The relationship between dN/dS and scaled selection coefficients

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, scaled selection coefficients, 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. This approach is really nice because it allows us to uncover properties of these metrics previously unidentified, etc. We found that codon-level models and metrics do not play nicely with amino-acid level models, and we found some interesting behaviors of the M0 model. Our study represents a very useful strategy - benchmarking and investigating model behavior be examining the intersection/relationship between distinct approaches.

we can gain insight into model behavior by comparing the extent to which distinct modeling frameworks relate, or do not relate, to each other.

There are two primary conclusions to our study: 1. it allows us to understand the relationship between distinct modeling frameworks. 2. It allows us to uncover properties of methods which would otherwise be impossible.

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 $S = 2Ns$, 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. Recently, however, 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 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 precisely infer dN/dS values from scaled

selection coefficients. Using a simulation approach, we verify that these derived dN/dS values correspond precisely to ω MLEs inferred using standard mechanistic codon models. Moreover, using this relationship, we are able to uncover certain biases or issues with modeling approaches. For instance, we find that the codon frequency parameter in mech codon models is ill-suited to accomplish its objective.

Moreover, our results here allow us to perform unprecendented methodological benchmarking.

Methods

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

$$f_{ij} = \frac{2Ns_{ij}}{1 - e^{2Ns_{ij}}}, \tag{2}$$

where the value $2N_e s_{ij}$ represents the scaled selection coefficient for a mutation from codon i to codon j [10, 14]. As shown by [10], the fixation probability

$$f_{ij} \propto ln\left(\frac{\pi_j \mu_{ij}}{\pi_i \mu_{ji}}\right) \Big/ \left(1 - \frac{\pi_i \mu_{ji}}{\pi_j \mu_{ij}}\right). \tag{3}$$

In this approximation, π_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, σ^2 effectively represents the strength of natural selection; larger values of σ^2 will correspond to greater fitness differences among amino acids, and thus more selective pressure. Moreover, these S_i values correspond to the relative amino acid fitness parameters as inferred by currently available MutSel inference methods [11, 12]. For simulations without codon bias, we directly assigned S_a values to codons such that all synonymous codons had the same scaled selection coefficient, 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 λ from $\mathcal{U}(0, 2)$.

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

$$\pi_i = \frac{e^{S_i}}{\sum_k e^{S_k}}, \tag{4}$$

where the denominator runs over all sense codons. Equation (4) directly relates codon equilibrium frequencies and ... using theory developed by Sella and Hirsh [15] Moreover, according to theory developed by Sella and Hirsh [15], these equilibrium frequencies are directly related to scaled selection coefficients according to

We calculated a global dN/dS for each alignment using the mathematical framework outlined in (4)–(10) 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 [16]. The M0 models uses the GY94 instantaneous rate matrix [1, 8], which includes the primary parameters ω, κ, and equilibrium codon frequencies. For simulations inferences, we inferred ω both by fixing κ to its true value, and maintaining κ 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 [13]. 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 equilibrium codon frequencies which would be expected in the absence of selection [17], 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 [18] and nucleotide mutation rates for either NP [18], yeast [19], or polio virus [20]. 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 [18], wherein the Metropolis acceptance criterion [21] 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 (1).

We inferred ω 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 ω using the Fequal [13] parameterization, which assigns equal codon frequencies of 1/61 each. Second, we inferred ω 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 F3x4 [9], CF3x4 [22], and F61 [8]. 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 three distinct sets of averaged codon frequencies, from which we directly calculated the parameters for

F3x4, CF3x4, and F61.

