

**From Genome-wide discoveries
of alternative splicing to
understanding its impact on the
entire proteome**

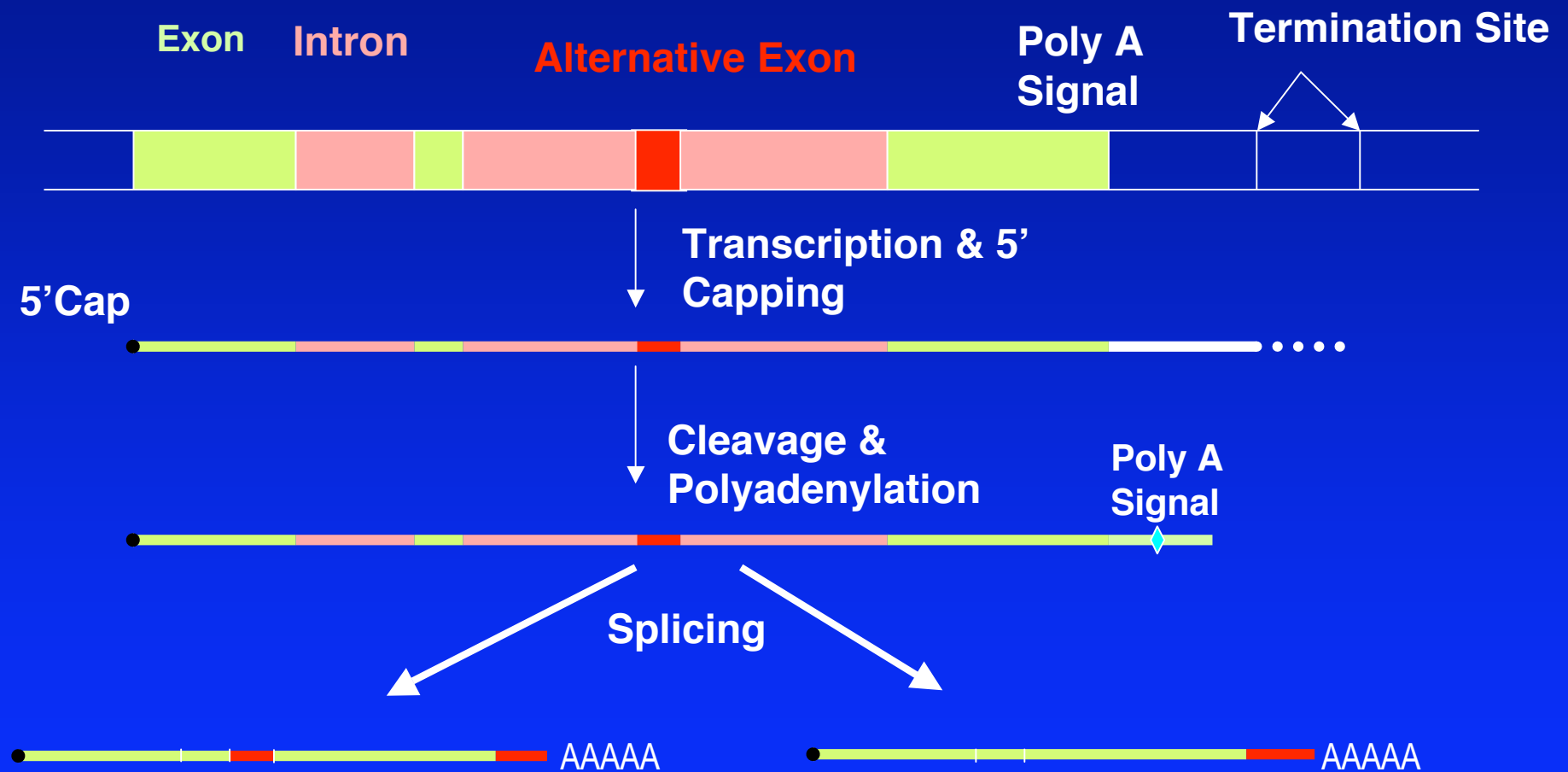
Yi Xing

Chris Lee Lab

Molecular Biology Institute

UCLA

RNA Processing & Alternative Splicing



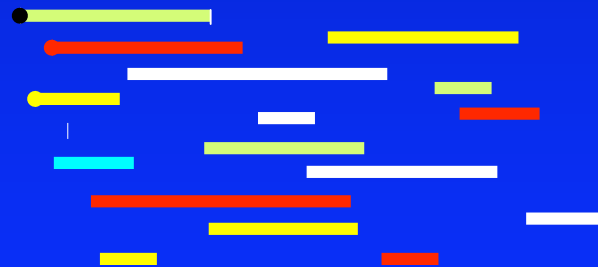
Regarded as a rare event (occurring in $< 5\%$ human genes) till late 90s

