Six-state amino acid recoding is not an effective strategy to offset the effects of compositional heterogeneity and saturation in phylogenetic analyses

Alexandra M. Hernandez[1,2] and Joseph F. Ryan[1,2]

[1] Whitney Laboratory for Marine Bioscience, University of Florida, St. Augustine, FL, USA
[2] Department of Biology, University of Florida, Gainesville, FL, USA

Corresponding author: Joseph F. Ryan[1] (joseph.ryan@whitney.ufl.edu)

Abstract

Six-state amino acid recoding strategies are commonly applied to combat the effects of compositional heterogeneity and substitution saturation in phylogenetic analyses. While these methods have been endorsed from a theoretical perspective, their performance has never been extensively tested. Here, we test the effectiveness of 6-state recoding approaches by comparing the performance of analyses on recoded and non-recoded datasets that have been simulated under gradients of compositional heterogeneity or saturation. In all of our simulation analyses, non-recoding approaches greatly outperformed 6-state recoding approaches. Our results suggest that 6-state recoding strategies are not effective in the face of high saturation. Further, while recoding strategies do buffer the effects of compositional heterogeneity, the loss of information that accompanies 6-state recoding outweighs its benefits, even in the most compositionally heterogeneous datasets. In addition, we evaluate recoding schemes with 9, 12, 15, and 18 states and show that these all outperform 6-state recoding. Our results have important implications for the more than 70 published papers that have incorporated 6-state recoding, many of which have significant bearing on relationships across the tree of life.

Keywords: six-state amino acid recoding, Dayhoff 6-state recoding, S&R 6-state recoding, compositional heterogeneity, substitution saturation

Introduction

Compositional heterogeneity and substitution saturation are major challenges to phylogenetic inference. Compositional heterogeneity stems from the tendency of genes or organisms to have unequal proportions of amino acids (Collins et al. 1994; Foster and Hickey 1999). These unequal amino acid frequencies are caused by mutational and selective pressures acting at the nucleotide level (Singer and Hickey 2000; Knight et al. 2001), as well as differences in translational efficiency (Akashi and Eyre-Walker 1998). The combination of evolutionary and biological processes results in different amino acid compositions across taxa on the tree. Consequently, challenges to phylogenetic analyses arise when distantly related taxa share sequence similarities due to homoplasy (convergence), rather than descent from a common ancestor (Foster and Hickey 1999; Tarrío et al. 2001).

Similarly, phylogenetic reconstruction artifacts emerge under substitution saturation of amino acids. Substitution saturation occurs when there have been multiple amino acid substitutions at the same site washing out the evolutionary signal (Ho and Jermiin 2004). Like compositional heterogeneity, sequence saturation can lead to long branch attraction, driving unrelated taxa to group together in a clade due to homoplasy (Felsenstein 1978; Hendy and Penny 1989).

Matrix recoding has been proposed as a solution for both compositional heterogeneity and substitution saturation. Under matrix recoding methods, nucleotides or amino acids are recoded into groups based on function (Blanquart and Lartillot 2006). For example, under the RY nucleotide recoding strategy, purines (i.e., A and G) are coded with the character R and pyrimidines (i.e., T and C) are coded with the character Y (Woese et al. 1991; Phillips et al. 2001). In this recoding scenario, only transversion events are meaningful in a phylogenetic analysis. A similar recoding strategy has been implemented for amino acids, the most well-known being Dayhoff 6-state recoding. In Dayhoff 6-state recoding, chemically related amino acids that frequently replace each other are pooled together into six groups based on similar substitution scores in the Dayhoff (or PAM250) matrix (Dayhoff et al. 1978): AGPST, DENQ, HKR, ILMV, FWY, and C (Embley et al. 2003; Hrdy et al. 2004). Thus, only amino acid changes between categories, and not within categories, are considered substitutions. Since the introduction of Dayhoff 6-state recoding, several other 6-state amino acid recoding strategies based around other scoring matrices have been developed. For example, S&R 6-state recoding (Susko and Roger 2007; Feuda et al. 2017) is based on the JTT matrix (Jones et al. 1992) and KGB 6-state recoding (Kosiol et al. 2004; Feuda et al. 2017) is based on the WAG matrix (Whelan and Goldman 2001).

To date, there are at least 77 phylogenetic studies that have implemented six-state amino acid recoding strategies (Table 1). While amino acid recoding has been valued from a theoretical perspective, the performance of 6-state recoding has

never been tested empirically against non-recoding methods. In this study, we simulate datasets with either a gradient of compositional heterogeneity or saturation and compare the performance of maximum-likelihood analyses on 6-state recoded datasets to the same analyses on non-recoded datasets. We also run a subset of these analyses using 9-, 12-, 15-, and 18-state recoding schemes and compare these results to those achieved with 6-state recoded and non-recoded matrices.

Results

The Efficacy of 6-state Recoding under a Compositional Heterogeneity Gradient

We simulated data with various levels of compositional heterogeneity by matching amino acid frequencies of four non-sister 5-taxa clades on a balanced 20-taxa tree and varying the length of the stem branches leading to those four clades (Figure 1A). We scored the ability of recoding and non-recoding approaches to recover the two compositionally heterogeneous 10-taxa clades (i.e., a clade containing all A and B taxa and a clade containing all C and D taxa). As compositional heterogeneity increased, the performance of the recoding approaches diminished at a slower rate than the non-recoding approaches (Figure 1B). However, in all cases tested, non-recoding approaches performed significantly better than recoding approaches, even under the highest levels of compositional heterogeneity and shortest stem branches (Table S2; Table S3).

The Efficacy of 6-state Recoding under a Saturation Gradient

We simulated datasets on the Chang and Feuda trees under the Dayhoff and JTT models with increasing levels of saturation. Under all tested levels of saturation, phylogenetic reconstructions using the Dayhoff and LG models on non-recoded data matrices that were simulated under the Dayhoff model produced trees with fewer errors on average (as measured by Robinson-Foulds distances from the starting tree) than those that used the Dayhoff 6-state recoded matrix (Figure 2A). The results were similar for data simulated under the JTT model, where trees reconstructed with the JTT and LG models on non-recoded data matrices contained fewer errors on average across all tested levels of saturation compared to reconstructions with the S&R 6-state recoded matrix (Figure 2B). The results were consistent regardless of which topology (i.e., Chang or Feuda) was used for data simulations (Figure S2). As saturation increased, the performance of recoding approaches decreased at a faster rate than non-recoding approaches (Figure S2). T-tests performed for each branch length scaling factor parameter showed that Robinson-Foulds distances were significantly higher for recoded datasets compared to non-recoded datasets (p-value < 2.2e-16).

We also simulated data under the GTR model using the amino acid rates of substitution, amino acid frequencies, and gamma rate heterogeneity parameters estimated from the Chang dataset. Phylogenetic analyses of data simulated under GTR resulted in fewer errors on average when reconstructed with non-recoded Dayhoff matrices compared to reconstructions with the Dayhoff 6-state recoded matrices (Figure 2C). T-tests carried out for each branch length scaling factor parameter indicated that recoded approaches performed significantly worse than non-recoded approaches (p-value < 2.2e-16).

