J Mol Evol (2007) 64:646–655
DOI: 10.1007/s00239-006-0207-7

JOURNAL OF MOLECULAR EVOLUTION

© Springer Science+Business Media, LLC 2007

How Common Are Intragene Windows with $K_A > K_S$ Owing to Purifying Selection on Synonymous Mutations?

Joanna L. Parmley, Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

Received: 11 September 2006 / Accepted: 7 March 2007 [*Reviewing Editor:* Dr. Manyuan Long]

Abstract. One method for diagnosing the mode of sequence evolution considers the ratio of nonsynonymous substitutions per nonsynonymous site (K_A) to the corresponding figure for synonymous substitutions (K_S). A ratio (K_A/K_S) greater than unity is taken as evidence for positive selection. This, however, need not necessarily be the case. Notably, there is one instance of a high intragenic K_A/K_S peak, revealed by sliding window analysis and observed in two pairwise comparisons, better accounted for by localised purifying selection on synonymous mutations that affect splicing. Is this example exceptional? To address this we isolate intragenic domains with $K_A/K_S > 1$ from more than 1000 long mouse-rat orthologues. Approximately one $K_A/K_S > 1$ peak is found per 12–15 kb of coding sequence. Surprisingly, low synonymous substitution rates underpin more incidences than do high nonsynonymous rates. Several reasons, however, prevent us from supposing that the low synonymous rates reflect purifying selection on synonymous mutations. First, for many peaks, the null that the peak is no higher than expected given the underlying rates of evolution, cannot be rejected. Second, of 18 statistically significant incidences with unusually low K_S values, only 3 are repeatable across independent comparisons. At least two of these are within alternatively spliced exons. We conclude that repeatable statistically significant intragenic domains of low intragenic K_S are rare. As so few K_A/K_S peaks reflect increased rates of protein evolution and so few hold statistical support, we additionally conclude that sliding window analysis to infer domains of positive selection is highly error-prone.

Key words: K_A/K_S ratio — Sliding Window analysis — Selection on synonymous mutations — Alternative transcripts

Correspondence to: Laurence D. Hurst; *email:* bssldh@bath.ac.uk

Introduction

With the recent proliferation of sequence data there is much interest in determining those genes on which positive selection has acted (see, e.g., Clark et al. 2003). Similarly, it is desirable to know where in the sequence selection might have acted and relate that to the biology of the gene/protein in question. The common method for diagnosing the mode of sequence evolution considers the ratio of nonsynonymous substitutions per nonsynonymous site (K_A), to the corresponding figure for synonymous substitutions (K_S). A ratio (K_A/K_S) greater than unity is taken as evidence for positive selection, promoting change at the protein level. This interpretation need not, however, be correct. In principle, if purifying selection is stronger on synonymous mutations than is purifying selection on the protein, $K_A/K_S > 1$ would also be found.

The latter possibility is typically considered to be so bizarre as to be effectively worth ignoring. In one case, however, *BRCA1*, a very high K_A/K_S intragenic peak, found using sliding window analysis, was

observed in both human-dog and mouse-rat orthologous gene alignments at the same location (Hurst and Pal 2001). The peak was associated not with an increased rate of protein evolution in the critical domain but rather with a strikingly reduced rate of synonymous evolution. This was interpreted as being consistent with purifying selection on synonymous mutations. Indeed, while K_A/K_S ratios greater than unity can be owing to forces other than positive selection—they can, for example, be recovered in simulations of background selection when assuming that synonymous mutations are neutral (Palsson 2004)—we are aware of no alternative interpretation, save for sampling artifact, for domains of low K_S but moderate K_A. While Hurst and Pal (2001) were unable to identify a possible cause, subsequently it was noted that this critical region was one containing splice enhancer domains associated with alternative splicing (Orban and Olah 200l). It was thus suggested that the K_A/K_S peak was owing to highly regionalized selection against synonymous mutations that affect alternative splicing. Since then, much evidence for selection against mutations that affect splicing and splice enhancer domains has emerged (Carlini and Genut 2006; Cartegni et al. 2002; Chamary et al. 2006; Chamary and Hurst 2005a; Chen et al. 2006; Ermakova et al. 2006; Fairbrother et al. 2004a; Parmley et al. 2006; Plass and Eyras 2006; Willie and Majewski 2004; Xing and Lee 2005a, 2005b, 2006a, 2006b). Other forms of selection on synonymous mutations are, however, possible. For example, selection for mRNA folding (Chamary and Hurst 2005b; Duan et al. 2003; Nackley et al. 2006), micro-RNA/mRNA binding (Hurst 2006), and translational pausing by use of rare codons (Kimchi-Sarfaty et al. 2007), all have some empirical support and could potentially be regionalized within genes. More generally, large spans of intragenic domains associated with low synonymous rates of divergence have been found (Schattner and Diekhans 2006).

The above case history of *BRCA1* tempts an obvious question: if one were to repeat the same form of sliding window analysis on very many genes, how commonly would one find $K_A/K_S > 1$ intragenic peaks best explained by localized selection on synonymous mutations? The problem, however, centers on what one means by "best explained", which in the case of sliding window analysis poses multiple difficulties. First it is necessary to isolate $K_A/K_S > 1$ peaks and attempt to classify them according to the pattern of rate variation around the region and in the gene more generally: are they regional K_S dips showing no increase in K_A, K_A peaks associated also with higher than average K_S, or might they be undeterminable?

