

# Transcription Factors

ChIP-seq measures TF binding to DNA.

ChIP-seq also measures histone modification, cofactor, and RNA Polymerase genomic locations—however, their occupancy are a consequence of TF binding.

Mike Guertin

# Broad lecture goals:

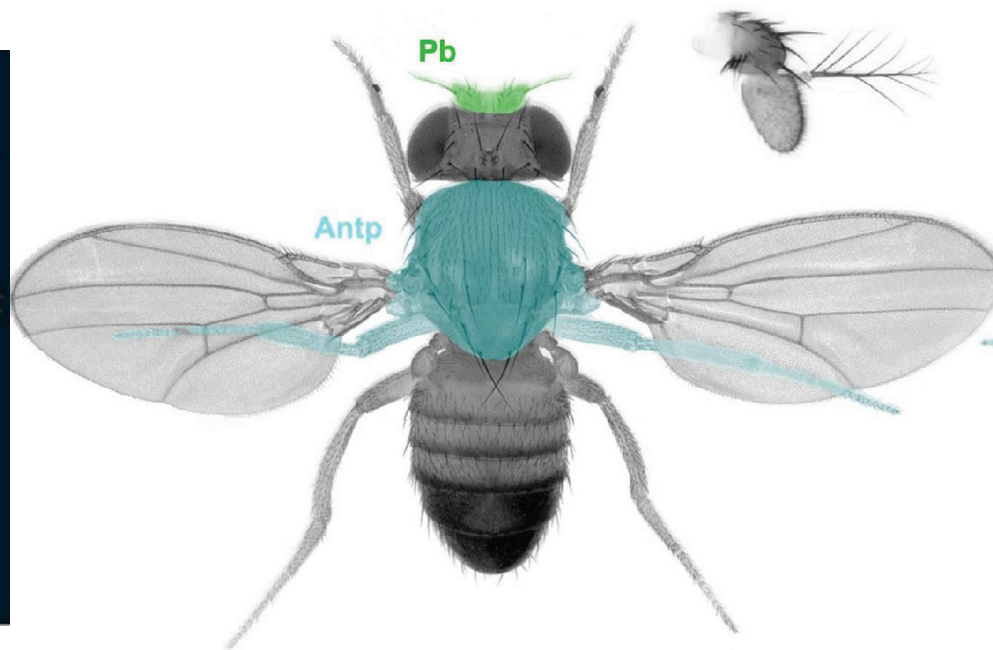
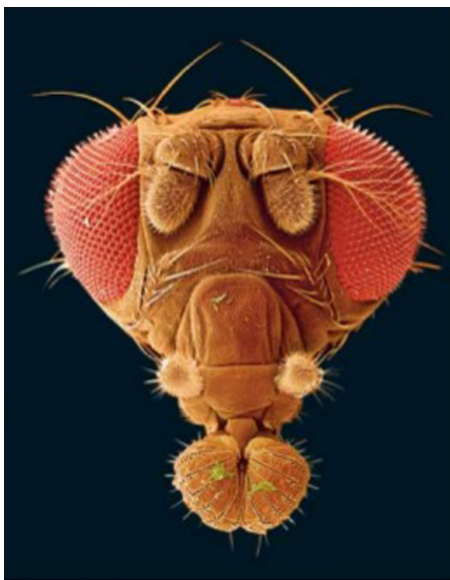
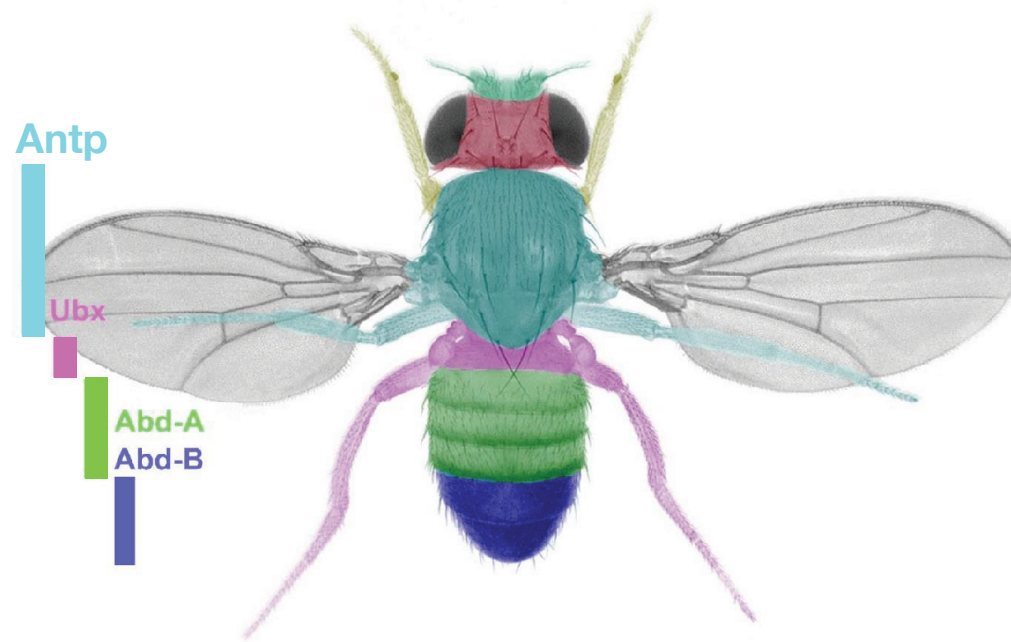
- Convince you of the importance of transcription factors in providing specificity in chromatin biology.
- Introduce classic experiments that defined principles of TF biology and provide references so one can follow up. Note that most molecular biology was interpreted through looking at bands on gels. **As a graduate student, you should aim to be able to take a well-written figure legend and figure and interpret the results.**
- Illustrate the point that biology is continuous, not discrete; relative quantification and controls are important.
- Emphasize the role of question-driven exploratory experiments (screens, molecular genomics, unbiased proximity label transfers, solving structures, etcetera) in defining principles of transcription factor biology.



# Transcription dysregulation alters developmental patterning



# Transcription dysregulation alters developmental patterning



pseudocolored flies: Justin Crocker, Ed Lewis, Nicolas Gompel, and Welcome Bender

pseudocolored SEM heads: Jürgen Berger



Classic genetics (perturb, observe, map) found that Transcription Factors control developmental patterning

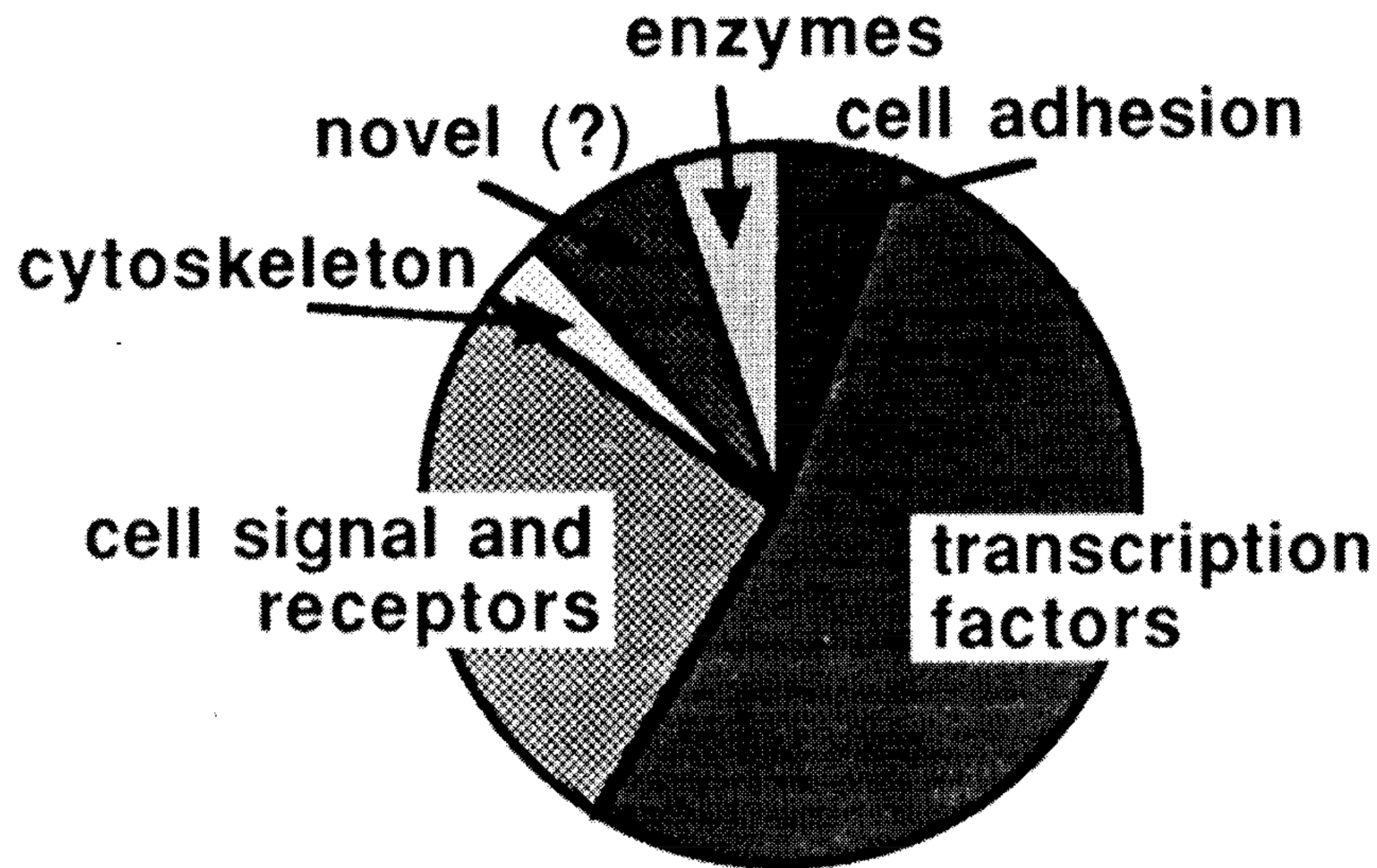
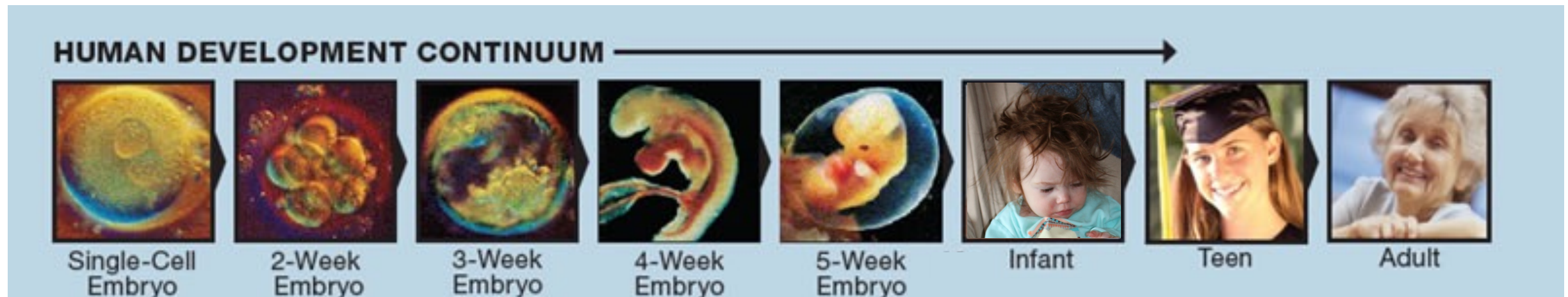
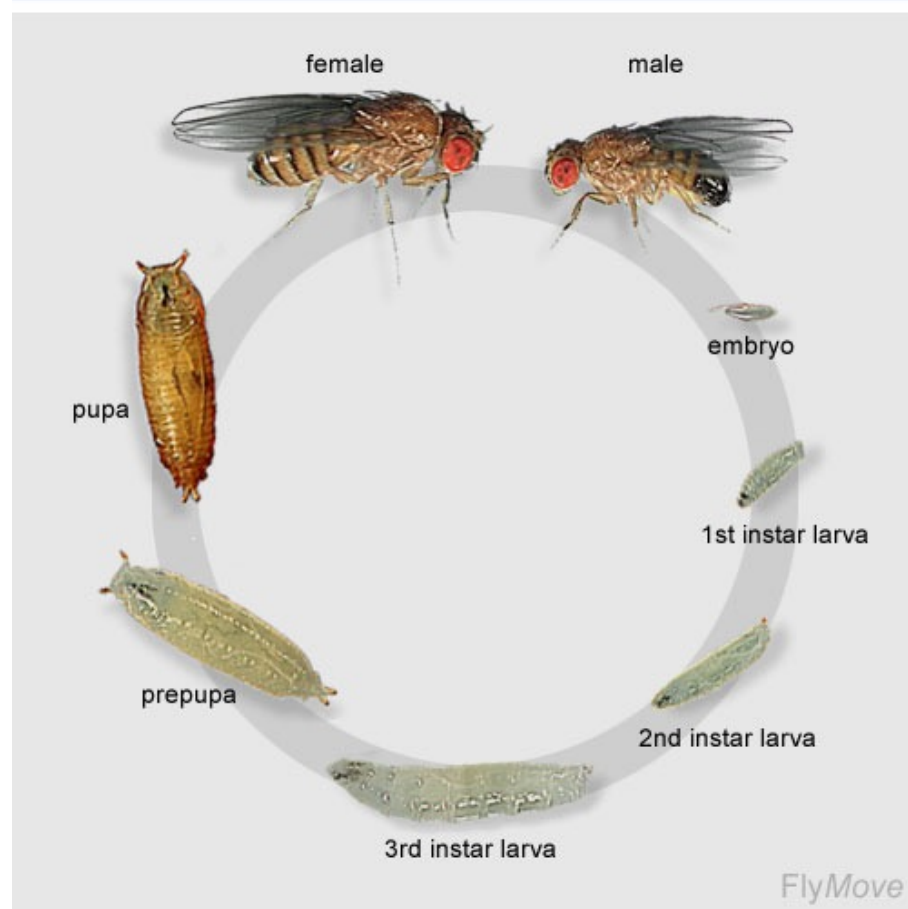


Figure 3. *Cellular Function of Heidelberg Mutations*. Based on the sequence of 75 cloned genes, most of the loci identified in Heidelberg encode transcription factors, or cell signals and receptors.

# Transcription control is key in development and homeostasis



The life cycle of *Drosophila melanogaster*



**Embryonic cells progress from totipotent to a spectrum of more specialized states.**

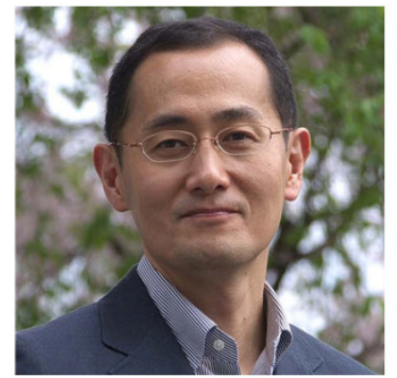
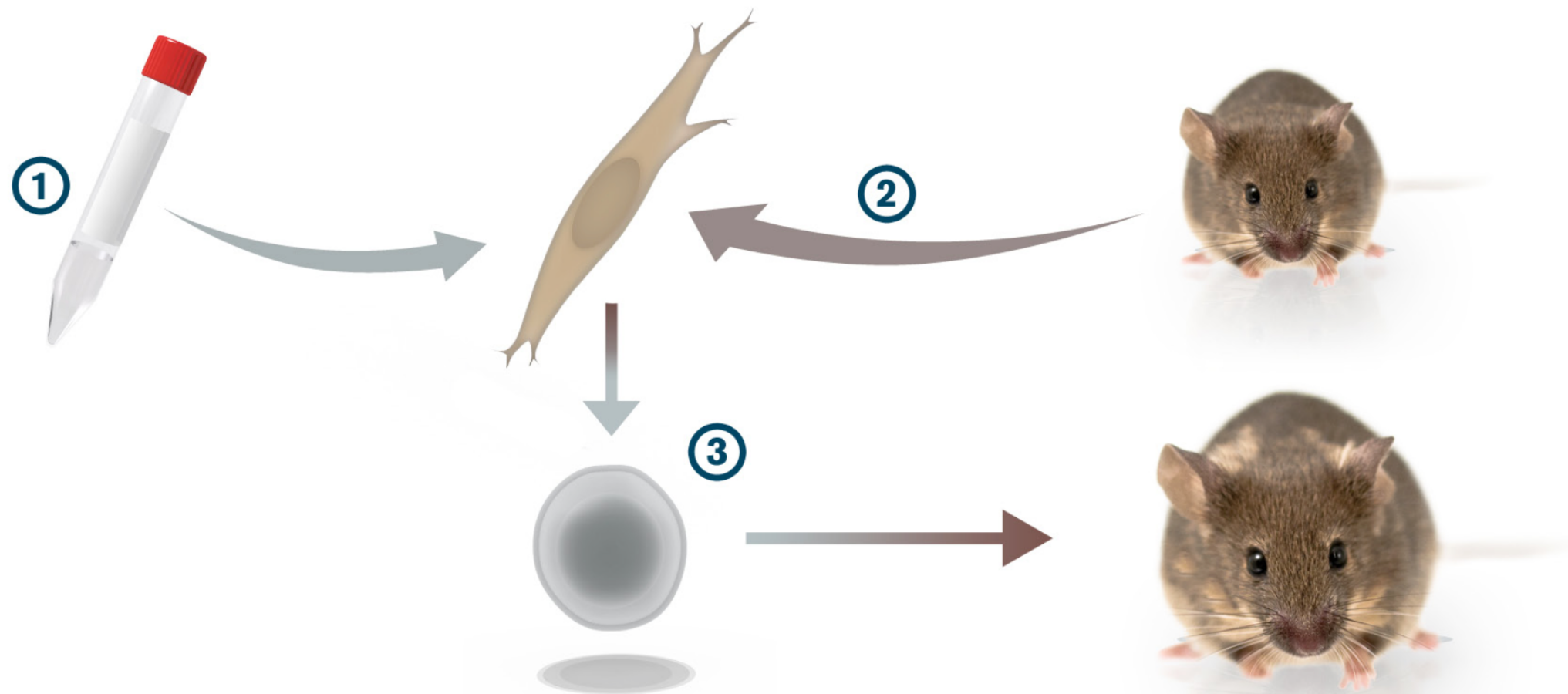
**Much of this developmental regulation starts at transcription.**

**Cells need to respond to changing nutrients and environments.**

**Organisms have sophisticated programs of transcription regulation.**

## 2012 Nobel in Physiology or Medicine:

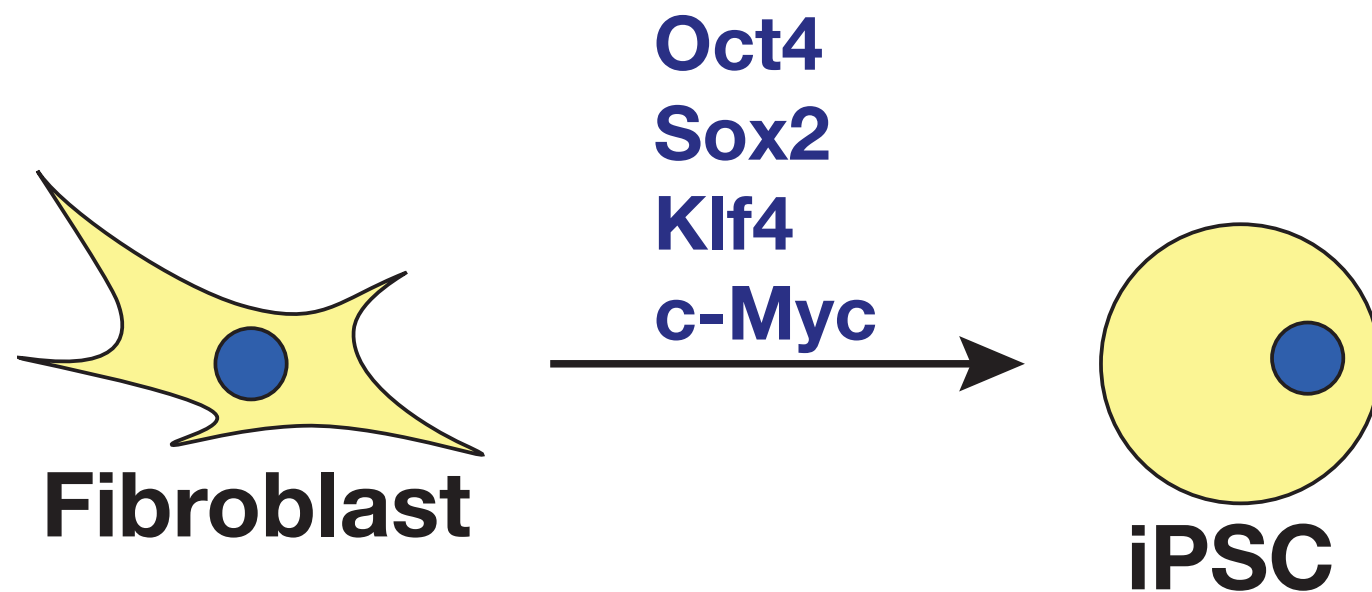
“for the discovery that mature cells can be reprogrammed to become pluripotent”



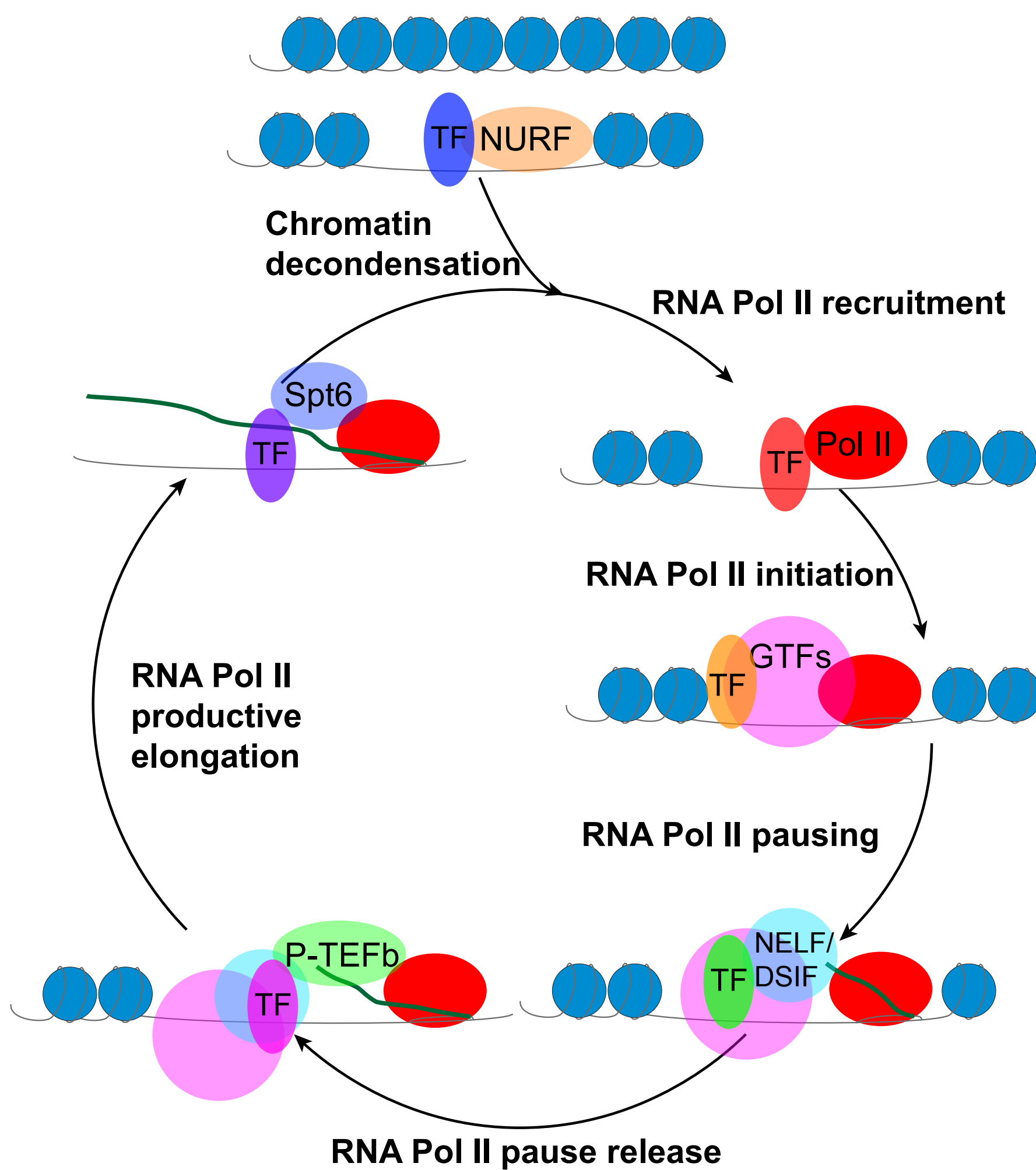
**Shinya Yamanaka**

Shinya Yamanaka studied genes that are important for stem cell function. When he transferred four such genes (1) into cells taken from the skin (2), they were reprogrammed into pluripotent stem cells (3) that could develop into all cell types of an adult mouse. He named these cells induced pluripotent stem (iPS) cells.

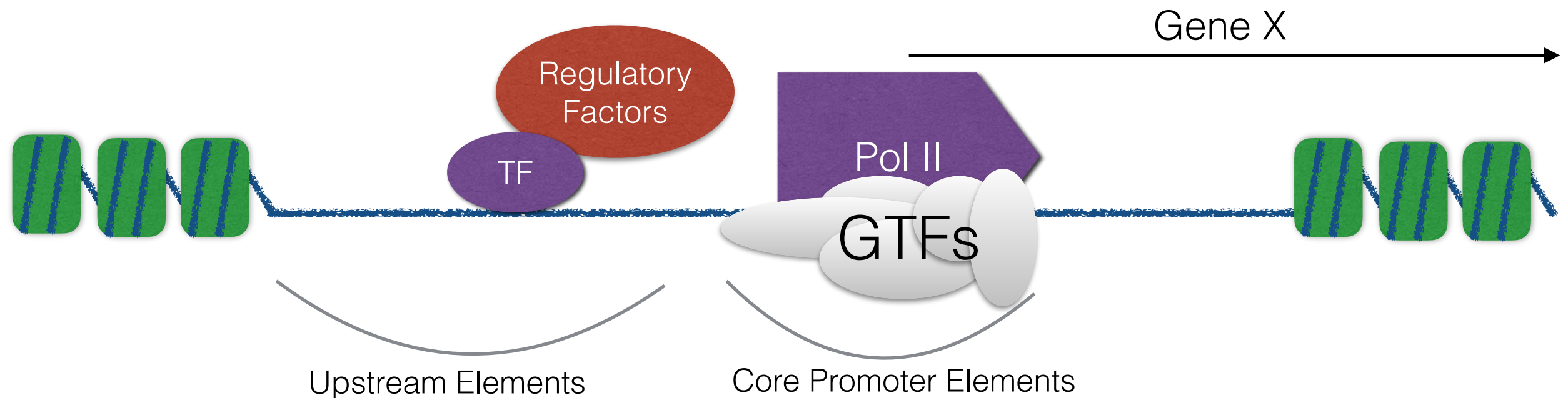
# Activating transcription factors changes cell identity





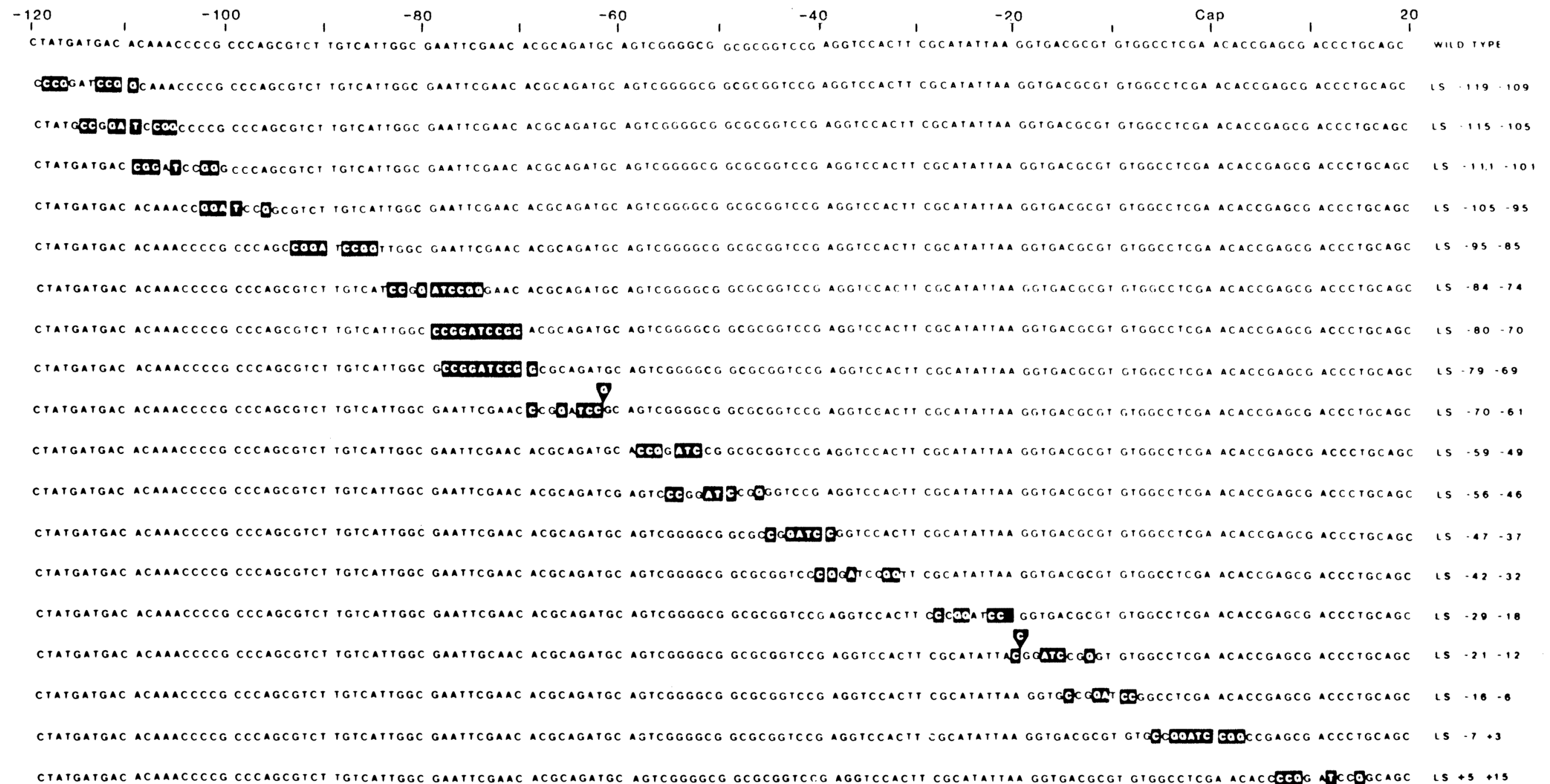


# Transcription Regulation by Transcription Factors (TFs) is determined by DNA sequence





# Linker Scanning Mutations of the thymidine kinase gene of HSV



Clusters of point mutants are generated at the point of joining 5' and 3' deletions, where linker sequence substitutes for tk sequence.

# Assay expression of tk promoter mutants

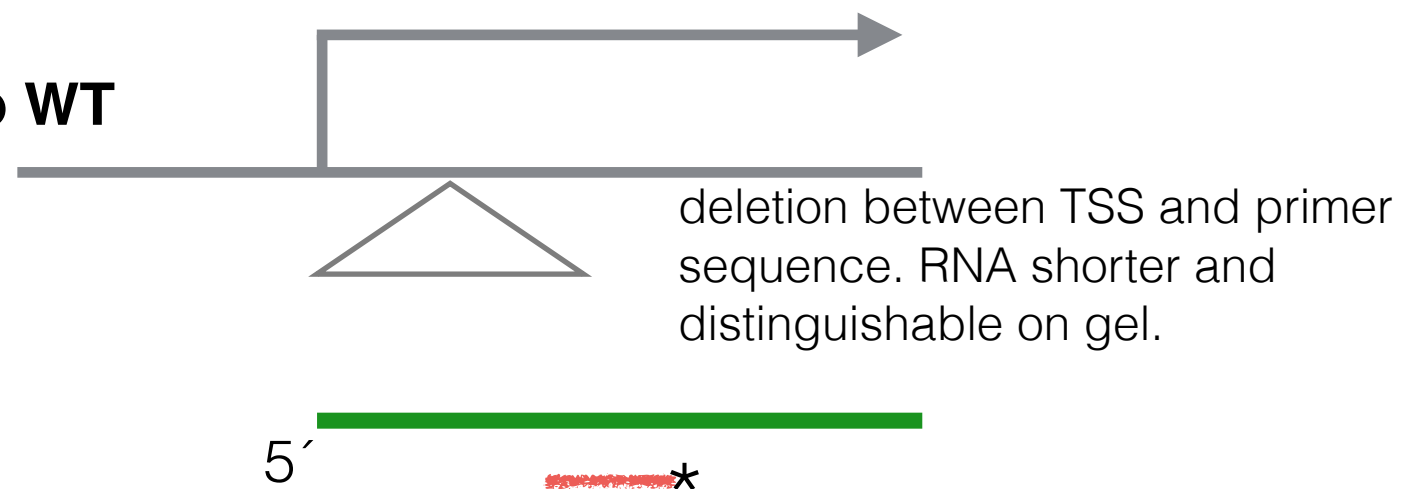
## Plasmid DNAs

Inject mutant DNA into frog oocyte nuclei, include **pseudo WT standard** as internal control

Isolate RNA

Primer Extension Assay

Gel Electrophoresis



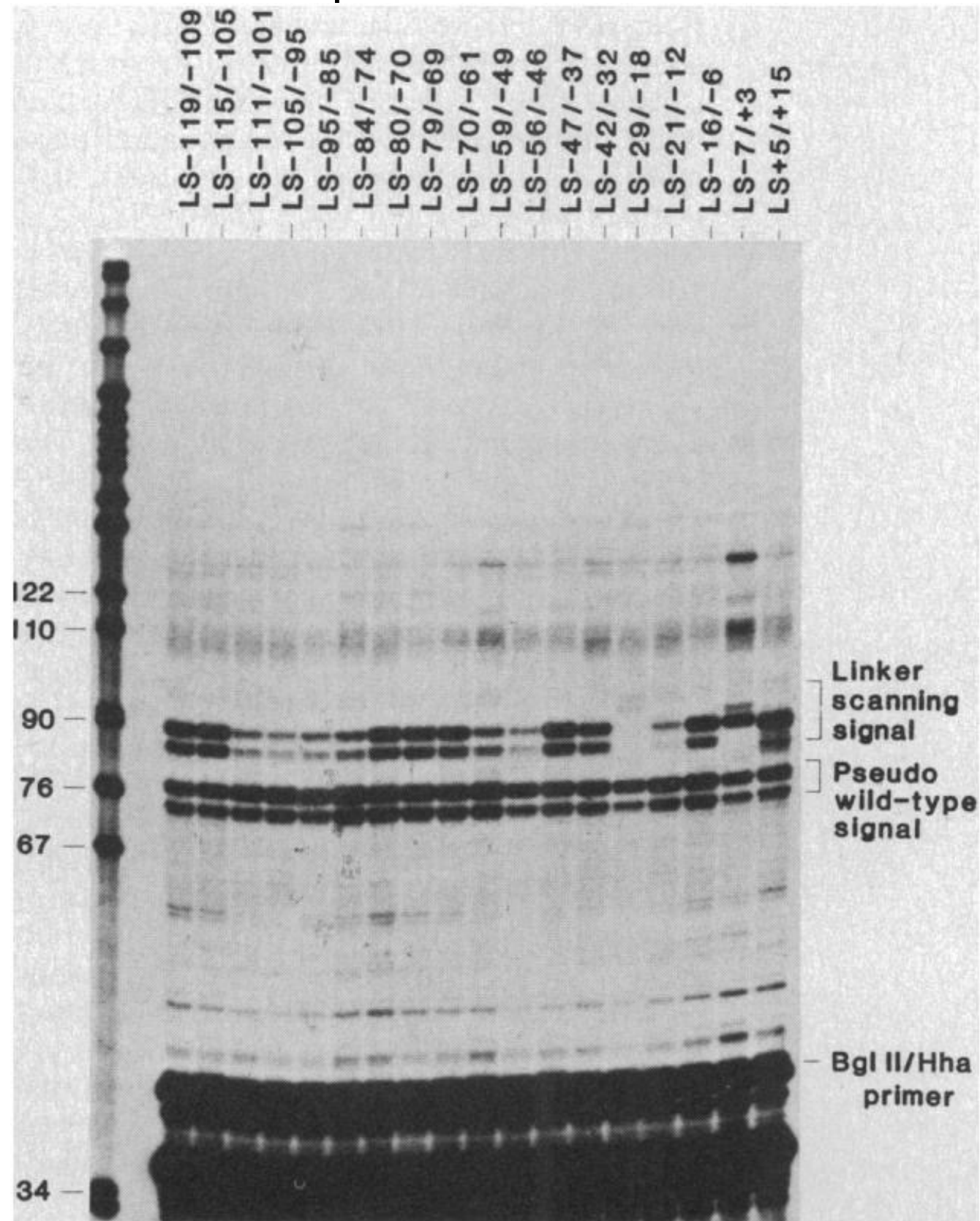
tk promoter mutant

tk promoter mutant

WT tk promoter; deletion of gene body

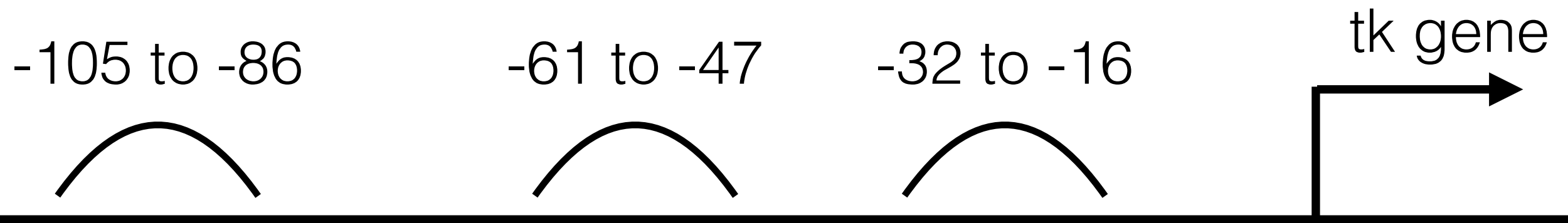
WT tk promoter; deletion of gene body

# Expression data from Linker Scanning Mutants



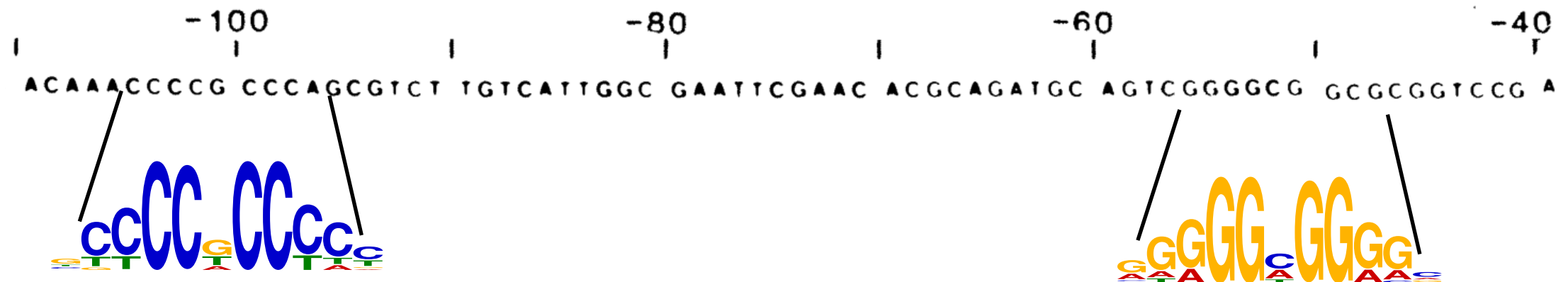
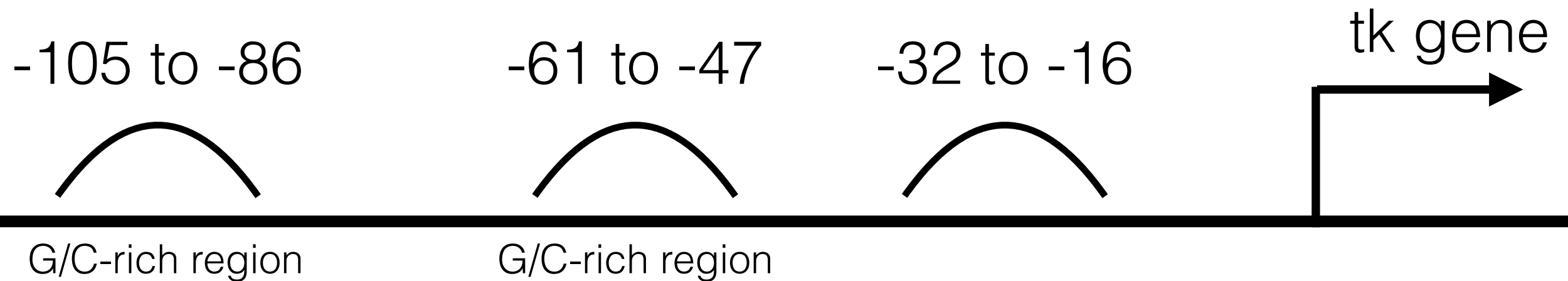
Short discontinuous regions of sequence are critical for basal expression.