The Effect of Alternative Recoding Strategies on Compositional Heterogeneity

We used the data simulated under inflation parameter 0.5 (mid-level of compositional heterogeneity) using the hypothetical tree 0.002 (short stem branches; Figure 1A) from the main compositional heterogeneity analysis to test Dayhoff 9-, 12-, 15-, and 18-state recoding strategies and compared the performance of these methods to Dayhoff 6-state recoding and non-recoding. As in the main compositional heterogeneity analysis outlined above, trees were assessed to determine if they recovered the two compositionally heterogeneous 10-taxa clades (i.e., AB and CD). The percentage of trees that passed these criteria increased as the number of Dayhoff states increased with Dayhoff 18-state recoding outperforming all other strategies including the non-recoding approach (Figure 3). Non-recoding outperformed all other recoding strategies except Dayhoff 12- and 15-state recoding under the highest level of compositional heterogeneity (inflation parameter 0.9; Figure 3C). Since the performance of Dayhoff-18 recoding surpassed the non-recoding method under all levels of compositional heterogeneity, we performed z-tests to determine if the differences in numbers of incorrect trees between analyses run with Dayhoff 18-state recoding and those run without recoding were significant. The difference was significant ($p \leq 0.05$) only under the highest level of compositional heterogeneity (p-values for inflation parameters 0.1, 0.5, and 0.9: 0.2338, 0.1205, and 4.125e-06 respectively).

Discussion

The philosophy underlying recoding strategies in phylogenetics is that sacrificing some information is beneficial in cases where homoplasy is high, as is the case when there is substantial heterogeneity in nucleotide or amino acid composition or when datasets are highly saturated. Six-state amino acid recoding has been proposed as a strategy to improve phylogenetic reconstruction in the presence of compositional heterogeneity and saturation (Embley et al. 2003; Hrdy et al. 2004; Martin et al. 2005). While there have been simulation analyses that compare different binning schemes (Susko and Roger 2007; Nesnidal et al. 2010), there are few if any studies that compare the accuracy of 6-state recoding to non-recoding approaches. In this study, we used simulations under gradients of compositional heterogeneity and saturation to compare the performance of 6-state amino acid recoding strategies. Remarkably, we found that non-recoding approaches outperformed 6-state recoding approaches in all of our comparisons. Our results show that while 6-state recoding seems to be less affected by increases in compositional heterogeneity, it

does not overcome the penalty of information loss even under the highest levels of compositional heterogeneity (Figure 1B). Further, we found that 6-state recoding performs poorly when applied to highly saturated datasets. As such, we conclude that the costs of information loss associated with the 6-state recoding schemes are too great to justify applying these strategies.

It is possible that not all recoding strategies are inappropriate. Specifically, we found that our Dayhoff 9-, 12-, 15-, and 18-state recoding strategies performed better than the standard Dayhoff 6-state recoding approach for all tested levels of compositional heterogeneity (Figure 3). Dayhoff-18 recoding performed the best under all gradients of compositional heterogeneity and may comprise the optimum balance of minimizing compositional heterogeneity while maximizing information retention. However, we do not advocate blindly applying Dayhoff 18-state recoding, especially since significant ($p \square 0.05$) improvement only occurs under the most extreme compositional heterogeneity setting (0.9), which in no way reflects a realistic level. Instead we suggest that further simulation experiments with challenging topologies and realistic datasets are needed before adopting any amino acid recoding approach.

Applying a recoding method that is dataset specific may be another tactic to handle compositional heterogeneity or saturation. Susko and Roger (2007) and Nesnidal et al. (2010) applied this strategy by testing several recoding binning schemes informed by their datasets of interest. Tailoring the level and/or type of recoding to the amount of compositional heterogeneity and saturation, perhaps on a column-by-column basis, may be a successful approach, but further testing using such a tailored method would be necessary. Since only a handful of studies have investigated different recoding schemes, it is clear that more analyses are required to gain an understanding of the impact of alternative recoding methods for compositionally heterogeneous and/or saturated datasets.

Implications

There are at least 77 publications that use 6-state amino acid recoding, with the first seven months of 2019 seeing more than any year to date (Table 1). Many of these studies have proposed controversial topologies with profound implications across the tree of life including bacteria, archaea, unicellular eukaryotes, fungi, animals, and plants. We have shown that 6-state recoding greatly reduces information content and therefore often results in suboptimal phylogenetic reconstructions. We therefore advocate caution when interpreting results stemming from analyses that have employed 6-state recoding and contend that publications in which 6-state recoding analyses had a substantial effect on the conclusions be revisited.

Materials and Methods

Reproducibility and Transparency Statement

Custom scripts, command lines, and data used in these analyses are available at https://github.com/josephryan/Hernandez_Ryan_2019_RecodingSim

To maximize transparency and minimize confirmation bias, all analyses were pre-planned using phylotocol (DeBiasse and Ryan 2018) and pre-registered using the Center for Open Science's pre-registration platform (https://osf.io/smj6k/ and https://osf.io/6ubgj/). We made three changes to our original plan during the life of this project, and these changes were documented and justified in the phylotocol available on our GitHub repository (URL above).

Overview of Empirical Datasets Employed

The following methods can be divided into two main analyses: compositional heterogeneity and saturation. Both analyses employ empirical data from the following papers: Chang et al. (2015) hereafter "Chang," and Feuda et al. (2017) hereafter "Feuda." The topologies from Chang and Feuda are based on the same dataset which is made up of 51,940 amino acid positions from 77 taxa representing a wide range of animals and 9 non-animal outgroups. Feuda extensively applied 6-state amino acid recoding to this dataset in a reanalysis of the Chang study, which did not use recoding.

For the compositional heterogeneity analysis, we use several hypothetical 20-taxa symmetrical trees which consist of 4 clades (named A, B, C, and D) made up of 5 taxa each (Figure 1A), and apply global parameters estimated from the Chang dataset. For the saturation analysis, we use the topologies reported in Chang and Feuda. More details on these analyses are provided below.

Testing 6-state Recoding Performance on Compositional Heterogeneity

We used the script comphet.pl (available in our GitHub repository) to simulate amino acid data in P4 (Foster 2004) on four hypothetical 20-taxa balanced trees (Figure 1A). We simulated sequences that were 1,000 amino acids in length under the GTR model using the amino acid rates of substitution from the Chang dataset. To introduce compositional heterogeneity, we generated one set of amino acid frequencies for clades A and C and a different set of frequencies for clades B and D. For clades B and D, we used the amino acid frequencies from the Chang dataset. For clades A and C, frequencies for the following pairs of amino acids ((A,L), (R,K), (N,M), (D,F), (C,P), (Q,S), (E,T), (G,W), (H,Y), (I,V); chosen based on an alphabetical pattern) were determined by adjusting each frequency by X, where X is the inflation parameter (i.e., 0.1, 0.5, 0.9) multiplied by the lowest frequency of the pair. The amino acid of the pair with the lowest frequency is incremented by X and the other is decremented by X.

For example, the Chang frequencies for the amino acids R and K are 0.063 and 0.080 respectively. These frequencies were used for clades B and D without adjustment. To determine the increment value X under the inflation parameter 0.1, we multiplied the frequency of R, which is the lowest of the pair, by 0.1 (X=0.0063). We then added X to the Chang frequency of R (0.063 + 0.0063) and subtracted X from the Chang frequency of K (0.080 - 0.0063). We rounded these values to 3 decimal places (to work with P4) for a final set of frequencies of R = 0.069 and K=0.074.

