W260–W264 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 27 April 2017*
doi: 10.1093/nar/gkx288

TraitRateProp: a web server for the detection of trait-dependent evolutionary rate shifts in sequence sites

Eli Levy Karin[1,2,†], Haim Ashkenazy[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 Molecular Biology and 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

Received February 28, 2017; Revised April 02, 2017; Editorial Decision April 10, 2017; Accepted April 26, 2017

ABSTRACT

Understanding species adaptation at the molecular level has been a central goal of evolutionary biology and genomics research. This important task becomes increasingly relevant with the constant rise in both genotypic and phenotypic data availabilities. The TraitRateProp web server offers a unique perspective into this task by allowing the detection of associations between sequence evolution rate and whole-organism phenotypes. By analyzing sequences and phenotypes of extant species in the context of their phylogeny, it identifies sequence sites in a gene/protein whose evolutionary rate is associated with shifts in the phenotype. To this end, it considers alternative histories of whole-organism phenotypic changes, which result in the extant phenotypic states. Its joint likelihood framework that combines models of sequence and phenotype evolution allows testing whether an association between these processes exists. In addition to predicting sequence sites most likely to be associated with the phenotypic trait, the server can optionally integrate structural 3D information. This integration allows a visual detection of trait-associated sequence sites that are juxtapose in 3D space, thereby suggesting a common functional role. We used TraitRateProp to study the shifts in sequence evolution rate of the RPS8 protein upon transitions into heterotrophy in Orchidaceae. TraitRateProp is available at http://traitrate.tau.ac.il/prop.

INTRODUCTION

Linking genomic regions with the variability of phenotypes seen in nature and deciphering the evolutionary processes responsible for this diversification has been the goal of numerous studies (e.g. 1–5). The constant rise in publicly available high-throughput sequencing data, as well as phenotypic trait data that span hundreds-of-thousands of species (e.g. 6–9) make it now more possible than ever to analyze the process of sequence evolution in light of whole-organism phenotypic changes. Connecting between these evolutionary processes bears the potential of pointing at functional roles played by specific genes (and sites within them) in shaping a phenotypic trait of interest.

We have previously developed a joint phenotype–genotype likelihood framework that enables the detection of genes or proteins whose rate of sequence evolution is in association with shifts in a binary phenotypic trait (10–11). Once an association is detected, specific sequence sites whose evolutionary rate most likely co-evolves with the phenotypic trait are inferred. This allows capturing scenarios in which not all sequence sites experience the same selective pressure due to, for example, differences in their functional role. Notably, our methodology, TraitRateProp, infers an association between the rate of sequence evolution with binary discrete traits (for continuous traits, see 12). These discrete traits may be genomic attributes, such as the presence/absence of a certain gene family as well as organismal traits, such as morphological/physiological/reproductive features or environmental preferences.

Here, we present the TraitRateProp web server, which provides statistical means to infer the strength of association between the rate of sequence evolution in a given genomic region and a phenotype of interest. Specifically, the web server allows: (i) testing the hypothesis of an association between the sequence evolutionary rate and

*To whom correspondence should be addressed. Tel: +972 3 640 7212; Fax: +972 3 640 9380; Email: itaymay@post.tau.ac.il
†These authors contributed equally to this work work as first authors.

