L4.2

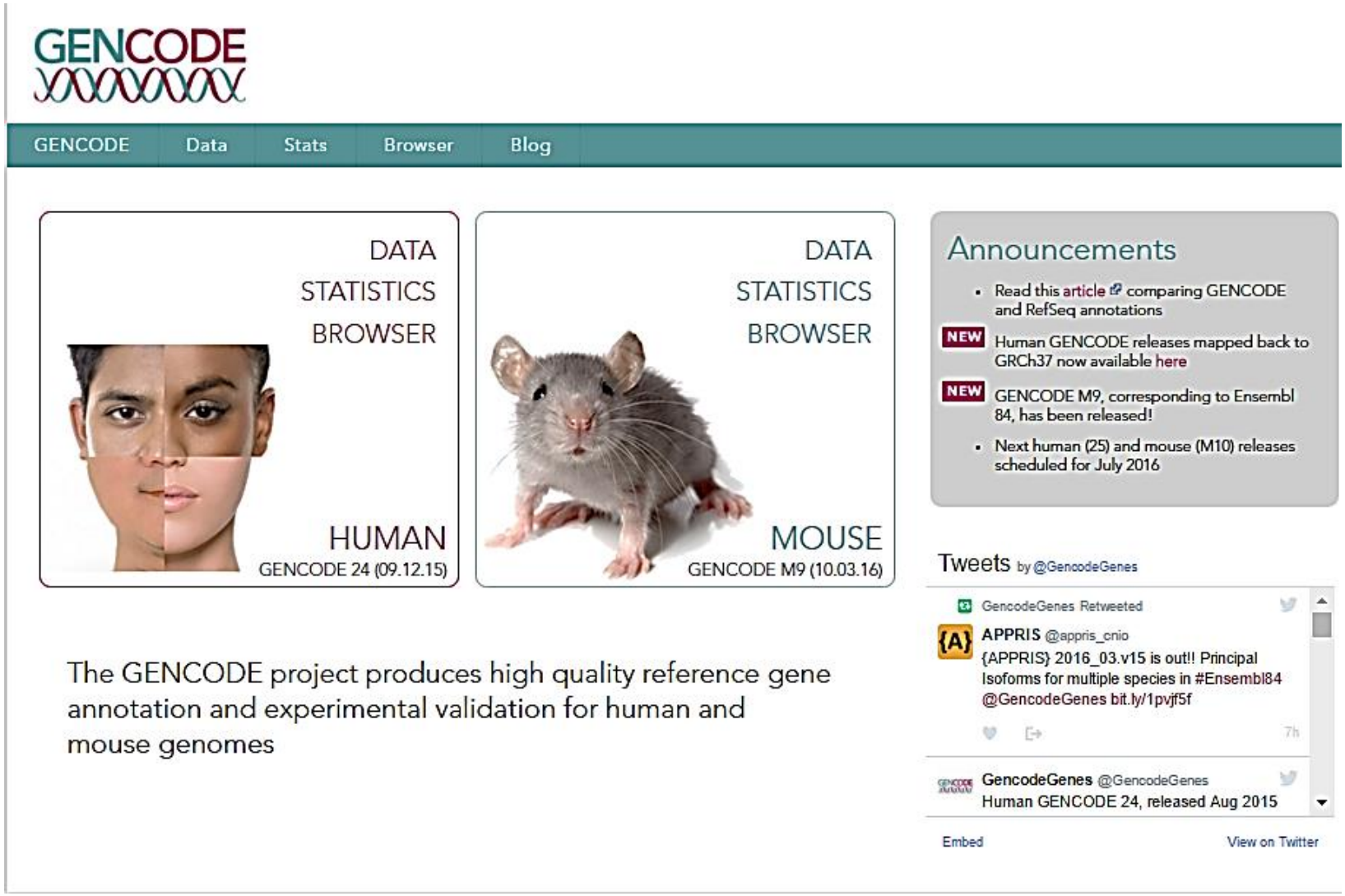
L4.2

Agenda

- other transcripts from the catalogue
- what is a gene today ?
- RNA Polymerase pausing
- extensive AS of lncRNAs
- mechanisms of exon splicing
- *cis*-regulatory sequences of splicing and AS
- trans-acting factors for AS (RBP)
- tissue-specific AS regulators

An RNA catalogue

It is worth spending few minutes on the Statistics to consider how many different types of long- and short-noncoding RNA have been catalogued



http://www.gencodegenes.org/

RNA Biotypes

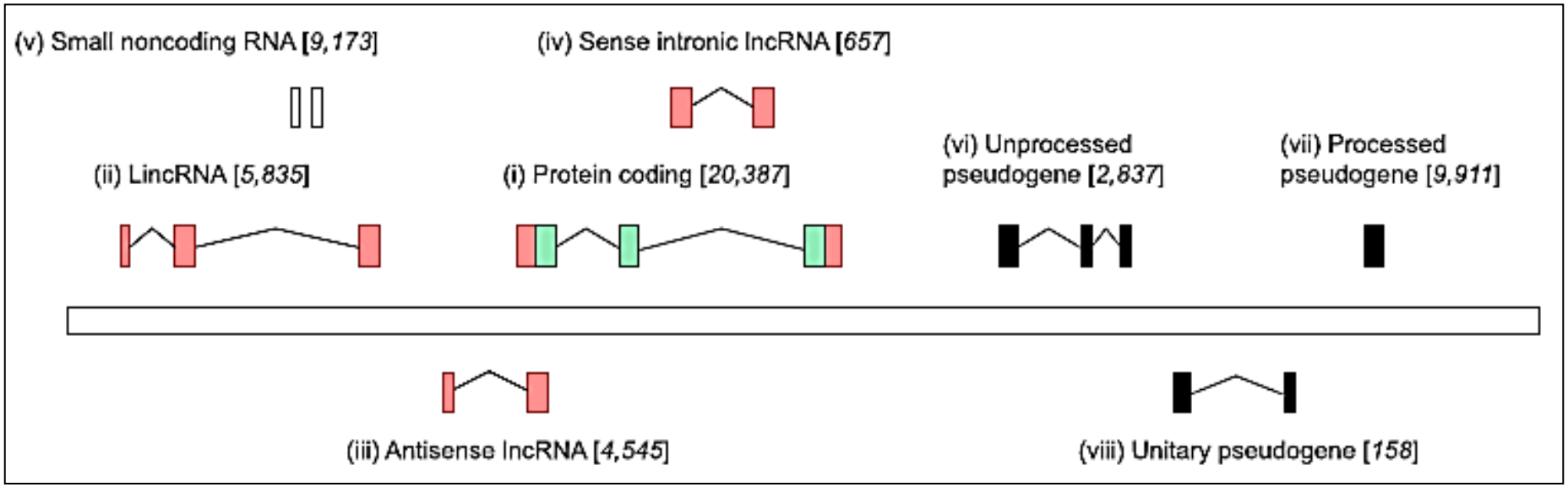


Figure 2. A summary of locus biotypes in GENCODE.

From Mudge et al., 2013

Perspective

Functional transcriptomics in the post-ENCODE era

Jonathan M. Mudge,[1] Adam Frankish, and Jennifer Harrow

Department of Informatics, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, United Kingdom

The last decade has seen tremendous effort committed to the annotation of the human genome sequence, most notably perhaps in the form of the ENCODE project. One of the major findings of ENCODE, and other genome analysis projects, is that the human transcriptome is far larger and more complex than previously thought. This complexity manifests, for example, as alternative splicing within protein-coding genes, as well as in the discovery of thousands of long noncoding RNAs. It is also possible that significant numbers of human transcripts have not yet been described by annotation projects, while existing transcript models are frequently incomplete. The question as to what proportion of this complexity is truly functional remains open, however, and this ambiguity presents a serious challenge to genome scientists. In this article, we will discuss the current state of human transcriptome annotation, drawing on our experience gained in generating the GENCODE gene annotation set. We highlight the gaps in our knowledge of transcript functionality that remain, and consider the potential computational and experimental strategies that can be used to help close them. We propose that an understanding of the true overlap between transcriptional complexity and functionality will not be gained in the short term. However, significant steps toward obtaining this knowledge can now be taken by using an integrated strategy, combining all of the experimental resources at our disposal.

23:1961–1973 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/13; www.genome.org

«Classical» versus modern view of a gene

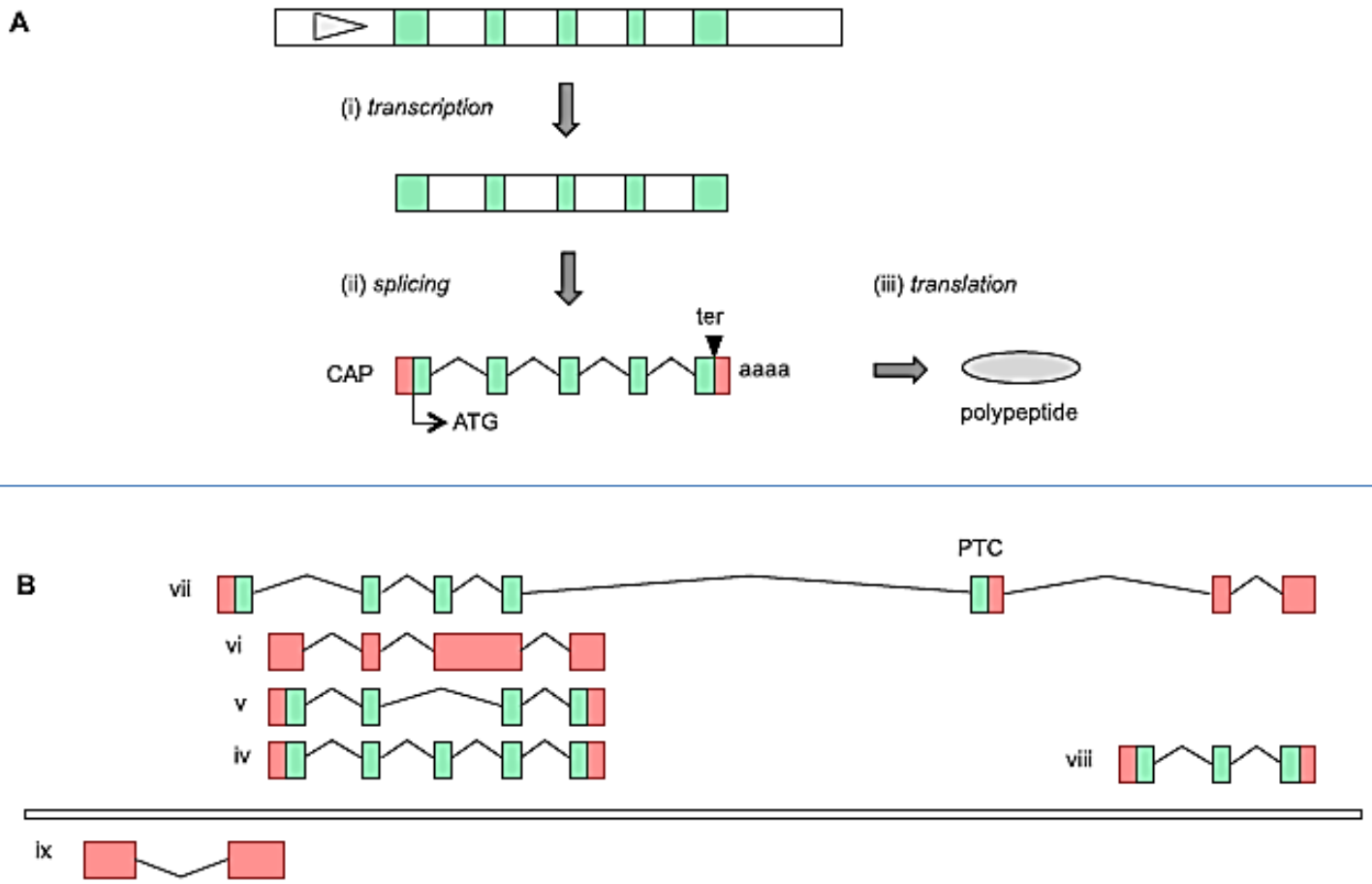
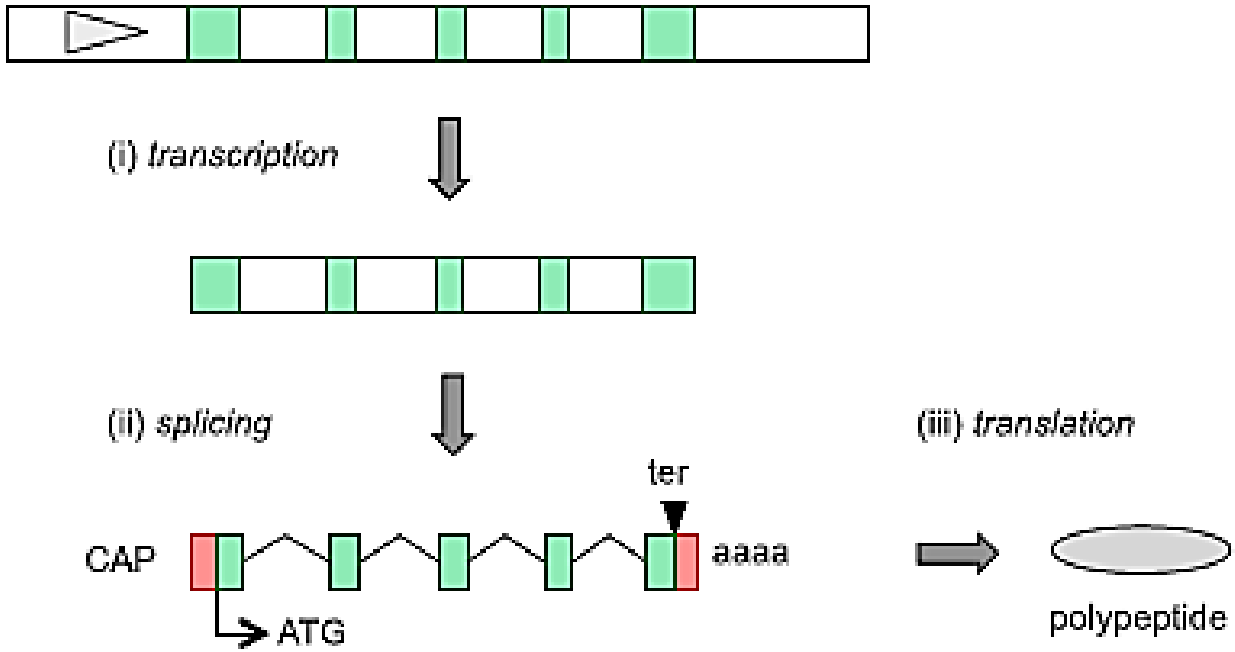


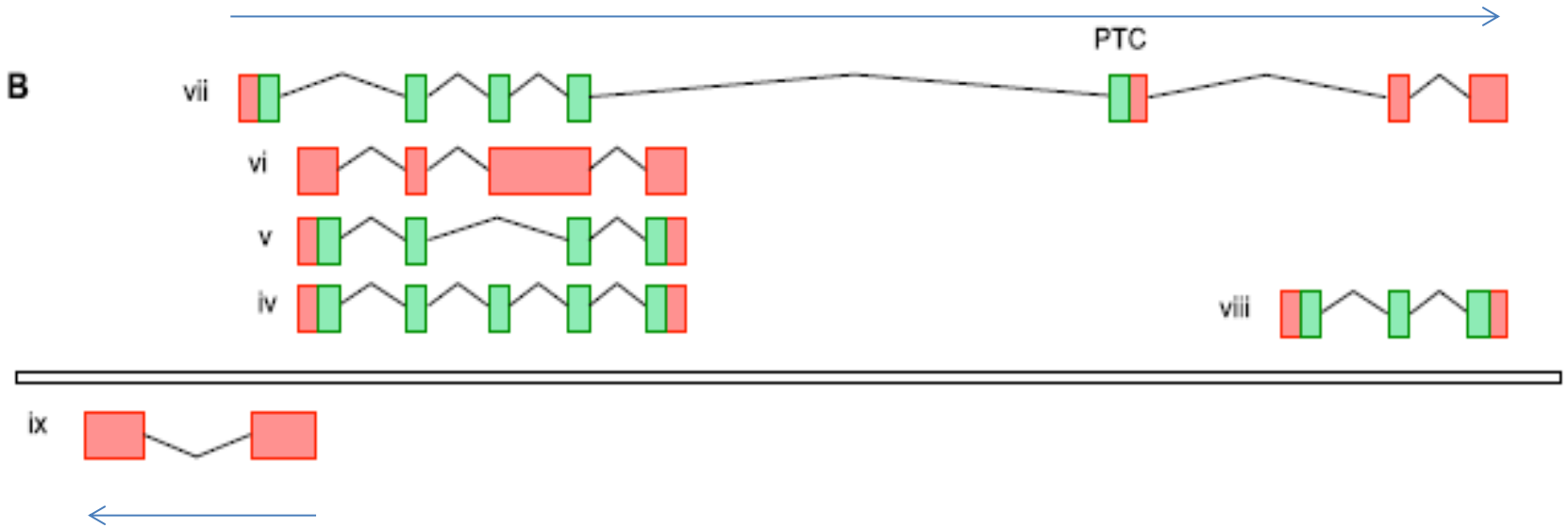
Figure 1. The evolving dogma of gene transcription.

