Weak selection on synonymous codons substantially inflates *dN/dS* estimates in bacteria

Shakibur Rahman[a], Sergei L. Kosakovsky Pond[b], Andrew Webb[a], and Jody Hey[a,1]

[a]Center for Computational Genetics and Genomics, Department of Biology, Temple University, Philadelphia, PA 19122; and [b]Institute for Genomics and Evolutionary Medicine, Department of Biology, Temple University, Philadelphia, PA 19122

Edited by Michael Lynch, Arizona State University, Tempe, AZ, and approved April 6, 2021 (received for review November 12, 2020)

Synonymous codon substitutions are not always selectively neutral as revealed by several types of analyses, including studies of codon usage patterns among genes. We analyzed codon usage in 13 bacterial genomes sampled from across a large order of bacteria, Enterobacterales, and identified presumptively neutral and selected classes of synonymous substitutions. To estimate substitution rates, given a neutral/selected classification of synonymous substitutions, we developed a flexible *dN*/*dS* substitution model that allows multiple classes of synonymous substitutions. Under this multiclass synonymous substitution (MSS) model, the denominator of *dN*/*dS* includes only the strictly neutral class of synonymous substitutions. On average, the value of *dN*/*dS* under the MSS model was 80% of that under the standard codon model in which all synonymous substitutions are assumed to be neutral. The indication is that conventional *dN*/*dS* analyses overestimate these values and thus overestimate the frequency of positive diversifying selection and underestimate the strength of purifying selection. To quantify the strength of selection necessary to explain this reduction, we developed a model of selected compensatory codon substitutions. The reduction in synonymous substitution rate, and thus the contribution that selection makes to codon bias variation among genes, can be adequately explained by very weak selection, with a mean product of population size and selection coefficient, *Ns* = 0.8.

natural selection | *dN/dS* | neutral model | codon-substitution models

The redundancy of the genetic code has long provided investigators a natural dichotomy in which two evolutionary rates are compared, one for the mutations that cause an amino acid change and a second for those mutations in coding regions that do not (1–3). Typically, it is assumed that synonymous changes are selectively neutral and that therefore they accrue at the underlying mutation rate (4). Thus, the ratio of nonsynonymous to synonymous substitutions dN/dS is expected to equal 1 if there is no selection on nonsynonymous substitutions. The ratio provides a simple way of assessing evolutionary pressure on protein coding sequences, as in principle it can reveal both the magnitude and the direction of selection on nonsynonymous variants, with values less than 1 reflecting selective constraint (i.e., negative selection) and values greater than 1 reflecting an overabundance of nonsynonymous substitutions (i.e., positive selection) (5). Investigators have a rich set of methods for estimating dN/dS ratios on the branches of phylogenies and sites in the alignment (6–10), and the fundamental approach, of assessing selective histories using a neutral, synonymous baseline, is fully baked into evolutionary genetics.

Notwithstanding the prevalence of dN/dS analyses, it has long been understood that for some organisms, the rate of synonymous changes varies in ways that are not consistent with selective neutrality. Early on, as the database of messenger RNA (mRNA) sequences began to grow, it was noticed that synonymous codon usage varied considerably among genes in a species-specific way and that this pattern correlated with gene expression levels (11–14). This link between codon usage variation and gene expression is largely mediated by variation in transfer RNA (tRNA) availability (15–17). These associations have been observed in prokaryotes and eukaryotes and have motivated a body of research premised upon many synonymous substitutions being affected by natural selection (10, 18–22). Consistent with variation in codon usage among genes, synonymous substitution rate is negatively correlated with the degree of unequal codon usage; that is, high codon bias corresponds with lower substitution rates (23–25).

Given the ubiquity of dN/dS analyses, and the widespread evidence of selection on synonymous substitutions in many species, we undertook an approach to dN/dS analysis that introduces a distinction between selected synonymous codons and nonselected (i.e., neutral) synonymous codons. This method allows the estimation of the ratio of the nonsynonymous substitution rate to the rate of strictly neutral synonymous substitutions. We find that in a large order of bacteria, the Enterobacterales, the strictly neutral synonymous rate is about 25% higher on average than the selected synonymous substitution rate. This difference causes a conventional dN/dS analyses to return estimates that are higher by a similar proportion than when the analysis is conducted using our method that incorporates rates for both neutral and selected synonymous codons.

Results

Multiclass Synonymous Substitution Model. The multiclass synonymous substitution (MSS) model is an extension of the Muse–Gaut 94 (MG94) codon-substitution model (7). This model allows for mutational biases that can have a large effect on base composition (26), on a per gene basis. Codon substitutions are modeled as a reversible time-homogeneous Markov process, with the infinitesimal rate generator matrix defined as follows. Assume that the set of all codons is partitioned into K disjoint

Significance

The ratio of the nonsynonymous codon-substitution rate to that for synonymous codons (*dN*/*dS*) is widely used to estimate the strength and direction of selection. However, many synonymous changes are under selection, particularly in genes with high expression levels. We developed a model to include both neutral and selected synonymous substitutions and applied it to a large order of bacteria. The analysis revealed that the conventional estimate of *dS* is usually well below the strictly neutral rate (80% on average). It follows that conventional *dN/dS* estimates that are less than 1 substantially underestimate the level of negative selection and that cases of apparent positive selection (*dN/dS* > 1), as well as other applications of *dN/dS*, should be reconsidered.

Author contributions: S.L.K.P. and J.H. designed research; S.R., S.L.K.P., A.W., and J.H. performed research; S.R., S.L.K.P., A.W., and J.H. analyzed data; and S.R., S.L.K.P., and J.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: hey@temple.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2023575118/-/DCSupplemental.

Published May 10, 2021.

nonempty subsets (classes), $C_1, \ldots C_K$; a valid partition further requires that each subset include at least two synonymous codons. Without loss of generality, designate the first codon subset, C_1, as strictly neutral. The instantaneous rate of substitution from codon X, composed of nucleotides $x_1x_2x_3$, to codon Y ($y_1y_2y_3$) is given by one of the following expressions:

1) 0, if X and Y differ by more than one nucleotide.
2) $\theta_{xy}\pi_y$, if $X, Y \in C_1$, X to Y is a synonymous change, and x and y are the nucleotides being substituted.
3) $\boldsymbol{\alpha_k}\theta_{xy}\pi_y$, if $X, Y \in C_k$, X to Y is a synonymous change.
4) $\boldsymbol{\alpha_{kl}}\theta_{xy}\pi_y$, if $X \in C_k$, $Y \in C_l$, X to Y is a synonymous change.
5) $\boldsymbol{\omega}\ \theta_{xy}\pi_y$, if X to Y is a nonsynonymous change.

