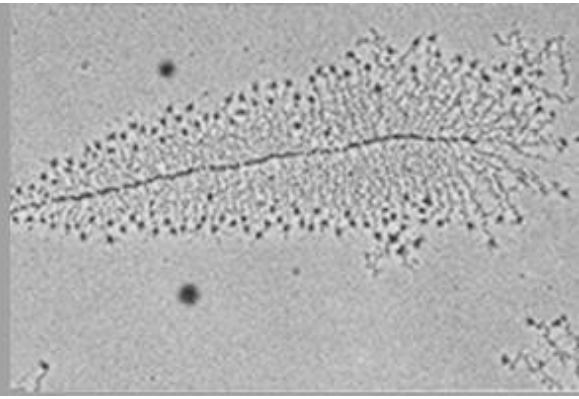
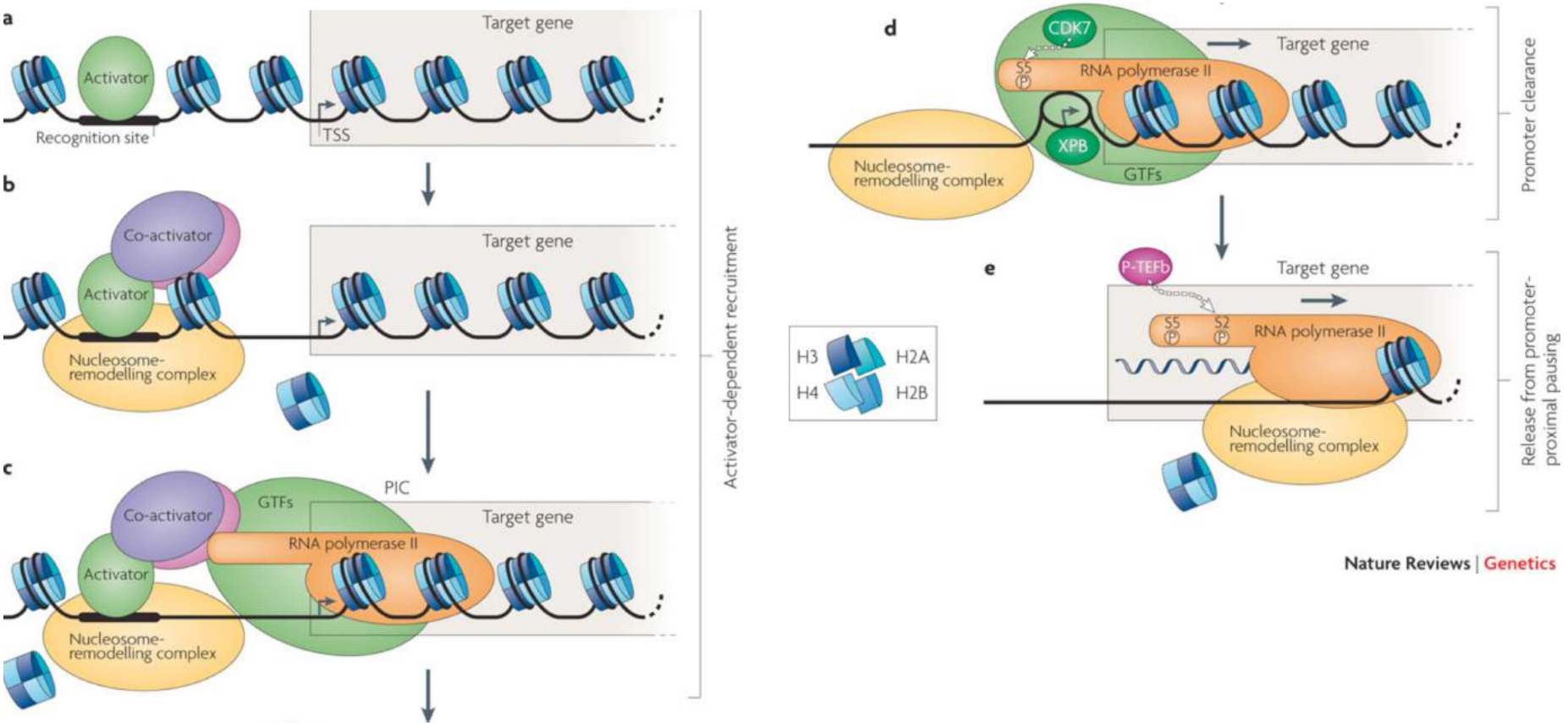
The regulation of eukaryotic transcription



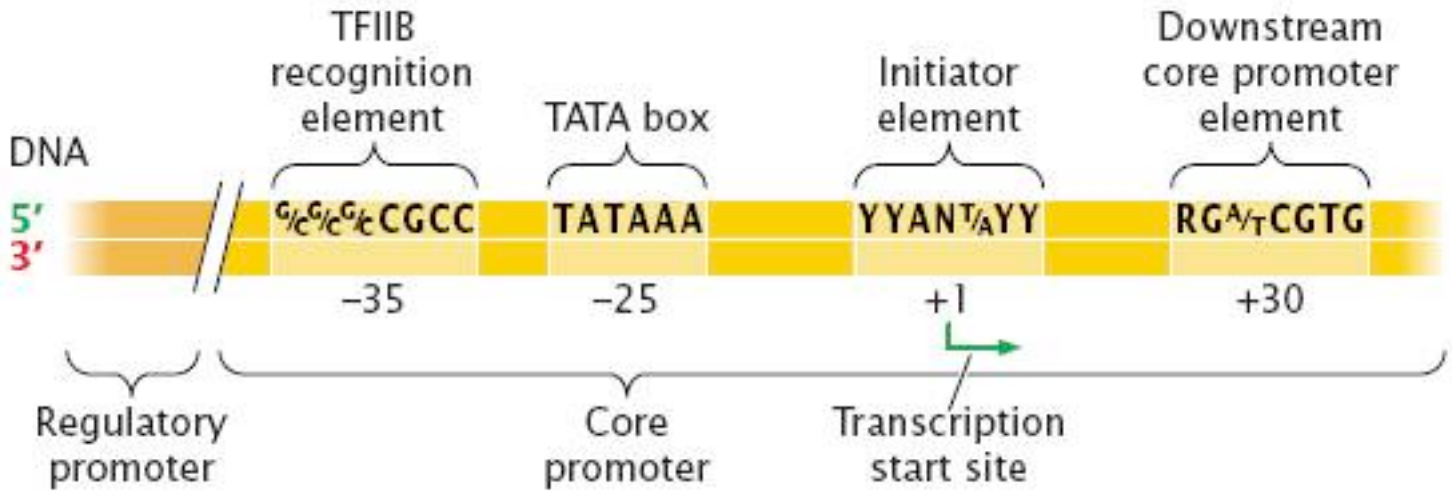
Máté Varga
mvarga@ttk.elte.hu

Transcription in eukaryotes

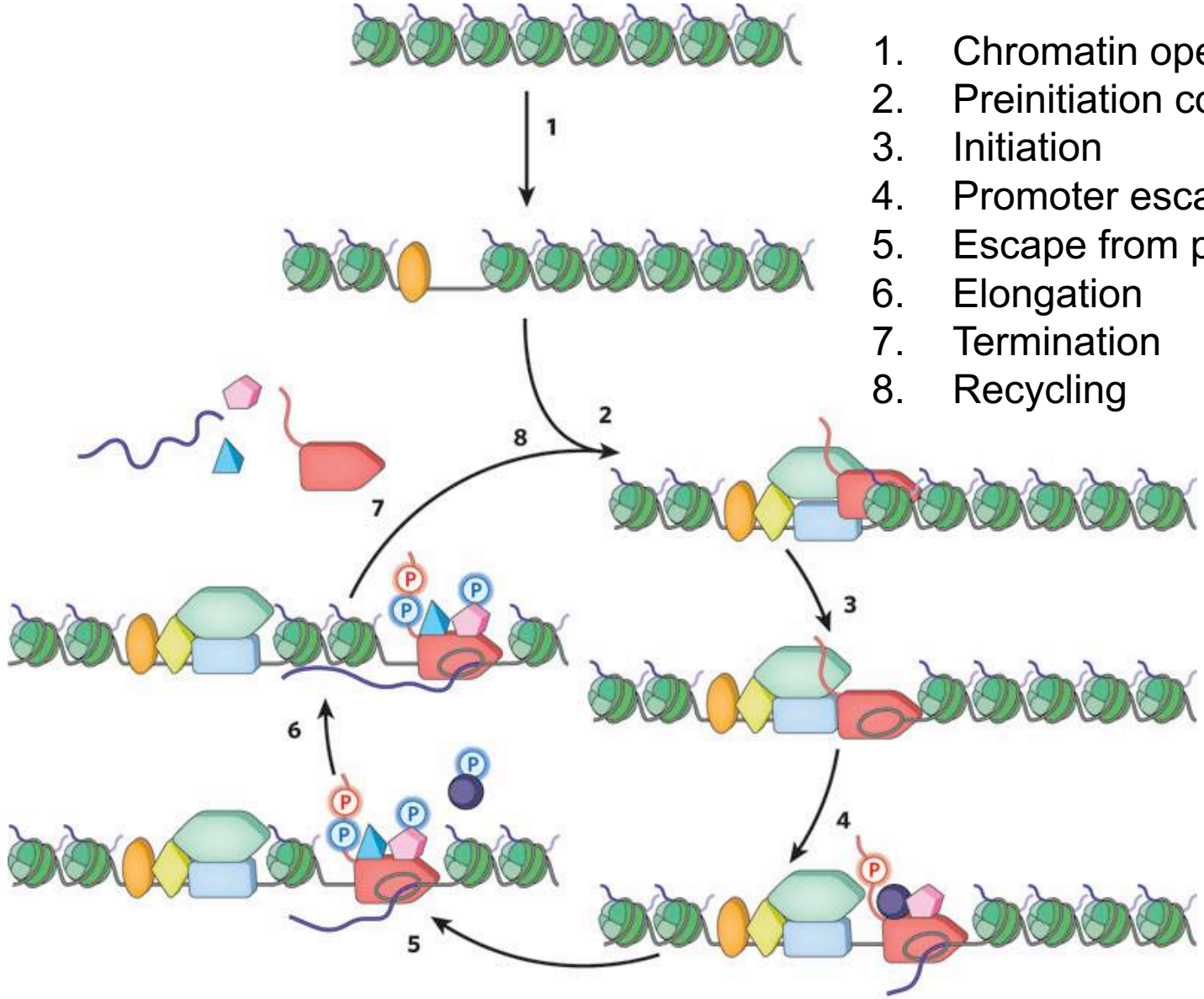


(Weake and Workman (2010) *Nat Rev Gen*)

The core promoter

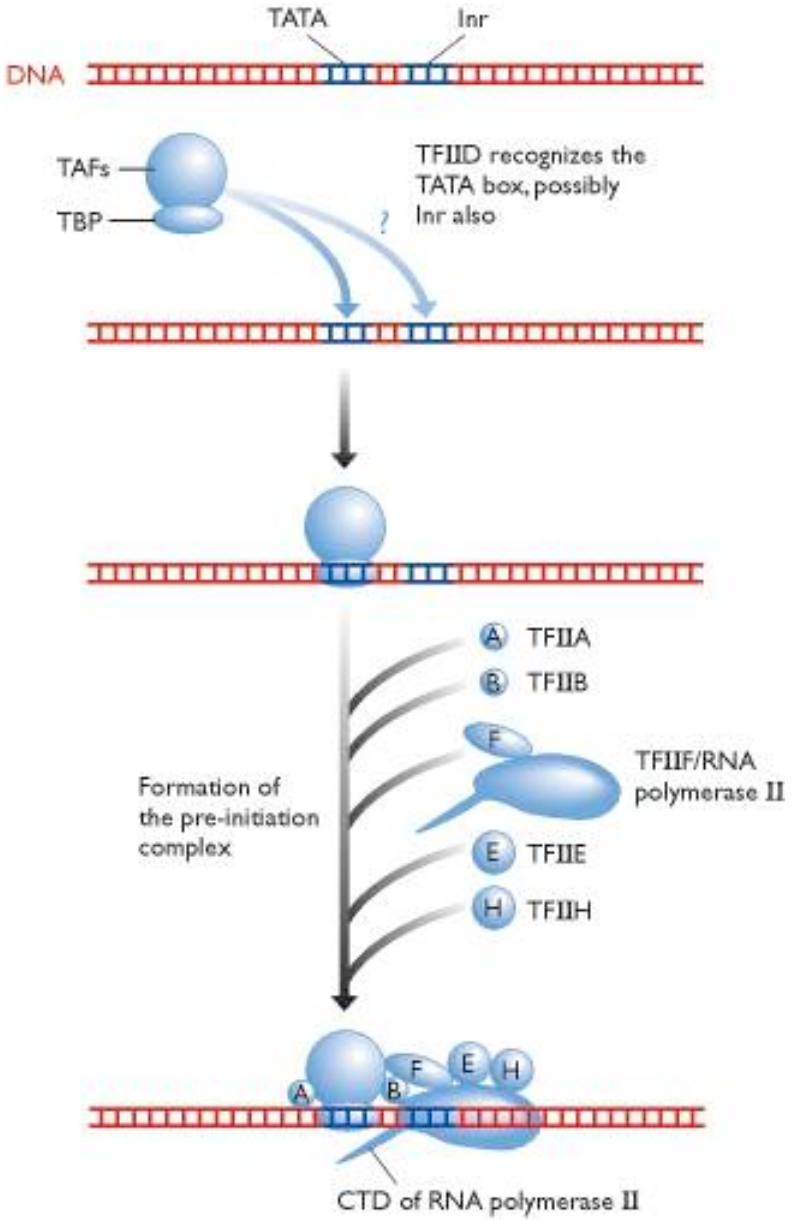


The transcription-cycle

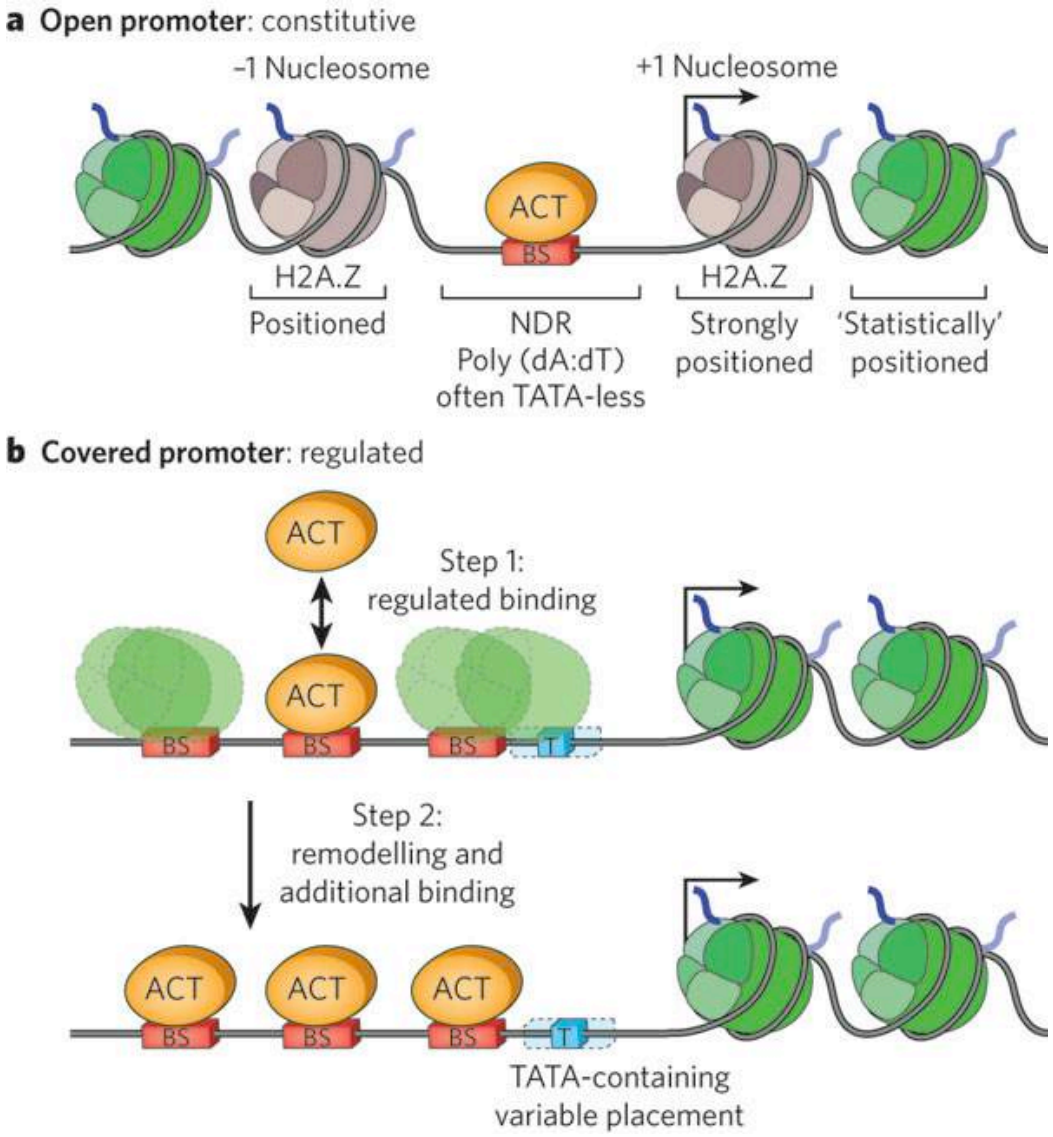


1. Chromatin opening
2. Preinitiation complex (PIC) forms
3. Initiation
4. Promoter escape
5. Escape from pausing
6. Elongation
7. Termination
8. Recycling

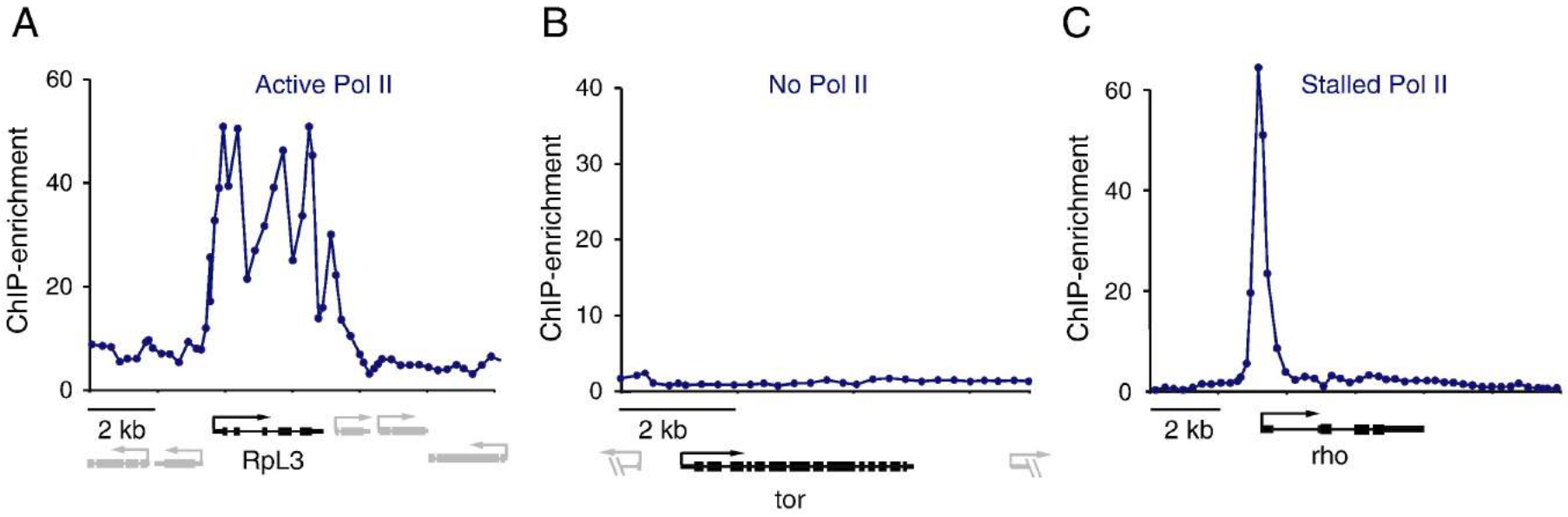
The assembly of the preinitiation complex (PIC)



Open and closed promoters

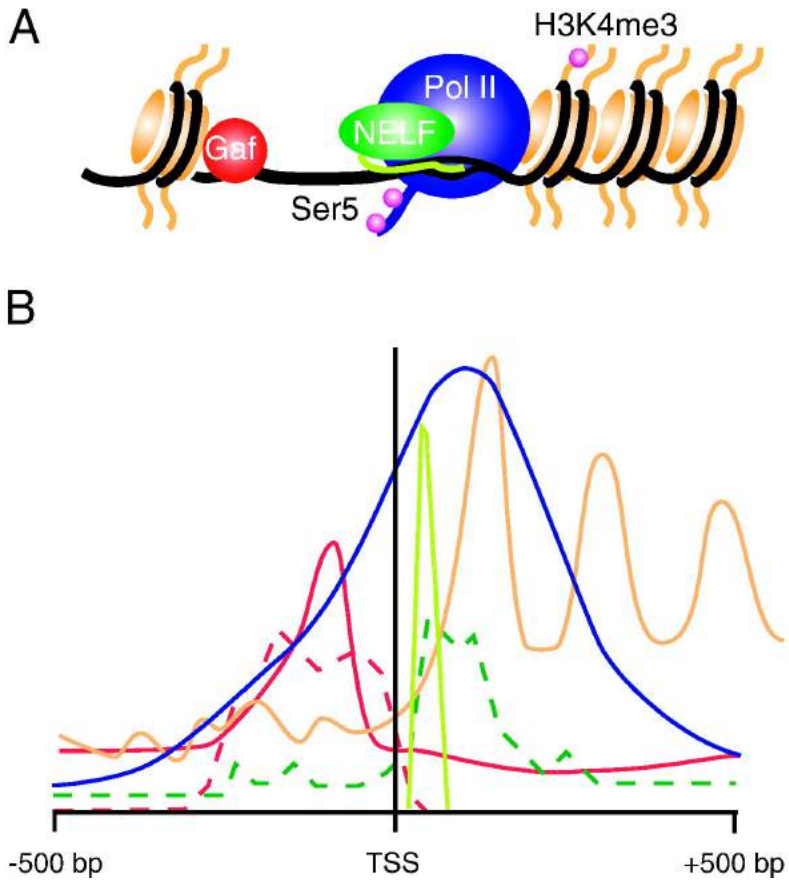


The regulation of transcription: different RNA pol II binding profiles



rpl13 = housekeeping gene (is transcribed continuously)
tor = not necessary for development
rho = developmental regulator

The regulation of transcription: the open promoter of developmental regulators



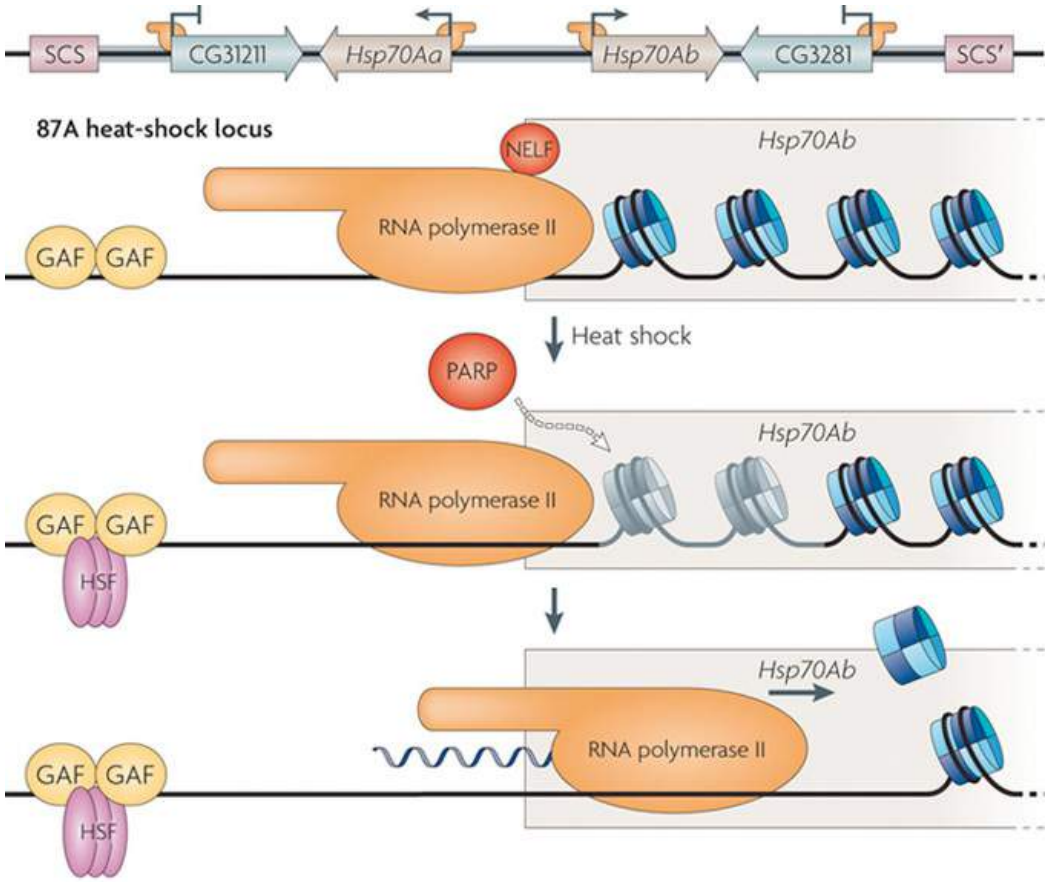
- developmental genes are under strict control

- the chromatin is open at these promoters and PolII can bind

- PolII is stalled, however, but can be released easily (by removing NELF), and transcription can commence

NELF = Negative ELongation Factor

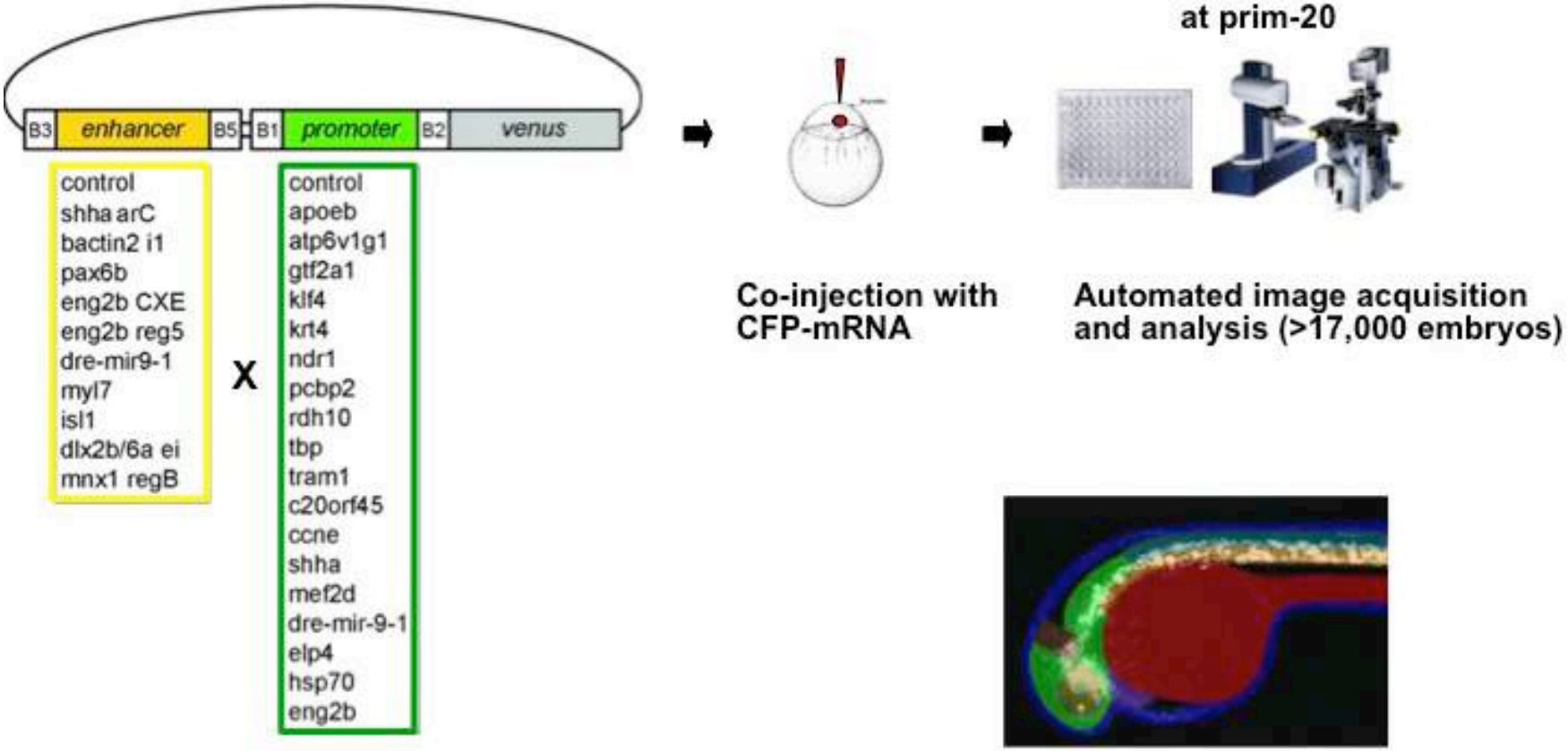
The regulation of transcription: the open promoter of heat-sock genes



NELF = Negative ELongation Factor

(Weake and Workman (2010) *Nat Rev Gen*)

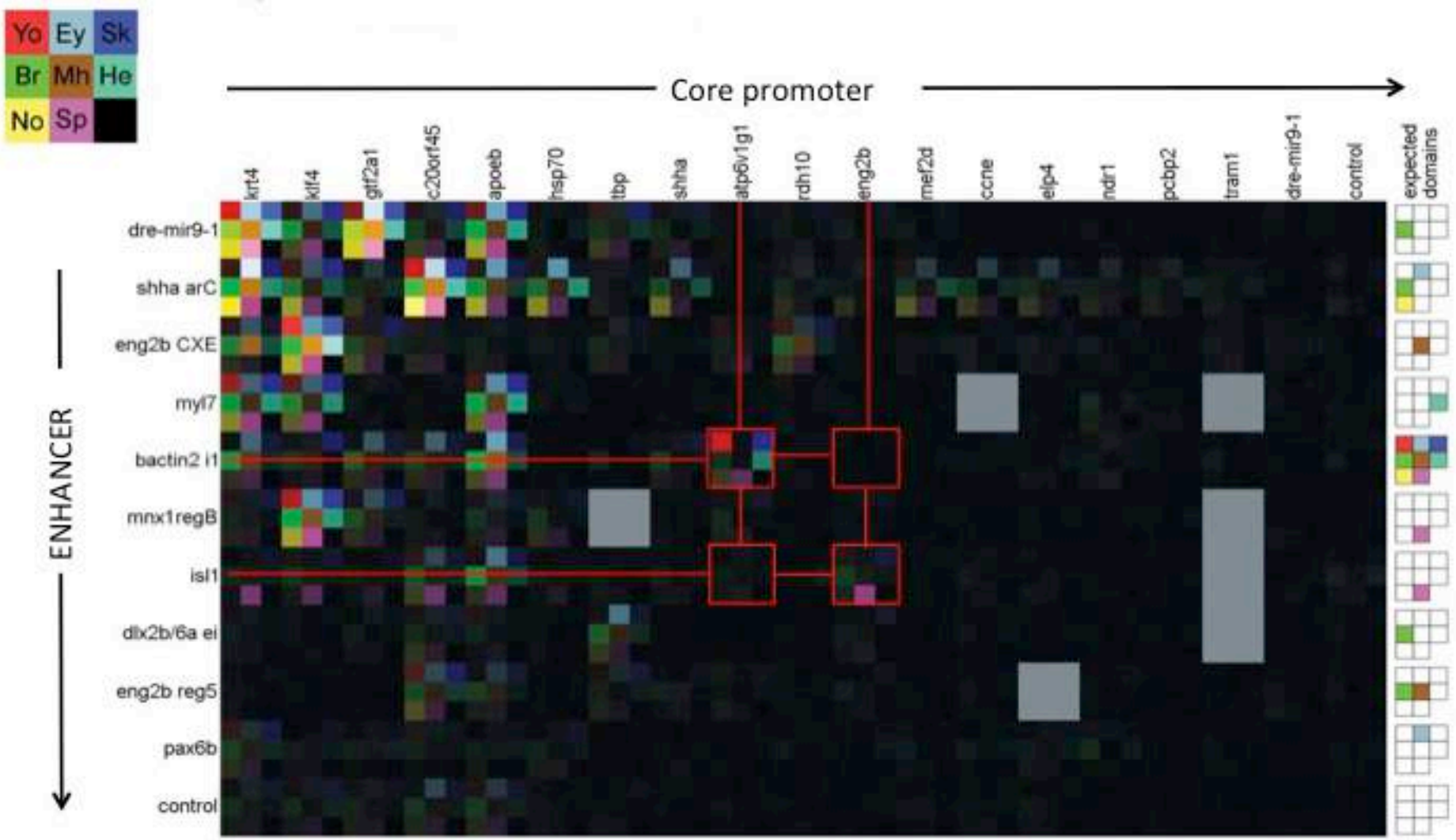
But is there really a “typical” core promoter?



