Lecture 5 - Promoters and transcription

contatto per studentessa cagliari

Analysis of transcriptome

-all kind of organisms from S.cerevisiae to human cells

-an armamentary of methods, from EST to microarrays to RNA-Seq

have led to the concept that the genome is transcribed much more than previously thought

This phenomenon was named “pervasive transcription”.

Most of “unusual” transcript are present at very low level and they are at least in part questioned. However, there are many RNA that have been characterized.

Conceptually, these RNA should be classified as:

a) Related to genes (promoters, intronic, antisense)

b) unrelated to genes (intergenic)

1) stable
2) unstable

I. coding
II. noncoding

Common problem:
How to distinguish “transcripts” from RNA fragments or processed RNAs ?

in addition

RNA transcripts (even capped) may arise from unexpected phenomena, such as cleavage of longer transcript, *trans*-splicing followed by processing, editing, and even by RNA-dependent RNA polymerization (even though such a polymerase, found in lower eukaryotes (e.g *C. elegans*), has never been observed in Vertebrates).

Other functional annotations may help, such as:

- the presence of RNA polymerase(s)
- the presence of other proteins necessary for transcription (and RNA processing)
- histone modifications that detect the local functional status of the chromatin
- DNA covalent modifications (e.g. CpG methylation)
- ....

Il problema della corretta mappatura delle T.U.

Definizione del 5’ e 3’ (CAGE, SAGE, RNA-Seq)

Definizione esatta degli esoni (Tiling arrays, RNA-Seq)

Definizione del promotore (CAGE, RNA-Seq + studies)

Promoters and transcription

- what we know from classical books

-How to define a TSS ?
    - gene by gene methods
    - genomic methods

- One study for the identification of promoters genome-wide

- Classification of non protein-coding RNAs (intragenic and intergenic transcripts)

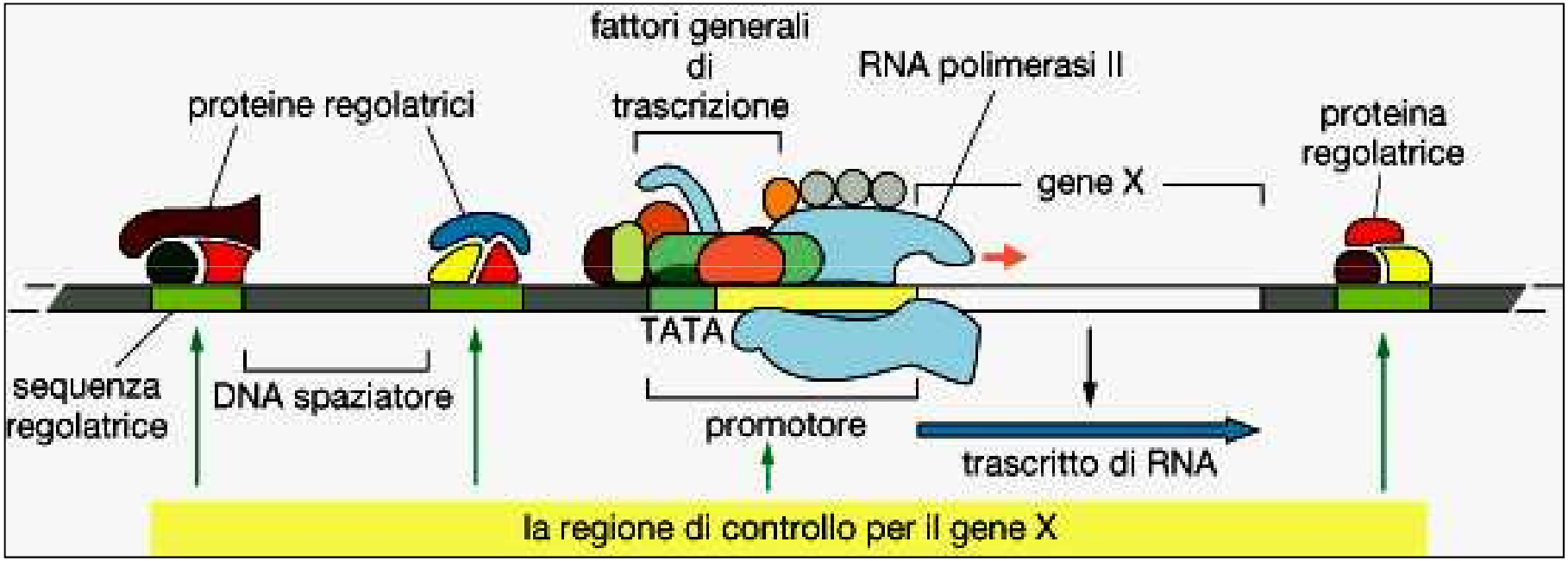
Mammalian RNA polymerase II core promoters: insights from genome-wide studies

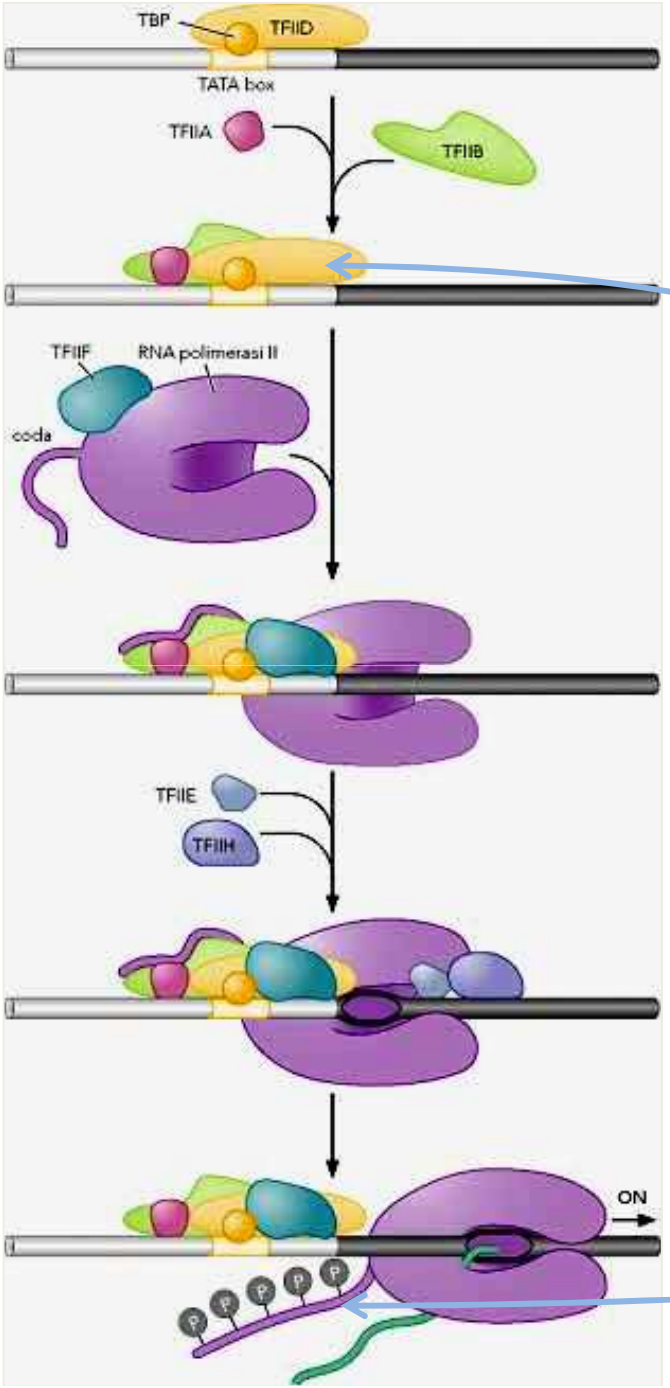
Review 1

Albin Sandelin[*‡], *Piero Carninci*[‡§], *Boris Lenhard*[‖], *Jasmina Ponjavic*[¶], *Yoshihide Hayashizaki*[‡§] *and David A. Hume*[#]

Abstract | The identification and characterization of mammalian core promoters and transcription start sites is a prerequisite to understanding how RNA polymerase II transcription is controlled. New experimental technologies have enabled genome-wide discovery and characterization of core promoters, revealing that most mammalian genes do not conform to the simple model in which a TATA box directs transcription from a single defined nucleotide position. In fact, most genes have multiple promoters, within which there are multiple start sites, and alternative promoter usage generates diversity and complexity in the mammalian transcriptome and proteome. Promoters can be described by their start site usage distribution, which is coupled to the occurrence of *cis*-regulatory elements, gene function and evolutionary constraints. A comprehensive survey of mammalian promoters is a major step towards describing and understanding transcriptional control networks.

How many important concepts of eukaryotic transcription are resumed in this figure ?





!

Transcription in Eukaryotes

General Transcription Factors

Regulatory elements in Eukaryotes

This is TFIID, a multisubunit protein composed of the TATA-binding protein (TBP) and associated factors (TAFs).

The RNA Polymerase II (RNAPolII) carboxy-terminal domain (CTD) has multiple 7-aa repeats pluriphoshorylated when active

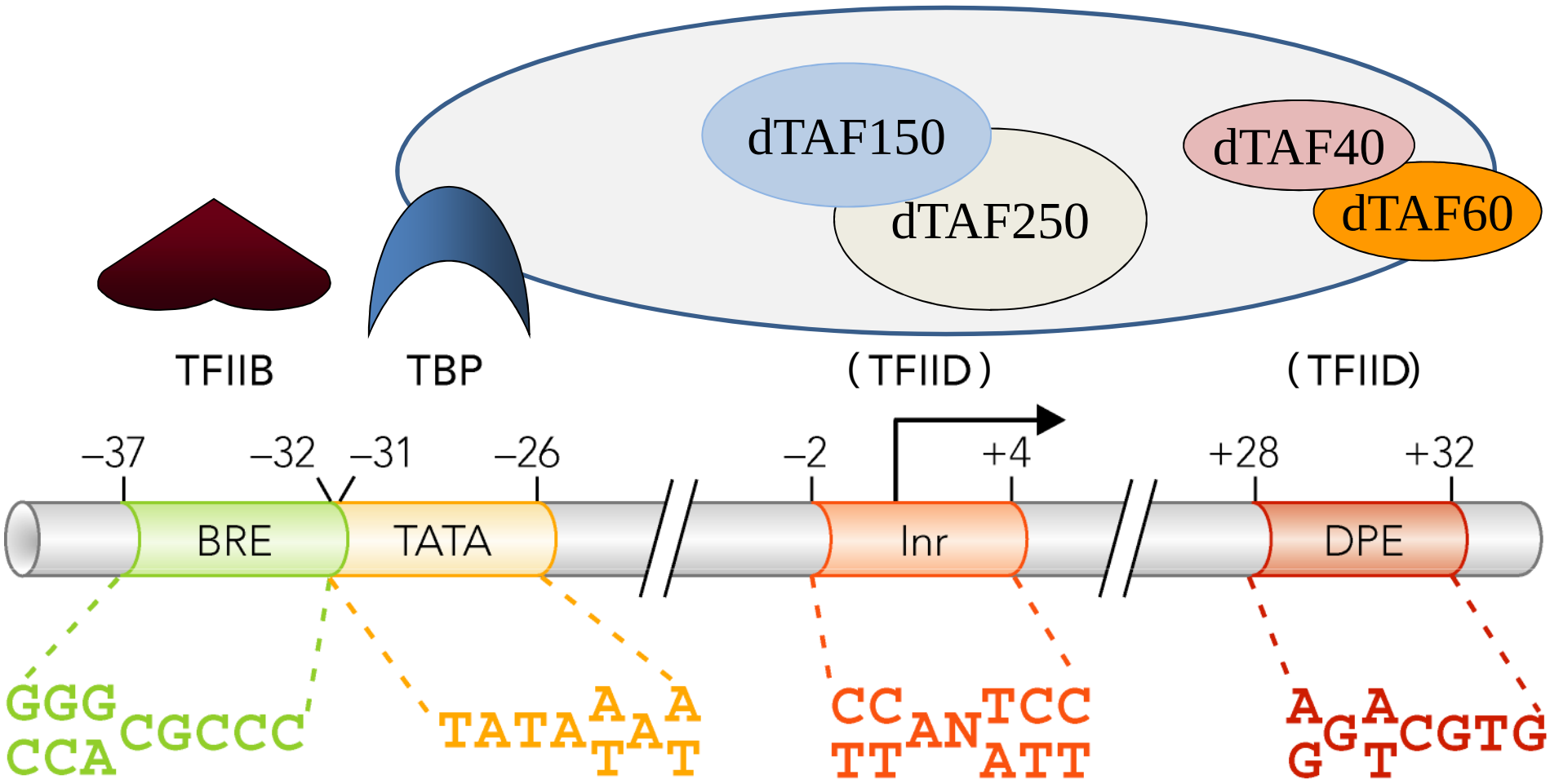
table 26–1

Proteins Required for Transcription at the RNA Polymerase II Promoters of Eukaryotes

Transcription factor	Number of subunits	Subunit M_r	Functions
Initiation			
RNA polymerase II	12	10,000–220,000	Catalyzes RNA synthesis
TBP (TATA-binding protein)	1	38,000	Specifically recognizes the TATA box
TFIIA	3	12,000, 19,000, 35,000	Stabilizes binding of TFIIB and TBP to the promoter
TFIIB	1	35,000	Binds to TBP; recruits RNA polymerase–TFIIF complex
TFIID	12	15,000–250,000	Interacts with positive and negative regulatory proteins
TFIIE	2	34,000, 57,000	Recruits TFIIH; ATPase and helicase activities
TFIIF	2	30,000, 74,000	Binds tightly to RNA polymerase II; binds to TFIIB and prevents binding of RNA polymerase to nonspecific DNA sequences
TFIIH	12	35,000–89,000	Unwinds DNA at promoter; phosphorylates RNA polymerase; recruits nucleotide-excision repair complex
Elongation*			
ELL†	1	80,000	
P-TEFb	2	43,000, 124,000	
SII (TFIIS)	1	38,000	
Elongin (SIII)	3	15,000, 18,000, 110,000	

*All elongation factors suppress the pausing or arrest of transcription by the RNA polymerase II – TFIIF complex.

†The name is derived from the term *e*leven-nineteen *l*ysine-rich *l*eukemia. The gene for the factor ELL is the site of chromosomal recombination events frequently associated with the cancerous condition known as acute myeloid leukemia.



The “textbook” promoter

“consensus” sequences are found by computing the frequencies of bases at any position in several genes. Example:

gene 1:		a	t	t	g	t	c	t	a
gene 2:		a	t	a	g	t	g	a	g
gene 3:		a	t	t	c	t	g	g	c
gene 4:		a	t	t	g	t	c	c	g
gene 5:		a	t	t	c	t	g	a	t
gene 6:		a	t	a	g	g	c	t	c
etc:		-	-	-	-	-	-	-	-
frequency	a	100	0	35	0	0	0	25	20
	t	0	100	65	0	90	0	25	20
	c	0	0	0	35	0	50	20	30
	g	0	0	0	65	10	50	30	30
Consensus:		**A**	**T**	**W**	**S**	**T**	**S**	**N**	**N**

Symbol	Name
A	Adenine
C	Cytosine
T	Thymine
G	Guanine
U	Uracil
W	A or T
R	A or G
K	G or T
Y	C or T
S	C or G
M	A or C
B	C, G or T
H	A, C, or T
D	A, G, or T
V	A, C, or G
N	A, C, G, or T

CCAAT

47	690	CAATGGGA	11.3
47	601	CAATCAGC	13.1
46	708	CAATCAGA	14.4
45	310	CCAATCGG	8.1
46	871	CCAATCCC	8.0
48	620	CCAATCAC	11.7
47	1061	CCAATCAG	23.6
47	306	CCAATCGC	13.3
47	770	CCAATGGG	31.1
47	896	GCCAATCA	22.5
46	361	GCCAATAG	9.2
48	357	GCCAATCG	12.4
46	578	GCCAATGA	17.0
47	775	GCCAATGG	26.8
46	553	GCCAATCC	9.7
47	537	TCCAATCA	7.0
47	220	ACCAATCG	14.7
47	469	ACCAATGG	17.8
46	583	ACCAATCA	17.4
47	384	GACCAATG	9.8
47	400	GACCAATC	19.2
47	893	AGCCAATC	19.3
46	748	AGCCAATG	13.8
47	680	GGCCAATG	11.7
48	658	GGCCAATC	24.0
47	547	GAGCCAAT	10.2
47	324	GGACCAAT	8.8
47	483	GGGCCAAT	12.4
48	509	CGGCCAAT	10.9
47	1039	CAGCCAAT	31.4
47	774	TCAGCCAA	10.5
47	1036	GCAGCCAA	7.0

SP1

48	1332	GCCACGCC	15.7
48	8136	GCCCCGCC	25.2
48	3078	CGCCCCTC	7.3
48	5248	CGCCCCGC	13.7
48	3141	CGCCCCCT	7.4
48	7055	CCGCCCCC	18.1
47	2106	CCGCCCAC	8.1
48	5783	CCGCCTCC	7.0
47	5204	CCGCCCCG	16.6
48	3688	CCGCCCCT	12.6
48	10767	CCCGCCCC	28.3
48	1170	ACGCCCCC	15.4
48	829	ACGCCCCG	7.9
48	1639	CACGCCCC	13.9
48	2890	CCCGCCCT	8.9
47	2334	CCCGCCCA	10.8
48	2462	TCCGCCCC	8.4
48	4767	CCCGCCTC	18.8
48	3366	CTCCGCCC	11.8
48	11029	CCCCGCCC	31.3
48	3190	CCCCGCCT	12.5
49	918	TTCCGCCC	17.8
48	2673	GCTCCGCC	7.2
49	1213	CTTCCGCC	7.9
48	4947	GGCCCCGC	7.1
47	5139	CCTCCCTC	8.1
48	7985	CCCCTCCC	7.4

CRE

50	484	TGACGTCA	18.4
49	282	ATGACGTC	8.5
50	503	CTGACGTC	9.3
48	635	GTGACGTC	13.5
50	313	GTGACGCA	7.4
49	345	AGTGACGT	9.4
49	294	CGTGACGC	8.0
49	280	CGTGACGT	10.2
48	379	GGTGACGT	7.1
50	264	TGTGACGT	11.4
49	241	ACGTGACG	10.3
49	472	ACGTGACC	8.4
49	332	AAGTGACG	23.9
50	769	GAAGTGAC	10.4

TATA

49	486	CCTATAAA	9.3
49	571	GCTATAAA	7.1
49	496	CTATAAAG	10.1
49	809	TTATAAAG	10.9
49	861!	TATAAAAG	11.4
49	417	TATATAAG	9.7
49	542!	TATAAAGG	28.0
49	860!	ATAAAAGG	17.1
49	630	TAAAAGGC	9.9

NRF-1

50	1240	TGCGCCTG	11.9
50	2300	GCGCCTGC	12.3
50	1767	CGCCTGCG	11.6
50	463	CGCATGCG	15.5
50	2154	GCCTGCGC	7.8
48	1205	GCGTGCGC	7.4
50	1041	CCTGCGCA	12.9
50	903	ACTGCGCC	8.0
50	572	TGCGCATG	8.5
49	386	CGCGCATG	11.1
50	1179	GCGCATGC	18.5

ETS

49	1546	AGGAAGTG	7.6
49	1324	GGAAGTGA	16.2
49	923	GGAAGTGC	11.9
50	1892	GGAAGTGG	7.5
49	284	CGGAAGTA	23.1
50	484	CGGAAGCA	13.8
50	426	CGGAAGTC	24.8
51	402	CGGAAGTT	8.0
50	991	CGGAAGTG	29.5
51	356	CGGAAATG	7.8
49	567	CGGAAGCT	8.4
50	824	CGGAAGCG	19.2
49	1150	CCGGAAGC	20.9
50	1030	CCGGAAGT	31.9
51	459	CCGGAAAC	13.1
50	600	ACCGGAAG	40.6
50	1096	GCCGGAAG	23.2
49	1224	CCCGGAAG	20.1
51	603	ACCCGGAA	7.8
50	382	CACCGGAA	12.9
49	401	GACCGGAA	7.4

ATG/KOZAK

52	960!	CCAAGATG	7.5
52	617!	GCAAGATG	13.7
51	543	GCGCCATG	9.3
53	688!	GCACCATG	9.9
52	1152	CAGCCATG	11.1
53	1005	CACCATGG	8.6
52	426	CGCCATGC	9.0
52	931	CGCCATGG	9.4
52	1081!	CAAGATGG	39.6
51	1018	CAAAATGG	9.8
51	1048	AAAATGGC	16.8
52	1202!	AAGATGGC	36.9
52	881!	AGATGGCG	40.2
51	654	ACATGGCG	13.5
51	436	AAATGGCG	23.4
51	414	AATGGCGG	12.9
52	1026!	GATGGCGG	27.2
52	920	CATGGCGG	18.4
54	291	CATGGCGT	11.1
51	583	ATGGCGCC	23.6
52	1125!	ATGGCGGC	27.7
52	619	ATGGCGGG	8.2
52	468!	ATGGCGGA	16.0
52	966	ATGGCTGC	15.8
54	791!	CCAGGTAA	7.1
56	307!	CGCAGGTA	8.2
51	443	CGCAGTCT	8.1
55	1638!	GGTGAGTG	7.6
53	848!	TGGTGAGT	7.9
52	1414	GAGAGCTG	7.4
53	3887!	CTGCTGCT	9.1
53	3570!	TGCTGCTG	8.0

Highlighted Sequences

CCAAT

*42-49	10.0	RRCCAATSR	1170

SP1

*44-50	8.8	CCCCGCCC	3424
*44-50	8.3	GCCCCGCC	2687
44-50	8.7	CCCGCCCC	2257

Clus1

*43-51	8.2	TCTCGCGA	211

USF

*46-50	7.3	CCACGTGA	123
*47-51	7.6	TCACGTGA	89

CRE

*45-50	9.5	TGACGTCA	190
*45-51	5.1	TGATGTCA	125
46-50	7.1	TTGCGTCA	48

TATA

*48-49!	7.7	TATAAAD	472
48-49!	2.4	TATATAD	349

FitzGerald PC et al. Clustering of DNA Sequences in Human Promoters. Genome Res. 2004 August; 14(8): 1562–1574.

