Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 240 (2006) 616–626
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites

Fyodor A. Kondrashov[a], Aleksey Y. Ogurtsov[b], Alexey S. Kondrashov[b,*]

[a]*Section of Ecology, Behavior and Evolution, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0346, USA*
[b]*National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA*

Received 20 June 2005; received in revised form 26 October 2005; accepted 27 October 2005
Available online 15 December 2005

Abstract

The impact of synonymous nucleotide substitutions on fitness in mammals remains controversial. Despite some indications of selective constraint, synonymous sites are often assumed to be neutral, and the rate of their evolution is used as a proxy for mutation rate. We subdivide all sites into four classes in terms of the mutable CpG context, nonCpG, postC, preG, and postCpreG, and compare four-fold synonymous sites and intron sites residing outside transposable elements. The distribution of the rate of evolution across all synonymous sites is trimodal. Rate of evolution at nonCpG synonymous sites, not preceded by C and not followed by G, is ~10% below that at such intron sites. In contrast, rate of evolution at postCpreG synonymous sites is ~30% above that at such intron sites. Finally, synonymous and intron postC and preG sites evolve at similar rates. The relationship between the levels of polymorphism at the corresponding synonymous and intron sites is very similar to that between their rates of evolution. Within every class, synonymous sites are occupied by G or C much more often than intron sites, whose nucleotide composition is consistent with neutral mutation–drift equilibrium. These patterns suggest that synonymous sites are under weak selection in favor of G and C, with the average coefficient $s \sim 0.25/N_e \sim 10^{-5}$, where N_e is the effective population size. Such selection decelerates evolution and reduces variability at sites with symmetric mutation, but has the opposite effects at sites where the favored nucleotides are more mutable. The amino-acid composition of proteins dictates that many synonymous sites are CpGprone, which causes them, on average, to evolve faster and to be more polymorphic than intron sites. An average genotype carries $\sim 10^7$ suboptimal nucleotides at synonymous sites, implying synergistic epistasis in selection against them.
Published by Elsevier Ltd.

Keywords: Mutation; Selection; Synonymous site; Evolution; Genetic drift

1. Introduction

Throughout all life, synonymous codons are used non-randomly (Grantham et al., 1980; see Li, 1997, Chapter 7, for review). There is a general agreement that selection plays a major role in this phenomenon (Andersson and Kurland, 1990; McVean and Vieira, 2001; Duret, 2002; Carlini and Stephan, 2003; Nielsen and Akashi, 2003). Synonymous substitutions affect mRNA translation (Ikemura, 1985; Sorensen et al., 1989; Sharp et al., 1995; Akashi, 1995, 1999a, b, 2003) and thus can cause translational selection which influences codon usage in many, although perhaps not in all (Kanaya et al., 1999), organisms. Synonymous substitutions also affect important properties of mRNAs which are "not directly related to the codon–anticodon interaction" (Duan and Antezana, 2003), in particular, their secondary structures (Hartl et al., 1994; Innan and Stephan, 2001; Duan et al., 2003; Katz and Burge, 2003; Chamary and Hurst, 2005a).

However, the importance of selection at synonymous sites in mammals remains unclear. Although their codon usage is obviously non-random, due to elevated frequencies of G and C at synonymous sites (Debry and Marzluff, 1994; Eyre-Walker, 1999), the causes of this pattern are controversial. Some authors argue for an important role of selection (Debry and Marzluff, 1994; Eyre-Walker, 1999; Keightley and Gaffney, 2003; Urrutia and Hurst, 2003;

*Corresponding author. Tel.: 1 301 435 8944; fax: 1 301 480 2288.
E-mail addresses: fkondras@ncbi.nlm.nih.gov (F.A. Kondrashov), ogurtsov@ncbi.nlm.nih.gov (A.Y. Ogurtsov), kondrashov@ncbi.nlm.nih.gov (A.S. Kondrashov).

0022-5193/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.jtbi.2005.10.020

Nielsen and Akashi, 2003; Chamary and Hurst, 2004; Comeron, 2004; Lu and Wu, 2005; Chamary and Hurst, 2005a), but others disagree (e. g., Sharp et al., 1995; Smith and Hurst, 1999; Duret and Hurst, 2001; Urrutia and Hurst, 2001; Duret, 2002; Subramanian and Kumar, 2003) and favor alternative explanations, such as biased mutation (Wolfe et al., 1989) or biased gene conversion (Duret, 2002).

The arguments for or against selection at synonymous sites in mammals are undermined by conflicting data on whether evolution at four-fold synonymous sites is slower than at presumably neutral intron or pseudogene sites, which is often thought to be the obligatory signature of any selection at synonymous sites. Hughes and Yager (1997) and Chamary and Hurst (2004) reported similar levels of rat–mouse divergence at synonymous and intron sites, Bustamante et al. (2002) found that synonymous sites evolve more slowly than homologous pseudogene sites, Subramanian and Kumar (2003) reported that in primates synonymous sites evolve faster than intron sites, and Hellman et al. (2003) reached the opposite conclusion.

Because effective neutrality of synonymous substitutions in mammals is widely accepted, mutation rates are routinely estimated through rates of synonymous substitution in mammalian evolution (Smith and Hurst, 1999; Keightley and Eyre-Walker, 2000; Kumar and Subramanian, 2002) and patterns in synonymous substitutions are generalized to the whole genome (Duret et al., 2002). Similarly, levels of intrapopulation variability and rates of interspecies divergence at synonymous sites have been accepted as the neutral point of reference in tests for positive selection (e.g. Fay et al., 2001).

We study selection at synonymous sites through patterns in human–chimpanzee divergence and in intrahuman polymorphism. Using such a close pair of species guarantees against errors caused by multiple substitutions at the same site (Li, 1997) and by ambiguous alignments of introns. Similar to several previous analyses, ours takes into account elevated mutability in mammals of the CpG context, i.e. of nucleotides within 5′CG3′ segments on the DNA sequence (see Li, 1997; Nachman and Crowell, 2000). However, the commonly used classification of sites into those residing and not residing within a CpG context (e.g. Hellman et al., 2003) may obscure the patterns in divergence, since substitutions at a site can affect its placement within this classification (Keightley and Gaffney, 2003).

Thus, we subdivide all sites into four non-overlapping classes: those not preceded by C and not followed by G (nonCpG, Keightley and Gaffney 2003), preceded by C but not followed by G (postC), followed by G but not preceded by C (preG), and preceded by C and followed by G (postCpreG). Sites from the last three classes are CpGprone, as they can reside within CpG context. This approach makes it possible to disentangle the impacts of mutation and selection and to show that weak selection in favor of G and C is a major factor of evolution of synonymous sites in mammals.

2. Materials and methods

2.1. Data

We obtained the human–chimpanzee (hg17-panTro1) alignments and annotation from the Genome Center at U.C. Santa Cruz (Karolchik et al., 2003). Transposable element (TE)-derived intron sites are those masked by RepeatMasker in these alignments. The first 40 and the last 40 nucleotides of an intron, as well as all sites preceded and/or followed by a human–chimpanzee mismatch were excluded from the analysis. Expression level was assayed by the number of ESTs. Mapped polymorphisms were taken from the U.C. Santa Cruz Genome Center annotation of dbSNP release 123 to assembly hg17 of the human genome. For our analyses we used polymorphisms that were obtained in genome-wide, non-exon targeted assays. Only SNPs with the following Submitter Handles in dbSNP flatfiles were used: CSHL-HAPMAP, BCM_SSAHASNP, SC_JCM, SSAHASNP, WI_SSAHASNP, TSC-CSHL, WUGSC_SSAHASNP, SC_SNP, SC. The data used are located as follows.