Using this algorithm to generate frequencies, we performed 1,000 simulations for each combination of the four hypothetical 20-taxa trees and the three inflation parameters (i.e., 0.1, 0.5, and 0.9) resulting in 12,000 total datasets. To verify that these datasets displayed compositional heterogeneity between clades A and C compared to clades B and D, we computed amino acid frequencies on the simulated datasets and summed the difference in frequencies across all replicates. We subtracted the average difference in amino acid frequencies between clades A and C and between clades B and D (clade pairs with homogeneous composition) from the average difference in amino acid frequencies between clades A and B, A and D, B and C, and C and D (clade pairs with heterogenous composition) to generate a compositional heterogeneity (comp-het) index value. Datasets with comp-het index values closer to 0 are characterized by low compositional heterogeneity, while datasets with higher comp-het index values are characterized by high compositional heterogeneity.

We recoded each simulated dataset with both Dayhoff 6-state recoding and S&R 6-state recoding, and then reconstructed maximum-likelihood trees of the recoded datasets using the GTR multi-state model and of the non-recoded datasets using the Dayhoff and JTT models in RAxML (Stamatakis 2014). In total we produced 48,000 phylogenies for testing compositional heterogeneity. We used the script is_mono.pl (available in our GitHub repository) to determine whether each tree recovered a monophyletic group that included all A and B taxa, which by definition would include a monophyletic group that included all C and D taxa. We did not test for more fine-scale relationships as our goal was to evaluate the degree to which the applied level of compositional heterogeneity was pulling together the compositionally homogenous clades A and C, and B and D. We calculated the percentage of incorrect trees using the above criteria for each combination of model, recoding type (including no recoding), and level of applied compositional heterogeneity (i.e., inflation parameter), and performed a z-test to compare the proportions of incorrect trees between non-recoding and recoding approaches.

Testing 6-state Recoding Performance on Saturation

We used Seq-Gen (Rambaut and Grass 1997) to simulate the evolution of amino acids on the Chang and Feuda topologies. First, we confirmed that increasing the branch length scaling factor parameter in Seq-Gen linearly increased levels of saturation (Figure S1) using the script seq-gen_saturation_test.pl (available in the accompanying GitHub repository). Next, we performed 1,000 simulations per combination of tree topology (Chang and Feuda), branch length scaling factor parameter (1–20), and model of amino acid substitution (either Dayhoff or JTT) for a total of 80,000 datasets. We simulated an additional 1,000 datasets on the Chang topology for a subset of branch length scaling factor parameters (1, 5, 10, 15, 20) under the GTR model using the amino acid rates of substitution, amino acid frequencies, and gamma rate heterogeneity from the Chang dataset, bringing the grand total to 85,000 datasets. Each dataset included 1,000 amino acid columns.

For simulations performed on the Chang topology, we increased the branch length scaling factor parameter from 1 to 20 in increments of 1. The Feuda topology was produced from recoding the Chang dataset (Feuda et al. 2017), and because trees produced from recoded data have substantially fewer substitutions and therefore shorter branch lengths, we incremented branch lengths by a factor of 2.6 for the Feuda tree (based on our calculation that the sum of branch lengths in the recoded tree was 2.6 shorter than the sum of branch lengths in the non-recoded Chang tree).

We performed maximum-likelihood analyses with RAxML for each set of sequences produced from simulations over the Chang and Feuda topologies. For the datasets simulated with Dayhoff and JTT substitution models, we reconstructed trees using the generating model, the 6-state recoding scheme derived from that model, and for a subset of branch length scaling factor parameters (1, 5, 10, 15, 20) we also reconstructed trees using LG, a sub-optimal model in this context, as it was not the model used for the simulations. For the datasets simulated with the GTR substitution model, we generated trees using Dayhoff and Dayhoff 6-state recoding. We produced 180,000 phylogenies in total to test saturation. To test the performance of each recoding (or non-recoding) scheme, we used TOPD/FMTS (Puigbo et al. 2007) to calculate Robinson-Foulds distances (Robinson and Foulds 1981) between the topology used for simulation (i.e., Chang or Feuda) and the reconstructed trees generated from simulated sequences. We used a t-test to determine if there were significant differences in Robinson-Foulds distances between recoded and non-recoded datasets for each branch length scaling factor.

Testing Alternative Recoding Strategies on Compositional Heterogeneity

To test the effect of number of states on recoding strategies, we developed alternative Dayhoff 9-, 12-, 15-, and 18-state recoding strategies. The first step in these analyses was to determine the optimal amino acid binning strategy for each number of tested states. Since the number of possible bins for each state is finite, ideally, we would use an exhaustive algorithm to identify the binning scheme that maximizes the sum of intra-bin substitution scores using the Dayhoff matrix. Unfortunately, as pointed out by Susko and Roger (2007), the number of possible bins is very large (e.g., there are roughly 1.5 x 1013 choices of bins under an 8-state recoding strategy) and an

exhaustive algorithm is computationally intractable. Instead, we generated scores (see score.pl in our GitHub repository) for several binning schemes that incorporated subsets of the Dayhoff 6-state recoding bins and chose the best-scoring binning strategies from this set (Table S1). We also compared our best binning strategies to those proposed in Susko and Roger (2007) and in all cases, the scores we generated were higher, except for one which had an equal score (not entirely surprising given that the Susko and Roger bins were optimized for JTT recoding).

We compared the binning schemes that scored the highest for each recoding strategy (Table 2) against the Dayhoff and Dayhoff 6-state recoded matrices by testing their performance under reasonably high levels of compositional heterogeneity. We recoded the data that we simulated for the compositional heterogeneity analysis (data simulated with inflation parameter 0.5 using the hypothetical tree 0.002 (Figure 1A)) using our Dayhoff 9-, 12-, 15-, and 18-state recoding strategies and reconstructed maximum-likelihood trees in RAxML. As in the main compositional heterogeneity analysis outlined above, we used the script is_mono_comphet.pl to test if the ten taxa labeled A and B were monophyletic and likewise the ten taxa labeled C and D were monophyletic. We also performed a z-test to compare the proportion of incorrect trees produced under Dayhoff-18 recoding (see Results for rationale) to those produced under non-recoding.

Supplementary Material

All commands and versions of software used in this study are provided in the supplementary material. All data and scripts are available in the following GitHub repository: https://github.com/josephryan/Hernandez_Ryan_2019_RecodingSim.

Funding

This work was supported by the National Science Foundation under Grant Number 1542597; and the Graduate Research Fellowship Program to A.M.H. Additional funding to A.M.H. was provided by the Florida Education Fund Mcknight Doctoral Fellowship Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We thank Melissa DeBiasse for providing comments on an earlier version of the manuscript and Gordon Burleigh, Christine Schnitzler, Marta Wayne, and Bryan Kolaczkowski for feedback on this project during A.M.H.'s committee meeting. The authors would like to express their thanks to David Swofford and Gavin Naylor for influential discussions at an early stage of the project. The views expressed in this paper do not necessarily reflect the views of those acknowledged.

References

Andersson J.O., Hirt R.P., Foster P.G., Roger A.J. 2006. Evolution of four gene families with patchy phylogenetic distributions: influx of genes into protist genomes. BMC Evol. Biol. 6:27.

Akashi H., Eyre-Walker A. 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. 8:688–693.

Aouad M., Taib N., Oudart A., Lecocq M., Gouy M., Brochier-Armanet C. 2018. Extreme halophilic archaea derive from two distinct methanogen Class II lineages. Mol. Phylogenet. Evol. 127:46–54.