Here, $\boldsymbol{\alpha_k}$ is the substitution rate for synonymous codons in class C_k relative to the strictly neutral substitution rate (for class C_1). $\boldsymbol{\omega}$ is the standard nonsynonymous to synonymous rate ratio, with the important distinction that the synonymous rate here is defined to encompass only strictly neutral codons (in class C_1). θ_{xy} represents the nucleotide substitution rate, and because of reversibility, $\theta_{xy} = \theta_{yx}$. Because of standard identifiability issues, we set one of the θ_{xy} ($e.g.$, θ_{AG}) = 1. π_y denotes the equilibrium frequency for the target nucleotide; each position in the codon gets its own set of frequencies. Frequencies are obtained from empirical counts using the CF3 × 4 estimator, which corrects biases introduced by the absence of stop codons in coding data (27).

When $K = 1$, the MSS model reduces to the standard MG94 model with the general time reversible–nucleotide component. When $K > 1$, the MSS model includes $K - 1$ additional parameters, including α_k for relative rates for synonymous substitutions within a "non-neutral" synonymous codon class C_k, and a variable number of parameters, which represent relative rates for synonymous substitutions between two synonymous codon classes C_k and C_l (i.e., α_{kl}). The number of the parameters is variable because the number of codon classes that can be connected using one nucleotide synonymous changes depends on the partitioning of the codons.

Given an alignment and a fixed tree topology, we can fit the MSS model together with branch lengths using maximum likelihood. With $K + L - 1$ additional synonymous class parameters, we can conduct $K + L$ likelihood ratio tests to determine which, if any, of the α rates are significantly different from 1 (the strictly neutral rate). For multiple tests, we apply the Holm–Bonferroni correction, which ensures that the family wise error rate is controlled. For each rate parameter, we also estimate an approximate CI using profile likelihood. The model is implemented in the HyPhy batch language (9). The computational complexity of the model is of the same order as that of the standard MG94 model, and it can be fitted to alignments of thousands of sequences on a standard multicore computer.

MSS Model Analyses. To identify the set of codon substitutions least affected by selection, we reasoned that selection for translational efficiency or accuracy will cause patterns of covariation in codon usage across genes, with the subset of codons that most strongly facilitate translation all tending to become more common in highly expressed genes and those that slow or otherwise impede translation becoming less common in highly expressed genes. Then, if translational selection is the primary cause of codon frequency covariation, those codons with frequencies that show the least covariation across genes will be those least impacted by selection. To identify the dominant pattern of covariation in codon frequencies, we conducted a factor analysis on the codon frequencies for the 18 amino acids with multiple codons for annotated genes in each genome of the study. This analysis reveals, for each genome, the loadings for each codon on the dominant axis of covariation in codon frequencies (18, 28). Codons with negative loadings become less frequent with respect to the dominant pattern of covariation across genes (Factor 1, denoted F1), while codons with positive loadings become more frequent. If the pattern of covariation is caused by selection, then it is the magnitude of the association that signals the strength of selection, and the actual sign of F1 loading is not informative (i.e., codons with positive loadings simply become more common in some genes and less in others, with the reverse pattern for codons with negative loadings). Codons were sorted by mean absolute value of F1 loading for each of the genomes to identify those that consistently showed little covariation and could thus serve as presumptive neutral codons.

Analyses were conducted with 13 genomes that were selected to sample the breadth and depth of the order Enterobacterales (29). The rankings of codons, by absolute F1 loadings, were similar across genomes, and they revealed two fourfold redundant amino acids, alanine and valine, that consistently had all codons among those with the lowest ranks. These eight codons were placed into the neutral group and the remainder in the selected group. Since there are no synonymous substitutions possible between the two groups defined this way, the MSS model as described above has just one additional parameter relative to the standard MG94 model, which is the relative rate of synonymous substitution in the selected codon group, α_s.

The MSS model invokes a partition of the synonymous substitution rate dS into two components: the neutral synonymous rate, dSn, and the selected synonymous rate, dSs; it yields estimates of the ratio of the nonsynonymous rate to the synonymous neutral rate, dN/dSn. For comparison, we also analyzed each gene under the MG94 model (7), which provides the "standard" dN/dS estimates, and we simulated data sets under the MG94 model for each gene using the estimated dN/dS values obtained with MG94 on the same gene. If the model performs as expected, then under a simulation in which all synonymous codons are neutral ($\alpha_s = 1$), the estimated dN/dSn rate should be the same as the estimated dN/dSn rate. We then estimated dN/dSn values under the MSS model for each set of simulated data. Fig. 1*A* shows dN/dSn estimates for each of the 1,613 genes plotted against the corresponding actual estimated dN/dS value, using robust Theil–Sen regression. The regression slope was effectively 1 (actual value 1.005, $r^2 = 0.85$), the mean of estimated α_s (i.e., dN/dSn) values was 1.02 (SD 0.15), and 4.8% of the simulated datasets rejected the null hypothesis at $\alpha_s = 1$ (at $P \leq 0.05$)—all nominal behaviors for the test. Note that because we are estimating dSn from a smaller "effective" sample size (only ALA and VAL synonymous changes), we expect there to be a loss of efficiency (higher sampling variance) under the MSS model when the data are generated with the MG94 model. Indeed, the root mean square error for MG94 in estimating dN/dS was 0.0069 and for MSS, it was 0.013.

