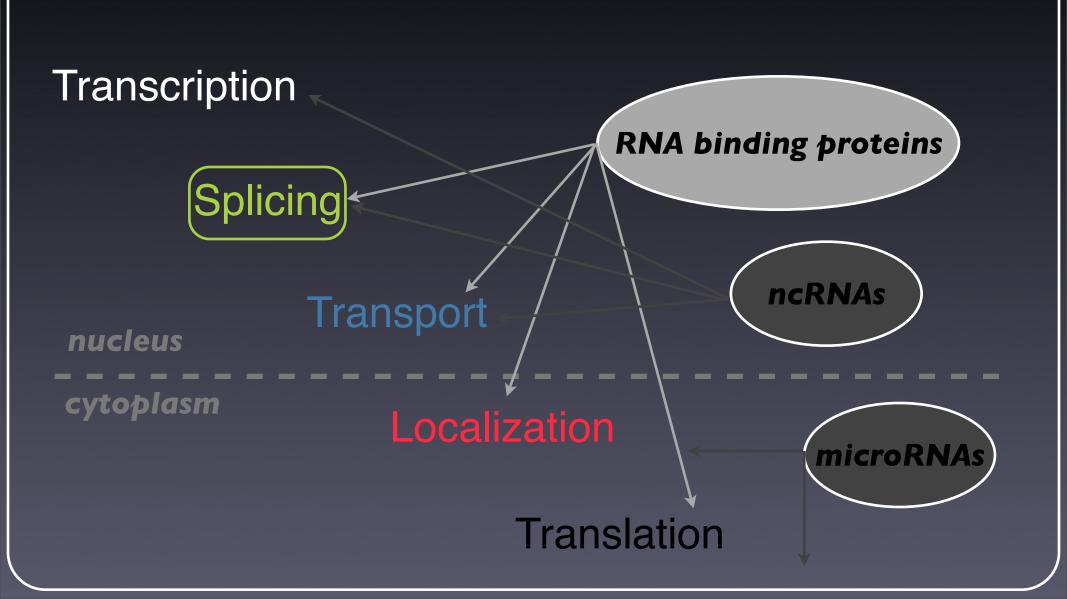
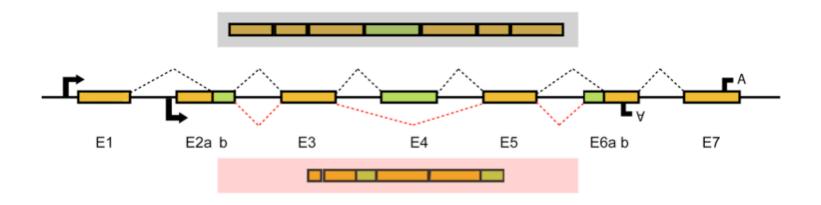
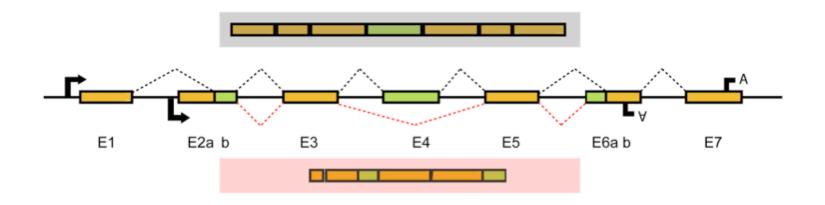


Different steps in gene regulation



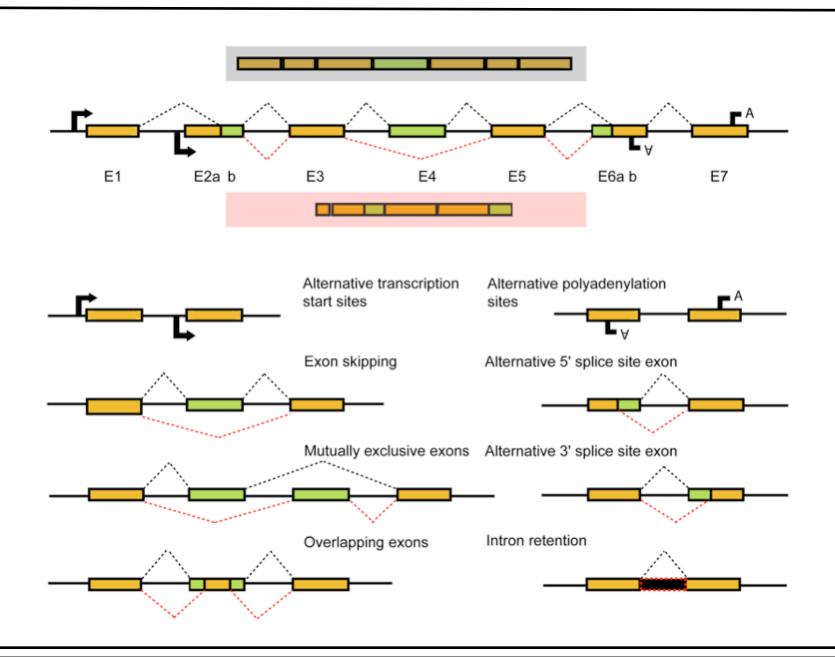


Decoding of primary transcript structure



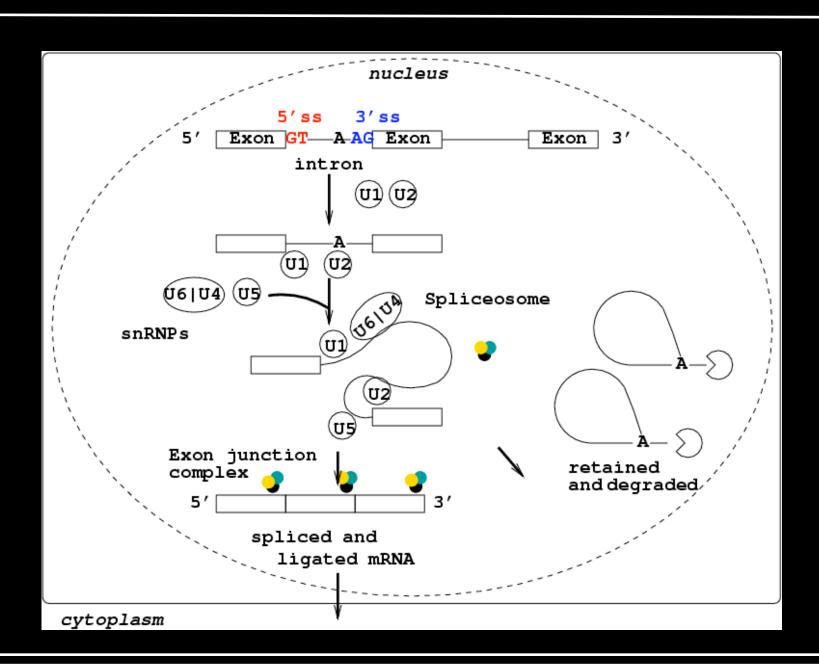
Presence of many overlapping codes

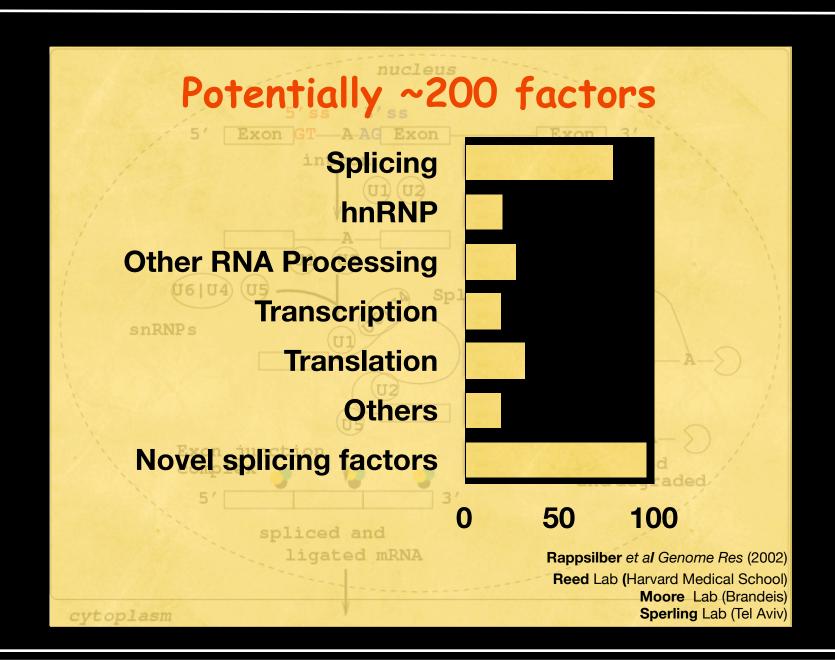
Processing of human primary transcripts to generate mRNAs that will direct protein synthesis is often variable, producing multiple alternatively spliced mRNA products



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

Nilson BioEssays (2004





Spliceosome function

- Precise recognition of splice sites among many pseudo-sites
- Removal of introns
- Production of correct message

Spliceosome function

- Precise recognition of splice sites among many pseudo-sites
- Removal of introns
- Production of correct message

RNA processing

- Coupling interaction of RNA splicing with gene expression:
- > Transcription
- Capping and Polyadenylation
- > mRNA export
- > Surveillance and mRNA degradation

Spliceosome function

- Precise recognition of splice sites among many pseudo-sites
- Removal of introns
- Production of correct message

RNA processing

- Coupling interaction of RNA splicing with gene expression:
- > Transcription
- Capping and Polyadenylation
- > mRNA export
- > Surveillance and mRNA degradation

Protein functional diversity

- Cell and tissue-specific mRNA isoforms
- Developmental stage regulated isoforms
- Inducible control and expression of isoforms

Spliceosome function

- Precise recognition of splice sites among many pseudo-sites
- Removal of introns
- Production of correct message

RNA processing

- Coupling interaction of RNA splicing with gene expression:
- > Transcription
- > Capping and Polyadenylation
- > mRNA export
- > Surveillance and mRNA degradation

Protein functional diversity

- Cell and tissue-specific mRNA isoforms
- Developmental stage regulated isoforms
- Inducible control and expression of isoforms

Splicing errors in human disease

- Inherited mutations and effects of genetic background affecting
- > Authentic splice sites
- > Alternative splice sites
- > Basal splicing machinery
- > Trans-acting regulators of AS
- Tumorigenic phenotypes and progression

Traditional approach of molecular dissection of genes and site-directed mutagenesis

Splicing factors in control of developmental or cell type specific switches, including the CELF, NOVA or PTB protein families

TONT gene

TONT 215389_s_at

TONT 215389_s_at

TONT 215389_s_at

ABBOTT ABBOTT

Expression

Adult and non-muscle muscle tissue

Embryonic muscle

CELF specific motifs

tt GATTGT gttctgacgtgt

Today's menu

I. Patterns: databases and frequency of occurrence

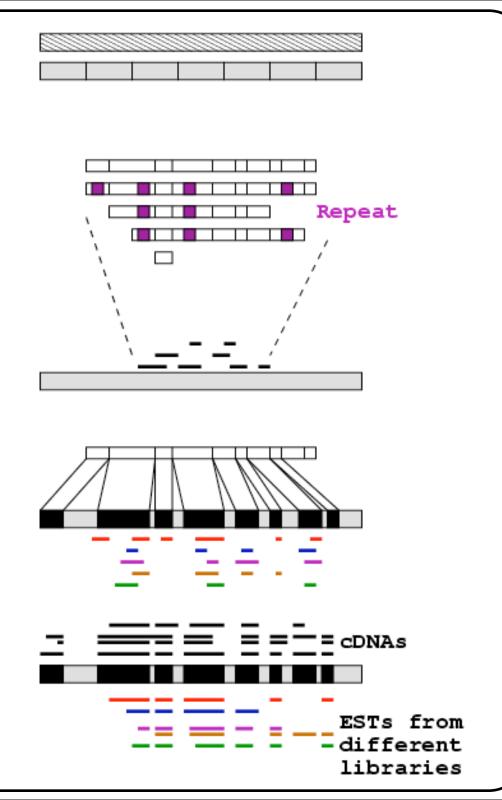
II. Prediction of evolutionary conserved alternative splicing across mammalian species

