Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L - 2022 | week 06

Pierre-Luc Germain

ETH Zürich

Plan

- Quick things (see slack) :
  - Polls
  - New channel for discuss project ideas
  - New packages to install

- Debriefing on last week’s assignment

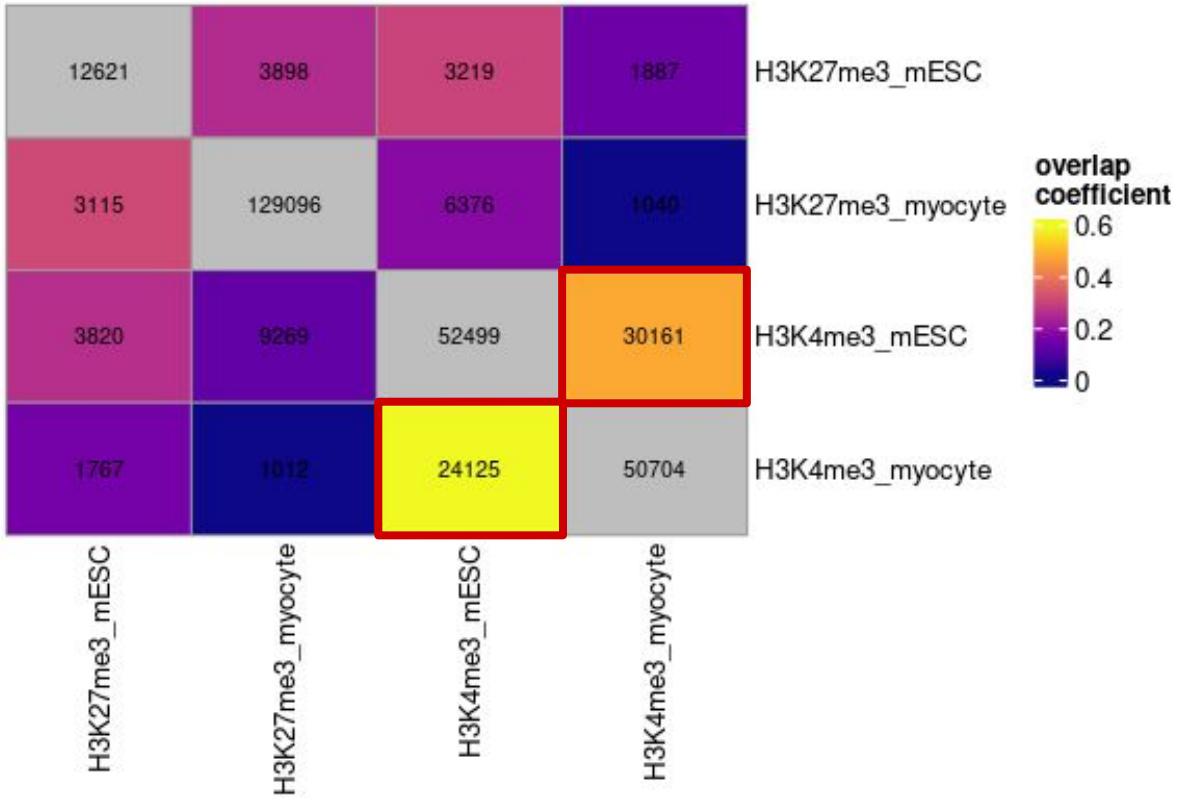
- Overview of transcription factors and their binding specificity

- Motifs and related analysis

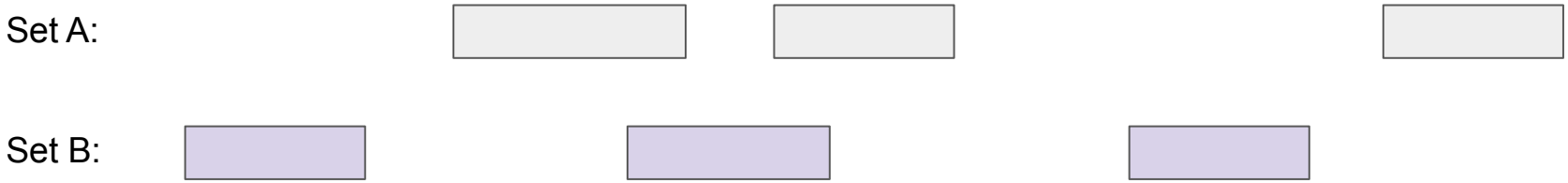
A few extra questions raised

- 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?

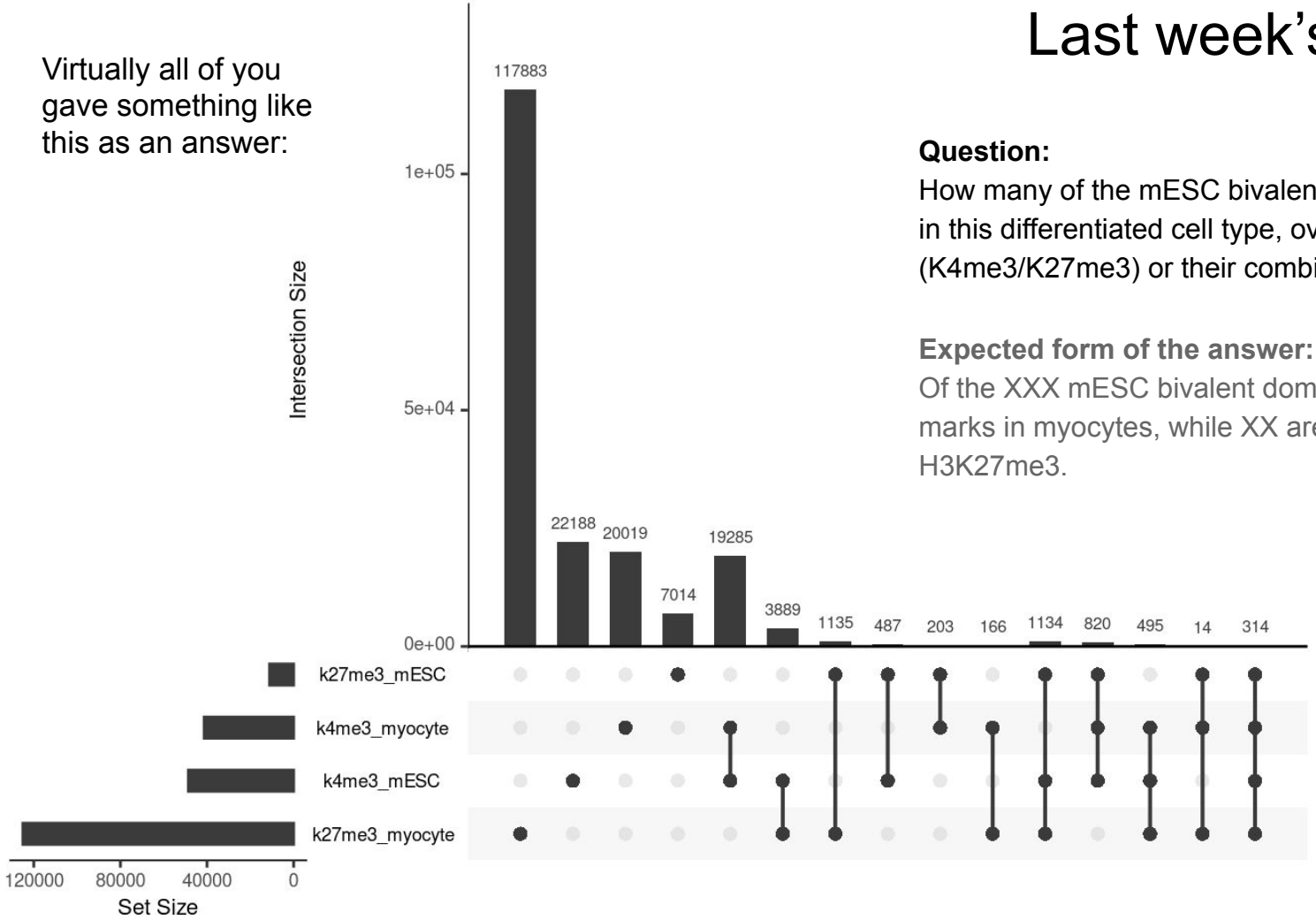


How many elements of A overlap elements of B? → 2/3

How many elements of B overlap elements of A? → 1/3

Last week's assignment

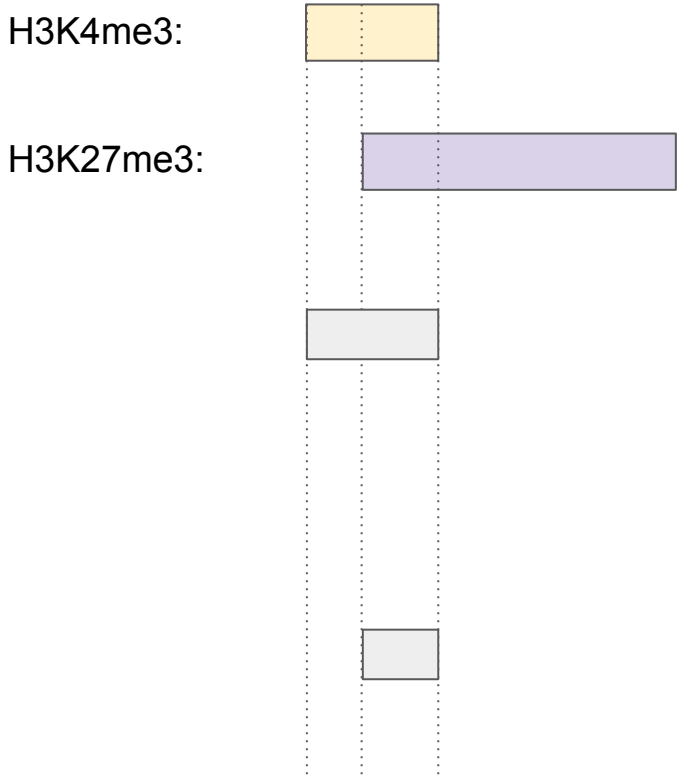
Virtually all of you gave something like this as an answer:



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

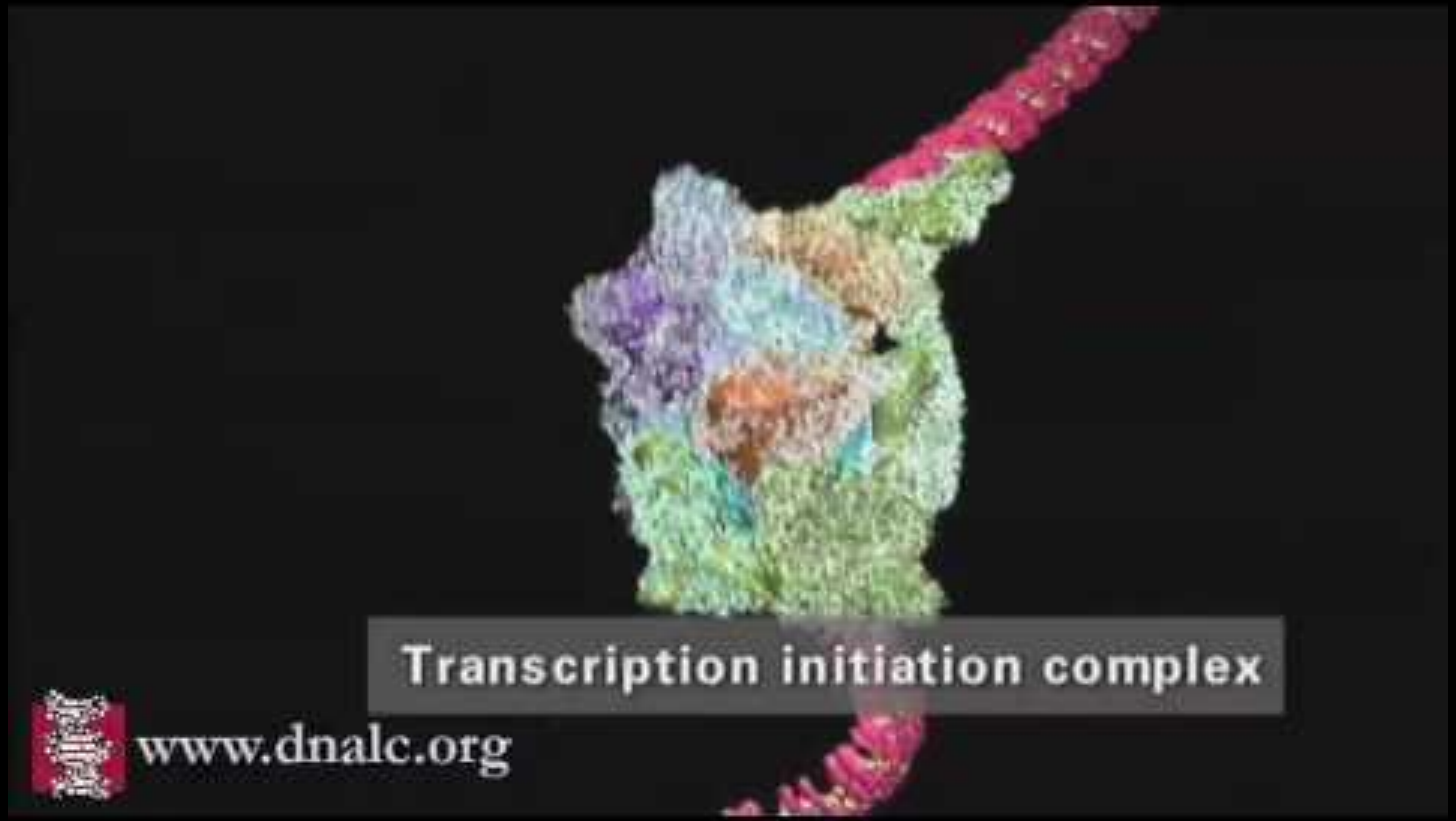
Expected form of the answer:
Of the XXX mESC bivalent domains, XX also have both marks in myocytes, while XX are H3K4me3 and XX are H3K27me3.

Intersection & overlap:
The example of bivalent domains



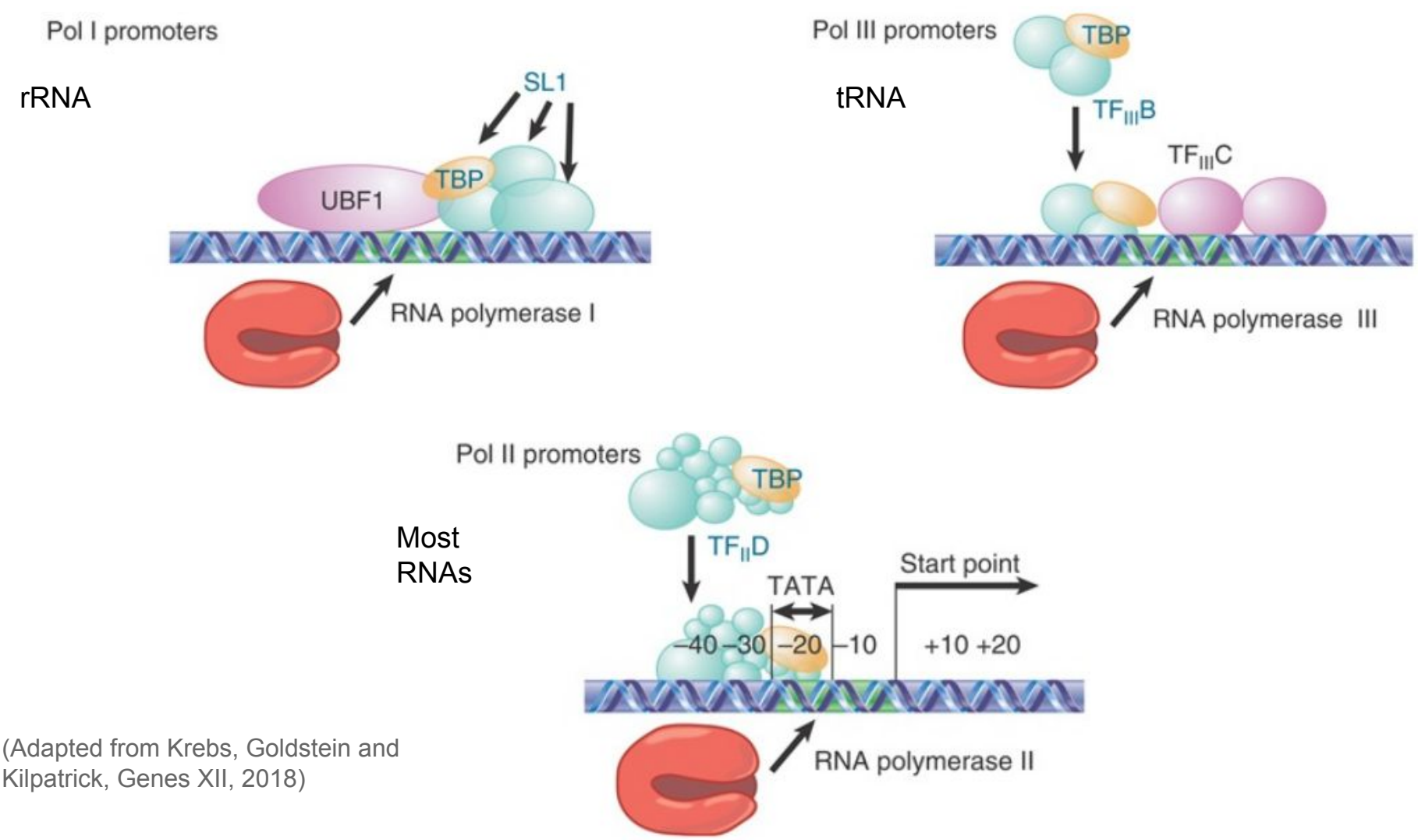
- **method one (overlapsAny):**
  find the H3K4me3 peaks that overlap a H3K27me3 domain