Having identified possible K_S dips we cannot, however, be confident that we are witnessing selection on synonymous mutations. Most of the problems stem from the fact that sliding window analysis has no formal statistical basis and is difficult to defend rigorously. Most notably, as the method requires multiple windows to be examined, the probability of spurious peaks is acute, made more so by nonindependence between overlapping windows. Note too that the requirement for many windows and lower limit on window size for accurate estimation of K_A and K_S mean that the method is also applicable only to long genes. To control for the multiple testing and nonindependence problems we assemble random genes by shuffling the codons in each alignment and determine how often a given peak is expected to be observed given the underlying rates of evolution, applying the same sliding window protocol to all the randomised versions. Even after making allowance for such error there remains the possibility that any significant peak is still just spurious, especially as multiple genes are being tested. To be more confident that the low K_S is associated with selection against synonymous mutations, it is best if, as in the case of *BRCA1*, some evidence for the same pattern is observed in an independent comparison.

Methods

Genes and Alignment

A file of 12634 orthologous mouse/rat genes was obtained from the Mouse Genome Informatics Web site (ftp://ftp.informatics.jax.org/pub/reports/index.html#orthology). The EntrezGene IDs were used to search the NCBI database for corresponding mouse and rat RefSeqs. Orthologous genes were discarded if the gene was only classed as predicted in either species. The mouse gene sequence and exon position data were obtained from Ensembl, whereas the rat sequence was obtained from NCBI. The orthologues were aligned using MUSCLE (Edgar 2004). Orthologues less than 1500 bp in length were discarded to allow for a sliding window analysis. Any alignments with indels of either high frequency (>5 indels per 1 kbp) or long length (>30 bp) were also discarded as potential poorly aligned sequence. For a list of the 1074 remaining genes and accession numbers of genes used in the study, see supplementary data 1.

Sliding Window

The synonymous and nonsynonymous rates of substitution were calculated by the Li method (Li 1993; Pamilo and Bianchi 1993) for windows of varying sizes moving three codons along the sequence for each "slide". In our application of the sliding window, if either extracted sequence contained an indel, then another codon was included in the window until an effective codon window size was achieved. The substitution rates and hence the K_A/K_S ratio for each window were calculated and those that were >1 had their cause assessed to be due to either a peak in local K_A or a dip in K_S. A ratio peak is deemed to be a peak in K_A under two conditions: first, if both K_A and K_S are higher than that of the gene average. This we refer to as the strict definition of a K_A peak. However, this strict definition, although rigorously defendable, will miss cases where K_A is much higher than the average but K_S is just below the average. To attempt to ensure that what is true for the strict set might be more

generally true, it is desirable to define a more generous definition. This we do by defining "much higher" and "little lower" by the deviation of the observed values from the genic mean in terms of the number of standard deviations (i.e., Z scores). The standard deviation in K_A and K_S was determined from the observed windows. If at the $K_A > K_S$ peak, the ratio of the Z score for K_A to the Z score for K_S is >2, we consider this to be a "generous" K_A peak. Likewise, the K_A/K_S peak is deemed due to a dip in K_S under two conditions: if both K_A and K_S are below the gene average (a strict K_S dip) or if K_A is a little higher than the mean and K_S is much lower than the mean, i.e., the ratio of the Z score for K_A to the Z score for K_S is <0.5 (a generous K_S dip). Any other permutations are considered to be a combination of factors and therefore have an undefined cause. The figure for defining a generous peak (2 or 0.5 ratio of Z scores) is an arbitrary choice but by visual inspection (Supplementary Fig. 4) appears to capture the appropriate forms of peak. It does, for example, capture the previously discussed K_S dip in *BRCA1*. Restricting analysis to only those peaks classified under the strict definition does not qualitatively affect conclusions (Supplementary Fig. 1).

Obtaining Exonic Splicing Enhancers

Exonic splicing enhancer sequences, for human and mouse, were downloaded from http://www.genes.mit.edu/burgelab/rescue-ese. These candidate exonic splicing enhancer (ESE) sequences were previously identified by the Burge group, by assaying whether oligonucleotide motifs exhibit splicing activity in vivo. The 238 human (Fairbrother et al. 2002) and 380 mouse (Yeo et al. 2004) ESE hexamers were determined using Relative Enhancer and Silencer Classification by Unanimous Enrichment (RESCUE), a computational approach followed by experimental validation. Briefly, the method identifies motifs that are (1) significantly enriched in exons relative to introns and (2) significantly more frequent in exons with strong consensus splice sites than in exons with strong consensus splice sites (Fairbrother et al. 2004b). Motifs that match these criteria are then grouped into clusters, after which representatives from each cluster are tested for ESE activity in vivo using a splicing reporter system.

SNP Analysis

The locations within the gene of synonymous SNPs in mouse were obtained by screening all the genes in our full long gene data set at dbSNP (Sherry et al. 2001) using each gene's unique unigene identification. The total number of SNPs in the full data set, along with the full length of all sequences, was then employed to define the expected SNP count within and outside the K_S dip windows.

