Synonymous Codon Usage in *Escherichia coli*: Selection for Translational Accuracy

*Nina Stoletzki**† and Adam Eyre-Walker*†‡

*Ludwig-Maximilan Universität, Biocenter, Planegg-Martinsried, Germany; †Center for Study of Evolution, School of Life Sciences, University of Sussex, Brighton, United Kingdom; and ‡National Evolutionary Synthesis Center, Durham, North Carolina

In many organisms, selection acts on synonymous codons to improve translation. However, the precise basis of this selection remains unclear in the majority of species. Selection could be acting to maximize the speed of elongation, to minimize the costs of proofreading, or to maximize the accuracy of translation. Using several data sets, we find evidence that codon use in *Escherichia coli* is biased to reduce the costs of both missense and nonsense translational errors. Highly conserved sites and genes have higher codon bias than less conserved ones, and codon bias is positively correlated to gene length and production costs, both indicating selection against missense errors. Additionally, codon bias increases along the length of genes, indicating selection against nonsense errors. Doublet mutations or replacement substitutions do not explain our observations. The correlations remain when we control for expression level and for conflicting selection pressures at the start and end of genes. Considering each amino acid by itself confirms our results. We conclude that selection on synonymous codon use in *E. coli* is largely due to selection for translational accuracy, to reduce the costs of both missense and nonsense errors.

Introduction

Synonymous codons are not used randomly, and in several organisms natural selection seems to bias codon use toward a certain subset of optimal codons. The evidence for this is two-fold. First, in several organisms, including *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Arabidopsis thaliana*, codon bias is correlated to gene expression levels (Gouy and Gautier 1982; Sharp and Li 1987a, 1987b; Duret and Mouchiroud 1999; Goetz and Fuglsang 2005). Second, in organisms for which we have the information, codon usage is biased toward codons that match the most abundant tRNA in the cell or that bind those tRNAs with optimal binding strength. This has been shown directly in *E. coli* and *S. cerevisiae* (Ikemura 1985). For most species, however, cellular tRNA abundances are unknown. In *E. coli*, *Bacillus subtilis*, and *S. cerevisiae*, cellular tRNA abundance correlates closely to tRNA gene copy numbers (Ikemura 1981; Dong et al. 1996; Percudani et al. 1997; Kanaya et al. 1999), and in *D. melanogaster* and *C. elegans*, preferentially used codons, in highly expressed genes, match tRNAs that have high gene copy numbers, suggesting that in these species also there is a correlation between optimal codon use and the abundance of tRNAs (Moriyama and Powell 1997; Duret 2000, 2002; Percudani 2001; although see Kanaya et al. 2001 who do not find this correspondence for *D. melanogaster* and humans).

Although it has been known for 20 years that there is selection on synonymous codon use to maximize some aspect of translation, it has remained unclear what the direct cause of selection for translational optimal codons is; is it to maximize the speed of elongation, minimize the costs of proofreading, or maximize the accuracy of translation (Bulmer 1991; Akashi and Eyre-Walker 1998)? Under all the 3 causes, we expect a correlation between codon bias and expression level, and the use of codons that match common tRNAs.

Here, we set out to investigate whether there is selection for translational accuracy on codon use in *E. coli*. There are 2 types of translational errors that can occur—missense errors in which an incorrect amino acid is incorporated into the growing peptide chain and nonsense errors in which peptide synthesis terminates prematurely. Both missense and nonsense errors that produce non- or misfunctional proteins are costly to the cell because they consume amino acids and energy both in their production and during breakdown. Additionally, missense errors may have other effects of larger impact, for example, a missense error in a DNA polymerase may temporally increase the mutation rate (Ninio 1991).

Although, synonymous codon use can potentially affect both the rate of missense and nonsense errors, we principally focus our attention on whether there is selection against missense errors that lead to proteins with an incorrectly incorporated amino acid. We only briefly consider nonsense errors that lead to a premature termination of peptide synthesis. However, note that both error types might in fact be linked, that is, the occurrence of missense errors might increase the chance of nonsense errors (see, e.g., Kurland and Ehrenberg 1987; Kurland et al. 1996; Farabough and Björk 1999).

We use 3 analyses to test whether selection acts upon synonymous codon use to minimize the number of missense translational errors. First, we use the test suggested by Akashi (1994). He pointed out that if selection was acting to maximize translational accuracy, then selection on synonymous codon bias should be strongest at the functionally most important amino acid sites and that those sites should therefore have higher codon bias. Akashi (1994) inferred the importance of a site by whether it was conserved between 2 species of *Drosophila*. He found that conserved amino acid sites did indeed have higher codon bias in *Drosophila*. However, a similar analysis performed on *E. coli* by Hartl et al. (1994) failed to find this correlation.

Second, we extend Akashi's test from within genes to between genes. We predict that genes that show a higher amino acid divergence between strains should have lower codon bias because genes with high divergence are likely to be those that have a large proportion of functionally

Key words: synonymous codon use, translational accuracy, translational selection.

