A population genetic model for the evolution of synonymous codon usage: patterns and predictions

GILEAN A. T. McVEAN* AND BRIAN CHARLESWORTH
Institute of Cell, Animal and Population Biology, Ashworth Laboratories, King's Building, West Mains Road, Edinburgh EH9 3JT, Scotland, UK

(*Received 12 October 1998 and in revised form 12 February 1999*)

Summary

Patterns of synonymous codon usage are determined by the forces of mutation, selection and drift. We elaborate on previous population genetic models of codon usage to incorporate parameters of population polymorphism, and demonstrate that the degree of codon bias expected in a single sequence picked at random from the population is accurately predicted by previous models, irrespective of population polymorphism. This new model is used to explore the relationships between synonymous codon usage, nucleotide site diversity and the rate of substitution. We derive the equilibrium frequency distribution of weakly selected segregating sites under the infinite-sites model, and the expected nucleotide site diversity. Contrary to intuition, levels of silent-site diversity can increase with the strength of selection acting on codon usage. We also predict the effects of background selection on statistics of synonymous codon usage and derive simple formulae to predict patterns of codon usage at amino acids with more than two synonymous codons, and the effects of variation in selection coefficient between sites within a gene. We show that patterns of silent-site variation and synonymous codon usage on the X chromosome and autosomes in *Drosophila* are compatible with recessivity of the fitness effects of unpreferred codons. Finally, we suggest that there still exist considerable discrepancies between current models and data.

1. Introduction

Patterns of synonymous codon usage vary considerably, both within genomes and between species. For a number of organisms, evidence suggests that biased codon usage reflects the action of selection to maximize translational efficiency and accuracy. In several prokaryotes, such as *E. coli*, there are clear trends relating the degree of codon bias to gene expression level (Gouy & Gautier, 1982), the relative abundance of iso-accepting tRNAs (Ikemura, 1981, 1985) and the rate of synonymous substitution (Sharp & Li, 1987; but see Eyre-Walker & Bulmer, 1995). For some amino acids, the effect on translational efficiency of alternative codons can be demonstrated experimentally (Sørensen & Pedersen, 1991). In eukaryotes, many of these patterns hold for yeast (Sharp *et al.*, 1986; Sharp & Cowe, 1991), and *Drosophila* (Shields *et al.*, 1988; Powell & Moriyama, 1997), though not for mammalian genes (Eyre-Walker, 1991).

The degree of codon bias can be interpreted as the result of an interaction between the forces of selection, mutation bias and drift (Kimura, 1983; Li, 1987; Bulmer, 1991). If the selective difference between alternative codons (s) is much less than the reciprocal of the effective population size (N_e), stochastic processes and mutational biases predominate. In contrast, for values of $|N_e s|$ much greater than one, selection dominates. For those genes where neither force dominates, models of the interaction between mutation, selection and drift are required to make predictions as to the relative strength of each.

A number of simple population genetic models (Kimura, 1983; Li, 1987; Bulmer, 1991; Kondrashov, 1995) have been studied, which relate the degree of codon bias for 2-fold degenerate amino acids to $|N_e s|$.

[*] Telephone: +44 (0)131 650 5543. Fax: +44 (0)131 650 6564. e-mail: g.mcvean@ed.ac.uk.

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

These are based on the distribution of allele frequencies under reversible mutation, at independently evolving sites, as originally formulated by Wright (1949). The expected proportion of sites fixed for the preferred codon type can be found either by integrating the distribution of allele frequencies over the interval $1-1/(2N)$ to 1 (Li, 1987; Kondrashov, 1995), or by considering the flux of substitutions at equilibrium (Bulmer, 1991). If it is assumed that (1) selection acts independently at all sites, (2) the fitness difference between homozygotes for the preferred and unpreferred codons is $2s$, and (3) heterozygotes are of intermediate fitness, then both approaches lead to the result that the proportion of sites fixed for the preferred codon at mutation–selection–drift equilibrium is

$$E(x) \approx \frac{1}{1+(\mu_{10}/\mu_{01})\,\mathrm{e}^{-4N_e s}}, \tag{1}$$

where μ_{10} and μ_{01} are (respectively) the rates of mutation away from and to the preferred codon.

This formula has two major limitations. First, it suggests that the proportion of sites fixed for the preferred codon type is independent of the actual mutation rates, and that only the relative values are of interest. Yet the absolute mutation rate will affect the dynamics through the number of sites found to be polymorphic in the population, and the expected pairwise differences between any two sequences (nucleotide site diversity). Hence a complete description of the situation at equilibrium must also include parameters of polymorphism. Secondly, the frequency distribution of Wright (1949) is a continuous approximation to a discrete distribution, which is inaccurate at extreme allele frequencies; hence the method of integrating over the interval $1-1/(2N)$ to 1 can lead to significant errors (Ewens, 1979, p. 158). Further, the agreement between the two derivations seems coincidental as both are estimating the frequency of sites fixed of the favoured codon, but polymorphic sites are only considered in the first derivation.

In the following section we consider the flux of mutations occurring at mutation–selection–drift equilibrium to provide an extension of a generalized form of this model, which includes both parameters of polymorphism and an arbitrary selection scheme. This allows us to consider the number of sites fixed for a particular codon type, the proportion of sites segregating at different allele frequencies, the expected nucleotide site diversity, the rate of substitution, and the relationship between these statistics, as employed by various tests of the neutral hypothesis (e.g. Hudson *et al.*, 1987; McDonald & Kreitman, 1991). We then develop this model to consider the effects of sex-linkage, background selection, the difference between 2-fold and greater than 2-fold degenerate amino acids, and variation in the selection coefficient across sites within the same gene. The model assumes that sites evolve independently of one another, as in previous analytical treatments of codon usage (Li, 1987; Bulmer, 1991). Relaxation of this assumption may have important consequences for patterns of codon usage evolution (Li, 1987), and is the focus of continuing research.

2. The model

(i) *Infinite-sites model*

We first consider the case of a 2-fold degenerate amino acid. The genome consists of three types of site: a proportion, m_1, fixed for the preferred codon (A), a proportion, m_0, fixed for the unpreferred codon (a) and a proportion, m_s, which are polymorphic within the population. The probabilities of fixation of novel variants are u_{01} and u_{10} for A and a respectively (the subscripts on u_{01} relate to the fixation of the preferred allele (1) in place of the unpreferred allele (0)). Mutation is reversible, such that A mutates to a with frequency μ_{10} and the reverse mutation occurs with frequency μ_{01}. Mutations always appear at sites which are not currently segregating (the infinite sites assumption: Kimura, 1971).

At mutation–selection–drift equilibrium, the number of mutations appearing at sites fixed for the preferred type must be equal to the number of segregating sites becoming fixed for the preferred type in the same generation:

$$m_1\mu_{10} = m_1\mu_{10}(1-u_{10})+m_0\mu_{01}u_{01}. \tag{2}$$

This gives the relationship

$$r = \frac{m_1}{m_0} = \frac{\mu_{01}}{\mu_{10}}\cdot\frac{u_{01}}{u_{10}}. \tag{3}$$

If sites evolve independently of one another, the fate of new mutations represents an exchangeable process (Ewens, 1979, p. 77) and equilibrium patterns of evolution can be predicted from the expected behaviour of single mutations (the principle of ergodicity). It follows that the proportion of segregating sites in the population is given by the relationship

$$m_s = 2N(m_1\mu_{10}\bar{t}_{10}+m_0\mu_{01}\bar{t}_{01}), \tag{4}$$

where $\bar{t}_{10}$ and $\bar{t}_{01}$ are the average times until loss or fixation for unpreferred and preferred mutations respectively. The proportions of sites fixed for each type are

$$m_1 = s[1+r'+2N(\mu_{10}\bar{t}_{10}+r'\mu_{01}\bar{t}_{01})]^{-1} \tag{5a}$$

and

$$m_0 = [1+r+2N(r\mu_{10}\bar{t}_{01}+\mu_{01}\bar{t}_{10})]^{-1}, \tag{5b}$$

where $r' = 1/r$.

