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

Keren Halabi[1], Eli Levy Karin[2*], Laurent Guéguen[3,4], Itay Mayrose[1*]

[1] School of Plant Sciences and Food Security, 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.

[4] Swedish Collegium for Advanced Study. Thunbergsvägen 2
752 38 Uppsala, Sweden.

* To whom correspondence should be addressed:

Itay Mayrose, Tel: +972-3-6407212

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

Eli Levy Karin, Tel: +49 551 201-2881

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

Running title: Coding sequence - phenotype integrated model

Keywords: Evolutionary selection; intensification, relaxation, genotype-phenotype; γ-proteobacteria; SEMG2;

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 patterns along specific branches of the phylogeny. These methods rely on a pre-specified partition of the phylogeny to branch categories, thus treating the course of trait evolution as fully resolved and assuming that transitions in phenotypic states have occurred only at speciation events. Here we present TraitRELAX, a new phylogenetic model that alleviates these strong assumptions by explicitly accounting for the uncertainty in the evolution of both trait and coding sequences. This joint statistical framework enables the detection of changes in selection intensity upon repeated trait transitions. We evaluated the performance of TraitRELAX using simulations and then applied it to two case studies. Using TraitRELAX, we found an intensification of selection in the SEMG2 gene in polygynandrous species of primates compared to species of other mating forms, as well as changes in the intensity of purifying selection operating on sixteen bacterial genes upon transitioning from free-living to an endosymbiotic lifestyle.

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 (Drory et al. 2004; Gertow et al. 2004; Melamed et al. 2004) or rapid accumulation of mutations in pathogenic genes evading the host immune system of drugs (Endo et al. 1996; Burch and Chao 1999). Therefore, analyzing selection fingerprints at the molecular level while considering phenotypic changes can reveal changes in the selective patterns upon repeated phenotype transitions and the identity of the genes that are associated with the phenotype. The ongoing advances in high-throughput sequencing and increasing efforts to collect phenotypic trait data (e.g., Kattge et al. 2011; Ashman et al. 2014; Parr et al. 2014; Rice et al. 2015), 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 on 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 demand for comparative methods for detection of such functionalities is growing. 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 positions, 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 partition of the branches into distinct categories (often termed background, BG, and foreground, FG, corresponding to the ancestral and derived states, respectively).

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 have 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). In evening primroses, positive selection has been associated with sexual reproduction in the chitinase gene, which takes part in construction of the cell wall and defense against pathogens (Hersch-Green et al. 2012).

The branch-site model RELAX (Wertheim et al. 2015) is designed to detect shifts in selection intensity in a predefined set of FG branches relative to BG branches. Intensification of selection (either purifying or positive) may be indicative of a shift 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 by 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 with fossorial species (Tavares and Seuánez 2018).

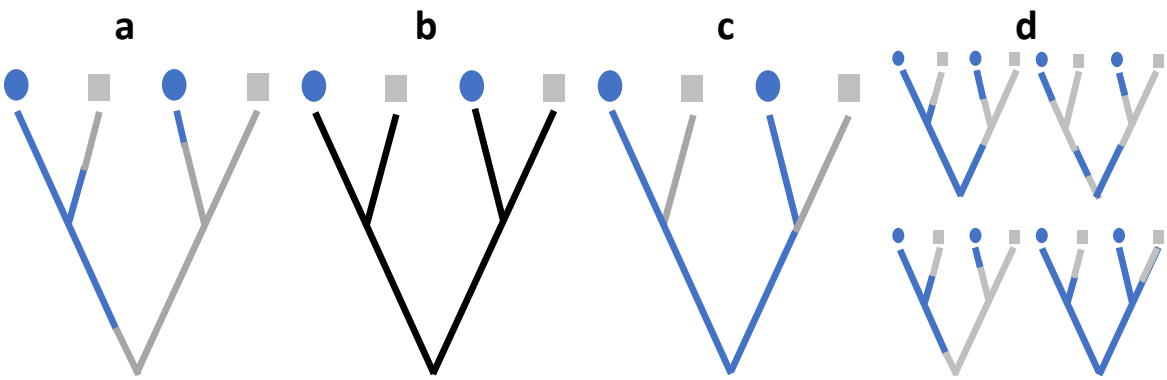
A notable shortcoming of existing branch-site models is the requirement for a prior specification of branches of the examined phylogeny into branch categories (e.g., BG and FG). Based on the observed phenotypes of the extant species, a partition of the branches is usually determined by reconstructing the ancestral phenotypic states using the maximum parsimony or the maximum likelihood 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. 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, and could lead to underestimation of the number of trait transitions. 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 phenotype 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) 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 trait transition, and thus cannot extract the differential effect on synonymous and nonsynonymous substitutions. Recently, Jones et al. (in press) extended this approach to codon sequences for the detection of adaptive changes in selective regime in association with a phenotype, rather than changes in the 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 data scenarios. We then demonstrate the use of the method by detecting evidence of relaxation in several house-keeping genes of γ-proteobacteria upon transitioning to endosymbiotic lifestyle, and an intensification of selection in the SEMG2 gene, involved in sperm competition in polygynandrous (i.e., multi-male multi-female mating system) primate lineages.

Figure 1 - Selection patterns in codon sequence evolution.



(a) The true history of a binary trait (square / circle) and 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 trait, incorporating uncertainty in its history by integrating over multiple possible histories, weighted by their probability.

Materials & Methods

TraitRELAX is a joint probabilistic model of phenotypic (termed throughout 'character trait') and coding-sequence evolution. The input consists 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 both processes in a joint statistical framework, TraitRELAX can detect relaxation or intensification of selection at the codon level in association with the character evolution. In our branch-site model, branch partition 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 coded as either '0' or '1'. M_C is defined by the rate matrix Q_C:

$$Q_C = \mu \begin{bmatrix} -\pi_1 & \pi_1 \\ \pi_0 & -\pi_0 \end{bmatrix} \tag{1}$$

where $\pi_1 = 1 - \pi_0$ represents the rate of change from state '0' to '1' and μ is a scaling parameter designated 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 any 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_{bm}, 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:

$$[Q_{bm}]_{IJ} = r \begin{pmatrix} \theta_{ij}\pi_J & \text{synonymous substitution} \\ \omega_m^{k_b}\theta_{ij}\pi_J & \text{nonsynonymous substitution} \\ 0 & I \text{ and } J \text{ differ by more than one nucleotide} \end{pmatrix} \tag{2}$$

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 equilibrium 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).