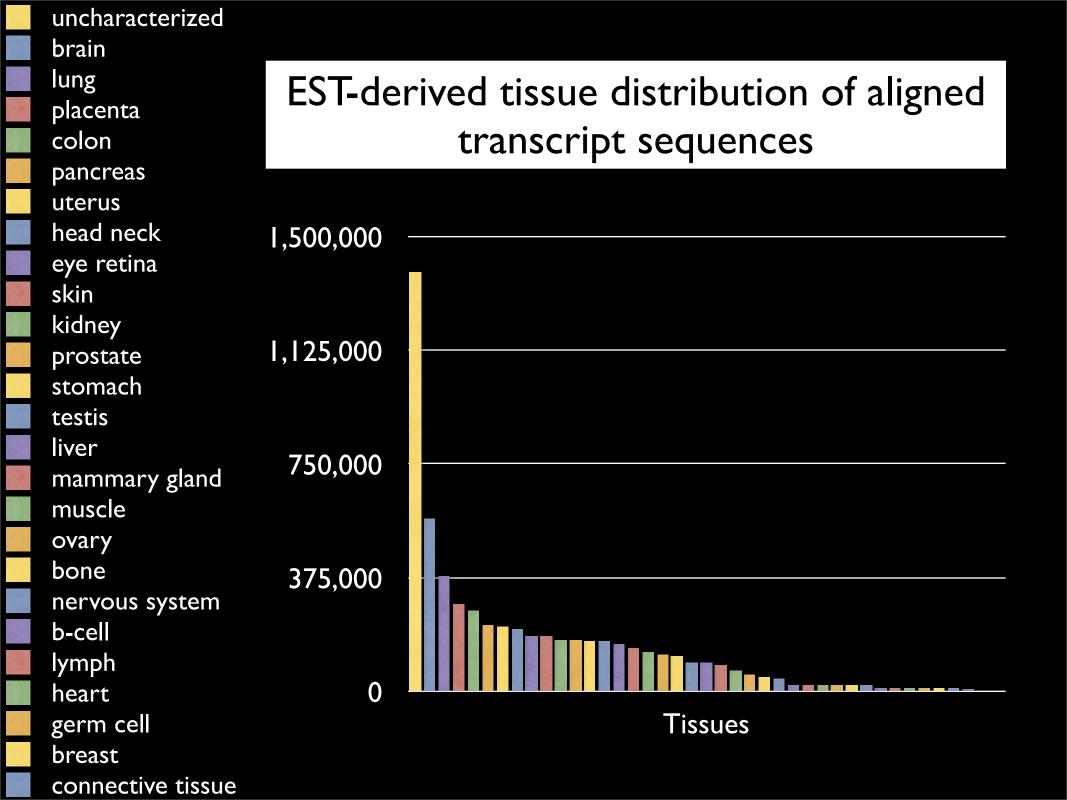
III. Alternative splicing and evolution

Developed and available data collections/ databases for alternative exons

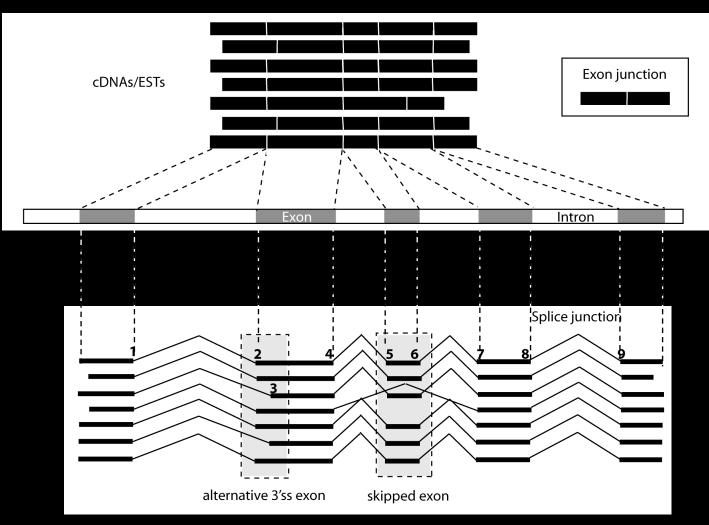
Database/Dataset	Web-based reference
Collection of EST-inferred splice sites	http://www.ebi.ac.uk/~thanaraj/splice.html
Collection of literature-based AS events	http://cgsigma.cshl.org/new_alt_exon_db2/
ASDB	http://cbcg.nersc.gov/asdb
ISIS (intron database)	http://isis.bit.uq.edu.au/a_splicers.html
ASMamDB	http://166.111.30.65/AsMamDB
SpliceDB (U2/U12 splice sites) SpliceNest	http://www.softberry.com http://splicenest.molgen.mpg.de
PALSdb	http://palsdb.ym.edu.tw
AltExtron	http://www.ebi.ac.uk/asd/altextron
AS graphs	http://www-cse.ucsd.edu/groups/bioinformatics/ESTs
ASAP	http://www.bioinformatics.ucla.edu/ASAP
ProSplicer	http://bioinfo.csie.ncu.edu.tw/ProSplicer
AS patterns in plants	http://pasdb.genomics.org.cn
ASD	http://www.ebi.ac.uk/asd
Splicing-conserved AS patterns	http://www.soe.ucsc.edu/~sugnet/psb2004/altGraphXCon.html
AS patterns across different tissues	http://genes.mit.edu/genoa
ASG, browsing splice patterns	http://statgen.ncsu.edu/asg
SpliceInfo	http://spliceinfo.mbc.nctu.edu.tw
ECgene	http://genome.ewha.ac.kr/Ecgene
STACKdb	http://www.sanbi.ac.za/Dbases.html
EASED	http://easedb.bioinfo.mdc-berlin.de
Collection of alternative poly-A sites	http://physics.nyu.edu/~jy272/altA
LSAT (literature-based)	http://www.bork.embl.de/LSAT
MAASE (literature-based)	http://maase.genomics.purdue.edu
HOLLYWOOD	http://hollywood.mit.edu

Spliced-alignments, filtering on sequence identity and splice site consensus



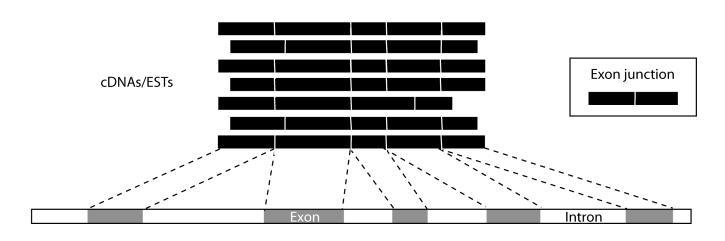


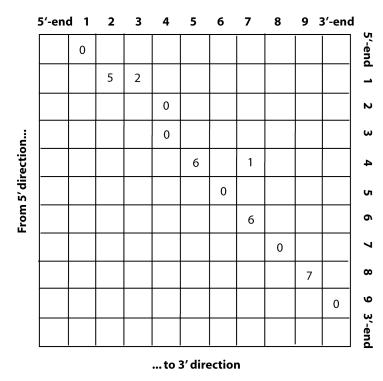
Objective - gene structure assembly





Objective - gene structure assembly

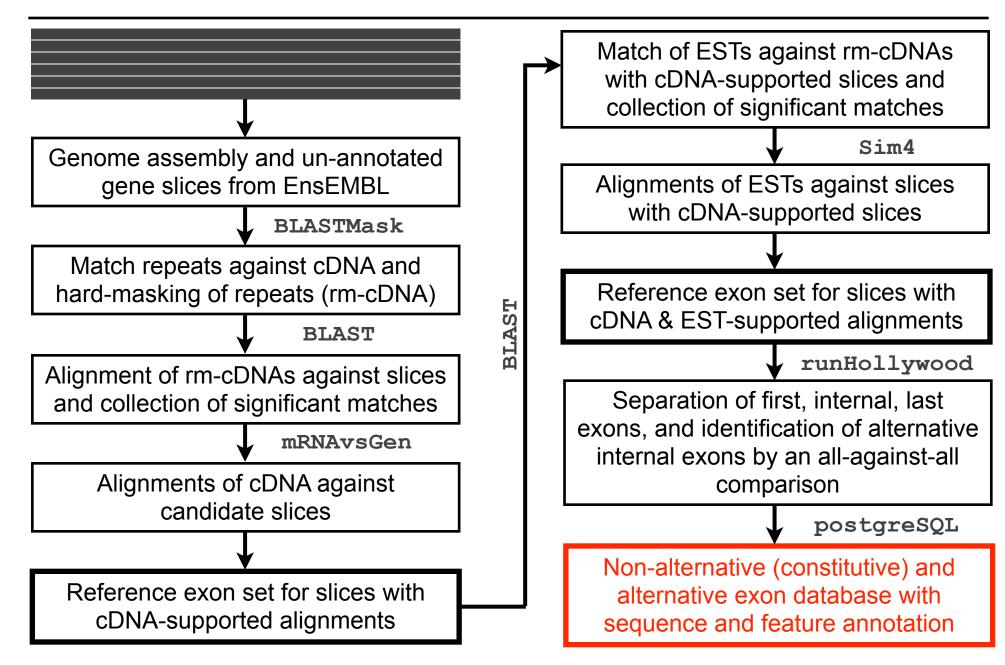




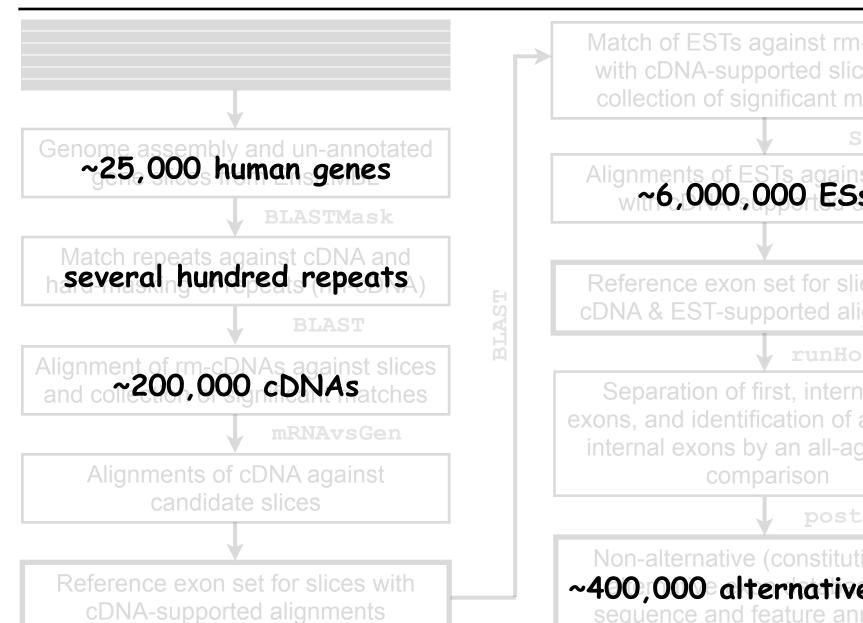
Connectivity matrix

```
"0" [5'end to 3'end of exon]
"#" [5'ss to 3'ss of intron support]
```

Pipeline for building an alternative exon database



Pipeline for building an alternative exon database



Match of ESTs against rm-cDNAs with cDNA-supported slices and collection of significant matches

