Ch 3 - L 5.1

Promoters

L3.5 - Agenda

1. Mapping of TSS
2. Chromatin marks – CpG methylaion
3. More than one type of gene promoter
4. Modes of transcriptional activation

Basics of Transcription initiation in Eukaryotes

Go to any Molecular Biology book for revising

One page in the **auxiliaries** in Moodle contains
- 1 video by Dr Robert Tjian
- 1 video by MDB.

https://cmb.i-learn.unito.it/mod/book/view.php?id=12867.

Textbook

Seminars in Cell & Developmental Biology 57 (2016) 11–23

Contents lists available at ScienceDirect

Seminars in Cell & Developmental Biology

journal homepage: www.elsevier.com/locate/semcdb

Review

Promoter architectures and developmental gene regulation

Vanja Haberle[a,b,1], Boris Lenhard[a,*]

[a] Institute of Clinical Sciences and MRC Clinical Sciences Center, Faculty of Medicine, Imperial College London, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK
[b] Department of Biology, University of Bergen, Thormøhlensgate 53A, N-5008 Bergen, Norway

ARTICLE INFO

Article history:
Received 5 November 2015
Received in revised form 8 January 2016
Accepted 11 January 2016
Available online 16 January 2016

Keywords:
Transcriptional regulation
Core promoter
Transcription start sites
CAGE
Promoter types
Overlapping codes

ABSTRACT

Core promoters are minimal regions sufficient to direct accurate initiation of transcription and are crucial for regulation of gene expression. They are highly diverse in terms of associated core promoter motifs, underlying sequence composition and patterns of transcription initiation. Distinctive features of promoters are also seen at the chromatin level, including nucleosome positioning patterns and presence of specific histone modifications. Recent advances in identifying and characterizing promoters using next-generation sequencing-based technologies have provided the basis for their classification into functional groups and have shed light on their modes of regulation, with important implications for transcriptional regulation in development. This review discusses the methodology and the results of genome-wide studies that provided insight into the diversity of RNA polymerase II promoter architectures in vertebrates and other Metazoa, and the association of these architectures with distinct modes of regulation in embryonic development and differentiation.

First essential issue:
the identification of Promoters genome-wide

Promoter identification

To identify promoters (genome-wide) we need essentially to have clear definition of TSS. This is not easy since:

- classical methods (i.e. cDNA sysnthesis using RT enzyme) rarely reach the very 5'-end due to secondary RNA structures.
- normal RNA-seq which is based on random fragmentation shows very poor enrichment of terminal fragments.

People have used different methods to map promoters genome-wide:

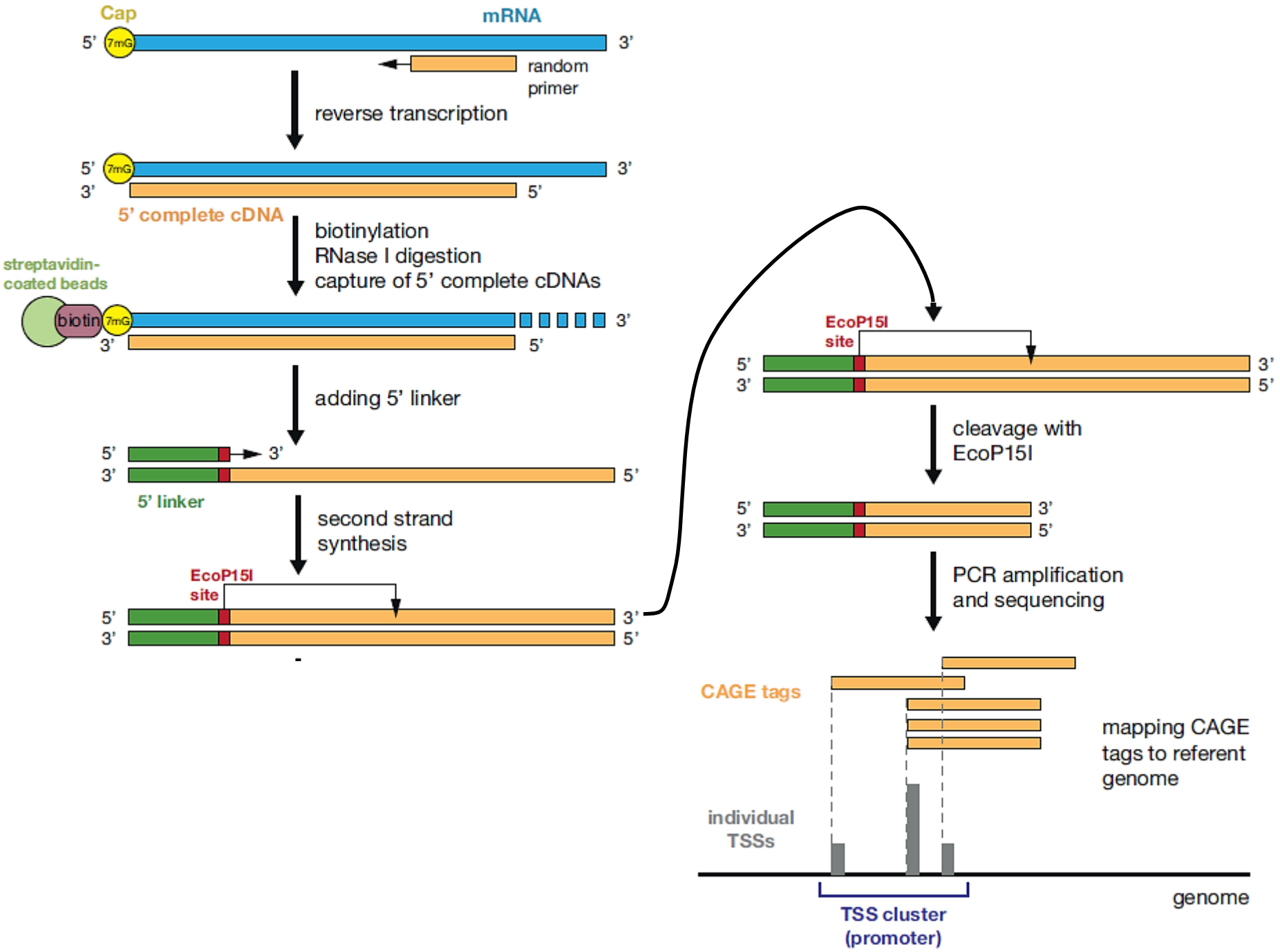
- CAGE (with both classical sequencing and later using NGS)
- 5'-SAGE (same)
- mapping PIC-component by ChIP-Seq
- mapping Histone PTMs/variants by ChIP-Seq
- Bioinformatics (prediction of basal promoter elements)