EST Sequencing: a shortcut to genes

Mix of mRNAs

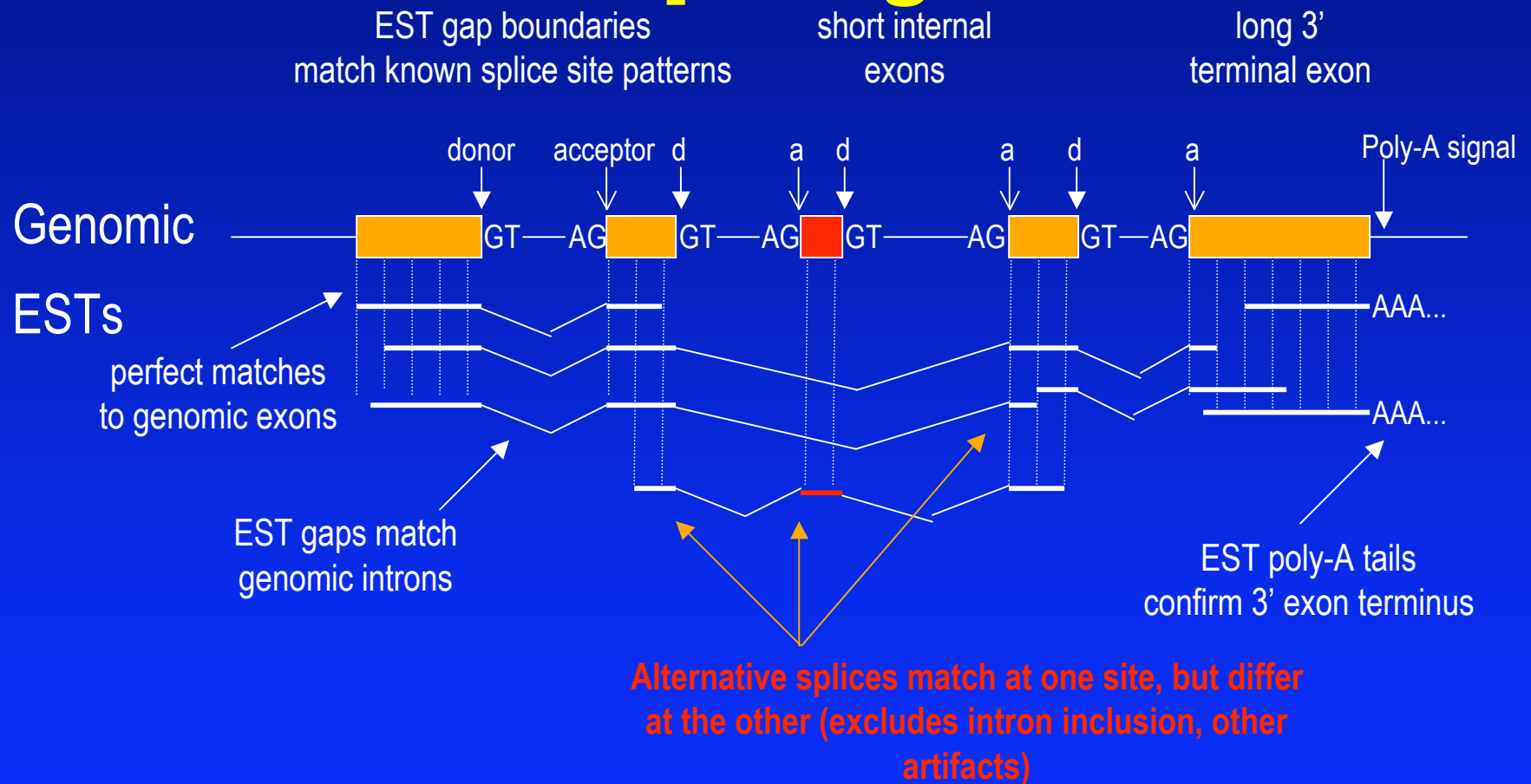


Shotgun sequencing
of random fragments
(ESTs)



Can be used for detecting SNPs, alternative splicing, etc.

mRNA/EST + Genomic Analysis of Splicing



Modrek ,Resch, Grasso and Lee, *Nucleic Acids Res.* 29, 2850 (2001)

Modrek & Lee, *Nature Genetics* 30:13-9 (2002).

Grasso , Modrek, Xing and Lee, *Pac. Symp. Biocomput.* (2004)

Alternative splicing expands human proteome diversity

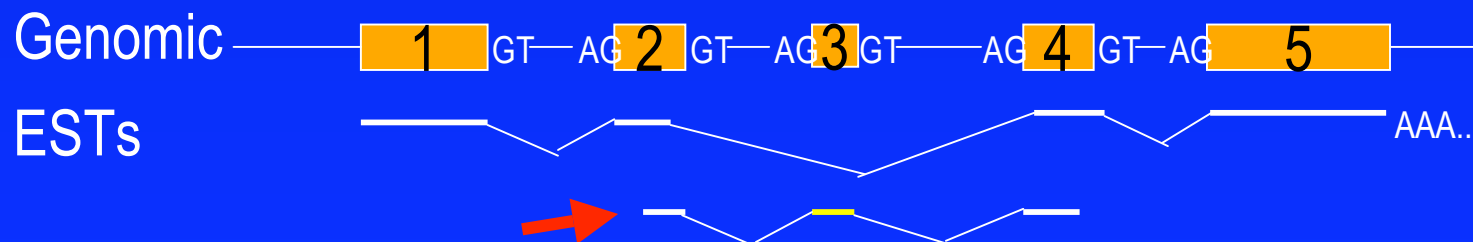
	All			Genes with mRNA		
	Number	Clusters		Number	Clusters	
Total UniGene Clusters		96109			20817	
Mapped to Draft Genome		68032	71%		17552	84%
Detected Splices	133369	18173	27%	121172	12537	71%
Alternative Splice Relationships	30793	7991	44%	28947	7143	57%

