Usual methods of phylogenetic tree topology reconstruction are known to have limitations. For example, maximum likelihood (ML) is known to fail when data violate some of the underlying model assumptions (Swofford et al. 2001; Ho and Jermiin 2004; Kück et al. 2012); maximum parsimony (MP) is statistically inconsistent in the Felsenstein zone (Felsenstein 1978); and neighbor-joining (NJ) is subject to the choice of an unbiased distance and it is not as accurate as ML when both methods can be applied (Tateno et al. 1994). When trying to estimate relationships among distantly related organisms, neglecting heterogeneity in the substitution process across lineages or heterogeneity across sites may result in inaccurate phylogenetic estimates (see Felsenstein 1978; Fitch 1986; Yang 1994; Yang and Roberts 1995; Galtier and Gouy 1998; Ho and Jermiin 2004; Foster 2004; Kolaczkowski and Thornton 2004; Stefankovic and Vigoda 2007, among others).

Phylogenetic invariants were first introduced by Cavender and Felsenstein (1987) and Lake (1987) as a nonparametric method of phylogenetic reconstruction: they are equations satisfied by any possible joint distribution of character patterns at the leaves of a tree evolving under an evolutionary Markov model. For example, if one considers the Jukes–Cantor model on the tree 1234, then any distribution $p$ of patterns at the leaves satisfies the equation $p_{AACC} = p_{CGGT}$. The potential of phylogenetic invariants was the ability to deal with more general models and of detecting the topology without estimating branch lengths or substitution parameters (see Felsenstein 2004, chapter 22). In particular, they can handle heterogeneity across lineages better than other methods (Casanellas and Fernández-Sánchez 2007; Holland et al. 2013) and (some) could deal with heterogeneity across sites, as Lake’s invariants did (Lake 1987). Nevertheless, only a few phylogenetic invariants were known by that time, it was not clear how to use them (Felsenstein 2004), they seemed useless for large trees, and the approach was laid aside due to the poor results obtained in simulations (Huson 1997).

Eriksson (2005) proposed a new topology reconstruction method, ErikSVD, based on the work on invariants of Allman and Rhodes (2007). The underlying idea is that organizing the joint distribution $p$ of character patterns according to a bipartition $A/B$ of the set of taxa gives a matrix $M_{A/B}(p)$ (called the bipartition matrix from now on, see (1) in the Methods section) of rank $\leq 4$ if the bipartition is induced by an edge of the tree (and otherwise, the rank is higher). This result holds for any set of DNA sequences evolving under the most general Markov (CM) model of nucleotide substitution, also known as the Barry–Hartigan model (Barry and Hartigan 1987; Jayaswal et al. 2005; Allman and Rhodes 2008). This is the most general model assuming heterogeneity across lineages, as it allows different instantaneous rate matrices and heterogeneous composition at different parts of the tree, even locally along each branch (Jayaswal et al. 2011). Given the vector $p$ of relative frequencies of columns in an alignment, ErikSVD does not use phylogenetic invariants directly but computes the distance of the bipartition matrix $M_{A/B}(p)$ to the set of matrices of rank $\leq 4$. The shorter the distance, the more likely is the bipartition $A/B$ to come from an edge of the tree underlining the data. However, the original ErikSVD method turned out not to be accurate enough to compete against standard methods (Eriksson 2005), especially in the presence of long-branches and short alignments. Here we revisit...
ErikSVD by adjusting the bipartition matrix: we normalize it by column (respectively, row) sums so that we obtain the transition matrix from the states of one side of the bipartition to the other. This adjustment is made to take into account that the rank of the bipartition matrix obtained from empirical distributions could be affected by the presence of long-branch attraction situations. Indeed, in these situations the probability of observing the same nucleotide at two non-sister species a and b can be large, so the corresponding row at the bipartition matrix $M_{\text{bipartition}}(p)$ has large entries, which can distort its theoretical rank (see Appendix 1 for an example). The original ErikSVD was already statistically consistent (i.e., as the empirical distribution approaches the theoretical distribution, the probability of correctly reconstructing the tree goes to one) and so is the method proposed here (see the “Methods” section) which will be called Erik+2.