Results

At $K_A/K_S > 1$ Peaks, K_S Dips Are More Common Than K_A Peaks

To investigate the relative contribution from a decrease in K_S and the increase in K_A to locally observed peaks in the K_A/K_S ratio, a sliding window analysis was implemented and applied to 1074 mouse-rat genes longer then 1500 base pairs (bp). For a given window size we considered overlapping windows in the gene and identified those windows showing $K_A/K_S > 1$. A peak was defined as a maximal point in K_A/K_S with at least six windows on either side of the peak having a lower ratio. All peaks were categorized as either being K_A peaks (when K_A was unusually high and K_S not unusually low), a K_S dip (when K_S was unusually low and K_A not unusually high), or undetermined (see Methods). We applied both a strict and a more generous set of definitions (see Methods).

Any results are likely to be sensitive to choice of window size, as there exists a compromise between the calculation accuracy of the substitution rate and the dilution of the signal from potentially selected regions by neutrally evolving neighbouring sequence. As the calculation of K_A/K_S for sequences shorter than circa 100 codons is thought to be error-prone, we set this as a lower limit but repeated the analysis using multiple larger window sizes.

As window size increases, so the absolute number of K_A/K_S peaks is reduced, as one might expect as larger windows potentially dilute a weak signal produced by selection on a small area (Fig. 1). Depending on window size, a $K_A/K_S > 1$ peak is found at a rate 0.15–0.2 per gene, with approximately 10% of genes showing at least one $K_A/K_S > 1$ peak. With a mean coding sequence size of around 2300 bp, this approximates to one peak per 12–15 kb of exonic sequence. Unexpectedly, at all window sizes we observe the same trend, namely, that K_S dips contribute to a larger proportion of peaks in K_A/K_S ratio than an increase in K_A (for generous and strict definition results see Fig. 1; for strict definition alone see Supplementary Fig. 1). The relative proportion of peaks that are K_S dips as opposed to K_A peaks varies from ~60% to just over 50% in the longer windows.

Allowing for False Positives

A problem with the sliding window analysis is the occurrence of spurious ("false-positive") peaks, not least because, even in randomly generated genes with no force producing intragenic heterogeneity in K_S, a high variance between windows is nonetheless possible. To minimize this possibility, and hence to control for spurious peaks owing to use of multiple windows in the same gene, a randomization was implemented. The codons of each gene were shuffled for 100 repetitions; the resulting sequences were assessed by the same sliding window analysis that determines the proportion of ratio peaks for a window size of 102 effective codons. A real peak was determined to be significant if $<5\%$ of the 100 simulants of each gene were able to produce a peak with a higher ratio. The sliding window analysis was repeated, as before, to identify the effects of increasing window sizes on the purged set of genes with no loss of magnitude in the contribution by K_S. Of 103 genes with at least one

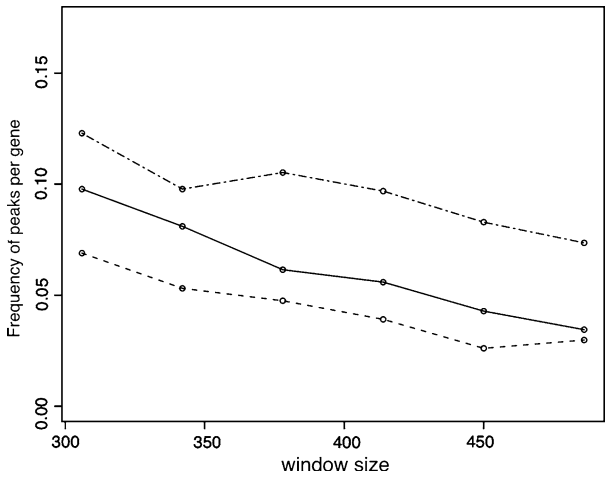

Fig. 1. The frequency of K_A/K_S peaks per gene that are due to either a peak in K_A, hence positive selection (dashedline), a dip in K_S, hence synonymous purifying selection (solidline), or an ambiguous cause where both events are concomitant (dash-dot line), as a function of the size of the sliding window.

intragenic K_A/K_S peak > 1, we find that 47 have at least one peak significantly higher than unity. Of the significant peaks, again, there are more that are K_S dips rather than K_A peaks (35 K_S dip in 25 genes versus 7 K_A dips in 6 genes). This qualitative finding is true under both the strict (Supplementary Fig. 2) and the more generous definitions (Supplementary Fig. 3). Sliding window plots of all genes with a $K_A/K_S > 1$ peak are shown in Supplementary Fig. 4.

Statistically Significant K_S Dips Are for the Most Part Not Repeatable

Can we be confident that the few K_A/K_S peaks that are significant and associated with reduced K_S are really the result of purifying selection on synonymous mutations? To examine this, we took the seven genes (Supplementary Table 1) in which we have K_S dips that are strictly defined and statistically significant and found, via homologene at NCBI, the sequence of human and dog (or pig) orthologues. We then performed a four-way alignment with the mouse-rat sequence and reimplemented the sliding window analysis.

