www.sciencemag.org/cgi/content/full/324/5934/1561/DC1

# Supporting Online Material for

## Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees

Kevin Liu, Sindhu Raghavan, Serita Nelesen, C. Randal Linder, Tandy Warnow*

*To whom correspondence should be addressed. E-mail: tandy@cs.utexas.edu

Published 19 June 2009, *Science* **324**, 1561 (2009)
DOI: 10.1126/science.1171243

**This PDF file includes:**

Materials and Methods
SOM Text
Figs. S1 to S10
Tables S1 to S34
References

# Supporting Online Materials

## Materials and Methods

### Sequence Data Generation and Acquisition

We used both simulated and biological sequence data in our analyses.

The simulated data were generated in two steps. Step 1 generated model trees with r8s (*S1*) version 1.7 and Step 2 generated sequence data on those trees using ROSE (*S2*).

**Step 1: Generation of model trees using r8s**

We used r8s to generate random birth-death model trees with 100, 500, or 1000 taxa using the following script:

```
begin rates;
simulate diversemodel=bdback seed=<integer random seed>
ntaxa=<100 or 500 or 1000> T=0;
describe tree=0 plot=tree_description;
end;
```

We then used a custom program (a modified version of tds2, originally developed by Daniel Huson and available at http://www.cs.utexas.edu/users/tandy/science-paper.html) that deviated the r8s trees from ultrametricity using the technique described in (*S3*) with deviation factor $c$ equal to 2.0. Next, we scaled all branches on the model trees by a variable factor, which we call "tree height". Finally, since ROSE cannot handle fractional branch lengths for its GTR simulation, we scaled all the branches in the model tree by a factor of 100 to ensure that the branch lengths were generally greater than 1.

**Step 2: Simulation of sequences using ROSE**

We simulated evolution using ROSE version 1.3 to evolve sequences with indels and substitutions on the model trees. For each model, we generated 20 different model trees (replicates), and for each model tree, we simulated one dataset, starting with a root DNA sequence of 1000 sites.

- Model of evolution - We used the GTR+Gamma model (*S4*) for site evolution with parameters (frequency of nucleotides at the root and the instantaneous rate matrix) obtained by estimating GTR+Gamma parameters on the NemATOL (*S5*) alignment of 682 species of nematodes, using PAUP* (*S6*). We used the ROSE transitional probability matrix $P(t) = e^{At}$ where $t = .001$ and $A$ is the transitional rate matrix.

  The frequencies of the nucleotides at the root were given by
  `TheFreq=[.300414,.191363,.196748,.311475]`

  and the transitional rate matrix for the GTR model was given by the following off-

diagonal entries:

```
A-C 1.24284
A-G 3.47484
A-T 0.48667
C-G 1.07118
C-T 4.38510
G-T 1.0
```

We set the shape parameter $\alpha$ of the gamma distribution, controlling rate variation across sites, to $1.0$.

- Gap length distribution - We used three geometric single-event gap-length distributions: short, medium, and long, each with finite tails. The long gap distribution had expected gap length 9.2 and median gap length 7; the medium gap distribution had expected gap length 5.0 and median gap length 4; and the short gap distribution had expected gap length 2.0 and median gap length 2. The gap length distributions used in our study are given below. The first element of each list is the probability of a gap of length one, given that a gap event occurs, the second is the probability of a gap of length two given a gap event occurs, and so on.

  Long Gap Length Distribution:

  ```
  [0.1028, 0.0899, 0.0792, 0.0702, 0.0627, 0.0565, 0.0514, 0.0470,
  0.0433, 0.0400, 0.0369, 0.0341, 0.0314, 0.0289, 0.0266, 0.0245,
  0.0225, 0.0206, 0.0188, 0.0171, 0.0155, 0.0141, 0.0127, 0.0114,
  0.0100, 0.0087, 0.0075, 0.0063, 0.0052, 0.0042]
  ```

  Medium Gap Length Distribution:

  ```
  [0.2012, 0.1600, 0.1280, 0.1024, 0.0819, 0.0655, 0.0524, 0.0419,
  0.0336, 0.0268, 0.0215, 0.0172, 0.0137, 0.0110, 0.0088, 0.0070,
  0.0056, 0.0045, 0.0036, 0.0029, 0.0023, 0.0018, 0.0015, 0.0012,
  0.0009, 0.0008, 0.0006, 0.0005, 0.0004, 0.0003, 0.0002]
  ```

  Short Gap Length Distribution:

  ```
  [0.4613, 0.2527, 0.1545, 0.0896, 0.0419]
  ```

- Probability of a gap event - We set insertion and deletion probabilities identically. The gap event probabilities we used for each model are in Table S1.

The ROSE script used to simulate data is given in Figure S1. Table S1 gives the parameters used to simulate each dataset, and Tables S2, S3, and S4 give the resulting true alignment statistics. The empirical statistics for a single replicate of the 100-taxon datasets (used in Experiment 6) is given in Table S5. Finally, an informative gap is defined as a gap with sites having between $2$

and $t - 2$ indels, where $t$ is the number of taxa. Table S6 gives informative gap information for the 500 and 1000-taxon datasets.

