"Next thing is how a cell's picking which GATs (stretches of nucleotides) get chosen, like Yogi in a picnic basket. Proteins and DNA? Some interesting chemistry. Cuz they getting jiggy with some different affinities."

–Tom McFadden

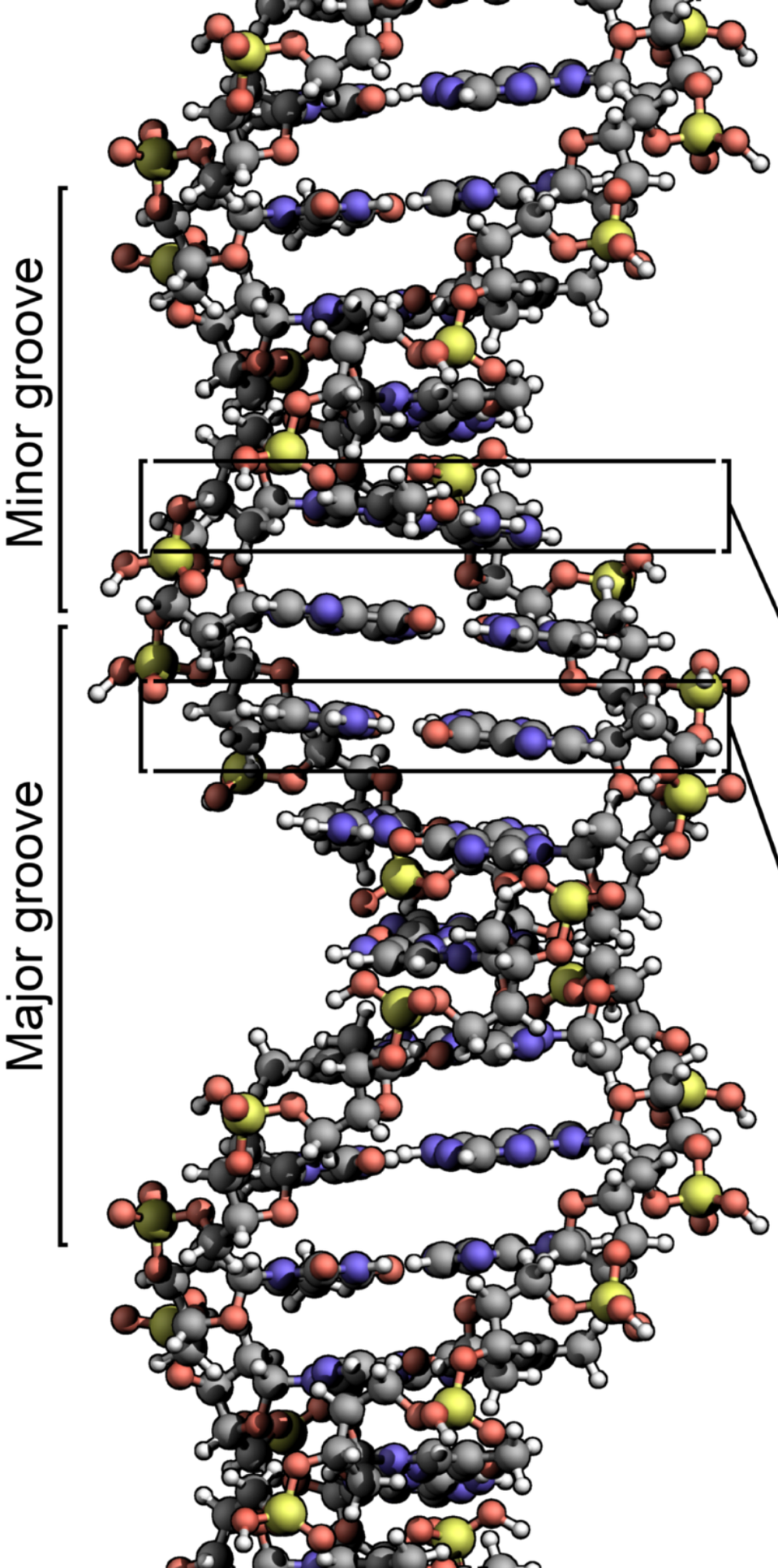
https://www.youtube.com/watch?v=9k_oKK4Teco&list=RD9k_oKK4Teco

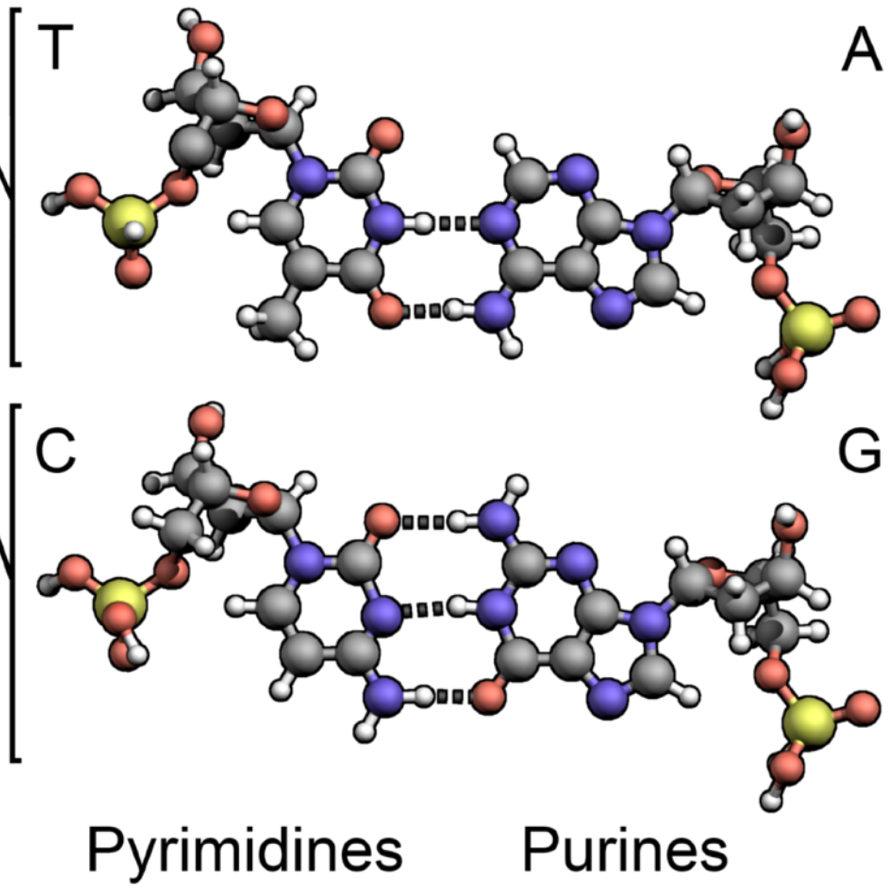
How do proteins interact with specific DNA sequences?

Hydrogen
Oxygen
Nitrogen
Carbon
Phosphorus

Minor groove

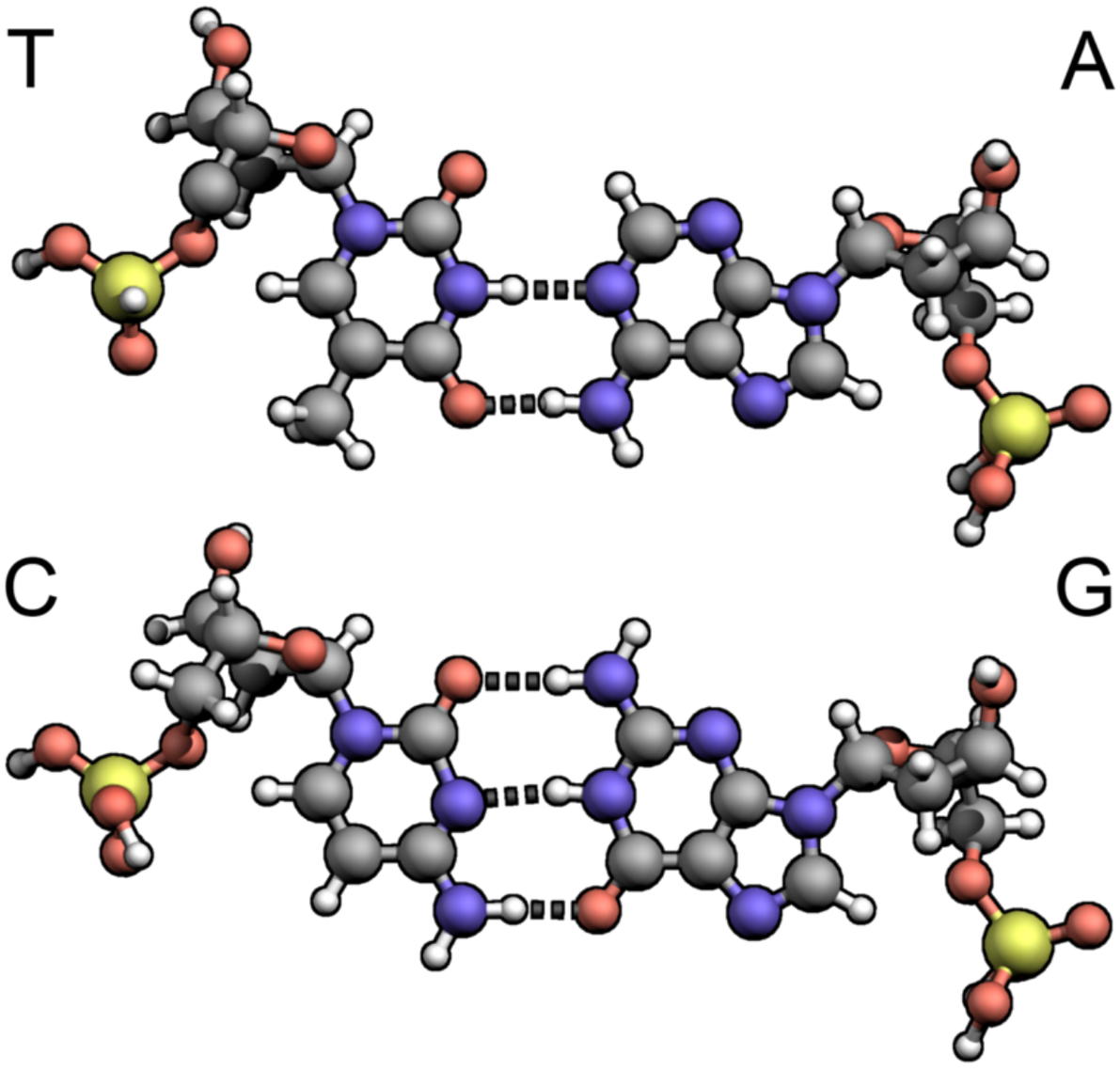
Major groove

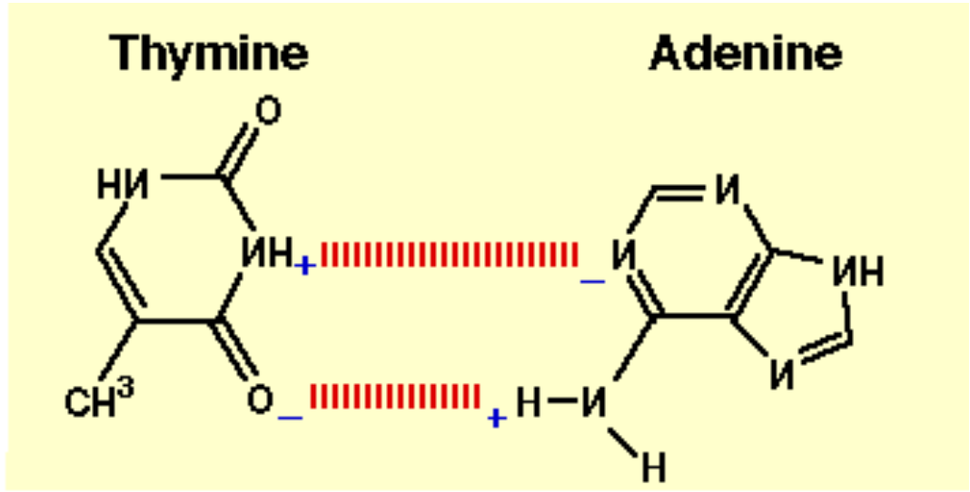


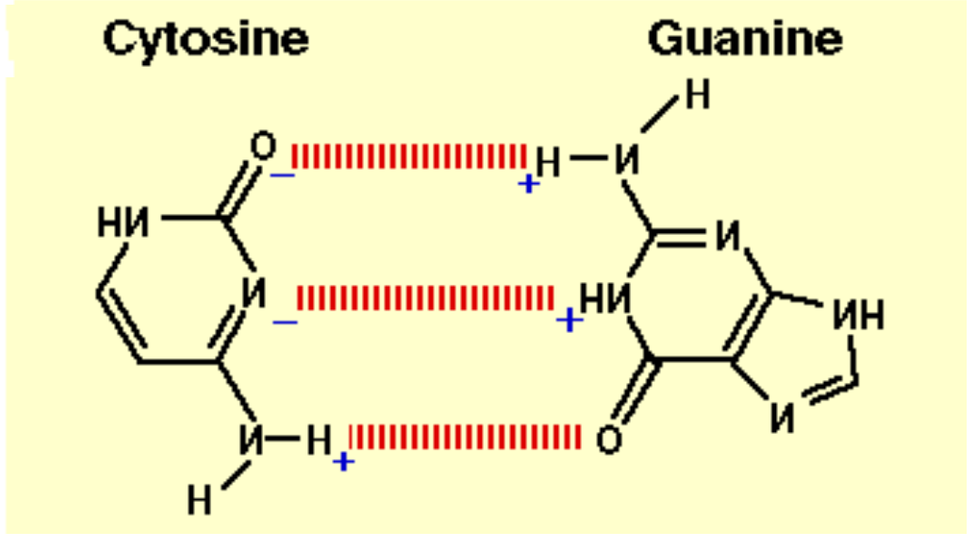


Pyrimidines Purines

Hydrogen bond is the electrostatic attraction between polar groups

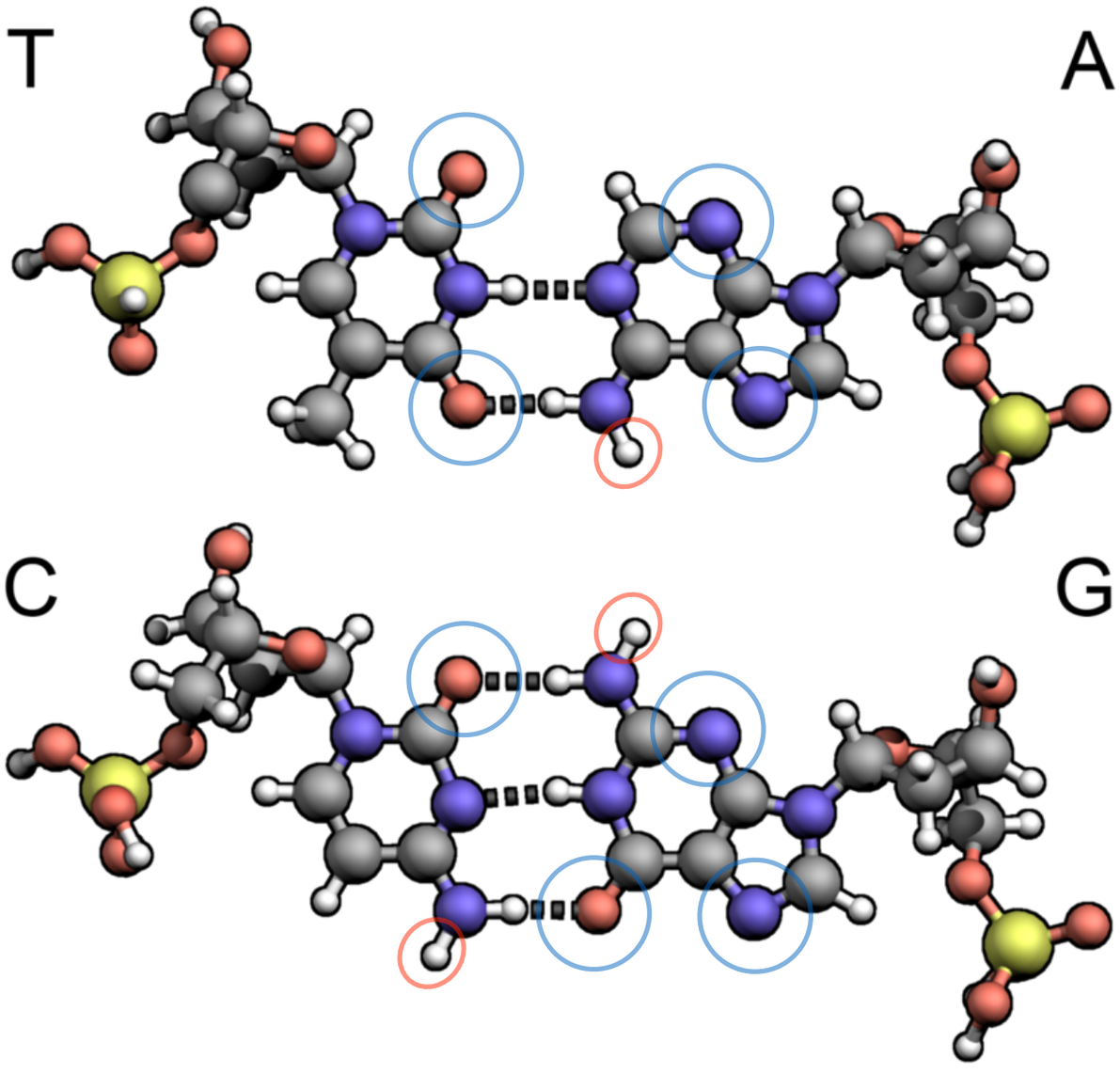


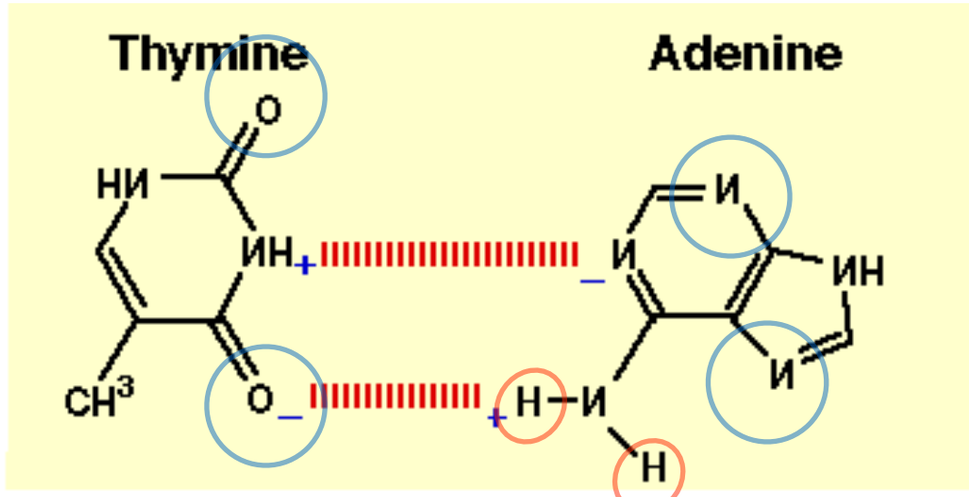


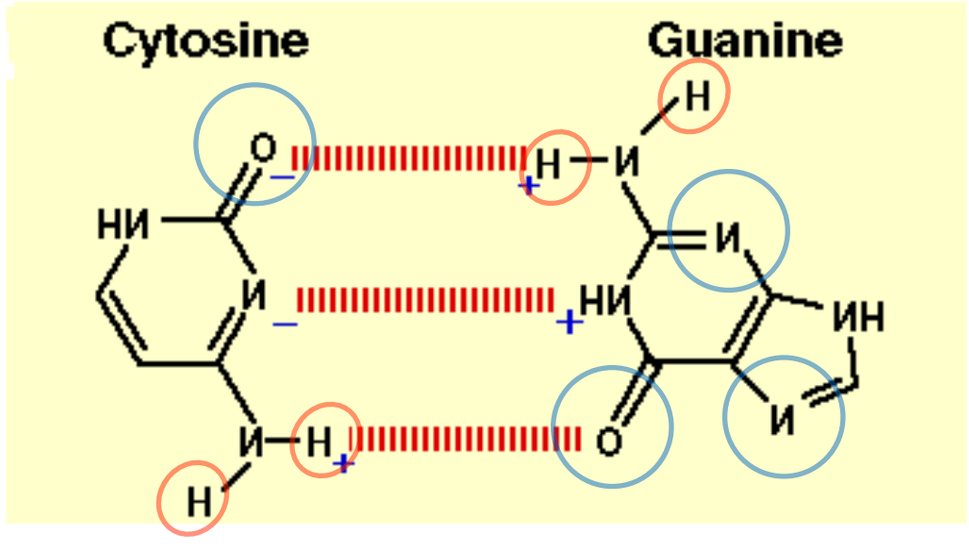


H-bond: a Hydrogen atom bound to a highly electronegative atom such as Nitrogen or Oxygen experiences attraction to another nearby highly electronegative atom.

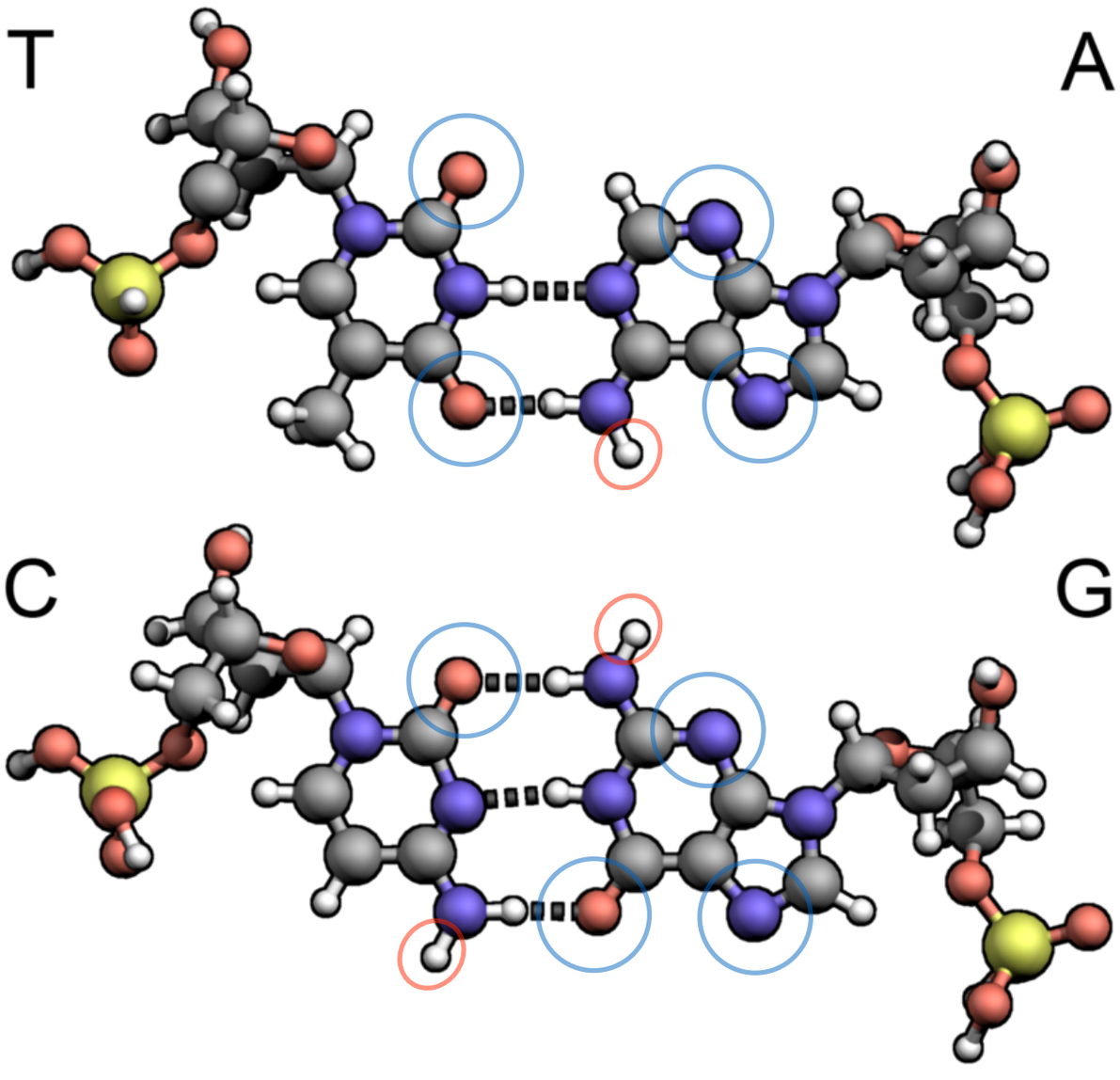
The atoms shown below are available to mediate protein/DNA interactions via H-bonds





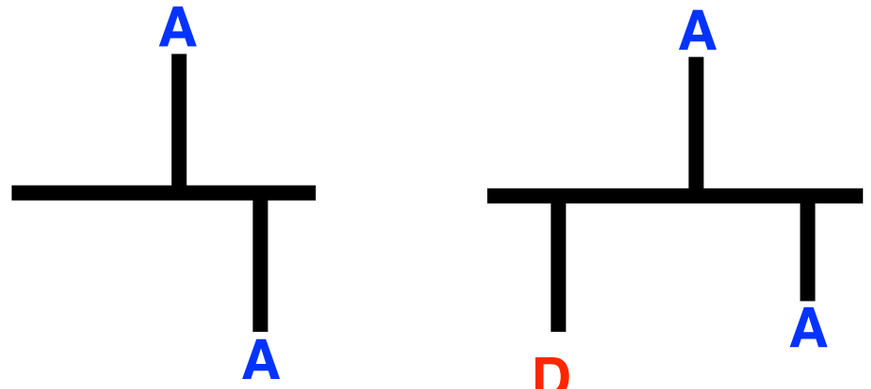


The atoms shown below are available to mediate protein/DNA interactions via H-bonds



T-A

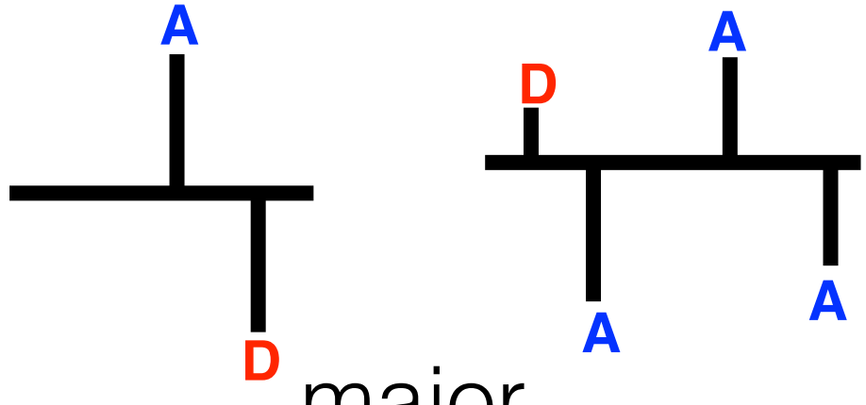
minor



major

C-G

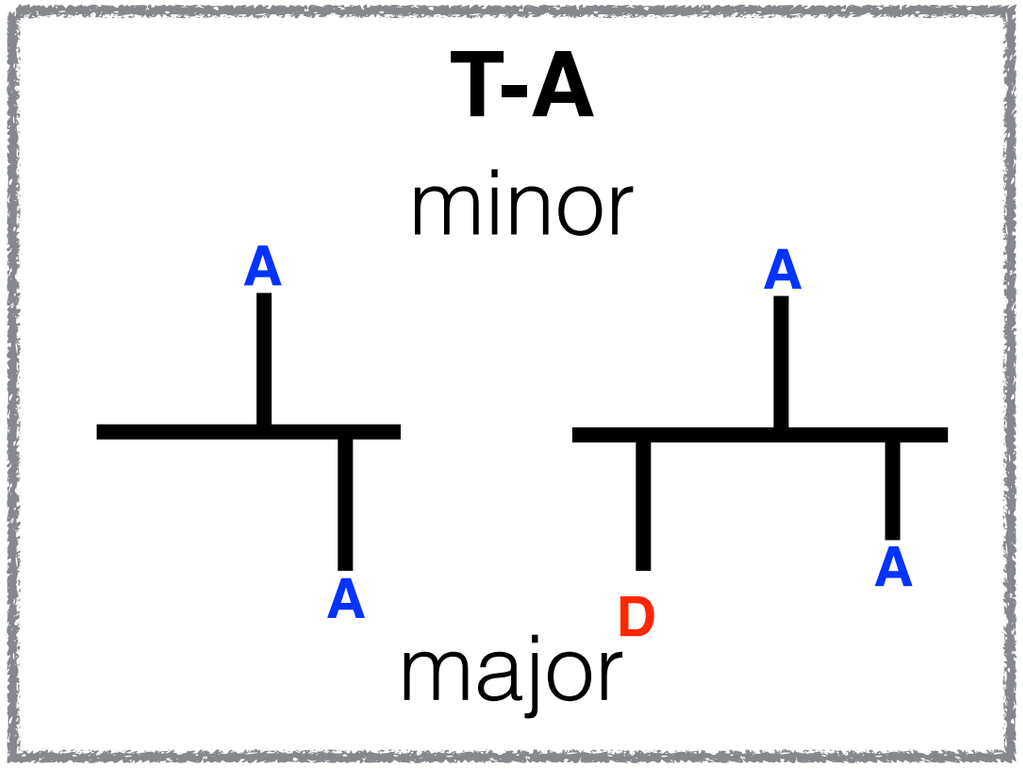
minor

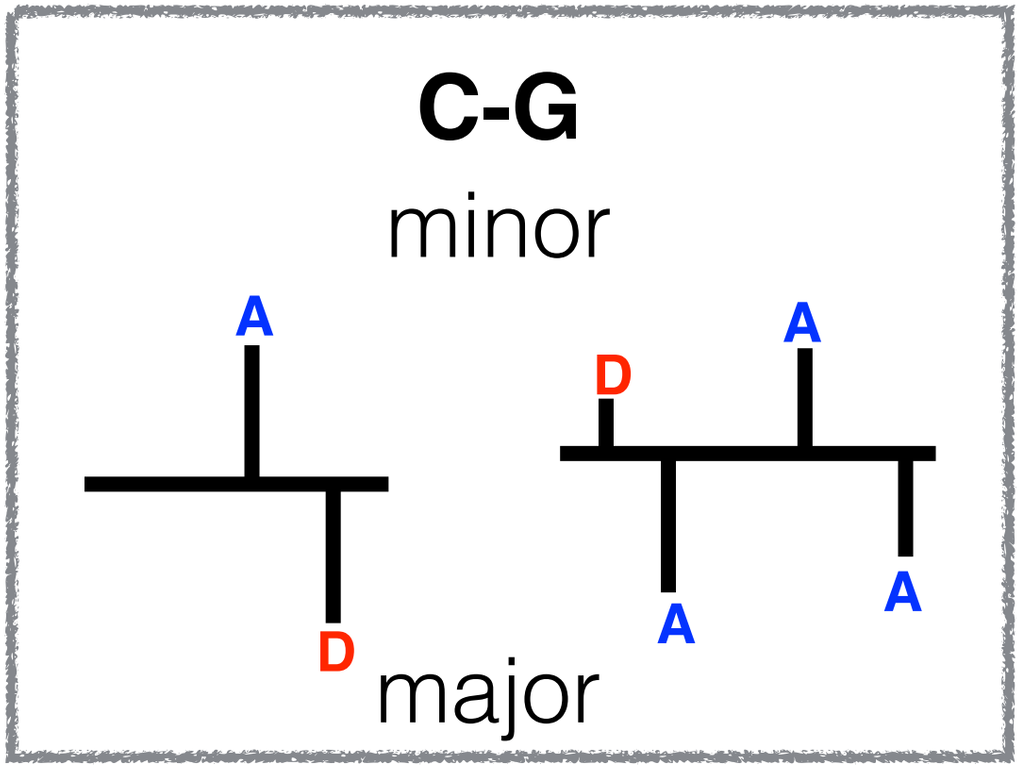


major

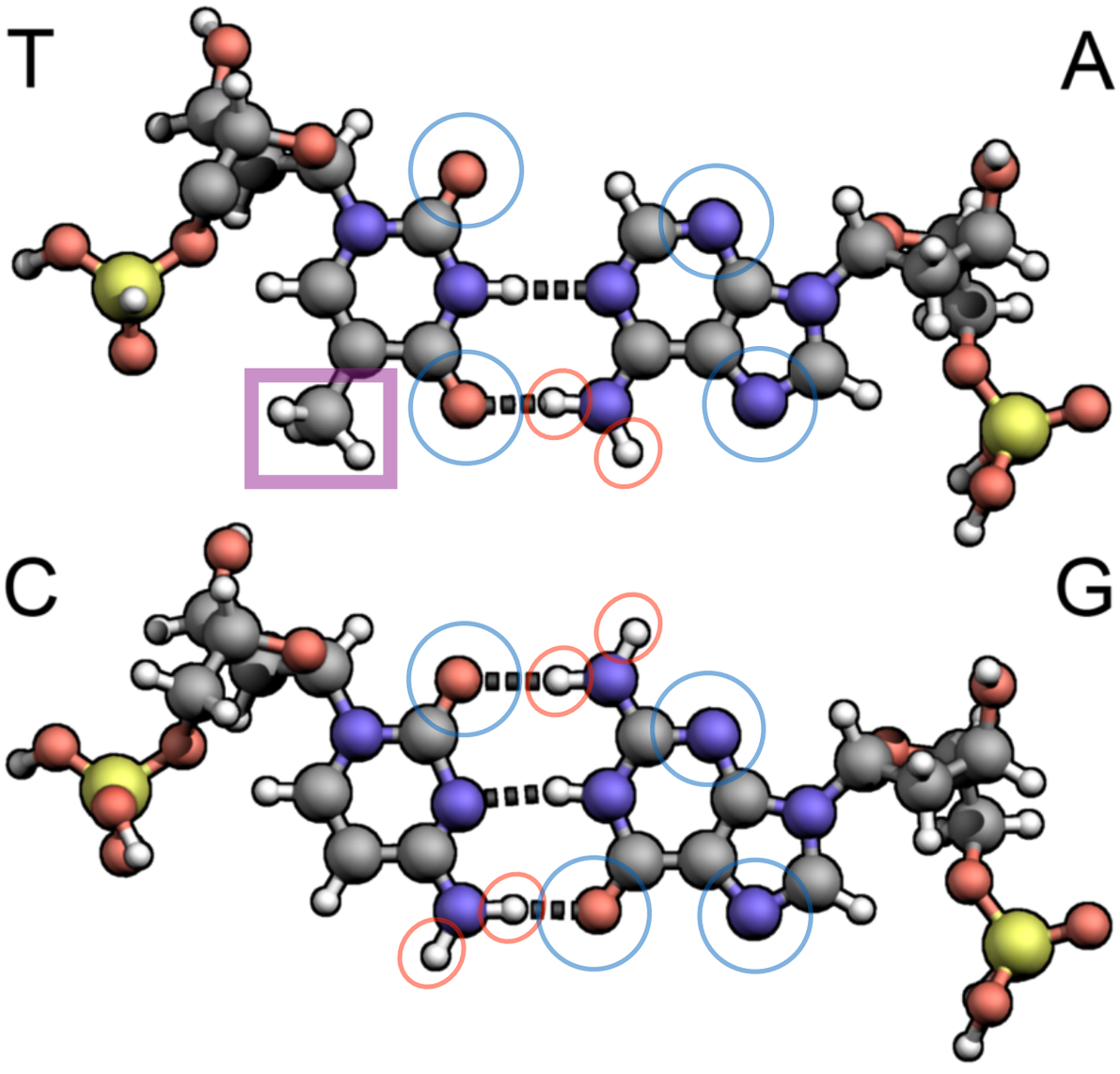
Hydrogen Bond Donors and Acceptors are exposed in the Major and Minor Grooves

