

Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L | week 06

Pierre-Luc Germain

Plan

- New packages to install (see slack)
- Debriefing on last week's assignment
- Overview of transcription factors and their binding specificity
- DNA motifs and related analysis

Recap

findOverlaps():

```
> gr1
GRanges object with 2 ranges and 0 metadata columns:
      seqnames      ranges strand
      <Rle> <IRanges> <Rle>
[1]   chr1      50-59      *
[2]   chr1      60-79      *
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths

> gr2
GRanges object with 2 ranges and 0 metadata columns:
      seqnames      ranges strand
      <Rle> <IRanges> <Rle>
[1]   chr1      50-55      *
[2]   chr1      57-59      *
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths

> ov <- findOverlaps(gr1,gr2)
> ov
Hits object with 2 hits and 0 metadata columns:
      queryHits subjectHits
      <integer> <integer>
[1]          1           1
[2]          1           2
-----
queryLength: 2 / subjectLength: 2

> gr1[queryHits(ov)]
GRanges object with 2 ranges and 0 metadata columns:
      seqnames      ranges strand
      <Rle> <IRanges> <Rle>
[1]   chr1      50-59      *
[2]   chr1      50-59      *
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths

> |
```

Recap

`findOverlaps()`:

Depending on what you
aim to do,
you do not want to have
the duplicates.

```
> gr1
GRanges object with 2 ranges and 0 metadata columns:
      seqnames      ranges strand
      <Rle> <IRanges> <Rle>
[1]      chr1      50-59      *
[2]      chr1      60-79      *
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths
> gr2
GRanges object with 2 ranges and 0 metadata columns:
      seqnames      ranges strand
      <Rle> <IRanges> <Rle>
[1]      chr1      50-55      *
[2]      chr1      57-59      *
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths
> ov <- findOverlaps(gr1,gr2)
> ov
Hits object with 2 hits and 0 metadata columns:
      queryHits subjectHits
      <integer>  <integer>
[1]           1           1
[2]           1           2
-----
queryLength: 2 / subjectLength: 2
> gr1[queryHits(ov)]
GRanges object with 2 ranges and 0 metadata columns:
      seqnames      ranges strand
      <Rle> <IRanges> <Rle>
