

WHAT DON'T WE KNOW?

Why Do Humans Have So Few Genes



Regolazione dello splicing alternativo (AS)

Nelle lezioni della parte 2.1 abbiamo imparato:

- risultati di analisi con deep-sequencing → forse fino a 98% dei geni con AS (hu)
- AS diviene sempre più importante nell'evoluzione verso i Vertebrati
- meccanismo biochimico dello splicing/spliceosoma e proteine associate
- sequenze che definiscono confine esone/introne e sequenze introniche
- modelli di AS (exon skipping, alternative 5' / 3', mutually excl., etc.)

- Alternative TSS, poly(A) signals, frequenze di uso alternativo di esoni.
- Esempi funzionali di AS (compreso il gene Dscam di *D. melanogaster*)
- Studi genome-wide con microarrays esonici
- Studi genome-wide con microarrays exon-junction
- Studi genome-wide eseguiti mediante "RNA-Seq"

ARTICLES

Alternative isoform regulation in human tissue transcriptomes

Eric T. Wang^{1,2*}, Rickard Sandberg^{1,3*}, Shujun Luo⁴, Irina Khrebtukova⁴, Lu Zhang⁴, Christine Mayr⁵, Stephen F. Kingsmore⁶, Gary P. Schroth⁴ & Christopher B. Burge¹

Through alternative processing of pre-messenger RNAs, individual mammalian genes often produce multiple mRNA and protein isoforms that may have related, distinct or even opposing functions. Here we report an in-depth analysis of 15 diverse human tissue and cell line transcriptomes on the basis of deep sequencing of complementary DNA fragments, yielding a digital inventory of gene and mRNA isoform expression. Analyses in which sequence reads are mapped to exon–exon junctions indicated that 92–94% of human genes undergo alternative splicing, ~86% with a minor isoform frequency of 15% or more. Differences in isoform-specific read densities indicated that most alternative splicing and alternative cleavage and polyadenylation events vary between tissues, whereas variation between individuals was approximately twofold to threefold less common. Extreme or ‘switch-like’ regulation of splicing between tissues was associated with increased sequence conservation in regulatory regions and with generation of full-length open reading frames. Patterns of alternative splicing and alternative cleavage and polyadenylation were strongly correlated across tissues, suggesting coordinated regulation of these processes, and sequence conservation of a subset of known regulatory motifs in both alternative introns and 3′ untranslated regions suggested common involvement of specific factors in tissue-level regulation of both splicing and polyadenylation.

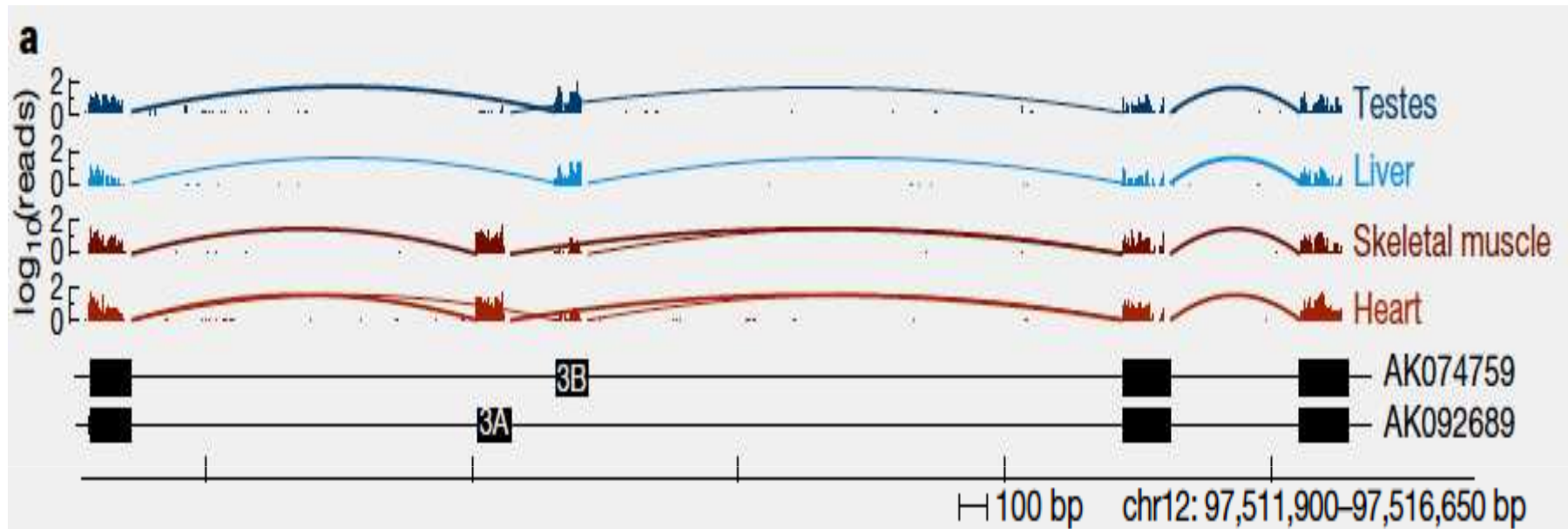
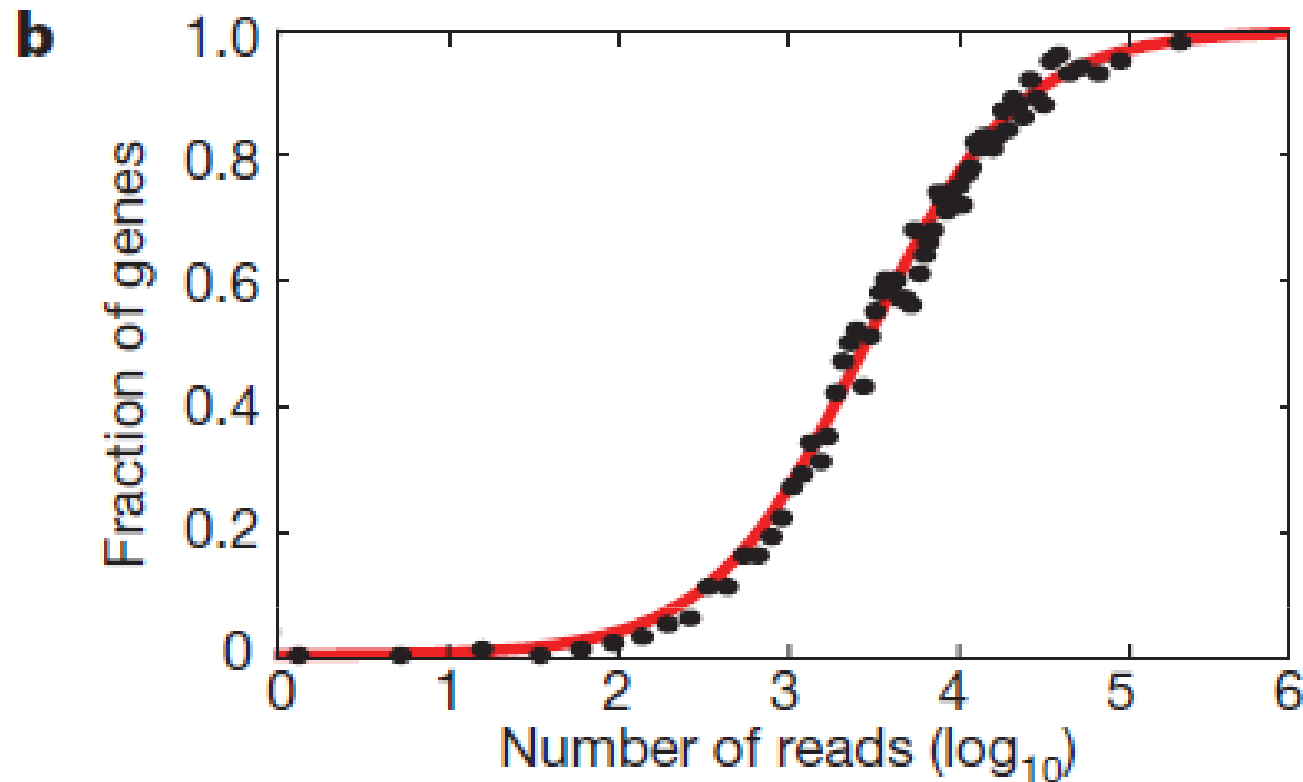


Figure 1 | Frequency and relative abundance of alternative splicing isoforms in human genes.



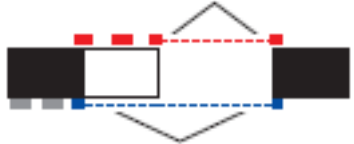
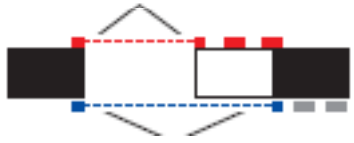
a, mRNA-Seq reads mapping to a portion of the SLC25A3 gene locus. The number of mapped reads starting at each nucleotide position is displayed (log₁₀) for the tissues listed at the right. Arcs represent junctions detected by splice junction reads. Bottom: exon/intron structures of representative transcripts containing mutually exclusive exons 3A and 3B (GenBank accession numbers shown at the right).



b, Mean fraction of multi-exon genes with detected alternative splicing in bins of 500 genes, grouped by total read count per gene. A gene was considered as alternatively spliced if splice junction reads joining the same 5' splice site (5'SS) to different 3' splice sites (3'SS) (with at least two independently mapping reads supporting each junction), or joining the same 3'SS to different 5'SS, were observed. The true extent of alternative splicing was estimated from the upper asymptote of the best-fit sigmoid curve (red curve). Circles show the fraction of alternatively spliced genes.

Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon		37	35	10,436	6,822	65	72
Retained intron		1	1	167	96	57	71
Alternative 5' splice site (A5SS)		15	15	2,168	1,386	64	72
Alternative 3' splice site (A3SS)		17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)		4	4	167	95	57	66
Alternative first exon (AFE)		14	13	10,281	5,311	52	63
Alternative last exon (ALE)		9	8	5,246	2,491	47	52
Tandem 3' UTRs		7	7	5,136	3,801	74	80
Total		105	100	37,782	22,657	60	68

Constitutive exon or region
 Body read
 Junction read
pA Polyadenylation site
 Alternative exon or extension
Inclusive/extended isoform
Exclusive isoform
Both isoforms

Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon		37	35	10,436	6,822	65	72
Retained intron		1	1	167	96	57	71
Alternative 5' splice site (A5SS)		15	15	2,168	1,386	64	72
Alternative 3' splice site (A3SS)		17	16	4,181	2,655	64	74


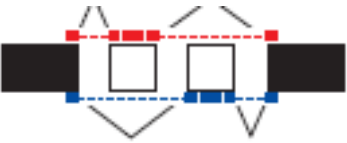
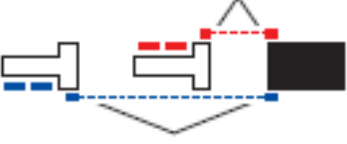
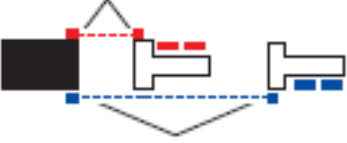

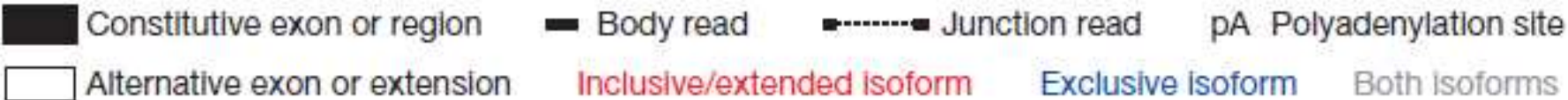


Figure 2 | Pervasive tissue-specific regulation of alternative mRNA isoforms. Rows represent the eight different alternative transcript event types diagrammed. Mapped reads supporting expression of upper isoform, lower isoform or both isoforms are shown in blue, red and grey, respectively. Columns 1-4 show the numbers of events of each type: (1) supported by cDNA and/or EST data; (2) with ≥ 1 isoform supported by mRNA-Seq reads; (3) with both isoforms supported by reads; and (4) events detected as tissue regulated (Fisher's exact test) at an FDR of 5% (assuming negligible technical variation¹⁰).

Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Mutually exclusive exon (MXE)		4	4	167	95	57	66
Alternative first exon (AFE)		14	13	10,281	5,311	52	63
Alternative last exon (ALE)		9	8	5,246	2,491	47	52
Tandem 3' UTRs		7	7	5,136	3,801	74	80
Total		105	100	37,782	22,657	60	68



Columns 5 and 6 show: (5) the observed percentage of events with both isoforms detected that were observed to be tissue-regulated; and (6) the estimated true percentage of tissue-regulated isoforms after correction for power to detect tissue bias (Supplementary Fig. 6) and for the FDR. For some event types, ‘common reads’ (grey bars) were used in lieu of (for tandem 39UTR events) or in addition to ‘exclusion’ reads for detection of changes in isoform levels between tissues.

Note that Aa use the following definition for “tissue-specific”:

at least 10% variation in isoforms

How is alternative splicing regulated ?

The first studies concentrated on the “transcriptional paradigm” i.e. on model reminiscent of transcriptional control....

Researchers started seeking for cis-regulatory elements and trans-regulatory proteins.

UNDERSTANDING ALTERNATIVE SPLICING: TOWARDS A CELLULAR CODE

Arianne J. Matlin[‡], Francis Clark^{} and Christopher W. J. Smith[‡]*

Abstract | In violation of the 'one gene, one polypeptide' rule, alternative splicing allows individual genes to produce multiple protein isoforms — thereby playing a central part in generating complex proteomes. Alternative splicing also has a largely hidden function in quantitative gene control, by targeting RNAs for nonsense-mediated decay. Traditional gene-by-gene investigations of alternative splicing mechanisms are now being complemented by global approaches. These promise to reveal details of the nature and operation of cellular codes that are constituted by combinations of regulatory elements in pre-mRNA substrates and by cellular complements of splicing regulators, which together determine regulated splicing pathways.

Nature Rev Mol Cell Biol (2005) 6:386.