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, 12, 23]. Therefore, it is possible to derive stationary equilibrium frequenies for all codons. These equilibrium codon frequencies result from the dynamic interplay between both mutational and selective pressures. 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 [15],

$$\pi_i = \frac{e^{S_i}}{\sum_k e^{S_k}}, \tag{5}$$

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 these equilibrium codon frequencies π_i, we can write the fixation probability for a mutation from codon i to codon j as [10, 15]

$$f_{ij} = \frac{1 - (\pi_i/\pi_j)^{1/N_e}}{1 - \pi_i/\pi_j} \approx \frac{1}{N_e} \frac{\ln \pi_j - \ln \pi_i}{1 - \pi_i/\pi_j}\,, \tag{6}$$

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_{N} as

$$K_{\mathrm{N}} = N_e \sum_i \sum_{j \in \mathcal{N}_i} \pi_i f_{ij} \mu_{ij}\,, \tag{7}$$

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_{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

$$L_{\mathrm{N}} = \sum_i \sum_{j \in \mathcal{N}_i} \pi_i \mu_{ij}\,, \tag{8}$$

and thus we find that

$$dN = \frac{K_{\mathrm{N}}}{L_{\mathrm{N}}} = \frac{N_e \sum_i \sum_{j \in \mathcal{N}_i} \pi_i f_{ij} \mu_{ij}}{\sum_i \sum_{j \in \mathcal{N}_i} \pi_i \mu_{ij}}\,. \tag{9}$$

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

$$dS = \frac{K_{\mathrm{S}}}{L_{\mathrm{S}}} = \frac{N_e \sum_i \sum_{j \in \mathcal{S}_i} \pi_i f_{ij} \mu_{ij}}{\sum_i \sum_{j \in \mathcal{S}_i} \pi_i \mu_{ij}}\,, \tag{10}$$

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_{S} and L_{S} are defined as in Eqs. (7) and (8) but summing over $j \in \mathcal{S}_i$ instead of $j \in \mathcal{N}_i$.

Equations (6)–(10) 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 $f_{ij} = 1/N_e$ [24]. 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 (6)–(10), we simulated protein-coding alignments according to a MutSel model [10,15]. 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 used 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 [15]. For each alignment, we calculated dN/dS using equations (6)–(10) as well as using the M0 mechanistic codon model [1], as implemented in the HyPhy batch language [16].

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.

paragraph here about how that figure contains results for kappa fixed to its true value in the M0 Markov matrix. We also performed the same inferences while allowing kappa to be a free parameter of the model. ω MLEs between kappa true and kappa free showed nearly perfect agreement (0.997 for no codon bias and 0.992 for codon bias), demonstrating the robust behavior of this model. We also examined the correlation between the simulated kappa and κ MLE and found a much weaker relationships (0.45 and something else). Thus, it appears that the M0 model's ability to infer ω is fairly robust to incorrect κ inferences.

Moreover, the strength of selection pressure scales fairly well with dN/dS. Figure 2 displays the relationship between dN/dS and the standard deviation, σ^2, of the distribution of amino acid selection coefficients. Higher values of σ^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, $r^2 = 0.45$). This weakened relationship emerges from the fact that fitness differences

among synonymous codons obscure the underlying amino acid fitness differences. Thus, while a significant negative correlation remains, the codon bias generates increased noise.

Importantly, Figure 2A shows that, in the limiting case when σ^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. In fact, in **SI 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 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*.

Using real data/ML doesn't work so well/Jesse Bloom is prolific

Results reported in the previous subsection were obtained from fully-simulated equilibrium codon frequencies, along with symmetric mutation rates. The latter assumption may not be entirely realistic; indeed, mutational biases, in particular transitions from $C/G \rightarrow T/A$ are known to contribute to uneven nucleotide compositions in real genomes [19, 20, 25, 26]. Therefore, we performed additional simulations which made use of realistic amino acid fitness and nucleotide mutation parameters. In particular, we used influenza nucleoprotein (NP) site-specific amino acid preference values, given by Bloom [18]. 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 [18], yeast [19], or polio virus [20]. 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. For each of the 498 amino acid fitness distributions, we calculated stationary codon frequencies π_i under detailed balance conditions, using the approach in [18, 27].

