Syst. Biol. 70(5):1046–1060, 2021

DOI:10.1093/sysbio/syab010
Advance Access publication February 22, 2021

QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution

BUI QUANG MINH[1,2], CUONG CAO DANG[3], LE SY VINH[3,*], AND ROBERT LANFEAR[2,*]

[1]*School of Computing, Australian National University, 145 Science Road, Acton, ACT 2601, Canberra, Australia;* [2]*Department of Ecology and Evolution, Research School of Biology, Australian National University, 145 Science Road, Acton, ACT 2601, Canberra, Australia;* [3]*Faculty of Information Technology, University of Engineering and Technology, Vietnam National University, Hanoi, 144 Xuan Thuy, Cau Giay, 10000 Hanoi, Vietnam*

Bui Quang Minh and Cuong Cao Dang contributed equally to this article.

**Correspondence to be sent to: University of Engineering and Technology, Vietnam National University, Hanoi, 144 Xuan Thuy, Cau Giay, 10000 Hanoi, Vietnam;*
E-mail: vinhls@vnu.edu.vn and
Department of Ecology and Evolution, Research School of Biology, Australian National University, 145 Science Road, Acton, ACT 2601, Canberra, Australia;
E-mail: rob.lanfear@anu.edu.au.

Received 24 August 2020; reviews returned 25 December 2020; accepted 10 February 2021
Associate Editor: Bryan Carstens

Abstract.—Amino acid substitution models play a crucial role in phylogenetic analyses. Maximum likelihood (ML) methods have been proposed to estimate amino acid substitution models; however, they are typically complicated and slow. In this article, we propose QMaker, a new ML method to estimate a general time-reversible Q matrix from a large protein data set consisting of multiple sequence alignments. QMaker combines an efficient ML tree search algorithm, a model selection for handling the model heterogeneity among alignments, and the consideration of rate mixture models among sites. We provide QMaker as a user-friendly function in the IQ-TREE software package (http://www.iqtree.org) supporting the use of multiple CPU cores so that biologists can easily estimate amino acid substitution models from their own protein alignments. We used QMaker to estimate new empirical general amino acid substitution models from the current Pfam database as well as five clade-specific models for mammals, birds, insects, yeasts, and plants. Our results show that the new models considerably improve the fit between model and data and in some cases influence the inference of phylogenetic tree topologies.[Amino acid replacement matrices; amino acid substitution models; maximum likelihood estimation; phylogenetic inferences.]

Amino acid substitution models are crucial for model-based phylogenetic analyses of protein sequences, including for maximum likelihood (ML) and Bayesian inference approaches. The most commonly used protein models are Markov processes summarized in a 20-by-20 replacement matrix, denoted as Q, which describes the rates of substitutions between pairs of amino acids. Because they have so many parameters, Q matrices are computationally very expensive to estimate. Together with the fear of model overfitting, they are not usually estimated during a phylogenetic analysis of a single amino-acid multiple sequence alignment (MSA). Instead, the best Q matrix for each locus in a multilocus MSA is usually selected from a set of pre-estimated Q matrices using model selection software such as ModelFinder (Kalyaanamoorthy et al. 2017), ModelTest (Darriba et al. 2019), or PartitionFinder (Lanfear et al. 2017). Estimating Q matrices from large collections of empirical MSAs, where one derives the so-called *empirical* Q matrix that jointly explains substitution patterns across all MSAs, remains challenging both because the task is computationally expensive, and because there is no user-friendly software implementation that facilitates the task. As a result, the publication of new empirical Q matrices remains infrequent, and empirical phylogeneticists rarely estimate their own Q matrix even in those cases where they have sufficient data.

The first empirical Q matrices, Dayhoff (Dayhoff et al. 1978) and JTT (Jones et al. 1992), were estimated using the Maximum Parsimony (MP) principle. This approach was based on counting the minimum number of amino-acid changes along a phylogeny required to explain the MSA. The MP approach was known for underestimation of the true number of multiple amino-acid substitutions on single branches of the tree (Whelan and Goldman 2001). Although Dayhoff et al. (1978) and Jones et al. (1992) employed some adjustments to reduce this limitation, the problem remained to some extent. The advent of ML methods led to further improvements in addressing this issue. In ML, one estimates the Q matrix that maximizes the joint likelihood of observing a large collection of MSAs given independently estimated tree topologies for each MSA. The most widely used Q matrices, WAG (Whelan and Goldman 2001) and LG (Le and Gascuel 2008), were estimated using the ML approach. These matrices substantially improved model fit on a range of MSAs compared with the older matrices. However, the methods used to estimate the LG and WAG matrices used several approximations to make the analyses computationally feasible. For example, Whelan and Goldman (2001) ignored rate heterogeneity across sites (RHAS), although this phenomenon is widely observed in empirical MSAs. Le and Gascuel (2008) later improved this method by incorporating RHAS with a discrete Gamma distribution (Yang 1994) but using a site-rate partition model instead of the originally designed mixture model. Moreover, the Pfam database (Bateman et al. 2002) used to estimate LG has now increased 8-fold (El-Gebali et al. 2019). As the most widely used Q matrices were estimated more than a decade ago, improvements in the available data and phylogenetic inference methods suggest that it might be possible to estimate improved Q matrices.

Recent attention has shifted towards mixture models (Le et al. 2008; Wang et al. 2008; Le and Gascuel 2010;

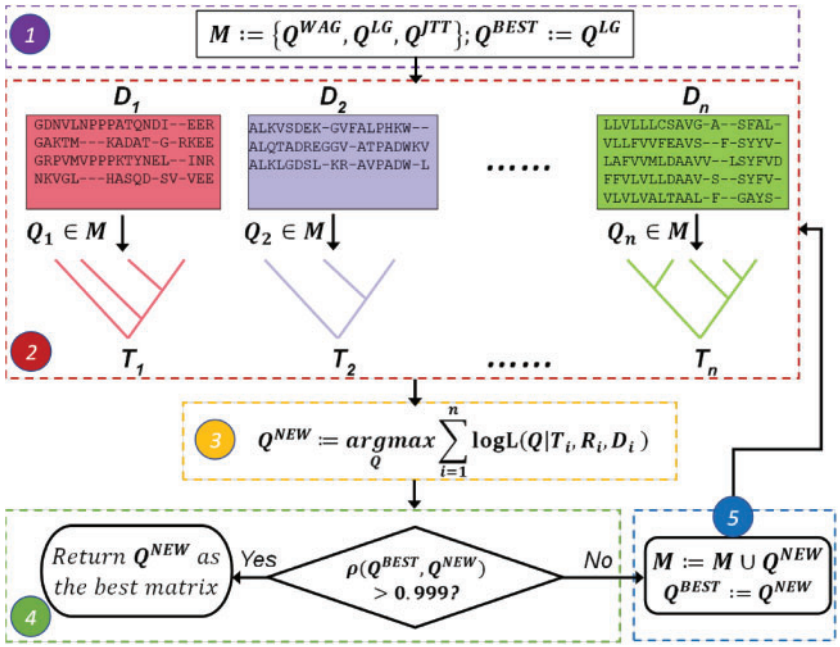

FIGURE 1. Schematic overview of QMaker consisting of five main steps: 1) initialize the current best replacement matrix Q^{BEST} as LG and the set of candidate matrices as WAG, LG, and JTT; 2) for each alignment D_i from the total n alignments, find the best-fit matrix Q_i (chosen from the list of candidate matrices), the rate heterogeneity across sites model R_i, the best-fit tree topology T_i and the estimated branch lengths Λ_i (not shown in the figure); 3) maximize the joint log-likelihood to obtain a new matrix Q^{NEW} that best explains all D_i; 4) if the Pearson correlation between Q^{BEST} and Q^{NEW} is higher than 0.999, return Q^{NEW} as the best matrix for the data; and 5) otherwise, replace Q^{BEST} by Q^{NEW}, extend the set of candidate matrices with Q^{NEW} and go back to Step 2.

Le et al. 2012). In standard (i.e., nonmixture) models, different Q matrices may be applied to different sites (often called 'partitioning'), but the likelihood of each site in the alignment is only ever calculated under a single Q matrix. In contrast, with a mixture model, the likelihood of every site in the alignment is calculated under a preassigned range of different Q matrices (the mixture), and these likelihoods are combined into a joint likelihood of the site under this mixture. However, despite their attractive properties, mixture models are still employed far more rarely in empirical analyses than their single Q matrix counterparts, probably because only a limited number of phylogenetic software tools support mixture models as standard, and because mixture models remain computationally more intensive to use than single-matrix models. For example, the papers that introduced the three most popular single-matrix models (LG, WAG, and JTT) were cited a total of 916 times in 2019, while the papers that introduced the three most popular mixture models (LG4X, CAT60, and EX) were cited a total of 39 times in the same year according to Google Scholar. In this study, we focus on improving the methods used to estimate single Q matrix models. We then estimate a suite of new single Q matrices, which can be combined together to form mixture models using the IQ-TREE software. Our methodological improvements for estimating single-matrix models lay the groundwork for future improvements in the estimation of explicit mixture models such as the LG4X model.