> 200 promoter-enhancer combinations

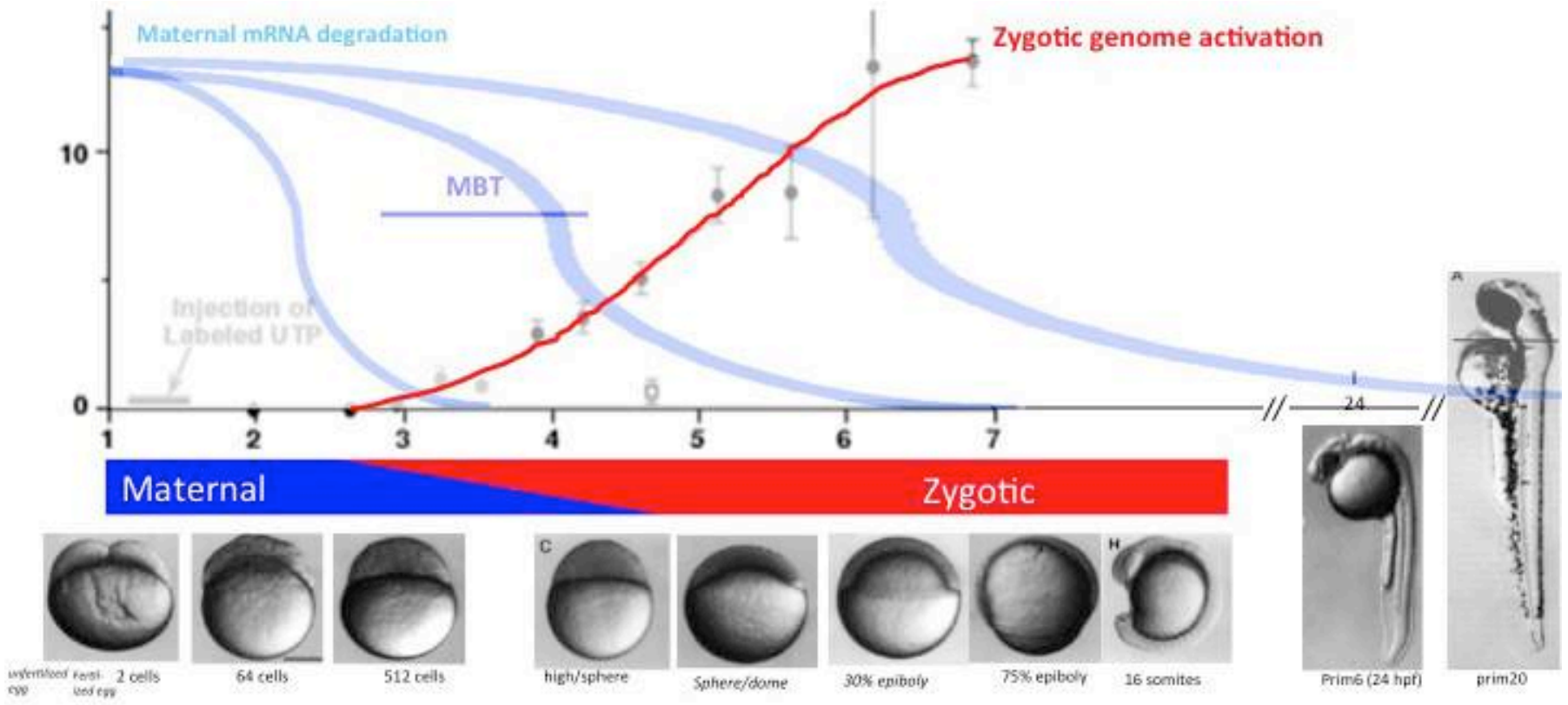
(Ferenc Müller)

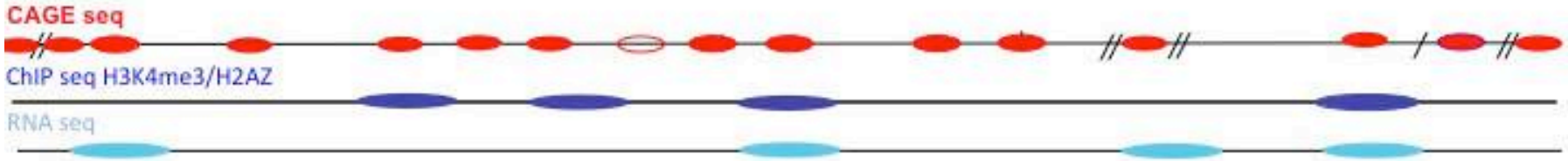
But is there really a “typical” core promoter?



(Müller Ferenc)

The activation of the zygotic genome



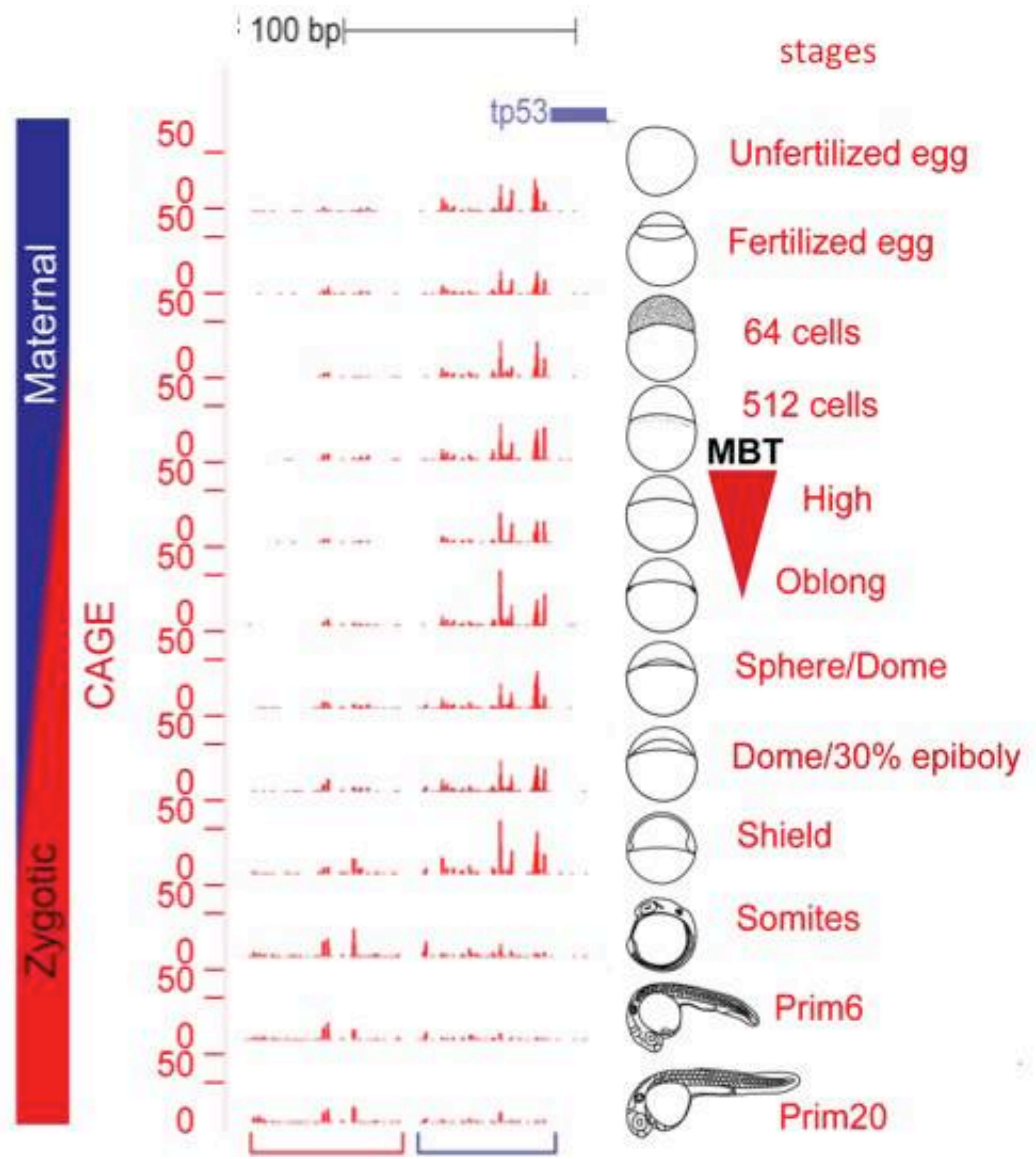


ZEPROME consortium: BHAM, ICL, RIKEN, UCL, KIT

(Ferenc Müller)

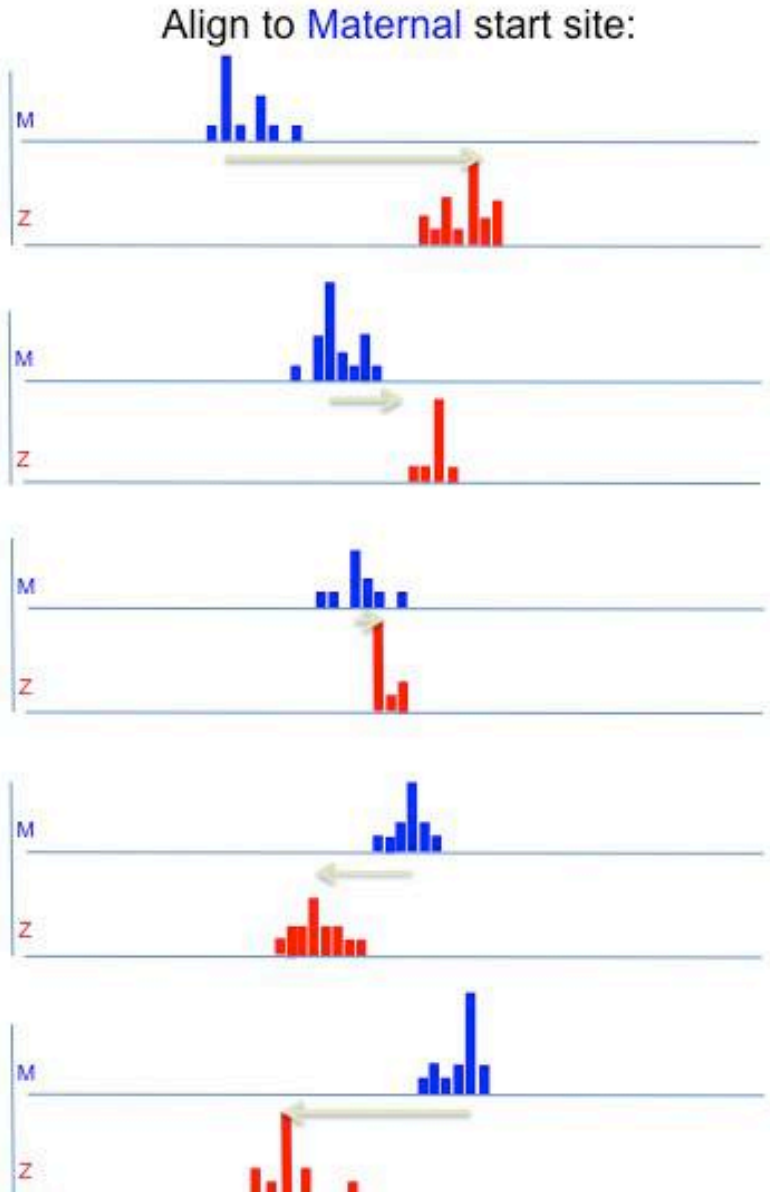
“Maternal” and “zygotic” promoters are different!

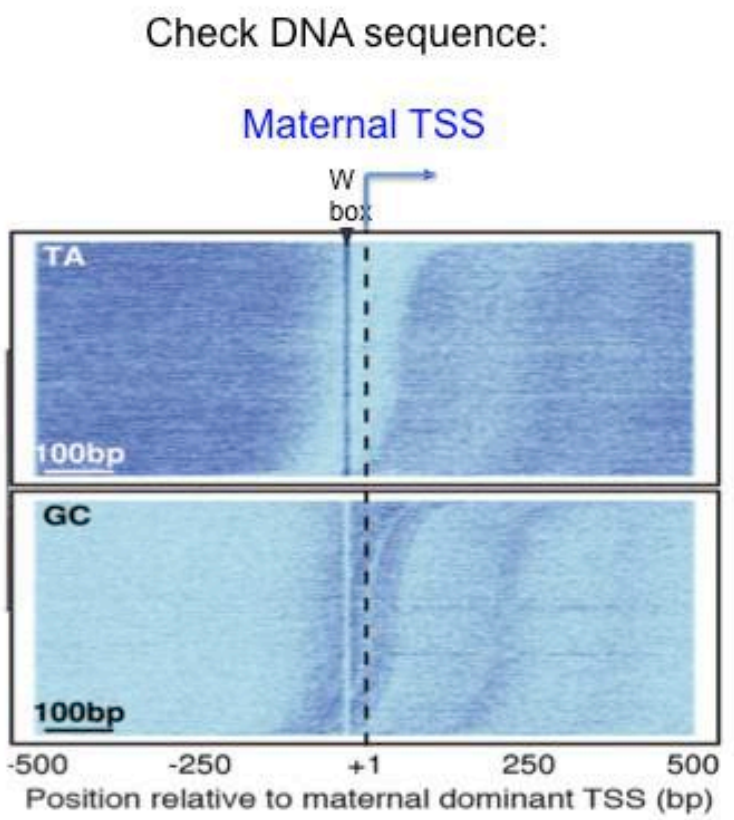
Cap-analysis gene expression (CAGE) – identifies the 5’ end of the mRNAs



(Ferenc Müller)

“Maternal” promoters are characterized by a W-box motif, whereas “zygotic”ones are GC-rich

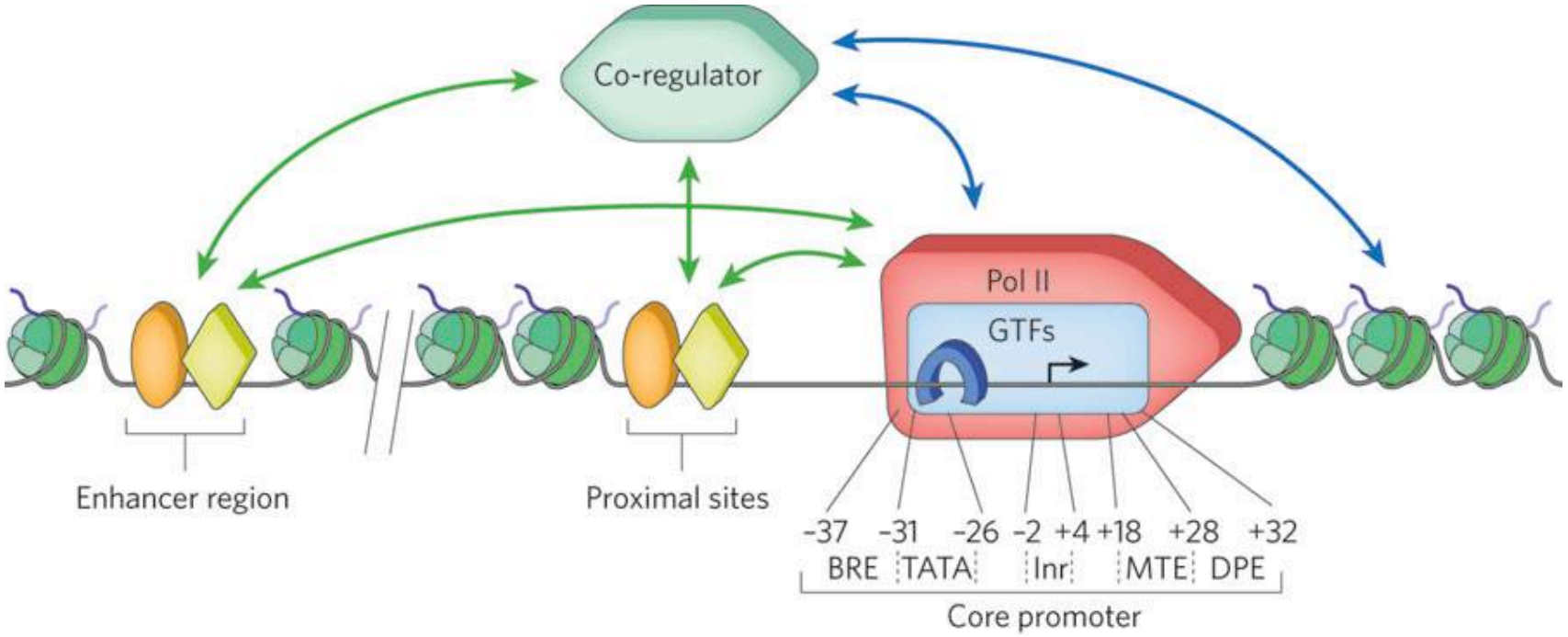




Vanja Haberle, Yavr Hadzhiev, Nan Li, Boris Lenhard

(Ferenc Müller)

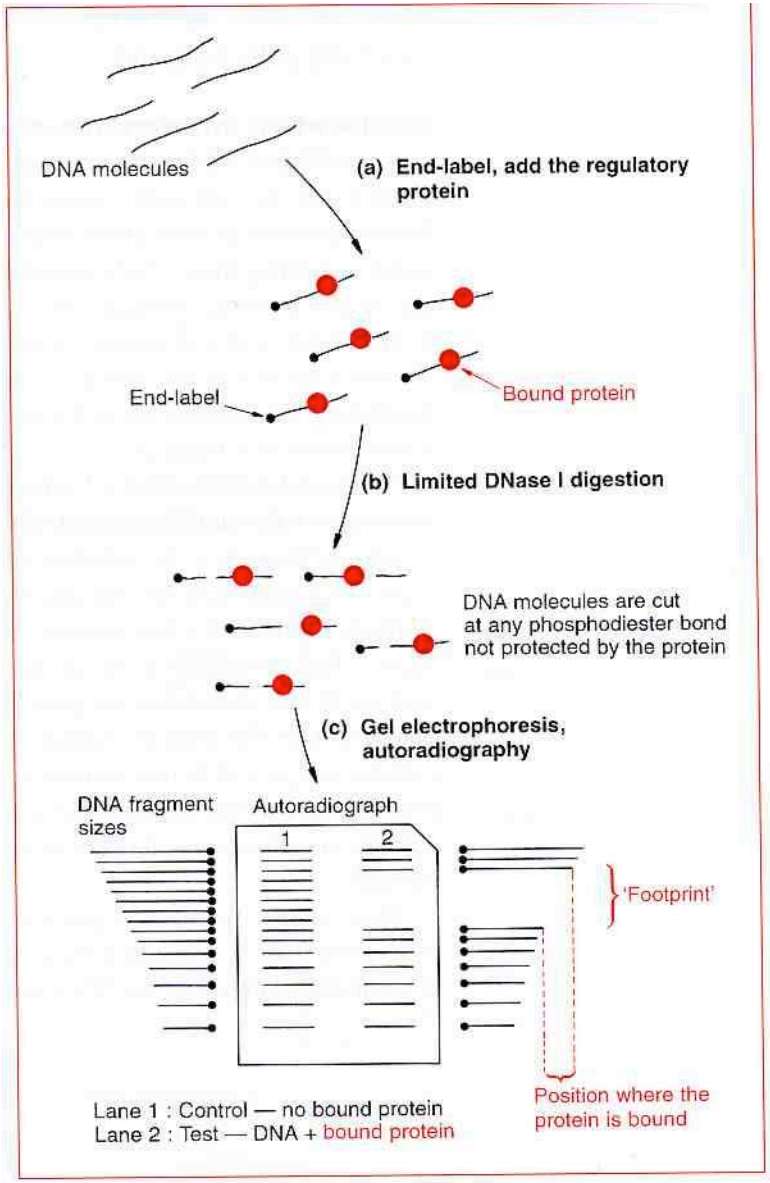
Interactions regulating transcription



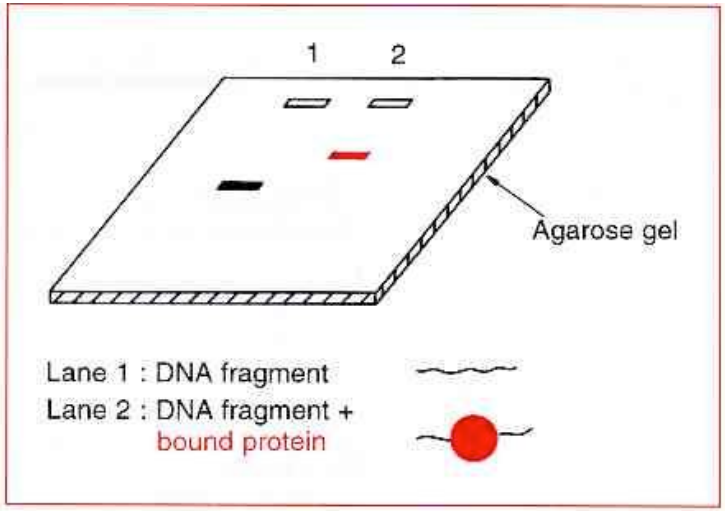
GTF = General Transcription Factors

Assays to examine TF binding

1. - DNase footprinting



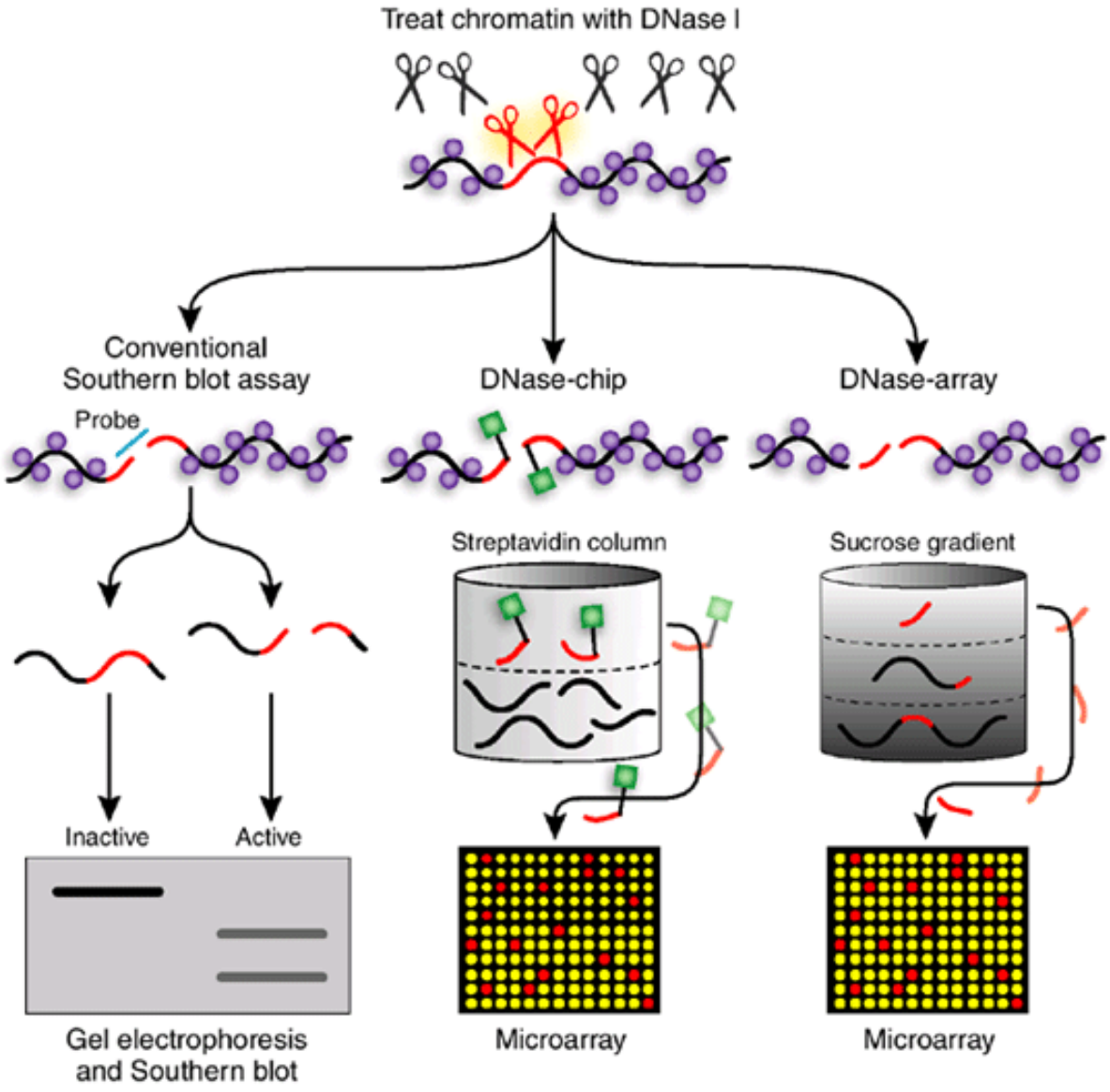
2. - EMSA/Band shift/Gel shift assay



3. - Chromatin immunoprecipitation (ChIP)

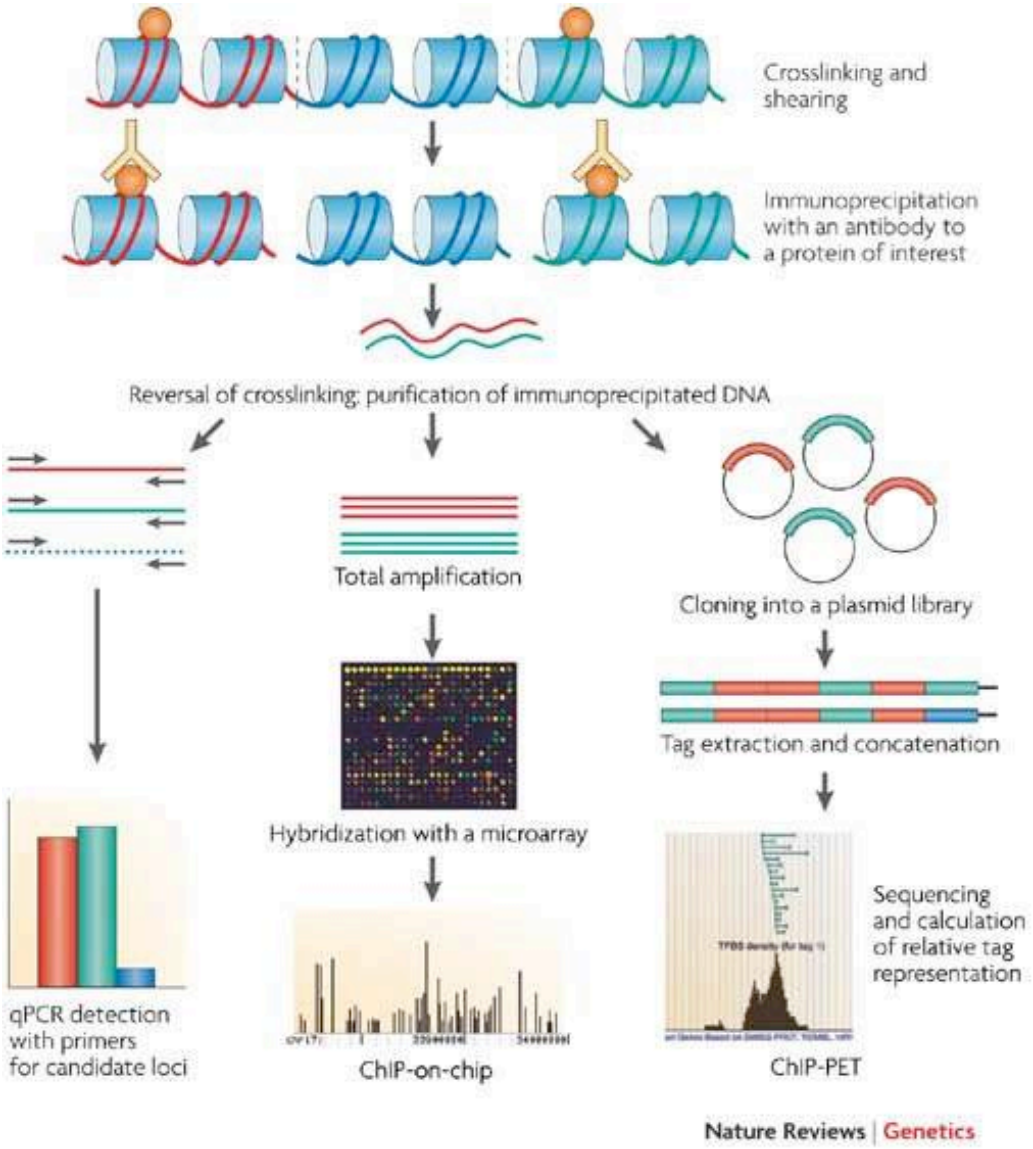
(http://bioweb.wku.edu/courses/biol350/Transcriptome17/Review.html)

Chromatin states (e.g. nucleosome and TF binding) can be examined using DNase hypersensitivity assays



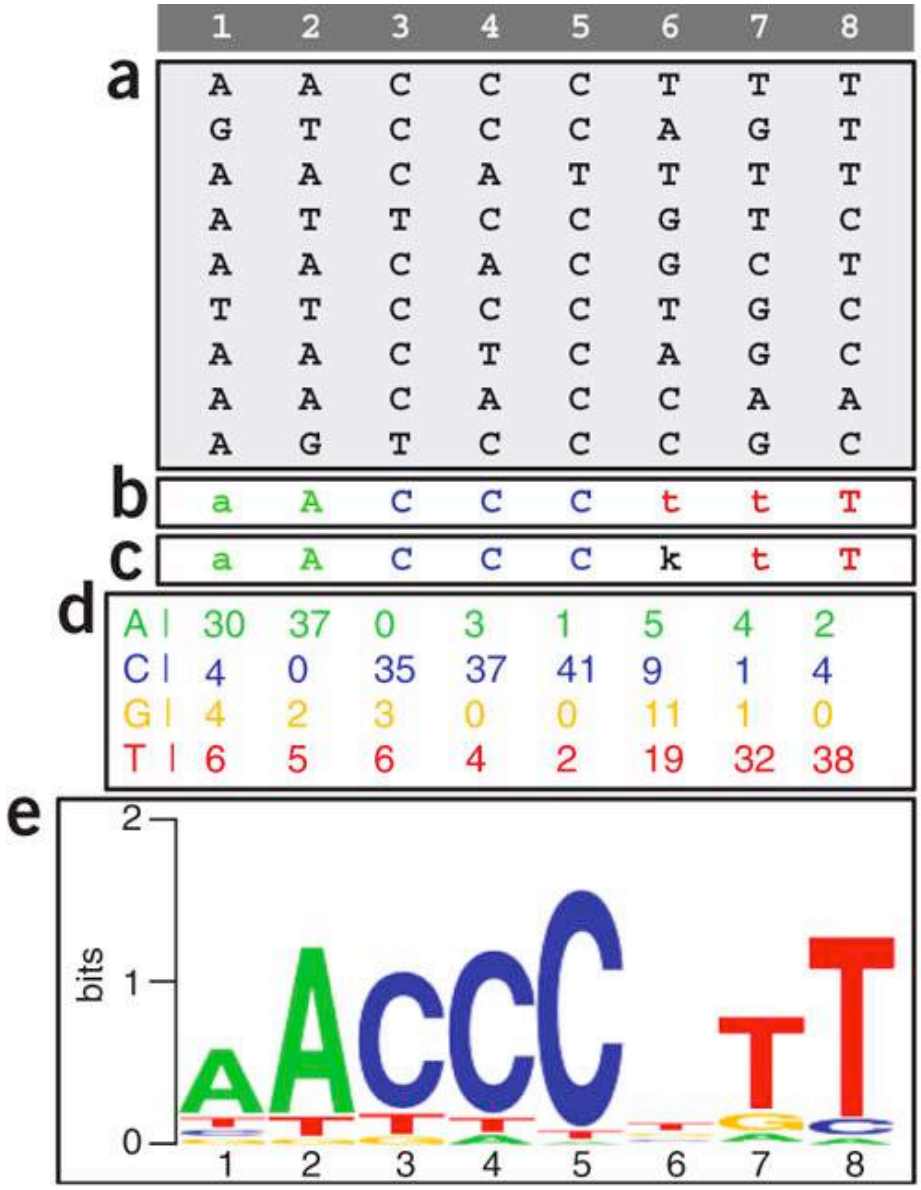
Katie Ris

Variations on chromatin-immunoprecipitation (ChIP)