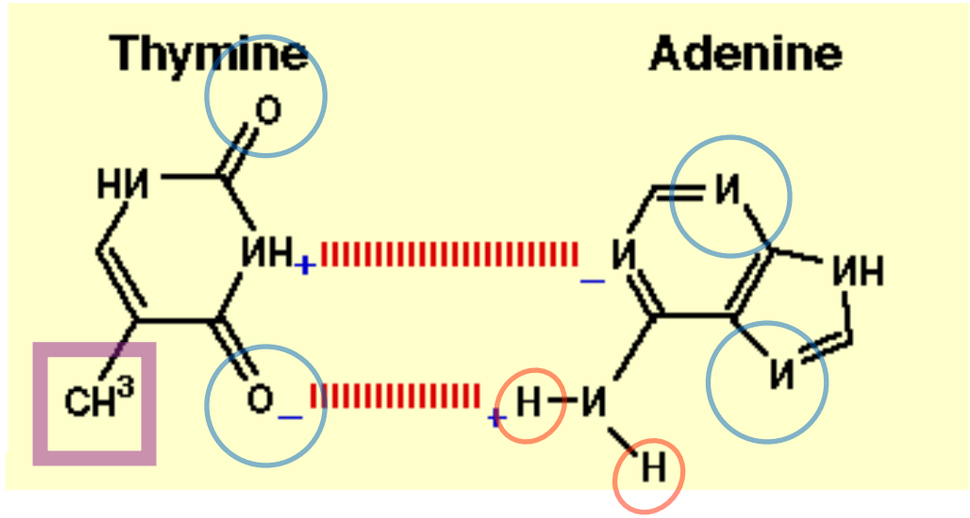
Base sequence is read an recognized by a protein probing the H-bonding possibilities in the Major and Minor Grooves

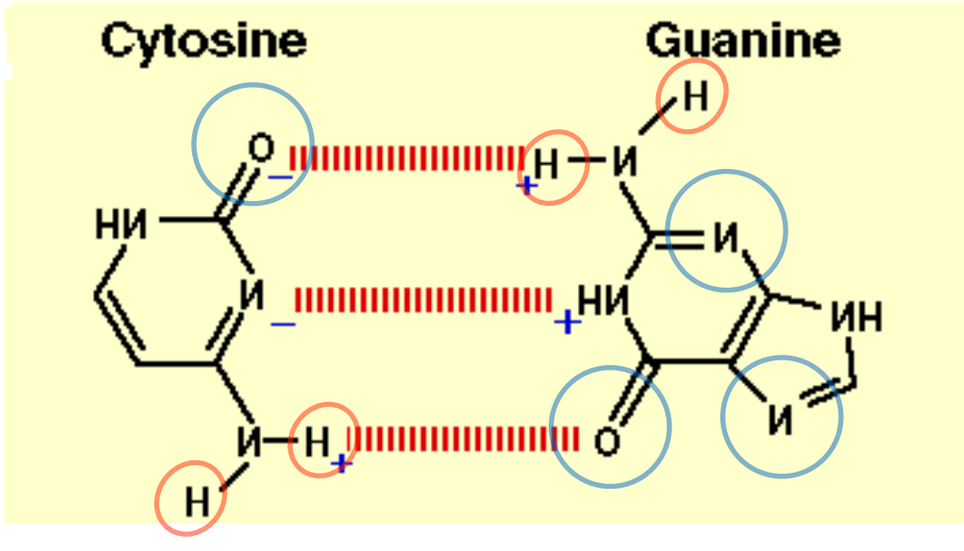




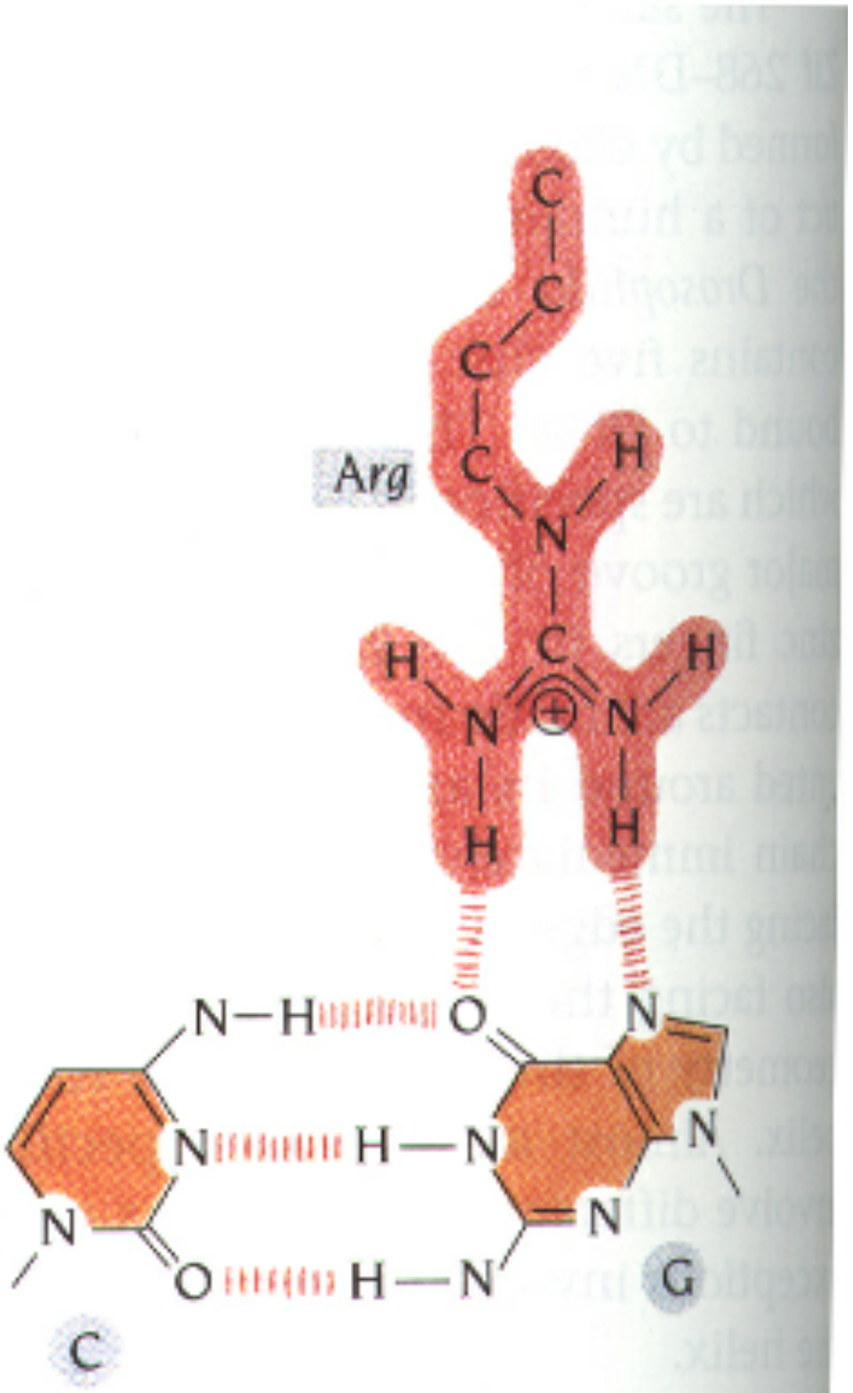
Thymine's methyl group provides an additional source of recognition/specificity







An example of an Amino acid/DNA base interaction



The interaction between arginine with its two hydrogen bond donors and a guanine base with its two acceptors in the major groove is an important component of many protein/DNA interactions.

How do proteins interact with specific DNA sequences?

Hydrogen
Oxygen
Nitrogen
Carbon
Phosphorus

Minor groove

Major groove

T

A

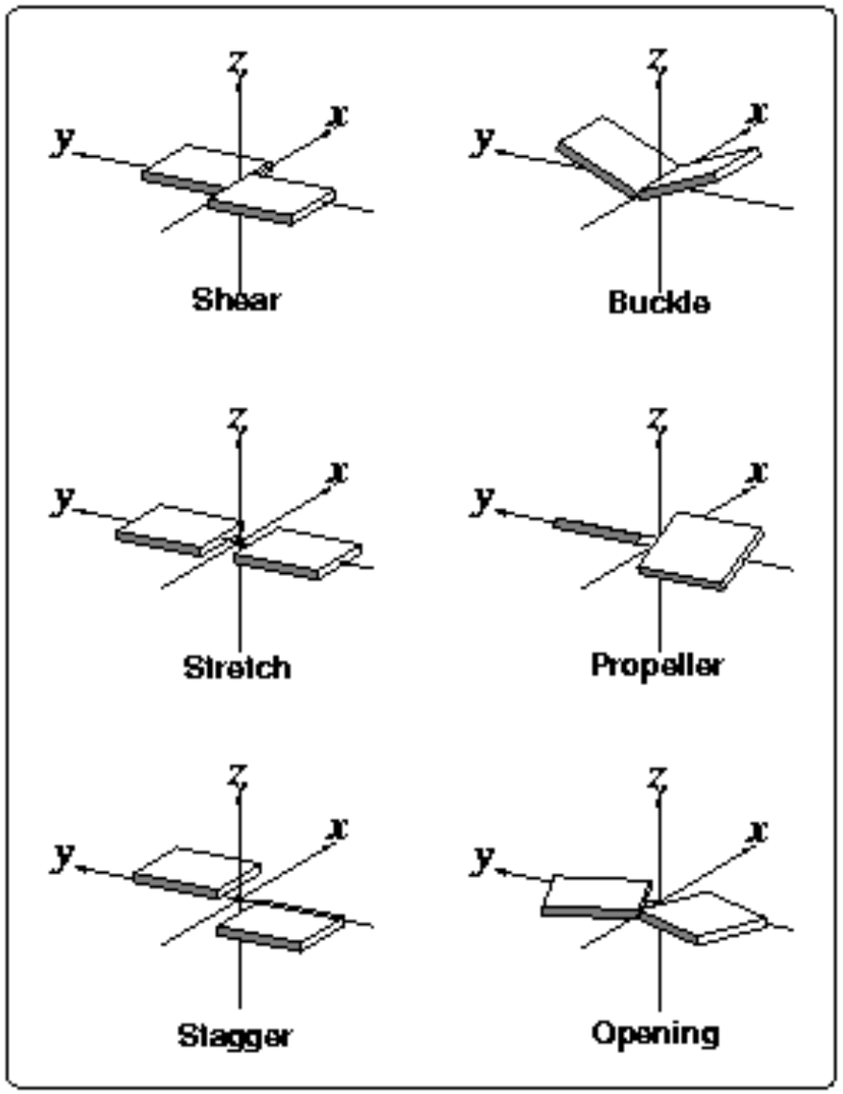
C

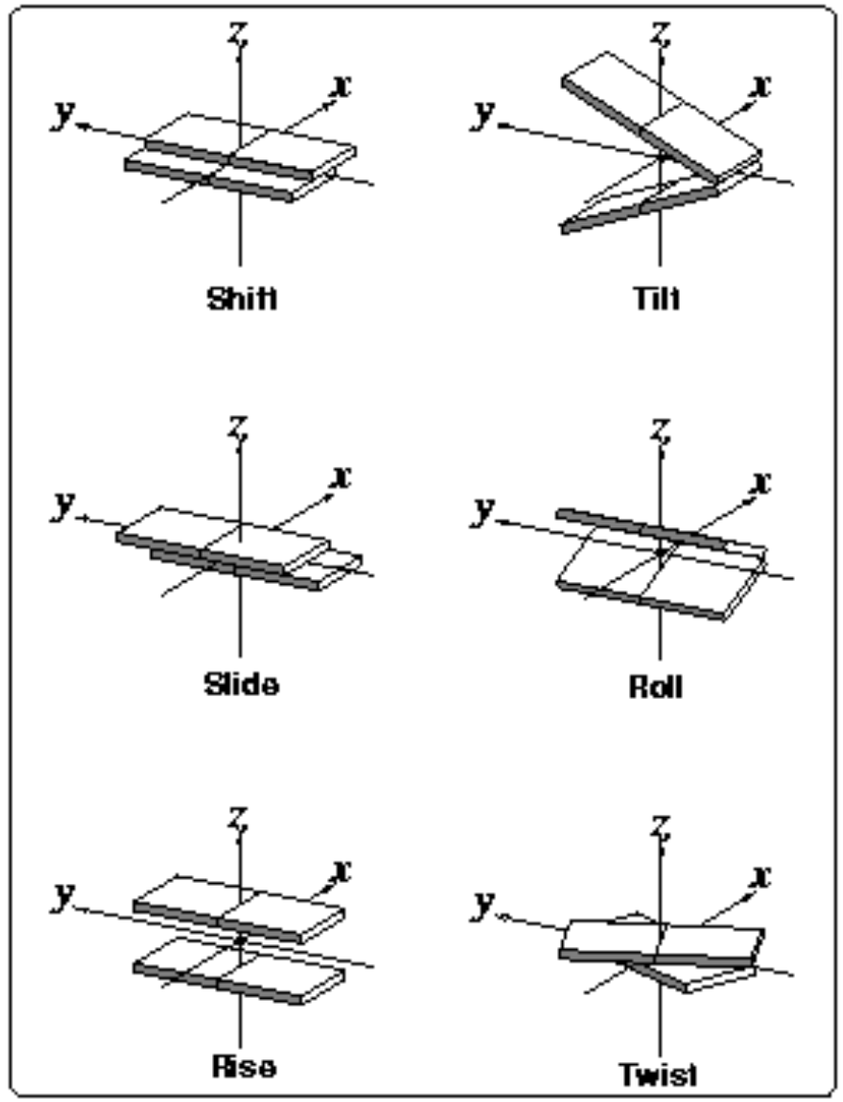
G

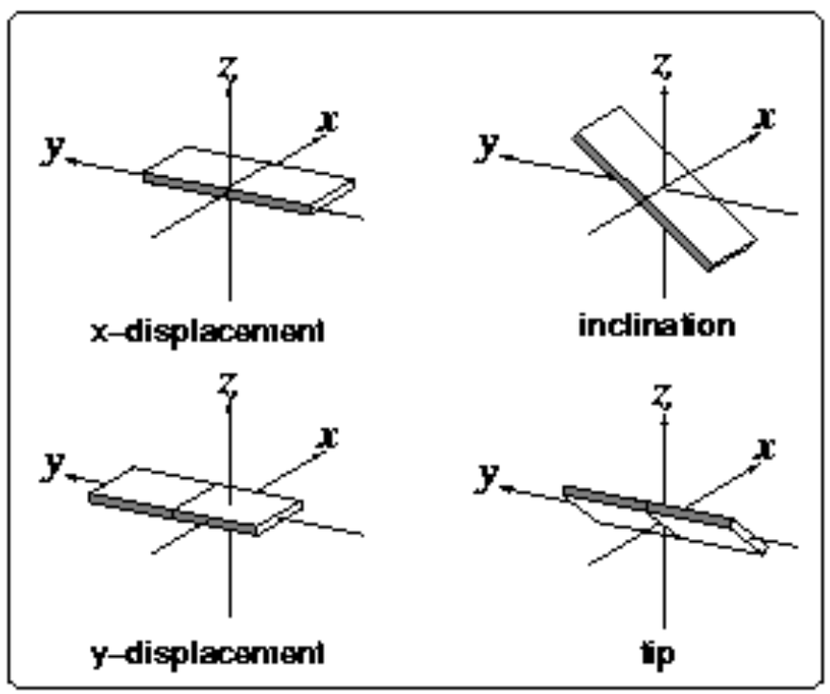
Pyrimidines

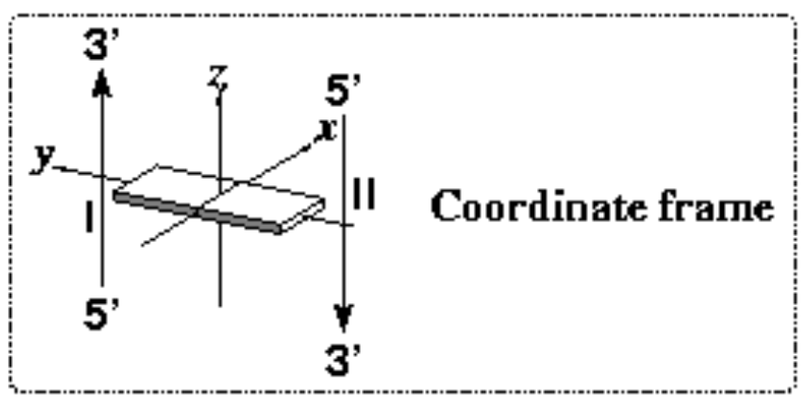
Purines

Base Pair Geometry



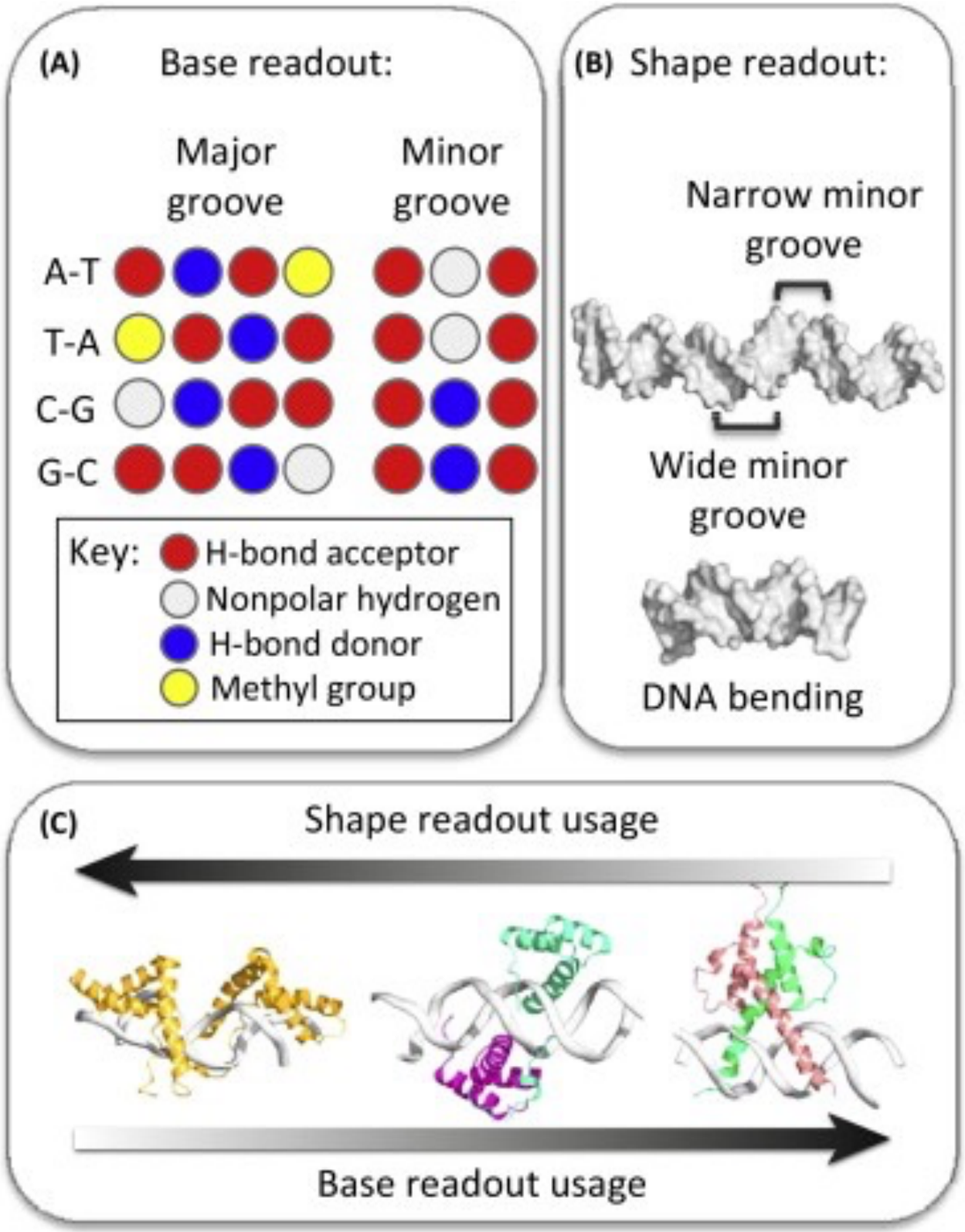






Backbone phosphates are differentially positioned based on the degree of tilt, buckle, twist, roll, etc. relative to the preceding base pair.

Base Composition and Shape Contribute to TF-DNA Specificity



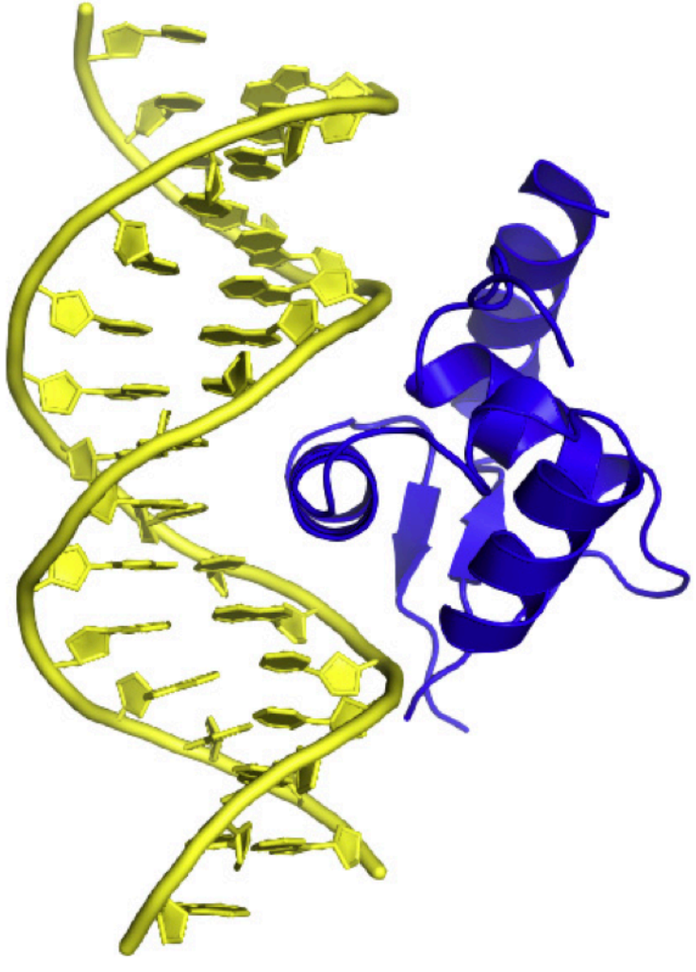
- Base readouts are specific for bp in major groove but degenerate for minor groove.
- Shape dominates for a minor groove-binding high motility group (HMG) box protein
- Base readout is a major contribution in DNA recognition by the bHLH protein Pho4
- Both readout modes are ~equally present in the DNA binding of a Hox–Exd heterodimer

Slattery et al. (2014) Trends Biochem Sci. 39: 381-99

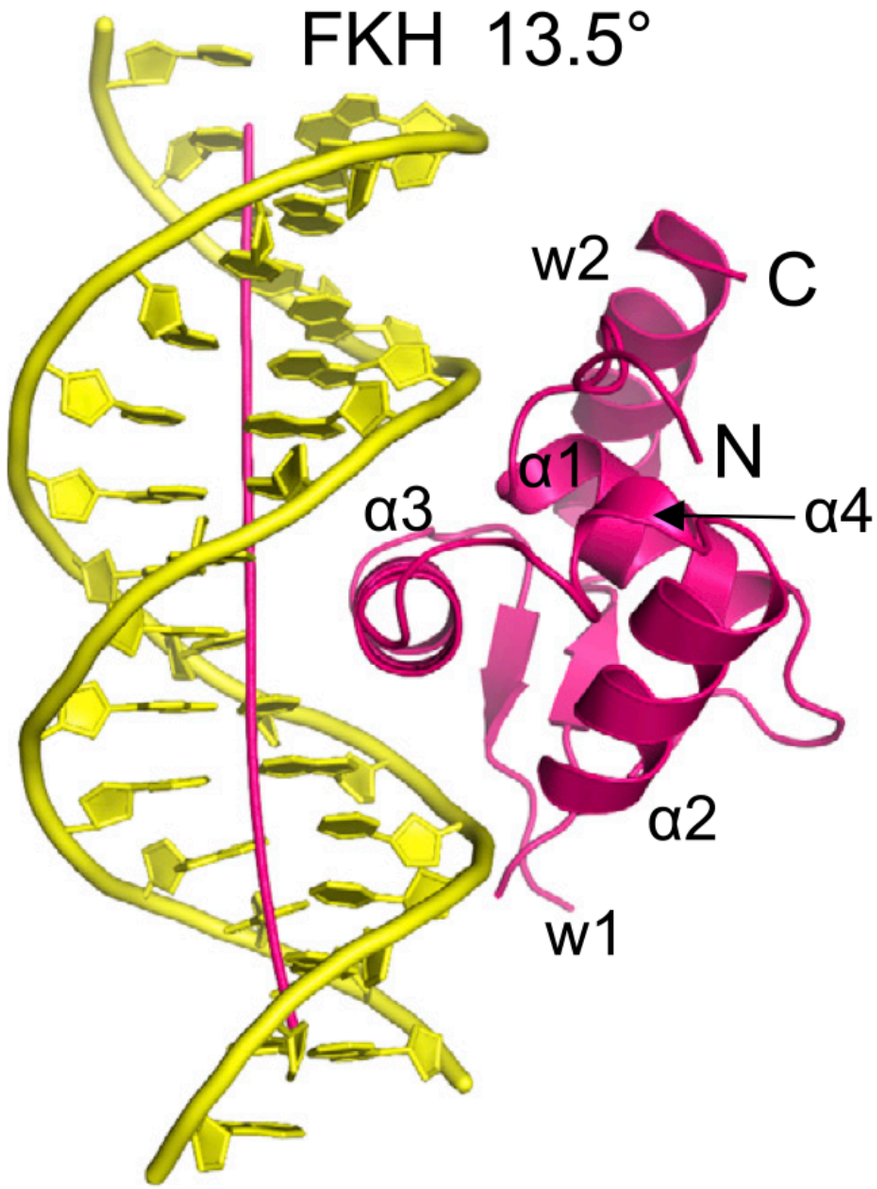
FoxN3 can bind to two distinct sequences that have distinct shapes

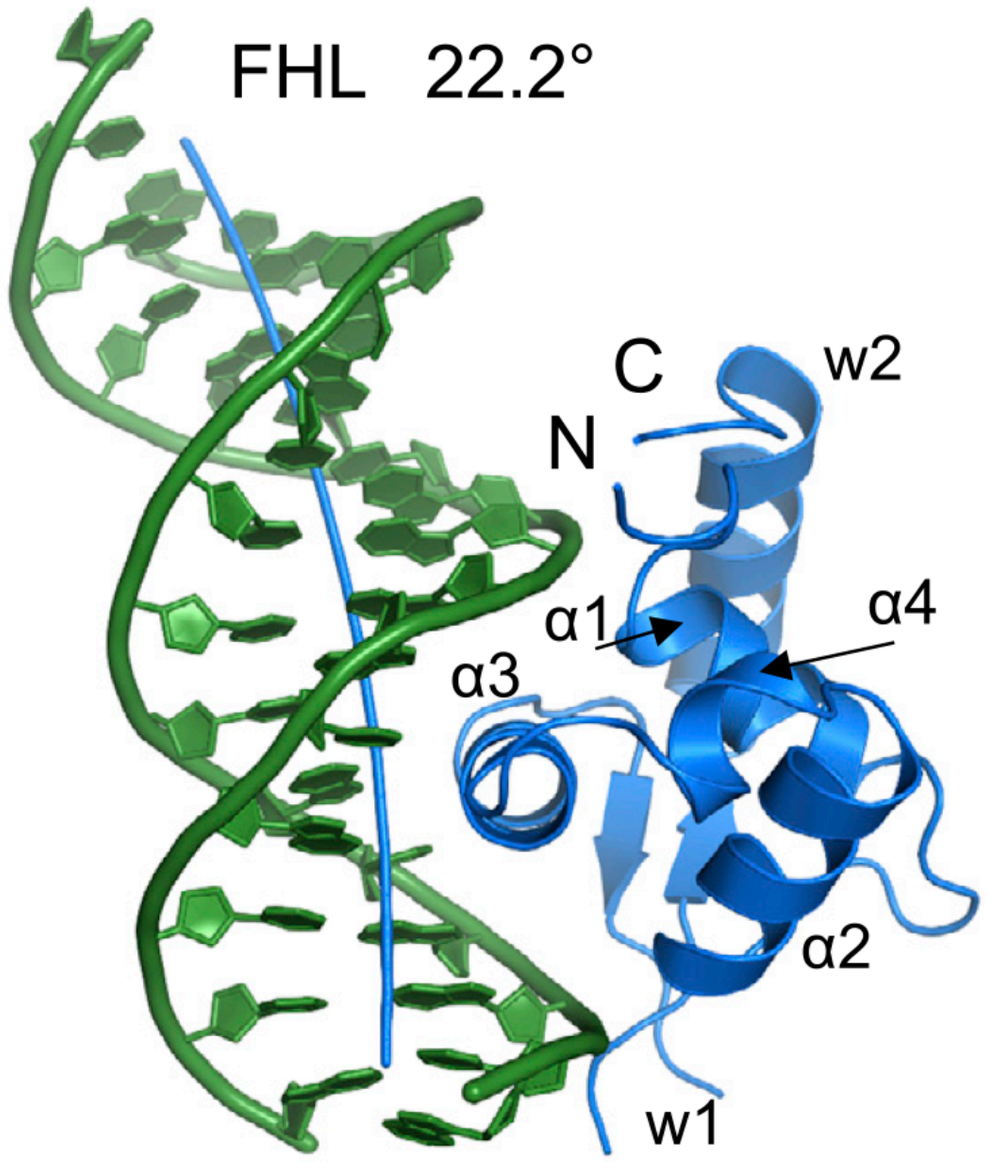
FKH

FHL



FoxN3 can bind to two distinct sequences that have distinct shapes





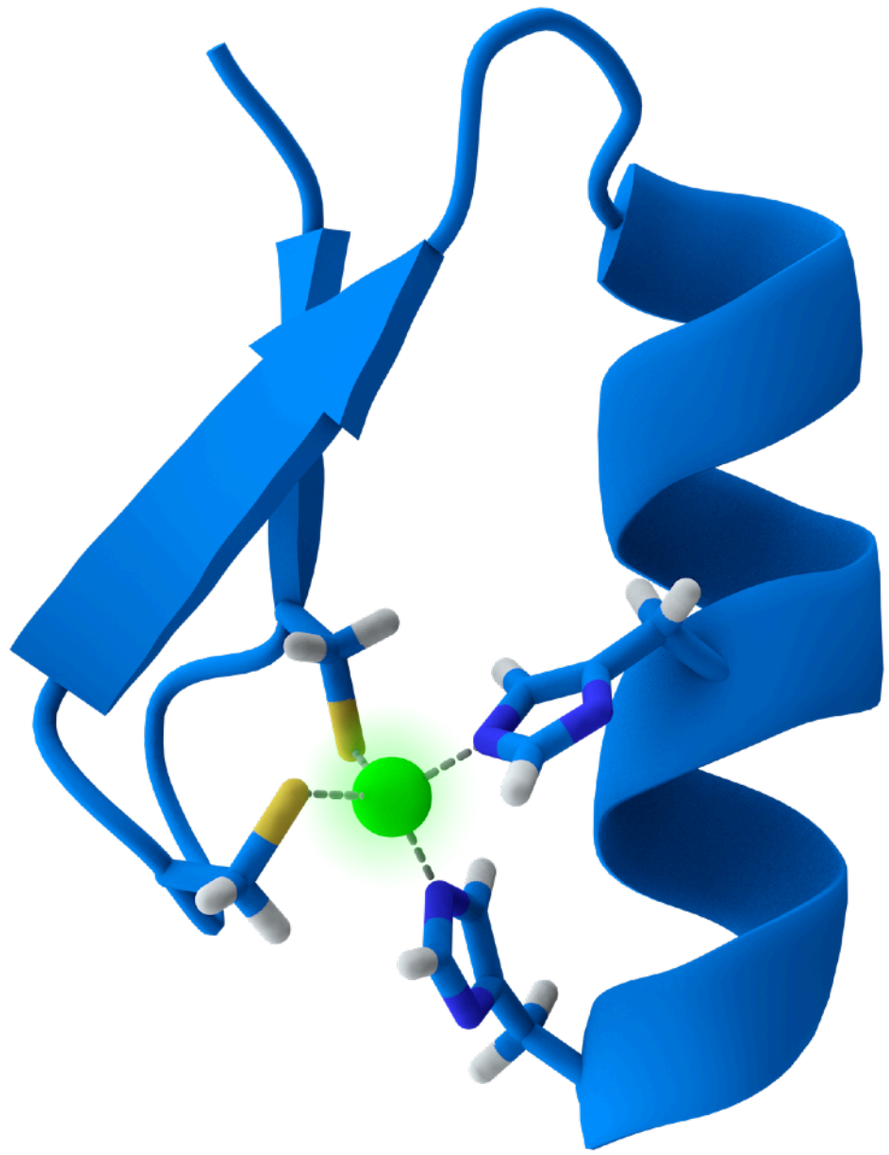
Rogers et al. (2019) Molecular Cell

What are the features of protein domains that bind DNA?

