Letter to the Editor

Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias

John A. Novembre

Department of Integrative Biology, University of California—Berkeley

The phenomenon of codon usage bias has been important in the study of evolution because it provides examples of weak selection working at the molecular level. During the last two decades, evidence has accumulated that some examples of codon usage bias are driven by selection, particularly for species of fungi (e.g., Bennetzen and Hall 1982; Ikemura 1985), bacteria (e.g., Ikemura 1981; Sharp and Li 1987), and insects (e.g., Akashi 1997; Moriyama and Powell 1997). This connection between codon usage bias and selection has been important in stimulating the development of alternatives to strictly neutral theories of molecular evolution (Ohta and Gillespie 1996; Kreitman and Antezana 2000). Uncertainty persists, however, regarding the phylogenetic distribution of codon usage bias (such as whether selection-based codon usage bias is present in mammals; e.g., Karlin and Mrazek 1996; Iida and Akashi 2000; Sueoka and Kawanishi 2000; Smith and Eyre-Walker 2001*a;* Urrutia and Hurst 2001). Questions also remain as to what models of selection underlie codon preferences (Kreitman and Antezana 2000), and specifically whether the presence of suboptimal codons is the result of mutation and drift, variation in selection pressure across sites, or antagonistic selection pressures (Smith and Eyre-Walker 2001*b*).

For investigating certain questions regarding the evolution of codon usage bias, it is useful to have a summary statistic describing the pattern of codon usage across all amino acids. Many summary statistics have already been developed to describe the patterns of codon usage. They can be divided roughly into two classes (Comeron and Aguade 1998). One class summarizes the usage of certain preferred codons, and the other compares every codon's usage to a null distribution (typically uniform usage of synonymous codons). The former class of methods has the disadvantage that it requires a prior knowledge of the preferred codons. With summary statistics one can explore general patterns, such as the relationship of codon usage bias to recombination rate, gene length, or synonymous substitution rate. Observing broad patterns such as these has already provided insight into the evolutionary dynamics of codon usage bias (Kliman and Hey 1993; Akashi and Eyre-Walker 1998; Kreitman and Comeron 1999).

Abbreviations: ENC, effective number of codons.

Key words: codon usage bias, background nucleotide composition, mutation bias.

Address for correspondence and reprints: John A. Novembre, Department of Integrative Biology, University of California—Berkeley, VLSB 3060, Berkeley, California 94720. E-mail: novembre@socrates.berkeley.edu.

Mol. Biol. Evol. 19(8):1390–1394. 2002

Of the summary statistics that do not require knowledge of preferred codons, the effective number of codons (ENC, or $\hat{N}_c$ as proposed by Wright 1990) has been found to be the most statistically well behaved, in that it is the least affected by short gene lengths (Comeron and Aguade 1998). $\hat{N}_c$ is inversely proportional to the extent of nonuniform codon usage. It takes the value of 61 when all codons are being used with equal frequency, and its value decreases as codon usage becomes less uniform. It is intuitively accessible because its value is intended to correspond to the number of codons being used in a sequence.

One limitation of $\hat{N}_c$ is that measuring the departures from uniform codon usage is not always desirable. Knowledge of background nucleotide composition patterns may suggest a null distribution of codon usage that is nonuniform. For instance, in *Drosophila,* where the background nucleotide composition is 60% AT, one might like to measure how far codon usage departs from 60% usage of AT-ending codons. Indeed, accounting for background nucleotide composition when studying codon usage has been recognized as being important (Akashi, Kliman, and Eyre-Walker 1998; Marais, Mouchiroud, and Duret 2001; Urrutia and Hurst 2001). Taking composition into account is particularly important for phylogenetic studies of codon usage bias where comparisons are made among species with differing background nucleotide compositions or for studies of codon usage bias in genomic regions that may differ in background nucleotide composition. If not taken into consideration, differences in background nucleotide composition might lead one to conclude that differences in codon preference exist when they do not.

The departure of $\hat{N}_c$ from 61 because of variation in background nucleotide composition was recognized by Wright and described as the following relationship between the third position GC content (f_{GC}) and $\hat{N}_c$:

$$\hat{N}_c = 2 + f_{GC} + \left(\frac{29}{f_{GC}^2 + (1 - f_{GC})^2} \right).$$

Using this relationship one can plot the expected value of $\hat{N}_c$ versus the GC content. Such plots are known as $\hat{N}_c$-plots (e.g., see fig. 2 in Wright 1990). Points that fall below the bell-shaped line suggest deviations from a null model of no-codon preferences. Unfortunately, studying such patterns does not have a clear statistical basis and detracts from the value of $\hat{N}_c$ as a quantitative summary statistic.

