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)

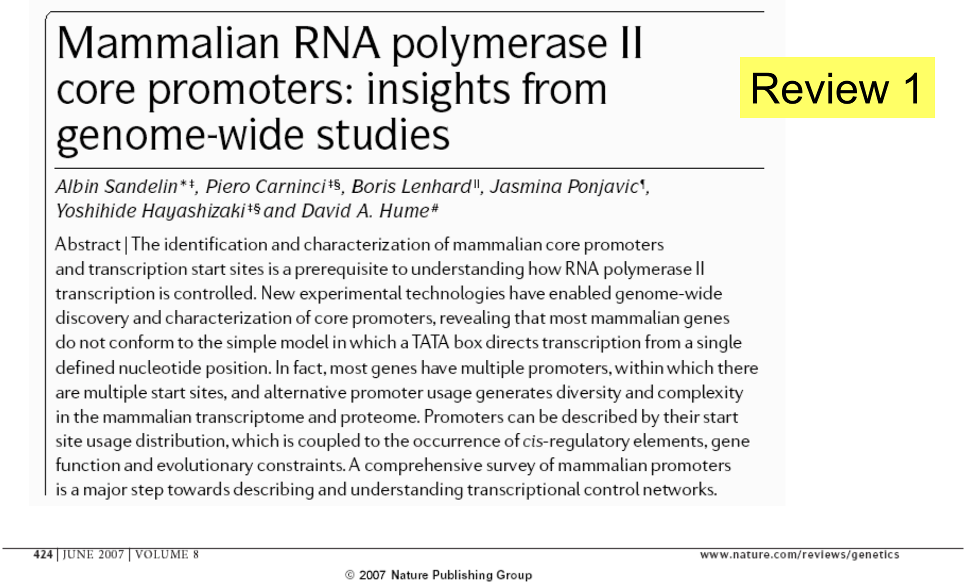
-unexpected variety of noncoding RNA transcripts of various size (from very short to very long, from intergenic to intragenic), whose functions are mostly unknown (with the exception of miRNA and few long ncRNA)

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

Most transcript use RNA Polymerase II

Today and in the following lessons, we will consider some important advances put forward by studies with CAGE, tiling arrays and RNA-Seq :

- Structure and variety of TU, with insight into functions
- structure and types of mammalian promoters
- the widespread use of alternative splicing in mammals



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

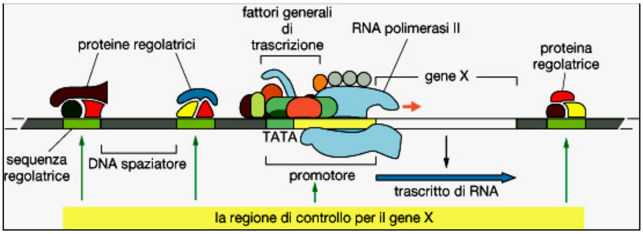
Albin Sandelin*‡, Piero Carninci‡§, Boris Lenhard||, Jasmina Ponjavic¶, Yoshihide Hayashizaki‡§ and David A. Hume#

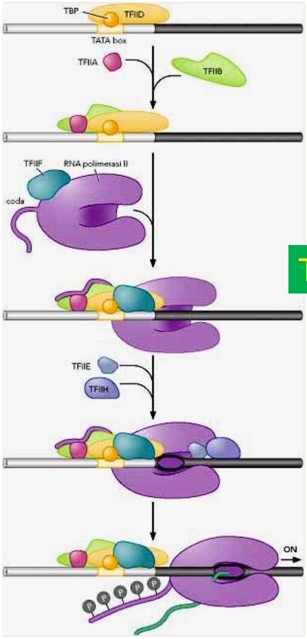
Abstract | The identification and characterization of mammalian core promoters and transcription start sites is a prerequisite to understanding how RNA polymerase II transcription is controlled. New experimental technologies have enabled genome-wide discovery and characterization of core promoters, revealing 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.

