A Codon Model for Associating Phenotypic Traits with Altered Selective Patterns of Sequence Evolution

Keren Halabi, Eli Levy Karin, Laurent Guéguen, Itay Mayrose

A codon model for associating phenotypic traits with altered selective patterns of sequence evolution

Keren Halabi[1], Eli Levy Karin[2*], Laurent Guéguen[3], Itay Mayrose[1*]

[1] Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

[2] Quantitative and Computational Biology, Max-Planck institute for biophysical Chemistry, Göttingen 37077, Germany.

[3] Laboratory of Biometry and Evolutive Biology, University of Lyon, CNRS, INRIA, Villeurbanne 69100, France.

* To whom correspondence should be addressed:

Itay Mayrose, Tel: 972-3-6407212; Fax: 972-3-6409380

E-mail: itaymay@tauex.tau.ac.il

Eli Levy Kain

E-mail: eli.levy.karin@gmail.com

Running title: Coding-sequence-phenotype integrated model

Keywords: Evolutionary selection; relaxation, genotype-phenotype; γ-proteobacteria;

ABSTRACT

Changes in complex phenotypes, such as pathogenicity levels, trophic lifestyle, and habitat shifts are brought on by multiple genomic changes: sub- and neo-functionalization, loss of function, and levels of gene expression. Thus, detecting the signature of selection in coding sequences and associating it with shifts in phenotypic state can unveil the genes underlying complex traits. Phylogenetic branch-site codon models are routinely applied to detect changes in selective pressures along specific branches of a phylogeny. This a-priori branch partitioning implies that the course of trait evolution is fully known and that transitions in phenotypic states occurred only at speciation events. Here we present TraitRELAX, a new phylogenetic model, that alleviates these strong assumptions by explicitly accounting for the evolution of both trait and coding sequences. This joint statistical framework enables the detection of changes in the selection intensity upon repeated trait transitions, while accounting for uncertainty in the trait transitions pattern by using a stochastic model to describe the trait evolution. We evaluated the performance of TraitRELAX using simulations and then applied it to two case studies. Using TraitRELAX, we found that 36 bacterial genes experienced significant relaxation or intensification of selective pressure upon transitioning from free-living to an endosymbiotic lifestyle, as well as intensification in the Semenogelin 2 gene in polygynandrous species of primates.

INTRODUCTION

The operation of selection on heritable traits leaves distinct signatures in the genes that code for them. These include, for example, depletions in amino-acid changing mutations in genes whose function is crucial. Therefore, analyzing selection fingerprints at the molecular level while considering phenotypic changes can reveal the identity of the genes that are associated with the phenotype, their novel functionalities, or which ones are no longer required. The ongoing advances in high-throughput sequencing and increasing efforts to collect phenotypic trait data (e.g., Parr et al. 2014, Tree Of Sex Consortium et al. 2014, Rice et al. 2015, Kattge et al. 2011) provide the opportunity to detect associations between evolutionary patterns at the genomic level and whole-organism phenotypic traits. Specifically, detecting associations between such traits and selective forces operating at the codon level can provide insight into the locations of functional domains in coding regions that shape the traits of interest and reveal functionalities of unknown genes. With the increased availability of large-scale genome sequence data, the need for comparative methods for detection of such functionalities is increasing. However, such methods are scarce (Nagy et al. 2020).

The nature of selection acting on a protein-coding gene can be revealed by computing the rate ratio between non-synonymous (amino acid altering) and synonymous substitutions, ω. Initial codon models (Goldman and Yang 1994; Muse et al. 1994) incorporated a single ω parameter, thus reflecting the assumption of a single selective pressure that operates across the entire sequence, be it purifying ($\omega < 1$), neutral ($\omega = 1$) or positive ($\omega > 1$). Further developments integrated multiple ω classes into site-models, thereby allowing variation in the selective regime across codon sites (Yang, Nielsen 2000). Moreover, branch-site models (Yang and Nielsen

2002), in which the selective pressure can vary not only across sites, but also among branches of the phylogeny, can be used to detect site-specific changes of selective patterns across the phylogeny based on a prior partitioning of the branches into distinct categories (often termed background, BG, and foreground, FG).

To date, branch-site models are often used to detect selective signatures at the codon level based on phenotypes of study. For example, using the branch-site model of Yang and Nielsen (2002), the color vision in butterflies and primates has been shown to be associated with positive selection in several sites of the opsin gene (Frentiu et al. 2007), and an evidence of connection between rice domestication and elevated ω in several genes has been found (Lu et al. 2006). Furthermore, the mating system in primates has been associated with positive selection in the NYD-SP12 gene that is involved in formation of the acrosome during spermatogenesis (Zhang et al. 2007). Positive selection in chitinase that takes part in construction of the cell wall and is implicated in defense against pathogens has been associated with sexual reproduction in evening primroses (Hersch-Green et al. 2012).

The branch-site model RELAX (Wertheim et al. 2015) is designed to detect shifts in selection intensity in FG branches relative to BG branches. Intensification of selection (either purifying or positive) may be indicative of a transition into more stressful conditions while relaxation of selective pressure may be the result of release of functional constraints upon phenotype transition. The latter may indicate loss of function (Wu et al. 1986; Go et al. 2005) or upcoming lineage extinction (Moran 1996). Analysis using this model allows distinguishing between three scenarios concerning the FG branches: (1) intensification of selection, reflected in ω values moving further away from 1 in the FG branches compared to the BG branches; (2) relaxation of selection, with ω values shifting towards 1; and (3) no change in

selection intensity. To achieve that, RELAX incorporates three ω parameters, $\omega_0 \leq \omega_1 \leq 1 \leq \omega_2$ that represent the site classes of the BG branches. The difference in the magnitude of selective pressure between the BG and FG branches is modeled using a selection intensity parameter k, such that each of the three ω values of the BG branches are raised to the power of k to obtain the selective pressures of the FG branches. Consequently, $k < 1$ implies relaxation of selection and $k > 1$ implies intensification of selection. Using RELAX, several phenotypes have been shown to be associated with a change in selection intensity at the codon level. In orchids, heterotrophic lifestyle has been associated with relaxed purifying selection on the plastid genome (Feng et al. 2016; Roquet et al. 2016; Wicke et al. 2016), and in rodents there is evidence of intensified selection in subterranean species, compared to fossorial species (Tavares and Seuánez 2018).

A notable shortcoming of existing branch-site models is the requirement for a prior specification of branches of the examined phylogeny into branch categories (e.g., BG and FG). Based on the observed phenotypes of the extant species, a partition of the branches is usually determined by reconstructing the ancestral phenotypic states using the maximum likelihood or the maximum parsimony principles. Either way, the obtained partition is assumed to represent the phenotypic history of the trait across the phylogeny. However, considering a single partition disregards any uncertainty in the reconstructed evolutionary history of the trait and tends to underestimate the number of trait transitions. In addition, such an approach unrealistically forces trait transitions to occur simultaneously with speciation events (i.e., at internal nodes of the tree), while in reality they could occur anywhere along a branch. Consequently, any misspecification of the branch assignments could potentially result in failure to detect

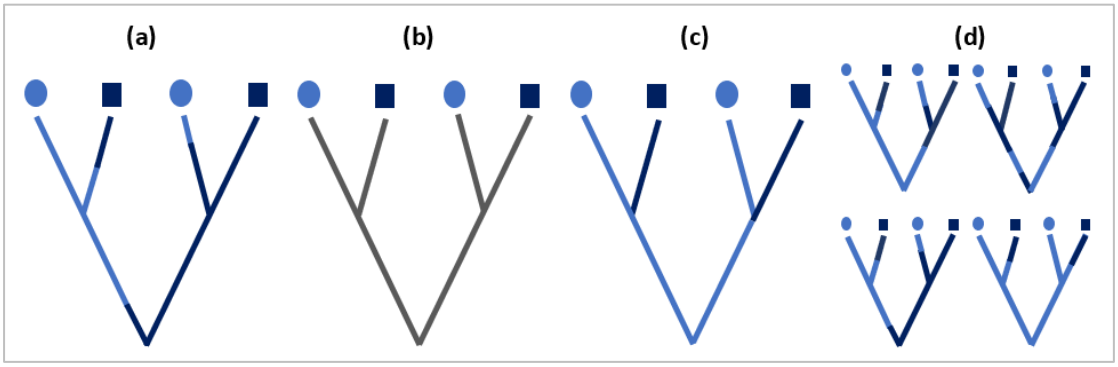
changes in selection patterns, as well as high false positive rate (Lu and Guindon 2014).