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

For a specific model of selection, the values of u_{01}, u_{10}, $\bar{t}_{01}$ and $\bar{t}_{10}$ can be calculated from the formulae of Kimura (1962) and Kimura & Ohta (1969*a*, *b*); see Appendix A. The expected nucleotide site diversity (π) can be calculated by summing the contribution to heterozygosity along the sample path for novel mutations (Kimura, 1971; Ewens, 1979, p. 239) and weighting by the frequency at which such mutations occur. This gives

$$\pi = 2N[m_1 \mu_{10} H_{10} + m_0 \mu_{01} H_{01}], \tag{6}$$

where H_{10} and H_{01} are the expected contribution to heterozygosity for unpreferred and preferred mutations respectively. The equilibrium rate of substitution is similarly given by

$$k = 2N[m_1 \mu_{10} u_{10} + m_0 \mu_{01} u_{01}]. \tag{7}$$

(ii) *Genic selection*

Consider the specific case where the fitness difference between homozygotes for preferred and unpreferred codons is $2s$ and between heterozygotes and homozygotes the fitness difference is s (i.e. the dominance coefficient is equal to one-half). Novel variants arise at autosomal loci as single mutations in a diploid population of N individuals with initial frequency $p = 1/2N$. The effective population size is N_e. The probability of fixation of a novel, preferred variant is (Kimura, 1962)

$$u_{01} = \frac{2\,(N_e/N)\,s}{(1-e^{-4N_e s})} \tag{8}$$

for $s \ll 1$. The sojourn time density for a preferred allele in the frequency interval x to $x+dx$ is

$$\Phi_{01}(x) = 2\frac{N_e}{N}\frac{(1-e^{-4N_e s(1-x)})}{(1-e^{-4N_e s})\,x(1-x)} \tag{9}$$

for $s \ll 1$. Similarly the expected contribution to heterozygosity for a novel, preferred variant is

$$H_{01} = 4\frac{N_e}{N}\left(1-\frac{1-e^{-4N_e s}}{4N_e\,s}\right)(1-e^{-4N_e s})^{-1} \tag{10}$$

for $s \ll 1$. Each of the quantities given above are for the preferred alleles; values for the unpreferred alleles are given by changing the sign of the selection coefficient. As $|N_e s|$ increases, the greatest contribution to heterozygosity is from mutations occurring at sites initially fixed for the preferred allele. For $|N_e s| \gg 1$, almost all sites are fixed for the preferred allele and the nucleotide site diversity approaches $2\mu_{10}/s$, i.e. twice the deterministic equilibrium frequency of the deleterious mutations. Formulae (8)–(10) can be substituted into equations (4)–(7) to generate results of the quantities of interest. For the case when $N_e s = 0$, the above equations recover the expected results for neutral alleles (Kimura, 1983).

(iii) *Multiple codons*

The model for 2-fold degenerate codons can readily be extended to the case of k synonymous codons. Codon i mutates to codon j with frequency μ_{ij}, the probability of fixation of codon j in place of codon i is u_{ij} where the selection coefficient is given by $s_i - s_j = s_{ij}$ (genic selection is assumed for convenience) and the proportion of sites fixed for codon i is m_i. At mutation–selection–drift equilibrium, the number of mutations appearing at any codon type is equal to the number becoming fixed for that type in the same generation. This gives k equations of the form

$$m_i \sum_{j \neq i} \mu_{ij} = m_i \sum_{j \neq i} \mu_{ij}(1-u_{ij}) + \sum_{j \neq i} m_j \mu_{ji} u_{ji}. \tag{11}$$

If we make the approximation $\sum_i m_i = 1$ (i.e. the proportion of segregating sites is assumed to be negligible), then this reduces to a set of $k-1$ simultaneous equations that can be solved explicitly. The time taken to loss or fixation could be included in a manner similar to that previously outlined, but the simplifying assumption does not appear to affect the result significantly. The frequency of each codon can then be used to predict the equilibrium nucleotide site diversity and rate of substitution:

$$\pi = 2N \sum_i m_i \sum_{j \neq i} \mu_{ij} H_{ij}, \tag{12}$$

$$k = 2N \sum_i m_i \sum_{j \neq i} \mu_{ij} u_{ij}. \tag{13}$$

The selective differences between multiple codons can be portrayed in two ways. First, multiple codons may form a smaller number of equivalence classes, where mutations within a class are effectively neutral. Secondly, selective differences may exist between all codons in a manner similar to that of the two-codon case.

In the results presented below, the patterns of mutation bias are reduced to two parameters; a transition–transversion ratio, and a general GC → AT mutation pressure. Codon bias can be measured by the effective number of codons, or ENC (Wright, 1990), such that for a k-fold degenerate site, where the frequency of each codon is f_i, we have

$$\text{ENC} = \left[\sum_i f_i^2\right]^{-1}$$

and bias is given by

$$\frac{k-\text{ENC}}{k-1}.$$

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

(iv) *Within-gene variation in the selection coefficient*

Given that there is evidence for selection on translational efficiency and error avoidance in *E. coli* (Gouy & Gautier, 1982; Eyre-Walker, 1996) and *D. melanogaster* (Shields *et al.*, 1988; Akashi, 1994), it seems likely that the strength of selection acting on codon usage will differ both between genes and between sites within a gene. Using the model presented here, we can consider how the relationship between the average selection coefficient within a gene and statistics of codon usage depend on the shape of the distribution of selection coefficients across sites within a gene.

We consider two variants of the gamma distribution as candidates for the variation in selection coefficients across sites within a gene. The gamma distribution has been used extensively in population genetics models where the selection coefficient varies across sites (Yang, 1996). The general form of the distribution is

$$\phi(y) = \frac{e^{-\beta y} y^{\alpha-1}}{\int_0^\infty e^{-\beta y} y^{\alpha-1}\, dy}. \tag{14}$$

We have considered the exponential distribution (shape parameter $\alpha = 1$), and the case when $\alpha = 2$, which gives a skewed, bell-shaped distribution. The value of β is chosen such that the mean of the distribution, α/β, is equal to the average value of $|4N_e s|$ across sites within a gene.

3. Results

(i) *Autosomal genes*

For the case of a 2-fold degenerate amino acid with genic selection, Fig. 1 shows the proportion of sites fixed for the preferred codon plotted against N_e times the selection coefficient between alternative alleles (N_e is held constant while s varies). Also shown are the values predicted by the formula used by previous authors (Equation 1) and the single-sequence bias (the expected frequency of the preferred codon in a single sequence picked at random from the population). This latter statistic is both the simplest empirical measure of bias, and fits the values predicted by (1) very closely, whereas the discrepancy between the proportion fixed for the favoured codon and the single-sequence bias increases as $|N_e s|$ increases.

When mutation bias is such that the rate of mutation away from the preferred codon is greater than that to it, a greater proportion of sites are fixed for the unpreferred type than for the preferred type for $|N_e s| \ll 1$ (Fig. 1*b*). As found previously (Li, 1987), selection coefficients of the order of $|N_e s| \approx 1$ are sufficient to create a large degree of bias.

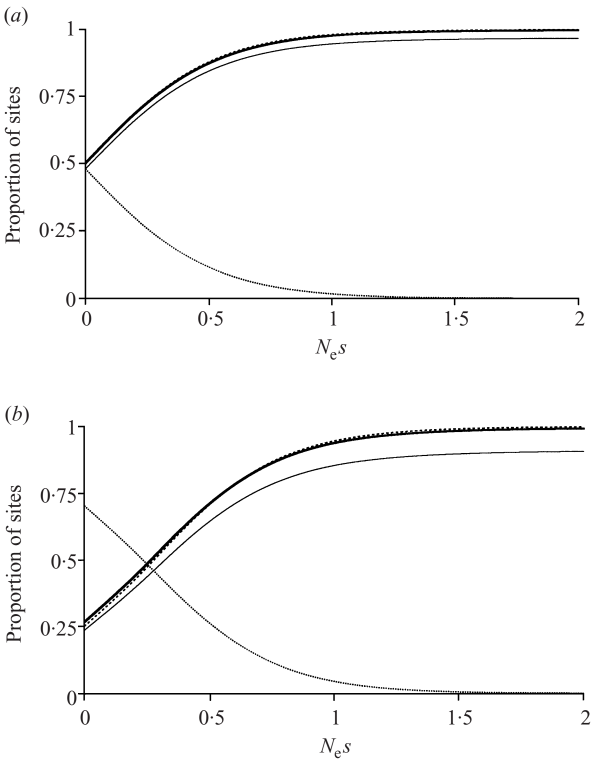

Fig. 1. Proportion of sites fixed for preferred codons (continuous line, from Equation 5*a*) and unpreferred codons (dotted line, from Equation 5*b*), single-sequence bias (bold line) and the expected bias as calculated by Equation 1 (dashed line). (*a*) Equal mutation rates, $N_e = N = 10^4$, $\mu_{10} = \mu_{01} = 10^{-7}$ (*b*) Unequal mutation rates $N_e = N = 10^4$, $\mu_{10} = 3 \times 10^{-7}$, $\mu_{01} = 10^{-7}$.