Alignments of ESTs against slices ~6,000,000 ESs ces

Reference exon set for slices with cDNA & EST-supported alignments

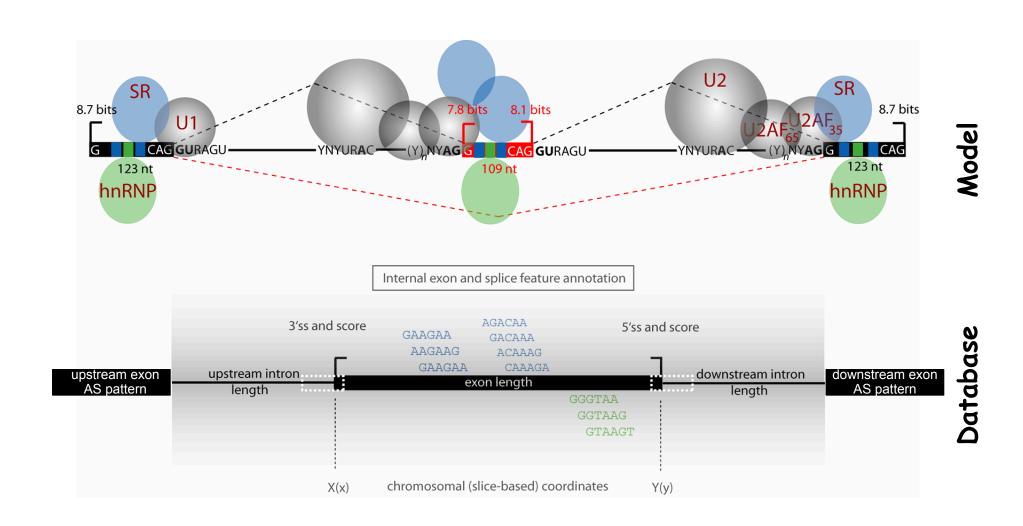
Separation of first, internal, last exons, and identification of alternative internal exons by an all-against-all

postgreSQL

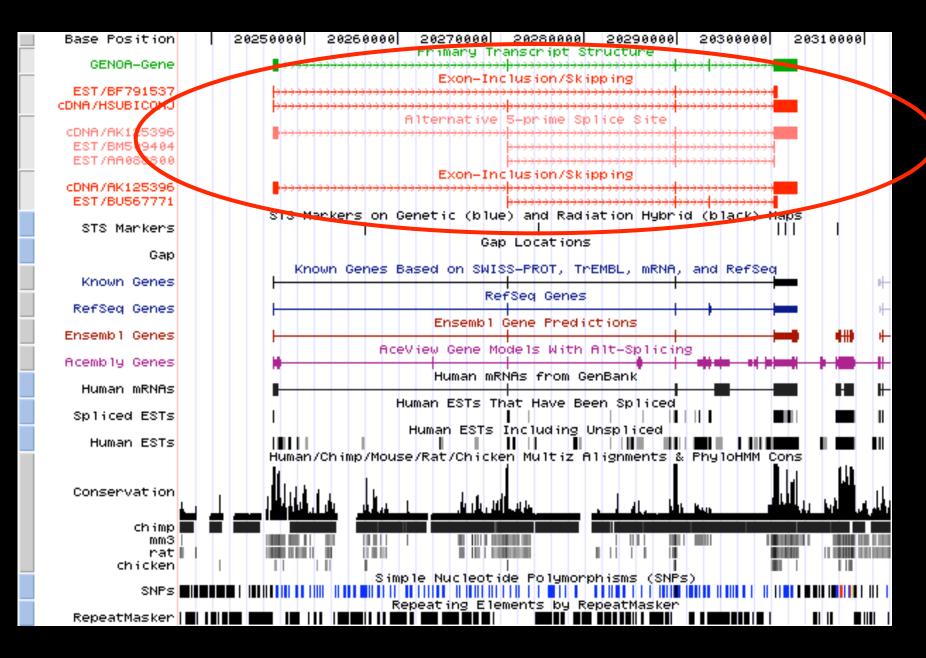
Non-alternative (constitutive) and

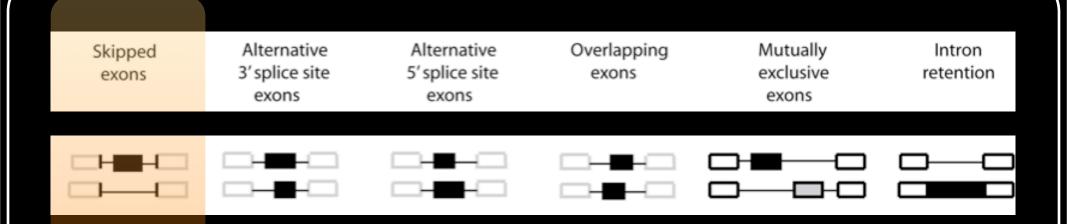
~400,000 alternative exons sequence and feature annotation