**Biological datasets**

We obtained several curated rRNA datasets from Robin Gutell's comparative RNA database (*S7*). These datasets are among the few biological nucleotide datasets for which there are highly reliable alignments – Gutell's alignments are obtained on the basis of RNA secondary structure – and therefore permit evaluation of the accuracy of an estimated alignment, and hence of an estimated tree.

We used datasets having more than 100 taxa and having fewer than 10 million cells in the curated alignment (since these fall within the dataset size of SATé's current scope). We excluded the 5S rRNA and tRNA datasets because the curated alignments had short sequence lengths. We also excluded the intronic datasets since curation is more reliable on non-intronic datasets.

We then "cleaned" the alignments by removing columns that contained only indel letters, and then removing taxa with more than 50% unsequenced letters. Table S7 gives the original curated alignment statistics on columns and taxa that were cleaned. Using these cleaned alignments, we produced corresponding unaligned sequence files by removing indel letters and computed ClustalW (*S8*), MAFFT (*S9*), Muscle (*S10*) and Prank+GT (*S11, S12*) alignments on the unaligned sequence files. We computed the alignment error (SP-FN) for each of these estimated alignments using the curated alignment as the reference alignment, and then ordered the datasets by average alignment error. We chose the 6 datasets with highest average alignment error, since these had much more error than the remaining datasets.

In order to assess the accuracy of the trees estimated on the biological datasets, we computed three reference trees by running RAxML (*S13*) on the curated alignment as follows. We performed 500 bootstrap replicates to determine the high support branches, and then created three reference trees of differing minimum bootstrap support: 50% bootstrap, 75% bootstrap and 90% bootstrap support. Table S8 gives the curated alignment statistics as well as the percent resolution of the 50%, 75% and 90% bootstrap reference trees.

**Two-phase alignment and tree generation**

**Alignment Generation**

Each dataset was aligned using ClustalW (*S8*) version 2.0.4, MAFFT (*S9*) version 6.240, Muscle (*S10*) version 3.7, and Prank+GT (*S11, S12*) (Prank version 080904). Prank+GT uses the RAxML (*S13*) tree on the MAFFT alignment as the guide tree to a Prank alignment estimation. The following commands were used:

- **ClustalW**:
  `clustalw2 -align -infile=<infile>`
  `-outfile=<output> -output=fasta`

- **MAFFT L-INS-i**:
  `mafft --localpair --maxiterate 1000 --quiet <input> > <output>`
- **Muscle**:
  `muscle -in <input> -out <output> -quiet`
- **Prank+GT**:
  `prank -d=<input> -o=<output> -t=<raxml(mafft) guide tree>`
  `-noxml -notree -nopost +F -matinitsize=5 -uselogs`

**Masked Alignment Generation**
We used four different masked alignments for each simulated dataset. We ran Gblocks version 0.91b (*S14*) in its two modes: relaxed and stringent. The commands we used were:

- Relaxed version
  `./Gblocks <input FASTA alignment> -t=d`
- Stringent version
  `./Gblocks <input FASTA alignment> -t=d`
  `-b2=<number of taxa * 0.5625> -b3=10 -b4=5 -b5=H`

We also produced two different masked alignments by eliminating columns from estimated alignments in which the proportion of taxa having gaps in the column was greater than 75% or 50%, respectively.

**Tree Generation**
Maximum likelihood trees were estimated on each estimated alignment and the true alignment using RAxML, version 7.0.4, with the following command:

- **RAxML default**:
  `raxmlHPC -m GTRMIX -w <working dir> -n <identifying suffix>`
  `-s <input PHYLIP file>`

To save a checkpoint best ML tree to disk (which allows us to implement a timeout function), we used the following command:

- **RAxML checkpoint to disk**:
  `raxmlHPC -m GTRMIX -w <working dir> -n <identifying suffix>`
  `-s <input PHYLIP file> -j`

All two-phase RAxML tree estimations were run to completion.

**SATé Tree and Alignment Generation**
Each run of SATé outputs a tree/alignment pair, and when multiple runs are performed, the

tree/alignment pairs from each are compared and the one with the best ML score is kept. We call the multiple run version SATé$^{BML}$.

An individual SATé run has the following structure. Stage 1 selects the tree/alignment pair for use in the second stage, and then Stage 2 continues with the hill-climbing strategy. In our experiments, the selection of the starting tree/alignment pair is typically performed by evaluating four different automated alignment methods, estimating ML trees on each using RAxML, and then keeping the one that had the best ML score.

The second stage is based on the CT-$i$ divide-and-conquer strategy. This strategy has two parameters: a branch around which to decompose the current best tree, and the decomposition level $i$. The branch is drawn from a collection of branches, including the "centroid branches" (that divide the tree roughly into equal numbers of taxa), the "midpoint branch" of the longest paths in the tree, and branches that are adjacent to these branches. The decomposition level $i$ specifies the maximum diameter of the center-tree computed. This center-tree defines the subproblems by separating the taxa into clades when the center-tree is removed. Taxa in a clade are then grouped into a subproblem. See Figure S2 for a CT-3 example.