To account for uncertainty in trait evolution and for possible associations between the processes of molecular and phenotypic evolution, several methods have been developed, in which phenotypic changes are explicitly modelled and analyzed jointly with sequence data. CoEvol (Lartillot and Poujol 2011) is designed for the analysis of continuous phenotypic traits, while TraitRate (Mayrose and Otto 2011), TraitRateProp (Levy Karin et al. 2017) and the method of O'Connor and Mundy (2009; 2013) are designed for discrete phenotypes. These methods use a rate matrix for nucleotide or amino acid substitutions to either compose a single phenotype-genotype rate matrix (O'Connor and Mundy 2009) or multiply it by a relative rate parameter according to the history of the phenotype (Mayrose and Otto 2011; Levy Karin et al. 2017). Conceptually, using a single phenotype-genotype rate matrix for the analysis of codon data is a straight-forward extension of O'Connor and Mundy's model but this approach can result in inconsistent phenotypic reconstructions (as described by Levy Karin et al, 2017). Both TraitRate and TraitRateProp are inadequate for the analysis of codon data due to their assumption that the entire rate matrix is scaled upon phenotype transition, and thus cannot extract a differential effect on synonymous and nonsynonymous substitutions. The PG-BSM method (Jones et al. 2019) extended this approach to codon sequences, but is used to detect changes in selective regime in association with a phenotype, rather than changes in selection intensity, which is the focus of the current study.

Here, we present TraitRELAX, which enables the detection of changes in the selection intensity upon transitions between phenotypic states. TraitRELAX considers many possible trajectories of trait evolution ("histories"), weighted by their

probabilities (Fig. 1). By doing so, TraitRELAX alleviates the need to rely on pre-specified branch partitions and allows trait transitions to occur anywhere along a branch. Using extensive simulations, we measure the sensitivity and specificity of TraitRELAX and examine the accuracy of the inferred parameter estimates under a range of scenarios. We then demonstrate the utility of the method by detecting relaxation in several house-keeping genes of γ-proteobacteria upon transitioning to endosymbiotic lifestyle, and intensification of selection in the SEMG2 gene, involved in sperm competition, in primate lineages with a polygynandrous mating system.

Figure1



Selection patterns in codon sequence evolution. (a) The true history of the binary phenotypic trait, and thus the true partition of the branches into categories. (b) codon models of the whole genome or site families do not allow selective regimes to vary across the phylogeny, assuming a time homogenous selection pattern. (c) Branch-site models allow the selective regime to vary between the branches of the tree, according to a -priory transitions pattern. (d) The suggested trait-related codon model co-infers the evolution of the phenotypic trait, incorporating uncertainty in its history by integrating over multiple possible histories, weighted by their probability.

MATERIALS AND METHODS

A joint model for character traits and coding sequences

TraitRELAX is a joint probabilistic model of phenotypic (termed throughout 'character trait') and coding-sequence evolution. The input data consist of coding sequence data (D_S) in the form of a multiple sequence alignment (MSA), character data (D_c) describing the trait states of the extant species, and a tree with specified branch lengths (T). The model integrates two continuous time Markov processes: one describing the evolution of the character trait (M_C) and the other is a branch site model that describes the evolution of the coding sequences (M_S). By considering the evolution of both processes in a joint statistical framework, TraitRELAX is able to detect relaxation or intensification of selection at the codon level in association with the character evolution. In our branch-site model, branch partitioning is determined based on the history of the trait of interest, where branch categories are mapped to different character states. Thus, in case of a binary trait, there are two branch categories '0' and '1'.

Character trait evolution

TraitRELAX considers character traits with two possible states (binary) coded as either '0' or '1'. M_C is defined by the rate matrix Q_C:

$$Q_C = \mu \begin{bmatrix} -\pi_1 & \pi_1 \\ \pi_0 & -\pi_0 \end{bmatrix} \quad (1)$$

where $\pi_1 = 1 - \pi_0$ represents the rate of change from state '0' to '1' and μ is a scaling parameter to adapt the branch lengths of T to the expected number of character changes. The model described here is not limited to binary character traits but can be applied generally for larger number of discrete states using the appropriate

rate matrix (Lewis 2001), but in such cases the model would contain additional free parameters.

Coding sequence evolution The TraitRELAX sequence model (M_S) is based on the branch-site model RELAX (Wertheim et al. 2015). This model consists of six rate matrices Q_{bc}, one for each combination of branch category $b \in \{0,1\}$ and site class $m \in \{0,1,2\}$, represented by the parameter ω_m:

$$[Q_{bm}]_{IJ} = r \begin{pmatrix} \theta_{ij}\pi_J & \text{for synonymous substitution between nucleotides } i \text{ and } j \\ \omega_m^{k_b}\theta_{ij}\pi_J & \text{for nonsynonymous substitution between nucleotides } i \text{ and } j \\ 0 & I \text{ and } J \text{ differ by more than one nucleotide} \end{pmatrix} \quad (1)$$

The diagonal elements are determined by the constraint that each row in Q sums to zero. Above, r represents a scaling parameter that is used to convert the branch lengths of the input tree to units of expected number of codon substitutions; π_J denotes the frequency of codon J, and θ_{ij} denotes the rate of substitution from nucleotide i to nucleotide j, which can be parameterized based on any time-reversible nucleotide substitution model (e.g., Tavaré 1986, Tamura 1992). In the analyses presented throughout this study, the K80 nucleotide substitution model (Kimura 1980) was used and thus the nucleotide model incldues a single parameter, which is the transition to transversion rate bias. The k_b parameter represents the selection intensity operating on branch category b relative to branch category 0, and thus $k_0 = 1$.

If we assume that a branch partitioning B is available, the likelihood at site x, $L_{S,B}^x$, can be computed by averaging over the conditional probabilities $P(D_S^x|B, M_S, m)$ of observing the sequence data at site x, D_S^x, in each site class m. Note that for each site class two rate matrices, Q_{0m} and Q_{1m}, are alternately used according to the branch assignments in B. The likelihood thus becomes:

$$L_{S,B}^x = \sum_{m=0}^{2} p_m \times P(D_S^x|B, M_S, m) \quad (3)$$

where p_0 and p_1 are two free parameters, describing the proportion of sites associated with site classes 0 and 1, respectively, and $p_2 = (1 - p_0 - p_1)$ those with site class 2.