We also wished to confirm that our factor analysis is identifying a pattern of covariation among codons that covaries with gene expression, which would be expected if gene expression were a key driver of F1. To do this, we used the mean expression level observed in an RNA sequencing (RNA-seq) study of *Escherichia coli* (30) and calculated the rank correlation between codon frequency and the reported measure of gene expression (the geometric mean of FPKM [Fragments Per Kilobase of transcript per Million mapped reads] across different samples) across all of the genes. Fig. 1*B* shows a strong correlation of these values with F1 loadings for the 59 codons ($r = -0.79$). A correlation is expected if the dominant pattern of covariation captured by factor analysis is associated with gene expression, as those codons that covary the most with gene expression should also show the most extreme F1 loadings (the actual sign of the association, which happens to be negative in Fig. 1*B*, is not informative). Highlighted in Fig. 1*B* are ALA and VAL codons, all of which are near the origin on both axes. Fig. 1*C* shows the FPKM values for each gene plotted against the F1 scores. Again, an association is expected if gene

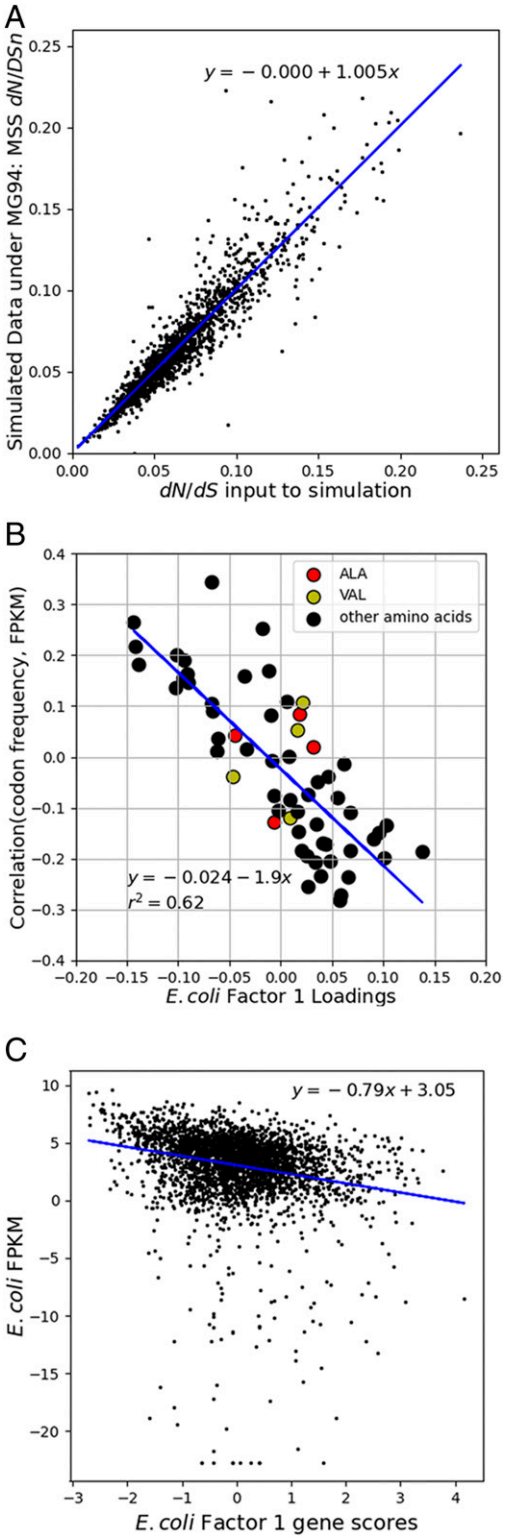

Fig. 1. (*A*) Results of MSS analysis of data simulated under MG94. (*B*) Association between the correlation of codon frequency and gene expression (FPKM) and F1 loadings for *E. coli* genes (r = −0.80, p < 0.00001). (*C*) Association between gene expression (FPKM) and *E. coli* F1 gene scores $r = 0.23$, $P < 0.000001$.

expression is driving the dominant pattern of codon frequency covariation, as high (or low) gene expression will cause a gene to use primarily a subset of codons with high absolute loading on F1, and this will cause that gene to have a high absolute F1 score. The association is not as clear as for codon frequencies but is highly statistically significant ($r = 0.23$, $p < 0.000001$). This pattern is fully consistent with previous reports of correlation between codon frequencies and gene expression in *E. coli* (31). Thus, despite high levels of linkage disequilibrium and clonal interference in natural populations (32), evidence strongly supports selection on codon usage in *E. coli*.

The MG94 model is nested within the MSS model, and we can assess the effect of including two classes of synonymous substitutions directly by comparing estimates of dN/dS (MG94) with dN/dSn (MSS). The nonsynonymous rate in the numerator of both models is the same (MSS and MG94 return nearly identical values of nonsynonymous tree lengths). Fig. 2*A* shows this relationship, revealing a strong linear component with a slope of 0.785 ($r^2 = 0.838$, using the Theil–Sen regression), and Fig. 2*B* shows the distribution of estimates under both models. We find that values of dS/dSn (calculated by taking the ratio of estimates of dN/dSn and dN/dS) have a mean value of 0.800 (95% CI 0.715 to 0.878, Fig. 2*C*). Values less than 1 are expected if in fact $dS < dSn$ because dS results from a mixture of neutral and selected synonymous substitutions. The hypothesis of neutrality of selected synonymous substitutions was rejected by the likelihood ratio test (MG94 versus MSS, 1 df, $p < 0.05$) for 919/1,613 genes (57%) or for 846/1,613 genes (52.4%) using the Benjamini–Hochberg false discovery rate procedure at $q = 0.05$ (33). In all but 6 of the 919 significant test cases, dSs/dSn was less than 1. The MSS model was also preferred to the MG94 model by an information theoretic goodness of fit criterion [small sample Akaike information criterion (AICc) (34)] for 1,222/1,613 (75.8%) of the alignments, with a median AICc improvement of 6.01 points (Fig. 2*D*).

The choice of codons to be placed in the neutral set has a strong impact on rate inference. To illustrate this, we applied an alternative MSS model using two other fourfold degenerate amino acids (threonine and proline) with high absolute F1 loadings (*SI Appendix*, Table S2). As expected, the mean estimate of dSs/dSn under this model is greater than 1 (*SI Appendix*, Fig. S3) because the proposed neutral set is now evolving at a rate slower than the average synonymous substitution; this model has a worse fit to the data (compared to the Alanine–Valine model using AICc) on 1,263/1,613 (78.3%) of the alignments.

