

Genome analysis

Assessing the application of Ka/Ks ratio test to alternatively spliced exons

Yi Xing and Christopher Lee*

Department of Chemistry and Biochemistry, Molecular Biology Institute, Center for Genomics and Proteomics, University of California, Los Angeles, Los Angeles, CA 90095-1570, USA

Received on July 5, 2005; revised and accepted on August 3, 2005

Advance Access publication August 9, 2005

ABSTRACT

Summary: Recently, the Ka/Ks ratio test, which assesses the protein-coding potentials of genomic regions based on their non-synonymous to synonymous divergence rates, has been proposed and successfully used in genome annotations of eukaryotes. We systematically performed the Ka/Ks ratio test on 925 transcript-confirmed alternatively spliced exons in the human genome, which we describe in this manuscript. We found that 22.3% of evolutionarily conserved alternatively spliced exons cannot pass the Ka/Ks ratio test, compared with 9.8% for constitutive exons. The false negative rate was the highest (85.7%) for exons with low frequencies of transcript inclusion. Analyses of alternatively spliced exons supported by full-length mRNA sequences yielded similar results, and nearly half of exons involved in ancestral alternative splicing events could not pass this test. Our analysis suggests a future direction to incorporate comparative genomics-based alternative splicing predictions with the Ka/Ks ratio test in higher eukaryotes with extensive RNA alternative splicing.

Contact: leec@mbi.ucla.edu**1 INTRODUCTION**

Comparative genomics has provided powerful tools for annotations of eukaryotic genomes (Kellis *et al.*, 2003). In a pioneering study, Nekrutenko *et al.* (2002) proposed the 'Ka/Ks ratio test' to assess the protein-coding potentials of predicted exons. This test is based on the assumption that the majority of protein-coding regions in the human genome are under strong purifying selection during evolution. As a result their rates of synonymous divergence (Ks) greatly exceed the rates of non-synonymous divergence (Ka), yielding Ka/Ks ratios of much less than one in human–mouse orthologous sequence comparisons. On a sample of 1244 exons from 153 protein-coding genes, the Ka/Ks ratio test gave an 8% false negative rate and a <5% false positive rate for internal exons, an accuracy which was better than most of the gene prediction tools (Nekrutenko *et al.*, 2002). Since its introduction, the Ka/Ks ratio test has been widely and successfully used for improving the annotations of human and other mammalian genomes (Miller *et al.*, 2004; Nekrutenko, 2004; Nekrutenko *et al.*, 2003b; Zhang and Gerstein, 2004).

One emerging question about the Ka/Ks ratio test relates to alternatively spliced exons in the eukaryotic genomes. Recent studies of

expressed sequences and microarray data have shown that alternative splicing is a widespread mechanism of gene regulation in higher eukaryotes (Lareau *et al.*, 2004; Modrek and Lee, 2002). Up to three quarters of human coding genes undergo alternative splicing (Johnson *et al.*, 2003). There is abundant evidence to suggest that alternative splicing is associated with relaxations of selection pressure during evolution (Boue *et al.*, 2003). For example, alternative splicing is observed to be associated with an accelerated rate of exon creation and loss (Modrek and Lee, 2003), new exon originations from *Alu* elements (Sorek *et al.*, 2002), tolerance of premature termination codons (Lewis *et al.*, 2003; Xing and Lee, 2004), and so on. Iida and Akashi (2000) investigated the sequence divergence patterns of 110 alternatively spliced protein-coding genes from human and *Drosophila*, and found that alternatively spliced regions of these genes had higher Ka/Ks values compared with constitutive regions. Other examples of elevated Ka/Ks in alternatively spliced exons have also been reported (Filip and Mundy, 2004; Hurst and Pal, 2001). These observations raise a question regarding the divergence from the Ka/Ks ratio test among alternatively spliced exons.

2 METHODS

We identified alternatively spliced exons by aligning human expressed sequences [mRNA/expressed sequence tag (EST)] to the human genome (Modrek *et al.*, 2001). To quantify the degree of alternative splicing for each alternatively spliced exon, we used a standard metric of alternative splicing—the exon inclusion level, defined as the number of ESTs that included an exon divided by the total number of ESTs that either included or skipped this exon. We subdivided alternatively spliced exons into three classes based on their inclusion levels: major-form (>2/3), medium-form (between 1/3 and 2/3) and minor-form (<1/3).

