

# Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L - 2022 | week 06

Pierre-Luc Germain

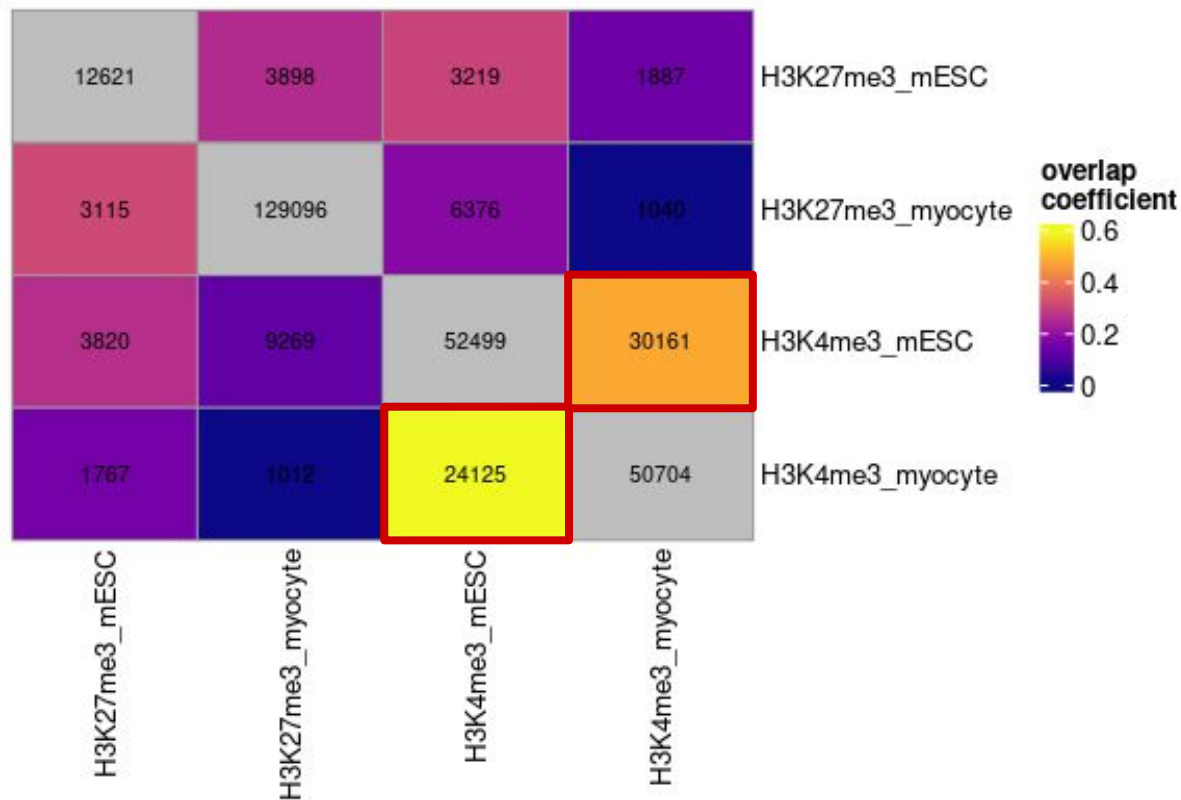
# Plan

- Quick things (see slack) :
  - Polls
  - New channel for discuss project ideas
  - New packages to install
- Debriefing on last week's assignment
- Overview of transcription factors and their binding specificity
- Motifs and related analysis

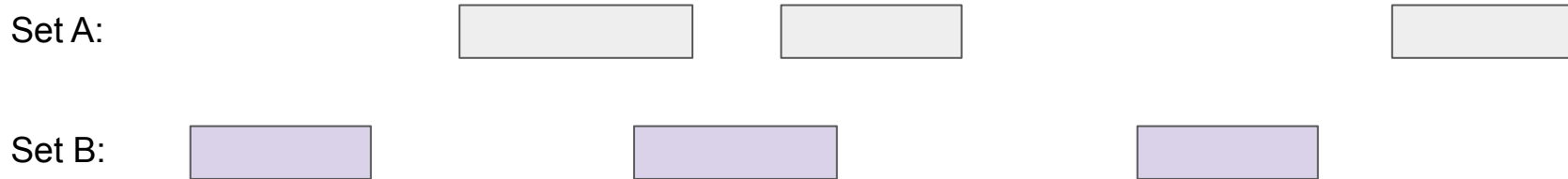
## A few extra questions raised

- Where do the files we export get saved? Can you move/copy files in your folders for them to be accessible by a new Markdown document?
- Why do we have to save the things from ENCODE with .gz in the end and not just e.g. .bed?
- What exactly are seqlevels?
- Why are overlaps asymmetric?

# Why are overlaps asymmetric?



# Why are overlaps asymmetric?



How many elements of A overlap elements of B?  $\rightarrow 2/3$

How many elements of B overlap elements of A?  $\rightarrow 1/3$

# Last week's assignment

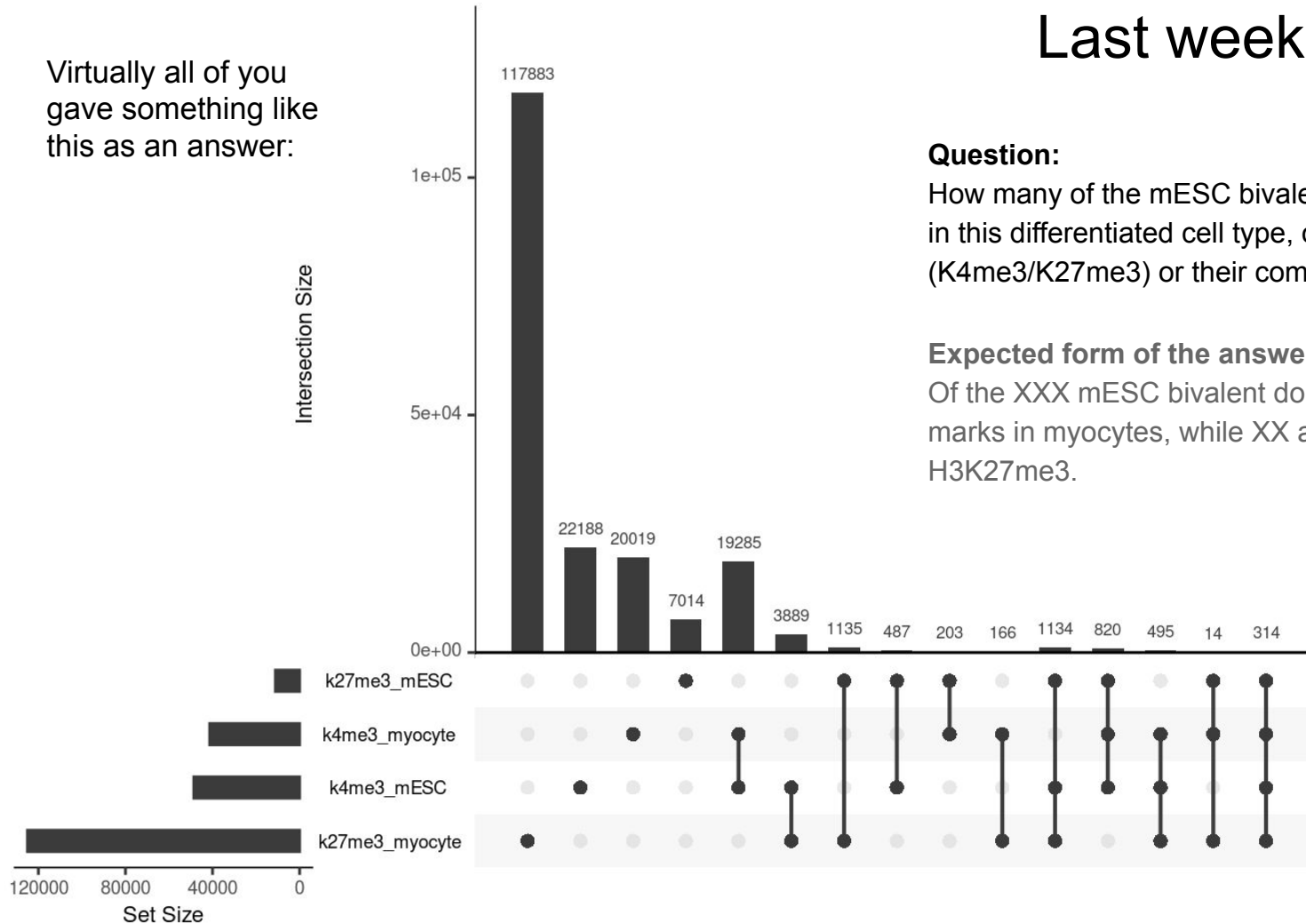
Virtually all of you gave something like this as an answer:

## Question:

How many of the mESC bivalent domains are, in this differentiated cell type, overlapping either mark (K4me3/K27me3) or their combination?

