# The relationship between dN/dS and scaled selection coefficients

**Stephanie J. Spielman** * **and Claus O. Wilke** *

*Department of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute of Cellular and Molecular Biology. The University of Texas at Austin, Austin, TX 78712, USA.

Submitted to Proceedings of the National Academy of Sciences of the United States of America

**Inferring the strength of natural selection in protein-coding sequences along a phylogeny is a major objective in the field of molecular evolution. Two broad modeling approaches have emerged to classify the selevtive pressure in protein-coding sequences. First, mechanistic codon models, which estimates the evolutionary rate ratio $dN/dS$, have seen tremendous success and are widely-used in the scientific community. A second class of models, mutation-selection-balance (MutSel) models, which explicitly model the evolutionary process as a dynamic interplay between mutation and selection, have recently emerged as a popular alternative to mechanistic codon models. These models estimate scaled selection coefficients, which classify the selective response to for all mutations. However, the extent to which these modeling frameworks relate to each other is largely unknown. Do $dN/dS$ estimates yield comparable or distinct information from scaled selection coefficients? To answer this question, we have derived a formal mathematical relationship between these two models' focal parameters. We find that $dN/dS$ can be precisely calculated from scaled selection coefficients, indicating that these models are in complete agreement. Importantly, this relationship allows us to uncover properties, limitations, and biases inherent to this modeling frameworks. In particular, we show that MutSel models are inherently unable to describe positive selection and/or adaptive evolution. However, we find that, in the presence of synonymous codon fitness differences, it is possible to achieve $dN/dS > 1$, even though positive selection is not occuring. Moreover, we have found that mechanistic codon models produce systematically biased $dN/dS$ estimates in the presence of asymmetric mutation rates, and therefore cannot properly accomodate nucleotide compositional bias. summary sentence about the awesomeness of our approach goes here.**

mechanistic codon models | dN/dS | mutation-selection-balance models | scaled selection coefficients | Markov models of sequence evolution

## Introduction

Over the years, various methods have been used to calculate the strength of natural selection acting on protein-coding sequences. Traditionally, the focus has been on estimating the evolutionary rate ratio, $dN/dS$, the rate of nonsynonymous to synonymous substitution rates. This metric indicates how quickly a protein's constituent amino acids change, and is widely used to identify cases of positive, diversifying selection ($dN/dS > 1$) [1, 2, 3, 4]. Following early counting methods for estimating $dN/dS$ (e.g. refs [5] and [6]), mechanistic codon models, which assume an explicit Markov-process model of sequence evolution (see ref. [7] for a comprehensive review), have taken a leading role as the inference method of choice since their introduction in the 1990s [8, 9, 1]. These models yield maximum likelihood estimates (MLEs) for the parameter $\omega$, which represents the quantity $dN/dS$, and have seen great success in the field of molecular evolution.

A second class of models, known as mutation-selection-balance (MutSel) models, are increasingly viewed as a popular alternative to mechanistic codon models. The MutSel framework, couched firmly in population genetics theory, models the dynamic interplay between mutation and selection in a protein-coding sequence. MutSel models yield estimates of site-wise scaled selection coefficients $S = 2Ns$, which indicate the extent to which natural selection favors, or disfavors, particular codons or amino acids at a given protein position [10, 11, 12, 13]. Although first introduced over 15 years ago [10], MutSel models have seen little use due to their high computational expense. Recently, however, several computationally tractable model implementations have emerged [14, 15], allowing for the first time the potential for widespread adoption.

Mechanistic codon models have undergone rigorous development in their 20 years of existence and have advanced to high levels of sophistication. These models can accomodate a variety of evolutionary scenarios, including episodic [16, 17] and lineage-specific selection pressures [18, 19, 3], as well as incorporate information regarding protein structure [20, 21, 22]. This flexibility, in addition to readily-available and accessible software implementations [23, 24, 25], make mechanistic codon models a very attractive modeling choice. On the other hand, have argued that MutSel models, given their firm grounding in population genetics theory and attention to site-specific amino acid fitness differences, offer a more fine-grained approach to studying protein evolution than do mechanistic codon models [10, 11, **?**, 13]. Recent phylogenetic studies have also demonstrated that evolutionary models which explicitly consider amino acid fitness values offer dramatic improvements over mechanistic codon models, suggesting that MutSel models may more aptly represent the evolutionary process [26, 27].

While both mechanistic codon models and MutSel models describe the same fundamental process of protein-coding sequence evolution along a phylogeny, it is unknown how these two modeling classes relate to one another. In particular, as these inference methods have been developed independently, it remains an open question whether or not parameter estimates from one model are comparable to those of the other model. Whether $dN/dS$ estimates from mechanistic codon models yield consistent information with scaled selection coefficients remains unknown. As a consequence, although certain rhetorical arguments may be made in favor of using one method over another, there is currently no formalized, concrete rationale to guide researchers in their methodological choices. Eludicating the relationship between these competing modeling frameworks will reveal under which circumstances the use of these models is justified.

## Reserved for Publication Footnotes

Here, we formalize the relationship between mechanistic codon and MutSel models by examining the extent to which their focal parameters, $dN/dS$ and scaled selection coefficients, yield overlapping information about the evolutionary process. To this end, we derive a mathematical relationship between these models' primary parameters, allowing us to calculate $dN/dS$ values from scaled selection coefficients. Using a simulation approach, we verify that these derived $dN/dS$ values correspond precisely to $\omega$ MLEs inferred using standard mechanistic codon models. In addition, we prove that, when synonymous codons have equal fitness values, $dN/dS$ calculated from selection coefficients is necessarily less than 1, demonstrating that MutSel models are inherently only able to model purifying selection, and therefore would be an inappropriate model choice if positive selection is expected. However, we have also found that, when synonymous codons have different fitnesses, it is possible to recover $dN/dS$ values above 1, even though no positive selection is occurring. Finally, we are able to use the robust relationship between $dN/dS$ and selection coefficients to identify previously unrecognized biases in mechanistic codon models. In particular, we have determined that the equilibrium codon frequency parameters used in these likelihood models are insufficient to account for underlying codon or nucleotide biases, but instead mechanistic codon models systematically underestimate $dN/dS$ in the presence of nucleotide compositional bias. By examining the relationship between distinct modeling frameworks, therefore, we were able to uncover certain properties and limitations of both mechanistic codon and MutSel models that would otherwise be unrecognized.

