# Population Genetics Based Phylogenetics Under Stabilizing Selection for an Optimal Amino Acid Sequence: A Nested Modeling Approach

JEREMY M. BEAULIEU[1,2,3], BRIAN C. O'MEARA[2,3], RUSSELL ZARETZKI[4], CEDRIC LANDERER[2,3], JUANJUAN CHAI[2,5], AND MICHAEL A. GILCHRIST[2,3,*]

[1]Department of Biological Sciences, University of Arkansas, Fayetteville, AR 72701

[2]Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610

[3]National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996-3410

[4]Department of Business Analytics & Statistics, Knoxville, TN 37996-0532

[5]Current address: 50 Main St, Suite 1039, White Plains, NY 10606

[*]Corresponding author. E-mail: mikeg@utk.edu

Version dated: Friday 11th May, 2018

# Abstract

We present a novel phylogenetic approach based of substitution rates generated from a nested population genetics model. Unlike many simpler codon models, which assume a single substitution matrix for all sites, our model more realistically represents the evolution of protein-coding DNA under the assumption of consistent, stabilizing selection for a gene specific, optimal amino acid sequence. By modeling the cost-benefit function of an amino acid sequence, our model naturally links the strength of stabilizing selection to protein synthesis levels, which, in turn, can be estimated. Our new set of models, which we collectively call SelAC (Selection on Amino acids and Codons), fit phylogenetic data much better than popular models, suggesting strong potential for more accurate inference of phylogenetic trees and branch lengths from existing data. SelAC also demonstrates that a large amount of biologically meaningful information is accessible when using a nested set of mechanistic models. For example, SelAC prediction of gene specific protein synthesis rates correlates well with both empirical ($r = 0.34 - 0.48$) and other theoretical predictions ($r = 0.59 - 0.64$) for multiple species. SelAC also provides estimates of which amino acid is optimal for a given site. Finally, because SelAC is a nested approach based on clearly stated biological assumptions, it can be simplified or expanded as needed, such as including shifts in the optimal amino acid sequence within or across lineages.

Phylogenetic analyses plays a critical role in most aspects of biology, particularly in the fields of ecology, evolution, paleontology, medicine, and conservation. While the scale and impact of phylogenetic studies has increased substantially over the past two decades, the realism of the mathematical models on which these analyses are based has changed relatively little by comparison. The most popular models of DNA substitution used by molecular phylogenetics are simple nucleotide models that are indifferent to the type of sequences to which they are applied. For example, when evaluating protein-coding sequences these models are inherently agnostic with regards to the different amino acid substitutions and their impact on gene function (e.g. F81, F84, HYK85, TN93, and GTR, see Yang (2014) for an overview) and, as a result, cannot describe the behavior of natural selection at the amino acid or protein level.

To address this critical shortcoming, Goldman and Yang (1994) and Muse and Gaut (1994) independently introduced models which assumed that differences in the physicochemical properties between amino acids, or physicochemical distances for short, could affect substitution rates. These physicochemical approaches as originally described have rarely been used for empirical data; instead these models have served as the basis for an array of simpler and, in turn, more popular models that, starting with Yang and Nielsen (1998); Nielsen and Yang (1998), typically assume an equal fixation probability for *all* non-synonymous mutations. Thus, these simpler models initially employed a single term $\omega$ to model the differences in fixation probability between nonsynonomous and synonomyous changes at all sites. To improve their realism, more complex forms have been developed that allow $\omega$ to vary between sites or branches (as cited in  Anisimova 2012) and include selection on different synonyms for the same amino acid (e.g. Yang and Nielsen 2008)

In Goldman and Yang (1994); Yang and Nielsen (1998); Nielsen and Yang (1998) and later studies based on their work, $\omega$ is suggested to indicate whether a given site within a protein sequence is under consistent 'stabilizing ($\omega < 1$) or 'diversifying' ($\omega > 1$) selection.

However, the model's actual behavior is inconsistent with how these terms are typically defined and understood (e.g. see Pellmyr 2002). Because synonymous substitutions have a higher substitution rate than any possible non-synonymous substitutions when $\omega < 1$, the model behaves as if the resident amino acid $i$ at a given site is favored by natural selection. Even when $\omega$ is allowed to vary between sites, the symmetrical nature of the model means that for any given site the strength of selection for the resident amino acid $i$ over its 19 alternatives is equally strong regardless of their physicochemical properties. Paradoxically, selection for amino acid $i$ persists *until* a substitution for another amino acid, $j$, occurs. As soon as amino acid $j$ fixes, but not before, selection now favors amino acid $j$ equally over all other amino acids, including amino acid $i$. This is now the opposite scenario from when $i$ was the resident. Similarly, when $\omega > 1$, synonymous substitutions have a lower substitution rate than any possible non-synonymous substitutions from the resident amino acid. Again due to the model's symmetry, the selection *against* the resident amino acid $i$ is equally strong relative to alternative amino acids. The selection against the resident amino acid $i$ persists until a substitution occurs at which point selection now *favors* amino acid $i$, as well as the 19 other amino acids, to the same degree $i$ was previously disfavored.

Thus, the simplest and most consistent interpretation of $\omega$ is that it represents the rate at which the *selective environment itself* changes, and this change in selection perfectly coincides with the fixation of a new amino acid. This, in turn, implies that the rate of shifts in the optimal (or pessimal) amino acid is on the time scale as the rate of substitution. Contrary to popular belief, $\omega$ does not describe whether a site is evolving under a constant regime of stabilizing or diversifying selection, but instead how a very particular selective environment changes over time. Given this behavior, $\omega$ based models are likely to only reasonably approximate a subset of scenarios such as perfectly symmetrical over-/under-dominance or positive/negative frequency dependent selection (Hughes and Nei 1988; Nowak 2006).

Here we propose an approach where selection is based explicitly on seeking to minimize the cost-benefit function $\eta$ of a protein where protein function is determined solely by the physicochemical properties of the primary amino acid sequence. Our approach, which we call SelAC (Selection on Amino acids and Codons), is developed in the same vein as previous phylogenetic applications of the Wright-Fisher process (e.g. Muse and Gaut 1994; Halpern and Bruno 1998; Yang and Nielsen 2008; Rodrigue et al. 2005; Koshi and Goldstein 1997; Koshi et al. 1999; Dimmic et al. 2000; Thorne et al. 2012; Lartillot and Philippe 2004; Rodrigue and Lartillot 2014). Similar to Lartillot's work (Lartillot and Philippe 2004; Rodrigue and Lartillot 2014), we assume there is a finite set of rate matrices describing the substitution process and that each position within a protein is assigned to a particular rate matrix category. Unlike Lartillot's work, we assume *a priori* there are 20 different families of rate matrices, one family for when a given amino acid is favored at a site. The key parameters underlying these matrices are shared across genes except for gene expression. As a result, SelAC allows us to quantitatively evaluate the support for a particular amino acid being favored at a particular position within the protein sequence.

JMB: At this point, this seems like a natural break. After this it sort of gives away the punchline of the model and it reads like a discussion. I suggest cutting this next bit or moving it elsewhere. As of now, the intro is way too long.
BCO: I agree. We're talking about the benefits of the model, but people have no idea what it is yet. I'm moving it to discussion.

# Materials & Methods

## *Overview*

We model the substitution process as a classic Wright-Fisher process which includes the forces of mutation, selection, and drift (Fisher 1930; Kimura 1962; Wright 1969; Iwasa 1988; Berg and Lässig 2003; Sella and Hirsh 2005; McCandlish and Stoltzfus 2014). For simplicity, we ignore linkage effects and, as a result of this and other assumptions, sequences evolve in a site independent manner.

Because SelAC requires twenty families of $61 \times 61$ matrices, the number of parameters needed to implement SelAC would, without further assumptions, be extremely large (i.e. on the order of 74,420 parameters). To reduce the number of parameters needed, while still maintaining a high degree of biological realism, we construct our gene and amino acid specific substitution matrices using a submodel nested within our substitution model, similar to approaches in Gilchrist (2007); Shah and Gilchrist (2011); Gilchrist et al. (2015).

One advantage of a nested modeling framework is that it requires only a handful of genome-wide parameters such as nucleotide specific mutation rates (scaled by effective population size $N_e$), side chain physicochemical weighting parameters, and a shape parameter describing the distribution of site sensitivities. In addition to these genome-wide parameters, SelAC requires a gene $g$ specific expression parameter $\psi_g$ which describes the average rate at which the protein's functionality is produced by the organism or a gene's 'average functionality production rate' for short (for notational simplicity, we will ignore the gene specific indicator $_g$, unless explicitly needed). Currently, $\psi$ is fixed across the phylogeny, though relaxing this assumption is a goal of future work. The gene specific parameter $\psi$ is multiplied by additional model terms to make a composite term $\psi'$ which scales the strength and efficacy of selection for the optimal amino acid sequence relative to drift (see Implementation below). In terms of the functionality of the protein encoded, we assume that for any given gene there exists an optimal amino acid sequence $\vec{a}^*$ and that, by definition, a complete, error free peptide consisting of $\vec{a}^*$ provides one unit of the gene's functionality. We also assume that natural selection favors genotypes that are able to synthesize their proteome more efficiently than their competitors and that each savings of an high energy phosphate bond per unit time leads to a constant proportional gain in fitness $A_0$. SelAC also requires the specification (as part of parameter optimization) of an optimal amino acid $a^*$ at each position within a coding sequence. This requirement of one $a^*$ per site makes our $\vec{a}^*$ the largest category of parameters SelAC estimates. Despite the need to specify $a^*$ for each site, because we use a submodel to derive our substitution matrices, SelAC estimates a relatively small number of the parameters when compared to more general approaches where the fitness of each amino acid is allowed to vary freely of any physicochemical properties (Halpern and Bruno 1998; Lartillot and Philippe 2004; Rodrigue and Lartillot 2014).

As with other phylogenetic methods, we generate estimates of branch lengths and nucleotide specific mutation rates. In addition, the method can also be used to make quantitative inferences on the optimal amino acid sequence of a given protein as well as the realized average synthesis rate of each protein used in the analysis. The mechanistic basis of SelAC also means it can be easily extended to include more biological realism and test more explicit hypotheses about sequence evolution.

## *Mutation Rate Matrix* $\boldsymbol{\mu}$

We begin with a 4x4 nucleotide mutation matrix $\boldsymbol{\mu}$ that describes mutation rates between different bases and, in turn, different codons. For our purposes, we rely on the general unrestricted model (UNREST from  Yang 1994) because it imposes no constraints on the instantaneous rate of change between any pair of nucleotides. More constrained models, such as the Jukes-Cantor (JC), Hasegawa-Kishino-Yano (HKY), or the general time-reversible model (GTR), can also be used.

The 12 parameter UNREST model defines the relative rates of change between a pair of nucleotides. Thus, we arbitrarily set the G→T mutation rate to 1, resulting in 11 free mutation rate parameters in the 4x4 mutation nucleotide mutation matrix. The nucleotide mutation matrix is also scaled by a diagonal matrix $\boldsymbol{\pi}$ whose entries, $\pi_{i,i}$, correspond to the equilibrium frequencies of each base. These equilibrium nucleotide frequencies are determined by analytically solving $\boldsymbol{\pi} \times \mathbf{Q} = 0$. We use this $\mathbf{Q}$ to populate a $61 \times 61$ codon mutation matrix $\boldsymbol{\mu}$, whose entries $\mu_{i,j}$ $i \neq j$ describes the mutation rate from codon $i$ to $j$ and $\mu_{i,i} = -\sum_j \mu_{i,j}$. We generate this matrix using a "weak mutation" assumption, such that evolution is mutation limited, codon substitutions only occur one nucleotide at a time. As a result, the rate of change between any pair of codons that differ by more than one nucleotide is zero.

MIKE: Is this definition of $\mu_{i,i}$ correct? JMB: Correct.

While the overall model does not assume equilibrium, we still need to scale our mutation matrices $\mu$ by a scaling factor $S$. As traditionally done, we rescale our time units such that at equilibrium, one unit of branch length represents one expected mutation per site (which equals the substitution rate under neutrality, but would not with selection). More explicitly, $S = -\left(\sum_{i \in \text{codons}} \mu_{i,i}\pi_{i,i}\right)$ where the final mutation rate matrix is the original mutation rate matrix multiplied by $1/S$.

MIKE: I updated the definition of $S$. Please ensure it is correct. JMB: Technically, the negative is shown as outside the equation (in every paper I've seen it in), but I don't think it matters. However I changed to be consistent.

## *Protein Synthesis Cost-Benefit Function $\eta$*

SelAC links fitness to the product of the cost-benefit function of a gene $\eta$ and the organism's average target synthesis rate of the functionality provided by gene $\psi$. This is because the average flux energy an organism spends to meet its target functionality provided by the gene is, by definition, $\eta \times \psi$. Compensatory changes that allow an organism to maintain functionality even with loss of one or both copies of a gene are widespread (reviewed in [1] ); here we assume that for finer scale problems than entire loss (for example, a 10% loss of functionality) the compensation is more production of the protein. In order to link genotype to our cost-benefit function $\eta = \mathbf{C}/\mathbf{B}$, we begin by defining our benefit function $\mathbf{B}$.

MIKE: Brian, please provide references or cut.