[1]      chr1      50-59      *
[2]      chr1      50-59      *
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths
> |
```

Recap

use either, depending
on the aim, `unique()`
or
`overlapsAny()` or
`subsetByOverlaps()`

```
> gr1
GRanges object with 2 ranges and 0 metadata columns:
      seqnames      ranges strand
   <Rle> <IRanges> <Rle>
[1]   chr1      50-59      *
[2]   chr1      60-79      *
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths
> gr2
GRanges object with 2 ranges and 0 metadata columns:
      seqnames      ranges strand
   <Rle> <IRanges> <Rle>
[1]   chr1      50-55      *
[2]   chr1      57-59      *
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths
> ov <- findOverlaps(gr1,gr2)
> ov
Hits object with 2 hits and 0 metadata columns:
      queryHits subjectHits
   <integer>   <integer>
[1]         1           1
[2]         1           2
-----
queryLength: 2 / subjectLength: 2
> gr1[queryHits(ov)]
GRanges object with 2 ranges and 0 metadata columns:
      seqnames      ranges strand
   <Rle> <IRanges> <Rle>
[1]   chr1      50-59      *
[2]   chr1      50-59      *
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths
> gr1[unique(queryHits(ov))]
GRanges object with 1 range and 0 metadata columns:
      seqnames      ranges strand
   <Rle> <IRanges> <Rle>
[1]   chr1      50-59      *
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

Finding bivalent domains

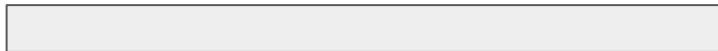
H3K4me3 peak



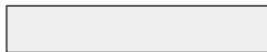
H3K27me3 peak



`reduce (C (H3K4me3, H3K27me3))`



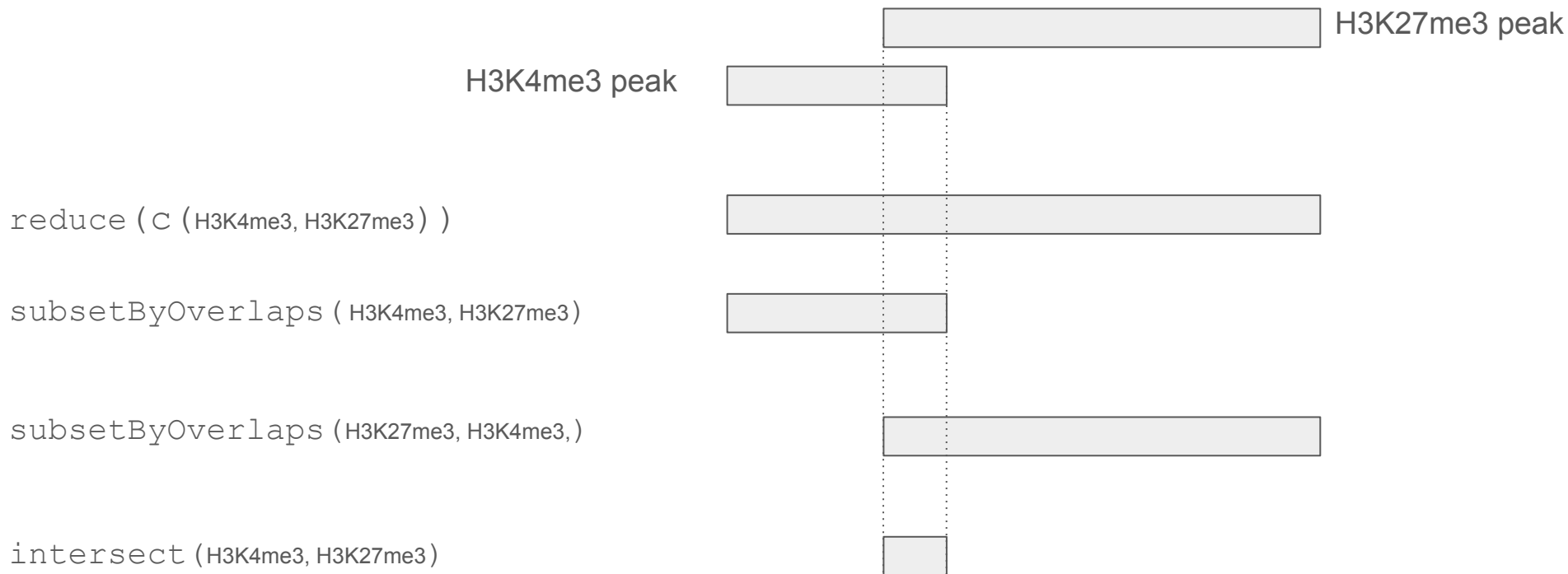
`subsetByOverlaps (H3K4me3, H3K27me3)`



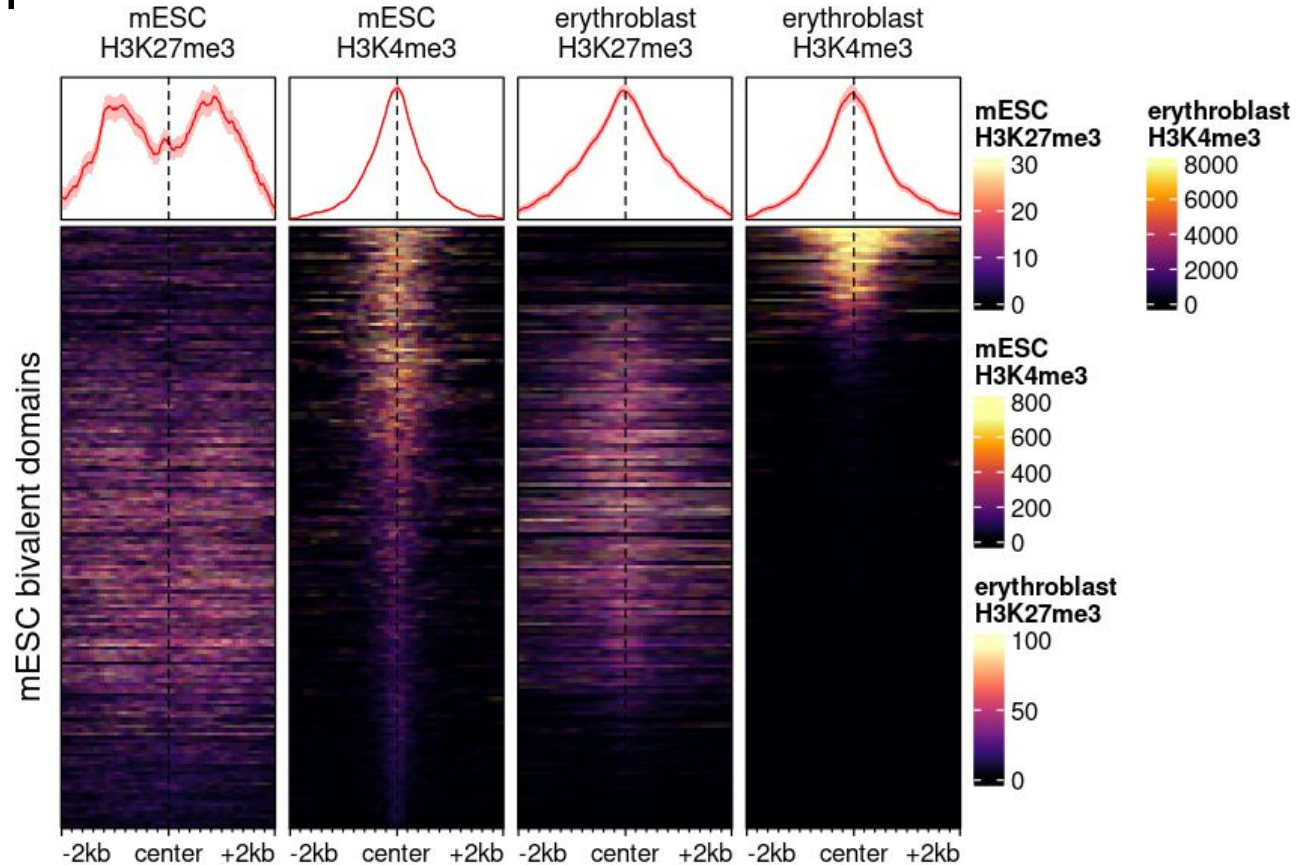
`subsetByOverlaps (H3K27me3, H3K4me3,)`



Finding bivalent domains



Embryonic bivalent domains binarize into active and inactive upon differentiation





Transcription initiation complex



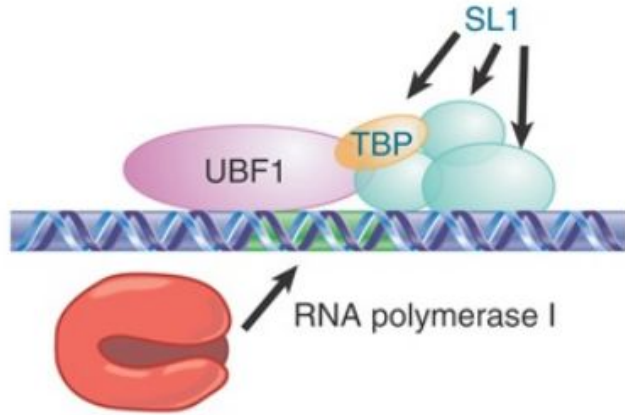
www.dnalc.org

<https://youtu.be/SMtWvDbfHLo>

(See also https://youtu.be/WW9IIYM_FC0)

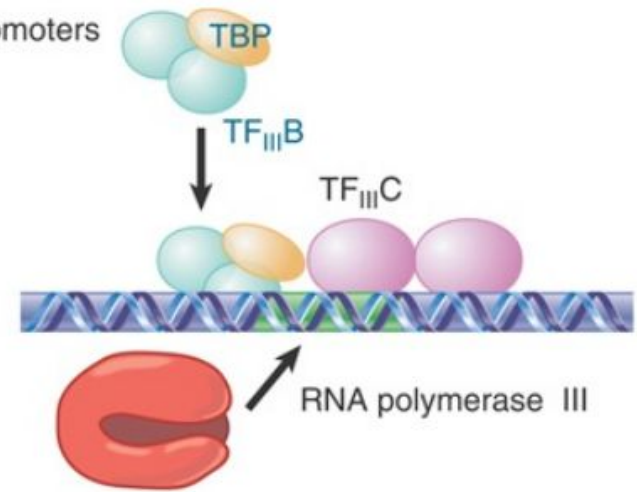
Pol I promoters

rRNA



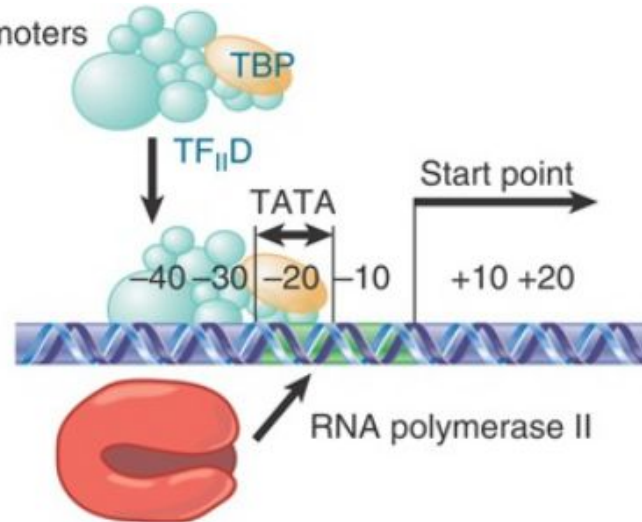
Pol III promoters

tRNA



Pol II promoters

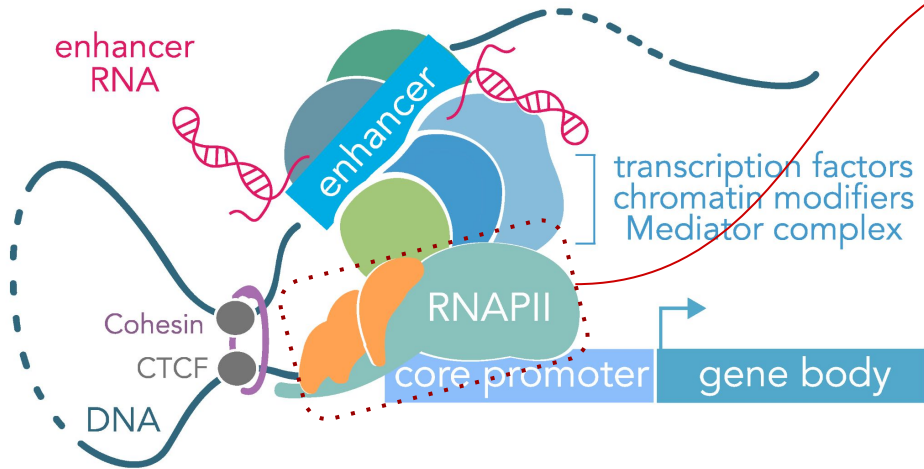