Erik+2 is model-based as it assumes a general Markov model of evolution (and it could also be redesigned to incorporate either more restrictive Markov models or amino acid substitution models), but is nonparametric in the sense that it does not attempt to recover the parameters of the model. Moreover, the theoretical background of Erik+2 allows it to be applied on heterogeneous data across sites evolved on the same tree topology under the GM model (Jayaswal et al. 2014); that is, a parameter $m$ can be introduced so that Erik+2 considers the sites of the alignment to be divided into $m$ categories, each evolving on the same topology but with (possibly) different Markov substitution matrices and nucleotide distribution at the root—this is often called an $m$-mixture (Stefankovic and Vigoda 2007). For example, discrete-gamma rates or the heterogeneous tree model of evolution (and it could also be extended to the case of continuous-time models with parameters to be estimated by the method) considering homogeneity or heterogeneity across alignments and sites depending on the situation.

The performance of Erik+2 on real data is analyzed on the eight species of yeast studied in Rokas et al. (2003) with the concatenated alignment provided by Jayaswal et al. (2014). We investigate whether the quartets output by Erik+2, ErikSVD, and ML support the tree $T$ of Rokas et al. (2003) or the alternative tree $T'$ of Phillips et al. (2004), and the mixture model proposed by Jayaswal et al. (2014).

**METHODS**

**ErikSVD and Erik+2 Methods**

Erik+2 arises as a variation of the method described by Eriksson (2005) by normalizing certain matrices obtained from an alignment of nucleotide sequences. As in the original method, the information contained in the alignment is recorded as a vector $\hat{p}$ whose coordinates are the observed relative frequencies of possible patterns at the leaves. In the case of an alignment of four taxa $1,2,3,4$, each possible (trivalent) topology is determined by a binary tree with parameters to be estimated by the method) considering homogeneity or heterogeneity across alignments and sites depending on the situation.

For example, the $(AG, CT)$-entry of $M_{1234}(p)$ is considered by Jayaswal et al. (2014).

Assume that the coordinates of $p$ are the empirical estimates of the theoretical joint distribution $p$ at the leaves of a tree $T$ evolving under the GM model, say

\begin{equation}
\begin{pmatrix}
\hat{p}_{AA} & \hat{p}_{AC} & \hat{p}_{AG} & \hat{p}_{CT} \\
\hat{p}_{AA} & \hat{p}_{AC} & \hat{p}_{AG} & \hat{p}_{CT} \\
\hat{p}_{AA} & \hat{p}_{AC} & \hat{p}_{AG} & \hat{p}_{CT} \\
\hat{p}_{AA} & \hat{p}_{AC} & \hat{p}_{AG} & \hat{p}_{CT} \\
\end{pmatrix}
\end{equation}

(1)
$T = 12|34$. Then the key point is Theorem 19.5 of Eriksson (2005) (see also Casanellas and Fernández-Sánchez (2010)) that claims that the rank of $M_{A,B}(p)$ is 4 if $A|B = 12|34$, and $I^2$ otherwise (if the substitution matrices that generated $p$ were general enough). Eriksson’s idea is to compute the Frobenius distance (i.e., the Euclidean distance if we view the matrices as elements in $\mathbb{R}^{2+2}$, Demmel 1997, 4 of the three matrices $M_{1234}(p)$, $M_{1324}(p)$, and $M_{1423}(p)$ to the space of matrices of rank $\leq 4$. In this manner, one derives which of the three matrices is closer to having rank $\leq 4$. Note that the Frobenius distance of a matrix $M$ to the set of matrices of rank $< k$, $d_k(M)$, is easily computed in terms of the singular values of $M$ (Eckart and Young 1936).