424 | JUNE 2007 | VOLUME 8 www.nature.com/reviews/genetics

© 2007 Nature Publishing Group

Review 1





Transcription in Eukaryotes

General Transcription Factors

Regulatory elements in Eukaryotes

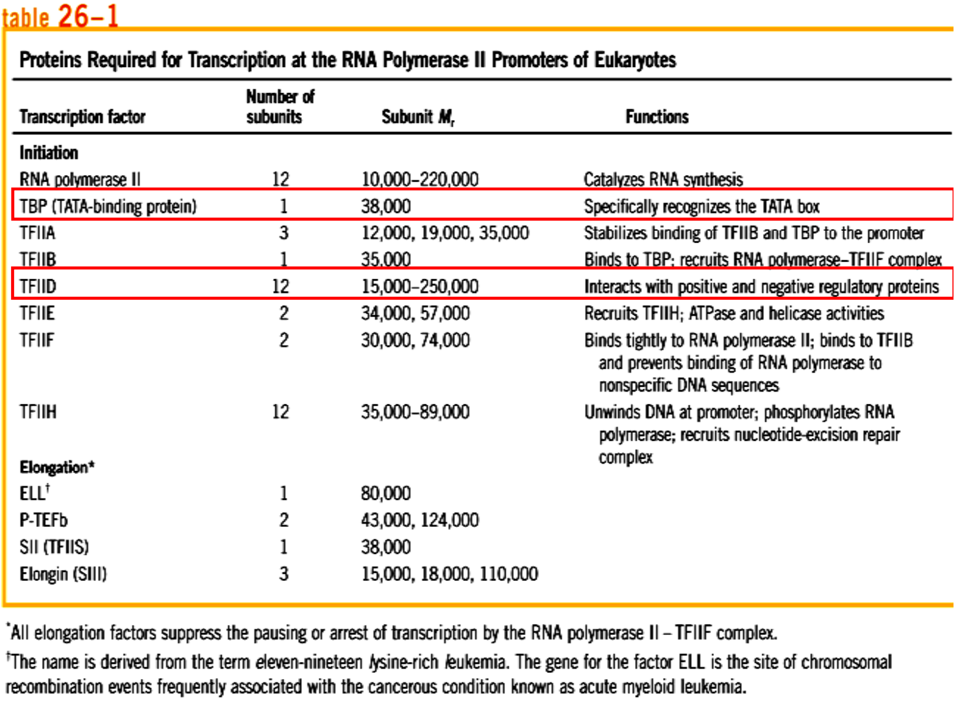


table 26–1

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.

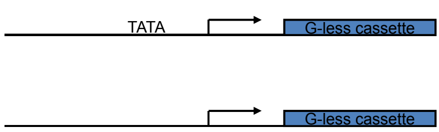
[†] The name is derived from the term *e*leven-nineteen *l*ysine-rich *l*eukemia. The gene for the factor ELL is the site of chromosomal recombination events frequently associated with the cancerous condition known as acute myeloid leukemia.

Although TFIID purifies as a multimeric protein, starting from recombinant TAFs and recombinant TBP expressed in E. coli, R. Tjian's group has shown that **partial complexes** can be assembled and tested in "*in vitro* transcription" experiments.

These experiments showed that transcription from TATA-box containing promoters can be correctly initiated using only TBP, TFIIB and RNA Pol II.

On the contrary, TATA-less promoters require TFIID (i.e. the TAFs) for initiation. In particular, a sub-complex containing TAFII-250 and TAFII-150 (drosophila) was enough to initiate transcription from a INR-containing promoter and TAFII40+TAFII60 was enough for a DPA-containing promoter.

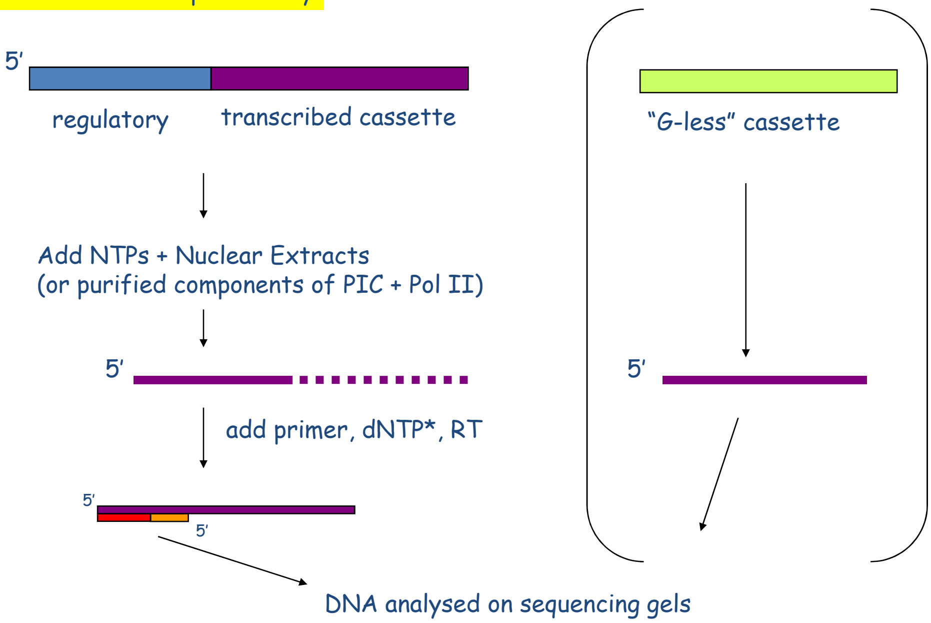
in vitro transcription experiment:



	TBP TFIIB PolII	TFIID TFIIB PolII	TBP TFIIB PolII TAFs (partial)
TATA promoter	+	+	+
TATA-less promoter	-	+	+/- (*)

(*) depending on TAFs added

In vitro transcription assay


5′
regulatory
transcribed cassette
Add NTPs + Nuclear Extracts
(or purified components of PIC + Pol II)
5′
add primer, dNTP*, RT
5′
5′
DNA analysed on sequencing gels
“G-less” cassette
5′

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

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

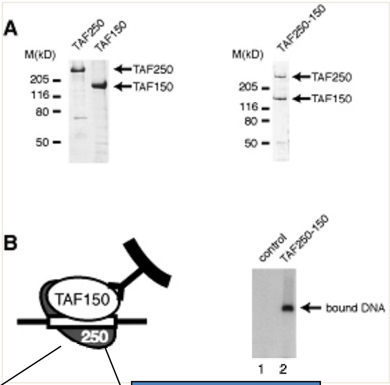
gene 1:		a	t	t	g	t	c	t	a
gene 2:		a	t	a	g	t	g	a	g
gene 3:		a	t	t	c	t	g	g	c
gene 4:		a	t	t	g	t	c	c	g
gene 5:		a	t	t	c	t	g	a	t
gene 6:		a	t	a	g	g	c	t	c
etc:		-	-	-	-	-	-	-	-
frequency	a	100	0	35	0	0	0	25	20
	t	0	100	65	0	90	0	25	20
	c	0	0	0	35	0	50	20	30
	g	0	0	0	65	10	50	30	30
Consensus:		**A**	**T**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

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?

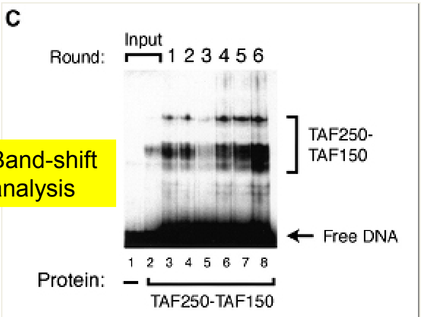
Da Chalklev e Tjian (1999), EMBO J., 18: 4835.



Recombinant TAFs

This complex was used for SELEX (see next slides)