Other codon usage bias summary statistics such as Akashi's scaled χ^2 (Akashi 1995), the maximum-likelihood codon bias statistic (MCB; Urrutia and Hurst

2001), and $B^*(a)$ (Karlin and Mrazek 1996) account for background nucleotide composition. However, as presented below, these statistics are affected strongly by the sequence lengths being studied. Clearly, it would be useful to have a statistic similar to $\hat{N}_c$ that has low sequence-length effects but that relaxes the assumption of equal usage of all synonymous codons.

Here, a statistic ($\hat{N}_c'$) is presented that accounts for background nucleotide composition and is minimally affected by sequence length. The statistic is based on Pearson's X^2 statistics and describes the departure of the observed codon usage from some expected distribution. The expected distribution can be derived from knowledge of the background nucleotide composition. When uniform usage of codons is expected, $\hat{N}_c'$ reduces to $\hat{N}_c$. Like $\hat{N}_c$, $\hat{N}_c'$ is relatively insensitive to sequence length. The properties of $\hat{N}_c'$ are tested on simulated sequences, and compared with related summary statistics. The results suggest that $\hat{N}_c'$ is useful for studying codon usage among sequences that vary in background nucleotide composition.

To understand the construction of $\hat{N}_c'$, it is first necessary to provide background on $\hat{N}_c$. $\hat{N}_c$, or the ENC (Wright 1990), is derived by making an analogy with the effective number of alleles n_e at a locus (Kimura and Crow 1964). An estimate of n_e is $\hat{n}_e = 1/\hat{F}$, where $\hat{F}$ is an estimate of the expected heterozygosity at the locus. For quantifying codon usage bias at an amino acid a, Wright defined an estimate of the homozygosity of codon usage as follows:

$$\hat{F}_a = \left(n_a \sum_{i=1}^{k} p_i^2 - 1\right) \Big/ (n_a - 1) \qquad (1)$$

Here, p_i is the frequency of the ith codon, k is the number of synonymous codons for the amino acid of interest, and n_a is the observed number of codons for the amino acid. The average of the $\hat{F}_a$ for each r-fold redundancy class (e.g., onefold, twofold, threefold, fourfold, sixfold) is then computed:

$$\bar{\hat{F}}_r = \frac{1}{n_{RC}} \sum_{a \in RC} \hat{F}_a \qquad (2)$$

where n_{RC} is the number of amino acids in the RC redundancy class. Finally, $\hat{N}_c$ is computed as follows:

$$\hat{N}_c = 2 + (9/\bar{\hat{F}}_2) + (1/\bar{\hat{F}}_3) + (5/\bar{\hat{F}}_5) + (3/\bar{\hat{F}}_6). \qquad (3)$$

The calculation of $\hat{N}_c$ can also be expressed more generally as

$$\hat{N}_c = \sum_{r \in ARC} n_r/\bar{\hat{F}}_r \qquad (4)$$

where ARC is the set of all redundancy classes, and n_r is the number of amino acids in the redundancy class. For any particular amino acid we can also compute an $\hat{N}_c$ value that is simply $1/\hat{F}$, where $\hat{F}$ is the homozygosity of codon usage for that particular amino acid.

To derive $\hat{N}_c'$ and show its relationship to $\hat{N}_c$, I use Pearson's X^2 statistics to quantify the departure of each codon's usage (p_i) from some expected usage (e_i) for each amino acid. The expected usage of codons can be calculated in a number of ways, including using mono-, di-, or trinucleotide frequencies. The χ^2 statistic for each amino acid, X_a^2, is calculated as

$$X_a^2 = \sum_{i=1}^{k} \frac{n_a(p_i - e_i)^2}{e_i}.$$

Using the X_a^2 values, $\hat{F}_a'$ is defined as follows:

$$\hat{F}_a' = \frac{X_a^2 + n_a - k}{k(n_a - 1)}. \qquad (5)$$

The calculations from here on mirror those of $\hat{N}_c$. The $\hat{F}_a'$ values are averaged for each redundancy class ($\bar{\hat{F}}_r'$) and used to calculate $\hat{N}_c'$:

$$\hat{N}_c' = \sum_{r \in ARC} n_r/\bar{\hat{F}}_r'. \qquad (6)$$

Note that because X_a^2 is χ^2 distributed, the expected value of $\hat{F}'$ will be $1/k$ when the codon usage matches the expected usage pattern. Using the δ-method, it can also be shown that the expected value of $\hat{N}_c'$ will be equal to 61 for the standard genetic code.