Different CT-$i$ divide-and-conquer analyses are attempted, each based upon a different branch, until one yields an improvement in the ML score or SATé times out. In our analyses, we first used a midpoint branch, then a centroid branch, and then other branches adjacent to the centroid branch. In general, when improvements to the ML score could be obtained, they were obtained from either a centroid or midpoint branch. We limited all second stage RAxML tree estimations to six hours.

For each stage there are several options, which result in different variants of SATé. The options include:

- **Initial tree chosen to pass to Stage 2.** By default, SATé computes alignments using methods that can analyze datasets with up to 1000 sequences: ClustalW, MAFFT, Muscle, and Prank+GT, which is Prank given the RAxML-on-MAFFT tree as a guide tree. For each estimated alignment, we ran RAxML under the GTRMIX model, and kept the alignment/tree pair that had the best ML score as input for the second stage. The SATé(C) variant uses only a RAxML-on-ClustalW starting tree.

- **The CT-i decomposition.** By default, SATé uses CT-5 decompositions and therefore divides the dataset into at most 32 subsets. We also performed experiments using CT-1, CT-2, CT-3 and CT-4 decompositions.

- **Stopping criterion for Stage 2.** By default, the second stage of SATé is terminated by a time limit of 24 hours, but any CT-$i$ divide-and-conquer analysis initiated before the 24 hour time limit is allowed to complete. Alternatively, the user can instead run SATé until no proposal results in an improved maximum likelihood score after 24 hours of search, which we call SATé*.

Our other defaults include MAFFT for subproblem alignment, Muscle for merging subalignments into an alignment on the entire set, and RAxML for ML tree estimation.

Many of our experiments focused on the following variants of SATé:

**SATé$^{24}$:** All the default settings for SATé are used (this is sometimes referred to as SATé).

**SATé$^{*}$:** All defaults are used, but the second stage continues until there has been no change in the ML score for 24 hours.

**SATé$^{24}$(C):** The ClustalW alignment/tree pair is used for the start of Stage 2; otherwise all the default settings are used.

**SATé$^{*}$(C):** The ClustalW alignment/tree pair is used for the start of Stage 2; the second stage runs until no change in the ML score has occurred for 24 hours; all other settings are the defaults.

Finally, we report results for a variant we call SATé$^{BML}$ (best maximum likelihood), which takes a collection of outputs from two or more other SATé variants, and returns the alignment/tree pair with the best maximum likelihood. Since this method depends upon the particular collection of SATé variants run, we give specific information about this variant for each experiment in which we ran it.

**ALIFRITZ, BAli-Phy and BEAST Tree and Alignment Generation**
We evaluated the performance of ALIFRITZ (*S15*), BAli-Phy (*S16*), and the method of Lunter *et al.* (*S17*) that is implemented in BEAST (*S18*) on a collection of 100-taxon datasets. We ran BAli-Phy version 2.0.1 with default settings with the following command:

```
baliphy --data-dir=<BAli-Phy installation data directory>
        --name <working directory> <unaligned FASTA input>
```

We ran ALIFRITZ version 1.0 under default settings for 10 million steps with a RAxML-on-ClustalW starting tree and alignment using the following command:

```
alifritz <configuration script> 2
```

The configuration script for ALIFRITZ is given in Figure S3.

Although we were successful in starting ALIFRITZ and BAli-Phy estimations on these datasets, we were unable to run Lunter *et al.*'s method within BEAST because a fundamental ClassCastException occurred which could not be resolved by the time of this writing. We attempted to run BEAST with the following command:

```
beast <XML configuration file specifying input dataset>
```

We ran BEAST on multiple datasets, including the small 10-taxon dataset originally published by Lunter et al. (*S17*) (available online as part of their online supplemental material). All

BEAST runs immediately terminated abnormally with the Java exception message shown in Figure S4.

**Performance evaluation**

For each simulated dataset, we compared all estimated trees to the PIMT. We calculated the missing branch rate for each estimated tree, i.e., the percentage of internal branches that appeared in the PIMT but did not appear in the estimated tree. We also compared all estimated alignments to the true alignment, and recorded the SP-FN error. For those estimated trees which were not guaranteed to be binary, we also computed the false positive rate, which is the percentage of the internal branches of the estimated tree that do not appear in the PIMT. Additionally, we performed a Spearman rank correlation analysis for the ML score and the missing branch rate and for the ML score and alignment accuracy on each dataset. We used MATLAB to calculate the Spearman rank correlation coefficient and associated p-value statistic for the datapoints in a given correlation calculation, using the commands shown in Figure S5. To test whether SATé's results were better than those of the next best two-phase method, we performed paired t-tests ($n$ = 40 for each test) on missing branch rate and SP-FN error, controlling for multiple tests using the Dunn-Ŝidák method (*S19*).

For the biological datasets, we compared estimated trees to reference trees computed on the curated alignments, where the reference trees are obtained by using bootstrap support values on the tree returned by RAxML. We compared estimated alignments to the curated alignments.

**Computing Resources**

For our experiments, we used three different sets of machines: a Condor (*S20*) pool, a cluster called Mastodon, and two very large memory machines.

