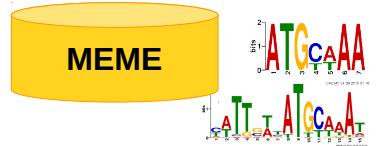
One similarity metric + linkage rule



# Matrix-clustering - algorithm

## Collection of Motifs

Motifs from several motif discovery tools



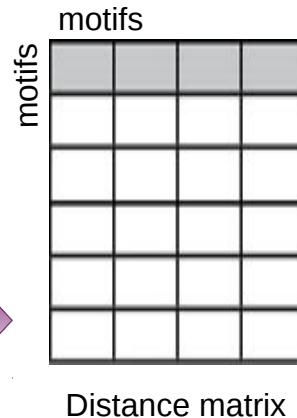
Motifs from different experiments/conditions



Motif Databases  
(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

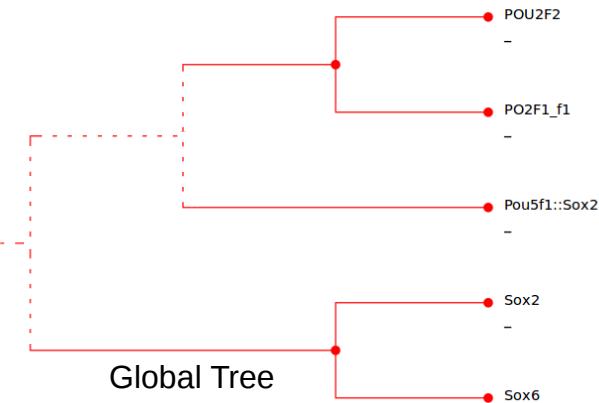
## Comparison of all motifs

Various metrics to calculate motif similarity



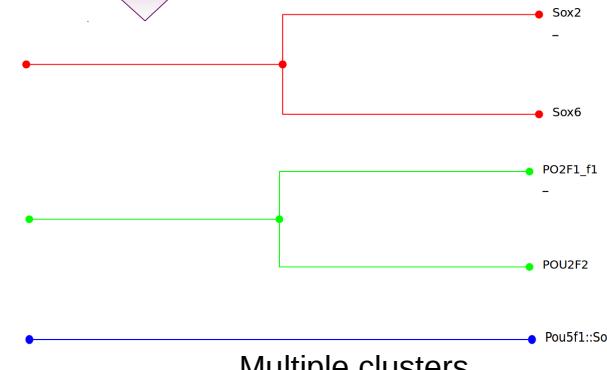
## Hierarchical Clustering

One similarity metric + linkage rule



## Partitioning

Integrative threshold with multiple metrics

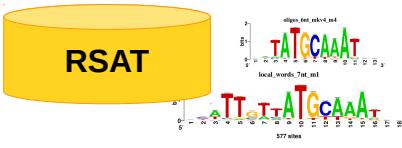
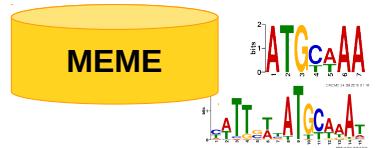


Multiple clusters

# Matrix-clustering - algorithm

## Collection of Motifs

Motifs from several motif discovery tools



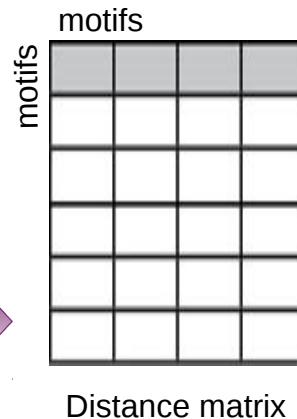
Motifs from different experiments/conditions



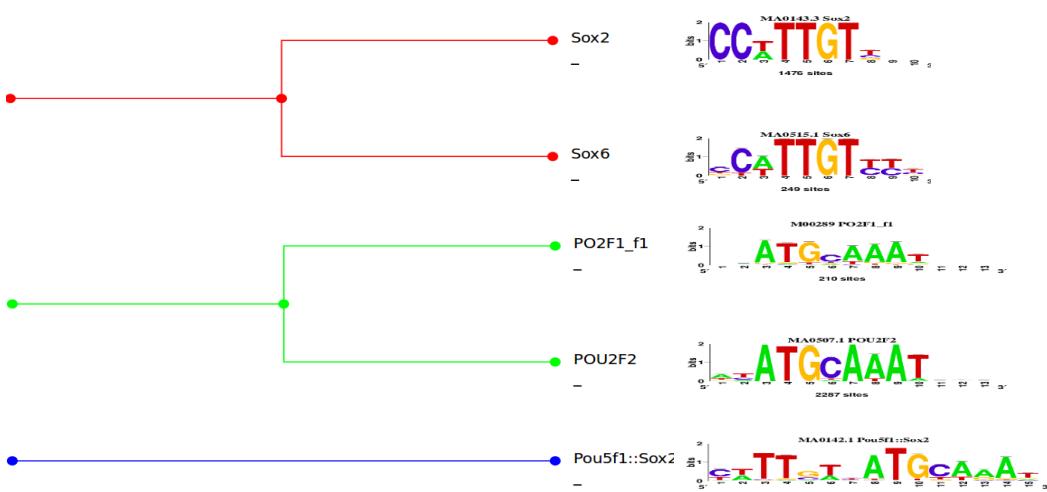
Motif Databases  
(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

## Comparison of all motifs

Various metrics to calculate motif similarity

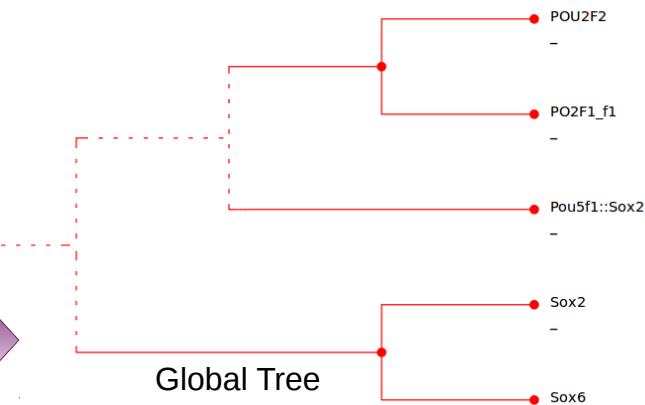


## Alignment



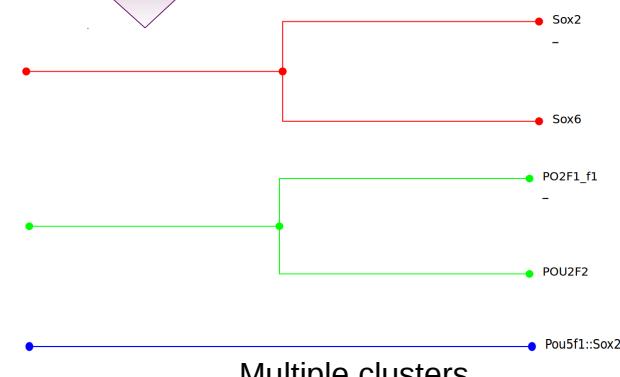
## Hierarchical Clustering

One similarity metric + linkage rule



## Partitioning

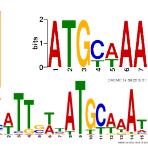
Integrative threshold with multiple metrics



# Matrix-clustering - algorithm

## Collection of Motifs

Motifs from several motif discovery tools



Motifs from different experiments/conditions

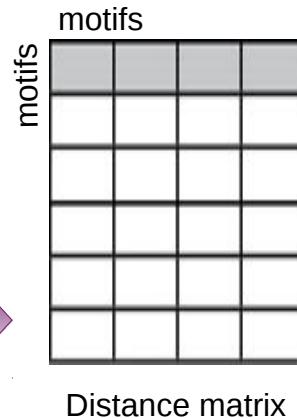


Motif Databases

(Jaspar, Transfac, CisBP, Hocomoco, FootprintDB)