Here, we present QMaker, an ML method and software implementation which estimates an empirical Q matrix for any set of protein MSAs. Figure 1 shows a schematic overview of the QMaker workflow (see Materials and Methods for full details). QMaker improves upon previously published ML procedures on a number of fronts (Table 1). These include the use of the efficient ML tree search algorithm of IQ-TREE (Nguyen et al. 2015), consideration of a distribution-free model of RHAS (Kalyaanamoorthy et al. 2017), full usage of the rate mixture model, support for multiple CPU cores, and an explicit separation of training and testing data. Furthermore, we provide an easy-to-use implementation of QMaker as part of the IQ-TREE software package (http://www.iqtree.org). We employed our new software to estimate and compare six new amino acid replacement matrices: one general matrix based on version 31 of the Pfam database, and five clade-specific matrices for mammals, birds, insects, yeasts, and plants. We show that the new matrices not only fit better to the data but also influence the topologies of inferred trees.

MATERIALS AND METHODS

Data Sets Used for Training and Testing

We downloaded a total of 16,712 Pfam MSAs from version 31 of the database (El-Gebali et al. 2019), removed

TABLE 1. Feature comparisons between QMaker and two previously published estimation procedures (Whelan and Goldman 2001; Le and Gascuel 2008)

Feature	Whelan and Goldman	Le and Gascuel	QMaker
Tree reconstruction	Neighbor-joining (Saitou and Nei 1987) with Dayhoff+F distances	PhyML (Guindon and Gascuel 2003) with WAG+Γ4 model	IQ-TREE (Nguyen et al. 2015) with best-fit model
Branch length estimation	Scaled on JTT+F	ML estimate	ML estimate
Rate heterogeneity across sites	No	Gamma model (Yang 1994)	Gamma (Yang 1994), invariant site (Gu et al. 1995), and free rate model (Kalyaanamoorthy et al. 2017)
Algorithm to optimize the ML parameters	Expectation maximization	Expectation maximization	Broyden–Fletcher–Goldfarb–Shanno (Fletcher 1987)
Multicore CPU support	No	No	Yes
Explicit separation of training and testing data	No	No	Yes

TABLE 2. Summary of the data sets used to estimate new amino-acid replacement matrices

Data set	References	Seqs	Sites	Loci	Training	Testing
Pfam	El-Gebali et al. (2019)	1,150,099	3,433,343	13,308	6654	6654
Plant	Ran et al. (2018)	38	432,014	1308	1000	308
Bird	Jarvis et al. (2015)	52	4,519,041	8295	1000 × 2	6295
Mammal	Wu et al. (2018)	90	3,050,199	5162	1000 × 2	3162
Insect	Misof et al. (2014)	144	595,033	2868	1000	1868
Yeast	Shen et al. (2018)	343	1,162,805	2408	1000 100 seqs	1408

For each data set, we randomly subsampled half (Pfam) or 1000 MSAs (others) as the training set and remaining loci as the test set. For bird and plant data sets, we used two nonoverlapping training sets to examine the effect of random subsampling. For the yeast data set, we additionally subsampled 100 sequences from the training set due to the excessive computational burden.

identical sequences from each MSA and only retained MSAs having between 5 and 1000 sequences and at least 50 sites. This leaves us with 13,308 remaining MSAs, denoted as the Pfam data set. We randomly divided the MSAs into two groups and used one half as the training set and the other half as the test set. Moreover, we downloaded five data sets for plant (Ran et al. 2018), bird (Jarvis et al. 2015), mammal (Wu et al. 2018), insect (Misof et al. 2014), and yeast (Shen et al. 2018). For each of these data sets, we randomly selected 1000 loci as the training set and used the remaining loci as the test set. For the bird and mammal data sets, we used two nonoverlapping training sets to examine the effect of random test-set selection. For the yeast training set, we additionally subsampled to 100 taxa from the alignment due to the excessive computational burden of estimating the Q matrix from the alignment containing all of the taxa in this data set. The training set was used to estimate Q and while the test set is used to compare the model fit between the estimated Q. Details of the data sets are summarized in Table 2. All data are available from the supplementary material available at https://doi.org/10.6084/m9.figshare.9768101.

Model of Amino Acid Substitutions

The model of amino acid substitutions follows the continuous Markov process that is stationary, reversible, and homogeneous. This process is summarized in a 20-by-20 rate matrix Q, that describes the rate of change from one amino acid to another per time unit. Because of the reversibility assumption, entries of Q can be written as the product of the symmetric exchangeability rates $\left(r_{ij}\right)$ and the amino-acid frequencies $\left(\pi_i\right)$:

$$Q=\begin{pmatrix} - & r_{1,2}\pi_2 & r_{1,3}\pi_3 & \dots & r_{1,20}\pi_{20} \\ r_{1,2}\pi_1 & - & r_{2,3}\pi_3 & \dots & r_{2,20}\pi_{20} \\ r_{1,3}\pi_1 & r_{2,3}\pi_2 & - & \dots & r_{3,20}\pi_{20} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{1,20}\pi_1 & r_{2,20}\pi_2 & r_{3,20}\pi_3 & \dots & - \end{pmatrix},$$

where the diagonal entries of Q are chosen such that its row sums equal 0. Before Q is used for tree inference, it is divided by the normalization factor μ (IQ-TREE will do this automatically):

$$\mu=-\sum_{i=1}^{20}\pi_i q_{i,i},$$

where $q_{i,i}$ is the diagonal entries of Q, so that the total number of substitutions is 1. Because of this constraint and also $\sum_{i=1}^{20}\pi_i=1$, we have 189 free exchangeability parameters r_{ij} and 19 free amino-acid frequency parameters, which can only be reliably estimated from a large amount of data, such as the BRKALN (Jones D., unpublished data) and Pfam (El-Gebali et al. 2019) alignment databases. Quite often the amino acid frequencies can be empirically estimated from the data set at hand, denoted by "+F." For example, the WAG+F model uses the exchangeability rates defined by WAG but amino acid frequencies from the current MSA.

TABLE 3. Existing amino-acid replacement matrices

Matrix	Reference	Genomic regions
Blosum62	Henikoff and Henikoff (1992)	General
Dayhoff	Dayhoff et al. (1978)	General
JTT	Jones et al. (1992)	General
LG	Le and Gascuel (2008)	General
PMB	Veerassamy et al. (2003)	General
VT	Muller and Vingron (2000)	General
WAG	Whelan and Goldman (2001)	General
mtArt	Abascal et al. (2007)	Mitochondrial
mtMam	Yang et al. (1998)	Mitochondrial
mtRev	Adachi and Hasegawa (1996)	Mitochondrial
mtZoa	Rota-Stabelli et al. (2009)	Mitochondrial
mtMet	Le et al. (2017)	Mitochondrial
mtVer	Le et al. (2017)	Mitochondrial
mtInv	Le et al. (2017)	Mitochondrial
cpRev	Adachi et al. (2000)	Chloroplast
FLU	Cuong et al. (2010)	Viral
HIVb	Nickle et al. (2007)	Viral
HIVw	Nickle et al. (2007)	Viral
rtREV	Dimmic et al. (2002)	Viral

Model of Rate Heterogeneity across Sites

It is well known that MSA sites may have evolved at different rates. The so-called rate heterogeneity across sites (RHAS) has been typically modeled by a discrete Gamma distribution (Yang 1994) w/o a proportion of invariable sites (Gu et al. 1995). For example, LG+I+Γ means that while all sites follow the LG exchangeability matrix, a fraction of sites is invariable (i.e., with zero evolutionary rates due to e.g. selective pressure) and the rates of the remaining variable sites follow a Gamma distribution.

Recently, it has been shown that a distribution-free rate model frequently provides a better fit than the Gamma model (Kalyaanamoorthy et al. 2017). The distribution-free rate model allows several site-rate categories, where the rates and proportions of each category are independent from one another and are estimated from the data. Hence, we do not assume any prior distribution of rates across sites.

Estimating a Joint Replacement Matrix from a Protein Data Set

We now introduce a new method to estimate a replacement matrix Q from a database of protein MSAs $D=\{D_1,\ldots,D_n\}$. Here, we want to find a single Q that best explains, in terms of ML, the pattern of amino acid substitutions for all MSAs. We denote by $M=\left\{Q^{\mathrm{WAG}},Q^{\mathrm{LG}},\ldots\right\}$ the set of candidate replacement matrices (Table 3).