Three promoter regions are critical for basal  
expression

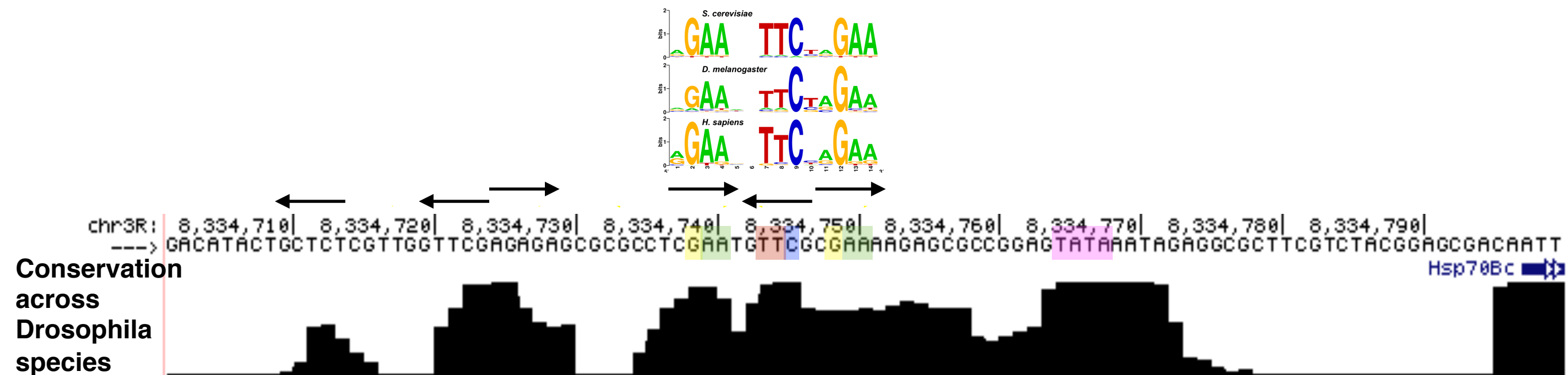


TATA element located  
approximately 30bp  
upstream of the TSS

Three promoter regions are critical for basal expression



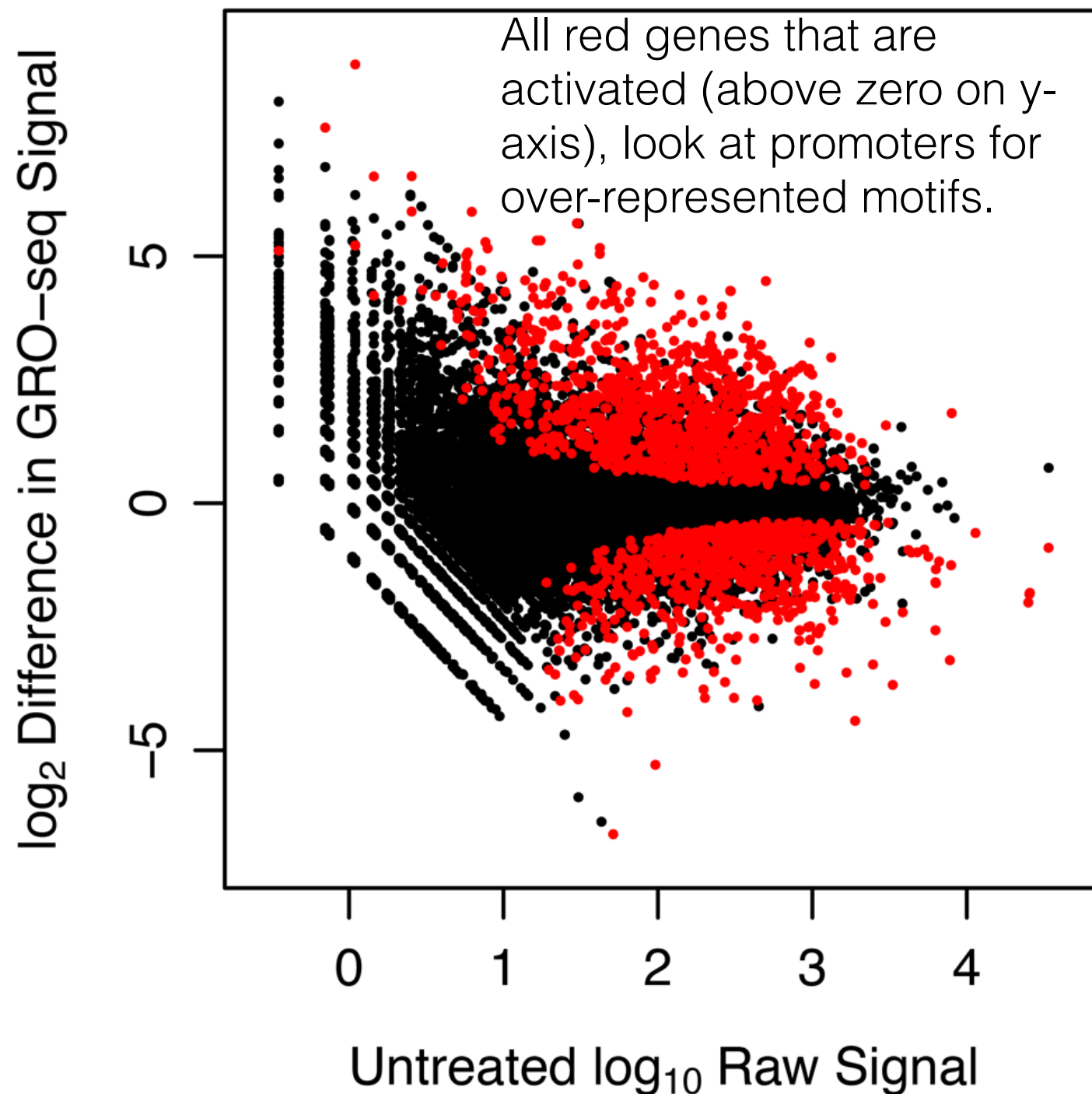
# Evolutionary conservation and comparative genomics can identify crucial elements



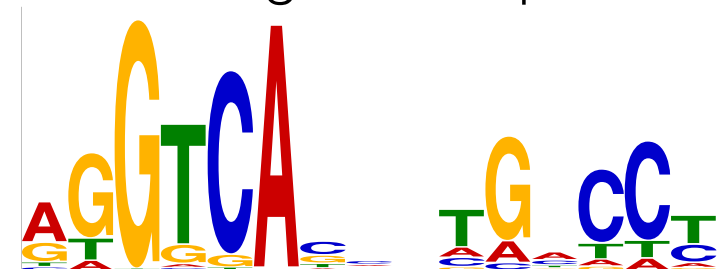
# Significant changes in nascent transcription upon estrogen treatment in breast cancer cells

Collect the sequences of multiple (co-regulated) promoters within a species, search for common sequence motifs

## Sustained Changes at both 10min and 40min



*de novo* motif analysis using MEME (or the alike) identifies the Estrogen Response Element, the known target of the Estrogen Receptor.



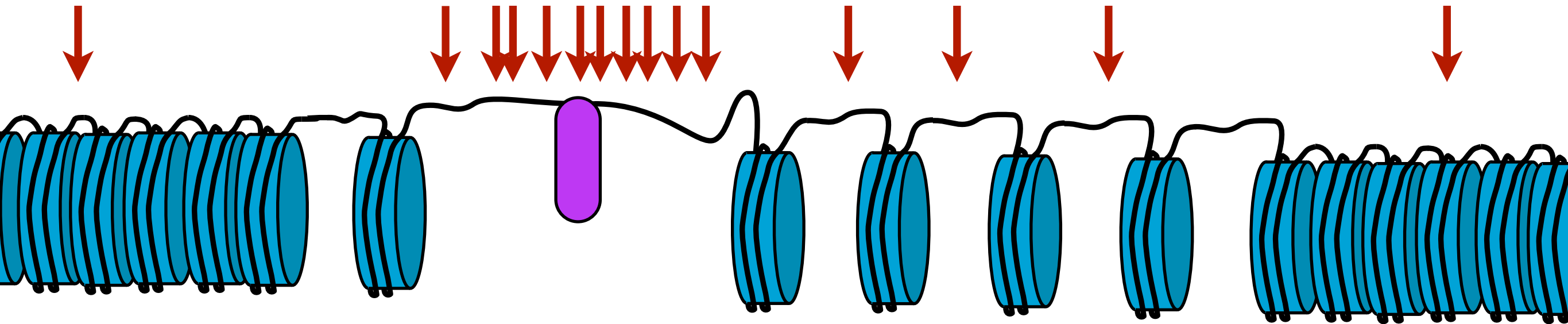
Note that not all regulatory elements bound by TFs are within promoters.

# Identify All Active Regulatory Elements in a Cell Type: Enzyme Hypersensitivity (DNase-seq & ATAC-seq)

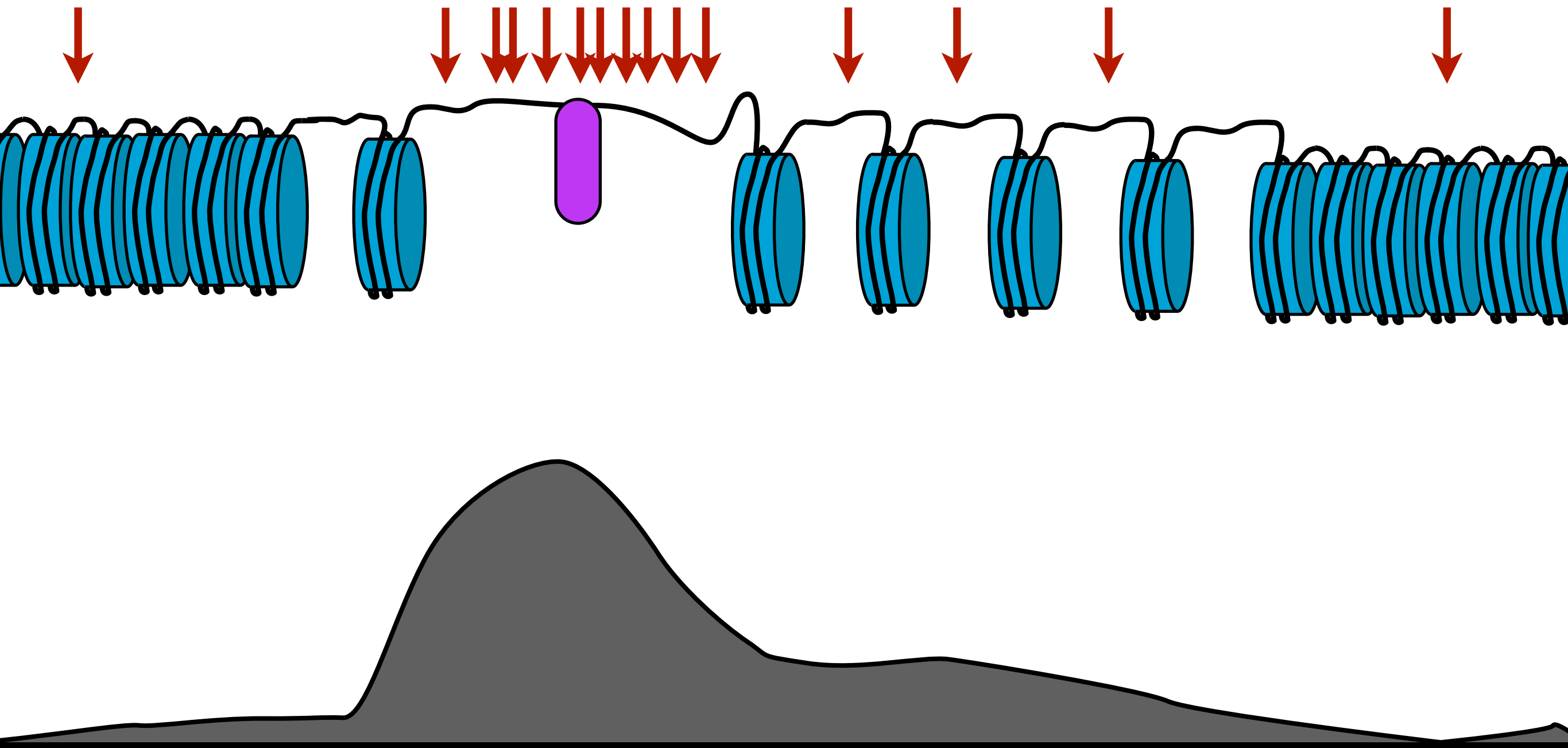
- A general measure of chromatin structure.
  - Factor/species-general
  - Changes in enzyme hypersensitivity landscape after drug treatment or throughout development can be used to identify novel regulatory elements and factors
  - Generally unbiased, but challenging to deconvolve
- **TFs controlling chromatin landscape can be inferred from the data**



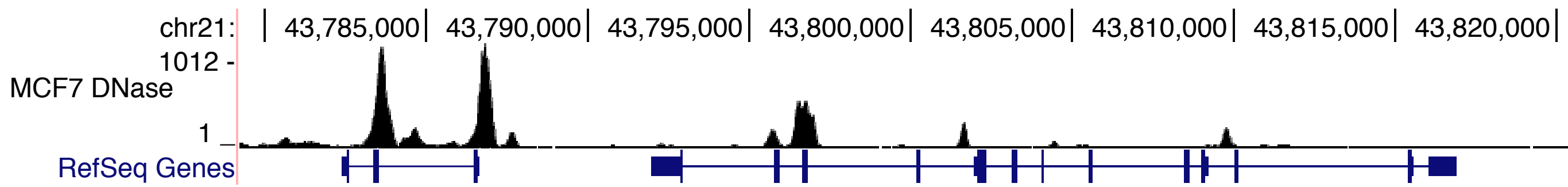
# Enzyme Hypersensitivity



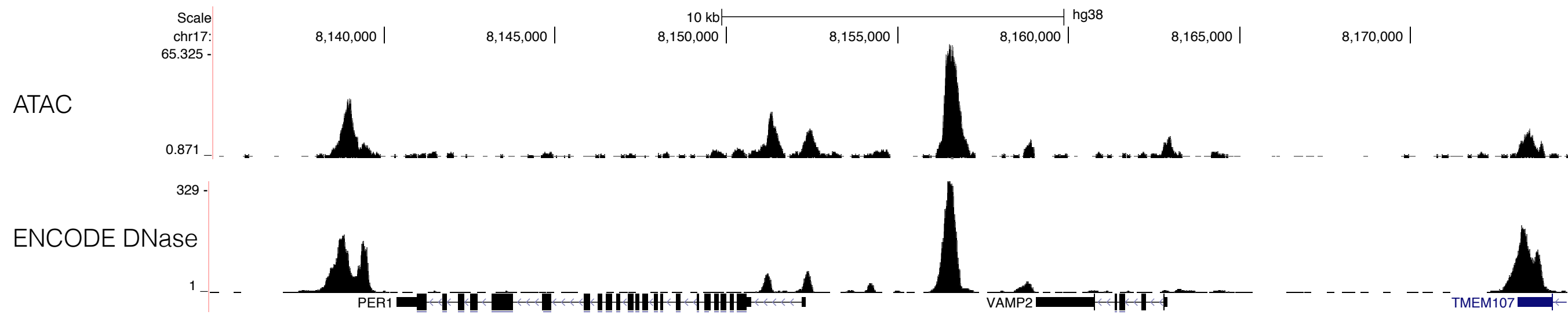
# Enzyme Hypersensitivity



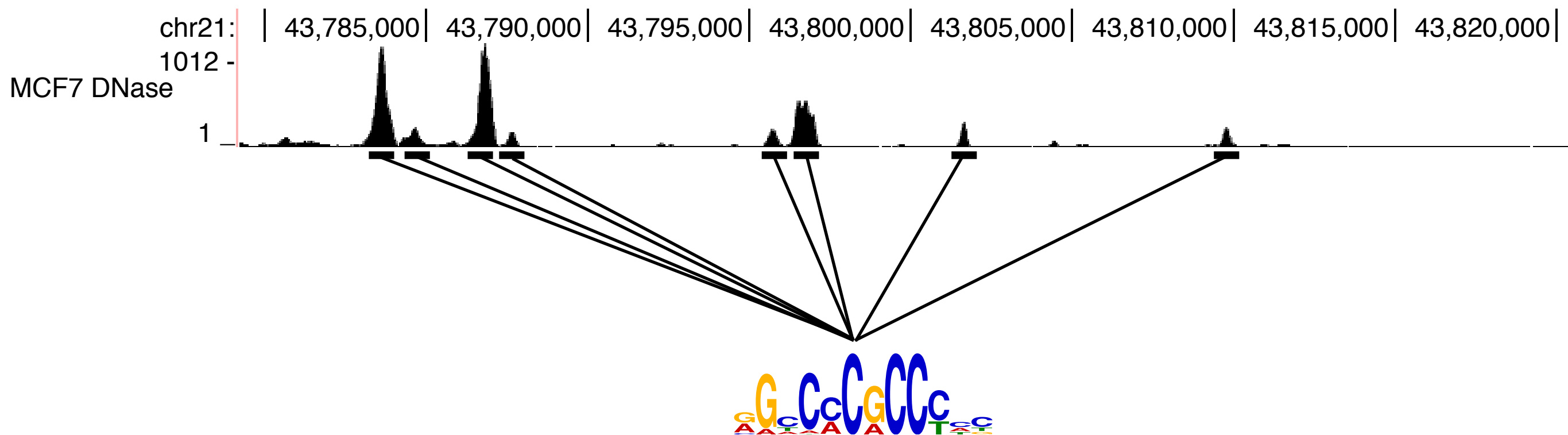
# DNase-seq Data



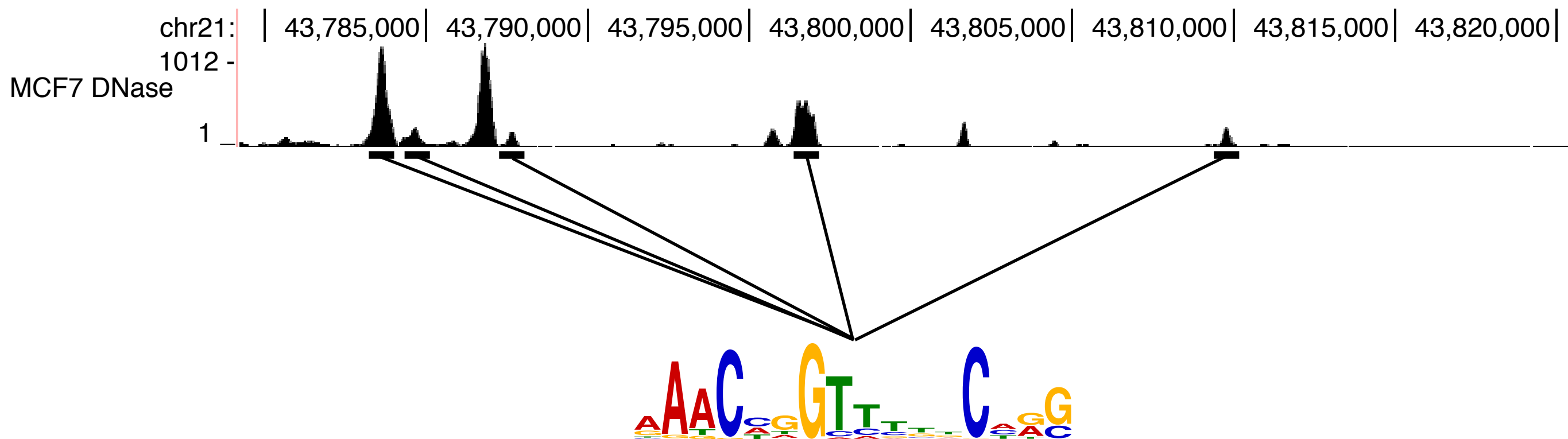
# ATAC-seq vs. DNase-seq



# DNase/ATAC identifies a repertoire of TF motifs

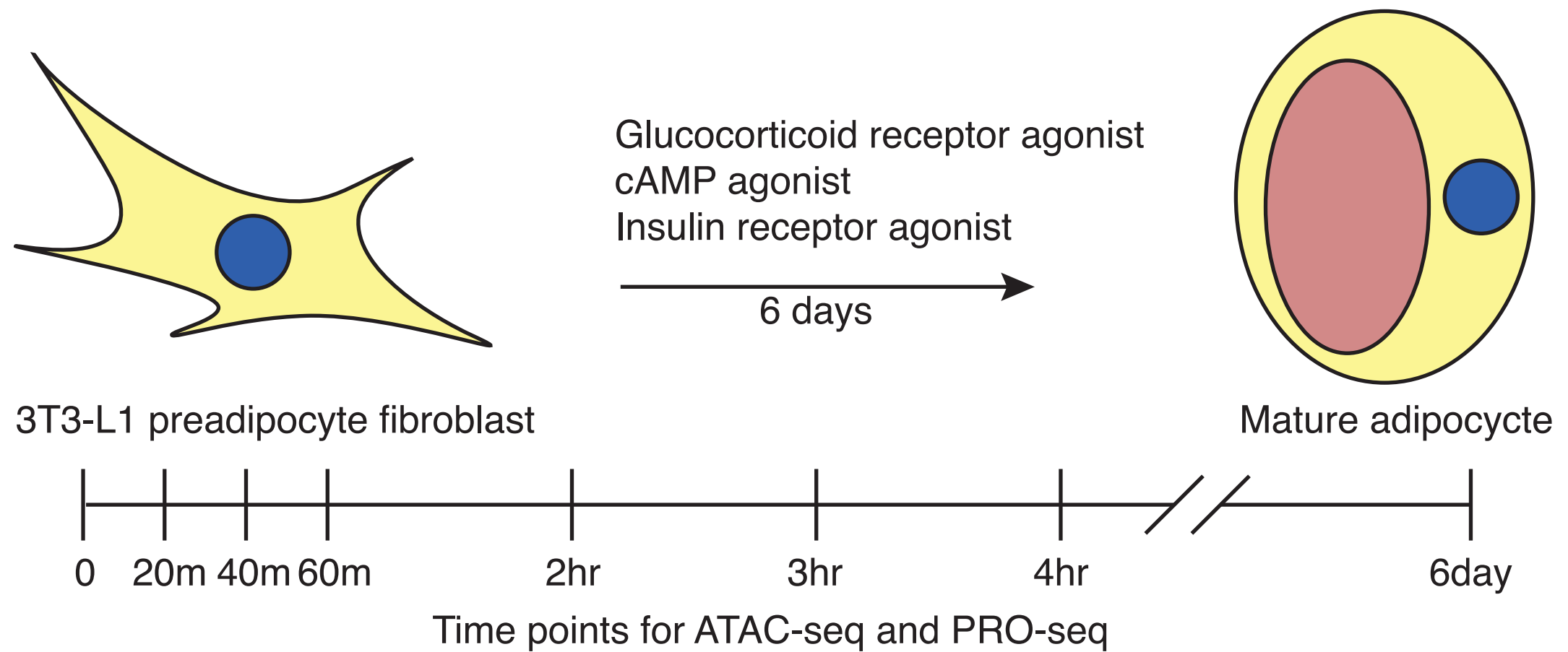


# DNase/ATAC identifies a repertoire of TF motifs

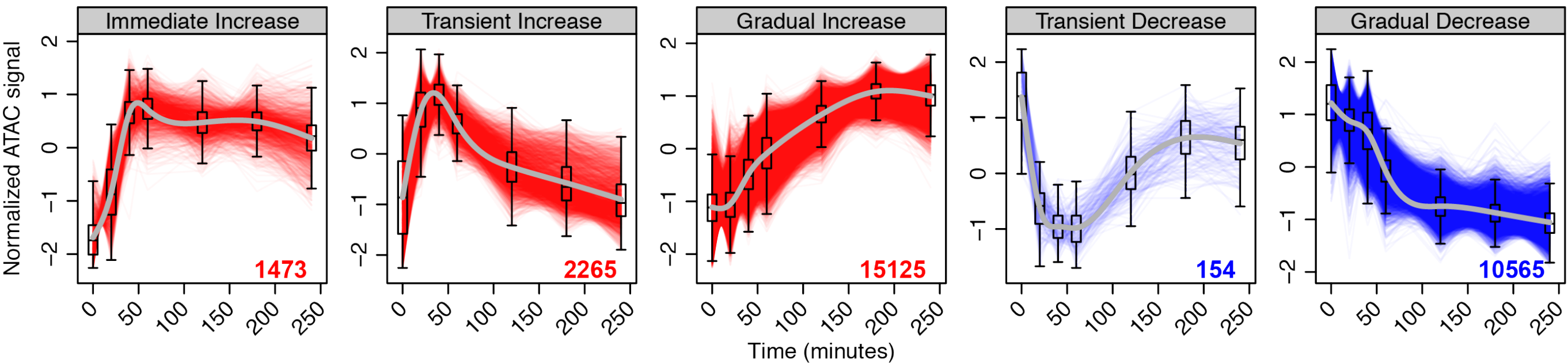


Identify sequence elements at hypersensitive site using iterative de novo motif analysis

# Experimental Design

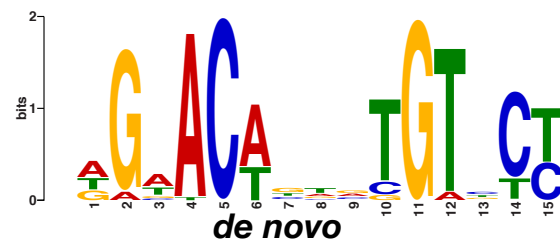
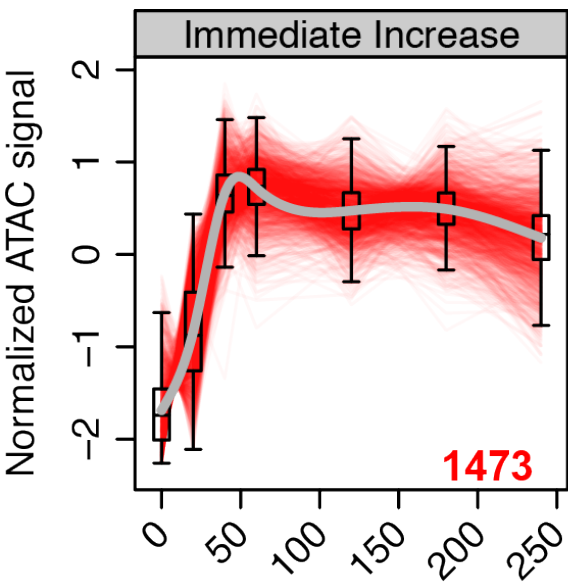


# ATAC-seq peaks have distinct accessibility kinetics

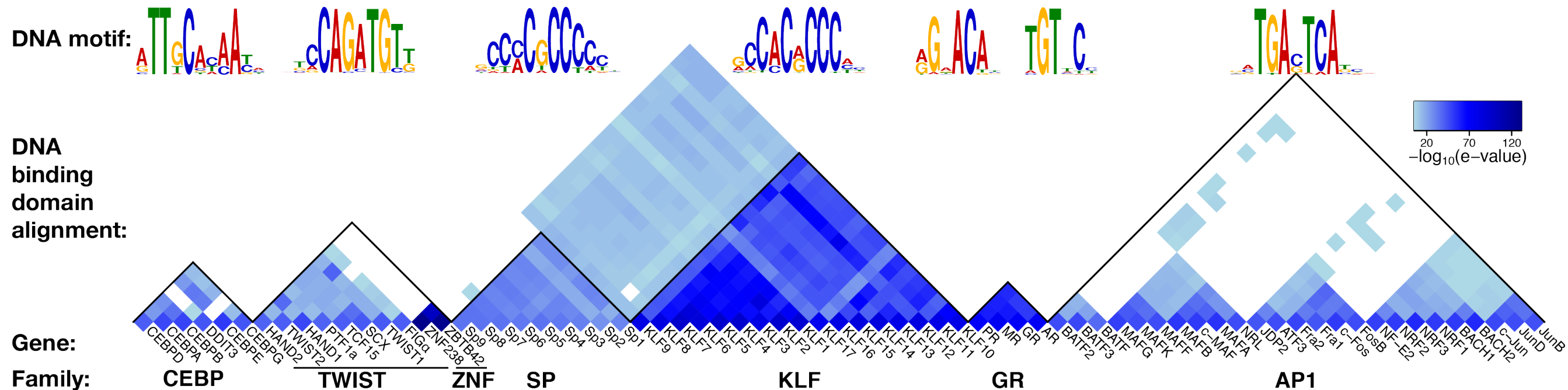




# *de novo* motif analysis identifies enriched sequence elements within dynamic ATAC peaks



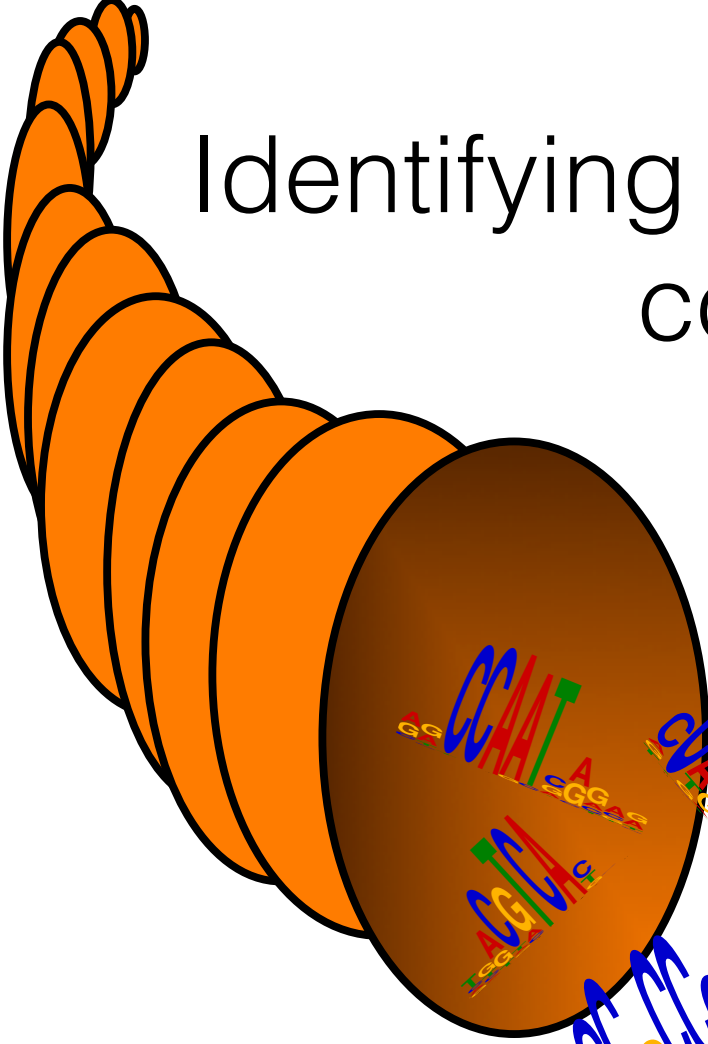
# 14 TF-family motifs (top 6 shown) drive early adipogenesis changes in accessibility



Paralogous TF DBD families that recognize each motif

# Identifying

co

A stylized orange DNA double helix structure, composed of several overlapping orange rings, curves from the top left towards the bottom right. The helix terminates in a large, brown circular base. This base contains several colorful letters representing nucleotides: 'A' (red), 'T' (blue), 'C' (green), and 'G' (yellow). Some letters are partially cut off by the edges of the frame. The background is plain white.

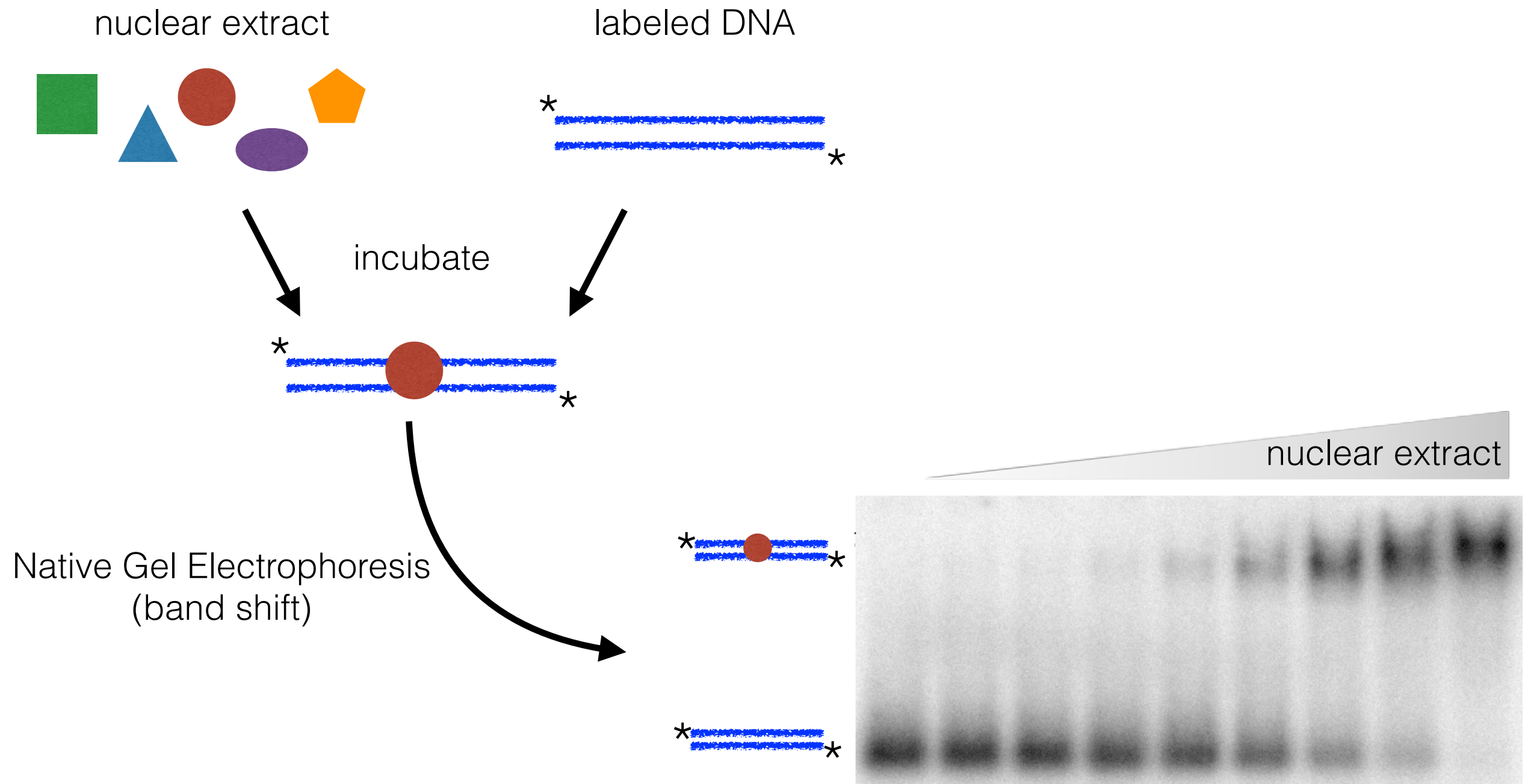
## How can we identify them?

How can we identify them?

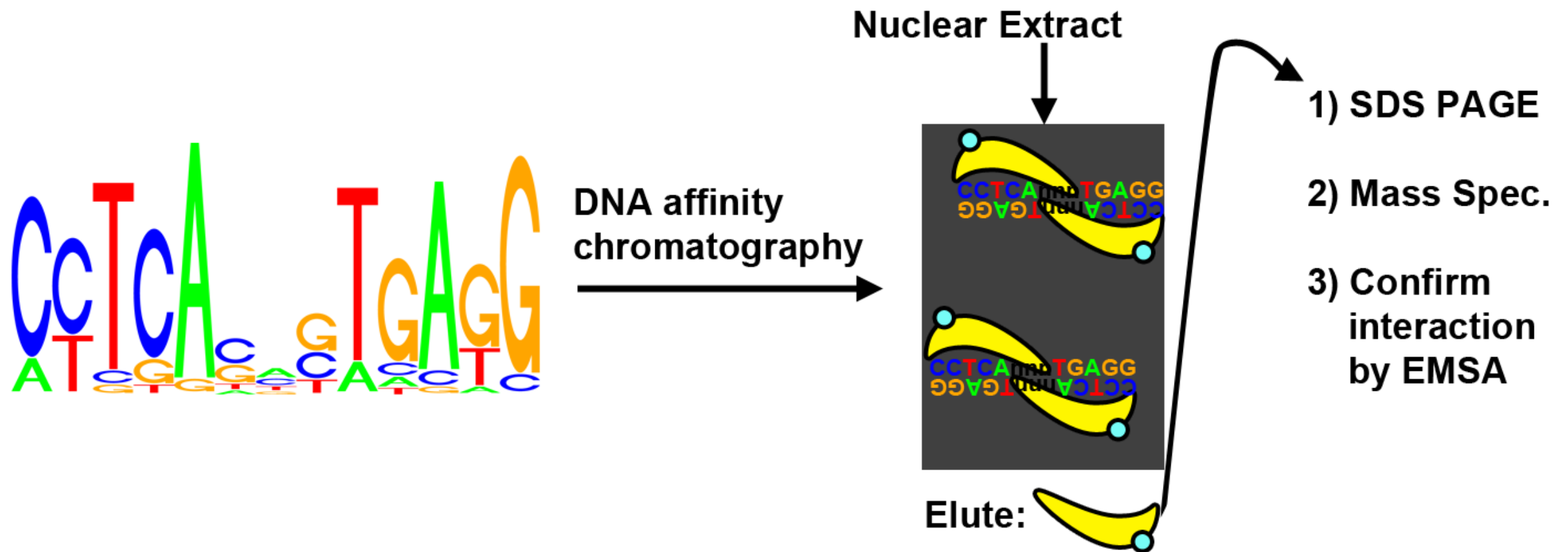
What is needed to purify such protein

**Knowing specific sequences helps create both an assay and tool for purifying.**

# Electrophoretic Mobility Shift Assay (EMSA) detect DNA binding factors



# Purification of sequence-specific DNA-binding proteins



Order oligos with modest variants of your consensus sequence (include random flanking DNA). Biotinylate the ends of the duplexed DNA, bind to streptavidin beads/column, elute, compare eluate to nuclear extract by PAGE, and mass spec.