From Mudge et al., 2013



(A) The **historical "central dogma"** of molecular biology. By this model, (i) transcription generates the primary transcript (exons in green, introns in white), with the initial interaction between the RNA polymerase complex and the genome being mediated by a promoter region (gray triangle). (ii) The introns of the primary transcript are removed by the spliceosome, and a mature mRNA is generated by 5' end capping (CAP) and polyadenylation (aaaa) (coding region [CDS] shown in green, untranslated 5' and 3' UTRs in red). (iii) The mRNA is translated into a polypeptide by the ribosome complex, with translation proceeding from the initiation codon (ATG) and ending at the termination codon (ter).

From Mudge et al., 2013



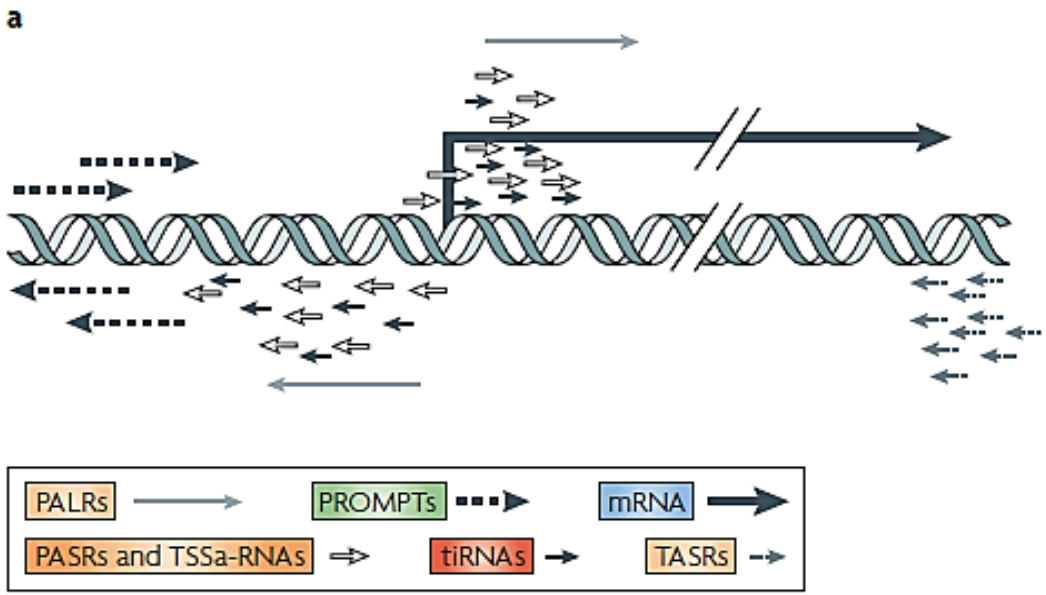
(B) An **updated model** reflecting a modern view of transcriptional complexity. Here, the same gene (iv) undergoes alternative splicing (AS), for example an exon skipping event that does not change the frame of the CDS (v); this event thus has the potential to generate an alternative protein isoform. However, products of AS cannot be assumed to be functional; this gene has generated a retained intron transcript (vi), perhaps due to the failure of the spliceosome to remove this intron. Further complexity comes from a read-through transcription event (vii), whereby a transcript is generated that also includes exons from a neighboring protein-coding locus (viii). In this example, the read-through transcript has an alternative first exon compared with the upstream gene that contains a potential alternative ATG codon, although the presence of a subsequent premature termination codon (PTC) prior to two splice junctions indicates that this transcript is likely subjected to the nonsense mediated decay (NMD) degradation pathway. Finally, model ix is a transcript that is antisense to the upstream gene; both loci are potentially generated under the control of a bidirectional promoter.

From Mudge et al., 2013

in addition....

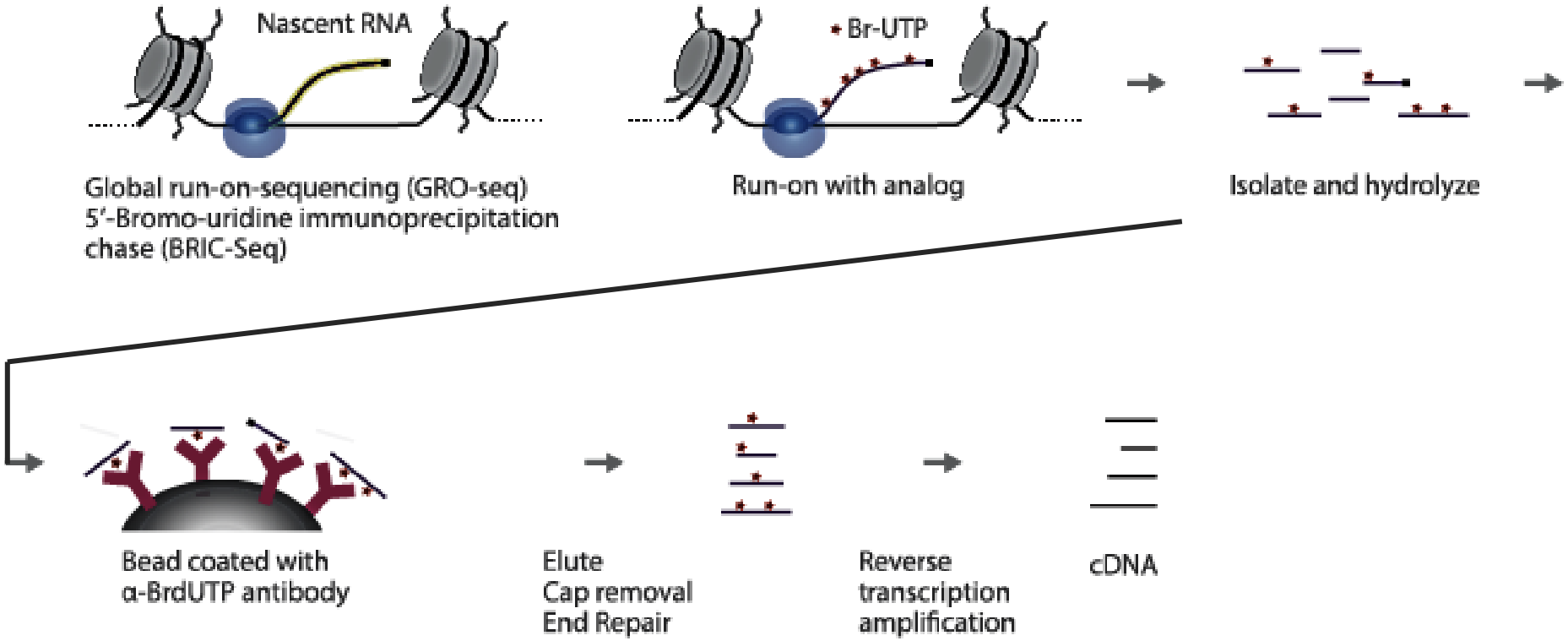
Using different technologies, including tiling microarrays, GRO-seq, CAGE, Sage and others, unstable short RNAs were also observed close to promoters

Unstable small RNA accompanying gene transcription



Short name of RNA classes	Full name of RNA classes	
PALRs	Promoter-associated long RNAs	Hundreds nt long RNAs spanning regions on proximal promoters to the first exon
PASRs	Promoter-associated short RNAs	20–70 nt long RNAs spanning regions around core promoters
TASRs	Termini-associated short RNAs	20–70 nt long RNAs spanning regions around transcription termination sites
PROMPTs	Promoter upstream transcripts	Unstable transcripts mapping 0.5–2 kb upstream the transcription starting sites
TSSa-RNAs	Transcription start sites antisense RNAs	RNAs, generally short and non-coding, generated from bidirectional activity of mammalian RNA Polymerase II
NRO-RNAs	Nuclear run-on assay derived RNAs	Short RNA detected by nuclear run-on assays, mapping 20 to 50 downstream to transcriptions starting sites of mRNAs
RE RNAs	Retrotransposon-derived RNAs	A heterogeneous class of RNAs which starting sites overlap retrotransposon elements
tiRNAs	Tiny transcription initiation RNAs	RNAs about 18 nt long, positioned about 20 bp after the transcription starting sites of highly expressed mRNAs