Blanquart S., Lartillot N. 2006. A Bayesian Compound Stochastic Process for Modeling Nonstationary and Nonhomogeneous Sequence Evolution. Mol. Biol. Evol. 23:2058–2071.

Bennett G.M., Mao M. 2018. Comparative genomics of a quadripartite symbiosis in a planthopper host reveals the origins and rearranged nutritional responsibilities of anciently diverged bacterial lineages. Environ. Microbiol. 20:4461–4472.

Borowiec M.L., Lee E.K., Chiu J.C., Plachetzki D.C. 2015. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics. 16:987.

Brochier-Armanet C., Forterre P., Gribaldo S. 2011. Phylogeny and evolution of the Archaea: one hundred genomes later. Curr. Opin. Microbiol. 14:274–281.

Burki F., Okamoto N., Pombert J.-F., Keeling P.J. 2012. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. Proc. R. Soc. B Biol. Sci. 279:2246–2254.

Chang E.S., Neuhof M., Rubinstein N.D., Diamant A., Philippe H., Huchon D., Cartwright P. 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proc. Natl. Acad. Sci. U. S. A. 112:14912–7.

Collins T.M., Wimberger P.H., Naylor G.J.P. 1994. Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. Syst. Biol. 43:482–496.

Cox C.J., Foster P.G., Hirt R.P., Harris S.R., Embley T.M. 2008. The archaebacterial origin of eukaryotes. Proc. Natl. Acad. Sci. U. S. A. 105:20356–61.

Cunha T.J., Giribet G. 2019. A congruent topology for deep gastropod relationships. Proc. R. Soc. B Biol. Sci. 286:20182776.

Davidson E.A., van der Giezen M., Horner D.S., Embley T.M., Howe C.J. 2002. An [Fe] hydrogenase from the anaerobic hydrogenosome-containing fungus Neocallimastixfrontalis L2. Gene. 296:45–52.

DeBiasse M.B., Ryan J.F. 2018. Phylotocol: Promoting Transparency and Overcoming Bias in Phylogenetics. Syst. Biol., syy090.

Derelle R., Lang B.F. 2012. Rooting the Eukaryotic Tree with Mitochondrial and Bacterial Proteins. Mol. Biol. Evol. 29:1277–1289.

Derelle R., Torruella G., Klimeš V., Brinkmann H., Kim E., Vlček Č., Lang B.F., Eliáš M. 2015. Bacterial proteins pinpoint a single eukaryotic root. Proc. Natl. Acad. Sci. 112:E693–E699.

Delsuc F., Brinkmann H., Chourrout D., Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature. 439:965–968.

Deschamps P., Moreira D. 2009. Signal Conflicts in the Phylogeny of the Primary Photosynthetic Eukaryotes. Mol. Biol. Evol. 26:2745–2753.

Domman D., Horn M., Embley T.M., Williams T.A. 2015. Plastid establishment did not require a chlamydial partner. Nat. Commun. 6:6421.

Eitel M., Francis W.R., Varoqueaux F., Daraspe J., Osigus H.-J., Krebs S., Vargas S., Blum H., Williams G.A., Schierwater B., Wörheide G. 2018. Comparative genomics and the nature of placozoan species. PLOS Biol. 16:e2005359.

Embley M., van der Giezen M., Horner D.S., Dyal P.L., Foster P. 2003a. Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. Philos. Trans. R. Soc. London. Ser. B Biol. Sci. 358:191–203.

Embley T.M., van der Giezen M., Horner D., Dyal P., Bell S., Foster P. 2003b. Hydrogenosomes, Mitochondria and Early Eukaryotic Evolution. IUBMB Life (International Union Biochem. Mol. Biol. Life). 55:387–395.

Feuda R., Dohrmann M., Pett W., Lartillot N., Wö G., Pisani D., De C.W. 2017. Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. Curr. Biol. 27:3864–3870.

Fitzpatrick D.A., Creevey C.J., McInerney J.O. 2006. Genome Phylogenies Indicate a Meaningful α-Proteobacterial Phylogeny and Support a Grouping of the Mitochondria with the Rickettsiales. Mol. Biol. Evol. 23:74–85.

Fitzpatrick D.A., Logue M.E., Stajich J.E., Butler G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol. Biol. 6:99.

Foster P.G., Cox C.J., Embley T.M. 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos. Trans. R. Soc. B Biol. Sci. 364:2197–2207.

Foster P.G., Hickey D.A. 1999. Compositional Bias May Affect Both DNA-Based and Protein-Based Phylogenetic Reconstructions. J. Mol. Evol. 48:284–290.

Fu C.-J., Sheikh S., Miao W., Andersson S.G.E., Baldauf S.L. 2014. Missing genes, multiple ORFs, and C-to-U type RNA editing in Acrasis kona (Heterolobosea, Excavata) mitochondrial DNA. Genome Biol. Evol. 6:2240–57.

Haen K.M., Lang B.F., Pomponi S.A., Lavrov D. V. 2007. Glass Sponges and Bilaterian Animals Share Derived Mitochondrial Genomic Features: A Common Ancestry or Parallel Evolution? Mol. Biol. Evol. 24:1518–1527.

He D., Sierra R., Pawlowski J., Baldauf S.L. 2016. Reducing long-branch effects in multi-protein data uncovers a close relationship between Alveolata and Rhizaria. Mol. Phylogenet. Evol. 101:1–7.

Heinz E., Williams T.A., Nakjang S., Noël C.J., Swan D.C., Goldberg A. V., Harris S.R., Weinmaier T., Markert S., Becher D., Bernhardt J., Dagan T., Hacker C., Lucocq J.M., Schweder T., Rattei T., Hall N., Hirt R.P., Embley T.M. 2012. The Genome of the Obligate Intracellular Parasite Trachipleistophora hominis: New Insights into Microsporidian Genome Dynamics and Reductive Evolution. PLoS Pathog. 8:e1002979.

Hill M.S., Hill A.L., Lopez J., Peterson K.J., Pomponi S., Diaz M.C., Thacker R.W., Adamska M., Boury-Esnault N., Cárdenas P., Chaves-Fonnegra A., Danka E., De Laine B.-O., Formica D., Hajdu E., Lobo-Hajdu G., Klontz S., Morrow C.C., Patel J., Picton B., Pisani D., Pohlmann D., Redmond N.E., Reed J., Richey S., Riesgo A., Rubin E., Russell Z., Rützler K., Sperling E.A., di Stefano M., Tarver J.E., Collins A.G. 2013. Reconstruction of Family-Level Phylogenetic Relationships within Demospongiae (Porifera) Using Nuclear Encoded Housekeeping Genes. PLoS One. 8:e50437.

Ho S.Y.W., Jermiin L.S. 2004. Tracing the Decay of the Historical Signal in Biological Sequence Data. Syst. Biol. 53:623–637.

Hrdy I., Hirt R.P., Dolezal P., Bardonová L., Foster P.G., Tachezy J., Martin Embley T. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature. 432:618–622.

Kayal E., Roure B., Philippe H., Collins A.G., Lavrov D. V. 2013. Cnidarian phylogenetic relationships as revealed by mitogenomics. BMC Evol. Biol. 13:5.

Knight R.D., Freeland S.J., Landweber L.F. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. 2:research0010.1.

Kosiol C., Goldman N., H. Buttimore N. 2004. A new criterion and method for amino acid classification. J. Theor. Biol. 228:97–106.

