Syst. Biol. 69(4):722–738, 2020

DOI:10.1093/sysbio/syz075
Advance Access publication December 6, 2019

A Phenotype–Genotype Codon Model for Detecting Adaptive Evolution

CHRISTOPHER T. JONES[1,*], NOOR YOUSSEF[2], EDWARD SUSKO[1,3] AND JOSEPH P. BIELAWSKI[1,2,3]

[1]*Department of Mathematics and Statistics, Dalhousie University, 1233 LeMarchant Street, B3H 4R2, Halifax, Nova Scotia, Canada;* [2]*Department of Biology, Dalhousie University, 1233 LeMarchant Street, B3H 4R2, Halifax, Nova Scotia, Canada; and* [3]*Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, 1233 LeMarchant Street, B3H 4R2, Halifax, Nova Scotia, Canada*

[*]*Correspondence to be sent to: Department of Mathematics and Statistics, Dalhousie University, 1233 LeMarchant Street, B3H 4R2, Halifax, Nova Scotia, Canada; E-mail: cjones2@dal.ca.*

Received 7 May 2019; reviews returned 9 November 2019; accepted 11 November 2019
Associate Editor: Jeremy Beaulieu

Abstract.—A central objective in biology is to link adaptive evolution in a gene to structural and/or functional phenotypic novelties. Yet most analytic methods make inferences mainly from either phenotypic data or genetic data alone. A small number of models have been developed to infer correlations between the rate of molecular evolution and changes in a discrete or continuous life history trait. But such correlations are not necessarily evidence of adaptation. Here, we present a novel approach called the phenotype–genotype branch-site model (PG-BSM) designed to detect evidence of adaptive codon evolution associated with discrete-state phenotype evolution. An episode of adaptation is inferred under standard codon substitution models when there is evidence of positive selection in the form of an elevation in the nonsynonymous-to-synonymous rate ratio ω to a value $\omega > 1$. As it is becoming increasingly clear that $\omega > 1$ can occur without adaptation, the PG-BSM was formulated to infer an instance of adaptive evolution without appealing to evidence of positive selection. The null model makes use of a covarion-like component to account for general heterotachy (i.e., random changes in the evolutionary rate at a site over time). The alternative model employs samples of the phenotypic evolutionary history to test for phenomenological patterns of heterotachy consistent with specific mechanisms of molecular adaptation. These include 1) a persistent increase/decrease in ω at a site following a change in phenotype (the pattern) consistent with an increase/decrease in the functional importance of the site (the mechanism); and 2) a transient increase in ω at a site along a branch over which the phenotype changed (the pattern) consistent with a change in the site's optimal amino acid (the mechanism). Rejection of the null is followed by *post hoc* analyses to identify sites with strongest evidence for adaptation in association with changes in the phenotype as well as the most likely evolutionary history of the phenotype. Simulation studies based on a novel method for generating mechanistically realistic signatures of molecular adaptation show that the PG-BSM has good statistical properties. Analyses of real alignments show that site patterns identified *post hoc* are consistent with the specific mechanisms of adaptation included in the alternate model. Further simulation studies show that the covarion-like component of the PG-BSM plays a crucial role in mitigating recently discovered statistical pathologies associated with confounding by accounting for heterotachy-by-any-cause. [Adaptive evolution; branch-site model; confounding; mutation-selection; phenotype–genotype.]

Statistical models for the evolution of phenotypes have traditionally been formulated independently of models for the evolution of gene sequences. Yet the two approaches share a common motivation, namely to provide a means to test various evolutionary hypotheses regarding apparent structural and/or functional novelties that might have occurred as a result of adaptation. Analyzing the two data types separately neglects any possible advantage of combining information and belies the fundamental objective of identifying individual genes whose evolution can be mechanistically linked to adaptive changes in phenotype. The centrality of this objective underlines the need for models that combine the two types of data under a common statistical framework. In this article, we propose such a model.

Among the first models for the evolution of phenotype were those developed to infer the rate and mode (e.g., gradual or punctuated) of phenotypic evolution, or to infer correlations between two phenotypic measures or between a phenotype and a contextual variable (for a brief review see Cornwell and Nakagawa 2017). Such models, which typically assume either a continuous phenotype that evolved via Brownian Motion (Felsenstein 1973) or a discrete phenotype that evolved via a Markov process (Pagel 1994; Lewis 2001), provide the basis for a wide variety of phylogenetic comparative methods. Sophisticated phylogenetic comparative methods now include models that assume an Ornstein–Uhlenbeck "mean-reverting" evolutionary process (Hansen 1997), models that account for temporal dynamics in the form of changes in the rate of change in a phenotype over the tree (Butler and King 2004; O'Meara et al. 2006; Eastman et al. 2011), and models that test for relationships between phenotype and diversification (e.g., the binary state speciation–extinction model, Maddison et al. 2007). More recently, several methods for the analysis of multivariate data have been proposed (for a candid assessment see Adams and Collyer 2018). The relevant point is that the majority of phylogenetic comparative methods use alignments of homologous protein-coding genes to estimate phylogentic relationships that are treated as fixed for the remainder of an analysis based on the phenotype data alone.

Codon substitution models were developed to detect evidence of adaptation at the molecular level. Under the current paradigm, the canonical signature of positive selection in the form of a nonsynonymous-to-synonymous rate ratio (typically denoted ω) greater than its neutral expectation (i.e., $\omega > 1$) is considered evidence of adaptation (e.g., Yang et al. 2000). Among the more sophisticated models of this type in use today are the branch-site models designed to detect evidence of adaptation at some sites along particular branches of

the tree (Yang and Nielsen 2002; Yang et al. 2005; Zhang et al. 2005). Amino acid substitution models formulated to detect clade-specific changes in the replacement rate (Type I functional divergence) or the preferred amino acid at a site (Type II functional divergence) (Gu 1999, 2001, 2006; Gaston et al. 2011) have also been proposed. Both approaches require *a priori* specification of the branches over which changes in the substitution process are thought to have occurred. This is often realized via informal use of external information such as phenotype.

Models that account for molecular and phenotypic evolution under a unified statistical framework have been proposed (Mayrose and Otto 2011; Lartillot and Poujol 2011; O'Connor and Mundy 2013; Karin et al. 2017). In CoEvolve (Lartillot and Poujol 2011), for example, $\log(\omega)$ is assumed to have evolved continuously over the tree via Brownian motion and the model objective is to estimate correlations between it and other continuous variables, such as body size, longevity, and metabolic rate. Similarly, in TrateRateProp (Karin et al. 2017) the objective is to determine whether a subset of nucleotide sites evolved under one of the two substitution rates depending on the state of a binary phenotype. Neither model appeals to mechanisms by which evolution of the phenotype might be linked to evolution of the gene. Here, we develop a novel approach in the form of a phenotype–genotype branch-site model (PG-BSM), the objective of which is to link phenomenological signatures of site-specific variations in ω (a.k.a. heterotachy, Lopez et al. 2002) to specific mechanistic processes, including those that occurred in association with changes in a discrete character state (e.g., a phenotype).

The mutation-selection framework of Halpern and Bruno (1998), and the notion of a site-specific fitness landscape that it implies (McCandlish 2011; Jones et al. 2017), provides a means to think about the mechanistic processes that can give rise to heterotachy in real alignments. Under this framework, each site is assumed to evolve independently with its own vector of fitness coefficients for the 20 amino acids (i.e., a site-specific fitness landscape). A site evolving on a static landscape can undergo chance fixation to a suboptimal amino acid followed by a period of positive selection that restores the site to its optimal state. This results in heterotachy via a process we call nonadaptive shifting balance (Jones et al. 2017). Wright introduced his theory of shifting balance to explain how a subpopulation might move from one fitness peak across a fitness valley to another higher peak on a fixed landscape and subsequently cause the entire population to move to the new peak (Wright 1932, 1982). Here, we use nonadaptive shifting balance to refer to the movement of an entire population away from and back to the same peak on a fixed site-specific landscape. Heterotachy can also be caused by episodic changes in site-specific landscapes congruent with molecular adaptation, such as a change in the optimal amino acid (i.e., a peak shift) or a change in the stringency of selection at a site. Nonadaptive shifting balance and episodic changes in site-specific landscapes can both be represented phenomenologically as a Markov-modulated or "covarion-like" Markov process (Galtier 2001) under which sites switch randomly between two rate ratios $\omega_1 < \omega_2$ over time. Significantly, nonadaptive shifting balance on static fitness landscapes and episodic adaptive changes in landscapes can both manifest as episodic elevations to $\omega_2 > 1$ (dos Reis 2015; Jones et al. 2017). It follows that the canonical $\omega > 1$ signature of positive selection does not necessarily provide unequivocal evidence of adaptation (Jones et al. 2017).

The PG-BSM was formulated to identify sites that likely underwent adaptation without appealing to evidence for positive selection (it will be argued that adaptive evolution and positive selection are not necessarily commensurate insofar as adaptation implies changes in a site-specific fitness landscape). Our approach is in some ways similar to the amino acid models developed to detect functional divergence (Gu 1999, 2001, 2006; Gaston et al. 2011) and other models formulated to infer adaptive evolution without requiring $\omega > 1$ (e.g., Tamuri et al. 2009; Parto and Lartillot 2018). The host-shift model, for example, was motivated in part by the recognition that heterotachy can be associated with changes in site-specific amino acid fitnesses, implying adaptation (Tamuri et al. 2009). The differential selection model takes a Bayesian approach to infer changes in amino acid proclivities at a site that imply adaptation (Parto and Lartillot 2018). Both models assume that the branches over which the phenotype changed are known. The PG-BSM, by contrast, explicitly accounts for uncertainty in the location of such branches by making formal use of a discrete phenotype assigned to the terminal nodes of the tree. Branches over which the phenotype might have changed are determined by a distribution of histories at the internal nodes of the tree derived from a model for phenotype evolution (cf. Karin et al. 2017). It is assumed under the null hypothesis that all heterotachous sites evolved independently of the phenotype and that their observed site patterns are consistent with the phenomenological covarion-like process of random shifts between $\omega_1 < \omega_2$. The alternative model permits specific modes of switching between $\omega_1 < \omega_2$ that occurred in coordination with changes in the discrete phenotype. The modes are specified to be consistent with either a change in the stringency of selection or a change in the optimal amino acid at a site. Rejection of the null is interpreted as evidence for the existence of sites where replacement substitutions apparently occurred in association with changes in phenotype, hereafter referred to as phenotype–genotype association. The PG-BSM represents a paradigm shift both in the information it uses (genotype and phenotype) and in the form of evidence for molecular adaptation (specific modes of heterotachy) it is meant to detect.

Materials and Methods

Background

The traditional way of characterizing codon evolution is to estimate from an alignment of homologous protein-coding genes the ratio of the nonsynonymous substitution rate dN to the synonymous substitution rate dS, accounting for differences in the rate at which

nonsynonymous and synonymous substitutions occur under neutral selection (see Jones et al. 2018 for an explanation of the way dN and dS are defined). Selection regimes are categorized according to $\omega = dN/dS$, such that $\omega < 1$ indicates a conservative regime, $\omega = 1$ a neutral regime, and $\omega > 1$ the canonical positive selection regime. Codon substitution models can be used to infer $\omega > 1$ by contrasting a null model that allows sites to evolve under a set of ω-categories all with $\omega \leq 1$ with an alternate model that includes an additional category for sites with $\omega > 1$. Rejection of the null is interpreted as evidence that positive selection occurred somewhere in the gene. Subsequent analysis can be conducted to identify sites at which positive selection is most likely to have occurred (e.g., Yang and Nielsen 1998).