SAGE (Serial Analysis of Gene Expression) and CAGE (Capped RNA Analysis) were originally based on making short fragments from either 3'- or 5'-end of mRNAs, concatamerize and sequence by Sanger.

Most important:

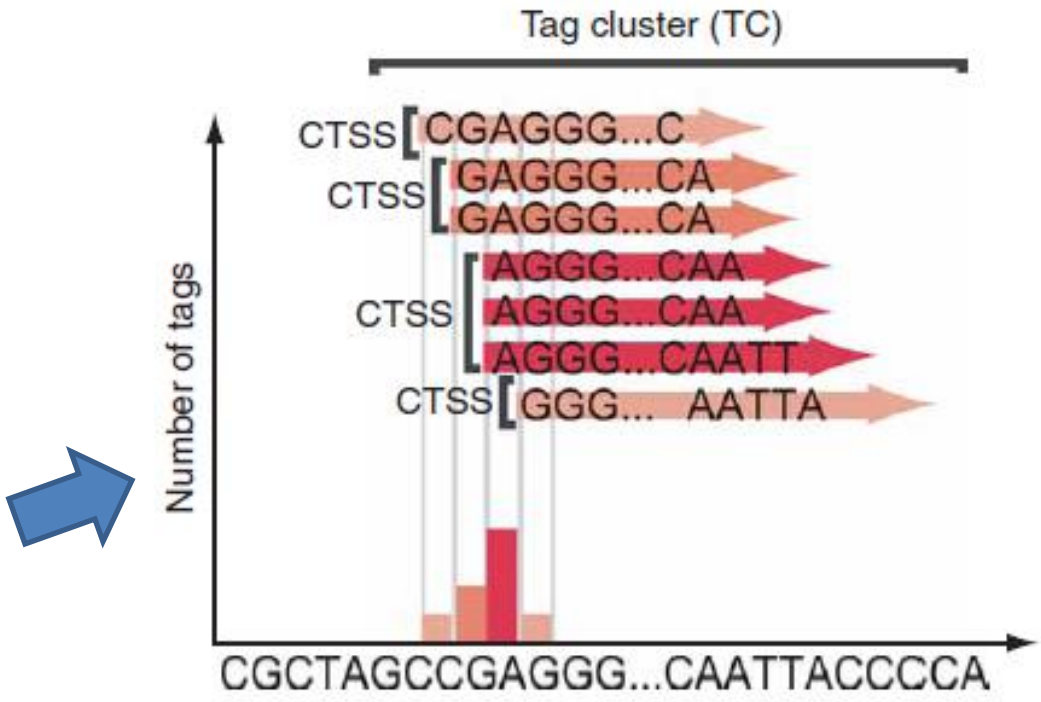
CAGE was developped specifically for 5'-end definition and is based on chemical modification of the the RNA Cap, allowing enrichment of correctly extended cDNAs.