Throughout this study we used the K80 nucleotide substitution model (Kimura 1980), which incldues a single parameter describing 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 $k_0 = 1$.

Given a branch partition B, 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. For each site class, two rate matrices, Q_{0m} and Q_{1m}, are alternately used according to the branch assignments in B. Considering all site classes, 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 free parameters, describing the fractions 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 as L_C, is the likelihood of M_C and is computed using the pruning algorithm (Felsenstein 1981). The second term, denoted $L_{S|C}$, is the likelihood of M_S, conditioned on M_C and D_C. 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 M_S for site x, L_S^x is computed by integrating over all possible character histories, weighted by their probabilities:

$$L_S^x = \int_h P(D_S^x \mid T, D_C, M_C, M_S, h) P(h \mid T, D_C, M_C, M_S) dh \quad (6)$$

Omitting parameters that do not affect the probability of D_S^x and h, we obtain:

$$L_S^x = \int_h P(D_S^x \mid h, M_S) P(h \mid T, D_C, M_C) dh \quad (7)$$

The computation of $L_{S,h}^x = P(D_S^x \mid h, M_S)$ follows from Equation 3. Of note, 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 (in such cases, each point of a trait transition along a branch defines an internal node with a single descendent lineage). Finally, assuming independence of sequence sites, the likelihood of the sequence model becomes:

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

Approximations

The number of possible character histories is infinite so full integration of Equation 7 is infeasible. Two alternatives were thus used to approximate the full integration in Equation 7: an exhaustive approach and an expected history based approach. In the exhaustive approach, we perform importance sampling using stochastic mappings (Nielsen 2002) and replace the integral by a summation over N mappings, each with a probability of 1/*N*:

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

The likelihood computation detailed above requires N computations of the sequence likelihood, which is computationally exhaustive. Alternatively, in the expected history

based approach, we use a single history $E(h)$ that represents the expected amount of time spent in each character state along each branch over all possible histories:

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

We compute the expected history $E(h)$ following the rewards method of Minin and Suchard (2008) (see Supp. Text 1). We used the expected history based approach throughout as it proved superior to the exhaustive approach both in terms of runtime as well as robustness (Supp. Text 2).

Estimating parameters

The TraitRELAX model includes a total of 19 parameters, assuming the K80 nucleotide substitution model is used and that codon frequencies are estimated using the F3x4 model. The nine codon frequencies are estimated from the observed sequence data, while other parameters are estimated using a maximum likelihood optimization procedure that is divided into two stages that are reiterated until convergence: (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, the branch partition induced by $E(h)$ 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) can be compared to an alternative model, in which k is free to vary (different intensities associated to trait states). Since the models are nested, a likelihood ratio test (LRT) can be used, with the distribution of twice the log-likelihood difference between the two models approaching a χ_1^2 distribution with infinite amount of data under null conditions. In phylogenetic contexts, however, data size is not well defined and is rarely large enough, such that likelihood-ratio test statistics are not always well behaved (Goldman 1993; Whelan and Goldman 1999). As a more exhaustive alternative, the distribution of the likelihood ratio was also estimated using parametric bootstrap. Specifically, sequence data were generated by simulating sequence evolution along the input phylogeny using the parameter estimates of the null model. Given the original character data at the tips of the tree and the simulated sequence data, the maximum likelihood values of the null and alternative models were computed. This procedure was repeated 100 times to produce a distribution for the difference in the maximum log-likelihood values obtained between the two models under null conditions. The critical value for the LR statistic was determined as the 95th percentile of this empirical distribution for a significance threshold of α = 0.05. In the simulation study, rather than simulating artificial data sets based on the maximum likelihood estimates obtained by each inference of the null simulated data set, we simulated 200 replicates under the null model with the same parameter values with which the data sets were simulated (except for the selection intensity parameter k which was fixed at 1).

Simulating data sets under the TraitRELAX model

Simulations were used to investigate the sensitivity and precision of our method and to assess its accuracy in inferring the selection intensity parameter k. We first generated random trees with different numbers of taxa (16, 32, and 64) according to a birth–death process using INDELible (Fletcher and Yang 2009) with default parameters (speciation rate 0.3 and extinction rate 0.1). We then simulated character evolution along a given tree using Bio++ (Guéguen et al. 2013). As part of these simulations, the exact locations of character state transitions were recorded, thereby yielding branch partition based on the simulated (i.e., true) character histories. This partition was then used to simulate codon sequences based on the sequence model of TraitRELAX using INDELible (Fletcher and Yang 2009). Sequence data were simulated with various numbers of positions (l = 150, 300, and 600), representing a typical range of protein lengths (Zhang 2000). Unless otherwise stated, all simulations were conducted with total branches length of 4, 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 $kappa$ = 2. Different values of the selection intensity parameter were simulated: k = 0.2, 0.5, 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 no trait-related change in selection intensity. For each simulated scenario, 50 independent replicates were generated (each based on a different tree), with the exception of cases with $k = 1$, for which 200 replicates were generated. The larger number of replicates in the null case was used to determine the empirical threshold for the LR bootstrap-based comparison.

Biological data analyses

TraitRELAX was used to examine the evolution of Semenogelin II (SEMG2) in 24 primate species with respect to changes in the mating system. Character state assignments (1 = polygynandrous, 0 = other systems) were based on published literature (Dixson 1997; Dixson and Anderson 2002; O'Connor and Mundy 2009). Nine species were assigned as polygynandrous (multi-male:multi-female) and 15 were assigned as 'other' (encompassing nine species that are monogamous, single-male:single-female; five species that are polygynous, single-male:multi-female; and one that is polygamous). Sequence data of the SEMG2 gene was 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 the *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 primate tree topology was obtained from TimeTree (Hedges et al. 2006).