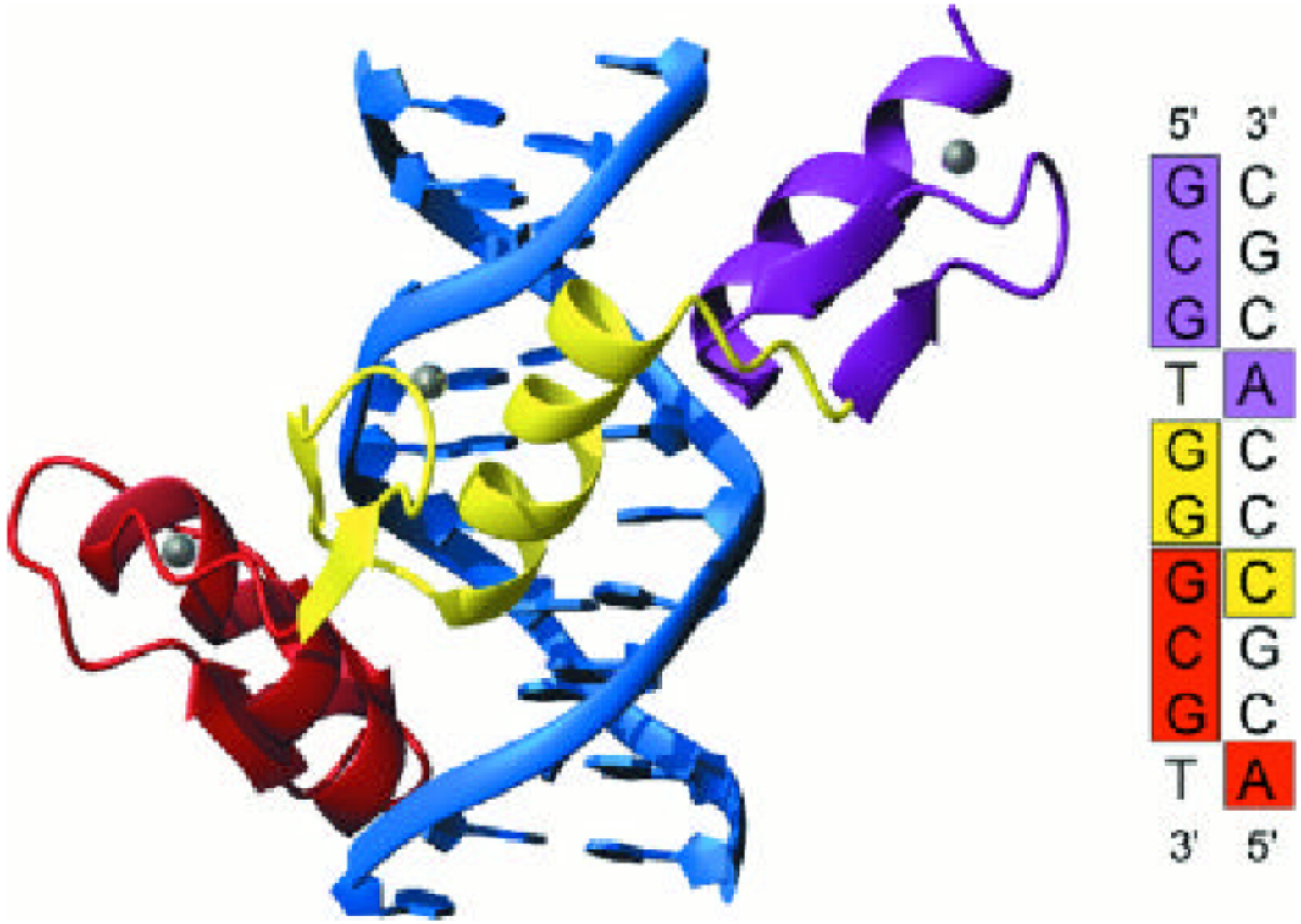
Zinc-containing DNA binding domains

(Zinc is coordinated with a combination of Cysteine and Histidine residues)

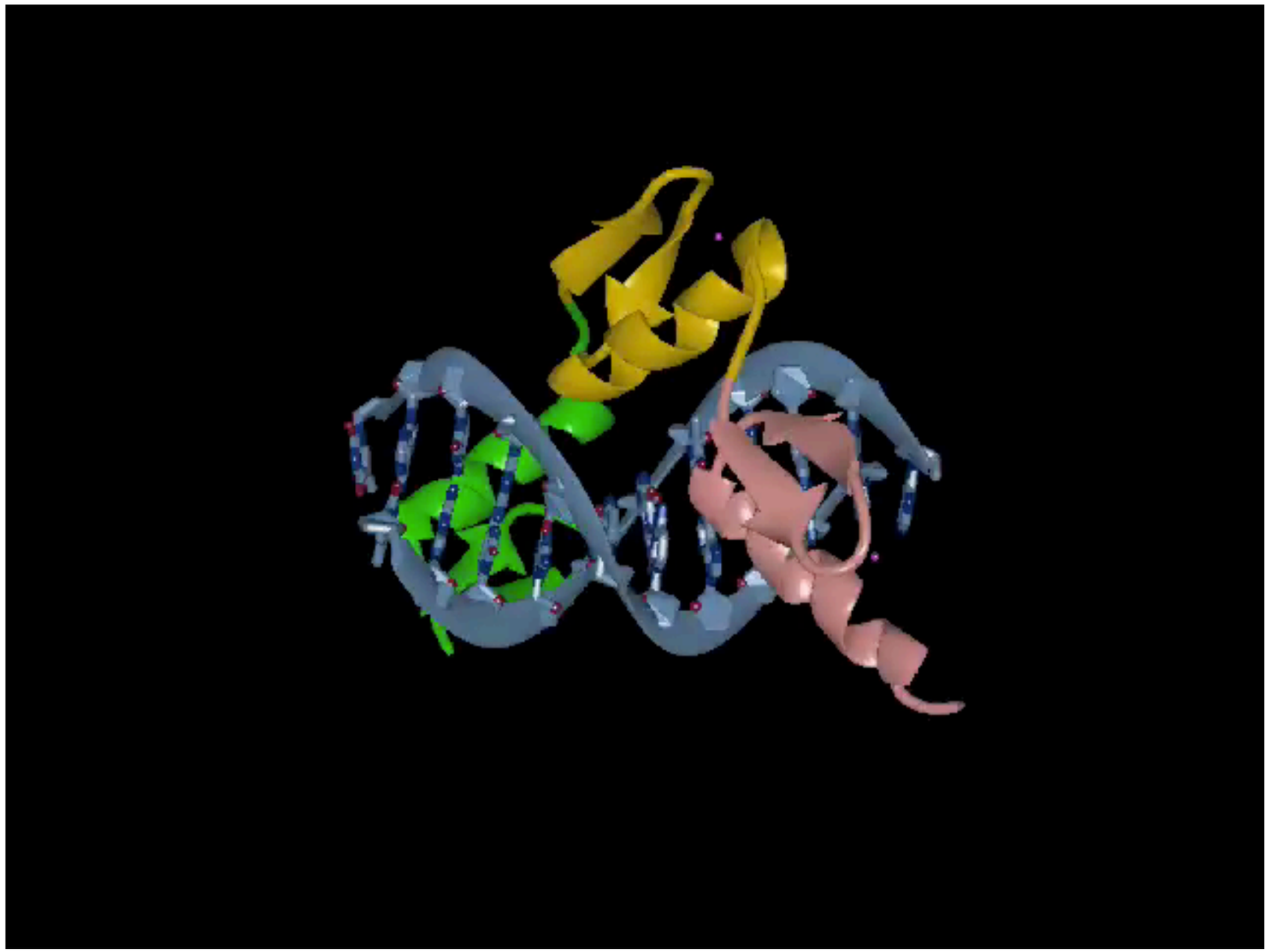
- Cys_2His_2 Class:
  - Beta-Beta-Alpha fold
    - $Cys\text{-}X_{2\text{-}4}\text{-}Cys\text{-}X_{12}\text{-}His\text{-}X_{3\text{-}5}\text{-}His$



Three zinc fingers of Zif268 follow the major groove with each fingers occupying ~3 bp.



Song Tan lab: Zif268



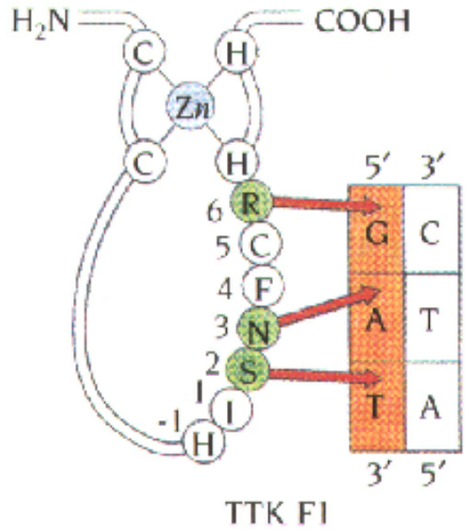
http://www.personal.psu.edu/sxt30/movies/zif268dna_h264.mov

Comparing Zn Fingers

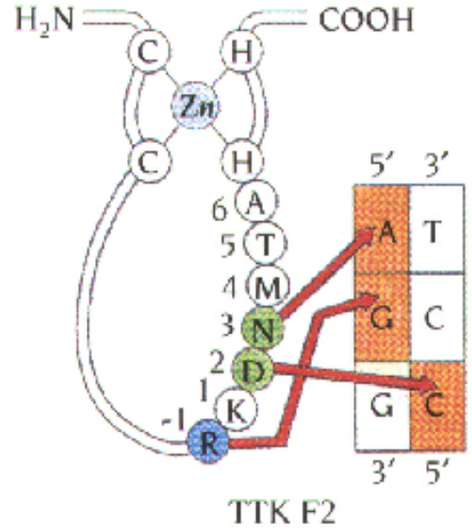
- Variations in a simple motif can provide a dramatic range of DNA sequence recognition.

- Zn Finger nucleases provide for directed mutagenesis. Geurts et al. Knockout rats via embryo microinjection of zinc-finger nucleases. Science. 2009 325: 433.

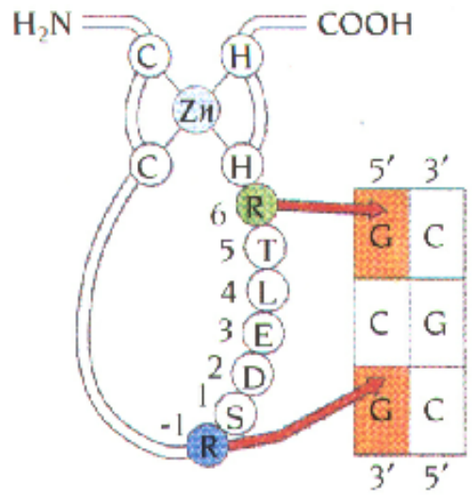
- Have you ever heard of TALE & TALENs? They were the rage prior to CRISPR.



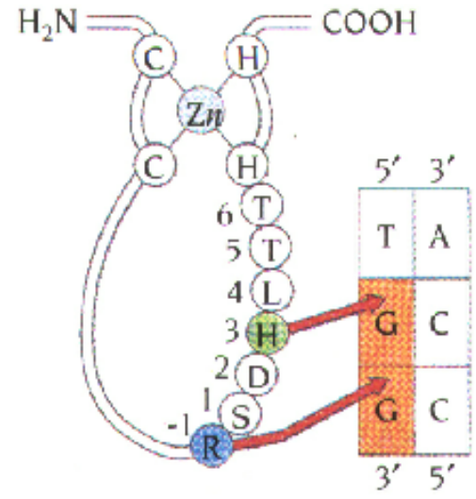
TTK F1



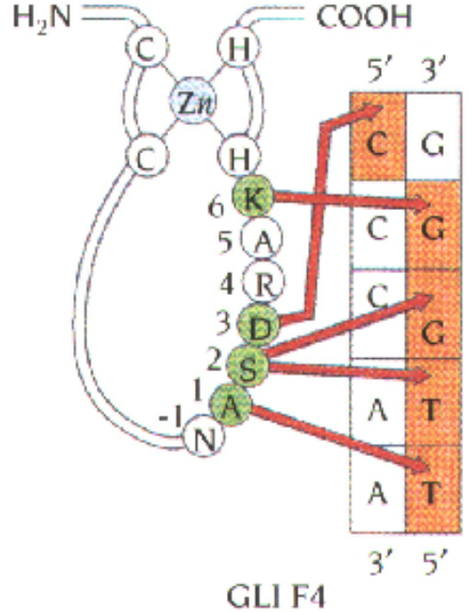
TTK F2



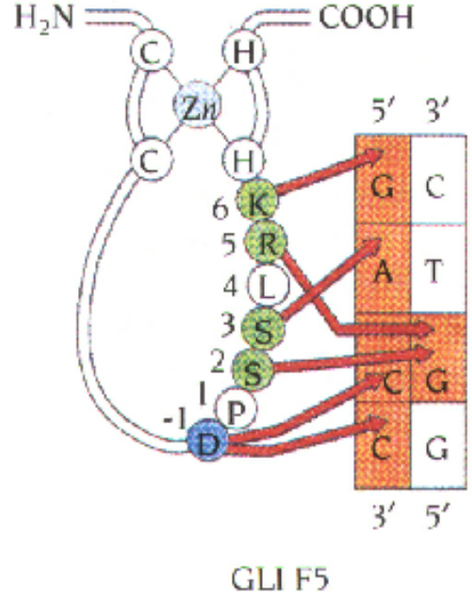
Zif F1



Zif F2



GLI F4

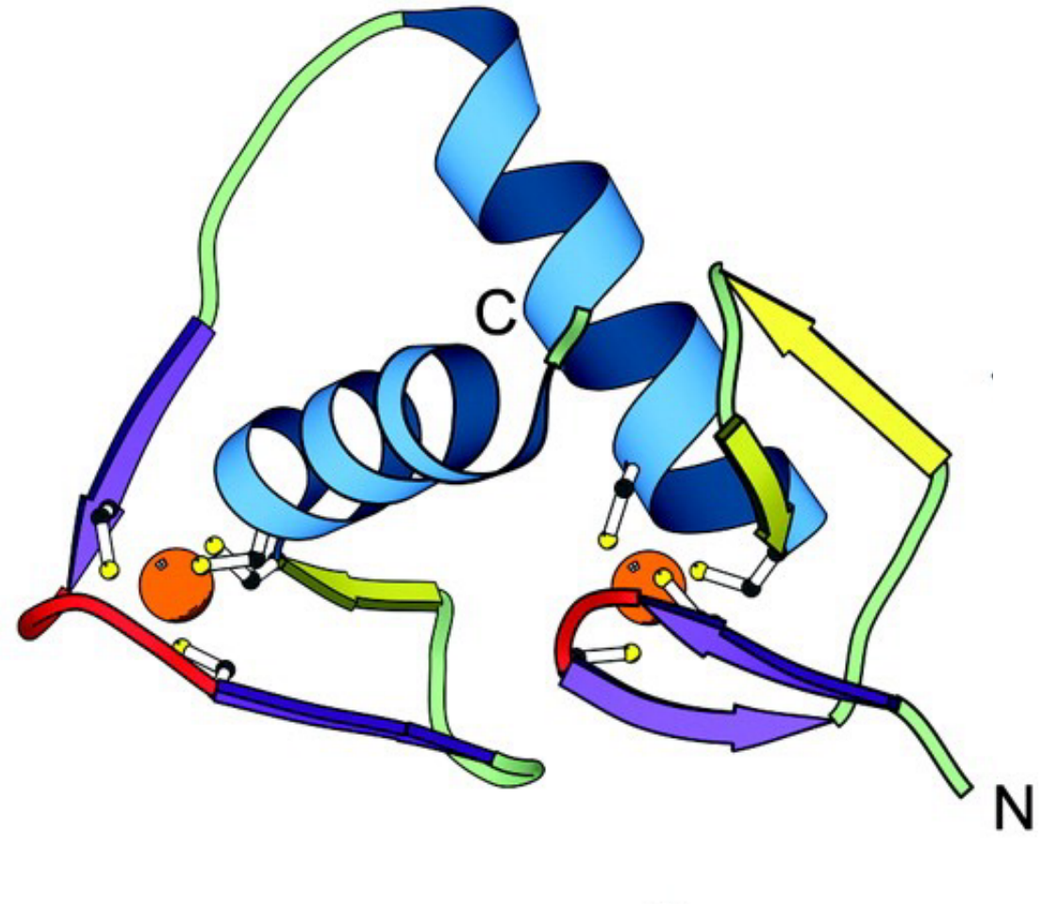


GLI F5

Zinc-containing DNA binding domains

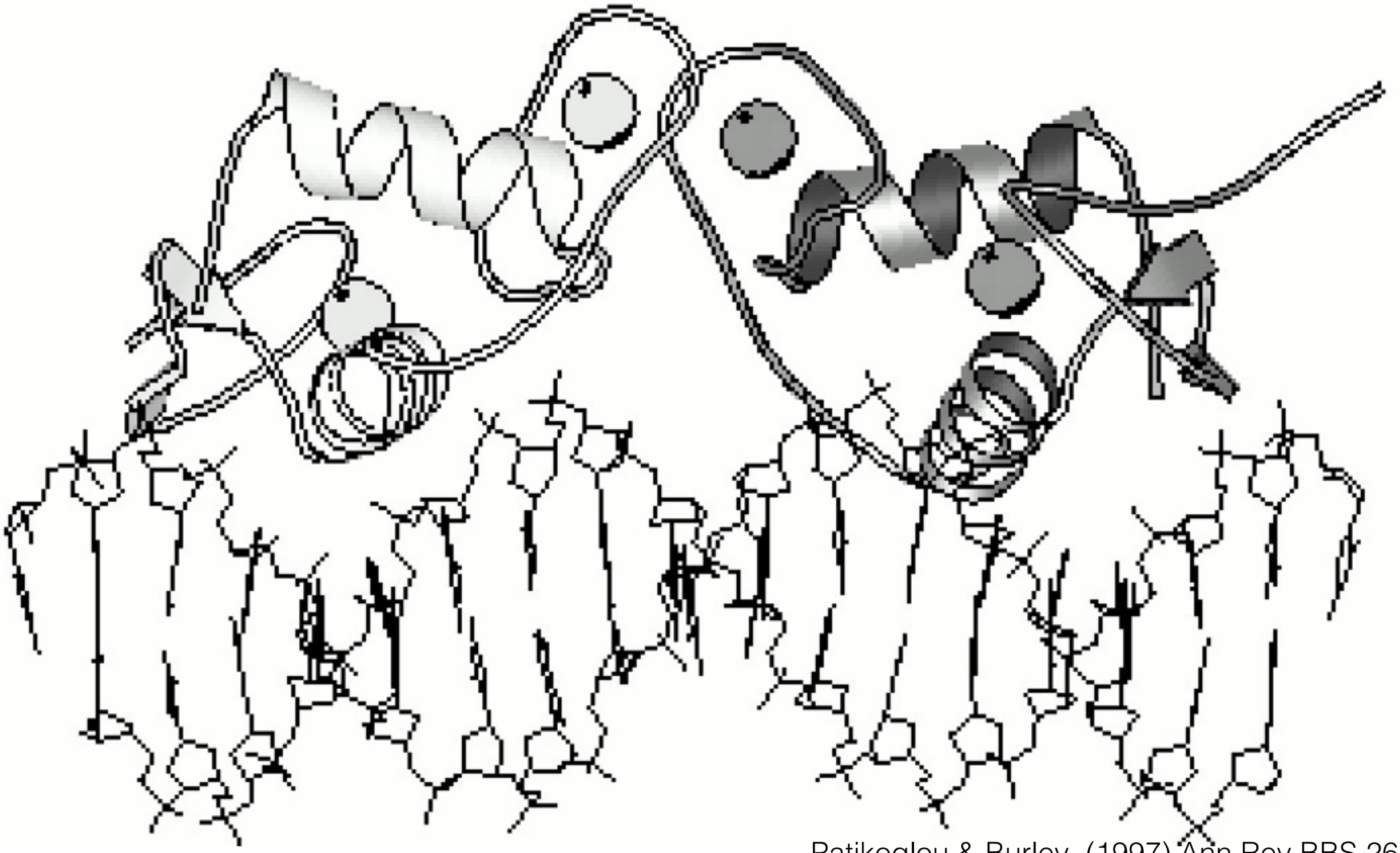
(Zinc is coordinated with a combination of four Cysteine residues)

- Treble-clef Class:
  - β-hairpin at the N-terminus and an α-helix at the C-terminus that each contribute two ligands for zinc binding (a loop and a second β-hairpin of varying length and conformation can be present between the N-terminal β-hairpin and the C-terminal α-helix)



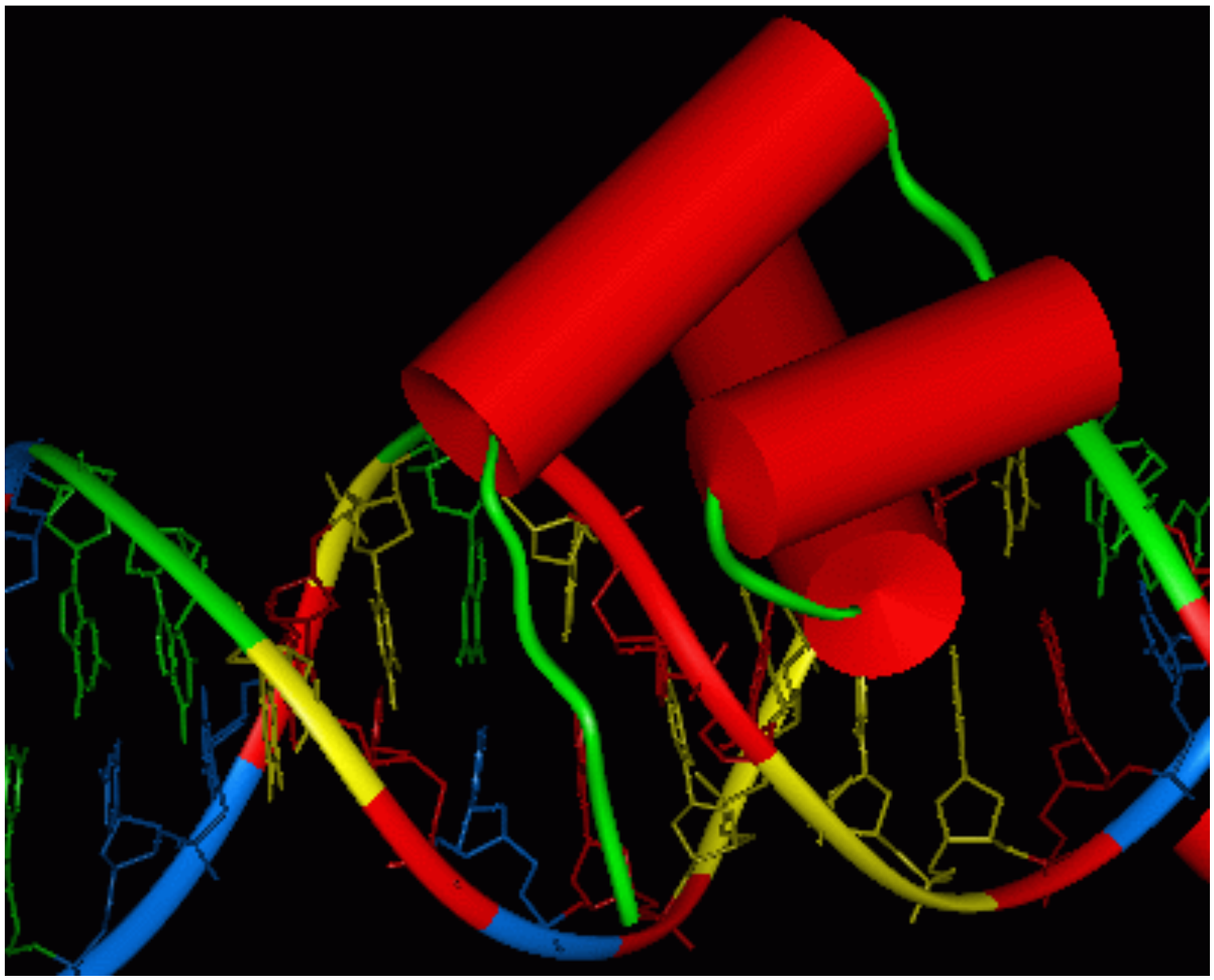
Estrogen Receptor/DNA Complex

(Zinc is coordinated with a combination of four Cysteine residues)



Patikoglou & Burley (1997) Ann Rev BBS 26:289

Engrailed Homeodomain/ DNA Complex

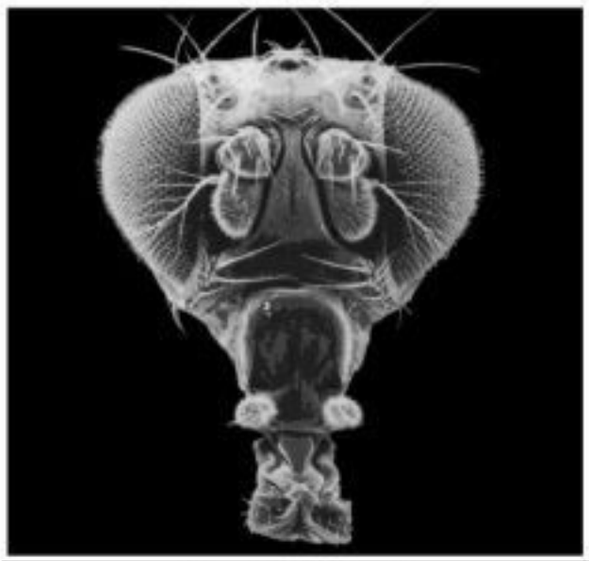


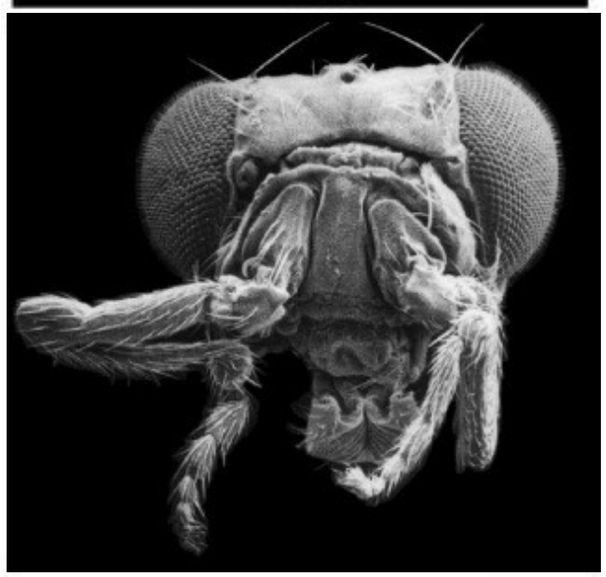
Both major and minor groove interactions by helix 3 and the N-terminal arm respectively

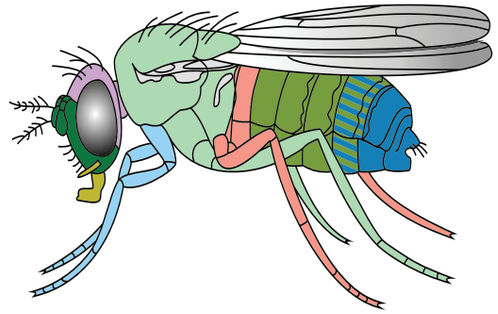
Song Tan lab: Engrailed

http://www.personal.psu.edu/sxt30/movies/engrdna_h264.mov

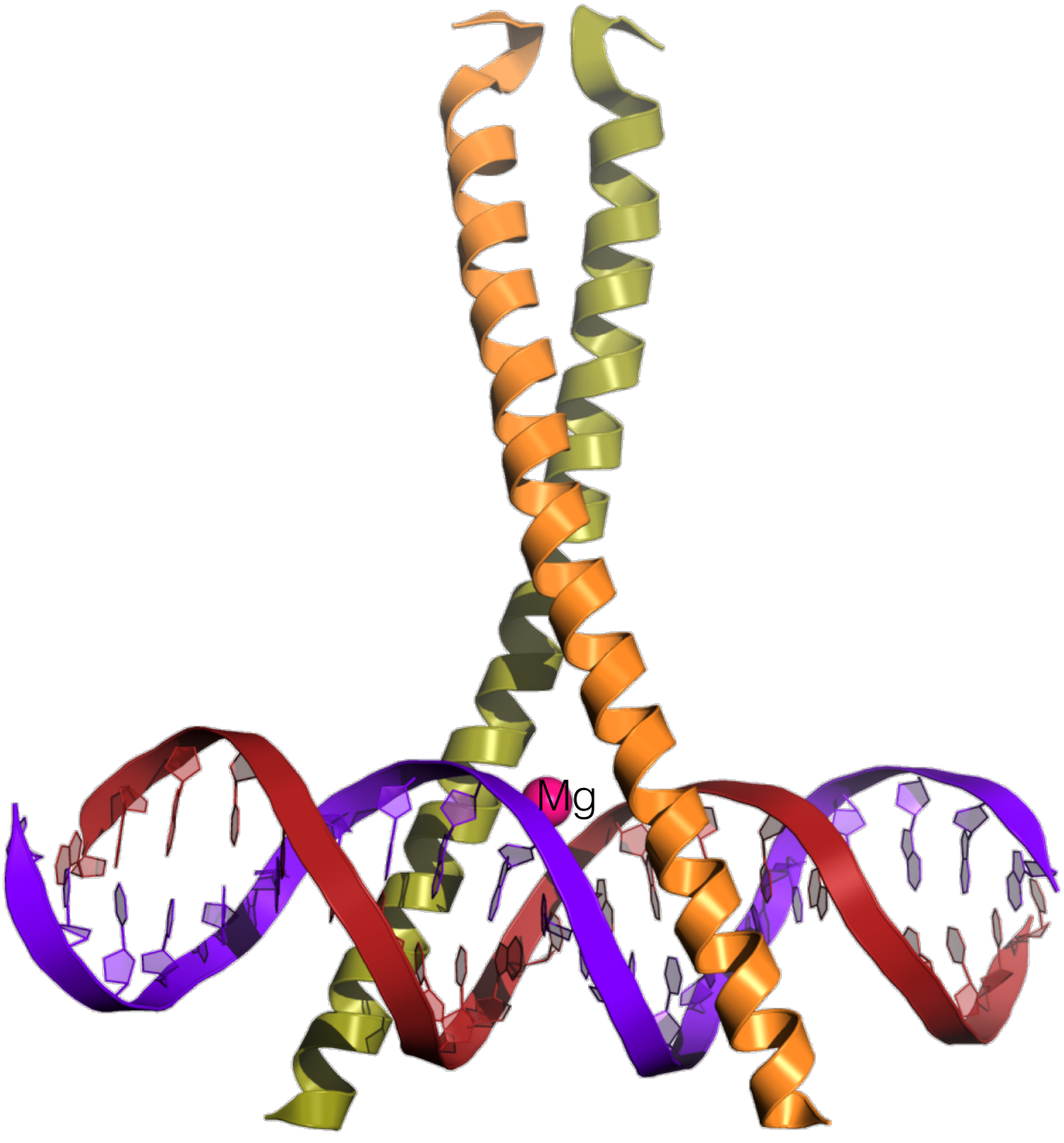
Engrailed Homeodomain (Hox genes also have homeodomains)