The majority of codon substitution models are based on a continuous-time homogeneous and time-reversible Markov process that describes the rate at which substitutions occur under neutral selection (i.e., with $\omega = 1$). This can be represented by a substitution rate matrix M, which in this study was constructed as follows (Jones et al. 2018):

$$M_{ij} \propto \begin{cases} \kappa^{s_t} \Pi_{i_k \neq j_k} \pi^*_{j_k} & \text{if } s = 1 \\ \alpha\kappa^{s_t} \Pi_{i_k \neq j_k} \pi^*_{j_k} & \text{if } s = 2 \\ \beta\kappa^{s_t} \Pi_{i_k \neq j_k} \pi^*_{j_k} & \text{if } s = 3 \end{cases}. \quad (1)$$

Equation (1) applies to all pairs of codons (i, j) that differ by $s \in \{1,2,3\}$ nucleotides, s_t of which are transitions (substitutions of the form T↔C or A↔G) and $s - s_t$ of which are transversions (substitutions of the form {T,C} ↔ {A,G}). $\pi^*_{j_k}$ is the frequency of the nucleotide in the kth $\in \{1,2,3\}$ position of the jth codon (i.e., j_k is a nucleotide $j_k \in$ {T, C, G, A}), κ is the transition/transversion rate ratio, and α and β are the rates at which double and triple substitutions occur. Diagonal elements M_{ii} are adjusted to make rows sum to zero. The selection process can be introduced via an element-wise matrix product:

$$Q(\omega) = M \circ (\boldsymbol{\ell}_S + \omega \boldsymbol{\ell}_N)/r_\omega, \quad (2)$$

$$\text{where } r_\omega = \sum_{j \neq i} \pi_i M_{ij} (\boldsymbol{\ell}_S(i,j) + \omega \boldsymbol{\ell}_N(i,j)) \{\ell_1 + 2\ell_2 + 3\ell_3\}.$$

Diagonal elements $Q_{ii}(\omega)$ are adjusted to make rows sum to zero. Here, $\boldsymbol{\ell}_S$ represents an indicator matrix whose (i,j)th element $\boldsymbol{\ell}_S(i,j)$ is one if i and j are synonymous and zero otherwise. $\boldsymbol{\ell}_N$ similarly indicates nonsynonymous codon pairs. The constant r_ω normalizes $Q(\omega)$ so that branch lengths give the expected number of single nucleotide substitutions per codon site and is computed using the stationary codon frequencies $\pi_i \propto \pi^*_{i_1} \pi^*_{i_2} \pi^*_{i_3}$. The scalar indicator ℓ_k is one if i and j differ by $k \in \{1,2,3\}$ nucleotides and zero otherwise. Note that M is a component of the analytic models fitted to data as well as the alignment-generating models used to simulate data. All of the analytic models used in this study assumed single nucleotide substitutions only (i.e., with $\alpha = \beta = 0$) and made neutral substitution rates proportional to the frequency of the nucleotide at the position where the substituting codon j differs from the incumbent codon i, consistent with Muse and Gaut (1994) (cf. Goldman and Yang 1994). Some of the alignment-generating models allowed fixation of double–triple mutations (i.e., with α and β both > 0).

The rate matrix $Q(\omega)$ is a useful phenomenological approximation of the evolutionary process at a codon site but is unsuitable as a means of *thinking* about the process. For example, the rate ratio ω, a proxy for the strength of selection for ($\omega > 1$) or against ($\omega < 1$) the i to j substitution, is assumed to be the same for all nonsynonymous (i,j) pairs. This is conceptually misleading for the majority of proteins because it implies that the fitness of an amino acid at a site is independent of its physicochemical properties. It is more useful to think of the evolutionary process at a codon site in terms of the dynamic on its site-specific fitness landscape (Jones et al. 2017) as characterized by the mutation-selection modeling framework (Halpern and Bruno 1998). If codon sites are assumed to evolve independently, a site-specific fitness landscape can be defined for the hth site by a vector of fitness coefficients $\mathbf{f}^h$ or its implied vector of equilibrium codon frequencies $\boldsymbol{\pi}^h$ (Sella and Hirsh 2005). These determine the evolutionary dynamic at the site, or the way it "moves" across its landscape over macroevolutionary time scales. Possible dynamic regimes include: nonadaptive shifting balance, under which the site moves episodically away from the peak of its static fitness landscape (i.e., the fittest amino acid) via drift and back again by positive selection; adaptive evolution, under which a change in the landscape in the form of a peak shift is followed by movement of the site toward its new fitness peak; and neutral or nearly neutral evolution, under which drift dominates and the site is free to move over a relatively flat landscape constrained primarily by biases in the mutation process.

The PG-BSM

Genetic information is assumed to consist of an alignment X of N homologous protein-coding sequences of length n with a known rooted topology τ. The phenotype, encoded by a vector $\mathbf{F}$, can be any discrete character state, such as a property of the gene's protein product (i.e., a molecular phenotype), some characteristic of the organism, or an environmental variable. The PG-BSM fitted to $(X, \mathbf{F})$ consists of three components : 1) a model for the evolution of the codon sequence; 2) a model for the evolution of a discrete phenotype; and 3) a model that accounts for the mechanism(s) by which 1) and 2) are associated. Details of all model components are provided in Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.rb4b420. Here, we provide a verbal/visual overview.

A Markov process is the natural choice to model the evolution of a discrete phenotype. It is possible to choose a parameter-rich model that allows a different substitution rate for each pair of phenotypes akin to the generalized time-reversible model for DNA

(Tavaré 1986), which would allow for possible asymmetries caused by canalization, for example. Instead, we chose to use the simpler proportional rates model under which the phenotype is assumed to have changed from i at a parent node to j at a daughter node via a continuous time Markov process at a constant rate proportional to the stationary frequency of state j with proportionality constant λ. This choice was motivated in part by the fact that the number of discrete phenotypic states was two for most of our simulations and real data analyses, and that the generalized time-reversible model for a 2-state system is equivalent to the proportional rates model.

The model for sequence evolution assumes that some proportion π_0 of sites evolved under $\omega_0 = 0$ over the tree while the remaining sites evolved via a covarion-like process with random switches between two rate ratios $\omega_1 < \omega_2$ over time at a rate of δ switches per unit branch length (i.e., under the simple covarion-like model hereafter referred to as CLM3($k=2$) Jones et al. 2017). Alignments typically exhibit variations in rate ratio across sites in addition to site-specific variations over time. It is possible to account for variations across sites using an M-series model such as M3($k=2$) (Yang et al. 2000), which assumes sites evolved under either ω_1 or ω_2 over the entire tree without heterotachy. However, by accounting for random switching between ω_1 and ω_2 over time, the covarion-like model CLM3($k=2$) implicitly accounts for variations in site-specific time-averaged rate ratios (cf. Wu and Susko 2009). Hence, with only one extra parameter (the switching rate δ), CLM3($k=2$) captures variations in rate ratio both across sites and over time. The covarion-like model consequently often provides a better fit to real alignments compared to M3($k=2$) (e.g., Jones et al. 2018). Furthermore, the CLM3($k=2$) component of the PG-BSM provides a means to account for heterotachy caused by processes unassociated with the evolution of the phenotype. Such processes can include not only nonadaptive shifting balance but also adaptive changes in site-specific landscapes not associated with changes in the phenotype. By accounting for what we call heterotachy-by-any-cause, the covarion-like model reduces the probability of falsely rejecting the null hypothesis, as will be demonstrated via simulation studies.

The alternative PG-BSM enforces dependencies between phenotype and genotype evolution at some fraction of sites, or what we call phenotype–genotype associations. Various mechanisms of dependency are amenable to phenomenological representation as distinct modes of heterotachy (Figure 1). Here we consider three. First, a change in phenotype along a branch can coincide with a reduction in the stringency of selection at a site in the descendant clade. This might be caused by a reduction in the site's role in the maintenance of the protein's tertiary structure (Pupko and Galtier 2002), or by a change in a life-history trait (Lartillot and Poujol 2011; Karin et al. 2017) such as a reduction in the reproductive population size. These mechanisms are expressed phenomenologically in the PG-BSM as the cladewise (CW) process under which a proportion π_{CW} of sites are assumed to have evolved under the smaller ω_1 prior to a change in phenotype and under the larger ω_2 over the entire clade descending from the branch over which a change in phenotype occurred (CW tree in Figure 1). Second, a change in phenotype along a branch can coincide with an increase in the stringency of selection at a site. Mechanisms that can produce this result are represented in the PG-BSM by the reverse cladewise (rCW) process under which a proportion π_{rCW} of sites are assumed to have evolved under the larger ω_2 prior to a change in phenotype and under the smaller ω_1 over the entire clade descending from the branch over which a change in phenotype occurred (rCW tree in Figure 1). Third, a change in phenotype can coincide with changes in site-specific fitness landscapes in the form of peak shifts. This mechanism is represented in the PG-BSM by the branchwise (BW) process under which a proportion π_{BW} of sites are assumed to have evolved under the larger ω_2 over branches along which the phenotype changed and under the smaller ω_1 everywhere else in the tree (BW tree in Figure 1).

Sites consistent with the phenomenological CW, rCW or BW processes (herein referred to as CW, rCW, or BW sites) represent a subset of those assumed by the null PG-BSM to have evolved under the covarion-like process. It is therefore assumed that the CW, rCW, BW, and covarion-like processes all share the same ω_1 and ω_2. This is in contrast to the approach taken by the branch-site model first introduced by Yang and Nielsen (2002). One version of that model (the YN-BSM A, Yang et al. 2005) partitions sites into four categories according to the way they are assumed to have evolved. Category 0 and 1 sites are assumed to have evolved under $0 < \omega_0 < 1$ and $\omega_1 = 1$ over the entire tree. Category 2a sites are assumed to have evolved under $0 < \omega_0 < 1$ and category 2b sites under $\omega_1 = 1$ everywhere in the tree except for the set of prespecified foreground branches, where they are both assumed to have evolved under $\omega_2 \geq 1$. The rate ratio ω_2 therefore applies to category 2 sites only. This approach gives the YN-BSM A the power to detect evidence of positive selection (i.e., $\omega_2 > 1$) at a small number of sites along foreground branches, but also introduces the risk of issues related to irregularity (e.g., Baker et al. 2016; Mingrone et al. 2018). When the proportion of category 2 sites (p_2) is small, for example, ω_2 becomes nearly unidentifiable. Its maximum likelihood estimate can consequently be very large and potentially misleading. The PG-BSM avoids potential irregularity issues by using the same parameters ω_1 and ω_2 for all four processes. Our approach undoubtedly reduces the statistical power to detect a small number of sites that evolved under an exceptionally large rate ratio. However, we argue that the potential impact of this loss is mitigated by the fact that the PG-BSM does not rely on evidence of positive selection to reject the null.