GRO-Seq



from the Illumina website

As secondary product, GRO-seq localize active Polymerase

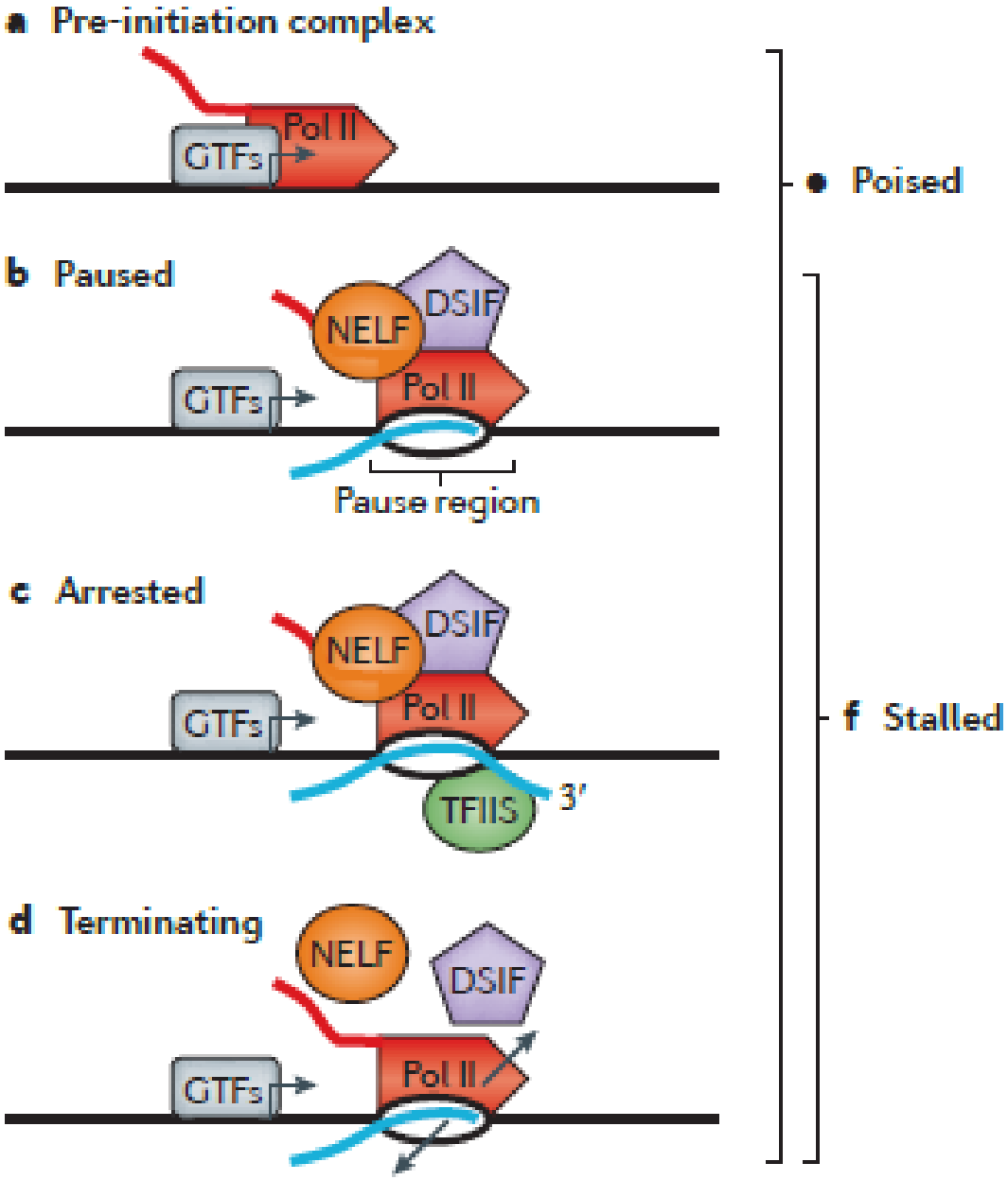
promoter-proximal pausing

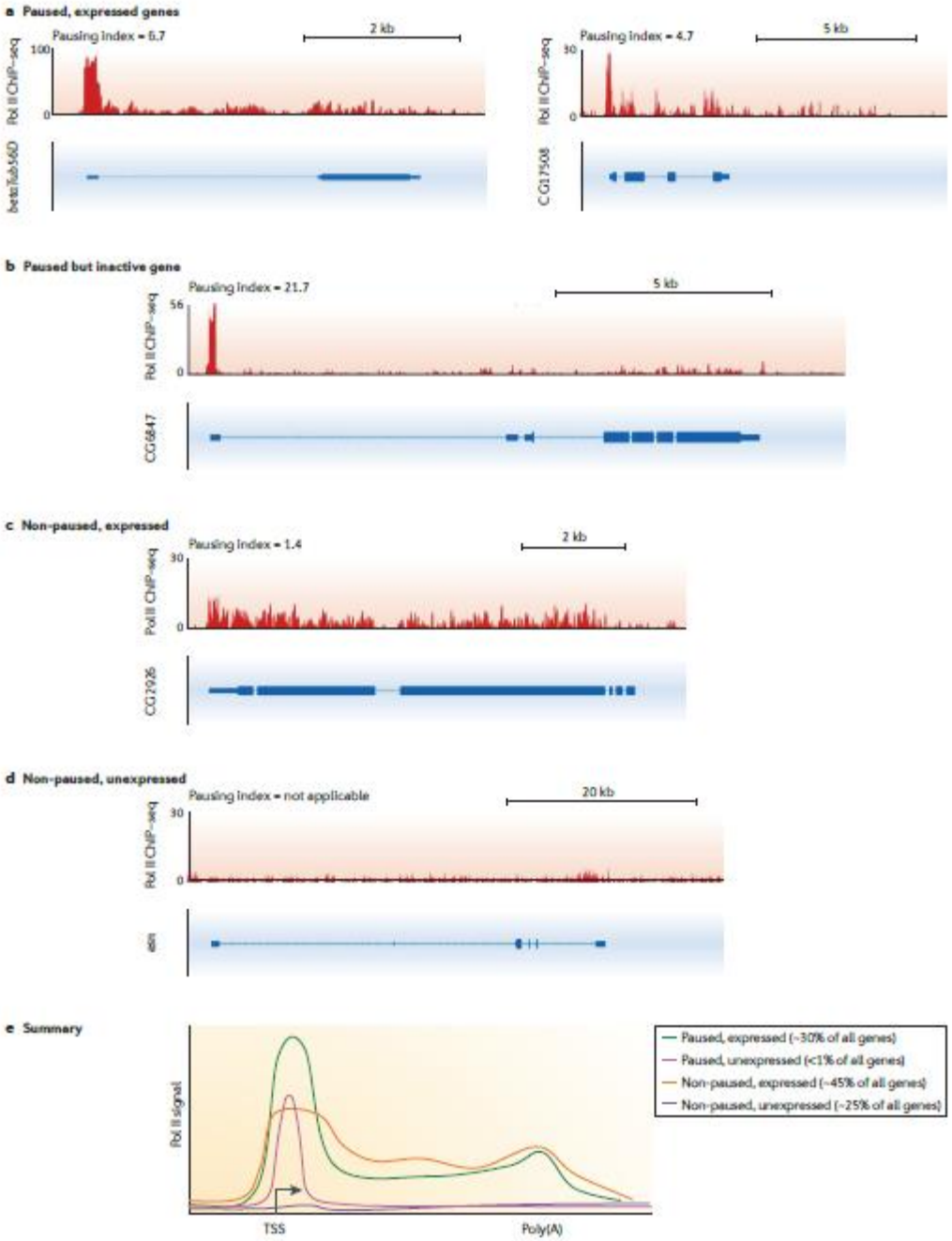
MODES OF TRANSCRIPTIONAL REGULATION

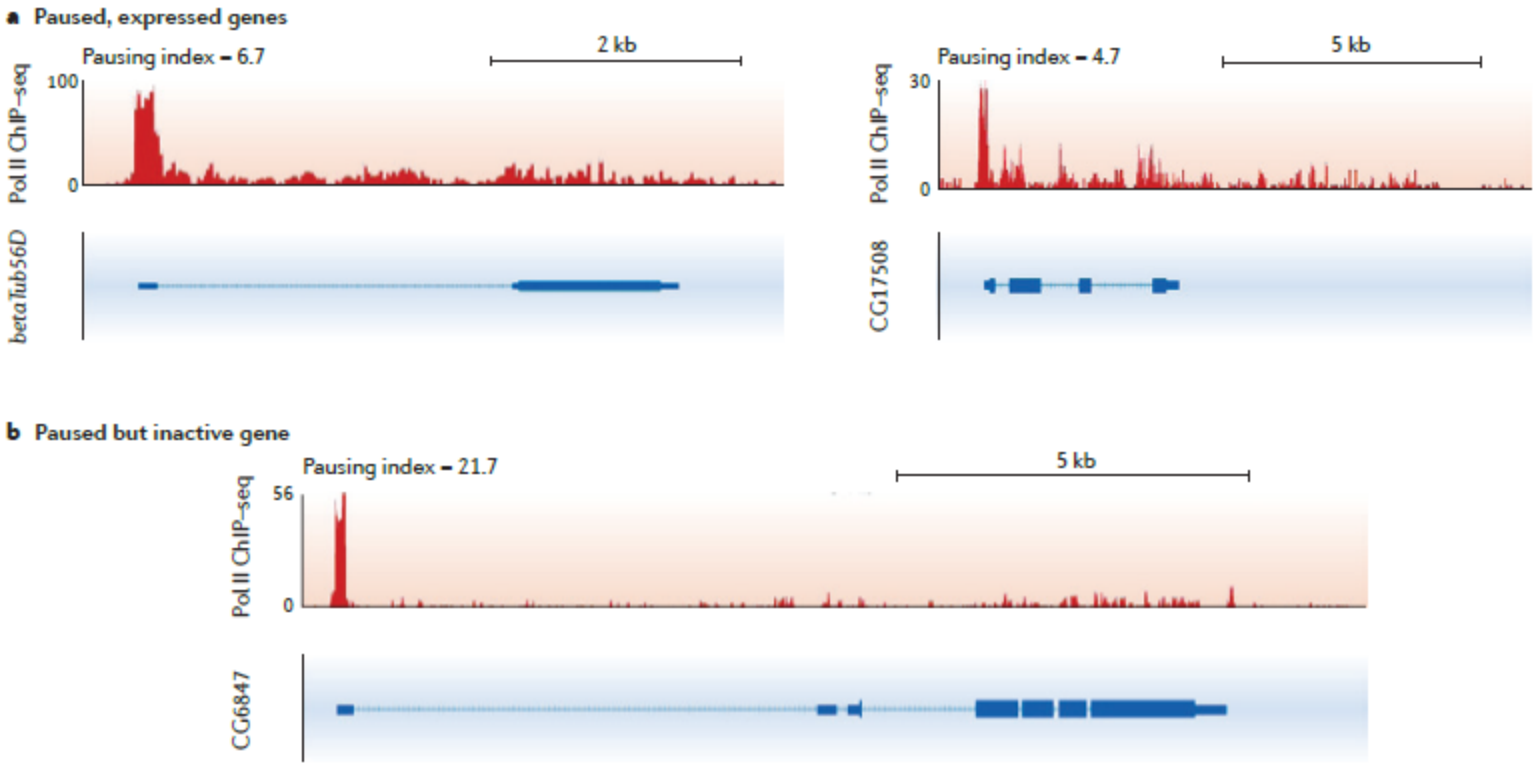
Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans

Karen Adelman[1] and John T. Lis[2]

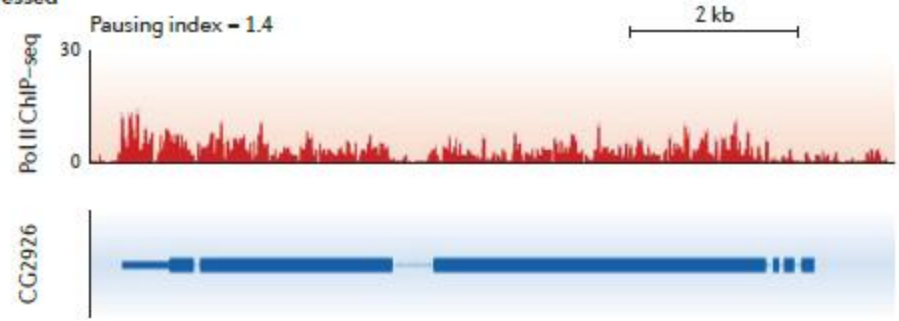
Abstract | Recent years have witnessed a sea change in our understanding of transcription regulation: whereas traditional models focused solely on the events that brought RNA polymerase II (Pol II) to a gene promoter to initiate RNA synthesis, emerging evidence points to the pausing of Pol II during early elongation as a widespread regulatory mechanism in higher eukaryotes. Current data indicate that pausing is particularly enriched at genes in signal-responsive pathways. Here the evidence for pausing of Pol II from recent high-throughput studies will be discussed, as well as the potential interconnected functions of promoter-proximally paused Pol II.


a Pre-initiation complex
GTFs
Pol II
b Paused
NELF
DSIF
GTFs
Pol II
Pause region
c Arrested
NELF
DSIF
GTFs
Pol II
TFIIS
3′
d Terminating
NELF
DSIF
GTFs
Pol II
Poised
f Stalled

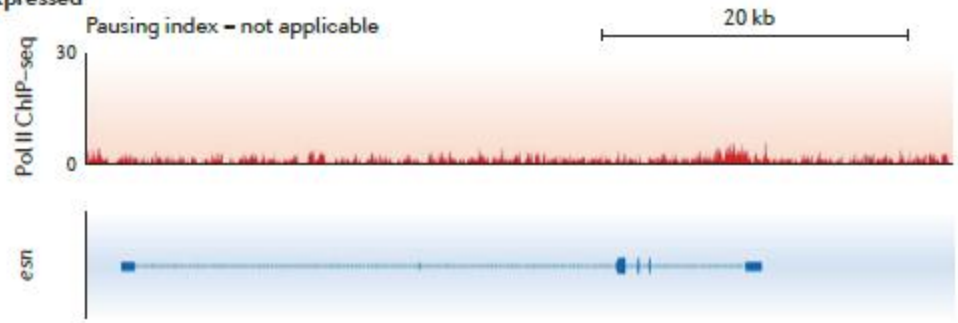

a Paused, expressed genes
Pausing Index = 6.7
2 kb
Pol II ChIP-seq
100
0
betaTub56D
Pausing Index = 4.7
5 kb
30
0
CG17508
b Paused but inactive gene
Pausing Index = 21.7
5 kb
56
0
CG6847
c Non-paused, expressed
Pausing Index = 1.4
2 kb
30
0
CG1926
d Non-paused, unexpressed
Pausing Index = not applicable
20 kb
30
0
elm
e Summary
Pol II signal
TSS
Poly(A)
Paused, expressed (~30% of all genes)
Paused, unexpressed (<1% of all genes)
Non-paused, expressed (~45% of all genes)
Non-paused, unexpressed (~25% of all genes)


a Paused, expressed genes
Pausing index = 6.7
2 kb
Pol II ChIP–seq
100
0
beta Tub56D
Pausing index = 4.7
5 kb
Pol II ChIP–seq
30
0
CG17508
b Paused but inactive gene
Pausing index = 21.7
5 kb
Pol II ChIP–seq
56
0
CG6847

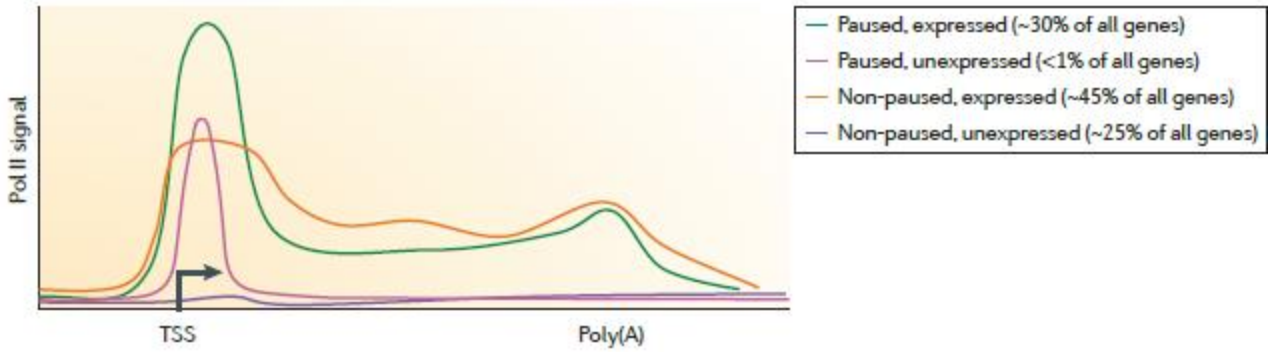
c Non-paused, expressed

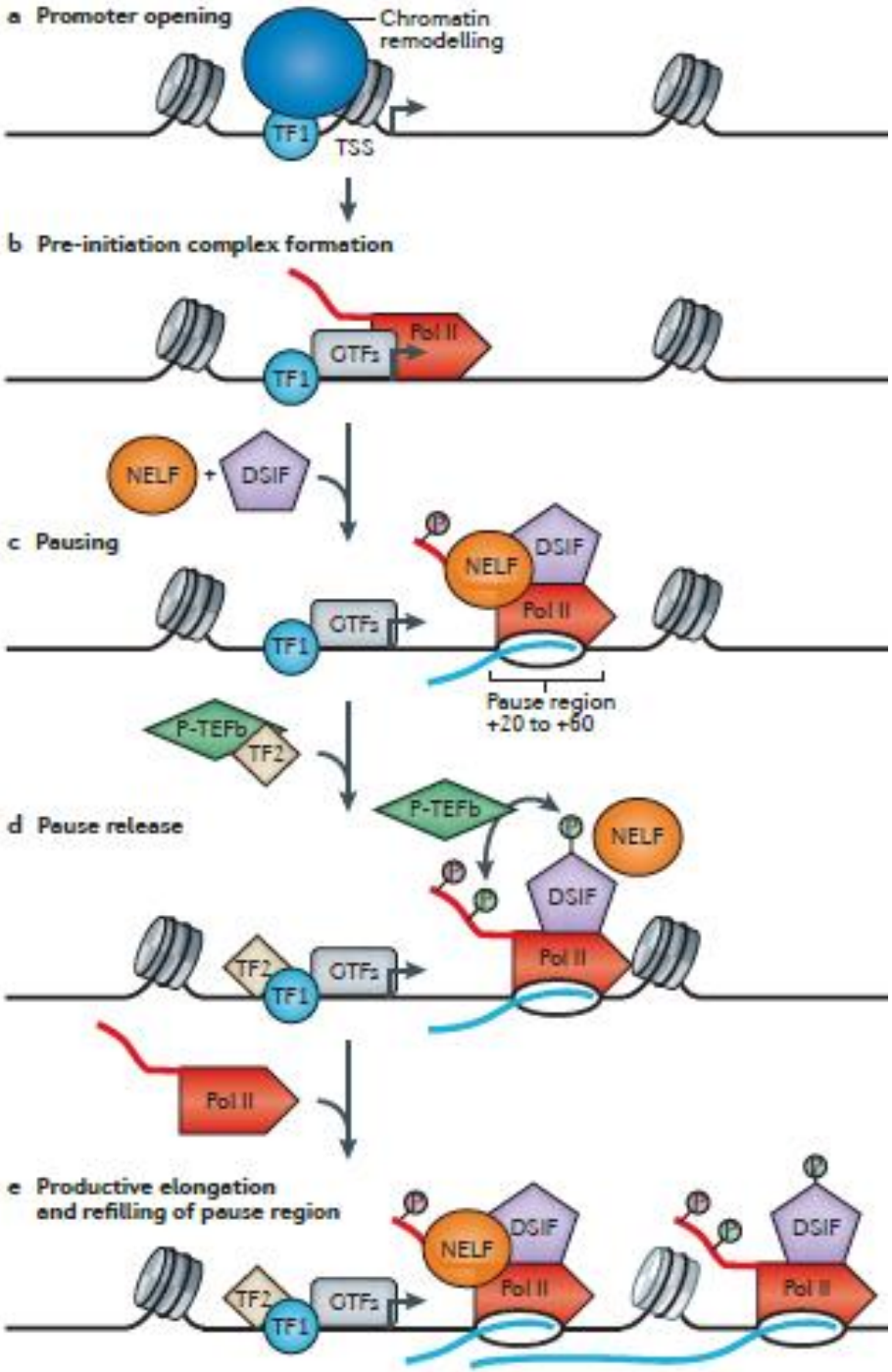


d Non-paused, unexpressed



e Summary

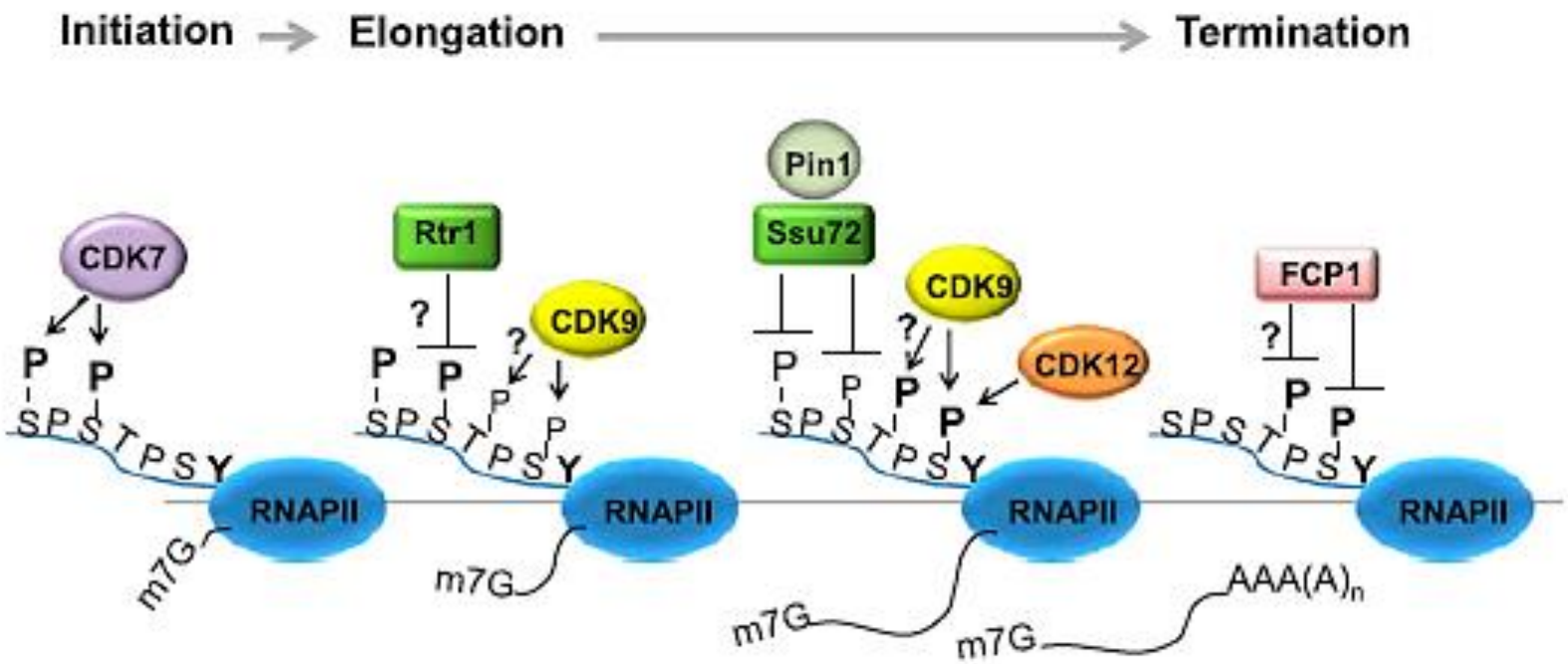




CTD phosphorylation in Ser-2 leads to SETD2 association. SET2D is the lysine Metyl Transferase specific for H3K36 methylation

Phosphorylation of DSIF leads to NELF dissociation, while DSIF is converted to elongation factor

RNA Polymerase cycle and CTD phosphorylation

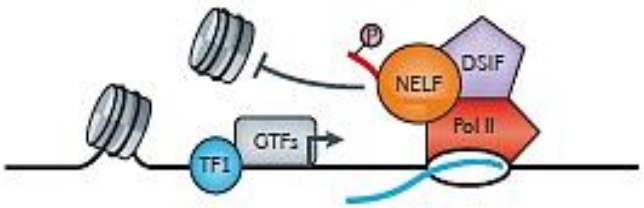


Dynamic modification of the CTD during the transcription cycle.

