# Introduction

Over the years, a variety of models have been proposed to describe the effects of natural selection on protein-coding sequences. Traditionally, the focus has been on estimating the evolutionary rate ratio, $\omega = dN/dS$, the rate of nonsynonymous to synonymous substitutions, which indicates how quickly a protein's constituent amino acids change. Following early counting methods for estimating $dN/dS$ (e.g. refs [1] and [2]), mechanistic codon substitution models (see ref. [3] for a comprehensive review) have taken a leading role as the inference method of choice since their introduction in the 1990s [4, 5]. These so-called $\omega$ models have seen great success in the field of molecular evolution and are widely used to examine the strength of selection pressure in protein-coding sequences.

More recently, a second class of models, known as mutation-selection-balance (MutSel) models, has emerged as a popular alternative to $\omega$ models. MutSel models explicitly model the dynamic balance between mutation and selection, lending a potentially more precise and realistic description of the evolutionary process than do $\omega$ models, which merely model the final outcome (e.g. substitution) of the underlying mutation-selection interplay [6–9]. Unlike $\omega$ models, MutSel models yield estimates of amino acid selection coefficients, which indicate the extent to which natural selection favors, or disfavors, particular amino acids at protein positions. These selection coefficients, which can in turn be scaled relative to a focal amino acid, the primary parameters of interest that MutSel models produce. Although MutSel models were first introduced over 15 years ago [6], they have seen virtually no use due to their high computational expense. However, recently, several computationally tractable model implementations have emerged [10, 11], allowing for the first time the potential for widespread use.

Although both $\omega$ 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. 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 aim to formalize the relationship between $\omega$ and MutSel models by examining the extent to which their focal parameters, $\omega$ and scaled amino acid 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 $\omega$ values from selection coefficients alone. Using a simulation approach, we verify that $\omega$ values estimated using selection coefficients alone correspond precisely to those inferred using the standard maximum likelihood $\omega$ inference approach. Further, we prove that 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 $dN/dS > 1$. This result has important implications for circumstances under which MutSel model use is justified.

Moreover, using this relationship as a benchmark, we have uncovered some potential biases inherent in $\omega$ inference methods. Typically, inference framework performance is assessed through simulating datasets 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 serves as a robust strategy to determine the accuracy of different modeling frameworks. Many frameworks for dnds inference have been proposed, and it is well-known that these inferences differ from one another (yang2006, that other 2006 paper). However, the only argument for preferring one model's results over another is some sort of statistical argument for "this model seems a lot better," but there is no way to truly know. Our approach presents a promising alternative. Here, it seems that we really know what this parameter's value is.

# Methods

## Sequence simulation and omega inference

We simulated protein-coding sequences as a continuous-time Markov process [12] according to the MutSel model proposed by [6]. This model's instantaneous rate matrix $Q = q_{ij}$, which describes the probability of substitution from codon $i$ to codon $j$, is given by

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

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

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

where $\pi_i$ is the equilibrium frequency of codon $i$.

For each simulation, we generated scaled amino acid selection coefficients, $s_a$, by fixing one coefficient to 0 and drawing the remaining 19 values from a normal distribution $\mathcal{N} \sim (0, \lambda)$. The standard deviation parameter $\lambda$ effectively represents the strength of selection; smaller $\lambda$ values will produce similar amino acid fitness values, whereas larger values will naturally lead to a stronger preference for certain amino acids. For each simulation, we selected a value for $\lambda$ from $\mathcal{U} \sim (0.5, 3.5)$. We converted these selection coefficients to steady-state amino acid frequencies $F(a)$ according to

$$F(a) = \frac{e^{s_a}}{\sum_b e^{s_b}} \tag{3}$$

, where the denominator sums over all 20 amino acids [13]. These frequencies were then assigned to amino acids, and subsequently converted to codon frequencies (assuming no codon bias).

We simulated protein-coding sequences along a 10-taxon phylogeny, with all branch lengths equal to 0.01, beginning with a root sequence selected using steady-state codon frequencies. For all simulations, we set a global symmetric mutation rate of $10^{-5}$, and we selected a value for each simulation's $\kappa$ from $\mathcal{U} \sim (1, 5)$. Unless otherwise stated, we simulated alignments of 500,000 codon positions. A single evolutionary model was applied to all positions in the simulated sequences, meaning that we did not incorporate any site-wise variation into the evolutionary process. 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 $\omega$ with a sufficiently sized data set.

