The relationship between dN/dS and ???

Stephanie J. Spielman[1] and Claus O. Wilke[1]

Address:
[1]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.

*Corresponding author
Email: ??????????

Keywords: mutation-selection-balance models, mechanistic codon models, dN/dS, equilibrium codon frequencies, natural selection, protein evolution, models of sequence evolution

Abstract

Two models which investigate strength of selection in protein-coding sequences are mechanistic codon models and mutsel models. We have measures of dN/dS and amino acid/codon "propensities", which correspond to equilibrium frequencies. Are they the same? Are they different? We don't know! But now we do. And they're the same. But there are some important consequences of this. For instance, codon bias. Also for instance, mech codon models seem ill-equipped to deal with asymmetric mutation rates and parameterizing them is really awkward.

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 ($dN/dS > 1$) [1–4]. 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 [1, 8, 9]. These models yield maximum likelihood estimates (MLEs) for the parameter ω, 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, have emerged recently 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, which indicate the extent to which natural selection favors, or disfavors, particular codons or amino acids at a given protein position. Although MutSel models were first introduced over 15 years ago [10], they have seen virtually no use due to their high computational expense. However, recently, several computationally tractable model implementations have emerged [11, 12], allowing for the first time the potential for widespread use.

Although both mechanistic codon models and MutSel models describe the same fundamental process of protein-coding sequence evolution along a phylogeny, it is largely unknown how these two classes of models 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 values have any correspondence with scaled selection coefficients remains an open question. Therefore, while 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.

Here, we show that dN/dS can be precisely determined based on MutSel model parameters alone. Specifically, the estimates of fitness/mutation rates allows one to compute equilibrium frequencies. These equilibrium frequencies can then accurately estimate dN/dS. Several important results

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 these models' primary parameters from which one can infer dN/dS values from selection coefficients alone. Using a simulation approach, we verify that dN/dS values estimated using

selection coefficients alone correspond precisely to ω MLEs inferred using standard mechanistic codon models.

(???)Further, we prove that, under conditions of symmetric mutation rates, this relationship holds only under regimes of purifying selection or neutral evolution ($dN/dS \leq 1$). This proof reveals that MutSel models are inherently unable to describe accurately protein evolution under a regime of positive diversifying selection, or when $dN/dS > 1$.

Moreover, our analyses incidentally have revealed certain biases inherent in the ML ω inference approach. (...)

Methods

Sequence simulation

We simulated protein-coding sequences as a continuous-time Markov process [13] according to the MutSel model proposed by [10]. This model's instantaneous rate matrix Q is given by

$$Q_{ij} = \begin{cases} f_{ij}\mu_{ij}\kappa & \text{single nucleotide transition} \\ f_{ij}\mu_{ij} & \text{single nucleotide transversion} \\ 0 & \text{multiple nucleotide changes} \end{cases}, \tag{1}$$

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

$$f_{ij} = ln\left(\frac{\pi_j\mu_{ij}}{\pi_i\mu_{ji}}\right)\Big/\left(1 - \frac{\pi_i\mu_{ji}}{\pi_j\mu_{ij}}\right), \tag{2}$$

where π_i is the equilibrium frequency of codon i.

First, we simulated general data sets to show the theory. To demonstrate the relationship between models, we simulated sequences with and without codon bias. For these simulations, we assumed a symmetric mutation matrix with $\mu_{xy} = 10^{-6}$ and $\kappa \sim \mathcal{U}(1, 6)$. We simulated equilibrium codon frequencies π_i according to a Boltzmann distribution,

$$\pi_i = \frac{e^{s_i}}{\sum_k e^{s_k}}, \tag{3}$$

where the sum in the denominator runs over all 61 sense codons [14]. The s_i exponents were drawn from $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 \sim \mathcal{U}(0, 4)$. Here, σ^2 effectively represents the strength of natural selection; larger values of σ^2 will correspond to greater equilibrium frequency differences among codons, and thus more selective pressure. Moreover, when $\sigma^2 = 0$, all amino acids will have the same equilibrium frequency, corresponding to neutral evolution. Note that, in the presence of a symmetric mutation matrix, equation (3) suffices to determine the precise codon equilibrium frequencies [14].

For simulations which did not consider codon bias, we assigned all synonymous codons the same value s_i, thus leading to equal steady-state frequencies among synonymous codons. For simulations which incorporated codon bias, we first derived the codon frequencies expected in the absence of bias, and then converted them as follows. Assume that the equilibrium frequency of a codon in a given amino acid family, in the absence of bias, is P, and k is the size of the codon family. We randomly selected one of these codons to be the preferred, and assigned it a frequency of $\pi_{preferred} = P(1 + \lambda(k - 1))$, and all remaining codons were assigned the frequency $\pi_{non-preferred} = P(1 - \lambda)$. Here, $\lambda \sim \mathcal{U}(0, 1)$ represents the extent of codon bias. When $\lambda = 0$, all synonymous codons have the same frequency, and when $\lambda = 1$, only one codon has a non-zero frequency. This codon bias model further ensures that each amino acid's equilibrium frequency remains unchanged, but while partitioning this value differently among its codons.

Second, we used real data! We used real amino acid preference and nucleotide mutation rate data. For these sets, we used Jesse's amino acid preferences and Jesse's mu as well as yeast mu. For these cases, as real, asymmetric mutation rates were used, we had to determine the codon equilibrium frequencies π_i under detailed balance, such that they satisfied the relationship $\pi_i \mu_i j = \pi_j \mu_j i$.