- **method two (intersect):**
  find the regions that are covered by both H3K4me3 and H3K27me3



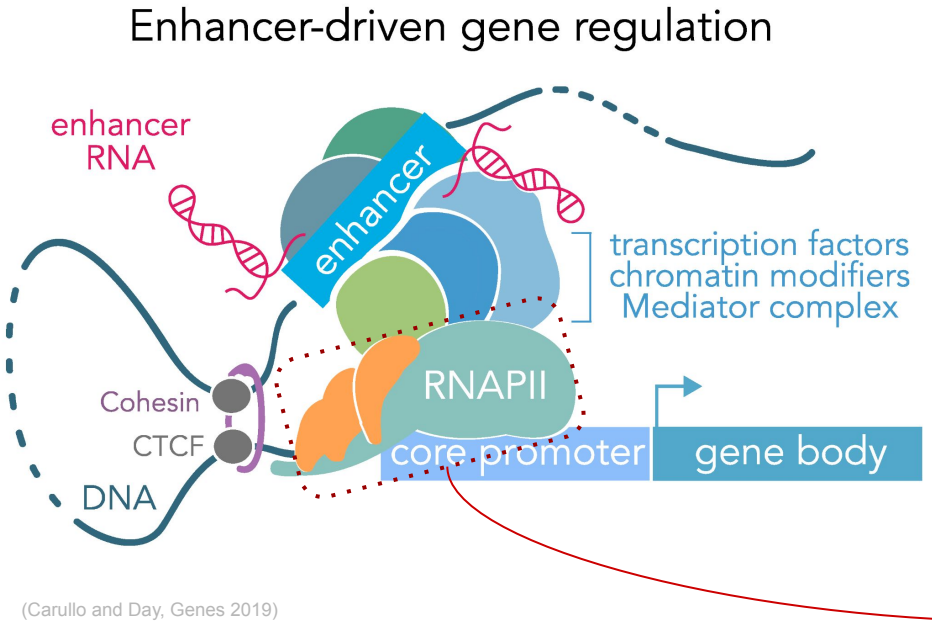
https://youtu.be/SMtWvDbfHLo

(See also https://youtu.be/WW9lIYM_FC0)

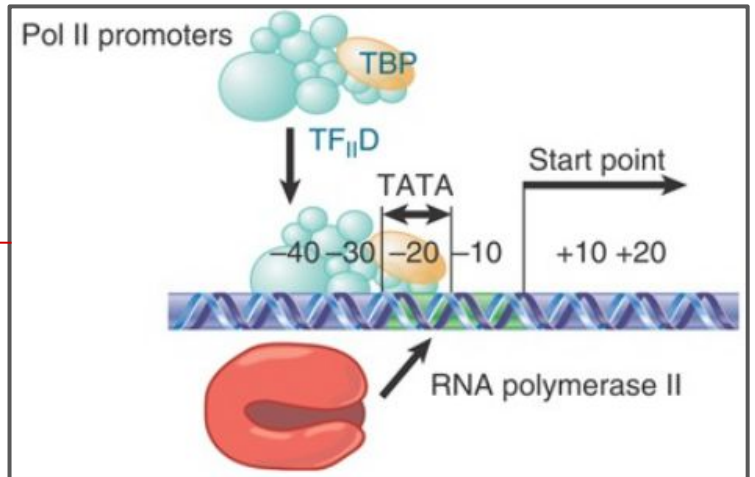


(Adapted from Krebs, Goldstein and Kilpatrick, Genes XII, 2018)

Additional regulatory elements



(Carullo and Day, Genes 2019)



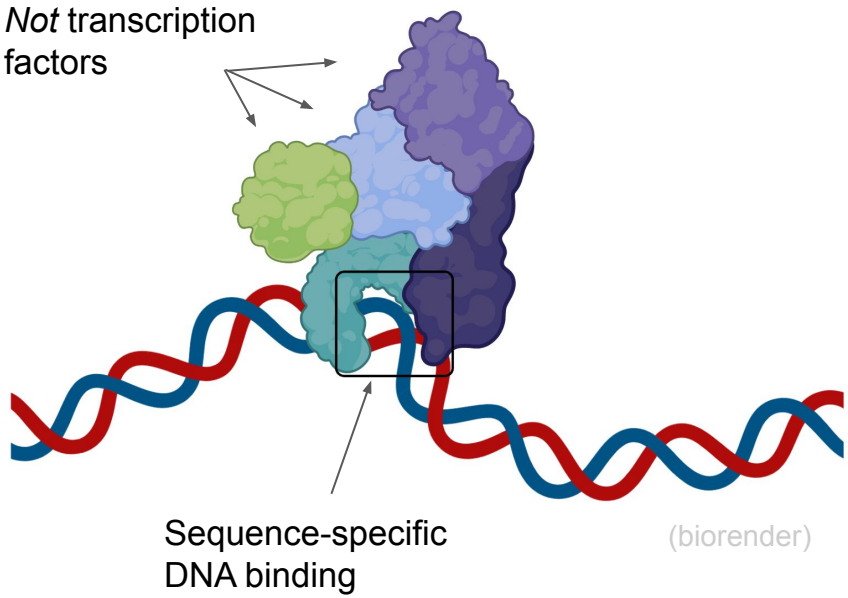
What is a transcription factor?

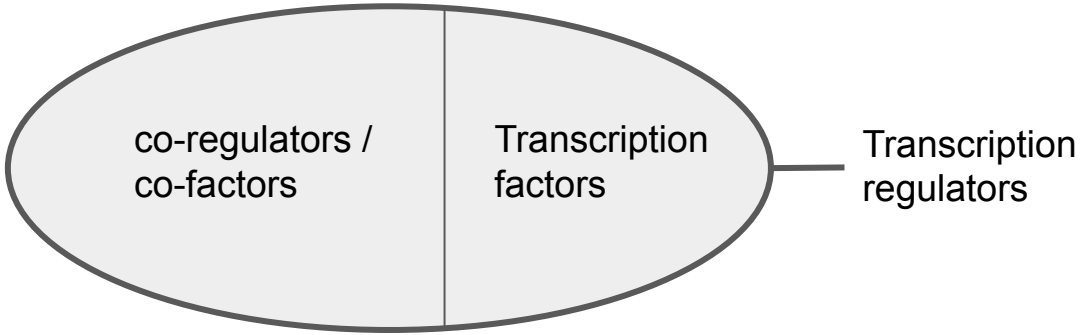
Proteins capable of both:

1) Binding DNA in a sequence-specific manner

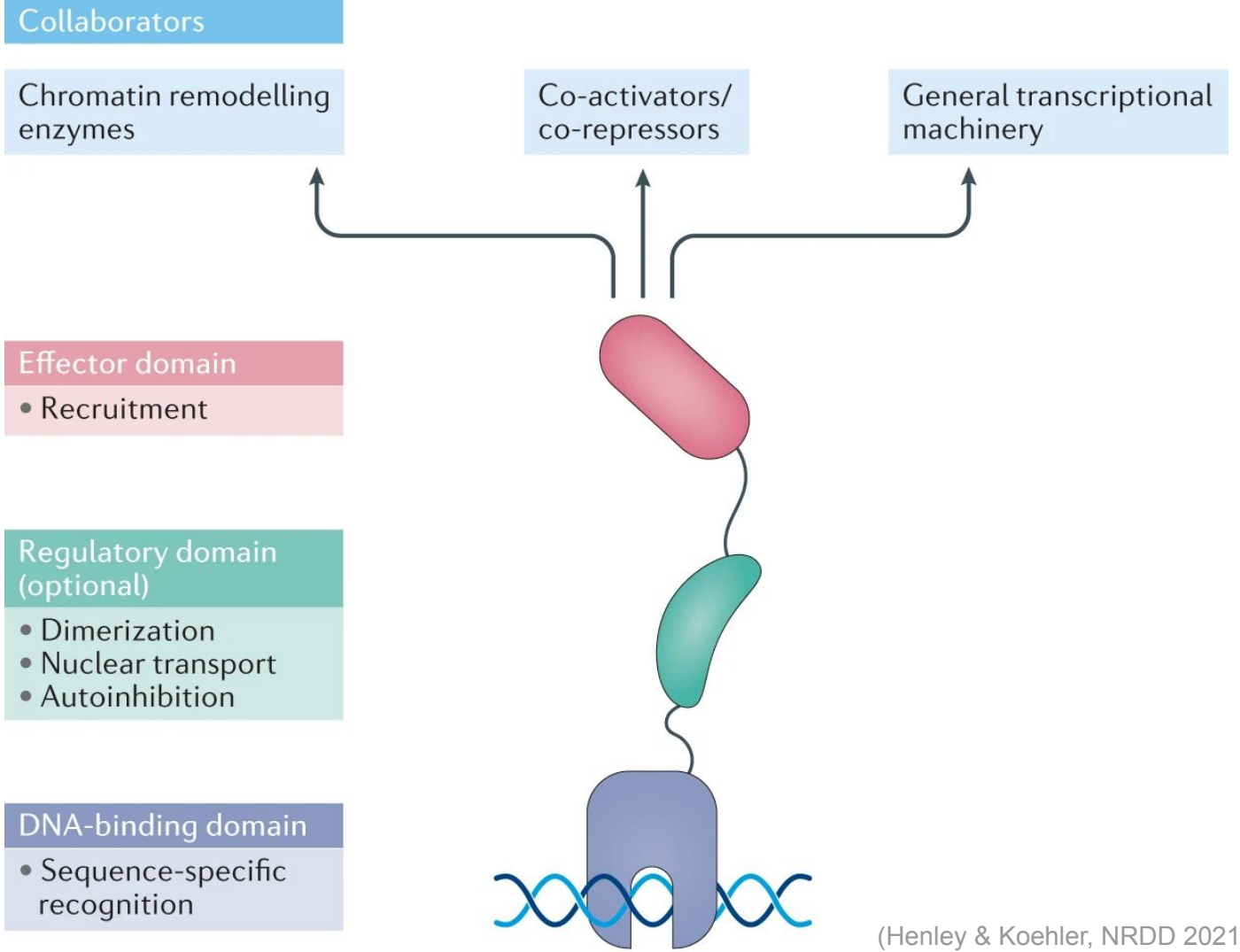
2) Regulating transcription

(Lambert et al., Cell 2018)

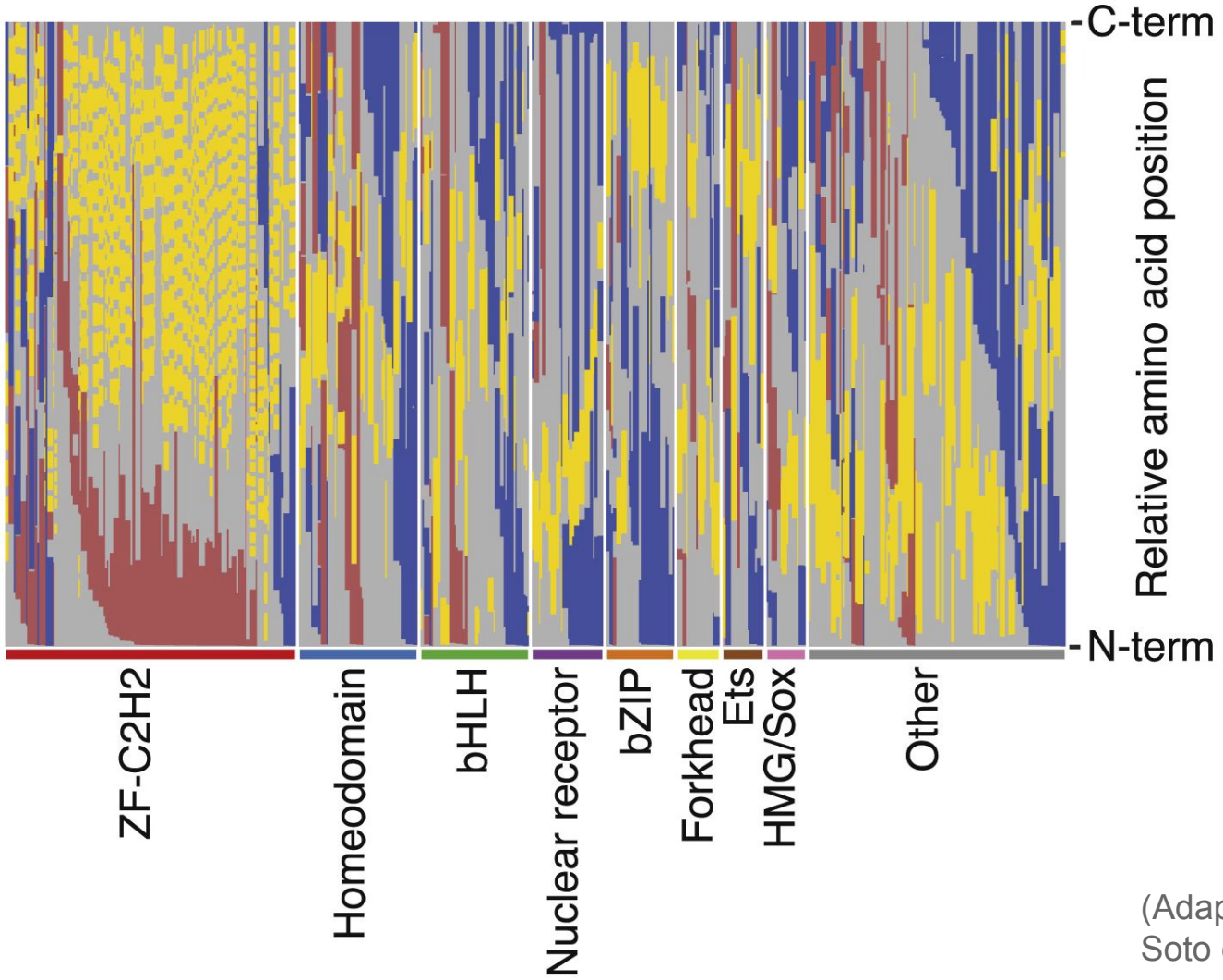




Anatomy of a transcription factor (TF)



(Henley & Koehler, NRDD 2021)



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

Domain
AD
RD
DBD
Other

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

Review (Cell 2018)

The Human Transcription Factors

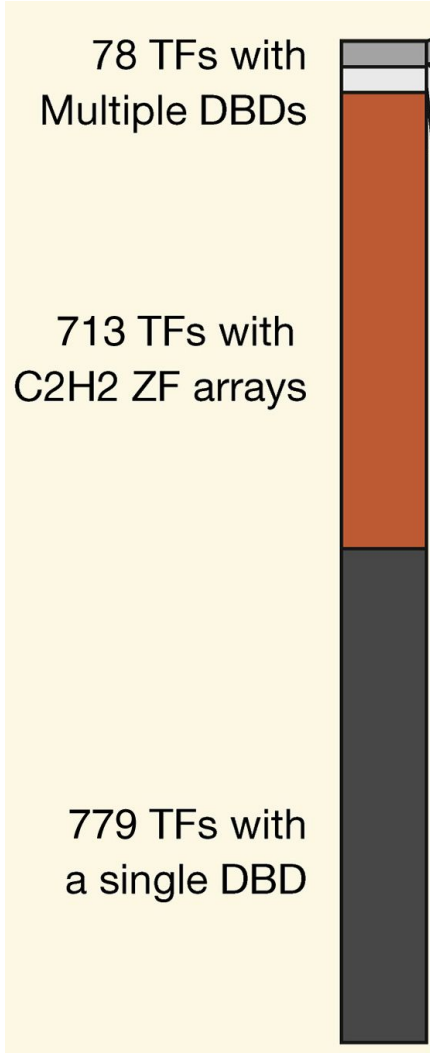
Samuel A. Lambert [1, 9], Arttu Jolma [2, 9], Laura F. Campitelli [1, 9], Pratyush K. Das [3], Yimeng Yin [4], Mihai Albu [2], Xiaoting Chen [5], Jussi Taipale [3, 4, 6] ✉, Timothy R. Hughes [1, 2] ✉, Matthew T. Weirauch [5, 7, 8] ✉

Proteins capable of both:

1) Binding DNA in a sequence-specific manner

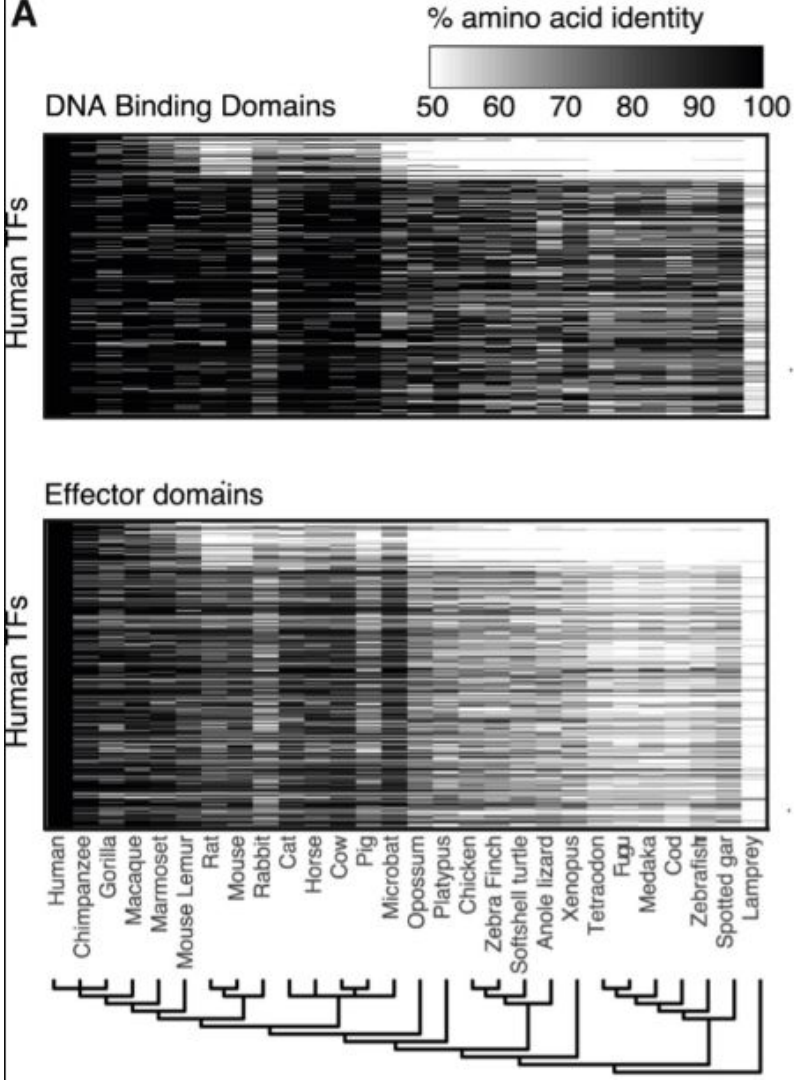
2) Regulating transcription

According to their census, humans have 1570 transcription factors

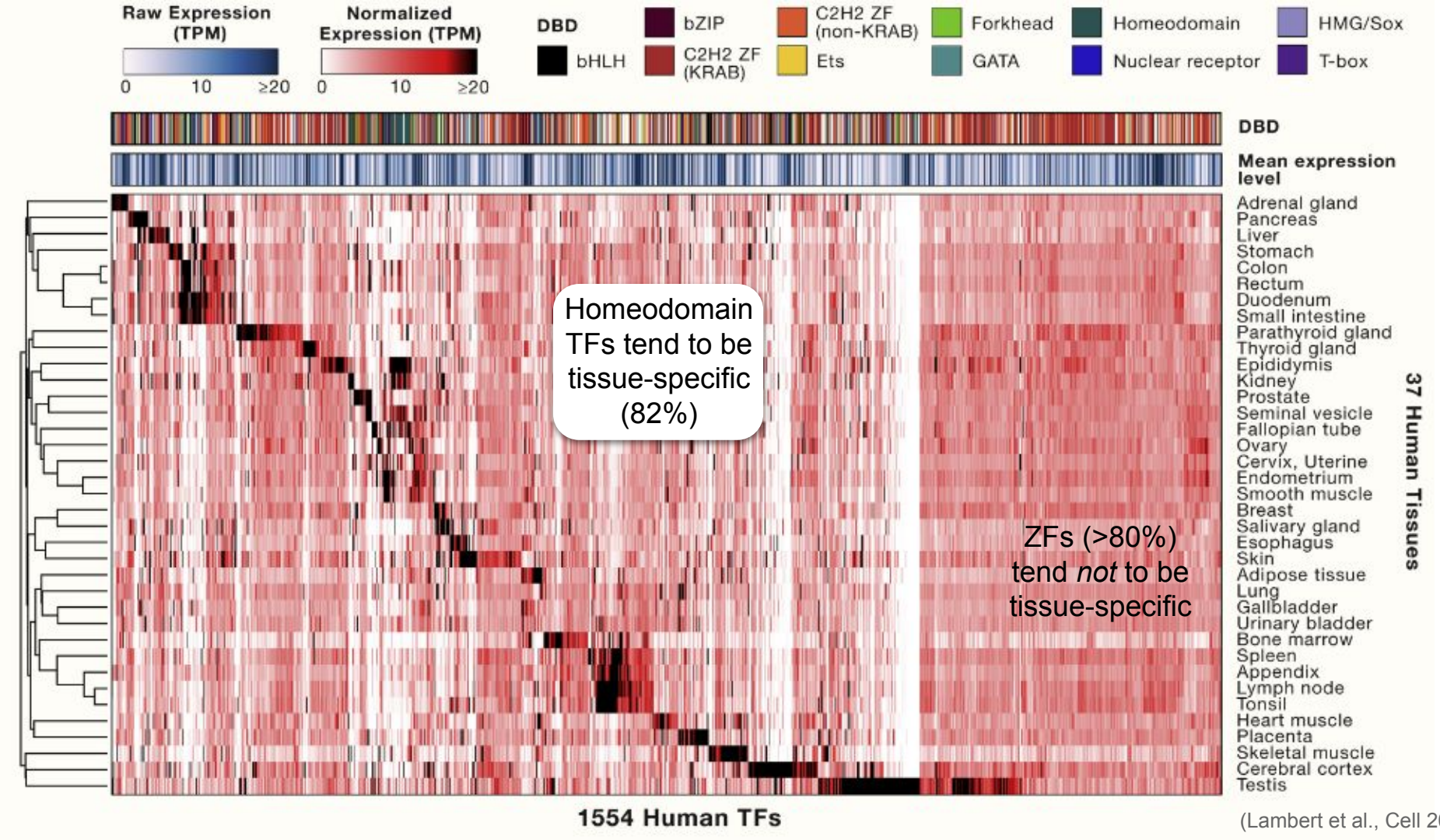


Transcription factors are highly conserved

DNA binding domains show much higher conservation than effector domains



(Soto et al., Molecular Cell 2021)