To assess the impact of using the MSS model on inferences of positive selection on nonsynonymous sites, we adapted the BUSTED (Bayesian UnreStricted Test for Episodic Diversification) method (35) for identifying episodic diversifying selection to include an alignment-wide parameter for the dSs/dSn ratio. BUSTED fits a random-effects model to sites and branches, where multiple dN/dS ratios are drawn independently from a three-bin discrete distribution of values. The likelihood function is computed by summing over all possible assignments of rates to branches/sites, and hyperparameters of the rate distribution (three values of dN/dS and two weight parameters) are estimated by maximum likelihood. Positive selection is inferred if the unrestricted model has a nonzero weight assigned to the positive rate class (denoted $dN/dS > 1$) and if the likelihood ratio test with a nested null model (in which the dN/dS values for all rate classes are not greater than 1) returns a significant result. Under BUSTED with MSS, these analyses are conducted with respect to dN/dSn rather than dN/dS. We assessed the false positive rate for 400 alignments simulated under strict neutrality ($dN/dS = 1$) under BUSTED-MSS using fits to 10 randomly selected bacterial alignments to obtain other simulation parameters (branch lengths, nucleotide biases, and alignment lengths). BUSTED-MSS did not show excessive false positive rates, with the empirical false positive rate of 0.037 at $P = 0.05$ and <0.003 at $P = 0.01$.

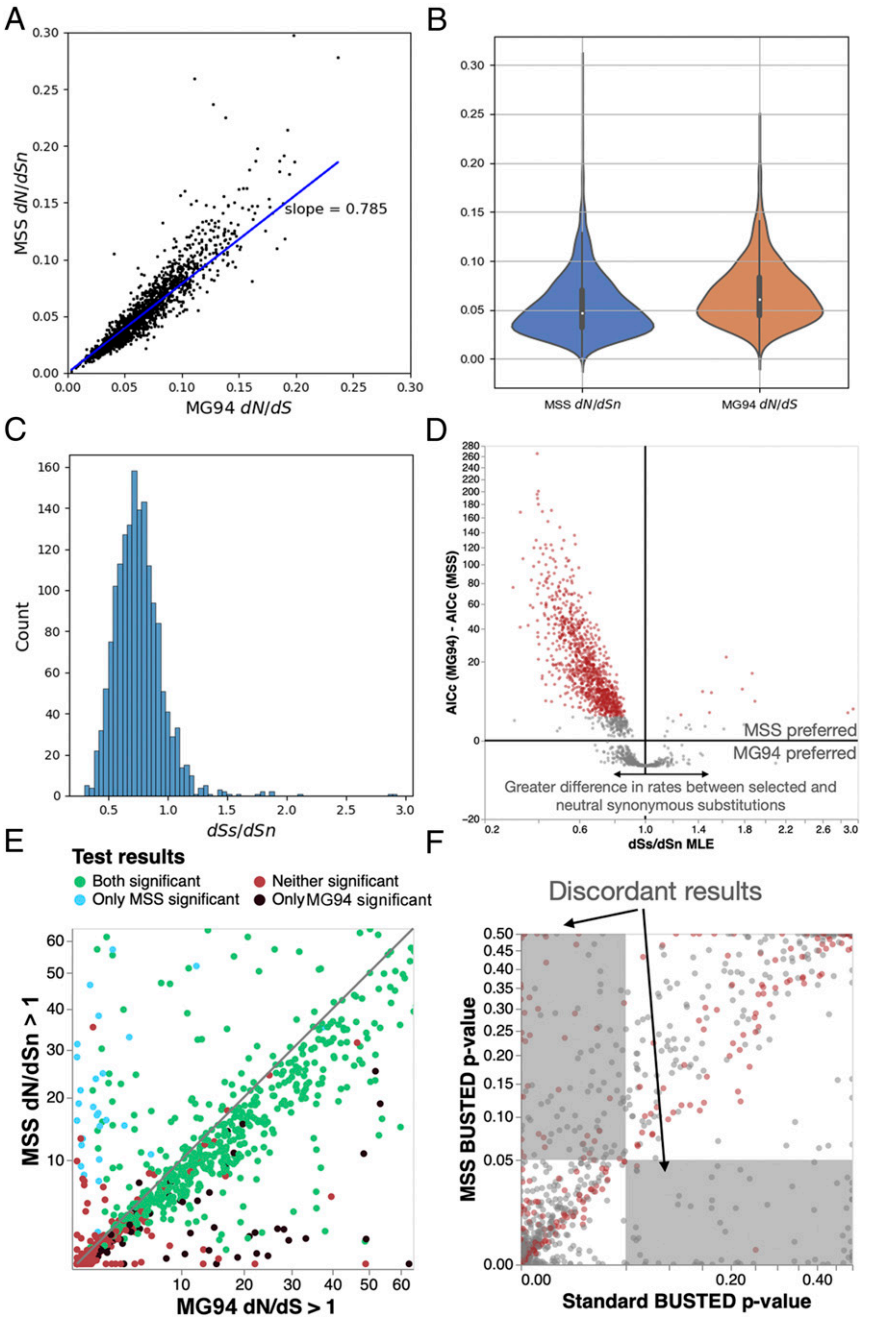

Fig. 2. (*A*) The results of the MSS model dN/dSn compared to the MG94 (dN/dS) model across 1,613 gene alignments. (*B*) Violin plots of the values used for *A*. (*C*) Histogram of the ratio of the selected synonymous rate to the neutral synonymous rate for 1,613 genes. (*D*) The relationship between dSs/dSn and small sample Akaike information criterion (AICc) differences between mG94 and MSS; red points represent datasets where the likelihood ratio test rejected the hypothesis of neutrality of selected synonymous substitutions ($P \leq 0.05$). (*E*) Point estimates for the dN/dS ratio of the positive selection component of the model, classified by how the two models classified the gene. (*F*) *P* values for the episodic selection likelihood ratio test. Points shown in red are for the datasets where the BUSTED under MG94 is preferred to BUSTED under MSS by AICc.

We compared gene classifications made by BUSTED under both MSS and MG94 and observed that consistent with expectations, the estimated positive selective component had a lower maximum likelihood point estimate of dN/dSn under MSS, relative to dN/dS under MG94, for the majority of genes (1,083 genes, 67.1%, Fig. 2*E*). There was also an increase in *P* values for the positive selection test under the MSS model, with 1,151 (71.4%) genes showing this pattern, suggesting that the evidence for positive selection is generally weaker under the MSS model.

Under BUSTED, the MG94 and MSS models differed on findings of selection in 164 (10.2%) of the datasets, with MSS finding no evidence of selection in 109 datasets, when BUSTED with MG94 did, and 55 datasets where the opposite held (Fig. 2*F*). The differences include a roughly equal number of genes in which very strong support was found under one model but not the other (*P* value went from $P < 0.001$ to $P > 0.2$). There were six cases where selection was indicated with MG94 and little support was found with MSS and nine cases in the other direction (*SI Appendix*, Table S3).