Notably, $\hat{N}_c'$ simplifies to $\hat{N}_c$ when synonymous codons are expected to be in equal frequency. In this case the e_i are all $1/k$, and the X_a^2 statistic for each amino acid reduces to

$$X_a^2 = n_a k \sum_{i=1}^{k} p_i^2 - n_a. \qquad (7)$$

Substitution of this X_a^2 into the equation for $\hat{F}_a'$ (eq. 5) produces

$$\hat{F}_a' = \left(n_a \sum_{i=1}^{k} p_i^2 - 1\right) \Big/ (n_a - 1). \qquad (8)$$

This value of $\hat{F}_a'$ is equal to the $\hat{F}_a$ used by Wright in the calculation of $\hat{N}_c$ (eq. 1). Because the calculation of $\hat{N}_c'$ matches that of $\hat{N}_c$ once $\hat{F}_a'$ is obtained, $\hat{N}_c'$ is equivalent to $\hat{N}_c$ when uniform usage is expected.

This derivation shows that $\hat{N}_c'$ is a generalization of $\hat{N}_c$ to the case in which expected codon usage is not uniform. Whereas $\hat{N}_c$ is implicitly based on an expected distribution of uniform usage, $\hat{N}_c'$ can have the expected proportions of synonymous codon usage given by the background nucleotide composition. $\hat{N}_c'$ will decrease from 61 when the codon usage does not match the distribution predicted by nucleotide composition. In contrast, $\hat{N}_c$ will decrease from 61 whenever the codon usage departs from uniform usage of each synonymous codon. The result is that the value of $\hat{N}_c'$ should be less dependent on nucleotide composition and more dependent on the codon preferences that go beyond those caused by unequal nucleotide composition.

In the practical implementation of both $\hat{N}_c$ and $\hat{N}_c'$, care must be taken to exclude $\hat{F}_a$ values that are undefined or equal to zero, which occurs when amino acids are rare or missing (see Wright 1990). The calculations presented subsequently follow Wright's suggestion that if no threefold redundant codons are observed, one should average $\bar{\hat{F}}_2$ and $\bar{\hat{F}}_4$ to obtain $\bar{\hat{F}}_3$. If other k-fold redundancy classes are unobserved, $\bar{\hat{F}}_k$ is assumed to equal $1/k$. Such

Table 1
Nucleotide Compositions Used for the Simulated Sequences

	None	Low-1	Low-2	Med-1	Med-2	High-1	High-2
$f_A \ldots$	0.25	0.20	0.20	0.125	0.125	0.05	0.05
$f_G \ldots$	0.25	0.30	0.20	0.375	0.125	0.45	0.05
$f_C \ldots$	0.25	0.30	0.40	0.375	0.625	0.45	0.85
$f_T \ldots$	0.25	0.20	0.20	0.125	0.125	0.05	0.05

an assumption is conservative with regard to measuring strong codon usage bias. Finally, in the calculation of $\hat{F}_k$ we exclude amino acids that are observed fewer than five times. Here, these methods of dealing with missing data are applied equally to the calculation of both $\hat{N}_c$ and $\hat{N}_c'$.

To explore the properties of the statistics $\hat{N}_c$ and $\hat{N}_c'$, pseudorandomly generated coding sequences were generated for various lengths and nucleotide compositions. Amino acids were chosen uniformly, and codon usage was assumed to be multinomial with proportions determined by the nucleotide composition. The background nucleotide composition estimates are given as f_A, f_C, f_G, and f_T for adenine, cytosine, guanine, and thymine, respectively. For example, for a codon i that has the sequence CAT, the expected frequency was calculated as $e_i = f_C f_A f_T / K$, where K is a renormalization constant for ensuring the e_i for an amino acid sum to one.