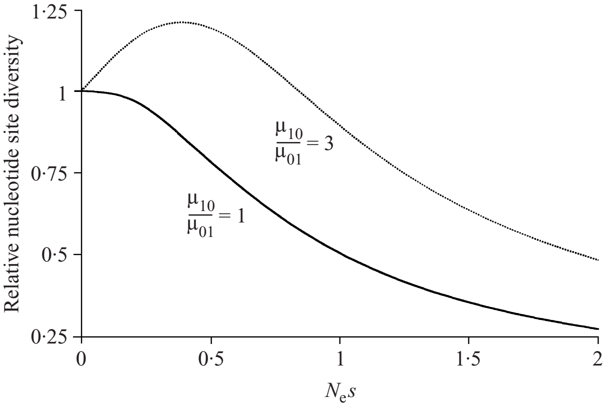

Fig. 2. Nucleotide site diversity relative to that expected for the neutral case, calculated from (6). $N_e = N = 10^4$. Continuous line, equal mutation rates $\mu_{10} = \mu_{01} = 10^{-7}$; dotted line, unequal mutation rates $\mu_{10} = 3 \times 10^{-7}$, $\mu_{01} = 10^{-7}$.

The interaction between nucleotide site diversity and the selection coefficient is complex (Fig. 2). For a given N_e, with no mutation bias, diversity decreases monotonically with increasing selection coefficient.

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

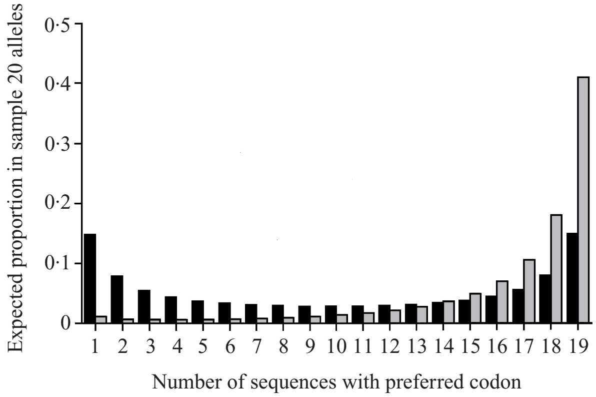

Fig. 3. Expected distribution of the number of preferred codons at sites observed to be segregating in a sample of 20 alleles, for neutral (black) and selected ($|N_e s| = 1$, grey) alleles at equilibrium (from Equation 20).

However, with mutation bias away from the preferred codon type, nucleotide site diversity reaches a maximum at an intermediate selection coefficient. If (1) is used as an approximation for the proportion of sites fixed for the preferred codon, then it can be shown that at equilibrium

$$\pi \approx 4N_e\mu \frac{2\kappa(e^S-1)}{S(\kappa+e^S)}, \qquad (15)$$

where $\mu = \mu_{01}$, $\kappa = \mu_{10}/\mu_{01}$ and $S = 4N_e s$. For $\kappa > 1$ (equivalent to a GC → AT bias of greater than one in *Drosophila*), the maximum diversity occurs at $S = \frac{3}{2}\ln\kappa$ at which, from (1), $m_1 = [1+\kappa^{-1/2}]^{-1}$. Hence, if the mutation rate away from the preferred codon is greater than the mutation rate to it, then for low selection coefficients an increase in s leads to an increase in nucleotide site diversity. Furthermore, the frequency of the preferred codon at which this maximum occurs can be considerably higher than 0·5 (e.g. for $\kappa = 2$, $\pi_{\max}$ occurs when the frequency of preferred codons is 0·59).

(ii) *Frequency distribution of segregating sites*

This model can be used to derive the frequency distribution of alleles at segregating sites at mutation–selection–drift equilibrium. While the frequency distribution of segregating sites under selection has been considered extensively (Fisher, 1930; Wright, 1931, 1938; Sawyer & Hartl, 1992), the equilibrium frequency distribution of segregating, weakly selected sites in the reversible mutation, infinite-sites context has not been explicitly derived (but see Wright, 1931, p. 131, for a related derivation). If sites are exchangeable (Ewens, 1979, p. 77), then, by the principle of ergodicity, the distribution for the infinite sites model can be derived from the expected time novel mutations spend in the frequency interval x to $x+dx$ en route to loss or fixation (Kimura, 1983, p. 239); see Appendix A. From (A 5), the frequency distribution for preferred alleles is given by

$$\phi(x) = C[m_1\mu_{10}\Phi_{10}(1-x)+m_0\mu_{01}\Phi_{01}(x)], \qquad (16)$$

where C is a constant such that

$$\int_{1/2N}^{1-1/2N} \phi(x)\,dx = 1.$$

If the proportion of sites fixed at equilibrium is m_f, then from (3), and assuming genic selection, we have

$$\frac{m_1}{m_f} = \frac{1}{1+(\mu_{10}/\mu_{01})\,e^{-4N_e s}}. \qquad (17)$$

Using the approximation $1-e^{-2s} \approx 2s$ for weak selection, it can be shown that

$$\phi(x) = \frac{C\,e^{4N_e sx}}{x(1-x)}. \qquad (18)$$

Two examples of this distribution are shown in Fig. 3. In the neutral case the distribution is symmetrical; at larger values of s, the distribution is skewed towards a higher frequency of preferred alleles at segregating sites. It is worth noting that (18) is the same as Wright's (1949) distribution of gene frequencies under selection and reversible mutation for the two-allele case, in the limit when mutation rates tend to zero. However, in contrast to Wright's distribution, (18) only considers segregating sites.

Most importantly, mutational biases do not affect the equilibrium frequency distribution. At equilibrium the product of the number of sites fixed for a type and the mutation rate away from that type must be equal to the rate of fixation for that type (Bulmer, 1991). Hence for 2-fold degenerate amino acids, the ratio $m_1\mu_{10}/m_0\mu_{01}$ is simply the ratio of the fixation probabilities and is independent of both the magnitude of the mutation rate and patterns of mutation bias. For the infinite-sites model, the expected time a novel variant spends within a given allele frequency interval is also independent of mutational parameters. Consequently, from (16), the equilibrium frequency distribution for variants at segregating sites is independent of the mutational process.

The analysis of polymorphism provides a potentially powerful method for estimating the strength of selection on codon usage. Previous applications of population genetic theory have considered only the frequency distribution under irreversible mutation (Sawyer & Hartl, 1992; Akashi & Schaeffer, 1997). However, as the equilibrium frequency distribution of segregating sites is independent of mutational parameters under the (infinite-sites) reversible mutation

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

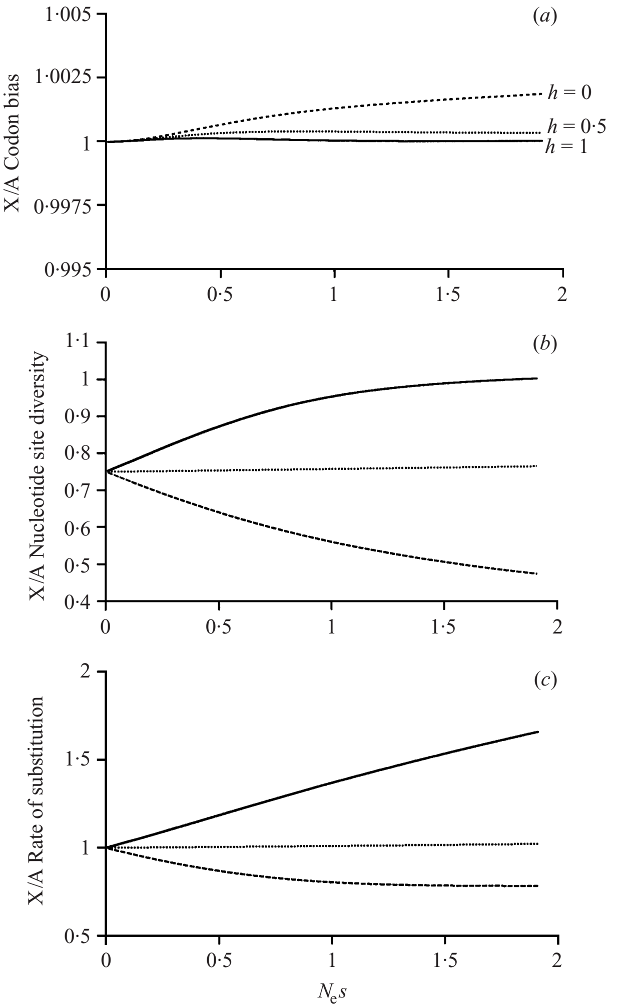

Fig. 4. The ratio of statistics of codon usage for X-linked and autosomal genes under different selection schemes: fully recessive ($h = 0$), genic ($h = 0{\cdot}5$) and fully dominant ($h = 1$). (*a*) Single-sequence bias. (*b*) Nucleotide site diversity. (*c*) Substitution rate. In all cases $N_e = N = 10^4$, $\mu_{10} = \mu_{01} = 10^{-7}$.

model, the hypothesis that selection is discriminating between different types of mutations can easily be tested if putative preferred and unpreferred status can be assigned to segregating alleles and where there is evidence that species are at equilibrium. It is therefore not essential to define the polarity of observed polymorphisms by outgroup methods (Akashi & Schaeffer, 1997), though the extent to which such analyses, with or without outgroup sequences, are sensitive to departures from equilibrium remains to be investigated.