Today, both methods were adapted to using NGS sequencing of the fragments obtained (e.g. CAGE-seq).


Cap
mRNA
5'
7mG
3'
random primer
reverse transcription
5' complete cDNA
biotinylation
RNase I digestion
capture of 5' complete cDNAs
streptavidin-coated beads
biotin
adding 5' linker
5' linker
second strand synthesis
EcoP15I site
cleavage with EcoP15I
PCR amplification and sequencing
CAGE tags
mapping CAGE tags to referent genome
individual TSSs
genome
TSS cluster (promoter)

Example of results from CAGE analysis

Carninci et al., 2006



(gene sequence)

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).

The FANTOM project used these methods to study a nuber of cell lines and tissues from Mouse and Human origin

These studies identified unprecedented numbers of TSSs therefore allowing intensive re-examination of Promoters features

It was clearly seen that, depending on the shape and dispersion of TSSs, Promoters could be grouped in (at least) four different groups, as exemplified in the following figures:

1. Single peak class (SP) (a single nucleotide or with few alternatives around it)
2. Broad TSS (various nucleotides in a range)
3. Bi- or multi-modal (some dominant peaks within broader initiation sites)
4. Broad with dominant peak (much like mixing type 1 and 2)

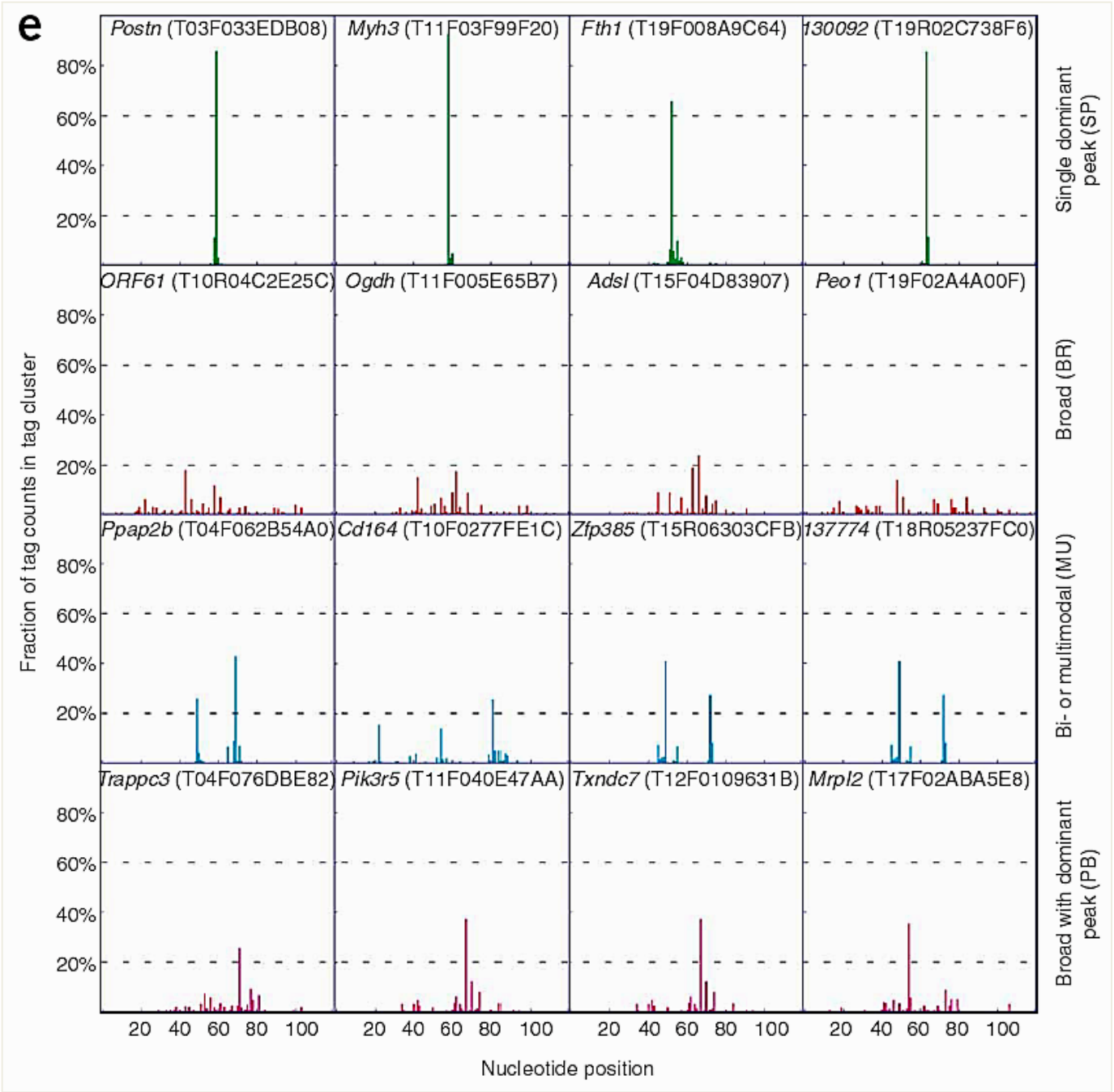
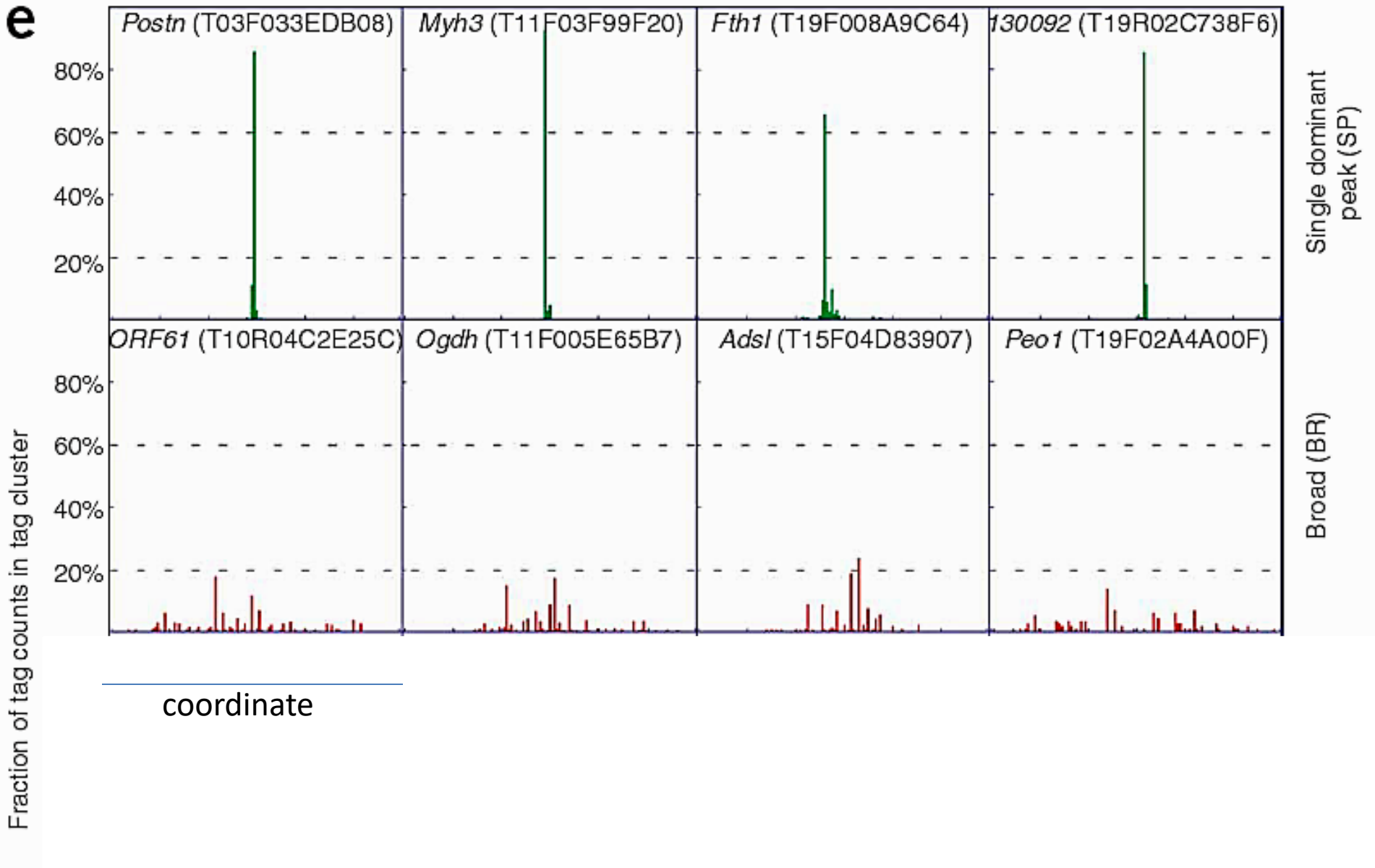