## Results

**Mathematical relationship between dN/dS and scaled selection coefficients.** We describe here how to calculate $dN/dS$ from scaled selection coefficients. MutSel models assume that both population size and selection pressure, and hence scaled selection coefficients, are constant across a given phylogeny [10, 29, 13]. Therefore, given reversibility, it is possible to derive stationary equilibrium frequencies, which result from the dynamic interplay between both mutational and selective pressures, for all codons. In the presence of symmetric nucleotide mutation rates, e.g. where $\mu_{xy} = \mu_{yx}$, we can compute analytically precise values for codon equilibrium frequencies according to theory developed by Sella and Hirsh [30],

$$\pi_i = \frac{e^{S_i}}{\sum_k e^{S_k}}, \qquad [\mathbf{1}]$$

where the sum in the denominator runs over all 61 sense codons, and $S_i$ corresponds to the scaled selection coefficient for codon $i$. Alternatively, if nucleotides mutation rates are asymmetric, equilibrium codon frequencies can be numerically calculated though detailed balance conditions, such that the relationships $\pi_i \mu_{ij} = \pi_j \mu_{ji}$ and $\sum \pi_i = 1$ are satisfied.

Using equilibrium codon frequencies $\pi_i$, we can write the fixation probability for a mutation from codon $i$ to codon $j$ as [10, 30]

$$f_{ij} = \frac{1 - (\pi_i/\pi_j)^{1/N_e}}{1 - \pi_i/\pi_j} \approx \frac{1}{N_e} \frac{\ln \pi_j - \ln \pi_i}{1 - \pi_i/\pi_j}, \qquad [\mathbf{2}]$$

where $N_e$ is the effective population size. Through this framework, we can calculate an evolutionary rate by summing over all substitution probabilities weighted by the frequency of the originating codon. Further, we can establish specific expressions for nonsynonymous and synonymous evolutionary rates, and then divide them in order to obtain a value for the evolutionary rate ratio $dN/dS$.

To begin, we can write the nonsynonymous rate $K_{\mathrm{N}}$ as

$$K_{\mathrm{N}} = N_e \sum_i \sum_{j \in \mathcal{N}_i} \pi_i f_{ij} \mu_{ij}, \qquad [\mathbf{3}]$$

where $\mathcal{N}_i$ is the set of codons that are nonsynonymous to codon $i$ and differ from it by one nucleotide. To normalize $K_{\mathrm{N}}$, we divide it by the number of nonsynonymous sites, which we calculate according to the mutational opportunity definition of a site [8, 28] as

$$L_{\mathrm{N}} = \sum_i \sum_{j \in \mathcal{N}_i} \pi_i \mu_{ij}, \qquad [\mathbf{4}]$$

and thus we find that

$$dN = \frac{K_{\mathrm{N}}}{L_{\mathrm{N}}} = \frac{N_e \sum_i \sum_{j \in \mathcal{N}_i} \pi_i f_{ij} \mu_{ij}}{\sum_i \sum_{j \in \mathcal{N}_i} \pi_i \mu_{ij}}. \qquad [\mathbf{5}]$$

Similarly, for $dS$, the synonymous evolutionary rate $K_{\mathrm{S}}$ per synonymous site $L_{\mathrm{S}}$, we find

$$dS = \frac{K_{\mathrm{S}}}{L_{\mathrm{S}}} = \frac{N_e \sum_i \sum_{j \in \mathcal{S}_i} \pi_i f_{ij} \mu_{ij}}{\sum_i \sum_{j \in \mathcal{S}_i} \pi_i \mu_{ij}}, \qquad [\mathbf{6}]$$

where $\mathcal{S}_i$ is the set of codons that are synonymous to codon $i$ and differ from it by one nucleotide substitution. The quantities $K_{\mathrm{S}}$ and $L_{\mathrm{S}}$ are defined as in Eqs. [**3**] and [**4**] but summing over $j \in \mathcal{S}_i$ instead of $j \in \mathcal{N}_i$.

Equations [**2**]–[**6**] establish a connection between the equilibrium codon frequencies and the evolutionary rate ratio $dN/dS$. Moreover, we note that, if we make the dual assumptions that nucleotide mutation rates are symmetric and that all synonymous codons have equal fitness (e.g. synonymous mutations are neutral), the synonymous fixation rate $f_{ij} = 1/N_e$ [36]. Under this circumstance, the value for $dS$ reduces to 1.

**dN/dS can be accurately predicted from scaled selection coefficients.** To validate the mathematical relationship between stationary codon frequencies and $dN/dS$ described in equations [**2**]–[**6**], we simulated protein-coding alignments according to the Halpern-Bruno [10] MutSel model framework, whose instantaneous rate matrix is given by equation [**??**]. We simulated 100 alignments in which synonymous codons had equal fitness values, and 100 alignments with codon bias, e.g. where the fitness values, and hence equilibrium frequencies, differed among synonymous codons (see Methods for details). All simulations described in this subsection assumed a symmetric nucleotide rate matrix, with the transition-tranversion bias ratio $\kappa \sim \mathcal{U}(1, 6)$. Given these symmetric mutation rates, the codon equilibrium frequencies in these 200 simulations are directly proportional to their fitnesses [30]. For each alignment, we calculated $dN/dS$ using equations [**2**]–[**6**] as well as using the M0 mechanistic codon model [1], as implemented in the HyPhy batch language [23].

The relationship between $dN/dS$ measurements is shown in Figure 1A (for simulations with no codon bias) and Figure 1B (for simulations with codon bias). It is clear that $dN/dS$ values derived using codon frequencies agree nearly perfectly with those inferred using standard maximum likelihood methods, and fitness differences among synonymous codons do not influence this robust relationship. Additionally, in Figure 1C, we demonstrate convergence of $dN/dS$ estimates as the size of the data set, represented by simulated alignment length, increases. Taken together, these results demonstrate that MutSel model parameters fully encapsulate information

regarding $dN/dS$, and that the results from MutSel and mechanistic codon models are in complete agreement.

We obtained the $\omega$ MLEs reported in Figure 1 by fixing the $\kappa$ parameter in the M0 model's instantaneous rate matrix to its true, simulated value. We additionally performed all $\omega$ inferences which considered $\kappa$ a free parameter of the model. The resulting $\omega$ MLEs computed for both $\kappa$ parameterization were in near perfect agreement, yielding $r^2 = 0.997$ for alignments without codon bias, and $r^2 = 0.992$ for alignments with codon bias (Figure S1). Moreover, we found that the correlation between $\kappa$ MLEs and the true $\kappa$ value, for likelihood inferences which considered $\kappa$ a free parameter, was somewhat weaker, with $r^2 = 0.5$ for alignments without codon bias, and $r^2 = 0.44$ for alignments with codon bias (Figure S2). However, these moderate correlation strengths were strongly influenced by the presence of a few outlying points, so the relationship between $\kappa$ MLEs and true $\kappa$ was actually quite strong. Thus, the mechanistic codon modeling framework appears to be extremely robust to estimating both $\omega$ and $\kappa$.

