Syst. Biol. 66(6):917–933, 2017

DOI:10.1093/sysbio/syx032
Advance Access publication February 16, 2017

An Integrated Model of Phenotypic Trait Changes and Site-Specific Sequence Evolution

ELI LEVY KARIN[1,2], SUSANN WICKE[3], TAL PUPKO[1,*], AND ITAY MAYROSE[2,*]

[1]*Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel;* [2]*Department of Molecular Biology & Ecology of Plants, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel; and* [3]*Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

**Correspondence to be sent to: Department of Molecular Biology & Ecology of Plants, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel. E-mails: itaymay@post.tau.ac.il; talp@post.tau.ac.il.*

Received 10 June 2016; reviews returned 23 January 2017; accepted 24 January 2017
Associate Editor: Luke Harmon

Abstract.—Recent years have seen a constant rise in the availability of trait data, including morphological features, ecological preferences, and life history characteristics. These phenotypic data provide means to associate genomic regions with phenotypic attributes, thus allowing the identification of phenotypic traits associated with the rate of genome and sequence evolution. However, inference methodologies that analyze sequence and phenotypic data in a unified statistical framework are still scarce. Here, we present TraitRateProp, a probabilistic method that allows testing whether the rate of sequence evolution is associated with a binary phenotypic character trait. The method further allows the detection of specific sequence sites whose evolutionary rate is most noticeably affected following the character transition, suggesting a shift in functional/structural constraints. TraitRateProp is first evaluated in simulations and then applied to study the evolutionary process of plastid plant genomes upon a transition to a heterotrophic lifestyle. To this end, we analyze 20 plastid genes across 85 orchid species, spanning different lifestyles and representing different genera in this large family of flowering plants. Our results indicate higher evolutionary rates following repeated transitions to a heterotrophic lifestyle in all but four of the loci analyzed. [Evolutionary models; evolutionary rate; genotype–phenotype; orchids; plastome; rate shift.]

Numerous studies have been devoted to identify phenotypic traits that are associated with the rate of genome and sequences evolution. Classic examples for such studies include reports on the correlation between the rate of nucleotide substitution and various factors such as body mass (as reviewed in Martin and Palumbi 1993), body size and temperature (Gillooly et al. 2005), metabolic rate (Martin 1995), and generation time (Li et al. 1996; Lehtonen and Lanfear 2014). More recently, it was proposed that in vertebrates, lineages with a preformation mechanism of primordial germ cell specification exhibit an accelerated rate of evolution as compared to lineages with an epigenesis mechanism (Evans et al. 2014), whereas in flowering plants rates of molecular evolution were inferred to be lower in woody compared to herbaceous taxa (Smith and Donoghue 2008).

In recent years, alongside the genomic revolution, there is a constant rise in the accumulation of character trait data, representing a range of phenotypes, including morphological and genomic features, ecological preferences, and life history characteristics. Databases, such as the Coleoptera Karyotype Database (Blackmon and Demuth 2015), Encyclopedia of Life – TraitBank (Parr et al. 2014), the Tree-of-Sex (Tree of Sex Consortium 2014), and the Chromosome Counts Database (Rice et al. 2015) catalog and document character traits for a wide variety of species. This abundance of phenotypic data provides means to associate alterations in genomic processes with phenotypic attributes of the organisms whose sequences are being analyzed.

However, to date, inference methods that consider sequence data as well as phenotypic data in a joint statistical framework are scarce. Notable exceptions include CoEvol (Lartillot and Poujol 2011), TraitRate (Mayrose and Otto 2011), and a method developed by O'Connor and Mundy (2009, 2013) (hereafter referred to as "OM"). Whereas CoEvol analyzes the correlation of sequence data and continuous trait data, TraitRate and OM focus on detecting associations between a discrete phenotypic state and the rate of evolution.

Specific implementation details put aside, one of the key requirements of a joint phenotype–genotype modeling framework (see "Materials and Methods" section, Fig. 1) is that the evolution of the phenotypic trait is consistent between sequence positions, assuming the same pattern of character state changes applies to all sequence positions. The OM method follows the coevolutionary model of Pagel (1994) by presenting a single Markov process where each state is a pair of sequence and trait characters (Fig. 2a). Yet, the OM model does not impose consistency of the phenotypic trait evolution among sites, that is, the ancestral probabilities of the phenotypic states in different sequence positions are not constrained to be the same. Figure 2b demonstrates this with an example where the marginal probability of phenotypic state "0" in an ancestral node of a phylogeny is not necessarily equal between two positions.

Alternatively, TraitRate combines models of sequence evolution and of phenotypic trait evolution into one likelihood framework by first reconstructing possible evolutionary histories of the phenotypic trait along the phylogeny (Mayrose and Otto 2011). Each such history is consistent with the observed phenotypic state values of the extant species. The method then assumes distinct processes of sequence evolution depending on the phenotypic state according to the reconstructed character history. Mayrose and Otto (2011) applied their method to demonstrate that in the crustacean

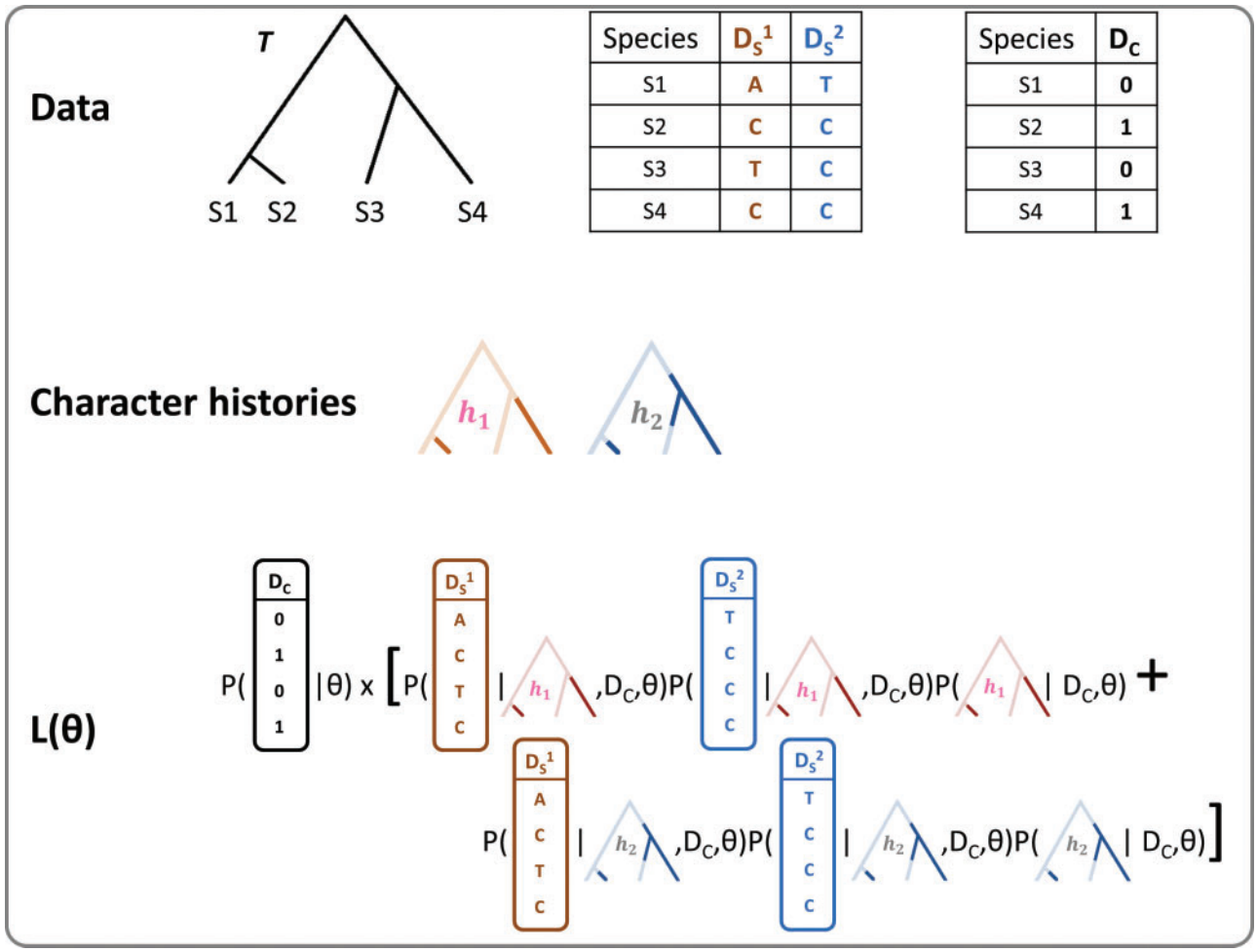

FIGURE 1. A joint phenotype–genotype likelihood framework. An example input for a joint likelihood model. For the sake of simplicity, only two possible character histories, h_1 and h_2, are accounted for, replacing the integral in the general likelihood function (see main text) with a sum.

genus *Daphnia*, habitat shifts from freshwater to saline environments are associated with an elevated substitution rate, probably due to the mutagenic effect of high salt concentration.

Notably, the original TraitRate model assumed that all sequence sites in the examined locus are influenced by the analyzed trait. However, it is quite often the case that not all positions in a given sequence are under the same selective regime (Yang 1994; Pupko and Galtier 2002; Guindon et al. 2004; Rodrigue et al. 2010; Kosakovsky Pond et al. 2011). Specifically, we expect positions that code for functionally important regions such as a protein binding domain or those within an active site to evolve under a stronger purifying selective pressure compared to positions that are not directly involved in the primary function of the protein. Similarly, we expect that the character trait effect is heterogenous across the sequence; phenotypes like those associated with domestication, mating system, or pathogenicity may cause shifts in the mutational and/or selective forces in specific sites within particular loci. This distinction between classes of positions motivated us to develop TraitRateProp, a methodology that allows the detection of cases in which the evolutionary rate of a certain proportion of sites within the analyzed genomic region depends on the phenotypic state.