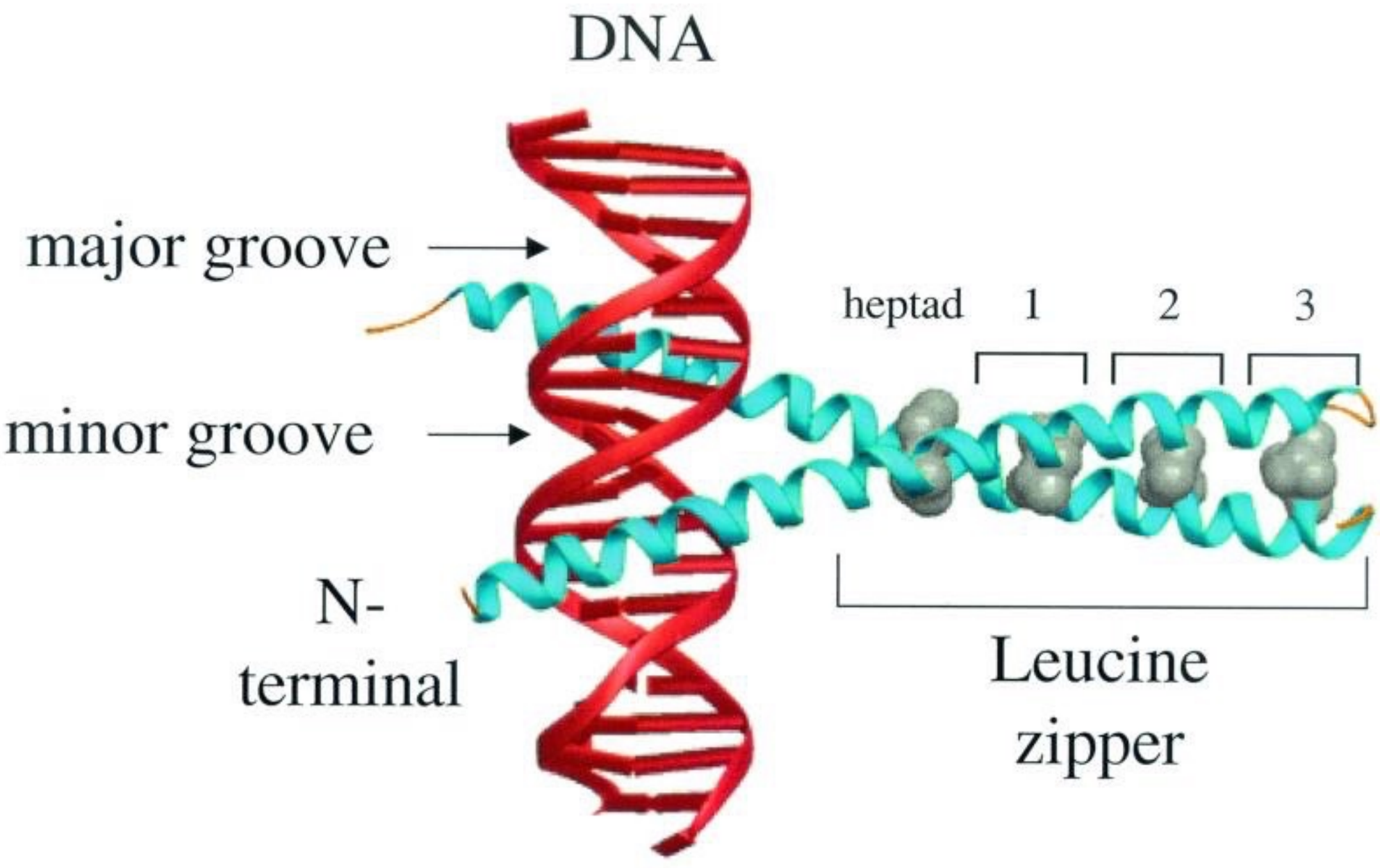


Leucine Zippers: dimerization of TFs

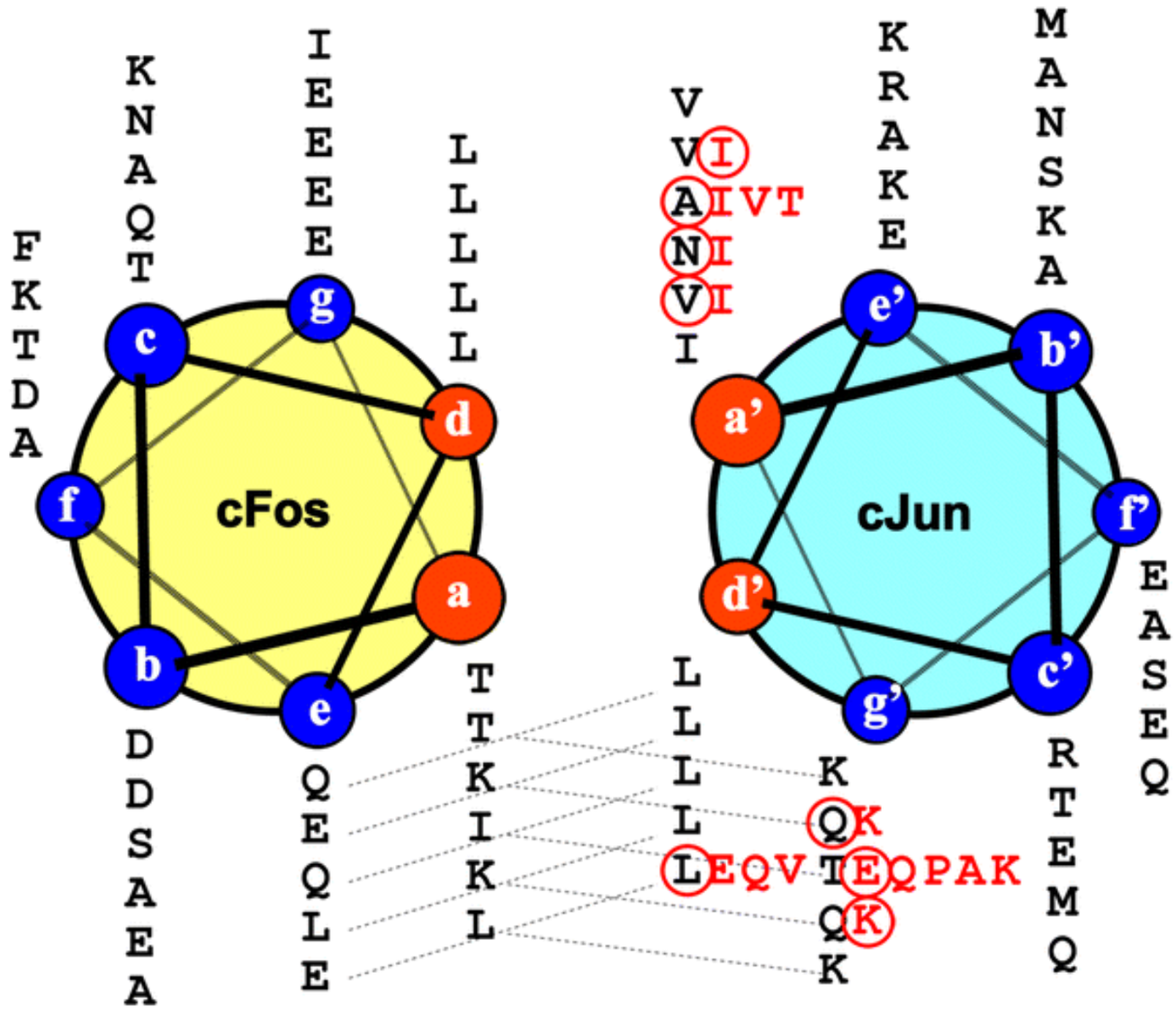


ATF1

Leucine Zippers: dimerization of TFs

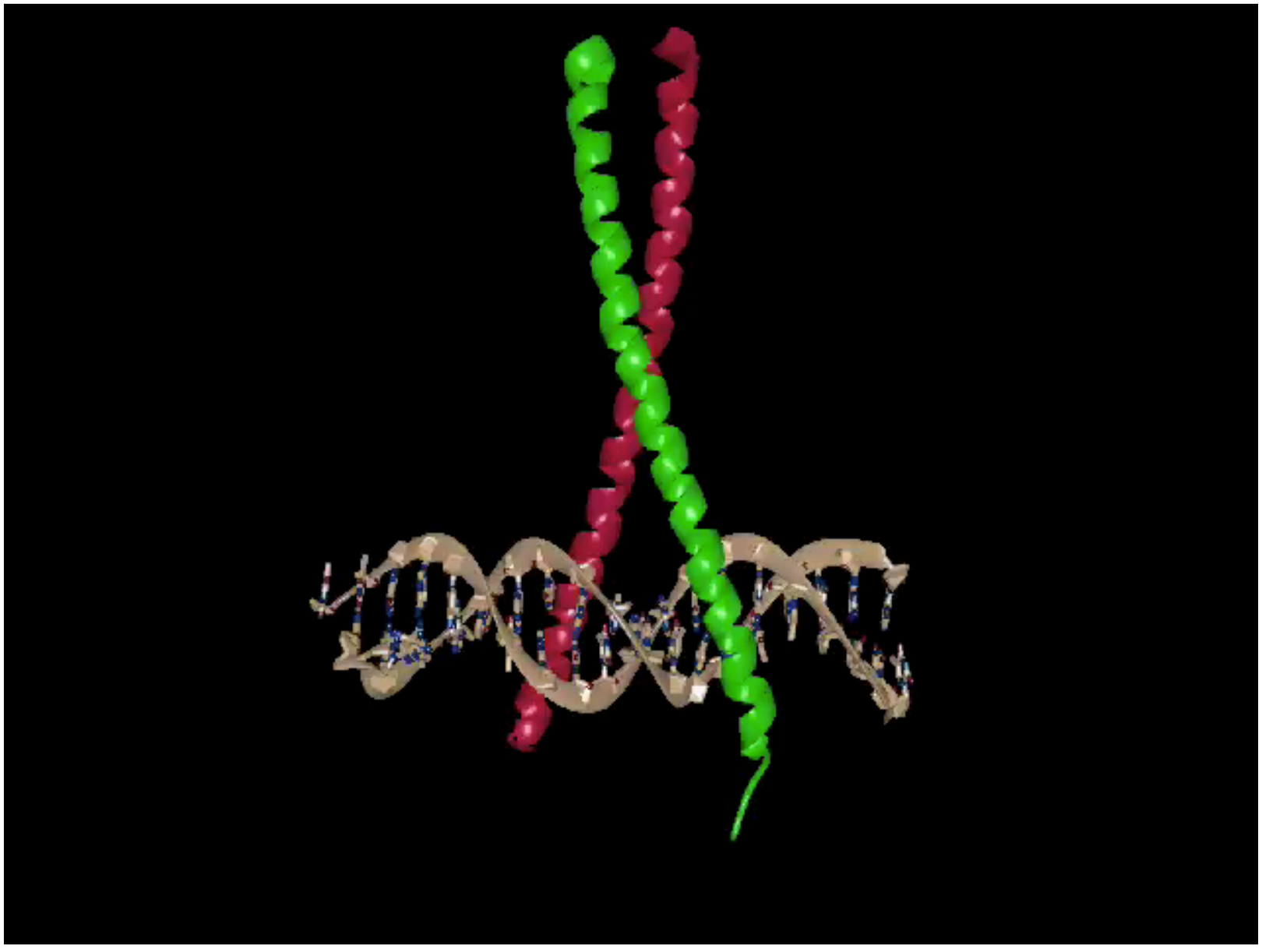


Leucine Zippers: dimerization of TFs



Azuma et. al., Chem Commun 2014 Jun 18;50(48):6364-7

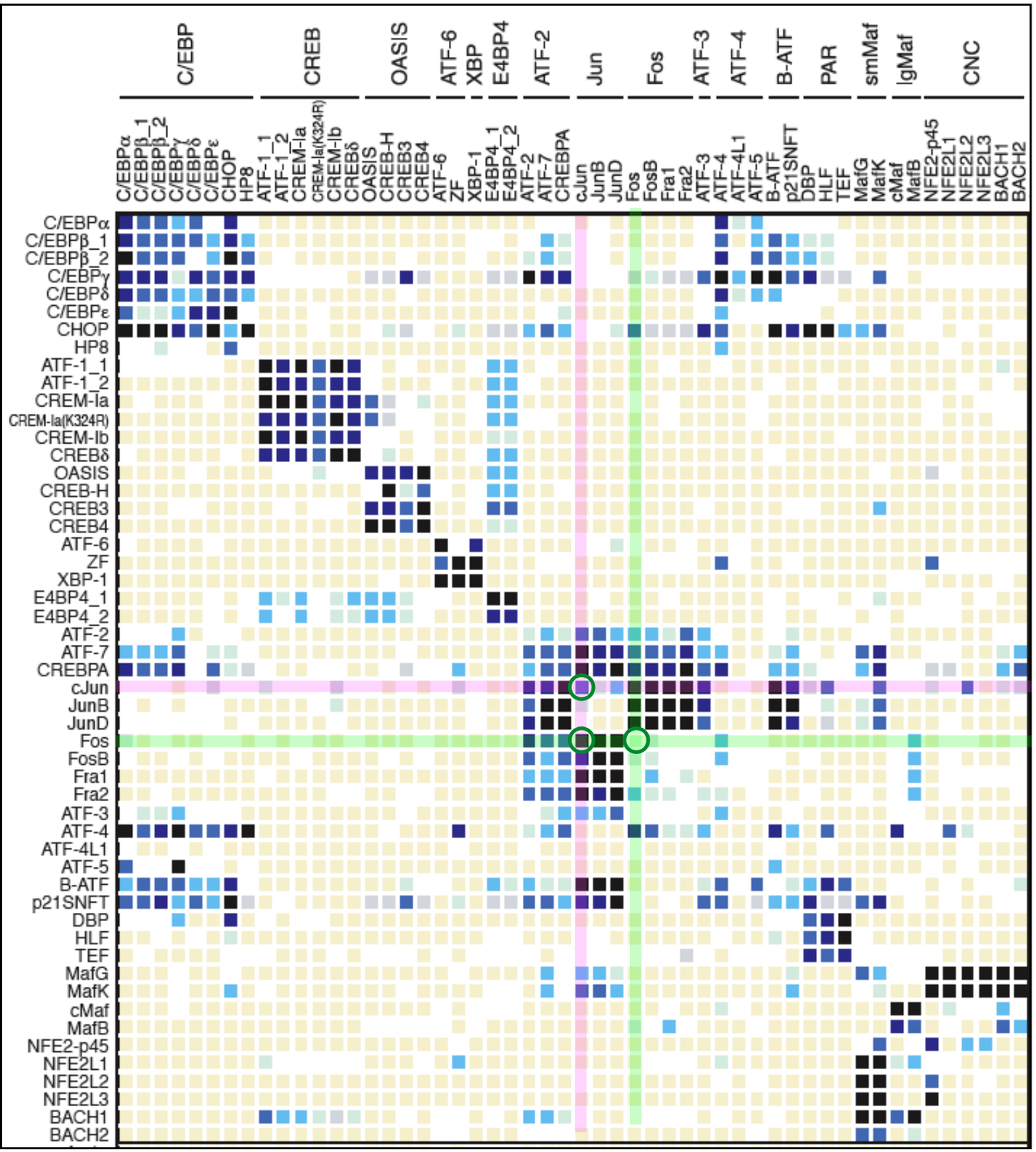
Song Tan Lab: bZIP



http://www.personal.psu.edu/sxt30/gallery_protdna.html

Have you heard of the transcription factor AP1?

Interaction Matrix for 49 human bZIP Peptides

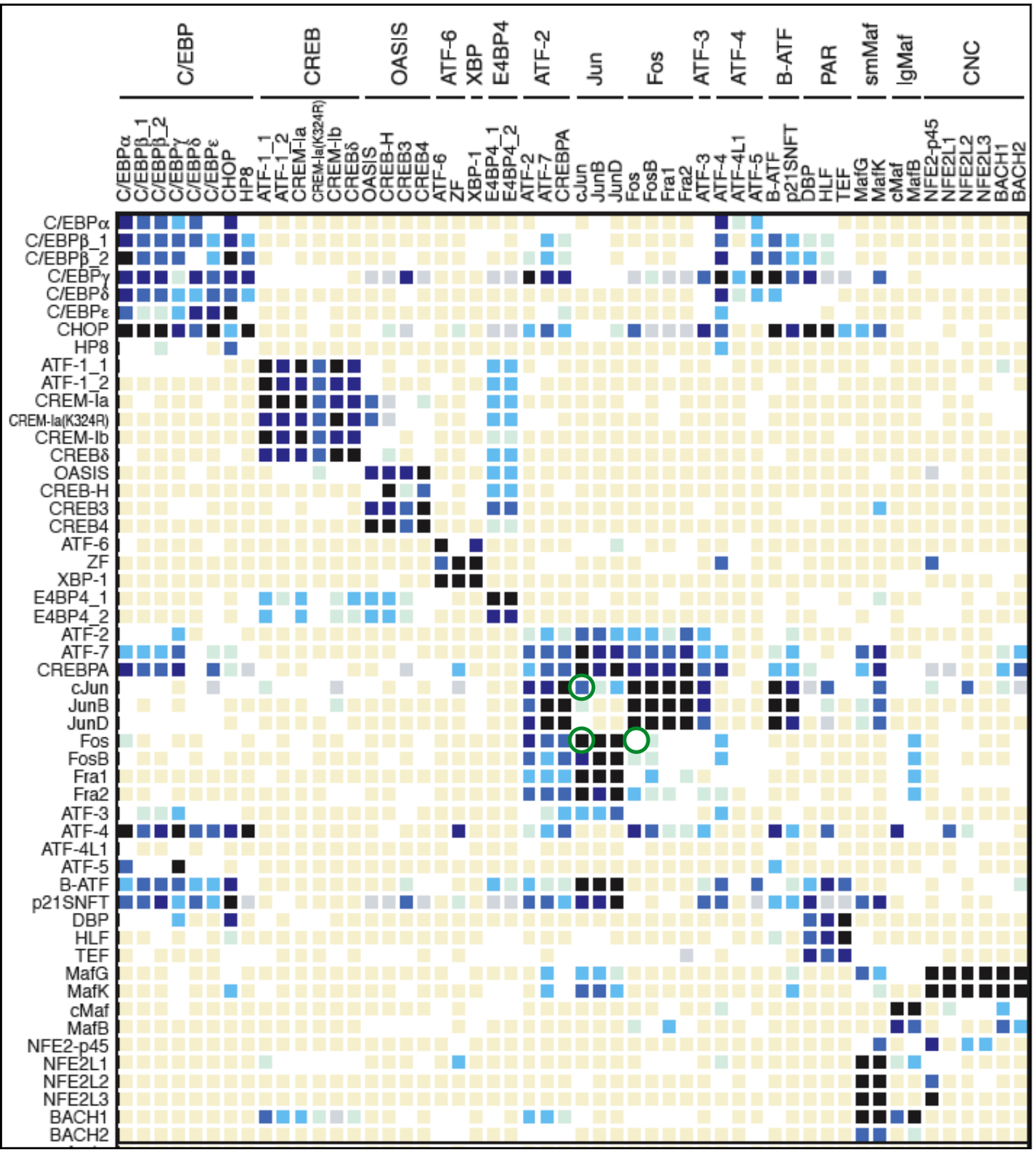


~14% of possible interactions detected.

Most between family members, but 136 between families.

Newman & Keating (2003) Science 300:2097

Leucine Zippers Provide Specific Dimerization Interactions



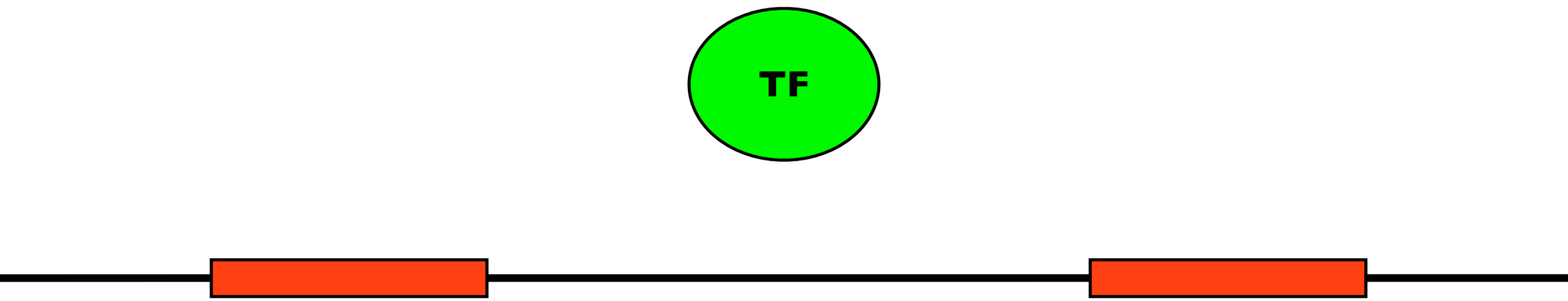
cJun/cFos heterodimer forms preferentially relative to homodimers

cJun/cJun forms but has two unfavorable charge interactions.

cFos/cFos does not form - four unfavorable charge interactions.

Newman & Keating (2003) Science 300:2097

Chromatin affects TF binding in vivo



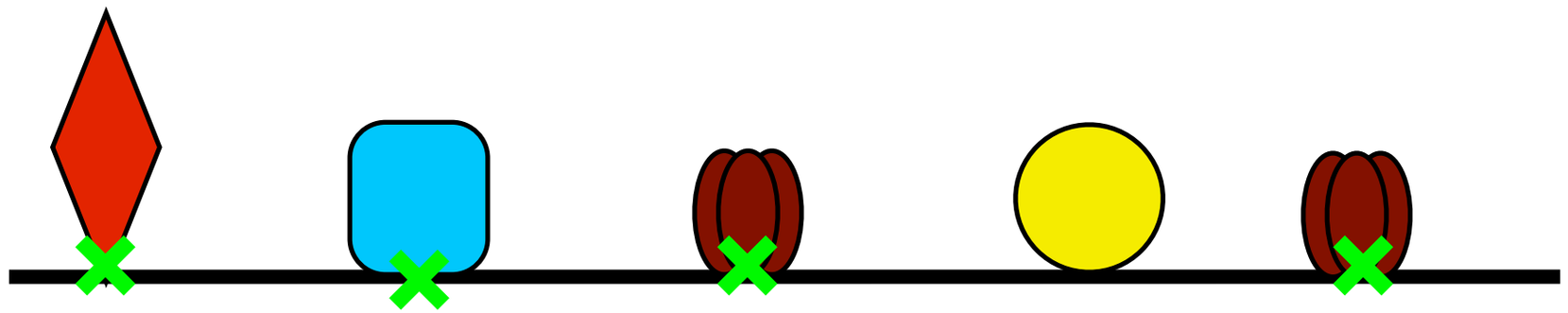
-Besides sequence, what influences TF binding?

HSF binds many sites after HS

Westwood, Clos & Wu Nature 1991

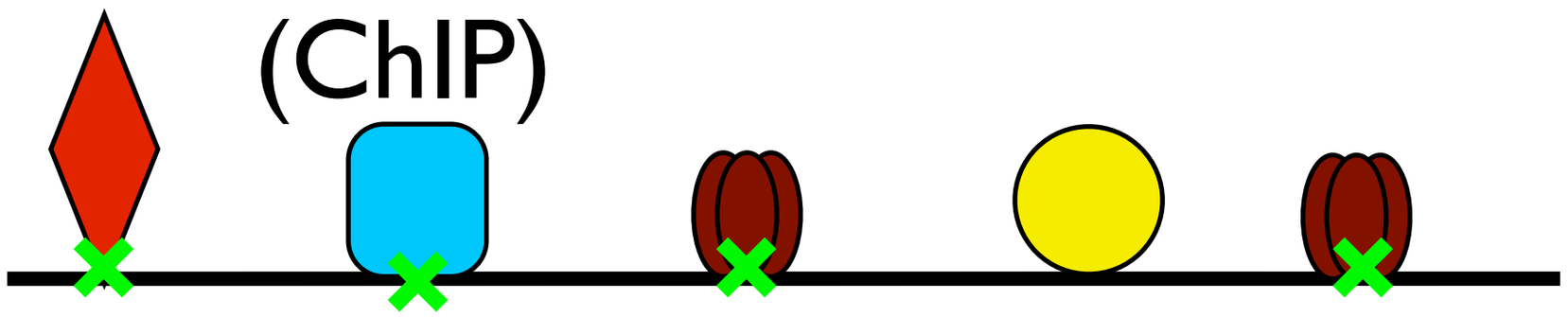
Chromatin Immunoprecipitation (ChIP)

Crosslink DNA and Proteins

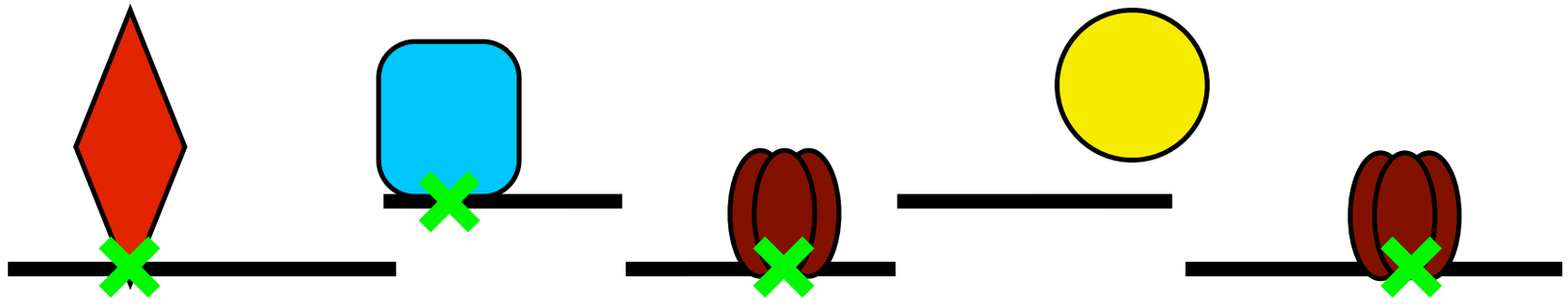


Chromatin Immunoprecipitation (ChIP)

Crosslink DNA and Proteins



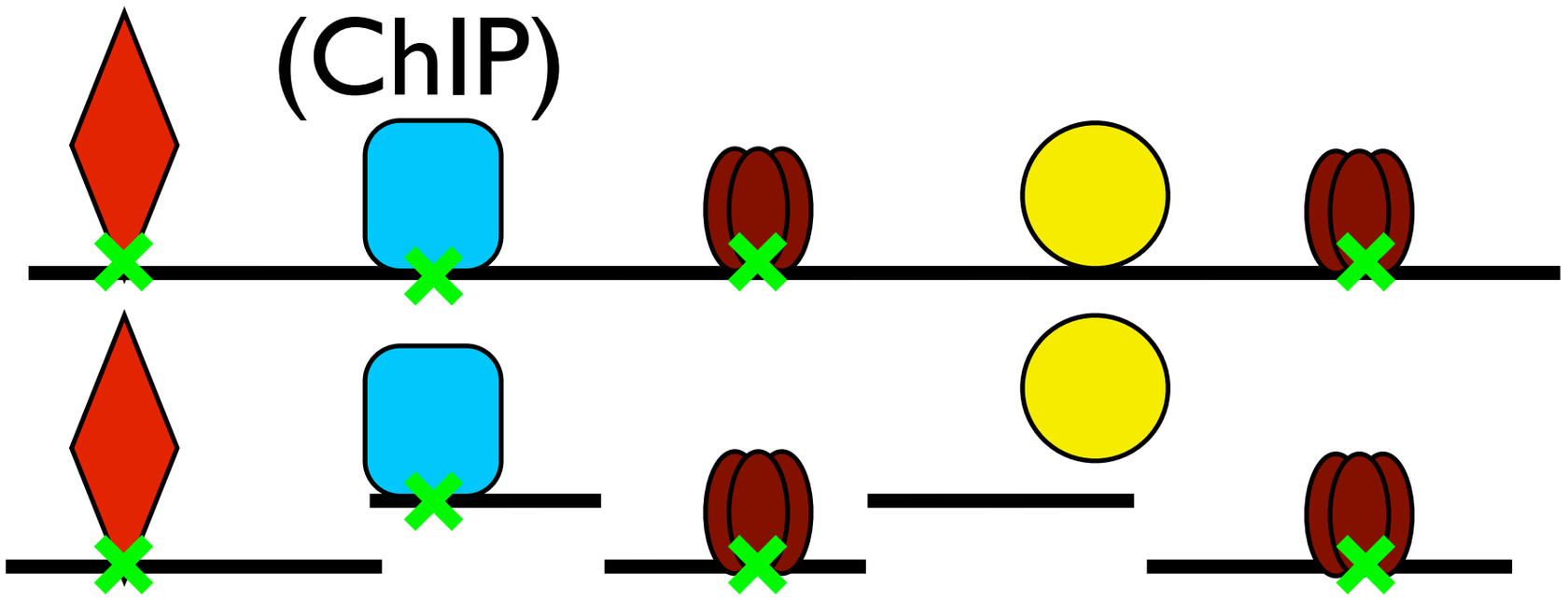
Shear DNA



Chromatin Immunoprecipitation (ChIP)

Crosslink DNA and Proteins

Shear DNA



Immunoprecipitate

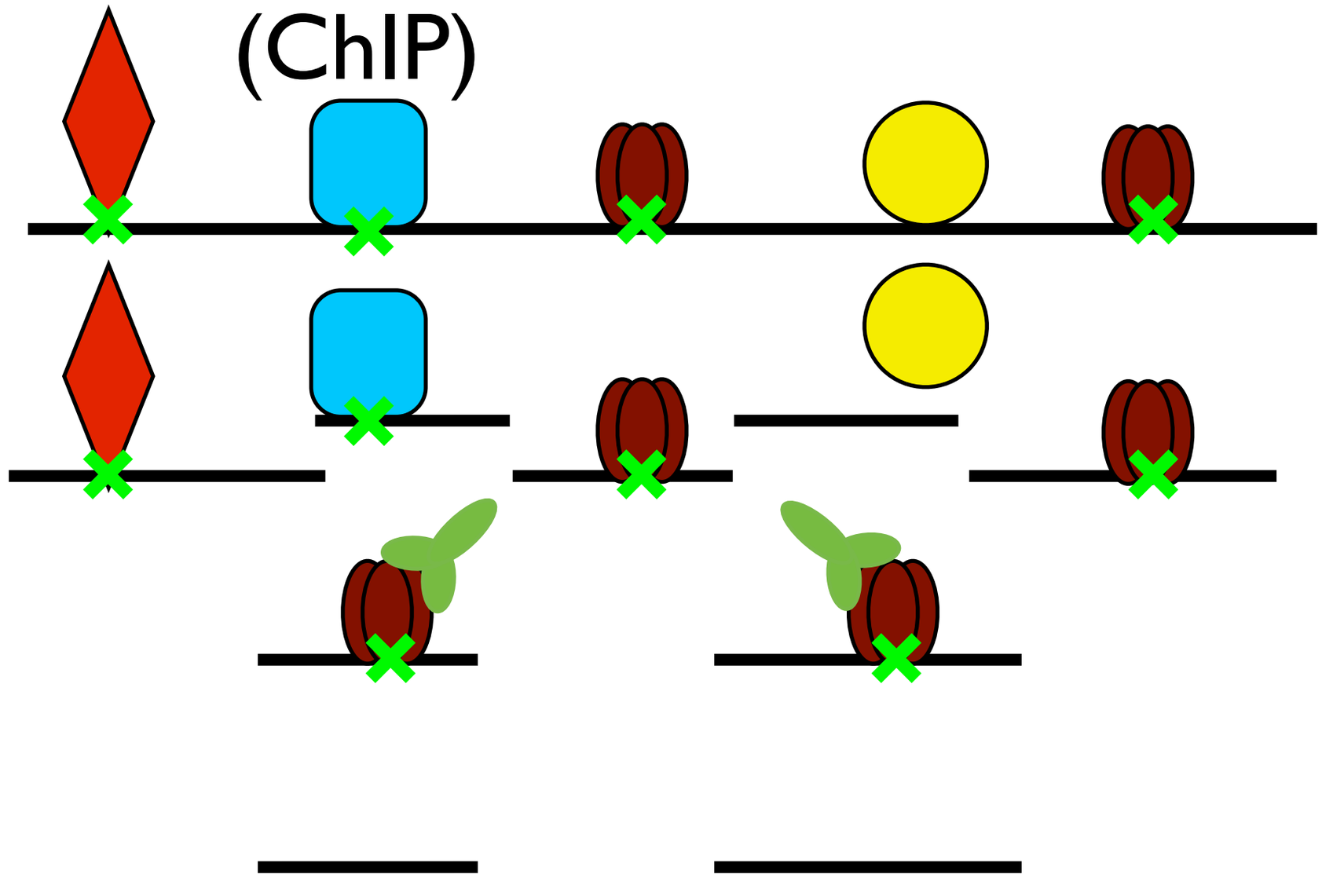
Chromatin Immunoprecipitation (ChIP)

Crosslink DNA and Proteins

Shear DNA

Immunoprecipitate

Purify DNA



Gilmour and Lis, Molecular and Cell Biology 1985

Chromatin Immunoprecipitation (ChIP-seq)

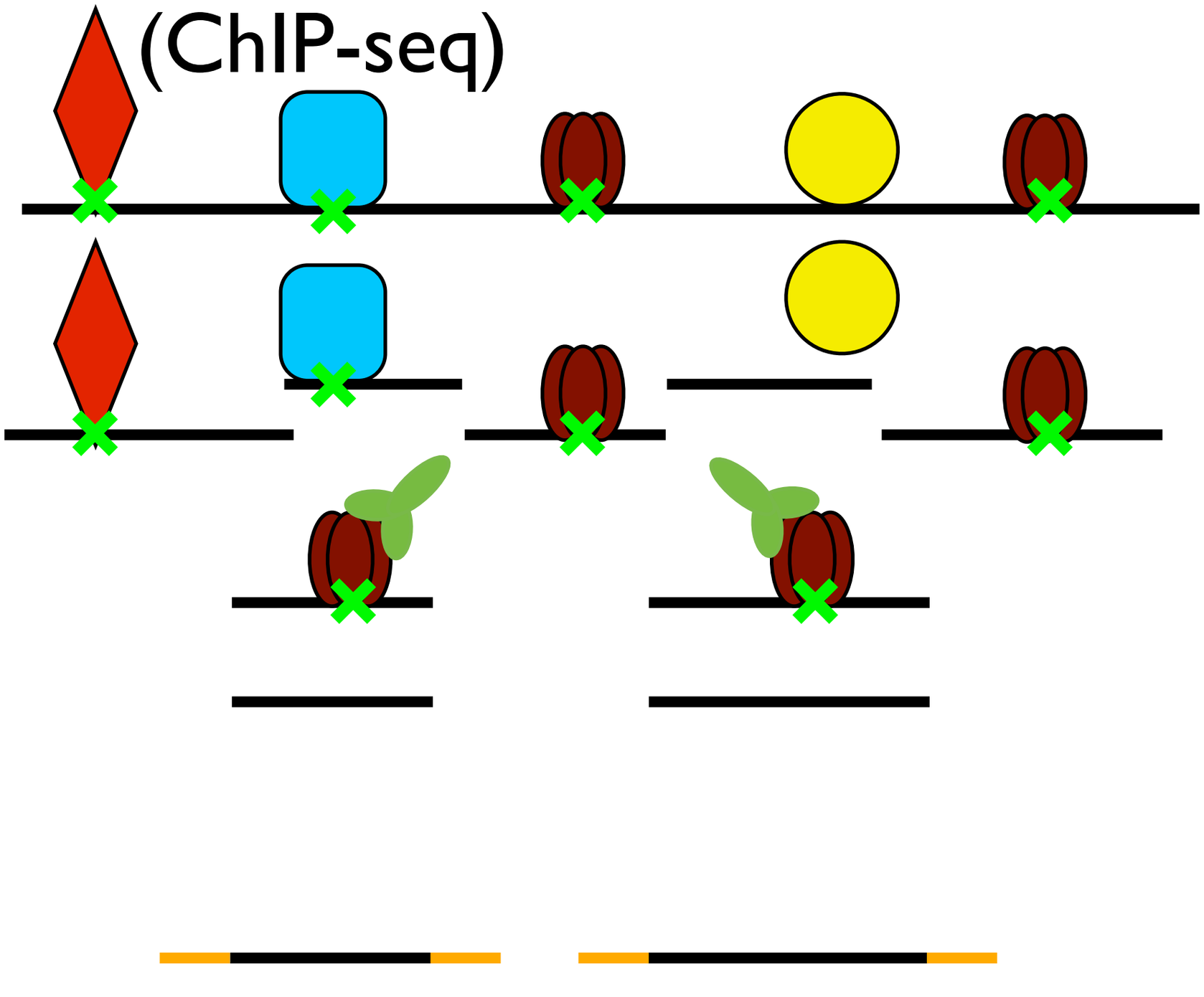
Crosslink DNA and Proteins

Shear DNA

Immunoprecipitate

Purify DNA

Ligate adapters



Mikkelsen et. al., Nature 2007; Johnson et al., Science 2007

Chromatin Immunoprecipitation (ChIP-seq)

Crosslink DNA and Proteins

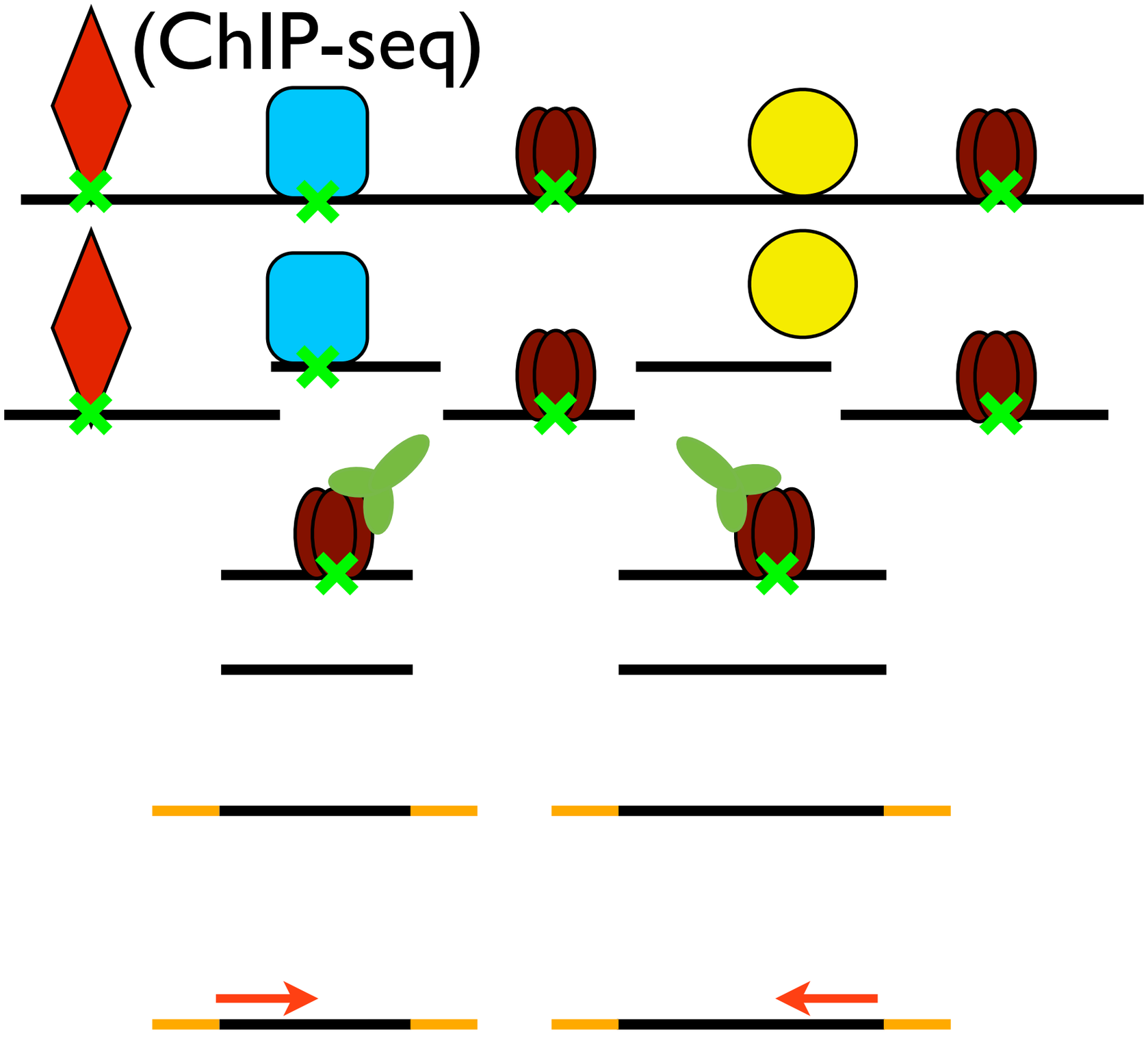
Shear DNA

Immunoprecipitate

Purify DNA

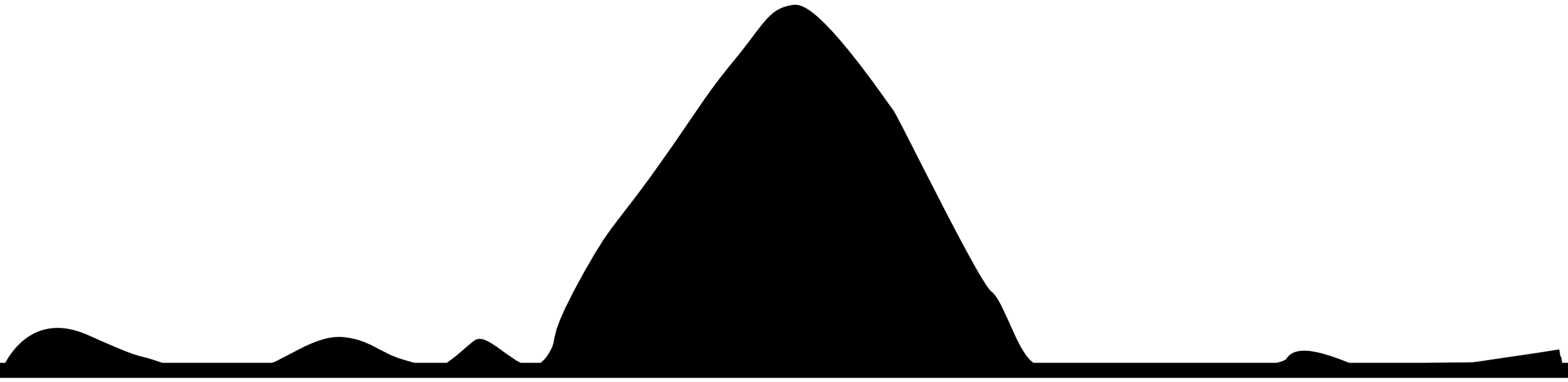
Ligate adapters

Sequence DNA ends



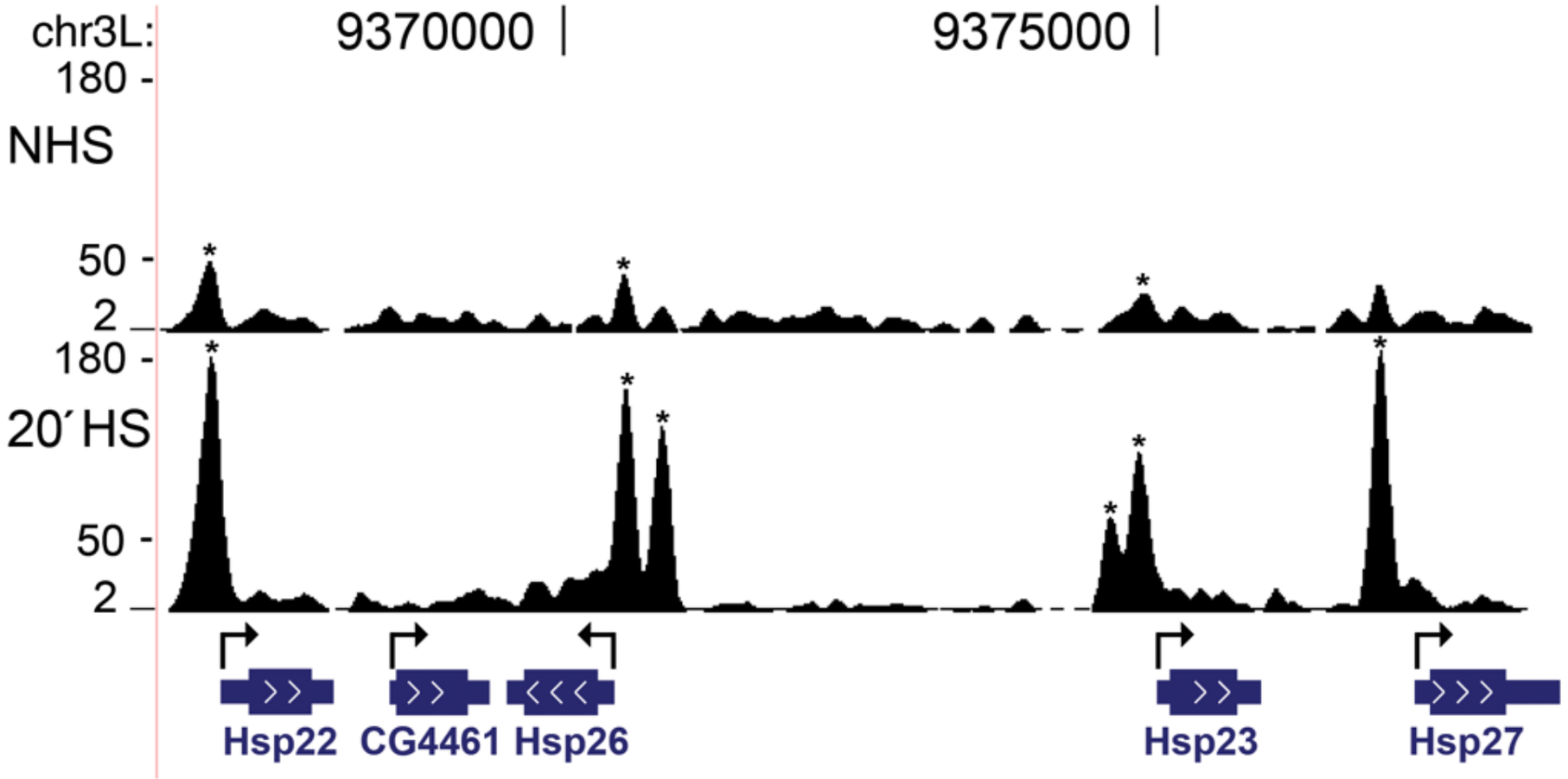
Mikkelsen et. al., Nature 2007; Johnson et al., Science 2007