We identified the orthologous exon sequence for each human exon in the genomic sequence of the mouse ortholog, as previously described (Modrek and Lee, 2003). For each human–mouse orthologous exon sequence pair, we performed the Ka/Ks ratio test following the protocol of Nekrutenko *et al.* (2003a). Briefly, orthologous exon sequences from human and mouse were translated and then aligned using CLUSTALW (Thompson *et al.*, 1994) under default parameters. This protein alignment was used to seed an alignment of corresponding nucleotide sequences, and gaps in the alignment were trimmed. We estimated the number of synonymous and non-synonymous substitutions/sites using the Yang–Nielsen estimates from the yn00 program of the PAML package (PAML 3.14) (Yang, 1997). We built a 2 × 2 contingency table using the numbers of changed and unchanged synonymous/non-synonymous sites, and tested whether the Ka/Ks ratio was significantly <1 using the Fisher's exact test. We defined an exon as passing the Ka/Ks ratio test if its Ka/Ks was significantly <1 at the $P < 0.05$ level.

*To whom correspondence should be addressed.

Table 1. Exons that pass or fail the Ka/Ks ratio test

Types of exons	Total #	Average length (bp)	# Fail	# Pass	% Fail	Mean (median) Ka/Ks
Constitutive	10 996	136	1077	9919	9.8	0.146 (0.070)
Alternative	925	122	206	719	22.3	0.199 (0.094)
Alt (Major-form)	630	121	101	529	16.0	0.162 (0.086)
Alt (Medium-form)	253	129	69	184	27.3	0.235 (0.113)
Alt (Minor-form)	42	81	36	6	85.7	0.649 (0.410)
Ancestral Alt	120	102	59	61	49.2	0.412 (0.182)
Alt (CpG/GpC < 0.8)	793	124	165	628	20.8	0.192 (0.097)

3 RESULTS AND DISCUSSION

We compiled a list of 925 human alternatively spliced exons that were conserved between human and mouse genomes, based on analyses of human expressed sequences (Modrek *et al.*, 2001). We also compiled a list of 10 996 human constitutive exons as a control. All these exons were internal exons flanked by introns at both ends. We performed the Ka/Ks ratio tests on these exons following the protocol of Nekrutenko *et al.* (2003a) (see Methods section). Of the constitutive exons 9.8% failed to pass the Ka/Ks ratio test, a ratio similar to what was reported by the initial study (8%) (Nekrutenko *et al.*, 2002) (Table 1). In contrast, 22.3% of alternatively spliced exons being tested could not pass the Ka/Ks ratio test, a more than 2-fold increase compared with constitutive exons. Because alternatively spliced exons with different exon inclusion levels (see definitions in the Methods section) exhibited different patterns of evolutionary divergence (Modrek and Lee, 2003; Pan *et al.*, 2004), we divided the 925 alternatively spliced exons into three classes based on their exon inclusion levels (see Methods section). The fraction of exons failing the test was 16.0% for major-form exons, and increased to 85.7% for minor-form alternative exons (included <1/3 in the transcripts). Since alternatively spliced exons were shorter on an average [Table 1; also see (Sorek *et al.*, 2004b)], we also subdivided exons based on their sizes (Fig. 1). In both constitutive and alternatively spliced exons, the fractions failing the test were higher for shorter exons, consistent with the original study (Nekrutenko *et al.*, 2002). However, the fraction was consistently higher in alternatively spliced exons after controlling for exon sizes (e.g. 5.4% for constitutive exons and 16.6% for alternatively spliced exons between 101 and 150 nt; Fig. 1). Analyses of mouse alternatively spliced exons in a mouse–human comparison produced similar results (data not shown).

Our result indicates that a significantly higher fraction of alternatively spliced exons in the human genome cannot pass the Ka/Ks ratio test. However, this does not immediately translate into an increased false negative rate of the Ka/Ks ratio test in alternatively spliced exons, since other interpretations are possible. Do these data actually imply that a considerable number of alternatively spliced exons observed in the human EST sequences do not represent real exons, but indeed come from artifacts in the EST data (e.g. rare spliceosomal errors) (Modrek and Lee, 2002; Sorek and Safer, 2003)? This explanation seems particularly plausible for minor-form exons (which are observed in a small fraction of EST sequences). To test this possibility, we analyzed a subset of alternatively spliced exons that were supported by full-length mRNA sequences. We observed similar fractions of alternatively spliced exons failing the Ka/Ks ratio test