Human–chimpanzee alignments:

http://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsPanTro1/axtNet/
Human genome annotation:
http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/knownGene.txt.gz
Human polymorphisms mapping to the human genome:
http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/snp.txt.gz
Human polymorphism annotation in dbSNP flatfiles:
ftp://ftp.ncbi.nlm.nih.gov/snp/human/ASN1_flat/

2.2. Review of theory

Consider stochastic mutation–selection–drift equilibrium at a locus (site) with four alleles: A, T, G, and C. Assuming that all mutation rates are low (well below N_e^{-1}, where N_e is the effective population size), a population (approximately) is fixed with one of the alleles most of the time, and occasionally undergoes switches between fixations of different alleles. The frequency of the ith allele, p_i, is the fraction of time when it is fixed. When the population is fixed for the ith allele, the flux of switches to fixation of the jth allele (the per generation probability of a switch), $f_{i>j}$, is the corresponding mutation rate $\mu_{i>j}$, times the population size N, times the probability $g_{i>j}$ that a mutant carrying the jth allele which appeared in a population fixed with the ith allele will reach fixation. The formula for $g_{i>j}$ can be found in Bulmer (1991, Eq. (7)). Equilibrium allele frequencies p_i^{EQ} can be obtained by solving the system of

linear equations which describes the equality of the total rates of switches from and to fixations of each allele (Bulmer, 1991, Eq. (10)):

$$\begin{aligned} p_A(f_{A>T}+f_{A>G}+f_{A>C}) &= p_T f_{T>A}+p_G f_{G>A}+p_C f_{C>A},\\ p_T(f_{T>A}+f_{T>G}+f_{T>C}) &= p_A f_{A>T}+p_G f_{G>T}+p_C f_{C>T},\\ p_G(f_{G>A}+f_{G>T}+f_{G>C}) &= p_A f_{T>G}+p_T f_{T>G}+p_C f_{C>G},\\ p_C(f_{C>A}+f_{C>T}+f_{C>G}) &= p_A f_{A>C}+p_T f_{T>C}+p_G f_{G>C}. \end{aligned} \quad (1)$$

The total rate of evolution (per generation probability of a switch between some allele fixations) at equilibrium is

$$R = p_{\mathrm{A}}^{EQ}(f_{\mathrm{A>T}}+f_{\mathrm{A>G}}+f_{\mathrm{A>C}})+\cdots+p_{\mathrm{C}}^{EQ}(f_{\mathrm{C>A}}+f_{\mathrm{C>T}}+f_{\mathrm{C>G}}) \quad (2)$$

and the total heterozygosity at equilibrium is

$$P = p_{\mathrm{A}}^{EQ}N(\mu_{\mathrm{A>T}}H_{\mathrm{A>T}}+\mu_{\mathrm{A>G}}H_{\mathrm{A>G}}+\mu_{\mathrm{A>C}}H_{\mathrm{A>C}})+\cdots, \quad (3)$$

where $H_{i>j}$ is the expected contribution to heterozygosity by a mutant carrying the jth allele which appeared in a population where the ith allele is fixed (see McVean and Charlesworth, 1999, Eq. (10)).

3. Results

3.1. Intron sites: data

Table 1 presents data on frequencies of the four nucleotides, rate of evolution R (assayed through human–chimpanzee divergence, i.e. the fraction of mismatches in the alignments), and the level of intrahuman polymorphism P (assayed through the density of SNPs) at four-fold synonymous sites and intron sites within 13 533 loci that contain 53 792 introns. First, let us consider introns.

At nonCpG sites, frequencies of G and C are only slightly below 25%. In contrast, postC sites are strongly depleted of G, preG sites are strongly depleted of C, and postCpreG sites are depleted of both G and C (this difference is highly statistically significant, as well as all the differences mentioned below). Of course, this is just another way of saying that introns are depleted of CpG contexts (Bird, 1980). R and P are the lowest at nonCpG sites, and the highest at postCpreG sites. At intron sites of TE origin, frequencies of G and C, as well as P and R, are higher than at nonTE intron sites from the corresponding classes.

Fig. 1 presents data on polarized polymorphisms, those where the ancestral allele is G or C and the derived allele is

Table 1
Properties of sites classified according to their possible CpG context

	All	nonCpG	postC	preG	postCpreG
(a) *Intron sites outside transposable elements*					
Number	52954891	33887011 (64%)	7978950 (15%)	8673858 (16%)	2415072 (5%)
Frequencies					
A	0.2827	0.2606	0.3007	0.3104	0.4331
T	0.3111	0.2883	0.3373	0.3347	0.4588
G	0.2097	0.2330	0.0423	0.3161	0.0538
C	0.1965	0.2180	0.3197	0.0387	0.0543
Divergence	0.01064	0.00932	0.01319	0.01178	0.01663
Polymorphism	0.001294	0.001165	0.001573	0.001416	0.001767
(b) *Intron sites inside transposable elements*					
Number	32287369	19894875 (62%)	5275940 (16%)	5401718 (17%)	1714836 (5%)
Frequencies					
A	0.2702	0.2414	0.3051	0.2924	0.4272
T	0.2923	0.2682	0.3039	0.3236	0.4367
G	0.2207	0.2479	0.0570	0.3290	0.0670
C	0.2169	0.2425	0.3340	0.0551	0.0690
Divergence	0.01274	0.01056	0.01568	0.01468	0.02280
Polymorphism	0.001574	0.001409	0.001815	0.001699	0.002351
(c) *Four-fold synonymous sites*					
Number	1949372	682032 (35%)	654300 (34%)	293573 (15%)	319467 (16%)
Frequencies					
A	0.2165	0.1478	0.2262	0.2039	0.3550
T	0.2320	0.1597	0.2399	0.2169	0.3840
G	0.2425	0.3479	0.0859	0.4751	0.1245
C	0.3090	0.3446	0.4480	0.1042	0.1365
Divergence	0.01282	0.00831	0.01351	0.01182	0.02195
Polymorphism	0.001441	0.001051	0.001529	0.001251	0.002267