Chromatin Immunoprecipitation (ChIP-seq)

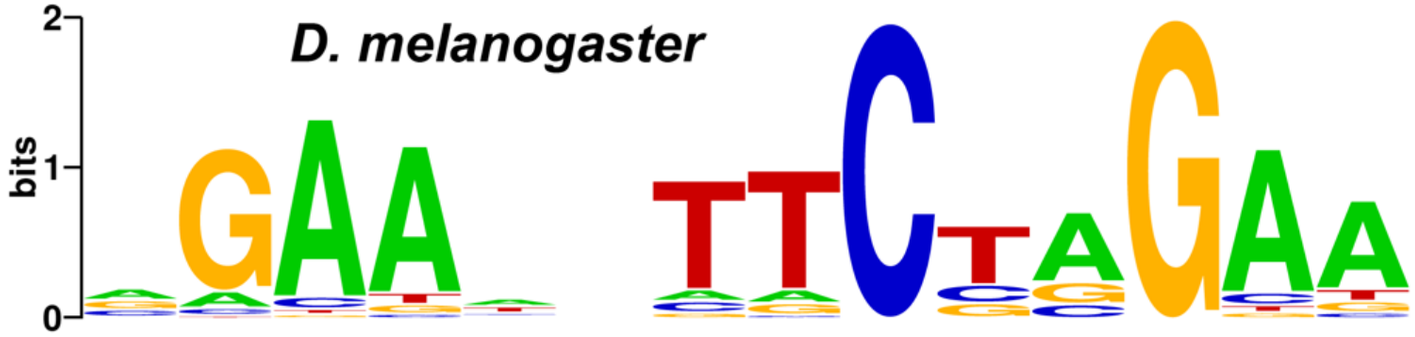


Peak calling: Zhang, et. al., Genome Biology 2008 (and many others)

HSF targets DNA inducibly

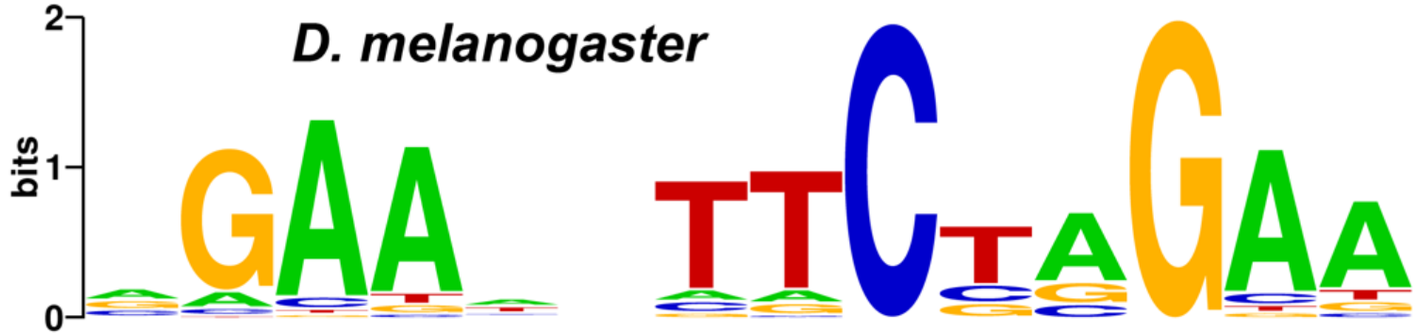


HSF targets a consensus motif



HSF binds a fraction of motifs in vivo

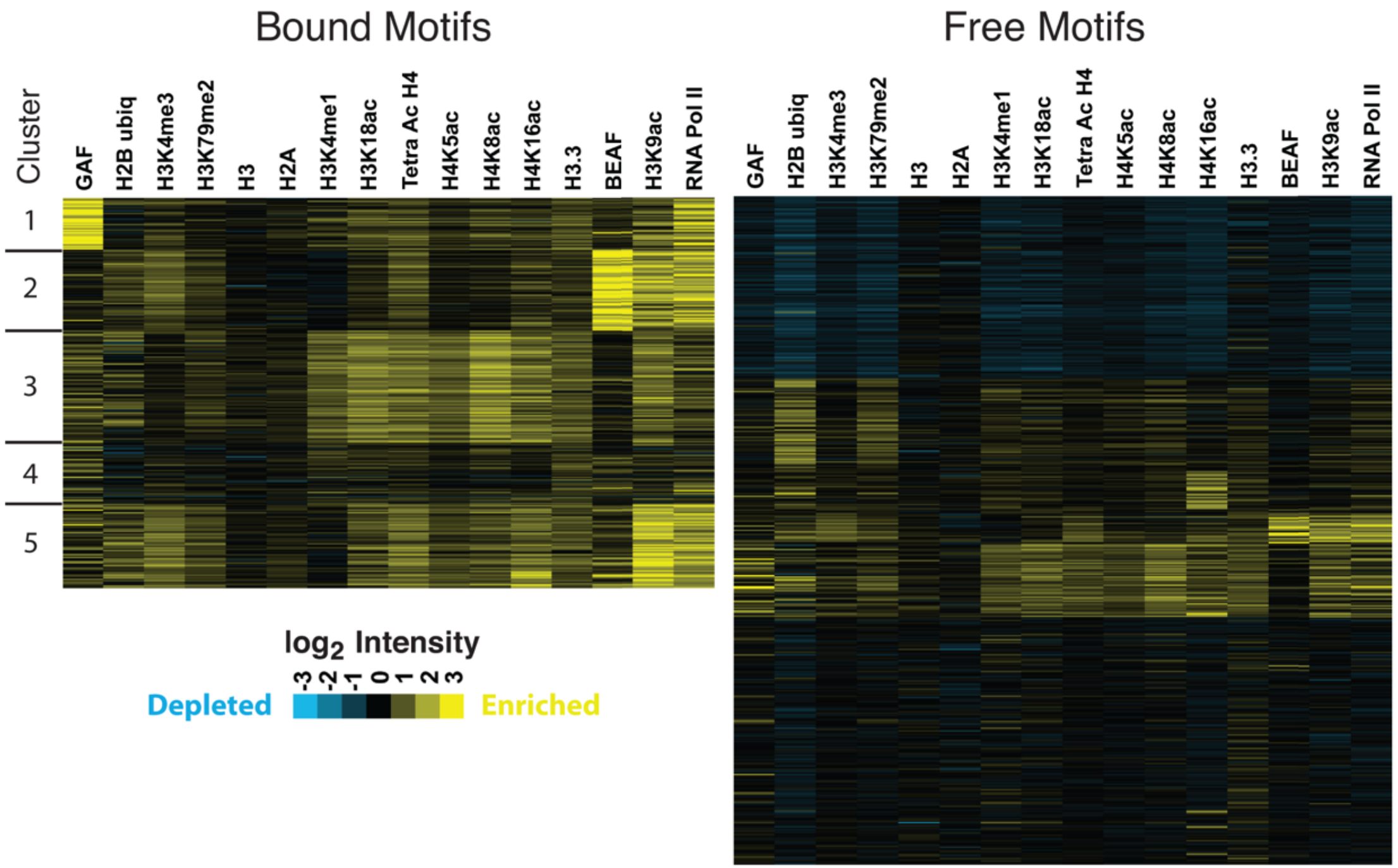
- Queried the *Drosophila* genome using this HSE matrix:



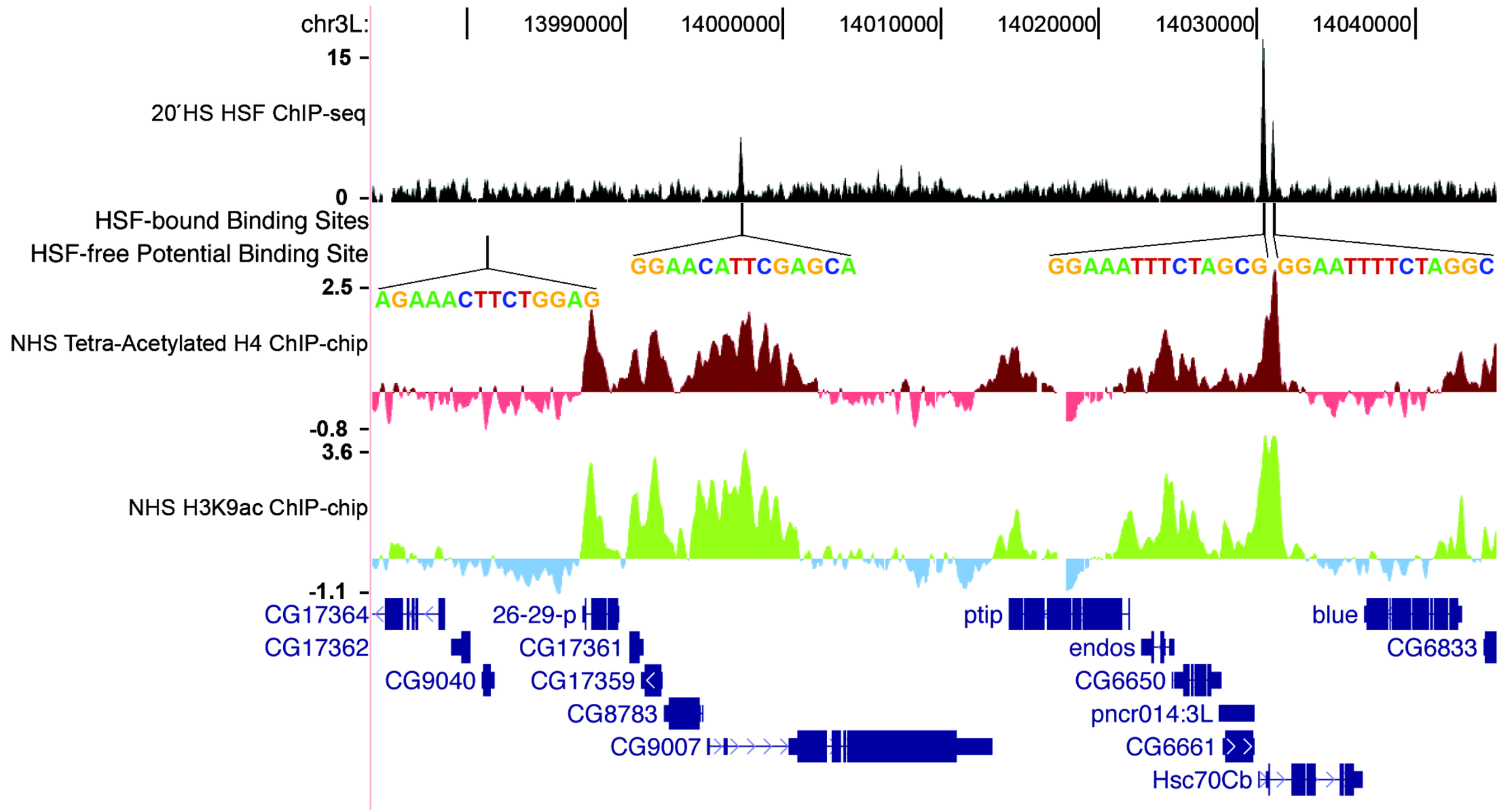
- Found 708 post-HS HSF-free motifs that conform stringently to this consensus HSE, compared to 442 HSF-bound motifs.

- Note that these are computational predictions of potential HSF binding sites

HSF targets motifs within active chromatin

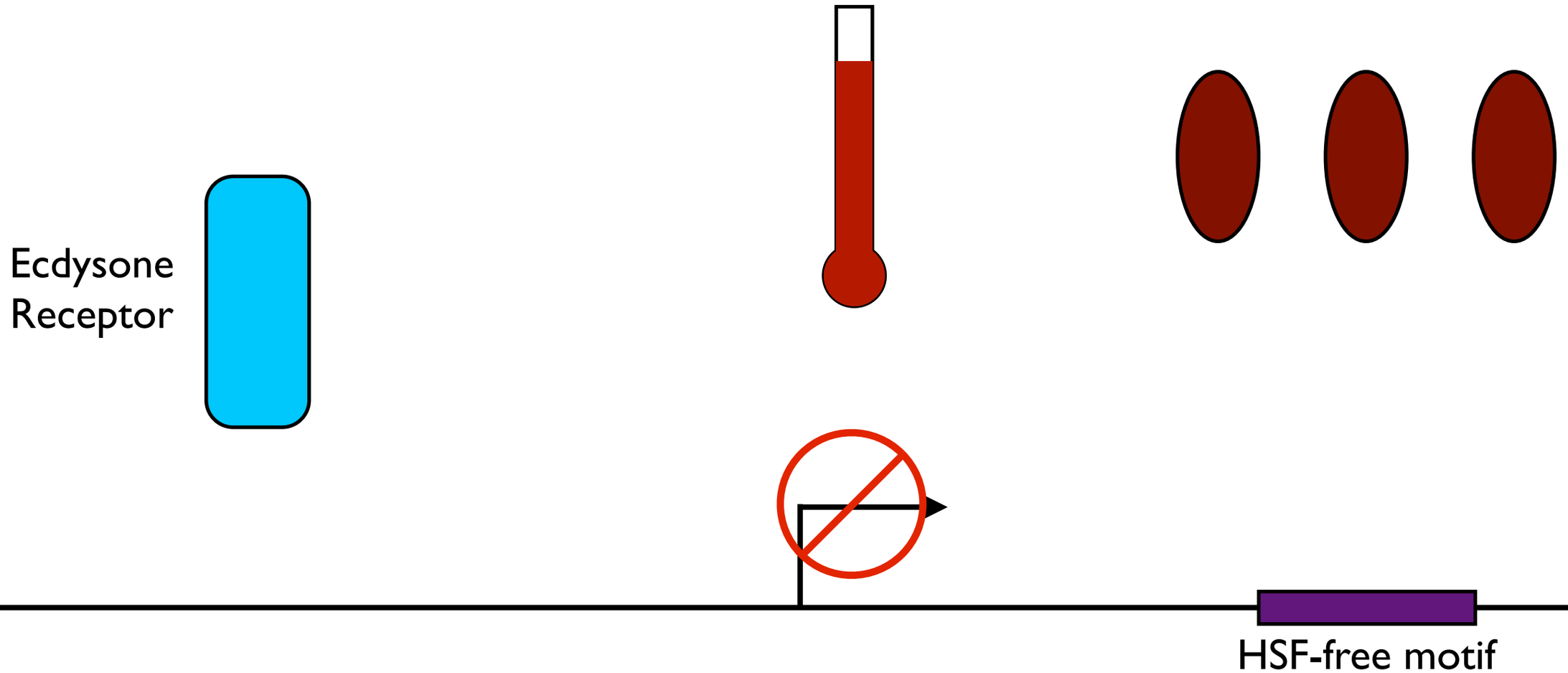


Correlative genomics allows one to develop a hypothesis; we hypothesize that active chromatin permits TF binding

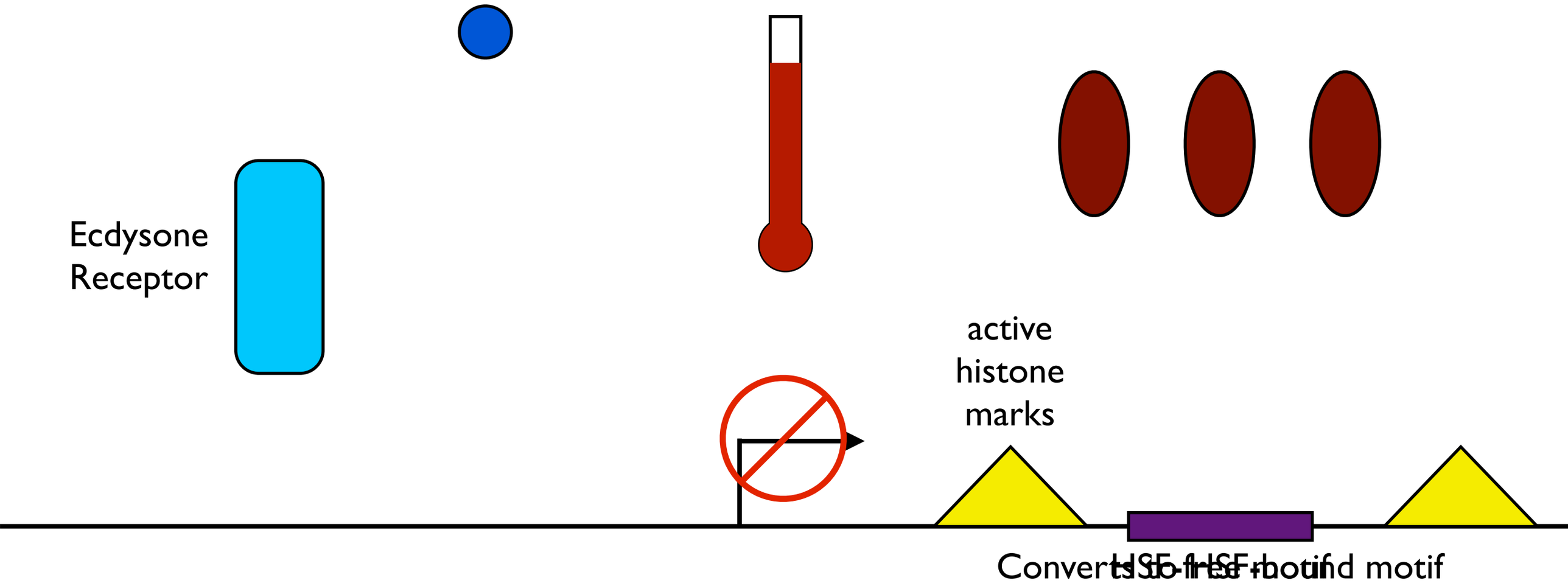


Changing the chromatin at an HSE

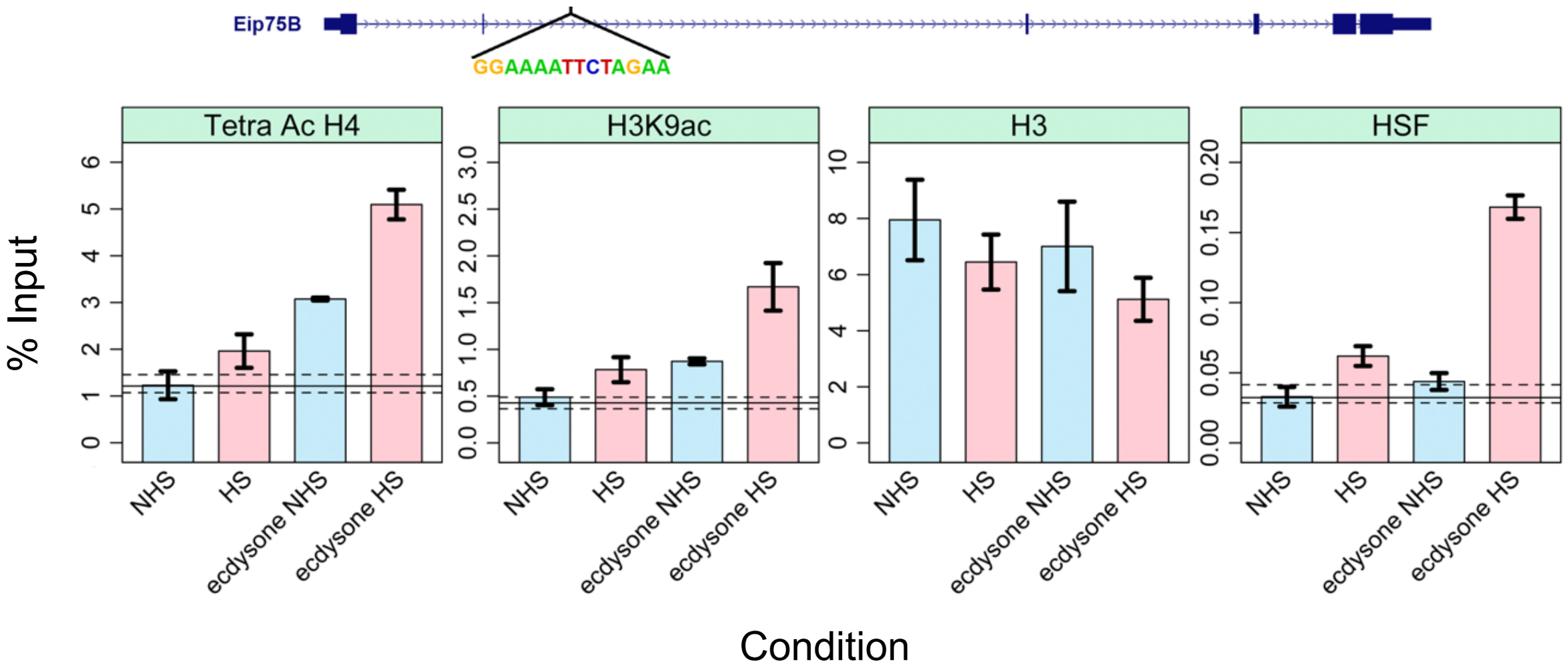
(now this is done trivially with CRISPR-dCas9 coupled to chromatin modifiers)



Changing the chromatin at an HSE



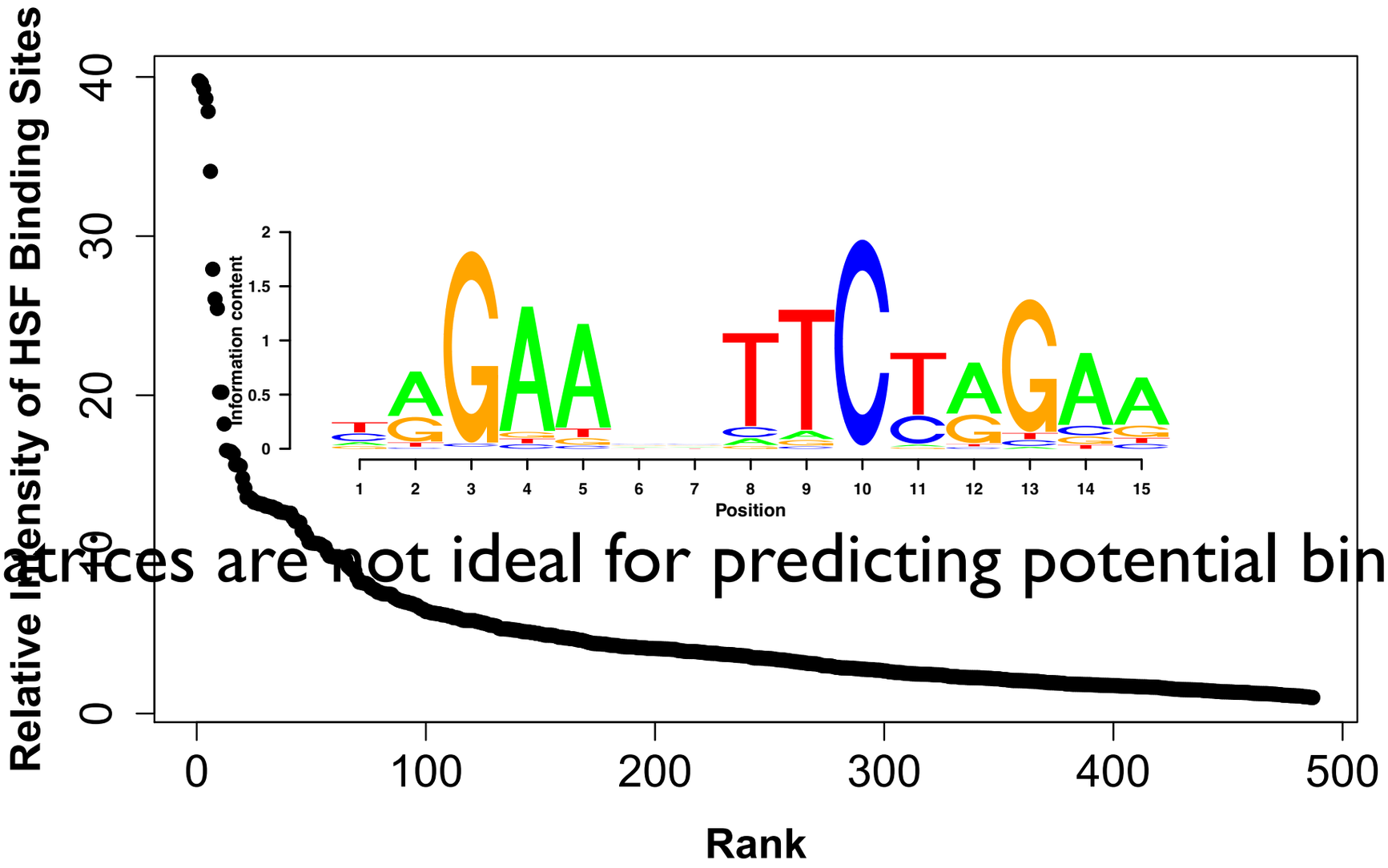
Converting and HSF-free to an HSF-bound motif



Biology is continuous, not discrete

- So far I have considered HSF-bound and HSF-free as two separate categories.
  - In reality, some low intensity binding sites look more like unbound regions and the intensity of binding is meaningful.
- Can we predict inducible TF binding intensity?

Predicting TF binding intensities



Weight matrices are not ideal for predicting potential binding affinity

in vitro nucleic acid/protein binding (PB-seq)

Isolate genomic DNA

in vitro nucleic acid/protein binding (PB-seq)

Isolate genomic DNA

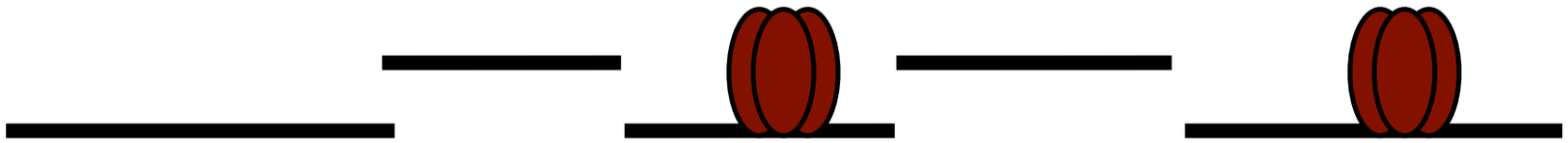
Shear DNA

in vitro nucleic acid/protein binding (PB-seq)

Isolate genomic DNA

Shear DNA

Incubate with Recombinant HSF



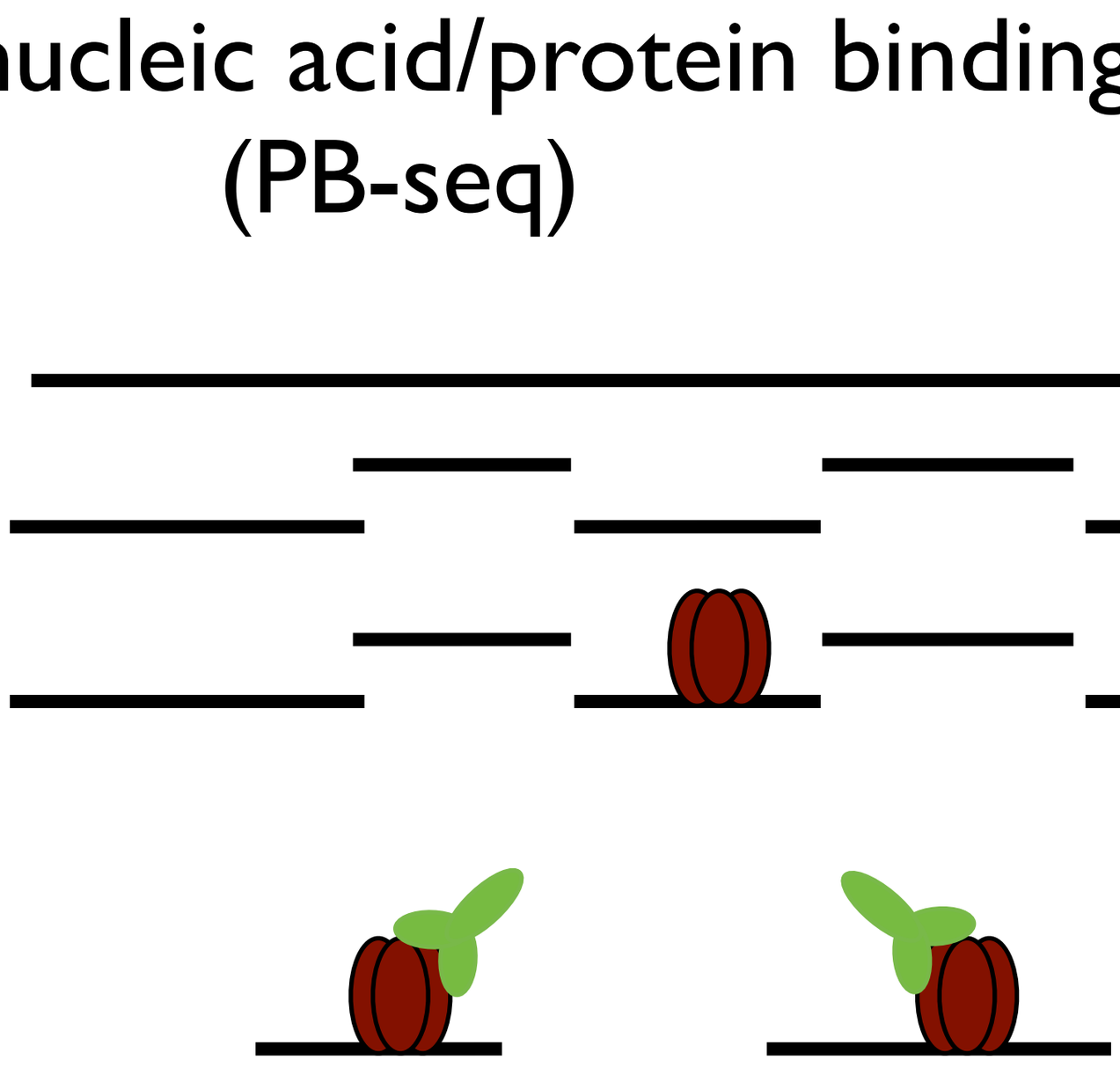
in vitro nucleic acid/protein binding (PB-seq)

Isolate genomic DNA

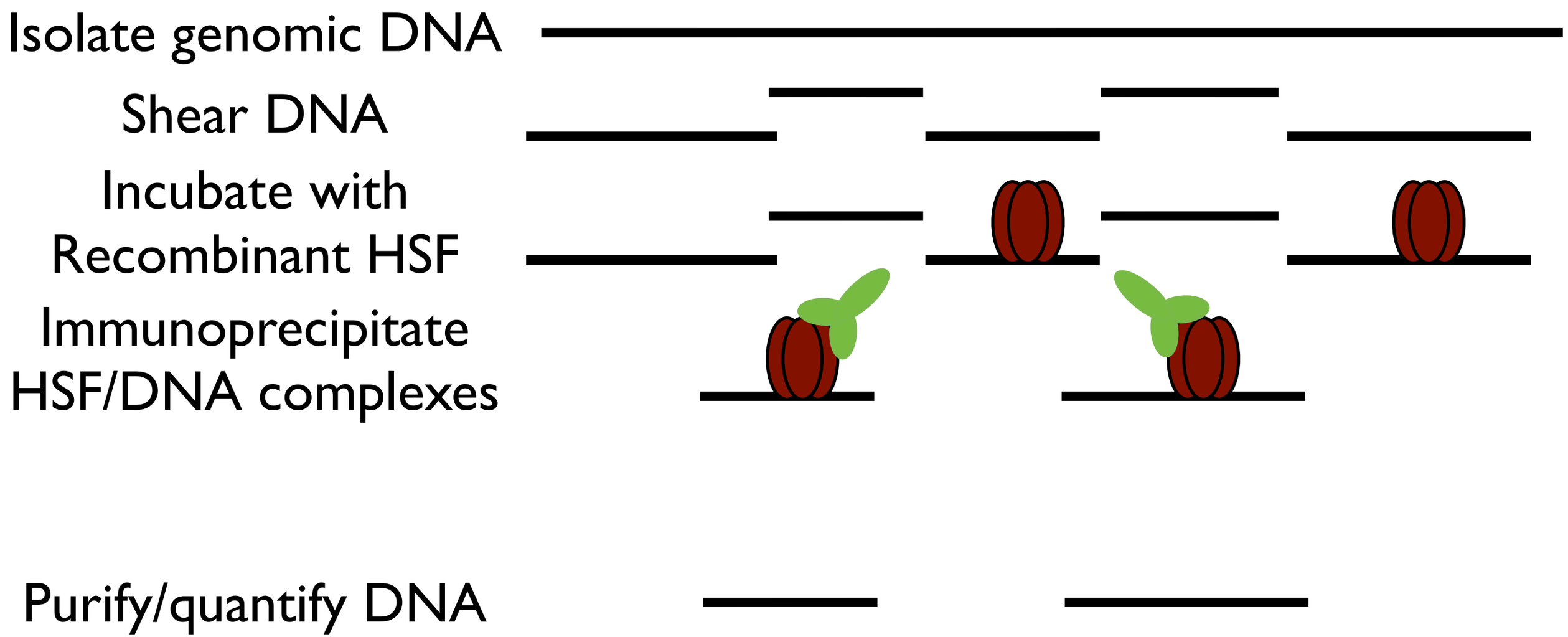
Shear DNA

Incubate with Recombinant HSF

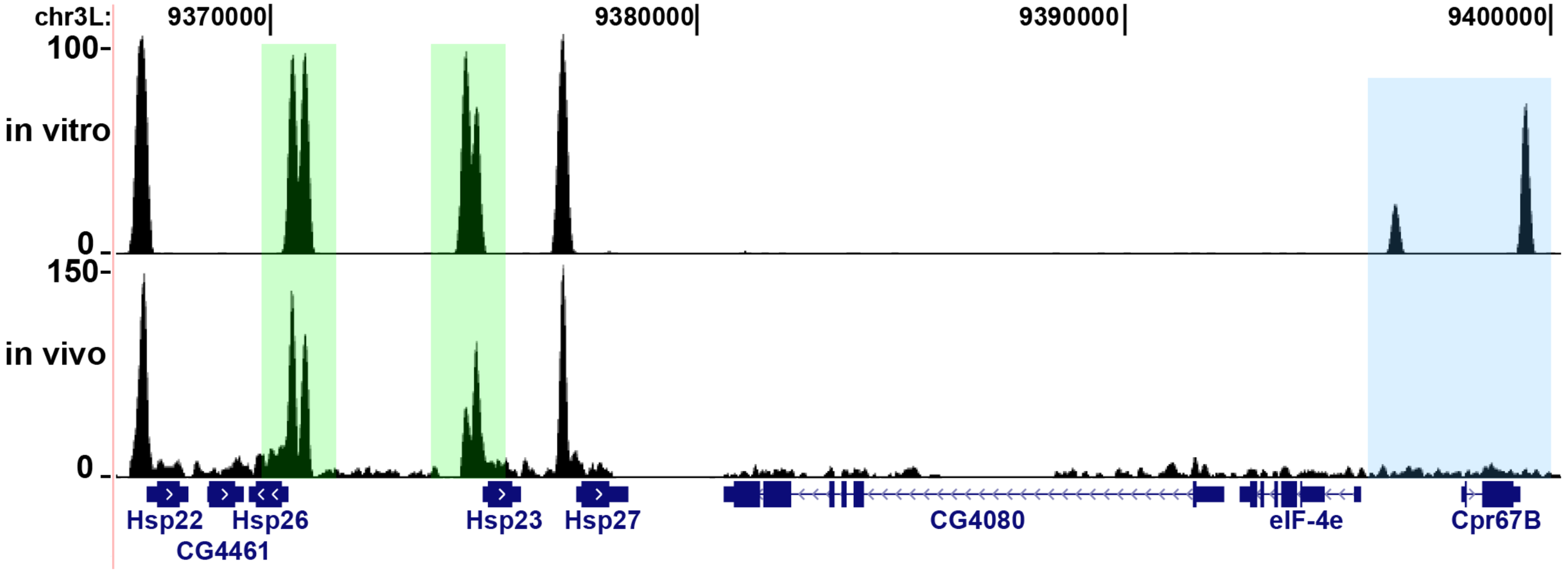
Immunoprecipitate HSF/DNA complexes



in vitro nucleic acid/protein binding (PB-seq)

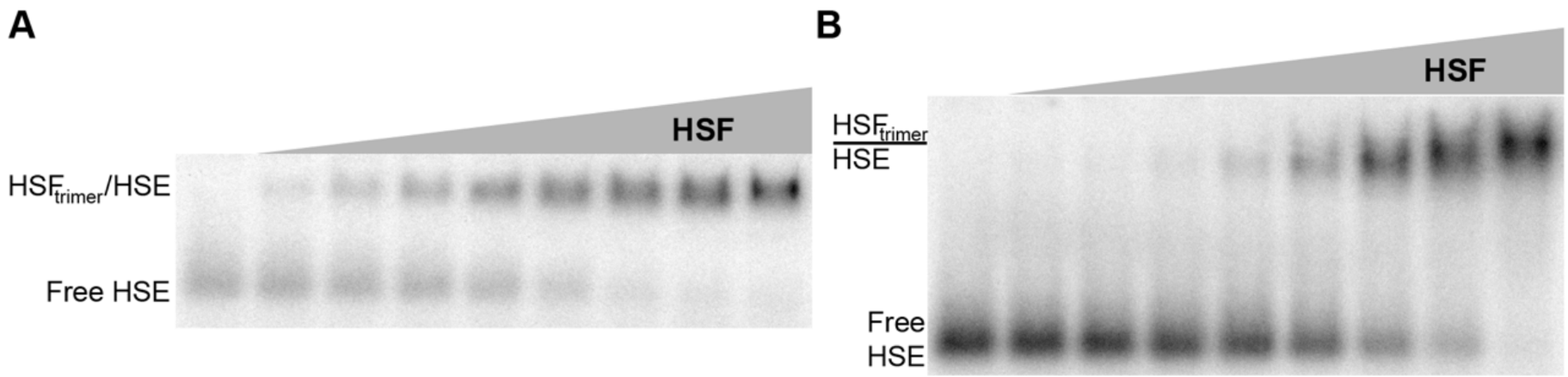


in vitro Binding Assay (PB-seq) reveals all potential binding sites and relative affinities



To transform these relative values into Kd measurements the absolute binding affinities for two genomic HSEs must be measured.