Details for each set follow:

- **Condor pool**: This is a heterogeneous cluster of machines using processors generally comparable to a 3.0 Ghz Intel 32-bit CPU with between 512 MB and 4 GB of main memory. This cluster was used to perform all simulations and to run all two-phase methods. Each two-phase method was run to completion.

- **Mastodon cluster**: This is a cluster of machines that have Intel 32-bit CPUs running at either 2.6 Ghz, 2.66 Ghz, or 3.06 Ghz and 4 GB main memory per CPU. These machines were used to run SATé and the first ALIFRITZ and BAli-Phy experiment (Experiment 5 below).

- **Large memory machines**: These two machines are 16-core 64-bit AMD Opteron(tm) 8360 SE processors with either 128 GB or 256 GB shared main memory per machine. They were used to run ALIFRITZ and BAli-Phy in experiment 6 below. Each BAli-Phy run had access to 128 GB of main memory per run, and each ALIFRITZ run had access to 256 GB of main memory per run.

## Experiments

We performed a number of experiments to determine the effects of parameter choices in SATé as well as to compare SATé to alternative methods.

**Experiment 1: SATé$^{24}$ compared to two-phase methods.** SATé$^{24}$ was run on the simulated datasets for the 500 and 1000-taxon model conditions. To understand how SATé$^{24}$ explores alignments and trees, we tracked how many CT-5 proposals occurred during the the second stage and how many of those proposals were accepted. We also computed the accuracy of the trees and alignments for the first accepted Stage 2 tree/alignment pair, and compared these error rates to those of both the starting tree/alignment pair for Stage 2, and the final accepted tree/alignment pair.

We performed several statistical tests to assess whether SATé$^{24}$'s performance was significantly better than other methods. To determine whether SATé$^{24}$returned more accurate trees and alignments than its closest competitor, we performed one-tailed, paired t-tests on SATé$^{24}$ and the two-phase method that came closest, on average, to matching SATé$^{24}$'s performance. To determine whether the correlations between ML scores and missing branch rates and between ML scores and alignment accuracy (SP-FN error) were significant, we computed p-values for their Spearman rank-correlations. All multiple tests were controlled for experimentwise error rate by examining both Bonferroni and Dunn-Ŝidák corrected p-values.

**Experiment 2: Impact of the starting tree** To assess the impact of a starting tree, we modified the default SATé so that Stage 2 started with the RAxML(ClustalW) tree/alignment pair because ClustalW usually produced the least accurate trees and alignments. We recorded the tree found at the end of 24 hours (SATé$^{24}$(C)), and then also continued to run this variant until no improvement in ML score was found in 24 hours (SATé*(C)). Finally, we compared the alignment and tree accuracy of these variants of SATé against the two-phase methods.

**Experiment 3: Biological dataset analysis** For each of the biological datasets we computed ClustalW, MAFFT, Muscle, and Prank+GT alignments, and also computed trees from these alignments using RAxML. We compared the error of these alignments and trees to those computed by SATé. We ran three variants of SATé: SATé*, SATé$^{24}$(C) and SATé*(C).

**Experiment 4: Impact of different CT-i decompositions** We considered the impact of the decomposition level by running a variant of SATé that used CT-1 decompositions instead of CT-5 for each of the simulated datasets. We also considered SATé* variants that used CT-1, CT-2, CT-3, CT-4 and CT-5 decompositions for the six real (biological) datasets.

**Experiments 5 and 6: Evaluating BAli-Phy and ALIFRITZ** We ran two experiments to understand the performance of BAli-Phy and ALIFRITZ. For the first experiment, we used the

datasets in Table S4 and ran both BAli-Phy and ALIFRITZ for 25 days on machines in the Mastodon cluster.

In the second experiment, we used the datasets in Table S5 and ran both BAli-Phy and ALIFRITZ for 337.1 hours on the large memory machines.

**Experiments 7 and 8: Impact of site removal techniques** We explored four techniques for removing unreliable sites from estimated alignments, with respect to the impact on phylogeny estimation. In Experiment 7, we evaluated the Gblocks software introduced by Talavera and Castresana (*S14*), while in Experiment 8 we evaluated techniques based upon removing sites that contain a large proportion of gaps (75% and 50%).

## Supporting Text

**Experiment 1: SATé[24] compared to two-phase methods** Figure S6 shows the results for the 500-taxon model conditions, while Figure 3 (of the main text) shows the results for the 1000-taxon model conditions. Table S9 gives running time information for each stage of SATé[24], and compares the maximum likelihood values from each stage. SATé[24] performs well compared to all two-phase methods, particularly on the moderate-to-difficult models. When compared to the best performing two-phase method, SATé[24]'s performance is not always significantly better in terms of alignment. However, the tree is significantly better for every moderate-to-difficult model except one (1000M3).