The alternate PG-BSM requires knowledge of where in the tree the phenotype changed. This information is

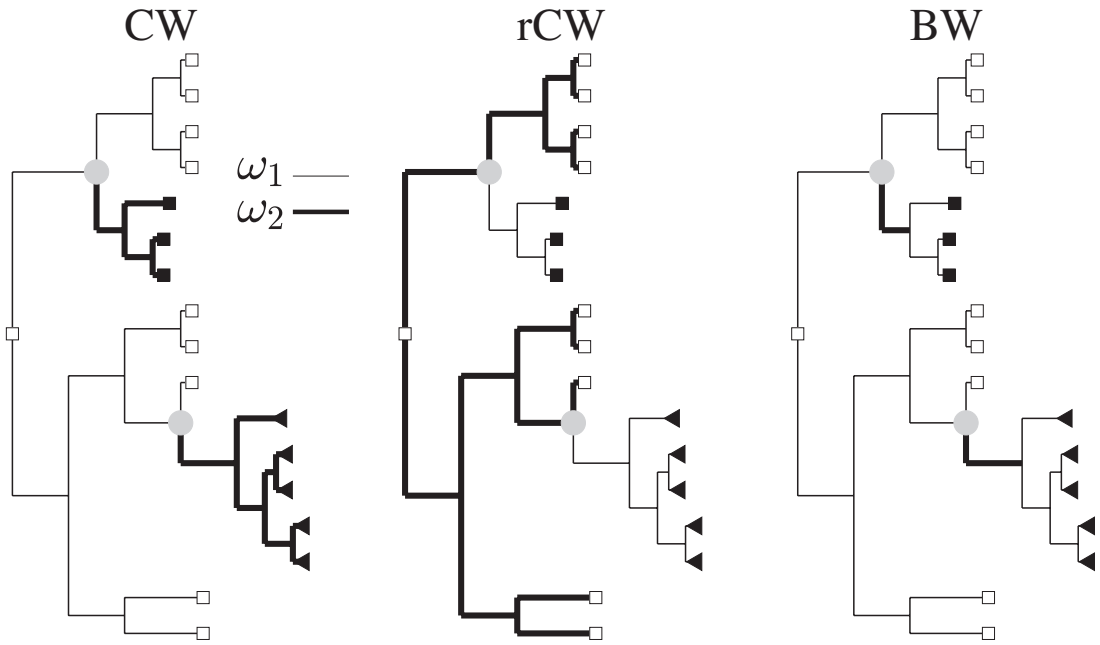

FIGURE 1. An illustration of the difference between the cladewise (CW and rCW) and branchwise (BW) evolutionary processes assumed under the alternative PG-BSM. Each process accounts for a specific form of heterotachy associated with changes in phenotype. The empty and filled markers at the terminal nodes indicate three phenotypic states. $\omega_1 < \omega_2$ are dN/dS rate ratios. The gray disks indicate the nodes at which a change in phenotype occurred. CW sites are assumed to have evolved under ω_1 prior to a change in phenotype and under ω_2 after a change. rCW sites are assumed to have evolved under ω_2 prior to a change in phenotype and under ω_1 after a change. BW sites are assumed to have evolved under ω_2 over branches along which a change in phenotype occurred and under ω_1 everywhere else in the tree. The model assumes a rooted tree because the interpretation of the CW and rCW processes require a particular order of change in rate ratio.

provided by realizations of ancestral phenotypes at the internal nodes of the tree generated using the model for phenotype evolution. Each set of realizations is converted to a change map $\mathbf{z}=(z_1,\ldots,z_{2N-2})$, a vector of zeros and ones where zero indicates a branch over which the phenotype remained constant and one a branch over which the phenotype changed. The likelihood function under the alternative hypothesis can in principle be computed by summing over all possible change maps each weighted by its probability under the model for phenotype evolution. However, the number of possible change maps can be very large depending on the number of taxa and phenotypic states. To make the summation feasible, a sample of 10^5 change maps is generated and the summation is over all unique change maps that appear in the sample each weighted by its relative frequency $\hat{\pi}_{\mathbf{z}}$. To further reduce computational load, change maps that occurred with $\hat{\pi}_{\mathbf{z}} < 10^{-3}$ are excluded. The relative frequencies $\hat{\pi}_{\mathbf{z}}$ of the remaining change maps are renormalized to sum to one.

An omnibus test is conducted to contrast the null and alternative components of the PG-BSM. The components can differ by $m \in \{1,2,3\}$ parameters among the proportions $\{\pi_{\text{CW}}, \pi_{\text{rCW}}, \pi_{\text{BW}}\}$ depending on which of the three processes are included in the alternate model. In all cases, the null PG-BSM is the same as the alternative when the proportions are on the boundary of the parameter space (i.e., when $\pi_{\text{CW}} = \pi_{\text{rCW}} = \pi_{\text{BW}} = 0$). The theoretical limiting distribution of the log-likelihood ratio is therefore a 50:50 mixture of the χ^2_0 and χ^2_1 distributions when $m=1$ and an unknown mixture of the χ^2_0, χ^2_1,...χ^2_m distributions when $m \in \{2,3\}$ (Self and Liang 1987). Although the mixture weights for the distribution are unknown when $m \in \{2,3\}$, the 95th percentile of a mixture of the χ^2_0, χ^2_1,...χ^2_m distributions is always at most that of the χ^2_m distribution (i.e., 3.84, 5.99, and 7.81 for $m=1$, $m=2$, and $m=3$, respectively). Using these as critical values for the omnibus test should therefore be conservative and produce <5% type I error rate.

Rejection of the null hypothesis provides evidence for phenotype–genotype associations at some sites in the alignment. Naive empirical Bayes analysis is then used to identify the most likely process under which each site evolved (i.e., the null covarion-like process or one of the alternatives, the CW, rCW, or BW processes). A false positive occurs when a site pattern x^h is incorrectly attributed to one of the three alternative processes. The false positive rate is usually controlled by attributing site patterns to process $c \in \{\text{CW}, \text{rCW}, \text{BW}\}$ only when the posterior $\text{P}(c \mid x^h)$ is greater than some threshold such as 0.95 (e.g., Yang et al. 2000). An alternative approach, also based on posteriors, is to aim to control the proportion of sites attributed to process c that in fact did not evolve under c (i.e., the false discovery rate, Benjamini and Hochberg 1995). To see the difference between the two approaches, consider an analysis of an alignment with 1000 codon sites. Suppose 10 sites were inferred

to have evolved under the CW process (i.e., there were 10 "discoveries") and that 5 of these were incorrect. Then the false positive rate would only be 0.5% (5 sites out of 1000), whereas the false discovery rate would be 50% (5 sites out of 10). Hence, a low false positive rate does not necessarily imply a low false discovery rate, particularly when the number of discoveries is small. The false discovery rate approach was used in all analyses but with one modification. Rather than controlling the *proportion* of false discoveries of a given category $c \in \{\text{CW}, \text{rCW}, \text{BW}\}$, it was decided to control the *number* of false discoveries or the "false discovery counts" (FDC) for each process.

To quantify the evidential support for branches over which the phenotype is thought to have changed, the probability of the most frequently sampled change map $\mathbf{z}^*$ conditioned on the combined data is estimated. The algorithm that generates change maps makes use of estimates of λ (the proportionality constant for the model of phenotype evolution) and $\mathbf{t}$ (a vector of branch lengths). But λ is independent of the alignment under the null hypothesis, meaning that the proportion $\hat{\pi}_{\mathbf{z}^*}$ of change maps that correspond to the most frequently sampled change map depends on X only through branch length estimates. The likelihood of the combined data under the alternative model $(X, \mathbf{F})$ given $\mathbf{z}^*$, by contrast, also depends on the existence of site patterns that match to greater or lesser degree patterns of heterotachy indicated by $\mathbf{z}^*$ that are consistent with either the CW, rCW, or BW process. It follows that the probability assigned to the most likely history of the phenotype can be increased under the alternate PG-BSM by accounting for such sites if they exist, as demonstrated in the analyses to follow.

Results

Rigorous Model Assessment Requires a Realistic Data-Generating Process

Accuracy and power are usually assessed by fitting a codon substitution model to alignments generated under a similar model (Anisimova et al. 2001, 2002; Wong et al. 2004; Zhang 2004; Kosakovsky Pond and Frost 2005; Yang et al. 2005; Zhang et al. 2005; Yang and dos Reis 2011; Kosakovsky Pond et al. 2011; Lu and Guindon 2013). One drawback of this approach is that standard models constructed from rate matrices of the form $Q(\omega)$ cannot mimic site-specific variations in ω caused by processes such as adaptation following episodic peak shifts (dos Reis 2015) and nonadaptive shifting balance (Jones et al. 2017). This is an issue because heterotachy generated by such processes may well be pervasive in real alignments (e.g., Fitch and Markowitz 1970; Fitch 1971; Lopez et al. 2002; Philippe et al. 2003; Wang et al. 2007; Whelan et al. 2011) and can engender statistical pathologies that will go unnoticed if models are tested using data simulated without such heterotachy (e.g., phenomenological load, Jones et al. 2018).

A direct way to mimic heterotachy is to base the generating model on the mutation-selection framework. Two such generating models were used in this study. The first, dubbed MSmmtDNA, was developed to mimic 12 concatenated H-strand mitochondrial DNA sequences (3331 codon sites) from 20 mammalian species as distributed in the PAML software package (Yang 2007). MSmmtDNA was shown to produce data similar to the real alignment by several measures of comparison, and with similar levels of heterotachy (Jones et al. 2018). Although MSmmtDNA is more realistic as a generating model, it represents only a small portion of the space of all distributions of vectors of site-specific fitness coefficients that might arise in nature. We therefore used a second generating model for some of our simulations that samples with replacement from a set of 3598 such vectors estimated from an alignment of 12 mitochondrial genes taken from 244 mammalian species (Tamuri et al. 2014). This generating process will be referred to as MSTGdR after the authors of that study (Tamuri, Goldman, and dos Reis). Substitutions between codons that differ by two or three nucleotides can only occur in single nucleotide steps under the PG-BSM, consistent with the majority of codon substitution models in common use today. The occasional fixation of double or triple mutations will therefore manifest as an additional source of heterotachy (Jones et al. 2018). It was not clear what effect such fixations might have on power and accuracy when left unaccounted for by the fitted model. We therefore included alignments generated using both MSmmtDNA with 0% double–triple mutations and MSmmtDNA with 6% double–triple mutations (recent studies suggest that double–triple mutations comprise between 1% and 3% of all mutations, Keightley et al. 2009; Schrider et al. 2014; De Maio et al. 2013; Harris and Nielsen 2014). The null PG-BSM itself was also used to generate alignments for the purpose of assessing the statistical properties of the PG-BSM without misspecification.

The mutation-selection framework has been used in several studies to simulate alignments (Holder et al. 2008; Spielman and Wilke 2015, 2016; Spielman et al. 2016; Jones et al. 2017, 2018). In all cases the substitution process was stationary, meaning that fitness coefficients and the stringency of selection were made to be constant at each site over the entire tree. In this study, MSmmtDNA and MSTGdR were formulated to include a subset of sites evolved under nonstationary processes in the form of changes in the stringency of selection and/or fitness coefficients at specific nodes of the tree (Figure 1). Changes in the stringency of selection starting along a single branch leading to a clade can manifest as a cladewise difference in the mean rate ratio ω to produce site patterns phenomenologically consistent with the CW or rCW processes. Similarly, changes in fitness coefficients (a peak shift) along a single branch can result in site patterns phenomenologically consistent with the BW process, particularly if they occur at sites otherwise evolved under stringent selection. In this way, the mutation-selection framework was used to produce

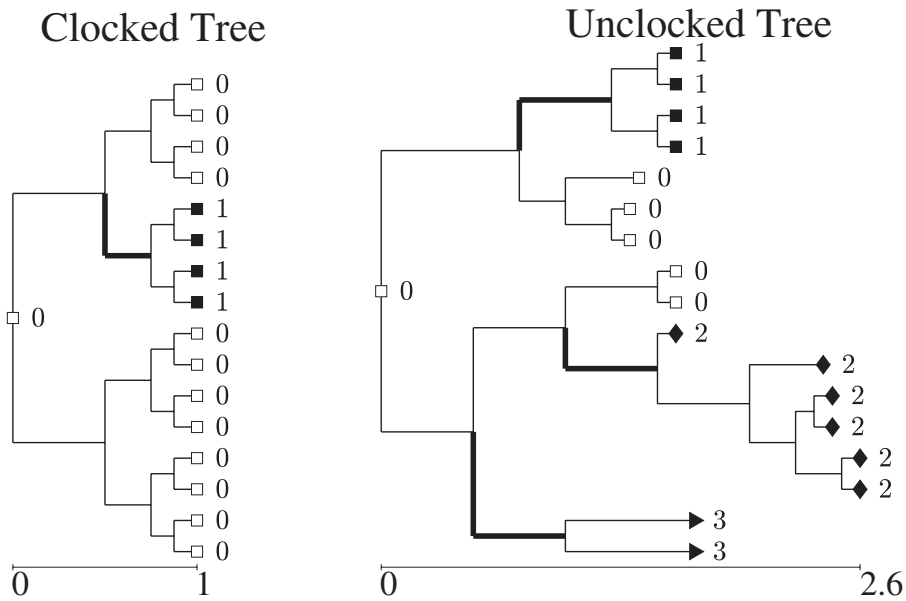

FIGURE 2. The clocked and unclocked trees used in Simulations 1, 2, and 3. Tree depths give the expected number of single nucleotide substitutions per codon. Symbols at the nodes indicate different phenotypes, with $k=2$ phenotypes (0 and 1) on the clocked tree and $k=4$ (0, 1, 2, and 3) on the unclocked tree.

alignments with realistic levels of heterotachy due to multiple processes. Our purpose was to assess the ability of the PG-BSM to identify among all variant site patterns those most consistent with one of the three modes of heterotachy represented in Figure 1. Our approach represents a significant improvement over traditional methods of model testing based on data generated using rate matrices of the form $Q(\omega)$. Details of all generating processes are provided in Supplementary material available on Dryad.