Review

Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches

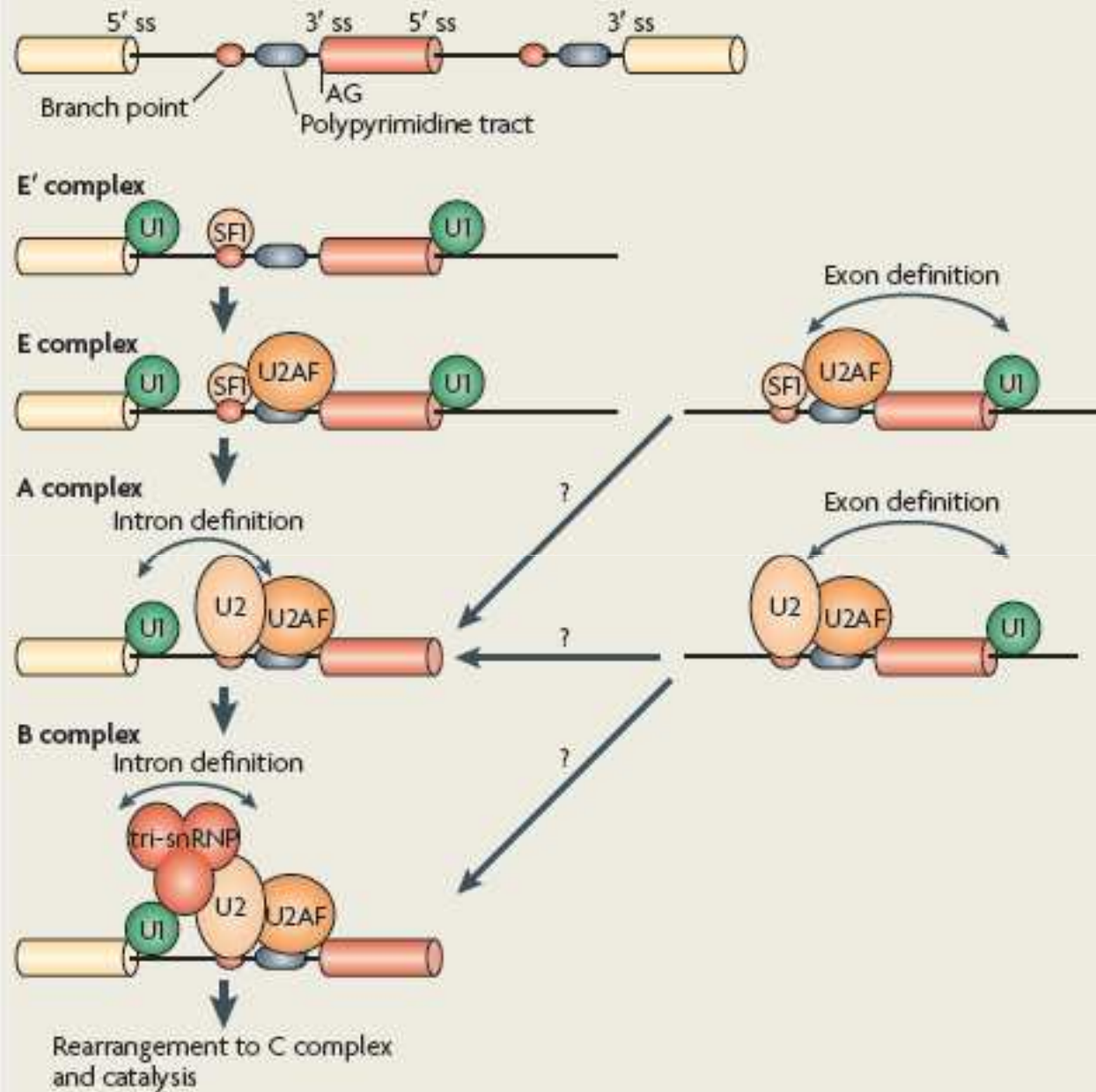
Mo Chen and James L. Manley

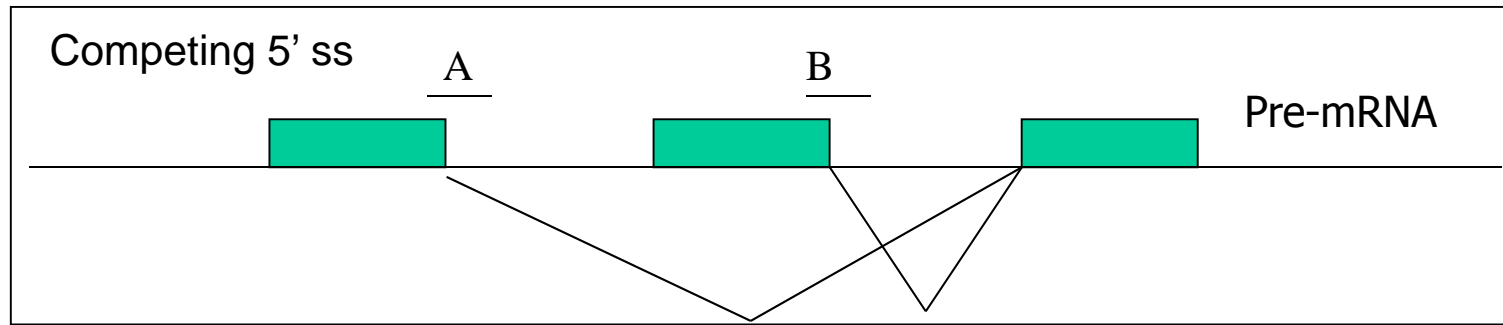
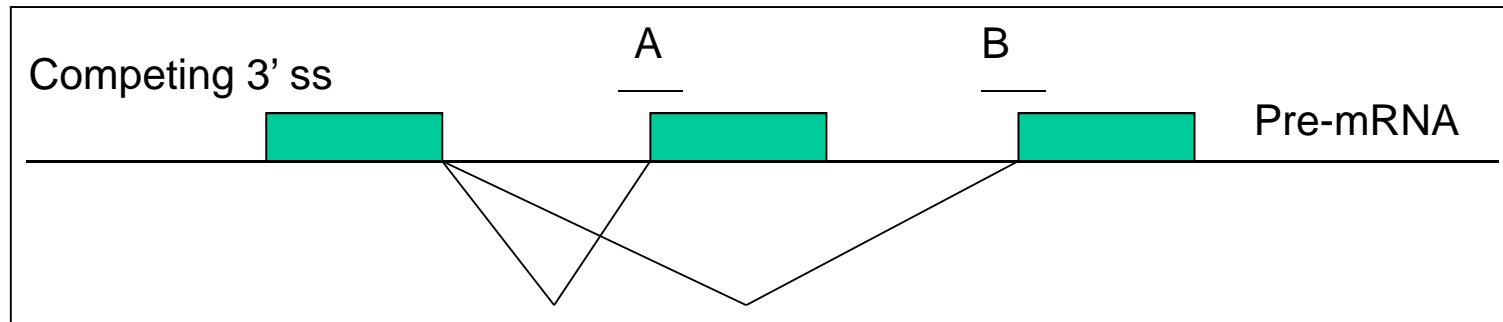
Abstract | Alternative splicing of mRNA precursors provides an important means of genetic control and is a crucial step in the expression of most genes. Alternative splicing markedly affects human development, and its misregulation underlies many human diseases.

Although the mechanisms of alternative splicing have been studied extensively, until the past few years we had not begun to realize fully the diversity and complexity of alternative splicing regulation by an intricate protein–RNA network. Great progress has been made by studying individual transcripts and through genome-wide approaches, which together provide a better picture of the mechanistic regulation of alternative pre-mRNA splicing.

Review

Box 1 | Splicing and spliceosome assembly

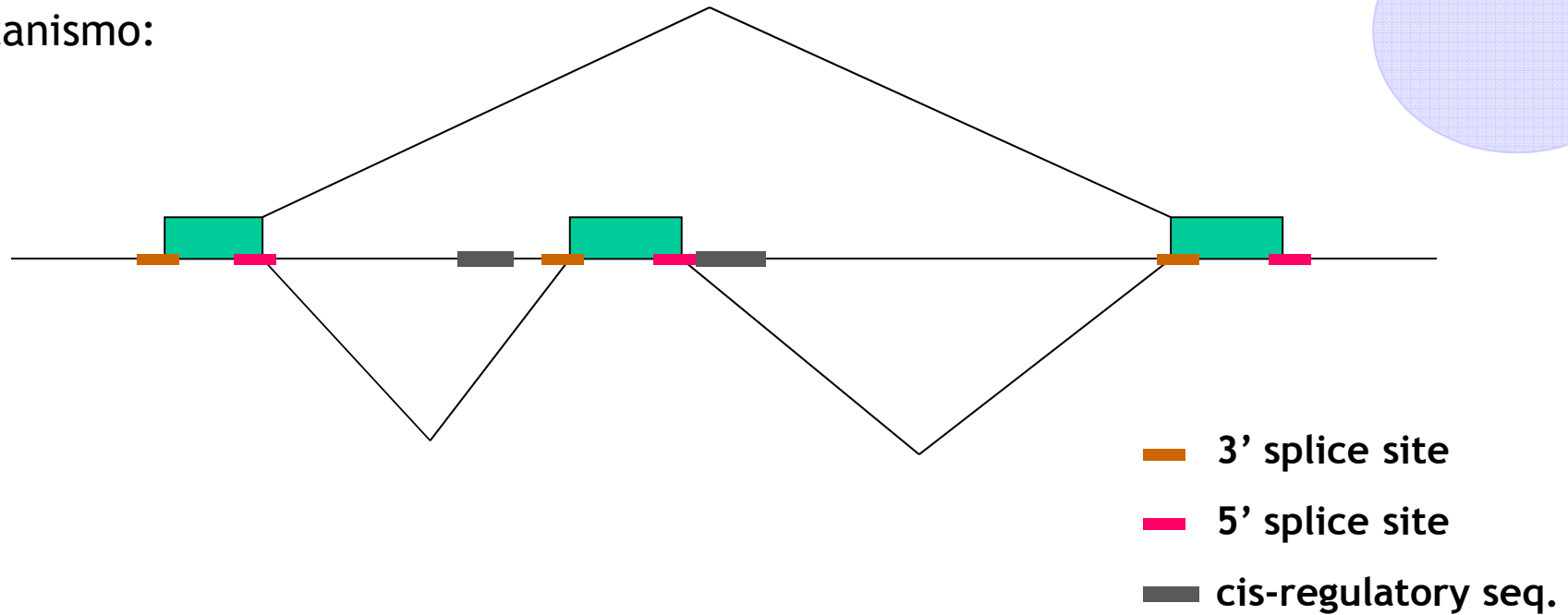




First point: What does mean "strong" or "weak" splice site?

- a) Of course the degree of complementarity to the RNA U comes first
- b) Are there additional sequences that contribute strength to the machinery? Are there specific proteins?

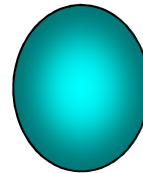
Meccanismo:



1) O i siti sono “belli”, ma vengono nascosti da un **inibitore**



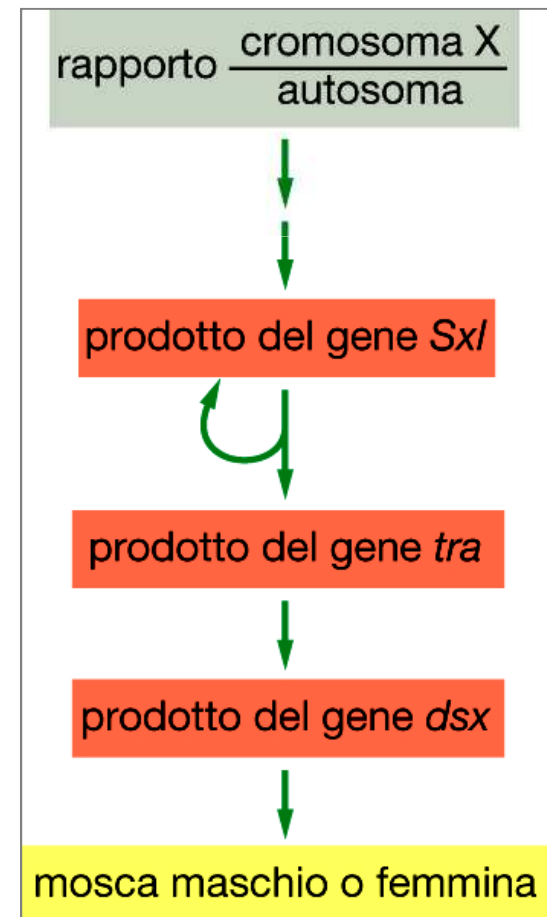
2) O i siti sono “brutti”, ma esistono sequenze accessorie che, mediante interazione con fattori *trans*, “aiutano” i siti a funzionare.

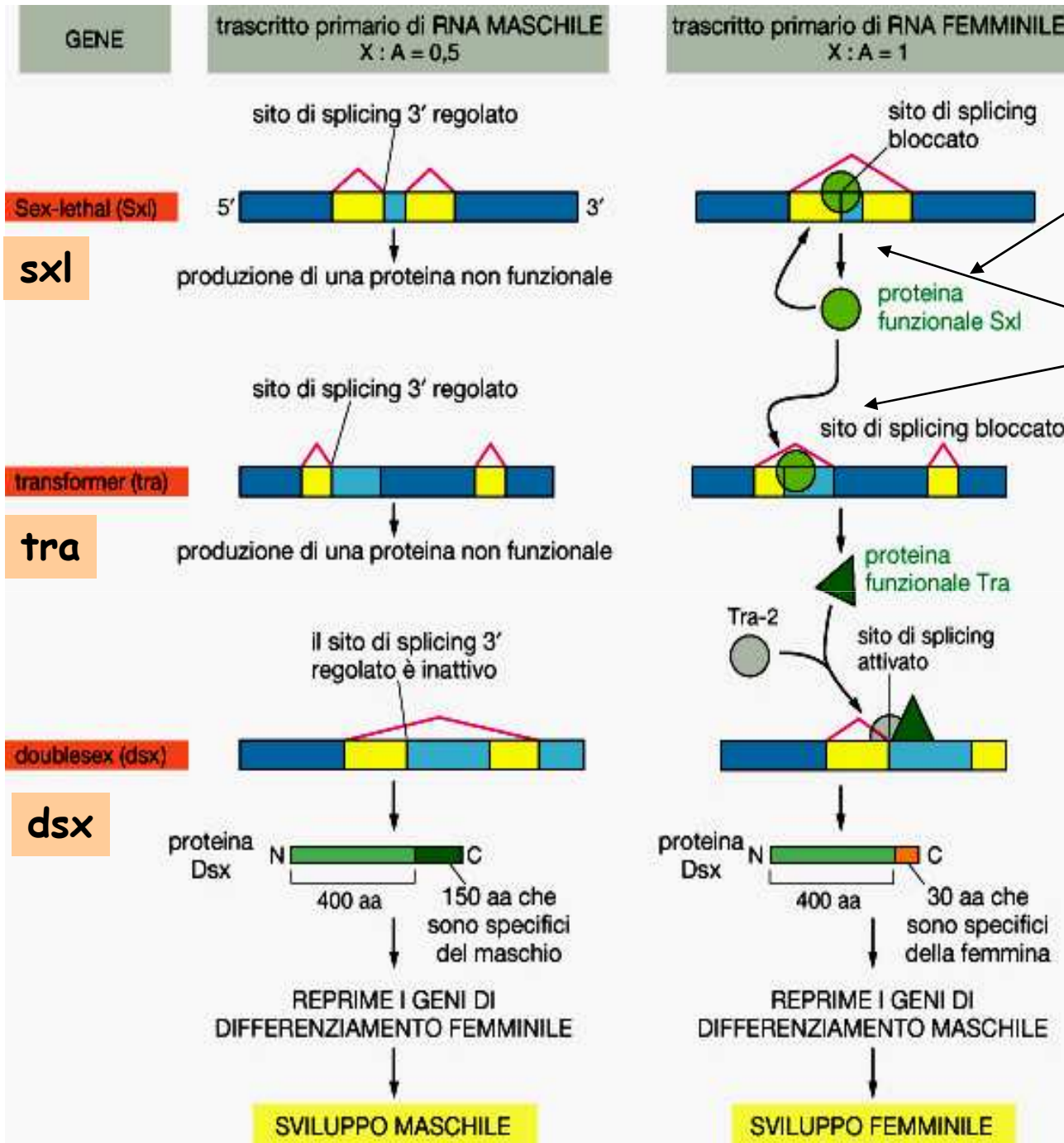


One of the first mechanisms studied illustrated exactly this simple situation.

Determination of sex in *Drosophila* gave the **first example** of exonic sequences enhancing a “poor” 3'-ss utilization as well as intronic sequences inhibiting 5'-ss

The first and most known model of regulated alternative splicing is the determination of sex in *Drosophila*. The primary determinant is the X:A chromosome ratio. This determines a cascade of splicing regulatory signals, resulting in the production of two alternative splicing isoforms of the *dsx* transcription factor, repressing either female-specific or male-specific genes.