[1] From Cruft: There is evidence of compensation for protein function. Metabolism with gene expression models (ME-models) link those factors to successfully make predictions about response to perturbations in a cell `https://www.nature.com/articles/ncomms1928`, `https://www.sciencedirect.com/science/article/pii/S0958166914002316`. For example, an ME-model for *E. coli* successfully predicted gene expression levels in vivo `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0045635`.

**Benefit:** Our benefit function $\mathbf{B}$ measures the functionality of the amino acid sequence $\vec{a}_i$ encoded by a set of codons $\vec{c}_i$, i.e. $a(\vec{c}_i) = \vec{a}_i$ relative to that of an optimal sequence $\vec{a}^*$. By definition, $\mathbf{B}(\vec{a}^*|\vec{a}^*) = 1$ and $\mathbf{B}(\vec{a}_i|\vec{a}^*) < 1$ for all other sequences. We assume all amino acids within the sequence contribute to protein function and that this contribution declines as an inverse function of physicochemical distance between each amino acid and the optimal one. Formally, we assume that

$$\mathbf{B}(\vec{a}|\vec{a}^*) = \left( \frac{1}{n} \sum_{p=1}^{n} \left(1 + G_p d(a_p, a_p^*)\right) \right)^{-1} \tag{1}$$

where $n$ is the length of the protein, $d(a_p, a_p^*)$ is a weighted physicochemical distance between the amino acid encoded at a given position $p$ and $a_p^*$ is the optimal amino acid for that position. For simplicity, we assume all nonsense mutations are lethal by defining the the physicochemical distance between a stop codon and a sense codon as $\infty$. The term $G_p$ describes the sensitivity of the protein's function to physicochemical deviation from the optimimum at site position $p$. There are many possible measures for physiochemical distance; we use Grantham (1974) distances by default, though others may be chosen. We assume that $G_p \sim \text{Gamma}\left(\text{shape} = \alpha_G, \text{rate} = \alpha_G\right)$ in order to ensure $\mathbb{E}(G_p) = 1$. Given the definition of the Gamma distribution, the variance in $G_p$ is equal to shape/rate$^2 = 1/\alpha_G$. Further, at the limit of $\alpha_G \to \infty$, the model becomes equivalent to assuming uniform site sensitivity where $G_p = 1$ for all positions $p$. Finally, we note that $\mathbf{B}(\vec{a}_i|\vec{a}^*)$ is inversely proportional to the average physicochemical deviation of an amino acid sequence $\vec{a}_i$ from the optimal sequence $\vec{a}^*$ weighted by each site's sensitivity to this deviation. $\mathbf{B}(\vec{a}_i|\vec{a}^*)$ can be generalized to include second and higher order terms of the distance measure $d$.

**Cost:** Protein synthesis involves both direct and indirect assembly costs. Direct costs consist of the high energy phosphate bonds $\sim P$ of ATP or GTP's used to assemble the ribosome on the mRNA, charge tRNA's for elongation, move the ribosome forward along the transcript, and terminate protein synthesis. As a result, direct protein assembly costs are the same for all proteins of the same length. Indirect costs of protein assembly are potentially numerous and could include the cost of amino acid synthesis as well the cost and efficiency with which the protein assembly infrastructure such as ribosomes, aminoacyl-tRNA synthetases, tRNAs, and mRNAs are used. When these indirect costs are combined with sequence specific benefits, the probability of a mutant allele fixing is no longer independent of the rest of the sequence (Gilchrist et al. 2015) and, as a result, model fitting becomes substantially more complex. Thus for simplicity, in this study we ignore indirect costs of protein assembly that vary between genotypes and define,

$$\begin{aligned}\mathbf{C}(\vec{c}_i) &= \text{Energetic cost of protein synthesis.} \qquad (2)\\ &= A_1 + A_2 n \qquad (3)\end{aligned}$$

where, $A_1$ and $A_2$ represent the direct cost, in high energy phosphate bonds, of ribosome initiation and peptide elongation, respectively, where $A_1 = A_2 = 4 \sim P$.

## *Defining Physicochemical Distances*

Assuming that functionality declines with an amino acid $a_i$'s physicochemical distance from the optimum amino acid $a^*$ at each site provides a biologically defensible way of mapping genotype to protein function that requires relatively few free parameters. In addition, SelAC naturally lends itself to model selection since we can compare the quality of SelAC fits using different mixtures of physicochemical properties. Following Grantham (1974), we focus on using composition $c$, polarity $p$, and molecular volume $v$ of each amino acid's side chain residue to define our distance function, but the model and its implementation can flexibly handle a variety of properties. We use the Euclidian distance between residue properties where each property $c$, $p$, and $v$ has its own weighting term, $\alpha_c$, $\alpha_p$, $\alpha_v$, respectively, which we refer to as 'Grantham weights'. Because physicochemical distance is ultimately weighted by a gene's specific average protein synthesis rate $\psi$, another parameter we estimate, there is a problem with parameter identifiablity. The scale of gene expression is affected by how we measure physicochemical distances which, in turn, is determined by our choice of Grantham weights. As a result, by default we set $\alpha_v = 3.990 \times 10^{-4}$, the value originally estimated by Grantham, and recognize that our estimates of $\alpha_c$ and $\alpha_p$ and $\psi$ are scaled relative to this choice for $\alpha_v$. More specifically,

MIKEG: Jeremy, is this still true? JMB: Yes it is. I thought the final state was assuming they summed to 1.

$$d(a_i, a^*) = \left(\alpha_c \left[c\left(a_i\right) - c\left(a^*\right)\right]^2 + \alpha_p \left[p\left(a_i\right) - p\left(a^*\right)\right]^2 + \alpha_v \left[v\left(a_i\right) - v\left(a^*\right)\right]^2\right)^{1/2}.$$

## *Linking Protein Synthesis to Allele Substitution*

Next we link the protein synthesis cost-benefit function $\eta$ of an allele with its fixation probability. First, we assume that each protein encoded within a genome provides some beneficial function and that the organism needs that functionality to be produced at a target average rate $\psi$. Again, by definition, the optimal amino acid sequence for a given gene, $\vec{a}^*$, produces one unit of functionality, i.e. $\mathbf{B}(\vec{a}^*) = 1$. Second, we assume that the actual average rate a protein is synthesized $\phi$ is regulated by the organism to ensure that functionality is produced at rate $\psi$. As a result, it follows that $\phi = \psi/\mathbf{B}(\vec{a}|\vec{a}^*)$ and the cost of a suboptimal amino acid increases the more it decreases the protein's functionality, $\mathbf{B}$. In other words, the average production rate of a protein $\vec{a}$ with relative functionality $\mathbf{B}(\vec{a}) < 1$ must be $1/\mathbf{B}(\vec{a}|\vec{a}^*)$ times higher than the production rate needed if the optimal amino acid sequence $\vec{a}^*$ was encoded since $\mathbf{B}(\vec{a}^*|\vec{a}^*) = 1$. For example, a cell with an allele $\vec{a}$ where $\mathbf{B}(\vec{a}|\vec{a}^*) = 9/10$ would have to produce the protein at rate $\phi = 10/9 \times \psi = 1.11\psi$. Similarly, a cell with an allele $\vec{a}$ where $\mathbf{B}(\vec{a}|\vec{a}^*) = 1/2$ will have to produce the protein at $\phi = 2\psi$. In contrast, a cell with the optimal allele $\vec{a}^*$ would have to produce the protein at rate $\phi = \psi$.

Third, we assume that every additional high energy phosphate bond, $\sim P$, spent per unit time to meet the organism's target function synthesis rate $\psi$ leads to a slight and proportional decrease in fitness $W$. This assumption, in turn, implies

$$W_i\left(\vec{c}\right) \propto \exp\left[-A_0\, \eta(\vec{c}_i)\psi\right]. \tag{4}$$

where $A_0$, again, describes the proportional decline in fitness with every $\sim P$ wasted per unit time. Because $A_0$ shares the same time units as $\psi$ and $\phi$ and only occurs in SelAC in conjunction with $\psi$, we do not need to explicitly identify our time units. Instead, we recognize that our estimates of $\psi$ share an unknown scaling term.

Correspondingly, the ratio of fitness between two genotypes is,

$$\begin{aligned} W_i/W_j &= \exp\left[-A_0\, \eta(\vec{c}_i)\psi\right] / \exp\left[-A_0\, \eta(\vec{c}_j)\psi\right] && (5)\\ &= \exp\left[-A_0\left(\eta(\vec{c}_i) - \eta(\vec{c}_j)\right)\psi\right] && (6)\\ & && (7) \end{aligned}$$

Given our formulations of $\mathbf{C}$ and $\mathbf{B}$, the fitness effects between sites are multiplicative and, therefore, the substitution of an amino acid at one site can be modeled independently of the amino acids at the other sites within the coding sequence. As a result, the fitness ratio for two genotypes differing at a multiple site simplifies to

$$W_i/W_j = \exp\left[-\left(\frac{A_0\left(A_1 + A_2 n_g\right)}{n_g}\right)\sum_{p\in\mathbb{P}}\left[d\left(a_{i,p}, a_p^*\right) - d\left(a_{j,p}, a_p^*\right)\right] G_p \psi\right]$$

where $\mathbb{P}$ represents the codon positions in which $\vec{c}_i$ and $\vec{c}_j$ differ. Fourth, we make a weak mutation assumption, such that alleles can differ at only one position at any given time, i.e. $|\mathbb{P}| = 1$, and that the population is evolving according to a Wright-Fisher process. As a result, the probability a new mutant, $j$, introduced via mutation into a resident population $i$ with effective size $N_e$ will go to fixation is,

$$\begin{aligned} u_{i,j} &= \frac{1 - \left(W_i/W_j\right)^b}{1 - \left(W_i/W_j\right)^{2N_e}} \\ &= \frac{1 - \exp\left\{-\frac{A_0}{n_g}\left(A_1 + A_2 n_g\right)\left[d\left(a_i, a^*\right) - d\left(a_j, a^*\right)\right] G_p \psi\, b\right\}}{1 - \exp\left\{-\frac{A_0}{n_g}\left(A_1 + A_2 n_g\right)\left[d\left(a_i, a^*\right) - d\left(a_j, a^*\right)\right] G_p \psi\, 2N_e\right\}} \end{aligned}$$

where $b = 1$ for a diploid population and 2 for a haploid population (Kimura 1962; Wright 1969; Iwasa 1988; Berg and Lässig 2003; Sella and Hirsh 2005). Finally, assuming a constant mutation rate between alleles $i$ and $j$, $\mu_{i,j}$, the substitution rate from allele $i$ to $j$ can be modeled as,

$$q_{i,j} = \frac{2}{b}\mu_{i,j} N_e u_{i,j}.$$

where, given the substitution model's weak mutation assumption, $N_e\mu \ll 1$. In the end, each optimal amino acid has a separate 64 x 64 substitution rate matrix $\mathbf{Q}_a$, which incorporates selection for the amino acid (and the fixation rate matrix this creates) as well as the common mutation parameters across optimal amino acids. This results in the creation of 20 **Q** matrices, one for each amino acid and each with $3,721$ entries which are based on a relatively small number of model parameters (one to 11 mutation rates, two free Grantham weights, the cost of protein assembly, $A_1$ and $A_2$, the gene specific target functionality synthesis rate $\psi$, and optimal amino acid at each position $p$, $a_p^*$). These model parameters can either be specified *a priori* and/or estimated from the data.

Given our assumption of independent evolution among sites, it follows that the probability of the whole data set is the product of the probabilities of observing the data at each individual site. Thus, the likelihood $\mathcal{L}$ of amino acid $a$ being optimal at a given site position $p$ is calculated as

$$\mathcal{L}\left(\mathbf{Q}_a|\mathbf{D}_p, \mathbf{T}\right) \propto \mathbf{P}\left(\mathbf{D}_p|\mathbf{Q}_a, \mathbf{T}\right) \tag{8}$$

In this case, the data, $\mathbf{D}_p$, are the observed codon states at position $p$ for the tips of the phylogenetic tree with topology $\mathbf{T}$. For our purposes we take $\mathbf{T}$ as given but it could be estimated as well. The pruning algorithm of Felsenstein (1981) is used to calculate $\mathcal{L}\left(\mathbf{Q}_a|\mathbf{D}_p, \mathbf{T}\right)$. The log of the likelihood is maximized by estimating the genome scale parameters which consist of 11 mutation parameters which are implicitly scaled by $2N_e/b$, and two Grantham distance parameters, $\alpha_c$ and $\alpha_p$, and the sensitivity distribution parameter $\alpha_G$. Because $A_0$ and $\psi_g$ always co-occur and are scaled by $N_e$, for each gene $g$ we estimate a composite term $\psi_g' = \psi_g A_0 b N_e$ and the optimal amino acid for each position $a_p^*$ of the protein. When estimating $\alpha_G$, the likelihood then becomes the average likelihood which we calculate using the generalized Laguerre quadrature with $k = 4$ points (Felsenstein 2001).

Finally, we note that because we infer the ancestral state of the system, our approach does not rely on any assumptions of model stationarity. Nevertheless, as our branch lengths grow the probability of observing a particular amino acid $a$ at a given site approaches a stationary value proportional to $W(a)^{2N_e-b}$ and any effects of mutation bias (Sella and Hirsh 2005).

