Phenomenological Load on Model Parameters Can Lead to False Biological Conclusions

Christopher T. Jones,*[,1] Noor Youssef,[2] Edward Susko,[1] and Joseph P. Bielawski[2]

[1]Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada

[2]Department of Biology, Dalhousie University, Halifax, NS, Canada

***Corresponding author:** E-mail: cjones2@dal.ca.

Associate editor: Jeffrey Thorne

Abstract

When a substitution model is fitted to an alignment using maximum likelihood, its parameters are adjusted to account for as much site-pattern variation as possible. A parameter might therefore absorb a substantial quantity of the total variance in an alignment (or more formally, bring about a substantial reduction in the deviance of the fitted model) even if the process it represents played no role in the generation of the data. When this occurs, we say that the parameter estimate carries phenomenological load (PL). Large PL in a parameter estimate is a concern because it not only invalidates its mechanistic interpretation (if it has one) but also increases the likelihood that it will be found to be statistically significant. The problem of PL was not identified in the past because most off-the-shelf substitution models make simplifying assumptions that preclude the generation of realistic levels of variation. In this study, we use the more realistic mutation-selection framework as the basis of a generating model formulated to produce data that mimic an alignment of mammalian mitochondrial DNA. We show that a parameter estimate can carry PL when 1) the substitution model is underspecified and 2) the parameter represents a process that is confounded with other processes represented in the data-generating model. We then provide a method that can be used to identify signal for the process that a given parameter represents despite the existence of PL.

Key words: **codon substitution models, mechanistic versus phenomenological, phenomenological load, false positives, reliability.**

Introduction

There are in general two ways to quantitatively describe a natural process. The phenomenological approach is to summarize relationships between variables with little or no reference to causation. The alternative is to specify a model based on known or hypothetical mechanistic links between variables that explain their relationships. For example, although Newton's law of universal gravitation provides a highly accurate description of the apparent force of attraction between objects, it does so without explaining the cause. Newton's law is therefore phenomenological. Einstein, by contrast, described gravitation mechanistically as the result of mass generating curvature in space-time. Biology is replete with models of complex processes that cannot be placed into either of these neat categories. On the one hand, there is a natural desire to build mechanistic models that capture as much of the complexity and richness of a process as possible. On the other hand, limitations in information and computational resources often make simplifying assumptions unavoidable, thereby forcing a more phenomenological approach. The result of this tension is that models of biological processes often fall somewhere on a continuum between phenomenological and mechanistic.

A key feature of a model characterized as mechanistic is that its parameters are interpretatable with repsect to the real data-generating process (Liberles et al. 2013). This underlines the fact that the terms mechanistic and phenomenological are more aptly applied to individual parameters. Indeed, mechanistic and phenomenological parameters often appear together in the same model (Rodrigue and Philippe 2010). However, the distinction between the two is not always clear. Codon substitution models (CSMs), for example, consist of two submodels, one for the effect of selection at the amino acid level (the selection submodel) and the other for DNA-level substitution processes (the DNA submodel). The processes described by these submodels, the appearance of a new mutation in an individual and its eventual fixation or elimination in the population, are necessarily probabilistic (Moran 1958; Kimura 1962). In this context, we define a mechanistic parameter as one meant to provide an explanation for differences in probability distributions rather than in specific outcomes. For example, a probabilistic bias for or against replacement substitutions is represented in the selection submodel by a nonsynonymous-to-synonymous rate ratio ω. An estimate $\hat{\omega}$ of this rate ratio has traditionally been used to support one of three possible explanations: stringent selection by $\hat{\omega} < 1$; neutrality by $\hat{\omega} = 1$; and positive selection by $\hat{\omega} > 1$. However, ω can only be estimated by combining the information contained in a number of sites, at least when the number of taxa is limited (but see Rodrigue et al. 2010;

Tamuri et al. 2012, 2014; Rodrigue and Lartillot 2014; Spielman and Wilke 2016). It follows that $\hat{\omega}$ only provides a summary of selection effects, which are typically heterogenous across sites, and does not represent the effects at any individual site. In this sense, we argue, the rate ratio is phenomenological or at least closer to the phenomenological end of the continuum. By contrast, it is generally assumed that the mutation process is the same across sites and over time. An estimate of the probabilistic bias in favor of mutations that are transitions over mutations that are transversions, represented in by κ in the DNA submodel, is therefore closer to the mechanistic end of the continuum because its interpretation applies equally to any specific site.

The distinction between mechanistic and phenomenological is further obfuscated by the way parameters are estimated. Under the maximum likelihood (ML) framework, the likelihood of a set of model parameters, such as the rate ratio ω and the vector of branch lengths *b*, is expressed in the form of a likelihood function $L(\omega, b|X, T)$, where *X* represents the alignment and *T* the assumed topology of the tree. The maximum likelihood estimate (MLE) for (ω, b) is the vector $(\hat{\omega}, \hat{b})$ that maximizes the likelihood $L(\omega, b|X, T)$ of the data. A key feature of the ML framework is that the likelihood of the data always increases when a new parameter is added to the fitted model. For example, $L(\hat{\omega}, \hat{b}, \hat{\psi}|X, T)$ must theoretically be greater than $L(\hat{\omega}, \hat{b}|X, T)$. The new parameter ψ is said to have improved the fit of the model in proportion to the size of the increase in likelihood it engendered, and is said to be statistically significant if the increase is larger than some prespecified threshold. Under this framework, it is possible to find that ψ is statistically significant even if the process it represents did not actually play a role in the generation of the data.

All CSMs are misspecified, meaning that they do not match the true generating process. For example, the selection submodel of the simplest M-series model M0 (Yang et al. 2000) characterizes the selection process with one rate ratio for all sites and branches. It is underspecified because it does not account for any heterogeneity in the selection process. In general, if the selection submodel of a CSM fails to absorb a substantial proportion of the variation in site patterns due to selection effects, some of this variation might be inappropriately absorbed by parameters of the DNA submodel. This is especially likely to occur when a parameter of the DNA submodel represents a process that is counfounded with selection effects. We say that two process are confounded if they can produce similar patterns or "signatures" in the data. Under the true evolutionary process, for example, the rate ratio at a site depends on the codon occupying the site, with higher values corresponding to codons that are less fit (Jones et al. 2017). Random changes in rate ratio (a.k.a. heterotachy) can therefore arise by episodic movement of the site away from and back to its optimal amino acid, provided neither selection nor drift dominates the site (e.g., by shifting balance, Jones et al. 2017). The process of shifting balance would be confounded with any other process that produces similar variations in rate ratio over time.

To assess the impact of confounding on the MLE of a model parameter, we introduce the concepts of percent reduction in deviance (PRD) and phenomenological load (PL). Formally, deviance is the difference between the maximum log-likelihood (LL) of a given CSM and the maximum log-likelihood of the saturated model (Ms) when both are fitted to the same alignment. The saturated model, analagous to a regression model in which there are as many predictor variables as observations, will always provide the largest log-likelihood of any CSM. The difference between this and the log-likelihood of M0 (i.e., the simplest M-series CSM) provides a baseline deviance score for comparison with differences between other pairs of models. The deviance under a model M can be reduced by the addition of a new parameter ψ. The PRD of the MLE $\hat{\psi}$ is the decrease in deviance it engenders normalized by the baseline score. A large PRD is generally considered to indicate that the new parameter improved model fit. However, better fit does not imply a better model. If ψ has a mechanistic interpretation, and if the process it represents did not actually occur when the data were generated, we equate PRD to the PL carried by $\hat{\psi}$. A large PL is a concern because it not only invalidates the mechanistic interpretation of $\hat{\psi}$ but also increases the likelihood that ψ will found to be statistically significant. Under this scenario, the model M with ψ will provide a better fit, but would also lead to false conclusions about the true data-generating process.

In this study, we test the hypothesis that, when the selection submodel of a CSM is both underspecified and confounded with processes specified by the DNA submodel, PL can arise and lead to false biological conclusions. We chose to focus on the fixation of simultaneous double and triple (DT) mutations. The majority of CSMs assume that sites evolve by a series of single nucleotide substitutions, despite evidence for fixation of DT mutations (Whelan and Goldman 2004; Kosiol et al. 2007; Tamuri et al. 2012). Several authors have argued that it would be beneficial to add a few extra parameters to the DNA submodel of any standard CSM to account for DT mutations (Miyazawa 2011; Zaheri et al. 2014). To investigate the utility of this recommendation, we added DT parameters (α and β, see Materials and Methods) to a variety of CSMs. Our main study consists of an analysis of the propensity of these models to detect the fixation of DT mutations in alignments generated with single nucleotide substitutions only. Two smaller studies were also conducted, including an analysis of the impact of PL on a parameter intended to measure changes in the intensity of selection (the RELAX model of Wertheim et al. 2014), and an analysis of the impact of PL on parameters intended to account for variations in the synonymous substitution rate (dS) across sites (Kosakovsky Pond and Muse 2005). We adduce the results of all three analyses as evidence of the universal applicability of the PL concept.

Paper Outline

The CSMs used in this study included M0, M3($k = 2$), and CLM3($k = 2$) (a covarion-like CSM described in Jones et al. 2017). These models are increasingly complex in their submodels for selection: M0 models selection with a single rate ratio ω_0, M3 with two rate ratios ω_0 and ω_1 and a proportion

p_0, and CLM3 with two rate ratios ω_0' and ω_1', a proportion p_0' and a switching rate δ. We extend these further by introducing a novel CSM called RaMoSS (for **Ra**ndom **M**ixture **o**f **S**tatic and **S**witching sites). RaMoSS combines M3 with CLM3 to account for a RaMoSS site in an alignment, as first suggested by Galtier (2001). M0, M3, CLM3, and RaMoSS all share the same DNA submodel restricted to allow the fixation of single nucleotide mutations only. A counterpart model was formulated for each CSM to allow fixation of DT mutations. These models, M0wDT, M3wDT, CLM3wDT, and RaMoSSwDT, provide a series of alternatives to test for evidence of the fixation of DT mutations, and represent a range of opportunities for PL that might potentially lead to incorrect conclusions about the data-generating process.

To illustrate that signatures for the fixation of DT mutations can be detected in real data using each of the four model-M versus model-MwDT contrasts, we start with an analysis of 12 concatenated H-strand mitochondrial DNA sequences (3,331 codon sites) from 20 mammalian species as distributed in alignment form by the PAML software package (Yang 2007). Fixation of DT mutations was detected in this alignment by all four contrasts. However, the inferred proportion of fixations that were DT, as well as the PRD attributed to the MLEs of the DT parameters, decreased with each incremental increase in the complexity of the submodel for selection. This trend is consistent with the hypothesis that DT parameters carry PL in proportion to the degree to which the submodel for selection is underspecified, and casts doubt on the veracity of the detection of DT in the real alignment.