Band-shift analysis

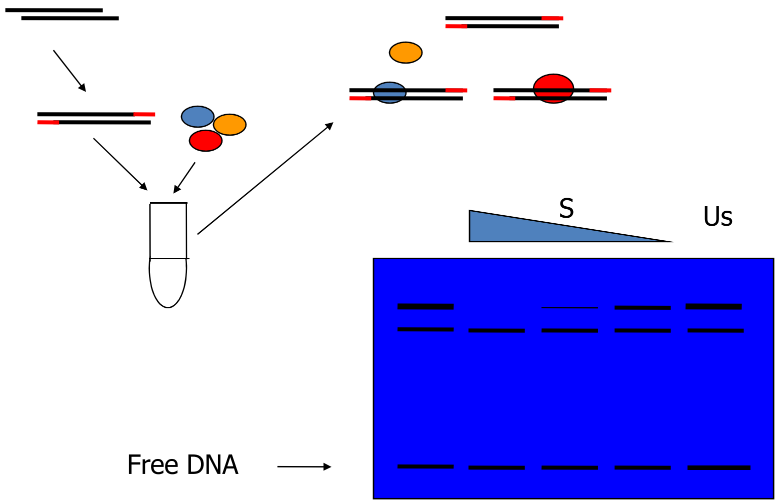


D

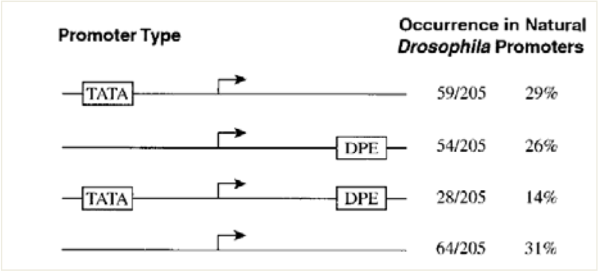
G	21	16	7	–	–	–	15	–	–	9	16	16	12
A	7	9	15	–	–	56	13	16	9	–	12	7	15
T	4	15	18	25	14	–	12	40	23	18	4	14	7
C	17	9	9	31	42	–	16	–	24	29	18	10	12
R	28	25	22	–	–	56	28	16	9	9	28	23	27
Y	21	24	27	56	56	–	28	40	47	47	22	24	19

INR consensus derived bioinformatically: Y $\overset{C}{t}$ A N $\overset{T}{a}$ Y Y

Band-shift assay or Electrophoretic Mobility Shift Assay EMSA



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.



Promoter Type	Occurrence in Natural *Drosophila* Promoters	
TATA	59/205	29%
DPE	54/205	26%
TATA DPE	28/205	14%
	64/205	31%

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.

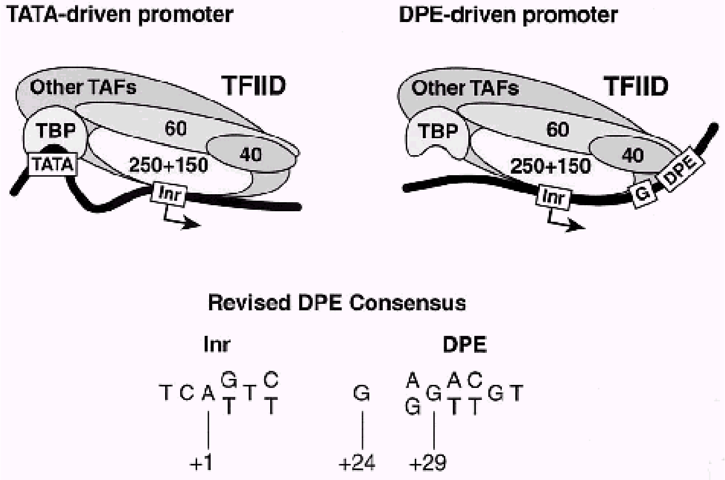
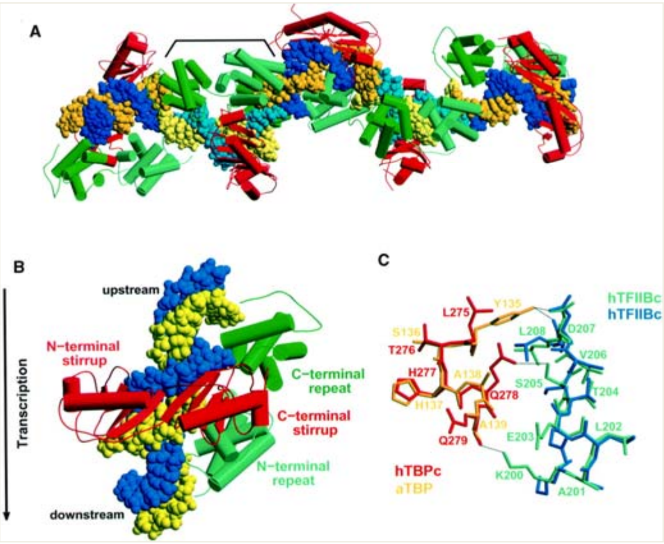


FIG. 7. A model of two distinct interactions of TFIID with TATA- versus DPE-driven core promoters. The model is discussed in the text. TAFs, TBP-associated factors.