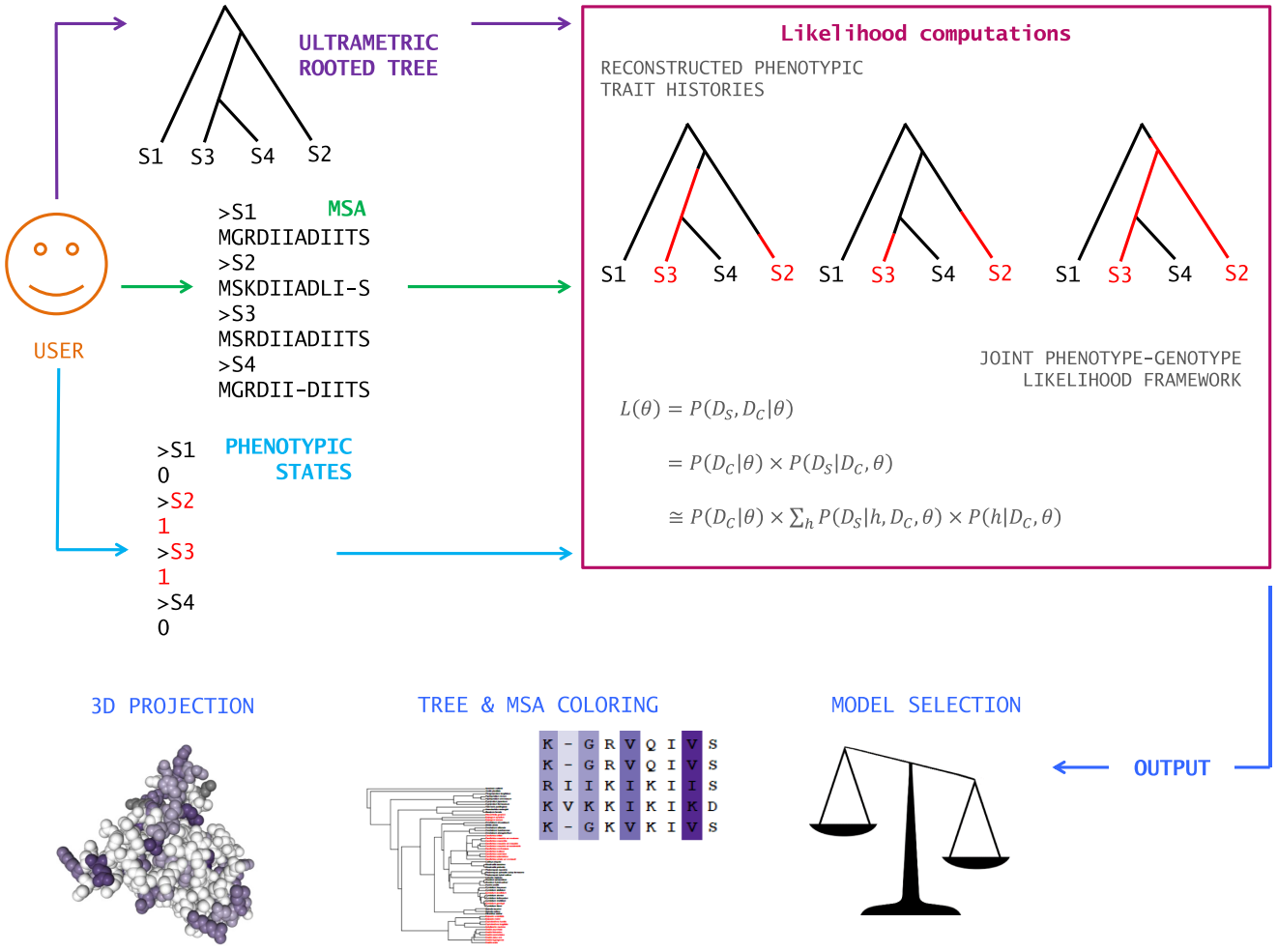

Figure 1. An illustration of the computational stages performed by the TraitRateProp web server. The user-provided input includes an ultrametric rooted tree, a multiple sequence alignment (MSA) and phenotypic states of the extant species (phenotypic states are visualized in red/black). The TraitRateProp algorithm analyzes the MSA data (D_S) and the phenotypic state data (D_C) in a joint likelihood framework. It does so by estimating the parameters (θ) of two models; a null model, which imposes no association between the rate of sequence evolution and the phenotype, and an alternative model, which allows such an association through the reconstruction of possible phenotypic trait histories (h) (transitions along the tree are visualized as a red/black color change).

the phenotypic trait; (ii) when an association is detected, TraitRateProp computes per-site predictions, where high scores indicate sequence sites whose rate is more likely to be in association with the phenotypic trait. These scores are graphically projected onto the analyzed sequence data; (iii) in case a 3D structure information of the protein is provided, the site-specific scores are mapped onto this structure, thus allowing the detection of putatively functional sequence sites that are spatially close to each other in 3D space.