For each resulting set equilibrium codon frequencies, we again computed dN/dS using equations (6)–(10) and simulated alignments according to equation (1). We inferred ω MLEs using the M0 mechanistic codon model according to five different codon frequency model parameterizations. These parameterizations included Fequal [13] and the common frequency estimators F3x4 [9], CF3x4 [22], and F61 [8]. The estimators F3x4 and CF3x4 approximate codon frequencies using positional nucleotide frequencies, and the F61 estimator simply uses an alignment's empirical codon frequencies. Additionally, we inferred ω using a fifth frequency parameterization which consisted of the codon frequencies that would arise strictly from mutational processes, in the absence of natural selection. As this specification is the intended purpose for this parameter, we term it Ftrue.

Figure 3 shows the resulting relationships between dN/dS and ω MLEs for each set of mutation rates (NP, yeast and polio), across M0 model codon frequency parameterizations (Figure **??** contains regression plots 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 ω, and Figure 3B displays r^2 values between dN/dS and ω. Note that a bias of 0 would indicate a perfect correlation between dN/dS and ω MLEs.

1. all specifications perform very well for alignments simulated with NP mutation rates, with a single exception of F61 (we will discuss this below). 2. Fequal really performs the best. Although it features more noise, it shows the least amount of bias for polio. This means that, while the other frequency parameterizations might show stronger correlations, the oemga estimates are systematically biased downwards. 3. We expect, if this parameter really suffices to accomodate underlying

nucleotide biases, Ftrue would outperform all other parameterizations.

The issue comes down to, "how to deal with asymmetric mutation or underlying nucleotide biases?"

Nearly all results shown in Figure 3A demonstrate negative bias, meaning that the M0 model yields, on average, ω MLEs which underestimate dN/dS. Moreoever, Figure 3B reveals a clear trend where the relationship between dN/dS and ω decreases in strength as mutation rates become increasingly asymmetric for all codon frequency parameterizations. Strikingly, we find that the ω MLEs computed using the Fequal codon frequency parameterization are the least biased, relative to the true dN/dS values.

the specifications with the least amount of bias are

We find that ω MLEs generally correlate extremely well with dN/dS values, but the strengths of these correlations differ among codon frequency model parameterizations (Figure 3B). Overall, the Fnull parameterization performs the best of all frequency specifications; Fnull features both the least amount of bias and extremely high correlations. The Fequal parameterization similarly yields relatively low levels of bias, but its r^2 values drop precipitously for simulations using the yeast and polio mutation rates. This trend highlights the expected result that increasing asymmetry in mutation rates render Fequal an incorrect parameterization, as asymmetric mutation rates generate substantial compositional bias that Fequal cannot accomodate. Finally, the frequency estimators F3x4, CF3x4, and F61 perform well, and produced comparable correlations to the Fnull specification, with only marginally more bias. The purpose of these frequency estimators is to approximate the codon frequencies in the absense of amino-acid level natural selection [13, 17] (i.e., the Fnull parameterization), and our results indicate that these estimators accomplish this objective fairly well.

Importantly, there is a clear trend that as asymmetric in mutation rates increase, the correlation strength systematically decreases, even for the Fnull parameterization (Figure 3B). Clearly, including codon frequency parameters in mechanistic codon models does dramatically improve ω inferences, as is evidenced by the marked improvements of F3x4, CF3x4, and F61 over Fequal. Even so, as the negative bias values in Figure 3A demonstrate, the M0 model systematically underestimates dN/dS in line with increasing mutation rate asymmetry. This result suggests that incoporating codon frequency parameters may not be the optimal way to incorporate information on nucleotide compositional bias.