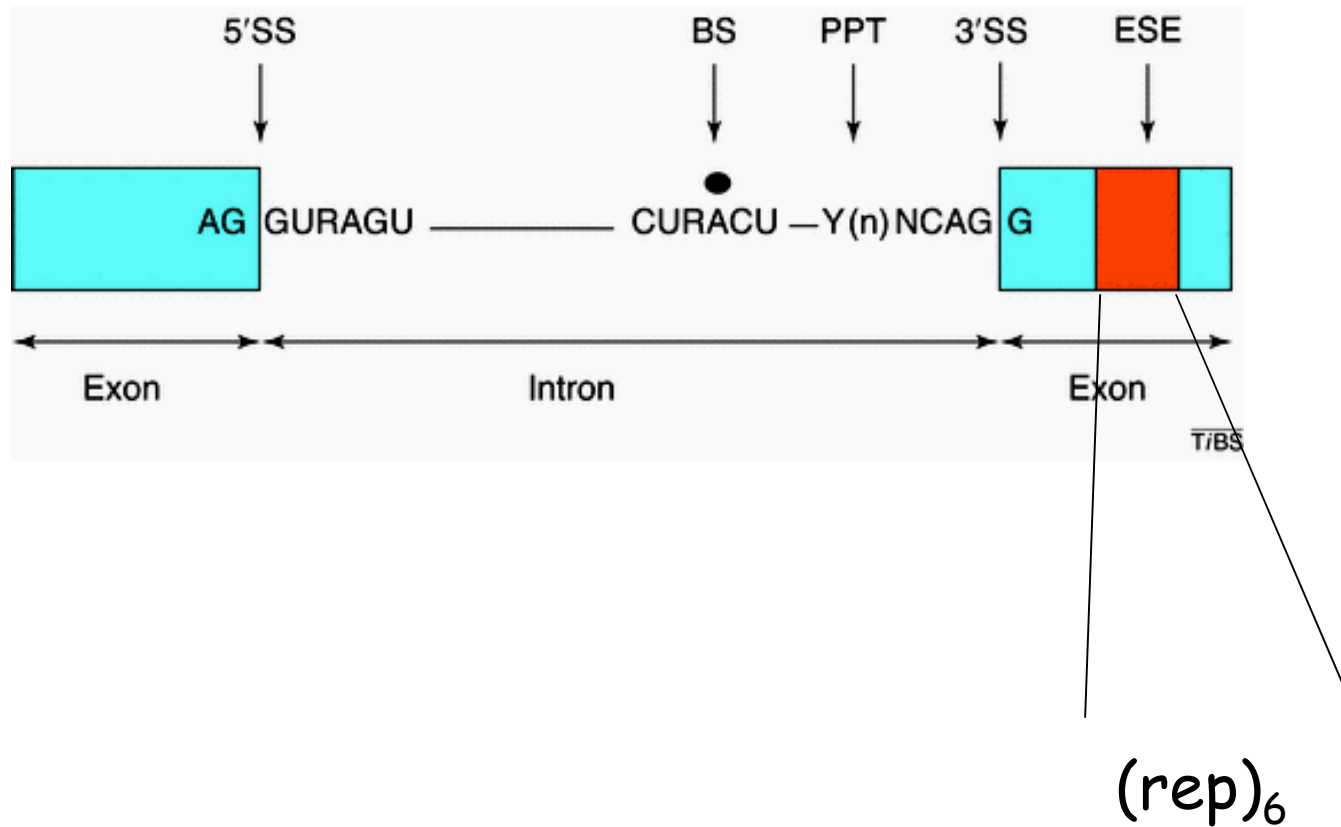


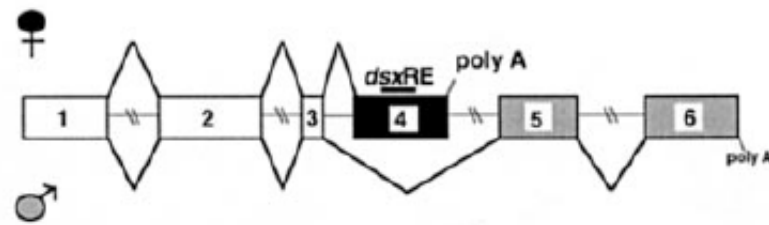
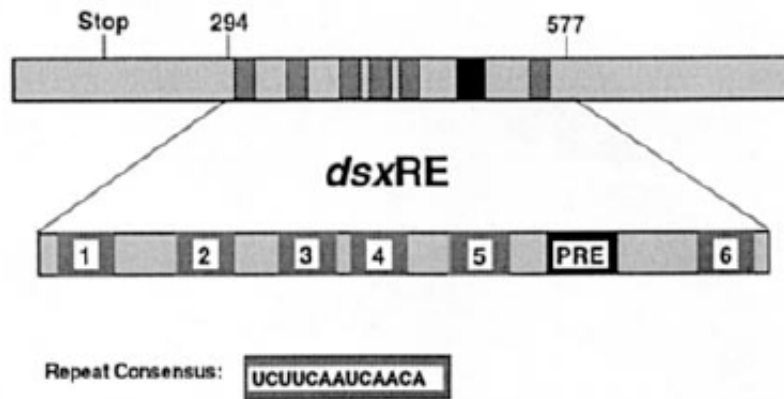
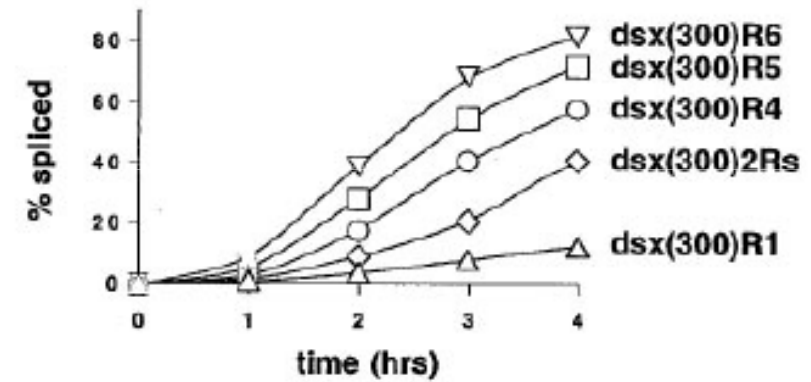
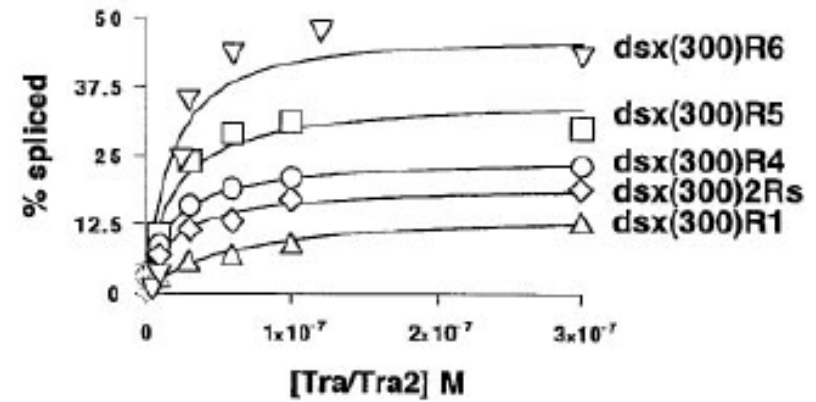
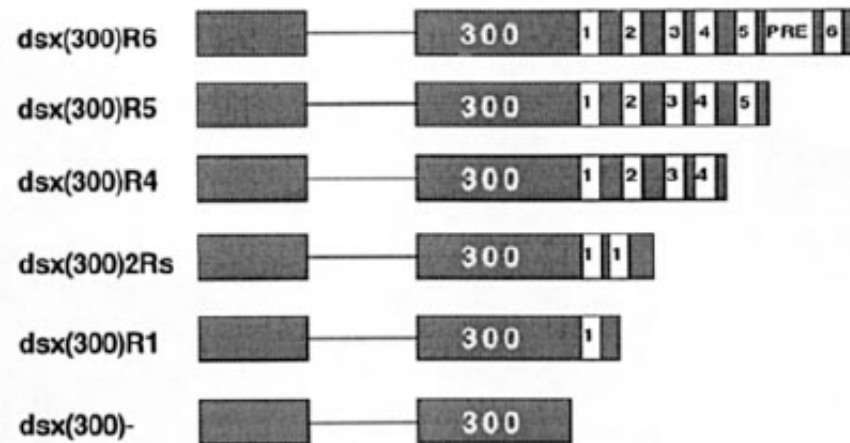


X:A=1 produces transient activation of an alternative promoter in the *sxl* gene, giving rise to a functional *sxl* protein

The *sxl* protein competes with U2AF for binding to the poly-pyrimidine tract

In the drosophila doublesex gene, the ESE element is present in exon 4 in addition to common intronic elements



A**Doublesex pre-mRNA****B****C**

From: Hertel & Maniatis, (1998)
Mol. Cell 1: 449.

Doublesex exon 4 repeats in *Drosophila* were the first discovered Exonic Splicing Enhancer (ESE)

ESE of diverse sequence were then recognized in a large number of exons in various species. Mutations in ESE were also found in disease, that lead to aberrantly spliced products.

ESE are the most frequent regulatory sequences found in pre-mRNAs

Which factors recognize ESEs ?

SR proteins = splicing regulators

The most typical domain is an alternating Arginine-Serine domain, called "RS domain": it is a protein-protein interaction domain.

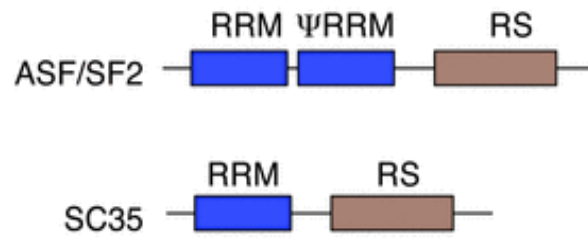
SR are phosphorylated at Ser by several kinases → regulates interaction with each other and with other proteins.

SR proteins are "proximalizing factors", whenever there is a "ss" choice, i.e. they have function in constitutive splicing, promoting the formation of complexes with pre-mRNA, snRNP U1 and U2.

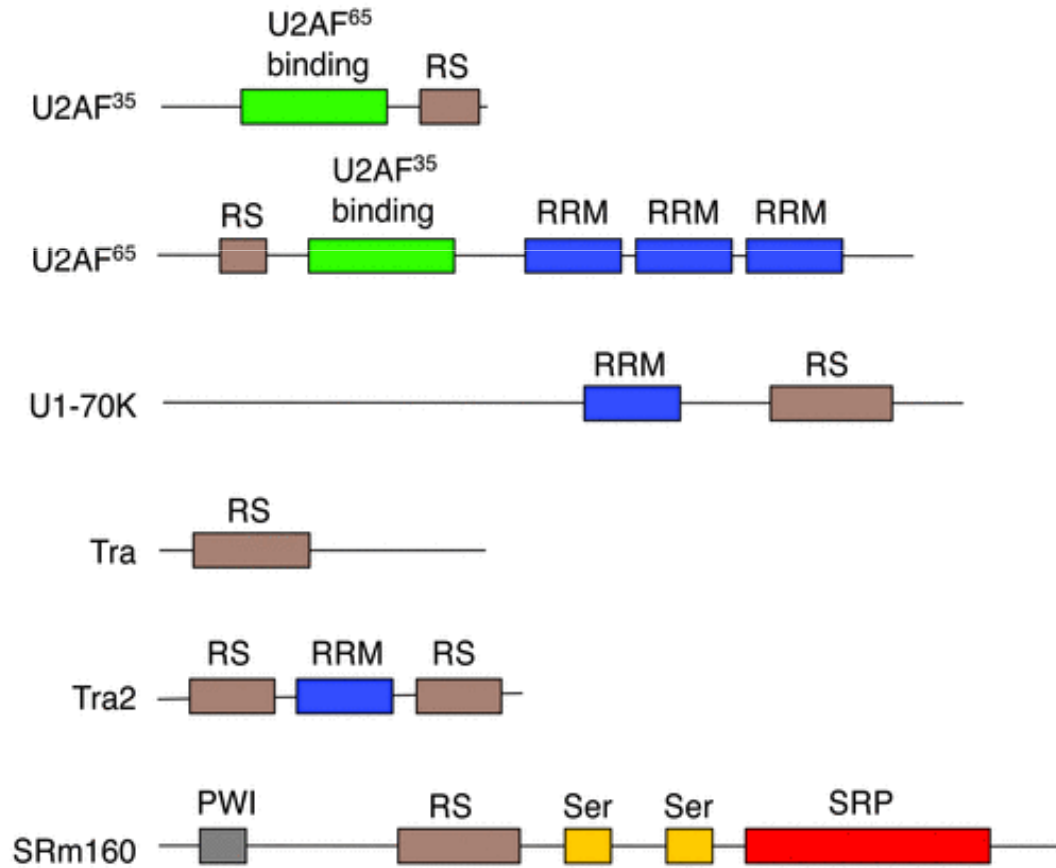
SR proteins also interact with the CAP-binding protein and with poly-A binding proteins.

Useful characteristics: they precipitate with 10 mM Mg⁺⁺, so that it is easy to deplete nuclear extracts of SR factors (e.g. the S100 extract)

SR-family proteins



SR-related proteins



SR proteins display an RS motif accompanied by one or more RRM domain.

Other related proteins possess RS domains.

Name*	Domains	Binding sequence	Target genes
<i>Canonical SR proteins</i>			
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	GGCAGCACCCUG	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
<i>Other SR proteins</i>			
SRp54	RRM and RS	ND	TAU
SRp46 (SFRS2B)	RRM and RS	ND	NA
RNPS1	RRM and Ser-rich	ND	TRA2B
SRp35	RRM and RS	ND	NA
SRp86 (SRp508 and SFRS12)	RRM and RS	ND	NA
TRA2 α	RRM and two Arg-rich	GAAARGARR	dsx
TRA2 β	RRM and two RS	(GAA) _n	SMN1, CD44 and TAU
RBM5	RRM and RS	ND	CD95
CAPER (RBM39)	RRM and RS	ND	VEGF

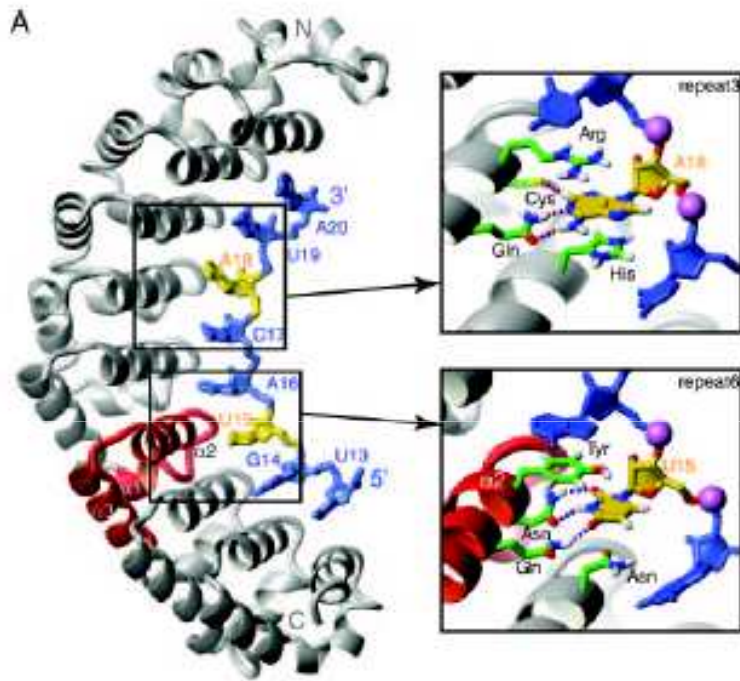
Table 1 | Ribonucleoproteins that are involved in pre-mRNA splicing

Name	Other names	Domains*	Binding sequences	Target genes
hnRNP A1	NA	RRM, RGG and G	UAGGGA/U	SMN2 and RAS
hnRNP A2	NA	RRM, RGG and G	(UUAGGG) _n	HIV tat and IKBKAP
hnRNP B1				
hnRNP C1	AUF1	RRM	U rich	APP
hnRNP C2				
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	DSEF1	RRM, RGG, GYR and GY	GGGA and G rich	PLP, HIV tat and BCL2L1
hnRNP H'				
hnRNP I	PTB	RRM	UCUU and CUCUCU	PTB, nPTB, SRC, CD95, TNTT2, CALCA and GRIN3B
hnRNPL	NA	RRM	C and A rich	NOS and CD45
hnRNPLL	SRRF	RRM	C and A rich	CD45
hnRNPM	NA	RRM and GY	ND	FGFR2
hnRN PQ	NA	RRM and RGG	ND	SMN2

Protein - RNA interaction

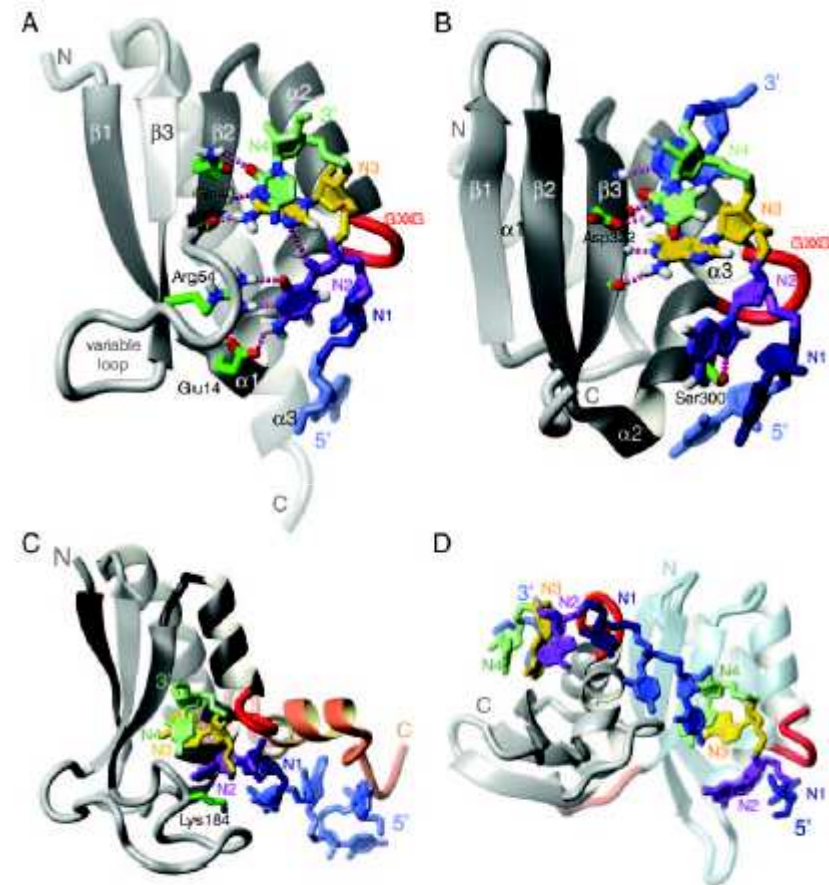
Given a RNA sequence, looking for binding protein (CLIP, EMSA)