A range of nucleotide compositions were used to generate the sequences (Table 1). The nucleotide compositions are modeled after those used in a study by Comeron and Aguade (1998) and contain conditions corresponding to different levels of background GC-AT content. The compositions also vary in their degree of GC skew (Lobry 1996; Shioiri and Takahata 2001). The Low-2, Med-2, and High-2 have higher GC skew relative to Low-1, Med-1, and High-1, respectively. For each composition and gene length, 10,000 sequences were generated. The mean and variance of the $\hat{N}_c$ and $\hat{N}_c'$ statistics were calculated for each set of sequences.

To compare the performance of $\hat{N}_c$ and $\hat{N}_c'$ with other summary statistics that incorporate the effects of background nucleotide composition, Akashi's scaled χ^2 statistic (Akashi 1995), the MCB statistic of Urrutia and Hurst (2001), as well as the $B^*(a)$ measure proposed by Karlin and Mrazek (1996) were also computed for each sequence. Each of these measures attempts to account for background nucleotide composition. In the implementation of these calculations the same expected values that are used for calculating $\hat{N}_c'$ are used for the scaled χ^2 and $B^*(a)$ measures. For calculating MCB the expected values were generated according to the method given by Urrutia and Hurst (2001). For the sequence lengths of 150 and 210 bp, there was insufficient data for the Urrutia and Hurst expectations to be computed, and so MCB was not calculated for these sequence lengths.

Figure 1 shows the mean values of the various bias statistics for sequences of 7,500 bp for the various background nucleotide compositions. The mean values are shown relative to their mean values with uniform nucleotide composition. As expected when the nucleotide frequencies are uniform, $\hat{N}_c$ and $\hat{N}_c'$ are equal, and both share a value of 61. However, as the unevenness in the background nucleotide composition increases, the mean value of $\hat{N}_c$ decreases, whereas that of $\hat{N}_c'$ remains relatively constant. The value of $\hat{N}_c'$ is nearly 61 for all nucleotide composition conditions, whereas that of $\hat{N}_c$ ranges from 61 with equal nucleotide compositions to nearly 26 when the compositions are highly nonuniform. It is also worth noting that $\hat{N}_c$ is particularly sensitive to GC skew. The value of $\hat{N}_c$ is on average nine units smaller for the composition sets with high GC skew than for those with none. Like $\hat{N}_c'$, scaled χ^2 and MCB remain constant over various nucleotide composition sets. The $B^*(a)$ statistic remains nearly constant for modest levels of unevenness in nucleotide composition but increases for the medium and high levels.

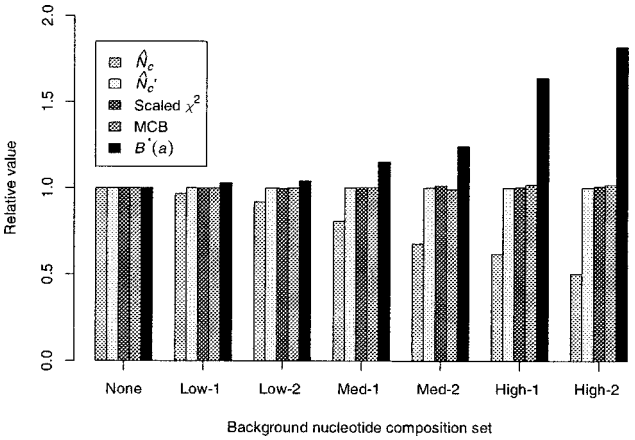

FIG. 1.—Summary statistic values relative to their values when nucleotide compositions are equal.

The statistics were also studied over a range of gene lengths to observe the effects of gene length. Figure 2 represents the values of each statistic relative to their values at 7,500 bp. At short sequence lengths (less

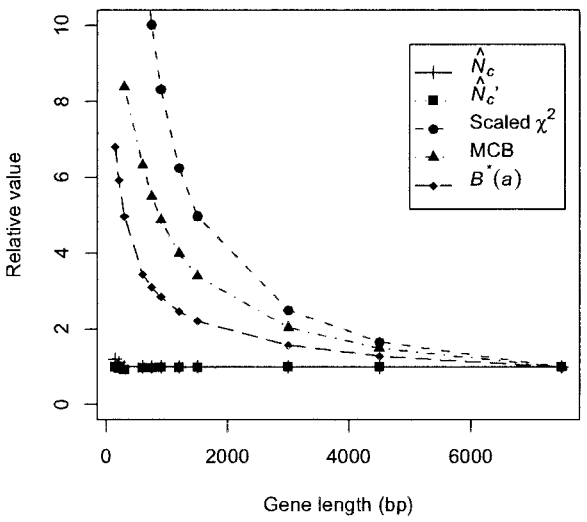

FIG. 2.—Summary statistic values relative to their values at 7,500 bp. Sequences were generated with the Medium-2 composition set. Results for other composition sets are qualitatively similar.