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

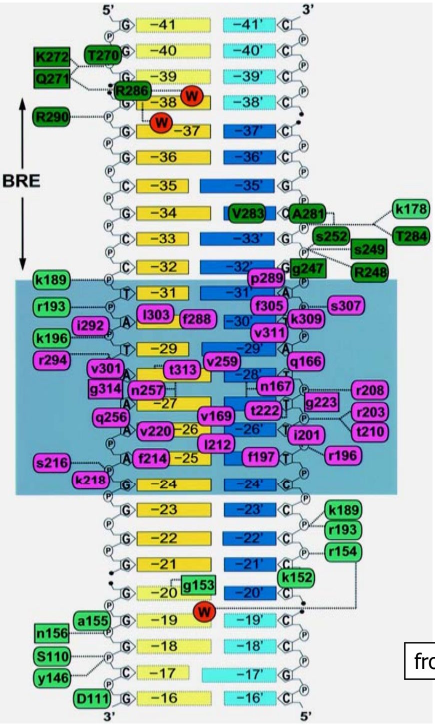
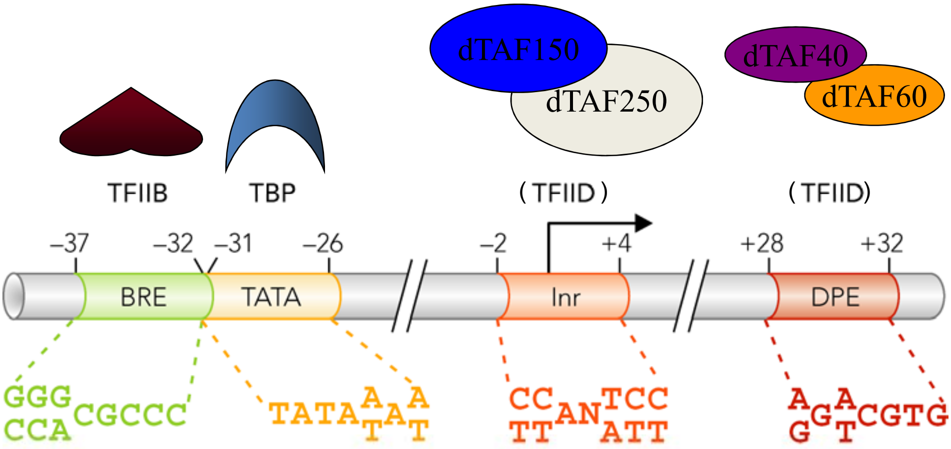


Fig. 3. Protein-DNA contacts that specify the orientation of the hTFIIBc-hTBPc-MLP complex. Complete schematic illustrating all protein-DNA interactions in the ternary complex.
Pink: TBP. Green: TFIIB.
Arrows indicate the location of the BRE.
An oval indicates an interaction between the promoter and 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



The “textbook” promoter

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.

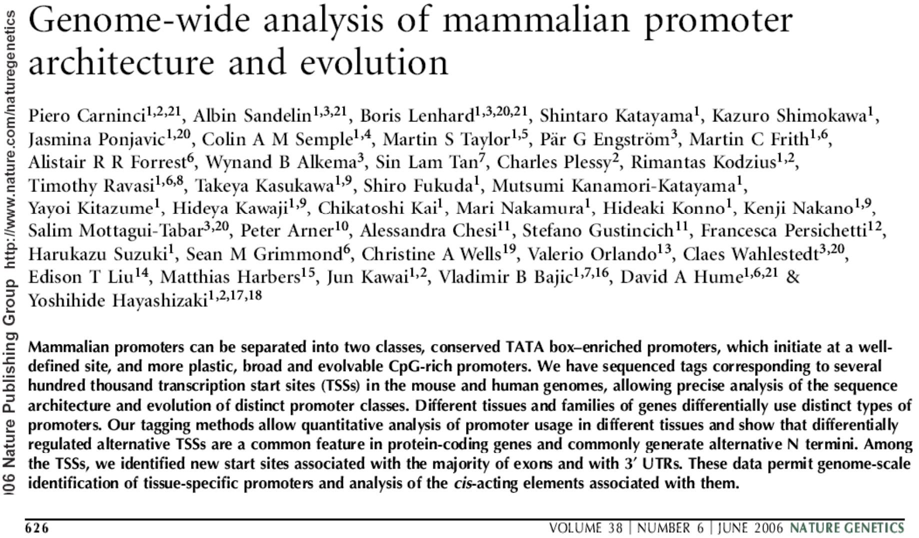
Several promoters, less defined and more difficult to study, were left apart, e.g. the CpG – type promoters.