Table S10 shows how many proposals were generated during Stage 2, as well as how many of those proposals were accepted. Tables S11 and S12 show the the error rates, for the moderate-to-hard and easy model conditions, respectively, for the first accepted tree/alignment pair, as well as both of these error rates for the starting tree/alignment pair for Stage 2 and the final accepted tree/alignment pair. For moderate-to-difficult datasets, fewer proposals are made, but the proposals are also more likely to be accepted. For the easy model conditions, many more proposals are made, but they are also much less likely to result in an improved ML score, and therefore are usually rejected. The tree and alignment errors change very little for the easy model conditions throughout Stage 2. For the moderate-to-difficult model conditions, the biggest improvement in both tree and alignment error is gained in the first accepted tree/alignment pair of Stage 2 though further improvements are made in tree/alignment pairs that are subsequently accepted.

Table S13 reports the results of the t-tests for both missing branch rates and alignment error rates. Tables S14, S15 and S16 show the p-values for the Spearman rank-correlations for the 1000-taxon moderate-to-difficult, 500-taxon moderate-to-difficult, and easy models, respectively. For all the moderate-to-difficult datasets, the correlation between ML and alignment error was significant, as was the correlation between ML and tree error (except for model 500S3). For the easy datasets, the alignment error was usually correlated with ML score, but the tree error was usually not correlated. Thus, as datasets became more difficult, using ML as a criterion to choose between tree/alignment pairs became more effective.

**Experiment 2: Impact of the starting tree** Figures S7 and S8 compare the tree and alignment errors of SATé$^{24}$(C), SATé*(C), SATé$^{24}$, and SATé$^{BML}$ for the 1000-taxon datasets. Figures S9 and S10 show the same statistics for the 500-taxon datasets. In this experiment, SATé$^{24}$(C) is the result of the first 24 hours of the search by SATe*(C), and SATé$^{BML}$ is the best of the other three methods. We present information on runtime, number of proposals, and number of accepted proposals for each of these variants in Table S17.

We observed that starting with the RAxML(ClustalW) alignment/tree pair had the following effects: (1) For the moderate-to-difficult model conditions, the trees returned after only 24 hours of analysis were generally only slightly less accurate than when SATé began with a better alignment/tree pair. (2) For the easy model conditions, trees found after 24 hours were essentially equally accurate, independent of the starting tree. (3) For SATé*, where SATé was allowed to run until no improvement in ML score was found for 24 hours, the alignment/tree pairs returned were of comparable (or better) accuracy than those found during a 24 hour analysis that began with the better alignments. (4) Using SATé$^{BML}$ produced trees with the best accuracy. Thus, the CT-5 search strategy was effective at exploring alignment space because even when the search began with a poor alignment/tree pair, it eventually produced trees with equivalent or improved topological accuracy.

**Experiment 3: Biological datasets** For each biological dataset, Table S18 reports the alignment error of each two-phase method, as well as the alignment error of the variant of SATé that gave the best maximum likelihood, i.e. SATé$^{BML}$. For these datasets, the MAFFT alignment was always most accurate, but SATé$^{BML}$ never had more than an additional 3% of SP-FN alignment error. However, the slight increase in SP-FN error for these datasets was not reflected in the trends for tree error, as Table S19 shows. In fact, a very different trend appeared for missing branch rates. Though SATé$^{BML}$ was not always the best in terms of alignment error, it was quite often the best with respect to any of the various reference trees when compared with the other two-phase methods and was consistently the most accurate method for each reference tree. SATé$^{BML}$ was best on average when the reference tree was the 90% bootstrap tree and the reference tree edges were most highly supported.

**Experiment 4: The impact of different CT-i decompositions** Results comparing the accuracy of using CT-5 versus CT-1 decompositions are shown in Tables S20 and S21 for the simulated datasets. These tables show that using CT-5 decompositions produced more accurate trees and alignments, as well as better ML scores for all the moderate-to-difficult datasets and trees and alignments of comparable accuracy on the easy datasets, as compared to CT-1 decompositions.

Table S22 shows the log likelihood values and tree errors for using each of CT-1 through CT-5 decompositions for the biological datasets. Tables S23 and S24 show the alignment error and running time information. These tables show that SATé$^{24}$ based upon CT-5 decompositions produced trees and alignments that were generally as good as those produced by SATé$^{24}$ using other CT-i decompositions, but on occasion slightly smaller values of $i$ produced slightly better

results. However, CT-5 seems to be a reasonably effective search strategy in comparison to the other CT-i searches we used on the real and simulated datasets we examined.

Since all our simulated datasets had 100, 500, or 1000 sequences, and our real datasets had at least 117 sequences, and up to 1028 sequences, this likely limits the generalizability of this observation. It is clear that the selection of the parameter $i$ for the CT-i decomposition must depend upon the dataset size, for the following reason: a CT-5 decomposition can result in 32 datasets. Thus, for small enough datasets, using CT-5 decompositions would amount, essentially, to producing an alignment on the current tree with Muscle only, and treating the current tree as a guide tree. Therefore, for very small datasets (fewer than 50 sequences), we conjecture that smaller values of $i$ would be more suitable than $i = 5$, while on very large datasets (with more than 5000 sequences), larger values of $i$ might be suitable. We do not yet know the consequences of varying $i$, but previous experience with Rec-I-DCM3 (*S21*) (a divide-and-conquer technique used to improve ML searches) strongly suggests that the dataset decomposition needs to be based upon the dataset size.