The extremely versatile orchid family includes an outstanding variety of lifeforms and lifestyles (e.g., Chase 2001; Givnish et al. 2015). Without exception, all orchids rely on a mycorrhizal association during germination to develop photosynthetic seedlings. While most orchids then grow autotrophically, some lineages pursue a heterotrophic lifestyle, obtaining nutrients through their mycorrhiza from another plant (e.g., McCormick et al. 2004, 2012). Earlier analyses focused on a small subset of nonphotosynthetic taxa or on sister taxa comparisons of arbitrarily selected nonphotosynthetic and photosynthetic species. These studies suggested that a heterotrophic lifestyle triggers convergent gene losses and occasionally rate accelerations in plastid genomes (plastomes), including in genes that are not primarily involved in photosynthesis (e.g., dePamphilis 1995; Logacheva et al. 2011; Bromham et al. 2013; Wicke 2013; Schelkunov et al. 2015; Cusimano and Wicke 2016). Here, we applied TraitRateProp to analyze 20 commonly retained plastid genes across 85 orchid

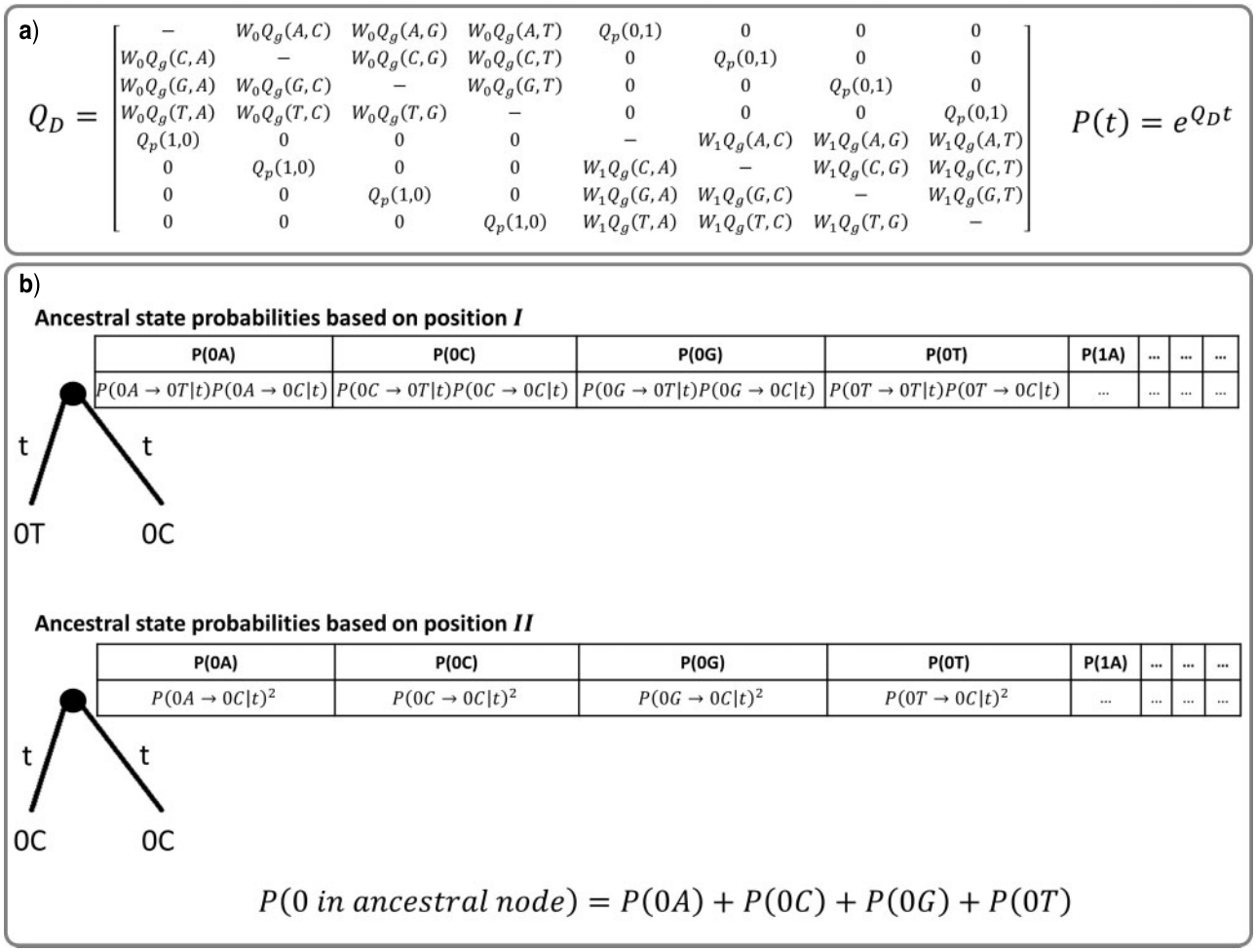

FIGURE 2. The OM model. The rate matrix Q_D of the OM model (a) and an example for inconsistent character history reconstruction for two different sequence positions (b). The marginal likelihood of character '0' at the ancestral node, as depicted by the equation at the bottom of the figure, is dependent on the observed nucleotides at each position.

species, representing 48 genera in this large family of flowering plants. This allowed us to study the evolutionary patterns exhibited by plants that transition from a photoautotrophic to a heterotrophic lifestyle in a broader phylogenetic context, within a unified statistical framework.

MATERIALS AND METHODS

The TraitRateProp Method

Input.—TraitRateProp requires as input sequence data (D_S) in the form of a multiple sequence alignment (MSA), a rooted ultrametric species tree with specified branch lengths (T), and the character data (D_C) describing the trait states of the extant species, coded as either "0" or "1." We note that T is assumed to be known prior to the analysis of D_S and is not reconstructed from the data. Further, the branch lengths of T are informative up to a multiplicative factor, as they intend to measure the relative time of divergence.

Phenotypic trait evolution.—We assume a two-state Markov model, M_C, to describe the evolution of the phenotypic trait along the tree. Specifically, this process is defined by the rate matrix Q_C:

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

where $\pi_1 = 1 - \pi_0$ is a parameter governing the rate of change from state "0" to "1" and μ is a factor designated to adapt the branch lengths of the phylogenetic tree to the expected number of character changes per unit time.

Sequence evolution.—Any of the widely used models of sequence evolution can be used to describe the process of sequence evolution (M_S) along the phylogenetic tree. In all analyses described here, the HKY $+\Gamma$ model was used (Uzzell and Corbin 1971; Wakeley 1993). The sequence rate matrix Q_S is controlled by the κ parameter describing the ratio between transitions and transversions and the nucleotide frequencies. The model implements among-site rate variation using a discretized

gamma distribution (with four rate categories), of mean 1 and shape parameter α (as in Yang 1994).

Connecting the trait and sequence evolutionary processes.—The model assumes that a proportion $(1-p)$ of sites evolves independently of the phenotypic character, and a proportion p of "phenotype-dependent positions" evolves so that their substitution rate varies depending on whether the phenotypic trait is in state "0" or "1." Specifically, a parameter r_1 is assumed when the character state is "1," and a parameter r_0 when the character state is "0." Thus, for a phenotype-dependent position, the sequence rate matrix, Q_S, is multiplied by either r_1 or r_0, according to the character state. Let r denote the ratio between r_1 and r_0. If $r>1$, a "0" → "1" transition in the character trait leads to an accelerated sequence evolution by the factor r relative to the rate in the "0" state.

Let H denote the history of character state transitions along the tree. For each branch i of length b_i, the fraction of time spent in state "1," f_{i1}, and the fraction of time spent in state "0," f_{i0}, can be extracted from H. In addition, we denote the fraction of time across the entire tree that was spent in state "1" as f_1, and the fraction of time in state "0" as f_0. These fractions can be computed as follows:

$$f_0=\frac{\sum_j b_j f_{j0}}{\sum_j b_j},\quad f_1=1-f_0=\frac{\sum_j b_j f_{j1}}{\sum_j b_j}$$

where the sum is over all branches in the phylogeny. Using the above notations, the average rate matrix along a branch i for a phenotype-dependent position is:

$$Q_{S_i}=r_0\times Q_S\times f_{i0}+r_1\times Q_S\times f_{i1}.$$

The average rate of phenotype-dependent positions across the entire tree is $r_0\times Q_S\times f_0+r_1\times Q_S\times f_1$. We require that this average rate matrix is equal to Q_S, which is the rate matrix of sequence positions whose evolution is not associated with phenotypic changes. Thus, we impose:

$$Q_S=r_0\times Q_S\times f_0+r_1\times Q_S\times f_1$$

As Q_S is invertible, we obtain:

$$r_0\times f_0+r_1\times f_1=1$$

In other words, the acceleration and deceleration of rates due to phenotypic change has no impact on the total number of changes along the tree. Rather, in phenotype-dependent positions, when $r>1$, most of sequence substitutions will occur in those parts of the tree that evolved under the state "1."

Replacing r_1 by $r\times r_0$ in the equation above, we obtain:

$$r_0=\frac{1}{f_0+r\times f_1},\quad r_1=\frac{r}{f_0+r\times f_1}$$

These equations allow us to express Q_{S_i} in terms of r, Q_S, f_{i0}, f_{i1}, f_0, and f_1.

More details regarding the average rate matrix Q_{S_i} and the specific implementation details are provided as Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.d4j55.2.

Likelihood computation.—The likelihood of the model is the joint probability of D_S and D_C given the free parameters θ. This expression can be termed as the probability to observe D_C times the probability to observe D_S conditioned on having observed D_C. Under these settings, likelihood computations based on the sequence data require the knowledge of the character state in each part of T, that is, the complete reconstructed history of character changes. This history is generally unknown and thus, the marginal probability of D_S given D_C and θ can be obtained by integrating over all possible character histories h. Finally, the assumption of sequence site independence is often integrated into the likelihood function. The likelihood expression for a simple case is presented schematically in Figure 1.