SIMULATIONS

In this section, we report the results of three simulation studies encompassing a wide variety of evolutionary scenarios. Simulation 1 was designed to test the statistical properties of the PG-BSM by fitting the model to alignments generated under the null PG-BSM. Simulation 2 was conducted to assess the impact of differences between the process assumed under the fitted model and the process used to generate the data (i.e., misspecification). For this purpose, alignments were generated using MSmmtDNA with 0% double–triple mutations or 6% double–triple mutations and MSTGdR with 0% double–triple mutations. In some cases, the alternate PG-BSM was fitted with the phenotype designated incorrectly. Simulation 3 was designed to assess the performance of the model under a scenario with four phenotypes. Increasing the number of phenotypes introduces greater uncertainty in the distribution of change maps and therefore represents a greater challenge to the model. Note that all simulations were conducted with $\pi_{rCW}=0$ (i.e., no sites were evolved under the rCW process), and in all that follows the alternate PG-BSM included the CW and BW processes unless otherwise indicated.

Simulation 1: Generating under the Null PG-BSM

According to maximum likelihood theory, when the PG-BSM is fitted to data generated under the null PG-BSM, and as information (i.e., the number of codon sites) increases without bound, 1) the distribution of the log-likelihood ratio for the contrast between the null and alternate PG-BSM will converge to some unknown mixture of the χ^2_0, χ^2_1 and χ^2_2 distributions (Self and Liang 1987), and 2) the distribution of the maximum likelihood estimate for each model parameter will converge to a normal centered on the parameter's generating value. The objective of the first simulation was to assess how well these expectations hold. To that end, 100 alignments 300 codons in length and 100 alignments 1000 codons in length were generated on the clocked and unclocked trees shown in Figure 2. The generating model was the null PG-BSM with the following parameters: $\pi_0=0.65$ (the proportion of sites evolving under $\omega_0=0$), $\omega_1=0.10$, $\omega_2=1.50$, $p_1=0.80$, $\delta=0.20$, $\pi_{CW}=\pi_{BW}=0$, where p_1 is the expected proportion of time a heterotachous site spends evolving under ω_1. The parameters for the mutation process, including position-specific nucleotide frequencies and the transition/transversion rate ratio, were set to values estimated from an alignment of 12 concatenated H-strand mitochondrial DNA sequences from 20 mammalian species (Yang 2007). The phenotypes assumed under the alternate PG-BSM were those indicated at the terminal nodes in Figure 2 (e.g., reading from top to bottom $\mathbf{F}=(0,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0)$ for the clocked tree and $\mathbf{F}=(1,1,1,1,0,0,0,0,0,2,2,2,2,2,2,3,3)$ for the unclocked tree).

Likelihood ratio tests were conducted assuming χ^2_2 to be the limiting distribution, when using 5.99 as the critical value would be expected to result in a nominal

TABLE 1. Simulation 1 log-likelihood ratio distributions

Scenario	0 to 0.50	0.50 to 5.99	5.99 to $+\infty$	False positive
PG-BSM C 300 sites	0.66	0.32	0.02	2/100
PG-BSM UC 300 sites	0.68	0.32	0.00	0/100
PG-BSM C 1000 sites	0.65	0.34	0.01	1/100
PG-BSM UC 1000 sites	0.78	0.22	0.00	0/100
Expectation under χ^2_2	0.22	0.73	0.05	5/100

Comparison of the empirical log-likelihood ratio distribution for Simulation 1 scenarios with the assumed χ^2_2 distribution. The last column shows the number of times the omnibus test incorrectly rejected the null to give a false positive. C indicates simulations using the clocked tree, UC simulations using the unclocked tree; 300 and 1000 indicate the number of simulated codon sites. One hundred alignments were generated under each simulation scenario.

TABLE 2. Site patterns identified by the YN-BSM

Clade	Site 64	Site 182	Site 153	Site 329	Site 239	Site 127
Marine (20)	S_{20}	S_{20}	N_{20}	$L_{12}M_5I_2V_1$	M_{20}	$L_{13}V_5A_1I_1$
Freshwater (8)	S_8	S_8	T_8	C_8	K_8	M_8
Terrestrial (9)	S_9	S_9	N_9	$L_7F_1A_1$	M_8L_1	L_7F_2
Intertidal (8)	S_8	S_8	N_8	$L_3A_3S_2$	M_8	L_6V_2
P(cat 2)	0.9970	0.9970	0.9610	0.9590	0.9530	0.9500

Amino acid composition for sites with P(cat 2) ≥ 0.95 as determined by the YN-BSM A using the branch leading to the freshwater clade as the foreground branch. Sites are shown in order of descending Bayes empirical Bayes posteriors. Letters represent amino acids and subscripts the number of taxa with that amino acid among the corresponding clade.

false positive rate of 5%. As the true limiting distribution of the log-likelihood ratio is an unknown mixture of χ^2_0, χ^2_1, and χ^2_2, the actual critical value corresponding to a 5% test is some unknown value less than 5.99. It follows that the expected false positive rate using 5.99 is less than 5%, and that assuming χ^2_2 makes the test conservative. The relative frequencies of the empirical log-likelihood ratio in each of the three intervals $[0, 0.50)$, $[0.50, 5.99)$, and $(5.99, +\infty)$ for all four simulation scenarios to that expected under the χ^2_2 distribution are shown in Table 1. Using the 5.99 cut-off gave a false positive rate of at most 2/100 among Simulation 1 scenarios. Furthermore, the relative frequencies in the $[0, 0.50)$ interval fell between 0.65 and 0.78 compared to the expected probability 0.22 for the χ^2_2 distribution. This result is not inconsistent with the fact that the actual limiting distribution places a substantial weight of 0.5 on the χ^2_0 component of the mixture (i.e., a point mass of 0.5 at zero; the weight is 0.5-p on the χ^2_1 distribution and p on the χ^2_2 distribution for some unknown $p \in [0, 0.5)$, Self and Liang 1987). The test therefore appears to be conservative as expected, at least under the scenarios considered. We nevertheless elected to use χ^2_2 for the remainder of our analyses as a buffer against inevitable misspecifications and/or issues associated with low information content, as is standard practice when an exact distribution is unknown (e.g., Wong et al. 2004; Zhang et al. 2005; Yang 2007, 2017).

The mean, median, and standard deviation of the maximum likelihood estimates of select model parameters for each generating scenario are shown in Table 1 in Supplementary material available on Dryad. In each case, the mean and median were either the same or nearly so, indicating symmetrical distributions. A one-sample Kolomogorov–Smirnov test for normality applied to each set of 100 maximum likelihood estimates failed to reject the null hypothesis of a normal distribution in all cases (P-value ≥ 0.16). Furthermore, the mean maximum likelihood estimate for each parameter was either the same or very close to its generating value in all four scenarios. And in each case the standard deviation was smaller for the 1000-codon scenario compared to its counterpart 300-codon scenario (see Figure 3). These results suggest that the PG-BSM is statistically well behaved when fitted to alignments generated under the scenarios considered.

Simulation 2: Generating under MSmmtDNA and MSTGdR

The second simulation study was conducted to assess the statistical accuracy and power of the PG-BSM when fitted to alignments simulated using a more complex generating model compared to the null PG-BSM. In particular, we aimed to generate alignments using the mutation-selection framework in such a way as to mimic realistic levels of heterotachy caused by nonadaptive shifting balance and episodic changes in site-specific fitness landscapes. The simulation is comprised of five scenarios, each of which was tested under three different sequence generating processes, yielding 15 cases in total (see Table 2 in Supplementary material available on Dryad). In the first scenario (denoted 2a $(\pi_{CW}, \pi_{BW}) = (0\%, 0\%)$), alignments were generated with no phenotype–genotype association, but with substantial heterotachy due to nonadaptive shifting balance. These alignments therefore contained signal for the covarion-like process that could potentially be misconstrued as signal for the CW and BW processes under the alternate PG-BSM. The second scenario (denoted 2b $(\pi_{CW}, \pi_{BW}) = (5\%, 0\%)$) included signal in the form of a small fraction ($\pi_{CW} = 5\%$ of 300 sites) of sites generated with a reduction in the stringency of selection. The third scenario (denoted 2c $(\pi_{CW}, \pi_{BW}) = (0\%, 5\%)$) included sites generated with peak shifts ($\pi_{BW} = 5\%$). In the fourth scenario (denoted 2d $(\pi_{CW}, \pi_{BW}) = (0\%, 0\%)$) we investigated the effect of phenotype error by using the data generated for 2c but with a misspecified vector of phenotypes. In this case, the data included heterotachy generated by peak shifts at 5% of sites that occurred independently of the assumed phenotype, and so was designated $(\pi_{CW}, \pi_{BW}) = (0\%, 0\%)$ to indicated no phenotype–genotype association. Signal for phenotype–genotype association was increased in the final scenario (denoted 2e $(\pi_{CW}, \pi_{BW}) = (5\%, 5\%)$) by including both sites generated with a reduction in the stringency of selection ($\pi_{CW} = 5\%$) and sites with peak

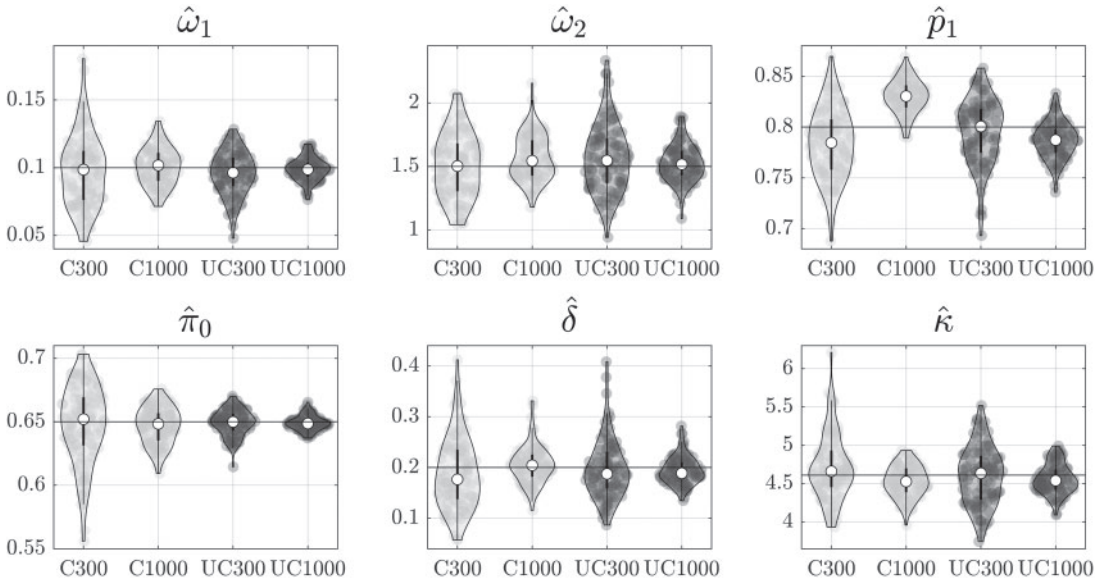

FIGURE 3. Violin plots for select maximum likelihood estimates obtained in Simulation 1. The first pair of plots in each panel (light grey) are for alignments generated on the clocked tree with 300 (C300) and 1000 (C1000) codon sites. The second pair of plots in each panel (dark grey) are for alignments generated on the unclocked tree with 300 (UC300) or 1000 (UC1000) codon sites. The varying width of each violin plot indicates a smoothed probability density. Data points are marked as disks with random horizontal offset to produce the cloud around each plot. The median parameter estimate is indicated by the white circle and the interquartile range by the thick vertical bar. The horizontal line in each panel shows the parameter value used to generate the alignments.

shifts ($\pi_{BW}=5\%$). The three generating processes used were: MSmmtDNA with 0% double–triple mutations, MSmmtDNA with 6% double–triple mutations, and MSTGdR with 0% double–triple mutations. In each case 50 alignments 300-codons in length were generated on the clocked tree in Figure 2. Changes in the stringency of selection and/or peak shifts were effected along the branch marked in bold. The correct phenotype designation was $\mathbf{F}=(0,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0)$, while the incorrect designation used in 2d was $\mathbf{F}=(0,0,0,0,0,0,0,0,0,1,1,1,1,1,0,0,0,0)$.

