One gene more products

- CAGE/SAGE, RNA-Seq and tiling arrays have revealed an unexpected number of protein-coding and noncoding transcripts.
- Focusing on protein-coding genes, we see a surprising veriety of RNAs from the same gene, and also a variety of mature, functional mRNAs.

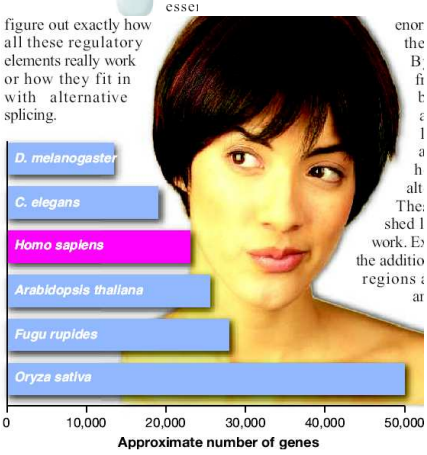
1 JULY 2005 VOL 309 SCIENCE www.sciencemag.org
Published by AAAS

WHAT DON'T WE KNOW?

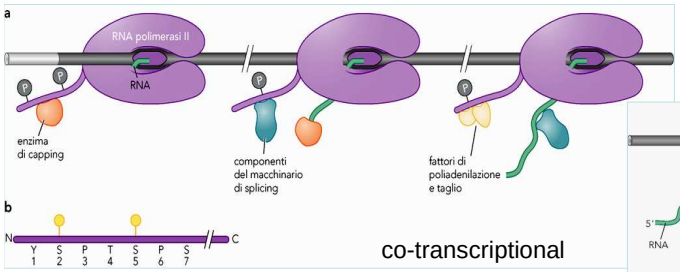
Why Do Humans Have So Few Genes

When leading biologists were unraveling the sequence of the human genome in the late 1990s, proteins. But how the tr ery decides which par any particular time is st

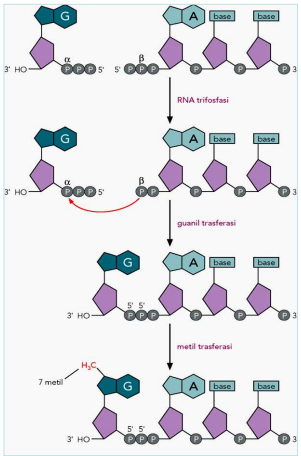
figure out exactly how all these regulatory elements really work or how they fit in with alternative splicing.



. I concetti generali di "processamento" dei trascritti
. I meccanismi dello splicing
. I concetti di base di splicing alternativo



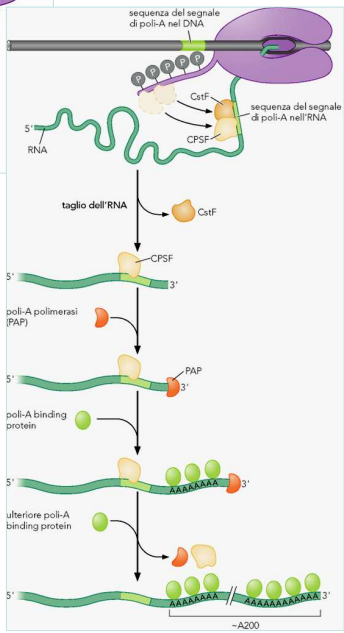
co-transcriptional

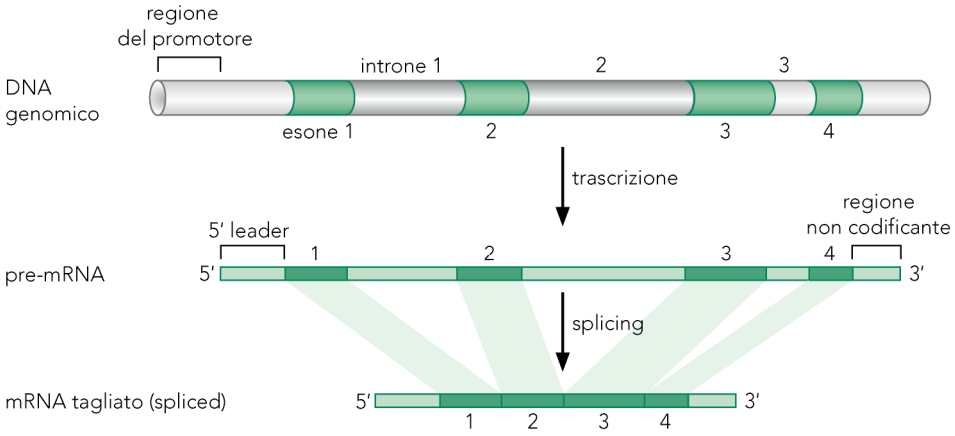


← capping

and

polyadenylation→





How prevalent is splicing (i.e. exon-intron gene organization) in different organisms ?

S. cerevisiae	has only 253 introns (3% of genes), only 6 genes have 2 introns.	(40-75 nt)
S. pombe	43% of the genes have introns, many of them contains >1 intron	
H. sapiens	>99% of genes contain multiple introns	

Average human gene:

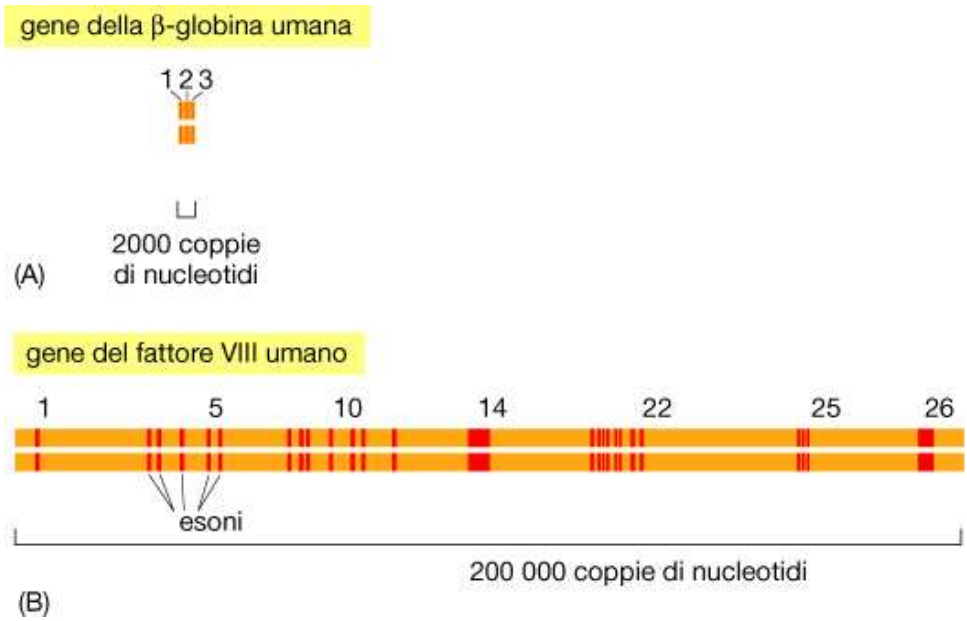
Length: 28,000 bp

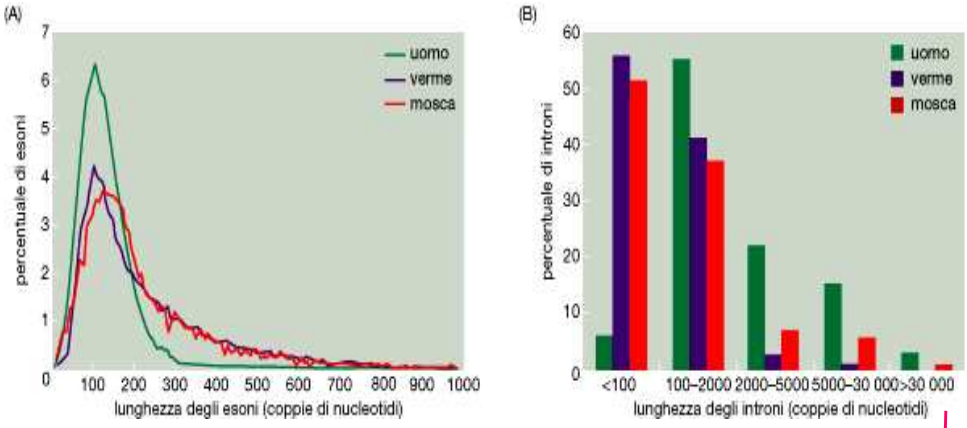
No. of exons: 8.8

Exon length: 120 bp

No. of introns: 7.8

Intron length: 10 to >100.000 bp





one intron in the human neurexin gene is approx. 480,000 nt !

Biochemical mechanism:

absolutely conserved in all organisms, derives from the group II autocatalytic introns (fungi organelles).