As a second biological example, we examined the utility of TraitRELAX in detecting changes in the selective type upon transitioning in the lifestyle 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) with default parameters. Positions with more than 90% gaps across sequences were filtered out.

In all analyses, branch lengths of the phylogeny were optimized using the M3 codon model (Yang and Nielsen 2002) as implemented in Bio++ (Guéguen et al. 2013). Comparisons between the null and alternative TraitRELAX models were conducted using the parametric bootstrap approach as detailed above.

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, as well as sequence and character data files. The program outputs the maximum-likelihood score for the null and alternative models as well as their inferred model parameters.

Results

Using TraitRELAX to infer associations between selection intensity and phenotypes

In this work we developed the TraitRELAX model, which enables the detection of associations between the evolution of phenotypic traits and the intensity of selective forces operating on coding sequences. The model emulates the joint evolution of a binary phenotypic trait and the coding sequences using two Markov processes. Character evolution is described by a two-state Markov process, while coding

sequence evolution is modeled using a variant of the branch-site model RELAX (Wertheim et al. 2015). In this joint model, each position in a coding sequence is subject to a selective regime (negative, neutral, or positive) whose intensity could increase (or decrease) upon a change in the phenotype. Namely, sequence evolution in segments of the phylogeny evolving under state '0' (background, BG) is described by a set of selection parameters: ω_0, ω_1 and ω_2, while segments evolving under state '1' are mapped to the foreground (FG) category. Shift in the selection intensity upon character transition is modeled using a selection intensity parameter k that modifies the selection parameters on the FG segments (${\omega_0}^k$, ${\omega_1}^k$ and ${\omega_2}^k$), such that k < 1 indicates relaxation of selection and k > 1 indicates intensification. A likelihood ratio test examines whether the coding sequences are subject to a trait-dependent shift in the selection intensity (i.e., whether k is significantly different from 1). The full details of this model and likelihood computations are given in “Materials and Methods”. In the following, we first evaluate the performance of the method on simulated data sets. We next apply it on real data to detect changes in the selection intensity in the primate SEMG2 gene upon transition to a polygynandrous mating system and in 68 γ-proteobacteria genes upon transition to endosymbiotic lifestyle.

Performance in simulations

Simulations were used to investigate the performance of TraitRELAX with regard to the false positive rate (FPR; i.e., detecting association between trait evolution and sequence evolution when no such association exists), sensitivity, and accuracy of parameter estimation.

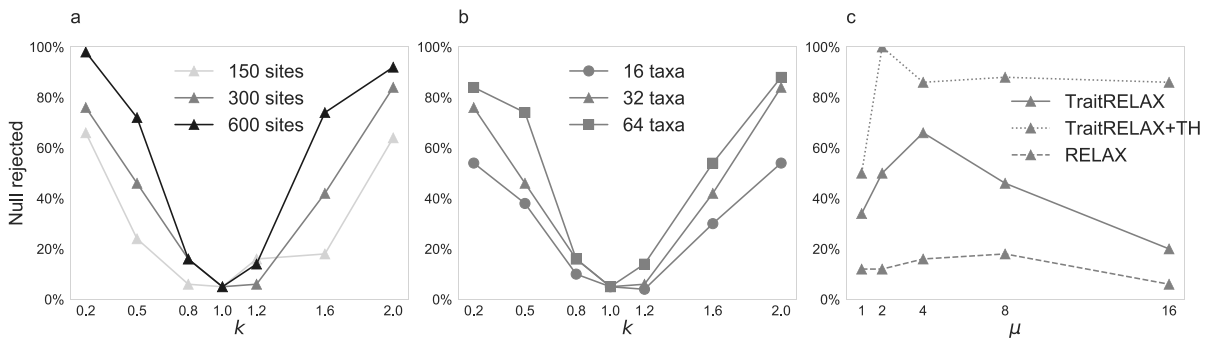
False positive rate and sensitivity

We analyzed the FPR by setting the selection intensity parameter k to 1, thereby simulating sequence data with no change in selection intensity along the phylogeny, and regarded as false positive any inference in which the null hypothesis (no association) is rejected according to the LR statistic. First, we examined the adequacy of p-values computed using the χ_1^2 approximation to the LR distribution under null conditions. Similar to the analysis of Levy Karin et al. (2017), we found deviations of the FPR from the expected 5%. For example, for alignments with 300 codons and a significant threshold of α = 0.05 (5% expected FPR), the FPR was 16.5%, 21% and 15.5%, for simulations with 16, 32, and 64 taxa, respectively (Simulation Set 4). For this reason, we empirically determined a threshold for the LR statistic using parametric bootstrapping, such that the FPR is fixed at 5% (see Materials and Methods). The thresholds determined this way for simulations with 300 codon positions were 6.57, 7.62, and 6.79, for 16, 32, and 64 taxa, respectively (compared with 3.84 using the χ_1^2 approximation). This procedure was performed for each combination of simulated parameters in the sensitivity analysis detailed below.

We analyzed the sensitivity of TraitRELAX based on simulations with various number of species, sequence lengths, and different values of the selection intensity parameter k. Our results indicated that the sensitivity of the method increases with data availability. This is true both for the number of sequence positions (**Error! Reference source not found.**2a) and the number of analyzed taxa (**Error! Reference source not found.**2b). For example, for simulations with 32 taxa and k = 0.5, TraitRELAX correctly rejected the null hypothesis in 46% of the simulations with 300 sequence positions and in 72% of the simulations with 600 positions. These numbers increased to 74% and 88% when using trees with 64 taxa. As may be

expected, increasing the magnitude of the selection intensity parameter also resulted in increased sensitivity (e.g., 74% correct rejection in simulations under k = 0.5 and 84% in simulations under k = 0.2; Fig. 2b). Usage of the empirical threshold instead of the theoretical one (derived from the χ_1^2 distribution) had little effect on the sensitivity of the method (Supp. Fig. 1). We next examined the effect of the number of character transitions by computing the sensitivity in simulations with increasing values of the character transition rate, μ. In these simulations, a bell-shape pattern was observed, where maximum sensitivity was obtained for intermediate values of μ and dropped when very small or very large number of character transitions were simulated (**Error! Reference source not found.**2c). In comparison, we also optimized the sequence model with branch partitions that were derived from the true (simulated) character histories, thus serving as a reference for the optimal performance that could be expected when eliminating any error in the reconstructed character history. As expected, the sensitivity in this case was consistently higher. The difference between using the true history rather than inferring it, was particularly apparent for high values of μ, representing simulations with a large number of simulated character transitions (Fig. 2c). Evidently, the complex character histories in such cases are harder to infer and reduce the ability of the method to detect genuine trait-related shifts in the selection intensity.