Given a protein, looking for binding sequence



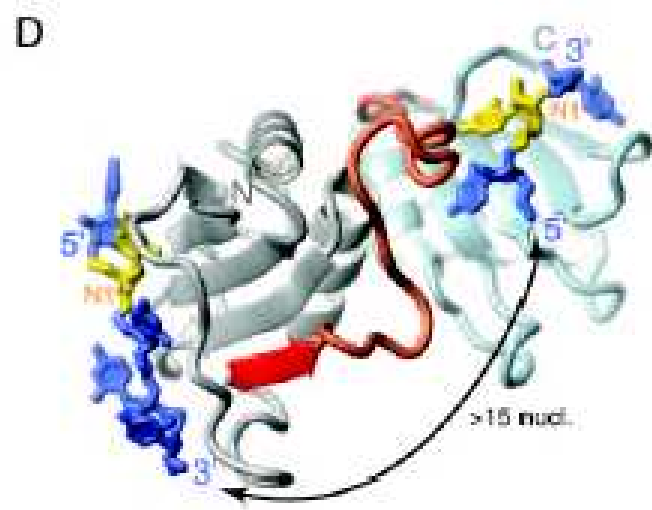
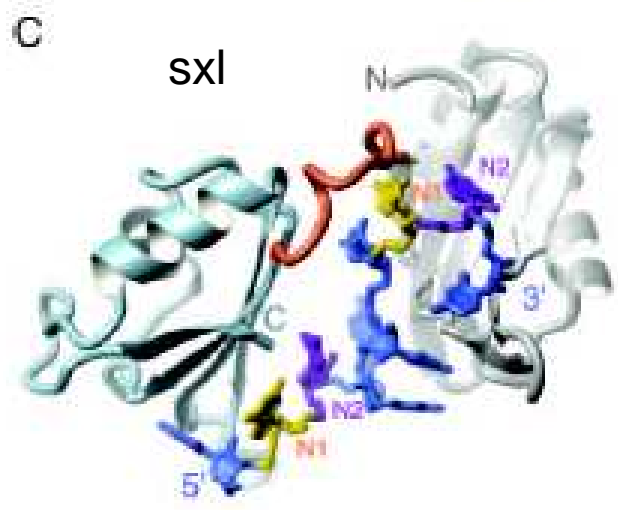
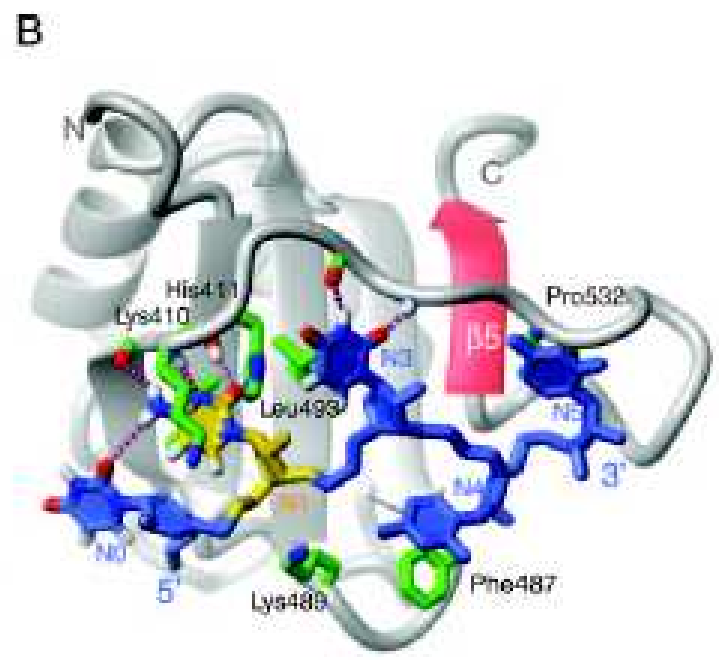
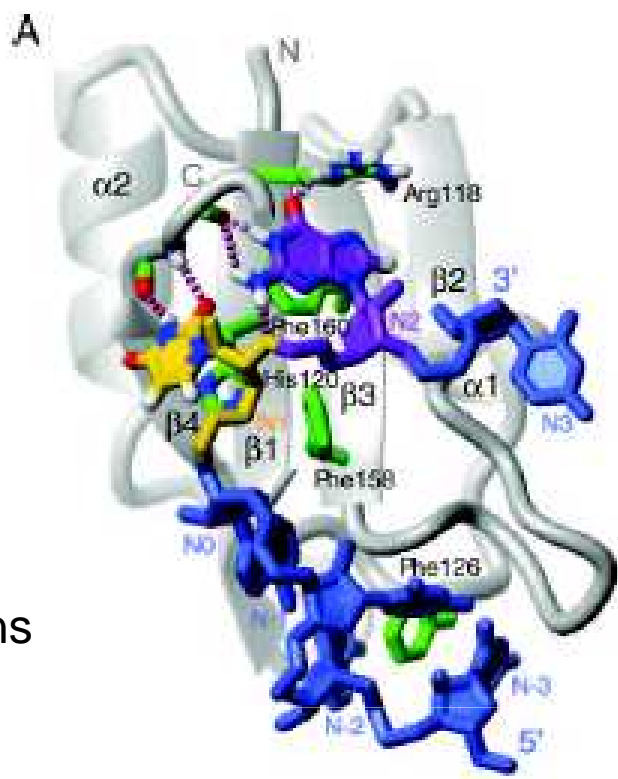
Pumilio repeat domain

Example of pure
sequence recognition



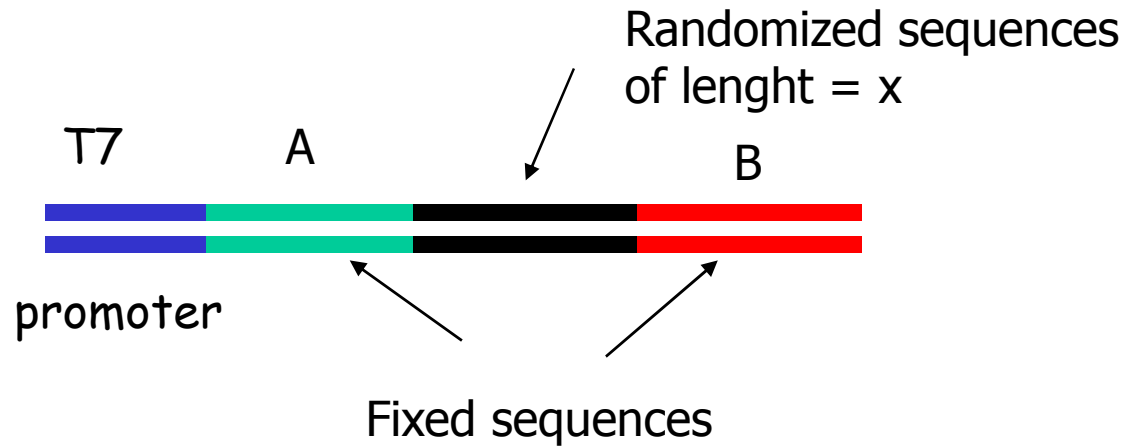
KH domains from hnRNPs

RRM domains

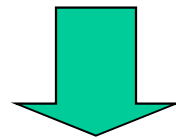


In vitro evolution of molecules (RNA)

RNA - SELEX

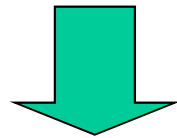


Starting with a pool of randomized DNA sequences ($N=4^x$)

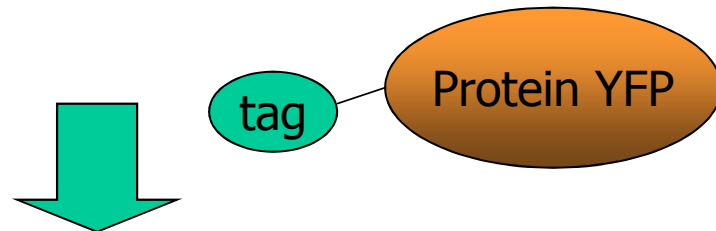


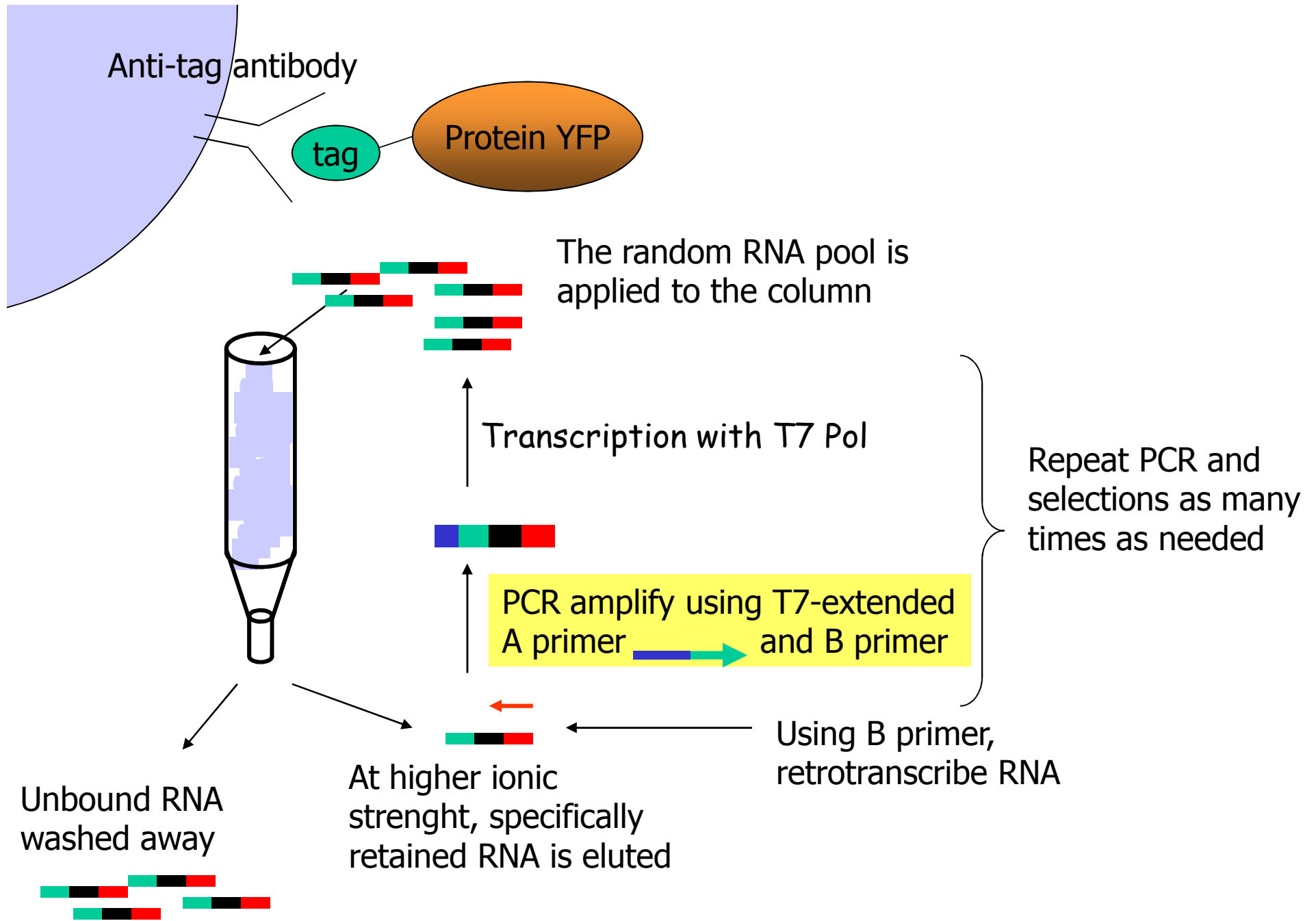
In vitro transcription with T7 RNA Polymerase + NTPs

In vitro transcribed RNA



Tagged YFP immobilized on Sepharose beads





After 6-10 cycles of SELEX, the eluted RNAs are RT, amplified, ligated, cloned and sequenced in series, or (today) directly sequenced, in order to read the sequences selected and their relative frequencies.

Extensive research on several model systems has led to definition of several sequences that regulate alternative splicing:

ESE – Exonic splicing enhancer

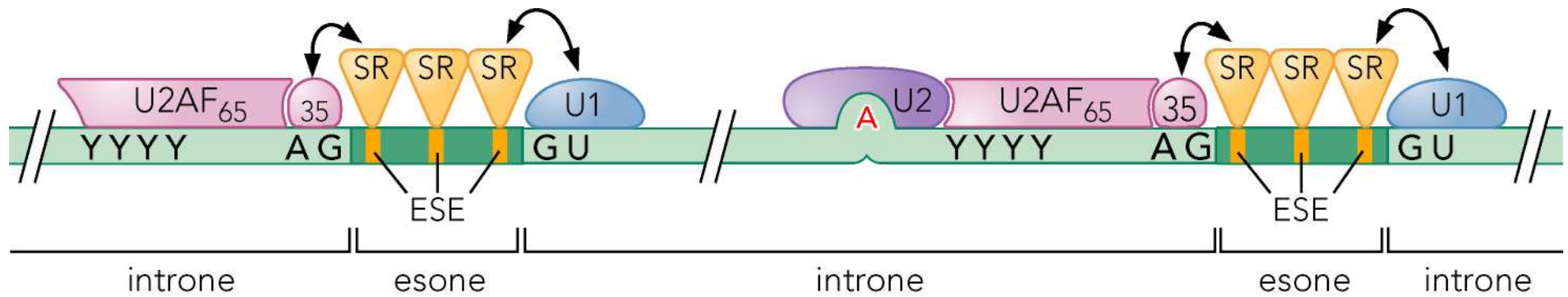
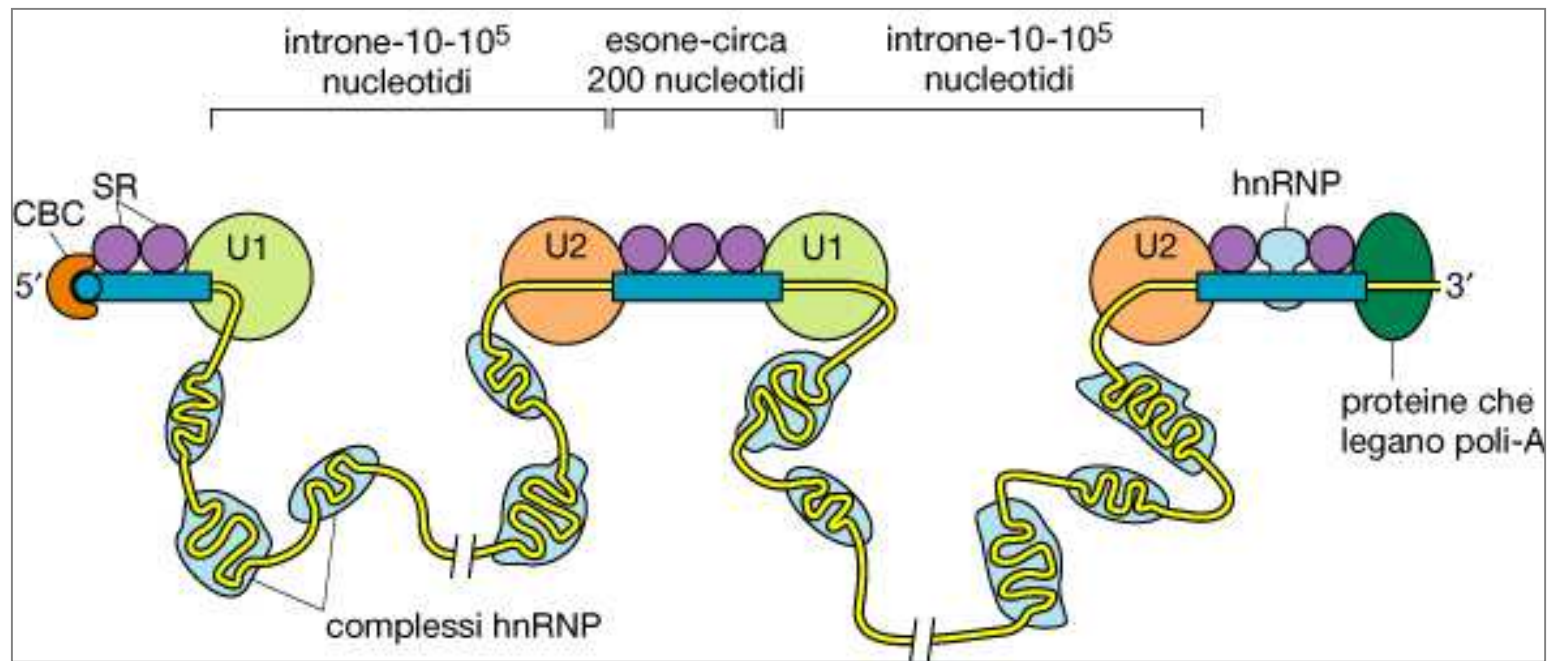
ESS – Exonic splicing silencer

ISE – Intronic splicing enhancer

ISS – Intronic splicing silencer

As a general rule, enhancers interact – directly or indirectly – with SR proteins or related, while silencers generally work through **hnRNPs**.

hnRNP = heterogeneous nuclear ribonucleoproteins



Regulatory sequences are found primarily close to the 5'-ss and 3'-ss i.e. around exons.