The omnibus test correctly failed to reject the null in all Scenario 2a ($\pi_{CW},\pi_{BW})=(0\%,0\%)$ trials (Table 2 in Supplementary material available on Dryad). In Scenario 2b ($\pi_{CW},\pi_{BW})=(5\%,0\%)$, the null was correctly rejected in 47/50, 46/50, and 42/50 trials, indicating good power, and the CW and BW processes were inferred at an average of (7%,1%), (7%,1%), and (4%,0%) of sites, in approximate agreement with their generating values. The agreement was also good in Scenario 2c ($\pi_{CW},\pi_{BW})=(0\%,5\%)$ where the null was rejected in 46/50, 38/50, and 50/50 trials, and the CW and BW processes were inferred at an average of (1%,9%), (2%,7%), and (1%,7%) of sites. The null was rejected in only 1/50, 1/50, and 2/50 trials, in Scenario 2d ($\pi_{CW},\pi_{BW})=(0\%,0\%)$, well below the expected 5% error rate. And in Scenario 2e ($\pi_{CW},\pi_{BW})=(5\%,5\%)$ the null was rejected in all trials and the CW and BW processes were inferred at an average of (6%,9%), (6%,8%), and (6%,6%) of sites.

Results of the *post hoc* analysis applied to alignments with signal for phenotype–genotype association (Scenarios 2b, 2c, and 2e) are summarized in Table 3 in Supplementary material available on Dryad. Analyses were conducted with the expected false discovery count limited to $E\{FDC\}=1$ for each process $c\in\{CW,BW\}$. For Scenario 2b ($\pi_{CW},\pi_{BW})=(5\%,0\%)$ MSmmtDNA 0% DT (first row of Table 3), for example, the average FDC was 0.80 CW sites and 0.86 BW sites per alignment, and the average power to detect the 15 (5% of 300) CW sites generated with a reduction in the stringency of selection was 0.31, corresponding to an average of 4.72/15 correctly identified sites per alignment. The FDCs among all scenarios were approximately normal in distribution and ranged between 0.62 and 2.12, with a mean of 1.32 and standard deviation of 0.44. The FDCs were therefore slightly biased toward values greater than the nominal expectation $E\{FDC\}=1$. Note that a FDC of 1.32 corresponds to a false positive rate of 1.32/270 sites $\times 100\%=0.49\%$ for the ($\pi_{CW},\pi_{BW})=(5\%,5\%)$ scenario.

TABLE 3. Site patterns identified by the PG-BSM

Clade	Site 153	Site 144	Site 25	Site 239	Site 182	Site 64
Marine (20)	N_{20}	G_{20}	L_{20}	M_{20}	S_{20}	S_{20}
Freshwater (8)	T_8	S_8	F_8	K_8	S_8	S_8
Terrestrial (9)	N_9	G_9	I_5L_4	M_8L_1	S_9	S_9
Intertidal (8)	N_8	G_8	L_8	M_8	S_8	S_8
P(BW)	0.9859	0.9799	0.9458	0.9433	0.9221	0.7509

Amino acid composition for the first six sites identified by the PG-BSM to be associated with the freshwater versus other phenotype. Sites are shown in order of descending P(BW). Letters represent amino acids and subscripts the number of taxa with that amino acid among the corresponding clade.

The model performed well with respect to identifying the correct evolutionary history of the phenotype (see the last two columns of Table 3 in Supplementary material available on Dryad). The change map $\mathbf{z}^*$ corresponding to the most frequently sampled history matched that used to generate the alignment in no less than 44/50 trials and often in 50/50 trials. Furthermore, the probability $\hat{P}(\mathbf{z}^* \mid X, \mathbf{F})$ of $\mathbf{z}^*$ conditioned on all of the data was always greater than its relative sampling frequency $\hat{\pi}_{\mathbf{z}^*}$, with average differences ranging between $0.22 < \hat{P}(\mathbf{z}^* \mid X, \mathbf{F}) - \hat{\pi}_{\mathbf{z}^*} < 0.27$. This result illustrates how accounting for phenotype–genotype associations can substantially reduce uncertainty in the inferred history of the phenotype whenever such associations exist.

Simulation 3: Generating under MSmmtDNA with Four Phenotypic States

The third simulation study was conducted to assess the statistical accuracy and power of the PG-BSM under scenarios with four phenotypic states. In this case, we used only MSmmtDNA with 0% double–triple mutations to generate data because the results of Simulation 2 indicated no substantial difference between the three mutation-selection generating processes used there (as reported in Supplementary material available on Dryad). Fifty 300-codon alignments were generated on the unclocked tree in Figure 2 under four scenarios. In the first (denoted 3a $(\pi_{CW}, \pi_{BW}) = (0\%, 0\%)$) alignments were generated with no phenotype–genotype association. In the second (denoted 3b $(\pi_{CW}, \pi_{BW}) = (5\%, 0\%)$) alignments were generated with a reduction in the stringency of selection at 5% of sites. In the third (denoted 3c $(\pi_{CW}, \pi_{BW}) = (0\%, 5\%)$) alignments were generated with peak shifts at 5% of sites. And in the last scenario (denoted 3d $(\pi_{CW}, \pi_{BW}) = (5\%, 5\%)$) alignments were generated with both a reduction in the stringency of selection at 5% of sites and peak shifts at 5% of sites. In all cases, the stringency of selection and/or peak shifts were effected along the branches marked in bold. The phenotype assumed by the alternate PG-BSM was $\mathbf{F} = (1,1,1,1,0,0,0,0,0,0,2,2,2,2,2,2,3,3)$ in all scenarios. The unclocked tree is arguably more consistent with real data in both its irregular topology and depth compared to the clocked tree in Figure 2, and was chosen, in combination with the increase in the number of phenotypes, to provide a more challenging test of model performance.

The omnibus test correctly failed to reject the null hypothesis in all Scenario 3a $(\pi_{CW}, \pi_{BW}) = (0\%, 0\%)$ trials under which alignments were generated with no phenotype–genotype association (Table 4 in Supplementary material available on Dryad). The null was correctly rejected in all Scenario 3c$(\pi_{CW}, \pi_{BW}) = (0\%, 5\%)$ and Scenario 3d $(\pi_{CW}, \pi_{BW}) = (5\%, 5\%)$ trials. However, in Scenario 3b $(\pi_{CW}, \pi_{BW}) = (5\%, 0\%)$ the null was correctly rejected in only 15/50 trials. The PG-BSM apparently had difficulty identifying sites generated with a reduction in the stringency of selection. Concordantly, the average power to detect CW sites was 0.16 for Scenario 3b $(\pi_{CW}, \pi_{BW}) = (5\%, 0\%)$ alignments and 0.36 for Scenario 3d $(\pi_{CW}, \pi_{BW}) = (5\%, 5\%)$ alignments, much less than the average power to detect BW sites, which was 0.67 for both Scenarios 3c $(\pi_{CW}, \pi_{BW}) = (0\%, 5\%)$ and 3d $(\pi_{CW}, \pi_{BW}) = (5\%, 5\%)$. The power to detect BW sites was substantially better in Simulation 3 (power = 0.67) compared to Simulation 2 ($0.24 \leq$ power ≤ 0.59), and the FDCs for both CW and BW sites were significantly lower (no more than 0.78 false discoveries per alignment among Simulation 3 scenarios compared to as much as 2.12 among Simulation 2 scenarios). There was also a marked increase in uncertainty in the ancestral phenotypes in Simulation 3, with an average $0.38 \leq \hat{\pi}_{\mathbf{z}^*} \leq 0.56$ compared to $0.62 \leq \hat{\pi}_{\mathbf{z}^*} \leq 0.78$ in Simulation 2. However, use of the combined information in the data made up the difference, as the probability $\hat{P}(\mathbf{z}^* \mid X, \mathbf{F})$ of $\mathbf{z}^*$ conditioned on all of the data was approximately twice as large as $\hat{\pi}_{\mathbf{z}^*}$ in all Simulation 3 scenarios with phenotype–genotype association—see (prior, post) in Table 5 in Supplementary material available on Dryad.

A possible explanation for the low power of the omnibus test in Scenario 3b $(\pi_{CW}, \pi_{BW}) = (5\%, 0\%)$ is confounding due to "branch-length effects." Two alignment-generating processes (real or simulated) are said to be nearly confounded if the site-pattern distributions they produce are approximately the same (Jones et al. 2018). Sites evolved under MSmmtDNA on the unclocked tree in Figure 2 with relaxation of selection pressure along the three branches marked in bold tend to produce site patterns consistent with the phenomenological CW process (i.e., with greater diversity among amino acids at the terminal nodes indicated by the filled markers and less diversity among terminal nodes indicated by the open marker). But similar patterns can arise on that tree at sites evolved on static fitness landscapes due to the fact that the distances from the root to the terminal nodes indicated by the filled markers are relatively long (increasing the probability of replacement substitutions) whereas the tip-to-root distances for terminal nodes indicated by the open marker are relatively short (decreasing the probability of replacement substitutions). Heterotachous site patterns x^h generated with relaxation of selection pressure in Scenario 3b $(\pi_{CW}, \pi_{BW}) = (5\%, 0\%)$ therefore tended to be approximately as likely under the covarion-like process as they were under the CW process. The log-likelihood ratio for the contrast between the null and alternate PG-BSM consequently tended to be small, and often something less than the critical value 5.99 (assuming the χ^2_2 distribution for the log-likelihood ratio and a 5% significance test). Note that modifying the alternate PG-BSM to test for the CW process only, which permits the use of the χ^2_1 distribution for the log-likelihood ratio and a critical value of 3.84 for a 5% test, increased the power of the omnibus test only slightly (19/50 rejections instead of 15/50).