Figure 2 - Assessment of FPR and sensitivity based on simulations with different data sizes.



The percent of replicates for which the null model was rejected is shown for simulations with (a) 32 taxa and different number of codon positions, and (b) 300 positions and different number of taxa. (c) Sensitivity is shown for different values of the simulated μ parameter. Results are shown for standard TraitRELAX (solid line), TraitRELAX given the true character history (TH) and sequence optimization only (dotted line) and RELAX given a maximum parsimony partition (dashed line). In all TraitRELAX cases, parametric bootstrap was used to determine an empirical LR cutoff for rejecting the null hypothesis. When using RELAX, the LR threshold was based on the χ_1^2. For both methods the cutoff for rejection was set at α = 0.05.

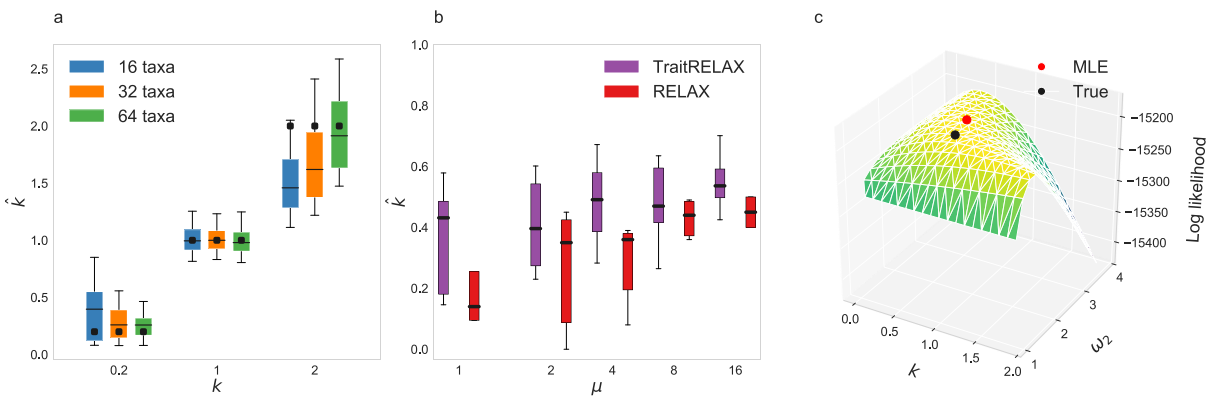
Accuracy of parameter estimation

We examined the accuracy of the method in estimating the selection intensity parameter k, as a function of the number of taxa and codon positions. In the null case of k = 1, the average inferred *k* was close to the simulated value, regardless of the number of taxa (Fig. 3a) or codon positions (Supp. Fig. 2d-f). In other cases, the inferred value of k was conservative with respect to the magnitude of the change in the intensity of selection, such that it was generally higher than the true value when relaxation of selection was simulated (k = 0.2) and lower than the true value when intensification was simulated (k = 2). This bias was most pronounced when only few taxa were available and nearly disappeared for the case of 64 taxa and 600 positions

(Fig. 3a, Supp. Fig. 2a,c,e). Similar to the sensitivity analysis, we found that the averaged error in the inferred value of *k* is most pronounced when the number of character transitions is very high or very low, and lowest when intermediate values of μ were simulated (Fig. 3b).

Finally, we studied the interplay between the inference of k and that of the ω parameters. The joint examination of the inferred values of k and the ω parameters suggests that underestimation of k (in the case of intensification of selection) is compensated by overestimation of ω_2 values, while the opposite is observed in the case of relaxation (i.e., overestimation of k and underestimation of ω_2) (Supp. Fig. 3). Indeed, examining the joint likelihood surface of k and ω_2 revealed that both the true (i.e., simulated) values and their maximum likelihood estimates are located along a ridge of high likelihood values, such that several (k, ω2) combinations receive similar log-likelihood scores (Fig. 3c), leading to difficulties in inferring the exact value of each parameter separately.

Figure 3 - Accuracy of inferring the selection intensity parameter.



(a) The distribution of the inferred values of k for simulations with 600 positions. For each simulated k value, the numbers of taxa are ordered from left to right: 16, 32, and 64 taxa. The horizontal lines within the box plots indicate the mean inferred

values and the simulated values are shown as black dots. (b) The inferred values of k are shown for simulations with 32 taxa, 300 positions, and k = 0.5 for different values of the simulated μ parameter. (c) Log likelihood surface over a grid of 300 combinations of ω_2 and k. The panel 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 black and red dots.

Comparison to RELAX

The simulated data were used to compare the performance of TraitRELAX to that of the branch-site model RELAX (Wertheim et al. 2015), as implemented in HyPhy (Kosakovsky Pond et al. 2005). To execute RELAX, a partition of the input phylogeny to background and foreground branches should be provided, corresponding to putative branches evolving under character state '0' and '1', respectively. To this end, we used the DELTRAN approach (Farris 1970; Swofford and Maddison 1987), implemented in Bio++, to infer the maximum parsimony solution given the character states of the extant taxa in each simulated data set.