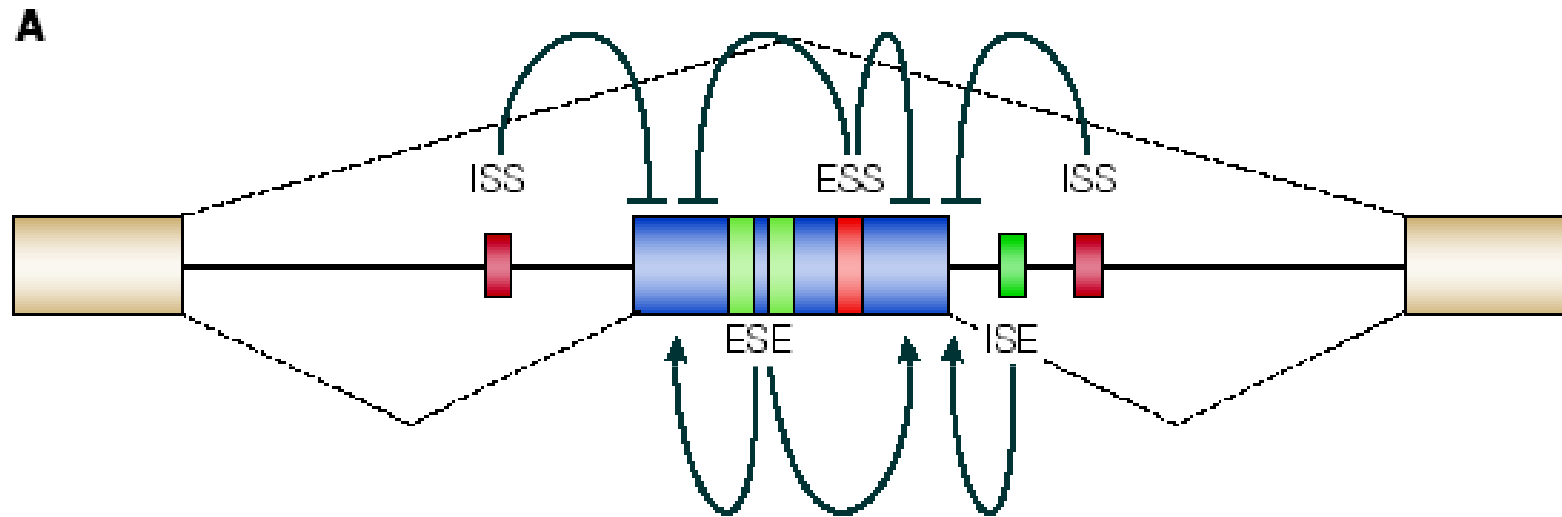
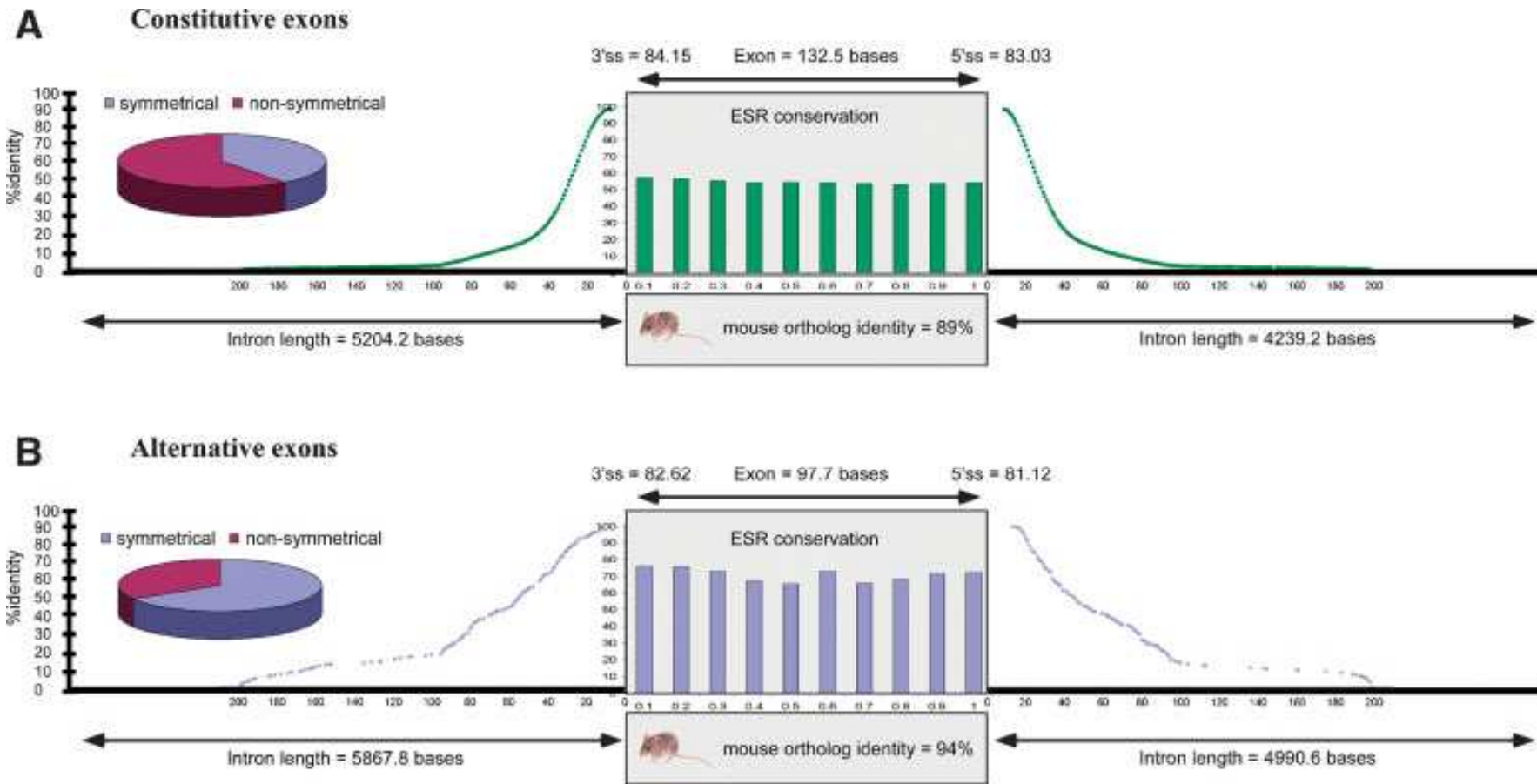


Figure 1 | **Elementary alternative splicing events and regulatory elements. A** | In addition to the splice-site consensus sequences, a number of auxiliary elements can influence alternative splicing. These are categorized by their location and activity as exon splicing enhancers and silencers (ESEs and ESSs) and intron splicing enhancers and silencers (ISEs and ISSs). Enhancers can activate adjacent splice sites or antagonize silencers, whereas silencers can repress splice sites or enhancers. Exon inclusion or skipping is determined by the balance of these competing influences, which in turn might be determined by relative concentrations of the cognate RNA-binding activator and repressor proteins.

From: Matlin et al. (2005), *Nature Rev Mol Cell Biol*, 6: 386.

Recent bioinformatic surveys have shown this very clearly:



from Kim et al., 2007. Bioessays 30:38-47.

Figure 2. Certain characteristics distinguish conserved alternative exons from constitutively spliced ones. The main features that differ between A: constitutively and B: alternatively spliced exons that are conserved in human and mouse are illustrated, namely, exon length, splice site strength, exonic splicing regulatory sequence (ESR) conservation, percent identity between human and mouse, length of flanking introns and their conservation level between human and mouse, and the fraction of symmetrical exons.

Splicing regulatory elements ESE, ISE, ESS, ISS

how do they look like ?

Very difficult to define, poor conservation, superposition with other sequence algorithms, possibly combinatorial interaction with many different RNA-binding proteins

We will go through one of the first paper addressing this question, starting from the known ESE in the dsx gene, then we will move to more “modern” approaches that put together bioinformatics and functional assays to explore the whole genome in search of sequence elements or “motifs” regulating AS.

D- melanogaster

Selection and Characterization of Pre-mRNA Splicing Enhancers: Identification of Novel SR Protein-Specific Enhancer Sequences

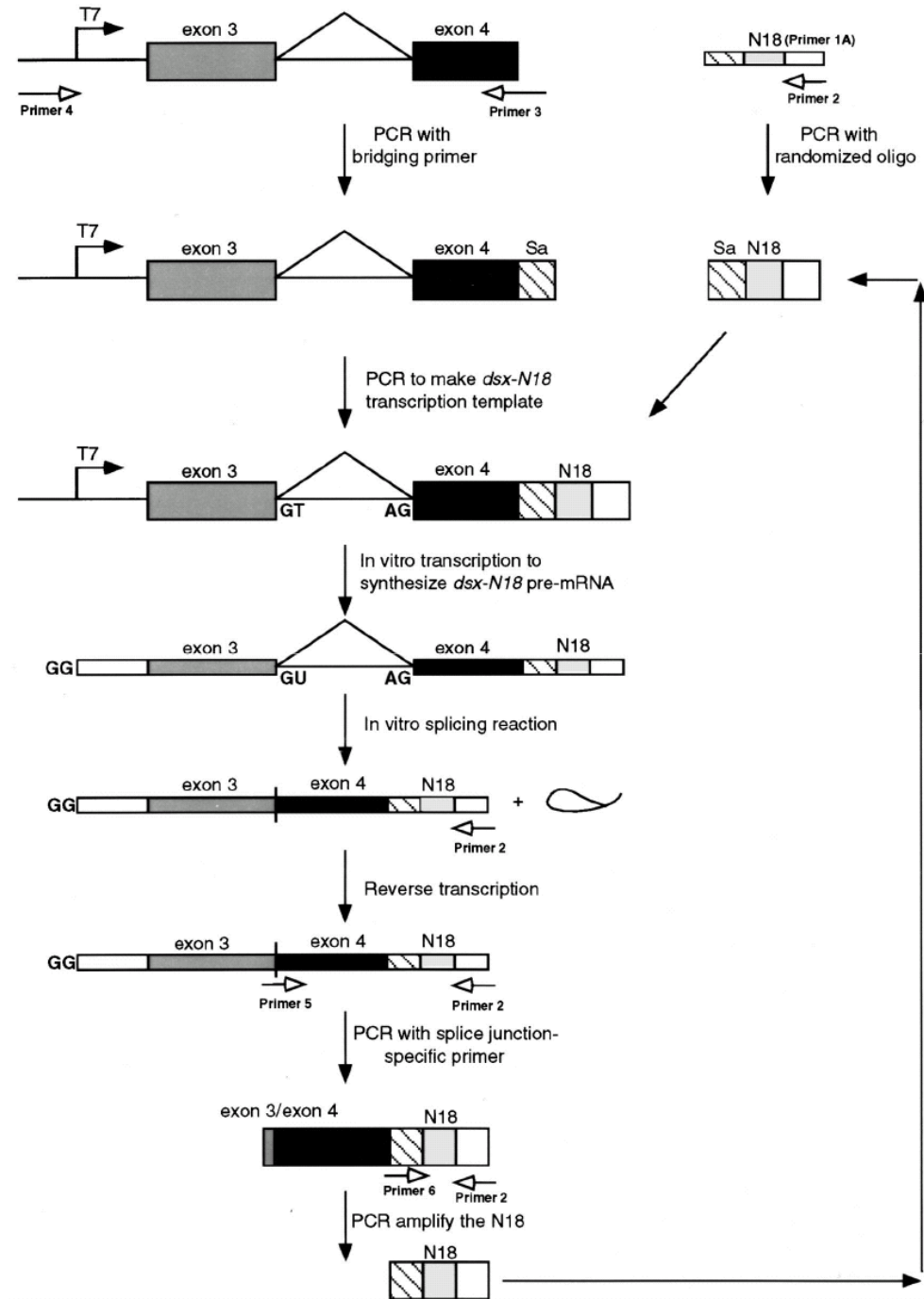
THOMAS D. SCHAAL AND TOM MANIATIS*

Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138

Received 14 September 1998/Returned for modification 28 October 1998/Accepted 23 November 1998

Splicing enhancers are RNA sequences required for accurate splice site recognition and the control of alternative splicing. In this study, we used an in vitro selection procedure to identify and characterize novel RNA sequences capable of functioning as pre-mRNA splicing enhancers. Randomized 18-nucleotide RNA sequences were inserted downstream from a *Drosophila doublesex* pre-mRNA enhancer-dependent splicing substrate. Functional splicing enhancers were then selected by multiple rounds of in vitro splicing in nuclear extracts, reverse transcription, and selective PCR amplification of the spliced products. Characterization of the selected splicing enhancers revealed a highly heterogeneous population of sequences, but we identified six classes of recurring degenerate sequence motifs five to seven nucleotides in length including novel splicing enhancer sequence motifs. Analysis of selected splicing enhancer elements and other enhancers in S100 complementation assays led to the identification of individual enhancers capable of being activated by specific serine/arginine (SR)-rich splicing factors (SC35, 9G8, and SF2/ASF). In addition, a potent splicing enhancer sequence isolated in the selection specifically binds a 20-kDa SR protein. This enhancer sequence has a high level of sequence homology with a recently identified RNA-protein adduct that can be immunoprecipitated with an SRp20-specific antibody. We conclude that distinct classes of selected enhancers are activated by specific SR proteins, but there is considerable sequence degeneracy within each class. The results presented here, in conjunction with previous studies, reveal a remarkably broad spectrum of RNA sequences capable of binding specific SR proteins and/or functioning as SR-specific splicing enhancers.