As can be seen (Figs. 2a and b) only two of the seven show both $K_A > K_S$ classified as a K_S dip at the same intragene location in mouse-rat as in human-dog (for the other five see Supplementary Fig. 5; for details of orthologues see Supplementary Table 1). The two genes with repeatable K_S dips are retinoid X receptor interacting protein 110 (*Rxrip110*: dip at ~600 nucleotides in alignment) and chloride channel CLIC-like 1 (*Clccl*: dip at ~1200 nucleotides in alignment). Only in the former case is the K_S dip in the human-dog comparison a strictly defined dip. These two constitute the strongest evidence that we have that some deterministic force is constraining K_S in these localized subdomains.

If we extend this final analysis to include those K_S dips that are significant and more generously defined (of which there are 18 individual dips, including the 7 strictly defined ones, for which we can obtain informative four-way alignments), we find one more instance of repeatability, this being in nuclear autoantigenic sperm protein (*Nasp*) (Fig. 2c). In another case, *Ltbpl*, the K_S shows striking reduction in the same domain in both comparisons, but only in the mouse-rat analysis does K_A exceed K_S (Supplementary Fig. 6g). Thus, we find that only 3, possibly 4, of 18 K_S dips show repeatability of the dip (see Supplementary Table 1, Supplementary Fig. 6). This finding suggests that K_S dips at K_A/K_S peaks can at best be a weak guide to domains of interest, if without the support of independent confirmation. Note, however, that our dual criteria of both statistical significance and repeatability may be too stringent and provide false negatives. For example, the peak in *Brca1*, associated with splice control, while repeatable in at least two independent contrasts, is not statistically significant, owing to the high variance in K_A and K_S of this gene.

K_S Dip Domains Are Not Associated with Low Synonymous SNP Counts

While the lack of repeatability of K_S dips is strongly suggestive of spurious significance, perhaps those that are nonrepeatable may yet be under selection, but just in rodents? To address this possibility, we assessed the SNP density within our critical windows. If there is an element within our critical windows at which selection is strong enough to reduce K_S, then we would expect the SNP density within this region also to be reduced.

SNP data for all genes in our sample were obtained from dbSNP at NCBI. This provides data on the location of synonymous SNPs within our genes (but not their frequency). We could then determine the number of synonymous SNPs within the critical K_S dip windows. We then compare this number to the number expected in and out of the critical windows using a chi-square test, given the number of SNPs in the sample as a whole and the relative proportion of sequence contained within the K_S dip windows. The test was repeated for the more stringent definitions of our critical windows. We also employed two different nulls, one employing the SNP density across all genes in the sample and a second applying the SNP density across only those genes within which we find K_S dips. Given the possibility of between genes deterministic differences in SNP density, the second is probably the more stringent.

Table 1. Chi-square analysis of SNP density in genes associated with K_s dips: (A) comparison of the grouped "critical" windows to a grouped set of their parent genes; (B) comparison of the grouped "critical" windows to the whole data set

	Observed window	Expected window	Window χ^2	Observed remainder	Expected remainder	Remainder	Final
A							
Generous	23	32.5	2.77	184	174.5	0.52	3.29
Generous + significant	14	19.1	1.36	130	124.9	0.21	1.57
Strict	9	13.9	1.75	63	58.1	0.42	2.17
Strict + significant	4	5.70	0.51	31	29.3	0.10	0.60
B							
Generous	23	34.7	3.92	3531	3519.3	0.04	3.96
Generous + significant	14	17.0	0.52	3540	3537.0	0.003	0.52
Strict	9	13.2	1.32	3545	3540.8	0.005	1.33
Strict + significant	4	3.70	0.03	3550	3550.3	0.00003	0.03