RNA and proteins in spliceosome:

quite conserved, but the complexity increases

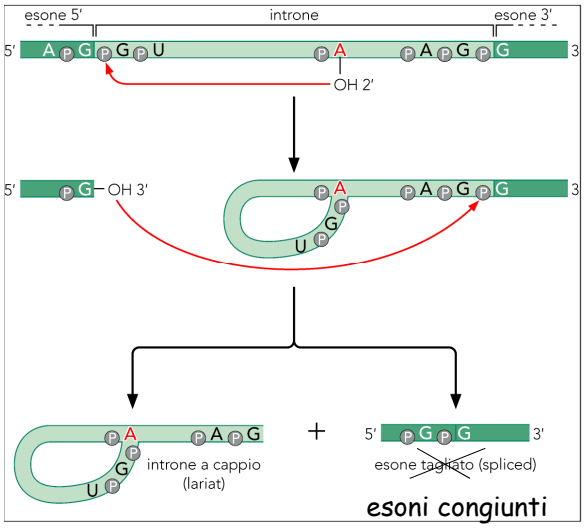
cis-elements in introns and exons defining splice sites:

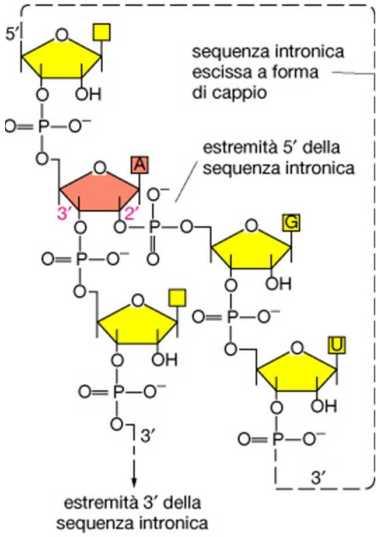
quite conserved, but with differences and increasing complexity

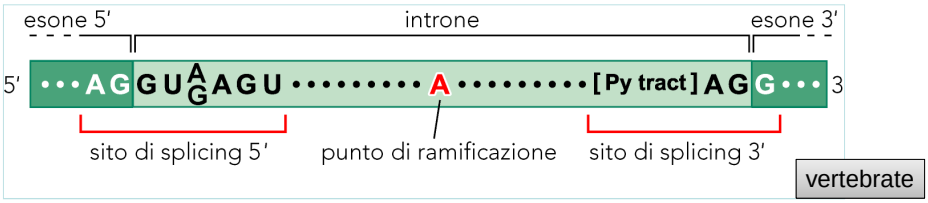
Pre-mRNA splicing occurs in two ATP-Independent transesterification reactions

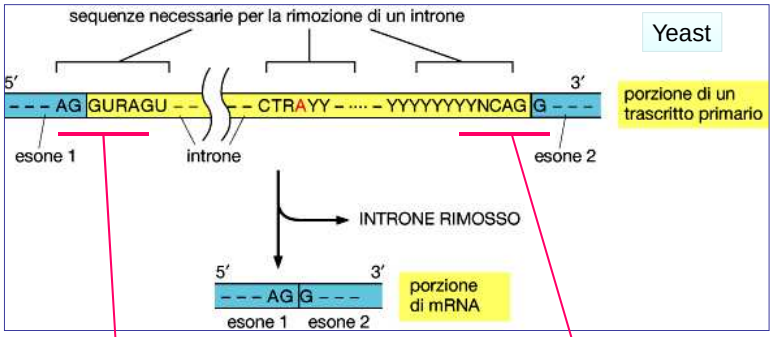
A) first transesterification

B) second transesterification









From: Ast G. (2004)
"How did alternative splicing evolve?"
Nature Rev Gen 5: 773-782.

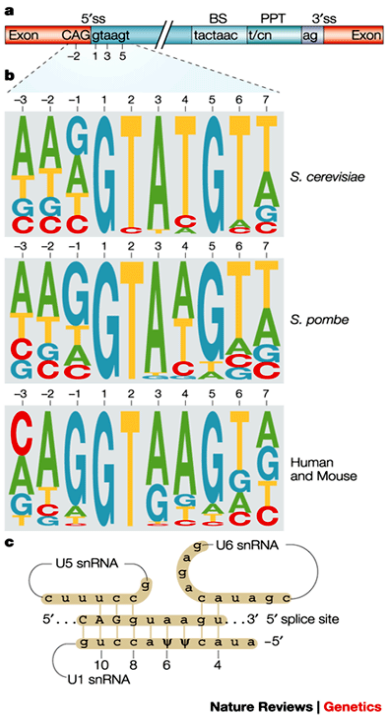
BS (branch site) in S. cerevisiae is very conserved (5'-UACUAAC-3') in S. pombe is much more variable, as in mammals (5'-CURAY-3')

Polypyrimidine tract is also variable, as well as the distance between BS and 3'-ss. (distance very short in S. pombe).

5'-ss has also major differences.

S.cerevisiae introns have 6 nt well conserved: 5'-GTATGT-3'

the -1 "G" increases in frequency from S.cerevisiae (37%) to humans (80%), positions +1 to +6 degenerate a little.



Spliceosome

snRNP = small nuclear ribonucleoproteins

RNA U + 8-12 peptides

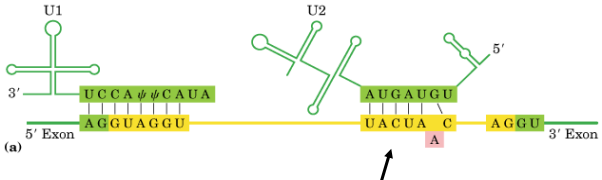
protein	U1	U2	U5	U4/U6
B,B',D,D',E,F,G	+	+	+	+
70k, A, C	+			
25k, 1BP			+	
RNA nt	164	187	116	145+106

Al riconoscimento delle sequenze di RNA necessarie allo splicing (i.e. 5'-ss, 3'-ss, branching point) partecipano sia interazioni RNA-RNA, sia proteine-RNA

Le proteine che interagiscono con l'RNA in modo specifico possiedono domini tipici, chiamati:

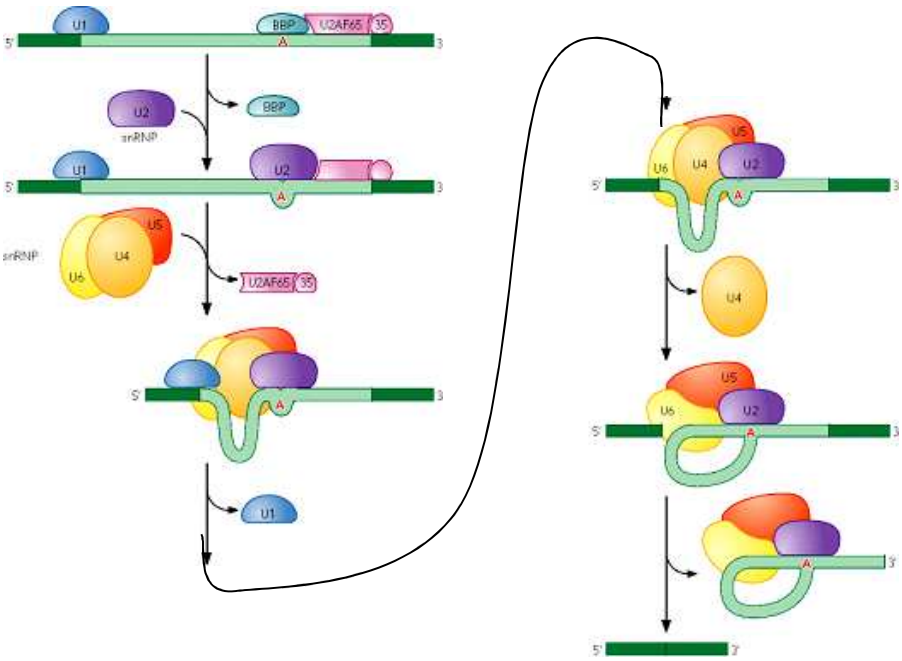
RBM – RNA binding motif

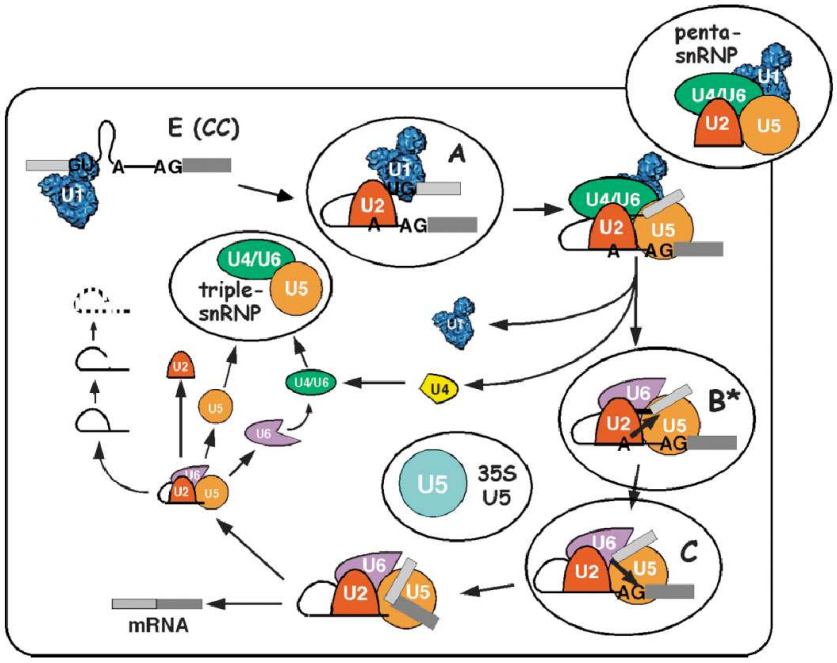
RRM – RNA recognition motif



In S. cerevisiae, the branching site sequence is very conserved (much less in higher eukaryotes)

Il modello sequenziale di montaggio e funzionamento dello spliceosome.