- More than half of human genes undergo RNA alternative splicing
- Alternative splicing greatly expands human proteome diversity

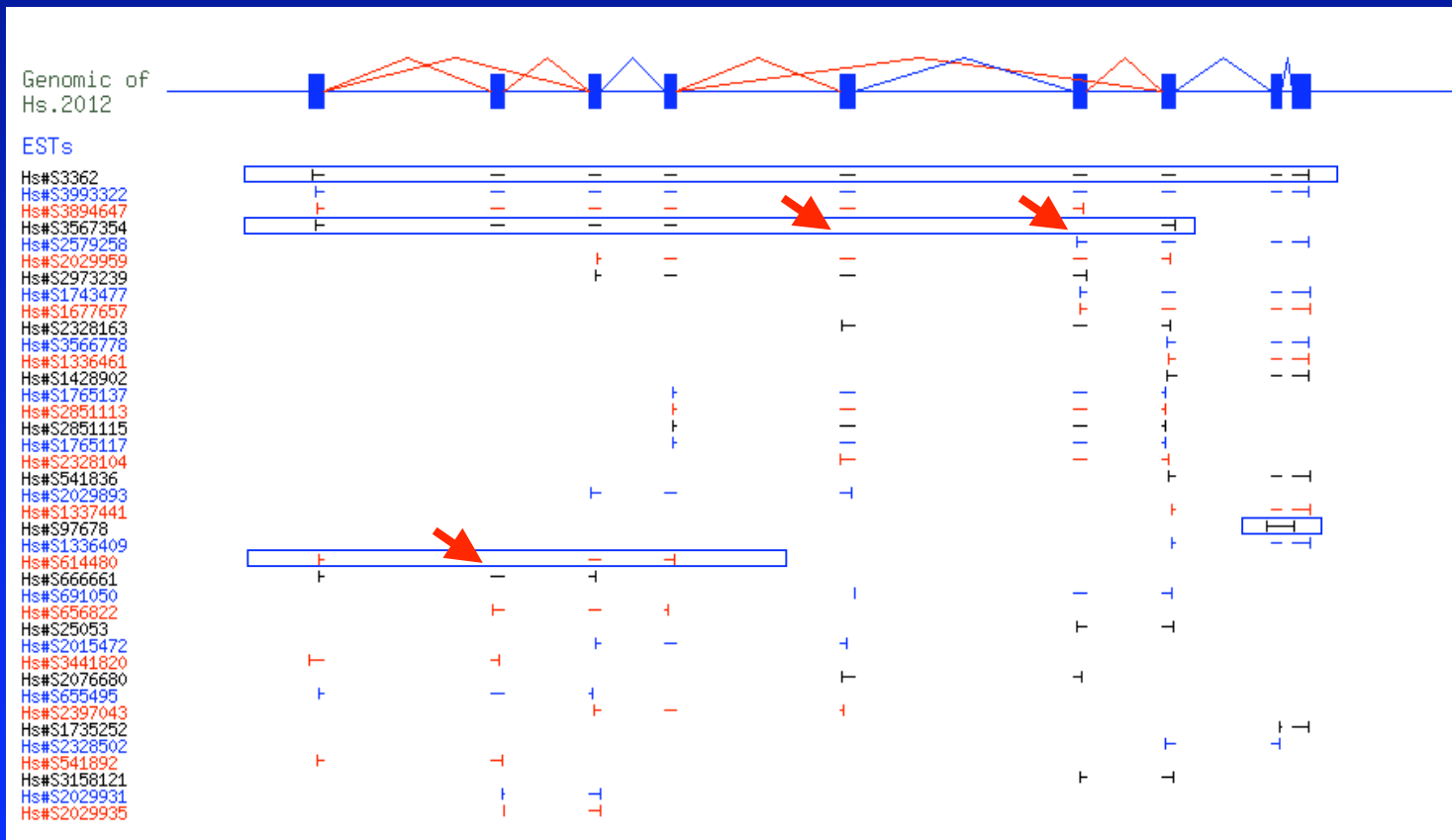
What are the functional impacts of those alternative splicing events?

Assessing the functional impact of alternative splicing requires *full length protein isoforms*

- The majority (over 80%) of alternative splicing events are detected from EST data
- ESTs are sequence fragments which can only tell us *local* information of the gene structure.
- To infer impact on protein, we need *full-length protein isoform* sequences



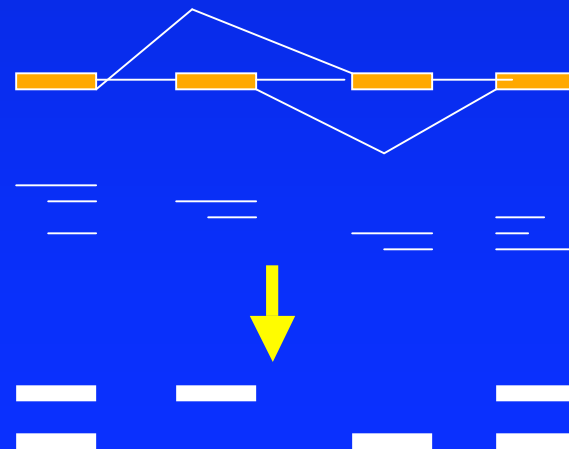
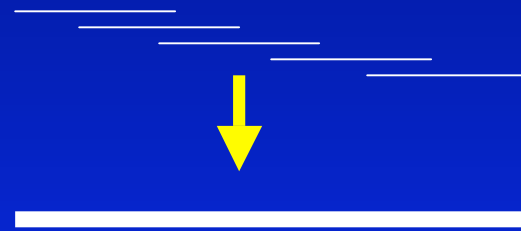
EST-Genomic Alignments of Hs.2012 (TCN1)