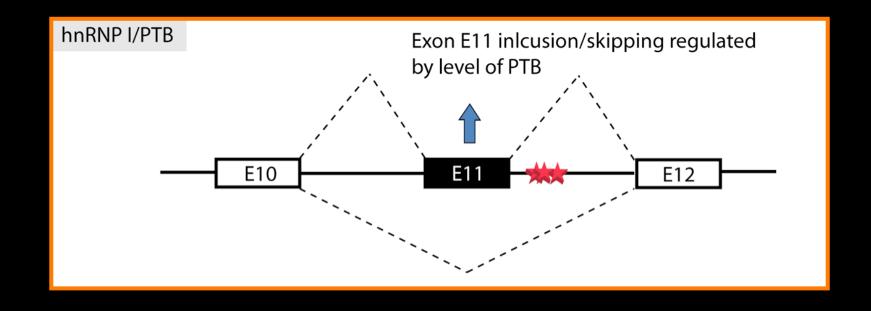
Data model

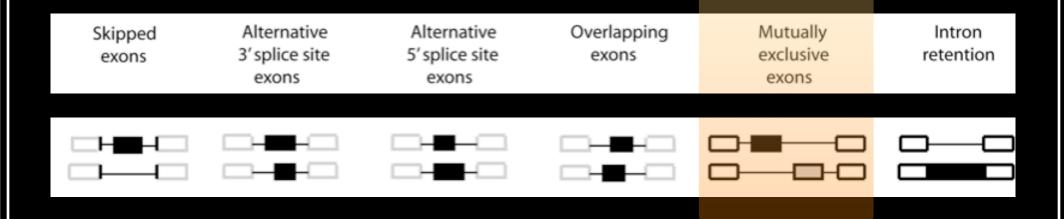


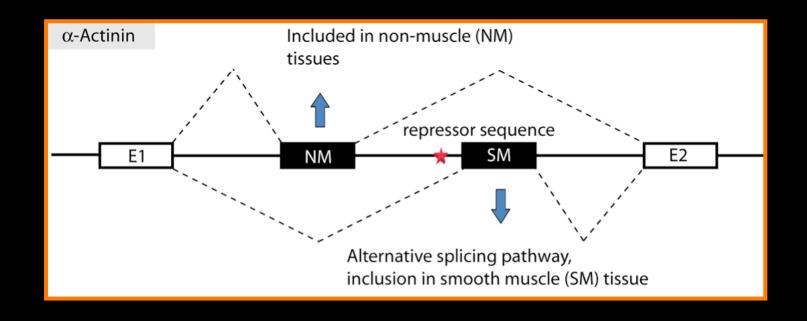
Layered onto UCSC genome browser

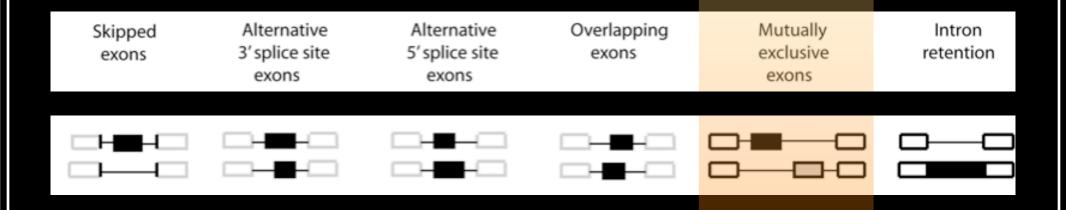


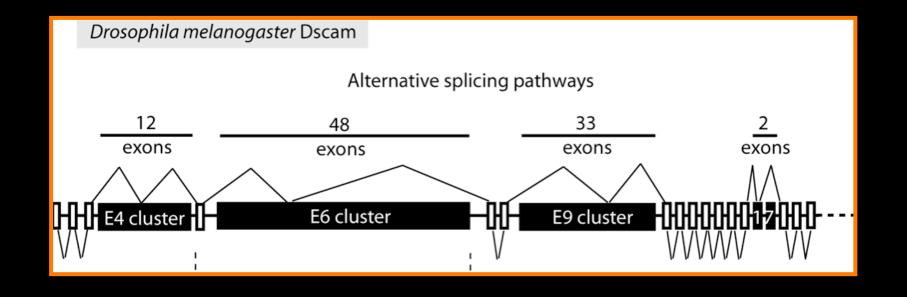


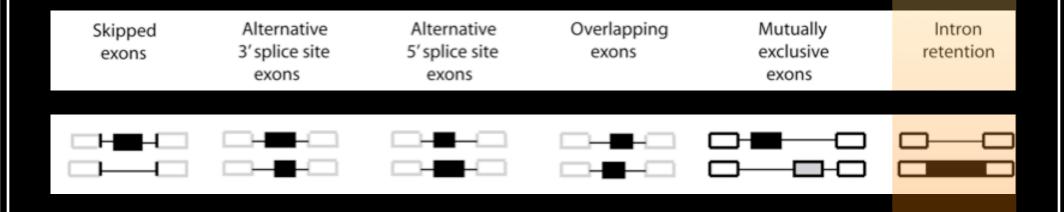


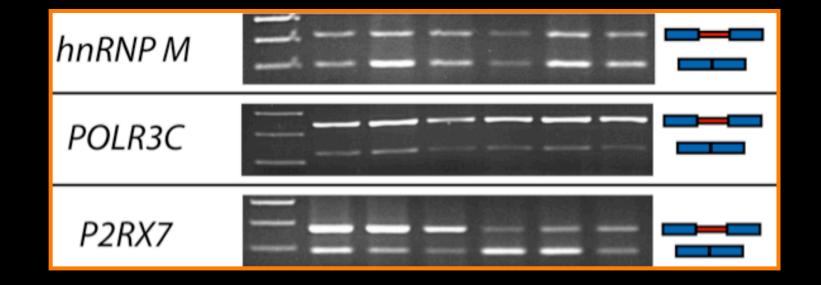


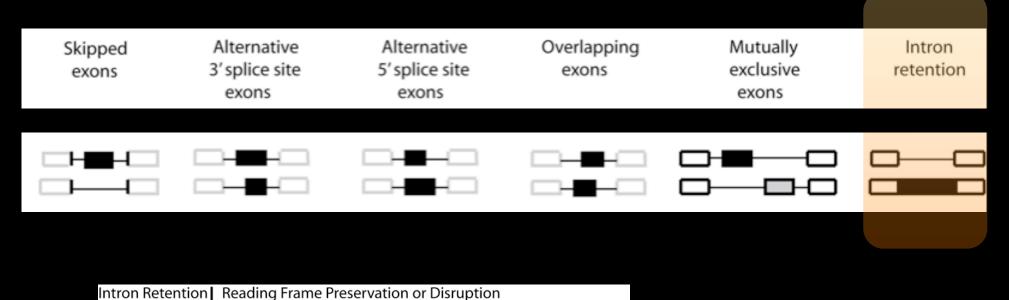


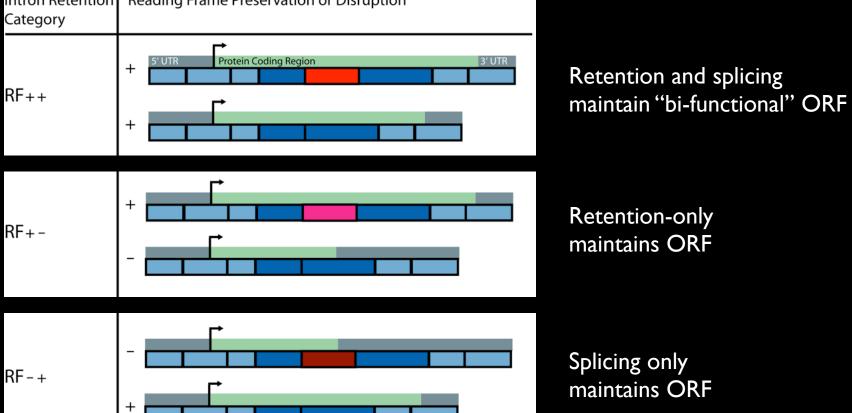


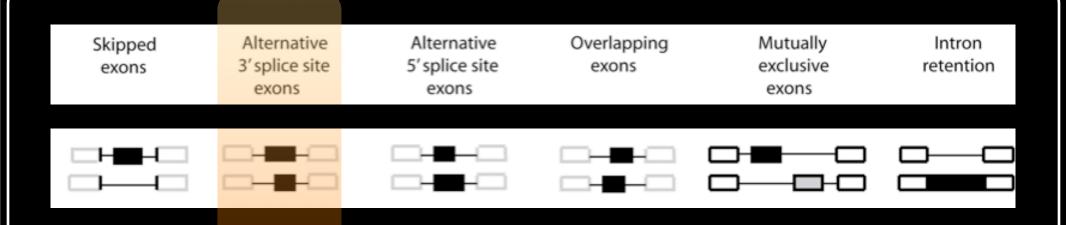


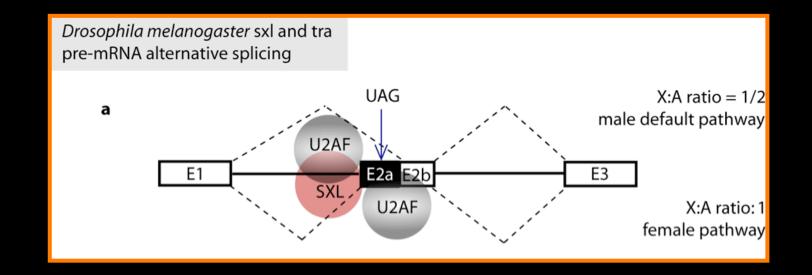




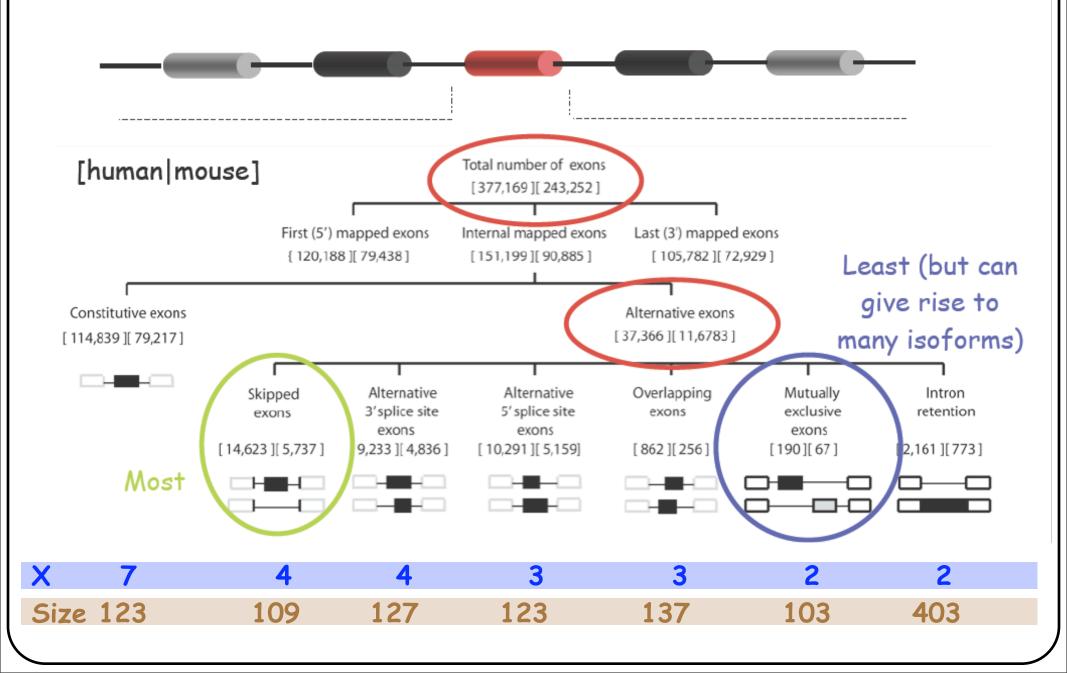




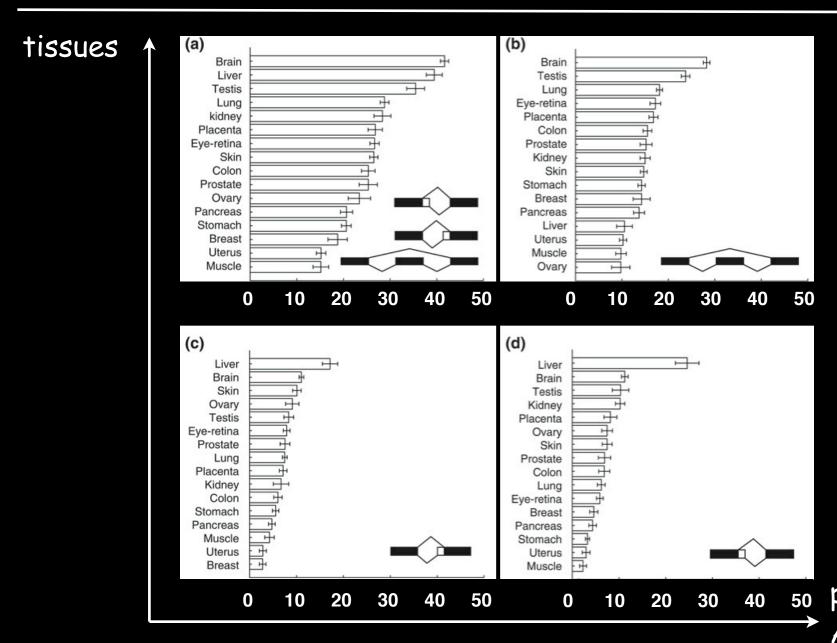




Types and abundance of mammalian AS patterns

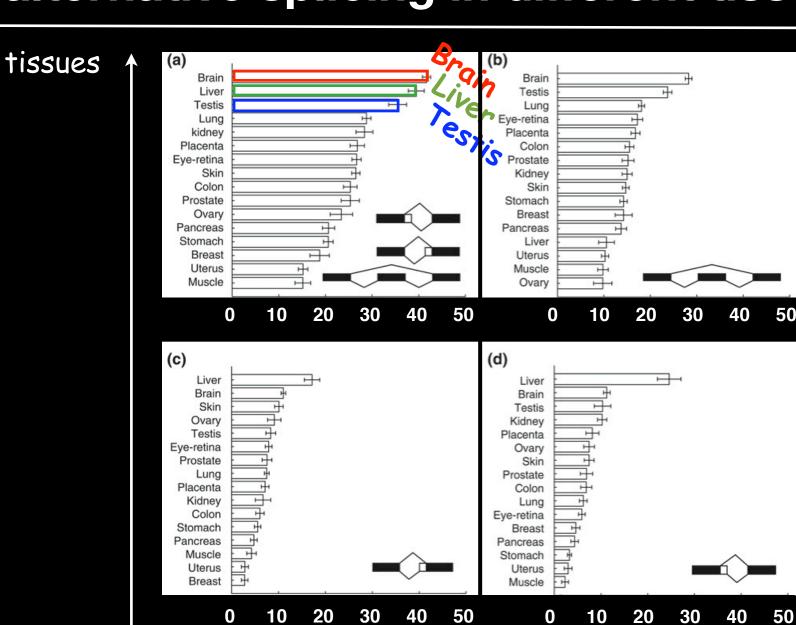


Differences in types and rates of alternative splicing in different tissues



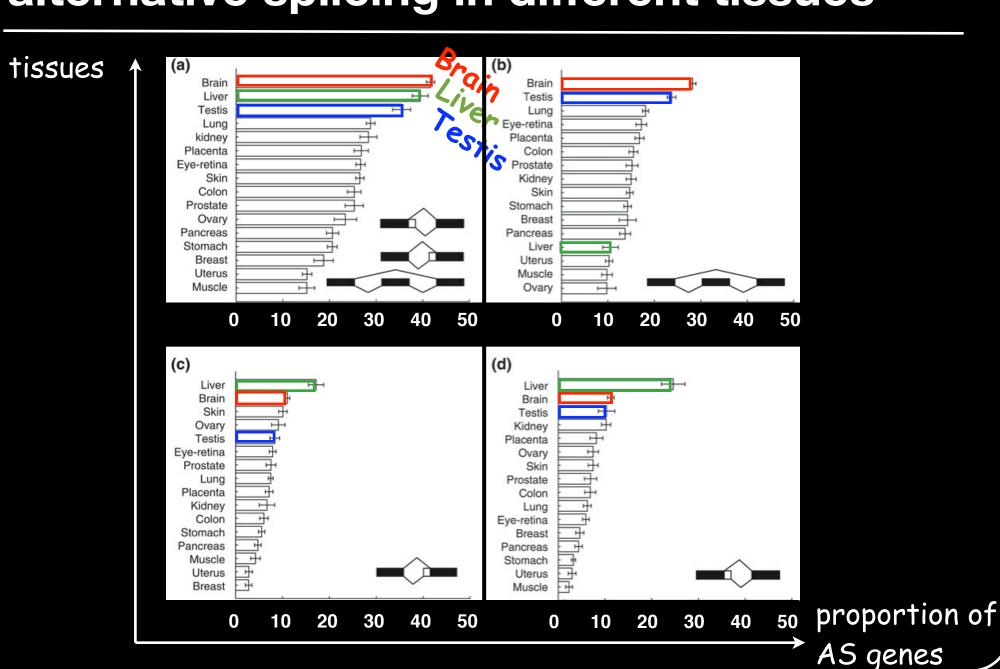
 $\xrightarrow{50}$ proportion of AS genes

Differences in types and rates of alternative splicing in different tissues

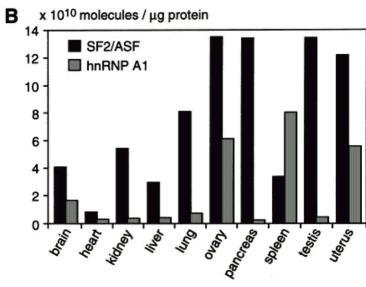


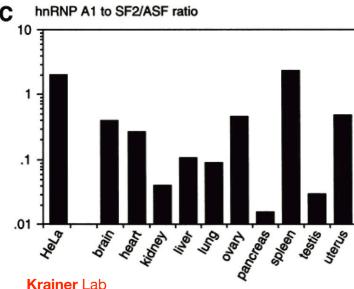
 $\xrightarrow{50}$ proportion of AS genes

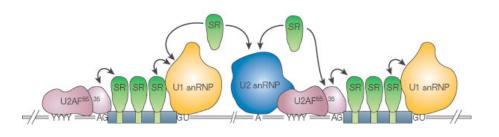
Differences in types and rates of alternative splicing in different tissues



Differences in splicing *trans*-factor expression between different tissues





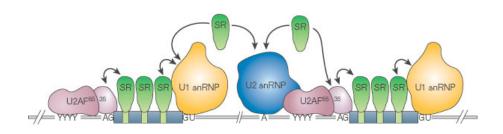






ASF/SF2	hnRNP A1
SRm300	hnRNP A2/B2
<i>SC</i> 35	hnRNPC
SRp40	hnRNP D
SRp55	hnRNP G
SRp30c	hnRNP H1
9 <i>G</i> 8	hnRNP K
SRp54	hnRNP L
SFRS10	hnRNP M
SRp20	RALY

Differences in splicing *trans*-factor expression between different tissues







ASF/SF2	hnRNP A1
SRm300	hnRNP A2/B2
<i>SC</i> 35	hnRNPC
SRp40	hnRNP D
SRp55	hnRNP G
SRp30c	hnRNP H1
968	hnRNP K
SRp54	hnRNP L
SFRS10	hnRNP M