Specifically, the likelihood of the model $L=P(D_S, D_C|T,r,p,M_C,M_S)$ can be presented as the following product: $L=P(D_C|T,M_C)P(D_S|T,r,p,D_C,M_C,M_S)$. In this equation, parameters that do not affect the probability of the character data, D_C, are omitted from the expression for the conditional probability of D_C. While the computation of $L_C=P(D_C|T,M_C)$ is straightforward, the computation of $L_S=P(D_S|T,r,p,D_C,M_C,M_S)$ is challenging due to the dependence on the evolutionary history of the phenotype, which is unknown and resides implicitly in D_C and the two states Markov process M_C. Thus, we first present L_S as an integration over all possible character histories h:

$$\begin{aligned}L_S&=P\left(D_S|T,r,p,D_C,M_C,M_S\right)\\&=\int_h P\left(D_S|T,r,p,D_C,M_C,M_S,h\right)\\&\quad P\left(h|T,r,p,D_C,M_C,M_S\right)dh\end{aligned}$$

When omitting parameters that do not affect the probability of D_S from the above equation we obtain:

$$L_S=\int_h P\left(D_S|T,r,p,M_S,h\right)P\left(h|T,D_C,M_C\right)dh$$

Under the assumption of independence between sites, the probability of D_S can be expressed as a product of probabilities over all sequence sites k, where D_S^k is the sequence data in position k:

$$L_S=\int_h\left[\prod_k P\left(D_S^k|T,r,p,M_S,h\right)\right]P\left(h|T,D_C,M_C\right)dh$$

Here we follow the same importance sampling approach to estimate the integral over character histories as proposed by Mayrose and Otto (2011), by replacing the

integral with an average over N stochastic mappings, each with a probability of $1/N$:

$$L_S \approx \frac{1}{N}\sum_{i=1}^{N}\prod_{k} P\left(D_S^k|T,r,p,M_S,h_i\right)$$

Finally, we denote $P\left(D_S^k|T,r,p,M_S,h_i\right)$ as $L_S^{k,i}$ and obtain:

$$L_S \approx \frac{1}{N}\sum_{i=1}^{N}\prod_{k} L_S^{k,i}$$

The likelihood based on position k, $L_S^{k,i}$, is computed using a mixture model of the likelihoods over two scenarios: either the position evolved independently of the character state, in which case, the likelihood is $L_S^k(I)$, or the position belongs to the phenotype-dependent positions category, in which case the likelihood is $L_S^{k,i}(D)$:

$$L_S^{k,i} = L_S^{k,i}(D)p + L_S^k(I)(1-p)$$

where p is a parameter that specifies the probability of a position to belong to the phenotype-dependent category.

$L_S^k(I)$ is computed using the standard pruning algorithm (Felsenstein 1981). $L_S^{k,i}(D)$ is computed in a similar manner, however, when computing transition probabilities along each branch, we explicitly account for the character changes by using Q_{S_i} as defined above. For more details regarding the average rate matrix Q_{S_i} and our implementation, see Supplementary Material available on Dryad.

Based on the likelihood model presented above, maximum likelihood estimates for all free parameters of the model are sought using a heuristic optimization scheme (Brent 1973). To avoid local maxima, the search starts from multiple starting points (in this study, 30 starting points).

Inference of position category.—The Bayes factor (BF) for position k is denoted B_k and is the ratio of likelihood scores for each of the categories (phenotype-dependent and phenotype-independent), based on the sequence position data:

$$B_k = \frac{L_S^k(D)}{L_S^k(I)}$$

The empirical Bayes posterior for position k is denoted Π_k and is computed as follows:

$$\Pi_k = \frac{L_S^k(D)\times p}{L_S^k(D)\times p + L_S^k(I)\times(1-p)}$$

The empirical Bayes posterior reflects the extent to which position k is likely to belong to the phenotype-dependent category. The computation of Π_k is approximated by plugging in the inferred maximum likelihood estimates for each of the free parameters as well as using a set of N stochastic mappings, as described above. We denote this approximation as $\hat{\Pi}_k$ (a similar approach adopted by Nielsen and Yang 1998). Furthermore, we note that in case the proportion parameter p is estimated to be 0 or 1 the category prediction is the same for all sequence positions.

Simulated Data Sets

Different performance aspects of the TraitRateProp model were evaluated on simulated data sets. In each such simulation, we provided our simulator with parameters of the sequence model, M_S; the character model, M_C; the ultrametric tree, T (see details below); the rate parameter, r; the proportion parameter, p; and the number of the sequence positions to simulate, L. The M_S parameters are the transition/transversion κ parameter (set to be 2 in all simulations) and the rate heterogeneity across sites parameter, α (set to be 1 in all simulations). The M_C parameters included π_1 (set to be 0.5 in all simulations) and μ (set to be 10 in all simulations). Based on this basic scheme, different simulation scenarios with various combinations of the number of species, p, and r were examined.

Course of simulation.—According to the provided M_C model parameters, a history of trait changes, h, was simulated along the input ultrametric tree T. Given h, sequences were simulated according to the sequence model along T, with two classes of positions: phenotype-dependent (proportion p of the positions) and phenotype-independent (proportion $1-p$ of positions). We examined the performance of the method on trees with 16, 32, 64, or 128 species. To this end, 100 ultrametric trees for each number of taxa were generated using Mesquite (Maddison and Maddison 2015), following a birth–death process with the default birth rate of 0.3 and death rate of 0.1. The tree height of each of these trees was then adjusted to be 0.2. Each of these trees was given as the input tree T in the simulations.

Models Comparison

Two methods for model comparison were examined. The first is based on the likelihood ratio test (LRT), in which the test statistic $D = 2\left(\mathrm{LL}_{\mathrm{alt}} - \mathrm{LL}_{\mathrm{null}}\right)$ is compared with the critical value C, according to χ_2^2 distribution for 0.95 statistical confidence. All cases in which $D > C$ were counted as cases in which the alternative model was preferred over the null model. As an alternative method, parametric bootstrapping (see details in the section below) is utilized. We note that the resolution of the empirical P values computed based on a parametric bootstrap is restricted by the number of bootstrap replicates. This limited resolution poses a challenge to correct for multiple hypotheses when several loci are analyzed. Under such conditions, one can compare the empirical P values based on the parametric bootstrap procedure to those approximated by the χ^2 distribution.

If the discrepancy between the two sets is not substantial, the multiple hypotheses correction can be applied to the χ^2 derived P values.

Two-stage parametric bootstrap.—In our simulation study, we detected slight deviations in the false positive rate based on the standard LRT (see "Results" section). Furthermore, for the orchid data set, we were interested in comparing three models (see details below), which necessitate correcting for multiple dependent comparisons also when a single locus is analyzed. These reasons prompted us to analyze the orchid data sets using a parametric bootstrap approach. Given a computed $D_{\text{real}} = 2\left(\text{LL}_{\text{alt}} - \text{LL}_{\text{null}}\right)$ test statistic, parametric bootstrapping can be used to assess the probability to observe a D value greater than or equal to D_{real} under null conditions. This provides an empirical P value for the D_{real} result while controlling the false positive rate. To this end, B simulations based on the null model and its estimated parameters are generated and the D value for each of these B simulations is computed. Next, D_{real} is compared with the distribution of D values under null conditions to obtain an empirical P value.

In the case of the orchid data sets, we employed this approach twice; first to select between the null model ($p=0$, $r=1$; no association between sequence and character evolution) and two alternative models: TraitRate model (TR) ($p=1$, $r=$free; all sequence positions are in association with character changes) and TraitRateProp model (TRP) ($p=$free, $r=$free; any proportion of sequence positions could be associated with character changes). In cases where the null model was rejected both when compared with the TR as well as with the TRP model, a subsequent analysis was performed to allow the selection between the TR and TRP models. This second stage analysis consisted of comparing the difference in the fit of the TRP and TR models to the observed data to the difference in the fit of these models to data simulated under the TR model conditions.

The required simulations in either stage of the parametric bootstrap procedure were performed using the INDELible sequence simulator (Fletcher and Yang 2009), with the HKY (Hasegawa et al. 1985) model of substitution and α and κparameters as estimated from the data by the null model. In both stages, the sequences were simulated while the real character data were kept constant. The full procedure details are given below.

Procedure details

Models compared:

1. Null: $p=0, r=1$. Its model parameters are denoted θ_{null}.
2. TR: $p=1$, $r=$free. Its model parameters are denoted θ_{TR}.
3. TRP: $p=$free, $r=$free. Its model parameters are denoted θ_{TRP}.

For each gene, we denote its observed data:

- $_{DS}$: sequence dat;
- $_{DC}$: character data; and
- T: ultrametric species tree.

Stage I

Test hypotheses:

- H_{0_TR}: Null model is correct (compared to TR).
- H_{0_TRP}: Null model is correct (compared to TRP).

Parametric bootstrap procedure:

1. Infer θ_{null}, θ_{TR}, θ_{TRP} based on the observed data. Obtain the maximum log-likelihood (LL) scores of each of the models and compute the following test statistics:
   - $D_{\text{TR}} = 2\left(\text{LL}_{\text{TR}} - \text{LL}_{\text{null}}\right)$
   - $D_{\text{TRP}} = 2\left(\text{LL}_{\text{TRP}} - \text{LL}_{\text{null}}\right)$
2. Simulate $B=100$ sequence data instances based on T and θ_{null}. Provide each simulated sequence data, D_S^b together with D_C and T for inference and obtain:
   - $D_{\text{TR}}^b = 2\left(\text{LL}_{\text{TR}}^b - \text{LL}_{\text{null}}^b\right)$
   - $D_{\text{TRP}}^b = 2\left(\text{LL}_{\text{TRP}}^b - \text{LL}_{\text{null}}^b\right)$
3. Determine a cutoff C such that the total ratio of rejections of H_{0_TR} and H_{0_TRP} is kept at 5%.
4. Reject H_{0_TR} if $D_{\text{TR}} > C$, Reject H_{0_TRP} if $D_{\text{TRP}} > C$.
5. If both null hypotheses were rejected continue to stage II.

Stage II

Test hypothesis:

- H_{0_TRP}: TR model is correct (compared to TRP).

Parametric bootstrap procedure:

1. Based on the already inferred θ_{null}, θ_{TR}, θ_{TRP} and LL scores of each of the models, compute the following test statistic:
   - $S_{\text{TRP}} = D_{\text{TRP}} - D_{\text{TR}}$
2. Simulate $Y=100$ sequence data instances based on T and θ_{TR}. Provide each simulated sequence data, D_S^y together with D_C and T for inference and obtain:
   - $S_{\text{TRP}}^y = D_{\text{TRP}}^y - D_{\text{TR}}^y$
3. Determine a cutoff C such that the ratio of rejections of H_{0_TRP} is kept at 5%.
4. Reject H_{0_TRP} if $S_{\text{TRP}} > C$.