(iii) *X-linked genes*

The model can be used to investigate how sex-chromosome linkage affects patterns of codon usage (Fig. 4). Two effects must be considered: the reduction in the total number of chromosomes and dominance relationships (we have ignored complications arising from unequal breeding numbers of males and females, which might be caused by strong sexual selection). For genic selection ($h = 1/2$) and no difference between the numbers of breeding males and females, patterns of evolution on the X chromosome and autosomes are identical (the same degree of bias and rate of substitution), although the nucleotide site diversity is 75% that on the autosomes for all values of $|N_e s|$. Surprisingly, the expected frequency of the preferred codon in a single sequence picked at random from the population is essentially independent of the degree of dominance of the fitness effects of unpreferred mutations. Hence, in the absence of large differences in the effective population size of males and females, the degree of codon bias on the X chromosome should be similar to that of the autosomes (Fig. 4*a*). In contrast, both patterns of polymorphism and substitution are affected by dominance levels (Fig. 4*b*, *c*). If the fitness effects of mutations to unpreferred codons are recessive, or partially recessive ($h < 1/2$), then the X chromosome has further reduced diversity compared with the autosomes (less than 75%), and a lower rate of substitution, due to the full expression of the trait in hemizygous males. Conversely, dominance of the fitness effects of unpreferred mutations ($h > 1/2$) has the effect of increasing diversity towards the level of the autosomes and increasing the rate of substitution on the X chromosome relative to the autosomes.

(iv) *Predicted effects of background selection*

The effect of simultaneous selection acting at linked sites (interference) can, under certain circumstances, be treated as a parametric transformation. Recurrent deleterious mutations occurring at linked loci, background selection (Charlesworth *et al.*, 1993), are of particular interest, as these may be responsible for the observed reduction in codon bias and nucleotide site diversity in regions of reduced recombination in the *Drosophila melanogaster* genome (Begun & Aquadro, 1992; Kliman & Hey, 1993). The effects of background selection on both fixation probabilities and the expected contribution to heterozygosity for weakly selected, linked variants can be approximated as a single change in N_e (Charlesworth, 1994; Stephan *et al.* 1999). Hence the model presented here can be manipulated to make specific predictions about the effects of background selection on evolution at silent and replacement sites in order to provide tests of the hypothesis.

For a given total population size, Fig. 5 shows the effects of reducing N_e on statistics of codon usage. For $Ns \geqslant 1$ (N is the total population size), a decrease in

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

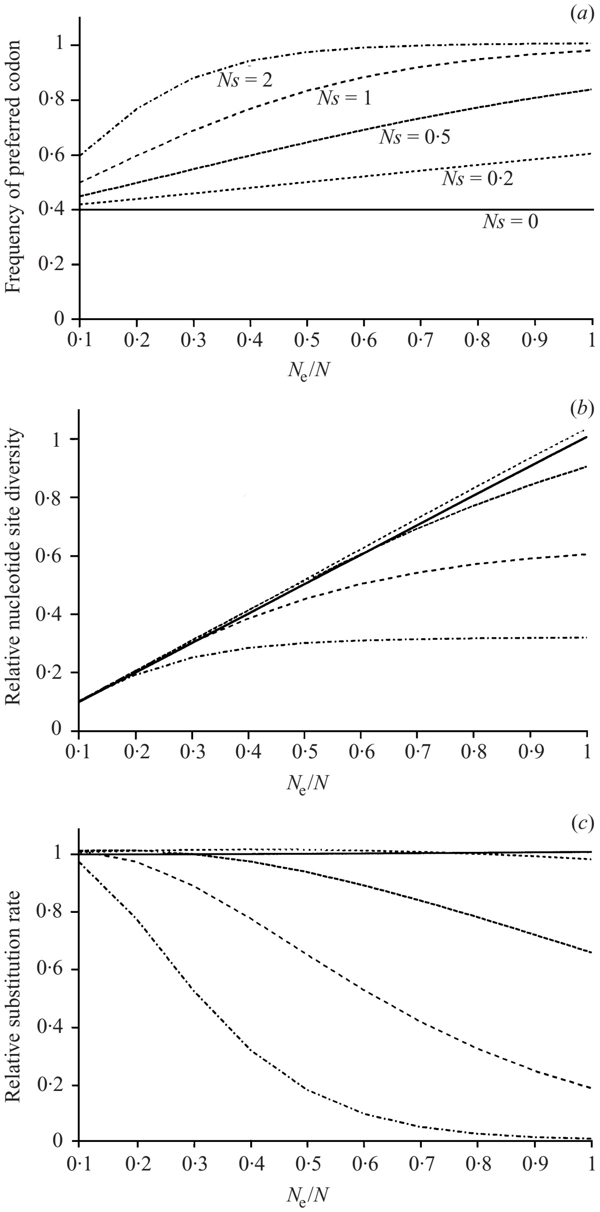

Fig. 5. The effect of variation in the ratio N_e/N (background selection) on (*a*) codon bias, (*b*) nucleotide site diversity (relative to that for a neutral locus when $N_e/N = 1$) and (*c*) rate of substitution (relative to the neutral rate) for different strengths of selection; $Ns = 2$, 1, 0·5 and 0·2. $N = 10^4$, $\mu_{10} = 1{\cdot}5 \times 10^{-7}$, $\mu_{01} = 10^{-7}$.

codon bias is most evident at low values of N_e (Fig. 5*a*). In models of background selection, the local N_e of a given genomic region can be calculated from parameters concerning the mutation rate and distribution of selection coefficients against novel mutations (Charlesworth, 1996), though there is considerable uncertainty as to the value of such parameters (Keightley, 1996). In general, the form of the relationship between the local recombination rate and N_e is such that for genes experiencing strong selection on synonymous codon usage ($Ns \geqslant 1$), a reduction in bias will only be seen in regions of strongly reduced recombination (as observed by Kliman and Hey). For genes under weaker selection, a correlation between recombination rate and bias should be apparent over a greater range of recombination rates. However, when selection on codon usage is very weak, the effects of regional variation in recombination rate will be hard to detect (and are absent for neutral sites). In *D. melanogaster* the effects of reduced recombination on codon bias are only observable across a range of recombination rates for genes less than 250 codons in length, which are also the most biased (Comeron *et al.*, 1999). Longer genes (less biased) only show reduced bias in regions of essentially no recombination. Our preliminary interpretation of these results is that they are contrary to the expectations of the simple background selection model.

The effect of background selection on patterns of diversity depends considerably on the strength of selection on codon usage (Fig. 5*b*). For $Ns < 1$ there is an approximately linear relationship between a regional decrease in N_e and the predicted decrease in nucleotide site diversity. For greater selection coefficients, reduced diversity only becomes apparent for severely reduced N_e (though such genes have low diversities even at $N_e/N \approx 1$). Assuming that the strength of selection acting on replacement polymorphism is greater than that on silent sites (such that replacement polymorphisms reflect sites near mutation–selection equilibrium), these results predict that the ratio of silent to replacement polymorphism will be lower in regions of reduced recombination (see also Charlesworth, 1994).

The effect of changing N_e on substitution rates shows the opposite relationship between strength of selection and magnitude of effect (Fig. 5*c*). For very weakly selected sites ($Ns < 0{\cdot}5$), a change in the effective population size has almost no impact on the rate of substitution, while for stronger selection, a reduced N_e can significantly increase the rate of substitution. These results predict that the rate of synonymous substitution should be higher in regions of low recombination.

