The regulation of splicing and the evolution of mammalian genes

Parmley, Joanna Louise

Award date:
2007

Awarding institution:
University of Bath

The regulation of splicing and the evolution of mammalian genes

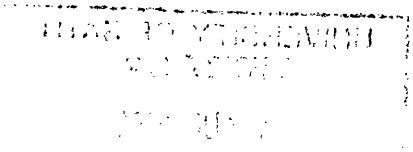
Joanna Louise Parmley

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

October 2007



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Contents

Acknowledgements

Collaborator contribution

Unless otherwise stated, all the work presented in this thesis is my own. Chapter 2 was co-written by myself along with Laurence D. Hurst and Jean-Vincent Chamary. Compilation of Table 1 was exlusively my responsibility. Jean-Vincent Chamary contributed the human-chimp comparison (that features in the supplement) to Chapter 3, otherwise, the majority of the work was my own (Jean-Vincent Chamary independently replicated these results). The analyses in Chapters 4 and 5 were my own. In Chapter 6, the dataset of functional retrogenes was provided by Henrik Kaessmann and Lukasz Potrzebowski, while the expression dataset was compiled by Araxi O. Urrutia.

Appreciation and dedication

Firstly, I would like to thank my supervisor, Laurence Hurst for giving me this opportunity and aiding me in getting the most out of my time in Bath. It is an experience I have thoroughly enjoyed and believe that I have learnt many things that I can take with me and that will be with me always. I would also like to thank Jean-Vincent Chamary who was very welcoming and, even though we no longer share a group, is still an important friend and colleague. Thanks also to my second supervisor, Martin Lercher, who gave me the opportunity to work in his groups in 2 other institutions and taught me valuable lessons in self-confidence for which I'm especially grateful. And finally, I would like to thank my family and friends who have supported me through any tough time I've faced and to whom I would like to dedicate this thesis.

My work was funded by a studentship from the Biotechnology and Biological Sciences Research Council.

Abbreviations

A	Adenine
Bp	Base pairs
C	Cytosine
ESE	Exonic splicing enhancers
ESS	Exonic splicing suppressors
G	Guanine
GFP	Green fluorescent protein
ISE	Intronic splicing enhancers
ISS	Intronic splicing suppressors
KA	The rate of substitutions at non-synonymous sites
KI	The rate of substitution at intronic sites
KS	The rate of substitution at synonymous sites
Ne	The effective population size
SNP	Single nucleotide polymorphism
SS	Splice site
T	Thymine
U	Uracil

Summary

Mammalian genes are unusual in being especially rich in introns. Does the need to remove these non-coding sequences from the immature mRNA impact on the evolution of the flanking coding exons? Until relatively recently, one might *a priori* have supposed not, as the information within the immature mRNA that specified the location of intron-exon boundaries was considered to dominantly be within the introns. More recent evidence, however, has highlighted the role of SR proteins, which bind to exonic splice enhancer domains, in facilitating intronic splicing. Does the need for exons to specify such binding domains impact on their rate of evolution and, if so, what is the magnitude of such effects? In this thesis I show that, in mammals, synonymous mutations within exonic splice enhancers are under stronger purifying selection than those not in enhancers and that codon usage bias is, in part, determined by the need to specify such enhancers. Comparably, amino acid choice is biased and rates of non-synonymous evolution are lower when associated with splice enhancers. These results are non-trivial in magnitude. The effect of selection near intron-exon boundaries is approximately as important a predictor of rates of protein evolution as are expression parameters. As regards the selection on synonymous mutations, by review of the literature I argue that it is no longer tenable to consider that synonymous mutations in mammals are neutrally evolving. The shift in our understanding stems from the provision of sound mechanistic underpinnings regarding the targets of such selection (rather than reliance on indirect inferences). The possibility that selection on synonymous mutations might mislead methods to detect positive selection, when sliding window protocols are employed, is shown to be acute. These results have implications for diagnosing the mode of splicing with nothing more than a well-annotated genome and for inferring positive selection. They also suggest the possibility of intelligent adjustment of transgenes so as to improve their efficacy.

Chapter1. **Introduction**

1.0 Introduction

In eukaryotic genomes, gene sequences are often interrupted by non-coding sequence, known as introns, which must be removed from the mRNA in order for the sequence to be translated (Berget, Moore, and Sharp 1977; Chow et al. 1977). The identification of the introns, their removal and the ligation of the resulting flanking (exonic) ends is collectively known as splicing. Classically, it was believed that all the information necessary for the identification and the regulation of the removal of these intronic sequences was contained within the introns themselves (for review see Burge, Tuschi, and Sharp 1999). These early models of splicing were derived from observations of mRNA processing in yeast (Burge, Tuschi, and Sharp 1999). However, recent evidence from studies in mammals has shown that this intronic information alone is insufficient for efficient and accurate splicing (Burge, Tuschi, and Sharp 1999; Blencowe 2000). In higher eukaryotes regulatory sequences within the exons are required to regulate the often complex use of splice sites (Blencowe 2000). In this thesis I examine the extent to which these sequence requirements affect the evolution of genes, the rates of synonymous and non-synonymous nucleotide substitution, amino acid choice and codon bias.

In this introduction I shall briefly review what is known of the mechanism of splicing. My thesis concentrates in no small part on deviations from neutral expectations as regards evolution at synonymous sites in mammals. Evidence for this can come from examination of codon usage bias and of rates of evolution at synonymous sites. As a necessary preamble to this, in this introduction I shall also be outlining the null neutral hypothesis for molecular evolution (Kimura 1983). I also examine the impact of splicing on amino acid content and rates of evolution of proteins. In this introduction I will also briefly discuss other correlates to protein evolutionary rates.

1.1 Control of splicing

Throughout eukaryotes, introns are defined by a combination of degenerate, yet essential, motifs: the 5' (donor) and 3' (acceptor) consensus site sequences (ss) at the intron ends; a branch point sequence, characterised by a highly conserved adenosine residue and a polypyrimidine tract, located between the branch point and the 3' ss. These sites are necessary for the localisation of the spliceosome, a large complex, which catalyses

the removal of the intron and the ligation of the flanking exonic sequences. The removal of an intron is a two-step process, involving two transesterification reactions, forming a lariat intermediate molecule. Firstly, the phosphodiester bond at the 5' ss is broken as the 5' end of the intronic sequence binds to the branch point, leaving a free 3' hydroxyl group on the 5' exon. This free end then attacks the 3' ss, ligating the two exons and displacing the intron as a separate lariat molecule. As, during these steps, the pre-mRNA sequence is divided into 2 separate molecules, the spliceosome is required to ensure their proximity for the ultimate stages. The formation of the spliceosome requires the coordination of 5 small nuclear (sn) RNAs and a large contingent of polypeptides, most recently thought to be greater than 300 in number. The mechanism of action of this vast complex, however, is yet poorly understood (for an overview of the mechanism of splicing and the spliceosome see Burge, Tuschi, and Sharp 1999). The initiation of spliceosome assembly is the most well characterised process and includes the recognition of the pre-mRNA by some essential factors. The snRNP U1 associates with the 5' ss as does the branch point sequence with U2 snRNP. The polypyrimidine tract (Py-tract) and the 3' ss associate with the U2 auxillary factor (U2AF), a heterodimer, the smaller fraction to the 3' ss, the larger to the Py-tract (Reed 1996).

Within plant and Metazoan genomes, the 5' and 3'ss sequences are degenerate, resulting in a large variation in the affinity of splice sites for those factors required to initiate splicing. By pure chance it is possible that sequences occur, with no intended splicing function, that have a stronger splicing signal than the true splice site. In fact, pseudo-exons greatly outnumber true exons, but are not included within the mature mRNA (Sun and Chasin 2000). It is also proposed that around 70 percent of human genes undergo some degree of alternative splicing (Johnson et al. 2003). For these reasons, the intronic definition of splice sites provides only half the information necessary for the regulation of splicing. The splicing information of intended splice sites are enhanced, or in some cases suppressed, by the binding of trans-factors to cis-regulatory elements within both constitutive and alternatively spliced exons. Serine-arginine rich proteins (SR proteins) interact with the pre-mRNA through an N-terminal RNA recognition motif to recognise Exonic Splicing Enhancer (ESE) sequences near the splice sites (Blencowe 2000). The SR proteins are capable of protein-protein interactions via a C-terminal RS domain enabling the enhancement of splicing by facilitating the assembly of the spliceosome (Wu and Maniatis 1993; Kohtz et al. 1994; Roscigno and Garcia-Blanco 1995; Zuo and Maniatis 1996; Hertel and Graveley 2005).

The RNA-recognition domains of SR proteins are, to some degree, degenerate, potentially allowing them to bind to several similar sequences (Tacke and Manley 1995; Liu, Zhang, and Krainer 1998; Tacke et al. 1998). It was observed in C. elegans that the individual targeted knockdown of 6 SR proteins did not cause an observable phenotype, indicating that other SR proteins can compensate for the loss of one (Longman, Johnstone, and Caceres 2000). The employment of specific SR proteins to different enhancer elements allows the flexibility needed to orchestrate the intricate temporal and spatial alternative splicing of some genes to be executed with reliable accurately. Exonic splicing enhancers (ESEs) are the best studied of the splicing cis-regulatory elements. Fairbrother *et al.* computationally derived human hexameric sequences that possess enhancer qualities by a process known as RESCUE-ESE (Fairbrother et al. 2004b). These sequences were accepted as enhancers if they were found significantly more often in exons than in intron sequences and were more common in exons that were defined by weak 5' or 3' ss. Other methods of identifying ESEs include SELEX, selection-based methods, which isolate ESE sequences by amplifying fractions in reporter assays (Tian and Kole 1995; Coulter, Landree, and Cooper 1997; Liu, Zhang, and Krainer 1998; Schaal and Maniatis 1999). Further work on the Fairbrother ESEs has supported their functional role in splicing in human genes. Upon scanning of exons, the ESE elements are found to be most dense near intron-exon boundaries; this trend was found to coincide with an inverse trend in SNP (single nucleotide polymorphisms) density (Fairbrother et al. 2004a; Willie and Majewski 2004). Other less well-defined cis-regulatory sequences also act to modify splice site usage, including Exonic Splicing Suppressors (ESSs) (Wang et al. 2004), Intronic Splicing Enhancers (ISEs) and Intronic Splicing Suppressors (ISSs), the latter two being generically intronic splicing regulatory elements (Yeo, Nostrand, and Liang 2007).

1.1.1 The neutral theory

What determines the fate of any given new mutation? In a strictly deterministic world, all that would matter would be whether the mutation was beneficial (i.e. under positive selection) or deleterious (i.e. under purifying selection). This deterministic view was however, importantly, challenged in the 1960s. First, from extrapolation from the rate of evolution of a few proteins (e.g. fibronectin), it was determined that the amount of selective mortality required to fix the proposed number of positively selected changes was just too high (Kimura 1968). Second, the observed levels of polymorphism were seen also to be too high (for review see Lewontin 1974). To explain these observations, Kimura proposed an alternative: the neutral hypothesis (Kimura 1983).

Kimura proposed that chance sampling must also be considered. As a consequence, mutations of no impact on fitness could also become fixed. Importantly, the rate of fixation of such mutations is independent of population size and depends only on the mutation rate. Is it, however, reasonable to suppose that a mutation has absolutely no effect on fitness? In part to address this concern, Kimura and Ohta provided an extension to the neutral model, this being the nearly neutral model (for review see Ohta 1992; Ohta and Gillespie 1996). While similar in name the two hypotheses differ considerably in their predictions. Under the nearly neutral model, deleterious mutations can be considered to fit into one of three classes: those for which the effects are so small that their evolutionary rate is effectively the same as perfectly neutral mutations (effectively neutral mutation), those that, despite being deleterious, can reach fixation but at rates below that expected from the neutral model (weakly or slightly deleterious mutations) and, finally, those with no prospect of reaching fixation. Importantly, in this model, which of the three classes a given mutation of deleterious effect belongs to is dependent on the effective population size (N_e). Effectively neutral mutations require $s<<1/2\ N_e$, weakly deleterious ones require $s\sim1/2N_e$ (for review see Ohta 1992; Ohta and Gillespie 1996). Importantly then, small populations are expected to have a higher proportion of mutations that are effectively neutral. It is this understanding that underpins the theory that in mammals those mutations that are weakly selected in flies, worms, yeast and bacteria should be effectively neutral (Sharp et al. 1995).

Consider, for example, the case of synonymous mutations. Due to the degeneracy of the genetic code, two denominations of mutations occur: those that change the encoded amino acid (non-synonymous) and those that maintain the protein sequence (synonymous). Given both this and the finding that the synonymous rate of evolution is much higher than the non-synonymous rate (Kimura 1977), Kimura suggested that synonymous mutations may well be neutrally (or effectively so) evolving. If so, the rate of mutations at synonymous sites can be used as a proxy for the genomic/local mutation rate, a technique often employed (Miyata et al. 1987; Wolfe, Sharp, and Li 1989; Smith and Hurst 1999; Keightley and Eyre-Walker 2000). Once this rate is determined it is a valuable tool for providing a benchmark of the background rate of evolution, thereby identifying proteins, or regions within proteins, that are subject to purifying selection or adaptive evolution (Yang and Bielawski 2000; Hurst 2002).

Is it necessarily the case that synonymous mutations will be of effective neutrality? In some species of fly, worm, yeast and bacteria selection has been shown to act on synonymous mutations (Ikemura 1985; Akashi and Eyre-Walker 1998; Duret 2002; Wright

et al. 2004). Although the mutations at these sites may be very weakly deleterious, the relatively large population size of these species allows selection to act with high efficiency (the reasons for this I discuss at greater length below). Conversely, in mammalian species, the effective population size is relatively low, potentially limiting the efficiency of selection (Sharp et al. 1995; Keightley, Lercher, and Eyre-Walker 2005). Since none of the patterns of synonymous selection observed in lower eukaryotes and prokaryotes have been identified in mammals (but see Iida and Akashi 2000; Lander et al. 2001; Comeron 2004; Lavner and Kotlar 2005) it is presumed that this effect allows synonymous mutations in mammals to evolve with effective neutrality (Sharp et al. 1995). However, these tests assumed just one cause of the selection and ignore the possible role of splicing imposed constraints. To examine these issues I have considered both codon usage bias and rate of evolution.

1.1.2 Codon usage bias

All things being equal, one might expect that, within a gene, all of the codons representing the same amino acid should be used with equal frequency. Any deviation from this is termed codon usage bias. Evidence from all taxa sampled indicates that such bias is the rule not the exception. Why does it occur?

While the neutral theory proposed that any bias is owing to neutral mutation coupled with mutational bias, through the 1970s and 1980s an alternative model gained ground. In this model, selection favours usage of codons that minimize the time required to translate an mRNA and/or to maximize the accuracy of the translation. Evidence for this came from several angles. First, it was found in numerous taxa (worm, fly, bacteria etc), that the codons preferred in a given taxa tended to correspond with the codons matching the more abundant iso-acceptor tRNAs (Ikemura 1985; Akashi and Eyre-Walker 1998; Duret 2002; Wright et al. 2004). Indeed, Bulmer proposed an evolutionary feedback loop in which skewed tRNA usage selects for skewed codon usage which in turn selects for more skewed tRNA usage etc (Bulmer 1987). Similarly, it was observed that the more highly expressed a gene the greater the skew in its codon usage (see e.g. Powell and Moriyama 1997). As also then expected, genes with higher rates of expression and higher codon bias have lower rates of synonymous evolution (see e.g. Powell and Moriyama 1997) (but also see Dunn, Bielawski, and Yang 2001).

For many years now, mammals have been assumed to be different (Sharp et al. 1995). With early limited sample sizes the same patterns as above could not be retrieved (for review see Duret 2002). Coupled with the expectations of the nearly neutral model this lack of evidence was considered as expected (Duret 2002). Even with much larger sample sizes it remains contentious as to whether there is even a very weak tendency in the predicted directions (Kanaya et al. 2001; Urrutia and Hurst 2001; Duret 2002; Urrutia and Hurst 2003; Comeron 2004; dos Reis, Savva, and Wernisch 2004; Lavner and Kotlar 2005). In no small part, the inability to detect signals of codon usage bias in mammals is owing to a massive signal derived from isochores, these being regions of homogeneous GC (Bernardi et al. 1985; Eyre-Walker 1991; Sharp et al. 1995; Smith and Hurst 1999).

Even if there were no translational selection operating in mammals, would it be fair to suppose that selection doesn't operate on synonymous mutations? The problem with the, often assumed, conflation of translational selection with selection on synonymous mutations is that it supposes that no other modes of selection might act. Does then the need to specify ESEs near splice sites induce a bias in codon usage? In chapter 3, I present evidence that selection operates on synonymous sites that function as part of an ESE. In chapter 4, I will present evidence that an observable trend in codon usage is present as a function of distance from intron-exon boundaries. It is also possible to predict the weight of the bias with our current knowledge of ESEs.

In chapter 2, I present a review outlining the above context and evidence, but discussing other potential factors that might act on synonymous mutations in mammals (see also appendix 1 for a brief update and discussion of two new pieces of evidence). One of the first important suggestions that selection might operate on synonymous sites in mammals came from the observation that in one sub-region of BRCA1 there was a very high K_A/K_S peak, observed in sliding window analysis, that was associated not with a high rate of protein evolution, but rather a profoundly low rate of synonymous evolution (Hurst and Pal 2001). This was subsequently argued to be owing to splice related constraints (Orban and Olah 2001). In chapter 5, I return to this issue to ask how often, using a sliding window approach, one might observe $K_A/K_S > 1$ peaks and whether these are better explained by local reductions in K_S or increases in K_A. To my surprise we find more incidences of the former rather than the latter. In incidences where the peak is both statistically significant and repeatable, I also observe that the domain tends to be associated with alternatively spliced exons. These results suggest that selection on synonymous mutations has the potential to provide misleading signals of positive selection, unless the results are handled carefully.

1.2 Protein evolution

For protein evolution, while neutral evolution could function as a null, it is not an especially helpful one, as we have known for a long time that non-synonymous mutations tend not to evolve neutrally ($K_A << K_S$ for most proteins) (Kimura 1977). For proteins then the issues are different. Rather the issue that I consider to be interesting, in the current context, is whether we should regard protein evolution as being driven exclusively by selection to optimise functionality of the protein. Do proteins employ the "optimal" amino acid for the job that the protein must perform? Does intra and inter-genic variation in rates of protein evolution reflect differences in selective constraint within and between proteins?

As regards intra-gene variation, it is logical to suppose that proteins should evolve slowly in regions where function must be strictly maintained (Pal, Papp, and Lercher 2006). Indeed, conservation of key residues for certain classes of functional domain is an important diagnostic of the existence of such domains. Likewise, one might expect that the rate of protein evolution should depend on the density of functionally important domains within the protein i.e. fitness density (Dobson 2004; Drummond et al. 2005). While this may indeed be partially so, the strongest factor for explaining the variation between proteins in their evolutionary rates is expression and translation rates (Pal, Papp, and Hurst 2001; Drummond, Raval, and Wilke 2006). Both those genes that are most highly expressed (Pal, Papp, and Hurst 2001; Rocha and Danchin 2004; Subramanian and Kumar 2004; Wright et al. 2004) and, in multicellular organisms, those that are expressed most widely, temporally and spatially (Duret and Mouchiroud 2000; Subramanian and Kumar 2004), evolve at lower rates than other genes. Why this might be remains unclear, although stronger selection for residues that minimize the impact of protein mistranslation is currently the strongest candidate (Dobson 2004; Drummond et al. 2005). Other factors that are claimed to correlate to protein evolutionary rate include essentiality (Wilson, Carlson, and White 1977; Hirsh and Fraser 2001) and number (Fraser et al. 2002) and type (Fraser 2005) of protein-protein interactions (Fraser et al. 2002), but these issues are not cut and dried (Hurst and Smith 1999; Batada, Hurst, and Tyers 2006), not least owing to covariance with expression rates (Bloom and Adami 2003; Pal, Papp, and Hurst 2003; Bloom and Adami 2004; Batada et al. 2006; Batada et al. 2007).

In this thesis (chapter 6), I examine an alternative possibility, what might be called dual coding. That is to say that the choice of amino acid and both intra and inter gene

variation in rates of evolution might be explained both by the need to specify ESEs near intron-exon junctions and by the needs of the protein. In chapter 6, I ask whether there are trends in amino acid usage near intron-exon junctions, whether these might be predicted by what is known about ESEs, whether rates of evolution are low near boundaries and whether genes with much sequence near boundaries evolve slowly (even allowing for expression parameters). The evidence uniformly supports the possibility of dual coding being an important force.

References

Akashi, H. and A. Eyre-Walker. 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. **8**:688-693.

Batada, N.N., L.D. Hurst, and M. Tyers. 2006. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol **2**:e88.

Batada, N.N., T. Reguly, A. Breitkreutz, L. Boucher, B.J. Breitkreutz, L.D. Hurst, and M. Tyers. 2007. Still stratus not altocumulus: further evidence against the date/party hub distinction. PLoS Biol **5**:e154.

Batada, N.N., T. Reguly, A. Breitkreutz, L. Boucher, B.J. Breitkreutz, L.D. Hurst, and M. Tyers. 2006. Stratus not altocumulus: a new view of the yeast protein interaction network. PLoS Biol **4**:e317.

Berget, S.M., C. Moore, and P.A. Sharp. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A **74**:3171-3175.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunierrotival, and F. Rodier. 1985. The mosaic genome of warm-blooded vertebrates. Science **228**:953-958.

Blencowe, B.J. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem.Sci. **25**:106-110.

Bloom, J. and C. Adami. 2003. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein--protein interactions data sets. BMC Evol Biol **3**:21.

Bloom, J.D. and C. Adami. 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: Response. BMC Evol Biol **4**:art. no.-14.

Bulmer, M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature **325**:728-730.

Burge, C.B., T. Tuschi, and P.A. Sharp. 1999. Splicing of Precursors to mRNAs by the Spliceosomes. Pp. 525-560 *in* R. F. Gesteland, T. R. Cech, and J. F. Atkins, eds. The RNA World. COLD SPRING HARBOUR LABORATORY PRESS, New York.

Chow, L.T., R.E. Gelinas, T.R. Broker, and R.J. Roberts. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell **12**:1-8.

Comeron, J.M. 2004. Selective and Mutational Patterns Associated With Gene Expression in Humans: Influences on Synonymous Composition and Intron Presence. Genetics **167**:1293-1304.

Coulter, L.R., M.A. Landree, and T.A. Cooper. 1997. Identification of a new class of exonic splicing enhancers by in vivo selection. Mol Cell Biol **17**:2143-2150.

Dobson, C.M. 2004. Principles of protein folding, misfolding and aggregation. Semin Cell Dev Biol **15**:3-16.

dos Reis, M., R. Savva, and L. Wernisch. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. **32**:5036-5044.

Drummond, D.A., J.D. Bloom, C. Adami, C.O. Wilke, and F.H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A **102**:14338-14343.

Drummond, D.A., A. Raval, and C.O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol **23**:327-337.

Dunn, K., J. Bielawski, and Z. Yang. 2001. Substitution rates in Drosophila nuclear genes: implications for translational selection. Genetics **157**:295-305.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12**:640-649.

Duret, L. and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**:68-74.

Eyre-Walker, A. 1991. An analysis of codon usage in mammals: selection or mutation bias? J. Mol. Evol. **33**:442-449.

Fairbrother, W.G., D. Holste, C.B. Burge, and P.A. Sharp. 2004a. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol **2**:E268.

Fairbrother, W.G., G.W. Yeo, R. Yeh, P. Goldstein, M. Mawson, P.A. Sharp, and C.B. Burge. 2004b. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res. **32**:W187-190.

Fraser, H.B. 2005. Modularity and evolutionary constraint on proteins. Nat Genet **37**:351-352.

Fraser, H.B., A.E. Hirsh, L.M. Steinmetz, C. Scharfe, and M.W. Feldman. 2002. Evolutionary rate in the protein interaction network. Science **296**:750-752.

Hertel, K.J. and B.R. Graveley. 2005. RS domains contact the pre-mRNA throughout spliceosome assembly. Trends Biochem Sci **30**:115-118.

Hirsh, A.E. and H.B. Fraser. 2001. Protein dispensability and rate of evolution. Nature **411**:1046-1049.

Hurst, L.D. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet **18**:486.

Hurst, L.D. and C. Pal. 2001. Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet. **17**:62-65.

Hurst, L.D. and N.G.C. Smith. 1999. Do essential genes evolve slowly? Curr. Biol. **9**:747-750.

Iida, K. and H. Akashi. 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene **261**:93-105.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13-34.

Johnson, J.M., J. Castle, P. Garrett-Engele, Z. Kan, P.M. Loerch, C.D. Armour, R. Santos, E.E. Schadt, R. Stoughton, and D.D. Shoemaker. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science **302**:2141-2144.

Kanaya, S., Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. **53**:290-298.

Keightley, P.D. and A. Eyre-Walker. 2000. Deleterious mutations and the evolution of sex. Science **290**:331-333.

Keightley, P.D., M.J. Lercher, and A. Eyre-Walker. 2005. Evidence for Widespread Degradation of Gene Control Regions in Hominid Genomes. PLoS. Biol. **3**:e42.

Kimura, M. 1983. The Neutral Theory of Evolution. Cambridge University Press, Cambridge.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature **217**:624-626.

Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267**:275-276.

Kohtz, J.D., S.F. Jamison, C.L. Will, P. Zuo, R. Luhrmann, M.A. Garcia-Blanco, and J.L. Manley. 1994. Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. Nature **368**:119-124.

Lander, E.S., L.M. Linton, B. Birren *et al.* 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.

Lavner, Y. and D. Kotlar. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene **345**:127-138.

Lewontin, R.C. 1974. The Genetic Basis of Evolutionary Change. Columbia University Press, New York.

Liu, H.X., M. Zhang, and A.R. Krainer. 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev **12**:1998-2012.

Longman, D., I.L. Johnstone, and J.F. Caceres. 2000. Functional characterization of SR and SR-related genes in Caenorhabditis elegans. Embo J **19**:1625-1637.

Miyata, T., H. Hayashida, K. Kuma, K. Mitsuyasu, and T. Yasunaga. 1987. Male-driven molecular evolution: a model and nucleotide sequence analysis. Cold Spring Harbor Symp. Quant. Biol. **52**:863-867.

Ohta, T. 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. System. **23**:263-286.

Ohta, T. and J.H. Gillespie. 1996. Development of Neutral and Nearly Neutral Theories. Theor Popul Biol **49**:128-142.

Orban, T.I. and E. Olah. 2001. Purifying selection on silent sites - a constraint from splicing regulation? Trends Genet. **17**:252-253.

Pal, C., B. Papp, and L.D. Hurst. 2001. Highly expressed genes in yeast evolve slowly. Genetics **158**:927-931.

Pal, C., B. Papp, and L.D. Hurst. 2003. Rate of evolution and gene dispensability. Nature **421**:496-497.

Pal, C., B. Papp, and M.J. Lercher. 2006. An integrated view of protein evolution. Nat Rev Genet **7**:337-348.

Powell, J.R. and E.N. Moriyama. 1997. Evolution of codon usage bias in *Drosophila*. Proc. Natl. Acad. Sci. U. S. A. **94**:7784-7790.

Reed, R. 1996. Initial splice-site recognition and pairing during pre-mRNA splicing. Curr Opin Genet Dev **6**:215-220.

Rocha, E.P.C. and A. Danchin. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol. Biol. Evol. **21**:108-116.

Roscigno, R.F. and M.A. Garcia-Blanco. 1995. SR proteins escort the U4/U6.U5 tri-snRNP to the spliceosome. RNA **1**:692-706.

Schaal, T.D. and T. Maniatis. 1999. Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol Cell Biol **19**:1705-1719.

Sharp, P.M., M. Averof, A.T. Lloyd, G. Matassi, and J.F. Peden. 1995. DNA-sequence evolution: the sounds of silence. Phil Trans R Soc Lond B **349**:241-247.

Smith, N.G.C. and L.D. Hurst. 1999. The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? Genetics **152**:661-673.

Subramanian, S. and S. Kumar. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics **168**:373-381.

Sun, H. and L.A. Chasin. 2000. Multiple splicing defects in an intronic false exon. Mol Cell Biol **20**:6414-6425.

Tacke, R. and J.L. Manley. 1995. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. Embo J **14**:3540-3551.

Tacke, R., M. Tohyama, S. Ogawa, and J.L. Manley. 1998. Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing. Cell **93**:139-148.

Tian, H. and R. Kole. 1995. Selection of novel exon recognition elements from a pool of random sequences. Mol Cell Biol **15**:6291-6298.

Urrutia, A.O. and L.D. Hurst. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics **159**:1191-1199.

Urrutia, A.O. and L.D. Hurst. 2003. The signature of selection mediated by expression on human genes. Genome Res **13**:2260-2264.

Wang, Z., M.E. Rolish, G. Yeo, V. Tung, M. Mawson, and C.B. Burge. 2004. Systematic identification and analysis of exonic splicing silencers. Cell **119**:831-845.

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

Wilson, A.C., S.S. Carlson, and T.J. White. 1977. Biochemical evolution. Ann. Rev. Biochem. **46**:573-639.

Wolfe, K.H., P.M. Sharp, and W.-H. Li. 1989. Mutation rates differ among regions of the mammalian genome. Nature **337**:283-285.

Wright, S.I., C.B. Yau, M. Looseley, and B.C. Meyers. 2004. Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. Mol. Biol. Evol. **21**:1719-1726.

Wu, J.Y. and T. Maniatis. 1993. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. Cell **75**:1061-1070.

Yang, Z.H. and J.P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol **15**:496-503.

Yeo, G.W., E.L.V. Nostrand, and T.Y. Liang. 2007. Discovery and Analysis of Evolutionarily Conserved Intronic Splicing Regulatory Elements. PLoS Genetics **3**:e85.

Zuo, P. and T. Maniatis. 1996. The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. Genes Dev **10**:1356-1368.

Chapter 2. **Hearing silence: non-neutral evolution at synonymous sites in mammals**

Jean-Vincent Chamary, Joanna L. Parmley and Laurence D. Hurst
Nature Reviews Genetics (2006) 7(2): 98-108

Hearing silence: non-neutral evolution at synonymous sites in mammals

J. V. Chamary, Joanna L. Parmley‡ and Laurence D. Hurst‡*

Abstract | Although the assumption of the neutral theory of molecular evolution — that some classes of mutation have too small an effect on fitness to be affected by natural selection — seems intuitively reasonable, over the past few decades the theory has been in retreat. At least in species with large populations, even synonymous mutations in exons are not neutral. By contrast, in mammals, neutrality of these mutations is still commonly assumed. However, new evidence indicates that even some synonymous mutations are subject to constraint, often because they affect splicing and/or mRNA stability. This has implications for understanding disease, optimizing transgene design, detecting positive selection and estimating the mutation rate.

Effective population size (N_e)
The number of individuals in a population that contribute to the next generation.

Codon usage
The relative frequency at which alternative codons specifying a particular amino acid are used.

Since its formulation in the 1960s, the neutral theory (BOX 1) has been a powerful null model for molecular evolution[1]. The unexpectedly high rate of evolution of genes indicates that most mutations have no effect on the fitness of an organism and so spread to fixation by chance[2] (drift). If all the mutations in putatively neutrally evolving DNA (for example, introns, intergene spacers and synonymous sites) really are neutral, then the rate of evolution of such a sequence can be used as a convenient measure of the mutation rate (for examples see REFS 3–5). This does not require that all such mutations have absolutely no fitness consequence, just that they must be of such a small effect that they evolve as if they were neutral (BOX 1). For an allele to be 'effectively neutral', the selective disadvantage that is associated with it must be considerably smaller than the inverse of the effective population size (N_e) (BOX 1). Consequently, we should expect neutral or effectively neutral evolution to be more common in species with small populations.

Although many sites in non-coding DNA in mammals are probably neutrally evolving, some intronic sequence is selectively constrained (see below), and up to 15% of non-coding DNA contains functionally important segments[6]. Is there then any class of sequence in which all mutations are likely to be neutral and from which we can therefore derive accurate estimates of the mutation rate? Taking an historical view, we note that mammals are relatively unusual in that it is still believed that all synonymous mutations in mammalian genomes are neutral. Mammals are often considered to be special owing to their small populations (rendering mutations of slight fitness effectively neutral; BOX 1) and because codon usage is largely dictated by patterns of base composition in the genomic region (isochore) within which a gene resides, rather than owing to forces that are specific to exonic regions. However, we argue that this position requires substantial revision, given that recent evidence indicates that synonymous sites are important in mRNA stability and for correct splicing, for example.

*Center for Integrative Genomics, University of Lausanne, Génopode building, CH-1015 Lausanne, Switzerland.
‡Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK.
Correspondence to J.V.C. or L.D.H.
e-mails: j.v.chamary@unil.ch; l.d.hurst@bath.ac.uk
doi:10.1038/nrg1770

The rise and fall of the neutral theory

The original neutral theory proposed that both mutations that have no effect on amino-acid content (non-coding and synonymous changes) and those that alter proteins (non-synonymous changes) could have no effect on fitness and so have their fate dictated by chance alone. The rise of neutralism was supported on two platforms. First, the arrival of protein electrophoresis data implied that polymorphism at the amino-acid level was common. This was not expected under selectionist population genetics, which predicted polymorphism only under special circumstances, such as cases in which heterozygotes are the most fit. By contrast, it was expected under the neutral theory. Second, Kimura[2] argued that the rate of protein evolution was such that, if all differences between species were due to selection, the total amount of selective death would be improbably high.

Although these findings brought the neutral theory to prominence, it has since largely been a theory in retreat. Neutrality alone cannot explain the number of observed polymorphisms[7]. The theory predicts that species that have large populations should show much higher levels of polymorphism than small populations;

Box 1 | The neutral theory, the nearly neutral theory and why mammals might be different

The strict neutral theory considers the fate of mutations that have no effect on fitness. If such mutations occur at a rate of μ per haploid genome per generation, where μ is the neutral mutation rate, then each generation there must have $2N\mu$ new neutral mutations, where N is the diploid population size. What is the fate of any such new mutation? Random fluctuations in allele frequency (drift) allow the new mutation to go up or down in frequency. The chance that the new mutation will become fixed in a population is $1/(2N)$, that is, the same as pulling one white ball from a collection of $2N$ balls where only one is white. Consequently, μ is the rate of fixation, as $(2N\mu)/(2N) = \mu$. Therefore, evolution at neutral sites can be used to estimate the mutation rate.

What if a mutation has only a small effect on fitness? The successor to strict neutrality, the nearly neutral theory[5], considers the fate of such mutations. The theory predicts that a mutation will be 'effectively neutral' if its selective disadvantage (s) is small compared with the effective population size (N_e) (more precisely, if $s << 1/(2N_e)$ (REF. 11)). By effectively neutral, we mean that the fixation rate is so close to μ that it makes no difference. By contrast, if a mutation is slightly deleterious it can be opposed by selection if the fitness effect is larger or the population size is smaller (with $s = 1/(2N_e)$), while still allowing substitutions to occur at some measurable rate (a fixation rate that is less than μ). If the mutation is even more deleterious ($s >> 1/(2N_e)$), then the mutation will not reach fixation. Mutations that cause evident disease are the more extreme examples of those that cannot reach fixation.

Note that what is classified as a slightly deleterious mutation is dependent on the effective population size. A mutation in a fly could be slightly deleterious ($s \approx 1/(2N_e)$), whereas one of the same fitness in a mammal could be effectively neutral ($s < 1/(2N_e)$). So it has been argued that it would be unlikely for selection to affect synonymous mutations in species that have small populations[23] such as mammals, where $N_e << 10^6$ (REF. 21), but would still affect codon usage in species such as bacteria and flies. The nearly neutral theory correctly predicts there to be lower levels of selective constraint in small populations[5].

however this is not observed[7]. Why the polymorphism levels are relatively invariant remains unclear, but such polymorphism is likely to be due to selection at linked sites, the effect of which is to reduce variation in the vicinity of a gene that is under positive selection[8].

Another body of evidence against neutrality comes from examining rates of protein evolution. According to the neutral theory, the number of mutations that become fixed within a population should be Poisson distributed with a mean μT, where T is the number of generations and μ is the mutation rate per sequence per generation. This makes two predictions. First, species with short generation times should have faster evolving proteins than those with long generation times. However, this is typically not so[9] and if a molecular clock is defined by rates of protein change it ticks per unit time, not per generation. Second, being Poisson distributed, the mean and variance in the number of substitutions should be equal. However, in general this is not observed[9]. For example, for non-synonymous (protein-changing) mutations in mammals, Ohta[10] estimated that the ratio of the variance to the mean is greater than five (see also REF. 11). Recent evidence[12] supports the suggestion that this might be due to episodic positive selection[13].

Perhaps it was unsurprising that protein evolution is not simply neutral. More surprising, however, were investigations of synonymous codon usage. As synonymous nucleotide changes do not alter the encoded amino acid, neutralists argued that they must be invisible to selection[14,15]. Although selectionists noted that, at least in theory, this need not necessarily be true[16], it was not until the early 1980s that evidence emerged for why selection should act at synonymous sites. Studies of some bacteria, plants, yeast, flies and worms have revealed that, especially in highly expressed genes, usage of synonymous codons is biased to maximize the rate of protein synthesis by matching skews in tRNA abundances[17–20].

Synonymous mutations in mammals are commonly assumed to be neutral. The above organisms all have large populations, so weakly deleterious mutations can be efficiently acted on by natural selection (BOX 1). However, when populations are small, as in mammals[21] or in species that are isolated on islands[22] the same mutations can be 'effectively neutral' (BOX 1). Therefore, synonymous sites in mammals have long been considered to be neutrally evolving[23].

Support for the idea that synonymous mutations in mammals are different is also based on the finding that the dominant factor dictating codon usage in mammals is the isochore effect[4,23,24]. Isochores are large (>300 kb) domains of relatively homogenous GC content[25]. For a given gene, by far the best predictor of nucleotide content at synonymous sites (FIG. 1a) and codon-usage bias (FIG. 1b) is the nucleotide content of the isochore (the flanking non-coding DNA)[26]. This strongly supports the view that the main force that operates on synonymous mutations in mammals is not selection that is specific to genes or exons.

The underlying cause of isochoric structure remains uncertain[26], but recent evidence[27–29] indicates that this is not simply a neutral process. The best current hypothesis (for an alternative see REFS 30,31) proposes that there is a mutation bias in favour of A and T, and a fixation bias whereby G and C frequency is increased through biased gene conversion, functioning either between sister chromosomes during meiotic recombination[32,33] or between tandem repeats in mitosis[34]. As a consequence, regions of the genome that have consistently high recombination rates tend to oppose GC>AT mutations, and therefore become GC-rich, whereas those that have low recombination rates have GC content that is closer to the AT-rich, mutationally driven equilibrium.

Are isochore effects alone adequate to explain synonymous codon usage in mammals? First, we address

Positive selection
Also known as Darwinian selection. Natural selection that promotes the spread of a new mutation through the population, resulting in a fixed difference between species.

Molecular clock
A model of sequence evolution in which the number of changes that occur between two lineages accumulate at a constant rate, therefore allowing the estimation of the time since lineage divergence from the number of changes that have occurred.

Biased gene conversion
Gene conversion is a process by which similar genomic fragments become identical. If, after the DNA-repair system recognizes GC:AT mismatches in a heteroduplex (for example, arising during recombination between paired sister chromosomes), mismatches are resolved in favour of certain bases, the process is considered to be biased. Typically, biased gene conversion favours GC over AT in GC:AT mismatches.

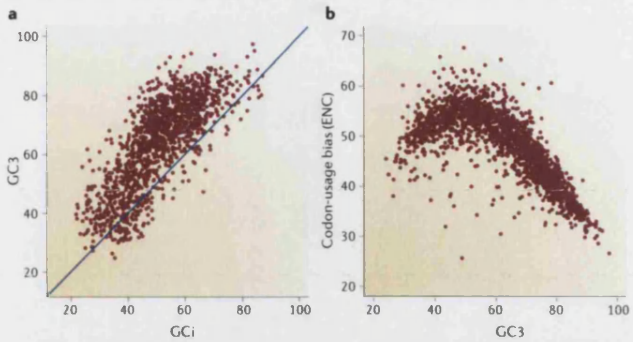


Figure 1 | **The effect of isochores on synonymous codon usage and codon-usage bias.** Mammalian genomes consist of relatively homogenous domains of GC content (>300 kb in size)[25,26]. Within these isochores, base composition of intergenic spacers, introns and coding sequence are all highly correlated. For example, panel **a** shows the correlation between GC content in introns (GCi) and at third sites of codons (GC3) of the same gene. The strength of the relationship indicates that whatever has driven the isochore effect is the dominant force that dictates nucleotide content at third (mainly synonymous) sites and so codon usage[4,23,24]. The plot shows 1,380 human genes (R^2 = 0.60; P < 0.0001), and the line indicates equality. Additionally, however, GC3 is consistently higher than GCi, particularly in GC-rich isochores. It has been suggested that, at synonymous sites, selection favours high G and C, but the lower GC content in introns can, at least in part, be explained by the presence of AT-rich transposable elements[43] (but see REF. 44). The isochore effect in skewing GC content at synonymous sites also has an effect on codon-usage bias. Panel **b** shows the correlation between GC content at third sites and codon-usage bias, which is measured by the effective number of codons (ENC) (a stronger bias is indicated by low ENC values). Codon bias is greatest (ENC is lowest) when GC content is most skewed away from equal usage of G and C compared with A and T. The same form of plot is found if intronic GC content or flanking GC content is used instead of GC content at third sites. This indicates that codon-usage bias is strongly determined not by exon-specific forces but by background isochore effects. This isochore effect underpins the need to correct for background nucleotide content when attempting to detect systematic codon-usage bias (translational selection) in mammals. The plot shows 2,030 human genes. The data for both panels are derived from REF. 96.

what might be considered indirect tests as they look for deviations from neutral expectations, while not necessarily specifying a mechanistic basis for the activity of selection. Following this, we review more recent lines of evidence, which we regard as direct evidence, in which specific mechanistic models of the cause of fitness effects of synonymous mutations are examined.