RALY

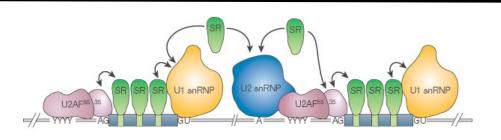
SRp20

Differences in splicing *trans*-factor expression between different tissues

Explore differences in splicing factor expression:

20 'classical' splicing factors of the SR, SR-related and hnRNP protein families

Quantify variation in gene expression between pairs of tissues by computing 20-dimensional correlation coefficient ρ between 16 plus 10 additional tissues







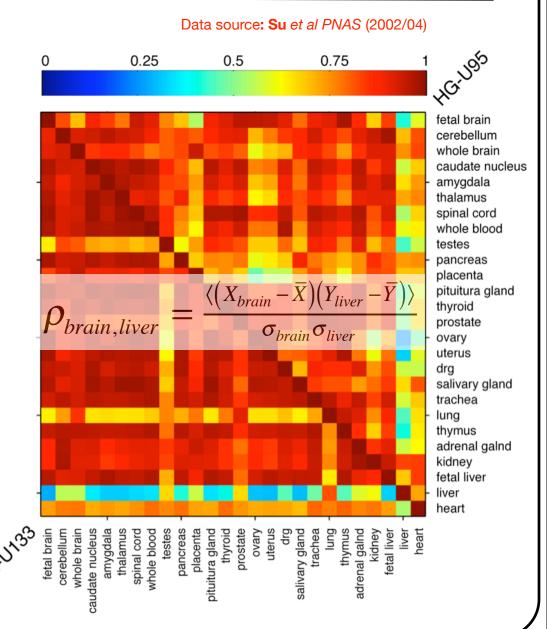
ASF/SF2	hnRNP A1
SRm300	hnRNP A2/B2
<i>SC</i> 35	hnRNPC
SRp40	hnRNP D
SRp55	hnRNP G
SRp30c	hnRNP H1
9 <i>G</i> 8	hnRNP K
SRp54	hnRNP L
SFRS10	hnRNP M
SRp20	RALY

Differences in splicing *trans*-factor expression between different tissues

Explore differences in splicing factor expression:

20 'classical' splicing factors of the SR, SR-related and hnRNP protein families

Quantify variation in gene expression between pairs of tissues by computing 20-dimensional correlation coefficient ρ between 16 plus 10 additional tissues

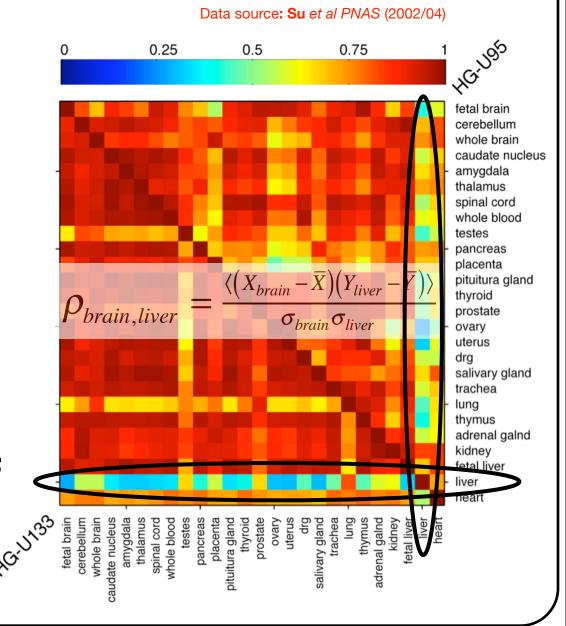


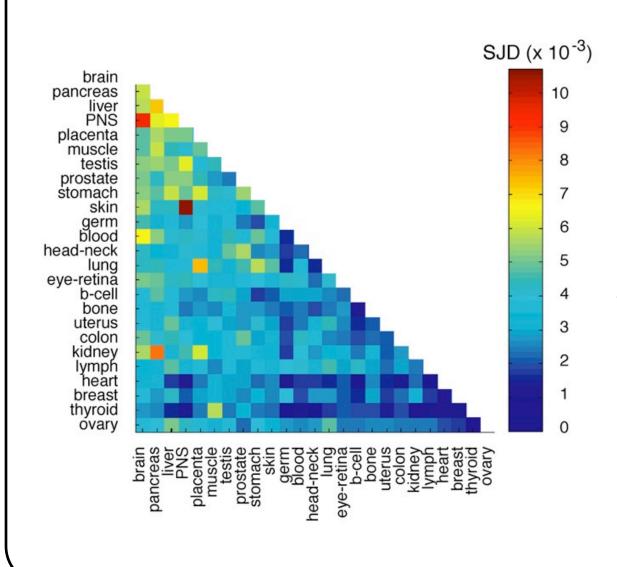
Differences in splicing *trans*-factor expression between different tissues

Explore differences in splicing factor expression:

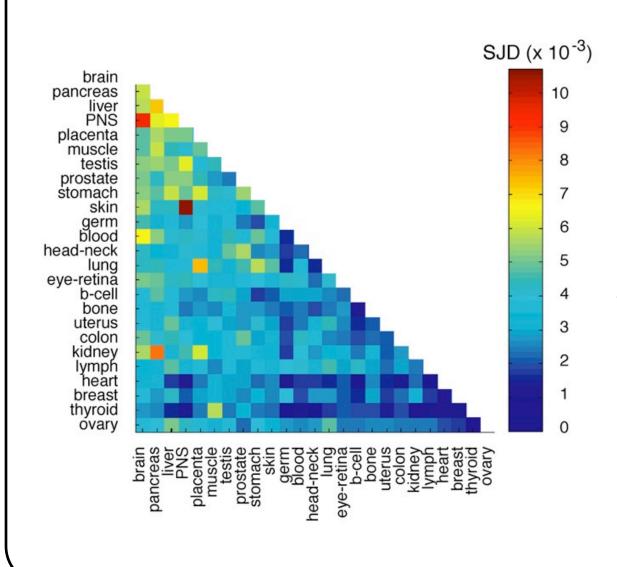
20 'classical' splicing factors of the SR, SR-related and hnRNP protein families

Quantify variation in gene expression between pairs of tissues by computing 20-dimensional correlation coefficient ρ between 16 plus 10 additional tissues

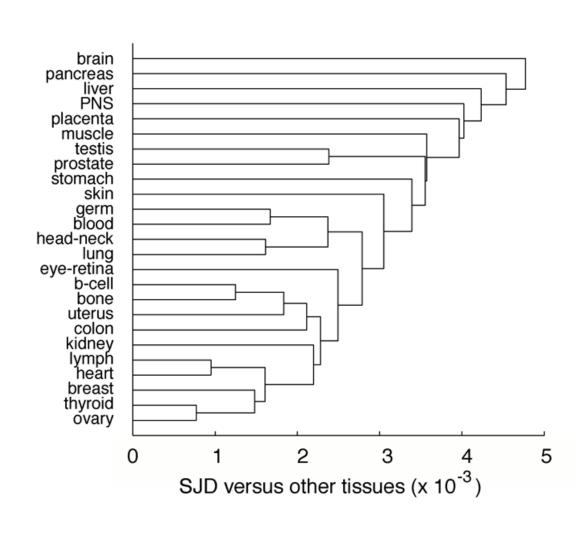




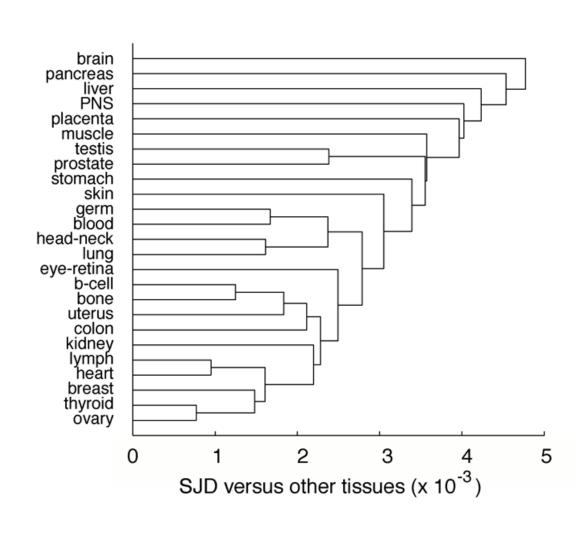
Capture the 'distance' between differentially expressed isoforms, by computing the number of splice junction differences normalized to total number of observed splice junctions (SJD)



Capture the 'distance' between differentially expressed isoforms, by computing the number of splice junction differences normalized to total number of observed splice junctions (SJD)



Can we re-identify relationships between tissues from isoform expression?



Can we re-identify relationships between tissues from isoform expression?

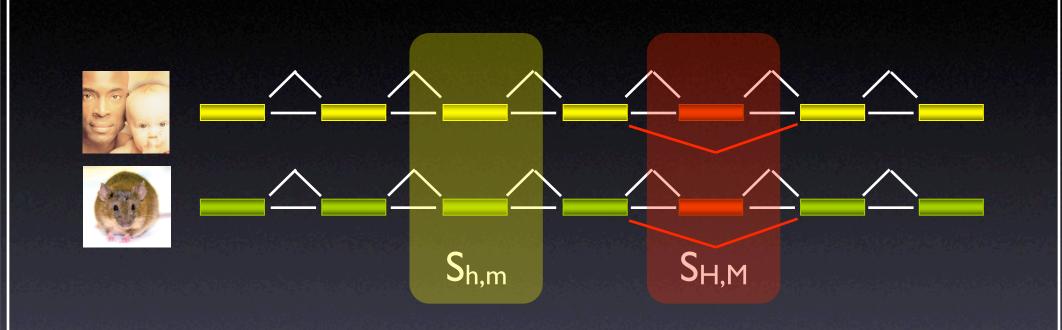
Today's menu

I. Patterns: databases and frequency of occurrence

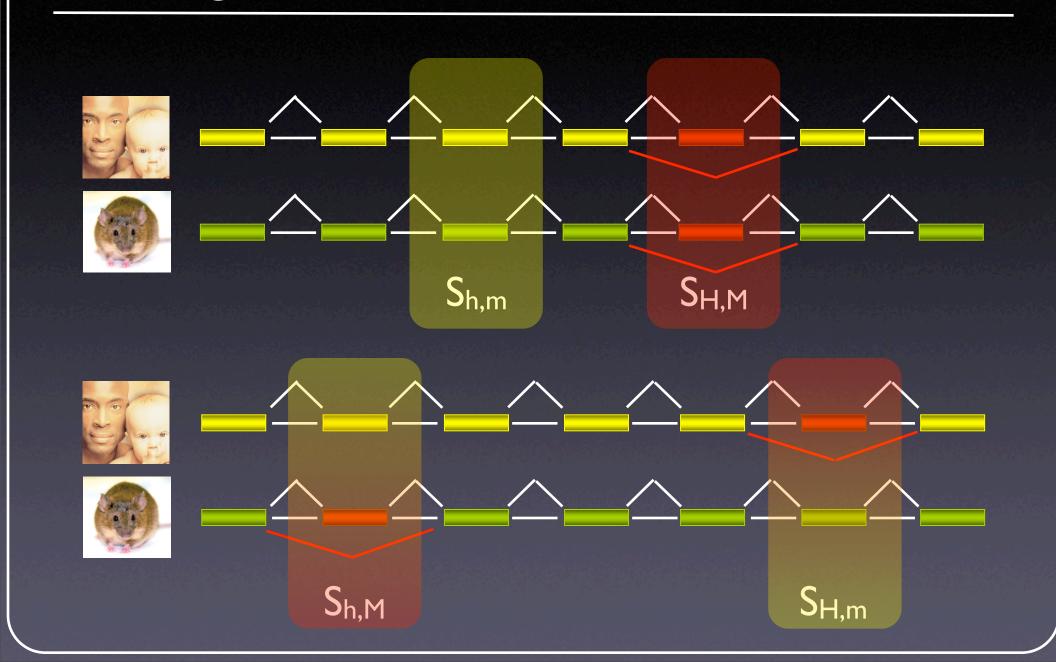
II. Prediction of evolutionary conserved alternative splicing across mammalian species

III. Alternative splicing and evolution

Evolutionary conservation of alternative splicing events



Evolutionary conservation of alternative splicing events



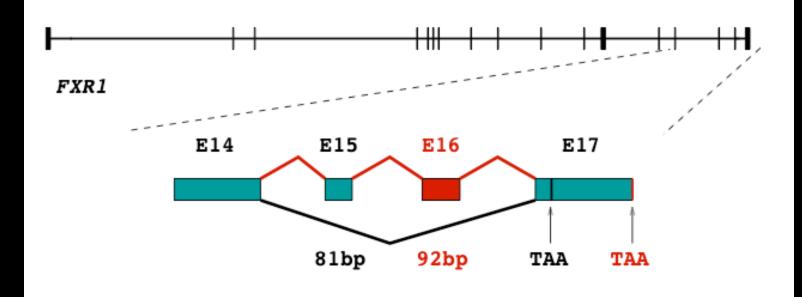
Evolutionary conservation of alternative splicing events



Is the encoded sequence information sufficient to distinguish between ACEs and constitutive exons?... can we identify & utilize intrinsic features and build a computational model that predicts ACEs?