Next, we report the results of three simulation studies, all of which used alignments generated with point mutations only. In the first, alignments were generated under RaMoSS with parameters set to those estimated from the real mtDNA alignment. The RaMoSS versus RaMoSSwDT contrast reliably failed to reject the null hypothesis of no DT fixations in all trials. In the second, alignments were generated to be misspecified with respect to RaMoSS using a CSM that assigned a different rate ratio to each site. The purpose was to exemplify the traditional approach of simulating with a more complex CSM to assess the impact of misspecification. The RaMoSS versus RaMoSSwDT contrast failed to reject the null hypothesis of no DT fixations in 95% these trials. In the past, this would have been considered satisfactory evidence for the reliability of the contrast, and used to support the veracity of the detection of DT in the real mtDNA alignment. However, the RaMoSS versus RaMoSSwDT contrast incorrectly rejected the null in >40% of trials in the third simulation study for which the mutation-selection (MutSel) framework (Halpern and Bruno 1998) was used to generate alignments. The MutSel generating procedure that was used (MutSel-mmtDNA) was designed to produce alignments to match the real mammalian mtDNA as closely as possible. The results of the third simulation suggest that the submodel for selection under RaMoSS was often insufficiently sensitive to signatures of heterotachy to protect DT parameters against PL.

Our analysis of DT is followed by two more examples that demonstrate the utility of the PL concept. When RELAX (Wertheim et al. 2014) was fitted to the real mtDNA alignment, it detected significant relaxation of selection pressure (via a single parameter m) in the primate clade. However, it also detected relaxation in the same clade in 31 out of 50 alignments generated under MutSel-mmtDNA with no relaxation. This suggests that the MLE for m has the potential to carry PL. By contrast, we found no evidence for variation in dS across sites when M3wdS (our name for the CSM that accounts for variations in dS across sites, Kosakovsky Pond and Muse 2005) was fitted to the same 50 alignments, despite detecting such variations in the real mtDNA alignment using the same model. This suggests that the parameters for dS variation under M3wdS do not carry PL, and that the signal detected in the real alignment was genuine. Taken together, our findings have broad implications about the formulation and experimental validation of CSMs. Specifically, we maintain that only alignments generated with realistic variations in the evolutionary process across sites and over time can reveal pathologies arising from PL, and that such pathologies are not necessarily evident when a model is tested using alignments generated under the traditional modeling framework.

New Approaches

Modeling a Mixture of Static and Switching Sites: RaMoSS

Many commonly used CSMs assume either that rate ratios vary across sites but not time (e.g., the M-series models of Yang et al. 2000), or that temporal variations occur at all sites (e.g., the branch-site models of Guindon et al. 2004; Kosakovsky Pond et al. 2011; Murrell et al. 2015). One exception is the branch-site model of Yang and Nielsen (2002), which allows some sites to evolve under the same rate ratio across the whole tree, and others to switch from a stringent or neutral selection regime to positive selection at a specific location in the tree. The location of the switch, based on prior information, is treated as a fixed effect. Although this approach is well suited for identifying episodic directional selection on a specific branch, it is inappropriate for detecting random site-specific variations in rate ratio. Since real alignments might include both static and switching sites, we propose a new model, RaMoSS, that combines the standard M-series model M3($k = 2$) (hereafter, simply M3) with the covarion-like model CLM3($k = 2$) (hereafter, CLM3) (cf., Galtier 2001; Guindon et al. 2004). Specifically, RaMoSS mixes (with proportion p_{M3}) one selection submodel with two rate-ratio categories $\omega_0 < \omega_1$ that are constant over the entire tree with a second selection submodel (with proportion $p_{CLM3} = 1 - p_{M3}$) under which sites switch randomly in time between $\omega_0' < \omega_1'$ at an average rate of δ switches per unit branch length. See Materials and Methods for additional details.

Quantifying PL

Phenomenological load can be quantified using the concept of statistical deviance. Let $P_M(x|T, \theta_M)$ be the probability of a site pattern x under a model M, topology T, and vector of model parameters θ_M. This defines a distribution on the set of

all possible site patterns. Under the usual assumption that sites are independently and identically distributed (iid), the probability of an alignment X is the product of the probabilities of its site patterns:

$$P_M(X|T,\theta_M) = \prod_{h=1}^{n} P_M(x^h|T,\theta_M) \quad (1)$$

where h is the site index for n sites. Equation (1) defines a distribution on the set of all possible alignments, meaning that X is the variable. Under the ML framework, the vector θ_M is considered to be the variable instead. To signify this, the probabilities in (1) are called likelihoods and are written with θ_M as the argument:

$$L_M(\theta_M|X,T) = \prod_{h=1}^{n} L_M(\theta_M|x^h,T) \quad (2)$$

It is standard practice to apply a natural-log transform to (2) to obtain what is called the log-likelihood (LL) of θ_M under the model M:

$$\begin{aligned}\ell_M(\theta_M|X,T) &= \ln\{L_M(\theta_M|X,T)\} \\ &= \sum_{i=1}^{k} y_i \ln\{L_M(\theta_M|x_i,T)\}\end{aligned} \quad (3)$$

The $x_i \in \{x_1,\ldots,x_k\}$ represent the unique site patterns in the alignment, each of which occurs y_i times.

The objective of the ML approach is to find the vector that maximizes equation (3). The resulting vector is called the maximum likelihood estimate (MLE) of θ_M, denoted $\hat{\theta}_M$. Speaking metaphorically, $\hat{\theta}_M$ accounts for or "absorbs" as much of the variance in the site-patterns of the alignment as possible. More formally, $\hat{\theta}_M$ minimizes the deviance of the fitted model, which is defined as the difference between the LL of the fitted model and the LL of the most general iid model (a.k.a., the saturated model, Ms). The MLE of the probability of a site pattern x_i under Ms can be shown to be its observed relative frequency y_i/n. Hence, the LL for the saturated model is:

$$\ell_{Ms}(X) = \sum_{i=1}^{k} y_i \ln(y_i/n) \quad (4)$$

To provide an interpretation, consider that any CSM fitted to an N-taxon alignment of mtDNA can be thought of as a multinomial distribution for the 60^N possible site patterns (or 61^N for nuclear DNA). Ms is the unique multinomial distribution defined by the vector of observed relative frequencies $(y_1/n,\ldots,y_k/n)$. In other words, the saturated model is specified by the empirical site-pattern distribution of X. Because it takes none of the mechanisms of mutation or selection into account, ignores the phylogenetic relationships between sequences (i.e., is independent of T), and excludes the possibility of site patterns that were not actually observed (i.e., the probability of a site pattern that was not observed is assumed to be 0), Ms can be construed as the maximally phenomenological explanation of X. The salient feature of (4) is that it is always larger than the LL for any CSM. It is in this sense that Ms is saturated, akin to a regression model with the same number of predictor variables as observations. For this reason, Ms provides a natural benchmark for model comparisons.

The selection submodel under M0 consists of one rate ratio for all sites, similar to a regression model that fits an intercept only. The deviance under M0 is defined as:

$$D_{M0} = -2\{\ell_{M0}(\hat{\theta}_{M0}|X,T) - \ell_{Ms}(X)\} \quad (5)$$

Equation (5) provides a baseline with which to compare changes in deviance for other model contrasts. For example, suppose M_ψ is the same model as M but with one extra parameter ψ. The statistical significance of this new parameter can be assessed by conducting a hypothesis test based on the log-likelihood ratio (LLR) statistic for the M versus M_ψ contrast:

$$\text{LLR} = D_M - D_{M_\psi} = -2\{\ell_M(\hat{\theta}_M|X,T) - \ell_{M_\psi}(\hat{\theta}_{M_\psi}|X,T)\} \quad (6)$$

Equation (6) is an absolute measure of the decrease in deviance caused by the addition of ψ to M. An alternative relative measure is what we call the percent reduction in deviance (PRD):

$$\text{PRD}(\hat{\psi}) = \frac{D_M - D_{M_\psi}}{D_{M0}} \times 100\% \quad (7)$$

This quantity can be construed as reflecting the strength of the signature for the process represented by ψ combined with random error and possibly PL. However, if an alignment is generated with ψ set to the value that precludes the process it represents (e.g., $\psi = 0$), then $\text{PRD}(\hat{\psi})$ is due to PL and random error only; we use the notation $\text{PL}(\hat{\psi})$ in place of $\text{PRD}(\hat{\psi})$ to emphasize that this is the case. It is this scenario that can lead to false biological conclusions.

Assessing the Realism of Alignments Simulated under MutSel

The standard way to assess the sensitivity of a CSM to misspecification has been to fit the model to alignments simulated using another, perhaps more complex, CSM (Anisimova et al. 2001, 2002; Wong et al. 2004; Zhang 2004; Kosakovsky Pond and Frost 2005; Yang et al. 2005; Zhang et al. 2005; Kosakovsky Pond et al. 2011; Yang and dos Reis 2011; Lu and Guindon 2014). The problem with this approach is that alignments generated under even a relatively complex CSM are not misspecified in the same way as real data. Off-the-shelf CSMs make the unrealistic assumption that all sites evolve under the same vector of stationary frequencies, and assume that all nonsynonymous substitutions have the same probability of fixation for a given rate ratio. These assumption preclude the generation of realistic levels of variation in rate ratio across sites and over time, and have until now prevented recognition of the problem that we call PL. The MutSel framework of Halpern and Bruno (1998) provides a way to evolve a codon site over a tree that is consistent with the dynamics of

an ideal Wright–Fisher population on a static fitness landscape. Under this framework, each site can be assigned its own vector of fitness coefficients. Amino acid proclivities and the stringency of selection reflected by the average rate ratio at a site can therefore be made to vary across sites in a way that is consistent with a real alignment. Alignments generated under MutSel can also exhibit heterotachy, which may comprise a significant proportion of the total variation in a real alignment (Lopez et al. 2002; Jones et al. 2017). MutSel therefore seems to be the ideal framework for generating realistic data with which to assess the reliability of a CSM.

The degree to which an alignment generated under MutSel mimics real data is in large part dependent on how site-specific fitness coefficients are specified. The most direct approach is to make use of site-specific amino acid frequencies derived from real data. For example, Spielman and Wilke (2016) estimated vectors of site-specific fitness coefficients from codon frequencies observed in structurally curated alignments of at least 150 taxa. These were then fed into Pyvolve (Spielman and Wilke 2015a), among the first open-source software packages with the option to evolve sites using the MutSel framework, to produce simulations of the original alignments. To explore how model misspecification might have influenced our analysis of the 20-taxon alignment of mammalian mtDNA, it was necessary to simulate alignments consistent with those data. Unfortunately, the methods presented in Spielman and Wilke (2016) are inappropriate for such a limited number of taxa. We therefore devised a new method for generating plausible vectors of site-specific fitness coefficients, which we call MutSel-m(ammalian)mtDNA. A detailed description of MutSel-mmtDNA is provided in the Supplementary Material online that accompanies this article.

More important than the new method of simulation is the way it was assessed for realism. We validated MutSel-mmtDNA by comparing distributions of summary statistics from simulated alignments to those of the real mtDNA alignment. The summary statistics considered were 1) the distribution of the number of amino acids per codon site, 2) the overall amino acid and codon frequencies, and 3) the frequency with which each pair of amino acids appeared together in the same site pattern. In addition, the expected distribution of simulated scaled selection coefficients for all mutations, all substitutions, all nonsynonymous mutations, and all nonsyonymous substitutions generated under MutSel-mmtDNA were compared with their empirical counterparts reported by Tamuri et al. (2012). The choice to use an alignment of mammalian mtDNA was largely motivated by the availability of these empirical distributions, which were derived from a concatenated alignment of 12 genes (3,598 codon sites) from 244 mammal species.