For each simulated alignment, we inferred $\omega$ in two main ways; first, we calculated $\omega$ using the mathematical framework described in (4)–(9), and second, we inferred $\omega$ used the standard maximum likelihood M0 model [14], which uses the GY94 rate matrix [4], as implemented in HyPhy [15]. The GY94 matrix includes the primary parameters $\omega$, $\kappa$, and equilibrium codon frequencies. As it is well-known that the manner of specification for the latter two parameters strongly influences $\omega$ estimates [12, 16], we inferred $\omega$ under a variety of model parameterizations, including three $\kappa$ parameterizations ($\kappa$ fixed to 1, $\kappa$ fixed to its true value, and $\kappa$ as a free parameter), and four codon frequency specifications (equal codon frequencies, F3x4 codon frequencies [5], CF3x4 codon frequencies [17] and empirical, or F61, codon frequencies [4]). These different specifications yielded twelve maximum likelihood $\omega$ inferences per simulated alignment. All code used is freely available at **github**.

# Results

## Mathematical relationship between selection coefficients and omega

We describe here how to calculate $dN/dS$ from the parameters of a MutSel model. We assume the following: (i) the mutational process is symmetric, such that $\mu_{xy} = \mu_{yx}$ for all nucleotide pairs $xy$; (ii) all synonymous codons for a given amino acid have the same fitness; there is no synonymous rate variation or codon bias.

In the framework of a MutSel model, we can write the steady-state frequency of amino acid $a$ as

$$f_a = \frac{e^{s_a}}{\sum_b e^{s_b}}, \tag{4}$$

where the sum in the denominator runs over all 20 amino acids [18]. Here, $s_a$ is the *scaled selection coefficient* for amino acid $a$; larger $s_a$ values correspond to higher frequencies of amino acid $a$, and corresponds precisely to MutSel model parameters [6]. Amino acid frequencies can be subsequently converted to codon frequencies, assuming no codon bias, as

$$f_i = F_a/n_a, \tag{5}$$

where $i$ is any codon coding for amino acid $a$ and $n_a$ is the total number of codons which code for $a$.

The fixation probability for a mutation from codon $i$ to codon $j$ is [6,18]

$$\pi_{i\to j} = \frac{1-(f_i/f_j)^{1/N_e}}{1-f_i/f_j} \approx \frac{1}{N_e}\frac{\ln f_j - \ln f_i}{1-f_i/f_j}\,, \tag{6}$$

where $N_e$ is the effective population size. We can calculate an evolutionary rate by summing over all fixation probabilities weighted by the frequency of the originating codon. For example, we can write the synonymous rate $K_\mathrm{S}$ as

$$K_\mathrm{S} = N_e \sum_i \sum_{j\in\mathcal{S}_i} f_i \pi_{i\to j}\mu_{ij}\,, \tag{7}$$

where $\mathcal{S}_i$ is the set of codons that are synonymous to codon $i$ and differ from it by one nucleotide substitution. To normalize $K_\mathrm{S}$, we divide it by the number of synonymous sites, which we calculate according to the mutational opportunity definition of a site [4,12] as

$$L_{\mathrm{S}} = \sum_{i} \sum_{j \in \mathcal{S}_i} f_i \mu_{ij} \,. \tag{8}$$

Under the assumption that all synonymous codons have equal fitness (all synonymous mutations are neutral), we have $\pi_{i \to j} = 1/N_e$ [19], and thus we find that $dS$, the synonymous rate per synonymous site, is equal to 1.

Similarly, we can derive an expression for $dN$, the non-synonymous rate per non-synonymous site, and we find

$$dN = \frac{K_{\mathrm{N}}}{L_{\mathrm{N}}} = \frac{N_e \sum_i \sum_{j \in \mathcal{N}_i} f_i \pi_{i \to j} \mu_{ij}}{\sum_i \sum_{j \in \mathcal{N}_i} f_i \mu_{ij}} \,, \tag{9}$$

where $\mathcal{N}_i$ is the set of codons that are not synonymous to codon $i$ and differ from it by one nucleotide substitution. The quantities $K_{\mathrm{N}}$ and $L_{\mathrm{N}}$ are defined as in Eqs. (7) and (8) but summing over $j \in \mathcal{N}_i$ instead of $j \in \mathcal{S}_i$. q Equations (4)–(9) establish a connection between the scaled selection coefficients $s_i$ (i.e., the primary parameters of a MutSel model) and the evolutionary rate ratio $dN/dS$.