"Despite the high fidelity of exon recognition in vivo, it is currently impossible to accurately predict alternative exons."

Thanaraj & Stamm Prog Mol Subcell Biol (2003)

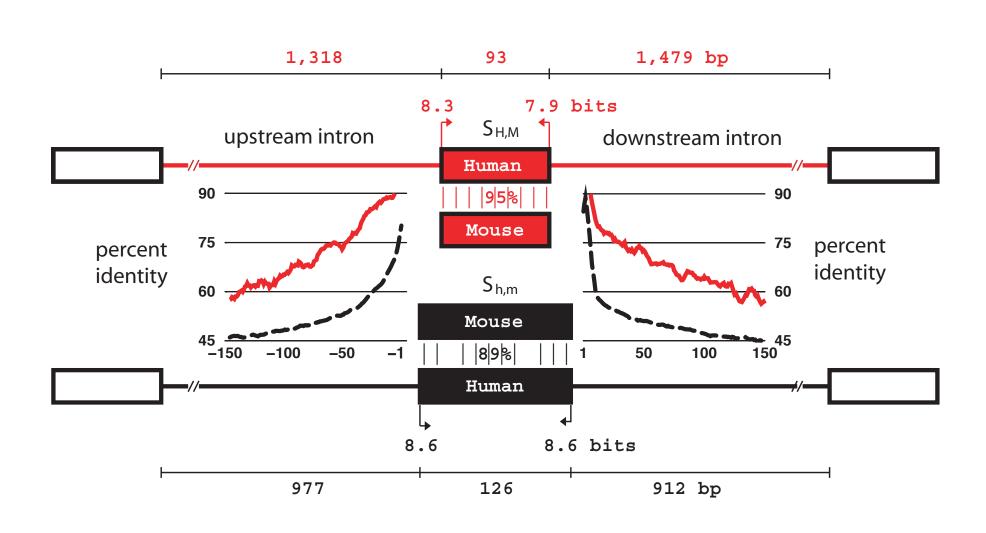




Generic features

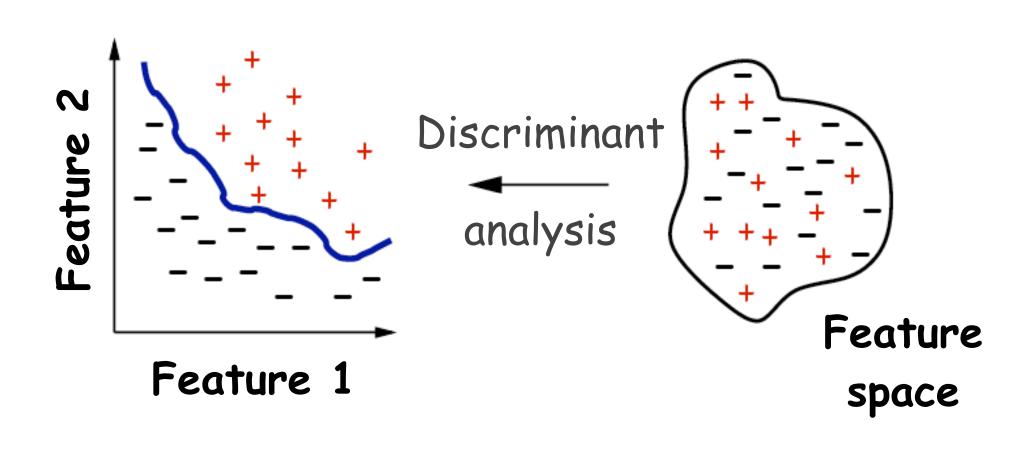
Oligonucleotide features

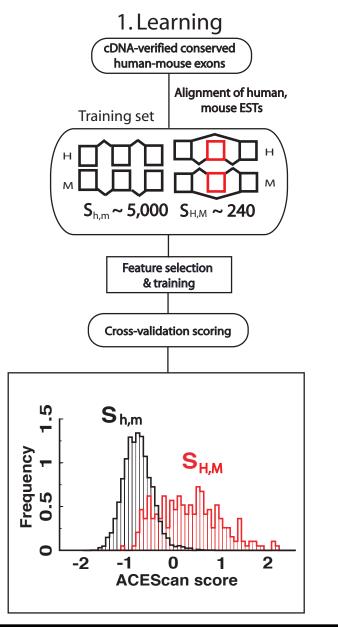
Oligonucleotide features

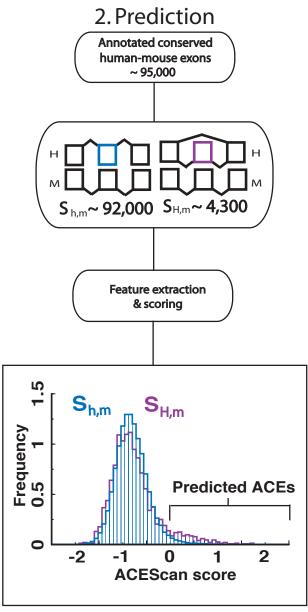


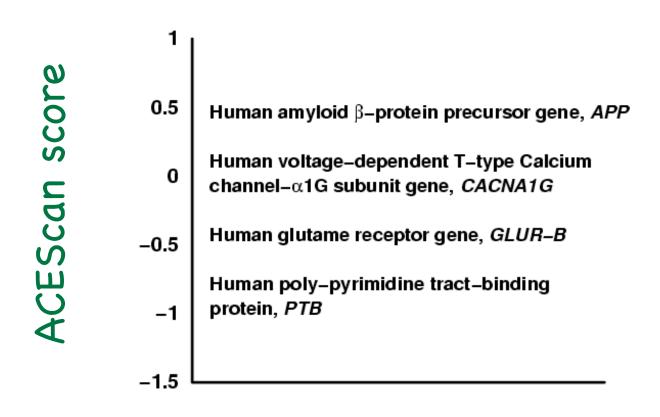
Over/underrepresented

Upstream intronic region		Exon	region		ownstrean	n intronic region	
Unaligned ———	Aligned	Exon 5' end	Alig	jned	Exon 3' end	Unaligned	Aligned
GCCGC UUUCU UGUCU UUGUC UUUU CUUU CUUU	UGCAU GCAUG CGGGG ACACU UUAAC ACUAC CCUAU UCCAU CAUAU CUAUU GUGAG	CCUCC CUCCC CCCUC CAAUC UCCCU CCCCC ACAUA CGGCG CCCGC UAACC UCUCU CAGGG CGUGG	UGUAG GUAGU ACUAG UAGAA CCUCC CGAAG CCCGC UAACG UAACG UAGUG UAGUG UAGUG CGCCG CGCCC UAGGC CACGA CGUAG CCCUA GUCGU UUACG AUCAA UGGAG	UAGGG CCUCG CGGCG UACGA UCUCU AGUAG CCGCG AGUUA UUAAU UAGGA ACUGA UCCUU GAAAG CUCGC UAAUU CAAUC CCGCU CCCCC GGCGG CUAGC CUCUC ACCUG	CCUCC CCAUC CUCCC UCCCU GUCCC UCCUU UCCCC GAAAG CCCGC CGAAG AAAAG UGAAA UGGCC CAUCA	GCAUG UGCAU CAUGC AUGCA CUAAC GUUUG ACUAA CACUA CACUU CAAAU CAUGG UAAGU UAAGA UGAGU GUAAG GUGAG	UGCAU GCAUG UCGCA UGUGC CAUUG UCGCG UCGCG CUAAC GUUUG GUGAG CAGGG GACA AUGAU GACAG
			CAUCA	GCUGG CUGGC			
						r.e	presented

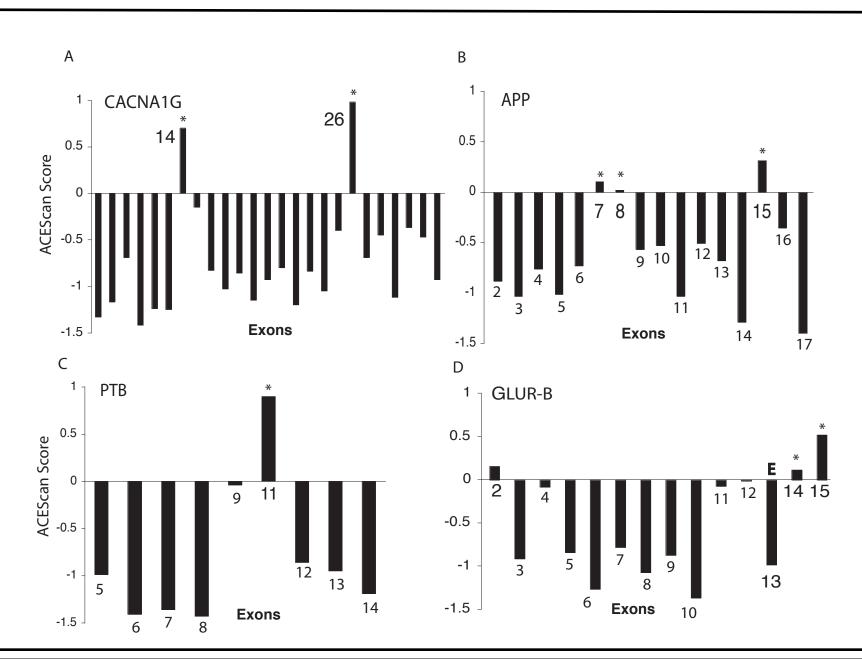








Exon number of orthologous genes



Today's menu

I. Patterns: databases and frequency of occurrence

II. Prediction of evolutionary conserved alternative splicing across mammalian species

III. Alternative splicing and evolution

Broader scope: cis-based versus trans-based hypothesis

- (1) *In-cis* emphasizes the change of produced splice patterns as an outcome of mutations in DNA
- (2) *In-trans* emphasizes the evolution and function of splicing factors

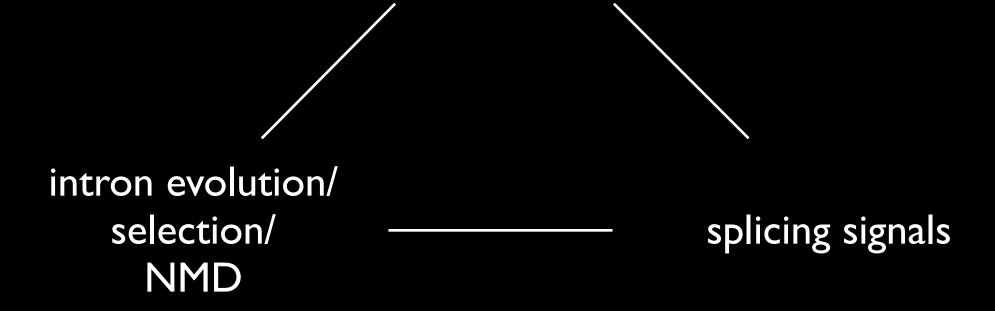
Following (I), the production of weak splice sites, deviating from the consensus that base-pairs with U1 snRNA, leads to suboptimal sites that open up differential pathways for the splicing machinery to skip an internal exon during several splicing evens --- the original isoform is kept, while an alternative isoform produces a new transcript