CRE			
*45-50	9.5	TGACGTCA	190
*45-51	5.1	TGATGTCA	125
46-50	7.1	TTGCGTCA	48
TATA			
*48-49!	7.7	TATAAAD	472
48-49!	2.4	TATATAD	349
48-50!	5.5	TATAAGD	217
NRF-1			
*46-51	7.2	CGCVTGCG	906
ETS			
*44-51	13.1	VCCGGAARY	1031

One hundred fifty-six DNA sequences are grouped into related sequences and arranged by their peak position relative to the TSS. From the left the table contains the most abundant bin, the number of times the sequence occurs in the distribution, the 8-mer sequence, and finally the *P* value (see text). The end of the table contains consensus sequences. Here the leftmost numbers are the bins defining the peak, followed by the clustering factor (CF), the consensus sequence, and finally the number of occurrences of the sequence in the bins that comprise the peak. Exclamation point (!) denotes sequences that are at least threefold more abundant in the maximum bin on the DNA strand presented in the table than on the opposite strand. The asterisk (*) denotes sequences used in Tables 2 and 3. IUPAC letters used to represent degenerate bases are R (G,A), W (A,T), Y (T,C), K (G,T), V (G, C, A), D (G,A,T), and N (A,T,G,C).

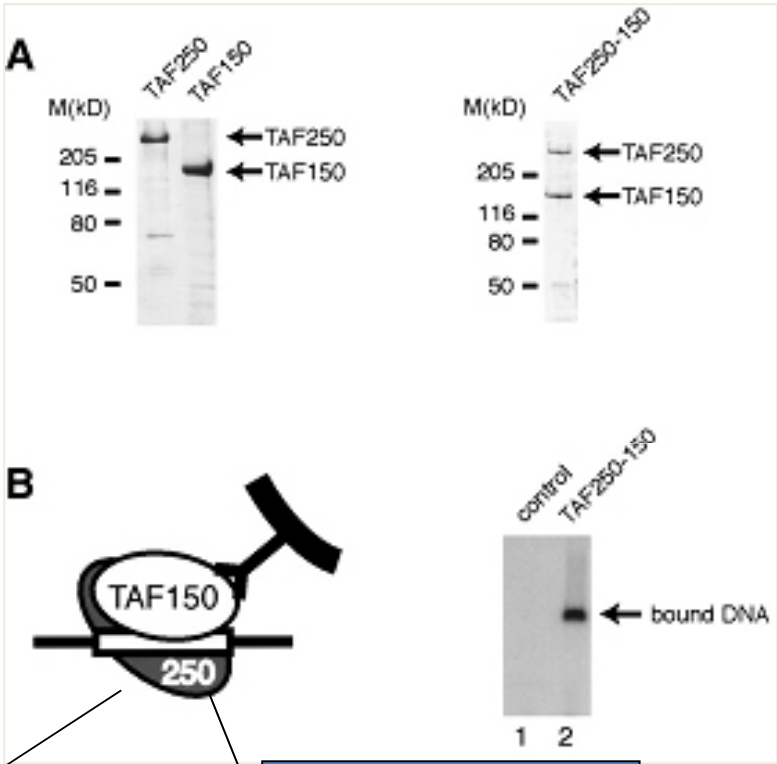
The “Initiator” element (INR) is ill-defined and much less conserved than the TATA-box

The INR was demonstrated by bioinformatics but also using direct interaction studies using isolated recombinant subunits of the TFIID protein, the so-called TAFs (TBP –associated factors).

In vitro transcripiton experiments using recombinant TBP, TFIIB and individual TAFs showed that TAF150+TAF250 (drosophila) can correctly initiate transcription from TATA-less promoters.

Question: do TAF150/TAF250 recognize INR ? How is this element composed?

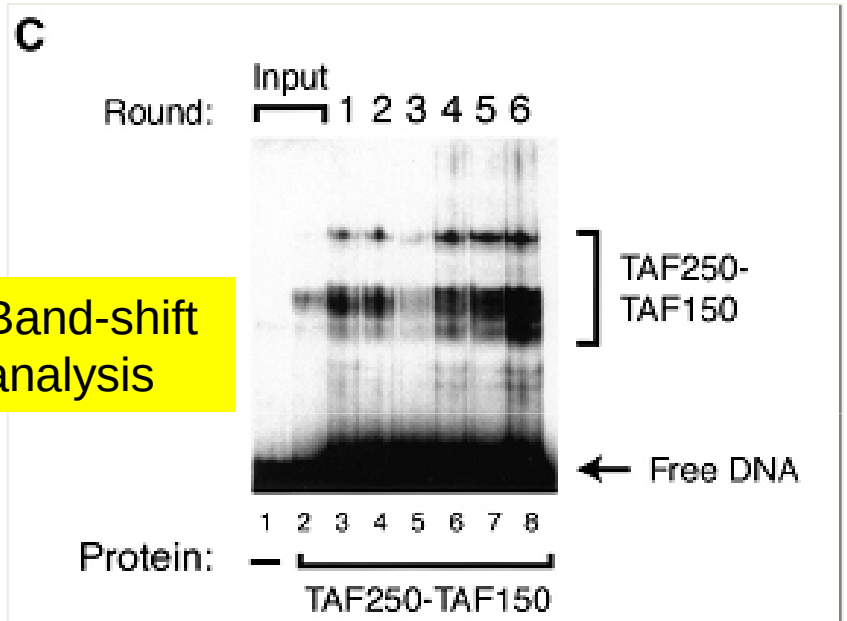
Da Chalklev e Tjian (1999), EMBO J., 18: 4835.



Recombinant TAFs

This complex was used for SELEX (see transactivation)

Band-shift analysis

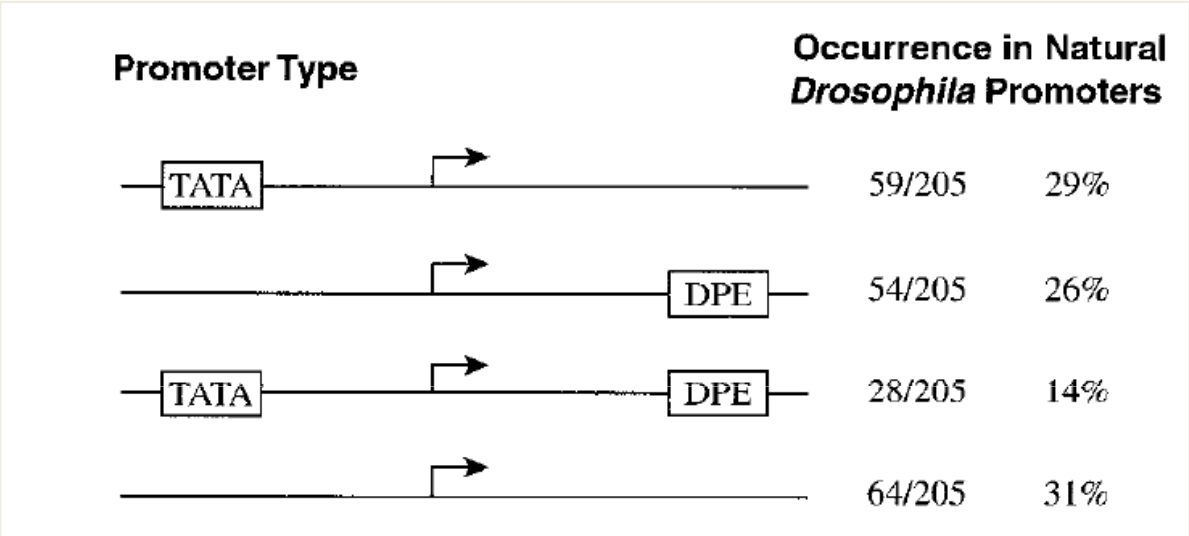


D

G	21	16	7	-	-	-	15	-	-	9	16	16	12
A	7	9	15	-	-	56	13	16	9	-	12	7	15
T	4	15	18	25	14	-	12	40	23	18	4	14	7
C	17	9	9	31	42	-	16	-	24	29	18	10	12
R	28	25	22	-	-	56	28	16	9	9	28	23	27
Y	21	24	27	56	56	-	28	40	47	47	22	24	19

INR consensus derived bioinformatically: Y C/t A N T/a Y Y

The **DPE** (downstream promoter element) was demonstrated by bioinformatic and functional analysis. TAFII40 and TAFII60 in reconstitution experiments were enough to sustain transcription initiation in TATA-less, DPE-containing promoters.



Very frequent in Drosophila

FIG. 4. The DPE appears to be present in many *Drosophila* promoters.
(A) The frequency of occurrence of the DPE appears to be comparable to that of the TATA box in *Drosophila* core promoters. A *Drosophila* core promoter database was created by aligning sequences of 205 *Drosophila* core promoters with accurately determined transcription start sites. The number of promoters that appear to possess a TATA box only, a DPE only, both elements, or neither element is shown. TATA boxes were defined as sequences with at least a 5 out of 6 match with the TATAAA sequence upstream of 220 relative to the transcription start site. DPE motifs were defined as sequences with at least a 5 out of 6 match with the DPE functional range set (Table 2) at exactly 128 to 133 relative to the start site. The *Drosophila* core promoter database is available at the website http://www-biology.ucsd.edu/labs/Kadonaga/DCPD.html .

from: Kutach & Kadonaga 2000, Mol Cell Biol 20: 4754-64.

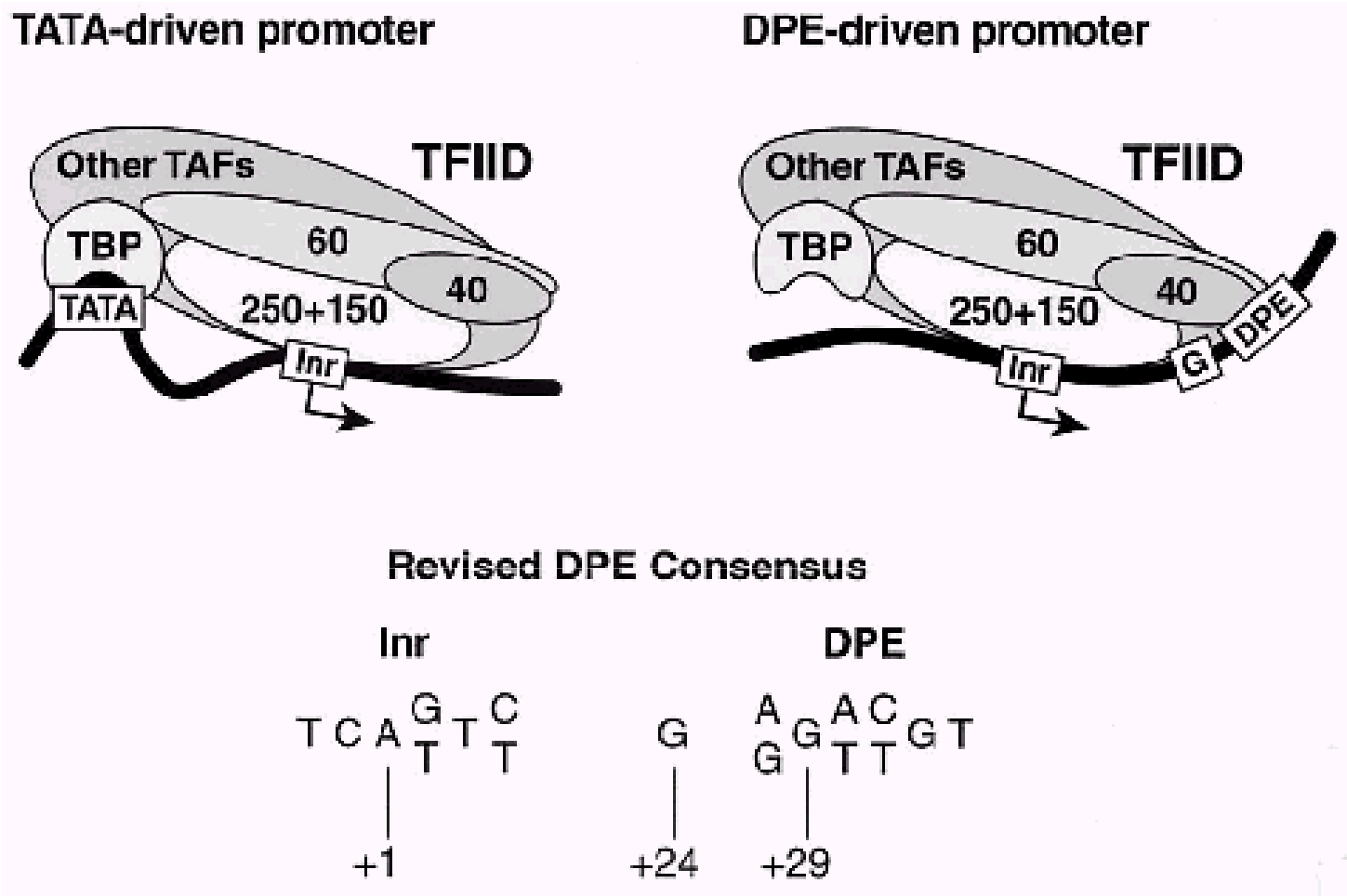
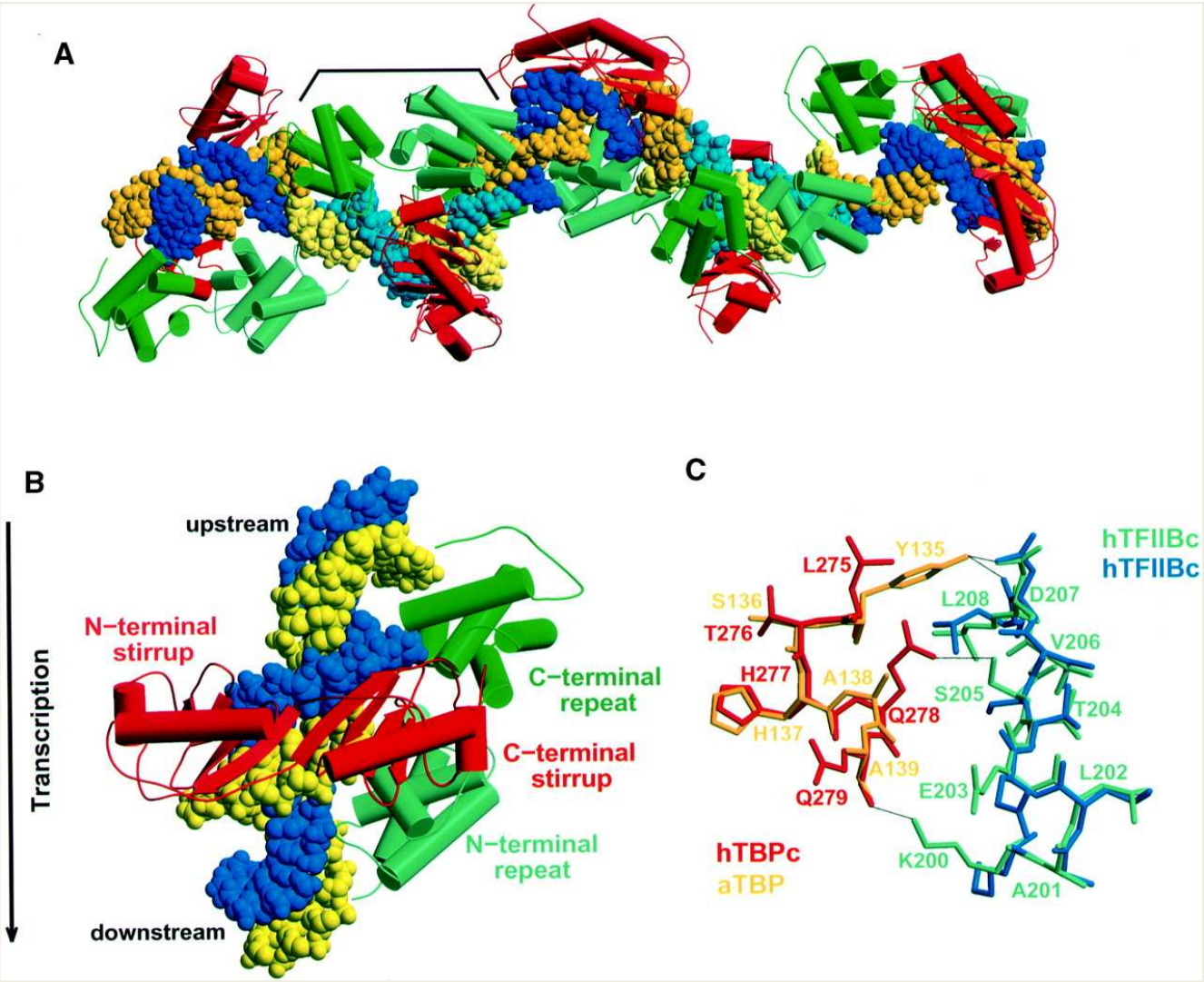


FIG. 7. A model of two distinct interactions of TFIID with TATA- versus DPE-driven core promoters. The model is discussed in the text. TAFs, TBP-associated factors.

from: Kutach & Kadonaga 2000, Mol Cell Biol 20: 4754-64.

A further promoter element, the **BRE** (TFIIB-response element) was unraveled essentially by structural analysis



Architecture of the human ternary TFIIBc-TBPc-MLP (Major late promoter) complex.
(**A**) Ribbons and space-filling representation of one asymmetric unit.
(**B**) An isolated ternary complex as viewed down the pseudo-2-fold axis of the hTBPc-TATA-box interaction.
(**C**) Species-specific differences in the TBP-TFIIB interface. Residues of hTBPc are colored in red, and those of *A.thaliana* (aTBP2) in yellow. Residues of hTFIIBc are shown in light green, those of the previously determined structure in blue.

Tsai FTF & Siegler PB. EMBO J., 19: 25-36, 2000. Structural basis of preinitiation complex assembly on human Pol II promoters

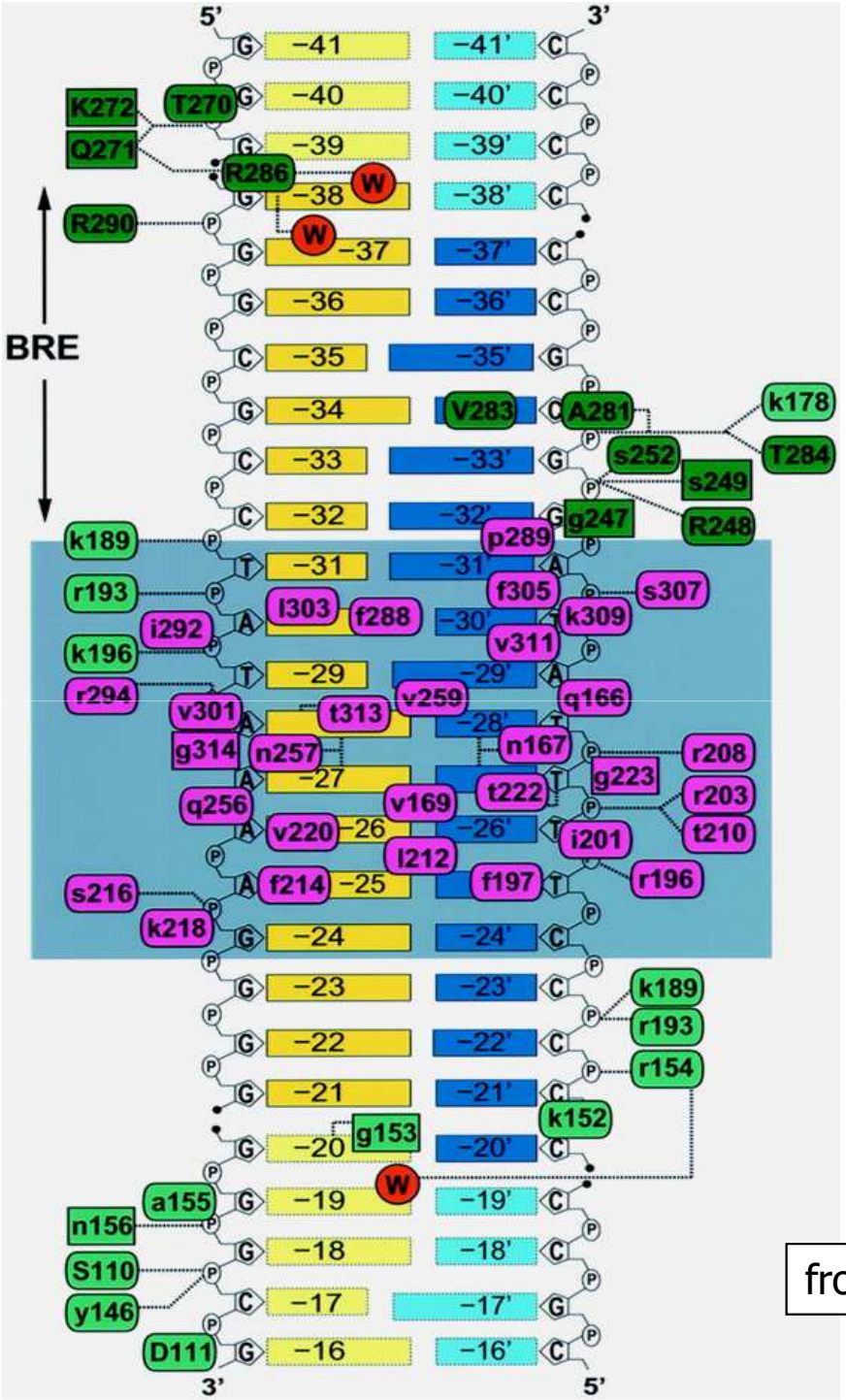


Fig. 3. Protein-DNA contacts that specify the orientation of the hTFIIBc-hTBPc-MLP complex. Complete schematic illustrating all protein-DNA interactions in the ternary complex.
Pink: TBP. Green: TFIIB.
Arrows indicate the location of the BRE.
An oval indicates an interaction between the promoter and the protein side chain, and a square an interaction with the protein main chain.
Amino acid residues that are in contact with the major groove are shown in upper-case letters, and those in contact with the minor groove in lower-case letters.
Hydrogen bonds are represented by dotted lines.