Results

Putative DT Mutations Are Detectable in a Real Alignment

Our objective was to use DT as a case study to test the hypothesis that an underspecified selection submodel combined with confounding can lead to false biological conclusions due to PL. The first step was to identify DT in a real alignment. To that purpose, models M0, M3, CLM3, and the new model RaMoSS, as well as their counterpart models that allow fixation of DT mutations, were fitted to the alignment of 20 mammalian mtDNA sequences. The tree with branch lengths obtained by ML under the best fitting model (RaMoSSwDT) is depicted in figure 1. Table 1 lists the log-likelihood (LL) and parameter estimates for each model. Table 2 shows the results for various model contrasts.

The log-likelihood ratios (LLR) were statistically significant for M0 versus M3, M3 versus CLM3, and CLM3 versus RaMoSS (table 2). Collectively, these contrasts provide evidence for variation in rate ratio across sites and branches, and support the existence of both static and temporally dynamic sites within the alignment. The four contrasts of the form model-M versus model-MwDT were also statistically significant, and therefore apparently detected fixation of DT mutations. However, the signal for DT became weaker with each increment in the complexity (i.e., number of parameters) of the selection submodel. The proportion of fixed mutations that were DT was inferred to be 23.6% under the simplest model contrast M0 versus M0wDT. Accounting for variations in rate ratio across sites (M3 vs. M3wDT) reduced this to 17.0%. By allowing sites to switch rate ratio (CLM3 vs. CLM3wDT), and allowing a mixture of static and switching sites (RaMoSS vs. RaMoSSwDT), the DT proportion was further reduced to 13.5% and 9.7%, respectively. Similarly, the $\mathrm{PRD}(\hat{\alpha}, \hat{\beta})$ decreased from 1.11% for the M0 versus M0wDT contrast to 0.36%, 0.14%, and 0.06% under M3 versus M3wDT, CLM3 versus CLM3wDT, and RaMoSS versus RaMoSSwDT, respectively.

Next, we conducted an investigation to determine which site patterns contributed the most to the 42-point difference in LL for RaMoSS ($\mathrm{LL} = -88{,}677$) as compared with RaMoSSwDT ($\mathrm{LL} = -88{,}635$). Of the 3,331 sites patterns, 83 were fixed, 25 had nonsynonymous differences only, 1,730 had synonymous differences only, and 1,493 were mixed with both synonymous and nonsynonymous differences. The contribution of each of these site-pattern categories to the total LL under RaMoSS and RaMoSSwDT is listed in table 3. RaMoSSwDT provided a slightly better fit to the 2.49% of site patterns that were fixed. This is because fixed sites become less likely as branch lengths increase, and RaMoSS produced larger branch lengths than RaMoSSwDT (supplementary fig. 1, Supplementary Material online). These sites accounted for only 3 the total difference of 42 LL points. Less than 1% of all sites patterns had nonsynonymous differences only. RaMoSSwDT fitted these sites slightly better, as expected given that allowing fixation of DT mutation increases the probability that a nonsynonymous substitution will occur. But since there were so few site patterns in this category, the total contribution was only 1 out of 42 LL points. Approximately 52% of site patterns had synonymous differences only. RaMoSSwDT provided a slightly worse fit to these sites. Most synonymous differences can be explained by a single nucleotide substitution at the third codon position. Allowing fixation of DT mutations, most of which are nonsynonymous, apparently reduces the probability of a site pattern with synonymous differences only. This effect was very

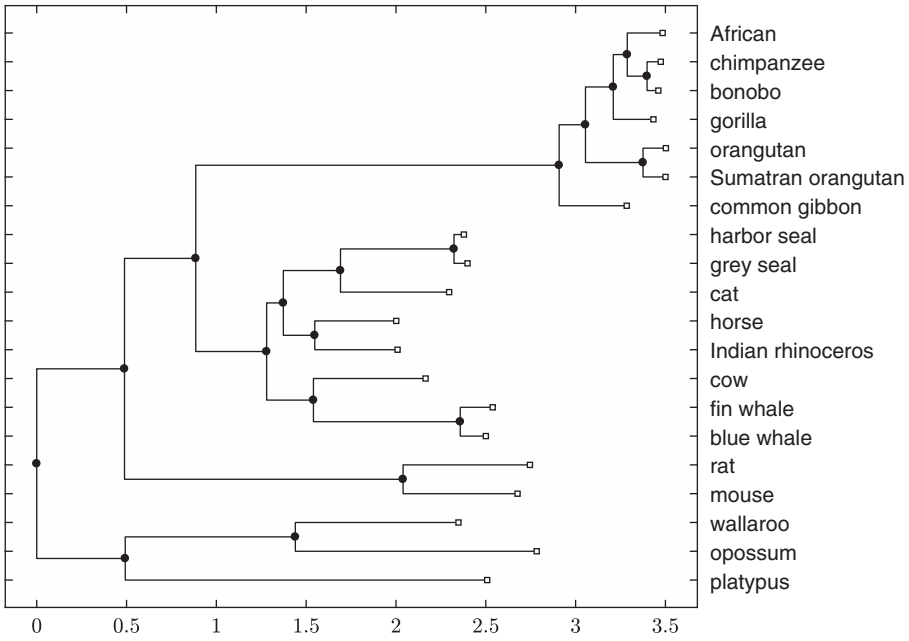

Fig. 1. The phylogeny for the concatenation of 12 H-strand mitochondrial DNA sequences (3,331 codon sites) from 20 mammalian species distributed by the PAML software package (Yang 2007). The topology is that reported in Cao et al. (1998). Branch lengths (expected number of single nucleotide substitutions per codon) were estimated using RaMoSSwDT (the best fitting of the models used in this study). The scale on the horizontal axis is the number of single nucleotide substitution per codon.

Table 1. Log-Likelihood (LL) and Parameter Estimates for Each Model Fitted to the Mammalian mtDNA Alignment Shown in figure 1.

Model	LL	Rate Ratios	Proportions	Switching	S, D, T
Ms	−26,752				
M0	−92,006	$\hat{\omega}_0 = 0.04$			
M3	−89,162	$(\hat{\omega}_0, \hat{\omega}_1) = (0.01, 0.15)$	$\hat{p}_0 = 0.71$		
CLM3	−88,880	$(\hat{\omega}_0, \hat{\omega}_1) = (0.00, 0.21)$	$\hat{p}_0 = 0.77$	$\hat{\delta} = 0.06$	
RaMoSS	−88,677	$(\hat{\omega}_0, \hat{\omega}_1) = (0.00, 0.08)$	$\hat{p}_{M3} = 0.73, \hat{p}_0 = 0.80$		
		$(\hat{\omega}'_0, \hat{\omega}'_1) = (0.01, 0.44)$	$\hat{p}'_0 = 0.66$	$\hat{\delta} = 0.21$	
M0wDT	−91,280	$\hat{\omega}_0 = 0.03$			76.4%, 21.5%, 2.1%
M3wDT	−88,930	$(\hat{\omega}_0, \hat{\omega}_1) = (0.01, 0.12)$	$\hat{p}_0 = 0.71$		83.0%, 16.7%, 0.3%
CLM3wDT	−88,786	$(\hat{\omega}_0, \hat{\omega}_1) = (0.00, 0.16)$	$\hat{p}_0 = 0.75$	$\hat{\delta} = 0.06$	86.5%, 13.5%, 0.0%
RaMoSSwDT	−88,635	$(\hat{\omega}_0, \hat{\omega}_1) = (0.00, 0.08)$	$\hat{p}_{M3} = 0.68, \hat{p}_0 = 0.80$		90.3%, 9.7%, 0.0%
		$(\hat{\omega}'_0, \hat{\omega}'_1) = (0.02, 0.34)$	$\hat{p}'_0 = 0.73$	$\hat{\delta} = 0.12$	

Table 2. Results for Model Contrasts Applied to the Mammalian mtDNA Alignment Shown in figure 1.

Contrast	Testing For. . .	LLR	Crit.val.	Detected	PRD
M0 vs. M3	var. across sites	5,668	5.99	Yes	4.36%
M3 vs. CLM3	var. across time	564	2.71	Yes	0.43%
CLM3 vs. RaMoSS	static and switching sites	406	9.49	Yes	0.31%
M0 vs. M0wDT	DT mutations	1,452	5.99	Yes	1.11%
M3 vs. M3wDT	DT mutations	464	5.99	Yes	0.36%
CLM3 vs. CLM3wDT	DT mutations	188	5.99	Yes	0.14%
RaMoSS vs. RaMoSSwDT	DT mutations	84	5.99	Yes	0.06%

small however, contributing a difference of only −1 LL points, despite the large number of site patterns in this category. Approximately 45% of sites had mixed site patterns, and these accounted for 39 out of the 42 LL points difference between RaMoSS and RaMoSSwDT. This demonstrates that mixed site patterns are more likely when the model permits fixation of DT mutations. Critically, mixed site patterns are also more likely to exhibit heterotachy (Jones et al. 2017). Of the 297 site patterns with a posterior probability of switching > 0.80 (computed using RaMoSSwDT), 289 had mixed site patterns. The remaining 8 were among the site patterns with nonsynonymous differences only. This suggests that the

Table 3. Each Row Reports the Number (and %) of Sites in the Corresponding Site-Pattern Category, the LL under RaMoSS and RaMoSSwDT, the Total Change in LL Associated with Each Category, and the Number of Sites for which the Posterior Probability of Switching was > 0.80.

Site-Pattern Cat.	Number (%)	RaMoSS LL	RaMoSSwDT LL	ΔLL	Post. > 0.80
fixed	83 (2.49)	−519	−516	3	0
Nonsynonymous	25 (0.75)	−497	−496	1	8
Synonymous	1,730 (51.94)	−34,214	−34,215	−1	0
Mixed	1,493 (44.82)	−53,447	−53,408	39	289
Total	3,331	−88,677	−88,635	42	297

process whereby rare episodic DT mutations are fixed can be confounded with the process of shifting balance (Jones et al. 2017), since both can produce signatures consistent with heterotachy.

A heuristic method for inferring DT mutations is to examine sites occupied by serine only (Averof et al. 2000). Codon aliases for serine include TCN, where N is any nucleotide, and AGY, where Y is a pyrimidine. Minimum paths between TCN and AGY by single nucleotide steps require substitution to cystine or threonine. But these amino acids are physicochemically different than serine, and can be assumed to be less fit than serine at a site observed to be occupied by serine only. The existence of serine sites with a mix of TCN and AGY would therefore suggest that some double mutations of the form TC ↔ AG were fixed. However, of the 112 serine sites in the real mtDNA alignment, one site was occupied by a single alias for serine, 19 were occupied by a combination of AGT and AGC, and 92 were occupied by a combination of TCC, TCT, TCA, and TCG. Aliases from the AGY and TCN groups did not appear together at any site. This result, combined with the observed decrease in the strength of the evidence for DT with each incremental increase in the complexity of the selection submodel, casts doubt on the veracity of the detection of fixed DT mutations in the real mtDNA under RaMoSS versus RaMoSSwDT. Simulation studies were therefore conducted to investigate the possibility of false detection of DT, as reported in the next section of this article.