We first determine, for each D_i, the best-fit matrix $Q_i\in M$, the best RHAS model R_i (e.g., I+Γ and R5) using ModelFinder (Kalyaanamoorthy et al. 2017), and the ML tree T_i with the set of branch lengths Λ_i using IQ-TREE (Nguyen et al. 2015). Next, we fix R_i,T_i and Λ_i to estimate the Q that maximizes the total log-likelihood across all MSAs in D:

$$\ell(Q)=\sum_{i=1}^{n}\log L\left(Q|T_i,\Lambda_i,R_i,D_i\right). \quad (1)$$

To estimate the 208 parameters of Q (189 exchangeabilities r_{ij} plus 19 frequencies π_i), we use the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, an iterative, hill-climbing method, specifically for solving unconstrained nonlinear optimization problem (Fletcher 1987). This optimization results in an estimate of a new rate matrix denoted as Q^{NEW}. Some MSAs may now show Q^{NEW} as the better-fit model. Therefore, we extend the set of candidate rate matrices by Q^{NEW} and repeat the procedure above to re-estimate Q_i,R_i,T_i,Λ_i

The overall workflow of QMaker is as follows (Fig. 1):

1. Initialize the set of candidate replacement matrices as $M:=\left\{Q^{\mathrm{WAG}},Q^{\mathrm{LG}},Q^{\mathrm{JTT}}\right\}$ and the current best matrix as $Q^{\mathrm{BEST}}:=Q^{\mathrm{LG}}$. Here, we use three commonly used matrices to initialize M, but in principle M can be initialized with any set of matrices.

2. For each MSA D_i, find the best-fit matrix $Q_i\in M$, rate heterogeneity across sites model R_i and estimate the ML tree T_i with branch lengths Λ_i based on Q_i and R_i (note that an edge-linked model sharing the topology is used for the five clade-specific data sets estimated below). Step 2 is necessary to obtain the initial topology and branch lengths for every locus, with which the new Q matrix are then estimated in Step 3.

3. Given R_i and T_i, jointly estimate Q and Λ_i that maximizes the log-likelihood function (1), resulting in a new replacement matrix Q^{NEW}. Specifically,

   (a) estimate Q given R_i,T_i, and Λ_i

   (b) estimate Λ_i given R_i,T_i, and Q

   (c) if the log-likelihood of Q and Λ_i is increased more than 0.1 (compared to the log-likelihood of previous Q and Λ_i), go to Step 3b, otherwise, go to Step 4.

4. Following Le and Gascuel, let ρ be the Pearson correlation coefficient between Q^{BEST} and Q^{NEW}. If $\rho>0.999$, report Q^{NEW} as the best replacement matrix for the database D and stop. Otherwise, go to Step 5.

5. Assign $Q^{\mathrm{BEST}}:=Q^{\mathrm{NEW}}$, add Q^{NEW} to the set of candidate matrices: $M:=M\cup Q^{\mathrm{NEW}}$, and go back to Step 2.

Comparisons with Previous Estimation Procedures

Compared with the ML procedures used to estimate WAG (Whelan and Goldman 2001) and LG (Le and Gascuel 2008), QMaker has a number of differences (Table 1). Among others, Whelan and Goldman (2001) omitted rate heterogeneity across sites and employed neighbor-joining trees for computational efficiency. Le and Gascuel (2008) improved this method by incorporating the Γ model of rate heterogeneity and inferring the ML tree with PhyML (Guindon and Gascuel 2003). However, they did not use the original Γ as a mixture model when estimating Q. Rather, they partitioned the sites in each D_i into their most likely rate category resulting in four sub-MSAs for each D_i, and essentially applied the method of Whelan and Goldman to derive the Q matrix from the expanded D.

Here, QMaker improves both methods by i) additionally considering the free rate and invariant site mixture models; ii) inferring the ML tree with IQ-TREE, which has been shown to outperform PhyML (Guindon and Gascuel 2003) and other software (Zhou et al. 2018); and iii) directly optimizing the log-likelihood function (1) to obtain Q instead of aforementioned approximations.

We note that because log-likelihoods are summed across sites, larger genes may have more influence on the model parameters than smaller genes. However, given that Q matrices are typically estimated from large data sets comprising a large number of genes, it is unlikely that any single gene will have an undue influence on the model. Regardless, summing likelihood across sites ensures that every site in the analysis is treated with a weight proportional to the amount of information it contains.

Software Implementation

We provided an implementation of QMaker as part of the IQ-TREE software version 2.0-rc1. The entire training stage for the Pfam data set can be accomplished with just two command lines. The first one is

```
iqtree -S ALN_DIR -nt 48
```

to find the best-fit models and ML trees for all MSAs residing in the folder `ALN_DIR`; `-nt` option is to specify the number of CPU cores. Note that for this study, due to the excessive size of the Pfam training set, we additionally used two options: `-mset LG,WAG,JTT` to consider only these three models and `-cmax 4` to restrict up to four categories for the rate heterogeneity across sites model. The second command line is

```
iqtree -S ALN_DIR.best_model.nex -nt
48 -te ALN_DIR.treefile --model-joint
GTR20+FO
```

to perform Step 3 of estimating the replacement matrix (GTR20 for general time reversible model with 20-state data), given the best models (`ALN_DIR.best_model.nex`) and best trees (`ALN_DIR.treefile`) found above.

For the five clade-specific data sets, to count best-fit models we performed an edge-linked partition model, which assumes a single tree topology and rescales edge lengths across the loci. This model was shown to best balance between schemes of model parameterization (Duchene et al. 2019). For this purpose, the `-S` option is changed to `-p`.

To test the model fit of the trained Q matrices we ran ModelFinder (Kalyaanamoorthy et al. 2017) as implemented in IQ-TREE:

```
iqtree -S TEST_DIR -m MF -mset
JTT,WAG,LG,Q.pfam,
Q.plant,Q.bird,Q.mammal,Q.insect,Q.yeast,
```

where `TEST_DIR` is a directory containing the testing MSAs.

Results

QMaker Outperforms Existing Estimation Methods

To establish whether the approach, we propose here improves upon previously suggested approaches, we compared QMaker (Fig. 1; Materials and Methods) to the method used to estimate the LG matrix (Le and Gascuel 2008) on the same training data. Because both methods use the same data to estimate a Q matrix, any differences between the matrices must be due solely to the estimation procedure.

To compare the two approaches, we first obtained the training set of 3412 Pfam MSAs originally used to estimate the LG matrix (http://www.atgc-montpellier.fr/models/index.php?model=lg) and then applied QMaker to estimate a new Q matrix from these data. We called the resulting matrix Q.LG to reflect the origin of the data set. QMaker took about 28 h wall-clock time using 36 CPU cores on a 2.3-GHz server to estimate Q.LG.

To compare the performance of the LG and Q.LG matrix, we asked how frequently each matrix was selected as the best-fit model for the 500 test MSAs originally used to test the LG matrix (Le and Gascuel 2008). To do this, we calculated the best-fit model for each MSA from a set of four candidate Q matrices comprised of the two comparator matrices LG and Q.LG, plus two other frequently used matrices: WAG (Whelan and Goldman 2001) and JTT (Jones et al. 1992). Q.LG was the most frequently selected matrix (246 MSAs), followed by LG (166 MSAs), WAG (55 MSAs), and JTT (33 MSAs). If we focus only on the relative fit of Q.LG and LG, we find that Q.LG is the best-fit model for 281 MSAs (56%) while LG is the best-fit model for 219 MSAs (44%), again confirming that the Q.LG model slightly outperforms the LG model on these data. Among the 246 MSAs that Q.LG is the best, 84 showed that Q.LG is significantly better the other three matrices (LG, WAG, and JTT) according to the approximately unbiased (AU, $P < 0.05$) test (Shimodaira 2002). The AU test is appropriate here because all of the models being compared have the same number of parameters. The AU compares the site

likelihoods calculated under each model to ask whether the likelihoods from the better model significantly better than those from the worse model. The AU test was done using *consel* package (Shimodaira and Hasegawa 2001; Shimodaira 2002).

To better understand which of the differences between the method used to estimate the LG matrix and QMaker most contributed to QMaker improved performance, we benchmarked three key differences between the two methods: i) tree reconstruction, ii) models of RHAS, and iii) parameter optimization technique (Table 1). To evaluate the contribution of each of these improvements, we replace each improvement with the approach previously used in the LG procedure. For example, to evaluate our improved approach to tree reconstruction, we changed the tree reconstruction method to use PhyML (Guindon and Gascuel 2003) instead of IQ-TREE while keeping the rest of the pipeline unchanged. This resulted in three new Q matrices, which we name for the component that they were estimated to benchmark: i) Q_{tree}, ii) Q_{RHAS}, and iii) Q_{EM}. On the 500 test MSAs, Q.LG is better than Q_{tree}, Q_{RHAS}, and Q_{EM} for 343 (69%), 331 (66%), and 294 (59%) MSAs, respectively. This reveals that the improvements in tree reconstruction and the modeling of rate heterogeneity across sites led to the largest improvements for QMaker, and that changing the optimization technique had the smallest but still nonnegligible influence.