We simulated protein-coding sequences along a 4-taxon phylogeny, with all branch lengths equal to 0.01, 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 selection coefficients and dN/dS with a sufficiently sized data set.

dN/dS Inference

We calculated a global dN/dS for each alignment using the mathematical framework outlined in (3)–(8) as well using standard maximum likelihood methods. Specifically, inferred dN/dS using the M0 mechanistic codon model [2], as implemented in the HyPhy batch language [15]. The M0 models uses the GY94 instantaneous rate matrix [1,8], which includes the primary parameters ω, κ, and equilibrium codon frequencies. As different κ and equilibrium codon frequency parameterizations can change ω estimates [13,16,17], we inferred ω under a variety of model parameterizations, including three κ parameterizations (κ fixed to 1, κ as a free parameter, and κ fixed to its true value), and four codon frequency specifications (Fequal (all sense codons have an equilibrium frequency of 1/61), F3x4 codon frequencies [9], CF3x4 codon frequencies [18] and F61, or empirical codon freqencies [8]), ultimately resulting in 12 ω MLEs per simulated alignment. Note that we additionally verified that the system was evolving under a state-state process by verifying that the true, simulated codon frequencies were the same as the empirical codon frequencies calculated from the simulated alignment. All code used is freely available at **github**.

Results

Mathematical relationship between dN/dS and equilibrium codon frequencies

We describe here how to calculate dN/dS from steady-state codon frequencies. The fixation probability for a mutation from codon i to codon j is [10,14]

$$f_{ij} = \frac{1 - (\pi_i/\pi_j)^{1/N_e}}{1 - \pi_i/\pi_j} \approx \frac{1}{N_e} \frac{\ln \pi_j - \ln \pi_i}{1 - \pi_i/\pi_j}\,, \tag{4}$$

where N_e is the effective population size. Using 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_{N} as

$$K_{\mathrm{N}} = N_e \sum_i \sum_{j \in \mathcal{N}_i} \pi_i f_{ij} \mu_{ij}\,, \tag{5}$$

where $\mathcal{N}_i$ is the set of codons that are nonsynonymous to codon i and differ from it by one nucleotide. To normalize $K_{\rm N}$, we divide it by the number of nonsynonymous sites, which we calculate according to the mutational opportunity definition of a site [8, 13] as

$$L_{\rm N} = \sum_i \sum_{j \in \mathcal{N}_i} \pi_i \mu_{ij} \,, \tag{6}$$

and thus we find that

$$dN = \frac{K_{\rm N}}{L_{\rm N}} = \frac{N_e \sum_i \sum_{j \in \mathcal{N}_i} \pi_i f_{ij} \mu_{ij}}{\sum_i \sum_{j \in \mathcal{N}_i} \pi_i \mu_{ij}} \,. \tag{7}$$

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

$$dS = \frac{K_{\rm S}}{L_{\rm S}} = \frac{N_e \sum_i \sum_{j \in \mathcal{S}_i} \pi_i f_{ij} \mu_{ij}}{\sum_i \sum_{j \in \mathcal{S}_i} \pi_i \mu_{ij}} \,, \tag{8}$$

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 $K_{\rm S}$ and $L_{\rm S}$ are defined as in Eqs. (5) and (6) but summing over $j \in \mathcal{S}_i$ instead of $j \in \mathcal{N}_i$.

Equations (4)–(8) establish a connection between the equilibrium codon frequencies and the evolutionary rate ratio dN/dS. Moreover, we note that, if we assume that all synonymous codons have equal fitness (e.g. synonymous mutations are neutral), the synonymous fixation rate $f_{ij} = 1/N_e$ [19]. Under this circumstance, the value for dS reduces to 1.

dN/dS can be accurately predicted from equilibrium codon frequencies

To validate the mathematical relationship between steady-state codon frequencies and dN/dS described in equations (4)–(8), we simulated protein-coding sequences along a 4-taxon phylogeny according to a mutation-selection model framework [10,14]. We simulated 100 alignments in which synonymous codons had equal frequencies, and 100 alignments with codon bias, e.g. where the equilibrium frequencies differed among synonymous codons (see Methods for details). All simulations assumed an underlying symmetric nucleotide rate matrix, with the transition-tranversion bias ratio $\kappa \sim \mathcal{U}(1, 6)$. For each alignment, we calculated dN/dS using equations (4)–(8) as well as using the M0 mechanistic codon model [1], as implemented in the HyPhy batch language [15].

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 frequency 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.

Moreover, dN/dS scales excellently with the strength of natural selection. For each simulated set of equilibrium codon frequencies, we computed the Shannon entropy

$$H = -\sum_i \pi_i \ln \pi_i, \tag{9}$$

where the sum runs over all 61 sense codons. When purifying selection acts strongly, relatively fewer codons will be selectively tolerated, thus leading to lower entropy values. Note that the

maximum value of $H = 4.11$ is achieved when all codons have an equal frequency of 1/61. Figure 2 displays the relationship between dN/dS and codon entropy for each simulated alignment. Here, we see that as codon entropy increases, dN/dS similarly increases, properly reflecting the decrease in selection pressure. Moreover, while this trend is present for both simulations with and without codon bias, the relationship between codon entropy and dN/dS is stronger for simulations without codon bias ($r^2 = 0.901$) than with codon bias ($r^2 = 0.493$). This difference emerges from the fact that simulations with codon bias featured frequency differences among synonymous codons, obscuring the overarching frequency differences among amino acids.