E-mail: nstoletzki@googlemail.com.

Mol. Biol. Evol. 24(2):374–381. 2007
doi:10.1093/molbev/msl166
Advance Access publication November 13, 2006

relatively unimportant sites. However, it is known that the rate of nonsynonymous substitution and the level of codon bias are both correlated to gene expression level; so any correlation between the divergence between strains and codon bias could be due to the fact that they are both correlated to gene expression. We therefore performed a partial correlation analysis between *d*N and codon bias, controlling for gene expression level.

Third, if selection is acting to reduce the costs of missense errors, we expect longer genes to have higher codon bias because the cost of producing a defective protein should be dependent on the total energy and resources that have been used in producing the protein; both should accrue with each added amino acid (Eyre-Walker 1996a). However, we also expect codon bias to increase along the length of the gene if there is selection to minimize nonsense errors because nonsense errors lead to the termination of protein synthesis and it is more costly to produce most of a protein than a small part of a protein. Hence, selection against nonsense errors also causes a positive correlation of codon bias and gene length. We can disentangle these 2 factors and independently test whether selection is acting specifically to minimize the number of missense errors by only considering genes that are greater than a certain length and only considering the level of bias in those codons up to that length. Again, we perform a partial correlation analysis controlling for gene expression.

As an alternative method of controlling gene expression, we use ribosomal genes, which, with the exception of L7/L12, we expect to be expressed at approximately similar levels, as they are coregulated and synthesized in equimolar amounts (Lindahl and Zengel 1986; Keener and Nomura 1996; Nomura 1999). Eyre-Walker (1996a) previously performed this analysis of ribosomal genes, but he did not disentangle selection against nonsense and missense errors and some of the codon bias values he used were incorrect (Higgs P, personal communication). We therefore repeat the analysis here.

We find evidence that selection acts upon synonymous codons to minimize both missense and nonsense errors. When disentangling codon bias into its contributing amino acids, we find our predictions to be confirmed, and we conclude that selection on synonymous codon use is at least in part due to selection for translational accuracy, to reduce the costs of missense and nonsense errors.

Materials and Methods

We performed 3 tests of whether selection acts to minimize missense errors. To test whether selection on synonymous codon bias is stronger at potentially functionally important amino acid sites, we use alignments between the *E. coli* strains K12, O157:H7, and CFT073 (Jordan et al. 2005). We use parsimony to infer along which lineage an amino acid mutation has occurred and designate codons in which the substitution occurs in any of the lineages as nonconserved; conserved codons are therefore those that are conserved in all 3 strains. We measure the level of codon bias of both conserved and nonconserved codons in K12.

As pointed out before, our observations might be confounded by other factors (Akashi 1994; Rocha and Danchin 2004). Nonsynonymous substitutions can convert an optimal to a nonoptimal codon, thereby contributing to less optimal codon use at less constrained sites. We exclude those sites if the change occurred along the K12 lineage because this is the lineage in which we measure codon bias. Doublet mutations can couple synonymous and nonsynonymous substitutions (Wolfe and Sharp 1993; Averof et al. 2000) and thereby generate an artificial correlation between codon conservation and the level of synonymous codon bias. As a consequence, we exclude doublet mutations from our data set.

The analysis yields a 2×2 contingency table (the codon either has a nonsynonymous substitution or does not and the codon is either optimal in K12 or suboptimal) for each amino acid in each gene. These contingency tables can be combined using the Mantel–Haenszel *Z* statistic as suggested by Akashi (1994; according to Sokal and Rohlf 1995). We excluded contingency tables with expected values that were zero and tested for homogeneity and computed the joint odds ratio (W_{MH}) and its significance, including the continuity correction. We orient the odds ratio such that when $W_{MH} > 1$, there is a greater frequency of optimal codons being used at conserved than nonconserved sites.

To test whether genes that have a higher amino acid divergence between strains have lower codon bias than genes with lower amino acid divergence, we used *E. coli* strains K12, O157:H7, and CFT073 as above (Jordan et al. 2005) and phylogenetic analyses using maximum likelihoods F3×4 model (Yang 1997) to compute the level of amino acid divergence (*d*N) between the 3 strains. Additionally, we use expression data from *E. coli* NCM 3416 (a derivate of *E. coli* K12) by Bernstein et al. (2002) and the GenProtEC Web site (http://genprotec.mbl.edu) to extract 327 genes from Jordan et al. (2005) for which Bernstein et al. report expression levels in lysogeny broth (LB) medium.

To test whether codon bias increases with gene length as a result of missense errors, we used the *E. coli* genes for which we had expression level measures (Bernstein et al. 2002); additionally, we downloaded 54 ribosomal protein genes of *E. coli* K12 from the Ribosomal Gene Database (http://ribosome.miyazaki-med.ac.jp/; Nakao et al. 2004). To test for selection against nonsense errors, we used a subset of 135 *E. coli* genes (Jordan et al. [2005] data) that are at least 2,000 nt in length; we exclude the first 50 codons and calculated codon bias at each codon position for the next 600 codons; and we used a regression analysis to test whether codon bias increases with gene position. To test for selection against missense errors specifically, excluding the contribution of nonsense errors, we extend Eyre-Walker's (1996a) test and take genes that are at least 2,000 nt long, exclude the first 50 codons, and measure codon bias for the next 600 codons.