The main motivation for the variation introduced in Erik+2 arises from the observation that the presence of short branches may seriously affect this distance when it is computed from short alignments. For example, for a tree with small $a$ and large $b$ (a tree in the Felsenstein zone; Fig 1a), the distance of $M_{1234}(p)$ to 4-rank matrices can be smaller than that of $M_{1324}(p)$ (see also Appendix 1 provided in Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.jk7t0). The reason is that a small $a$ implies that the probability of observing the same nucleotide at leaves 2 and 4 is high, so columns in $M_{1234}(p)$ indexed by $\{A\rightarrow B, C\rightarrow C, G\rightarrow G, \text{ or } T\}T$ capture most of the nonzero entries in the matrix, whereas other columns may only have few nonzero entries. Thus $M_{1324}(p)$ is not far from having rank 4, even if 13/24 is not the correct topology.

Moreover, if $p$ evolves on the 12|34 topology, the sum of the entries of a column of $M_{1234}(p)$ depend on the substitution matrices at the branches 3,4. In the spirit of producing a more robust score that is less influenced by the peripheral substitution processes (see e.g., the idea of using Markov invariants in Holland et al. (2013)), we normalize these sums. By dividing any nonzero column by the sum of its entries, all the nonzero columns become of the same weight. As the same situation may occur with rows, we also need to normalize the matrix by row sums. In this way, each matrix $M_{A,B}(p)$ gives rise to a pair of transition matrices $M_{A,G}(p)$ and $M_{A,B}(p)$, obtained by column and row sum adjustment, respectively:

$$M_{12\rightarrow 34}(p) = \begin{pmatrix} \frac{PAAS}{FAAS} & \frac{PAAC}{FAAC} & \cdots & \frac{PAAT}{FAAT} \\ \frac{PAAC}{PAAC} & \frac{PAAC}{PAAC} & \cdots & \frac{PAAT}{PAAT} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{PAAS}{PAAS} & \frac{PAAC}{PAAC} & \cdots & \frac{PAAT}{PAAT} \end{pmatrix}$$

$$M_{12\rightarrow 34}(p) = \begin{pmatrix} \frac{PAAS}{PAAS} & \frac{PAAC}{PAAC} & \cdots & \frac{PAAT}{PAAT} \\ \frac{PAAC}{PAAC} & \frac{PAAC}{PAAC} & \cdots & \frac{PAAT}{PAAT} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{PAAS}{PAAS} & \frac{PAAC}{PAAC} & \cdots & \frac{PAAT}{PAAT} \end{pmatrix}.$$

We give a score to any tree $T_{A,B}$ as

$$sc(T_{A,B}) := \frac{d_4(M_{A,G}(p)) + d_4(M_{A,B}(p))}{2}$$

Erik+2 outputs the topology with smallest score (notice that the smaller the score, the more reliable the topology $T_{A,B}$ is). As the empirical distribution $p$ approaches the theoretical distribution $p$, the transition matrices $M_{A,G}(p)$ and $M_{A,B}(p)$ approach the theoretical transition matrices. These have rank 4 for the correct topology because they have the same rank as the peripheral bipartition matrices (as they are obtained from them by dividing rows/columns by scalars).

Therefore $d_4(M_{A,G}(p))$ and $d_4(M_{B,A}(p))$ tend to 0 when $p$ approaches the theoretical distribution (as the Frobenius distance is a continuous function) and thus Erik+2 is statistically consistent.

In order to avoid bias in the rank of the bipartition matrices due to insufficient data, in the adjustment by rows we disregard rows whose sum is not larger than $2/N$ where $N$ is the alignment length (same for columns).

Erik+2 also provides normalized weights that can be used as input for weighted quartet-based methods (Strimmer and von Haeseler 2006). Indeed, the score above is turned into a confidence weight by inverting and normalizing so that the overall sum of weights is 1:

$$w(T_{A,B}) := sc(T_{A,B})^{-1} \sum_7 sc(T)^{-1}.$$
We had also developed different modifications of the original method of Eriksson, all of them showing lower success than the version considered here. For example, we have tried different options to provide a single value from the distances $d_A$ of the two transition matrices: the 2-norm, the $\infty$-norm, or even the minimum. We also tried normalizing by rows and columns at the same time (with poor results), and also considering the 2-norm instead of the Frobenius norm (similar results). Therefore in this article we only present the results corresponding to the method Erik+2 explained here.