Most
RNAs



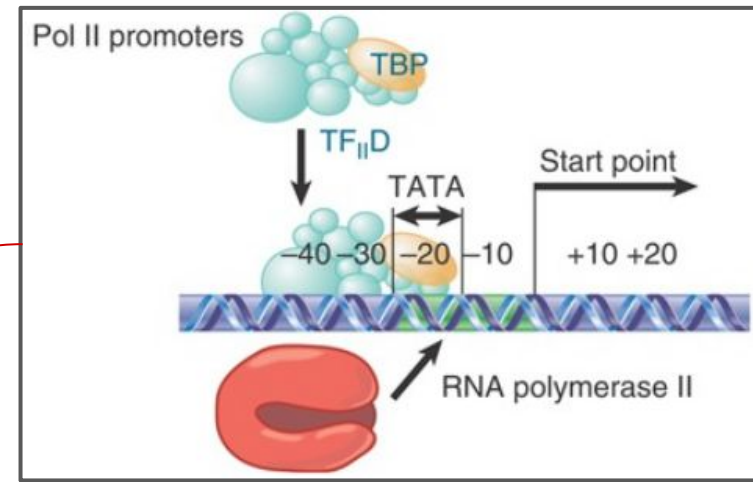
(Adapted from Krebs, Goldstein and Kilpatrick, Genes XII, 2018)

Additional regulatory elements

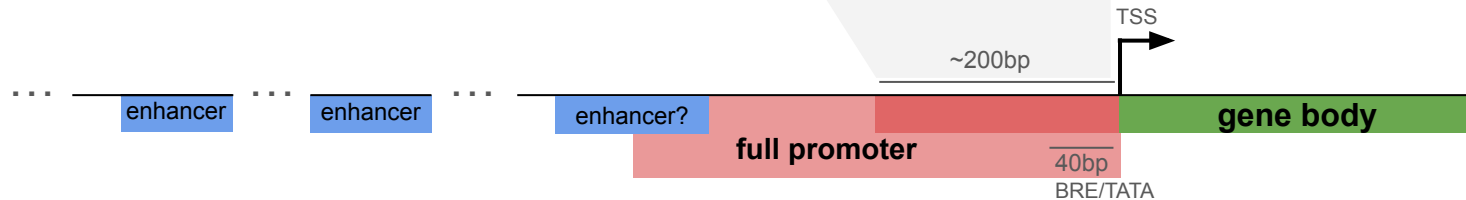
Enhancer-driven gene regulation



(Carullo and Day, Genes 2019)



"function as non-cell-type-specific 'on switches' providing similar expression levels to their associated gene"
([Agarwal et al., biorxiv 2023](#))

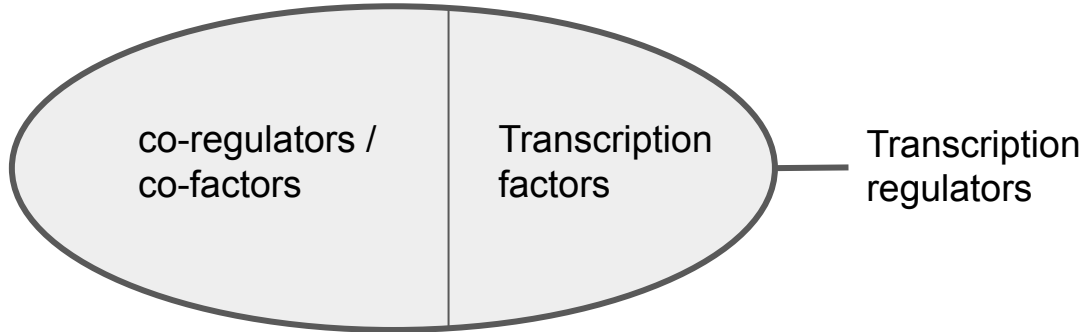
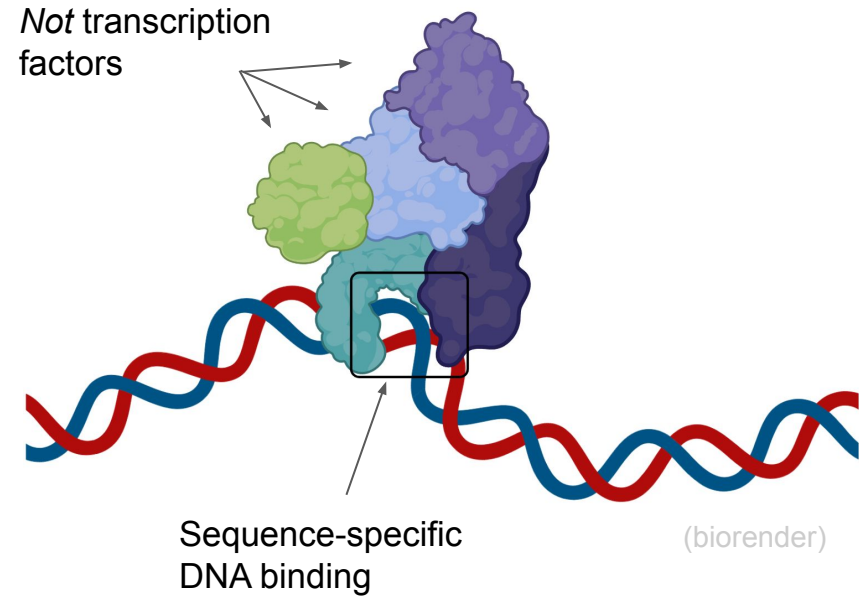


What is a transcription factor?

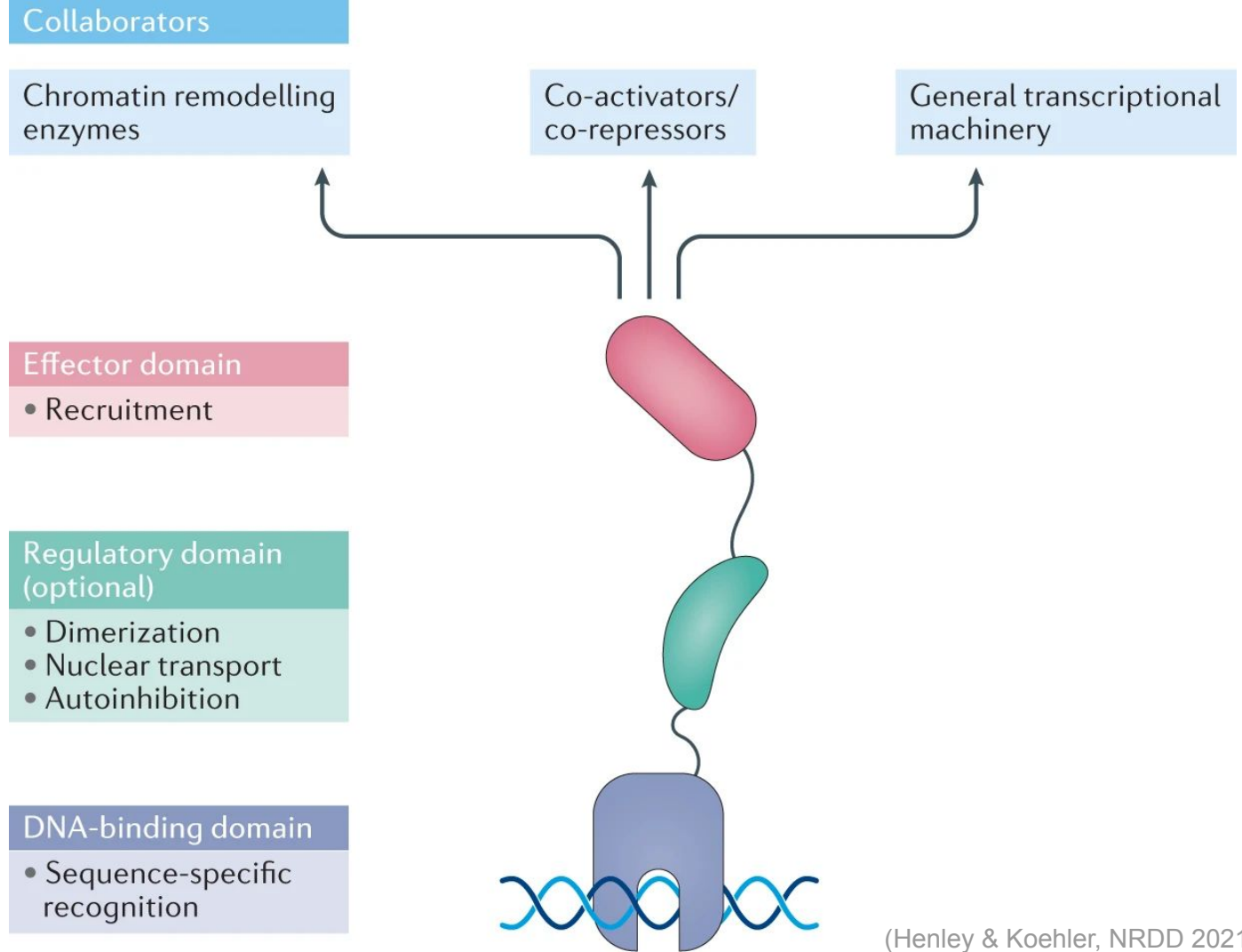
Proteins capable of both:

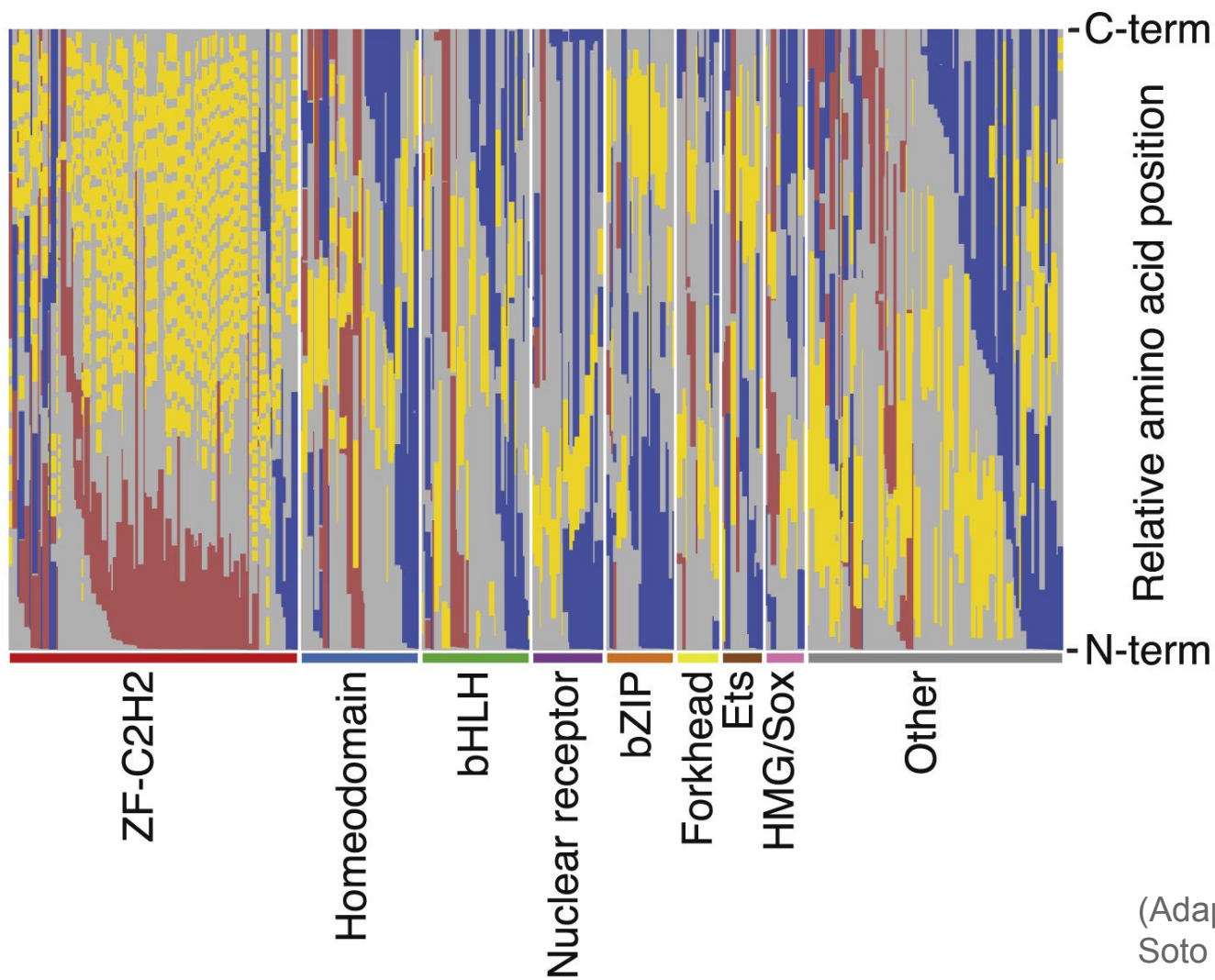
- 1) Binding DNA in a sequence-specific manner
- 2) Regulating transcription

(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

Domain

- AD
- RD
- DBD
- Other

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

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,*} and 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

78 TFs with
Multiple DBDs

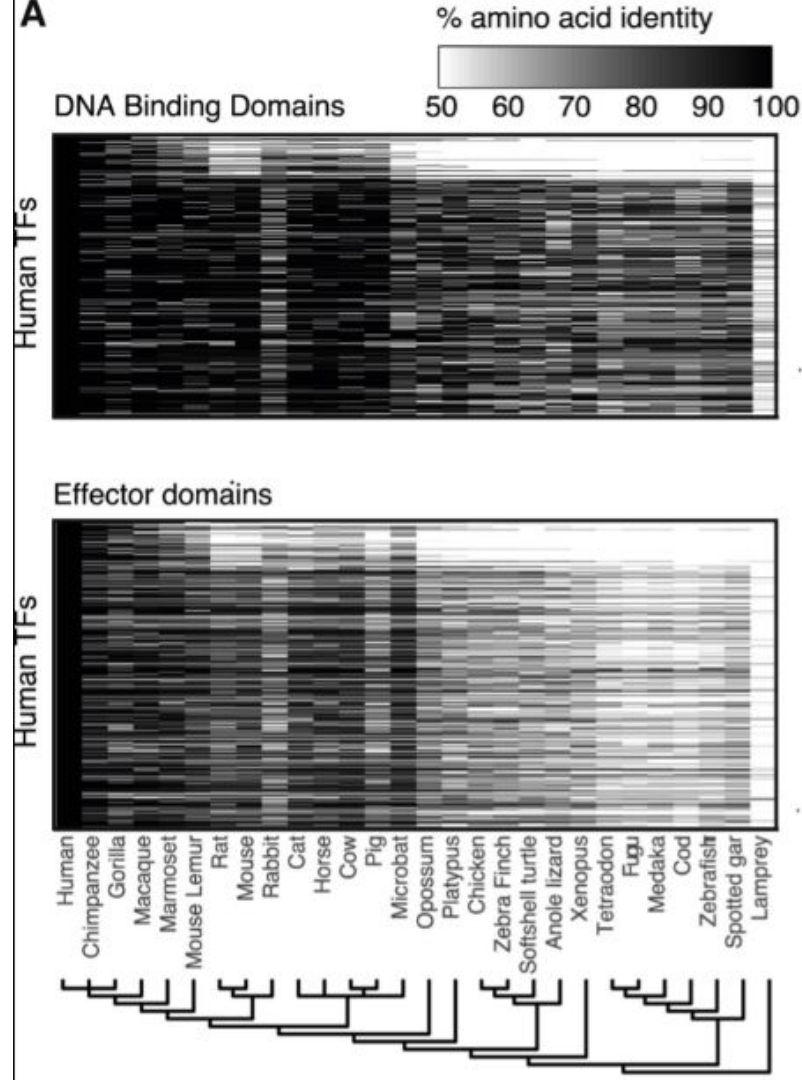
713 TFs with
C2H2 ZF arrays

779 TFs with
a single DBD