Similar to the results obtained by TraitRELAX (Fig. 2), the sensitivity of RELAX increased with the intensity of the simulated value of k and with the size of the data (Supp. Fig. 4). Here too, RELAX had higher sensitivity in simulations with intermediate values of the character transition rate parameter μ, although the trend was shallower compared to that observed for TraitRELAX (Fig. 2c). Noticeably, the sensitivity of RELAX was consistently lower than that of TraitRELAX across all simulation scenarios (Fig. 2c, Supp. Fig. 4). For example, in simulations with 64 taxa,

600 positions, and k = 0.2, TraitRELAX correctly rejected the null in 98% of the cases, while RELAX in 44%.

Next, we evaluated the accuracy of both methods in inferring k in simulations with varying data size and with different number of character transitions. In all simulated conditions, the error and bias of RELAX was consistently higher than that of TraitRELAX (Fig. 3a-b, Supp. Fig. 5). Taken together, these analyses suggest that by integrating character trait data into a joint statistical framework, TraitRELAX is more robust than RELAX to uncertainty in trait evolution.

Association between selection type and intensity upon transition to polygynandry in primates and to endosymbiosis in γ-proteobacteria

We demonstrate the use of TraitRELAX to detect trait-related changes in selection intensity using two biological data sets: one concerning transitions in mating systems among primates, and the other between free living and endosymbiont bacteria. Transition among mating systems in primates has been previously associated with shifts in the selection patterns operating on several reproductive genes, suggesting that sperm competition among males in polygynandrous societies, in which multiple males complete for the same female, is an underlying force of sexual selection (Ramm et al. 2008). Elevated nucleotide substitution rate and dN/dS ratio in polygynandrous lineages has been detected in the Semenogelin II (SEMG2) gene, whose encoded protein takes part in the formation of a gel matrix that encases ejaculated spermatozoa (Dorus et al. 2004; Hurle et al. 2007; O'Connor and Mundy 2009; Levy Karin et al. 2017). Here, we used TraitRELAX to investigate associations between transitions in primates mating systems and shifts intensity of selection in

SEMG2. The SEMG2 gene coding sequences of 24 primate species were collected (see Material and Methods), nine of which are polygynandrous and 15 that participate in other mating systems (including monogamous, polygamous, and polygamous systems). The distribution of mating systems along the phylogeny indicated that there were at least six transitions among this group of analyzed species (based on the maximum parsimony reconstruction, which can be considered as lower bound). A comparison between the null and alternative models of TraitRELAX revealed significant intensification of selection under the polygynandrous state (k = 2.13, p-value = 0.03). We found that most of the SEMG2 sites are dominated by neutral selection and are not affected by transitions in the mating system ($\widehat{\omega_1}$ = 1, $\widehat{p_1}$ = 0.76), while approximately 17% of the sites are subject to positive selection, which is intensified along the polygynandrous lineages (with ω_2 = 2.33 and 6.08 in the 'other' and polygynandrous states, respectively). This proportion is in agreement with a previous study that detected positive selection in the SEMG2 gene with a proportion 0.15 of sites and ω_2 = 3.27 (Ramm et al. 2008).

Next, we investigated whether transitions in the lifestyle of γ-proteobacteria are associated with shifts in the selection intensity operating on 68 housekeeping genes. Unlike free-living bacteria, bacterial endosymbionts rely on their host for mandatory functions, such as translation and metabolite decomposition (Husník et al. 2011). The smaller population size of the endosymbionts, caused by their inability to replicate independently from the host, reduces the efficiency of natural selection and can result in accumulation of deleterious mutations and overall elevated rates of nonsynonymous substitutions and radical amino acid replacements (Muller 1964; Moran 1996; Funk et al. 2001; Herbeck et al. 2003; Wernegreen 2011). Indeed, a relaxation of selection in endosymbiotic bacteria, relative to their free-living

progenitors, was detected using the RELAX model (Wertheim et al. 2015). However, that analysis was performed on an alignment of 69 single-copy genes that were treated as a single concatenated gene region and thus did not provide information concerning the identity of specific genes that are subject to differential selection in endosymbionts and the magnitude by which they are affected. Here, we used TraitRELAX to analyze 68 of these genes separately (one gene was excluded from analysis due to the sparse sequence data). The analysis was performed on 50 species of γ-proteobacteria (16 endosymbionts – state 1, and 34 free living – state 0), with at least seven transitions between the two states according to maximum parsimony reconstruction. Comparing the null and alternative TraitRELAX models revealed significant change in selection intensity under endosymbiotic lifestyle in sixteen out of the 68 examined genes (p-*value* < 0.05 following the Benjamini-Hochberg false discovery rate correction for multiple testing (Benjamini and Hochberg 1995), with relaxation of selection in ten genes (with $\hat{k}$ ranging from 0.78 to 0.94) and intensification in six genes (with $\hat{k}$ ranging from 1.06 to 1.36). The relaxed genes were either involved in metabolite decomposition or in tRNA selection processes, while the intensified genes were involved in transcription and translation (Supp. Table 1). In all cases, the inferred value of ω_2 was 1, suggesting that in either state positive selection has not operated on these genes and the change in selection intensity is with respect to purifying selection only. For the ten relaxed genes, the mean estimated values of ω_0 and $\omega_1$were 0.016 and 0.140 in the free-living state and 0.026 and 0.178 in the endosymbiotic state, respectively and for the six intensified genes the mean values of ω_0 and ω_1 were 0.013 and 0.137 in the free-living state and 0.006 and 0.097 in the endosymbiotic state.

Discussion

In this study we developed TraitRELAX, a new method for detecting associations between organismal traits and selection intensity at the codon level. Within the framework of the method, an association is expressed as a dependency between phenotypic trait evolution and changes in patterns of selection operating on coding sequences. 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 codon site should not be treated independently. Since the exact timings of phenotypic transitions are unknown, we consider possible scenarios that led to the observed phenotypic states at the extant taxa by using stochastic mappings, such that each mapping induces a possible partition as input for the sequence model and all mappings are considered simultaneously. This, however, entails heavy computations since likelihood computation of the sequence data is performed multiple times for each parameter combination examined.