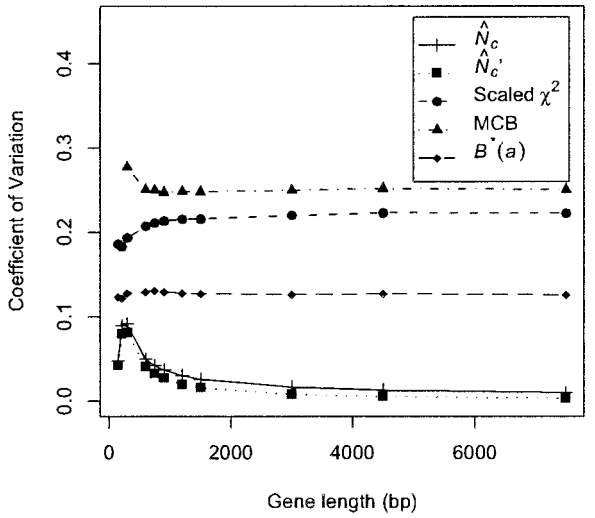

FIG. 3.—CVs for the various summary statistics applied to sequences generated with the Medium-2 composition set. Results for other composition sets are qualitatively similar.

than 600 bp), the values of $\hat{N}_c$ and $\hat{N}_c'$ are higher than their asymptotic values. In general, however, the asymptotic value is well approximated when the gene length reaches 600 bp. In contrast, the values of scaled χ^2, MCB, and $B^*(a)$ do not closely match their asymptotic values even at long sequence lengths (3,000 bp). Figure 3 shows the coefficients of variation (CVs) for the various summary statistics. For sequence lengths of above 600 bp the order of the statistics in terms of increasing CV is $\hat{N}_c'$, $\hat{N}_c$, $B^*(a)$, scaled χ^2, and MCB. Below 600 bp the behavior of each method's CV becomes complicated by the ways in which each measure handles the missing values.

These results show that the statistic $\hat{N}_c'$ performs well in terms of accounting for nucleotide composition, being the least affected by gene length and having a low CV.

Although $\hat{N}_c'$ differs from $\hat{N}_c$, they are similar enough in some cases so that using $\hat{N}_c'$ rather than $\hat{N}_c$ may not always produce different results. For example, a reanalysis with $\hat{N}_c'$ of data from Dunn, Bielawski, and Yang (2001) on the relationship between *Drosophila* substitution rates and codon usage bias produced results that are qualitatively similar to those from the original analysis with $\hat{N}_c$ (data not shown).

However, in other cases the difference between $\hat{N}_c$ and $\hat{N}_c'$ will be important. For example, a difference in $\hat{N}_c$ of 10 may be seen as biologically significant (e.g., Moriyama and Powell 1997, 1998). At that resolution the results presented here show that using $\hat{N}_c'$ with nonuniform expectations will result in a considerable difference if the GC or AT content is on the order of 85%. Such nucleotide composition is extreme in nature and is found genome-wide in protozoan species, such as *Plasmodium falciparum* (Gardner et al. 1998), or in mitochondrial genomes of arthropods, such as *Drosophila*, *Anopheles*, *Bombyx*, and *Apis* (Shioiri and Takahata 2001). In addition, the results presented here show that the presence of nucleotide composition skew will affect the value of $\hat{N}_c$, to change it even more from that of $\hat{N}_c'$.

If differences of less than 10 in the value of $\hat{N}_c$ are of interest, then using $\hat{N}_c'$ rather than $\hat{N}_c$ is important for even lower levels of unevenness in nucleotide composition. For instance, a nucleotide composition of approximately 35% GC or AT can cause the value of $\hat{N}_c$ to change by five units relative to equal usage. In addition, modest levels of GC skew, such as in composition set Low-2, can cause the value of $\hat{N}_c$ to change by five units relative to no skew.

Using $\hat{N}_c'$ with nonuniform expectations will be particularly advantageous when comparing codon usage bias among sequences from genomic regions that vary significantly in background nucleotide composition. Such variation is especially notable in mammals and birds, where variation in nucleotide composition is structured into isochores (Bernardi et al. 1985). In addition, $\hat{N}_c'$ may also be useful in comparative studies of codon usage bias in which the species being studied have substantially different nucleotide compositions.