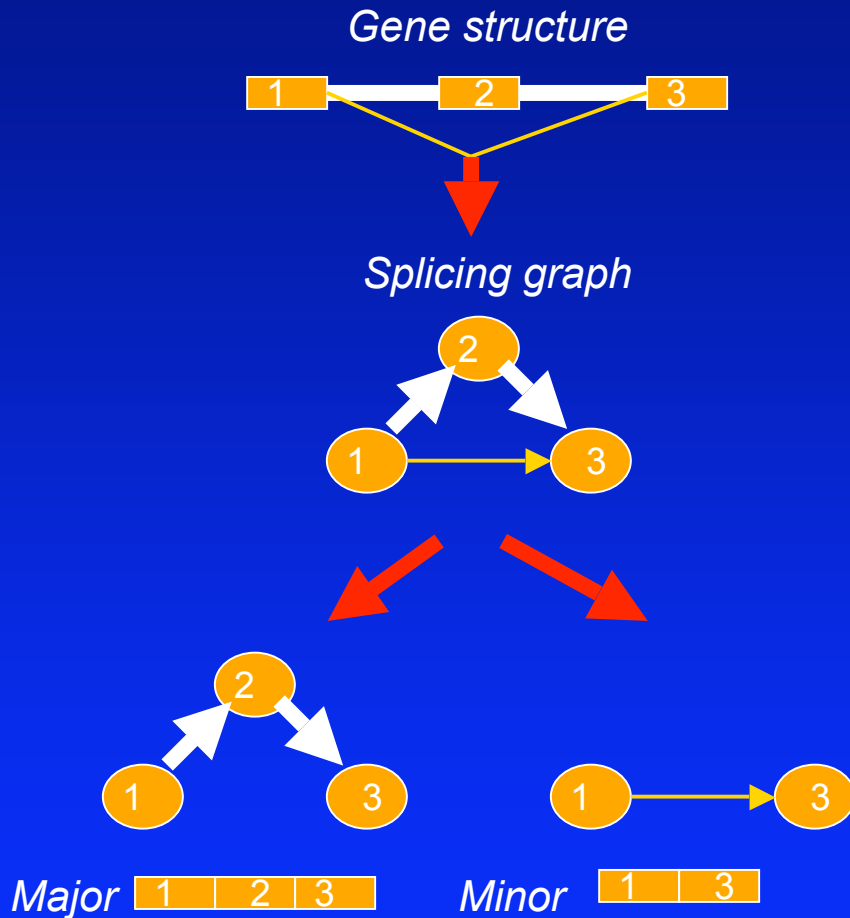
This is an EST assembly problem!

Assembling Multiple Consensus Sequences from ESTs

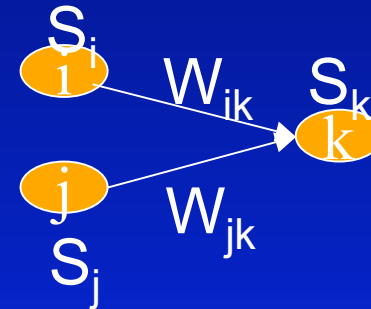
- Traditionally, a gene structure is represented as a linear string of exons; the purpose of EST assembly is to generate a linear path across all the sequences as the consensus sequence
- Alternative splicing represents branches in the gene structure, which breaks the assumption for a single consensus sequence