To speed up the likelihood computation, we adopted a similar approach to that presented in Levy Karin et al. (2017), in which multiple stochastic mappings are represented by a single expected history (Equation 10). Here, unlike in their paper, all possible mappings were summarized analytically using the rewards method (Minin and Suchard 2008). This approximation limits the number of transitions along each branch to at most two (see Supp. Text 1). When many transitions have occurred along a branch, this approach is expected to yield inferior approximation to the exhaustive approach that uses multiple stochastic mappings (Equation 9). However, most organismal traits of interest are key features of a lineage and are expected to evolve rather slowly, such that multiple or back transitions on the same branch are expected to be rare. In such cases, the expected history approximation should

perform well. In addition to reducing running times, the approximation approach increases the stability of the computation as it is robust to the stochasticity of the sampled mappings. Another alternative approximation can be obtained by clustering the set of stochastic mappings into several groups and perform the computation on a representative of each, weighted by its relative frequency. Such an approximation was adopted by Jones et al. (in press). The method groups stochastic mappings in case phenotype transitions have occurred on the same branches and ancestral state assignments are identical. While allowing for multiple unique phenotype histories, this approach practically restricts phenotype transitions to occur at internal nodes and limits the number of transitions to at most one per branch.

TraitRELAX can receive as input an ultrametric tree, in which branch lengths are proportional to time. Such trees well describe traits whose rate of change is time-dependent, such as organismal habitat or lifestyle. Alternatively, a non-ultrametric tree can be provided, better reflecting traits whose rate of change is proportional to the amount of genetic change or those whose transition rate is likely associated with the generation time. Because the reconstruction of time-calibrated phylogenies suffers from high degrees of uncertainty (Dornburg et al. 2012; Wertheim et al. 2012; Dos Reis et al. 2015; Bromham et al. 2018), one may prefer the use of uncalibrated trees, unless there is a strong evidence that the evolution of the trait in question is time dependent.

Our simulation study demonstrated that the sensitivity and accuracy of the method depend on the size of the data, the magnitude of selection-intensity shift upon phenotype transition, and the number of character trait transitions. When there are only few character state transitions across the phylogeny, there is little information to

accurately infer changes in selection intensity. When the character history is too complex (high μ values), there is not enough information in the character data of extant species to accurately reconstruct the character history, and thus, reducing the ability to detect phenotype effect on patterns of selection at the sequence level. Notably, in small data sets, failure to reject the null hypothesis could indicate no association between the examined trait and coding sequence evolution, but could also indicate low statistical power to detect such an association, especially if the inferred value of the selection intensity parameter k is far from 1. To test whether the data are sufficient for reliable inference, sensitivity analysis should be conducted (Boettiger et al. 2012). To this end, simulations should be used to generate artificial data sets with similar data characteristics as the empirical data (in terms of number of taxa, sequence sites, and number of character transitions), and with various values of the k parameter that are increasingly distant from 1. Such artificial data sets, as generated in this study, can guide us on the possible values of k for which the null hypothesis could be reliably rejected.

When compared to the existing branch-site model RELAX (Wertheim et al. 2015), TraitRELAX demonstrated higher sensitivity to detect trait-related changes in selection intensity and higher accuracy in estimating the k parameter governing the direction and magnitude of the selection shift. Multiple differences between the methods can explain the observed results. First, unlike RELAX, TraitRELAX considers the uncertainty in the trait history and additionally allows trait transitions to occur anywhere along a branch rather than only at hypothesized speciation events (i.e., internal nodes of the phylogeny). Second, in TraitRELAX we constrain the selective regimes assigned to each position to remain the same across the phylogeny, while the implementation of RELAX in HyPhy allows changes in the

selective regime of a site (e.g., from purifying to positive selection) when transitioning across branches (for more details on the difference between these implementations see Supp. Text 3). This latter approach could result in confounding effects between the selection intensity parameter k and the ω parameters and could lead to erroneous estimates of k (Supp. Fig. 5a). For example, when the inferred value of k is very small, $\widehat{\omega_0{}^{FG}}$ may resemble $\omega_1{}^{BG}$ (i.e., $\widehat{\omega_1} \approx \widehat{\omega_0^k}$). In such a case, the null scenario in which a sequence position evolves under neutral selection throughout the entire phylogeny yields similar likelihood to an alternative scenario in which the same position shifts from neutral selection in the BG branches to purifying selection in the FG branches. Finally, we note that the sensitivity and accuracy of RELAX observed here were lower than those reported by Wertheim et al. (2015), most probably due to the smaller data size examined in our simulations (reaching up to 600 codon positions, compared to 21,154 positions in theirs).

A main benefit of the statistical framework presented here is that it simultaneously emulates the evolution of both the examined phenotype and that of the coding sequences. As such, it deviates from many widely used models that rely on a two-step inference scheme, in which the evolution of the phenotype is first reconstructed, and this inference is then incorporated into models of sequence evolution while ignoring any uncertainty in the preceding inference step. The TraitRELAX joint phenotype-sequence approach can be extended in various ways, most notably, by selecting a different branch-site codon model as the underlying sequence model. For example, a generalization of our approach can be applied by mapping independent codon models (Yang et al. 2000) to each phenotypic state, while having some of the parameters shared between the models to reduce the overall number of parameters. This will enable the detection of more subtle associations between the evolution of

the phenotype and that of the examined sequences. One such example was recently presented by Jones et al. (in press) who developed the PG-BSM method that, in common with the approach presented here, integrates trait and coding sequence data in a joint statistical framework. The PG-BSM method is designed to detect phenotype-dependent shifts in adaptation, that is, persistent or transient increase of ω at a site dictated by the trait under study. The PG-BSM method further allows for a proportion of the sequence positions to evolve in a trait-independent manner. Incorporating such a parameter has been proved useful in other models (Levy Karin et al. 2017), and thus may be a beneficial extension to TraitRELAX. Lastly, a useful extension of TraitRELAX would allow consideration of categorical phenotypes that consist of more than two states, which will also allow the examination of multiple, perhaps correlated, traits. 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. Such modelling would provide more realistic scenarios regarding the many pathways by which a combination of organismal phenotypes governs evolutionary processes at the genomic level.