In summary, this paper has introduced a summary statistic of codon usage bias ($\hat{N}_c'$) that incorporates variation in background nucleotide composition among sequences. Simulations show that relative to the available statistics, $\hat{N}_c'$ effectively adjusts for background nucleotide composition, is the least affected by gene length, and has a low CV. With its use the analysis of codon usage bias can be accomplished without the confounding effects of variation in nucleotide composition.

A freely distributable program that calculates both $\hat{N}_c$ and $\hat{N}_c'$ is available on the Web page: http://ib.berkeley.edu/labs/slatkin/software.html.

Acknowledgments

The author thanks Josh Herbeck and Montgomery Slatkin for helpful discussions, Araxi Urrutia for discussion and for providing the script to calculate MCB, and two reviewers for their comments. The research was supported by a Howard Hughes Medical Institute Predoctoral Fellowship and the National Institutes of Health (NIH GM-40282 to M. Slatkin).

LITERATURE CITED

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

———. 1997. Distinguishing the effects of mutational biases and natural selection on DNA sequence variation. Genetics **147**:1989–1991.

AKASHI, H., and A. EYRE-WALKER. 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. **8**: 688–693.

AKASHI, H., R. KLIMAN, and A. EYRE-WALKER. 1998. Mutation pressure, natural selection, and the evolution of base composition in Drosophila. Genetica **102–103**:49–60.

BENNETZEN, J., and B. HALL. 1982. Codon selection in yeast. J. Biol. Chem. **257**:3026–3031.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS, G. CUNY, M. MEUNIER-ROTIVAL, and F. RODIER. 1985.

The mosaic genome of warm-blooded vertebrates. Science **228**:953–958.

COMERON, J., and M. AGUADE. 1998. An evaluation of measures of synonymous codon usage bias. J. Mol. Evol. **47**: 268–274.

DUNN, K., J. BIELAWSKI, and Z. YANG. 2001. Substitution rates in Drosophila nuclear genes: implications for translational selection. Genetics **157**:295–305.

GARDNER, M., H. TETTELIN, D. CARUCCI et al. (24 co-authors). 1998. Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum. Science **282**:1126–1132.

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

IKEMURA, T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. **146**:1–21.

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

KARLIN, S., and J. MRAZEK. 1996. What drives codon choices in human genes? J. Mol. Biol. **262**:459–472.

KIMURA, M., and J. CROW. 1964. The number of alleles that can be maintained in a finite population. Genetics. **49**:725–738.

KLIMAN, R., and J. HEY. 1993. Reduced natural selection associated with low recombination in Drosophila melanogaster. Mol. Biol. Evol. **10**:1239–1258.

KREITMAN, M., and M. ANTEZANA. 2000. The population and evolutionary genetics of codon bias. Pp. 82–101 *in* R. SINGH and C. KRIMBAS, eds. Evolutionary genetics: from molecules to morphology. Cambridge University Press, Cambridge, U.K.

KREITMAN, M., and J. COMERON. 1999. Coding sequence evolution. Curr. Opin. Genet. Dev. **9**:637–641.

LOBRY, J. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13**:660–665.

MARAIS, G., D. MOUCHIROUD, and L. DURET. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl. Acad. Sci. USA **98**:5688–5692.

MORIYAMA, E., and J. POWELL. 1997. Codon usage bias and tRNA abundance in Drosophila. J. Mol. Evol. **45**:514–523.

———. 1998. Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res. **26**:3188–3193.

OHTA, T., and J. GILLESPIE. 1996. Development of neutral and nearly neutral theories. Theor. Popul. Biol. **49**:128–142.

SHARP, P., and W. LI. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4**:222–230.

SHIOIRI, C., and N. TAKAHATA. 2001. Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. J. Mol. Evol. **53**:364–376.

SMITH, N., and A. EYRE-WALKER. 2001*a*. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol. Biol. Evol. **18**:982–986.

———. 2001*b*. Why are translationally sub-optimal synonymous codons used in Escherichia coli? J. Mol. Evol. **53**: 225–236.

SUEOKA, N., and Y. KAWANISHI. 2000. DNA G+C content of the third codon position and codon usage biases of human genes. Gene **261**:53–62.

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

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

ADAM EYRE-WALKER, reviewing editor

Accepted April 5, 2002