However, there is a noteworthy exception to the trend of ω underestimation; the F61 frequency parameterization actually overestimated ω for simulations which employed NP mutation rates. We attribute this result to the fact that the NP mutation rates were only minimally asymmetric, featuring an average $\mu_{xy}/\mu_{yx} = 1.04$. Moreover, when nucleotide mutation rates are symmetric, steady-state frequencies are controlled only by selection, as there is no opportunity to generate compositional bias through mutation [15]. As the F61 estimator directly uses empirical codon frequencies, the resulting M0 codon frequency parameters contain information about selection. Therefore, selective pressures which should be incorporated in the ω parameter are inadvertently contained within the codon frequency parameters, leading the model infer the strength of selection to be weaker than it truly is and produce elevated ω MLEs. Although the ω overestimatation was relatively minimal in this particular example, it serves as an important illstrative example of the necessity to properly parameterize mechanistic codon models. If these models are parameterized improperly, the ω parameter will no longer accurately represent the dN/dS evolutionary rate ratio.

Discussion

The oldest and most-widely used method to infer selection pressure in protein-coding genes calculates the evolutionary rate ratio of non-synonymous (dN) to synonymous (dS) 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 that experience positive, diversifying selection ($dN/dS > 1$) [1–3]. By contrast, MutSel models estimate scaled selection coefficients for amino acids, [10, 12, 14, 23, 28], for codons [14], 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 mechanistic codon models, suggesting that MutSel models may more aptly represent the process of coding-sequence evolution [18, 27].

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 scaled selection coefficients. 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 [29–33]. 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 will always be less than 1. This restriction does not, however, apply when synonymous codons have different fitness values (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. Experimental evidence in the Hsp90 protein, for instance, demonstrates that while, while there are some fitness differences among synonymous codons, these differences are exceedingly minimal relative to fitness differences among amino acids [34,35]. 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 [30,31,33,36], may not be true cases of positive selection, but rather simply signals of strong codon bias.

That the dN/dS values calculated using equations (6) - (10) agree precisely with ω 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 [13, 17, 37]. Previously proposed frameworks for calculating dN/dS have broadly fallen into two camps: heuristic counting methods [5, 6, 17, 38, 39] and maximum likelihood methods [7–9, 13]. Unlike these frameworks, the dN/dS calculations we have proposed here are solidly grounded in population genetics theory. That ω 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.

However, while the relationship between dN/dS values and ω MLEs was remarkably strong under symmetric nucleotide mutation rates, the relationship weakened somewhat when we introduced asymmetric nucleotide mutation rates. As our dN/dS calculations explicitly consider nucleotide mutation rates, we contend that this weakened relationship resulted from incorrect maximum likelihood dN/dS inferences. In particular, the weakened relationship very likely resulted from model misspecifications in the M0 equilibrium codon frequency parameters. Mechanistic codon models attempt to deal with mutation-induced nucleotide compostional bias through equilibrium codon frequency parameters [13]. Unlike the equilibrium steady-state frequencies we have focused on 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 biased gene conversion, alone [8, 9, 13, 17]. If this parameter performs its task correctly, we would expect that a codon frequency parameterization which used precisely the correct parameter values would produce ω MLEs which agreed precisely with dN/dS. We did just this when inferring ω on alignments simulated with asymmetric mutation rates through the frequency parameterization Fnull. This specification gave the precise codon frequencies which mutation alone would generate, in the absence of selection. While this frequency specification did perform relatively well, and dramatically outperformed a model with equal codon frequencies, M0 still systematically underestimated dN/dS; as asymmetry in mutation rates increased, M0 underestimated ω more and more, and the correlation between dN/dS and ω decreased. We can be sure that this trend did result entirely from asymmetric mutation rates, and simply from nucleotide compositional biases. Indeed, our alignments simulated with symmetric mutation rates featured a wide array of GC-contents, ranging from 0.18-0.72. Given these alignments' symmetric mutation rates, natural seletion favoring particular codons alone generated all compositional biases in those data sets, and maximum likelihood methods inferred ω perfectly under the Fequal frequency parameterization.