# Summary: Part I

- Transcription and its regulation is specified by short DNA sequence elements.
- These elements interact with particular transcription factors.
- See Lambert et. al., The Human Transcription Factors, Cell 2018 for a review of TF/DNA binding

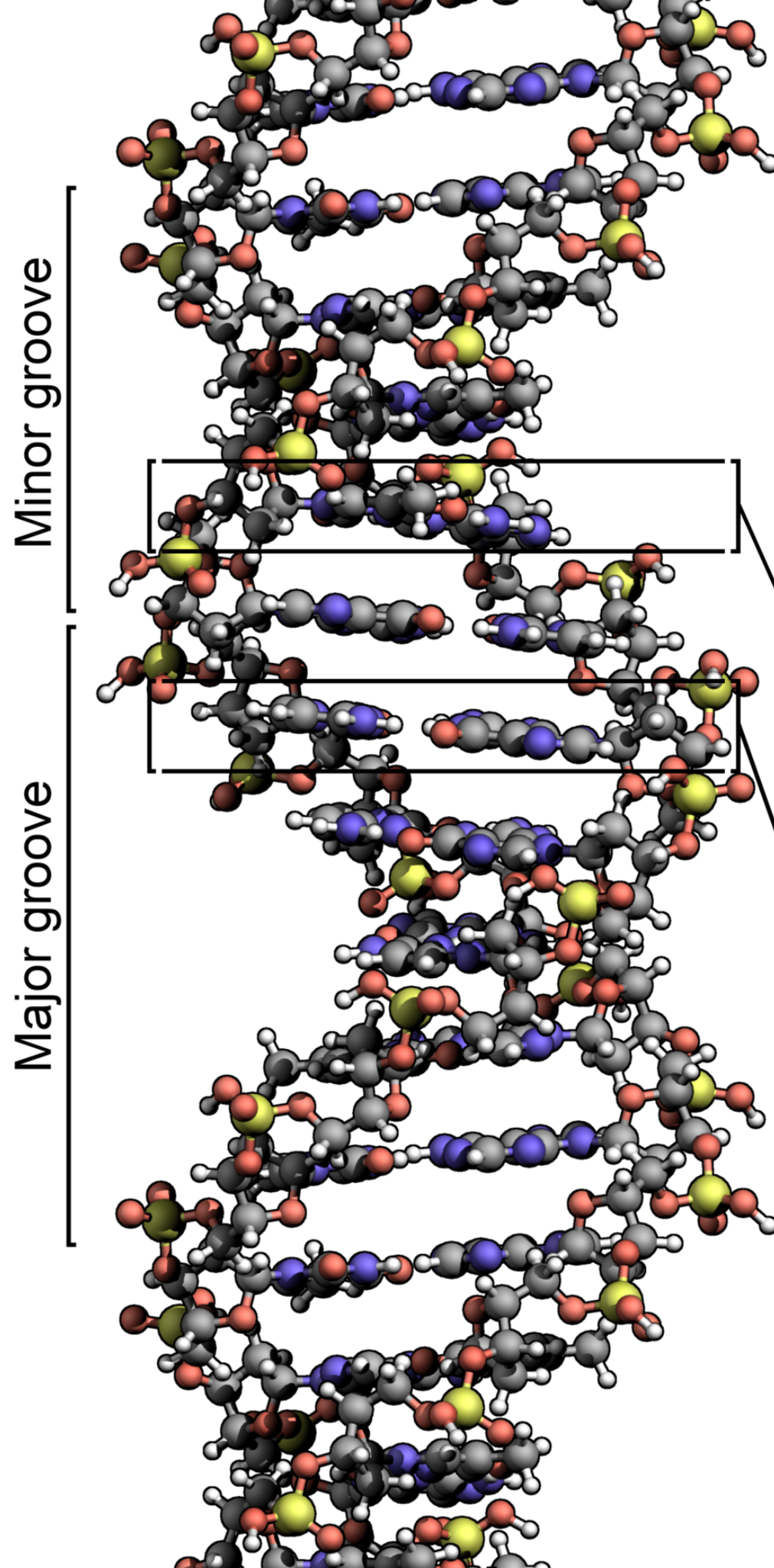
“Next thing is how a cell’s picking which GATs (stretches of nucleotides) get chosen, like Yogi in a picnic basket. Proteins and DNA? Some interesting chemistry. Cuz they getting jiggy with some different affinities.”

–Tom McFadden

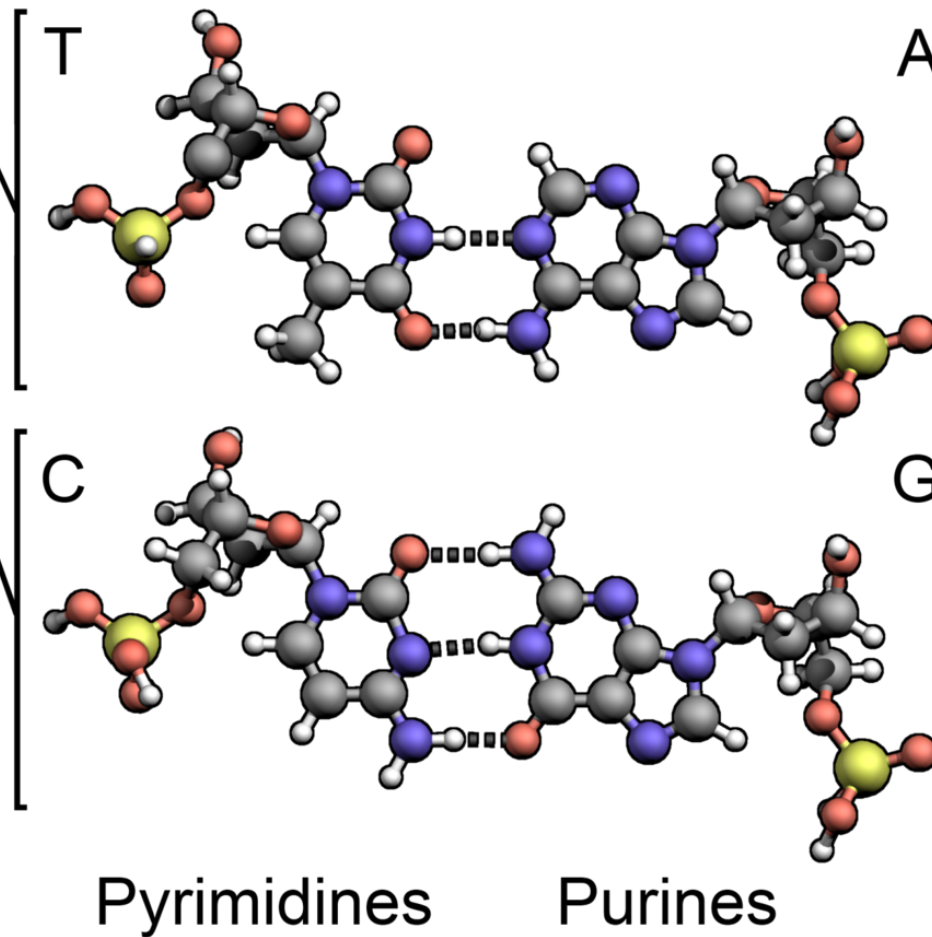
[https://www.youtube.com/watch?v=9k\\_oKK4Teco&list=RD9k\\_oKK4Teco](https://www.youtube.com/watch?v=9k_oKK4Teco&list=RD9k_oKK4Teco)



# How do proteins interact with specific DNA sequences?

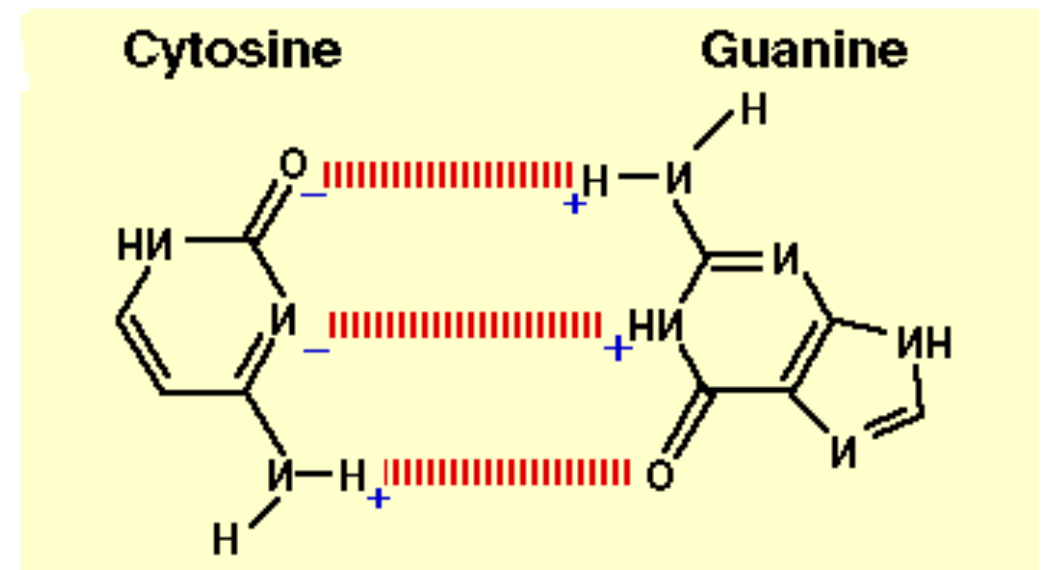
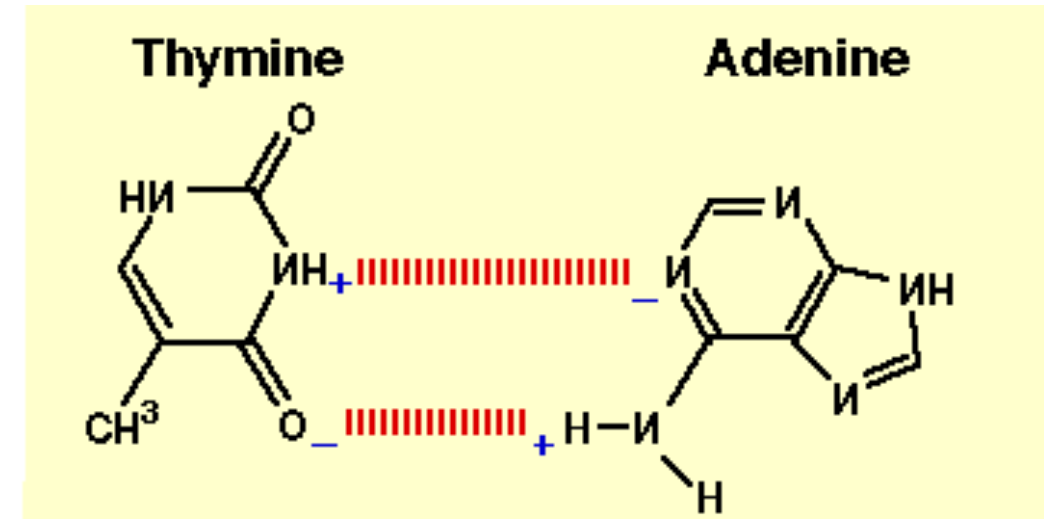
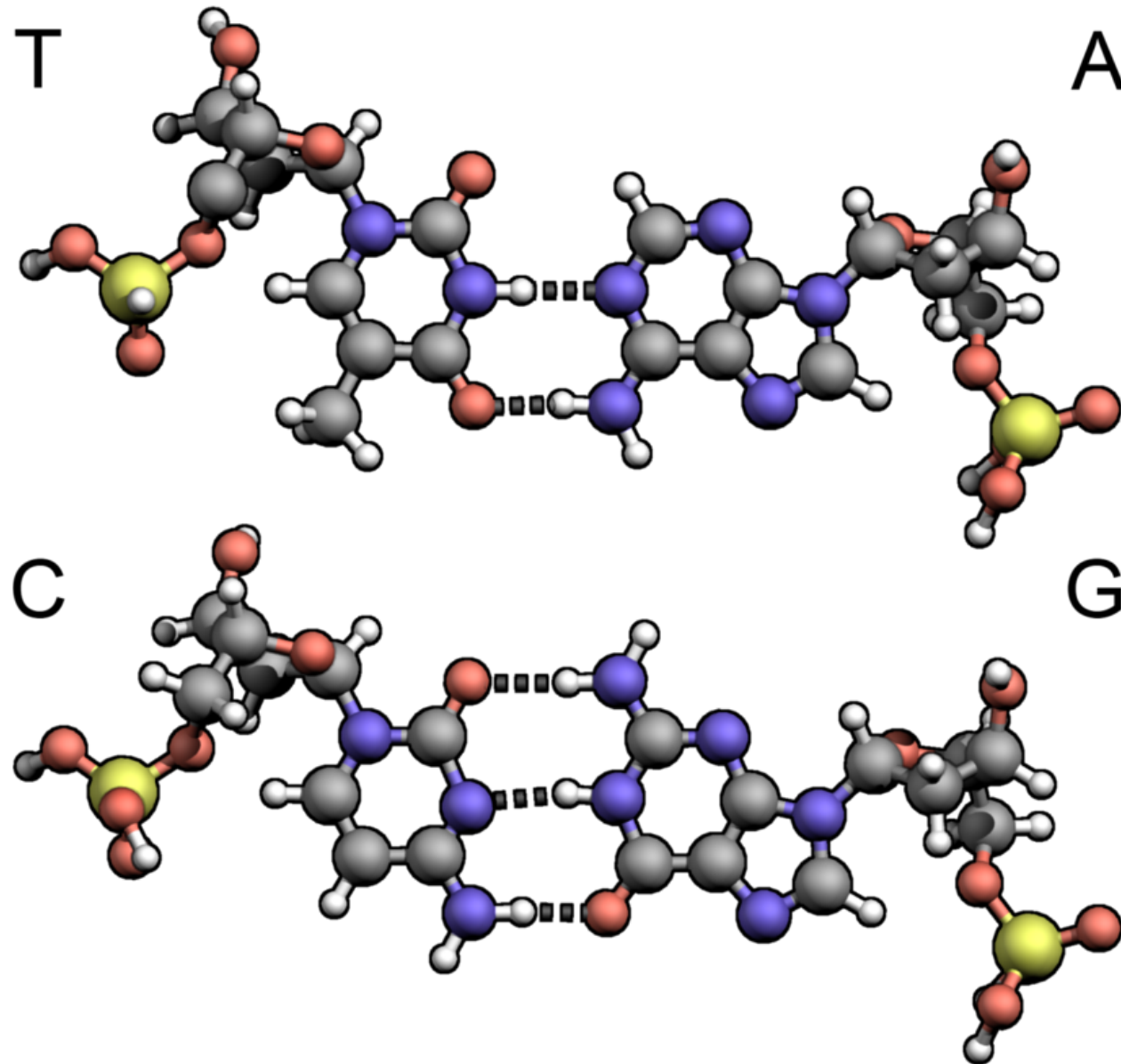


- Hydrogen
- Oxygen
- Nitrogen
- Carbon
- Phosphorus



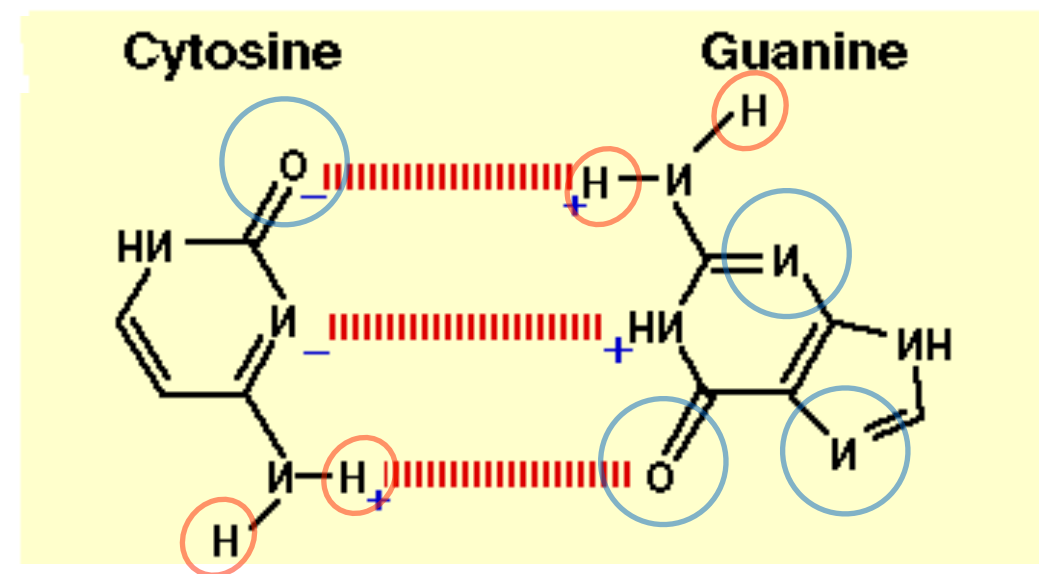
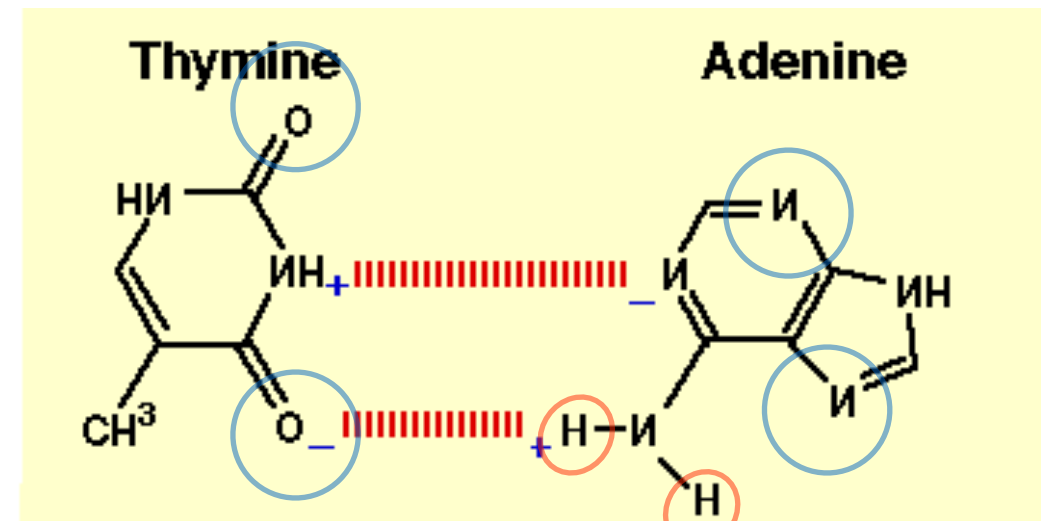
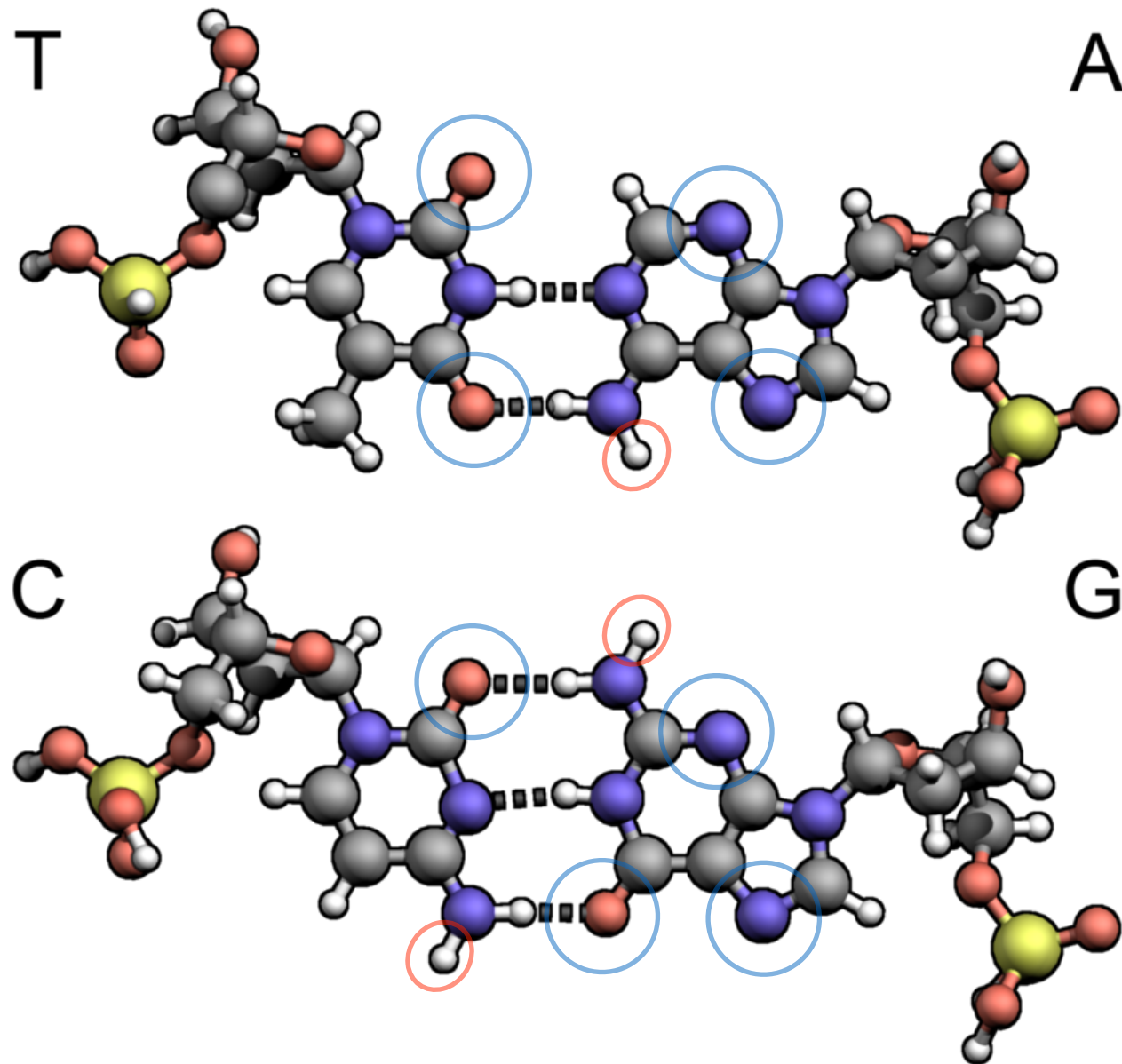


# Hydrogen bond is the electrostatic attraction between polar groups

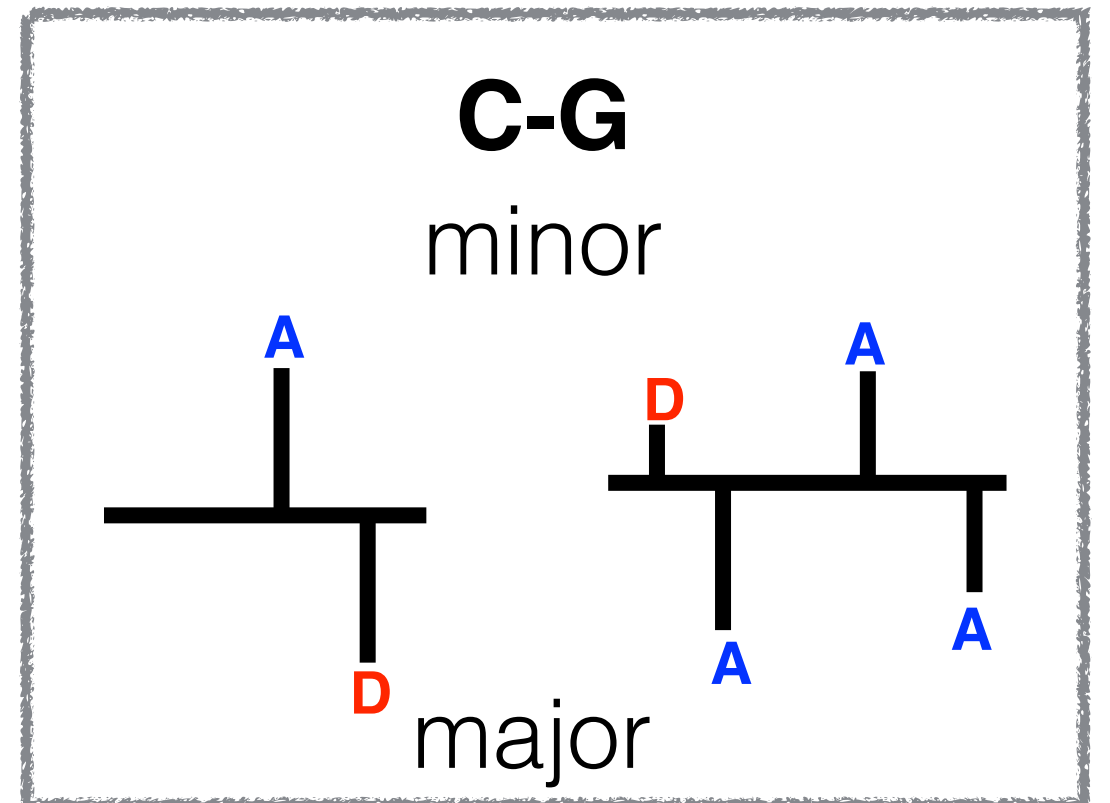
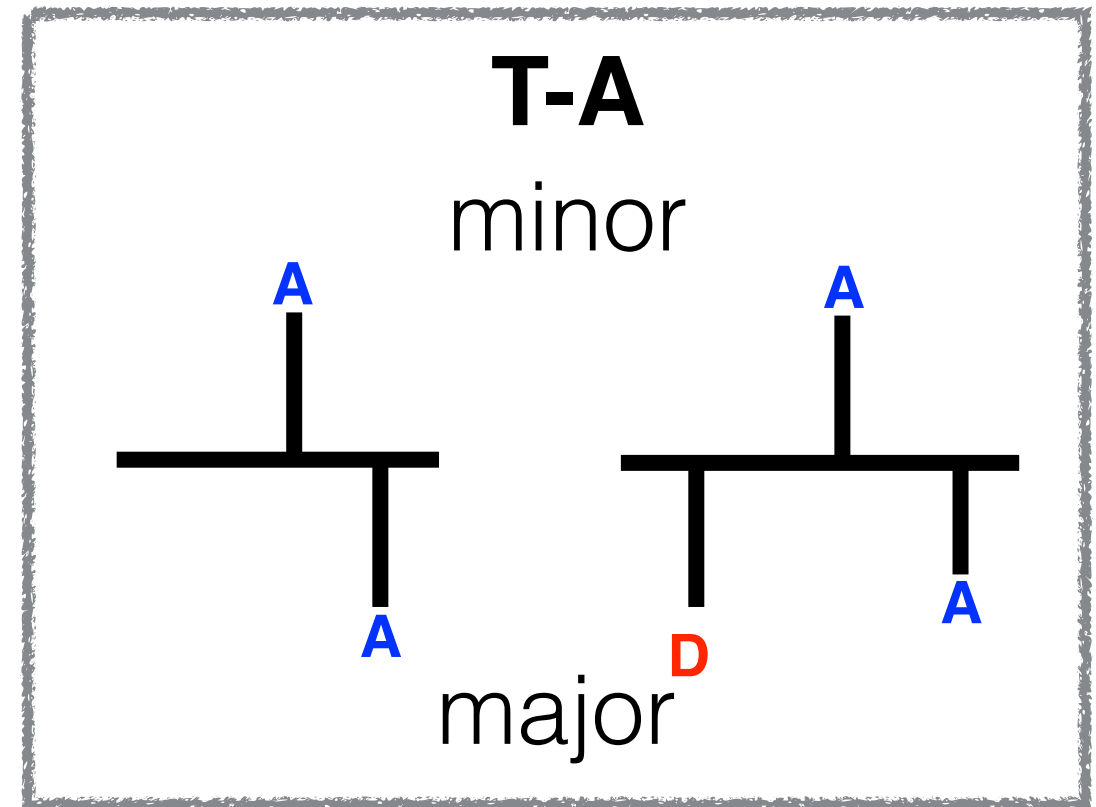
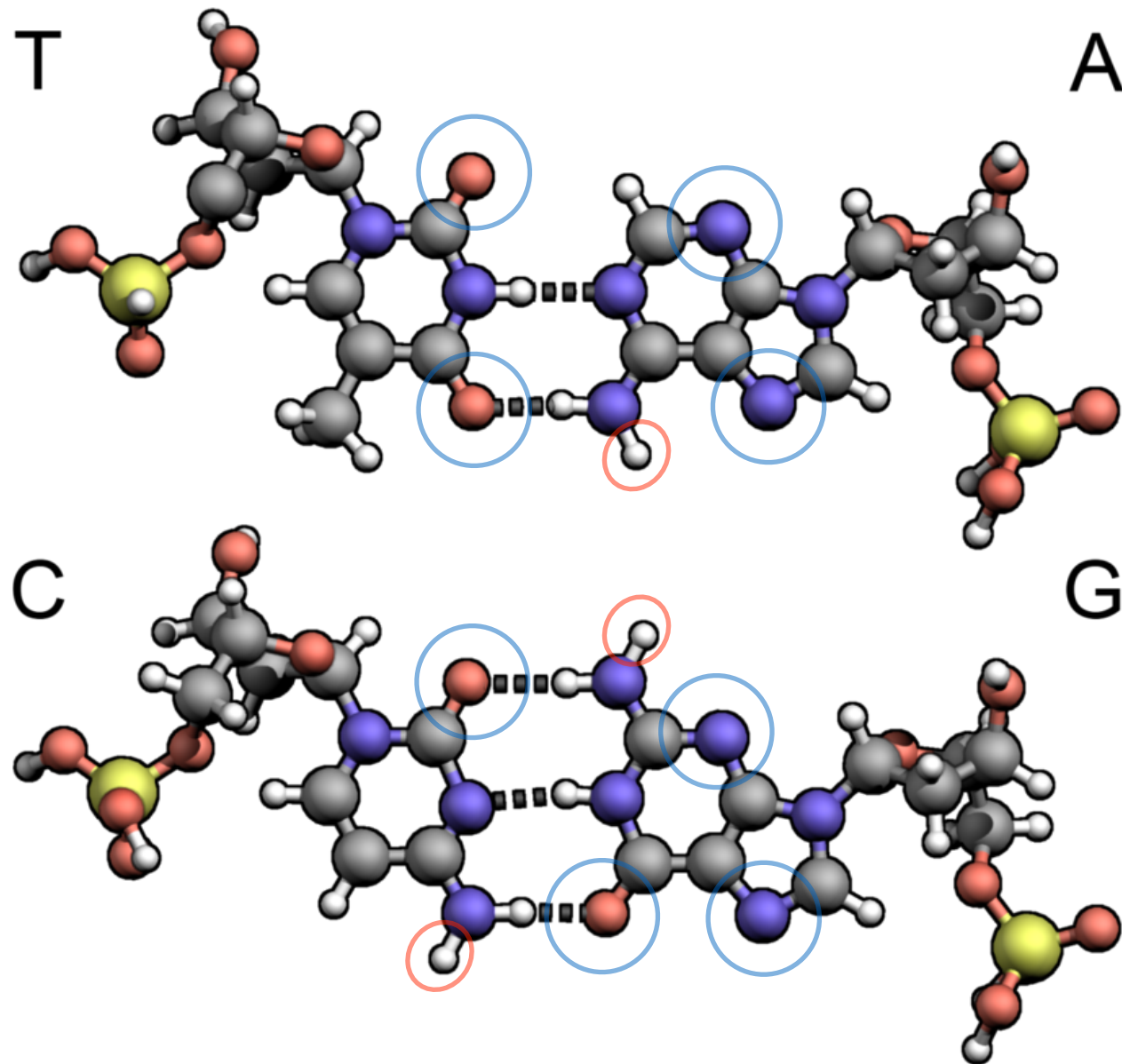


H-bond: a Hydrogen atom bound to a highly electronegative atom such as Nitrogen or Oxygen experiences attraction to another nearby highly electronegative atom.

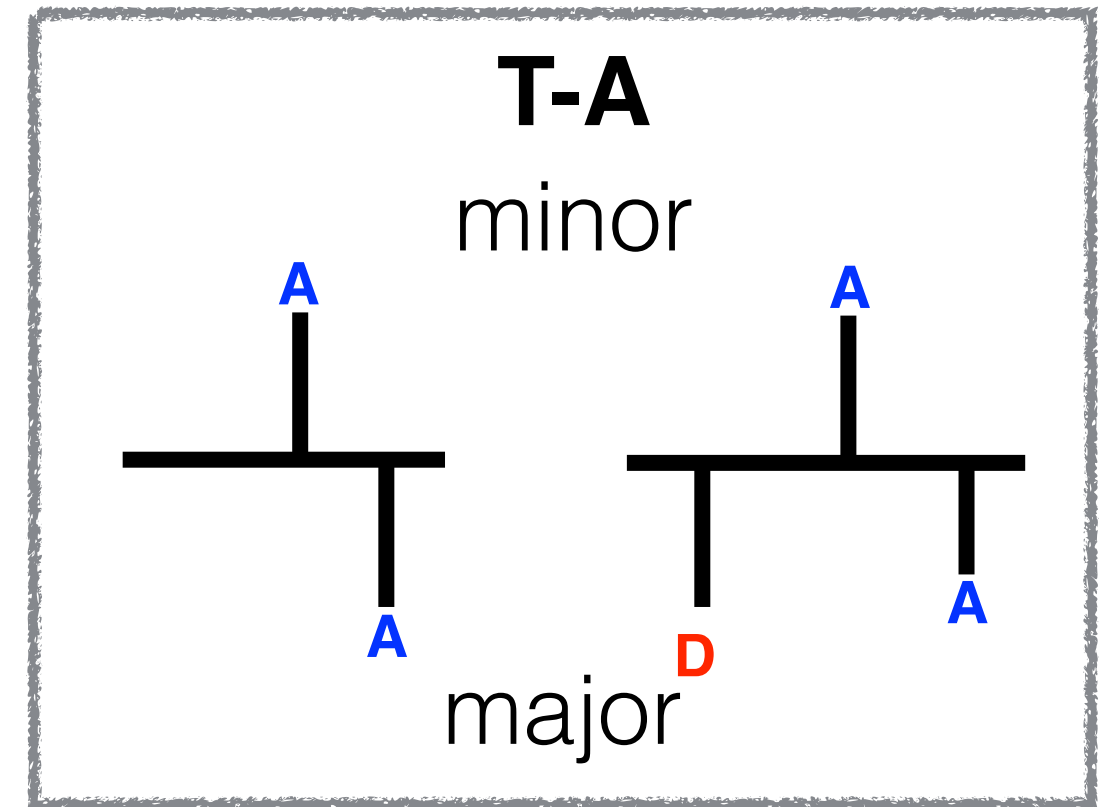
**The atoms shown below are available to mediate protein/DNA interactions via H-bonds**



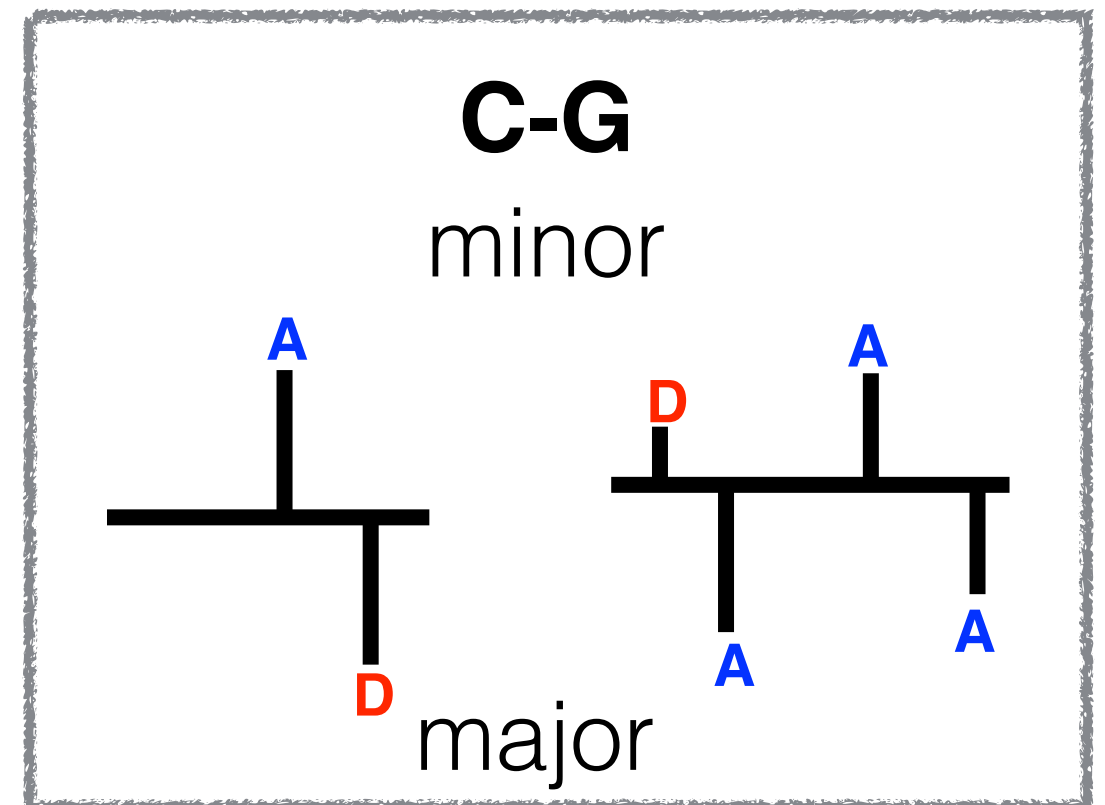
The atoms shown below are available to mediate protein/DNA interactions via H-bonds