## *Implementation*

All methods described above are implemented in the new R package, `selac` available through GitHub (`https://github.com/bomeara/selac`) [it will be uploaded to CRAN once peer review has completed]. Our package requires as input a set of fasta files that each contain an alignment of coding sequence for a set of taxa, and the phylogeny depicting the hypothesized relationships among them. In addition to the SelAC models, we implemented the GY94 codon model of Goldman and Yang (1994), the FMutSel mutation-selection model of Yang and Nielsen (2008), and the standard general time-reversible nucleotide model that allows for $\Gamma$ distributed rates across sites. These likelihood-based models represent a sample of the types of popular models often fit to codon data.

For the SelAC models, the starting guess for the optimal amino acid at a site comes from 'majority' rule, where the initial optimum is the most frequently observed amino acid at a given site (ties resolved randomly). Our optimization routine utilizes a four stage hill climbing approach. More specifically, within each stage a block of parameters are optimized while the remaining parameters are held constant. The first stage optimizes the block of branch length parameters. The second stage optimizes the block of gene specific composite parameters $\psi_g' = A_0\psi_g N_e b$. The third stage optimizes SelAC's parameters shared across the genome $\alpha_c$ and $\alpha_p$, and the sensitivity distribution parameter $\alpha_G$. The fourth stage estimates the optimal amino acid at each site $a^*$. This entire four stage cycle is repeated six more times, using the estimates from the previous cycle as the initial conditions for the new one. However, if there is no improvement in the log-likelihood between any successive optimization cycles, we consider the ML solution found and the search is terminated. For optimization of a given set of parameters, we rely on a bounded

MIKE: I thought we updated this to be based on the change in LLik function. JMB: Is this clear now?

subplex routine (Rowan 1990) in the package `NLopt` (Johnson 2012) to maximize the log-likelihood function. To help the optimization navigate through local peaks, we perform a set of independent analyses with different sets of naive starting points with respect to the gene specific composite $\psi'$ parameters, $\alpha_c$, and $\alpha_p$. Confidence in the parameter estimates can be generated by an 'adaptive search' procedure that we implemented to provide an estimate of the parameter space that is some pre-defined likelihood distance (e.g., 2 lnL units) from the maximum likelihood estimate (MLE), which follows Beaulieu and OMeara (2016) and Edwards (1984).

We note that our current implementation of SelAC is painfully slow, and is best suited for data sets with relatively few number of taxa (i.e. $< 10$). This limitation is largely due to the size and quantity of matrices we create and manipulate to calculate the log-likelihood of an individual site. Ongoing work will address the need for speed, with the eventual goal of implementing SelAC in popular phylogenetic inference toolkits, such as RevBayes (Hhna et al. 2016), PAML (Yang 2007) and RAxML (Stamatakis 2006).

## *Simulations*

We evaluated the performance of our codon model by simulating datasets and estimating the bias of the inferred model parameters from these data. Our 'known' parameters under a given generating model were based on fitting SelAC to the 106 gene data set and phylogeny of Rokas et al. (2003). The tree used in these analyses is outdated with respect to the current hypothesis of relationships within *Saccharomyces*, but we rely on it simply as a training set that is separate from our empirical analyses (see section below). Bias in the model parameters were assessed under two generating models: one where we assumed a model of SelAC assuming uniform sensitivity across sites (i.e. $G_p = 1$ for all sites, i.e. $\alpha_G = \infty$), and one where we used the Gamma distribution joint shape and rate parameter $\alpha_G$ estimated from the empirical data. Under each of these two scenarios, we used parameter estimates from the corresponding empirical analysis and simulated 50 five-gene data sets. For the gene specific composite parameter $\psi'_g$ the 'known' values used for the simulation were five evenly spaced points along the rank order of the estimates across the 106 genes. The MLE estimate for a given replicate were taken as the fit with the highest log-likelihood after running five independent analyses with different sets of naive starting points with respect to the composite $\psi'_g$ parameter, $\alpha_c$, and $\alpha_p$. All analyses were carried out in our `selac` R package.

## *Analysis of yeast genomes & tests of model adequacy*

We focus our empirical analyses on the large yeast data set and phylogeny of Salichos and Rokas (2013). The yeast genome is an ideal system to examine our phylogenetic estimates of gene expression and its connection to real world measurements of these data within individual taxa. The complete data set of Salichos and Rokas (2013) contain 1070 orthologs, where we selected 100 at random for our analyses. We also focus our analyses on *Saccharomyces sensu stricto* and their sister taxon *Candida glabrata*, and we used the phylogeny depicted in Fig. 1 of Salichos and Rokas (2013) for our fixed tree. We fit the two SelAC models described above (i.e., SelAC and SelAC+$\Gamma$), as well as two codon models, GY94 and FMutSel, and a standard GTR + $\Gamma$ nucleotide model. The FMutSel model assumes that the amino acid frequencies are determined by functional requirements of the protein while the other models make no assumptions about amino acid frequencies. In all cases, we assumed that the model was partitioned by gene, but with branch lengths linked across genes.

MIKE: Can someone verify this last statement about FMutSel? JMB: This seems right based on what I understand about this model.

For SelAC, we compared our estimates of $\phi' = \psi'/\mathbf{B}$, which represents the average protein synthesis rate of a gene, to estimates of gene expression from empirical data. Specifically, we obtained gene expression data for five of the six species used - four species were measured during log-growth phase, whereas the other was measured at the beginning of the stationary phase (*S. kudriavzevii*) from the Gene Expression Omnibus (GEO). Gene expression in this context corresponds to mRNA abundances which were measured using either microarrays (*C. glabrata*, *S. castellii*, and *S. kudriavzevii*) or RNA-Seq (*S. paradoxus*, *S. mikatae*, and *S. cerevisiae*).

For further comparison, we also predicted the average protein synthesis rate for each gene $\phi$ by analyzing gene and genome-wide patterns of synonymous codon usage using ROC-SEMPPR (Gilchrist et al. 2015) for each individual genome. While, like SelAC, ROC-SEMPPR uses codon level information, it does not rely on any inter-specific comparisons and, unlike SelAC, uses only the intra- and inter-genic frequencies of synonymous codon usage as its data. Nevertheless, ROC-SEMPPR predictions of gene expression $\phi$ correlates strongly (Pearson $r = 0.53 - 0.74$) with a wide range of laboratory measurements of gene expression (Gilchrist et al. 2015).

While one of our main objectives was to determine the improvement of fit that SelAC has with respect to other standard phylogenetic models, we also evaluated the adequacy of SelAC. Model fit, measured with assessments such as the Akaike Information Criterion (AIC), can tell which model is least bad as an approximation for the data, but it does not reveal whether a model is actually doing a good job of representing the data. An adequate model does the latter, one measure of which is that data generated under the model resemble real data (Goldman 1993). For example, Beaulieu et al. (2013) assessed whether parsimony scores and the size of monomorphic clades of empirical data were within the distributions of simulated data under a new model and the best standard model; if the empirical summaries were outside the range for each, it would have suggested that neither model was adequately modeling this part of the biology.

In order to test adequacy for a given gene we first remove a particular taxon from the data set and the phylogeny. A marginal reconstruction of the likeliest sequence across all remaining nodes is conducted under the model, including the node where the pruned taxon attached to the tree. The marginal probabilities of each site are used to sample and assemble the starting coding sequence. This sequence is then evolved along the branch, periodically being sampled and its current functionality assessed. We repeat this process 100 times and compare the distribution of trajectories against the observed functionality calculated for the gene. For comparison, we also conducted the same test, by simulating the sequence under the standard GTR + $\Gamma$ nucleotide model, which is often used on these data but does not account for the fact that the sequences are protein coding, and under FMutSel, which includes selection on codons but in a fundamentally different way as our model.

### *The appropriate estimator of bias for AIC*

As part of the model set described above, we also included a reduced form of each of the two SelAC models, SelAC and SelAC+$\Gamma$. Specifically, rather than optimizing the amino acid at any given site, we assume the the most frequently observed amino acid at each site is the optimal amino acid $a^*$. We refer to these 'majority rule' models as SelAC$_M$ and SelAC$_M + \Gamma$ and the majority rule parameterization accelerates model fitting.

Since these majority rule models assume that the optimal amino acids are known prior to fitting of our model, it is tempting to reduce the count of estimated parameters in the model by the number of parameters estimated using majority rule. Despite having become standard behavior in the field of phylogenetics, this reduction is statistically inappropriate unless one uses an additional dataset for this purpose, something we have not seen. Thus, although using majority rule doesn't necessarily give you the most likely parameter estimate, it still uses the data to generate the estimate and, thus, represents a parameter estimated from the data. Because the difference in the number of parameters $K$ when counting or not counting the number of nucleotide sites drops out when comparing nucleotide models with AIC, this statistical issue does not apply to nucleotide models. It does, however, matter for AICc, where $K$ and the sample size $n$ combine in the penalty term. This also matters in our case, where the number of estimated parameters for the majority rule estimation differs based on whether one is looking at codons or single nucleotides.

In phylogenetics two variants of AICc are used. In comparative methods (e.g. Butler and King 2004; O'Meara et al. 2006; Beaulieu et al. 2013) the number of data points, $n$, is taken as the number of taxa. More taxa allow the fitting of more complex models, given more data. However, in DNA evolution, which is effectively the same as a discrete character model used in comparative methods, the $n$ is taken as the number of sites. Obviously, both cannot be correct.

The original derivation of AICc by Hurvich and Tsai (1989) assumed a regression model, where the true model was in the set of examined models, as well as approximations in the derivation itself. The appropriateness of this approximation for phylogenetic data, where shared evolutionary history means data points between taxa lack independence is unclear. In any case, we argue that for phylogenetic data, a good estimate of data set size is number of taxa multiplied by number of sites. First, this is what is conventionally seen as the size of the dataset in the field. Second, when considering how likelihood is calculated, the likelihood for a given site is the sum of the probabilities of each observed state at each tip, which is then multiplied across sites. It is arguable that the conventional approach in comparative methods is calculating AICc in the same way. That is, if only one column of data (or ”site”) is examined, as remains remarkably common in comparative methods, when we refer to sample size, it is technically the number of taxa multiplied by number of sites, even though it is referred to simply as the number of taxa. One notable exception to this appoach to calculating AICc is the program SURFACE implemented by Ingram and Mahler (2013), which uses multiple characters and taxa.

MIKE: This sentence doesn't make any sense to me. JMB: Me neither. But I know what it is trying to say. Modified for clarity.

Recently, Jhwueng et al. (2014) performed an analysis that investigated what variant of AIC and AICc worked best as an estimator, but the results were inconclusive. Here, we have adopted and extended the simulation approach of Jhwueng et al. (2014) in order to examine a large set of different penalty functions and how well they approximate the remaining portion of the Kullback-Liebler (KL) divergence between two models after accounting for the deviance (i.e., $-2\mathcal{L}$) (see Appendix 1 for more details).

# Results

By linking transition rates $q_{i,j}$ to gene expression $\psi$, our approach allows use of the same model for genes under varying degrees of stabilizing selection. Specifically, we assume the strength of stabilizing selection for the optimal sequence, $\vec{a}^*$, is proportional to the average protein synthesis rate $\phi$, which we can estimate for each gene. In regards to model fit, our results clearly indicated that linking the strength of stabilizing selection for the optimal sequence to gene expression substantially improves our model fit. Further, including the shape parameter $\alpha_G$ for the random effects term $G \sim \text{Gamma}(\text{shape} = \alpha_G, \text{rate} = \alpha_G)$ to allow for heterogeneity in this selection between sites within a gene improves the $\Delta$AICc of SelAC+$\Gamma$ over the simpler SelAC models by over 22,000 AIC units. Using either $\Delta$AICc or AIC$_\text{w}$ as our measure of model support, the SelAC models fit extraordinarily better than GTR + $\Gamma$, GY94, or FMutSel (Table 1). This is in spite of the need for estimating the optimal amino acid at each position in each protein, which accounts for 49,881 additional model parameters. Even when compared to the next most parameter rich codon model in our model set, FMutSel, SelAC+$\Gamma$ model shows over 160,000 AIC unit improvement over FMutSel.

With respect to estimates of $\phi$ within SelAC, they were strongly correlated with both our empirical measurements (Pearson $r = 0.34 - 0.48$) and theoretical predictions (Pearson $r = 0.59 - 0.64$) of gene expression (Figure 1 and Figures S1-S2, respectively). In

JMB: I know that the details of the AIC analyses are in the appendix, but this seems like a great place to mention the main highlights here. A reader may get here and just say poo-poo the other words, using only codon sequences, our model can predict which genes have high or low expression levels. The estimate of the $\alpha_G$ parameter, which describes the site-specific variation in sensitivity of the protein's functionality, indicated a moderate level of variation in gene expression among sites. Our estimate of $\alpha_G = 1.36$, produced a distribution of sensitivity terms $G$ ranged from 0.342-7.32, but with more than 90% of the weight for a given site-likelihood being contributed by the 0.342 and 1.50 rate categories. In simulation, however, of all the parameters in the model, only $\alpha_G$ showed a consistent bias, in that the MLE were generally lower than their actual values (see Supporting Materials). Other parameters in the model, such as the Grantham weights, provide an indication as to the physicochemical distance between amino acids. Our estimates of these weights only strongly deviate from Grantham's 1974 original estimates in regards to composition weight, $\alpha_c$, which is the ratio of noncarbon elements in the end groups to the number of side chains. Our estimate of the composition weighting factor of $\alpha_c$=0.459 is 1/4th the value estimate by Grantham which suggests that the substitution process is less sensitive to this physicochemical property when shared ancestry and variation in stabilizing selection are taken into account.