Indirect evidence for selection

Comparing base composition between synonymous sites within the same gene. Iida and Akashi[35] proposed that, because constitutively expressed exons are translated more frequently than alternative exons, a difference in nucleotide content would indicate selection for the use of optimal codons in constitutive exons (see below). They found that both GC3 (GC content at the mostly synonymous third sites of codons) and the rate of synonymous evolution are higher in human exons that are expressed constitutively[35] (see also REF. 36). More generally, intragenic heterogeneity in synonymous evolution seems to be common[37].

Expression breadth
The proportion of tissues in which a given gene is expressed.

Expression rate
The average level of gene expression across all tissues in which a given gene is expressed.

An alternative to comparing constitutive and alternative exons is to assay codon bias within a gene, in a manner that attempts to correct for potential isochore effects[38,39]. For example, Urrutia and Hurst[39] extended a previous method that measures the expected codon usage for each set of synonymous codons, on the basis of the within-gene usage in all the other synonymous sets that have the same level of degeneracy. They found that, although isochoric effects do explain much of the biased codon usage (as expected), they could not explain all of the skew. After correcting for the relationship between codon bias and gene length, the observed codon usage is not associated with expression breadth[39] but, consistent with selection, is correlated with expression rate[40].

Comparing base composition at synonymous sites with flanking introns. The observation that GC content at synonymous sites is greater than GC in the flanking introns (FIG. 1a), at least for relatively GC-rich regions, could indicate selection at synonymous sites[27,41], not least because the effect might be most pronounced in highly expressed genes, notably histones[42]. However, histones typically occur in tandem arrays, and biased gene conversion between genes, restricted to the exons, can at least in part account for their high GC content[34]. Moreover, the higher GC3 in most exons can at least in part result from the insertion of AT-rich transposable elements into introns within GC-rich isochores[43]. Although reduced, the difference still remains after masking transposable elements[43]. This remaining difference might be due to the presence of old elements, which would be hidden because transposable elements can only be identified if they have diverged <40% from their progenitor sequence[43]. Nonetheless, this is unlikely to be a complete explanation, as masking elements that have diverged up to 20% gives almost identical figures[44].

Comparing evolutionary rates at synonymous sites with pseudogenes. If synonymous sites are neutral, they should evolve at the same rate as other putatively neutral sequences. The earliest tests found that the rate of nucleotide substitution at synonymous sites is much lower than in pseudogenes[45]. Bustamante *et al.*[46] later estimated evolution at synonymous sites to be 70% of that in pseudogenes. Unfortunately, however, such analyses suffer from at least two confounding factors that render interpretation difficult. First, only transcribed genes will experience biases that are associated with transcriptionally coupled mutation and repair[47,48]. Second, substitution rates vary within the genome[3,49,50], such that related pseudogenes in different locations also evolve at different rates[51,52]. It remains unclear whether either of these factors fully account for the 30% difference between synonymous sites and pseudogenes[46].

Comparing evolutionary rates at synonymous sites with flanking introns. Carrying out within-gene analyses[35], such as comparing synonymous substitution rates (K_s) with flanking intronic substitution rates (K_i), avoids the problems of the regional variation in substitution rates and transcription-associated biases.

However, not all intronic sequence evolves neutrally. Both first introns and sequences that are near intron–exon junctions are conserved by selection[53–55]. Although these are relatively easy to exclude, it is hard to define *a priori* those functional regions that are towards the interior of introns. Consequently, comparing intron evolution with flanking synonymous sites might not prove to be definitive. Moreover, the hypermutability of CpGs and their differing densities in introns and exons[55] renders comparisons even more problematic. Attempts to exclude CpGs come to different conclusions[56,57], which might be related to difficulties in identifying sites that are prone to hypermutation[54]. Furthermore, in the human–chimpanzee comparison, differences between rates might be obscured by low divergence, whereas mouse–rat analyses suffer from problems of intron alignment. Given these difficulties, perhaps it is unsurprising that every possible result has been obtained. Various studies claim that $K_i < K_s$ (REFS 57,58), others that $K_i = K_s$ (REFS 41,55,59) and others still that $K_i > K_s$ (REFS 56,60). Although it has been suggested that an increased sample size resolves disagreements[61], the discrepancy probably reflects methodological differences. Some researchers suggest that, as K_s is so much lower than K_i, 40% of synonymous mutations have been opposed by selection[56].

Altogether, these studies indicate that evolutionary rates alone do not tell us the whole story. Closer analysis is more informative. Notably, even if the overall rates are similar[41,55,59], the patterns of nucleotide substitution at synonymous sites and in introns are quite different[54,55]. For example, C residues are both more common at fourfold degenerate (synonymous) sites than in introns, and also are relatively less likely to be associated with a substitution, after controlling for relative abundance[55] (see also REF. 54). This indicates that the action of selection that is particular to silent changes in exons cannot be accounted for by isochore effects alone. Furthermore, it has been claimed that a reduced rate of synonymous evolution ($K_i > K_s$) is most pronounced on the X chromosome[62], on which purifying selection is more efficient owing to hemizygous expression in males. The unusually low rate of synonymous evolution in imprinted genes[4] is also then expected.

Direct tests of specific models of selection

The above evidence, although sometimes contradictory, is nonetheless indicative of a role for selection. However, an understandable reluctance to accept selection at synonymous sites in mammals must remain until any putative effect is allied with a plausible model.

Maximized translational efficiency. For any given set of synonymous codons, the relevant iso-acceptor tRNAs might not be equally abundant. Consequently, if tRNA abundances are skewed and selection favours rapid translation, there might be a pressure to use the codon that matches the most abundant tRNA. This model predicts that for any given amino acid there is a 'best' (optimal) codon, which is defined by the skew in tRNA usage, and so there must also be a preferred set of codons if translation rate is to be maximized. Use of codons that are specified by rare tRNAs might also be a selectively favourable means to slow translation in genes that are expressed at a low level[63]; however, here the case is less clear as this class of genes is also expected to be under weaker selection. Co-evolution between non-random codon usage and skewed tRNA abundance is possible, leading to a positive-feedback loop that exaggerates codon bias and corresponding tRNA skews[64]. Another prediction is that the bias to favour preferred codons should be most pronounced in highly expressed genes and that experimentally adjusted codon usage should affect expression rates. As mentioned above, these patterns are seen in many organisms[17–20]. Consequently, translational selection is considered the dominant model and has become all but exclusively identified with systematic codon-usage bias. However, note that 30% of bacterial species show no evidence of such translational selection[65]. This might reflect low effective population sizes, but might also be due to an absence of selection for fast growth[65].

Some data support a weak relationship between gene expression and codon usage in mammals[40,63,66]. For example, the lower GC content of alternative exons[35] has been proposed as support for translational selection. However, that certain classes of alternatively spliced exons have low flanking intronic evolution[36,67] indicates that differences between constitutive and alternative exons might also reflect variation in the density and composition of splicing control elements (see below).

As mentioned above, highly expressed genes show the strongest codon bias[40]. However, correlating bias and expression fails to directly associate codon usage with tRNA abundance (which is reliably assayed by the copy number of tRNA genes[19]). Results of such analyses are contradictory.

Kanaya *et al.*[68] did not find evidence for skews in putative tRNA genes, whereas Lander *et al.*[69] found "only a very rough correlation of human tRNA gene number with either amino-acid frequency or codon bias". Duret[19] interpreted these results as having no detectable relationship. Similarly, dos Reis *et al.*[70] developed a measure of translational selection, *S*, which is the extent to which tRNA copy-number and codon usage are co-adapted across genomes. They found that organisms in which selectively driven codon-usage bias has previously been described (for example, *Escherichia coli*, *Saccharomyces cerevisiae* and *Caenorhabditis elegans*) have high *S*-values ($S > 0.45$), whereas humans possessed low values ($S = 0.03$), indicating that selection does not maximize translational efficiency in mammals.

Conversely, two recent studies have found a correlation between tRNA skews and codon usage in humans. Comeron[66], using the data from Lander *et al.*[69], reports that tRNA copy-number matched his proposed set of preferred codons for 14 out of 17 amino acids. Likewise, Lavner *et al.*[63] show that iso-accepting tRNA numbers positively correlate with expression-weighted frequencies of both amino acids and codons.

Does this mean that adjusting codon usage can modify the rate of translation in mammals, as it does in *Drosophila melanogaster*, for example[71]? Numerous studies

Synonymous substitution rate (K_s)
The ratio of the number of synonymous differences (corrected for multiple hits) between two orthologous genes to the number of sites in the gene at which synonymous mutations could occur.

Intronic substitution rate (K_i)
The number of differences per site (corrected for multiple hits) between orthologous introns.

Purifying selection
Also known as negative selection. Selection that eliminates a new mutation from the population, therefore removing changes from the population and maintaining the *status quo*.

Iso-acceptor tRNA
Any tRNA molecule that is charged by the single aminoacyl-tRNA synthetase which is specific to a given amino acid. The entire complement of tRNAs is divided into 20 iso-accepting groups, with each group being associated with a particular synthetase.

have demonstrated that modified codon choice can affect net expression levels. For example, early attempts to express jellyfish GFP in human cell lines were more successful after codon usage was adjusted[72,73] (see also REF 74).

However, even if in principle translational efficiency can be experimentally maximized by adjusting numerous sites within a gene, it is inappropriate to extrapolate this to supposing that a single synonymous mutation must be under selection, as any given single mutation is unlikely to have a substantial effect on translation rates. Moreover, these experimental results do not always directly show that it is translation rate that modulates any effect. For example, the transcript must be efficiently transcribed, have the introns successfully removed and the resulting mRNA must be stable enough to be exported and successfully dock with a ribosome for translation. All these stages might be sensitive to codon choice. However, in the first possibility support for a relationship between transcript levels and GC content at silent sites is currently weak[75] and contentious[76] (but see REF. 77). Evidence for involvement in mRNA stability and splicing is stronger.

Optimized mRNA stability. If a stable mRNA secondary structure confers resistance to premature degradation, selection might oppose synonymous mutations that disrupt base pairing[78]. Under this hypothesis, a transcript folds into the optimal conformation given the available sequence, which will for the most part be dictated by protein-coding requirements (note that highly conserved stem-loop sub-structures, as seen in tRNAs, for example, are probably unlikely in mammalian mRNAs[79]). Several cases have highlighted the significance of synonymous mutations that affect mRNA secondary structure[80–82], which in some cases are associated with disease[81,82]. Moreover, this model would be consistent with clustering of substitutions within genes[83].

Determining whether synonymous mutations might generally affect fitness, mediated by effects on mRNA folding, is difficult because structures cannot be observed directly. However, some studies have investigated the importance of synonymous sites on computationally predicted mRNA structure and stability in various organisms (for examples see REFS 84,85). As even *in vitro* foldings might not reflect those that are formed *in vivo*[86], it is likely that structures that are predicted *in silico* feature an even larger error component[78]. Nonetheless, recent *in silico* tests in the mouse indicate that selection does occur at synonymous sites[78]. One particularly intriguing result is that, as previously described in histone genes[87], there is a skew towards G at the first two sites within codons. This can therefore potentially explain the C preference at fourfold sites[55], as strong G:C pairs create stable mRNAs. Consistent with this are the findings that the stability of wild-type mRNAs relative to artificial transcripts is highest when there is a strong third-site skew towards C, and mRNAs are also less stable when Gs and Cs are interchanged[78]. Moreover, had the synonymous mutations observed in the mouse lineage occurred elsewhere within genes, transcripts would have been less stable[78]. Secondary structure therefore provides a possible explanation for C being in excess at third sites.

Transcript stability can also arise from preferring or avoiding particular sequence motifs. Notably, introducing synonymous substitutions that increase C|G dinucleotide content (where | is the codon boundary) decreases the rate of degradation, whereas increasing U|A enhances transcript decay[88]. This avoidance of UA dinucleotides[39,88] might prevent recognition by proteins that cleave AU-rich elements[88]. This provides another potential explanation for the C preference at third sites.

Efficient splicing control. Most of the recent evidence indicates that synonymous mutations can be under selection because they upset intron removal. There are abundant examples of synonymous mutations that cause disease by disrupting the splicing process[89,90] (TABLE 1). Nonetheless, such disease-associated mutations are probably much rarer than non-synonymous changes that are associated with disease, indicating that only a small fraction of synonymous mutations might have a significant effect on splicing. Disease-associated synonymous mutations might create new 'cryptic' splice sites[91] or affect splicing-control elements, such as exonic splicing enhancers (ESEs)[92] and silencers (ESSs)[93]. Splicing modulators are oligomeric motifs that recruit spliceosomal proteins to facilitate splice-site recognition[93]. These tend to be purine-rich[94] and so are unlikely to explain the C excess, or its potential association with translation[66] or mRNA stability[78].

Importantly, exonic splicing modulators tend to reside near intron–exon junctions. Much recent evidence has documented the aspects in which the ends of exons are unusual. For example, the codon GAA is common in ESEs and is increasingly preferred over its synonym GAG towards the intron–exon junction[95] (FIG. 2). However, a preference for ESEs, although a robust model, might not explain all the observed gradients in nucleotide content across exons[96,97]. Alternatively, such biases might reflect an avoidance of codons that contain potentially cryptic splice sites[91] — those dinucleotides that could be inappropriately identified as intronic ends. However, if this pressure exists it seems to be much weaker than a preference for ESEs[96].

Consistent with gradients of biased codon choice, some genes show a marked reduction in the rate of synonymous evolution in regions that contain an ESE — for example, breast cancer 1, early onset (*BRCA1*) (REFS 98,99) (FIG. 3) and cystic fibrosis transmembrane-conductance regulator ATP-binding cassette subfamily C member 7 (*CFTR*)[100]. More generally, SNP density decreases towards the ends of exons[51], which could be explained by increasing ESE density[101]. Moreover, consistent with purifying selection on ESEs, SNP frequency is lower at synonymous sites in putative ESE hexamers than in non-exonic sequences[102]. Similarly, synonymous evolution in putative ESEs is slower than in non-ESE sequences, which explains the reduced synonymous substitution rate near exon ends[97]. Selection on exonic splicing modulators might even be

Table 1 | **Synonymous mutations that are associated with aberrant splicing, which lead to human diseases**

Gene	Mutation	Exon	Mechanism	Disease	References
ALG3	G55G	1	ESE activates upstream cryptic SS?	Congenital disorder of glycosylation type Id	118
APC	R623R H652H; R653R	14	ESE disrupted?	Familial adenomatous polyposis	89 119
AR	S888S	8	5′ SS created	Androgen-insensitivity syndrome	89
ATM	S706S S1135S	16 26	5′ SS disrupted	Ataxia telangiectasia	89
ATR	G677G	9	mRNA structure?	Seckel syndrome	120
CYBB	A84A	3	5′ SS disrupted	Chronic granulomatous disease	121
CYP27A1	G112G	2	5′ SS created	Cerebrotendinous xanthomatosis	89
FAH	N232N	8	Unknown	Hereditary tyrosinaemia type 1	89
FBN1	I2118I	51	Unknown	Marfan syndrome	89
GLDC	P869P	22	ESE?	Glycine encephalopathy	122
HBA2	G22G	1	5′ SS created	Unknown α-thalassaemia disease	123
HEXA	L187L V324V	5 8	5′ SS disrupted 5′ SS created	Tay–Sachs disease G_{M2} gangliosidosis	89 124
HMBS	R28R	3	ESE disrupted?	Acute intermittent porphyria	89
HPRT1	F199F	8	Unknown	Lesch–Nyhan syndrome?	89
ITGB3	T420T G605G	9 11	mRNA structure? 5′ SS created	Glanzmann thrombasthenia	89 125
LAMB3	H1003H	20	5′ SS created	Junctional epidermolysis bullosa	126
L1CAM	G308G	8	5′ SS created	X-linked hydrocephalus	127
LIPA	Q277Q	8	Unknown	Cholesteryl ester storage disease	89
MAPT	L284L N296N S305S	10	ESE or ESS disrupted ESS disrupted 5′ SS disrupted	Frontotemporal dementia with Parkinsonism — chromosome 17 type Familial dementia with swollen achromatic neurons and corticobasal inclusion bodies Supranuclear palsy	89
MLH1	S577S	16	Unknown	Hereditary non-polyposis colorectal cancer	89
NF1	K354K	7	5′ SS disrupted	Neurofibromatosis type 1	89
OPA1	R590R	18	Unknown	Autosomal dominant optic atrophy	128
PAH	V399V	11	ESE disrupted?	Phenylketonuria	89
PDHA1	G185G	6	ESE disrupted	X-linked Leigh syndrome	89
PKLR	A423A	9	Unknown	Pyruvate kinase deficiency	89
PTPRC	P48P	4	Unknown	Multiple sclerosis	89
PTS	E81E	4	5′ SS disrupted	PTPS (6-pyrovoyltetrahydropterin synthase) deficiency	89
PYGM	K608K	15	Unknown	McArdle disease	129
RET	I647I	11	ESE?	Hirschsprung disease	89
SMN1	F280F	7	ESE disrupted	Spinal muscular atrophy	89
TGFBR2	Q508Q	6	5′ SS disrupted	Marfan syndrome	130
TNFRSF5	T136T	5	ESE disrupted	Immunodeficiency with hyper IgM	89
UROD	E314E	9	5′ SS disrupted	Familial porphyria cutanea tarda	89

REF. 89 provides a similar table. For those incidences in the present table that are cited as being from REF. 89 the full citation details can be found by reference to this paper. ESE, exonic splicing enhancer; SS, splice site.

more important than the encoded protein[90]. Consistent with this idea, splicing can also affect non-synonymous evolution[101] and amino-acid usage[97]. However, selection on ESEs seems not to explain the reduced synonymous rate of evolution in alternatively spliced exons[97].

Implications

The above evidence fails to support the assumption that all synonymous sites in mammalian exons are neutrally evolving. Although it remains probable that most synonymous mutations are neutral (or effectively

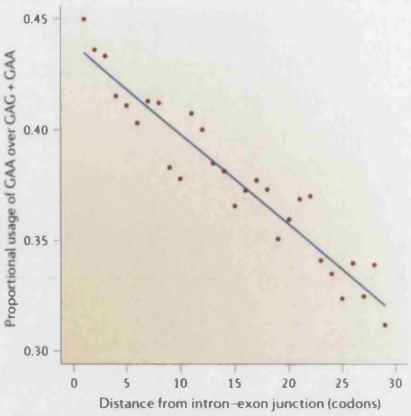


Figure 2 | **Usage of certain codons is more biased near intron–exon junctions, owing to synonyms being differentially common in exonic splicing enhancers.** The example depicted shows the proportional usage of the codon GAA versus its synonym GAG, as a function of the distance from intron–exon junctions[95]. The trend is mostly explained by high exonic splicing enhancer (ESE) density near exon ends[101] and the prevalence of GAA in ESEs[95]. The action of purifying selection on synonymous mutations that affect splicing is supported by decreased SNP density[53] and substitution rates[97] in close proximity to intron–exon junctions. The plot combines data from both the 5′ and 3′ ends of 14,407 human exons in 1,802 genes ($R^2 = 0.88$; $P < 0.0001$). The line of best fit was derived by regression and weighted by the total number of codons compared at each position. Data are from REF. 96.

neutral), the finding that selection does operate on a significant proportion, possibly up to 40% (REF. 56), has important implications. First and foremost, given the evidence for the involvement of synonymous sites in disease, especially when mediated by splicing defects (TABLE 1), the assumption of a lack of phenotype caused by synonymous mutations, like the assumption of neutrality, can no longer be sustained.

Instead of the neutral model, we should be considering synonymous mutations in the framework of the nearly neutral model (BOX 1). In retrospect, the assumption that synonymous mutations must all be neutral because they do not affect protein sequence[14,15] probably reflects the earlier incomplete understanding of the pathway from gene to protein. Indeed, we might still be missing important constraints. For example, it is possible that microRNAs that bind to sense mRNA as a mode of gene regulation might impose constraint on sites in the mRNA to ensure efficient pairing. Synonymous sites might also be under seletion to enable efficient RNA editing[131]. Furthermore, synonymous mutations can affect protein folding. For example, in *E. coli* the use of rare codons can induce translational pauses[104] that allow a newly synthesized polypeptide strand enough time to fold into the correct secondary structure[105]. Suggestively, stretches of rare codons correspond to turns, loops and links between protein domains[106,107]. Preventing co-translational misfolding might be even more important in eukaryotes[108] and could explain the preference for GAT over GAC at the N termini of α-helices in humans[107]. We also do not yet fully understand why genes that are expressed uniquely in a given tissue have a GC content that is prototypical for genes that are expressed in that tissue[31]. Note that claims that the GC content of tissue-specific genes is independent of isochore effects[109] are not robust[132].

Detecting positive selection. One leading use for K_s is as a background evolutionary rate to detect positive selection[110]. If selection favours adaptive non-synonymous changes, the protein should evolve faster than expected under neutral evolution. To this end, the number of non-synonymous substitutions per non-synonymous substitution rate (K_a) is compared with K_s. If $K_a > K_s$ then positive selection is inferred; that is, $K_a/K_s > 1$.

A very low K_s that is due to purifying selection on synonymous sites could, in principle, also give rise to $K_a/K_s > 1$ (REFS 37,98). This possibility is usually not even considered. However, a few examples have recently been given for intragenic dips in synonymous evolution, which are probably associated with splicing regulation[98–100] (FIG. 3). Are these simply oddities or is it the case that an intragenic $K_a/K_s > 1$ often reflects low K_s rather than high K_a? To assess this we examined long (>3,000 nucleotides) mouse–rat orthologues and constructed sliding-window plots across alignments to search for $K_a/K_s > 1$ peaks. Such peaks are relatively rare, occurring in only 15 of 143 genes. Of the 15, only 11 could be best interpreted as peaks owing to very high K_a with normal K_s or *vice versa*. The striking conclusion is that 6 could be classified as K_s peaks and 5 as K_s dips (L.D.H., unpublished observations). This indicates that the K_a/K_s ratio, applied within genes, is not a safe way to identify positive selection, unless purifying selection on synonymous sites can be discounted. In principle, this might be achieved by examining synonymous evolution in a region that has a high K_a/K_s peak to see whether the synonymous rate is unusually low (see also REF. 37).

Underestimating the mutation rate. If synonymous evolution in mammals is not neutral and K_s is used as a measure of the mutation rate, by how much might we be underestimating the true mutation rate? Is it possible to quantify non-neutral effects and so still use K_s after adjusting for the contribution of selection?

Lu and Wu[62] estimated the proportion of synonymous mutations that are deleterious by comparing rates of evolution between introns and synonymous sites on the X chromosome and the autosome. Remarkably, they estimated that >90% of synonymous mutations are under weak selection. However, for the most part, the selection is so weak that it has a negligible effect on substitution rates. Whether this quantitatively agrees with the 30% lower divergence at synonymous sites compared with pseudogenes[46] or the 40% reduction compared with non-coding DNA[56] is unclear.

MicroRNAs
Short non-coding RNAs (~22 nucleotides long) that can repress gene expression by base pairing to target mRNAs.

Non-synonymous substitution rate (K_a)
The ratio of the number of non-synonymous differences (corrected for multiple substitutions at the same site) between two orthologous genes to the number of sites at which non-synonymous mutations could occur.

Sliding-window plot
A graphical representation of a sequence in which subsections, sometimes overlapping, of a given size (a window) are successively analysed.

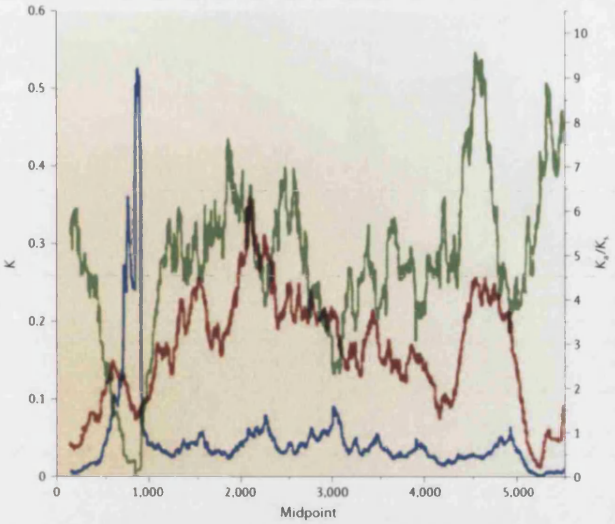


Figure 3 | **Fluctuation in rates of evolution across the *BRCA1* gene.** The sliding-window plot compares sequences between human and dog orthologues. The x-axis shows the midpoint in base pairs of the 306-nucleotide window. The y-axis shows, on the left, the rate of non-synonymous substitution (K_a, red), the rate of synonymous evolution (K_s, green) and, on the right, shows the K_a/K_s ratio (blue). Note that the very high K_a/K_s peak that is near the 5′ end of the gene is associated with a marked dip in K_s rather than a peak in K_a. Such K_s dips might represent half of all K_a/K_s peaks (see main text), and significant heterogeneity in synonymous evolution across genes seems to be common[37]. Consequently, some proteins and peptide regions are erroneously identified as being under positive selection.

An alternative approach is to examine each model individually. However, if the reduction in K_s that is associated with each model were to be quantified, the relative contributions of each need not be additive. In flies and yeast there are trade-offs between codon bias for translational efficiency and mRNA secondary-structure requirements[111,112]. This caveat aside, given the proportion of exons that specify putative splicing enhancers and the extent to which their rate of evolution is slower than non-ESE sequence, the mutation rate seems to have been underestimated by no more than about 10% (REF. 97), although in one well-characterized example[100], about 30% of synonymous mutations in a given exon are associated with mis-splicing. A similar quantitative assessment has yet to be carried out for other modes of selection, although their effects are probably weak. With selection at synonymous sites for mRNA stability, only a minority of genes show strong evidence of selection[78] and it probably affects only specific sites. Likewise, codon bias for translational efficiency in mammals, if present, is only detectable in the most highly expressed genes[40]. This indicates that mutation-rate estimates are unlikely to increase substantially. Ambiguity about the number of generations that separate taxa, owing to uncertainty about generation times and time since common ancestry, would potentially force adjustments of a much higher order. For example, the time since the mouse and rat shared a common ancestor might be anywhere between 5 and 42 million years (for discussion see REF. 113).

In short, it is unlikely that the assumption of neutrality of synonymous mutations has grossly misled us in estimates of the genomic mutation rate. Perhaps this is unsurprising, given that some signals of selection, which are seen in species that have a large N_e using a handful of genes[114], have only been detected in mammals through the use of large data sets[63], although this is not universally true[78]. An upwards correction to the mutation rate will have a greater effect on the estimated number of new deleterious mutations per genome per generation, as we must now allow for some proportion of synonymous mutations to be deleterious. However, the extent to which these impinge on fitness will depend on whether there is interaction between mutations. For example, Akashi[115] argued that individual synonymous mutations might have a small effect on fitness, but that they might show a cumulative effect through synergistic epistasis (which would also apply to non-coding DNA[52,116]). This provides a potentially important explanation to account for the fact that synonymous SNPs are both relatively common and potentially deleterious.

The conclusion that our estimates of the mutation rate are not greatly misleading comes, however, with a strong proviso. Above we asked about selection that might be peculiar to synonymous mutations. However, apart from the presence of functional residues, there might be reason to suppose that substitution rates at all silent sites (intronic, intergenic and synonymous) could be misleading. Notably, biased gene conversion will affect substitution rates of all forms of silent DNA[117]. As this process accelerates the fixation of AT>GC mutations and diminishes the rate of fixation of GC>AT mutations, regardless of their coding status, the net rate of evolution will not be equal to the mutation rate, even if the mutations would otherwise be neutral. If the effect is profound, then mutation rates cannot safely be extracted from any sequence comparison.

Optimizing transgene expression. Understanding the mode of action of selection on synonymous mutations should allow us to improve transgenes without altering the encoded protein. Although transgene expression is often more efficient when constructs retain the first intron (as these contain regulatory elements), the other introns tend to be dispensable (for citations see REF. 55). In principle, as codon choice near intron–exon junctions is biased to allow efficient splicing[95,96], synonymous sites near junctions could be modified with potentially beneficial effects for transgenes that lack non-first introns. As ESEs tend to be A-rich and third sites of codons might be C-rich for mRNA stability[78], swapping A for C at synonymous sites might well decrease transcript-decay rates. Moreover, a high GC content might also be compatible with the proposed set of preferred codons[66] and will minimize deleterious UA usage[88]. We can foresee that this procedure for transgene optimization could be incorporated into a sophisticated *in silico* tool.

Synergistic epistasis
The interaction between mutations that causes their combined effect on fitness to be greater than would be expected from their individual (multiplicative) effects.

Transgene
Foreign DNA that is experimentally inserted into totipotent embryonic cells or into unicellular organisms.

Conclusion

The idea that synonymous mutations must all be neutral, as they have no effect on the encoded protein, might at first seem both seductive and intuitive. However, the recently discovered knowledge of what really determines the fate of synonymous mutations in mammals has brought to our attention the unexpected strength of natural selection and a plethora of previously unrecognized selective forces. Although many synonymous mutations are no doubt free from selection, the assumption that they all are neutral no longer seems safe. Acknowledging the various mechanisms will be important for understanding and potentially combating genetic disease. Importantly, understanding how synonymous codon choice makes for efficient expression of a gene will aid in the engineering of better transgenes.

1. Kreitman, M. The neutral theory is dead — long live the neutral theory. *Bioessays* **18**, 678–683 (1996).
2. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
3. Wolfe, K. H., Sharp, P. M. & Li, W. H. Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283–285 (1989).
4. Smith, N. G. C. & Hurst, L. D. The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? *Genetics* **152**, 661–673 (1999).
5. Keightley, P. D. & Eyre-Walker, A. Deleterious mutations and the evolution of sex. *Science* **290**, 331–333 (2000).
6. Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A. & Kondrashov, A. S. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17**, 373–376 (2001).
7. Lewontin, R. C. *The Genetic Basis of Evolutionary Change* (Columbia Univ. Press, New York, 1974).
8. Gillespie, J. H. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* **155**, 909–919 (2000).
9. Ohta, T. & Gillespie, J. H. Development of neutral and nearly neutral theories. *Theor. Pop. Biol.* **49**, 128–142 (1996).
10. Ohta, T. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* **40**, 56–63 (1995).
11. Nielsen, R. Robustness of the estimator of the index of dispersion for DNA sequences. *Mol. Phyl. Evol.* **7**, 346–351 (1997).
12. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. Erratic overdispersion of three molecular clocks: GPDH, SOD, and XDH. *Proc. Natl Acad. Sci. USA* **98**, 11405–11410 (2001).
13. Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford, 1991).
14. King, J. L. & Jukes, T. H. Non-Darwinian evolution. *Science* **164**, 788–798 (1969).
15. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276 (1977).
16. Clarke, B. Darwinian evolution of proteins. *Science* **168**, 1009–1011 (1970).
17. Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 13–34 (1985).
A review of how skews in tRNA abundance correspond to biased codon usage in *E. coli* and *S. cerevisiae*. This relationship showed that synonymous sites can be functional, with selection in this model promoting efficient protein synthesis.
18. Akashi, H. & Eyre-Walker, A. Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8**, 688–693 (1998).
19. Duret, L. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**, 640–649 (2002).
An excellent review of evolution at synonymous sites in *D. melanogaster*, *C. elegans* and vertebrates, with particular emphasis on the effects that make it difficult to disentangle the neutral and selective forces that impinge on codon usage, particularly in mammals.
20. Wright, S. I., Yau, C. B., Looseley, M. & Meyers, B. C. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Mol. Biol. Evol.* **21**, 1719–1726 (2004).
21. Keightley, P. D., Lercher, M. J. & Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3**, e42 (2005).
22. Woolfit, M. & Bromham, L. Population size and molecular evolution on islands. *Proc. Biol. Sci.* **272**, 2277–2282 (2005).
23. Sharp, P. M., Averof, M., Lloyd, A. T., Matassi, G. & Peden, J. F. DNA sequence evolution: the sounds of silence. *Philos. Trans. R. Soc. Lond. B* **349**, 241–247 (1995).
24. Eyre-Walker, A. An analysis of codon usage in mammals: selection or mutation bias? *J. Mol. Evol.* **33**, 442–449 (1991).
25. Bernardi, G. *et al.* The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958 (1985).
26. Eyre-Walker, A. & Hurst, L. D. The evolution of isochores. *Nature Rev. Genet.* **2**, 549–555 (2001).
27. Eyre-Walker, A. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* **152**, 675–683 (1999).
28. Lercher, M. J., Smith, N. G., Eyre-Walker, A. & Hurst, L. D. The evolution of isochores: evidence from SNP frequency distributions. *Genetics* **162**, 1805–1810 (2002).
29. Duret, L., Semon, M., Piganeau, G., Mouchiroud, D. & Galtier, N. Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162**, 1837–1847 (2002).
30. Vinogradov, A. E. Bendable genes of warm-blooded vertebrates. *Mol. Biol. Evol.* **18**, 2195–2200 (2001).
31. Vinogradov, A. E. Isochores and tissue-specificity. *Nucleic Acids Res.* **31**, 5212–5220 (2003).
32. Galtier, N., Piganeau, G., Mouchiroud, D. & Duret, L. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159**, 907–911 (2001).
33. Meunier, J. & Duret, L. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21**, 984–990 (2004).
34. Galtier, N. Gene conversion drives GC content evolution in mammalian histones. *Trends Genet.* **19**, 65–68 (2003).
35. Iida, K. & Akashi, H. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene* **261**, 93–105 (2000).
By comparing exons within the same gene, this paper provided strong evidence for selection at synonymous sites in humans, controlling for isochore effects, regional variation in rates of evolution and transcription-associated biases.
36. Xing, Y. & Lee, C. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl Acad. Sci. USA* **102**, 13526–13531 (2005).
37. Pond, S. K. & Muse, S. V. Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* **22**, 2375–2385 (2005).
Using a new method, the authors found evidence that significant heterogeneity in the synonymous substitution rate within mammalian genes is common. Consequently, certain sites are being erroneously identified as being under positive selection.
38. Karlin, S. & Mrazek, J. What drives codon choices in human genes? *J. Mol. Biol.* **262**, 459–472 (1996).
39. Urrutia, A. O. & Hurst, L. D. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* **159**, 1191–1199 (2001).
40. Urrutia, A. O. & Hurst, L. D. The signature of selection mediated by expression on human genes. *Genome Res.* **13**, 2260–2264 (2003).
The first report of a broad correlation between codon-usage bias (corrected for the isochore effect) and expression rate in human genes.
41. Hughes, A. L. & Yeager, M. Comparative evolutionary rates of introns and exons in murine rodents. *J. Mol. Evol.* **45**, 125–130 (1997).
42. DeBry, R. W. & Marzluff, W. F. Selection on silent sites in the rodent H3 histone gene family. *Genetics* **138**, 191–202 (1994).
43. Duret, L. & Hurst, L. D. The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. *Mol. Biol. Evol.* **18**, 757–762 (2001).
44. Vinogradov, A. E. Within-intron correlation with base composition of adjacent exons in different genomes. *Gene* **276**, 143–151 (2001).
45. Miyata, T. & Hayashida, H. Extraordinarily high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. *Proc. Natl Acad. Sci. USA* **78**, 5739–5743 (1981).
46. Bustamante, C. D., Nielsen, R. & Hartl, D. L. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Mol. Biol. Evol.* **19**, 110–117 (2002).
In mammals, the rate of nucleotide substitution at synonymous sites is 70% of that in the cognate pseudogenes.
47. Green, P. *et al.* Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genet.* **33**, 514–517 (2003).
48. Majewski, J. Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am. J. Hum. Genet.* **73**, 688–692 (2003).
49. Matassi, G., Sharp, P. M. & Gautier, C. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**, 786–791 (1999).
50. Lercher, M. J., Chamary, J. V. & Hurst, L. D. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res.* **14**, 1002–1013 (2004).
51. Casane, D., Boissinot, S., Chang, B. H., Shimmin, L. C. & Li, W. H. Mutation pattern variation among regions of the primate genome. *J. Mol. Evol.* **45**, 216–226 (1997).
52. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
53. Majewski, J. & Ott, J. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12**, 1827–1836 (2002).
54. Keightley, P. D. & Gaffney, D. J. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl Acad. Sci. USA* **100**, 13402–13406 (2003).
55. Chamary, J. V. & Hurst, L. D. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol. Biol. Evol.* **21**, 1014–1023 (2004).
Using mouse–rat alignments, the rate and patterns of evolution were compared between introns and synonymous sites, two classes of supposedly neutral sequence. Relative to intronic sites, C residues were found to be abundant and relatively stable at synonymous sites.
56. Hellmann, I. *et al.* Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13**, 831–837 (2003).
By comparing human–chimpanzee divergence at fourfold synonymous sites relative to the intergenic spacer, this study indicates that nearly 40% of (non-CpG) synonymous mutations have been eliminated by purifying selection.
57. Subramanian, S. & Kumar, S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* **13**, 838–844 (2003).
58. Chen, F. C. & Li, W. H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444–456 (2001).

