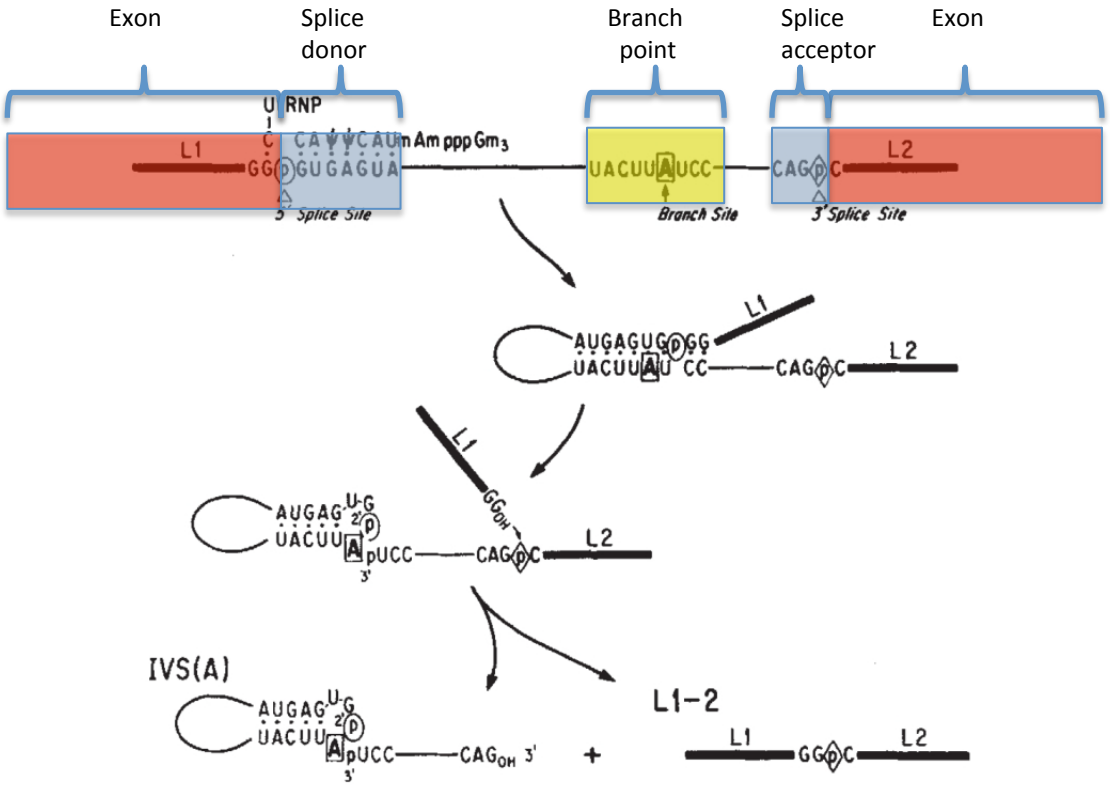
Loss of function annotation (LOF) Splice sites

Pablo Cingolani

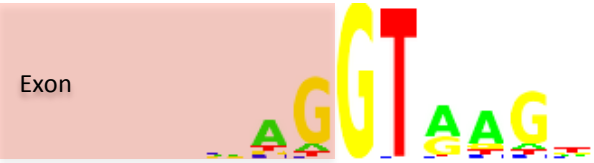
Mathieu Blanchette

Rob Sladek

Splicing


Exon
Splice donor
Branch point
Splice acceptor
Exon
U RNP
L1
GGpGUGAGUA
m Am ppp Gm3
UACUUAUCC
CAGpC
L2
5' Splice Site
Branch Site
3' Splice Site
AUGAGUGpGG
UACUUAU CC
L1
L2
AUGAG U-G
UACUU pUCC
CAGpC
L2
IVS(A)
AUGAG U-G
UACUU pUCC
CAGOH 3'
+
L1-2
L1
GGpC
L2