Comparison to the OM Method

We obtained the code for the OM method (O'Connor and Mundy 2009), two sequence data sets analyzed in their paper, the primate species topology, and the character states for each of the analyzed primate species from the authors. The sequence data sets were those of the *SEMG2* gene (16 primate species and an MSA of 4245 positions) and of the *ZAN* gene (16 primate species and an MSA of 555 positions). To analyze these data sets by TraitRateProp, we used MrBayes (Ronquist et al. 2012) with a birth–death relaxed clock model to compute an ultrametric tree based on each of these data sets while imposing the topological constraints as defined in the species topology provided by O'Connor and Mundy (2009). These trees together with the sequence and character data sets were then provided as input to TraitRateProp (The MSAs of the two genes, the primate species tree, the primate character states, the MrBayes configuration file, and the resulting ultrametric trees are available as Supplementary Material on Dryad).

Orchidaceae Data Analysis

Sequence data.—We extracted the 20 most commonly present coding regions of all published plastid genomes of Orchidaceae available in the NBCI Genomes database. These data were complemented with the plastid gene data sets of Givnish et al. (2015), yielding a final data set of 20 plastid genes over 85 Orchidaceae species. These data included ten additional recently sequenced species of the widely distributed Neottieae tribe (Supplementary Table S1 available on Dryad), which itself comprises a great variety of heterotrophic lifestyles from autotrophic to completely heterotrophic orchids. Where necessary, intron/exon boundaries of *rpl16*, *rpl2*, *rps12*, *rps16*, and *clpP* were corrected manually using verified coding sequences of these genes from *Nicotiana tabacum* (GenBank Accession: Z00044.2), *Arabidopsis thaliana* (NC_000932.1), and *Triticum aestivum* (KJ592713.1). The taxon sampling and accession numbers for the plastid sequences are provided in Supplementary Table S1 available on Dryad.

Species tree reconstruction.—Codon MSAs for each of the 20 plastome genes were computed using a Perl script, by first aligning the translated protein sequences using MAFFT v7.182 (with parameters: –localpair –maxiterate 1000) and then back-translating this MSA to nucleotide-based alignment (Katoh et al. 2009; Katoh and Standley 2013). Each of these MSAs contained sequences for at least 75 out of the 85 species. We next applied MrBayes (Ronquist et al. 2012) using a birth–death relaxed clock model to obtain a set of ultrametric species trees, of which the tree with the highest likelihood score was selected. The input to MrBayes was a concatenation of gene MSAs. Each gene was defined as a partition. The nucleotide substitution model for each partition was identified using jModelTest (Darriba et al. 2012), according to the Akaike information criterion score. A few topological constraints based on Givnish et al. (2015). were set; fixing the root at *Mapania palustris*, then setting *Apostasia wallichii* as a sister taxon to all other 83 species, fixing *Pogonia ophioglossoides* and *Vanilla planifolia* as a monophyletic group, and *Phragmipedium longifolium*, *Cypripedium japonicum*, *Cypripedium formosanum*, *Paphiopedilum niveum*, and *Paphiopedilum armeniacum* as another monophyletic group. Finally, we set the Vanilla monophyletic group as a sister clade to all other 81 species. The MSA files, the MrBayes configuration file as well as the resulting ultrametric tree are provided as Supplementary Material available on Dryad.

Phenotypic trait character states.—The phenotypic trait character state of heterotrophism was determined with respect to each of the 85 orchidaceae species. Heterotrophic species were coded as character state "1," whereas autotrophic species were coded as "0." The character states were determined according to the presence or absence of chlorophyll indicating a fully heterotrophic lifestyle, as well as physiological studies of the carbon flow between an adult heterotrophic (chlorophyllous) orchid and its host. The trait states and the decision criteria are provided in Supplementary Table S1 available on Dryad.

Code Implementation and Availability

The TraitRateProp software, the C++ source code, and a short manual are provided at http://www.tau.ac.il/~itaymay/cp/TraitRateProp. The input to the program is an ultrametric species tree in Newick format and the sequence and character data in Fasta format. The program outputs the maximum likelihood estimates of the parameters for the null and alternative models as well as the computed BF for each sequence position.

RESULTS

Inferring Associations Between Evolutionary Rates and Phenotypes with the TraitRateProp Model

In this work, we present the TraitRateProp model, which allows studying possible associations between a binary phenotypic trait and the rate of sequence evolution. This method extends a previous work by Mayrose and Otto (2011), which aimed at detecting whole genes (proteins) whose evolutionary rate correlates with the state of a binary phenotypic trait. In that model it is assumed that in cases where such a correlation exists, all sequence positions correlate with the state of the phenotypic trait. Here, we relax this assumption by allowing the evolutionary rate of only a proportion of positions in the analyzed genomic region to depend on the phenotypic state. TraitRateProp allows for: (i) testing whether the evolutionary rate of the input sequence data is correlated with the given trait data; (ii) in case a correlation is detected, the method infers the positions most likely associated with the trait data.

TraitRateProp is based on the maximum likelihood paradigm (see "Materials and Methods" section), and provides two maximum likelihood estimates regarding the coevolution of sequence and trait data: the relative rate parameter, r, describing the ratio between the sequence evolutionary rates under states "1" and "0," the parameter, p, which is the proportion of sequence positions whose evolutionary rate is associated with the phenotypic state. Moreover, TraitRateProp ranks sequence sites according to their likelihood of being at the trait-dependent category. The full details of the model, the likelihood estimation procedures, and the associated statistical tests are detailed in the "Materials and Methods" section. We examined the performance of TraitRateProp in simulations and we used it to detect plastid genes in Orchidaceae whose rate varies depending on the lifestyle of adult plants.

Number of Stochastic Mappings and Likelihood Computation Mode

Inference with TraitRateProp relies on an approximated likelihood computation using a set of stochastic mappings drawn based on the character model (see "Materials and Methods" section and Supplementary Materials and Methods available on Dryad for full details). The number of stochastic mappings, N, is a parameter that tunes the accuracy of this approximation. In an exhaustive computation mode, each such stochastic mapping is used to evaluate the likelihood score based on the sequence data, which is computationally demanding. Next, these likelihood scores are averaged to obtain the total likelihood score of the joined model. Notably, this exhaustive scheme entails a large number of likelihood computations, which becomes prohibitively long for a large value of N. As a heuristic alternative to the exhaustive approach, the stochastic mappings in the set are first summarized by taking the average time each branch spent in each character state. Then the likelihood score of the sequence model is computed only once, based on the average stochastic mapping. While offering a speed-up factor of up to N in running times, the LL score computed based on the average stochastic mapping may not be an adequate approximation of the average of the LL scores computed based on each single stochastic mapping in the set. Thus, we first studied the impact of N and computation mode (exhaustive/heuristic) on the accuracy and stability of the likelihood computation. We then studied their impact on the accuracy of a complete inference procedure, in terms of power and parameter estimation. Our results indicate that the heuristic mode with $N=10,000$ provides similar accuracy to the exhaustive mode with $N=200$ while reducing running times by an average factor of over 60 (full details concerning these investigations are given in Supplementary Material and figures therein available on Dryad). We thus chose to continue with the heuristic mode with $N=10,000$ from now on, as it provides sufficient accuracy alongside considerable improvements in running times.

Performance in Simulations

We studied the performance of TraitRateProp using simulations to evaluate its false positive rate (the tendency of the method to detect correlation when no such correlation exists), power (i.e., its ability to detect correlation between phenotypic evolution and sequence evolution when such a correlation exists), and accuracy of parameter estimation.

False positive rate and power analysis.—We analyzed the false positive rate by setting the proportion parameter to zero, thus simulating sequences without any dependence on the character trait. We first estimated the false positive rate when the null (and correct) model is rejected according to the LRT test statistic, D, as approximated using the χ^2 distribution (see "Materials and Methods" section). We found slight deviations of the false positive rates from the expected 5% (3%, 2%, 10%, and 9% for 16, 32, 64, and 128 species, respectively). This suggested that the χ^2 approximation is not accurate in this case, which prompted us to determine alternative cutoffs to which the test statistic D should be compared, such that the false positive rate in the null scenarios is fixed to be 5%. To this end, we set the cutoff in each of the null scenarios with 16, 32, 64, and 128 species by taking the D value of the 95th percentile in the simulated null data set. The cutoffs determined this way were 4.38, 5.2, 7.8, and 6.9 for 16, 32, 64, and 128 species, respectively (compared with 5.99 using the χ^2 approximation). These cutoffs were subsequently used for power analysis. Of note, the deviations in the χ^2 cutoffs reported here are specific to the examined simulated data sets and could differ when real data are analyzed. We thus used a parametric bootstrap procedure to analyze the orchid data set as detailed below.

We analyzed the power of TraitRateProp when simulating sequences of 1000 base pairs in length and the rate parameter r fixed to 3. This value of r is within the distribution of r values inferred from real data sets (see below). In these simulations, we varied the number of species analyzed and the proportion of positions whose rate is in association with the phenotypic trait. When the number of species analyzed was 64 or higher, and the proportion of sites affected by the phenotypic state was 50% or higher, the method correctly detected the existence of the association between sequences and phenotypic evolution in all cases (Fig. 3). As expected, when data are limited, the power decreases. For example, with a moderate number of species (32), the power is still greater than 90% when the proportion parameter is 0.5 or higher. However, when the number of species is only 16, or the proportion is 0.25, the power is substantially lower (Fig. 3). These results suggest that TraitRateProp is expected to perform well when trait and sequence data are available for at least a few dozen species.