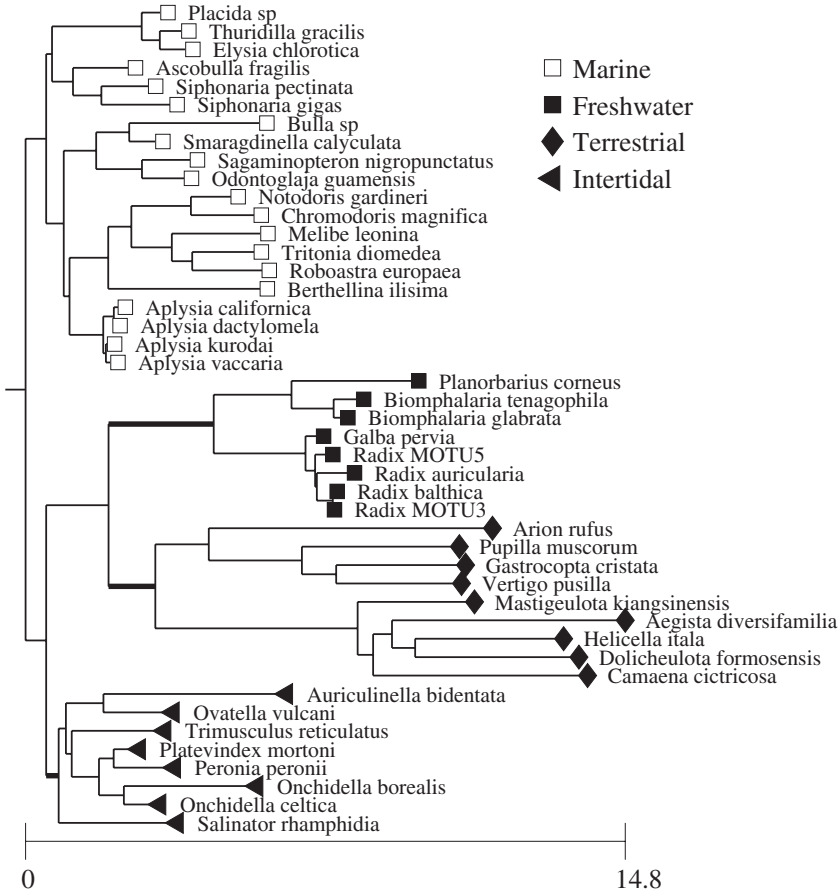

FIGURE 4. Branch lengths (the expected number of single nucleotide substitutions per codon) estimated by fitting the null PG-BSM to the cytochrome B alignment.

INVERTEBRATE CYTOCHROME B

We now turn to an analysis of real data. Euthyneura (snails and slugs) have adapted to diverse habitats, including marine, intertidal, terrestrial, and freshwater. Their mitochondrial genome includes cytochrome B (cyt B), an essential component of the electron transport chain common to most life forms on Earth. Given its crucial role, one would expect cytB to be highly conserved. It is nevertheless reasonable to suspect that transition from the marine to the other three environments might have required some adaptations (e.g., for differences in osmotic pressure or the risk of dessication). To test this hypothesis, the PG-BSM and YN-BSM A were fitted to an alignment consisting of 45 cytB sequences 341 codons in length. The sequences were selected from a larger published data set (Romero et al. 2016) to produce four homogeneous clades. Tree topology was estimated from the DNA sequences using RAxMLv0.6.0 with default settings, and the tree was rooted to produce that shown in Figure 4.

The PG-BSM was initially fitted to the alignment using the different environments to define a phenotype with four states, but no signal for phenotype–genotype association was found. The data were then reanalyzed using three different binary phenotypes: terrestrial versus nonterrestrial, freshwater versus nonfreshwater, and intertidal versus nonintertidal. Furthermore, rather than seeking to detect CW and BW sites simultaneously, we elected to attempt to detect either CW, rCW, or BW sites alone. The YN-BSM A was also fitted to the alignment using the branch leading to the terrestrial, freshwater, and intertidal clade each in turn as the foreground (marked in bold in Figure 4). Signal was detected for BW sites by the PG-BSM when phenotype was set to freshwater versus nonfreshwater, with log-likelihood ratio $=2(24,537-24,527)=20$ compared to a critical value of 5.73 (assuming that the log-likelihood ratio follows a χ^2_1 distribution and using a level of significance $\alpha=0.05/3$ to adjust for the fact that three tests were conducted on the alignment with freshwater vs. nonfreshwater as the phenotype). The maximum likelihood estimates were $\hat{\omega}_1=0.00$, $\hat{\omega}_2=0.08$, $\hat{p}_1=0.58$, $\hat{\delta}=0.06$, and $(\hat{\pi}_0,\hat{\pi}_{CL},\hat{\pi}_{BW})=(0.34,0.60,0.06)$. The model therefore inferred that 34% of sites evolved under $\omega_0=0$, 60% of sites evolved under the covarion-like process with random switching between $\hat{\omega}_1=0.00$ and $\hat{\omega}_2=0.08$, and 6% of sites evolved under the BW process with $\hat{\omega}_1=0.00$ everywhere in the tree except along

the branch leading to the freshwater clade, where the rate ratio was elevated slightly to $\hat{\omega}_2 = 0.08$. The YN-BSM A detected evidence of positive selection in two cases, once along the branch leading to the terrestrial clade (log-likelihood ratio $= 2(25,819 - 25,812) = 14$, $\hat{\omega}_2 = 999$, $\hat{p}_2 = 0.04$) and again along the branch leading to the freshwater clade (log-likelihood ratio $= 2(25,811 - 25,807) = 8$, $\hat{\omega}_2 = 999$, $\hat{p}_2 = 0.12$). Akaike's Information Criterion (AIC $= 2m - 2$LL, where m is the number of estimated model parameters and LL is the log-likelihood of the data under the fitted model) is frequently used to compare non-nested models under the maximum likelihood framework. The better of any two models is the one with the smaller AIC. The PG-BSM provided the better fit by this criterion: the criterion was 51,700 for YN-BSM A but 49,146 for the alternate PG-BSM with BW sites alone using the freshwater versus nonfreshwater phenotype.

It is instructive to compare the distribution of amino acids among each clade at sites identified by the YN-BSM A and PG-BSM via *post hoc* analysis. The YN-BSM A identified 11 sites with strong evidence of having undergone episodic positive selection on the branch leading to the freshwater clade after *post hoc* analysis was conducted with the expected FDC set to E{FDC} = 1. Table 2 shows the distribution of the amino acids at six of those sites for which P(cat 2) ≥ 0.95, where P(cat 2) is the Bayes empirical Bayes posterior probability (Yang et al. 2005). Site 329, for example, is occupied by four amino acids among the 20 taxa in the marine clade, 12 by L (Leucine), 5 by M (Methionine), 2 by I (Isolucine), and 1 by V (Valine). A comparison of distributions across clades gives some clue as to the processes that might have generated the data. Sites 153 and 239 exhibit one amino acid among the freshwater clade (T at site 153 and K at site 239) but are dominated by a different amino acid among the other three clades (N at site 153 and M at site 239). These patterns are consistent with peak shifts along the branch leading to the freshwater clade (i.e., the BW process). Sites 329 and 127 show one amino acid among the freshwater clade and two or more different amino acids among each of the remaining clades. These sites are consistent with intensification of selective constraint in the freshwater clade (i.e., the rCW process) possibly accompanied by a peak shift (since the amino acid in the freshwater clade, C at site 329 and M at site 127, does not occur in any of the other three clades).

The PG-BSM fitted using freshwater versus nonfreshwater as the phenotype detected seven sites with $0.52 \leq$ P(BW) ≤ 0.99 after *post hoc* analysis was conducted with the expected FDC set to E{FDC} = 1. The first six of these are shown in Table 3. The first four sites (153, 144, 25, and 239) are highly consistent with the BW process, being dominated by one amino acid among the freshwater clade and a different amino acid among the nonfreshwater clades. Sites 182 and 64 are both occupied by serine only. There are eight codon aliases for serine in the invertebrate mtDNA code, including TCN and AGN where N is any nucleotide. Paths between TCN and AGN by single nucleotide substitutions require a minimum of one nonsynonymous change to either tryptophan, cystine, or threonine. The fact that sites 64 and 182 were identified in the *post hoc* analysis is explained by the codons that appear within each clade: both sites are occupied by AGN everywhere in the freshwater clade but are dominated by TCN among all remaining taxa. This suggests that substitutions to intermediate amino acids occurred along the branch leading to the freshwater clade. Note that the YN BSM A assigned the largest posterior to sites 64 and 182, and also assigned them equal probability P(cat 2) = 0.9970 despite the fact that the two sites have different codon substitution patterns (data not shown). The PG-BSM, in comparison, placed less weight on these sites and was apparently sensitive to their differences, since P(BW) = 0.9221 for site 182 but only P(BW) = 0.7509 for site 64.

Accounting for heterotachy prevents false positives

The PG-BSM was used to analyze two other real alignments. The first consisted of 12 concatenated sequences of mammalian mtDNA taken from 20 species (Yang, 2007). This alignment is characterized by a long branch separating 7 primate species from 13 nonprimate species. Our analysis of the mtDNA alignment revealed the potential impact of accounting for nonstationary CW and BW processes on branch-length estimates (i.e., by making the tree more clock-like), and also that the PG-BSM can be robust to confounding by what we call branch-length effects. Details of that analysis can be found in Supplementary material available on Dryad. The second alignment consisted of genes for various forms of phytochrome. The same data were used to illustrate the efficacy of the YN-BSM (Yang and Nielsen 2002; Zhang et al. 2005) and therefore has historical significance. Our analysis of the phytochrome data motivated simulations that revealed that accounting for heterotachy using the covarion-like component of the PG-BSM can be essential to prevent false inference of phenotype–genotype association. The results of that simulation study are presented here, whereas additional components of that analysis are reported in Supplementary material available on Dryad.

The PG-BSM did not detect phenotype–genotype association in the phytochrome data (15 sequences 1072 codons in length) despite the presence of site patterns consistent with both the CW and BW processes. We speculated that the data might contain true signal that went undetected due to the unusually large proportion (approximately 70%) of variable sites (i.e., site patterns with 2 or more amino acids). To test this hypothesis, a fourth simulation was conducted under which MSmmtDNA was used to generate sets of 50 alignments 1072 codons in length on the phytochrome tree. The proportion of variable sites can be controlled under MSmmtDNA by changing the proportion of sites with landscapes that admit nonadaptive shifting balance (i.e., landscapes with a selection regime somewhere between stringent and neutral, Jones et al. 2017). Alignments under scenarios 4a and 4b were generated with either

≈40% or ≈70% variable sites, including 5% CW and 5% BW sites. Alignments under Scenarios 4c and 4d were generated with either ≈40% or ≈70% variable sites with no phenotype–genotype association.

The PG-BSM correctly detected phenotype–genotype association in 50/50 alignments generated with 40% variable sites (scenario 4a $(\pi_{CW}, \pi_{BW}) = (5\%, 5\%)$), but in only 42/50 alignments generated with 70% variable sites (scenario 4b $(\pi_{CW}, \pi_{BW}) = (5\%, 5\%)$). Although 42/50 indicates substantial statistical power, the reduction in the number of detections is consistent with our hypothesis that the power of the PG-BSM can be reduced when the proportion of variable sites is high. The YN-BSM A inferred positive selection at some sites along the foreground branch in all trials under both of these scenarios. The PG-BSM produced 0/50 false positives when there was no phenotype–genotype association regardless of the proportion of variable sites. The YN-BSM A, by contrast, inferred positive selection (i.e., $\omega_2 > 1$) at some sites in 11/50 alignments generated with 40% variable sites (scenario 4c $(\pi_{CW}, \pi_{BW}) = (0\%, 0\%)$) and in 31/50 alignments generated with 70% variable sites (scenario 4d $(\pi_{CW}, \pi_{BW}) = (0\%, 0\%)$). Some of these might be true evidence of $\omega > 1$ at some sites over the foreground branch since positive selection due to shifting balance is expected to occur some of the time (Jones et al. 2017). However, they are all false positives when interpreted as evidence of adaptive evolution (e.g., a change in the protein's function) because the data were generated with static site-specific fitness landscapes.