These results demonstrate that the QMaker method substantially improves the fit between models and data compared to previous estimation procedures. For the sake of reproducibility, we provide the Q.LG matrix in the Supplementary material. We do not, however, intend for the Q.LG matrix to be widely used, as it is estimated from a now-outdated version of the Pfam database.

Larger Amino Acid Databases Improve Model Fit, but Primarily to Target Alignments

Given these improvements we applied QMaker to estimate a new amino-acid substitution matrix from the latest version of the Pfam database, and call the resulting matrix Q.pfam. We estimated Q.pfam from a training set of half of the MSAs (6654 MSAs in total) available in Pfam database version 31 (El-Gebali et al. 2019), reserving the remaining half of the database as a test set with which to compare Q.pfam to three previously estimated matrices (LG, WAG, and JTT).

Q.pfam outperformed other matrices on the test MSAs. Q.pfam was the most frequently selected matrix in 40.7% of the test MSAs, followed by LG (35.5%), JTT (14.7%), and WAG (9.1%).

We further tested the new matrix on a collection 13,041 single-locus MSAs from five recently published phylogenomic data sets (Table 2). To do this, we compared the fit of the same four models (Q.pfam, LG, WAG, and JTT) to each of the 13,041 empirical MSAs. Surprisingly, the most commonly selected matrix across all 13,041 MSAs was JTT (74.9%), followed by LG (5.9%), Q.pfam (3.0%), and WAG (2.0%). The JTT matrix was the most commonly selected matrix for three out of the five data sets (birds, plants, and mammals; Appendix Fig. 1), and the LG matrix was the most commonly selected matrix for the remaining two data sets (insects and yeasts; Appendix Fig. 1). This shows that amino acid models estimated from the Pfam database (Q.pfam, LG) often fail to provide the best fit to MSAs used for phylogenomic inference on commonly studied clades.

Five New Clade-Specific Q Matrices Improve Model Fit on Phylogenomic Data

The surprisingly poor fit of Pfam-based matrices (Q.pfam, LG) to the empirical MSAs of birds, mammals and plants, combined with the high variation in the identity of the best model in each data set, suggests that there may be substantial between-clade variation in the way that proteins used for phylogenetic inference evolve. If this is the case, then accounting for this by estimating independent Q matrices for each clade should improve model fit. To test this, we estimated a clade-specific Q matrix for each of the five phylogenomic data sets of nuclear sequences (Table 2): Q.plant, Q.bird, Q.mammal, Q.insect, and Q.yeast. For each data set, we used 1000 training MSAs to estimate the Q matrix, and the remaining MSAs from each data set as test sets (see Materials and Methods for more details). The time taken to estimate these matrices depended on the size of the loci and the number of taxa in the data set, but using 15 CPU cores on a 2.3 GHz server the times ranged from 68 h for the plant data set to 385 h for the insect data set.

Figure 2 shows the frequency with which each of the six new (Q.pfam, Q.plant, Q.bird, Q.mammal, Q.insect, and Q.yeast) and three existing (JTT, WAG, and LG) matrices were selected as the best-fit for the six test sets. As expected, the best fit Q matrix for each test set was the Q matrix estimated from the corresponding training set, although the strength of the association varied widely among test data sets. For example, Q.plant was the best model for 92.2% of plant test MSAs, with the next best model selected for fewer than 5% of test MSAs. But Q.pfam was only selected as the best model for 34.1% of the Pfam test MSAs, with the next best model (LG) selected for 24.2% of MSAs. These results are likely driven in part by the fact that the set of models we considered included many models that are similar to Q.pfam (e.g., LG and WAG), but few that are similar to Q.plant.

We then applied the approximately unbiased (AU, $P < 0.05$) test (Shimodaira 2002) to count how frequently each clade-specific model is significantly better than Q.pfam on the corresponding test set. The results (Table 4) show that in the overwhelming majority of cases, the clade-specific model fits the data significantly better than the Q.pfam model. Remarkably, this is even the case for the Q.yeast matrix, which is highly similar to the Q.pfam matrix (Pearson correlation $= 0.982$) but has 135/190 exchangeabilities smaller than those of Q.pfam (Appendix Fig. 2). This reveals that even small differences in the Q matrix can lead to statistically significant differences in model fit.

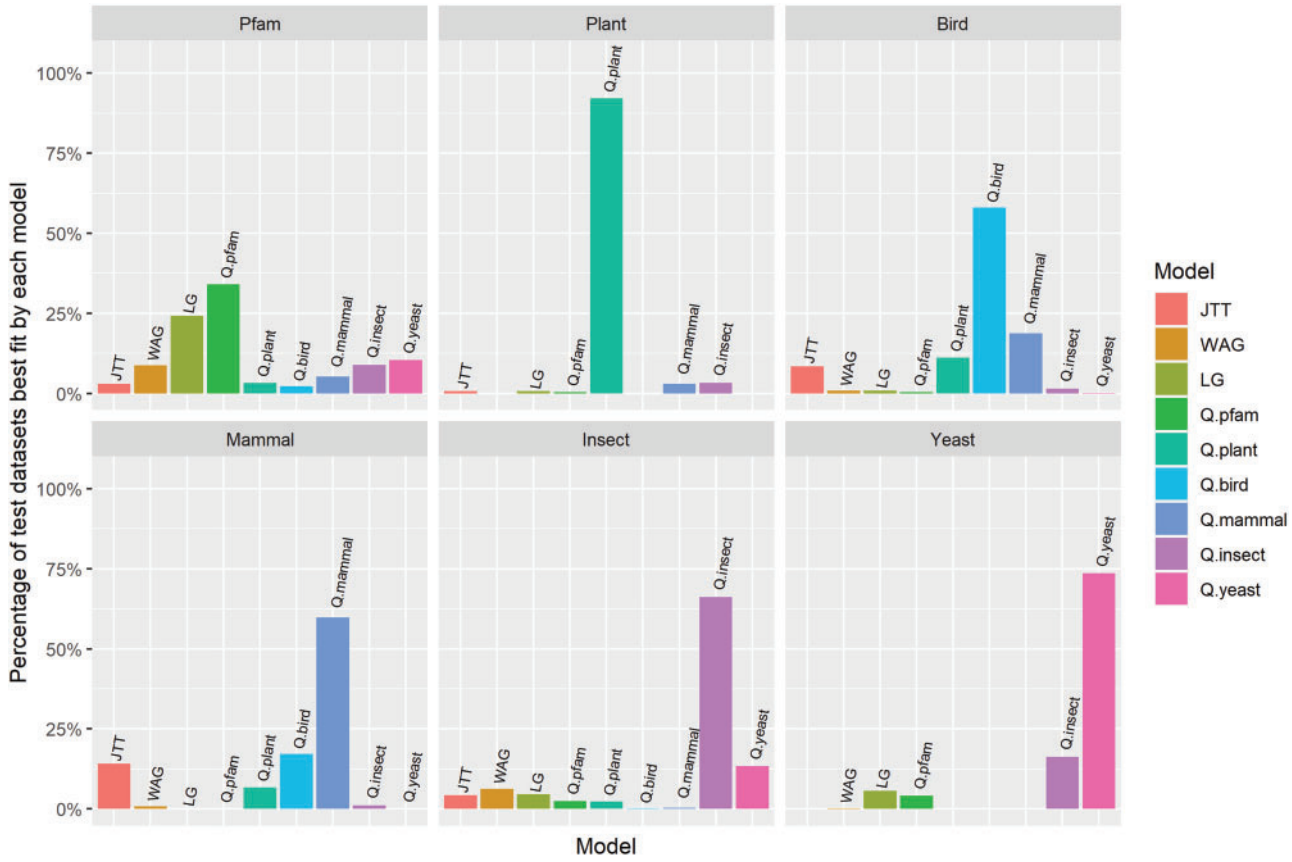

FIGURE 2. Frequency of best fitting for nine amino-acid replacement matrices on six test data sets: Pfam, Plant (Ran et al. 2018), Bird (Jarvis et al. 2015), Mammal (Wu et al. 2018), Insect (Misof et al. 2014), and Yeast (Shen et al. 2018).

TABLE 4. Likelihood comparisons between Q.pfam, Q.plant, Q.bird, Q.mammal, Q.insect, and Q.yeast

Compared model	Q.pfam is better	Compared model is better
Q.plant	4 (1)	304 (288)
Q.bird	468 (110)	5827 (5106)
Q.mammal	94 (24)	3068 (2872)
Q.insect	222 (62)	1646 (1291)
Q.yeast	72 (25)	1336 (1227)

Note: The second column shows the number of alignments having higher likelihood with Q.pfam than the comparing model and the number of alignments with significantly higher likelihood in parentheses (AU test, $P < 0.05$). The third column shows the number of alignments where the compared model has higher likelihood than Q.pfam, with the number of significant cases in parentheses.