Following (II), the evolution of positive and negative regulatory splicing factors, such as SR proteins and hnRNPs, might have released the natural selection pressure on splice sites, while applying selection pressure on constitutive exons to undergo alternative splicing

- Splice sites of constitutive exons of higher eukaryotes are similar to lower eukaryotic organisms (yeasts)
- Splice sites of AS exons are predominantly found in higher eukaryotes
- Longer splice forms are often observed as the ancestral one

Evolution of splicing and alternative splicing

alternative exon classes/ exon birth processes



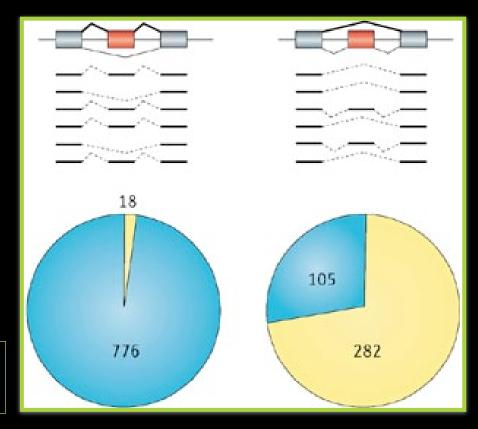
- Splicing-conserved across species (ACEs)
- Conserved, but species-specific in their splicing patterns
- Newly created alternative exons that are present in orthologous genes of other lineages

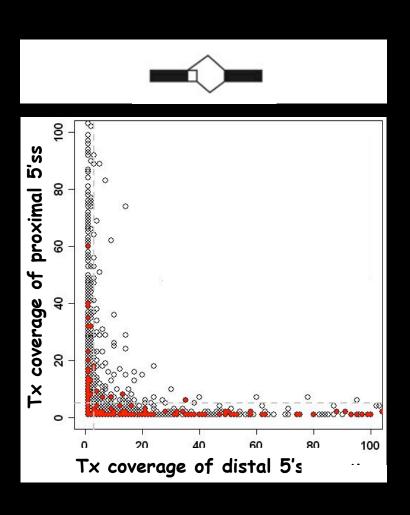
Based on transcript-inferred and predicted ACEs, the proportion of ACEs is estimated to be larger than 15%, while other estimates determine about 50% or more "conservation" (message: conservation matters)

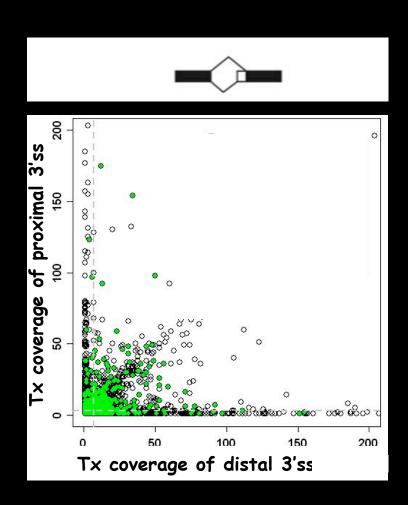
- "Major-form" exons are predominantly expressed over
- "Minor-form" exons, which are lowly expressed

Major- and minorform exons possess different properties

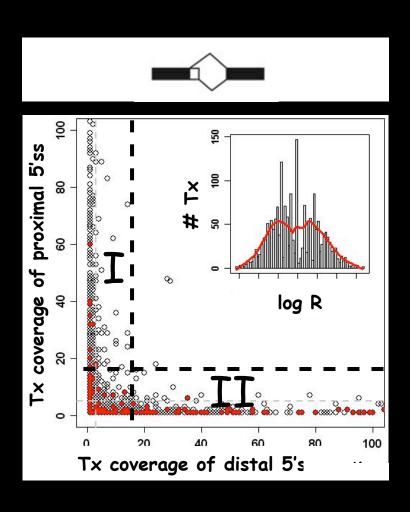
H.sapiens - M.musculus (Modrek & Lee)

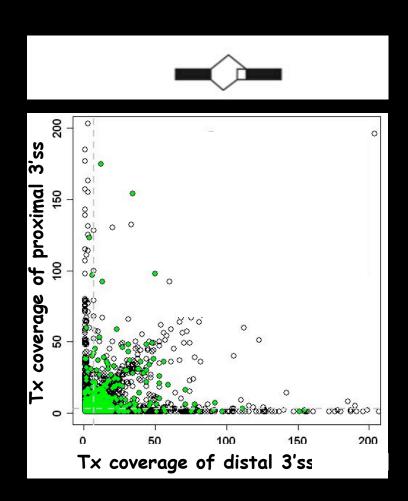




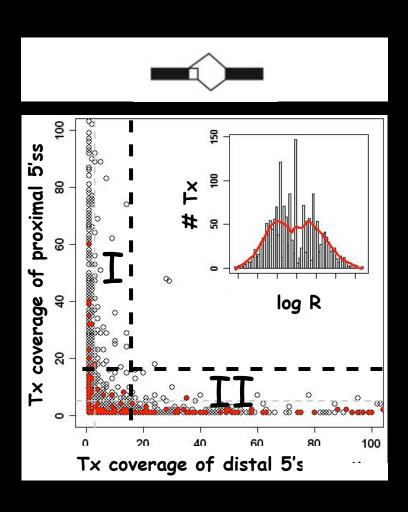


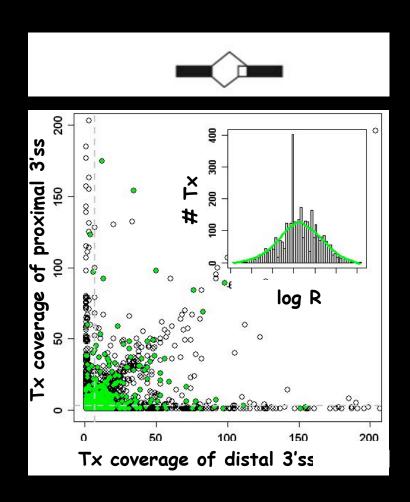
log R = transcript coverage of proximal donor / distal donor





log R = transcript coverage of proximal donor / distal donor





log R = transcript coverage of proximal donor / distal donor

Major- and minor-form exons

Modrek & Lee observed a correlation between inclusion levels of skipped exons and their splicing-conservation between orthologous human and mouse genes:

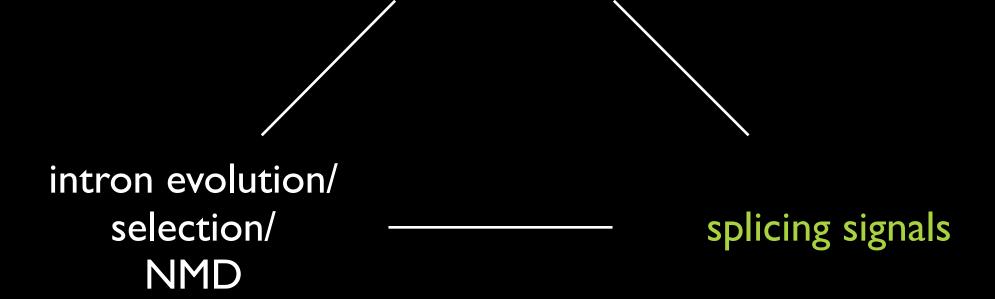
Skipped exons of the **major-form** were found to be highly splicing-conserved, while exons of the **minor-form** were often found to be lineage-specific

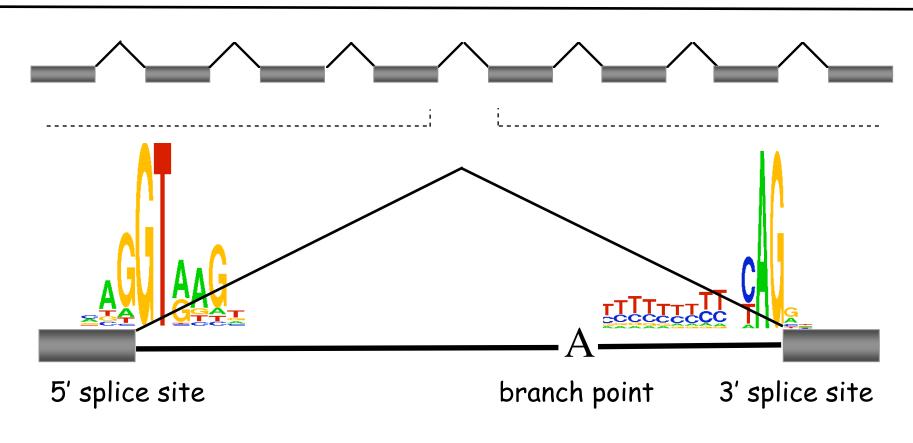
Splicing-conserved skipped exons (ACEs) show a small preference for being "modular", i.e., the reading-frame preservation ratio (number of modular SEs/ non-modular SEs) is slightly about 1/2, while minor-form coding ACEs exhibit a larger frame preservation ratio than the major-form

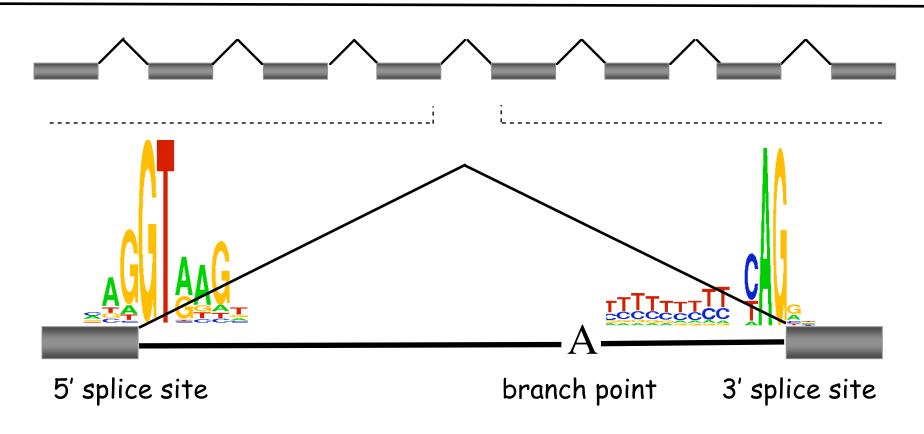
Modular skipped exons exhibit changes in their percentage of inclusions in a tissue-enriched pattern --- with the exception of minor-form exons (low conservation of AS

Evolution of splicing and alternative splicing

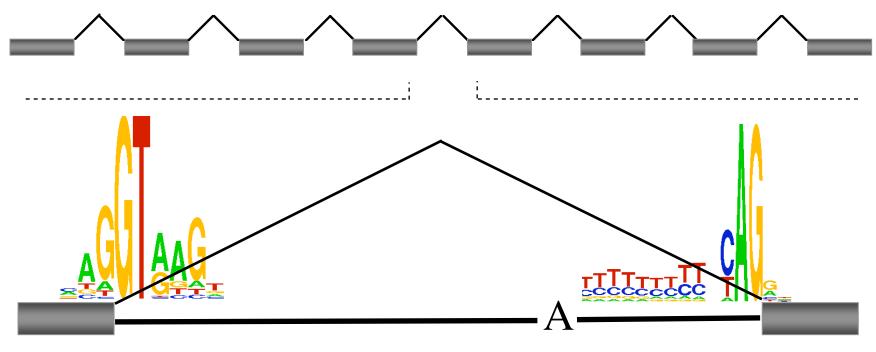
alternative exon classes/ exon birth processes







Signal/Species	5'ss	BP	3'ss	
S.cerevisiae	11	12	7	30
C.elegens	8	5	11	24
D.melanogaster	9	5	10	24
H.sapiens	8	5	8	21