SELEX procedure for selecting 18-mers with exon splicing enhancer properties



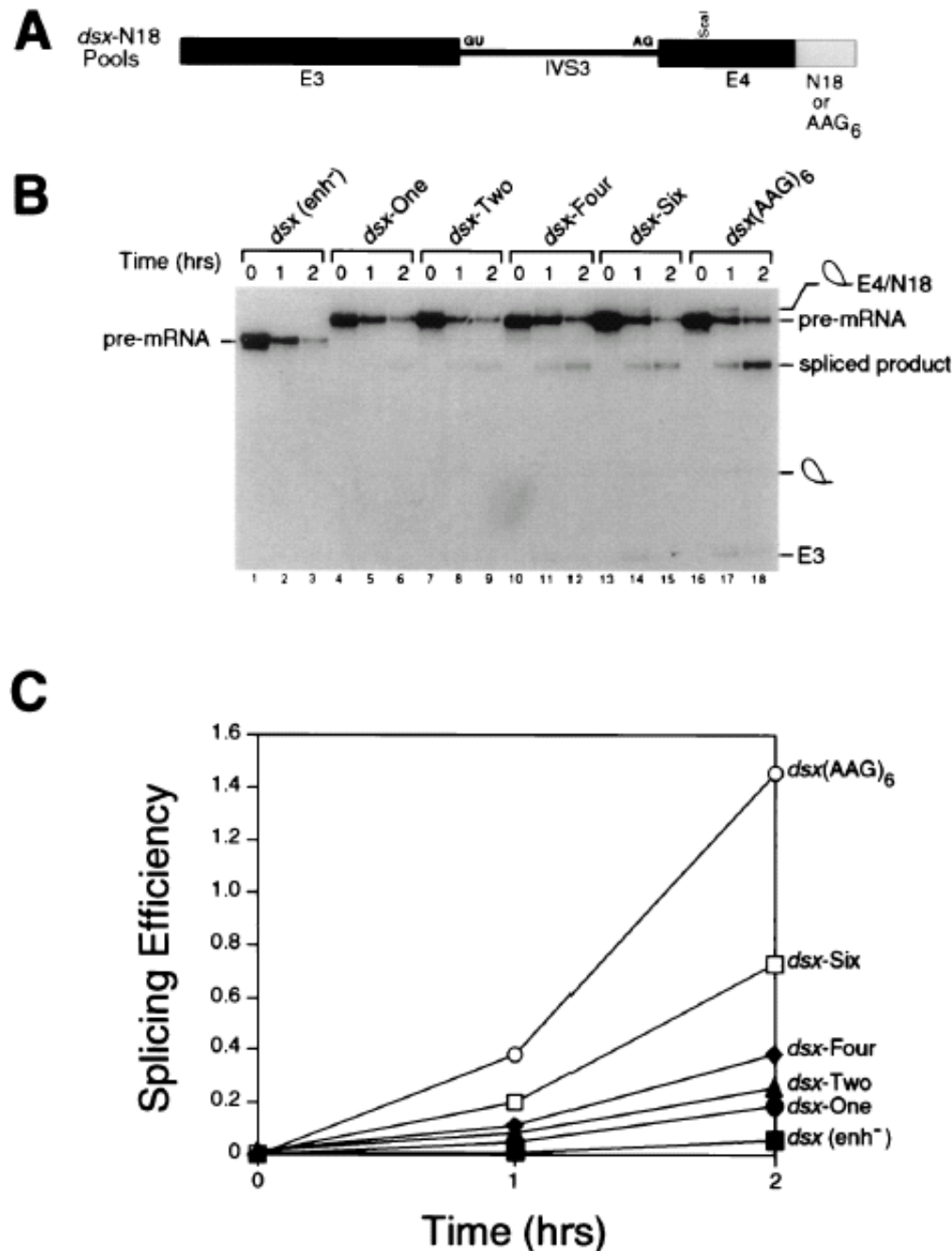


FIG. 2. Evolution of the *dsx-N18* pool. (A) The *dsx-N18* and *dsx-(AAG)*₆ constructs are shown schematically. Exon 3, intron 3, exon 4, and the enhancer(s) are indicated by E3, IVS3, E4, and N18 or AAG₆, respectively. The 59 and 39 splice sites are indicated by GU and AG, respectively. (B) Kinetic analysis showing in vitro splicing assays performed with HeLa cell nuclear extracts and uniformly labeled pre-mRNA splicing substrates comprising the total pool of *dsx-N18* pre-mRNAs after various rounds of the selection (rounds 1, 2, 4, and 6 are shown in lanes 4 to 6, 7 to 9, 10 to 12, and 13 to 15, respectively). The negative control pre-mRNA (lanes 1 to 3) is a *dsx* pre-mRNA lacking an enhancer [*dsx(enh2)*]. The positive control pre-mRNA (lanes 16 to 18) is a *dsx* pre-mRNA activated by six consecutive copies of a multimerized AAG trinucleotide splicing enhancer (modeled after a synthetic polypurine splicing enhancer in reference 66) that is otherwise isogenic to the *dsx-N18* construct. In the kinetic analysis shown, the reaction mixtures were incubated for the number of hours indicated at the top, and positions of the precursors, intermediates, and products of the splicing reaction are indicated to the left and right. The RNAs were analyzed on a 10% denaturing gel in order to resolve the lariat-exon 4 intermediate from the spliced product. (C) Quantitation of the in vitro splicing reactions in panel B. The splicing efficiency (ratio of spliced product to precursor) is calculated from quantitation of individual bands after subtraction of background using a BAS2000 phosphorimager.

TABLE 1. Purine- and pyrimidine-rich sequences in the selected splicing enhancers

Clone no.	Splicing enhancer sequence ^a	Splicing efficiency ^b (%)
Class I, purine-rich enhancers ($\geq 65\%$ purine content)		
Motif A, GGGGA		
3-7	GCAACGGGGACGCGGC	40
3-1	AGCGGUCGCGGUUGGGG gag	32
6-43	GCGGAGGAGGCCCGUGG gag	50
Motif B, GGAGGA		
6-43	GCGGAGGAGGCCCGUGG gag	50
6-19	GCCAGCGGAGGAUGCGG	53
Motif C, GGAGA		
3-35	CUGGAAUACGGAGACCGG	36
6-40	GGUGAGCGGAGAUGCUGC	31
Others		
3-36	GGACCUAGAGGUGGCGAC	40
6-29	GACCGUCGGACAGGAGC	36
Class II, pyrimidine-rich enhancers ($\geq 67\%$ pyrimidine content)		
Motif D, UCUC		
6-13	au CUCC ACGUCGCCUGCUGC	38
6-16	au CUCC ACGUCGCCUGCUGC	37
6-24	UUUGCGG UCUCCGGCCUCC	56
Motif E, UCUUC		
6-5	UGCCACCCGCGG UCUUC	26
6-12	UCGUCG UCUUC GCGGCC	49
3-32	CCUGCUGCG UCUUC GUCC	27
Motif F, UCCUC		
6-7	CCUG UCCUC GUGUUGC	36
6-22	CG UCCUC GUGUACCGCC	37
6-6	GGU UCCUC GCGCCGCC	41
Controls (reference)		
<i>dsx</i> , enhancerless		≤ 1
h β -globin (51)		81
<i>dsx</i> -ASLV (60)		58

TABLE 2. Recurring motifs in selected splicing enhancers and other strong enhancers

Clone	Selected enhancer sequence ^a	Splicing efficiency ^b (%)
Class III, enhancers containing permutations of the sequence (U)GGACCNG		
6-14	GCCGCCGCUUCGUGGACCag	53
6-25	CACGCUCCUCGCUGGACCag	53
6-38	GCCGCCGUGGUGGACCGGag	50
6-26	CCGAGCUACAGGACCGGag	35
6-29	GACCGUCGGACAGGAGC	36
3-35	CUGGAAUACGGAGACCGGag	36
3-36	uGGACCUAGAGGUGGCGAC	40
6-9	uGGACCGCCUGCCAUACC	34
3-3	CAGGCGGGACCGCGACG	17
Class IV, enhancers containing the sequence (C)CACC(C)		
6-28	CCGAGCCACCCGGUACC	29
6-5	UGCCACCCGCGGUCUCC	26
6-2	CGUCGCACCCUGUCUGCC	29
6-22	CGUCCUCGUGUCACCGCC	37
6-35	UCCUGGCGUCACCGUAC	27
Class V, enhancers containing the sequence YGCCGCC		
6-14	uGCCGCCGCUUCGUGGACC	53
6-38	uGCCGCCGUGGUGGACCGG	50
6-45	uGCCGCCGCGAGUUGGGGC	32
6-8	GCCAGUAGUUGCCGCCGC	24
6-6	GGUCCUGUCGCCGCCCC	41
6-1	GGACACCUGUGCGCCGCCag	43
Class VI, enhancers containing the sequence RGAACYU		
3-25	CCACGUGGAACCUUGUCC	35
6-44	ACGGCGCGGAACCUUCC	47
6-23	GCCCGAGAACUUCUUGCC	40
Class VII, other strong enhancers		
6-18	CCGACGCCAUGGACGACGag	55
6-3	GGCUGCCAGUCGGAUUGG	52
6-47	CCGUGACAGCAUCGGCGG	50
3-23	CGUCGGCAGGUGGUCCCG	47
6-39	UCUGGAUCCUGCGGAUGG	44

Exploration of known alternative exons in *Drosophila* genome confirmed the presence of these motifs with variable frequency.

This first work on ESE discovered the first set of sequences showing enhancer function *in vitro* and *in vivo*.

Some of these classes of sequence were successively verified in other organisms as well, up to Mammals.

Other Authors have afforded identification of ESE starting from a pure theoretical point of view:

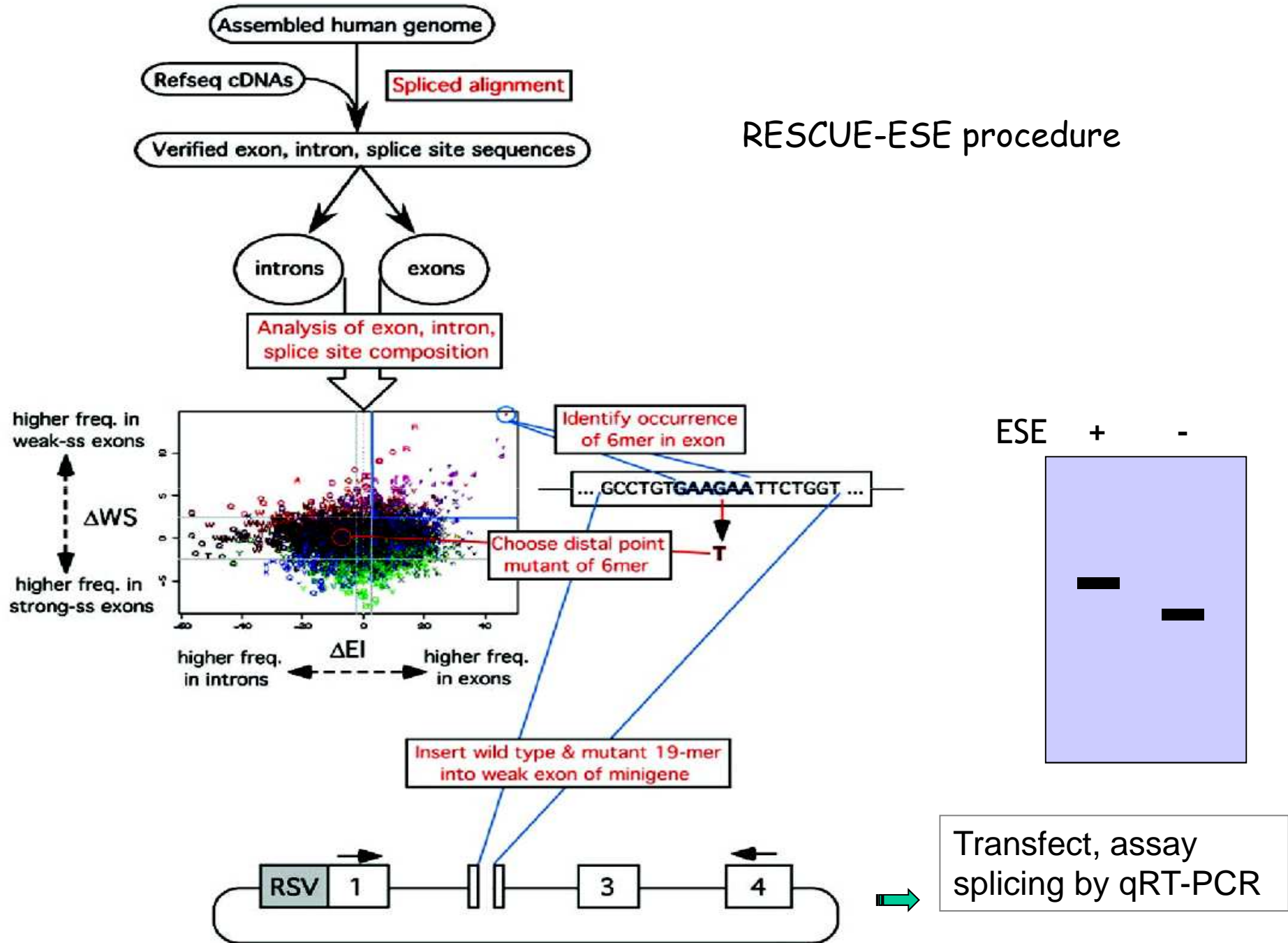
Predictive Identification of Exonic Splicing Enhancers in Human Genes

**William G. Fairbrother,^{1,2*} Ru-Fang Yeh,^{1*} Phillip A. Sharp,^{1,2}
Christopher B. Burge^{1†}**

Specific short oligonucleotide sequences that enhance pre-mRNA splicing when present in exons, termed exonic splicing enhancers (ESEs), play important roles in constitutive and alternative splicing. A computational method, RESCUE-ESE, was developed that predicts which sequences have ESE activity by statistical analysis of exon-intron and splice site composition. When large data sets of human gene sequences were used, this method identified 10 predicted ESE motifs. Representatives of all 10 motifs were found to display enhancer activity *in vivo*, whereas point mutants of these sequences exhibited sharply reduced activity. The motifs identified enable prediction of the splicing phenotypes of exonic mutations in human genes.

Science (2002) 297: 1007-1013.