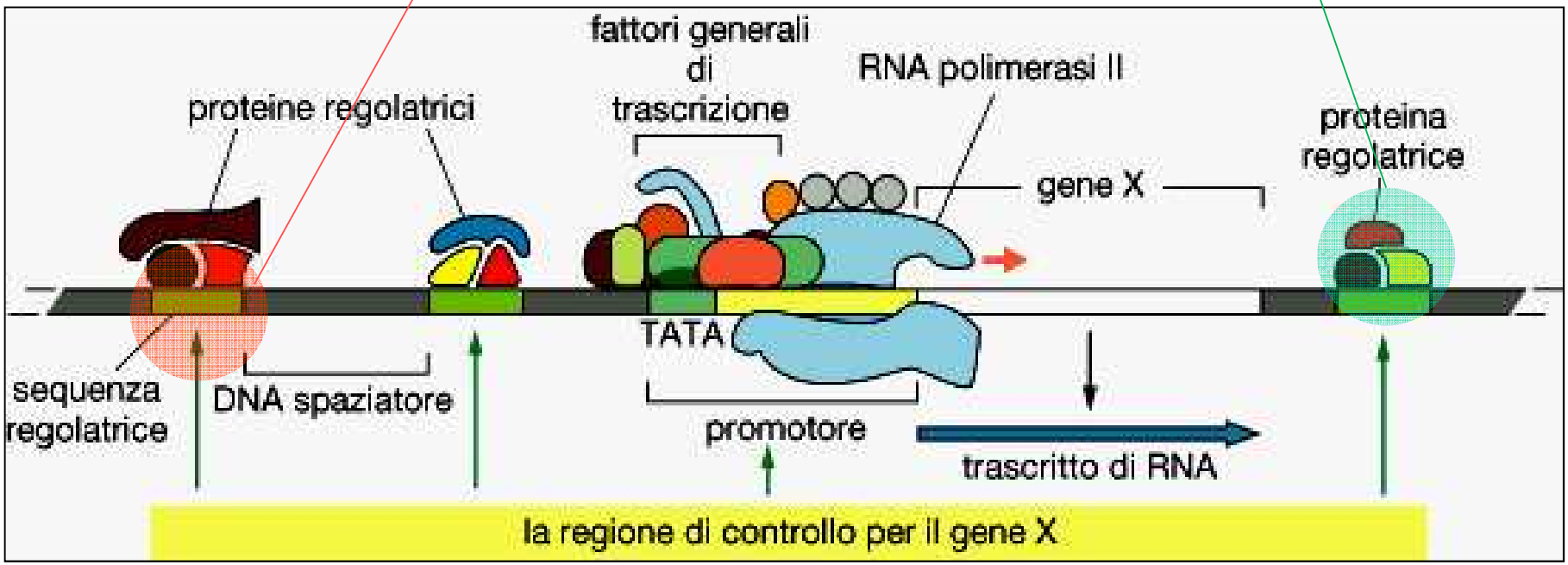
from: Tsai FTF & Siegler PB. EMBO J., 19: 25-36, 2000

Regulatory factors

The Human Genome contains genes encoding 1700-1900 transcription factors.

Regulatory elements (cis): regulatory modules

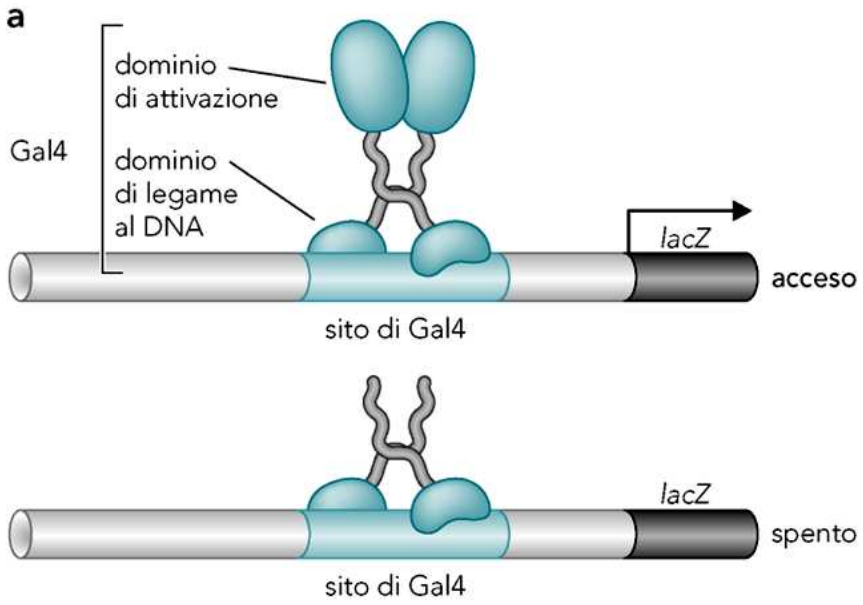
Regulatory factors: regulatory complexes (DNA-binding proteins + coactivators/corepressors



Transcription factors are composed of DNA-binding domains (DBA) plus trans-activating domains (TA)

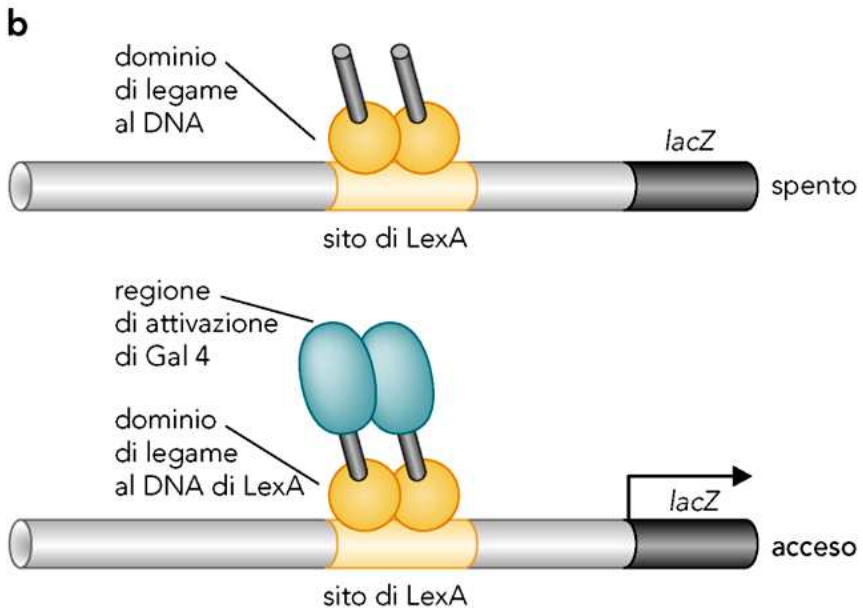
These domains are independent from each other

DBD is sequence-specific



The simple DNA binding per se does not influence transcription

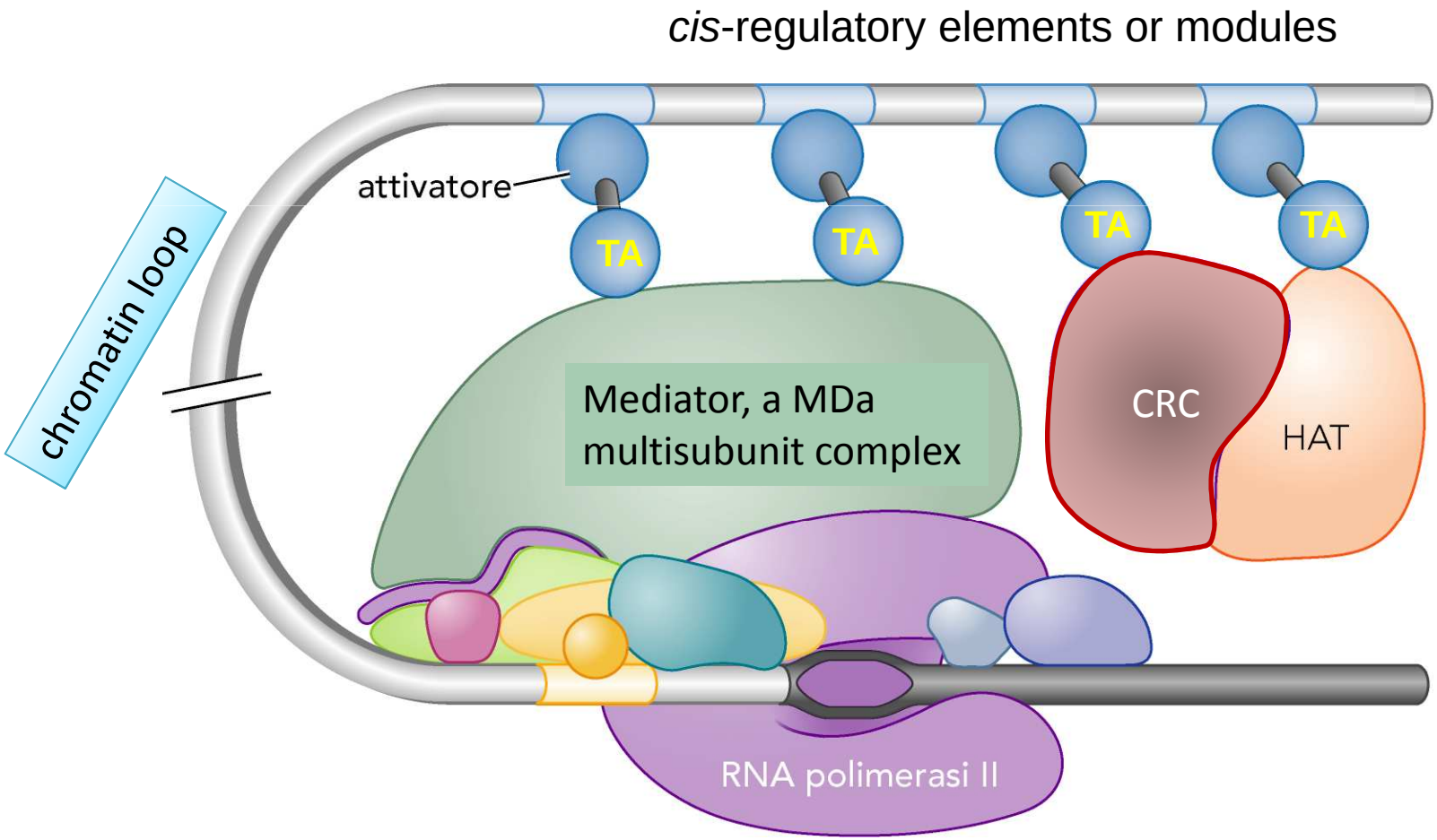
DBDs and TA act independently as demonstrated by domain-switching experiments

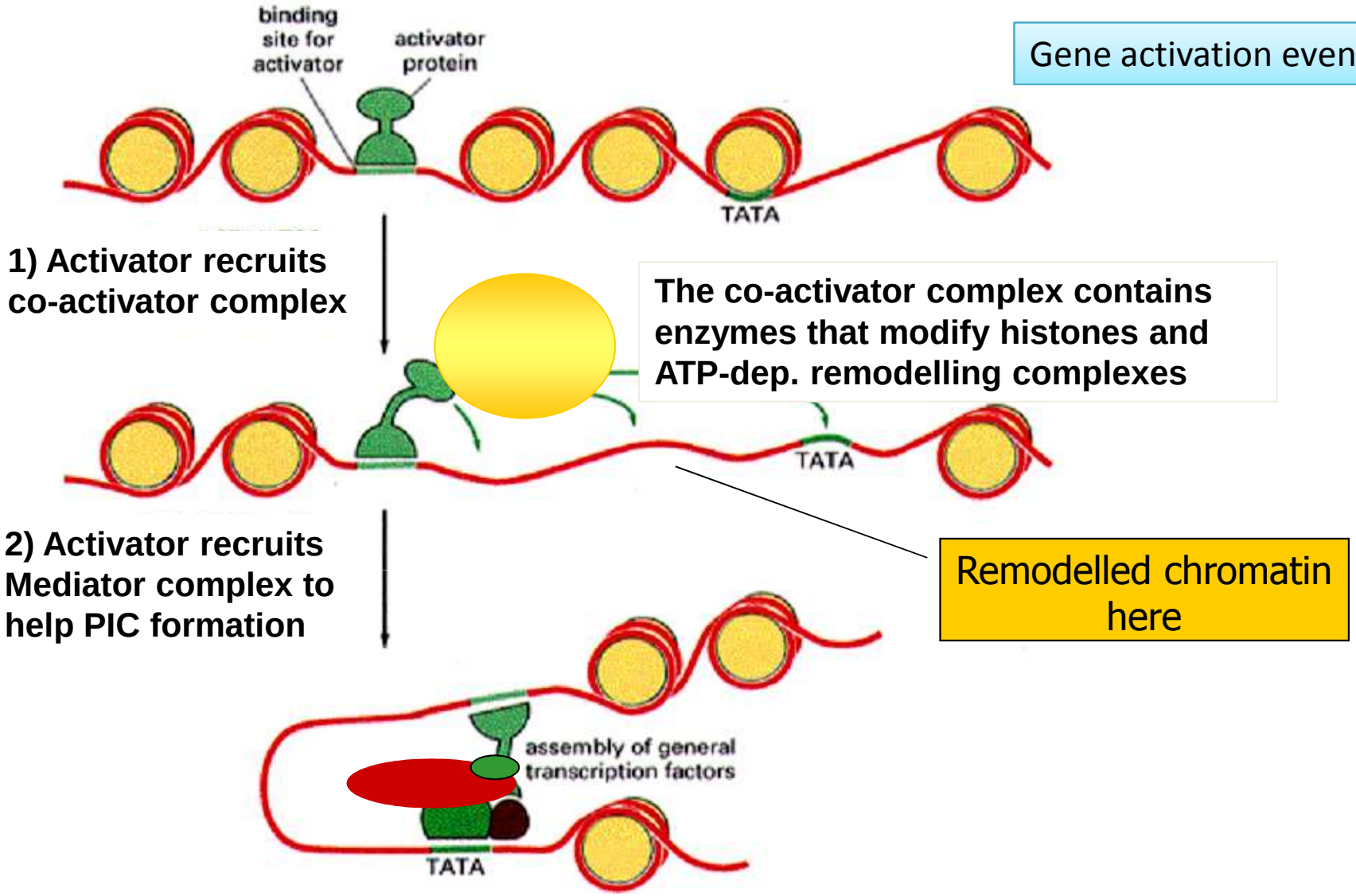


Transcription factors have multiple roles:

1. Protein-protein interaction with the basal complex (PIC) through Mediator
2. Recruitment of histone modifying enzymes
3. Recruitment of Chromatin Remodeling Complexes

Distant regulatory regions contact basal promoters by looping





Quindi, quando un gene deve essere attivato, la sequenza di eventi è:

il riconoscimento del gene da parte di uno o più fattori di trascrizione regolatori, seguito da associazione di enzimi che rimodellano la cromatina localmente (HAT+ATP-remodeling complexes), che permette il legame di ulteriori fattori regolatori, del Mediatore e del PIC, ed infine la trascrizione del gene

Let’s go back to the promoter story....

This kind of information ("**textbook promoter**") was obtained historically by studying a limited number of promoters with well-defined TSS, clear promoter activity and defined regulatory elements.

Several promoters, less defined and more difficult to study, were left apart, typical example the CpG – type promoters.

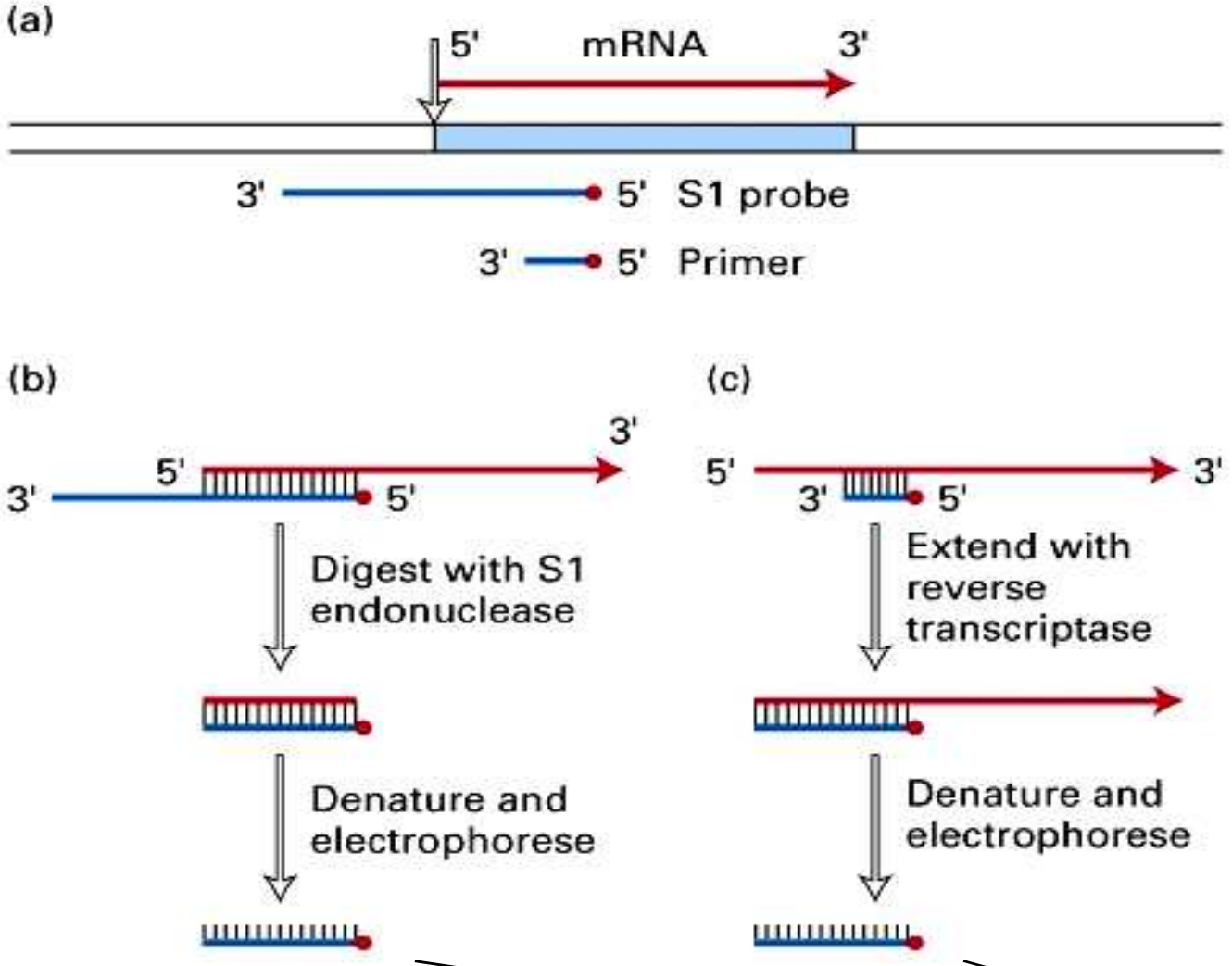
Genomic studies have partially changed our knowledge of promoters.

Studies oriented to define the **TSS** genome-wide, such as CAGE and 5' - SAGE, were especially instructive.

These studies demonstrated, first, that the "textbook promoter" is present at no more that 10-20% of mammalian genes (17% in human and mouse), which represent a group of inducible, tissue-specific genes.