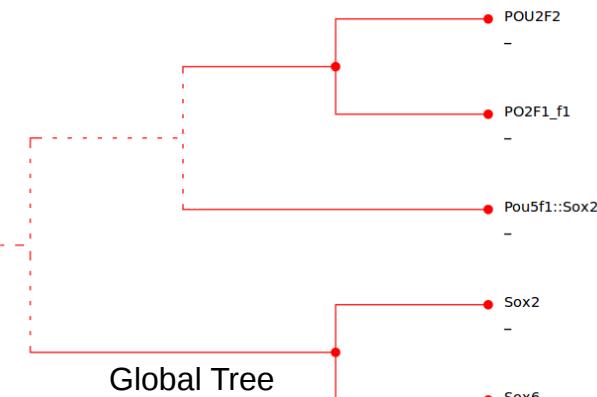
## Comparison of all motifs

Various metrics to calculate motif similarity



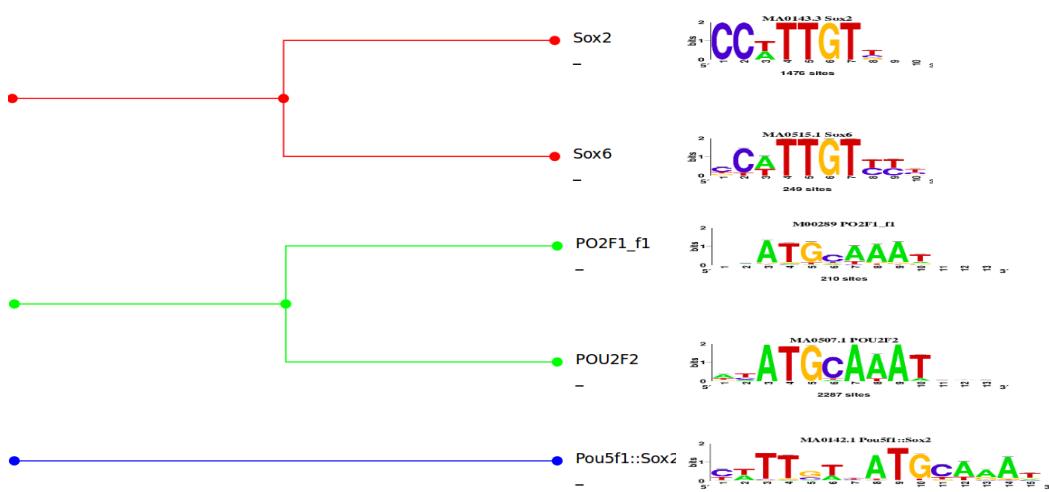
## Hierarchical Clustering

One similarity metric + linkage rule

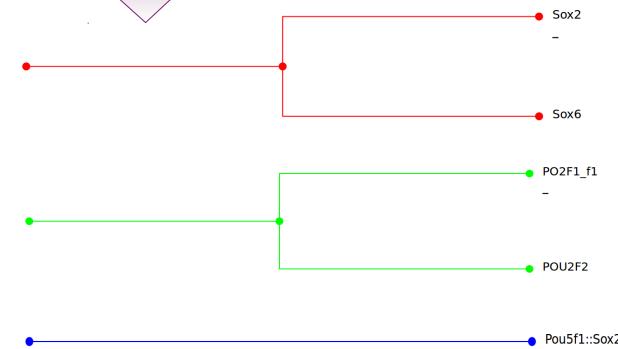


## Partitioning

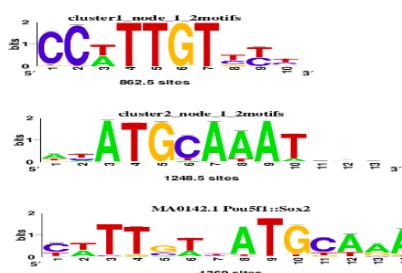
Integrative threshold with multiple metrics



## Alignment



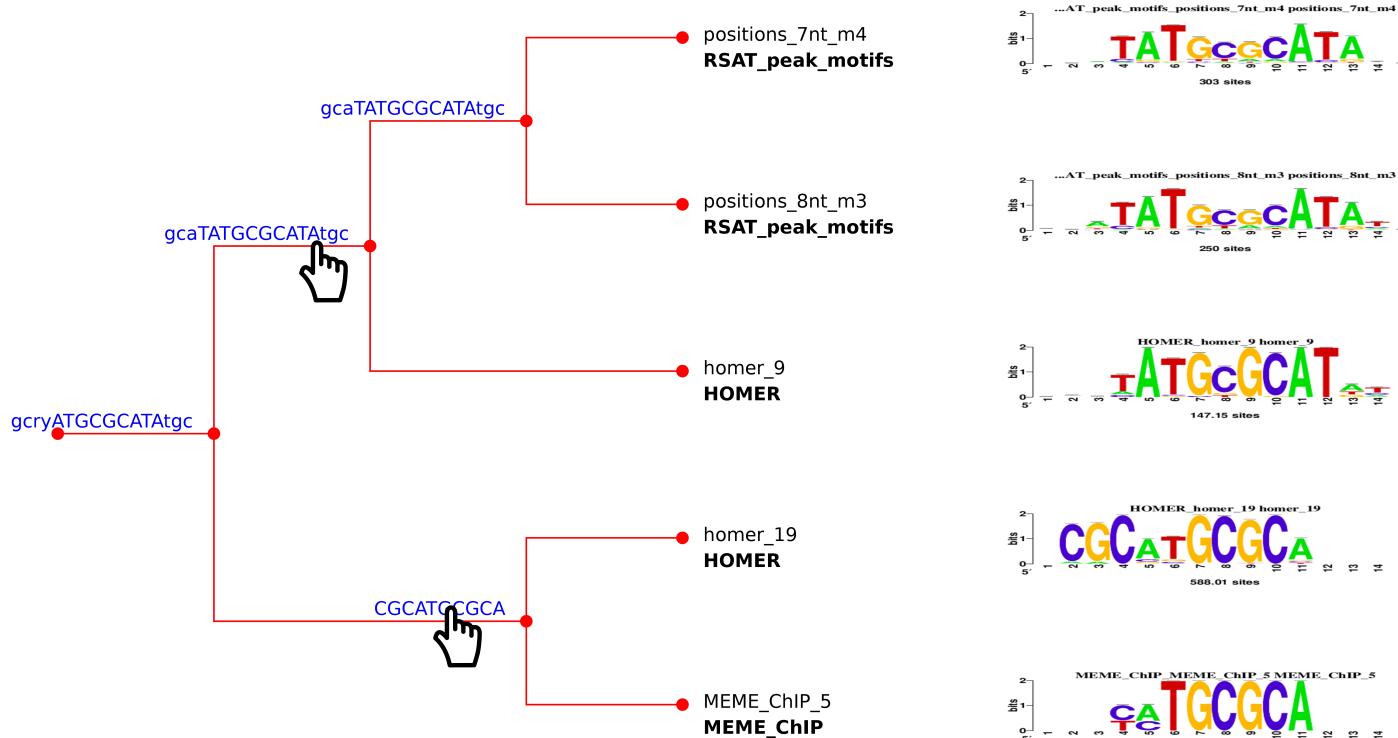
Non-redundant Motifs 22



22

# Matrix-clustering : output example

## Expanded Tree



The **Familial Binding Profiles** are calculated at each branch.

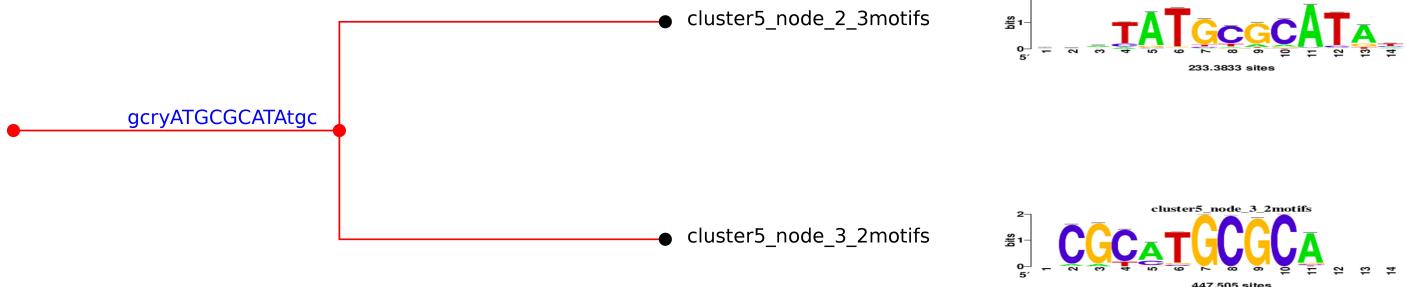
The '**collection**' tag indicates the provenance of the motif.

Click on each branch to collapse it



The hierarchical tree is collapsed/expanded dynamically.

## Collapsed Tree

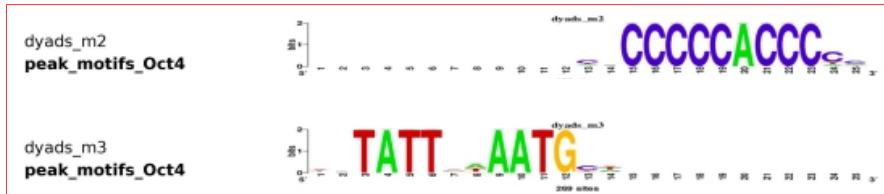


**Non-Redundant Motifs:** they can be selected manually by the user.

# *Choosing a metric - Drawbacks*

- Spurious alignments (alignment of non-informative columns)

**cor >= 0.8**



Cluster 1

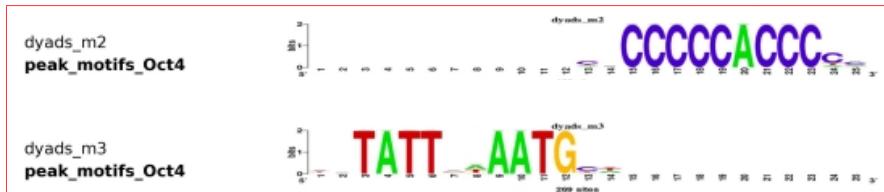


Cluster 2

# *Choosing a metric - Drawbacks*

- Spurious alignments (alignment of non-informative columns)