(iv) *Multiple codons*

The models presented so far have all considered two types of allele: preferred and unpreferred. This represents codon usage for 2-fold degenerate amino acids. Where more than two codons code for the same amino acid, but the selective difference between the most preferred and least preferred is the same as in the two-codon case, then, for a proportion of new mutations, the selection coefficient discriminating between alternative codons will be lower than in the two-codon case, leading to lower codon bias, higher

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

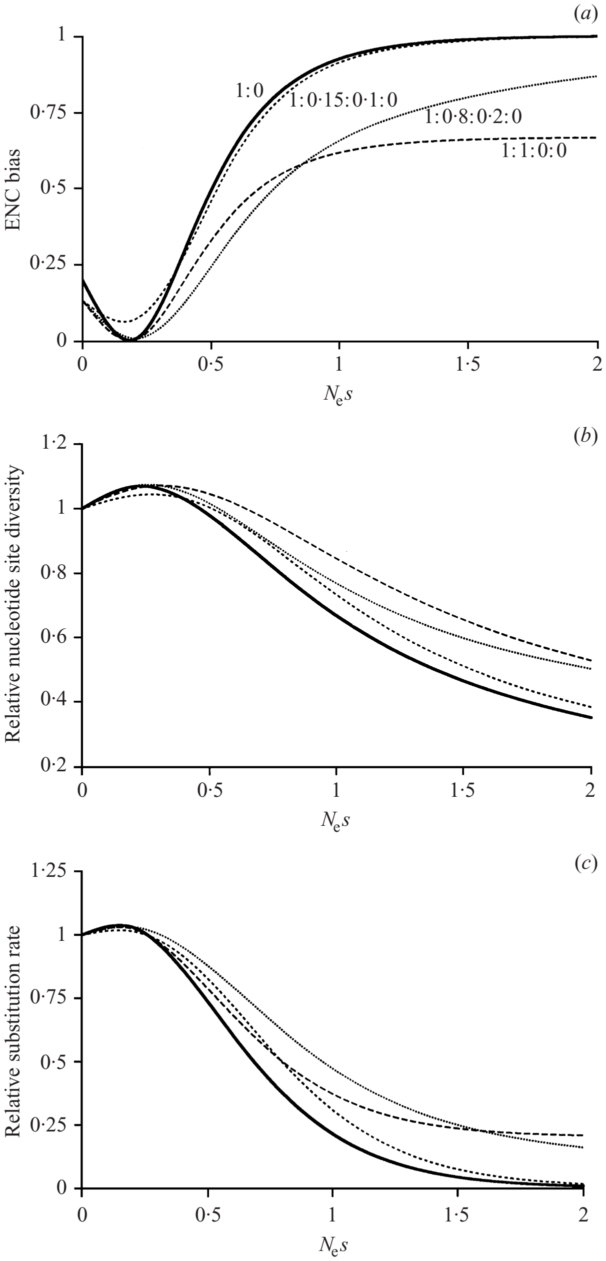

Fig. 6. The difference in patterns of evolution between 2-fold degenerate (continuous line) and 4-fold degenerate amino acids under different selection schemes. (*a*) ENC bias. (*b*) Nucleotide site diversity. (*c*) Substitution rate. The ratio s_1:s_2:s_3:s_4 for each scheme is 1:1:0:0 (long dashes), 1:0·8:0·2:0 (dots) and 1:0·15:0·1:0 (short dashes). $N_e = N = 10^4$, mutation rates represent a general GC → AT mutation pressure of factor 2 and a transition–transversion ratio of 2, where G and C represent the two most favoured codons; $\mu_{AT} = 5 \times 10^{-8}$. Results derived from numerical analysis.

nucleotide site diversity and higher rates of substitution (Fig. 6). At the extreme, if there are four codons in two equivalence classes, then mutations within any one class are neutral. When there is a single preferred codon, with others low and approximately equal selective value, there is little difference between the predicted statistics for 2-fold and 4-fold degenerate amino acids.

These results have implications for the analysis of codon usage at greater than 2-fold degenerate amino acids. When there are four synonymous codons (or more), the frequency of certain suboptimal codons can reach a maximum at intermediate values of $|N_e s|$, if they are also disfavoured by mutation bias (data not shown). Such codons may not be identified as preferred by correlations between the degree of bias in the gene and the use of particular codons (see, e.g., Moriyama & Powell, 1997; Powell & Moriyama, 1997). In addition, when mutation bias is such that the mutation rate away from the preferred codon is greater than that to it, for low values of $|N_e s|$ the degree of bias, as measured by ENC, decreases with increasing s (Fig. 6*a*). This will confuse the identification of preferred codons by correlations between ENC-X (the ENC of a particular amino acid) and the usage of individual codons (Moriyama & Powell, 1997). Correlations between the use of individual codons and the ENC of the gene as a whole are likely to be less sensitive to such error. Preferable methods, which employ explicit population genetics models of synonymous codon usage, are currently being developed (McVean & Vieira, 1999).

Two results should be useful in examining patterns of codon bias at more than 2-fold degenerate amino acids. First, whatever the distribution of selection coefficients between alternative codons, numerical analysis (data not shown) shows that the relative use of any two codons closely approximates the 2-fold degenerate pattern:

$$\frac{m_i}{m_i+m_j} \approx \left[1+\frac{\mu_{ji}}{\mu_{ij}}\,\mathrm{e}^{-4N_e[s_i-s_j]}\right]^{-1}. \qquad (19)$$

Secondly, for 4-fold degenerate codons, where there is a general GC → AT mutation bias of $\mu_{GC\rightarrow AT}/\mu_{AT\rightarrow GC} = \kappa_{GC}$, then it can be shown (Appendix B) that the combined use of G- and C-ending codons is accurately approximated by the relationship

$$m_{GC} \approx [1+\kappa_{GC}\,\mathrm{e}^{-2N_e s[1+x_1-x_2]}]^{-1}, \qquad (20)$$

where s is the selective difference between C- and T-ending codons, $s(1-x_1)$ is the fitness difference between C- and G-ending codons and $s(1-x_2)$ is the difference between C- and A-ending codons. Each of the parameters can be estimated by comparing the use of different codons across genes with different degrees of overall codon bias.

(v) *Within-gene variation in the selection coefficient*

The effects of variation in the selection coefficient across sites within individual genes are shown in Fig.

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

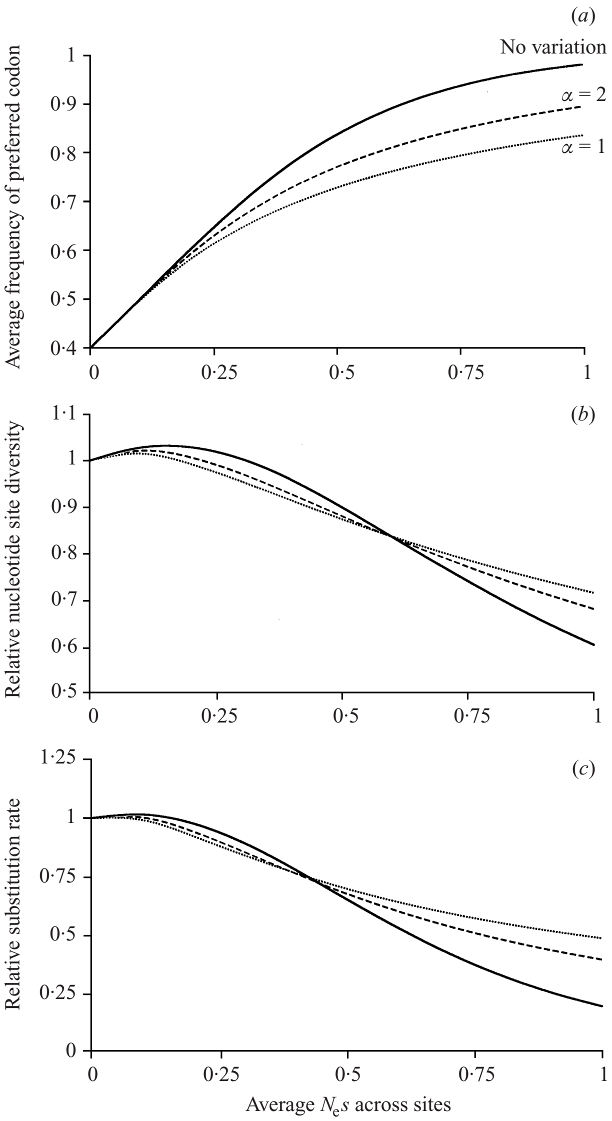

Fig. 7. The effect of variation in selection coefficient across sites within a gene. (*a*) Frequency of preferred codon. (*b*) Nucleotide site diversity. (*c*) Substitution rate. Distributions represented are: no variation (bold line), negative exponential distribution ($\alpha = 1$; dotted line) and gamma distribution with $\alpha = 2$ (dashed line). $N_e = N = 10^4$, $\mu_{10} = 1{\cdot}5 \times 10^{-7}$, $\mu_{01} = 10^{-7}$.

7 for two different models. For a given mean selection coefficient across sites within a gene, variation in the selection coefficient decreases the frequency of the preferred codon, relative to the case of no variation, for all values of $|N_e s|$. Variation also reduces both the relative diversity and substitution rate at low values of $|N_e s|$, but increases them at higher values. For all statistics, the magnitude of effect is greater for the exponential distribution than the gamma distribution with $\alpha = 2$, since a fraction of sites remain neutral. In short, between-sites variation in the selection coefficient reduces the effect of between-gene variation in selection coefficient.