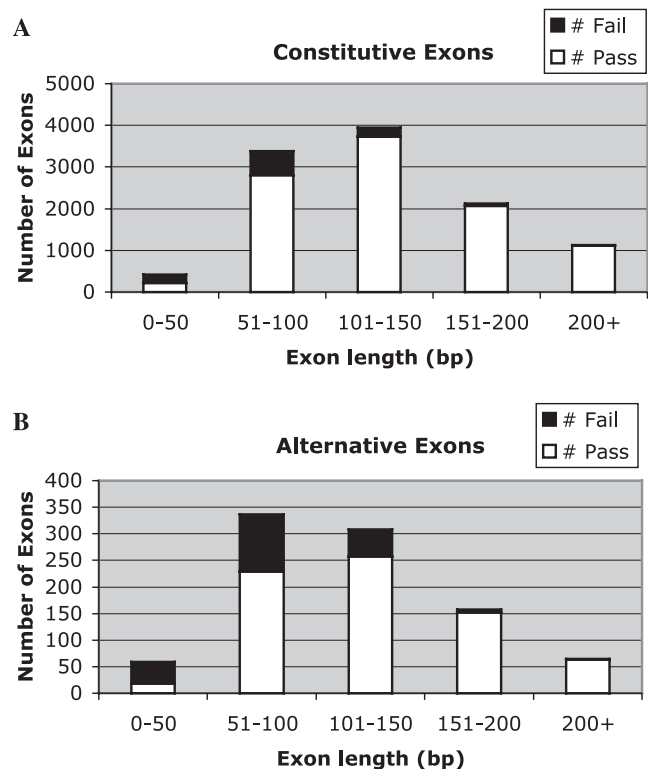


Fig. 1. The Ka/Ks ratio test on constitutive and alternatively spliced exons with different sizes. (A) Constitutive exons and (B) alternatively spliced exons.

(Table 2). Therefore, spurious exons originated from EST artifacts cannot explain our data. Do alternatively spliced exons failing the Ka/Ks ratio test largely represent non-functional splice forms? To answer this question, we restricted our analysis to a set of 120 exons that were alternatively spliced in both human and mouse transcriptomes. Such a pattern of ‘ancestral alternative splicing’ was widely adopted as a criterion for functional alternative splicing events (Resch *et al.*, 2004; Sorek *et al.*, 2004a). In these exons an even higher 49.2% (versus 22.3% of all alternatively spliced exons) could not pass the Ka/Ks ratio test (Table 1), consistent with another recent study on such exons (Ohler *et al.*, 2005). Therefore the hypothesis for non-functional splice forms cannot explain our data either. Finally, to

Table 2. Exons supported by human mRNAs that pass or fail the Ka/Ks ratio test

Types of exons	Total #	# Fail	# Pass	% Fail
Constitutive	10 524	1017	9507	9.7
Alternative	811	158	653	19.5
Alt (Major-form)	618	99	519	16.0
Alt (Medium-form)	170	39	131	22.9
Alt (Minor-form)	23	20	3	87.0

rule out the potential influence of CpG islands, we calculated the frequency of CpG over GpC in each exon, and restricted our analysis to a subset of alternatively spliced exons whose CpG/GpC ratios were <0.8 (Iida and Akashi, 2000). Of these exons 20.8% could not pass this test, similar to the percentage for the total set of alternatively spliced exons (Table 1). Although in principle an increased Ka/Ks ratio might reflect various underlying mechanisms [e.g. relaxations of protein-level selection pressure; translational selections (Iida and Akashi, 2000) and selections at silent sites (Hurst and Pal, 2001)], which is not the focus of this manuscript, our control analyses do indicate that a large fraction of functional alternative exons in the human genome fail the Ka/Ks ratio test.

Constitutive exons outnumber alternatively spliced exons in most protein-coding genes. Since in many organisms transcript sequence coverage (e.g. ESTs) is still quite low, the Ka/Ks ratio test is a powerful tool for refining computational gene structure predictions. However, the majority of mammalian protein-coding genes are alternatively spliced, and a small number of alternatively spliced exons might have profound functional and regulatory impacts, as recently illustrated by the alternative splicing of the C2A domain of Piccolo (Garcia *et al.*, 2004) and many others. Our analysis suggests that in organisms with extensive alternative splicing (e.g. mammals) it is preferable to combine the Ka/Ks ratio test with other metrics that indicate the probability of alternative splicing. Fortunately evolutionary genomics has also shed light on the typical traits of functional alternatively spliced exons [such as an increased conservation at their flanking introns (Sorek and Ast, 2003) and an increased preference for protein reading frame preservation (Resch *et al.*, 2004; Sorek *et al.*, 2004a)], which have been successfully used in predictions (Philipps *et al.*, 2004; Sorek *et al.*, 2004b; Yeo *et al.*, 2005). Such information can be integrated with the Ka/Ks ratio test for a more accurate assessment of protein-coding potentials of genomic regions.