(Spivakov and Fisher (2007) *Nature*)

Consensus TF binding sites



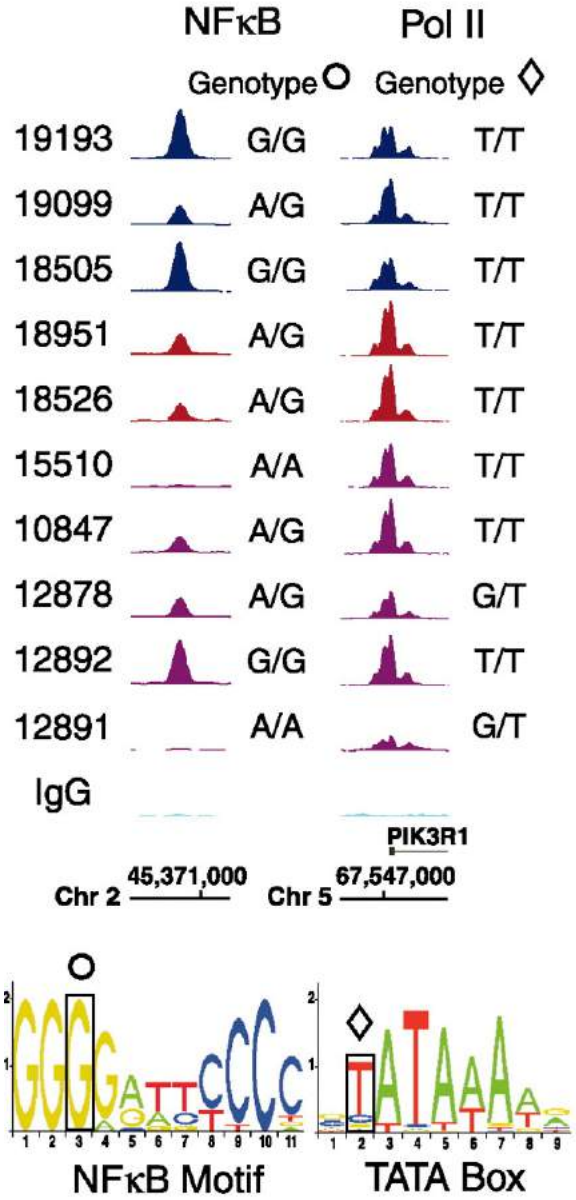
- a few examples for the binding sites of the *Drosophila* Krüppel TF

- strict consensus
- degenerated consensus
- PSSM (Position-Specific Scoring Matrix)

-Sequence Logo
(e.g.:
http://weblogo.berkeley.edu/logo.cgi)

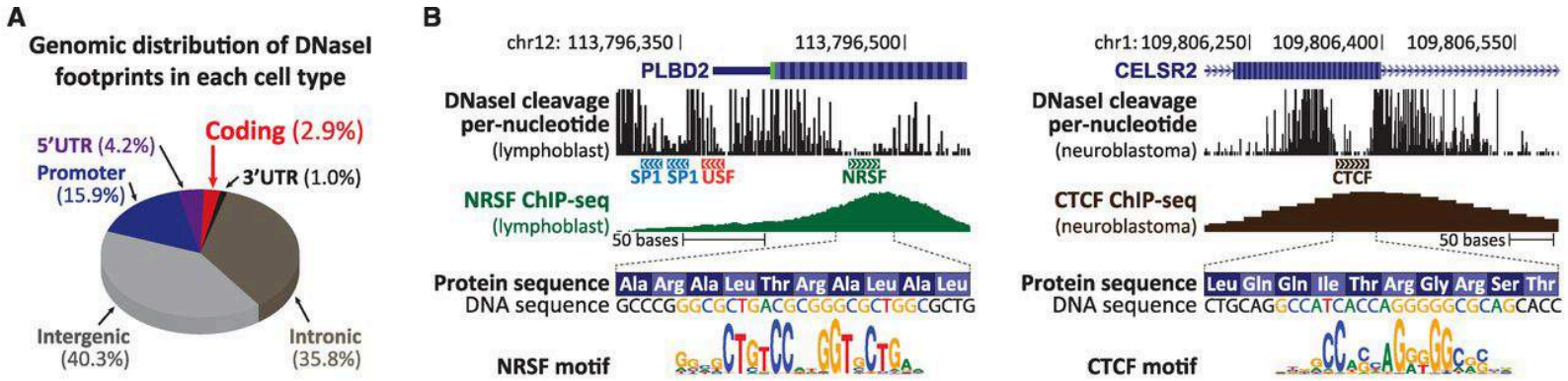
(Turatsinze et al. (2008) *Nat Prot*)

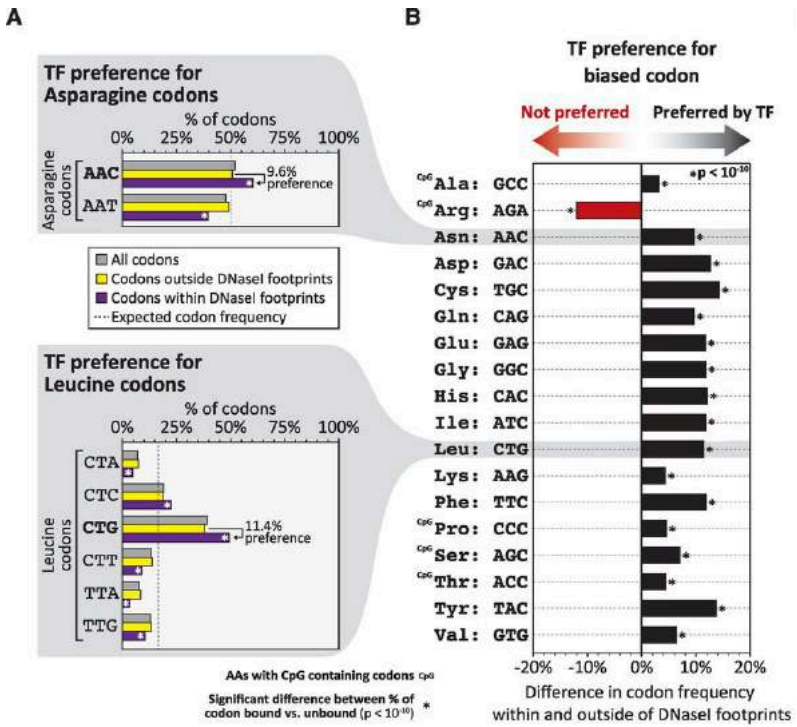
Smaller changes in consensus TF-binding sites can have dramatic transcriptional consequences



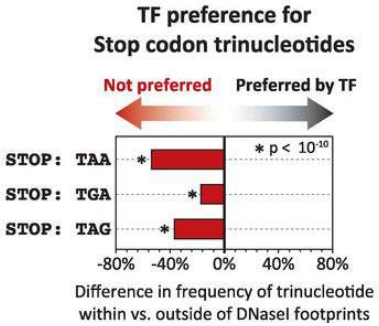
(Kasowski et al. (2010) *Science*)

Duons – triplets in the CDS that also bind TFs





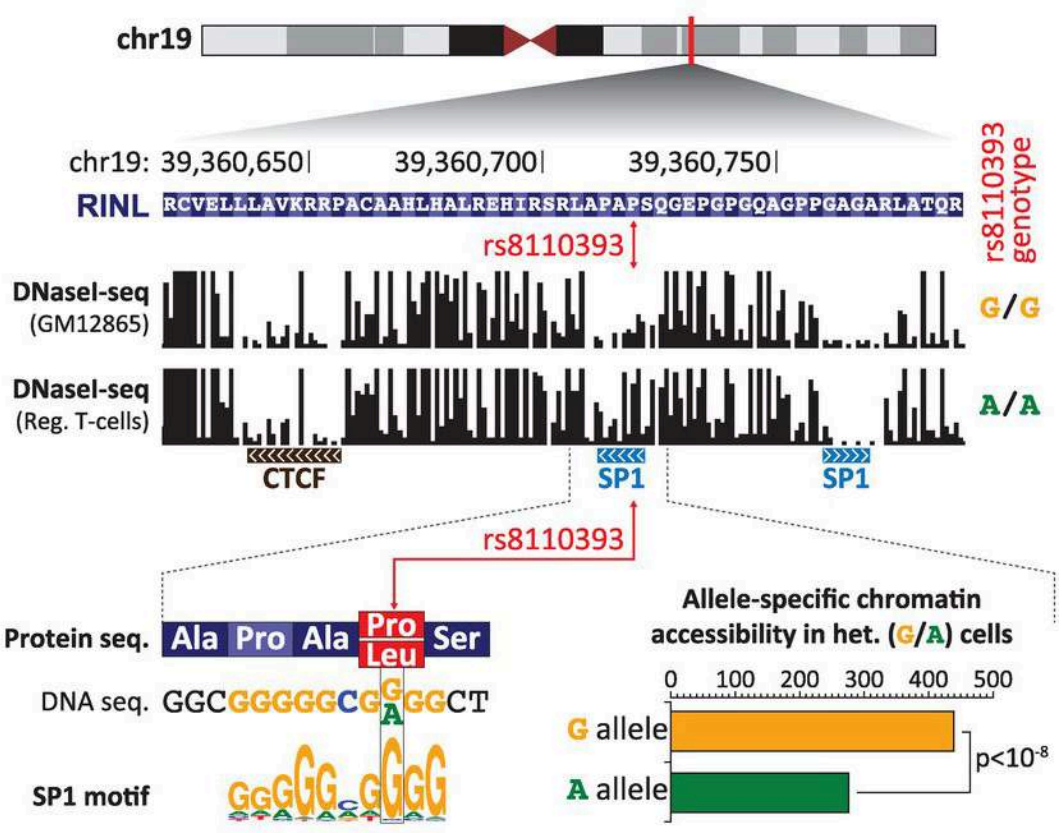
- Some triplets are favoured because of TF binding site conservation



… STOP codons are underrepresented in TF binding sites

(Stergachis et al. (2013) *Science*)

Mutations in duons chance aminoacid sequence AND alter TF-binding

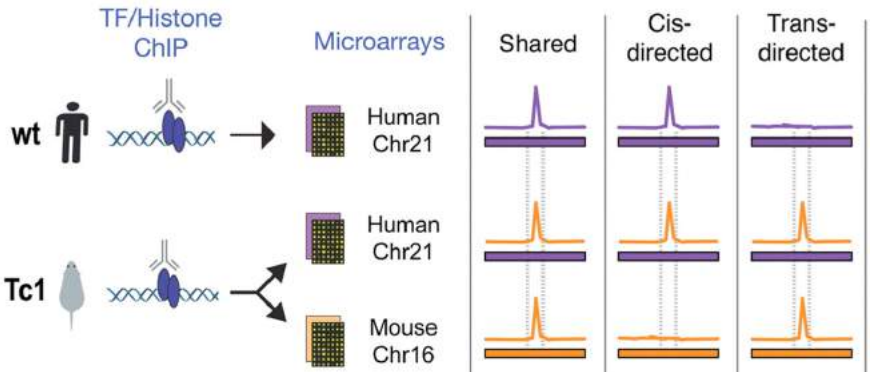


(Stergachis et al. (2013) *Science*)

TF-binding site turnover is high

- There are big differences in the TF-binding sites of orthologous sequences in functionally conserved liver cells

Chromatin IP			
HNF1α	76	38	300
HNF4α	137	119	146
HNF6	28	60	115



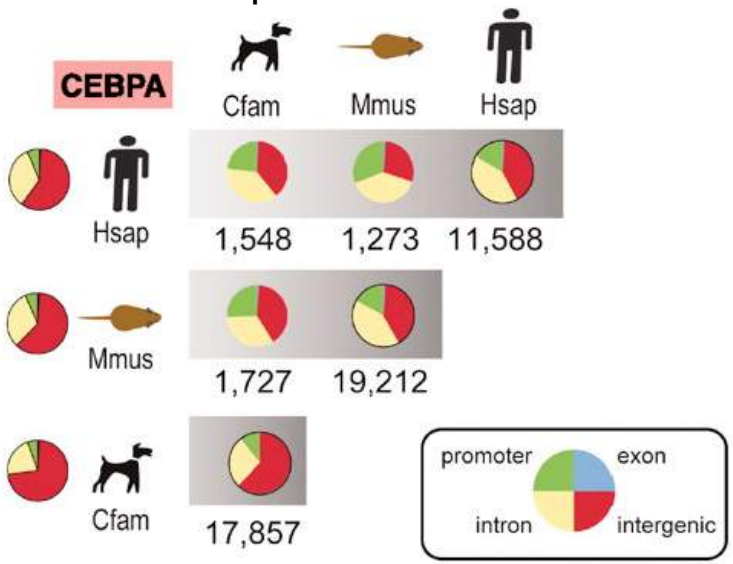
	Shared	Cis-directed	Trans-directed
HNF1α	38	39	1
HNF4α	58	93	4
HNF6	11	36	0

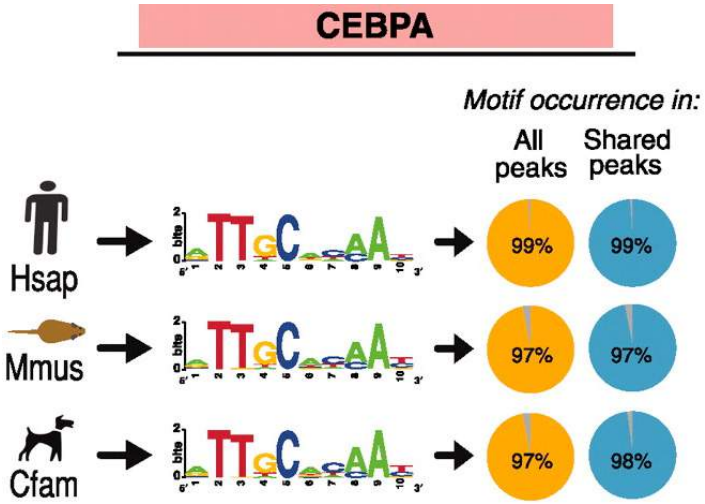
- Most of these differences are genetic in their origin as a piece of a human chromosome will have a similar TF-binding profile in a mouse to its endogenous binding profile

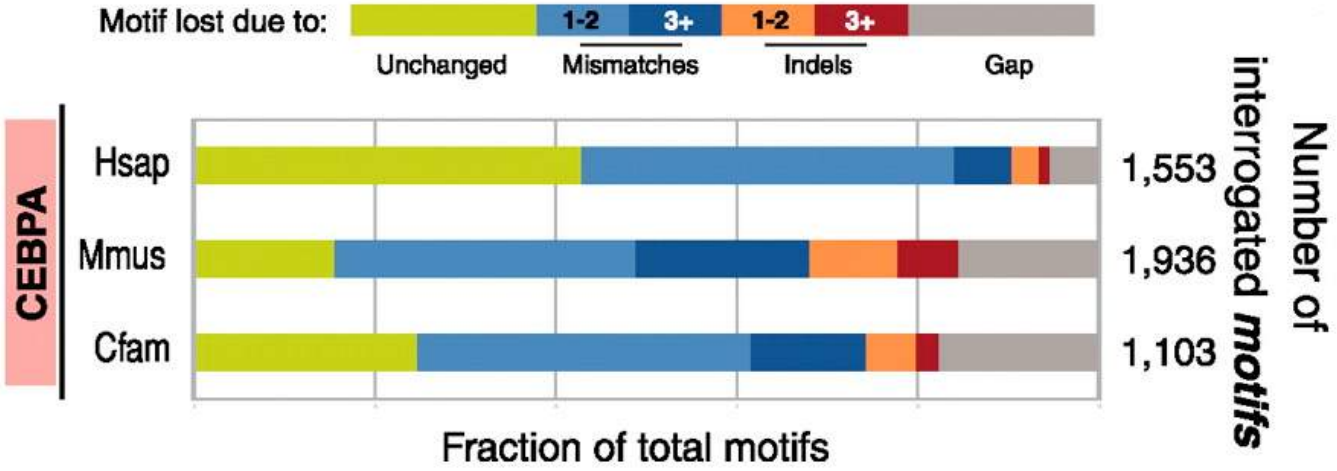
(Wilson et al. (2008) *Science*)

Conserved binding site does not mean binding!

CEBPA: liver specific TF

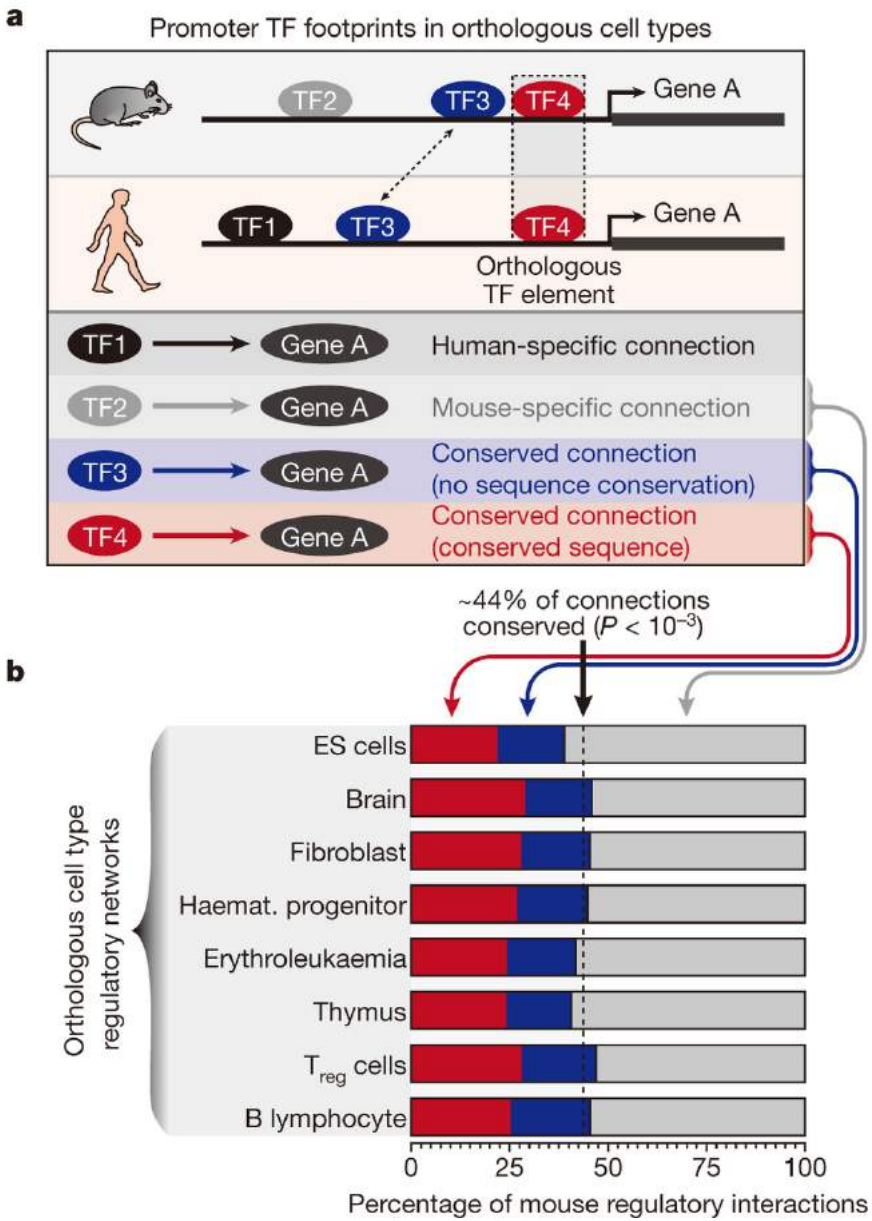






(Schmidt et al. (2010) *Science*)

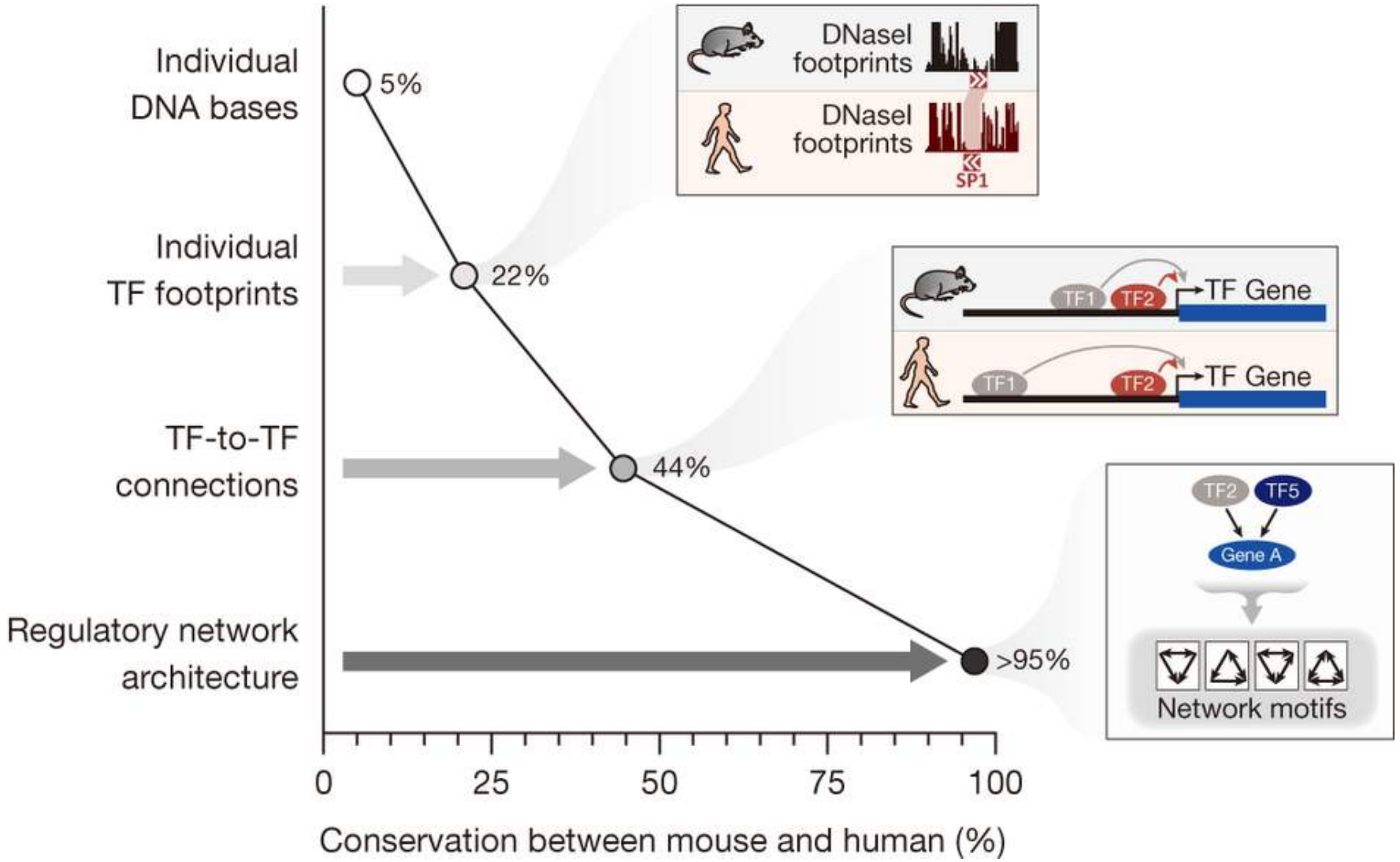
High TF-binding site turnover is present in many cell types



- In average half of the TF binding sites in every tissue are species-specific, and even in the case of “conserved” sites, for many absolute position is not conserved

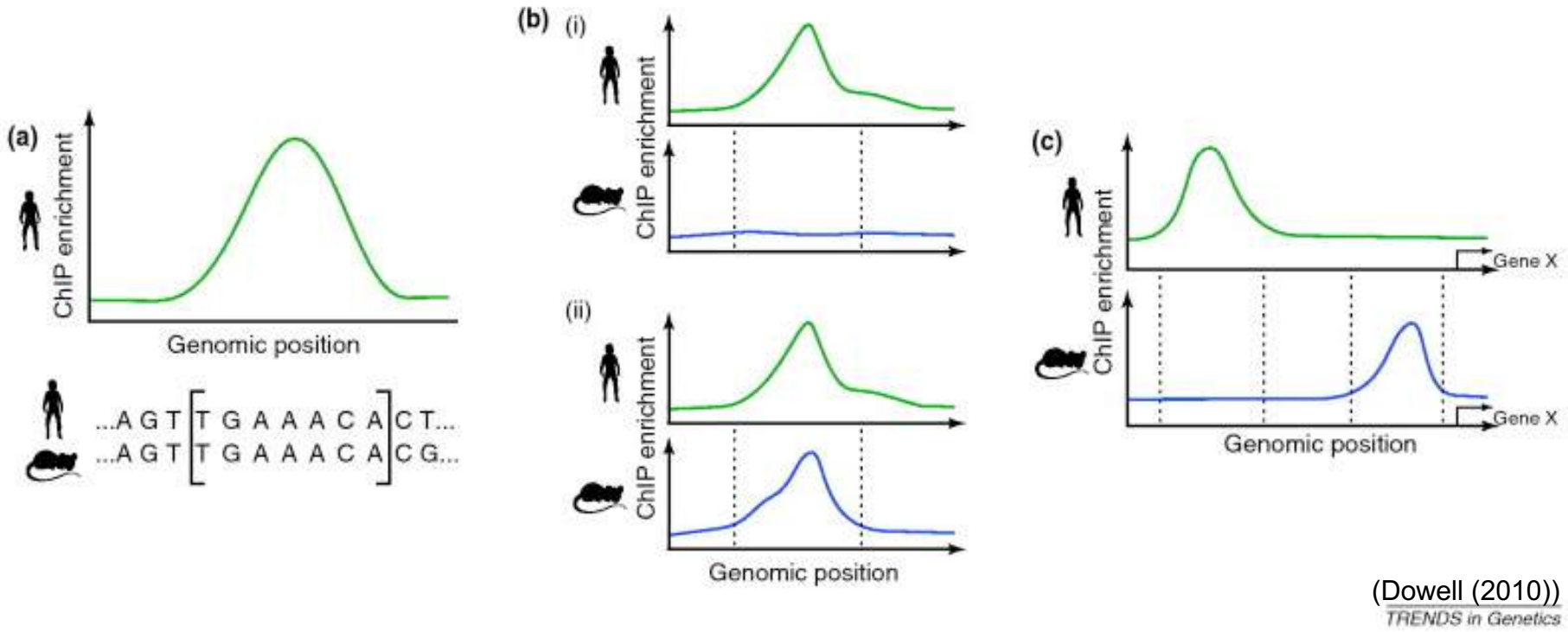
(Stergachis et al. (2014) *Nature*)

The hierarchy of cis-regulatory conservation



(Stergachis et al. (2014) *Nature*)

TF-binding site summary



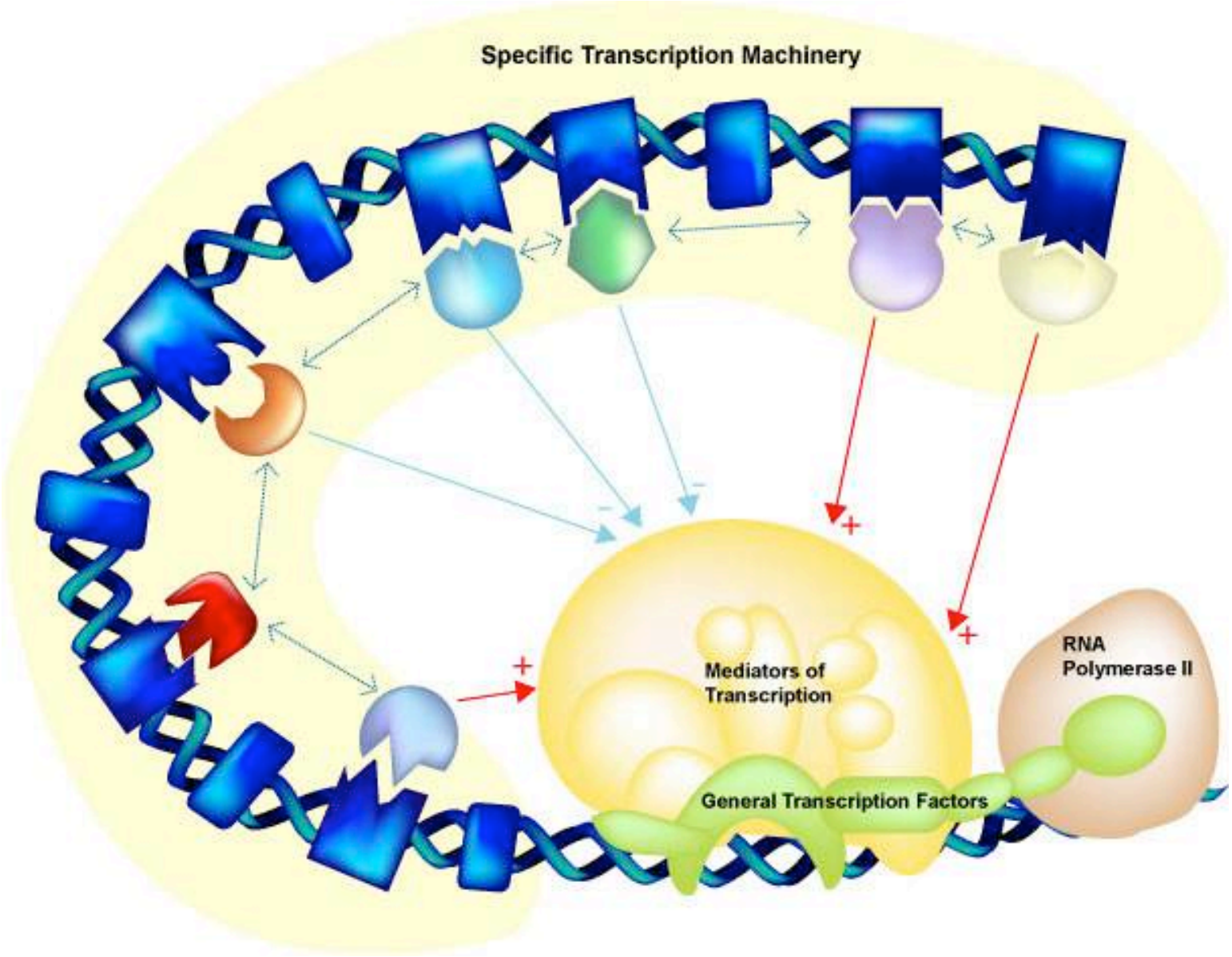
(Dowell (2010))
TRENDS in Genetics

1. - a conserved binding site can be used to suspect conserved DNA-protein interaction

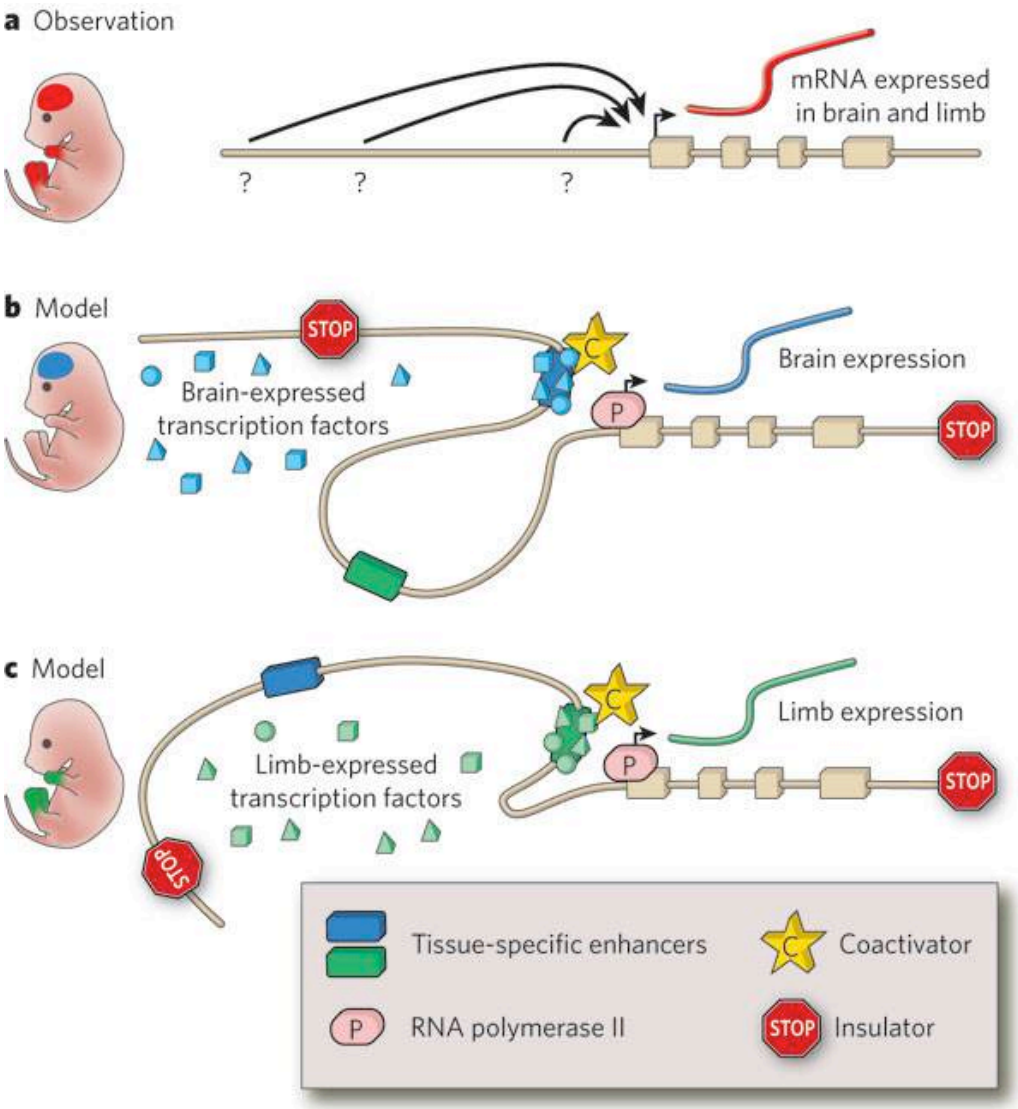
2. - but it is NOT proof (proximal sequences can change so the TF looses physical access to the binding site)