We can use the relationship between MSS and MG94 models to estimate a quantity θ; the proportion of synonymous substitutions that are neutral. Thus, we let dS be apportioned into neutral and selected components, $dS = \theta\, dSn + (1-\theta)dSs$. For each gene, we have direct estimates of dSs/dSn (Fig. 2*C*), which are given by the α_s parameter, and, with ω ratio estimates from the MSS (dN/dSn) and MG94 (dN/dS) models, we can compute $\frac{dN}{dSn}/\frac{dN}{dS} = \frac{dS}{dSn} = \frac{\theta\, dSn + (1-\theta)dSs}{dSn}$. An overall estimate of θ can be obtained by casting this as a linear regression of form $\frac{dS}{dSn} = \left(\theta - \frac{dSs}{dSn}\right)/(\theta - 1)$, which yields a nearly perfect line with $\theta = 0.1896$ (Fig. 3*A*). Thus, about 81% of the synonymous substitutions that occur are in the selected class, on average. The same regression performed on 10,000 bootstrap samples yields a 95% percentile interval on the θ estimate of 0.1844 to 0.1958. The value of 81% appears reasonable given the fact that the eight codons in the neutral class are for two common fourfold degenerate amino acids (i.e., alanine and valine, 12 total synonymous pairs) and that the remaining degenerate amino acids are a mix of two-, four-, six- and threefold redundancies (75 total synonymous pairs); that is, about 84% of all synonymous pairs are in the selected class.

Estimating the Strength of Selection on Synonymous Substitutions. Unlike nonsynonymous substitutions, synonymous substitutions have only a small number of alternative states at any position. This, and the fact that a codon can only be favored (or disfavored) in relation to another synonymous codon, suggests that the mode of selection on non-neutral synonymous substitutions will be roughly equal parts positive and negative. Here, we develop expressions for relative (selected to neutral) substitution rates for a simple haploid model. Consider a twofold redundant amino acid that alternates substitutions between a favored codon (selection coefficient s) and a disfavored codon ($-s$). Then, assuming that $s \ll 1$, the probability of fixation is approximately $p(s) = \frac{2s}{1-e^{-2Ns}}$ (36), where N is the effective population size. Given a mutation rate u, the mean waiting time to the next substitution, with alternating favored and disfavored substitutions, will be $\frac{1}{2}\left(\frac{1}{u\,p(s)} + \frac{1}{u\,p(-s)}\right)$. The corresponding substitution rate for a strictly neutral twofold redundant site will be u/N. Thus, the ratio of selected to neutral substitution rates for twofold sites is given by the following:

$$\frac{dSs_{(2fold)}}{dSn_{(2fold)}} = \frac{2}{\left(\frac{1}{u\,p(s)} + \frac{1}{u\,p(-s)}\right)} \Bigg/ \frac{u}{N} = \frac{4Nse^{2Ns}}{e^{4Ns}-1}.$$

For fourfold sites, we consider the case of two favored and two disfavored codons, so that one-third of the mutations would be strictly neutral (i.e., mutations within either class) and two-thirds would be selected. The mean waiting time to the next neutral change would be $3\,N/u$, while the mean waiting time to the next selected change would be $\frac{3}{4}\left(\frac{1}{u\,p(s)} + \frac{1}{u\,p(-s)}\right)$. Then, total substitution rate at this site (neutral and selected) would be the sum of the inverses of these values, and the ratio of this to the corresponding rate at a strictly neutral fourfold site is given by the following:

$$\frac{dSs_{(4fold)}}{dSn_{(4fold)}} = \left(\frac{u}{3N} + \frac{1}{\frac{3}{4}\left(\frac{1}{u\,p(s)} + \frac{1}{u\,p(-s)}\right)}\right) \Bigg/ \frac{u}{N} = \frac{e^{4Ns} + 8Nse^{2Ns} - 1}{3(e^{4Ns}-1)}.$$

For both twofold and fourfold sites, this simple model predicts the rate of synonymous substitution under selection relative to that of a strictly neutral site as a function of Ns, which is the product of effective population size and selection coefficient.

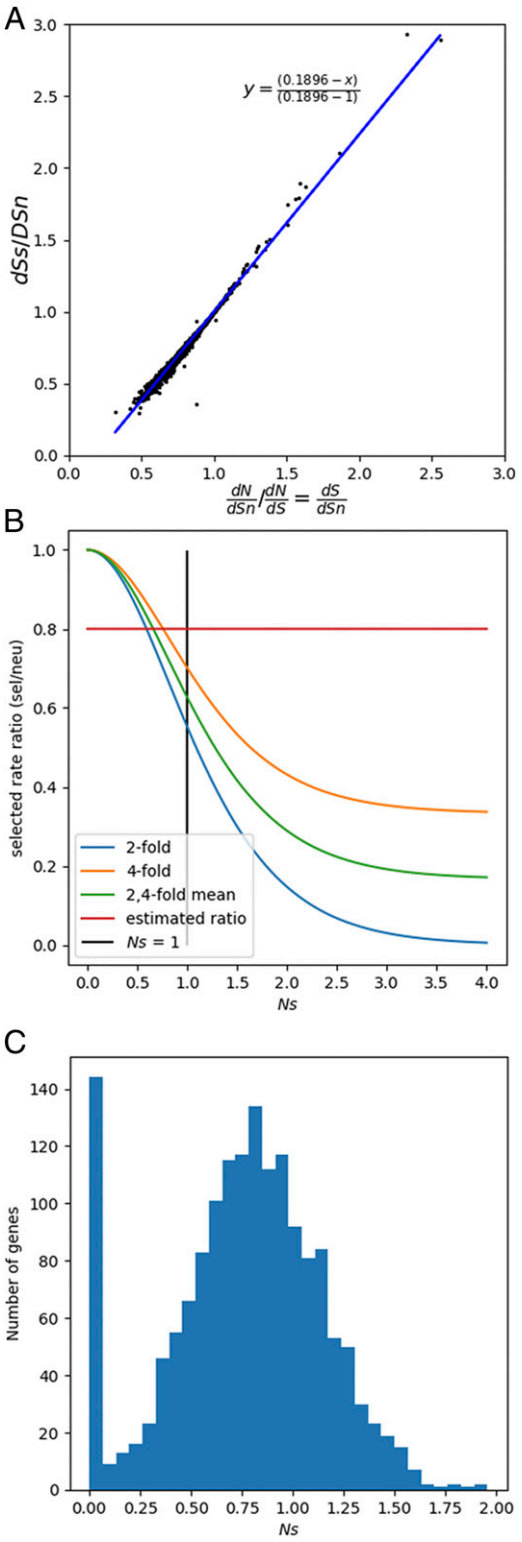

Fig. 3. (*A*) The ratio of selected to neutral rates from the MSS model plotted against the ratio of the MSS model dN/dSn to the MG94 dN/dS. (*B*) Plots of the selected rate ratio for three classes of codons (twofold, fourfold, and mixture). (*C*) Histogram of estimated Ns values obtained using the mixture model as a function of observed estimates of dSs/dSn for 1,613 genes.