We can also derive an approximate expression for the relationship between the expected frequency of the preferred codon type and the average and variance of the selection coefficient between sites within a gene, for an arbitrary, unimodal and unskewed distribution. Using the delta technique (Bulmer, 1979, p. 78), if the average value of $|4N_e s|$ across sites within a gene is $\bar{S}$, the variance in this quantity is σ_S^2, and the relationship between the selection coefficient and expected frequency of the preferred codon (Equation 1) is written as $\phi(S)$, it follows that the expected frequency of the preferred codon in a gene is approximately

$$E(x) \approx \phi(\bar{S}) + \tfrac{1}{2}\sigma_S^2 \phi''(\bar{S})$$

$$= \frac{1}{1+\kappa\, e^{-\bar{S}}}[1 - C\sigma_S^2(e^{\bar{S}} - \kappa)], \qquad (21)$$

where κ is as in (15) and $C = \tfrac{1}{2}\kappa(e^{\bar{S}} + \kappa)^{-2}$. Therefore, the effect of variation in the selection coefficient between sites within a gene is to decrease the average frequency of the preferred codon type relative to that predicted by the average selection coefficient, when $\bar{S} > \ln \kappa$. Numerical analysis demonstrates that this approximation works well for a range of distributions, even those which are moderately skewed (e.g. gamma distributions with the shape parameter $\alpha \geqslant 2$).

4. Discussion

For sites which are thought to be under no selective constraint, the neutral theory (Kimura, 1983) provides a series of predictions against which to test observed data. In contrast, there exists no comparable body of theory to predict patterns of evolution at sites thought to be experiencing weak selection. The model presented here is an attempt to provide a theoretical framework from which to explore patterns of evolution at weakly selected sites. Such a model can be used to predict the expected relationships between various statistics relating to synonymous codon usage, such as the frequencies of alternative codons, patterns of polymorphism and rates of substitution. By comparing the observed relationships between such statistics and the model's predictions we can ask whether empirical data are compatible with particular models. Furthermore, simple models can be expanded to incorporate heterogeneity in the parameters determining codon usage, in order to examine how expected relationships are affected. In the following sections we discuss four important points raised by this work.

(i) *The dominance level of fitness effects acting on mutations to unpreferred codons*

Data on the relative patterns of silent-site variation and evolution on the X chromosome and autosomes

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

can be used to assess the dominance level of selection acting on codon usage. In several *Drosophila* species, the X chromosome has a reduced level of nucleotide site diversity compared with the autosomes (Moriyama & Powell, 1996). Diversity at silent sites on the X chromosome of *D. melanogaster* is 75% that of the autosomes, while it is much further reduced (46% at silent coding sites) in *D. simulans*. In addition, the ratios of silent to replacement polymorphisms on the X chromosome are approximately equal in the two species, but the ratio is much greater in *D. simulans* for autosomal genes. However, the degree of synonymous site divergence between *D. melanogaster* and *D. simulans* X-linked genes is indistinguishable from that of autosomal ones (Bauer & Aquadro, 1997) and there is no detectable difference in the degree of codon bias on the X chromosome and the autosomes in *D. melanogaster* (excluding regions of low recombination: average ENC autosomes ($\pm$SEM) = 45·1 $\pm$ 0·4, X chromosome = 44·8 $\pm$ 0·8; data in Kliman & Hey, 1993).

The fact that the silent diversity for the X chromosome is less than 75% that of the autosomes in *D. simulans* could be interpreted as evidence for recessivity (or partial recessivity) of the fitness effects of mutations leading to unpreferred codons, if, as is suspected, the effective population size of *D. simulans* is much larger than that of *D. melanogaster* (Moriyama & Powell, 1996), and silent sites in *D. melanogaster* are effectively neutral (Akashi, 1996). Recessivity of unpreferred mutations is also supported by the finding that the absolute levels of diversity on the *D. simulans* and *D. melanogaster* X chromosomes are more similar than the autosome levels ($\pi_{\text{Xsim}}/\pi_{\text{Xmel}} = 1{\cdot}71$, $\pi_{\text{Asim}}/\pi_{\text{Amel}} = 2{\cdot}91$; Moriyama & Powell, 1996). For a given selection coefficient, an increase in N_e generates greater diversity, but if deleterious mutations are recessive, such an increase will be less for X chromosomes than autosomes (Fig. 4*b*). Consider a numerical example (with the caveat that, while these values are derived assuming that both species are at equilibrium, there is evidence that this is not true for *D. melanogaster*; Akashi, 1996). Assuming that unpreferred mutations are fully recessive, $N_e s$ in *D. melanogaster* is 0·4 and N_e in *D. simulans* is 5-fold greater (s is the same), we predict (for 2-fold degenerate codons) that $\pi_{\text{Xsim}}/\pi_{\text{Xmel}} = 1{\cdot}88$, $\pi_{\text{Asim}}/\pi_{\text{Amel}} = 2{\cdot}60$, $\pi_{\text{Xsim}}/\pi_{\text{Asim}} = 0{\cdot}47$ and $\pi_{\text{Xmel}}/\pi_{\text{Amel}} = 0{\cdot}66$.

The data on the ratios of silent to replacement polymorphisms in the two species (Begun, 1996) are also compatible with recessivity of selection against unpreferred codons. If the majority of replacement mutations are strongly deleterious, levels of replacement polymorphism largely reflect sites close to mutation–selection equilibrium, and should be similar in species with different N_e. If such mutations are also recessive ($h < 1/2$), the level of replacement polymorphism on the X chromosome should be less than 75% that on the autosomes (in all species). In contrast, levels of silent-site diversity will be sensitive to variation in N_e. For species where the levels of dominance acting on unpreferred silent and replacement mutations are similar, and, due to large N_e, selection on codon usage is strong, the ratio of silent to replacement polymorphisms on the X chromosome and autosomes should be approximately equal. But in species with a small N_e, such that variants at silent sites are effectively neutral, the ratio of silent to replacement polymorphism should be greater on the X chromosome. When comparing species, we expect a greater ratio of silent to replacement polymorphism in the species with the higher N_e, though this effect should be less marked on the X chromosome. All these predictions are met by the data on *D. melanogaster* and *D. simulans*, the only exception being that there is no detectable difference in the ratio of silent to replacement polymorphism on the X chromosomes between the two species, whereas we expect a higher ratio in *D. simulans* (though lower than for the autosomes). It should be noted that these data represent different numbers of alleles sampled from different genes in the two species (Begun, 1996), hence conclusions are preliminary.

In contrast, the lack of difference in silent substitution rates between the X chromosome and autosomes might be taken to argue for genic selection on codon usage (we expect a lower rate on the X chromosome if unpreferred mutations are recessive). However, *D. melanogaster* is evolving reduced codon bias compared with *D. simulans*, which appears to be at mutation–selection–drift equilibrium (Akashi, 1996). Hence a large proportion of the substitutions that have accumulated between the two species are due to the fixation of effectively neutral 'unpreferred' codons in the *D. melanogaster* lineage, on the approach to a new equilibrium. Any difference in the rate of substitution between the X chromosome and autosomes of codons currently experiencing selection will be masked by the excess of neutral substitutions in the *D. melanogaster* lineage. Finally, a lack of difference in the degree of codon bias between the X chromosome and autosomes in *D. melanogaster* is predicted by all levels of dominance against unpreferred mutations (see Fig. 4*a*).

The alternative explanation to recessivity for the data is that the *D. simulans* X chromosome has suffered a reduced N_e relative to the autosomes. Possible reasons for a lower N_e of the X chromosome are obscure, since genetic forces which reduce N_e, such as background selection, should be weaker on the X chromosome than on the autosomes, as it has a higher effective recombination rate. In *D. melanogaster* the ratio of map length (centimorgans) to physical length of euchromatin (Mb), corrected for sex-limited cross-

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

ing over frequencies, is 1·94 for the X chromosome and 1·27 for the combined autosomes (excluding chromosome 4: data from Heino *et al.*, 1994). Hitch-hiking effects should also be weaker for the same reason, although this might be counteracted by the greater chance of fixation of advantageous recessive mutations on the X chromosome (Aquadro *et al.*, 1994).

