Alignment Errors Strongly Impact Likelihood-Based Tests for Comparing Topologies

Eli Levy Karin,[1] Edward Susko,*[,2] and Tal Pupko*[,1]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

[2]Department of Mathematics and Statistics, Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, NS, Canada

***Corresponding author:** E-mail: susko@mathstat.dal.ca; talp@post.tau.ac.il.

Associate editor: Jeffrey Thorne

Abstract

Estimating phylogenetic trees from sequence data is an extremely challenging and important statistical task. Within the maximum-likelihood paradigm, the best tree is a point estimate. To determine how strongly the data support such an evolutionary scenario, a hypothesis testing methodology is required. To this end, the Kishino–Hasegawa (KH) test was developed to determine whether one topology is significantly more supported by the sequence data than another one. This test and its derivatives are widely used in phylogenetics and phylogenomics. Here, we show that the KH test is biased in the presence of alignment error and can lead to erroneous conclusions. Using simulations we demonstrated that due to alignment errors the KH test often rejects one of the competing topologies, even though both topologies are equally supported by the data. Specifically, we show that the KH test favors the guide tree used to align the analyzed sequences. Further, branch length optimization renders the test too conservative. We propose two possible corrections for these biases. First, we evaluated the impact of removing unreliable alignment columns and found out that it decreases the bias at the cost of substantially reducing the test's power. Second, we developed a parametric test that entirely abolishes the biases without data filtering. This test incorporates the alignment construction step into the test's hypothesis, thus removing the above guide tree effect. We extend this methodology for the case of multiple-topology comparisons and demonstrate the applicability of the new methodology on an exemplary data set.

Key words: **alignment, alignment uncertainty, KH test, SOWH test, phylogeny, likelihood, tree comparisons, branch length optimization.**

Introduction

The Kishino–Hasegawa (KH) test (Kishino and Hasegawa 1989) is one of the most commonly used likelihood-based statistical tests of competing evolutionary hypotheses (tree topologies). It has been broadly used in thousands of evolutionary studies. For example, Teeling et al. (2000) used it to show Microchiroptera are not a monophyletic group, Lister et al. (2005) used it to determine the position of the giant deer (*Megaloceros giganteus*) among other deer species, and Daubin et al. (2003) used it to investigate lateral gene acquisition events in bacteria. Moreover, various popular implementations for the KH test are available: TREE-PUZZLE (Schmidt et al. 2002), PAML (Yang 2007), and CONSEL (Shimodaira and Hasegawa 2001).

The KH test is designed to compare two a-priori specified trees with respect to a multiple sequence alignment (MSA). Its null hypothesis states that the two trees are equally supported by the MSA data. The alternative hypothesis, in the one-sided version of the test, states that T_1 is better supported. The *P* value computed by the test represents the probability of seeing, under null conditions, a difference in the log-likelihoods of the two trees which is at least as big as the observed difference.

In order to compute this *P* value, the KH test receives as input two vectors: $\mathrm{LL}(T_1)$ and $\mathrm{LL}(T_2)$, such that $\mathrm{LL}_h(T_1)$ is the log-likelihood score of T_1 based on position h in the MSA. The log-likelihoods of each tree, $S(T_1)$ and $S(T_2)$, are simply the sum of the values in each of the LL vectors. Let δ be the difference in the log-likelihood values of the two trees ($\delta = S(T_1) - S(T_2)$). Under the null hypothesis, $E(\delta) = 0$. In order to estimate whether the observed value of δ is significantly greater than 0, independence between sites is assumed which allows estimating the variance of δ in the following way:

$$\widehat{\sigma^2_\delta} = \frac{n}{n-1}\sum_{h=1}^{n}\left\{\mathrm{LL}_h(T_1) - \mathrm{LL}_h(T_2) - \frac{1}{n}\sum_{h'=1}^{n}\left[\mathrm{LL}_{h'}(T_1) - \mathrm{LL}_{h'}(T_2)\right]\right\}^2.$$

The KH test next makes use of the central limit theorem to note that $S(T_1)$ and $S(T_2)$ approximately follow a normal distribution and therefore the difference between them, δ, is also approximately normally distributed. A standardized test statistic z is then calculated as follows:

$$z = \frac{\delta - 0}{\sqrt{\widehat{\sigma^2_\delta}}}.$$

For which a *P* value can then be calculated:

$$P\,\text{value} = 1 - \phi(z).$$

Article

The RELL variation of the KH test replaces the normal approximation by a bootstrap resampling method (Kishino et al. 1990). In this approach, the position-based likelihood scores are resampled and a distribution of log-likelihood differences is obtained. A *P* value is computed based on the proportion of time the bootstrap log-likelihood difference is greater than the observed log-likelihood difference.

Shimodaira and Hasegawa (1999) and Goldman et al. (2000) discuss the limitations of applying the KH test. Specifically, they emphasize that the KH test is only valid for comparing two a-priori specified topologies (i.e., it is sensitive to selection bias). These limitations motivated the development of a new generation of tests for phylogeny comparison, such as the nonparametric Shimodaira–Hasegawa test (Shimodaira and Hasegawa 1999) and the Approximately Unbiased (AU) test (Shimodaira 2002), which adjust for the selection bias. While offering improvements to the original KH test, these more recent tests still rely on the KH test principles: A nonparametric bootstrapping of a single MSA. These principles form the very foundations for likelihood-based hypotheses testing in phylogeny. Alternative to the nonparametric tests, the SOWH test (Swofford et al. 1996; Goldman et al. 2000) is a parametric bootstrap approach that adjusts for the selection bias.

Without exception, all these tests rely on an MSA as input. Currently, all available test procedures do not account for alignment uncertainty; treating the input MSA as free from errors and unbiased. Likelihood-based tests for topology comparison are one of many downstream biological analyses for which an MSA is a prerequisite. Such downstream analyses include the inference of positive selection, phylogeny reconstruction, and the detection of lateral gene transfer events. It has been recently shown that errors in the MSA can lead to erroneous phylogenetic inference (Ogden and Rosenberg 2006; Talavera and Castresana 2007; Wang et al. 2011) and may affect the inference of positive selection (Fletcher and Yang 2010; Jordan and Goldman 2012; Privman et al. 2012). Notably, the guide tree used in progressive alignment methods has been shown to have a strong impact on the resulting alignments (Nelesen et al. 2008; Landan and Graur 2009; Penn, Privman, Landan, et al. 2010; Capella-Gutierrez and Gabaldon 2013; Toth et al. 2013).

In this study, we characterize the bias introduced by alignment errors to the KH test. In addition, we present the effect of branch length optimizations on the test. We then propose a nonparametric method to reduce this bias as well as a novel parametric methodology that takes into account the impact of the guide tree on the alignment as well as branch length optimization.