The spliceosome: the most complex macromolecular machine in the cell?

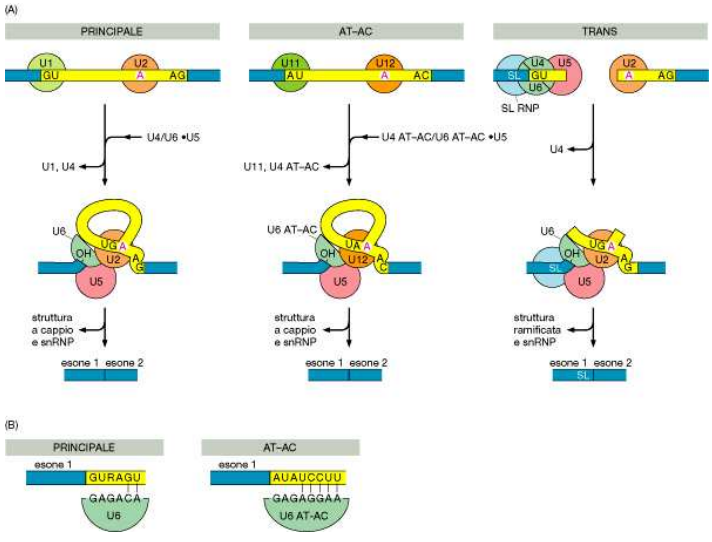
Timothy W. Nilsen

Bioessays (2003), 25: 1147-1149.

Summary
The primary transcripts, pre-mRNAs, of almost all protein-coding genes in higher eukaryotes contain multiple non-coding intervening sequences, introns, which must be precisely removed to yield translatable mRNAs. The process of intron excision, splicing, takes place in a massive ribonucleoprotein complex known as the spliceosome. Extensive studies, both genetic and biochemical, in a variety of systems have revealed that essential components of the spliceosome include five small RNAs–U1, U2, U4, U5 and U6, each of which functions as a RNA, protein complex called an snRNP (small nuclear ribonucleoprotein). In addition to snRNPs, splicing requires many non-snRNP protein factors, the exact nature and number of which has been unclear. Technical advances, including new affinity purification methods and improved mass spectrometry techniques, coupled with the completion of many genome sequences, have now permitted a number of proteomic analyses of purified spliceosomes. These studies, recently reviewed by Jurica and Moore,[1] reveal that the spliceosome is composed of as many as 300 distinct proteins and five RNAs, making it among the most complex macromolecular machines known. *BioEssays* 25:1147–1149, 2003.

Esistono due forme di splicing un po' diverse: la prima si riferisce ad un sottotipo di introni detti AT-AC dai dinucleotidi di confine, molto poco frequenti, che hanno snRNP dedicate.

La seconda è detta "trans-splicing" ed è un fenomeno raro in cui un esone presente in un pre-mRNA viene ligato ad un altro esone presente in un secondo pre-mRNA



H. sapiens	Estimated number of genes: < 25,000 Estimated number of proteins: > 90,000

Complexity of higher organisms (animals→vertebrates) is estimated to require more functions than the number of genes detected in sequeced genomes.

This can be accounted for if we assume a more extensive than previously suspected role for **alternative splicing**

S. cerevisiae:	253 genes contain introns only **3 genes** shown <u>experimentally</u> to undergo alternative splicing

H. sapiens:	>99% predicted to have exon-intron structure >60% predicted to undergo alternative splicing .....but we will see this by RNA-Seq experiments !

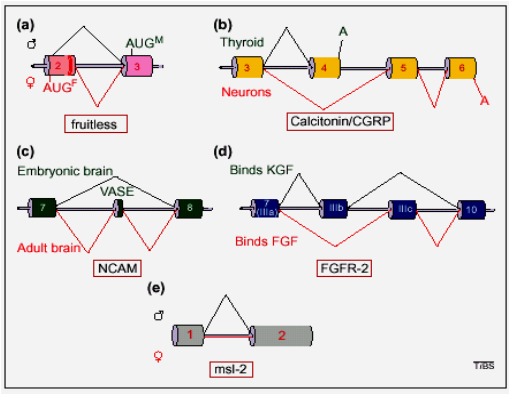
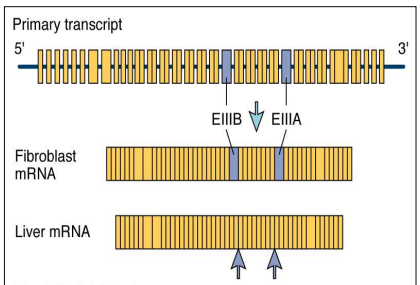
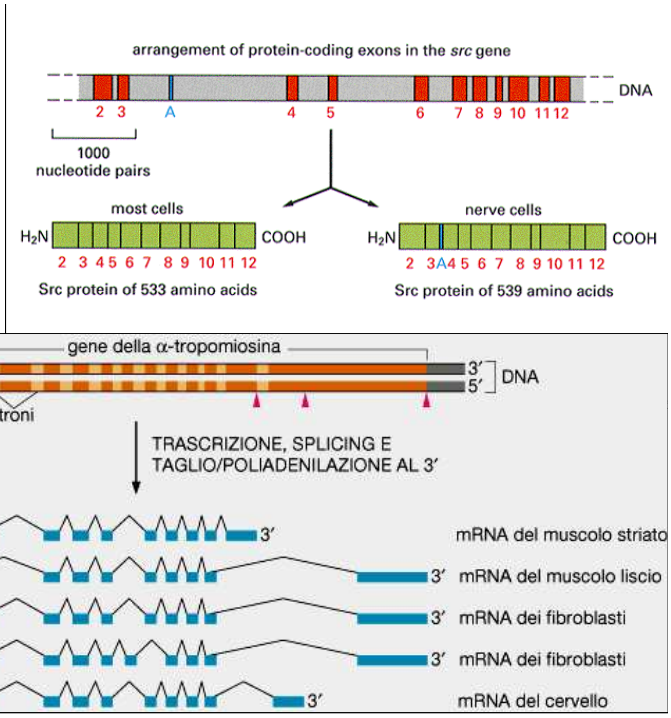


Figure 1
Different modes of alternative splicing and examples of its biological consequences. **(a)** Alternative 5′ splice-site use in the *Drosophila* gene *fruitless* governs sexual orientation and behaviour. Male (green) and female (red) patterns of splicing, as well as translation initiation codons giving rise to long open reading frames, are indicated. Red lines in exon 2 represent binding sites for Tra (for 'Transformer') and Tra-2. **(b)** Alternative 3′ splice-site usage, associated with differential use of polyadenylation sites (represented by A) in the vertebrate gene for calcitonin and calcitonin-gene-related peptide (CGRP) generates a calcium homeostatic hormone in the thyroid gland or a vasodilator neuropeptide in the nervous system. Processing patterns in green are found in thyroid, those in red are found in neurons. **(c)** Differential inclusion or skipping of the variable alternatively spliced exon (VASE) in the gene for neural cell adhesion molecule (NCAM) in embryonic (green) versus adult (red) rat brain, represses or promotes axon outgrowth during development. **(d)** Mutually exclusive use of exons IIIb and IIIc in mammalian fibroblast growth factor receptor 2 (FGFR-2) changes its binding specificity for growth factors during prostate cancer progression. The pattern of splicing represented in green generates an mRNA encoding a receptor with high affinity for keratinocyte growth factor (KGF), whereas that in red generates a receptor with high affinity for FGF. **(e)** Female-specific retention of an intron at the 5′ untranslated region (UTR) of the gene *male-specific-lethal 2* (*msl-2*) allows export of the unspliced RNA to the cytoplasm. The protein Sex-lethal facilitates both intron retention in the nucleus and translational repression in the cytoplasm, thereby switching off *msl-2* expression, which controls X-chromosome dosage compensation.