RESCUE-ESE procedure



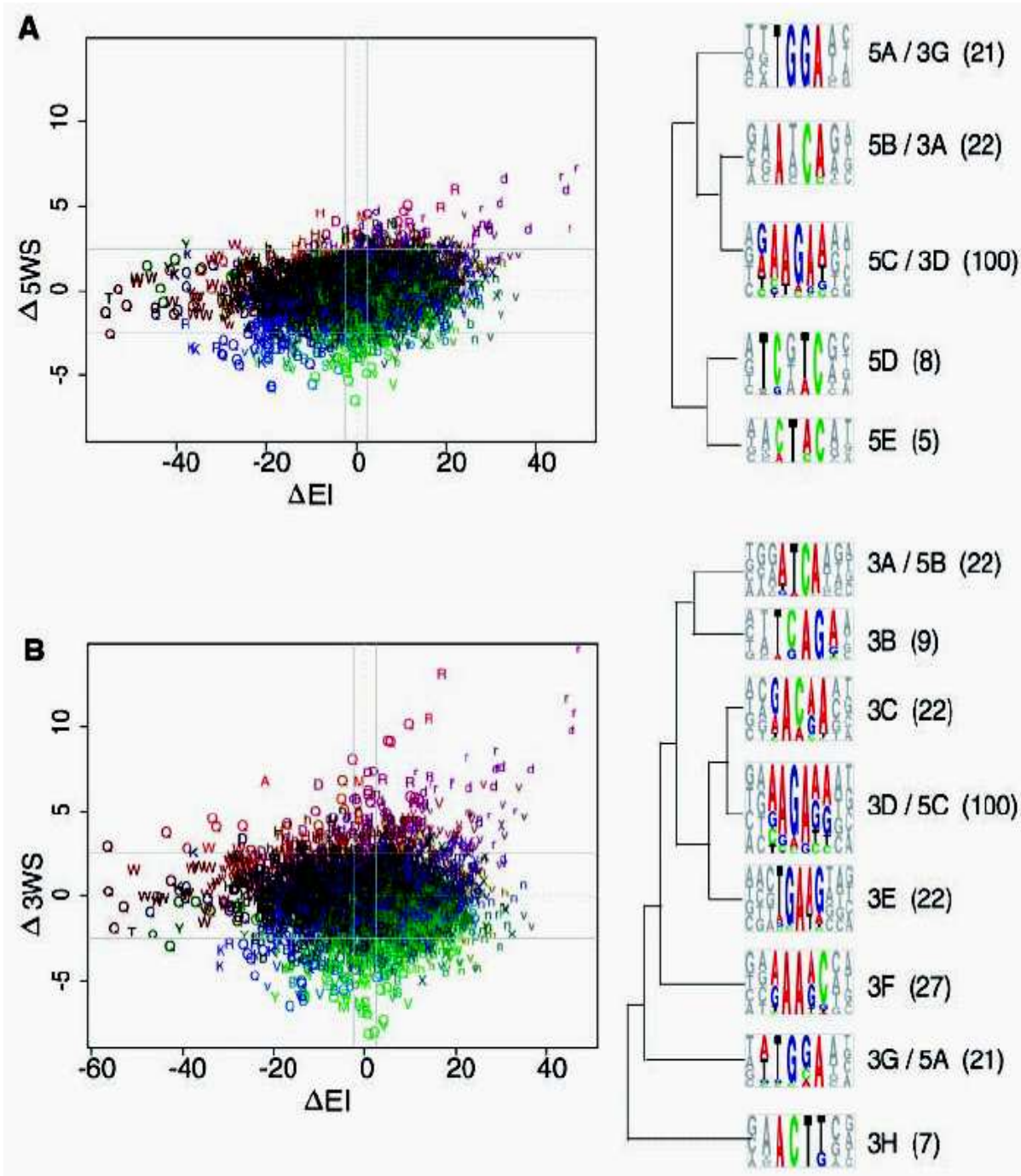


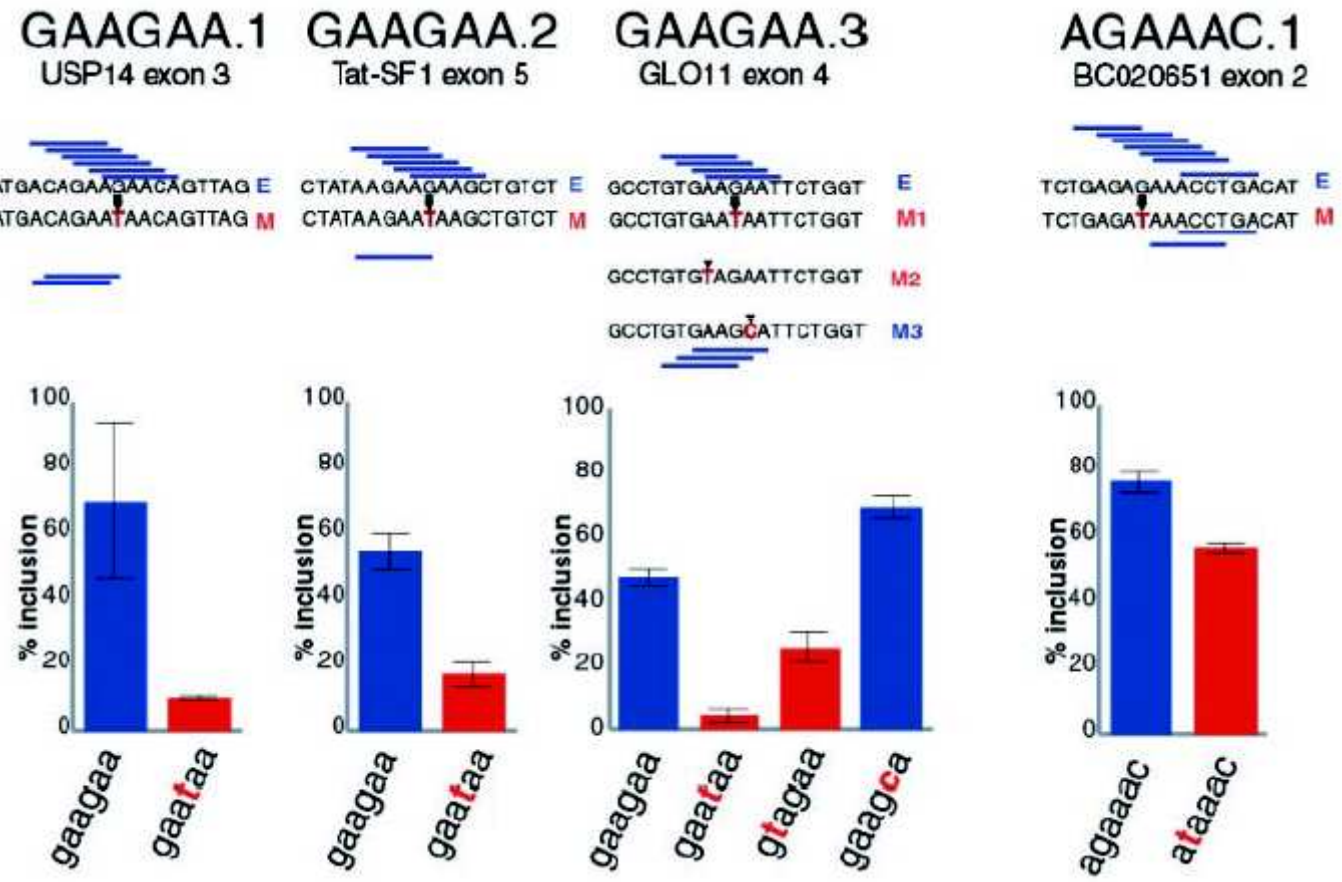
Fig. 2. RESCUE-ESE prediction of 5' and 3' ESEs in human genes.

(A) Scatterplot for prediction of 5'ESE activity. Hexamers are represented by colored letters as described in Fig. 1. Simplified dendrogram shows clustering of 5'ESE hexamers (total of 103 hexamers with $\Delta EI > 2.5$ and $\Delta 5WS > 2.5$) into five clusters of four or more hexamers.

(B) Scatterplot for prediction of 3'ESE activity. Simplified dendrogram shows clustering of 3'ESE hexamers (total of 198 hexamers with $\Delta EI > 2.5$ and $\Delta 3WS > 2.5$) into eight clusters of four or more hexamers. Complete dendrograms of all hexamers are shown in fig. S3. The aligned sequences in each cluster are represented as Pictograms (<http://genes.mit.edu/pictogram.html>).

Cluster labels (e.g., 3B, 5A/3G) are listed to the right of each Pictogram, with the total number of hexamers in the cluster indicated in parentheses.

Sequences taken from representative examples of ESE containing identified “words” were singularly cloned into a reporter vector, together with a mutated version, and tested for splicing in transfected cells. Most of them, indeed, showed ESE function.



Examers + 19-b upstream and 6-b downstream for each “exemplar”

An additional approach is that of addressing directly the biological function.

Cell, Vol. 119, 831–845, December 17, 2004, Copyright ©2004 by Cell Press

Systematic Identification and Analysis of Exonic Splicing Silencers

Zefeng Wang,¹ Michael E. Rolish,^{1,2}
Gene Yeo,^{1,3} Vivian Tung,¹
Matthew Mawson,¹ and Christopher B. Burge^{1,*}

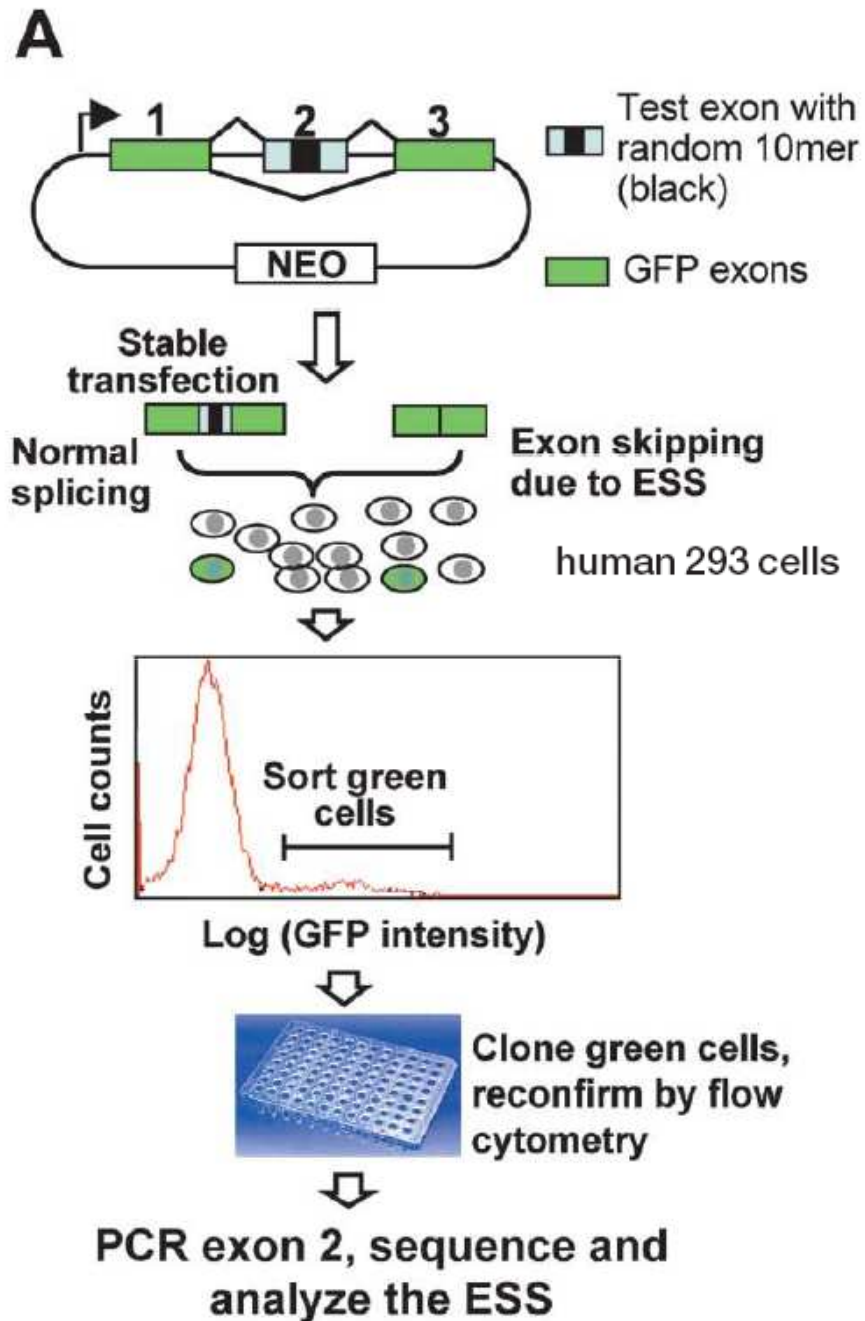
¹Department of Biology

²Department of Electrical Engineering
and Computer Science

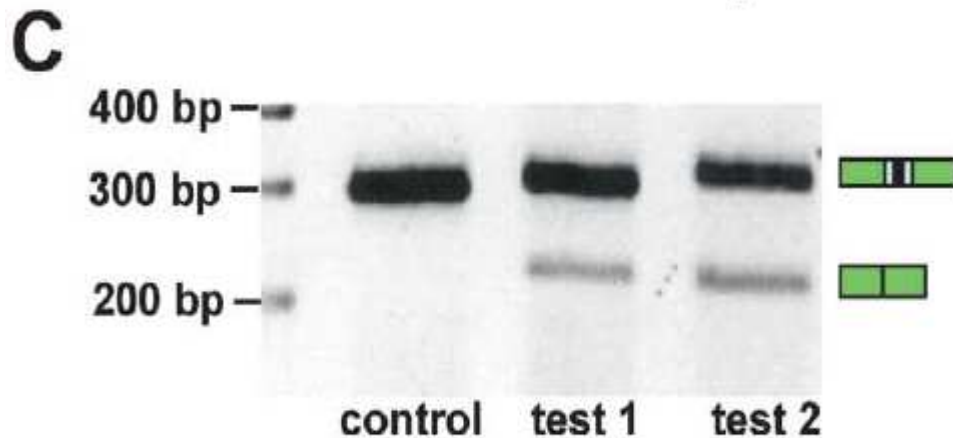
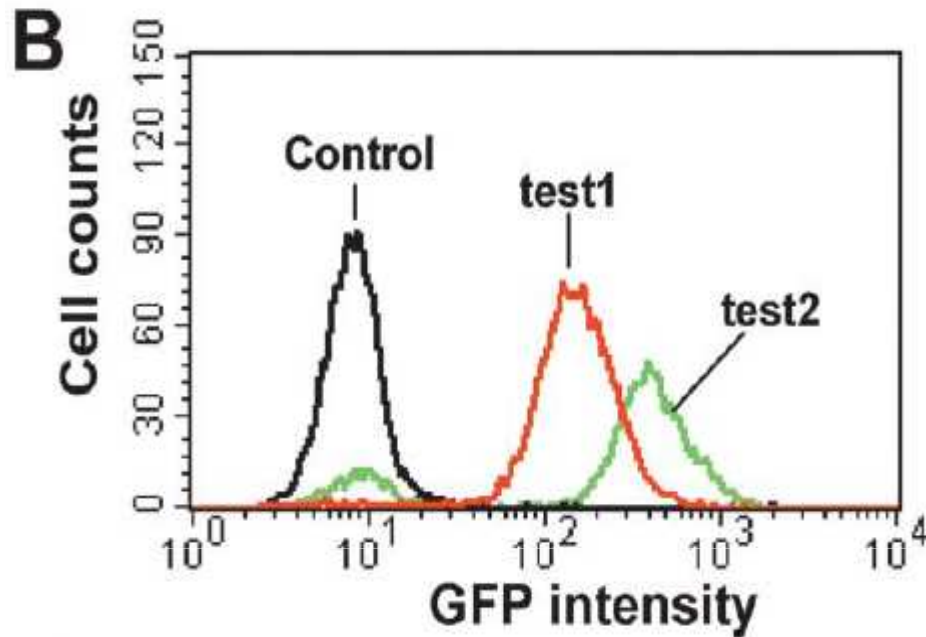
³Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Summary

Exonic splicing silencers (ESSs) are *cis*-regulatory elements that inhibit the use of adjacent splice sites, often contributing to alternative splicing (AS). To systematically identify ESSs, an *in vivo* splicing reporter system was developed to screen a library of random decanucleotides. The screen yielded 141 ESS decamers, 133 of which were unique. The silencer activity of over a dozen of these sequences was also confirmed in a heterologous exon/intron context and in a second cell type. Of the unique ESS decamers, most could be clustered into groups to yield seven putative ESS motifs, some resembling known motifs bound by hnRNPs H and A1. Potential roles of ESSs in constitutive splicing were explored using an algorithm, ExonScan, which simulates splicing based on known or putative splicing-related motifs. ExonScan and related bioinformatic analyses suggest that these ESS motifs play important roles in suppression of pseudoexons, in splice site definition, and in AS.



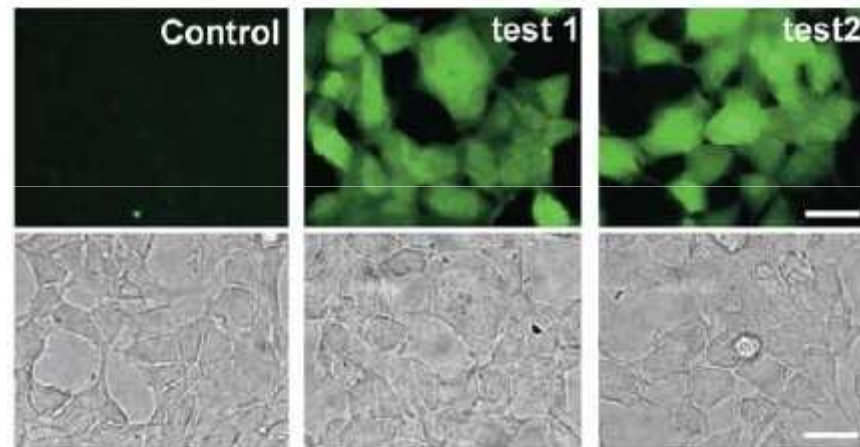
exon 2 of the Chinese hamster dihydrofolate reductase (*DHFR*) gene, was used as the test exon.



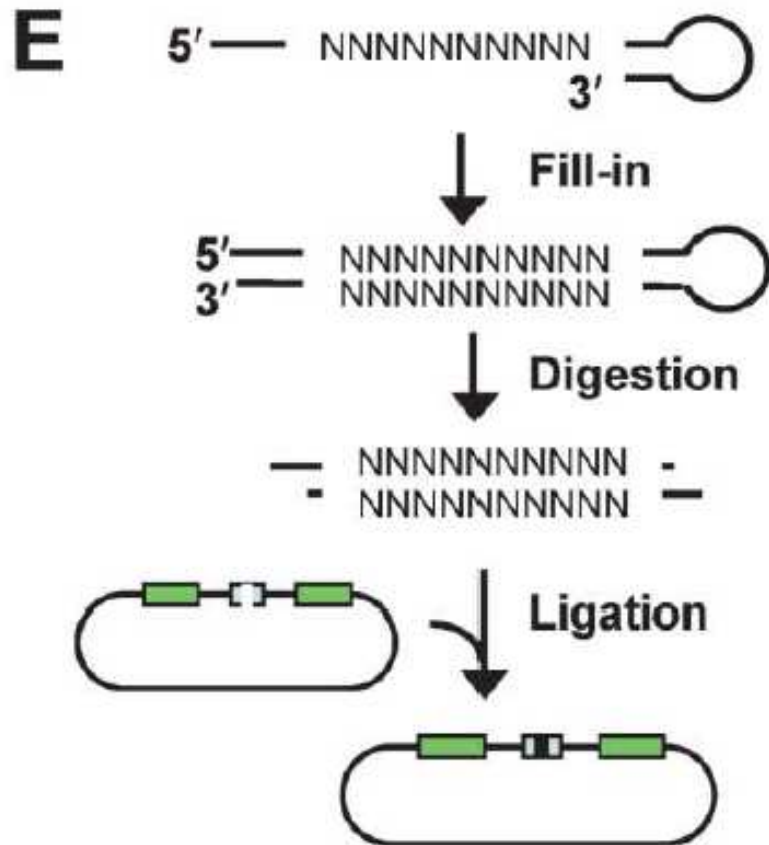
(B) Test of the reporter system with two known ESS sequences. Test 1 (hnRNP A1 binding site, ATGATAG GGACTTAGGGT [Burd and Dreyfuss, 1994]) and test 2 (U2AF65 binding site, TTTTTTTTCCTTTT TTTTCCTTTT [Singh et al., 1995]) were inserted into the pZW4 reporter construct and transfected into 293 Flp-In cells, and positive transfectants were pooled for flow cytometry. The “Control” was a randomly chosen 10-mer sequence (ACCTCAGGCG) inserted into the same vector.