Figure 2A also shows that, as codon entropy approaches its maximum value of 4.11, dN/dS approaches, but never exceeds, the value of 1. This result properly reflects the case of neutral evolution, in which amino acids . In fact, in **SI proof**, we prove that, when synonymous codons have equal equilibrium frequencies, dN/dS is necessarily always less than or equal to 1. This restriction 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*.

Neutral evolution happens when dN/dS ¡ 1.

Actually, something bigger is going on here. Neutral evolution implies that fitness differences among amino acids are nonexistent, and natural selection will tolerate any amino acid. However, the equilibrium codon frequencies which would corresponds to equal amino acid frequencies of 0.05, properly normalized based on codon family size, do not yield $dN/dS = 1$, the canonical value for neutral evolution. Instead, for codon frequencies which correspond to 0.05, again in the absence of synonymous codon fitness differences, the maximum value that dN/dS can achieve is 0.9215, far less than 1. This maximum value is reached when there is a single nucleotide mutation rate, which provides the mathematical max. However, if one computes dN/dS according to equal fitness values for *codons*, we arrive at the canonical $dN/dS = 1$. This finding is a little disturbing, and reveals that codon-based models, in general, might not suffice to measure selection pressure. When all codons have a an equal fitness, this necessarily implies fitness differences among amino acids given the redundancy of the genetic code. Fitness differences among amino acids should not, by definition, yield to estimates of neutral evolution, but there you have it.

This result suggests that tests for positive selection may be far too conservative. Well, bear in mind, this result is only when mutation is symmetric, and the value for dN/dS depends on kappa (given mutational opportunity site definition). dN/dS here is bounded by 0.912 as kappa approaches infinity and 0.925 as kappa approaches 0. A very small range, but again entirely depend on symmetric mutation rates. This doesn't terribly bother me, because we should NEVER want to see neutral evolution with a dN/dS != 1, even under unrealistic conditions. I'm sure that given the right mutational scheme and equal amino acid fitnesses, you could get dN/dS=1, but that doesn't detract from the fact that it's not 1 under the most basic conditions.

Use of realistic data is also ok, but actually not really.

We have shown that, under general conditions, this relationship holds up. We now investigate the relationship between codon frequencies and dN/dS Results reported in the previous subsection were computed from fully simulated data sets, and importantly an underlying symmetric mutation matrix. However, these parameters may not be entirely realistic, given well-known nucleotide compositional biases.

Therefore, we implemented additional simulations using realistic parameters taken directly from

experimental evidence of amino acid fitness values and mutation rates. In particular, we used influenza nucleoprotein (NP) site-specific amino acid preference values, given by Bloom [20]. These data consisted of experimental determination of the extent to which natural selection favors each individual amino acid across all 498 sites in NP, thus consisting 498 sets of amino acid fitness parameters. We combined these experimental fitness parameters with three sets of experimentally determined mutation rates, either for NP [20], yeast [21], or polio virus [22]. For each of the 498 distributions of amino acid preferences, we calculated steady-state codon frequencies π_i under detailed balance, such that the relationships $\pi_i \mu_{ij} = \pi_j \mu_{ji}$ and $\sum \pi_i = 1$, where the sum runs across all 61 sense codons, were satisfied.

We simulated alignments using these resulting equilibrium frequencies, and inferred ω using the M0 mechanistic codon model, as described in Methods. Again, we inferred ω according to a variety of codon frequency model parameterizations. We sought to mimic real data analysis to the extent possible by using the frequency estimators based on global "alignment" frequencies. In other words, for each set of mutation rates, we pooled the 498 site-based equilibrium codon frequencies to derive theoretical protein-wide equilibrium frequencies, as it typical of real analysis. We additionally

This is typically how analyses work. It is expected that this pooling will accurately reflect the underlying mutational process. We also had site-wise F3x4, CF3x4, and F61. We used two additional frequency specifications of Fequal and what we term Fnull. Fnull are the true codon frequencies for this data set which would be expected in the absence of amino-acid level selection but which would be produced in the presence of mutational/compositional biases alone.

Results here are fairly clean. We see that ML does an excellent job of inferring for the most part, but clearly not all frequency specifications behave the same. The color figure shows the bias from each frequency specification, where bias represents the systematic deviation from a 1:1 relationship between dN/dS and the ω parameter. We see distinct trends related to the extent of asymmetry in the mutation rates. We can calculate an asymmetry "factor" for each set of mutation rates by finding the average value of μ_{xy}/μ_{yx}, where we further ensure that this quantity is greater than 1 for mathematical goodness. The asymmetry factor for NP, yeast, and polio mutation rates are 1.029, 1.655, and 4.45, respectively.

Trends - overall, the least bias is found for global, but there is much more noise for fequal than for anything else. As expected, null performs the best as it has the least bias and the least noise. However, F3x4 and CF3x4 appear to do a decent job approximating this null distribution of codon frequencies. F61 also does decently, but markedly less well for the NP data set and slightly worse than the yeast data set. Now we come to the site-based frequencies. These performed relatively poorly, featuring substantial bias and poor correlations, in particular for the F61. This is easy to explain, as it is in fact a stupid way to parameterize the model. Remember, the goal of frequency estimators such as F3x4 is to estimate the null values. It's also really nice to see that bias for the NP is the highest. This is because np had the least amount of asymmetry in its mutation rates, and thus the frequencies more closely represent the actual fitnesses. Thus, all selection info is inside the F61, effectively a total model misspecification.