We assigned preferred codons according to Sharp and Li (1987a) and computed the frequency of optimal codons as F_{OP} = number of optimal codons/all codons per gene (excluding trp, met, and stop codons) according to Ikemura (1981).

This general measure, however, does not differentiate among amino acids, and the contribution of different amino acids to F_{OP} depends not only on their respective levels of

Table 1
Spearman's Rank Correlations and Partial Correlations, Controlling for Levels of Gene Expression, between (1) Codon Bias and *d*N or (2) Codon Bias and Gene Length; (3) Regression Analysis of Codon Bias with Position in Gene

	Codon Bias-*d*N	Codon Bias-Gene Length	Codon Bias-Position in Gene
F_{OP}	Spearman's rho = −0.4036***	Spearman's rho = +0.3027***	Gradient = +0.2100***
	Spearman's partial = −0.3100***	Spearman's partial = +0.3054***	
F_{OPaa}	Spearman's rho = −0.3538***	Spearman's rho = +0.3130***	Gradient = +0.188***
	Spearman's partial = −0.2745***	Spearman's partial = +0.3145***	

NOTE.—After removal of 50 codons at start and 20 codons at the end of genes.
***$P < 0.001$.

codon bias but also on their relative frequency, degeneracy, and optimal codon numbers. Relative frequencies, however, will vary among genes, for example, with expression level (Lobry and Gautier 1994; Akashi and Gojobori 2002; Akashi 2003).

We also compute F_{OPi}, the optimal codon use for each contributing amino acid separately, and additionally F_{OPaa}, the average of F_{OPi} per gene. To reduce sampling errors, we use only amino acids that are present at least 4 times in a gene. We exclude cys as it is only present in very small frequencies in many genes, and we again exclude trp and met as they only have 1 codon each. We only use nonribosomal genes in which all 17 contributing amino acids are present at least 4 times. Note that this reduces the number of contributing genes for F_{OPaa} to 192; when excluding the first 50 and last 20 codons, the number of contributing genes is further reduced to 160. We weighted the contribution of each amino acid by dividing the expected optimal codon use x/n, where x is the number of assigned optimal codons and n is the total number of synonymous codons for the amino acid, that is, the degeneracy of the amino acid; for example, there are 4 glycine codons of which 2 are regarded as optimal in *E. coli*; we therefore, divided the F_{OP} value for glycine by 2/4. Note that F_{OPaa} controls for amino acid frequency but that amino acid contributions will still differ depending on their degeneracy and optimal codon numbers, that is, the maximum contribution of leucine can increase to 6, whereas for glycine it can increase to 2.

For ribosomal genes, the F_{OPaa} codon bias measure was not feasible as too few genes would be left with 17 contributing amino acids. We compute standard Codon Adaptation Index (CAI) and F_{OP}, and following Eyre-Walker (1996a) we exclude genes, which are less than 100 codons (i.e., the gene has at least 30 codons when codons at the start and end are removed—see below).

We computed CAI (Sharp and Li 1987a) and the effective number of codons (Wright 1990) using CodonW (Peden 1999).

We remove the codons at the start and at the end of *E. coli* genes because they show lower codon bias and are thought to be under conflicting or different selection pressures (Eyre-Walker and Bulmer 1993; Eyre-Walker 1996b; Hooper and Berg 2000). Eyre-Walker (1996a) removed 50 codons at the start and 20 codons at the end, whereas Comeron et al. (1999) removed 100 codons at the start and 50 codons at the end. We used both criteria for all analyses except those involving the ribosomal genes where it was impractical to remove the 100 codons at the start and 50 codons at the end because this left only a few genes with sufficient codons. For the other analyses, our results were generally unaffected by how many codons were removed and so we just present the results obtained by removing 50 at the start and 20 at the end.

We use nonparametric Spearman's rank and Spearman's partial correlations; we use standard least squares regression to evaluate the relationship between codon bias and the position in the gene (SPSS and R programs).

Results

We performed 3 tests of whether selection is acting upon synonymous codon use to maximize translational accuracy. First, we performed the test suggested by Akashi (1994); if selection is acting to minimize translational errors, then codon bias is expected to be highest in codons that encode the most important amino acid sites. We judged the functional importance of a codon by whether it was conserved between *E. coli* strains K12, O157:H7, and CFT073. As expected, we find that optimal codons occur significantly more frequently at codons in which the amino acid is conserved (presumably functionally constrained) than at nonconserved sites within the same gene—the frequency of optimal codons is approximately 1.7-fold higher than expected (W_{MH} = 1.729***). This result remains qualitatively unchanged if we remove codons at the start and end of genes that are thought to be subject to conflicting selection pressures (removing a total of 70 codons: W_{MH} = 1.664***). Nonsynonymous substitutions may change an optimal codon to a suboptimal codon and vice versa. To test whether nonsynonymous substitutions are responsible for the lower codon bias in the nonconserved sites, we removed nonconserved amino acid sites that differed in optimal codon status (W_{MH} = 1.748***; removing 70 codons at the start and end: W_{MH} = 1.683***). The test is not significant when we remove the first 100 and last 50 codons, but there are very few nonconserved amino acid sites in this test.