The Extent to Which DT Parameters Carry PL Is Related to Model Misspecification

There is substantial heterogeneity in selection pressure across sites within the mammalian mitochondrial genome (Garvin et al. 2015). It is therefore likely that the single rate ratio of M0 provides a highly inadequate summary of variations due to selection effects in the real mtDNA alignment. Our analysis resulted in a substantial PRD for the M0 versus M0wDT contrast (1.11% PRD, corresponding to a highly significant LLR of 1,452) and a relatively large estimated proportion of fixed mutations that were DT (23.6%). If those estimates were influenced by PL, we would expect a reduction in both with an increase in the complexity of the selection submodel. This is exactly what was observed. RaMoSS versus RaMoSSwDT resulted in a much smaller PRD (only 0.06%) and indicated that a smaller proportion of fixed mutations were DT (9.7%). However, even the selection submodel under RaMoSS is likely to be underspecified compared with the actual data-generating process.

Three simulations studies were conducted to assess the relationship between model misspecification and the PL detected by the four M versus MwDT contrasts. Each study was conducted using a different alignment generating model, all of which did not include fixation of DT mutations. In the first simulation study, alignments were generated under RaMoSS. This study included the scenario for which the selection submodel of the RaMoSS versus RaMoSSwDT contrast was not misspecified, although the selection submodel of the other three model contrasts were underspecified to some extent. In the second simulation study, alignments were generated under a substantially more complex CSM, having an independent rate ratio for each site. Alignments were therefore generated with more variation in rate ratio across sites than accounted for by any of the M versus MwDT contrasts. In the third simulation study, alignments were generated under MutSel-mmtDNA to have variation across sites and over time comparable (as will be shown) to the real mtDNA alignment. We demonstrate that, although the pathology we call PL was readily identified in all simulation studies under the contrast with the simplest selection submodel (M0 vs. M0wDT), it was only detected under the most complex contrast (RaMoSS vs. RaMoSSwDT) in alignments generated using the more realistic MutSel-mmtDNA generating model.

Simulation Study 1: MLEs for the DT Process Carry Substantial PL When the Selection Submodel Is Underspecified, but False Conclusions Are Avoided When the Selection Submodel Is Correctly Specified

In the first simulation study, one-hundred 300-codon alignments were generated on the tree depicted in figure 1 using RaMoSS as the generating model. A starting sequence was constructed by selecting codons in proportion to their empirical frequencies estimated from the real mtDNA. All alignments were generating starting with this same sequence. Parameters for the selection submodel (including $\omega_0, \omega_1, p_0, \omega_0', \omega_1', p_0', p_{M3}$ and δ) were set to values estimated from the real mtDNA alignment using RaMoSSwDT (i.e., the best fitting model; see table 1 for parameter values). Similarly, parameters for the DNA submodel, including κ and position-specific nucleotide frequencies, were set to values estimated from the real alignment, except that α and β (i.e., the parameters that specify the rate of double and triple mutations, see Materials and Methods) were set to 0 to exclude fixation of DT mutations. Table 4 shows median results for the various likelihood ratios tests

Table 4. Median Values for Log-Likelihood Ratios (LLR) and the Number of Times DT was Detected from 100 Alignments Generated under (RaMoSS, M3($k = n$), MutSel-mmtDNA) with $\alpha = \beta = 0$.

Contrast	Trials	Testing For...	Median LLR	Crit.val.	Detected
M0 vs. M3	100	var. across sites	(171, 867, 767)	5.99	(98, 100, 100)
M3 vs. CLM3	100	var. across time	(34.0, 5.88, 30.6)	2.71	(99, 75, 100)
CLM3 vs. RaMoSS	100	static and switching sites	(19.6, 62.9, 57.1)	9.49	(95, 98 100)
M0 vs. M0wDT	100	DT mutations	(29.7, 63.3, 147)	5.99	(99, 100, 100)
M3 vs. M3wDT	100	DT mutations	(5.69, 1.03, 25.7)	5.99	(49, 10, 97)
CLM3 vs. CLM3wDT	100	DT mutations	(0.01, 0.20, 12.3)	5.99	(3, 3, 76)
RaMoSS vs. RaMoSSwDT	100	DT mutations	(0.00, 0.01, 4.34)	5.99	(0, 5, 41)

(see supplementary table 1, Supplementary Material online, for median parameter estimates).

The contrast with the simplest selection submodel (M0 vs. M0wDT) incorrectly rejected the null hypothesis and inferred DT mutations in almost all trials (99/100). Improving the selection submodel by accounting for variations in rate ratio across sites (M3 vs. M3wDT) yielded a substantial reduction in the false positive rate (49/100). Accounting for heterotachy (CLM3 vs. CLM3wDT and RaMoSS vs. RaMoSSwDT) further reduced the number of false positives to 3/100 and 0/100, respectively. RaMoSS provided the best fit in all trials, and produced median parameter estimates similar to their generating values: $\hat{\omega}_0 = 0.00 (\text{generating } \omega_0 = 0.00), \hat{\omega}_1 = 0.03 (0.08), \hat{p}_0 = 0.86 (0.80), \hat{\omega}'_0 = 0.01 (0.01), \hat{\omega}'_1 = 0.44 (0.44), \hat{p}'_0 = 0.88\ (0.66), \hat{p}_{M3} = 0.72 (0.73)$ and $\hat{\delta} = 0.17 (0.21)$. It was no surprise to find that RaMoSS produced reliable parameter estimates, and that RaMoSS versus RaMoSSwDT did not falsely detect the fixation of DT mutations, since RaMoSS was an exact match to the generating process. However, it was interesting to find that the MLEs $\hat{\alpha}$ and $\hat{\beta}$ for the DT process carry substantial PL when the selection submodel is underspecified (as indicated by the high false detection rate). In particular, DT was only detected by the two models that did not account for heterotachy (M0 and M3). This demonstrates that random variations in site-specific rate ratios, produced in this simulation study by the CLM3 component of the generating model, can create false signal for the episodic fixation of DT mutations when heterotachy is not accounted for by the selection submodel.

Simulation Study 2: Improving the Selection Submodel Reduces PL Even When the Submodel Is Substantially Underspecified

In the second simulation study, one-hundred 300-codon alignments were generated on the tree depicted in figure 1 using what we call M3($k = n$) as the generating model (where n is the number of codon sites). The objective was to produce the same level of variation in rate ratio across sites as in the real mtDNA, but without heterotachy, using the following procedure. First, a vector of codon fitness coefficients $\mathbf{f}^h$ was drawn for each site using the MutSel-mmtDNA model (see Supplementary Material online). The MutSel rate matrix A^h was then constructed with $\alpha = \beta = 0$ and used to determine the expected rate ratio for the site:

$$\omega^h = \frac{\sum_{(i,j)} \pi_i^h A_{ij}^h \ell_{ij}}{\sum_{(i,j)} \pi_i^h M_{ij} \ell_{ij}} \quad (8)$$

where ℓ_{ij} is an indicator for nonsynonymous codon pairs (i, j), the π_i^h are site-specific stationary frequencies for the 60 codons, and M_{ij} is the rate of mutation from i to j (Spielman and Wilke 2015b; Jones et al. 2017). The rate matrix Q^h for the site-specific generating model was then constructed using equation (10) in Materials and Methods. Note that the rate ratio at a site evolving under Q^h (e.g., under M0 with rate ratio ω^h) is always ω^h regardless of the incumbent codon. An alignment generated using the set of Q^h will therefore have the same level of variation in the expected rate ratio across sites as an alignment generated using the A^h (e.g., using MutSel-mmtDNA), but without heterotachy. Furthermore, all of the Q^h share the same vector of stationary frequencies (whereas each A^h generated under MutSel-mmtDNA has its own vector of site-specific frequencies).

Table 4 shows median results for the various likelihood ratios tests (see supplementary table 2, Supplementary Material online, for median parameter estimates). As expected, the M0 versus M3 contrast detected substantial signal for variations in rate ratio across sites in all trials. Quite unexpected was the result that the M3 versus CLM3 contrast implied signal for heterotachy in 75/100 trials. This is in apprent contradiction to the design of the generating process, which precluded heterotachy. However, the signal for changes in rate ratio over time was relatively weak: the median switching rate was only $\hat{\delta} = 0.02$ or one switch per 50 single nucleotide substitutions. Furthermore, the median LLR for M3 versus CLM3 was only 5.88 (compared with the critical value of 2.71 for a 5% test) with a corresponding P value of 0.008. Given that CLM3 is equivalent to M3 when $\delta = 0$, these results are not entirely inconsistent with sites evolving under fixed rate ratios. Nevertheless, they seem to indicate that $\hat{\delta}$ carried some PL in three-quarters of the trials. The CLM3 versus RaMoSS contrast similarly implied a fraction of sites with signal for heterotachy. The LLR for this contrast was significant in 98/100 trials (median LLR = 62.9), but with a very small switching rate ($\hat{\delta} = 0.00$). RaMoSS is the same as M3($k = 4$) when $\delta = 0$, so in this case it seems that RaMoSS provided the better fit not because of PL on $\hat{\delta}$, but because four ω-categories provided a significantly better fit than two, apparently reflecting the generated level of variation in ω across sites.

Turning to the tests for fixation of DT mutations, the contrast involving the simplest selection submodel (M0 vs. M0wDT) incorrectly inferred DT mutations in all 100 trials (table 4). Again, improving the selection submodel substantially reduced the false positive rate. Even limited accommodation of variations in rate ratio across sites using M3 (e.g., with only two rate-ratio categories) reduced the false positive rate to 10/100. This was reduced further to only 3/100 and 5/100 by CLM3 versus CLM3wDT and RaMoSS versus RaMoSSwDT. These rates are consistent with the 5% level of significance of the likelihood ratio test, and seem to imply that both CLM3 versus CLM3wDT and RaMoSS versus RaMoSSwDT will reliably fail to detect the fixation of DT mutations when they do not occur. However, the generating model M3($k = n$) is unrealistic, and in particular does not simulate heterotachy or variations in site-specific amino acid propensities. A more rigorous test of the reliability of the RaMoSS versus RaMoSSwDT contrast for detecting DT requires use of a more realistic alignment-generating process.

Simulation Study 3: RaMoSS versus RaMoSSwDT Is Unreliable When Fitted to Data Generated Using MutSel-mmtDNA

The M3($k = n$) generating model reflects the traditional approach of testing the impact of model misspecification by simulating alignments using a more complex CSM. However, the absence of heterotachy and site-specific stationary frequencies means that the simulated distribution of site patterns can only be unrealistic compared with the real mtDNA alignment. In the third simulation study, one-hundred 300-codon alignments were generated on the tree depicted in figure 1 using the generating process we call MutSel-mmtDNA, which was formulated to produce alignments that match the real mtDNA alignment as closely as possible. In this section, we report the results of the model fits; the results of comparisons between alignments generated using MutSel-mmtDNA and the real alignment are reported in the next section. Comparison with table 1 shows that median parameter estimates (reported in supplementary table 3, Supplementary Material online) under RaMoSS were similar to those estimated from the real mtDNA alignment using the same model (we use RaMoSS rather than RaMoSSwDT for this comparison because the alignments were simulated without DT substitutions). The median values under RaMoSS were: $\hat{\omega}_0 = 0.00$ (compared to $\hat{\omega}_0 = 0.00$ for the real mtDNA), $\hat{\omega}_1 = 0.12$ (0.08), $\hat{p}_0 = 0.82$ (0.80), $\hat{\omega}'_0 = 0.00$ (0.01), $\hat{\omega}'_1 = 0.56$ (0.44), $\hat{p}'_0 = 0.60$ (0.66), $\hat{p}_{M3} = 0.80$ (0.73) and $\hat{\delta} = 0.20$ (0.21). These results suggest a substantial degree of "phenomenological similarity" between the real and simulated alignments. Note that this was not by design, since the MLEs derived from the real mtDNA alignment were not used in the formulation of MutSel-mmtDNA; the similarity was a consequence of the method used to generate site-specific fitness coefficients (see Supplementary Material online).