Maximum Likelihood

Heterogeneity across lineages.—In order to estimate a heterogeneous across lineages model we used the software bppml of the Bio++ package (Dutheil and Boussau 2008). The software had to infer a GTR rate matrix per edge, the branch lengths, and the equilibrium distribution on a rooted quartet tree with either

- homogeneity across sites, ML(GTR) henceforth, or
- discrete gamma rates with two/three categories, ML(GTR+2/3), or
- two categories plus invariable sites, ML(GTR+2/3+I).

Homogeneity across lineages.—When we restricted the likelihood computations to homogeneous across lineages continuous-time models, we used PAML (Yang 1997) to estimate the rate matrix $Q$ (a unique rate matrix for the whole tree), the equilibrium distribution, and the branch lengths on a rooted quartet tree. Depending on the setting, the method infers either

- an unrestricted rate matrix $Q$ (model UNREST in PAML documentation) leading to the most general continuous-time homogeneous Markov model, which is denoted as ML(homGMc), or
- a rate matrix $Q$ restricted to the GTR model, ML(homGTR), with possibly auto-discrete gamma rates, ML(homGTR+I).

For the ML computations, we waited up to 60 seconds for convergence on each tree topology and if it did not converge, we treated it as failed (because we cannot compare likelihoods in this case). It is worth pointing out that, usually, ML was not convergent only for the incorrect topologies.

Neighbor-Joining

As far as neighbor-joining is concerned, the paralinear distance (Lake 1994) was used throughout the article to estimate pairwise divergences in NJ. This distance is based on the GM model and has shown to be very useful for phylogenetic inference when nucleotide sequences

![FIGURE 1. Description of the quartet tree and the tree space used in our simulations. a) 4-leaf tree where the length of two opposite branches is represented by $a$, the other two peripheral branches have length $b$, and the length of the interior branch is denoted by $c$. The root $r$ is located at the parent node of leaves 3 and 4. Branch lengths will be measured in the expected number of substitutions per site. b) Tree space obtained form the tree in a) when the branch length $c$ is set equal to $a$ and branch lengths $a$ and $b$ are varied from 0.01 to 1.5 in steps of 0.02 (see “Description of the data” in the Methods section).]

Felsenstein zone.—We consider trees subject to long-branch attraction, also known as trees in the Felsenstein zone. To this end, on the tree of Figure 1a we fix $a = 0.05$, $b = 0.75$, and let the internal branch length $c$ vary in the range [0.01, 0.4] (or either we fix $c = 0.05$ and call it the Felsenstein tree). For each set of branch lengths, we generated one hundred alignments of different lengths.

Tree space.—We adopt a similar approach to Huelsenbeck (1995) to test different methods. More precisely, we evaluate the methods on a tree space (see Fig. 1b) where the quartets are as in Figure 1a with $c = a$, and the branch lengths $a$ and $b$ vary between 0 and 1.5 in steps of 0.02. For each pair $a, b$ we generate one hundred alignments of a fixed length and represent the success of different methods in recovering the right topology.

Farris tree.—We also consider trees with two incident pendant edges of length equal to 0.5 and the other three edges of equal length varying from 0.01 to 0.05 (Figure S5 in Appendix 2 provided in Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.jk7j0).
Kolaczkowski and Thornton (2004). We consider two categories of the same size both evolving under the GM model on the tree of Figure 1A: the first category corresponds to branch lengths $a=0.05$, $b=0.75$, whereas the second corresponds to $a=0.75$ and $b=0.05$. The internal branch length takes the same value in both categories and varies from 0.01 to 0.4 in steps of 0.05.

**Simulations under the GM model.**—To generate data under the general Markov model, we have used GenNon-h (Kedzierska and Casanellas 2012). Given a set of branch lengths (understood as the expected number of substitutions per site) and the tree topology of Figure 1 and Appendix 2 provided in Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.jk7t0 had rates 2 ($A\rightarrow C$), 7 ($A\rightarrow G$), 4 ($A\rightarrow T$), 3 ($C\rightarrow G$), 1 ($C\rightarrow T$), 5 ($G\rightarrow T$). The rate matrix underlying GTR+$\Gamma$ had rates 2 ($A\rightarrow C$), 5 ($A\rightarrow G$), 3 ($A\rightarrow T$), 4 ($C\rightarrow G$), 1 ($C\rightarrow T$), 2 ($G\rightarrow T$) and the sites were varied according to a gamma distribution with parameter $\alpha=\beta$ in the range $[0.1,2]$ varying in steps of 0.1. Small values of this parameter indicate a lot of variation across sites (Yang 1993).