**dN/dS and selection strength are strongly related.** Moreover, the strength of selection pressure scales fairly well with $dN/dS$. Figure 2 displays the relationship between $dN/dS$ and the standard deviation, $\sigma^2$, of the distribution of amino acid selection coefficients. Higher values of $\sigma^2$ indicate larger fitness differences among amino acids, ultimately leading to stronger selection pressure acting on nonsynonymous substitutions. Figure 2 demonstrates that when fitness differences among amino acids are very high, $dN/dS$ takes on lower values, properly reflecting stronger purifying selection. As expected, this trend is more robust for alignments without codon bias (Figure 2A, $r^2 = 0.83$) than for alignments with codon bias (Figure 2B, $r^2 = 0.45$). The weaker relationship for simulations with codon bias emerges from the fact that fitness differences among synonymous codons obscure the underlying amino acid fitness differences. Even so, synonymous fitness differences did not affect the significant negative correlation between $dN/dS$ and selection strength.

Importantly, Figure 2A shows that, in the limiting case when $\sigma^2$ approaches 0, and thus all codons have virtually the same fitness values, $dN/dS$ converges to a value of 1. This result properly reflects the case of neutral evolution, futher verifying the accuracy of these $dN/dS$ estimates. In fact, in **PROOF**, we prove that, when synonymous codons have equal fitness values, $dN/dS$ is necessarily always less than or equal to 1. Indeed, the largest $dN/dS$ value recovered for simulations with equal synonymous codon fitness was 0.997, which featured a $\sigma^2 = 0.08$. This restriction of $dN/dS < 1$ does not, however, hold in the face of codon bias, which can readily yield $dN/dS$ values greater than 1 (Figures 1B and 2B), even though the protein sequence is evolving under equilibrium conditions. We discuss the implications of these findings in depth in *Discussion*.

**Incorporation of asymmetric mutation rates reveals bias in mechanistic codon model inferences.** Results reported in the previous subsection were obtained using fully-simulated scaled selection coefficients, along with symmetric nucleotide mutation rates (e.g. where $\mu_{xy} = \mu_{yx}$). The latter assumption may not be entirely realistic; indeed, mutational biases, particularly transitions from $C/G \rightarrow T/A$, are known to contribute to uneven nucleotide compositions in real genomes [37, 38, 32, 33]. Therefore, we performed additional simulations which made use of realistic amino acid fitness and nucleotide mutation rate parameters. In particular, we used influenza nucleoprotein (NP) site-specific amino acid preference values, given by Bloom [26]. These data consisted of experimentally-determined fitness values for each individual amino acid across all sites in NP, yielding 498 distinct amino acid propensity distributions. We combined these experimental fitness parameters with three sets of experimentally determined mutation rates, either for NP [26], yeast [32], or polio virus [33]. Importantly, while all of these mutation matrices is asymmetric, they feature differing degrees of asymmetry, with NP mutation rates being the most symmetric and polio mutation rates the most asymmetric. More precisely, the average ratios $\mu_{xy}/\mu_{yx}$ for the NP, yeast, and polio mutation rates are 1.03, 1.66, and 4.45, respectively. Finally, using each of the 498 amino acid fitness distributions, we calculated stationary codon frequencies $\pi_i$ under detailed balance conditions, using the approach in [26, 27].

We again computed, for each resulting set of stationary codon frequencies, $dN/dS$ using equations **[2]**–**[6]**, and we simulated alignments using the Halpern-Bruno MutSel model (equation **[7]**). We inferred $\omega$ MLEs using the M0 mechanistic codon model according to five different codon frequency model parameterizations. These parameterizations included Fequal [28] and the common frequency estimators F61 [8], F3x4 [9], and CF3x4 [35]. The F61 estimator approximates these model parameters using an alignment's empirical codon frequencies, while the F3x4 and CF3x4 estimators approximate codon frequency parameters using positional nucleotide frequencies. Additionally, we inferred $\omega$ using a fifth frequency parameterization which consisted of the codon frequencies that would arise strictly from mutational processes, in the absence of natural selection. We term this parameterization "Ftrue," as its values are precisely those intended for the model's codon frequency parameters [8, 9, 31, 28].

Figure 3 shows the resulting relationships between $dN/dS$ and $\omega$ MLEs for each set of mutation rates (NP, yeast and polio), across M0 model codon frequency parameterizations (Figure S3 and Tables S1-2 contains regression plots and numerical data for all simulated data sets and frequency parameterizations). Figure 3A displays the bias, or systematic deviation from a 1:1 relationship, between $dN/dS$ and $\omega$, and Figure 3B displays $r^2$ values between $dN/dS$ and $\omega$. Note that a bias of 0 would indicate a perfect correlation between $dN/dS$ and $\omega$ MLEs.

Overall, Figure 3 reveals a clear trend where the relationship between $dN/dS$ and $\omega$ decreases in accuracy and strength as mutation rates become increasingly asymmetric for all codon frequency parameterizations. Across all frequency parameterizations, $\omega$ inferences on alignments with NP mutation rates have the least amount of bias, followed by yeast and finally polio mutation rates, and a parallel trend exists for the correlation strengths. Strikingly, even though Ftrue takes on the exact values intended for the model's codon frequency parameters, Ftrue does not outperform the other frequency parameterizations. Instead, Ftrue features the same general trend of decreasing accuracy with increasing mutational asymmetry. Surprisingly, it seems as though the Fequal parameterization yields the most accurate $\omega$ MLEs; it features the least amount of bias, and has very high correlations for both NP and yeast mutation rates. Although Fequal features a lower $r^2$ for polio mutation rates than do the estimators F61, F3x4, and CF3x4, the relative strength of the $r^2$ values for these three estimators is misleading. In fact, their increased bias demonstrates that these estimators systematically underestimate $dN/dS$. Taken together, these results suggest that the codon frequency parameters in mechanistic codon models cannot adequately account for nucleotide compositional bias as generated by mutation.