It is important to note that the nonsynonymous/synonymous mutation ratio, or $\omega$, which we estimated for each gene under the FMutSel model strongly correlated with our estimates of $\phi' = \psi'/\mathbf{B}$ where $\mathbf{B}$ depends on the sequence of each taxa. In fact, $\omega$ showed similar, though slightly reduced correlations, with the same empirical estimates of gene expression described above (Figure 2) This would give the impression that the same conclusions could have been gleaned using a much simpler model, both in terms of the number of parameters and the assumptions made. However, as we discussed earlier, not only is this model greatly restricted in terms of its biological feasibility, SelAC clearly performs better in terms of its fit to the data and biological realism.

For example, when we simulated the sequence for *S. cervisieae*, starting from the ancestral sequence under both GTR + Γ and FMutSel, the functionality of the simulated sequence moves away from the observed sequence, whereas SelAC remains near the functionality of the observed sequence (Figure 3b). This is somewhat unsurprising, given that both GTR + Γ and FMutSel are agnostic to the functionality of the gene, but it does highlight the improvement in biological realism in amino acid sequence evolution that SelAC provides. We do note that the adequacy of the SelAC model does vary among individual taxa, and does not always match the observed functionality. For instance, our simulations of *S. castellii* gene function is consistently higher than estimated from the data (Figure 3c). We suspect this is an indication that assuming a single set of optimal amino acid across all taxa may be too simplistic, but we cannot rule out other potential simplifying assumptions in our model, such as a single set of Grantham weights and $\alpha_G$ values or the simple, inverse relationship between physicochemical distance $d$ and benefit $\mathbf{B}$.

Finally, we note that our simulation analysis suggested that the best measure of dataset size for estimating KL distance uses a scaled value of the product of number of sites and number of characters. The model comparison approach described above included this assumption. For more details on the simulation approach, see Appendix 1.

# Discussion

Biologically, we know protein-coding DNA sequences largely evolve through the introduction of new mutations that either become fixed or lost due to selection and/or drift. Selection on protein coding regions can take many forms, but can generally be related to the cost of producing the protein and the functional benefit (or potential harm) caused by the protein product. The gene specific cost of protein synthesis can be affected by the amino acids used, the direct and indirect costs of peptide assembly by the ribosome, the use of chaparones to aid in folding, and even the expected lifespan of the protein.

Importantly, these costs can be computed to varying degrees of realism (e.g. Wagner 2005; Lynch and Marinov 2015). We have previously presented models of protein synthesis costs that, alternatively, take into account the cost of ribosome pausing (Shah and Gilchrist 2011) or premature termination errors (Gilchrist and Wagner 2006; Gilchrist 2007; Gilchrist et al. 2009).

Protein benefit or 'function' can be affected by the amino acids at each site and their interactions. As a result, amino acid substitutions can affect the functionality at key catalytic sites or, more broadly, the probability of a particular protein fold and, in turn, the expected functionality of the protein. Linking amino acid sequence to protein function is a daunting task; thus for simplicity, we assume that for any given desired biological function to be carried out by a protein, that (a) the biological importance of this protein function is invariant across the tree, (b) single optimal amino acid sequence that carries out this function best, and (c) the functionality of alternative amino acid sequences declines with their physicochemical distance from the optimum on a site by site basis.

We readily acknowledge that sequence space may have more than one local optimum, that physicochemical distance from the optimal primary amino acid sequence is likely a poor model of protein function, and that the biological importance of a function can vary over time. Nevertheless, we believe our cost-benefit approach to be a substantial advance of the more simplistic $\omega$ models, is complementary to the work of others in the field (e.g. Thorne et al. 2012; Rodrigue and Lartillot 2014), and, in turn, lays the foundation for more realistic work in the future.

For instance, by assuming there is an optimal amino acid for each site, SelAC naturally leads to a non-symmetrical and, thus, more cogent model of protein sequence evolution. Because the strength of selection depends on an additive function of amino acid physicochemical properties, an amino acid more similar to the optimum has a higher probability of replacing a more dissimilar amino acid than the converse situation. Further, SelAC does not assume the system is always at the optimum or pessimum point of the fitness landscape, as occurs when $\omega < 1$ or $> 1$, respectively.

Importantly, the cost-benefit approach underlying SelAC allows us to link the strength of selection on a protein sequence to its gene's expression level. Despite its well recognized importance in determining the rate of protein evolution (e.g. Drummond et al. 2005, 2006), phylogenetic models have ignored the fact that expression levels vary between genes. In order to link gene expression and the strength of stabilizing selection on protein sequences, we simply assume that the strength of selection on a gene is proportional to the average protein synthesis rate of the gene.

One possible mechanism that generates a linear relationship between the strength of selection and gene expression is the assumption of compensatory gene expression. That is, the assumption that any reduction in protein function is compensated for by an increase in the protein's production rate and, in turn, abundance. For example, a mutation which reduces the functionality of the protein to 90% of the optimal protein, would require $1/0.9 = 1.11$ of these suboptimal proteins to be produced relative to the optimal protein in order to maintain the same amount of that protein's functionality in the cell. Because the energetic cost of an 11% increase in a protein's synthesis rate is proportional to its target synthesis rate, our assumptions naturally link changes in protein functionality and changes in gene expression and its associated costs. Under what circumstances cells actually respond in this manner, remains to be determined. The fact that our method allows us to explain 13-23% of the variation in gene expression measured using RNA-Seq, suggests that this assumption is a reasonable starting point. More importantly, by linking the strength of stabilizing selection for an optimal amino acid sequence to gene expression, we can effectively weight the phylogenetic information encoded in high and low expression genes which tend to evolve at different rates.

Because SelAC infers the optimal amino acid for each site, it is substantially more parameter rich than more commonly used models such as GTR+Γ, GY94, and FMutSel. Despite this increase in number of model parameters, SelAC drastically outperforms these models with AICc values on the order of 10,000s to 100,000s. We predict that SelAC's performance could be improved even further if we use a hierarchical approach where the optimal amino acid is not estimated on a per site basis, but rather as a vector of probability an amino acid is optimal at the gene level.

SelAC makes inferences about the tree, but also about population genetic parameters, and we can validate the assumptions indirectly by comparing our inferences to other empirical data, such as we do with protein synthesis data. More generally, SelAC's assumptions lead to mechanistic and, thus, testable hypothesis about the relationship between mutation, protein function, gene expression, and rates of evolution. More importantly, alternative hypotheses could be used in place of ours and, in turn, phylogenetic and other types of data could be used to evaluate the support of these alternative models. Our hope is that by moving away from the more phenomenological models we can better connect population genetics, molecular biology, and phylogenetics allowing each area inform the others more effectively.

A central goal in evolutionary biology is to quantify the nature, strength, and, ultimately, shifts in the forces of natural selection relative to genetic drift and mutation. As data set size and complexity increase, so does the amount of potential information on these forces and their dynamics. As a result, there is a need for more complex and realistic models (Goldman et al. 1996; Thorne et al. 1996; Goldman et al. 1998; Halpern and Bruno 1998; Lartillot and Philippe 2004) to accomplish this goal. Although extremely popular due to their elegance and computational efficiency, the utility of $\omega$ based models in helping us reach this goal is substantially more limited than commonly recognized. Because these $\omega$ models use a single substitution matrix, they are only applicable for situations in which the substitution process and shifts in the selective environment are intrinsic to the sequence, such as with positive or negative frequency dependent selection; these models do not describe stabilizing or diversifying selection as commonly envisioned (Endler 1986; Pelmyr 2002).

Starting with Halpern and Bruno (1998), a number of researchers have developed methods for linking site-specific selection on protein sequence and phylogenetics (e.g. Koshi et al. 1999; Dimmic et al. 2000; Koshi and Goldstein 2000; Robinson et al. 2003; Lartillot and Philippe 2004; Thorne et al. 2012; Rodrigue and Lartillot 2014) Halpern and Bruno (1998) calculated a vector of 19 expected amino acid frequencies for each amino acid site, making it the most general and most parameter rich of these methods. This generality, however, comes at the cost of being purely descriptive; there is no explicit biological mechanism proposed to explain the site specific amino acid frequencies estimated. By grouping together amino sites with similar evolutionary behaviors, Lartillot and colleagues retained the descriptive nature of Halpern and Bruno (1998) work while greatly reduced the number of model parameters needed (Lartillot and Philippe 2004; Rodrigue and Lartillot 2014). SelAC follows in this tradition of using multiple substitution matrices, but includes some key advances.

First, by nesting a model of a sequence's cost-benefit function $\mathbf{C}/\mathbf{B}$ within a broader model, SelAC allows us to formulate and test a hierarchical, mechanistic models of stabilizing selection. More precisely, our nested approach allows us to relax the assumption that physicochemical deviations from the optimal sequence $\vec{a}^*$ are equally disruptive at all sites within a protein. We found strong support for SelAC's hypothesis that the strength of stabilizing selection against physicochemical deviations from $\vec{a}^*$ varies between sites ($\Delta$AICc = 20,983). Second, because our substitution matrices are built on a formal description of a sequence's cost-benefit function $\mathbf{C}/\mathbf{B}$, we are able to efficiently parameterize 20 different matrices using a relatively small number of genome-wide parameters – e.g. our physicochemical weighting and $G$ distribution shape parameters, and one gene specific gene expression parameter $\psi$. While the $\mathbf{C/B}$ function on which SelAC currently rests is very simple, nevertheless, it leads to a dramatic increase in our ability to explain the sequence data we analyzed. Importantly, because SelAC uses a formal description of a sequence's $\mathbf{C/B}$, replacing our assumptions with more sophisticated ones in the future is relatively straightforward. Conceptually, our work lies in between that of Lartillot's and Thorne's, where the latter is utilizing even more detailed models of protein structure as a means of linking amino acid substitutions and stabilizing selection. Third, our use of nested models also allows us to make biologically meaningful and testable predictions. By linking a gene's expression level to the strength of purifying selection it experiences, we are able to provide coarse estimates of gene expression. This also suggests that $\omega$ is best explained as a proxy for gene expression, rather than the nature of selection on a sequence.

One simplifying assumption we make is that the organism can and does compensate for any reduction in protein function by simply increasing the protein's production rate. While this production compensation assumption will clearly not hold in many situations, it does allow us to connect protein function and energetic costs in a simple and biologically plausible manner. Of course, researchers could employ and test other assumptions within our framework, namely, by utilizing more detailed, gene specific knowledge about the relationship between protein function and organism fitness. For example, suppose a protein for a glucose transporter is far less efficient than usual. One possible response and the one envisioned here is that the protein is thus produced at a higher rate to compensate. This would leave the overall ability to transport glucose unchanged. An alternative is that the cell is just less able to transport glucose across membranes. In biology, it is likely that a mixture of such effects exists. However, the production compensation mechanism is likely to have the same costs across proteins, making it a useful first approximation, while the same expression but reduced functionality will have gene specific effects more difficult to model generally (e.g., how does the cost of having glucose transport slow by half compare to the cost of underproducing an anthocyanin for flower color or fewer taste receptor proteins?). Moreover, there is evidence that cells do compensate for lower protein function by increasing gene expression. Nevertheless, by assuming that fitness declines with extraneous energy flux, SelAC explicitly links the variation in the strength of stabilizing selection for the optimal protein sequence among genes, to the variation among genes in their target expression levels $\psi$.

MIKE: Cut unless we provide citations.

Furthermore, by linking expression and selection, SelAC provides a natural framework for combining information from protein coding genes with very different rates of evolution; from low expression genes providing information on shallow branches to high expression genes providing information on deep branches. This is in contrast to a more traditional approach of concatenating gene sequences together, which is equivalent to assuming the same average protein synthesis rate $\psi$ for all of the genes, or more recent approaches where different models are fitted to different genes. Our results indicate that including a gene specific $\psi$ value vastly improves SelAC fits (Table 1). Perhaps more convincingly, we find that the target expression level $\psi$ and realized average protein synthesis rate $\phi$ are reasonably well correlated with laboratory measurements and theoretical predictions of gene expression (Pearson $r = 0.34 - 0.64$; Figures 1, S1, and S2). The idea that quantitative information on gene expression is embedded within intra-genomic patterns of synonymous codon usage is well accepted; our work shows that this information can also be extracted from comparative data at the amino acid level.

Of course, given the general nature of SelAC and the complexity of biological systems, other biological forces besides selection for reducing energy flux likely contribute to intergenic variation in the magnitude of stabilizing selection. Similarly, other physicochemical properties besides composition, volume, and charge likely contribute to site specific patterns of amino acid substitution. Thus, a larger and more informative set of physicochemical weights might improve our model fit and reduce the noise in our estimates of $\phi$. Even if other physicochemical properties are considered, the idea of a consistent, genome wide physicochemical weighting of these terms seems highly unlikely. Since the importance of an amino acid's physicochemical properties likely changes with its position in a folded protein, one way to incorporate such effects is to test whether the data supports multiple sets of physicochemical weights for either subsets of genes or regions within genes, rather than a single set.