Principle Components Analysis Reveal the Landscape of Amino Acid Models

We used principle components analysis (PCA) to compare the properties of the six new amino acid models presented here to 19 previously estimated models (Table 3). The PCA plot of the *Q* matrices (Fig. 3a) shows a clear separation between matrices inferred from the nuclear, mitochondrial, chloroplast and viral genomes and the clade-specific matrices, with the clade-specific matrices falling between the mitochondrial and viral matrices, and the two Pfam-based matrices (LG, Q.pfam) in close proximity. The PCA plot of the models' amino acid frequencies (Fig. 3b) reveals that most of the variation among frequency vectors comes from differences between and within the viral and mitochondrial models, with more limited separation between the clade-specific matrices (Q.bird, Q.plant, Q.insect, Q.mammal, and Q.yeast) and the general-purpose matrices (LG, WAG, JTT, and Q.pfam).

Figure 4 shows visual comparison of exchangeabilities between LG and six new models. We found that, as expected, Q.pfam and LG are highly correlated (Pearson correlation = 0.990) but many exchangeabilities of Q.pfam are larger than those of LG (137/190). In particular, all exchangeabilities involving Cysteine (C) or Proline (P) are larger for Q.pfam while many entries involving Tryptophan (W) are a little larger for LG. Q.pfam is also highly correlated to Q.yeast (Pearson correlation = 0.982). Supplementary Table S1 shows correlation values between six new matrices and 20 existing matrices.

Incorporating the New Matrices into Model Selection Changes Locus-Tree Inference

To examine whether the six new matrices, we propose here affect the inference of phylogenetic trees, we asked how often the new matrices affected the phylogenetic tree when they were selected as the best model. For each single-locus MSA in each data set, if one of the

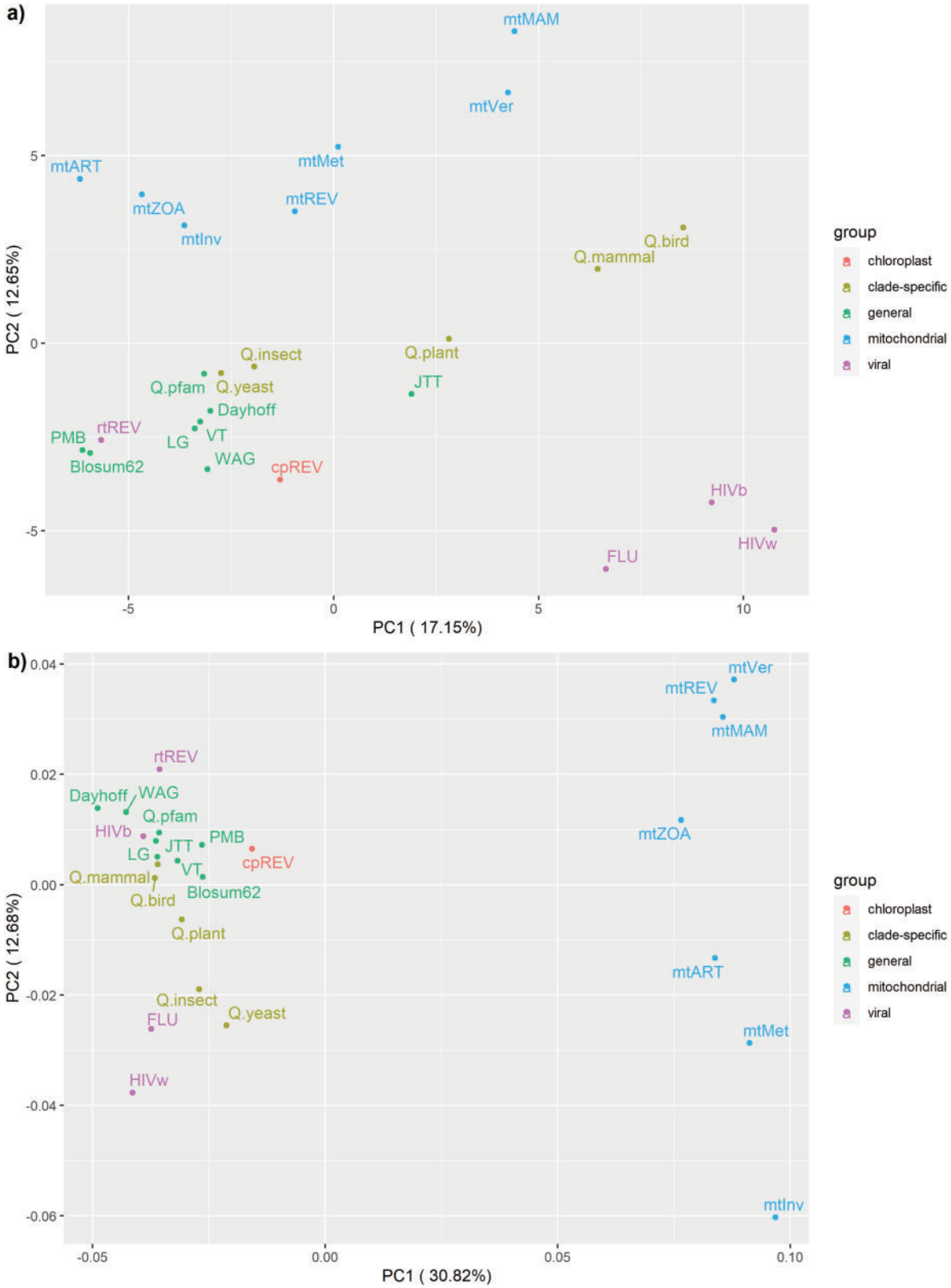

FIGURE 3. Principle component analysis (PCA) of all matrices with respect to a) the amino acid exchangeabilities and b) the amino acid frequencies.

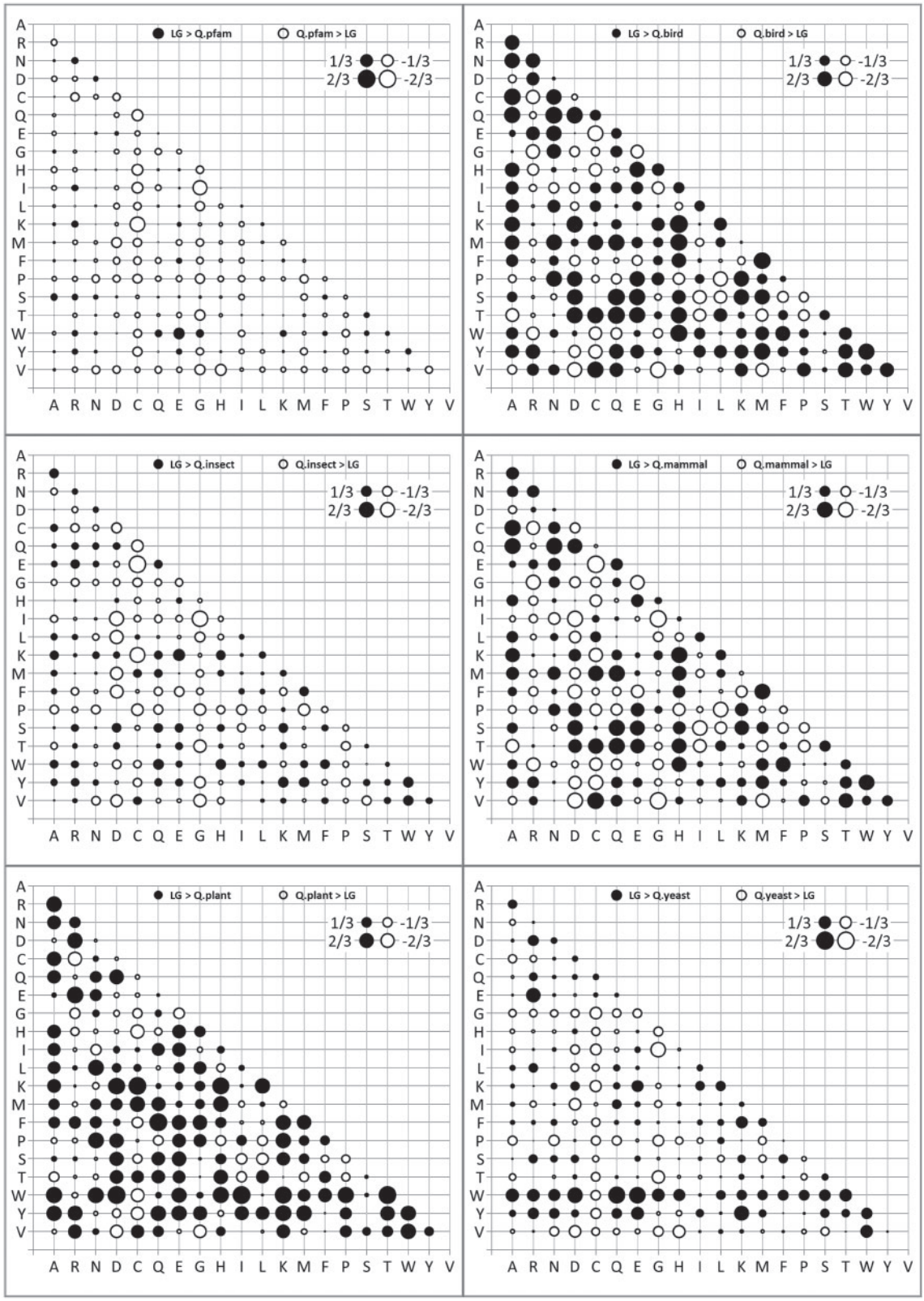

FIGURE 4. The bubble plots show relative differences between amino acid exchangeability coefficient in LG and those in six new models (Q.pfam, Q.plant, Q.bird, Q.mammal, Q.insect, and Q.yeast). Each bubble represents the value of $(r_1^{ij} - r_2^{ij})/(r_1^{ij} + r_2^{ij})$, where $r_1^{ij}(r_2^{ij})$ is the exchangeability in Q_1 (Q_2). Values 1/3 and 2/3 mean that the Q_1 exchangeability is 2 and 5 times as large as that of Q_2, respectively. Values -1/3 and -2/3 mean that Q_2 is 2 and 5 times larger than Q_1, respectively. The explanations are the same for all six subfigures, where Q_1 always is LG while Q_2 is Q.pfam, Q.plant, Q.bird, Q.mammal, Q.insect, or Q.yeast.