Joint likelihood computation

The likelihood of the combined model is the joint probability of D_c and D_S given all model parameters and the phylogeny:

$$L = P(D_C, D_S | T, M_C, M_S) \quad (4)$$

which can be decomposed to:

$$L = P(D_C | T, M_C) \times P(D_S | T, D_C, M_C, M_S) \quad (5)$$

where the first term, denoted here L_C is the likelihood of the character model and is computed using the pruning algorithm (Felsenstein 1981). The second term, denoted $L_{S|C}$, is the likelihood of the sequence model, conditioned on M_C and D_C. Thus, under these settings the computation of $L_{S|C}$ requires knowledge of the character state in each part of T. This history is generally unknown, but the probability of a given history $P(h|T, D_C, M_C)$ can be computed based on D_C and M_C. Thus, the likelihood of the sequence model for site x, $L^x_{s|c}$ is computed by integrating over all possible character histories, weighted by their probabilities:

$$L^x_{S|C} = \int_h P(D^x_S | T, D_C, M_C, M_S, h) P(h | T, D_C, M_C, M_S) dh \quad (6)$$

Omitting parameters that do not affect the probability of D^x_S, we obtain:

$$L^x_{S|C} = \int_h P(D^x_S | h, M_S) P(h | T, D_C, M_C) dh \quad (7)$$

The computation of $L^x_{S,h} = P(D^x_S | h, M_S)$ follows from Equation 3. Note, however, that in the branch partition induced by *h*, character transitions are allowed to occur anywhere along a branch, not necessarily at the internal nodes of the tree. Thus, some branches of the input phylogeny may be divided into several segments, each mapped to a different branch class.

Finally, assuming independence of sequence sites x:

$$L_S = \prod_x L_S^x \quad (8)$$

Approximations

The number of possible character histories is infinite so the full integration in Equation 7 is infeasible. We thus approximate it by importance sampling using stochastic mappings (Nielsen 2002) and replace the integral by summation over N mappings, each with a probability of 1/N:

$$L_S^x \approx \frac{1}{N} \sum_{i=0}^{N} P(D_S^x | h_i, M_S) \quad (9)$$

The likelihood computation detailed above requires N computations of the sequence likelihood, which is costly. Thus, as a second approximation, we replace the costly summation over N mappings with a single likelihood computation given a single history that represents the expected amount of time spent in each character state along each branch, $E(h)$:

$$L_S^x \approx P(D_S^x | E(h), M_S) \quad (10)$$

Obtaining the expected history $E(h)$ can be performed either by averaging N sampled stochastic mappings (Mayrose and Otto 2011, Levy Karin et al 2017) or analytically, following the rewards method of Minin and Suchard (2008). For more details on both approaches, see supplementary materials.

Estimating parameter values

The TraitRELAX model includes a total of 19 parameters, assuming that the K80 nucleotide substitution model is used and that codon frequencies are estimated using the F3x4 model. The nine codon frequencies parameters are estimated from the observed sequence data, while the rest are estimated using a maximum likelihood

search that is divided into two stages: (1) searching for the character model parameters, given a fixed set of sequence model parameters, and (2) searching for the sequence model parameters, given a fixed set of character model parameters. The first stage yields an assignment for the character model parameters, with which the expected character history, $E(h)$, is computed. In this stage, the local search is conducted using either a sequential one-dimensional search (Brent 1974) or a two-dimensional search (Hestenes and Stiefel 1952). In the second stage, branch partitioning that is induced from $E(h)$ is used for likelihood computations. In this stage, the values for the sequence model parameters are being searched simultaneously using the conjugate gradient method (Hestenes and Stiefel 1952).

Testing for a trait-related change in selection intensity

A null model, in which the selection intensity parameter, k, is fixed at 1 (i.e., imposing the same selection intensity throughout the tree, with no effect of the character trait) allows statistical testing of the hypothesis that the evolution of the character trait is not associated with the selection intensity operating on the coding sequences. This is compared to an alternative model, in which k is free to vary. Since the models are nested, the null and alternative models can be compared using a likelihood ratio test (LRT) with one degree of freedom. As an alternative, the distribution of the likelihood ratio was also estimated using parametric bootstrapping (see Results).

Simulating data sets under the TraitRELAX model

Simulations were used to investigate the power and precision of our method and to assess its accuracy in inferring the selection intensity parameter *k*. In general, the simulations were performed by first simulating character evolution along a given

tree. As part of these simulations, the exact locations of character state transitions were recorded, thereby yielding branch partitioning based on the simulated (i.e., true) character histories. This partitioning was used to simulate codon sequences using the RELAX sequence model. Specifically, random trees with different numbers of taxa (16, 32, and 64) were generated according to a birth–death process using INDELible (Fletcher and Yang 2009) with default parameters (speciation rate 0.3 and extinction rate 0.1) and were scaled so that the distance from the root to the tips, defined as the tree height, is 1. Trait histories were simulated along the generated trees and the specified model of character evolution using Bio++ (Guéguen et al. 2013). These histories were then used to generate the codon sequence data using INDELible (Fletcher and Yang 2009) based on the specified sequence model of TraitRELAX. Sequence data were simulated with various number of positions (l = 150, 300, and 600), representing typical range of protein lengths (Anon 2016). Unless otherwise stated, all simulations were conducted with the parameters of the character model set to π_0= 0.5 and μ = 8 and parameters of the sequence model set to $\omega_0 = 0.1$, $\omega_1 = 0.8, \omega_2 = 2, p_0 = 0.5$, $p_1 = 0.4$ and the transition/transverion rate ratio kappa = 0.2. Different values of the selection intensity parameter were simulated: k = 0.2 and 0.8 represent relaxation of selection along lineages with character state 1, k = 1.2, 1.6, and 2 represent intensification of selection, and k = 1 represents null conditions of no trait-related change in the selection intensity. For each simulated scenario, 50 independent runs were conducted (each based on a different tree), with the exception of cases with $k = 1$, for which 200 runs were conducted. The larger number of simulations in the latter case was used to determine the empirical threshold for the LRT using parametric bootstrapping. The different scenarios analyzed in this study are summarized in table 1.

TABLE 1

Values of simulation parameters						
Tree length	**Character model parameters**		**# Taxa**	**Alignment length (codons)**	**Selection intensity**	**Simulation scenario**
	μ	π_0			k	
4	1,4,8,16	0.5	32	300	0.5 1	1
1,4,8,16,32	varies*	0.5	32	300	0.5 1	2
4	8	0.1,0.3,0.5,0.7,0.9	32	300	0.5	3
4	8	0.5	16,32,64	150,300,600	0.2,0.8,1,1.2,1.6,2	4

Simulation parameters used to evaluate TraitRELAX.

*The value of μ was adjusted to the tree length to obtain approximately the same number of character transitions in all the simulations (for example, in simulations with tree length 1, μ was set as 32, and in simulations with tree length 4 it was set as 8).

Empirical Data analyses

Endosymbiont and free living γ-proteobacteria

We first examined the utility of TraitRELAX in detecting changes in the selective pressure upon transitions in the life style of bacteria. To this end, 68 genes from 50 species of γ-proteobacteria, of which 36 are free living and 14 are endosymbionts were examined. The coding sequences, the phylogeny, and character state assignments of extant taxa (free living = 0, endosymbiont = 1) were retrieved from Husník et al. (2011). The sequences were aligned with RevTrans version 2.0b (Wernersson and Pedersen 2003) using default parameters. Positions with more than 90% gaps across sequences were filtered out.