**Real data.**—We consider the data provided by Jayaswal et al. (2014) with 42,337 second codon positions of 106 orthologous genes of Saccharomyces cerevisiae, S. paradoxus, S. mikatae, S. kudriavzevii, S. castellii, S. kluveri, S. bayanus, and Candida albicans. The phylogenetic tree of these species was originally studied in Rokas et al. (2003), where a tree topology $T$ was identified with its correct inference is known to depend on the consideration of heterogeneity across lineages (Rokas et al. 2003; Phillips et al. 2004; Jayaswal et al. 2014). For example, Phillips et al. (2004) obtain an alternative $T'$ with 100% bootstrap support for the concatenated alignment of these genes. This tree is widely accepted by the community but its correct inference is known to depend on the consideration of heterogeneity across lineages (Rokas et al. 2003; Phillips et al. 2004; Jayaswal et al. 2014). For example, Phillips et al. (2004) obtain an alternative $T'$ with 100% bootstrap support for the concatenated alignment of these genes. This tree is widely accepted by the community but its correct inference is known to depend on the consideration of heterogeneity across lineages (Rokas et al. 2003; Phillips et al. 2004; Jayaswal et al. 2014).

**Homogeneity across Sites**

First of all we evaluate Erik+2, ML, NJ, and ErikSVD on the Felsenstein zone on data generated under the GM model (Fig. 1). In this figure, two models underlying ML computations have been considered: the most general homogeneous continuous-time model, $ML_{(homGMc)}$, and a (heterogeneous across lineages) GTR model, $ML(GTR)$. ML has a similar performance with both models. We observe that Erik+2 is more accurate than ErikSVD in general and especially when the interior branch length is short (only for length 1000 and $c\in(0.13,0.25)$ ErikSVD outperforms slightly Erik+2). Both versions of ML perform better than Erik+2 for 1000 bp, Erik+2 and ML perform similarly for 10,000 sites, but Erik+2 outperforms ML (in its both versions) for 100,000 sites and short internal branch. Notice incidentally that the accuracy of ML does not seem to increase as the length of the alignment grows.

In a more complete study, the methods Erik+2, ML, and NJ have been tested on the tree space for data generated under the GM model (Fig. 2) and under a homogeneous GTR model (Fig. 4). Both $ML_{(homGMc)}$ and NJ have lower accuracy than Erik+2 (Fig. 2), as was expected under data that violates the assumptions of ML and NJ. Although Erik+2 and NJ drastically increase their accuracy when the alignment length is multiplied by 10, ML does not (Fig. 4 and, very specially, Fig. 3 where the substitutition model assumed by ML is incorrect). Under the homogenous GTR model and for length 10,000 (Fig. 4), Erik+2 performs slightly better than ML (homGTR) (although the difference does not seem significant). One possible explanation is suggested by the presence of long branches, which may distort the performance of ML even when the correct model is assumed (Kück et al. 2012). Erik+2 also outperforms ErikSVD on this tree space (Fig. 4 and Fig. S1 in Appendix 2 provided in Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.jk7t0). We also present the average success achieved by these methods on this tree space for alignments of length 500 bp and 1000 bp (Table 1).