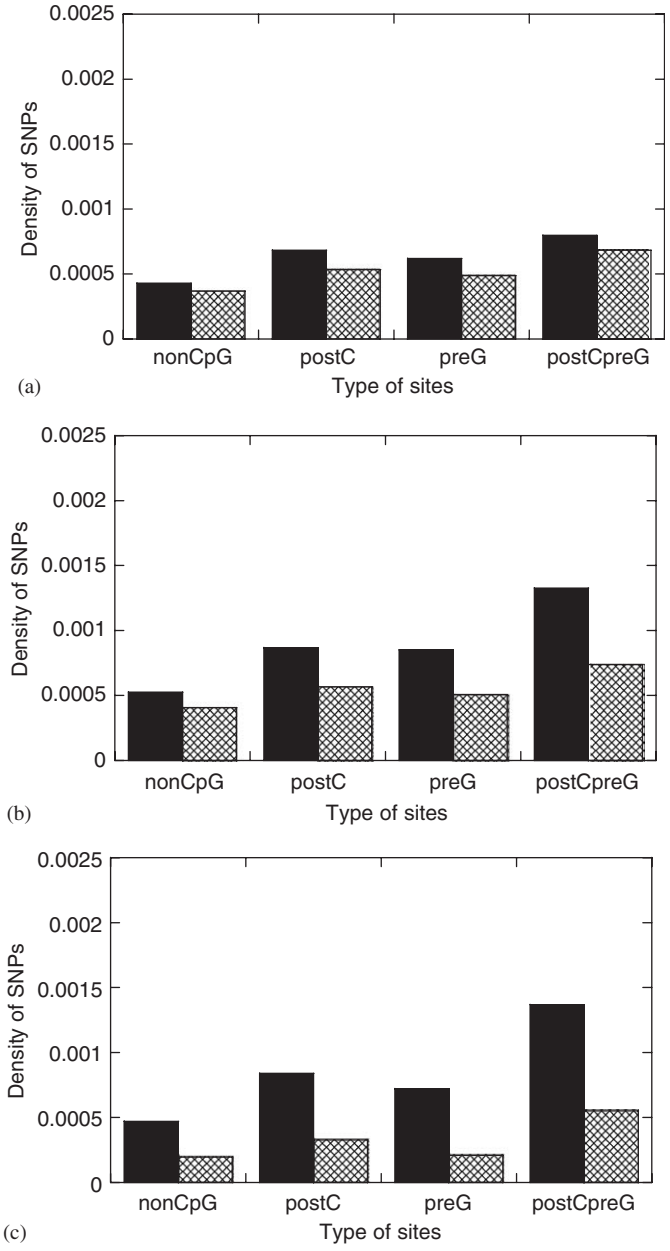

Fig. 1. Densities of reciprocal polymorphisms, GC > AT (black bars) vs. AT > GC (gray bars), at intron nonTE sites (a), intron TE sites (b), and four-fold synonymous sites (c).

A or T (GC > AT, Lercher et al., 2002b), and the reciprocal (AT > GC) (currently, the lack of a close enough outgroup for *Homo* and *Pan* genomes make it impossible to obtain the analogous data on polarized substitutions). At nonTE intron sites, there is only a small excess of GC > AT polymorphisms over AT > GC polymorphisms. In contrast, at TE intron sites this excess is much larger, especially at CpGprone sites.

Fig. 2 presents data on human–chimpanzee divergence at orthologous intron sites located within TEs from different families. On average, TEs which were inserted more recently evolve much faster.

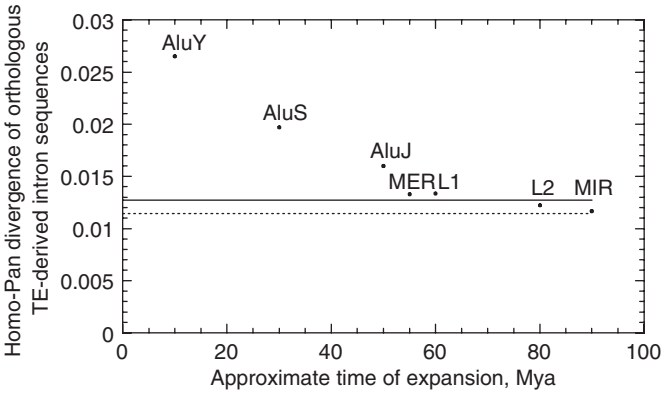

Fig. 2. The dependence of human–chimpanzee divergence between orthologous within-intron copies of transposable elements from different families and subfamilies on the approximate ages of their expansions (Kapitonov and Jurka, 1996; International Human Genome Sequencing Consortium, 2001). The solid line shows the average human–chimpanzee divergence of all intron sequences identified as TEs by RepeatMasker, and the dotted line shows the divergence of intron sequences which remain unmasked.

3.2. *Intron sites: equilibrium with asymmetric mutation at nonTE sites*

The patterns observed at intron sites are consistent with their selective neutrality. NonTE sites appear to be close to mutation–drift equilibrium, while sites of TE origin are losing G's and C's at CpGprone sites.

The ratio of $\mu_{\text{trv-nonCpG}}$, $\mu_{\text{tri-nonCpG}}$, $\mu_{\text{trv-CpG}}$, and $\mu_{\text{tri-CpG}}$, the rates of transversions (of each of the two possible ones) outside CpG, transitions outside CpG, transversions within CpG, and transitions within CpG, is ~1:3:5:30 among mammalian nucleotide substitutions (Nachman and Crowell, 2000; Ebersberger et al., 2002; Kondrashov, 2003; our data; $\mu_{\text{trv-nonCpG}} \sim 0.4 \times 10^{-8}$). Thus, at mutation–drift equilibrium, postC (preG) sites must be depleted of G (C), and postCpreG sites must be depleted of both G and C. CpGprone sites also must evolve faster and be more polymorphic than nonCpG sites.

Indeed, at CpGprone sites some mutation rates are the same as at nonCpG cites but other mutation rates are higher. At a selectively neutral site (locus) with only two alleles, B_1 and B_2, the equilibrium frequency of B_1 (i.e. the probability that B_1 is fixed at a random moment) is $v/(u+v)$ (e.g. Sueoka, 1962), and R, defined as the per generation frequency of switches between fixations of B_1 and of B_2, is $2uv/(u+v)$, where u and v are rates of mutation from B_1 to B_2 and back (e.g. Bulmer, 1991, Eqs. (6) and (7)). Thus, R doubles when v increases from u to infinity. P, defined as heterozygosity, is $4N_e uv/(u+v)$ (McVean and Charlesworth, 1999, Eq. (10)). Without selection, P always (with any number of alleles) changes with the mutation rates in exactly the same way as does R.

This analysis can be extended to mutation–drift equilibrium at a site with four nucleotides (alleles), A, T, G, and C (Bulmer, 1991, Eq. (10), see Methods). Fig. 3 shows how

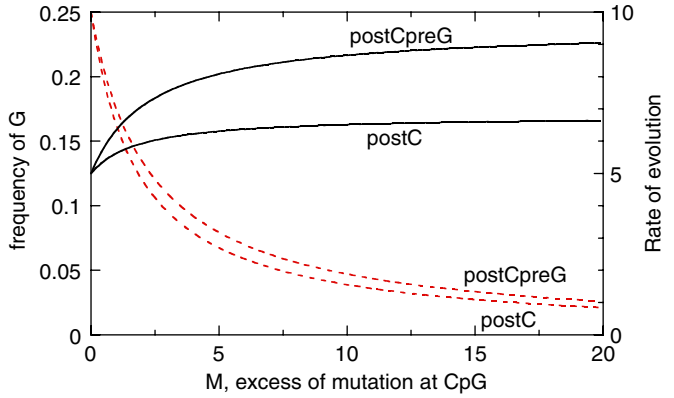

Fig. 3. Frequencies of G (broken red lines) and rates of evolution R in the units of $\mu_{\text{trv-nonCpG}}$ (solid black lines) at the mutation–drift equilibrium without selection as functions of M, where $\mu_{\text{trv-nonCpG}} = 1$, $\mu_{\text{tri-nonCpG}} = 3$, $\mu_{\text{trv-CpG}} = (1+4M/9)\,\mu_{\text{trv-nonCpG}}$, and $\mu_{\text{tri-CpG}} = (1+M)\,\mu_{\text{tri-nonCpG}}$. PostC sites (shown) and preG (not shown) sites evolve at identical rates. Frequency of C at preG sites is the same as frequency of G at postC sites, and at postCpreG sites frequencies of G and C are identical. Properties of nonCpG sites correspond to $M = 0$.

the predicted parameters of such sites depend on M, the relative excess of transitions within the CpG context. For intron nonTE sites, the frequencies of G observed at postCpreG and postC sites, 5.4% and 4.2% (Table 1), imply $M = 8.4$ and $M = 9.0$, respectively (Fig. 3), in excellent agreement with $M \sim 9.0$ which follows from direct data on mutation rates (Kondrashov, 2003). Excesses of R (of P) at postCpreG or at postC sites over the corresponding parameters of nonCpG sites are (Table 1) 1.784 (1.517) or 1.415 (1.350) and imply $M = 14.5$, $M = 3.2$, $M \to \infty$, or $M = 20$, respectively (Fig. 3). However, since under selective neutrality R and P are essentially independent of M when $M > 5$ (Fig. 3), frequencies of mutable nucleotides are more suitable for indirect estimates of high values of M.