Mating system evolution in primates

As a second empirical example, we examined the evolution of SEGM2, a gene known to be involved in the male reproduction system, in 24 primate species with respect to changes in the mating system. Character state assignments (1 = multimale-multifemale system, 0 = other) were collected from the literature (Dixson 1997; Dixson and Anderson 2002; O'Connor and Mundy 2009). Sequence accessions of the SEMG2 gene were collected from previous papers (Ulvsbäck and Lundwall 1997; Jensen-Seaman and Li 2003; Hurle et al. 2007; Roan et al. 2011; Isshiki and Ishida 2019). The accession of *Otolemur garnettii* was excluded due to exceedingly larger sequence length. In case more than one accession per species was available, a representative accession was selected (the one with maximal pairwise similarity relative to all other accessions). Finally, the assembled sequences were aligned with RevTrans version 2.0b (Wernersson and Pedersen 2003) with default parameters. Positions with more than 90% gaps across sequences were filtered out. The topology of the tree was obtained from the TimeTree knowledge-base (Hedges et al. 2006) and the lengths of the branches were optimized based on the M3 codon model (Yang and Nielsen 2002). The assembled sequence data is provided in the supplementary materials.

Availability

TraitRELAX was implemented as an open-source program in Bio++ (Guéguen et al. 2013) and its code is available at https://github.com/halabikeren/TraitRELAX/. The input to the program is a tree in Newick format, as well as sequence and character data files in Fasta format. The program outputs the maximum-likelihood score for the null and alternative models as well as their inferred model parameters.

RESULTS

Inferring Associations Between Selection Intensity and Phenotypes with the TraitRELAX Model

In this work we developed TraitRELAX to detect associations between the evolution of a binary phenotypic trait and the intensity of selective forces operating on coding sequences. TraitRELAX models their joint evolution by combining two stochastic processes: a two-state Markov process of character changes and a variant of the branch-site model RELAX (Wertheim et al. 2015) of changes in the coding sequences (see "Materials and Methods" for full details). In this joint model, each position in a coding sequence is subject to a selective regime (negative, neutral, or positive) whose intensity could increase (decrease) upon a change in the binary phenotypic trait. Namely, sequence evolution in parts of branches of the phylogenetic tree evolving under state '0' (background, BG) is described by a set selection parameters: ω_0, ω_1 and ω_2, while in parts evolving under state '1' (foreground, FG) an intensification (relaxation) parameter k yields ${\omega_0}^k$, ${\omega_1}^k$ and ${\omega_2}^k$ to describe sequence evolution. TraitRELAX allows testing whether k is significantly different from 1 (indicating intensification/relaxation) and estimating all relevant model parameters (fully detailed in the "Materials and Methods"). We evaluated TraitRELAX on simulated data sets and then applied it to discover intensification in the SEMG2 gene of primates with polygynandrous mating system.

Assessing Performance Using Simulations

Simulations were used to investigate the performance of TraitRELAX with regard to (1) accuracy of phenotypic trait evolution ("character history reconstruction"), (2) false positive rate (FPR): the tendency of the method to detect association between trait evolution and sequence evolution when no such association exists, power: the ability of the method to correctly detect association between trait and sequence evolution when such an association exists, (3) accuracy of parameters estimation (e.g., the selection intensity parameter, k), and (4) analysis of running time of the method. As a comparison to all these analyses, RELAX was executed with partitions that were derived from the true character histories used in the simulation procedure, thus serving as a reference for the optimal performance that could be expected from the method when eliminating any error in character history reconstruction. To compare our method to an existing one, RELAX was also executed with partitions that were induced by applying the maximum parsimony principle on the simulated character data.

Accuracy of phenotypic trait evolution

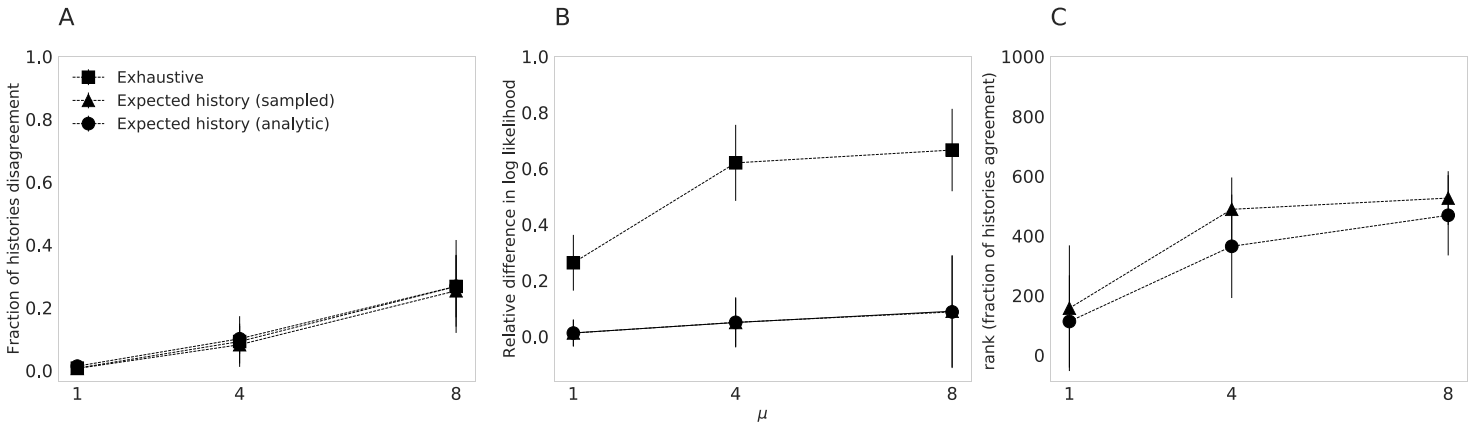
The full likelihood function of TraitRELAX (Equation 6) integrates over all possible character histories. However, since this integration is not feasible, the computation is approximated by a set of N sampled stochastic mappings (Equation 9, denoted here exhaustive approximation). Another approximation uses a single expected history (Equation 10) that is either based on the stochastic mappings or computed analytically. We estimated the accuracy of these approximations by using two measures: (1) the distance between the estimated character histories and the true (i.e., simulated) character history, measured by either the degree of dissimilarity between their induced partitions or by the difference in the computed likelihood values and (2) the rank of similarity between the expected history and the true history, across stochastic mappings. The latter serves as a measure of the effect of the reduction of multiple stochastic mappings to a single expected history on the accuracy of estimated character evolution. More details on these two measures are available in the supplementary materials. The mean values of each of these measures were computed for simulations from scenario 1. For each simulated dataset, three executions of TraitRELAX were carried out; one for each approximation approach. In the first two executions, 1000 stochastic mappings were sampled.

In all approximations, the distance between the estimated character histories and the true history increases with the number of transitions in the true history (i.e., higher values of μ). When comparing the derived partitions (Fig. 2a), all three approximations exhibit similar patterns. However, when comparing these three with regards to the computed likelihood, the approximations based on a single history exhibit lower distance from the true history compared with the exhaustive approximation (

Figure 2b). In addition, the mean rank of the expected history among stochastic mappings increases with the simulated value of μ (

Figure 2c), suggesting that the accuracy of approximations using a single history, compared with the exhaustive approximation, increases with the complexity (i.e., number of character transitions) of the true history. The results suggest that the single-history approximations have similar accuracies and are superior to the exhaustive approximation. However, since the expected history that is reconstructed analytically is robust to stochasticity in sampled mappings by computing the locations of transitions based on the rewards methods (Minin and Suchard 2008), we recommend using it.