Splice Graph



heaviest bundling (HB) algorithm



If $W_{ik} > W_{jk}$, $S_k = S_i + W_{ik}$

If $W_{ik} < W_{jk}$, $S_k = S_j + W_{jk}$

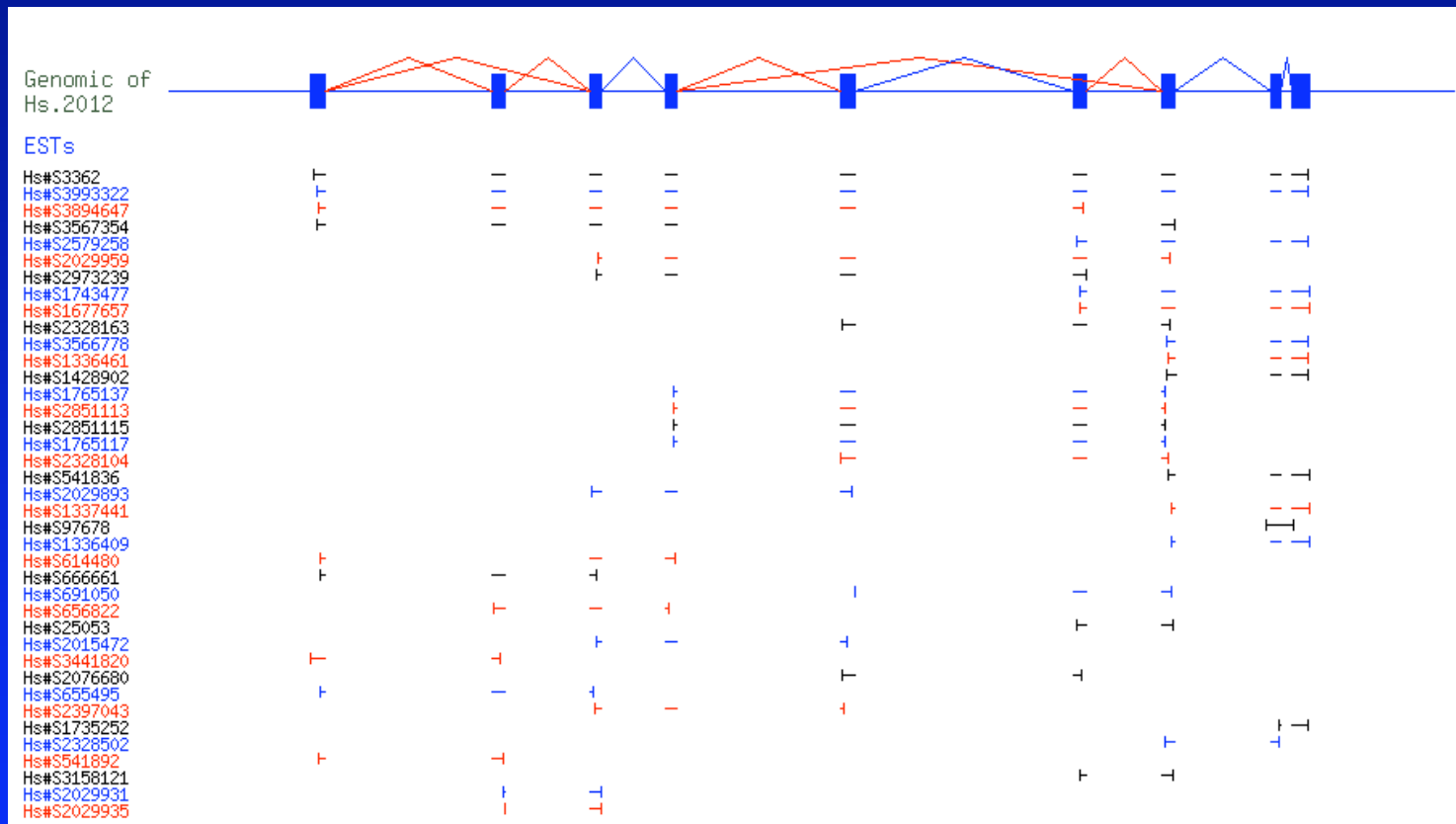
If $W_{ik} = W_{jk}$, $S_k = W_{ik} + \max(S_i, S_j)$

Lee, *Bioinformatics* 19, 999 (2003)

Isoforms are defined as start-end traversals across the splice graph

Heber S et al., *Bioinformatics*, 2002 18(Suppl 1):S181-8
 Lee C et al., *Bioinformatics*, 2002 18(3):452-464

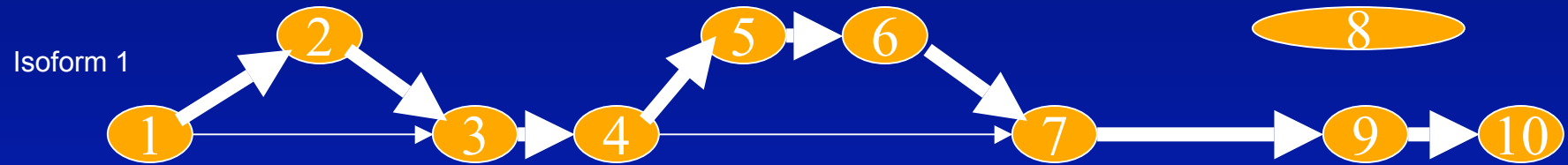
How can we construct isoforms from the splice graph?



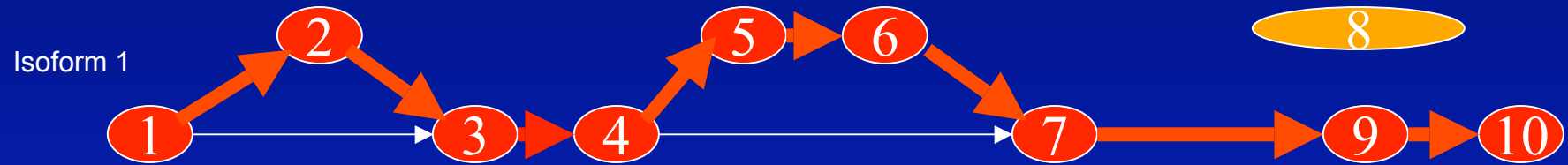
Liberal vs Conservative Strategies for Isoform Assembly

- Liberal
 - Enumerating all possible traversals across the splice graph (Heber *et.al.*)
 - Maximize coverage; High false positives for generating combinations of exons that don't really occur
- Conservative
 - Generating the minimal set of isoforms sufficient to explain all the EST data (Xing *et.al.*)
 - A maximum likelihood solution