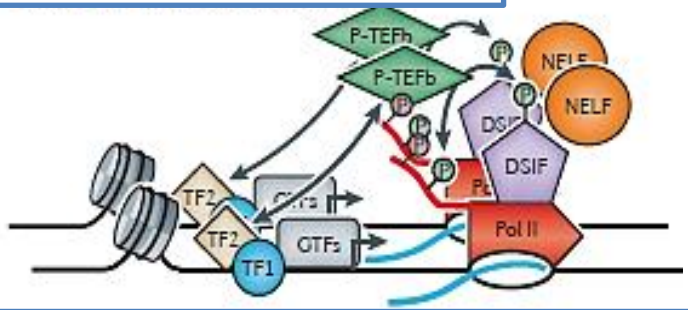
From Hsin et al., Genes Dev (2011) 26:2119–2137

Possibe functions of PolII pausing

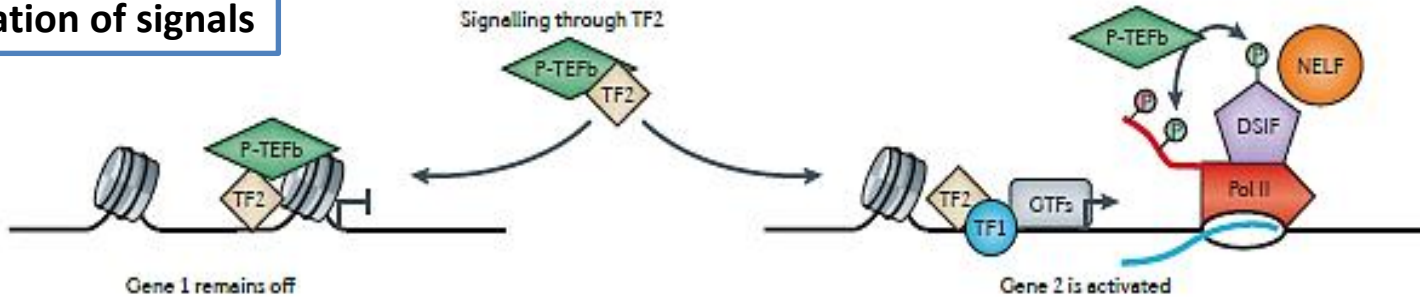
establishing permissive chromatin



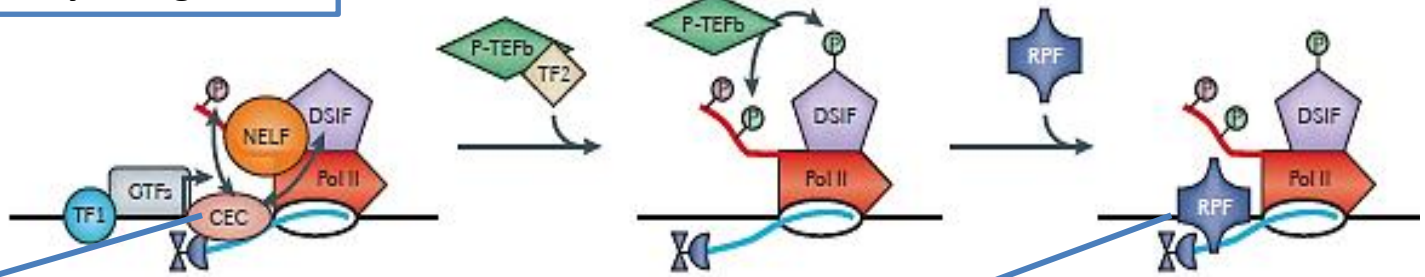
synchronous rapid activation



integration of signals



checkpoint in early elongation ?



Capping Enzyme Complex

RNA Processing Factors

GENCODE

Human | Mouse | How to access data | FAQ | Documentation | About us

Human

Statistics about the current GENCODE Release (version 30)

The statistics derive from the gtf file that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the README_stats.txt file.

General stats

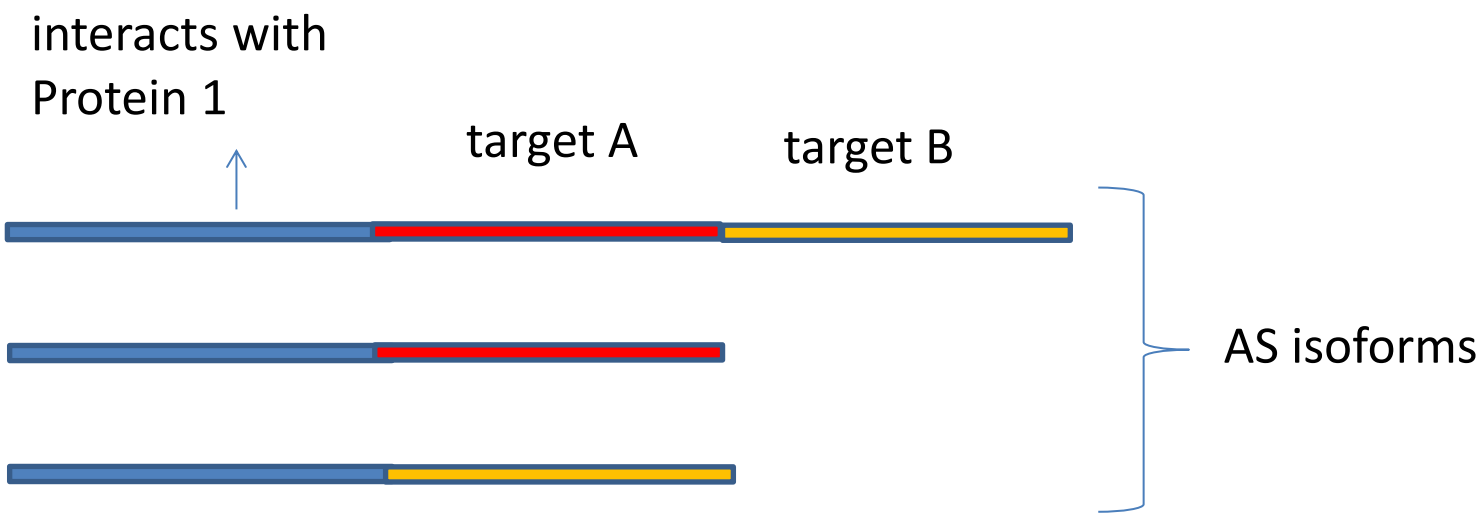
Total No of Genes	58870	Total No of Transcripts	208621
Protein-coding genes	19986	Protein-coding transcripts	83688
Long non-coding RNA genes	16193	- full length protein-coding	57687
Small non-coding RNA genes	7576	- partial length protein-coding	26001
Pseudogenes	14706	Nonsense mediated decay transcripts	15550
- processed pseudogenes	10663	Long non-coding RNA loci transcripts	30369
- unprocessed pseudogenes	3525		
- unitary pseudogenes	221		
- polymorphic pseudogenes	42		
- pseudogenes	18	Total No of distinct translations	61870

LncRNA undergo Alternative Splicing

They are capped and polyadenylated

What is the sense of making AS ?

Their function can be modulated by including/excluding certain parts.



Alternative Splicing of lncRNAs is guided by the same elements as protein-coding RNAs

However, while in protein-coding RNA alternative exons are few (average one-two on an average of 9 exons), lncRNA tend to have more alternatives.

Note that lncRNAs do not have the constraint of the coding sequence.

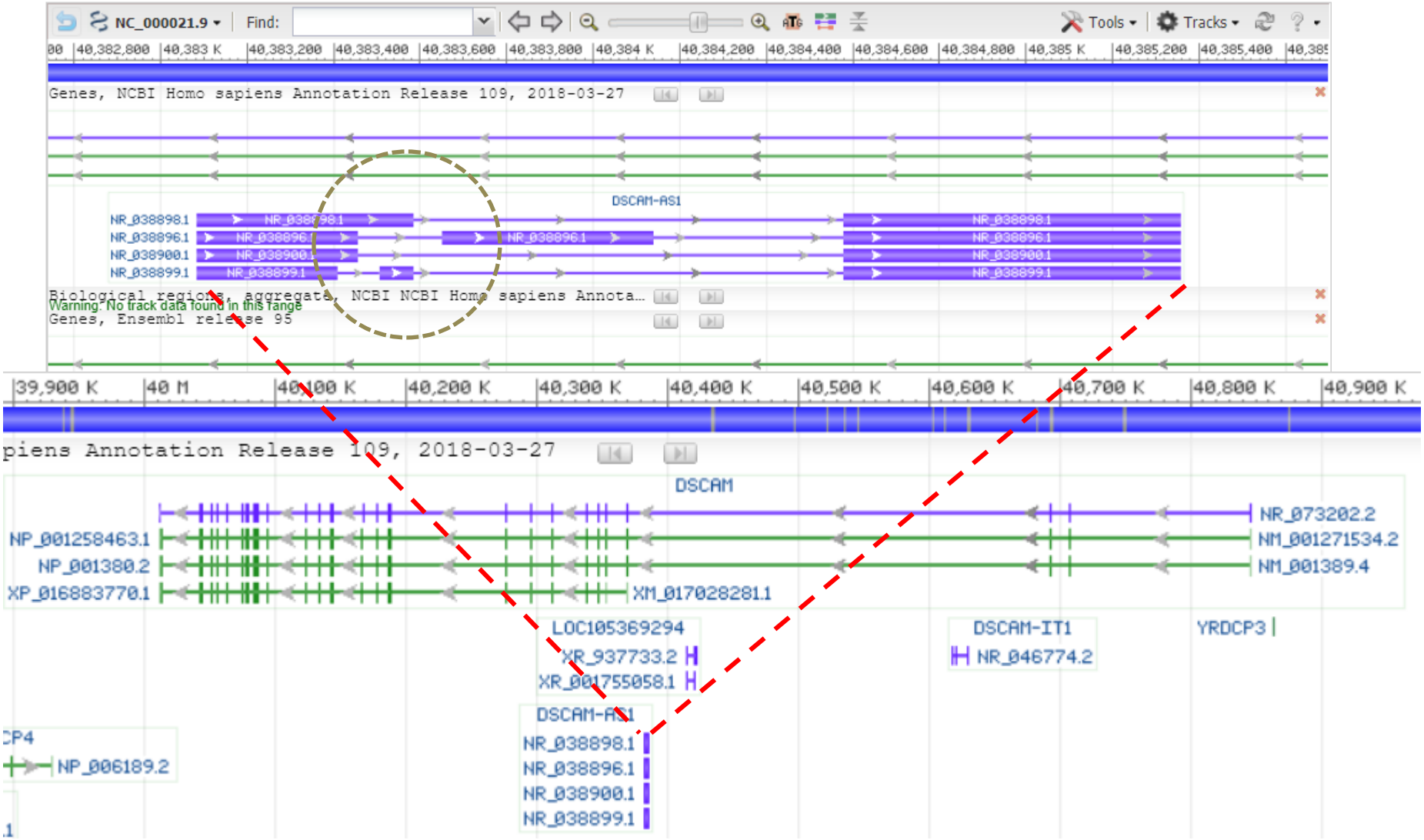
DSCAM-AS1

Genomic regions, transcripts, and products

Go to reference sequence details

Genomic Sequence: NC_000021.9 Chromosome 21 Reference GRCh38.p12 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank



Back to splicing

What do we know of splicing ?

When, where, how ?

Co-transcriptional or post-transcriptional ?

This is a long-standing, wearying discussion on this, lasting at least 25 years

let’s see how ENCODE has approached it

The transcriptome of nuclear subcompartments

For the K562 cell line, we also analysed RNA isolated from three subnuclear compartments (chromatin, nucleolus and nucleoplasm.

Almost half (18,330) of the GENCODE (v7) annotated genes detected for all 15 cell lines (35,494) were identified in the analysis of just these three nuclear subcompartments. In addition, there were as many novel unannotated genes found in K562 subcompartments as there were in all other data sets combined.

Djebali et al., 20120

The interrogation of different subcellular RNA fractions provides snapshots of the status of the RNA population along the RNA processing pathway.

Thus, by analysing short and long RNAs in the different subcellular compartments, we confirm that splicing predominantly occurs during transcription.

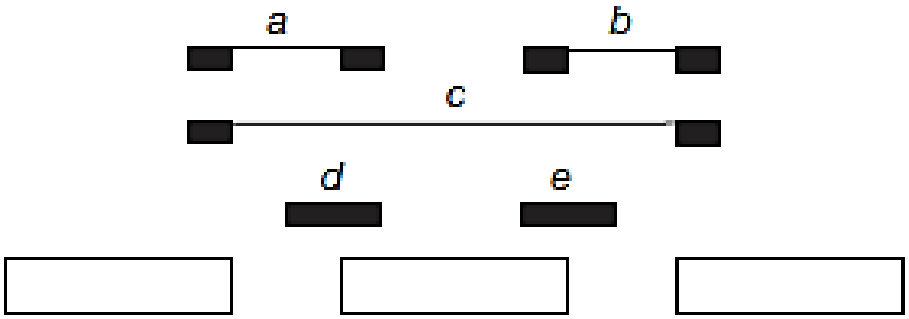
By using RNA-seq to measure the degree of completion of splicing (Fig. 2a), we observed that around most exons, introns are already being spliced in chromatin-associated RNA—the fraction that includes RNAs in the process of being transcribed (Fig. 2b).