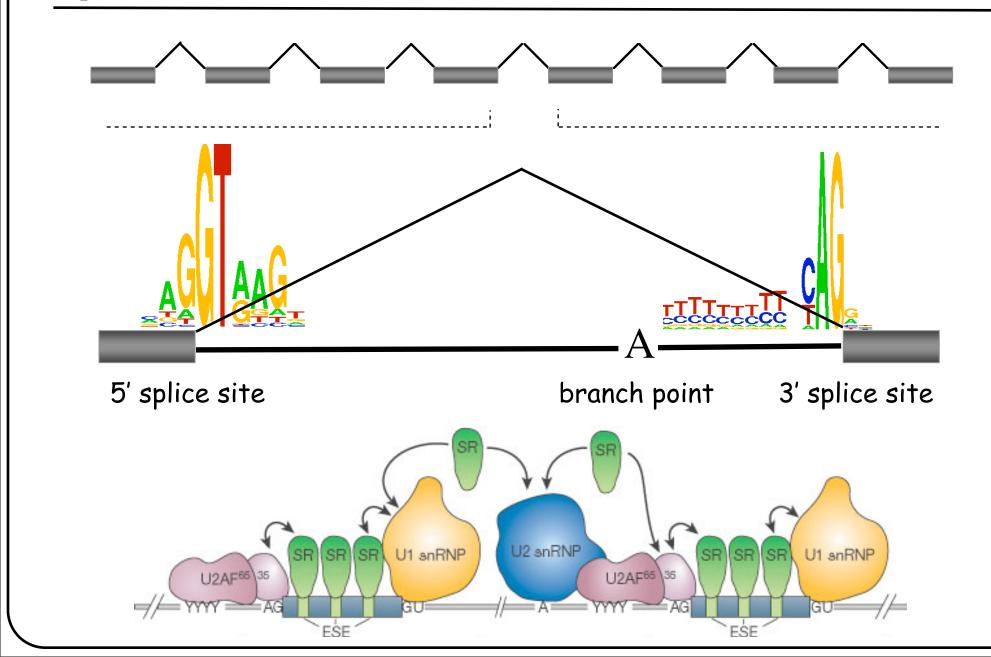
5' splice site

branch point

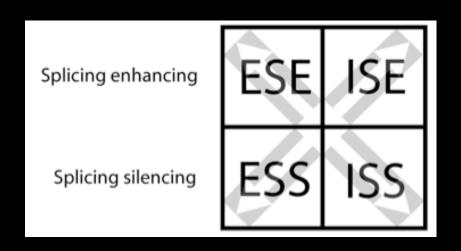
3' splice site

Signal/Species	5'ss	BP	3'ss	
S.cerevisiae	11	12	7	30
C.elegens	8	5	11	24
D.melanogaster	9	5	10	24
H.sapiens	8	5	8	21

Weaker splice sites, yet more more signals enhancing (ESE) or silencing (ESS) exons in higher eukaryotes



Splicing-enhancing and suppressing activity of nucleic elements

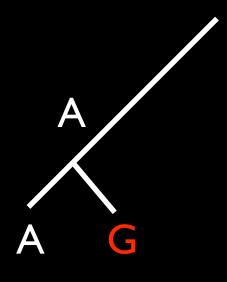


ESE: promote utilization of proximal splice site through exon-definition; R-rich ESEs recognized by members of SR protein family

ESS: may interfere with access of activators to enhancers via competitive binding; Recognized by hnRNP protein family

Splicing-regulatory signals and factors also play a role in ensuring the correct linear order of exons, after intron removal, and ESEs can act as barriers to prevent exon skipping ---

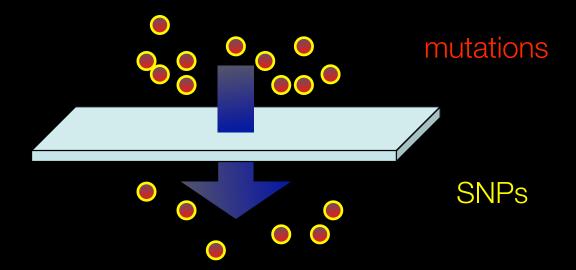
however, ESEs are also found in intronless genes, introns, and are implicated in other functions besides recruiting SR proteins

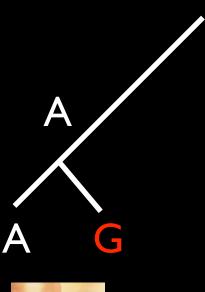




cggcgRcgata

Single nucleotide polymorphisms (SNPs) were created by mutations that survived pressure by natural selection



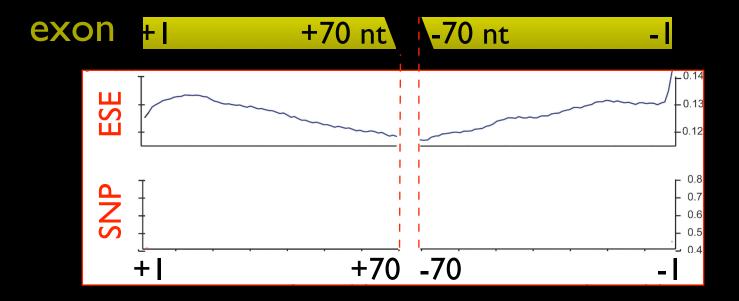




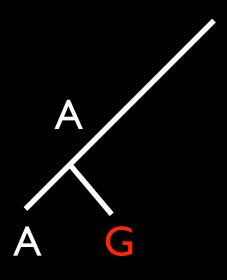
cggcgRcgata

Previously identified set of 238 hexanucleotides predicted to possess ESE activity

Fairbrother et al Science (2002)



Direction ...?

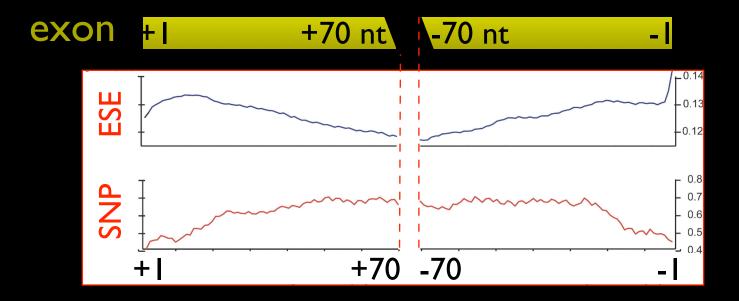




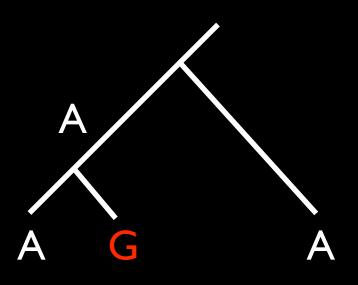
cggcgRcgata

Previously identified set of 238 hexanucleotides predicted to possess ESE activity

Fairbrother et al Science (2002)



Assess extent of natural selection acting on ESE nucleotides in human constitutive exons



Allele	Disruption	Alteration	Neutral	Creation
wildtype	+	+	_	_
mutant	-	+	-	+





cggcgRcgata					
cggcgacgata					

Assess extent of natural selection acting on ESE nucleotides in human constitutive exons

Direction ... "A" identified as ancestral allel (A6G)

Allele	Disruption	Alteration	Neutral	Creation
wildtype	+	+	_	_
mutant	-	+	-	+

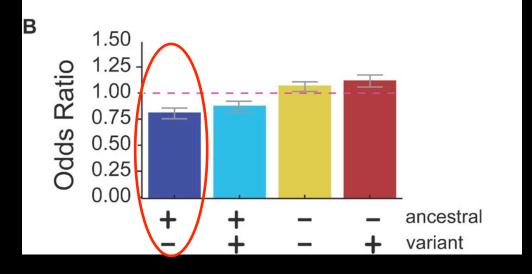
ESE Disruption

	SNP	Simulated
Yes	а	b
No	С	d

Odds ratio = a/c/b/d

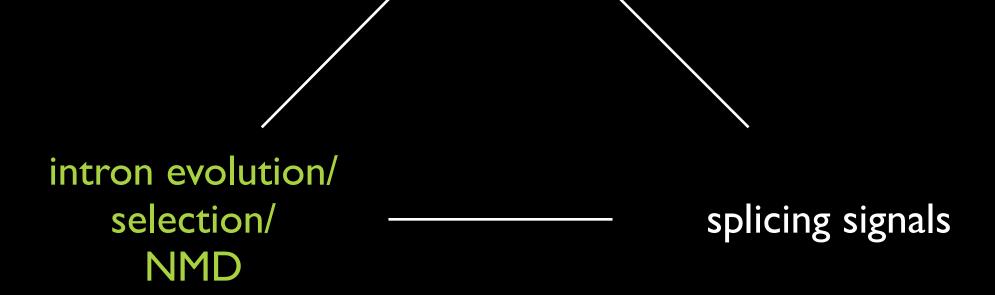
- ► About one-fifth of mutations that disrupt predicted ESEs have been eliminated by natural selection, with odds-ratio = 0.82 ± 0.05
- Selection is strongest for ESEs that are located near splice sites

Allele	Disruption	Alteration	Neutral	Creation
wildtype	+	+	_	_
mutant	_	+	_	+



Evolution of splicing and alternative splicing

alternative exon classes/ exon birth processes



Population-genetics perspective of intron evolution

Model **creation** and **retention** of introns by *random genetic drift* and *weak mutation pressure* against intron-containing alleles

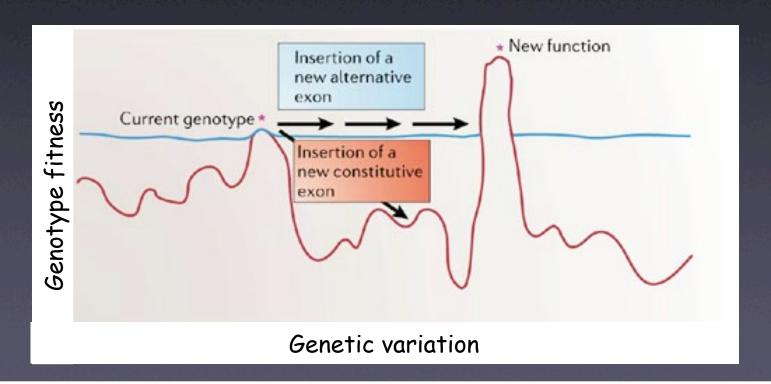
Predictions:

- Rarity of introns in unicellular organisms with large population sizes, and their expansion after origin of multicellular organisms (weak mutation pressure prevents establishment of functional introns in taxa with sufficiently large population sizes)
- Association of intron abundance and size with weak recombination frequency
- U2/U12 intron splicing
- Intron bias and NMD (enhancing retention time)

Fitness landscape adaptive walks

Differential splicing of a newly created exon can "dampen" the impact on the transcript by creating different isoforms, and convert an isoform with a low fitness value toward higher values, given that accumulated mutations become beneficial

AS may serve as "evolutionary tunneling" and provide an organism with a mechanism to explore new isoforms, only few of which eventually get fixed



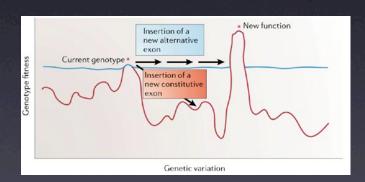
Fitness landscape adaptive walks

Exonization:

- One evolutionary pathway to create specific isoform diversity
- Acts by acquiring mutational changes to create a 5'ss, which gave rise to about ~5% of AS exons, and is estimated to affect several thousand Alus

Exon duplication:

- Estimated to affect 7-10% of genes in the human, worm, or fly genomes (often involving MXEs)
- Expect AS to be associated with the recent creation/loss of exons (potential for species-specific exons), and gene expression levels of splicing-conserved exons are correlated across different species and cell- or tissue types



Thanks to...

Stefan Schuster

Gene Yeo Eric van Notstrand Will Fairbrother Phil Sharp Chris Burge 92 bp Vivian Tung **Uwe Ohler** Ralf Bortfeld