While null performs the best, the ω -dN/dS show a marked decrease as asymmetry increases. This suggests that the codon frequency parameters do not entirely capture what they are meant to. Remember, the only way for ω to equal dN/dS is if the rest of the parameters are correct. Otherwise, it represents some quantity ω.

The asymmetry-induced trends are also apparent in the relationship between codon entropy and dN/dS. Recall, when mutation rates are symmetric, equilibrium codon frequencies correspond precisely to codon fitness values [14], and thus selective strength. The relationship between entropy and dN/dS gets noisier as mutation rates become more asymmetric. This result perfectly reflects the dual influences of selection and mutation on equilibrium frequencies. The more that uneven

mutation rates influence frequencies, the more noise there is.

Discussion

The oldest and most-widely used method to infer selection pressure in protein-coding genes calculates calculates the ratio of non-synonymous (dN) to synonymous (dS) substitution rates dN/dS to identify sites that experience negative selection ($dN/dS < 1$), sites that evolve neutrally ($dN/dS \approx 1$), and sites that experience positive diversifying selection ($dN/dS > 1$). By contrast, MutSel models equilibrium amino acid and/or codon steady state frequencies and selection coefficients [**?**, 10, 12, 23, 24], for codons [25], or for both. 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. Some 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, 23]. Recent phylogenetic studies have also demonstrated that evolutionary models which explicitly consider amino acid fitness values offer dramatic improvements over other models, including mechanistic codon models, suggesting that MutSel models may more aptly represent the process of coding-sequence evolution [20, 26].

Here, we have derived a formal mathematical relationship between the quantities dN/dS and scaled codon selection coefficients, the primary parameters of mechanistic codon and MutSel models, respectively. Through a simulation approach, we find that these two models are in full agreement, and that the value for dN/dS can be precisely calculated from selection coefficients alone. Furthermore, this mathematical equivalency between selection coefficients and dN/dS values is robust to fitness differences among synonymous codons. However, it is important to note that our implementation of codon bias explicitly assumed that frequency differences among synonymous codons resulted from fitness differences alone. In other words, the sole source of codon bias in our simulations was selection, not mutation. This implementation might not be entirely biologically realistic, as both mutational and selective forces likely contribute to codon bias in real genomes [27–30]. 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.

Our results rest on the key assumptions that the protein sequence is evolving under steady-state, or equilibrium, conditions, and the nucleotide mutation rates are symmetric (e.g. $\mu_{xy} = \mu_{yx}$). The first assumption recapitulates the population genetics theory behind MutSel models, which assume that selection coefficients remain constant over the phylogeny, and therefore the protein is evolving along a static fitness landscape [10, 23, 24]. While the latter assumption of symmetric mutation rates may not be biologically realistic **cite papers which show asym mu**, it allowed us to investigate the mathematical relationship between dN/dS and selection coefficients under broad conditions. In particular, we have proven that, when synonymous codons have equal fitness and mutation rates are symmetric, dN/dS will always be less than 1. However, when synonymous codons were allowed to have different selection coefficients, dN/dS can easily be greater than 1, and indeed frequently was (Figures 1B and **??**B). In fact, when synonymous codons have different fitnesses, 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. We additionally expect that an asymmetric mutation rate matrix could yield $dN/dS > 1$.

In other words, even if the protein is evolving along to a static fitness landscape, it is possible

to that the sequence will feature dN/dS values greater than 1. This finding seems paradoxical to classic interpretations of $dN/dS > 1$. Typically, these values are viewed as hallmarks of positive or diversifying selection, which is assumed to occur when the protein sequence experiences strong selective pressure to change its constituent amino acids. Positive selection, therefore, necessarily implies that the protein does not evolve at equilibrium, but rather a shift in selective constraint caused new amino acids to be favored. However, one must also recognize that the contention that $dN/dS > 1$ represents positive selection assumes that synonymous substitutions are selectively neutral, which is likely not the case when codon bias exists. Thus, what is classically termed positive selection can result simply from strong synonymous fitness differences. Therefore, it is entirely possible that estimates of positive selection as inferred via dN/dS estimates in species with high levels of codon bias, such as many bacterial, *Drosophila*, or certain mammalian species [28–31], may not be true cases of positive selection, but rather simply signals of strong codon bias.

Incidentally, our study recovered that mechanistic codon models can produce strongly biased inferences when parameters are incorrectly specified. In particular, ω MLE values only corresponded to the true dN/dS value when the equilibrium codon frequency parameters were specified as equal (e.g. each codon had an equilibrium frequency of 1/61). Alternatively, the common approaches of using F61 (empirical) frequencies [8]) or frequency estimators such as F3x4 [9] and CF3x4 [18] always yielded incorrect and highly elevated ω MLEs. We explain this phenomenon by recognizing that the rationale for including codon frequency parameters in mechanistic codon models is to account for unequal nucleotide frequencies specifically caused by mutational and not selective forces [13,16]. The proper values for these parameters, then, should be the codon frequencies which would exist *in the absence of natural selection*. This approach is the only way to ensure that ω is the sole model parameter which contains information about natural selection. Otherwise, the ω parameter will no longer represent the true dN/dS evolutionary rate ratio.