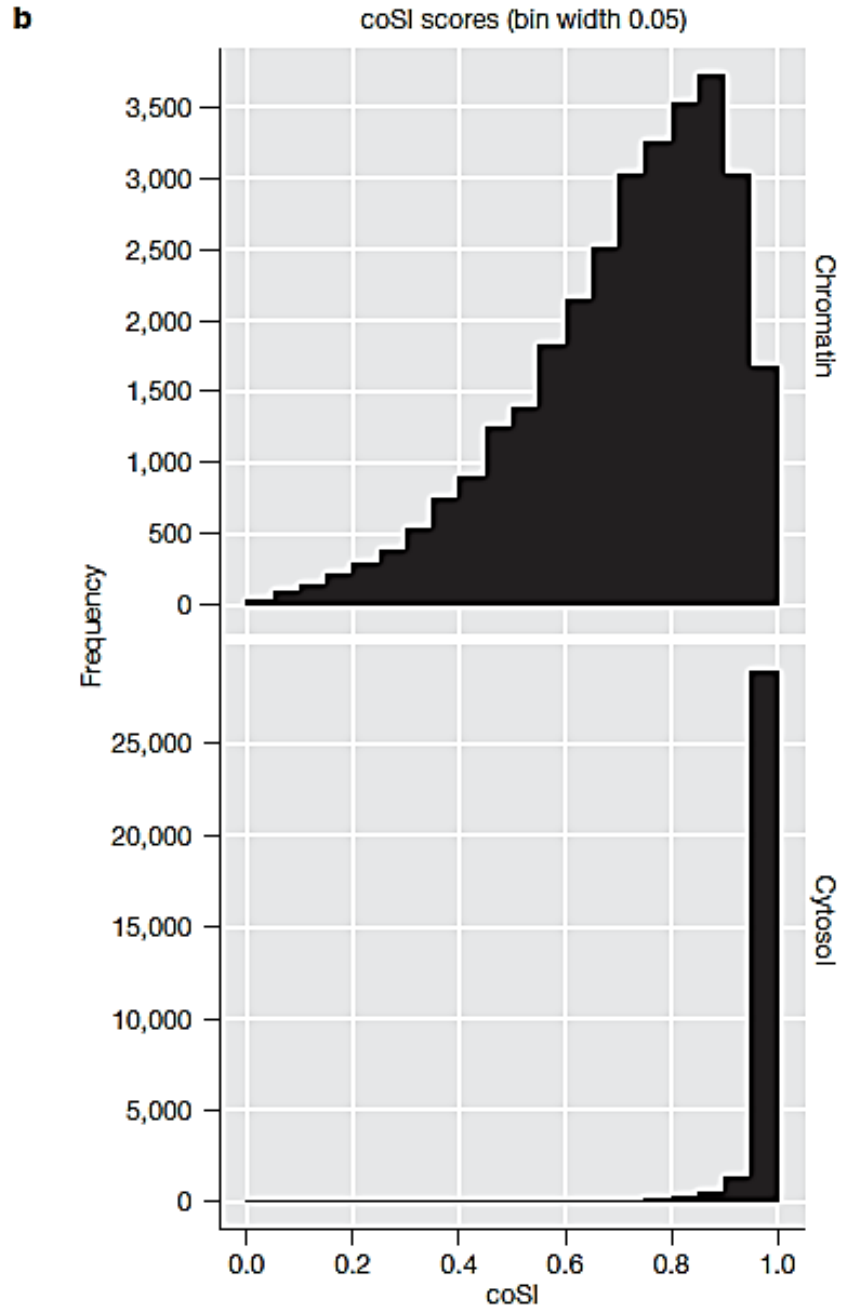
Concomitantly, we found strong enrichment specifically of spliceosomal small nuclear RNAs (snRNAs) in this RNA fraction

Co-transcriptional splicing provides an explanation for the increasing evidence connecting chromatin structure to splicing regulation, and we have observed that exons in the process of being spliced are enriched in a number of chromatin marks

Djebali et al., 20120

Figure 2
Co-transcriptional splicing evaluation





coSI = the ratio between junction reads and exon-intron reads

Djebali et al., 20120

AS a matter of fact your Textbook, written by Alberto Kornblhitt, one of the major scientists in the field, bases discussion non mechanisms on splicing being mostly co-transcriptional.

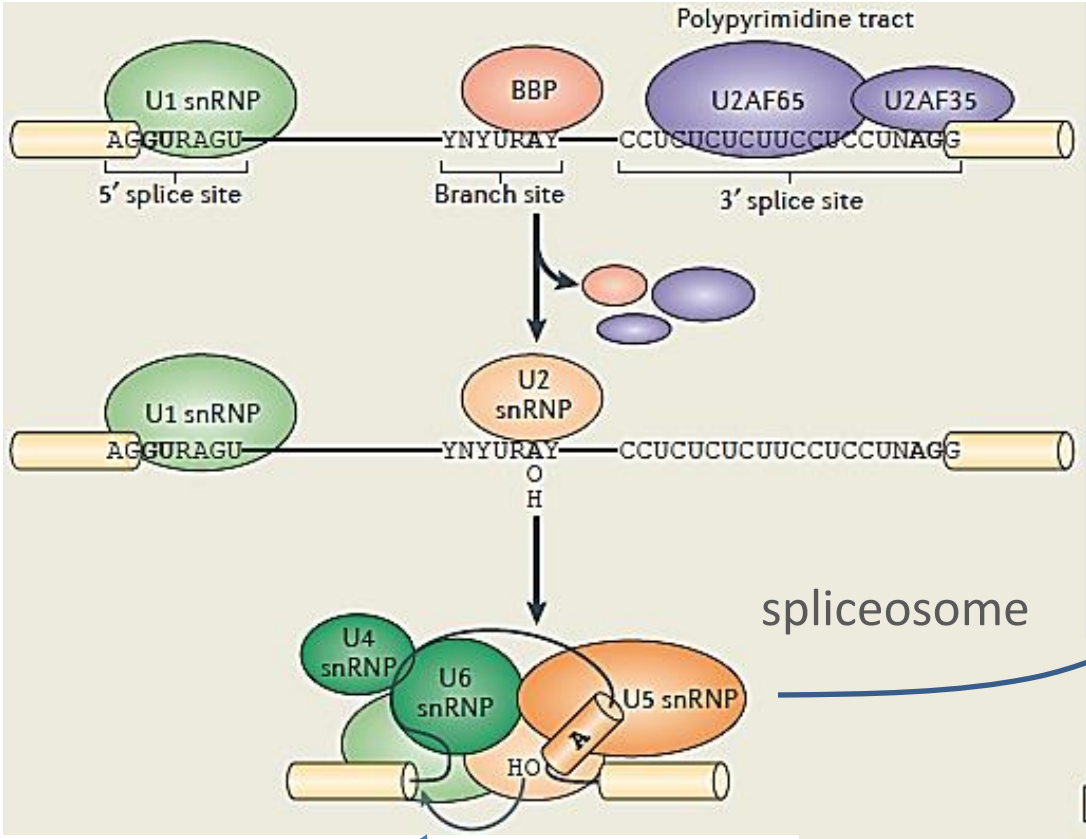
Note:
saying that splicing occurs co-transcriptionally does not mean that all catalytic events of splicing occur immediately as soon as the RNA progressively emerges from RNA PolII.

Textbook

Alternative splicing: a pivotal step between eukaryotic transcription and translation

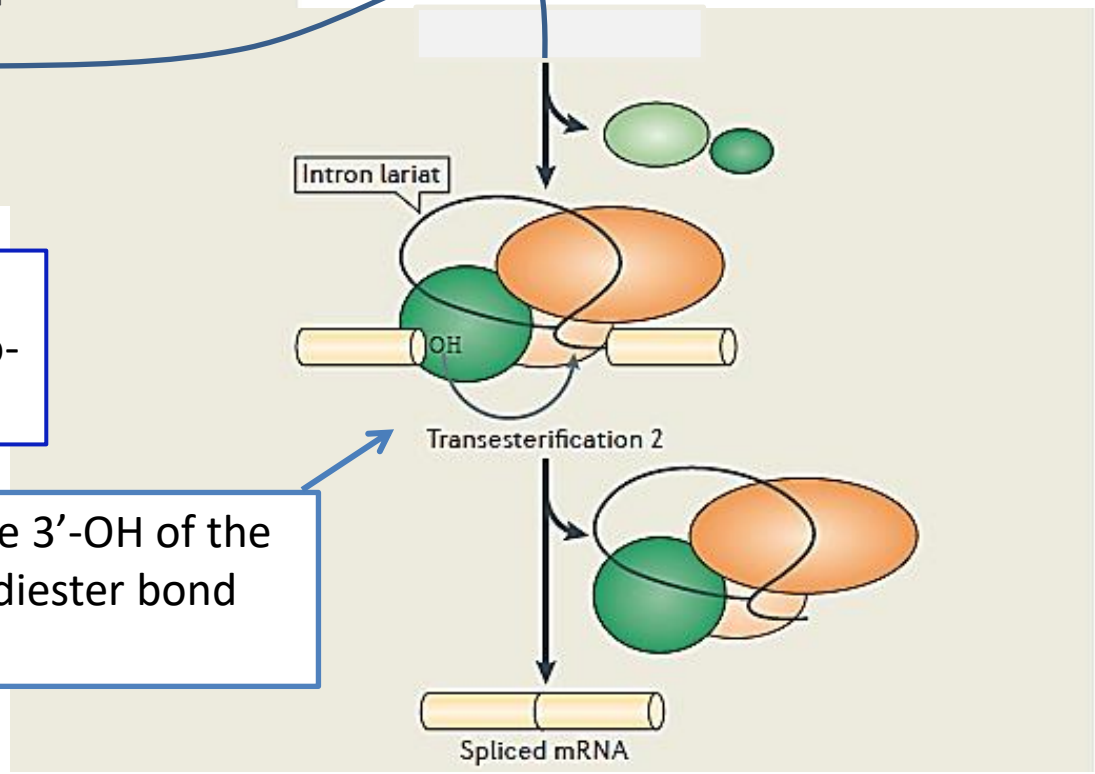
Alberto R. Kornblihtt, Ignacio E. Schor, Mariano Alló, Gwendal Dujardin, Ezequiel Petrillo and Manuel J. Muñoz

Abstract | Alternative splicing was discovered simultaneously with splicing over three decades ago. Since then, an enormous body of evidence has demonstrated the prevalence of alternative splicing in multicellular eukaryotes, its key roles in determining tissue- and species-specific differentiation patterns, the multiple post- and co-transcriptional regulatory mechanisms that control it, and its causal role in hereditary disease and cancer. The emerging evidence places alternative splicing in a central position in the flow of eukaryotic genetic information, between transcription and translation, in that it can respond not only to various signalling pathways that target the splicing machinery but also to transcription factors and chromatin structure.



exon intron exon

spliceosome



1st transesterification: the 2′-OH of the branching point «A» attacks the phospho-diester bond betwee exon and intron

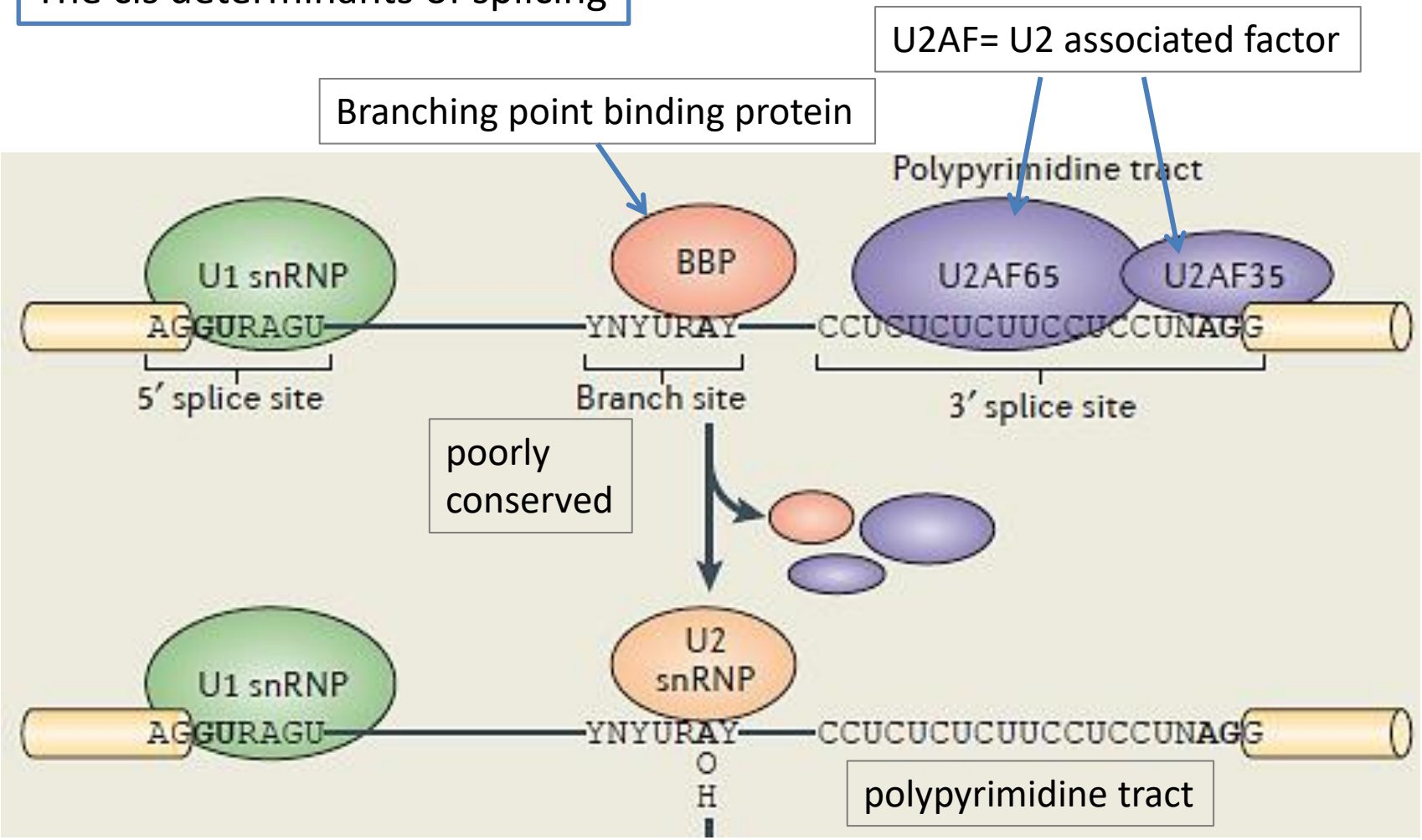
2nd transesterification: the 3′-OH of the exon attacks the phosphodiester bond betwee intron and exon

From: Van der Feltz et al., 2012

Table 1. Core Subunits of Human U snRNPs

U snRNP	subunit gene	common subunit name(s)[a]	molecular mass (kDa)[b]	% U snRNP	recognizable domain/ functional site[c]
U1 (248.1 kDa)	RNU1	U1 snRNA	53.5	21.6	
	SNRPB, -B2, -D1, -D2, -D3, -E, -F, -G	seven Sm proteins	94.3	38.0	Sm
	SNRNPA	U1-A	31.3	12.6	RRM
	SNRNP70	U1-70k	51.6	20.8	RRM; SR repeat
	SNRNPC	U1-C	17.4	7.0	Znf
U2 (987.4 kDa)	RNU2	U2 snRNA	61.2	6.2	
	SNRPB, -B2, -D1, -D2, -D3, -E, -F, -G	seven Sm proteins	94.3	9.6	Sm
	SNRPA1	U2A′	28.4	2.9	LRR
	SNRPB2	U2B″	25.4	2.6	RRM
	SF3A1	SF3a120	88.9	9.0	SWAP; UBQ domain
	SF3A2	SF3a66	49.3	5.0	Znf
	SF3A3	SF3a60	58.6	5.9	Znf; SAP
	SF3B1	SF3b155	145.8	14.8	HEAT repeat
	SF3B2	SF3b145	100.2	10.1	SAP
	SF3B3	SF3b130	135.5	13.7	DExH/D
	SF3B4	SF3b49	44.4	4.5	RRM
	SF3B5	SF3b10	10.1	1.0	
	SF3B14	SF3b14a; p14	14.6	1.5	RRM
	PHF5A	SF3b14b; Rds3	12.4	1.3	PHD-like
	DDX46	DDX46; hPrp5p	117.4	11.9	DExH/D; SR repeat
	SMNDC1	SPF30/SMNrp	26.7	2.7	Tudor domain
U5 (1055.7 kDa)	RNU5	U5 snRNA	37.6	3.6	
	SNRPB, -B2, -D1, -D2, -D3, -E, -F, -G	seven Sm proteins	94.3	8.9	Sm
	TXNL4A	U5-15K	16.9	1.6	TRX
	SNRNP40	U5-40K	39.3	3.7	WD40
	CD2BP2	U5-52K	37.6	3.6	GYF
	DDX23	U5-100K; hPrp28	95.6	9.1	DExH/D; SR repeat
	PRPF6	U5-102K; hPrp6	106.9	10.1	HAT/TPR repeats
	EFTUD2	U5-116K; hSnu114	109.4	10.4	EF2-like fold; GTPase
	SNRNP20	U5-200K; hBrr2	244.5	23.2	DExH/D
	PRPF8	U5-220k; hPrp8	273.6	25.9	RNase H-fold; RRM; Jab1/MPN
U4/U6 (589.1 kDa)	RNU4	U4 snRNA	46.9	8.0	
	RNU6	U6 snRNA	34.6	5.9	
	SNRPB, -B2, -D1, -D2, -D3, -E, -F, -G	seven Sm proteins	94.3	16.0	Sm
	LSM2, -3, -4, -5, -6, -7, -8	seven LSm proteins	78.9	13.4	Sm
	NHP2L1	15.5K	14.2	2.4	
	PPIH	U4/U6-20K; SnuCyp-20	19.2	3.3	cyclophilin-like
	PRPF31	U4/U6-61K; hPrp31	55.5	9.4	Nop
	PRPF4	U4/U6-60K; hPrp4	58.4	9.9	WD40
	PRPF3	U4/U6-90K; hPrp3	77.5	13.1	PWI
	SART3	p110; SART3; hPrp24	109.6	18.6	HAT repeats; RRM

The cis determinants of splicing



This is GT-AG introns (by far the most frequent)
A secondary type exist (xx-xx), requiring U11 and U12 snRNP

Code	Represents	Complement
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N
-	Gap	-

DNA base code

Alternative splicing

Sequences at the borders of exon-intron and within the intron are similar but can vary.