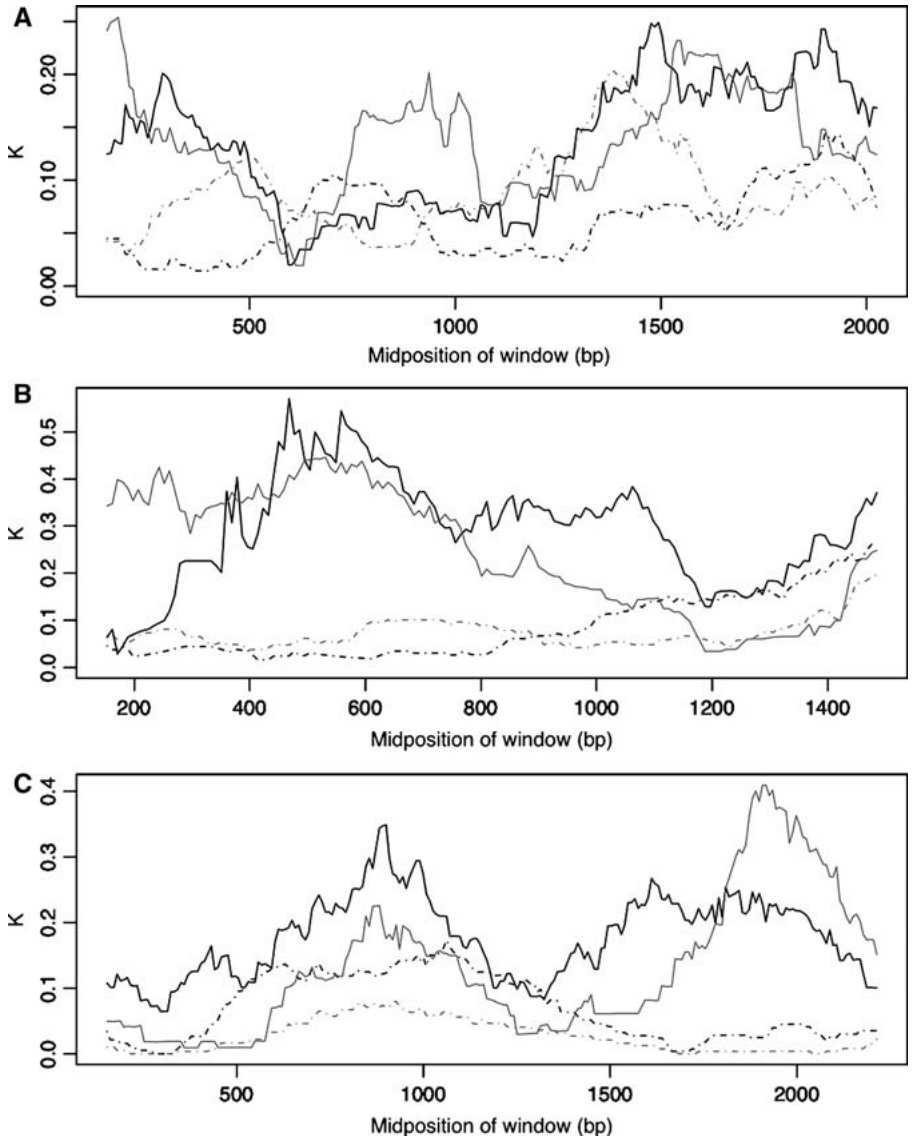

Fig. 2. Three genes with repeatable K_s dips at K_A/K_s peaks. **a** Retinoid X receptor interacting protein 110 (dip at ~600 nucleotides in alignment). **b** Chloride channel CLIC-like 1 (dip at ~1200 nucleotides in alignment). **c** Nuclear autoantigenic sperm protein (Nasp; dip at ~1300 nucleotides in alignment). Mouse/rat sequences are shown in gray; human/other, in black. K_s is a solid line: K_A, a dotted line.

Although the observed numbers of SNPs within our windows is consistently lower than the predicted value, this value was only significant in the most liberally defined set (generous dip definition including nonsignificant peaks) when applying a null derived from the complete gene set. We would also expect

that the chi-square value would increase with the stringency of the definition, but this was not observed (Table 1). Moreover, if we use the number of synonymous SNPs seen in the genes containing the K_S dips as the basis for the null expectation, then we see no significant reduction in the K_S dip domains. We conclude that we find no evidence that domains identified as K_S dips have unusually low SNP rates, be they significant or not.

Discussion

We find that K_A/K_S intragenic peaks are relatively common, occurring about once for every 10,000 bp of exonic sequence, this figure being dependent on the window size employed. Many of these, if not most, are not statistically significant and may be regarded as spurious. When all the peaks, be they significant or not, are divided into those that are likely owing to regional increases in rates of protein evolution or regional reduction in the rate of synonymous evolution, unexpectedly we find an excess of the latter. Employing stricter definitions of the form of the peak and requiring the peak to be significantly high does not alter this conclusion but does render the number of incidences much lower. Assuming that our sample of well-described, long genes is representative of genes more generally, this suggests that it is unwise to make the inference that a $K_A/K_S > 1$ peak is in and of itself indicative of a region of positive selection. The lack of evidence for repeatability and the lack of evidence for selection from SNP data also suggest that many, if not most, of those few K_S dips that are statistically significant are also possibly spurious.

In sum, from over 1000 long genes we have found only three examples of K_S dips for which we can provide prima facie evidence that the low synonymous substitution rate associated with the dip requires a special explanation (other than spurious noise). Can we say anything about the cause of the reduced K_S in the three repeatable cases? An association between alternative splicing and high K_A/K_S (Chen et al. 2006; Ermakova et al. 2006; Plass and Eyras 2006; Xing and Lee 2005a, 2005b, 2006a, 2006b) and low K_S (Parmley et al. 2006) has been suggested in several studies. Might there be a correlation here as well? Given the difficulties associated with providing any definitive statements as to the presence/absence of alternative transcripts, to investigate this we examined three resources: the alternative splicing database (http://www.ebi.ac.uk/asd/) (Stamm et al. 2006), the alternative transcript database (http://www.ebi.ac.uk/atd/) (Le Texier et al. 2006), and Ensembl. In two cases (*Rxrip 110* and *Nasp*) we find that the repeatable K_S dips are in "cleanly" described alternative exons, by which we mean that there are long transcripts having nearly all the same exons, but excluding the K_S dip-containing one (Table 2, Supplementary Fig. 7). In the other case (*Clcc1*) we find evidence for a long transcript that omits the final two exons, the K_S dip being in the last but one exon (Supplementary Fig. 7). Analysis of the orthologous human sequences, within the same reference databases, reveals evidence that this alternative splicing is maintained in the three human genes (data not shown).