59. Chen, F. C., Vallender, E. J., Wang, H., Tzeng, C. S. & Li, W. H. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J. Hered.* **92**, 481–489 (2001).
60. Smith, N. G. C. & Hurst, L. D. Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. *J. Mol. Evol.* **47**, 493–500 (1998).
61. Mikkelsen, T. S. *et al.* Initial sequence of the chimpanzee genome and comparison with the human genome. **437**, 69–87 (2005).
62. Lu, J. & Wu, C. I. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proc. Natl Acad. Sci. USA* **102**, 4063–4067 (2005).
Using human–chimpanzee alignments, the authors compared rates of evolution between autosomes and the X chromosome to measure the strength of selection at synonymous sites. They found that more than 90% of synonymous mutations are under weak selection, but suggest that, for the most part, selection seems to be too weak to influence substitution rates.
63. Lavner, Y. & Kotlar, D. Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene* **345**, 127–138 (2005).
64. Bulmer, M. Coevolution of codon usage and transfer RNA abundance. *Nature* **325**, 728–730 (1987).
65. Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F. & Sockett, R. E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* **33**, 1141–1153 (2005).
66. Comeron, J. M. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* **167**, 1293–1304 (2004).
References 63 and 66 provide evidence that human tRNA gene-copy numbers match a proposed set of preferred codons and correlate with expression-weighted frequencies of optimal codons.
67. Kaufmann, D., Kenner, O., Nurnberg, P., Vogel, W. & Bartelt, B. In *NF1, CFTR, PER3, CARS* and *SYT7*, alternatively included exons show higher conservation of surrounding intron sequences than constitutive exons. *Eur. J. Hum. Genet.* **12**, 139–149 (2004).
68. Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. & Ikemura, T. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.* **53**, 290–298 (2001).
Unlike yeast, flies and worms, codon usage in the genes that encode ribosomal proteins and histones is not significantly biased in humans, which indicates that the primary factor influencing codon-usage diversity in these species is not translation efficiency.
69. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
70. dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).
By measuring the extent to which tRNA copy-number and codon usage are co-adapted across genomes, the authors find no evidence for translational selection in humans.
71. Carlini, D. B. & Stephan, W. *In vivo* introduction of unpreferred synonymous codons into the *Drosophila Adh* gene results in reduced levels of ADH protein. *Genetics* **163**, 239–243 (2003).
72. Levy, J. P., Muldoon, R. R., Zolotukhin, S. & Link, C. J. Jr. Retroviral transfer and expression of a humanized, red-shifted green fluorescent protein gene into human tumor cells. *Nature Biotechnol.* **14**, 610–614 (1996).
73. Zolotukhin, S., Potter, M., Hauswirth, W. W., Guy, J. & Muzyczka, N. A 'humanized' green fluorescent protein cDNA adapted for high-level expression in mammalian cells. *J. Virol.* **70**, 4646–4654 (1996).
74. Kim, C. H., Oh, Y. & Lee, T. H. Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells. *Gene* **199**, 293–301 (1997).
75. Lercher, M. J., Urrutia, A. O., Pavlicek, A. & Hurst, L. D. A unification of mosaic structures in the human genome. *Hum. Mol. Genet.* **12**, 2411–2415 (2003).
76. Semon, M., Mouchiroud, D. & Duret, L. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Hum. Mol. Genet.* **14**, 421–427 (2005).
77. Vinogradov, A. E. Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. *Trends Genet.* **21**, 639–643 (2005).
78. Chamary, J. V. & Hurst, L. D. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* **6**, R75 (2005).
Provides evidence that the preference for C at synonymous sites (as shown in reference 55) could be explained by selection that favours thermodynamically stable mRNA secondary structures. Moreover, had synonymous substitutions occurred at locations other than those that were observed in the mouse lineage, the mRNA would have been less stable.
79. Buratti, E. & Baralle, F. E. Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol. Cell. Biol.* **24**, 10505–10514 (2004).
80. Shen, L. X., Basilion, J. P. & Stanton, V. P. Jr. Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc. Natl Acad. Sci. USA* **96**, 7871–7876 (1999).
81. Duan, J. *et al.* Synonymous mutations in the human dopamine receptor D2 (*DRD2*) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* **12**, 205–216 (2003).
A well-worked example of how a synonymous mutation can affect mRNA stability. Of six naturally occurring synonymous SNPs in the *DRD2* gene, only the mutation that decreases mRNA half-life and induced a conspicuous change in predicted secondary structure was linked to disease.
82. Capon, F. *et al.* A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Hum. Mol. Genet.* **13**, 2361–2368 (2004).
83. Smith, N. G. & Hurst, L. D. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153**, 1395–1402 (1999).
84. Seffens, W. & Digby, D. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.* **27**, 1578–1584 (1999).
85. Cohen, B. & Skiena, S. Natural selection and algorithmic design of mRNA. *J. Comp. Biol.* **10**, 419–432 (2003).
86. Schroeder, R., Barta, A. & Semrad, K. Strategies for RNA folding and assembly. *Nature Rev. Mol. Cell Biol.* **5**, 908–919 (2004).
87. Huynen, M. A., Konings, D. A. & Hogeweg, P. Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure. *J. Mol. Evol.* **34**, 280–291 (1992).
88. Duan, J. & Antezana, M. A. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. *J. Mol. Evol.* **57**, 694–701 (2003).
89. Cartegni, L., Chew, S. L. & Krainer, A. R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet.* **3**, 285–298 (2002).
90. Pagani, F. & Baralle, F. E. Genomic variants in exons and introns: identifying the splicing spoilers. *Nature Rev. Genet.* **5**, 389–396 (2004).
References 89 and 90 are excellent reviews of how exonic mutations can disrupt the pre-mRNA splicing process.
91. Eskesen, S. T., Eskesen, F. N. & Ruvinsky, A. Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. *Genetics* **167**, 543–550 (2004).
92. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1013 (2002).
93. Wang, Z. *et al.* Systematic identification and analysis of exonic splicing silencers. *Cell* **119**, 831–845 (2004).
94. Blencowe, B. J. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* **25**, 106–110 (2000).
95. Willie, E. & Majewski, J. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* **20**, 534–538 (2004).
The first demonstration that codons associated with splicing are increasingly preferred near intron–exon junctions.
96. Chamary, J. V. & Hurst, L. D. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? *Trends Genet.* **21**, 256–259 (2005).
97. Parmley, J. L., Chamary, J. V. & Hurst, L. D. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* 12 October 2005 (doi:10.1093/molbev/msj035).
98. Hurst, L. D. & Pal, C. Evidence for purifying selection acting on silent sites in *BRCA1*. *Trends Genet.* **17**, 62–65 (2001).
The first evidence from mammals that a $K_s/K_s > 1$ peak is due to a dip in the synonymous substitution rate, which reference 99 later revealed to coincide with the location of an ESE.
99. Orban, T. I. & Olah, E. Purifying selection on silent sites — a constraint from splicing regulation? *Trends Genet.* **17**, 252–253 (2001).
100. Pagani, F., Raponi, M. & Baralle, F. E. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl Acad. Sci. USA* **102**, 6368–6372 (2005).
About 30% of synonymous mutations in exon 12 of CFTR are associated with splicing disruption.
101. Fairbrother, W. G., Holste, D., Burge, C. B. & Sharp, P. A. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* **2**, e268 (2004).
As one approaches intron–exon junctions in humans, predicted ESE density increases while SNP density decreases. Additionally, the authors suggest that one-fifth of mutations that might potentially disrupt ESEs have been eliminated by purifying selection.
102. Carlini, D. B. & Genut, J. E. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Evol.* 30 November 2005 (doi:10.1007/s00239-005-0055-x).
103. Cusack, B. P. & Wolfe, K. H. Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. *Mol. Biol. Evol.* **22**, 2198–2208 (2005).
104. Purvis, I. J. *et al.* The efficiency of folding of some proteins is increased by controlled rates of translation *in vivo*. A hypothesis. *J. Mol. Biol.* **193**, 413–417 (1987).
105. Cortazzo, P. *et al.* Silent mutations affect *in vivo* protein folding in *Escherichia coli*. *Biochem. Biophys. Res. Comm.* **293**, 537–541 (2002).
106. Thanaraj, T. A. & Argos, P. Ribosome-mediated translational pause and protein domain organization. *Protein Sci.* **5**, 1594–1612 (1996).
107. Oresic, M. & Shalloway, D. Specific correlations between relative synonymous codon usage and protein secondary structure. *J. Mol. Biol.* **281**, 31–48 (1998).
108. Netzer, W. J. & Hartl, F. U. Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* **388**, 343–349 (1997).
109. Plotkin, J. B., Robins, H. & Levine, A. J. Tissue-specific codon usage and the expression of human genes. *Proc. Natl Acad. Sci. USA* **101**, 12588–12591 (2004).
110. Hurst, L. D. The *Ka*/*Ks* ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18**, 486 (2002).
111. Carlini, D. B., Chen, Y. & Stephan, W. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*. *Genetics* **159**, 623–633 (2001).
112. Carlini, D. B. Context-dependent codon bias and mRNA longevity in the yeast transcriptome. *Mol. Biol. Evol.* **22**, 1403–1411 (2005).
113. Adkins, R. M., Gelke, E. L., Rowe, D. & Honeycutt, R. L. Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. *Mol. Biol. Evol.* **18**, 777–791 (2001).
114. Grantham, R., Gautier, C. & Gouy, M. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res.* **8**, 1893–1912 (1980).
115. Akashi, H. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* **144**, 1297–1307 (1996).
116. Kryukov, G. V., Schmidt, S. & Sunyaev, S. Small fitness effect of mutations in highly conserved non-coding regions. *Hum. Mol. Genet.* **14**, 2221–2229 (2005).
117. Piganeau, G., Mouchiroud, D., Duret, L. & Gautier, C. Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. *J. Mol. Evol.* **54**, 129–133 (2002).

118. Denecke, J., Kranz, C., Kemming, D., Koch, H. G. & Marquardt, T. An activated 5′ cryptic splice site in the human *ALG3* gene generates a premature termination codon insensitive to nonsense-mediated mRNA decay in a new case of congenital disorder of glycosylation type Id (CDG-Id). *Hum. Mut.* **23**, 477–486 (2004).
119. Aretz, S. *et al.* Familial adenomatous polyposis: aberrant splicing due to missense or silent mutations in the *APC* gene. *Hum. Mut.* **24**, 370–380 (2004).
120. O'Driscoll, M., Ruiz-Perez, V. L., Woods, C. G., Jeggo, P. A. & Goodship, J. A. A splicing mutation affecting expression of ataxia-telangiectasia and Rad3-related protein (ATR) results in Seckel syndrome. *Nature Genet.* **33**, 497–501 (2003).
121. Ishibashi, F. *et al.* Improved superoxide-generating ability by interferon γ due to splicing pattern change of transcripts in neutrophils from patients with a splice site mutation in *CYBB* gene. *Blood* **98**, 436–441 (2001).
122. Flusser, H. *et al.* Mild glycine encephalopathy (NKH) in a large kindred due to a silent exonic *GLDC* splice mutation. *Neurology* **64**, 1426–1430 (2005).
123. Harteveld, C. L. *et al.* An α-thalassemia phenotype in a Dutch Hindustani, caused by a new point mutation that creates an alternative splice donor site in the first exon of the α2-globin gene. *Hemoglobin* **28**, 255–259 (2004).
124. Wicklow, B. A. *et al.* Severe subacute G_{M_2} gangliosidosis caused by an apparently silent *HEXA* mutation (V324V) that results in aberrant splicing and reduced *HEXA* mRNA. *Am. J. Med. Genet. Part A* **127A**, 158–166 (2004).
125. Xie, J. L., Pabon, D., Jayo, A., Butta, N. & Gonzalez-Manchon, C. Type I Glanzmann thrombasthenia caused by an apparently silent β3 mutation that results in aberrant splicing and reduced β3 mRNA. *Thromb. Haemost.* **93**, 897–903 (2005).
126. Buchroithner, B. *et al.* Analysis of the *LAMB3* gene in a junctional epidermolysis bullosa patient reveals exonic splicing and allele-specific nonsense-mediated mRNA decay. *Lab. Invest.* **84**, 1279–1288 (2004).
127. Du, Y. Z., Dickerson, C., Aylsworth, A. S. & Schwartz, C. E. A silent mutation, C924T (G308G), in the *L1CAM* gene results in X linked hydrocephalus (HSAS). *J. Med. Genet.* **35**, 456–462 (1998).
128. Amati-Bonneau, P. *et al.* Sporadic optic atrophy due to synonymous codon change altering mRNA splicing of *OPA1*. *Clin. Genet.* **67**, 102–103 (2005).
129. Fernandez-Cadenas, I. *et al.* Splicing mosaic of the myophosphorylase gene due to a silent mutation in McArdle disease. *Neurology* **61**, 1432–1434 (2003).
130. Mizuguchi, T. *et al.* Heterozygous *TGFBR2* mutations in Marfan syndrome. *Nature Genet.* **36**, 855–860 (2004).
131. Hoopengardner, B., Bhalla, T., Staber, C. & Reenan, R. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301**, 832–836 (2003).
132. Sémon, M., Lobry, J. R. & Duret, L. No evidence for tissue-specific adaption of synonymous codon usage in human. *Mol. Biol. Evol.* 9 November 2005 (doi:10.1093/molbev/msj053).

Acknowledgements
The authors wish to thank K. Wolfe, F. Kondrashov and an anonymous reviewer for helpful comments on the manuscript. J.V.C. and J.L.P. were funded by the UK Biotechnology and Biological Sciences Research Council.

Competing interests statement
The authors declare no competing financial interests.

DATABASES
The following terms in this article are linked online to:
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
BRCA1 | *CFTR*

FURTHER INFORMATION
Laurence Hurst's homepage: http://www.bath.ac.uk/bio-sci/hurst.htm
Access to this interactive links box is free online.

Chapter 3. **Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers**

Joanna L. Parmley, Jean-Vincent Chamary and Laurence D. Hurst

Molecular Biology and Evolution (2006) 23(2): 301-309

Evidence for Purifying Selection Against Synonymous Mutations in Mammalian Exonic Splicing Enhancers

Joanna L. Parmley, J. V. Chamary, and Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

Silent sites in mammals have classically been assumed to be free from selective pressures. Consequently, the synonymous substitution rate (K_s) is often used as a proxy for the mutation rate. Although accumulating evidence demonstrates that the assumption is not valid, the mechanism by which selection acts remain unclear. Recent work has revealed that the presence of exonic splicing enhancers (ESEs) in coding sequence might influence synonymous evolution. ESEs are predominantly located near intron-exon junctions, which may explain the reduced single-nucleotide polymorphism (SNP) density in these regions. Here we show that synonymous sites in putative ESEs evolve more slowly than the remaining exonic sequence. Differential mutabilities of ESEs do not appear to explain this difference. We observe that substitution frequency at four-fold synonymous sites decreases as one approaches the ends of exons, consistent with the existing SNP data. This gradient is at least in part explained by ESEs being more abundant near junctions. Between-gene variation in K_s is hence partly explained by the proportion of the gene that acts as an ESE. Given the relative abundance of ESEs and the reduced rates of synonymous divergence within them, we estimate that constraints on synonymous evolution within ESEs causes the true mutation rate to be underestimated by not more than ~8%. We also find that K_s outside of ESEs is much lower in alternatively spliced exons than in constitutive exons, implying that other causes of selection on synonymous mutations exist. Additionally, selection on ESEs appears to affect nonsynonymous sites and may explain why amino acid usage near intron-exon junctions is nonrandom.

Introduction

At least in mammals, synonymous (silent) sites have long been assumed to be free from the pressures of natural selection (Eyre-Walker 1991; Sharp et al. 1995). If synonymous mutations are neutral (King and Jukes 1969; Kimura 1977) then the rate of synonymous substitution can be employed to measure the point mutation rate (e.g., Eyre-Walker and Keightley 1999; Keightley and Eyre-Walker 2000). Recently, however, there has been mounting evidence against this line of thought (Iida and Akashi 2000; Bustamante, Nielsen, and Hartl 2002; Hellmann et al. 2003; Keightley and Gaffney 2003; Urrutia and Hurst 2003; Chamary and Hurst 2004; Comeron 2004; Chamary and Hurst 2005*b*; Lavner and Kotlar 2005; Lu and Wu 2005). For example, constitutively and alternatively spliced exons differ in GC content at third (largely synonymous) sites (Iida and Akashi 2000).

What might be the mechanism for selection at so-called silent sites in exons? The classical model, that selection favors efficient translation (e.g., Ikemura 1985; Bulmer, Wolfe, and Sharp 1991; Akashi and Eyre-Walker 1998; Duret 2002), may not apply in mammals (Duret 2002; dos Reis, Savva, and Wernisch 2004) (but see Urrutia and Hurst 2003; Comeron 2004; Lavner and Kotlar 2005). Some evidence suggests that synonymous sites might be of importance in mRNA secondary structure and stability (Duan and Antezana 2003; Duan et al. 2003; Capon et al. 2004; Chamary and Hurst 2005*b*). Here we consider the possibility that purifying selection acts at synonymous sites to ensure efficient pre-mRNA splicing (Willie and Majewski 2004; Chamary and Hurst 2005*a*).

Exons are classically thought to be defined by sequence located within introns: the 5′ splice site, branch point, and 3′ splice site (Robberson, Cote, and Berget 1990). However, this tripartite signal (Fairbrother and Chasin 2000) is often necessary but not sufficient for intron excision. In human introns, these signals contain only half the required information for accurate splicing (Lim and Burge 2001). The polypyrimidine tract is important for regulating alternative splicing (Spellman et al. 2005). Exonic splicing enhancers (ESEs) are oligonucleotide sequences that are abundant in both constitutively and alternatively spliced exons (Tian and Kole 1995; Coulter, Landree, and Cooper 1997; Liu, Zhang, and Krainer 1998; Schaal and Maniatis 1999; Fairbrother et al. 2002). Most ESEs are thought to function though the binding of serine/arginine-rich proteins, which help instigate spliceosome assembly and localization (Wang et al. 2004). The Burge/Sharp group recently developed a computational method (Fairbrother et al. 2002; Fairbrother et al. 2004*b*) that identifies candidate hexameric sequences with ESE activity (for a brief summary of how these are defined, see *Materials and Methods*). The density of these ESE hexamers increases as one approaches intron-exon junctions (Supplementary Fig. 1, Supplementary Material online; Fairbrother et al. 2004*a*). ESE activity is optimal within ~70 nucleotides of splice sites, although the effect is dependent on the strength of the enhancer, with potent enhancers exerting an influence at double this distance (Graveley, Hertel, and Maniatis 1998).

Prior evidence suggests that codon choice is biased owing to the presence of ESEs and biased against intronic splicing enhancers (Willie and Majewski 2004; Chamary and Hurst 2005*a*), e.g., the codon GAA is common in ESEs and is increasingly preferred over its synonym GAG near intron-exon boundaries. It is unclear, however, whether this explains all the trends in codon bias as a function of distance from exonic ends (S. T. Eskesen, F. N. Eskesen, and Ruvinsky 2004; Chamary and Hurst 2005*a*). Consistent with a preference for ESEs at particular exonic locations, at least two genes exhibit a marked reduction in the synonymous rate of evolution in regions containing an

Key words: codon usage bias, mutation rate, purifying selection, splicing, synonymous sites.

E-mail: l.d.hurst@bath.ac.uk.

Mol. Biol. Evol. 23(2):301–309. 2006
doi:10.1093/molbev/msj035
Advance Access publication October 12, 2005

ESE (BRCA1: Hurst and Pal 2001; Liu et al. 2001; Orban and Olah 2001; CFTR: Pagani, Raponi, and Baralle 2005). More generally, it has been reported that single-nucleotide polymorphism (SNP) density decreases as one approaches the ends of exons (Majewski and Ott 2002) and that this can be explained by increasing ESE density (Fairbrother et al. 2004*a*; see also Carlini and Genut 2005). Although some ESEs appear to be conserved over the course of evolution (Yeo et al. 2004), it has not previously been demonstrated that the fixation of certain mutations have been opposed by natural selection because they occur within ESEs. Consequently, here we ask whether putative ESEs are associated with a lower rate of synonymous evolution and, if they are, what impact this might have had on estimates of the mutation rate (μ) derived from the rate of synonymous nucleotide substitution (K_s).

Materials and Methods

Alignments of Orthologous Mammalian Genes

We downloaded the 7,645 human-chimpanzee-mouse orthologues used by Clark et al. (2003) from http://www.sciencemag.org/cgi/content/full/302/5652/1960/DC1, using only those alignments where each of the three sequences contained a start codon and a terminal stop codon. Alignment trios containing sequences with lengths that were not multiples of three or contained internal stop codons were discarded. Sequences from the remaining trios were translated and aligned at the amino acid level using MUSCLE, http://www.drive5.com/muscle, after which the peptide sequences were used to reconstruct the nucleotide alignment.

Determining the Location of Intron-Exon Junctions

The GeneID (LocusLink) numbers in the annotation file were used to derive the human RefSeq identifiers at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene. We then compared the human sequences in the alignments to those in the RefSeq files, retaining only those which were of same length and >99% identical. The RefSeq identifier was then used to identify genomic sequence (hence exon structure of the human coding sequence [CDS]) at Ensembl, http://www.ensembl.org/Homo_sapiens/exportview. We justify the use of the exon structure from human genes to define intron-exon junctions in other mammals because such structures are highly conserved (Roy, Fedorov, and Gilbert 2003). We ignored Ensembl genomic files where the CDS of the associated RefSeq was not the same length as that derived from the genomic annotation. For the 972 genes remaining, the intron-exon junctions in the alignments were reconstructed from the genomic sequence.

Obtaining Exonic Splicing Enhancers and Silencers

Candidate ESEs and exonic splicing silencer (ESS) sequences were identified 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 weak nonconsensus splice sites than in exons with strong consensus splice sites (Fairbrother et al. 2004*b*). 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. ESS motifs were identified by screening a library of random decamers for splicing activity in an in vivo reporter system (Wang et al. 2004). Human and mouse ESEs were downloaded from the RESCUE-ESE Web Server, http://genes.mit.edu/burgelab/rescue-ese, while human ESSs came from the supplementary data of Wang et al. (2004), http://www.download.cell.com/supplementarydata/cell/119/6/831/DC1index.htm.

Identification of ESEs and ESSs Within CDS

Defining sequence as ESE or ESS is nontrivial, so we took several different approaches. In principle, a putative ESE within an alignment could be defined as sequence present in one, either, or both species. Although one might imagine that the latter is the best definition because it is the most restrictive, human and mouse ESEs are very similar (e.g., 175/238 human hexamers are also found in mouse) and so this protocol may well end up isolating slow evolving sequence, rather than ESE. Consider the following hypothetical human-mouse alignment:

Human GAAGAATTT
Mouse CCCGAAGAA

If the hexamer GAAGAA is only identified in one species (by "human masking" or "mouse masking"), six of the nine sites are considered to be associated with the ESE (underlined) and 3 nucleotide substitutions have occurred. Under our most stringent definition of an ESE ("human + mouse masking"), only the three sites (GAA) that are within hexamers in both species are considered. Note that it is not the alignments but the sequences themselves that are scanned for the presence of putative ESEs/ESSs (gaps are collapsed and then later reinserted). Non-ESE regions were defined as the remaining unmasked sequence.

Evolutionary Rate Estimation

Nonsynonymous (K_a) and synonymous (K_s) substitution rates were estimated with the Li method (Li 1993) using the Kimura 2-parameter model. Whenever possible, to control for heterogeneity in mutation/substitution rates between genes (e.g., Lercher, Chamary, and Hurst 2004), differences in rates between putative ESE and non-ESE were performed by paired analyses using *t*-tests or one-sample Wilcoxon signed-rank tests. To minimize the effect of noise when sampling short sequence, we only considered pairs of sequences (ESE vs. non-ESE) where neither rate estimate was unusually high for the comparison (human-chimpanzee $K_a < 0.01$ and $K_s < 0.03$; human-mouse $K_a < 0.2$ and $K_s < 0.75$).

Frequency of Substitutions as a Function of Distance from Intron-Exon Junctions

Each exon was divided in two, with the first half being considered the 5′ end and the second the 3′ end. Under this

protocol no given site can be counted more than once. Running toward the interior of an exon, the distance from the intron-exon junction is the number of nucleotides (including gaps) from the junction pertinent to the half-exon. If a given site was fourfold degenerate in both species, we incremented the count of the number of sites at that distance and the number of substitutions where appropriate.

We also obtained ESE hexamers predicted to be predominantly active at the 5′ and 3′ ends of exons. The human ESE clusters were kindly provided by Will Fairbrother and the mouse 5′ and 3′ ESEs by Gene Yeo. Masking 5′ ends using ESEs with 5′ activity and 3′ ends with 3′ ESEs does not qualitatively affect our results (data not shown).

Comparison of Alternative and Constitutive Exons

We obtained the "training" set of exons (Yeo et al. 2005) from ACEScan, http://genes.mit.edu/acescan, where we have high confidence that exons have been conserved as being alternative or constitutive between human and mouse. The mouse and human exons were aligned at the nucleotide level using ClustalX. Exons in which the number of single-base indels in the alignment was not a multiple of three were eliminated (16 of the alternative exons and 24 of the constitutive ones). For the remainder we calculated the Tamura-Nei distance (Tamura and Nei 1993). For each of the three possible reading frames, we followed the method of Xing and Lee (2005) to ascribe the correct frame. After translating all exons in each of the three frames and eliminating those containing a stop codon, for each exon we calculated K_a for each of the remaining frames and employed the frame with the lowest K_a as the reading frame.

Results

Synonymous Evolution Is Slower in ESEs

If selection acts to preserve splicing activity (Yeo et al. 2004), the rate of synonymous substitution (K_s) should be lower in putative ESEs when compared with non-ESE sequence. To investigate this we scanned a data set of chimpanzee-human-mouse orthologues (Clark et al. 2003) for the presence of 238 putative human (Fairbrother et al. 2002) and 380 mouse (Yeo et al. 2004) ESE hexamers. As ESEs have yet to be identified in chimpanzees, here we report data for the human-mouse comparison, although the use of human hexamers as a "chimpanzee" set yields qualitatively the same results (Supplementary Table 1, Supplementary Material online; additional data available upon request). Similarly, as many ESEs are conserved (Yeo et al. 2004), one can also identify "mammalian" enhancers. This too gives similar results (Supplementary Table 2, Supplementary Material online).

As it is unclear on a priori grounds whether we should consider putative ESEs as being present in one or both species, we employ various masking protocols to identify sites that might be associated with putative ESEs. The first method identifies ESE sites as those that occur within human hexamers in human sequence (human masking). The second considers ESE sites to be those that are within mouse hexamers (mouse masking). Using more stringent definitions, we can also define ESE sites to be those present within hexamers in both sequences (human + mouse masking). This involves masking human hexamers in human sequence and mouse hexamers in mouse sequence, realigning the masked sequences (based on the original unmasked alignment), and then identifying those sites in the alignment where both sequences are putatively ESE.

In all masking permutations, we find that the synonymous substitution rate in putative ESEs is lower than that in non-ESEs (table 1; Supplementary Table 1, Supplementary Material online). The magnitude of the reduction in K_s is dependent on the masking protocol. The difference in K_s is relatively modest when masking hexamers in single species (~5%) but quite large in the more stringent double masking (~35%).

Reduced K_s Within ESEs Is Not Due to a Skewed CpG Distribution

Sites within CpG dinucleotides are known to be hypermutable (Bird 1980; Cooper and Krawczak 1989; Sved and Bird 1990), and ESEs are typically purine rich (Blencowe 2000) (in combined human/mouse hexamers A = 42.5%, G = 25.7%, C = 17.9%, and T = 13.9%). Consequently, it is possible that the reduction in K_s is an artefact owing to non-ESE sequence having a higher concentration of CpGs. However, after repeating the above analysis, this time omitting CG/GC pairs in either sequence, we again find that putative ESEs evolve more slowly than non-ESEs

Table 1
Differences in the Rate of Synonymous Evolution Between Putative ESE and Non-ESE Sequence in Human-Mouse Alignments

Masking Protocol[a]	Non-ESE[b]	ESE[b]	N[c]	P[d]
Human	0.4484 ± 0.0042	0.4117 ± 0.0054	812	8×10^{-11}
Human non-CpG	0.3378 ± 0.0041	0.3006 ± 0.0053	848	1×10^{-12}
Mouse	0.4440 ± 0.0040	0.4377 ± 0.0048	854	0.0538
Mouse non-CpG	0.3343 ± 0.0041	0.3184 ± 0.0048	889	8×10^{-5}
Human + mouse	0.4701 ± 0.0042	0.2896 ± 0.0053	815	3×10^{-103}
Human + mouse non-CpG	0.3488 ± 0.0041	0.2157 ± 0.0048	797	3×10^{-77}

[a] The sequences in which putative ESE motifs are masked. For human + mouse, these are the sites that are identified as being associated with ESEs in both species.

[b] The mean synonymous substitution rate (±SEM).

[c] The number of genes analyzed in pairwise comparisons.

[d] The significance of the difference between ESE and non-ESE (P values from paired t-tests).

(table 1). In fact, the previously marginally nonsignificant difference in the mouse masking now becomes significant. We conclude that the decreased K_s in ESEs cannot be explained by differential abundances of hypermutable CpGs.

Reduced K_s Within ESEs Is Not Due to a Skewed Nucleotide Distribution

The above test considers a class of well-known hypermutable sites. However, different nucleotides may themselves have different mutabilities (see e.g., Chamary and Hurst 2004). More generally, we can ask whether, controlling for skewed nucleotide contents, ESEs still have unusually low synonymous rates of evolution. Moreover, it is also possible that the reduction in K_s is a result of searching for relatively little sequence (particularly in human + mouse masking) which will artificially isolate slowly evolving sequences.

To examine these possibilities we performed a simulation. In each of 1,000 randomizations, we generated a set of simulated hexamers of the same average nucleotide composition as the real ESE hexamers. These simulated sets are then used to carry out human, mouse, and the human + mouse (stringent) maskings. For each gene, the difference between the real and the simulants was expressed as a Z-score, the number of standard deviations the observed K_s (from real ESEs) is away from the mean K_s of the simulated ESEs. Under a null hypothesis that the reduced K_s in ESE is due to the masking protocol and/or skewed nucleotide content in ESEs, the Z-score distribution should have an average that is not significantly different from zero. Alternatively, if putative ESEs evolve slowly, then their K_s should be significantly lower than the average of the simulants, i.e., a negative Z-score. Under the three protocols studied, we found that this was indeed the case (human masking median $Z = -0.293$, $P < 0.0001$; mouse median $Z = -0.214$, $P < 0.0001$; human + mouse median $Z = -0.17$, $P = 0.015$). We conclude that the low K_s in putative ESEs is not owing to skewed nucleotide content or any bias introduced by the masking process.

Substitution Frequency at Fourfold Degenerate Sites Declines Near Intron-Exon Junctions, Which Is Partially Explained by the Presence of ESEs

While the above results are consistent with a model in which ESE sequence is under selection to retain their function, there exists a further possibility. ESE density is known to be highest near intron-exon junctions. If, for some other reason, sequence in the near vicinity of such junctions are under stronger selection (or experience low mutation rates), then ESEs would have lower rates of evolution than either non-ESE sequence or our simulated ESEs, both of which may be relatively more common in exonic interiors. For example, exon-exon junctions tend to occur at or around the position of nucleosome formation (Kogan and Trifonov 2005). If nucleosomal or perinucleosomal sequence is more conserved than the average, then we may expect ESEs to be slow evolving, but only because they tend to be near nucleosomes. Note too that there may well be patterns of nucleotide usage across exons that are not explained by ESE presence/absence (S. T. Eskesen, F. N. Eskesen, and Ruvinsky 2004; Chamary and Hurst 2005*a*). We can therefore ask whether, given their location in proximity to the junctions, ESEs evolve slower than non-ESEs and whether this alone is adequate to explain the reduced SNP density near intron-exon junctions (Fairbrother et al. 2004*a*).

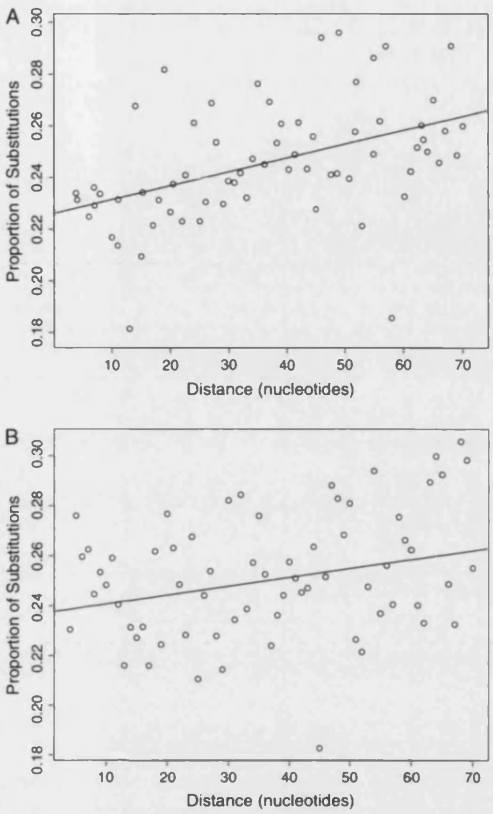


FIG. 1.—Frequency of substitutions at fourfold degenerate sites in human-mouse alignments as a function of distance from intron-exon junctions, at (*A*) the 5′ end of exons (slope = 0.2260; $R^2 = 0.1995$; $P = 9 \times 10^{-05}$) and (*B*) the 3′ end (slope = 0.2372; $R^2 = 0.0660$; $P = 0.0203$). The lines of best fit are derived by linear regression and weighted by the number of sites.

The frequency of substitutions at fourfold degenerate sites was assessed as a function of distance from both the 5′ and 3′ ends of exons, without masking ESE/non-ESE but ignoring CpGs. This analysis strongly suggests that synonymous mutations are increasingly opposed as one approaches the end of an exon (fig. 1). Studies looking at SNP density have suggested that such selection only extends about 30 nt into exons (Majewski and Ott 2002; Fairbrother et al. 2004*a*), but we observe an effect that is closer to the biased codon choice data (~100 nt, Willie and Majewski 2004; Chamary and Hurst 2005*a*).

Given the possible discrepancy in the scale of the effect, we then asked whether it is likely owing to a reduced rate of evolution in ESEs coupled with their greater

Table 2
ANCOVA Between Putative ESE and Non-ESE Sequences for the Substitution Frequency at Fourfold Synonymous Sites as a Function of Distance from Intron-Exon Junctions in Human-Mouse Alignments

		5′ End of Exons		3′ End of Exons	
Masking Protocol[a]	Parameter	Estimate[b]	P[c]	Estimate[b]	P[c]
Human non-CpG	Distance	0.0005 ± 0.0001	7×10^{-5}	0.0003 ± 0.0001	0.0137
	Level	0.0254 ± 0.0054	7×10^{-6}	0.0214 ± 0.0060	0.0005
Mouse non-CpG	Distance	0.0005 ± 0.0001	0.0002	0.0003 ± 0.0001	0.0123
	Level	0.0231 ± 0.0053	3×10^{-5}	0.0376 ± 0.0051	2×10^{-11}
Human + mouse non-CpG	Distance	0.0005 ± 0.0001	0.0001	0.0003 ± 0.0001	0.0244
	Level	0.0886 ± 0.0061	$<2 \times 10^{-16}$	0.1036 ± 0.0067	$<2 \times 10^{-16}$

[a] The sequences in which putative ESEs are masked.

[b] The "Estimate" for "Distance" is the slope of the regression line (±SEM) for the substitution frequency at fourfold sites in ESEs plotted against the distance from the intron-exon junction. There is no difference between the slopes derived from ESE and non-ESE sequences ($P > 0.05$). The estimate for "Level" is the difference between the slopes (±SEM) for ESE and non-ESE.

[c] For Distance, the P value indicates whether the common slope (ESE was used) is significant. For Level, the P value indicates whether there is a difference between ESEs and non-ESEs while controlling for the distance from the junction, i.e., to determine whether, at a given distance from the junction, the proportion of substitutions at fourfold sites differs between ESE and non-ESE.

proximity to intron-exon junctions or to some more general underlying cause. Under the first model, we expect both ESE rates of evolution and non-ESE rates of evolution to show no trend as a function of the distance from the junction, but with the ESE synonymous rates lower than those of the non-ESEs. In the second case, we might expect ESE and non-ESE to show the same trend of increasing synonymous divergence as a function of distance from the junction and no difference in the rates of evolution controlling for distance from junction.

These hypotheses were tested by analysis of covariance (ANCOVA) in which the distance from the junction was the covariate, and ESE and non-ESE sequence were the two factors/groups (NB there is no significant interaction term, so the assumptions of ANCOVA are upheld, $P > 0.05$). The difference in rates between the groups was always significant controlling for the distance from the junction ("Level" in table 2). This strongly suggests that ESEs are slow evolving even controlling for their differential abundance near junctions (table 2 and fig. 2). In all cases, there remains an effect whereby all sequences evolve marginally slower if closer to the junction ("Distance" in table 2). This suggests the presence of some weak force affecting substitution rates as a function of the distance from the junction independent of ESE presence or absence. As the effect is weak, however, we cannot rule out the possibility that it arises as a consequence of missing true ESEs in our classification.

The Effect of ESEs on Evolution at Nonsynonymous Sites

Here we have concentrated on how conservation of ESEs can influence synonymous mutations and codon usage. In principle, however, ESEs could also affect nonsynonymous mutations. This may well be the case as K_a is lower in putative ESEs (table 3). Moreover, as ESEs are generally purine rich (Blencowe 2000), it is interesting to ask whether amino acids specified by purine-rich codons are also more abundant near junctions. If so, we should expect the effect to be most strikingly seen for usage of lysine (AAA and AAG), A being the most common nucleotide in ESEs followed by G. This is indeed observed (fig. 3). However, while AG-rich codons tend to be employed near boundaries, at least for the 3′ end, the effect is more striking for AT-rich codons (Supplementary Fig. 2, Supplementary Material online). This suggests a pressure toward A and T rather than A and G and might hint at some other force (e.g., Chamary and Hurst 2005*a*). This is unlikely to be nucleosome associated as in mouse and human these are associated with G and C (Kogan and Trifonov 2005).

Discussion

Our analyses demonstrate that ESEs are under purifying selection. As the enhancer regions do not discriminate between synonymous and nonsynonymous sites, it is perhaps unsurprising that both classes of site are under constraint due to the presence of ESEs, most profoundly at the end of exons. This finding tempts several questions. First, assuming selection on splicing enhancers is the only mode of selection on synonymous mutations, to what extent might one underestimate the mutation rate when extrapolating from synonymous divergence? Second, is it likely that this is the only mechanism of selection on synonymous mutations? To address the latter issue, we examine alternative exons, these being known to have lower synonymous substitution rates than constitutive ones from the same gene (Iida and Akashi 2000; Xing and Lee 2005). Finally, we ask about implications of the finding that codon usage and rates of evolution are unusual in the vicinity of intron-exon junctions.

Selection on ESEs Has a Modest Effect on Underestimation of the Mutation Rate

Under the supposition that synonymous sites evolve neutrally, their rate of evolution has been used as a measure of the mutation rate (see e.g., Eyre-Walker and Keightley 1999; Keightley and Eyre-Walker 2000). Assuming selection on ESEs to be the only form of selection at synonymous sites, how much might this method underestimate the real mutation rate? To address this issue we need to know what proportion of the sequence is functional splicing

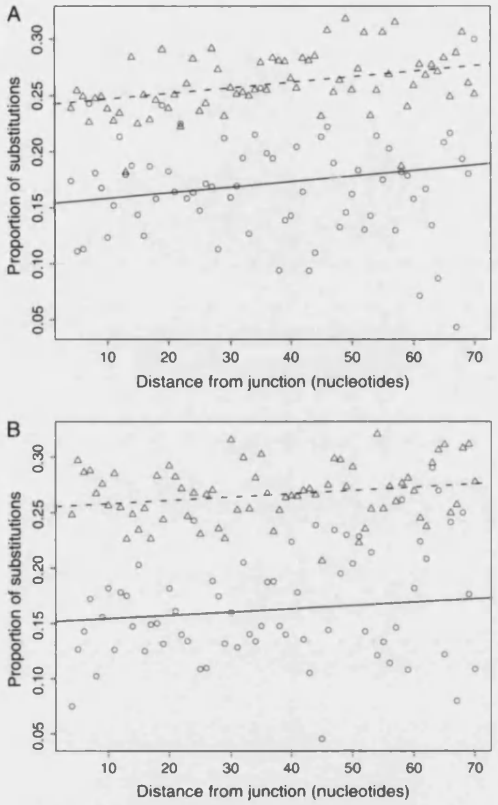


FIG. 2.—Frequency of substitutions at fourfold degenerate sites in human-mouse alignments as a function of distance from intron-exon junctions in ESE (circles and solid lines) and non-ESE (triangles and dashed lines) sequences, at (*A*) the 5′ end of exons and (*B*) the 3′ end. The weak trends are shown for sites within ESEs at the 5′ (*A*, slope = 0.1658; $R^2 = 0$; $P = 0.6831$) and 3′ end (*B*, slope = 0.1369; $R^2 = 0.0773$; $P = 0.0130$), and non-ESE sequence at the 5′ (*A*, slope = 0.2396; $R^2 = 0.1625$; $P = 0.0004$) and 3′ end (*B*, slope = 0.2575; $R^2 = 0.0143$; $P = 0.1664$). The lines of best fit are derived by linear regression and weighted by the number of nucleotide sites. The ESE masking is by the human + mouse protocol.

enhancer and what, on the average, is the reduction in the rate of evolution within ESEs.

We have employed three different methods to define putative ESEs. Enhancers identified within a single species (mouse or human) show a modest 1%–11% reduction in their rate of evolution (depending on whether we ignore CpGs, table 4). Sequence defined as ESE in both mouse and human have a more striking ~38% reduction in their rate compared with non-ESE regions (table 4). However, the more stringent definition defines less of the sequence as being in enhancer. When we factor in the proportion of sequence that is putatively ESE, the three methods all suggest that the net reduction in K_s, owing to the presence of ESEs, is modest. It may be as low as 2% and unlikely to be much more than 8% (table 4). This suggests that correction for the presence of ESEs will not have a major effect on estimates of the mutation rate, not least because the margin of error associated with estimates of the number of generations between any two mammalian taxa is vastly more error prone and alterations here will have a much more profound effect.

Selection on ESEs Is Only One Form of Selection on Synonymous Mutations

Conservation of ESEs is unlikely to be the only form of selection at synonymous sites. In terms of splicing, biased codon usage may also reflect an avoidance of certain sequences that might be associated with cryptic splice sites (S. T. Eskesen, F. N. Eskesen, and Ruvinsky 2004; but see Chamary and Hurst 2005*a*). Additionally, we have not considered the contribution of ESS sequence, although we find that masking the 133 decamers that have been systematically identified in humans (Wang et al. 2004) does not alter our conclusions (Supplementary Table 3, Supplementary Material online). Importantly, the strongest signal for selection that has been seen so far is a high stability of cytosine at third sites (Chamary and Hurst 2004). This is not obviously explained by a role in the splicing process (Chamary and Hurst 2005*a*) because ESEs are AG rich and C poor. The cause of the C preference remains unclear, but a role in mRNA stability is supported by some data (Chamary and Hurst 2005*b*). There may also be other factors that constrain synonymous evolution, such as the need to bind antisense transcripts. Therefore, we cannot conclude that selection on silent sites has not lead to a significant underestimate of the mutation rate.

Selection on ESEs Does Not, for the Most Part, Explain Low Synonymous Rates in Alternative Transcripts

Another way to address whether other forms of selection act at synonymous mutations is to ask whether it is a greater abundance of and/or stronger selection on ESEs that might explain why alternatively spliced exons have unusually low rates of synonymous evolution (Iida and Akashi 2000; Xing and Lee 2005). To address this, we examined a carefully curated set of conserved alternative and constitutive exons (Yeo et al. 2005). We see that mean substitution rates in alternative exons (Tamura-Nei distance = 0.069 ± 0.004; $N = 225$) is lower ($P < 0.0001$ by Mann-Whitney *U*-test) than that in constitutive exons (0.123 ± 0.001, $N = 5{,}045$). This is owing to a much lower rate of evolution at both synonymous sites and, in contrast to prior analyses (Iida and Akashi 2000; Xing and Lee 2005), nonsynonymous sites, although the effect is more dramatic for the former. Examining exons with a minimum of 30 codons, for example, we find that the mean K_s is lower in alternative exons (0.115 ± 0.02; $N = 51$) compared to constitutive exons (0.311 ± 0.009; $P < 0.0001$ by Mann-Whitney *U*-test) while K_a in alternative exons (0.058 ± 0.008) is lower than that in constitutives (0.103 ± 0.002; $P = 0.0003$ by Mann-Whitney *U*-test). The reduced K_s is not due to alternative exons possessing more ESEs, as we find that there is no consistent difference in the proportion of putative enhancer sequence between the two classes

Table 3
Differences in the Rate of Amino Acid Evolution Between Putative ESE and Non-ESE Sequence in Human-Mouse Alignments

Masking Protocol[a]	Non-ESE[b]	ESE[b]	N^c	P^d
Human	0.0526 ± 0.0015	0.0473 ± 0.0015	862	5×10^{-9}
Human non-CpG	0.0394 ± 0.0013	0.0404 ± 0.0015	874	0.5685
Mouse	0.0524 ± 0.0015	0.0503 ± 0.0015	890	0.0147
Mouse non-CpG	0.0396 ± 0.0013	0.0402 ± 0.0014	908	0.4211
Human + mouse	0.0545 ± 0.0016	0.0343 ± 0.0013	838	2×10^{-68}
Human + mouse non-CpG	0.0418 ± 0.0015	0.0298 ± 0.0013	815	1×10^{-34}

[a] The sequences in which putative ESEs are masked.
[b] The mean nonsynonymous substitution rate (±SEM).
[c] The number of genes analyzed in pairwise comparisons.
[d] The significance of the difference between ESE and non-ESE (*P* values from paired *t*-tests).

of exons (Supplementary Table 4, Supplementary Material online). Is then the reduced rate of evolution especially noticeable in ESEs, and is it seen in non-ESE parts of alternative transcripts?