new models was selected as best model, we inferred the ML tree using the new model which we denoted T_{new}. We then compared this tree to the tree inferred for the same MSA using the best-fit model among JTT, WAG, and LG, which we denoted T_{old}. Differences between T_{new} and T_{old} could come from two sources: the effects of using a different amino acid substitution model or the stochasticity in tree search. The stochasticity in tree search arises because each gene is short, hence the phylogenetic signal islow, therefore the estimate of the ML tree can be somewhat unstable such that two different runs may produce different trees. To decouple these two factors, we performed another independent tree search to infer T_{old2} in the same way as we inferred T_{old} but using a different random-number seed. If T_{old} is different from T_{old2}, then the difference is merely due to tree search stochasticity. For each data set, we then compared the distribution of normalized Robinson–Foulds (nRF) (Robinson and Foulds 1981) distances between T_{new} and T_{old} to the distribution of the nRF distances between T_{old} and T_{old2} (Fig. 5, left-hand column). The extent to which nRF distances between T_{new} and T_{old} are larger than those between T_{old} and T_{old2} indicates the extent to which the new model affects tree inference, independently of stochasticity in the tree search. We used normalized RF distance, which is RF divided by 2*($n-3$), where n is the number of taxa, because this procedure allows us to compare tree distances between data sets with different numbers of taxa. nRF distances always scale between 0 and 1, regardless of the number of taxa. To ensure that this procedure works as we expect, we also made the same comparisons in cases where the best-fit model was *not* one of the new matrices we infer here. In this case, T_{new}, T_{old}, and T_{old2} are all inferred from the same model such that all differences between the trees are due to stochasticity in the tree search. As a result, we expect in this case that the distribution of nRF distances between T_{new} and T_{old} should be the same as the distribution between T_{old} and T_{old2} (Fig. 5, right-hand column).

Figure 5 confirms that the nRF distances between T_{new} and T_{old} are moderately higher than those between T_{old} and T_{old2}, indicating that using the newly proposed (and better fit) models changes locus tree topologies in every data set. In fact, the two distributions are significantly different for all data sets ($P < 0.001$ from a Kolmogorov–Smirnov test comparing the two distributions in each data set), indicating that the new models of evolution affect a nontrivial number of single-locus tree topologies in every data set. Of note, Figure 5 shows that in some data sets the topologies of a large number of loci differ considerably depending only on the random-number seed. For example, in both the Bird and Insect data sets, a considerable fraction of the topologies differ by more than half of the splits in the tree (nRF > 0.5). Manual inspection of a subset of these loci revealed that they tend to be very short and uninformative loci, such that many splits in the gene trees are supported by no substitutions and so resolved randomly depending on the random-number seed.

We also looked at the tree lengths of T_{new} and T_{old} and found that T_{new} is often longer than T_{old} (Table 5). For example, on the Plant test set, T_{new} is longer than T_{old} in 293/308 cases. The average length of T_{new} and T_{old} are 5.104 and 4.665, respectively, suggesting that the new matrices allow us to infer, on average, 9.4% more substitutions than the existing matrices. This increase in branch length results from the combination of the changes in the transition rates and the amino acid frequencies, because the new trees were estimated using both the new transition matrix and its associated empirical amino acid frequencies.

The cases we discuss here are those in which the information-theoretic approaches (here we use the BIC) suggest that the new models are a better fit to the data than the old models. Since, we should prefer the tree topologies and branch lengths estimated under the best model we have, to the extent that information-theoretic approaches to model selection in phylogenetics are accurate (Sullivan and Joyce 2005), our results suggest that the custom-matrices we infer with QMaker make meaningful improvements to both the topologies and branch lengths of a large number of inferred single-locus trees. These differences are likely to impact downstream analyses such as molecular dating and species-tree estimates. For example, species trees estimated with the Multi-Species Coalescent (MSC) model in software such as ASTRAL can be sensitive to relatively small changes in the gene trees used as input data (Sayyari et al. 2017). It is therefore plausible that using improved models of sequence evolution like those that we estimate in this study could have effects on the species trees and divergence dates estimated from amino acid data.

Discussion

In this study, we describe and implement QMaker, an easy-to-use tool to estimate an amino acid replacement matrix Q for any data set of one or more amino-acid alignments. Phylogenetic inference from amino acid alignments relies heavily on precomputed Q matrices. It is a little surprising, therefore, that new Q matrices are published relatively infrequently (e.g., Table 3), particularly in the age of phylogenomics when an increasing number of studies collect sufficient data to reliably estimate a Q matrix. We hope that the development of QMaker will democratize the inference of Q matrices, and that this will lead to concomitant improvements in phylogenetic inference and our understanding of molecular evolution. Similarly, we hope that the publication of five new Q matrices for some highly studied clades will improve phylogenetic inference for those clades.

The approach we implement in QMaker builds on previously described approaches (Whelan and Goldman 2001; Le and Gascuel 2008), and our analyses reveal that it improves on them in terms of model fit to the data. We applied QMaker to estimate one general-purpose Q matrix and five clade-specific Q matrices for mammals, plants, birds, insects, and yeasts. We showed that they