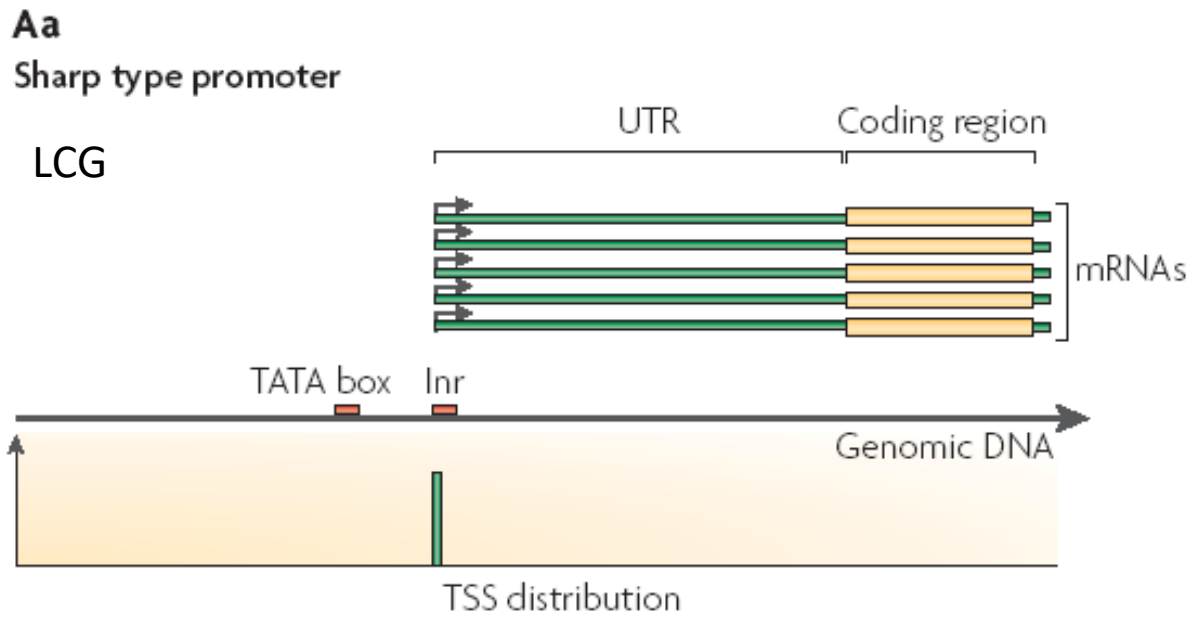
Figure 1. (e) Arrays of representative tag clusters for different shape classes. Histograms indicate the fraction of tags in the tag cluster mapping into each position in a 120-bp window centered on the tag cluster. The single peak (SP) class is characterized by a sharp peak, indicative of a single,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

Carninci et al., 2006


e
Postn (T03F033EDB08)
Myh3 (T11F03F99F20)
Fth1 (T19F008A9C64)
130092 (T19R02C738F6)
ORF61 (T10R04C2E25C)
Ogdh (T11F005E65B7)
Adsl (T15F04D83907)
Peo1 (T19F02A4A00F)
80%
60%
40%
20%
Single dominant peak (SP)
Broad (BR)
Fraction of tag counts in tag cluster
coordinate

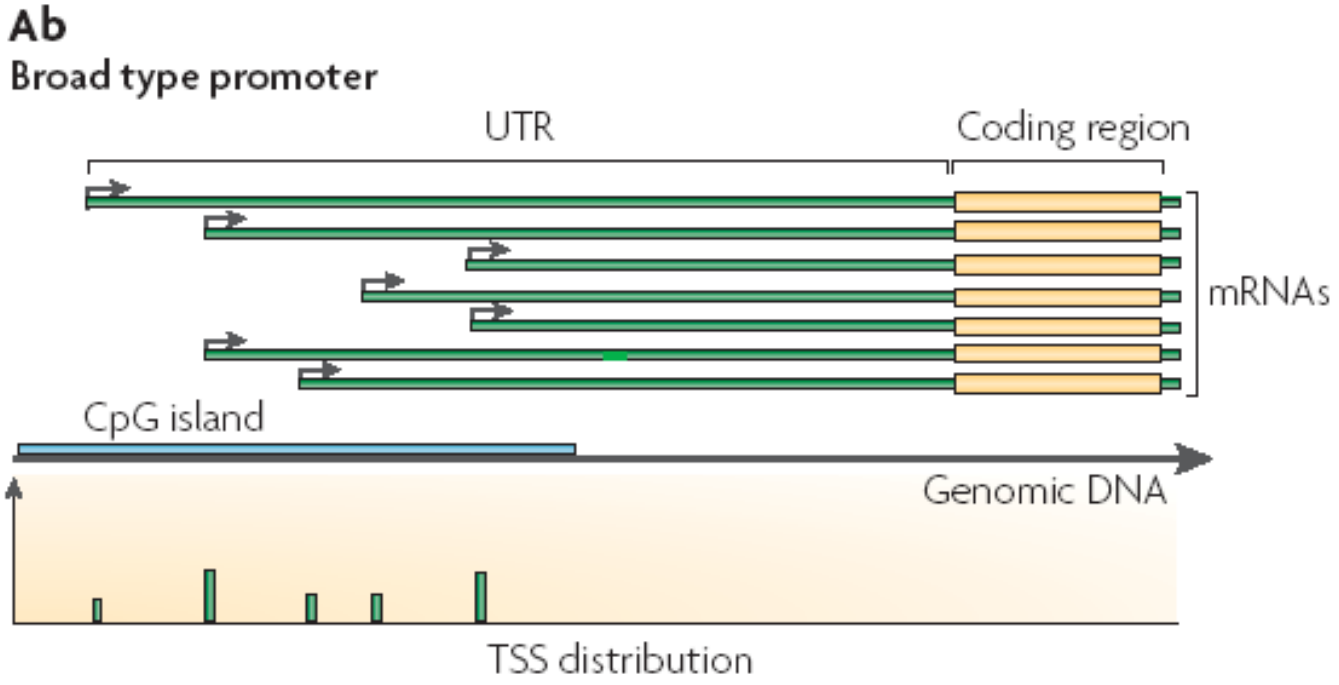
For common usage, we classify today in only two classes,
as
«sharp-type promoters»
and
«broad-type promoters»