3. - even if a TF has a conserved role in regulating one particular gene, that does not mean it will have a conserved bindig-site: TF binding site turnover is very high (in the genes of liver-specific enzymes only 7-48% of binding sites are conserved)

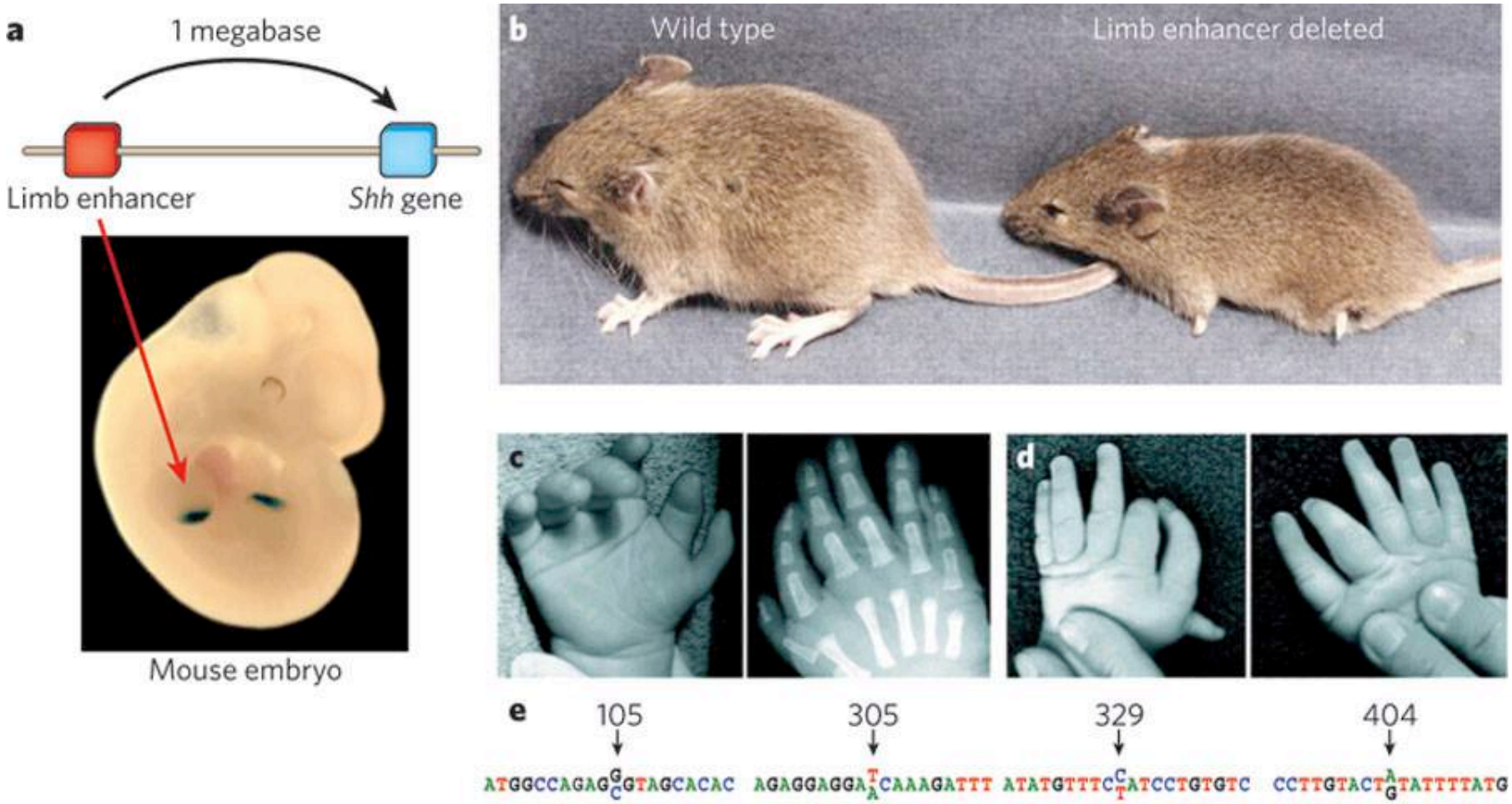
The regulation of transcription: enhancers



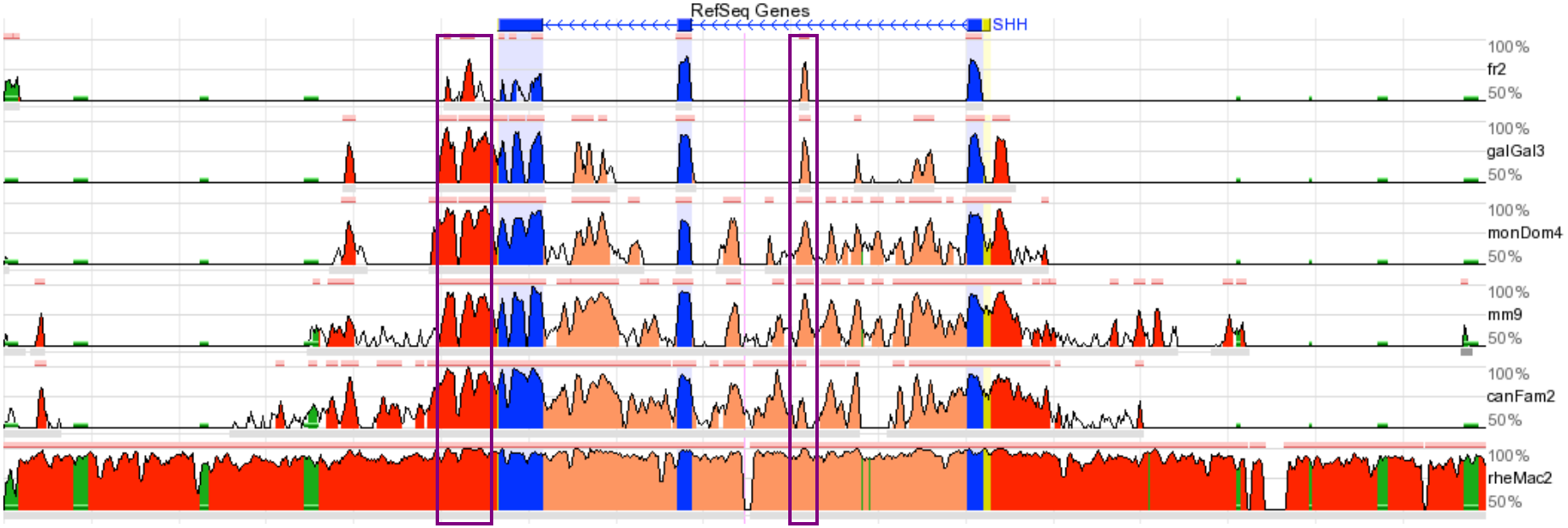
Long range enhancers



Long range enhancers: the *Shh* gene

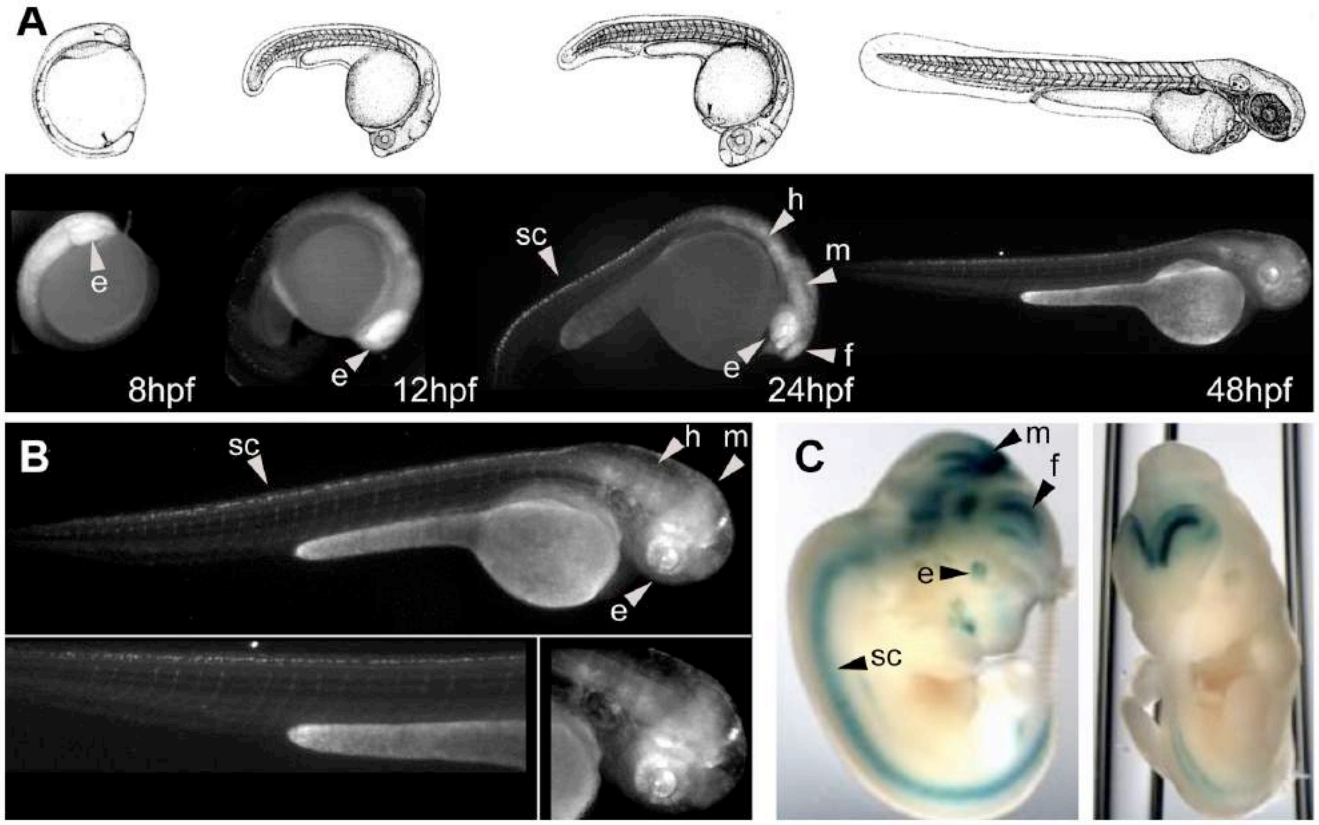


Conserved Non-coding Elements (CNE)



CNE: DNA pieces of several hundred basepair that show higher (!!) conservation that protein coding sequences

In transgenic reporter assays CNEs act as conserved enhancers

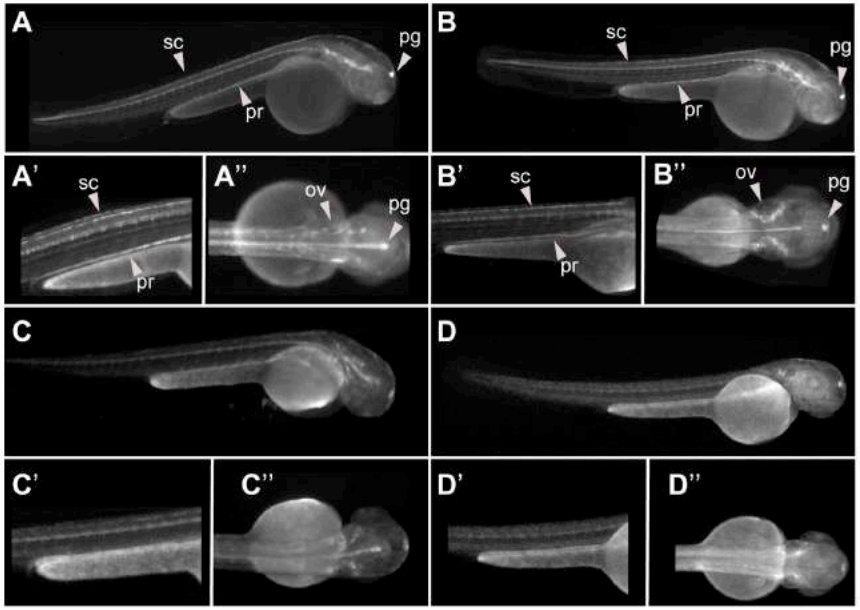


HCNR C81 from human chromosome 16 shows similar enhancer activity in zebrafish and mice

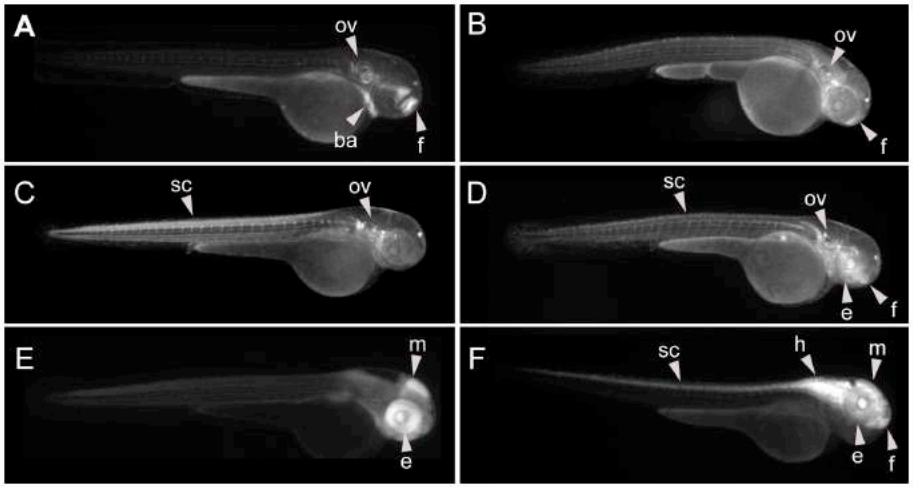
(Royo et al. (2011) *PLoS One*)

De nincs egyértelmű funkció ami CNE-hez rendelhető

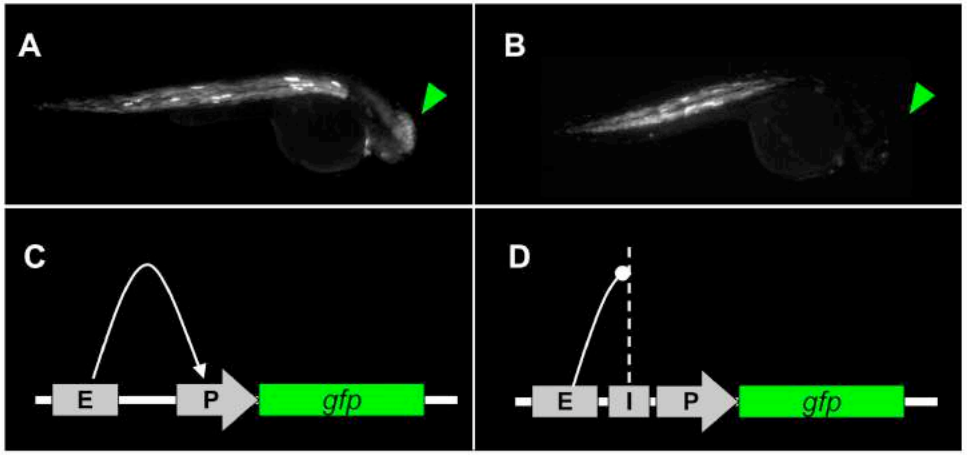
Reprodukálható enhancer: HCNR C32



Genomi régió függő HCNR C60

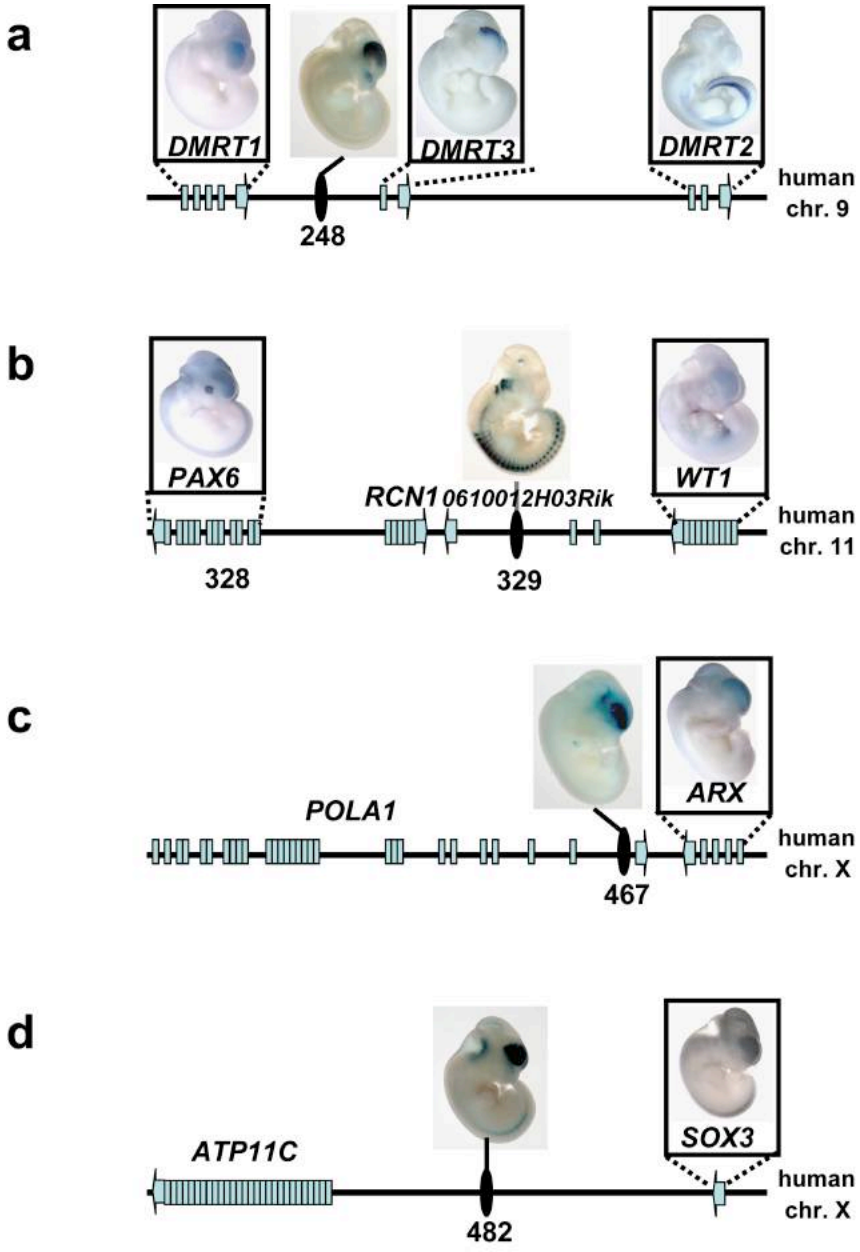


Inzulátor: HCNR C91



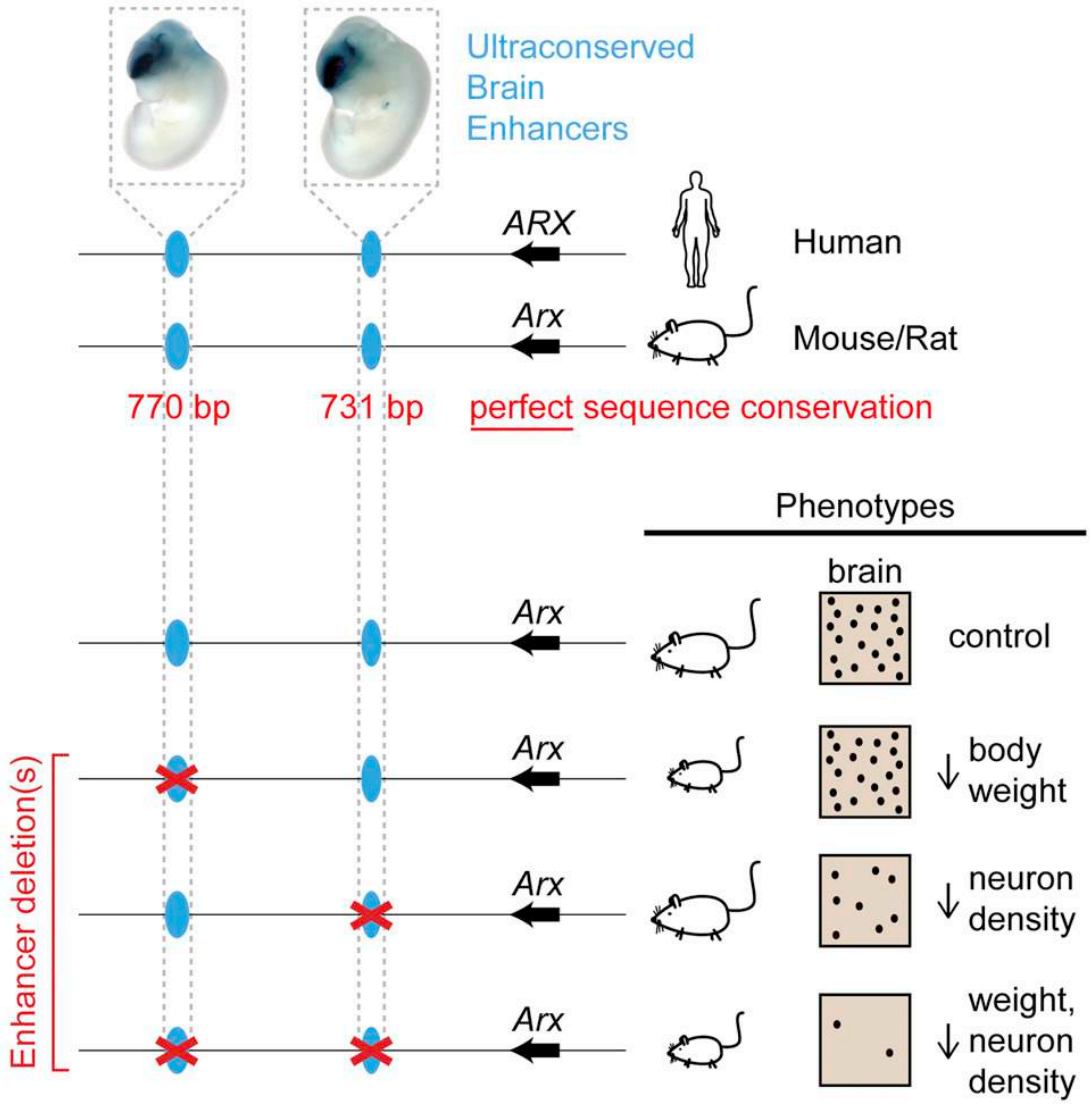
(Royo et al. (2011) *PLoS One*)

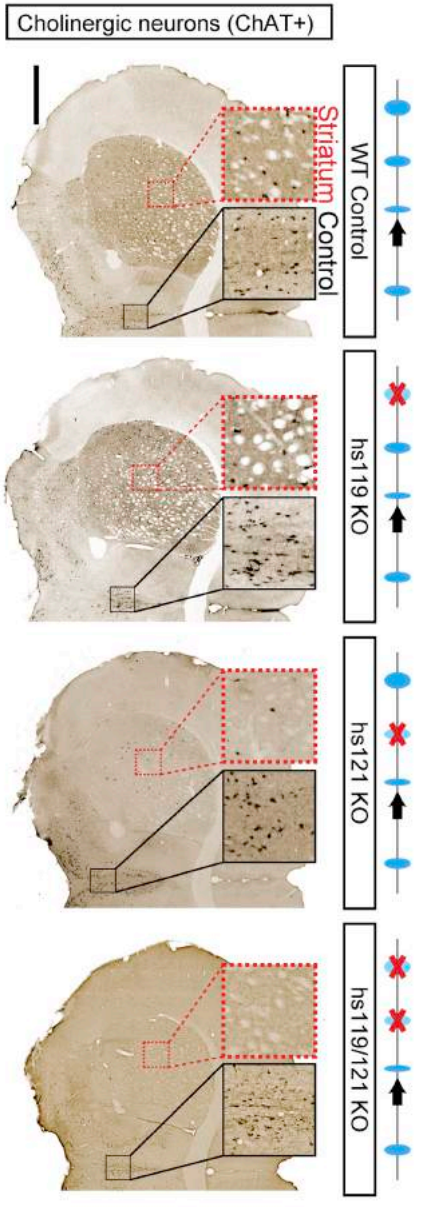
Importantly, the deletion of the CNEs is not lethal…



(Ahituv et al. (2007) *PLoS Biol*)

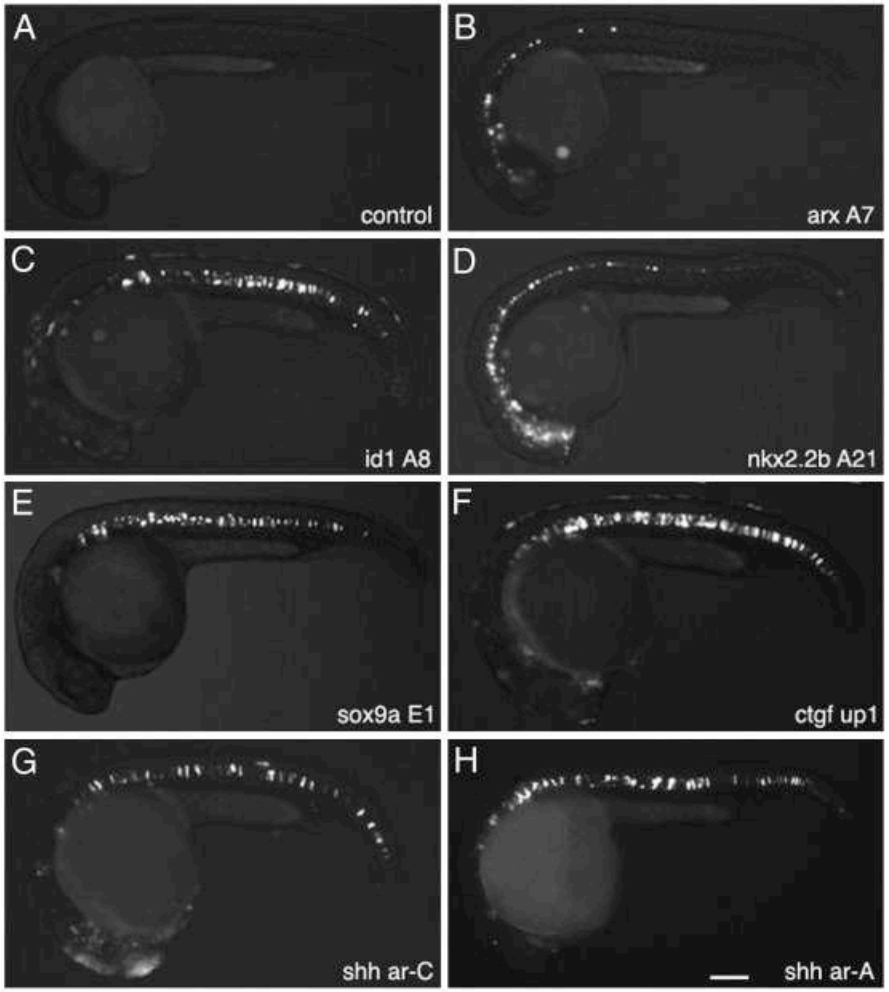
... it still has consequences, though

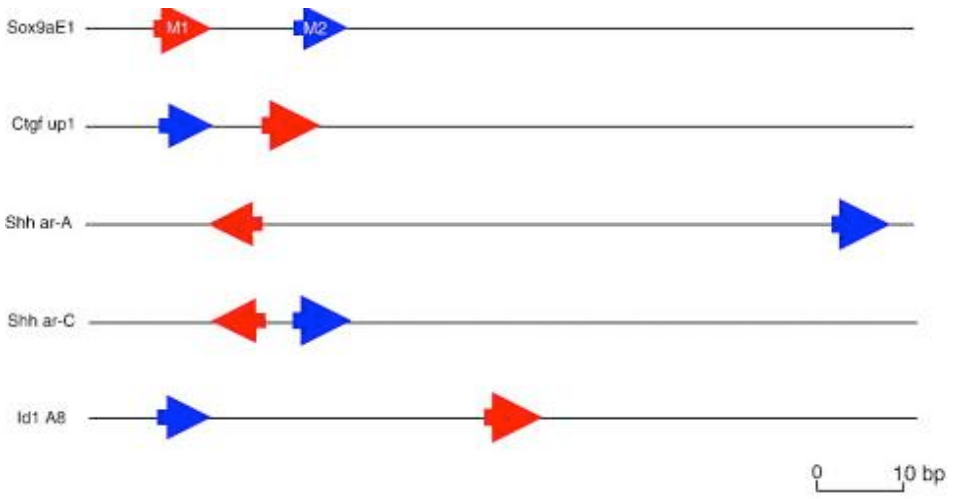




(Dickel et al. (2018) *Cell*

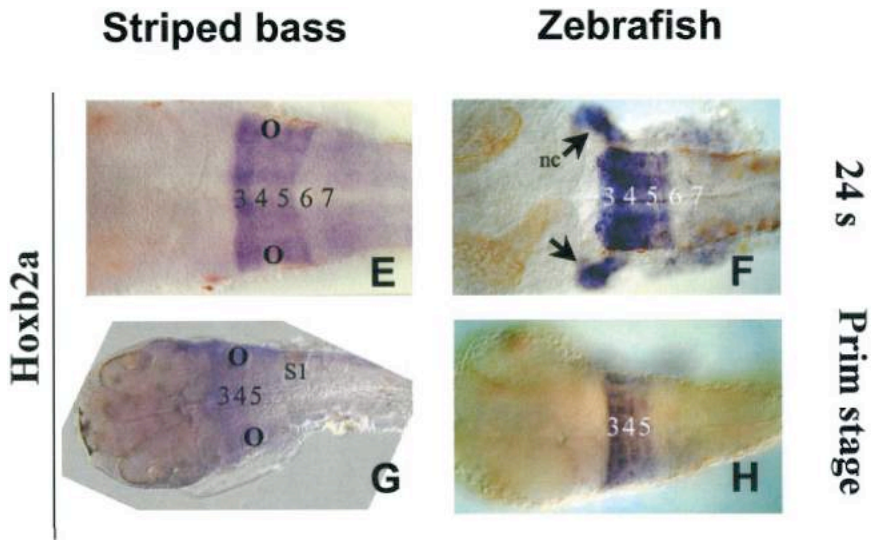
The position of essential TF-binding sites is not conserved in homologous CNEs

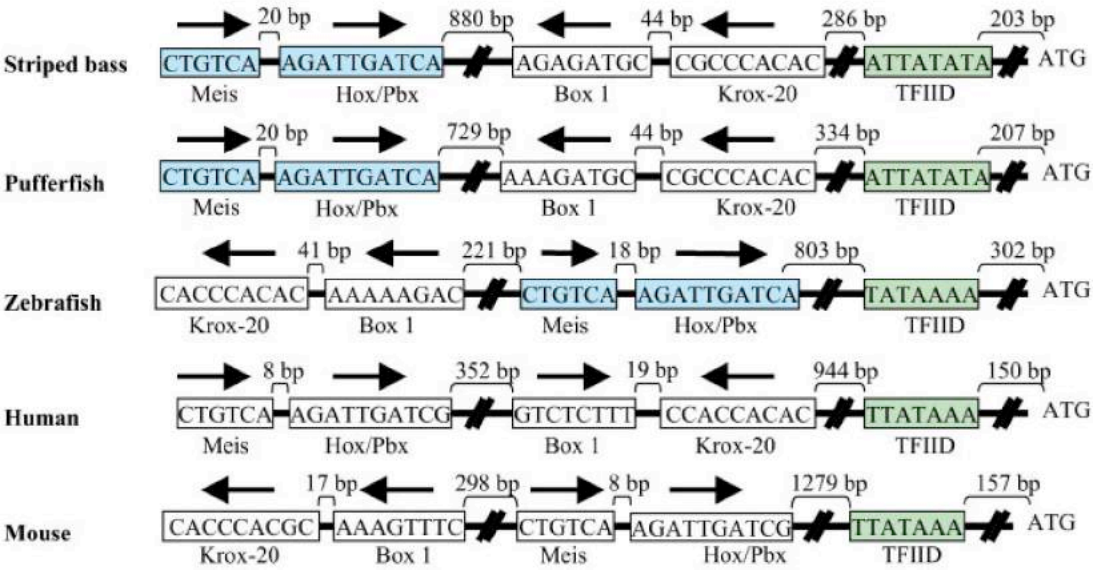




(Rastegar et al. (2008) *Dev Bio*)