**cor >= 0.8**



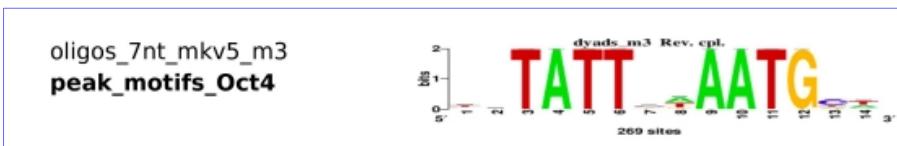
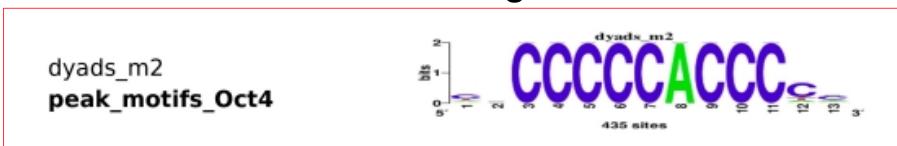
## Cluster 1



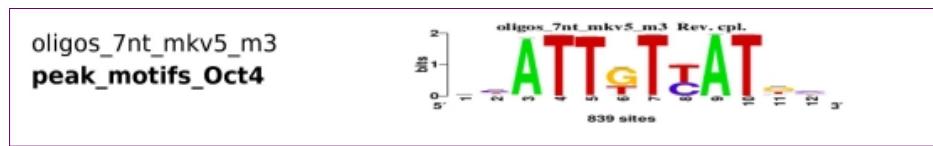
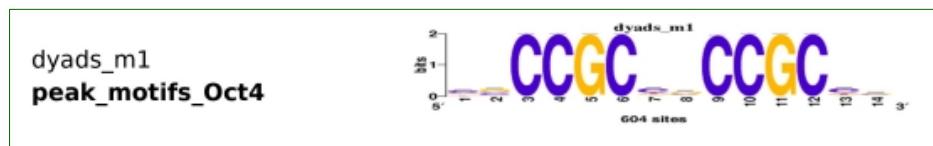
## Cluster 2

- Including the width of the aligned segment avoid the spurious alignments. But the motifs are not aligned.

**cor >= 0.8 + w >= 5**



## Cluster 1 and 2

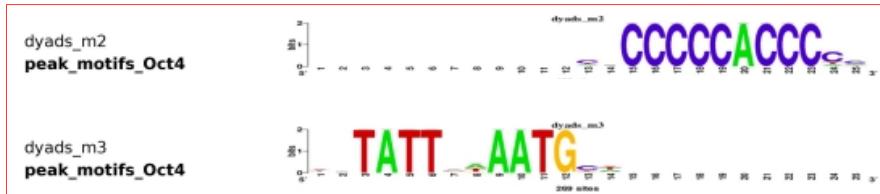


## Cluster 3 and 4

# Choosing a metric - Drawbacks

- Spurious alignments (alignment of non-informative columns)

$\text{cor} \geq 0.8$



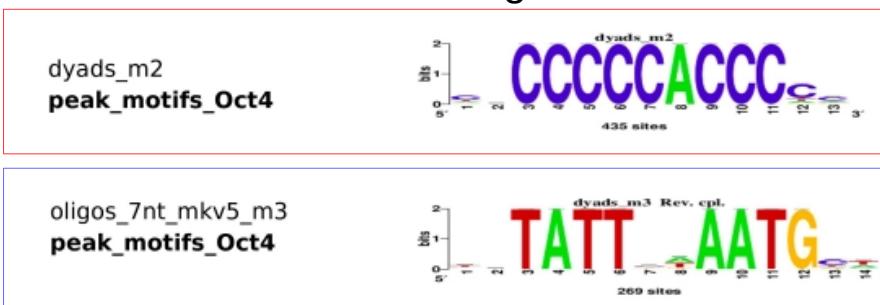
Cluster 1



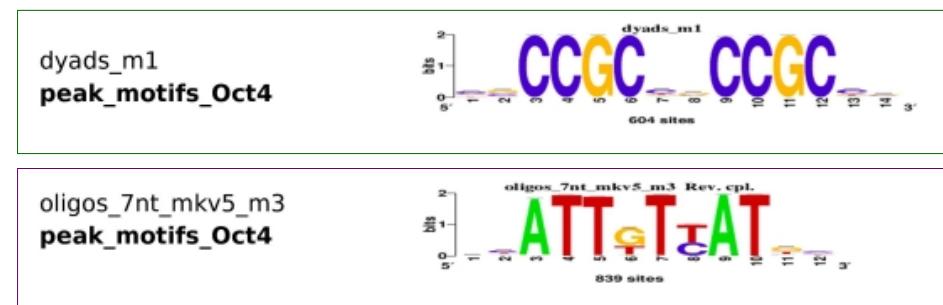
Cluster 2

- Including the width of the aligned segment avoid the spurious alignments.  
But the motifs are not aligned.

$\text{cor} \geq 0.8 + w \geq 5$



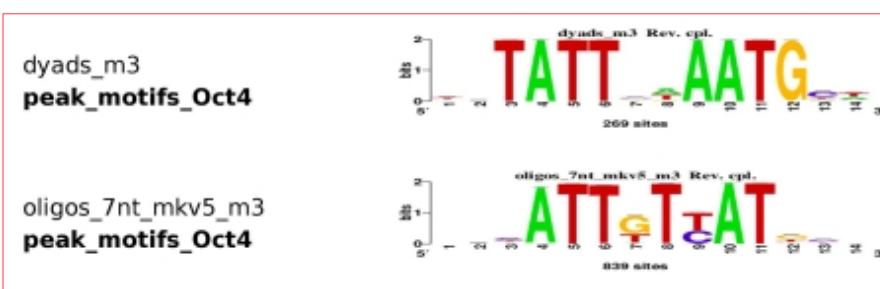
Cluster 1 and 2



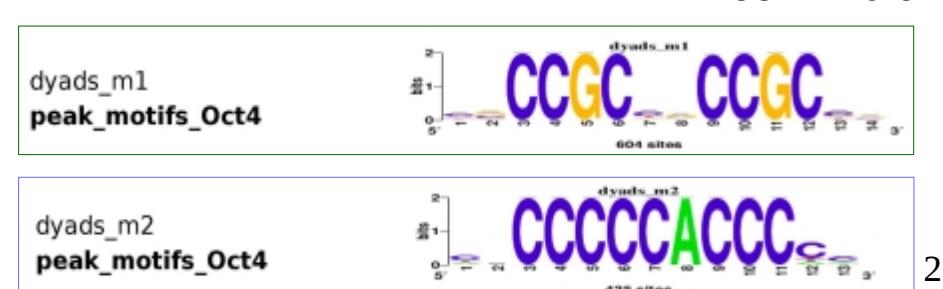
Cluster 3 and 4

- A correction in these metrics (e.g. normalization with the aligned segment) improves the results.

$\text{Ncor} \geq 0.6$



Cluster 1



Cluster 2 and 3

# *Example 1 : clustering several motif sets*

- **Main Objective:**

Discover and Cluster motifs in Oct4 ChIP-seq peaks with different algorithms (RSAT + MEME + HOMER).

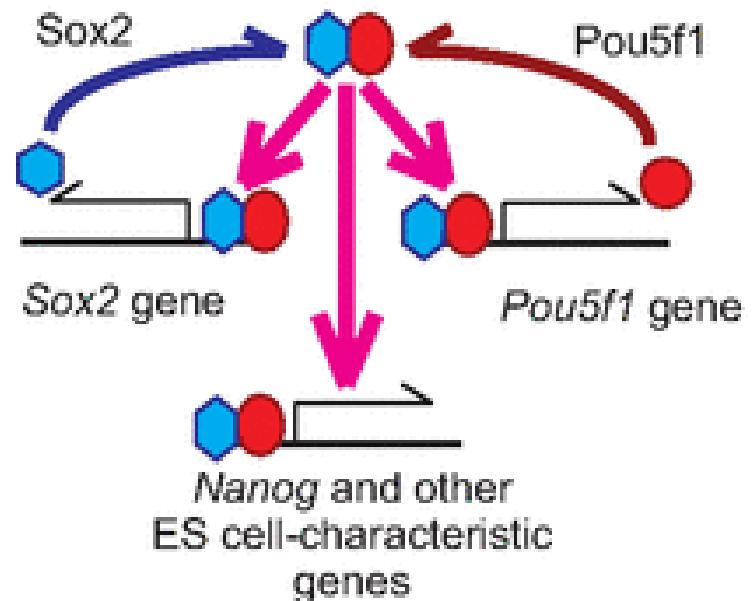
- **Specific Objectives:**

- Cluster several collections (files) of motifs coming from different sources (Databases, Algorithms, etc).
- Reduce the redundancy in multiple collections of motifs.
- Obtain a representative (non-redundant) set of motifs.
- Identify the Oct variants (Monomer, homodimer, heterodimer, etc).
- Use STAMP<sup>1</sup> as alternative tool to cluster the motifs.

# *Oct4 and Sox2*

- Forced expression of four TFs expressed in ES Cells (c-Myc, Klf4, Oct4, Sox2) can reprogram mouse embryonic fibroblast to pluripotent cells.
- OCT-TFs bind to target sequences either by homodimerisation or by heterodimerisation with other TFs.
- Oct4 (Pou5f1) can heterodimerise with Sox2. (SOx + OCt = SOCT).