It is worth pointing out that, for alignments of 1000 bp evolving under the homogeneous GTR model, ML seems to outperform Erik+2 in the Felsenstein zone when it estimates exactly this model, $ML_{(homGTR)}$ (Fig. 4). However, for length 10,000, Erik+2 already outperforms ML (homGTR). Moreover, the global accuracy of $ML_{(homGMc)}$ drastically drops when applied to data obtained under the GM model.
FIGURE 2. Performance in the Felsenstein zone. Percentage of correctly reconstructed topologies by Erik+2, ErikSVD, maximum likelihood ML, and maximum-parsimony MP on data generated under the general Markov model (GM) on the Felsenstein zone tree (see “Description of the data” in the Methods section). Two types of ML inference have been applied here: ML(homGMc) estimating the most general homogeneous continuous-time model, and ML(GTR) estimating a GTR model (due to the execution time of this last method, we could only test it for 1000 bp and 10,000 bp).
FIGURE 3. Data generated under GM. Performance in the tree space of Figure 1b on data generated under the general Markov model. Left: 1000 bp; Right: 10,000 bp. Black is used to represent 100% of successful topology reconstruction, white to represent 0%, and different tones of gray the intermediate frequencies. The 95% contour line is drawn in white, whereas the 33% contour line is drawn in black. Top: Erik+2; Middle: neighbor-joining (paralinear distance); Bottom: ML(homGMc) estimating the most general homogeneous across lineages continuous-time model.
FIGURE 4. Data generated under homGTR. Performance in the tree space of Figure 1b on data generated under the (homogeneous across lineages) GTR model. Left: 1000 bp; Right: 10,000 bp. Black is used to represent 100% of successful topology reconstruction, white to represent 0%, and different tones of gray the intermediate frequencies. The 95% contour line is drawn in white, whereas the 33% contour line is drawn in black. Top: Erik+2; Middle: neighbor-joining (paralinear distance); Bottom: ML(homGTR) estimating homogeneous GTR model.
with two categories of evolutionary rates, we have the corresponding 0% line in the figure).

We also generated data under the homogeneous GTR model (homGTR) for different lengths. In parentheses, we show the standard deviation of the set of percentages of success of each method in each tree space. In each row, the highest success is indicated in bold font. ML(homGTR) is applied when data are generated under the continuous GM model (i.e., it estimates the most general homogeneous continuous-time model), whereas ML(homGMc) is applied when data are generated under the general time-reversible model homogeneous across lineages and sites.

Table 1. Average success of different quartet methods on the tree space of Figure 1b

| Simulations | Base pairs | ErikSVD | Erik+2 | NJ | ML |
|-------------|------------|---------|--------|----|----|
| GM          | 1000       | 0.856 (0.21) | 0.803 (0.17) | 0.797 (0.18) | 0.726 (0.17) |
|             | 10,000     | 0.958 (0.13) | 0.971 (0.04) | 0.943 (0.09) | 0.754 (0.17) |
| homGTR      | 500        | 0.732 (0.21) | 0.748 (0.22) | 0.729 (0.23) | 0.840 (0.11) |
|             | 1000       | 0.796 (0.30) | 0.843 (0.19) | 0.805 (0.20) | 0.934 (0.06) |
|             | 10,000     | 0.940 (0.22) | 0.992 (0.04) | 0.945 (0.10) | 0.980 (0.02) |

Notes: Average success of ErikSVD, Erik+2, NJ and ML obtained in the results of Figures 3, 4, and Figure S2 in Appendix 2 provided in Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.jk7j0. This is data simulated on the tree space of Fig. 1b according to the general Markov model (GM) or the time-reversible model homogeneous across lineages and sites (homGTR for different lengths). In parentheses, we show the standard deviation of the set of percentages of success of each method in each tree space. In each row, the highest success is indicated in bold font. ML(homGTr) is applied when data are generated under the continuous GM model (i.e., it estimates the most general homogeneous continuous-time model), whereas ML(homGMc) is applied when data are generated under the general time-reversible model homogeneous across lineages and sites.

Table 2. Performance of Erik+2 and ML on the Farris tree with data simulated under the GM model (Fig. S5 in Appendix 2 provided in Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.jk7j0). In this case, ML performs slightly better than Erik+2.

Performance on Real Data

We applied ErikSVD, Erik+2 with $m=1, 2, 3, 4$, ML (GTR+2$^T$I), and ML (GTR+3$^T$I) to 4-taxon subalignments of the real data described above and investigated the proportion of output quartets that are compatible with $T$ or $T'$. Erik+2 and ML support the tree $T$ and the model suggested by Jayaswal et al. (2014) ($m=3$ for Erik+2 and GTR+2$^T$I for ML), whereas ErikSVD gives more support to the alternative tree $T'$. (Table 2).