Donor & acceptor sites



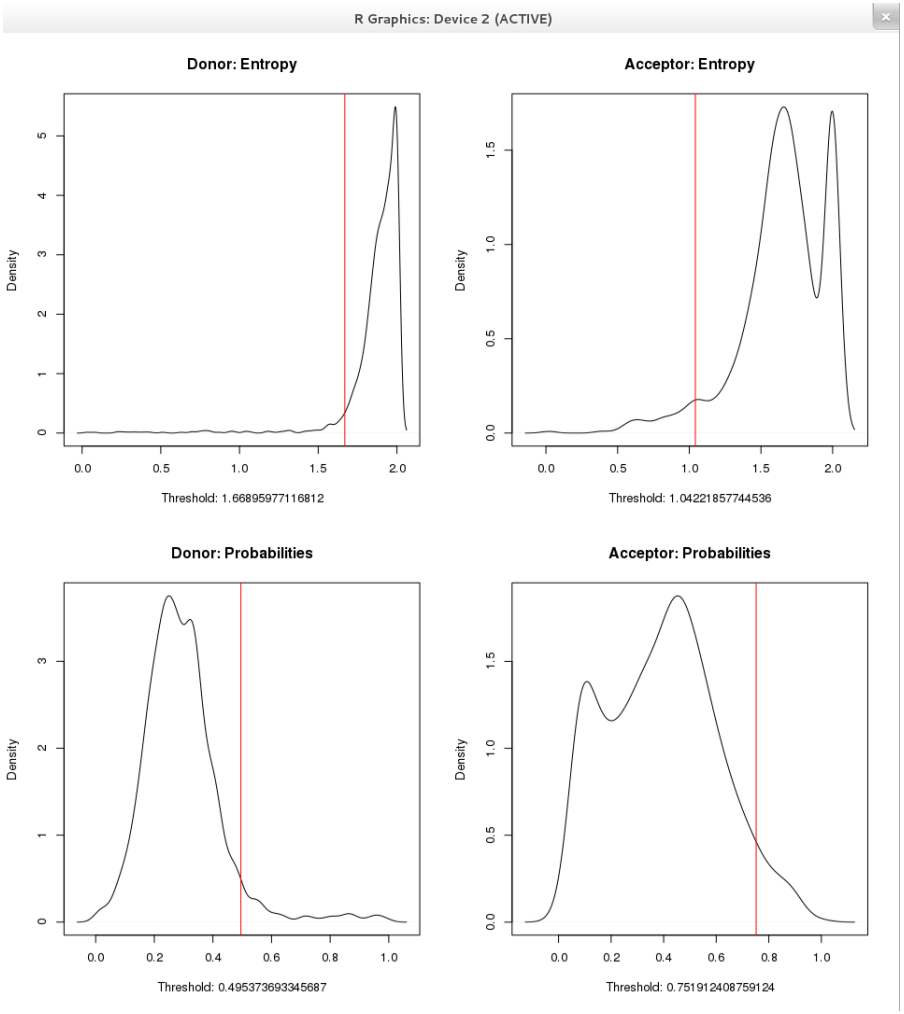
- “Essential splice sites” is only two bases in the intronic part (usually donor “GT”, acceptor “AG”)
- There is conservation beyond those two bases.
- There is conservation in the exonic part.
- There is no “gold standard” to predict branch point.

Analysis

Step 1

Goal: Find conserved donor & acceptor pairs

- Create quaternary trees counting occurrences of donor (or acceptor) sites.
- Find low entropy nodes in the trees (defined as bottom 0.05 quantile)
- For those nodes, find child nodes have high probability (i.e. top 0.95 quantile).
- Select those child nodes and create new quaternary trees only for those sites.
- Repeat the process selecting paired acceptor (or donor).
- Create a list of "significant donor-acceptor pairs"