We demonstrate the utility of the TraitRateProp web server by analyzing the chloroplast protein RPS8, across 60 orchid species that are either photoautotrophic or heterotrophic. The results of the TraitRateProp analysis are projected onto the recently published 3D structure of the protein (13) revealing sites whose rate of sequence evolution is predicted to be in association with the phenotypic trait and are in direct contact with other parts of the ribosomal complex.

MATERIALS AND METHODS

Input

The TraitRateProp web server requires three types of inputs: (i) a rooted ultrametric phylogenetic tree with informative branch lengths, describing the evolutionary relationships among the analyzed species; (ii) a multiple sequence alignment (MSA) of the examined sequence data (either nucleotide or protein) and (iii) the phenotypic states of the extant species coded as either '0' or '1'. In addition, the user can provide a 3D structural model in the form of a Protein DataBank (PDB) file. In this case, the site-specific predictions of TraitRateProp are projected onto the provided 3D protein structure. An illustration of the computational

stages performed by the TraitRateProp web server is presented in Figure 1.

Algorithm

The TraitRateProp web server implements the methods to infer phenotype–genotype associations described by Mayrose and Otto (11) and Levy Karin *et al*. (10). More details about these methods can be found in the Overview section of the web server and in the references therein. Briefly, these methods combine models of sequence evolution and phenotypic trait evolution into one likelihood framework by first reconstructing a large number of possible evolutionary histories of the phenotypic trait along the phylogeny. Each such history is inferred using the stochastic mapping approach (14) and is consistent with the observed phenotypic state values of the extant species. Two models are examined by TraitRateProp: a null model, which imposes no association between the rate of sequence evolution and the phenotype, and an alternative model in which such an association is allowed. A likelihood ratio test is then conducted to select between the null and alternative models. For the alternative model, TraitRateProp computes the site-specific Bayes factor and the posterior probability of an association between the rate of sequence evolution and the examined phenotypic trait. The performance of the TraitRateProp method was extensively assessed in a simulation study, showing high power and sensitivity for datasets in which 20 species or more are analyzed (10).

Output

The likelihood framework of TraiRateProp is computationally demanding, particularly when large inputs (in terms of sequence length and the number of taxa) are analyzed, such that the results are not instantly presented. Thus, once a process is submitted, an estimate of the running time (pre-calculated through simulations) is provided to the user. Upon completion of the TraitRateProp computations, the results of the likelihood ratio test are presented, indicating whether the alternative model can be preferred over the null model. In case an association between the rate of sequence evolution and the phenotype is detected, the ratio between the rates of sequence evolution in phenotypic state '1' and in phenotypic state '0' (relative rate parameter) and the proportion of sequence sites in association with the phenotypic trait are reported. Furthermore, the user can view the input MSA where each column is colored according to the TraitRateProp per-site score, indicating how likely its rate of evolution is in association with the phenotypic trait. In addition, the input phylogeny is presented graphically, with each species colored according to its phenotypic trait state. Finally, in case a protein structure was provided, the TraitRateProp per-site scores are projected onto it using the RCSB-PDB NGL viewer (15), for a visual inspection of how these sites relate to each other and to other functional sites in 3D space. All outputs of the TraitRateProp web server concerning the inferred association are offered as downloadable files.

Implementation

The TraitRateProp web server runs on a Linux cluster of 2.6 GHz AMD Opteron processors, equipped with 4GB RAM per quad-core node. The TraitRateProp algorithm was implemented in C++. The source code, a pre-compiled version for UNIX/Linux systems, and a short user manual are provided in the Source Code section of the web server.

Case study

With thousands of species, the Orchidaceae is one of the largest families of flowering plants. While most orchid species are autotrophs and rely on photosynthesis for carbon fixation, some orchid lineages exhibit a heterotrophic lifestyle, parasitizing on fungi for nutrients (16–19). The photosynthetic reaction is a result of a joint operation of different proteins, some of which are encoded in the chloroplast genome. It is thus interesting to examine the rate of sequence evolution of chloroplast proteins, taking into account trait shifts between autotrophic and heterotrophic lifestyles. In this system, it can be hypothesized that elevated rates of sequence evolution in the whole gene or in certain sequence sites are indicative of a relaxation of the functional constraints that occurred with the transition to a nonphotosynthetic lifestyle.