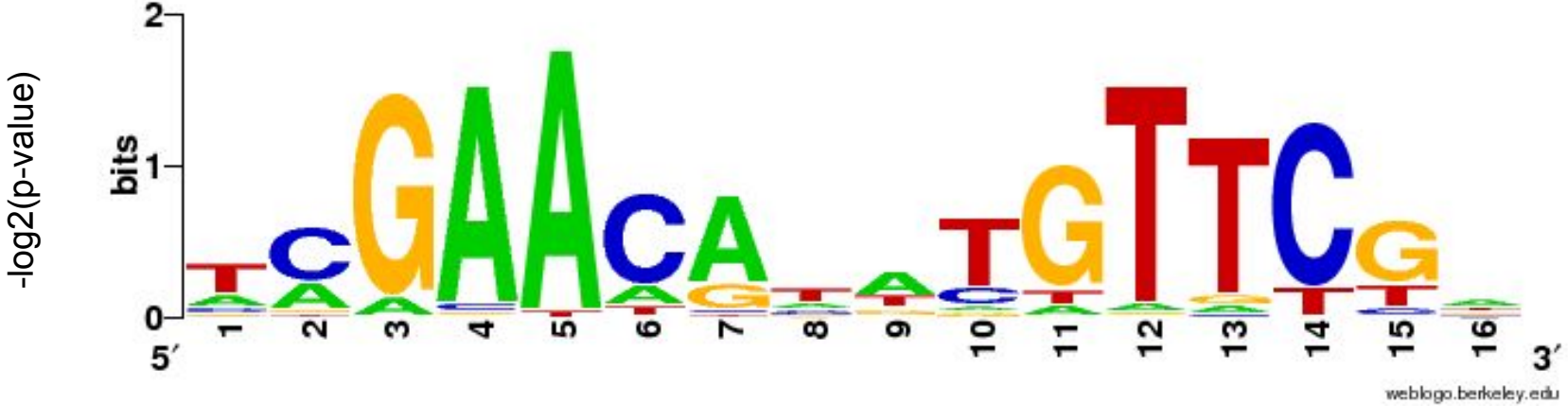


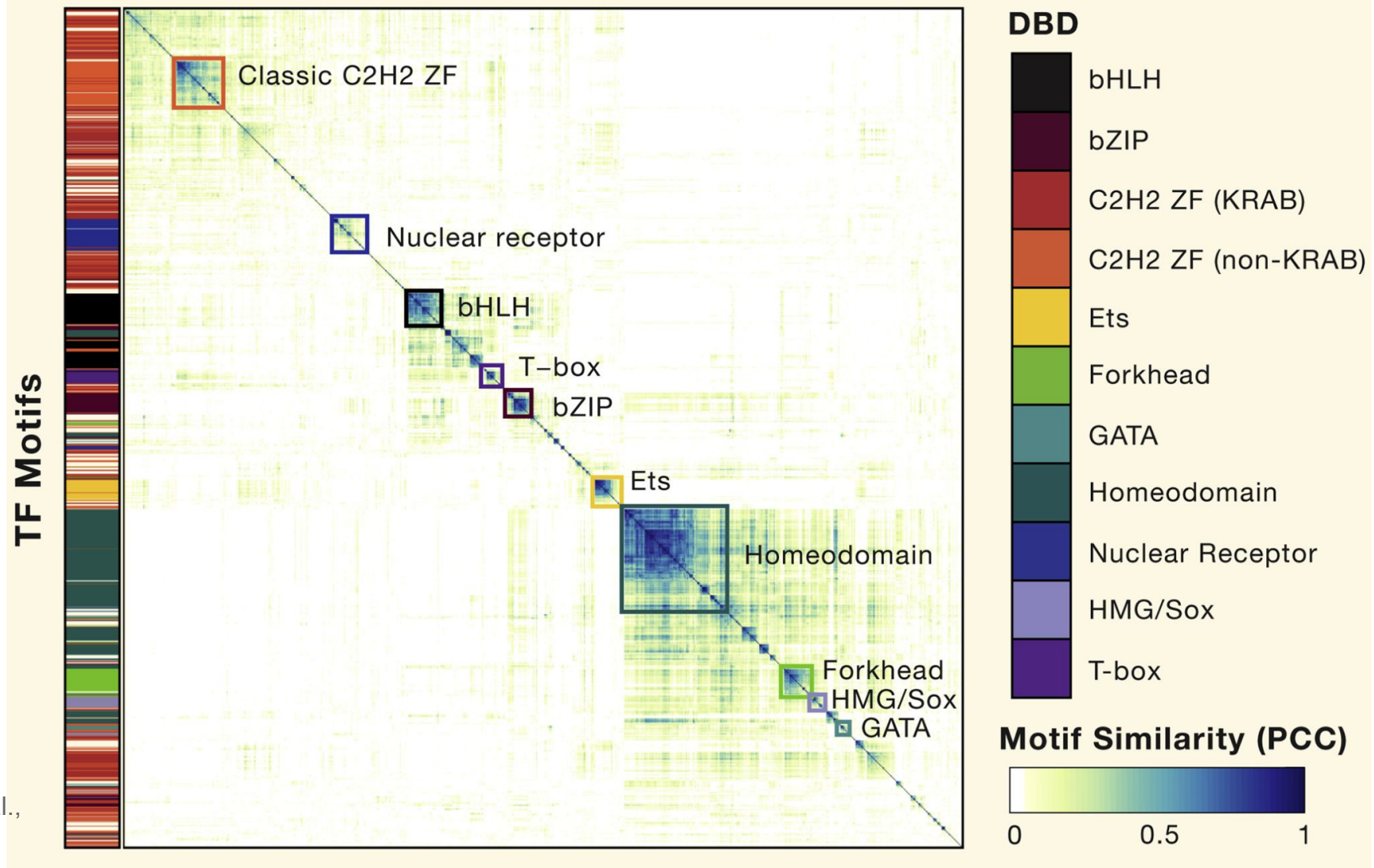
(Lambert et al., Cell 2018)

Sequence-specificity

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

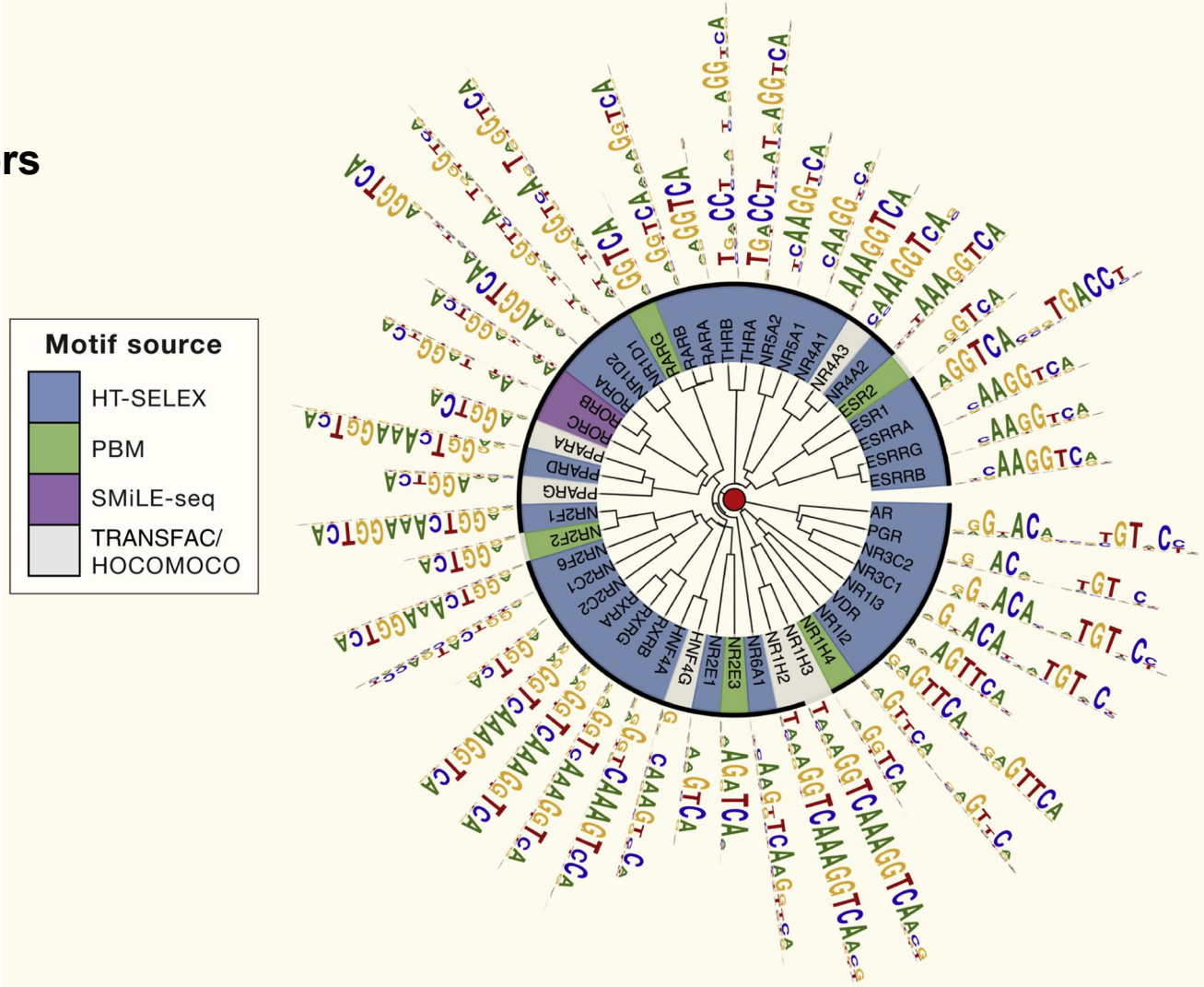
5'-GAACAnnTGTTC-3'