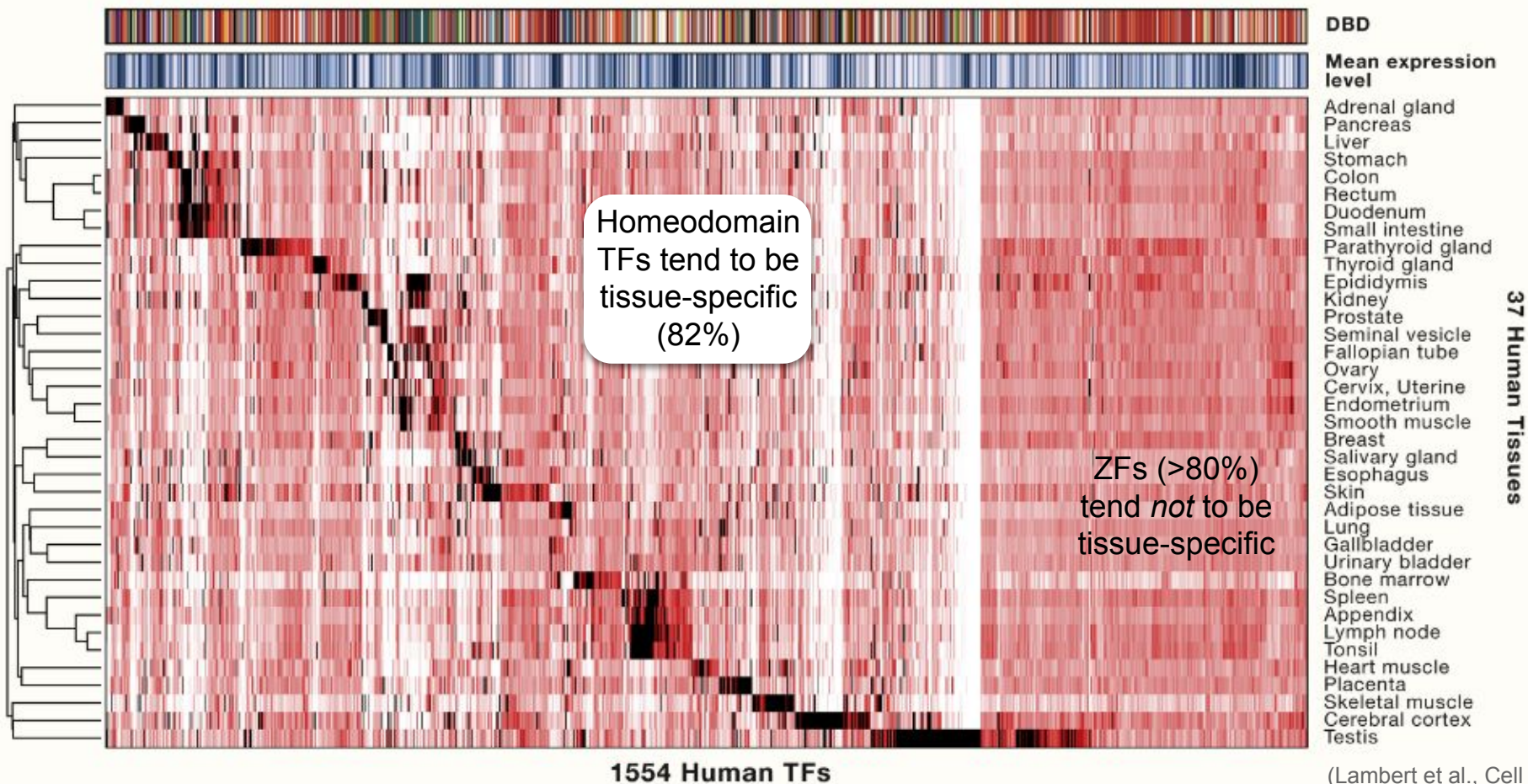


Transcription factors are highly conserved

DNA binding domains show much higher conservation than effector domains



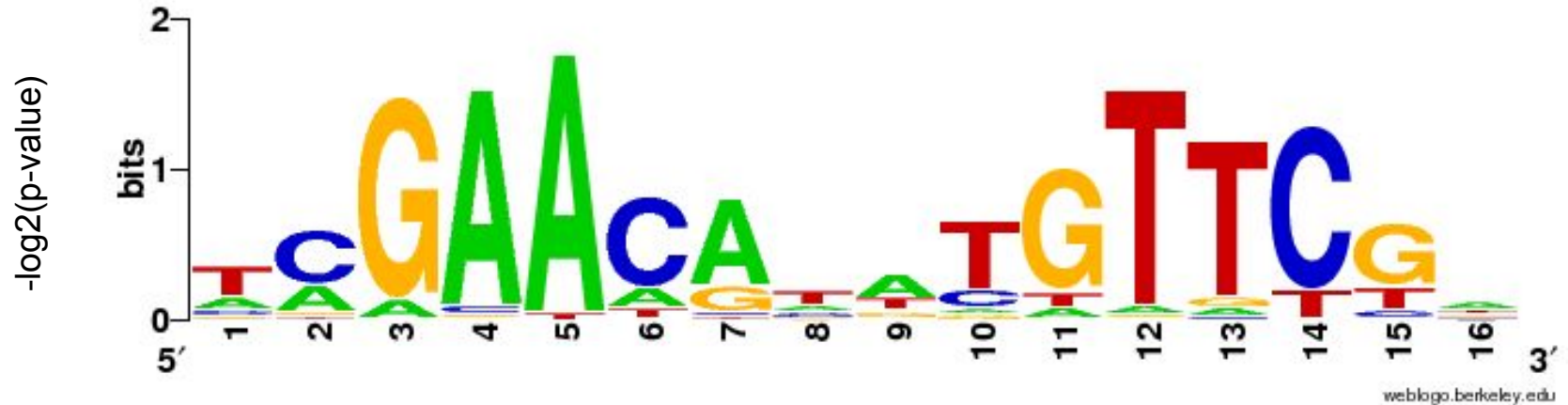
(Soto et al.,
Molecular Cell 2021)



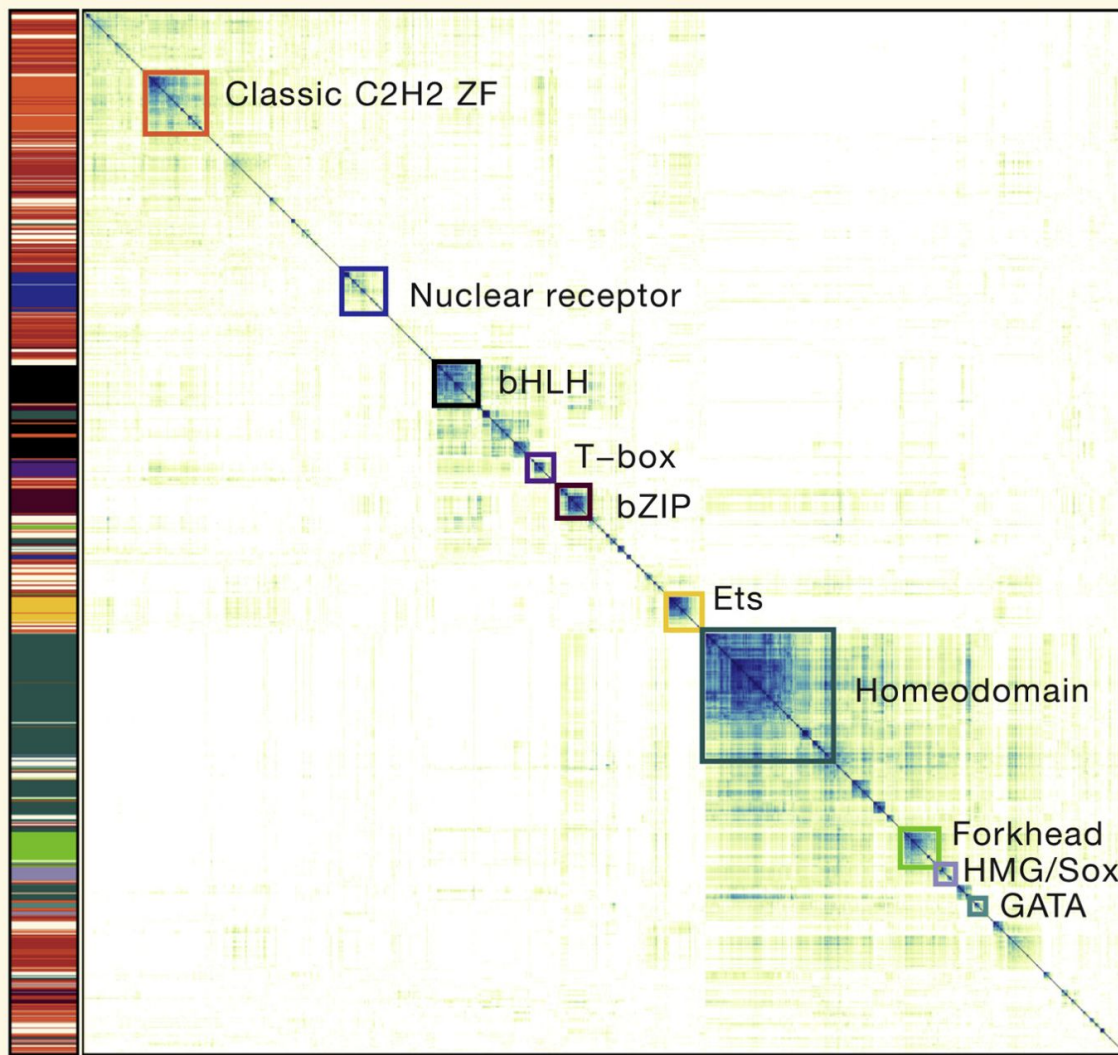
Sequence-specificity

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

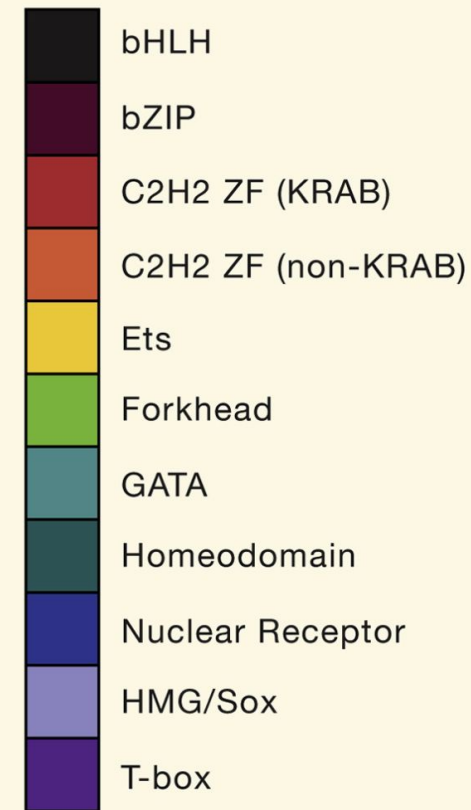
5' -GAACAnnTGTTTC-3'



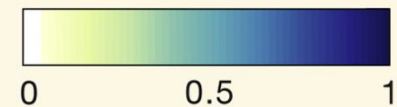
TF Motifs



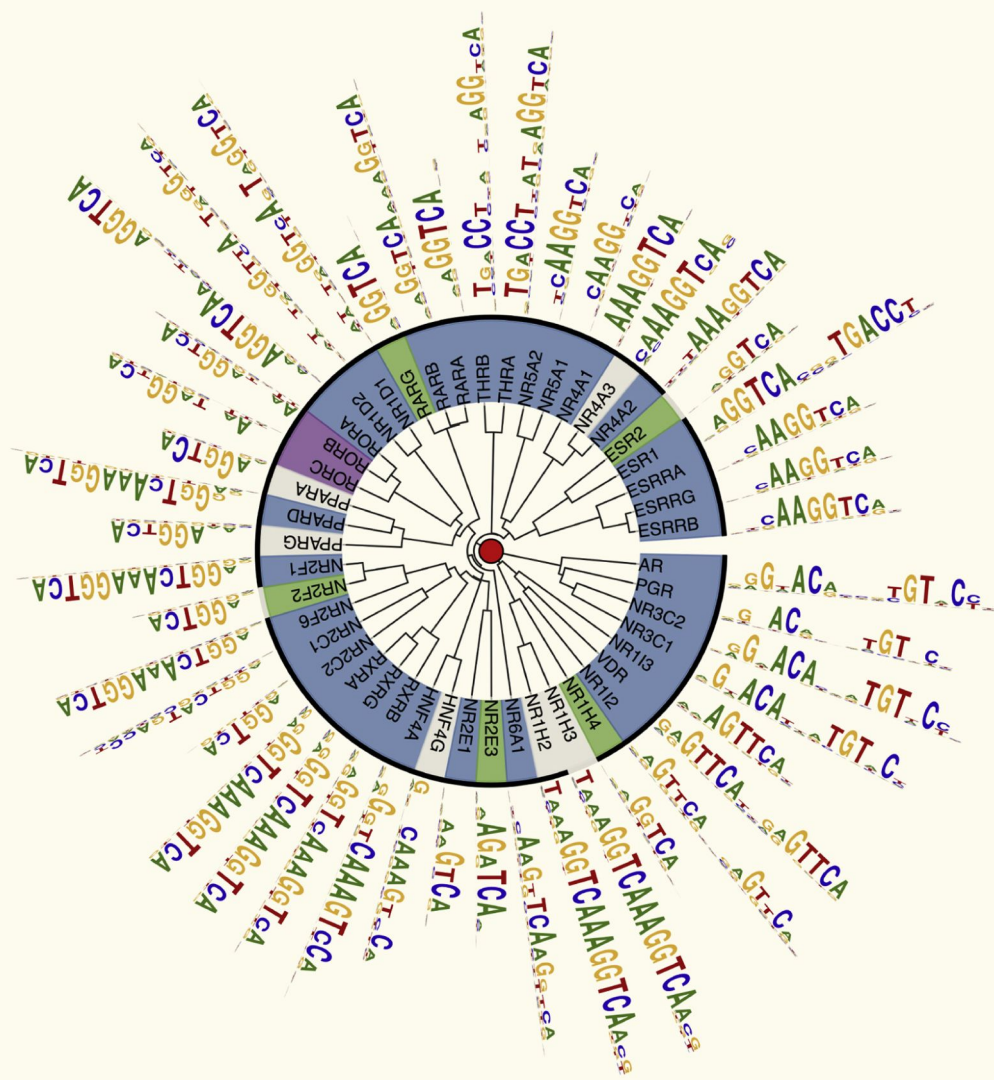
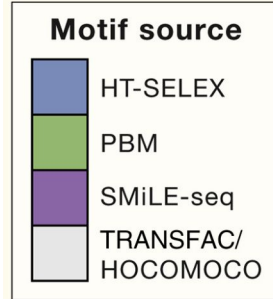
DBD



Motif Similarity (PCC)



An example of TF motif degeneracy: Nuclear hormone receptors

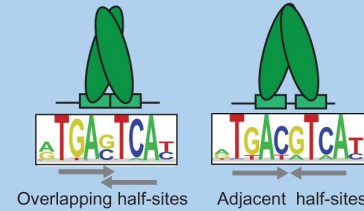


Variations in DNA binding specificity

Multiple Modes of DNA Binding

A

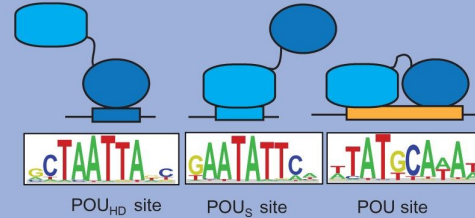
Variable Spacing



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

B

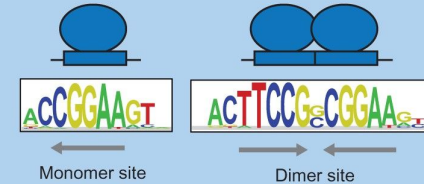
Multiple DBDs



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

C

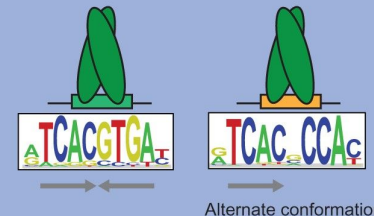
Multi-meric Binding



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)