Taken together, these results strongly suggest that the mechanistic codon model's codon frequency parameters might not be adequate 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.

Moreover, we emphasize that improper model parameterizations lead to spurious ω MLEs which do not accurately represent dN/dS, but instead some meaningless quantity. If other model parameters (e.g. κ or the equilibrium codon frequencies) are specified incorrectly or inadvertently contain information about amino-acid level natural selection, the resulting ω MLE will not represent the true dN/dS evolutionary rate ratio. Only by ensuring that ω 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 [22] 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.

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] 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.

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

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

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

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

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

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

[21] Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equation of state calculations by fast computing machines .

[22] 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.

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

[25] 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.

[26] Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6: e1001115.

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

[28] 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.

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

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

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

[32] 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.

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

[34] Hietpas RT, Jensen JD, Bolon DNA (2011) Experimental illumination of a fitness landscape. Proc Natl Acad Sci USA 108: 7896–7901.

[35] Hietpas RT, Bank C, Jensen JD, Bolon DNA (2013) Shifting fitness landscapes in response to altered environments. Evolution 67: 3512–3522.

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

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

[38] Pamilo P, Bianchi NO (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol 10: 271–281.

[39] Ina Y (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. Journal of Molecular Evolution 40: 190–226.

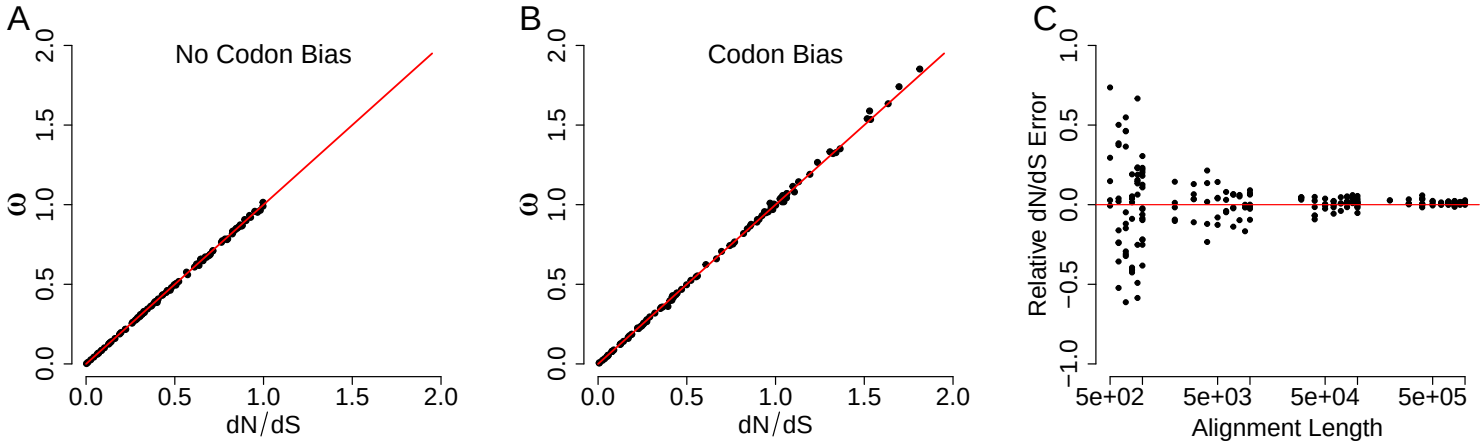

Figure 1: Relationship works exceedingly well. Left panel shows 100 points, each of which corresponds to single simulation. Note that here the ml inference is shown for equal codon frequency specs and kappa fixed to true value (a similar plot for free kappa is shown in suppfigs, but results are qualitatively 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. note that convergence data was generated using the same approach as for the no codon bias data set, except by varying the alignment size.