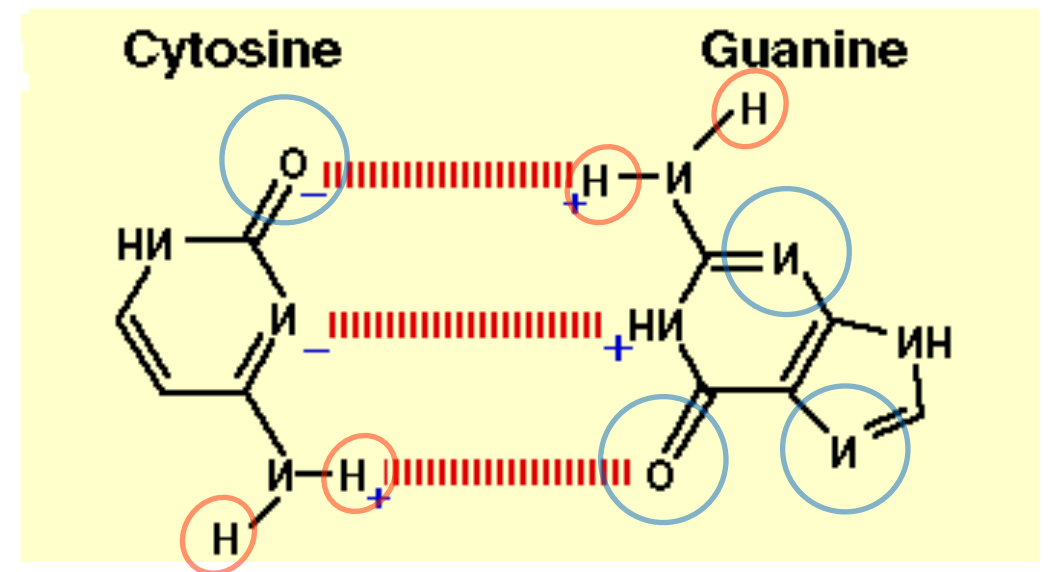
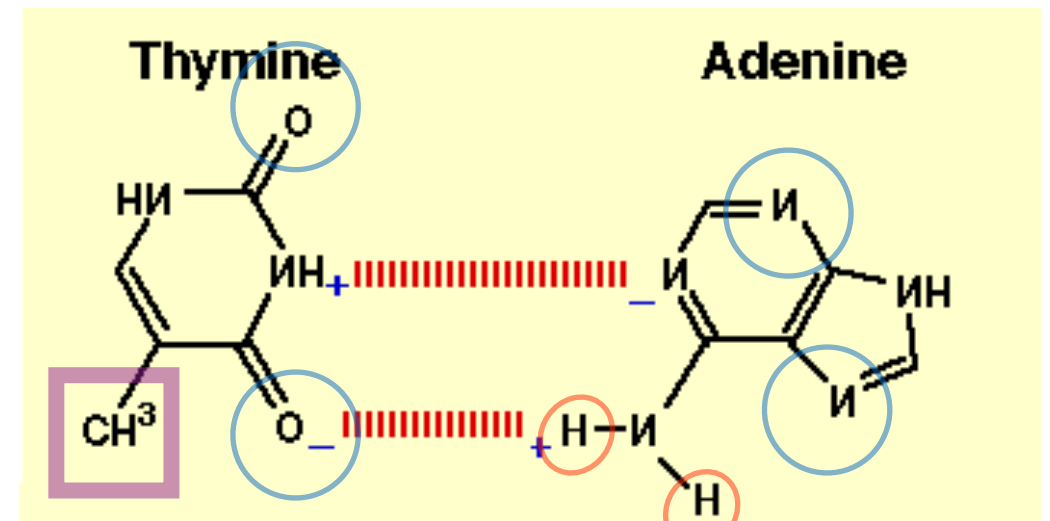
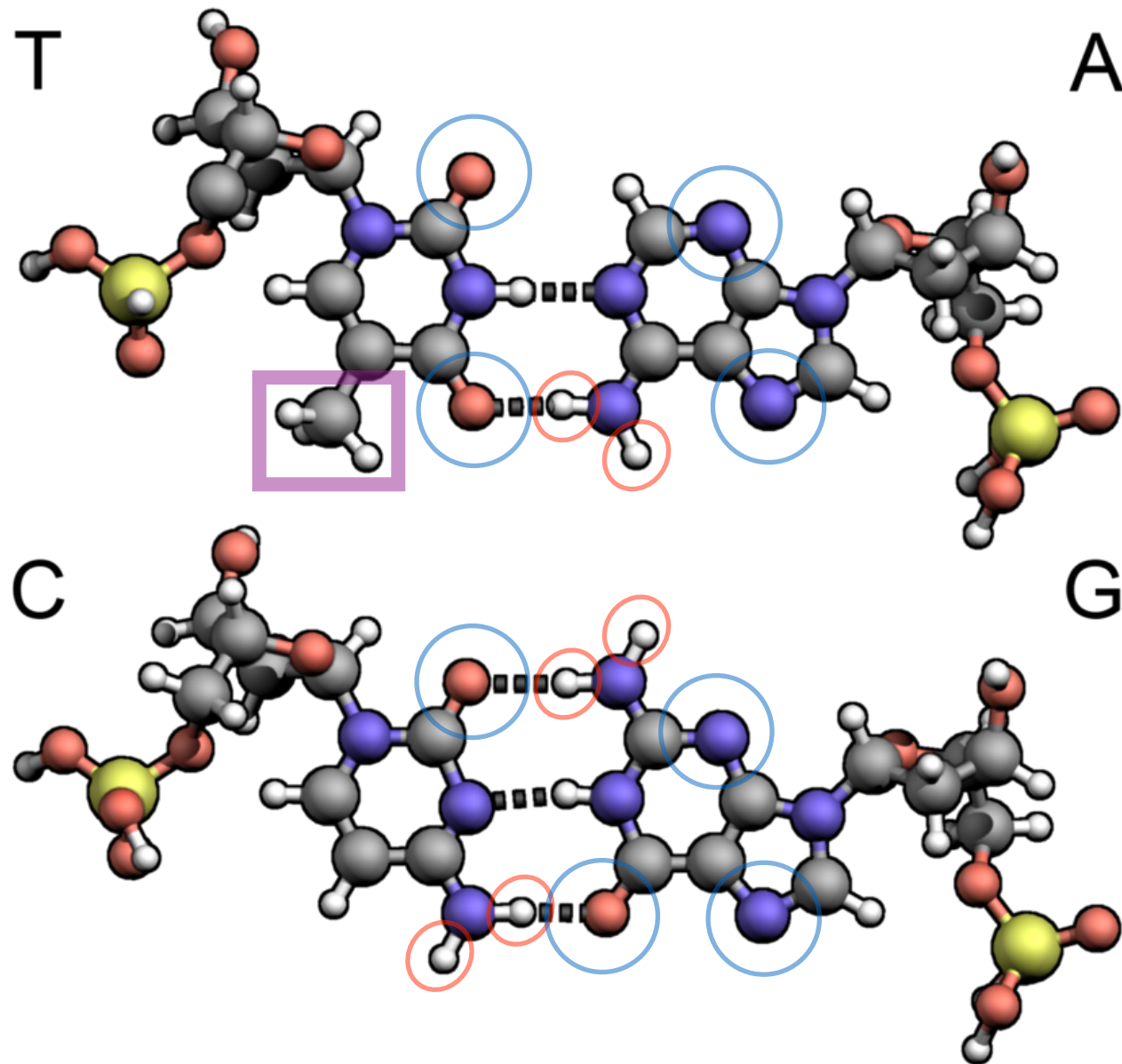
# Hydrogen Bond Donors and Acceptors are exposed in the Major and Minor Grooves



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

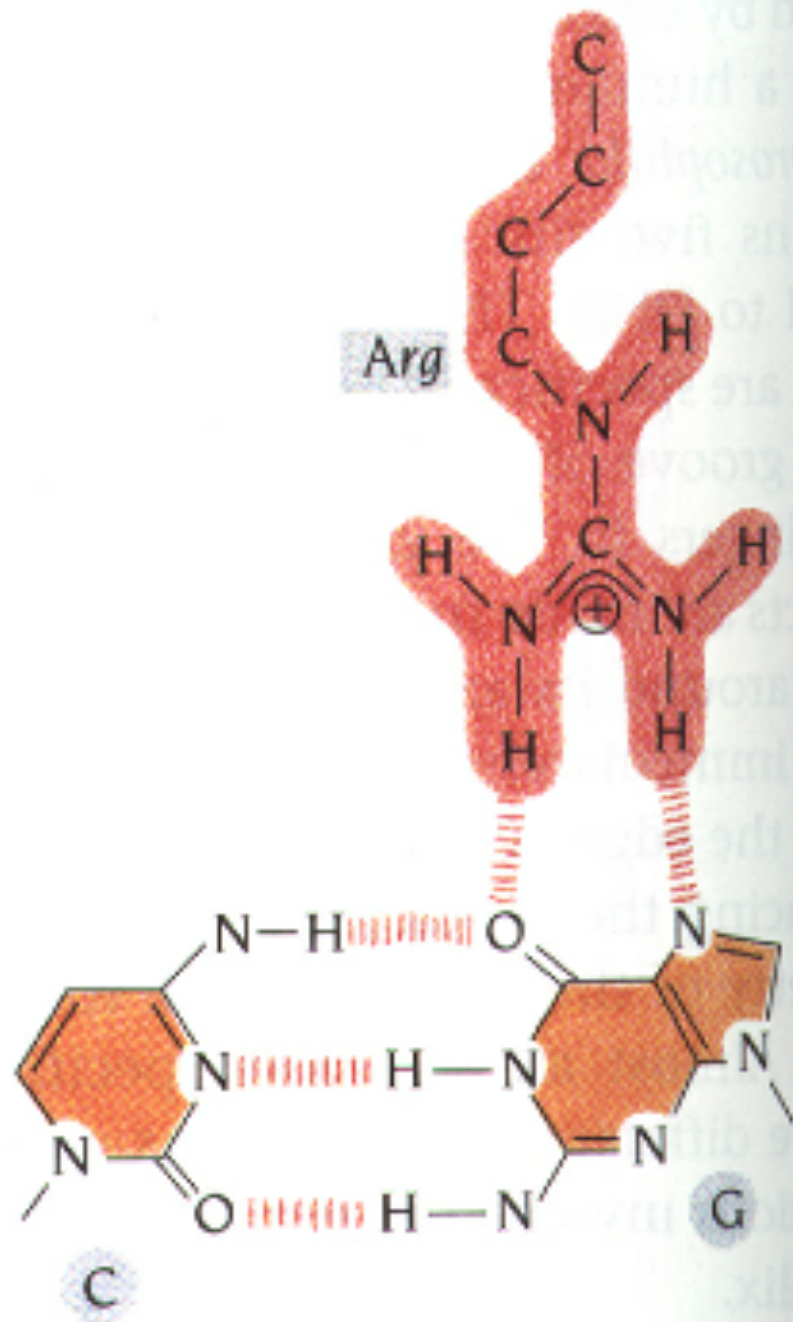


# Thymine's methyl group provides an additional source of recognition/specificity



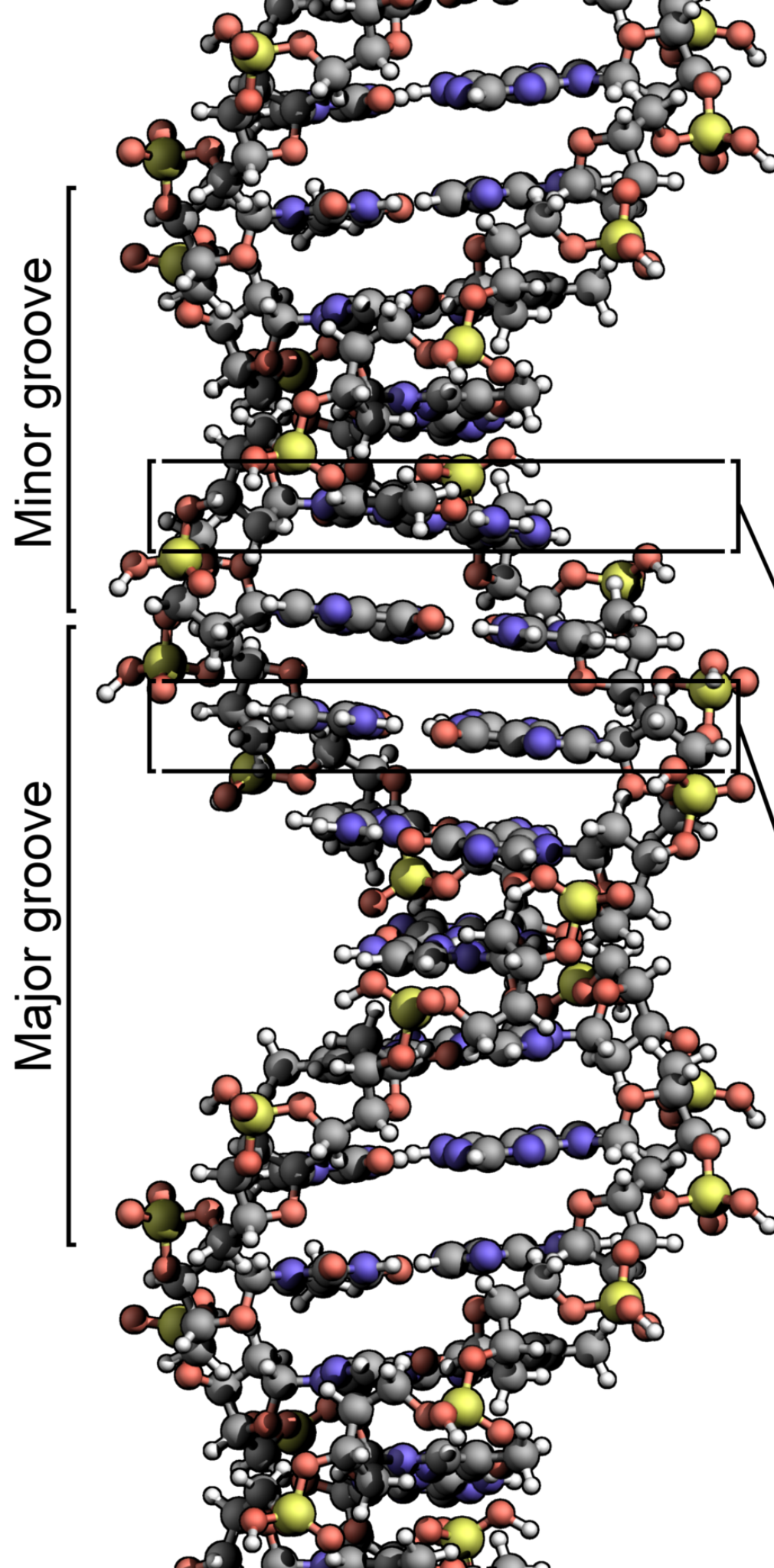


# An example of an Amino acid/DNA base interaction

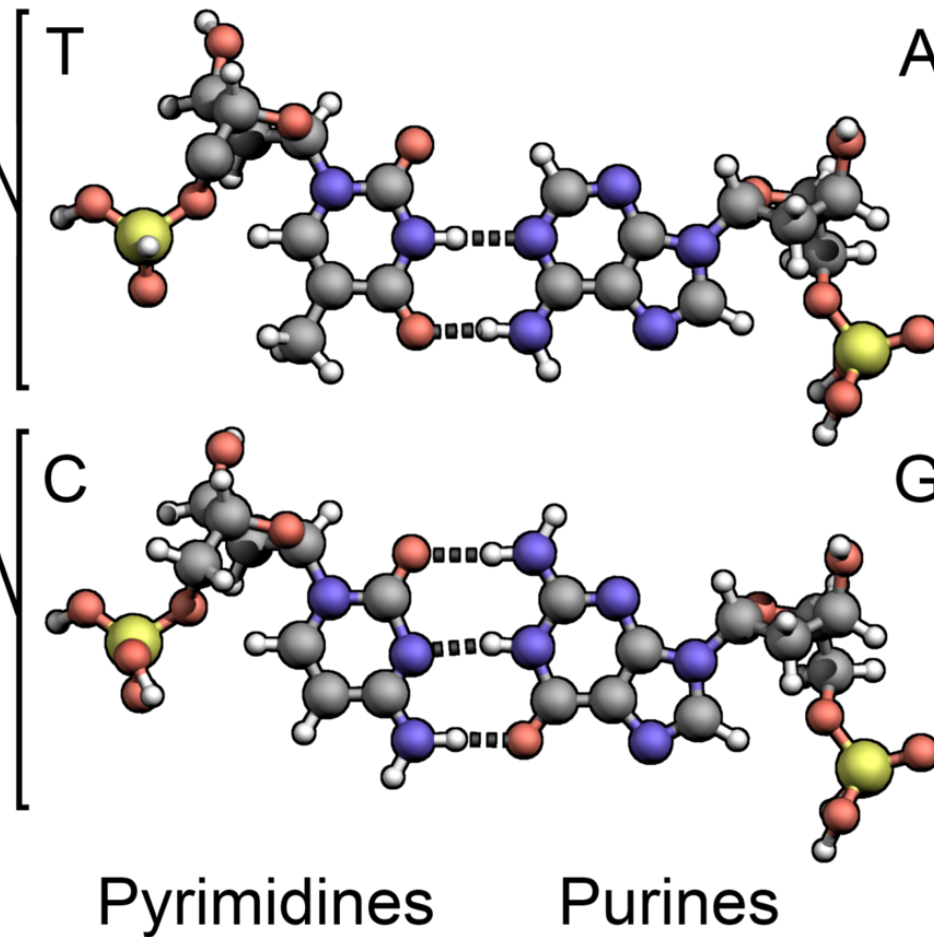


The interaction between arginine with its two hydrogen bond donors and a guanine base with its two acceptors in the major groove is an important component of many protein/DNA interactions.

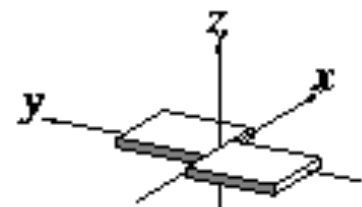
How do proteins interact with specific DNA sequences?



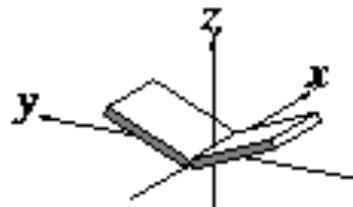
- Hydrogen
- Oxygen
- Nitrogen
- Carbon
- Phosphorus



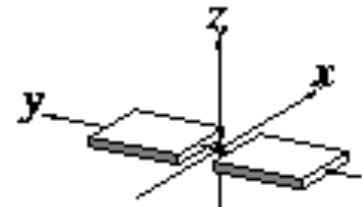
# Base Pair Geometry



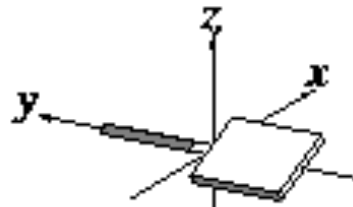
Shear



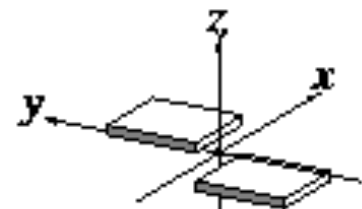
Buckle



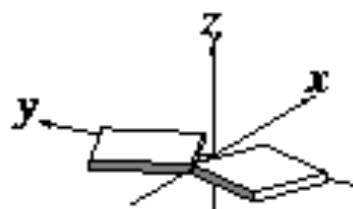
Stretch



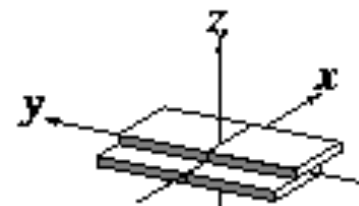
Propeller



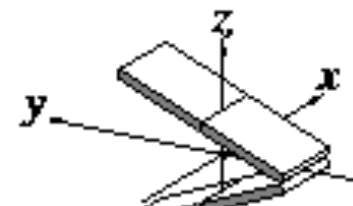
Stagger



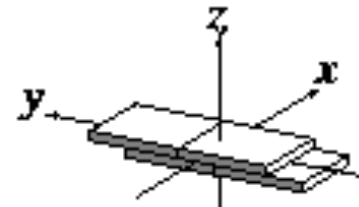
Opening



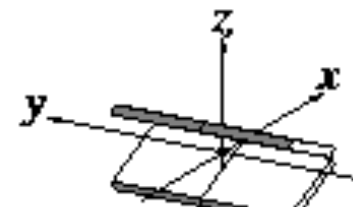
Shift



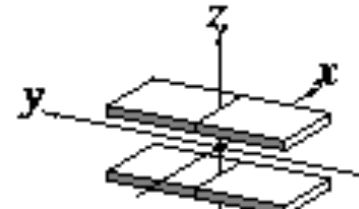
Tilt



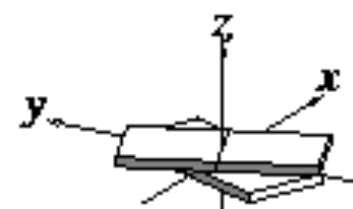
Slide



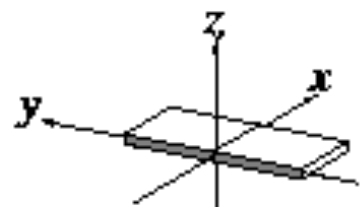
Roll



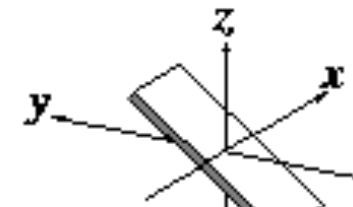
Rise



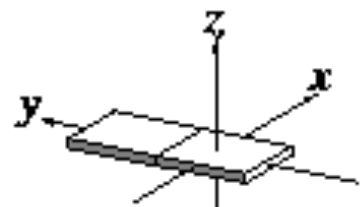
Twist



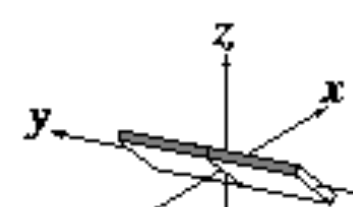
x-displacement



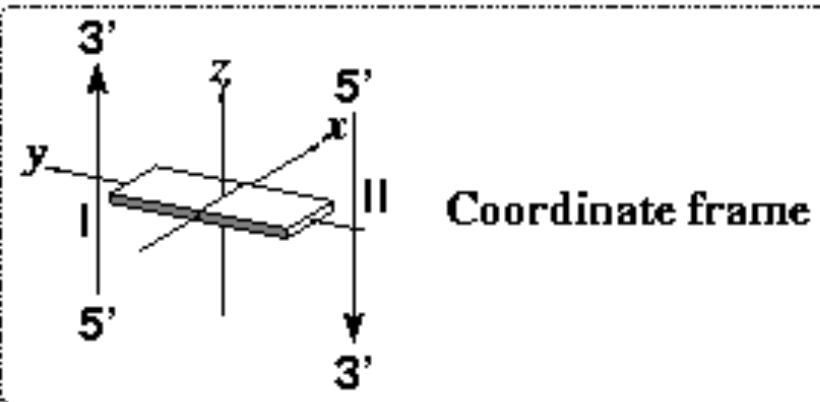
inclination



y-displacement



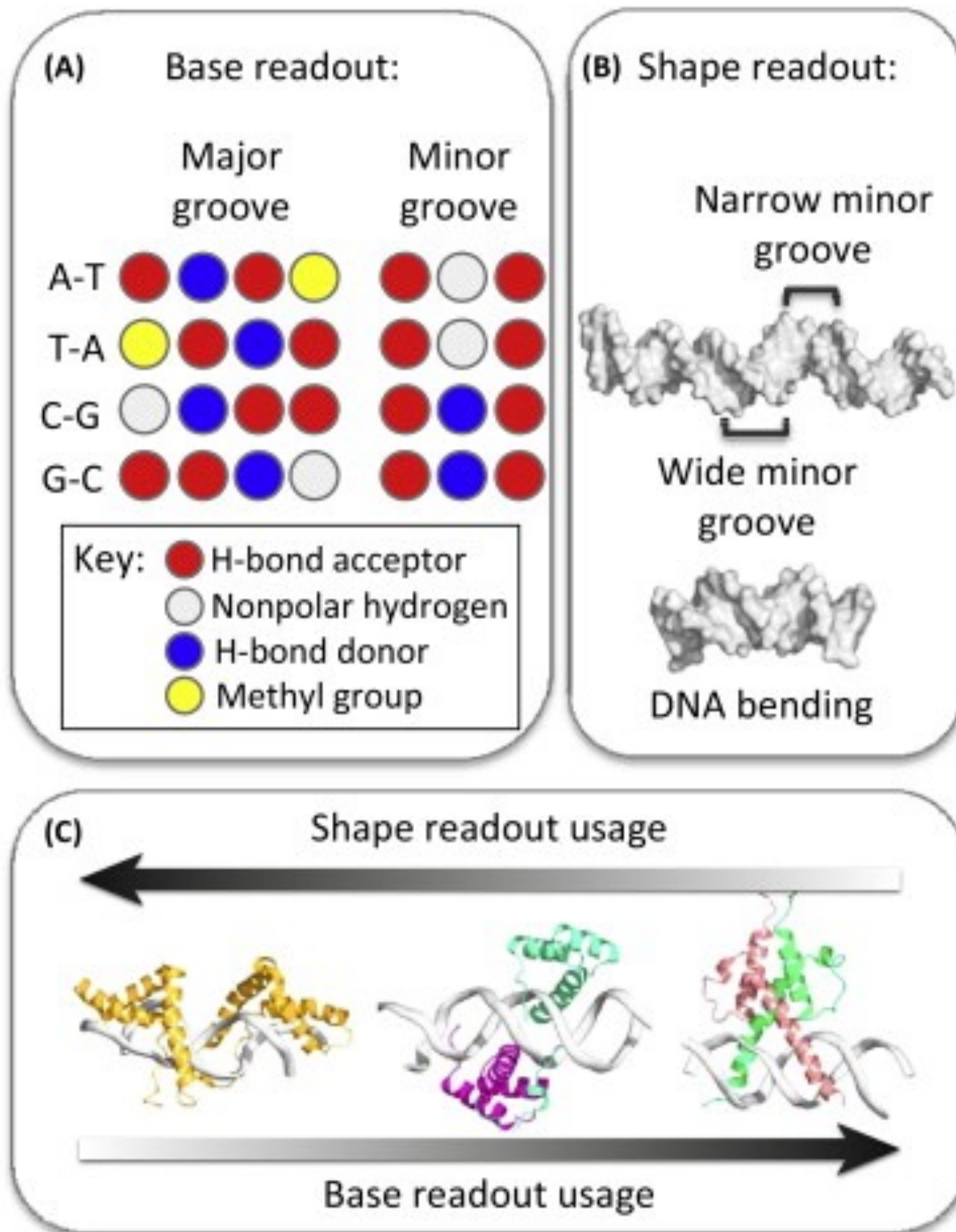
tip



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



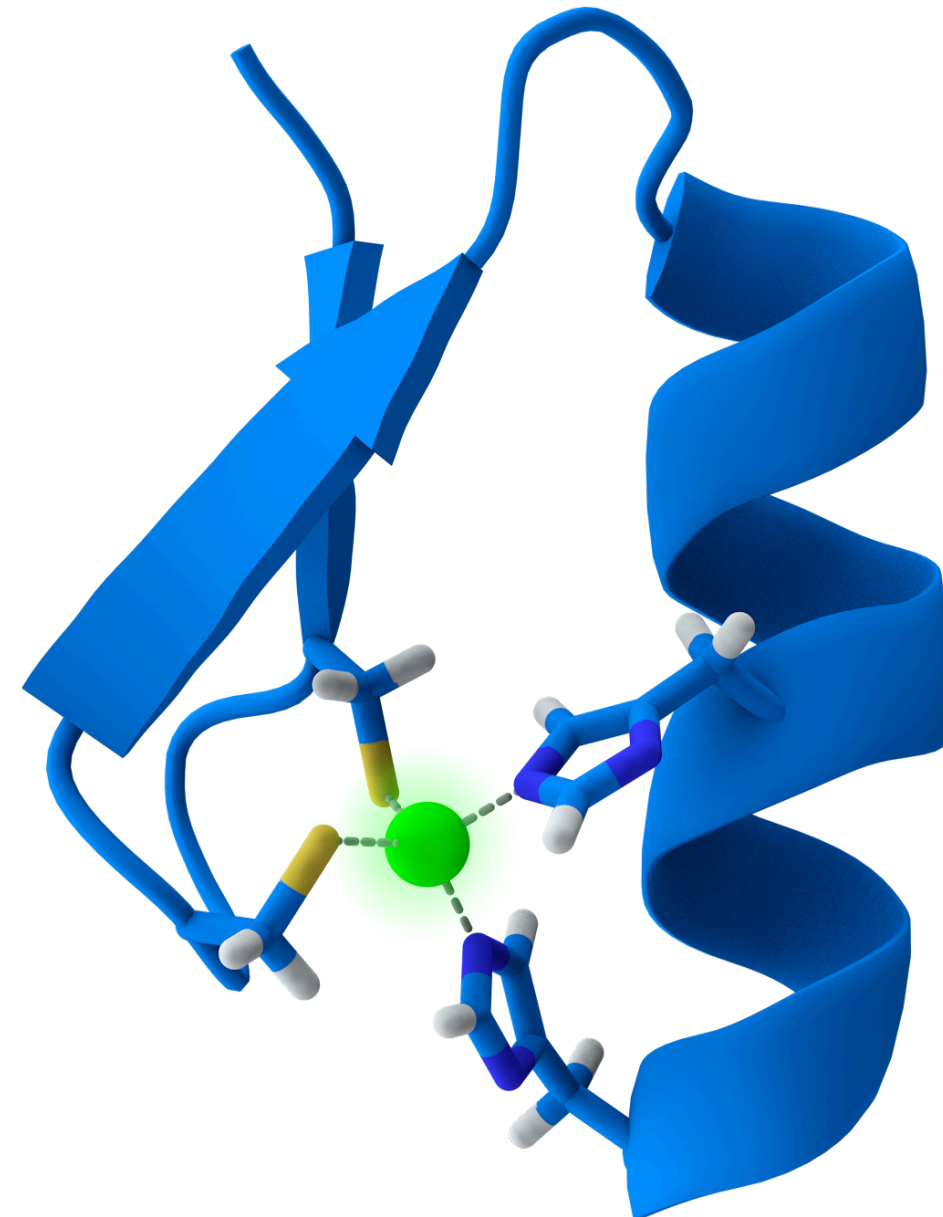
- Base readouts are specific for bp in major groove but degenerate for minor groove.
- Shape dominates for a minor groove-binding high motility group (HMG) box protein
- Base readout is a major contribution in DNA recognition by the bHLH protein Pho4
- Both readout modes are ~equally present in the DNA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

What are the features of protein domains that bind DNA?

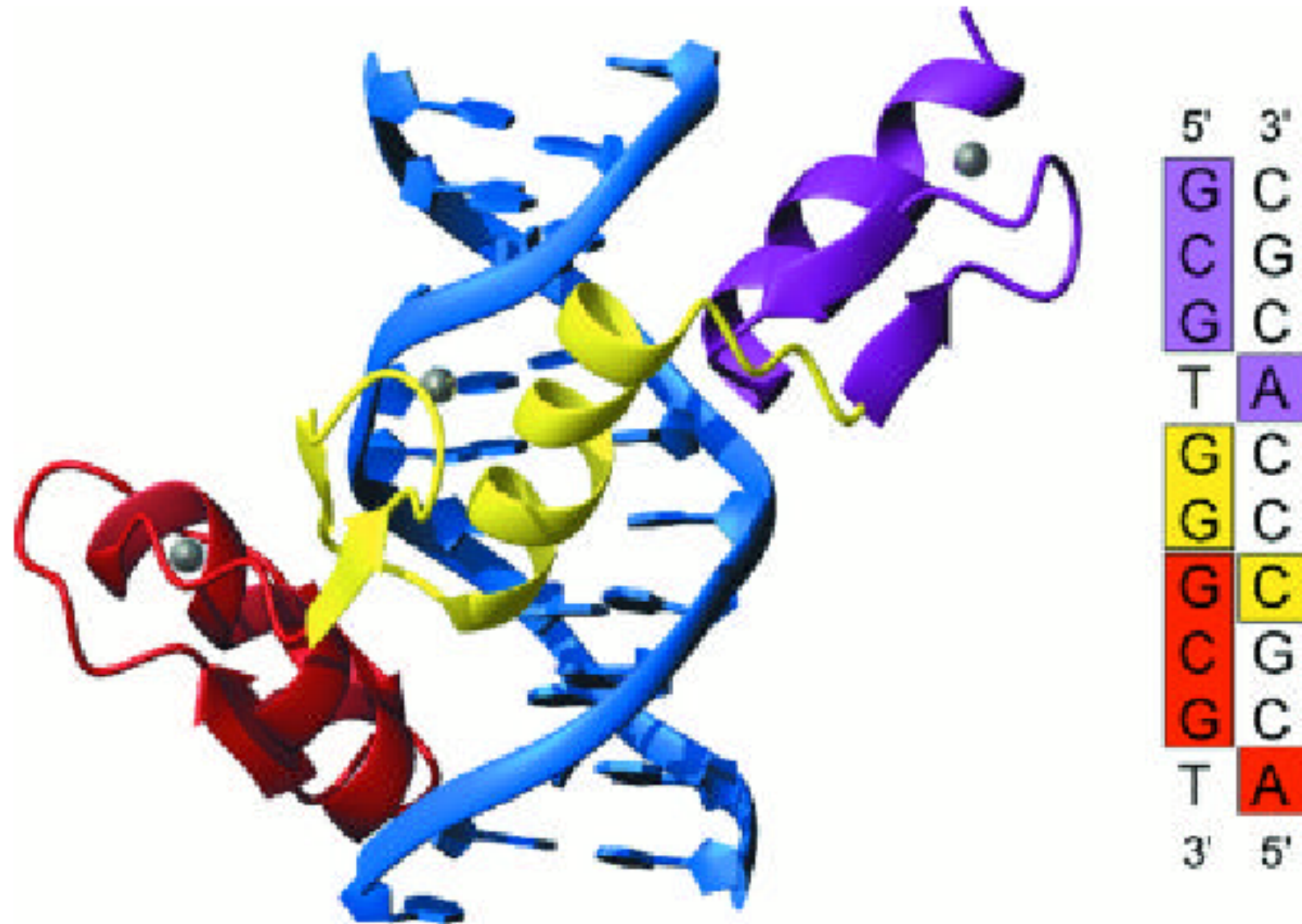
# Zinc-containing DNA binding domains

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

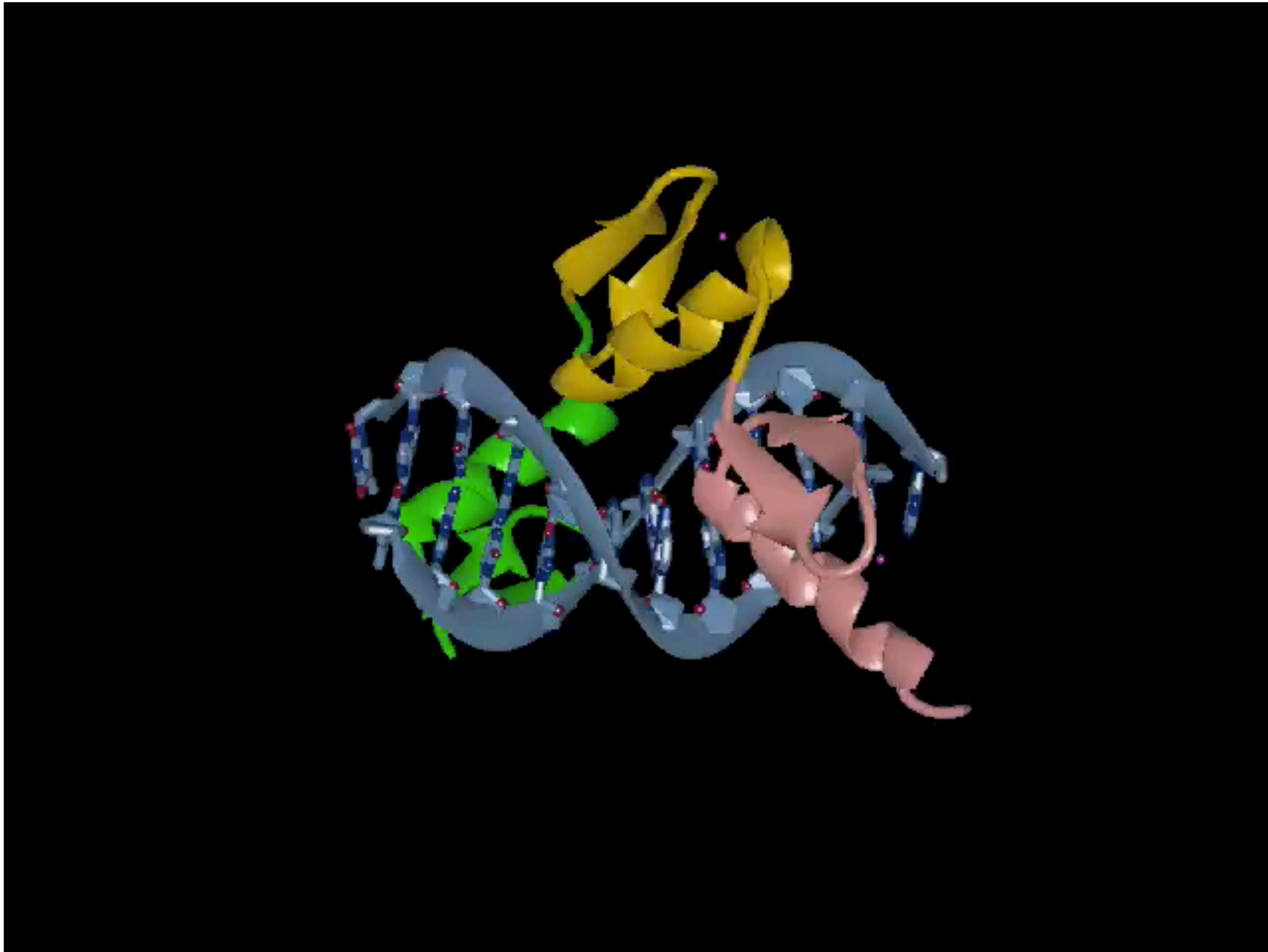
- Cys<sub>2</sub>His<sub>2</sub> Class:
  - Beta-Beta-Alpha fold
  - Cys-X<sub>2-4</sub>-Cys-X<sub>12</sub>-His-X<sub>3-5</sub>-His