Moreover, frequency estimators such as F3x4 and CF3x4, which use positional nucleotide frequencies to calculate codon frequencies, must make the implicit assumption that observed unequal base frequencies result from biased mutation rates. This assumption, however, may not be fully justified. Indeed, our simulated alignments featured a wide array of nucleotide compositions, with GC-contents ranging from 0.22-0.79. Given that we simulated sequences according to a symmetric mutation matrix, all compositional biases in our data sets resulted entirely from natural selection favoring particular codons, not by any bias towards unequal base frequencies. Therefore, the proper equilibrium frequency parameterization for our alignments was indeed equal codon frequencies, which would be expected in the absence of natural selection and when mutation rates are symmetric.

These results emphasize that it is crucial to parameterize mechanistic codon models properly. Indeed, their primary parameter ω will only truly represent dN/dS when all other parameters are properly specified. If codon frequencies are not properly specified, which we suspect is the case in most analyses, then the ω MLEs are virtually meaningless and do not represent selective pressure. Therefore, we contend that there is hardly ever a justification to specify empirical codon frequencies, also known as the F61 frequency estimator [8], as natural selection has clearly produced the observed frequencies. Unfortunately, the F61 frequencies are the default parameterization in the widely-used PAML software's codeml implementation [32], so we strongly recommend that users take great care when using this package. In addition, the only robust way to ensure that codon frequencies are properly specified is through experimentally calculating mutation rates. Luckily, this data already exists for a variety of taxa, including **citations for papers uncovering mutation rates**. We recommend that, if experimental data is absent, users err on the side of caution and specify equal codon frequencies to reduce the possibility of false positives.

Finally, we contend that this methods presented in this paper reveal 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.

References

[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 SD (2005) Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22: 1208–1222.

[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, Lartillot N (2014) Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. Bioinformatics : 1020–1021.

[12] 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.

[13] Yang Z (2006) Computational Molecular Evolution. Oxford University Press.

[14] Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA 102: 9541–9546.

[15] Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenetics. Bioinformatics 21: 676–679.

[16] Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32–42.

[17] Zhang Z, Yu J (2006) Evaluation of six methods for estimating synonymous and nonsynonymous substitution rates. Geno Prot Bioinfo 4: 173–181.

[18] 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.

[19] Crow JF, Kimura M (1970) An Introduction to Population Genetics Theory. California: Burgess Pub. Co.

[20] Bloom JD (2014) An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol Biol Evol : To appear.

[21] Zhu YO, Siegal ML, Hall DW, Petrov DA (2014) Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci USA : FORTHCOMING.

[22] Acevedo A, Brodsky L, Andino R (2014) Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505: 686 – 690.

[23] 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.

[24] 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.

[25] 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.

[26] Bloom JD (2014) An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. Mol Biol Evol 31: 1956-1978.

[27] Blumer M (1991) The selection-mutation-drift theory of synonymous codon usage 129: 897–907.

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

[29] Hershberg R, Petrov D (2008) Selection on codon bias. Annu Rev Genet 42.

[30] Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nature Rev Genet 12: 32–42.

[31] Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Rev Genet 7: 98–108.

[32] Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586–1591.

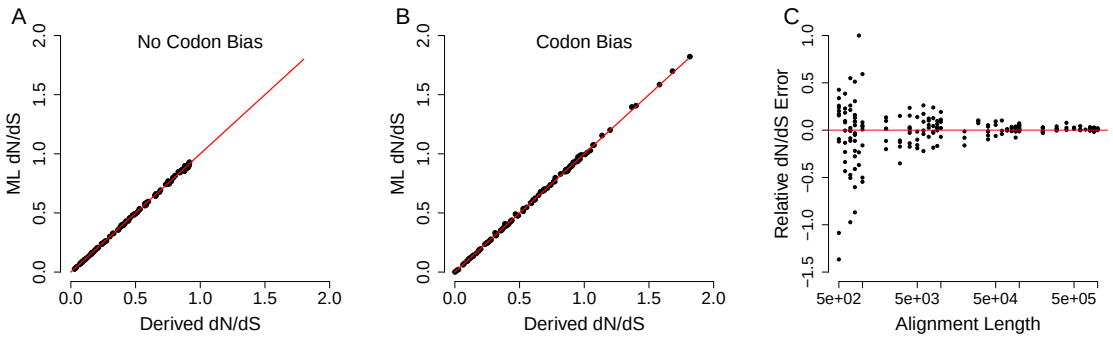

Figure 1: Relationship works exceedingly well. Left panel shows 100 points, each of which corresponds to single simulation. Note that here the ml inference is shown for equal codon frequency specs and kappa fixed to true value (a similar plot for free kappa is shown in suppfigs, but results are qualitatively identical.) Right panels shows convergence of omega values as data set size (represented as simulated alignment length) increases. The y-axis indicates relative error of the ML dN/dS estimates, and the x-axis indicates sequence length on a log-scale. As the sequence length, or the data set size, increases, the two dN/dS estimates converge to the same value.