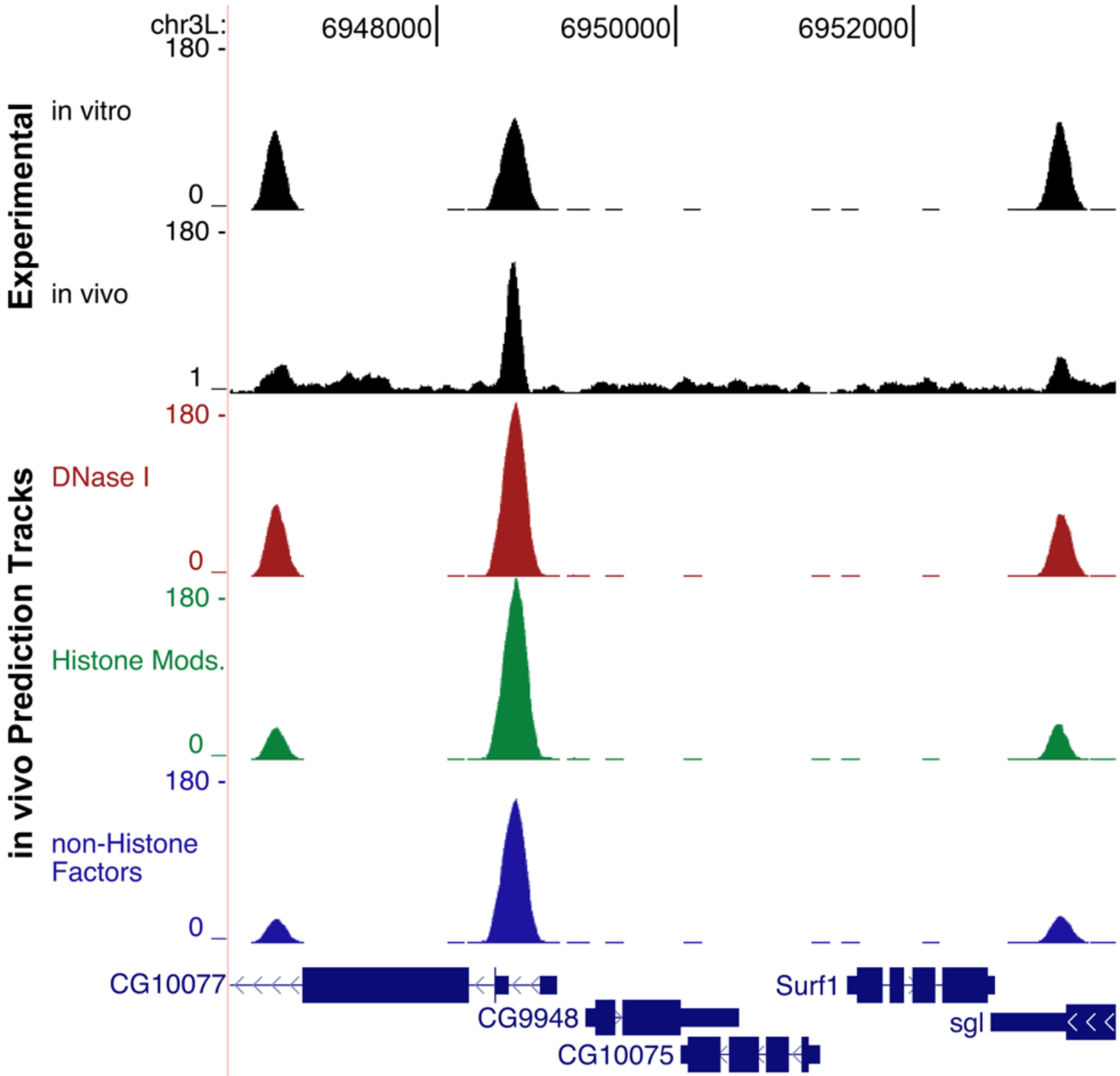
HSF binds to HSEs with picomolar to nanomolar affinity in vitro



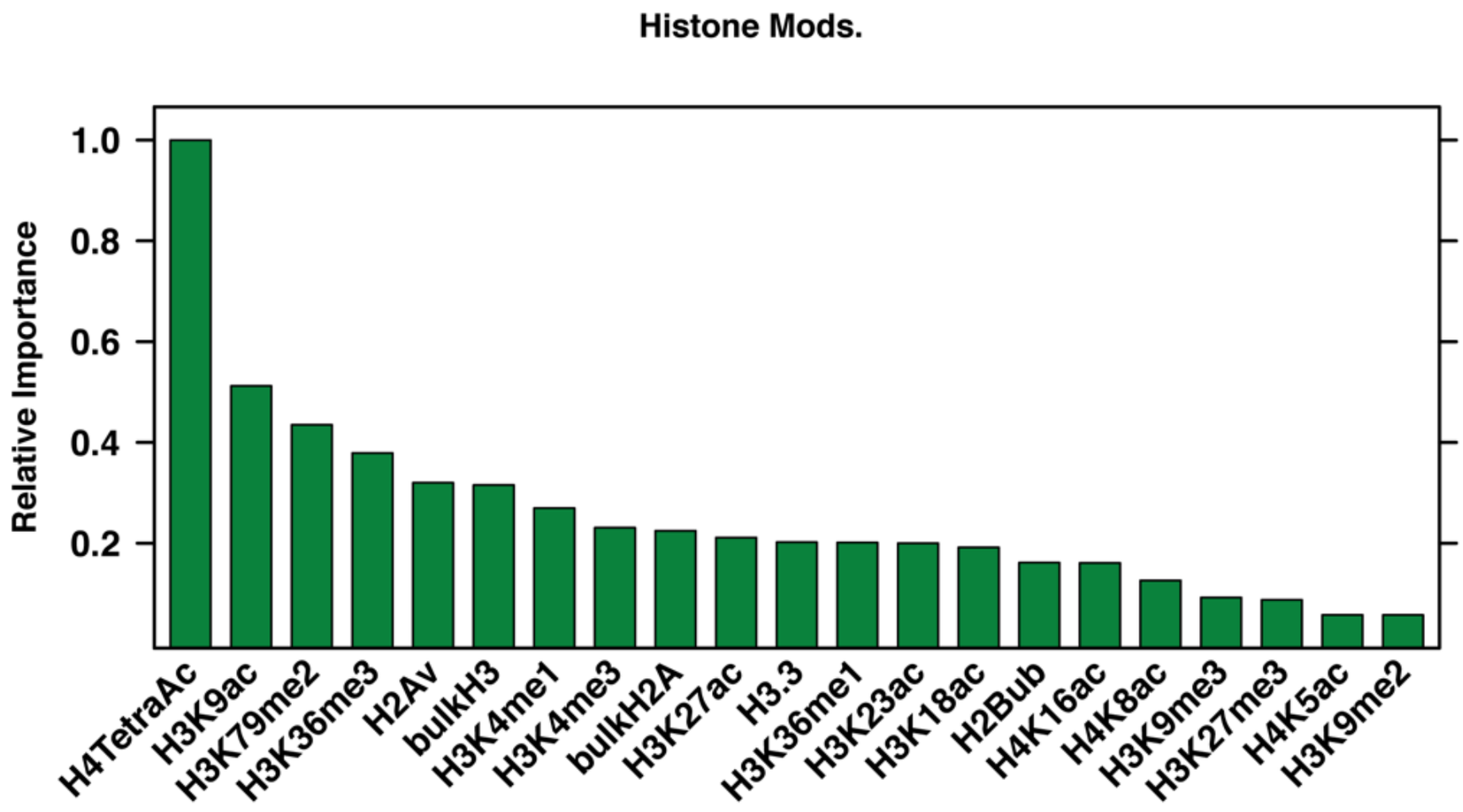
Prediction of Binding Profiles using available PB-seq Data and Genomic Chromatin Data from modENCODE

DNase I	H3K9me2
GAF	H3K36me1
H4K16ac	H3K36me3
H4TetraAc	CP190
MNase	SuHw
Chro(Chriz)	Ez
BEAF	H4K8ac
H3K4me3	H3.3
H3K27ac	H3K23ac
H3K9ac	H3K4me1
H4K5ac	H3K9me3
HP1	CTCF
H3K27me3	bulkH3
H2Bub	bulkH2A
H3K79me2	Pc
H2Av	H3K18ac

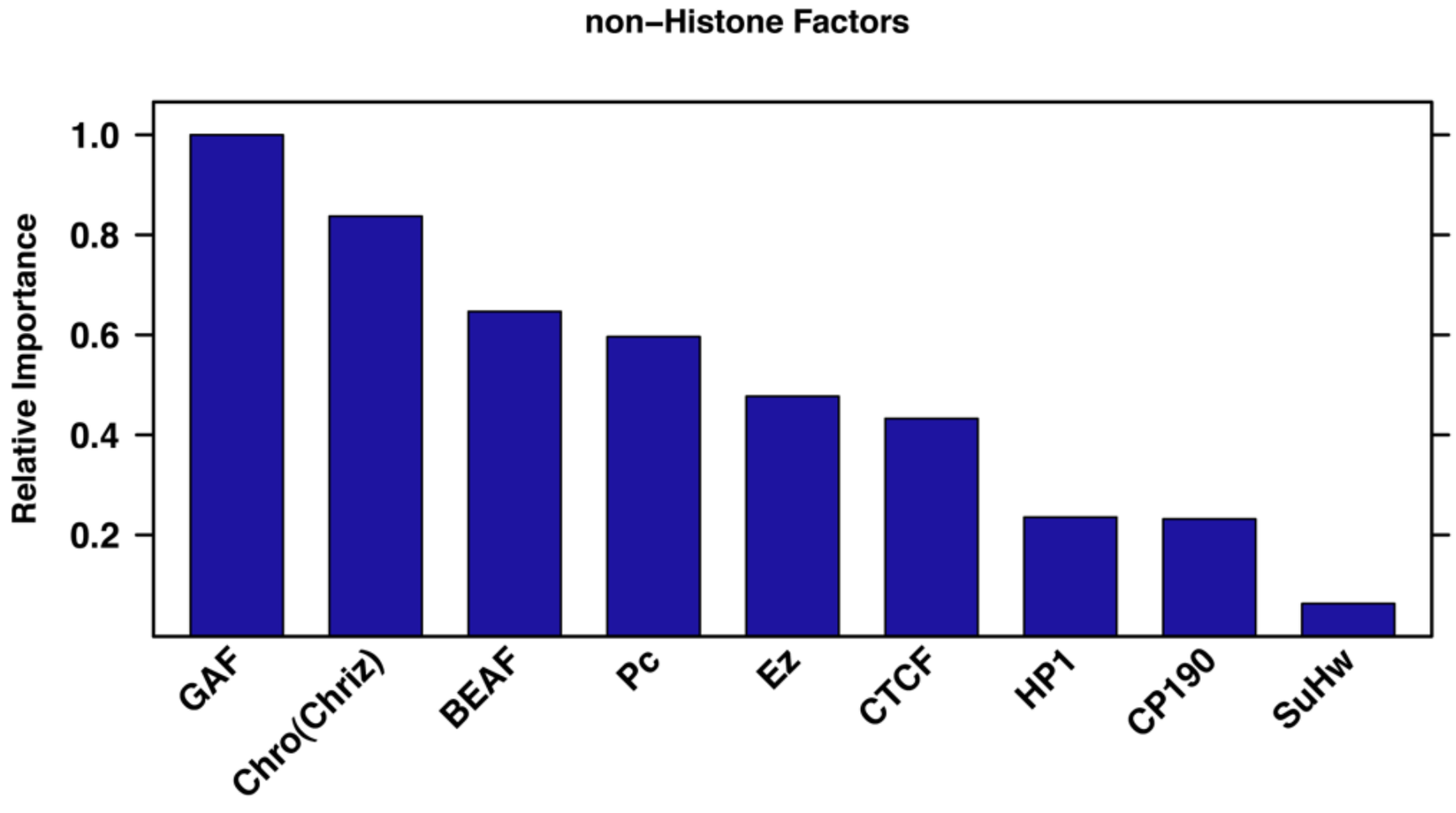
Regression models predict in vivo (ChIP) binding signal using the in vitro (PB-seq) data and NHS chromatin landscape



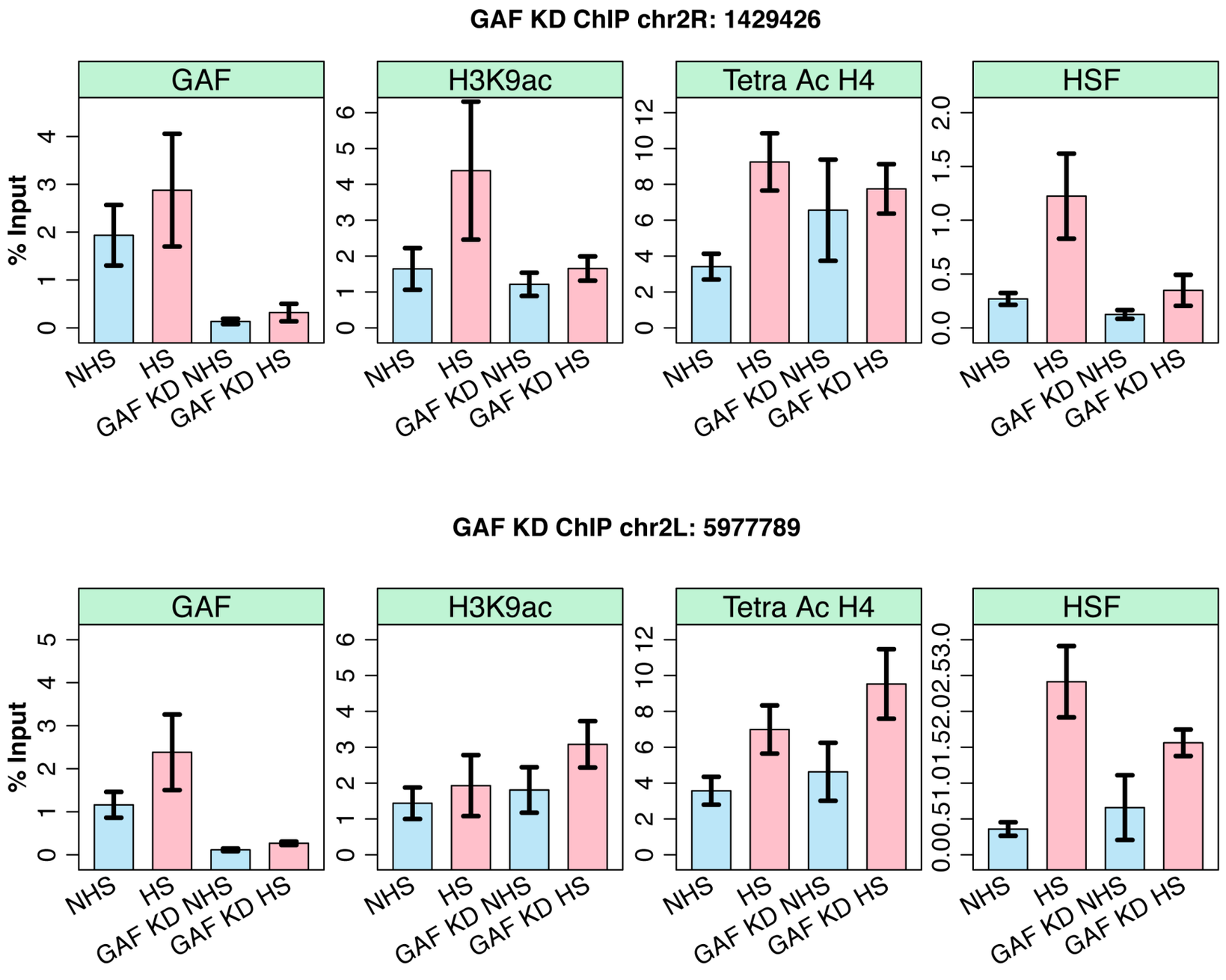
Histone Acetylation is the most influential modification for predicting HSF binding intensity



GAF is the most influential non-histone covariate in the predictive model



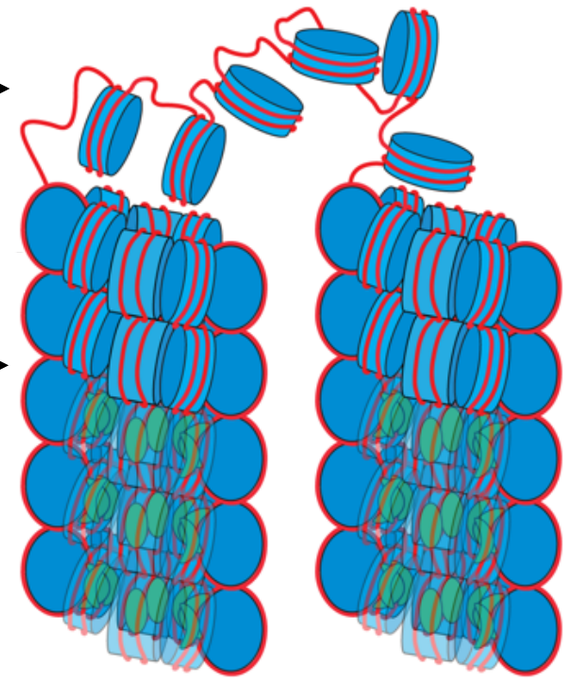
GAF depletion compromises HSF binding intensity



How do accessibility and active marks originate?
working model:

Accessible →

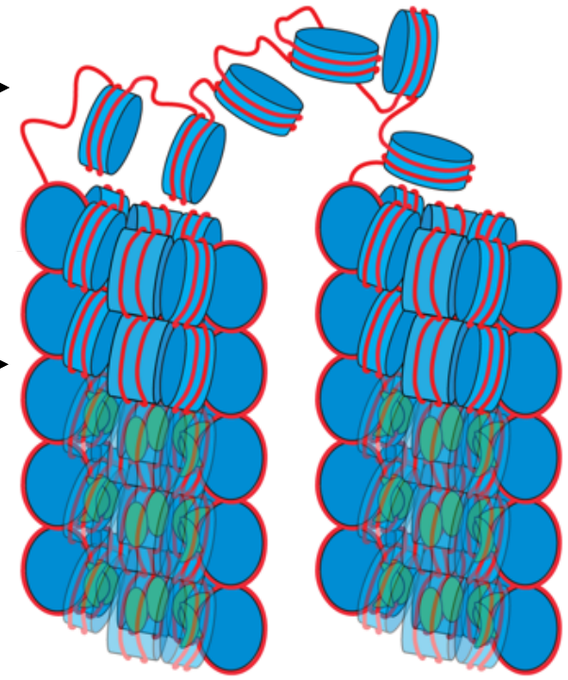
Not accessible →



DNA sequence directs accessibility

Accessible →

Not accessible →

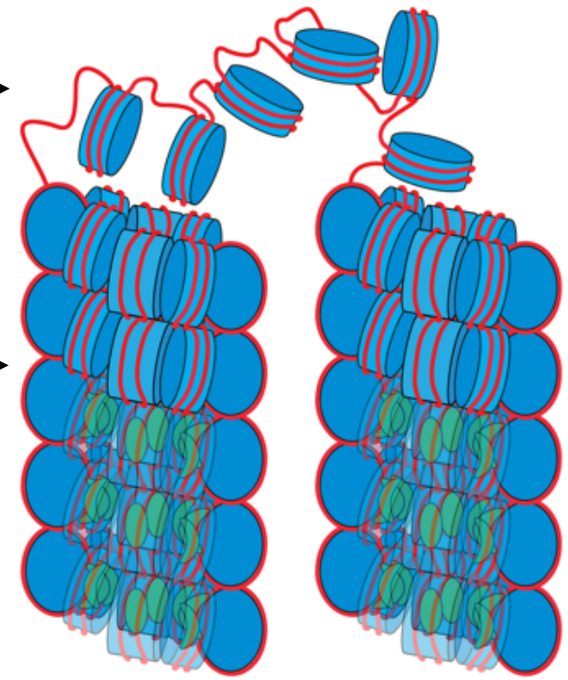


DNA Sequence (poly-AT tracts)
keep yeast promoters
free of nucleosomes
(Struhl & Rando labs)

DNA sequence directs accessibility

Accessible ⟶

Not accessible ⟶



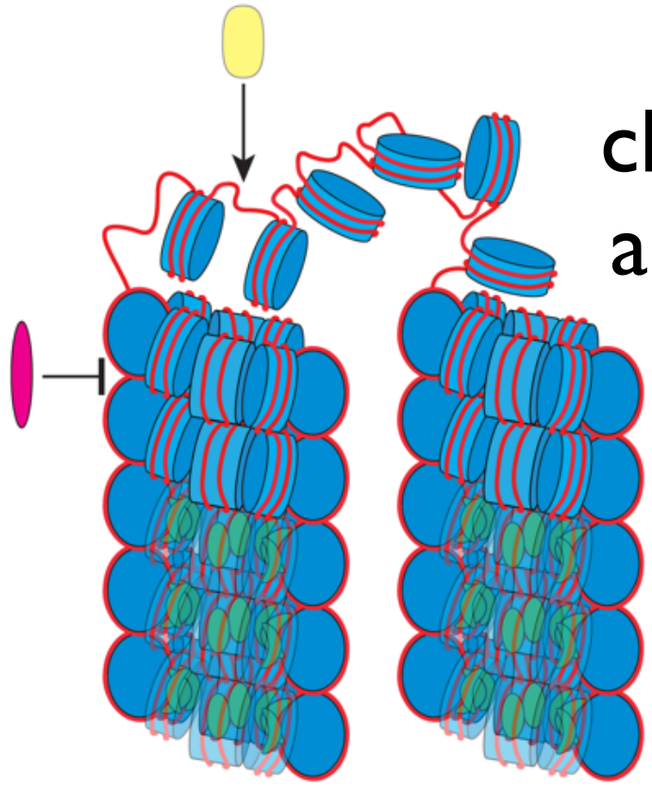
CpG islands favor nucleosome formation in vitro, but are depleted in vivo.

H1 preferentially binds AT-rich linker DNA.

(Reviewed in Zlatanova and Yaneva, DNA cell Biol. 1991)

Hypothesis: CpG islands are inherently refractory to higher order compaction by H1, which maintains the chromatin is a transiently accessible state.

Factors target linker DNA between nucleosomes



TFs target uncompacted chromatin and CpG islands are highly occupied in vivo.

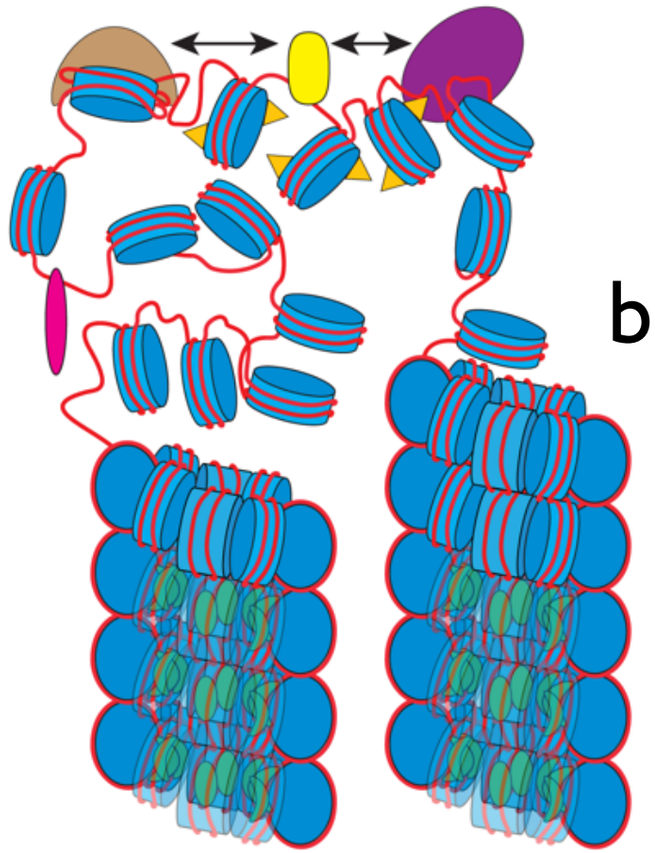
The Set1 H3K4 methylation complex and a H3K36 demethylase has been shown to interact with unmethylated CpG-rich DNA in vitro.

(Ooi et al. Nature 2007, Zhou et al. Mol Cell Biol 2012)

Mammalian sequence-specific TFs, as a class, have a GC-bias in their cognate binding sites.

(Deaton et al. Genes Dev 2011)

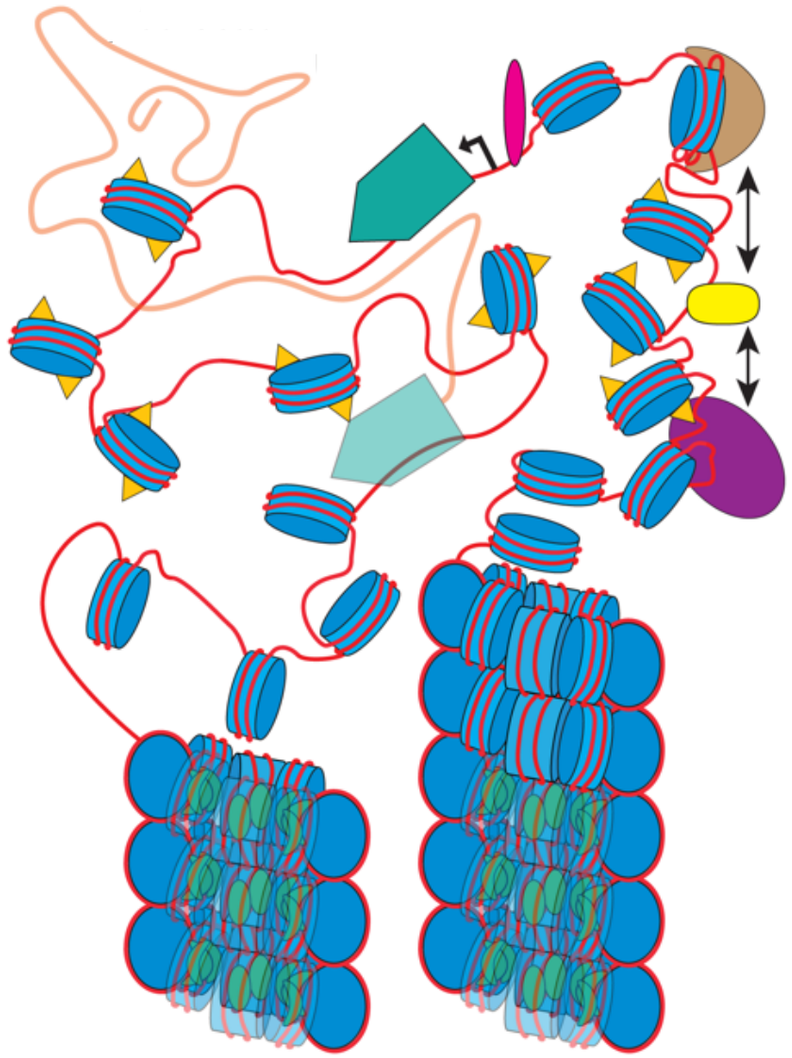
Binding expands accessible regions



TF binding increases and expands the boundaries of accessibility and active marks.

Binding expands accessible regions

Then the pink TF can bind and active **transcription**, which ultimately controls cell fate.



Is there a paradigm shift in biology, away from overly hypothesis-driven research?

- Starting today, here is how I would approach my PhD:
  - Identify a relevant question.
  - Design experiments that are as unbiased (and controlled!) as possible to address this question.
  - Analyze data and look for correlations.
  - Formulate hypotheses from the correlations.
  - Test the hypotheses.
- With an open mind and competently designed experiments/analyses, one can develop hypotheses that were inconceivable at the onset.