Here, we focused on the 30S ribosomal protein S8 (RPS8) that is encoded in the chloroplast. Directly binding to the 16S rRNA central domain, RPS8 is one of the rRNA binding proteins comprising the small ribosomal subunit (13), which together with the large ribosomal subunit gives rise to the unique translation machinery of the chloroplast. As opposed to other plastid-encoded genes, RPS8 has been previously reported to be retained in heterotrophic orchids (19,20). We were thus interested in examining RPS8 in a finer resolution, searching for specific sites within it, which experience relaxed selection upon repeated transitions to heterotrophy. To this end, we queried GenBank (21) to collect RPS8 orthologous protein sequences across 60 orchid species (accessions are available from the Gallery section of the web server) and used MAFFT v7.182 (22) to compute their MSA. Out of the 60 examined orchid species, 26 are heterotrophs and 34 are autotrophs (10). The MSA and phenotypic trait states together with the orchid ultrametric species tree (10) were given as input to the TraitRateProp web server. In order to augment the TraitRateProp analysis with structural information, we sought a structural model for the orchid RPS8 using blastp (23). To this end we used all unaligned orchid RPS8 sequences as queries against the PDB database (24). For all orchid RPS8 queries, the best blast hit was PDB ID 5MMJ, chain 'h'. This PDB ID is an atomic structure of the 70S ribosomal complex, including RPS8, in chloroplasts isolated from *Spinacia oleracea* leaves (13). As a representative sequence to the TraitRateProp web server, we chose *Dendrobium chrysotoxum*, which had the highest sequence identity to 5MMJ chain 'h' (80.60%).

Using the TraitRateProp web server, we identified an association between the rate of protein sequence evolution in RPS8 and transitions between an autotrophic and heterotrophic lifestyle (P-value < 0.00001, chi-squared likelihood ratio test). Specifically, 70% of sequence sites were