We also represented the weights of each quartet output by Erik+2 with $m=3$ and ML (GTR+3$^T$I) (Fig. S6 in the Appendix 2 provided in Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.jk7j0). It is interesting to observe that Erik+2 provides powerful discrimination between quartets and that the weights of the quartets that are incompatible with $T$ are the lowest. Indeed, while Erik+2 exhibits strong support for the correct topology and its weights vary from 0.3066 to 0.6700, ML weights are essentially equal and vary from 0.3314 to 0.3365. It is interesting to note that both methods tend to fail with the same quartets.
FIGURE 5. Accuracy under gamma distribution and mixture data. Percentage of correctly reconstructed topologies by different methods on alignments of lengths 1000, 10,000 and 100,000 bp. shown from top to bottom. a) Data simulated under (homogeneous across lineages) GTR model with continuous gamma-rates and parameter \( \alpha \) varying between 0.1 and 2 on the Felsenstein tree (see “Description of the data” in the Methods section). \( \text{Erik}+2 \), \( \text{ErikSVD} \), and ML estimating the GTR model (homogeneous across lineages) with auto-discrete gamma-rates and denoted as ML(homGTR+Γ) (MP had 0% success on this data, so we do not show it). b) Data generated under the GM model with two categories according to the test designed in (Kolaczkowski and Thornton 2004), varying the internal branch length, and recovering with \( \text{Erik}+2 \) with \( m=2 \), \( \text{Erik}+2 \) with \( m=1 \), ML(homGMc) estimating the most general homogeneous across lineages model, and ML(GTR+2Γ) estimating a time-reversible model with 2 discrete-gamma categories (due to the time of execution of this last method, we could only test it for 1000 bp). The plot represents the logistic regression curve of the output of each method. In all cases, ML had to estimate all parameters.
may obtain better execution times.

Notes: Percentage of quartets output by ErikSVD, Erik+2 (with different mixture assumptions), and ML (assuming a GTR model with either two discrete rate categories and invariant sites, or three discrete rate categories across sites) that are compatible with the yeast tree $T$ of Rokas et al. (2003) and the alternative tree $T'$ of Phillips et al. (2004). In each column, the highest success is indicated in bold font.

| Topology | ErikSVD | Erik+2 ($m=1$) | Erik+2 ($m=2$) | Erik+2 ($m=3$) | ML (GTR+2I) | ML (GTR+3I) |
|----------|---------|----------------|----------------|----------------|-------------|-------------|
| $T$      | 91.42   | 84.29          | 87.14          | 92.86          | 97.14       | 91.42       |
| $T'$     | 94.26   | 82.86          | 77.14          | 85.71          | 90.00       | 84.29       |

**DISCUSSION AND CONCLUSIONS**

The simulation studies show that Erik+2 is an accurate and robust topology reconstruction method on quartets, especially in situations where other methods systematically fail (model heterogeneity across sites or lineages, or long-branch attraction). In such scenarios, Erik+2 outperforms the method of Eriksson, ErikSVD, and common methods like MP, NJ, and ML based on models that cannot accommodate these assumptions. Erik+2 is based on the most general Markov model and hence accounts for heterogeneous data across lineages, even locally at each edge. When its assumptions are violated, for example in the presence of continuous gamma-distributed rates among sites, we have shown that it is highly accurate if there is enough data. As observed, Erik+2 can also deal with $n$-mixtures on the same tree topology (although for quartets the limit is $m=3$). Even more, the simulations presented suggest that Erik+2 with $m=2$ seems to be a good option to deal with large alignments when the mixture status of the data is unknown.