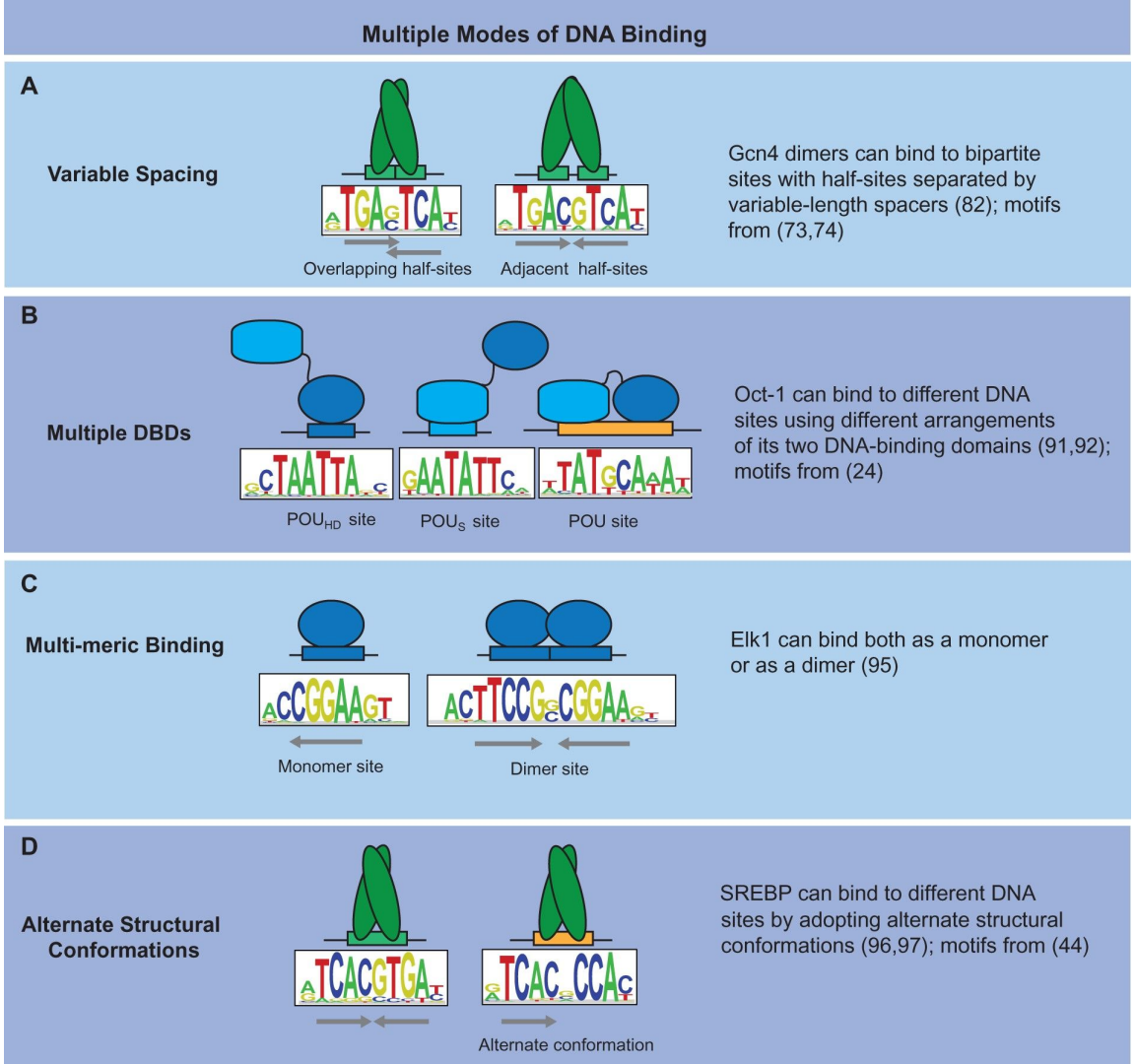
(Lambert et al., Cell 2018)

An example of TF motif degeneracy: Nuclear hormone receptors



(Lambert et al., Cell 2018)

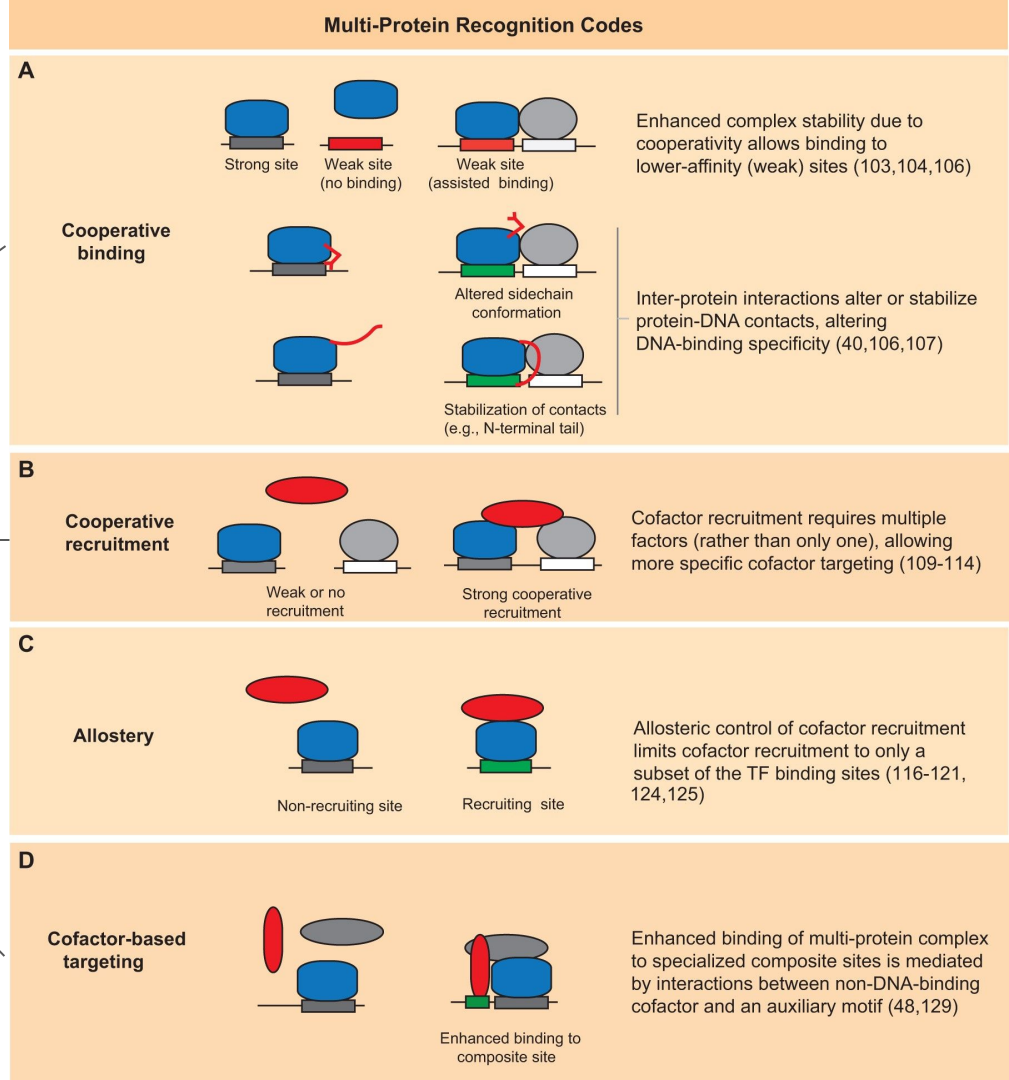
Variations in DNA binding specificity



(Siggers and Gordân, NAR 2014)

Cooperative binding

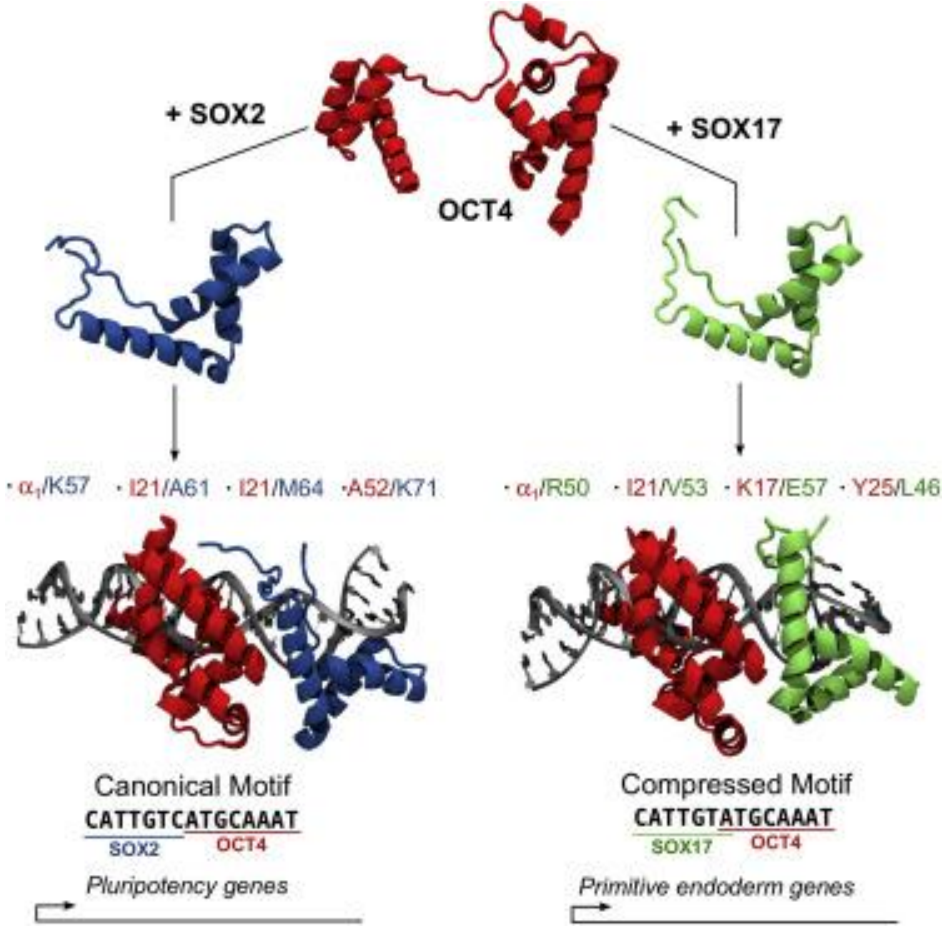
Highly combinatorial binding of TFs



(Siggers and Gordân, NAR 2014)