Figure 2



Assessment of approximations for TraitRELAX likelihood computation. (a) The mean distance between the derived partitions from the estimated character histories and the true (i.e., simulated) history, measured by the dissimilarities between them along the phylogeny. (b) The mean distance between the likelihood values based on of the estimated character histories and the true (i.e., simulated) history, measured by the relative difference in log likelihood values. (c) The mean rank of the expected histories across stochastic mappings. All measurements are shown for increasing values of μ and for the 3 approximation approaches; Exhaustive (squares), Expected history obtained based on sampled stochastic mappings (rectangles) and analytically estimated expected history (circles).

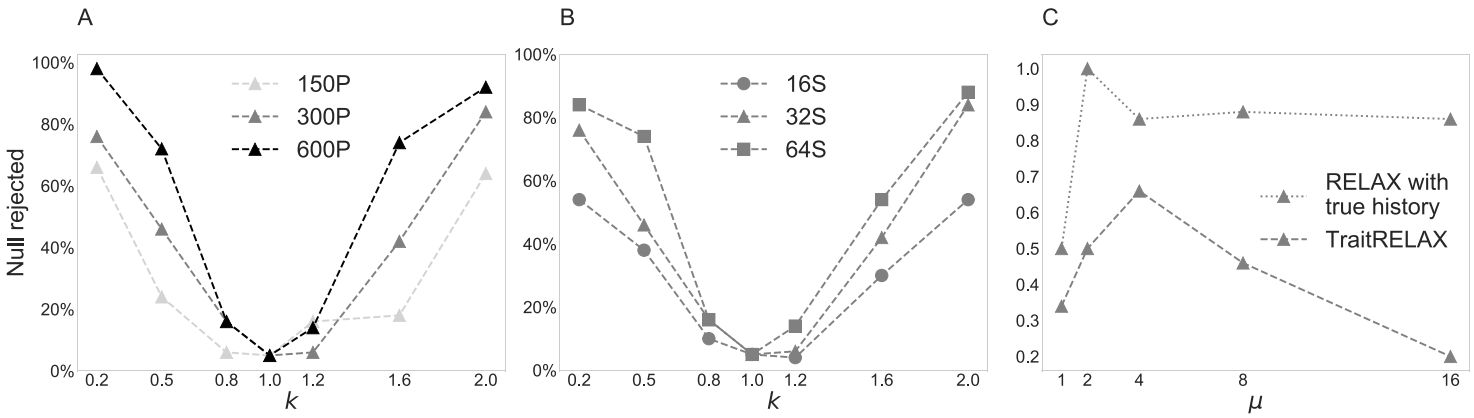
False positive rate and power We analyzed the FPR by setting the selection intensity parameter k to 1, thereby simulating sequence data with no changes in selection intensity along the phylogeny. Similar to Levy Karin et al. (2017), we found deviations of the FPR from the expected 5% when LRT test statistic is compared to a χ_1^2 distribution (FPR of 16.5%, 21% and 15.5%, for simulations with 16, 32, and 64 taxa, respectively, when the number of codon positions was 300). Importantly, when given the true (i.e., simulated) character histories as input, TraitRELAX still exhibited slight derivations from the expected FPR of 5% (FPR of 5%, 5.6% and 6.1%, for simulations with 16, 32, and 64 taxa, respectively, when the number of codon positions was 300), thereby suggesting that the high FPR cannot be solely attributed to error in estimation of character history. We thus developed an alternative parametric bootstrap procedure for corrected p-value computation. To obtain a fixed FPR of 5%, we determined a threshold for the LRT statistic as the 95th percentile of the LR values that TraitRELAX yielded for the simulated null data. The thresholds determined this way for simulations with 300 codon positions were 6.57, 7.62, and 6.79, for 16, 32, and 64, respectively (compared to 3.84 using the χ_1^2 approximation). These thresholds were then used for power analysis.

We analyzed the power of TraitRELAX based on simulations from scenario 4. In these simulations, we varied the number of species, the lengths of the simulated codon alignments and the value of selection intensity parameter k.

The power of the method increases with the number of sequence positions (Fig 3a); the power for simulations with 32 taxa and k=0.2 is 0.76 and 0.98 for 300 positions and 600 positions in the sequence alignment) and the number of taxa in the tree (Fig 3b; power for simulations with 300 positions and k=0.2 is 0.76 and 0.84 for

32 taxa and 64 taxa). Furthermore, increasing the magnitude of the selection intensity parameter also resulted in increased power (The power was 0.74 for simulations under k=0.5, and 0.84 for simulations under k=0.2 in simulations of 64 taxa and 300 positions). To examine the effect of the number of transitions in character state of the trait on the power of the method, the power was computed for simulations with increasing values of μ (i.e., the character transition rate). A maximum power value of 0.66 was observed for scenario of a mild μ value of 4 and decreased (as low as 0.2) for extreme values (Fig 3c). As expected, the reference (i.e., TraitRELAX when given the true history) consistently exhibits higher power than standard execution of TraitRELAX, and the difference between the two approaches increases with the value of μ, that is, with increasingly large number of transitions (The power was 0.2 and 0.86 in standard executions and when given the true history in simulations with μ=16).

Figure 3

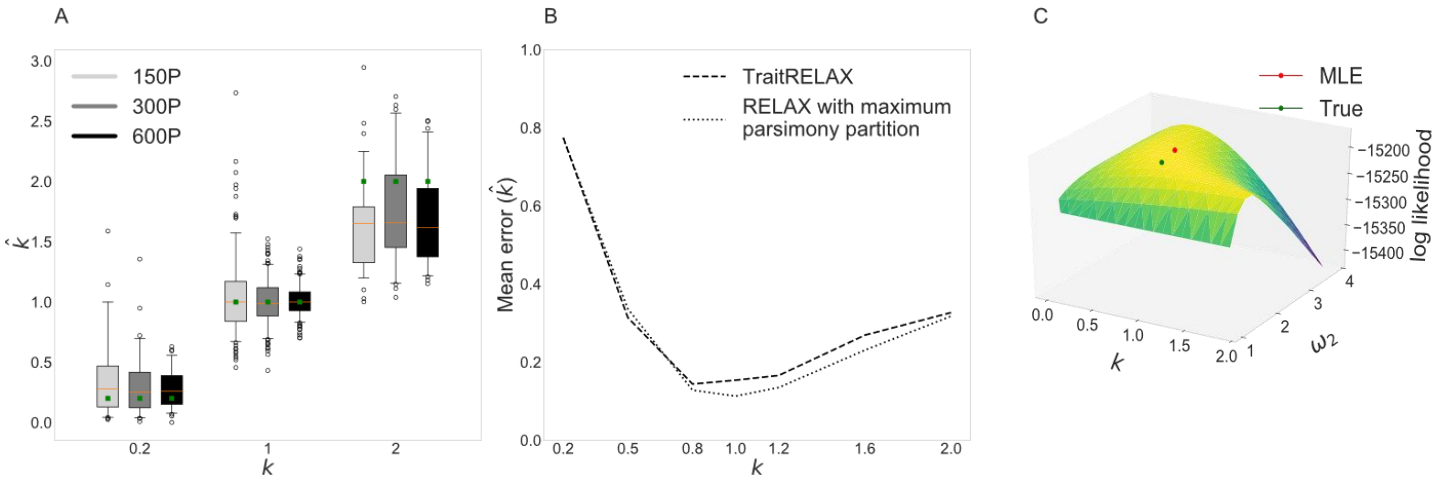


Assessment of the power and false positive rate (FPR) of TraitRELAX. The percent of replicates from simulation scenario 4 for which TraitRELAX rejected the null model based on parametric bootstrapping is shown is shown for simulations of 32 taxa with increasing number of codon positions (d) and for simulations of 300 codon positions with increasing number of taxa; 150, 300 and 600 codon positions in blue, orange and green, and similarly, 16, 32 and 64 taxa are in blue, orange and green. (c) The power of the method is shown for simulations with increasing simulated values μ from scenario 1. Results are shown for standard executions of TraitRELAX and executions of RELAX given the simulated character histories in full and dashed lines.