In rejecting the 15 statistically significant but nonrepeatable examples, might we have been too stringent? Might it be the case that the SNP data are too sparse to be informative and the repeatability assay unnecessarily restrictive, excluding real examples of selection unique to rodents? One way to address this is to test for particular mechanisms by which selection acts on synonymous sites. Such tests are by necessity weak, as they require the majority of dips to be associated with the same form of selection. We shall, however, examine the two dominant and related explanatory variables: alternative splicing and splice control. For all the genes with significant peaks we hence scrutinized the same alternative transcript resources as above to look for an association with alternative splicing (Table 2). While there is evidence that the three repeatable cases are associated with alternative exons, we find no similar evidence to imply that the remainder are. For a few of the genes there is evidence for very small transcripts missing most exons (including the one with the K_S dip) but no evidence for any association with cleanly described alternative exons. To further check we also consulted the Hollywood database of alternative splicing (http://hollywood.mit.edu/Login.php) and, again, found no evidence for an association between the nonrepeatable dips and alternative exons (data not shown).

Perhaps the K_S dips are not the result of alternative exons but are more generally owing to selection on synonymous mutations associated with splicing control, as evidenced by the reduced SNP density and synonymous rates of evolution in exonic splice enhancers (ESEs) near intron-exon boundaries (Carlini and Genut 2006; Fairbrother et al. 2004a; Parmley et al. 2006). Are, then, the K_S dip domains enriched for splice enhancer elements, and are they associated with intron-exon boundaries? To deduce whether the windows defined by K_S dips have significantly greater proportions of ESE than we would expect, we compared the windows in which the peak in K_A/K_S occurred, that are putative K_S dips, with all other windows from the same gene. We find no evidence for enrichment of K_S dip regions with splice enhancer hexamers (for significant K_S dips, $Z = 0.09 \pm 1.08$, $p > 0.339$). The same is true if we analyze all possible K_S dips, be they significant or not (mean

Table 2. Association of the exon containing or spanned by the statistically significant K_s dip domains with alternative exons

Mus_num	Gene	Mus refseq	Alt transcripts database	Alt splice database	Ensembl annotation
3885	Cypla2	NM_009993	No information	No information	1 transcript
6176	Il21r	NM_021887	No information	4 transcripts: K_s dip in exons seen in 2 long transcripts, absent in 2 3′ truncated short transcripts	2 transcripts: K_s dip in constitutive exon
7057	Clccl	NM_145543	5 transcripts: all but 1 with final 2 exons. K_s dip in last exon but 1	5 transcripts only 1 has last 2 exons, K_s dip in last but 1	1 transcript
7572	Nasp	NM_016777	4 transcripts: 2 transcripts almost identical, differing by large exon. Repeatable K_s dip in this alt exon. Nonrepeatable dip in the 5′ exon and this alternative exon.	4 transcripts: 1 K_s dip within and 1 K_s dip adjacent to exon that is whole in 1 transcript, truncated in 1 transcript, and absent in remaning transcripts	3 transcripts: Repeatable K_s dip in large at exon
8715	Pde3a	NM_018779	No information	No information	1 transcript
10130	Rxrip110	NM_011307	4 transcripts: 2 long ones have exon with K_s dip. The short ones truncate at the K_s dip-containing exon.	4 transcripts: K_s dip in exon seen in 2 long transcripts, absent in 2 3′ truncated transcripts	3 long transcripts: K_s dip in alt exon
10579	Slc22a5	NM_011396	No information	No information	1 transcript
1122	Lrrc56	NM_153777	No information	No information	3 transcripts: K_s dip in constitutive exon
3769	Cspg3	NM_007789	No information	No information	1 transcript
4888	Fbxo7	NM_153195	No information	Incomplete data set: whole transcript not found in 4 transcripts	1 transcript
5179	Gabrq	NM_020488	No information	No information	1 transcript
5620	Grn	NM_008175	No information	12 transcripts: K_s dip in exon seen in 6 transcripts, 2 transcripts have alternative 5′ exon	1 transcript
6001	Hspa4	NM_008300	A very short transcript lacking the exons with the K_s dip is also seen	5 transcripts: K_s dip spans exons in long transcript: neither seen in short transcripts	1 transcript
6317	Itpkc	NM_181593	No information	2 transcripts: K_s dip in first exon of long transcript, absent in 5′ truncated short transcript	1 transcript
6854	Ltbp1	NM_019919	No information	Incomplete data set; whole transcript not found in 11 transcripts	2 transcripts: a very short transcript lacking the exon with the K_s dip is also seen.
10711	Slc5a6	NM_177870	No information	No information	2 transcripts: K_s dip in constitutive exon
11865	Trpv2	NM_011706	2 very short transcripts lacking the exon with the K_s dip also seen	3 transcripts: K_s dip in exon seen in 1 transcript, absent in 5′ truncated short transcripts	1 transcript

$Z = 0.03 \pm 1.08$, $p = 0.308$). Similarly, for those K_S dips in which most of the sequence is near an intron-exon junction (>90% within 70 bp), we see no evidence for enrichment in ESEs compared with other windows in the same gene equally close to junctions (mean $Z = 0.04$, $p \gg 0.05$).