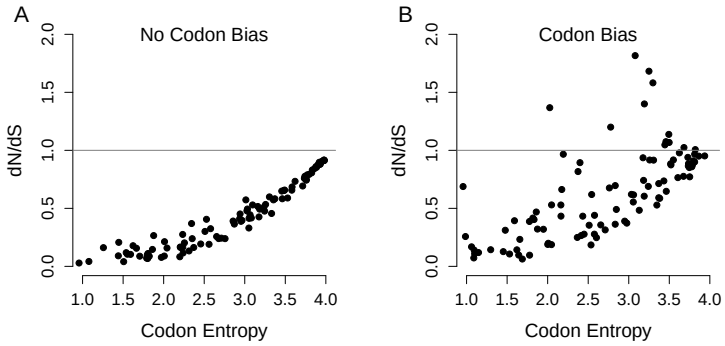

Figure 2: codon frequency entropy scales well with dnds but the strength of the relationship diminishes with codon bias as synonymous now have frequency differences, so dnds is less of a reliable indicator of selection strength.

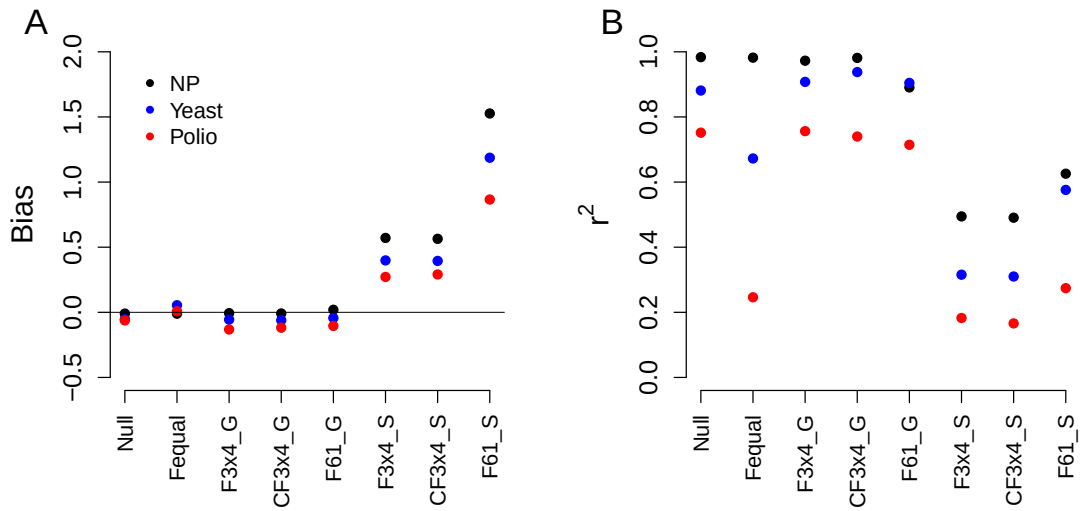

Figure 3: Global outperform, obviously. Clear that the null is probably the best bet. The question is, how well to commonly used estimators approximate this null specification? Decently, but clearly F3x4 and CF3x4 do better than F61. Also, Fequal does reasonably well, but it gets noisy as the asymmetry in mutation rates grows. Site-wise sucks, and we have a good explanation for it - a lot of selection information is in there!! Bad news.

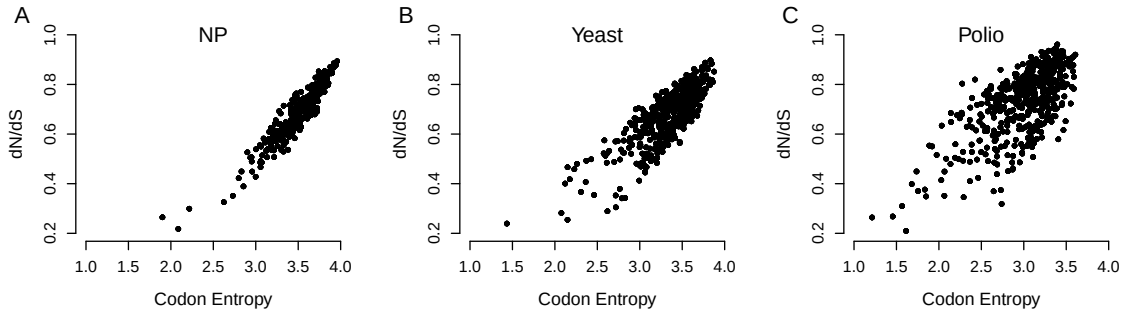

Figure 4: This is nice because it shows that entropy and dnds still scale really wonderfully, and noise ups as you get more asymmetric. This is because non-selection processes end up contributing to codon frequencies.

Supplementary Figures and Tables

Codon frequencies	κ parameterization	dN/dS correlation	dN/dS error	κ correlation	κ error
Fequal	True	1	0.02		
Fequal	Free	0.971	0.105	0.699	0.262
Fequal	1	0.974	0.263		
F3x4	True	0.646	2.134		
F3x4	Free	0.552	2.342	0.212	0.774
F3x4	1	0.597	1.847		
CF3x4	True	0.642	2.105		
CF3x4	Free	0.553	2.316	0.212	0.765
CF3x4	1	0.581	1.845		
F61	True	-0.591	19.067		
F61	Free	-0.281	27.714	0.315	0.927
F61	1	-0.56	12.272		

Results from runs with codon bias. Codon frequency specifications were either set as equal (1/61 per codon), calculated from the F3x4 estimator [9], calculated from the CF3x4 estimator [18], or set equal to the simulated alignment's empirical frequencies. κ was specified as either a fixed value, its true simulated value or 1, or as a free parameter of the model. Correlations given are between the ML ω estimate and our derived ω values. Error refers to the mean absolute error between these two ω estimates. Similar values for κ are shown for those inferences where κ was a free parameter of the model. Note that all is significant.