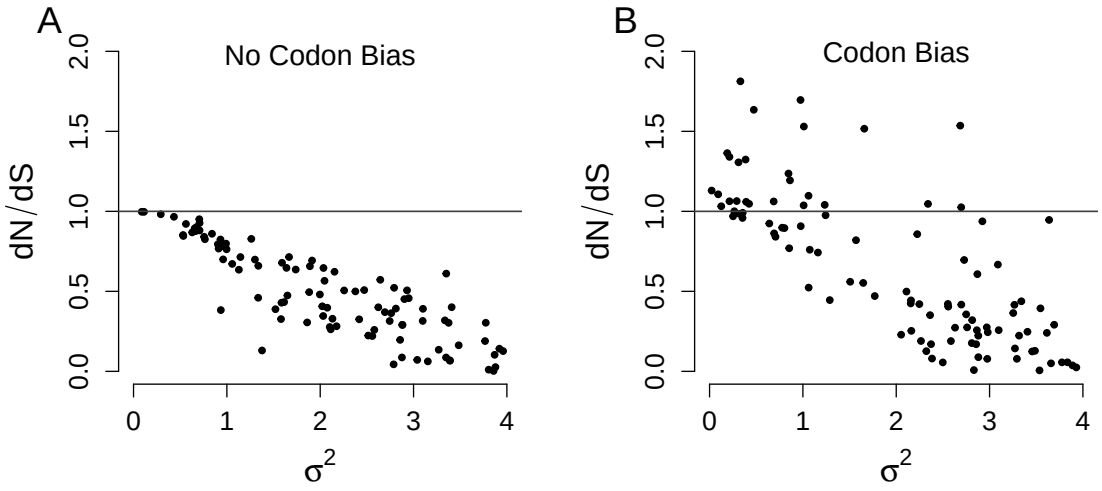

Figure 2: strength of selection scales well with dnds but the strength of the relationship diminishes with codon bias as synonymous now have frequency differences, so dnds is less of a reliable indicator of selection strength.

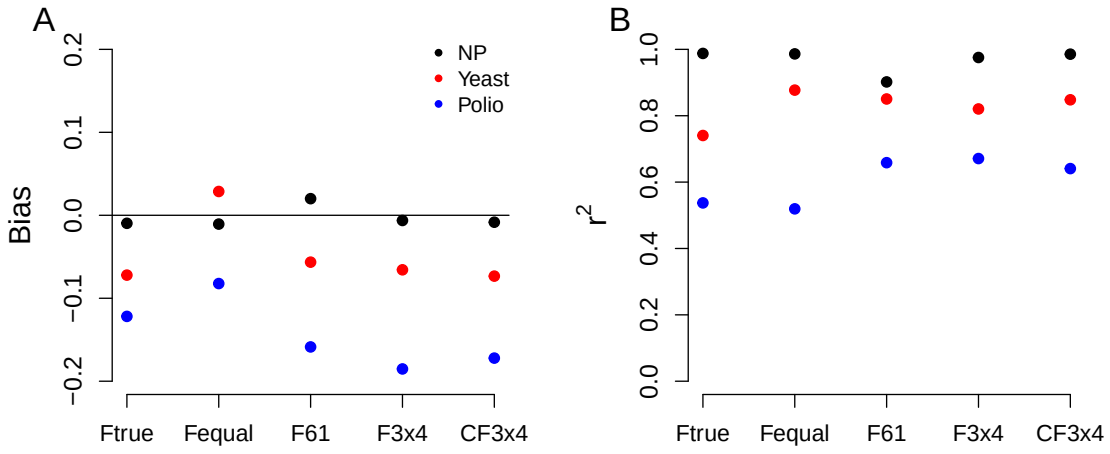

Figure 3: Global outperform, obviously. Clear that the null is probably the best bet. The question is, how well to commonly used estimators approximate this null specification? Decently, but clearly F3x4 and CF3x4 do better than F61. Also, Fequal does reasonably well, but it gets noisy as the asymmetry in mutation rates grows.