There is a noteworthy exception to the overall trend of $\omega$ underestimation; results using F61 actually overestimated $\omega$ for simulations which employed NP mutation rates, leading to a relatively low correlation. We attribute this result to the fact that the NP mutation rates were only minimally asymmetric, with an average $\mu_{xy}/\mu_{yx} = 1.03$. Importantly, when nucleotide mutation rates are symmetric, steady-state codon frequencies are controlled only by selection, as there is no opportunity to generate compositional bias through mutation [30]. Because the F61 estimator directly uses empirical codon frequencies, the resulting M0 codon frequency parameters actually contain information about the strength of natural selection. Therefore, selection pressures which should be strictly incorporated in the $\omega$ parameter are inadvertently contained within the codon frequency parameters. The ultimate effect is that the model infers selection to be weaker than it actually is and thus produces elevated $\omega$ MLEs. Although the $\omega$ overestimation was relatively small in this particular case, it highlights that it is crucial to properly parameterize mechanistic codon models. If these models are incorrectly parameterized, the $\omega$ parameter will no longer accurately represent the $dN/dS$ evolutionary rate ratio, but will instead be a meaningless quantity.

## Discussion

The oldest and most-widely used method to infer selection pressure in protein-coding genes calculates the evolutionary rate ratio $dN/dS$, which represents the ratio of non-synonymous to synonymous substitution rates. In turn, $dN/dS$ is commonly used to identify proteins or protein sites that experience negative selection ($dN/dS < 1$), evolve neutrally ($dN/dS \approx 1$), or experience positive, diversifying selection ($dN/dS > 1$) [1, 2, 3]. By contrast, MutSel models estimate scaled selection coefficients for amino acid and/or codons [10, 11, 12, 13, 15]. Thus, while mechanistic codon models describe the how quickly a protein's constituent amino acids change, MutSel models calculate the strength of natural selection operating on the specific amino-acid changes.

Until now, however, it has been an open question how these two modeling frameworks relate to one another. Whether these models' focal parameters, $dN/dS$ and scaled selection coefficients, yield comparable or distinct information about the evolutionary process has unknown. To solve this question, we have derived a formal mathematical relationship between $dN/dS$ and codon scaled selection coefficients. Through a simulation approach, we confirm that for $dN/dS$ can be precisely calculated from scaled selection coefficients, demonstrating that mechanistic codon models and MutSel models are in full agreement. Furthermore, this relationship is robust to fitness differences among synonymous codons. However, we note that our implementation of codon bias explicitly assumed that selection alone, and not mutation, was the sole source of codon bias. This implementation might not be entirely biologically realistic, as both mutational and selective forces likely contribute to codon bias in real genomes [39, 40, 41, 42, 43]. However, the key finding that we present is that fitness differences among synonymous codons do not affect the robust mathematical equivalency between scaled selection coefficients and $dN/dS$.

We have also proven that, when synonymous codons have equal fitnesses and mutation rates are symmetric, $dN/dS$ as calculated from scaled selection coefficients will always be less than 1. This proof formalizes the underlying assumption of MutSel models that the selection pressure is constant over the phylogeny, and confirms that MutSel models are inherently unable to describe positive, diversifying selection. This result implies that MutSel models should not be used when positive selection is expected to be occuring. Although this proof assumes symmetric nucleotide mutation rates, we do not expect that deviations from this assumption will have dramatic effects on $dN/dS$ estimates.

Alternatively, when the assumption of equal fitness among synonymous codons is violated, we find that it is theoretically possible to achieve $dN/dS > 1$ (Figures 1B and 2B). In fact, when selection induces codon bias, it is possible to have arbitrarily high $dN/dS$ values; in the most extreme case of codon bias, in which only a single codon per amino acid is selectively tolerated, the number of synonymous sites $L_S = 0$, and thus the value for $dN/dS$ approaches infinity. Given that all simulations here assumed an overarching regime of purifying selection, the finding that $dN/dS$ can still be greater than 1 might seem paradoxical. However, the logical argument that $dN/dS > 1$ represents positive selection assumes that synonymous substitutions are selectively neutral, an assumption which is violated when synonymous codons have different fitnesses. Thus, in theory, what is classically termed positive selection can result simply from strong synonymous fitness differences. Even so, it is unlikely that this possibility will strongly influence real analyses, as selection on synonymous codons has been shown to be relatively weak in most taxa [41]. Experimental evidence from the yeast Hsp90 protein, for instance, showed that, while there are some fitness differences among synonymous codons, these differences are exceedingly minimal relative to fitness differences among amino acids [44, 45]. However, it is possible that estimates of positive selection in species with high levels of codon bias, such as bacterial, *Drosophila*, or certain mammalian species [40, 46, 43], may not be true cases of positive selection, but rather simply signals of strong codon bias.

That the $dN/dS$ values calculated using equations **[2]** - **[6]** agree precisely with $\omega$ estimates inferred from the M0 mechanisic codon model lends firm support for the validity of our $dN/dS$ calculations. It has been long-recognized that different $dN/dS$ inference methods yield different $dN/dS$ estimates, as do different parameterizations of maximum likelihood mechanistic codon models [31, 28, 47]. Previously proposed frameworks for calculating $dN/dS$ have broadly fallen into two camps: heuristic counting methods [5, 6, 48, 49, 31] and maximum likelihood methods [8, 9, 28, 7]. Unlike these frameworks, the $dN/dS$ calculations we have proposed here are solidly grounded in population genetics theory. That $\omega$ MLEs broadly agree with our $dN/dS$ calculations lend robust support to the accuracy of these $dN/dS$ values, and indeed to the methodological accuracy of mechanistic codon models. We emphasize, however, that the $dN/dS$ calculations we have proposed are only suitable when the protein is evolving under steady-state conditions, or in other words when selective pressure remains constant over time.

Mechanistic codon models inferred $dN/dS$ with extremely high accuracy only when nucleotide mutation rates were symmetric. In the presence of asymmetric mutation rates the relationship between $dN/dS$ and mechanistic codon model $\omega$ MLEs weakened substantially. As our $dN/dS$ calculations explicitly consider nucleotide mutation rates, we contend that this weakened relationship resulted when the M0 systematically underestimated $dN/dS$. This trend definitively resulted from asymmetric mutation rates, and not from uneven nucleotide compositions. Indeed, our alignments simulated with symmetric mutation rates featured a wide array of GC-contents, ranging from 0.22-0.82. Given these alignments' symmetric mutation rates, unequal nucleotide compositions arose strictly from natural selection favoring particular amino

acids, and maximum likelihood methods inferred $dN/dS$ perfectly.