As regards the second issue, the rate of synonymous evolution in non-ESE sequence of alternative exons is over 50% lower than that for non-ESE parts of constitutive exons (Supplementary Table 5, Supplementary Material online). This strongly suggests that selection on ESEs cannot fully explain why alternative exons are slow evolving. Although the data are noisy, our best evidence suggests that ESEs in alternative transcripts have K_s values that are slightly lower than that of non-ESE in the same alternative exon (Supplementary Table 6, Supplementary Material online). The causes of the unusually low rates of evolution in conserved alternative exons deserve further scrutiny.

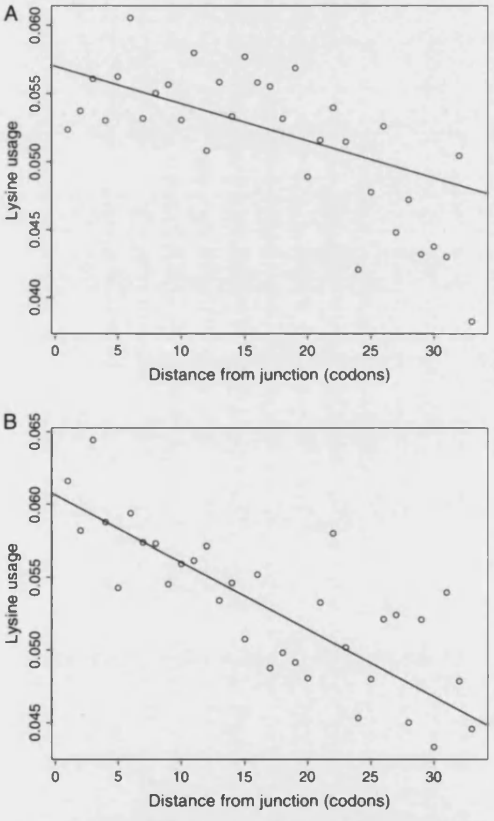


FIG. 3.—Lysine residue usage as a function of distance from intron-exon junctions, at (*A*) the 5′ end of exons ($R^2 = 0.2936$; $P = 0.0007$) and (*B*) the 3′ end ($R^2 = 0.6364$; $P = 2 \times 10^{-8}$). The lines of best fit are derived by linear regression and weighted by the number of codons.

Implications of Stronger Selection Near Intron-Exon Junctions

One consequence of all the evidence for skewed nucleotide composition (Louie, Ott, and Majewski 2003; S. T. Eskesen, F. N. Eskesen, and Ruvinsky 2004) and biased codon usage (Willie and Majewski 2004; Chamary and Hurst 2005*a*) near intron-exon boundaries is that it adds layers of complexity to the interpretation of prior results. First, the conventional application of $K_a/K_s > 1$ as an indication of positive selection should be treated with caution as this may be owing to reduced K_s rather than elevated K_a (Pond and Muse 2005), as previously described in at least two genes (BRCA1 [Hurst and Pal 2001; Liu et al. 2001; Orban and Olah 2001] and CFTR [Pagani, Raponi, and Baralle 2005]). Further, several recent reports find evidence for systematic codon bias that is not explained by background nucleotide content (Urrutia and Hurst 2003; Comeron 2004; Lavner and Kotlar 2005). For example, highly expressed genes exhibit the greatest bias (Urrutia

Table 4
The Contribution of Purifying Selection at Synonymous Sites in Putative ESEs to Underestimates of the Mutation Rate (μ) in Mammals

Masking Protocol[a]	K_s Reduction[b] (%)	ESE Coverage[c] (%)	μ Underestimation (%)
Human	8.19	30.42	2.49
Human non-CpG	11.03	30.42	3.36
Mouse	1.41	40.30	0.57
Mouse non-CpG	4.74	40.30	1.91
Human + mouse	38.39	21.77	8.36
Human + mouse non-CpG	38.15	21.77	8.31

[a] The sequences in which putative ESEs are masked.
[b] The difference in the synonymous substitution rate between ESE and non-ESE.
[c] The proportion of sequence covered by ESE sites.

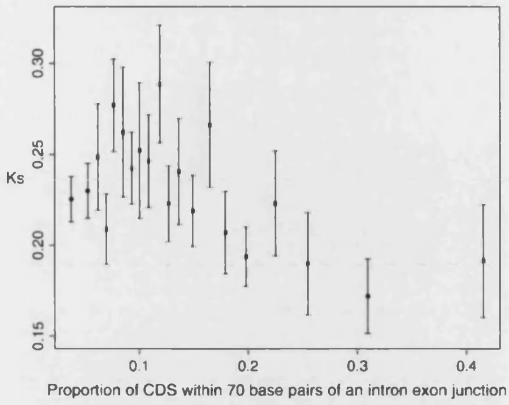


FIG. 4.—The synonymous substitution rate (K_s) as a function of the proportion of CDS within 70 bp of an intron-exon junction. The data is split into 20 bins with equal numbers of genes ($N = 48$) in each bin.

and Hurst 2003). As intron density also varies with expression parameters (Comeron 2004), these results may be artefacts of biased codon usage in the proximity of intron-exon junctions. Indeed, when we consider the relationship between K_s and the proportion of the CDS within 70 nt of the junction, we observe a significant negative correlation (fig. 4; Spearman rank correlation $\rho = -0.15$, $P < 0.0001$). To factor out any such effects, we recommend that one should exclude those regions of exons within about 70 nt on either side of junctions.

The potential impact of ESE presence on nonsynonymous substitution rates has numerous corollaries. First, this makes it difficult to ask whether a certain protein domain is under purifying selection. A low K_a may be evidence for this, but it could also be explained by selection on an ESE rather than the protein. To examine in detail such claims, one should also ask whether the DNA specifying the domain is near an intron-exon junction and matches known ESEs. The skewed amino acid usage near intron-exon boundaries has two possible interpretations. First, that at the time of insertion, a viable intron can only be tolerated if there are already ESEs present in the near vicinity. Second, that after insertion, the process of splicing is subject to selection, with choice of amino acids around junctions being determined in part by the efficiency of splicing of flanking introns. These are not mutually incompatible. To establish whether the first is true, one would need to identify new introns within the mammal lineage. These are remarkably rare (Roy, Fedorov, and Gilbert 2003) (see also Sry in marsupials, O'Neill et al. 1998). Conversely, if loss of an intron is not followed by adjustment of amino acid content, this would suggest that amino acid content was dictated by the protein level considerations rather than splicing regulation.

Supplementary Material

Supplementary Figs. 1 and 2 and Supplementary Tables 1–6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank the anonymous referees for suggestions. J.L.P. and J.V.C. are funded by the United Kingdom Biotechnology and Biological Sciences Research Council.

Literature Cited

Akashi, H., and A. Eyre-Walker. 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. **8**:688–693.

Bird, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. **8**:1499–1504.

Blencowe, B. J. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem. Sci. **25**:106–110.

Bulmer, M., K. H. Wolfe, and P. M. Sharp. 1991. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc. Natl. Acad. Sci. USA **88**:5974–5978.

Bustamante, C. D., R. Nielsen, and D. L. Hartl. 2002. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. Mol. Biol. Evol. **19**:110–117.

Capon, F., M. H. Allen, M. Ameen, A. D. Burden, D. Tillman, J. N. Barker, and R. C. Trembath. 2004. A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. Hum. Mol. Genet. **13**:2361–2368.

Carlini, D. B., and J. E. Genut. 2005. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J. Mol. Evol. (in press).

Chamary, J. V., and L. D. Hurst. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol. Biol. Evol. **21**:1014–1023.

———. 2005*a*. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet. **21**:256–259.

———. 2005*b*. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. **6**:R75.

Clark, A. G., S. Glanowski, R. Nielsen et al. (17 co-authors). 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302**:1960–1963.

Comeron, J. M. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics **167**:1293–1304.

Cooper, D. N., and M. Krawczak. 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. **83**:181–188.

Coulter, L. R., M. A. Landree, and T. A. Cooper. 1997. Identification of a new class of exonic splicing enhancers by in vivo selection. Mol. Cell. Biol. **17**:2143–2150.

dos Reis, M., R. Savva, and L. Wernisch. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. **32**:5036–5044.

Duan, J., and M. A. Antezana. 2003. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J. Mol. Evol. **57**:694–701.

Duan, J., M. S. Wainwright, J. M. Comeron, N. Saitou, A. R. Sanders, J. Gelernter, and P. V. Gejman. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum. Mol. Genet. **12**:205–216.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12**:640–649.

Eskesen, S. T., F. N. Eskesen, and A. Ruvinsky. 2004. Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. Genetics **167**:543–550.

Eyre-Walker, A. 1991. An analysis of codon usage in mammals: selection or mutation bias? J. Mol. Evol. **33**:442–449.

Eyre-Walker, A., and P. D. Keightley. 1999. High genomic deleterious mutation rates in hominids. Nature **397**:344–347.

Fairbrother, W. G., and L. A. Chasin. 2000. Human genomic sequences that inhibit splicing. Mol. Cell. Biol. **20**:6816–6825.

Fairbrother, W. G., D. Holste, C. B. Burge, and P. A. Sharp. 2004*a*. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. **2**:e268.

Fairbrother, W. G., R. F. Yeh, P. A. Sharp, and C. B. Burge. 2002. Predictive identification of exonic splicing enhancers in human genes. Science **297**:1007–1013.

Fairbrother, W. G., G. W. Yeo, R. Yeh, P. Goldstein, M. Mawson, P. A. Sharp, and C. B. Burge. 2004*b*. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res. **32**:W187–W190.

Graveley, B. R., K. J. Hertel, and T. Maniatis. 1998. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. EMBO J. **17**:6747–6756.

Hellmann, I., S. Zollner, W. Enard, I. Ebersberger, B. Nickel, and S. Paabo. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. **13**: 831–837.

Hurst, L. D., and C. Pal. 2001. Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet. **17**:62–65.

Iida, K., and H. Akashi. 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene **261**:93–105.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.

Keightley, P. D., and A. Eyre-Walker. 2000. Deleterious mutations and the evolution of sex. Science **290**:331–333.

Keightley, P. D., and D. J. Gaffney. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proc. Natl. Acad. Sci. USA **100**:13402–13406.

Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267**:275–276.

King, J. L., and T. H. Jukes. 1969. Non-Darwinian evolution. Science **164**:788–798.

Kogan, S., and E. N. Trifonov. 2005. Gene splice sites correlate with nucleosome positions. Gene **352**:57–62.

Lavner, Y., and D. Kotlar. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene **345**:127–138.

Lercher, M. J., J. V. Chamary, and L. D. Hurst. 2004. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. Genome Res. **14**:1002–1013.

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

Lim, L. P., and C. B. Burge. 2001. A computational analysis of sequence features involved in recognition of short introns. Proc. Natl. Acad. Sci. USA **98**:11193–11198.

Liu, H. X., L. Cartegni, M. Q. Zhang, and A. R. Krainer. 2001. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. Nat. Genet. **27**:55–58.

Liu, H. X., M. Zhang, and A. R. Krainer. 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev. **12**:1998–2012.

Louie, E., J. Ott, and J. Majewski. 2003. Nucleotide frequency variation across human genes. Genome Res. **13**:2594–2601.

Lu, J., and C. I. Wu. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc. Natl. Acad. Sci. USA **102**: 4063–4067.

Majewski, J., and J. Ott. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res. **12**: 1827–1836.

O'Neill, R. J., F. E. Brennan, M. L. Delbridge, R. H. Crozier, and J. A. Graves. 1998. De novo insertion of an intron into the mammalian sex determining gene, SRY. Proc. Natl. Acad. Sci. USA **95**:1653–1657.

Orban, T. I., and E. Olah. 2001. Purifying selection on silent sites—a constraint from splicing regulation? Trends Genet. **17**:252–253.

Pagani, F., M. Raponi, and F. E. Baralle. 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc. Natl. Acad. Sci. USA **102**:6368–6372.

Pond, S. K., and S. V. Muse. 2005. Site-to-site variation of synonymous substitution rates. Mol. Biol. Evol. (in press).

Robberson, B. L., G. J. Cote, and S. M. Berget. 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. **10**:84–94.

Roy, S. W., A. Fedorov, and W. Gilbert. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc. Natl. Acad. Sci. USA **100**:7158–7162.

Schaal, T. D., and T. Maniatis. 1999. Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol. Cell. Biol. **19**:1705–1719.

Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi, and J. F. Peden. 1995. DNA-sequence evolution: the sounds of silence. Phil. Trans. R. Soc. Lond. B **349**:241–247.

Spellman, R., A. Rideau, A. Matlin, C. Gooding, F. Robinson, N. McGlincy, S. N. Grellscheid, J. Southby, M. Wollerton, and C. W. Smith. 2005. Regulation of alternative splicing by PTB and associated factors. Biochem. Soc. Trans. **33**:457–460.

Sved, J., and A. Bird. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc. Natl. Acad. Sci. USA **87**:4692–4696.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512–526.

Tian, H., and R. Kole. 1995. Selection of novel exon recognition elements from a pool of random sequences. Mol. Cell. Biol. **15**:6291–6298.

Urrutia, A. O., and L. D. Hurst. 2003. The signature of selection mediated by expression on human genes. Genome Res. **13**:2260–2264.

Wang, Z., M. E. Rolish, G. Yeo, V. Tung, M. Mawson, and C. B. Burge. 2004. Systematic identification and analysis of exonic splicing silencers. Cell **119**:831–845.

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

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

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

Yeo, G. W., E. Van Nostrand, D. Holste, T. Poggio, and C. B. Burge. 2005. Identification and analysis of alternative splicing events conserved in human and mouse. Proc. Natl. Acad. Sci. USA **102**:2850–2855.

Kenneth Wolfe, Associate Editor

Accepted October 10, 2005

Chapter 4. **Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals**

Joanna L. Parmley and Laurence D. Hurst

Molecular Biology and Evolution (2007) 24(8): 1600-1603

Exonic Splicing Regulatory Elements Skew Synonymous Codon Usage near Intron-exon Boundaries in Mammals

Joanna L. Parmley and Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

In mammals there is a bias in amino acid usage near splice sites that is explained, in large part, by the high density of exonic splicing enhancers (ESEs) in these regions. Is there a similar bias for the relative use of synonymous codons, and can any such bias be predicted by their abundance in ESEs? Prior reports suggested that such trends may exist. From analysis of human exons, we find that 47 of the 59 codons with at least one synonym show differential usage in the proximity of exon ends, of which 42 remain significant after correction for multiple testing. Within sets of synonymous codons those more preferred near splice sites are generally those that are relatively more abundant within the ESEs. However, the examples given previously appear exceptionally good fits and there exist many exceptions, the usage of lysine's codons being a case in point. Similar results are observed in mouse exons. We conclude that splice regulation impacts on the choice of synonymous codons in mammals, but the magnitude of this effect is less than might at first have been supposed.

Whether selection acts to induce and maintain codon usage bias is a question that is central to the neutralist/selectionist debate and has recently led to the identification of several convincing causative parameters. In bacteria, yeast, *Caenorhabditis* and *Drosophila*, biased codon usage is largely explained by selection for translational efficiency (including accuracy) (Ikemura 1985; Akashi and Eyre-Walker 1998; Duret 2002; Wright et al. 2004). The story in vertebrates, however, is far more complex (Kanaya et al. 2001; Lander et al. 2001; Duret 2002; Comeron 2004; dos Reis et al. 2004; Lavner and Kotlar 2005). Recent studies, in mammals, suggest that the dominant force in amino acid and codon usage is not selection for translational efficiency (dos Reis et al. 2004). Splice related biases are, however, evident (Willie and Majewski 2004). In particular, selection for the preservation of exonic splicing regulatory elements, most notably exonic splicing enhancers (ESEs), explains the low synonymous substitution rates (Parmley et al. 2006), low SNP density (Fairbrother et al. 2004a; Carlini and Genut 2006) and low protein evolutionary rates (Parmley et al. 2007) near intron-exon boundaries, this being where there is the highest concentration of regulatory elements. It also introduces a predictable amino acid bias: those amino acids encoded by codons that occur frequently in ESE sequences are preferred near intron-exon boundaries (Parmley et al. 2007). These studies accord with several in depth analyses of individual exons and genes (Pagani et al. 2005; Baralle et al. 2006; Raponi et al. 2007) indicating that both synonymous and nonsynonymous mutations can have fitness effects via the modification of splicing, in some instances leading to genetic disorders (Cartegni et al. 2002; Chamary et al. 2006).

To what extent does selection for the preservation of exonic splice regulatory elements affect codon bias? In a few anecdotal cases it has been argued that a codon that is common in ESEs is preferred near boundaries relative to the synonymous codons that feature less commonly in ESEs (Willie and Majewski 2004; Chamary and Hurst 2005). Willie and Majewski (2004) highlighted the usage of GAA compared with GAG, both specifying glutamic acid. GAA features very much more often in splice enhancers, compared with GAG, and is greatly enriched, relative to its synonym, near intron-exon boundaries. Here we ask about the generality of this observation. In particular we ask whether, given what is known of ESEs, it is possible to predict which codons are preferred near intron-exon boundaries.

A data set of over 170,000 internal human exons (see methods; Parmley et al. 2007) was assessed to determine the usage of each codon up to a distance of 30 codons from intron-exon boundaries. At any given distance from the exon end, summing across all exons, we determine the proportion of each codon amongst the set of codons seen at this distance. For each codon we then derive a plot of the proportional usage of that codon as a function of the distance from exon ends. This trend is captured by 2 statistics. First, the correlation between the proportional codon usage and distance from the exon end was assessed by Spearmans rank correlation (Rho). Second, the slope (α) on the line relating the distance from the exon boundary to the proportional usage (i.e. proportional usage = α distance from boundary + β). Naturally the 2 measures are strongly correlated. For subsequent analysis we consider only those 59 codons that specify the amino acids with at least 2-fold degeneracy.

From the significance values derived from the Spearmans rank correlation we find that there are significant trends in the usage of synonymous codons near intron-exon boundaries for 47 of the 59 codons, of which 42 are significant after sequential Bonferonni correction (supplementary table 1). We repeated the analysis for 115,466 exons from 14,005 mouse genes. The trends found in mice (supplementary table 2) are very similar to those reported in humans.

Can these trends be explained by the presence of splice control elements? A set of putative hexameric ESE sequences has been determined for several species including human and mouse (Fairbrother et al. 2004b). From these we were able to identify a set of 176 high confidence ESEs, by employing only those present in both the human and the mouse set. To identify those codons that are most common in splice regulatory elements, a hexamer preference index (HPI) was calculated, a modification of that previously developed (Parmley et al. 2007) (see supplementary methods 1). A high HPI value indicates that a given codon is

Key words: codon usage bias, mammals, exonic splicing enhancers, splicing.

Email: jlp21@bath.ac.uk

Mol. Biol. Evol. 24(8):1600–1603. 2007
doi:10.1093/molbev/msm104
Advance Access publication May 24, 2007

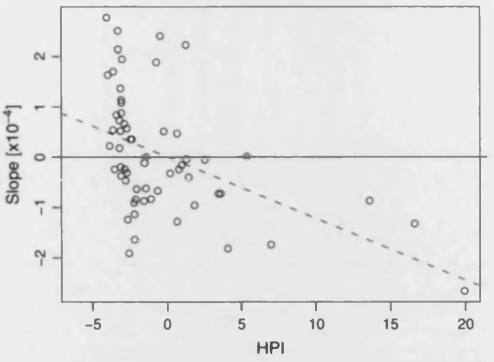


FIG. 1.—The relationship between HPI and trend to enrichment (negative slope) or avoidance (positive slope) of the codon near intron exon boundaries (Spearman rank correlation, $r = -0.5520$, $P = 7.025 \times 10^{-6}$).

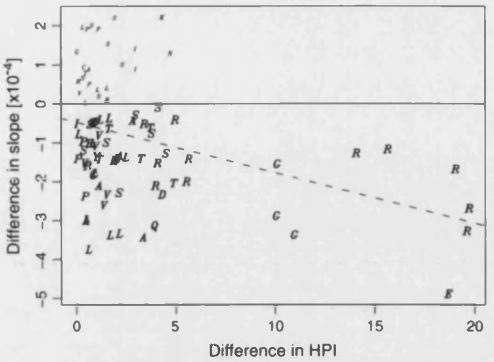


FIG. 2.—Pairwise analysis for all synonymous codon pairs comparing difference in HPI with difference in slope. Data for 3 and 5 prime exonic ends are merged. The amino acids of the analysis are represented by their own single letter symbol. Those in large font (below the solid line) are behaving as we would expect, where those codons more common in ESEs (greater HPI) are preferred near splice sites (more negative slope). By contrast, those in small font (above the solid line) are those that oppose this expectation. Spearman rank correlation, $r = -0.3098$, $P = 0.0036$.

enriched in ESEs compared with that expected given their content in the genome and given the underlying variance expected given the number of hexamers as input.

Is it generally the case that codons found commonly in splice enhancers are preferred near exon boundaries? To determine this we considered the correlation between HPI and α, the slope of the line describing codon abundance versus distance from the exon boundary. We find a robust general trend of this variety (Spearman Rank correlation, $r = -0.5520$, $P = 7.025 \times 10^{-6}$; fig. 1). This analysis, however, conflates amino acid preferences with codon preferences. To control for biases in amino acid usage near boundaries, a series of pairwise analyses were performed between synonymous codons. Under the splice enhancer model we would expect that the synonymous codon that is more abundant in ESEs would have a greater preference for usage near the splice sites (conversely, the one more profoundly avoided in enhancers should be more profoundly avoided near boundaries).

For a synonymous codon pair we considered the difference in slopes ($\Delta\alpha$), this difference being plotted against the difference in HPI (ΔHPI). We oriented the comparisons so as to ensure that the difference in HPI was always positive. For 2-fold degenerate amino acids a simple pairwise comparison was implemented, with one comparison for each amino acid. For amino acids with greater degeneracy, every pairwise permutation was assessed. If the ESE model is correct we expect that a negative correlation between these 2 parameters should exist. This indeed we observed: those codons more common in ESEs are more common near splice sites (Spearman rank correlation between $\Delta\alpha$ and ΔHPI, $r = -0.3098$, $P = 0.0036$; fig. 2). More generally, we expect that the codon relatively preferred near boundaries should be relatively enriched in the ESE data set. Indeed, 63 of 87 comparisons accord with this prediction.

Might the above result stem from the fact that small exons contribute more data to the slope calculation at positions near the boundaries? To examine this possibility, we restrict analysis to only those exons longer than 60 codons; so all exons contribute equally to all distances. We find no important differences in the results (see supplementary table 3, Supplementary fig. 1). The same data set also permits estimation of the magnitude of the difference in a codon's usage near and far from exon boundaries. To do this we consider the average of a codon's usage up to a distance of 5 codons from the boundary (excluding the first codon), and compare this to codons 30–33 inclusive (see supplementary methods 2). At the extreme, some codons (e.g. TAT, AGA) are approximately 30% more common close to boundaries, while others (e.g. CGC, CGG) are 30% less common (see supplementary table 4).

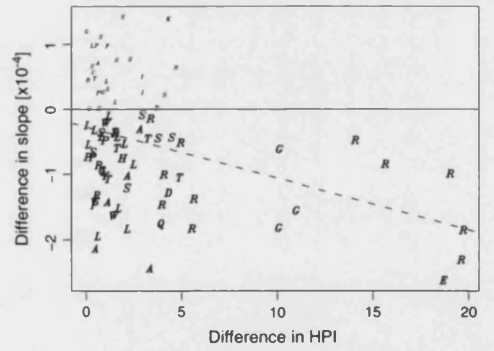


FIG. 3.—Pairwise analysis for all synonymous codon pairs comparing difference in SCPI with difference in slope. Data for 3 and 5 prime exonic ends is merged. The amino acids of the analysis are represented by their own single letter symbol. Those in large font (below the solid line) are behaving as we would expect, where those codons more common in ESEs and ESSs (greater SCPI) are preferred near splice sites (more negative slope. By contrast, those in small font, above the solid line are those that oppose this expectation. Spearman rank correlation, $r = -0.3398$, $P = 0.0014$.

The above results suggest that the need to specify efficient splice enhancers near intron-exon boundaries explains some of the variation in relative codon usage as one approaches intron-exon boundaries. There are, however, numerous exceptions (any data point with a difference in slope of greater than zero in figure 2 is unexpected). Are we able to explain the behaviour of those synonymous codons that go against our expectations by controlling for the presence of other splicing control elements within our analysis? Although ESEs are the most well studied and most prolific splice regulatory element, exonic splicing suppressors (ESSs) are also an important splice regulator (Wang et al. 2004), especially for alternative splicing events (Wang et al. 2006). A list of 133 decameric putative ESSs has been described in human (Wang et al. 2004) from which we were able to produce a DPI (Decamer Preference Index) for synonymous codons, in the same way as the HPI with minor necessary changes. The difference in DPI does not correlate with the difference in slope ($P = 0.23$). The indexes were then combined (the mean of the 2) to produce an overall Splice Control Preference Index (SCPI). The pairwise analysis of synonymous codon usage was repeated using the new SCPI as the indicator of codon representation in splice control elements (fig. 3). Although there are still comparisons that do not behave as we would expect, the overall trend is now a little stronger and more significant (Spearman rank correlation, $r = -0.3398$, $P = 0.0014$; fig. 3), suggesting that a combined suppressor/enhancer model provides a better fit. Under these conditions, 64 of 87 comparisons fit the expectation that the codon preferred near boundaries is more common in ESEs than their synonym.

Prior analysis comparing GAA with GAG argued that codons needed in splice enhancers are strikingly more abundant near exon boundaries (Willie and Majewski 2004; Chamary and Hurst 2005). More generally, in human and mouse genes, we commonly see trends in the usage of synonymous codons near intron-exon boundaries. However, from 2-fold degenerate amino acids, the previously noted GAA/GAG comparison (E on fig. 2) is by far the best support for the hypothesis. While there exists a general trend for those codons that are preferred near splice sites to be more common in human ESEs (higher HPI), the GAA/GAG comparison is perhaps misleading in the extent to which the trends are predictable.

Perhaps the most striking exception is lysine. This has 2 codons, AAA and AAG, but the one that is more abundant in splice enhancers (AAG) is the one that is relatively avoided near boundaries: both AAA and AAG are preferred near boundaries, in line with the observation that lysine is greatly preferred, but the slope for AAG is less negative than the slope for AAA. This may well be explained by the cryptic splice site avoidance model predicting, as it does, a force against AG ending codons that might be mistaken for exon ends (Eskesen et al. 2004; Chamary and Hurst 2005). In this regard it is notable that the GAA/GAG comparison matches both models, it just so happens that the synonym that might act as a cryptic splice site (GAG) is also the one less employed in splice enhancers.

In sum, we find that patterns of preference of codon usage as one approaches intron-exon boundaries are modulated in a manner that, to a first order approximation, is explained by splice control elements. This model however fails to explain all of the variation and numerous outliers remain.

Supplementary Material

Supplementary tables 1, 2, 3 and 4, Supplementary figure 1 and supplementary methods 1 and 2 are available at Molecular Biology and Evolution on-line (http://www.mbe.oxfordjournals.org/).

Acknowledgments

J.L.P. is funded by the United Kingdom Biotechnology and Biological Sciences Research Council.

Literature Cited

Akashi H, Eyre-Walker A. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev. 8:688–693.

Baralle M, Skoko N, Knezevich A, De Conti L, Motti D, Bhuvanagiri M, Baralle D, Buratti E, Baralle FE. 2006. NF1 mRNA biogenesis: effect of the genomic milieu in splicing regulation of the NF1 exon 37 region. FEBS Lett. 580: 4449–4456.

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

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. 2005. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet. 21: 256–259.

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

Comeron JM. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics. 167: 1293–1304.

dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 32:50365–044.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 12:640–649.

Eskesen ST, Eskesen FN, Ruvinsky A. 2004. Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. Genetics. 167:543–550.

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:W187–190.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2:13–34.

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with

CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol. 53:290–298.
Lander E, Linton SLM, Birren B, et al. (254 co-authors) 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.
Lavner Y, Kotlar D. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene. 345:127–138.
Pagani F, Raponi M, Baralle FE. 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci USA. 102:6368–6372.
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.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the Evolution of Proteins in Mammals. PLoS. Biol. 5:e14.
Raponi M, Baralle FE, Pagani F. 2007. Reduced splicing efficiency induced by synonymous substitutions may generate a substrate for natural selection of new splicing isoforms: the case of CFTR exon 12. Nucleic Acids Res. 35:606–613.
Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. Cell. 119:831–845.
Wang Z, Xiao X, Van Nostrand E, Burge CB. 2006. General and Specific Functions of Exonic Splicing Silencers in Splicing Control. Mol Cell. 23:61–70.
Willie E, Majewski J. 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet. 20: 534–538.
Wright SI, Yau CB, Looseley M, Meyers BC. 2004. Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. Mol Biol Evol. 21: 1719–1726.

Aoife McLysaght, Associate Editor

Accepted May 15, 2007

Chapter 5. **How common are intragene windows with Ka>Ks owing to purifying selection on synonymous mutations?**

Joanna L. Parmley and Laurence D. Hurst
Journal of Molecular Evolution (2007) 64(6): 646-655

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

JOURNAL OF MOLECULAR EVOLUTION

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 2001). 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 weak nonconsensus 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

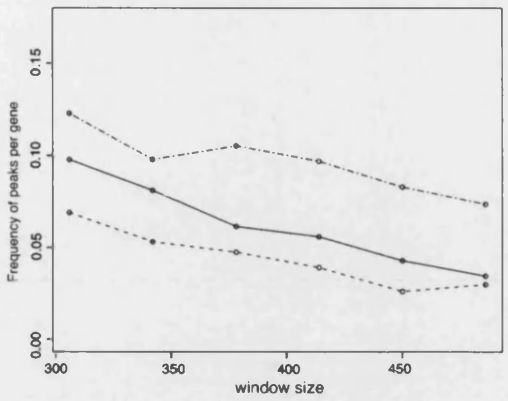


Fig. 1. The frequency of K_A/K_S peaks per gene that are due to either a peak in K_A, hence positive selection (dashedline), a dip in K_S, hence synonymous purifying selection (solidline), or an ambiguous cause where both events are concomitant (dash-dot line), as a function of the size of the sliding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

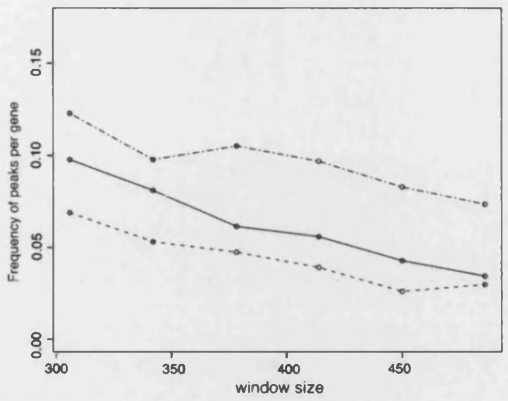


Fig. 1. The frequency of K_A/K_S peaks per gene that are due to either a peak in K_A, hence positive selection (dashedline), a dip in K_S, hence synonymous purifying selection (solidline), or an ambiguous cause where both events are concomitant (dash-dot line), as a function of the size of the sliding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

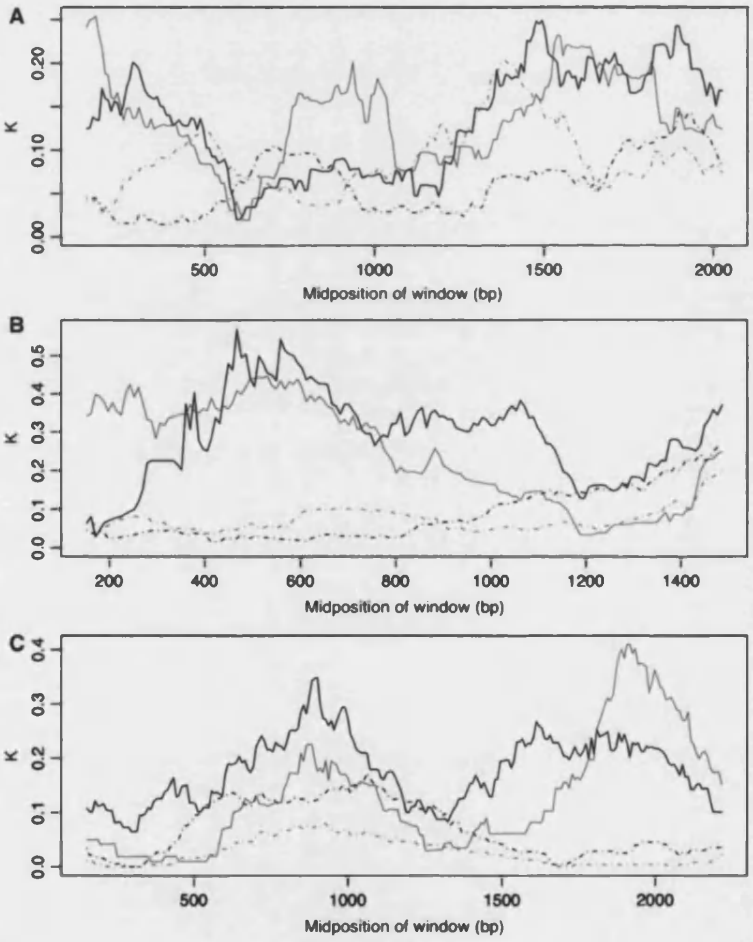


Fig. 2. Three genes with repeatable K_s dips at K_a/K_s peaks. **a** Retinoid X receptor interacting protein 110 (dip at ~600 nucleotides in alignment). **b** Chloride channel CLIC-like 1 (dip at ~1200 nucleotides in alignment). **c** Nuclear autoantigenic sperm protein (Nasp; dip at ~1300 nucleotides 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, $p = 0.88$).

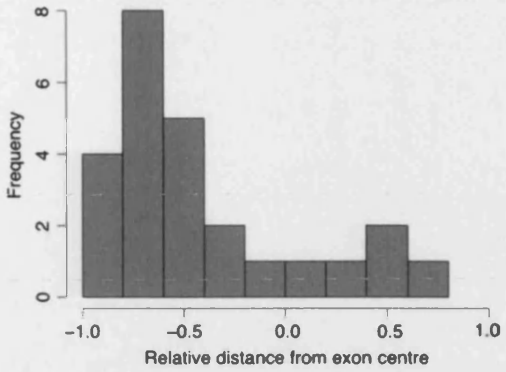


Fig. 3. The relative location of the center of the K_s dip window as a function of proximity to intron-exon junctions for those windows with no part of the sequence within 70 bp of an intron-exon window.

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 297:1007–1013

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 17:62–65

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 34:1700–1710

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

Chapter 6. **Splicing and the evolution of proteins in mammals**

Joanna L. Parmley, Araxi O. Urrutia, Lukasz Potrzebowski, Henrik Kaessmann and Laurence D. Hurst

PloS Biology (2007) 5(2): 98-108

OPEN ACCESS Freely available online PLOS BIOLOGY

Splicing and the Evolution of Proteins in Mammals

Joanna L. Parmley[1], Araxi O. Urrutia[1], Lukasz Potrzebowski[2], Henrik Kaessmann[2], Laurence D. Hurst[1*]

1 Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom, 2 Center for Integrative Genomics, Genopode, University of Lausanne, Lausanne, Switzerland

It is often supposed that a protein's rate of evolution and its amino acid content are determined by the function and anatomy of the protein. Here we examine an alternative possibility, namely that the requirement to specify in the unprocessed RNA, in the vicinity of intron–exon boundaries, information necessary for removal of introns (e.g., exonic splice enhancers) affects both amino acid usage and rates of protein evolution. We find that the majority of amino acids show skewed usage near intron–exon boundaries, and that differences in the trends for the 2-fold and 4-fold blocks of both arginine and leucine show this to be owing to effects mediated at the nucleotide level. More specifically, there is a robust relationship between the extent to which an amino acid is preferred/avoided near boundaries and its enrichment/paucity in splice enhancers. As might then be expected, the rate of evolution is lowest near intron–exon boundaries, at least in part owing to splice enhancers, such that domains flanking intron–exon junctions evolve on average at under half the rate of exon centres from the same gene. In contrast, the rate of evolution of intronless retrogenes is highest near the domains where intron–exon junctions previously resided. The proportion of sequence near intron–exon boundaries is one of the stronger predictors of a protein's rate of evolution in mammals yet described. We conclude that after intron insertion selection favours modification of amino acid content near intron–exon junctions, so as to enable efficient intron removal, these changes then being subject to strong purifying selection even if nonoptimal for protein function. Thus there exists a strong force operating on protein evolution in mammals that is not explained directly in terms of the biology of the protein.

Citation: Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD (2007) Splicing and the evolution of proteins in mammals. PLoS Biol 5(2): e14. doi:10.1371/journal.pbio.0050014

Introduction

Why do some parts of proteins evolve more slowly than others? Why, in turn, do some proteins evolve more slowly than others? Intragenic conserved regions are typically considered to reflect domains of functional importance to the protein [1]. Similarly, proteins with a high density of important functional sites should evolve slowly. There are, however, potentially multiple other correlates to rates of protein evolution [1]. The expression parameters of a gene (rate of expression, protein abundance, and number of tissues in which a gene is expressed) are consistently reported to be important predictors [2–5]. This may in part reflect selection to resist mistranslation [6]. Other possible covariates include essentiality and the number of protein interactions, but the issues here are more contentious, not least because of covariance with expression parameters [7–17]. Here we test the hypothesis that selection acting to ensure that introns are correctly removed skews amino acid content in predictable ways and imposes constraints on rates of protein evolution.

In mammalian genes, which are rich in introns [18], correct removal of introns often requires the presence, in the flanking exons, of splice-enhancer domains, these being short (six nucleotide) blocks required for binding of serine/arginine-rich proteins [19]. The need for splice enhancers can impact the use of synonymous codons in the domains flanking intron–exon junctions, such that when a synonymous codon is used commonly in splice enhancers it is preferred over its less commonly used synonym [20,21]. Moreover, selection to preserve splice enhancers affects both the synonymous single nucleotide polymorphism profile [22,23] and the rate of evolution at synonymous sites of splice-enhancer-associated domains [24].

Might the same forces also act to cause skews in amino acid usage in the vicinity of intron–exon junctions? In a preliminary analysis, we showed that there is a tendency for enrichment near boundaries of an amino acid whose codons are common in splice enhancers: lysine is coded by AAA and AAG, both of which are common in splice enhancers, and at both 5′ and 3′ ends of exons, lysine's proportional usage increases [24]. Is it more generally the case that an amino acid's usage increases near intron–exon junctions if it commonly features in splice enhancers? Conversely, are some amino acids avoided near such boundaries if they are rare in splice-enhancer domains? To address these issues, we derive patterns of amino acid preference in the vicinity of intron–exon boundaries and compare these patterns with a metric of enrichment of amino acids in splice enhancers relative to rates of usage in the genome. In turn, we ask whether selective constraints are stronger near intron–exon boundaries, and

Academic Editor: Kenneth H. Wolfe, University of Dublin, Ireland

Received August 9, 2006; **Accepted** November 13, 2006; **Published** February 6, 2007

Abbreviations: DPI, decamer preference index; ESE, exonic splice enhancer; HPI, hexamer preference index; K_A, the number of nonsynonymous substitutions per nonsynonymous site; K_S, the number of synonymous substitutions per synonymous site

* To whom correspondence should be addressed. E-mail: l.d.hurst@bath.ac.uk

Author Summary

Most of the DNA in our genes is actually not involved in the specification of proteins. Rather, the bits with the protein-coding information (exons) are separated from each other by noncoding bits, introns. Before a gene can be translated into protein these introns are removed and the exons are spliced back together to be translated into protein. While information about which DNA to remove is largely in the introns themselves, parts of the exons near the intron–exon boundary can, for example, function as splice enhancer elements. In principle, then, these parts of exons have two functions: to specify the amino acids of the resulting protein and to enable the correct removal of introns. What impact might this have on a gene's evolution? We show that near intron–exon boundaries, amino acid usage is biased towards nucleotides involved in splice control. Moreover, these parts of genes evolve especially slowly. Indeed, we estimate that a gene with many exons would evolve at under half the rate of the same gene with no introns, simply owing to the need to specify where to remove introns. Likewise, genes that have lost their introns evolve especially fast near the former intron's location. Thus, human proteins may not be as optimised as they could be, as their sequence is serving two conflicting roles.

whether such constraints explain much of the variation between proteins in their rate of evolution.

Results

Amino Acid Preferences near Intron–Exon Junctions Are Common

For 178,382 human exons we considered the trends in amino acid composition as one approaches the intron–exon boundary, as assayed by the rank correlation, rho, between distance from the boundary and proportional usage of the amino acid. Considering the 2-fold and 4-fold blocks of the 6-fold degenerate amino acids as different groupings, we found that of 46 independent comparisons (23 amino acid groups 5′ and 3′ prime), 34 showed significant trends for enrichment or avoidance near intron–exon boundaries (Table 1). After Bonferroni correction 26 remained significant (with 46 comparisons $p < 0.001$ indicates significance). For all plots for individual amino acids see Figure S1. We repeated the analysis for 115,466 exons from 14,005 mouse genes and found that patterns of preference are strikingly similar between the two species (Table 1). In mice, 34 amino acids again showed significant trends, and the correlation of rho values for 46 comparisons in mice versus human was extremely high (Pearson product moment correlation, $r = 0.96$, $p < 0.0001$).