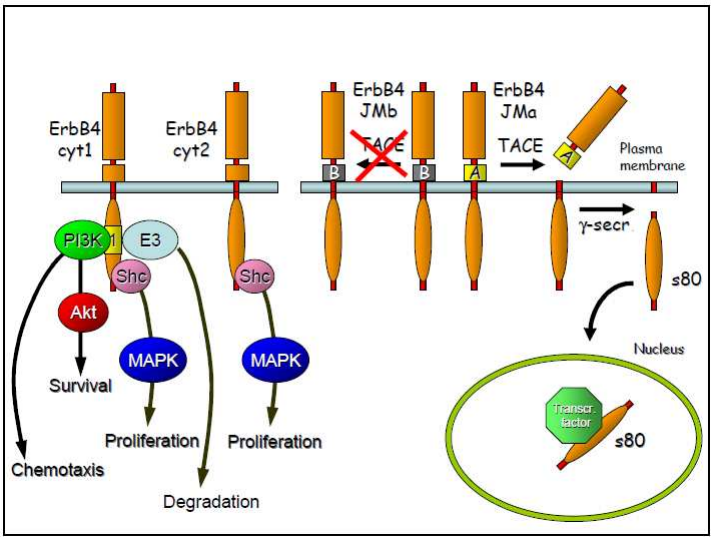
Splicing alternativo della fibronectina

Quite often, AS is tissue-specific



AS in many cases give rise to proteins with differential functions and roles.

One example is already well-known in this course: ERBB4



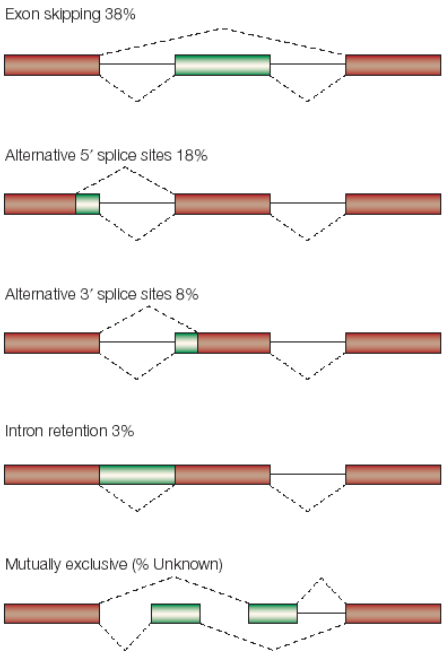


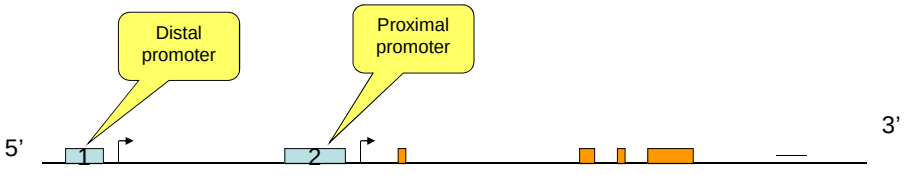
Figure 3

Types of alternative splicing.

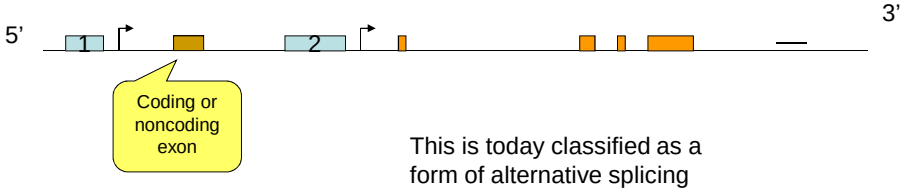
In all five examples of alternative splicing, constitutive exons are shown in red and alternatively spliced regions in green, introns are represented by solid lines, and dashed lines indicate splicing activities. Relative abundance of alternative splicing events that are conserved between human and mouse transcriptomes are shown above each example (in % of total alternative splicing events).

From: Ast G. (2004)
Nature Rev Genetics 5: 773.

Some genes display “alternative promoters”

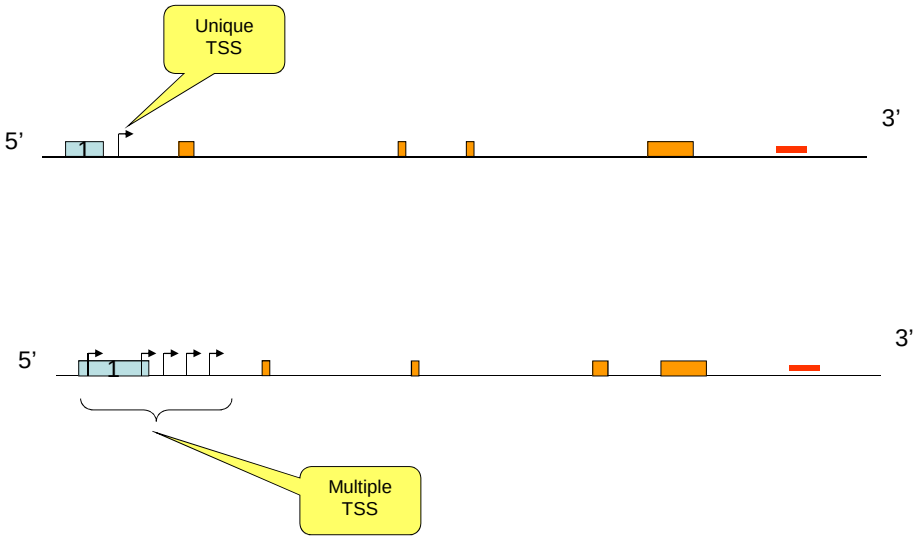


Sometimes an exon is present between the two promoters

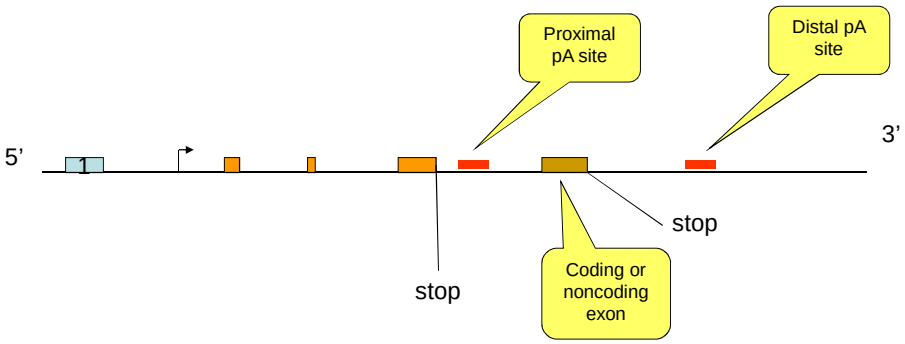


This is today classified as a form of alternative splicing

This is different from the story of multiple TSS



Other genes possess “alternative polyadenylation sites”



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

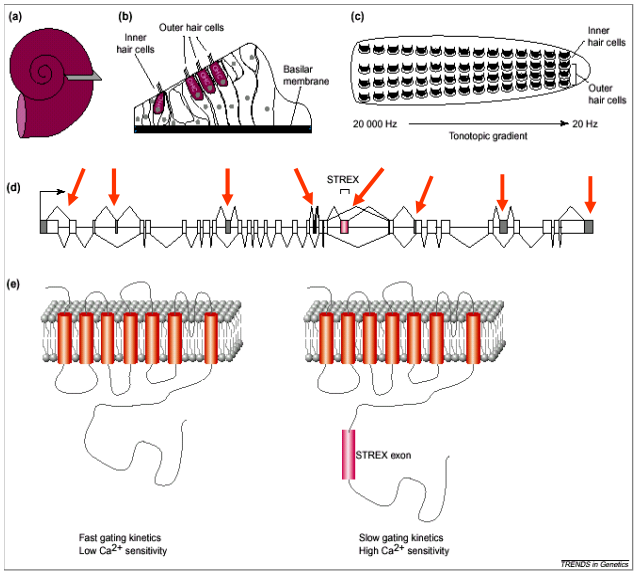


Fig. 1. Alternative splicing of the *slo* gene. (a) The mammalian cochlea. The cochlea is a snail-like structure of the inner ear that contains hair cells organized along a basilar membrane. The basilar membrane traverses the length of the curled-up cochlea. (b) The cochlea is sliced transversely as shown in (a) and the section of the cochlea containing the basilar membrane and the hair cells depicted. There are four rows of hair cells, one inner hair cell and three outer hair cells, situated above the basilar membrane. (c) The cochlea is unrolled to reveal the basilar membrane viewed from above. The four hair cells are arranged in rows along the length of the basilar membrane. The hair cells are tuned to unique narrow sound frequencies along the basilar membrane creating a tonotopic gradient. At one end of the membrane, hair cells are tuned to respond to a frequency of 20 Hz, where as hair cells at the other end respond to 20 000 Hz. (d) Organization of the human slo gene. The exon–intron organization of the slogene (determined by an analysis of draft sequence of the human genome) is depicted. The constitutive splicing events are indicated below the gene and alternative splicing events are depicted above the gene. The constitutive exons are white and the alternative exons are shaded. The STREX exon is purple.