Remaining transcription units have different structures, more often relying on CpG islands.

gene by gene



Primer extension is an alternative mapping method

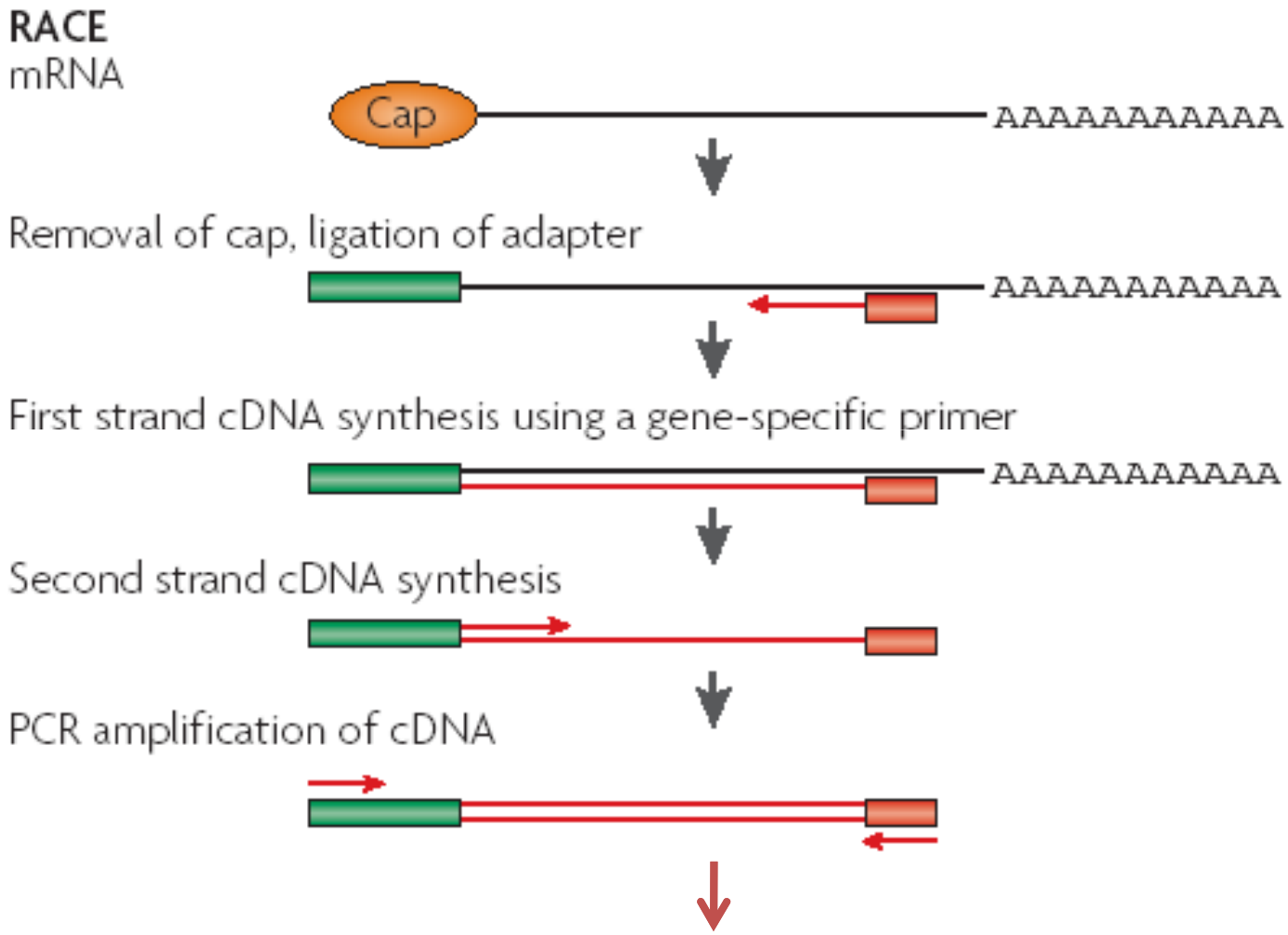
S1 nuclease mapping (RNase protection quite similar)

Primer extension

Single-stranded DNA is evaluated by polyacrylamide gel electrophoresis

gene by gene

RACE= rapid amplification of cDNA ends

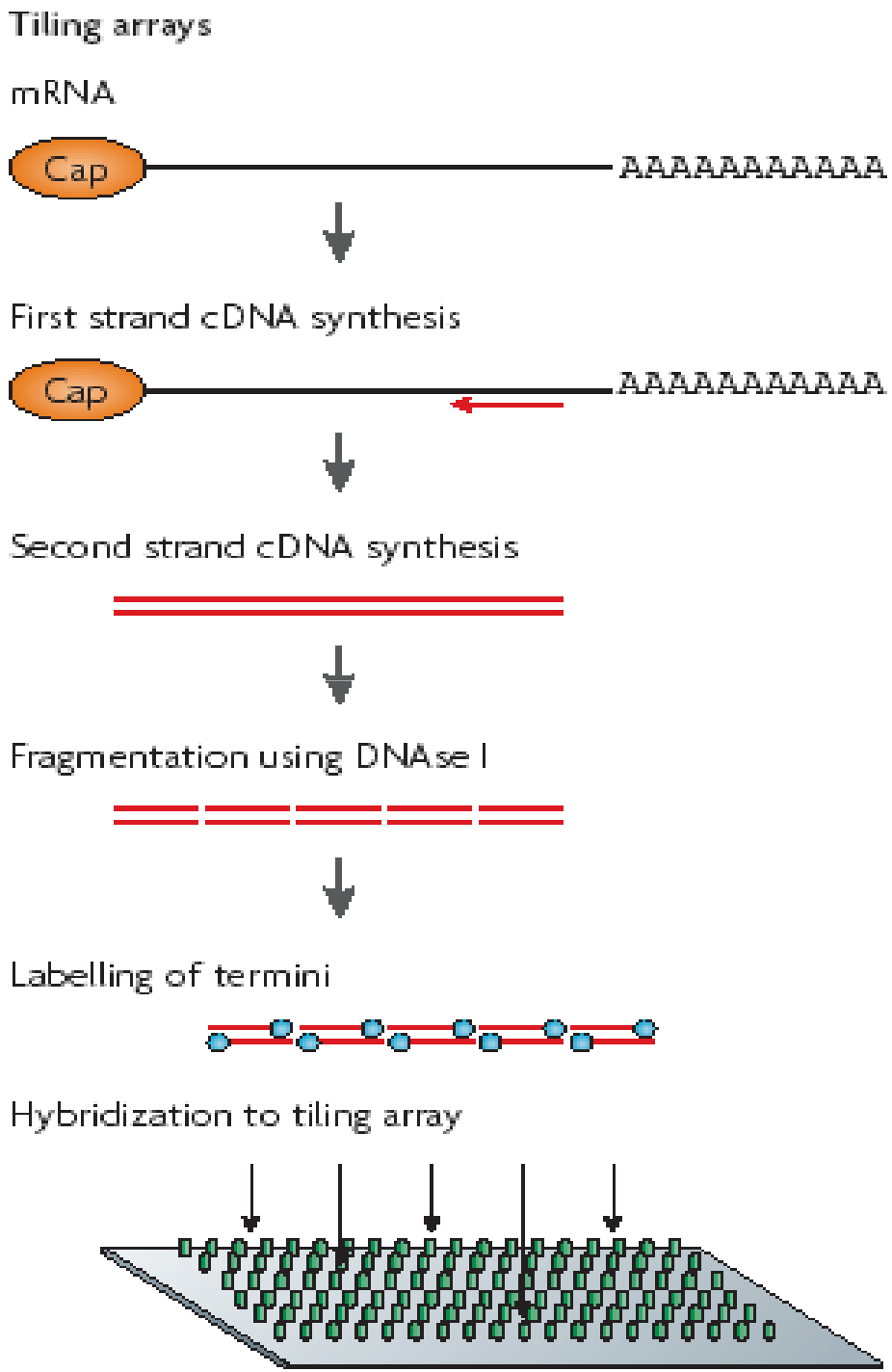


sequencing

genome-wide

essentially CAGE

But also tiling arrays have been used after selection for CAP



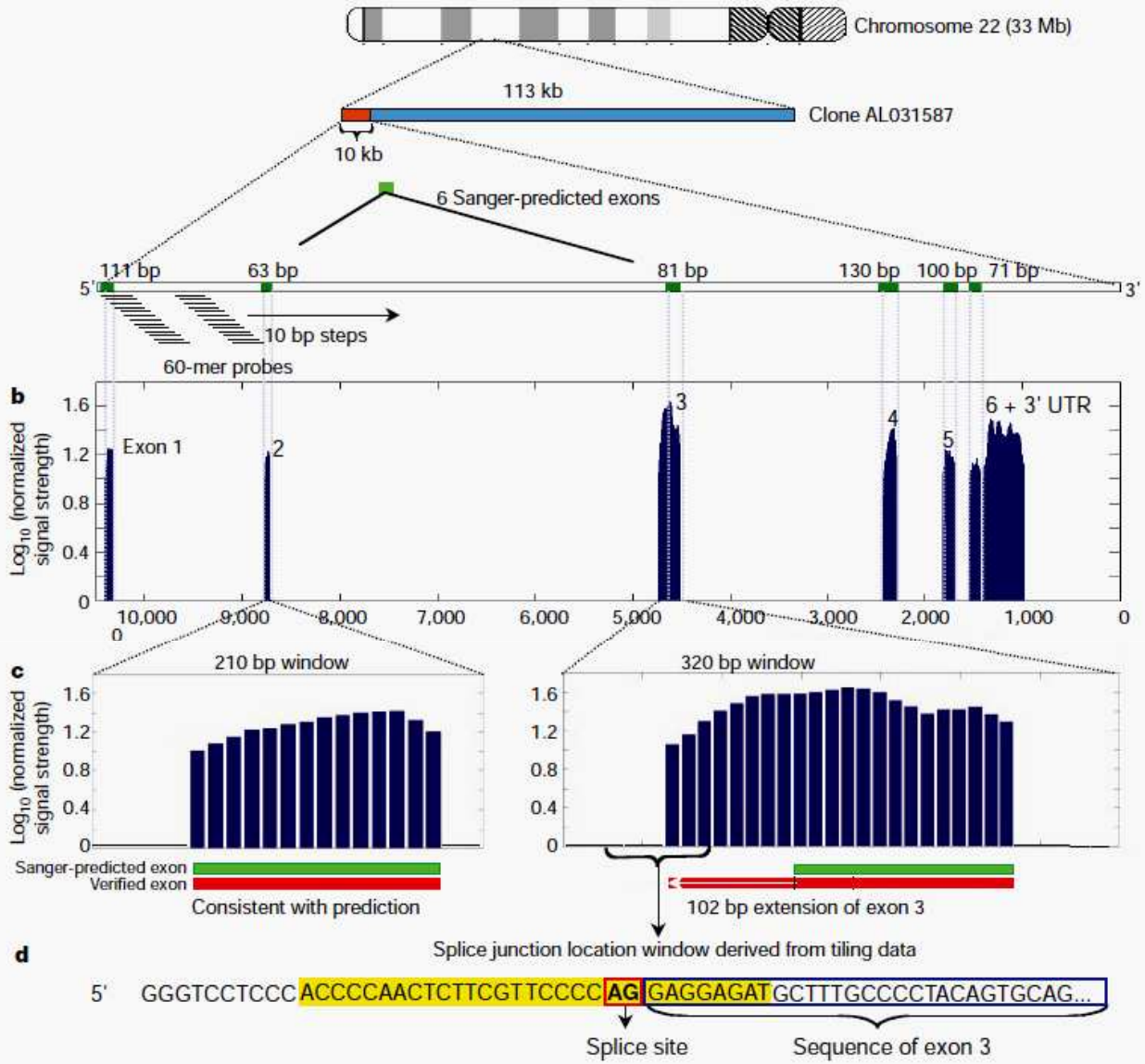


Figure 3 Characterization of a novel testis transcript using tiling arrays. **a**, An EVG discovered in the analysis of chromosome 22 (Fig. 2e) was localized to a 10-kb region at one end of the insert of BAC clone AL031587. Both strands of this 113-kb genomic interval were tiled with 60-mer probes at 10-bp steps. The tiling array was hybridized with RNA from human testis. **b**, Hybridization signals corresponding to tiling probes from this region were filtered and plotted as $\log_{10}$ values of the normalized signal strengths. Of the six Genscan predicted exons in this region, two (exons 3 and 6) were at variance with the hybridization data. **c**, Detailed views of tiling data showing one correctly predicted exon and one incorrectly predicted exon. **d**, Typically, tiling data narrow the search window for an intron/exon boundary to 20–30-bp. The exact splice junction is then identified using consensus sequences (GT-AG rule) and ORF information. The exact splice junction can also be determined by sequencing RT–PCR products.

Direct isolation and identification of promoters in the human genome

Tae Hoon Kim,[1] Leah O. Barrera,[1] Chunxu Qu,[1] Sara Van Calcar,[1]
Nathan D. Trinklein,[4] Sara J. Cooper,[4] Rosa M. Luna,[2] Christopher K. Glass,[2]
Michael G. Rosenfeld,[3] Richard M. Myers,[4] and Bing Ren[1,2,5]

[1]Ludwig Institute for Cancer Research, [2]Department of Cellular and Molecular Medicine, and [3]Howard Hughes Medical Institu University of California, San Diego, La Jolla, California 92093, USA; [4]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

Transcriptional regulatory elements play essential roles in gene expression during animal development and cellular response to environmental signals, but our knowledge of these regions in the human genome is limited despite the availability of the complete genome sequence. Promoters mark the start of every transcript and are an important class of regulatory elements. A large, complex protein structure known as the pre-initiation complex (PIC) is assembled on all active promoters, and the presence of these proteins distinguishes promoters from other sequences in the genome. Using components of the PIC as tags, we isolated promoters directly from human cells as protein–DNA complexes and identified the resulting DNA sequences using genomic tiling microarrays. Our experiments in four human cell lines uncovered 252 PIC-binding sites in 44 semirandomly selected human genomic regions comprising 1% (30 megabase pairs) of the human genome. Nearly 72% of the identified fragments overlap or immediately flank 5′ ends of known cDNA sequences, while the remainder is found in other genomic regions that likely harbor putative promoters of unannotated transcripts. Indeed, molecular analysis of the RNA isolated from one cell line uncovered transcripts initiated from over half of the putative promoter fragments, and transient transfection assays revealed promoter activity for a significant proportion of fragments when they were fused to a luciferase reporter gene. These results demonstrate the specificity of a genome-wide analysis method for mapping transcriptional regulatory elements and also indicate that a small, yet significant number of human genes remains to be discovered.

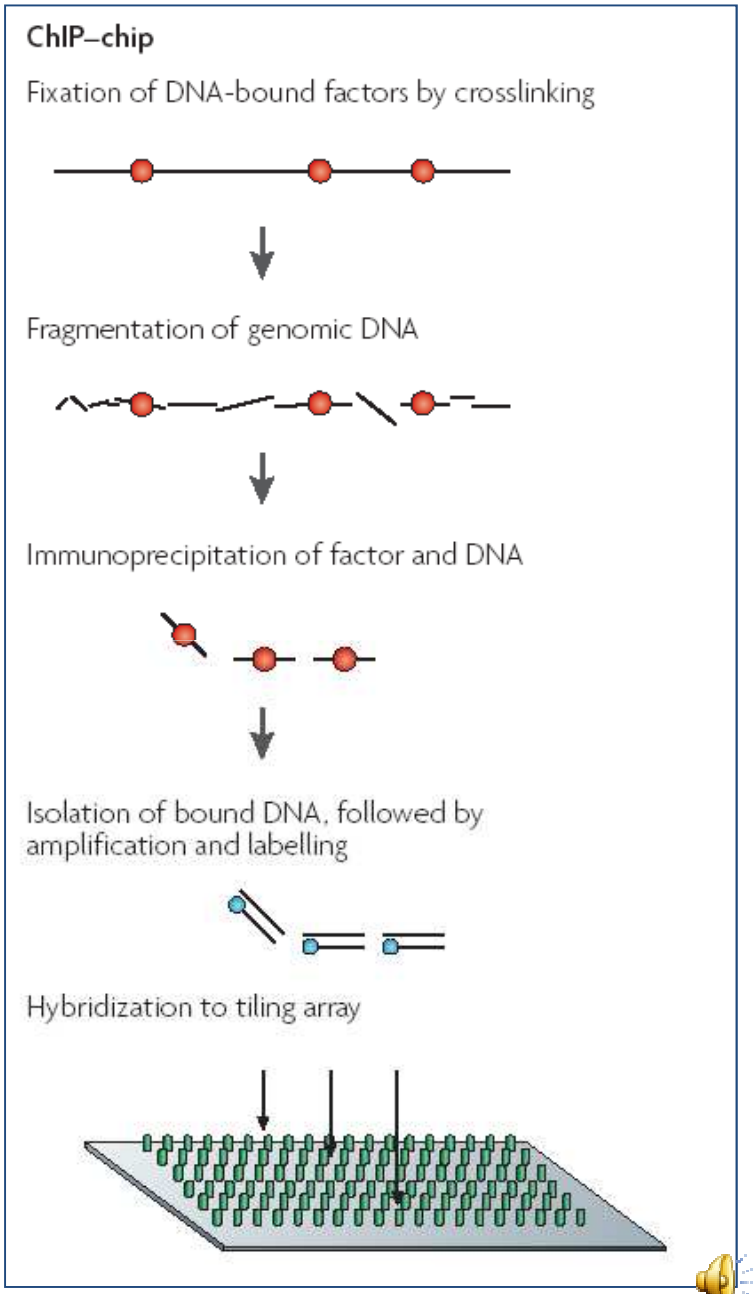
[Supplemental material is available online at www.genome.org.]

15:830–839 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

Other methods to identify genome-wide promoters have been used.

One strategy is to identify all the sequences that are bound by RNA Polymerase II and by basal transcription factors, such as TAFs, by using ChIP-on-chip.

As a part of the ENCODE project, for example, 1%-human genome coverage tiling arrays were used to identify sequences bound by RNA PolII and TAF1 (see TAF nomenclature).