Are K_S dip domains especially close to intron-exon junctions? We determined the proportion of the window containing the K_A/K_S peak, due to reduced K_S, that was within 70 nucleotides of any intron-exon boundary, this being thought to be the approximate span of splice regulating elements. To determine whether there was any skew compared to that we would expect by random chance, a simulation was employed. Each critical window was compared to 100 randomly sampled windows from the same gene. A p-value was determined as the fraction of randomly sampled windows that had a greater than or equal proportion of sequence within 70 bp of an intron-exon boundary, compared to the critical window. Of the 128 dips under the broadest definition, seven reside closer to boundaries than expected by chance, at $p = 0.05$. None of these are significant peaks. Note too that the proportion of p-values falling below 0.05 was $7/128 = 0.054$: more or less as might be expected by chance (we confirmed this also by simulation using a randomly picked window as a pseudo-K_S dip; data not shown).

From the above tests we surmise that there is little reason to suppose that the K_S dips that we rejected as probably being spurious were falsely rejected. Employing the direct tests, however, we found one noteworthy aspect to the data: those K_S dip windows with no sequence within 70 bp of an intron-exon boundary avoid the center of the exon (Fig. 3). While the dip test for bimodality (Hartigan and Hartigan 1985; Hartigan 1985) fails to reject the null of unimodality, there is a tendency for the K_S dip windows to be non-randomly located. If we section the interior of the exon into quarters, we can confirm this skew by chi-square analysis. If peaks were to occur evenly throughout the exon (stochastic variance), then we would expect 6.25 ratio peaks per quarter; what we actually see is a distribution of 12:8:4:1, $\chi^2 = 11.00$, $p < 0.001$. In addition, there is a strong correlation observed between the position of the window within the exon and the position of the window within the gene (see Supplementary Fig. 8, Spearman rank correlation $= -0.5523$, $p = 0.0042$): exons near the beginning of the gene tend to have ratio peaks in the 3′ region, whereas those exons nearer the 3′ end of genes tend to have peaks at their 5′ ends. Why this might be is far from transparent. The windows away from boundaries are not especially enriched for exonic splice enhancers: of 42 K_S dips, 22 have a higher ESE density than the other windows from the same gene, and the remaining 20 having a lower density (Sign test, $p = 0.88$).

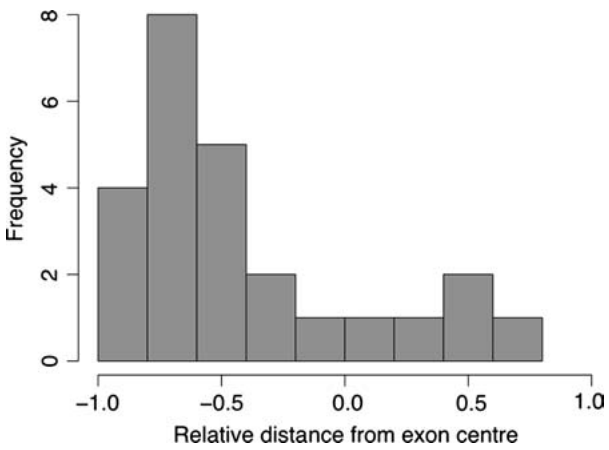

Fig. 3. The relative location of the center of the K_S dip window as a function of proximity to intron-exon junctions for those windows with no part of the sequence within 70 bp of an intron-exon window.

From the above we conclude that, after attempting to make some allowance for problems inherent in sliding window analysis, domains for which we can find coherent evidence that synonymous mutations are under selection appear to be rare. One could equally well conclude that sliding window analysis, while a common method (see, e.g., Endo et al. 1996; Fares et al. 2002; Gissen et al. 2005; Huttley et al. 2000; Talbert et al. 2004) and featured in several computer applications (e.g., Fares 2004; Filatov 2002; Liang et al. 2006; Rozas and Rozas 1999), is both a weak and hard-to-defend mode of analysis. It is striking that for so many of the $K_A/K_S > 1$ peaks we cannot eliminate the possibility of spurious occurrence owing to multiple sampling. That there are more $K_A/K_S > 1$ peaks resulting from lowered K_S (spurious or otherwise) than raised K_A also suggests that it is very unwise to take $K_A/K_S > 1$ peaks as prima facie evidence for positive selection.

Recently, there have also been a number of genome scans to identify positive selection using polymorphism data alone or in addition to divergence data (e.g., Carlson et al. 2005; Hanchard et al. 2006; Hutter et al. 2006; Voight et al. 2006). Whether the same false-positive problems apply in these instances remains to be seen. It is also unclear whether similar problems will affect more sophisticated sliding window analyses. For example, Liang et al. (2006) have developed a sliding window K_A/K_S procedure that allows windows to be defined by reference to the three-dimensional structure of the protein. Given our results, we would suggest that, to be conservative, even in this more directed approach, interpretation of an intragenic K_A/K_S ratio > 1 is best treated with caution.