Fig. 3*A* shows a plot of these relative rates as a function of Ns for both models and for an average of the two models. The average can be taken as an approximation for what would be expected for

a gene that is a roughly even mixture of twofold and fourfold sites, as well as sixfold sites (which are themselves similar to a mix of twofold and fourfold sites).

The striking feature of the model predictions is that the amount of selection needed to account for our estimated values of the ratio of the selected synonymous substitution rate to the neutral synonymous substitution rate is quite small. Those estimated rates had a mean of about 0.8, corresponding to rate ratios less than 1. If we treat the rate ratio for actual genes as falling under the mixture model (i.e., a mix of twofold and fourfold sites), we can estimate *Ns* for each of the genes using the maximum likelihood estimates (MLEs) of the selected to neutral ratio. As shown in Fig. 3*B*, the mode of the inferred ratio distribution is at 0.8, with a few genes approaching $Ns = 2.0$ (the peak at $Ns = 0$ corresponds to those genes with an estimate of dSs/dSn greater than or equal to 1). By applying the same mapping of the 95% CI of the selected to neutral ratios to the mixture model, the corresponding CI of the *Ns* estimates had a mean width of 0.24.

Discussion

Our understanding of selection on nonsynonymous substitutions is based on properly defining the neutral reference. Here, we show that a neutral reference based on a carefully selected subset of synonymous substitutions provides a significant improvement of fit for the majority of genes and an overall reduction in the estimated relative rate of nonsynonymous substitutions. If the overall synonymous rate in these bacteria is actually lower than for strictly neutral substitutions (e.g., 80% of the rate that has been estimated under the MG94 model where all synonymous codons are neutral), then the necessary corollary is that purifying selection on nonsynonymous changes is about 25% more effective than previously thought (i.e., $1/0.8 = 1.25$). This calibration will also affect some cases of apparently high adaptive substitution rates (i.e., $dN/dS > 1$), which may fall below 1 if dSn were the denominator. We did not observe any cases of such high gene-wide dN/dS in this study, but we did find that an analysis of episodic positive selection had a substantial effect, with about 1 in 11 findings of positive selection found under MSS that were not inferred under MG94, as well as many findings of the reverse pattern.

The approach of using patterns of covariation in codon frequencies across genes will classify codons without regard to the underlying causes of that covariation. These are expected to include selection on gene expression levels, with high expression genes having more selection on codon usage. It may also include effects related to gene length, such that longer genes are selected for a certain pattern of codon usage, consistent with translational accuracy (37, 38). However, to the extent that substitutions between alanine codons and between valine codons are not strictly neutral, our estimate of dN/dSn will be an overestimate because the denominator, dSn, is an underestimate of the strictly neutral rate. Our approach does not address the fact that individual synonymous changes can, under some circumstances, come under strong selection. For example, synonymous substitutions may come under strong selection due to transcript secondary structure (39, 40), or in the case of eukaryotes, the effects on intron splicing (41, 42).

The difference we have estimated between dS and dSn can be accounted for by selection on synonymous variation that is extremely weak, with a mean *Ns* value of 0.8. It is noteworthy that this value is less than 1, the value of *Ns* at which selection is conventionally taken to become effective (43, 44). Our results are consistent with previous modeling, which show that even selection as weak as $Ns \sim 1$ can have a large effect on synonymous substitution patterns and thus codon bias (45, 46).

One class of analyses that may suffer from a strong reliance on an assumption of strict neutrality for synonymous substitutions are those that take dN/dS estimates as indicators of the effective population size, N (47–49). As population size increases, more efficacious selection is expected to lower dN and thus dN/dS if selection on nonsynonymous mutations is largely purifying. But if synonymous substitutions are also selected, then the ratio may not be very sensitive to changes in population size. If most synonymous substitutions actually have slight effects on fitness, then both dN and dS will covary negatively with effective population size, and the response of dN/dS to changes in population size will depend on the relative numbers of the different types of mutations that fall near the margins of what can be modulated by selection. We can estimate the impact on estimates of N by drawing upon the relation $\frac{dN}{dS} = \frac{2s}{1-e^{-2Ns}}$ (50, 51). Then, a scalar f can be estimated from $\frac{dN}{dS}/\frac{dN}{dSn} = \left(\frac{2Ns}{1-e^{-2Ns}}\right)/\left(\frac{f\,2Ns}{1-e^{-f\,2Ns}}\right)$. With our estimated mean of the ratio on the left of 0.8, we obtain $\hat{f} = 1.069$. Thus, the reduction in the estimated substitution ratio corresponds, on average, to about a 7% inflation of the *Ns* when using the MG94 compared to what is obtained under the MSS model.

Another type of analysis that depends strongly on dN/dS is the estimate of the adaptive substitution rate or α, using the relation $\alpha = 1 - \frac{dS\,pN}{dN\,pS}$ (52), where pN and pS are counts of polymorphic sites (nonsynonymous and synonymous, respectively) within species. An estimate of $\alpha = 0.56$ has previously been reported for *E. coli* (53). Given our mean estimates of dN/dS under MG94 [0.067, similar to the value of 0.074 estimated along with α for *E. coli* (53)], we obtain an estimate of polymorphism ratio, $pN/pS = 0.0386$. With this ratio and using the mean of dN/dSn under MSS (0.052), we obtain the following:

$$\alpha_{MSS} = 1 - \frac{dSn\,pN}{dN\,pS} = 0.26.$$

This example yields a 54% reduction in α (i.e., 0.56 versus 0.26), suggesting that an improved estimate of relative nonsynonymous substitution rates can have a large effect on estimates of the rate of adaptive amino acid substitutions.