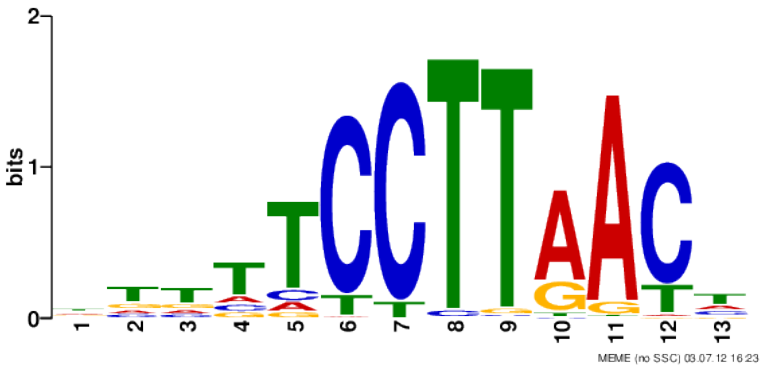


Analysis

Step 2

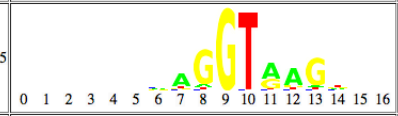
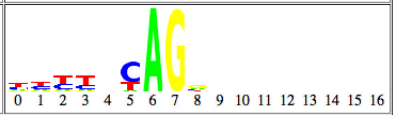
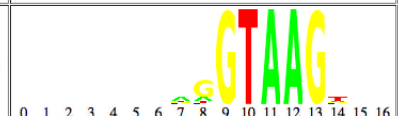
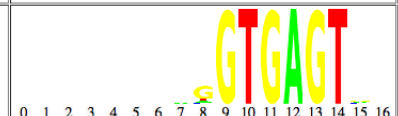
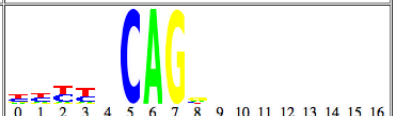
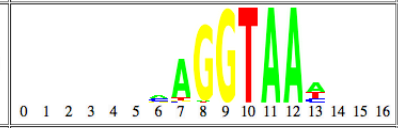
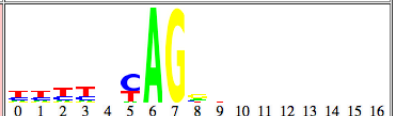
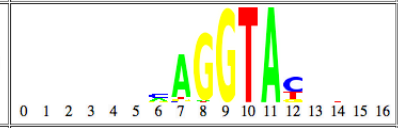
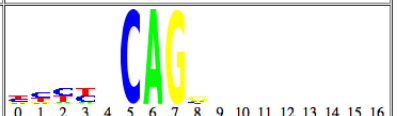
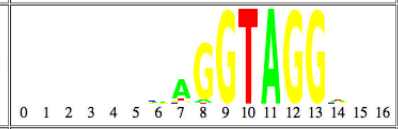
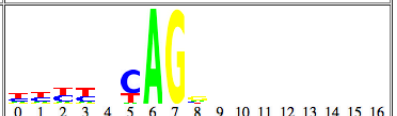
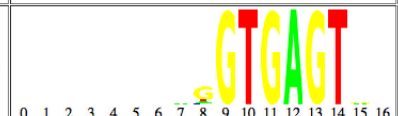
Goal: Find Branch points

- For all “significant donor-acceptor pairs”:
- Create sets of sequences 50 bases before the acceptor.
- Run motif finding using EM algorithm.
- Find overexpressed motifs
- Found U12 motif: Confirmation that the analysis was OK.