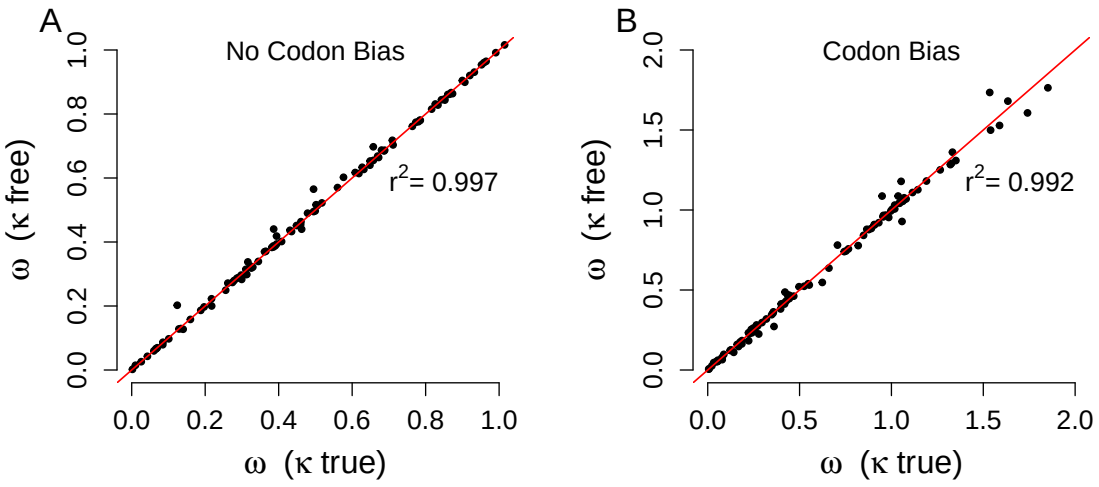

Fig. S1 Regressions in between omegas estimated with kappa as free and kappa as true.

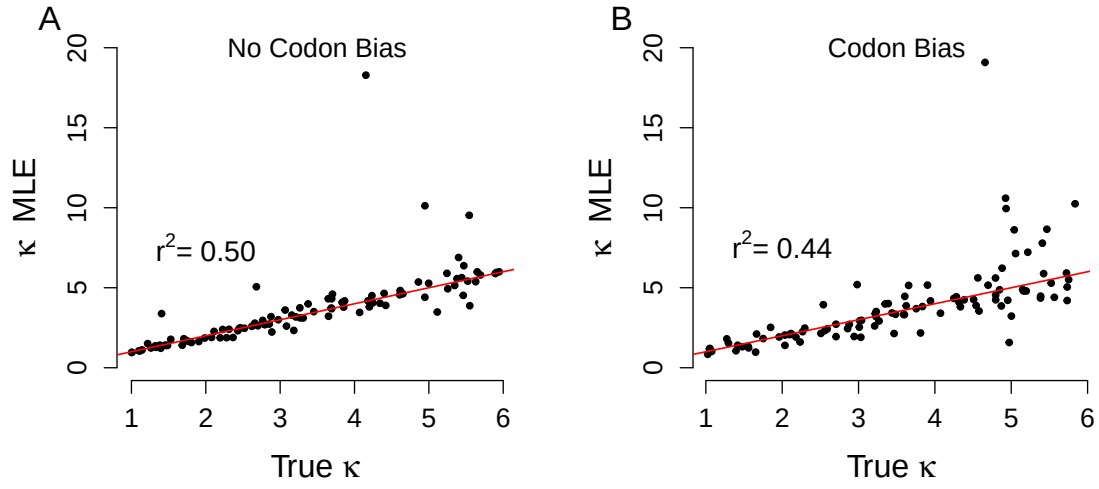

Fig. S2 Regressions in between true and MLE kappa.