Results

Sensitivity to Branch Length Optimization and Alignment Errors

We first tested whether the KH test is biased under "ideal" null conditions, in which the two tested topologies are equally supported by the data. To assure such ideal conditions of no alignment errors, the "true" MSA as generated by the simulation program was used for computing the site-specific log-likelihood scores used by the KH test. Moreover, the two topologies to be compared were given a priori with their associated branch lengths (fig. 1). Here, as is conventional, we use the one-sided version of the KH test, such that the alternative hypothesis (H1) to the null hypothesis (H0) is that a prespecified topology (T_1) is better supported by the data than an alternative topology (T_2). Under these conditions, as expected, the obtained *P* values are uniformly distributed between 0 and 1 (turquoise columns in fig. 2 for phylogeny set *A* and supplementary fig. S1, Supplementary Material online, for phylogeny sets *B* and *C*). The null hypothesis of a Kolmogorov–Smirnov test (*P* values derived from a uniform distribution) could not be rejected for all three topology sets.

In reality, both the MSA and the branch lengths of the two competing tree topologies are unknown and are inferred from the data (the set of unaligned homologous sequences). We first tested the impact of optimizing branch lengths on the KH test. We repeated the above simulations, optimizing the branch lengths of the two alternative topologies from the MSA. The distribution of *P* values substantially deviates from the expected uniform distribution (Kolmogorov–Smirnov test; $P < 10^{-15}$; fig. 3). The obtained distribution is centered around 0.5 and has less mass near 0.05. Thus the test is more conservative than it should be, a consequence of additional variability in log likelihoods due to the addition of branch length optimization.

We next tested the impact of alignment errors on the KH test. To this end, we repeated the above simulations, only this time the simulated sequences were aligned by MAFFT. Examining the *P*-value distribution revealed a deviation from the uniform distribution for the MAFFT alignment (Kolmogorov–Smirnov test; $P < 10^{-4}$ for all three phylogeny sets; pink columns in fig. 2 for phylogeny set *A* and supplementary fig. S1, Supplementary Material online, for phylogeny sets *B* and *C*). The proportion of *P* values smaller than 0.05 was found to be higher than 0.07 in each of the three phylogeny sets. This deviation from 0.05 is statistically significant (Binomial test; $P < 0.0001$). The meaning of this deviation is that alignment errors render the KH test too permissive, that is, it has a 40% higher-than-acceptable probability to reject the null hypothesis, wrongly preferring one topology over the other, even though there is no significant difference between the topologies. Notably, when we performed this analysis using PRANK as the alignment program, similar results were obtained (supplementary fig. S2, Supplementary Material online). We thus chose to focus on MAFFT in our analyses as it is more computationally efficient than PRANK.

It was previously shown that the guide tree used in progressive alignment methods, such as MAFFT, has a strong impact on the resulting alignment (Nelesen et al. 2008; Landan and Graur 2009; Penn, Privman, Landan, et al. 2010; Capella-Gutierrez and Gabaldon 2013; Toth et al. 2013). We hypothesized that the KH test's sensitivity to alignment errors demonstrated above may stem from the guide tree used to direct the alignment. Specifically, we suspected that the test would tend not to reject the topology of the tree used in guiding the alignment and overreject the competing

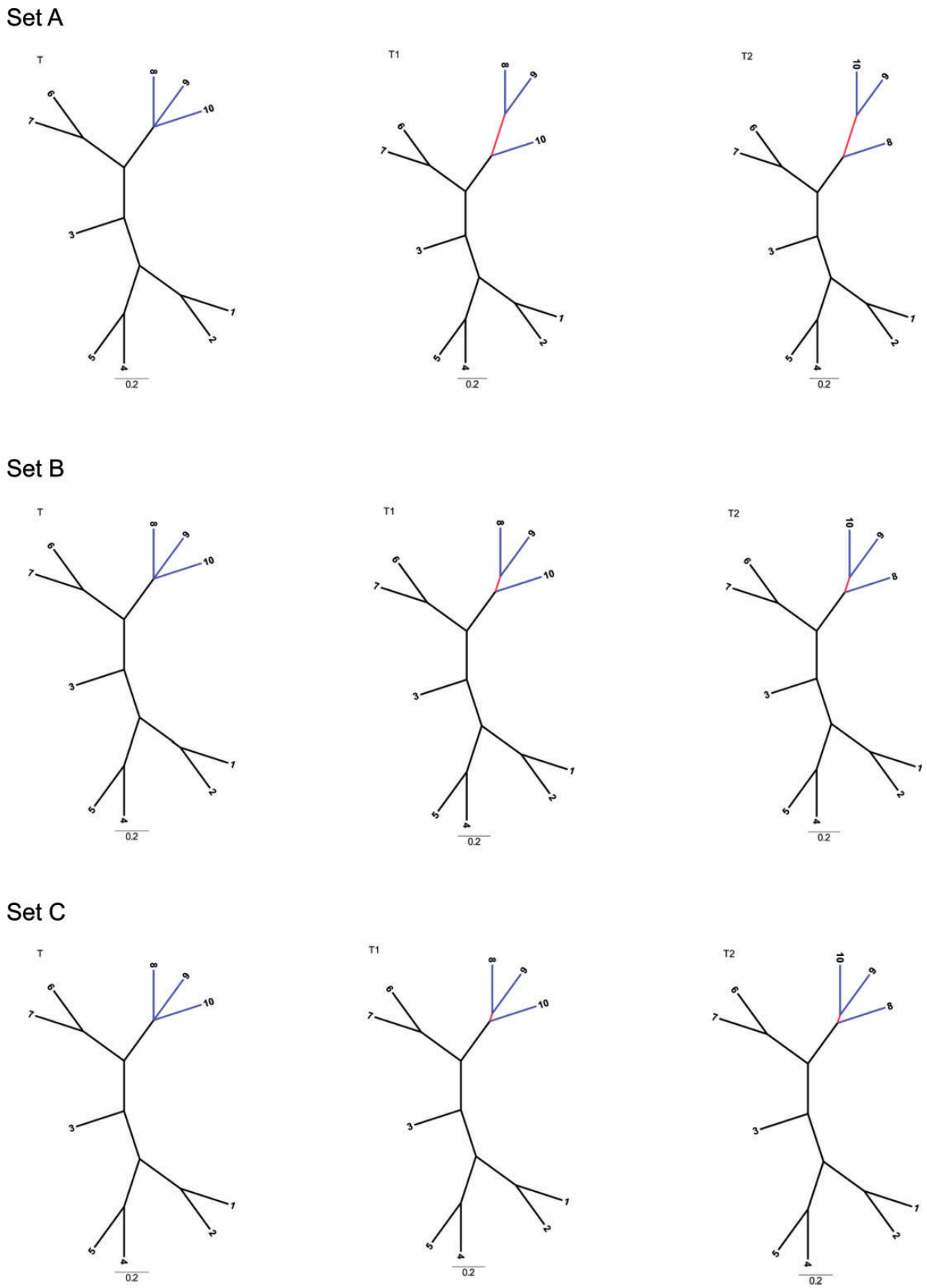

FIG. 1. A multifurcating topology T and two derived symmetric bifurcating options, T_1 and T_2. Three such sets were used in this study, which differ in the length of the branch that separates the species 8, 9, and 10 to a pair and a singleton (this branch is colored red). A long red branch (set *A*, top) corresponds to competing trees that are very different from the multifurcating tree, while shorter red branches (sets *B* and *C*, middle and bottom) correspond to competing trees that are not very different from the multifurcating tree.