## $\omega$ values fully encapsulated by scaled selection coefficients

To validate our derived relationship between $\omega$ values and scaled selection coefficients, we simulated protein-coding sequences along a 10-taxon phylogeny according to the Halpern-Bruno mutation-selection model [6]. Simulations were conducted with varying degrees of selective constraint and mutational parameterizations, although all simulated assume a symmetric mutation scheme. We calculated $\omega$ for each simulation set using both standard maximum likelihood methods, according to the GY94 [4] model, and the derivation given in equations (4)–(9).

As shown in Figure 2A, $\omega$ values derived using selection coefficients agree nearly perfectly with those inferred using standard maximum likelihood methods. We additionally demonstrate convergence of these values with increasing amounts of data, represented by simulated alignment length (Figure 2B). Taken together, these results clearly show that MutSel model parameters fully encapsulate information regarding the evolutionary rate ratio, $\omega$, and that the results from MutSel and $\omega$ models are largely in agreement.

Moreover, as seen in Figure 2A, estimates for $\omega$ never exceed 1, but rather all reflect a regime of purifying selection. In fact, in SuppMat, we prove that, when calculated using amino acid selection coefficients, $\omega$ is always less than or equal to 1. Thus important insight reveals that, while MutSel models fully agree with $\omega$ models, they only do so under conditions of purifying selection or neutral evolution. MutSel models, therefore, are inherently unable to describe protein evolution under positive, diversifying selection ($\omega > 1$).

mutational process. we had a symmetric process, so gc bias is not expected. even though we have diff gc content, this was driven by amino acid propensities, not by an overarching mutational process.

## Influence of maximum likelihood $\omega$ model parameterizations

It is well-known that different $\omega$ model calculations or parameterizations can influence the estimated $\omega$ value [12, 16, 20]. In the previous subsection, we reported results obtained when the maximum likelihood parameterization was set to $\kappa$ as true and codon frequencies as equal, or $1/61$ for each codon. Estimating $\kappa$ as a free parameter of the model yielded the same broad results as presented above ($r^2 = 0.995$), albeit with slightly more noise (see **FigureS1**). However, when different frequency specifications were used

The previous section demonstrating the excellent agreement between $\omega$ values contained

A valuable application of our this established relationship is benchmarking.

We have two possible explanations for this trend. The first is that these results reflect a broad misinterpretation of how the equilibrium codon frequencies parameters should be specified in mechanistic codon models. Typically, these values are taken directly from the given data and fixed in the model to reduce the number of parameters inferred. However, the codon frequencies observed in the data set represent those which exist *after* natural selection has acted on the sequence. Instead, the model should use the codon frequencies which would exist *in the absence of natural selection.* Selection, alternatively, acts to tune this frequencies to increase protein fitness. If one specifies equilibrium frequencies which exist after natural selection has acted on the protein sequence (i.e., frequencies measured directly from the data), then the influence of selection pressure is incorporated into these values. The desirable outcome, however, is that the $dN/dS$ parameter measures the selective strength. If other parameters in this model contain selection information, estimates for $dN/dS$ will not accurately capture selective effects.

The second explanation is that a single parameter $dN/dS$ is in effect representing two distinct values; on one hand, $dN/dS$ indicates the strength of selective constraint. On the other hand,

## Discussion

Here, we have derived a formal mathematical relationship between the parameter estimates of MutSel and mechanistic codon models. Through a simulation approach, we validated this relationship and demonstrated that these models yield nearly identical results. However, we additionally found that MutSel models necessarily cannot describe scenarios in which positive or diversifying selection occurs, or when $dN/dS > 1$. Although it is generally acknowledged that MutSel models carry the assumption of purifying selection, we formalize and prove this result. These findings have important implications for how and when these models should be applied. In cases of purifying selection or neutral evolution, these two competing models are virtually no different from one another, and thus use of either model is fully justified. Alternatively, if positive selection is occurring, MutSel models cannot be used as they mathematically cannot capture the actual evolutionary process.