Therefore, it seems as though, when nucleotide compositional bias results from mutational processes, mechanistic codon models yield biased $dN/dS$ estimates. Mechanistic codon models attempt to deal with mutation-induced nucleotide compostional bias through the use of equilibrium codon frequency parameters [28]. Unlike the equilibrium steady-state frequencies on which we have focused throughout this paper, these frequency parameters are intended to represent the codon frequencies which would exist in the absence of amino-acid level selection, but from mutational or other biological processes, such as GC-biased gene conversion [50, 51], alone [8, 9, 31, 28]. However, we found that even a codon frequency parameterization which used precisely these values (Ftrue) still suffered from a decrease in accuracy as mutation-induced nucleotide compositional bias increased. Instead, the Fequal frequency parameterization, which assigns parameter values of 1/61 for all sense codons and is therefore no different from a matrix-scaling factor, performed just as well, if not better, than all other frequency parameterizations, including F61, F3x4, and CF3x4. Taken together, these results strongly suggest that the mechanistic codon model's codon frequency parameters are ill-suited to accomodate compositional biases which result from forces other than amino-acid level selection. We therefore suggest that future work investigate the utility of novel parameters for mechanistic codon models which better account for asymmetry in the mutational process.

We additionally emphasize that improper model parameterizations lead to spurious $\omega$ MLEs which do not accurately represent $dN/dS$. If other model parameters (e.g. $\kappa$ or the equilibrium codon frequencies) are specified incorrectly, or inadvertently contain information about amino-acid level natural selection, the resulting $\omega$ MLE will not represent the true $dN/dS$ evolutionary rate ratio. Only by ensuring that $\omega$ is the only model parameter which contains information about natural selection will it assuredly represent $dN/dS$. This finding calls into question the use of the F61 frequency estimator, which assigns codon frequency parameter values based on empirical codon frequencies. If, by chance, a given alignment's protein-coding sequences evolved under symmetric mutation rates, these empirical codon frequencies will contain substantial information regarding the strength of natural selection, ultimately leading to incorrect $dN/dS$ inferences. Therefore, we recommend that users employ either the F3x4 [9] or the CF3x4 [35] frequency estimators in their analyses.

In sum, we have garnered several important insights into the behavior of mechanistic codon models, the $dN/dS$ metric, and selection coefficients. These results were only made possible through establishing a formal mathematical relationship between distinct modeling frameworks. We believe that the approach presented in this paper represents a promising future avenue for methodological benchmarking. Typically, researchers assess the performance of a given inference framework through simulations which adhere to the underlying model's assumptions. However, this strategy can only confirm that inference methods are behaving as expected; it cannot confirm that the underlying model accurately represents the evolutionary process. Instead, we suggest an alternate approach to benchmark inference methods, and indeed evolutionary models: assessing the extent to which distinct models agree may serve as a novel, robust strategy to determine the accuracy of different modeling frameworks and reveal previously unrecognized model properties or biases.

## Methods

We simulated protein-coding sequences as a continuous-time Markov process [28] according to the MutSel model proposed by [10]. This model's instantaneous rate matrix $Q$ is given by

$$Q_{ij} = \begin{cases} f_{ij}\mu_{ij}\kappa & \text{single nucleotide transition} \\ f_{ij}\mu_{ij} & \text{single nucleotide transversion} \\ 0 & \text{multiple nucleotide changes} \end{cases}, \quad [7]$$

. Here, $\mu_{ij}$ is the nucleotide mutation rate and $f_{ij}$, the fixation probability from codon $i$ to $j$, is defined as

$$f_{ij} = \frac{2Ns_{ij}}{1 - e^{2Ns_{ij}}}, \quad [8]$$

where the value $2N_e s_{ij}$ represents the scaled selection coefficient for a mutation from codon $i$ to codon $j$[10, 29]. As shown by [10], the fixation probability

$$f_{ij} \propto ln\left(\frac{\pi_j \mu_{ij}}{\pi_i \mu_{ji}}\right) \Big/ \left(1 - \frac{\pi_i \mu_{ji}}{\pi_j \mu_{ij}}\right). \quad [9]$$

In this approximation, $\pi_i$ is the steady-state, or equilibrium, frequency of codon $i$. Importantly, these equilibrium frequency values are those which result from the joint effects of both mutation and selection.

All alignments presented here were simulated along a 4-taxon phylogeny, beginning with a root sequence selected using steady-state codon frequencies. Unless otherwise stated, all simulated alignments contained 500,000 codon positions. A single evolutionary model was applied to all positions in the simulated sequences. While this lack of site-wise heterogeneity is unrealistic for real sequence evolution, it allows us to verify our derived relationship between equilibrium codon frequencies and $dN/dS$ with a sufficiently sized data set.

To demonstrate the relationship between $dN/dS$ and scaled selection coefficients, we simulated 100 sequences in which all synonymous codons have equal fitness (no codon bias), and 100 alignments in which synonymous codons featured different equilibrium frequencies (codon bias). For both sets of simulations, we assumed symmetric nucleotide mutation rates of $\mu_{xy} = 10^{-6}$ and $\kappa \sim \mathcal{U}(1, 6)$. We generated relative amino acid scaled selection coefficients $S_a$ for each simulation, by fixing one coefficient to 0 and drawing the remaining 19 values from a normal distribution $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 \sim \mathcal{U}(0, 4)$. Here, $\sigma^2$ effectively represents the strength of natural selection; the higher $\sigma^2$ is, the greater the fitness difference among amino acids, and thus selection acts more strongly. Moreover, these $S_a$ values correspond to the relative amino acid fitness parameters as inferred by currently available MutSel inference methods [14, 15]. For simulations without codon bias, we directly assigned $S_a$ values to codons such that all synonymous codons had the same scaled selection coefficient $S_i$, and thus the same fitness. For simulations with codon bias, we randomly selected a preferred codon for each amino acid. We then assigned the preferred codon a selection coefficient of $S_a + \lambda$ and all non-preferred codons a selection coefficient of $S_a - \lambda$. For each codon bias simulation, we drew $\lambda$ from $\mathcal{U}(0, 2)$.

Finally, we computed equilibrium frequencies for all codons according to a Boltzmann distribution,

$$\pi_i = \frac{e^{S_i}}{\sum_k e^{S_k}}, \quad [10]$$

where the denominator runs over all sense codons. Note that equation **[10]** reveals the precise equilibrium codon frequencies in the presence of a symmetric mutation matrix [30].

We calculated a global $dN/dS$ for each alignment using the mathematical framework outlined in **[??]**–**[6]** as well using standard maximum likelihood methods. Specifically, we

inferred $dN/dS$ using the M0 mechanistic codon model [2], as implemented in the HyPhy batch language [23]. The M0 model uses the GY94 instantaneous rate matrix [8, 1], which includes the primary parameters $\omega$, $\kappa$, and equilibrium codon frequencies. For simulations inferences, we inferred $\omega$ both by fixing $\kappa$ to its true value, and maintaining $\kappa$ as a free parameter of the model. We used the Fequal equilibrium codon frequency model parameterization, which assigns equal frequencies of 1/61 to all sense codons [28]. Codon frequency parameters, unlike the steady-state codon frequencies of the underlying evolutionary model, are meant to capture mutational biases, and these parameters should correspond to the codon frequencies which would be expected in the absence of selection [31], and a symmetric mutation process would produce equal frequencies of 1/61.