The PG-BSM was specifically designed to account for sources of heterotachy other than the three mechanisms of phenotype–genotype association by including the covarion-like component of the model. The importance of this component is illustrated by fixing $\delta = 0$ and fitting the resulting modified PG-BSM to Scenario 4d alignments (70% variable sites, no phenotype–genotype association). Setting the switching rate to zero has the effect of making CLM3($k = 2$) equivalent to M3($k = 2$), since $\omega_1 < \omega_2$ are still estimated but sites can no longer switch between them. Sites most consistent with M3($k = 2$) are those that evolved at a constant rate over the tree. The modified version of the alternate PG-BSM can therefore accommodate heterotachous sites only by appealing to the CW and BW processes. The modified PG-BSM incorrectly inferred phenotype–genotype association in 33/50 trials when fitted to Scenario 4d alignments compared to 0/50 for the regular PG-BSM. This result demonstrates the utility of accounting for heterotachy with the covarion-like process as a mean to mitigate false detection of phenotype–genotype associations.

Discussion

Traditional branch-site codon substitution models (Yang et al. 2005) provide a means to detect evidence that a codon site underwent positive selection along a specified foreground branch of a phylogeny. Such evidence, in the form of an estimated rate ratio $\omega > 1$, is widely considered sufficient to infer adaptive evolution at a codon site. However, $\omega > 1$ does not necessarily imply adaptation. It is true that the dynamic at a codon site following a peak shift is characterized by a transient increase in the expected rate ratio, and that the increase can sometimes be to $\omega > 1$ (dos Reis 2015). But the same can also occur on a static fitness landscape following chance fixation to a less-than-optimal amino acid (i.e., by nonadaptive shifting balance, Jones et al. 2017). It is therefore not possible to distinguish an episodic change in a site-specific landscape from nonadaptive shifting balance on a static landscape using estimates of ω alone. Furthermore, adaptation does not necessarily imply $\omega > 1$. The increase in the rate ratio following a peak shift rapidly diminishes as the site moves toward its new fitness peak (dos Reis 2015). This suggests that the initial elevation in rate ratio can be more difficult to detect as sequences become more divergent. A previous analysis of the relationship between branch length and $\hat{\omega}$ estimated from pairs of sequences simulated under the mutation-selection framework supports this intuition. Peak shifts were implemented by simultaneously changing the fitness coefficients at all 1000 codon sites in an initial sequence S_1 that was subsequently evolved over a branch of length b to obtain a second sequence S_2. The codon substitution model M0 (i.e., a model that estimates a single rate ratio for all sites in an alignment, (Nielsen and Yang 1998) was then fitted to (S_1, S_2) to obtain $\hat{\omega}$. The median estimate across 200 trials was $\hat{\omega} \approx 1.4$ when $b = 0.2$, but $\hat{\omega} \approx 1.0$ when $b = 1.0$ (Jones et al. 2017). Hence, $\omega > 1$ does not imply adaptive evolution and nor does adaptive evolution imply $\omega > 1$.

The PG-BSM provides an approach for inferring adaptation that does not rely on the canonical $\omega > 1$ signature of positive selection. The method is based on the supposition that mechanisms of adaptation at the molecular level consist of changes in site-specific fitness landscapes. The mechanisms considered in this study consisted of either a persistent change in the stringency of selection at a site or a peak shift at a site along a branch of the tree. Changes in stringency are represented as CW changes in rate ratio, whereas a peak shift is represented by the BW process as a transient elevation in rate ratio along specific branches of the tree. The locations of branches over which these processes may have occurred are informed by a discrete character state (e.g., a phenotype) via a model for the evolution of that character state. This constraint provides additional information that makes it possible to identify among all variable sites those with replacement patterns that imply phenotype–genotype association. It is possible to use a model with asymmetric rates something like the generalized time-reversible model for the evolution of the discrete phenotypic state. However, the generalized time-reversible model is equivalent to the proportional rates model when there are only two phenotypic states, as was the case for most of the data sets used in this study. The simpler proportional rates model was therefore used throughout. Nevertheless, the addition of parameters to account for asymmetric transition rates

can potentially be useful for data sets with many taxa and more than two phenotypic states. The PG-BSM also includes a covarion-like component to account for variant site patterns inconsistent with phenotype–genotype association. This component provides the null hypothesis, which is rejected by the presence of site patterns that are more likely to have occurred under one of the CW, rCW, or BW processes.

The PG-BSM framework offers several advantages over its predecessor, the YN-BSM. First, it includes a model for the evolution of a discrete phenotype that not only frees the analyst from the task of specifying foreground branches but also automatically takes into account less likely but nevertheless possible evolutionary histories of the phenotype. Second, it includes a covarion-like component to account for random shifts between $\omega_1 < \omega_2$ consistent with all processes that can potentially result in heterotachy, including changes in site-specific landscapes not associated with changes in phenotype. Covarion-like models (e.g., Galtier 2001; Guindon et al. 2004) were originally intended to account for epistatic interactions between codons sites thought to be the cause of the covarion (i.e., concomitantly variable codons, Fitch and Markowitz 1970; Fitch 1971) phenomenon. It is now understood that potential sources of heterotachy include nonadaptive shifting balance and the fixation of double–triple mutations in addition to episodic changes in site-specific fitness landscapes (Jones et al. 2018; Venkat et al. 2018). The utility of using the covarion-like model as the null hypothesis was illustrated in our simulations of the phytochrome alignment, where its inclusion as a component of the PG-BSM was instrumental in reducing the false positive rate of the omnibus test. Third, pathologies such as false positives that can sometimes arise under the YN-BSM due to statistical irregularities (e.g., Baker et al. 2016; Mingrone et al. 2018) are avoided under the PG-BSM. The YN-BSM assumes that category 2 sites evolved under a separate rate ratio ω_2 on the foreground. The rate ratio ω_2 is consequently nearly unidentifiable when p_2 is small. Under this irregular condition, the maximum-likelihood estimate $\hat{\omega}_2$ is sometimes very large and potentially misleading (e.g., in our analysis of cytB, the YN-BSM A yielded $\hat{\omega}_2 = 999$ with $\hat{p}_2 = 0.04$ or $\hat{p}_2 = 0.12$). This issue is avoided under the PG-BSM because estimates of ω_1 and ω_2 make use of information contained in all variable sites. Fourth, the PG-BSM can identify sites consistent with specific mechanisms of adaptation without a test for positive selection. This key feature was empirically validated by our simulation studies, where the null hypothesis was correctly rejected for the majority of alignments generated with changes in site-specific landscapes. Moreover, a fair proportion of sites generated under specific mechanisms (relaxation or intensification in the stringency of selection, a peak shift) were correctly identified via *post hoc* analysis.

The chance fixation into the tail of a static site-specific landscape and an adaptive change in a site-specific landscape both cause a site to be temporarily occupied by a less-than-optimal amino acid, say B. In either case the result is a transient increase in rate ratio to some value ω_B that decays exponentially while positive selection drives the site from B to the fittest amino acid A. Once A is fixed the rate ratio stabilizes to some value $\omega_A < \omega_B$. These processes can manifest across the sites in an alignment as covarion-like switching between $\hat{\omega}_1 < \hat{\omega}_2$. The magnitude of $\hat{\omega}_2$ depends on the distribution of the ω_B, which in turn depends on the magnitude of the selection coefficients $s_{BA} = f_A - f_B > 0$. In Simulations 2 and 3, where sites were evolved using models based on the mutation-selection framework, the mean value of $\hat{\omega}_2$ was never less than one. This indicates that s_{BA} tended to be large enough to make the $\omega_B > 1$. In the cytochrome data, the rate ratio was only $\hat{\omega}_2 = 0.08$. Sites in real proteins are undoubtedly subject to both intragenic (e.g., Pollock et al. 2012; Starr and Thornton 2016) and intergenic (e.g., Phillips 2008) epistatic constraints. These can be difficult to model because they depend on unique aspects of the structure and function of a given protein as well as the nature of its interactions with other proteins. These and other potential sources of constraint are therefore absent in the majority of generating models used in simulation studies (e.g., Anisimova et al. 2001, 2002; Wong et al. 2004; Zhang 2004; Kosakovsky Pond and Frost 2005; Yang et al. 2005; Zhang et al. 2005; Yang and dos Reis 2011; Kosakovsky Pond et al. 2011; Lu and Guindon 2013), including those used in our study. Such constraints might have the effect pushing the s_{BA} closer to zero. For example, there is evidence that epistasis can cause the magnitude of s_{BA} at a site to diminish over time due to compensating substitutions at other sites (e.g., via an evolutionary Stokes shift, Pollock et al. 2012). This can have the overall effect of reducing the ω_B. Differences in the depth of the tree might also have played a role in lowering $\hat{\omega}_2$, since the cytochrome alignment was considerably more divergent than the simulated alignments, and estimates of ω tend to diminish with larger divergences (dos Reis and Yang 2013; Jones et al. 2017). It is noteworthy, however, that the PG-BSM detected evidence of adaptive evolution in the cytochrome alignment despite the small estimate of ω_2. This was possible only because the model was designed to identify patterns of change in ω consistent with specific mechanisms of adaptation without imposing bounds on the magnitude of ω_2.

Like the vast majority of codon substitution models, the YN-BSM framework assumes evolution occurs via a series of single nucleotide substitutions. Consequently, whether or not a site is inferred to have undergone positive selection depends in part on the codon distribution implicitly inferred by the pruning algorithm (Felsenstein 1981) at the two nodes of the foreground branch. Positive selection is more often inferred when the codons that most likely occupied those two nodes differ by more than one nucleotide. Indeed, it was recently shown that the majority of support for positive selection in real data under the YN-BSM A consists of sites patterns that suggest multiple single nucleotide

substitutions along the foreground (Venkat et al. 2018). Yet instantaneous double and triple mutations can occur at a rate recently estimated to be roughly 1% to 3% of all mutations (Keightley et al. 2009; Schrider et al. 2014; De Maio et al. 2013; Harris and Nielsen 2014). The chance fixation of a double–triple mutation along the foreground can only be misconstrued by the YN-BSM A as evidence of multiple single nucleotide substitutions. Hence, positive selection was often falsely inferred by the YN-BSM A in alignments generated with the fixation of rare double–triple mutations (Venkat et al. 2018). It follows that positive selection due to genuine episodic peak shifts can be confounded not only by nonadaptive shifting balance (Jones et al. 2017), but also by the fixation of double–triple mutations (Venkat et al. 2018). The PG-BSM was specifically formulated with the understanding that evidence of positive selection in the form of $\omega > 1$ can result from multiple processes, some of which are nonadaptive. This was the point of the move away from the standard $\omega > 1$ paradigm. The PG-BSM is apparently robust to double–triple mutations since, although the inclusion of 6% double–triple mutations resulted in larger $\hat{\omega}_2$ compared to simulations with 0% DT, the omnibus test never incorrectly rejected the null.

The current trend in model development is toward greater realism via the addition of parameters that represent specific mechanistic processes (e.g., Liberles et al., 2013; Zaheri et al., 2014; Pollock et al., 2017; Venkat et al., 2018). It is gradually becoming clear that this approach is not guaranteed to give better models. Under the maximum likelihood framework, the addition of any parameter ψ to a null model M will always result in a better fit (i.e., a larger likelihood). To guard against a spurious increase in likelihood, the null is rejected only if the log-likelihood ratio comparing the null model without ψ with the alternate model that includes ψ is greater than a prespecified threshold chosen to limit the false positive rate to some maximum upper bound (e.g., 5%). The trend toward realism implicitly assumes that rejection of the null can be interpreted to mean that the model with ψ provides a better representation of the actual data-generating process than the model without ψ. It has been recently pointed out by several authors that this assumption can be invalidated by hidden variables and confounding (e.g., Beaulieu and O'Meara 2016; Caetano et al. 2018; Jones et al. 2018). Suppose ψ represents process P_1, which did not occur when the data were generated. Further, suppose process P_2 did occur when the data were generated, and that P_2 tends to produce patterns in the data similar to P_1 (i.e., P_1 and P_2 are confounded, Jones et al. 2018). Rejection of the null under this scenario is likely because ψ can account for variations in the data generated by process P_2. And rejection would be correct as an indication that the inclusion of ψ improved model fit. But it would also lead to the false conclusion that process P_1 actually occurred. When this happens we say that $\hat{\psi}$ carries phenomenological load (Jones et al. 2018).