(e) Isoforms of the Slo protein lacking sequences encoded by the STREX exon have fast deactivation kinetics and low Ca 2+ sensitivity, whereas isoforms containing STREX-encoded sequences have slower deactivation kinetics and higher Ca 2+ sensitivity.

From: Graveley BR (2001) Trends Genet., 17:100-106.

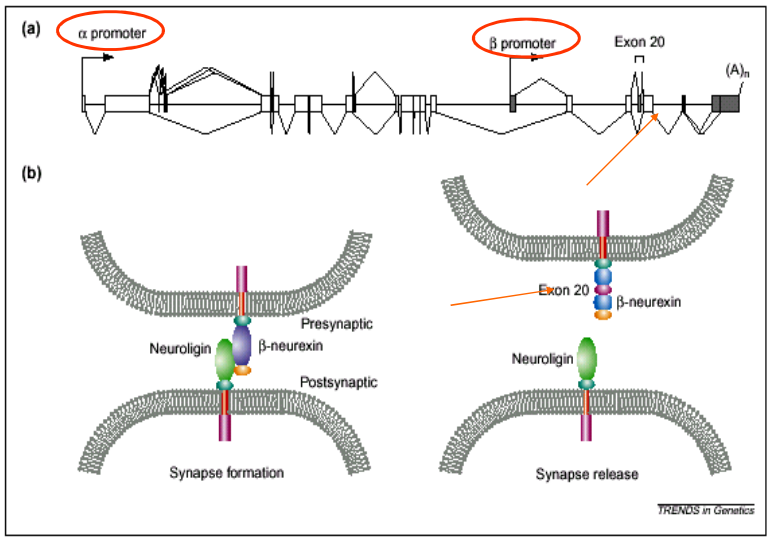


Fig. 2. Alternative splicing of the neurexin genes. (a) Organization of the human gene encoding neurexin I. The exon–intron structure of the human neurexin I gene is depicted (L. Rowen and B. Graveley, unpublished). The constitutive splicing events are indicated below the gene and alternative splicing events are depicted above the gene. The constitutive exons are white and the alternative exons are shaded. Exon 20 is indicated. Human neurexins II and III have a very similar exon–intron organization (L. Rowen and B. Graveley, unpublished). (b) Model for the function of the alternative splicing of exon 20 in b-neurexin I. b-neurexin I (present in the presynaptic cell) lacking sequences encoded by exon 20 can interact with neuroligin present in the postsynaptic cell, and thus function to initiate synaptogenesis. In contrast, b-neurexin I containing exon 20 encoded sequences can not interact with neuroligins. This form of b-neurexin I might indirectly function in releasing synapses.

Neurexins

From: Graveley BR (2001) Trends Genet., 17:100-106.

Drosophila Dscam gene provides probably the **extreme** example of alternative splicing.

Perhaps the most complex event that takes place during development is the migration and connection of neurons. Even in a 'simple' organism such as *Drosophila melanogaster*, which contains only ~250 000 neurons, accurately wiring neurons together would appear to be a daunting task.

In flies, the gene encoding the Down syndrome cell adhesion molecule (*Dscam*) appears to fulfill at least part of this role. *Dscam* encodes an axon guidance receptor with an extracellular domain that contains ten immunoglobulin (Ig) repeats. The most striking feature of the *Dscam* gene is that it's pre-mRNA can be alternatively spliced into over 38,000 different mRNA isoforms (Fig. 3a). This is 2–3 times the number of predicted genes in the entire organism !

Each mRNA encodes a distinct receptor with the potential ability to interact with different molecular guidance cues, directing the growing axon to its proper location.

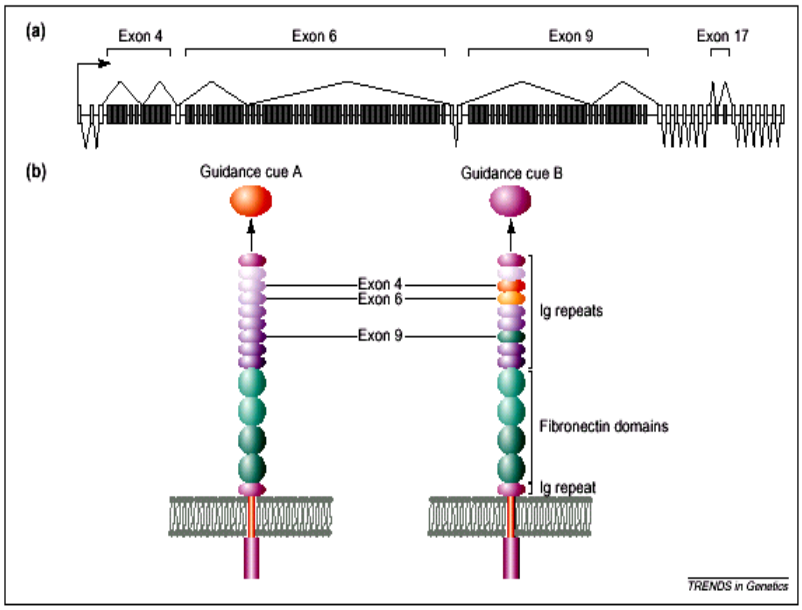
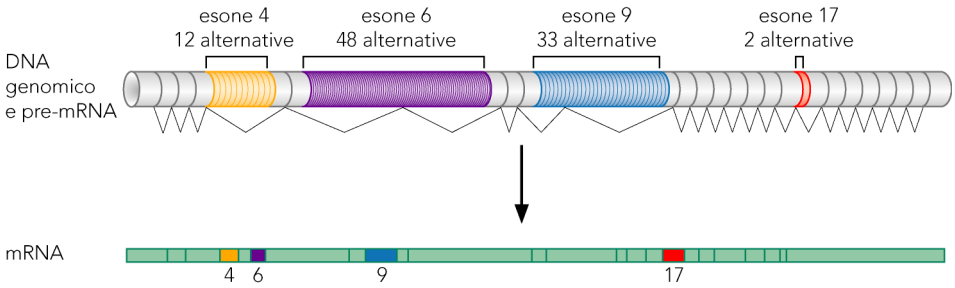


Fig. 3. Alternative splicing of the gene encoding *Drosophila Dscam*. (a) The organization of the *Dscam* gene. The constitutive splicing events are indicated below the gene and alternative splicing events are depicted above the gene. The constitutive exons are white and the alternative exons are shaded. The *Dscam* gene contains four sites of alternative splicing at exons 4, 6, 9 and 17. There are 12 variants of exon 4, 48 variants of exon 6, 33 variants of exon 9 and 2 variants of exon 17. Only one variant exon from each position is included in the *Dscam* mRNAs. Alternative exons 4, 6, and 9 encode alternative versions of immunoglobulin repeats. (b) Functional consequences of *Dscam* alternative splicing. The Dscam protein functions as an axon guidance receptor. It is thought that each Dscam variant will interact with a unique set of axon guidance cues. The form of Dscam shown on the left will interact with guidance cue A. The form of Dscam shown on the right contains different sequences encoded by exons 4, 6 and 9 and thus interacts with guidance cue B, rather than guidance cue A. Neurons expressing the form of Dscam shown on the right will be attracted in a different direction than neurons expressing the form shown on the left.

From: Graveley BR (2001) Trends Genet., 17:100-106.



Potentially 38,000 splicing variants

Very different situations, but “on average” only one-two exons per gene appear as “alternative”

So, what do we know at the genomic scale?

articles 922 © 2001 Macmillan Magazines Ltd NATURE | VOL 409 | 15 FEBRUARY 2001 |

Experimental annotation of the human genome using microarray technology

D. D. Shoemaker*, E. E. Schadt*, C. D. Armour, Y. D. He, P. Garrett-Engele, P. D. McDonagh, P. M. Loerch, A. Leonardson, P. Y. Lum, G. Cavet, L. F. Wu, S. J. Altschuler, S. Edwards, J. King, J. S. Tsang, G. Schimmack, J. M. Schelter, J. Koch, M. Ziman, M. J. Marton, B. Li, P. Cundiff, T. Ward, J. Castle, M. Krolewski, M. R. Meyer, M. Mao, J. Burchard, M. J. Kidd, H. Dai, J. W. Phillips, P. S. Linsley, R. Stoughton, S. Scherer & M. S. Boguski

Rosetta Inpharmatics, Inc., 12040 115th Avenue N.E., Kirkland, Washington 98034, USA
** These authors contributed equally to this work*

The most important product of the sequencing of a genome is a complete, accurate catalogue of genes and their products, primarily messenger RNA transcripts and their cognate proteins. Such a catalogue cannot be constructed by computational annotation alone; it requires experimental validation on a genome scale. Using 'exon' and 'tiling' arrays fabricated by ink-jet oligonucleotide synthesis, we devised an experimental approach to validate and refine computational gene predictions and define full-length transcripts on the basis of co-regulated expression of their exons. These methods can provide more accurate gene numbers and allow the detection of mRNA splice variants and identification of the tissue- and disease-specific conditions under which genes are expressed. We apply our technique to chromosome 22q under 69 experimental condition pairs, and to the entire human genome under two experimental conditions. We discuss implications for more comprehensive, consistent and reliable genome annotation, more efficient, full-length complementary DNA cloning strategies and application to complex diseases.