In addition to describing genome-wide selective patterns on synonymous substitutions, the MSS framework is easily extended to other applications. If the partitioning of codons into groups is available using external sources of information, the MSS model can be incorporated into dN/dS based selection tests at the level of genes (35), sites (54), and lineages (55). Recent work in this area has revealed that many standard assumptions, such as constant rates of synonymous substitution among sites (56), absence of instantaneous multinucleotide changes (57), and single dN/dS ratio for all pairs of amino acids (58), are often rejected by biological data. A more precise definition of the neutral subset of synonymous substitutions is likely to have an effect on fine-grain selection inference. Another avenue for development could be learning the sets of synonymous substitutions from the data, via a suitable technique for exploring large sets of discrete models, for example, genetic algorithms (58).

Methods

Fourteen genomes were selected from across the breadth and depth of a phylogeny for 179 members of the Enterobacterales (29), a large order of gram-negative bacteria that includes many enteric species, including *E. coli*. For each genome, a factor analysis was carried out on a table of codon frequencies across genes in order to identify the largest factor of covariation in codon frequencies across genes (18, 28). To prepare this table, each gene for which coding sequences were listed in the genome GenBank file (files links are given in *SI Appendix*, Table S1) was reduced to a set of codon frequencies for the 59 codons of the 18 amino acids that utilize multiple codons (i.e., excluding tryptophan and methionine). Genes were included only if the following criteria applied: 1) coding sequence length was a multiple of 3; 2) no early stop codons; 3) all 18 amino acids with multiple codons were represented at least once; and 4) there were at least 100 codons in the sequence. Factor analysis was carried out using the factoranalysis module of Scikit-learn version 0.22.1 for the Python programming language,

version 3.7.6. For each codon, the absolute value of the loading on the principal factor was taken to reflect the overall level to which the codon covaried in frequency with other codons and thus reflect the action of those sources of natural selection that cause such patterns of covariation. *SI Appendix*, Fig. S1 shows these factor loadings for all of the genomes, with codons sorted by their mean loading across species. For the most part, the loadings are strongly correlated across different genomes (*SI Appendix*, Fig. S2). As shown in *SI Appendix*, Figs. S1 and S2, the loadings for *Shimwellia blattae* showed much less similarity to the others (lower pairwise correlations), and this genome was excluded from further analyses.

If, as often reported, variation among genes in expression levels is a major driver of patterns of covariation in codon frequencies among genes (19, 59), then the F1 loadings should be correlated with gene expression. To examine this, we used FPKM values recorded from RNA-seq analysis in *E. coli*. We focused on data from a study that recorded RNA-seq data for *E. coli* K-12 substrain MG1655 under 28 conditions (30). For each of 3,233 genes, the geometric mean of FPKM was calculated across the 28 RNA-seq analyses. Then, for each of the 59 codons, the Spearman correlation between the geometric mean FPKM for each gene and codon frequency for each gene was calculated and plotted against F1 loadings for each codon (Fig. 1*B*).

From the F1 loadings, we identified those amino acids that had multiple codons showing the lowest mean absolute factor loadings. Synonymous substitutions between these codons are candidates for not being subject to the kinds of selection that affect codon usage, while in contrast synonymous substitutions among codons with high absolute loadings are expected to be those that drive the patterns of covariation detected by factor analysis. However, an important limitation of this approach is that it highlights codons themselves and not substitutions between codons. A codon per se can be neither neutral nor selected. For this reason, we focused on amino acids for which all of the codons have low rankings, for it is in these cases that we expect that all of the substitutions between the codons would be neutral. As shown in *SI Appendix*, Table S2, the two amino acids with the lowest mean ranking among their respective codons were both fourfold degenerate (alanine and valine).

For each gene that was represented in at least four of the 13 genomes (1,613 genes), alignments were generated for the amino acid sequences using MAFFT (60) with the Blosum80 substitution matrix and the "globalpair" option. Phylogenies were then estimated for each gene using RAXML-NG (61) with the GTR+G model and the MSS model fitted to the alignment and the phylogeny using HyPhy 2.5 (9, 27). To assess the quality of rate estimates under the MSS model, data were simulated under the MG94 model with true dN/dS values using HyPhy (Fig. 1*A*).

Data Availability. The MSS code is available at https://github.com/veg/hyphy-analyses/tree/master/MulticlassSynonymousSubstitutions (62). Alignments, tree topologies, parameter estimates and other values used in preparation of this paper are available from https://data.hyphy.org/web/MSS (63).

ACKNOWLEDGMENTS. J.H. was supported by US NSF Grant 1564659. S.L.K.P. was supported by the following grants from the US NIH: R01 AI134384 (NIH/National Institute of Allergy and Infectious Diseases [NIAID]), R01 AI140970 (NIH/NIAID), and R01 GM093939 (NIH/National Institute of General Medical Sciences [NIGMS]).