Both of these points highlight the advantage of the detailed, mechanistic modeling approach underlying SelAC. Because there is a clear link between protein expression, synthesis cost, and functionality, SelAC can be extended by increasing the realism of the mapping between these terms and the coding sequences being analyzed. For example, SelAC currently assumes the optimal amino acid for any site is fixed along all branches. This assumption can be relaxed by allowing the optimal amino acid to change during the course of evolution along a branch. From a computational standpoint, the additive nature of selection between sites is desirable because it allows us to analyze sites within a gene largely independently of each other. From a biological standpoint, this additivity between sites ignores any non-linear interactions between sites, such as epistasis, or between alleles, such as dominance. Thus, our work can be considered a first step to modeling these more complex scenarios.

For example, our current implementation ignores any selection on synonymous codon usage bias (CUB) (c.f. Yang and Nielsen 2008; Pouyet et al. 2016). Including such selection is tricky because introducing the site-specific cost effects of CUB, which is consistent with the hypothesis that codon usage affects the efficiency of protein assembly or $\mathbf{C}$, into a model where amino acids affect protein function or $\mathbf{B}$, results in a cost-benefit ratio $\mathbf{C}/\mathbf{B}$ with epistatic interactions between all sites. These epistatic effects can likely be ignored under certain conditions or reasonably approximated based on an expectation of codon specific costs (e.g. Kubatko et al. 2016). Nevertheless, it is difficult to see how one could identify such conditions without modeling the way in which codon and amino acid usage affects $\mathbf{C}/\mathbf{B}$.

This work also points out the potential importance of further investigation into model choice in phylogenetics. For likelihood models, use of AICc has become standard. However, how one determines the appropriate number of parameters estimated in a model is more complicated than generally recognized. Common sense suggests that dataset size is increased by adding taxa and/or sites. In other words, a dataset of 1000 taxa and 100 sites must have more information on substitution models than a dataset of 4 taxa and 100 sites. Our simple analyses agree that the number of observations in a dataset (number of sites × number of taxa) should be taken as the sample size for AICc, but this conclusion likely only applies when there is sufficient independence between taxa. For instance, one could imagine a phylogeny where one taxon is sister to a polytomy of 99 taxa that have zero length terminal branches. Absent measurement error or other intraspecific variation, one would have 100 species but only two unique trait values, and the only information about the process of evolution comes from what happens on the path connecting the lone taxon to the polytomy. Although this is a rather extreme example, it seems prudent for researchers to use a simulation based approach similar to the one we take here to determine the appropriate means for calculating the effective number of data points in their data.

There are still significant shortcomings in the approach outlined here. Most worrisome are biological oversimplifications in SelAC. For example, at its heart, SelAC assumes that suboptimal proteins can be compensated for, at a cost, simply by producing more of them. However, this is likely only true for proteins reasonably close to the optimal sequence. Different enough proteins will fail to function entirely: the active site will not sufficiently match its substrates, a protein will not properly pass through a membrane, and so forth. Yet, in our model, even random sequences still permit survival, just requiring more protein production. Other oversimplifications include the assumption of no selection on codon usage, no change of optimal amino acids through time, and no change of the effect of physiochemical properties on fitness through time. However, because we take a mechanistic approach, all of these assumptions can be relaxed through further extension of our model.

There are also deficiencies in our implementation. Though reasonable to use for a given topology with a modest number of species, it is currently too slow for practical use for tree search. Our work serves as a proof of concept, or of utility for targeted questions where a more realistic model may be of use (placement of particular taxa, for example). Future work will encode SelAC models into a variety of mature, popular tree-search programs. SelAC also represents a challenging optimization problem: the nested models reduce parameter complexity vastly, but there are still numerous parameters to optimize, including the discrete parameter of the optimal amino acid at each site. A different implementation, more parameter-rich, would optimize values of three (or more) physiochemical properties per site. This would have the practical advantage of continuous parameter optimization rather than discrete, and biologically would be more realistic (as it is the properties that selection ”sees”, not the identity of the amino acid itself).

In spite of these difficulties, SelAC represents an important step in uniting phylogenetic and population genetic models. While Koshi et al. (1999); Dimmic et al. (2000); Koshi and Goldstein (2000); Robinson et al. (2003); Lartillot and Philippe (2004); Thorne et al. (2012); Rodrigue and Lartillot (2014) are all models of constant, stabilizing selection, SelAC can be generalized further to include diversifying selection. Specifically, by letting SelAC's Grantham weighting term $G$, which we now assume is $\geq 0$, to take on negative values, SelAC will behave as if there is a pessimal, rather than optimal, amino acid for the given site. In this diversifying selection scenario, amino acids with physicochemical qualities more dissimilar to the pessimal amino acid are increasingly favored, potentially resulting in multiple fitness peaks.

This ability to extend our model and, in turn, sharpen our thinking about the nature of natural selection on amino acid sequences illustrates the value of moving from descriptive to more mechanistic models in general and phylogenetics in particular. How frequently diversifying selection of this nature occurs is an open, but addressable, question. Regardless of the frequency at which diversifying selection occurs, it leads to the question, "How often does the optimal/pessimal amino sequence change along any given branch?" Due to its mechanistic nature, SelAC can also be extended to include changes in the optimal/pessimal sequence over a phylogeny using a hidden markov modelling approach. Extending SelAC in these ways, will allow researchers to explicitly model shifts in selection on protein sequences and, in turn, quantify their frequency and magnitude.

In summary, SelAC allows biologically relevant population genetic parameters to be estimated from phylogenetic information, while also dramatically improving fit and accuracy of phylogenetic models. By explicitly modeling the optimal/pessimal sequence of a gene, SelAC can be extended to include shifts in the optimal/pessimal sequence over evolutionary time. Extending this model in this way will allow researchers to describe not only the dynamic shifts in natural selection, but evaluate how well a given dataset supports such a model. Moreover, it demonstrates that there remains substantially more information in the coding sequences used for phylogenetic analysis than other methods can access. Given the enormous amount of efforts expended to generate sequence datasets, it makes sense for researchers to continue developing more realistic models of sequence evolution in order to extract the biological information embedded in these datasets. The cost-benefit model we develop here is just one of many possible paths of mechanistic model development.

## Acknowledgements

This work was supported in part by NSF Awards MCB-1120370 (MAG and RZ) and

DEB-1355033 (BCO, MAG, and RZ) with additional support from The University of Tennessee Knoxville and University of Arkansas (JMB). JJC and JMB received support as Postdoctoral Fellows and CL received support as a Graduate Student Fellow at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation through NSF Award DBI-1300426, with additional support from UTK. The authors would like to thank Premal Shah, Todd Oakley, and two anonymous reviewers for their helpful criticisms and suggestions for this work.

*

## References

Anisimova, M. 2012. Parametric models of codon evolution. Pages 12–33 *in* Codon Evolution: Mechanisms and Models (G. M. Cannarozzi and A. Schneider, eds.). Oxford University Press, Oxford, UK.

Beaulieu, J. M., B. C. O'Meara, and M. J. Donoghue. 2013. Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms. Systematic Biology 62:725–737.

Beaulieu, J. M. and B. C. OMeara. 2016. Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction. Systematic Biology 65:583–601.

Berg, J. and M. Lässig. 2003. Stochastic Evolution and Transcription Factor Binding Sites. Biophysics 48:S36–S44.

Butler, M. A. and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. American Naturalist 164:683–695.

Dimmic, M. W., D. P. Mindell, and R. A. Goldstein. 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. Pacific Symposium on Biocomputing 5:18–29.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences of the United States of America 102:14338–14343.

Drummond, D. A., A. Raval, and C. O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Molecular Biology and Evolution 23:327–337.

Edwards, A. 1984. Likelihood. Cambridge science classics Cambridge University Press.

Endler, J. A. 1986. Natural Selection in the Wild Pages 16–17. No. 21 in Monographs in Population Biology Princeton University Press, Princeton, NJ reference for definition of diversifying selection.

Felsenstein, J. 1981. Evolutionary trees from DNA-sequences - a maximum-likelihood approach. Journal of Molecular Evolution 17:368–376.

Felsenstein, J. 2001. Taking Variation of Evolutionary Rates Between Sites into Account in Inferring Phylogenies. Journal of Molecular Evolution 53:447–455.

Fisher, S., Ronald A. 1930. The Genetical Theory of Natural Selection. Oxford University Press, Oxford.

Gilchrist, M., P. Shah, and R. Zaretzki. 2009. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. Genetics 183:1493–1505.

Gilchrist, M. A. 2007. Combining Models of Protein Translation and Population Genetics to Predict Protein Production Rates from Codon Usage Patterns. Molecular Biology and Evolution 24:2362–2373.

Gilchrist, M. A., W.-C. Chen, P. Shah, C. L. Landerer, and R. Zaretzki. 2015. Estimating Gene Expression and Codon-Specific Translational Efficiencies, Mutation Biases, and Selection Coefficients from Genomic Data Alone. Genome Biology and Evolution 7:1559–1579.

Gilchrist, M. A. and A. Wagner. 2006. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. Journal of Theoretical Biology 239:417–434.

Goldman, N. 1993. Statistical tests of models of DNA substitution. Journal of molecular evolution 36:182–198.

Goldman, N., J. L. Thorne, and D. T. Jones. 1996. Using Evolutionary Trees in Protein Secondary Structure Prediction and Other Comparative Sequence Analyses. Journal of Molecular Biology 263:196 – 208.

Goldman, N., J. L. Thorne, and D. T. Jones. 1998. Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution. Genetics 149:445–458.

Goldman, N. and Z. H. Yang. 1994. Codon-based model of nucleotide substitution for protein-coding DNA-sequences. Molecular Biology and Evolution 11:725–736.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862–864.

Halpern, A. L. and W. J. Bruno. 1998. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. Molecular Biology And Evolution 15:910–917.

Hughes, A. L. and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class-i loci reveals overdominant selection. Nature 335:167–170.

Hurvich, C. M. and C.-L. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76:297–307.

Hhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. Systematic Biology 65:726.

Ingram, T. and D. L. Mahler. 2013. SURFACE: detecting convergent evolution from data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. Methods in ecology and evolution 4:416–425.

Iwasa, Y. 1988. Free fitness that always increases in evolution. Journal of Theoretical Biology 135:265–281.

Jhwueng, D.-C., H. Snehalata, B. C. O'Meara, and L. Liu. 2014. Investigating the performance of AIC in selecting phylogenetic models. Statistical applications in genetics and moleculr biology 13:459–475.

Johnson, S. G. 2012. The NLopt nonlinear-optimization package. Version 2.4.2 – Released 20 May 2014.

Kimura, M. 1962. on the probability of fixation of mutant genes in a population. Genetics 47:713–719.

Koshi, J. M. and R. A. Goldstein. 1997. Mutation matrices and physical-chemical properties: Correlations and implications. Proteins-Structure Function And Genetics 27:336–344.

Koshi, J. M. and R. A. Goldstein. 2000. Analyzing site heterogeneity during protein evolution. Pages 191–202 *in* Biocomputing 2001. World Scientific.

Koshi, J. M., D. P. Mindell, and R. A. Goldstein. 1999. Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes. Molecular biology and evolution 16:173–179.

Kubatko, L., P. Shah, R. Herbei, and M. A. Gilchrist. 2016. A codon model of nucleotide substitution with selection on synonymous codon usage. Molecular Phylogenetics and Evolution 94:290 – 297.

Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology And Evolution 21:1095–1109.

Lynch, M. and G. K. Marinov. 2015. The bioenergetic costs of a gene. Proceedings Of The National Academy Of Sciences Of The United States Of America 112:15690–15695.

Mayrose, I., N. Friedman, and T. Pupko. 2005. A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21:ii151–ii158.

McCandlish, D. M. and A. Stoltzfus. 2014. Modeling evolution using the probability of fixation: History and implications. The Quarterly Review of Biology 89:225–252.

Muse, S. V. and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Molecular Biology and Evolution 11:715–724.

Nielsen, R. and Z. H. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.

Nowak, M. A. 2006. Evolutionary Dynamics: Exploring the Equations of Life. Belknap of Harvard University Press, Cambridge, MA.

O'Meara, B. C., C. Ane, M. J. Sanderson, and W. P.C. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Pellmyr, O. 2002. Microevolution. Pages 731–732 *in* Encyclopedia of Evolution (M. Pagel, ed.). Oxford University Press, Oxford, UK.

Pelmyr, O. 2002. Microevolution. Pages 731–732 *in* Encyclopedia of Evolution (M. Pagel, ed.) vol. 2. Oxford University Press, Oxford, UK.

Pouyet, F., M. Bailly-Bechet, D. Mouchiroud, and L. Guguen. 2016. SENCA: A Multilayered Codon Model to Study the Origins and Dynamics of Codon Usage. Genome Biology and Evolution 8:2427–2441.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Molecular Biology And Evolution 20:1692–1704.

Rodrigue, N. and N. Lartillot. 2014. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. Bioinformatics 30:1020–1021.

Rodrigue, N., N. Lartillot, D. Bryant, and H. Philippe. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347:207–217.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Rowan, T. 1990. Functional Stability Analysis of Numerical Algorithms. Ph.D. thesis University of Texas, Austin.