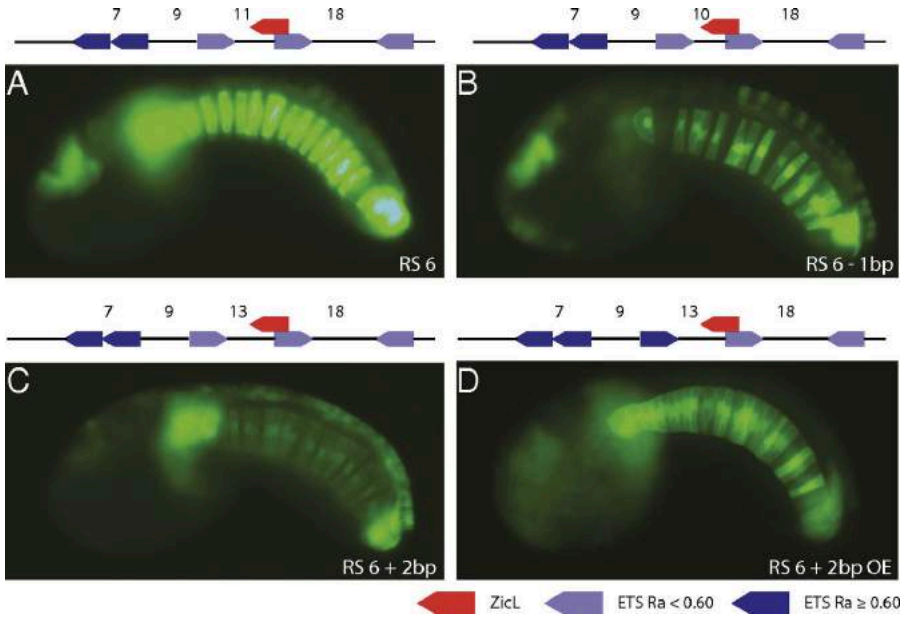
The position of essential TF-binding sites is not conserved in functionally homologous enhancers

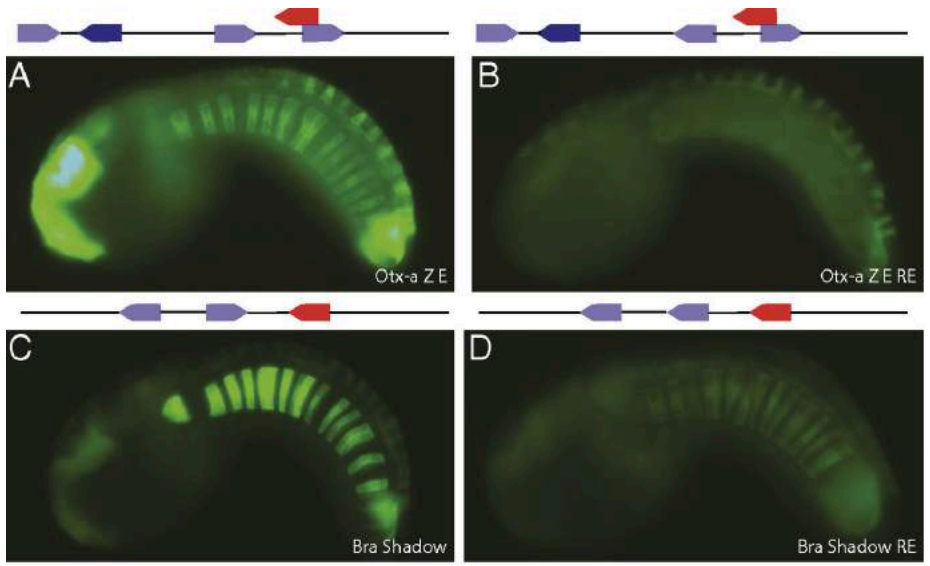




(Scemema et al. (2002) *J Exp Zool B*)

Importance of distance and relative directionality of TF binding sites for gene expression



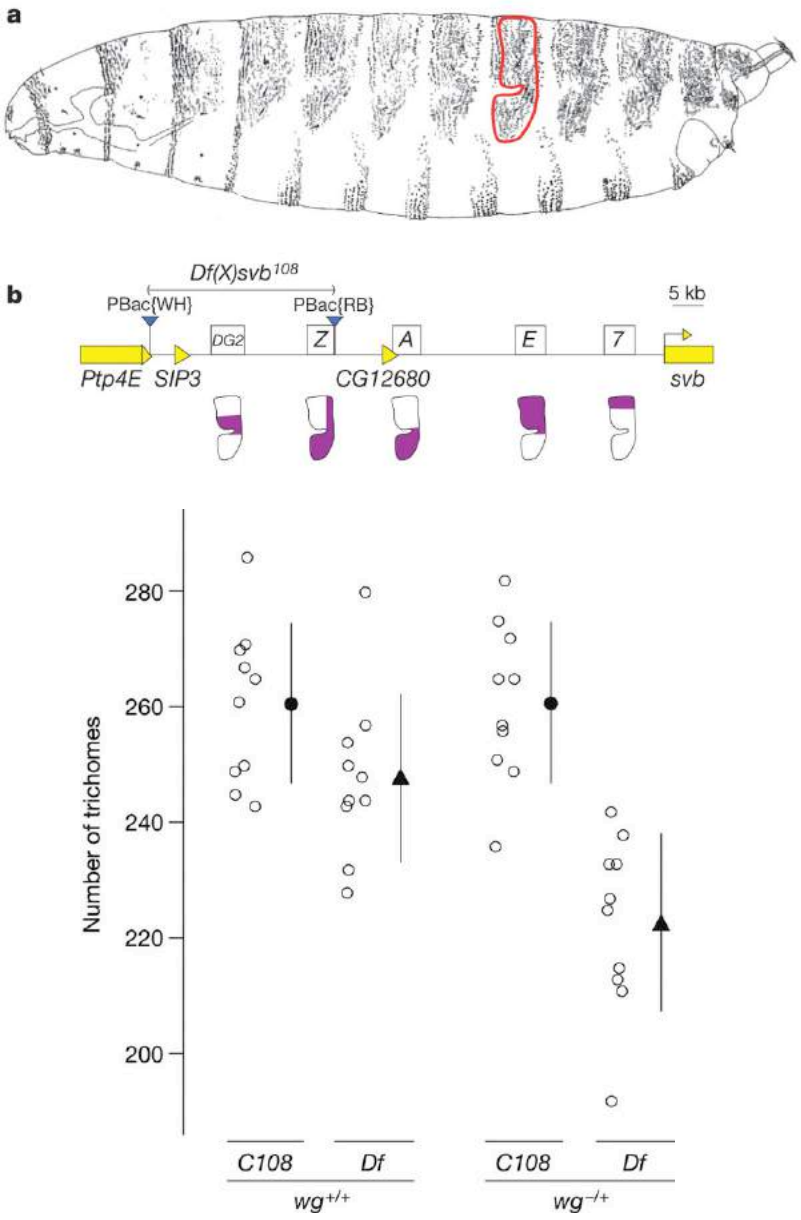


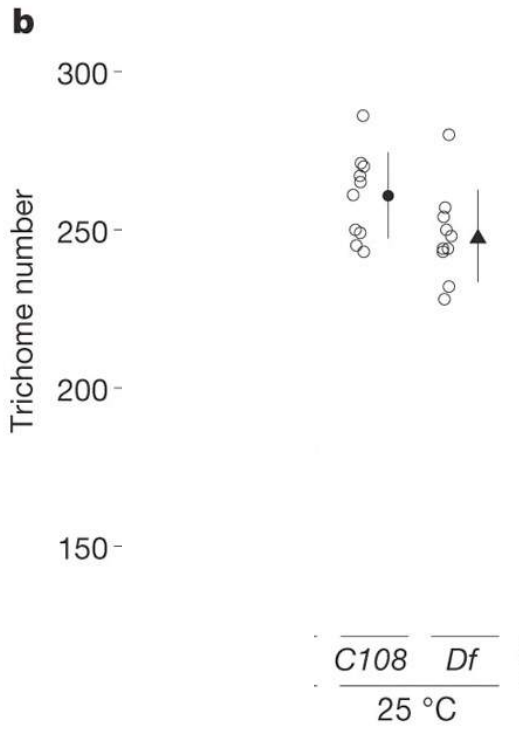
- Distance between binding sites can affect gene expression

- Similarly, the directionality of the binding sites can be important as well

(Farley et al. 2016 *PNAS*)

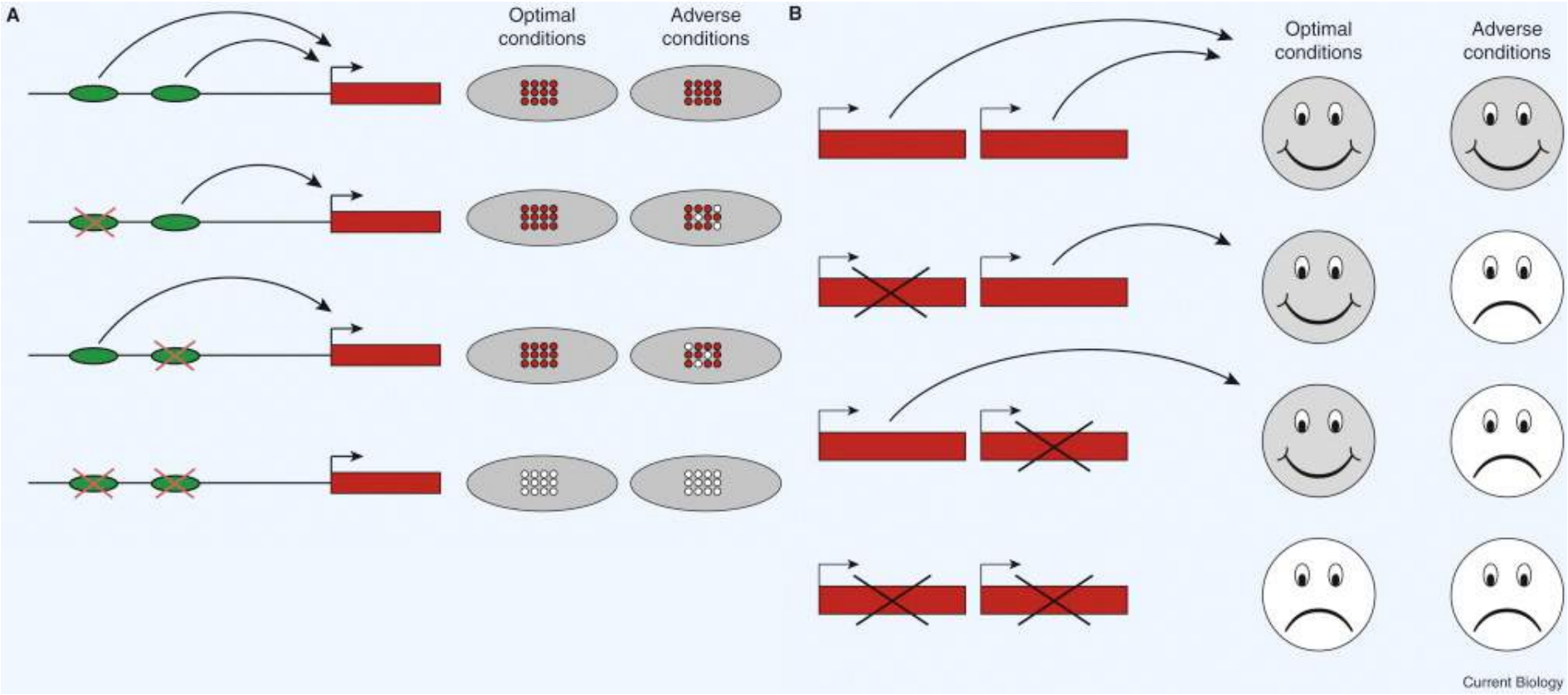
“Shadow” enhancers ensure the robustness of developmental processes





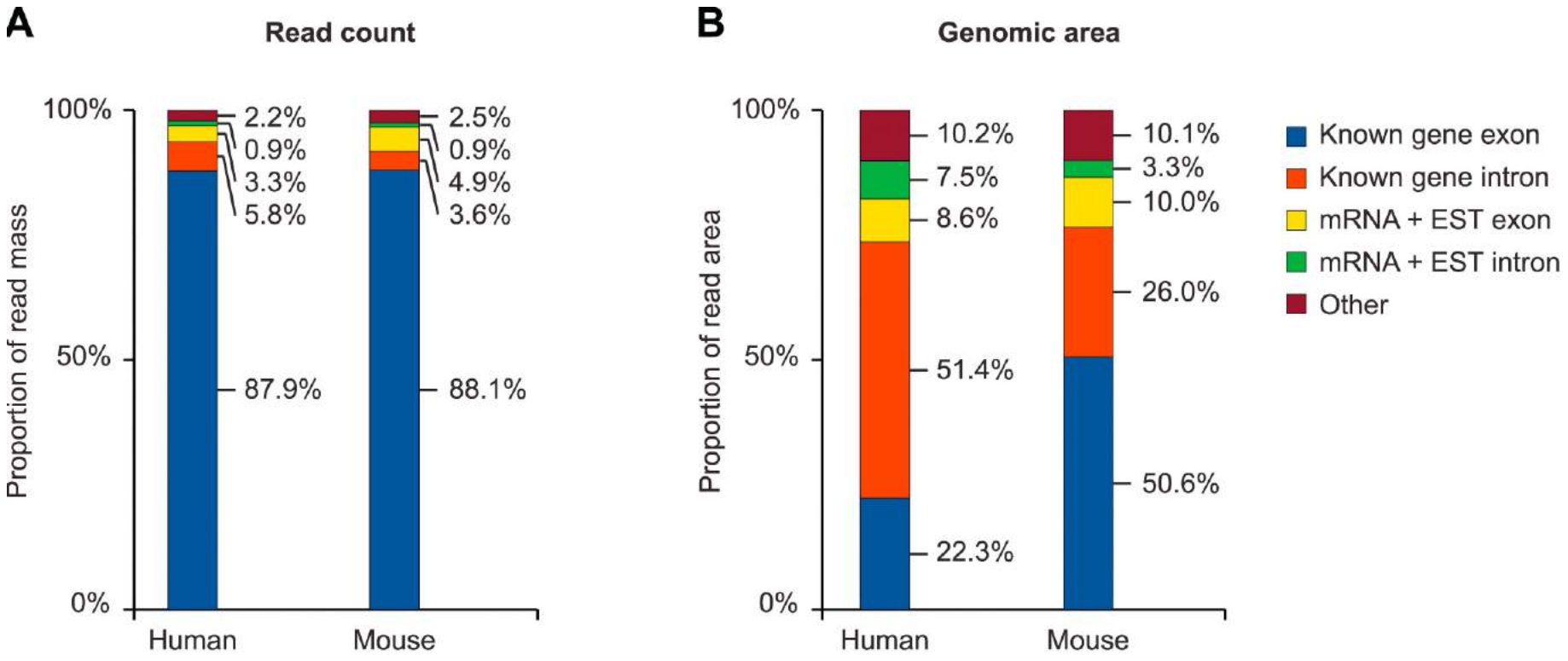
(Frankel et al. (2010) *Nature*)

The logic of “shadow” enhancers is similar to that of paralog genes



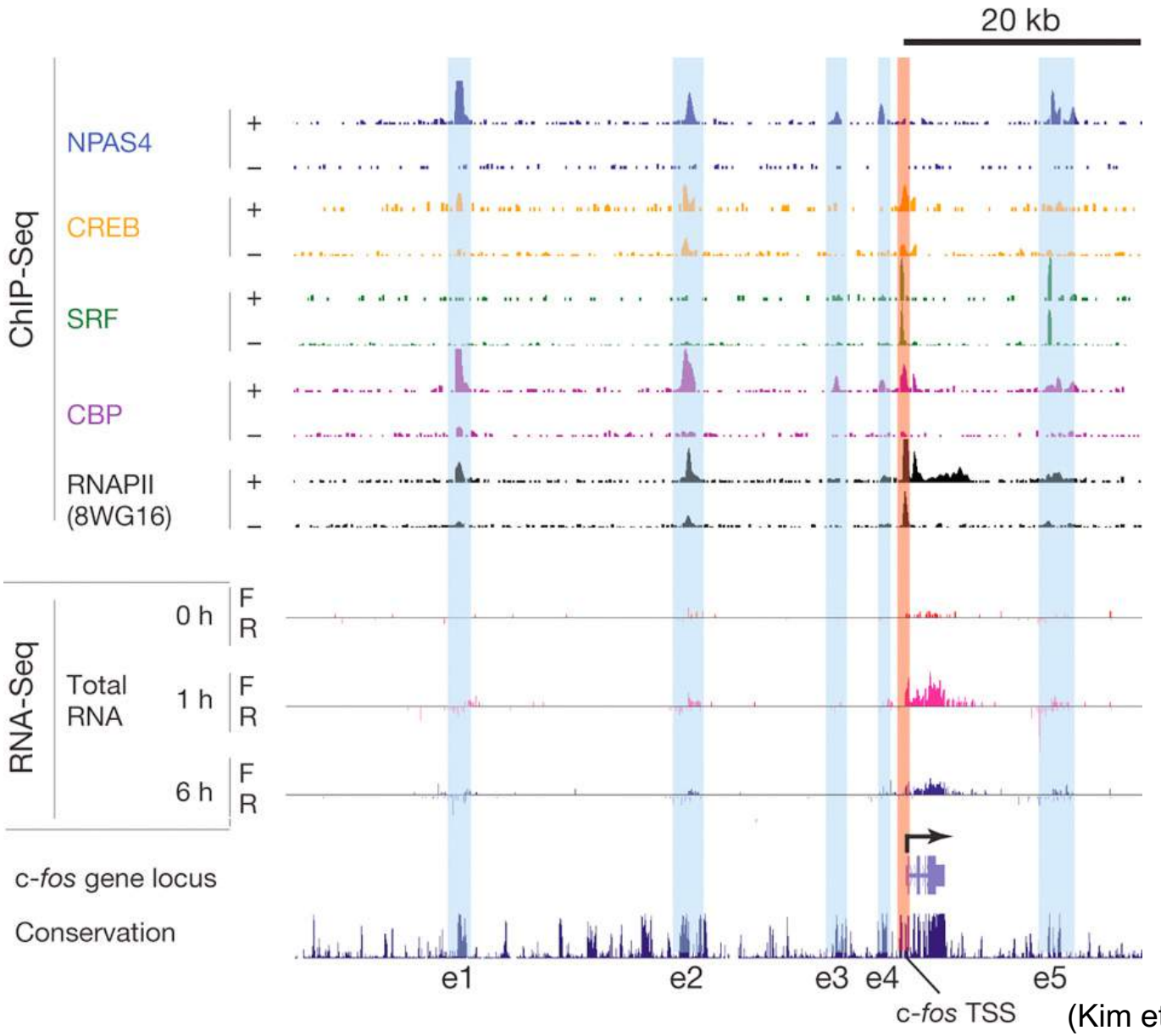
(Holbert (2010) *Curr Bio*)

Not only genes can be transcribed



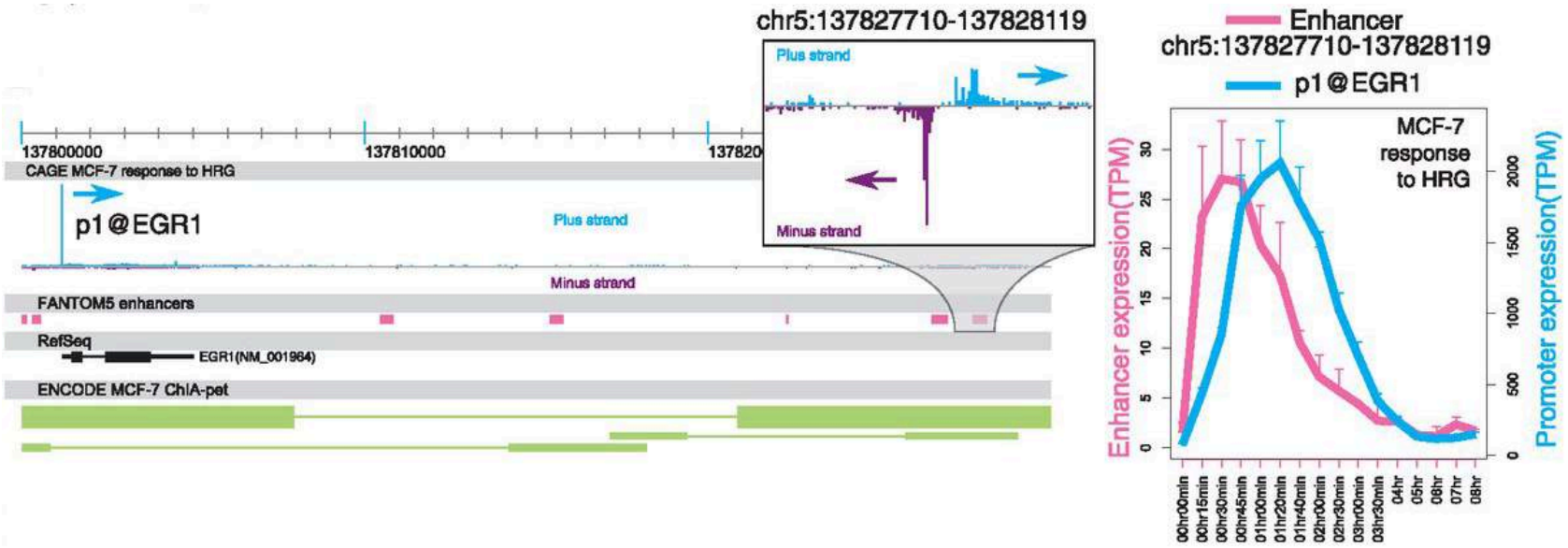
(van Bakel et al. (2010) *PLoS Bio*)

eRNA: transcription around enhancers



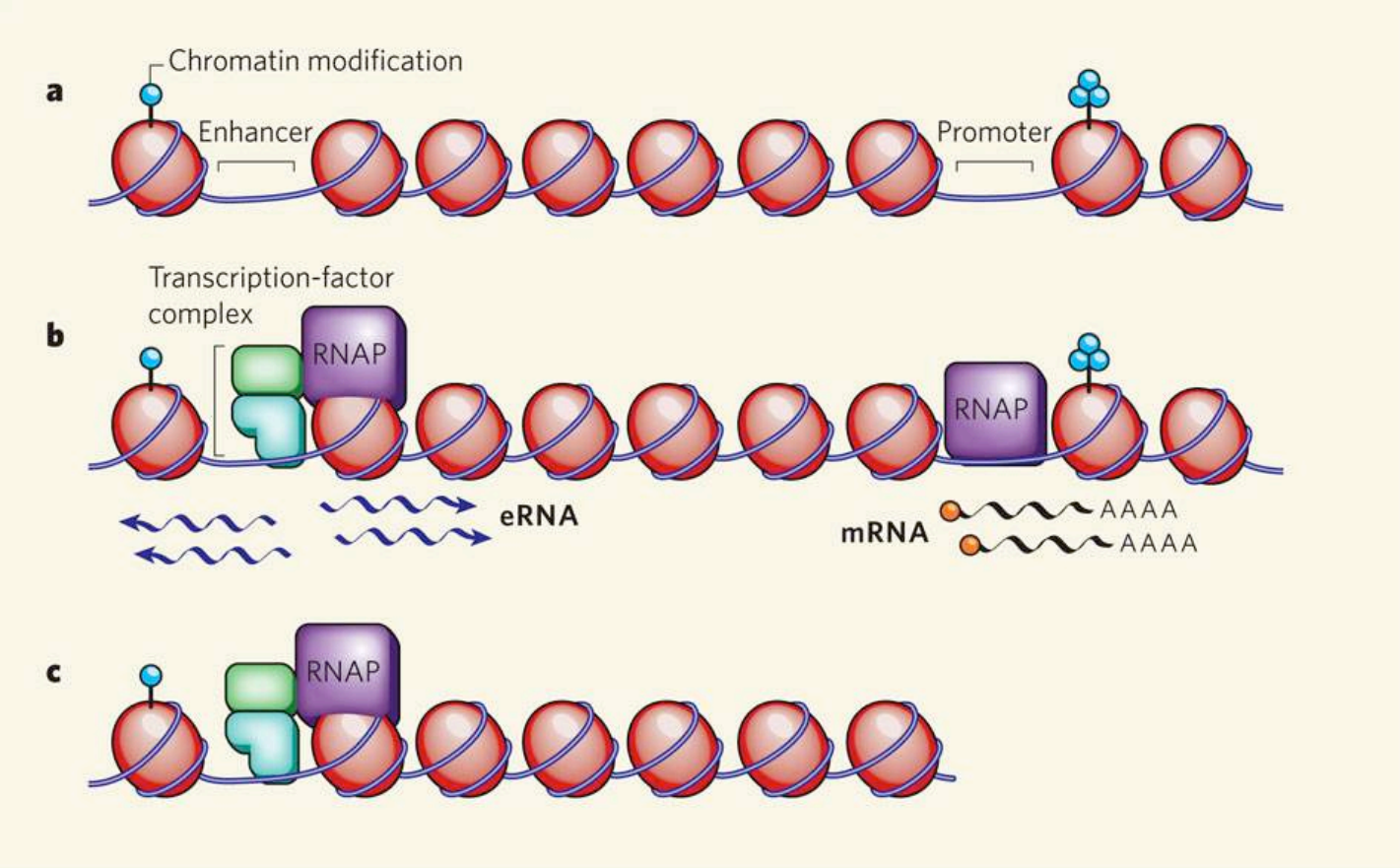
(Kim et al. (2010) *Nature*)

Transcription at enhancers precedes transcription at promoters



(Arner et al. (2015) *Science*)

eRNA: transcription around enhancers

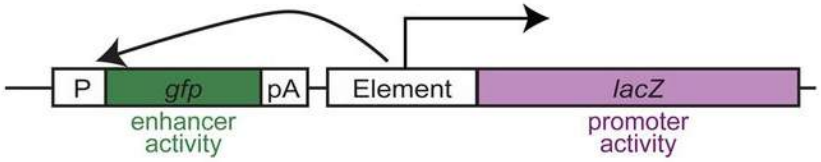


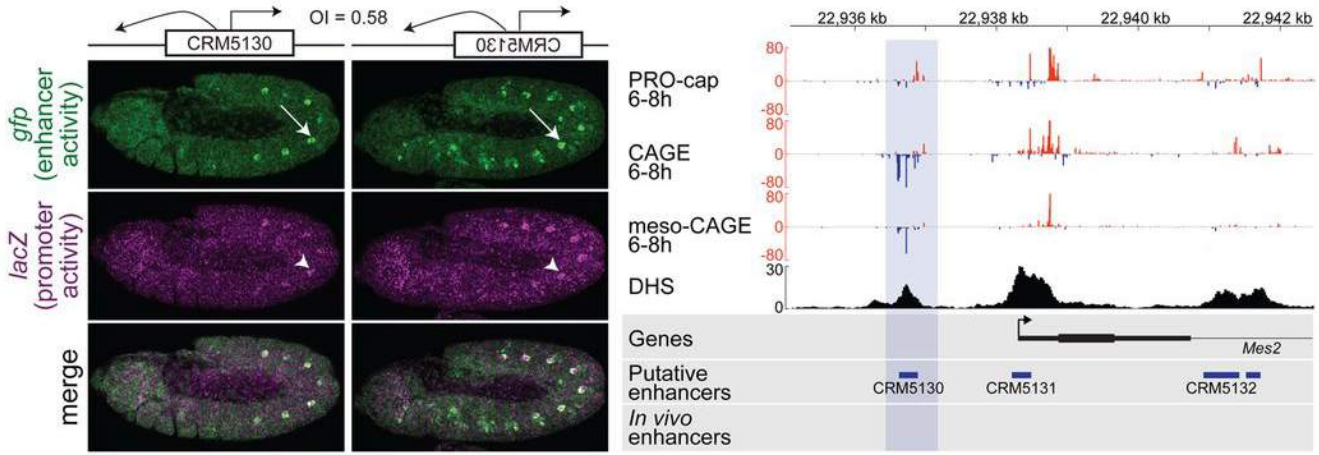
We do not understand:
- if eRNAs have a function themselves?
- or only the loosening of the chromatin at enhancers is important?