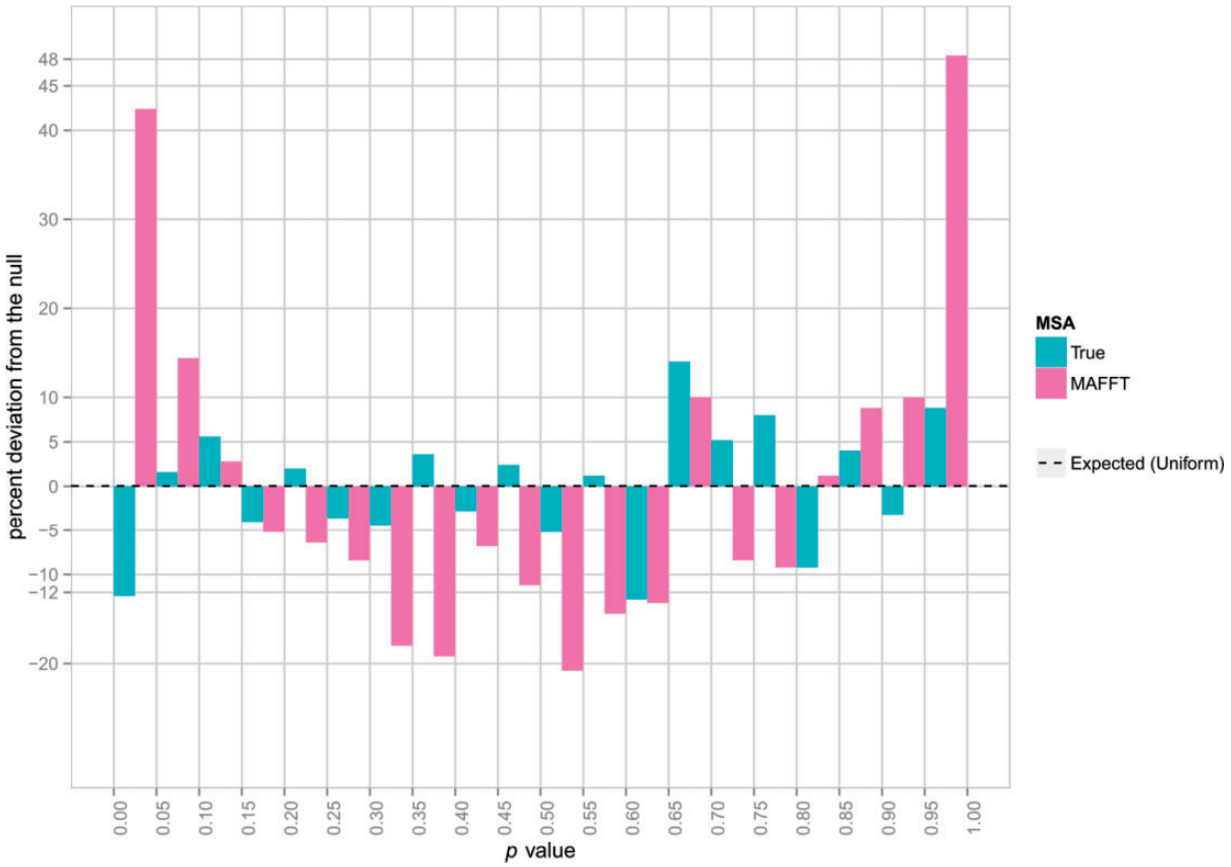

FIG. 2. Testing the bias of the KH test under the null conditions—MSA effect. Sequences were simulated using INDELible based on the multifurcating tree in set *A* of figure 1. Under these conditions, among the 5,000 simulations, 5% of the *P* values should fall in each of the 20 *P* value bins, as expected from a uniform distribution. Shown is the deviation from expected, computed as the difference between the observed number of *P* values in each bin minus the expected number (250), divided by the expected number times 100. For any given bin, a percent deviation of 12% or less is expected 95% of the time. In turquoise, computations were based on the true MSA, whereas in pink the computations were based on the MSA inferred using MAFFT.

topology. Our analysis clearly supports this hypothesis. The average rejection rate for each of the competing topologies (at the 5% level) was approximately 7% when no specific guide tree was given as input to MAFFT (pink columns in fig. 2 and supplementary fig. S1, Supplementary Material online). In contrast, the proportion of tests in which the null hypothesis was rejected ($\alpha = 0.05$) ranged from 15% to 50% (depending on the phylogeny set) when H1 has T_1 equal to the guide tree. Moreover, when H1 has a T_1 different from the guide tree, the null hypothesis was almost never rejected (ranging from 0.2% to 3.26%, depending on the phylogeny set) (fig. 4). Although the result holds for the three phylogeny sets examined, the bias is stronger when the two compared topologies and their associated branch lengths are most dissimilar (fig. 4). Taken together, the KH is highly biased, having a large false positive rate when H1 has T_1 equal to the guide tree.

A Nonparametric Solution to Reduce Bias

Our results suggest that the guide tree has a strong impact on the KH test. One possible solution is thus to remove any alignment column that occurs according to one guide tree but not according to the other. This is exactly the rationale of GUIDANCE (Penn, Privman, Landan, et al. 2010), a tool for detecting unreliable alignment regions. We thus next tested whether filtering out all alignment columns that are unreliable according to GUIDANCE cancels the alignment bias. The observed bias in the KH test (as measured by the proportion of *P* values smaller than 0.05) was reduced from around 0.07 for the unfiltered MAFFT alignments to around 0.05 after filtering out positions with GUIDANCE (turquoise columns in fig. 5 for phylogeny set *A* and supplementary fig. S3, Supplementary Material online, for phylogeny sets *B* and *C*). This reduction is statistically significant (Fisher's exact test; $P < 0.0001$).

However, using GUIDANCE raises two major problems. First, as can be seen in figure 5, the KH test with GUIDANCE filtering is still biased: The distribution of *P* values significantly deviates from the uniform distribution (Kolmogorov–Smirnov test; $P < 10^{-5}$ for all three phylogeny sets). Secondly, removing positions from the alignment reduces the signal. We examined the power of the KH test with the true alignment, the unfiltered MAFFT alignment, and the alignment filtered by GUIDANCE. This was done by simulating sequences by T_1 and examining the proportion of KH tests in which T_2 was rejected (at $\alpha = 0.05$). Our results clearly indicate that GUIDANCE reduces the power of the KH test. This reduction correlates with the similarity among the compared trees (table 1).