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.



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.

626 VOLUME 38 | NUMBER 6 | JUNE 2006 NATURE GENETICS

146 mouse cDNA libraries

41 human cDNA libraries

!

By CAGE analysis (see genomics lesson)

Students you must go back to the “genomics” lesson to find out how a CAGE library is generated and what a CAGE tag is.

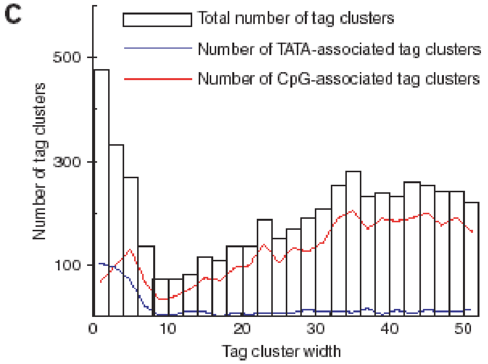


Figure 1. (c) Association of tag cluster **width** (minimal length of the sequence fragment containing >80% of all tags in the cluster) with TATA boxes and CpG islands for tag clusters with >100 tags.

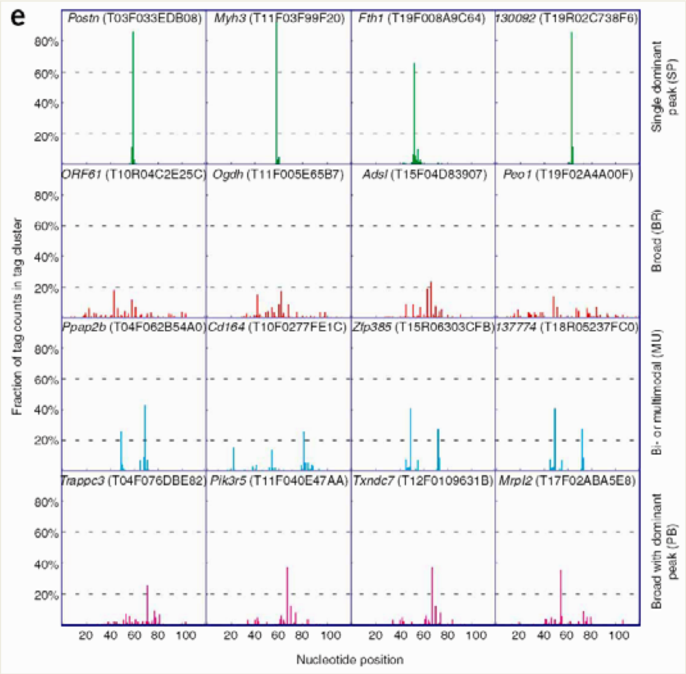
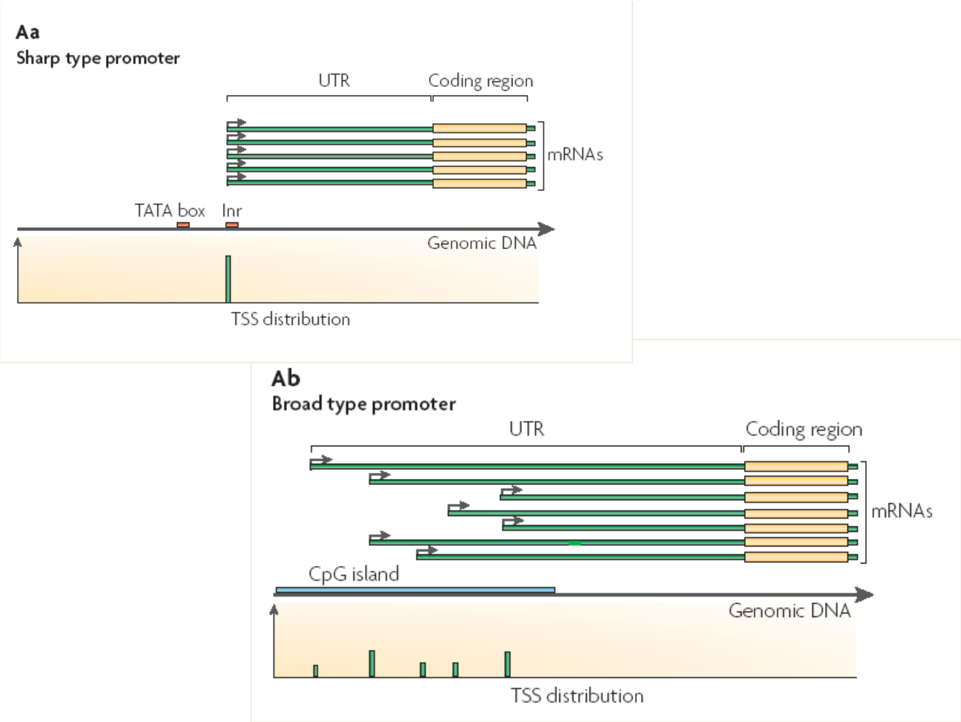