Do these effects necessarily relate to the nucleotide content of the codons, as the splice-regulation model requires? One might conjecture instead that these effects reflect some coincidence of exon boundaries with protein substructures having unusual amino acid contents. Several facts strongly support the hypothesis that the trends seen are at least in part driven by effects at the nucleotide level. Notably, while the 2-fold block of arginine (amino acid r in Table 1) was strongly preferred near boundaries at both 3′ and 5′ ends, the 4-fold redundant block (amino acid R) showed the reverse pattern. A comparable difference was seen for the 2-fold (amino acid l) and 4-fold (amino acid L) blocks for leucine. The same pattern was seen in mouse genes. A preference for certain amino acids, regardless of the nucleotide content of their codons, would not have predicted this.

Amino Acid Preferences near Intron–Exon Junctions Are Predicted by Involvement in Splice-Enhancer Domains

If splice-enhancer domains impact amino acid usage near intron–exon boundaries, we expect that those amino acids preferred in splice enhancers should be preferred near junctions (i.e., rho < 0). To test this we developed a metric of involvement of codons in splice-enhancer hexamers, which we term the hexamer preference index (HPI). Using hexamers found both in mouse and human to define the HPI (and ignoring 3′ and 5′ differences), we found a striking predictability of patterns of preference near boundaries (Spearman rank correlation between HPI and rho for preference/avoidance near boundaries, rho = –0.54, $p < 0.0001$, $n = 46$). As an alternative to rho, we can employ the slope of the best-fit regression line between proportional usage of an amino acid and distance from intron–exon junctions. A negative slope, like a negative rho, indicates preferential usage near junctions. Using this slope on the best-fit regression line revealed, as expected, the same trend (Spearman rank correlation, slope versus HPI = –0.57, $p < 0.0001$; Figure S2). The trend for preference of high HPI amino acids near boundaries was also seen in mice (e.g., using mouse–human overlap set of hexamers, correlation of rho with HPI = –0.49, $p = 0.0005$; correlation of slope with HPI = –0.52, $p = 0.0002$).

These results are not greatly affected by considering 5′ and 3′ ends separately (Spearman rank correlation between rho 5′ and HPI 5′ using human 5′-specific hexamers = –0.59, $p = 0.003$, $n = 23$, Figure 1A; between rho 3′ and HPI 3′ using human 3′-specific hexamers = –0.57, $p = 0.004$, $n = 23$, Figure 1B). This is reflected in the fact that trends in usage (rho) and patterns of HPI are similar 5′ and 3′ (Pearson correlation, r, between rho 5′ and rho 3′ for the 23 amino acid classes = 0.80, $p < 0.0001$; Pearson correlation between HPI 5′ and HPI 3′ for the 23 amino acid classes = 0.95, $p < 0.0001$).

One might suppose that our measure of HPI might be biased by incomplete knowledge of enhancers. We can control for this, in part, by recognizing that splice enhancers tend to be adenine rich and cytosine poor. Consider then the composite measure AC bias = frequency of adenine in synonymous codon set – frequency of cytosine. For example, in the 4-fold degenerate set for alanine (GCN), of the 12 bases in four possible synonymous codons, adenine and thymine both featured 1/12 of the time, and guanine and cytosine both featured 5/12 of the time. So AC bias for alanine is 1/12 – 5/12 = –1/3. This AC bias was a robust predictor of preference/avoidance near boundaries (Spearman rank correlation, AC bias versus rho = –0.67, $p < 0.0001$) (Figure 2). Avoidance of cytosine in the synonymous codons appeared to be a somewhat stronger predictor of patterns of avoidance or preference of amino acids than was preference for adenine (Spearman rank correlation, cytosine content of codons versus rho = 0.67, $p < 0.0001$; adenine content of codons versus rho = –0.37, $p = 0.01$). Neither thymine nor guanine content showed any trends ($p >> 0.05$). These results suggest that the general profile of enhancers and the specifics employed to define HPI are about equally good predictors of patterns of preference/avoidance.

Table 1. Trends in Avoidance of (rho > 0) or Preference for (rho < 0) Amino Acids as a Function of Distance from the Intron–Exon Junction

Amino Acid	DPI	HPImh	5′					3′				
			Human		Mouse			Human		Mouse		
			rho	*p*	rho	*p*	HPI	rho	*p*	rho	*p*	HPI
A	2.535	−5.81	0.866	1.36E−07	0.8118	4.81E−07	−4.89	0.661	4.32E−05	0.6404	8.63E−05	−5.35
C	−3.18	−3.92	0.095	0.59	−0.187	0.30	−2.83	0.140	0.436	0.0495	0.78	−3.99
D	6.664	2.596	−0.499	0.0035	−0.496	0.0037	1.852	−0.578	0.0005	−0.59	0.0004	2.85
E	20.07	13.69	−0.642	8.31E−05	−0.636	9.87E−05	8.125	0.125	0.48	0.0996	0.58	12.42
F	−12.4	−2.53	−0.520	0.002	−0.652	5.97E−05	−2.2	−0.757	1.40E−06	−0.768	1.07E−06	−2.4
G	−17.1	−1.33	−0.058	0.75	0.1624	0.36	−0.57	0.301	0.0886	0.3168	0.073	−2.06
H	0.528	−3.39	0.607	0.0002	0.6628	4.08E−05	−1.49	−0.202	0.26	−0.194	0.28	−3.85
I	1.211	−1.83	−0.830	3.54E−07	−0.784	7.71E−07	−1.57	−0.839	2.88E−07	−0.783	7.85E−07	−1.1
K	17.23	13.93	−0.881	6.95E−08	−0.88	7.61E−08	10.28	−0.936	0	−0.891	3.48E−08	12.45
L	−1.1	−5.83	0.279	0.115	0.2102	0.24	−3.91	0.505	0.003	0.1705	0.34	−5.4
M	5.054	3.471	−0.628	0.00013	−0.528	0.0018	3.358	−0.446	0.00980	−0.53	0.0018	1.943
N	8.652	4.355	−0.582	0.0005	−0.699	1.09E−05	2.625	−0.590	0.0004	−0.572	0.00063	5.846
P	−1.03	−5.83	0.617	0.00018	0.618	0.00017	−4.14	0.660	4.42E−05	0.6731	2.83E−05	−5.43
Q	7.914	1.758	0.874	9.77E−08	0.8078	5.14E−07	0.186	0.440	0.011	0.5084	0.0028	3.941
R	−1.2	−3.81	0.875	9.34E−08	0.9358	0	−2.96	0.959	0	0.8971	1.59E−08	−3.89
S	−1.91	−3.41	0.476	0.005	0.4174	0.016	−1.58	0.450	0.0091	0.4856	0.0046	−2.82
T	6.044	−0.27	0.723	4.45E−06	0.5993	0.0003	−0.14	−0.257	0.15	−0.109	0.54	1.698
V	−23.8	−5.7	−0.175	0.33	−0.293	0.010	−3.29	0.391	0.025	0.4081	0.0191	−5.35
W	−1.42	1.253	−0.069	0.71	−0.238	0.18	2.002	−0.125	0.49	−0.153	0.392	0.32
Y	−3.22	−3.55	−0.055	0.759	−0.443	0.01	−1.32	−0.376	0.033	−0.218	0.222	−2.89
l	−17.4	−2.47	−0.958	0	−0.951	0	−1.2	−0.728	3.67E−06	−0.805	5.41E−07	−2.79
s	−1.26	−1.56	0.795	6.32E−07	0.7985	5.98E−07	−2.85	0.791	6.75E−07	0.877	8.63E−08	−1.6
r	12.41	13.15	−0.696	1.22E−05	−0.582	0.00050	9.352	−0.840	2.84E−07	−0.717	5.54E−06	10.17

Also specified is the HPI for each amino acid using hexameric data specific to human exonic ends (HPI) and, alternatively, using the set of hexamers reported in both mouse and human regardless of end (HPImh). The figures for HPImh were derived using human codon frequencies as expected. Using mouse frequencies shows a highly similar pattern (Pearson *r* between HPImh using human versus mouse codon frequencies, $r = 0.999$). rho and *p* were calculated from Spearman rank correlation with 31 degrees of freedom (i.e., from 33 data points, representing the codons up to 34 away from the boundary but excluding the first). DPI is the comparable index but for decameric splice suppressors.

doi:10.1371/journal.pbio.0050014.t001

Rates of Evolution Are Reduced near Intron–Exon Boundaries and in Genes Rich in Introns

The above results suggest that selection acts to prefer nucleotides that permit efficient intron removal. Does this in turn affect rates of protein evolution? Were there such an effect, we should expect that smaller exons should evolve more slowly, as a higher proportion of sequence is near (e.g., within 70 bp) boundaries. Indeed, from a set of 36,683 mouse–human aligned exons, we found that small exons do tend to have low rates of evolution (Spearman rank correlation between the number of nonsynonymous substitutions per nonsynonymous site [K_A] and exon length, rho = 0.15, $p < 0.0001$). This might, however, be owing to a trend for genes with small exons to be disproportionately in functional classes of protein that have intrinsically low rates of evolution. To control for this we considered, for all genes with more than two internal exons, the Spearman rank correlation between exon K_A and the size of the exon. As each correlation coefficient is derived from a given gene, between-gene variation in K_A (and indeed the number of synonymous substitutions per synonymous site [K_S]) is controlled for in any such analysis. If splice control impacts rates of exon evolution we expect that on the average this correlation should be positive, while the null hypothesis, that small exons have low rates of evolution because they derive from classes of genes with intrinsically low K_A, predicts a mean rho of zero. The distribution of rho was very strongly skewed to positive values (median rho = +0.14, Wilcoxon rank test, $p < 0.0001$, $n = 3{,}629$). Restricting analysis to genes with ten or more exons only strengthened this conclusion (median rho = +0.16, $p < 0.0001$, $n = 1{,}286$).

Is there also a trend for lower rates of evolution near boundaries? Using all exons, asking about the proportion of all sites a given distance from a boundary (5′ or 3′) in which we see a nonsynonymous change, we observed the predicted low rate of amino acid evolution near boundaries (Spearman rank correlation, proportion of aligned sites showing non-synonymous change versus distance from boundary, rho = 0.955, $p < 0.0001$) (Figure 3, circles). Might this result simply be an artefact of the possibility that small exons might both come disproportionately from a class of slow-evolving genes and contribute more data to the estimate of divergence near the exon–intron junctions than they do to the more distant sites? To control for this, we again considered divergence rates within 40 codons of boundaries (5′ and 3′) but considered only the 1,836 exons that are at least 80 codons long. This way all exons contribute approximately the same amount of data at all distances from the junction. We found that the lower rate of evolution near the boundary remained highly robust (rho = 0.7685, $p < 0.0001$) (Figure 3, squares).

Note, however, that absolute rates of evolution, at any given distance from the boundary, were higher in this long exon set. This is consistent either with reduced density of splice-control elements near boundaries in long exons or with a splice-unrelated force acting more profoundly on long exons. There is good evidence for the former. When we examined

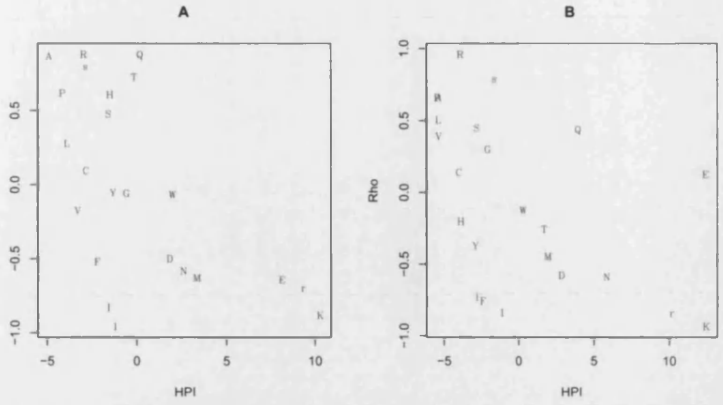


Figure 1. The Relationship between Tendency for an Amino Acid to Be Preferred near Exon–Intron Junctions (rho < 0) or Avoided (rho > 0) and the HPI
(A) 5′ exonic ends and (B) 3′ ends.
doi:10.1371/journal.pbio.0050014.g001

the density of putative exonic splice enhancers (ESEs) in the exon span within 100 bp of a boundary at either end (or all of the exon in the case of exons shorter than 200 bp), we found a robust negative correlation between enhancer density and exon size (rho = −0.18, $p < 0.0001$). Comparably, when we considered exons longer than 200 bp to be long exons and those shorter than this to be short exons, we found that ESEs occupy a median of 31% of the short exons, but only 21% of the 200 bp near the boundaries (100 bp 5′ and 100 bp 3′) of the long exons. This is consistent with the idea that there is less space in short exons to pack in the information necessary to enable proper splicing.

As expected, K_A was lower in ESEs than in nonenhancers (Figure 4) (see also [24]). This was also true if we restricted analysis to exons longer than 200 bp (paired test, $p < 0.0001$) (Figure S3). These results tally with the finding that genes with long introns tend to have low rates of evolution [12], as exons flanked by long introns tend to be richest in ESEs [25].

As expected from the above results, genes with a high proportion of sequence within, for example, 70 bp of an intron–exon junction showed lower K_A (Table 2; Figure 5). Using alternative bounds (50 or 100 bp) did not qualitatively affect conclusions (Table 2). The difference between a gene with all sequence within 70 bp of exon boundaries and one with very little (<10%) was striking (mean $K_A = 0.032$ for those with small exons and 0.083 for those with less than 10% of sequence near junctions). Were all things equal, this result suggests that the rate of evolution of an intron-rich gene is on average approximately under 40% that of an intron-poor gene.

It is, however, unlikely that all things are equal. To allow for this, we performed a paired test. For each gene we concatenated all sequences in the alignment flanking (within 72 nucleotides) intron–exon boundaries, both 5′ and 3′, and concatenated all of the middle sections of exons (defined as anything beyond 72 nucleotides). As before we considered only internal exons. We then calculated K_A for the concatenated flanks and the concatenated middle sections and considered the gene-specific ratio of the two. We then considered the mean of the gene-specific ratio for all genes.

By necessity we had to eliminate all genes with no exon larger than 144 bp, leaving 3,058 genes. Moreover, as accurate estimation of K_A probably requires a minimum of 100 codons, we restricted analysis to those genes with at least 300 bp in the concatenated flanks and in the concatenated middle of exons. We found that the mean ratio of the rate of evolution (K_A) of the middle part of exons to the flanks within the same gene was 1.93 (Wilcoxon signed rank test, $p < 0.0001$, $n = 666$). Requiring at least 600 bp in both flanks and middle sections, the middle was estimated to evolve 2.3 times faster than the flanks. When we considered the exon flanks to be 102 bp, the mean ratio of middle to flank was 2.5 when requiring a minimum of 300 bp in each class ($n = 368$). Requiring a minimum of 600 bp, the middle parts of exons evolved on average 2.7 times faster than the exon flanks from the same genes ($n = 167$). Overall, then, it seems safe to conclude that exon centres evolve at about 2.3 times the rate of exon flanks from the same gene, the precise estimate depending on parameter choices.

These results demonstrate that exon flanks evolve more slowly than exon centres, regardless of the functional class of the protein. The mean K_A of flanking domains was around 0.04 in the above samples. A gene with short exons should then have approximately a K_A of 0.04, controlling for between-gene heterogeneity. By contrast one with 90% of sequence not near boundaries should have a K_A of on average around 0.086, assuming exon centres of such long exons evolve 2.3 times faster than flanks ($0.04 \times 2.3 \times 0.9 + 0.04 \times 0.1 = 0.086$). Controlling then for functional class, we estimated that a gene with all sequence near intron–exon boundaries should evolve at about 46% (0.04/0.086) the rate of one with proportionally little sequence near boundaries.

This estimate can be downwardly adjusted if we consider that some of the genes with long exons have more than 90% of sequence near boundaries: at the limit intronless genes should evolve with $K_A \cong 0.092$, i.e., at 2.3 times the rate of small exon genes. Likewise, if our estimate of the ratio of rates of evolution is higher, then the discrepancy between intron-poor and intron-rich genes will be greater. Using the 2.7 ratio, for example, intron-rich genes evolve at 37% of the

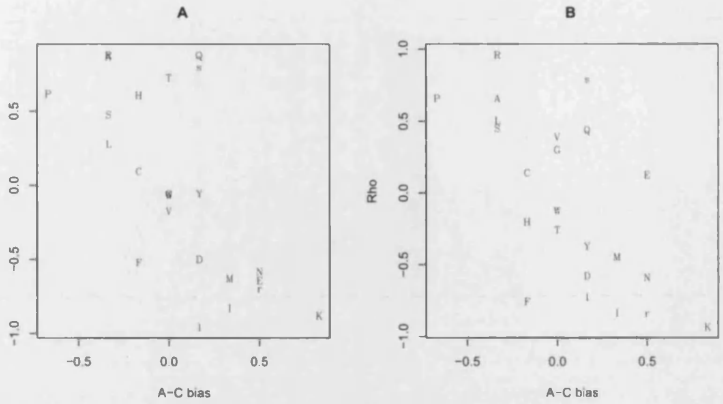


Figure 2. AC Bias in the Codon Set of a Given Amino Acid and Its Relationship to Amino Acid Usage near Exon–Intron Junctions
(A) 5′ exonic ends and (B) 3′ ends.
doi:10.1371/journal.pbio.0050014.g002

rate of intronless genes, controlling for protein function. Equally, the estimate can be upwardly adjusted if we presume a more modest ratio of rates of evolution of internal parts of exons to flanks. Overall, it seems fair to suppose that constraints imposed in the proximity of intron–exon boundaries can reduce the rate of evolution of a gene by a half or more, if the gene is full of small exons rather than lacking introns. That this is similar to the prior estimate, not controlling for between-gene heterogeneity, suggests that selection on exon flanks is a major determinant of rates of evolution.

Comparing Constraints Owing to Splicing with Other Correlates of Rates of Evolution

How does the effect of selection in the vicinity of intron–exon junctions compare with and covary with other strong predictors of rates of protein evolution? In principle any relationship between rate of protein evolution and proportion of sequence near a boundary might in part be because genes with many introns tend to be housekeeping genes [26], and housekeeping genes (those expressed in many tissues) tend to have low rates of evolution [4,27,28]. The two parameters (expression breadth and proportion of sequence near boundaries) both appear, however, to be good predictors when controlling one for the other (Table 2). Use of alternative metrics of gene expression (mean rate and peak rate) (see Table 2) make no qualitative difference to the conclusion that, before and after control for covariates, the proportion of sequence near intron–exon junctions is at least as strong a predictor of rates of evolution as expression parameters, if not stronger.

After expression parameters, the dispensability of a protein may, in mammals, also be a good predictor [12]. From a sample of 1,198 mouse genes for which knockout experiments have resolved whether they are essential or not, and for which we have orthologues, we can ask whether essential and nonessential genes (a) differ in their proportion of sequence near intron–exon junctions and (b) differ in their rate of evolution. Confirming the prior report [12], we found that essential proteins evolve at about two-thirds the rate of nonessential ones (mean K_A for nonessential proteins, 0.07; for essential proteins, 0.049; $p < 0.0001$, Mann-Whitney U test). However, the two classes are no different as regards the proportion of sequence near intron–exon boundaries (mean proportion of sequence near boundaries for nonessential proteins, 0.618; for essential proteins, 0.607; $p = 0.67$, Mann-Whitney U test). There is, therefore, no reason to suppose that the lower rate of evolution of genes with much sequence near intron–exon boundaries is owing to their being more likely to be essential. Equally, there is no reason to suppose that the lower rate of evolution of essential genes is owing to their having more sequence near intron–exon boundaries. Note too that the difference in evolutionary rate between essentials and nonessentials is more modest than that between genes with high and low proportion of sequence near intron–exon junctions. The majority of our sample is of unknown dispensability. These genes have a mean K_A of 0.059, more or less as expected, given the means for the essential and nonessential genes and assuming that 30% of mouse genes are essential [12].

Retrogenes and Loss of Selective Constraint near Intron–Exon Junctions

Let us now consider two models for what might happen after a new intron has been inserted. In the first, a new intron might be favoured only if enough splice-enhancer domains in adequate proximity are already present to enable efficient removal of the intron (model 1). An alternative model (model 2) might suppose that immediately after introduction of a new intron, proper excision, owing to a dearth of local splice enhancers, is not always possible. If, then, some transcripts preserve the original mRNA by proper excision, but others fail to so do, the new intron would effectively reduce the rate of protein production for a given transcription rate. Such a mutation might be weakly deleterious such that fixation through drift is still possible. Selection may then favour shifts in amino acid usage to enable more efficient splicing. The second model is especially interesting as it suggests that intra-protein amino acid usage is not dictated simply by protein requirements alone.

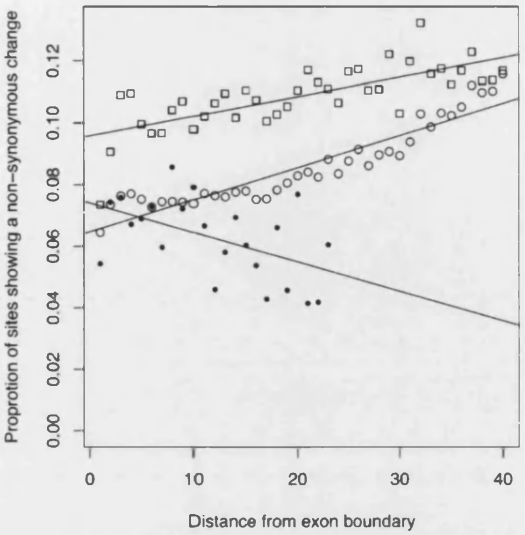


Figure 3. Rate of Nonsynonymous Evolution as a Function of the Distance from an Intron–Exon Boundary

The proportion of informative sites in intron-containing genes showing a nonsynonymous change in the human–mouse comparison (all exons, circles; exons > 80 codons, squares), and the proportion of informative sites in retrogene sequences showing retrogene-specific changes as a function of distance from what was originally the exon–intron boundary (black spots) and as a function of the distance from the real exon boundary.

doi:10.1371/journal.pbio.0050014.g003

Both models predict that should enhancer domains be employed, they may then be under selection to preserve functionality. Both also predict that amino acids that feature commonly in the hexameric sequences describing splice enhancers should be more common near intron–exon junctions, as observed. How they differ is in the prediction of subsequent evolution following gain/loss of introns. Model 1 supposes that if an intron inserts but is not successfully removed owing to a dearth of splice-enhancer domains in the near vicinity, the insertion may simply be too deleterious to be tolerated and is hence lost from the population. By contrast, model 2 considers the possibility that compensatory nonsynonymous changes can further occur that permit more efficient intron removal.

To discriminate these two classes, one needs a sufficiently sized dataset of intron losses or gains in humans. Unfortunately, intron gain appears to be vanishingly rare in humans and mammals more generally. However, functional retroposed genes do provide a means to ask about the consequences of intron loss. Is it then the case that, after retroposition, the residues that, in the original parental copy of the gene, flanked intron–exon junctions are more prone to change?

We examined a set of 49 old functional retroposed genes for which, in all cases, there existed mouse and human parent and retroposed sequences. For all sites in the alignment that specified an amino acid in all four lineages, we considered the proportion of retrogene-specific changes (see Materials and Methods). We then considered how this varied as a function of the distance from what was, in the parental gene, the intron–exon boundary. Merging figures for 3′ and 5′ ends, we found that the rate of evolution in retrogenes is higher close to what was the boundary (Spearman rank correlation, proportion of sites subject to change in retrogenes versus distance from ancient boundary, rho = −0.48, $p = 0.019$) (Figure 3). Moreover, retrogenes that are derived from genes in which a high proportion of the sequence was near exon boundaries (genes with predominantly small exons) tended to have higher overall rates of evolution (proportion of parent sequence 70 bp from boundary versus number of retrogene-specific changes per base pair, rho = +0.38, $p = 0.008$, $n = 49$).

The difference in behaviour between genes that have lost their introns and intron-containing genes (Figure 3) suggests that constraints that exist near intron–exon boundaries have

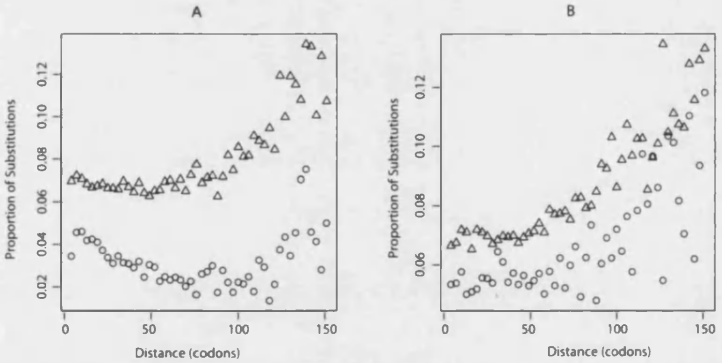


Figure 4. Rate of Nonsynonymous Evolution as a Function of the Distance from an Intron–Exon Boundary for ESS and Non-ESS Sequence

The rate of evolution of sequences defined as part of ESSs (circles) and those not in enhancers (triangles) is shown as a function of the distance from exon boundaries in the mouse–human analysis at (A) the 5′ end of exons and (B) the 3′ end of exons. To define putative enhancer sequence the mouse and human sequence was matched to the set of species-specific, exon-end-specific set of hexamers. Any part of the alignment not found to be enhancer in either species was considered nonenhancer. Any part of the alignment found to be enhancer in both was considered to be enhancer sequence. As can be seen, exonic enhancer sequence evolves more slowly than nonenhancer. Given that functional splice enhancers are rare more than 100 bp from a boundary, it is expected that the further into the exon, the less the difference between enhancer and nonenhancer.

doi:10.1371/journal.pbio.0050014.g004

Table 2. Correlations and Partial Correlations between Rate of Protein Evolution (K_A or K_A/K_S), Proportion of Sequence within 50, 70, or 100 bp of an Intron–Exon Junction, and Measures of Expression of the Relevant Gene in Humans

X	Y	Z	r_{XY}	$r_{XY\|Z}$	$p_{XY\|Z}$	r_{XZ}	$r_{XZ\|Y}$	$p_{XY\|Z}$	r_{YZ}
K_A	Proportion 50	Breadth	−0.1984	−0.1603	1.00E−04	−0.2019	−0.1647	1.00E−04	0.2250
K_A	Proportion 50	Median	−0.2055	−0.2015	1.00E−04	−0.0942	−0.0849	1.00E−04	0.0549
K_A	Proportion 50	Peak rate	−0.2055	−0.2037	1.00E−04	−0.0368	−0.0246	0.12669	0.0621
K_A/K_S	Proportion 50	Breadth	−0.2064	−0.1719	1.00E−04	−0.1858	−0.1462	1.00E−04	0.2250
K_A/K_S	Proportion 50	Median	−0.2175	−0.2140	1.00E−04	−0.0827	−0.0726	1.00E−04	0.0549
K_A/K_S	Proportion 50	Peak rate	−0.2175	−0.2165	1.00E−04	−0.023	−0.0097	0.32067	0.0621
K_A	Proportion 70	Breadth	−0.2007	−0.1639	1.00E−04	−0.2019	−0.1654	1.00E−04	0.2181
K_A	Proportion 70	Median rate	−0.2066	−0.2015	1.00E−04	−0.0942	−0.082	1.00E−04	0.0685
K_A	Proportion 70	Peak rate	−0.2066	−0.2046	1.00E−04	−0.0368	−0.0219	5.00E−04	0.0748
K_A/K_S	Proportion 70	Breadth	−0.2115	−0.1783	1.00E−04	−0.1858	−0.1464	1.00E−04	0.2181
K_A/K_S	Proportion 70	Median rate	−0.2219	−0.2175	1.00E−04	−0.0827	−0.0694	0.00060	0.0685
K_A/K_S	Proportion 70	Peak rate	−0.2219	−0.2208	1.00E−04	−0.023	−0.0066	0.3817	0.0748
K_A	Proportion 100	Breadth	−0.2030	−0.1646	1.00E−04	−0.2019	−0.1633	1.00E−04	0.2278
K_A	Proportion 100	Median	−0.2068	−0.1989	1.00E−04	−0.0942	−0.0747	0.00030	0.104
K_A	Proportion 100	Peak rate	−0.2068	−0.2041	1.00E−04	−0.0368	−0.0152	0.23318	0.1065
K_A/K_S	Proportion 100	Breadth	−0.2138	−0.1793	1.00E−04	−0.1858	−0.1441	1.00E−04	0.2278
K_A/K_S	Proportion 100	Median	−0.2227	−0.2160	1.00E−04	−0.0827	−0.0614	0.00220	0.104
K_A/K_S	Proportion 100	Peak rate	−0.2227	−0.2216	1.00E−04	−0.023	7.00E−04	0.4889	0.1065

The first three columns in each row indicate which variables are the *X*, *Y*, and *Z* variables. The subsequent columns indicate the correlations between *X* and the other two variables (r_{XY}, r_{XZ}) and the partial correlation ($r_{XY|Z}$ indicates the partial of *X* versus *Y* controlling for *Z*). *p*-Values indicate the significance of the partial correlation determined by 10,000 randomizations. Spearman rank correlation was employed throughout. The expression data were derived from Su et al.'s array-based analysis [35]. Breadth is the number of tissues in which a gene was expressed (defined by presence/absence calls). The median rate for a gene is the median value of the signal sampled across all tissues in which the gene is considered to be expressed. The peak rate is the highest level of expression for a given gene across all tissues. For the comparable data employing mouse expression data see Table S1.

doi:10.1371/journal.pbio.0050014.t002

been released in the retrogenes, and, hence, that these sites are now free to change. This evidence, therefore, lends some support to the converse possibility, namely that, after intron insertion, exonic domains flanking the new boundary changed, probably to permit better splicing. The result does not specifically show that all the change involved the evolution of new splice enhancers; however, with the data showing that the HPI predicts trends in amino acid usage near junctions and low nonsynonymous rates in ESEs (Figure 4) [24], this is likely to explain much of the effect.

Discussion

We have found that, in both mouse and human, most amino acids show skewed usage in the vicinity of intron–exon junctions. These patterns appear owing to preference at the nucleotide level, as evidenced by the different behaviours of the 2-fold and 4-fold blocks of leucine and arginine. To a first approximation, the patterns are well explained by the abundance of the relevant codons, relative to levels in the genome, in splice enhancers. The preferences are also reflected in reduced rates of evolution near intron–exon boundaries and in intron-rich genes more generally. Indeed, the proportion of sequence near intron–exon boundaries is, to the best of our knowledge, one of the strongest predictors to date of rates of protein evolution (for analysis of alternatives see [12]). That in retrogenes the domains that used to be near intron–exon junctions show increased rates of evolution supports the view that intron–exon junctions are domains on which constraint operates. Were it the case that new introns are only tolerated if the full repertoire of splice-control elements is already in place, we would not expect that, on loss of introns, these domains would show unusually high rates of evolution. Although by necessity our sample size of retrogenes is small, we suggest that model 2, evoking evolution to modify amino acid content after intron insertion, is more parsimonious.

Whether the elements being preferred are necessarily and exclusively splice enhancers remains uncertain. First, as can be seen in Figure 4, sequence putatively not in enhancers is more highly constrained near boundaries, at least at the 3′ end. This suggests the possibility of constraint imposed near boundaries independent of splice enhancers and/or inaccuracy in the definition of enhancers. Further, there are a few strong outliers in the distribution of HPI versus preference near boundaries (Table 1). In human sequences, of 46 comparisons, 14 fail to match with the expectation that if HPI is negative, rho should be positive and vice versa, of which nine are significant and six significant after Bonferroni correction: I5′, I5′, Q5′, F3′, I3′, and I3′ (Table 1). Glutamine (CAA and CAG) is unique in being preferred in splice enhancers and avoided both 3′ and 5′ at boundaries. Three amino acids are strongly preferred near boundaries (rho << 0) but disfavoured in splice enhancers (HPI < 0), these being the 2-fold degenerate codons of leucine (TTA and TTG), isoleucine (ATC, ATA, and ATT), and phenylalanine (TTC and TTT). Tyrosine (TAC and TAT) may be a weaker outlier (rho < 0 both 5′ and 3′, HPI < 0). The same outliers are seen in mouse genes (Table 1).

Are these apparent exceptions instructive of some other force driving amino acid choice near boundaries, or might they reflect limitations in our understanding of splice-enhancer hexamers? Were the latter the case we might expect that a surrogate measure of involvement in splice enhancers might reveal these exceptions to simply have poorly described roles in splice enhancers. As noted above adenine and cytosine content of the synonymous codon blocks of each amino acid well predicts HPI (Figure 2). Fitting the best-fit

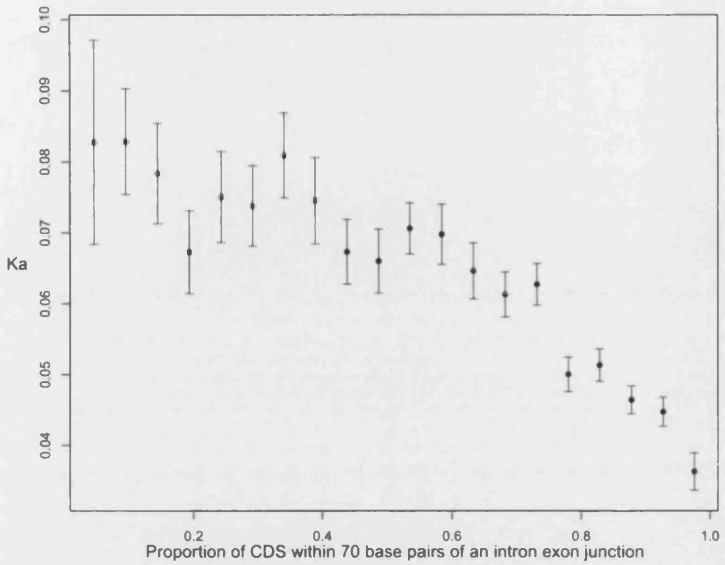


Figure 5. The Relationship between K_A in the Mouse–Human Comparison and the Proportion of Sequence within 70 bp of an Exon–Intron Junction
Error bars show standard error of the mean. CDS, coding sequence.
doi:10.1371/journal.pbio.0050014.g005

regression of AC bias to rho (using both 5′ and 3′ data), we indeed find from inspection of the standardised residuals (Figure S4) that, both 3′ and 5′, isoleucine and leucine usage now sit within the 95% confidence intervals, as does phenylalanine usage 5′. However, phenylalanine usage 3′ is a little outside the line, as is glutamine 5′ usage.

Another possibility is that the presence of exonic splice suppressors may impact amino acid usage. Wang et al. [29] have identified 131 decamers that function as splice suppressors. We therefore adapted our method to calculate a decamer preference index (DPI) to correspond with these splice suppressors (Table 1). DPI and HPI are not themselves correlated (for mouse–human set for HPI, Spearman rank correlation between HPI and DPI = –0.05, $p = 0.7$). Relating DPI scores to either the slope or the rho values for amino acid preference, we find only a marginal tendency for DPI to explain amino acid preferences (Spearman rank correlation, rho versus DPI, –0.27, $p = 0.07$; slope versus DPI, –0.26, $p = 0.07$). Splice suppressors hence appear to have less impact on amino acid usage than do splice enhancers. Taking a combined measure, the mean of DPI and HPI, marginally improves the fit between amino acid preference and involvement in splice regulation (Spearman rank correlation between mean of DPI and HPI and rho, –0.61, $p < 0.0001$; for HPI alone, –0.54, $p < 0.0001$). AC bias remains a better predictor. Involvement in splice suppressors may, however, explain some of our apparent exceptions. Notably, phenyalanine and the 2-fold block of leucine, while having a negative HPI, have a strongly positive DPI (9.8 and 14.9, respectively). Similarly, glutamine, while having a positive HPI, has a strongly negative DPI (–6.1). The converse roles of these amino acids in splice enhancers and splice suppressors may hence explain their apparently aberrant behaviour. Indeed, on a plot of the mean of DPI and HPI these amino acids no longer appear as outliers (Figure S5). Isoleucine remains an exception, being negative for both HPI and DPI but preferred near boundaries.

The only other model for selection near intron–exon junctions, the so-called cryptic splice-site avoidance model [21,30], does not predict any tendency for cytosine avoidance near boundaries. The relevance of this model is unclear as both AG[A|G] (arginine) and AG[C|T] (serine) appear to have patterns of usage near boundaries at both 5′ and 3′ ends as expected given their HPI scores, whereas the cryptic splice-site avoidance model would predict avoidance at 5′ ends. This model cannot also obviously explain why 3′ usage of phenyalanine might be discordant.

One further striking peculiarity is notable. The profile of usage of glycine (GGN) shows a curious pattern at both 3′ and 5′ ends (also seen in mouse, data not shown) (Figure 6): at every third codon the usage is much higher than at the intervening distances from the boundary. With the sample sizes in question (~10,000 glycines at these positions), this is not a sample-size artefact. The effect is highly repeatable, being found regardless of the phase of the exon (Figure S6). At both the 3′ and 5′ ends, it is found for all of the four (GGN) codons when analysed separately, although it may be most pronounced for GGA (data not shown). This appears to reflect a pattern at the protein level, at least in part owing to collagens, whereby glycines are very commonly three apart (see Figure S7). Given that introns tend to prefer G|G insertion sites, codons starting GG may well be hot spots for insertion, potentially at all positions. This together with the apparent periodicity in the occurrence of glycine might explain the observations. We leave this to future analysis. Whatever the cause, it points to a limitation of our method,

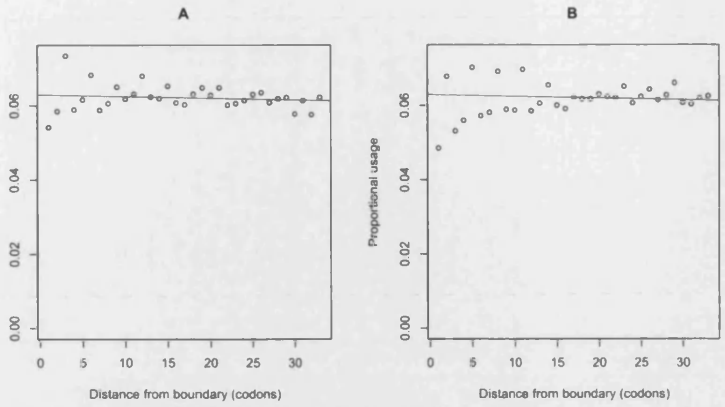


Figure 6. Glycine Usage as a Function of Distance from 5′ and 3′ Exonic Ends
(A) 5′ and (B) 3′ exonic ends.
doi:10.1371/journal.pbio.0050014.g006

which assumes that trends towards boundaries are monotonic and consistent. For the most part (see Figure S1) these assumptions appear relatively sound, although 5′ usage of proline suggests a U-shaped function.

The hypothesis that the domains under constraint are uniquely splice enhancers might also predict that amino acids not having a role in splice enhancers tend to be gained in retrogenes in boundary proximal domains. Unfortunately, from a sample of 803 gains/losses for retrogenes and 229 in parental genes in regions near intron–exon junctions (<30 codons), we find no amino acid showing statistically significant differences between parental and retrogenes. However, the top three most discordant amino acids (judged by the chi-squared value) all show net gain in the retrogenes and net loss in the parental genes, and, as might be predicted, are all avoided in splice enhancers. These are the 4-fold block of leucine (49 gains to 39 losses in retrogenes; nine losses to 18 gains in parental genes; chi-squared = 4.13), histidine (20 gains to 13 losses in retrogenes; three gains to seven losses in parental genes; chi-squared = 2.89), and the 4-fold block of serine (48 gains to 29 losses in retrogenes; 14 gains to 17 losses in parental genes, chi-squared = 2.66; N.B., for three degrees of freedom $p < 0.05$ occurs at chi-squared > 7). It would be unwise to read too much into this observation, not least because there are several other amino acids with strong avoidance in splice enhancers that show no evidence of switching substitutional profile (notably alanine, cysteine, phenylalanine, and valine). No amino acids show any good evidence for being gainers in the parental gene but losers in the retrogene. Firmer conclusions regarding the patterns of amino acid loss and gain will require larger sample sizes.

Given the outliers being possibly explained by splice-suppressor roles and the strange behaviour of glycine, we do not wish to suggest that the need for splice enhancers determines all amino acid bias, nor all constraint, seen near intron–exon boundaries. Constraints operating near intron–exon boundaries not explained by splice enhancers may nonetheless reflect selection on splice regulation of some form (e.g., exonic splice suppressors). These caveats aside, it is notable that constraints in the vicinity of intron–exon boundaries appear to be one of the stronger, if not the strongest, predictors of rates of protein evolution in mammals. Naturally, for intron-poor genomes the same will not apply.

Materials and Methods

Amino acid preferences near intron–exon junctions. We established a dataset of 178,382 human exons derived from the RefSeq track at the University of California Santa Cruz genome browser (http://genome.cse.ucsc.edu/cgi-bin/hgTables), March 2006 release. We obtained a set of 21,990 RefSeq files with the exon structure of the CDS specified. All files were checked to ensure that the coding sequence started with ATG, finished with a stop codon, had no internal stop codons, had no codons of uncertain translation, and was a multiple of three. This resolved to a dataset of 19,384 RefSeq files. We eliminated all first and last exons, leaving a sample of 178,382 exons. We trimmed all exons so that the first base was the first base of the first complete codon, and the last base the last of the final complete codon. As, to ensure correct splicing, first and last codons are by necessity highly skewed in usage, these too were eliminated. For each codon and in turn each amino acid, we considered proportional usage of that amino acid at a given distance from the junction both 3′ and 5′. All exons were divided in two, so a given codon never featured in both 3′ and 5′ calculations. This sample was not purged for duplicates. However, we repeated the analysis on a more stringently defined set of over 2,000 genes and 14,000 exons, previously purged for duplicates [21]. We confirmed that all qualitative trends are identical (data not shown).