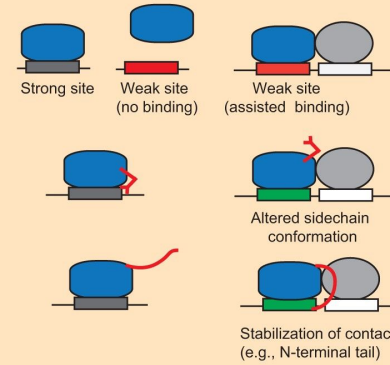
Cooperative binding

Highly combinatorial
binding of TFs

Multi-Protein Recognition Codes

A

Cooperative binding

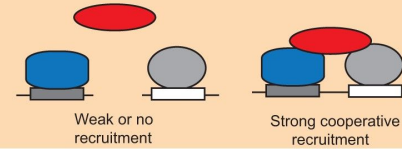


Enhanced complex stability due to cooperativity allows binding to lower-affinity (weak) sites (103,104,106)

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

B

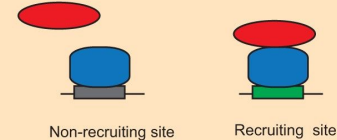
Cooperative recruitment



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

C

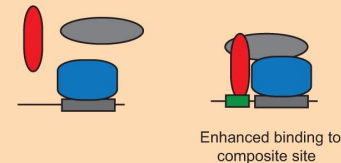
Allostery



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

D

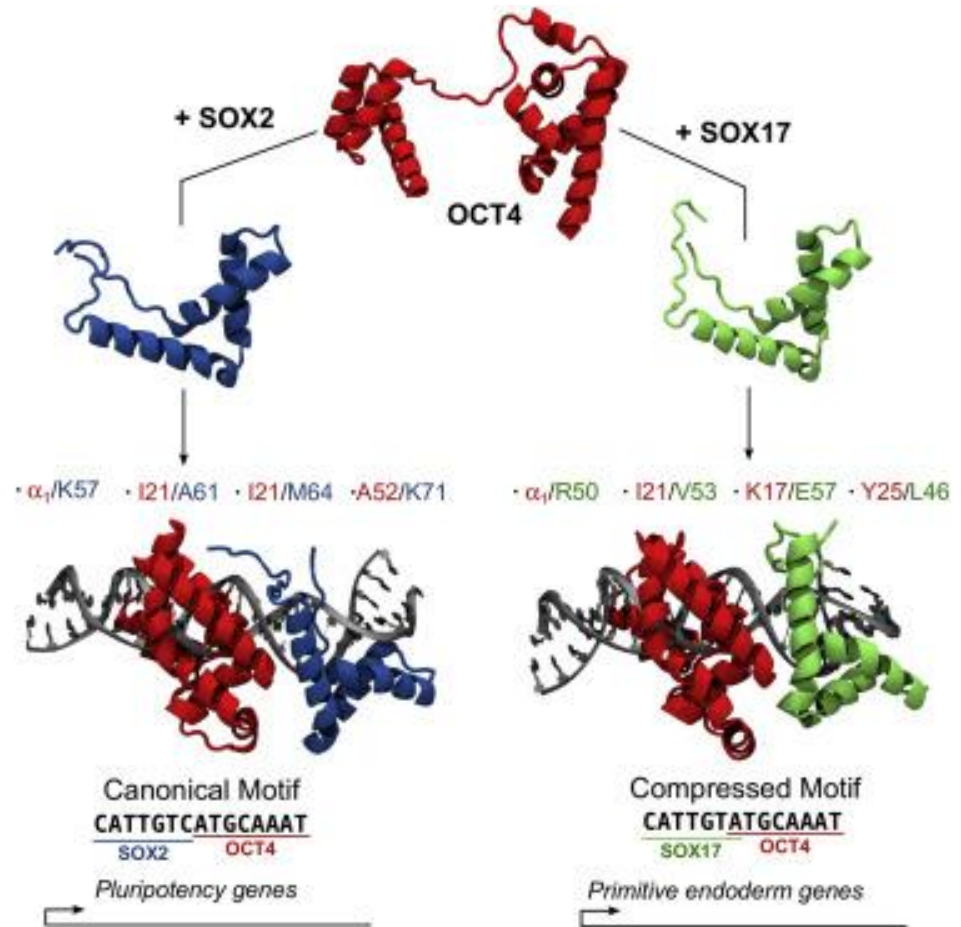
Cofactor-based targeting



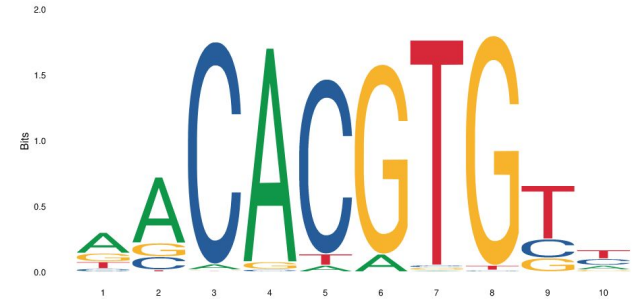
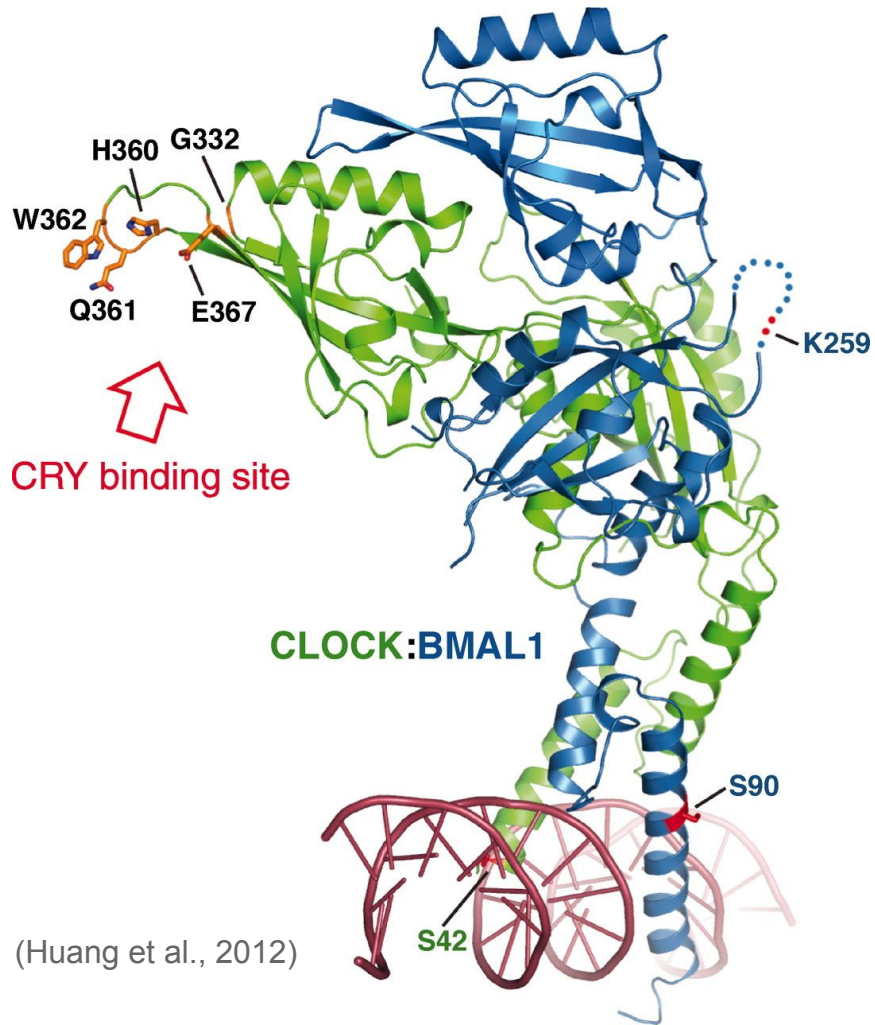
Enhanced binding of multi-protein complex to specialized composite sites is mediated by interactions between non-DNA-binding cofactor and an auxiliary motif (48,129)

Two examples of Cooperative binding

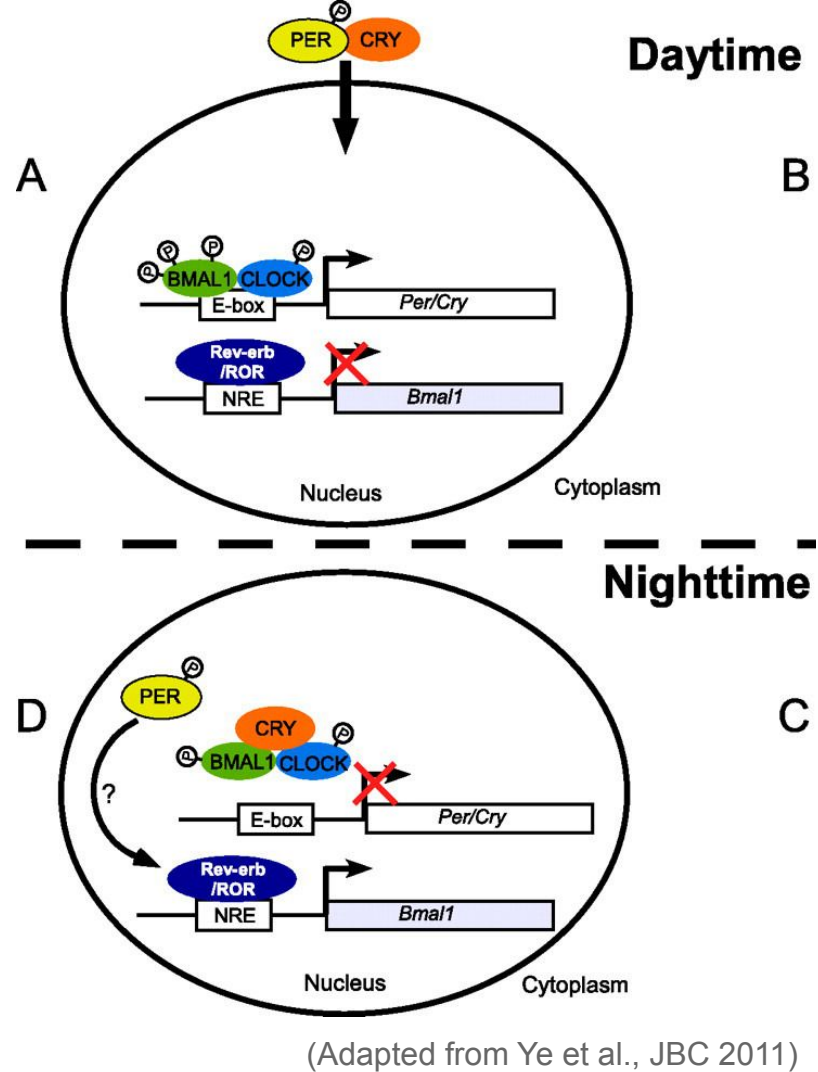
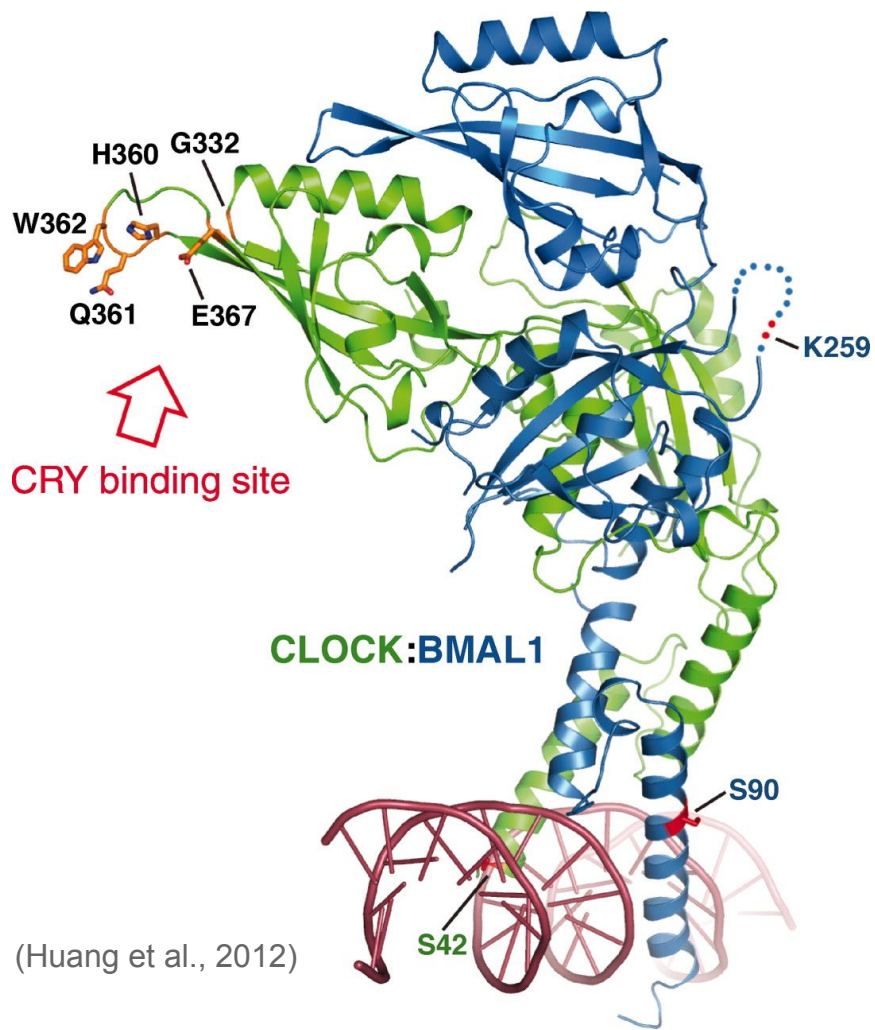
OCT4 (POU5f1) binding upon differentiation



Clock-Bmal-Cry during circadian rhythm



(Huang et al., 2012)



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 R 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)
 - Bioconductor method: motifmatchr
 - (other options are the TFBSTools R package and FIMO of the Meme suite)

Genetic variation at TF binding sites

- Genetic variation at TF binding sites can affect the binding of the protein, and hence impact development and health
- Nevertheless, while most coding sequences show evidence of **evolutionary constraint** (e.g. purifying selection), only a small fraction of TF binding sites (11.6% of footprints) show evidence of constraint – the vast majority appears to be evolving neutrally

(Vierstra et al., Nature 2020)

- This suggests a degree of (at least partial) redundancy between regulatory elements

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 peaks for that factor (whatever organism/cell type, just make sure you use the corresponding genome!)
- 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 (or in one chromosome), 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](#) !