Appearance of new splice sites on human lineage is associated with positive selection and splicing changes

Stepan Denisov

Institute for Information Transmission Problems (IITP RAS), Bolshoy Karetny per. 19, build.1, Moscow 127051 Russia, stepan@bioinf.fbb.msu.ru

Svetlana Iarovenko

Lomonosov Moscow State University, Leninskiye Gory, 1-73, Moscow 119234, Russia

svetyalana@gmail.com

Mikhail Gelfand

Institute for Information Transmission Problems (IITP RAS), Bolshoy Karetny per. 19, build.1, Moscow 127051 Russia, gelfand@iitp.ru

Eukaryotic genes are splitted into exons end introns. Exon-intron boundaries are marked by splice sites (SSs) – special sequences recognized by spliceosome (Fig. 1).

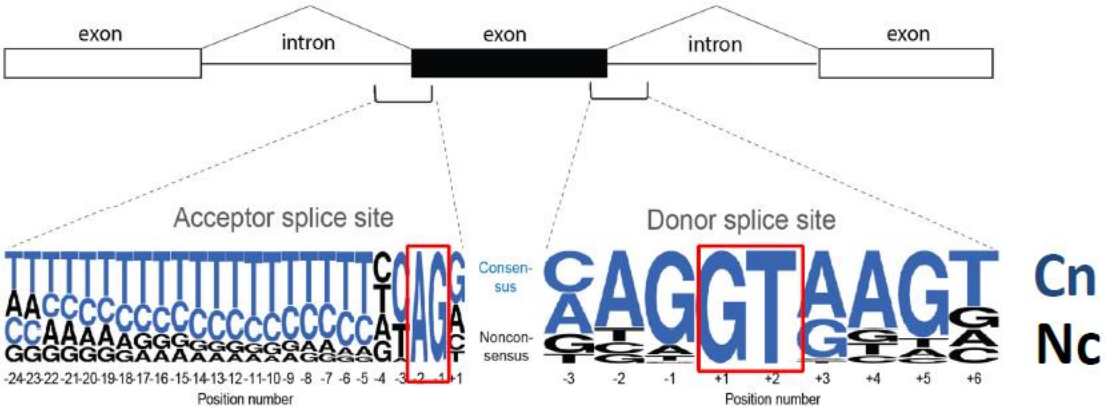

Fig. 1. Splice sites. Key dinucleotides (AG and GT) are shown in red rectangles. In each position nucleotides are subdivided into frequently used consensus nucleotides (Cn, shown in blue) and rarely used nonconsensus ones (Nc, shown in black).

It is well known that exon-intron structure of orthologous genes varies between different eukaryotic and even mammalian species [1]. Evolution of exon-intron structures and spicing

patterns is of great scientific interest. The most easy-to-research events like gain and loss of complete exons were described previously [2,3]. Exon loss occurs by several mechanisms including disabling of splice sites and genomic deletions [2]. Exons have been gained in evolution by acquisition of previously unused splice sites from intronic sequences [2] and by duplication of exons [4]. Changes in splicing enhancers and silencers could also "turn on" new exons or "turn off" previously used exons [3]. New exons originated from introns can be unique sequences or parts of repeats. The latter case was observed for primate-specific exons originated from Alu repeats. Alu repeats contain sequences similar to SSs, so only several mutations could convert them into real SSs and sequence between them into a new exon [5]. Patterns of intron gain and loss were studied in large phylogenetic distances [6] but underling mechanisms are still hypothetical [7]. Here we focused on appearance on new SSs from scratch within introns as well as within exons.

Both donor splice site (DSS, at 5' end of intron) and acceptor splice site (ASS, at 3' end of intron) contain key dinucleotide (GT in DSS and AG in ASS) which is surrounded by more variable context (Fig. 1). Key dinucleotides are necessary but not sufficient for splicing. We considered events of new splice site formation (where new key dinucleotide appears) on human lineage after divergence from rhesus macaque. We found 1500 of ASSs and 1244 DSSs created *de novo* on human lineage. About one fifth of these SSs lie within protein coding regions, the remaining ones are located within non-coding RNAs and UTRs. Formation of key dinucleotides is strongly associated with positive selection of wider SS context to increase adherence of SS to spliceosome. In particular, positive selection for substitutions from nonconsensus to consensus nucleotides is 2 – 8 times stronger in new DSSs than in old ones (compare Fig 2A and 2B). Appearance of new SSs changes splicing pattern of corresponding gene. We classified observed changes in splicing and found that new splice sites frequently create new exons from intronic sequences, extend or shorten existing exons (exonization and intronization, respectively), and create new introns within exons (Fig. 3). New splice sites also change first exon of the gene. In cases of new exon creation and extension of the existing exons newly acquired coding sequences tend to be multiple of three by length and seem to be under positive selection on amino acid sequence.