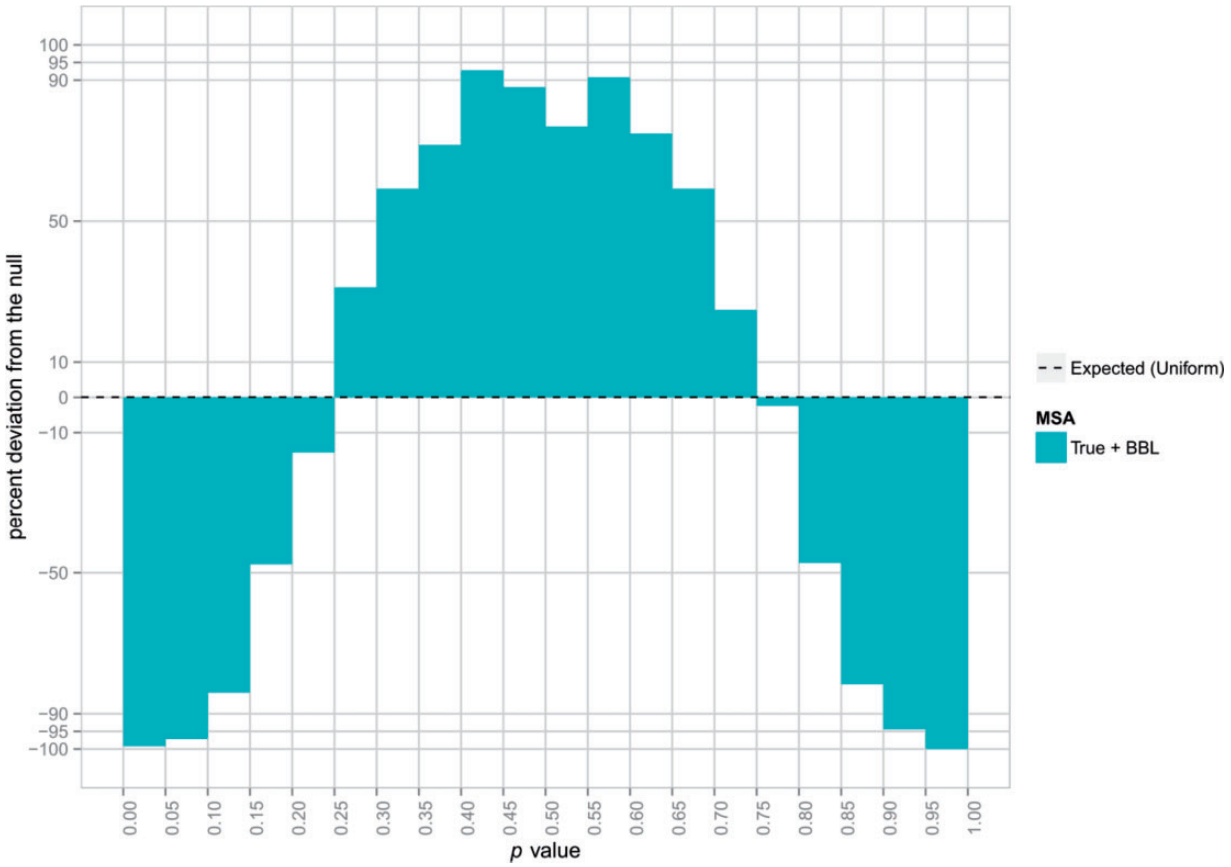

Fig. 3. Testing the bias of the KH test under the null conditions—branch length optimization effect. Sequences were simulated similarly to figure 2. In turquoise, deviations from expected (computed as in fig. 2) based on the true MSA, when the branch lengths of each of the competing topologies were optimized, denoted True+BBL, where BBL stands for best branch lengths.

A Parametric Solution to Reduce Bias

The strong impact of the guide tree on the KH test led us to seek a solution in which the guide tree is integrated into the test hypothesis. We thus present here a novel parametric test scheme for comparing two topologies. Our test utilizes the concept of parametric bootstrap, as suggested in the SOWH test (Swofford et al. 1996; Goldman et al. 2000). However, unlike the SOWH test, in this newly proposed test, alignment aspects are explicitly accounted for, and the above described biases due to guide tree and alignment uncertainty (and branch length optimization) are removed.

As in the SOWH test, the null hypothesis states that T_2 is the correct topology. Rejection of T_2 suggests that T_1 is better supported by the data compared with T_2.

The first step in the SOWH test is to infer the set of parameters under the null hypothesis. In the SOWH test, this is a straightforward task because the same alignment is assumed for both T_1 and T_2. However, as we have shown above, this is problematic as tree comparison tests are sensitive depending on which topology guided the alignment.

Our proposed one-sided test examines the possibility of rejecting T_2 while taking into account alignment uncertainty and branch length optimization. Here, two alignments are constructed, based on each of the competing topologies, that is, two alignments obtained using the two different competing phylogenies as guide trees. Branch lengths for each topology are optimized with regards to its corresponding MSA. Let T_1, t_1, and MSA_1 be the topology, optimized branch lengths, and MSA of topology 1, respectively (T_2, t_2, and MSA_2 for topology 2). As the components of such a triplet are highly dependent on each other, we propose an iterative procedure to estimate these components simultaneously.

The Iterative Optimization Procedure

1. Start with two competing topologies: T_1 and T_2. For each of the topologies, set arbitrary branch lengths to generate initial branch length estimates: $t_1^{(1)}$ and $t_2^{(1)}$.
2. Use $t_1^{(1)}$ and $t_2^{(1)}$ to guide the two alignments under each topology. Denote the resulting alignments $MSA_1^{(1)}$ and $MSA_2^{(1)}$.
3. Optimize the branch lengths of T_1 (T_2) with regards to $MSA_1^{(1)}$ ($MSA_2^{(1)}$) to obtain $t_1^{(2)}$ ($t_2^{(2)}$).
4. Use the obtained trees (topology and branch lengths) to recompute the alignments.
5. Repeat steps 3–4 until no significant increase in log-likelihood is achieved.