In our second test of accuracy, we extended Akashi's test from within genes to between genes, expecting that genes that show a higher amino acid divergence between strains have lower codon bias because genes with high divergence are likely to be those that have a large proportion of relatively unimportant sites. We find a negative correlation of codon bias and *d*N per gene, which remains when we control for expression levels (table 1). The correlation is also apparent when we control for amino acid composition (by using F_{OPaa}) and when we remove codons potentially under different selective constraints at the beginning and

Table 2
Test Statistics for Each Amino Acid after Removal of the First 50 and Last 20 Codons of Each Gene

	(1) CB-Expression	(2) CB-*d*N|Expression	(3) CB-GL|Expression	(4) CB-Position	(5) W_{MH}
Arg	+0.3943***	−0.2855***	+0.1563**	+0.037, NS	3.64***
Phe	+0.3812***	−0.2455***	+0.2096**	+0.174, ***	1.11, NS
Leu	+0.3627***	−0.1980***	+0.1658**	+0.152, ***	1.57, ***
Tyr	+0.3461***	−0.0165, NS	+0.1786**	+0.011, NS	1.06, NS
Gly	+0.3325***	−0.1656**	+0.1204*	+0.046, NS	1.19*
Asn	+0.3108***	−0.1761**	+0.2516***	+0.047, NS	1.48***
Asp	+0.2851***	−0.0972, NS	+0.1129, NS	−0.023, NS	1.00, NS
His	+0.2792***	−0.2016**	+0.1370, NS	+0.046, NS	1.00, NS
Ile	+0.2757***	−0.0179, NS	+0.1836**	−0.062, NS	1.46***
Pro	+0.2506***	−0.2350***	+0.2163***	+0.126**	1.36**
Val	+0.2335***	−0.0379, NS	+0.0087, NS	+0.005, NS	0.88*
Gln	+0.1854**	−0.0835, NS	+0.1790**	0.083, NS	1.04, NS
Glu	+0.1767**	−0.0380, NS	−0.0446, NS	−0.008, NS	1.32***
Ala	+0.0870, NS	−0.1527**	+0.1337*	+0.042, NS	1.04, NS
Thr	+0.0186, NS	−0.1537*	+0.1504*	+0.036, NS	1.35, ***
Ser	−0.0208, NS	+0.0463, NS	+0.0099, NS	+0.037, NS	0.96, NS
Cys	−0.0221, NS	−0.2008, NS	−0.1506, NS	+0.248, NS	0.78, NS
Lys	−0.0282, NS	−0.0116, NS	+0.0553, NS	−0.081, NS	NS

NOTE.—(1) Spearman's rank correlation, (2 and 3) partial correlations controlling for expression, (4) regression analysis of codon bias position in gene, and (5) comparison of codon bias at conserved and nonconserved amino acid sites with Mantel–Haenszel statistics. CB = codon bias, expr = expression, GL = gene length, Pos = codon position within the gene, and NS = nonsignificant.

***$P < 0.001$, **$P < 0.01$, and *$P < 0.05$.

end of the gene (table 1). However, note that because of the relatively high level of noise in current expression measurements (for yeast see Coghlan and Wolfe 2000; for *E. coli* see Dos Reis et al. 2003), it is difficult to control for gene expression (see also Drummond, Raval, et al. 2005) and the negative partial correlation of codon bias and *d*N could be an artifact of not controlling for gene expression adequately.

In our third test, we followed the rationale of Eyre-Walker (1996a); if selection is acting to minimize missense translational errors, then codon bias is expected to be highest when production costs of the protein are high. For the nonribosomal genes, we find a positive correlation of codon bias and gene length (table 1). This is unlikely to be an artifact of correlations to expression: although gene length and codon bias are both correlated to expression, the correlations are in opposite directions; gene length is negatively correlated to expression, whereas codon bias is positively correlated.

However, we expect codon bias to increase with gene length if selection is acting to minimize nonsense errors during translation (see Introduction). Selection against nonsense errors and the corresponding codon bias is expected to increase in strength along the length of the gene which is, indeed, what we observe (table 1) and what has been observed before in *E. coli* (Qin et al. 2004). Note, however, that the slope is 0.000051, that is, F_{OP} will increase by 0.051 every 1,000 bp; so it is not a very strong effect. We disentangle selection against missense and nonsense errors by only considering genes that are greater than a certain length and measuring codon bias across the same region in the gene (codons 51–650); the remaining positive correlation indicates selection specifically against missense errors (nonribosomal genes controlling expression: $+0.1585, P = 0.0181$).