intended that many situations in between exist.



The two extreme models

HCG



Sandelin et al., 2007

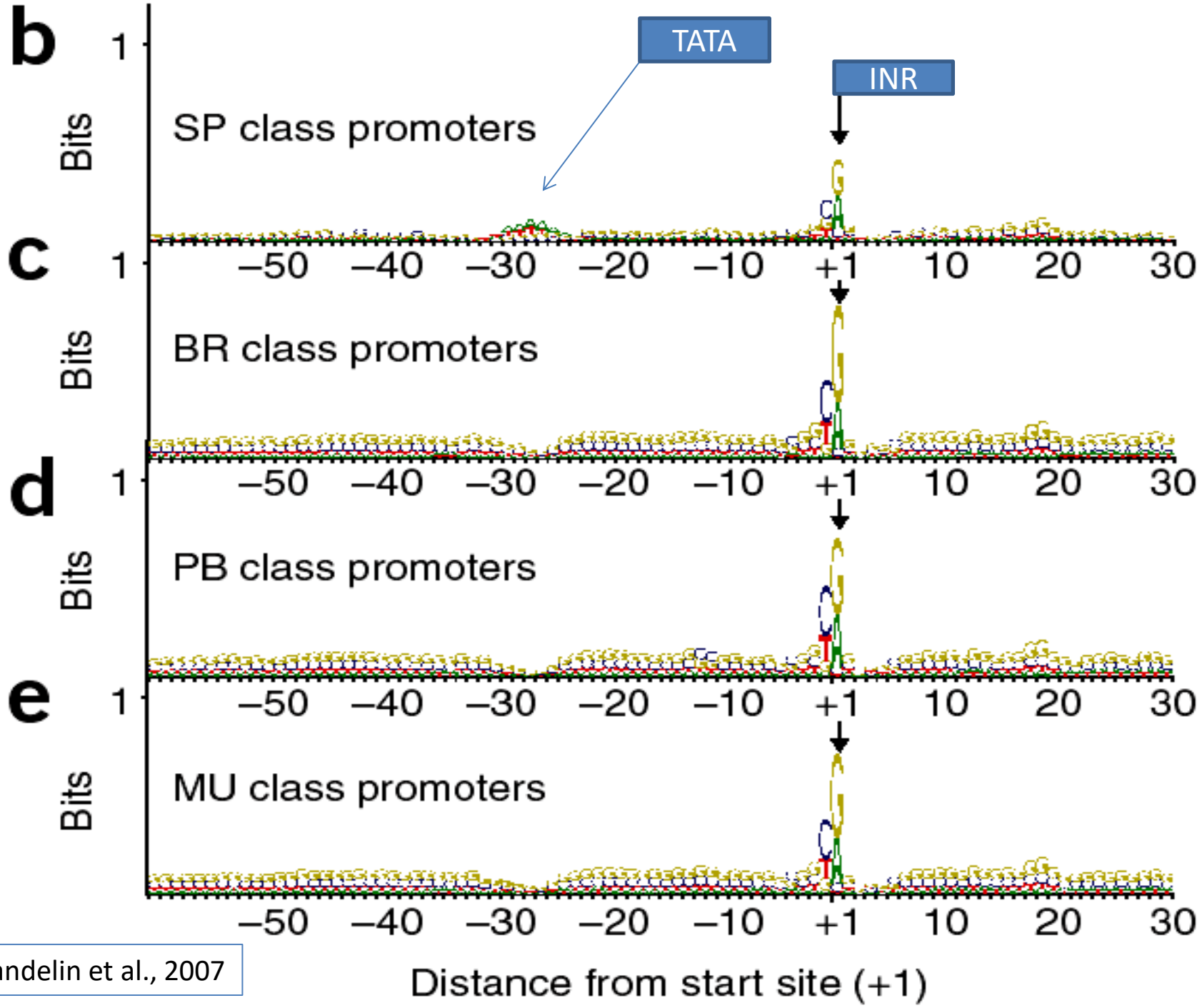
Genomic studies have partially changed our knowledge of promoters.