**Experiments 5 and 6: Evaluating BAli-Phy and ALIFRTIZ** ALIFRITZ and BAli-Phy are each computationally intensive, requiring substantial running time and memory to even begin to return trees and alignments on moderate sized datasets. For example, ALIFRITZ will only return trees after it has performed 1000 iterations. Furthermore, each method is only likely to produce reliable estimates of trees if run to convergence, which is likely to be substantially beyond the few weeks of analysis that our study permitted. For example, the recommended minimum number of iterations for ALIFRITZ on smaller datasets than the ones we examined is 10,000,000. Thus, our investigation here is only suggestive of how well these methods perform if running times are limited.

The first experiment explored performance on the Mastadon cluster, involving machines with only 4G main memory per CPU, and letting the methods run for 25 days. In this experiment, 39 BAli-Phy runs crashed due to memory limitations (Table S25), and none had completed. ALIFRITZ analyses were similarly problematic: none had completed more than 124,000 iterations, which is less than 1.25% of the number of recommended iterations for smaller datasets. Due to the cluster configuration, we were unable to access any intermediate results, such as the trees, number of iterations, or logs for BAli-Phy; we were unable to access any intermediate results except for the number of iterations for ALIFRITZ.

We then performed an experiment to evaluate the performance of BAli-Phy and ALIFRITZ on one dataset from each of the seven 100 taxon model conditions explored in the first experiment, over two weeks of analysis on our large memory machines. In this experiment we stored intermediate results so we could evaluate the topological accuracy of the trees and alignments obtained by these methods. For ALIFRITZ, we report the ML tree and ML alignment found during the two week period. For BAli-Phy, we extracted the majority consensus tree, MAP tree, MAP alignment, and posterior-decoding-estimated alignment (*S22–S24*) from its two-week run results. We also computed trees and alignments on these datasets using two-phase methods and SATé$^{BML}$. Alignment and tree errors for these alignments and trees are reported in Table S26,

Table S27, and Table S28.

BAli-Phy's performance on these seven datasets was mixed. On four datasets (100L2, 100M2, 100M3, and 100S2) it produced excellent results, with estimated trees that are equal in accuracy to those of the best two-phase methods on these datasets. On the other hand, BAli-Phy produced trees that were less accurate than the best two-phase methods on 100L1, 100M1, and 100S1. The four datasets on which it produced good trees were also datasets that we would characterize as "easy", since most of the two-phase methods produced trees nearly as accurate as those produced by RAxML on the true alignment. Thus, BAli-Phy was able to perform well given large memory machines and relatively easy model conditions, but its performance on more difficult datasets was not as good. Since BAli-Phy was far from stationarity on these datasets, we conjecture that it may simply need much more time on these datasets than the two weeks we permitted it. Furthermore, multiple MCMC walks by BAli-Phy on each dataset would be needed to assess convergence.

The performance of ALIFRITZ was generally not as good as BAli-Phy. ALIFRITZ succeeded in reporting trees on only the three easy datasets because on only these had it completed the minimum number of iterations. Furthermore, on these datasets, the trees produced had lower accuracy than those obtained by BAli-Phy. It should be noted that ALIFRITZ was far from convergence, and that the trees it returned should not be considered reasonable estimates.

A comparison of BAli-Phy to SATé$^{BML}$ is interesting. On every dataset, SATé$^{BML}$ produced trees that were of equal or better accuracy than those of BAli-Phy, with substantial improvements on the harder datasets.

Table S29 has running times, and Table S30 shows the number of iterations completed for each BAli-Phy run after 337.1 hours.

**Experiments 7 and 8: Impact of site removal techniques** Tables S31 and S32 show the results for masking using Gblocks stringent and relaxed variants, respectively. Tables S33 and S34 show the results for masking columns 75% and 50% gapped, respectively. The techniques which removed sites based upon the percentage of gaps had very little impact on the phylogenetic tree, while the Gblocks technique actually made all phylogeny estimations less accurate. Thus, the two techniques differed in their impact, but neither resulted in improving phylogenetic estimation. Both techniques removed many sites; the stringent Gblocks technique removed more sites than the relaxed Gblocks technique, which removed more sites than the other simple masking techniques. More generally, the impact on the phylogenetic accuracy of removing sites seems to depend on the number of sites that remain for the phylogenetic analysis, so that if too few sites are left then the resultant phylogenies will tend to have a high missing branch rate.

Talavera and Castresana observed that Gblocks improved phylogenetic estimation, in contrast to our observations here. While the two studies would seem to be contradictory, we suggest that they are not. The model conditions studied by Talavera and Castresana started with a very large number of sites and very few taxa, so that even after masking, a sufficient number of sites remained to ensure an accurate phylogeny. By contrast, our study started with many more taxa (up to 1000) and proportionally fewer sites (only one model condition had more than 5500 sites,

and most had less than 4000 sites).