(Ren (2010) *Nature*)

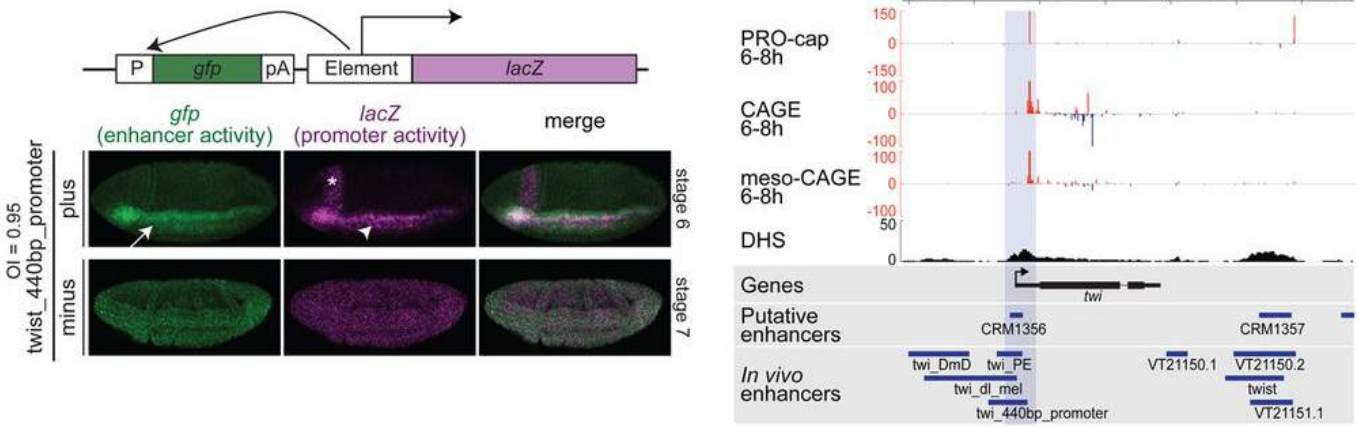
The boundary between enhancers and promoters is blurred

- Enhancers with high eRNA transcription can act as promoters



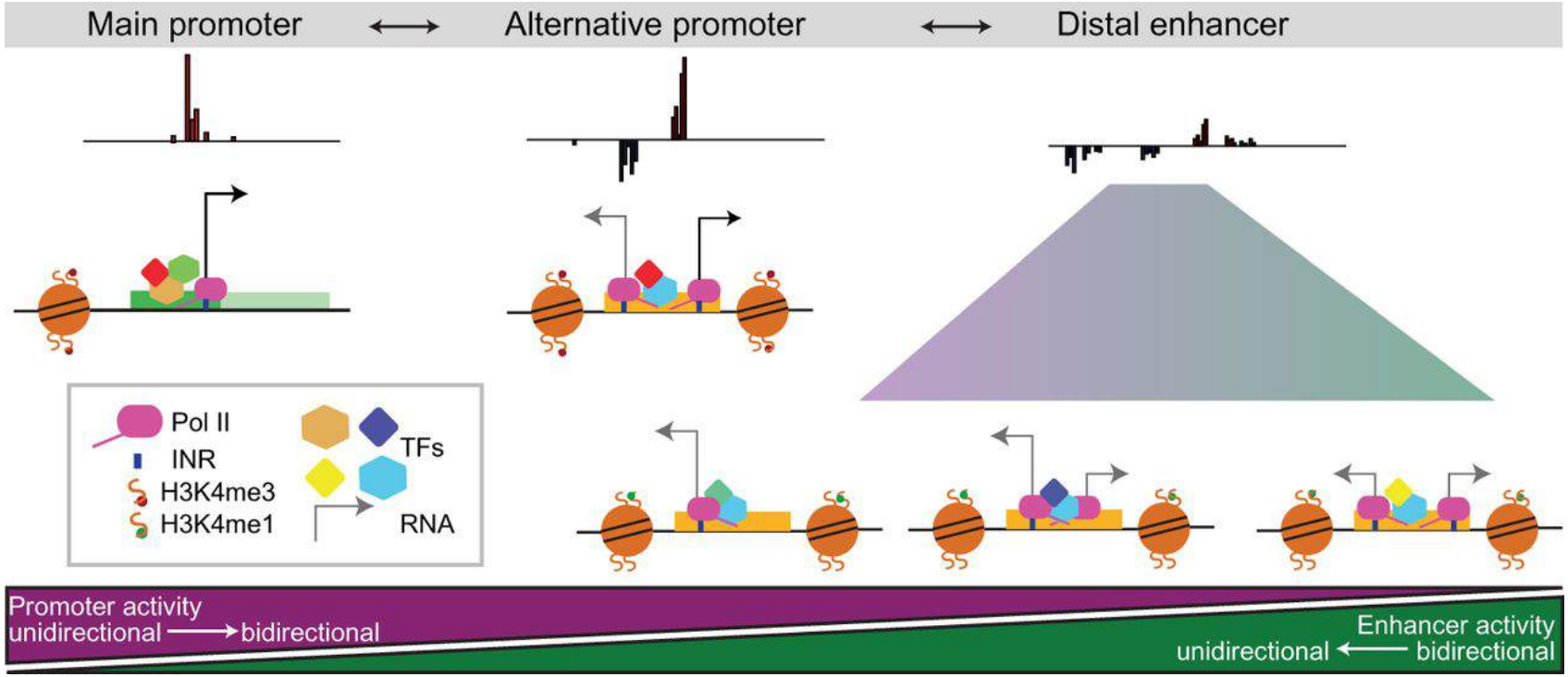


- Alternative promoters can also act as enhancers



(Mikhaylichenko et al. 2018 *GenesDev*)

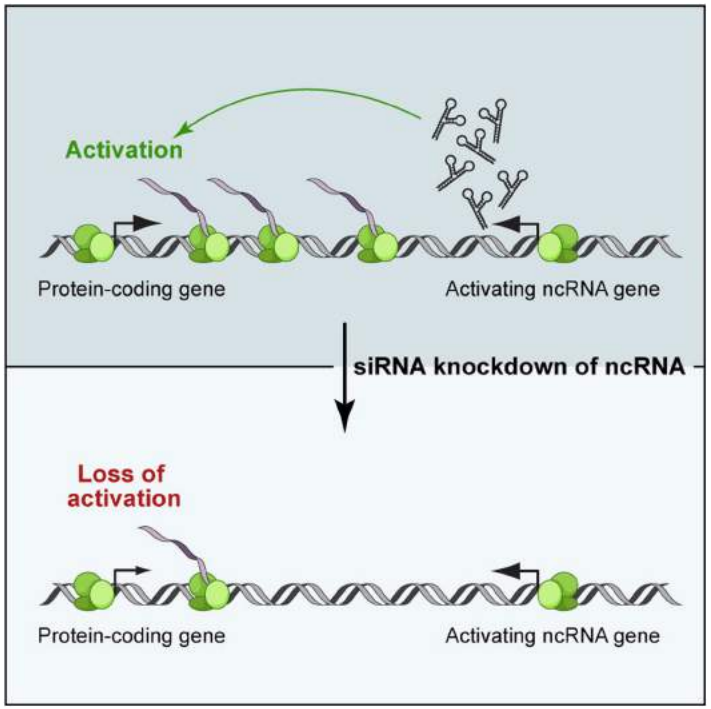
The boundary between enhancers and promoters is blurred



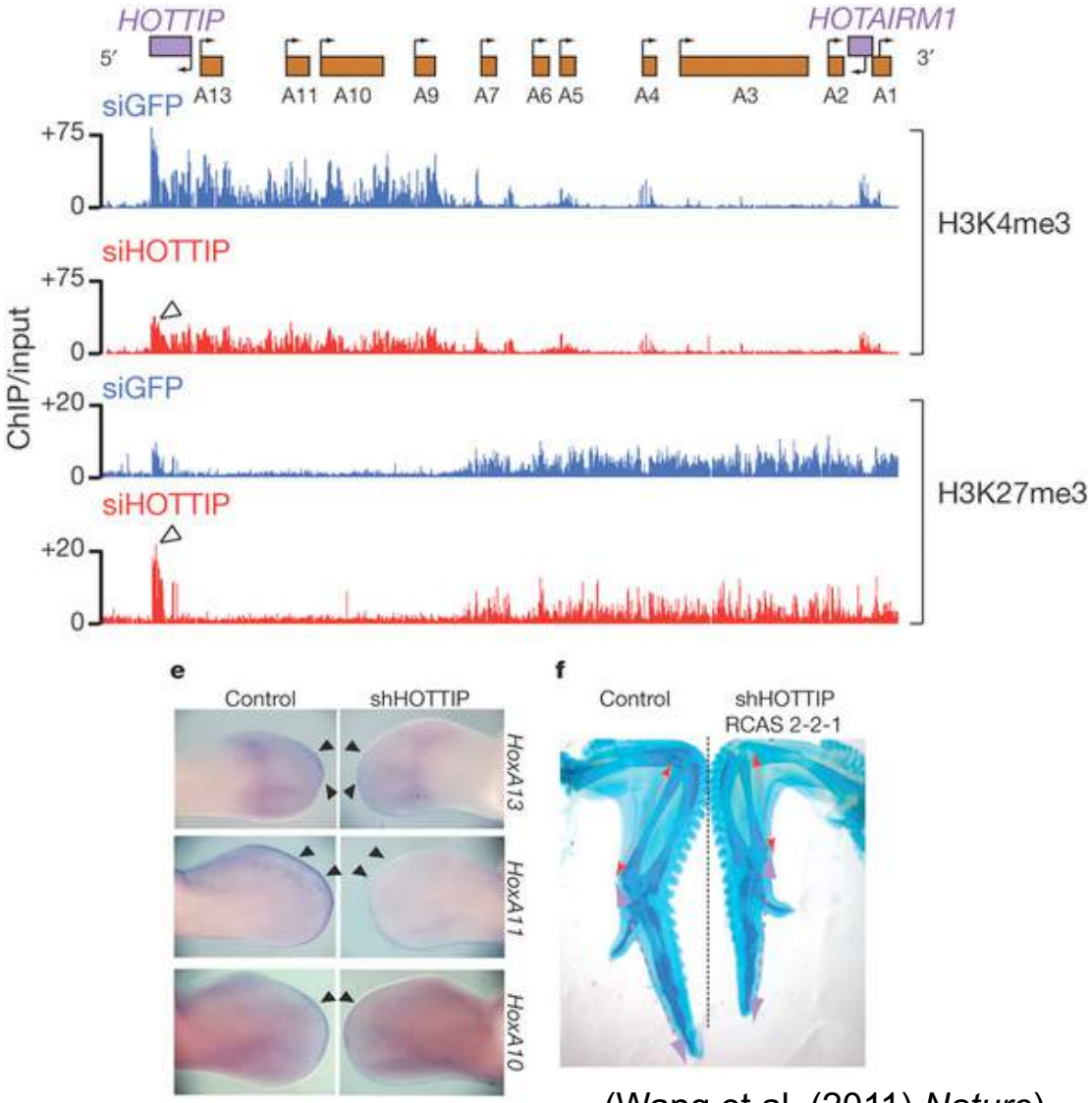
- It is still not quite clear how these things evolve

(Mikhaylichenko et al. 2018 *GenesDev*)

lncRNAs could have a role in helping the transcription in nearby genes

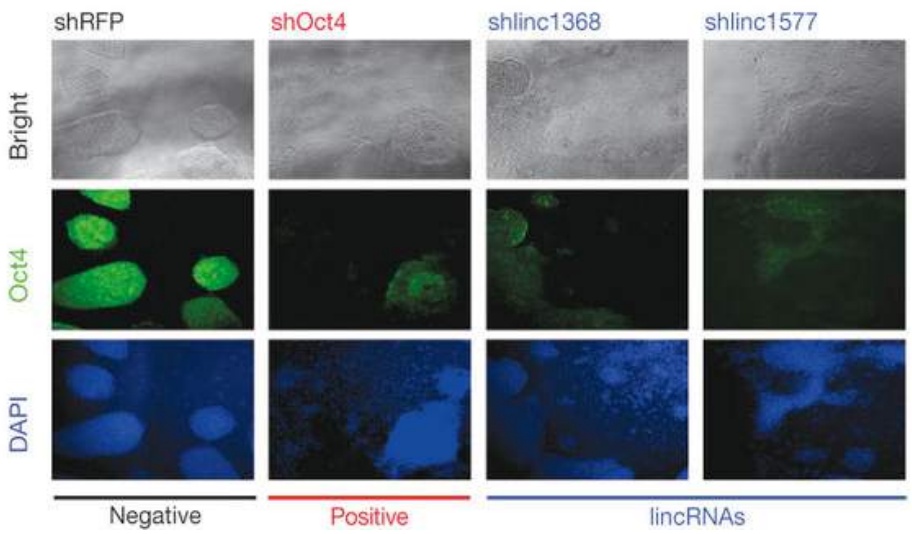


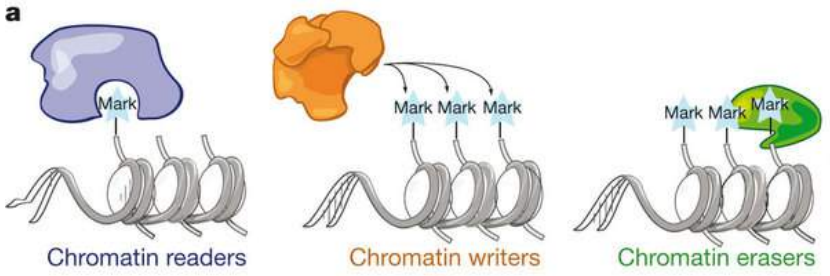
(Ørom et al. (2010) *Cell*)

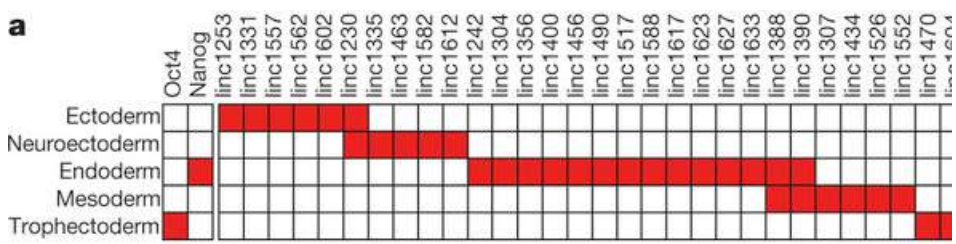


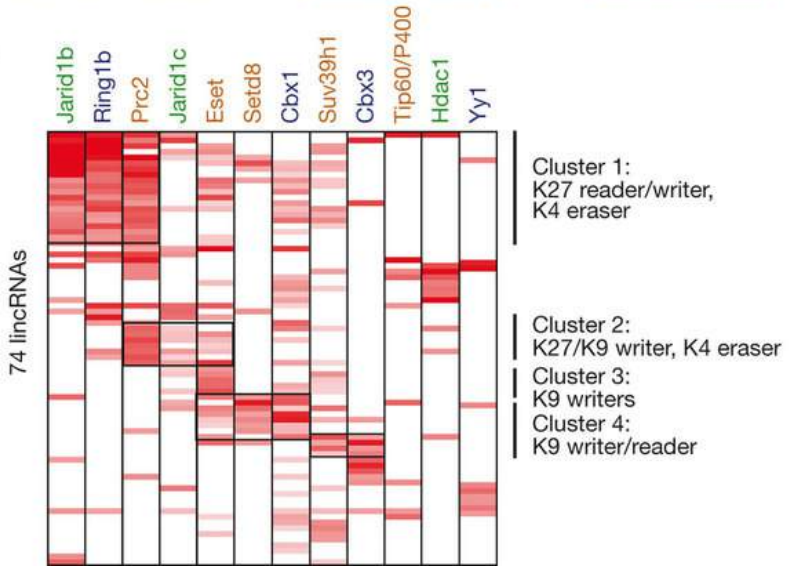
(Wang et al. (2011) *Nature*)

Some lncRNAs can act as specific adapters in ES cells



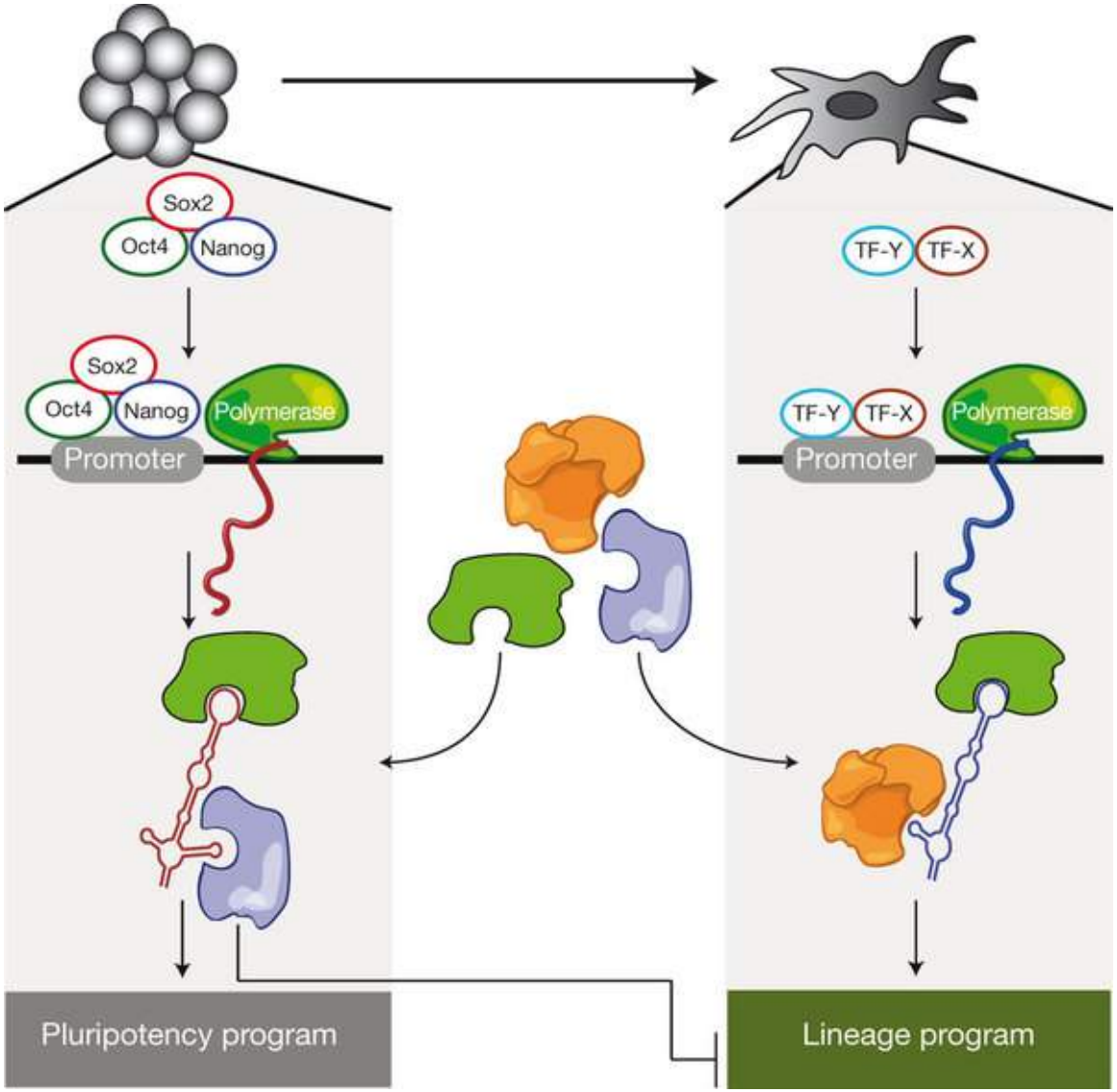






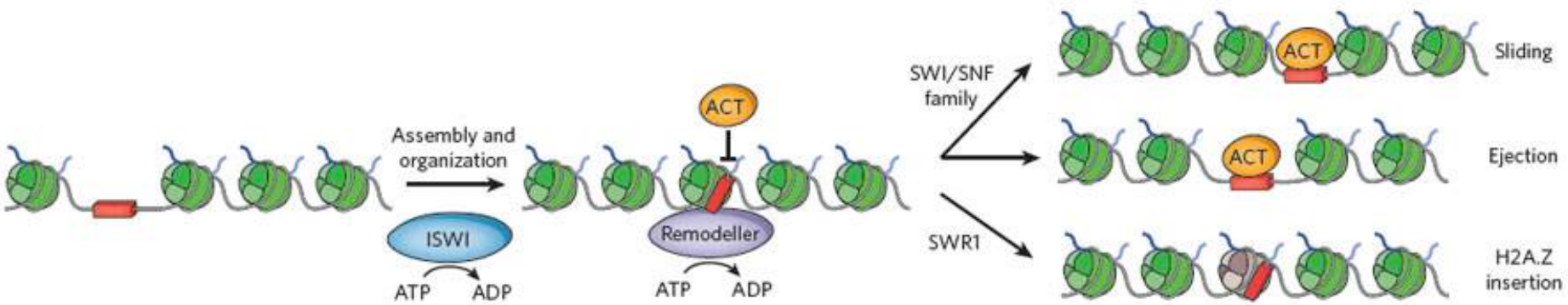
(Guttman et al. (2011) *Nature*)

Some lncRNAs can act as specific adapters in ES cells



(Guttman et al. (2011) *Nature*)

Chromatin remodeling is necessary for transcriptional activation

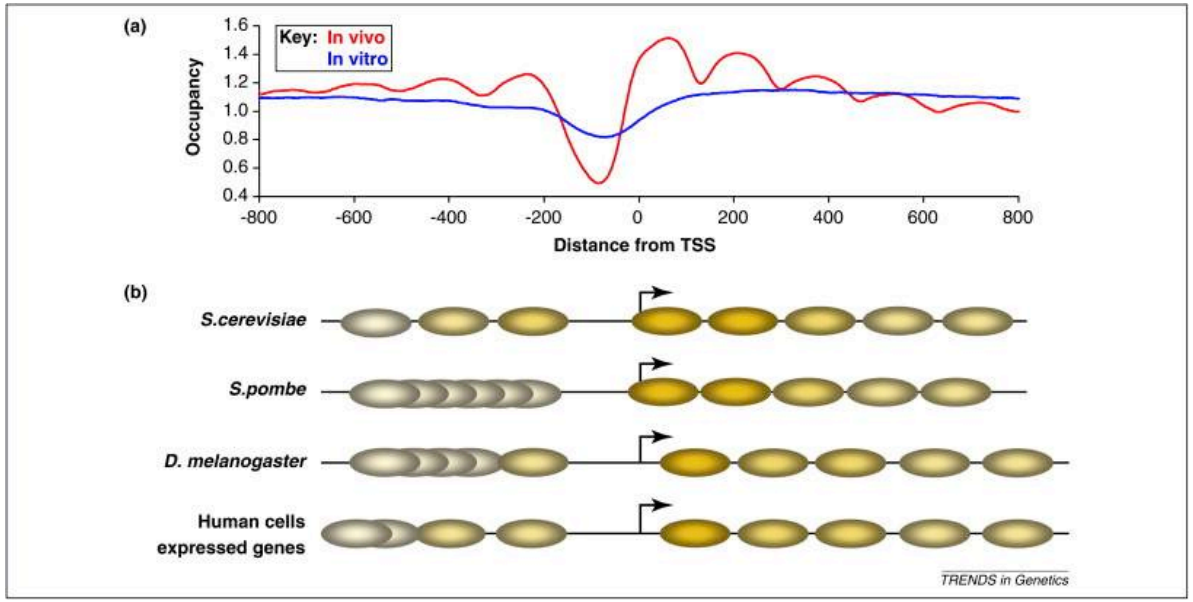


ISWI - help to conduct chromatin assembly and organization and provide consistent spacing of nucleosomes

SWI/SNF - provide access to binding sites in nucleosomal DNA, mainly through nucleosome movement or ejection

SWR1 - reconstruct nucleosomes by inserting the histone variant H2A.Z into nucleosomes, specializing their composition and leading to an unstable nucleosome

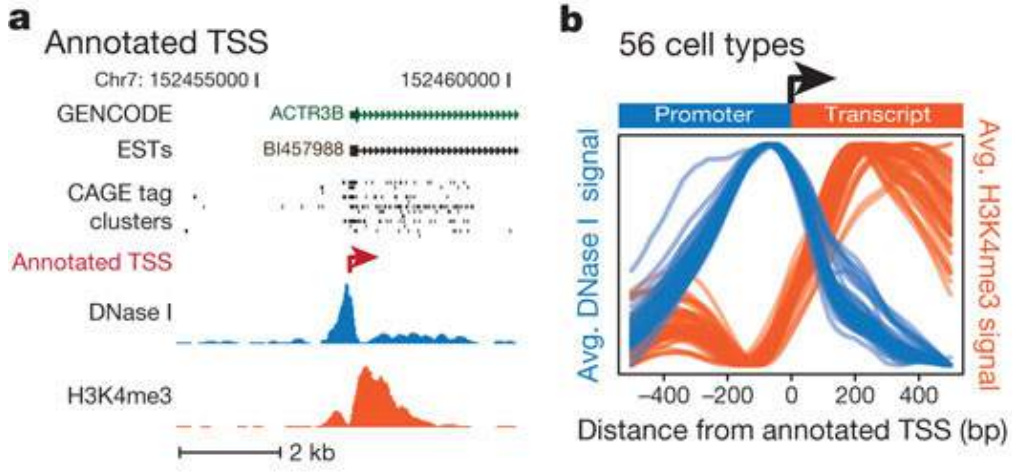
In the proximity of the transcriptional start site (TSS) nucleosome position is stereotypic



-Before the TSS a nucleosome-free region can be observed

- distal to the TSS nucelosome position becomes less and less stereotypical

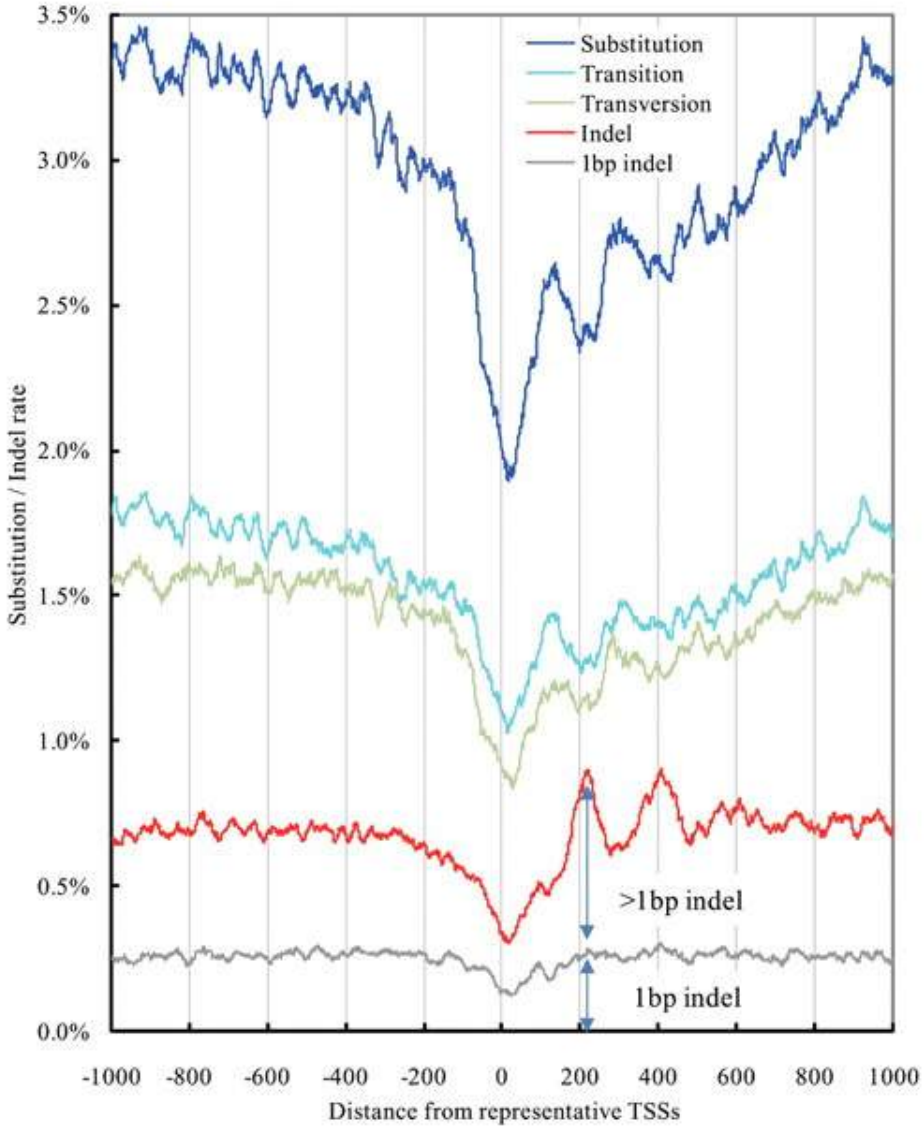
(Bai and Morozov (2010) *TiG*)



- nucleosome-free regions show DNAase hypersensitivity

(Thurman et al. (2012) *Nature*)

Sequence conservation in the proximity of the transcriptional start site (TSS)



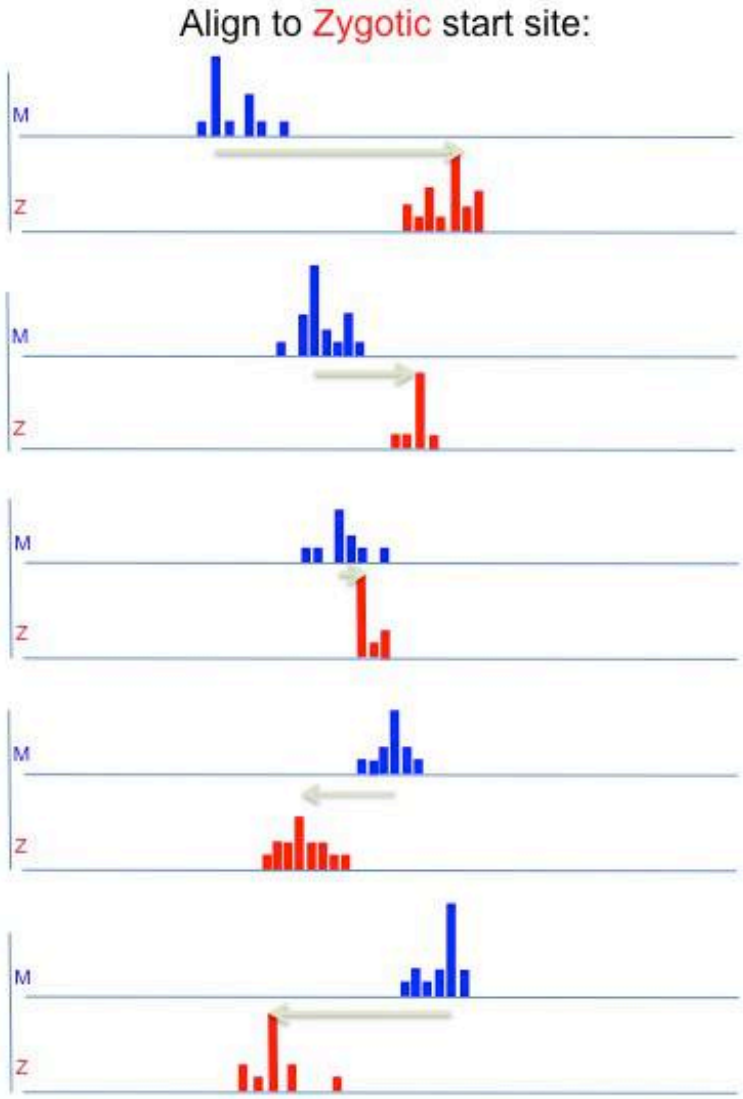
- there are fewer mutations in the immediate proximity of the TSS

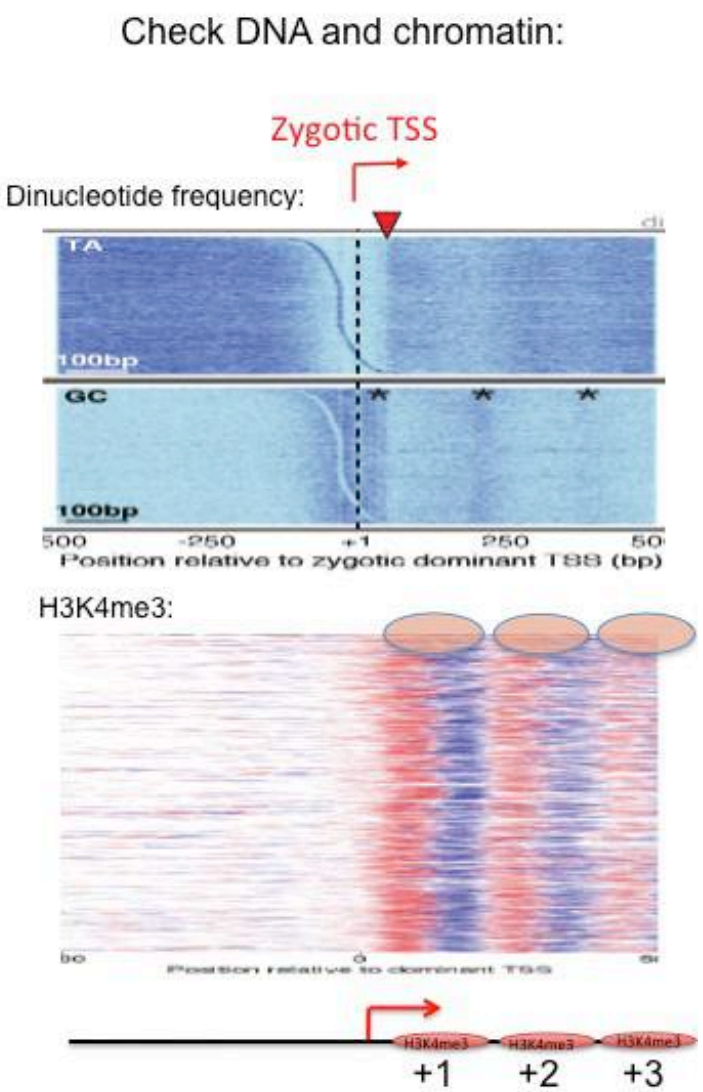
- interestingly, a bit further away, just as the periodicity of histone placement would imply, we see this conservation reoccurring (although it gets weaker)

- one hypothesis is that the position of the nucleosomes is coded into the DNA

(Sasaki et al. (2009) *Science*)

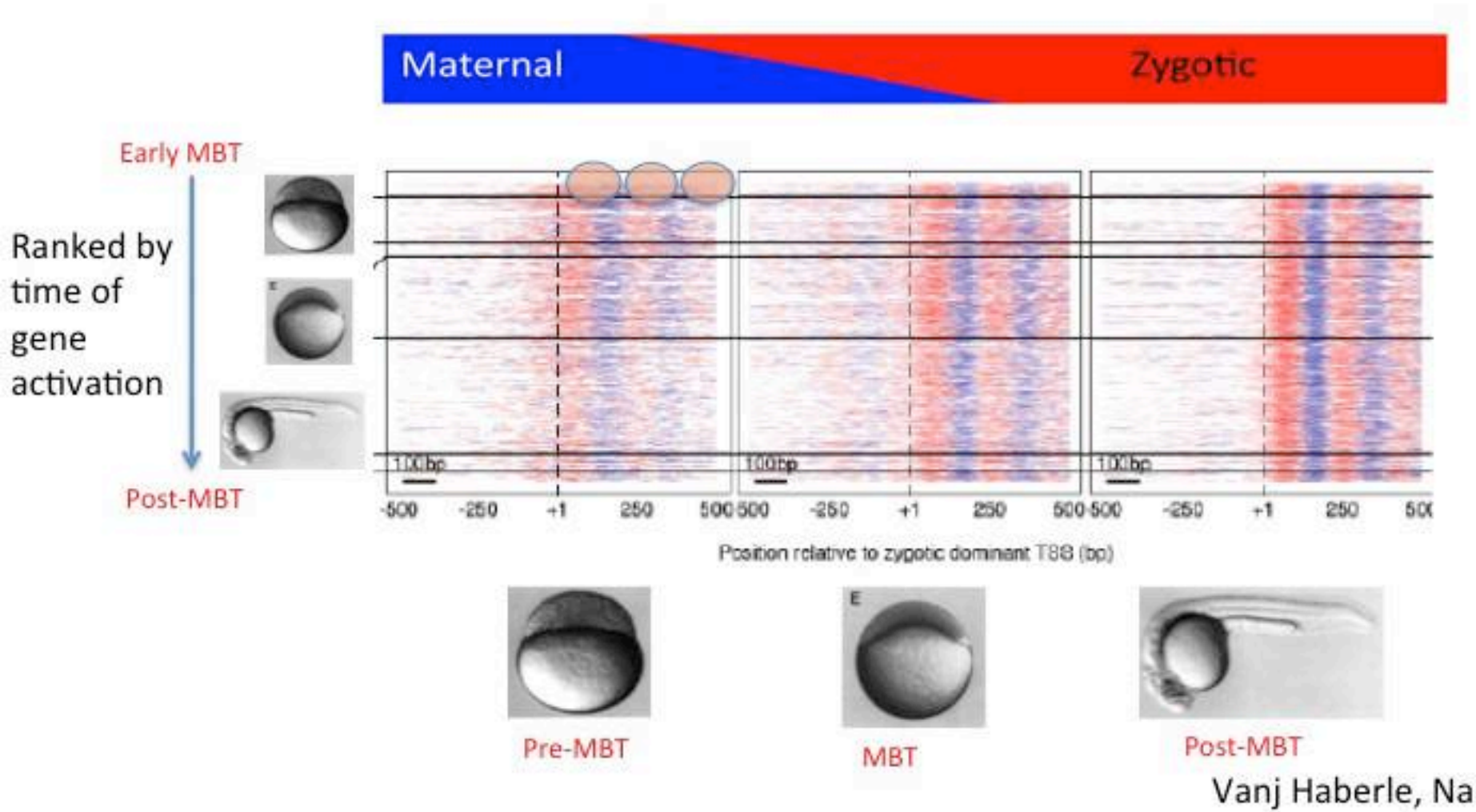
Az “anyai” és “zigotikus” promóterek máshol találhatók!