Additionally, we simulated alignments which made use of experimentally-determined amino acid fitness and mutation rate data. We used site-wise influenza nucleoprotein (NP) amino acid preferences from Bloom 2014 [26] and nucleotide mutation rates for either NP [26], yeast [32], or polio virus [33]. Note that all of these experimental mutation rate matrices were asymmetric. We combined each the 498 amino acid preference distributions with each set of nucleotide mutation rates to determine a total of $493 \times 3 = 1494$ unique experimental evolutionary Markov models, using the approach in Bloom [26], wherein the Metropolis acceptance criterion [34] was used to calculate amino acid fixation rates. We calculated each model's equilibrium, or steady-state, codon frequencies such that detailed balance $\pi_i \mu_i j = \pi_j \mu_j i$ and $\sum \pi_i = 1$, where the sum runs across all 61 sense codons, was satisfied. Finally, for each set of equilibrium codon frequencies, we simulated alignments according to equation **[7]**.

We inferred $\omega$ for the simulations which employed experimental data with 5 different M0 model parameterizations. All inferences considered *kappa* a free parameter of the model, but 5 different equilibrium codon frequency parameterizations were used. First, we inferred $\omega$ using the Fequal [28] parameterization, which assigns equal codon frequencies of 1/61 each. Second, we inferred $\omega$ by specifying codon frequencies which would arise strictly from mutational processes in the absence of natural selection. We computed these codon frequency values using the same approach as we did in calculating the true steady-state codon frequencies, except instead of using the experimental amino acid preference data, we assigned all amino acids the same preference value of 0.05, thus eliminating any amino-acid level fitness differences. This eliminated any amino-acid level selection and allowed mutation rates alone to determine equilibrium codon frequencies. We term this frequency parameterization "Ftrue." Finally, we used the common frequency estimators F61 [8], F3x4 [9], and CF3x4 [35]. As typical analyses consider model frequency parameters as protein-wide (not site-specific) parameters, we computed these parameter values by pooling, for each set of mutation rates, all 498 steady-state codon frequencies to derive average codon frequencies. This approach yielded a set of global equilibrium frequencies for each set of mutation rates, and we calculated the F61, F3x4, and CF3x4 frequencies from these distributions.

**ACKNOWLEDGMENTS.** This work was supported by the army and by NIH.


1. Nielsen R, Yang Z (1998) Likelihood models for detecting positive selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.
2. Yang ZH, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.
3. Kosakovsky Pond S, Frost S (2005) A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol 22:478–485.
4. Huelsenbeck JP, Jain S, Frost SWD, Kosakovsky Pond SL (2006) A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc Natl Acad Sci USA 103:6263–6268.
5. Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution consider the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174.
6. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426.
7. Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26:255–271.
8. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736.
9. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724.
10. Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15:910–917.
11. Rodrigue N, Philippe H, Lartillot N (2010) Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc Natl Acad Sci USA 107:4629–4634.
12. Tamuri AU, dos Reis M, Goldstein RA (2012) Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. Genetics 190:1101–1115.
13. Thorne JL, Lartillot N, Rodrigue N, Choi SC (2012) Codon models as vehicles for reconciling population genetics with inter-specific data. In Cannarozzi G, Schneider A, eds., Codon evolution: mechanisms and models (Oxford University Press, New York).
14. Rodrigue N, Lartillot N (2014) Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. Bioinformatics :1020–1021.
15. Tamuri AU, Goldman N, dos Reis M (2014) A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. Genetics 197:257–271.
16. Kosakovsky Pond S, et al. (2011) A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28:3033–3043.
17. Murrell B, et al. (2012) Detecting individual sites subject to episodic diversifying selection. PloS Genet 8:e1002764.
18. Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908–917.
19. Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472–2479.
20. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20:1692–1704.
21. Scherrer MP, Meyer AG, Wilke CO (2012) Modeling coding-sequence evolution within the context of residue solvent accessibility. BMC Evol Biol 12:179.
22. Meyer AG, Wilke CO (2012) Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol 30:36–44.
23. Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenetics. Bioinformatics 21:676–679.
24. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24:1586–1591.
25. Delport W, Poon A, Frost S, Pond S Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology .
26. Bloom JD (2014) An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol Biol Evol :To appear.
27. Bloom JD (2014) An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. Mol Biol Evol 31:1956–1978.
28. Yang Z (2006) Computational Molecular Evolution (Oxford University Press).
29. Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25:568–579.
30. Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA 102:9541–9546.
31. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–42.
32. Zhu YO, Siegal ML, Hall DW, Petrov DA (2014) Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci USA :FORTHCOMING.
33. Acevedo A, Brodsky L, Andino R (2014) Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505:686 – 690.
34. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087 – 1092.
35. Kosakovsky Pond SL, Delport W, Muse SV, Scheffler K (2010) Correcting the bias of empirical frequency parameter estimators in codon models. PLoS One 5:e11230.
36. Crow JF, Kimura M (1970) An Introduction to Population Genetics Theory (Burgess Pub. Co., California).
37. Hernandez RD, Williamson SH, Zhu L, D BC (2007) Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. Mol Biol Evol 24:2196 – 2202.
38. Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6:e1001115.

39. Blumer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907.
40. Duret L (2002) Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12:640–649.
41. Hershberg R, Petrov D (2008) Selection on codon bias. Annu Rev Genet 42.
42. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH (2009) Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci USA 101:3480–3485.
43. Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nature Rev Genet 12:32–42.
44. Hietpas RT, Jensen JD, Bolon DNA (2011) Experimental illumination of a fitness landscape. Proc Natl Acad Sci USA 108:7896–7901.
45. Hietpas RT, Bank C, Jensen JD, Bolon DNA (2013) Shifting fitness landscapes in response to altered environments. Evolution 67:3512–3522.
46. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Rev Genet 7:98–108.
47. Zhang Z, Yu J (2006) Evaluation of six methods for estimating synonymous and nonsynonymous substitution rates. Geno Prot Bioinfo 4:173–181.
48. Pamilo P, Bianchi NO (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol 10:271–281.
49. Ina Y (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. Journal of Molecular Evolution 40:190–226.
50. Duret L, Galtier N (2005) Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10:285–311.
51. Webster M, Hurst L (2012) Direct and indirect consequences of meiotic recombination: implications for genome evolution. Trends Genet 28:101–109.