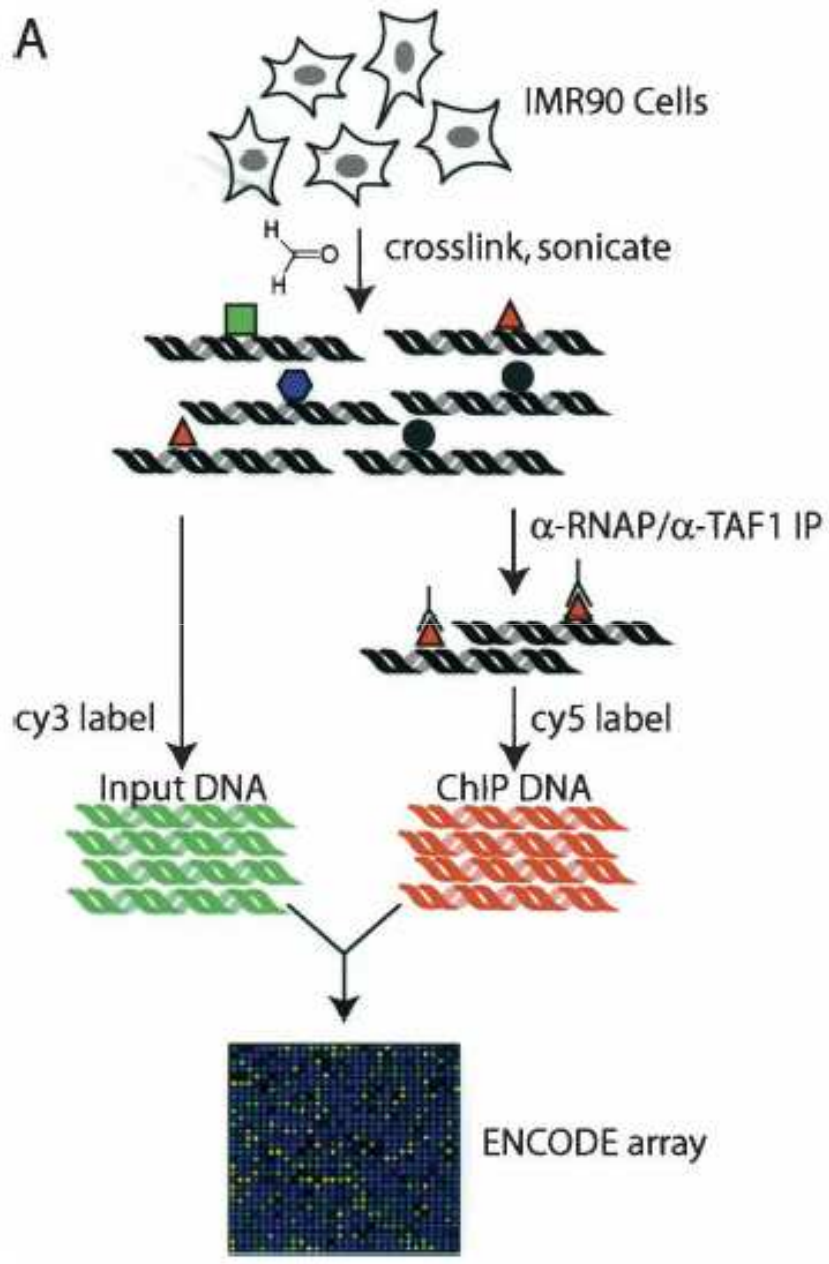
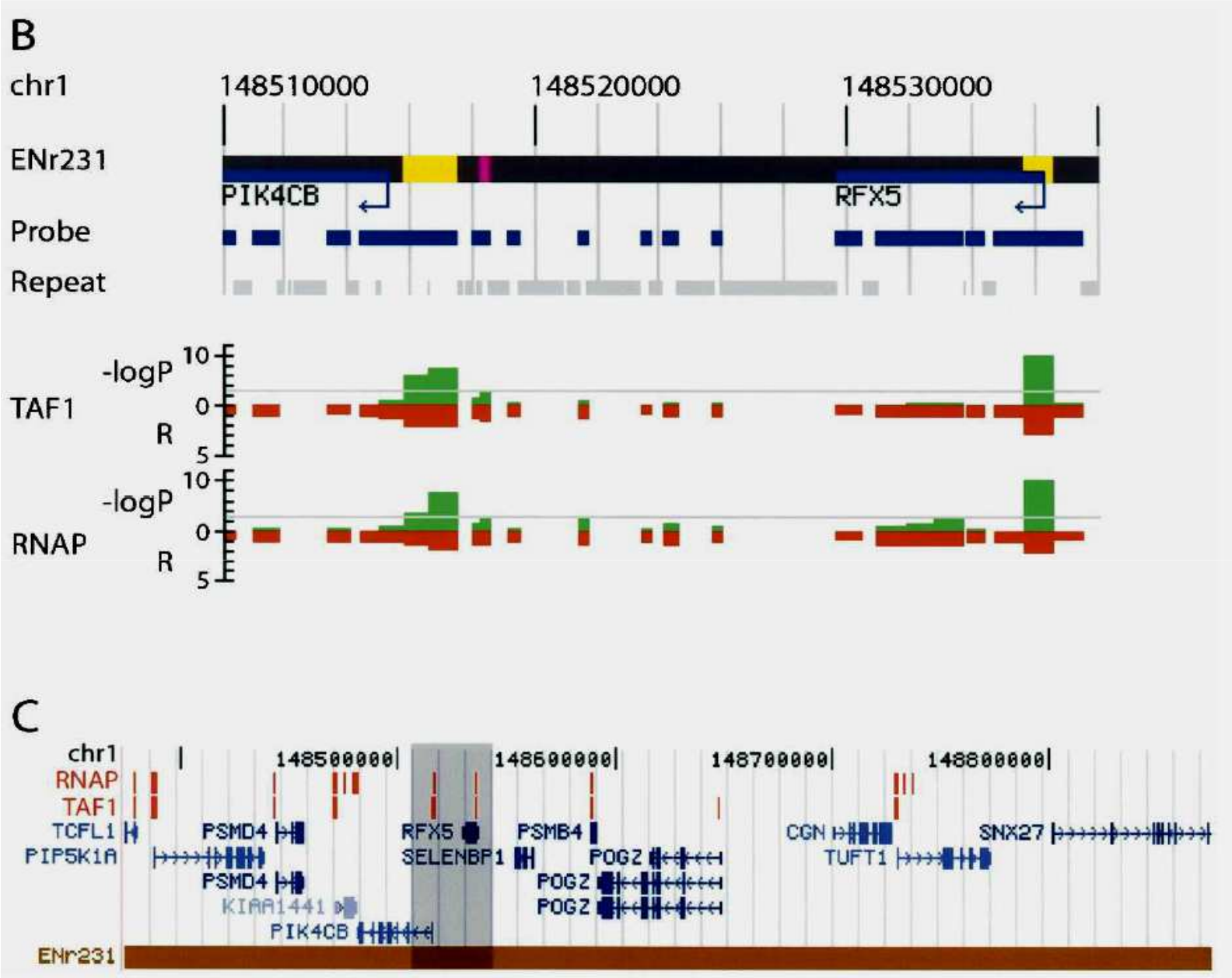


Figure 1. Direct isolation and identification of promoters in the human genome.
(*A*) A schematic of GWLA for mapping RNAP and TAF1-binding sites.
Growing cells are cross-linked with formaldehyde, and their nuclei are isolated , the chromatin extracted and sonicated.
The resulting chromatin (protein–DNA) complexes are incubated with either anti-RNAP (-RNAP) or anti-TAF1 (-TAF1) antibody. The immunoprecipitated DNA is subjected to ligation-mediated PCR, labeled with Cy5 dye, and competitively hybridized to the ENCODE array with the Cy3-labeled unenriched chromatin
(Li et al. 2003).



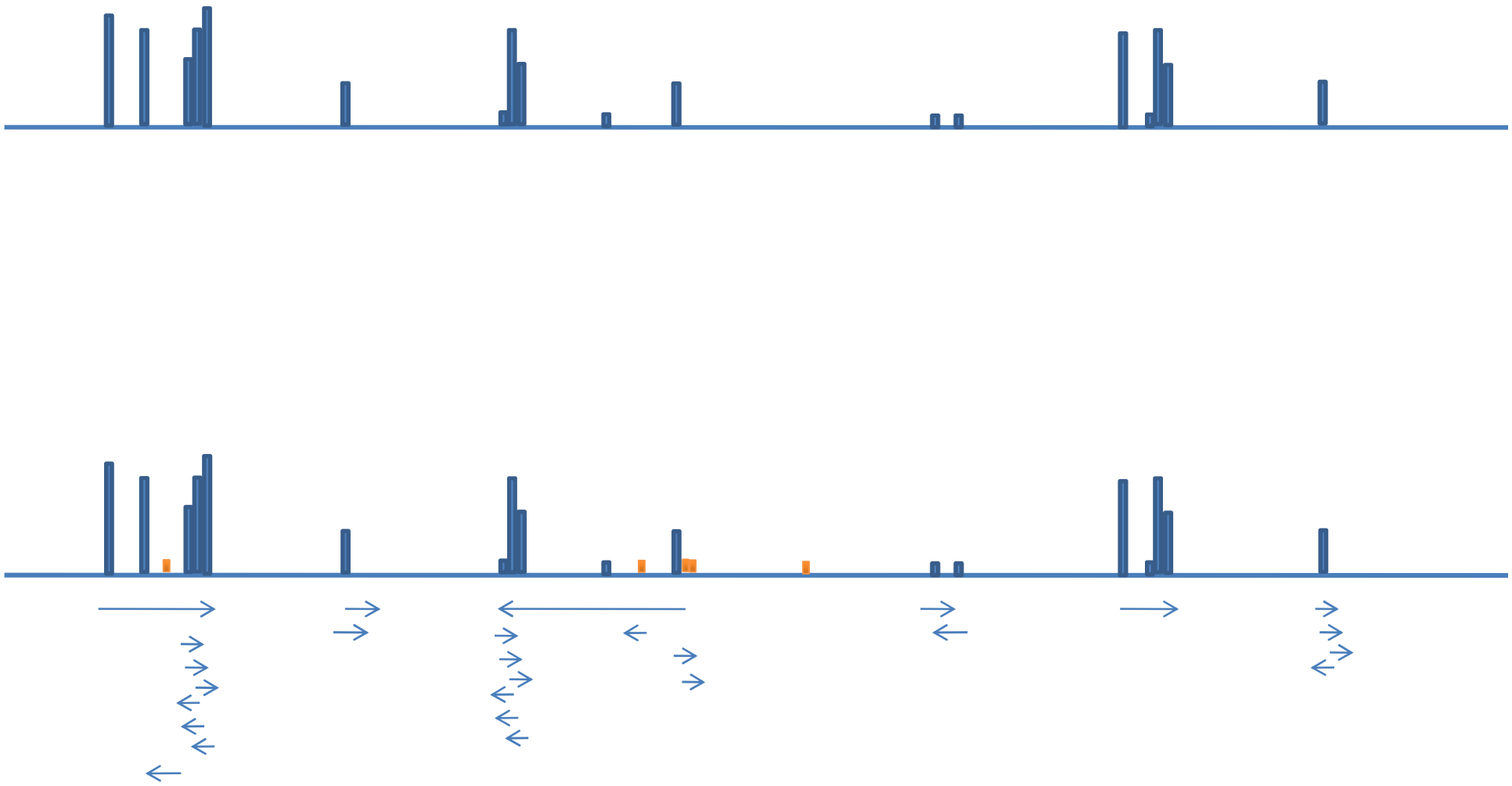
(*B*) A typical detailed view of the TAF1- and RNAP-binding data (within the ENCODE region, ENr231). Negative logarithmic *P* values of enrichment by RNAP or TAF1 ChIP for each probe are plotted in green. Relative enrichment (R) values by RNAP or TAF1 ChIP for each probe fragment are plotted in red on the inverted axis. (*C*) A representative view of an entire ENCODE locus (ENr231) with annotated RNAP- and TAF1-binding sites that have *P* values <0.0001, noted in red blocks (detailed view in *B* is highlighted in gray).

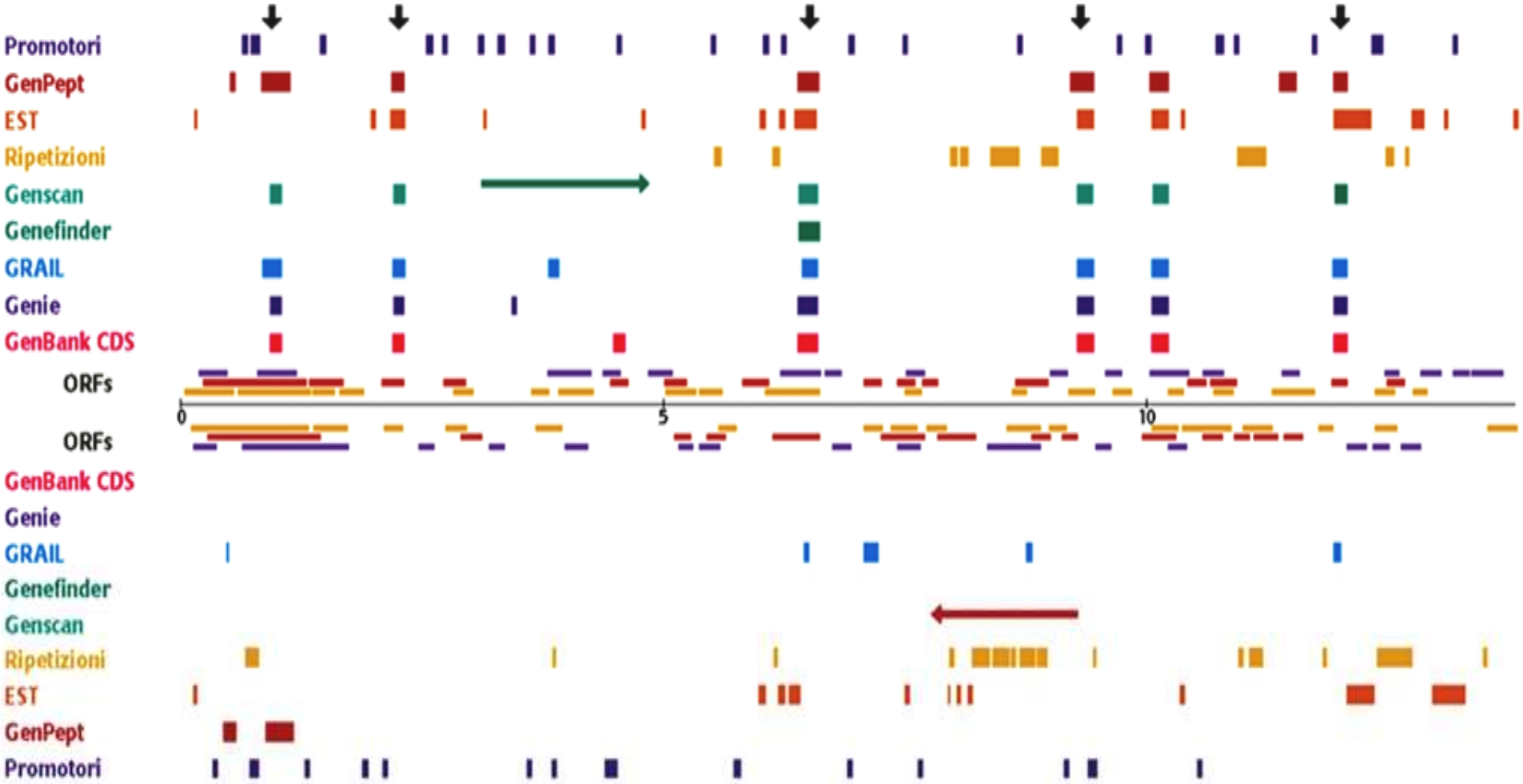
This is a modern way to show information on a genome scheme, and is used today by most genomic browsers

Genome annotation taken by multiple studies, is aligned to genome maps, thus giving a representational picture of “what is known” on each section of the genome in a continuous fashion.

Genome annotation

http://genome.ucsc.edu/cgi-bin/hgGateway





Annotazione bioinformatica di un segmento di 15 kb del genoma umano, contenente un gene, utilizzando il browser Genotator.

Tutto il genoma viene continuamente ri-annotato considerando le nuove evidenze sperimentali, se liberamente disponibili.

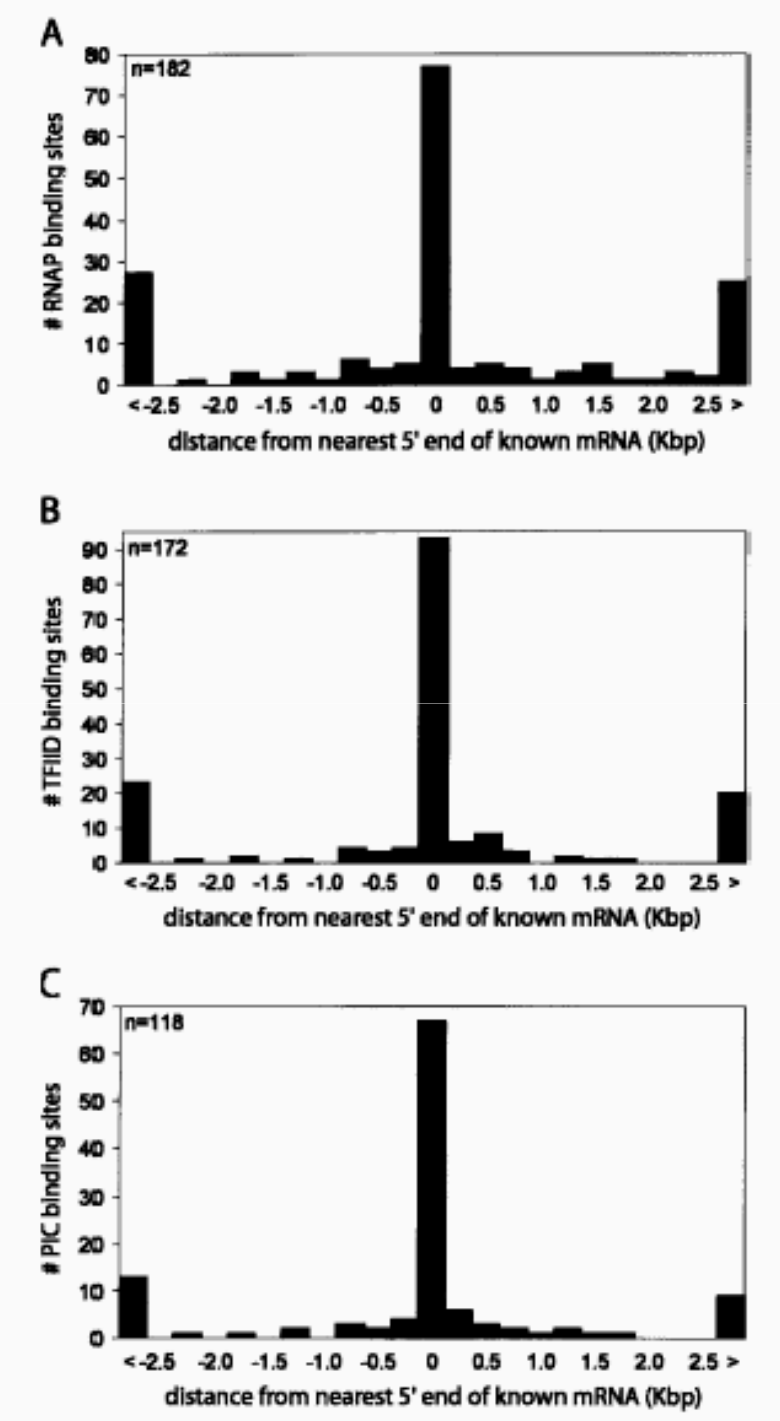


Figure 2. Summary of identified PIC-binding sites matched to transcripts

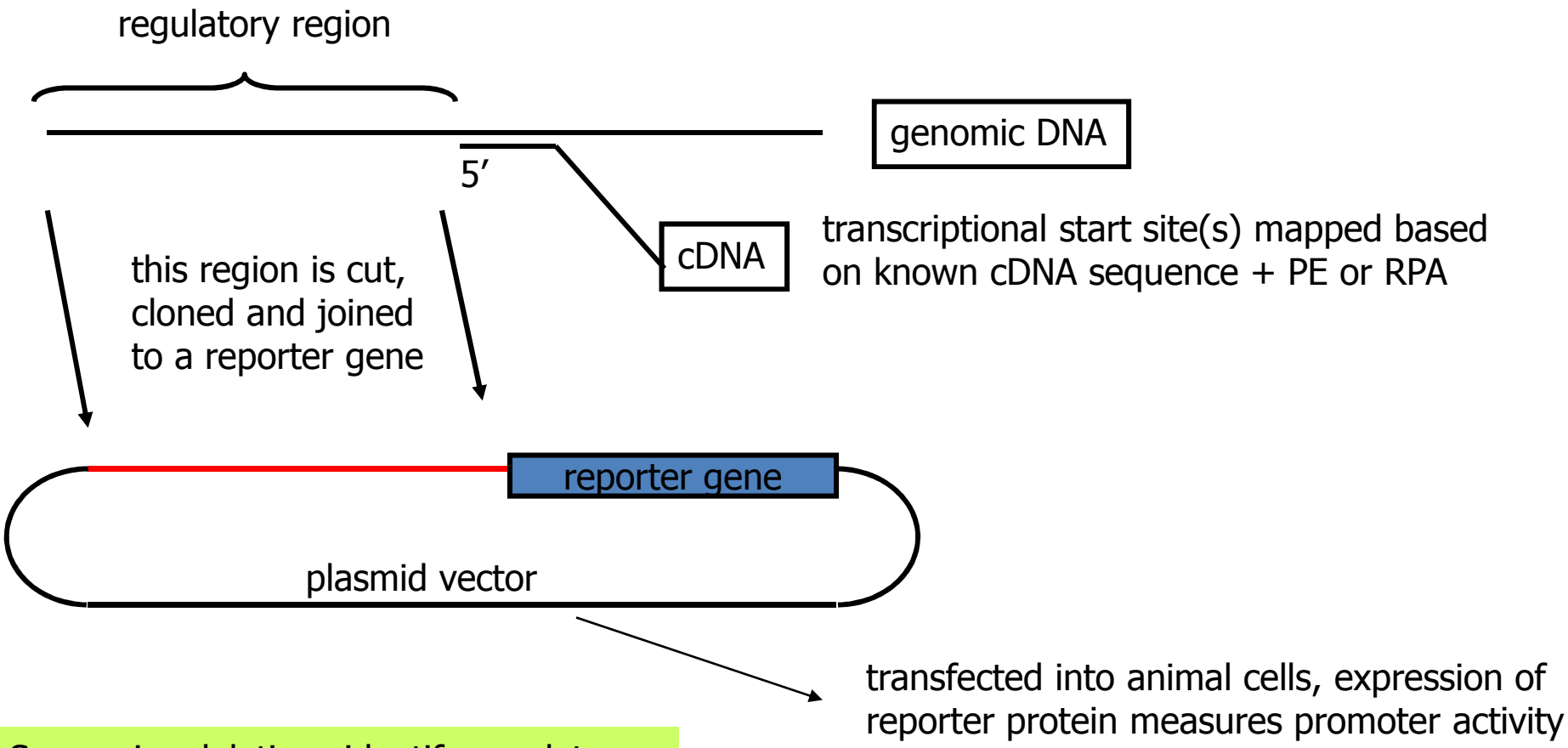
in IMR90. Histograms plotting relative distance of the RNAP-, TFIID-, or PIC-binding sites to the nearest 5 ends of full-length GenBank RefSeq transcripts. (*A) Distribution (number of RNAP-binding sites, y-axis)* of relative distances (in Kbp, *x-axis) of the RNAP-binding sites to the* nearest 5 ends of full length mRNA. The first and last bars are counts for those RNAP-binding sites that are >2.5 Kbp upstream (<) or downstream (>) from the nearest 5 end. (*B) Distribution (number of TFIID-binding* sites, *y-axis) of relative distances (in Kbp, x-axis) of the TFIID-binding sites* to the nearest 5 ends of full-length mRNA. The first and last bars are counts for those TFIID-binding sites that are >2.5 Kbp upstream (<) or downstream (>) from the nearest 5 end. (*C) Distribution (number of* PIC-binding sites, *y-axis) of relative distances (in Kbp, x-axis) of the PICbinding* sites to the nearest 5 ends of full-length mRNA. The first and last bars are counts for those PIC-binding sites that are >2.5 Kbp upstream (<) or downstream (>) from the nearest 5 end.