## Expected form of the answer:

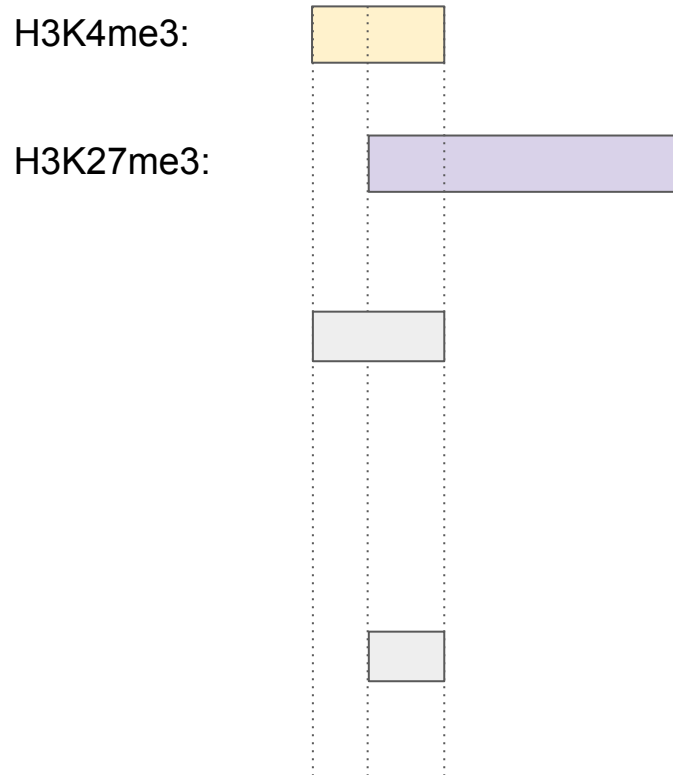
Of the XXX mESC bivalent domains, XX also have both marks in myocytes, while XX are H3K4me3 and XX are H3K27me3.



# Intersection & overlap:

## The example of bivalent domains

- **method one (overlapsAny):**  
find the H3K4me3 peaks that overlap a H3K27me3 domain
- **method two (intersect):**  
find the regions that are covered by both H3K4me3 and H3K27me3





**Transcription initiation complex**



[www.dnalc.org](http://www.dnalc.org)

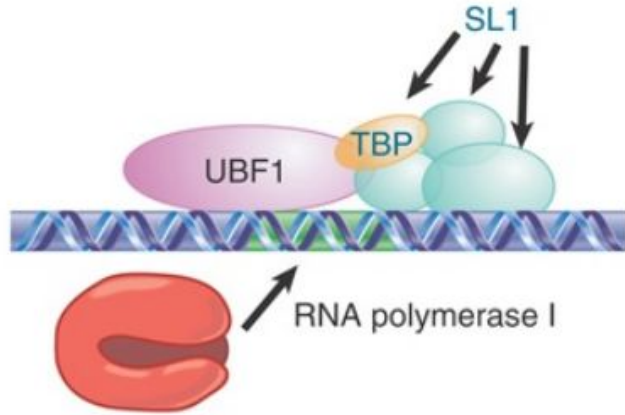
<https://youtu.be/SMtWvDbfHLo>

( See also [https://youtu.be/WW9IIYM\\_FC0](https://youtu.be/WW9IIYM_FC0) )



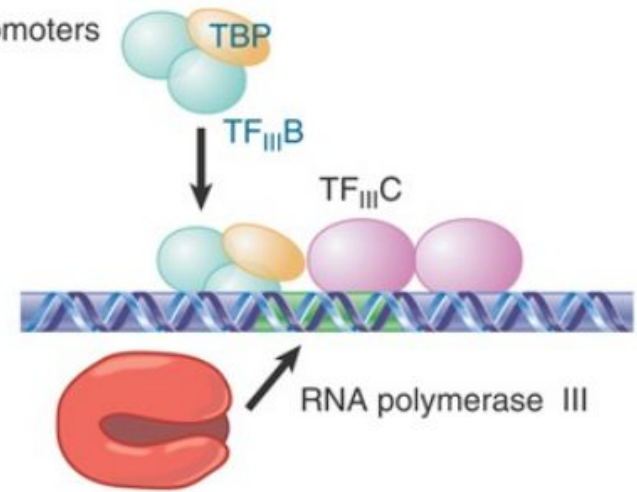
Pol I promoters

rRNA



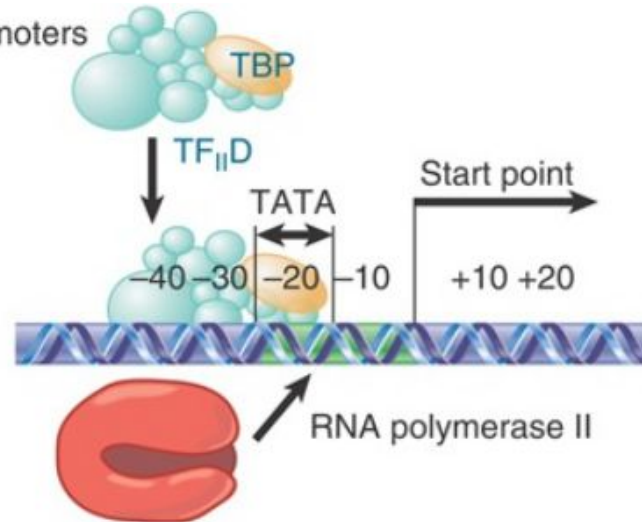
Pol III promoters

tRNA



Pol II promoters

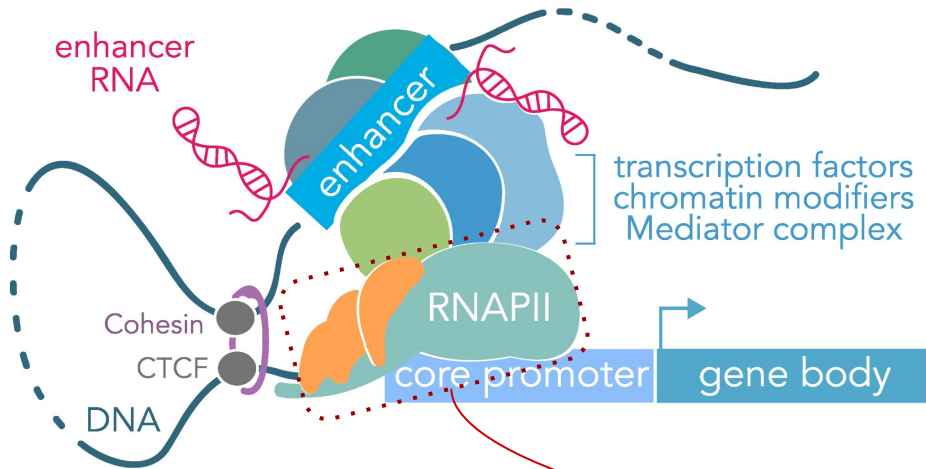
Most  
RNAs



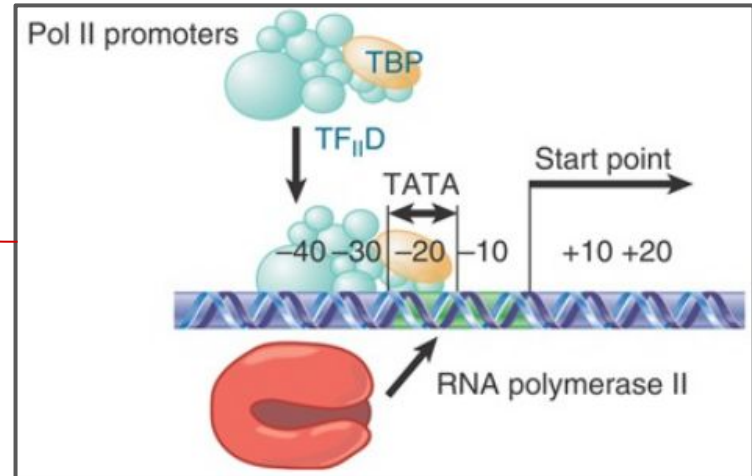
(Adapted from Krebs, Goldstein and Kilpatrick, Genes XII, 2018)