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



# Song Tan lab: Zif268

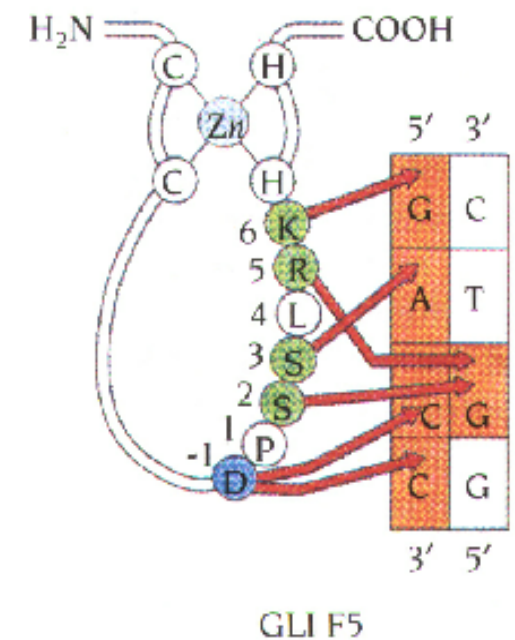
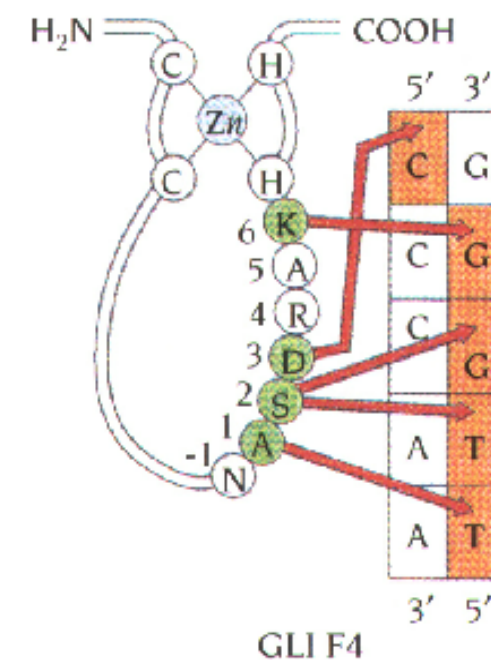
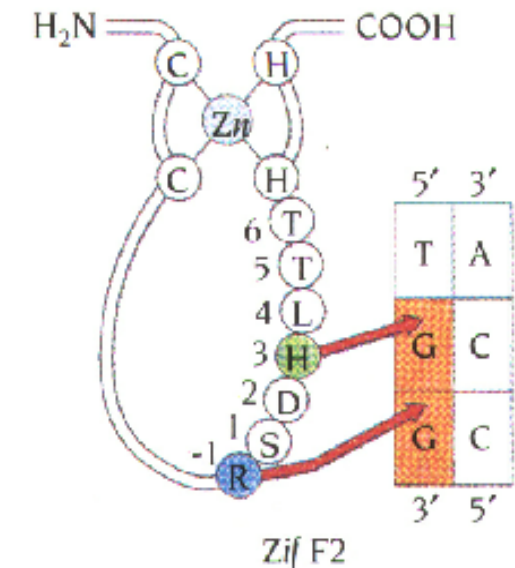
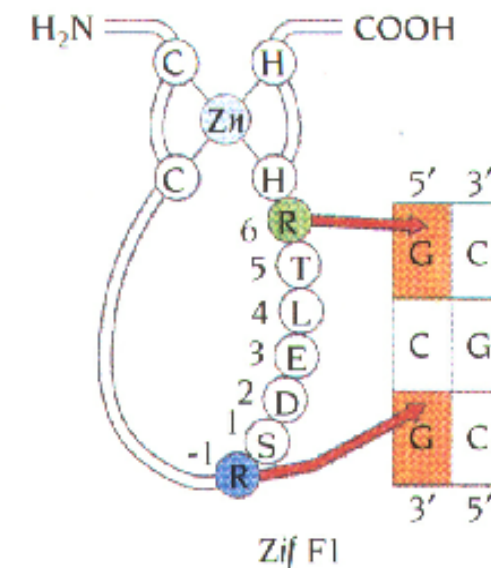
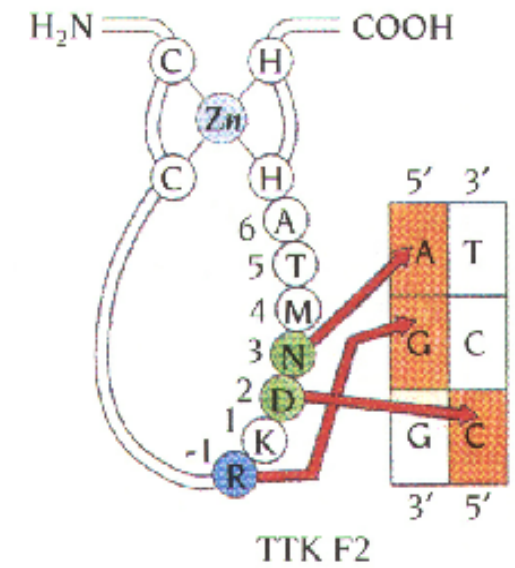
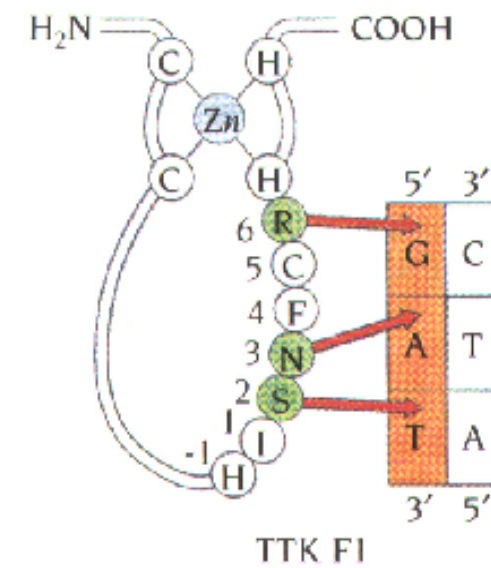


[http://www.personal.psu.edu/sxt30/movies/zif268dna\\_h264.mov](http://www.personal.psu.edu/sxt30/movies/zif268dna_h264.mov)



# Comparing Zn Fingers

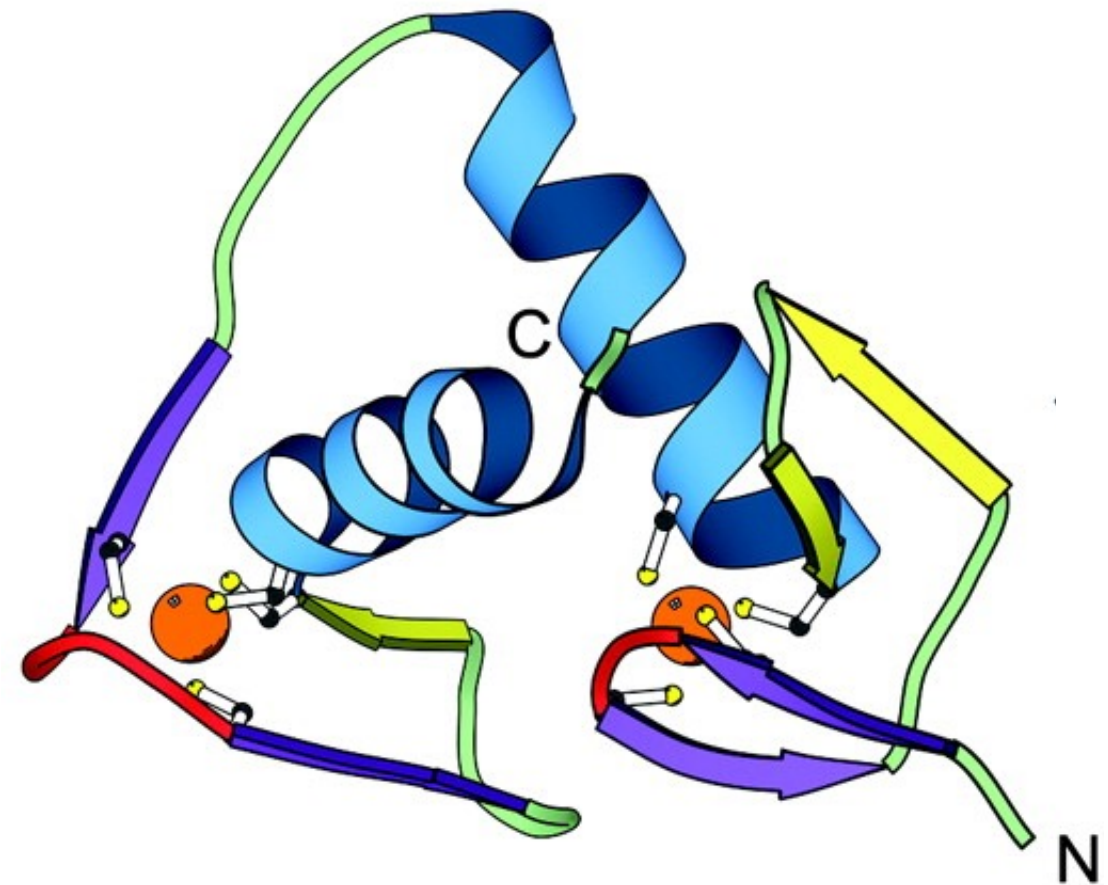
- Variations in a simple motif can provide a dramatic range of DNA sequence recognition.
- Zn Finger nucleases provide for directed mutagenesis. Geurts et al. Knockout rats via embryo microinjection of zinc-finger nucleases. Science. 2009 325: 433.
- Have you ever heard of TALE & TALENs? They were the rage prior to CRISPR.



# Zinc-containing DNA binding domains

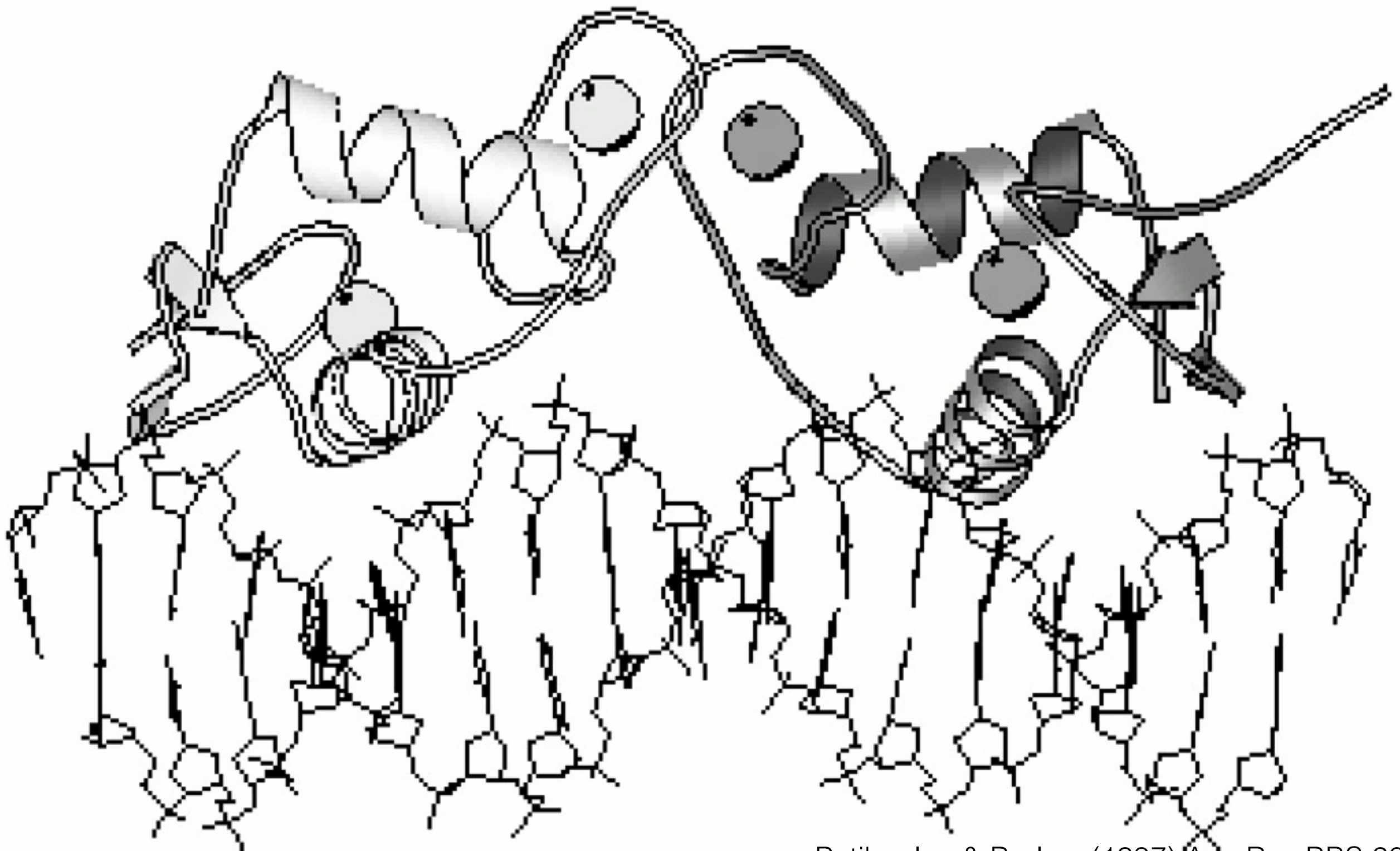
(Zinc is coordinated with a combination of four Cysteine residues)

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



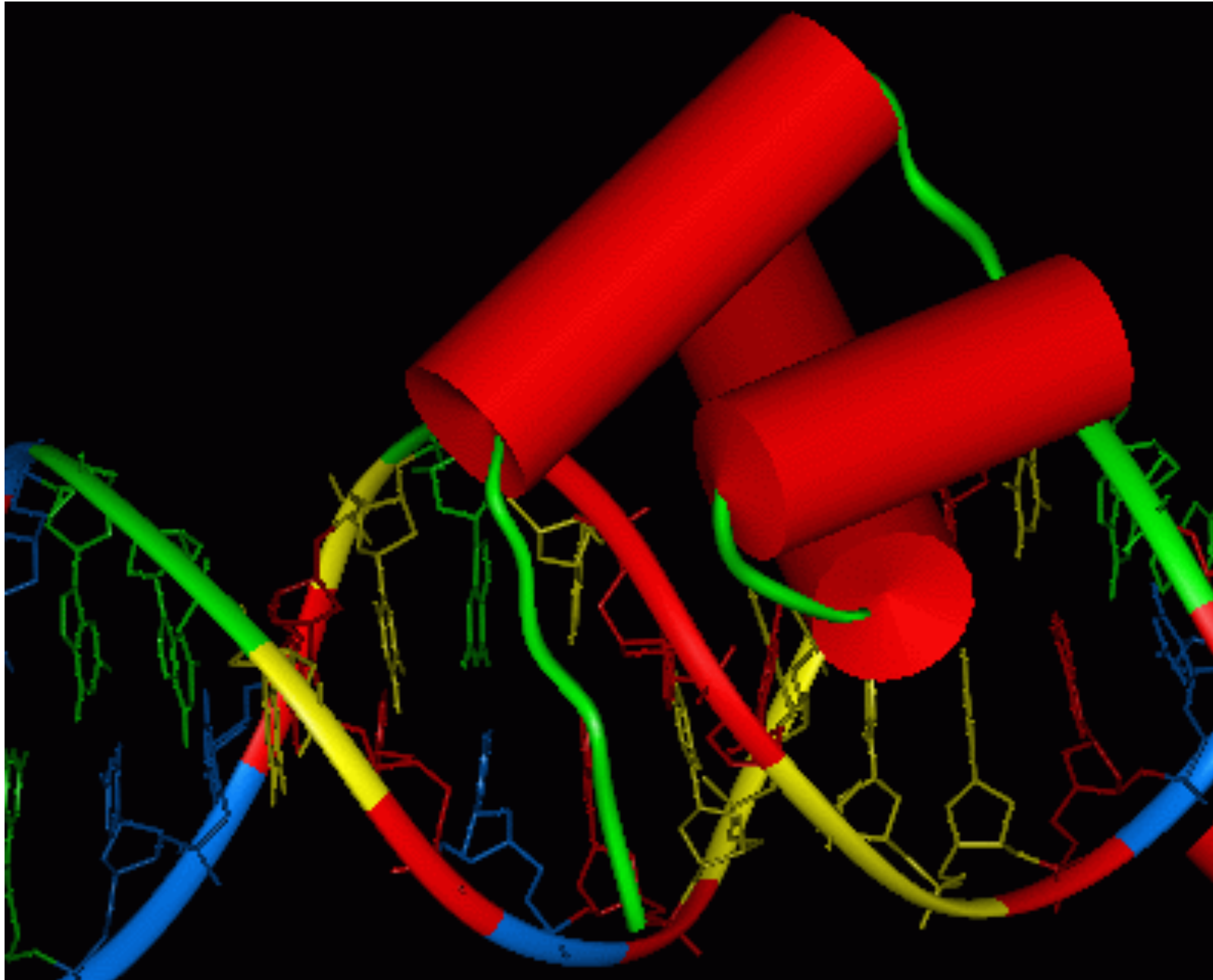
# Estrogen Receptor/DNA Complex

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





# Engrailed Homeodomain/ DNA Complex

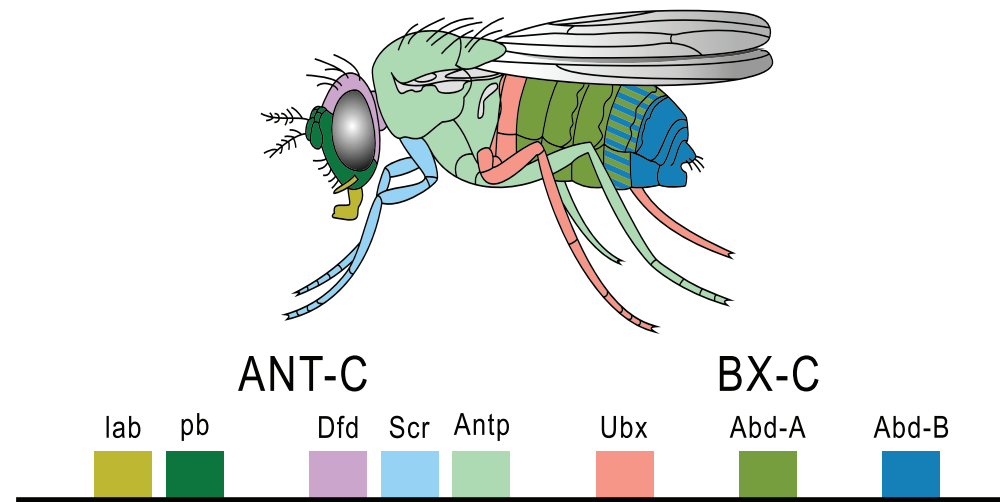
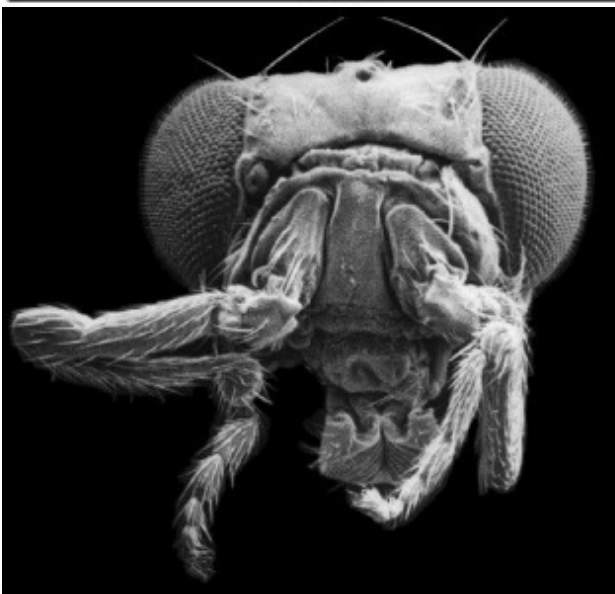
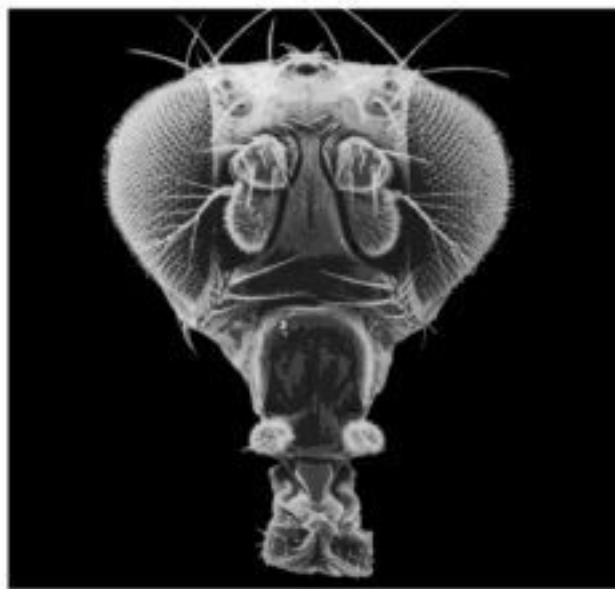


Both major and minor groove interactions by helix 3 and the N-terminal arm respectively

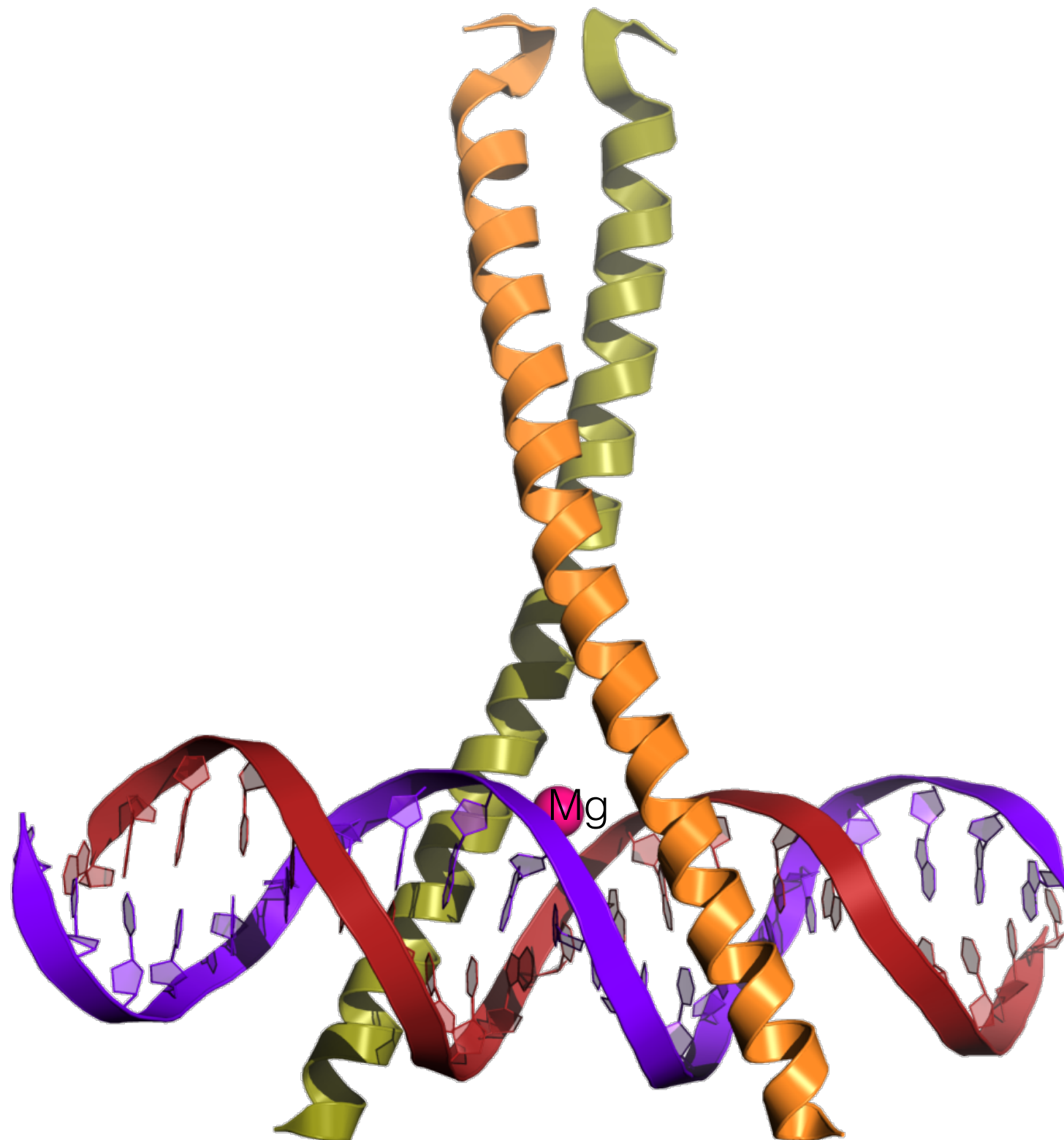
# Song Tan lab: Engrailed



# Engrailed Homeodomain (Hox genes also have homeodomains)



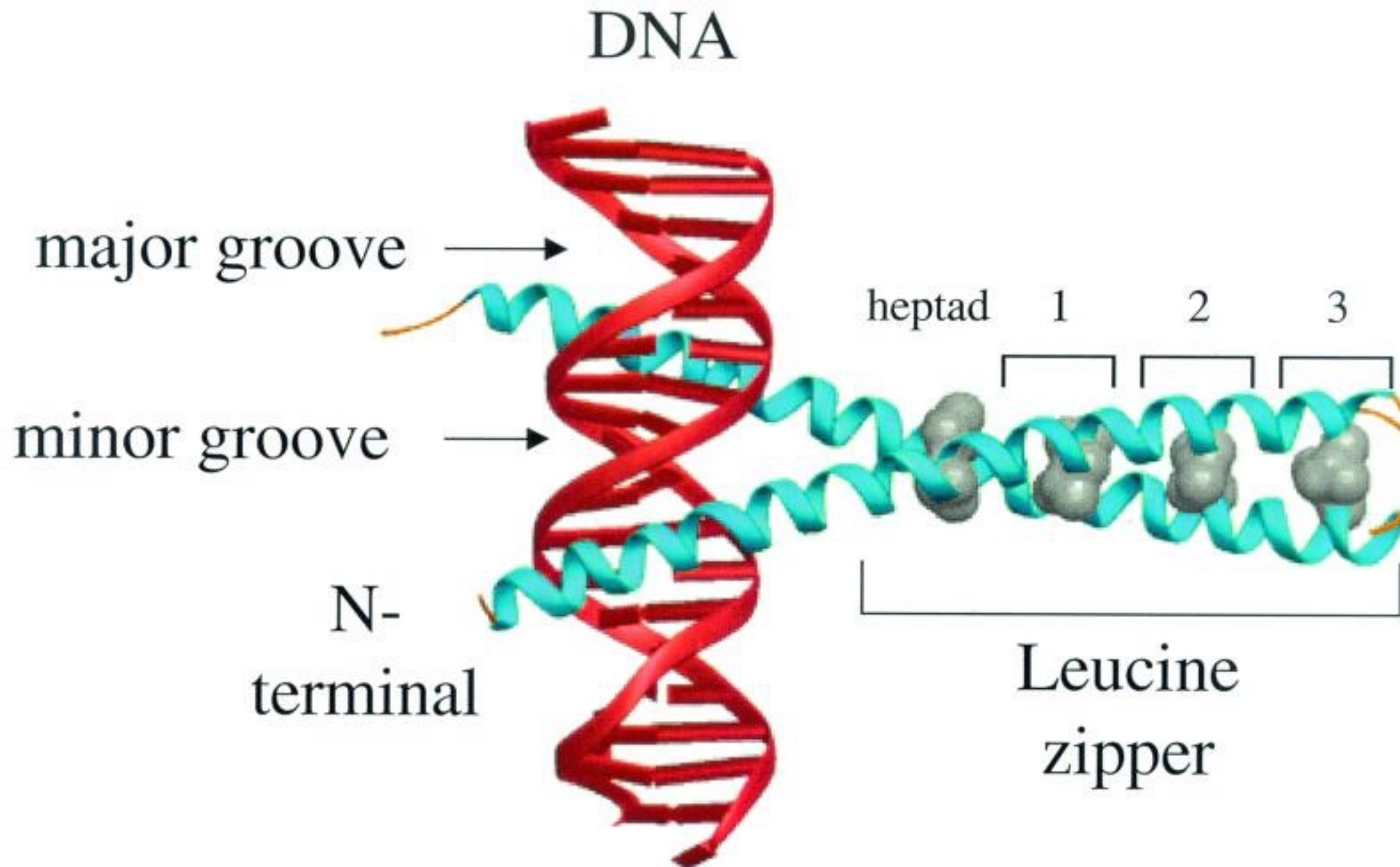
# Leucine Zippers: dimerization of TFs



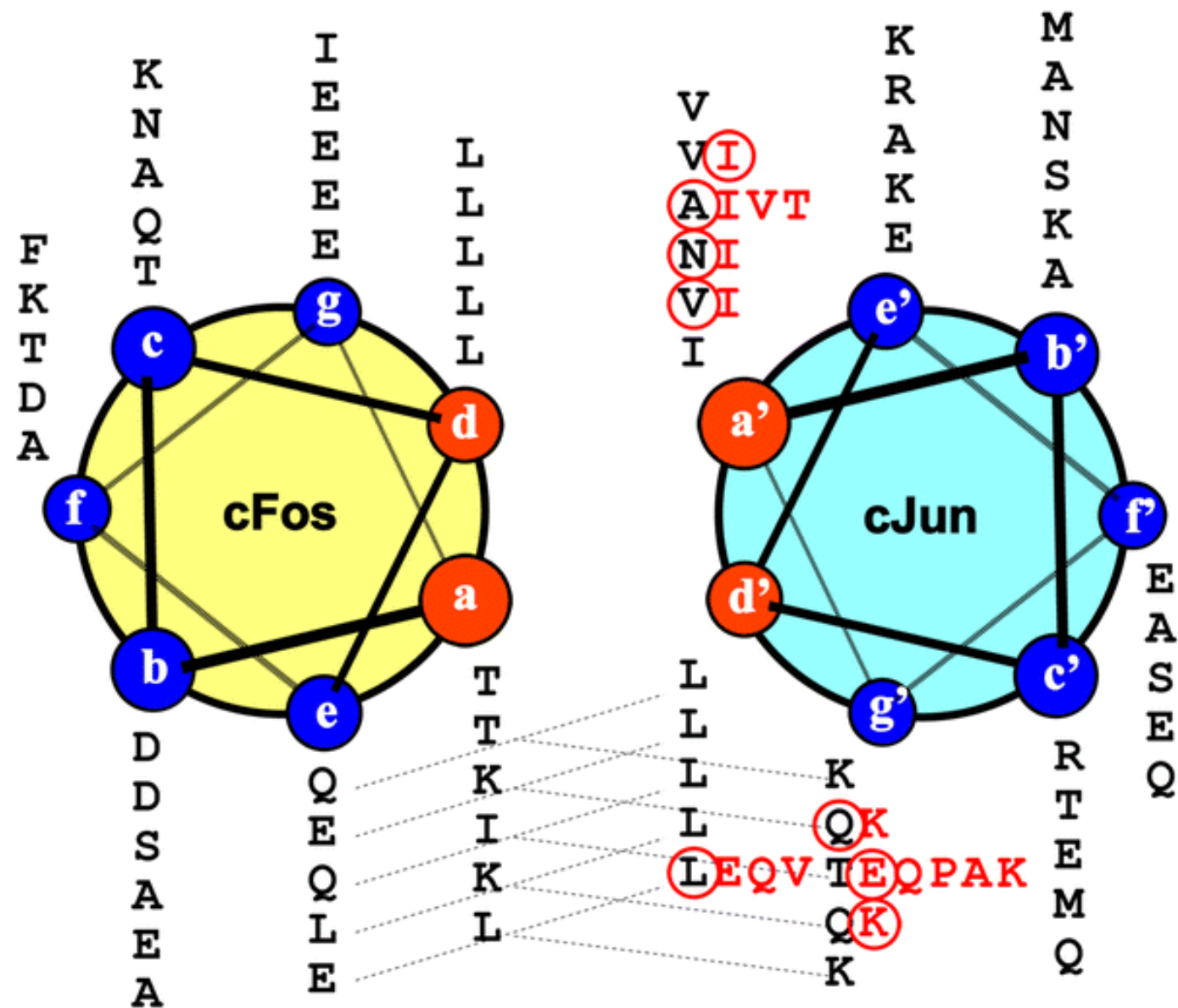
ATF1



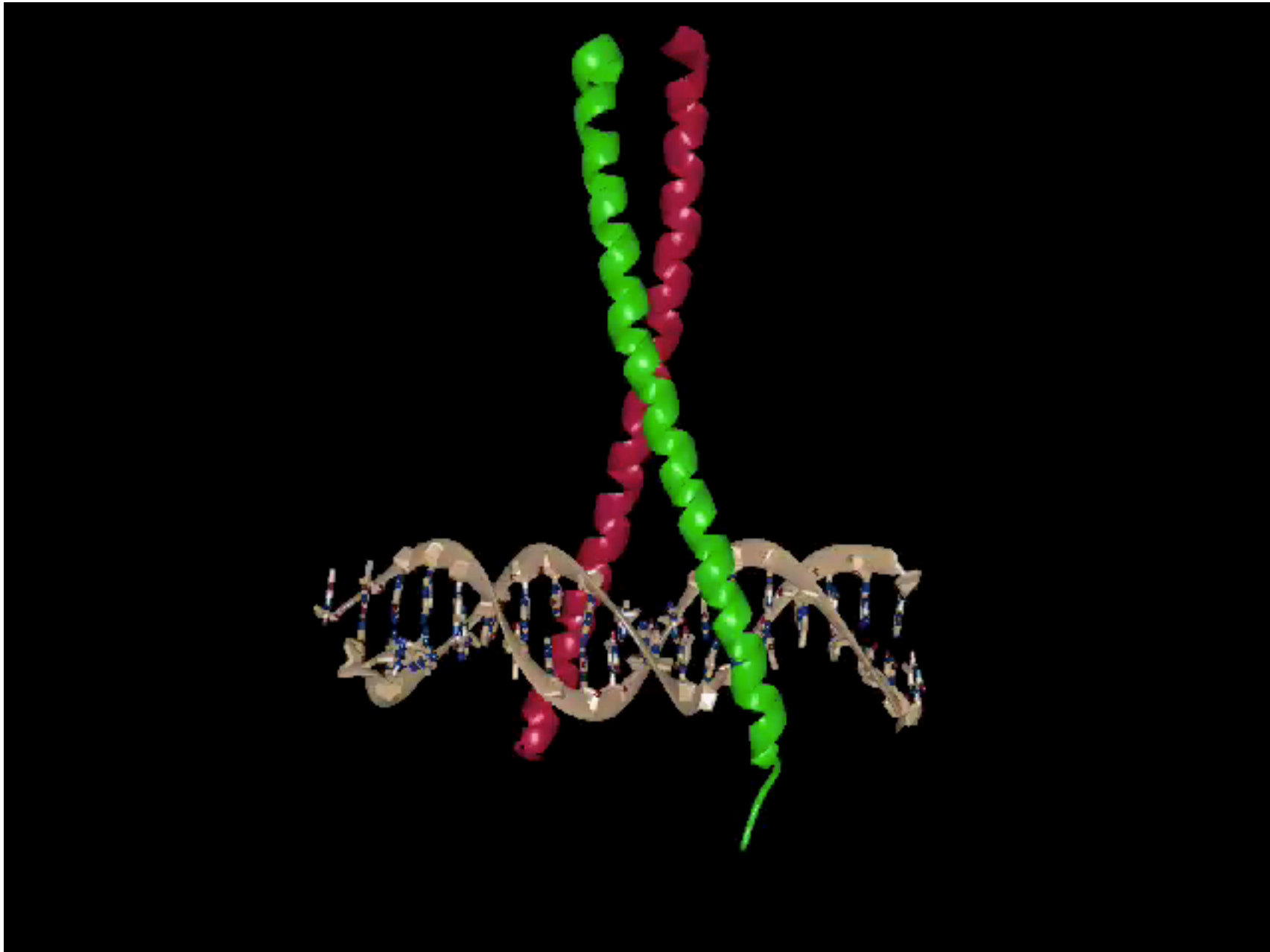
# Leucine Zippers: dimerization of TFs



# Leucine Zippers: dimerization of TFs



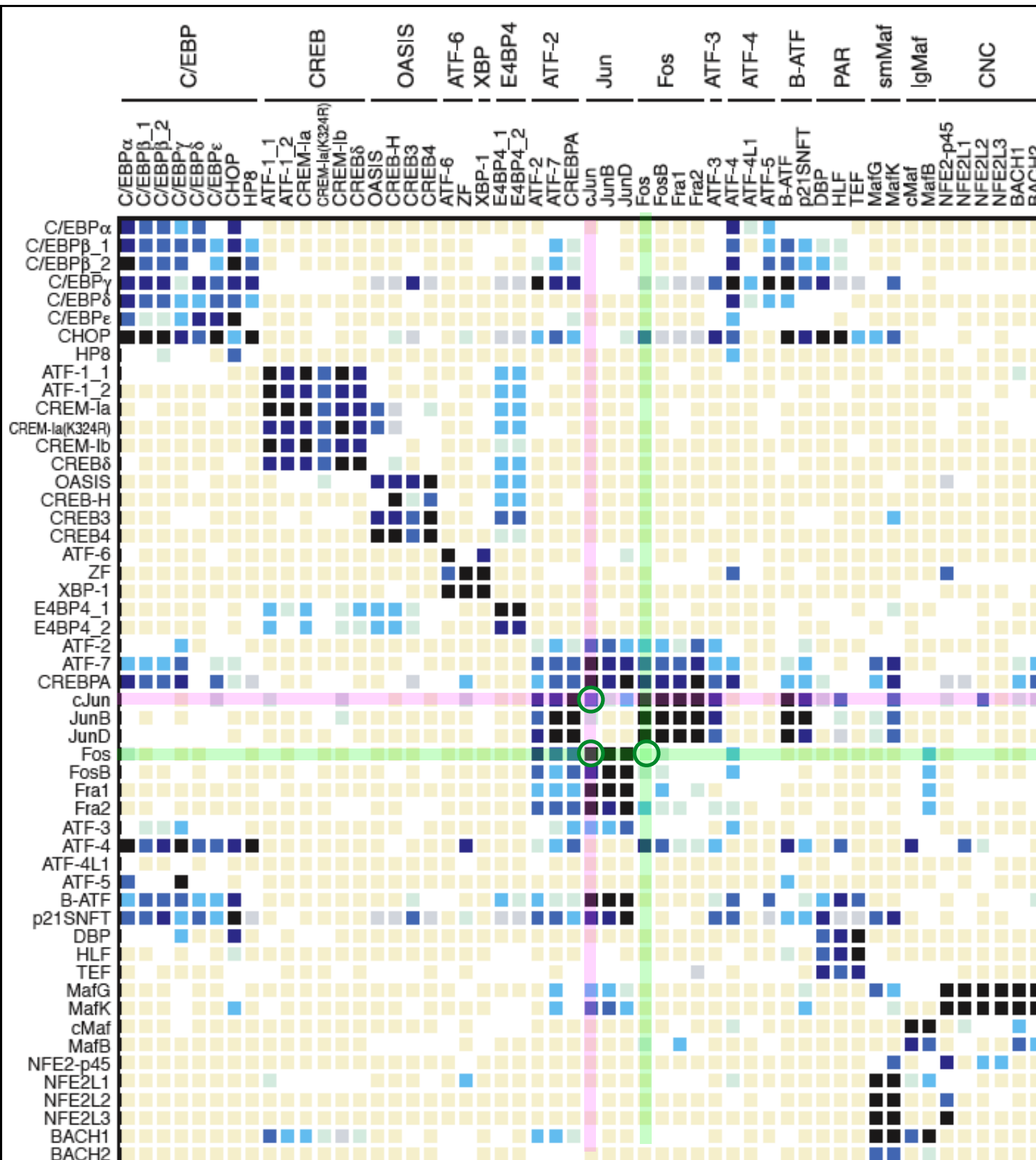
# Song Tan Lab: bZIP





# Have you heard of the transcription factor AP1?

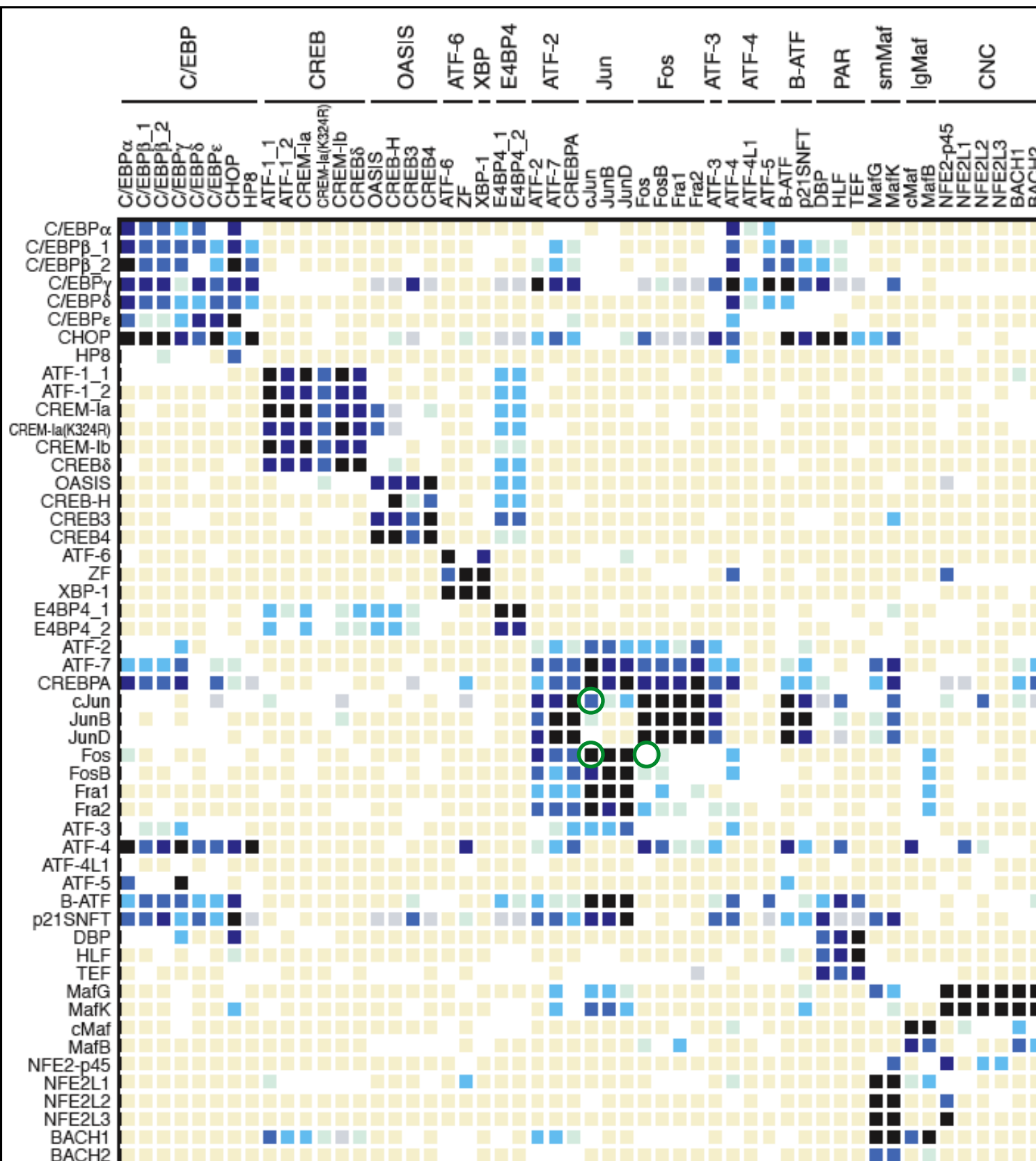
## Interaction Matrix for 49 human bZIP Peptides



~14% of possible interactions detected.