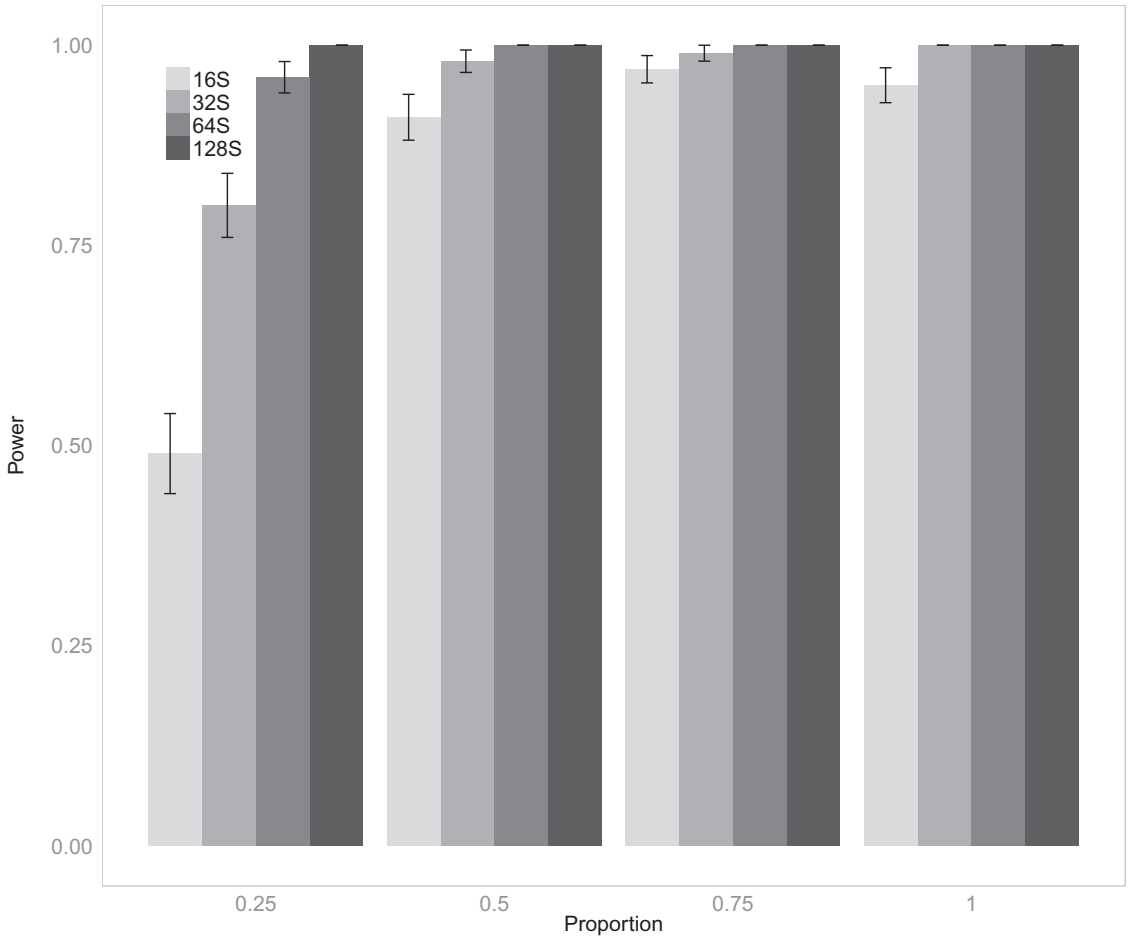

FIGURE 3. Power assessment. Results based on simulated data sets with varied proportion parameter values and number of species. For the power analysis, a cutoff for rejecting the null model for each number of species was determined by fixing the false positive rate in the null scenario ($p=0$) for that number of species at 5%. The height of the bars reflects the percentage of simulated data sets in which the null model was rejected and the whiskers show the standard error based on a binomial distribution.

In cases where the total number of positions affected by the phenotype is small, additional species should be included in the analysis.

Accuracy of parameter estimation.—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 proportion of positions whose rate is associated with the phenotypic trait (Fig. 4). In addition, we found a negative correlation between the inferred r and P values (e.g., Spearman coefficient of correlation $\rho=-0.66$, $P<10^{-10}$ for the 32 species, and $p=0.5$ data set. Similar results were observed for other data sets, data not shown). This suggests a tradeoff between the inference of the relative rate and the proportion parameters, where overestimation of one of them can cause an underestimation of the other, and vice versa. We studied this tradeoff in more detail by examining the likelihood surface as a function of p and r. We found a rather flat likelihood surface for r and p combinations whose product is close to the product of the real parameter values, $p=0.5$ and $r=3$ (Fig. 5). However, parameter combinations in which $r>1$ received a higher likelihood score than parameter combinations in which $r<1$ (all 46 top scoring points out of 100 points examined had $r>1$). This result suggests that, despite the inference tradeoff between the relative rate and proportion parameters, the method can correctly detect the direction of the phenotypic state effect.

Comparison to OM

We first used TraitRateProp to analyze the publicly available data of O'Connor and Mundy (2009). In agreement with their findings, we detected an association between a phenotype of a multimale–multifemale mating system and the rate of evolution in semenogelin II ($D=26.7$, empirical $P<0.01$, $r=4$, $p=0.42$), but no such association was detected in sperm ligand zonadhesin ($D=1.6$, empirical $P=0.24$). We next compared the performance of TraitRateProp with that of OM on all simulated data sets described in the previous section, where we varied the number of species analyzed as well as the proportion of positions whose rate is

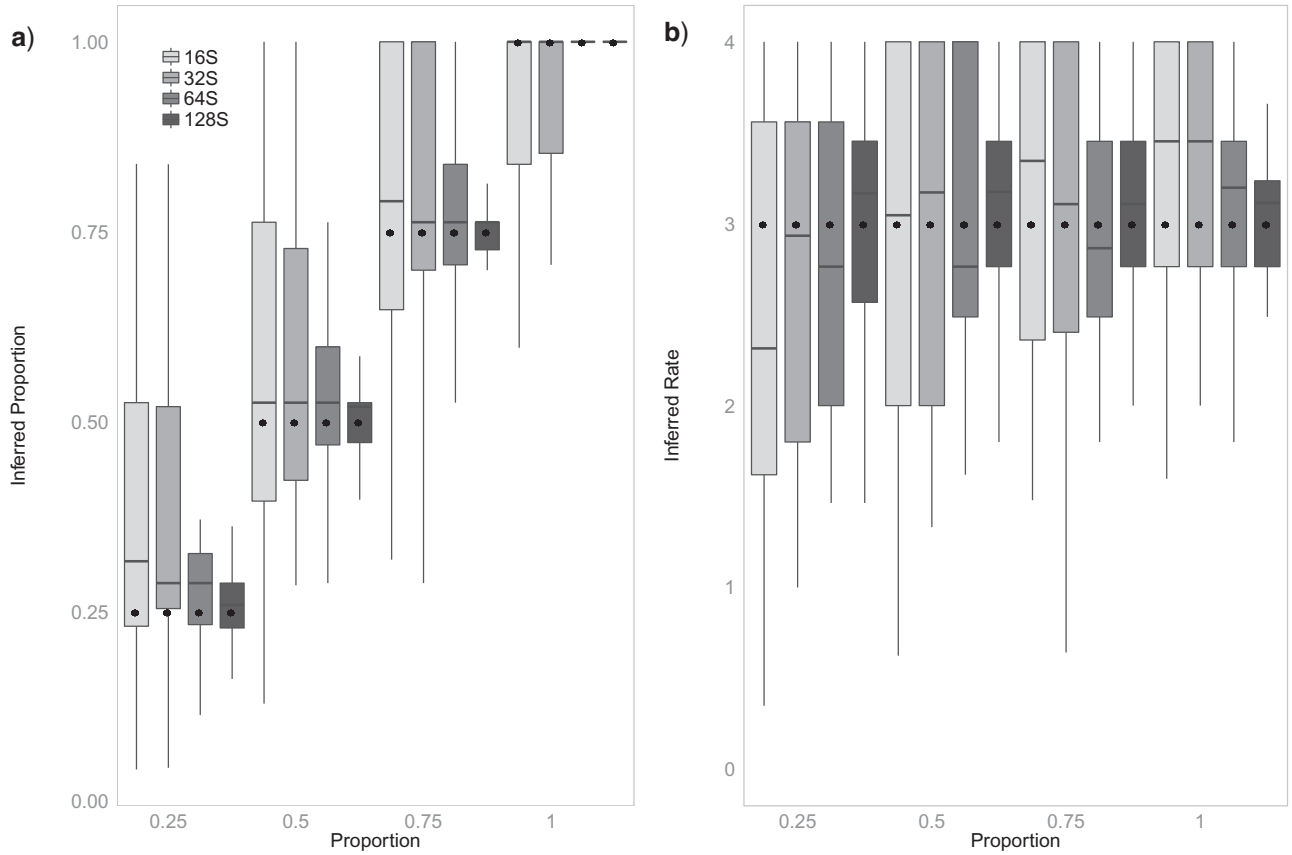

FIGURE 4. Parameter inference assessment. Results are based on simulated data sets with varied proportion parameter values and number of species. a) Inferred proportion parameter. b) Inferred relative rate parameter.

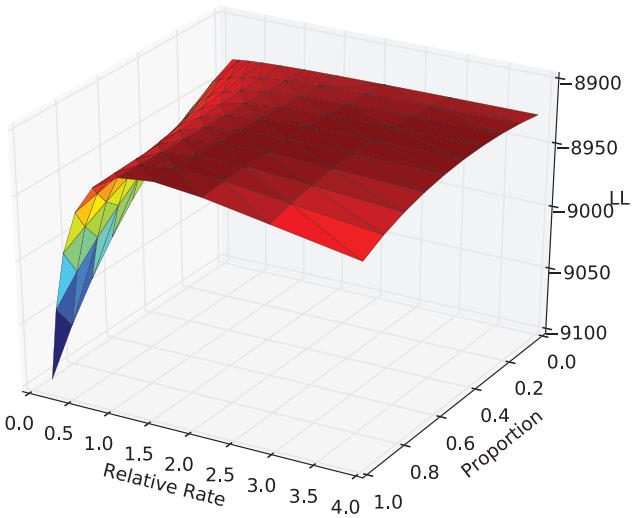

FIGURE 5. LL surface over a grid of 100 combinations of the relative rate and proportion parameters. The figure depicts a single simulation instance in which $p=0.5$, $r=3$, and the number of species was 32.

in association with the phenotypic trait. The maximal power measured for the OM method was on a data set of 64 species where 25% of the positions were in association with the phenotypic state; on this data set, the OM method had a power of 6% (in comparison, TraitRateProp had a power of 98% on this data set). In their original paper, the power of the OM method was reported for data sets in which the sequence-evolution rate ratio between phenotypic state "0" and "1" was either 10 or 1000 (O'Connor and Mundy 2009). We, therefore, analyzed additional simulated data sets where the relative rate parameter, r, was set to be either 10 or 1000 and the proportion parameter, p, was set to be 0.5. Under these scenarios, the OM method had a higher power compared to its performance on simulations with $r=3$, reaching 8.2% and 14% on the 64 and 128 species data sets for r values of 10 and 1000, respectively. In comparison, the power of TraitRateProp was 100%, on both these data sets.