# Additional regulatory elements

## Enhancer-driven gene regulation



(Carullo and Day, Genes 2019)

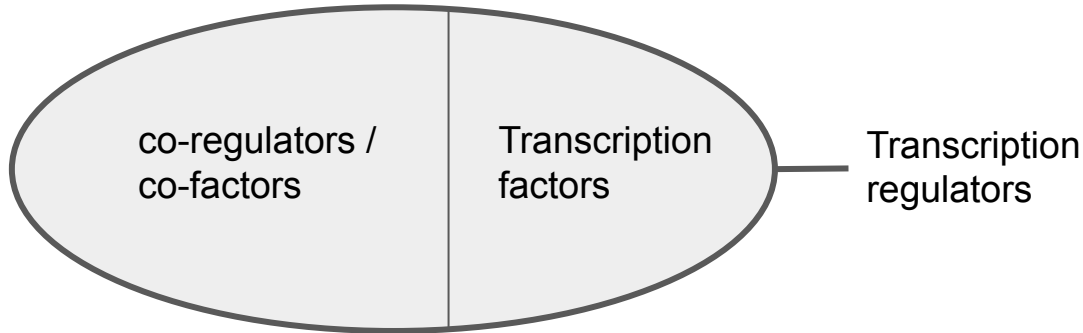
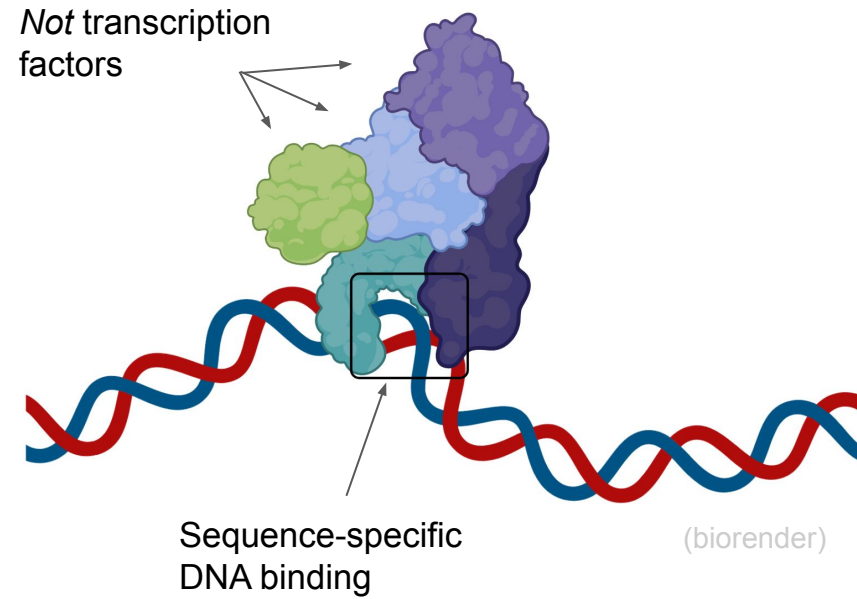


# What is a transcription factor?

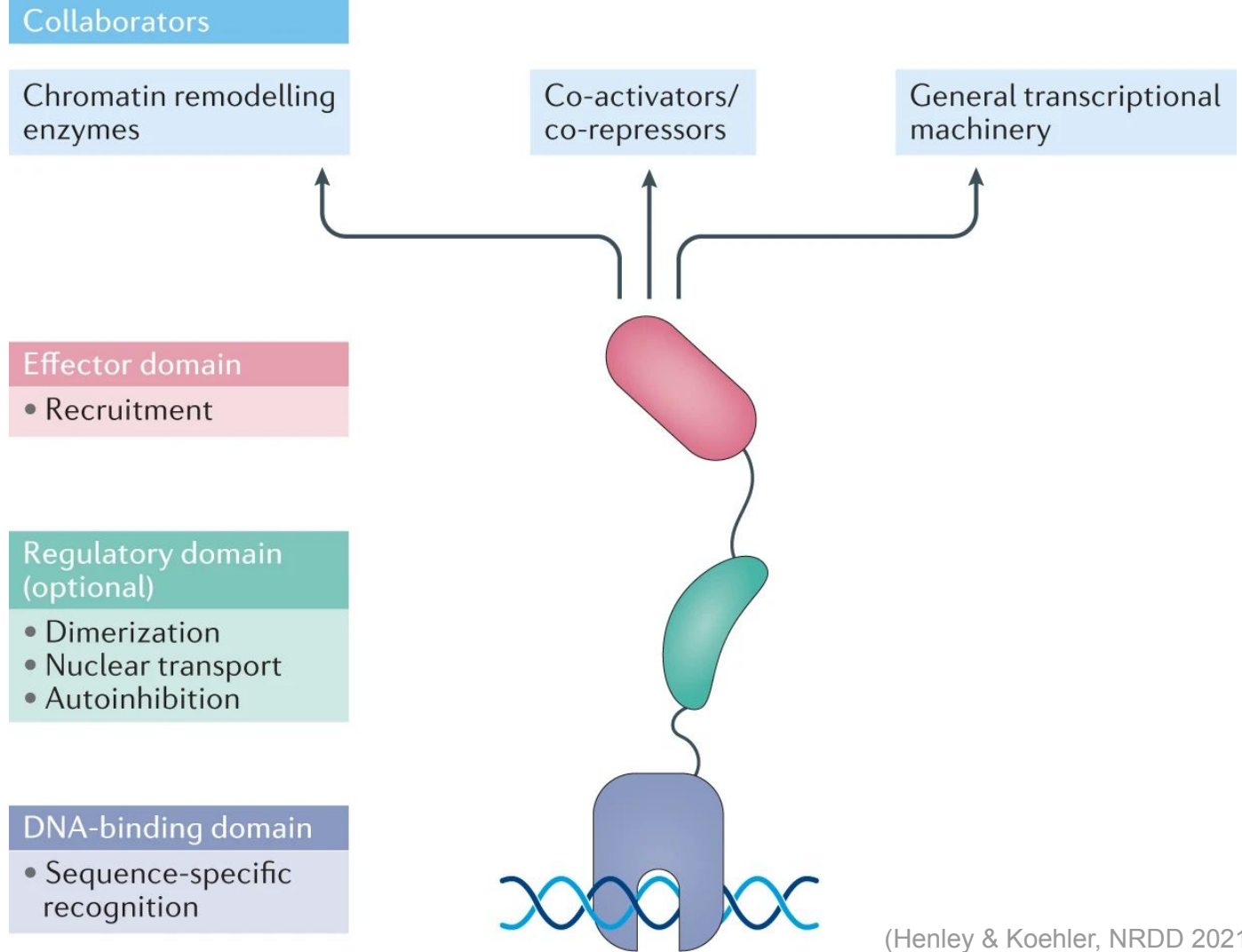
Proteins capable of both:

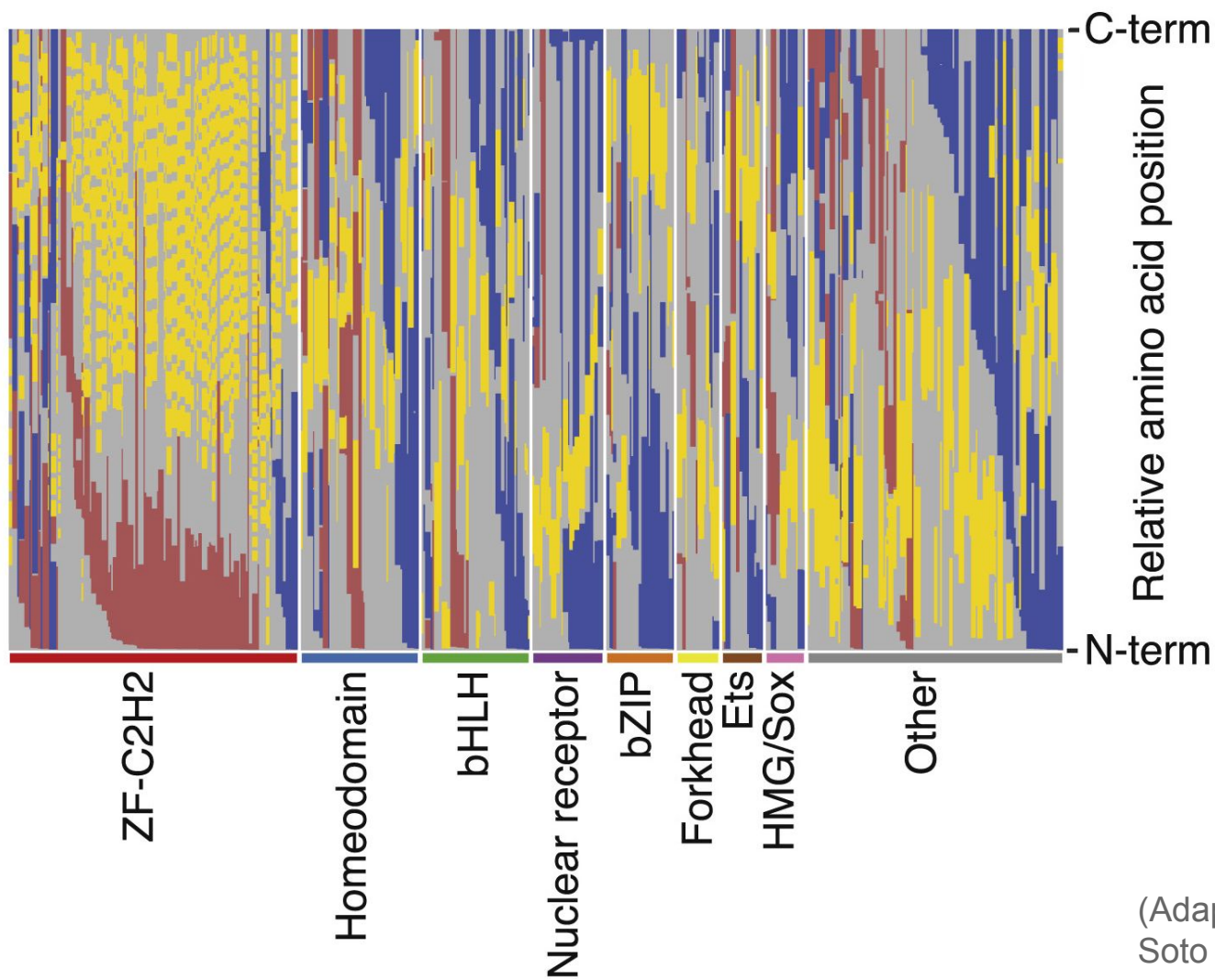
- 1) Binding DNA in a sequence-specific manner
- 2) Regulating transcription

(Lambert et al., Cell 2018)



# Anatomy of a transcription factor (TF)





While most TF have either an activating (AD) or repressive (RD) domain, some have both

(Adapted from  
Soto et al., Molecular Cell 2021)

Review (Cell 2018)

# The Human Transcription Factors

Samuel A. Lambert <sup>1, 9</sup>, Arttu Jolma <sup>2, 9</sup>, Laura F. Campitelli <sup>1, 9</sup>, Pratyush K. Das <sup>3</sup>, Yimeng Yin <sup>4</sup>, Mihai Albu <sup>2</sup>, Xiaoting Chen <sup>5</sup>, Jussi Taipale <sup>3, 4, 6</sup>  , Timothy R. Hughes <sup>1, 2</sup>  , Matthew T. Weirauch <sup>5, 7, 8</sup>  

Proteins capable of both:

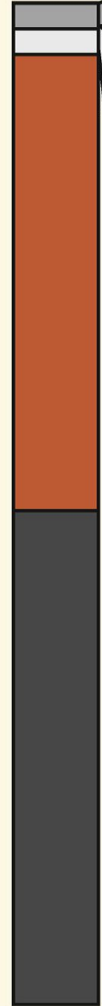
- 1) Binding DNA in a sequence-specific manner
- 2) Regulating transcription