When a sequence is identified as a putative promoter, the functional validation consists in the classical **reporter assay**

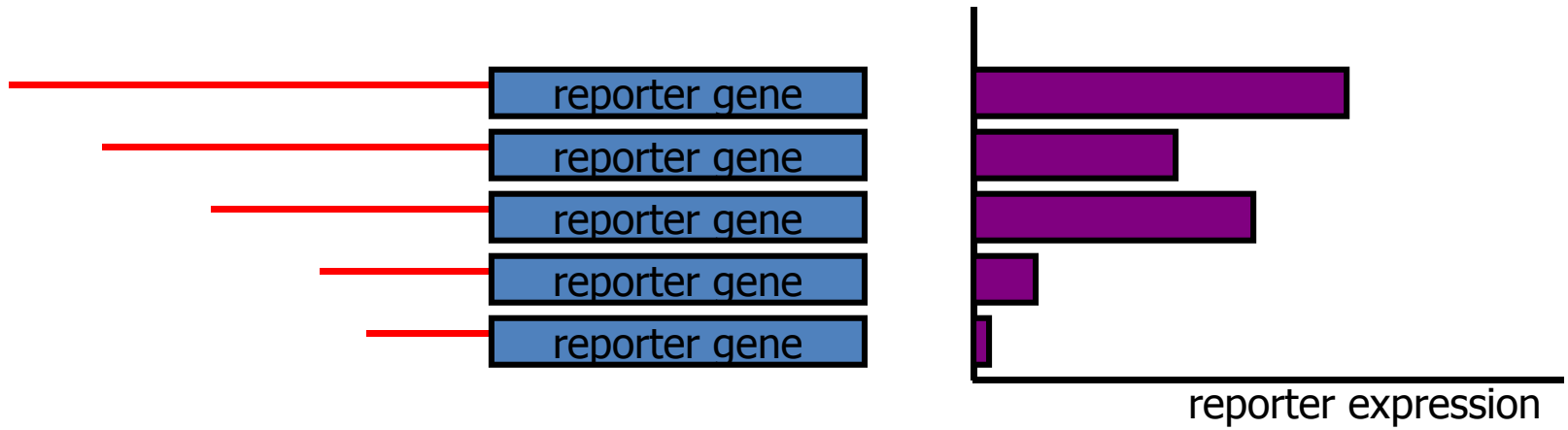
It this assay, a minigene is constructed fusing the putative promoter sequence to a **reporter** gene, i.e. a gene whose product can be easily measured.

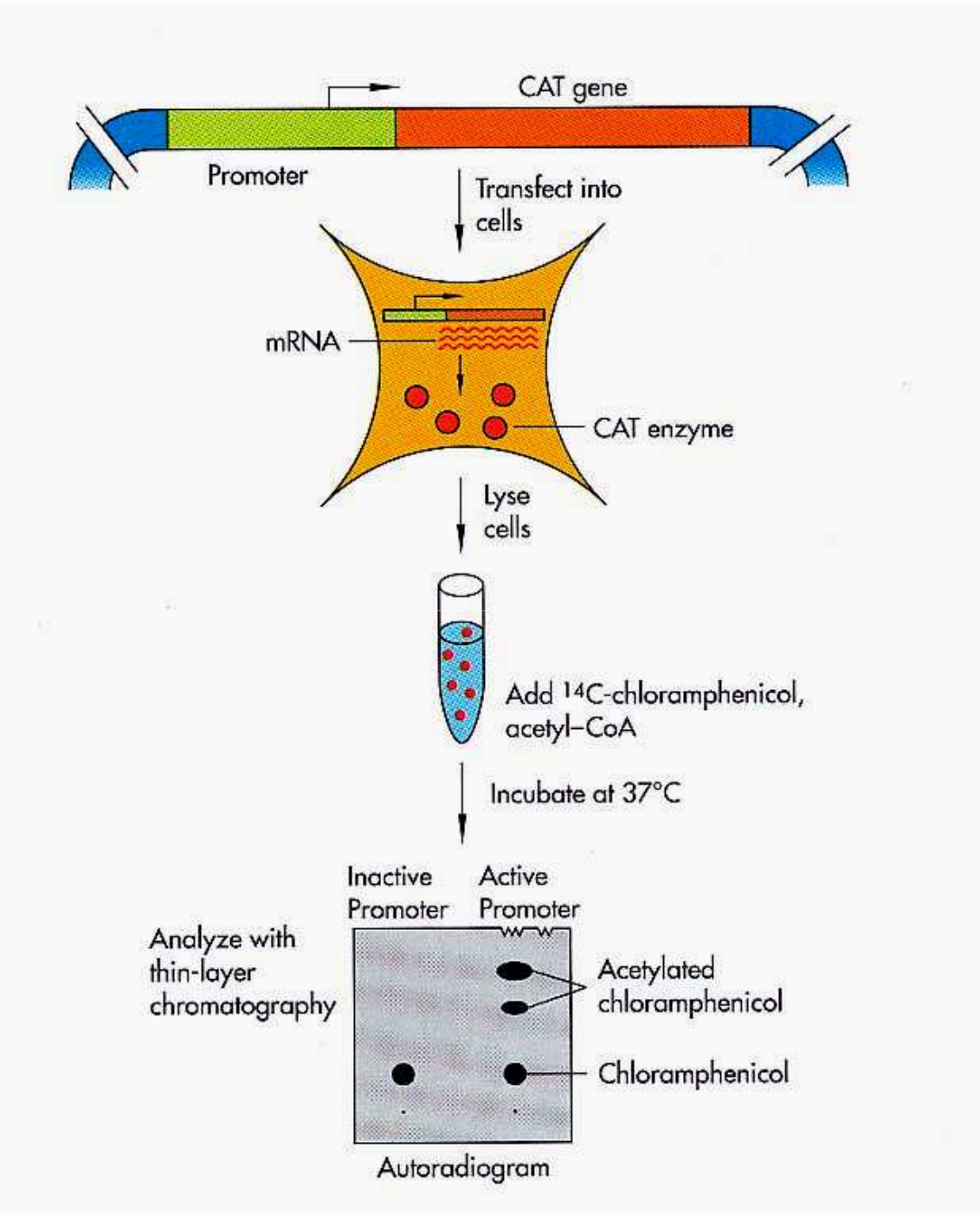
This construct is then transfected into cultured cells and the product measured after a period of time necessary for transgnene import and expression.

Mammalian **reporter genes** are usually nonmammalian genes, such as
- CAT cloramphenicol acetyltransferase
- Luc firefly luciferase
- GFP Jellyfish green fluorescent protein
- β-gal beta-galactosidase



Successive deletions identify regulatory elements and minimal promoter





The reporter gene assay for promoter and enhancers

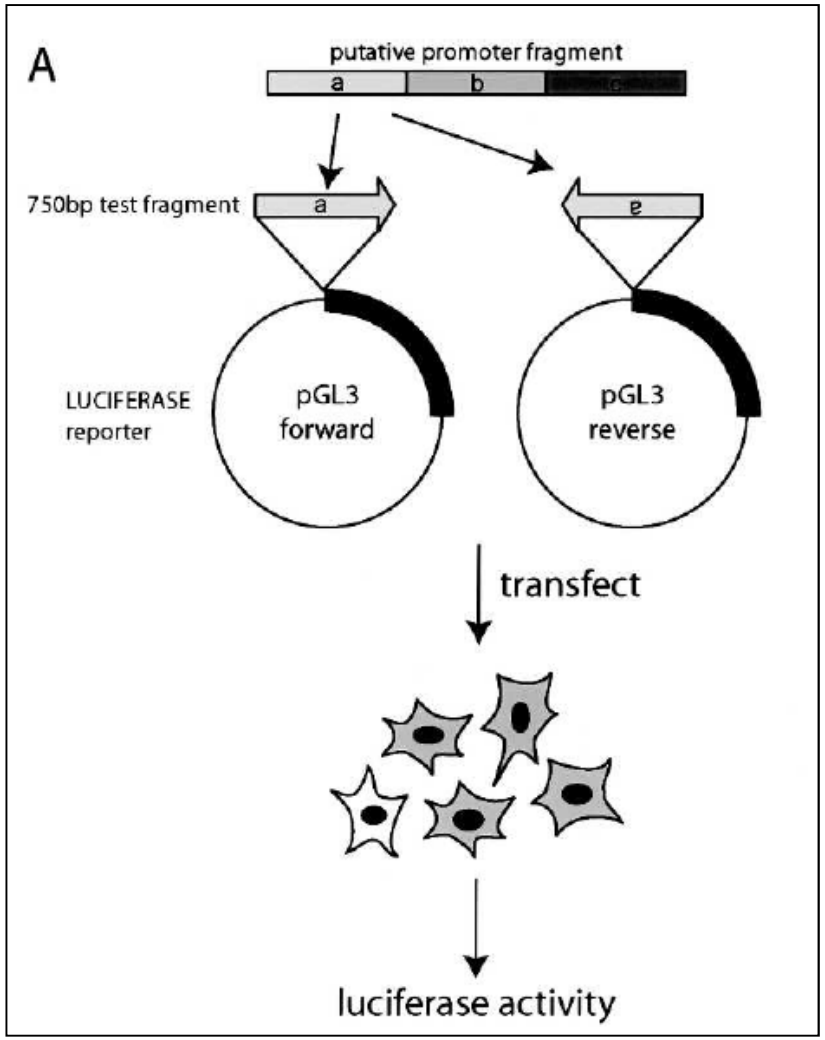


Figure 3. Experimental validation of the putative promoters by reporter assays.
(*A*) A schematic of reporter assay used to determine whether the identified putative promoter fragment can support transcription. Each putative promoter fragment was segmented into 750-bp fragments by PCR and cloned into the luciferase reporter construct, pGL3, in either forward or reverse orientations. The resulting reporter constructs were individually transfected into HT1080 cells and the resulting luciferase activity was measured.

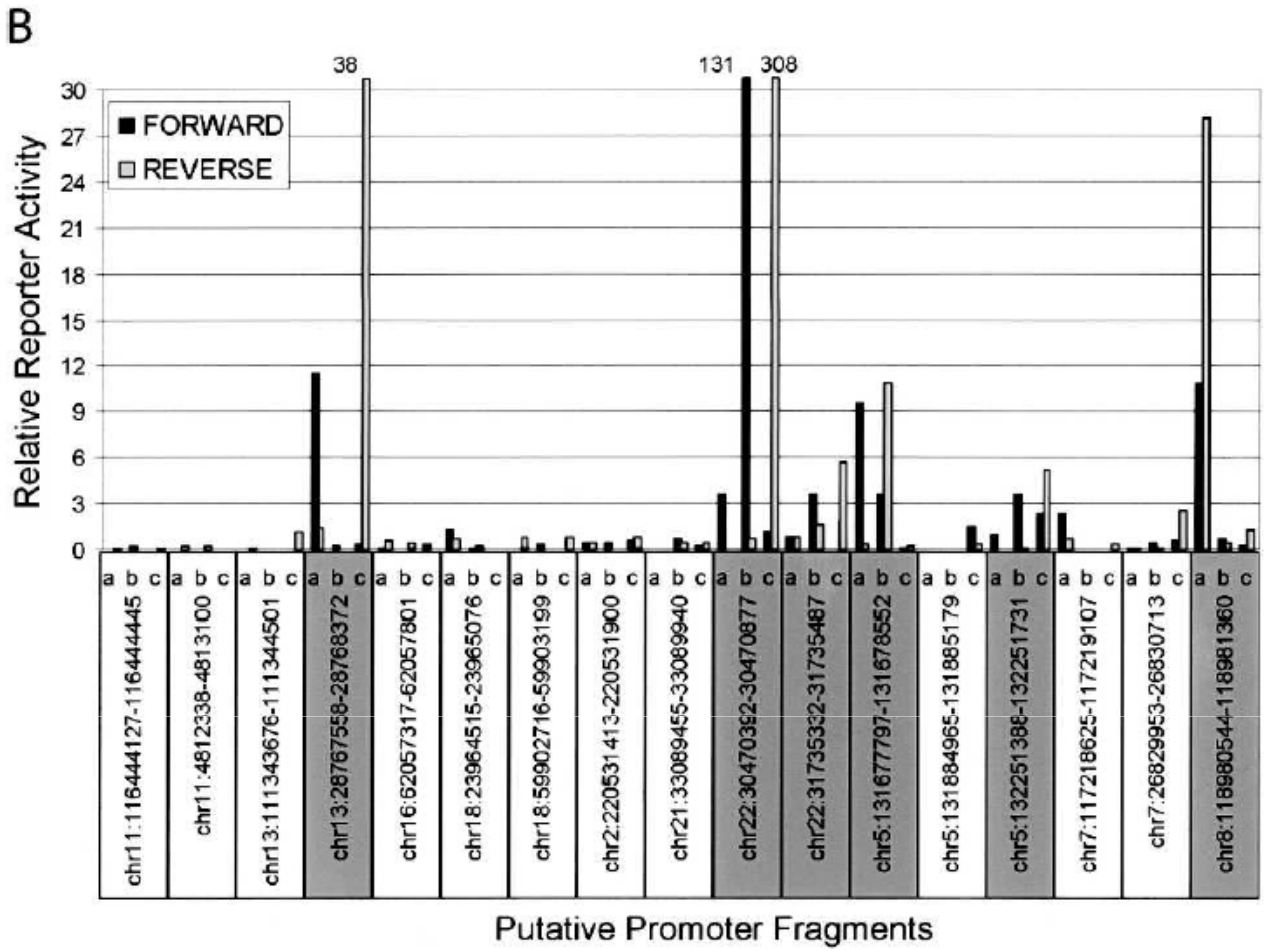
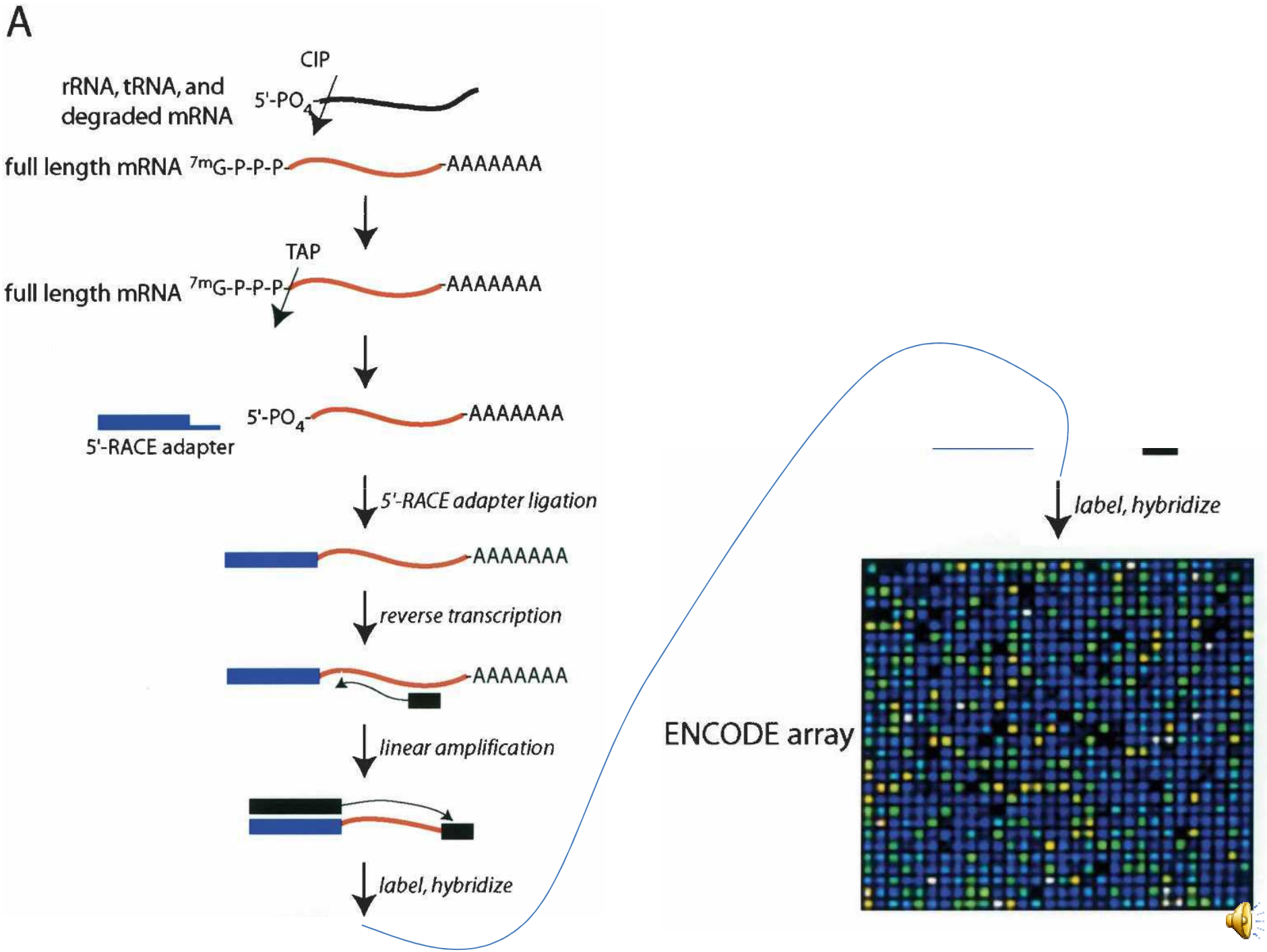
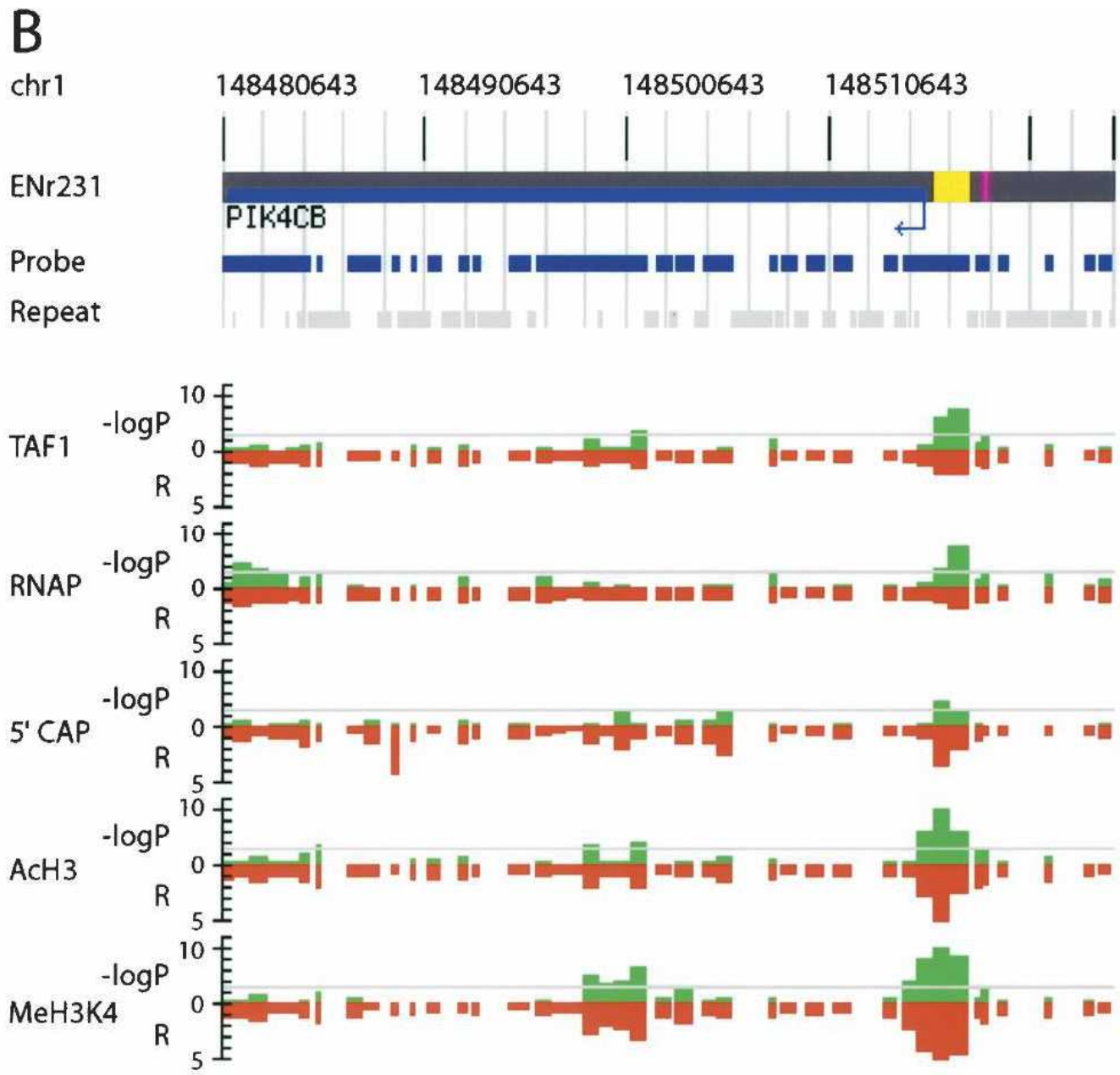
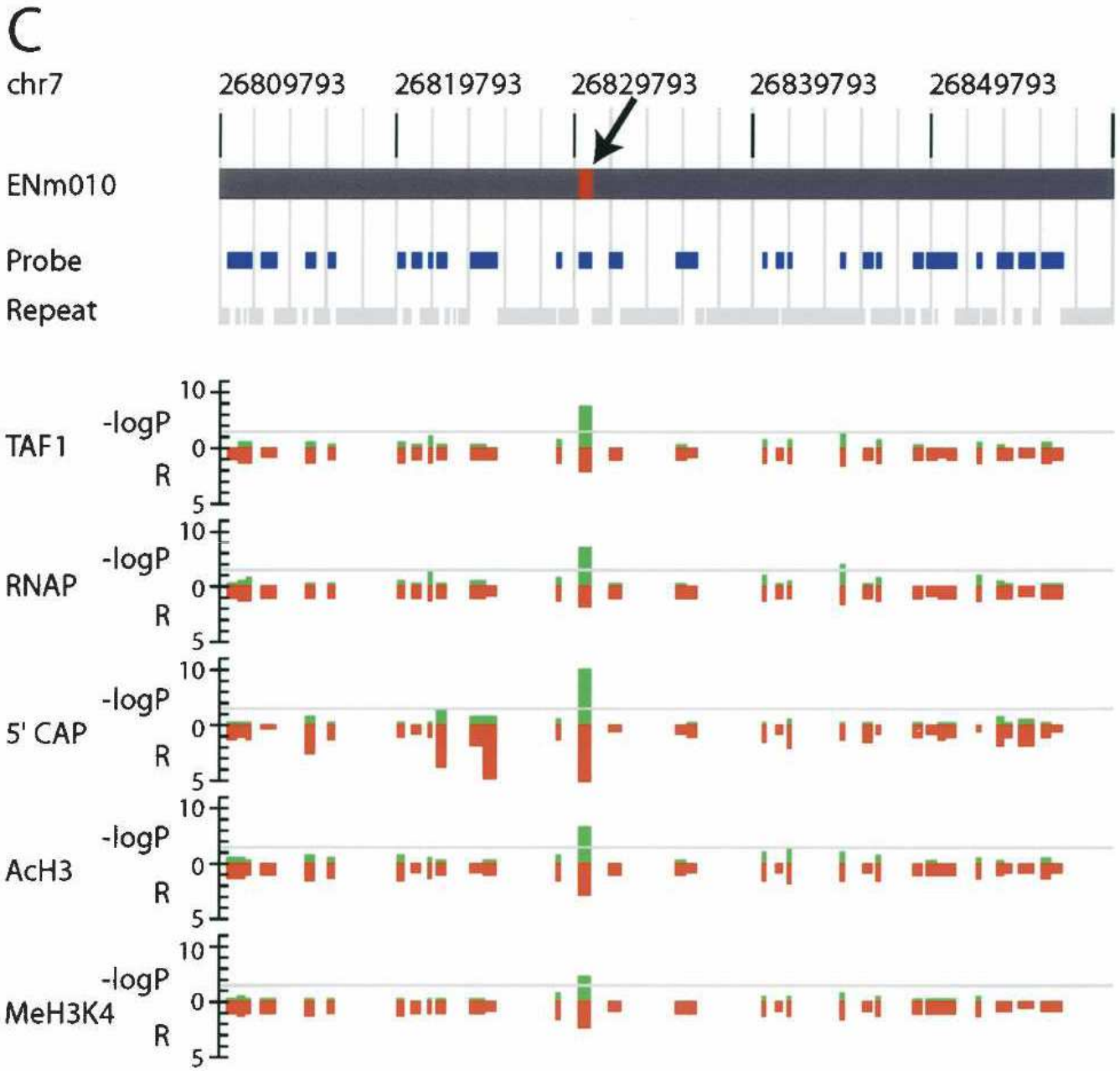