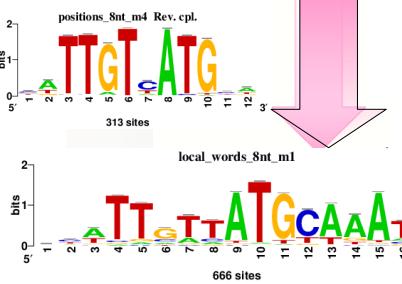
Sox2 = CATTGTT  
SOCT = CATTGTTATGCAAAT  
OCT4 = ATGCAAAT
- SOCT is the top motif identified with *de novo* motif discovery algorithms independently in Sox2 and Oct4 ChIP-seq peaks from mouse ES cells.



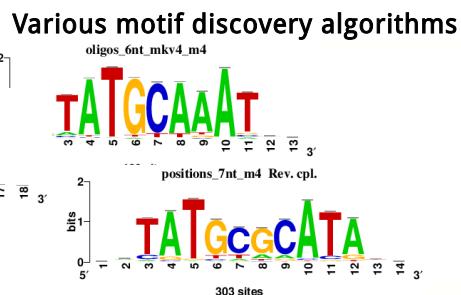
# Motifs found in Oct4 ChIP-seq peaks

>mm9\_chr1\_3473041\_3473370\_+  
 ctgtctcttatcttgcttaataaaggat  
 ctcttgtattggaaattgggttgttggg  
 tatatcctgtgcctaattgcataatgga

## Oct4 ChIP-seq peaks<sup>2</sup>



**Motif Discovery** RSAT

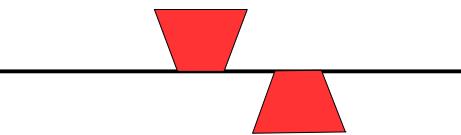


## Identification of Oct variants

### Monomer



### Plaindromic Homodimer



### Heterodimer



### Alternative Motif



cluster1\_node\_8\_9motifs

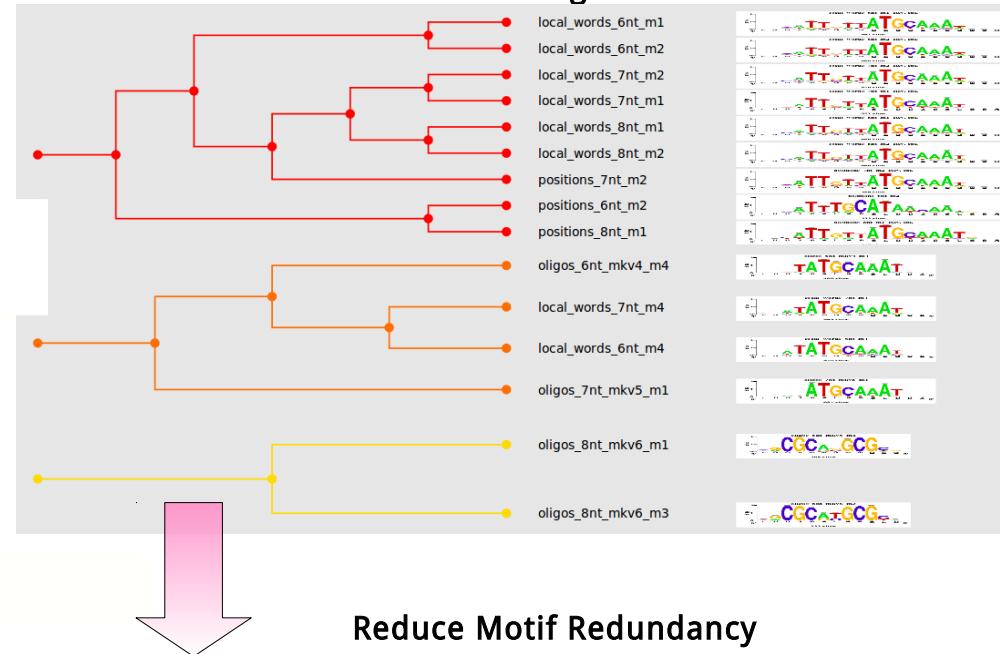
cluster2\_node\_3\_4motifs

cluster3\_node\_1\_2motifs

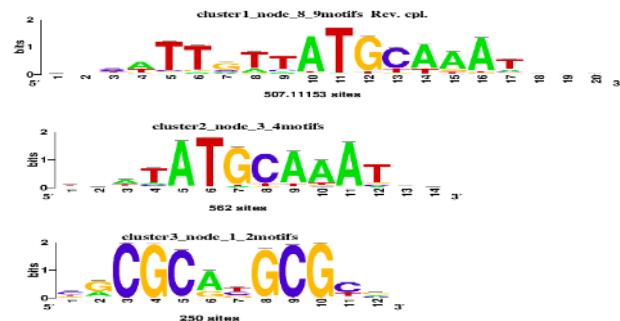
Non-redundant motifs compared with DataBases



## Clustering of motifs



## Reduce Motif Redundancy



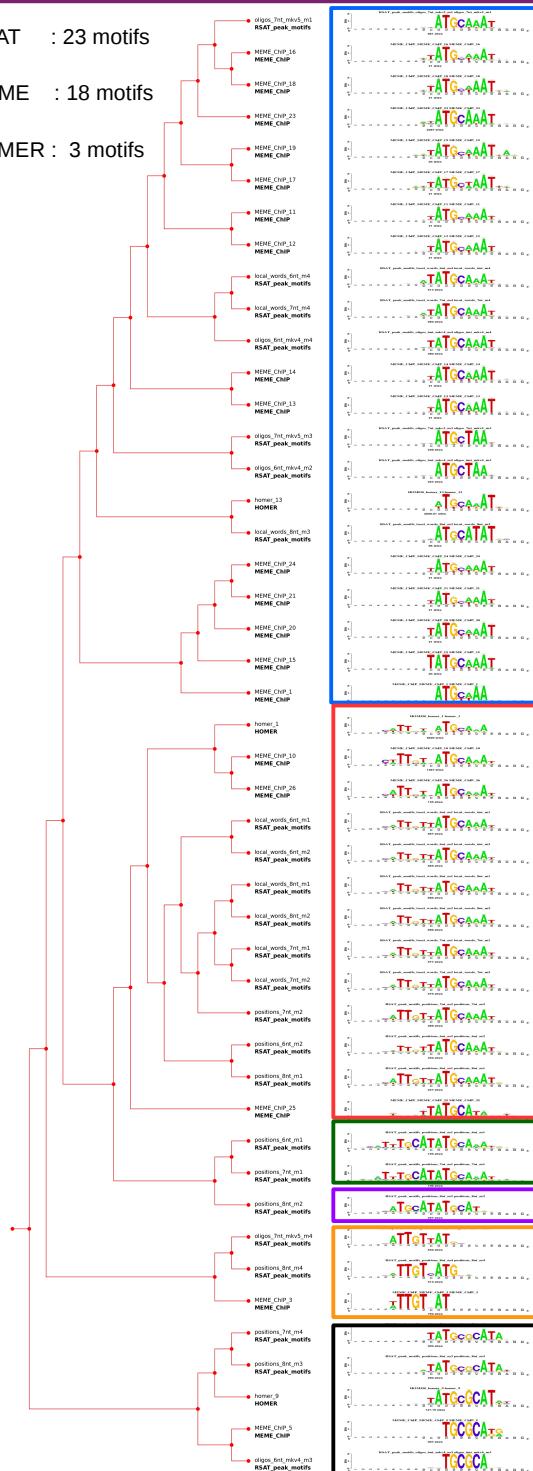
- 1.- Adapted from Thomas-Chollier et al. (2011) **RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets**. Nucleic Acids Research.  
 2.- Chen X et al. (2008). **Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells**. Cell