According to their census, humans have 1570 transcription factors

78 TFs with  
Multiple DBDs

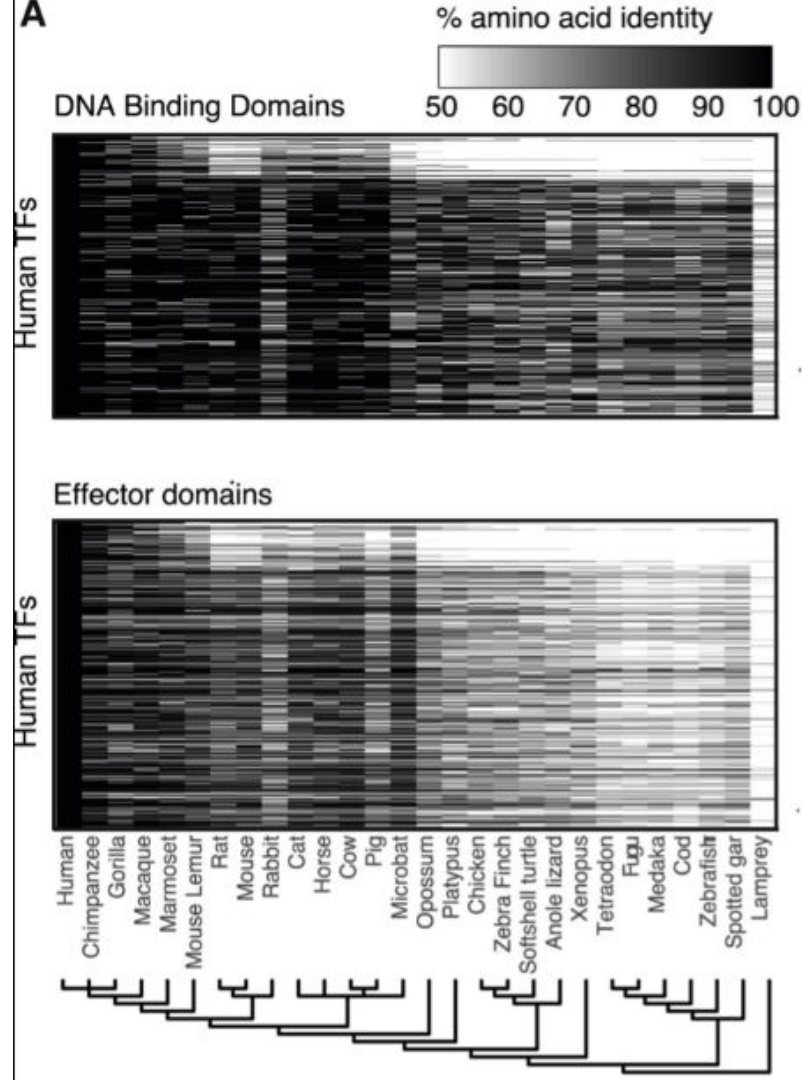
713 TFs with  
C2H2 ZF arrays

779 TFs with  
a single DBD



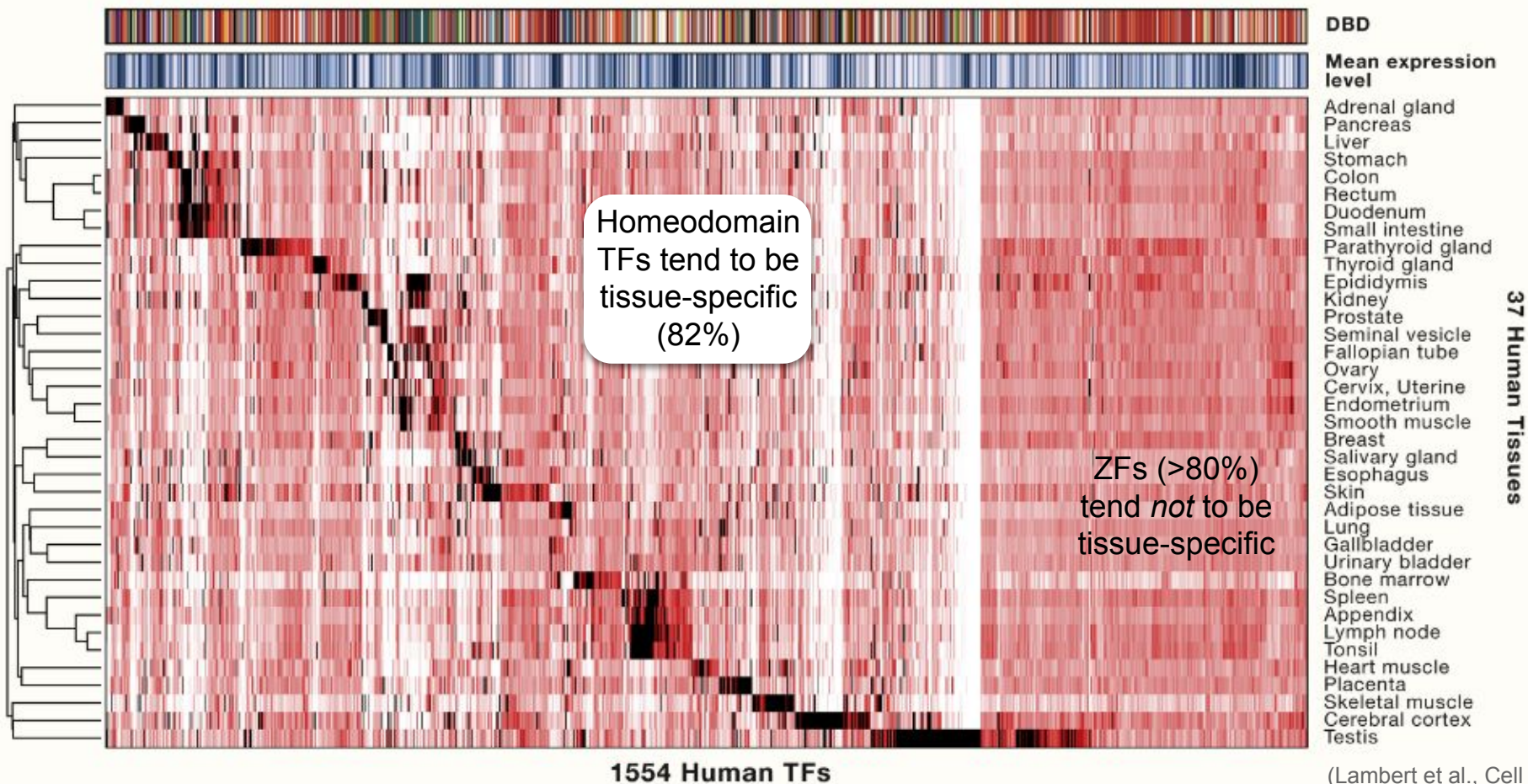
## Transcription factors are highly conserved

DNA binding domains show much higher conservation than effector domains



(Soto et al.,  
Molecular Cell 2021)