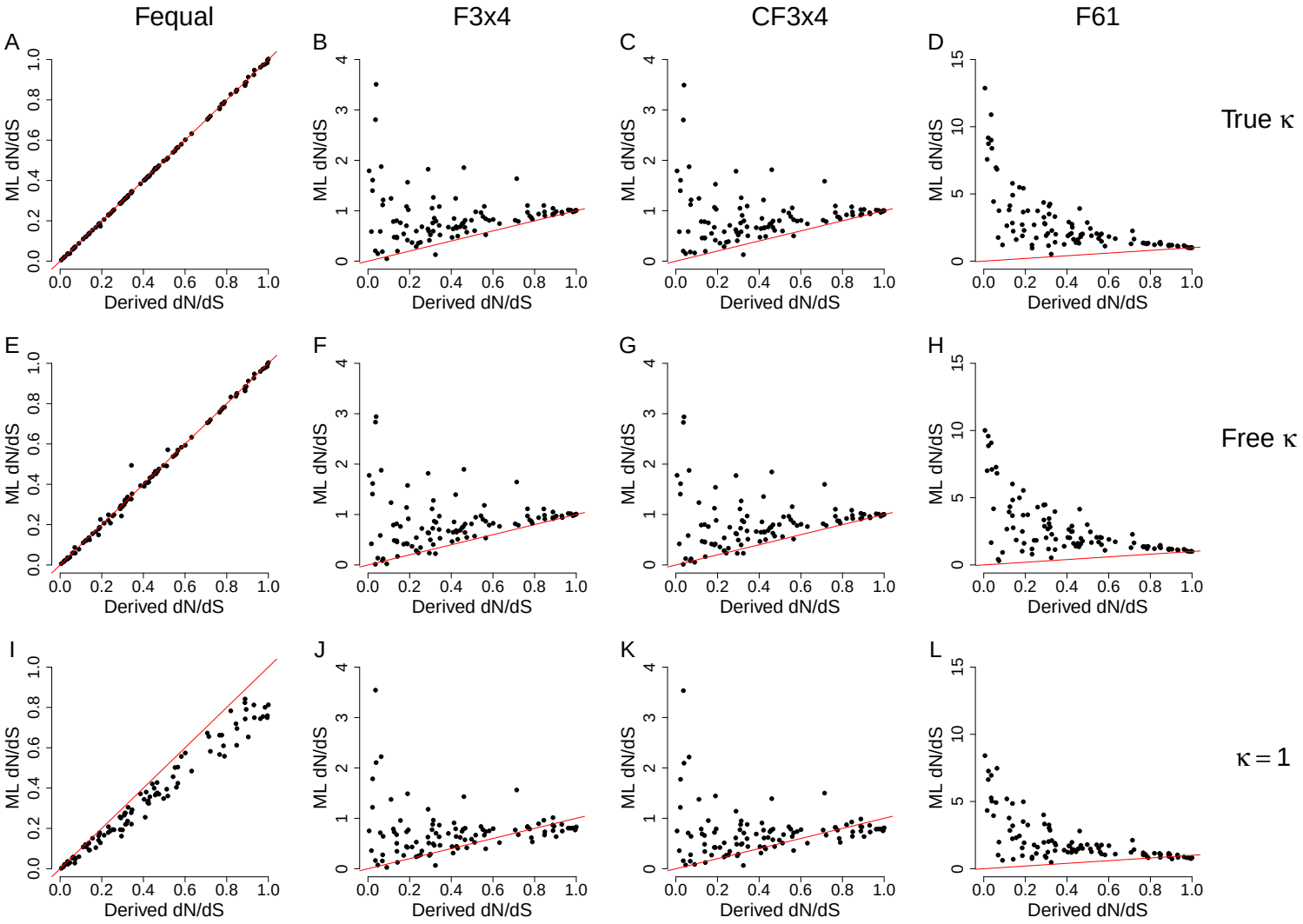

Fig. S1 Omega regression for all ML parameterizations, without codon bias.