Funding

This study was supported in part by a fellowship to KH from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University and by the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics to IM. ELK is a recipient of a FEBS long-term fellowship and is an EMBO non stipendiary long-term fellow.

Acknowledgments

The authors thank Prof. Sergei L. Kosakovsky Pond for his useful advice during the development process of the method and Julien Y. Dutheil for his assistance in the implementation of the method.

References

Ashman T.L., Bachtrog D., Blackmon H., Goldberg E.E., Hahn M.W., Kirkpatrick M., Kitano J., Mank J.E., Mayrose I., Ming R., Otto S.P., Peichel C.L., Pennell M.W., Perrin N., Ross L., Valenzuela N., Vamosi J.C. 2014. Tree of Sex: A database of sexual systems. Sci. Data. 1:1–8.

Benjamini Y., Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B. 57:289–300.

Boettiger C., Coop G., Ralph P. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution (N. Y). 66:2240–2251.

Brent R.P. 1974. Algorithms for Minimization Without Derivatives. .

Bromham L., Duchêne S., Hua X., Ritchie A.M., Duchêne D.A., Ho S.Y.W. 2018. Bayesian molecular dating: opening up the black box. Biol. Rev. 93:1165–1191.

Burch C.L., Chao L. 1999. Evolution by small steps and rugged landscapes in the RNA virus φ6. Genetics. 151:921–927.

Dixson A.F. 1997. Evolutionary perspectives on primate mating systems and behavior. Ann. N. Y. Acad. Sci.

Dixson A.F., Anderson M.J. 2002. Sexual selection, seminal coagulation and copulatory plug formation in primates. Folia Primatol.

Dornburg A., Brandley M.C., McGowen M.R., Near T.J. 2012. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). Mol. Biol. Evol. 29:721–736.

Dorus S., Evans P.D., Wyckoff G.J., Sun S.C., Lahn B.T. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nat. Genet. 36:1326–1329.

Drory O., Frolow F., Nelson N. 2004. Crystal structure of yeast V-ATPase subunit C reveals its stator function. EMBO Rep. 5:1148–1152.

Endo T., Ikeo K., Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. 13:685–690.

Farris J.S. 1970. Methods for Computing Wagner Trees. Syst. Biol.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Feng Y.L., Wicke S., Li J.W., Han Y., Lin C.S., Li D.Z., Zhou T.T., Huang W.C., Huang L.Q., Jin X.H. 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 F.D., Bernard G.D., Cuevas C.I., Sison-Mangus M.P., Prudic K.L., Briscoe A.D. 2007. Adaptive evolution of color vision as seen through the eyes of

butterflies. In the Light of Evolution. p. 187–204.

Funk D.J., Wernegreen J.J., Moran N.A. 2001. Intraspecific variation in symbiont genomes: bottlenecks and the aphid-Buchnera association. Genetics. 157:477–489.

Gertow K., Bellanda M., Eriksson P., Boquist S., Hamsten A., Sunnerhagen M., Fisher R.M. 2004. Genetic and structural evaluation of fatty acid transport protein-4 in relation to markers of the insulin resistance syndrome. J. Clin. Endocrinol. Metab. 89:392–399.

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. 170:313–326.

Goldman N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

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

Guéguen L., Gaillard S., Boussau B., Gouy M., Groussin M., Rochette N.C., Bigot T., Fournier D., Pouyet F., Cahais V., Bernard A., Scornavacca C., Nabholz B., Haudry A., Dachary L., Galtier N., Belkhir K., Dutheil J.Y. 2013. Bio++: Efficient extensible libraries and tools for computational molecular evolution. Mol. Biol. Evol. 30:1745–1750.

Hedges S.B., Dudley J., Kumar S. 2006. TimeTree: A public knowledge-base of divergence times among organisms. Bioinformatics. 22:2971–2972.

Herbeck J.T., Funk D.J., Degnan P.H., Wernegreen J.J. 2003. A conservative test of genetic drift in the endosymbiotic bacterium buchnera: slightly deleterious mutations in the chaperonin groel. Genetics. 165:1651–1660.

Hersch-Green E.I., Myburg H., Johnson M.T.J. 2012. Adaptive molecular evolution of a defence gene in sexual but not functionally asexual evening primroses. J. Evol. Biol. 25:1576–1586.

Hestenes M.R., Stiefel E. 1952. Methods of conjugate gradients for solving linear systems. Journal of Research of the National Bureau of Standards. p. 409–436.

Hurle B., Swanson W., Green E.D. 2007. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. Genome Res. 17:276–286.

Husník F., Chrudimský T., Hypša V. 2011. Multiple origins of endosymbiosis within the Enterobacteriaceae (γ-Proteobacteria): Convergence of complex phylogenetic approaches. BMC Biol. 9.

Isshiki M., Ishida T. 2019. Molecular evolution of the semenogelin 1 and 2 and mating system in gibbons. Am. J. Phys. Anthropol. 168:364–369.

Jensen-Seaman M.I., Li W.H. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. J. Mol. Evol. 57:261–270.

Jones C.T., Youssef N., Susko E., Bielawski J.P. In press. A phenotype-genotype codon model for detecting adaptive evolution. Syst. Biol.:1–49.