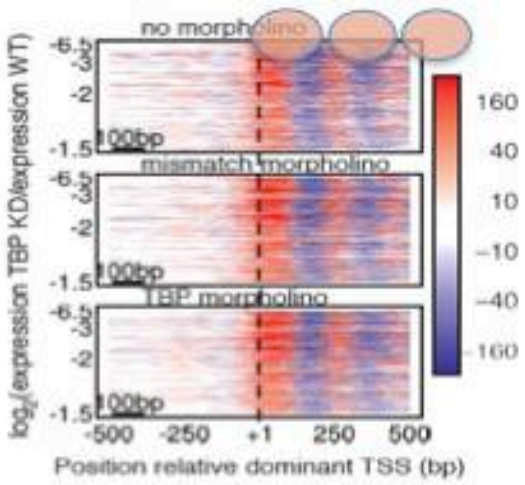
(Müller Ferenc)

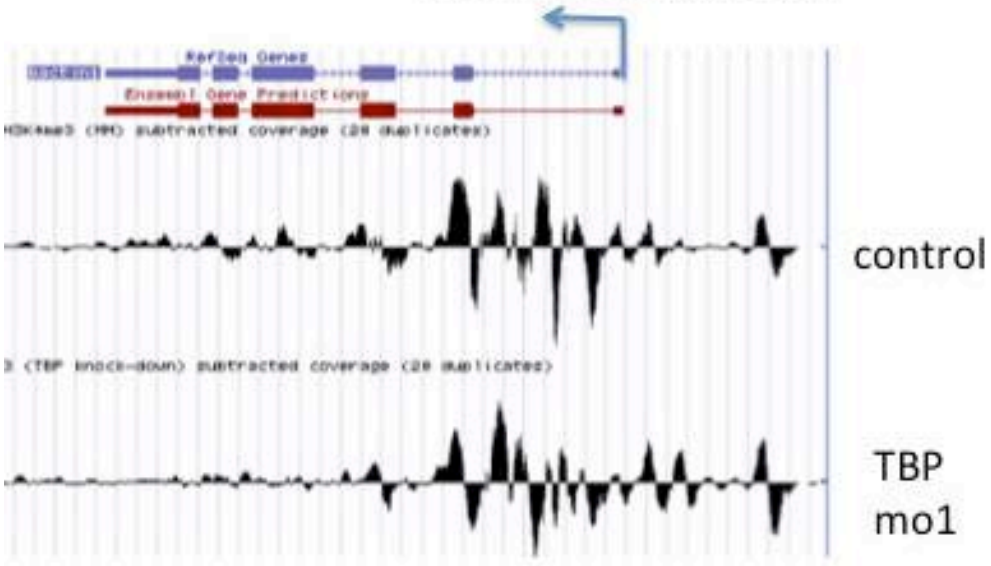
A hiszton-mintázat már a transzkripció megindulása előtt észlelhető



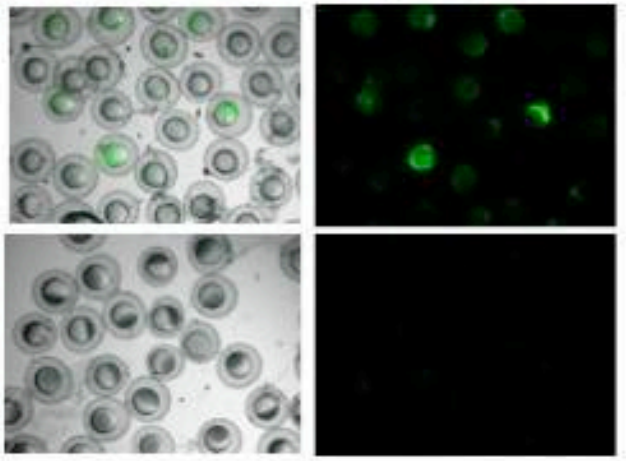
(Müller Ferenc)

… és független a transzkripciós apparátustól!





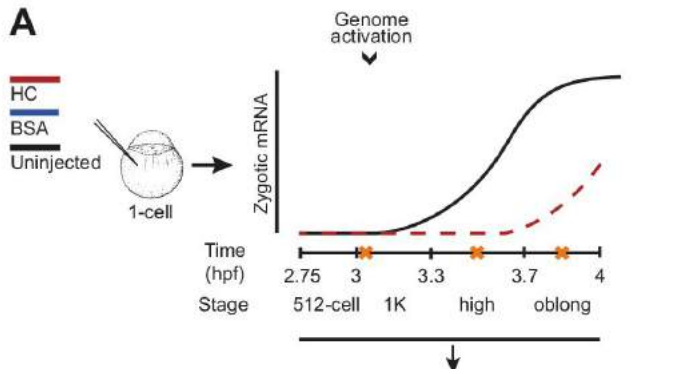
β-actin:YFP

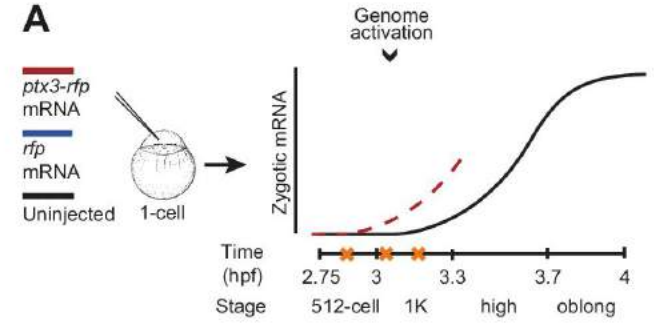


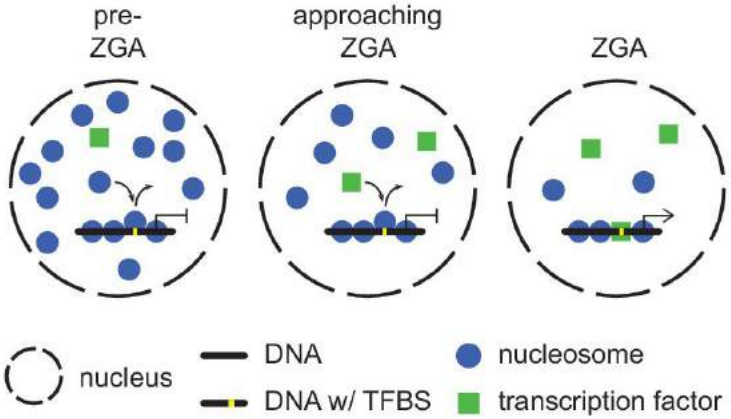
Vanja Haberle Nan li

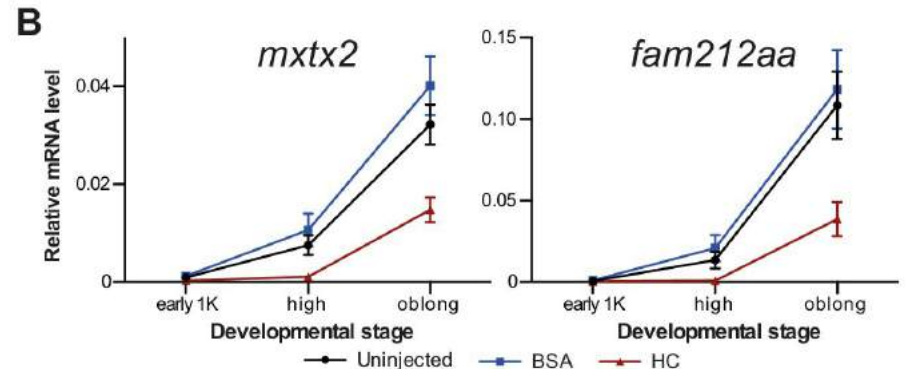
(Müller-Ferreiro)

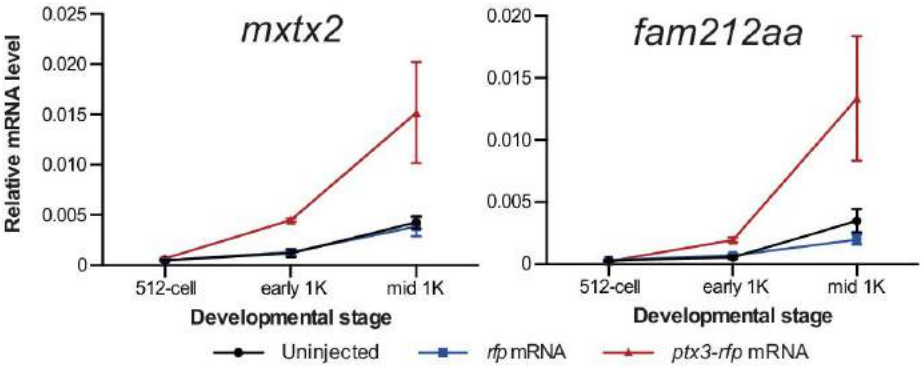
Histones and TF-s compete for DNA







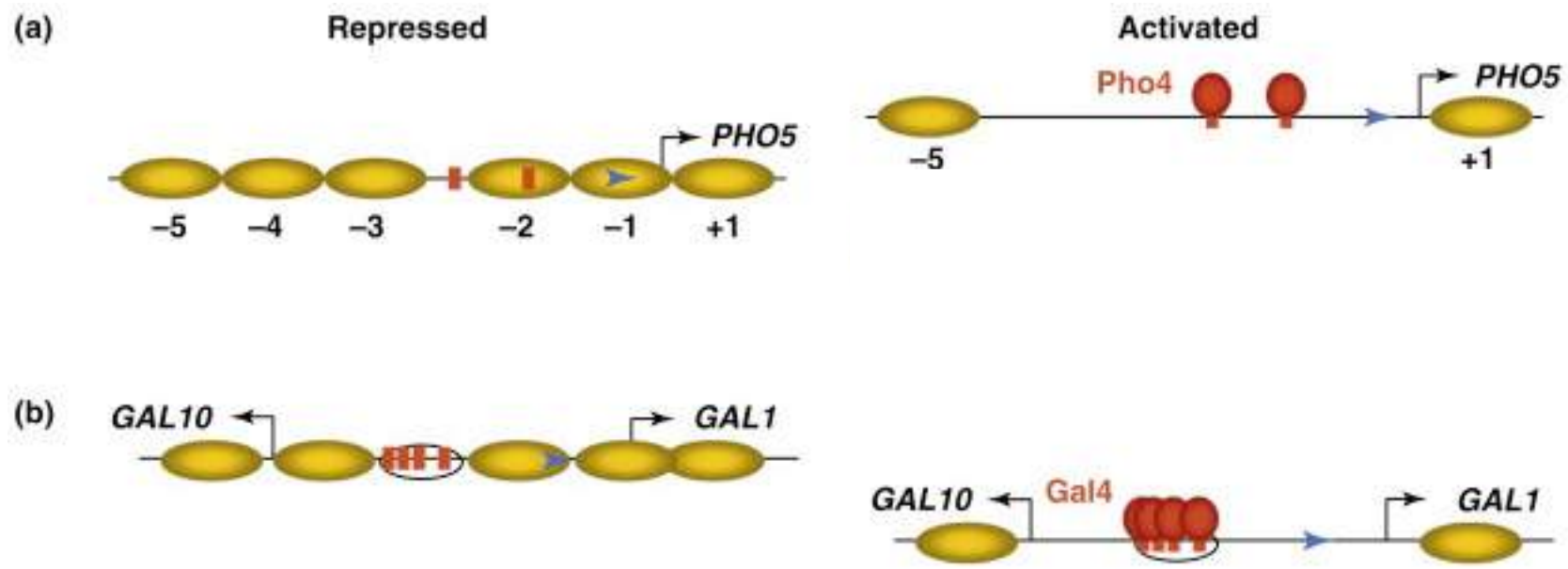




HC – histone mixture
Ptx3 – irreversible histone binder

(Joseph et al. (2017) *eLife*)

Nucleosome rearrangements occur during transcription



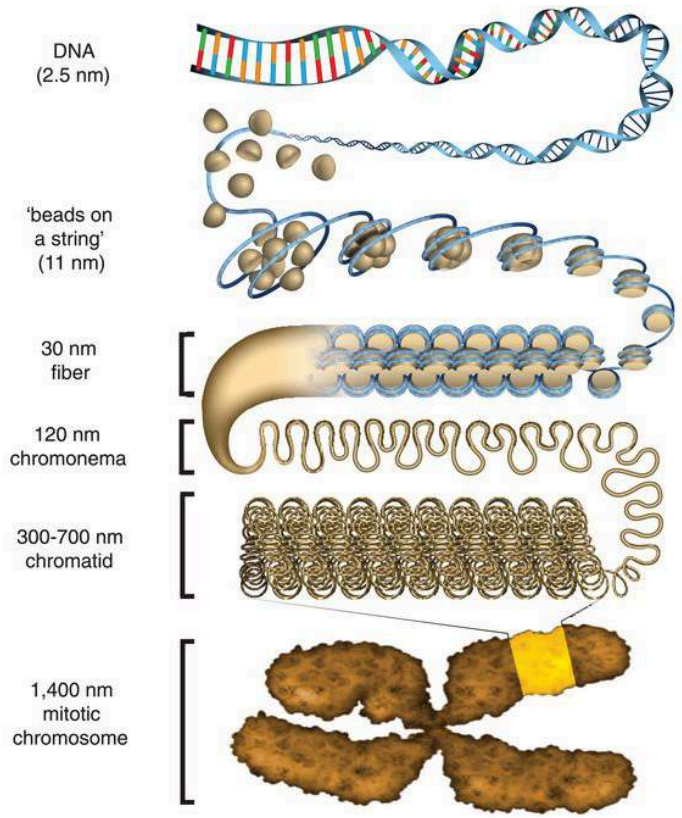
(Bai and Morozov (2010) *TiG*)

- If the right TFs bind to their consensus sites in the nucleosome-free region, the nearby nucleosomes will rearrange themselves and the TSS becomes accessible

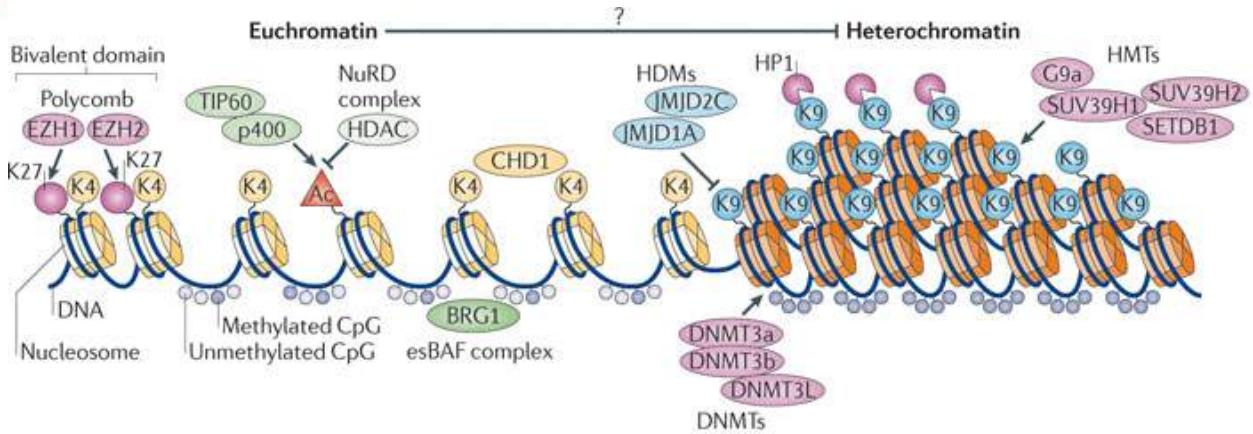
- In the absence of histones nucleosomes can not reform and transcription becomes permanent even after TF-binding is gone for some genes

- For other genes nucleosome rearrangement is a necessary but not sufficient prerequisite for successful transcription

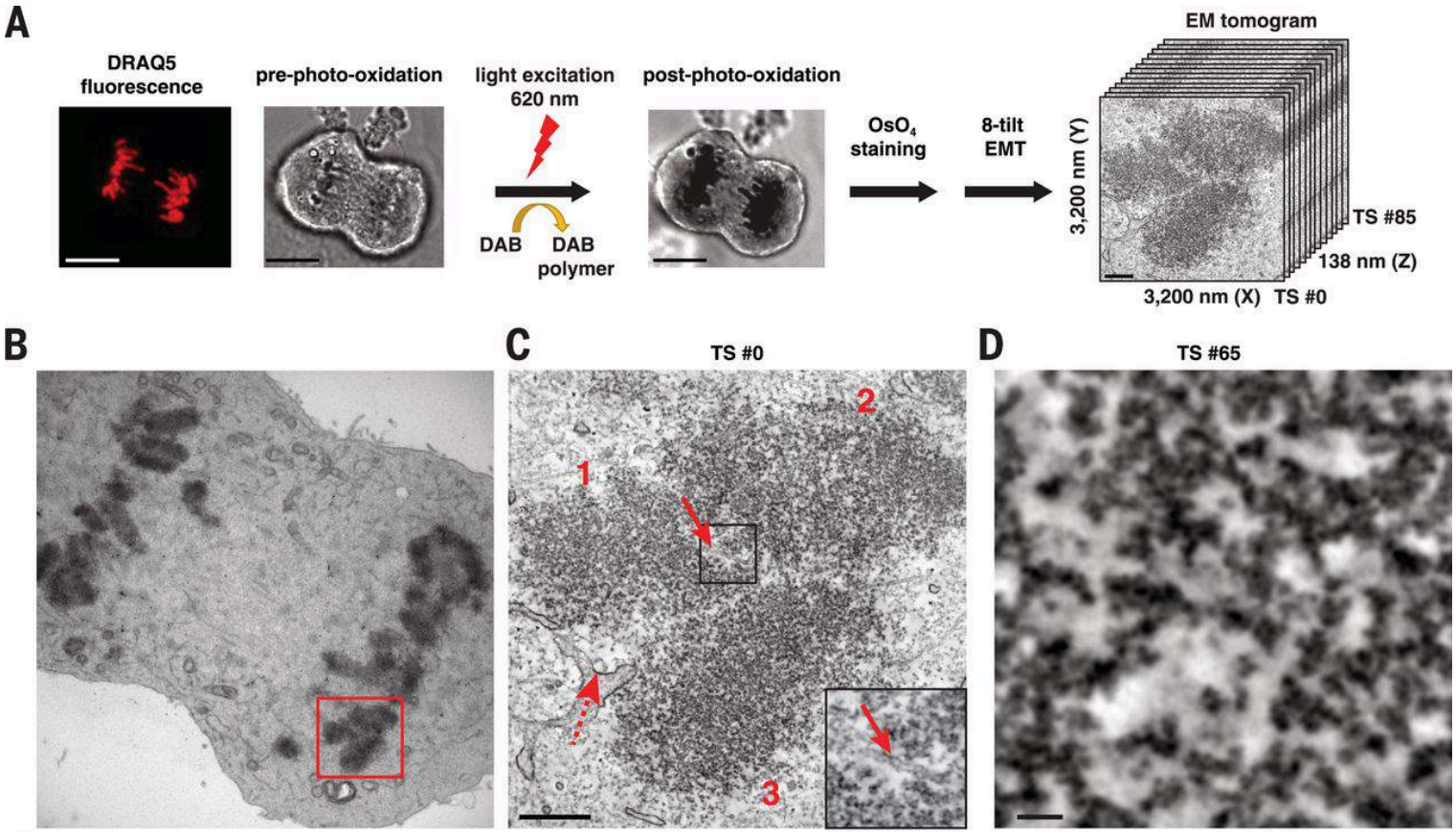
Chromatin organization



Historically we distinguish inactive **heterochromatin** and active **euchromatin**

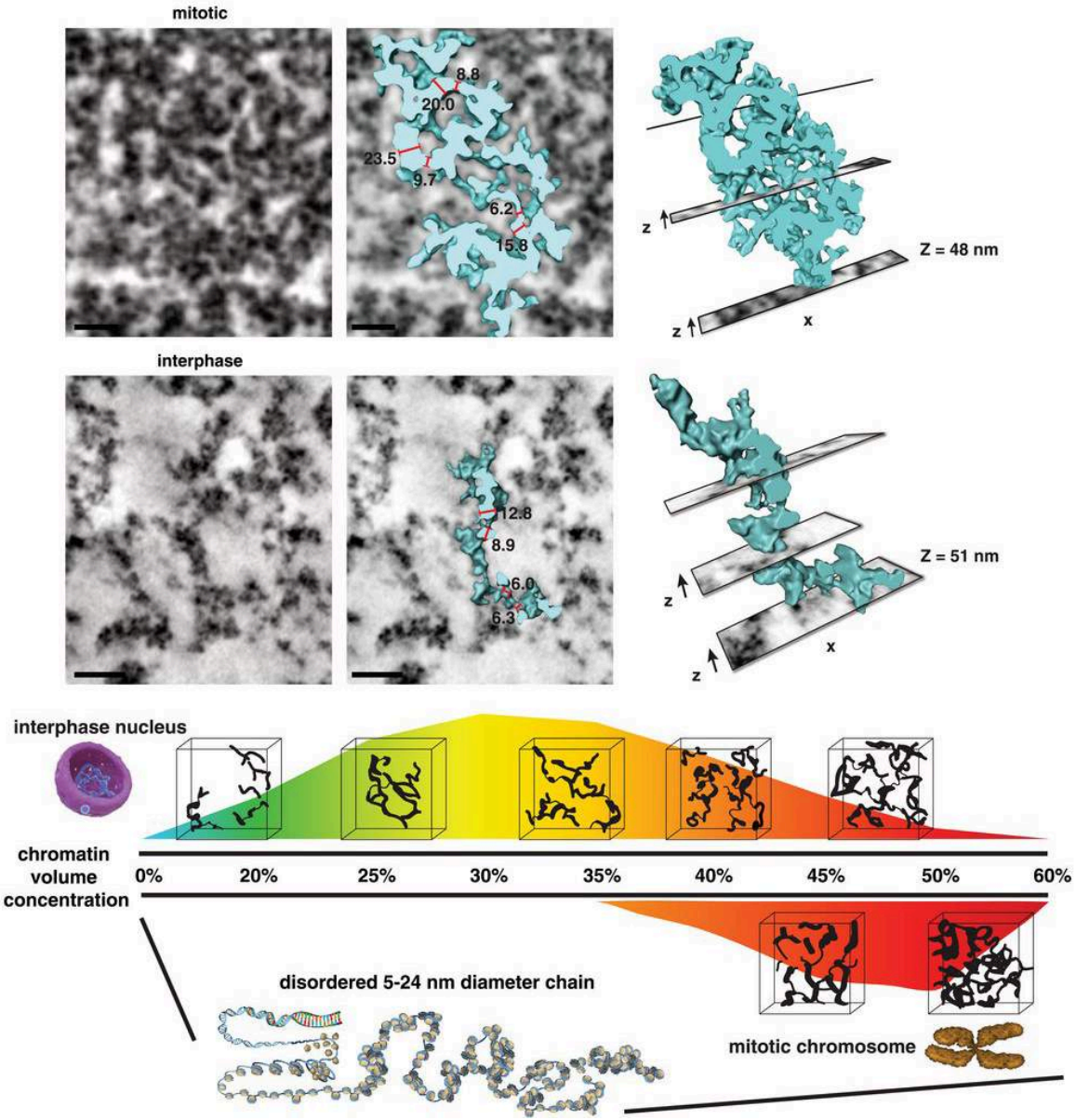


Chromatin organization



(Ou et al. (2017) *Science*)

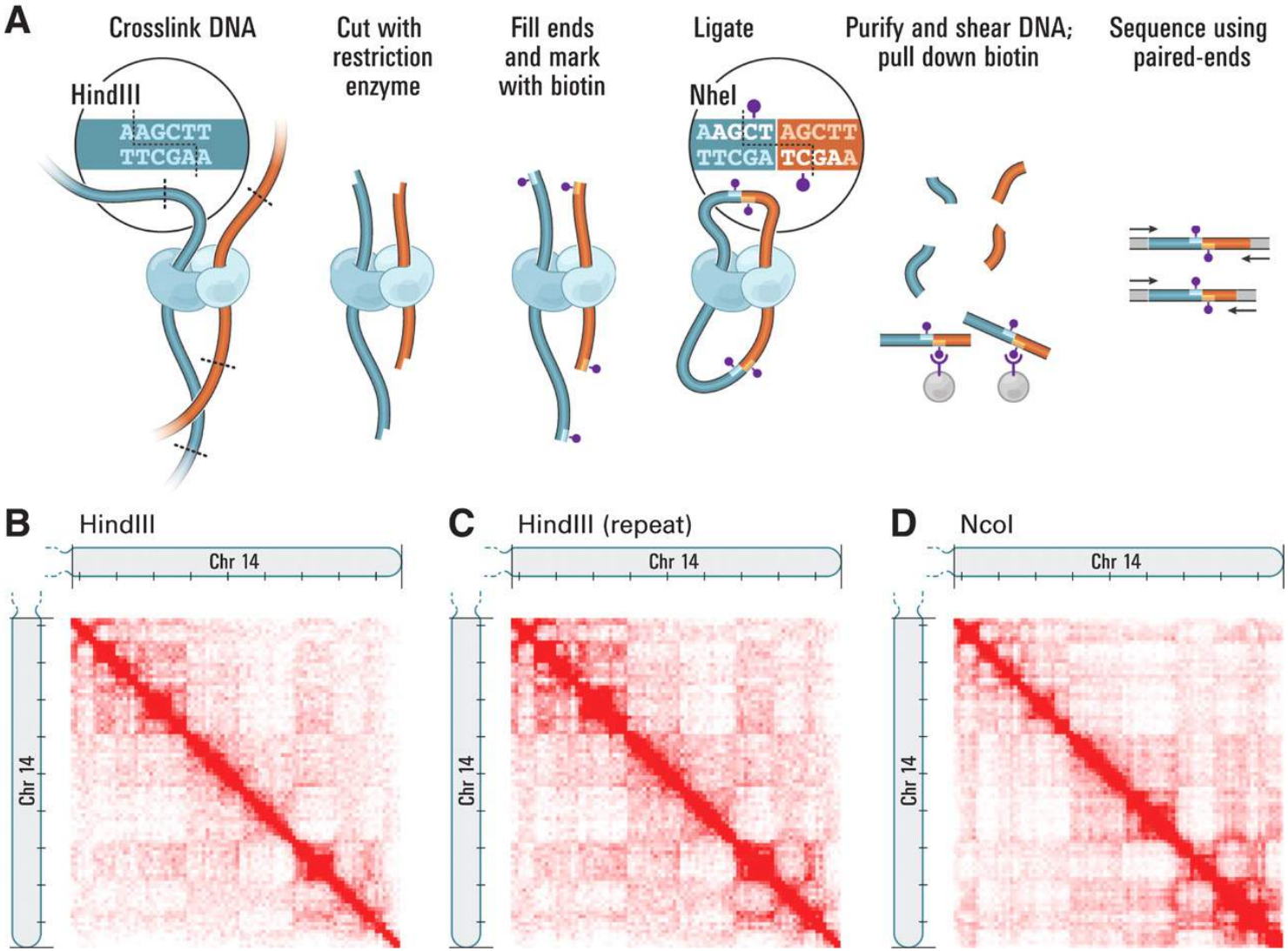
Chromatin organization



- on ChromEMT pictures one can't see the higher level chromatin structures posited by earlier *in vitro* experiments – DNA on histones is just more packed

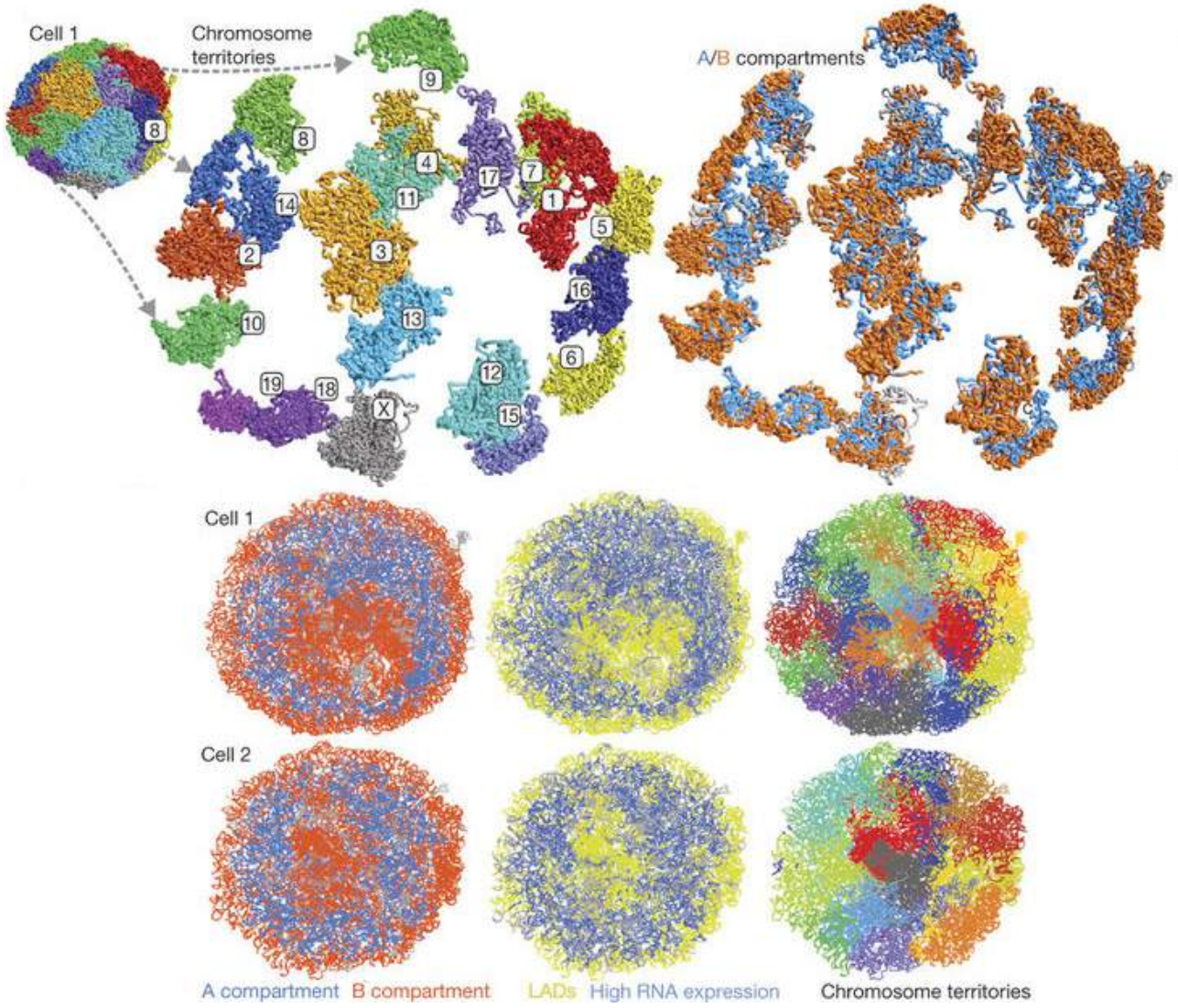
(Ou et al. (2017) *Science*)