Acknowledgments. We wish to thank the editor and two referees for constructive comments on an early version of the manuscript. J.L.P. is funded by the Biotechnology and Biological Sciences Research Council.

References

Carlini DB, Genut JE (2006) Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J Mol Evol 62:89–98

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA (2005) Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 15:1553–1565

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298

Chamary JV, Hurst LD (2005a) Biased codon usage near mtron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet 21:256–259

Chamary JV, Hurst LD (2005b) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol 6:R75

Chamary J-V, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98–108

Chen FC, Wang SS, Chen CJ, Li WH, Chuang TJ (2006) Alternatively and constitutively spliced exons are subject to different evolutionary forces. Mol Biol Evol 23:675–682

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963

Duan JB, Wainright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, Gejman PV (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum Mol Genet 12:205–216

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13:685–690

Ermakova EO, Nurtdinov RN, Gelfand MS (2006) Fast rate of evolution in alternatively spliced coding regions of mammalian genes. BMC Genomics 7:84

Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. Science 297:1007–l013

Fairbrother WG, Holste D, Burge CB, Sharp PA (2004a) Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2:E268

Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB (2004b) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res 32:W187–W190

Fares MA (2004) SWAPSC: sliding window analysis procedure to detect selective constraints. Bioinformatics 20:2867–2868

Fares MA, Elena SF, Ortiz J, Moya A, Barrio E (2002) A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. J Mol Evol 55:509–521

Filatov DA (2002) PROSEQ: A software for preparation and evolutionary analysis of DNA sequence data sets. Mol Ecol Notes 2:621–624

Gissen P, Johnson CA, Gentle D, Hurst LD, Doherty AJ, O'Kane CJ, Kelly DA, Maher ER (2005) Comparative evolutionary analysis of VPS33 homologues: genetic and functional insights. Hum Mol Genet 14:1261–1270

Hanchard NA, Rockett KA, Spencer C, Coop G, Pinder M, Jallow M, Kimber M, McVean G, Mott R, Kwiatkowski DP (2006) Screening for recently selected alleles by analysis of human haplotype similarity. Am J Hum Genet 78:153–159

Hartigan JA, Hartigan PM (1985) The dip test of unimodality. Ann Stat 13:70–84

Hartigan PM (1985) Computation of the dip statistic to test for unimodality. J Roy Stat Soc C App Stat 34:320–325

Hurst LD (2006) Preliminary assessment of the impact of micr-oRNA-mediated regulation on coding sequence evolution in mammals. J Mol Evol 63:174–182

Hurst LD, Pal C (2001) Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet l7:62–65

Hutter S, Vilella AJ, Rozas J (2006) Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinform 7:409

Huttley GA, Easteal S, Southey MC, Tesoriero A, Giles GG, McCredie MRE, Hopper JL, Venter DJ (2000) Adaptive evolution of the tumour suppressor BRCA 1 in humans and chimpanzees. Nat Genet 25:410–413

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315:525–528

Le Texier V, Riethoven JJ, Kumanduri V, Gopalakrishnan C, Lopez F, Gautheret D, Thanaraj TA (2006) AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. BMC Bioinform 7:169

Li W-H (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Liang H, Zhou W, Landweber LF (2006) SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. Nucleic Acids Res 34:W382–384

Nackley AG, Shabalina SA, Tchivileya IE, Satterfield K Korchynskyi O, Makarov SS, Maixner W, Diatchenko L (2006) Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science 314:1930–1933

Orban TI, Olah E (2001) Purifying selection on silent sites—a constraint from splicing regulation? Trends Genet 17:252–253

Palsson S (2004) On the effects of background selection in small populations on comparisons of molecular variation. Hereditas 141:74–80

Pamilo P, Bianchi NO (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol 10:271–281

Parmley JL, Chamary JV, Hurst LD (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol 23:301–309

Plass M, Eyras E (2006) Differentiated evolutionary rates in alternative exons and the implications for splicing regulation. BMC Evol Biol 6:50

Rozas J, Rozas R (1999) Dna SP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

Schattner P, Diekhans M (2006) Regions of extreme synonymous codon selection in mammalian genes. Nucl Acids Res 34:1700–l710

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, Tang Y, Barbosa-Morais NL, Thanaraj TA (2006) ASD: a bioinformatics resource on alternative splicing. Nucleic Acids Res 34:D46–D55

Talbert P, Bryson T, Henikoff S (2004) Adaptive evolution of centromere proteins in plants and animals. J Biol 3:18

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:e72

Willie E, Majewski J (2004) Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet 20:534–538

Xing Y, Lee C (2005a) Assessing the application of Ka/Ks ratio test to alternatively spliced exons. Bioinformatics 21:3701–3703

Xing Y, Lee C (2005b) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proc Natl Acad Sci USA 102:13526–13531

Xing Y, Lee C (2006a) Alternative splicing and RNA selection pressure—evolutionary consequences for eukaryotic genomes. Nat Rev Genet 7:499–509

Xing Y, Lee C (2006b) Can RNA selection pressure distort the measurement of Ka/Ks? Gene 370:1–5

Yeo G, Hoon S, Venkatesh B, Burge CB (2004) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc Natl Acad Sci USA 101:15700–15705