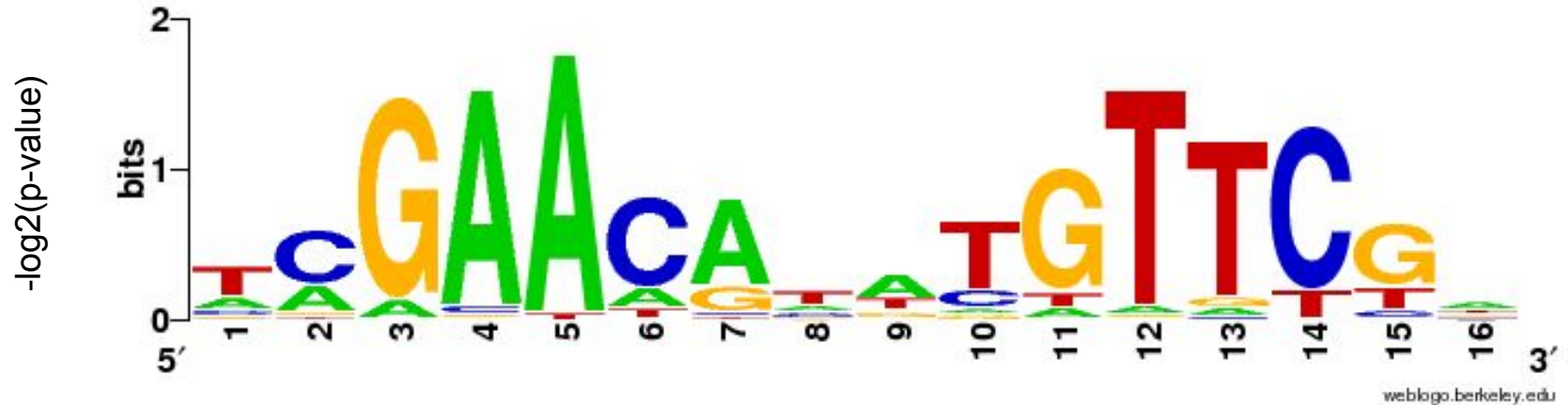




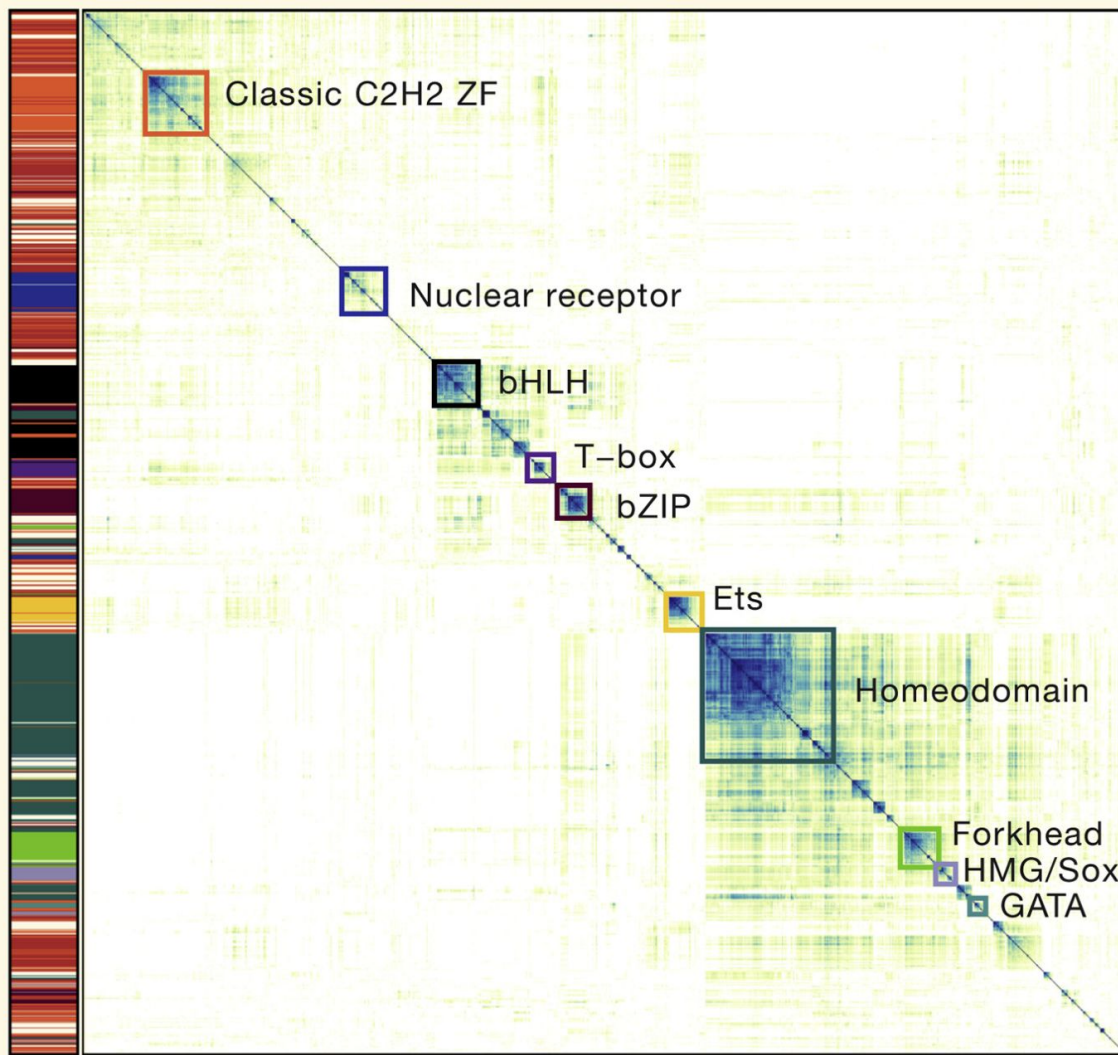
# Sequence-specificity

E.g. The LexA bacterial TF recognizes the consensus sequence

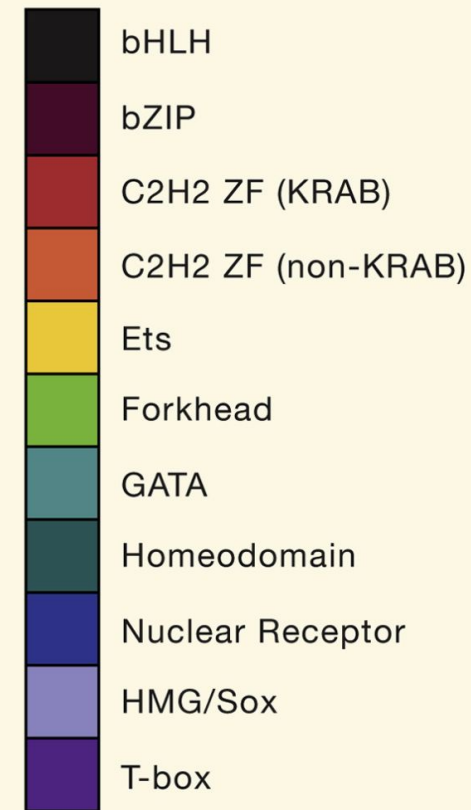
5' -GAACAnnTGTTTC-3'



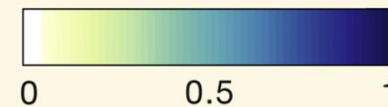
**TF Motifs**



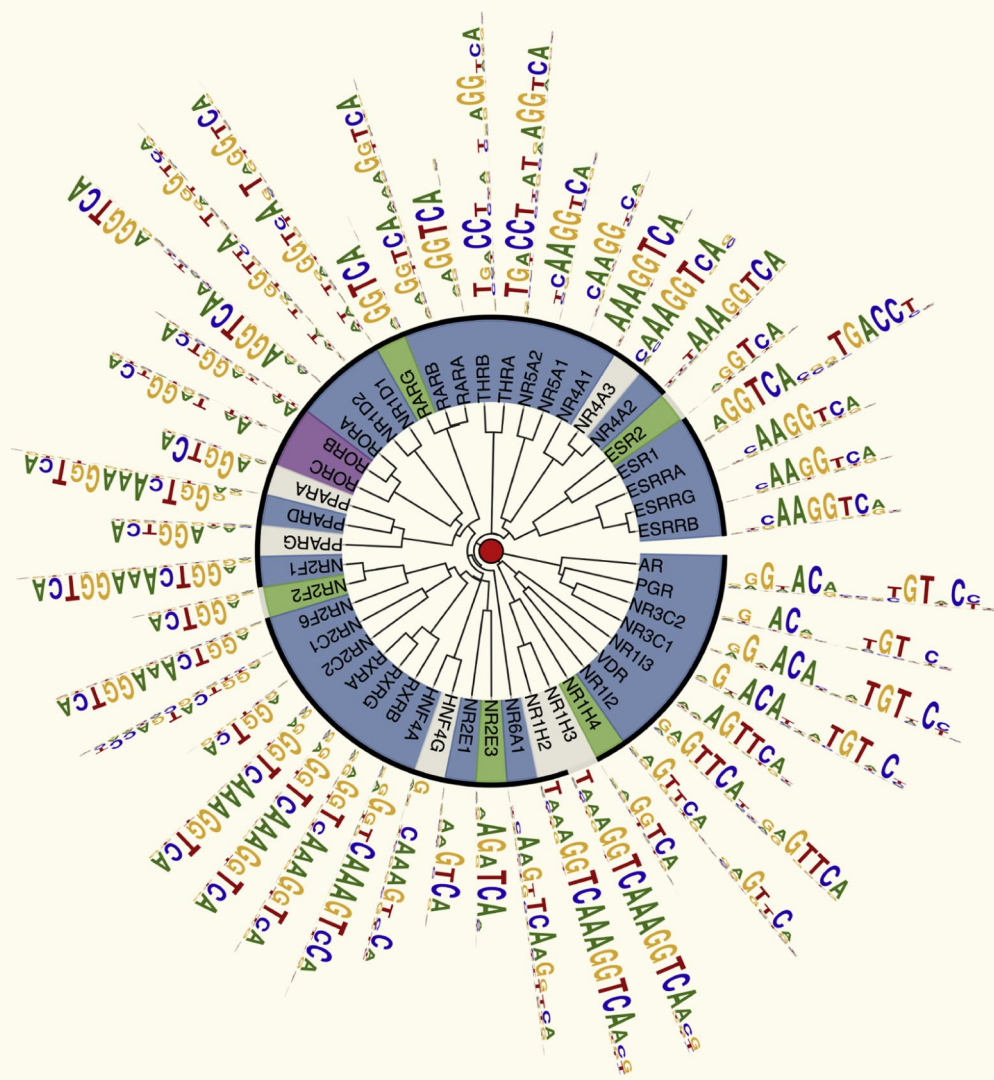
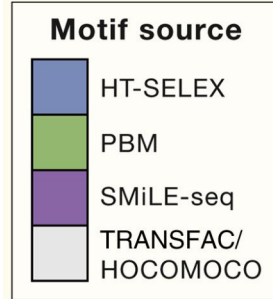
**DBD**



**Motif Similarity (PCC)**



# An example of TF motif degeneracy: Nuclear hormone receptors

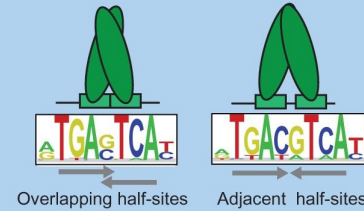


# Variations in DNA binding specificity

## Multiple Modes of DNA Binding

A

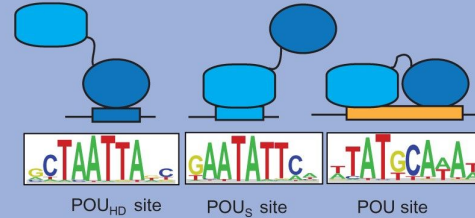
Variable Spacing



Gcn4 dimers can bind to bipartite sites with half-sites separated by variable-length spacers (82); motifs from (73,74)

B

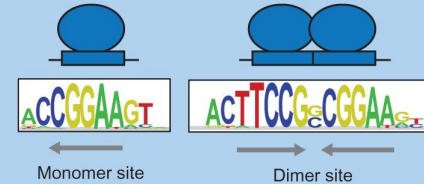
Multiple DBDs



Oct-1 can bind to different DNA sites using different arrangements of its two DNA-binding domains (91,92); motifs from (24)