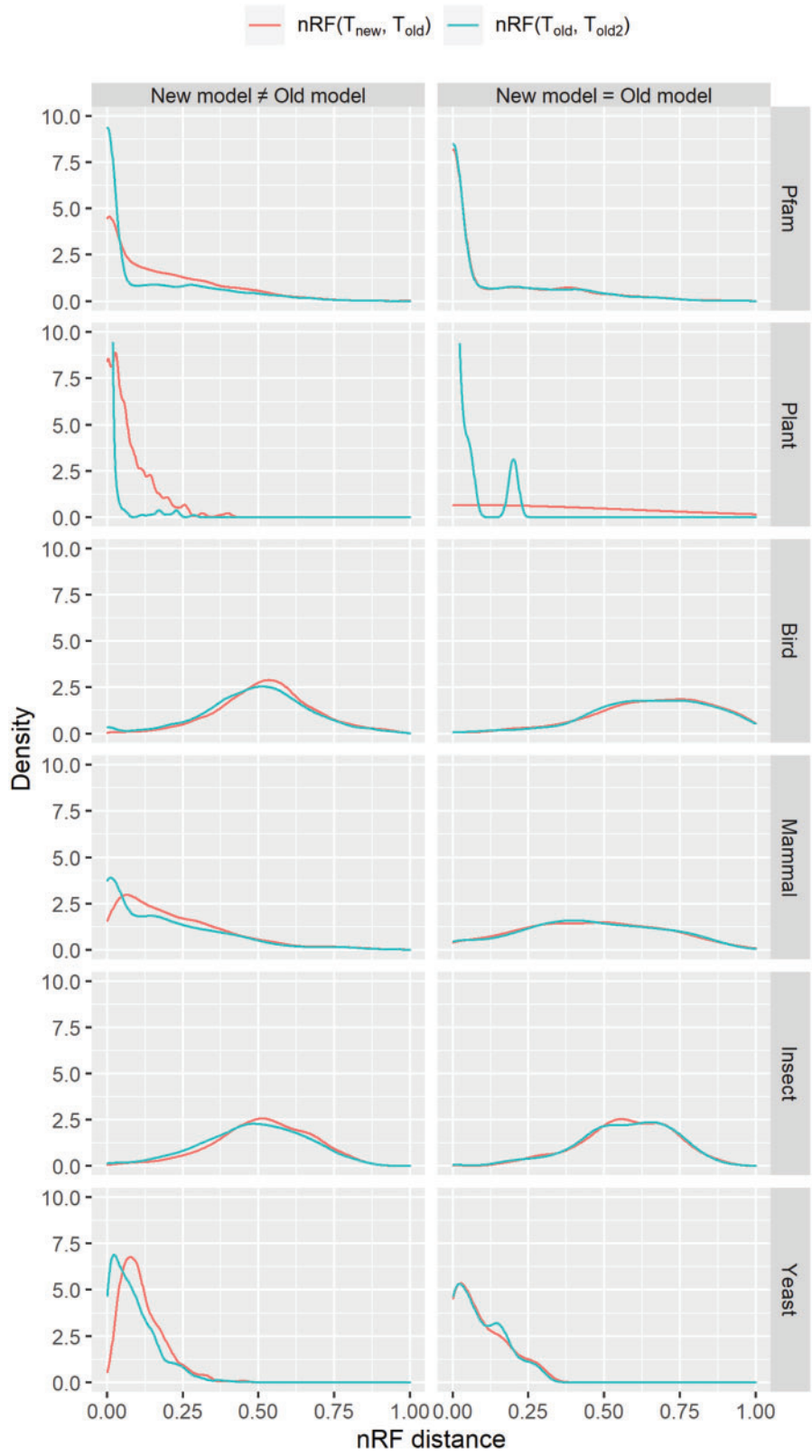

FIGURE 5. Distributions of normalized Robinson–Foulds (nRF) distances between the trees inferred by new and existing (JTT, WAG, or LG) models. The left-hand column indicates distributions where the best-fit model is one of the new models inferred in this study and shows that the new model has an effect on the tree (red distribution) that is larger than the effect of stochasticity in tree search alone (blue distribution). The distributions are much more similar to each other, as expected, if the best-fit model is one of the existing models (JTT, WAG, or LG; right-hand column of the figure).

TABLE 5. The average lengths of trees inferred using new Q matrices (T_{new}) and that using existing matrices (T_{old})

Data set	T_{new}	T_{old}	$T_{new} > T_{old}$
Plant	5.104	4.665	293/308
Bird	2.341	2.276	4698/6295
Mammal	6.091	5.786	2509/3162
Insect	39.873	38.291	1198/1868
Yeast	86.761	83.606	1057/1408

Note: The last column shows the number of cases when T_{new} is longer than T_{old}/ the total number of cases.

not only improve the fit between the model and the data but also influence the tree topologies. We note that inferring topologies for some specific organisms is not the main focus of this article. Practitioners are encouraged to use QMaker to infer new Q matrices and phylogenies for their own data of interest. All of the new matrices are now implemented in IQ-TREE version 2 (Minh et al. 2020) and incorporated as part of the model selection procedure, and the data necessary to implement all matrices in other phylogenetic software packages are provided in Supplementary material.

When estimating a new empirical model of amino acid substitution, it is important to consider how much data is required. There is no single answer to this question that can apply across data sets because different data sets will contain different amounts of information. Nevertheless, we can ask whether a Q Matrix estimated from a data set is useful. An amino acid replacement matrix requires the estimation of 208 parameters, representing the transition rates between pairs of amino acids and the frequencies of each amino acid in the data. Thus, a general rule of thumb is that the data should contain enough estimated substitutions of each type to reliably estimate relative transition rates. A pragmatic approach to establishing whether this is the case is to follow the approach we outline in the methods of this article. Namely, first split the loci in the data set into training and testing sets, then apply QMaker to the training set to obtain a new replacement matrix. Next estimate the AICc or BIC scores of the new matrix and the standard precomputed matrices using IQ-TREE. If the training data contain sufficient information to estimate a useful amino-acid replacement matrix, this will be revealed by the new matrix being the best-fit matrix for a substantial fraction of the loci in the testing set.

The relationships among the 19 existing Q matrices and the 7 new matrices we present here (Fig. 3) reveal a number of interesting patterns. As expected, there is a clear distinction between Q matrices estimated from different genomes, with the general-purpose matrices estimated from large data sets of protein alignments from the nuclear genome tending to cluster tightly together (Fig. 3). More surprising is the observation that the five new clade-specific Q matrices we estimate here tend to be quite distinct from all other Q matrices and are also remarkably distinct from one another. This result, combined with our observations that the clade-specific Q matrices tend to improve model fit and affect tree inference, highlight the potential benefits of inferring a clade-specific Q matrix before inferring a phylogeny. The differences among the clade-specific matrices also hint at potentially significant differences between the molecular evolutionary processes driving protein evolution in different clades of organisms.

Our results underscore previous recommendations that software for estimating phylogenetic trees should be run multiple times to help ensure that the estimate of the ML tree is as accurate as possible (Zhou et al. 2018). As part of our approach to comparing matrices, we ran IQ-TREE twice on the each of thousands of single-locus alignments, using different random-number seeds. We routinely observed differences in the phylogenies estimated from these replicate runs, as is expected for any heuristic search algorithm that is provided with limited information (in this case, single-locus alignments) with which to search for the optimal solution among an astronomical number of potential solutions.

The sometimes substantial variation in best-fit model for different loci from a single data set (Fig. 2) confirms that there can also be substantial variation in molecular evolution among loci. Thus, although QMaker allows researchers to infer a single Q matrix from a collection of alignments, it still seems sensible to infer phylogenies in a framework that allows for different Q matrices to be applied to different loci, such as by using a partitioned model (Lanfear et al. 2012; Chernomor et al. 2015) or mixture models. This can be achieved in the most commonly used phylogenetic inference software, including in IQ-TREE (Nguyen et al. 2015), RaxML (Stamatakis 2014), and PhyML (Guindon and Gascuel 2003).

The QMaker framework opens new avenues of research by simplifying the process of inferring a single Q matrix but is currently limited to estimating a single reversible Q matrix from one or more amino acid alignments. In principle, both of these limitations could be relaxed, for example by extending the QMaker approach to infer nonreversible Q matrices (e.g., Minh et al. 2020) and/or mixtures of Q matrices from amino acid alignments, for example, as was done to infer the LG4M and LG4X mixtures of matrices (Le et al. 2012). Both of these approaches have the potential to further improve phylogenetic inference beyond the developments that we present here.

SUPPLEMENTARY MATERIAL

Supplementary materials are available online from: https://doi.org/10.6084/m9.figshare.9768101.

FUNDING

This research was funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED; [102.01.2019.06 to B.Q.M., C.C.D., and L.S.V.], an Australian National University Futures Grant to R.L., an Australian Research Council Discovery Grant [DP200103151 to R.L. and B.Q.M.], and a Chan-Zuckerberg Initiative Grant for Essential Open Source Software for Science to B.Q.M. and R.L.