Accuracy of parameters estimation The accuracy of the method was evaluated by the error in the inferred value of selection intensity parameter k, which is of major importance since it is reflective of the magnitude of trait-related change in selection intensity. The results exhibit an increase in the error for extreme values of k, and higher error for relaxation values than for intensification values (Figure a; The mean error is 0.77 and 0.17 for simulations of 32 taxa and 300 codon positions when k is 0.2 and 1.2). The influence of the trait evolution on the accuracy in inference of k was also examined on simulations with increasing values of μ (scenario 1) and π_0 (scenario 3). The lowest error is obtained for mild values of μ=4 but the accuracy is robust to the simulated value π_0 (Figure 3Sa-b). In addition, TraitRELAX exhibits lower error than RELAX when given a partition inspired by the maximum parsimony principle, for simulation with μ higher than 4 and tree length larger than 4. This suggests that a more refined modeling of the character evolution of the trait leads to higher accuracy in the inference of k, when the true history of the trait is reach with transitions. Lastly, similar to the results of the power analysis, we found that the accuracy of inferring the model parameters increases with the number of species in the data set and with the number of positions in the sequence data (Fig. 4S).

Due to the nature of expression of selection intensity parameter k via the ω values of branch category 1, we studied the interplay between the inference of k and that of the ω parameters by examining the likelihood surface as a function of k and ω_2 (Figure c). Both the true (i.e., simulated) values and maximum likelihood estimators are located along a ridge of high likelihood value, such that a large number of combinations of $\{k, \omega 2\}$ pairs receive similar log-likelihood scores. The results exhibit apparent deterioration in likelihood as the parameter values grow farther than the simulated ones, with a more prominent effect upon changes in the value of k.

Figure 4



Assessment of the accuracy of TraitRELAX. (a)The distribution of inferred values of k across simulations from scenario 4 with 32 taxa and $k \in \{0.2,1,2\}$, (b) the mean error in the inferred values of k are shown for standard executions of TraitRELAX and executions of RELAX given a partition inspired by a maximum parsimony solution in full and dotted lines, for simulations from scenario 4 with 32 taxa and 300 codon positions with increasing simulated values of k. The error is measured as $\left|log\left(\hat{k}\right) - log(k)\right|$ to account for the exponential effect of k on the ω values of branch category 1. (c) Log likelihood surface over a grid of 300 combinations of ω_2 and k. The figure depicts a single simulation instance in which k=0.8, ω_2=2 (with p_2=0.99) with 32 species and 600 codon positions. The simulated values and maximum likelihood estimators are shown in green and red.

Duration

The mean run time of the method was computed for while allocating each run a single core in a Linux server. The run time of the method varies both with the number of taxa (mean values of 10.5 and 14.77 hours for simulations of 32 and 64 taxa with 300 positions) and with the number of positions in the coding sequence alignment (mean values of 9.15 and 10.5 hours for simulations of 150 and 300 positions with 32 taxa). Importantly, the average number of cycles required to converge to optimal assignment of the alternative model parameters remains approximately 2 consistently throughout the simulation study.

Association Between the intensification in SEMG2 Gene and Polygynandry in Primates

We used TraitRELAX to investigate associations between a polygynandrous mating in primates the intensity of selection in the Semenogelin II (SEMG2) gene. We fitted two models to the collected data; in the first model, we fixed the value of selection intensity parameter k to 1, thus imposing that the sequence evolution is in no association to the character evolution. In the second model, we set the selection intensity parameter k as free, thus allowing the sequence evolution to be in association with character changes. To compare our method to existing ones, we also fitted the RELAX model described earlier based on two alternative partition that were inspired by the maximum parsimony and maximum likelihood ancestral assignments of character states of the trait along the phylogeny. Here, we also fitted two models accordingly; one in which k is fixed to 1 and another where it is free to vary. Following the FPR control approach used in simulations, we tested the hypothesis of polygynandry-related change in selection intensity using parametric bootstrapping, both in RELAX and in TraitRELAX. The result in RELAX varied based on the given partition; while the results given the partition inspired by the maximum likelihood

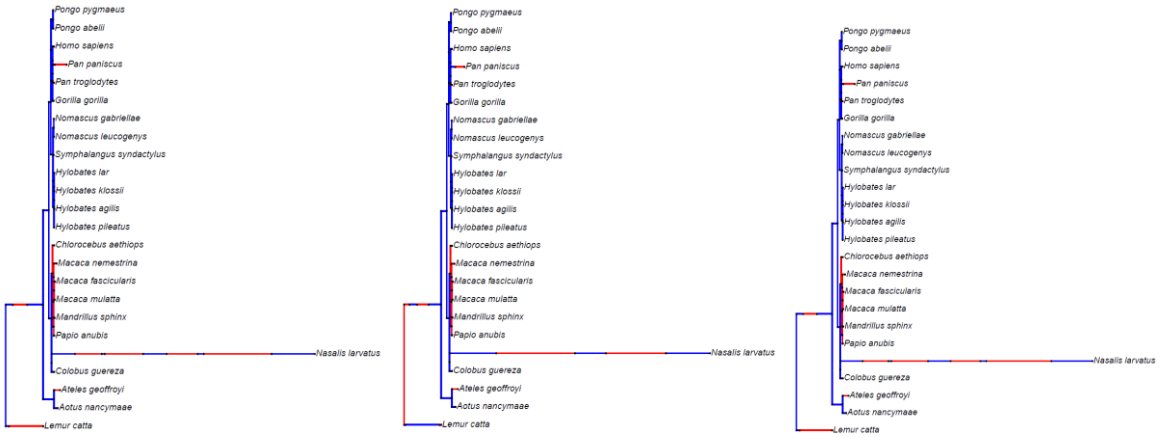
ancestral assignments gave no significant evidence of trait-related change in selection intensity (k=1.44, $p-value$=0.066), the one when given the partition based on maximum parsimony principle showed significant evidence of intensification of selection under the polygynandrous state (k= 2.00, $p-value$=0.002). TraitRELAX also yielded significant result (k=2.55, $p-value$=0.001). This suggests that by accounting for uncertainty in trait evolution, TraitRELAX is robust to misleading starting point. This is in opposite to RELAX which is susceptible to misspecifications in the input partition. Importantly, the results of TraitRELAX obtained lower p-value than the ones in RELAX, and obtained better likelihood of the sequence model, which corresponds to RELAX (by a difference of 0.63 log likelihood units). Examination of the partitions used by RELAX and the estimated character histories by TraitRELAX suggests that the more refined pattern of transitions in the latter plays a role in yielding different inference in the two approaches (Figure 5).

Figure 5

Maximum parsimony

Maximum likelihood

Stochastic mappings



Partitions used in RELAX and TraitRELAX on a species tree of primates.