We then considered the trend in usage of each amino acid as a function of the distance from the boundary. This we did by calculating Spearman rank correlations (rho) between the distance from the boundary (5′ or 3′) and proportional usage of the amino acid (i.e., in proportion to the number of residues at that given distance). Note that a negative rho implies an amino acid that is preferred near boundaries, and a positive rho implies a tendency to be avoided. To simplify numbering on the plots, we refer to amino acid positions by reference to the number of full codons between the given position and the relevant end of the trimmed exon. We split the three 6-fold degenerate amino acids into a block of four and a block of two. The block of two is specified by the usage of the lowercase letter (i.e., "S" implies TCA, TCC, TCG, and TCT, while "s" implies AGC and AGT). In relevant circumstances, the 2-fold and 4-fold blocks were treated as separate amino acids. Changes between the 2- and 4-fold blocks were not, however, treated as nonsynonymous changes.

Mouse–human orthologous exon set. As with the derivation of the human exon set, we obtained a set of mouse exons via the RefSeq track at the University of California Santa Cruz genome browser. For analysis of trends in amino acid preference near junctions, these exons were handled as described above. For analysis of orthologous

exons, we obtained the human–mouse orthologue list from Mouse Genome Informatics (ftp://ftp.informatics.jax.org/pub/reports/index.html). We identified all pairs for which both mouse and human sequence had a RefSeq entry. As before, we eliminated all full coding sequences that were not well translated (more than one stop, ambiguous codons, etc.). We further eliminated those in which the number of exons differed between the orthologues. We then compared the phases of the putatively orthologous exons. Gene pairs in which any orthologous exon did not have the same phase in mouse and human were eliminated, leaving 7,767 genes. Any genes in which any orthologous exon differed by more than 5% in size were also eliminated, leaving 5,057 genes. First and last exons were removed, and all remaining orthologous exons were trimmed to start at the first full codon and end at the end of the last complete codon. They were then aligned at the peptide level using muscle v3.6 [31]. This left 36,683 aligned orthologous internal exons.

HPI. Burge and colleagues have characterised numerous hexameric sequences that function as splice enhancers [22,25,32,33]. For each hexamer we can then define a series of full codons that could potentially be present in the hexamer. If we consider a series of six nucleotides, $n_1n_2n_3n_4n_5n_6$, then codons $n_1n_2n_3$, $n_2n_3n_4$, $n_3n_4n_5$, and $n_4n_5n_6$ are specified in their entirety. We sum all such possible codons for all specified splice-enhancer hexamers. This provides a measure of ESE hexameric involvement of all possible codons, within any given hexamer dataset. The three stop codons were removed, and the proportions renormalised. To provide a metric of involvement of an amino acid in ESEs, we compared rates of involvement of codons in the hexamers with those in the genome as a whole. To this end, we normalised (after stop codon removal) the relative abundances of all codons as specified in the appropriate codon usage database (http://www.kazusa.or.jp/codon). We then generated 10,000 sets of random hexamers, each set being the same size as the input hexamer list. Hexamers were generated by joining two codons selected at random in proportion to their frequency in the appropriate genome. We parsed each random hexamer in the same manner as we parsed the input list, extracting all non-stop codons.

For each amino acid, given the frequencies of the relevant synonymous codons, we then determined the mean and standard deviation in relative abundance in the 10,000 random sets. The difference between the observed frequency of an amino acid in the real hexamer set and in the randomised sets, normalised by the standard deviation in the randomised sets, then is our HPI (i.e., a Z score). A high HPI value indicates that a given amino acid is enriched in ESEs compared with what is expected given its content in the genome, and given the underlying variance expected based on the number of hexamers used as input. Source code to calculate HPI is freely available from L. D. H.

In principle, the HPI score for an amino acid will change as a function of both input codon frequencies and with the input set of known ESE hexamers. In practice, we find that employing mouse rather than human codon frequencies makes little or no difference (data not shown). In this analysis we thus employed human codon frequencies to assemble random hexamers. As regards the input list for hexamers, we considered three sets: two sets specific to human 5′ and 3′ exonic ends (95 5′ enhancers and 177 3′ enhancers) and a set of 175 hexamers found both in mouse and human at either exonic end. We found that scores for 5′ and 3′ ends were very similar to each other. Unless otherwise stated, we employed the mouse–human conserved set. Use of this latter set is advantageous as it is most probably enriched for strong enhancers.

The splice-enhancing hexamers in all datasets have two striking features, notably an abundance of adenine and a dearth of cytosine, relative to their usage in the human genome. In the human genome, cytosine constitutes 26.0% of all nucleotides in coding sequences (derived from table of codon usage as noted above) but only 12.5% in splice enhancers, while adenine is 25.6% of all nucleotides in coding sequences but is 49.0% of the nucleotides in splice enhancers. Guanine is used in approximately the same amount in hexamers and in the genome (26.4% in genome and 25.7% in hexamers). Thymine is, like cytosine, underused in hexamers (12.4%), but its usage in the genome is just 22.0%, so its relative reduction in hexamers is less dramatic than that of cytosine. As expected, amino acids with few cytosine nucleotides in their codon set and many adenine residues tend to have positive HPI values (Spearman rank correlation, HPI versus cytosine content of codons, rho = −0.63, $p = 0.0012$; HPI versus adenine content of codons, rho = +0.71, $p = 0.0002$, $n = 23$). A composite measure of adenine and cytosine bias of codons (frequency of adenine in synonymous codon set minus frequency of cytosine) is a good predictor of HPI (Spearman rank correlation = 0.85, $p < 0.0001$, $n = 23$).

For the DPI pertinent to splice suppressors we extracted the 131 decamers provided by Wang et al. [29] from http://www.cell.com/cgi/content/full/119/6/831/DC1. The protocol to define DPI scores was identical to that to calculate HPI, except that random decamers were made by random selection of four codons and trimming off of the final two bases. The eight full codons in the decamers were employed to define expected frequencies.

Establishing a set of ancient functional retroposed genes. Mouse retroposed gene copies were identified using the procedure described in Vinckenbosch et al. [34]. For humans, we used a previously established retrocopy dataset [34]. To identify orthologous retrocopies shared between humans and mouse, we used human–mouse chained alignments available from the University of California Santa Cruz (hg17 versus Mm6). Similar to our previous procedure [34], we first extracted the best alignments that overlapped with the genomic location of human retrocopies and that were >15 kb (this length ensures that the alignment also covers surrounding, nonretrocopy-derived sequences in the two species). If no such alignments could be identified, presence/absence in mouse was not determined. We then scanned the chained alignments for an aligned block (putative orthologous sequence in the chain) that overlapped with the human retrocopy. If such a block was found, its corresponding mouse coordinates were compared to the mouse retrocopy set. Mouse retrocopies overlapping with these coordinates were considered orthologues of human retrocopies. In total, we identified 56 orthologous retrocopy pairs, of which 49 showed intact open reading frames in both species. The fact that these retrocopies emerged in the common ancestor of humans and mice (at least approximately 75–90 million years ago) and possess intact open reading frames strongly suggests that they have been selectively preserved by natural selection. Thus, they likely represent functional retroposed gene copies (retrogenes). Functionality of these human–mouse retrocopies is further supported by their generally higher transcription levels and lower K_A/K_S values relative to younger, lineage-specific retrocopies [34].

To infer retrogene-specific changes, the sets of four sequences were aligned at the protein level using Muscle [31]. The sequences were then cut into individual exons by reference to the human annotation of parental genes. Exons were trimmed so as to contain only complete codons. The 5′ end of the first exon and the 3′ end of the last exon were ignored. All sites in the amino acid alignment that specified the same amino acid in three of the four sequences but a different amino acid in the third were considered, by parsimony, to be informative. That is, if the two human sequences specify amino acid X, as does the mouse parent gene at a given position, while the mouse retrogene is amino acid Y, then an X→Y change is inferred to have occurred in the mouse retrogene. The total number of retrogene changes is simply the sum of those in the mouse and those in the human retrogene, employing this strict 3:1 criterion.

Expression data. Gene expression estimates were obtained from Su and colleagues [35], employing the March 2006 annotation (http://wombat.gnf.org/index.html). Mas5 files with Affymetrix present/absent calls were used. Human gene expression data were obtained by merging U133A and GNF1h chip datasets. In both mouse and human, average expression was obtained from samples of the same tissues. Probes matching to more than one gene were eliminated from further analyses. Indexes of gene activity were obtained only from samples obtained from normal adult tissues. Levels and breadth of expression were calculated. Three indexes for expression levels were obtained: peak, average, and median expression. The peak level was the highest score across all analysed tissues. Breadth of expression was calculated from present/absent calls. For the analysis of mean/median levels, for each gene we considered only those tissues in which a gene was expressed (judged by present/absent call). When multiple probes matched the same gene we considered a gene to be expressed in a given tissue if half or more of the probes indicated presence.

Supporting Information

Figure S1. Trends in Relative Levels of Amino Acid Usage as a Function of the Distance from Intron–Exon Boundaries at Both 5′ and 3′ Ends

Found at doi:10.1371/journal.pbio.0050014.sg001 (183 KB PDF).

Figure S2. Relationship between Slope of Regression Line (between Proportion of Amino Acid and Distance from Boundary) and HPI Score

For (A) 5′ and (B) 3′ ends.

Found at doi:10.1371/journal.pbio.0050014.sg002 (47 KB DOC).

Figure S3. Rates of Evolution in Enhancer and Nonenhancer Domains as a Function of Distance from the Boundary for Exons Longer than 200 bp

All exons contribute equally to all data points. Here we merge 3′ and 5′ data.

Found at doi:10.1371/journal.pbio.0050014.sg003 (55 KB DOC).

Figure S4. Plot of Standardised Residuals for the Regression of AC Content Versus rho

Grey lines indicate top and bottom 95% confidence intervals.

Found at doi:10.1371/journal.pbio.0050014.sg004 (41 KB DOC).

Figure S5. Relationship between the Correlation between Proportion of Amino Acid and Distance from Boundary (rho) and the Mean of HPI and DPI

For (A) 5′ and (B) 3′ ends.

Found at doi:10.1371/journal.pbio.0050014.sg005 (43 KB DOC).

Figure S6. Glycine Usage as a Function of Frame of Exon and Exonic End

The first number in the title is the exonic end (5′ or 3′); the second is the phase (0, 1, or 2).

Found at doi:10.1371/journal.pbio.0050014.sg006 (115 KB DOC).

Figure S7. Periodicity Analysis of Glycine and Proline

For (A) glycine and (B) proline the distribution of homologous residues in the flanking sequence was determined. The first such residue in the sequence was taken as the reference point 0; once frequency data for the same amino acid were obtained, for the 150 flanking residues, the reference point moved along to the next homologous residue. The frequency of the residue in the flanking sequence was then determined by the absolute occurrence of the residue at this distance, divided by the number of informative sites. Glycine exhibits an unusual pattern where, following the use of a glycine, there is a preference for glycine to be used every third amino acid (top series of points in [A]). This is not an artefact of contamination by collagen transcripts (GPXn), as the distribution of proline indicates no such trend. This pattern in glycine usage is still strong over 500 residues away from the reference point.

Found at doi:10.1371/journal.pbio.0050014.sg007 (79 KB DOC).

Table S1. Correlations and Partial Correlations between Rate of Protein Evolution (K_A or K_A/K_S), Proportion of Sequence within 50, 70, or 100 bp of an Intron–Exon Junction, and Measures of Expression of the Relevant Gene in Mouse

Found at doi:10.1371/journal.pbio.0050014.st001 (36 KB DOC).

Acknowledgments

We thank Fedya Kondrashov and two anonymous referees for comments that substantially improved the manuscript, Nicolas Vinckenbosch for helpful discussions, and the Vital-IT team at the University of Lausanne for computational support.

Author contributions. JLP, HK, and LDH conceived and designed the experiments and analyzed the data. JLP and LDH performed the experiments. AOU, LP, HK, and LDH contributed reagents/materials/analysis tools. JLP and LDH wrote the paper.

Funding. JLP is funded by the Biotechnology and Biological Sciences Research Council, United Kingdom. This work was also partly funded by Swiss National Science Foundation grant 3100A0–104181 and the Center for Integrative Genomics (University of Lausanne, Lausanne, Switzerland).

Competing interests. The authors have declared that no competing interests exist.

References

1. Pal C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. Nat Rev Genet 7: 337–348.
2. Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158: 927–931.
3. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23: 327–337.
4. Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Mol Biol Evol 17: 68–74.
5. Subramanian S, Kumar S (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168: 373–381.
6. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102: 14338–14343.
7. Pal C, Papp B, Hurst LD (2003) Genomic function (communication arising): Rate of evolution and gene dispensability. Nature 421: 496–497.
8. Hirsh AE, Fraser HB (2003) Genomic function (communication arising): Rate of evolution and gene dispensability. Nature 421: 497–498.
9. Bloom JD, Adami C (2004) Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: Response. BMC Evol Biol 4: 14.
10. Fraser HB, Wall DP, Hirsh AE (2003) A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 3: 11.
11. Batada NN, Hurst LD, Tyers M (2006) Evolutionary and physiological importance of hub proteins. PLoS Comput Biol 2: e88. doi:10.1371/journal.pcbi.0020088
12. Liao BY, Scott NM, Zhang J (2006) Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol 23: 2072–2080.
13. Zhang JZ, He XL (2005) Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol 22: 1147–1155.
14. Hurst LD, Smith NGC (1999) Do essential genes evolve slowly? Curr Biol 9: 747–750.
15. Rocha EPC, Danchin A (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol 21: 108–116.
16. Jordan IK, Wolf YI, Koonin EV (2003) No simple dependence between protein evolution rate and the number of protein-protein interactions: Only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3: 1.
17. Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res 12: 962–968.
18. Logsdon JM (1998) The recent origins of spliceosomal introns revisited. Curr Opin Genet Dev 8: 637–648.
19. Blencowe BJ (2000) Exonic splicing enhancers: Mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci 25: 106–110.
20. Willie E, Majewski J (2004) Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet 20: 534–538.
21. Chamary JV, Hurst LD (2005) Biased codon usage near intron-exon junctions: Selection on splicing enhancers, splice-site recognition or something else? Trends Genet 21: 256–259.
22. Fairbrother WG, Holste D, Burge CB, Sharp PA (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2: e268. doi:10.1371/journal.pbio.0020268
23. Carlini DB, Genut JE (2006) Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J Mol Evol 62: 89–98.
24. 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.
25. 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 U S A 101: 15700–15705.
26. Comeron JM (2004) Selective and mutational patterns associated with gene expression in humans: Influences on synonymous composition and intron presence. Genetics 167: 1293–1304.
27. Lercher MJ, Chamary JV, Hurst LD (2004) Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. Genome Res 14: 1002–1013.
28. Zhang LQ, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol 21: 236–239.
29. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, et al. (2004) Systematic identification and analysis of exonic splicing silencers. Cell 119: 831–845.
30. Eskesen ST, Eskesen FN, Ruvinsky A (2004) Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. Genetics 167: 543–550.
31. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
32. Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, et al. (2004) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res 32: W187–W190.
33. Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. Science 297: 1007–1013.
34. Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A 103: 3220–3225.
35. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062–6067.

Chapter 7. **Discussion**

7.0 Discussion

In this thesis I have analysed and presented evidence for the impact of splicing regulation on the evolution of mammalian genes. The mechanisms of the intricate regulation of splicing are still being elucidated, but it would appear that their effects are in no way trivial. Previously, the best predictor of the rate of non-synonymous evolution was gene expression, with low rates of evolution in the most abundantly expressed genes (Drummond, Raval, and Wilke 2006). The reasons for this correlation are far from clear. We can now infer that intron density (measured by the proportion of sequence within 70 nucleotides of a splice site) is at least equal in strength at predicting the evolution rates of non-synonymous sites as the strongest known correlate. Perhaps more notably, unlike expression parameters, the correlation to splicing has an obvious mechanism of selection. The importance of these results does not end here. Due to the relatively low effective population size of mammalian genomes, synonymous sites were thought to evolve neutrally or with effective neutrality (Sharp et al. 1995). Previous observations have shown codon biases that could suggest that selection does occur at these sites, but there have been contrasting, yet not definitive, explanations including selection for splicing of introns (Willie and Majewski 2004) and cryptic splice site avoidance (Eskesen, Eskesen, and Ruvinsky 2004; but also see Chamary and Hurst 2005).

Previous work highlighted the change in the employment of synonymous codons GAA:GAG approaching splice sites (Willie and Majewski 2004). The observed trend, favouring GAA near boundaries was interpreted as evidence that ESEs impact on codon usage (Willie and Majewski 2004). However, the same fact is also consistent with the cryptic splice site avoidance model (Eskesen, Eskesen, and Ruvinsky 2004). In this thesis I have shown that regions thought to possess splicing enhancer properties are selectively maintained, as evidenced by slower rates of evolution. I also asked whether patterns of codon bias near intron-exon boundaries could more generally be explained by the presence of ESEs. I found that trends in codon usage were prevalent in human genes and, what's more, the direction and extent of the bias was correlated to the comparative usage of the codons in the ESE data set. In addition to other recent work (see chapter 2 and appendix) there now seems little doubt that selection does act on synonymous sites in mammals. The remaining questions concern the magnitude and commonality of such effects.

Understanding the causes of protein evolution is the key to many fields, including molecular evolution, comparative genomics and structural biology. Protein domains of

high functional importance are likely subject to purifying selection and thus, the analysis of protein evolution can facilitate the identification of such, yet unknown, functionally important sites. In order to determine regions of proteins under selective constraint or diversification (positive) selection, the non-synonymous substitution rate must be adjusted to account for local mutation rates. Traditionally this is done by the K_A/K_S ratio. Under the assumption that synonymous sites evolve with effective neutrality, a ratio greater than unity is interpreted as diversifying selection. Given that selection can act on synonymous mutations, how might this affect interpretation of K_A/K_S ratios? By analysis of substitution rates in sliding windows, I revealed that at least half, if not more, of the peaks in $K_A/K_S > 1$ are caused, not by an increase in protein evolution, but by low rates of evolution at the synonymous sites. After the identification of strictly significant peaks in the ratio, it was found that all examples that were in addition repeatable in independent taxa, were related to alternatively spliced exons.

Given the presence of ESE sequences near intron-exon boundaries (Fairbrother et al. 2004), many non-synonymous sites proximal to such boundaries will have a dual coding function: to encode a protein but also to ensure that splicing occurs with a minimal required level of accuracy. How then, do these two factors interact? Surprisingly, it was found that amino acid choice was not totally dependant on protein function, and that, potentially, the protein function is compromised so that ESE sequences can be present. It was observed that amino acids whose codons are more common within the ESE dataset are preferred near intron-exon boundaries and that the removal of introns, as in functional retrogenes, is linked to an increase in amino acid substitutions in this region.

Given that we observe such a large impact at the protein level would we not expect to see a larger influence at the synonymous sites? When we compare the effect of purifying selection for accurate splicing on the evolution of synonymous (chapter 3) and non-synonymous sites (chapter 6), at first glance, the impact on the bias of protein usage seems much greater than that of synonymous codon usage. In chapter 3, it was stated that the value of the genic synonymous evolution rate could be underestimated by much less than 10%, whereas in chapter 6, the non-synonymous evolution rate near splice sites was said to be 50% that of the evolution rate in the centre of exons. This is counter-intuitive and would go against our current understanding of the characteristics of these two classes of site. This may likely be an artefact of the different parameters that were calculated, resulting in neither outcome being directly comparable with the other. There is one common observation between the two chapters: those contained in figure 4 (chapter 3) and figure 5 (chapter 6). The direct comparison of these two figures shows us two main features.

Firstly, as expected, there is a large difference between the scales for the substitution rates, with the synonymous substitution rate starting at 0.275 in genes with few introns to 0.17 in genes with short exons. Similarly for non-synonymous sites, the substitution rates range from 0.08 in genes with few introns to 0.04 in genes with the majority of sequence near a splice site. Looking at the data in this way shows that the effect of purifying selection on splicing regulation has an effect of much greater magnitude on the evolution of synonymous sites, as would be predicted. The second observable feature, which may account for the strength of the relationship seen in chapter 6 that is absent in chapter 3, is a much lower level of variance among the rates of non-synonymous evolution, a trend we may expect given the flexibility associated with synonymous sites.

Are there any alternative interpretations to these results that we have not yet considered?

In this thesis I have regularly examined patterns (of evolutionary rates, of codon and amino acid usage) in the vicinity of intron-exon boundaries. The trends I have shown are consistent with selection for efficient splicing. But might there be alternative explanations? In species, such as Drosophila, in which there is selection for optimal codon usage modulated by selection for translational efficiency (accuracy/rate), weaker Hill-Robertson interference near intron-exon boundaries could act to increase usage of such optimal codons (for references see Warnecke and Hurst 2007). However, the trends seen in Drosophila are not consistent with such a model (Warnecke and Hurst 2007) and instead favour the splice regulation model examined in this thesis. For example, the translationally optimal codons tend to be avoided near intro-exon boundaries, while those rich in A and poor in C (as typical of ESEs) are favoured (Warnecke and Hurst 2007). Given too the absence of robust evidence for translational optimization in mammals, I consider greater efficiency of selection for translationally optimal codons near introns to be a model of little relevance to mammals.

A further alternative explanation concerns nucleosome binding. The packaging of DNA requires the binding of several histone proteins, some of which are the most conserved proteins among animal species. The DNA helix loops around a complex of central histone proteins, and is secured in place by external histones. This structure is termed a nucleosome and occurs periodically along the DNA. Recent work into the positioning of nucleosomes onto specific sequences has suggested that nucleosomes might localise near intron-exon boundaries (Denisov, Shpigelman, and Trifonov 1997). This may modulate

evolutionary rates for a variety of reasons. First, as the nucleosomes are thought to associate with RR(YY) dinucleotides (Trifonov and Sussman 1980; Mengeritsky and Trifonov 1983; Uberbacher, Harp, and Bunick 1988; Kato et al. 2003) selection to favour or avoid nucleosome binding could result in biased sequence composition and altered evolutionary rates. Second, it has been proposed that the presence of nucleosomes confer a level of protection to the bound sequence, as the DNA is not exposed to potential mutagenic elements. If these sequences happen to be intron-exon boundaries then we would expect a lower mutation rate, thus lower SNP density and lower synonymous and non-synonymous substitution rates near splice sites.

The presence of nucleosomes near splice sites is, however, unlikely to cause the correlation we observe between enhancer sequences and constraints on sequence surrounding splice sites. In human, unlike yeast, genes it was found that the dinucleotides most favoured for periodicity, and thus, most likely to participate in nucleosome formation were GG and CC (Kogan and Trifonov 2005). Since C residues are the most rare in exonic splicing enhancer elements it is unlikely that the skews we observe in codon usage near intron-exon boundaries is due to the maintenance of nucleosome binding residues.

Prospects

Does the work presented here have any potential utility and what should be the optimal future avenues of research? The work presented here has lead on to further investigations, featured in the appendices to this thesis, which delve into the breadth of the effect of splicing on evolution across species and taxa (Appendix 2). Here, evidence is given that those genomes that have more SR protein family members have more codon bias near intron-exon boundaries. More simple eukaryotes, such as yeast, do not require SR proteins to facilitate splicing, as the intron-definition of splice sites is adequate (Burge, Tuschi, and Sharp 1999). It is notable that in these species we see no robust trends in amino acid usage near intron-exon boundaries. This suggests that it may be possible, therefore, to determine the level of SR protein involvement in the regulation of splicing by identifying the level of codon and amino acid bias in species with nothing more than an annotated genome.

Conversely, we can ask whether our observations of codon bias and amino acid bias enable us to predict the binding motifs of SR proteins in a species-specific manner? ESE sequences are not thought to be identical across species. Indeed, those sets derived by RESCUE-ESE for human and mouse genomes, even though these species are relatively

close, only share 174 hexameric sequences from a full human set of 238 (Fairbrother et al. 2004). Might it be possible, given only sequence data, to derive those SR mRNA binding motifs depending on the bias in codon usage near intron-exon boundaries? Both of the above may only be viable if there is little or no competition by translational selection for codon usage in the genome of interest. Recent results (Warnecke and Hurst 2007), however, suggest that this is not an issue.

A further possibility, suggested by the work in this thesis, is the potential to develop more efficient transgenes. In particular, the results suggest the potential for identifying those sites that are needed for splicing in intron containing genes, but thus possibly amenable to beneficial change in an intronless transgene. Due to the further revelation that splicing control elements affect the protein sequence, potentially compromising gene function, can this idea be expanded? Can we increase the potency of a therapeutic protein? Any such developments will require extensive *in vitro* analysis.

How extensive is dual coding in the genome and can we use sequence divergence to estimate the mutation rate?

One corollary of the work presented here is that estimation of the mutation rate from synonymous substitution rates may well, even in mammals, provide an underestimate. However, by masking ESE related sequences, or abolishing sequences near intron-exon boundaries, we might mitigate such biases. This, however, begs the question as to what the quantitative effect of other modes of selection on synonymous mutations might be. Recent work has highlighted both translational pausing and modification of RNA structure as mechanisms of human disease and selection on synonymous mutations (for discussion see appendix 1). But just how common and important are these mechanisms?

Perhaps, given these problems, we might suppose the estimates from both intronic and synonymous sites might be unsafe. In intergenic regions, far from known genes, can we be confident that evolution is neutral? I should like to suggest that, aside from the existence of ultra-conserved domains and the abundant transcription all over the human genome (Kapranov, Willingham, and Gingeras 2007), a further problem may prove important. Just as mRNA needs to specify ESEs so might DNA be under selection to enable proper nucleosome positioning? As noted above, nucleosomes are a core component of DNA packaging. The regular spacing of conserved motifs that correspond to nucleosome

formation are at intervals of roughly 10.5 nucleotides, approximately one turn of the DNA double helix. If selection to ensure nucleosome binding (or non-binding (Lee et al. 2007)) constrains the evolution of synonymous mutations, can we truly assume that any genomic sequence evolves neutrally? This begs the issue of whether it is therefore possible to estimate neutral evolution rates from sequence data. Whether these are issues requiring minor correction or a fundamental challenge to the enterprise of estimating mutation rates from sequence data, remains to be seen.

References

Burge, C.B., T. Tuschi, and P.A. Sharp. 1999. Splicing of Precursors to mRNAs by the Spliceosomes. Pp. 525-560 *in* R. F. Gesteland, T. R. Cech, and J. F. Atkins, eds. The RNA World. COLD SPRING HARBOUR LABORATORY PRESS, New York.

Chamary, J.V. and L.D. Hurst. 2005. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet. **21**:256-259.

Denisov, D.A., E.S. Shpigelman, and E.N. Trifonov. 1997. Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. Gene **205**:145-149.

Drummond, D.A., A. Raval, and C.O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol **23**:327-337.

Eskesen, S.T., F.N. Eskesen, and A. Ruvinsky. 2004. Natural selection affects frequencies of AG and GT dinucleotides at the 5' and 3' ends of exons. Genetics **167**:543-550.

Fairbrother, W.G., G.W. Yeo, R. Yeh, P. Goldstein, M. Mawson, P.A. Sharp, and C.B. Burge. 2004. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res. **32**:W187-190.

Kapranov, P., A.T. Willingham, and T.R. Gingeras. 2007. Genome-wide transcription and the implications for genomic organization. Nat Rev Genet **8**:413-423.

Kato, M., Y. Onishi, Y. Wada-Kiyama, T. Abe, T. Ikemura, S. Kogan, A. Bolshoy, E.N. Trifonov, and R. Kiyama. 2003. Dinucleosome DNA of human K562 cells: experimental and computational characterizations. J Mol Biol **332**:111-125.

Kogan, S. and E.N. Trifonov. 2005. Gene splice sites correlate with nucleosome positions. Gene **352**:57-62.

Lee, W., D. Tillo, N. Bray, R.H. Morse, R.W. Davis, T.R. Hughes, and C. Nislow. 2007. A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet **39**:1235-1244.

Mengeritsky, G. and E.N. Trifonov. 1983. Nucleotide sequence-directed mapping of the nucleosomes. Nucleic Acids Res **11**:3833-3851.

Sharp, P.M., M. Averof, A.T. Lloyd, G. Matassi, and J.F. Peden. 1995. DNA-sequence evolution: the sounds of silence. Phil Trans R Soc Lond B **349**:241-247.

Trifonov, E.N. and J.L. Sussman. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci U S A **77**:3816-3820.

Uberbacher, E.C., J.M. Harp, and G.J. Bunick. 1988. DNA sequence patterns in precisely positioned nucleosomes. J Biomol Struct Dyn **6**:105-120.

Warnecke, T. and L.D. Hurst. 2007. Evidence for a Trade-Off between Translational Efficiency and Splicing Regulation in Determining Synonymous Codon Usage in Drosophila melanogaster. Mol Biol Evol. in press.

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

Appendices

Appendix 1 **How do synonymous mutations affect fitness?**

Joanna L. Parmley and Laurence D. Hurst

Bioessays (2007) 29(6): 515-519

How do synonymous mutations affect fitness?

Joanna L. Parmley and Laurence D. Hurst*

Summary

While it has often been assumed that, in humans, synonymous mutations would have no effect on fitness, let alone cause disease, this position has been questioned over the last decade. There is now considerable evidence that such mutations can, for example, disrupt splicing and interfere with miRNA binding. Two recent publications suggest involvement of additional mechanisms: modification of protein abundance most probably mediated by alteration in mRNA stability[1] and modification of protein structure and activity,[2] probably mediated by induction of translational pausing. These case histories put a further nail into the coffin of the assumption that synonymous mutations must be neutral. *BioEssays* 29:515–519, 2007. © 2007 Wiley Periodicals, Inc.

Introduction

It is seductive to think that, owing to the redundancy in the genetic code, a point mutation in a protein-coding exon that changes the DNA but not the protein sequence (a synonymous mutation), would have no discernible fitness consequences. Indeed, even a decade ago such an assumption looked relatively sound. Since then, however, there has been a plethora of evidence to indicate that synonymous mutations can, indeed, have important fitness consequences, with over 40 genetic diseases now associated with such "silent" mutations.[3] How do apparently innocuous base changes have such an effect?

Codon usage bias puts the neutral theory in retreat

Since the introduction of the neutral theory and the finding that synonymous substitutions happen much faster than non-synonymous ones,[4] the neutrality of synonymous mutations was initially widely assumed. For species with large population sizes (worms, flies, yeast, bacteria etc.), however, this position was gradually eroded through the 1980s by the finding that, especially in highly expressed genes, the choice of which synonymous codon is employed to specify a given amino acid was not random.[5,6] Rather the codon that matched the most-abundant iso-acceptor tRNA species was preferentially employed (see Refs. 7–9). Indeed, it was conjectured that the skew in tRNA pool and codon usage should co-evolve so as to ensure that the most highly expressed genes could be translated as fast and as accurately as possible.[10]

Deptartment of Biology and Biochemistry, University of Bath, Bath, BA2 7AY.
*Correspondence to: Laurence D. Hurst, Department of Biology and Biochemistry, University of Bath, Bath, BA2 7AY.
E-mail: l.d.hurst@bath.ac.uk
DOI 10.1002/bies.20592
Published online in Wiley InterScience (www.interscience.wiley.com).

A mammal is not an invertebrate

In this translation rate modification model, selection on a synonymous mutation that specifies an un-preferred rather than a preferred codon is likely to be weak.[11] Given that, in the framework of the nearly neutral model of molecular evolution,[12] selection is less efficient in species with small effective population sizes, it was supposed that selection of this variety would be all but irrelevant in mammals.[13] This was given credence by a variety of studies that failed to find any evidence of the expected forms of codon bias in mice and humans.[14,15] More recently, however, with the vast datasets now available and improvements in methods to detect codon bias (for example, those that allow for the overwhelming influence of regional nucleotide differences around mammalian genomes, see Ref. 16), there have been some small indications of this mode of selection [17–20] and, more generally, of selection of some form on synonymous mutations.[21–24]

Selection for translational accuracy/rate appears, however, to be weak if at all present in humans,[25] and cannot obviously explain why large tracts in exons containing highly conserved synonymous positions exist.[26,27] What then might be the mechanism or mechanisms of selection on synonymous mutations in mammals?

One early clue came from the finding that alternatively spliced exons have unusually low rates of evolution at synonymous sites;[28] this has since been verified on numerous occasions.[29] Combining this with evidence that synonymous rates of evolution can be especially low in exonic domains associated with splice control,[30,31] has led to the understanding that most selection on synonymous mutations in mammals is associated with perturbation of splicing. Remarkably, in one well-studied example, exon 12 of CFTR, a quarter of synonymous variations result in exon skipping.[32] More generally, most of the 40 or so genetic diseases

associated with synonymous mutation appear owing to disruption of splicing.[3] Likewise, many of the large exonic tracts of low synonymous substitution rate are associated with alternative exons.[26,27]

An association with splicing need not simply reflect mutation in the few base pairs immediately adjacent to the intron–exon boundary. Rather, the role of exonic splice enhancer (ESE) domains has been highlighted in several incidences.[3] These are sequences necessary for the binding of SR proteins to the immature mRNA, which, in turn, are needed for specification of the location of the intron–exon boundary. Importantly, ESEs have low synonymous SNP densities[33,34] and synonymous sites in ESEs evolve significantly slower than the flanking non-ESE synonymous sites.[35] Selection favouring ESEs in the vicinity of the intron–exon boundary has striking effects both on genic synonymous[35] and non-synonymous[36] rates of evolution in mammals.

It would then be tempting to suppose that, in humans, with their very high density of introns, selection on synonymous mutations is different to that which occurs in yeast, fly and worm, and is all associated with control of splicing. It appears premature to suppose that, in mammals, splicing explains all of the selection on synonymous mutations. For one thing, miRNA binding within coding exons appear to impose selective constraint on synonymous mutations within the binding sites,[37] as might be expected. Importantly, two recent papers highlight further different modes of selection. In one instance, the stability of mRNA is affected, which, in turn, affects protein concentration and net enzymatic rate. In the other, the synonymous mutations appear to affect protein folding, possibly by causing translational pausing while rare tRNAs are recruited. This in turn affects the activity of the protein.

mRNA stability and the case of *COMT*

Nackley et al. focused on the single nucleotide polymorphisms (SNPs) that affect the activity of the cathechol-*O*-methyltransferase (*COMT*). This gene is responsible for the degradation of catecholamines and is associated with responsiveness to pain in humans. There are three common haplotypes that are associated with levels of pain sensitivity: low (LPS), average (APS) and high (HPS).[38] The three haplotypes are composed of varying combinations of four SNPs: one in the promoter (A/G), two synonymous changes (C/T and C/G) and one non-synonymous valine-to-methionine change (A/G). It had been widely accepted, in humans, that the cause of the variation in *COMT* activity is due only to the non-synonymous SNP. Evidence for this, however, is weak as the two haplotypes with the most-extreme phenotypes (LPS and HPS), aside from differing in the promoter, only differ within the coding sequence at one synonymous SNP. These differences between the haplotypes are paralleled by differences in enzyme activity (reduced in cells expressing the HPS haplotype, in comparison to the LPS haplotype). Importantly, it was shown that this was due to reduced protein abundance and not reduced mRNA abundance.[1] As mRNA abundance does not parallel enzyme activity levels,[1] it seems unlikely that the promoter in the SNP could explain the differences between haplotypes (see also Supplementary Table 3 of Ref. 38). It was thus proposed that the underlying cause of the pain phenotype was implemented at the mRNA/translation level, and that a change in mRNA secondary structure could lead to a perturbation in protein synthesis.

To test this, Nackley et al. computationally analysed the stability of the mRNA secondary structures across the three haplotypes. Results of such in silico analyses should always be taken with caution, as the methods of assessing mRNA structure tend not to allow for features such as the proteins that are left on the mature mRNA at splice junctions. Nonetheless, the team report that (1) the predicted least-stable structure was that of the LPS haplotype mRNA, forming the shortest stem-loop structure, (2) the most-stable structure was encoded by the HPS haplotype (Fig. 1) and (3) the APS haplotype formed a mRNA secondary structure of intermediate stability. This mechanism was given credence by site-directed mutagenesis analyses in which new nucleotide changes were introduced that disrupted the predicted stem-loop structure observed in HPS haplotype, creating the LPS-like mRNA structure. Then, a secondary compensatory nucleotide change was introduced that converted the LPS-like structure back to a HPS-like structure. These predicted structural changes resulted in the same protein expression and enzymatic activity levels associated with the newly acquired haplotype mRNA structure.

Assuming mRNA stability is at the heart of the different activity and expression levels, why might this be? Several hypotheses could be considered. Perhaps the ultra-stable mRNA is hard for the ribosome associated helicases[39] to unwind? Perhaps, the stable structures are more prone to attack by RNAases?[40] RNA levels and degradation rates did not parallel protein levels[1] suggesting that a mechanism more like the former than the latter probably applies. Whatever the mechanism, this result runs counter to that previously proposed by mRNA stability studies[41–43] which find that, generally speaking, more stable mRNAs are selectively favoured as they enable mRNA persistence, potentially increasing the rate of protein expression. However, ultra-stable structures have been suggested as a means to limit expression in special cases by limiting scanning the 5′ end of the mRNA.[44]

Perhaps what is more remarkable in this case history is that the impact of the synonymous changes on enzyme activity is vastly greater than the modest thermostability effect of the non-synonymous change. If this case history tells us anything, it is that we have probably been too fast in ascribing all phenotypic effects to non-synonymous changes simply because the only other SNPs in a haplotype are synonymous.

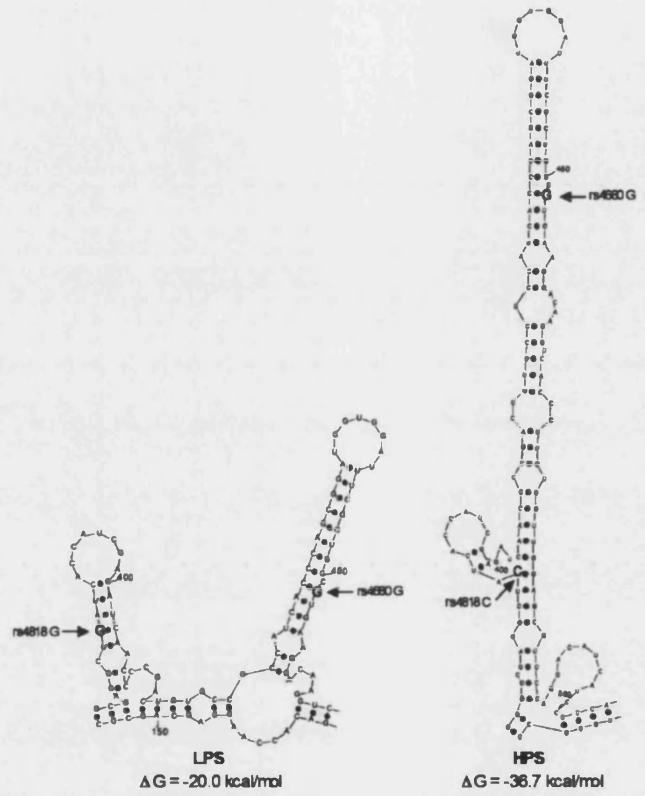


Figure 1. The predicted structures of the LPS and HPS haplotype mRNAs and their Gibbs free energy (ΔG). Note that these two extreme haplotypes differ only at the synonymous rs4818 SNP and not at the non-synonymous rs4680 SNP (courtesy of Andrea G. Nackley Neely).

Translational pausing, protein folding and the case of *MDR1*

The *Multidrug Resistance 1* (*MDR1*) gene encodes an ATP-driven efflux pump (P-gp) that has been associated with the multidrug resistance of cancer cells, though in many instances, the molecular mechanisms of such resistance are unknown. The variation within this gene is high, with over 50 reported SNPs. One synonymous SNP (C3435T) has been linked to a change in P-gp activity and is further associated, when present with a greater combination of SNPs, with reduced functionality. Kimchi-Sarfaty et al.[2] endeavoured to find out how the presence of a silent polymorphism can induce such a fitness effect.

The underlying mechanism for drug resistance is very complicated. Assays analysing the function of P-gp on single mutants, and haplotypes from combinations of these polymorphic variants, revealed no reduction in transporter function compared to the wild type. There was, however, an alteration in drug specificity in only those haplotypes containing the synonymous C3435T variant, even though this was not observed with this SNP alone. How, then, can this silent SNP cause altered drug specificity when in combination with other synonymous and non-synonymous variants? Neither mRNA nor protein levels were found to be diminished in these haplotypes and the protein sequence was as expected, ruling out the possibility that aberrant splice forms were involved.

Perhaps, then, a conformational change has occurred that allows P-gp to function, but inhibits the drug–protein interaction? Assays of trypsin digestion of the common (C1236T-G2677T-C3435T) P-gp haplotype required more than a three-fold increase in trypsin concentration, from the wild-type protein, to reach 50% degradation, indicating that the two

proteins have different tertiary structures. This conclusion was supported by the differential recognition of the haplotype protein compared to wild type using a conformation-sensitive monoclonal antibody. The mechanism by which these two isoforms were produced was attributed to the formation of a cluster of rare codons. One model supposes that rare codons are specified by rare tRNAs (which may[18–19] or may not[25] be the case) and that this ensures that the translational machinery must pause to enable tRNA recruitment. In line with both theory[45] and experiment,[46] such pausing in turn could enable the protein to find new structures. Closer inspection of the common haplotype associated with drug resistance revealed that all three SNPs involved a codon that was more rare than that of the wild type. The pausing mechanism was reinforced as the cause of drug resistance when an artificial haplotype was produced, employing a codon yet more rare than that originally found in this synonymous SNP, that reduced the sensitivity to the drugs yet further.

Conclusions

We have then, in mammals, at least four relatively well-resolved mechanisms by which synonymous mutations can have an effect on fitness: splice regulation, miRNA binding, mRNA folding and protein folding. If we add the possibility of weak effects of translational rate/accuracy and an otherwise mysterious effect of synonymous nucleotide content on mRNA levels,[47] mediated at either the transcriptional or RNA-processing level, that brings the current possible mechanisms to six. It is also likely that overlapping transcripts, which may well be much more common than once thought,[48] will impose some form of extra constraint[49] on mutations that are synonymous but only in one of the two genes.