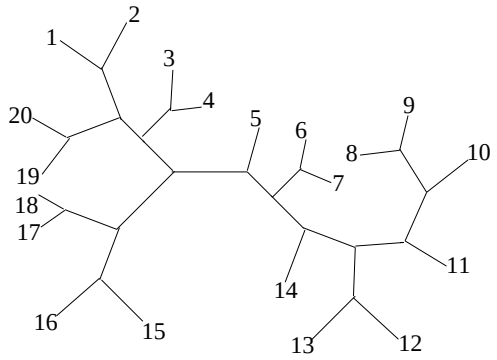
We conjecture that the main reason for Gblock's deletion of such a large proportion of the sites in our case is that our analysis is on datasets with many taxa, which would tend to increase the probability of any single site being masked. By contrast, their study was on datasets with very few taxa but many thousands of sites. As a result, their technique not only removed proportionally fewer sites (due to the smaller number of taxa), but still left a very large number of sites for phylogenetic analysis. Our findings on our datasets are therefore not contradictory, but rather point to the unsuitability of their masking techniques on large datasets with many gaps.

While SP-FN is a standard measurement of alignment error (in that its complement is a standard measurement of alignment accuracy), there are problems with this measurement. For example small reductions in the SP-FN error do not necessarily produce better estimated trees. Also, the SP-FN error rate measures only the percentage of the truly homologous pairs of nucleotides that are missing in the estimated alignment, and so is a false negative rate. Measurements of accuracy that would take false positives into account, and which would include indels, have also been proposed (*S25*), but are not yet standard in the field. For these reasons, although we showed the SP-FN error rate, we consider it a general indication of the quality of the estimated alignment but do not consider small differences in SP-FN error to be necessarily important.

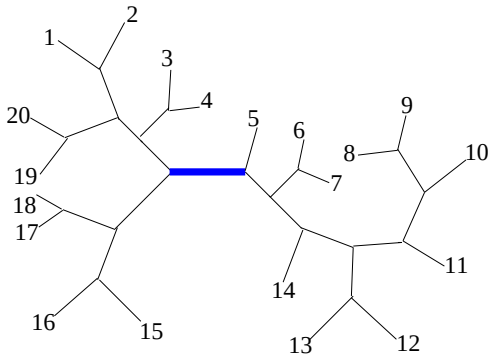
## Supporting Figures

```
TheInsFunc = <Insert event gap length distribution -
               long, medium or short>
TheDelFunc = <Delete event gap length distribution -
               long, medium or short>
InputType = 4
TheAlphabet = "ACGT"
TheFreq = [.300414,.191363,.196748,.311475]
ThePAMMatrix =
[[0.9948, 0.0012, 0.0035, 0.0005],
[0.0012, 0.9933, 0.0011, 0.0044],
[0.0035, 0.0011, 0.9944, 0.0010],
[0.0005, 0.0044, 0.0010, 0.9941]]
TheInsertThreshold = <Insertion event probability>
TheDeleteThreshold = <Deletion event probability>
SequenceLen = 1000
TheTree =<birth-death model tree in Newick format with branch
            lengths, deviated from ultrametricity>
ChooseFromLeaves = False
AlignmentWithAncestors = True
TreeWithAncestors = True
SequenceNum = <99 or 499 or 999>
SeedVal = <random seed integer>
TheMutationProbability=<site-by-site vector listing rate
                          multipliers for that site>
```

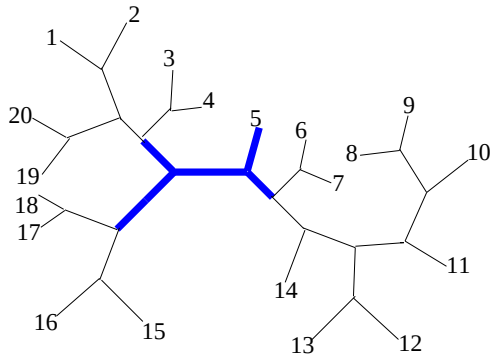
Figure S1: **ROSE script.**



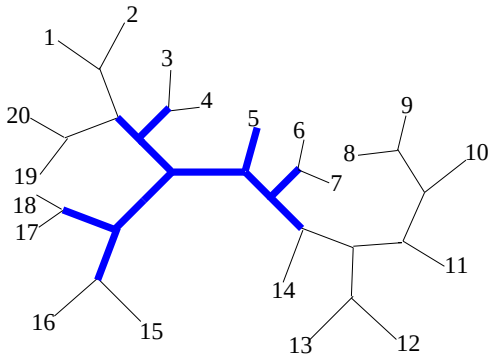

(a) Estimated tree




(b) CT-1 tree in blue




(c) CT-2 tree in blue




(d) CT-3 tree in blue

Figure S2: **A CT-3 decomposition resulting in 7 subproblems.** The subproblems have the taxon sets {1,2,19,20}, {3,4}, {5}, {6,7}, {8,9,10,11,12,13,14}, {15,16}, and {17,18}.

```
SUBST :
model = GENREV
parameter= (1.0,1.0,1.0,1.0,1.0,1.0)
freqs = estimate
INDEL :
model=TKF2
parameter= (0.1,0.11,0.5)
FILES :
infile=<initial FASTA alignment>
format = 0
guidetree=<initial NEWICK tree>
outtree=<output NEWICK tree>
outalign=<output FASTA alignment>
logfile=<log file>
KEEP :
COOL :
burnin=100
maxsteps=10000000
resamprange=60
reportafter=1000
```