# **One cluster with Oct4 motifs**

RSAT : 23 motifs

MEME : 18 motifs

## HOMER : 3 motifs

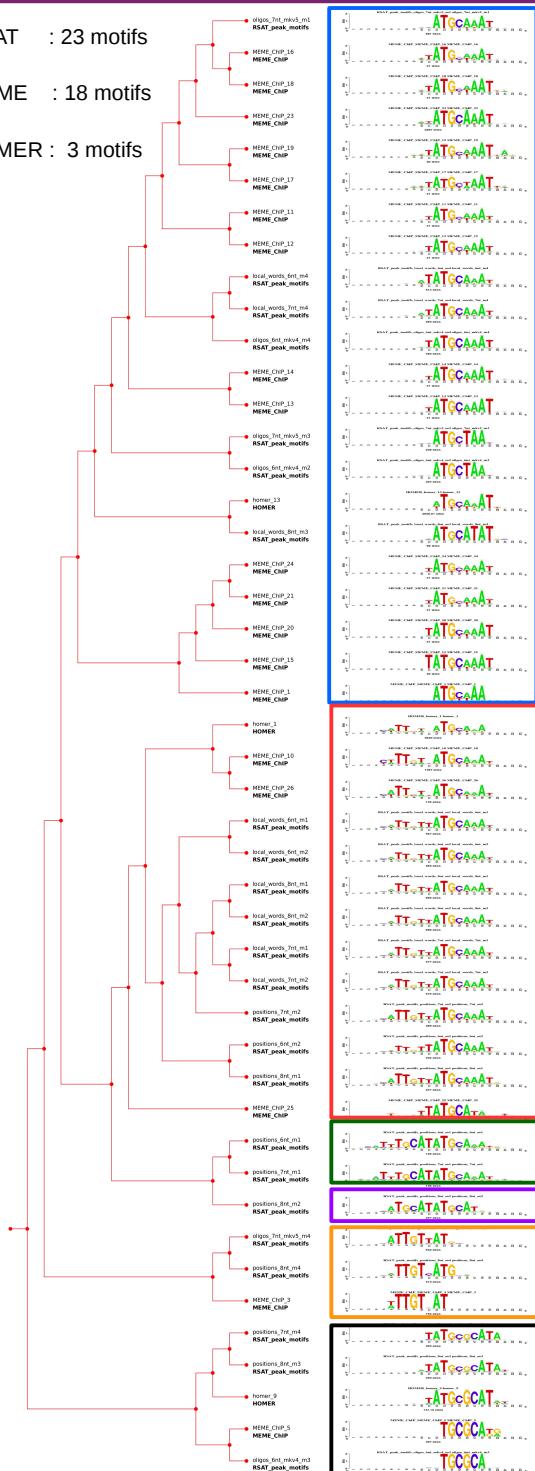


# **One cluster with Oct4 motifs**

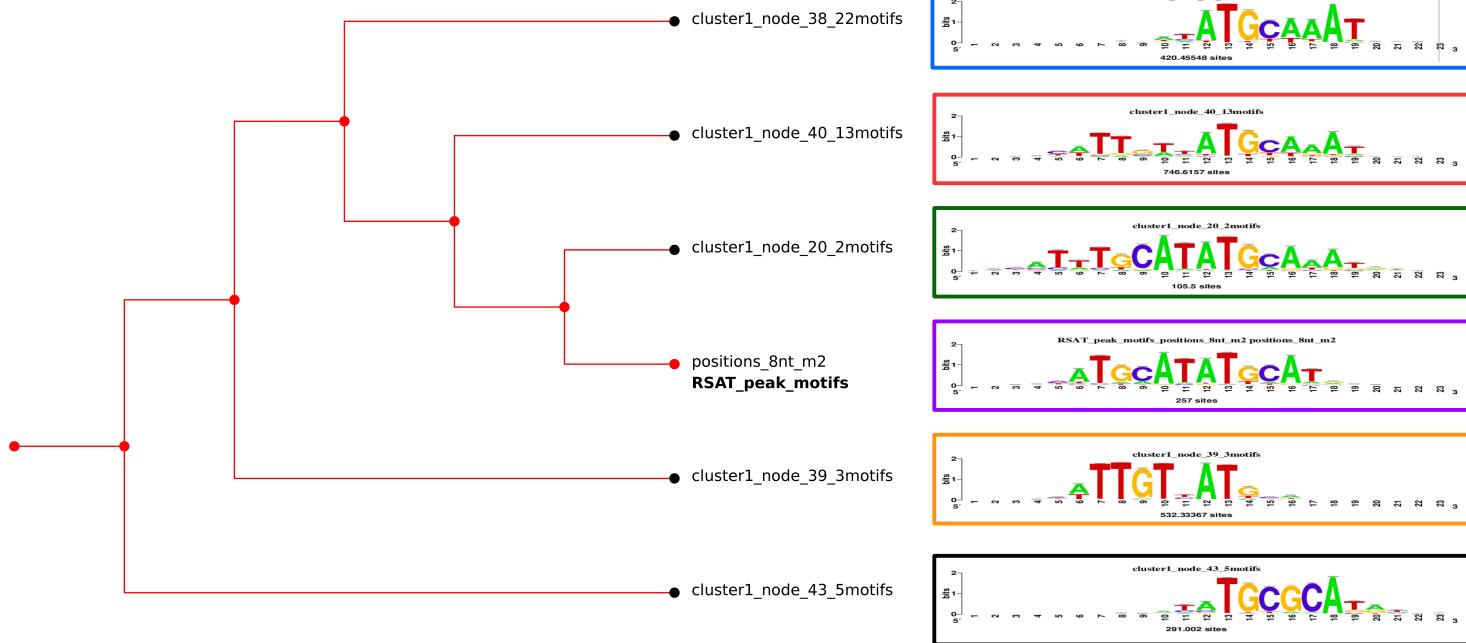
RSAT : 23 motifs

MEME : 18 motifs

## HOMER : 3 motifs



# **Collapsed tree with non-redundant motifs**

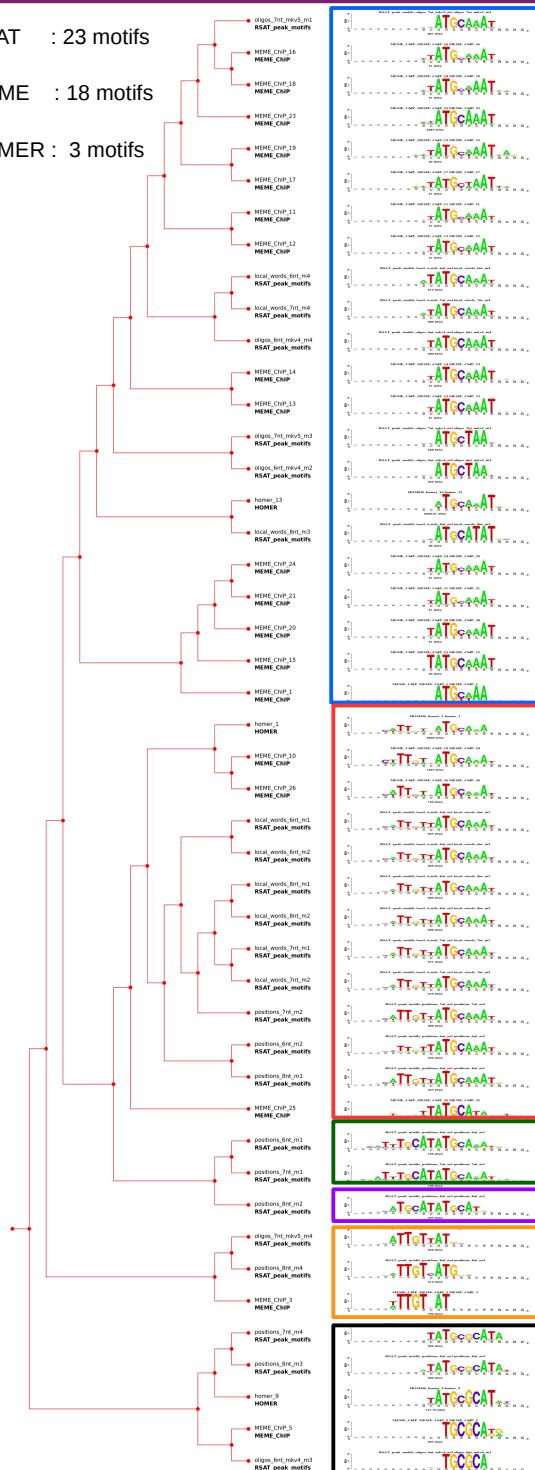


# **One cluster with Oct4 motifs**

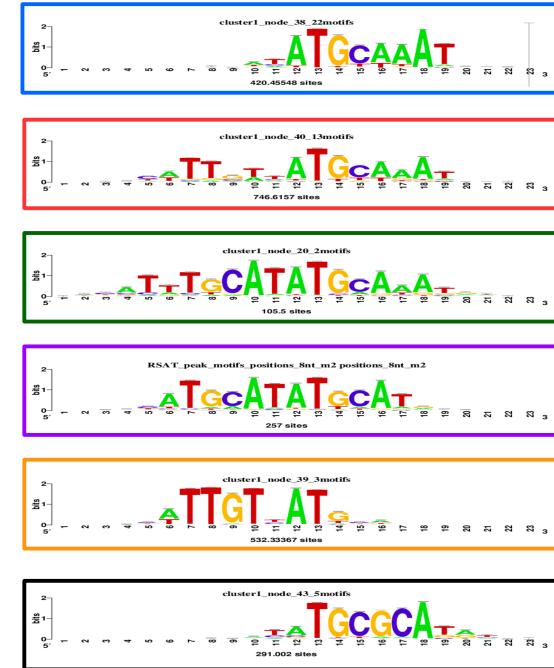
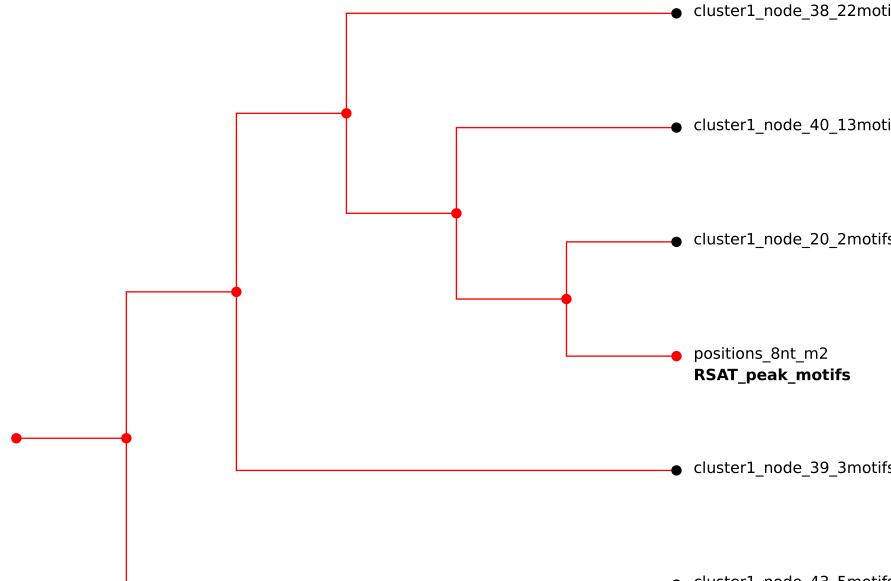
RSAT : 23 motifs