The impact of PL when the models were fitted to alignments generated under MutSel-mmtDNA is apparent in table 4. The contrast involving the simplest selection submodel (M0 vs. M0wDT) incorrectly inferred DT in all 100 trials, as might be expected given previous results. However, unlike the previous two simulation studies, accounting for variations in rate ratio across sites (M3 vs. M3wDT) had negligible impact on the false positive rate (97/100). Although accounting for heterotachy (CLM3 vs. CLM3wDT and RaMoSS vs. RaMoSSwDT) reduced the number of false positives (to 76/100 and 41/100, respectively), the lowest rate was still too large given the 5% level of significance of the test. We conclude that the selection submodel for RaMoSS is underspecified with respect to MutSel-mmtDNA, with the result that substantial PL was conferred onto $\hat{\alpha}$ and $\hat{\beta}$ in a large number of trials.

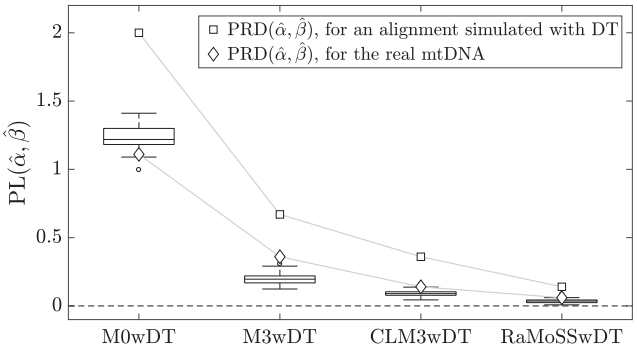

FIG. 2. Boxplots show the distribution of PL($\hat{\alpha}, \hat{\beta}$) for each of the M versus MwDT model contrasts fitted to 50 full-scale alignments (20 taxa, 3,331 codon sites) generated under MutSel-mmtDNA with $\alpha = \beta = 0$. Diamonds show PRD($\hat{\alpha}, \hat{\beta}$) for each contrast fitted to the real mtDNA. Squares show the PRD($\hat{\alpha}, \hat{\beta}$) for each contrast fitted to a full-scale alignment generated under MutSel-mmtDNA with α and β set to values estimated from the real mtDNA using RaMoSSwDT. Circles indicate outliers in PL($\hat{\alpha}, \hat{\beta}$) for the corresponding boxplot.

It now seems plausible that the detection of DT in the real mtDNA was a false positive due to PL. If we can assume that MutSel-mmtDNA produces alignments consistent with the real data, then it can be used to estimate the distribution of PL($\hat{\alpha}, \hat{\beta}$) for each of the M versus MwDT model contrasts. To this end, MutSel-mmtDNA was used to generate 50 full-scale alignments, each with 3,331 codon sites, without fixation of DT mutations. Each model contrast was fitted to produce distributions of PRD($\hat{\alpha}, \hat{\beta}$). Because α and β were set to 0 in the generating process, we can equate PRD to PL. The resulting distributions are shown as boxplots in figure 2, where the previously described decline in the PRD($\hat{\alpha}, \hat{\beta}$) obtained by fitting the contrasts to the real mtDNA (last column of table 2) is reflected by a similar decline in the median PL($\hat{\alpha}, \hat{\beta}$) with each incremental increase in the complexity of the selection submodel.

The diamond in each boxplot of figure 2 marks the PRD($\hat{\alpha}, \hat{\beta}$) for the corresponding contrast fitted to the real mtDNA alignment. This value falls just within the upper tail of the estimated distribution of PL($\hat{\alpha}, \hat{\beta}$) for the RaMoSS versus RaMoSSwDT contrast. For comparison, a single full-sized alignment was generated using MutSel-mmtDNA with α and β set to the values estimated by RaMoSSwDT fitted to

the real mtDNA (e.g., with 9.7% double and 0.0% triple mutations, see table 1). The small square in each boxplot marks the PRD$(\hat{\alpha}, \hat{\beta})$ obtained by fitting each contrast to this alignment. As the signal for DT mutations was real in this case, PRD$(\hat{\alpha}, \hat{\beta})$ cannot be not equated to PL$(\hat{\alpha}, \hat{\beta})$, but can be interpreted as an indication of real signatures for DT, possibly combined with random error and PL. The decrease in PRD with each increase in the complexity of the selection submodel is still evident, and suggests that $\hat{\alpha}$ and $\hat{\beta}$ carry some PL. These comparisons, combined with the large number of false detections reported in table 4, suggest that the detection of fixations of DT mutations in the real mtDNA might have been a false positive.

Alignments Generated under MutSel-mmtDNA Are Realistic by Several Measures of Comparison

The design of the third simulation study represents a substantial departure from the first two. We maintain that the role PL might have played in the analysis of the real mtDNA can be assessed only insofar as simulated alignments match real data. Hence, rather than using an M-series CSM as the generating process, we used the more realistic mutation-selection (MutSel) framework of Halpern and Bruno (1998). Under MutSel, each site can be assigned its own vector of fitness coefficients $\mathbf{f}^h$. This determines the stringency of selection (the average rate ratio) and temporal dynamics (heterotachy) at the site. Our objective was to simulate alignments with heterogeneity in rate ratio across sites and time, and in site patterns, consistent with the real mammalian mtDNA. Here, we report results that show that the method we developed to achieve this (MutSel-mmtDNA) can produce alignments similar to the real mtDNA alignment by several measures of comparison.

Empirical distributions of scaled selection coefficients for all mutations, all substitutions, all nonsynonymous mutations and all nonsyonymous substitutions derived from mammalian mtDNA have already been published (Tamuri et al. 2012). We therefore adjusted our formulation of MutSel-mmtDNA to make the estimated probability density functions (PDFs) of our scaled selection coefficients $s_{ij}^h = 2N_e(f_j^h - f_i^h)$ match those as closely as possible. The predicted distributions derived from 10^5 sites simulated under the resulting MutSel-mmtDNA model were similar in shape to their empirical counterparts (cf. supplementary fig. 2, Supplementary Material online vs. figure 2 in Tamuri et al. 2012) and had similar probabilities $p(s_{ij} < -2)$, $p(-2 < s_{ij} < 2)$ and $p(s_{ij} > 2)$ (table 5). Further comparisons between MutSel-mmtDNA and the real mtDNA were based on a full-sized simulated alignment of 3,331 codon sites. Amino acid frequencies for the simulated alignment were highly correlated with those in the real data (correlation = 0.91, P value $<<$ 0.001, supplementary fig. 3, Supplementary Material online), as were the codon frequencies (correlation = 0.83, P value $<<$ 0.001). The frequencies with which each pair of amino acids was observed within a given site pattern were found to be strongly concordant (correlation = 0.91, P value $<<$ 0.001, supplementary fig. 4, Supplementary Material online). The distributions of the number of amino acids realized at each site were also very similar (supplementary fig. 5, Supplementary Material online). And the simulated alignment had a similar number of fixed, nonsynonymous, synonymous, and mixed site patterns compared with the real data: (87, 18, 2,052, 1,174) in the simulated alignment versus (83, 25, 1,730, 1,493) as reported in table 3.

Table 5. Comparison of Interval Probabilities for Scaled Selection Coefficients s_{ij} under the Generating Model MutSel-mmtDNA versus Those Derived Empirically by Tamuri et al. (2012).

	$p(s_{ij} < -2)$	$p(-2 < s_{ij} < 2)$	$p(s_{ij} > 2)$
All mutations	**0.61**	**0.39**	**0.00**
Tamuri et al. (2012)	**0.65**	**0.34**	**0.01**
Nonsyn. mutations	**0.90**	**0.09**	**0.01**
Tamuri et al. (2012)	**0.89**	**0.10**	**0.01**
All substitutions	**0.03**	**0.94**	**0.03**
Tamuri et al. (2012)	**0.03**	**0.94**	**0.03**
Nonsyn. substitutions	**0.18**	**0.64**	**0.18**
Tamuri et al. (2012)	**0.14**	**0.72**	**0.14**

Evidence of Confounding

Our simulation studies demonstrate that PL$(\hat{\alpha}, \hat{\beta})$ is related to the degree to which the selection submodel is underspecified with respect to the data-generating process. As was stated in the introduction, misspecification alone is insufficient to produce PL. There must also be some measure of confounding between the processes governed by the mechanistic parameters in the DNA submodel with processes that generate variations in selection effects. To further illustrate this issue, we examined the effects of changes in κ and α (both of which are parameters of the DNA submodel) on the expected number of nonsynonymous substitutions per unit branch length E(rN) and the predicted switching rate δ (measures that reflect variations in selection effects). Specifically, we examined the changes in E(rN) and δ when κ was increased from 1 to 10 with $\alpha = \beta = 0$, and changes in the same when α was increased from 0.015 (corresponding to 2.5% double mutations) to 0.075 (11% double mutations) with κ fixed at 4 and β fixed at 0.

Vectors of site-specific fitness coefficients were first generated using MutSel-mmtDNA with $\alpha = \beta = 0$. Each was used to compute site-specific values for E(rN) and δ, once with $\kappa = 1$ and again with $\kappa = 10$, using previously published methods (Jones et al. 2017). The resulting distributions for the change in E(rN) and δ [ΔE (rN) and $\Delta\delta$, respectively] with κ were both roughly symmetric and centered at 0 (fig. 3A and B). Hence, the same change in κ sometimes increased, and sometime decreased, both E(rN) and δ. The net effect of changes in κ on these two quantities is therefore negligible when averaged across sites. This result explains why κ carried little or no PL in a subsequent simulation study under which 100 alignments generated using MutSel-mmtDNA with $\kappa = 1$ were fitted to both M0($\kappa = 1$) (i.e., M0 with κ fixed at 1) and M0 (under which κ is estimated): the M0($\kappa = 1$) versus M0 contrast reliably failed to reject the null hypothesis of no transition bias in all trials, despite the fact that the submodel for selection under M0 is highly underspecified with respect to MutSel-mmtDNA. In the

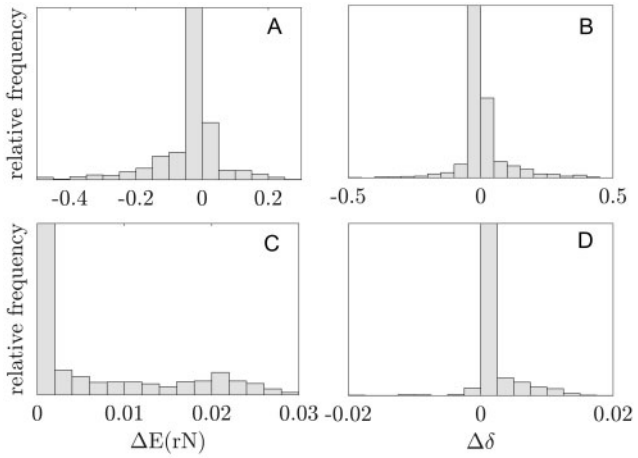

FIG. 3. Distributions of the change in the expected number of nonsynonymous substitutions per unit branch length [ΔE (*rN*)] and the expected switching rate ($\Delta\delta$) for 1,000 sites with fitness coefficients generated using MutSel-mmtDNA. (*A*) ΔE (*rN*) when κ is increased from 1 to 10 with $\alpha = \beta = 0$, (*B*) $\Delta\delta$ when κ is increased from 1 to 10 with $\alpha = \beta = 0$, (*C*) ΔE (*rN*) when α is increased from 0.015 to 0.075 with $\kappa = 4$ and $\beta = 0$, (*D*) $\Delta\delta$ when α is increased from 0.015 to 0.075 with $\kappa = 4$ and $\beta = 0$.

second analysis, vectors of fitness coefficients were generated under MutSel-mmtDNA with $\kappa = 4$ and $\beta = 0$ (to prohibit fixation of triple mutations). Each vector was used to compute site-specific values for $\mathrm{E}(rN)$ and δ, with α set to either 0.015 (corresponding to 2.5% double mutations) or 0.075 (11% double mutations). The distributions for ΔE (*rN*) and $\Delta\delta$ indicated that these values are almost always nonnegative (fig. 3*C* and *D*). Hence, an increase in α generally results in an increase in the expected nonsynonymous substitution rate and an increase in the level of heterotachy when its effects are averaged across sites. These simulations support the view that the process of episodic fixation of DT mutations can be confounded with selection effects, and show how the potential for a parameter to take on PL might be assessed using the MutSel framework.