Kattge J., Díaz S., Lavorel S., Prentice I.C., Leadley P., Bönisch G., Garnier E., Westoby M., Reich P.B., Wright I.J., Cornelissen J.H.C., Violle C., Harrison S.P., Van Bodegom P.M., Reichstein M., Enquist B.J., Soudzilovskaia N.A., Ackerly D.D., Anand M., Atkin O., Bahn M., Baker T.R., Baldocchi D., Bekker R., Blanco C.C., Blonder B., Bond W.J., Bradstock R., Bunker D.E., Casanoves F., Cavender-Bares J., Chambers J.Q., Chapin F.S., Chave J., Coomes D., Cornwell W.K., Craine J.M., Dobrin B.H., Duarte L., Durka W., Elser J., Esser G., Estiarte M., Fagan W.F., Fang J., Fernández-Méndez F., Fidelis A., Finegan B., Flores O., Ford H., Frank D., Freschet G.T., Fyllas N.M., Gallagher R. V., Green W.A., Gutierrez A.G., Hickler T., Higgins S.I., Hodgson J.G., Jalili A., Jansen S., Joly C.A., Kerkhoff A.J., Kirkup D., Kitajima K., Kleyer M., Klotz S., Knops J.M.H., Kramer K., Kühn I., Kurokawa H., Laughlin D., Lee T.D., Leishman M., Lens F., Lenz T., Lewis S.L., Lloyd J., Llusià J., Louault F., Ma S., Mahecha M.D., Manning P., Massad T., Medlyn B.E., Messier J., Moles A.T., Müller S.C., Nadrowski K., Naeem S., Niinemets Ü., Nöllert S., Nüske A., Ogaya R., Oleksyn J., Onipchenko V.G., Onoda Y., Ordoñez J., Overbeck G., Ozinga W.A., Patiño S., Paula S., Pausas J.G., Peñuelas J., Phillips O.L., Pillar V., Poorter H., Poorter L., Poschlod P., Prinzing A., Proulx R., Rammig A., Reinsch S., Reu B., Sack L., Salgado-Negret B., Sardans J., Shiodera S., Shipley B., Siefert A., Sosinski E., Soussana J.F., Swaine E., Swenson N., Thompson K., Thornton P., Waldram M., Weiher E., White M., White S., Wright S.J., Yguel B., Zaehle S., Zanne A.E., Wirth C. 2011. TRY - a global database of plant traits. Glob. Chang. Biol. 17:2905–2935.

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 S.L., Frost S.D.W., Muse S. V. 2005. HyPhy: Hypothesis testing using phylogenies. Bioinformatics. 21:676–679.

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 P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.

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 C.I. 2006. The accumulation of deleterious mutations in rice genomes: A hypothesis on the cost of domestication. Trends Genet. 22:126–131.

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.

Melamed D., Mark-Danieli M., Kenan-Eichler M., Kraus O., Castiel A., Laham N., Pupko T., Glaser F., Ben-Tal N., Bacharach E. 2004. The conserved carboxy

terminus of the capsid domain of human immunodeficiency virus type 1 gag protein is important for virion assembly and release. J. Virol. 78:9675–9688.

Metropolitan T., Genetics M.E. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol.:678–687.

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

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

Muller H.J. 1964. The role of recomination to mutational advance. Mutat. Res. Mol. Mech. Mutagen. 1:2–9.

Muse S. V, Gaut B.S., 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 L.G., 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. 48:2209–2219.

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

O'Connor T.D., Mundy N.I. 2009. Genotype-phenotype associations: Substitution models to detect evolutionary associations between phenotypic variables and genotypic evolutionary rate. Bioinformatics. 25.

Parr C.S., Wilson N., Leary P., Schulz K.S., Lans K., Walley L., Hammock J.A., Goddard A., Rice J., Studer M., Holmes J.T.G., Corrigan R.J. 2014. The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. Biodivers. Data J. 2.

Ramm S.A., Oliver P.L., Ponting C.P., Stockley P., Emes R.D. 2008. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. Mol. Biol. Evol. 25:207–219.

Dos Reis M., Thawornwattana Y., Angelis K., Telford M.J., Donoghue P.C.J., Yang Z. 2015. Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. Curr. Biol. 25:2939–2950.

Rice A., Glick L., Abadi S., Einhorn M., Kopelman N.M., Salman-Minkov A., Mayzel J., Chay O., Mayrose I. 2015. The Chromosome Counts Database (CCDB) - a community resource of plant chromosome numbers. New Phytol. 206:19–26.

Roan N.R., Müller J.A., Liu H., Chu S., Arnold F., Stürzel C.M., Walther P., Dong M., Witkowska H.E., Kirchhoff F., Münch J., Greene W.C. 2011. Peptides released by physiological cleavage of semen coagulum proteins form amyloids that enhance HIV infection. Cell Host Microbe. 10:541–550.

Roquet C., Coissac É., Cruaud C., Boleda M., Boyer F., Alberti A., Gielly L., Taberlet P., Thuiller W., Van Es J., Lavergne S. 2016. Understanding the evolution of holoparasitic plants: The complete plastid genome of the holoparasite Cytinus hypocistis (Cytinaceae). Ann. Bot. 118:885–896.

Swofford D.L., Maddison W.P. 1987. Reconstructing ancestral character states under

Wagner parsimony. Math. Biosci. 87:199–229.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Am. Math. Soc. Lect. Math. Life Sci. 17:57–86.

Tavares W.C., Seuánez H.N. 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. 245:25–31.

Wernegreen J.J. 2011. Reduced selective constraint in endosymbionts: Elevation in radical amino acid replacements occurs genome-wide. PLoS One. 6.

Wernersson R., Pedersen A.G. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 31:3537–3539.

Wertheim J.O., Fourment M., Kosakovsky Pond S.L. 2012. Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. Mol. Biol. Evol. 29:451–456.

Wertheim J.O., Murrell B., Smith M.D., Pond S.L.K., 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. 16:1292–1299.

Wicke S., Müller K.F., DePamphilis C.W., Quandt D., Bellot S., Schneeweiss G.M. 2016. Mechanistic model of evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants. Proc. Natl. Acad. Sci. U. S. A. 113:9045–9050.

Wu C.I., Li W.H., Shen J.J., Scarpulla R.C., Limbach K.J., Wu R. 1986. Evolution of cytochrome c genes and pseudogenes. J. Mol. Evol. 23:61–75.

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

Yang Z., Nielsen R., Goldman N., Pedersen A.M.K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431–449.

Zhang J. 2000. Protein-length distributions for the three domains of life. Trends Genet. 16:107–109.

Zhang Q., Zhang F., Chen X.H., Wang Y.Q., Wang W.Q., Lin A.A., Cavalli-Sforza L.L., Jin L., Huo R., Sha J.H., Li Z., Su B. 2007. Rapid evolution, genetic variations, and functional association of the human spermatogenesis-related gene NYD-SP12. J. Mol. Evol. 65:154–161.