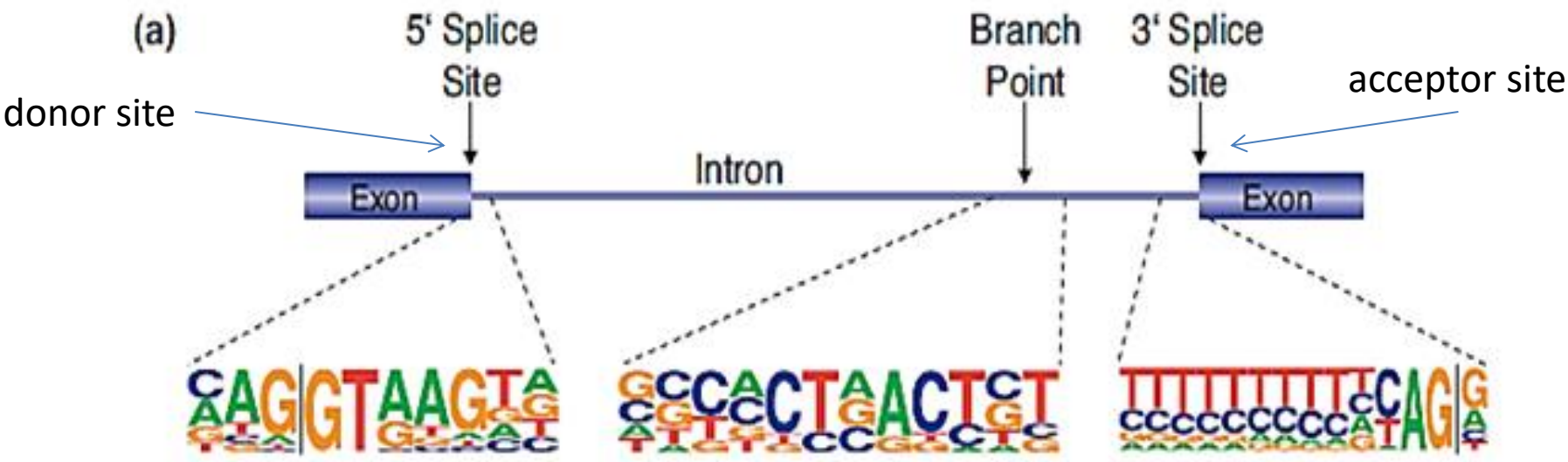
Splice sites can be **strong** or **weak** depending on how far their sequences diverge from the consensus sequence.

This determines their affinities for cognate splicing factors

In general, strong splice sites lead to constitutive splicing and full usage of the site

Pay attention to this concept, it will ground the discussion on PolII elongation rate as determinant of AS

The first chance to obtain regulation derives from how exons are recognized



Variations of these sequences can give «stronger» and «weaker» splicing sites

From: McManus & Graveley, COGD, 2011

Indeed, **in addition** to the sequences directly regulating binding of spliceosome components, in both Exons and Introns sequence motifs exist that regulate the use of splice sites.

Named after their location and effect:

ESE: exonic splicing enhancers
ESS: exonic splicing silencers
ISE: intronic splicing enhancers
ISS: intronic splicing silencers

They are therefore ***cis*-elements** for splicing regulation

Trans-acting factors of splicing are proteins binding to these elements.
They belong to the general class of RNA Binding Proteins (**RBP**)

Three categories:

1. SR proteins and SR-like

2. hnRNP

3. tissue-specific and context-specific factors

SR proteins = splicing regulators

Domains:
The most typical domain is an alternating Arginine-Serine-rich domain, called “RS domain”: it is a protein-protein interaction domain.

Regulation:
SR are phosphorylated at Ser by several kinases → regulates interaction with each other and with other proteins.

Other interactants:
SR proteins also interact with the CAP-binding protein and with poly-A binding proteins.

Binding sites:
Mostly at Exons, sometimes also to ISE (intronic splicing enhancers)

Activity:
Mostly activatory toward the most proximal exon.
Exon definition.

Canonical SR proteins

SR-like
(other protein
containing an RS
domain

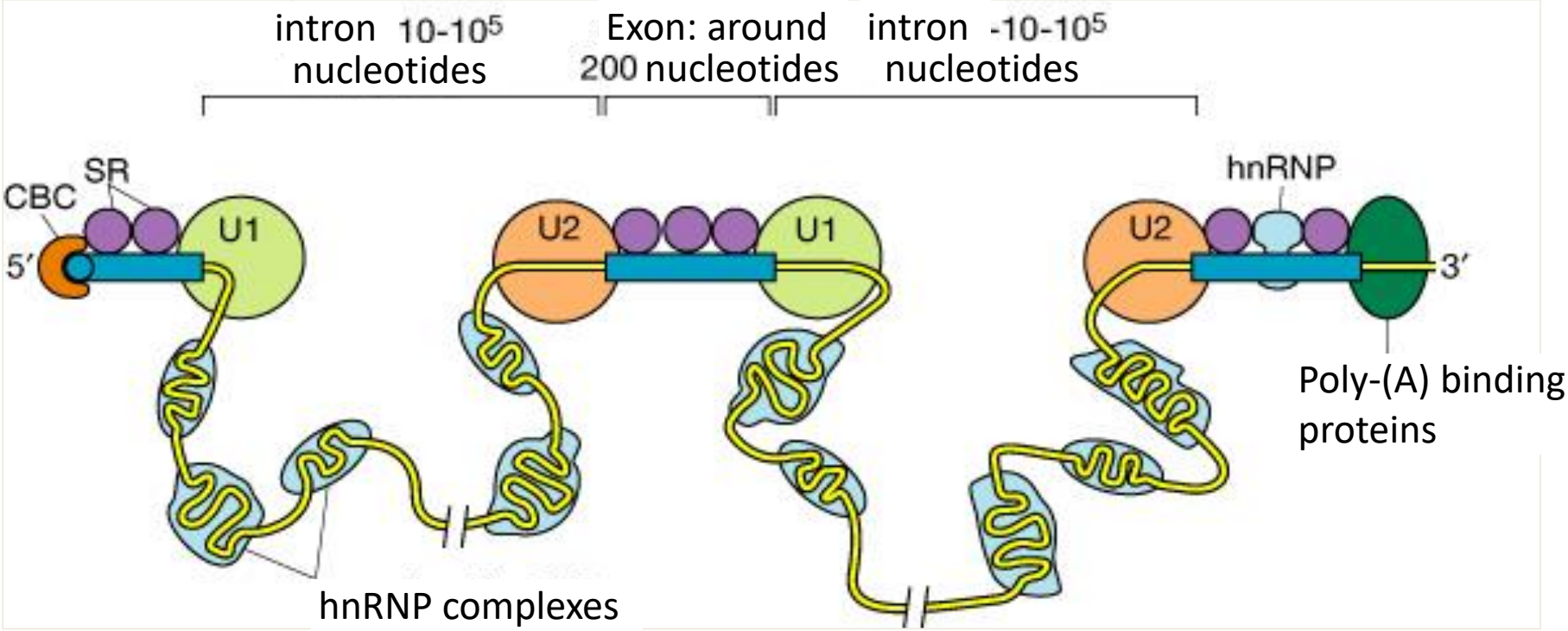
Name*	Domains	Binding sequence	Target genes
Canonical SR proteins			
SRp20 (SFRS3)	RRM and RS	GCUCCUCUUC	*SRP20*, *CALCA* and *INSR*
SC35 (SFRS2)	RRM and RS	UGCUGUU	*ACHE* and *GRIA1–GRIA4*
ASF/SF2 (SFRS1)	RRM, RRMH and RS	RGAAGAAC	*HIPK3*, *CAMK2D*, HIV RNAs and *GRIA1–GRIA4*
SRp40 (SFRS5)	RRM, RRMH and RS	AGGAGAAGGGA	*HIPK3*, *PRKCB* and *FN1*
SRp55 (SFRS6)	RRM, RRMH and RS	GGCAGCACCUG	*TNNT2* and *CD44*
SRp75 (SFRS4)	RRM, RRMH and RS	GAAGGA	*FN1*, *E1A* and *CD45*
9G8 (SFRS7)	RRM, zinc finger and RS	(GAC)n	*TAU*, *GNRH* and *SFRS7*
SRp30c (SFRS9)	RRM, RRMH and RS	CUGGAUU	*BCL2L1*, *TAU* and *HNRNPA1*
SRp38 (FUSIP1)	RRM and RS	AAAGACAAA	*GRIA2* and *TRD*
Other SR proteins			
SRp54	RRM and RS	ND	*TAU*
SRp46 (SFRS2B)	RRM and RS	ND	NA
RNPS1	RRM and Ser-rich	ND	*TRA2B*
SRrp35	RRM and RS	ND	NA
SRrp86 (SRrp508 and SFRS12)	RRM and RS	ND	NA
TRA2α	RRM and two Arg-rich	GAAARGARR	dsx
TRA2β	RRM and two RS	(GAA)n	*SMN1*, *CD44* and *TAU*
RBM5	RRM and RS	ND	*CD95*
CAPER (RBM39)	RRM and RS	ND	*VEGF*

hnRNP proteins (heterogeneous nuclear Ribo Nucleic Protein)

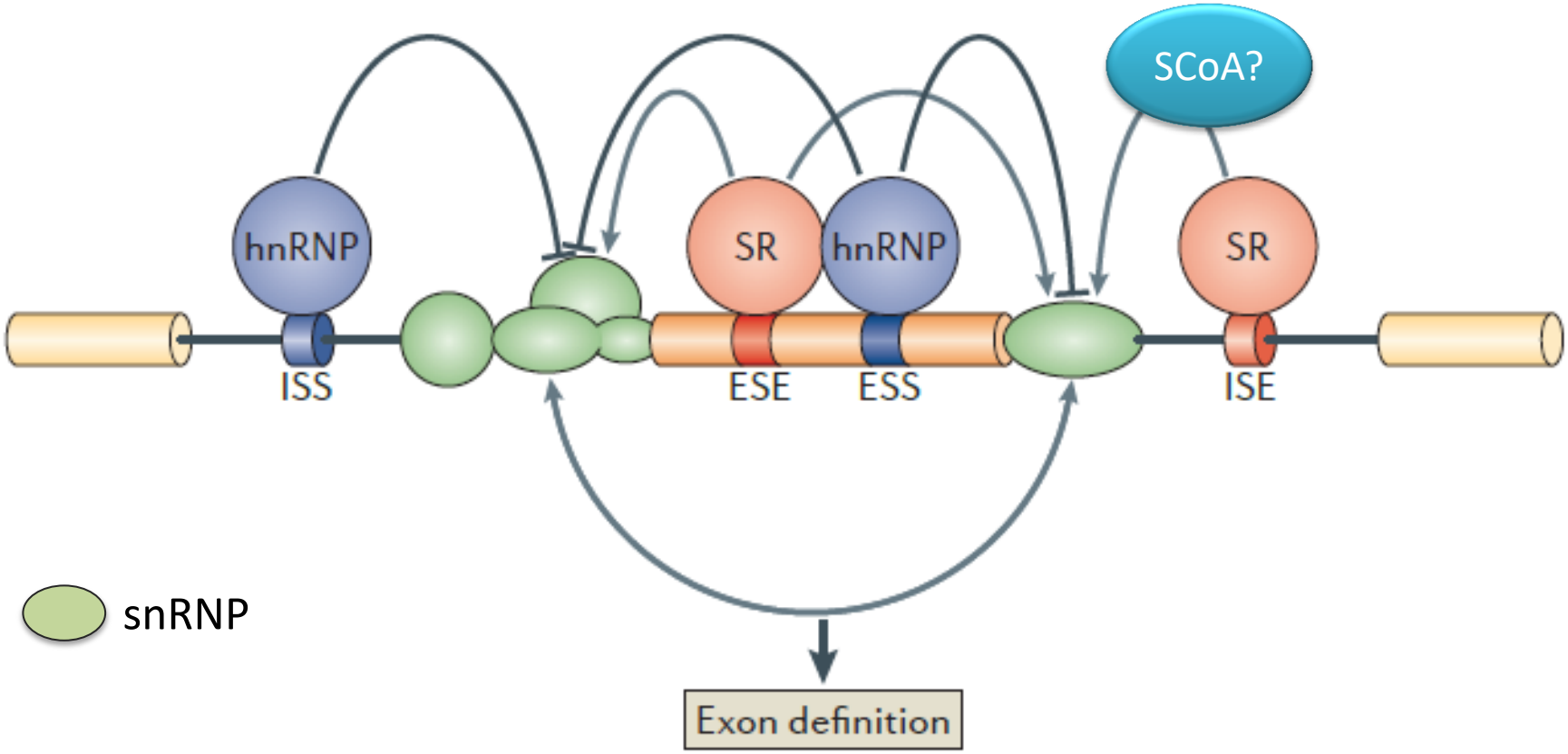
- Many different families
- Usually bind intronic sites
- Intron definition
- Several other roles have been ascribed to individual members, e.g. cytoplasmic localization.

Table 1 | **Ribonucleoproteins that are involved in pre-mRNA splicing**

Name	Other names	Domains*	Binding sequences	Target genes
hnRNP A1	NA	RRM, RGG and G	UAGGGA/U	*SMN2* and *RAS*
hnRNP A2 hnRNP B1	NA	RRM, RGG and G	(UUAGGG)n	HIV *tat* and *IKBKAP*
hnRNP C1 hnRNP C2	AUF1	RRM	U rich	*APP*
hnRNP F	NA	RRM, RGG and GY	GGGA and G rich	*PLP*, *SRC* and *BCL2L2*
hnRNP G	NA	RRM and SRGY	CC(A/C) and AAGU	*SMN2* and *TMP1*
hnRNP H hnRNP H′	DSEF1	RRM, RGG, GYR and GY	GGGA and G rich	*PLP*, HIV *tat* and *BCL2L1*
hnRNP I	PTB	RRM	UCUU and CUCUCU	*PTB*, *nPTB*, *SRC*, *CD95*, *TNTT2*, *CALCA* and *GRIN3B*
hnRNP L	NA	RRM	C and A rich	*NOS* and *CD45*
hnRNP LL	SRRF	RRM	C and A rich	*CD45*
hnRNP M	NA	RRM and GY	ND	*FGFR2*
hnRNP Q	NA	RRM and RGG	ND	*SMN2*


intron 10-10^5 nucleotides
Exon: around 200 nucleotides
intron -10-10^5 nucleotides
CBC
SR
5′
U1
U2
U1
U2
hnRNP
3′
Poly-(A) binding proteins
hnRNP complexes

The effect of SR and hnRNP binding is either to stabilize or destabilize the interaction of basal splicing factors (snRNPs) with the splicing sites
This action can be direct protein-protein contact, or mediated by splicing co-activators.