We have also repeated Eyre-Walker's analysis of ribosomal proteins because there were some errors in his codon adaptation values. Eyre-Walker also used 2 measures of codon bias, CAI, and a CAI_u controlling for amino acid composition, both supporting a positive correlation of gene length and codon bias. We find Eyre-Walker's original observation is supported if we follow his procedure and exclude genes that are less than 100 codons, that is, the gene has at least 30 codons when codons at the start and end are removed (CAI with gene length: $+0.5125$, $P = 0.0019$; F_{OP}: $+0.4113$, $P = 0.0157$). We disentangle selection against missense errors by measuring codon bias of the first 30 codons after codons at the start and end are removed; the remaining positive correlation indicates selection specifically against missense errors (CAI: $+0.4773$, $P = 0.0043$; F_{OP}: $+0.3110$, $P = 0.0734$).

Performing our 3 tests on individual amino acids, we confirm our predictions (table 2): most amino acids show the expected patterns, that is, 1) conserved codons have higher codon bias than unconstrained codons, 2) a negative correlation between codon bias and *d*N, and 3) a positive correlation to gene length. We ranked amino acids by how strongly their codon usage correlated to expression level. As expected, amino acids that show little or no correlation in their codon usage with expression level (ser, cys, and lys) show no evidence of selection for translational accuracy, whereas amino acids that have a strong correlation with expression level (arg, phe, leu, tyr, gly, asn, and asp) show strong evidence for selection for translational accuracy. In the rare cases where these predictions are contradicted, the test statistic is generally nonsignificant.

Discussion

We find strong support for selection for translational accuracy using all 3 tests. Highly conserved sites and genes have higher codon bias than less conserved sites and genes, and codon bias is positively correlated to gene length and production costs, both indicating selection against missense

errors; additionally, codon bias increases along the length of genes, indicating selection against nonsense errors. We control for expression level, amino acid composition, and for codons under different selective constraints at the start and end of genes. We find our predictions confirmed across individual amino acids. It is difficult to think of an alternative explanation for all these results. However, a number of the analyses we have performed have been considered before, sometimes with different results.

In contrast to us, Hartl et al. (1994) did not find significant differences in the degree of codon bias between codons in which the amino acid was conserved and nonconserved. They concluded that selection for translational accuracy was not effective in enteric bacteria. This could have been due to any one of several reasons; the test statistic used, the sample size, or it could have been due to the fact that Hartl et al. compared *E. coli* with *Salmonella*. Enteric bacteria appear to have very high rates of adaptive amino acid substitution (Charlesworth and Eyre-Walker 2006), and if many amino acid substitutions are due to adaptive evolution and not to random genetic drift, then constraint may not be a good indicator of whether a site is important or not. Rather than 2 species of bacteria, we compare 3 relatively closely related strains of *E. coli* that undergo recombination (Charlesworth and Eyre-Walker 2006), which suggests that they are strains from the same population genetic species—that is, a group of strains that can undergo genetic drift together. As such, amino acid sites that differ between the strains are likely to be neutral and therefore a better indicator that the site is relatively unimportant. We find that more important amino acids sites have higher bias than less conserved sites.

Like us, Sharp and Li (1987b) and Sharp (1991) also found a negative correlation between codon bias and *d*N for *E. coli* compared with *Salmonella typhimurium*. However, they did not control for expression level and it is known in *E. coli*—as in other organisms—that the rate of nonsynonymous substitution and the level of codon bias are both correlated to gene expression level (Gouy and Gautier 1982; Sharp and Li 1987a, 1987b; Sharp 1991; Rocha and Danchin 2004), so any correlation between the divergence between strains and codon bias could also be due to the fact that they are both correlated to gene expression, and controlling for gene expression is crucial in an analysis of this kind. We attempted to control for expression level by using partial correlations, and we still find that functionally more constrained genes have higher bias than less constrained genes; however, as highlighted before, the relatively high level of noise in current expression measurements makes controlling for gene expression difficult (Drummond, Ravel, et al. 2005).