The null hypothesis is then tested based on the obtained triplets (T_1, t_1, MSA_1) and (T_2, t_2, MSA_2). Without loss of generality, we assume that topology 1 has the higher log-likelihood. Denote δ as the log-likelihood difference between T_1 and T_2 (each with regards to its own MSA). H0: T_2 is better

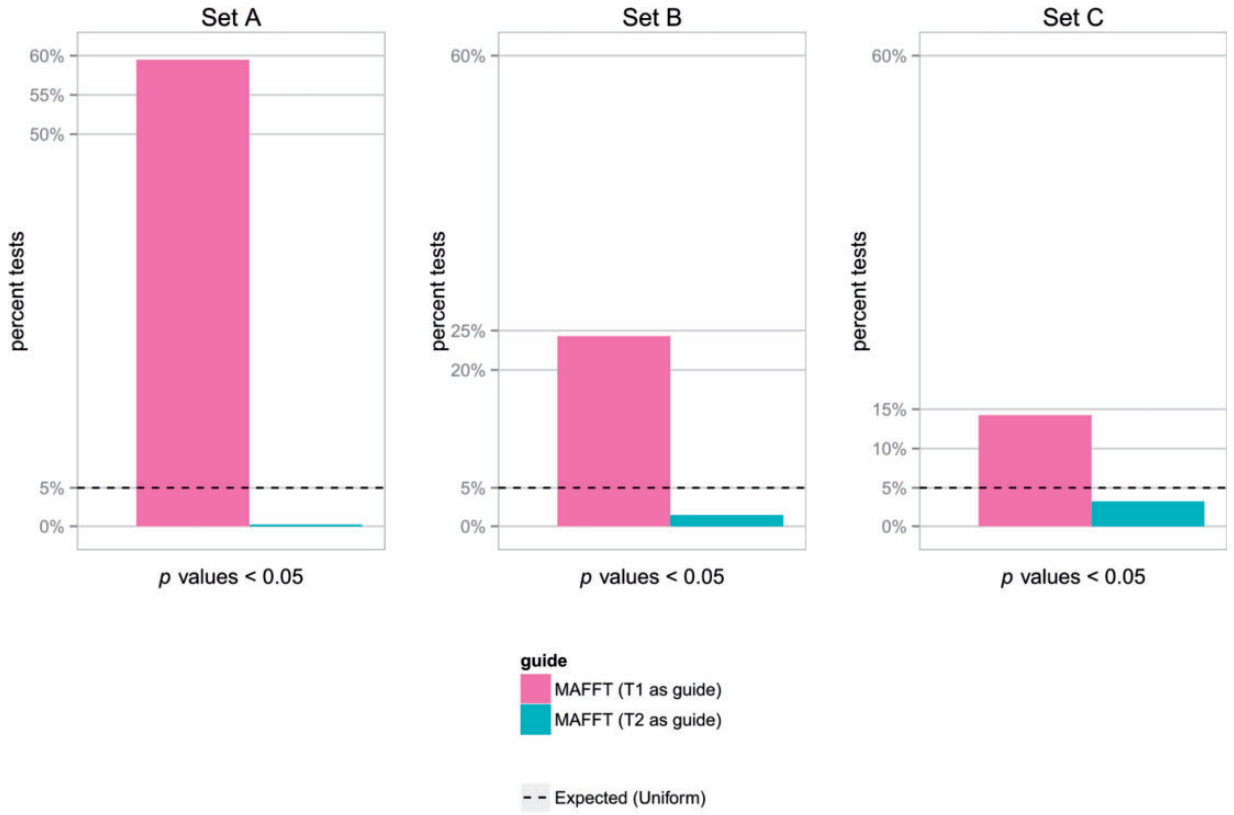

FIG. 4. Testing the bias of KH test under the null conditions. Sequences were simulated using INDELible based on the multifurcating tree in each of the three sets presented in figure 1. Under these conditions, 5% of the obtained *P* values should be smaller than 0.05. Shown here is the percent of *P* values smaller than 0.05 obtained when computations were based on the MSA inferred using MAFFT with either of the competing topologies given as guide tree. In pink, (T_1, t_1) was provided as guide tree; in turquoise, (T_2, t_2) was given as guide tree.

or equally supported by the data (rejecting H0 would indicate that T_2 can be rejected).

The Test Procedure

1. Simulate N sequence data sets according to (T_2, t_2). Repeat the iterative optimization procedure (steps 1–5 above) for each of the simulated data sets. Obtain the empirical distribution of log-likelihood differences.
2. Calculate the *P* value: The proportion of simulations for which the log-likelihood difference was greater than δ.
3. Reject the null hypothesis if the obtained $P < \alpha$.

Note that this procedure is computationally intensive and thus, in practice we used $N = 100$ and the iterative procedure was limited to constructing initial alignments, full optimization of branch lengths, recomputing the alignments, and calculating the log-likelihood of the topology and previously obtained branch-lengths with respect to the recomputed alignment.

Testing the Iterative Parametric Test

As a correct implementation of parametric bootstrapping principles, the proposed iterative parametric test is guaranteed to have approximately correct type I error (i.e., for a confidence level $\alpha = 0.05$, the null hypothesis is wrongly rejected in 5% of the cases). In order to test the power of the iterative parametric test, we examined the test's ability to reject T_2 for data simulated by (T_1, t_1). To this end, we ran the test 100 times on simulated sequence data and examined the proportion of *P* values <0.05. This was performed three times with increasing difficulty to reject the null hypothesis. This was achieved by simulating the sequence data according to T_1 with different associated branch lengths (fig. 1). The results suggest that at least for the topologies and branch lengths tested here, the power ranges between 39% for the most difficult case and 57% for the easiest case (table 2). Comparing the results to the power analysis in table 1, it is evident that removing the bias comes with decreased power.

The power analysis described above is based on a specific set of parameters used for the simulations. For example, we used root sequences of length 2,000 amino acids and a relatively high indel rate of 0.1. As expected, reducing the sequence length decreases the test's power (supplementary table S1, Supplementary Material online). In addition, we studied the effect of the indel rate parameter and the number of sequences on the test's power. In general, when the number of sequences is increased, the alignment becomes longer as a result of additional gaps being opened. We expect that the test's power in such cases will be reduced. Our simulations reveal that indeed, the addition of more sequences decreases power for high indel rate values. However, when the indel rate is low, the test's power increases, reaching a power