APPENDIX

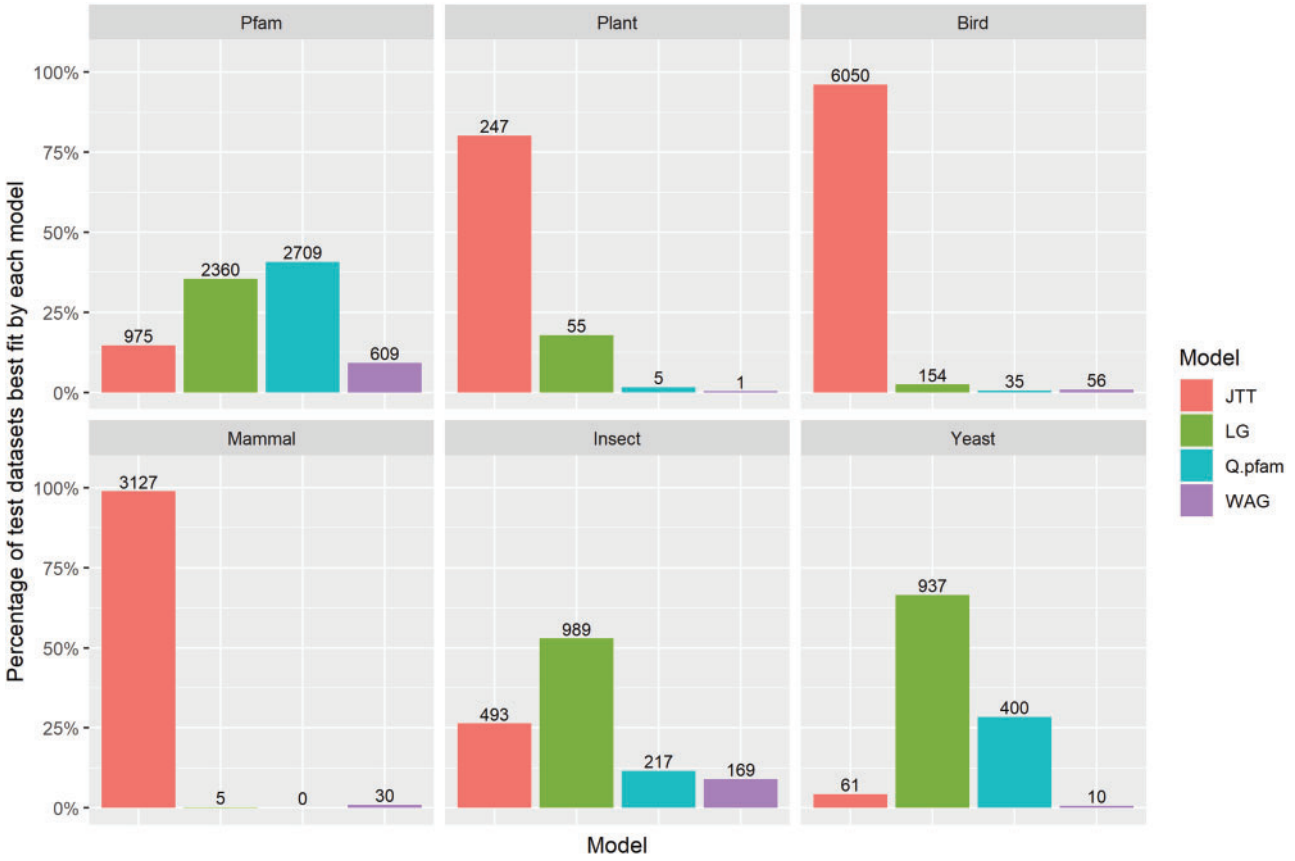

FIGURE 1. The performance of four matrices Q.pfam, JTT, LG, WAG on Pfam, bird, plant, insect, yeast, and mammal data sets.

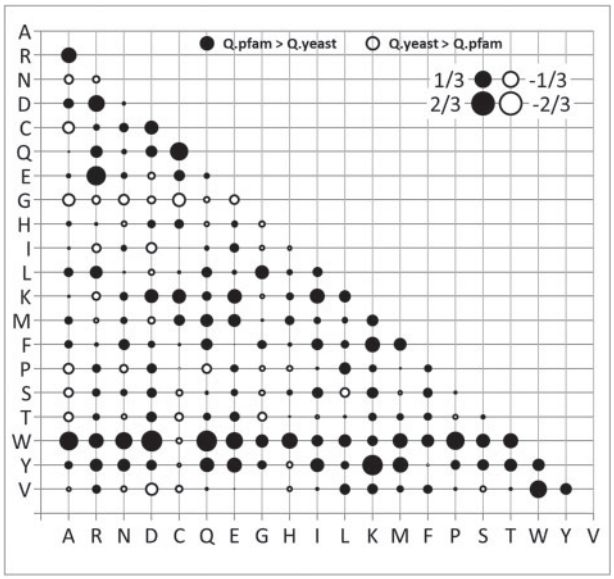

FIGURE 2. The bubble plot show relative differences between amino acid exchangeability rates in Q.pfam and Q.yeast. The explanations as similar as in Figure 4.

References

Abascal F., Posada D., Zardoya R. 2007. MtArt: a new model of amino acid replacement for arthropoda. Mol. Biol. Evol. 24(1):1–5.

Adachi J., Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. 42(4):459–468.

Adachi J., Waddell P.J., Martin W., Hasegawa M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. 50(4):348–358.

Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M., Sonnhammer E.L. 2002. The Pfam protein families database. Nucleic Acids Res. 30(1):276–280.

Chernomor O., Minh B.Q., von Haeseler A. 2015. Consequences of common topological rearrangements for partition trees in phylogenomic inference. J. Comput. Biol. 22(12):1129–1142.

Cuong C.D., Le Q.S., Gascuel O., Vinh S.L. 2010. FLU, an amino acid substitution model for influenza proteins. BMC Evol. Biol. 10(99):1–11.

Darriba D., Posada D., Kozlov A.M., Stamatakis A., Morel B., Flouri T. 2019. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol. Biol. Evol. 37(1):291–294.

Dayhoff M.O., Schwartz R.M., Orcutt B.C. 1978. A model for evolutionary change in proteins. Atlas Protein Sequence Struct. 5:345–352.

Dimmic M.W., Rest J.S., Mindell D.P., Goldstein R.A. 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J. Mol. Evol. 55(1):65–73.

Duchene D.A., Tong K.J., Foster C.S.P., Duchene S., Lanfear R., Ho S.Y.W. 2019. Linking branch lengths across sets of loci provides the highest statistical support for phylogenetic inference. Mol. Biol. Evol. 37(4):1202–1210.

El-Gebali S., Mistry J., Bateman A., Eddy S.R., Luciani A., Potter S.C., Qureshi M., Richardson L.J., Salazar G.A., Smart A., Sonnhammer E.L.L., Hirsh L., Paladin L., Piovesan D., Tosatto S.C.E., Finn R.D. 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47(D1):D427-D432.

Fletcher R. 1987. Practical methods of optimization. New York (NY): Wiley.

Gu X., Fu Y.X., Li W.H. 1995. Maximum-likelihood-estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12(4):546–557.

Guindon S., Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52(5):696–704.

Henikoff S., Henikoff J.G. 1992. Amino-acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89(22):10915–10919.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Alfaro-Nunez A., Narula N., Liu L., Burt D., Ellegren H., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G.J., The Avian Phylogenomics Consortium. 2015. Phylogenomic analyses data of the avian phylogenomics project. Gigascience 4(1):1–9.

Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8(3):275–282.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14(6):587–589.

Lanfear R., Calcott B., Ho S.Y., Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29(6):1695–1701.

Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott B. 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34(3):772–773.

Le S.Q., Dang C.C., Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. 29(10):2921–2936.

Le S.Q., Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25(7):1307–1320.

Le S.Q., Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Syst. Biol. 59(3):277–287.

Le S.Q., Lartillot N., Gascuel O. 2008. Phylogenetic mixture models for proteins. Philos. Trans. R. Soc. Lond. B 363(1512):3965–3976.

Le V.S., Dang C.C., Le Q.S. 2017. Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. BMC Evol. Biol. 17(136):1–13.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37(5):1530–1534.

Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer A.J., Aspock U., Aspock H., Bartel D., Blanke A., Berger S., Bohm A., Buckley T.R., Calcott B., Chen J., Friedrich F., Fukui M., Fujita M., Greve C., Grobe P., Gu S., Huang Y., Jermiin L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schutte K., Sekiya K., Shimizu S., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan M., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune T., Walzl M.G., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K., Wu Q., Wu G., Xie Y., Yang S., Yang Q., Yeates D.K., Yoshizawa K., Zhang Q., Zhang R., Zhang W., Zhang Y., Zhao J., Zhou C., Zhou L., Ziesmann T., Zou S., Li Y., Xu X., Zhang Y., Yang H., Wang J., Wang J., Kjer K.M., Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346(6210):763–767.

Muller T., Vingron M. 2000. Modeling amino acid replacement. J. Comput. Biol. 7(6):761–776.

Nguyen L.T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32(1): 268–274.

Nickle D.C., Heath L., Jensen M.A., Gilbert P.B., Mullins J.I., Pond S.L.K. 2007. HIV-specific probabilistic models of protein evolution. PLoS One 2(6):1–11.

Ran J.H., Shen T.T., Wang M.M., Wang X.Q. 2018. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. Proc. R. Soc. B 285(1881):1–9.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53(1–2):131–147.

Rota-Stabelli O., Yang Z.H., Telford M.J. 2009. MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies. Mol. Phylogenet. Evol. 52(1):268–272.

Saitou N., Nei M. 1987. The neighbor-joining method - a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4(4):406–425.

Sayyari E., Whitfield J.B., Mirarab S. 2017. Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. Mol. Biol. Evol. 34(12):3279–3291.

Shen X.X., Opulente D.A., Kominek J., Zhou X., Steenwyk J.L., Buh K.V., Haase M.A.B., Wisecaver J.H., Wang M., Doering D.T., Boudouris J.T., Schneider R.M., Langdon Q.K., Ohkuma M., Endoh R., Takashima M., Manabe R., Cadez N., Libkind D., Rosa C.A., DeVirgilio J., Hulfachor A.B., Groenewald M., Kurtzman C.P., Hittinger C.T., Rokas A. 2018. Tempo and mode of genome evolution in the budding yeast subphylum. Cell 175(6):1533.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51(3):492–508.

Shimodaira H., Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 12:1246–1247.

Sullivan J., Joyce P. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36(1):445–466.
Veerassamy S., Smith A., Tillier E.R.M. 2003. A transition probability model for amino acid substitutions from blocks. J. Comput. Biol. 10(6):997–1010.
Wang H.-C., Li K., Susko E., Roger A.J. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol. Biol. 8(1):331.
Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18(5):691–699.
Wu S.Y., Edwards S., Liu L. 2018. Genome-scale DNA sequence data and the evolutionary history of placental mammals. Data Brief 18:1972–1975.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39(3):306–314.
Yang Z.H., Nielsen R., Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15(12):1600–1611.
Zhou X.F., Shen X.X., Hittinger C.T., Rokas A. 2018. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. Mol. Biol. Evol. 35(2):486–503.