Salichos, L. and A. Rokas. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Sella, G. and A. E. Hirsh. 2005. The application of statistical physics to evolutionary biology. Proceedings of the National Academy of Sciences of the United States of America 102:9541–9546.

Shah, P. and M. A. Gilchrist. 2011. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proceedings of the National Academy of Sciences of the United States of America 108:10231–10236.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Thorne, J. L., N. Goldman, and D. T. Jones. 1996. Combining protein evolution and secondary structure. Molecular Biology and Evolution 13:666–673.

Thorne, J. L., N. Lartillot, N. Rodrigue, and S. C. Choi. 2012. Codon models as a vehicle for reconciling population genetics with inter-specific sequence data. Codon Evolution: Mechanisms And Models Pages 97–110 D2 10.1093/acprof:osobl/9780199601165.001.0001 ER.

Wagner, A. 2005. Energy constraints on the evolution of gene expression. Molecular Biology and Evolution 22:1365–1374.

Wright, S. 1969. Evolution and the genetics of populations. Vol. 2. The theory of gene frequencies. vol. 2. University of Chicago Press.

Yang, Z. 2014. Molecular Evolution: A Statistical Approach. Oxford University Press, New York.

Yang, Z. H. 1994. Maximum-likelihood phylogenetic estimation from DNA-sequences with variable rates over sites - approximate methods. Journal Of Molecular Evolution 39:306–314.

Yang, Z. H. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology And Evolution 24:1586–1591.

Yang, Z. H. and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. Journal Of Molecular Evolution 46:409–418.

Yang, Z. H. and R. Nielsen. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Molecular Biology and Evolution 25:568–579.

# Table

| Model | logLik | Parameters<br>Estimated | AIC | AICc | $\Delta$AICc | Model<br>Weight |
|:--|--:|--:|--:|--:|--:|:--|
| GTR+$\Gamma$ | -655,166.4 | 610 | 1,311,553 | 1,311,554 | 284,240 | <0.001 |
| GY94 | -612,670.4 | 111 | 1,225,563 | 1,225,563 | 198,249 | <0.001 |
| FMutSel | -597,140.7 | 178 | 1,194,637 | 1,194,638 | 167,324 | <0.001 |
| SelAC$_M$ | -478,302.4 | 50,004 | 1,056,613 | 1,076,674 | 49,360 | <0.001 |
| SelAC | -464,114.8 | 50,004 | 1,028,238 | 1,048,299 | 20,985 | <0.001 |
| SelAC$_M + \Gamma$ | -465,106.9 | 50,005 | 1,030,189 | 1,050,286 | 22,972 | <0.001 |
| SelAC+$\Gamma$ | -453,620.8 | 50,005 | 1,007,252 | 1,027,314 | 0 | >0.999 |

Table 1: Comparison of model fits using AIC, AICc, and AIC$_\text{w}$. Note the subscripts $M$ indicate model fits where the most common or 'majority rule' amino acid was fixed as the optimal amino acid $a^*$ for each site. As discussed in text, despite the fact that $a^*$ for each site was not fitted by our algorithm, its value was determined by examining the data and, as a result, represent an additional parameter estimated from the data and are accounted for in our table. Also, the sample size used in the calculation of AICc is assumed to be equal to the size of the matrix (number of taxa x number of sites).

# Figures

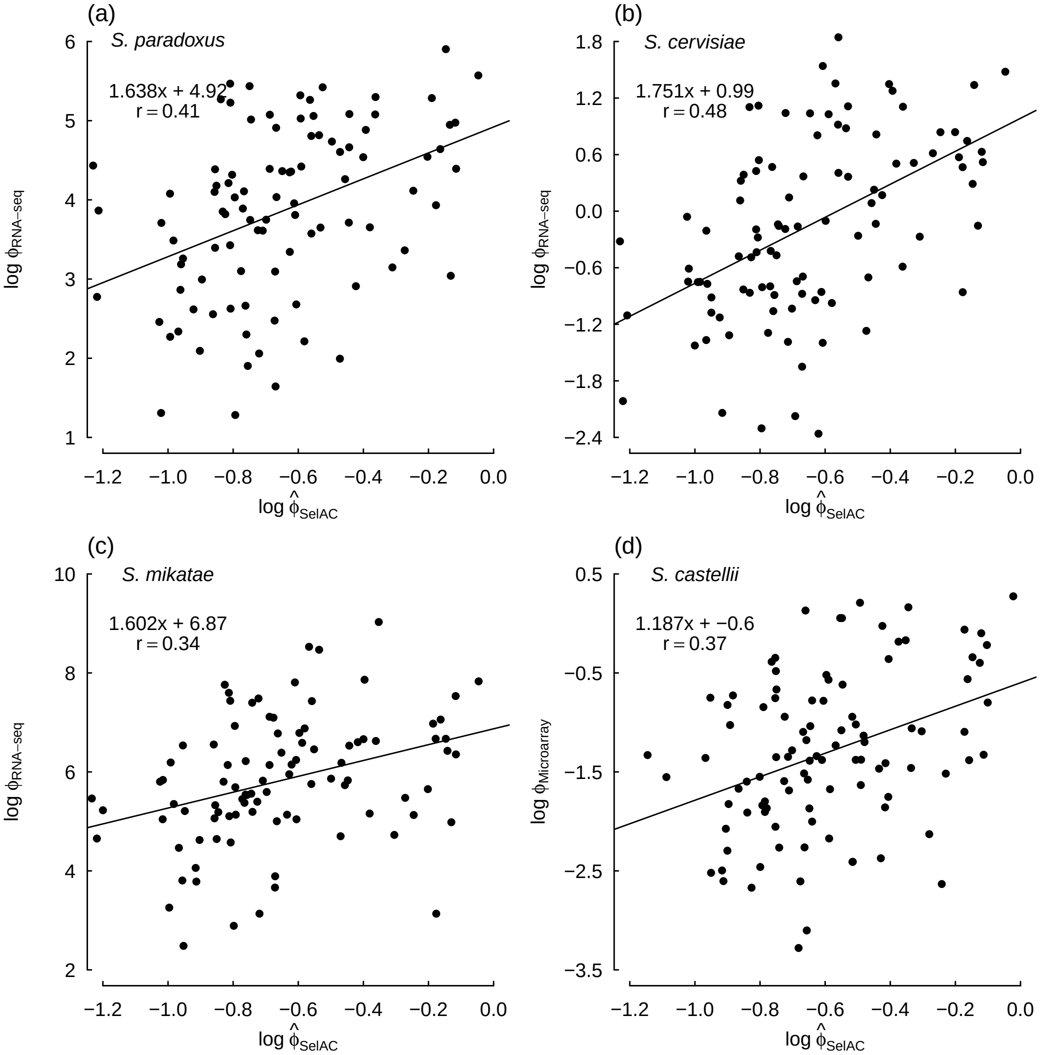

Figure 1: Comparisons between estimates of average protein translation rate $\hat{\phi}_{\text{SelAC}}$ obtained from SelAC+$\Gamma$ and direct measurements of expression for individual yeast taxa across the 100 selected genes from Salichos and Rokas (2013). Estimates of $\hat{\phi}_{\text{SelAC}}$ were generated by dividing the composite term $\psi'$ by $\mathbf{B}(\vec{a}_i|\vec{a}^*)$. Gene expression was measured using either RNA-Seq (a)-(c) or microarray (d). The equations in the upper left hand corner of each panel represent the regression fit and the Pearson correlation coefficient $r$.

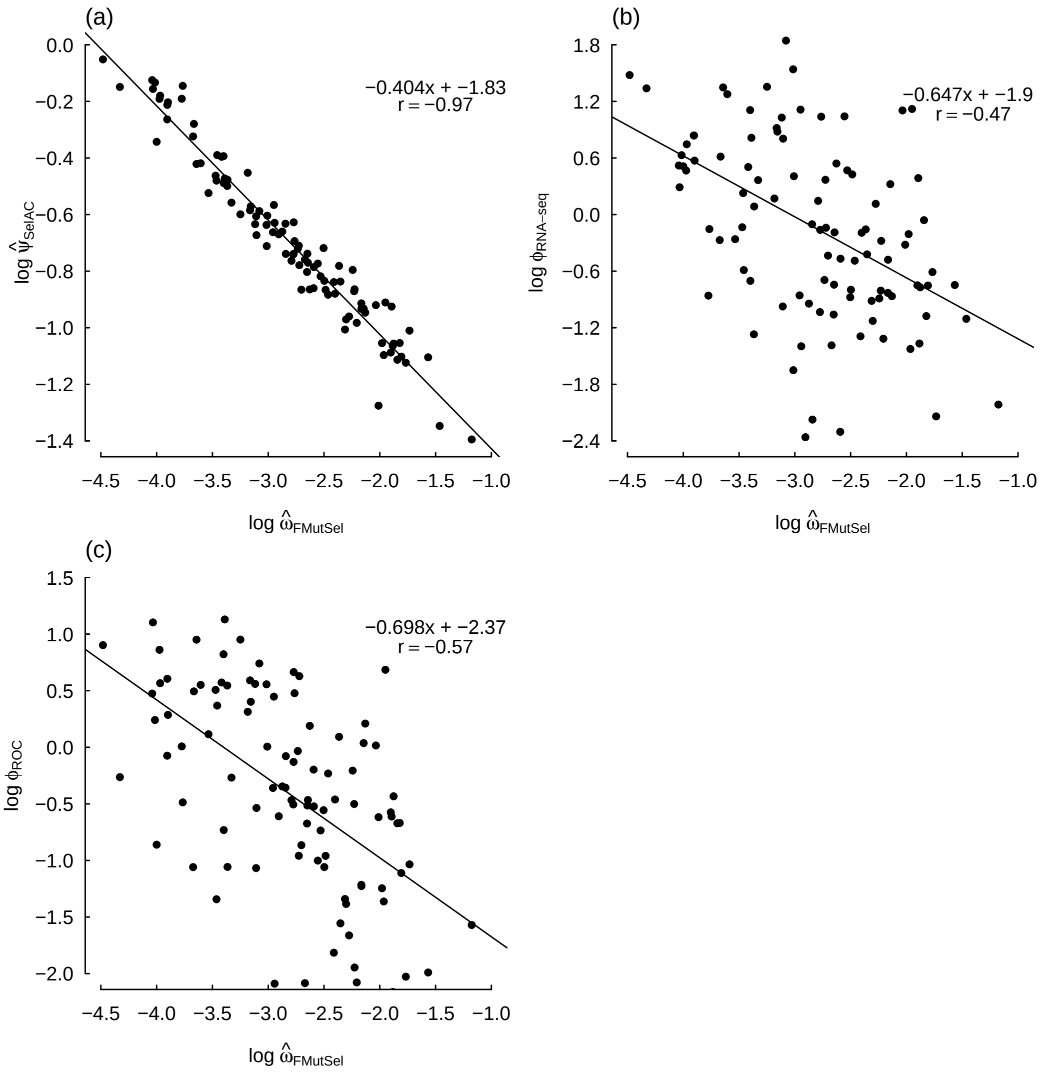

Figure 2: Comparisons between $\omega_{\text{FMutSel}}$, which is the nonsynonymous/synonymous mutation ratio in FMutSel, SelAC+Γ estimates of protein functionality production rates $\hat{\psi}_{\text{SelAC}}$ (a), RNA-Seq based measurements of mRNA abundance $\phi_{\text{RNA-seq}}$ (b), and ROC-SEMPPER's estimates of protein translation rates $\phi_{\text{ROC}}$, which are based solely on *S. cerevisiae*'s patterns of codon usage bias (c), for *S. cerevisiae* across the 100 selected genes from Salichos and Rokas (2013). As in Figure 1, the equations in the upper right hand corner of each panel provide the regression fit and correlation coefficient.