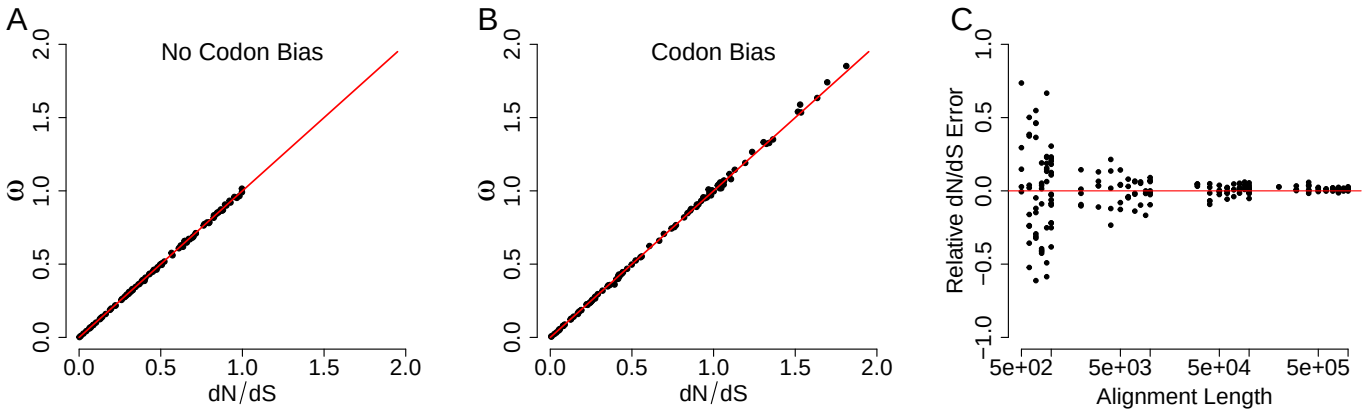


**Fig. 1.** Regressions between $dN/dS$ values as calculated from scaled selection coefficients and as inferred using the M0 mechanistic codon model. Each point corresponds to a single simulated alignment. All $\omega$ values shown here were inferred using by parameterizing the M0 model with $\kappa$ fixed to its true, simulated value as well as the Fequal codon frequency specification [28]. The red line in panels (A-B) is the $x = y$ line. (A) Simulations which assigned equal fitness values to synonymous codons. (B) Simulations which allowed different fitness values among synonymous codons. (C) Convergence of $\omega$ MLEs to the true $dN/dS$ value. The y-axis indicates the relative error of the maximum likelihood $dN/dS$ estimate, and the x-axis indicates the number of positions in the simulated alignment. As the number of positions, and hence the size of the data set, increases, the maximum likelihood estimates converge to the $dN/dS$ values calculated using equations $[\mathbf{2}]$-$[\mathbf{6}]$. The red line in panel (C) is the line $y = 0$, indicating no error.

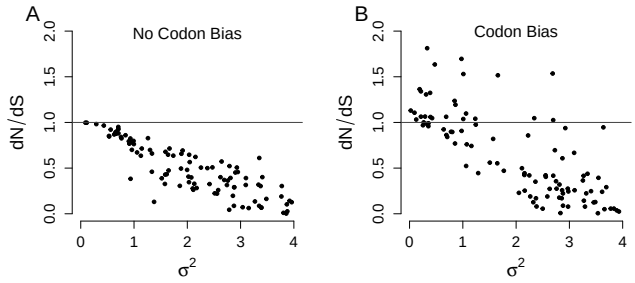


**Fig. 2.** $dN/dS$ decreases in proportion to amino-acid level selection. $dN/dS$ is plotted against the $\sigma^2$ of the simulated distribution of amino-acid scaled selection coefficients. Higher values of $\sigma^2$ indicate larger fitness differences among amino acids, whereas the limiting value of $\sigma^2 = 0$ means that all amino acids have the same fitness. (A) Simulations which assigned equal fitness values to synonymous codons. (B) Simulations which allowed different fitness values among synonymous codons. Note that, when synonymous codons have different fitness values, the relationship between amino-acid level selection and $dN/dS$ expectedly weakens.

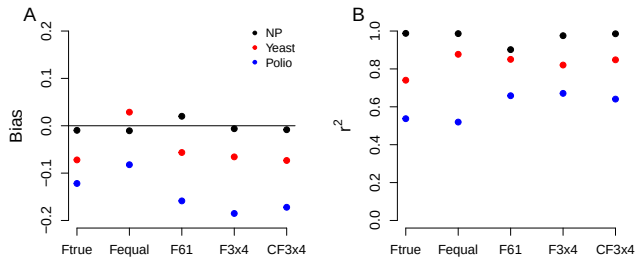


**Fig. 3.** Bias and correlation between $dN/dS$ and $\omega$ MLEs. (A) Bias, or systematic deviation from a 1:1 relationship, between $dN/dS$ and $\omega$ across M0 codon frequency parameterizations and mutation rates. (B) $r^2$ values between $dN/dS$ and $\omega$ across M0 codon frequency parameterizations and mutation rates.

## Supplementary Information

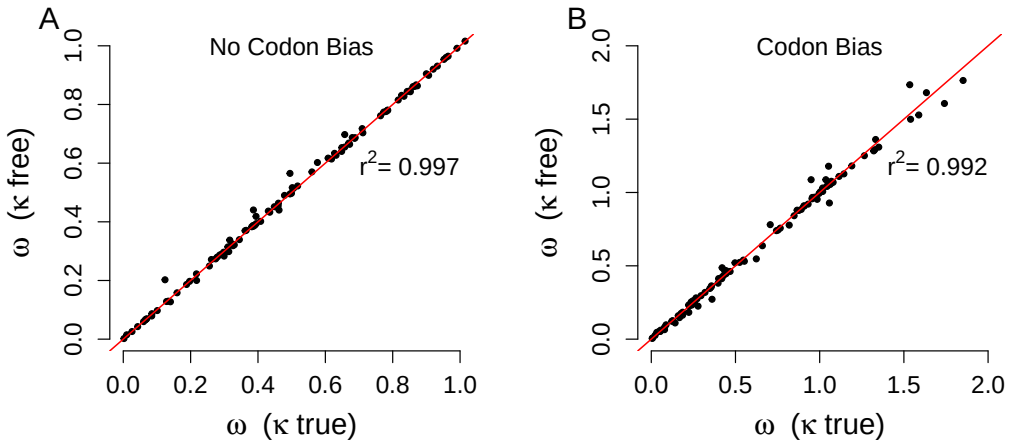


Figure S1. Caption!