DISCUSSION

We developed TraitRELAX, a new method for detecting associations between organismal phenotypes and selection intensity at the codon level. Within the framework of the method, an association is expressed as a dependency between the evolutionary processes of phenotypic changes and patterns of sequence change along branches of a phylogeny. An important consideration of the method is that transitions in the phenotypic states should affect all sequence positions simultaneously, and thus the evolution of each site cannot be treated independently. Since the exact timings of phenotypic transitions are unknown, we use the stochastic mapping approach to sample possible scenarios of change that led to the observed phenotypic states at the extant taxa, which in turn induce multiple partitions that are provided to the RELAX model simultaneously. TraitRELAX can receive as input an ultrametric tree, which reflects an assumption in which branch lengths are proportional to time (e.g., millions of years) and better fits traits whose rate of change is time-dependent, such as organismal habitat or lifestyle. Alternatively, TraitRELAX can also consider non-ultrametric trees, in which the branch lengths may correspond to the expected number of sequence substitutions per site, and better fit traits whose rate of change is proportional to the amount of genetic change, like pathogenicity or generation time. Because the inference of time calibrated (or ultramertric) trees is known to be error-prone (Graur and Martin 2004; Hedges and Kumar 2004), one may prefer as a general rule of thumb the use of uncalibrated trees unless there is strong evidence that the evolution of trait in question is time dependent.

The performance of TraitRELAX was evaluated with simulations which included various scenarios of dependency between phenotypic and sequence changes as well as scenarios in which these were independent of each other. We found that

larger number of species in the input tree contributes to the increase in power and accuracy of TraitRELAX, as well as large number of sequence positions. This result well reflects an expected pattern in which the increase in the size of the input data increases the statistical power of a method. Since the number of sequence positions is fixed and is determined by the size of the analyzed protein coding gene, it will beneficial to collect sequences from as many species as possible to assure adequate performance. The results demonstrated that TraitRELAX performs well when provided data with 300 sequence positions and 64 species, although shorter sequences may suffice when larger trees are available. We also expected optimal results when large proportion of the sequence data is associated with the trait of interest, to avoid the potential artifact of non-associated positions affecting the overall fit of the model to the data. The results further showed that for the TraitRELAX model an empirical likelihood ratio (LR) test is better suited to test trait-related change in selection intensity. An alternative that is more compatible with the case of using real data, yet time consuming, approach is to use a full parametric bootstrapping procedure (Whelan and Goldman 1999), which can obtain for each examined dataset a more compatible distribution of LR values and consequently result in reduction of the FPR. The results also indicated that the method is sensitive to the confounding effect between k and the ω parameters (underestimation of the inferred values of the ω_0 parameter may be compensated by overestimation in the inference of the k parameter, such that the induced value of ${\omega_0}^k$ is similar and thus the obtained likelihood score is nearly unchanged). TraitRELAX was able to detect trait-related relaxation of selective pressures in a biological example that have been hypothesized to exhibit such pattern in the past. In the reproductive gene SEMG2, TraitRELAX detected polygynandry-related intensification.

To speed up the computation of the likelihood of the TraitRELAX model, which is meant to integrate over multiple possible character histories of the trait of study, we adopted the approach presented in Levy Karin et al. (2017) that relies on a single character history which represents the sampled character histories, denoted as the "expected history". A possible shortcoming of this approximation concerns with the reduction of the number of transitions of character state in each branch of the phylogeny to two transitions at most. When many transitions have occurred along a branch, this approach is expected to yield inferior approximation of the likelihood function than the one based on multiple histories. However, most organismal traits of interests are key features of a lineage and are expected to evolve rather slowly, such that multiple or back transitions on the same branch should be rare. In such cases, we expect the single history approximation to perform well. While the nature of transitions in the expected history may pose an issue with regard to the accuracy of the character history, this approximation can be beneficial both in duration (since only once history is considered in the computation) and in stability of the likelihood computation (because the analytic expected history is deterministic and the sampling-based expected history is more robust to stochasticity than the stochastic mappings used in the exhaustive approximation). Moreover, the heavy computations of TraitRELAX mainly stem from likelihood computation of the sequence data, while the time needed to create the large sample of histories, as well as computing the expected history based on this sample, is substantially lower. This allows applying the sampling-based expected history approach using much more histories (i.e., larger value of N) compared to the more exhaustive computation, which further improves the stability of the joint likelihood computation.

An important, yet overlooked, consideration when computing the likelihood of a branch-site codon model, such as the sequence model in TraitRELAX, concerns with the nature of transitions between selective regimes along the phylogeny (i.e., transition in ω categories across branches). Consider a branch-site model with two branch categories (e.g., BG and FG), such that each is associated with three ω parameters ($\omega_0, \omega_1, \omega_2$ for purifying, neutral and positive selection). There are three ways to compute the likelihood for two consecutive branches b_1 and b_2. The first option restricts the ω category in both branches to be the same, whether they belong to the same branch category or not. Thus, if a site evolved under ω_0 in b_1, it will evolve under ω_0 also in b_2. Because each branch category has its own value of ω for each category (e.g., ${\omega_0}^{BG} \neq {\omega_0}^{FG}$), the site can still experience a change in the selective pressure when transitioning between BG and FG branches. Alternatively, we can assume that the ω category assigned to each site may vary between branches only upon transition from FG to BG along the phylogeny. Thus, despite having a site evolving under ω_0 in b_1, it may evolve under any ω category in b_2. In a third and most permissive approach, also known as the random effects approach (Kosakovsky Pond et al. 2011), one can assume that the ω category assigned to each site may vary between branches, regardless of their branch assignment. While this is the most general approach, it allows rapid shifts in ω categories that are not necessarily realistic, such as transitioning from purifying to positive selection along very short branches, and can lead to high rate of false positives (Kosakovsky Pond and Frost 2005). The second approach also poses a theoretical shortcoming of favoring certain partitioning of branches that consider more combinations of ω assignments. Specifically, when a transition from a FG branch to BG branches occurs in an internal branch, the restriction of the ω assignment for all the direct BG branches beneath it is

relaxed, while transitions that occur in external branches relax the ω assignment restriction only for the respective external branch. This issue is prominent in the common multiple testing procedure offered by Anisimova and Yang (2007), in which each test corresponds to a different partitioning that classifies a single branch to the FG category. Because this approach is the one most commonly used by the community (for example, it is the approach used in the popular software package PAML; (Yang 1997)), further research is needed to explore its behavior. Due to the consideration of more realistic selective patterns and the added benefit of computation simplicity, in the implementation of TraitRELAX I decided to follow the first and most restrictive computation approach.

Here we introduced a combined model of phenotypic changes and codon changes. Specifically, the model allows detection of associations between phenotypes and changes in the selection intensity at the codon level. However, the same procedure can be applied for a wide range of analyses, by selecting a different branch-site codon model as the underlying sequence model. For instance, setting the standard branch-site model (Zhang et al. 2005) as the sequence model will allow the detection of positive selection at the codon level upon repeated transitions in a specific character state. In such case, segments in the phylogeny corresponding to character state '0' could be develop under purifying or neutral selection, while segments corresponding to state '1' could also develop under positive selection. Previous work has attempted to integrated phenotypic evolution into a covarion-like codon model to detect adaptation (Jones et al. 2019). The method was compared to the standard branch-site model. However, it may be interesting to compare it to a model which interrogated phenotypic evolution into the same statistical framework of the branch-site model. A generalization of the same approach can be applied by mapping

independent codon models (Ziheng Yang,* Rasmus Nielsen 2000) to each phenotypic character state, while having some of the parameters shared between the models to reduce the overall number of parameters and, consequently, the running time of the optimization.

Lastly, despite having TraitRELAX consider only binary phenotypic traits, many categorical traits are not binary, such that the probability of transitions between states will account of graduality in the discretized phenotypes. Additionally, the examination of multiple, perhaps correlated, traits, can provide new insights about the characteristics of selective patterns exhibited at the codon level. The combination of multiple traits can be perceived as a single, categorical, trait, whose states correspond to the various combinations of character states of the traits of interest. Another possible extension would be to generalize the approach used in TraitRELAX to consider categorical traits with more than two states, by adjusting the number of states in the character model and the number of branches categories in the codon sequence model.