The present-day predominance in the literature of the splice-associated mechanisms accounting for disease phenotypes[3] may reflect the relative ease of determining that an alternative splice form is found, as opposed to showing, for example, that a protein or mRNA structure is different (see Ref. 50). Suggestive of greater than previously recognized importance of the alternative mechanisms, we note that neither of the two new case histories is without precedent. A synonymous mutation was, for example, previously shown to be associated with disease mediated via its effects on mRNA stability.[51] More generally, several computational analyses have indicated a role for selection acting on synonymous mutations that affect mRNA stability,[41,42,52] although, as noted, these suggest that high stability is preferred. RNA half-life need be associated not only with stem-loop structures but also with residues that enable RNAases to digest mRNA, notably UA residues, which in turn are both avoided and are possibly under selection.[40] Likewise a role for usage of rare codons in enabling translational pausing, which alters protein folding, has been noted previously[53] and may indeed explain why stretches of rare codons correspond to turns, loops and links between protein domains.[54,55] It is notable that different protein structures may well be translated at different rates owing to skews in codon usage,[56] although these skews may instead relate to mRNA stability.[57] Preventing co-translational misfolding has been suggested to be especially important in mammals[58] and could explain why GAT is preferred over GAC at the N termini of alpha-helices in humans.[55,59]

Whether these new case histories are the tip of the iceberg or just rare curiosities remains to be seen. What is clear, however, is that in mammals not only are many synonymous mutations under selection, but the mechanisms by which selection acts on such changes are more diverse than commonly appreciated.

Acknowledgments

We thank Luda Diatchenko, Chava Kimchi-Sarfaty amd Michael Gottesman for comments on an earlier versions of the manuscript. We particularly thank Andrea Nackley Neely for providing us with Fig. 1.

References

1. Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, et al. 2006. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science 314: 1930–1933.
2. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, et al. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315:525–528.
3. Chamary J-V, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98–108.
4. Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267:275–276.
5. Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12:640–649.
6. Powell JR, Moriyama EN. 1997. Evolution of codon usage bias in Drosophila. Proc Natl Acad Sci USA 94:7784–7790.
7. Ikemura T. 1992. Correlation between codon usage and tRNA content in microrganisms. In: Hatfield DL, Lee B, Pirtle PM, editors. Transfer RNA in protein synthesis. Boca Raton: CRC Press. p 87–111.
8. Duret L. 2000. tRNA gene number and codon usage in the C-elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet 16:287–289.
9. Moriyama EN, Powell JR. 1997. Codon usage bias and tRNA abundance in Drosophila. J Mol Evol 45:514–523.
10. Bulmer M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature 325:728–730.
11. McVean GAT, Vieira J. 1999. The evolution of codon preferences in Drosophila: A maximum- likelihood approach to parameter estimation and hypothesis testing. J Mol Evol 49:63–75.
12. Ohta T. 1992. The nearly neutral theory of molecular evolution. Annu Rev Ecol System 23:263–286.
13. Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF. 1995. DNA-sequence evolution—the sounds of silence. Philos Trans R Soc B-Biol Sci 349:241–247.
14. Eyre-Walker A. 1991. An analysis of codon usage in mammals: selection or mutation bias? J Mol Evol 33:442–449.
15. Smith NGC, Hurst LD. 1999. The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? Genetics 152:661–673.
16. Urrutia AO, Hurst LD. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159:1191–1199.

17. Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. Genome Res 13:2260–2264.
18. Lavner Y, Kotlar D. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene 345:127–138.
19. Comeron JM. 2004. Selective and mutational patterns associated with gene expression in humans: Influences on synonymous composition and intron presence. Genetics 167:1293–1304.
20. Comeron JM. 2006. Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans. Proc Natl Acad Sci USA 103:6940–6945.
21. Chamary JV, Hurst LD. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: Evidence for selectively driven codon usage. Mol Biol Evol 21:1014–1023.
22. Kondrashov FA, Ogurtsov AY, Kondrashov AS. 2006. Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. J Theor Biol 240:616–626.
23. Lu J, Wu CL. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc Natl Acad Sci USA 102:4063–4067.
24. Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B, et al. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res 13:831–837.
25. dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32:5036–5044.
26. Schattner P, Diekhans M. 2006. Regions of extreme synonymous codon selection in mammalian genes. Nucleic Acids Res 34:1700–1710.
27. Parmley JL, Hurst LD. 2007. How common are intragene windows with KA > KS owing to purifying selection on synonymous mutations? J Mol Evol: (in press).
28. Iida K, 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.
29. Xing Y, Lee C. 2006. Alternative splicing and RNA selection pressure—evolutionary consequences for eukaryotic genomes. Nat Rev Genet 7: 499–509.
30. Hurst LD, Pal C. 2001. Evidence for purifying selection acting on silent sites in BR CA1. Trends Genet 17:62–65.
31. Orban TI, Olah E. 2001. Purifying selection on silent sites - a constraint from splicing regulation? Trends Genet 17:252–253.
32. Pagani F, Raponi M, Baralle FE. 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci USA 102:6368–6372.
33. Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2:E268.
34. Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J Mol Evol 62: 89–98.
35. 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.
36. Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. PLoS Biol 5:343–353.
37. Hurst LD. 2006. Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. J Mol Evol 63:174–182.
38. Diatchenko L, Slade GD, Nackley AG, Bhalang K, Sigurdsson A, et al. 2005. Genetic basis for individual variations in pain perception and the development of a chronic pain condition. Hum Mol Genet 14:135–143.
39. Takyar S, Hickerson RP, Noller HF. 2005. mRNA helicase activity of the ribosome. Cell 120:49–58.
40. Duan J, Antezana MA. 2003. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J Mol Evol 57:694–701.
41. Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol 6:R75.
42. Cohen B, Skiena S. 2003. Natural selection and algorithmic design of mRNA. J Comput Biol 10:419–432.
43. Seffens W, Digby D. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res 27:1578–1584.
44. Kozak M. 2002. Pushing the limits of the scanning mechanism for initiation of translation. Gene 299:1–34.
45. Purvis IJ, Bettany AJ, Santiago TC, Coggins JR, Duncan K, et al. 1987. The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. J Mol Biol 193:413–417.
46. Komar AA, Lesnik T, Reiss C. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Lett 462:387–391.
47. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol 4:933–942.
48. Sun M, Hurst LD, Carmichael GG, Chen J. 2006. Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcription and organismic complexity. Genome Res 16:922–933.
49. Lipman DJ. 1997. Making (anti)sense of non-coding sequence conservation. Nucleic Acids Res 25:3580–3583.
50. Shen LX, Basilion JP, Stanton VP Jr. 1999. Single-nucleotide polymorphisms can cause different structural folds of mRNA. Proc Natl Acad Sci USA 96:7871–7876.
51. Duan JB, Wainwright MS, Comeron JM, Saitou N, Sanders AR, et al. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum Mol Genet 12:205–216.
52. Shabalina SA, Ogurtsov AY, Spiridonov NA. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res 34:2428–2437.
53. Cortazzo P, Cervenansky C, Marin M, Reiss C, Ehrlich R, et al. 2002. Silent mutations affect in vivo protein folding in Escherichia coli. Biochem Biophys Res Commun 293:537–541.
54. Thanaraj TA, Argos P. 1996. Ribosome-mediated translational pause and protein domain organization. Protein Sci 5:1594–1612.
55. Oresic M, Shalloway D. 1998. Specific correlations between relative synonymous codon usage and protein secondary structure. J Mol Biol 281:31–48.
56. Thanaraj TA, Argos P. 1996. Protein secondary structural types are differentially coded on messenger RNA. Protein Sci 5:1973–1983.
57. Jia MW, Luo LF. 2006. The relation between mRNA folding and protein structure. Biochem Biophys Res Commun 343:177–182.
58. Netzer WJ, Hartl FU. 1997. Recombination of protein domains facilitated by co-translational folding in eukaryotes. Nature 388:343–349.
59. Oresic M, Dehn M, Korenblum D, Shalloway D. 2003. Tracing specific synonymous codon-secondary structure correlations through evolution. J Mol Evol 56:473–484.

Appendix 2. **Splicing related constraints on protein composition and evolution are common within the metazoa**

Tobias Warnecke, Joanna L. Parmley and Laurence D. Hurst

Abstract

Background

Splice-regulatory domains in mammalian exons, associated with SR-protein binding, have been found to impose marked trends on the relative abundance of certain amino acids as one approaches the exon-intron boundary. But are such splicing-related constraints restricted to mammals or a common feature of eukaryotic genomes?

Results

We find preference and avoidance of certain amino acids near exon-intron boundaries throughout the metazoa. There is extensive cross-species concordance as to which amino acids are affected, the relative strength and direction of trends. Indicative of functional importance, rates of protein evolution are lower near exon-intron boundaries and, consequently, in genes with small exons. Patterns of composition bias are typically well predicted with knowledge of exonic splice enhancers. This fits with a dearth of significant abundance trends in two yeast species (*S. pombe, S. cerevisiae*), believed to lack SR-protein involvement in splicing. The analysis also indicates that 5' ends of nematode exons deviate radically from norm: amino acids strongly preferred near boundaries are strongly avoided in other species, and vice versa. This we suggest is a measure to avoid attracting *trans*-splicing machinery, which processes 5' ends of numerous nematode genes.

Conclusion

Amino acid usage near exon-intron boundaries exhibits largely similar biases across the metazoa. Absent in yeasts, these biases accord with sequence preferences of SR proteins, suggesting that splicing has imposed constraints upon protein-coding sequence, unrelated to its biological function, across the metazoa. These results have implications for inferring aspects of the mechanism of splicing given nothing more than a well-annotated genome.

Background

The maxim that "form follows function", dogmatically adhered to in some early 20th century design and architecture, refers to the idea that the final function of a product should be the only determinant of its design. Phenotypic products of evolutionary processes have also frequently been analyzed in this seductively simple framework. However, costs of production, the availability of raw materials, and other factors regularly lead to marketable goods being suboptimally designed as far as their immediate function is concerned. Likewise, proteins tend to employ metabolically cheap amino acids [1] and in for example mammals, amino acid content of a protein reflects localized GC content [2]. The need to encode, in exonic sequence, information relevant for correct splicing is another factor that might have the potential to influence protein composition [3]. Located in the exonic parts of primary mRNA transcripts, exonic splicing enhancers (ESEs) are short (6-8nt) nucleotide motifs, which have been established as a core component of the pre-mRNA splicing mechanism in metazoans [4]. Playing a critical role in constitutive as well as alternative splicing [5], they function at multiple stages of spliceosome assembly by interacting with corresponding RNA recognition motifs harbored in the N-terminal end of SR (Serine-Arginine) proteins. The exact *modus operandi* of SR proteins in splicing has yet to be fully resolved but they appear to be critical for establishing, in conjunction with other proteins, cross-exon complexes that enable faithful communication between splice sites (see Blencowe 2000 for a review of functional hypotheses).

Recognition of exonic alongside intronic sequence motifs has been proposed to be particularly pivotal in organisms where a majority of exons are flanked by much larger introns, allowing exons to be efficiently identified and not lost in a sea of intronic sequence [6]. Furthermore, whereas in *Saccharomyces cerevisiae* splice sites and branch point sequences show a high degree of conservation to ensure the intron is correctly targeted by the splicing machinery, these recognition motifs tend to be less well conserved in multicellular organisms [7].

Experimentally raising the number of natural exonic enhancer sites leads to an additive increase in splicing activity [8]. Importantly, ESEs function in a position-dependent manner, their efficiency in catalyzing splicing decreasing with increasing distance from the splice site [9, 10] The significant enrichment for GAA (a codon known to be

overrepresented in ESEs) compared with the synonymous GAG near exon-intron boundaries is consistent with this finding [10, 11].

A recent study by Parmley et al. (2007) suggests ESEs have also left an imprint on the amino acid composition of proteins. Exploring exonic sequences adjacent to exon-intron boundaries in human and mouse, the authors reported marked trends in the relative abundance of certain amino acids when one moves away from the boundary. Some amino acids, such as lysine (K) and isoleucine (I), are strongly preferred near boundaries whereas others, such as proline (P) and alanine (A), are significantly avoided (for a full list see Table 1). This is the case for both 5' and 3' ends of exons. Considering separately the two-fold and four-fold blocks of the six-fold degenerate amino acids, the authors also showed that these trends are owing to avoidances/preferences at the nucleotide level and that there is a high degree of correspondence between the codons preferred and their involvement in computationally predicted and experimentally verified ESEs.

But are these trends a peculiarity of mammals or are they common in other taxa? Does the presence or absence of trends correspond to what is known about the significance of exonic splicing regulation in each species? For example, a recent survey of several eukaryote genomes showed the SR protein family to be greatly expanded in metazoans but scarcely represented in unicellular genomes [12]. A failure to find preference trends in *S. cerevisiae*, an organism lacking SR proteins [13], might corroborate the hypothesis that preference patterns are indeed caused by ESEs. Moreover, if there are discernible trends in other species, do we repeatedly see the same amino acids avoided or preferred or are trends largely unique to each species? Also, are mammals unusual in showing a tight correlation between 5' and 3' trends, and may divergent results bear implications for the workings of the splicing machinery? Here we examine these issues with exon data from a broad range of animals and fungi.

Results

a) Preference trends are rare in yeasts but widespread in multicellular species

For nine metazoan species (Human (Hs), mouse (Mm), *Danio rerio* (Dr), *Caenorrhabditis elegans* (Ce), *Caenorrhabditis briggsae* (Cb), *Anopheles gambiae* (Ag), *Drosophila melanogaster* (Dm), *Apis mellifera* (Am)), one plant *(Arabidopsis thaliana* (At)) and two fungi (*Saccharomyces cerevisiae* (Sc), *Schizosaccharomyces pombe* (Sp)) we examined trends in amino acid composition as one approaches the exon-intron boundary. As splice site signals can extend into exons and our main focus is on ESE-dependent splicing regulation, we removed the first full codon at the exon-intron boundary (see Materials and Methods). Thereafter, rank correlations (rho) between distance from the boundary (34 codons into the exon, see Materials and Methods) and proportional usage of the amino acid were computed independently for 5' and 3' regions of exons. Further, for all amino acids independently we fitted a linear regression extracting the slope of the line to be used as a crude diagnostic for the strength of amino acid preference/avoidance. Figure 1 illustrates the different types of relationship recovered from the data.

Two-fold and four-fold blocks of the six-fold degenerate amino acids were considered as distinct groupings so that a total of 46 tests (23 amino acid groups 5' and 3') were carried out for each species. Table 1 gives a comprehensive by-species overview of amino acid preferences/avoidances, significant after Bonferroni correction (N=46 comparisons, $P<0.0011$). Supplementary Table 1 contains the complete set of rank correlations for all amino acids for all 11 species.

The most conspicuous feature of Table 1 is arguably the commonality of trends in the metazoa and the scarcity of trends in the yeast species. The two-fold block of leucine (L*) in *S. cerevisiae* is the only amino acid grouping exhibiting a significant preference trend (rho= -0.4482, $P< 0.0003$). This is in stark contrast to the suite of metazoan eukaryotes where an extensive range of avoidance and preference trends is observed, covering the complete set of

amino acids. Only three multicellular species display fewer than 13 significant trends (Dm, Ag, At) whereas five (Hs, Mm, Ce, Cb, Am) display more than 20. These quantitative differences are not a function of the different number of exons obtained from each species (a Spearman rank correlation between the number of significant trends with the number of exon ends (combined 5' and 3') yields: rho=0.2259, P=0.5517). For Dm and Ce, we also tested whether the results might be biased as a result of exon homology, but in either case found amino acid abundance patterns at exon ends to be virtually identical in a set of homology-reduced genes (Dm: N=8840; Ce: N=11790), with minor differences in amino acid spectra owing to small changes in the P values for amino acids originally close to the significance threshold (Supplementary Table 2, Supplementary Document 1).

b) Cross-species patterns

Whilst the spectra of amino acids preferred/avoided by individual species are ultimately unique in breadth (how many trends) and composition (which amino acids are affected), there is considerable cross-specific overlap in terms of whether a particular trend is present at all, its direction, and relative strength (as measured by the slope of the line of best fit): A tight agreement in breadth and composition was already reported by Parmley et al. (2007) comparing human and mouse; Table 1 illustrates that this particular agreement is virtually perfect, with marginal differences in the relative strength of individual trends, and that directionality is conserved throughout. Considering zebrafish (*Danio rerio*) as the only other vertebrate in our sample alongside these species, we notice that its spectrum is slightly diminished in breadth and contains a few trends not seen in the two mammals (G(3'), V (5',3')). However, overall concordance in composition and strength is still remarkably good, and the "mammalian pattern of directionality" perfectly adhered to.
Regarded as a pair the two nematode species almost match the human-mouse dyad in terms of overall concordance of preference patterns, with the preference spectrum of *C. briggsae* slightly broader and strength marginally different but directionality conserved between the two.

For the most part, the patterns of preference/avoidance are repeatable across species. Again considering 5' and 3' ends separately, Table 2 contains data of pairwise comparisons between species giving rank correlations (rho) for the slopes derived from all 23 amino acid groupings. For the vertebrate group both 5' and 3' correlations are very high (all rho>0.9, all P<1.81E-06; 72 tests, significance threshold: P<6.94E-04), with human and mouse in almost perfect agreement. More remarkably, however, some correlations of substantial strength also exist 3' between the vertebrates and, for example, Anopheles (all rho>0.87, all P<2.94E-06) and Drosophila (all rho>0.75, all P<2.9E-05). 3' correlations are less impressive for the remaining species (Am, At) but Apis boasts remarkably strong 5' correlations with the vertebrates (all rho>0.75, all P<4.11E-05). Focusing on specific amino acid trends, isoleucine (I) stands out in that it is strongly preferred near 3' boundaries across all species. Although no other individual trend can issue the claim to be universal across all species in our sample, some are well represented through the entire phylogeny, for example 5' avoidance of glutamine (Q), 3' preference of phenylalanine (F), and 3' avoidance of the four-fold block of arginine (R).

c) Deviant nematodes

The striking cross-species concordance in preference patterns makes one observation all the more puzzling. The nematode 5' spectra behave in a highly counterintuitive manner in that the "mammalian pattern of directionality" is violated on several occasions: Where we do find significant trends in nematodes and other species (E, K, L*, Q, R, R*, T), all but glutamine (Q) show discrepant directionality (Table 1). For example, whereas lysine (K) is strongly preferred near boundaries in vertebrates and some insects (Dm, Am), it appears to be strongly avoided in the 5' region of nematode exons (Figure 2). Table 2 also underlines the exceptional position of nematodes: 5' correlations between either nematode and any other species are pervasively negative with some of sizable magnitude (rho(Cb~Hs)≅-0.6). Although no single correlation is significantly different from zero applying the adjusted significance threshold (P<6.17E-04), the pervasiveness of this pattern is nonetheless noteworthy. This is especially

true given that, strikingly, the same is not the case for the 3' spectra where we find a coherent agreement between nematodes and vertebrates (minimum rho>0.65, all significant at P<5.92E-04) and only the two-fold block of serine (S*) shows a reverse pattern of directionality among the significant trends for individual amino acids.

d) Most species obey an approximately symmetric pattern of preference trends 5' and 3'

This curious discrepancy between the 5' and 3' spectra of amino acid trends in nematodes led us to investigate further the relationship of 5' and 3' patterns across species. Considering all amino acid trends simultaneously, rank correlations between slope coefficients (5'~3') were computed. Furthermore, we wanted to explicitly test the hypothesis that preference trends show a "symmetric" behaviour, i.e. that individual amino acids exhibit preference trends of similar strength and direction at 5' and 3' ends. To this end, we carried out standardized major axis regressions (SMA) for 5' versus 3' trends in each species and compared the resulting regression lines with one expected under perfect symmetry (y=x). The results are given in Table 3 and graphically represented in Figure 3. Human and mouse show very substantial positive correlations between 5' and 3' preference trends (Hs: rho=0.8528, P=1.96E-06; Mm: rho=0.8626, P=2.28E-06). Although diminished in strength, we also see significant correlations for Drosophila and Danio, with Anopheles almost significant. As expected from the previous analysis, correlations for nematodes are negative, albeit not significantly so (Ce: rho=-0.1413, P=0.5185; Cb: rho=-0.4358, P=0.0388). However, the SMA results allow us to confidently reject any notion of *C. elegans* or *C. briggsae* adhering to a symmetric pattern of amino acid usage, the respective confidence intervals ruling out a symmetry slope of β=1 (CI (Ce): [-1.118;-0.7309]; CI (Cb): [-0.7474; -0.5139]). In contrast, no other species for which an SMA could be carried out (see Table 3) significantly deviates from a symmetric model.

e) Amino acid trends are largely consistent with participation in ESEs motifs

Intriguingly, asymmetries in the amino acid composition of nematode exon ends appear to be mirrored by a corresponding asymmetry of regulatory motifs. Robinson (2005), using a computational approach to characterize candidate ESEs in *C. elegans,* found that 5' and 3' ends were distinguished by different classes of consensus motifs [14]. Crucially, he found purine-rich human-like candidate motifs to be associated with 3' ends but not 5' ends of nematode exons, which is broadly consistent with our observation that amino acids encoded by purine-rich codons tend to be, in contrast to other animals, disfavored at 5' ends (Table 1, Figure 3).

For mammals, the prediction that amino acids preferred near boundaries should be disproportionately frequent in ESEs was tested by Parmley et al (2007). The authors defined a measure of relative involvement of amino acids in splice enhancer hexamers. As predicted, these hexamer preference indices (HPI), computed for each amino acid grouping, were found to correlate with preference trends, strongly preferred amino acids on average associated with higher HPI values.

This relationship held true for human as well as murine ESE sets and amino acid trends, considering either rank correlation coefficients (rho_x) (Hs HPI~ rho_x: rho=-0.54, P<0.00001, N=46; Mm HPI~ rho_x: rho= -0.49, P=0.0005, N=46) or the slope (β) of the fitted linear model (Hs HPI~β: rho=-0.57, P<0.0001, N=46; Mm HPI~β: rho=-0.52, P=0.0002, N=46).

However, when we derived HPIs for Danio (zebrafish), using a set of ESEs obtained from the same source (http://genes.mit.edu/burgelab/rescue-ese/), we were surprised to find a significant correlation of reverse sign (Dr HPI~ rho_x (5'): rho=0.6, P<0.003, N=46; HPI~rho_x (3'): rho=0.59, P<0.0033, N=46). This was particularly unexpected given the substantial agreement between mammalian and zebrafish preference spectra (Tables 1 and 2). Many experimentally verified ESEs have been characterized as A-rich and C-poor relative to the background frequency of these nucleotides in coding sequence. Whilst we found this to be the case for putative human ESE motifs *not* shared with zebrafish (A: 47.38% (ESE) v 25.57% (exonic); C: 15.28% v 25.99%, N(ESE)=204), and for ESEs present in *both* species (A: 50% v 25.57%; C:

6.37% v 25.99%, N=34), unique zebrafish ESEs (i.e. ESEs not present in human) from this dataset were unusually enriched in C (39.47% v 25.99%, N=288) and relatively poor in A (18.40% v 25.57%). Re-examination of these putative zebrafish ESEs may be worthwhile.

f) Reduced rates of evolution near the exon-intron boundary in species where ESEs are essential components of the splicing machinery.

To further advance the hypothesis that gradients in amino acid abundance near exon-intron boundaries are a critical feature of exon ends in metazoans, we examined the degree of amino acid conservation as a function of distance from the boundary. For three pairs of species (*S. cerevisiae – Saccharomyces castelli; D. melanogaster – Drosophila pseudoobscura (Dps); C. elegans – C. briggsae*) sets of orthologous internal exons were derived from various sources and aligned at the amino acid level (see Materials and Methods). Supporting results reported in Parmley et al. (2007) of markedly higher levels of amino acid conservation near boundaries for a set of orthologous vertebrate exons (Hs~Mm comparison), we found strong and highly significant positive correlations of strikingly linear character (Figure 4) between distance from the boundary and amino acid substitution rate for the Drosophila pair and the Caenorhabditis pair, whilst proximity to the boundary did not appear to confer a higher level of amino acid conservation in the *Saccharomyces* comparison. Restricting the analysis to exons of at least 70 codons in length, we obtain qualitatively equivalent results (Drosophilae: 5': rho=0.53, $P<0.002$; 3': rho=0.77, P=9.70E-07; N=3690; Caenorhabdites: 5': rho=0.74, P=2.33E-06; 3': rho=0.58, P=4.5E-04; N=6273). This restriction ensures that all exons contribute an approximately equal share of information to each codon position from the boundary and eliminates the potential confounder that short exons might, for reasons unrelated to splicing, feature more frequently in highly conserved genes and consequently create an artefactual trend by virtue of their disproportionate contribution to substitution rate information closer to the boundary.
Given that the set of aligned Saccharomyces exons consisted entirely of terminal exons (see Materials and Methods), we repeated the analysis for a set

of 5352 orthologous pairs of terminal exons from our Drosophila dataset in order to rule out that differences are caused by any special characteristics of terminal exons. Correlations observed for terminal exons closely resemble those for internal exons (5': rho=0.83, P=3.8E-07; 3': rho=0.75, P=1.95E-06), alleviating any such concerns.

The above results appear consistent with greater functional significance of boundary-proximal amino acid composition in metazoans, proposed to be at least in part owing to their more extensive utilization of exonic splice regulatory sequences. However, after repeated (k=10000) random sampling of 90 aligned terminal exons from the Drosophila dataset and subsequent statistical analysis, we cannot reject the possibility that the Saccharomyces statistics were sampled from the same underlying distribution (see Supplementary Figure 1 for a detailed explanation), implying that differences in conservation near exon-intron boundaries cannot be ultimately established from the data at hand.

Having detected higher levels of amino acid conservation near exon-intron boundaries, we might expect genes with a high proportion of sequences near boundaries ("flank-heavy") to evolve more slowly. This is indeed what we found when we considered K_A as a function of the proportion of sequence within 70 base pairs (bp) of the boundary (Drosophilae: rho=-0.26, P=2.2E-16, N=4132; Caenorhabdites: rho=-0.08, P=6.18E-09, N=5248; Fig. 5). We report K_A rather than K_A/K_S, more commonly used as a measure of selection on protein sequence, because we know that the underlying premise of K_A/K_S, namely that K_S reflects neutral rates of evolution, is violated for sequence encoding ESEs [15].

The results are not qualitatively affected by contracting (50bp) or expanding (100bp) the region considered to constitute the boundary flank (Supplementary Table 3). Focusing on the terminal bins in Figure 5A, it appears that between *D. melanogaster* and *D. pseudoobscura* a gene with less than 10% of coding sequence near an exon-intron boundary evolves on average almost twice as fast (mean K_A = 0.195) as a gene with more than 70% of boundary-proximal sequence (mean K_A = 0.099). Discrepancies in evolutionary rate between "flank-heavy" and "core-heavy" bins appear less marked for the nematode pair (mean K_A (%CDS near boundary>0.9) = 0.12; mean K_A (%CDS near boundary

<0.3) = 0.18). However, Figure 5B suggests that this is principally owing to curiously elevated levels of conservation for genes with a small proportion of sequence near the boundary, i.e. genes with very large exons, a feature we did not encounter in the analysis of either insect (Dm-Dps) or mammalian (Hs-Mm) orthologues [3].

Importantly, this anomaly highlights a more general reservation, namely that any measure capturing the proportion of sequence near the boundary will strongly covary with exon length, which in turn might covary with underlying functional determinants of evolutionary rate entirely unrelated to splicing control. Thus, in order to control for any putatively distorting effects of functional class on K_A, we employed the following strategy: For each aligned gene, we concatenated the flanking regions of all exons, 5' and 3', defined as the first 72bp bordering the exon-intron junction of trimmed exons. By implication, genes with no exon larger than 144bp had to be excluded from this analysis. Concurrently, we concatenated the core sections of all exons of sufficient length in the respective gene, defined as the sequence block enclosed by the two 72bp flanking regions. As accurate estimation of K_A probably requires a minimum of 100 codons, we further restricted analysis to those genes with at least 300bp in the concatenated flanks and in the concatenated cores of exons. For each gene meeting the above criteria we then determined the rates of amino acid evolution in the concatenated core sections (K_{Ac}) and flanking sections (K_{Af}) (Figure 5). We find that more Drosophila orthologous genes than expected by chance have faster evolving core regions (median $(K_{Ac}-K_{Af})/ K_{Af}$) = 0.14, Wilcoxon signed rank test $P<0.0001$, N=1237), consistent with the evidence, presented above, for additional sequence constraint operating on flanking regions. A significant tendency towards more rapid evolution in core sections is also evident when we confine the sample to genes with at least 600bp in flanking as well as core regions (median $(K_{Ac}-K_{Af})/ K_{Af}$) = 0.14, Wilcoxon signed rank test $P<0.0001$, N=785). Despite exhibiting the expected shift towards average higher K_A in the core of exons, this trend is much less pronounced than in a previously reported comparison of human-mouse orthologues (median $(K_{Ac}-K_{Af})/ K_{Af}$) = 0.68, Wilcoxon signed rank test $P<0.0001$, N=360, Figure 6, see Parmley et al. (2007) for details). Curiously,

for the nematode pair, we find significant evidence for a reverse correlation (300bp: median $(K_{Ac}-K_{Af})/K_{Af}) = -0.07$, Wilcoxon signed rank test $P<0.0001$, N=1102; 600bp: median $(K_{Ac}-K_{Af})/K_{Af}) = -0.014$, $P<0.038$, N=496), i.e. in the majority of genes, flanking regions evolve at a marginally higher rate than core regions.

Discussion

a) General trends

Parmley et al. (2007) recently presented evidence that, in mammals, amino acid usage in the vicinity of exon-intron boundaries is affected by factors unrelated to protein function but to sequence-based information required for correct splicing. The objective of the present study was to elucidate whether such requirements have left an evolutionary imprint on exonic sequence composition across a phylogenetically diverse set of species. To this end, we systematically compared trends in relative amino acid abundance near exon-intron boundaries in eleven eukaryotic species. Our analysis revealed that preference for or avoidance of certain amino acids near boundaries is a common phenomenon among metazoan species. Species-specific spectra of significant preferences/avoidances show unmistakable signs of conservation along several dimensions: composition, relative strength, and directionality. The concordance in directionality (whether an amino acid is preferred or avoided) is particularly impressive in that we observe very few deviations from the mammalian pattern even in only distantly related species.

We do not claim that the systematic patterns we observe are solely caused by a selected preference for codons involved in ESEs. In fact, composite trends are almost certain to be the result of multiple functional constraints, including the need to incorporate in the sequence various enhancer and suppressor elements, but also to avoid intron-specific enhancer motifs (for example GGG in

mammals, [16]). Furthermore, abundance trends could partially be the result of cryptic splice site avoidance as suggested by Eskesen and colleagues [17]. However, many of the trends observed, for example cytosine avoidance near boundaries, are not predicted by this model [3, 11].
Introns associate non-randomly with the codon in direct proximity to the splice site in a phase-specific manner, an observation often described as insertional preference [18]. Trimming and elimination of the first full codon should guard against picking up such insertional preferences or an extended splice site consensus. We cannot exclude the possibility that some boundary-proximal codons have slipped into our dataset owing to poor splice site annotation. However, it must be pointed out that this reservation applies only to the subset of amino acid trends that show biased usage directly adjacent to introns and might be more relevant to the interpretation of local discontinuities (see Materials and Methods). Also, if the above-mentioned explanations were of major relevance, we would expect cryptic splice site avoidance, insertional preference, and (to a lesser extent) poor splice site annotation to cause similar patterns in yeasts, in particular *S. pombe* for which a dataset of reasonable size is available. This is not the case.

Establishing to what extent these trends are caused by preference for ESEs will ultimately depend on characterizing species-specific catalogues of ESE/ESS motifs together with their corresponding *trans*-factors and relating these to the observed spectra of preferred/avoided amino acids. This work, in particular relating to tissue- and stage-specific splicing patterns, is still in its infancy [19], the catalogues currently available restricted to a small number of vertebrates and yet to be fully verified experimentally [20, 21].
However, the dearth of significant trends in *S. cerevisiae* and *S. pombe* strengthens the proposition that preference trends principally reflect requirements to accommodate ESEs: Although the *S. cerevisiae* genome codes for an SR protein kinase (Sky1p) with the capacity to phosphorylate mammalian arginine-serine rich (RS) domains, the likely endogenous substrate (SR-like protein Npl3p) does not appear to be involved in pre-mRNA splicing [4, 22]. Importantly, no splicing factors homologous to metazoan SR proteins have been

discovered in *S. cerevisiae* (Graveley 2000), consistent with the classical view that splicing in budding yeast is regulated intronically, whilst the SR protein family is greatly expanded in metazoans [12]. This is further consistent with the observation that splice site consensus is generally highly conserved. The fact that our analysis revealed a significant 3' trend for the two-fold block of leucine (L*) might hint at the presence of recognition motifs in yeast exonic sequence. However, at present there is no evidence supporting the regular involvement of an ESE-like binding motif in *S. cerevisiae* splicing and alternative explanations for this pattern, both functional and neutral, should be considered.

Splicing in *S. cerevisiae* is moderately common in quantitative terms because many highly expressed genes, notably encoding ribosomal proteins, contain introns, so that over 25% of the mRNA population are spliced [23]. However, in over 6000 *S. cerevisiae* genes we find less than 300 introns in total, so that splicing can hardly be considered a representative processing stage on a genomic scale. In contrast, splicing is much more prevalent in *S. pombe* where ~40% of genes contain introns [24]. Basal splicing proteins show an enhanced similarity to their mammalian homologues and two SR protein homologues (Srp1p, Srp2p) have been identified [25-27]. Unlike in budding yeast, there is recent evidence that Srp2p binds to specific exonic elements and interacts with the fission yeast orthologue of human splice factor U2AF [28]. Why then, given that SR-ESE-like interactions seem to exist in *S. pombe*, do we not find any trends for amino acid or codon preference in this species? We suggest that trends may be lacking for two reasons: firstly, given the comparatively low level of splice site consensus degeneracy, a minimal number of ESEs might be sufficient to ensure correct splicing. On a genomic level, we might then fail to register biased abundance patterns on the spatial scale investigated in this study. Secondly, for clear-cut preference trends to evolve a minimum level of splice-regulatory complexity might be required: Alternative splicing or a more complex gene structure comprising several introns, where regulatory elements would frequently compete for precedence if arranged close to each other, could be envisaged as an evolutionary pressure initially driving the diversification of ESEs and corresponding *trans*-factors, thereby creating an environment in

which strong trends might be required to attract or repel the correct set of *trans*-factors, both for constitutively and alternatively spliced genes. Consistent with this tentative hypothesis, reports of alternative splicing in *S. cerevisiae* [29] and *S. pombe* [30] are restricted to singular cases, for which functionality of the recovered alternative splice products remains to be shown [31]. However, attempts to link diversity and density of ESEs to alternative splicing have so far yielded ambiguous results [32].

The absence of preference patterns in yeasts has an important practical implication. Determining whether amino acid trends are present near exon-intron boundaries can be used as a reliable indicator for whether a particular species employs ESE-based splicing regulation, certainly on a genomic scale, without prior knowledge of specific binding motifs or *trans*-factors.

b) Nematode exceptionalism in an ESE framework – is *trans*-splicing to blame?

The fundamental deviation from the "mammalian pattern of directionality" shown by the 5' amino acid trends in nematode exons (Table 1) is, at first sight, unexpected. There are extensive homologies between vertebrate and nematode basal splicing machineries on the protein level [12]. Furthermore, splicing in SR-depleted cells of the *Caenorrhabditis* relative *Ascaris lumbricoides* can be rescued by adding SR proteins derived from non-nematode (HeLa) whole cell extracts, supporting at least a minimum degree of functional overlap [33]. Thirdly, the high level of conservation between SR and SR-like proteins identified in each species explicitly includes the RNA recognition motifs, tentatively suggesting similar binding specificities [34].
There is, however, one feature of the nematode splicing process that sets it apart from the other species in our sample: A substantial proportion (~70%) of *C. elegans* (and *C. briggsae*) genes are *trans*-spliced [35]. In this process a short (22nt) 5' snRNA fragment, the spliced leader (SL), which is transcribed from a different genomic locale, is added at the 5' end of the pre-mRNA [36]. It would, we suggest, be highly disadvantageous for this *trans*-splicing machinery to act at the 5' end of exons, where *cis*-splicing should occur. Indeed, were trans-

splicing to occur where intron removal should occur, a gene would in effect be broken in two. Thus, we suggest that 5' ends of internal exons have evolved to ensure that they do not attract the *trans*-splicing machinery. Given that the trans-splicing machinery is ubiquitously present in a cell, all 5' ends of internal exons, be they from trans-spliced genes or not, should be equally under pressure to avoid trans-splicing occurring where cis-splicing should happen. Consistent with this expectation, the trends seen at 5' and 3' ends in internal exons are the same in genes from operons and those not in operons (data not shown).

What might be the proteins involved in trans-splicing? There is good evidence that several stages of the *trans*-splicing process are, like *cis*-splicing, critically supported by SR proteins [33, 37]. Furthermore, whilst mammalian and *Ascaris* SR extracts are equally efficient in catalyzing *cis*-splicing in vitro, *Ascaris* SR extracts engender an approximately five-fold higher *trans*-splicing activity (Sanford and Bruzik 1999). Although the use of whole cell extracts in these experiments precludes an analysis of the differential contribution of individual SR proteins, the above observations are consistent with the hypothesis that a subset of splice-regulatory proteins in these species is dedicated to *trans*-splicing.

Given the above, we envisage *trans*-splicing specific SR and other proteins to interact primarily with intergenic sequence upstream of the first exon of the pre-mRNA to provide further guidance for the *trans*-splicing apparatus or mediate other functions crucial to *trans*-splicing such as protecting downstream RNA from degradation [35, 38]. A prediction derived from this model is that we should find in nematodes proteins (likely to be SR proteins) participating in *trans*-splicing should bind to nucleotide motifs depleted of codons from amino acids avoided near the 5' end of exons.

c) Symmetric exons?

Owing to their deviant 5' trends nematodes stand out in another aspect of systematic amino acid biases. Parmley et al. (2007) observed no significant differences in preference trends between 5' and 3' ends of exons in human and mouse. Similarly, approximate symmetry has been reported for ESE distribution

in human exons [20]. Conversely, standardized major axis regressions strongly suggest that nematodes do not conform to a symmetric pattern of preference trends.

An assessment of this situation very much depends on how we expect ESE-guided splicing regulation to work on a mechanistic level. If SR proteins are assumed to interact directly with specific components of the basal splicing machinery, as is probably the case for U2AF [4], we would not automatically expect the same ESEs (and by implication amino acid trends) to be represented at similar frequencies 5' and 3' where different spliceosomal proteins are present. Predictions of whether symmetry might be of functional relevance, however, especially for scenarios of indirect interaction, are difficult to derive from the data at hand. Nematodes seem to cope just fine without symmetry, perhaps indicating that symmetry might be an incidental by-product of equal representation of ESEs at both exon ends rather than functional in itself. Confidence intervals in our exploration of symmetry are large so that we do not want to suggest that symmetry is a dominant pattern throughout our species sample. However, some best estimates of SMA slopes (β) are tantalizingly close to perfect symmetry (Mm: β=0.9907, Hs: β=1.0362, Dr: β=1.0439, Ag: β=1.0788), warranting more detailed examination of this potentially functional signature in the future.

d) Patterns of amino acid evolution

Consistent with the proposition that trends in relative amino acid abundance near exon-intron boundaries are functionally important, we observe lower rates of nonsynonymous evolution near those junctions in insects (Dm-Dps), nematodes (Ce-Cb) and mammals (Hs-Mm), indicative of higher selective constraint in this region. Furthermore, for the above species pairs, the proportion of coding sequence that is located near boundaries is a partial predictor of K_A (Fig. 5). Genes with a higher share of sequence partaking in exon flanks tend to show reduced rates of evolution. Nematode genes, again, stand out in that they do not conform to the negative linear relationship between K_A and flank-heaviness established by the analysis of other species pairs (Hs-

Mm and Dm-Dps), but show unexpectedly high levels of conservation for genes with very large exons. The causal mechanisms behind this currently remain elusive. Similarly, we would not have predicted that in worms gene-specific differences between evolutionary rate in flanking and core section of exons are biased (if only slightly) towards more rapid evolution of flanking regions. However, the distribution of core-flank evolutionary rate differentials in worms appears more comparable to the one for flies, a higher median evolutionary rate of core regions in the latter notwithstanding (Figure 6). Human-mouse orthologous genes on the other hand show a much more dramatic distributional shift towards faster evolution in exon cores (see distributions in Figure 6). Between-taxa differences in gene composition, especially relating to the presence of more and longer introns in mammals, might account for these differences: on a speculative note, splice-relevant information – or indeed all information necessary to distinguish an exon from surrounding non-coding sequence – might require a unique degree of conservation under these circumstances, perhaps severely restricting the leeway for non-synonymous changes to occur in flanking regions. Alternatively, restrictions imposed by our experimental set-up, especially relating to minimum sequence length requirements, might have resulted in the selection of gene sets with divergent splicing characteristics in the different species pairs. We leave a closer dissection of these questions to further analysis.

Conclusion

Biased usage of amino acids in the vicinity of exon-intron boundaries is a common feature in metazoan genes, with the direction of biases largely consistent between taxa. That the biases accord with sequence preferences of SR proteins and that such biases do not exist in yeasts, support the view that dual coding of DNA in exons, to specify both which amino acids to employ and where introns are to be removed, is a common feature of metazoan species. In nematodes, the possible relationship between trans-splicing and the exceptional departure from the mammalian pattern of amino acid trends at the 5' end of exons deserves further scrutiny. This exception aside, the results presented here

suggest a simple diagnostic for the involvement of SR proteins in splicing given nothing more than a well-annotated genome.