This study highlights the importance of examining the similarities and differences among different evolutionary model classes. While some may prefer one model over another, it is important to have a full understanding of why one model might be applied over another. Our results demonstrate that, for circumstances of purifying selection or neutral, the models are robust and in agreement. However, be careful, because sometimes they don't agree, and this is equally important.

Now, for the frequency specification: does this really matter? Typically, when one measures codon frequencies from the data set, codon frequencies are treated as global parameters rather than site-specific. Thus, equilibrium codon frequency parameters effectively represent an average across the entire data set, naturally lead to a flatter distribution of codon frequencies. Our simulated data sets, on the other hand, were evolved according to a single parameter site, leading to more constrained global codon frequencies. Therefore, the extent to which bias introduced by incorrect frequency specifications was driven by how far from equal we're talking.

Caveat about how our math relies on symmetric mutation rates and/or equal mutation rates, depending on which one ends up happening.

### Mutation-selection-balance models are only valid for purifying selection

**I really have no idea how to write up proofs, so I've virtually just latex'd the mathematica document. But at least the equations are there! ... next day: I'd better stop. This section might have to be yours, unfortunately.**

To show that mutation-selection models only corresponds to $dN/dS \leq 1$, we make use of that fact that each calculation of $dN/dS$, as described in equations (4)–(9), entails summing the forward and backward fixation probabilities between codons, which are in turn divided by the codon frequency sums. We additionally assume that all mutation rates $\mu_{ij}$ are equal, and as all values for $N$ will ultimately cancel, we exclude them from the following proof. We will deal with the case of two nonsynonymous codons, $i$ and $j$, and we write their frequencies as $x$ and $y$, respectively.

As follows from (6), the sum of the probabilities going from codon $i$ to codon $j$ and from codon $j$ to codon $i$ is

$$xP(x,y) + y(P(y,x) = \frac{2xy[\ln x - \ln y]}{x-y} \tag{10}$$

, and we will demonstrate that this value is necessarily $\leq x+y$ for $x, y \geq 0$ and $x \leq y$.

To this end, we define the function

$$F(x,y) = x + y - \frac{2xy[\ln x - \ln y]}{x-y} \tag{11}$$

. Thus, we show that $F(x,y) \geq 0$. For the condition $x = y$, this is straightforward to show, as $\lim_{x \to y} F(x,y) = 0$. We now show that the first derivative of $F(x,y)$ is negative throughout $x \in (0,y)$, thus proving that $F(x,y)$ has to be monotonically decreasing, and hence $\geq 0$, in this interval.

a possible explanation: the dnds parameter, when greater than one, has 2 interpretations. one interpretation is

# References

[1] 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.

[2] Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426.

[3] Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26: 255–271.

[4] Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736.

[5] 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.

[6] Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15: 910–917.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] Ramsey DC, Scherrer MP, Zhou T, Wilke CO (2011) The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188: 479–488.

[14] 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.

[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] 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.

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

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

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

Table 1: Effect of ML parameterizations on inference.

| Codon frequencies | $\kappa$ parameterization | $\omega$ correlation | $\omega$ error | $\kappa$ correlation | $\kappa$ error |
|---|---|---|---|---|---|
| Equal | True | 0.999 | 0.012 | | |
| Equal | 1 | 0.951 | 0.22 | | |
| Equal | Free | 0.993 | 0.045 | 0.869 | 0.148 |
| F3x4 | True | 0.077 | 5.284 | | |
| F3x4 | 1 | 0.032 | 3.553 | | |
| F3x4 | Free | 0.147 | 2.152 | 0.772 | 0.215 |
| CF3x4 | True | 0.076 | 5.265 | | |
| CF3x4 | 1 | 0.022 | 3.519 | | |
| CF3x4 | Free | 0.149 | 2.141 | 0.764 | 0.22 |
| Empirical | True | -0.62 | 51.851 | | |
| Empirical | 1 | -0.61 | 28.06 | | |
| Empirical | Free | -0.608 | 59.792 | 0.76 | 0.244 |

Codon frequency specifications were either set as equal (1/61 per codon), calculated from the F3x4 estimator [5], calculated from the CF3x4 estimator [17], or set equal to the simulated alignment's empirical frequencies. $\kappa$ 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 $\omega$ estimate and our derived $\omega$ values. Error refers to the mean absolute error between these two $\omega$ estimates. Similar values for $\kappa$ are shown for those inferences where $\kappa$ was a free parameter of the model.