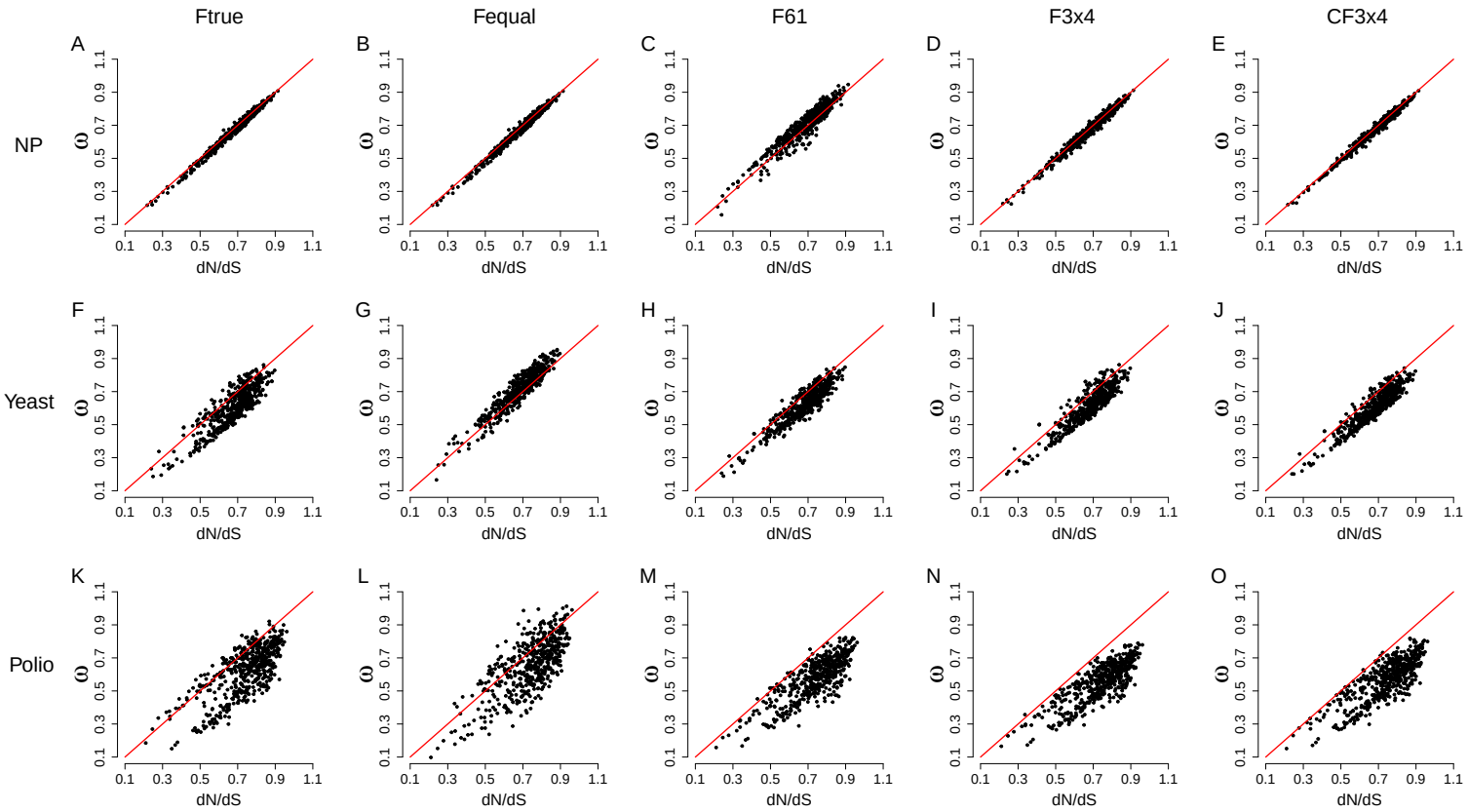

Fig. S3 Omega regression for all ML parameterizations for the np, yeast, and polio mutation rates.