A positive correlation of codon bias and gene length has been reported before, but in contrast to us, Eyre-Walker (1996a) and Moriyama and Powell (1998) did not disentangle selection against mis- and nonsense errors; so their results could have been entirely due to selection against nonsense errors. Indeed, Qin et al. (2004) have shown that codon bias increases along the length of genes in *E. coli*. We find selection against both non- and missense errors. Surprisingly, Comeron et al. (1999) found that the positive correlation of codon bias and gene length in *E. coli* disappears when excluding sites at the start and end of genes that might be under different selective constraints. This again could have been for several reasons; we find that the choice of codon bias statistic has a dominant effect. If we use the effective number of codons (Ec) (Wright 1990) as Comeron et al. did, we find no correlation, whereas by using F_{OP} or CAI, we do (Ec: $r = +0.0304$, $P = 0.4933$; F_{OP}: $+0.2408$***; CAI: $r = +0.2101$, $P = 0.0008$). This may be due to the fact that Ec does not take into account the direction of codon bias toward translationally optimal codons, but simply measures the overall bias. Statistics, such as Ec and Chi/L must be used with caution in species such as *E. coli* because for some amino acids lowly expressed genes can be biased in the opposite direction to highly expressed genes—for example, consider the 2-fold degenerate codons of phe: in highly expressed genes, the frequency of the optimal codon (TTC) is 78.5%, whereas in lowly expressed genes, the frequency of the nonoptimal codon (TTT) is 70.5% (see table 1 given by Smith and Eyre-Walker 2001).

More recently, Marquez et al. (2005) tested for selection against translational errors and concluded, in contrast to us, that selection against translational errors does not affect codon use. Marquez et al., however, infer that codons with small "error values" would be selected for translational accuracy, whereas we look for selection on translationally optimal codons as defined by cellular tRNA content in the cell and by correlation with expression level. An error value is the sum of differences in amino acid properties when changing from one codon to another that can be reached by a single base substitution (see also Haig and Hurst 1991; Freeland and Hurst 1998). Hence, it is not surprising that they reached a different conclusion.

Our results show that the rate of translational errors for *E. coli* is apparently sufficiently high and that associated costs are sufficiently strong to affect synonymous codon use. As such, we hereby demonstrate an example of some form of selection upon what Bürger et al. (2006) have termed the phenotypic mutation rate.

Note that selection for translational accuracy, of course, does not exclude other forces affecting synonymous codon use, such as selection with respect to initiation, elongation rate, mRNA secondary structure, ribosome stalling, or SsrA tagging (see, e.g., Klionsky et al. 1986; Gross et al. 1990; Bulmer 1991; Hayes et al. 2001; Sunohara et al. 2004).

If there is selection upon translational accuracy, then this has a number of implications. First, we expect the strength of selection to vary between sites. This might effectively explain the observation of apparent site-specific codon bias by Maynard Smith and Smith (1996a, 1996b). Maynard Smith and Smith found that certain sites appeared to be fixed for a particular codon across highly divergent enteric bacteria. This pattern might be explained by site-specific selection or by site-specific mutation rates (Maynard Smith and Smith 1996b; Berg 1999); site-specific selection for translational accuracy could contribute to the patterns.

Second, variation in the strength of selection will mean that the substitution rate varies between synonymous sites; hence, synonymous substitution rates may be underestimated. This may go someway in explaining why the

divergence between *E. coli* and *Salmonella enterica* is far below what one might expect, given the apparent nucleotide mutation rate and divergence time of the 2 taxa (Eyre-Walker and Bulmer 1995; Ochman 2003), although again this could be due to variation in the mutation rate (Berg 1999).

Third, selection for translational accuracy has often been considered to be of less importance than selection for increased elongation rate, and models of synonymous codon evolution have reflected this (Bulmer 1987, 1991; Berg and Kurland 1997; Xia 1998). The evidence presented here, in *Drosophila* (Akashi 1994) and in *C. elegans* (Marais and Duret 2001), suggests that selection for accuracy may be an important force shaping codon use. It would be interesting to develop models incorporating selection for translational accuracy and see how such models fit the levels and patterns of codon bias and tRNA frequencies observed in data. Gilchrist and Wagner (2006) recently incorporated selection against nonsense errors in a model of translation; however, selection against missense errors is not considered.

Selection against missense errors is also interesting due to its association with functional constraints; sites that are functionally constrained and consequently conserved at the amino acid level are also likely to experience stronger selection for translational accuracy and hence higher codon bias. This might explain why there is a negative correlation between *d*N and codon bias as we and others have observed in enteric bacteria (Sharp and Li 1987b; Sharp 1991; Rocha and Danchin 2004), *Drosophila* (Akashi 1994; Betancourt and Presgraves 2002; Marais et al. 2005), and yeast (Pal et al. 2001; Drummond, Bloom, et al. 2005; Drummond, Ravel, et al. 2005; Stoletzki et al. 2005). Plotkin et al. (2006) have also recently reported a negative correlation between *d*N/*d*S and codon bias, which due to the way in which they "correct" *d*S for codon usage is effectively a correlation between *d*N and codon bias.

The correlation of selection strengths may also contribute to the correlation between the rates of synonymous and nonsynonymous substitution. Eyre-Walker and Bulmer (1995) and Berg and Martelius (1995) originally suggested, based on an analysis of synonymous codon bias and the synonymous substitution rate, that this correlation was probably due to a negative correlation between the mutation rate and gene expression levels. However, most direct estimates of mutation rates suggest quite the opposite patterns: that the mutation rate increases with gene expression (Hudson et al. 2003; Ochman 2003). Thus, the correlation between the rates of nonsynonymous and synonymous substitution does not appear to be due to variation in the mutation rate. However, selection on translational accuracy could produce the correlation by giving correlated strengths of selection at synonymous and nonsynonymous sites. Selection for translational accuracy could act alone or, as recently suggested, in combination with selection for translational robustness (Drummond, Bloom, et al. 2005).