Assessing PL in Other CSMs

The utility of the PL framework for assessing the validity of the interpretation of model parameters in other CSMs is illustrated in this section by applying our methods to two other inferential scenarios. The first is a test for changes in selection intensity in one clade compared with the remainder of the tree (RELAX, Wertheim et al. 2014). Under the RELAX model, it is assumed that each site evolved with a rate ratio randomly drawn from $\boldsymbol{\omega}_\mathrm{R} = \{\omega_1, \ldots, \omega_k\}$ on a set of prespecified reference branches, and from a modified set of rate ratios $\boldsymbol{\omega}_\mathrm{T} = \{\omega_1^m, \ldots, \omega_k^m\}$ on test branches, where m is an exponent. A value $0 < m < 1$ moves the rate ratios in $\boldsymbol{\omega}_\mathrm{T}$ closer to one compared with their corresponding values in $\boldsymbol{\omega}_\mathrm{R}$, consistent with relaxation of selection pressure at all sites on the test branches. Relaxation is indicated when the contrast of the null hypothesis that $m = 1$ versus the alternative that $m < 1$ is statistically significant. RELAX was fitted to the real mtDNA with three ω-categories using the HyPhy software

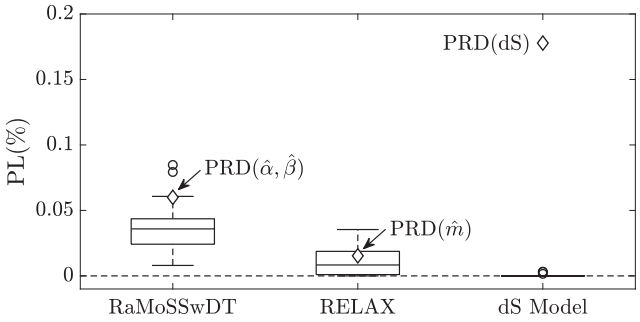

FIG. 4. Boxplots show the distributions of PL for parameters in models fitted to the same 50 full-scale alignments generated under MutSel-mmtDNA (20 taxa, 3,331 codon sites). Diamonds show PRD for each contrast fitted to the real mmtDNA. Circles indicate outliers. PL was statistically significant in 48/50 trials under RaMoSS versus RaMoSSwDT contrast, in 31/50 trials under RELAX, and in 0/50 trials under M3($k = 4$) versus M3wdS($k = 4$).

package (Kosakovsky Pond et al. 2005). Test branches were set to all of those in the primate clade, including the long branch leading to that clade (see fig. 1). The test revealed significant evidence for relaxation of selection pressure among the branches in the primate clade ($\hat{m} = 0.81$, LLR = 18, *P* value = 2.2×10^{-5}, $\mathrm{PRD}(\hat{m}) = 0.015\%$). The model was also fitted to the 50 full-scale alignments generated using MutSel-mmtDNA, under which no relaxation occurred. The null was falsely rejected in 31/50 trials. Furthermore, $\mathrm{PRD}(\hat{m})$ estimated from the real alignment fell well within the distribution of $\mathrm{PL}(\hat{m})$ from the 50 simulated alignments (fig. 4). These results suggest that PL provides a plausible explanation for the detection of relaxation in selection pressure in the primate clade of the real mtDNA.

The vast majority of CSMs assume that the synonymous substitution rate is constant across sites, despite evidence that dS can vary (particularly in mitochondrial DNA, e.g., Bielawski and Gold 2002). The second scenario is a test for variation in dS across sites (Kosakovsky Pond and Muse 2005). This test has no moniker that we are aware of, so it will be designated here as M3wdS (M3 with changes in dS) due to its similarity to the M-series model M3 (Yang et al. 2000). Under M3wdS, it is assumed that there are *k* dS categories and *k* dN categories that combine to produce k^2 ω-categories. M3wdS(*k*) is contrasted with the null model M3(*k*) that assumes dS is constant across sites. Rejection of the null is interpreted as evidence for variations in dS across sites. M3(*k*) was first fitted to the real mtDNA using HyPhy with $k \in \{3, 4, 5\}$. It was found that four categories were sufficient to account for all of the variation in rate ratio across sites (i.e., four categories fit the alignment better than three and just as well as five). M3($k = 4$) was then contrasted with M3wdS($k = 4$) using HyPhy. The contrast was found to be significant (LLR = 252, $\mathrm{PRD}(dS) = 0.19\%$). The M3($k = 4$) versus M3wdS($k = 4$) contrast was then fitted to the 50 full-scale alignments generated using MutSel-mmtDNA. The null was rejected in 0/50 trials. Furthermore, the phenomenological load $\mathrm{PL}(dS)$ associated with the parameters for dS variation was very close to 0 in all 50 trials (fig. 4). These results

support the interpretation of M3wdS as detecting genuine variations in dS in the real mtDNA alignment. Interestingly, there is biochemical support for the notion of spatial variation in dS within the mitochondrial genome: due to the different amount of time that mtDNA spends in the single-strand state during its replication process (Clayton 1982), it will be subject to different probabilities of spontaneous mutational damage (Tanaka and Ozawa 1994), which is expected to lead to different synonymous substitution rates (Reyes et al. 1998; Bielawski and Gold 2002; Raina et al. 2005).

Discussion

Codon substitution models have evolved toward ever increasing complexity since their introduction by Muse and Gaut (1994) and Goldman and Yang (1994), motivated in part by the rapid increase in the quantity of information available. With greater information comes greater opportunity to tease out the effects of subtle processes. This can be achieved by adding parameters for such processes to a standard CSM. Or so it would seem. Sites for which mutation and selection are in balance can exhibit signatures consistent with random changes in site-specific rate ratios or heterotachy (e.g., mixed site patterns, table 3) cause by shifting balance. But signatures of heterotachy can also be produced by the occasional fixation of double or triple mutations. Hence, shifting balance and DT are confounded processes. Consequently, if a CSM accounts for DT in its DNA submodel but fails to account for shifting balance in its selection submodel, the rate parameters (α, β) will be forced to account for signatures of heterotachy alone. Our analyses demonstrate that this can result in false inference for DT, and provides an example of a general principle, namely that it is possible for a parameter meant to support a specific mechanistic interpretation to be inferred to be statistically significant even if the process it represents did not occur. When this happens, we say that the parameter's MLE carries PL.

The basal cause of PL was shown to be confounding. We said in the introduction that two processes are confounded if they produce a common signature in the data. This implies that the generating process is the ultimate source of confounding. But whether or not counfounding manifests depends on the relationship between the fitted model and the data. This is in part because different CSMs are sensitive to different signatures. For example, the existence of multiple processes that generate heterotachy has little or no impact on the estimation of the parameters in M0 because this model ignores temporal dynamics. Our analysis suggests that RaMoSSwDT overestimated the proportion of fixed mutations that were DT ($\approx$ 10%), and that this occurred because the parameter for site-specific shifts in rate ratio (δ), and the parameters for the rates of double (α) and triple (β) mutation, are all sensitive to the same signatures in the alignment, those consistent with heterotachy. By contrast, in an analysis of a similar although larger set of mammalian mtDNA (244 taxa with 3,598 codon sites), Tamuri et al. (2012) inferred DT rates an order of magnitude smaller ($\sim$1% DT). Their model (swMutSel) utilizes signatures in the alignment that RaMoSSwDT is insensitive to in the form of empirical site-specific codon frequencies. The temporal dynamic at a site is to a large degree characterized by its site-specific frequencies (Jones et al. 2017), so their inclusion in swMutSel likely captured some variation in selection effects due to shifting balance. This apparently faciliated the detection of distinct signatures for DT (if DT was real), or else reduced the PL carried by DT parameters (if it was not). Hence, the degree to which confounding impacts inference is dependent on the signatures present in the data that the model is sensitive to, or in other words on the relationship between model and data.

It can happen that two processes produce signatures that differ only slightly and in such a way that they are confounded under a given CMS when information is sparse, but readily disentangled when information is rich. Such a scenario might not be uncommon, particularly among mixture models (Mingrone et al. 2018), but is an issue only if the amount of data required to ameliorate associated pathologies (e.g., false positives due to PL) is prohibitively large. Under this scenario, we say that the processes are only nearly confounded. In our analyses, by contast, the reduction in deviance engendered by the inclusion of the parameters for DT was associated with mixed site patterns in the real mtDNA alignment. A larger taxonomic sample or the addition of more genes to the concatenation would result in more mixed site patterns, and would presumably increase the probablity of falsely inferring DT. This is supported by our observation that the false positive rate for DT under RaMoSS versus RaMoSSwDT increased from 41% (41/100) to 96% (48/50) among alignments generated using MutSel-mmtDNA when the number of sites was increased from 300 to 3,331 (fig. 4). We therefore maintain that the false detection of fixation of DT mutations by RaMoSS versus RaMoSSwDT was not driven by lack of information, but by an abundance of information (cf. Kumar et al. 2012). Under the scenario where more site patterns (or more taxa) only worsen PL, we say that the two processes are perfectly confounded. To be clear, the introduction of information of a different type into the analysis, such as site-specific codon frequencies in the case of swMutSel, can potentially allay pathologies associated with perfect confounding (see previous paragraph).