MEME : 18 motifs

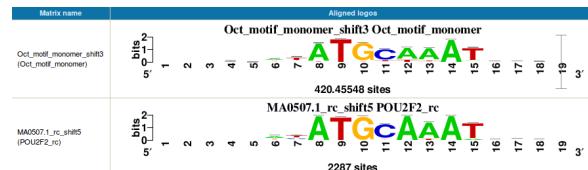
## HOMER : 3 motifs



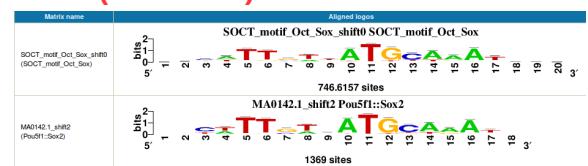
# **Collapsed tree with non-redundant motifs**



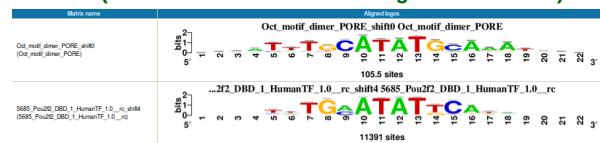
# Canonical Oct



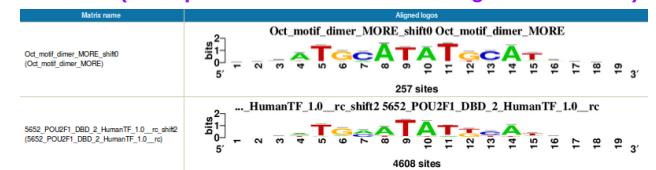
## SOCT (Sox + Oct)



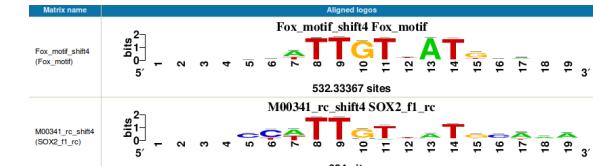
### **PORE (Palindromic Oct factor Recognition Element)**



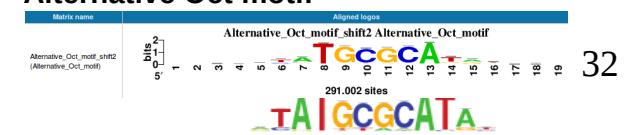
**MORE** (More palindromic Oct factor Recognition Element)



Sox



## Alternative Oct motif



# (The same) Motifs clustered with STAMP<sup>1</sup>

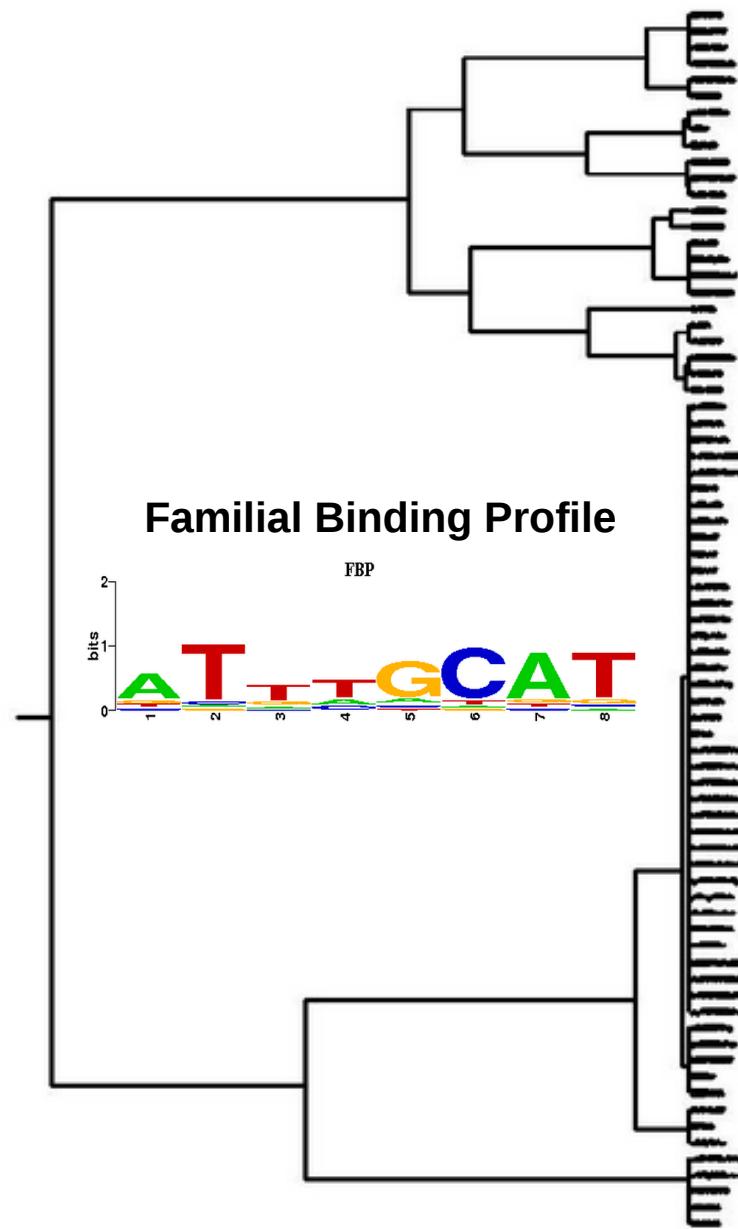
## Alignment

```

cgGCAAGCgg: -----CGCACGC-----
gwdATGCTAAtk: -----TTAGCAT-----
nwTATGCCAAATrw: -----ATTTCATATA-----
kaywTTgTTATGCCAAAtkw: -----ATTTCATATAAANAM-
wsaTTTGCTATAcAAwrr: -----ATTTCATATAAANAM-
dcaTATGCCAAATrw: -----ATTTCATATA-----
ywyATtTGCATATGCCaaAtrr: -----TTTCATATGCCAAAN-
twcATTTCATATAAAnrr: -----ATTTCATATAAANAM-
wtttTtwTTTTtaAAAAnw: -----TTTTTTTTTTAAAAAAAM-
wtATGCCAAATrw: -----ATTTCATATA-----
csCGGCCCTCbs: -----CGGCCCTC-----
wsaATGCCAAAtk: -----TTAGCAT-----
nsATACAAATrr: -----ATTGGTTAT-----
aywTTgTATGCCAAAtkw: -----ATTTCATATAAANAM-
wsaTTTGCTATAcAAwrr: -----ATTTCATATAAANAM-
kxaTATGCCAAATrw: -----ATTTCATATA-----
wyaTtTGCATATGCCaaAtrw: -----ATTTCATATAAANAM-
wsaTTTGCTATAcAAAtrg: -----ATTTCATATAAANAM-
caTATGCCATATtg: -----TATGGCTATA-----
crCGCAaGCCcr: -----GCCGASGCGC-----
csCGGCTCCGss: -----CGCTTCCG-----
yrCGCRYGCgyr: -----GCCGATGCCG-----
CGGGCAGGCCcr: -----CGGGCAGGCCG-----
aywTTgTTATGCCAAATdw: -----ATTTCATATAAANAM-
whATTTGCTATAcAAwrr: -----ATTTCATATAAANAM-
rcATATGCCATTrw: -----ATATGCAT-----
cyATGTTTATGCCAAATgur: -----ATTTCATATAAANAM-
ywytATGCATATGCATTrw: -----ATGCATATGCAT-----
gcaATGEGCATAtgc: -----TATGGCTATA-----
wrCATgACAwwr: -----TTTGTGATG-----
ATGyvNA: -----TTTGCAT-----
CCMCdCCC: -----GGGGGGGG-----
ATrACAAw: -----TTTGTGAT-----
rGGAr: -----TTTCCY-----
TGGGCAKkr: -----TGGGCACTR-----
CAAGGCTCA: -----TGACCTTG-----
CwGFR: -----TCCCGW-----
CCKCCKCC: -----CCNCCKEC-----
CCCCwCycCc: -----KGGGGGGGGGG-----
ywTTsawATGCCaaAt: -----ATTTCATATAAANAM-
wtATGcwRATt: -----ATTTCATATA-----
nwTATGcwAATk: -----ATTTCATATA-----
twkTATGcwAATkarm: -----ATTTCATATA-----
ywtdTATGcwAATkarm: -----ATTTCATATA-----
TATGCCAAAT: -----ATTTCATATA-----
wtATGywRATlw: -----ATTTCATATA-----
nwTATGcwAATTw: -----ATTTCATATA-----
wtATGcwAATTlw: -----ATTTCATATA-----
nwTATGcwAATkAg: -----TNTATTCATATA-----
TATGcwAAT: -----ATTTCATATA-----
TATGyvAAT: -----ATTTCATATA-----
mKTTTkyTkTkTwkTwkG: -----TTTTKYTTTTNTNTNTNG-----
ykwATTTGCTTrw: -----ATTTCATATA-----
TATGyvAAT: -----ATTTCATATA-----
waddTATGCCATAkkdw: -----WNNTTATGCCATANNT-----
aTbwGCTawsAAwr: -----ATTTCATATAAANAM-
CATTGNTATGCCAA: -----TNTGATCNINNAM-
AGGCTGGCTGGRA: -----YCCAGSCAACCCN-
SCTGCTGCTGCYBC: -----RGCAAGCAGCAAG-
YGCCCATGCCAASN: -----GCCATGCCGGE-----
RGGGGGGGGGCGNG: -----GGGGGGGGGGCGNS-
TTTGTGTTGTTTG: -----TTTGTGTTGTTTG-----
TCCCAAGCACCCACA: -----TGTGGGGGHTGGG-
KKAGGTGTTGCCCTT: -----NRNTGTTGGYYNN-
NSDTATGCCATAT: -----TATGGCTATNN-----
CCTGKCTCTG: -----NTGNCNTNG-----
CYCAGCYCTS: -----NCACGNYNN-----
RRGRGRGRGRG: -----CYCYCYCYCY-----
KMATTGCTAT: -----ATTTCATATA-----
YTCCCTGGAAD: -----NONG-----
CWGCGWCWGN: -----CNGCGWCWG-----
MAAAAAGAAA: -----TTTTTTTTTT-----
TGTGTGTGTG: -----TGTGTGTGTG-----
AAAAAAA: -----TTTTTTTTTT-----
CGCATGCCA: -----TGCCTATGCCG-----