Figure 1. (e) Arrays of representative tag clusters for different shape classes. Histograms indicate the fraction of tags in the tag cluster mapping into each position in a 120-bp window centered on the tag cluster. The single peak (SP) class is characterized by a sharp peak, indicative of a single, 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.



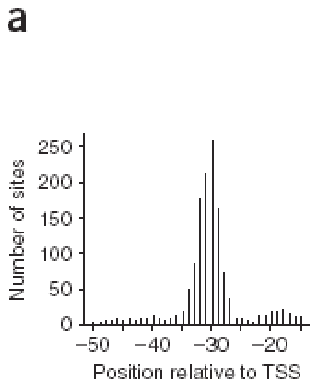
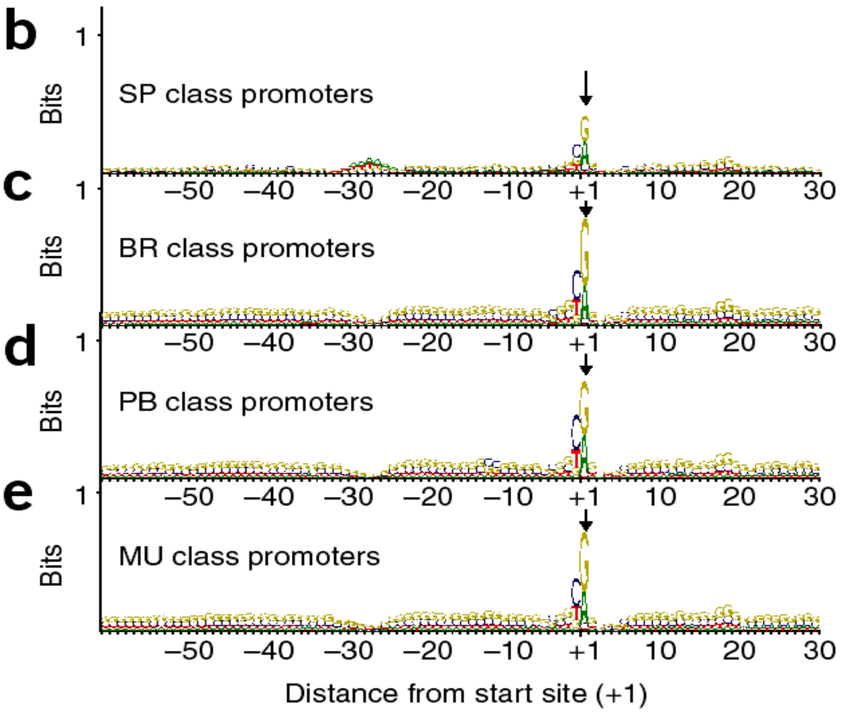


Figure 2. TATA-box and TSS spacing definition and consensus. (a) Accurate distribution of the spacing between TATA-box promoter and initiation sites.



← 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.