Albeit SR protein can be regulated by signalling pathways (*e.g. by phosphorylation, and several examples are given in your Kornblhitt Textbook*), as well as some hnRNP proteins, the fact that most SRs and most hnRNPs have ubiquitous expression suggests that additional tissue-specific factors should be involved in tissue-specific AS.

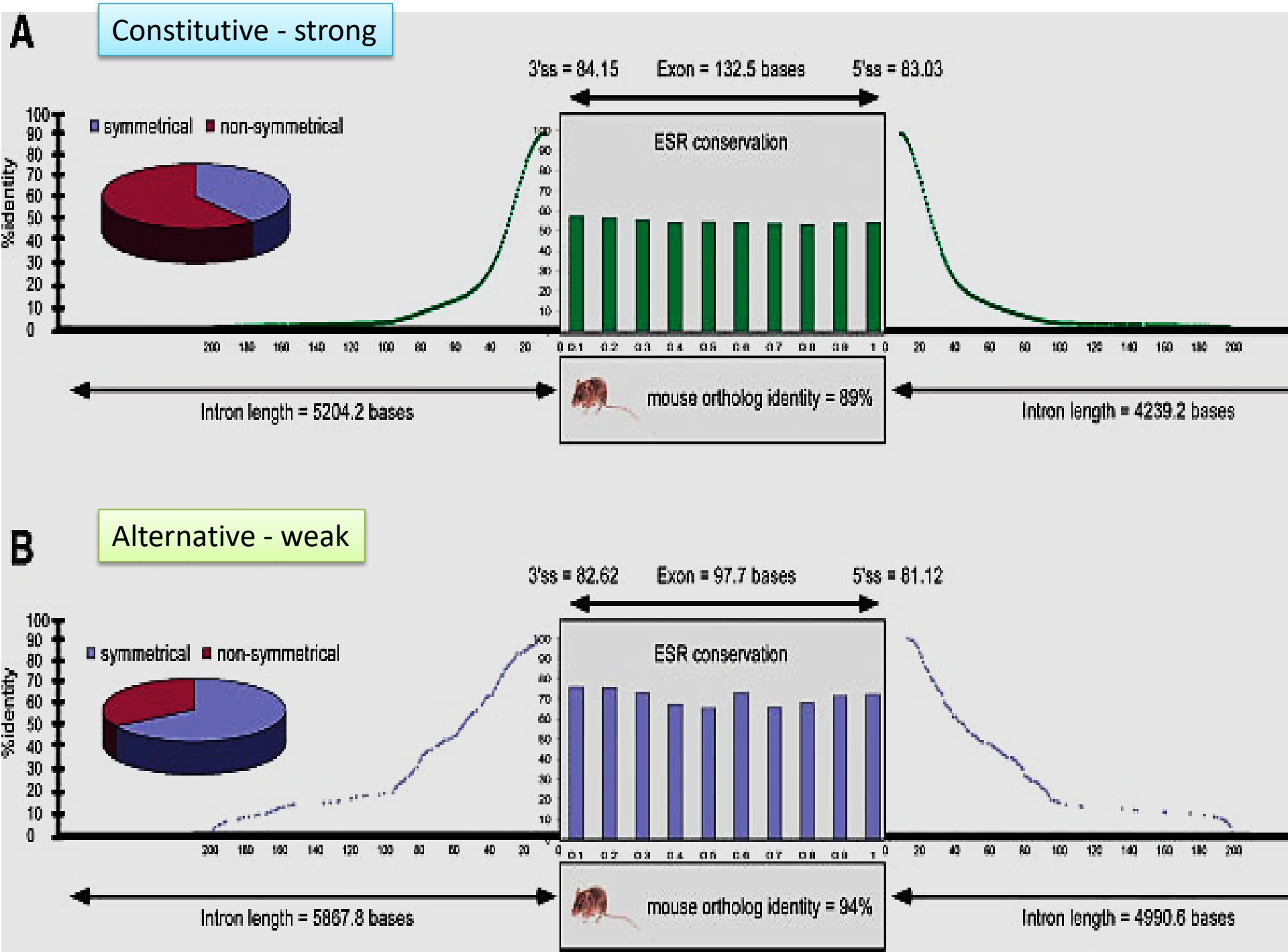
Hence, additional cis-elements sould be present in regulated pre-mRNAs.

One way to answer this question is to compare sequences around **alternative** exons to those of **constitutive** exons.

Some Authors used comparative genomics to obtain insight.
Human and mouse transcriptomes are quite well characterized, making it possible to classify exons as constitutive or alternative based on real expression data (microarrays, RNA-seq).

Sequences were then compared. Exons were normalized in lenght and flanking introns were explored within 200 pb.

From Kim et al., BioEssays 30:38–47


A
Constitutive - strong
3'ss = 84.15
Exon = 132.5 bases
5'ss = 83.03
symmetrical
non-symmetrical
%identity
ESR conservation
Intron length = 5204.2 bases
mouse ortholog identity = 89%
Intron length = 4239.2 bases
B
Alternative - weak
3'ss = 82.62
Exon = 97.7 bases
5'ss = 81.12
symmetrical
non-symmetrical
%identity
ESR conservation
Intron length = 5867.8 bases
mouse ortholog identity = 94%
Intron length = 4990.6 bases

2.1 Features of Alternative *versus* Constitutive Exons

More conserved

Exons that are alternatively spliced in both human and mouse are **more conserved** than constitutive exons

Conservation is higher toward exon edges and extends farther in introns: **cis-regulatory sequences** ?

Weak splice sites

Cassette exons[1] have **weak splice sites**, compared to the strong ones in constitutive exons

Shorter

Alternative cassette exons are also **shorter** and are flanked by longer introns than constitutively spliced ones.

Symmetric

The percentage of **symmetrical** exon is definitely higher in alternative exons (symmetrical means "divisible-by-three" number of base pairs)

[1]*Cassette exons: refers to exon skipping*

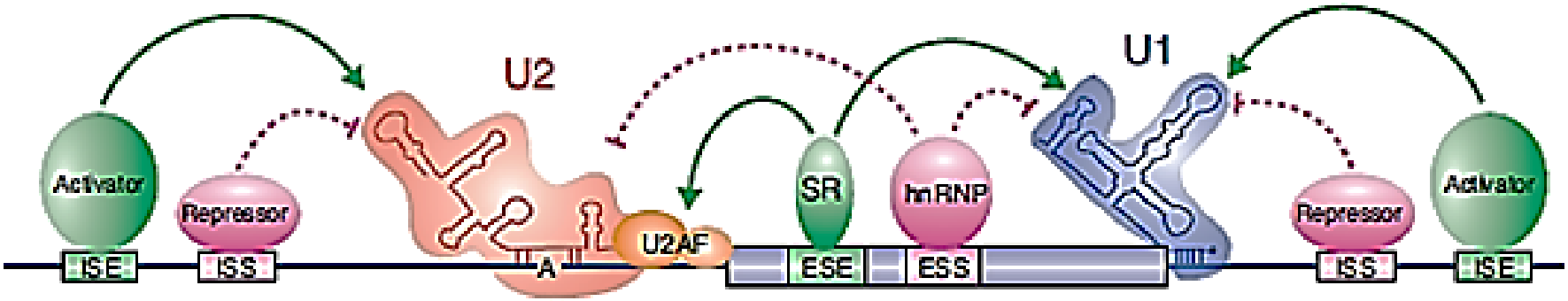
From Kim et al., 2007, BioEssays 30:38–47

All this points to the existence of additional AS regulators

Regulatory

- Tissue-specific splicing
- Regulated splicing
- Epigenetic establishment of splicing patterns

1) Tissue-specific splicing factors
2) Signal transduction regulated factors
3) Chromatin effects on splicing choice



Major SR proteins and hnRNP can hardly explain tissue-specific splicing

How to identify tissue-specific AS regulators

Appendix: The search for tissue-specific splicing factors

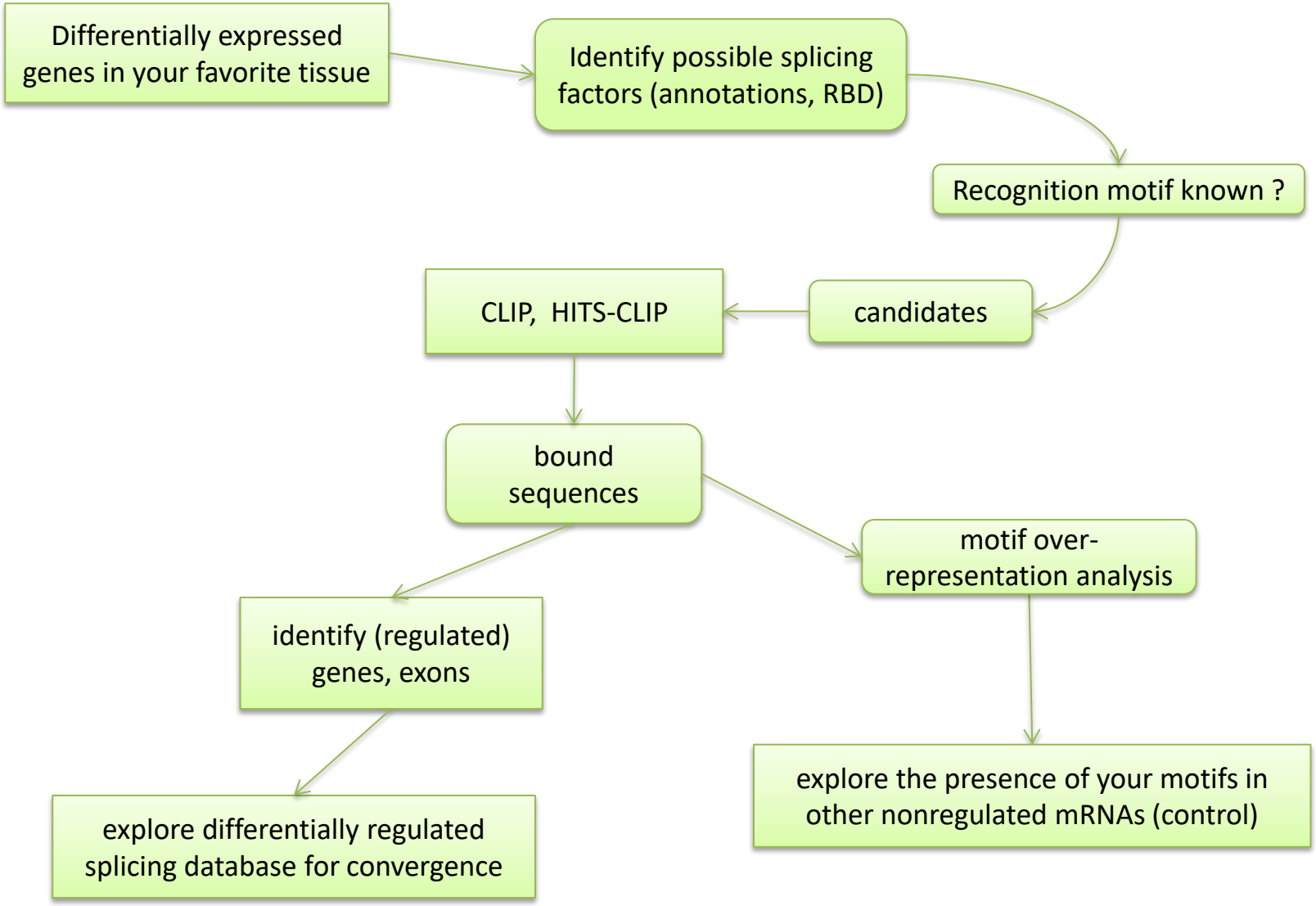


Table 2 | **Tissue-specific alternative splicing factors**

Name	Other names	Binding domain	Binding motif	Tissue expression	Target genes
nPTB	brPTB and PTBP2	RRM	CUCUCU	Neurons, myoblasts and testes	*BIN1*, *GLYRA2*, *ATP2B1*, *MEF2*, *NASP*, *SPAG9* and *SRC*
NOVA1	NA	KH	YCAY	Neurons of the hindbrain and spinal cord	*GABRG2*, *GLYRA2* and *NOVA1*
NOVA2	NA	KH	YCAY	Neurons of the cortex, hippocampus and dorsal spinal cord	*KCNJ*, *APLP2*, *GPHN*, *JNK2*, *NEO*, *GRIN1* and *PLCB4*
FOX1	A2BP1	RRM	(U)GCAUG	Muscle, heart and neurons	*ACTN*, *EWSR1* , *FGFR2*, *FN1* and *SRC*
FOX2	RBM9	RRM	(U)GCAUG	Muscle, heart and neurons	*EWS*, *FGFR2*, *FN1* and *SRC*
RBM35a	ESRP1	RRM	GU rich	Epithelial cells	*FGFR2*, *CD44*, *CTNND1* and *ENAH*
RBM35b	ESRP2	RRM	GU rich	Epithelial cells	*FGFR2*, *CD44*, *CTNND1* and *ENAH*
TIA1	mTIA1	RRM	U rich	Brain, spleen and testes	*MYPT1*, *CD95*, *CALCA*, *FGFR2*, *TIAR*, *IL8*, *VEGF*, *NF1* and *COL2A1*
TIAR	TIAL1 and mTIAR	RRM	U rich	Brain, spleen, lung, liver and testes	*TIA1*, *CALCA* , *TIAR*, *NF1* and *CD95*
SLM2	KHDRBS3 and TSTAR	KH	UAAA	Brain, tests and heart	*CD44* and *VEGFA*
Quaking	QK and QKL	KH	ACUAAY[...]UAAY	Brain	*MAG* and *PLP*
HUB	HUC, HUD and ELAV2	RRM	AU rich	Neurons	*CALCA*, *CD95* and *NF1*

Chen & Manley 2009. Nat Rev Mol Cell Biol., 10:741.

MBNL	NA	CCCH zinc finger domain	YGCU(U/ G)Y	Muscles, uterus and ovaries	*TNTT2*, *INSR*, *CLCN1* and *TNNT3*
CELF1	BRUNOL2	RRM	U and G rich	Brain	*TNTT2* and *INSR*
ETR3	CELF2 and BRUNOL3	RRM	U and G rich	Heart, skeletal muscle and brain	*TNTT2*, *TAU* and *COX2*
CELF4	BRUNOL4	RRM	U and G rich	Muscle	*MTMR1* and *TNTT2*
CELF5	BRUNOL5 and NAPOR	RRM	U and G rich	Heart, skeletal muscle and brain	*ACTN*, *TNTT2* and *GRIN1*
CELF6	BRUNOL6	RRM	U and G rich	Kidney, brain and testes	*TNTT2*

A2BP1, ataxin 2-binding protein 1; *ACTN*, α-actinin; *APLP2*, amyloid-β precursor-like protein 2; *ATP2B1*, ATPase, Ca^{2+} transporting, plasma membrane 1; *BIN1*, bridging integrator 1; *CALCA*, calcitonin-related polypeptide-α; CELF, CUGBP- and ETR3-like factor; *CLCN1*, chloride channel 1; *COL2A1*, collagen, type II, α1; *COX2*, cytochrome c oxidase II; *CTNND1*, catenin δ1, *EWSR1*, Ewing sarcoma breakpoint region 1; *FGFR2*, fibroblast growth factor receptor 2; *FN1*, fibronectin 1; *GABRG2*, GABA A receptor, γ2; *GLYRA2*, glycine receptor, α2 subunit; *GPHN*, gephyrin; *GRIN1*, glutamate receptor, ionotropic, NMDA 3B; *IL8*, interleukin-8; *INSR*, insulin receptor; *JNK2*, Jun N-terminal kinase 2; *KCNJ*, potassium inwardly-rectifying channel, subfamily; KHDRBS3, KH domain-containing, RNA-binding, signal transduction-associated protein 3; *MAG*, myelin associated glycoprotein; MBNL, muscleblind; *MEF2*, myocyte enhancing factor 2; *MTMR1*, myotubularin-related protein 1; *NASP*, nuclear autoantigenic sperm protein; *NEO*, neogenin; *NF1*, neurofibromin 1; *NOVA*, neuro-oncological ventral antigen; *PLCB4*, phospholipase C β4; *PLP*, proteolipid protein; PTB, polypyrimidine-tract binding protein; RBM, RNA-binding protein; RRM, RNA recognition motif; *SLM2*, SAM68-like mammalian protein 2; *SPAG9*, sperm associated antigen 9; TIA1, T cell-restricted intracellular antigen 1; TIAR, TIA1-related protein; *TNTT2*, troponin T type 2; *VEGF*, vascular endothelial growth factor.