(C) RT-PCR results using RNA purified from the transfected cells as template, with primers targeted to exons 1 and 3 of pZW4.

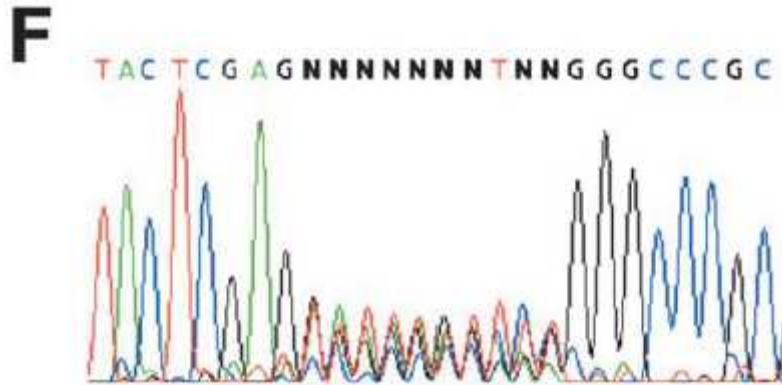
D



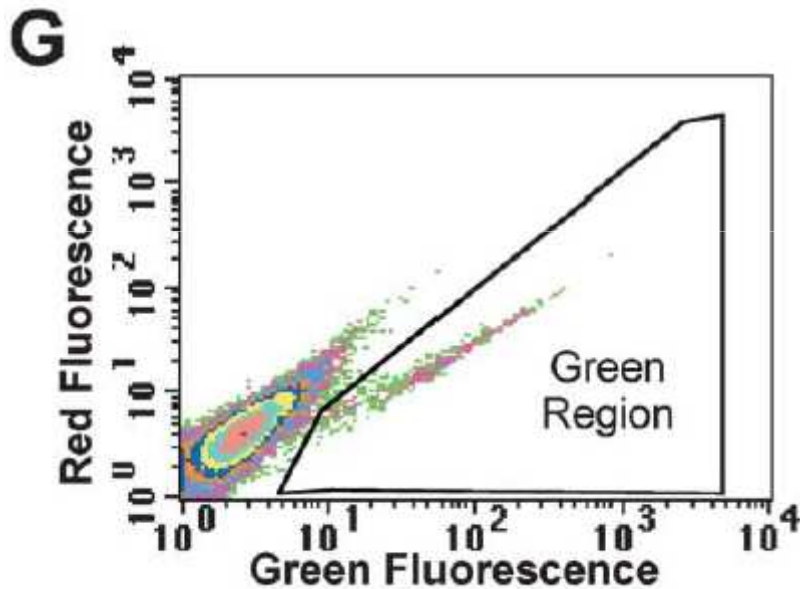
D) Microscopic images of transfected cells. Upper panel, GFP fluorescence. Lower panel, phase images. Scale bar, 50 μ M.



(E) Construction of **random decamer** library. The foldback primer was synthesized with a random sequence of 10 bp, then extended with Klenow fragment, digested, ligated into pZW4, and transformed into *E. coli*.



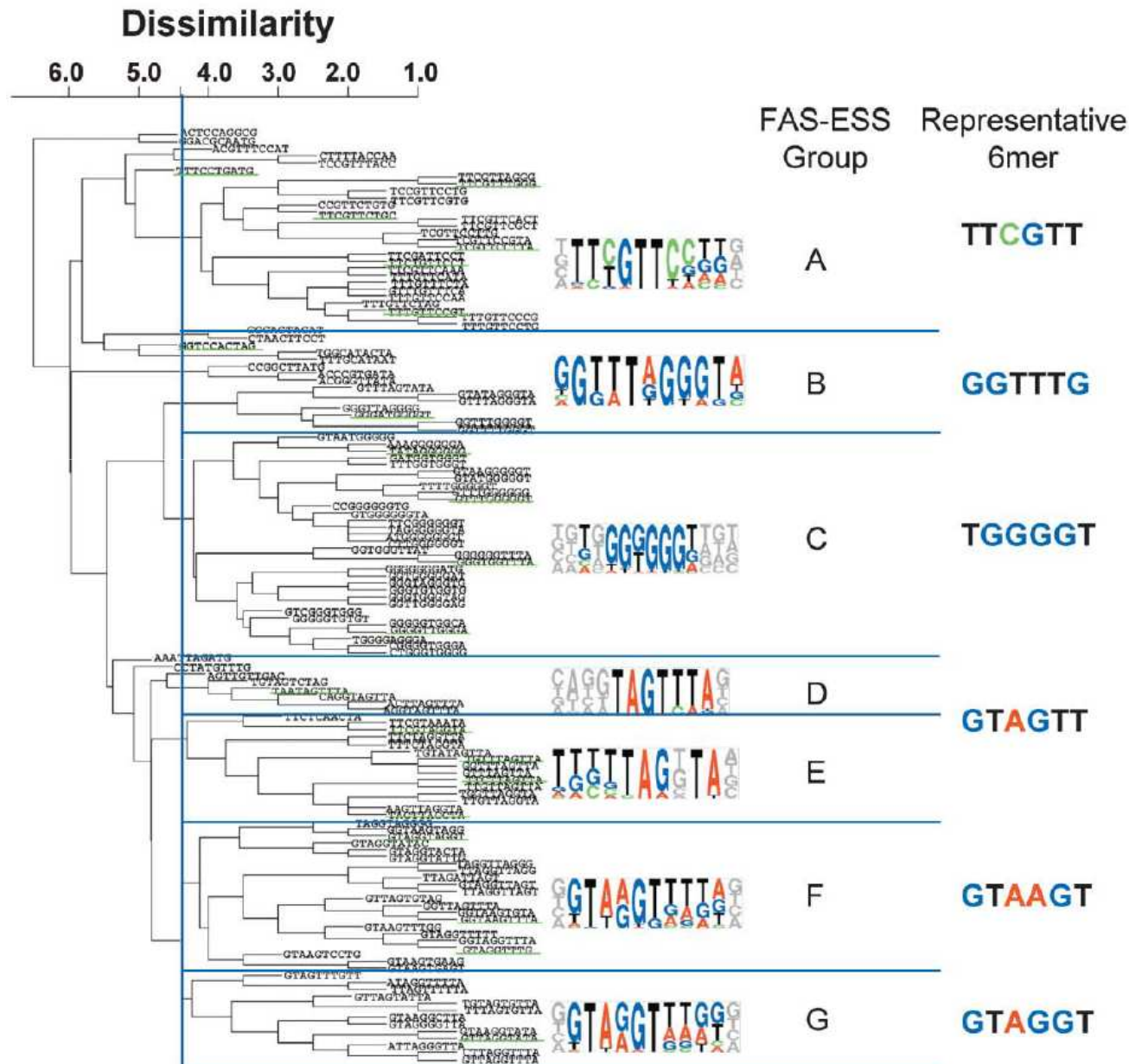
(F) Sequencing of the random decamer region. 293 cells stably transfected with the pZW4 library were pooled to purify total DNAs, from which minigene fragments were amplified by PCR and sequenced. Sequences around the insertion region are shown.



(G) Flow cytometry profile of single transfection using pZW4 random decamer library.

Sequences in the green cell clones, cloned and sequenced

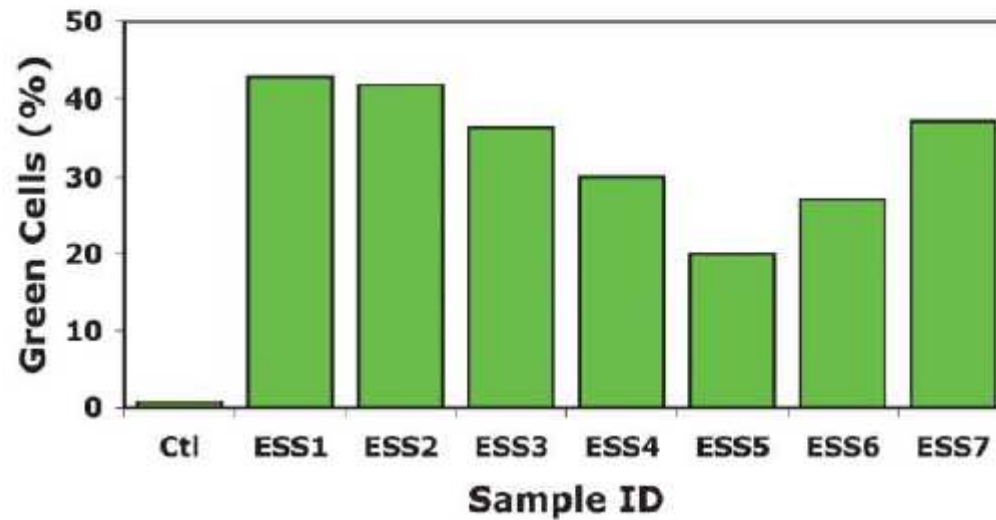
The ca. 100 sequences found were clustered using a similarity algorithm to give 6 classes of consensus sequence:



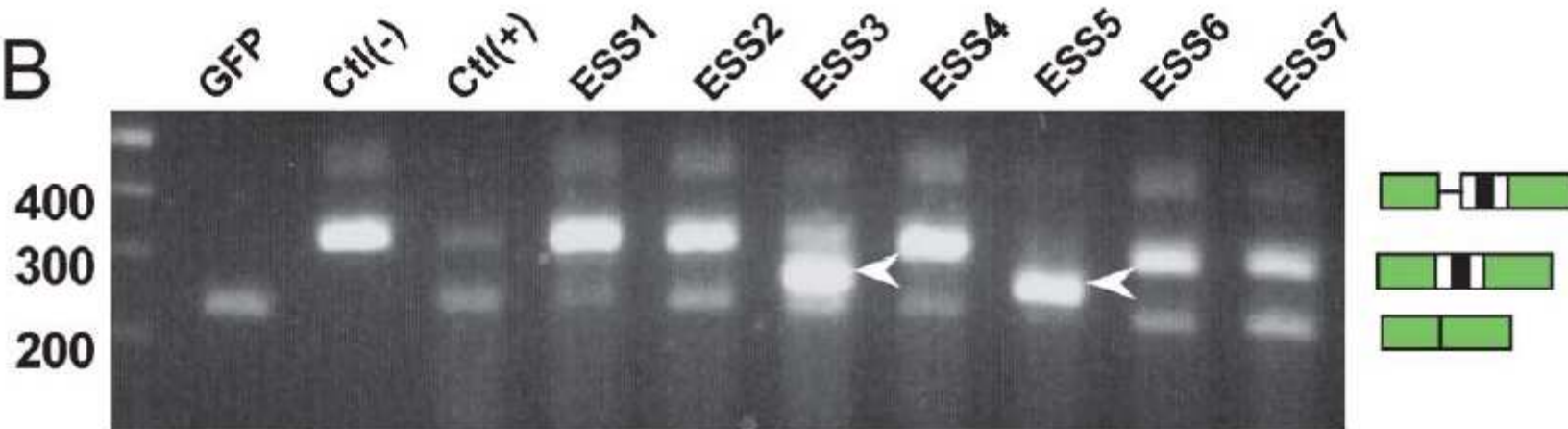
The 6 representative exons were inserted into reporters and splicing analyzed

A

ESS1 **TTTGTTCCGT**
ESS2 **GGGTGGTTTA**
ESS3 **GTAG|GTAGGT**
ESS4 **TTCGTTCTGC**
ESS5 **G|GTAAGTAGG**
ESS6 **GGTTAGTTTA**
ESS7 **TTCGTAGGTA**



B



These 6 classes of functionally derived ESE sequences are extremely frequent in alternative exons in H. sapiens and other genomes.

The number of ESE, ESS, ISE and ISS found is not really very large.

Especially in the light of diffuse “tissue-specific” regulation, the problem of alternative splicing regulation is difficult to solve with SR proteins and hnRNP alone, interacting with a limited number of RNA sequences. Combinatorial regulation can be taken into account, but the relatively low number of SR+hnRNP and their almost ubiquitous expression make it difficult to make up a model.

Tissue-specific splicing regulators, interacting with additional RNA sequences, are an attractive model. However, despite many efforts, relatively few of these have been identified.

The following Table reviews data available (to 2009).

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	<i>BIN1</i> , <i>GLYRA2</i> , <i>ATP2B1</i> , <i>MEF2</i> , <i>NASP</i> , <i>SPAG9</i> and <i>SRC</i>
NOVA1	NA	KH	YCA Y	Neurons of the hindbrain and spinal cord	<i>GABRG2</i> , <i>GLYRA2</i> and <i>NOVA1</i>
NOVA2	NA	KH	YCA Y	Neurons of the cortex, hippocampus and dorsal spinal cord	<i>KCNJ</i> , <i>APLP2</i> , <i>GPHN</i> , <i>JNK2</i> , <i>NEO</i> , <i>GRIN1</i> and <i>PLCB4</i>
FOX1	A2BP1	RRM	(U)GCAUG	Muscle, heart and neurons	<i>ACTN</i> , <i>EWSR1</i> , <i>FGFR2</i> , <i>FN1</i> and <i>SRC</i>
FOX2	RBM9	RRM	(U)GCAUG	Muscle, heart and neurons	<i>EWS</i> , <i>FGFR2</i> , <i>FN1</i> and <i>SRC</i>
RBM35a	ESRP1	RRM	GU rich	Epithelial cells	<i>FGFR2</i> , <i>CD44</i> , <i>CTNND1</i> and <i>ENAH</i>
RBM35b	ESRP2	RRM	GU rich	Epithelial cells	<i>FGFR2</i> , <i>CD44</i> , <i>CTNND1</i> and <i>ENAH</i>
TIA1	mTIA1	RRM	U rich	Brain, spleen and testes	<i>MYPT1</i> , <i>CD95</i> , <i>CALCA</i> , <i>FGFR2</i> , <i>TIAR</i> , <i>IL8</i> , <i>VEGF</i> , <i>NF1</i> and <i>COL2A1</i>
TIAR	TIAL1 and mTIAR	RRM	U rich	Brain, spleen, lung, liver and testes	<i>TIA1</i> , <i>CALCA</i> , <i>TIAR</i> , <i>NF1</i> and <i>CD95</i>
SLM2	KHDRBS3 and TSTAR	KH	UAAA	Brain, tests and heart	<i>CD44</i> and <i>VEGFA</i>
Quaking	QK and QKL	KH	ACUAA Y[...]UAAY	Brain	<i>MAG</i> and <i>PLP</i>
HUB	HUC, HUD and ELAV2	RRM	AU rich	Neurons	<i>CALCA</i> , <i>CD95</i> and <i>NF1</i>
MBNL	NA	CCCH zinc finger domain	YGCU(U/G)Y	Muscles, uterus and ovaries	<i>TNNT2</i> , <i>INSR</i> , <i>CLCN1</i> and <i>TNNT3</i>
CELF1	BRUNOL2	RRM	U and G rich	Brain	<i>TNNT2</i> and <i>INSR</i>
ETR3	CELF2 and BRUNOL3	RRM	U and G rich	Heart, skeletal muscle and brain	<i>TNNT2</i> , <i>TAU</i> and <i>COX2</i>
CELF4	BRUNOL4	RRM	U and G rich	Muscle	<i>MTMR1</i> and <i>TNNT2</i>
CELF5	BRUNOL5 and NAPOR	RRM	U and G rich	Heart, skeletal muscle and brain	<i>ACTN</i> , <i>TNNT2</i> and <i>GRIN1</i>
CELF6	BRUNOL6	RRM	U and G rich	Kidney, brain and testes	<i>TNNT2</i>

A2BP1, ataxin 2-binding protein 1; *ACTN*, α -actinin; *APLP2*, amyloid- β precursor-like protein 2; *ATP2B1*, ATPase, Ca²⁺ 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, $\alpha 1$; *COX2*, cytochrome c oxidase II; *CTNND1*, catenin $\delta 1$; *EWSR1*, Ewing sarcoma breakpoint region 1; *FGFR2*, fibroblast growth factor receptor 2; *FN1*, fibronectin 1; *GABRG2*, GABA A receptor, $\gamma 2$; *GLYRA2*, glycine receptor, $\alpha 2$ 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 $\beta 4$; *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; *TNNT2*, troponin T type 2; *VEGF*, vascular endothelial growth factor.