(ii) *Relationship between strength of selection on codon usage and patterns of polymorphism*

It is generally assumed that there is an inverse relationship between the strength of selection on codon usage and the expected nucleotide site diversity (Moriyama & Powell, 1996). However, we find that if selection and mutation bias are acting in opposite directions, then for low values of $N_e s$, an increase in the selection coefficient actually leads to an increase in the expected diversity (Fig. 2). In *Drosophila*, preferred codons typically have G or C at the third position (Shields *et al.*, 1988; Akashi, 1995), while the general bias is towards A and T in low bias genes and introns (Shields *et al.*, 1988; Kliman & Hey, 1994). In accordance with the predictions of our model, in *D. melanogaster* there is a significant positive relationship between the degree of codon bias and silent-site diversity, even after controlling for recombination ($P < 0{\cdot}05$; data in Moriyama & Powell, 1996). However, the relationship also applies to non-coding regions ($P < 0{\cdot}01$). If mutations in non-coding regions are strictly neutral, this would suggest that the correlation between silent-site diversity and codon bias is a product of regional genomic factors affecting both diversity and codon bias (e.g. hitch-hiking). However, there is evidence of considerable selection on non-coding regions in *Drosophila*; the overall level of diversity is higher at synonymous sites than in non-coding regions (Moriyama & Powell, 1996), as is the rate of substitution (Li & Graur, 1991; Bauer & Aquadro, 1997). If parameters of selection and mutation bias are similar at silent sites and adjacent non-coding regions (for example, if the strength of selection on both classes of site is a function of gene expression level), a correlation between codon bias and nucleotide diversity at both silent sites and adjacent non-coding regions is predicted.

(iii) *Insensitivity of codon bias to population polymorphism*

We have shown that (1) provides an accurate prediction of the expected frequency of the preferred codon type in a single gene selected at random from the population, even though it was derived without considering polymorphic sites. This greatly simplifies further analysis of patterns of codon bias, but is perhaps surprising. There are two possible explanations. First, the proportion of segregating sites is usually so low that it has little effect on the estimation of codon bias. Secondly, the form of the relationship between the average frequency of the preferred codon and the selection coefficient is the same for both fixed and segregating sites.

The expected frequency of the preferred codon type can be calculated from the sojourn time functions for novel preferred and unpreferred mutations. From (18), the equilibrium frequency distribution for the preferred codon is given by

$$\phi(x) = \frac{C\,\mathrm{e}^{Sx}}{x(1-x)},$$

where $S = 4N_e s$. Numerical analysis shows that the average frequency of the preferred codon is closely approximated by an exponential relationship with S (data not shown). To consider what the form of this may be, we can examine the behaviour of (18) for small selection coefficients. The average frequency of the preferred codon at segregating sites is

$$\bar{x} = \frac{\displaystyle\int_{1/2N}^{1-1/2N} x\phi(x)\,dX}{\displaystyle\int_{1/2N}^{1-1/2N} \phi(x)\,dx} = \frac{1}{1+\mathit{\Lambda}\,\mathrm{e}^{-S}},$$

where

$$\mathit{\Lambda} = -\frac{\displaystyle\int_{-S(1-1/2N)}^{-S/2N} \frac{\mathrm{e}^{-t}}{t}\,dt}{\displaystyle\int_{S/2N}^{S(1-/2N)} \frac{\mathrm{e}^{-t}}{t}\,dt}. \tag{22}$$

Approximating the exponential function as a Taylor expansion around 0 gives for $|S| < 1$

$$\mathit{\Lambda} \approx \frac{\ln(2N)+S}{\ln(2N)-S} \approx 1+\frac{2S}{\ln(2N)}. \tag{23}$$

Given the numerical results suggesting an exponential relationship, (23) further suggests the approximation

$$\bar{x} \approx \frac{1}{1+\mathrm{e}^{-S[1-2/\ln(2N)]}}, \tag{24}$$

which gives values very similar to exact numerical results. For large N, this equation is almost identical to the one which predicts the unconditional frequency of the preferred codon when there is no mutation bias. The relationship between the expected frequency of the preferred codon and the selection coefficient is therefore similar for fixed and segregating sites, which means that (1) provides a very accurate approximation for the expected frequency of the preferred codon in a single sequence picked at random from the population.

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

(iv) *The application of single-site models to empirical data*

Finally, it is worth noting that simple models of synonymous codon usage do not appear to provide a particularly close overall fit to the data. Explicit population genetics models can be used to estimate underlying parameters of codon bias (Bulmer, 1991; Sawyer & Hartl, 1992; Akashi & Schaeffer, 1997; McVean & Vieira, 1999). By analysing the frequency distribution of segregating variants in *D. simulans*, the average selection coefficient acting on codon usage is estimated to be $|N_e s| \approx 1$–2 (Akashi & Schaeffer, 1997). If mutation bias is such that GC content is 40% in the absence of selection, $|N_e s| \approx 1$ predicts that the frequency of preferred codons at 2-fold degenerate amino acids should be 0·97. In contrast, the observed frequency of the preferred codon at 2-fold degenerate amino acids in 29 genes from *D. simulans* is 0·69 (data in Akashi & Schaeffer, 1997). Potential explanations for this discrepancy include non-equilibrium codon usage (but see Akashi, 1996), synergistic selection (Li, 1987), variation in selection coefficients across sites (see Fig. 7), selective constraints such as secondary structure or the effects of interference between linked, weakly selected sites (Hill & Robertson, 1966; Li, 1987). The importance of these complications remains to be investigated.

We would like to thank Philip Awadalla, Jody Hey, Laurence Hurst, Molly Przeworski and Mikkel Schierup for discussion and comments on the manuscript, and two anonymous referees for additional comments. G.M. is funded by the NERC; B.C. is a Royal Society Research Professor.

Appendix A. Formulae for predicting statistics of synonymous codon usage

If the mean change in allele frequency (x) per generation is $M_{\delta x}$ and the variance in change in allele frequency per generation is $V_{\delta x}$, then the probability of fixation of an allele at initial frequency p is (Kimura, 1962)

$$u(p) = \frac{\int_0^p G(x)\,dx}{\int_0^1 G(x)\,dx} \quad \text{where } G(x) = \exp\left\{-\int \frac{2M_{\delta x}}{V_{\delta x}}\right\}. \tag{A 1}$$

The average time to loss or fixation of a single novel mutant allele is (Kimura & Ohta, 1969*b*)

$$\bar{t}(p) = u(p)\int_p^1 \psi(\xi)[1-u(\xi)]\,d\xi + [1-u(p)]\int_0^p \psi(\xi)\,u(\xi)\,d\xi, \tag{A 2}$$

where

$$\psi(\xi) = 2\int_0^1 \frac{G(x)\,dx}{V_{\delta\xi}\,G(\xi)}.$$

If sites are exchangeable (Ewens, 1979, p. 77), then, by the principle of ergodicity, the stationary distribution of allele frequencies at polymorphic sites and the expected nucleotide site diversity can be calculated from the expected time a novel mutation spends in the frequency range x to $x+dx$, where $dx = 1/2N$ (Kimura, 1983, p. 239). The sojourn time density for an allele en route to either loss of fixation is (Kimura, 1983, p. 228)

$$\Phi(p,x) = \frac{2u(p)[1-u(x)]}{V_{\delta x}\,u'(x)}, \quad \text{for } x \geqslant p, \tag{A 3}$$

where

$$u'(x) = \frac{d[u(x)]}{dx}, \quad \text{i.e. } u'(x) = \frac{G(x)}{\int_0^1 G(x)\,dx}.$$

If the numbers of breeding males and females are equal, this can be written as

$$\Phi(p,x) = 2\alpha N_e\,u(p)\,\frac{\int_x^1 G(y)\,dy}{x(1-x)\,G(x)}, \tag{A 4}$$

where α is the relationship between the number of diploid adults and alleles at the locus of interest (for an autosomal locus $\alpha = 2$, for an X-linked locus $\alpha = 3/2$). If x is the frequency of the preferred allele, $\Phi_{10}(1-x)$ is the sojourn time for mutations at sites initially fixed for the preferred type, and $\Phi_{01}(x)$ is the sojourn time for mutations at sites initially fixed for the unpreferred type. The expected proportion of polymorphic sites where the preferred allele is in the frequency interval x to $x+dx$ is

$$\phi(x) = C[m_1\,\mu_{10}\,\Phi_{10}(1-x) + m_0\,\mu_{01}\,\Phi_{01}(x)], \tag{A 5}$$

where C is a normalizing factor such that

$$\int_{1/2N}^{1-1/2N} \phi(x)\,dx = 1.$$

The expected contribution to heterozygosity for an autosomal allele with initial frequency p is

$$H(p) = \int_0^1 2x(1-x)\,\Phi(x)\,dx = 4\alpha N_e\,u(p) \times \int_0^1 \left\{\frac{\int_x^1 G(y)\,dy}{G(x)}\right\} dx. \tag{A 6}$$

For an autosomal locus, the mean change in allele frequency per generation is approximately

$$M_{\delta x} = 2sx(1-x)[h + x(1-2h)], \tag{A 7a}$$

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

where the difference in fitness between the homozygotes is $2s$ and h is the degree of dominance. The variance in gene frequency change per generation is

$$V_{\delta x} = \frac{x(1-x)}{2N_e}. \tag{A 7b}$$

For an X-linked gene, the approximate mean and variance in change of allele frequency per generation are respectively

$$M_{\delta x} = \tfrac{2}{3} sx(1-x)[1-2\{h+x(1-2h)],$$
$$V_{\delta x} = \frac{2x(1-x)}{3N_e}. \tag{A 8}$$

The initial frequency for a single novel mutation (hence also the probability of fixation in the neutral case) is $2/3N$ for X-linked genes (assuming a sex ratio of unity) and $1/2N$ for autosomal genes.