It would be helpful to have a means to assess in advance whether confounding might impact inference under a given CSM. This was attempted in the section "Evidence of Confounding," where it was shown that an increase in κ (a property of the mutation process) is not correlated with an increase in the rate of fixation of nonsynonymous mutations (a property of the substitution process), but that an increase in the double mutation rate is. Such a result is intuitive, since the fixation of a double mutation at a site along a short branch [e.g., TTA(L) $\rightarrow$ GCA(A)] can be consistent with the fixation of two single mutations in rapid succession [e.g., TTA(L) $\rightarrow$ GTA(V) $\rightarrow$ GCA(A)], and therefore manifest as a transient elevation in the nonsynonymous to synonymous rate ratio at that site under a model that does not allow DT. Indeed, such intuition might have been used to predict the possibility of confounding between episodic elevations in dN/dS and episodic fixation of DT mutations. However, our

analysis was based on predictions derived from a mechanistic model, not by fitting a CSM to data. Given our supposition that the impact of confounding on inference depends on the relationship between a CSM and the actual data it is to be fitted to, it would seem that the only currently available method to identify PL is a case-by-case application of the approach illustrated in figure 4. To reiterate: suppose a mechanistic parameter ψ were introduced into a substitution model M to give the model M_ψ. Further suppose that the M versus M_ψ contrast indicated a significant reduction in deviance, $PRD(\hat{\psi})$, when fitted to a real alignment. To determine whether the cause of the balance of this reduction was real signal or PL, one can first generate alignments in such a way as to resemble the real alignment as closely as possible, but without the mechanistic process represented by ψ. These would be fitted to M versus M_ψ to produce a null distribution for $PL(\hat{\psi})$. Confounding would be inferred to have influenced the analysis when the $PRD(\hat{\psi})$ computed from the real alignment is no greater than that 95% percentile of the $PL(\hat{\psi})$ distribution (cf. fig. 4). This approach requires a method mimic the real alignment. The Pyvolve software package (Spielman and Wilke 2015a) provides a way to generating alignments consistent with a large real alignment. The methods used in this article (described in Supplementary Material online) provide an alternative approach for smaller alignments.

Our results have implications about how the performance of a CSM should be assesesed. Early efforts to test the reliability of CSMs made use of the comparatively simplistic generating models available at the time under the assumption that the findings of such analyses would be applicable to real alignments (Anisimova et al. 2001, 2002; Wong et al. 2004; Zhang 2004; Yang et al. 2005; Zhang et al. 2005; Yang and dos Reis 2011; Lu and Guindon 2014). Implicit in this methodology is the presupposition that the reliability of a CSM has little to do with the data. In its original instantiation, for example, the Yang–Nielsen Branch-Site Model (YN-BSM, Yang and Nielsen 2002) was evaluated using real data only. It was later shown via simulation that the original YN-BSM is prone to falsely infer positive selection under certain testing scenarios (Zhang 2004). A modified verison of the model was subsequently shown to be reliable under the same scenarios (Zhang et al. 2005). Hence the problem was implicitly assumed to be with the model, with little consideration of the role the data might have played in the observed pathology. The problem with this approach is that it leaves open the possiblity that the modified YN-BSM might still be unreliable when fitted to alignments simulated using an alternative, more realistic, generating scenario.

This possibility was illustrated by our three simulation studies. In the first study, alignments were generated using RaMoSS to assess the reliability of the RaMoSS versus RaMoSSwDT contrast in the absence of any model misspecification. The fixation of double and triple mutations was not inferred in any of the 100 simulated alignments (bottom row of table 4). To assess performance in the presence of some misspecification, alignments in the second simulation study were generated using M3($k = n$), a model that typifies traditional methods to assess model reliability. The false postive rate for DT under the RaMoSS versus RaMoSSwDT contrast was only 5/100. In the past might, this result might have been sufficient to conclude that the contrast is a reliable instrument with which to detect signatures of DT in real data, similar to the conclusion implicit in Zhang et al. (2005) about the modified YN-BSM. However, alignments in the third simulation study were generated to have variations in selection effects across sites and over time, and variations in site-specific codon frequencies, that mimic the real mtDNA alignment. Under this generating scenario, RaMoSS versus RaMoSSwDT falsely detected DT in 41/100 of the 300-codon alignments and 48/50 of the full-scale alignments. These results illustrate that pathologies associated with confounding might only be realized by fitting a contrast to be applied to a real data set to alignments that are comparable with that data. It was shown that the generating model MutSel-mmtDNA can produce alignments that are similar in many respects to the real mtDNA alignment used in this study. However, MutSel-mmtDNA neglects many important aspects of molecular evolution that might further impact inference. For example, MutSel-mmtDNA does not include changes in site-specific fitness coefficients that initiate site-specific dynamics consistent with adaptive evolution (e.g., a peak shift, dos Reis 2015), and does not take into account effects such as epistasis or selection on thermodynamic stability (Pollock et al. 2012). It might therefore be necessary to continue to work toward generating models of greater realism by including these and other such processes.

Materials and Methods

The Mutation Model

The model of the mutation process used in this article is similar to the mechanistic mutation model presented in Tamuri et al. (2012). The mutation rate from codon $i = i_1 i_2 i_3$ to codon $j = j_1 j_2 j_3$ was specified as:

$$M_{ij} \propto \begin{cases} \kappa^{n_t} \prod_{i_k \neq j_k} \pi^*_{j_k} & \text{if } n{=}1 \\ \alpha\kappa^{n_t} \prod_{i_k \neq j_k} \pi^*_{j_k} & \text{if } n{=}2 \\ \beta\kappa^{n_t} \prod_{i_k \neq j_k} \pi^*_{j_k} & \text{if } n{=}3 \end{cases} \quad (9)$$

Equation (9) applies to all pairs of codons (i, j) that differ by $n \in \{1, 2, 3\}$ nucleotides, n_t of which are transitions. The $\pi^*_{j_k}$ are position-specific nucleotide frequencies; κ is the transition/transversion rate ratio; α and β determine the rate of double and triple mutations, respectively. Diagonal elements M_{ii} were adjusted to make rows sum to 0.

The New Model RaMoSS

RaMoSS is a mixture of two standard CSMs: M3 to account for static sites (those evolving under one of two rate ratios ω_0 or ω_1 across the tree) and CLM3 to account for switching sites (those that change between ω_0' and ω_1' randomly in time). Both M3 and CLM3 are based on substitution rate matrices constructed from the mutation rate matrix M defined in equation (9) and a nonsynonymous-to-synonymous substitution rate ratio ω:

$$Q_\omega = M \circ (I_S + \omega I_N) \tag{10}$$

where $\circ$ represents the entrywise matrix product, I_S is an indicator matrix whose $(i,j)^{th}$ element is one if i and j are synonymous and 0 otherwise, and I_N similarly indicates nonsynonymous codon pairs (the diagonal elements of Q_ω are adjusted to make its rows sum to 0). The row vector of stationary codon frequencies $\boldsymbol{\pi} = \langle \pi_1, \ldots, \pi_{60} \rangle$ associated with equation (10) is independent of ω and can be found by solving $\boldsymbol{\pi}\mathrm{M} = 0$. Hence, $\boldsymbol{\pi}$ is determined by the mutation process alone and is the same for all Q_ω. It is convenient to specify M3 using a compound rate matrix defined as follows:

$$Q_{\mathrm{M3}} = \frac{1}{r}\begin{bmatrix} Q_{\omega_0} & 0 \\ 0 & Q_{\omega_1} \end{bmatrix} \tag{11}$$

The state space for M3 consists of 120 $(\text{codon}, \text{rate ratio})$ pairs. CLM3 can be specified in a similar way, but requires two compound rate matrices, one for substitutions and the other for switches between ω_0' and ω_1' (cf. Guindon et al. 2004):

$$Q_{\mathrm{CLM3}} = \frac{1}{r}\begin{bmatrix} Q_{\omega_0'} & 0 \\ 0 & Q_{\omega_1'} \end{bmatrix} + \frac{\delta}{c}\begin{bmatrix} -(1-p_0')I & (1-p_0')I \\ p_0'I & -p_0'I \end{bmatrix} \tag{12}$$

By this formulation, a site evolving under CLM3 is allowed to change either its codon or its rate ratio at any instant, but never both.

The value of the scaling constant r in equations (11) and (12) can be specified to make branch length equivalent to the expected number of single nucleotide substitutions per codon. Assuming a common mutation process M, the scaling factor for any individual rate matrix depends only on ω and can be specified as follows:

$$r_\omega = \sum_{(i,j)} \pi_i Q_\omega(i,j)\{\ell_1 + 2\ell_2 + 3\ell_3\} \tag{13}$$

The indicator ℓ_n is one if i and j differ by $n \in \{1,2,3\}$ nucleotides and 0 otherwise. The scaling constant r is constructed by taking into account both the proportion p_0 of static sites evolving under ω_0 and the proportion $1 - p_0$ of sites evolving under ω_1, as well as the expected proportion of time a switching spends evolving under ω_0' (i.e., p_0') or ω_1' $(1 - p_0')$:

$$r = p_{\mathrm{M3}}(p_0 r_{\omega_0} + (1-p_0) r_{\omega_1}) + (1-p_{\mathrm{M3}})(p_0' r_{\omega_0'} + (1-p_0') r_{\omega_1'}) \tag{14}$$

The matrix governing the switching process in equation (12) can be scaled to make δ the expected number of switches per unit branch length. The scaling parameter c is determined by multiplying the diagonal matrix D whose entries are equal to the vector of stationary frequenices for the state pairs under CLM3:

$$\langle p_0'\pi_1, \ldots, p_0'\pi_{60}, (1-p_0')\pi_1, \ldots, (1-p_0')\pi_{60} \rangle \tag{15}$$

with the switching matrix in equation (12), and then summing over all values corresponding to a switch in rate ratio (i.e., summing over all but the elements on the main diagonal). It can be shown that the resulting scaling factor is:

$$c = 2\sum_{i=1}^{60} p_0'(1-p_0')\pi_i = 2p_0'(1-p_0') \tag{16}$$

Scaling the switching matrix in this way was first proposed by Jones et al. (2017). However, their equation for the scaling parameter c (labeled r_2 in eq. 22 of that article) contained an error that made $r_2 = 0$. Equation (16) corrects this.

The likelihood for RaMoSS is a weighted average of the likelihoods for the M3 and CLM3 components, each of which is computed using the pruning algorithm (Felsenstein 1981):

$$\mathrm{L}_{\mathrm{RaMoSS}}(\theta_{\mathrm{RaMoSS}}|\mathrm{X},\mathrm{T}) = p_{\mathrm{M3}}\mathrm{L}_{\mathrm{M3}}(\theta_{\mathrm{M3}}|\mathrm{X},\mathrm{T}) + (1-p_{\mathrm{M3}})\mathrm{L}_{\mathrm{CLM3}}(\theta_{\mathrm{CLM3}}|\mathrm{X},\mathrm{T}) \tag{17}$$

where X represents the alignment, T the topology of the tree, and θ_{RaMoSS} a vector that includes the model parameters for both M3 and CLM3, as well as an additional parameter p_{M3} for the proportion of sites evolving under M3. The posterior probability of heterotachy at the h^{th} site (see table 3) can be computed from the MLE for θ_{RaMoSS} using the standard naive Bayesian approach:

$$P(\text{switching}|x^h, \hat{\theta}_{\mathrm{RaMoSS}}) = \frac{L_{\mathrm{CLM3}}(x^h|\hat{\theta}_{\mathrm{CLM3}}, T)(1-\hat{p}_{\mathrm{M3}})}{L_{\mathrm{M3}}(x^h|\hat{\theta}_{\mathrm{M3}}, T)\hat{p}_{\mathrm{M3}} + L_{\mathrm{CLM3}}(x^h|\hat{\theta}_{\mathrm{CLM3}}, T)(1-\hat{p}_{\mathrm{M3}})} \tag{18}$$

where x^h is the site pattern.