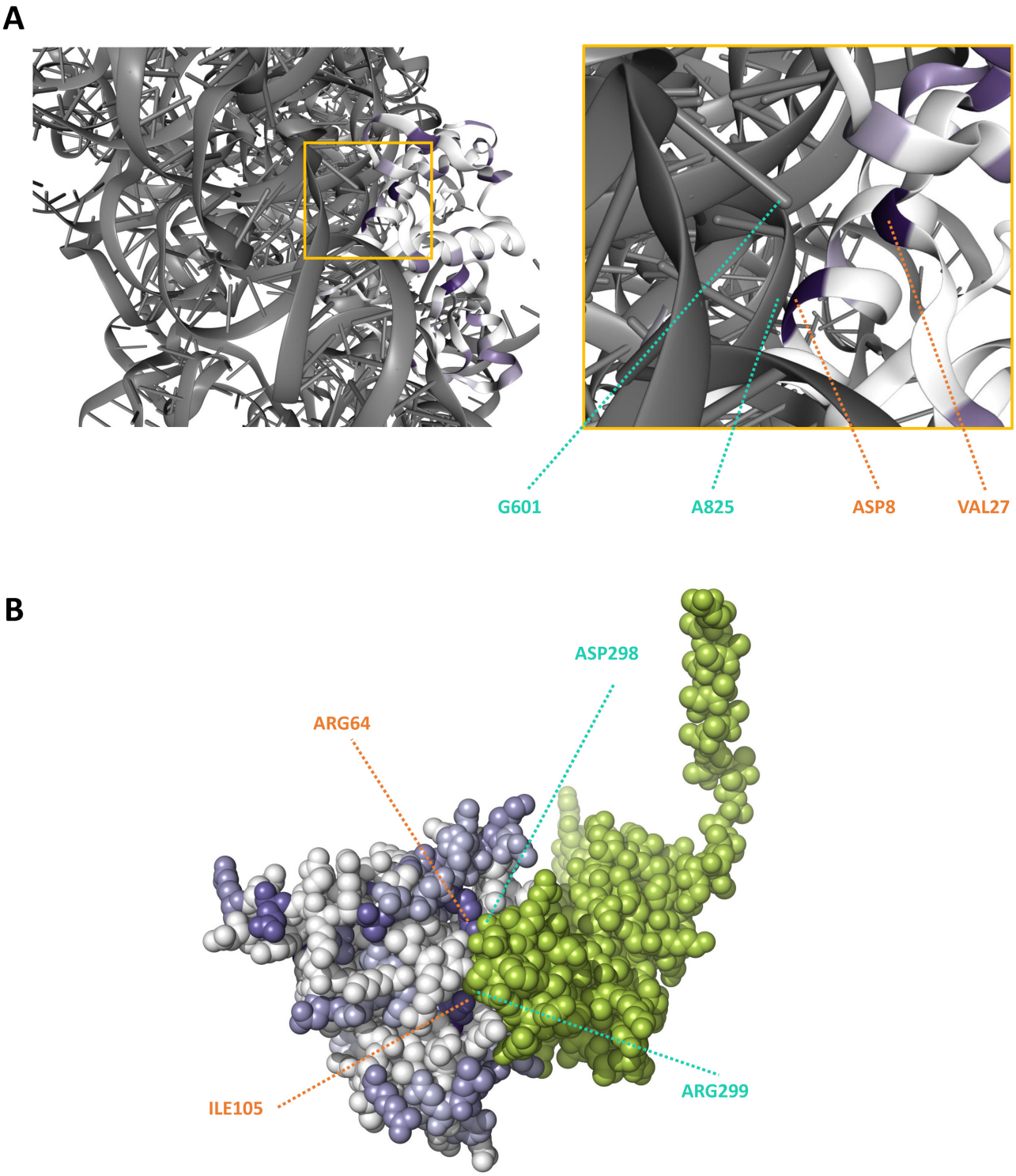

Figure 2. TraitRateProp analysis of the RPS8 chloroplast protein across 60 orchid species. Sequence site prediction for an association with shifts between autotrophic and heterotrophic trait states is indicated by white-to-purple coloring scale, where a darker shade reflects a stronger association. (**A**) The TraitRateProp per-site predictions are projected onto the 3D structure of the protein. High scoring residues ASP8 and VAL27 are in close proximity (<6 Å) to nucleotides A825 and G601 of the 16S rRNA (gray). The right-hand side is a zoom-in of the interaction interface between RPS8 and the 16S rRNA, whose overview is shown to the left. (**B**) Residues ARG64 and ILE105 are close to residues ASP298 and ARG299 of the protein RPS5 (green), which is part of the 30S ribosomal complex.

detected to be trait-associated with a relative rate parameter of 7.6, indicating a higher evolutionary rate associated with the heterotrophic lifestyle. Next, we examined the projection of the TraitRateProp per-site predictions onto the protein 3D structure. We identified four residues (ASP8, VAL27, ARG64 and ILE105) with very high TraitRateProp site scores. Interestingly, ASP8 and VAL27 are in close proximity (< 6 Å) to nucleotides A825 and G601 of the 16S rRNA (Figure 2A), while ARG64 and ILE105, which are also clustered together in 3D space, are in close proximity to residues ASP298 and ARG299 of the protein RPS5 (chain 'e') of the 30S ribosomal complex (Figure 2B). This result may indicate a relaxation of the selective pressure to maintain plastid-specific ribosomal complex stability or translation efficiency upon transition to a heterotrophic lifestyle in orchids.

ACKNOWLEDGEMENTS

We would like to thank Vasili Galka for his advice concerning the integration of the NGL protein structure viewer.

FUNDING

BSF [2013286 to I.M.]; ISF [1265/12 to I.M.; 1092/13 and 802/16 to T.P.]; E.L.K. is a fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. Funding for open access charge: BSF [2013286 to I.M.].
Conflict of interest statement. None declared.