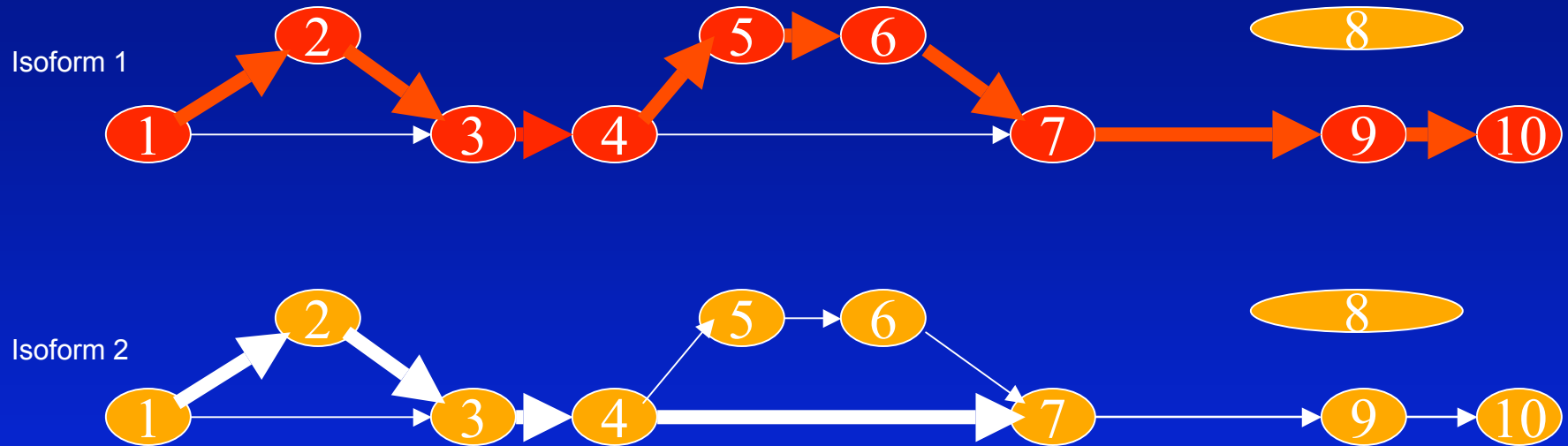
Splice Graph and Isoform Generation of TCN1



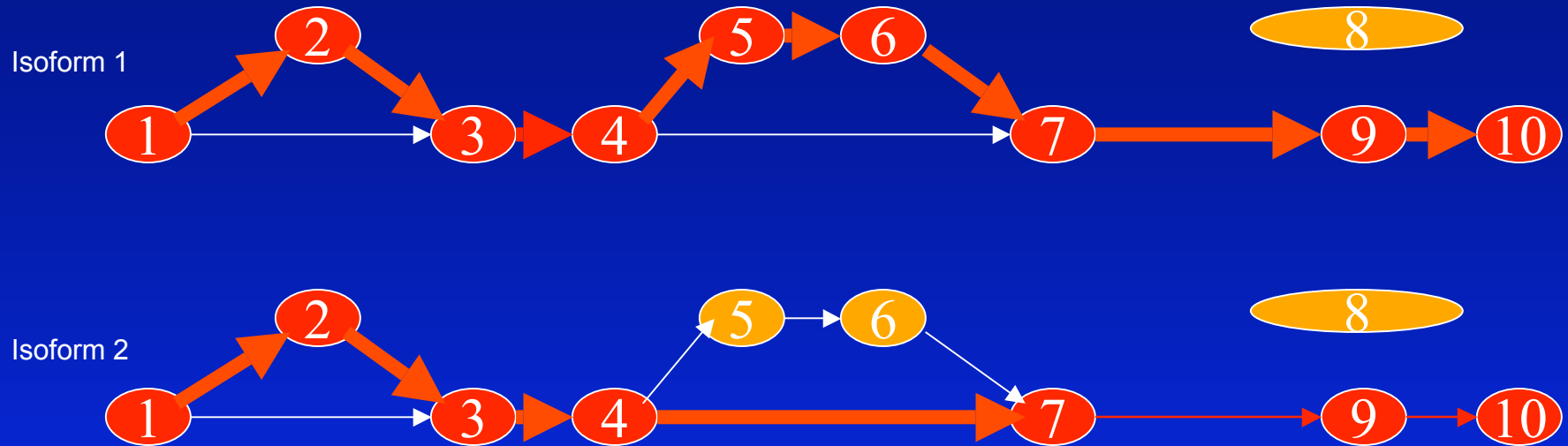
Splice Graph and Isoform Generation of TCN1