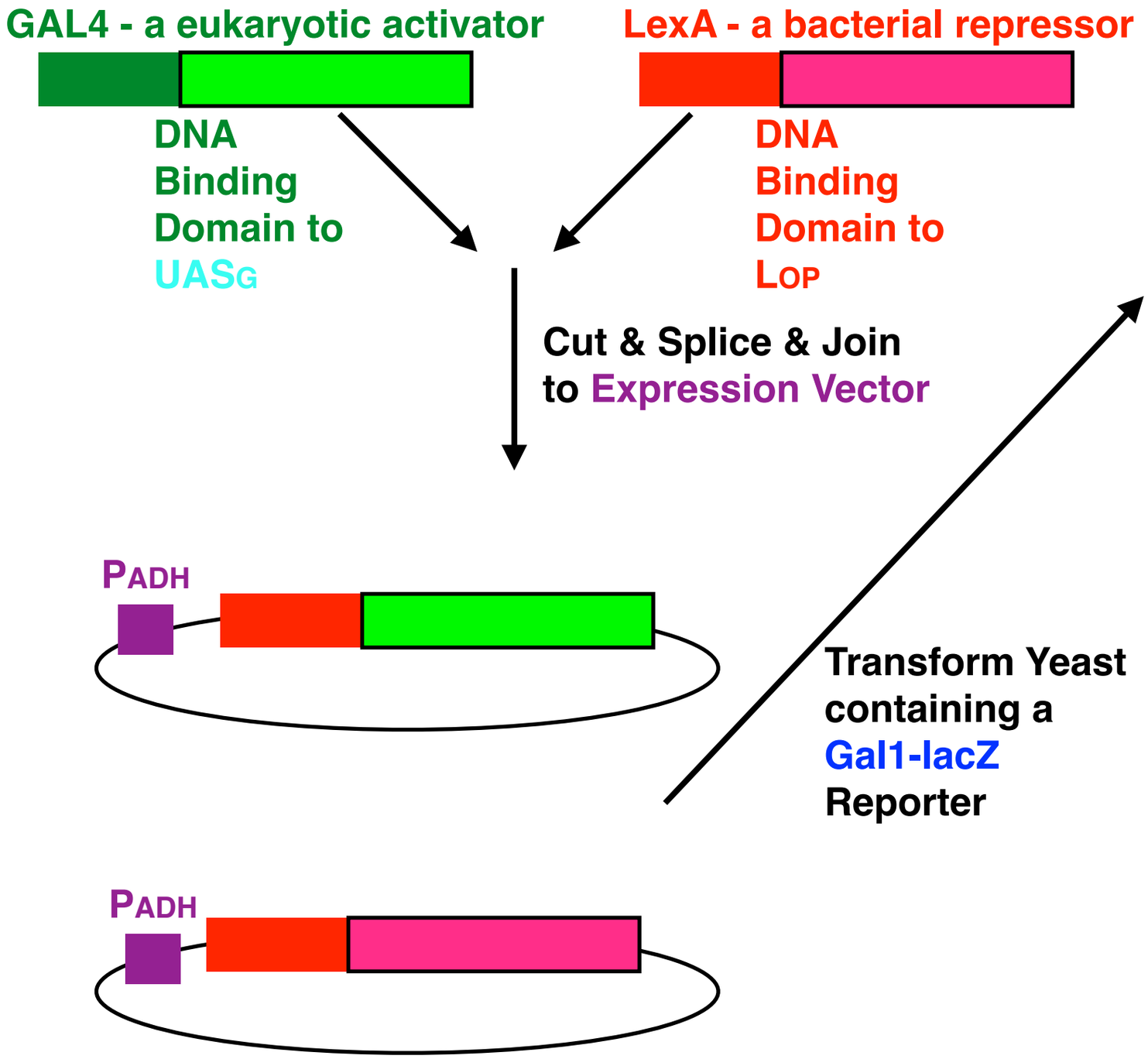
Summary: Part II

- Features of double-stranded DNA sequence can provide recognition features for proteins.
- DNA-binding transcription factors (TFs) represent a fairly large fraction of the proteome.
- TFs have domains that bind specific DNA elements and fall into distinct classes.
- These domains employ a variety of strategies to build a molecular protein complement to the DNA element.
- The repertoire of target DNA sequences that can be specifically recognize by these proteins is further enriched by heterodimerization and cooperative interactions.
- Chromatin state dictates TF binding, which can in turn influence chromatin structure

My favorite Transcription Factor:
Drosophila HSF

DNA binding **Trimerization** **Activation**

DNA binding and activation domains of transcription factors are distinct and separable



Remove and Substitute the UASG

Vector Expression Units of β-Gal

	LexA	LexA- GAL4
ΔUASG	0	0
LOP	<1	520

Brent and Ptashne, Cell, 1985

Transcription Activation Activity can reside in one or more regions

Generate deletions in vitro

Clone into Expression Vector

Transform gal4⁻ strain w/ GAL1-lacZ reporter

Construct	Units of β-Gal
GAL4 WT	1900
GAL4 1-147	<1
GAL4 1-238	110
1-147/768-881	110
1-238/768-881	1400

Ma and Ptashne, Cell, 1987

Transcription activation regions occur frequently and are often acidic

Library of plasmids containing random bits of E. Coli DNA

Transform gal4- strain w/ a GAL1-lacZ reporter

Select Blue colonies: ~1%

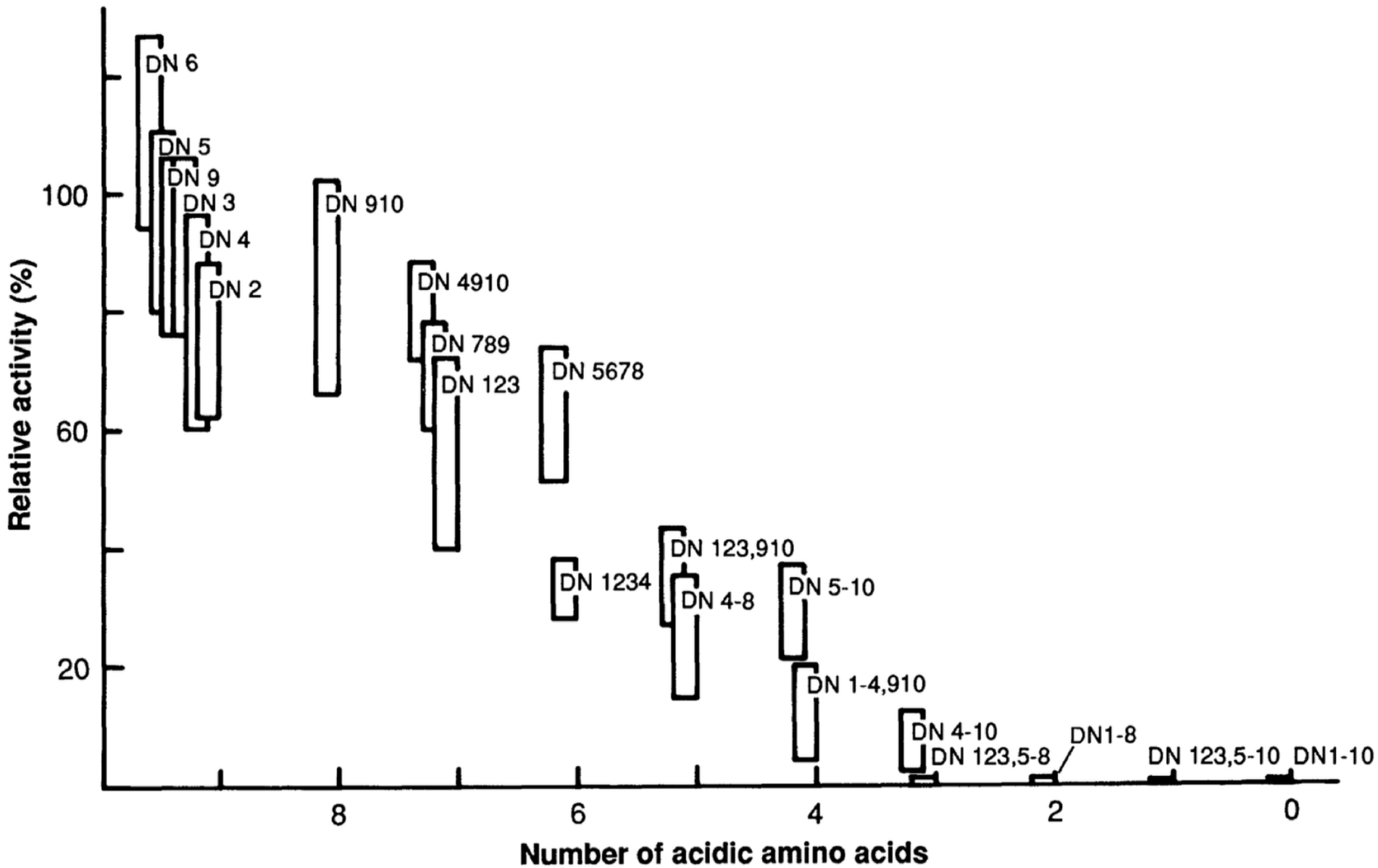
Sequence

	Units of β-Gal
	<1
	10-80

Random E. Coli DNA encoded 12-81 amino acids and their common feature was net charge of minus 1-10

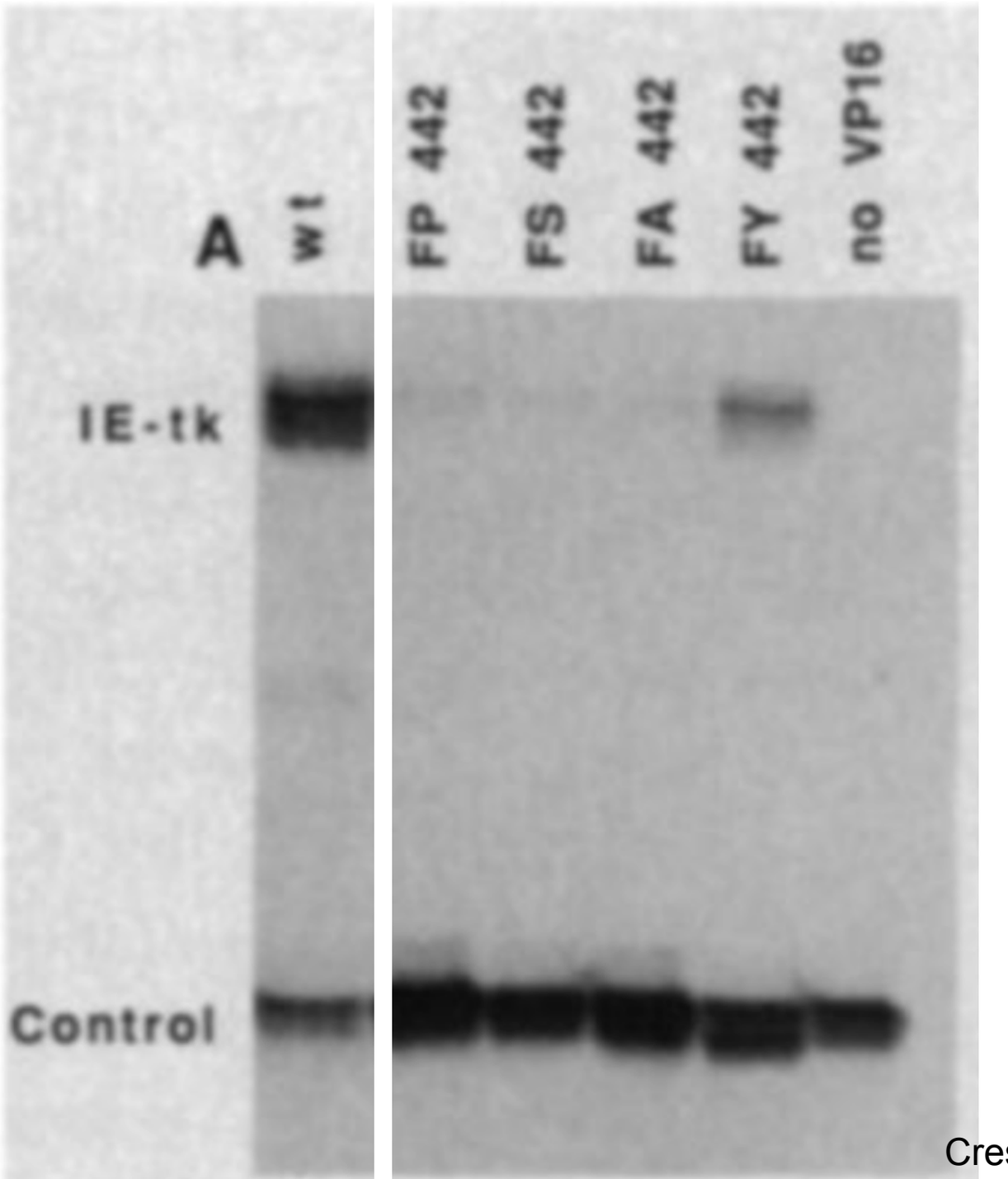
Ma and Ptashne, Cell, 1987 (not the same paper as the previous slide)

Negative charge correlates with transcriptional activity of VP16



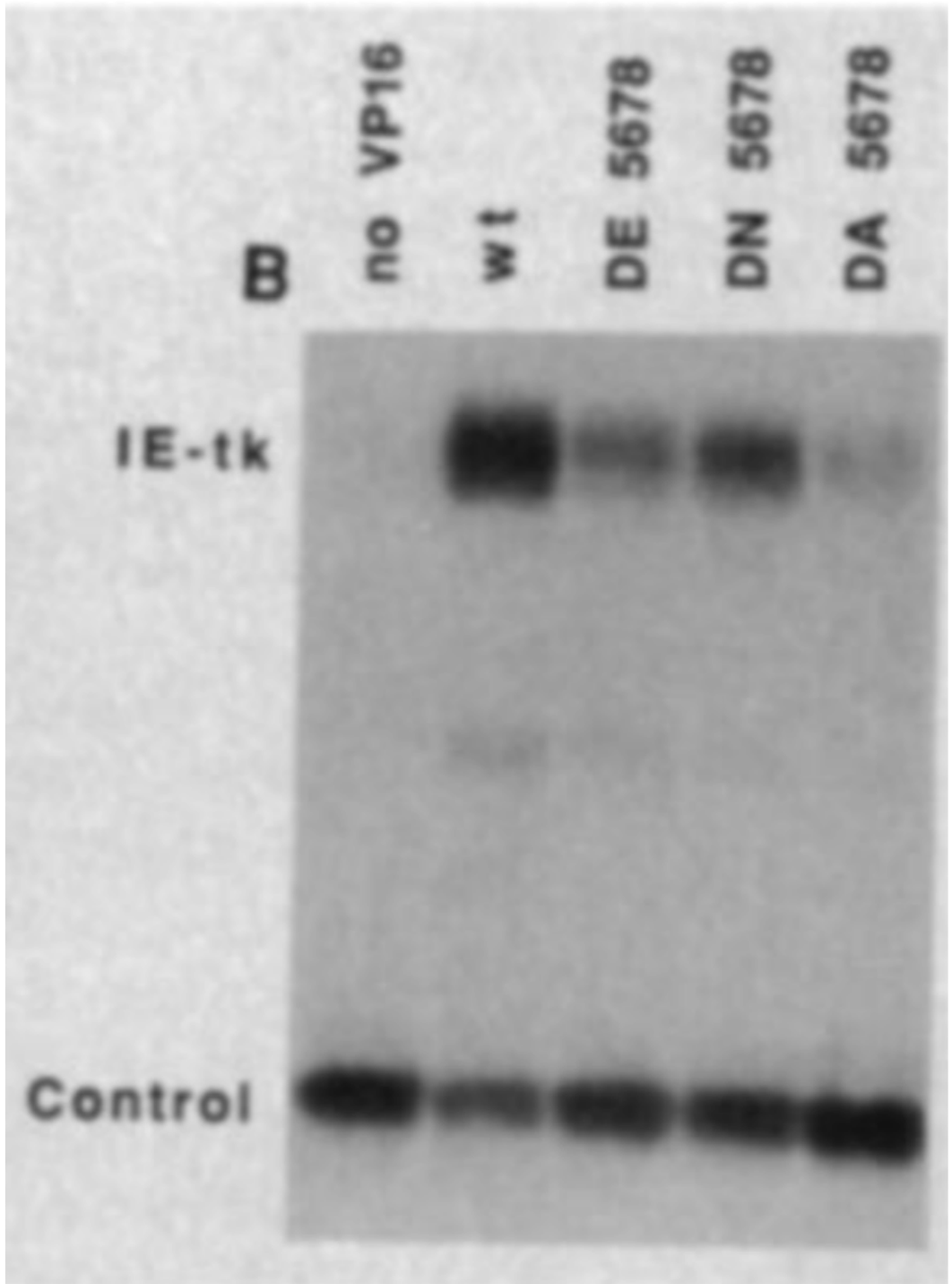
Cress & Trezenberg, Science 1991

A single Phenylalanine is critical to VP16 function



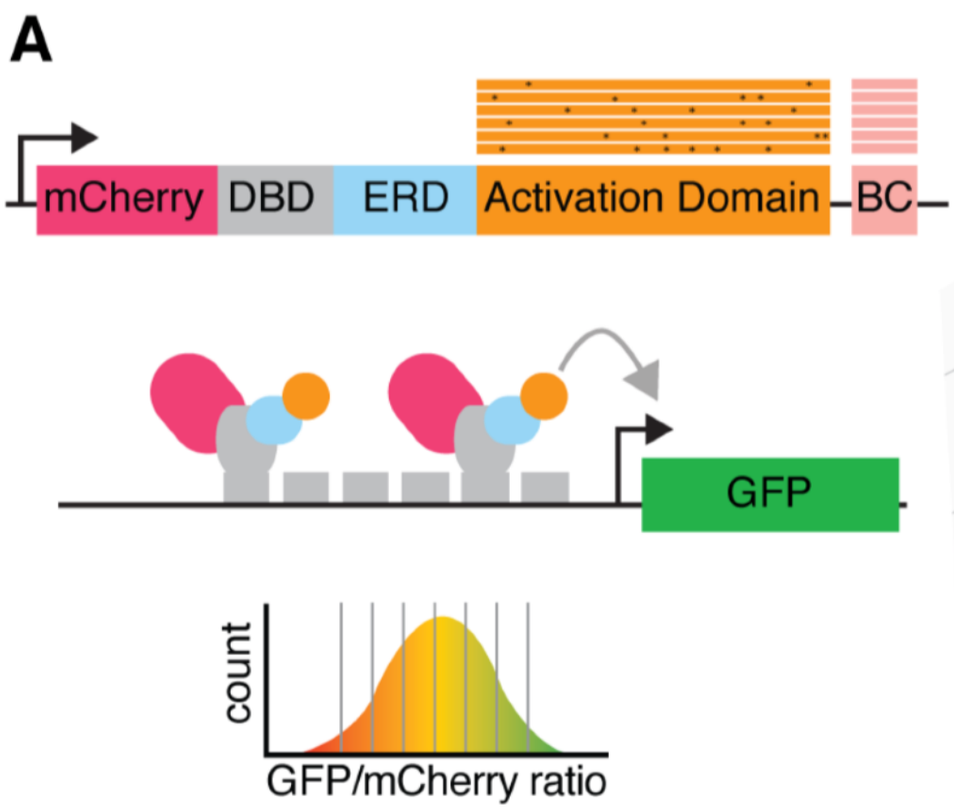
Cress & Trezenberg, Science 1991

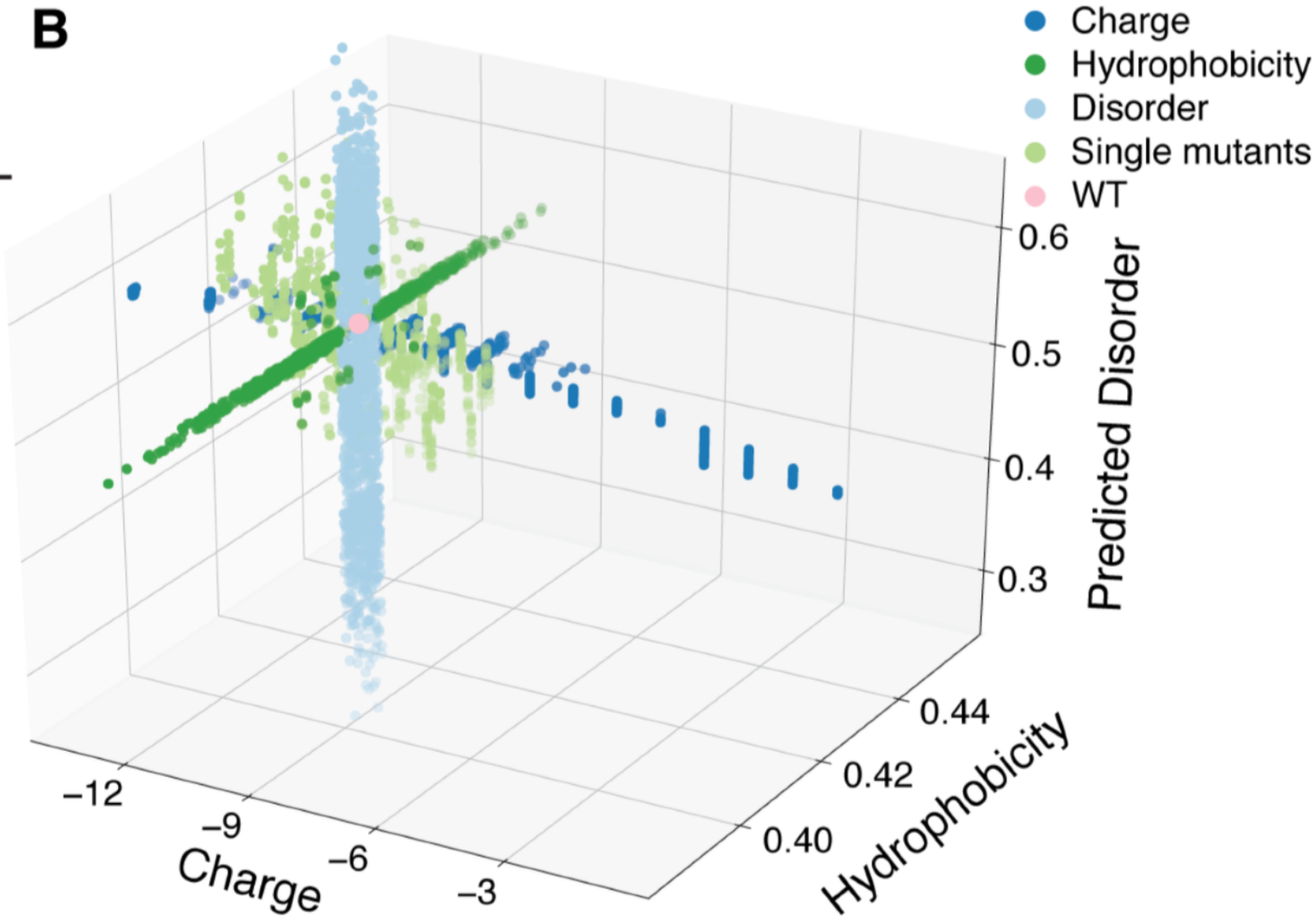
The environment surrounding Phe^{442} affect VP16 function



Cress & Trezenberg, Science 1991

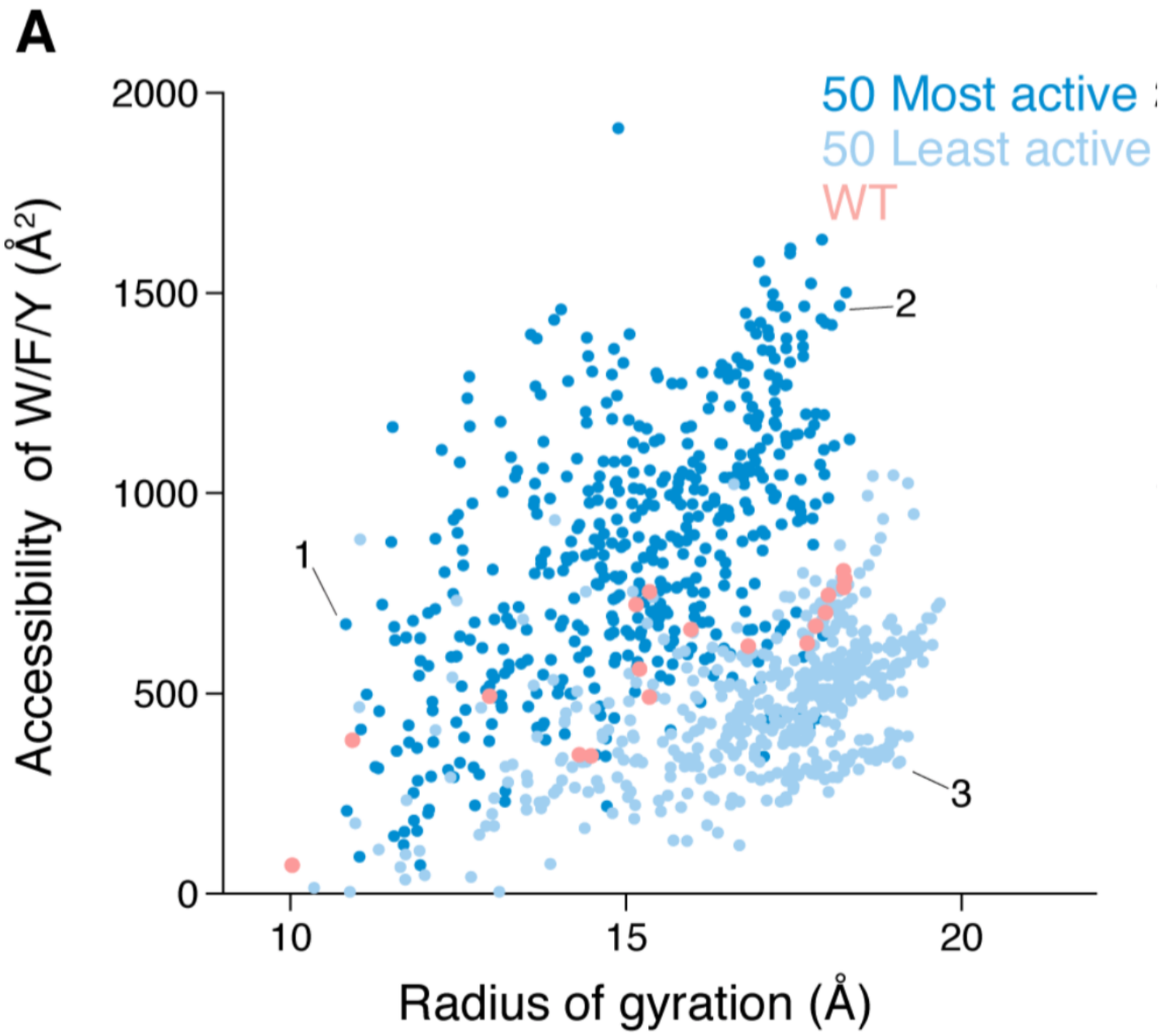
Measuring thousands of designed GCN4 activation domain mutants in parallel





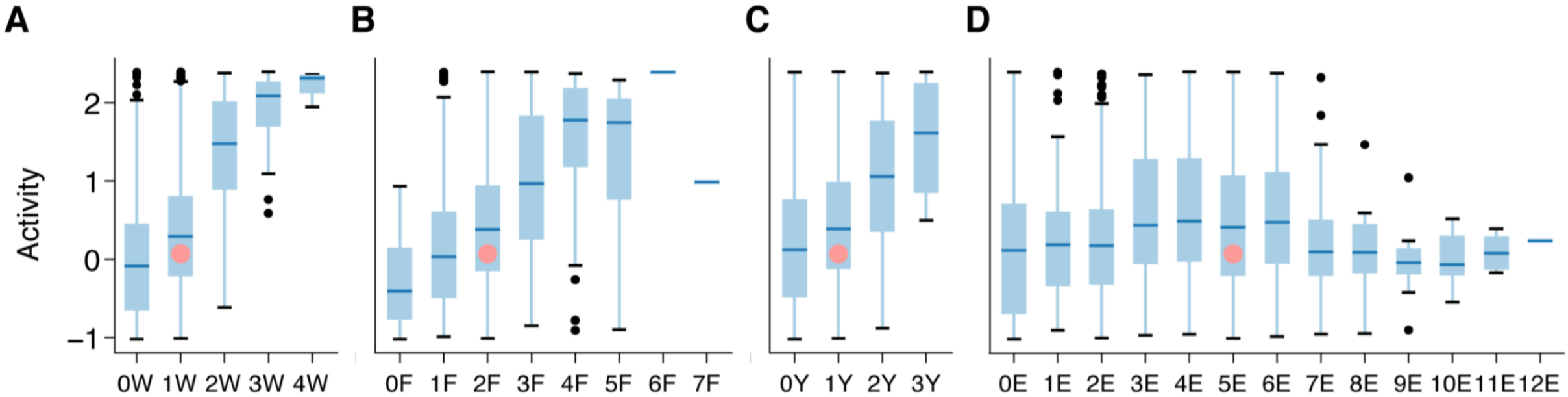
Staller, et. al., Cell Systems, 2018

Simulations reveal that highly active variants keep aromatic residues exposed to solvent



Staller, et. al., Cell Systems, 2018

Aromatic residues control Gcn4 Activity



Conclusions: acidic residues regions keep two hydrophobic motifs exposed to solvent to mediate activity.

Staller, et. al., Cell Systems, 2018

Activation domains come in several flavors

- SP1 - Q-rich domain (polar)
- CTF - P-rich activation domain (hydrophobic)
- NTF1 - I-rich activation domain (hydrophobic)

Summary: Part III

- DNA-binding transcription factors (TFs) often have distinct and separable domains (DBD and activation)
- Hydrophobic and Acidic residues are often critical for TF activation function, perhaps acidic residues keep hydrophobic solvent-exposed.
- There are many types of activation domains

Part IV

- Most TF binding events do not result in changes in gene expression.

- As a corollary, just because something exists does not mean it is functional.
- Too often people ask the question "what is the function of X", when there is no evidence that X is functional.
- Many lncRNAs serve as contemporary examples.

How can we identify which cofactors interact with your favorite activation domain?

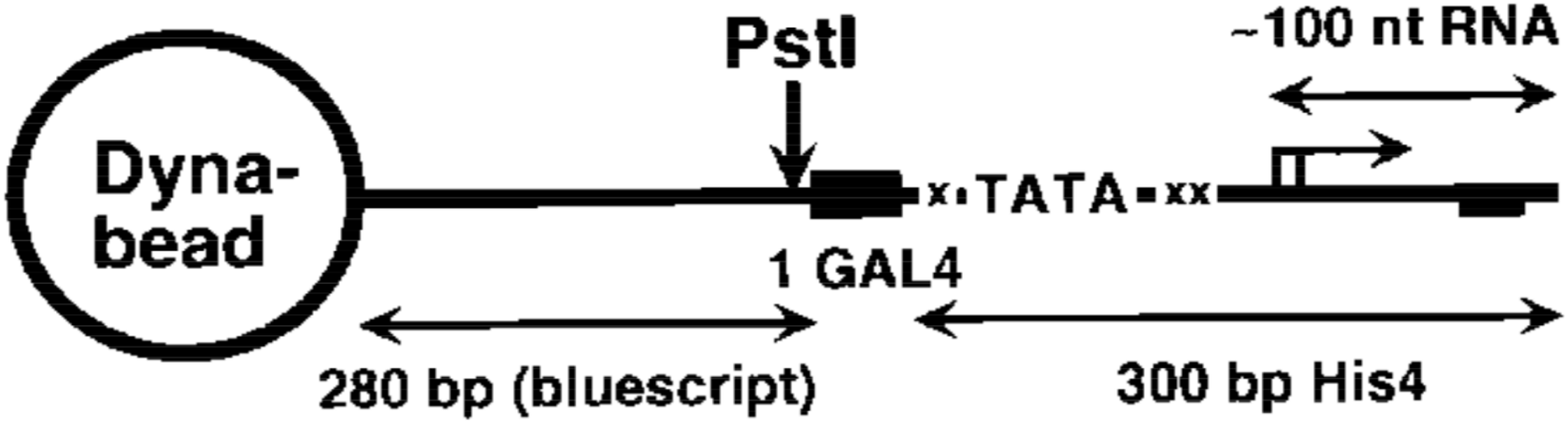
- Conventional chromatography of Mediator: Näär, et. al., Science, 1999; follow up work identified the KIX domain of Med15: Yang, et. al., 2006.

- immobilized template and label transfer: Fishburn, et. al., Molecular Cell 2005.

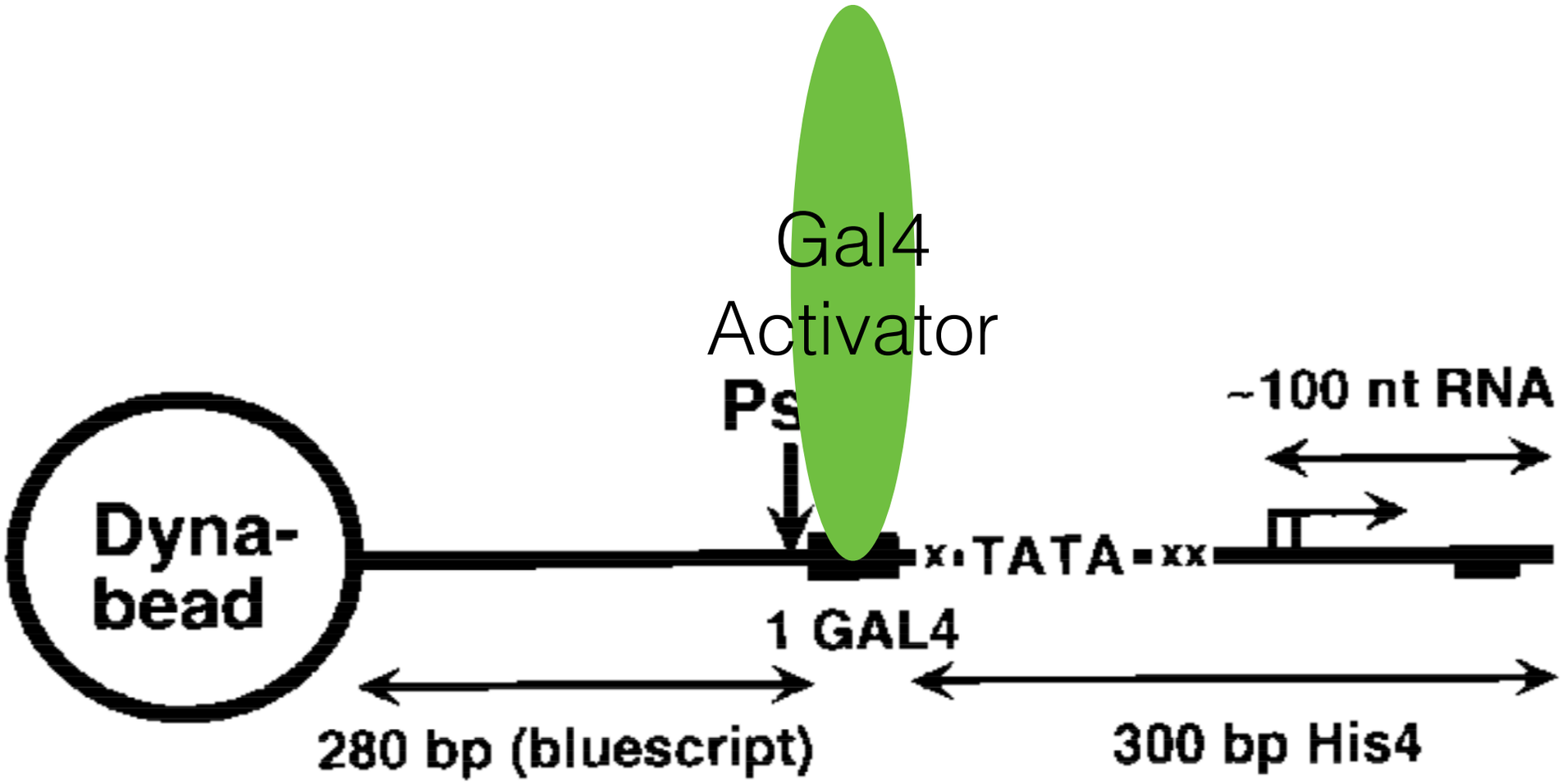
- Unbiased approach: BioID or APEX tagging of TFs —Roux, et. al., J Cell Biol. 2012; Lam, et. al., Nature Methods 2014 (there are newer versions of each)