REFERENCES

1. Martin,A.P. and Palumbi,S.R. (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 4087–4091.
2. Gillooly,J.F., Allen,A.P., West,G.B. and Brown,J.H. (2005) The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 140–145.
3. Martin,A.P. (1995) Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. *Mol. Biol. Evol.*, **12**, 1124–1131.
4. Li,W.H., Ellsworth,D.L., Krushkal,J., Chang,B.H. and Hewett-Emmett,D. (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylogenet. Evol.*, **5**, 182–187.
5. Lehtonen,J. and Lanfear,R. (2014) Generation time, life history and the substitution rate of neutral mutations. *Biol. Lett.*, **10**, 20140801.
6. Blackmon,H. and Demuth,J.P. (2015) Coleoptera karyotype database. *Coleopt. Bull.*, **69**, 174–175.
7. Parr,C.S., Wilson,N., Leary,P., Schulz,K.S., Lans,K., Walley,L., Hammock,J.A., Goddard,A., Rice,J., Studer,M. *et al.* (2014) The encyclopedia of life v2: providing global access to knowledge about life on earth. *Biodivers. Data J.*, doi:10.3897/BDJ.2.e1079.
8. Tree of Sex Consortium (2014) Tree of sex: a database of sexual systems. *Sci. Data*, **1**, 140015.
9. Rice,A., Glick,L., Abadi,S., Einhorn,M., Kopelman,N.M., Salman-Minkov,A., Mayzel,J., Chay,O. and Mayrose,I. (2015) The chromosome counts database (CCDB)—a community resource of plant chromosome numbers. *New Phytol.*, **206**, 19–26.
10. Levy Karin,E., Wicke,S., Pupko,T. and Mayrose,I. (2017) An integrated model of phenotypic trait changes and site-specific sequence evolution. *Syst. Biol.*, doi:10.1093/sysbio/syx032.
11. Mayrose,I. and Otto,S.P. (2011) A likelihood method for detecting trait-dependent shifts in the rate of molecular evolution. *Mol. Biol. Evol.*, **28**, 759–770.
12. Lartillot,N. and Poujol,R. (2011) A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.*, **28**, 729–744.
13. Bieri,P., Leibundgut,M., Saurer,M., Boehringer,D. and Ban,N. (2017) The complete structure of the chloroplast 70S ribosome in complex with translation factor pY. *EMBO J.*, **36**, 475–486.
14. Nielsen,R. (2002) Mapping mutations on phylogenies. *Syst. Biol.*, **51**, 729–739.
15. Rose,A.S. and Hildebrand,P.W. (2015) NGL viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
16. Givnish,T.J., Spalink,D., Ames,M., Lyon,S.P., Hunter,S.J., Zuluaga,A., Iles,W.J.D., Clements,M.A., Arroyo,M.T.K., Leebens-Mack,J. *et al.* (2015) Orchid phylogenomics and multiple drivers of their extraordinary diversification. *Proc. Biol. Sci.*, **282**, 20151553.
17. McCormick,M.K., Whigham,D.F. and O'Neill,J. (2004) Mycorrhizal diversity in photosynthetic terrestrial orchids. *New Phytol.*, **163**, 425–438.
18. McCormick,M.K., Lee Taylor,D., Juhaszova,K., Burnett,R.K., Whigham,D.F. and O'Neill,J.P. (2012) Limitations on orchid recruitment: not a simple picture. *Mol. Ecol.*, **21**, 1511–1523.
19. Barrett,C.F., Freudenstein,J.V, Li,J., Mayfield-Jones,D.R., Perez,L., Pires,J.C. and Santos,C. (2014) Investigating the path of plastid genome degradation in an early-transitional clade of heterotrophic orchids, and implications for heterotrophic angiosperms. *Mol. Biol. Evol.*, **31**, 3095–3112.
20. Delannoy,E., Fujii,S., Colas des Francs-Small,C., Brundrett,M. and Small,I. (2011) Rampant gene loss in the underground orchid Rhizanthella gardneri highlights evolutionary constraints on plastid genomes. *Mol. Biol. Evol.*, **28**, 2077–2086.
21. Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2016) GenBank. *Nucleic Acids Res.*, **44**, D67–D72.
22. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
23. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
24. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.