Materials and Methods

Relative amino acid abundance near exon-intron boundaries

For eleven species (Human (Hs), mouse (Mm), *Danio rerio* (Dr), *Caenorrhabditis elegans* (Ce), *Caenorrhabditis briggsae* (Cb), *Anopheles gambiae* (Ag), *Drosophila melanogaster* (Dm), *Apis mellifera* (Am), *Arabidopsis thaliana* (At), *Saccharomyces cerevisiae* (Sc), *Schizosaccharomyces pombe* (Sp)) we established independent exon data sets derived from a small number of databases (Supplementary Table 4). Pre-established CDS tracks were followed in all but two cases (At, Sp), for which annotated chromosome sequences were downloaded from the relevant database and exons extracted subsequently. Exons with identical locus IDs were then sorted into individual files, only retaining files with at least three exons in total (i.e. at least one internal exon). All locus files were subsequently checked to ensure coding sequence started with ATG, finished with a stop codon (TAA, TAG, TGA), had no internal stop codons, and was a multiple of three nucleotides. Locus files where one of the above prerequisites was violated were removed from the final data set. We also eliminated exons containing one or more ambiguous nucleotides ("n"). The remaining exons were trimmed so that the first nucleotide was the first nucleotide of the first complete codon and the last nucleotide the last of the final complete codon. Then, we discarded all terminal exons to obtain the final exon sets, the species-specific sizes of which are provided in Supplementary Table 4.

After splitting individual exons in half to ensure that no codon featured in both 5' and 3' analyses, we considered the trend in usage of each amino acid as a function of the distance from the boundary up to a maximum distance of 34 codons. Importantly, the codon in direct proximity to the boundary was also eliminated.

We then calculated Spearman rank correlations (rho) between the distance from the boundary (5' or 3') and proportional usage of the amino acid (i.e. in proportion to the number of residues at that given distance) for the remaining 33 data points for each species. The three 6-fold degenerate amino acids we split into a block of 4 and a block of 2. The block of 2 is denoted by the use of an

asterisk (i.e. "S" signifies, TCA, TCC, TCG and TCT, while "S*" signifies AGC and AGT). In relevant circumstances, the two-fold and four fold blocks were treated as separate amino acids, yielding a total of 23 amino acid groupings.

For each amino acid grouping independently we fitted unweighted linear models and extracted the slope of the regression line to be used as a basic measure of the strength of individual preference trends. Note that a negative rho/slope implies an amino acid that is preferred near boundaries and a positive rho/slope implies a tendency to be avoided. Unless otherwise stated, results are reported as significant only if they remain significant after correction for multiple testing (see Results for adjusted P values).

For the most part, trends are approximately monotonic and linear and hence adequately captured by simple linear models. For certain amino acids departures from linearity, some recurrent across species and typically highly localized, do exist however. Unusual U-shaped 5' trends for proline, originally noted for human and mouse by Parmley et al. (2007), are also present in other species (Ce, Dr). Further, some amino acids, notably isoleucine and the 2-fold block of leucine, are disproportionately preferred in direct proximity to the boundary (after trimming) at 3' exon ends in several species. "Popping out" from otherwise linear trends (see Supplementary Figure 2), these patterns are perhaps caused by participation of the relevant codons in an extended splice site consensus relevant for U5 snRNA-mediated exon joining (see Supplementary Document 2 for a more detailed discussion of recurrent, locally confined preference/avoidance patterns and potential functional explanations). As a corollary of discontinuities more generally, comparative interpretation of slope coefficients as an index of relative strength ought to be done with care. In particular, our rank ordering of slopes derives its value from providing another dimension through which congruence in preference spectra can be asserted, rather than being easily translated into differential functional impact on a mechanistic level.

Modifications in the analysis of Saccharomyces cerevisiae exons

Given the small number of internal exons in *S. cerevisiae* (only 8 genes have more than one intron), we decided to include terminal exons in the final data set (417 exons) for this species. The one end of each terminal exon that did not border the intron was excluded from the final data set. Otherwise, the removal of irregularities (internal stop codons etc.) proceeded as described above. Restricted sample size also indirectly prompted a re-examination of the results obtained from Spearman's rank correlations because the presence of multiple tied ranks led to concerns about the adequacy of this statistic. However, using the more appropriate Kendall's tau statistic did not return any qualitatively different results.

Cross-species patterns in preference across all amino acid groupings

For 5' and 3' data sets independently, non-parametric (Spearman's) correlations were computed between the previously derived slope coefficients of all 23 amino acid groupings for every possible metazoan species pair. 72 tests (with the number of species N=9, N^2-N=72) were carried out and significance threshold adjusted accordingly (P=0.05/72=6.94E-04). We initially included both yeast species in the analysis but, as expected from the absence of significant individual amino acid trends, we found no significant correlations for the global amino acid set (data not shown). No loss of relevant information is incurred whilst clarity of presentation is enhanced when these species are excluded from the analysis and, in particular, the accompanying table (Table 2).

Comparison of orthologous exons

S. cerevisiae – S. castelli

A set of *S. cerevisiae-S. castelli* orthologous genes, based on a re-annotation of the *S. castelli* genome by Wolfe and colleagues, were obtained from the Yeast Gene Order Browser (http://wolfe.gen.tcd.ie/ygob/). For each *S. cerevisiae* gene that contributed exons to our analysis of amino acid abundance, we checked

whether a homologous *S. castelli* gene was present on the same positional track, the rationale being to compare true orthologues rather than outparalogues. If putatively orthologous gene pairs were found on both tracks, implying the retention of two post-genome duplication paralogues in both species, only the pair on track 1 was considered. This procedure yielded 164 orthologue pairs. *S. castelli* ORF structure downloaded from the same source was used to eliminate all *S. castelli* genes that lacked any introns, did not have a regular start or stop codon, or whose exon sequence was not a multiple of three nucleotides. Further discarding all genes with unequal exon number or unequal intron phase between species 51 gene pairs (102 exons) remained. We further eliminated all exons shorter than 8 amino acids in length as these were considered uninformative. After exons were trimmed so that each exon only contained full codons, codons were translated into amino acids and orthologous exons aligned using MUSCLE (version 3.6) [39]. After alignment, the first and last amino acid of each exon was removed. Exons were then split in half so that any one amino features exclusively in either 5' or 3' analysis. We then calculated the number of amino acid changes over the total number of informative (amino acid present in both species) sites for each amino acid position from the boundary, including only exon ends that bordered an intron (i.e. only the 3' end for the first exon and only the 5' end for the last exon)

Considering 5' and 3' ends separately, Spearman's and Kendall's test of rank correlation between distance from the boundary and the proportion of amino acid changed were computed in R; the latter method, used in response to an appreciable proportion of tied ranks in the data, supports entirely the conclusions of the former. Given the small sample sizes for end-specific analyses (N(5')=51, N(3')=39), we also computed rank correlations for 5' and 3' ends pooled. Linear models were fitted for each analysis, weighting by the number of informative sites at distance x from the boundary.

D. melanogaster – D. pseudoobscura

A list of *D. melanogaster-D. pseudoobscura* orthologous genes was obtained from the Inparanoid database

(http://inparanoid.sbc.su.se/download/current/sqltables/sqltable.flyDROPS.fa-modDROME.fa). *D. pseudoobscura* exons were downloaded from the flybaseGene track on the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgTables) and sorted into files by gene locus, eliminating genes with irregularities as described above. Using the orthologue list we established a set of 4165 orthologue pairs for which genes were present in the cleaned data sets of both species. 2677 gene pairs (comprising 7545 orthologous internal exon pairs, and 5352 orthologous terminal exon pairs) remain after checking for equal exon number and intron phase. Trimming of exons, alignment and statistical analysis were carried out as described for *S. cerevisiae-S. castelli*. 3' and 5' end were considered for each internal exon, whereas only exon ends bordering an intron were included in the analysis of terminal exons.

C. elegans – C. briggsae

For each *C. elegans* locus file concatenated exons (i.e. before trimming and with the ORF intact) were translated into protein and thereafter used to query a database of all translated *C. briggsae* locus files using BLAST (blastp), and vice versa. Only reciprocal best hits with an eigenvalue E≤1 were retained. After checking for equal exon number and intron phase, 5359 orthologous gene pairs (19347 orthologous internal exon pairs) remained. Trimming and alignment were carried out as described above for Drosophila.

Intraspecific 5'~3' correlations and symmetry analysis

Covering all 23 amino acid groupings Spearman's rank correlations were computed between 5' and 3' trends within each species (N=11, P=0.05/11=4.54E-03).

Standardized major axis regressions (SMAs) were computed in R using the SMATR package [40, 41] applying standard confidence limits (95% CI). As symmetry of the type x=y was to be tested, the regression line was forced through the origin. SMA requires estimates of the slope of the regression line to have a consistently positive or negative sign so that major and minor axis can be identified unambiguously. This is not the case for *A. thaliana*, which is hence

not amenable to this type of analysis and was not included. Further, residual distribution for *S. cerevisiae* shows significant deviation from normality so that results for this species should be interpreted with care.

List of abbreviations

SR proteins (Serine-arginine proteins); ESE (exonic splicing enhancer); SMA (standard major axis); HPI (hexamer preference index); bp (base pairs); RS domains (Arginine-Serine rich domains); SL (spliced leader)

Acknowledgements

We would like to thank Max Robinson (University of Washington) for kindly providing us with results from his PhD thesis. This work was funded by the Medical Research Council (T.W.) and the Biotechnology and Biological Sciences Research Council (J.L.P.).

Fig. 1

Nature and diversity of amino acid abundance trends near exon-intron boundaries. Relative abundance of glutamine (Q), methionine (M), and lysine (K) as a function of distance from the boundary across 5' ends of *D. melanogaster* exons is shown. Glutamine is significantly *avoided* near the boundary (rho=0.86, P<1.84E-7), lysine is *preferred* (rho=-0.65, P<6.2E-5), whilst no significant trend is evident for methionine (rho=0.096, P=0.59). Note that a negative slope/rho value indicates a *preference* near the exon-intron boundary. Typically, where patterns of preference/avoidance are evident, we observe quasi-monotonic decreases/increases in relative abundance across the sequence range analyzed.

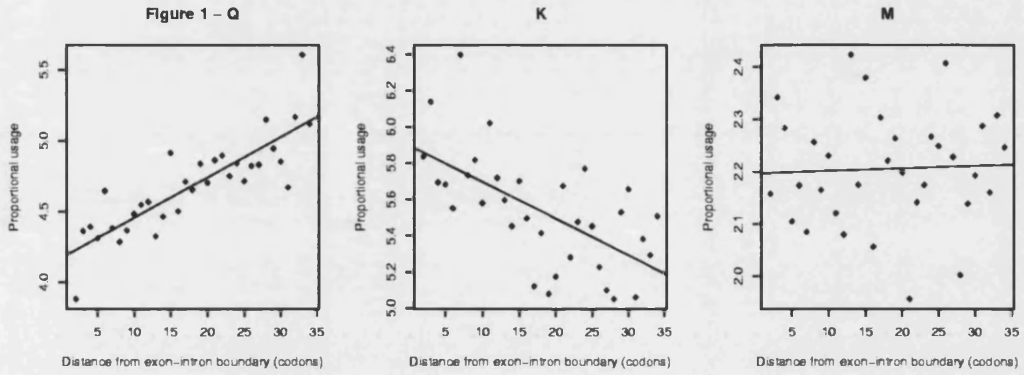


Fig 2.

Relative amino acid abundance of lysine (K) at 5' ends of exons in six species.
Proportional usage of lysine vis-à-vis all other amino acids is plotted against distance from the exon-intron boundary measured in amino acids. Variable degrees of preference for lysine near the boundary are evident for non-nematode species (Am: rho=-0.67, P=2.71E-05, β(slope)=-0.017; Dr: rho=-0.79, P=6.51E-07, β=-0.035; Dm: rho=-0.65, P=6.11E-05, β=-0.020; Hs: rho=-0.90, P=3.67E-09, β=-0.041) whereas nematodes show strong avoidance trends (Ce: rho=0.89, P=5.26E-08, β=0.030; Cb: rho=0.92, P=0, β=0.033)

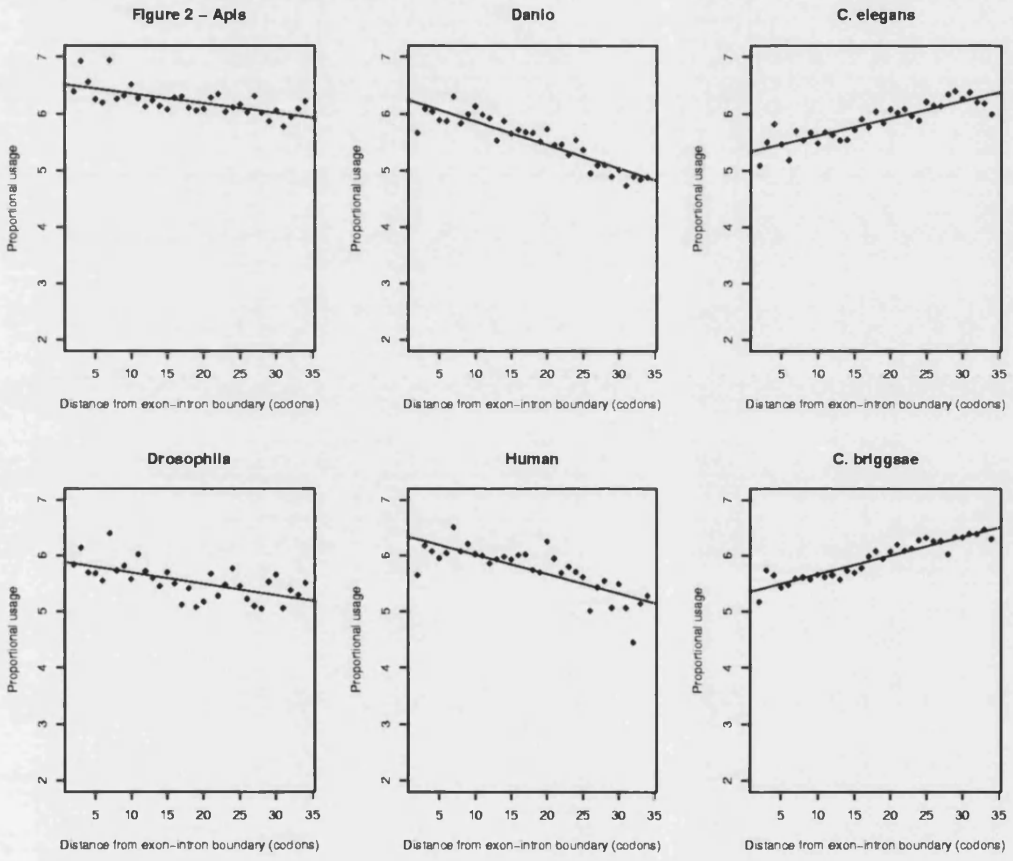


Fig 3.

Variable symmetry in amino acid abundance trends comparing 5' and 3' exon ends within species.

Intraspecific correlations between the 5' and 3' slopes as extracted from individually fitted linear models considering all 23 amino acid groupings are shown. Approximately symmetric arrangements are particularly evident for some species (notably vertebrates) whereas nematode species (Ce, Cb) show a clearly asymmetric arrangement. Further notable is the higher variability of slope coefficients in some species (vertebrates and nematodes) vis-à-vis others (Am, At). Amino acids are represented by their one letter code (two-fold blocks are denoted by an asterisk). The regression lines are from standardized major axis regressions. Lines were not fitted for *Arabidopsis* and *S. cerevisiae* given concerns about the adequacy of this technique for these datasets (see Materials and Methods). For associated statistics consult Table 3.

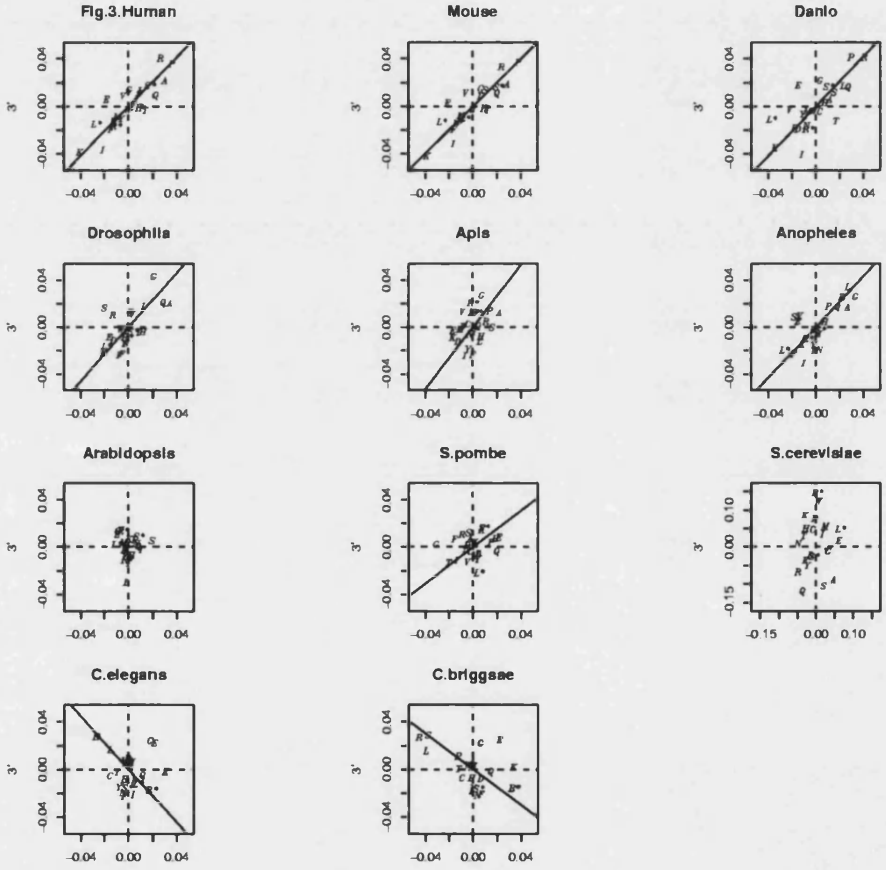


Fig 4.

Frequency of nonsynonymous change as a function of distance from the exon-intron boundary. Amino acids are significantly more likely to be conserved near the exon-intron boundary comparing (A) *C. elegans – C. briggsae* (5': rho=0.957, P=0; 3': rho=0.96. P=0; N=19347 exons) and (B) *D. melanogaster – D. pseudoobscura* (5': rho=0.87, P=1.02E-07; 3': rho=0.95, P=0; N=7545 exons). The trends appear approximately monotonous and linear. Location-dependent conservation levels also appear slightly higher near the boundary comparing (C) *S. cerevisiae – S. castelli* but this is not significant (5': rho=0.11, P=0.55, N=51; 3':rho=0.11, P=0.55, N=39; pooled 3'/5': rho=0.12, P=0.51, N=90) or of comparable monotony (but see Suppl. Fig 1.).

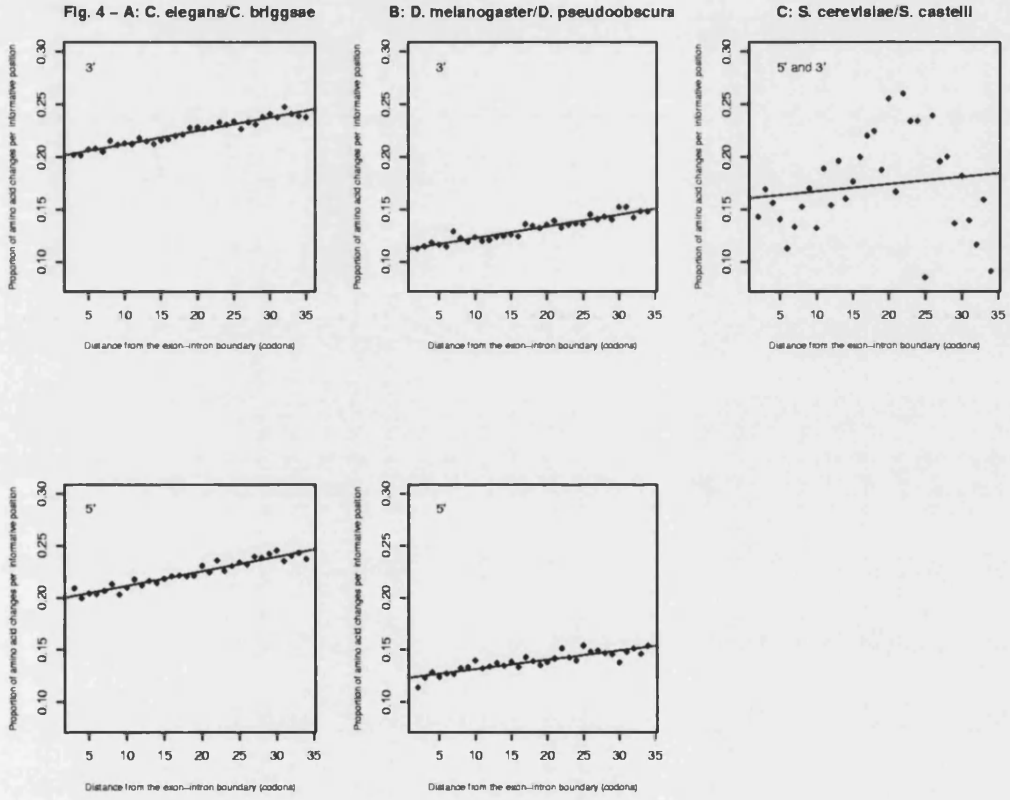


Fig 5.

The rate of nonsynonymous evolution correlates negatively with the proportion of boundary-proximal sequence.

K_A is plotted as a function of the proportion of coding sequence located within 70bp of an exon-intron boundary for (A) *D. melanogaster-D. pseudoobscura* orthologous genes (rho=-0.26, P=2.2E-16, N=4132) and (B) *C. elegans – C. briggsae* orthologous genes (rho=-0.08, P=6.18E-09, N=5248).

The data has been divided into bins along regular decimal intervals (0.1, 0.2, etc) and the mean K_A within each bin plotted against the mean proportion of sequence near the boundary. The last (A) and first (B) three bins, respectively, have been pooled to obtain approximately equal bin sizes. Negative trends are present for both sets of aligned genes, but a departure from the general trend is evident for nematode genes with a low proportion of boundary-proximal sequence.

Fig. 5 A: D. melanogaster/D. pseudoobscura

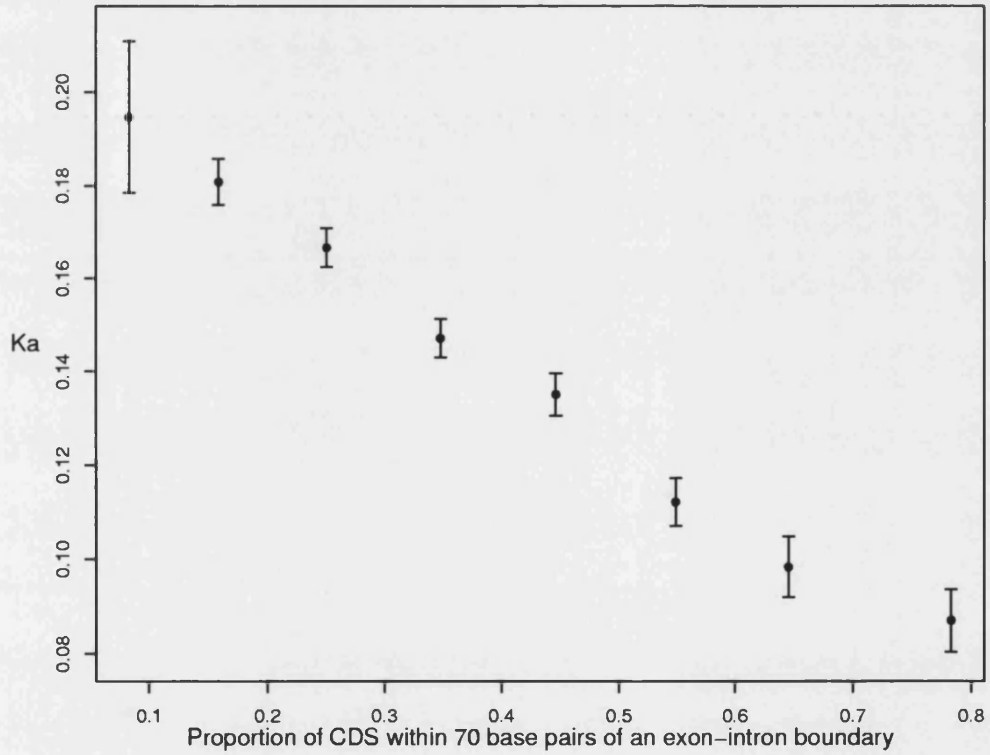
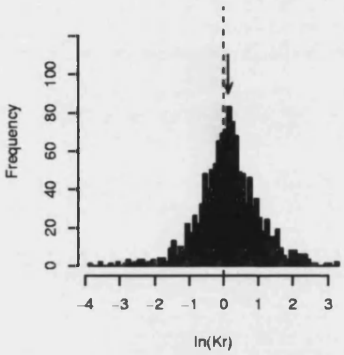


Fig. 6

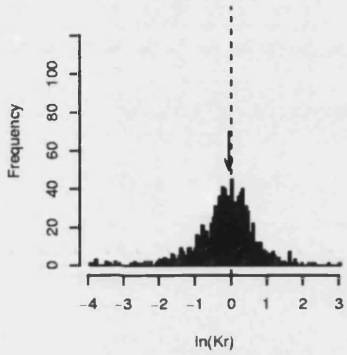
Exon cores and flanks evolve at different rates.

Histograms of logged *Kr* ratios (K_{Ac}/K_{Af}), using 100 bins, for (A) *D. melanogaster-D. pseudoobscura* orthologous genes (N=1237), (B) *C. elegans-C. briggsae* orthologous genes (N=1102), and (C) human-mouse orthologous genes (N=360) with a minimum of 300bp of concatenated middle and flanking sequence of exons are plotted. The dashed line in each graph indicates ln(*Kr*)=0, the point at which middle and flanking sections evolve at the same average rate. The arrows indicate the median logged *Kr* ratios of (A) =0.128, (B) = -0.065, and (C) =0.559 respectively. All three are significantly different from the null expectation of ln(*Kr*)=0 ($P<0.0001$). Note the much more marked departure from the null expectation in the mammalian data set.

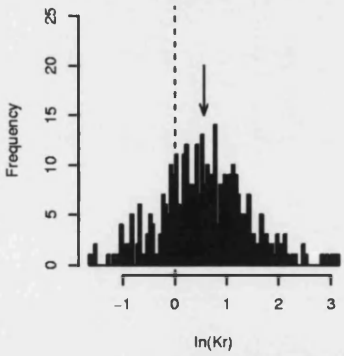
Fig. 6 A: D. melanogaster/D. pseudoobscura



B: C. elegans/C. briggsae



C: Human/Mouse



Tables

Table 1. a) Amino acids significantly preferred (-) or avoided (+) at 3' ends of exons across species

Amino acids[a, b]																							
A	C	D	E	F	G	H	I	K	L	L*	M	N	P	Q	R	R*	S	S*	T	V	W	Y	Species (number of exons)[c]
$+_3$		$-_7$		$-_3$			$-_2$	$-_1$		$-_5$		$-_6$	$+_2$		$+_1$	$-_4$		$+_4$					**Human** (178438)
$+_3$		$-_6$		$-_3$			$-_2$	$-_1$		$-_5$		$-_4$	$+_1$		$+_2$	$-_7$	$+_5$	$+_4$					**Mouse** (126268)
		$-_4$		$-_5$	$+_3$		$-_1$	$-_2$				$-_6$	$+_2$		$+_1$	$-_3$				$+_4$			D. rerio *(41264)*
			$+_4$	$-_1$	$+_3$	$-_6$	$-_2$		$+_5$			$-_3$			$+_1$	$-_4$	$+_2$	$-_5$					***C. elegans*** (79958)
		$-_6$	$+_3$	$-_2$	$+_4$	$-_8$	$-_3$		$+_5$	$-_5$		$-_1$	$+_6$		$+_2$	$-_7$	$+_1$	$-_4$					***C. briggsae*** (74178)
							$-_1$			$-_3$		$-_2$	$+_2$		$+_1$	$-_4$							***A. gambiae*** (7930)
				$-_2$	$+_1$		$-_1$			$-_3$				$+_2$									***D. melanogaster*** (48933)
				$-_2$	$+_1$	$-_5$	$-_1$		$-_4$	$+_5$			$+_3$		$+_2$			$+_6$		$+_4$		$-_3$	***A. mellifera*** (45426)
					$+_2$		$-_2$		$-_1$						$-_3$	$+_3$				$+_1$			***A. thaliana*** (109900)
																							S. pombe (2403)
										$-_1$													***S. cerevisiae*** (417)

[a] Indices signify rank order of slope coefficients, separately for negative and positive trends

[b] L*, R*, S* signify the two-fold degenerate blocks of leucine, arginine, and serine, respectively

[c] for *S. cerevisiae* terminal exons were retained given the small number of genes with more than one intron (8)

Table 1. b) Amino acids significantly preferred (-) or avoided (+) at 5' ends of exons across species

Amino acids[a, b]																							Species (number of exons)[c]
A	C	D	E	F	G	H	I	K	L	L*	M	N	P	Q	R	R*	S	S*	T	V	W	Y	
$+_2$			$-_4$	$-_5$		$+_7$	$-_3$	$-_1$		$-_2$	$-_8$	$-_6$	$+_1$	$+_4$	$+_3$	$-_7$		$+_5$	$+_6$				**Human** (178438)
$+_2$			$-_4$	$-_5$		$+_7$	$-_3$	$-_1$		$-_2$	$-_7$	$-_6$	$+_1$	$+_4$	$+_3$			$+_5$	$+_6$				**Mouse** (126268)
								$-_2$		$-_1$			$+_2$	$+_3$	$+_1$			$+_5$	$+_4$	$-_3$			D. rerio ***(41264)***
	$-_3$		$+_2$		$+_4$			$+_1$						$+_5$	$-_1$	$+_3$	$-_2$		$-_4$			$-_5$	***C. elegans*** (79958)
	$-_5$		$+_4$					$+_3$	$-_2$	$+_2$				$+_5$	$-_1$	$+_1$	$-_3$		$-_4$		$-_6$		***C. briggsae*** (74178)
										$-_1$													***A. gambiae*** (7930)
$+_1$						$+_3$		$-_1$		$-_3$				$+_2$			$-_2$				$-_4$		***D. melanogaster*** (48933)
$+_1$		$-_3$	$-_2$			$+_4$		$-_1$			$-_4$				$+_3$		$+_2$			$-_5$	$-_6$		***A. mellifera*** (45426)
																	$+_1$	$+_3$	$+_2$				***A. thaliana*** (109900)
																							S. pombe (2403)
																							S. cerevisiae (417)

[a] Indices signify rank order of slope coefficients, separately for negative and positive trends

[b] L*, R*, S* signify the two-fold degenerate blocks of leucine, arginine, and serine, respectively

[c] for *S. cerevisiae* terminal exons were retained given the small number of genes with more than one intron (8)

Table 2. Cross-species correlations of preference slope coefficients considering all 23 amino acid groupings, 5' (bottom-left) and 3' (top-right)[a,b]

	Hs	Mm	Dr	Ce	Cb	Ag	Dm	Am	At
Hs	1	0.9852^{++}	0.9308^{++}	0.7065^{++}	0.6749^{++}	0.8834^{++}	0.8439^{++}	0.5316^{+}	0.1126
Mm	0.9852^{++}	1	0.9160^{++}	0.6917^{++}	0.6729^{++}	0.8824^{++}	0.8498^{++}	0.6047^{+}	0.1957
Dr	0.9167^{++}	0.9042^{++}	1	0.7441^{++}	0.7075^{++}	0.8706^{++}	0.7678^{++}	0.4792^{+}	0.1640
Ce	**-0.4338**$^{+}$	**-0.3864**	**-0.4051**	1	0.9832^{++}	0.8370^{++}	0.7204^{++}	0.3725	0.2391
Cb	**-0.5998**$^{+}$	**-0.5634**	**-0.6463**$^{+}$	0.7767^{++}	1	0.8221^{++}	0.7065^{++}	0.3449	0.2125
Ag	0.6235^{+}	0.5958^{+}	0.6146^{+}	0	**-0.2589**	1	0.8933^{++}	0.4951^{+}	0.1769
Dm	0.6423^{+}	0.6117^{+}	0.5148^{+}	**-0.0425**	**-0.1403**	0.6354^{+}	1	0.5702^{+}	0.2085
Am	0.7589^{++}	0.7895^{++}	0.7658^{++}	**-0.3192**	**-0.4140**	0.4812^{+}	0.4605^{+}	1	0.6581^{+}
At	0.4407^{+}	0.4427^{+}	0.4960^{+}	**-0.3646**	**-0.3636**	0.0593	0.1887	0.3972	1

[a] *S. pombe* and *S. cerevisiae* omitted for clarity given the absence of significant correlations.

[b] negative correlations in bold

+ significant at P=0.05

++ significant at P=0.05/72=6.94E-04 (N=72 tests)

Table 3. Intraspecific 5'~3' correlations of preference slopes for all 23 amino acid groupings

			SMA		
	Rho	**P value**[a]	**Slope (β)**	**Lower CI**[b]	**Upper CI**[b]
Human	0.8528	1.96E-06	1.0362	0.8312	1.2918
Mouse	0.8626	2.28E-06	0.9907	0.7965	1.2322
D. rerio	0.6591	8.3E-04	1.0439	0.7796	1.3979
C. elegans	-0.1413	0.5185	-1.1118	-0.7309	-1.6913
C. briggsae	-0.4358	0.0388	-0.7474	-0.5139	-1.0869
A. gambiae	0.5702	5.16E-03	1.0788	0.7886	1.4757
D. melanogaster	0.6087	2.49E-03	1.1519	0.8207	1.6167
A. mellifera	0.3943	0.0633	1.3173	0.8844	1.9622
A. thaliana	-0.2233	0.3042	NA[c]	NA[c]	NA[c]
S. pombe	0.2213	0.3075	0.7689	0.5042	1.1726
S. cerevisiae	0.1611	0.4597	2.417[d]	1.5774	3.7035

[a]with 11 species significance is indicated by P=0.05/11=4.54E-03

[b]CI=0.95, the regression line was forced through the origin

[c]see Materials and Methods

[d]Adequacy of SMA regression analysis is seriously in doubt for S. cerevisiae because normal distribution of residuals is strongly violated

Additional Material

Table of amino acid trends for all species and associated statistics (Supplementary Table 1 in Warnecke_SupplTabl1.xls)

Amino acid trends and associated statistics for homology-reduced gene sets of *D. melanogaster* and *C. elegans* (Supplementary Table 2 in Warnecke_SupplTabl2.xls)

Rank correlation between K_A and proportion of sequence near the exon-intron boundary (Supplementary Table 3)

Sources of exon datasets (Supplementary Table 4)

Patterns of discontinuous preference in direct proximity to the exon-intron boundary (Supplementary Table 5)

Sampling distributions (Supplementary Figure 1)

Examples of locally discontinuous preference (Supplementary Figure 2)

Methods for homology reduction (Supplementary Document 1)

Examination of local discontinuities (Supplementary Document 2)

Supplementary Tables 3-5, Supplementary Figures 1 and 2, and Supplementary Documents 1-2 can be found in the file Warnecke_SupplMaterial.doc

References

1. Akashi H, Gojobori T: Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc Natl Acad Sci U S A* 2002, 99(6):3695-3700.
2. Clay O, Caccio S, Zoubak S, Mouchiroud D, Bernardi G: Human coding and noncoding DNA: compositional correlations. *Mol Phylogenet Evol* 1996, 5:2-12.
3. Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD: Splicing and the evolution of proteins in mammals. *PLoS Biol* 2007, 5(2):e14.
4. Blencowe BJ: Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* 2000, 25(3):106-110.
5. Zheng: Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *J Biomed Sci* 2004, 11(4):538-538.
6. Ram O, Ast G: SR proteins: a foot on the exon before the transition from intron to exon definition. *Trends Genet* 2007, 23(1):5-7.
7. Berget SM: Exon Recognition in Vertebrate Splicing. *J Biol Chem* 1995, 270(6):2411-2414.
8. Hertel KJ, Maniatis T: The function of multisite splicing enhancers. *Mol Cell* 1998, 1(3):449-455.
9. Graveley BR, Hertel KJ, Maniatis T: A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. *Embo Journal* 1998, 17(22):6747-6756.
10. Willie E, Majewski J: Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet* 2004, 20(11):534-538.
11. Chamary JV, Hurst LD: Biased codon usage near exon-intron junctions: selection on splicing enhancers, splice-site recognition or something else? *Trends Genet* 2005, 21(5):256-259.
12. Barbosa-Morais NL, Carmo-Fonseca M, Aparicio S: Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res* 2006, 16(1):66-77.
13. Graveley BR: Sorting out the complexity of SR protein functions. *RNA-Publ RNA Soc* 2000, 6(9):1197-1211.

14. Robinson RM: Splicing Signals in Caenorhabditis elegans: Candidate Exonic Splicing Enhancer Motifs. *PhD thesis*. Washington: University of Washington; 2005.
15. Parmley JL, Chamary JV, Hurst LD: Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular Biology and Evolution* 2006, 23(2):301-309.
16. Yeo G, Hoon S, Venkatesh B, Burge CB: Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A* 2004, 101(44):15700-15705.
17. Eskesen ST, Eskesen FN, Ruvinsky A: Natural selection affects frequencies of AG and GT dinucleotides at the 5 ' and 3 ' ends of exons. *Genetics* 2004, 167(1):543-550.
18. Whamond GS, Thornton JM: An analysis of intron positions in relation to nucleotides, amino acids, and protein secondary structure. *J Mol Biol* 2006, 359(1):238-247.
19. Blencowe BJ: Alternative splicing: New insights from global analyses. *Cell* 2006, 126(1):37-47.
20. Fairbrother WG, Yeh RF, Sharp PA, Burge CB: Predictive identification of exonic splicing enhancers in human genes. *Science* 2002, 297(5583):1007-1013.
21. Wang ZF, Xiao XS, Van Nostrand E, Burge CB: General and specific functions of exonic splicing silencers in splicing control. *Mol Cell* 2006, 23(1):61-70.
22. Siebel CW, Feng LN, Guthrie C, Fu XD: Conservation in budding yeast of a kinase specific for SR splicing factors. *Proc Natl Acad Sci U S A* 1999, 96(10):5440-5445.
23. Ares M, Grate L, Pauling MH: A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA-Publ RNA Soc* 1999, 5(9):1138-1139.
24. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S *et al*: The genome sequence of Schizosaccharomyces pombe. *Nature* 2002, 415(6874):871-880.
25. Gross T, Richert K, Mierke C, Lutzelberger M, Kaufer NF: Identification and characterization of srp1, a gene of fission yeast encoding a RNA binding domain and a RS domain typical of SR splicing factors. *Nucl Acids Res* 1998, 26(2):505-511.

26. Lutzelberger M, Gross T, Kaufer NF: Srp2, an SR protein family member of fission yeast: in vivo characterization of its modular domains. *Nucl Acids Res* 1999, 27(13):2618-2626.
27. Kuhn AN, Kaufer NF: Pre-mRNA splicing in Schizosaccharomyces pombe: regulatory role of a kinase conserved from fission yeast to mammals. *Curr Genet* 2003, 42(5):241-251.
28. Webb CJ, Romfo CM, van Heeckeren WJ, Wise JA: Exonic splicing enhancers in fission yeast: functional conservation demonstrates an early evolutionary origin. *Genes Dev* 2005, 19(2):242-254.
29. Davis CA, Grate L, Spingola M, Ares M, Jr.: Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res* 2000, 28(8):1700-1706.
30. Okazaki K, Niwa O: mRNAs encoding zinc finger protein isoforms are expressed by alternative splicing of an in-frame intron in fission yeast. *DNA Res* 2000, 7(1):27-30.
31. Ast G: How did alternative splicing evolve? *Nat Rev Genet* 2004, 5(10):773-782.
32. Xing Y, Lee C: Alternative splicing and RNA selection pressure - evolutionary consequences for eukaryotic genomes. *Nat Rev Genet* 2006, 7(7):499-509.
33. Sanford JR, Bruzik JP: SR proteins are required for nematode trans-splicing in vitro. *RNA* 1999, 5(7):918-928.
34. Longman D, Johnstone IL, Caceres JF: Functional characterization of SR and SR-related genes in Caenorhabditis elegans. *Embo J* 2000, 19(7):1625-1637.
35. Blumenthal T: Trans-splicing and operons. In. Edited by Community" T*Ce*R: WormBook; 2005.
36. Hastings KE: SL trans-splicing: easy come or easy go? *Trends Genet* 2005, 21(4):240-247.
37. Furuyama S, Bruzik JP: Multiple roles for SR proteins in trans splicing. *Mol Cell Biol* 2002, 22(15):5337-5346.
38. Huang T, Kuersten S, Deshpande AM, Spieth J, MacMorris M, Blumenthal T: Intercistronic region required for polycistronic pre-mRNA processing in Caenorhabditis elegans. *Mol Cell Biol* 2001, 21(4):1111-1120.
39. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004, 32(5):1792-1797.
40. Warton DI, Wright IJ, Falster DS, Westoby M: Bivariate line-fitting methods for allometry. *Biol Rev Camb Philos Soc* 2006, 81(2):259-291.

41. **Warton DI, Weber NC: Connnon slope tests for bivariate errors-in-variables models. *Biom J* 2002, 44(2):161-174.**