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.

**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

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.

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[9], 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. 1)). 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 $s \approx 1/(2N_e)$), 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.

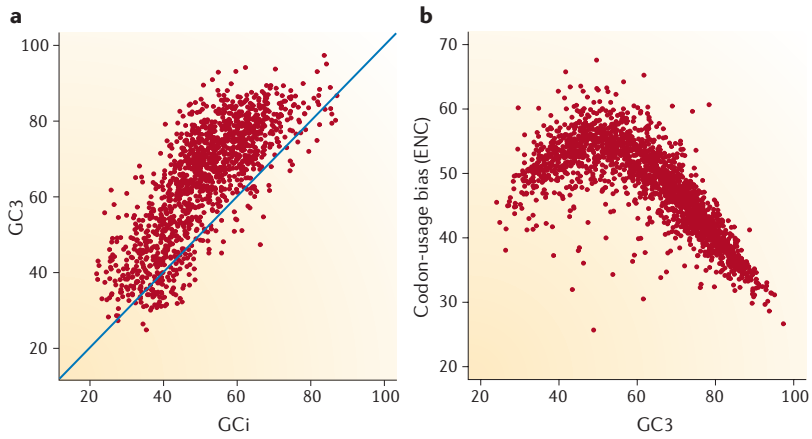

Figure 1 | **The effect of isochores on synonymous codon usage and codon-usage bias.** Mammalian genomes consist of relatively homogenous domains of GC content (>300 kb in size)[25,26]. Within these isochores, base composition of intergenic spacers, introns and coding sequence are all highly correlated. 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[53], 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[103] 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

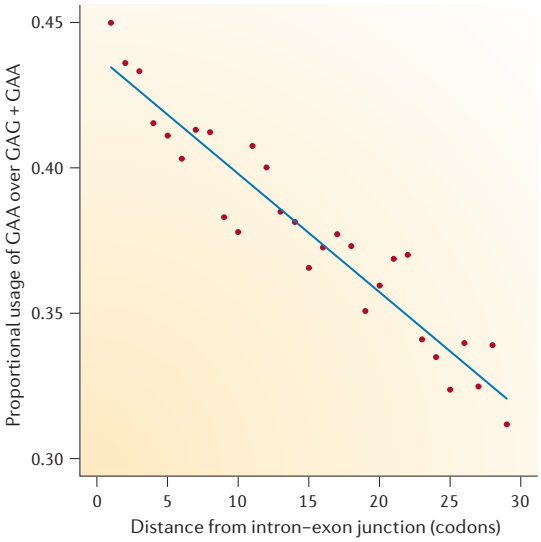

Figure 2 | **Usage of certain codons is more biased near intron–exon junctions, owing to synonyms being differentially common in exonic splicing enhancers.** The example depicted shows the proportional usage of the codon GAA versus its synonym GAG, as a function of the distance from intron–exon junctions[95]. 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_a 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.

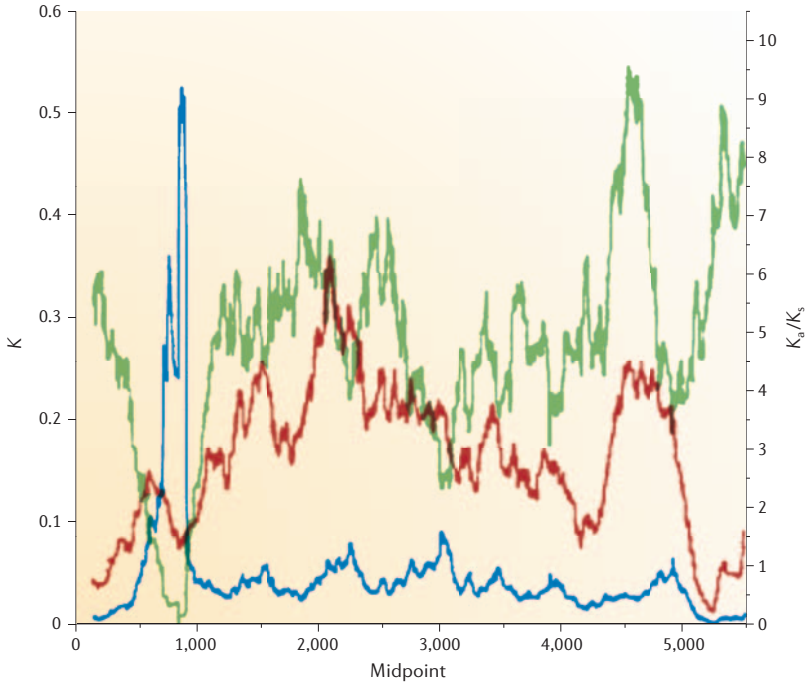

Figure 3 | **Fluctuation in rates of evolution across the *BRCA1* gene.** The sliding-window plot compares sequences between human and dog orthologues. The *x*-axis shows the midpoint in base pairs of the 306-nucleotide window. The *y*-axis shows, on the left, the rate of non-synonymous substitution (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 $K_a/K_s > 1$ 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_{M2} 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. Sēmon, 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.