Appendix B. Derivation of Equation (20)

For a 4-fold degenerate amino acid, the selective value of alternative codons (with genic selection) can be represented as

$$s_1 = 1+s, \quad s_2 = 1+x_1 s,$$
$$s_3 = 1+x_2 s, \quad s_4 = 1, \tag{A 9}$$

where codons 1–4 respectively represent C, G, A and T at the silent site. Represent the mutation parameters as a single GC→AT bias of κ_{GC} and a transition–transversion ratio of α. Using the same logic about the flux of substitutions at equilibrium as in the derivation of (2), and assuming that $\Sigma_i m_i = 1$, it follows that

$$m_{GC} = \frac{1}{1+\kappa_{GC}\dfrac{(u_{43}+u_{34})}{(u_{21}+u_{12})}\cdot\dfrac{u_{21}(u_{13}+\alpha u_{14})+u_{12}(\alpha u_{23}+u_{24})}{u_{43}(u_{31}+\alpha u_{32})+u_{34}(\alpha u_{41}+u_{42})}}. \tag{A 10}$$

Numerical analysis of (A 10) shows that this closely follows an exponential relationship in s (data not shown). To obtain a suitable approximation for what this may be, consider small values of s. Using a Taylor approximation around $s = 0$ for the exponential function, the approximations

$$u_{ij} = \frac{2(s_j-s_i)}{1-e^{-4N_e(s_j-s_i)}} \quad \text{and} \quad \frac{1}{1+x} \approx 1-x \quad (\text{for } x \ll 1)$$

and discarding terms of order s^2, (A 10) approximates to

$$m_{GC} = \frac{1}{1+\kappa_{GC}\{1-2N_e s(1+x_1-x_2)\}}. \tag{A 11}$$

Given the numerical results suggesting an exponential relationship, (A 11) suggests the form

$$m_{GC} = \frac{1}{1+\kappa_{GC}\, e^{-2N_e s[1+x_1-x_2]}}. \tag{A 12}$$

This agrees well with exact numerical results for $S \ll 1$. If mutational parameters are more complex, such that the two major codons (C- and G-ending) have different mutational biases (respectively, κ_C and κ_G), then the relative use of C in C- and G-ending codons is approximately

$$\frac{m_C}{m_C+m_G} = \frac{1}{1+(\kappa_C/\kappa_G)\, e^{-4N_e s(1-x_1)}} \tag{A 13}$$

and the combined frequency of C- and G-ending codons is approximately

$$m_{GC} = \frac{1}{1+(2\kappa_C\kappa_G/[\kappa_C+\kappa_G])\, e^{-4N_e s[1+x_1-x_2]}}. \tag{A 14}$$

References

Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**, 927–935.

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*. *Genetic* **144**, 1297–1307.

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

Aquadro, C. F., Begun, D. J. & Kindahl, E. C. (1994). Selection, recombination and DNA polymorphism in *Drosophila*. In *Non-Neutral Evolution: Theories and Molecular Data* (ed. B. Golding), pp. 46–56. New York: Chapman and Hall.

Bauer, V. L. & Aquadro, C. F. (1997). Rates of DNA sequence evolution are not sex biased in *Drosophila melanogaster* and *D. simulans*. *Molecular Biology and Evolution* **14**, 1252–1257.

Begun, D. J. (1996). Population genetics of silent and replacement variation in *Drosophila simulans* and *D. melanogaster*: X/autosome differences? *Molecular Biology and Evolution* **13**, 1405–1407.

Begun, D. J. & Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520.

Bulmer, M. G. (1979). *Principles of Statistics*. New York: Dover.

Bulmer, M. G. (1991). The selection–mutation–drift theory of synonymous codon usage. *Genetics* **129**, 897–907.

Charlesworth, B. (1994). The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research* **63**, 213–227.

Charlesworth, B. (1996). Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genetical Research* **68**, 131–149.

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press

Charlesworth, B., Morgan, M. T. & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.

Comeron, J., Kreitman, M. & Aguadé, M. (1999). Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. *Genetics* **151**, 239–249.

Ewens, W. J. (1979). *Mathematical Population Genetics*. Berlin: Springer-Verlag.

Eyre-Walker, A. (1991). An analysis of codon usage in mammals: selection or mutation bias? *Journal of Molecular Evolution* **33**, 442–449.

Eyre-Walker, A. (1996). Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Molecular Biology and Evolution* **13**, 864–872.

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

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Oxford University Press.

Gouy, M. & Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research* **10**, 7055–7074.

Heino, T. I., Saura, A. O. & Sorsa, V. (1994). Maps of the salivary gland chromosomes of *Drosophila melanogaster*. In *Drosophila Information Service 73*, part B, pp. 619–738.

Hill, W. G. & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294.

Hudson, R. R., Kreitman, M. & Aguadé, M. (1987). A test for neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology* **151**, 389–409.

Ikemura, T. (1985). Codon usage and transfer-RNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2**, 13–34.

Keightley, P. (1996). The distribution of mutation effects on viability in *Drosophila melanogaster*. *Genetics* **138**, 1–8.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719.

Kimura, M. (1971). Theoretical foundation of population genetics at the molecular level. *Theoretical Population Biology* **2**, 174–208.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Kimura, M. & Ohta, T. (1969*a*). The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics* **63**, 701–709.

Kimura, M. & Ohta, T. (1969*b*). The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61**, 763–771.

Kliman, R. M. & Hey, J. (1993). Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Molecular Biology and Evolution* **10**, 1239–1258.

Kliman, R. M. & Hey, J. (1994). The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. *Genetics* **137**, 1049–1056.

Kondrashov, A. (1995). Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *Journal of Theoretical Biology* **175**, 583–594.

Li, W.-H. (1987). Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *Journal of Molecular Evolution* **24**, 337–345.

Li, W.-H. & Graur, D. (1991). *Fundamentals of Molecular Evolution*. Sunderland, Mass.: Sinauer Associates.

McDonald, J. H. & Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654.

McVean, G. A. T. & Vieira, J. (1999). The evolution of codon preferences in *Drosophila*. *Journal of Molecular Evolution*, in press.

Moriyama, E. N. & Powell, J. R. (1996). Intraspecific nuclear DNA variation in *Drosophila*. *Molecular Biology and Evolution* **13**, 261–277.

Moriyama, E. N. & Powell, J. R. (1997). Codon usage bias and tRNA abundance in *Drosophila*. *Journal of Molecular Evolution* **45**, 514–523.

Powell, J. R. & Moriyama, E. N. (1997). Evolution of codon usage bias in *Drosophila*. *Proceedings of the National Academy of Sciences of the USA* **94**, 7784–7790.

Sawyer, S. A. & Hartl, D. L. (1992). Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176.

Sharp, P. M. & Cowe, E. (1991). Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* **7**, 657–678.

Sharp, P. M. & Li, W.-H. (1987). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Molecular Biology and Evolution* **4**, 222–230.

Sharp, P. M., Tuohy, T. M. E. & Mosurski, K. R. (1986). Codon usage in yeast: cluster-analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* **14**, 5125–5143.

Shields, D. C., Sharp, P. M., Higgins, D. G. & Wright, F. (1988). 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution* **5**, 704–716.

Sørensen, M. A. & Pedersen, S. (1991). Absolute *in vivo* translation rates of individual codons in *Escherichia coli*: the 2 glutamic-acid codons GAA and GAG are translated with a threefold difference in rate. *Journal of Molecular Biology* **222**, 265–280.

Stephan, W., Charlesworth, B. & McVean, G. (1999). The effect of background selection at a single locus on weakly selected, linked variants. *Genetical Research* **73**, 133–146.

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene* **87**, 23–29.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1938). The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences of the USA* **24**, 253–259.

Wright, S. (1949). Adaptation and selection. In *Genetics, Palaeontology and Evolution* (ed. G. Jepson, G. Simpson & E. Mayr), pp. 365–391. Princeton: Princeton University Press.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution* **11**, 367–372.

https://doi.org/10.1017/S0016672399003912 Published online by Cambridge University Press