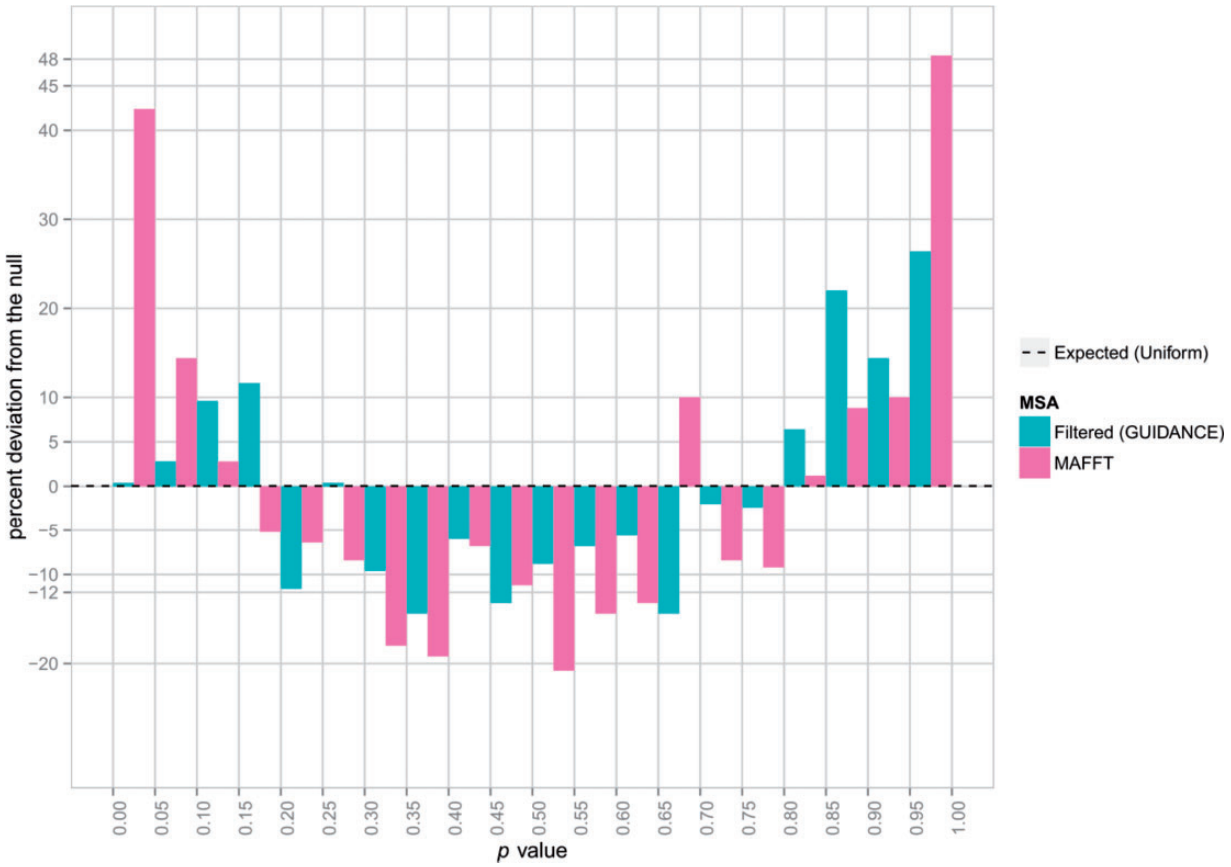

FIG. 5. Testing position filtering as a method for bias reduction in the KH test. Sequences were simulated using INDELible based on the multifurcating tree in set A of figure 1. Under these conditions, among the 5,000 simulations, 5% of the *P* values should fall in each of the 20 *P* value bins, as expected from a uniform distribution. Shown is the deviation from expected, computed as the difference between the observed number of *P* values in each bin minus the expected number (250), divided by the expected number times 100. In turquoise, computations were based on the GUIDANCE filtered MSA, whereas in pink the computations were based on the MSA inferred using MAFFT.

Table 1. Effect of Column Filtering on Power.

Set	A			B			C		
MSA	True	MAFFT	GUIDANCE	True	MAFFT	GUIDANCE	True	MAFFT	GUIDANCE
Power (%)	100	100	99.58	99.94	99.20	71.82	89.20	75.22	34.62

NOTE.—Percent cases in which T_2 was rejected in data sets simulated by T_1.

Table 2. Parametric Test's Power.

Data Simulated by Set	A (easy case)	B (medium case)	C (hard case)
Power (%)	57	41	39

NOTE.—Percent cases in which T_2 was rejected in data sets simulated by T_1.

of 100% for alignments without indels (supplementary table S2, Supplementary Material online).

Multiple Topologies Iterative Parametric Test

The iterative parametric test presented above can be extended to deal with *k* topologies, one of which is the most likely tree topology obtained through a search in the tree topology space. Notably, this scenario is commonly used in practice when constructing a confidence set of tree topologies and analyzing a set of sequences. The challenge in constructing tests for multiple topologies obtained in a tree search procedure is to account for the selection bias (Goldman et al. 2000; Shimodaira 2002): The observation that the log-likelihood of the maximum-likelihood tree topology will always be at least as large as that of a fixed T_1 making larger KH test statistics likely, even under the null hypothesis.

We propose the following parametric test procedure for comparing *k* topologies:

1. The input to this test is a list of *k* candidate topologies, with initial branch lengths, obtained through a tree search procedure.

2. For each topology, perform the iterative optimization procedure described above to get its associated alignment, its optimal branch lengths, and recompute the log-likelihood score. Rank the tree topologies according to their newly computed log-likelihood score and for each topology, keep the difference between the maximum-likelihood score and the topology's score. This value serves as the test statistic to determine significance.
3. Based on each topology (and its associated optimal branch lengths), simulate a set of sequences. Repeat this simulation N times. For each set of simulated sequences, repeat step 2 above to obtain the empirical distribution under the null (the data can be explained by the current tested topology).
4. Compare the difference calculated in step 2 with the empirical distribution of differences obtained in step 3, to compute a P value for each topology.
5. Reject all topologies whose P value is smaller than α.

As this test is extremely computationally intensive, we implemented a few numerical approximations to reduce run times. In addition to those described above for the pairwise test, we also determine the maximum-likelihood tree topology (both on the real data and on each of the simulated data sets) without the reranking stage.

Examining the Multiple Topologies Iterative Parametric on BRCA2

To illustrate our procedure, we performed the test on ten primate sequences for the BRCA2 gene. Although the primate tree is considered to be known (Perelman et al. 2011), our purpose is to demonstrate the application of our test in typical research scenarios in which the goal is to find the species tree based on a small sample of sequences. Further, we show how the test can be used to statistically determine which topologies can be rejected and which cannot.

A bootstrap analysis suggested five candidate topologies that are supported by the data to some extent (see Materials and Methods). We next examined our proposed multiple topologies iterative parametric test procedure on these five topologies. Among these five topologies the maximum-likelihood tree was indeed the known species tree. At a confidence level $\alpha = 0.05$, we were able to reject one of the four incorrect topologies.

We next compared this performance with the popular nonparametric AU test. To this end, we used PhyML to optimize the branch lengths of each of the five competing topologies with regards to the MAFFT alignment. Finally, we used the CONSEL package to perform the AU test on these data. Similarly to our proposed test, at a confidence level of $\alpha = 0.05$, the P values calculated by the AU test allowed the rejection of one of the four incorrect topologies. It should be noted, though, that the topology rejected by the two tests was not the same one; using our proposed test topology 4 was rejected, whereas using the AU test topology 1 was rejected (fig. 6).