Exonic arrays

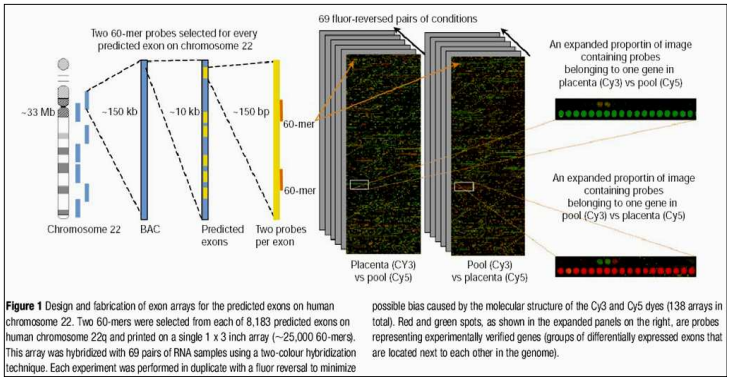


Figure 1 Design and fabrication of exon arrays for the predicted exons on human chromosome 22. Two 60-mers were selected from each of 8,183 predicted exons on human chromosome 22q and printed on a single 1 x 3 inch array (~25,000 60-mers). This array was hybridized with 69 pairs of RNA samples using a two-colour hybridization technique. Each experiment was performed in duplicate with a fluor reversal to minimize possible bias caused by the molecular structure of the Cy3 and Cy5 dyes (138 arrays in total). Red and green spots, as shown in the expanded panels on the right, are probes representing experimentally verified genes (groups of differentially expressed exons that are located next to each other in the genome).

from: Shoemaker et al. (2001) Nature, 409: 922.

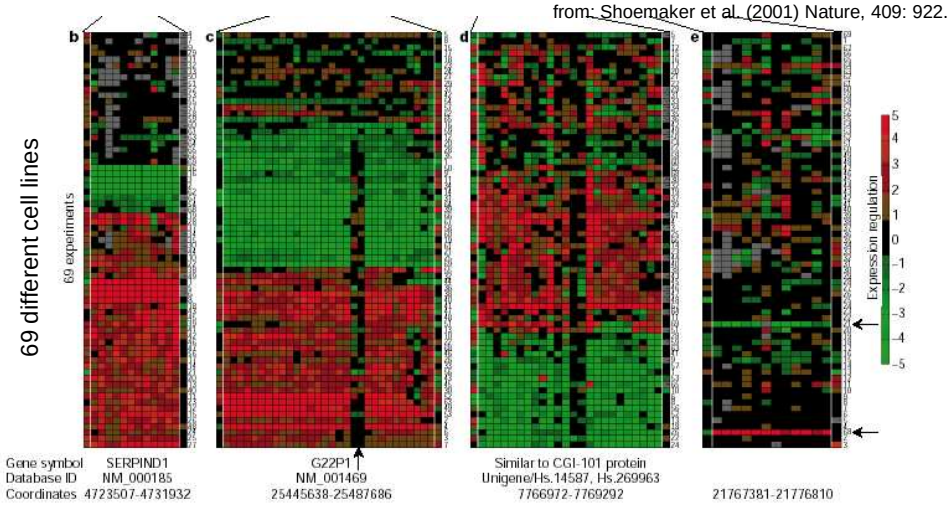


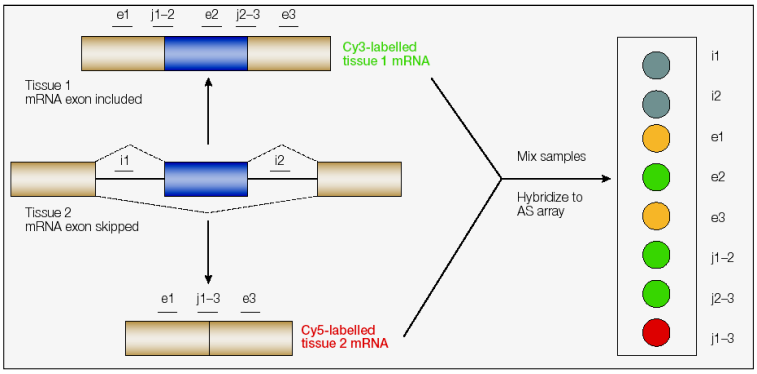
Figure 2 Using expression data from multiple conditions to validate exons and define gene boundaries on chromosome 22. **a**, Pseudocolour image showing error-weighted $\log_{10}$ expression ratios (red/green) for each of the ~8,000 exons (*x*-axis) across the 69 fluor-reversed experiments (*y*-axis). A brief description of each experiment is listed on the right side of the image; the numbers (1–69) are reference points for the Table in the Supplementary Information. The 15,511 probes representing the 8,183 predicted exons are arranged linearly across the 33 Mb of chromosome 22. **b**, Expanded region showing a known gene (SERPIND1, NM_ 000185). The experiments on the *y*-axis have been clustered to emphasize how co-regulation across diverse experiments can be used to group exons into genes. The vertical white lines indicate the boundaries predicted by our gene finding algorithm; numbers on *y*-axis indicate experimental conditions. **c**, Expanded region showing a set of co-regulated exons from another known gene (G22P1, NM_001469), illustrating the detection of potential false positives (arrow) made by the Genscan prediction program. **d**, Expanded region representing an EVG that collapses two Unigene EST clusters (HS.269963 and HS.14587) into a single transcript. **e**, Expanded region showing an EVG containing six exons that are part of a novel testis-expressed transcript (arrows, two experiments involving testis RNA samples).

Exon arrays give often “difficult-to-interpret” data

Tiling arrays could give the best information, provided that they have really one-nucleotide-tiling probes (one nucleotide resolution)

Exon-junction arrays also have been extensively used

Exon-exon junction arrays



Oligonucleotide probes, typically 25–60 nucleotides in length, can be designed to hybridize to isoform-specific mRNA regions. Recently, alternative splicing microarrays have been designed with probes that are specific to both exons and exon–exon junctions. Probes e1, e2 and e3 are exon specific, whereas j1–2, j2–3 and j1–3 are isoform-specific junction probes. Some arrays also contain intron probes (i1 and i2) to indicate signals from pre-mRNA. Various array design and data processing strategies facilitate the quantitative analysis of alternative splicing patterns, some of which have been subsequently confirmed by PCR after reverse transcription of RNA (RT-PCR). Johnson et al.(2003) used arrays with probes for all adjacent exon–exon junctions in 10,000 human genes and hybridized these with samples from 52 human tissues and cell lines. This revealed cell-type-specific clustering of alternative splicing events, and allowed the discovery of new alternative splicing events. Pan et al.136 analysed 3,126 known cassette-type alternative splicing events in mouse using exon-specific and exon–exon junction probes. Analysis of RNAs in ten tissues showed clustering of alternative splicing events by tissue type, and further revealed that tissue-specific programmes of transcription and alternative splicing operate on different subsets of genes. A direct comparison also showed that computational prediction of tissue-specific alternative splicing based on ESTs and cDNAs performed poorly compared with the alternative splicing microarray and RT-PCR. From: Matlin et al. (2005), Nature Rev Mol Cell Biol, 6: 386.

A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome

Marc Sultan,[1]* Marcel H. Schulz,[2,3]* Hugues Richard,[2]* Alon Magen,[1] Andreas Klingenhoff,[4] Matthias Scherf,[4] Martin Seifert,[4] Tatjana Borodina,[1] Aleksey Soldatov,[1] Dmitri Parkhomchuk,[1] Dominic Schmidt,[1] Sean O'Keeffe,[2] Stefan Haas,[2] Martin Vingron,[2] Hans Lehrach,[1] Marie-Laure Yaspo[1]†

The functional complexity of the human transcriptome is not yet fully elucidated. We report a high-throughput sequence of the human transcriptome from a human embryonic kidney and a B cell line. We used shotgun sequencing of transcripts to generate randomly distributed reads. Of these, 50% mapped to unique genomic locations, of which 80% corresponded to known exons. We found that 66% of the polyadenylated transcriptome mapped to known genes and 34% to nonannotated genomic regions. On the basis of known transcripts, RNA-Seq can detect 25% more genes than can microarrays. A global survey of messenger RNA splicing events identified 94,241 splice junctions (4096 of which were previously unidentified) and showed that exon skipping is the most prevalent form of alternative splicing.

Paper discussed in part in Lesson 4

956 15 AUGUST 2008 VOL 321 SCIENCE

Table 1. Summary of genes, splice junctions, and previously unrecognized TUs identified by RNA-Seq; mapping of the read for the merged lanes.