Similarity of the levels of the reciprocal polymorphisms GC > AT and AT > GC suggest that nonTE intron sites are close to mutation–drift equilibrium without selection (Eyre-Walker, 1997; Smith and Eyre-Walker, 2001), although, on average, these sites are slowly losing G and C (Fig. 1a; Lercher and Hurst, 2002). This conclusion is supported by direct data on their evolution, obtained for a small fraction of *Homo-Pan* genome alignments for which a suitable outgroup is available (Webster et al., 2003).

3.3. Intron sites: loss of CpG context in transposable elements

Intron sites of TE origin deviate substantially from mutation–drift equilibrium and rapidly lose G and C at CpGprone sites (Fig. 1b). At the moment of insertion, many TEs have a higher (and not a lower, Duret and Hurst, 2001) GC-content and a higher proportion of mutable CpG contexts than nonTE intron sites (Chen et al., 2001). It takes almost 100 Myr for a TE-derived intron segment to reach mutation–drift equilibrium (Fig. 2) and, before this happens, the segment remains more GC-rich and CpG-rich and, thus, evolves faster and is more polymorphic than at equilibrium.

However, even within the nonCpG class, R and P at TE intron sites are ~15% higher than at nonTE intron sites, although for CpGprone classes these excesses are higher (Table 1). This pattern can be caused by slow dissolution of other (different from CpG) mutable contexts within TE-derived intron segments (Hwang and Green, 2004) and/or by negative selection affecting ~10% of nonTE intron sites (Shabalina et al., 2001). Still, we will use nonTE intron sites as a neutral mutation–drift equilibrium point of reference, since they appear to be much closer to this equilibrium than any other sites.

3.4. Four-fold synonymous sites: data

The average rate of evolution at four-fold synonymous sites is similar to that at intron sites (Hughes and Yager, 1997; but see Smith and Hurst, 1998; Chamary and Hurst, 2004) apparently suggesting the lack of selective constraint. However, this overall similarity is misleading (Chamary and Hurst, 2004) and hides a complex pattern.

Synonymous sites from all the four classes are strongly enriched by G and C, relative to the corresponding intron sites (Table 1). In particular, frequencies of G and C at postCpreG synonymous sites are 2.5 times above those expected at mutation–drift equilibrium with $M = 9$ and observed at postCpreG nonTE sites within introns.

In contrast, there is no uniform relationship between R or P at synonymous and the corresponding nonTE intron sites. NonCpG synonymous sites evolve 10% slower, postC and preG synonymous sites evolve at approximately the same rate, and postCpreG synonymous sites evolve ~30% faster, with the levels of polymorphism displaying a very similar pattern (Table 1). As the result of this diversification of P and R at synonymous sites, the ratios of their values at nonCpG, postC or preG, and PostCpreG synonymous sites are ~1:1.5:2.5.

In contrast to nonTE intron sites, GC > AT polymorphisms at synonymous sites are ~2.5 times more common than AT > GC polymorphisms (Fig. 1c; Smith and Eyre-Walker, 2001).

3.5. Four-fold synonymous sites: selection for G and C

There is no reason to assume that context-dependent mutation rates are different between exons and introns. However, the observed contrasts between nonTE intron sites and four-fold synonymous sites can be readily explained by weak selection. Let us make an oversimplified assumption that uniform, constant selection with the coefficient $s \sim 0.25 N_e^{-1}$ favors G or C over A or T at all

synonymous sites. Intermediate dominance will be assumed, with selective advantage $2s$ to homozygotes for the favored allele. Naturally, such selection will always increase frequencies of G and C. In contrast, R and P will be affected differently at sites from different classes, being reduced at nonCpG sites, but elevated at postCpreG sites.

Indeed, constant selection always reduces R and P if all mutation rates are equal or if less mutable alleles are favored (e.g. Akashi, 1999b, c). However, constant selection favoring more mutable allele(s) may increase R (Eyre-Walker, 1992; Eyre-Walker and Bulmer, 1995; McVean and Charlesworth, Figs. 5c and 6c) and P (McVean and Charlesworth, 1999, Fig. 2). At a site with two alleles, B_1 and B_2, and selection for B_2 with coefficient s, at equilibrium $R = 2Suv/[(1-e^{-S})(u+ve^{S})]$ (Bulmer, 1991, Eq. (6) and (7)) and

$$P = \frac{4N_e uv}{u+ve^{-S}}\left(\frac{e^{-S}}{1+e^{-S}}+\frac{1-e^{-S}}{2S}\right) \quad (4)$$

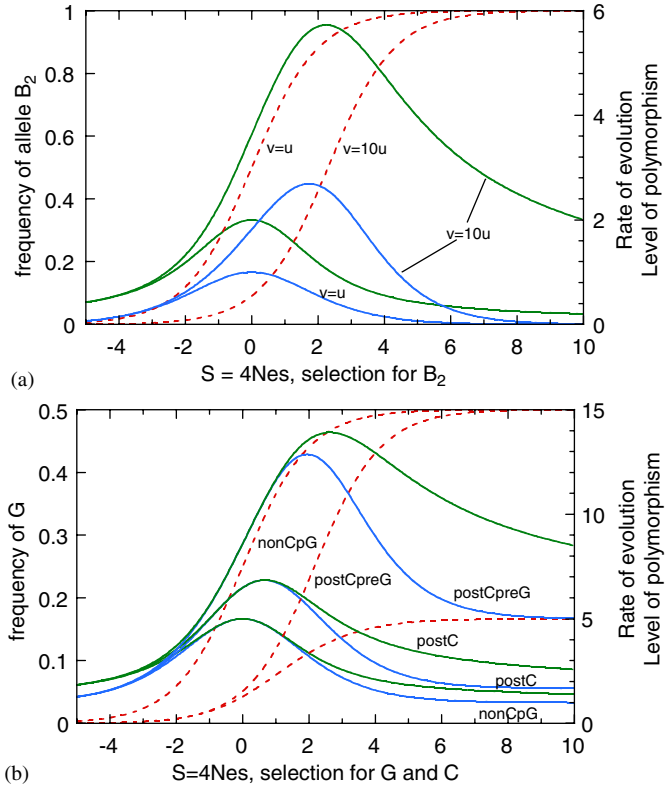

Fig. 4. Equilibrium allele frequencies (broken red lines), rates of evolution in the units of $\mu_{trv\text{-}nonCpG}$ (solid blue lines), and levels of polymorphism in the units of $2N_e\mu_{trv\text{-}nonCpG}$ (solid green lines) as functions of $S = 4N_e s$. (a) Two alleles, B_1 and B_2, under selection with coefficient s in favor of B_2. (b) Four alleles, A, T, G, and C, under selection with coefficient s in favor of G and C, and $M = 9$. PostC sites (shown) and preG sites (not shown) evolve at identical rates. Frequency of C at preG sites is the same as frequency of G at postC sites, and frequencies of G and C are identical at nonCpG and postCpreG sites.