Immobilized template and label transfer to determine activator targets

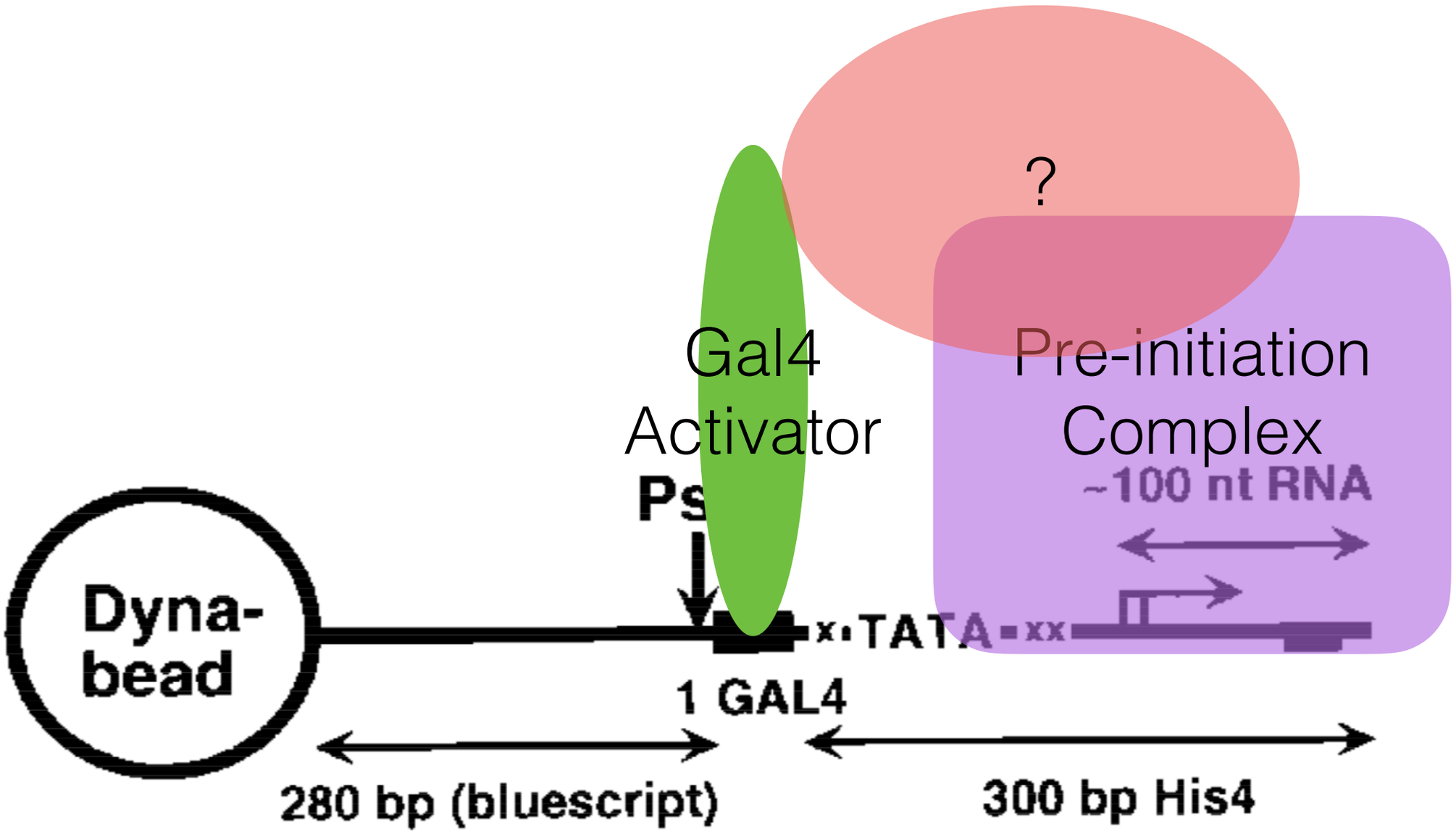
Fishburn, Mohibullah, and Hahn, 2005



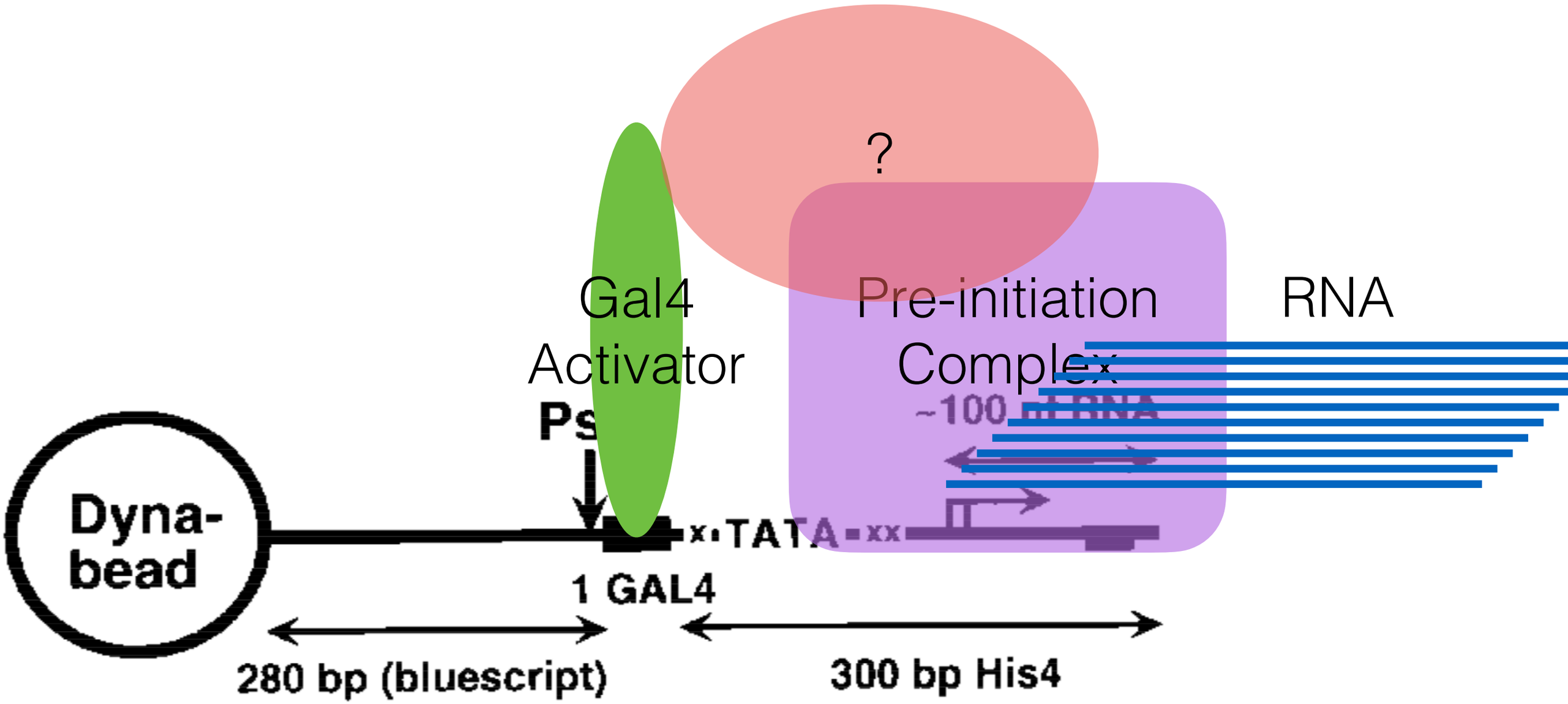
Add a recombinant Gal4 transcription activator



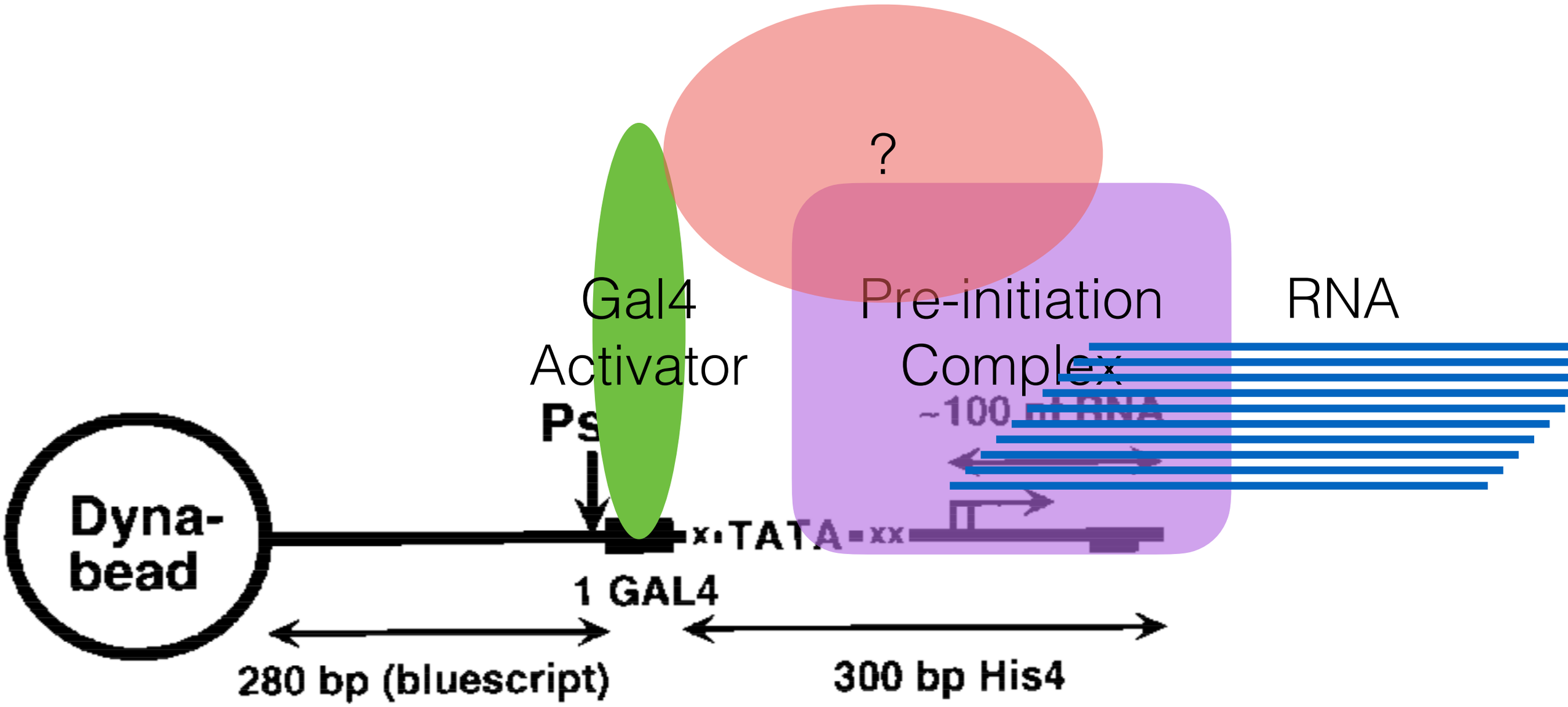
Incubate with yeast nuclear extract to establish transcription-competent complexes



Add NTPs to stimulate transcription

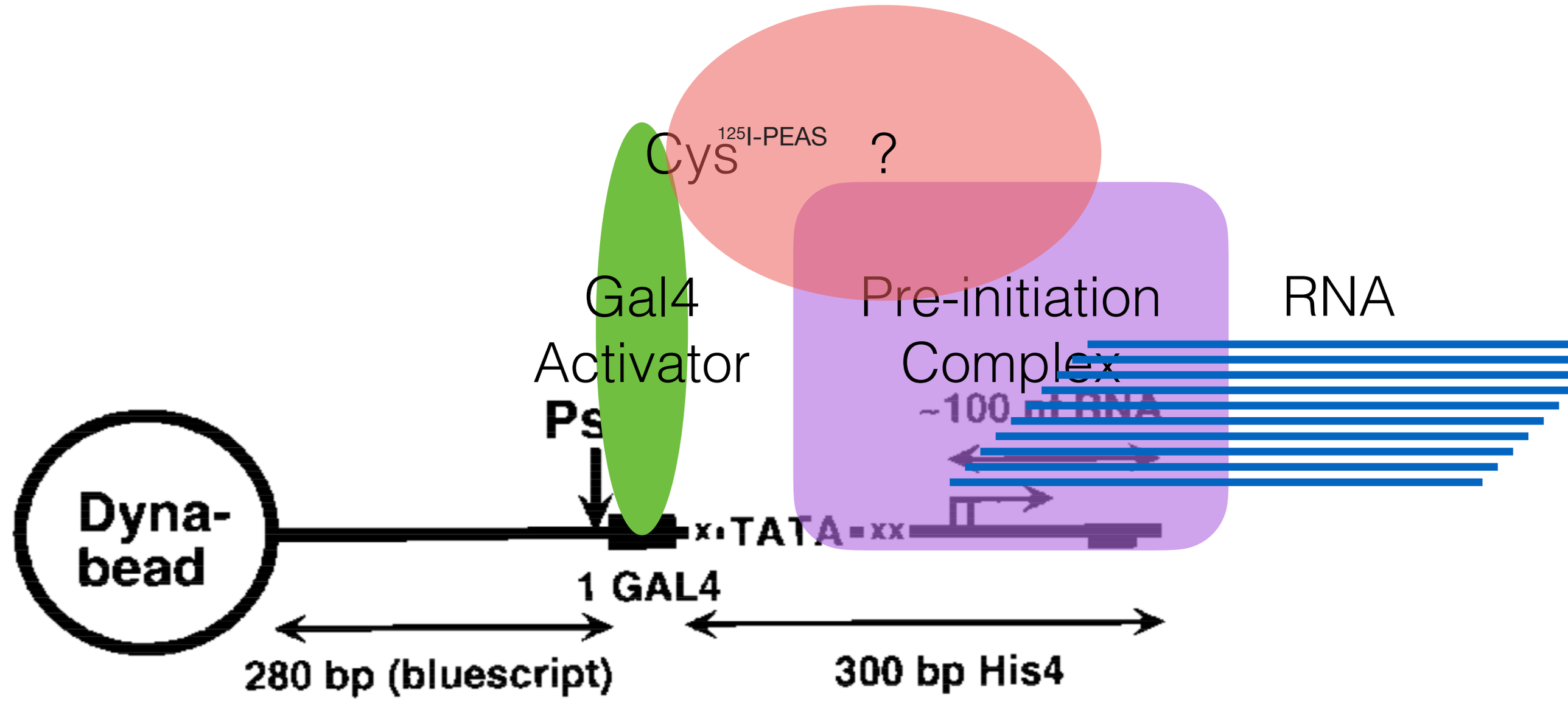


Does Gal4 interact with unknown cofactors (?) from the extract?

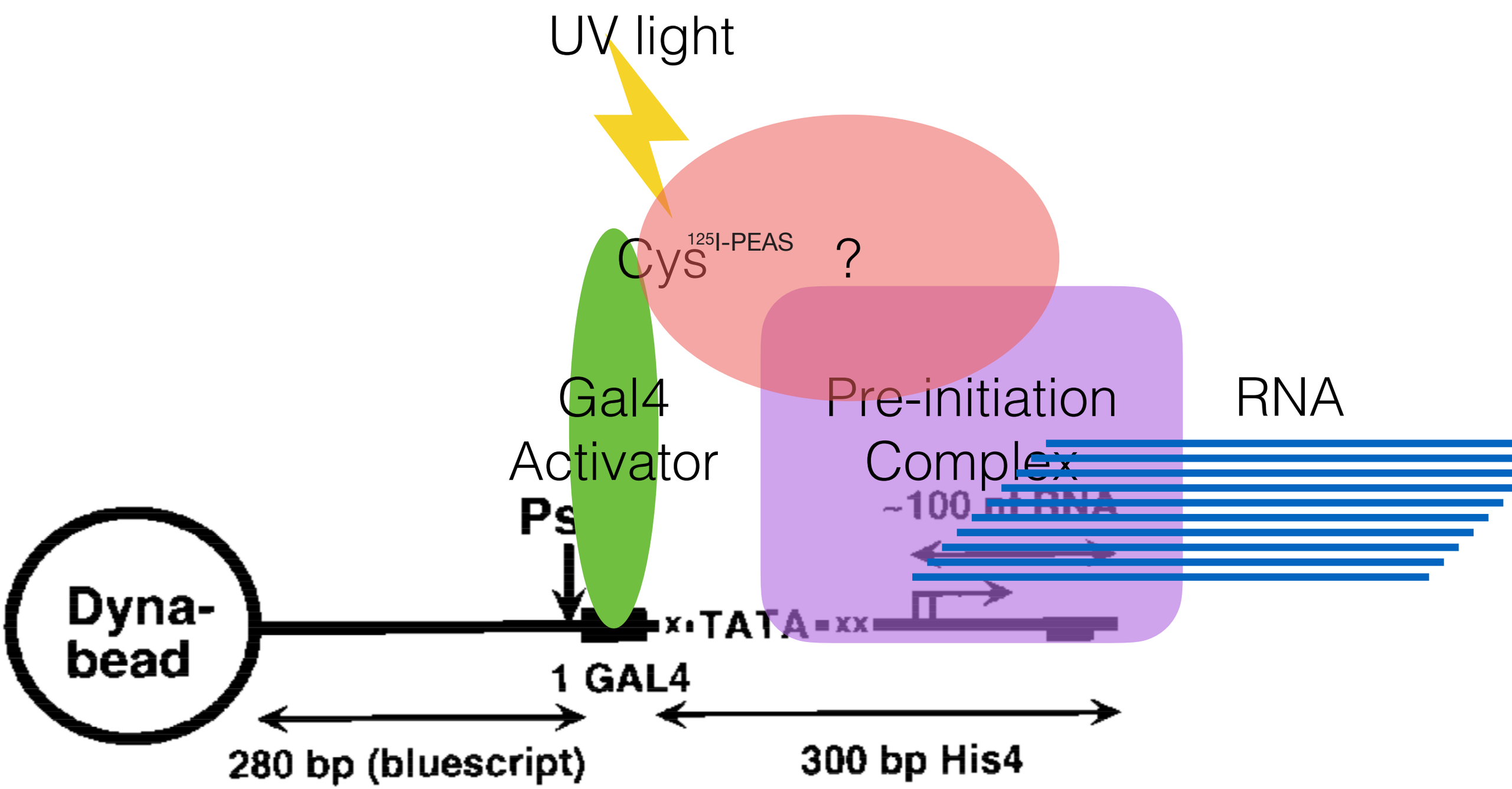


Make modified Gal4 with Cysteine substitutions and a UV-reactive radiolabeled l-PEAS

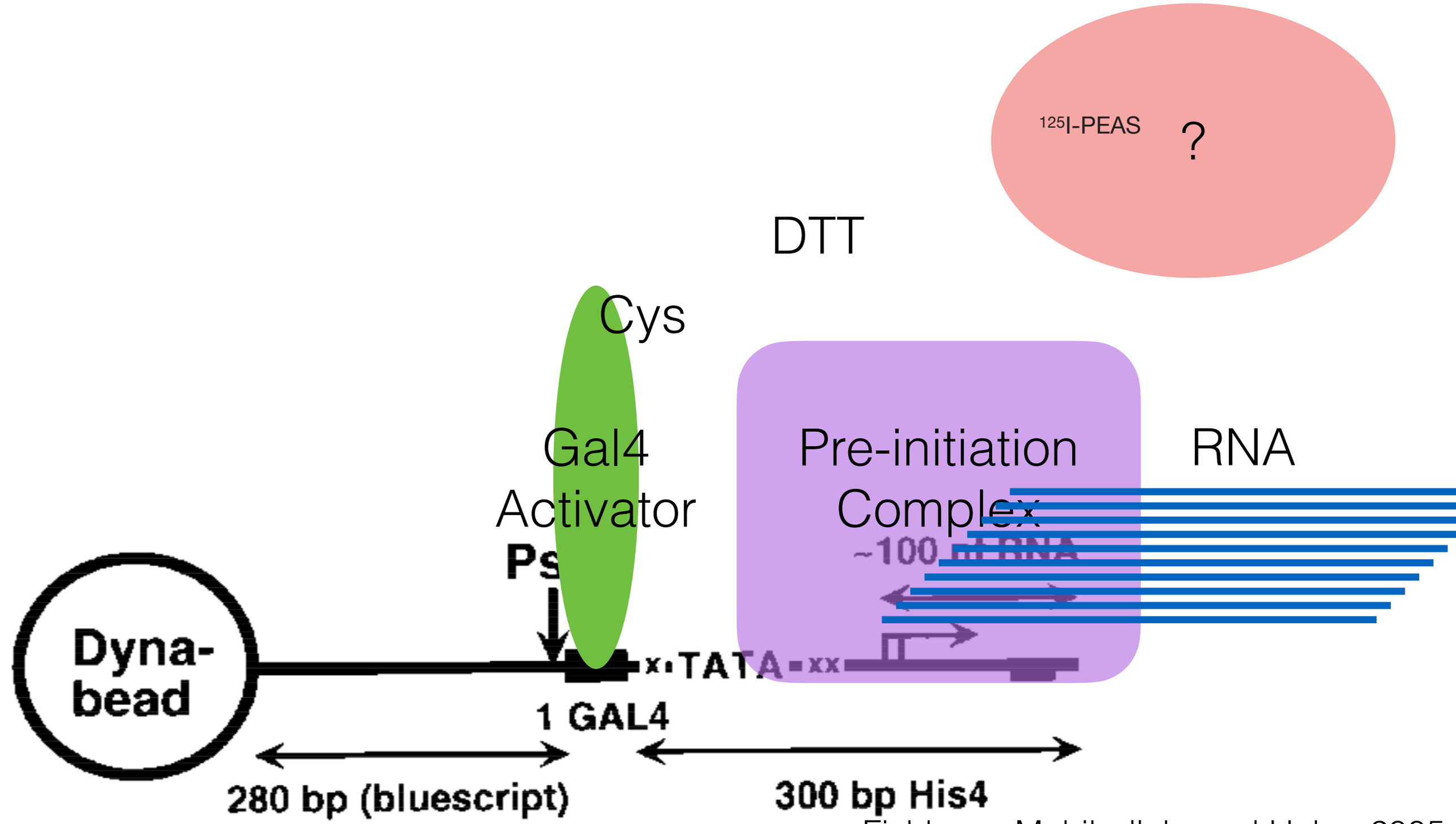
(Confirm that modified Gal4 complexes are transcription-competent)



UV light to crosslink

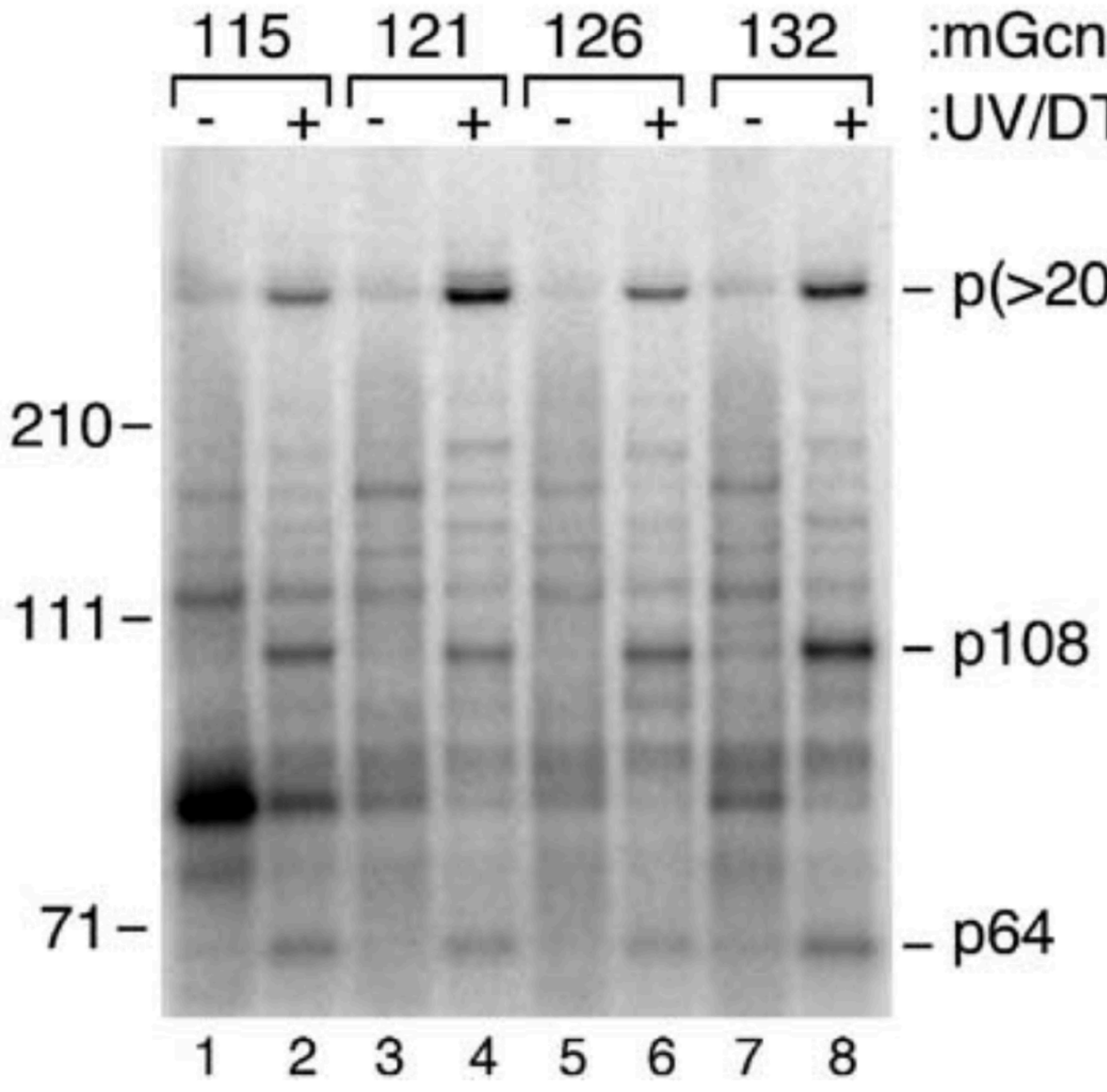


Liberate and identify labeled proteins on a gel



Fishburn, Mohibullah, and Hahn, 2005

Liberate and identify labeled proteins on a gel



Tra1 — Subunit of SAGA and NuA4 histone acetyltransferase complexes

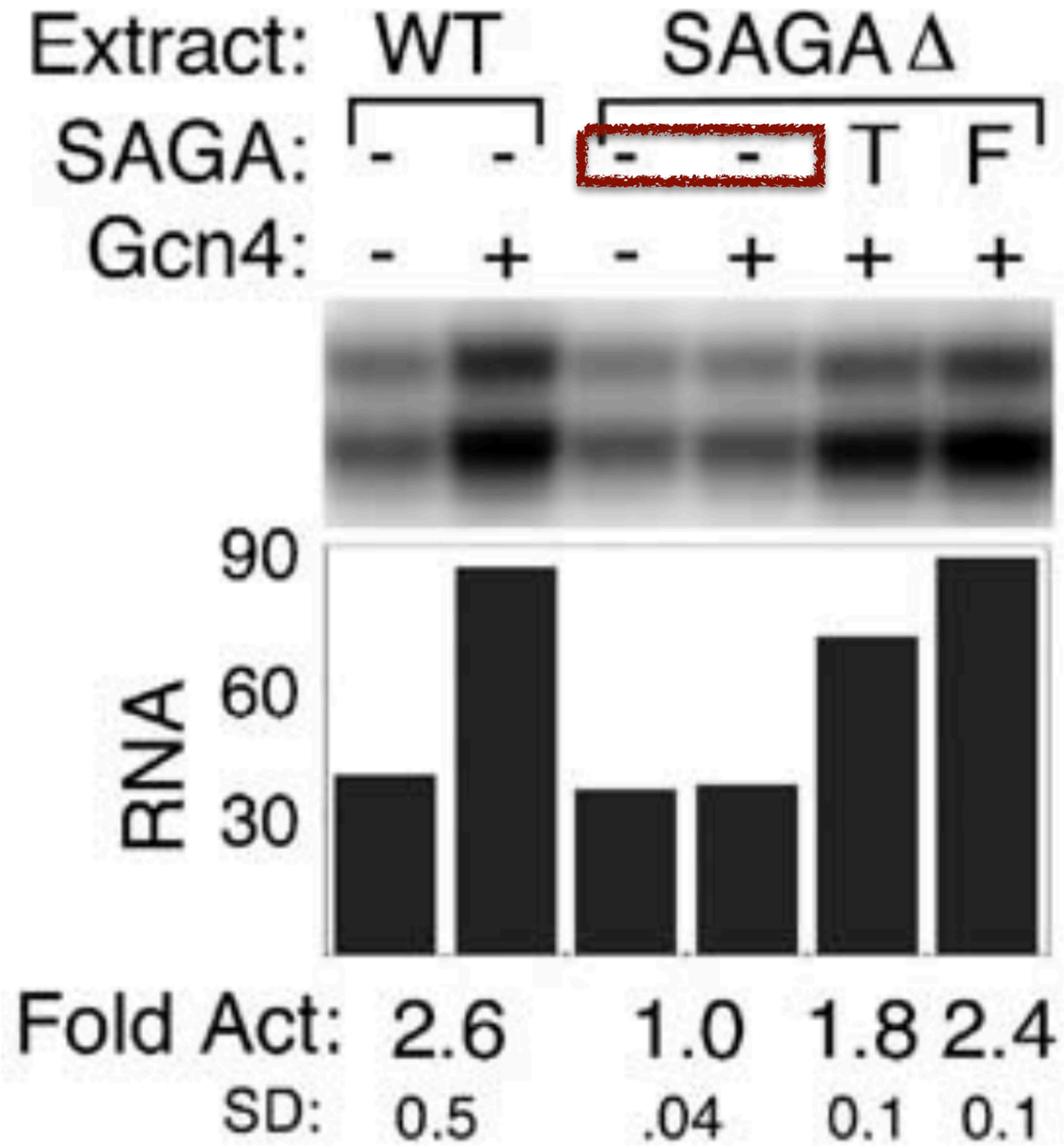
Gal11 — Subunit of the RNA polymerase II mediator complex; facilitates polymerase recruitment by interfacing with activators

Taf12 — Subunit of the TFIID, which is a component of the pre-initiation complex

Fishburn, Mohibullah, and Hahn, 2005

The authors performed functional follow up to confirm that these interactions mediate activation

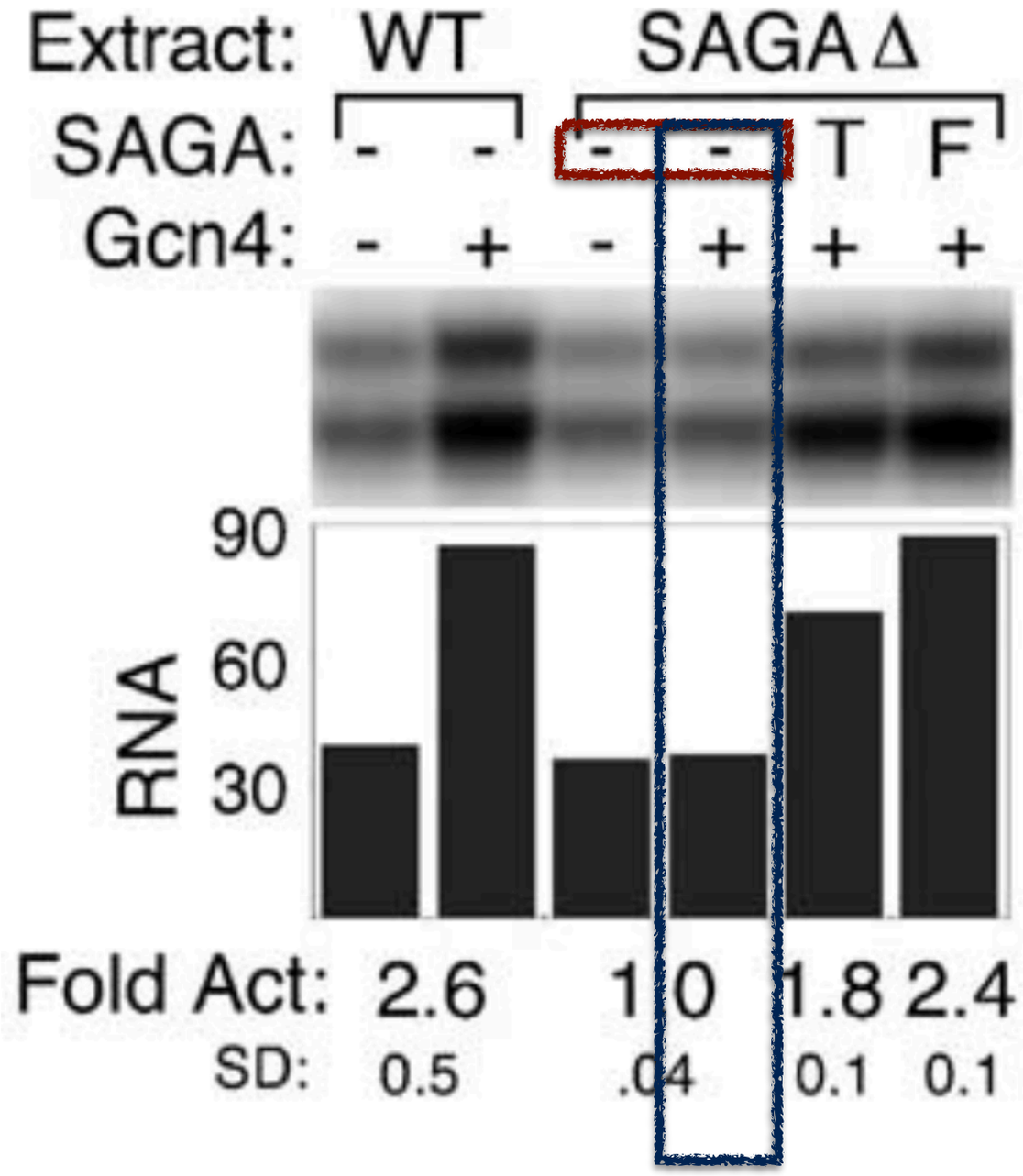
Deplete SAGA complex from NE



Fishburn, Mohibullah, and Hahn, 2005

The authors performed functional follow up to confirm that these interactions mediate activation

Deplete SAGA complex from NE

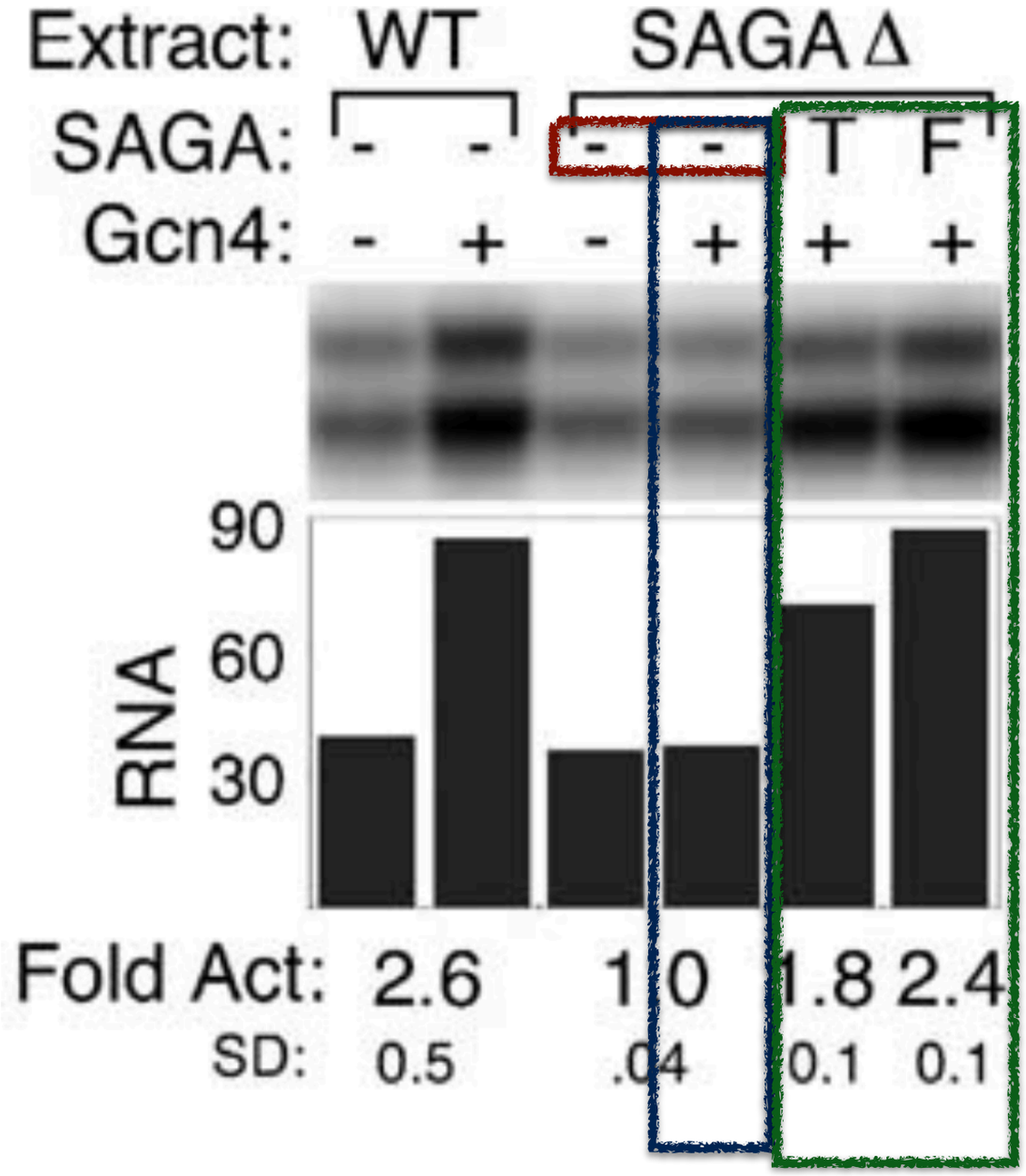


Activation is abrogated

Fishburn, Mohibullah, and Hahn, 2005

The authors performed functional follow up to confirm that these interactions mediate activation

Deplete SAGA complex from NE



Activation is abrogated

Add back SAGA to restore activation

Fishburn, Mohibullah, and Hahn, 2005

Elegant activator/cofactor interaction studies

- c-Myc recruits P-TEFb (Eberhardy and Farnham, 2002; Rahl, et. al, 2010)

- SREBP1 recruits the acetyltransferases CBP/p300 and Mediator through interactions with the cofactor's KIX domains (Yang, et. al, 2006)

- What contemporary "unbiased" methods that capture transient interactions can we employ to identify targets of activators?

Cofactor Review

- How are cofactors that lack DNA binding domains directed to various places in the genome?

- These interactions may be transient and not stably detected by conventional protein/protein interaction experiments.