Predicting Sites Whose Rate Is Associated with the Phenotypic State

Under the TraitRateProp model, each sequence position is either associated or not with the examined phenotypic state. Based on the maximum likelihood estimates of the model parameters, the likelihood of each sequence position can be estimated twice; once ascribing the position to the phenotype-dependent category and once ascribing it to the phenotype-independent category. The ratio between these likelihood scores is an approximation of the BF per position and can be used

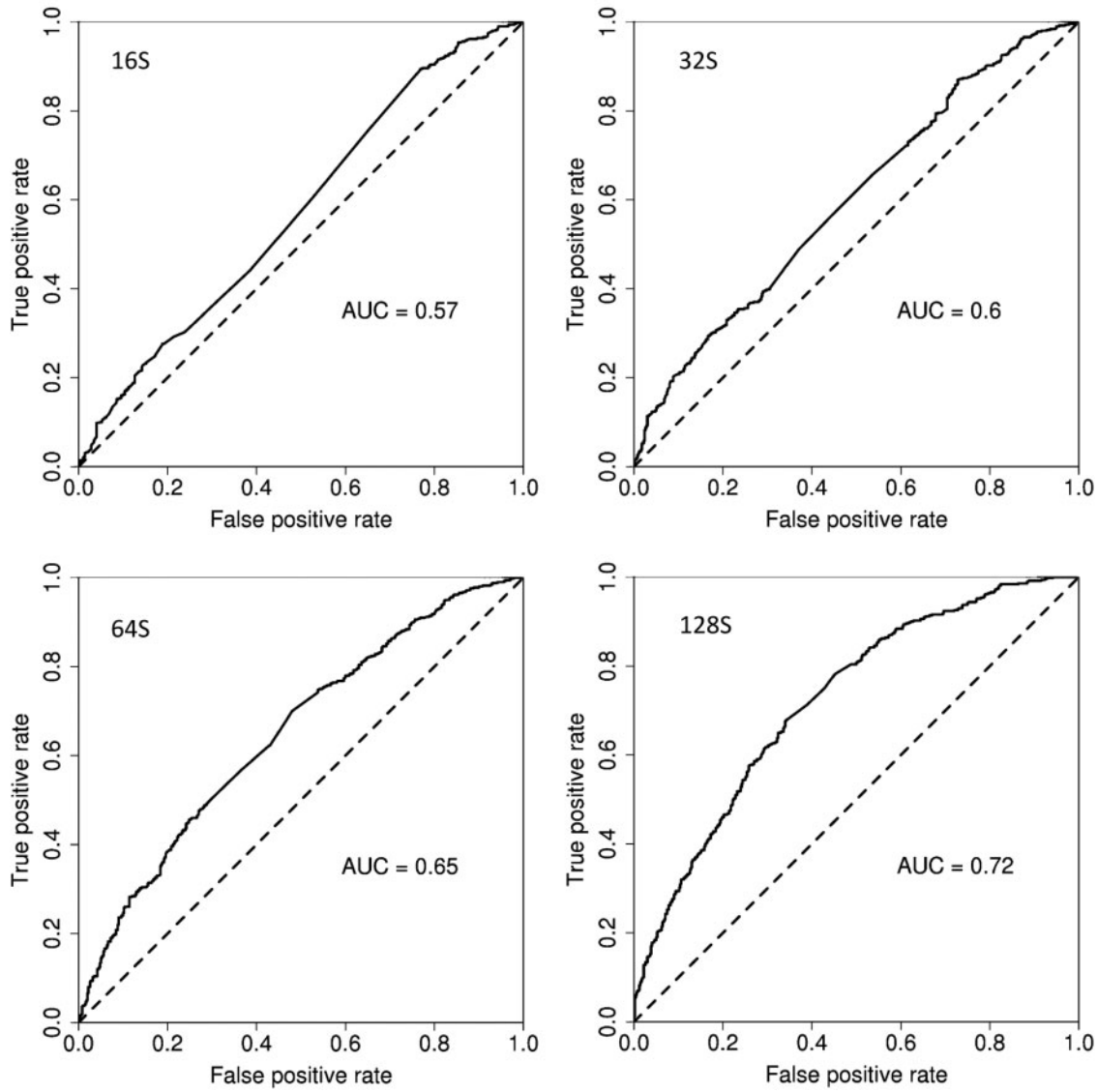

FIGURE 6. Prediction of association with the phenotypic trait per position. ROC curve of the simulation with the median AUC value for each number of species. Simulations were conducted with $p = 0.5$ and $r = 3$.

to classify positions into these two categories. To test the classification accuracy of TraitRateProp, we focused on simulated data sets with a proportion parameter of 0.5. We used the estimated BF scores to compute the area under the ROC curve (AUC-ROC) (Fawcett 2006). As expected, we found the classification accuracy to increase with the number of analyzed species; starting with a moderate median AUC-ROC value of 0.57 for 16 species and reaching a median AUC-ROC value of 0.72 in data sets with 128 species (Fig. 6).

We hypothesized that the slowly evolving positions are less informative and thus should be more challenging to classify: a position that was simulated assuming it is character-dependent but is completely invariant holds no information for classification. Thus, we examined the effect of filtering out slowly evolving positions, prior to categorizing positions. To this end, we applied rate4site (Mayrose et al. 2004) to infer the evolutionary rate of each position in the simulated MSA. The rate4site scores were used to exclude 10%, 20%, or 30% of the slowest positions from the analysis. We found that excluding slowly evolving positions resulted in a slight, yet statistically significant increase in the ability of TraitRateProp to classify the remaining sites, reaching average AUC-ROC values of 0.59 and 0.75 for 16 and 128 species, respectively, when 30% of the slowest positions are filtered prior to classification (Fig. 7). In addition, we examined the effect of filtering the fastest evolving positions as these positions have the potential to become saturated and less informative. In our simulation scheme, filtering such positions reduced the evolutionary signal, causing a reduction in the classification accuracy (Supplementary Fig. S1 available on Dryad), suggesting that in the evolutionary scenarios examined in our simulation

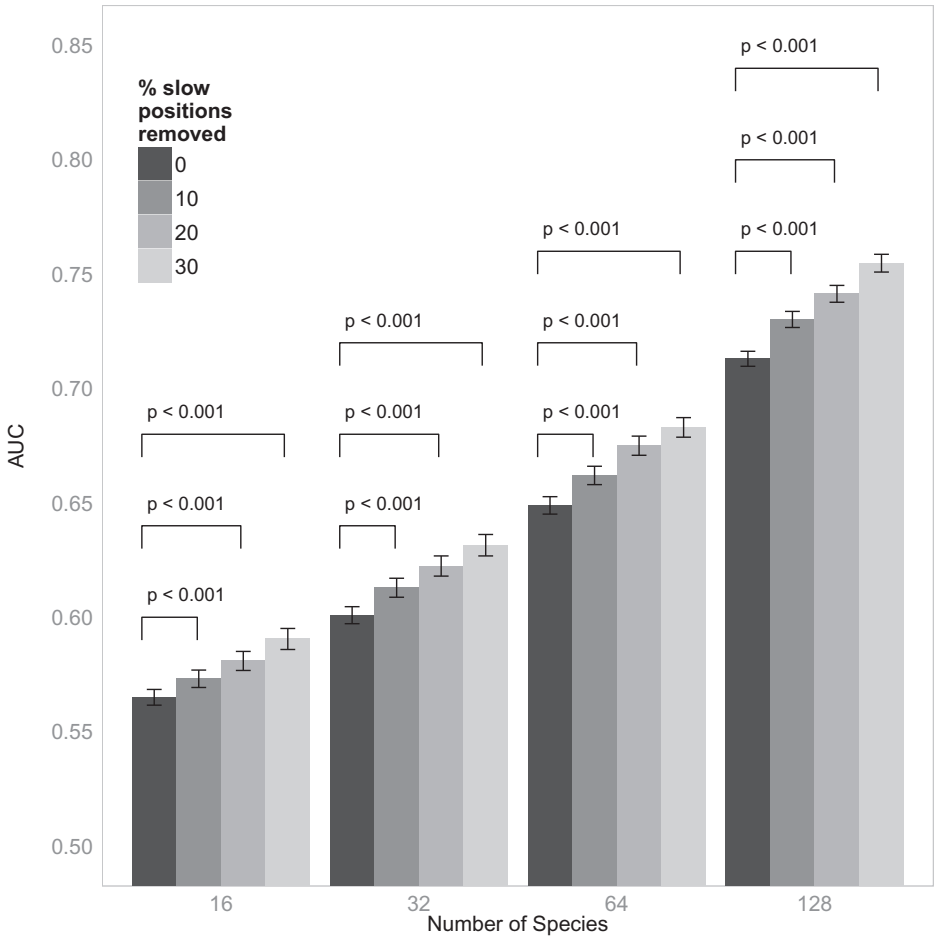

FIGURE 7. Prediction of association with the phenotypic trait per position with position filtration. Average AUC values across simulations for each number of species. The AUC values were computed after 0%, 10%, 20%, or 30% of the slowest positions as detected by rate4site were removed. The *P* values indicate a paired Mann–Whitney test (Wilcoxon test) between each filtered set and the nonfiltered one. Simulations were conducted with $p=0.5$ and $r=3$.

scheme, these fast evolving have more phylogenetic signals than noise.

Notably, the AUC-ROC value measures the discriminative power across the whole range of BFs while most often the primary interest is in the correct classification of sites with the strongest signal (highest BF values), thus focusing on the region with the lowest false positive rate. We thus examined the classification accuracy of TraitRateProp by considering positions whose BF is above a certain cutoff (10, 8, 5, and 2). We found that higher BF cutoffs yield more accurate predictions (higher true positive rate). Notably, in the smaller data sets of 16 and 32 species, less than 0.1% of positions receive a high BF value, indicating a weak discriminative signal in these cases (Table 1).