(McVean and Charlesworth 1999, Eq. (15)), where $S = 4N_e s$. If B_2 is more mutable than B_1 ($v > u$), R and P are maximal at some $S > 0$ (Fig. 4a).

The analogous patterns persist in the case of four alleles (Bulmer, 1991, Eq. (10)); McVean and Charlesworth, 1999, Eq. (10)), see Methods). Thus, at nonCpG sites, where mutation is symmetric, R and P are maximal at $S = 0$. In contrast, at CpGprone sites R and P are maximal at a positive S, i.e. under selection favoring more mutable allele G and/or C (Fig. 4b).

These patterns can be used to estimate S roughly. Frequencies of G at nonCpG, postC or preG, and postCpreG synonymous sites imply $S \sim 0.8$, ~ 1.25, and 1.1, respectively (Table 1 and Fig. 4b). The values of R and P at synonymous sites deviate from their values at the corresponding nonTE intron sites by -10%, 0%, and $+30\%$ at nonCpG, postC or preG, and postCpreG sites, respectively, which implies $S \sim 0.9$, $S \sim 1.3$ (or, alternatively, $S = 0$), and $S \sim 0.9$ (or $S > 3.0$), respectively (Fig. 4b). Thus, it appears that $S \sim 1$, so that a typical value of s at a synonymous site is $\sim 0.25N_e^{-1}$.

The 2.5-fold difference between the levels of reciprocal GC > AT and AT > GC polymorphisms at synonymous sites (Maside et al., 2004) of all classes (Fig. 1c) suggests a higher $s \sim 0.55N_e^{-1}$ (data not reported). However, the different levels of reciprocal polymorphisms may be to some extent caused by factors other than selective advantage of G and C (Smith and Eyre-Walker, 2001; Lercher et al., 2002a, b), which work even at nonTE intron sites (Fig. 1a). Thus, $s \sim 0.55N_e^{-1}$ is probably an overestimation.

3.6. Heterogeneity of the observed patterns across genes

The position of a mammalian gene within isochores, genome regions with different GC-contents (see Eyre-Walker and Hurst, 2001) affects the patterns described above. Not surprisingly, genes located within GC-rich genome regions (as assayed by GC-content of their introns) have proportionally more G (Fig. 5a) and C (data not reported) at their synonymous sites. Still, the relationships between rates of evolution at nonTE intron sites and fourfold synonymous sites from the corresponding classes remain the same for genes with all GC-contents, except for those which are very GC-poor, where synonymous sites do not evolve faster than intron sites (Fig. 5b). Perhaps, a factor which favors G and C at synonymous sites is counterbalanced, in genes residing within the most GC-poor genome regions, by another factor responsible for the low regional GC-content. The patterns in P depend on the regional GC-content similarly (data not reported).

In contrast, the nucleotide frequencies (Fig. 6a), R (Fig. 6b), and P (data not reported) depend very little on the expression level of a gene (Duret and Mouchiroud, 2000). A slight increase of the rate of evolution at CpGprone sites with the expression may by due to positive correlation of expression with the GC-content of the gene

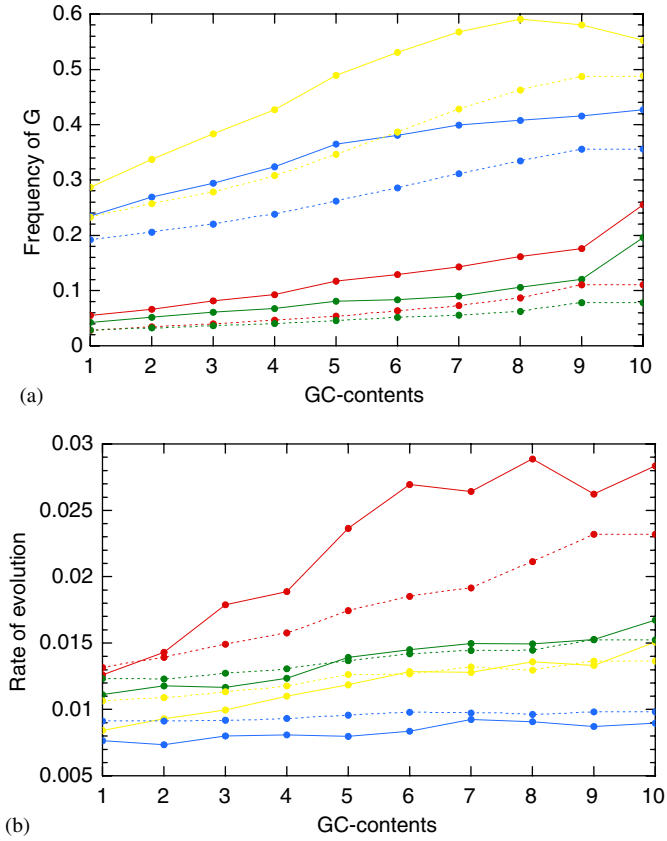

Fig. 5. Frequencies of nucleotide G (a) and the rates of evolution (b) at four-fold synonymous (solid lines) and nonTE intron sites (broken lines) from the four classes (nonCpG—blue, postC—green, preG—yellow, postCpreG—red) in genes split into ten bins of equal sizes according to the GC-content of their introns.

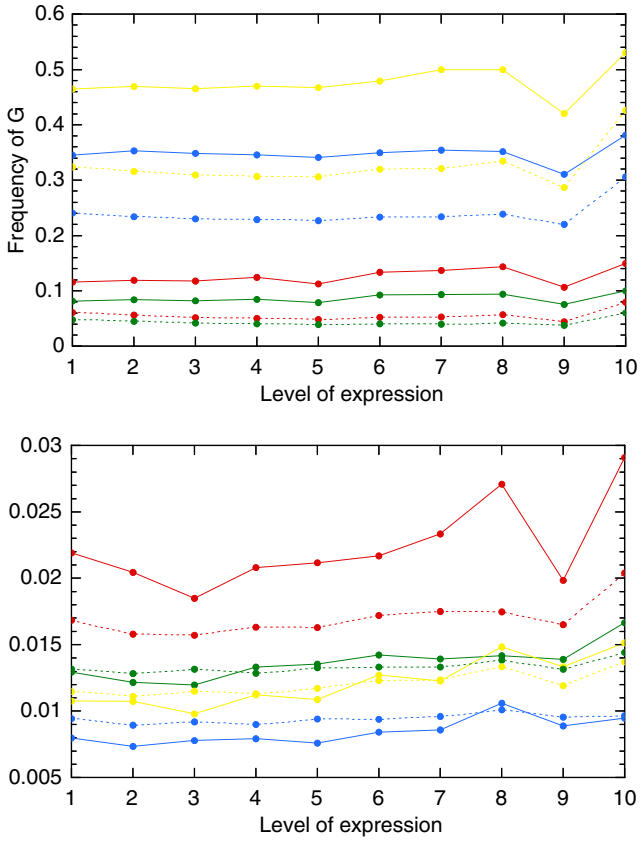

Fig. 6. Frequencies of nucleotide G (a) and the rates of evolution (b) at four-fold synonymous (solid lines) and nonTE intron sites (broken lines) sites from the four classes (nonCpG—blue, postC—green, preG—yellow, postCpreG—red) in genes split into ten bins of equal sizes according to their expression levels.

in mammals (Lercher et al., 2002a; Urrutia and Hurst, 2003).