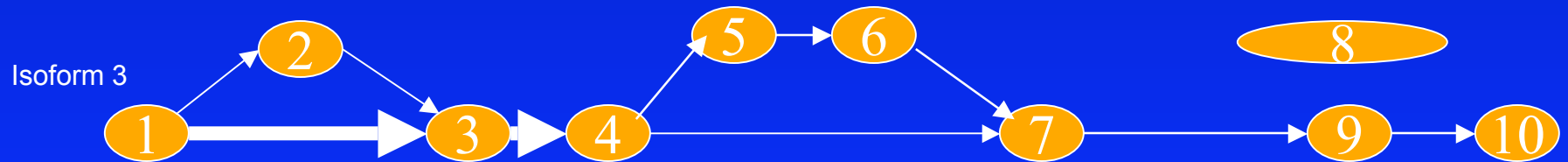
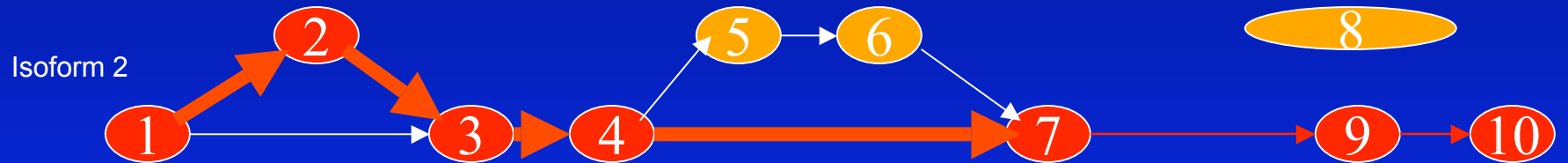
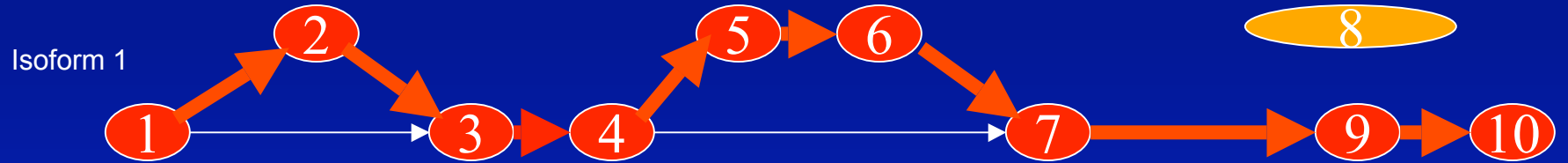
Splice Graph and Isoform Generation of TCN1



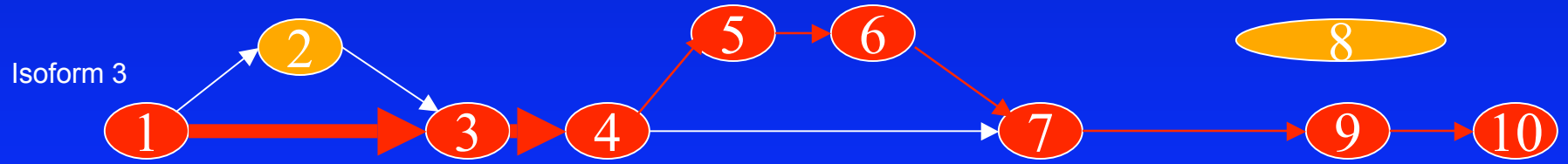
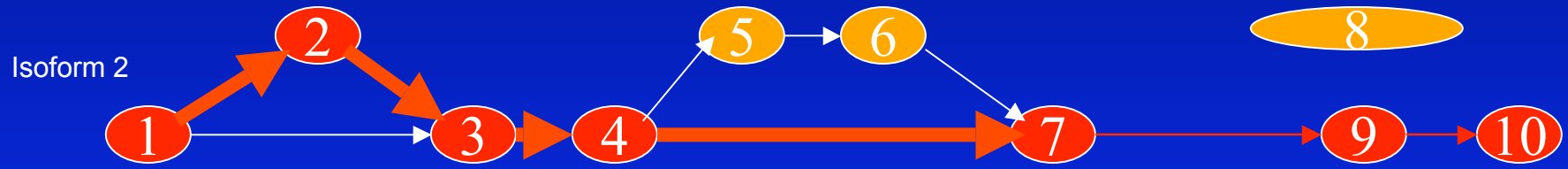
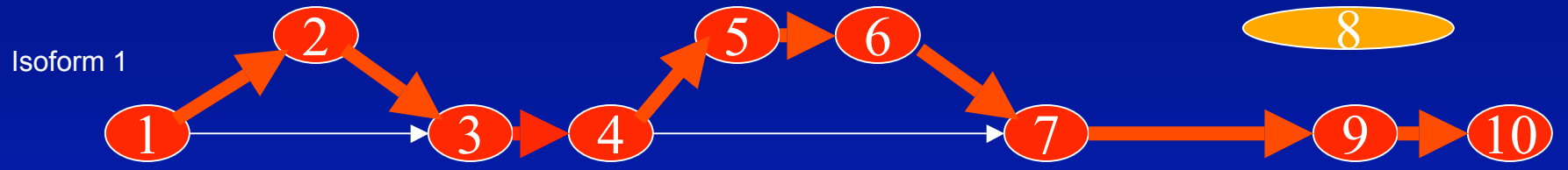
Splice Graph and Isoform Generation of TCN1



Splice Graph and Isoform Generation of TCN1



Splice Graph and Isoform Generation of TCN1



Database of Alternatively Spliced Proteins in Human

- We constructed a database of alternatively spliced proteins (ASP) for human, consisted of 13384 distinct protein isoforms for 4422 human genes, many of them are novel isoforms.
- A useful resource for validation and functional studies of novel full-length isoforms , and for large scale analyses to assess the impact of alternative splicing on the human proteome (e.g. domain architecture, transmembrane association)