The covarion-like component of the PG-BSM confers some robustness against phenomenological load. Under the null PG-BSM, the covarion-like model accounts for all mechanisms that might generate heterotachy, whether adaptive (i.e., episodic peak shifts) or nonadaptive (shifting balance, fixation of double–triple mutations, epistasis). Hence, the parameters $\langle \omega_1, \omega_2, p_1, \delta \rangle$ for the covarion-like process account for multiple mechanisms. By contrast, the parameters α and β in equation (1) have specific mechanistic interpretations as the rate at which double and triple nucleotide mutations occur. It was recently suggested that existing codon substitution model should be modified to account for the possible fixation of double–triple mutations (Venkat et al. 2018). However, models that include α and β as estimated parameters can result in false detection of fixation of double–triple mutations due to phenomenological load (Jones et al. 2018). This is avoided under the PG-BSM by allowing that the maximum likelihood estimates for $\langle \omega_1, \omega_2, p_1, \delta \rangle$ result from an unknown combination of mechanisms, including the fixation of double–triple mutations. Hence, for example, finding that $\hat{\delta}$ is significantly > 0 in a contrast between the null PG-BSM with $\delta = 0$ versus the null PG-BSM with δ estimated need not be interpreted as evidence for any particular mechanism of heterotachy, but only for "heterotachy-by-any-cause." In this way, the possibility of the fixation of double–triple mutations is subsumed in the parameters for the covarion-like process, and false conclusions due to the confounding of processes are avoided. The PG-BSM framework therefore not only provides a means to identify site patterns consistent with specific adaptive mechanisms, but through the addition of external phenotypic information also offers a solution to several recently discovered problems associated with confounding and phenomenological load (Jones et al. 2017, 2018; Venkat et al. 2018).

Software

Software for the methods is available at: https://www.mathstat.dal.ca/~tsusko/software.html.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.rb4b420.

Funding

This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to E.S. and J.P.B.

References

Adams, D.C., Collyer M.L. 2018. Multivariate phylogenetic comparative methods: evaluations, comparisons, and recommendations. Syst. Biol. 67:14–31.

Anisimova M., Bielawski J.P., Yang Z.H. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18:1585–1592.
Anisimova M., Bielawski J.P., Yang, Z.H. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. 19:950–958.
Baker J.E., Dunn K.A., Mingrone J.M., Wood B.A., Karpinski B.A., Sherwood C.C., Wildman D.E., Maynard T.M., Bielawski J.P. 2016. Functional divergence of the nuclear receptor nr2c1 as a modulator of pluripotentiality during hominid evolution. Genetics 203:905–922.
Beaulieu J.M., O'Meara B.C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst. Biol. 65:583–601.
Benjamini Y., Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B 57:289–300.
Butler M.A., King A.A. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am. Nat. 164:683–695.
Caetano D.S., O'Meara B.C., Beaulieu J.M. 2018. Hidden state models improve state-dependent diversification approaches, including biogeographical models. Evolution 72:2308–2324.
Cornwell W., Nakagawa S. 2017. Phylogenetic comparative methods. Curr. Biol. 27:327–338.
De Maio N., Holmes I., Schlötterer C., Kosiol C. 2013. Estimating empirical codon hidden Markov models. Mol. Biol. Evol. 30:725–736.
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.
dos Reis M., Yang Z.H. 2013. Why do more divergent sequences produce smaller nonsynonymous/synonymous rate ratios in pairwise sequence comparisons. Genetics 195:195–204.
Eastman J.M., Alfaro M.E., Joyce P., Hipp A.L., Harmon L.J. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65:3578–3589.
Felsenstein J.J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. 25:471–492.
Felsenstein J.J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.
Fitch W. 1971. The nonidentity of invariable positions in the cytochrome c of different species. Biochem. Genet. 5:231–241.
Fitch W., Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4:579–593.
Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18:866–873.
Gaston D., Susko E., Roger A.J. 2011. A phylogenetic mixture model for the identification of functionally divergent protein residues. Bioinformatics 27:2655–2663.
Goldman N., Yang Z.H. 1994. Codon-based model of nucleotide substitution for protein-coding DNA-sequences. Mol. Biol. Evol. 11:725–736.
Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. 16:1664–1674.
Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol. 18:453–464.
Gu X. 2006. A simple statistical model for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol. Biol. Evol. 23:1937–1945.
Guindon S., Rodrigo A.G., Dyer K.A., Huelsenbeck J.P. 2004. Modeling the site-specific variation of selection patterns along lineages. Proc. Natl. Acad. Sci. USA 101:12957–12962.
Halpern A.L., Bruno, W.J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. 15:910–917.
Hansen T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.
Harris K., Nielsen R. 2014. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Res. 9:1445–1554.
Holder M.T., Zwickl D.J., Dessimoz C. 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. Philos. Trans. R. Soc. B. 363:4013–4021.
Jones C.T., Youssef N., Susko E., Bielawski J.P. 2017. Shifting balance on a static mutation-selection landscape: a novel scenario of positive selection. Mol. Biol. Evol. 34:391–407.
Jones C.T., Youssef N., Susko E., Bielawski J.P. 2018. Phenomenological load on model parameters can lead to false biological conclusions. Mol. Biol. Evol. 35:1473–1488.
Karin E.L., Wicke S., Pupko T., Mayrose I. 2017. An integrated model of phenotypic trait changes and site-specific sequence evolution. Syst. Biol. 66:917–933.
Keightley P., Trivedi U., Thomson M., Oliver F., Kumar S., Blaxter M. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genet. Res. 19:1195–1201.
Kosakovsky Pond S.L., Frost S.D.W. 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 S.L., Murrell B., Fourment M., Frost S.D.W., Delport W., Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol. Biol. Evol. 28:3033–3043.
Lartillot N., Poujol R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotyic characters. Mol. Biol. Evol. 28:729–744.
Lewis P.O. 2001. A likelihood approach for estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.
Liberles D.A., Teufel A.I., 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., Phillipe H. 2002. Heterotachy, and important process of protein evolution. Mol. Biol. Evol. 19:1–7.
Lu A., Guindon S. 2013. Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. Mol. Biol. Evol. 31:484–495.
Maddison W.P., Midford P.E., Otto S.P. 2007. Estimating a binary character's effect on speciation and extinction. Syst. Biol. 56:701–710.
Mayrose I., Otto S.P. 2011. A likelihood method for detecting trait-dependent shifts in the rate of molecular evolution. Mol. Biol. Evol. 28:759–770.
McCandlish D.M. 2011. Visualizing fitness landscapes. Evolution 65:1544–1558.
Mingrone J., Susko E., Bielwaski J.P. 2018. ModL: exploring and restoring regularity when testing for positive selection. Bioinformatics 35:2545–2554.
Muse S.V., Gaut B.S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. Mol. Biol. Evol. 11:715–724.
Nielsen R., Yang Z.H. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.
O'Connor T.D., Mundy N.I. 2013. Evolutionary modeling of geneotype-phenotype association and application to the primate coding and non-coding mtDNA rate variation. Evol. Bioinformatics 9:301–316.
O'Meara B.C., Ané C., Sanderson M.J., Wainwright P.C. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.
Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. Lond. B 255:37–45.
Parto S., Lartillot N. 2018. Molecular adaptation of Rubisco: discriminating between convergent evolution and positive selection using mechanistic and classical codon models. PLoS One 13:1–16.
Philippe H., Casane D., Gribaldo S., Lopez P., Meunier J. 2003. Heterotachy and functional shift in protein evolution. IUBMB Life 55:257–265.
Phillips P.C. 2008. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat. Rev. Genetics 9:855–867.
Pollock D.D., Thiltgen G., Goldstein R.A. 2012. Amino acid coevolution induces an evolutionary Stokes shift. Proc. Natl. Acad. Sci. USA 109:E1352–E1359.
Pollock D.D., Pollard S.T., Shortt J.A., Goldstein R.A. 2017. Mechanistic models of protein evolution in evolutionary biology: self/nonself

evolution, species and complex traits evolution, methods and concepts. Gewerbestrasse, Switzerland: Springer.
Pupko T., Galtier N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc. R. Soc. Lond. 269:1313–1316.
Romero P.E., Weigand A.M., Pfenninger M. 2016. Positive selection on panpulmonate mitogenomes provide new clues on adaptations to terrestrial life. BMC Evol. Biol. 16:1–13.
Schrider D., Hourmozdi J., Hahn M. 2014. Pervasive multinucleotide mutational events in eukaryotes. Curr. Biol. 21:1051–1054.
Self S.G., Liang K.Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio test under nonstandard conditions. J. Am. Stat. Assoc. 82:605–610.
Sella G., Hirsh A.E. 2005. The application of statistical physics to evolutionary biology. Proc. Natl. Acad. Sci. USA 102:9541–9546.
Spielman S., Wilke C.O. 2015. The relationship between dN/dS and scaled selection coefficients. Mol. Biol. Evol. 34:1097–1108.
Spielman S., Wilke C.O. 2016. Extensively parameterized mutation-selection models reliably capture site-specific selective constraints. Mol. Biol. Evol. 33:2990–3001.
Spielman S., Suyang W., Wilke C.O. 2016. A comparison of one-rate and two-rate inference frameworks for site-specific dn/ds estimation. Genetics 204:499–511.
Starr T.N., Thornton J.W. 2016. Epistasis in protein evolution. Protein Sci. 25:1204–1218.
Tamuri A.U., dos Reis M., Hay A.J., Goldstein R.A. 2009. Identifying changes in selective constraints: host shifts in influenza. PLoS Comput. Biol. 5:1–14.
Tamuri A.U., Goldman N., dos Reis M. 2014. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. Genetics 197:257–271.
Tavaré S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life. Sci. 17:57–86.
Venkat A., Hahn M.W., Thorton J.W. 2018. Multinucleotide mutations cause false inferences of lineage-specific positive selection. Nat. Ecol. Evol. 2:1280–1288.
Wang H., Spencer M., Susko E., Rodger A.J. 2007. Testing for covarion-like evolution in protein sequences. Mol. Biol. Evol. 24: 294–305.
Whelan S., Blackburne B.P., Spencer M. 2011. Phylogenetic substitution models for detecting heterotachy during plastid evolution. Mol. Biol. Evol. 28:449–458.
Wong W.S.W., Yang Z.H., 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 168:1041–1051.
Wright S. 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proceeding of the Sixth International Congress on Genetics, Vol. 1. p. 355–366.
Wright S. 1982. The shifting balance theory and macroevolution. Annu. Rev. Genetics 16:1–19.
Wu J., Susko E. 2009. General heterotachy and distance method adjustments. Mol. Biol. Evol. 26:2689–2697.
Yang Z.H. 2007. PAML4: phylogentic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.
Yang Z.H. 2017. PAML: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.
Yang Z.H., dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. Mol. Biol. Evol. 28:1217–1228.
Yang Z.H., Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46:409–418.
Yang Z.H., Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908–917.
Yang Z.H., Nielsen R., Goldman N., Pedersen A.M.K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.
Yang Z.H., Wong S.W.S., Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107–1118.
Zaheri M., Dib L., Salamin N. 2014. A generalized mechanistic codon model. Mol. Biol. Evol. 31:2528–2541.
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 Z.H. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472–2479.