3.7. Heterogeneity of the observed patterns within genes

The observed patterns are not exactly uniform within genes. Synonymous sites are more GC-enriched within first exons than within last exons of genes. The difference is particularly substantial, 40% vs. 24%, for postCpreG sites. Indeed, first exons are often covered by CpG islands, located in the 5′ ends of genes (Takai and Jones, 2003). However, the rate of evolution of postCpreG sites within first exons is not higher, and even slightly lower than within last exons (data not reported). Perhaps, coefficients of selection in favor of G and C are higher within first exons, and exceed, at some sites, the values which leads to the maximal R. Alternatively, CpG contexts may be less mutable within CpG islands. Different patterns in codon bias at the beginnings vs. the ends of genes have also been observed in bacteria (Hartl et al., 1994). In contrast, there is no difference between GC-contents of first and last introns of genes, although postCpreG sites located close to edges of all introns are more GC-rich and evolve faster than such sites deep inside introns (data not reported).

3.8. More detailed classification of sites

Table 2 presents data on nonTE intron sites and four-fold synonymous sites subdivided into classes according to all 4×4 immediate contexts (the genetic code does not admit postA four-fold synonymous sites). Not surprisingly (e.g. Hess et al., 1994; Hwang and Green, 2004), there is some heterogeneity within sites lumped into nonCpG, postC, or preG inclusive classes of our 2×2 classification. In particular, there is a strong tendency for postApreA sites to be occupied by A more often than by T, and postTpreT sites are occupied by T more often than by A, implying the lack of strand asymmetry in this pattern. Apparently, at postApreA (postTpreT) sites $A>T$ ($T>A$) substitutions are rarer than $T>A$ ($A>T$) substitutions.

Still, CpG is by far the most important context, which is not surprising since its impact on the mutation rate is an order of magnitude higher than that of all other contexts (Hwang and Green, 2004). Also, predictions based on the full 4×4 classification of sites are currently impossible, due to lack of data on the impacts of contexts, other than CpG, at the mutation rate in primates.

Table 2
Properties of sites classified according to all 4 × 4 immediate contexts

	postA		postT		postG		postC	
	Intron	4-fold	Intron	4-fold	Intron	4-fold	Intron	4-fold
preA								
A	0.365		0.259	0.118	0.327	0.265	0.332	0.239
T	0.213		0.281	0.130	0.199	0.090	0.274	0.164
G	0.235		0.239	0.405	0.256	0.283	0.039	0.060
C	0.186		0.220	0.346	0.219	0.362	0.354	0.537
Divergence	0.0090		0.0084	0.0073	0.0076	0.0064	0.0111	0.0108
preT								
A	0.280		0.184	0.095	0.234	0.186	0.288	0.207
T	0.307		0.400	0.213	0.283	0.171	0.349	0.275
G	0.217		0.206	0.346	0.220	0.164	0.046	0.073
C	0.196		0.210	0.346	0.263	0.480	0.316	0.446
Divergence	0.0106		0.0104	0.0108	0.097	0.0091	0.0154	0.0156
preG								
A	0.325		0.245	0.125	0.344	0.311	0.431	0.342
T	0.306		0.364	0.208	0.323	0.230	0.450	0.384
G	0.326		0.349	0.600	0.284	0.306	0.059	0.124
C	0.043		0.042	0.068	0.050	0.153	0.060	0.150
Divergence	0.0132		0.0107	0.0108	0.0113	0.0132	0.0166	0.0220
PreC								
A	0.251		0.180	0.093	0.202	0.179	0.298	0.218
T	0.242		0.321	0.206	0.220	0.143	0.363	0.274
G	0.283		0.246	0.446	0.289	0.309	0.052	0.133
C	0.225		0.253	0.255	0.289	0.368	0.288	0.376
Divergence	0.0095		0.0085	0.0085	0.0091	0.0085	0.0122	0.0143

4. Discussion

Elevated frequencies of nucleotides G and C at synonymous sites, as well as complex relationships between the rates of divergence and levels of polymorphism at synonymous sites vs. intron sites suggests that the majority of synonymous sites of human and chimpanzee genes are under weak selection that favors nucleotides G and C. Comparison of the properties of such sites with those of intron sites of nonTE origin (Table 1), which appear to be close to selectively neutral mutation–drift equilibrium (Fig. 1), and to theoretical predictions (Fig. 4b) implies that the average coefficient of selection s in favor of nucleotides G and C at a human synonymous site is $\sim 0.25N_e^{-1}$, with not too much variation across individual sites. The data are clearly inconsistent with strong selection for G and/or C at some synonymous sites and selective neutrality at other sites: elevated frequencies of G and C can be generated in this way, but elevated rates of evolution (Eyre-Walker, 1992) and levels of polymorphism (McVean and Charlesworth, 1999) at CpGprone sites cannot (Fig. 4b).

A variety of methods produced the following estimates for s: $\sim 1.3N_e^{-1}$ in *Escherichia coli* (Hartl et al., 1994), $\sim 2.2N_e^{-1}$ in *Drosophila simulans* (Akashi, 1995), $\sim 4.6N_e^{-1}$ in *D. pseudoobscura* (Akashi and Schaeffer, 1997), and $\sim 0.65N_e^{-1}$ in *D. americana* (Maside et al., 2004) and *D. miranda* (Bartolomé et al., 2005). Thus, in the units of the corresponding $1/N_e$ values, selection at synonymous sites is apparently weaker in hominids than in *Drosophila*. In *D. simulans*, $N_e \sim 5 \times 10^6$ (Ayala and Hartl, 1993). Estimates of N_e in hominids are to some extent controversial: in modern humans and chimpanzees $N_e \sim (1-2) \times 10^4$ (Yu et al., 2003); however, in the human–chimpanzee common ancestor it was either the same (Rannala and Yang, 2003) or 2–5 times higher (Satta et al., 2004). Thus, the absolute strength of selection at synonymous sites in hominids, $s \sim 10^{-5}$, is close to or even higher than in *Drosophila*, where $s \sim 5 \times 10^{-6}$.

Since coefficients of selection at synonymous sites can vary over many orders of magnitude, their concentration within a narrow range may appear unlikely (Gillespie, 1994). A plausible cause for this concentration is synergistic epistasis (Li, 1987; Akashi, 1995, p. 1074; Akashi, 1996, p. 1305), expected, for example, if synonymous sites are involved in maintaining the structure of mRNA (Innan and Stephan, 2001; Katz and Burge, 2003; Chamary and Hurst, 2005a). With synergistic epistasis, selection against a deleterious nucleotide is negligible when most of the sites of the molecule are occupied by beneficial nucleotides; however, selection gradually gets stronger when the number of deleterious nucleotides increases and eventually becomes sufficient to arrest their further accumulation (Kondrashov, 1994; Piganeau et al., 2001; Berg et al., 2004). This happens when s grows past $\sim 0.1N_e^{-1}$ (Akashi, 1996; Ohta, 2002), and the further growth of s (past $\sim 5.0N_e^{-1}$, Maside et al., 2004; Fig. 2b) can eventually eliminate almost all deleterious nucleotides, making selection negligible again. Thus, at mutation–selection–drift equilibrium, coefficients of selection against deleterious nucleotides at the majority of sites must be confined between $\sim 0.1N_e^{-1}$ and $\sim 5.0N_e^{-1}$ (Akashi, 1996). High values of s in hominids probably suggest that their mRNAs are far from optimal.

Two factors differentially affect the rates of evolution at mammalian nonTE intron vs. four-fold synonymous sites. First, synonymous sites are CpGprone much more often, which is dictated by amino-acid composition of proteins and the genetic code. In particular, highly mutable postCpreG sites are 3 times more common among synonymous sites than among intron sites (Table 1). Second, the interplay of mutation biases and constant selection for G and C reduces R and P at synonymous nonCpG sites, but increases them at such postCpreG sites.