```

## Tree



- Pros:

- a) The fastest
- b) Several metrics can be selected
- c) Several alignment types allowed.

- Cons:

- a) The tree is not partitioned in a forest.
- b) The motifs are forced to be aligned in a single alignment.
- c) Only one FBP can be obtained.

## *Example 2 : clustering a complete motif database*

- **Main Objective:**

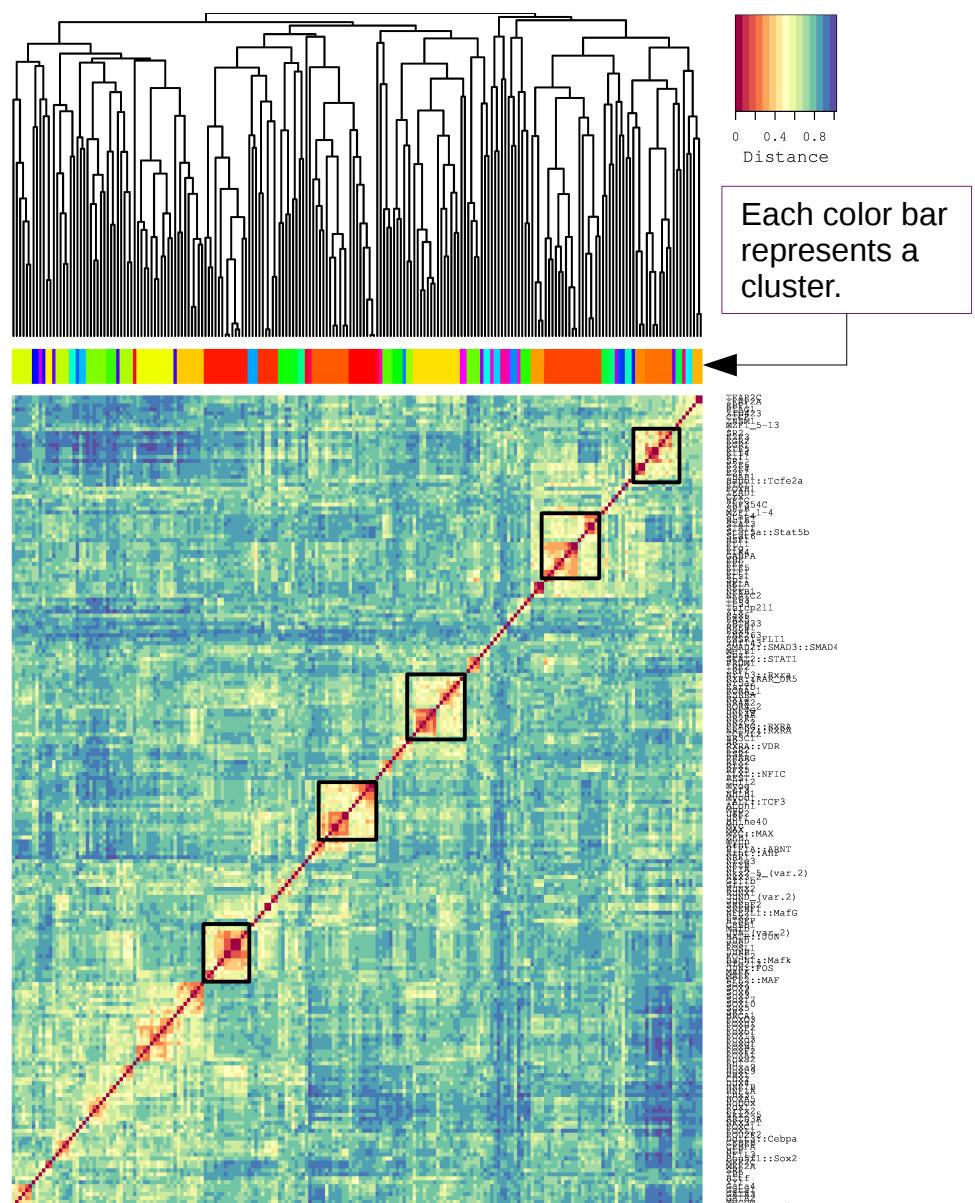
Cluster the full collection of motifs in different Databases (Jaspar vertebrates and insects, Hocomoco, RegulonDB, etc)

- **Specific Objectives:**

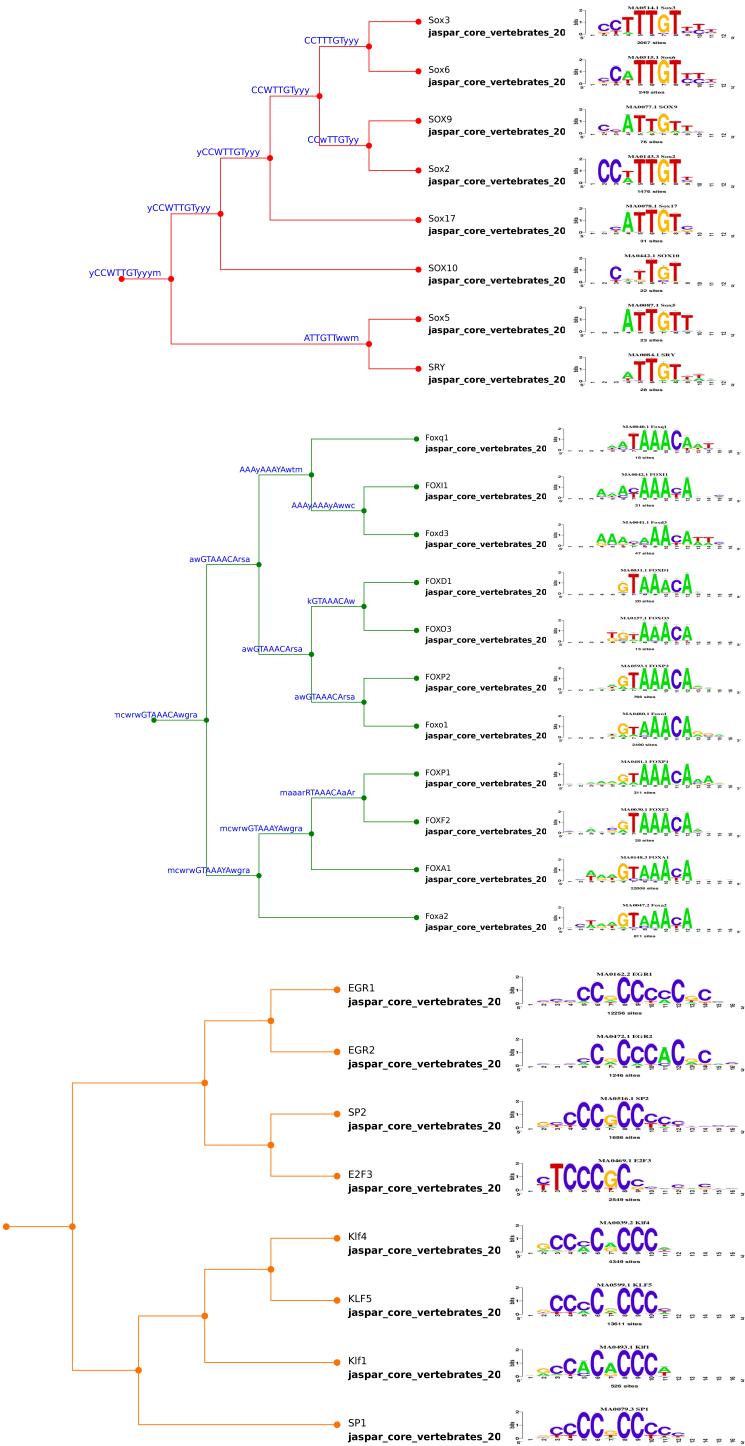
- Identify groups of motifs belonging to the same TF family.
- Find a threshold to separate the motifs in Families.