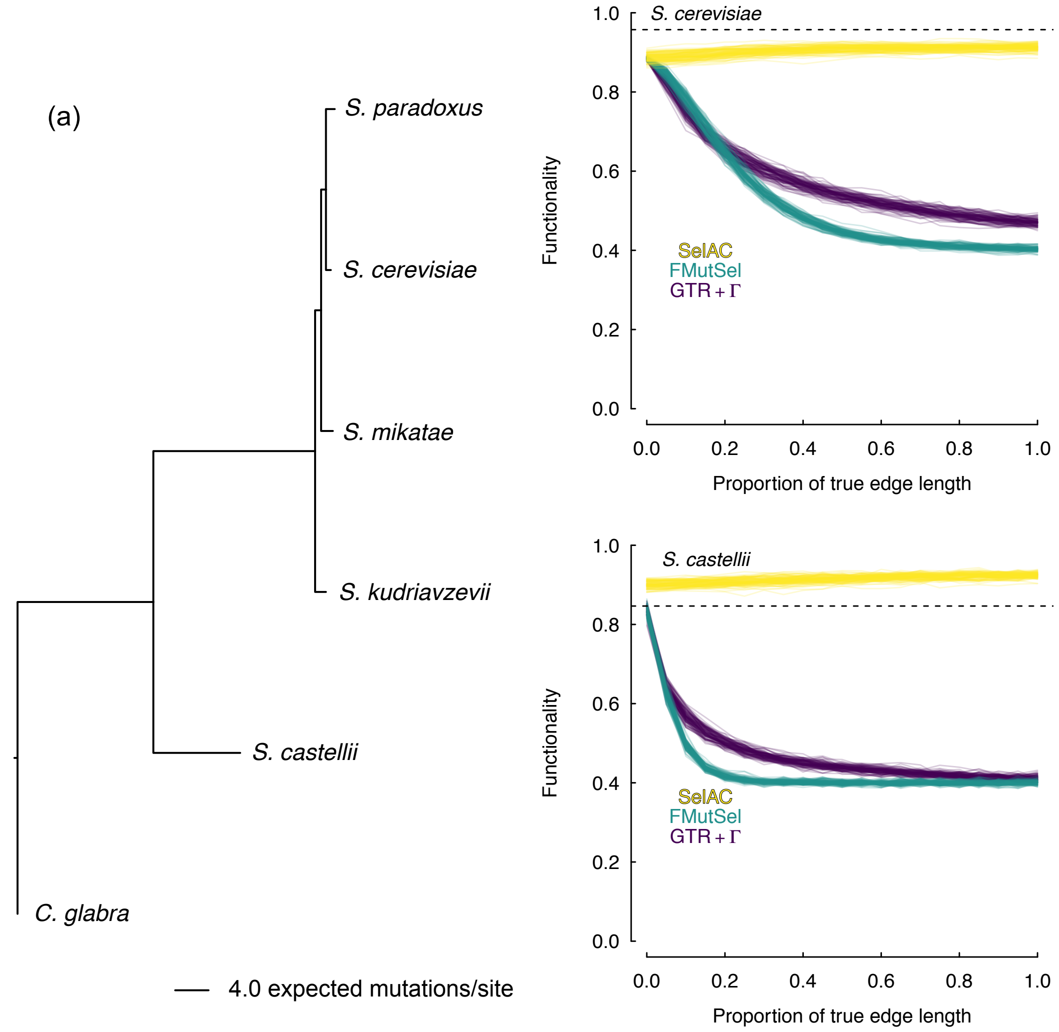

Figure 3: (a) Maximum likelihood estimates of branch lengths under SelAC+Γ for 100 selected genes from Salichos and Rokas (2013). Tests of model adequacy for *S. cerevisiae* (b) and *S. castellii* (c) indicated that, when these taxa are removed from the tree, and their sequences are simulated, the parameters of SelAC+Γ exhibit functionality $\mathbf{B}(\vec{a}_{\text{obs}}|\vec{a}^{*})$ that is far closer to the observed (dashed black line) than data sets produced from parameters of either FMutSel or GTR + Γ.

# Supporting Materials

Supporting Materials for *Population Genetics Based Phylogenetics Under Stabilizing Selection for an Optimal Amino Acid Sequence: A Nested Modeling Approach* by Beaulieu *et al.* (In Review).

## *Comparisons of SelAC gene expression estimates with empirical measurements*

In our model, the parameter $\phi$ measures the realized average protein synthesis rate of a gene. We compared our estimates of $\phi$ to two separate measures of gene expression, one empirical (Figure S1), and one model-based prediction that does not account for shared ancestry, for individual yeast taxa across the same set of genes. Our estimates of $\phi$ are positively correlated with both measures, which are also reasonably well correlated with each other (Figure 1 - S2) On the whole, these comparisons indicate not only a high degree of consistency among all three measures, but also, importantly, that estimates of $\phi$ obtained from SelAC provide real biological insight into the expression level of a gene.

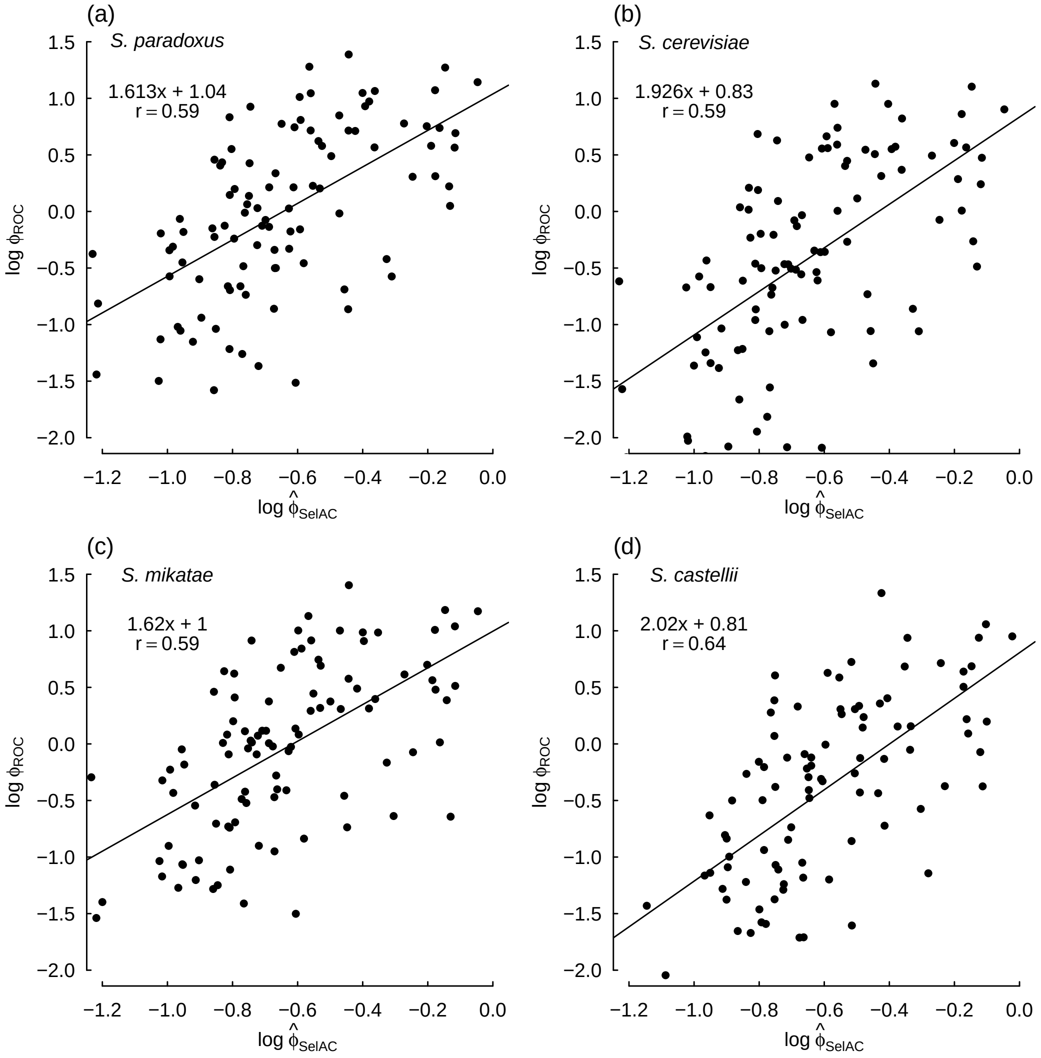

Figure S1: Comparisons between estimates of $\phi$ obtained from SelAC+$\Gamma$ and the predicted gene expression from the ROC SEMPER model (Gilchrist et al. (2015)) for individual yeast taxa across the 100 selected genes from Salichos and Rokas (2013). As with figures in the main text, estimates of $\phi$ were obtained by solving for $\psi$ based on estimates of $\psi'$, and then dividing by $\mathbf{B}(\vec{a}_i|\vec{a}^*)$. The equations in the upper left hand corner of each panel represent the regression fit and correlation coefficient.

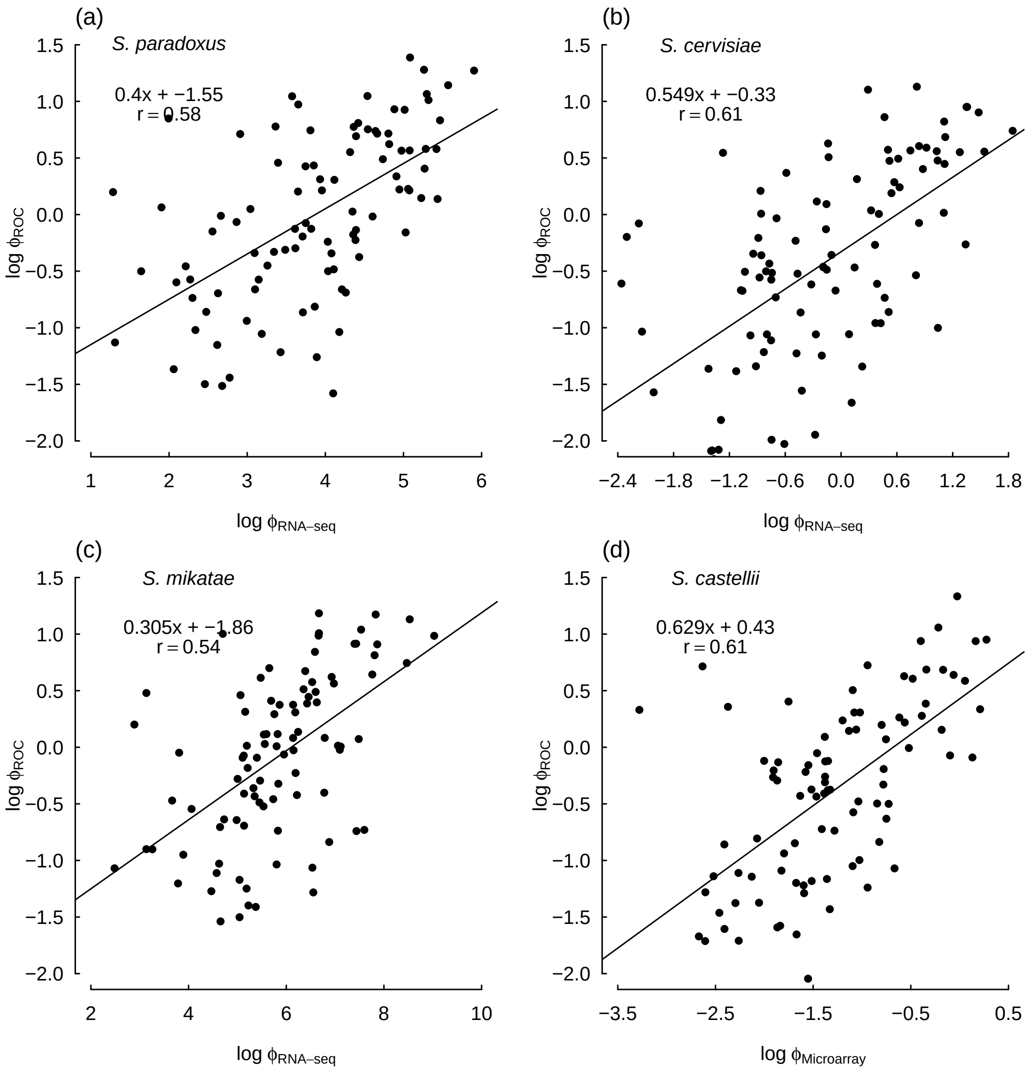

Figure S2: Comparisons of predicted gene expression from the ROC SEMPER model (Gilchrist et al. (2015)) and direct measurements of expression from RNA-Seq or microarray data for individual yeast taxa across the 100 selected genes from Salichos and Rokas (2013). The equations in the upper left hand corner of each panel represent the regression fit and correlation coefficient.

## *Simulations*

Overall, the simulation results indicate that the SelAC model can reasonably recover the known values of the generating model (Figure S3 - S6). This includes not only the parameters in SelAC, but also the optimal amino acids for a given sequence as well as the estimates of the branch lengths. There are a few observations to note. First, the ability to accurately recover the true optimal amino acid sequence will largely depend on the magnitude of the realized average protein synthesis rate of the gene $\phi$. This is, of course, intuitive, given that $\phi$ sets the strength of stabilizing selection towards an optimal amino acid at a site. However, the inclusion of $\alpha_G$ into SelAC, appears to generally increase values of $\phi$ and generally improves the ability to recover the optimal amino acids even for the gene with the lowest baseline $\phi$. Second, we found a strong downward bias in estimates of $\alpha_G$, which actually translates to greater variation among the rate categories. The choice of a gamma distribution to represent site-specific variation in sensitivity was based on mathematical convenience and convention, rather than on biological reality. Nevertheless, we suspect that this bias is in large part due to the difficulty in determining the baseline $\psi$ for a given gene and the value of $\alpha_G$ that globally satisfies the site-specific variation in sensitivity across all genes, as indicated by the slight upward bias in estimates of $\psi$. A reviewer pointed out that it may also be difficulty for SelAC to account for changing amino-acid, which we agree may also play a role. It has been suggested, in studies of the behavior of the gamma distribution in applications of nucleotide substitution model, that increasing the number of rate categories can often improve accuracy of the shape parameter (Mayrose et al. (2005)). Future work will address this issue.