Lasek-Nesselquist E. 2012. A Mitogenomic Re-Evaluation of the Bdelloid Phylogeny and Relationships among the Syndermata. PLoS One. 7:e43554.

Lasek-Nesselquist E., Gogarten J.P. 2013. The effects of model choice and mitigating bias on the ribosomal tree of life. Mol. Phylogenet. Evol. 69:17–38.

Laumer C.E., Fernández R., Lemer S., Combosch D., Kocot K.M., Riesgo A., Andrade S.C.S., Sterrer W., Sørensen M. V., Giribet G. 2019. Revisiting metazoan phylogeny with genomic sampling of all phyla. Proc. R. Soc. B Biol. Sci. 286:20190831.

Laumer C.E., Gruber-Vodicka H., Hadfield M.G., Pearse V.B., Riesgo A., Marioni J.C., Giribet G. 2018. Support for a clade of Placozoa and Cnidaria in genes with minimal compositional bias. Elife. 7.

Leliaert F., Tronholm A., Lemieux C., Turmel M., DePriest M.S., Bhattacharya D., Karol K.G., Fredericq S., Zechman F.W., Lopez-Bautista J.M. 2016. Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. Sci. Rep. 6:25367.

Lemer S., Bieler R., Giribet G. 2019. Resolving the relationships of clams and cockles: dense transcriptome sampling drastically improves the bivalve tree of life. Proc. R. Soc. B Biol. Sci. 286:20182684.

Lemieux C., Otis C., Turmel M. 2014. Chloroplast phylogenomic analysis resolves deep-level relationships within the green algal class Trebouxiophyceae. BMC Evol. Biol. 14:211.

Luo H. 2015. Evolutionary origin of a streamlined marine bacterioplankton lineage. ISME J. 9:1423–1433.

Luo H., Csűros M., Hughes A.L., Moran M.A. 2013. Evolution of Divergent Life History Strategies in Marine Alphaproteobacteria. MBio. 4:e00373-13.

Luo H., Swan B.K., Stepanauskas R., Hughes A.L., Moran M.A. 2014. Evolutionary analysis of a streamlined lineage of surface ocean Roseobacters. ISME J. 8:1428–1439.

Manzano-Marín A., Coeur d'acier A., Clamens A.-L., Orvain C., Cruaud C., Barbe V., Jousselin E. 2018. A Freeloader? The Highly Eroded Yet Large Genome of the Serratia symbiotica Symbiont of Cinara strobi. Genome Biol. Evol. 10:2178–2189.

Marlétaz F., Peijnenburg K.T.C.A., Goto T., Satoh N., Rokhsar D.S. 2019. A New Spiralian Phylogeny Places the Enigmatic Arrow Worms among Gnathiferans. Curr. Biol. 29:312-318.e3.

Martin W., Deusch O., Stawski N., Grünheit N., Goremykin V. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10:203–209.

Masta S.E., Longhorn S.J., Boore J.L. 2009. Arachnid relationships based on mitochondrial genomes: Asymmetric nucleotide and amino acid bias affects phylogenetic analyses. Mol. Phylogenet. Evol. 50:117–128.

Matsumoto T., Shinozaki F., Chikuni T., Yabuki A., Takishita K., Kawachi M., Nakayama T., Inouye I., Hashimoto T., Inagaki Y. 2011. Green-colored Plastids in the Dinoflagellate Genus

Lepidodinium are of Core Chlorophyte Origin. Protist. 162:268–276.

Morgan C.C., Foster P.G., Webb A.E., Pisani D., McInerney J.O., O'Connell M.J. 2013. Heterogeneous Models Place the Root of the Placental Mammal Phylogeny. Mol. Biol. Evol. 30:2145–2156.

Narayanan Kutty S., Meusemann K., Bayless K.M., Marinho M.A.T., Pont A.C., Zhou X., Misof B., Wiegmann B.M., Yeates D., Cerretti P., Meier R., Pape T. 2019. Phylogenomic analysis of Calyptratae: resolving the phylogenetic relationships within a major radiation of Diptera. Cladistics.

Nesnidal M.P., Helmkampf M., Bruchhaus I., Hausdorf B. 2010. Compositional Heterogeneity and Phylogenomic Inference of Metazoan Relationships. Mol. Biol. Evol. 27:2095–2104.

Nishimura Y., Kamikawa R., Hashimoto T., Inagaki Y. 2012. Separate Origins of Group I Introns in Two Mitochondrial Genes of the Katablepharid Leucocryptos marina. PLoS One. 7:e37307.

O'Halloran D.M., Fitzpatrick D.A., McCormack G.P., McInerney J.O., Burnell A.M. 2006. The Molecular Phylogeny of a Nematode-Specific Clade of Heterotrimeric G-Protein α-Subunit Genes. J. Mol. Evol. 63:87–94.

Ometto L., Cestaro A., Ramasamy S., Grassi A., Revadi S., Siozios S., Moretto M., Fontana P., Varotto C., Pisani D., Dekker T., Wrobel N., Viola R., Pertot I., Cavalieri D., Blaxter M., Anfora G., Rota-Stabelli O. 2013. Linking Genomics and Ecology to Investigate the Complex Evolution of an Invasive Drosophila Pest. Genome Biol. Evol. 5:745–757.

Otero-Bravo A., Goffredi S., Sabree Z.L. 2018. Cladogenesis and Genomic Streamlining in Extracellular Endosymbionts of Tropical Stink Bugs. Genome Biol. Evol. 10:680–693.

Parfrey L.W., Grant J., Tekle Y.I., Lasek-Nesselquist E., Morrison H.G., Sogin M.L., Patterson D.J., Katz L.A. 2010. Broadly Sampled Multigene Analyses Yield a Well-Resolved Eukaryotic Tree of Life. Syst. Biol. 59:518–533.

Petitjean C., Deschamps P., López-García P., Moreira D., Brochier-Armanet C. 2015. Extending the Conserved Phylogenetic Core of Archaea Disentangles the Evolution of the Third Domain of Life. Mol. Biol. Evol. 32:1242–1254.

Philip G.K., Creevey C.J., McInerney J.O. 2005. The Opisthokonta and the Ecdysozoa May Not Be Clades: Stronger Support for the Grouping of Plant and Animal than for Animal and Fungi and Stronger Support for the Coelomata than Ecdysozoa. Mol. Biol. Evol. 22:1175–1184.

Philippe H., Brinkmann H., Copley R.R., Moroz L.L., Nakano H., Poustka A.J., Wallberg A., Peterson K.J., Telford M.J. 2011. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature. 470:255–258.

Philippe H., Poustka A.J., Chiodin M., Hoff K.J., Dessimoz C., Tomiczek B., Schiffer P.H., Müller S., Domman D., Horn M., Kuhl H., Timmermann B., Satoh N., Hikosaka-Katayama T., Nakano H., Rowe M.L., Elphick M.R., Thomas-Chollier M., Hankeln T., Mertes F., Wallberg A., Rast J.P., Copley R.R., Martinez P., Telford M.J. 2019. Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria. Curr. Biol. 0.

Phillips M.J., Lin Y.-H., Harrison G., Penny D. 2001. Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials. Proc. R. Soc. London. Ser. B Biol. Sci. 268:1533–1538.

Pons J., Ribera I., Bertranpetit J., Balke M. 2010. Nucleotide substitution rates for the full set of mitochondrial protein-coding genes in Coleoptera. Mol. Phylogenet. Evol. 56:796–807.