Discussion

In this study, we have shown that alignment accuracy and branch length optimization procedures have a major effect on the KH test. These results indicate that much like other downstream analyses, such as detecting positive selection and ancestral sequence reconstruction, alignment reliability is an aspect that should not be ignored when performing likelihood-based tests for phylogeny comparisons.

We first proposed and examined a nonparametric solution that uses GUIDANCE to filter unreliable alignment positions before performing the KH test. Though this fast and simple method reduces the bias significantly, it does not rid of it altogether. In addition, alignment filtering causes a severe loss in data, which was evident in the reduced power observed when using GUIDANCE in our simulations.

As demonstrated in this article and in previous studies (Liu et al. 2009), reconstructing the MSA and inferring the correct phylogeny are two tasks which heavily depend on each other. For this reason, our proposed parametric solution is designed to break the linear pipeline in which the alignment is first computed and only then the phylogeny is inferred. This is achieved by treating the MSA and the phylogeny as two sides of the same coin, optimizing them with regards to each other in an iterative procedure, and obtaining alternative pairs of phylogenies and MSAs. In addition to breaking this circularity, the parametric procedure is free from the reported bias.

However, it should be noted that this procedure is computationally intensive as it requires several stages of branch length optimizations and alignment recalculations both on the real data and on each of the simulated data sets. To reduce the run time, we introduced several numerical approximations to the general test scheme. Mainly, we performed a single step of branch length optimization instead of waiting for full convergence. As this approximation is performed for all competing topologies, this approximation most likely does not favor any specific topology and thus does not create a bias in the test. Moreover, on a few test examples we examined, we found that full convergence was achieved after 3–4 steps (data not shown), suggesting that this approximation is reasonable.

Naturally, an important aspect of the parametric test is the parametric model and its assumptions. In this study, we used INDELible as a tool to simulate the sequences. INDELible offers control over many parameters by which the sequences should be simulated. Among these parameters are the indel rate and maximal length. Currently, there is no methodology to properly estimate these parameters from the data. Such a methodology will enable performing the parametric test with accurate parameters, which we believe, will increase the power of the parametric test.

Materials and Methods

Simulation Scheme

General Scheme

In each simulation, we generated sequence data, aligned the generated set of sequences, and computed the log-likelihood of each of two bifurcating trees with regards to each position

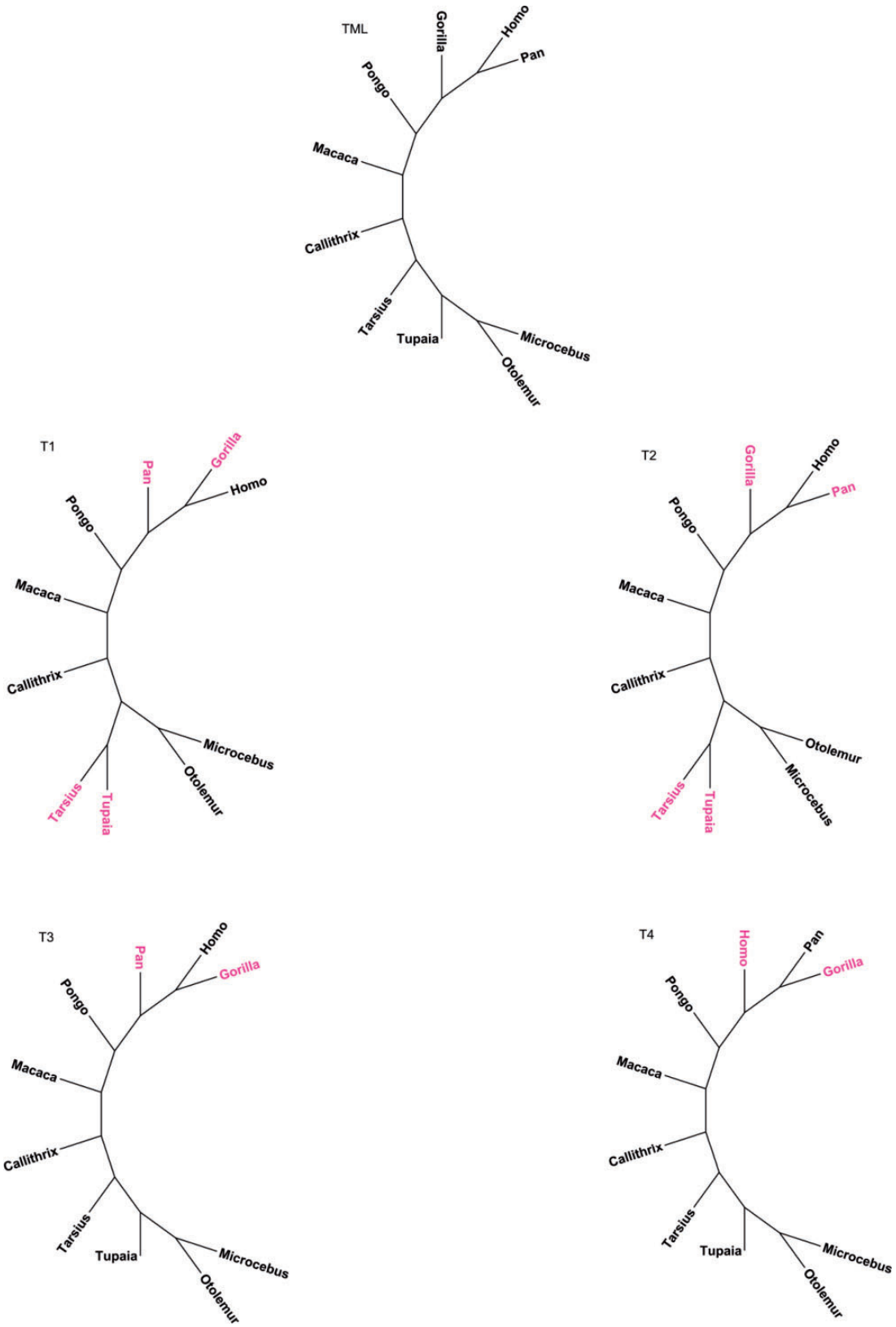

Fig. 6. Competing primate topologies. Five competing topologies inferred based on the MAFFT MSA of the BRCA2 protein sequences. The ML topology is the correct topology. The four alternative topologies were obtained by bootstrap analysis. Differences of each topology from the ML topology are highlighted in pink.

in the calculated alignment. We then either performed the KH test on the calculated log-likelihoods to obtain a *P* value or computed the difference in the total tree log-likelihoods.