Most between family members, but 136 between families.

# Leucine Zippers Provide Specific Dimerization Interactions

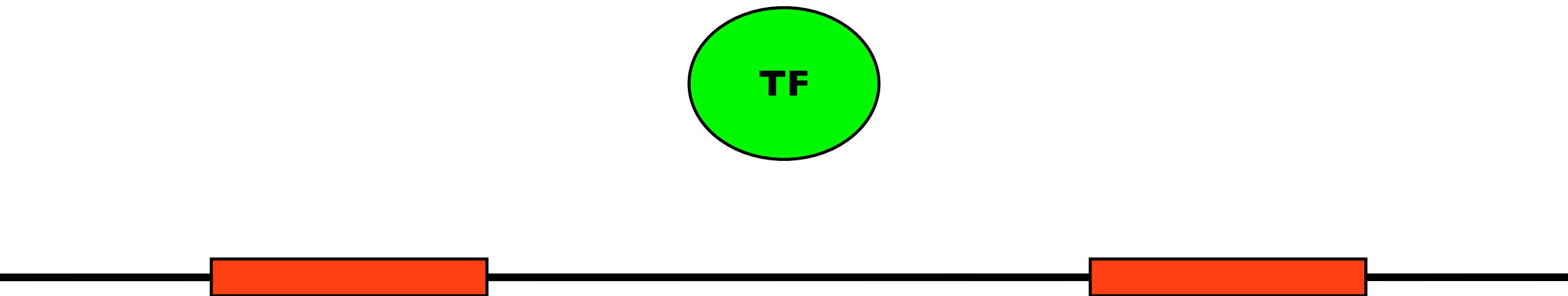


cJun/cFos heterodimer forms preferentially relative to homodimers

cJun/cJun forms but has two unfavorable charge interactions.

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

# Chromatin affects TF binding in vivo



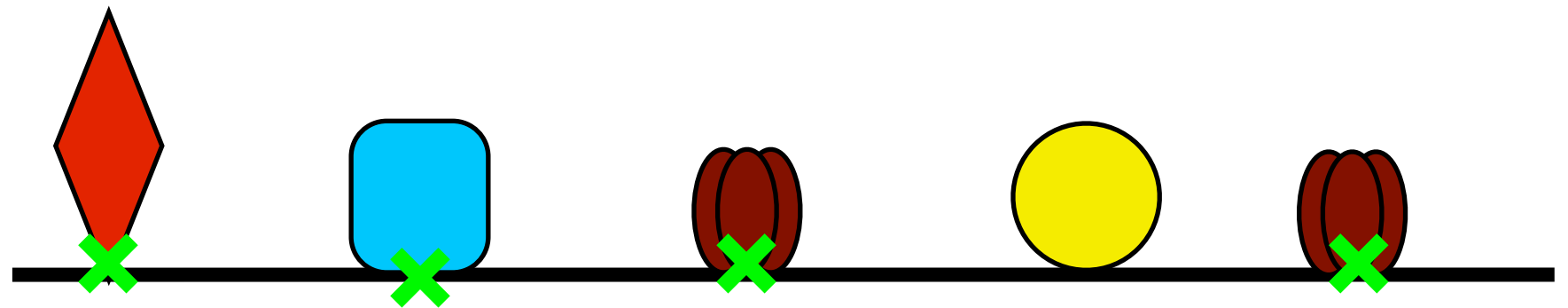
-Besides sequence, what influences TF binding?

# HSF binds many sites after HS



# Chromatin Immunoprecipitation (ChIP)

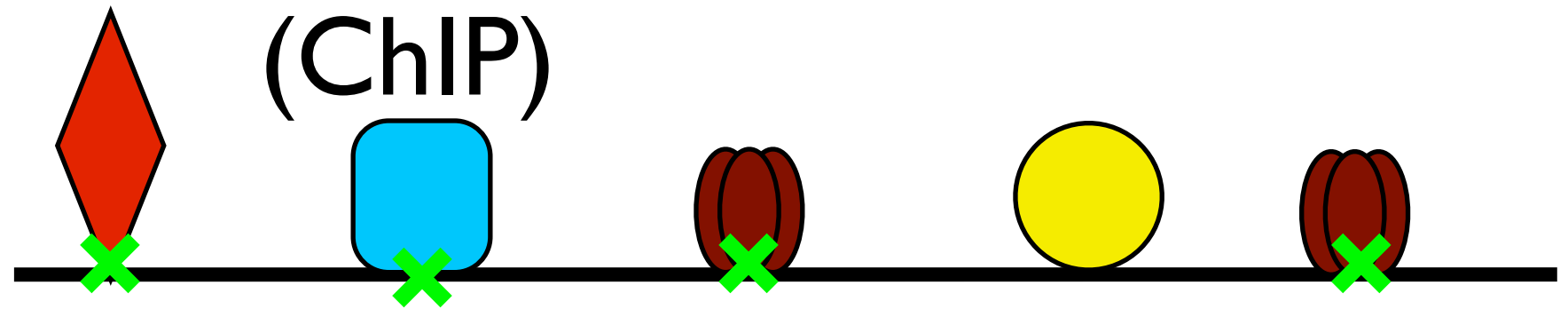
Crosslink DNA and  
Proteins



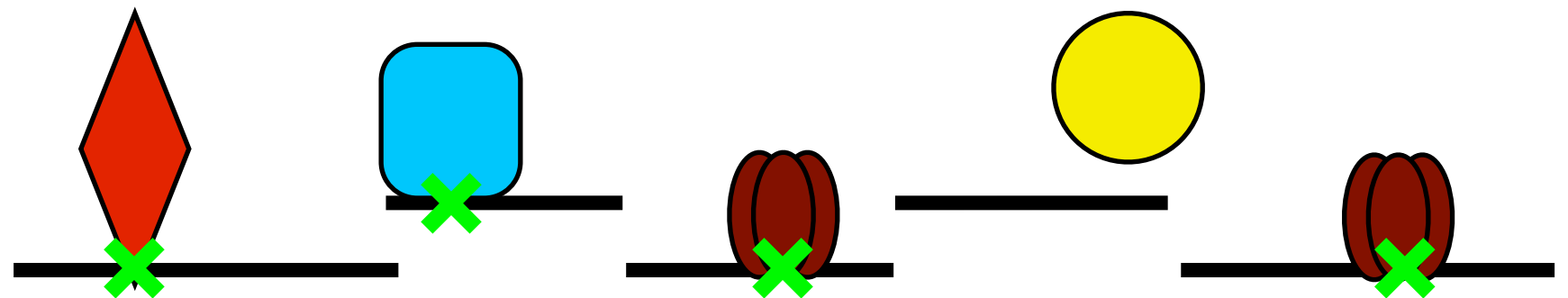
# Chromatin Immunoprecipitation

(ChIP)

Crosslink DNA and  
Proteins



Shear DNA

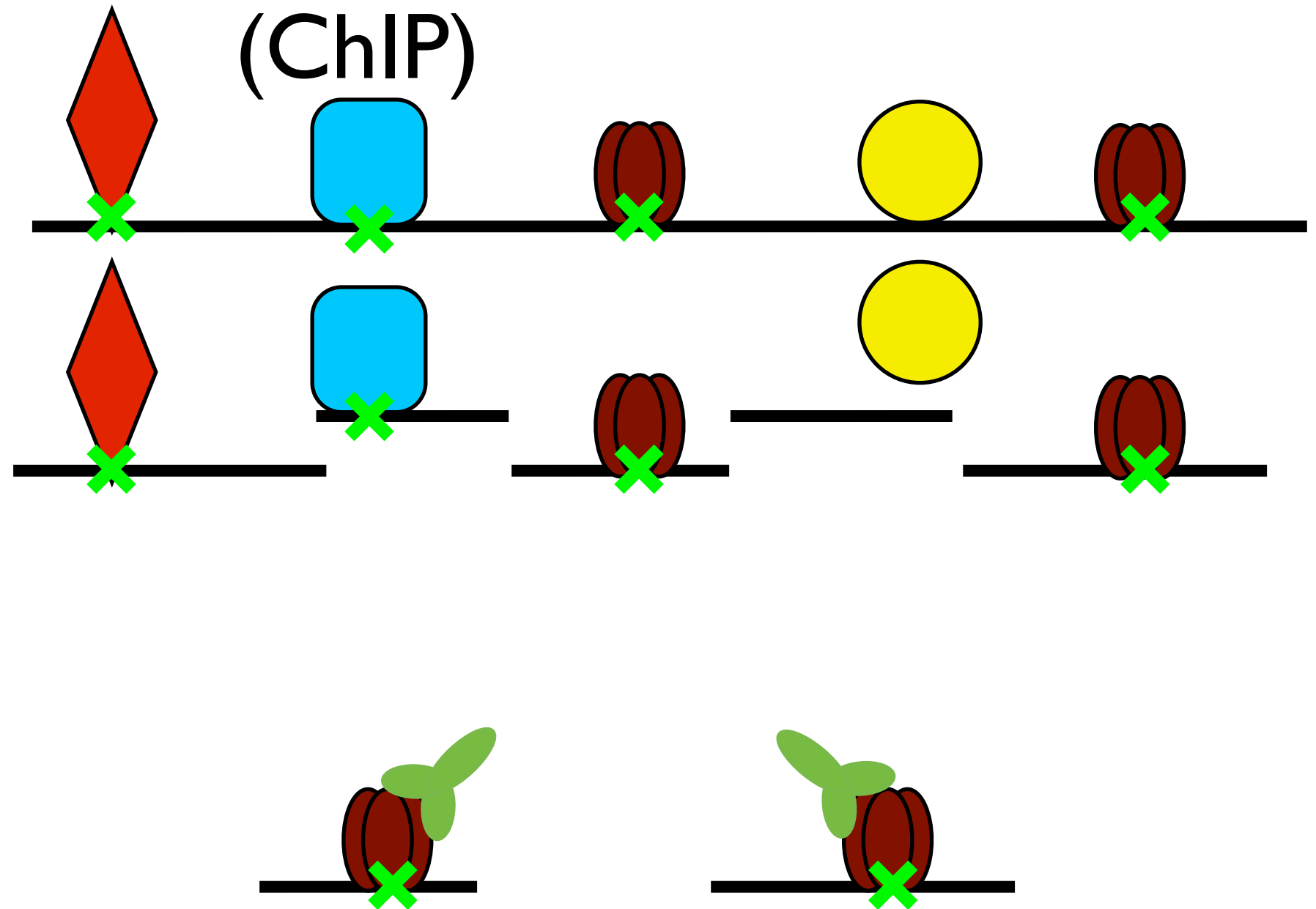


# Chromatin Immunoprecipitation (ChIP)

Crosslink DNA and  
Proteins

Shear DNA

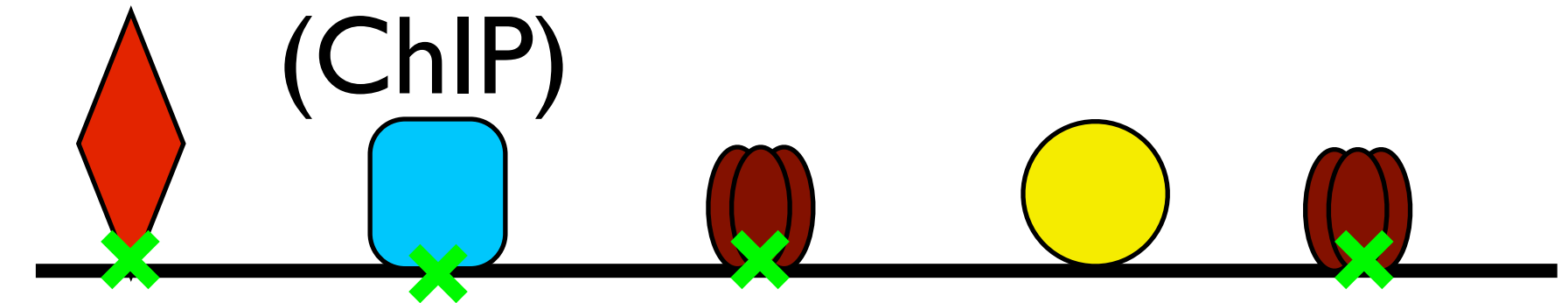
Immunoprecipitate



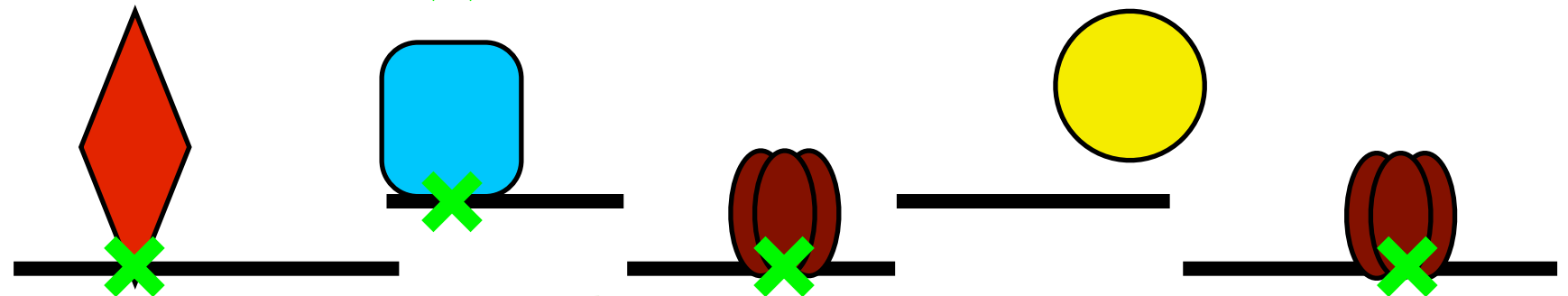


# Chromatin Immunoprecipitation (ChIP)

Crosslink DNA and  
Proteins



Shear DNA



Immunoprecipitate



Purify DNA



# Chromatin Immunoprecipitation

(ChIP-seq)

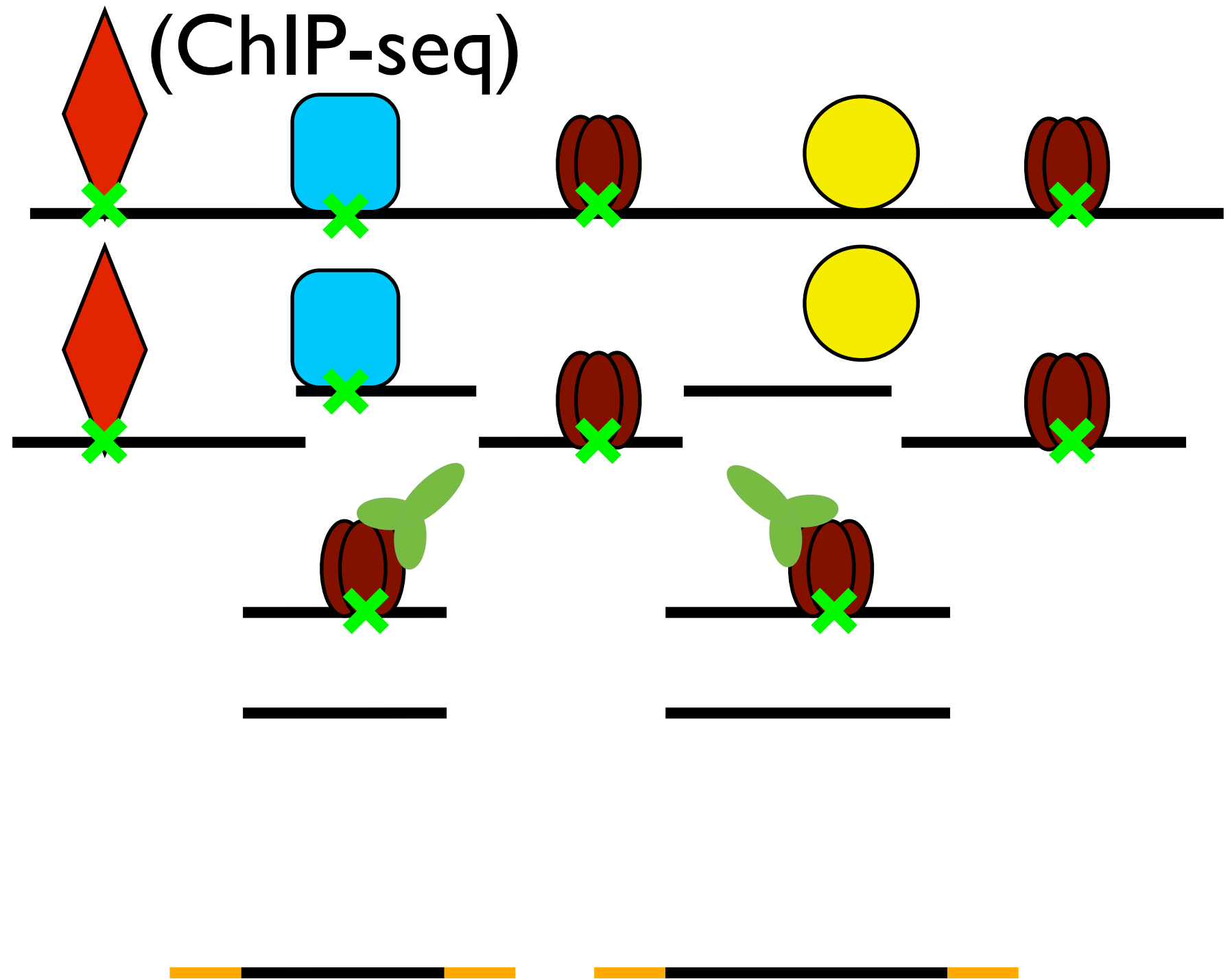
Crosslink DNA and  
Proteins

Shear DNA

Immunoprecipitate

Purify DNA

Ligate adapters



# Chromatin Immunoprecipitation

(ChIP-seq)

Crosslink DNA and  
Proteins

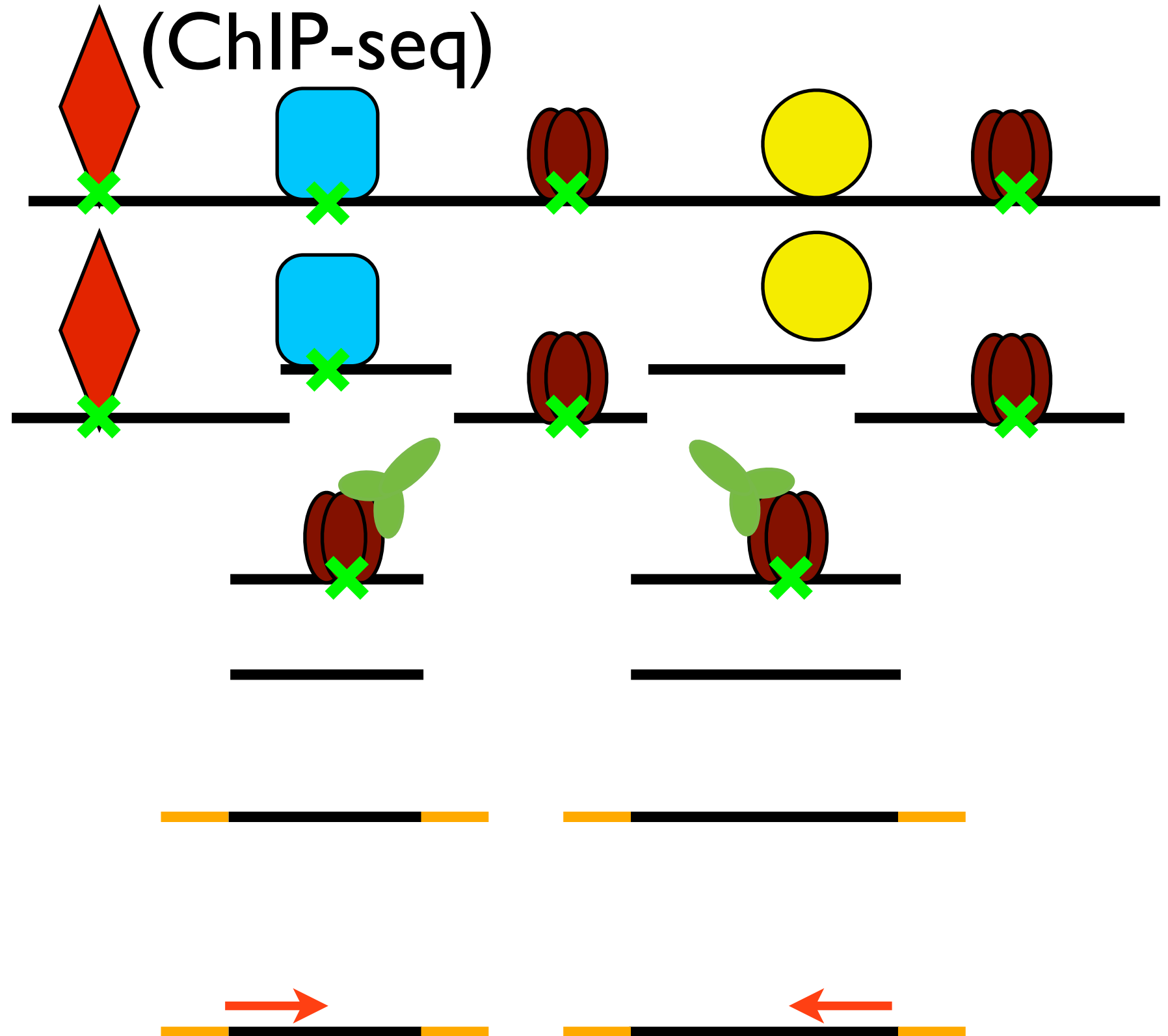
Shear DNA

Immunoprecipitate

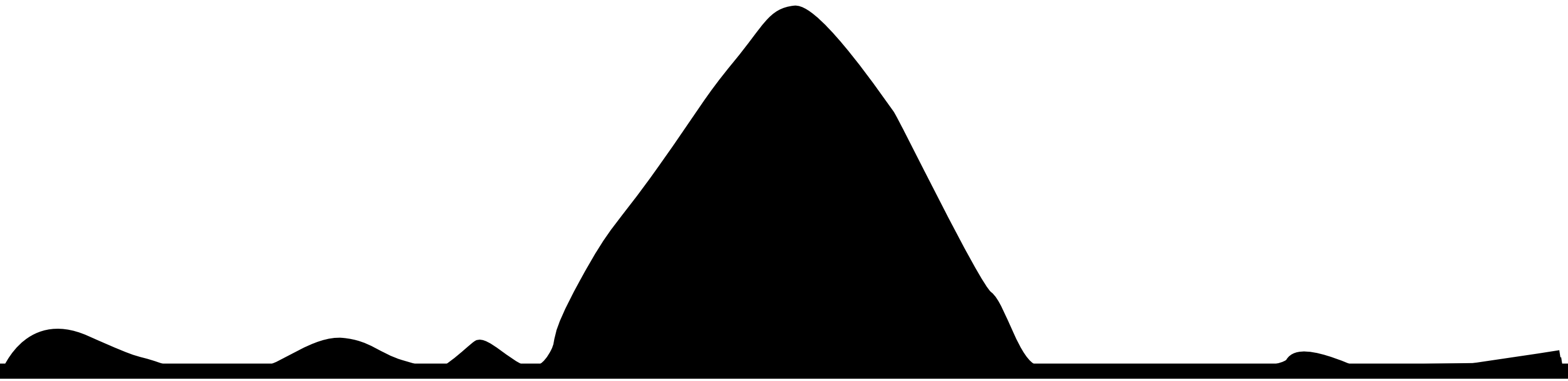
Purify DNA

Ligate adapters

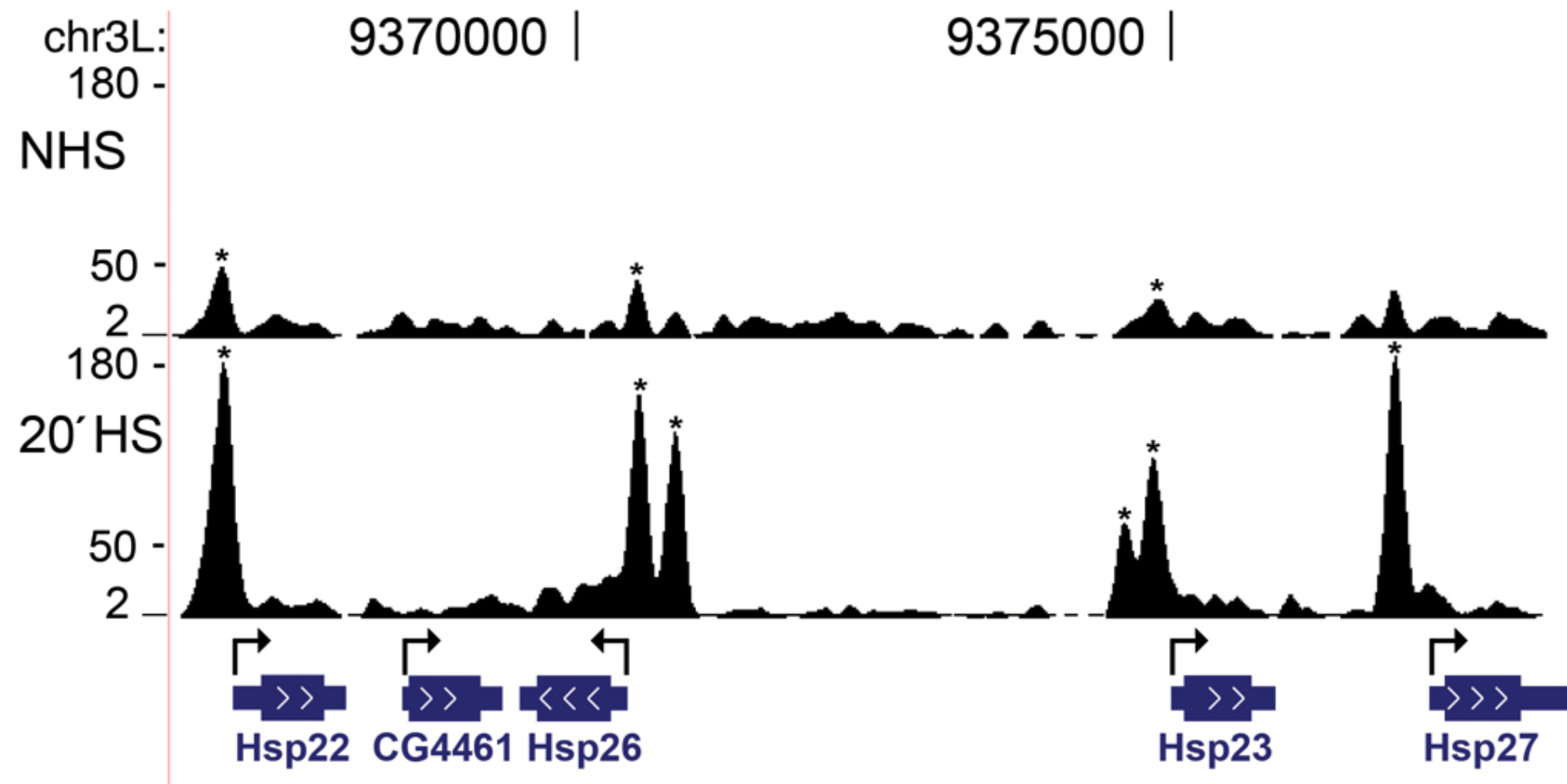
Sequence DNA ends



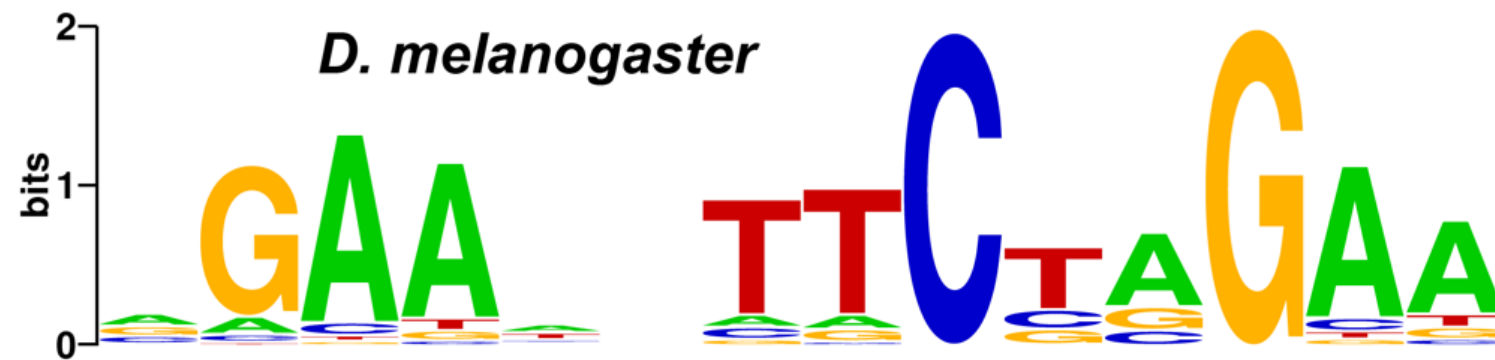
# Chromatin Immunoprecipitation (ChIP-seq)



# HSF targets DNA inducibly



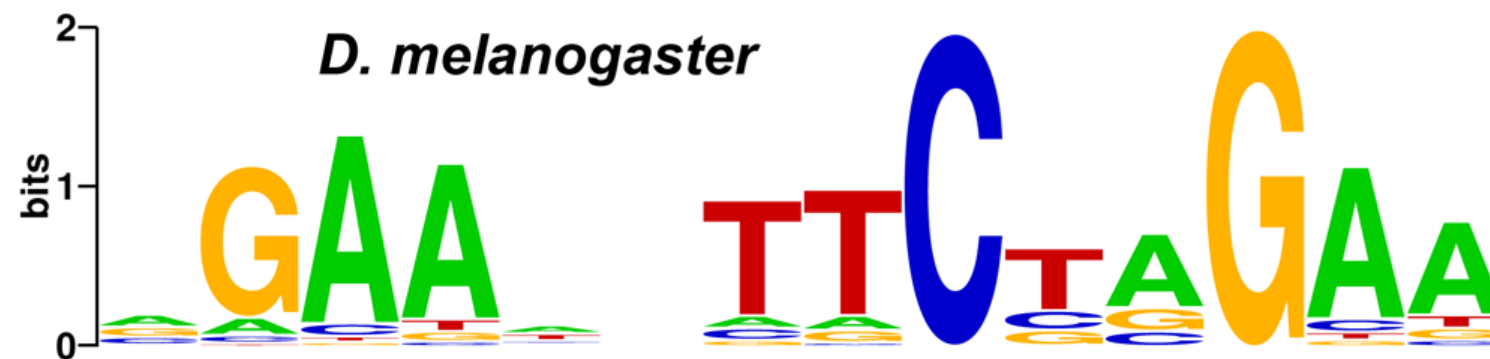
# HSF targets a consensus motif





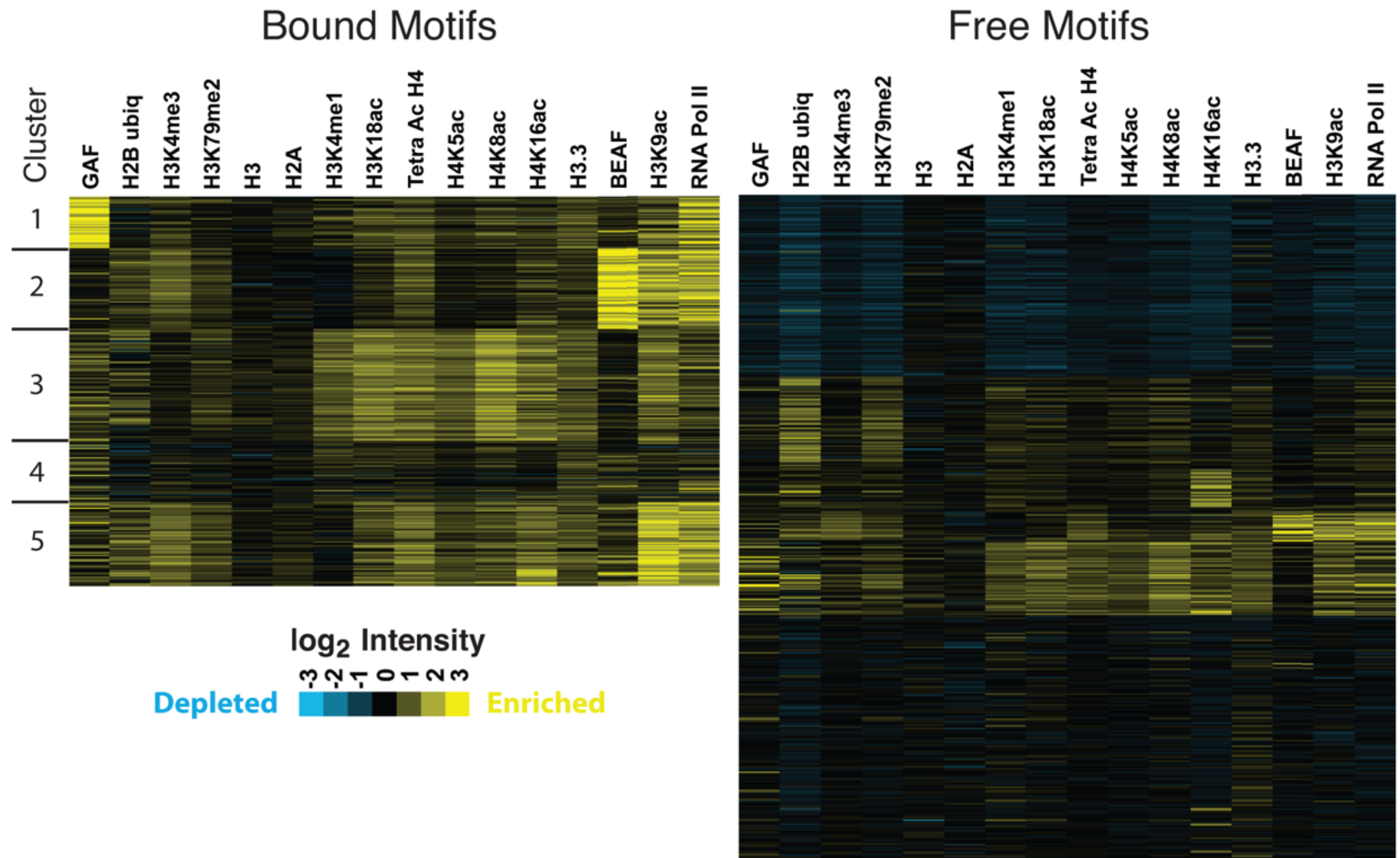
# HSF binds a fraction of motifs in vivo

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

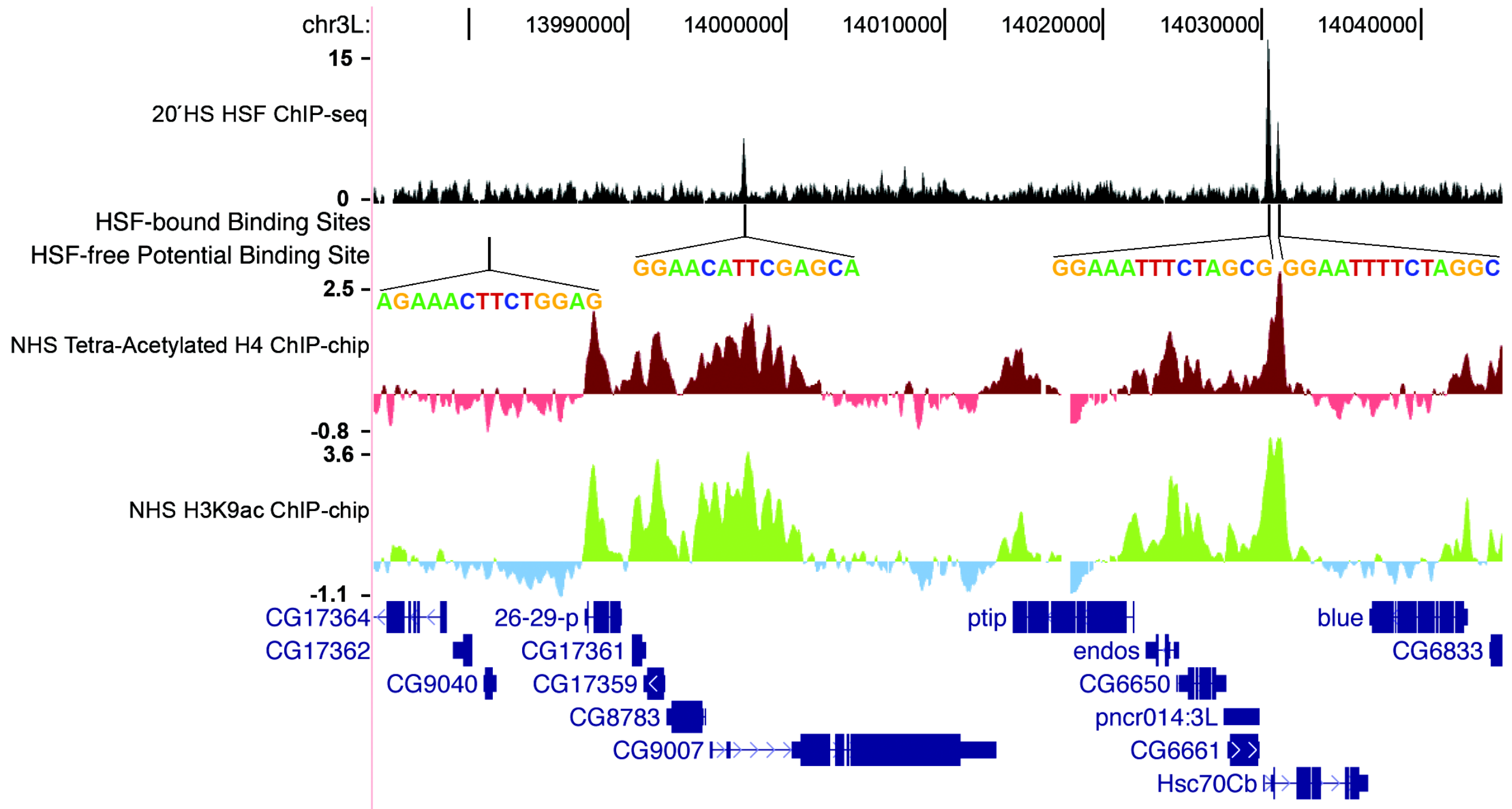


- Found 708 post-HS HSF-free motifs that conform stringently to this consensus HSE, compared to 442 HSF-bound motifs.
- Note that these are computational predictions of potential HSF binding sites