Puigbo P., Garcia-Vallve S., McInerney J.O. 2007. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics. 23:1556–1558.

Puttick M.N., Morris J.L., Williams T.A., Cox C.J., Edwards D., Kenrick P., Pressel S., Wellman C.H., Schneider H., Pisani D., Donoghue P.C.J. 2018. The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. Curr. Biol. 28:733-745.e2.

Rambaut A., Grass N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics. 13:235–238.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rodríguez-Ezpeleta N., Embley T.M. 2012. The SAR11 Group of Alpha-Proteobacteria Is Not Related to the Origin of Mitochondria. PLoS One. 7:e30520.

Rota-Stabelli O., Lartillot N., Philippe H., Pisani D. 2013. Serine Codon-Usage Bias in Deep Phylogenomics: Pancrustacean Relationships as a Case Study. Syst. Biol. 62:121–133.

Schwentner M., Combosch D.J., Pakes Nelson J., Giribet G. 2017. A Phylogenomic Solution to the Origin of Insects by Resolving Crustacean-Hexapod Relationships. Curr. Biol. 27:1818-1824.e5.

Schwentner M., Richter S., Rogers D.C., Giribet G. 2018. Tetraconatan phylogeny with special focus on Malacostraca and Branchiopoda: highlighting the strength of taxon-specific matrices in phylogenomics. Proc. R. Soc. B Biol. Sci. 285:20181524.

Shin S., Clarke D.J., Lemmon A.R., Moriarty Lemmon E., Aitken A.L., Haddad S., Farrell B.D., Marvaldi A.E., Oberprieler R.G., McKenna D.D. 2017. Phylogenomic Data Yield New and Robust Insights into the Phylogeny and Evolution of Weevils. Mol. Biol. Evol. 35:823–836.

Simion P., Philippe H., Baurain D., Jager M., Richter D.J., Di Franco A., Roure B., Satoh N., Quéinnec É., Ereskovsky A., Lapébie P., Corre E., Delsuc F., King N., Wörheide G., Manuel M. 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Curr. Biol. 27:958–967.

Singer G.A.C., Hickey D.A. 2000. Nucleotide Bias Causes a Genomewide Bias in the Amino Acid Composition of Proteins. Mol. Biol. Evol. 17:1581–1588.

Song N., An S.-H., Yin X.-M., Zhao T., Wang X.-Y. 2016. Insufficient resolving power of mitogenome data in deciphering deep phylogeny of Holometabola. J. Syst. Evol. 54:545–559.

Sousa F., Foster P.G., Donoghue P.C.J., Schneider H., Cox C.J. 2018. Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.). New Phytol. 222: 565-575.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Susko E., Roger A.J. 2007. On Reduced Amino Acid Alphabets for Phylogenetic Inference. Mol. Biol. Evol. 24:2139–2150.

Szabó G., Schulz F., Toenshoff E.R., Volland J.-M., Finkel O.M., Belkin S., Horn M. 2017. Convergent patterns in the evolution of mealybug symbioses involving different intrabacterial symbionts. ISME J. 11:715–726.

Tarrío R., Rodríguez-Trelles F., Ayala F.J. 2001. Shared Nucleotide Composition Biases Among Species and Their Impact on Phylogenetic Reconstructions of the Drosophilidae. Mol. Biol. Evol. 18:1464–1473.

Torruella G., Derelle R., Paps J., Lang B.F., Roger A.J., Shalchian-Tabrizi K., Ruiz-Trillo I. 2011. Phylogenetic Relationships within the Opisthokonta Based on Phylogenomic Analyses of

Conserved Single-Copy Protein Domains. Mol. Biol. Evol. 29:531–544.

Uribe J.E., Irisarri I., Templado J., Zardoya R. 2019. New patellogastropod mitogenomes help counteracting long-branch attraction in the deep phylogeny of gastropod mollusks. Mol. Phylogenet. Evol. 133:12–23.

Wang X., Lavrov D. V. 2006. Mitochondrial Genome of the Homoscleromorph Oscarella carmela (Porifera, Demospongiae) Reveals Unexpected Complexity in the Common Ancestor of Sponges and Other Animals. Mol. Biol. Evol. 24:363–373.

Wang Z., Wu M. 2015. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. Sci. Rep. 5:7949.

Williams T.A., Embley T.M., Heinz E. 2011. Informational Gene Phylogenies Do Not Support a Fourth Domain of Life for Nucleocytoplasmic Large DNA Viruses. PLoS One. 6:e21080.

Williams T.A., Szöllősi G.J., Spang A., Foster P.G., Heaps S.E., Boussau B., Ettema T.J.G., Embley T.M. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc. Natl. Acad. Sci. U. S. A. 114:E4602–E4611.

Wodniok S., Brinkmann H., Glöckner G., Heidel A.J., Philippe H., Melkonian M., Becker B. 2011. Origin of land plants: Do conjugating green algae hold the key? BMC Evol. Biol. 11:104.

Woese C.R., Achenbach L., Rouviere P., Mandelco L. 1991. Archaeal Phylogeny: Reexamination of the Phylogenetic Position of Archaeoglohus fulgidus in Light of Certain Composition-induced Artifacts. Syst. Appl. Microbiol. 14:364–371.

Wolfe J.M., Breinholt J.W., Crandall K.A., Lemmon A.R., Lemmon E.M., Timm L.E., Siddall M.E., Bracken-Grissom H.D. 2019. A phylogenomic framework, evolutionary timeline and genomic resources for comparative studies of decapod crustaceans. Proc. R. Soc. B Biol. Sci. 286:20190079.

Yoshida Y., Koutsovoulos G., Laetsch D.R., Stevens L., Kumar S., Horikawa D.D., Ishino K., Komine S., Kunieda T., Tomita M., Blaxter M., Arakawa K. 2017. Comparative genomics of the tardigrades Hypsibius dujardini and Ramazzottius varieornatus. PLOS Biol. 15:e2002266.

Zhang Y., Sun Y., Jiao N., Stepanauskas R., Luo H. 2016. Ecological Genomics of the Uncultivated Marine Roseobacter Lineage CHAB-I-5. Appl. Environ. Microbiol. 82:2100–2111.

Zverkov O.A., Mikhailov K. V., Isaev S. V., Rusin L.Y., Popova O. V., Logacheva M.D., Penin A.A., Moroz L.L., Panchin Y. V., Lyubetsky V.A., Aleoshin V. V. 2019. Dicyemida and Orthonectida: Two Stories of Body Plan Simplification. Front. Genet. 10:443.