Together, these two factors cause the average values of R and P across all four-fold synonymous sites to be $\sim$20% and $\sim$10%, respectively, above the R and P values for intron sites of nonTE origin. An elevated rate of evolution of synonymous sites, where selection favors more mutable G:C pairs, has been reported for *Drosophila* (McVean and Vieira, 2001). However, in hominids, R and P are also elevated at intron sites of TE origin, due to their deviation from mutation–drift equilibrium, and not to selection (Table 1, Figs. 1 and 2). It is a mere coincidence

(Chamary and Hurst, 2004) that the average rate of evolution at all four-fold synonymous sites is very close to that at all intron sites (Hughes and Yager, 1997; Subramanian and Kumar, 2003). Similarly, all four nucleotide frequencies at four-fold synonymous sites are close to 25% (Table 1) due to selection in favor of G and C being counterbalanced by their elevated mutability, and not to selective neutrality.

With $s \sim 0.25 N_e^{-1}$, selection is weak enough to allow fixations of many slightly deleterious nucleotides. If suboptimal nucleotides (mostly, A and T) with $s \sim 10^{-5}$ occupy ~30% from $\sim 3 \times 10^7$ synonymous sites in the diploid mammalian genome, an organism carries $\sim 10^7$ deleterious nucleotides at such sites, which constitute ~100 lethal equivalents. The survival of a population of such organisms requires synergistic epistasis among loci (Kondrashov, 1995). A substantial fraction of new mutations replaces a suboptimal nucleotide with the optimal one and, thus, are slightly beneficial.

Our analysis makes it possible to explain several observations. At synonymous sites, the frequency of C is higher than the frequency of G (Chamary and Hurst, 2004) because four-fold postC sites, where G is rare, are >2 times more common than preG sites, where C is rare, (Table 1). This fact, dictated by the genetic code, where all codon families with C at the second position are four-fold degenerate, could be responsible for strand asymmetry in evolution at synonymous sites (Webster and Smith, 2004). Synonymous sites of constitutive exons have higher frequencies of G and C and evolve more rapidly than such sites of alternatively spliced exons (Iida and Akashi, 2000) because selection in favor of G and C at synonymous sites is stronger in constitutive exons. If synonymous sites are released from selective constraint, for example after a gene turns into a pseudogene, this leads to a large, temporary increase in the rate of their evolution (Bustamante et al., 2002), due to the above mutation–drift equilibrium frequencies of mutable C and G at CpGprone synonymous sites. The same mechanism leads to temporarily elevated rates of evolution of newly inserted transposons (Fig. 2).

Since biased gene conversion can lead to the same dynamics as selection (e.g. Lercher et al., 2002a), we cannot formally discriminate between the two. However, an important role of biased gene conversion in creating the patterns described above is unlikely (Eyre-Walker, 1999), because of the contrasts between exons and introns of the same genes. While selection can obviously be very different at synonymous exon sites vs. intron sites, it is unclear how the rate of biased gene conversion could change drastically at exon/intron boundaries. The increased probability of fixation of AT > GC mutations (Webster and Smith, 2004), as well as the excess of GC > AT over AT > GC polymorphisms (Fig. 1c) can be due to selection for G and C (Smith and Eyre-Walker, 2001; Lercher et al., 2002a, b; Webster et al., 2003). The exon–intron contrasts also argue against transcription-coupled repair bias (Green et al., 2003; Majewski, 2003) as a cause of patterns reported here.

Widespread, weak advantage of nucleotides G and C at synonymous sites supports selection on mRNA stability as an important factor in the dynamics of such sites (Chamary and Hurst, 2005a). In contrast, this advantage appears to be inconsistent with the possible involvement of such sites in splice regulation (Eskesen et al., 2004; Fairbrother et al., 2004; Willie and Majewski, 2004), since GC-content at such sites diminishes near exon–intron junctions (Chamary and Hurst, 2005b).

The available methods of estimating K_s, the evolutionary distance at synonymous sites (Yang, 1997), do not accommodate trimodal distributions of the rates of evolution and nucleotide frequencies at individual sites (Table 1). Since the context of a synonymous site is mostly determined by the surrounding non-synonymous sites, a synonymous site remains within the same class for a long time. Even for a relatively close mouse–rat pair, divergences at CpGprone four-fold synonymous sites, estimated using the Tamura–Nei formula (Tamura and Nei, 1993), are well below, relative to the divergence at nonCpG sites, of what is expected from the 1:1.5:2.5 ratios observed in the human–chimpanzee pair (data not reported). This indicates that multiple substitutions occurred at CpGprone synonymous sites since rat–mouse divergence, and that the Tamura–Nei formula, which assumes equal rates of reciprocal substitutions, underestimates divergences at such sites. In the case of mouse–human divergence, saturation at CpGprone sites is much more pronounced (data not reported). Thus, estimates of K_s between distant mammals are unreliable.

Even the correct values of K_s should not be used to estimate mutation rates in mammals, due to lack of neutrality (Kondrashov, 2001). Probably, neutral divergence between a pair of mammals (and, thus, the mutation rates outside CpG context multiplied by the number of generations of their independent evolution) can be approximated as ~1.1 times the (correctly estimated) K_s at nonCpG four-fold synonymous sites. Whether synonymous sites are a suitable neutral point of reference (Fay et al., 2001, 2002; Anisimova et al., 2002; Smith and Eyre-Walker, 2002; Eyre-Walker, 2002) in tests for positive selection, is not clear (Akashi, 1995), although the answer may be affirmative for hominids, since selection with $s \sim 0.25 N_e^{-1}$ at synonymous sites affects R and P in almost the same way (Fig. 4b).

Supporting information

C codes for calculating the properties of mutation–drift–selection equilibrium are available from ftp://ftp.ncbi.nih.gov/pub/kondrashov/k4.

Acknowledgment

This research was supported in part by the Intramural Research Program of the NIH, National Library of Medicine.

References

Akashi, H., 1995. Inferring weak selection from patterns of polymorphism and divergence at silent sites in Drosophila DNA. Genetics 139, 1067–1076.

Akashi, H., 1996. 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.

Akashi, H., 1999a. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151, 221–238.

Akashi, H., 1999b. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene 238, 39–51.

Akashi, H., 1999c. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151, 221–238.

Akashi, H., 2003. Translational selection and yeast proteome evolution. Genetics 164, 1291–1303.

Akashi, H., Schaeffer, S.W., 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics 146, 295–307.

Andersson, S.G., Kurland, C.G., 1990. Codon preferences in free-living microorganisms. Microbiol. Rev. 54, 198–210.

Anisimova, M., Bielawski, J.P., Yang, Z., 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. 19, 950–958.

Ayala, F.J., Hartl, D.L., 1993. Molecular drift of the bride of sevenless (boss) gene in Drosophila. Mol. Biol. Evol. 10, 1030–1040.

Bartolomé, C., Maside, X., Yi, S., Grant, A.L., Charlesworth, B., 2005. Patterns of selection on synonymous and nonsynonymous variants in Drosophila miranda. Genetics 169, 1495–1507.

Berg, J., Willmann, S., Lässig, M., 2004. Adaptive evolution of transcription factor binding sites. BMC Evol. Biol. 4: Research42.

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

Bulmer, M., 1991. The selection-mutation–drift theory of synonymous codon usage. Genetics 129, 897–907.

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

Carlini, D.B., Stephan, W., 2003. In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. Genetics 163, 239–243.