Chen & Manley 2009. Nat Rev Mol Cell Biol., 10:741.

RNA Binding Proteins are a growing class ...

Generic name: RBP (RNA binding Proteins) GO category: RNA-Binding

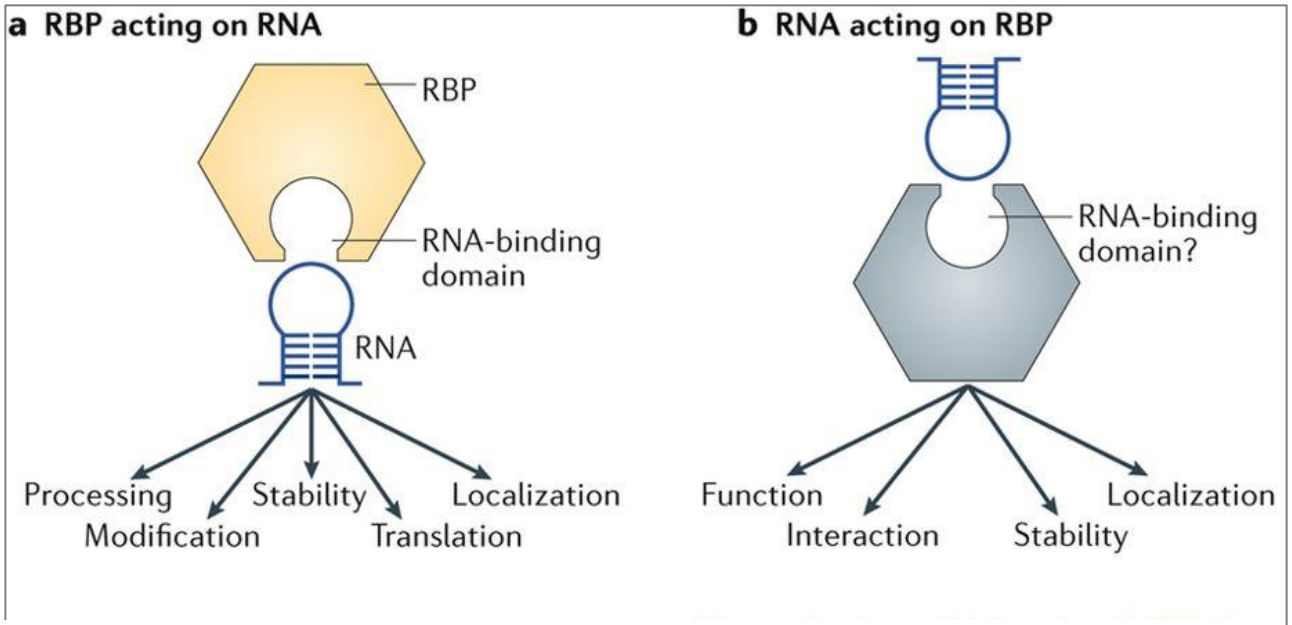
Recent studies used RIC (RNA Interactome Capture) identified an exceptional number of RBPs (860 from HeLa and 791 from HEK293). Many of these do not carry any of the known domains:

-RRM - RNA Recognition Motif

-KH

-DEAD box helicase

-Zn-fingers motifs



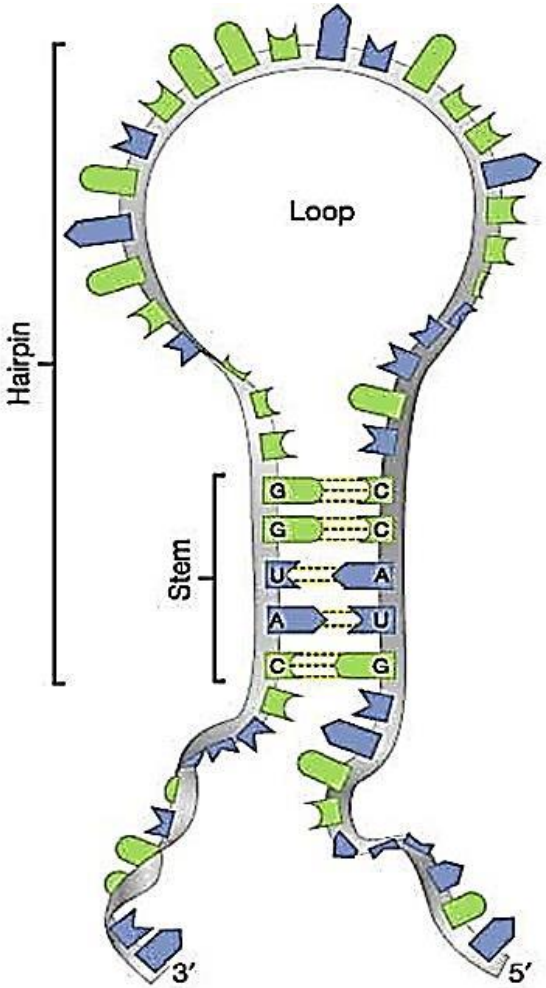
Form Hentze et al. Nature Reviews Molecular Cell Biology **volume 19**, pages 327–341 (2018)

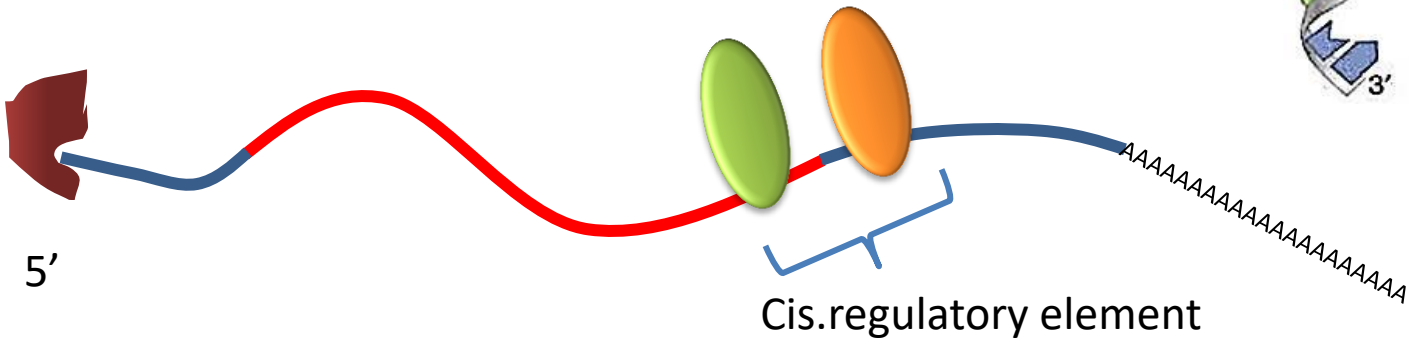
Identification of RNA binding proteins motifs

Specificity of RNA binding: both «sequence» and «structure» elements

Problems in predicting regulatory motifs:

- Localization (intron length)
- Sometimes dispersed elements
- Sometimes the structural component prevails upon pure sequence



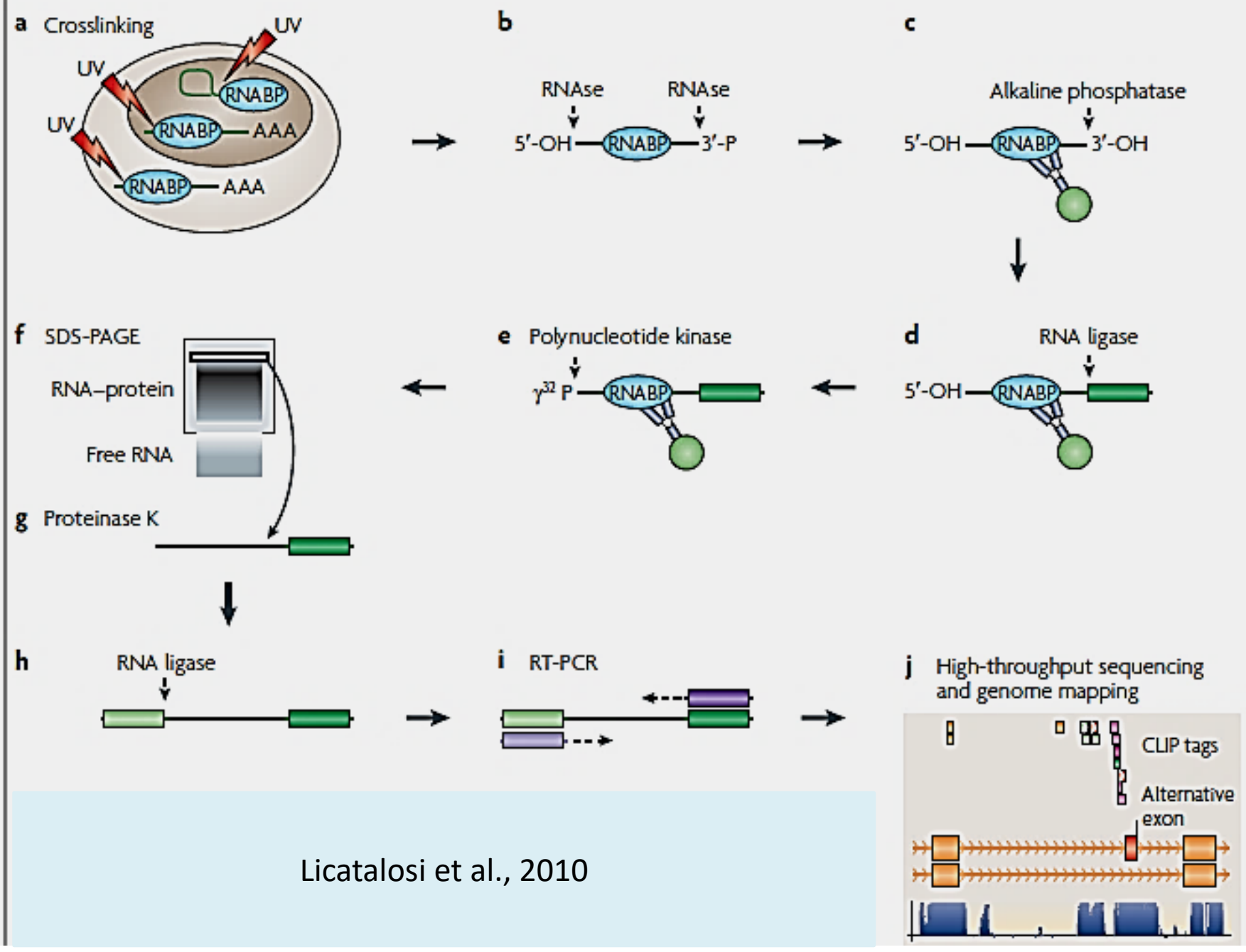


RRM: Sex-lethal http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1b7f

RRM: PTB http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=2adc

KH: NOVA1 http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1dt4

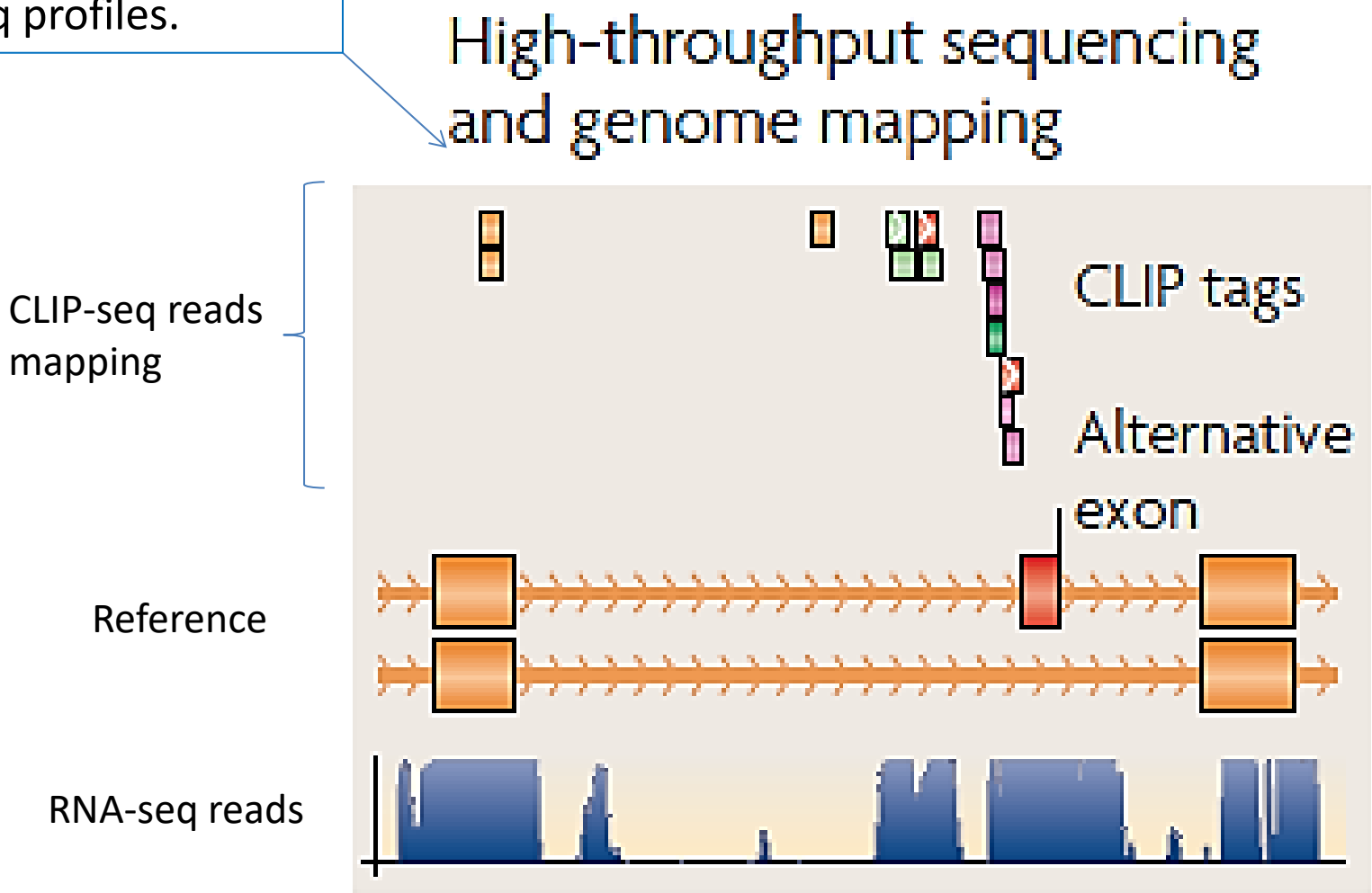
Box 2 | **CLIP and HITS-CLIP methods**



Licatalosi et al., 2010

Example:

Individual Splicing Factors HITS-CLIP profiles mapped to genome and compared to RNA-Seq profiles.



The mechanisms of alternative splicing regulation