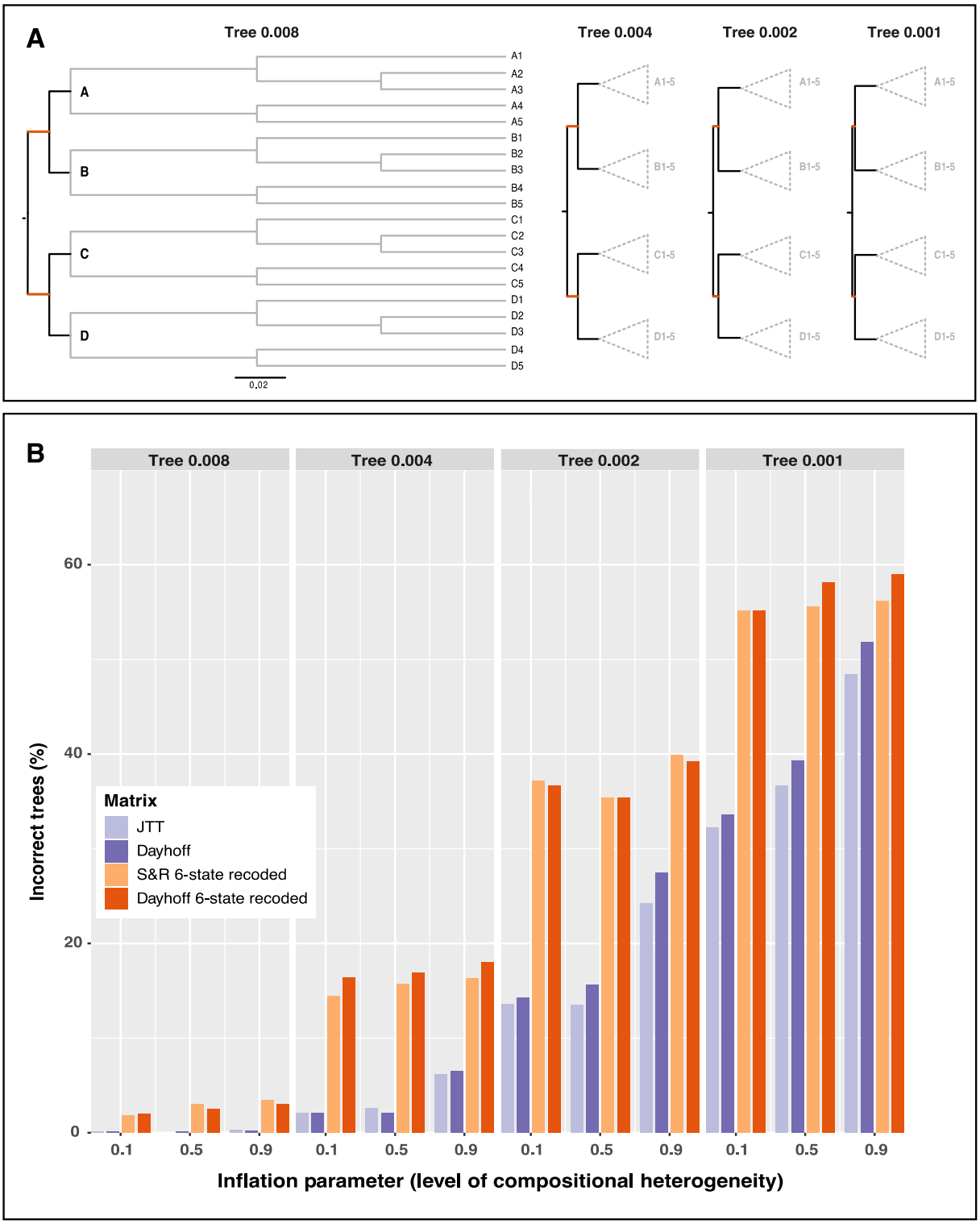

Figure 1. Six-state recoding approaches produce more incorrect trees under various levels of compositional heterogeneity. (a) Trees used for simulations. The value in the name of the

tree (e.g., 0.008 in Tree 0.008) denotes the length in substitutions per site of the stem branches of the AB and CD clades (highlighted in orange). Decreasing the lengths of these branches increased the effect of compositional heterogeneity (Figure S3). **(b)** Percentage of 1000 trees that did not reconstruct a monophyletic group of taxa from clades A and B and monophyletic group of taxa from clades C and D.

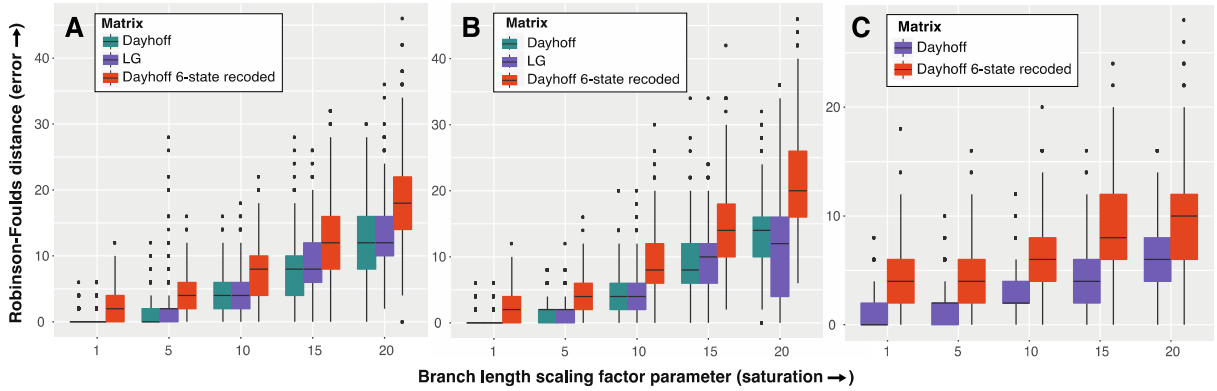

Figure 2. Six-state recoding approaches produce more errors under increasing levels of saturation. Robinson-Foulds distances of all 1,000 runs for each branch length scaling factor parameter. All data were simulated on the Chang tree topology. **(a)** Datasets simulated under the Dayhoff model. **(b)** Datasets simulated under the JTT model. **(c)** Datasets simulated under the GTR model using the amino acid rates of substitution, amino acid frequencies, and gamma rate heterogeneity estimated from the Chang dataset.