Figure 3. Experimental validation of the putative promoters by reporter assays.
(*B*) Reporter activities of 17 putative promoter fragments are shown (remaining five of 22 were not tested). Relative reporter activity was determined by comparing the luciferase activity of the test fragment and the control genomic DNA fragments. Promoter fragments with significant reporter activity (exceeding three times the standard deviation of all control fragments) are highlighted in gray.


A
CIP
rRNA, tRNA, and degraded mRNA
5'-PO4
full length mRNA 7mG-P-P-P-
AAAAAAA
TAP
full length mRNA 7mG-P-P-P-
AAAAAAA
5'-PO4
AAAAAAA
5'-RACE adapter
5'-RACE adapter ligation
AAAAAAA
reverse transcription
AAAAAAA
linear amplification
label, hybridize
label, hybridize
ENCODE array


B
chr1
148480643
148490643
148500643
148510643
ENr231
PIK4CB
Probe
Repeat
TAF1
RNAP
5' CAP
AcH3
MeH3K4
-logP
R
10
0
5


C
chr7
26809793
26819793
26829793
26839793
26849793
ENm010
Probe
Repeat
TAF1
RNAP
5' CAP
AcH3
MeH3K4
-logP
R
10
0
5

However, the most impressive data came from the FANTOM project carried out using CAGEX

The most extensive core promoter identification study undertaken so far used CAGE tags to identify 184,379 human and 177,349 mouse core promoters, many of which might contain a cluster of individual TSSs[24]. A previous analysis that involved full-length cDNA sequencing identified 30,964 human and 19,023 mouse promoters[42]. But even the most recent figures are likely to be a substantial underestimate. First, sequencing 50–100,000 tags in each library can reliably detect only those transcripts that are expressed at a level of at least 10 copies in each cell (as there are at least 400,000 mRNAs in an average mammalian cell[43]). Many transcripts are not present at this level, either because they are of low abundance in individual cells or are expressed in only a small subset of cells in the tissues that have been studied.

Genome-wide analysis of mammalian promoter architecture and evolution

Piero Carninci[1,2,21], Albin Sandelin[1,3,21], Boris Lenhard[1,3,20,21], Shintaro Katayama[1], Kazuro Shimokawa[1], Jasmina Ponjavic[1,20], Colin A M Semple[1,4], Martin S Taylor[1,5], Pär G Engström[3], Martin C Frith[1,6], Alistair R R Forrest[6], Wynand B Alkema[3], Sin Lam Tan[7], Charles Plessy[2], Rimantas Kodzius[1,2], Timothy Ravasi[1,6,8], Takeya Kasukawa[1,9], Shiro Fukuda[1], Mutsumi Kanamori-Katayama[1], Yayoi Kitazume[1], Hideya Kawaji[1,9], Chikatoshi Kai[1], Mari Nakamura[1], Hideaki Konno[1], Kenji Nakano[1,9], Salim Mottagui-Tabar[3,20], Peter Arner[10], Alessandra Chesi[11], Stefano Gustincich[11], Francesca Persichetti[12], Harukazu Suzuki[1], Sean M Grimmond[6], Christine A Wells[19], Valerio Orlando[13], Claes Wahlestedt[3,20], Edison T Liu[14], Matthias Harbers[15], Jun Kawai[1,2], Vladimir B Bajic[1,7,16], David A Hume[1,6,21] & Yoshihide Hayashizaki[1,2,17,18]

Mammalian promoters can be separated into two classes, conserved TATA box–enriched promoters, which initiate at a well-defined site, and more plastic, broad and evolvable CpG-rich promoters. We have sequenced tags corresponding to several hundred thousand transcription start sites (TSSs) in the mouse and human genomes, allowing precise analysis of the sequence architecture and evolution of distinct promoter classes. Different tissues and families of genes differentially use distinct types of promoters. Our tagging methods allow quantitative analysis of promoter usage in different tissues and show that differentially regulated alternative TSSs are a common feature in protein-coding genes and commonly generate alternative N termini. Among the TSSs, we identified new start sites associated with the majority of exons and with 3′ UTRs. These data permit genome-scale identification of tissue-specific promoters and analysis of the *cis*-acting elements associated with them.

146 mouse cDNA libraries

41 human cDNA libraries

!

By CAGE analysis (see previous lectures)

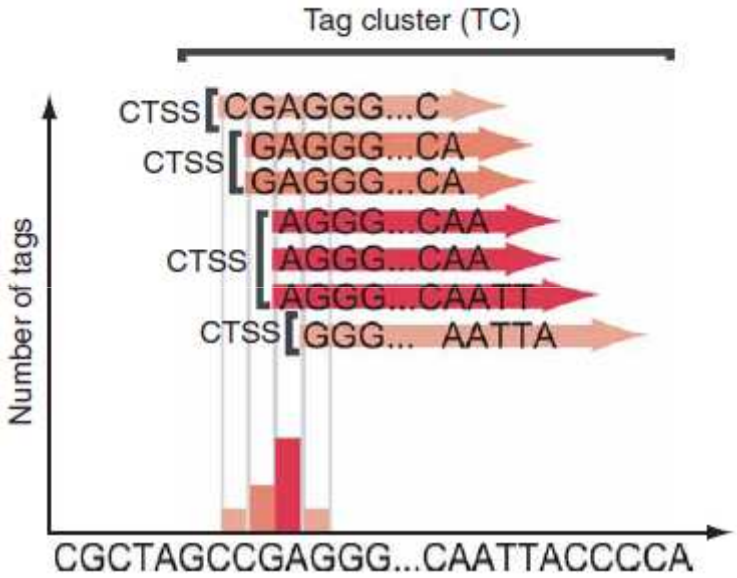
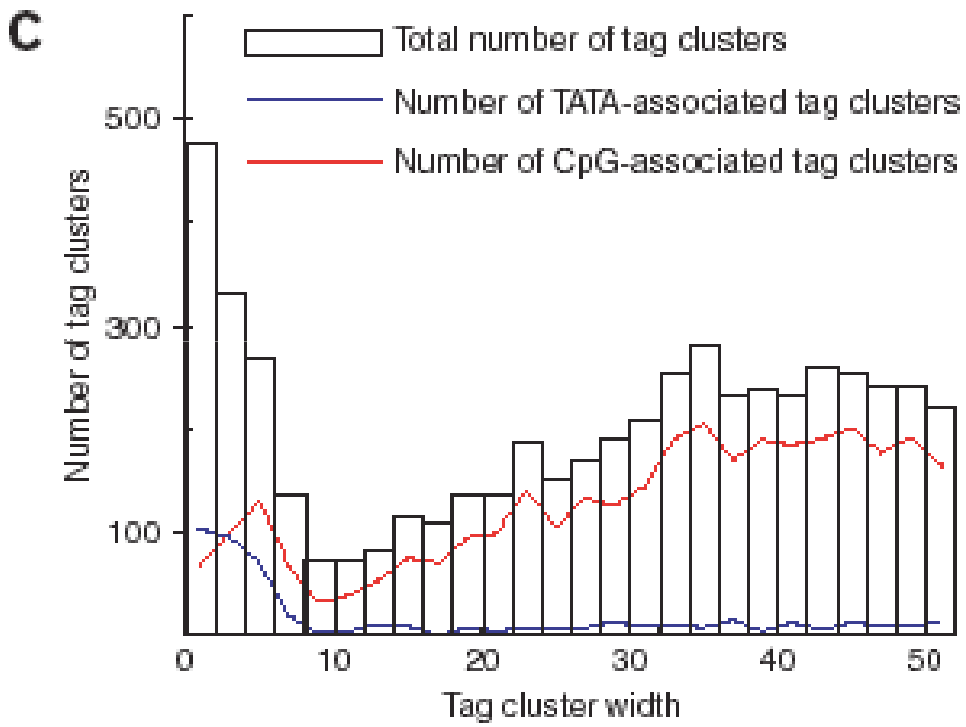


Figure 1 Definition and characteristics of CAGE tag clusters. (**a**) Tag clusters are produced by grouping overlapping tags on the same strand. Hence, tag clusters are defined by a start and end position, a count of tags and a distribution of these counts. Unique tag starts within the tag cluster form CAGE tag starting sites (CTSSs).

Students you must go back to previous lectures to find out how a CAGE library is generated and what a CAGE tag is.



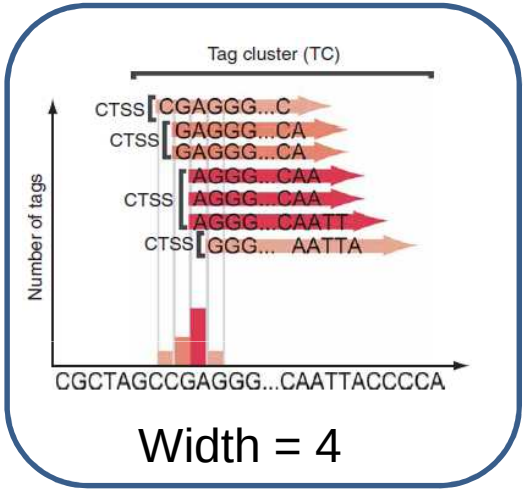


Figure 1. (c) Association of tag cluster **width** (minimal length of the sequence fragment containing >80% of all tags in the cluster) with TATA boxes and CpG islands for tag clusters with >100 tags.

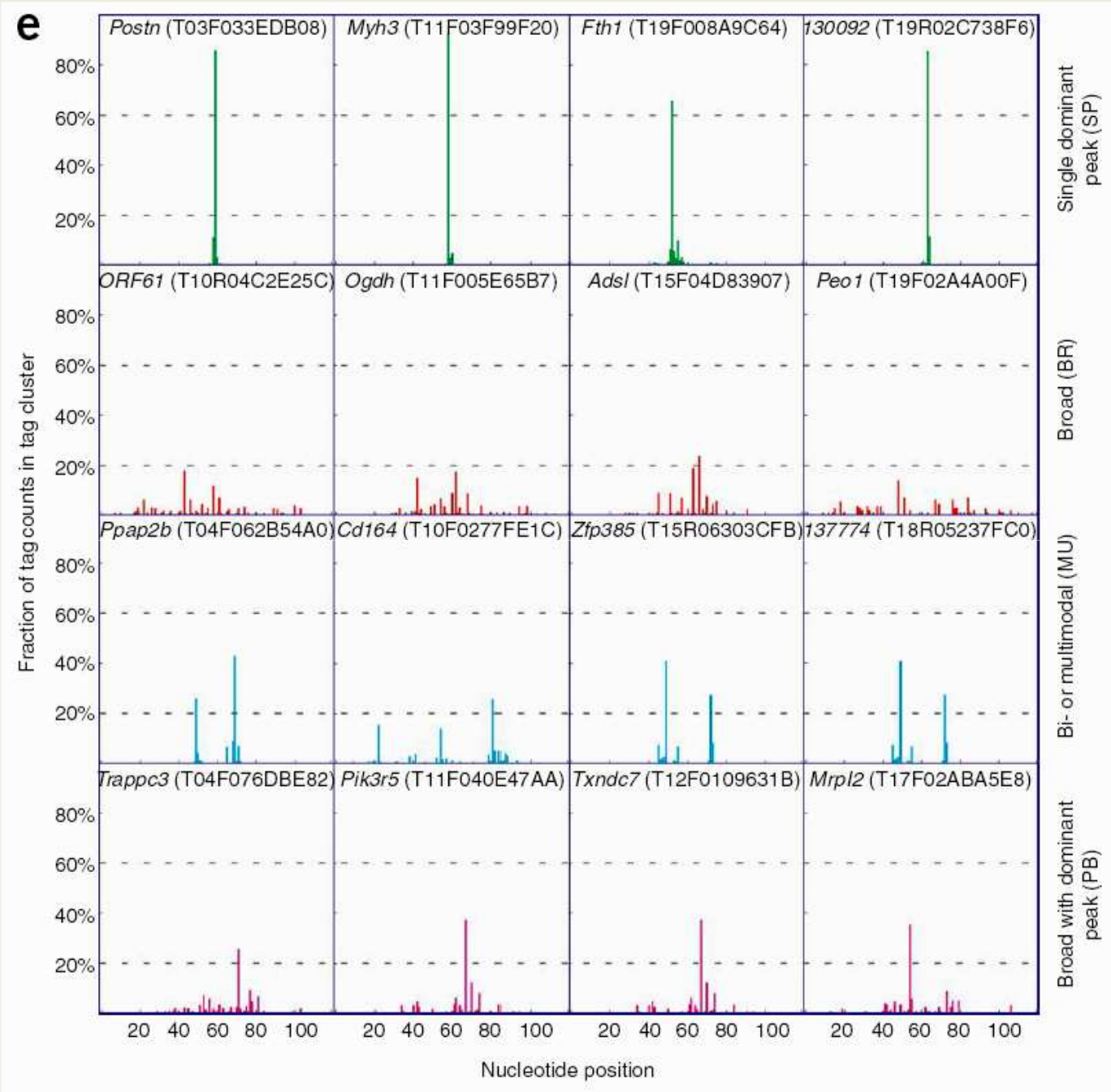
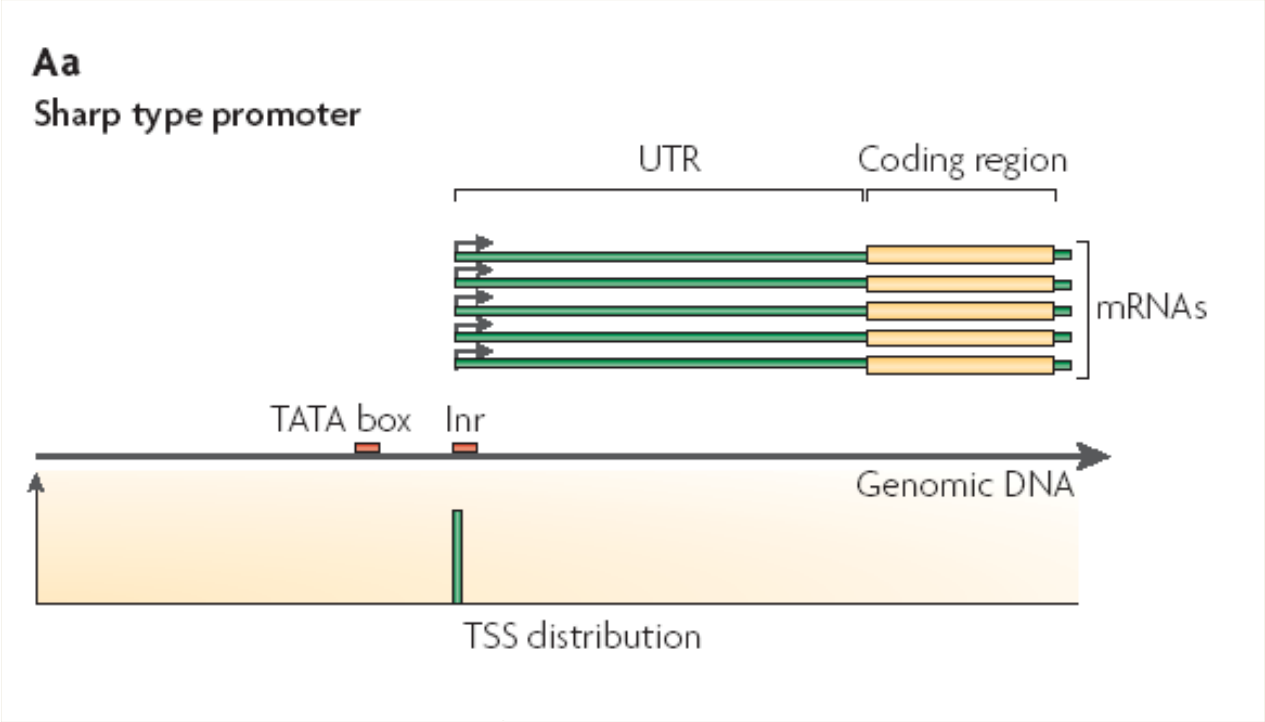
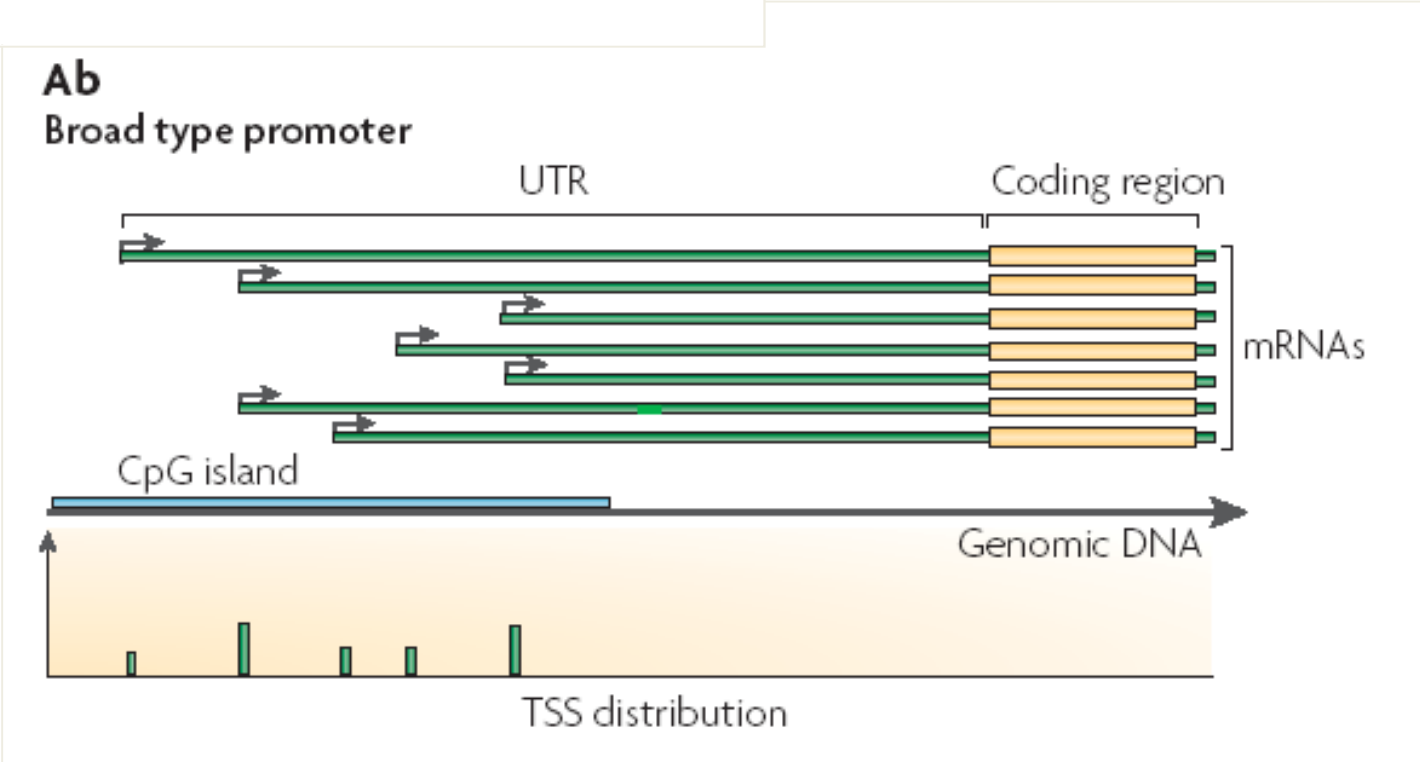


Figure 1. (e) Arrays of representative tag clusters for different shape classes. Histograms indicate the fraction of tags in the tag cluster mapping into each position in a 120-bp window centered on the tag cluster. The single peak (SP) class is characterized by a sharp peak, indicative of a single, well-defined TSS. The broad (BR) shape indicate multiple, weakly defined TSSs. The bimodal/multimodal (MU) shape class implies multiple welldefined TSSs within one cluster. Combination of a well-defined TSS surrounded by weaker TSSs results in a broad with dominant peak shape (PB). HUGO gene names or transcriptional unit identifiers for cognate genes and tag cluster identifiers are shown above each tag cluster.