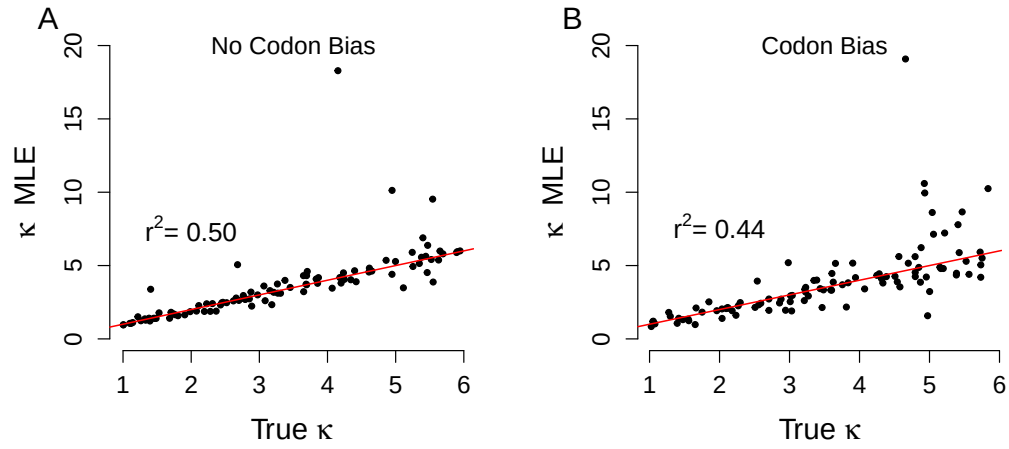


Figure S2. Caption!

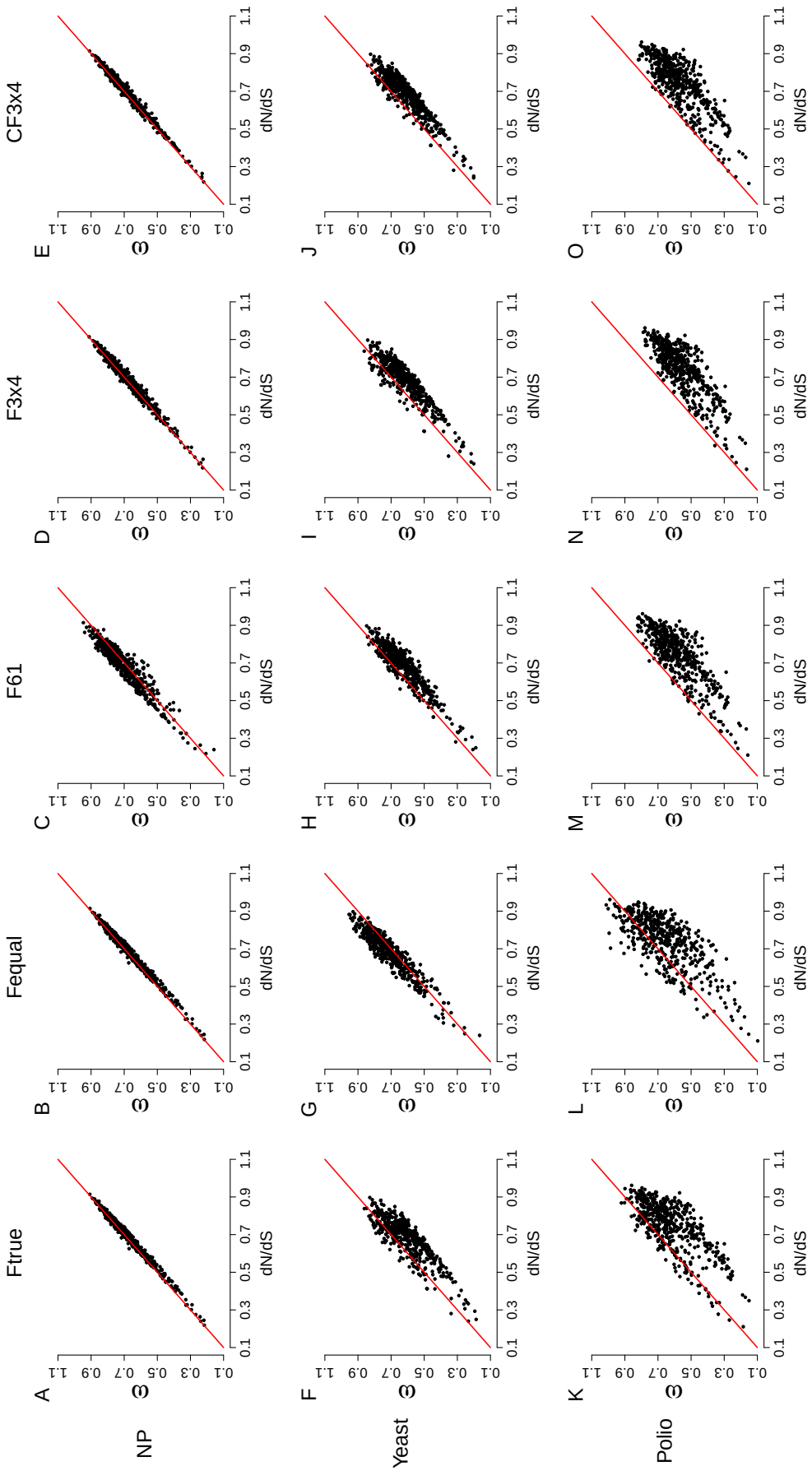


Figure S3. Caption!

Table S1. NYP BIAS.

| | M0 model codon frequency parameterization | | | | |
|---|---|---|---|---|---|
| Mutation rate | Ftrue | Fequal | F61 | F3x4 | CF3x4 |
| NP | -0.0097 | -0.011 | 0.02 | -0.0063 | -0.0084 |
| Yeast | -0.072 | 0.029 | -0.056 | -0.066 | -0.073 |
| Polio | -0.12 | -0.082 | -0.16 | -0.18 | -0.17 |

Table S2. NYP $r^2$.

| | M0 model codon frequency parameterization | | | | |
|---|---|---|---|---|---|
| Mutation rate | Ftrue | Fequal | F61 | F3x4 | CF3x4 |
| NP | 0.988 | 0.986 | 0.902 | 0.975 | 0.986 |
| Yeast | 0.740 | 0.877 | 0.850 | 0.820 | 0.848 |
| Polio | 0.537 | 0.519 | 0.659 | 0.671 | 0.641 |