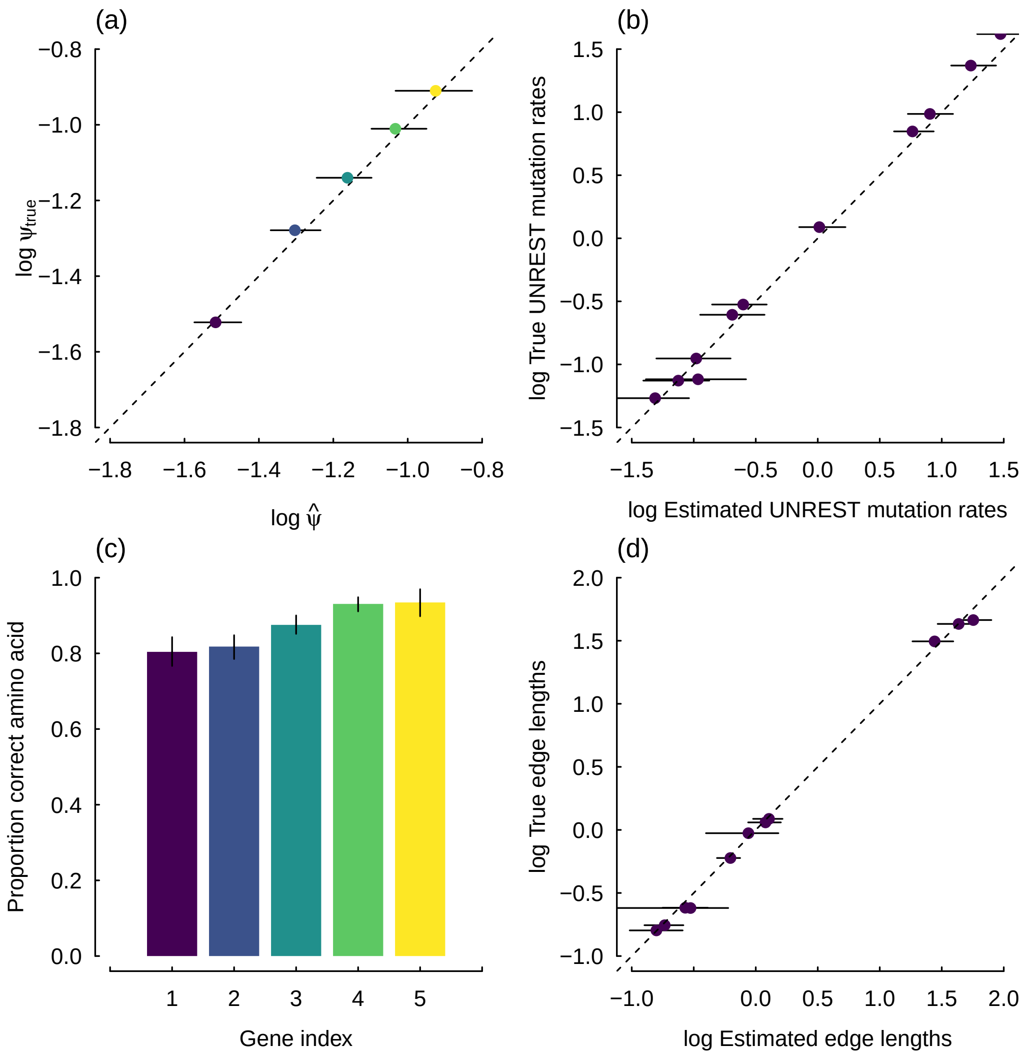

Figure S3: Summary of a 5-gene simulation for a SelAC model where we assume $\alpha_G = \infty$, and thus, no site-specific sensitivity in the generating model. The 'known' parameters were based on fitting the same SelAC to the 106 gene data set and phylogeny of Rokas et al. (2003), with gene choice being based on five evenly spaced points along the rank order of the gene specific composite parameter $\psi'_g$. The points and associated uncertainty in the estimates of the gene-specific average protein synthesis rate, or $\psi$ (calculated from $\psi'$)(a), nucleotide mutation rates under the UNREST model (b), proportion of correct optimal amino acids for a given gene (c), and estimates of the individual edge lengths are based the mean and 2.5% and 97.5% quantiles across all 50 simulated datasets (d). Gene index on the x-axis refers to the arbitrary number assigned to the simulated gene.

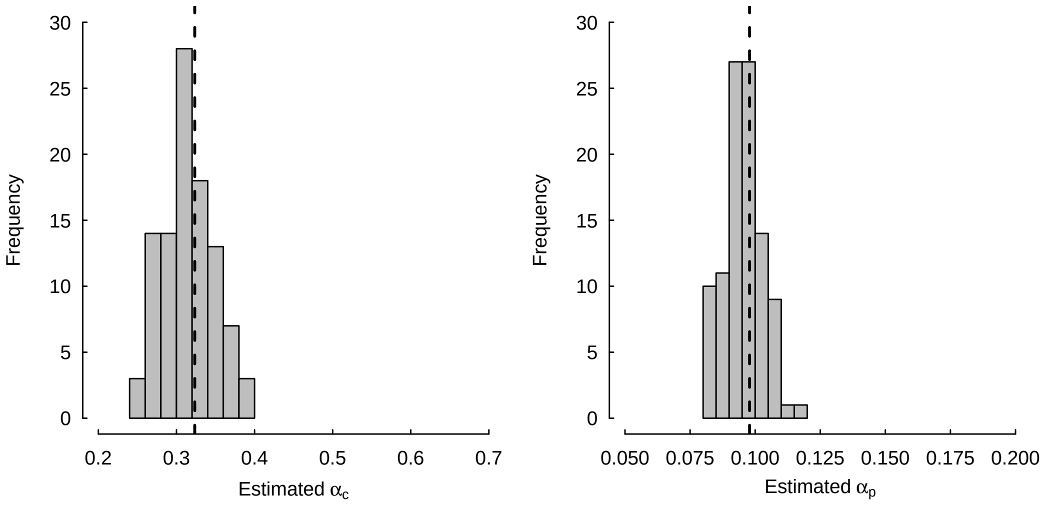

Figure S4: The distribution of estimates of the Grantham weights, $\alpha_c$ and $\alpha_p$, in a SelAC model, where we assume $\alpha_G = \infty$, and thus no site-specific sensitivity in the generating model. The dashed line represents the value used in the generating model.

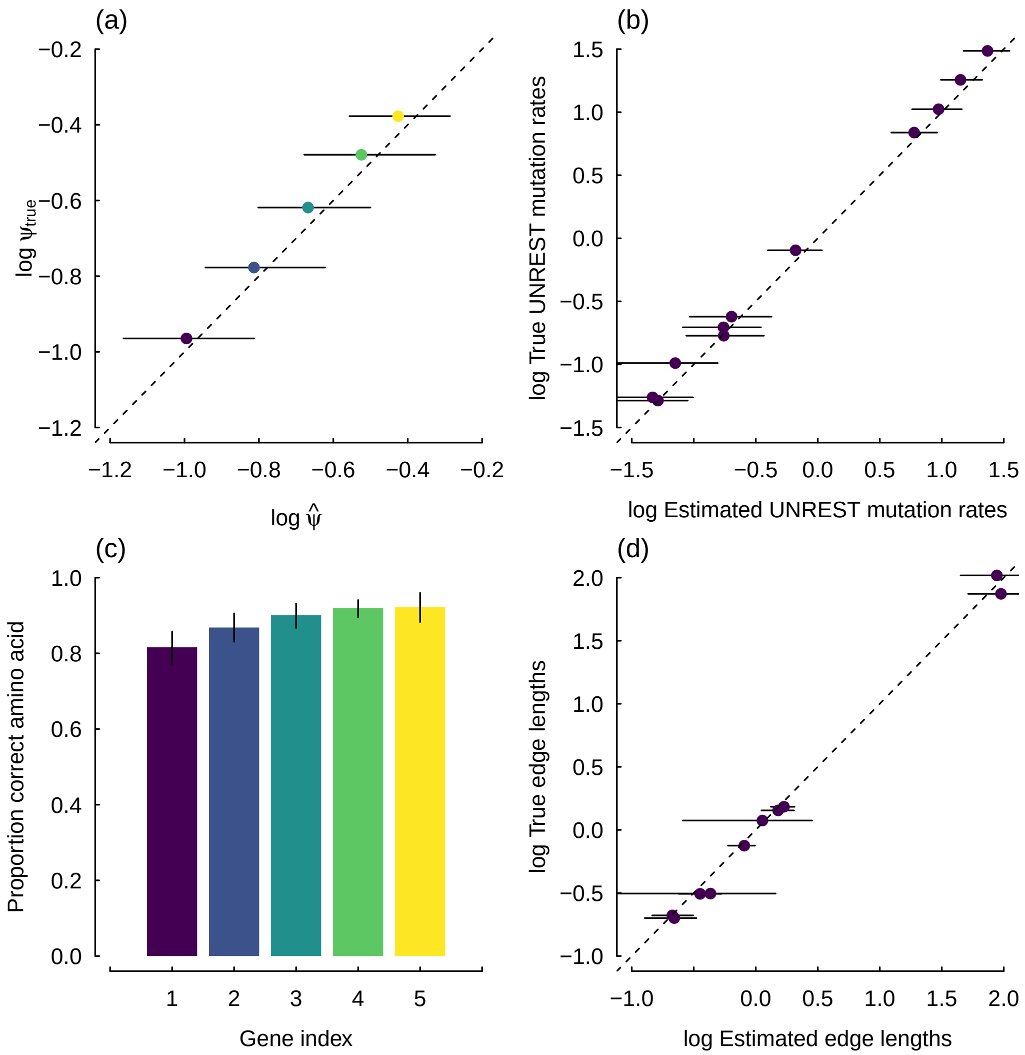

Figure S5: Same figure as in Figure S3, except the generating model includes site-specific sensitivity in the generating model (i.e., $\alpha_G$).

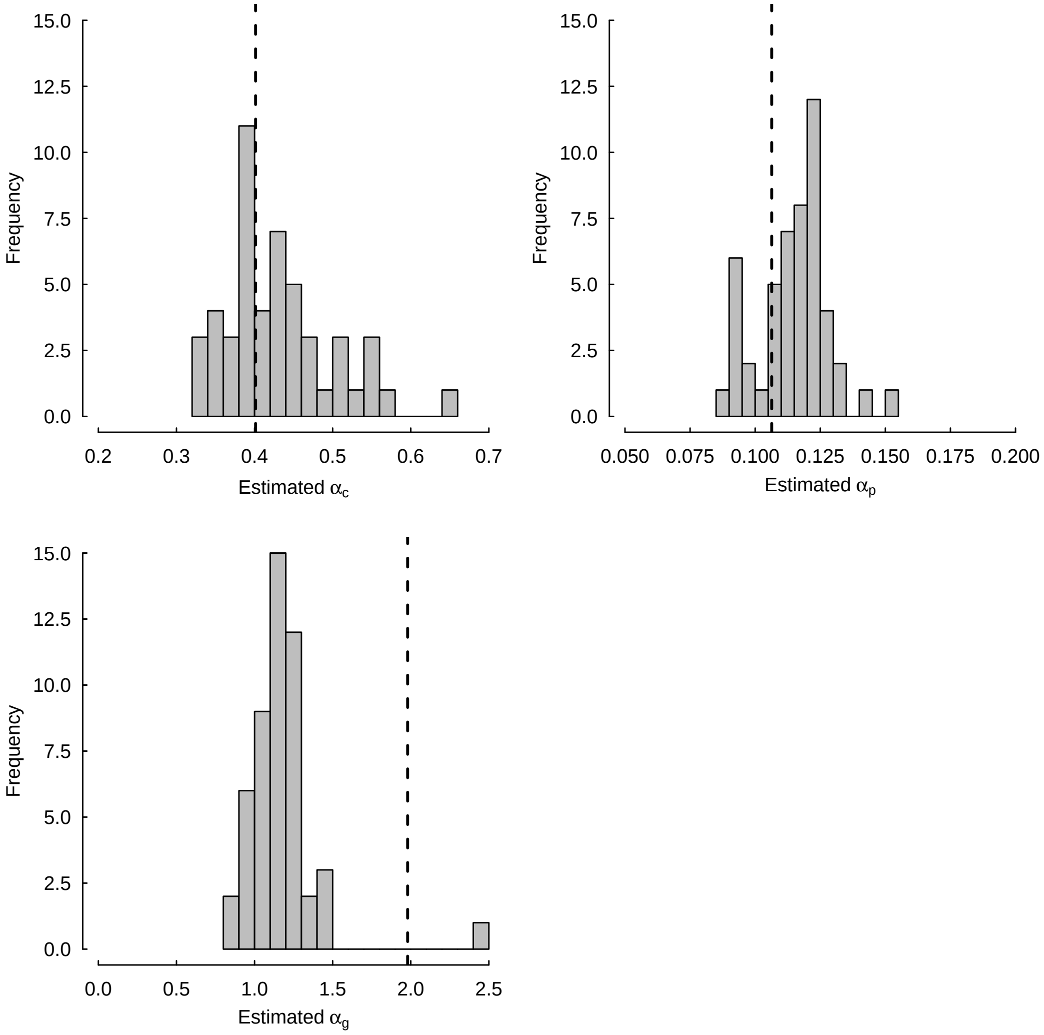

Figure S6: Same figure as in Figure S4, except the generating model includes site-specific sensitivity in the generating model (i.e., $\alpha_G$). Unlike, Grantham weights, which showed no systematic bias, there is a downward bias in estimates of $\alpha_G$.