Despite the potential of such extensions to improve the performance of the method, we have shown that the current implementation of TraitRELAX is able to detect trait-related changes in selection-intensity both in simulations in the real biological examples. TraitRELAX offers the added value of a probably-based approximation of the history of the trait that is used to study the pattern of change in selection intensity, that has been shown to be more refined and accurate in some cases compared to methods like maximum parsimony, thereby reducing the rate of false positive and increasing the power to detect such patterns, compared to existing approaches that are mostly based on prior information on the history of the trait.

ACKNOWLEDGEMENTS

This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University

REFERENCES

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol. Biol. Evol.

Anon. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol.

Anon. 2016. Cell biology by the numbers. Choice Rev. Online.

Brent RP. 1974. Algorithms for Minimization Without Derivatives. IEEE Trans. Automat. Contr.

Dixson AF. 1997. Evolutionary perspectives on primate mating systems and behavior. In: Annals of the New York Academy of Sciences.

Dixson AF, Anderson MJ. 2002. Sexual selection, seminal coagulation and copulatory plug formation in primates. Folia Primatol.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol.

Feng YL, Wicke S, Li JW, Han Y, Lin CS, Li DZ, Zhou TT, Huang WC, Huang LQ, Jin XH. 2016. Lineage-specific reductions of plastid genomes in an orchid tribe with partially and fully mycoheterotrophic species. Genome Biol. Evol. 8:2164–2175.

Fletcher W, Yang Z. 2009. INDELible: A flexible simulator of biological sequence evolution. Mol. Biol. Evol. 26:1879–1888.

Frentiu FD, Bernard GD, Cuevas CI, Sison-Mangus MP, Prudic KL, Briscoe AD. 2007. Adaptive evolution of color vision as seen through the eyes of butterflies. Proc. Natl. Acad. Sci. U. S. A.

Go Y, Satta Y, Takenaka O, Takahata N. 2005. Lineage-specific loss of function of bitter taste receptor genes in humans and nonhuman primates. Genetics.

Graur D, Martin W. 2004. Reading the entrails of chickens: Molecular timescales of evolution and the illusion of precision. Trends Genet.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V, et al. 2013. Bio++: Efficient extensible libraries and tools for computational molecular evolution. Mol. Biol. Evol. 30:1745–1750.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: A public knowledge-base of divergence times among organisms. Bioinformatics.

Hedges SB, Kumar S. 2004. Precision of molecular time estimates. Trends Genet.

Hersch-Green EI, Myburg H, Johnson MTJ. 2012. Adaptive molecular evolution of a defence gene in sexual but not functionally asexual evening primroses. J. Evol. Biol.

Hestenes MR, Stiefel E. 1952. Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur. Stand. (1934).

Hurle B, Swanson W, Green ED. 2007. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. Genome Res.

Husník F, Chrudimský T, Hypša V. 2011. Multiple origins of endosymbiosis within the convergence of complex phylogenetic approaches. BMC Biol. 9:87.

Isshiki M, Ishida T. 2019. Molecular evolution of the semenogelin 1 and 2 and mating system in gibbons. Am. J. Phys. Anthropol.

Jensen-Seaman MI, Li WH. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. J. Mol. Evol.

Jones CT, Youssef N, Susko E, Bielawski JP. 2019. A phenotype-genotype codon model for detecting adaptive evolution. Syst. Biol.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol.

Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol. Biol. Evol. 28:3033–3043.

Lartillot N, Poujol R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Mol. Biol. Evol. 28:729–744.

Levy Karin E, Wicke S, Pupko T, Mayrose I. 2017. An Integrated Model of Phenotypic Trait Changes and Site-Specific Sequence Evolution. Syst. Biol. 66:917–933.

Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol.

Lu A, Guindon S. 2014. Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. Mol. Biol. Evol. 31:484–495.

Lu J, Tang T, Tang H, Huang J, Shi S, Wu CI. 2006. The accumulation of deleterious mutations in rice genomes: A hypothesis on the cost of domestication. Trends

Genet.

Mayrose I, Otto SP. 2011. A likelihood method for detecting trait-dependent shifts in the rate of molecular evolution. Mol. Biol. Evol. 28:759–770.

Minin VN, Suchard MA. 2008. Fast, accurate and simulation-free stochastic mapping. Philos. Trans. R. Soc. B Biol. Sci. 363:3985–3995.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. 93:2873–2878.

Muse S V, Gaut BS, Carolina N. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol 1.

Nagy LG, Merényi Z, Hegedüs B, Bálint B. 2020. Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. Nucleic Acids Res.

Nielsen R. 2002. Mapping mutations on phylogenies. Syst. Biol. 51:729–739.

O'Connor TD, Mundy NI. 2009. Genotype-phenotype associations: Substitution models to detect evolutionary associations between phenotypic variables and genotypic evolutionary rate. Bioinformatics 25:94–100.

O'Connor TD, Mundy NI. 2013. Evolutionary modeling of genotype-phenotype associations, and application to primate coding and non-coding mtDNA rate variation. Evol. Bioinforma. 2013:301–316.

Roan NR, Müller JA, Liu H, Chu S, Arnold F, Stürzel CM, Walther P, Dong M, Witkowska HE, Kirchhoff F, et al. 2011. Peptides released by physiological cleavage of semen coagulum proteins form amyloids that enhance HIV infection. Cell Host Microbe.

Roquet C, Coissac É, Cruaud C, Boleda M, Boyer F, Alberti A, Gielly L, Taberlet P,

Thuiller W, Van Es J, et al. 2016. Understanding the evolution of holoparasitic plants: The complete plastid genome of the holoparasite Cytinus hypocistis (Cytinaceae). Ann. Bot. 118:885–896.

Tamura. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Am. Math. Soc. Lect. Math. Life Sci.

Tavares WC, Seuánez HN. 2018. Changes in selection intensity on the mitogenome of subterranean and fossorial rodents respective to aboveground species. Mamm. Genome 29:353–363.

Ulvsbäck M, Lundwall Å. 1997. Cloning of the semenogelin II gene of the rhesus monkey. Duplications of 360 bp extend the coding region in man, rhesus monkey and baboon. Eur. J. Biochem.

Wernersson R, Pedersen AG. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res.

Wertheim JO, Murrell B, Smith MD, Pond SLK, Scheffler K. 2015. RELAX: Detecting relaxed selection in a phylogenetic framework. Mol. Biol. Evol. 32:820–832.

Whelan S, Goldman N. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol. Biol. Evol.

Wicke S, Müller KF, dePamphilis CW, Quandt D, Bellot S, Schneeweiss GM. 2016. Mechanistic model of evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants. Proc. Natl. Acad. Sci. [Internet] 113:9045–9050. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1607576113

Wu CI, Li WH, Shen JJ, Scarpulla RC, Limbach KJ, Wu R. 1986. Evolution of

cytochrome c genes and pseudogenes. J. Mol. Evol.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics 13:555–556.

Yang Z, Nielsent R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908–917.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472–2479.

Zhang Q, Zhang F, Chen XH, Wang YQ, Wang WQ, Lin AA, Cavalli-Sforza LL, Jin L, Huo R, Sha JH, et al. 2007. Rapid evolution, genetic variations, and functional association of the human spermatogenesis-related gene NYD-SP12. J. Mol. Evol.

Ziheng Yang,* Rasmus Nielsen NG and A-MKP. 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Mol. Biol. Evol. 19:49–57.