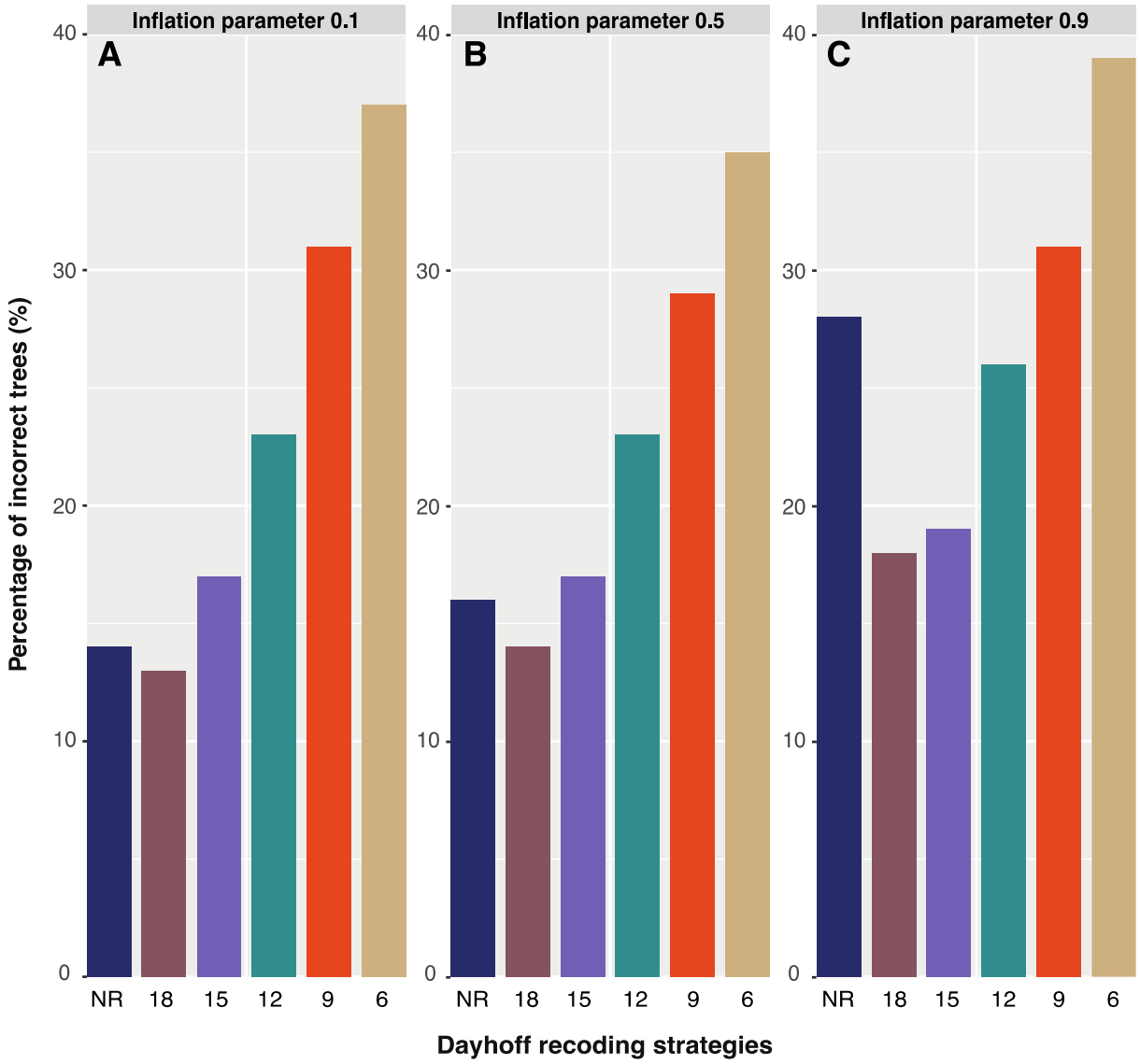

Figure 3. Dayhoff 9-, 12-, 15-, and 18-state recoding produce fewer incorrect trees than Dayhoff 6-state recoding under various levels of compositional heterogeneity. Trees were reconstructed by applying the non-recoded (NR) Dayhoff matrix or alternative Dayhoff recoding strategies (the number of states in the recoding strategy is indicated by digits). Incorrect trees did not include a monophyletic group of taxa from clades A and B and monophyletic group of taxa from clades C and D. The Y-axis refers to percentage out of 1,000 trees.

Citation	Recoding in main figure	Organismal scope or featured taxon
Cunha & Giribet (2019)	yes	Gastropods
Laumer et al. (2019)	yes	Animals
Lemer et al. (2019)	yes	Bivalves
Lozano-Fernandez et al. (2019)	yes	Chelicerates
Marlétaz et al. (2019)	yes	Spiralia
Philippe et al. (2019)	yes	Bilateria
Narayanan Kutty et al. (2019)	no	Calyptratae
Uribe et al. (2019)	no	Gastropods
Wolfe et al. (2019)	no	Decapod crustaceans
Zverkov et al. (2019)	no	Dicyemida and Orthonectida
Aouad et al. (2018)	yes	Archaea
Laumer et al. (2018)	yes	Placozoa
Otero-Bravo et al. (2018)	yes	*Pantoea*
Puttick et al. (2018)	yes	Land plants
Schwentner et al. (2018)	yes	Pancrustacea
Sousa et al. (2018)	yes	Land plants
Bennett & Mao (2018)	no	Fulgoroidea symbionts
Eitel et al. (2018)	no	Placozoa
Manzano-Marín et al. (2018)	no	*Cinara strobi* symbionts
Feuda et al. (2017)	yes	Animals
Szabó et al. (2017)	yes	Pseudococcidae symbionts
Williams et al. (2017)	yes	Archaea
Schwentner et al. (2017)	no	Pancrustacea
Shin et al. (2017)	no	Curculionoidea
Simion et al. (2017)	no	Animals
Yoshida et al. (2017)	no	Tardigrades
Leliaert et al. (2016)	yes	Viridiplantae
Zhang et al. (2016)	yes	Roseobacter CHAB-I-5 lineage
He et al. (2016)	no	Rhizaria
Song et al. (2016)	no	Holometabola
Domman et al. (2015)	yes	Plastids
Luo (2015)	yes	SAR11
Petitjean et al. (2015)	yes	Archaea

Borowiec et al. (2015)	no	Animals
Derelle et al. (2015)	no	Eukaryotes
Wang & Wu (2015)	no	Mitochondria
Luo et al. (2014)	yes	Roseobacter
Fu et al. (2014)	no	Discoba
Lemieux et al. (2014)	no	Trebouxiophyceae
Luo et al. (2013)	yes	Marine Alphaproteobacteria
Morgan et al. (2013)	yes	Placental mammals
Rota-Stabelli et al. (2013)	yes	Pancrustacea
Hill et al. (2013)	no	Demospongiae
Kayal et al. (2013)	no	Cnidaria
Lasek-Nesselquist & Gogarten (2013)	no	3 domains (eukaryotes, archaea, bacteria)
Ometto et al. (2013)	no	*Drosophila suzukii*
Lasek-Nesselquist (2012)	yes	Syndermata
Rodríguez-Ezpeleta & Embley (2012)	yes	SAR11
Burki et al. (2012)	no	Plastids
Derelle & Lang (2012)	no	Eukaryotes
Heinz et al. (2012)	no	*Trachipleistophora hominis*
Nishimura et al. (2012)	no	Mitochondria
Brochier-Armanet et al. (2011)	yes	Archaea
Williams et al. (2011)	yes	Nucleocytoplasmic large DNA virus
Matsumoto et al. (2011)	no	Plastids
Philippe et al. (2011)	no	Xenacoelomorpha
Wodniok et al. (2011)	no	Streptophyte algae and land plants
Torruella et al. (2011)	no	Opisthokonta
Parfrey et al. (2010)	no	Eukaryotes
Pons et al. (2010)	no	Coleoptera
Deschamps & Moreira (2009)	yes	Archaeplastida
Foster et al. (2009)	yes	Eukaryotes
Masta et al. (2009)	yes	Arachnida
Cox et al. (2008)	yes	Eukaryotes
Haen et al. (2007)	no	Hexactinellida
Andersson et al. (2006)	yes	Eukaryotes

Fitzpatrick et al. (2006)	yes	Mitochondria
Fitzpatrick et al. (2006)	yes	Fungi
O'Halloran et al. (2006)	yes	*Caenorhabditis elegans*
Delsuc et al. (2006)	no	Chordates
Wang & Lavrov (2006)	no	Homoscleromorpha
Martin et al. (2005)	yes	Land plants
Philip et al. (2005)	no	Eukaryotes
Hrdy et al. (2004)	yes	Hydrogenosomes
Embley et al. (2003a)	yes	Hydrogenosomes
Embley et al. (2003b)	yes	Hydrogenosomes
Davidson et al. (2002)	yes	Hydrogenosomes

Table 1. Publications that use 6-state amino acid recoding.

Dayhoff recoding	Binning scheme
9-state	DEHNQ ILMV FY AST KR G P C W
12-state	DEQ MLIV FY KHR G A P S T N W C
15-state	DEQ ML IV FY G A P S T N K H R W C
18-state	ML FY I V G A P S T D E Q N H K R W C

Table 2. Best scoring binning schemes optimized on the Dayhoff matrix.