These studies demonstrated, first, that the “textbook promoter” with a clear TATA-box and a single TSS, is present at no more that 10-20% of mammalian genes (<15% in human and mouse), which represent a group of tissue-specific genes.

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

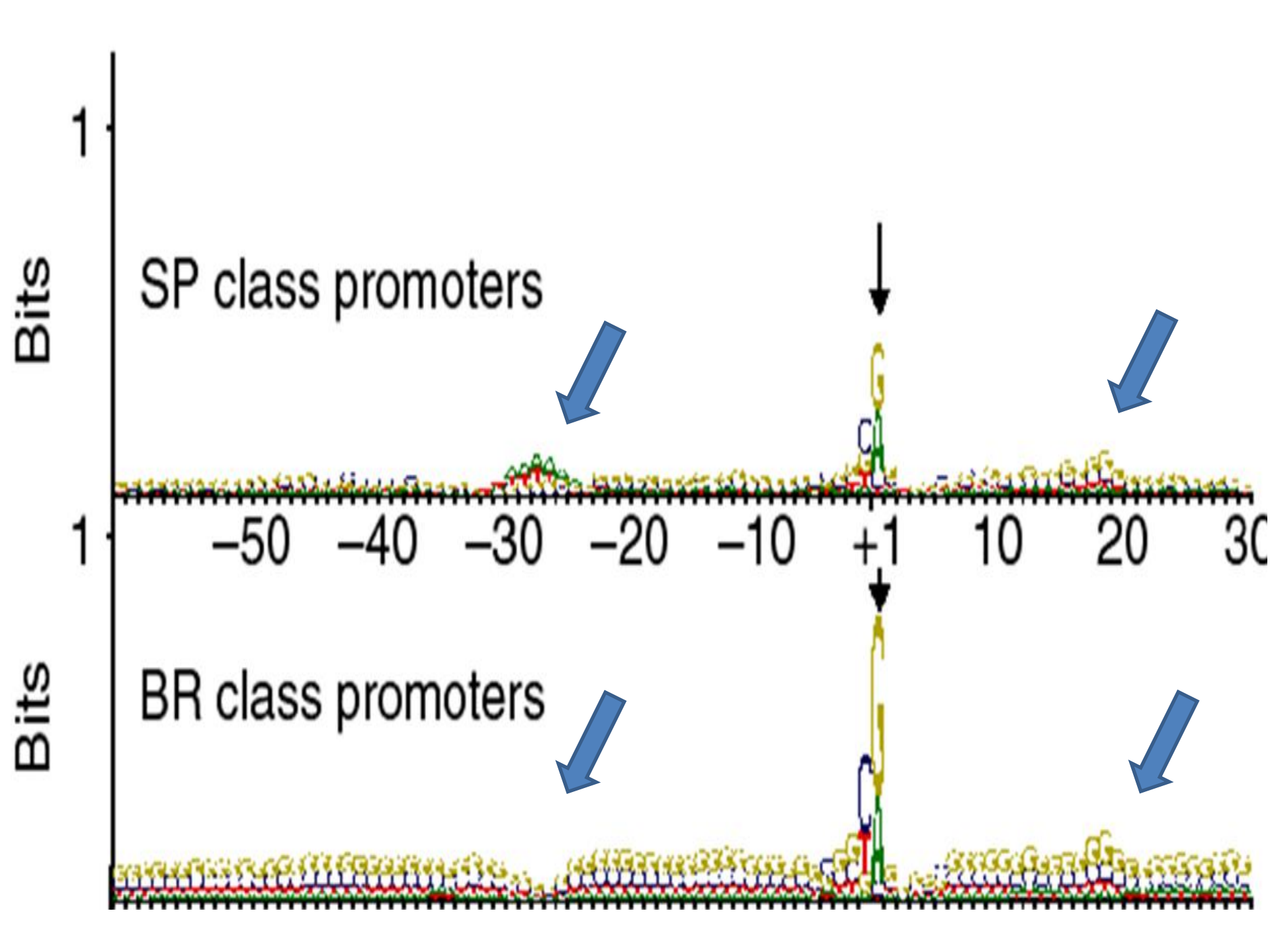
Alignment of thousands Promoters has allowed appreciation of strict geometrical constraint in the position of Promoter elements, like TATA-box, Initiator (INR), downstream promoter element (DPE)