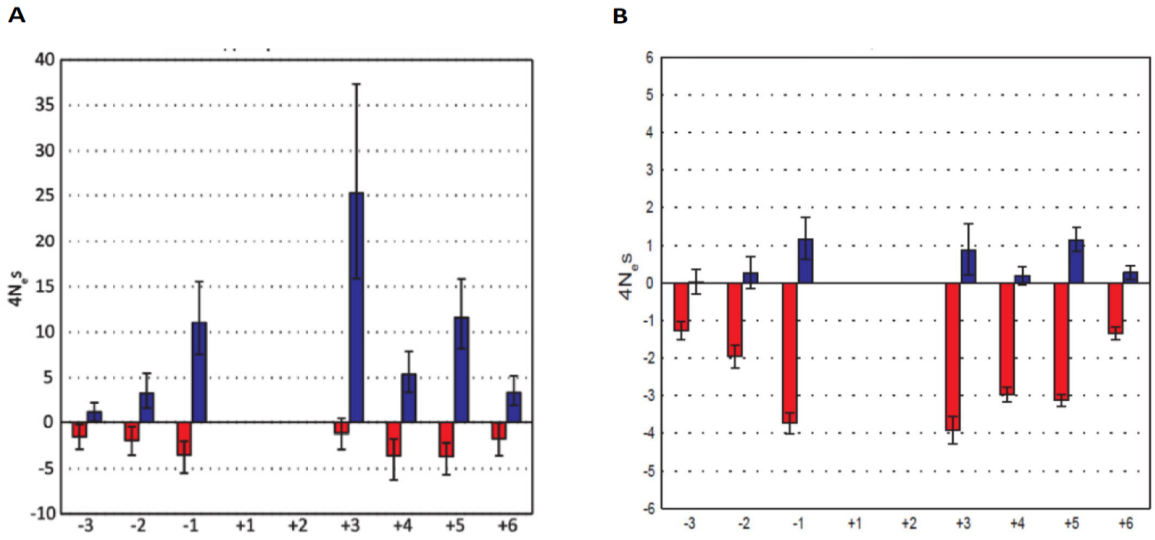

Fig. 2. Strength of selection in new (A) and old (B) splice sites. Strength of selection measured in $4N_es$ units (vertical axis) acting on Nc-to-Cn substitutions (blue bars), and Cn-to-Nc substitutions (red bars). Positive and negative values of $4N_es$ correspond to positive and negative selection, respectively. Horizontal axis, position within the SS corresponding to Fig.1

Both donor splice site (DSS, at 5' end of intron) and acceptor splice site (ASS, at 3' end of intron) contain key dinucleotide (GT in DSS and AG in ASS) which is surrounded by more variable context (Fig. 1). Key dinucleotides are necessary but not sufficient for splicing. We considered events of new splice site formation (where new key dinucleotide appears) on human lineage after divergence from rhesus macaque. We found 1500 of ASSs and 1244 DSSs created *de novo* on human lineage. About one fifth of these SSs lie within protein coding regions, the remaining ones are located within non-coding RNAs and UTRs. Formation of key dinucleotides is strongly associated with positive selection of wider SS context to increase adherence of SS to spliceosome. In particular, positive selection for substitutions from nonconsensus to consensus nucleotides is 2 – 8 times stronger in new DSSs than in old ones (compare Fig 2A and 2B). Appearance of young SSs changes splicing pattern of corresponding gene. We classified observed changes in splicing and found that young splice sites frequently create new exons from intronic sequences, extend or shorten existing exons (exonization and intronization, respectively), and create new introns within

exons (Fig. 3). Young splice sites also change first exon of the gene. In cases of new exon creation and extension of the existing exons newly acquired coding sequences tend to be multiple of three by length and seem to be under positive selection on amino acid sequence. Most of events not overlap any known repetitive sequences (according to Repeat Masker).

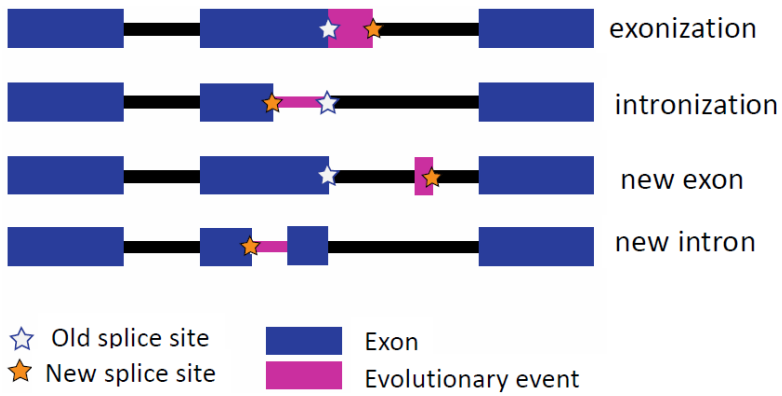

Fig. 3. Classification of evolutionary events.

1. I.I. Artamonova, M.S. Gelfand (2007) Comparative genomics and evolution of alternative splicing: the pessimists′ science, *ChemInform*, **38(46):** 3407-3430.

2. A.V. Alekseyenko et al. (2007) Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes, *RNA*, **13:** 661–670.

3. J.J. Merkin et al. (2015) Origins and impacts of new mammalian exons, *RNA*, **10:** 1992–2005.

4. F.A. Kondrashov, E.V. Koonin (2001) Origin of alternative splicing by tandem exon duplication, *Hum Mol Genet*, **10:** 2661–2669.

5. G. Lev-Maor et al. (2003) The birth of an alternatively spliced exon: 3’ splice-site selection in Alu exons, *RNA*, **13:** 661–670.

6. M. Csuros et al. (2011) A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes, *PLOS Comput Biol*, **7:** e1002150.

7. S.W. Roy, W. Gilbert (2006) The evolution of spliceosomal introns: patterns, puzzles and progress, *Nat Rev Genet*, **7:** 211–221.