Association Between the Rate of Orchid Plastid Housekeeping Genes and Heterotrophic Lifestyle

We used TraitRateProp to investigate associations between a heterotrophic lifestyle in adult developmental stages and the rate of evolution in 20 plastid genes in the Orchidaceae plant family. Seventeen of these genes encode for subunits relevant for protein biosynthesis and three function in other, photosynthesis-unrelated pathways (*clpP*, *matK*, and *ycf2*). To this end, we reconstructed the MSA for each gene using MAFFT (Katoh et al. 2009) and then used MrBayes (Ronquist et al. 2012) to obtain an ultrametric species tree based on the concatenation of all 20 MSAs (see "Materials and Methods" section for full details). The resulting ultrametric tree (Fig. 8) as well as the character state for each orchid species (autotrophic or heterotrophic) were then provided as input to TraitRateProp together with each of the MSAs. Using TraitRateProp we fitted three models; the first model is the null model in which the proportion parameter is fixed to 0, imposing that the sequence evolution is in no association to the character evolution. In the second model, denoted "TR," we fixed the proportion parameter to 1, imposing the rate of all sequence positions to be associated with the character evolution. In the last model, denoted "TRP," the proportion parameter is free to vary, allowing any number of positions to be in association with character changes. For each gene, we selected between these models using a two-stage parametric bootstrap

TABLE 1. Classification of positions by BF cutoffs

	Percent true positives in positions above cutoff			
Number of species/BF cutoff	16	32	64	128
10	100 ($n=2$)	85.4 ($n=41$)	90.6 ($n=351$)	93.7 ($n=2149$)
8	83.3 ($n=6$)	87.6 ($n=89$)	89.8 ($n=557$)	93.1 ($n=2869$)
5	79.7 ($n=64$)	82.6 ($n=344$)	86.1 ($n=1451$)	89.7 ($n=5220$)
2	69.6 ($n=2011$)	73 ($n=4675$)	75.3 ($n=9671$)	78 ($n=17,934$)

Note: For each number of species, a total of 100,000 positions were examined.

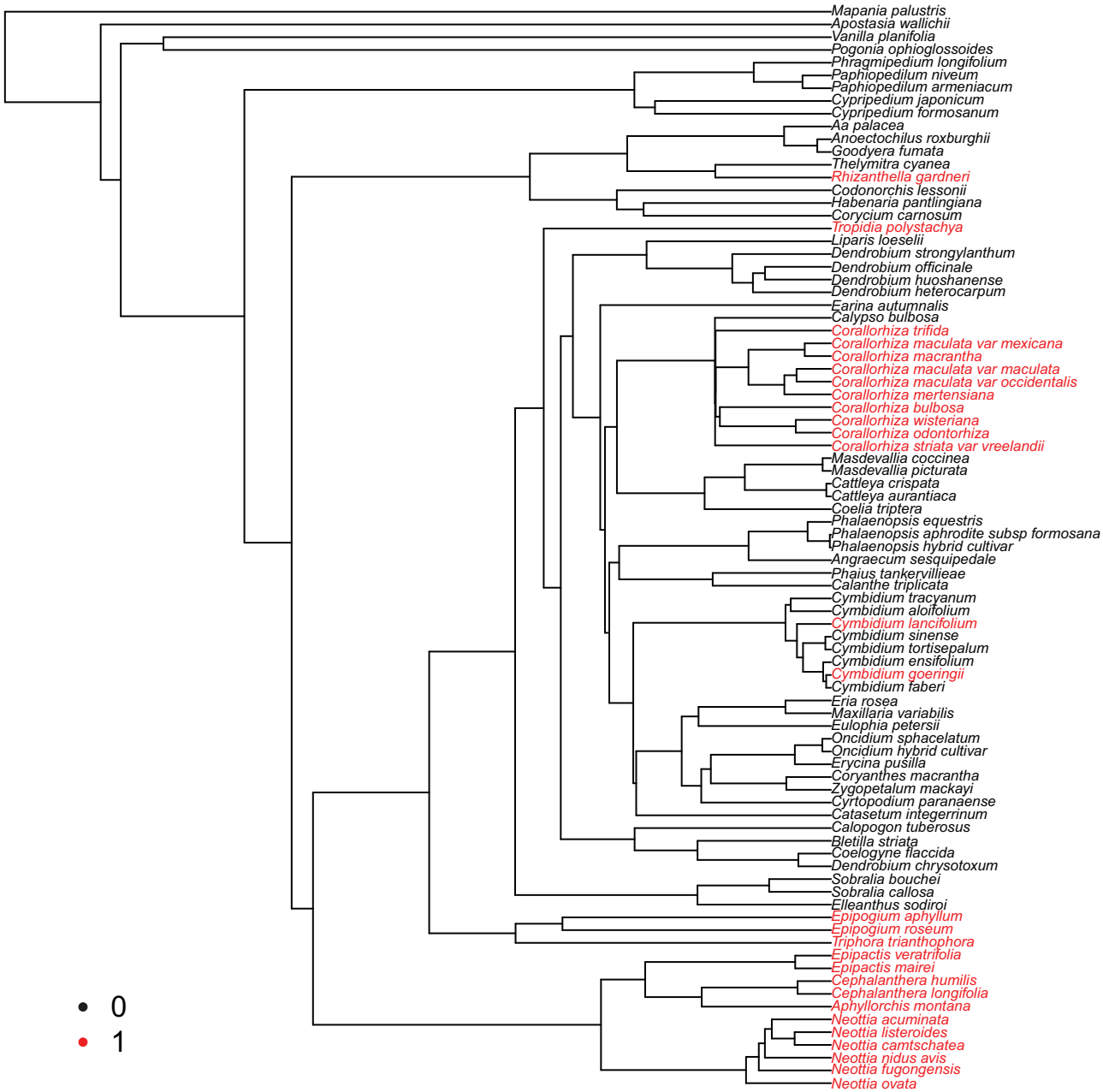

FIGURE 8. The maximum likelihood species tree for the 85 Orchidaceae species. In red are heterotrophic species (labeled as "1") and in black are autotrophic (labeled as "0").

TABLE 2. Orchid plastid genes analysis results

Gene	Selected model, two-stage BS	*p selected model*	*r selected model*	*P* value TR, BS stage I	*P* value TRP, BS stage I	*P* value TRP, BS stage II
clpP	TR	1	2.51	<0.01	<0.01	0.41
infA	TR	1	3.71	<0.01	<0.01	0.4
matK	TR	1	0.79	0.03	<0.01	0.09
rpl14	TR	1	2.67	<0.01	<0.01	0.92
rpl16	TR	1	3.12	<0.01	<0.01	1
rpl2	TRP	0.891	10	<0.01	<0.01	<0.01
rpl20	TR	1	2.08	<0.01	<0.01	0.97
rpl22	Null	NA	NA	0.5	0.53	NA
rpl23	TR	1	2.6	<0.01	<0.01	1
rpl32	Null	NA	NA	0.13	0.17	NA
rpl33	TRP	0.3	3.22	0.01	0.01	0.04
rpl36	TR	1	2.71	<0.01	<0.01	0.57
rps11	TR	1	2.66	<0.01	<0.01	0.23
rps12	TR	1	4.44	<0.01	<0.01	0.98
rps14	TR	1	3.44	<0.01	<0.01	0.09
rps15	TR	1	1.64	<0.01	<0.01	0.09
rps16	Null	NA	NA	0.13	0.08	NA
rps18	TR	1	3.52	<0.01	<0.01	0.19
rps19	TR	1	5.44	<0.01	<0.01	0.13
ycf2	TRP	0.63	3.2	<0.01	<0.01	<0.01

NA, not applicable.

approach (see "Materials and Methods" section). For 17 out of the 20 analyzed genes, we could reject the null model, preferring either the TR or TRP model. Notably, in all genes except for *matK*, the inferred trait-rate association pointed in the same direction, showing an inferred relative rate parameter greater than 1.0, that is, indicating a higher evolutionary rate associated with a heterotrophic lifestyle (Table 2). Furthermore, for three genes, *rpl2*, *rpl33*, and *ycf2*, the TRP model was preferred over the TR model, indicating that in these genes only some of the sequence sites are associated with the phenotypic trait.

DISCUSSION

In this study, we presented TraitRateProp, a likelihood framework for the joint analysis of trait and sequence data that enables the detection of specific sites exhibiting a rate shift upon repeated character trait transitions. Using a simulation study, we showed that the power and parameter estimation accuracy of TraitRateProp increase with the number of species being analyzed and with the proportion of sequence positions in association with changes in the character state. This comes as no surprise as both these factors contribute to the strength of the association signal. Based on our simulation study, we conclude that TraitRateProp is most suitable for analyzing data sets composed of a few dozen species or more. It may also be suitable for the analysis of smaller data sets if there is a good reason to expect a large proportion of sequence positions to be associated with the character trait. We then studied the ability of TraitRateProp to classify sequence positions as either trait-dependent or trait-independent, according to their estimated BF. Examining all sequence sites, classification accuracy is rather limited; particularly for data sets with a small number of species. However, when focusing on positions with the most discriminative power (i.e., those with high BF values), we found high true positive rates ($\geqslant$80% for positions with BF $\geqslant$5; Table 1).

We next used TraitRateProp to analyze 20 plastid housekeeping genes across 85 species from the Orchidaceae plant family. For 16 genes, we detected that the transition to a heterotrophic lifestyle is correlated with a higher rate of molecular evolution. Out of the 20 genes analyzed in this study, three genes—*rps16*, *rpl22*, and *rpl32*—were not detected to be associated with a heterotrophic lifestyle. Of these, *rpl22* and *rpl32* were previously shown to be essential plastid genes, which are required even under heterotrophic conditions in model plants (Fleischmann et al. 2011), which might explain why no difference in rates between the autotrophic and heterotrophic lifestyles was detected.

Recently, Maddison and FitzJohn (2015) discussed possible pitfalls in detecting phylogenetic associations. Their main concerns relate to determining causality and possible caveats stemming from the distribution of the characters along the phylogeny (i.e., few evolutionary events giving rise to numerous modern taxa). Particularly, because the statistical support for correlated evolution is drawn from the total amount of time being in each state, regardless of the number of independent trait transitions, phylogenetic methods for the detection of coevolution are susceptible to infer significant association even when a single trait transition has occurred. These concerns are of relevance to TraitRateProp, as much as they are to the methods discussed by Maddison and FitzJohn (2015). First, it is important to note that any detected association does not

imply causality. Second, we recommend examining the distribution of the character states along the tree, prior to the analysis with TraitRateProp to reveal potentially problematic scenarios, such as the "Darwin's scenario" and the "Unreplicated burst," as discussed by Maddison and FitzJohn (2015). Specifically, the orchid phylogeny analyzed in this article does not display such patterns (Figure 8).