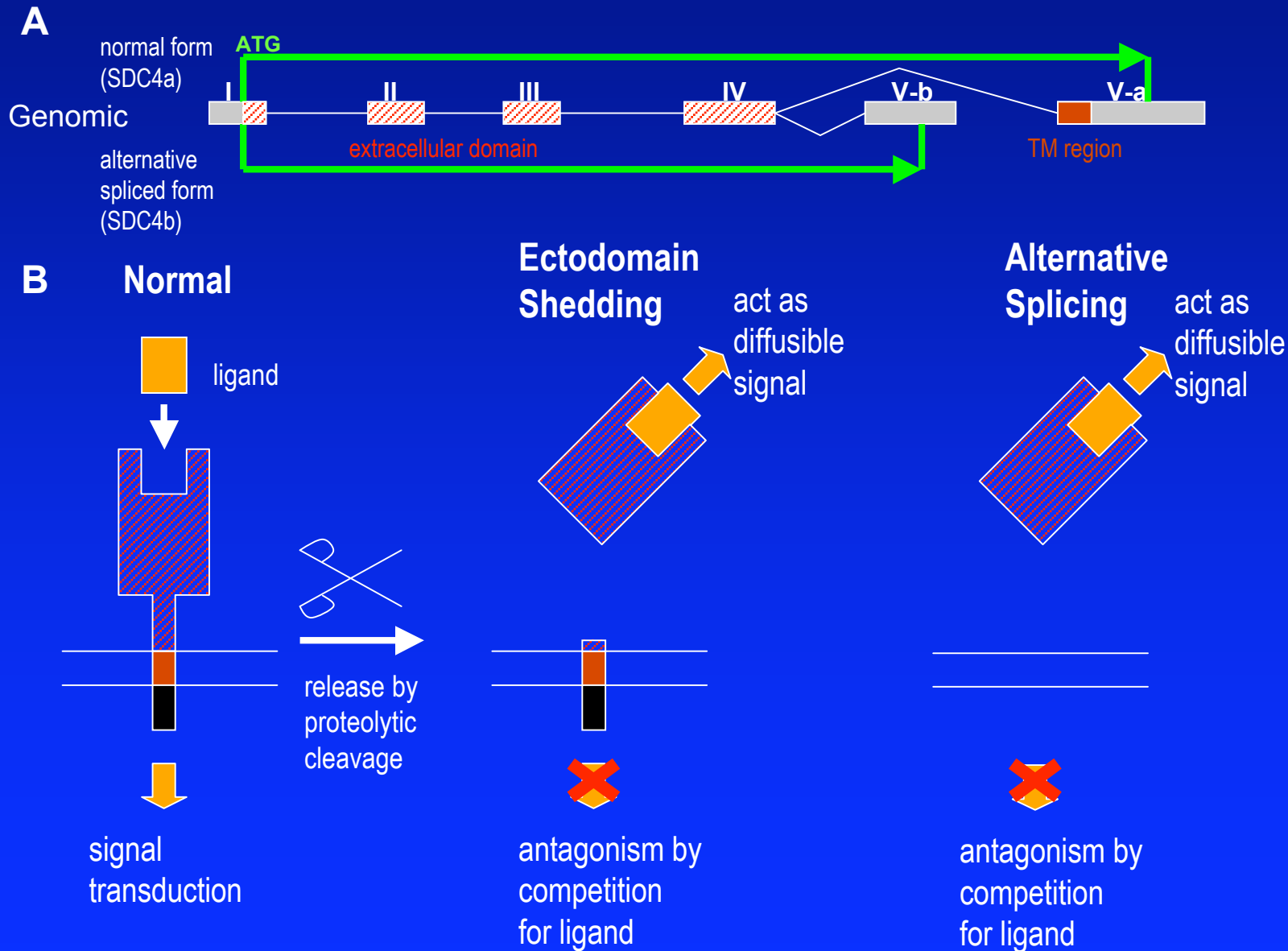
Xing, Resch and Lee, *Genome Research* (2004)

Frequent Removal of Transmembrane Domains by Alternative Splicing

- Production of soluble fragments of membrane proteins (e.g. via proteolysis) is an important regulatory mechanism for membrane proteins.
- Out of a total of 464 alternatively spliced genes encoding single-pass transmembrane (TM) proteins, in 188 (40%) we observed a splice form that specifically removed the TM domain
- TM alternative splicing showed a notable asymmetry in tissue specificity : membrane-anchored forms were more likely to be the ubiquitous forms , while the soluble forms were often localized to a single tissue (P-value <0.01 in 57 genes)
- Some genes are previously known to undergo proteolytic cleavage (e.g. ectodomain shedding) to release soluble fragments

Xing,Xu and Lee,*FEBS Letters* (2003)

Alternative splicing of Syndecan-4 parallels its ectodomain shedding event



Acknowledgement

- Christopher Lee
- Lee Lab, UCLA
- Molecular Biology PhD Program, UCLA

Figure 1

Alternative Splicing Annotation Project

ASAP

Search
By ID
By Tissue

Help
FAQ
Questions
Papers

Intro
Splicing

Glossary
Schema
Download

LeeLab Home

Gene View for Cluster Hs.2012 On Click Show: Sequence View

UniGene Cluster Hs.2012 (TCN1)

mRNA Isoform 11891

mRNA Isoform 11892

Sequence View for Exon 37143 On Click Show: Sequence View

4450	tctgccacag	AGGTAAGTGA	AGAAAACTAC	ATCCGCCTAA	AACCTCTGTT	GAATACAATG
4510	ATCCAGTCAA	ACTATAACAG	GGGAACCAGC	GCTGTCAATG	TTGTGTTGTC	CCTCAAACTT
4570	GTTGGAATCC	AGATCCAAAC	CCTGATGCAA	AAGATGATCC	AACAAATCAA	ATACAATGTG
4630	AAAAGCAGAT	gtaagttgct				

<http://www.bioinformatics.ucla.edu/ASAP/>