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

Chamary, J.V., Hurst, L.D., 2005a. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. 6, R75.

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

Chen, F.-C., Vallender, E.J., Wang, H., Tzeng, C.-S., Li, W.-H., 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. J. Hered. 92, 481–489.

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.

Debry, R.W., Marzluff, W.F., 1994. Selection on silent sites in the rodent H3 histone gene family. Genetics 138, 191–202.

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

Duan, J.B., Wainwright, M.S., Comeron, J.M., Saitou, N., Sanders, A.R., Gelernter, J., Gejman, P.V., 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.

Duret, L., Hurst, L.D., 2001. The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Mol. Biol. Evol. 18, 757–762.

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.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., Galtier, N., 2002. Vanishing GC-rich isochores in mammalian genomes. Genetics 162, 1837–1847.

Ebersberger, I., Metzler, D., Schwarz, C., Paabo, S., 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. Am. J. Hum. Genet. 70, 1490–1497.

Eskesen, S.T., Eskesen, F.N., Ruvinsky, A., 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., 1992. The effect of constraint on the rate of evolution in neutral models with biased mutation. Genetics 131, 233–234.

Eyre-Walker, A., 1997. Differentiating selection and mutation bias. Genetics 147, 1983–1987.

Eyre-Walker, A., 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics 152, 675–683.

Eyre-Walker, A., 2002. Changing effective population size and the McDonald-Kreitman test. Genetics 162, 2017–2024.

Eyre-Walker, A., Bulmer, M., 1995. Synonymous substitution rates in enterobacteria. Genetics 140, 1407–1412.

Eyre-Walker, A., Hurst, L.D., 2001. The evolution of isochores. Nat. Rev. Genet. 2, 549–555.

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

Fay, J.C., Wyckoff, G.J., Wu, C.I., 2001. Positive and negative selection on the human genome. Genetics 158, 1227–1234.

Fay, J.C., Wyckoff, G.J., Wu, C.-I., 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature 415, 1024–1026.

Gillespie, J.H., 1994. Substitutional processes in molecular evolution. III. Deleterious alleles. Genetics 138, 943–952.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pave, A., 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, r49–r62.

Green, P., Ewing, B., Miller, W., Thomas, P.J., Green, E.D., NISC Comparative Sequencing Program, 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. 33, 514–517.

Hartl, D.L., Moriyama, E.N., Sawyer, S.A., 1994. Selection intensity for codon bias. Genetics 138, 227–234.

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

Hess, S.T., Blake, J.D., Blake, R.D., 1994. Wide variations in neighbor-dependent substitution rates. J. Mol. Biol. 236, 1022–1033.

Hughes, A.L., Yager, M., 1997. Comparative evolutionary rates of introns and exons in murine rodents. J. Mol. Evol. 45, 125–130.

Hwang, D.G., Green, P., 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl Acad. Sci. USA 101, 13994–14001.

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.

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

Innan, H., Stephan, W., 2001. Selection intensity against deleterious mutations in RNA secondary structure and rate of compensatory nucleotide substitutions. Genetics 159, 389–399.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T., 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238, 143–155.

Kapitonov, V., Jurka, J., 1996. The age of Alu subfamilies. J. Mol. Evol. 42, 59–65.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D., Kent, W.J., 2003. The UCSC genome browser database. Nucleic Acids Res. 31, 51–54.

Katz, L., Burge, C.B., 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. 13, 2042–2051.

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

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

Kondrashov, A.S., 1994. Muller's ratchet under epistatic selection. Genetics 136, 1469–1473.

Kondrashov, A.S., 1995. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? J. Theor. Biol. 175, 583–594.

Kondrashov, A.S., 2001. Sex and U. Trends Genet. 17, 75–77.

Kondrashov, A.S., 2003. A direct estimate of human per nucleotide spontaneous mutation rate. Hum. Mutat. 21, 12–27.

Kumar, S., Subramanian, S., 2002. Mutation rates in mammalian genomes. Proc. Natl Acad. Sci. USA 99, 803–808.

Lercher, M.J., Hurst, L.D., 2002. Can mutation or fixation biases explain the allele frequency distribution of human single nucleotide polymorphisms (SNPs)? Gene 300, 53–58.

Lercher, M.J., Smith, N.G., Eyre-Walker, A., Hurst, L.D., 2002a. The evolution of isochores: evidence from SNP frequency distributions. Genetics 162, 1805–1810.

Lercher, M.J., Urrutia, A.O., Hurst, L.D., 2002b. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. 31, 180–183.

Li, W.H., 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. 24, 337–345.

Li, W.-H., 1997. Molecular Evolution. Sinauer, Sunderland.

Lu, J., Wu, C.I., 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., 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. Am. J. Hum. Genet. 73, 688–692.

Maside, X., Lee, A.W., Charlesworth, B., 2004. Selection on codon usage in *Drosophila americana*. Curr. Biol. 14, 150–154.

McVean, G., Charlesworth, B., 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res. 74, 145–158.

McVean, G.A., Vieira, J., 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics 157, 245–257.

Nachman, M.W., Crowell, S.L., 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297–304.

Nielsen, R., Akashi, H., 2003. Action of Purifying Selection at Silent Sites. Encyclopedia of the Human Genome. Nature Publishing Group, London.

Ohta, T., 2002. Near-neutrality in evolution of genes and gene regulation. Proc. Natl Acad. Sci. USA 99, 16134–16137.

Piganeau, G., Westrelin, R., Tourancheau, B., Gautier, C., 2001. Multiplicative versus additive selection in relation to genome evolution: a simulation study. Genet. Res. 78, 171–175.

Rannala, B., Yang, Z., 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656.

Satta, Y., Hickerson, M., Watanabe, H., O'hUigin, C., Klein, J., 2004. Ancestral population sizes and species divergence times in the primate lineage on the basis of intron and BAC end sequences. J. Mol. Evol. 59, 478–487.

Shabalina, S.A., Ogurtsov, A.Y., Kondrashov, V.A., Kondrashov, A.S., 2001. Selective constraint in intergenic regions of human and mouse genomes. Trends Genet. 17, 373–376.

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

Smith, N.G.C., Eyre-Walker, A., 2001. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol. Biol. Evol. 18, 982–986.

Smith, N.G.C., Eyre-Walker, A., 2002. Adaptive protein evolution in *Drosophila*. Nature 415, 1022–1024.

Smith, N.G.C., Hurst, L.D., 1998. Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. J. Mol. Evol. 47, 493–500.

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

Sorensen, M.A., Kurland, C.G., Pedersen, S., 1989. Codon usage determines translation rate in *E. coli*. J. Mol. Biol. 207, 365–377.

Subramanian, S., Kumar, S., 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. 13, 838–844.

Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl Acad. Sci. USA 48, 582–592.

Takai, D., Jones, P.A., 2003. The CpG island searcher: a new WWW resource. Silico Biol. 3, 235–240.

Tamura, K., Nei, M., 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.

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

Webster, M.T., Smith, N.G.C., 2004. Fixation biases affecting human SNPs. Trends Genet. 20, 122–126.

Webster, M.T., Smith, N.G.C., Ellegren, H., 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. Mol. Biol. Evol. 20, 278–286.

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

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

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13, 555–556.

Yu, N., Jensen-Seaman, M.I., Chemnick, L., Kidd, J.R., Deinard, A.S., Ryder, O., Kidd, K.K., Li, W.-H., 2003. Low nucleotide diversity in chimpanzees and bonobos. Genetics 164, 1511–1518.