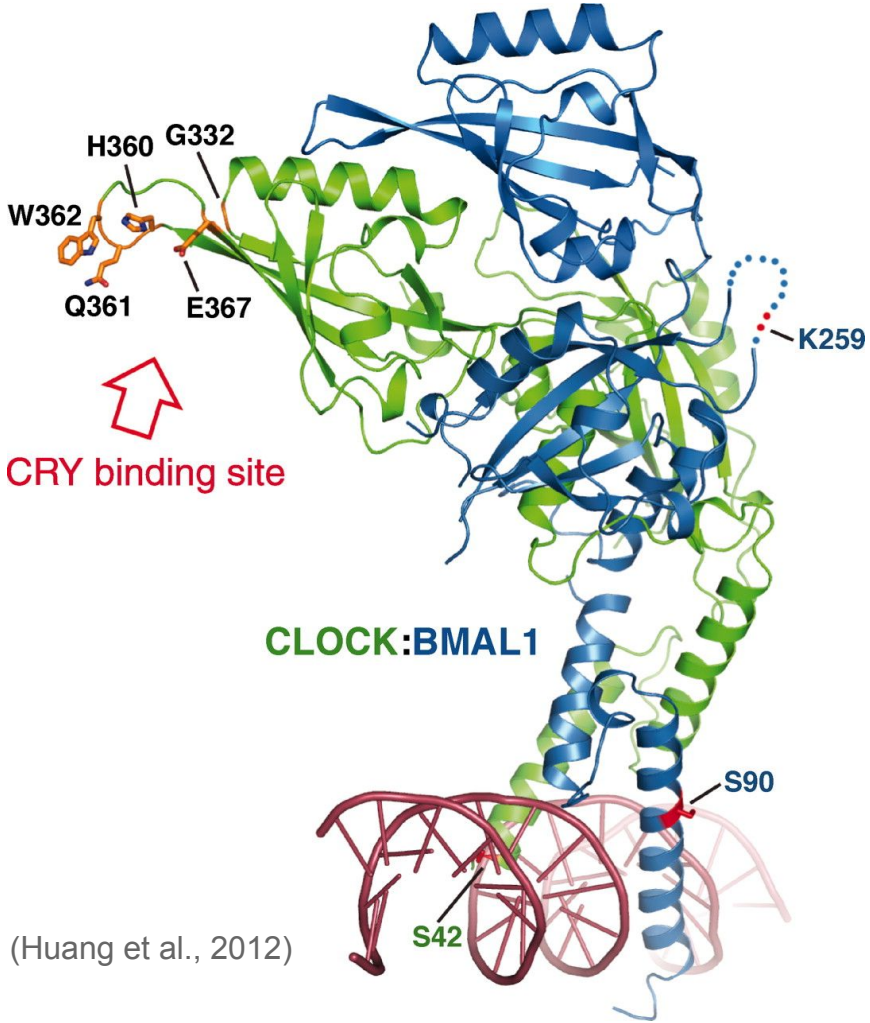
Two examples of Cooperative binding

OCT4 (POU5f1) binding upon differentiation

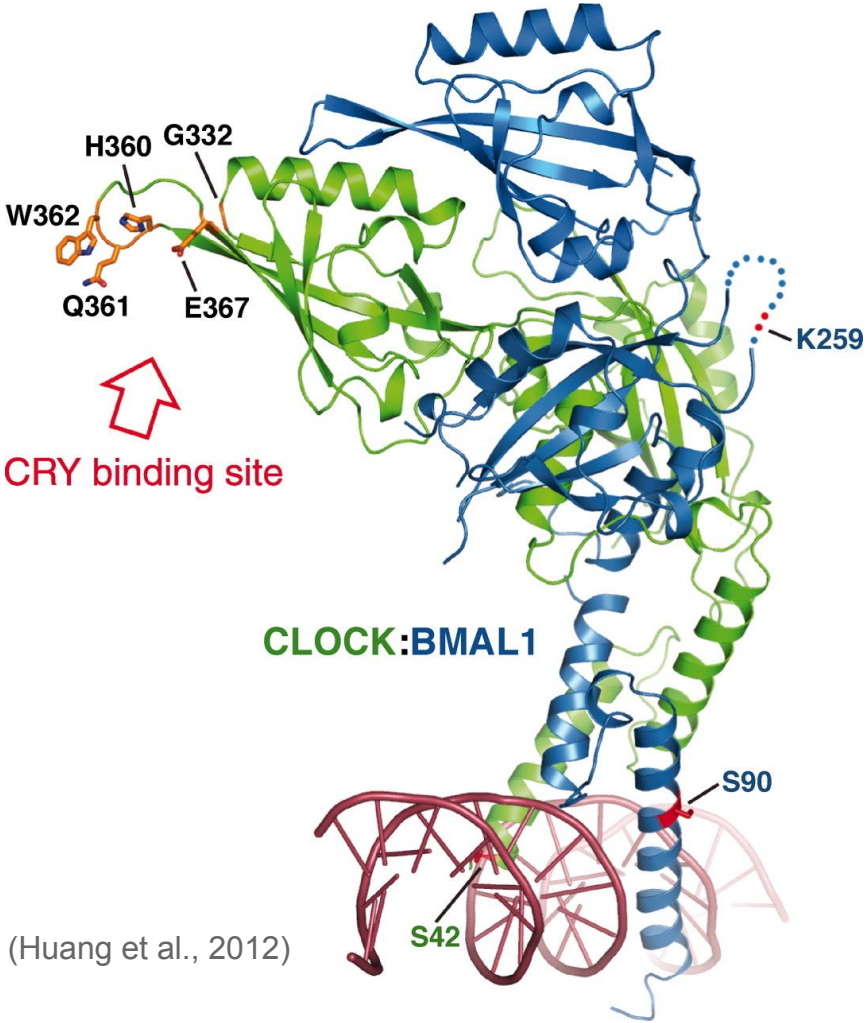


(Merino et al., *Structure* 2014)

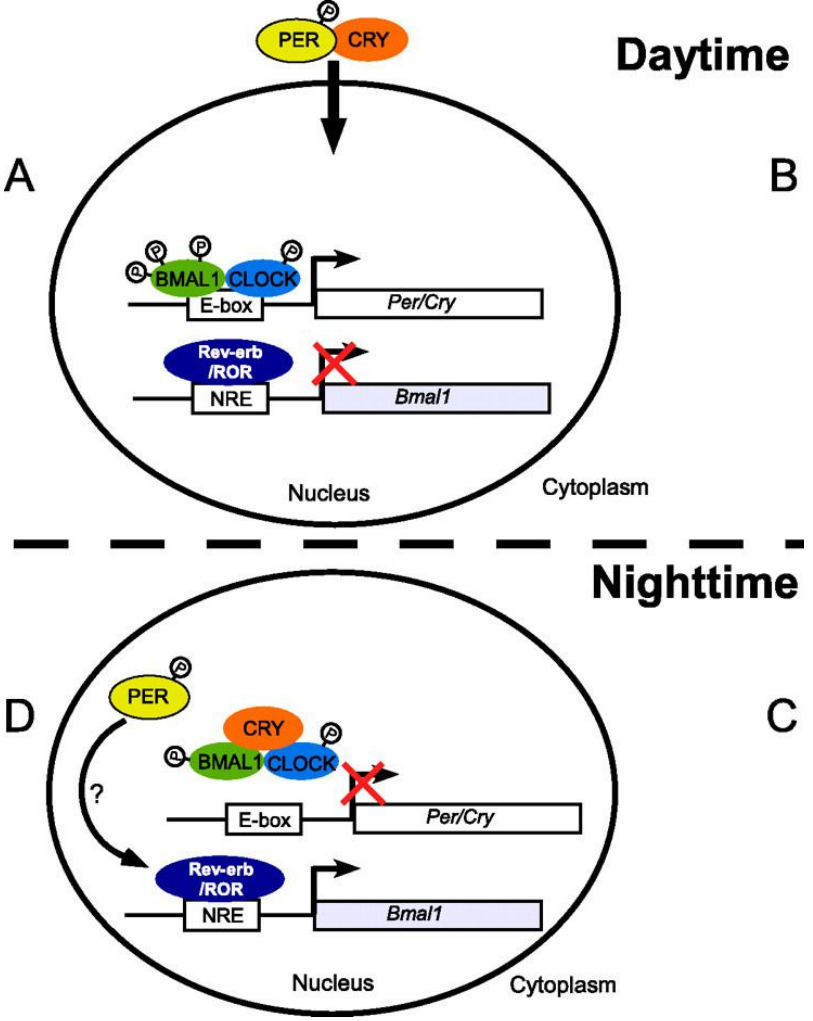
Clock-Bmal-Cry during circadian rythm



(Huang et al., 2012)



(Huang et al., 2012)



(Adapted from Ye et al., JBC 2011)

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` !