(*that are recognized by different subunits of TFIID*)

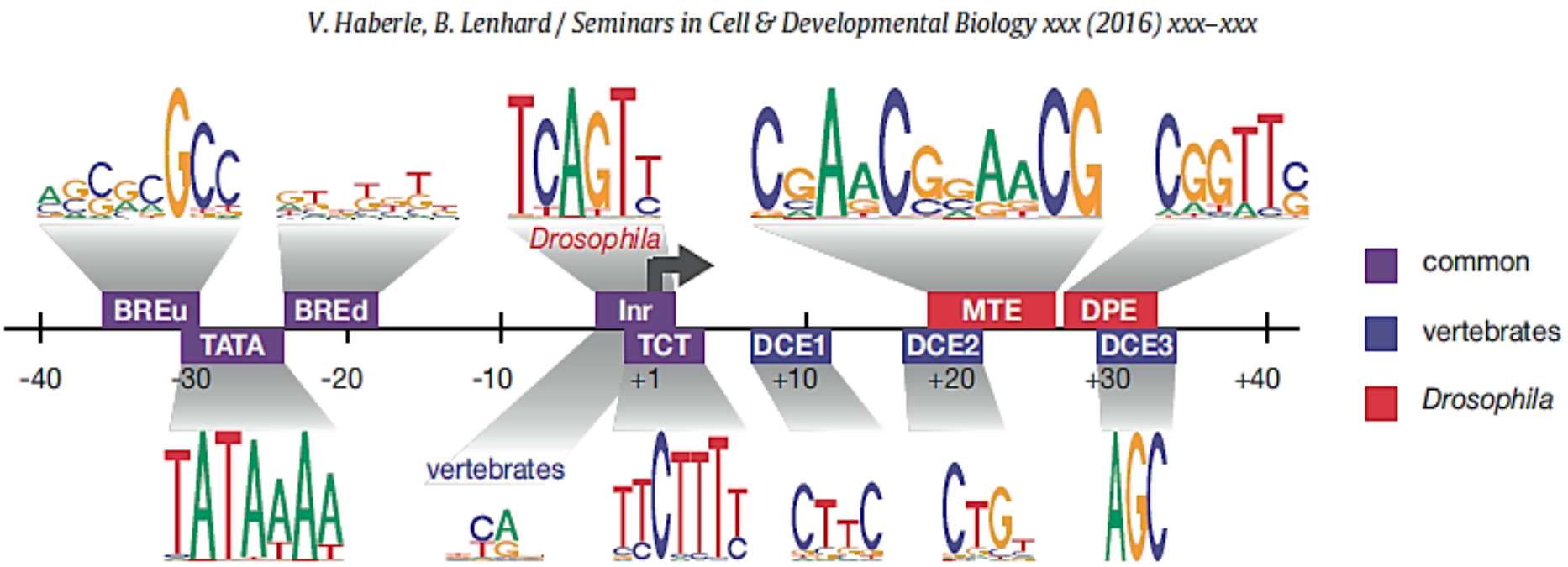


Sandelin et al., 2007

← Figure legend


Bits
1
SP class promoters
1
−50
−40
−30
−20
−10
+1
10
20
30
Bits
BR class promoters

Basic promoter elements in insects and Vertebrates are positionally constrained



Associated sequence logos are based on motifs from described for Drosophila and motifs from the JASPAR database for Vertebrates.
The initiator motif (Inr). BRE, TFIIB recognition element; DCE, downstream core element; DPE, downstream promoter element; Inr, initiator; MTE, motif ten element; TATA, TATA-box element; TCT, TCT initiator.
IMPORTANT: hardly any real promoter contains all or even most of the above elements – on the contrary, different elements are associated with different promoter architectures and their co-occurrence in individual promoters are strongly underrepresented compared to chance.

These elements are assorted in various combinations in vertebrate promoters, with some rules (e.g. INR associates with either a TATA or a DPE, very rarely all together)

BRE (TFIIB response element) was identified essentially by cristallography, very weak consensus

Different promoters exist that are recognized by sets of different proteins (e.g. TFR instead of TBP)(*see Levine's Textbook*)

Some promoters have «mixed sequences» that are recognized by different sets of proteins (on a developmental or tissue-specific basis)(*see Levine's Textbook*).