Aa
Sharp type promoter
UTR
Coding region
mRNAs
TATA box
Inr
Genomic DNA
TSS distribution


Ab
Broad type promoter
UTR
Coding region
mRNAs
CpG island
Genomic DNA
TSS distribution

a

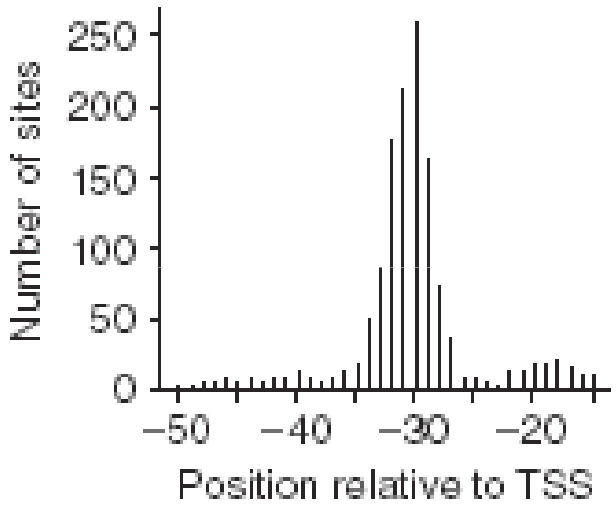
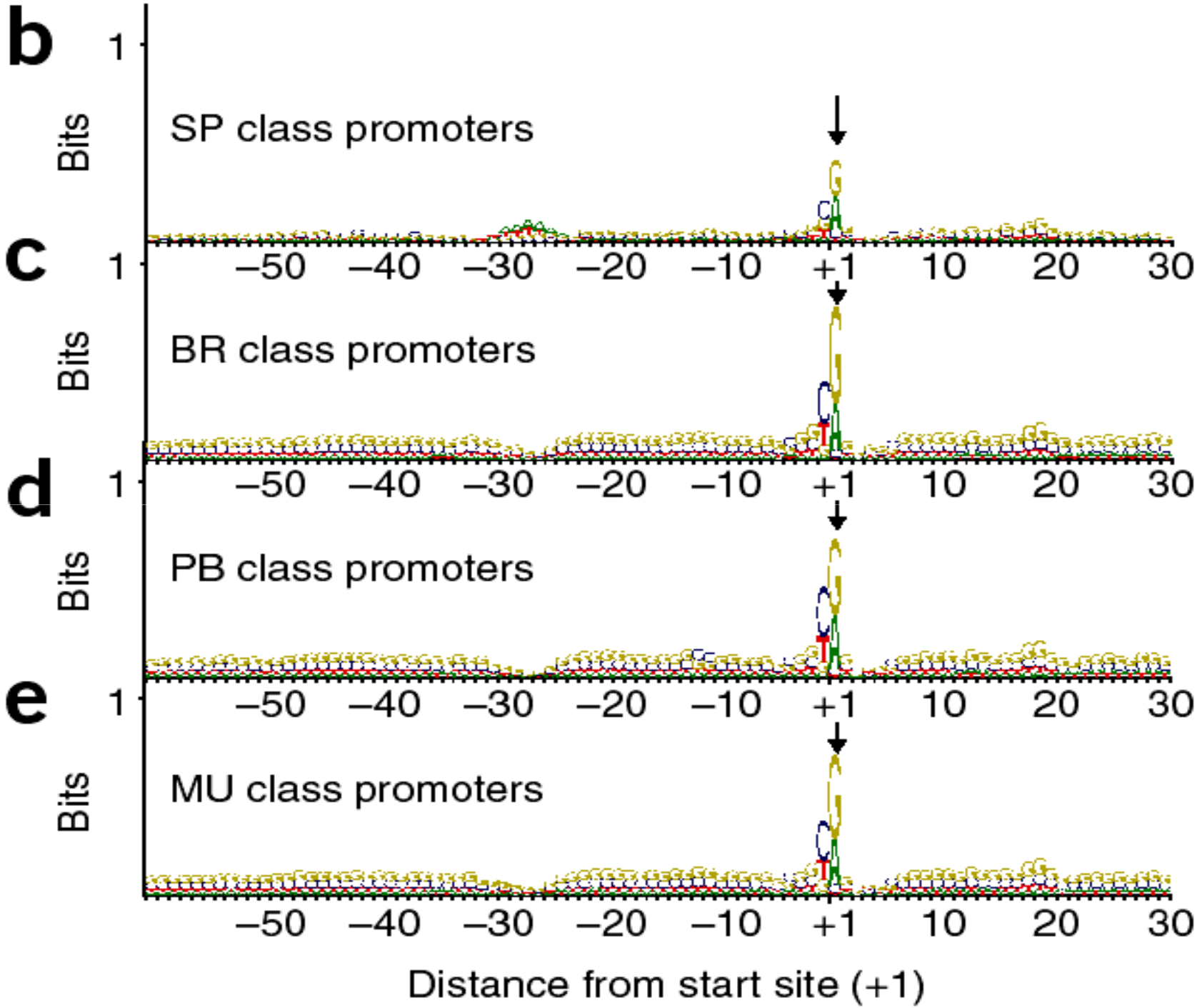


Figure 2. TATA-box and TSS spacing definition and consensus. (a) Accurate distribution of the spacing between TATA-box promoter and initiation sites.



← Figure legend

Legend to the previous slide

Figure 2. TATA-box and TSS spacing definition and consensus.
(b–e) Sequence logos for promoter sequences aligned at the TSSs constructed by counting each tag and its flanking region as one sequence, divided by promoter shape class. The y axis shows the information content (measured in bits), reviewed in ref. 15. In all cases, there is a clear preference for a pyrimidine-purine initiation site at –1,+1. A TATA-like motif is visible around the –30 position in the SP class promoters (b). In the BR class promoters, as most of those promoters are overlapped by CpG islands, the entire region is GC-rich; there is anisotropy of nucleotide content: there are more guanine than cytosine nucleotides in the plus strand upstream of the TSS (c). The logos of PB (d) and MU (e) class promoters look similar to this, indicating that these two ambiguous two categories are more likely to share the common initiation mechanism with BR promoters than with the SP ones. The PB class has a certain proportion of mixed cases, with both a CpG island and a TATA-box.

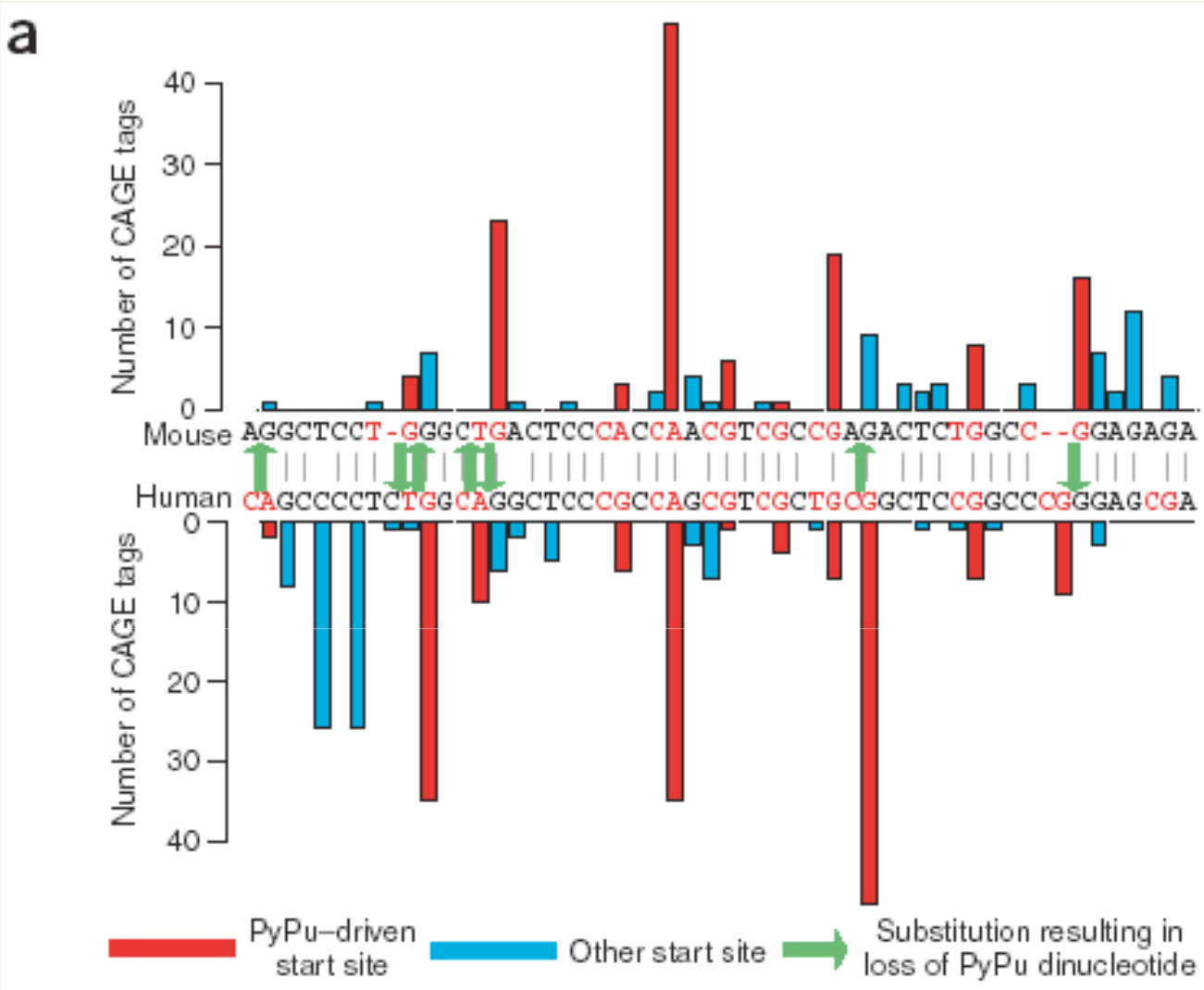
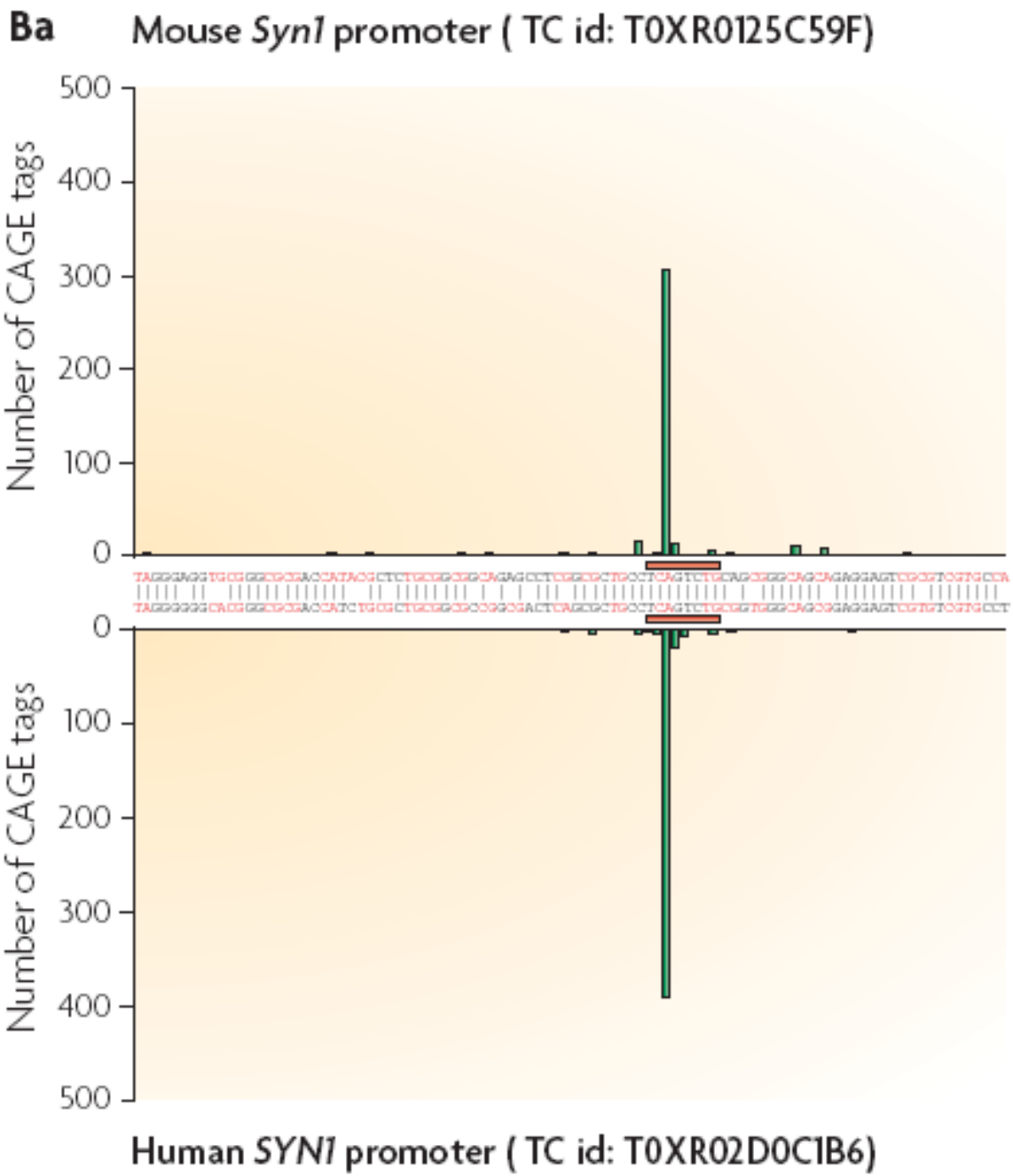
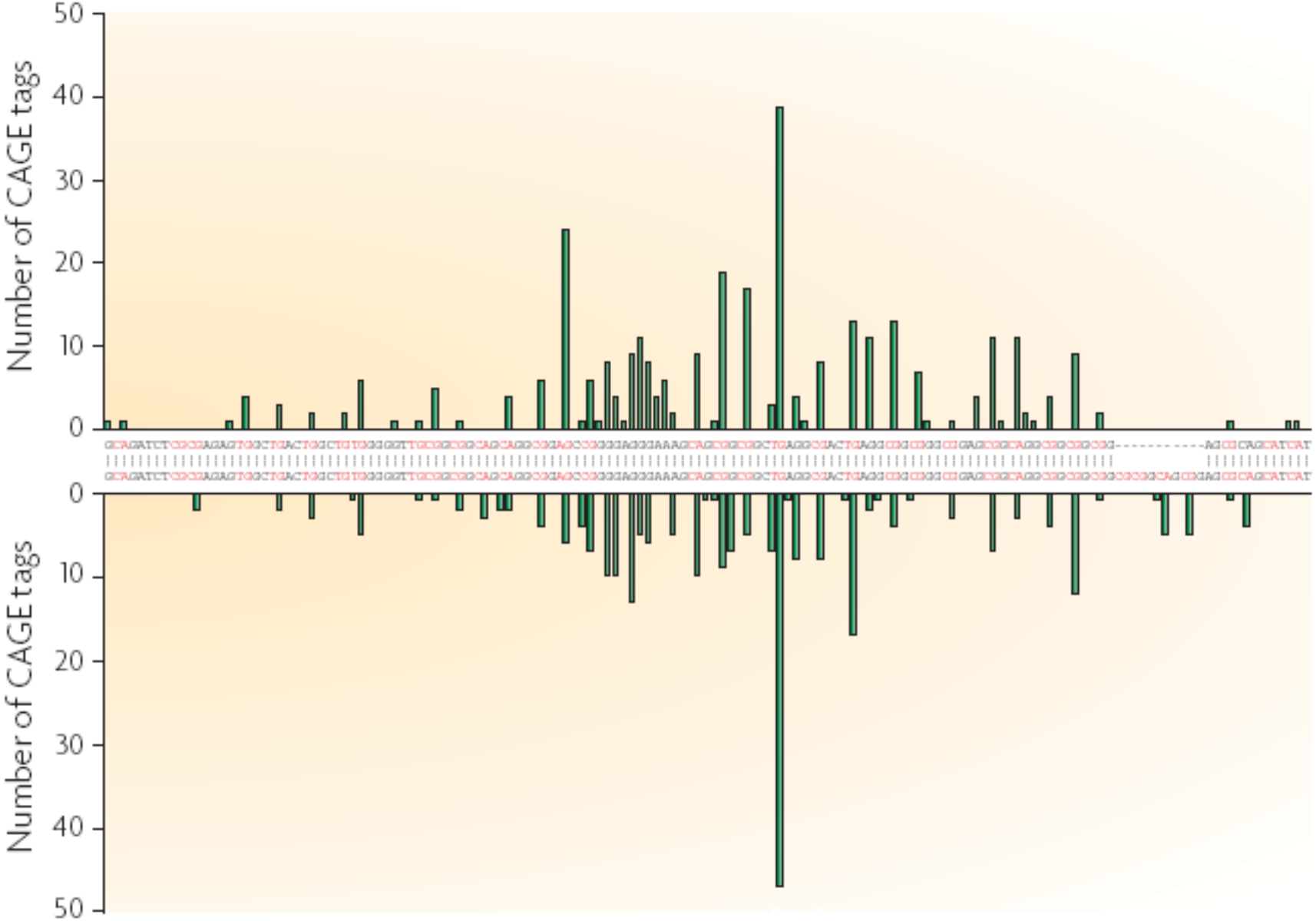


Figure 4 Pyrimidine-purine dinucleotides drive expression.
(a) A detailed view of the core promoter of the mouse Ptprn gene (TC 73140) and corresponding human region illustrates the usage of pyrimidine-purine dinucleotides as dominant start sites and the expression changes resulting from mutations at these positions.

Promoter structures are conserved



Bb Mouse *Pura* promoter (TC id: T18F0230753D)



Human *PURA* promoter (TC id: T05F085033E0)

What have we learnt from genome-wide expression analysis ?

-more protein-coding genes than known

-most of genes produce more RNA transcripts than expected (derived both from nonclassical sense or antisense transcription and by alternative RNA processing)

-unexpected variety of noncoding RNA transcripts of various size (from very short to very long, from intergenic to intragenic), whose functions are mostly unknown (with the exception of miRNA and few long ncRNA)

-tissue- , cell- and stage-specific expression of both protein-coding and noncoding RNA widely confirmed

Most transcript use RNA Polymerase II