Model Contrasts

Nested models (a null model vs. an alternative, e.g., M0 vs. M0wDT) can be compared using a log-likelihood ratio test. The null hypothesis is that the data were generated under the simpler of the two models (e.g., M0). This is rejected if the log-likelihood ratio (LLR) for the test is larger than a critical value determined by the limiting distribution of the log-likelihood ratio statistic and the level of significance of the test. In this article, the models M0, M3, CLM3, and RaMoSS were fitted to real and simulated alignments. Each allows single nucleotide mutations only (e.g., $\alpha = \beta = 0$ in eq. 9). The four models have counterparts that allow double and triple mutations (e.g., α and β in eq. 9 are estimated): M0wDT, M3wDT, CLM3wDT, and RaMoSSwDT. The contrast between M and MwDT provides a test for DT mutations, where M $\in \{\mathrm{M0},\ \mathrm{M3},\ \mathrm{CLM3},\ \mathrm{RaMoSS}\}$. In a similar fashion, the M0-M3 contrast provides a tests for variation in the rate ratio across sites; M3-CLM3 provides a test for variations in the rate ratio over time; and CLM3-RaMoSS provides a test for a combination of static and switching sites in the same alignment compared with switching sites only. The limiting distribution of the LLR statistic is often unknown. In such cases, it is standard practice to use a distribution that is thought to be more conservative (i.e., less likely to reject the null hypothesis) than the unknown true distribution.

Table 6. Critical Values Used for the Log-Likelihood Ratios Tests in This Article.

Contrast	df	Theoretical Distribution	Implemented	Crit. val.
M vs. MwDT	2	n/a	χ^2_2	5.99
M0 vs. M3	2	n/a	χ^2_2	5.99
M3 vs. CLM3	1	$0.5\chi^2_0 + 0.5\chi^2_1$	$0.5\chi^2_0 + 0.5\chi^2_1$	2.71
CLM3 vs. RaMoSS	4	n/a	χ^2_4	9.49

df, the number of extra parameters in the larger model compared with its nested counterpart.

The distributions used for the tests in this study are listed in table 6, along with the corresponding critial values for 5% level of significance. The null hypothesis for all of the M versus MwDT contrasts places both $\alpha = 0$ and $\beta = 0$ on the boundary of the parameter space. The theoretical limiting distribution is therefore a mixture of a χ^2_0, a χ^2_1 and a χ^2_2 distribution (Self and Liang 1987). The mixing weights are unknown however, so we assumed a χ^2_2 distribution to be conservative. The proportion of sites in the ω_1 category under the null for the M0 versus M3 contrast is $p_1 = 1 - p_0 = 0$, making ω_1 unidentifiable. Similarly, the proportion of sites evolving with constant rate ratio under the null for the CLM3 versus RaMoSS contrast is $p_{M3} = 0$, making ω_0', ω_1' and p_0' unidentifiable. The theoretical limiting distributions for these contrasts are not available from Self and Liang (1987). We therefore used the conventional χ^2_{df} distribution with degrees of freedom (df) equal to the difference in the number of parameters (table 6). The theoretical distribution for the M3 versus CLM3 contrast is known to be an equal mixture of a χ^2_0 and a χ^2_1 (Self and Liang 1987).

Constructing PDFs for Scaled Selection Coefficients

The probability density functions (PDFs) for the scaled selection coefficients depicted in Figure 2 in Supplementary Material online and used to compute the values in Table 5 were approximated by discrete probability mass functions (PMFs). This section explains how the PMFs were constructed (cf., Tamuri et al. 2014). We started with a fixed set of $n = 10^5$ vectors of site-specific fitness coefficients from which a fixed set $S = \{s^h_{ij}\}^n_{h=1}$ of scaled selection coefficients was produced. The PMF for all mutations was then constructed as follows:

(1) $p^h_{ij} = \pi^h_i M_{ij}$ was computed for each s^h_{ij}; p^h_{ij} is proportional to the long-run probability that a mutation will occur at site h and correspond to $i \rightarrow j$ with associated scaled selection coefficient s^h_{ij}.

(2) The elements of S were then partitioned into 50 bins. The left-most bin was the interval $(-\infty, -10)$ and the right-most bin was $(10, +\infty)$. The remaining bins between ± 10 were constructed with bin width ≈ 0.4.

(3) Each bin was assigned a sum $c_b = \sum_{i,j,h} p^h_{ij} \ell(s^h_{ij} \in$ the b^{th} bin$)$ where $\ell(s^h_{ij} \in$ the b^{th} bin$)$ is one if s^h_{ij} is in the b^{th} bin and 0 otherwise.

(4) Each c_b was then divided by $\sum_{b=1}^{50} c_b$.

(5) The resulting values were plotted against the bin centers, except for the end points c_1 and c_{50}, for which the abscissa was -10 and $+10$, respectively.

The PMF for all substitutions was constructed by first setting $p^h_{ij} = \pi^h_i A^h_{ij}$, where A^h_{ij} is the site-specific substitution rate matrix, followed by the same steps 2 to 5. The PMFs for nonsynonymous mutations and nonsynonymous substitutions were similarly constructed using s^h_{ij} and p^h_{ij} corresponding to nonsynonymous pairs of codons i and j. The resulting PMFs approximate continuous distributions of scaled selection coefficients s_{ij}, and can be used to approximate integrals. For instance, $p(-2 < s_{ij} < 2)$ in the first row of table 5 is approximated by the sum of c_b correponding to bin centers between -2 and 2, and gives the expected proportion of mutations across sites and over time that would have a selection coefficient between -2 and 2.

Acknowledgments

Funding for this research was provided by the Natural Sciences and Engineering Council of Canada (J.P.B. and E.S.). We thank the reviewers for their questions and comments, which greatly improved the quality of this work.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

References

Anisimova M, Bielawski JP, Yang ZH. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol*. 18:1585–1592.

Anisimova M, Bielawski JP, Yang ZH. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol*. 19(6):950–958.

Averof M, Rokas A, Wolfe KH, Sharp PM. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* 287:1283–1286.

Bielawski JP, Gold JR. 2002. Mutation patterns of mitochondrial H- and L-strand dna in closely related cyprinid fishes. *Genetics* 161:1589–1597.

Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Paabo S, Hasegawa M. 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J Mol Evol*. 47:307–322.

Clayton DA. 1982. Replication of animal mitochondrial DNA. *Cell* 28(4):693–705.

dos Reis M. 2015. How to calculate the non-synonymous to synonymous rate ratio protein-coding genes under the Fisher-Wright mutation-selection framework. *Biol Lett*. 11:1–4.

Felsenstein JJ. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol*. 17(6):368–376.

Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol Biol Evol*. 18(5):866–873.

Garvin MR, Bielawski JP, Sazanov LA, Gharrett AJ. 2015. Review and metaanalysis of natural selection in mitochondrial complex I in metazoans. *J Zool Syst Evol Res*. 53:1–17.

Goldman N, Yang ZH. 1994. Codon-based model of nucleotide substitution for protein-coding dna-sequences. *Mol Biol Evol*. 11:725–736.

Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP. 2004. Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci U S A*. 101:12957–12962.

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

Jones CT, Youssef N, Susko E, Bielawski JP. 2017. Shifting balance on a static mutation-selection landscape: a novel scenario of positive selection. *Mol Biol Evol*. 34:391–407.

Kimura M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 47:713–719.
Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*. 22:1208–1222.
Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. Hyphy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.
Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol*. 28:3033–3043.
Kosakovsky Pond SL, Muse SV. 2005. Site-to-site variations of synonymous substitution rates. *Mol Biol Evol*. 22(12):2375–2385.
Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. *Mol Biol Evol*. 24(7):1464–1479.
Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenetics. *Mol Biol Evol*. 29(2):457–472.
Liberles DA, Teufel AI, Liu L, Stadler T. 2013. On the need for mechanistic models in computational genomics and metagenomics. *Genome Biol Evol*. 5:2008–2018.
Lopez P, Casane D, Philippe H. 2002. Heterotachy, and important process of protein evolution. *Mol Biol Evol*. 19(1):1–7.
Lu A, Guindon S. 2014. Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. *Mol Biol Evol*. 31(2):484–495.
Mingrone J, Susko E, Bielwaski JP. 2018. Modified likelihood ratio tests for positive selection. The penalized likelihood paper.
Miyazawa S. 2011. Advantages of a mechanistic codon substitution model for evolutionary analysis of protein-coding sequences. *PLoS One* 6(12):e28892.
Moran PAP. 1958. Random processes in genetics. *Math Proc Camb Philos Soc*. 54: 60–71.
Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, et al. 2015. Gene-wide identification of episodic selection. *Mol Biol Evol*. 32(5):1365–1371.
Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. *Mol Biol Evol*. 11:715–724.
Pollock DD, Thiltgen G, Goldstein RA. 2012. Amino acid coevolution induces an evolutionary Stokes shift. *Proc Natl Acad Sci U S A*. 109:E1352–E1359.
Raina SZ, Faith JJ, Disotell TR, Seligmann H, Stewart CB, Pollock DD. 2005. Evolution of base-substitution gradients in primate mitochondrial genomes. *Genomes* 15(5):665–673.
Reyes A, Gissi C, Pesole G, Saccone C. 1998. Asymmetric directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol*. 15(8):957–966.
Rodrigue N, Lartillot N. 2014. Site-heterogeneous mutation-selection models with the PhyloBayes-MPI package. *Bioinformatics* 30:1020–1021.
Rodrigue N, Philippe H. 2010. Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends Genet*. 26:248–252.
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 U S A*. 107:4629–4634.
Self SG, Liang KY. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio test under nonstandard conditions. *JASA* 82:605–610.
Spielman S, Wilke CO. 2015a. Pyvolve: a flexible Python module for simulating sequences along phylogenies. *PLoS One* 10(9):e0139047.
Spielman S, Wilke CO. 2015b. The relationship between dN/dS and scaled selection coefficients. *Mol Biol Evol*. 34:1097–1108.
Spielman S, Wilke CO. 2016. Extensively parameterized mutation-selection models reliably capture site-specific selective constraints. *Mol Biol Evol*. 33:2990–3001.
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.
Tamuri AU, Goldman N, dos Reis M. 2014. A penalized-likelihood method to estimate the distribution of seletion coefficients from phylogenetic data. *Genetics* 197(1):257–271.
Tanaka M, Ozawa T. 1994. Strand asymmetry in human mitochondrial mutations. *Genomics* 22:327–335.
Wertheim JO, Murrell B, Smith MD, Pond SLK, Scheffler K. 2014. Relax: detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol*. 32:820–832.
Whelan S, Goldman N. 2004. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics* 167:2027–2043.
Wong WSW, Yang ZH, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168(2):1041–1051.
Yang ZH. 2007. PAML4: phylogentic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.
Yang ZH, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol*. 28(3):1217–1228.
Yang ZH, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 19:908–917.
Yang ZH, Nielsen R, Goldman N. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.
Yang ZH, Wong SWS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 22(4):1107–1118.
Zaheri M, Dib L, Salamin N. 2014. A generalized mechanistic codon model. *Mol Biol Evol*. 31(9):2528–2541.
Zhang J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol*. 21:1332–1339.
Zhang J, Nielsen R, Yang ZH. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*. 22(12):2472–2479.