Mapping summary	HEK 293	B cells
Total reads	8,638,919	7,682,230
Low-quality reads	234,160	194,999
Reads with multiple matches	1,546,361	1,324,770
Reads with unique matches	4,640,112	3,895,643
Reads mapping to annotated RNAs (ENSEMBL + Eldorado)	3,712,476	2,902,387
ENSEMBL genes with at least five reads	12,567	10,668
ENSEMBL genes with at least one read	14,963	13,739
Reads in intronic clusters	38,598	44,781
ENSEMBL genes with intronic read clusters	1445	1409
Introns with read clusters	1862	1847
Reads with no match to the genome	2,218,286	2,266,818
Reads aligned to splice junctions	307,904	229,453
Identified junctions	78,880	62,596
(expected)	(81,302)	(66,981)
Genes (at least five reads) with junctions	10,292	8655
Genes (at least one read) with junctions	10,558	8910
Genes (at least one read) with previously unknown junctions	2078	1732
Previously unknown junctions	2397	1965
Previously unknown junctions identified by less than one read	203	182

In totale, 64% dei trascritti poliadenilati corrisponde a geni noti, mentre il restante 34% mappa su regioni non annotate.

14% di questi possono essere attribuiti a “splice junctions”

In totale, ci sono 7,2 giunzioni per gene e in media 3,8 reads per giunzione.

Si sono osservate tantissime giunzioni con un solo read, ma questo non sembra dovuto al caso, perché è un numero troppo grande.

Il 95% degli splicing attesi è stato effettivamente visto.

Inoltre, si sono visti 4096 splice sites non noti in 3106 geni.

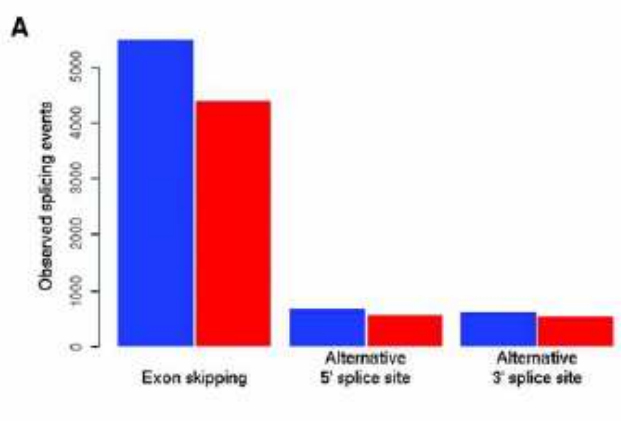


Fig. 3. AS events observed by junction reads. (A) Distribution of the three major types of AS: (i) cassette exons, (ii) alternative 5′ splice sites, and (iii) alternative 3′ splice sites. Blue HEK293 cells. Red B cells.

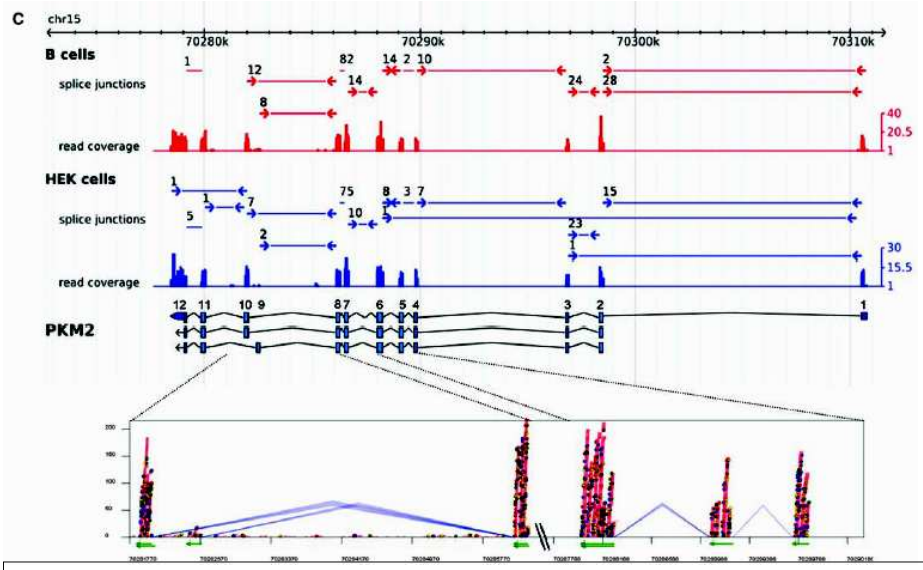


Fig.3- (C) Example of AS in the PKM2 gene. Three isoforms annotated in ENSEMBL (ENST00000335181, ENST00000389092, ENST00000389091) are shown next to the gene name, and exons are numbered. The read coverage is shown for each exon (blue for HEK and red for B cells). Splice junction reads are shown as arrows; the numbers above the arrows represent the number of reads at junctions. The bottom box shows basepair resolution coverage in HEK cells of the gene's regions containing exons 8 to 10 (green arrows at left) and 4 to 6 (green arrows at right). The blue lines denote splice junctions. (Left) Two different sequenced junctions connecting either exon 9 or exon 10 and identifying alternative transcripts with mutually exclusive exons in HEK and in B cells. Colored dots represent sequence differences.

Review articles

Alternative splicing: current perspectives

Eddo Kim,† Amir Goren,† and Gil Ast*

Summary
Alternative splicing is a well-characterized mechanism by which multiple transcripts are generated from a single mRNA precursor. By allowing production of several protein isoforms from one pre-mRNA, alternative splicing contributes to proteomic diversity. But what do we know about the origin of this mechanism? Do the same evolutionary forces apply to alternatively and constitutively splice exons? Do similar forces act on all types of alternative splicing? Are the products generated by alternative splicing functional? Why is "improper" recognition of exons and introns allowed by the splicing machinery? In this review, we summarize the current knowledge regarding these issues from an evolutionary perspective. *BioEssays* 30:38–47, 2008.

An evolutionary point of view

REVIEW

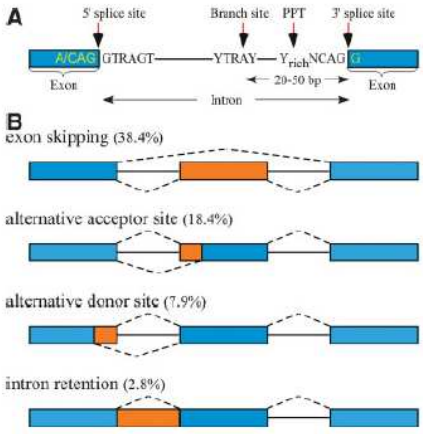


Figure 1. Types of alternative splicing. **A:** The four basal splice signals are depicted: 5′ splice site, branch site, polypyrimidine tract (PPT), and 3′ splice site. **B:** The four main types of alternative splicing are illustrated: exon skipping, alternative acceptor site selection, alternative donor site selection, and intron retention. The relative prevalence of each type in alternative exons conserved in human and mouse is shown in parenthesis. The remaining 32.5%, which are not shown, represent more complex alternative splicing events. Constitutive exons are shown in blue; alternatively spliced regions in orange; introns are represented by solid lines; and dashed lines indicate splicing options.

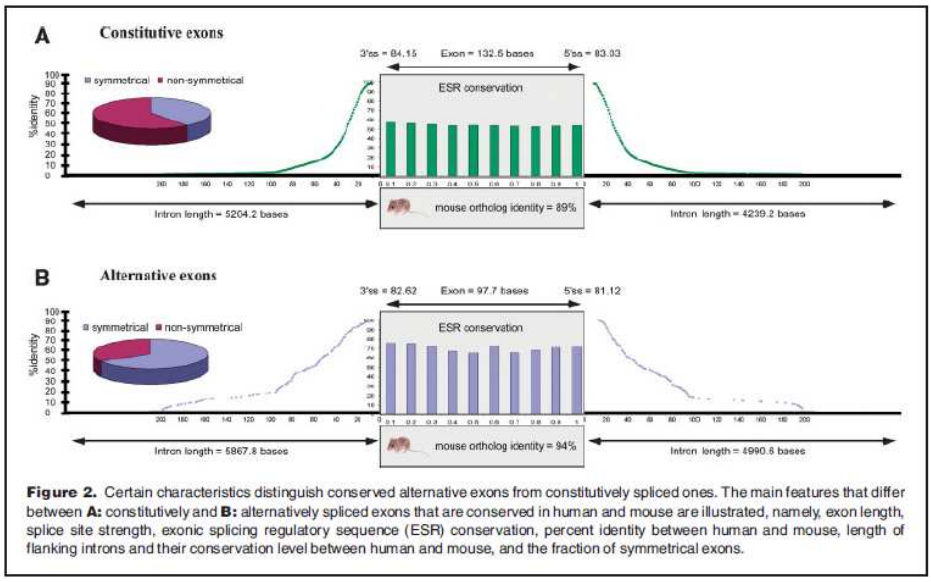


Figure 2. Certain characteristics distinguish conserved alternative exons from constitutively spliced ones. The main features that differ between **A:** constitutively and **B:** alternatively spliced exons that are conserved in human and mouse are illustrated, namely, exon length, splice site strength, 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.