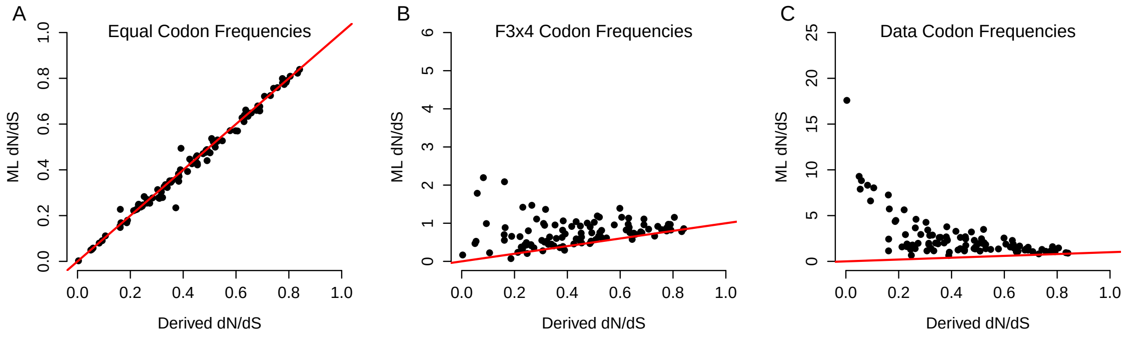

Figure 1: Issues with frequency specifications abound. Relationship between omega values only really exists when equal codon frequencies are specified. When f3x4 or true freqs used, there is the potential to end up with dramatically inflated values. As we have shown that omegas from steady-state evolution can only

strongly related to the codon frequencies in the data set. Issue is more egregious when there are relatively few codons, based on entropy. As entropy increases (more permissive, and thus data set codon frequencies are flatter), the error decreases and ML more approximates the true omega value.

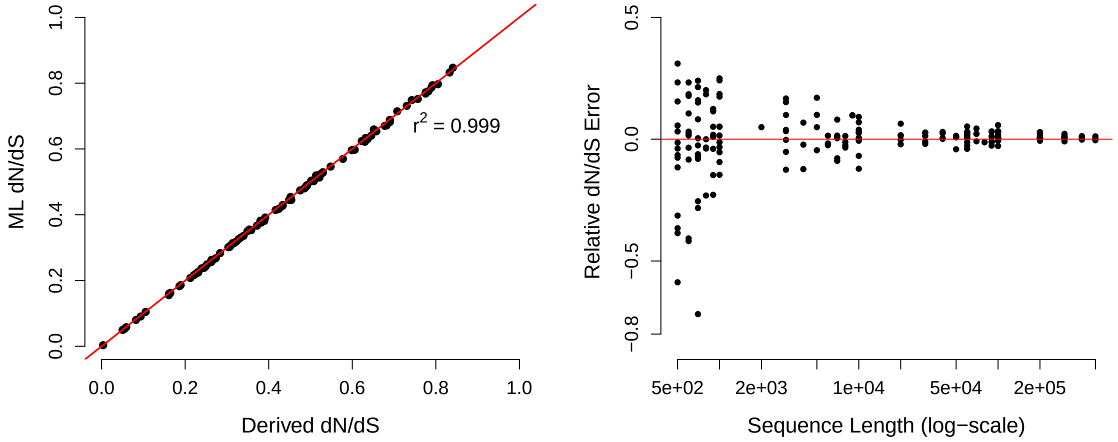

Figure 2: Relationship works exceedingly well. Left panel shows 100 points, each of which corresponds to single simulation. Note that here the ml inference is shown for equal codon frequency specs and kappa fixed to true value (a similar plot for free kappa is shown in suppfigs, but results are qualitatively 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.