ACKNOWLEDGEMENTS

The authors thank Anton Nekrutenko for reading of our manuscript and for the helpful comments. This work was supported by NIH Grant U54-RR021813, a Teacher-Scholar award to C.J.L. from the Dreyfus Foundation, a DOE grant DE-FC02-02ER63421. Y.X. is supported by a Ph.D. dissertation fellowship from UCLA.

Conflict of Interest: none declared.

REFERENCES

- Boue, S. *et al.* (2003) Alternative splicing and evolution. *Bioessays*, **25**, 1031–1034.
- Filip, L.C. and Mundy, N.I. (2004) Rapid evolution by positive Darwinian selection in the extracellular domain of the abundant lymphocyte protein CD45 in primates. *Mol. Biol. Evol.*, **21**, 1504–1511.
- Garcia, J. *et al.* (2004) A conformational switch in the Piccolo C2A domain regulated by alternative splicing. *Nat. Struct. Mol. Biol.*, **11**, 45–53.
- Hurst, L.D. and Pal, C. (2001) Evidence for purifying selection acting on silent sites in BRCA1. *Trends Genet.*, **17**, 62–65.
- Iida, K. and Akashi, H. (2000) A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene*, **261**, 93–105.
- Johnson, J.M. *et al.* (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
- Kellis, M. *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
- Lareau, L.F. *et al.* (2004) The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol.*, **14**, 273–282.
- Lewis, B.P. *et al.* (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA*, **100**, 189–192.
- Miller, W. *et al.* (2004) Comparative genomics. *Annu. Rev. Genomics Hum. Genet.*, **5**, 15–56.
- Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.
- Modrek, B. and Lee, C. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased rate of exon creation/loss. *Nat. Genet.*, **34**, 177–180.
- Modrek, B. *et al.* (2001) Genome-wide analysis of alternative splicing using human expressed sequence data. *Nucleic Acids Res.*, **29**, 2850–2859.
- Nekrutenko, A. (2004) Reconciling the numbers: ESTs versus protein-coding genes. *Mol. Biol. Evol.*, **21**, 1278–1282.
- Nekrutenko, A. *et al.* (2002) The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.*, **12**, 198–202.
- Nekrutenko, A. *et al.* (2003a) ETOPE: evolutionary test of predicted exons. *Nucleic Acids Res.*, **31**, 3564–3567.
- Nekrutenko, A. *et al.* (2003b) An evolutionary approach reveals a high protein-coding capacity of the human genome. *Trends Genet.*, **19**, 306–310.
- Ohler, U. *et al.* (2005) Recognition of unknown conserved alternatively spliced exons. *PLoS Comp. Biol.*, **1**, e15.
- Pan, Q. *et al.* (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell.*, **16**, 929–941.
- Philipps, D.L. *et al.* (2004) A computational and experimental approach toward a priori identification of alternatively spliced exons. *RNA*, **10**, 1838–1844.
- Resch, A. *et al.* (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.*, **32**, 1261–1269.
- Sorek, R. and Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, **13**, 1631–1637.
- Sorek, R. and Safer, H.M. (2003) A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.*, **31**, 1067–1074.
- Sorek, R. *et al.* (2002) Alu-containing exons are alternatively spliced. *Genome Res.*, **12**, 1060–1067.
- Sorek, R. *et al.* (2004a) How prevalent is functional alternative splicing in the human genome? *Trends Genet.*, **20**, 68–71.
- Sorek, R. *et al.* (2004b) A non-EST-based method for exon-skipping prediction. *Genome Res.*, **14**, 1617–1623.
- Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
- Xing, Y. and Lee, C. (2004) Negative selection pressure against premature protein truncation is reduced by both alternative splicing and diploidy. *Trends Genet.*, **20**, 472–475.
- Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
- Yeo, G.W. *et al.* (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl Acad. Sci. USA*, **102**, 2850–2855.
- Zhang, Z. and Gerstein, M. (2004) Large-scale analysis of pseudogenes in the human genome. *Curr. Opin. Genet. Dev.*, **14**, 328–335.