The fact that *d*N is correlated to codon bias and that this is probably associated with selection for translational accuracy suggests that codon bias might be used as a measure of the level of constraint upon a site or a gene (Plotkin et al. 2004; Stoletzki et al. 2005; Plotkin et al. 2006).

Acknowledgments

We are grateful to Claus Wilke and 2 other anonymous referees for comments. N.S. was supported by a Emmy Noether grant to J. Hermisson and A.E.-W. was supported by the National Evolutionary Synthesis Centre and the Biotechnology and Biological Sciences Research Council.

Literature Cited

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

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

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

Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci USA. 99:3695–3670.

Averof M, Rokas A, Wolfe KH, Sharp PM. 2000. Evidence for a high frequency of simultaneously double-nucleotide mutations. Science. 287:1283–1286.

Berg OG. 1999. Synonymous nucleotide divergence and saturation: effects of site-specific variations and mutation rates. J Mol Evol. 48:398–407.

Berg OG, Kurland CG. 1997. Growth rate-optimised tRNA abundance and codon usage. J Mol Biol. 270:544–550.

Berg OG, Martelius M. 1995. Synonymous substitution rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship to gene expression and selection pressure. J Mol Biol. 41:449–456.

Bernstein JA, Khodursky AB, Lin P-H, Lin-Chao S, Cohen SN. 2002. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proc Natl Acad Sci USA. 99(15):9697–9702.

Betancourt A, Presgraves D. 2002. Linkage limits the power of natural selection in Drosophila. Proc Natl Acad Sci USA. 99(21):13616–13620.

Bulmer M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature. 325(6106):728–730.

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

Bürger R, Willensdorfer M, Nowak MA. 2006. Why are phenotypic mutation rates much higher than genotypic mutation rates? Genetics. 172:197–206.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. Mol Biol Evol. 23(7): 1348–1356.

Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast. 16:1131–1145.

Comeron JM, Kreitman M, Aguade M. 1999. Natural selection on synonymous sites is correlated with gene length and recombination rate in Drosophila. Genetics. 151:239–249.

Dong H, Nielsen L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J Mol Biol. 260:649–663.

Dos Reis M, Wernisch L, Sava R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K 12 genome. Nucleic Acids Res. 31(23):6976–6985.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed genes evolve slowly. Proc Natl Acad Sci USA. 102(40):14338–14343.

Drummond DA, Raval A, Wilke CO. 2005. A single determinant dominates the rates of yeast protein evolution. Mol Biol Evol. 23(3):327–337.

Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of high expression genes. Trends Genet. 16:287–289.

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

Duret L, Mouchiroud D. 1999. Expression pattern, and surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci USA. 96:4482–4487.

Eyre-Walker A. 1996a. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol Biol Evol. 13:864–872.

Eyre-Walker A. 1996b. The close proximity of *Escherichia coli* genes: consequences for stop codon and synonymous codon use. J Mol Evol. 42:73–78.

Eyre-Walker A, Bulmer M. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. 21:4599–4603.

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

Farabough PJ, Björk GR. 1999. How translational accuracy influences reading frame maintenance. EMBO J. 18(6):1427–1434.

Freeland SJ, Hurst LD. 1998. The genetic code is one in a million. J Mol Evol. 47:238–248.

Gilchrist MA, Wagner A. 2006. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. J Theor Biol. 239:417–434.

Goetz R, Fuglsang A. 2005. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. Biochem Biophys Res Commun. 327:4–7.

Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10:7055–7074.

Gross G, Mielke C, Hollatz I, Blöcker H, Frank R. 1990. RNA primary sequence or secondary structure in the translational initiation region controls expression of two variant interferon-β genes in *Escherichia coli*. J Biol Chem. 265(29):17627–17636.

Haig D, Hurst LD. 1991. A quantitative measure of error-minimization in the genetic code. J Mol Evol. 33:412–417.

Hartl DL, Moiyama EN, Sawyer S. 1994. Selection intensity for codon bias. Genetics. 138:227–234.

Hayes CS, Bose B, Sauer RT. 2001. Stop codon preceded by rare arginine codons are efficient determinants of SsrA tagging in *Escherichia coli*. Proc Natl Acad Sci USA. 99(6):3440–3445.

Hooper SD, Berg OG. 2000. Gradients in nucleotide and codon usage along *Escherichia coli* genes. Nucleic Acids Res. 28(18):3517–3523.

Hudson RE, Bergthorsen U, Ochman H. 2003. Transcription increases multiple spontaneous point mutations in *Salmonella enterica*. Nucleic Acids Res. 31(15):4517–4522.

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.

Ikemura T. 1985. Review codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2(1): 13–34.