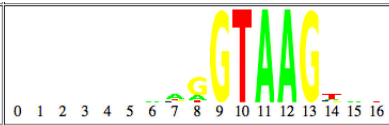
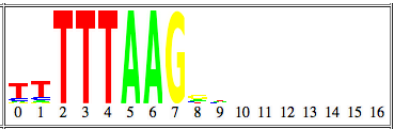
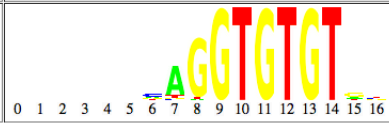
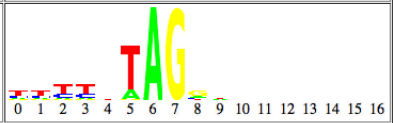
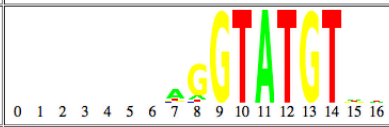
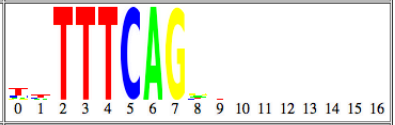
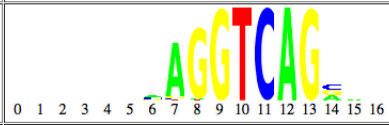
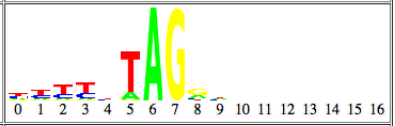
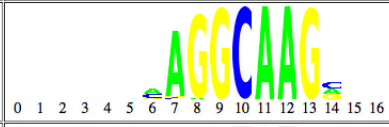
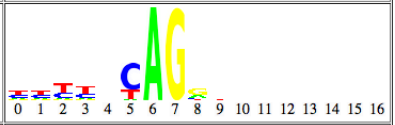
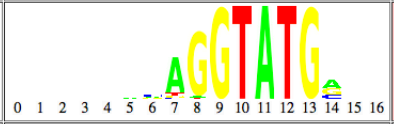
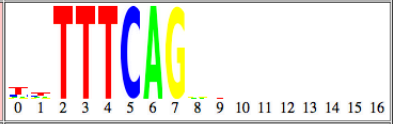
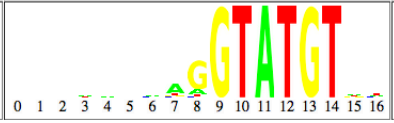
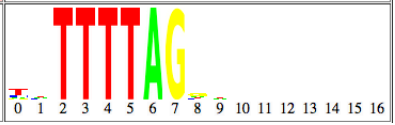
Results

i-) Some donor-acceptor pairs have higher degree of conservation in the donor and exon.

Donor type	Count	Donor Motif	U12 matches (Observed / Expected)	Acceptor Motif
ALL	346355		0 (0.00)	
GTAAG_AG	76416		4163 (1.09)	
GTGAGT_CAG	30901		1278 (0.83)	
GTAA_AG	29477		1808 (1.23)	
GTA_AG	28146		1625 (1.15)	
GTGAG_CAG	27933		989 (0.71)	
GTAGG_AG	17625		880 (1.00)	
GTGAGT_AG	13443		734 (1.09)	

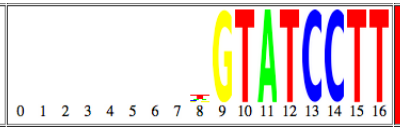
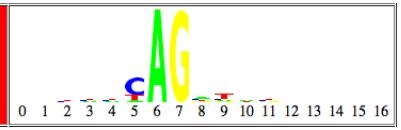
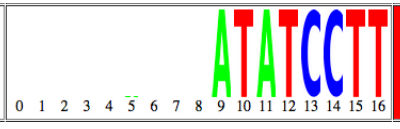
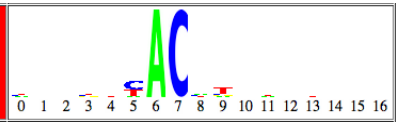
Results

i-) Some donor-acceptor pairs have higher degree of conservation in acceptor.

GTAAG_TTTAAG	842		38 (0.90)	
GTGTGT_AG	840		38 (0.90)	
GTATGT_TTTCAG	836		40 (0.96)	
GTCAG_AG	782		35 (0.90)	
GCAAG_AG	696		29 (0.83)	
GTATG_TTTCAG	674		41 (1.22)	
GTATGT_TTTTAG	621		28 (0.90)	

Results

i-) Branch point associated with donor-acceptor pairs: Over 100 times the expected values (red square).

GTATCCTT_AG	516		320 (12.40)	
ATATCCTT_AC	163		138 (16.93)	

Proposed changes: LOF

- Extend splice site donor & acceptor definitions to include highly conserved sequences (i.e. low entropy).
- Extend sites to exons whenever conservation is significant.
- Add U12 branch points predictions, but only in donor-acceptor subsets where enrichment is significant.
- Use PWMs scores (or similar scores) to “predict” LOF annotation of those sites.