# HSF targets motifs within active chromatin

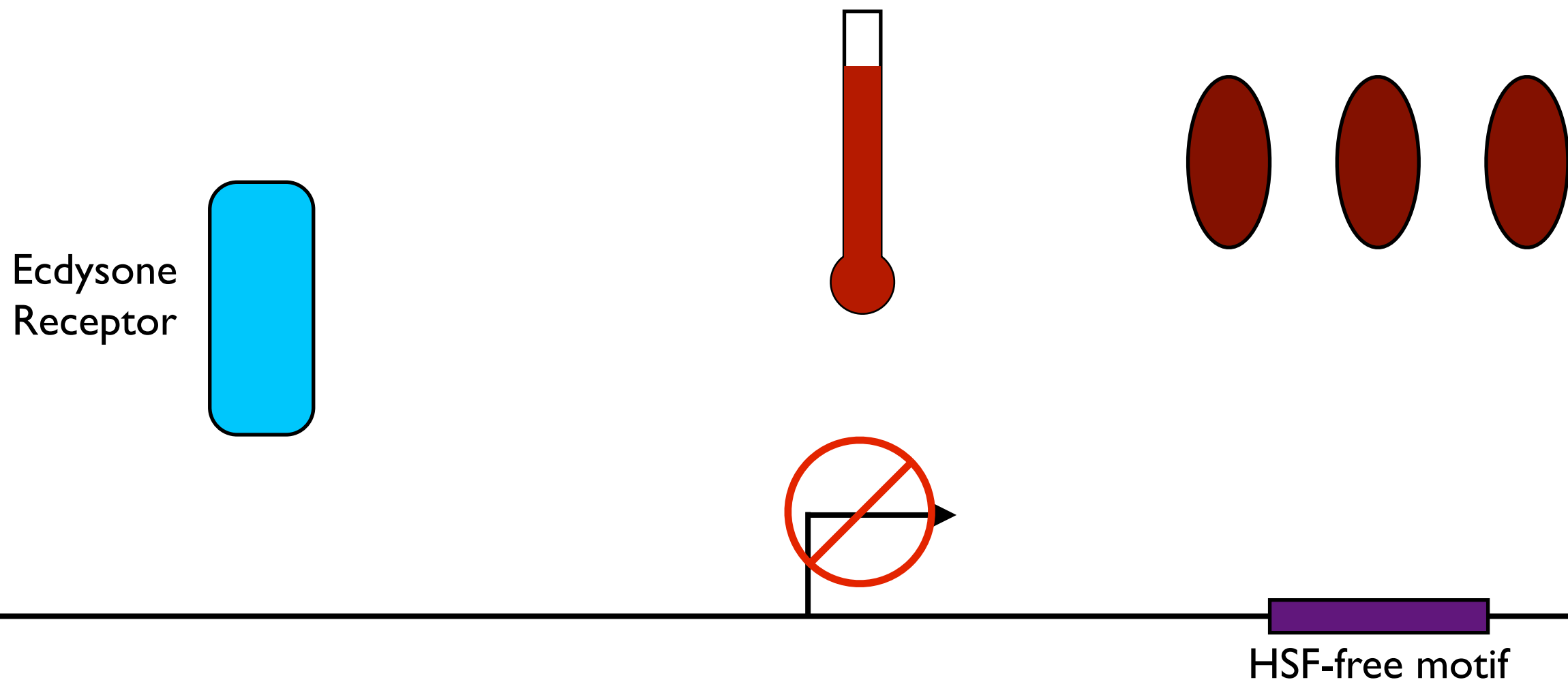


Correlative genomics allows one to develop a hypothesis;  
we hypothesize that active chromatin permits TF binding

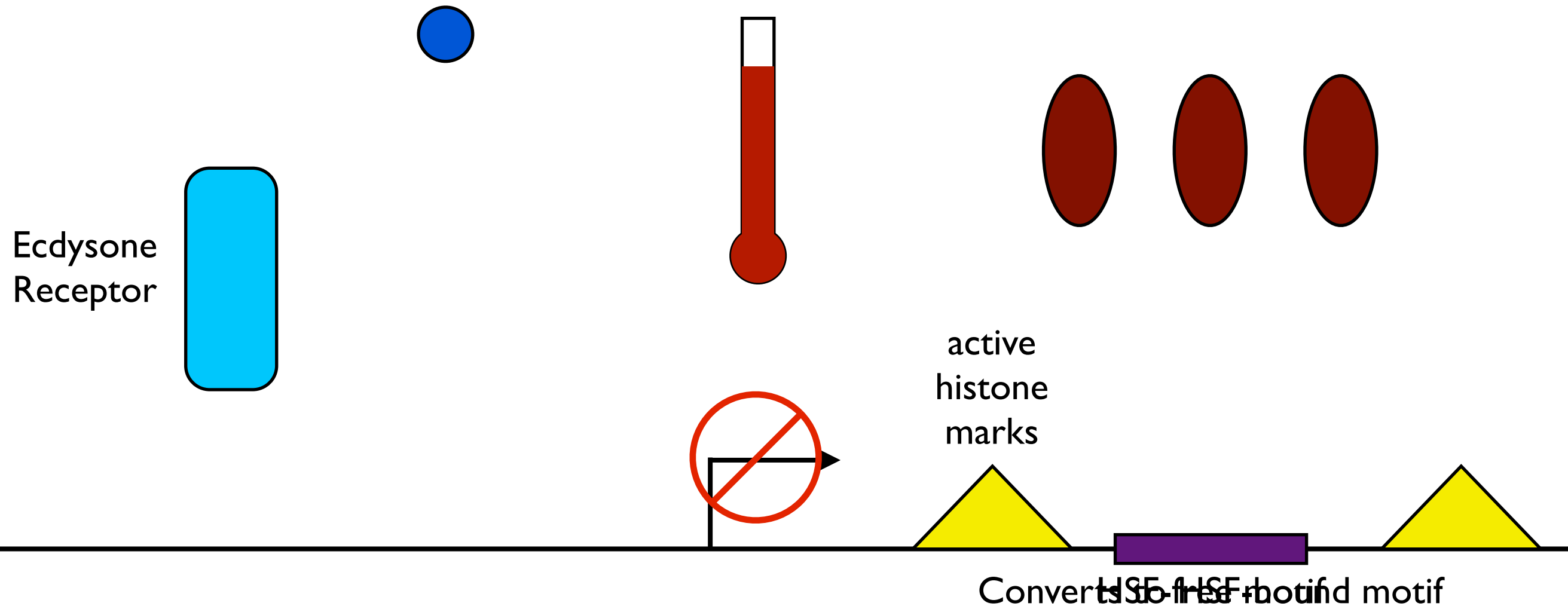


# Changing the chromatin at an HSE

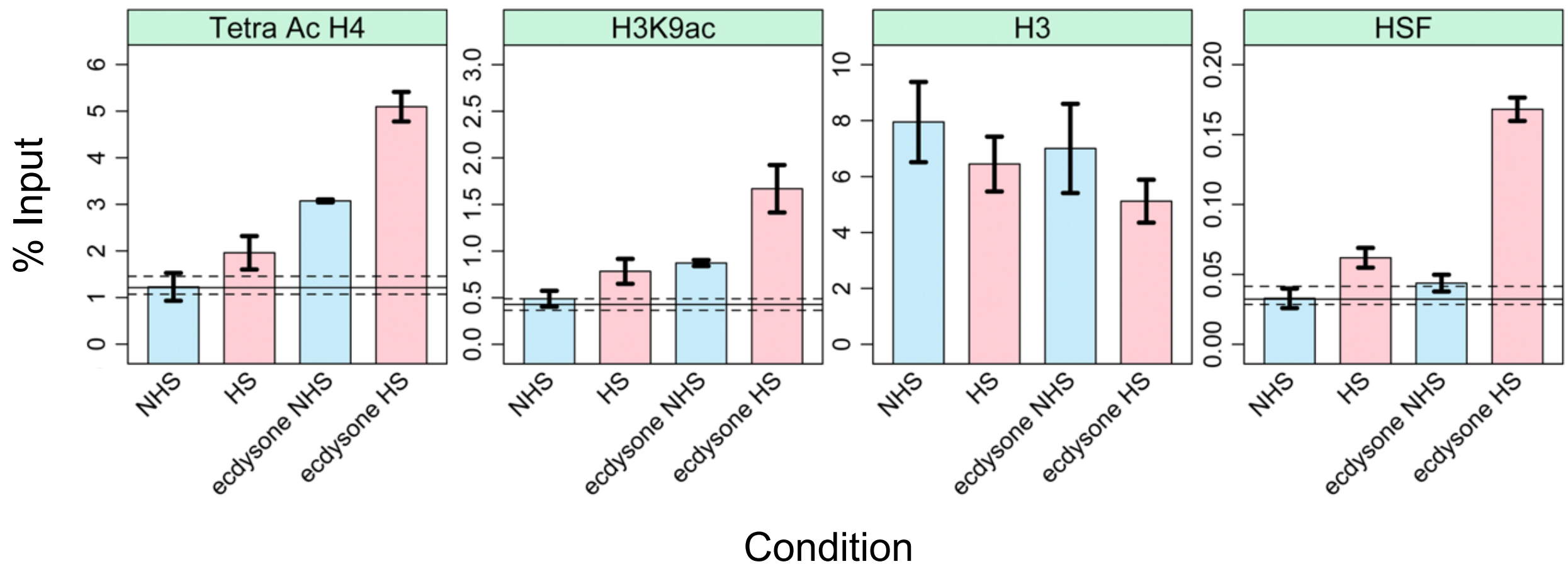
(now this is done trivially with CRISPR-dCas9 coupled to chromatin modifiers)



# Changing the chromatin at an HSE



# Converting and HSF-free to an HSF-bound motif

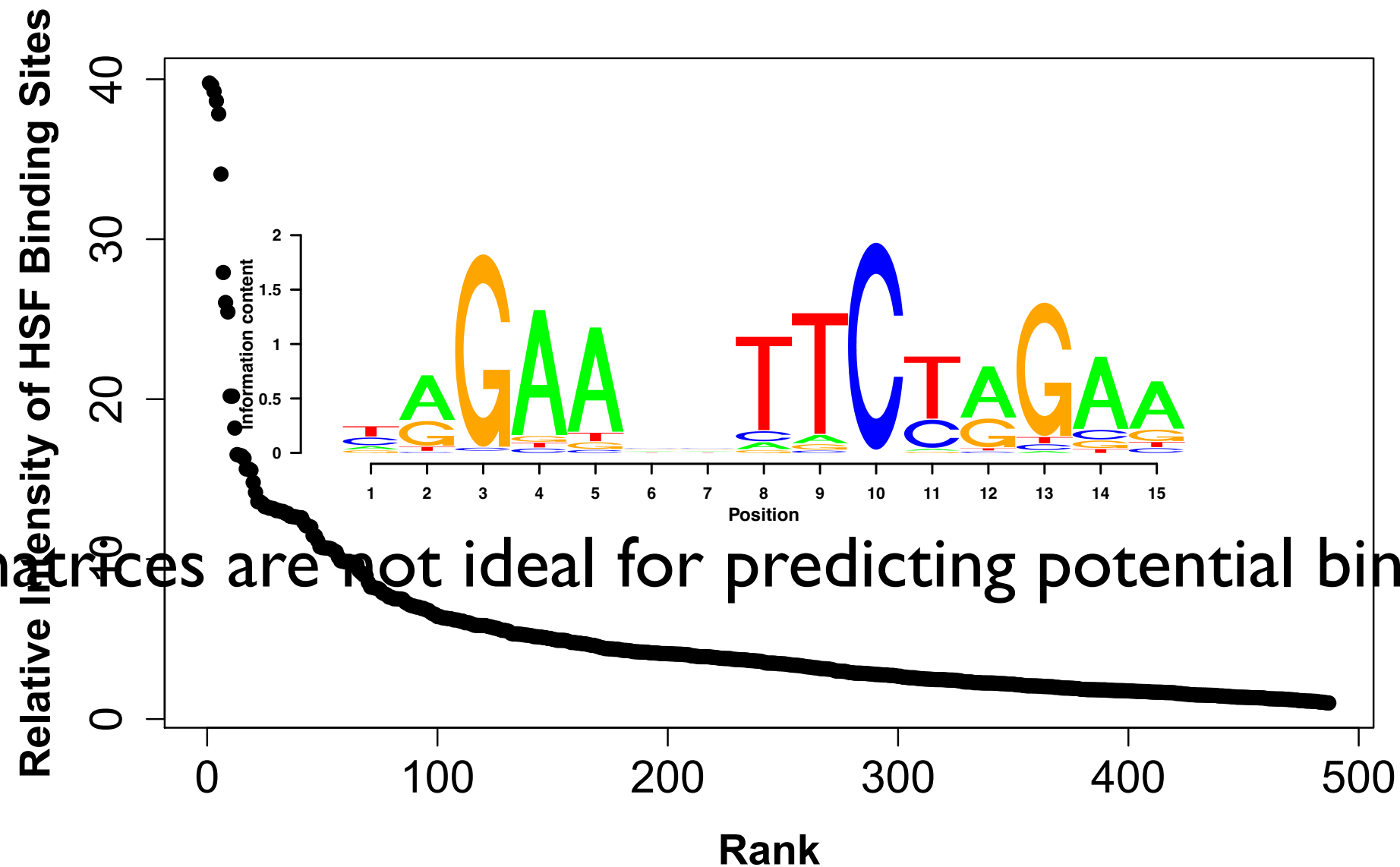




# Biology is continuous, not discrete

- 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)

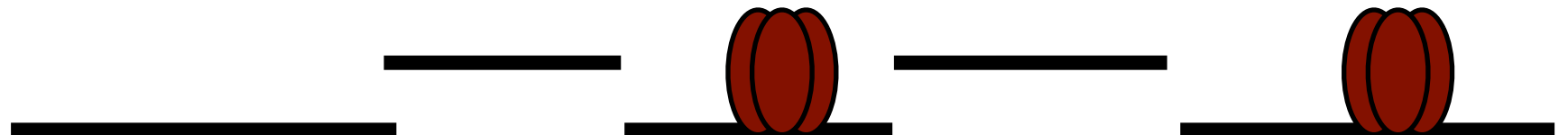
Isolate genomic DNA



Shear DNA



Incubate with  
Recombinant HSF



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

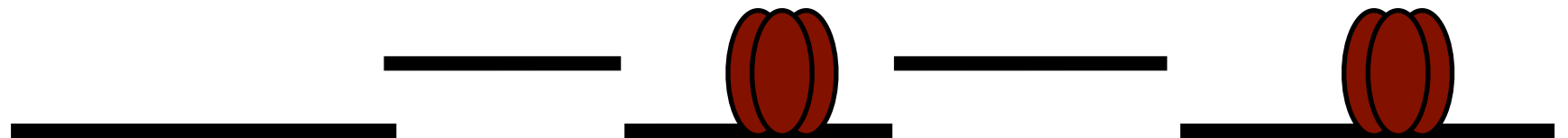
Isolate genomic DNA



Shear DNA



Incubate with  
Recombinant HSF



Immunoprecipitate  
HSF/DNA complexes





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

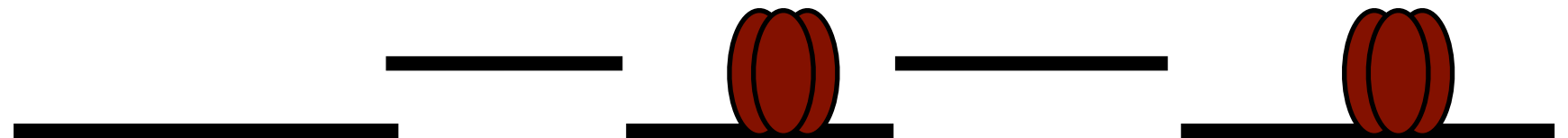
Isolate genomic DNA



Shear DNA



Incubate with  
Recombinant HSF



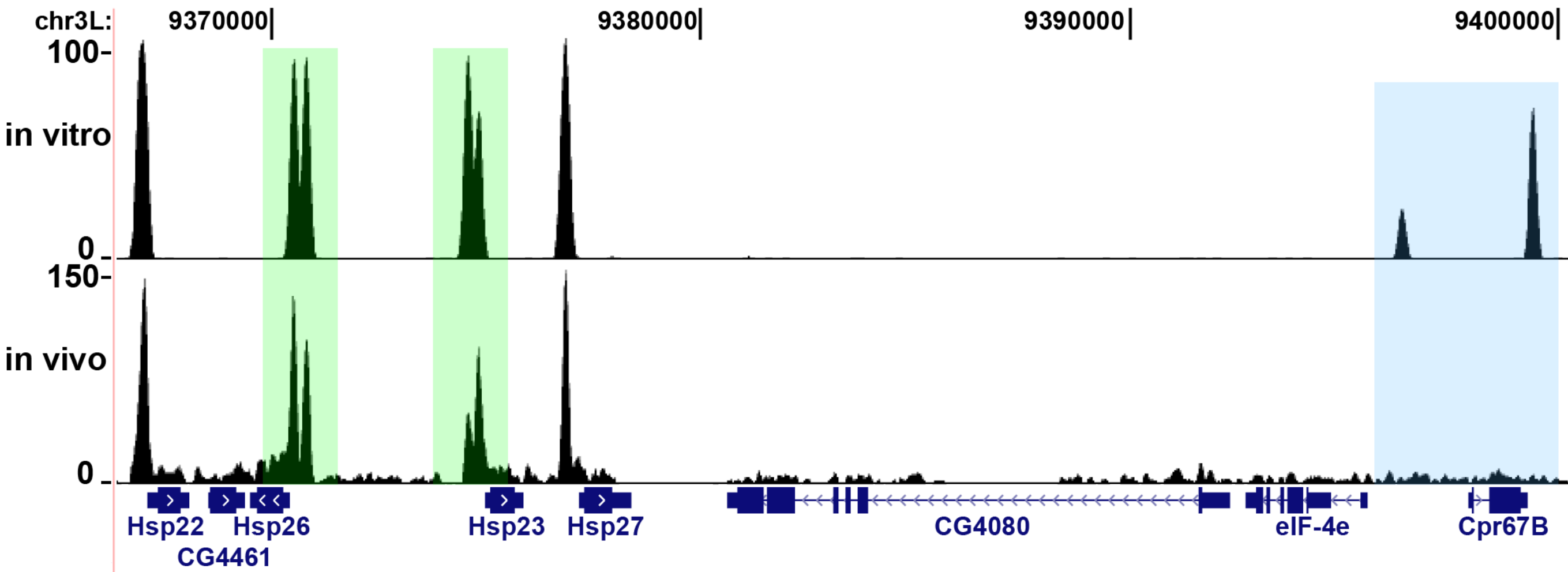
Immunoprecipitate  
HSF/DNA complexes



Purify/quantify DNA

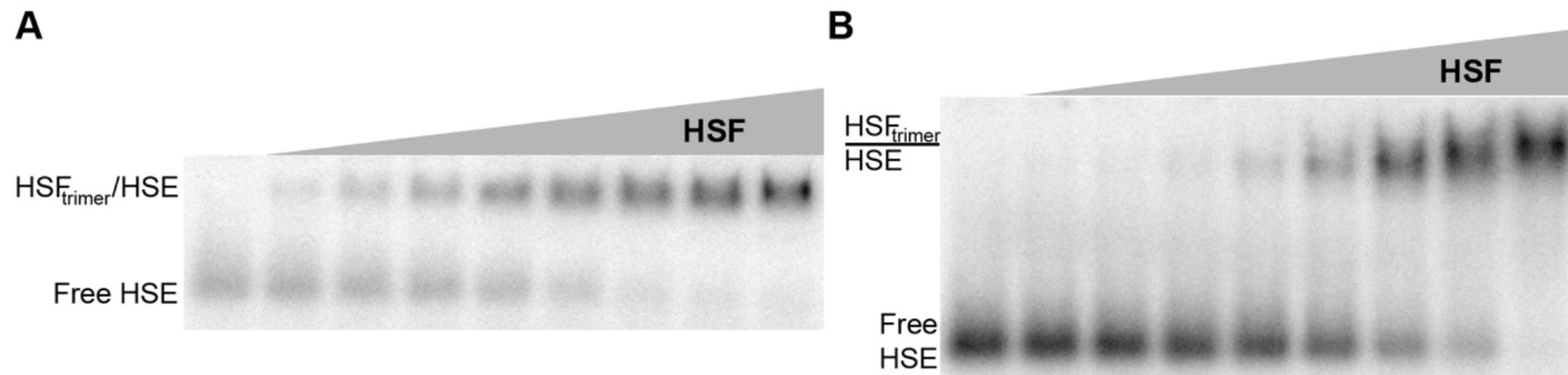


# in vitro Binding Assay (PB-seq) reveals all potential binding sites and relative affinities



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

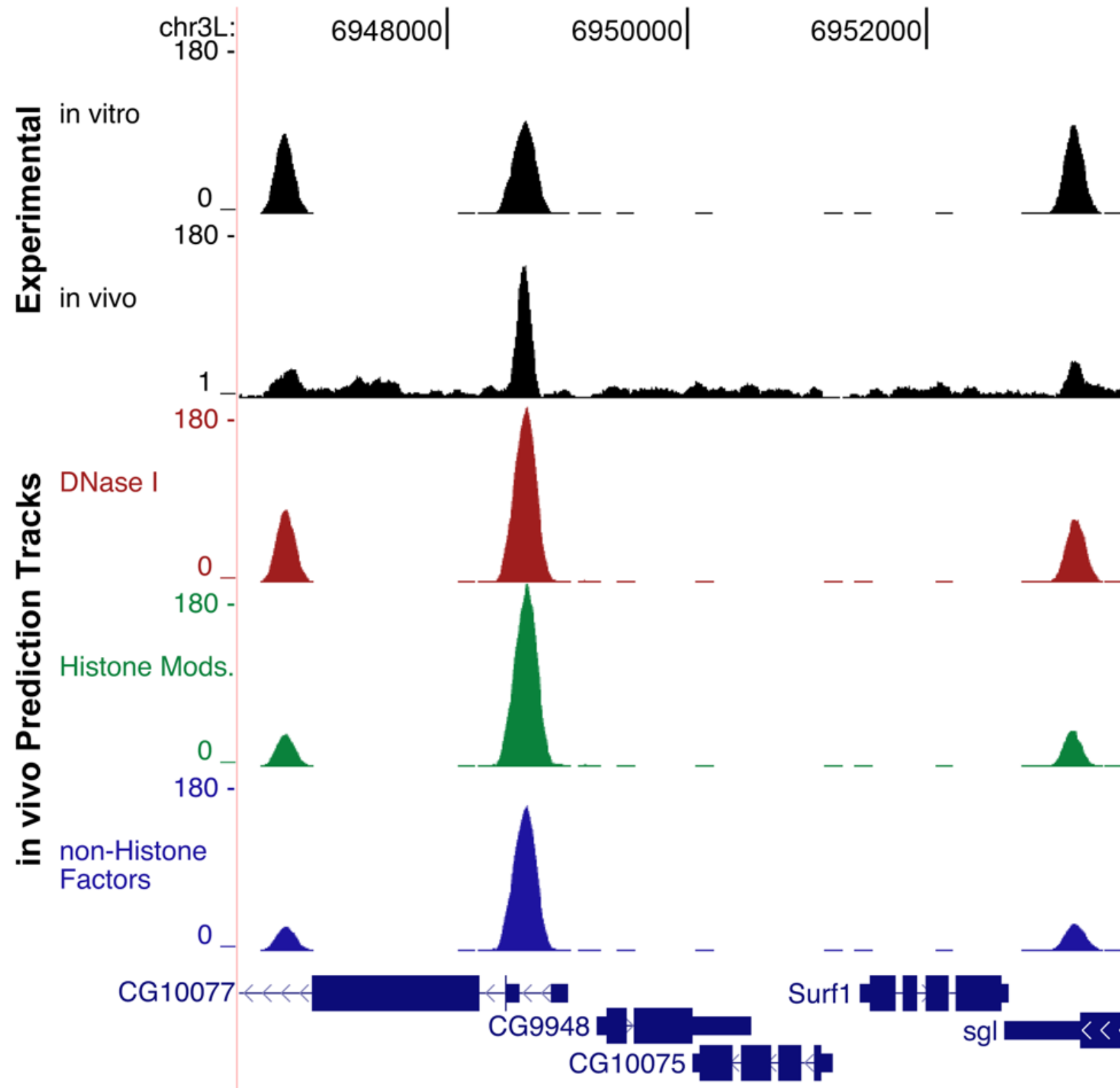
# HSF binds to HSEs with picomolar to nanomolar affinity in vitro



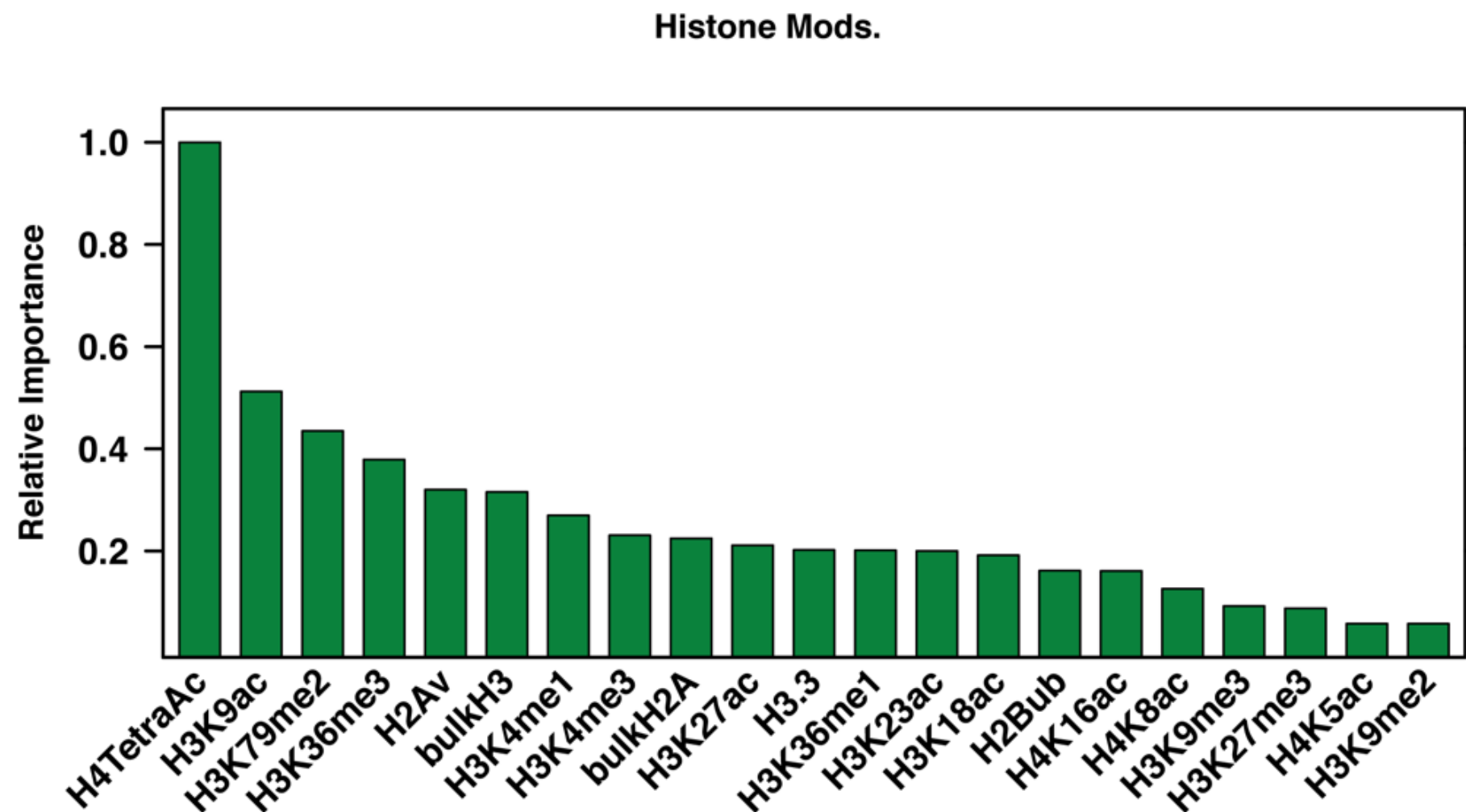
# Prediction of Binding Profiles using available PB-seq Data and Genomic Chromatin Data from modENCODE

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

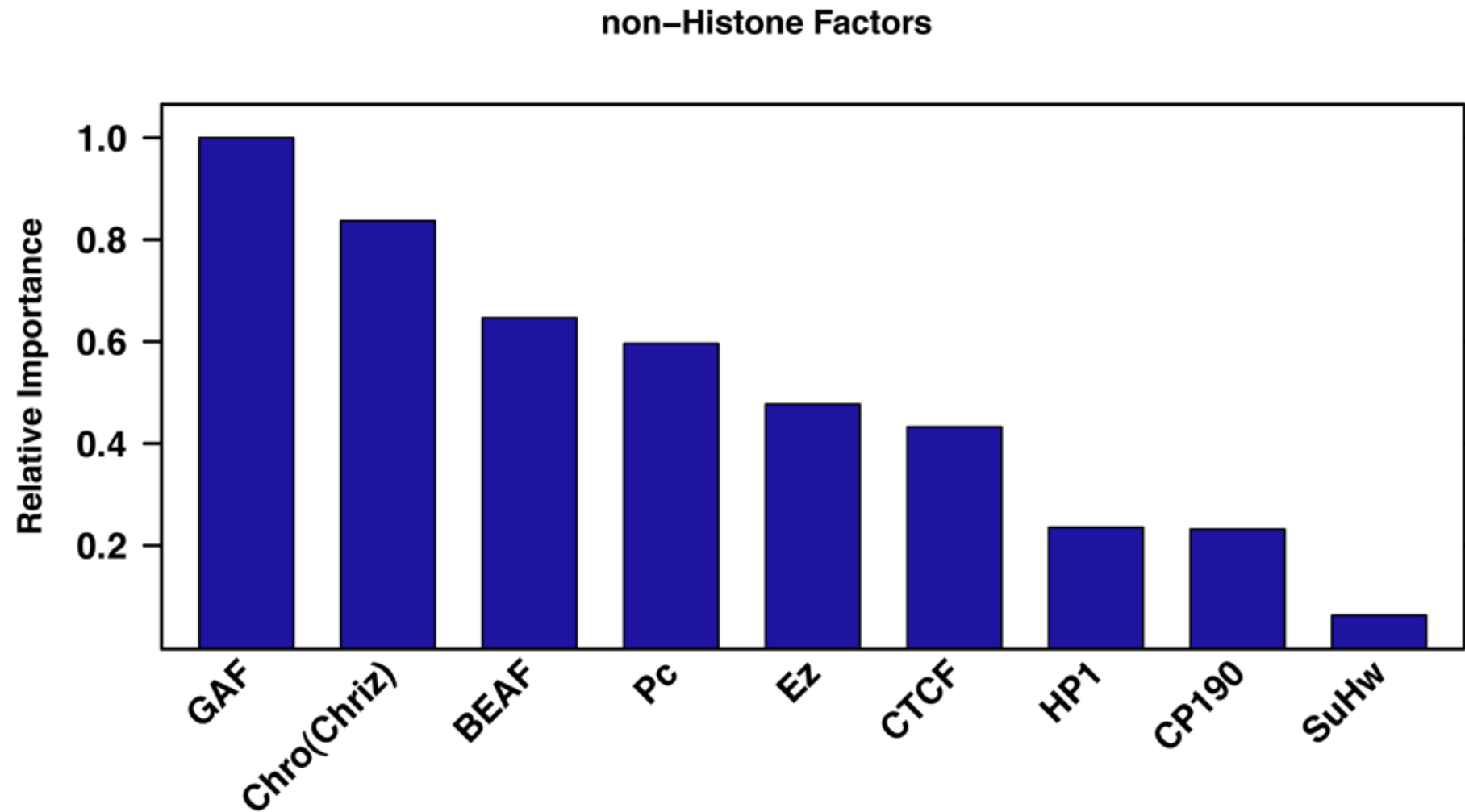


Histone Acetylation is the most influential modification for predicting HSF binding intensity

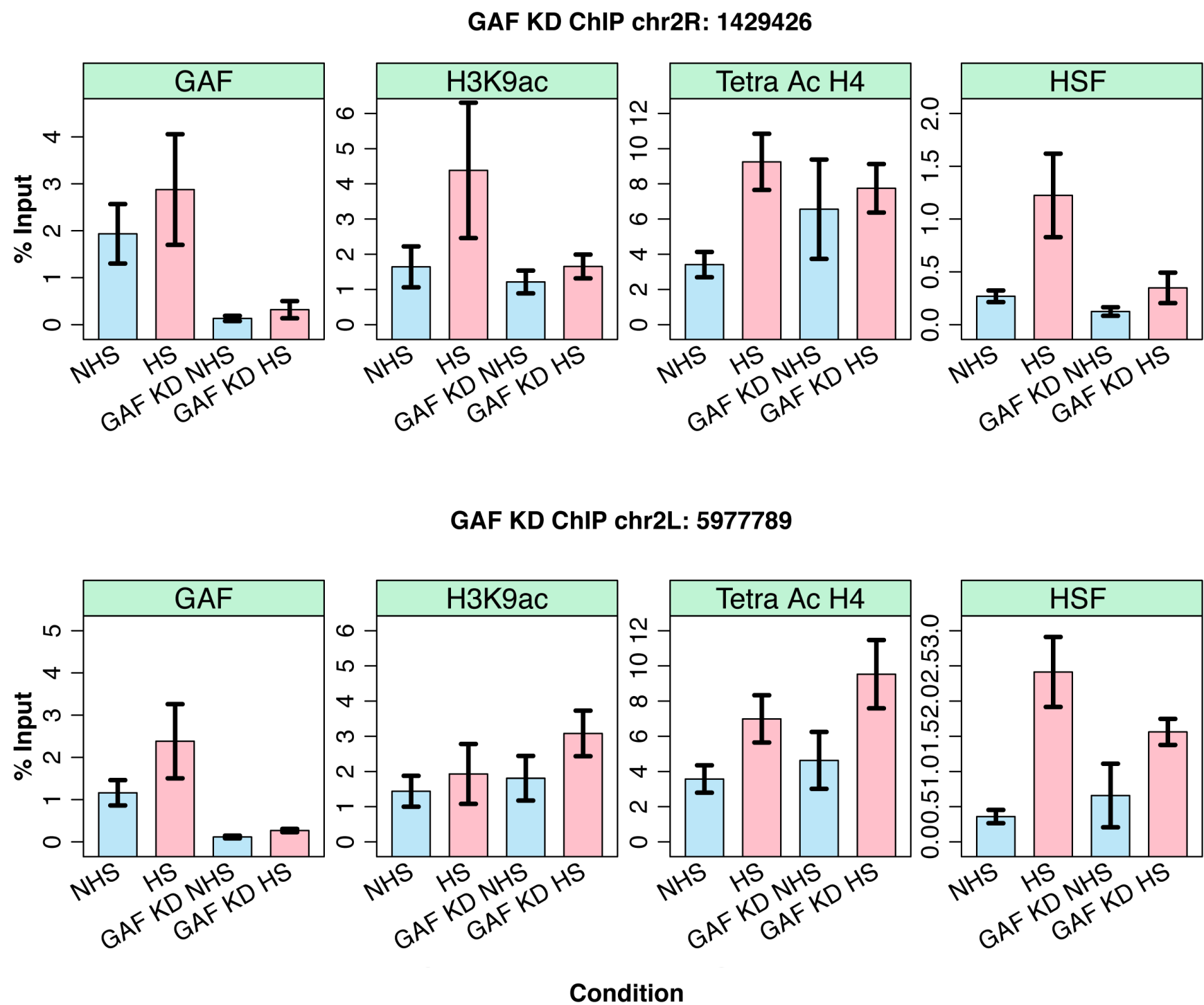




GAF is the most influential non-histone  
covariate in the predictive model



# GAF depletion compromises HSF binding intensity

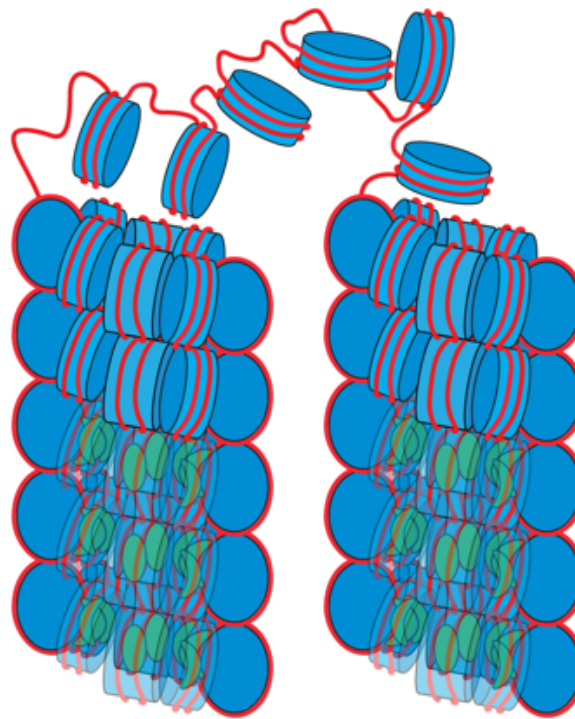


How do accessibility and active marks originate?  
working model:

Accessible



Not accessible

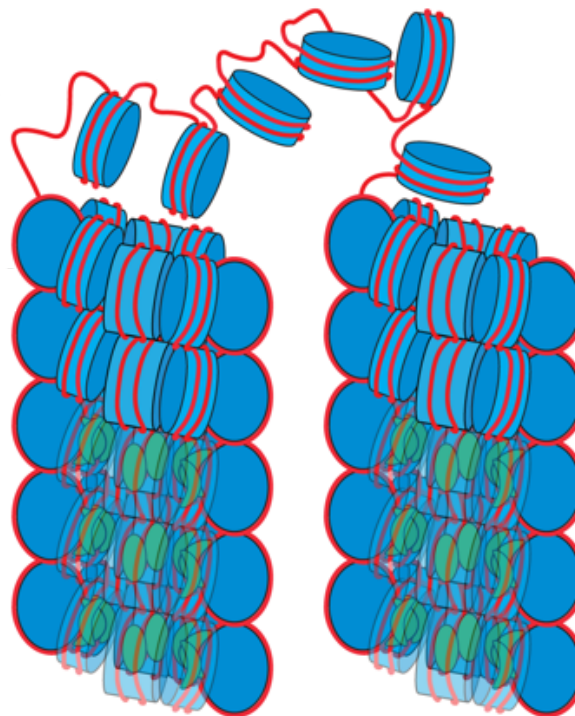


# DNA sequence directs accessibility

Accessible



Not accessible



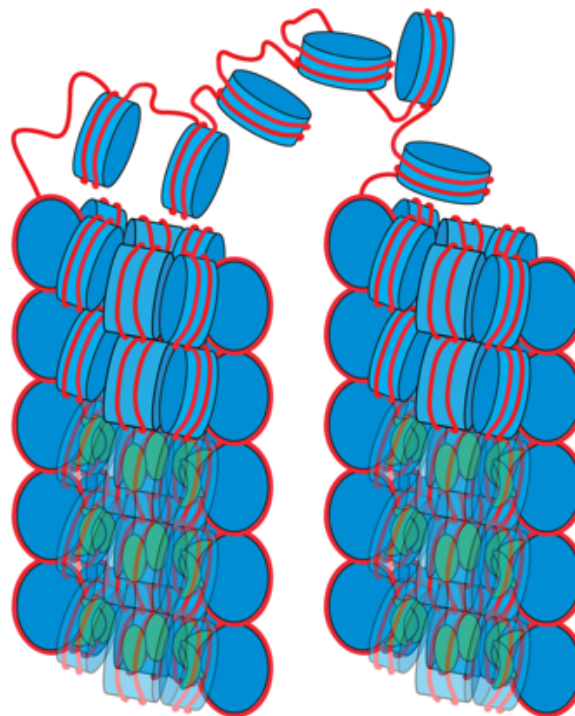
DNA Sequence (poly-AT tracts)  
keep yeast promoters  
free of nucleosomes  
(Struhl & Rando labs)

# DNA sequence directs accessibility

Accessible



Not accessible



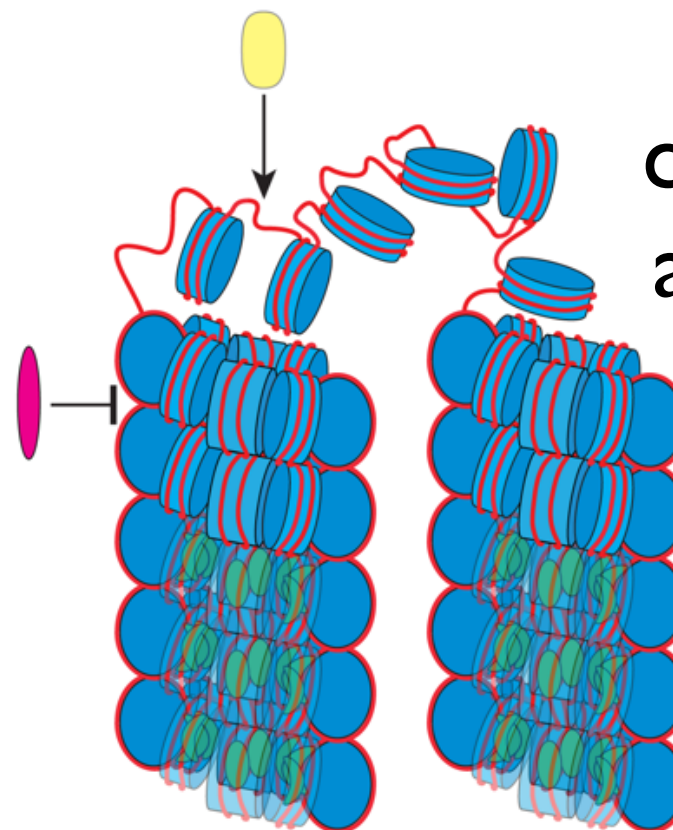
CpG islands favor nucleosome formation in vitro, but are depleted in vivo.