On the experiments we presented, the overall performance of ML is quite accurate if model assumptions are not violated, confirming the conclusions of Kuck et al. (2012). Also in line with these papers, we corroborate that long sequences do not improve ML performance on data that do not satisfy the hypothesis of the underlying model. Moreover, ML is by far the slowest among the methods tested here, whereas Erik+2 is not as fast as NJ, but still fast as its computing time is nearly twice NJ's. Another drawback of ML is that, quite often, it does not converge when it is computed on the incorrect topology, which makes the comparison of likelihoods impossible. Whereas the goal of Erik+2 is to reconstruct the topology, ML is designed to estimate the parameters of the substitution matrices and it would probably be a good choice to use first Erik+2 and then ML to estimate the parameters. In our simulation study, NJ (with paralinear distance) and MP have been the methods with least success, which is not so surprising if one takes into account that they are also less adaptable to general data. It is worth pointing out that in some cases (e.g., short alignments), ErikSVD might be a better option than Erik+2, but in order to decide which method is best in each case, a detailed statistical study should be done.

We have only developed Erik+2 for quartets with the aim of validating it as a successful method, and it is still a work in progress to further develop it for larger numbers of taxa. One possibility is to apply some quartet method based on the weights provided by Erik+2. Similar studies have been carried out by Rusinko and Hipp (2012) using weights obtained from phylogenetic invariants. In any case, the weighting method for the quartets plays a key role in this approach (Ranwez and Gascuel 2001). A different possibility would be to use Erik+2 to evaluate the confidence of particular bipartitions of large sets of taxa, and in this case one can deal with a larger number of categories (the maximum $m$ allowed depends on the size of the subsets $A, B$ of taxa involved in the bipartition: $4^{|\text{min}\{|A|,|B|\}}-1-1$). However, some more work has to be done to adapt Erik+2 to more taxa because the lack of data (short alignments relative to the size of the bipartition matrices) might affect the rank of the bipartition matrices. Extending Erik+2 to amino acid data is also viable, and in this case the maximum number of allowed categories is equal to 19 for quartets (although we still have not addressed the possible natural biases that have to be considered in this case).

**Supplementary Material**

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.jk70l.
ACKNOWLEDGMENTS

We are indebted to B. Misof and C. Mayer for encouraging us to pursue this project and for their very useful comments. We wish to thank them and their group for their warm hospitality during our stay at the Alexander Koering Zoological Museum. We also thank F. Anderson, B. Boussau, O. Cascuelo, B. Holland, P. Jarvis, and L. Jermiin for useful comments and suggestions.

REFERENCES

Allman E.S., Rhodes J.A. 2006. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. J. Comput. Biol. 13:1101–1113.

Allman E.S., Rhodes J.A. 2007. Phylogenetic invariants. In: Gascuel O., Steel M.A., editors. Reconstructing evolution. New York: Oxford University Press.

Allman E.S., Rhodes J.A. 2008. Phylogenetic ideals and varieties for the general Markov model. Adv. Appl. Math. 40(2):127–148.

Barry D., Hartigan J.A. 1987. Statistical analysis of hominoid molecular evolution in the Bio++ suite of libraries and programs. BMC Bioinformatics 13(1):216.

Batschelet E. 1981. Complex analysis for mathematicians and engineers. New York: Academic Press.

Batschelet E. 1983. Topics in mathematics for the life sciences. New York: Academic Press.

Batschelet E. 1984. A primer of statistics for psychological research. New York: Academic Press.

Batschelet E. 1985. A primer of statistical applications for psychologists. New York: Academic Press.

Batschelet E. 1986. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1987. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1988. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1989. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1990. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1991. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1992. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1993. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1994. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1995. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1996. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1997. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1998. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 1999. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2000. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2001. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2002. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2003. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2004. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2005. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2006. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2007. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2008. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2009. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2010. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2011. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2012. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2013. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2014. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2015. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2016. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2017. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2018. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2019. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2020. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2021. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2022. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2023. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2024. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2025. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2026. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2027. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2028. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2029. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2030. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2031. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2032. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2033. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2034. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2035. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2036. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2037. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2038. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2039. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2040. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2041. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2042. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2043. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2044. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2045. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2046. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2047. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2048. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2049. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2050. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2051. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2052. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2053. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2054. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2055. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2056. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2057. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2058. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2059. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2060. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2061. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2062. A primer of statistics for psychologists. New York: Academic Press.

Batschelet E. 2063. A primer of statistics for psychologists. New York: Academic Press.