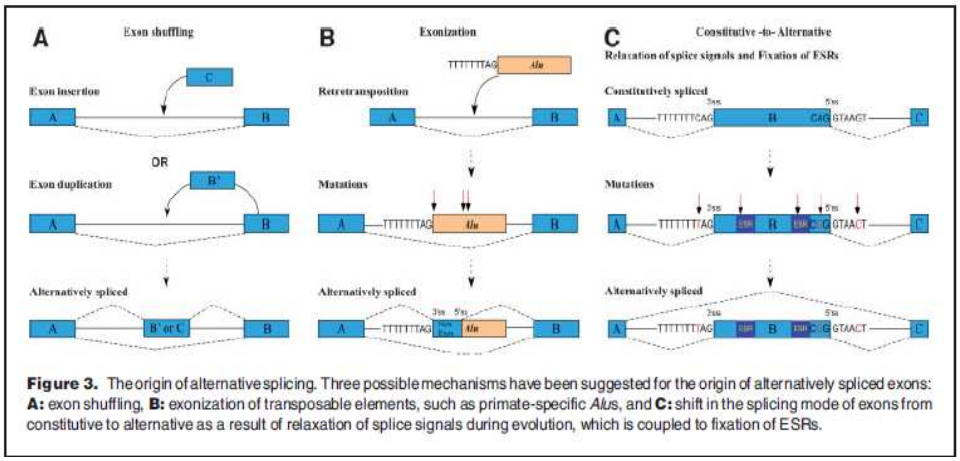


Figure 3. The origin of alternative splicing. Three possible mechanisms have been suggested for the origin of alternatively spliced exons: **A:** exon shuffling, **B:** exonization of transposable elements, such as primate-specific *Alu*s, and **C:** shift in the splicing mode of exons from constitutive to alternative as a result of relaxation of splice signals during evolution, which is coupled to fixation of ESRs.

Origin of alternative cassette exons

Until recently, only two mechanisms were suggested to be responsible for the origin of alternatively spliced exons. Both mechanisms describe the appearance of new exons, which are then spliced alternatively. One of these mechanisms is known as exon shuffling, in which a new exon is inserted into an existing gene or an exon is duplicated within the same gene, and becomes alternatively spliced (Fig. 3A).[64,65] About 10% of all genes contain tandemly duplicated exons and about 10% of mutually exclusive alternatively spliced exons originated from tandemly duplicated exons.[2,66]

The second mechanism for the origin of alternatively spliced exons involves the emergence of alternatively spliced exons following exonization of intronic sequences (Fig. 3B). For example, the primate-specific *Alu* retroelement, which is highly abundant in intronic sequences,[67] contains multiple sites that are similar, but not identical, to real splice sites.[68,69]

The pressure on alternative exons to maintain the reading frame

Many of the human cassette exons with high inclusion levels are also alternatively spliced in the mouse orthologous gene. This indicates that the alternative form emerged before the human/mouse lineages diverged and that the alternative state remained (fixated) during evolution. This conservation implies that there are functional roles for both the exon inclusion and skipping isoforms and that the alternative form is not merely a splicing error. About 66% of alternative cassette exons conserved between human and mouse are symmetrical—that is, the total number of nucleotides in the exon is divisible by three—compared with only 40% of constitutively spliced ones.[56,62] Hence, functional alternative exons tend to maintain the open reading frame.

A similar trend was observed for alternative 5′ and 3′ splice sites selection—the alternative extension was usually symmetrical (63–72%), whereas the exon as a whole exhibited symmetry levels identical to constitutive exons.[63] But why

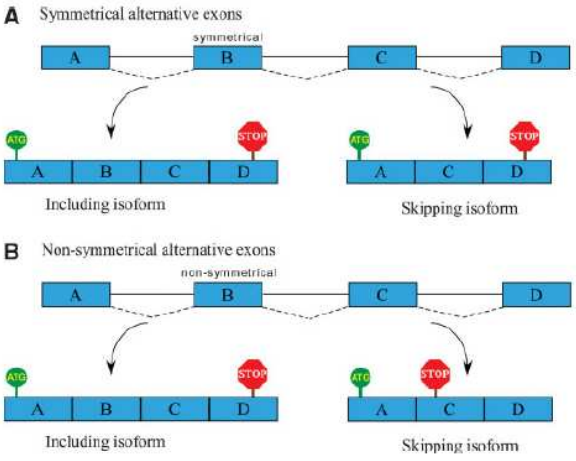
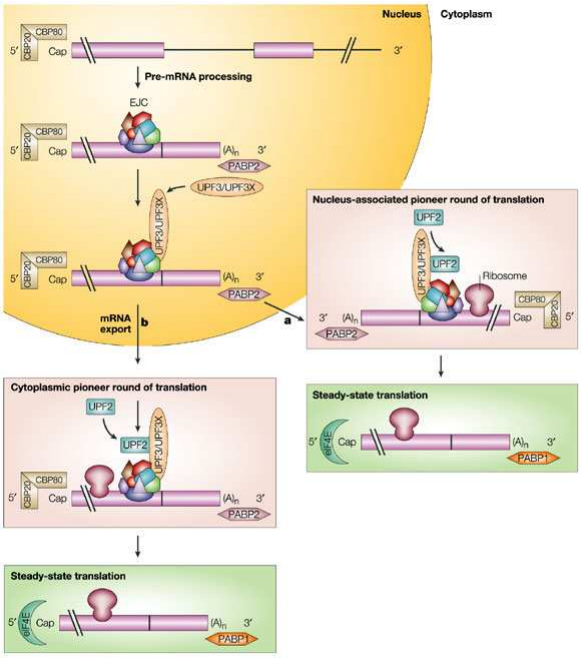


Figure 4. Effects of symmetrical and non-symmetrical alternatively spliced exons on the reading frame. **A:** A schematic of a gene structure consisting of four exons, where the second exon is a symmetrical cassette exon, is depicted. Inclusion or exclusion of the symmetrical exon does not alter the open reading frame of the downstream exons. **B:** A gene consisting of four exons, where the second exon is a non-symmetrical cassette exon, is illustrated. Inclusion or exclusion of the non-symmetrical exon results in a frame-shift and a PTC downstream of the alternative exon.

Nonsense-mediated decay

Pre-mRNA, which consists of exons (pink boxes) and introns (black lines between boxes), is bound by the cap-binding proteins CBP80 and CBP20 at the 5' cap and, after 3'-end formation, poly(A)-binding protein PABP2 at the 3' poly(A) tail. Pre-mRNA processing generates spliced mRNA that is likewise bound by CBP80, CBP20 and PABP2, as well as an exon junction complex (EJC) of proteins 20–24 nucleotides (nt) upstream of each exon–exon junction. This EJC consists minimally of RNPS1, SRm160 and UAP56, as well as Y14, REF/ALY and NXF1/TAP–p15 . The EJC acquires further proteins, including UPF2 and UPF3 or UPF3X, which function in nonsense-mediated mRNA decay (NMD). UPF3 or UPF3X, which is mostly nuclear but shuttles to the cytoplasm, is thought to recruit UPF2, which concentrates along the cytoplasmic side of the nuclear envelope. The resulting messenger ribonucleoprotein particle (mRNP) constitutes the pioneer translation initiation complex. This complex is thought to undergo a pioneer round of translation either in association with nuclei, in the case of mRNAs that are subject to nucleus-associated NMD (a), or in the cytoplasm, in the case of mRNAs that are subject to cytoplasmic NMD (b). If the mRNP lacks a premature termination codon (PTC) or has a PTC that fails to elicit NMD, then the mRNP is remodelled to the steady-state translation initiation complex. During remodelling, the EJC and associated UPF proteins are removed, CBP80 and CBP20 are replaced by eukaryotic initiation factor eIF4E, and PAPB2 is replaced by PABP1. Whether translation is required for all steps of mRNP remodelling is unclear. So far, translation has been reported to remove Y14



REVIEW

A more complete review on the subject

Alternative splicing and evolution: diversification, exon definition and function

Hadas Keren, Galit Lev-Maor and Gil Ast

Abstract | Over the past decade, it has been shown that alternative splicing (AS) is a major mechanism for the enhancement of transcriptome and proteome diversity, particularly in mammals. Splicing can be found in species from bacteria to humans, but its prevalence and characteristics vary considerably. Evolutionary studies are helping to address questions that are fundamental to understanding this important process: how and when did AS evolve? Which AS events are functional? What are the evolutionary forces that shaped, and continue to shape, AS? And what determines whether an exon is spliced in a constitutive or alternative manner? In this Review, we summarize the current knowledge of AS and evolution and provide insights into some of these unresolved questions.