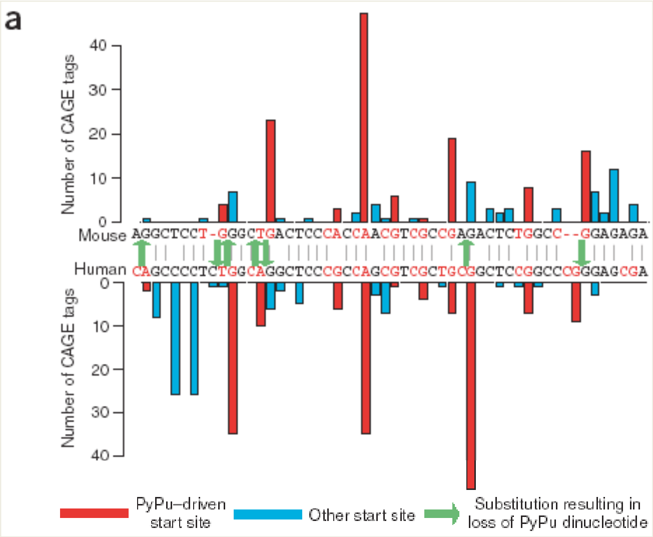
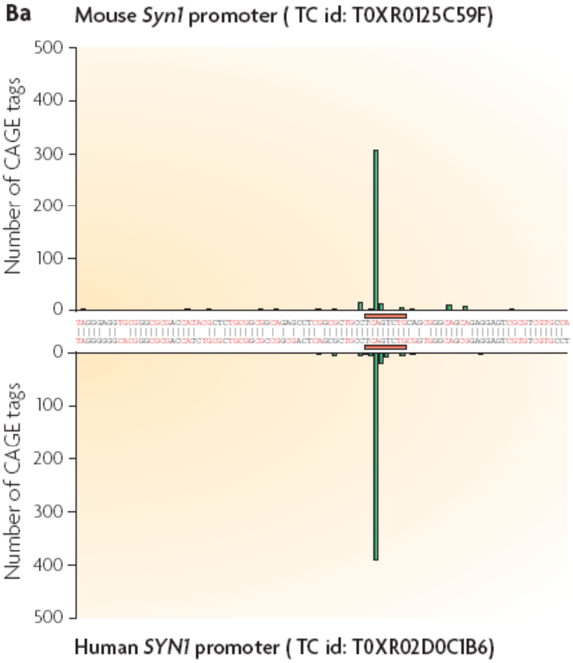
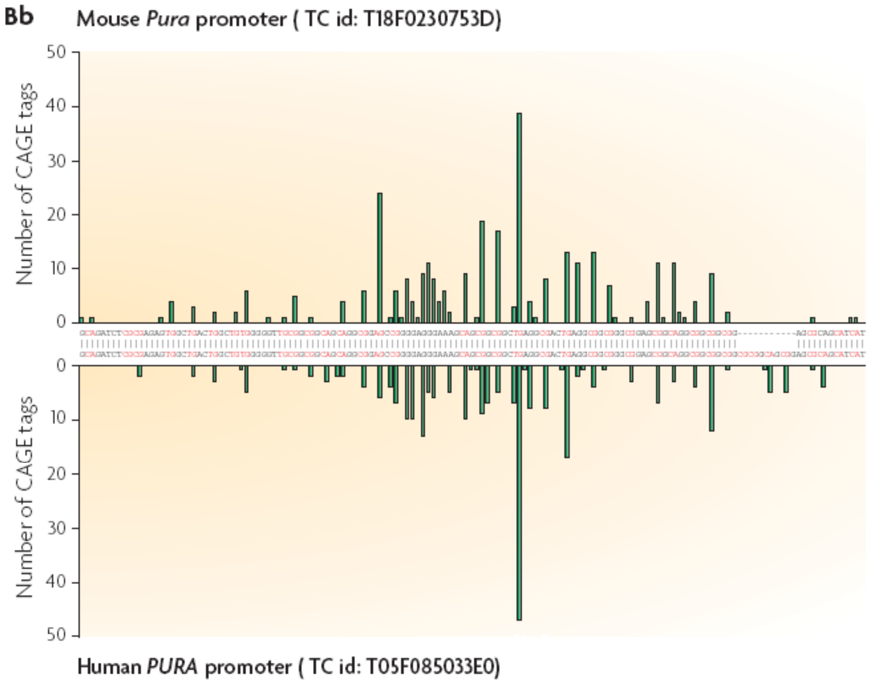


Figure 4 Pyrimidine-purine dinucleotides drive expression.
(a) A detailed view of the core promoter of the mouse Ptprn gene (TC 73140) and corresponding human region illustrates the usage of pyrimidine-purine dinucleotides as dominant start sites and the expression changes resulting from mutations at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

Promoter structures are conserved





Bb
Mouse *Pura* promoter (TC id: T18F0230753D)
Number of CAGE tags
50
40
30
20
10
0
0
10
20
30
40
50
Number of CAGE tags
Human *PURA* promoter (TC id: T05F085033E0)