Jordan KI, Kondrashov FA, Adzhubel IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S. 2005. A universal trend of amino acid gain and loss in protein evolution. Nature. 433:633–638.

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

Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies on 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.

Keener J, Nomura M. 1996. Regulation of ribosome synthesis. Chapter 90. In: Neidhardt FC, editor. *Escherichia coli* and *Salmonella typhimurium*. Cellular and Molecular Biology. Washington (DC): ASM Press. p. 1417–1431.

Klionsky DJ, Skalnik DG, Simoni RD. 1986. Differential translation of the genes encoding proton-translocating ATPase of *Escherichia coli*. J Biol Chem. 261:8096–8099.

Kurland CG, Ehrenberg M. 1987. Growth optimizing accuracy of gene expression. Annu Rev Biophys Biophys Chem. 16:291–317.

Kurland CG, Hughes D, Ehrenberg M. 1996. Limitations of translational accuracy. Chapter 65. In: Neidhardt FC, editor. *Escherichia coli* and *Salmonella typhimurium*. Cellular and Molecular Biology. Washington (DC): ASM Press. p. 979–1004.

Lindahl L, Zengel JM. 1986. Ribosomal genes in *Escherichia coli*. Annu Rev Genet. 20:297–326.

Lobry JR, Gautier C. 1994. Hydrophobicity, expressivity and aromacity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res. 22(15):3174–3180.

Marais G, Domazet-Loso D, Tautz D, Charlesworth B. 2005. Correlated evolution of synonymous and nonsynonymous sites in Drosophila. J Mol Evol. 59:771–779.

Marais G, Duret L. 2001. Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J Mol Evol. 52:275–280.

Marquez R, Smit S, Knight R. 2005. Do universal codon-usage patterns minimize the effects of mutation and translation error? Genome Biol. 6(11):R91.

Maynard Smith J, Smith NH. 1996a. Synonymous nucleoticle divergence: what is "saturation"? Genetics. 142:1033–1034.

Maynard Smith J, Smith NH. 1996b. Site-specific codon bias in bacteria. Genetics. 142:1037–1043.

Moriyama EN, Powell JR. 1997. Codon usage and tRNA abundance in Drosophila. J Mol Evol. 45:514–523.

Moriyama EN, Powell JR. 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res. 26(13):3188–3193.

Nakao A, Yoshihama M, Kenmochi N. 2004. RPG: the Ribosomal Protein Gene database. Nucleic Acids Res. 32:D168–D169.

Ninio J. 1991.Transient mutators: a semiquantitative analysis of the influence of translation and transcription errors on mutation rates. Genetics. 129:957–962.

Nomura M. 1999. Regulation of ribosome biosynthesis in *Escherichia coli* and *Saccharomyces cerevisiae*: diversity and common principles. J Bact. 181(22):6857–6864.

Ochman H. 2003. Neutral mutations and neutral substitutions in bacterial genomes. J Biol Evol. 20(12):2091–2096.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics. 158:927–931.

Peden JF. 1999. Analysis of codon usage [dissertation]. [Nottingham (UK)]: University of Nottingham.

Percudani R. 2001. Restricted wobble rules for eukaryotic genomes. Trends Genet. 17:133–135.

Percudani R, Pavesi, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J Mol Biol. 268:322–330.

Plotkin JB, Dushoff J, Desai MM, Fraser HB. 2006. Estimating selection pressures from limited comparative data. Mol Biol Evol. 23(8):1457–1459.

Plotkin JB, Dushoff J, Fraser HB. 2004. Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. Nature. 428:942–945.

Qin H, Wu WB, Comeron JM, Kreitman M, Li W-H. 2004. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics. 168:2245–2260.

Rocha EPC, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol. 21(1):108–116.

Sharp PM. 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J Mol Evol. 33:23–33.

Sharp PM, Li W-H. 1987a. The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.

Sharp PM, Li W-H. 1987b. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol. 4:222–230.

Smith NGC, Eyre-Walker A. 2001. Why are translationally suboptimal codons used in Escherichia coli? J Mol Evol. 53: 225–236.

Sokal RR, Rohlf FJ. 1995. Biometry. New York: WH Freeman and company.

Stoletzki N, Welch J, Hermisson J, Eyre-Walker A. 2005. A dissection of volatility in yeast. Mol Biol Evol. 22:2022–2026.

Sunohara T, Jojima K, Tagami H, Inada T, Aiba H. 2004. Ribosome stalling during translation elongation induces cleavage of mRNA being translated in *Escherichia coli*. J Biol Chem. 279(15):15368–15375.

Wolfe KH, Sharp PM. 1993. Mammalian gene evolution—nucleotide sequence divergence between mouse and rat. J Mol Evol. 37:441–456.

Wright S. 1990. The 'effective number of codons' used in a gene. Gene. 87:23–29.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–565.

Xia X. 1998. How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*? Genetics. 149:37–44.

Kenneth Wolfe, Associate Editor

Accepted November 2, 2006