HI 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 HI, which maintains the chromatin in a transiently accessible state.

# Factors target linker DNA between nucleosomes



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

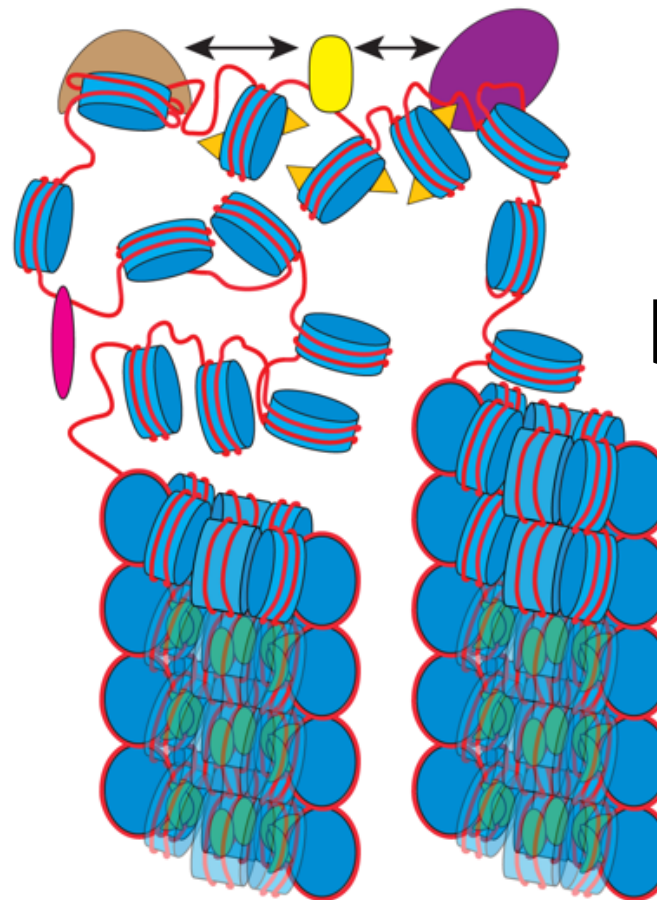
The Set1 H3K4 methylation complex and a H3K36 demethylase has been shown to interact with unmethylated CpG-rich DNA in vitro.

(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

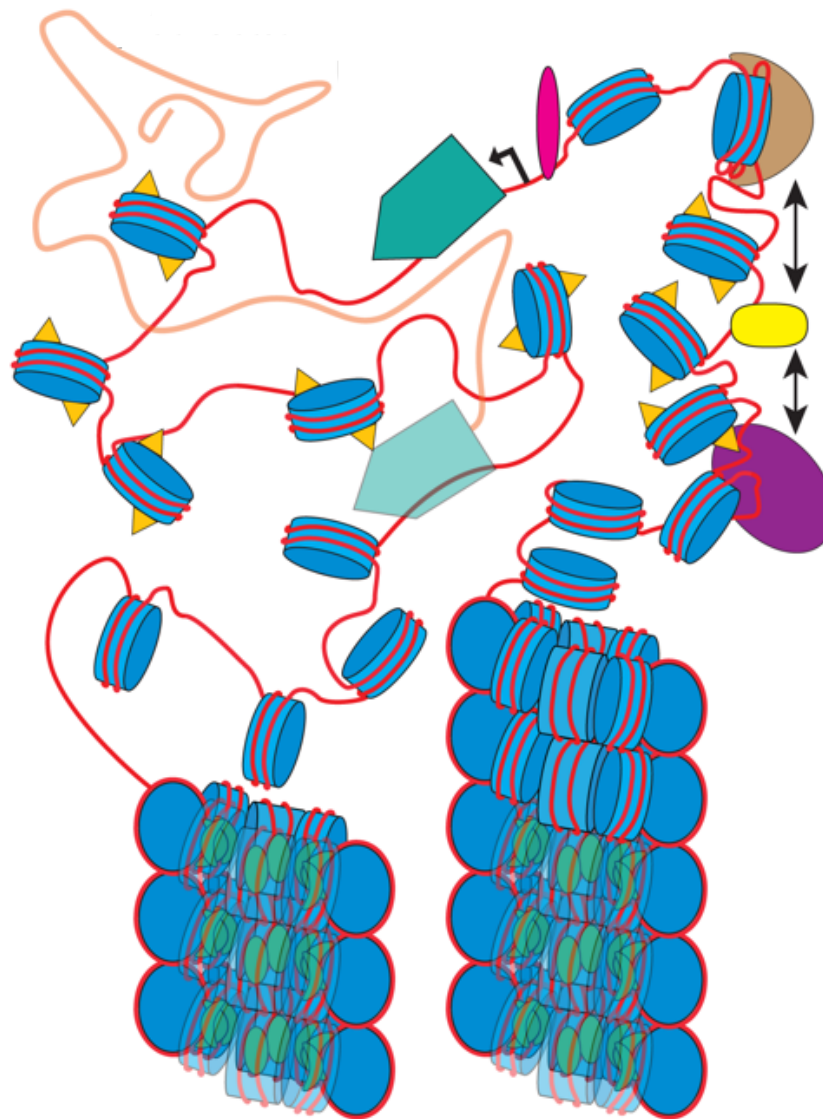


TF binding increases and expands the boundaries of accessibility and active marks.



# Binding expands accessible regions

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



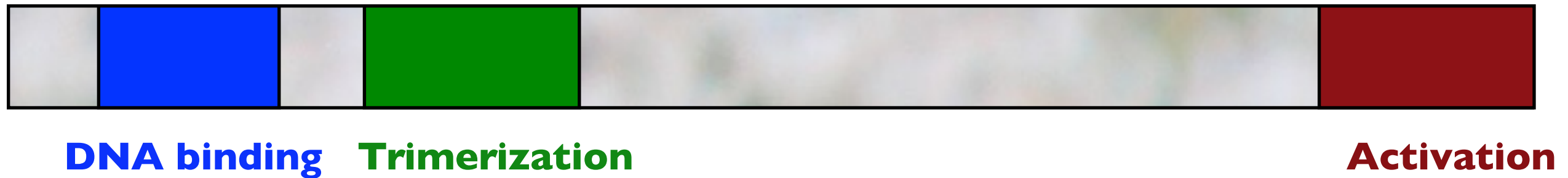
Is there a paradigm shift in biology, away from overly hypothesis-driven research?

- 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 and activation domains of transcription factors are distinct and separable

**GAL4 - a eukaryotic activator**



DNA  
Binding  
Domain to  
UAS<sub>G</sub>

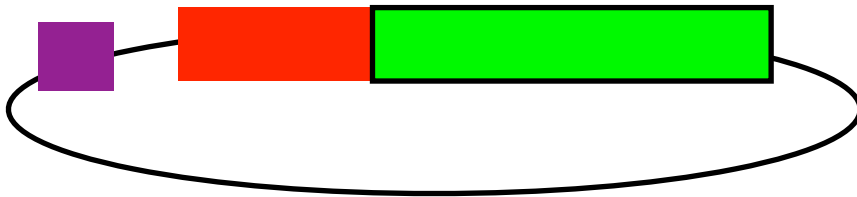
**LexA - a bacterial repressor**



DNA  
Binding  
Domain to  
LOP

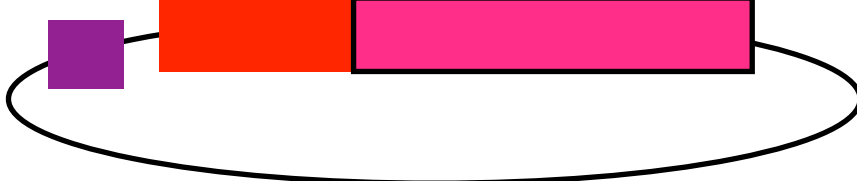
Cut & Splice & Join  
to Expression Vector

PADH



Transform Yeast  
containing a  
**Gal1-lacZ**  
Reporter

PADH



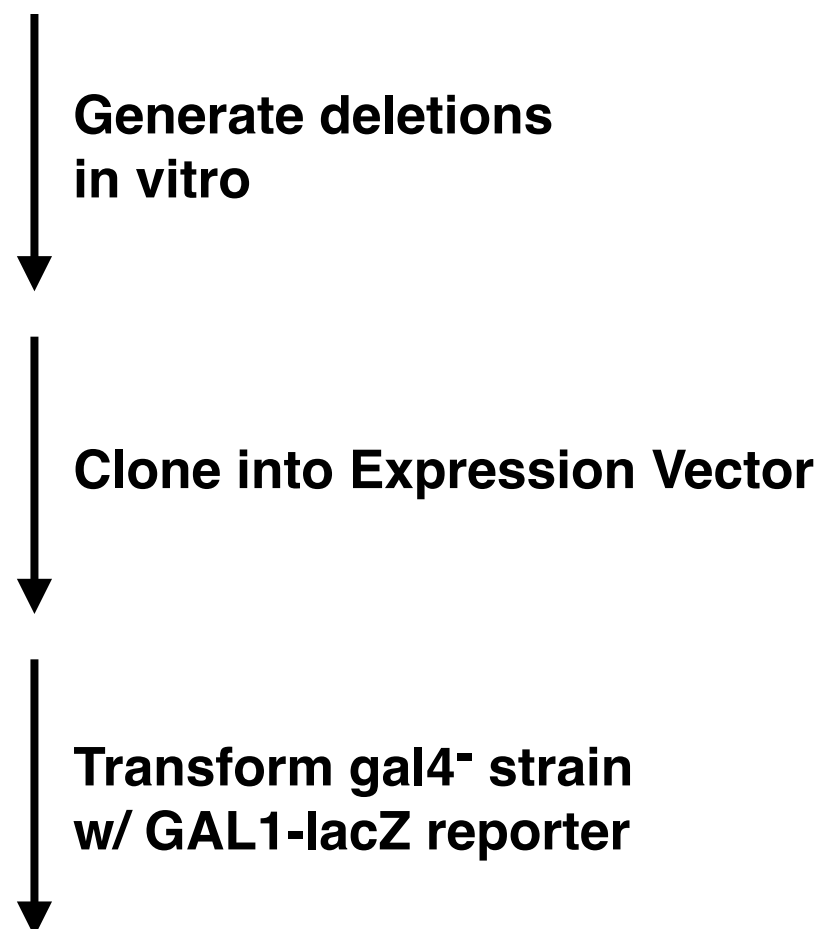
Remove and  
Substitute the  
UAS<sub>G</sub>

**ΔUAS<sub>G</sub>**

**LOP**

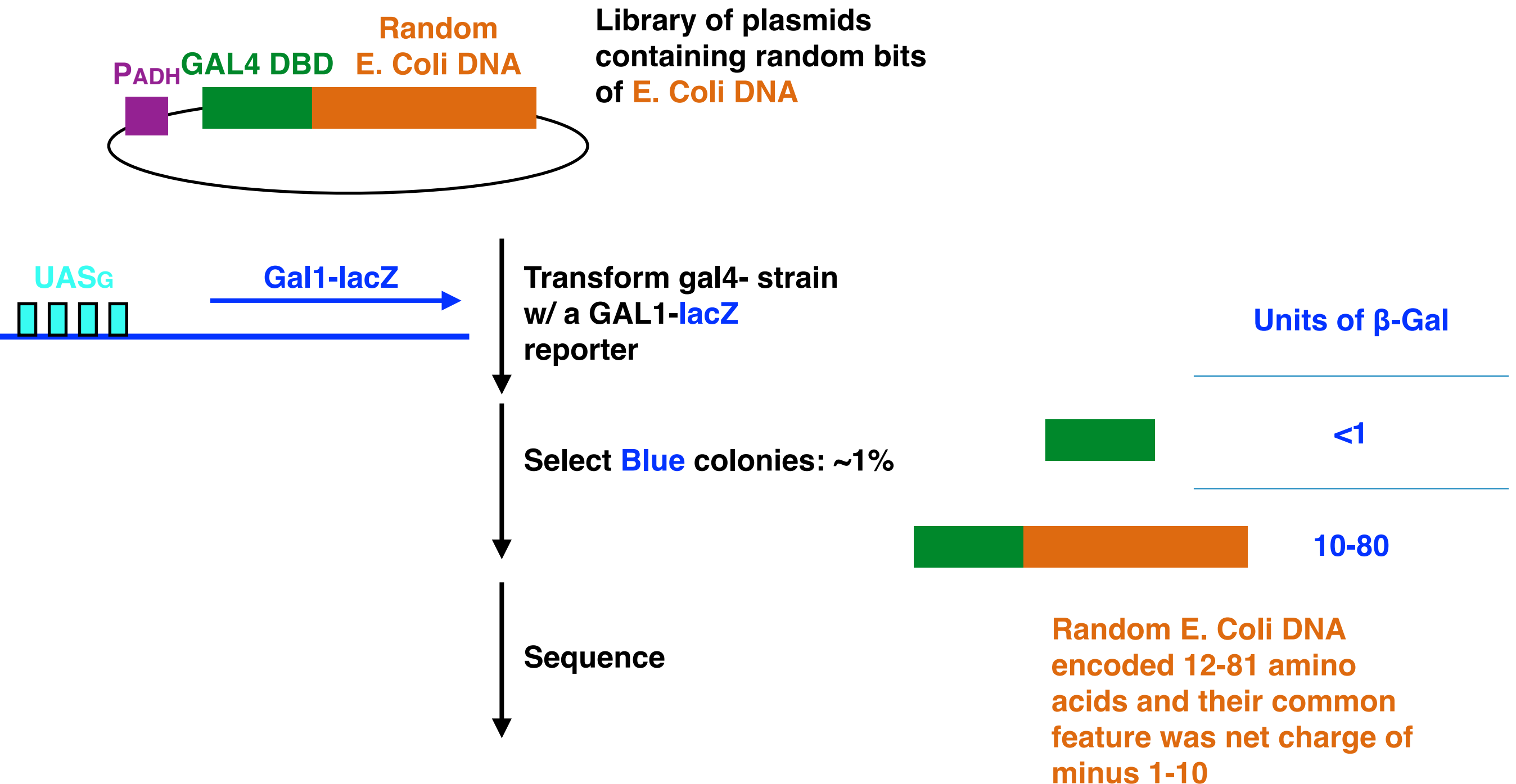
Vector Expression Units of β-Gal	
LexA	LexA- GAL4
0	0
<1	520

# Transcription Activation Activity can reside in one or more regions



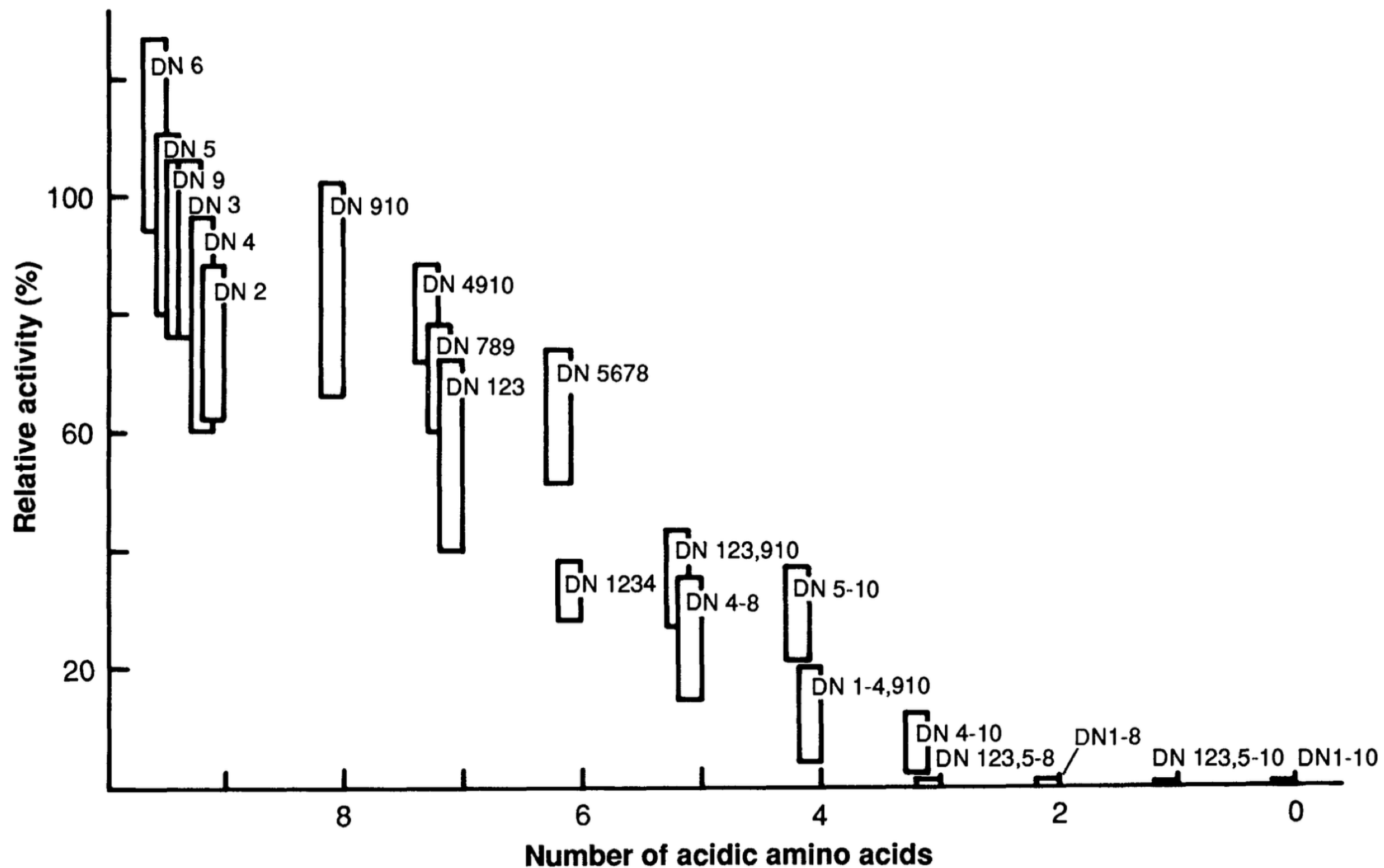
Construct	Units of $\beta$ -Gal
<b>GAL4 WT</b>	<b>1900</b>
<b>GAL4 1-147</b>	<b>&lt;1</b>
<b>GAL4 1-238</b>	<b>110</b>
<b>1-147/768-881</b>	<b>110</b>
<b>1-238/768-881</b>	<b>1400</b>

# Transcription activation regions occur frequently and are often acidic

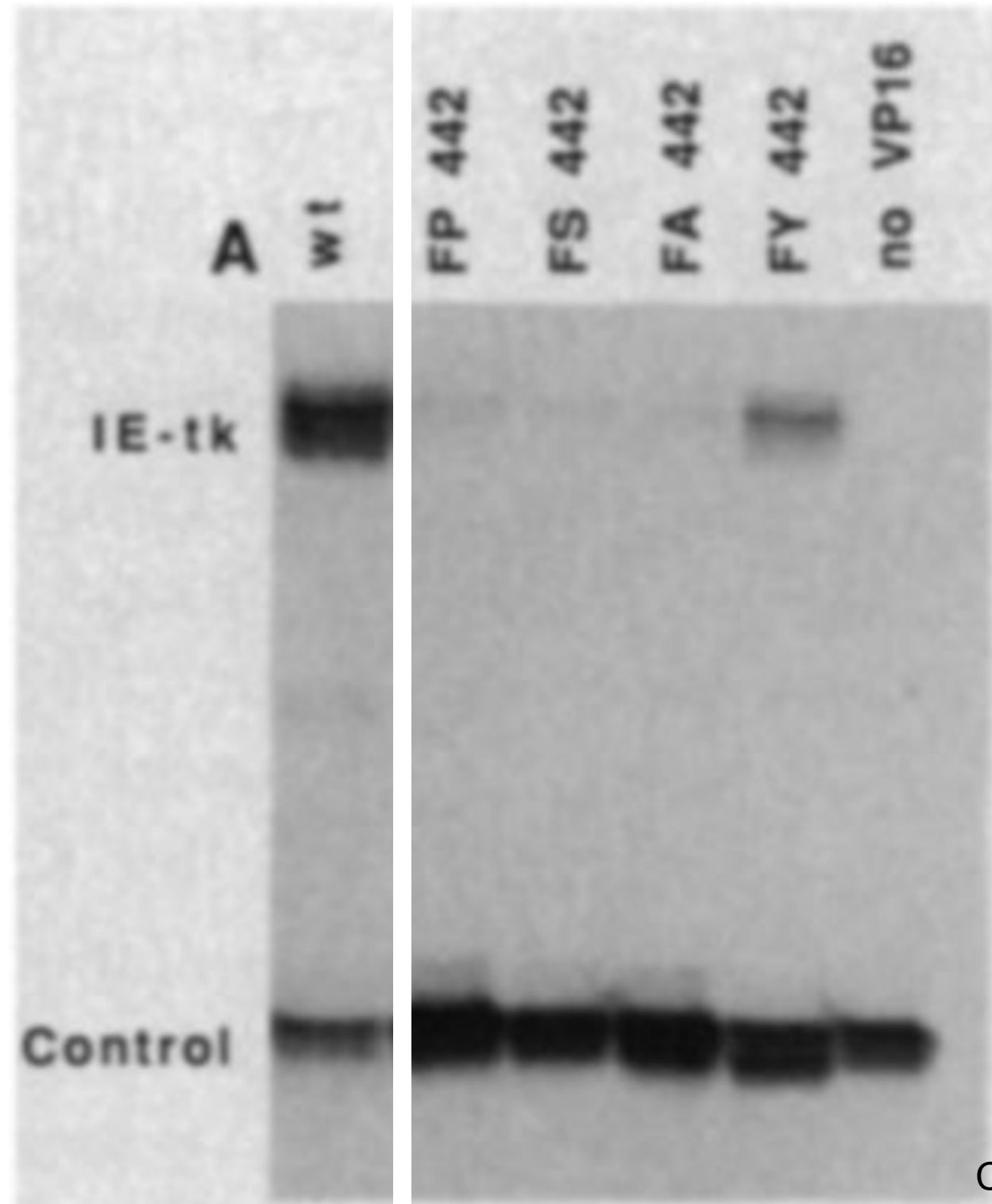




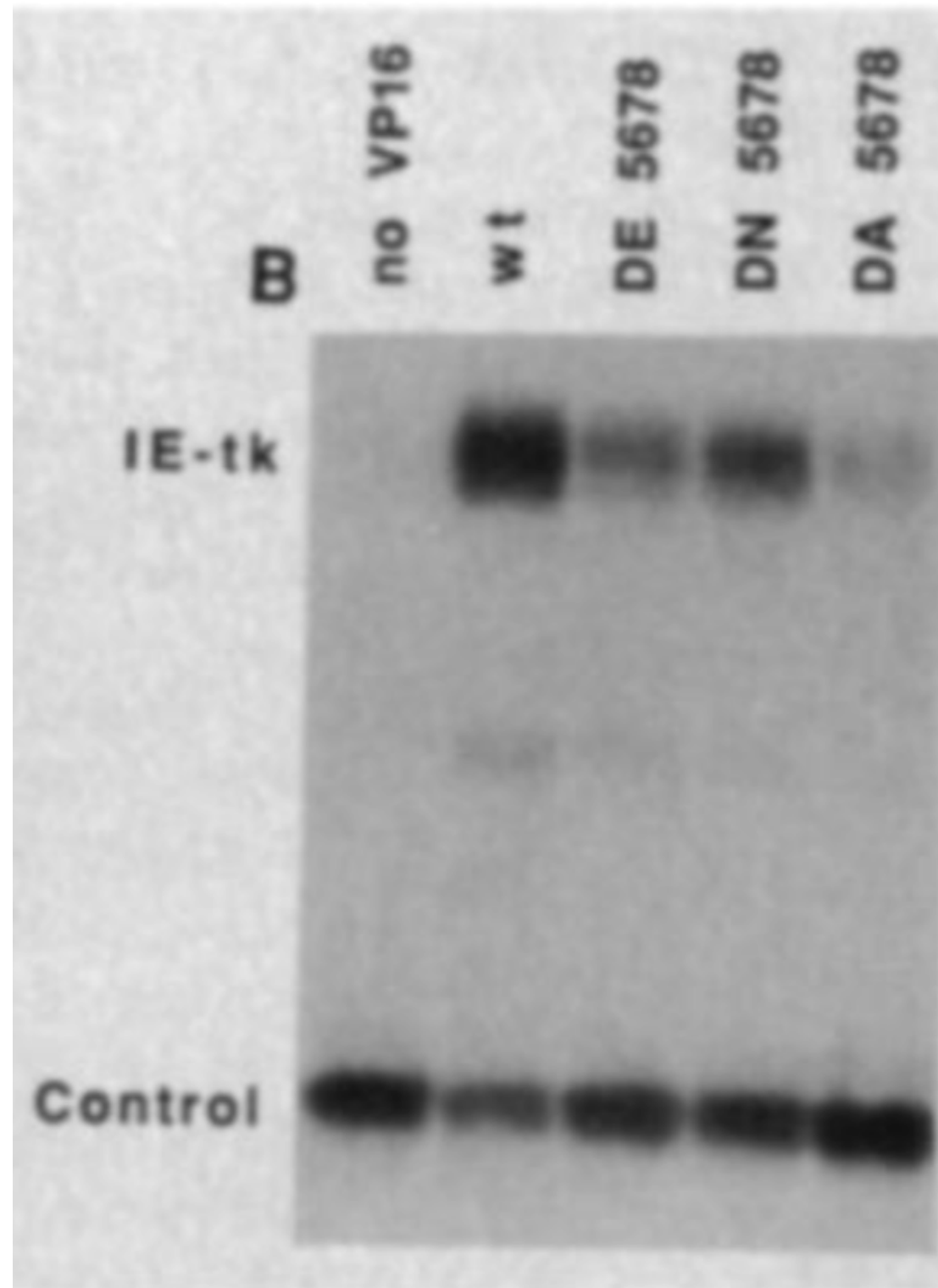
# Negative charge correlates with transcriptional activity of VP16



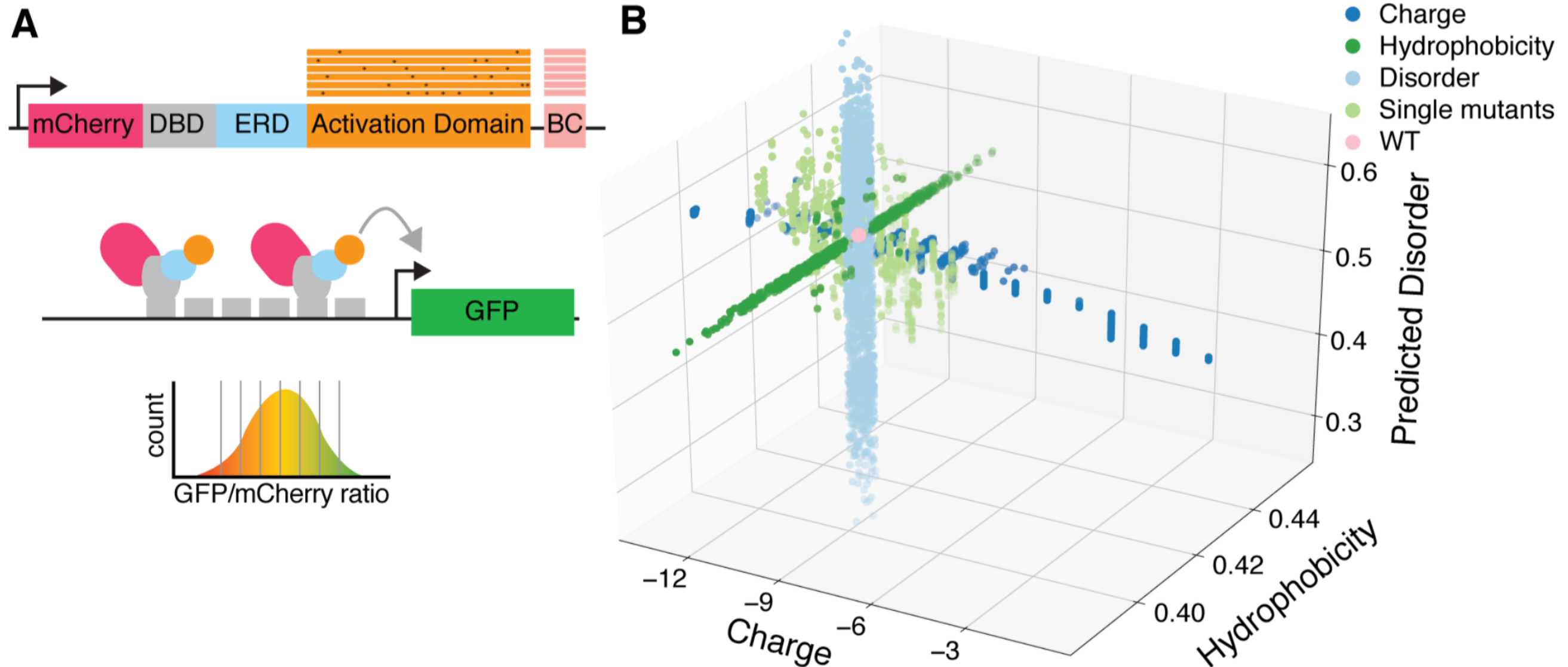
# A single Phenylalanine is critical to VP16 function



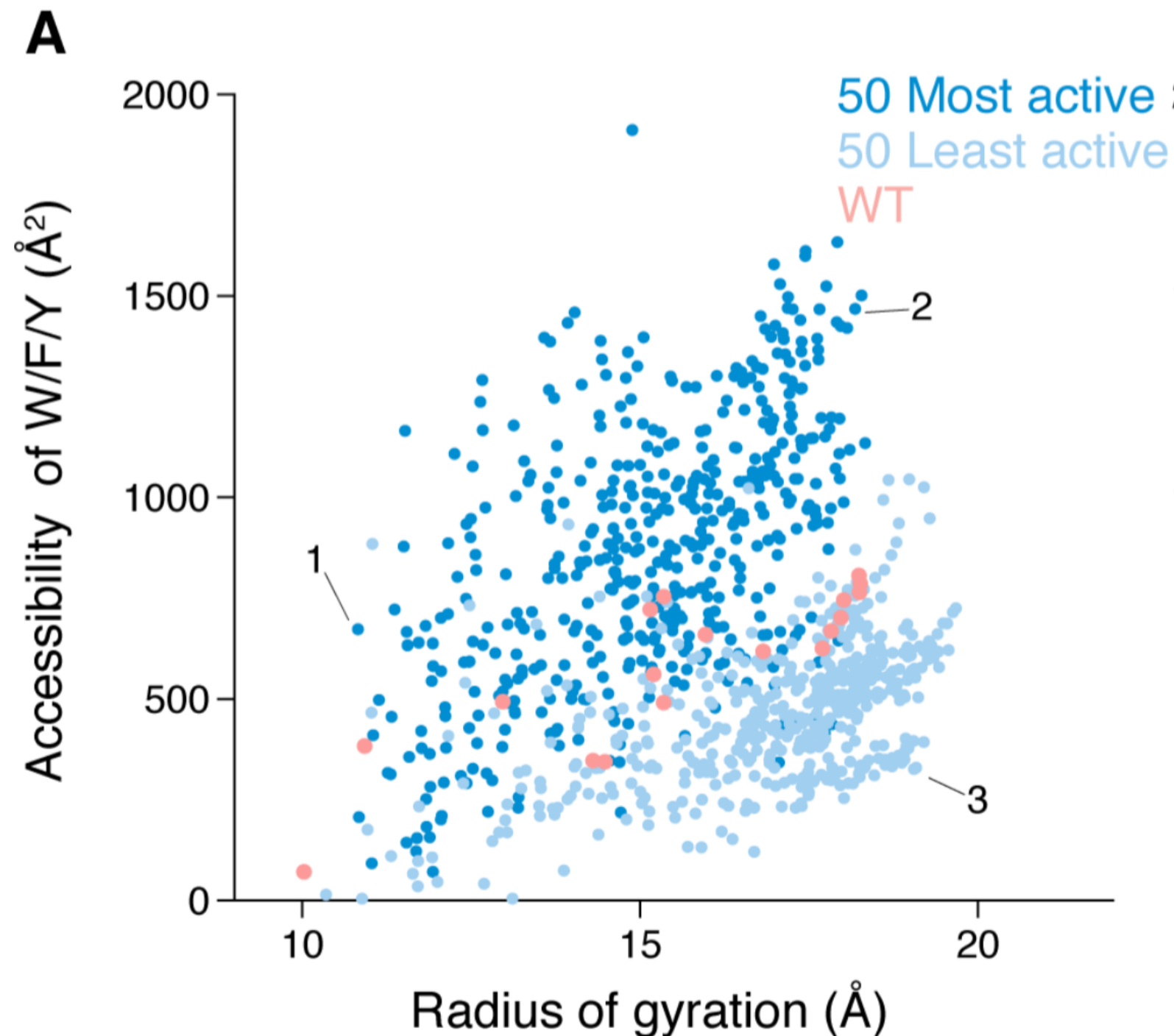
# The environment surrounding Phe<sup>442</sup> affect VP16 function



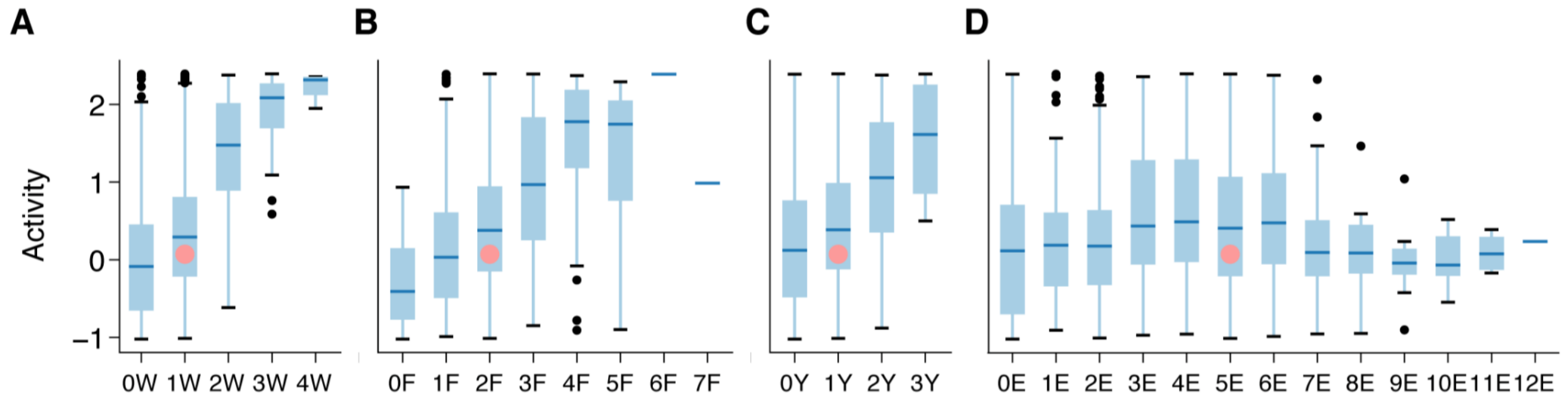
# Measuring thousands of designed GCN4 activation domain mutants in parallel



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



# Aromatic residues control Gcn4 Activity



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

# **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.
- A contemporary example is lncRNAs.

# **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)