Hi-C method to test chromosomal organization



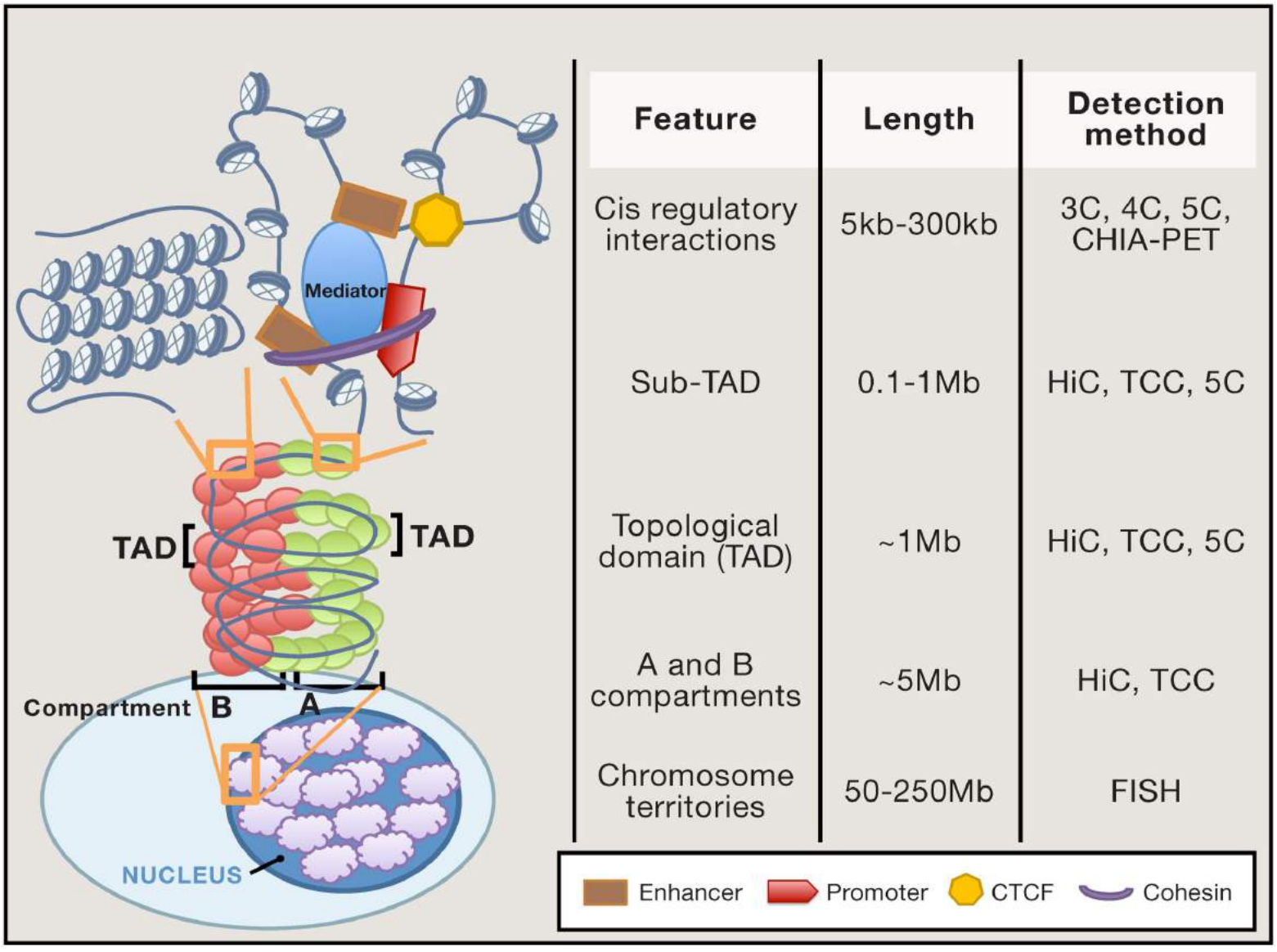
(Lieberman- Aiden et al. (2009) *Science*)

Chromatin architecture



- A and B compartments separate – A is mostly active, while B is mostly inactive chromatin
- LAD – Lamin Associated Domain (inactive)
- TAD – Topologically Associated Domain

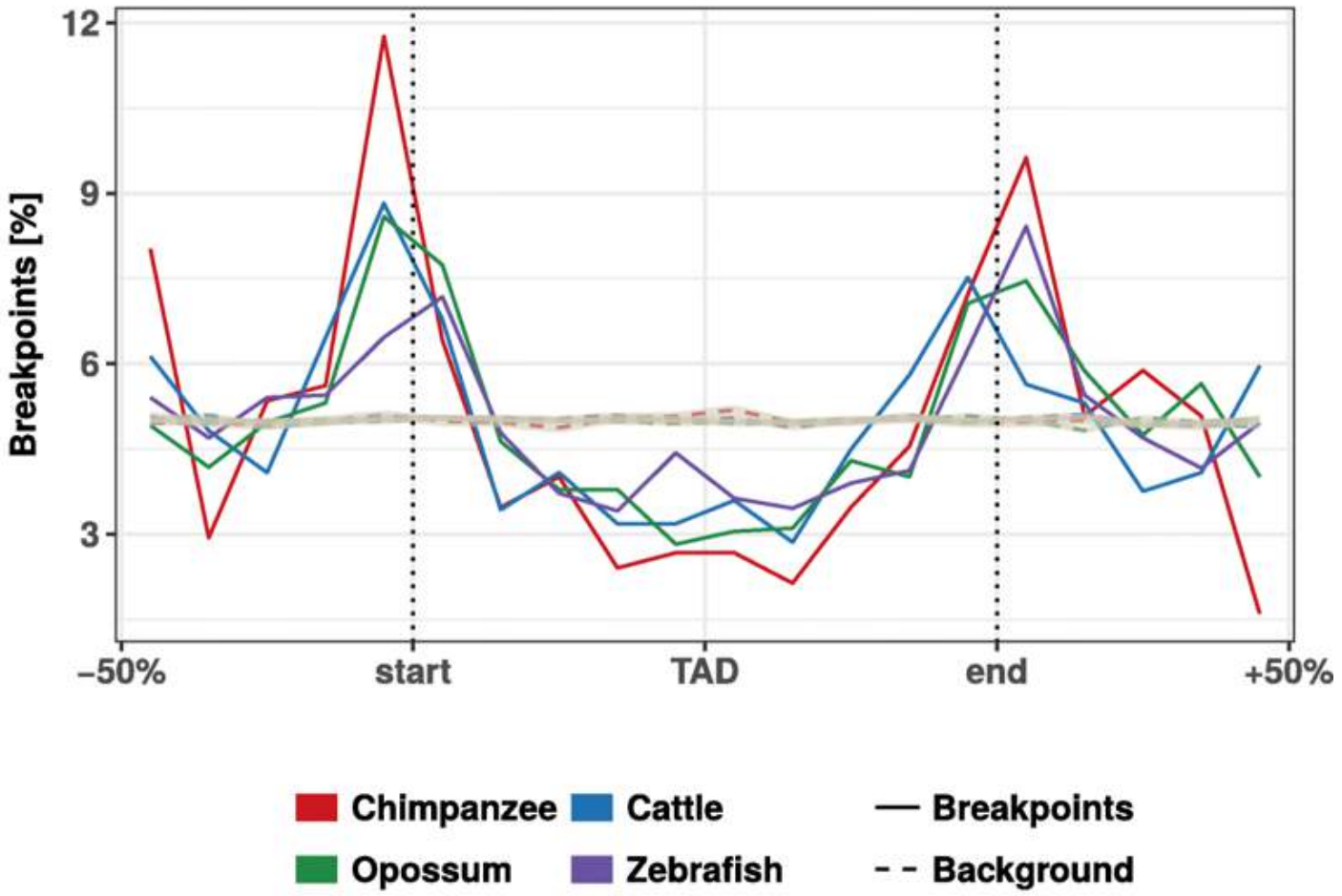
The organization of the chromatin



Feature	Length	Detection method
Cis regulatory interactions	5kb-300kb	3C, 4C, 5C, CHIA-PET
Sub-TAD	0.1-1Mb	HiC, TCC, 5C
Topological domain (TAD)	~1Mb	HiC, TCC, 5C
A and B compartments	~5Mb	HiC, TCC
Chromosome territories	50-250Mb	FISH

(Rivera and Ren (2013) *Science*)

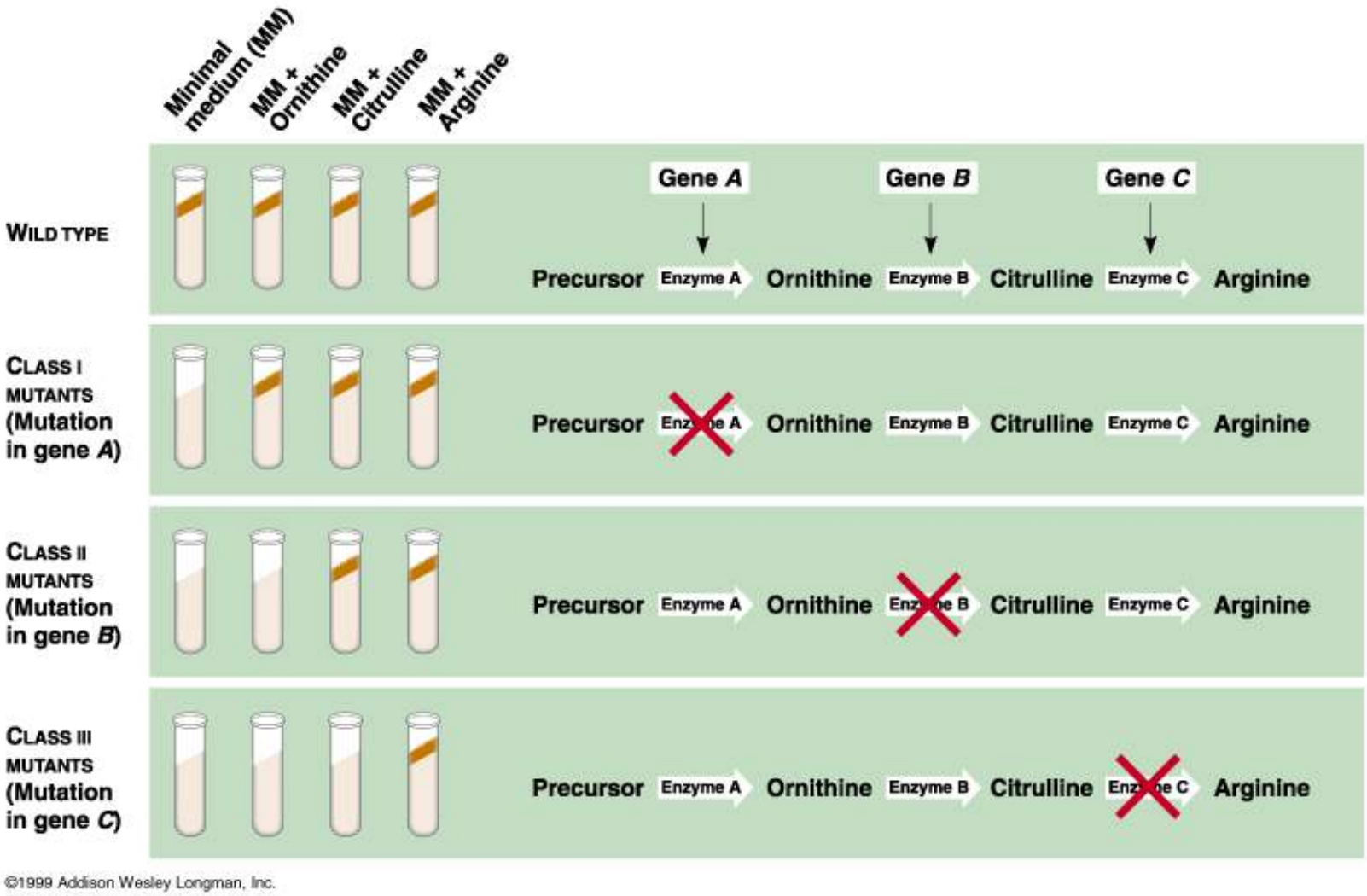
TADs function as evolutionary units



- Genome rearrangement breakpoints are enriched at TAD boundaries

(Krefting et al. (2018) *BMC Bio*)

What is a gene (in it's physical form)?



Beadle – Tatum experiment: "one gene - one enzyme"

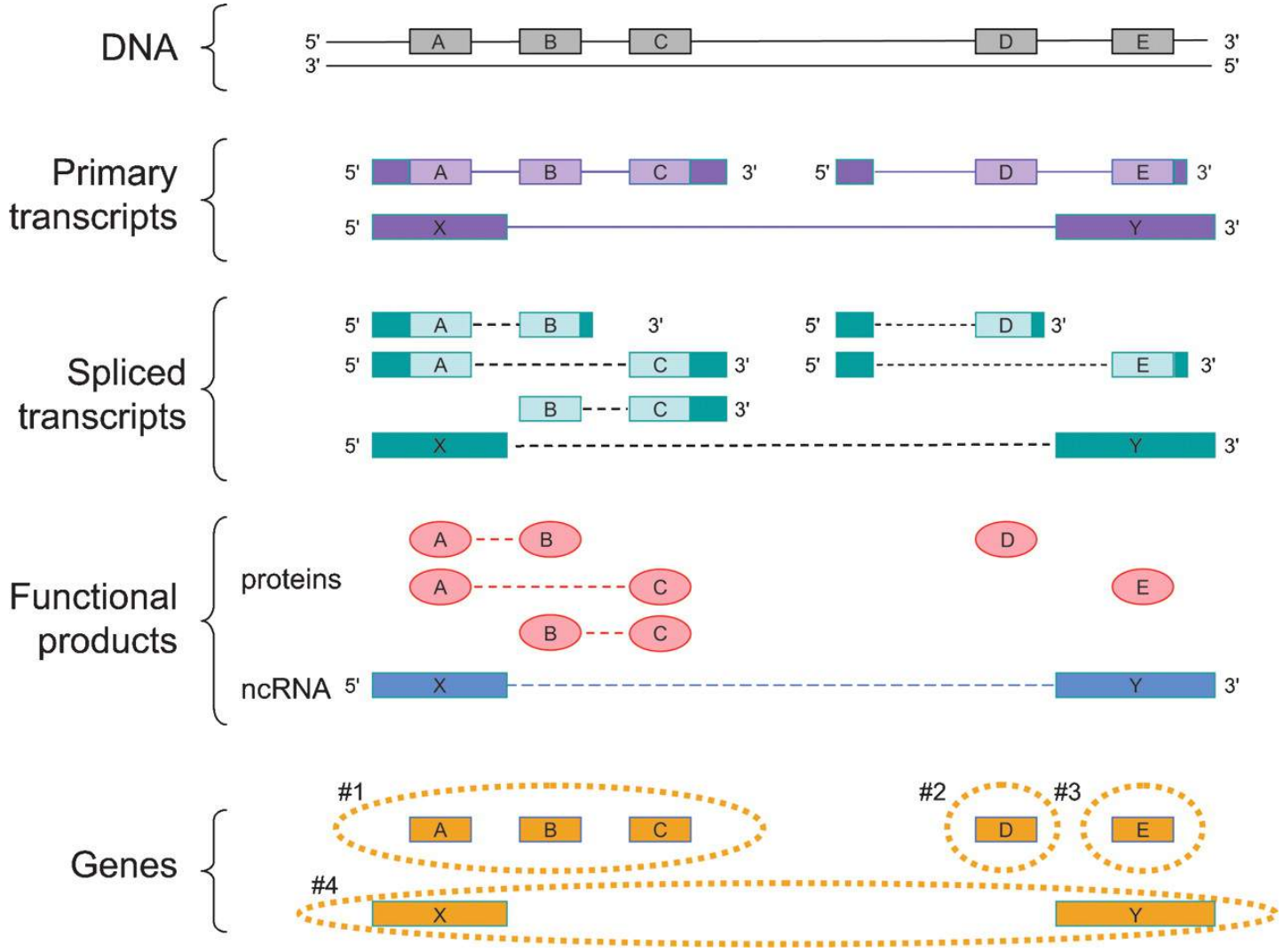
Some problems with the "one gene - one enzyme" definition of the gene

Table 1. Phenomena complicating the concept of the gene

Phenomenon	Description	Issue
Gene location and structure		
Intronic genes	A gene exists within an intron of another (Henikoff et al. 1986)	Two genes in the same locus
Genes with overlapping reading frames	A DNA region may code for two different protein products in different reading frames (Contreras et al. 1977)	No one-to-one correspondence between DNA and protein sequence
Enhancers, silencers	Distant regulatory elements (Spilianakis et al. 2005)	DNA sequences determining expression can be widely separated from one another in genome. Many-to-many relationship between genes and their enhancers.
Post-transcriptional events		
Alternative splicing of RNA	One transcript can generate multiple mRNAs, resulting in different protein products (Berget et al. 1977; Gelinas and Roberts 1977)	Multiple products from one genetic locus; information in DNA not linearly related to that on protein
Alternatively spliced products with alternate reading frames	Alternative reading frames of the INK4a tumor suppressor gene encodes two unrelated proteins (Quelle et al. 1995)	Two alternative splicing products of a pre-mRNA produce protein products with no sequence in common
RNA *trans*-splicing, homotypic *trans*-splicing	Distant DNA sequences can code for transcripts ligated in various combinations (Borst 1986). Two identical transcripts of a gene can *trans*-splice to generate an mRNA where the same exon sequence is repeated (Takahara et al. 2000).	A protein can result from the combined information encoded in multiple transcripts
RNA editing	RNA is enzymatically modified (Eisen 1988)	The information on the DNA is not encoded directly into RNA sequence

(Gerstein et al. (2007) *Genome Res*)

Some problems with the “one gene - one enzyme” definition of the gene



(Gerstein et al. (2007) *Genome Res*)

So what is the “gene” after all?

“The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.”

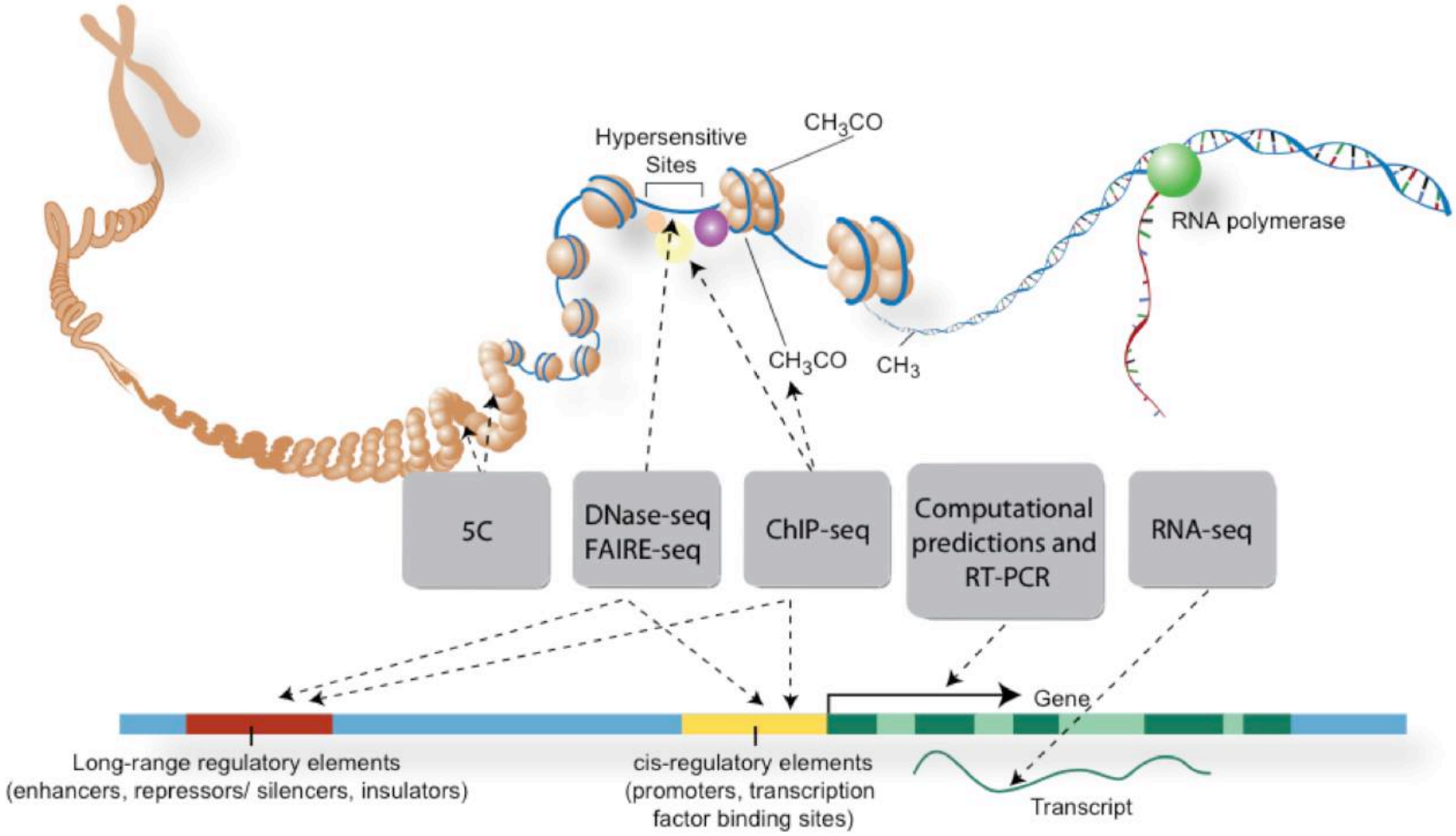
(Gerstein et al. (2007) *Genome Res*)

“A gene is a DNA sequence (whose component segments do not necessarily need to be physically contiguous) that specifies one or more sequence-related RNAs/proteins that are both evoked by GRNs and participate as elements in GRNs, often with indirect effects, or as outputs of GRNs, the latter yielding more direct phenotypic effects.”

(GRN - Gene Regulatory Network)

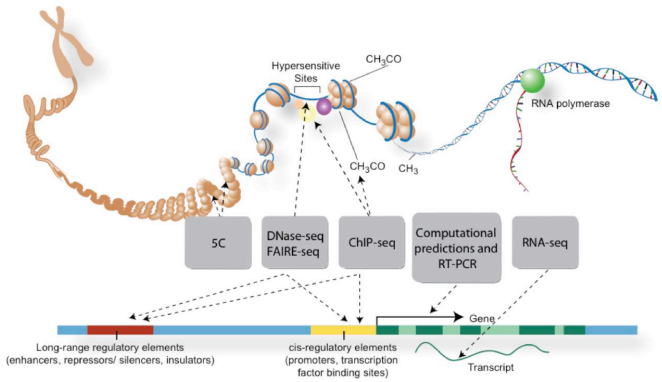
(Portin and Wilkins (2017) *Genetics*)

ENCODE – Encyclopedia of DNA Elements



(Rinn et al. (2007) *Cell*)

ENCODE – Encyclopedia of DNA Elements



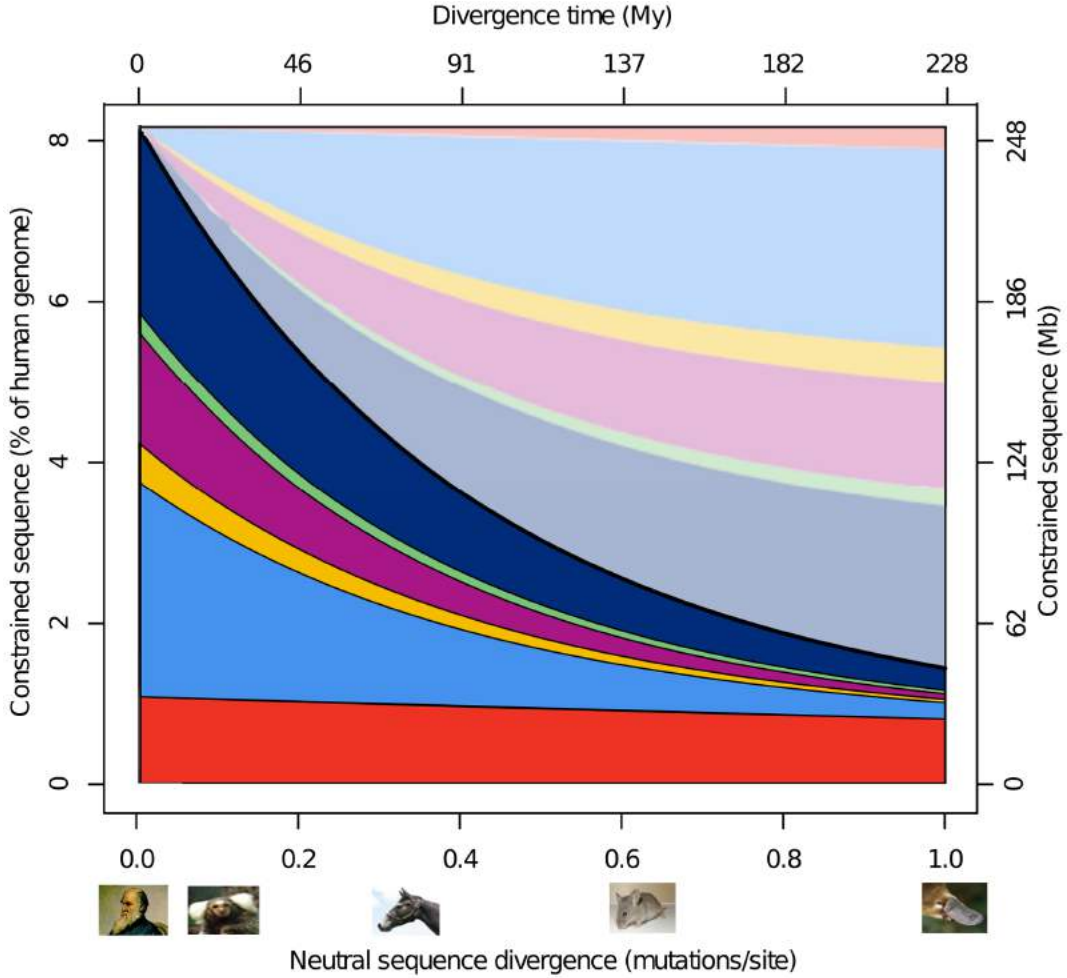
"The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type. Much of the genome lies close to a regulatory event: 95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE."

(ENCODE Consortium (2012) *Nature*)

BUT: <2% of the human genome encodes proteins, thus if the majority of the non-coding sequence is functional, it has to code fro ncRNAs or has to be regulatory sequence

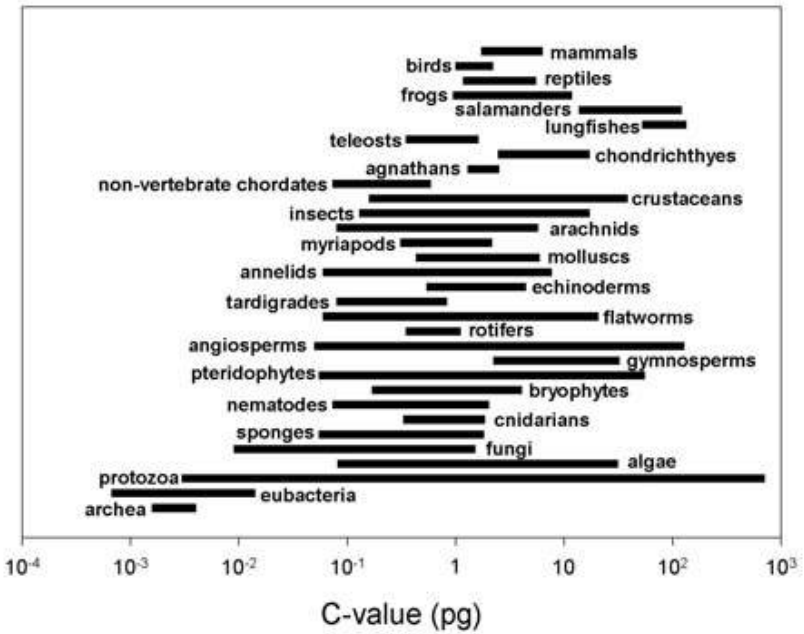
BUT2: If indeed >80% of the genome is functional, how come that only ~8% is under selection?

The coding amount of the human genome: ~8.2%



(Rands et al. 2014 *PLOS Genet*)

What does “function” per ENCODE really mean?



C-value paradox: genome size can change enormously even within taxa, and does not correlate with organismal complexity

- For example the genome for lungfish is 130 Gb körüli (the human genome is 3Gb), while the Fugu genome is only 400 Mb

- If 80% of a genome would be functional such reduction (with essentially identical physiology) would not be possible

- Animal genomes are rich in “jumping genes” (transposons) - 30-70% of the genome – and these once active sequences sometimes can reactivate

- Genes are born and die contiuously, and in many phases of this process we will detect transcription
- Similaryl, TF s can bind the genome randomly (at non-functional binding sites) and in these cases transcription could be detected in the loose chromatin

- The original definition of ENCODE considers noise functional!

A population genetics argument for less than 25% of the human genome being important

Table 1

Replacement Level Fertility Values in Humans As a Function of the Deleterious Mutation Rate (μ_{del}) and the Fraction of the Genome that is Functional[a,b]

μ_{del}	Functional Fraction of the Genome										
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.50	0.80	1.00
4.0×10^{-10}	1.1	1.3	1.4	1.6	1.8	2.1	2.4	2.7	3.4	7.1	12
5.0×10^{-10}	1.2	1.4	1.6	1.8	2.2	2.5	2.9	3.4	4.6	12	22
6.0×10^{-10}	1.2	1.4	1.7	2.1	2.5	3.0	3.6	4.4	6.3	19	40
7.0×10^{-10}	1.2	1.5	1.9	2.4	2.9	3.6	4.5	5.6	8.6	31	74
8.0×10^{-10}	1.3	1.6	2.1	2.7	3.4	4.4	5.6	7.1	12	51	136
9.0×10^{-10}	1.3	1.7	2.3	3.0	4.0	5.3	6.9	9.1	16	83	252
1.0×10^{-09}	1.4	1.8	2.5	3.4	4.6	6.3	8.6	12	22	136	466
2.0×10^{-09}	1.8	3.4	6.3	12	22	40	74	136	466	1.9×10^{04}	2.2×10^{05}
3.0×10^{-09}	2.5	6.3	16	40	100	252	633	1.6×10^{03}	1.0×10^{04}	2.5×10^{06}	1.0×10^{08}
4.0×10^{-09}	3.4	12	40	136	466	1.6×10^{03}	5.4×10^{03}	1.9×10^{04}	2.2×10^{05}	3.5×10^{08}	4.7×10^{10}
5.0×10^{-09}	4.6	22	100	466	2.2×10^{03}	1.0×10^{04}	4.7×10^{04}	2.2×10^{05}	4.7×10^{06}	4.7×10^{10}	2.2×10^{13}
6.0×10^{-09}	6.3	40	252	1.6×10^{03}	1.0×10^{04}	6.4×10^{04}	4.0×10^{05}	2.5×10^{06}	1.0×10^{08}	6.4×10^{12}	1.0×10^{16}
7.0×10^{-09}	8.6	74	633	5.4×10^{03}	4.7×10^{04}	4.0×10^{05}	3.4×10^{06}	3.0×10^{07}	2.2×10^{09}	8.8×10^{14}	4.8×10^{18}
8.0×10^{-09}	12	136	1.6×10^{03}	1.9×10^{04}	2.2×10^{05}	2.5×10^{06}	3.0×10^{07}	3.5×10^{08}	4.7×10^{10}	1.2×10^{17}	2.2×10^{21}
9.0×10^{-09}	16	252	4.0×10^{03}	6.4×10^{04}	1.0×10^{06}	1.6×10^{07}	2.5×10^{08}	4.0×10^{09}	1.0×10^{12}	1.6×10^{19}	1.0×10^{24}
1.0×10^{-08}	22	466	1.0×10^{04}	2.2×10^{05}	4.7×10^{06}	1.0×10^{08}	2.2×10^{09}	4.7×101^{0}	2.2×10^{13}	2.2×10^{21}	4.8×10^{26}
2.0×10^{-08}	466	2.2×10^{05}	1.0×10^{08}	4.7×10^{10}	2.2×10^{13}	1.0×10^{16}	4.8×10^{18}	2.2×10^{21}	4.8×10^{26}	4.9×10^{42}	2.3×10^{53}

[a]Values above 1.8 are unrealistically high in humans.

[b]A more comprehensive table can be found in supplementary material, Supplementary Material online.

- μ_{del}=4x10^{-10}/nucleotide/generation – if only nonsense mutations are deleterious
- M_{del}=2x10^{-8}/nucleotide/generation – if missense mutations are deleterious, too

(Graur 2017 *Genome Bio Evol*)