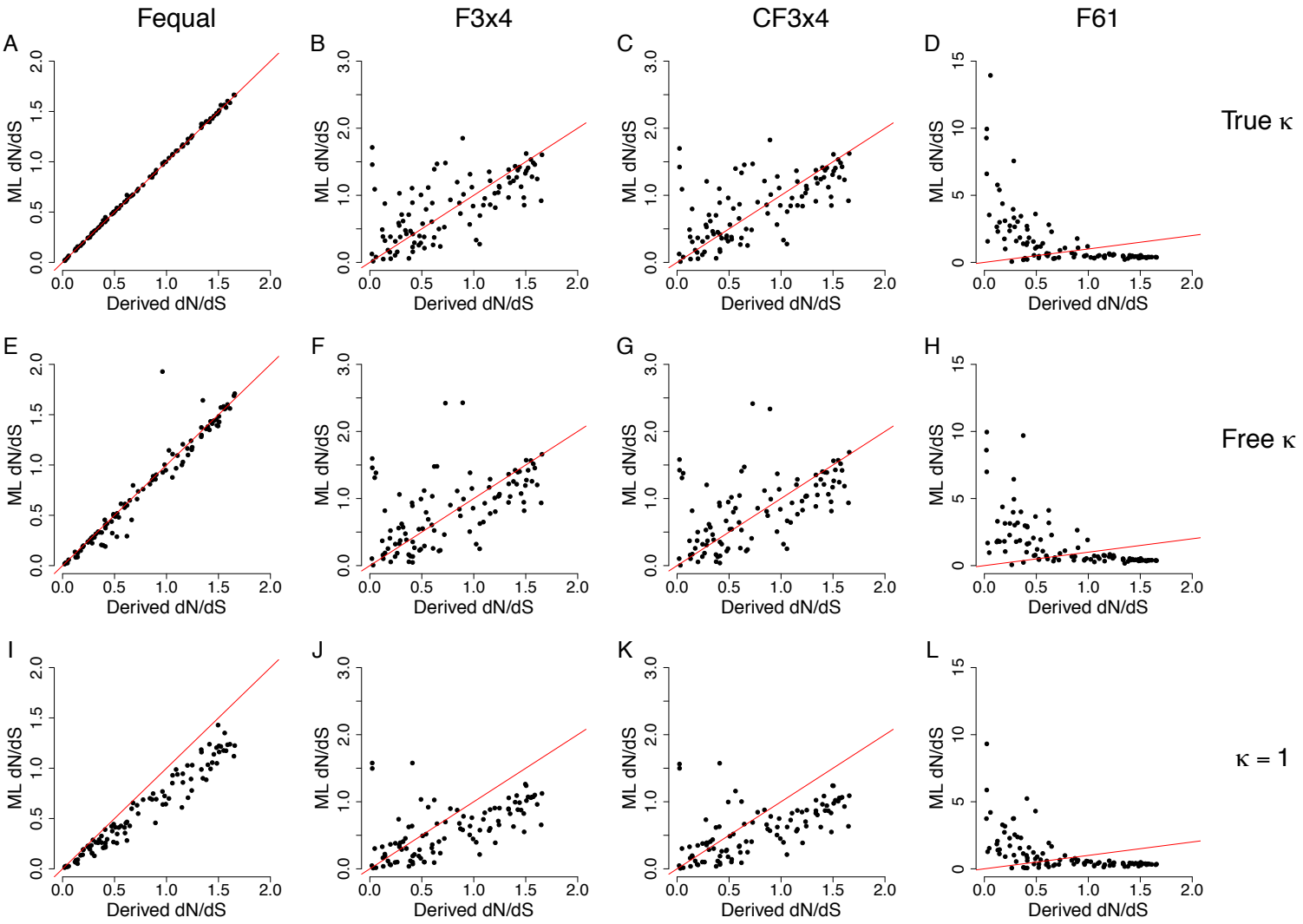

Fig. S2 Omega regression for all ML parameterizations, with codon bias.

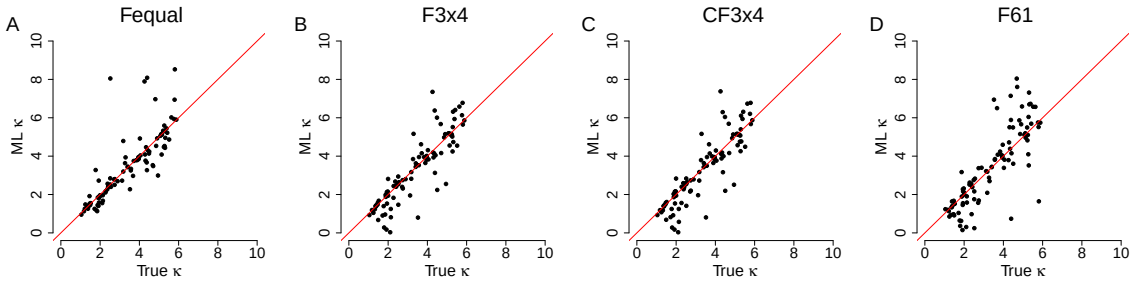

Fig. S3 Kappa regression for all ML freqspec parameterizations where kappa is a free parameter, without codon bias.

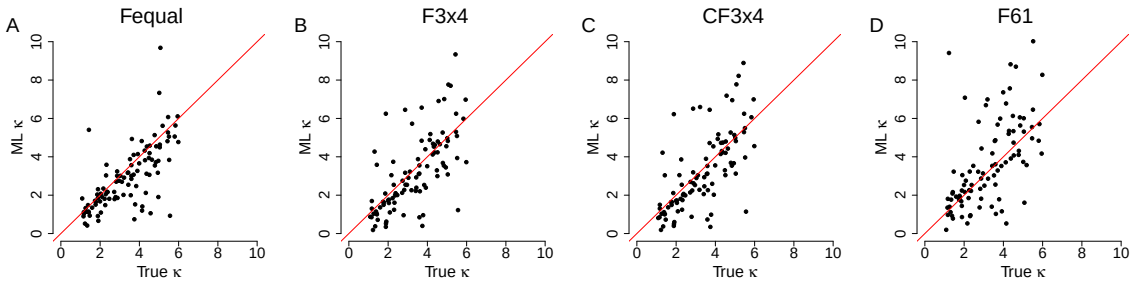

Fig. S4 Kappa regression for all ML freqspec parameterizations where kappa is a free parameter, without codon bias.