Here we concentrated on modeling shifts in the rate of sequence evolution upon character transition while assuming that other aspects of the substitution process are unaffected. However, additional trait effects on the evolutionary process can be incorporated within the TraitRate framework. For example, Halpern and Bruno (1998) proposed a model that accounts for a varied selective pressure across sites by incorporating site-specific substitution matrices in which a mutation fixation factor is integrated. However, their model assumes that at a given position the selection pressure is constant across all parts of the phylogeny. Using the TraitRateProp framework, this model can be extended to associate two rate matrices with each site, alternating their usage according to the character state. Similarly, aspects concerning codon bias (Palidwor et al. 2010) and varying stationary base frequencies (as discussed, e.g., in Gojobori 1983) can be integrated in an analogous manner.

The current TraitRateProp model assigns a single rate parameter to each character state (r_0 and r_1). However, even if the association between evolutionary rates and the analyzed trait is factual, it is rather unlikely that a single rate typifies a certain character state. Alternatively, the modeling of the effect the trait has on the evolutionary rate can be refined by considering a prior distribution of rate values dictated by each character state. To this end, defining two separate gamma distributions, one for each state is a possible way to generate a distribution of rate values under each state while avoiding over-parametrization. To this end, two alternative implementations are possible. First, we can assume two distinct gamma distributions, each governed by a single parameter, α (i.e., forcing the shape and scale parameters of the gamma distribution to be equal such that the expectation is 1.0). This could imply a change in the selective pattern across sequence sites. For example, character state 0 could be characterized by a gamma distribution with a low α value (high heterogeneity among sites), whereas character state 1 with a high α value (low heterogeneity among sites). Notably, since the expectations of both distributions are equal to 1.0, this implementation implies that the overall rate of substitution is similar for the two character states. Alternatively, two distinct gamma distributions, without imposing that the shape and scale parameters be equal, additionally allow for a shift in the overall rate of change.

Currently, TraitRateProp focuses on phenotypic traits with two states. It is often the case that more than two categories exist for a certain phenotype. Extending TraitRateProp to handle such data could broaden the spectrum of analyses possible with TraitRateProp. In this regard, we note that TraitRateProp is not limited to phenotypic attributes but can also be used to detect rate shifts that are associated with genomic attributes. For example, using genomic scans we can identify shifts in the rate of sequence evolution of a particular gene that are associated with the presence or absence of a certain gene family.

Finally, branch-site codon models (Yang and Nielsen 2002; Zhang et al. 2005) allow the detection of changes in selective pressure along a particular set of branches. By explicitly distinguishing between synonymous and nonsynonymous substitutions, the occurrence of positive, diversifying, selection along a specific set of branches can be detected. Although the use of branch-site methods is often inspired by lineage-speci?c attributes, the categorization of branches prior to analysis requires knowledge that is not always available (Lu and Guindon 2014), and ignores uncertainties concerning the different possible pathways by which the character state proceeds. It is, therefore, interesting to harness the advantages of codon models into the trait-rate framework, by a separate examination of the association of either synonymous or nonsynonymous substitutions with the phenotypic trait. This would allow distinguishing between the release of functional constraints and adaptive evolution while accounting for the phenotypic trait evolution.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.d4j55.2.

Funding

This study was supported by ISF grants [802/16 to T.P., 1265/12 to I.M.] and by a BSF grant 2013286 to I.M. E.L.K. is a fellow of the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. A travel fund to E.L.K. was granted by the Manna Institute for Plant BioSciences.

Acknowledgments

We wish to thank Saharon Rosset, Shiran Abadi, Haim Ashkenazy, and Dafna Shkedy for useful advice and comments. We would also like to thank Luke Harmon, Alex Pyron, and Nicolas Lartillot for reviewing the manuscript and providing constructive feedback and advice.

References

Blackmon H., Demuth J.P. 2015. Coleoptera karyotype database. Coleopt Bull 69:174–175.

Brent R.P. 1973. Algorithms for minimization without derivatives: Courier Corporation.

Bromham L., Cowman P.F., Lanfear R. 2013. Parasitic plants have increased rates of molecular evolution across all three genomes. BMC Evol. Biol. 13:126. http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-13-126.

Chandra Sanyal S., Liljas A. 2000. The end of the beginning: structural studies of ribosomal proteins. Curr. Opin. Struct. Biol. 10:633–636.
Chase M.W. 2001. The origin and biogeography of Orchidaceae. In: A.M. Pridgeon, P.J. Cribb, M.W. Chase, F. Rasmussen, editors. Orchidoideae (Part 1). New York, US: Oxford University Press. p. 1–5.
Cusimano N., Wicke S. 2016. Massive intracellular gene transfer during plastid genome reduction in nongreen Orobanchaceae. New Phytol. 210:680–693.
Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9:772.
dePamphilis C. 1995. Genes and genomes. Parasitic plants. p. 177–205.
Evans T., Wade C.M., Chapman F.A., Johnson A.D., Loose M. 2014. Acquisition of germ plasm accelerates vertebrate evolution. Science 344:200–203.
Fawcett T. 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27:861–874.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.
Fleischmann T.T., Scharff L.B., Alkatib S., Hasdorf S., Schottler M.A., Bock R. 2011. Nonessential plastid-encoded ribosomal proteins in tobacco: a developmental role for plastid translation and implications for reductive genome evolution. Plant Cell 23: 3137–3155.
Fletcher W., Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. Mol. Biol. Evol. 26:1879–1888.
Gillooly J.F., Allen A.P., West G.B., Brown J.H. 2005. The rate of DNA evolution: effects of body size and temperature on the molecular clock. Proc. Natl Acad. Sci. USA 102:140–145.
Givnish T.J., Spalink D., Ames M. et al. 2015. Orchid phylogenomics and multiple drivers of their extraordinary diversification. Proc. Biol. Sci. 282. http://rspb.royalsocietypublishing.org/content/royprsb/282/1814/20151553.full.pdf.
Gojobori T. 1983. Codon substitution in evolution and the "saturation" of synonymous changes. Genetics 105:1011–1027.
Guindon S., Rodrigo A.G., Dyer K.A., Huelsenbeck J.P. 2004. Modeling the site-specific variation of selection patterns along lineages. Proc. Natl Acad. Sci. USA 101:12957–12962.
Halpern A.L., Bruno W.J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. 15:910–917.
Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.
Katoh K., Asimenos G., Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. Methods Mol. Biol. 537:39–64.
Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.
Kosakovsky Pond S.L., Murrell B., Fourment M., Frost S.D., 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.
Lehtonen J., Lanfear R. 2014. Generation time, life history and the substitution rate of neutral mutations. Biol. Lett. 10:20140801. http://rsbl.royalsocietypublishing.org/content/roybiolett/10/11/20140801.full.pdf.
Li W.H., Ellsworth D.L., Krushkal J., Chang B.H., Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol. Phyl. Evol. 5:182–187.
Logacheva M.D., Schelkunov M.I., Penin A.A. 2011. Sequencing and analysis of plastid genome in mycoheterotrophic orchid *Neottia nidus-avis*. Genome Biol. Evol. 3:1296–1303.
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.
Maddison W.P., Maddison D.R. 2015. Mesquite: a modular system for evolutionary analysis. Version 3.02 http://mesquiteproject.org.
Maddison W.P., FitzJohn R.G. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. Syst. Biol. 64:127–136.
Martin A.P. 1995. Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. Mol. Biol. Evol. 12:1124–1131.
Martin A.P., Palumbi S.R. 1993. Body size, metabolic rate, generation time, and the molecular clock. Proc. Natl Acad. Sci. USA 90: 4087–4091.
Mayrose I., Graur D., Ben-Tal N., Pupko T. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol. Biol. Evol. 21:1781–1791.
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.
McCormick M., Whigham D., O'Neill J. 2004. Mycorrhizal diversity in photosynthetic terrestrial orchids. New Phytol. 163: 425–438.
McCormick M.K., Lee Taylor D., Juhaszova K., Burnett R.K. Jr., Whigham D.F., O'Neill J.P. 2012. Limitations on orchid recruitment: not a simple picture. Mol. Ecol. 21:1511–1523.
Nielsen R., Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.
O'Connor T.D., Mundy N.I. 2009. Genotype-phenotype associations: substitution models to detect evolutionary associations between phenotypic variables and genotypic evolutionary rate. Bioinformatics 25:i94–i100.
O'Connor T.D., Mundy N.I. 2013. Evolutionary modeling of genotype-phenotype associations, and application to primate coding and non-coding mtDNA rate variation. Evol. Bioinform. Online 9: 301–316.
Pagel M. 1994. Detecting correlated evolution on phylogenies - a general method for the comparative analysis of discrete characters. Proc. R. Soc. B 255:37–45.
Palidwor G.A., Perkins T.J., Xia X. 2010. A general model of codon bias due to GC mutational bias. PLoS One 5:e13431.
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., Corrigan R.J. Jr. 2014. The encyclopedia of life v2: providing global access to knowledge about life on earth. Biodivers. Data J. e1079. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4031434/.
Pupko T., Galtier N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc. Biol. Sci. 269:1313–1316.
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.
Rodrigue N., Philippe H., Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc. Natl Acad. Sci. USA 107:4629–4634.
Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Hohna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61: 539–542.
Schelkunov M.I., Shtratnikova V.Y., Nuraliev M.S., Selosse M.A., Penin A.A., Logacheva M.D. 2015. Exploring the limits for reduction of plastid genomes: a case study of the mycoheterotrophic orchids *Epipogium aphyllum* and *Epipogium roseum*. Genome Biol. Evol. 7:1179–1191.
Smith S.A., Donoghue M.J. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322:86–89.
Tree of Sex Consortium. 2014. Tree of Sex: a database of sexual systems. Sci. Data 1:140015. https://www.ncbi.nlm.nih.gov/pubmed/25977773.
Uzzell T., Corbin K.W. 1971. Fitting discrete probability distributions to evolutionary events. Science 172:1089–1096.
Wakeley J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. 37:613–623.

Wicke S. 2013. Genomic evolution in Orobanchaceae. In: Joel D., Gressel J., Musselman L., editors. Parasitic Orobanchaceae. Berlin Heidelberg, Germany: Springer. p. 267–286. http://link.springer.com/chapter/10.1007%2F978-3-642-38146-1_15.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang Z, Nielsen 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.