C

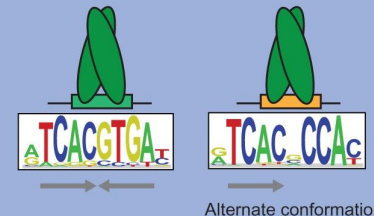
Multi-meric Binding



Elk1 can bind both as a monomer or as a dimer (95)

D

Alternate Structural Conformations



SREBP can bind to different DNA sites by adopting alternate structural conformations (96,97); motifs from (44)

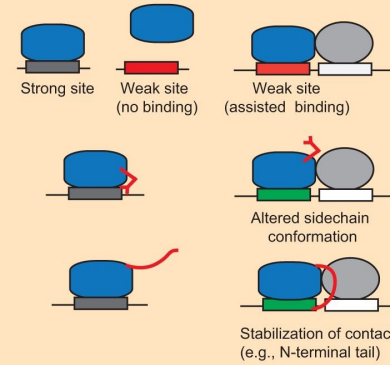
# Cooperative binding

Highly combinatorial  
binding of TFs

## Multi-Protein Recognition Codes

A

Cooperative binding

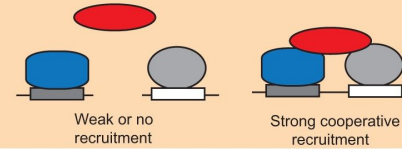


Enhanced complex stability due to cooperativity allows binding to lower-affinity (weak) sites (103,104,106)

Inter-protein interactions alter or stabilize protein-DNA contacts, altering DNA-binding specificity (40,106,107)

B

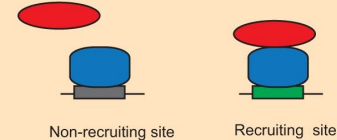
Cooperative recruitment



Cofactor recruitment requires multiple factors (rather than only one), allowing more specific cofactor targeting (109-114)

C

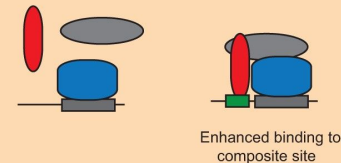
Allostery



Allosteric control of cofactor recruitment limits cofactor recruitment to only a subset of the TF binding sites (116-121, 124,125)

D

Cofactor-based targeting

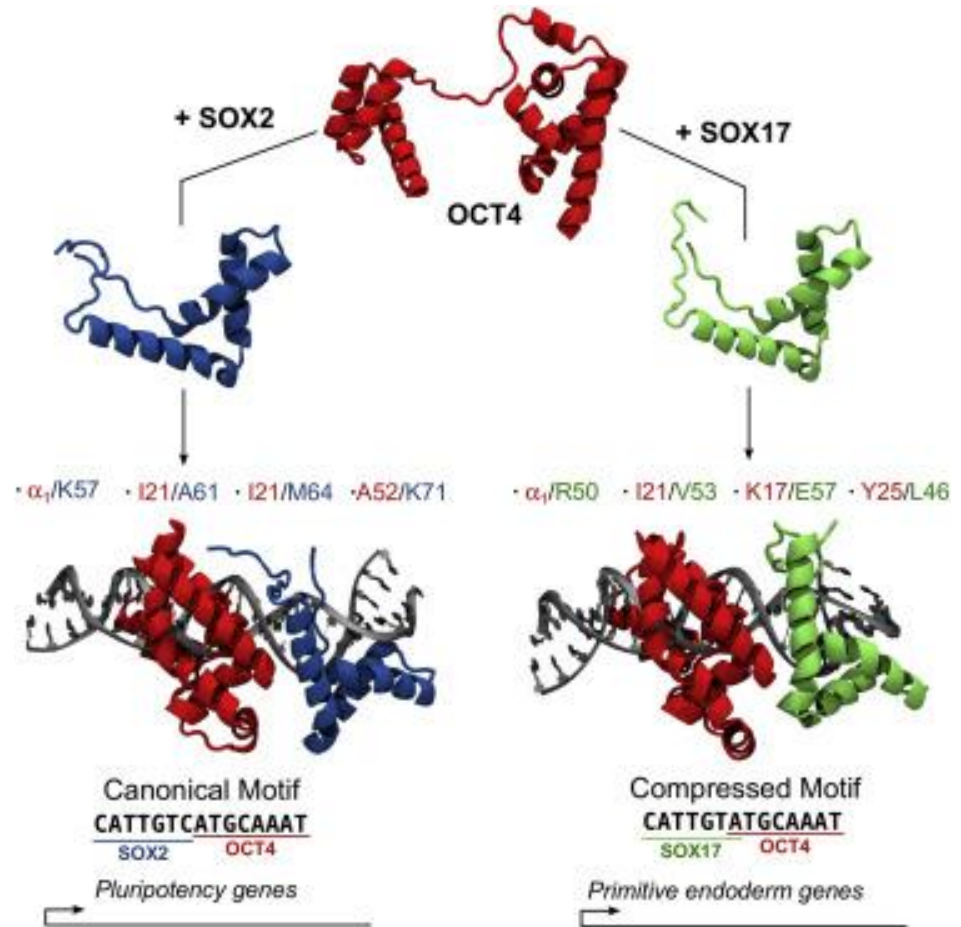


Enhanced binding of multi-protein complex to specialized composite sites is mediated by interactions between non-DNA-binding cofactor and an auxiliary motif (48,129)

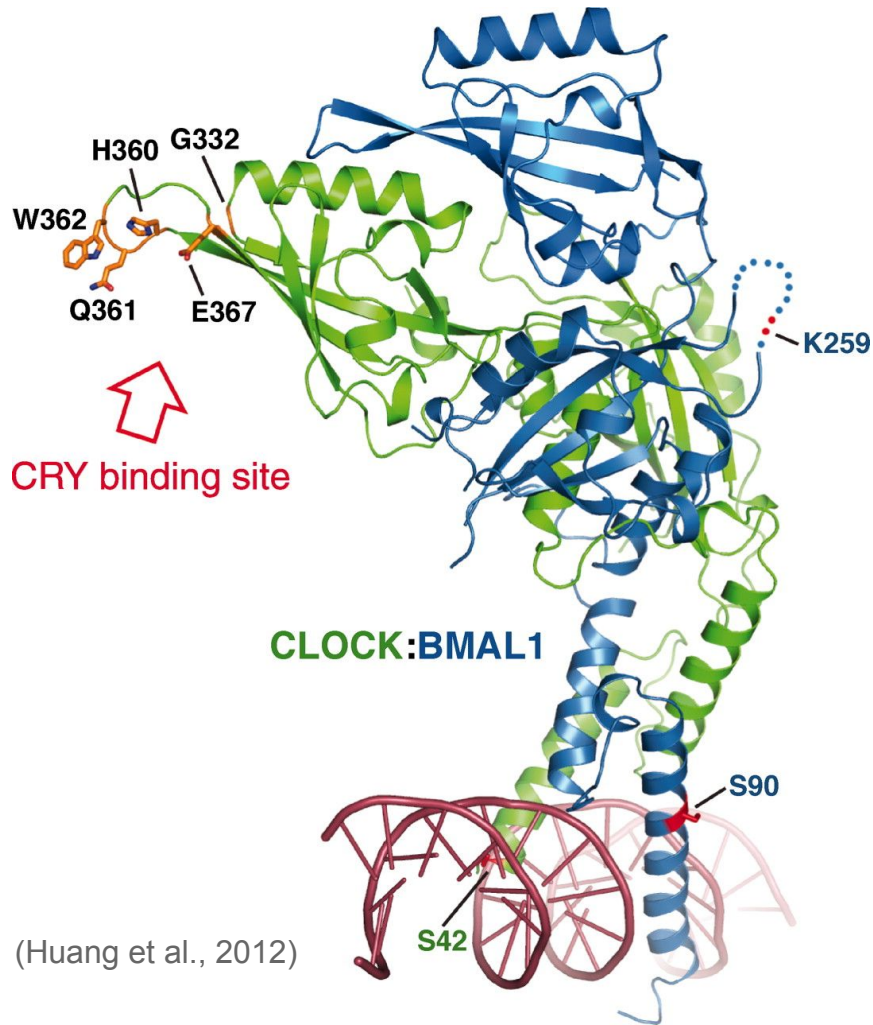


# Two examples of Cooperative binding

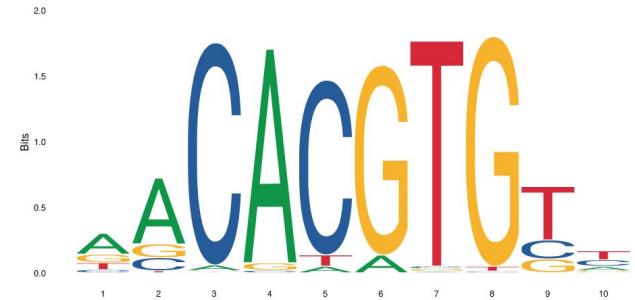
OCT4 (POU5f1) binding upon differentiation