Tree Sets

Three sets of trees were used in this study (fig. 1). Each of these sets is composed of a multifurcating tree T and two bifurcating trees, (T_1, t_1) and (T_2, t_2) which are symmetric with regards to T. In set *A*, the branch that differs between T_1 and T_2 is long (length = 0.3), thus T_1 and T_2 are very different from each other in this set. In set *B* we have shortened the branch that differs between T_1 and T_2 (length = 0.1), making T_1 and T_2 resemble each other more. In set *C*, the differing branch is the shortest (length = 0.05), which means that T_1 and T_2 are most similar to each other in this set.

Sequence Generation

In each of the performed simulations, INDELible (Fletcher and Yang 2009) was used to simulate the course of evolution along a tree. When performing the KH test under the null hypothesis conditions, the multifurcating topology T was used with associated branch lengths. For the power calculations and for the parametric test, T_1 and T_2 were used to simulate the course of evolution. An INDELible simulation results in a set of simulated sequences, as well as in the correct (true) MSA. Each INDELible simulation was performed with the following parameters: 2,000 amino acid positions at the root and the JTT substitution model (Jones et al. 1992). All simulations except for the simulations designed to test the bias in KH caused by optimizing branch lengths included the following parameters as well: Gamma-distributed among-site rate variation (proportion of invariant positions = 0, alpha = 0.7, and 16 rate categories), power law model of indel distribution ($a = 1.7$ and $M = 30$), and an indel rate of 0.1 (INDELible's default parameters). Empirical distributions of *P* values were obtained by repeating the simulation 5,000 times.

Alignments

In this study, we examined five methods to align the generated set of sequences:

1. INDELible true alignment.
2. MAFFT (Katoh and Standley 2013) alignment with the following parameters: localpair, maxiterate = 1,000.
3. MAFFT alignment guided by T_1 with following parameters: localpair, maxiterate = 1,000 and treein = (T_1, t_1).
4. MAFFT alignment guided by T_2 with following parameters: localpair, maxiterate = 1,000 and treein = (T_2, t_2).
5. GUIDANCE (Penn, Privman, Ashkenazy, et al. 2010) filtered alignment with following parameters: msaProgram = MAFFT, localpair, maxiterate = 1,000; with column filtering set to default (0.93).

Log-Likelihood Calculation

The log-likelihood for each of the two bifurcating topologies with regards to each position in each of the alignments was calculated using PhyML (Guindon et al. 2010). In addition, the following parameters were used: Model = JTT, $\nu = 0$, $a = 0.7$, and $c = 16$. The value of the "o" parameter was either set to "n" (no optimizations) when testing for the impact of alignment errors (in this case the branch lengths presented in fig. 1 were used) or it was set to "l" (length optimization) when examining the effect of branch length optimizations and for the iterative parametric test simulations.

KH Test P Value Calculation

For a given MSA, the computed sitewise log-likelihoods for each of the bifurcating trees were given as input to the KH test as implemented in CONSEL (Shimodaira and Hasegawa 2001). Its *P* value output was obtained.

Biological Example: The BRCA2 Gene from Primates

We obtained the ENSG00000139618 BRCA2 sequences for nine primates and Tupaia (outgroup) from OrthoMAM (Ranwez et al. 2007). We unaligned the sequences, translated them into amino acids, and realigned them using MAFFT (no guide tree provided). We then ran PhyML on the alignment with 100 bootstrap replicates to obtain a set of competing topologies supported by the data. This resulted in five topologies (including the ML topology which, in this case, matched the correct primate tree). Phylogenetic trees were visualized using FigTree version 1.4 (http:/tree.bio.ed.ac.uk/software/figtree/, last accessed August 10, 2014).

Supplementary Material

Supplementary figures S1–S3 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This study was supported by an Israeli Science Foundation (ISF) grant 1092/13 to T.P. E.L.K. is a fellow of the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University and a Discovery Grant to E.S. from the Natural Sciences and Engineering Research Council of Canada. The authors thank Haim Ashkenazy for insightful remarks and two anonymous reviewers for their valuable comments and suggestions to improve the manuscript.

References

Capella-Gutierrez S, Gabaldon T. 2013. Measuring guide-tree dependency of inferred gaps in progressive aligners. *Bioinformatics* 29: 1011–1017.

Daubin V, Moran AN, Ochman H. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science* 301:829–832.

Fletcher W, Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol.* 26:1879–1888.

Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol.* 27:2257–2267.

Goldman N, Anderson PJ, Rodrigo GA. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst Biol.* 49:652–670.

Guindon S, Dufayard FJ, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.

Jones DT, Taylor RW, Thornton MJ. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8: 275–282.

Jordan G, Goldman N. 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol.* 29:1125–1139.

Katoh K, Standley MD. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772–780.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol.* 29: 170–179.

Kishino H, Miyata T, Hasegawa M. 1990. Maximum-likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol.* 31: 151–160.

Landan G, Graur D. 2009. Characterization of pairwise and multiple sequence alignment errors. *Gene* 441:141–147.

Lister AM, Edwards JC, Nock AD, Bunce M, van Pijlen AI, Bradley GD, Thomas GM, Barnes I. 2005. The phylogenetic position of the 'giant deer' *Megaloceros giganteus*. *Nature* 438:850–853.

Liu K, Raghavan S, Nelesen S, Linder RC, Warnow T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324:1561–1564.

Nelesen S, Liu K, Zhao D, Linder RC, Warnow T. 2008. The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses. *Pac Symp Biocomput.* 13: 25–36.

Ogden TH, Rosenberg SM. 2006. Multiple sequence alignment accuracy and phylogenetic inference. *Syst Biol.* 55:314–328.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010. GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res.* 38:W23–W28.

Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol.* 27:1759–1767.

Perelman P, Johnson EW, Roos C, Seuánez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A molecular phylogeny of living primates. *PLoS Genet.* 7:e1001342.

Privman E, Penn O, Pupko T. 2012. Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol Biol Evol.* 29:1–5.

Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak KM, Douzery JE. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol Biol.* 7:241.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.

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

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol.* 16: 1114–1116.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17: 1246–1247.

Swofford DL, Olsen GJ, Waddell JP, Hillis MD. 1996. Phylogenetic inference. In: Hillis DM, Moritz C, Mable KB, editors. Molecular systematics. Sunderland (MA): Sinauer. p. 407–514.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 56:564–577.

Teeling EC, Scally M, Kao JD, Romagnoli LM, Springer SM, Stanhope JM. 2000. Molecular evidence regarding the origin of echolocation and flight in bats. *Nature* 403:188–192.

Toth A, Hausknecht A, Krisai-Greilhuber I, Papp T, Vagvolgyi C, Nagy GL. 2013. Iteratively refined guide trees help improving alignment and phylogenetic inference in the mushroom family Bolbitiaceae. *PLoS One* 8:e56143.

Wang LS, Leebens-Mack J, Kerr Wall P, Beckmann K, dePamphilis WC, Warnow T. 2011. The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE/ACM Trans Comput Biol Bioinform.* 8:1108–1119.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.