1. T. Miyata, T. Yasunaga, Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16**, 23–36 (1980).
2. M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
3. W. H. Li, C. I. Wu, C. C. Luo, A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**, 150–174 (1985).
4. M. Kimura, Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276 (1977).
5. R. Nielsen, Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218 (2005).
6. N. Goldman, Z. Yang, A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).
7. S. V. Muse, B. S. Gaut, A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**, 715–724 (1994).
8. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
9. S. L. Kosakovsky Pond *et al.*, HyPhy 2.5-A customizable platform for evolutionary hypothesis testing using phylogenies. *Mol. Biol. Evol.* **37**, 295–299 (2020).
10. S. L. K. Pond, S. V. Muse, "HyPhy: Hypothesis testing using phylogenies" in *Statistical Methods in Molecular Evolution*, R. Nielsen, Ed. (Springer, 2005), pp. 125–181.
11. R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, R. Mercier, Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9**, r43–r74 (1981).
12. G. Chavancy, A. Chevallier, A. Fournier, J. P. Garel, Adaptation of iso-tRNA concentration to mRNA codon frequency in the eukaryote cell. *Biochimie* **61**, 71–78 (1979).
13. T. Ikemura, 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. *J. Mol. Biol.* **151**, 389–409 (1981).
14. J. L. Bennetzen, B. D. Hall, Codon selection in yeast. *J. Biol. Chem.* **257**, 3026–3031 (1982).
15. R. Sabi, T. Tuller, Modelling the efficiency of codon-tRNA interactions based on codon usage bias. *DNA Res.* **21**, 511–526 (2014).
16. G. Hanson, J. Coller, Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.* **19**, 20–30 (2018).
17. E. P. Rocha, Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* **14**, 2279–2286 (2004).
18. J. Hey, R. M. Kliman, Interactions between natural selection, recombination and gene density in the genes of Drosophila. *Genetics* **160**, 595–608 (2002).
19. J. B. Plotkin, G. Kudla, Synonymous but not the same: The causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42 (2011).
20. L. Duret, Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**, 640–649 (2002).
21. E. Lebeuf-Taylor, N. McCloskey, S. F. Bailey, A. Hinz, R. Kassen, The distribution of fitness effects among synonymous mutations in a gene under directional selection. *eLife* **8**, e45952 (2019).
22. N. Galtier *et al.*, Codon usage bias in animals: Disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. *Mol. Biol. Evol.* **35**, 1092–1103 (2018).
23. P. M. Sharp, W. H. Li, The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* **4**, 222–230 (1987).
24. G. Marais, T. Domazet-Lošo, D. Tautz, B. Charlesworth, Correlated evolution of synonymous and nonsynonymous sites in Drosophila. *J. Mol. Evol.* **59**, 771–779 (2004).
25. A. E. Hirsh, H. B. Fraser, D. P. Wall, Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Mol. Biol. Evol.* **22**, 174–177 (2005).
26. H. Long *et al.*, Evolutionary determinants of genome-wide nucleotide composition. *Nat. Ecol. Evol.* **2**, 237–240 (2018).
27. S. L. Kosakovsky Pond, W. Delport, S. V. Muse, K. Scheffler, Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS One* **5**, e11230 (2010).
28. R. M. Kliman, N. Irving, M. Santiago, Selection conflicts, gene expression, and codon usage trends in yeast. *J. Mol. Evol.* **57**, 98–109 (2003).
29. M. Adeolu, S. Alnajar, S. Naushad, R. S. Gupta, Genome-based phylogeny and taxonomy of the 'Enterobacteriales': Proposal for Enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morganellaceae fam. nov., and Budviciaceae fam. nov. *Int. J. Syst. Evol. Microbiol.* **66**, 5575–5599 (2016).
30. Y. Gao *et al.*, Systematic discovery of uncharacterized transcription factors in Escherichia coli K-12 MG1655. *Nucleic Acids Res.* **46**, 10682–10696 (2018).
31. P. M. Sharp, W. H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**, 28–38 (1986).
32. T. S. Whittam, H. Ochman, R. K. Selander, Multilocus genetic structure in natural populations of Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1751–1755 (1983).
33. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
34. N. Sugiura, Further analysts of the data by Akaike's information criterion and the finite corrections: Further analysts of the data by Akaike's. *Commun. Stat. Theory Methods* **7**, 13–26 (1978).
35. B. Murrell *et al.*, Gene-wide identification of episodic selection. *Mol. Biol. Evol.* **32**, 1365–1371 (2015).
36. M. Kimura, On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719 (1962).
37. A. Eyre-Walker, Synonymous codon bias is related to gene length in Escherichia coli: Selection for translational accuracy? *Mol. Biol. Evol.* **13**, 864–872 (1996).
38. N. Stoletzki, A. Eyre-Walker, Synonymous codon usage in Escherichia coli: Selection for translational accuracy. *Mol. Biol. Evol.* **24**, 374–381 (2007).
39. J.-V. Chamary, L. D. Hurst, Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* **6**, R75 (2005).
40. L. Katz, C. B. Burge, Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.* **13**, 2042–2051 (2003).
41. R. Savisaar, L. D. Hurst, Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res.* **28**, 1442–1454 (2018).
42. J. L. Parmley, L. D. Hurst, Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol. Biol. Evol.* **24**, 1600–1603 (2007).
43. T. Ohta, Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor. Popul. Biol.* **10**, 254–275 (1976).
44. T. Ohta, Population size and rate of evolution. *J. Mol. Evol.* **1**, 305–314 (1972).

45. W. H. Li, Models of nearly neutral mutations with particular implications for non-random usage of synonymous codons. *J. Mol. Evol.* **24**, 337–345 (1987).
46. G. A. McVean, B. Charlesworth, A population genetic model for the evolution of synonymous codon usage: Patterns and predictions. *Genet. Res.* **74**, 145–158 (1999).
47. L.-M. Bobay, H. Ochman, Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol. Biol.* **18**, 153 (2018).
48. P. S. Novichkov, Y. I. Wolf, I. Dubchak, E. V. Koonin, Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. *J. Bacteriol.* **191**, 65–73 (2009).
49. C.-H. Kuo, N. A. Moran, H. Ochman, The consequences of genetic drift for bacterial genome complexity. *Genome Res.* **19**, 1450–1454 (2009).
50. R. Nielsen, Z. Yang, Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol. Biol. Evol.* **20**, 1231–1239 (2003).
51. S. Kryazhimskiy, J. B. Plotkin, The population genetics of dN/dS. *PLoS Genet.* **4**, e1000304 (2008).
52. A. Eyre-Walker, The genomic rate of adaptive evolution. *Trends Ecol. Evol.* **21**, 569–575 (2006).
53. J. Charlesworth, A. Eyre-Walker, The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.* **23**, 1348–1356 (2006).
54. B. Murrell *et al*., Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**, e1002764 (2012).
55. M. D. Smith *et al*., Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32**, 1342–1353 (2015).
56. S. R. Wisotsky, S. L. Kosakovsky Pond, S. D. Shank, S. V. Muse, Synonymous site-to-site substitution rate variation dramatically inflates false positive rates of selection analyses: Ignore at your own peril. *Mol. Biol. Evol.* **37**, 2430–2439 (2020).
57. A. Venkat, M. W. Hahn, J. W. Thornton, Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat. Ecol. Evol.* **2**, 1280–1288 (2018).
58. W. Delport *et al*., CodonTest: Modeling amino acid substitution preferences in coding sequences. *PLoS Comput. Biol.* **6**, e1000885 (2010).
59. R. Hershberg, D. A. Petrov, Selection on codon bias. *Annu. Rev. Genet.* **42**, 287–299 (2008).
60. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
61. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
62. S. L. Kosakovsky Pond, Hyphy analyses - multiclass synonymous substitutions. GitHub. https://github.com/veg/hyphy-analyses/tree/master/Multi-classSynonymousSubstitutions. Accessed 29 April 2021.
63. S. Rahman, S. L. Kosakovsky Pond, A. Webb, J. Hey, Weak selection on synonymous codons substantially inflates dN/dS estimates in bacteria. https://data.hyphy.org/web/MSS/. Accessed 29 April 2021.