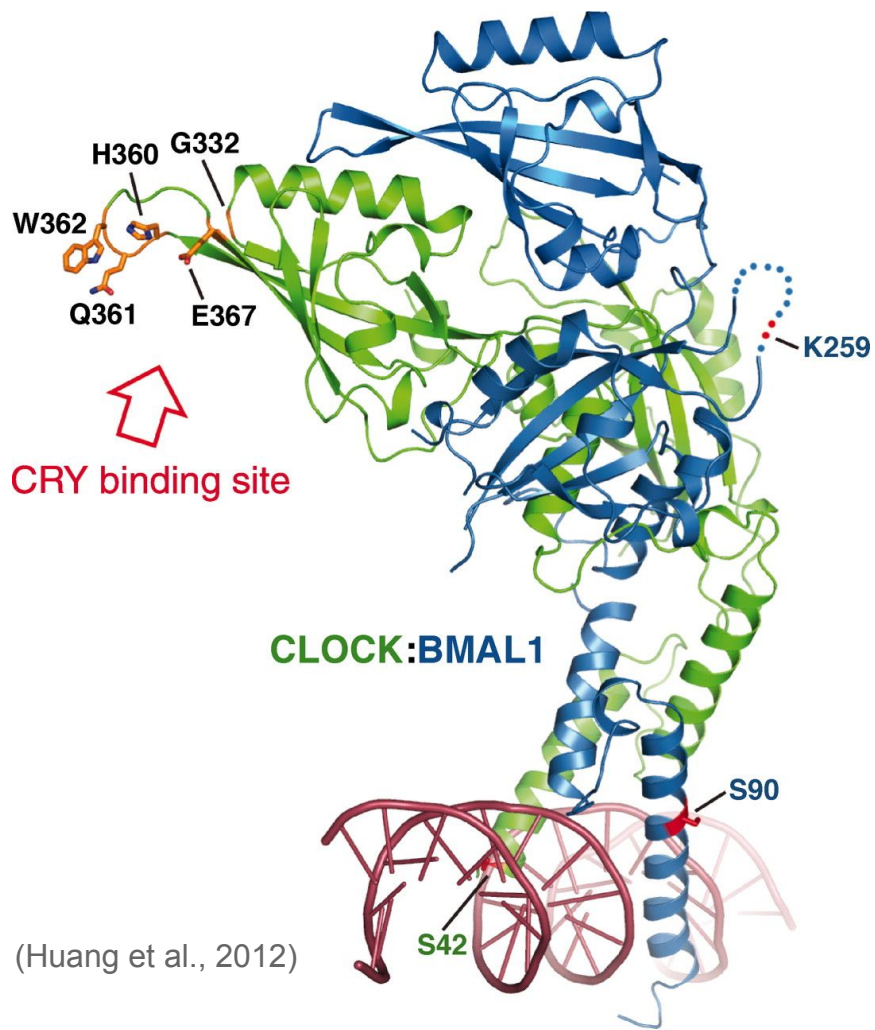


## Clock-Bmal-Cry during circadian rhythm

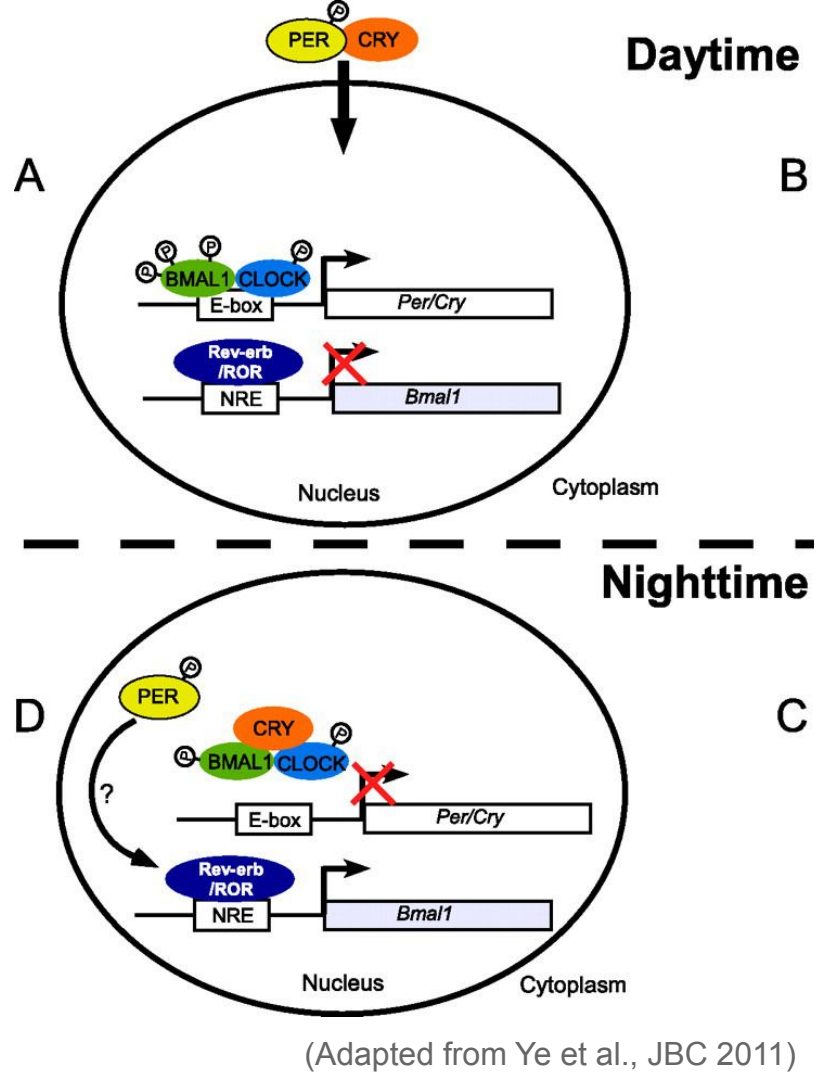


(Huang et al., 2012)





(Huang et al., 2012)





# Motif analysis

- **Motif discovery** aims at finding **new** motifs that are enriched in a set of sequences (e.g. peaks) versus a background
  - Example method: meme (Meme suite)
  - Bioconductor method: rGADEM package (see also the memes package)
- **Motif enrichment** analysis aims at finding **known** motifs that are enriched in a set of sequences (e.g. peaks) versus a background
  - Example method: AME (Meme suite)
  - Bioconductor method: PWMEnrich package
- **Motif scanning** aims at finding the **occurrences of known** motifs in a set of sequences (methodologically fairly simple – which method doesn't matter much)
  - Example method: fimo (Meme suite)
  - Bioconductor method: searchSeq function of the TFBSTools package

# Assignment

- Choose a transcription factor, e.g. CREB1, REST, GATA5, EGR1, GCR (or any of your choice that has a motif and available ChIPseq data)
- Download the (e.g. Mouse) peaks for that factor (whatever cell type)
- Identify the instances of the factor's motif
- Answer the following questions:
  - Of all the peaks, what proportion contains a motif for the factor?
    - Expected form of an answer: of the XX peaks, XX (XX%) contain a motif
  - Of all instances of that motif in the genome, what proportion is bound by the factor (i.e. has a peak)?
    - Expected form of an answer: of the XX motif instances, XX (XX%) overlap a peak

Don't forget to *render* your markdown and push it as [assignment.html](#) !