# Example 3 : clustering Jaspar vertebrates

- 205 Motifs => 68 clusters

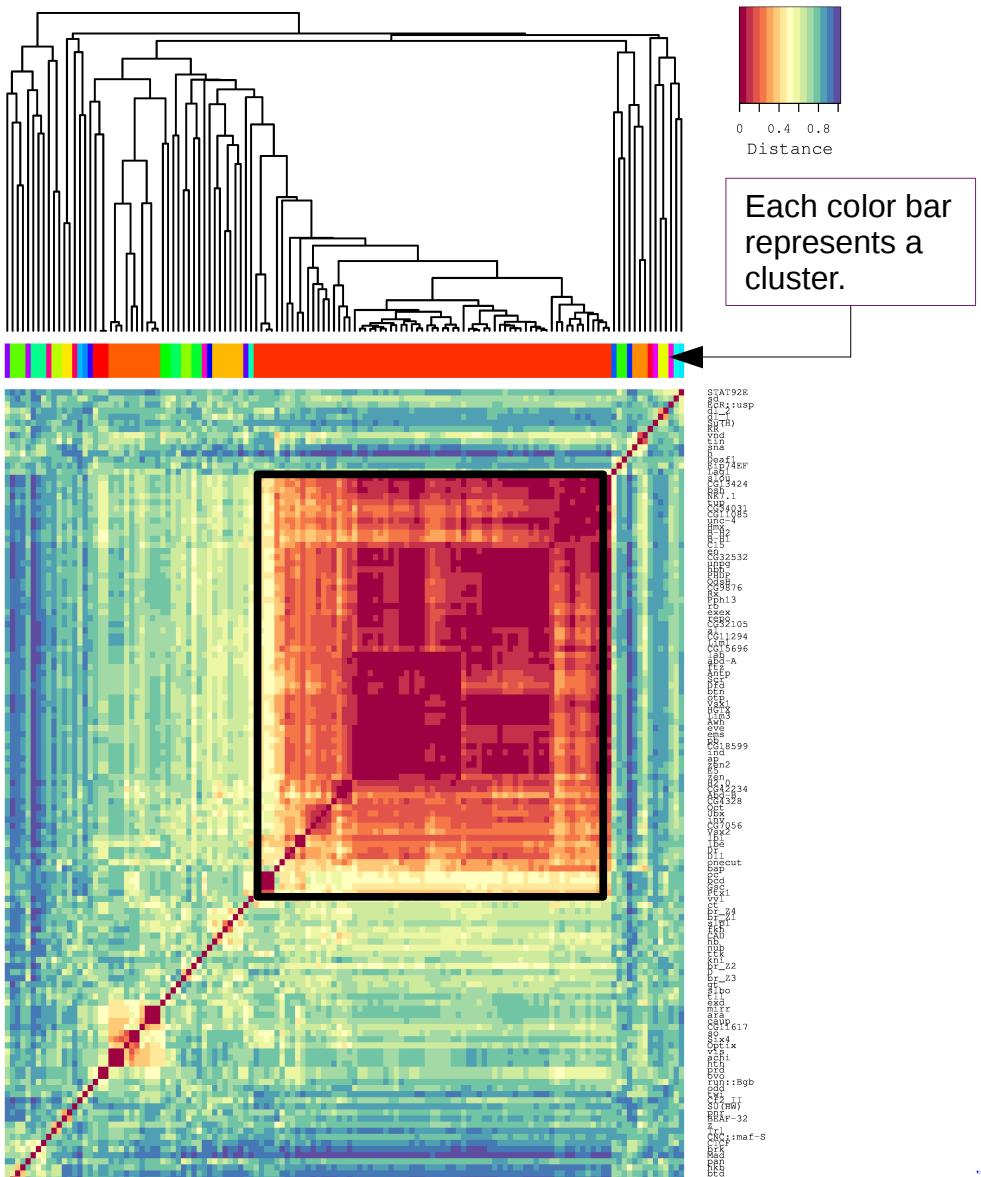


Many small clusters which correspond to different TF Families (e.g. Fox, Gata, Sox, etc).

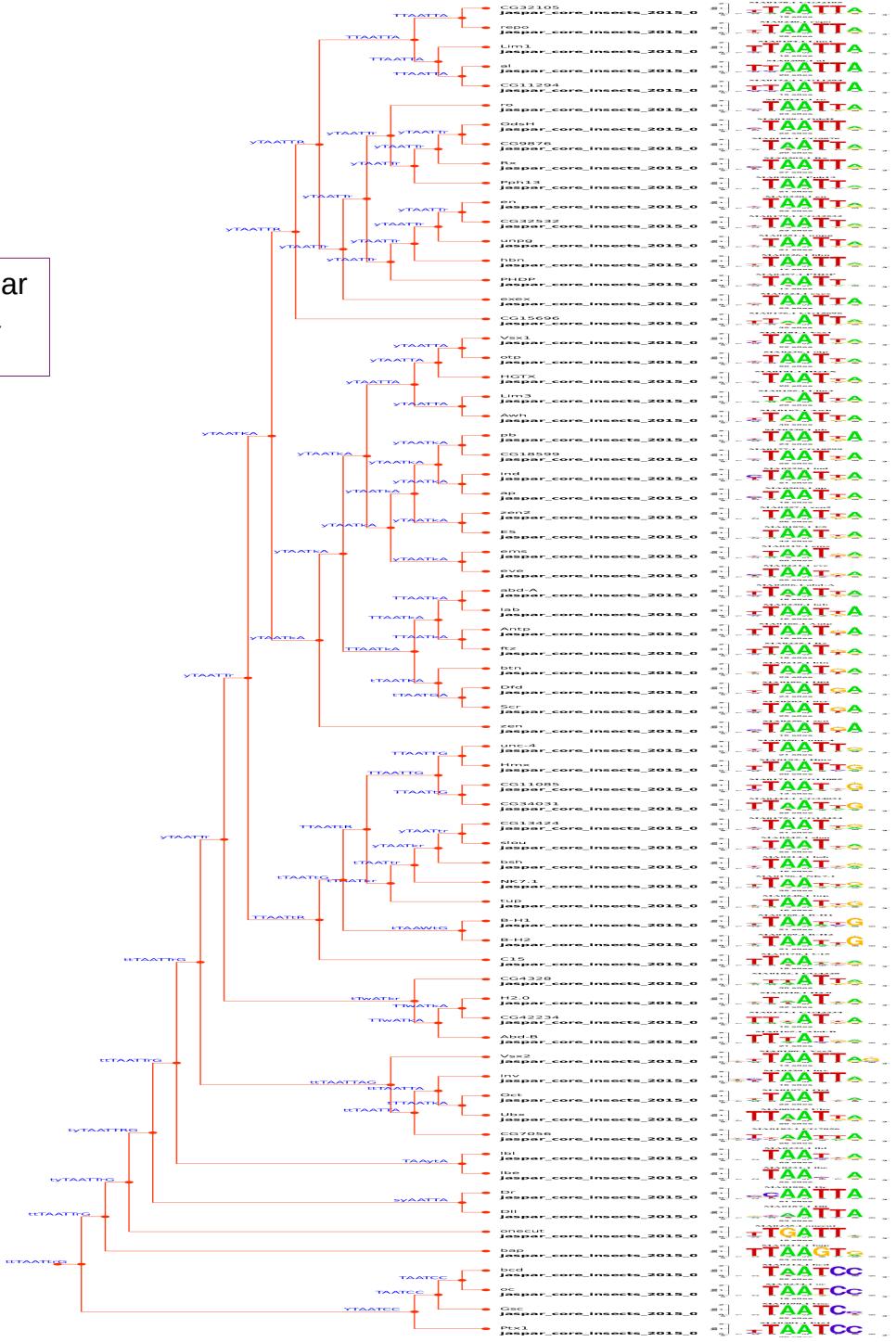


# Example 3 : clustering Jaspar insects

- 205 Motifs => 68 clusters



Almost half of the motifs in the collection are grouped in a big cluster, that corresponds to the Hox motifs.



# *Take Home Messages*

- There is no a 'best' metric to measure motif similarity. (But combining several metrics can help to properly compare or classify the motifs).
- The motif redundancy issue after using motif discovery tools can be faced with clustering of motifs.
- A large collection of redundant motifs can be reduced to a small non-redundant set, thus simplifying the analysis.

# Acknowledgments

- Jacques van Helden
- Morgane Thomas-Chollier
- Denis Thieffry
- Nishant Thakur
- Bruno Conteras