Figure S3: **ALIFRITZ configuration script.**

```
Exception in thread "Thread-1" java.lang.ClassCastException:
    dr.evomodel.indel.TKF91Likelihood
    cannot be cast to
    dr.inference.model.CompoundLikelihood
    at dr.inference.markovchain.MarkovChain.chain(Unknown
Source)
    at dr.inference.mcmc.MCMC.chain(Unknown Source)
    at dr.inference.mcmc.MCMC.run(Unknown Source)
    at java.lang.Thread.run(Thread.java:619)
```

Figure S4: **BEAST error message**

```
load <input file in x vs.  y space-delimited columnar format>
x = <input file>(:,1)
y = <input file>(:,2)
[rho pval] = corr(x,y,'type','Spearman');
save('<output file for correlation coefficient>', 'rho','-ASCII')
save('<output file for associated p-value>', 'pval','-ASCII')
```

Figure S5: **MATLAB commands to compute Spearman rank correlation coefficient and associated p-value statistic.**

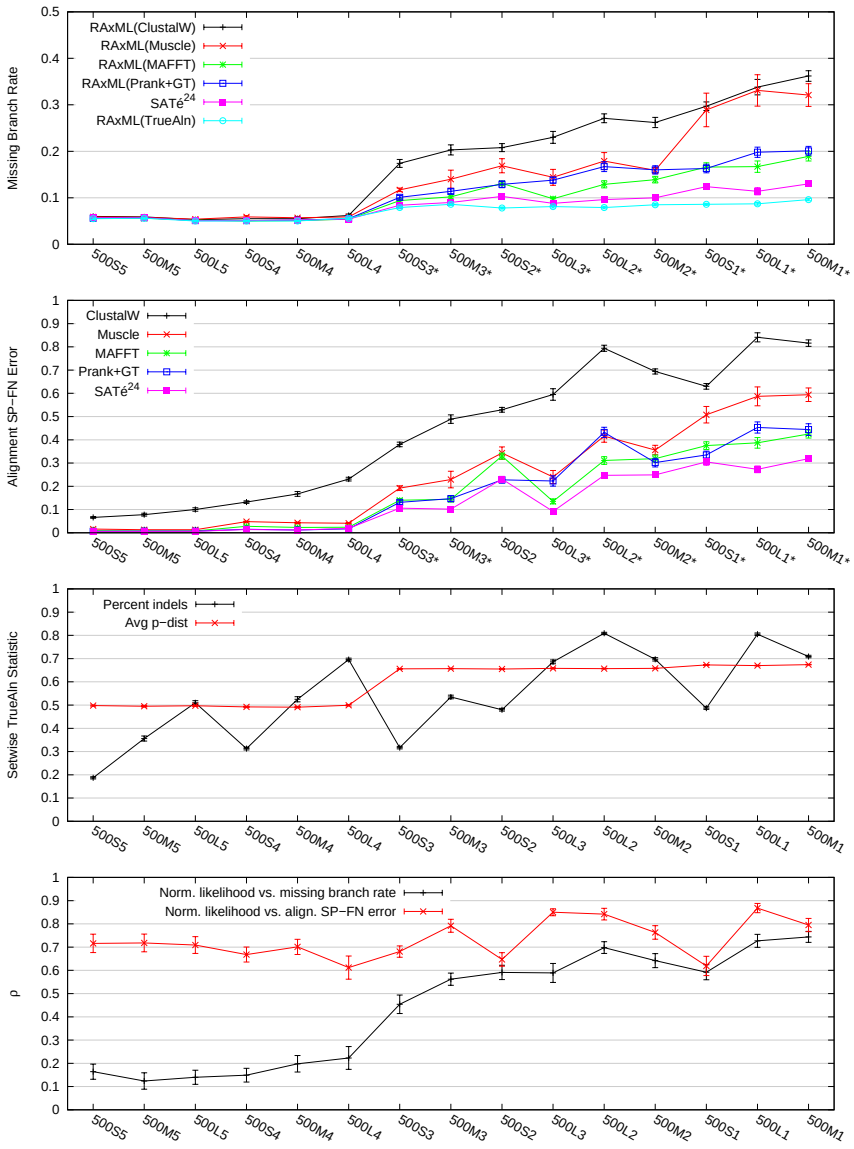



Figure S6: **Results for SATé$^{24}$ on 500-taxon datasets.** All x-axes show the fifteen 500-taxon models from easy to hard, based on missing branch rates. From top-to-bottom panels, the y-axes are missing branch rate (calculated with respect to the PIMT), alignment SP-FN error, true alignment setwise statistics, and Spearman rank correlation coefficients ($\rho$). All data points include standard error bars. For the top two panels, models on the x-axis followed by an asterisk indicate that SATé$^{24}$'s performance was significantly better than the nearest two-phase method (paired t-tests, setwise $\alpha = 0.05$, $n = 40$ for each test).

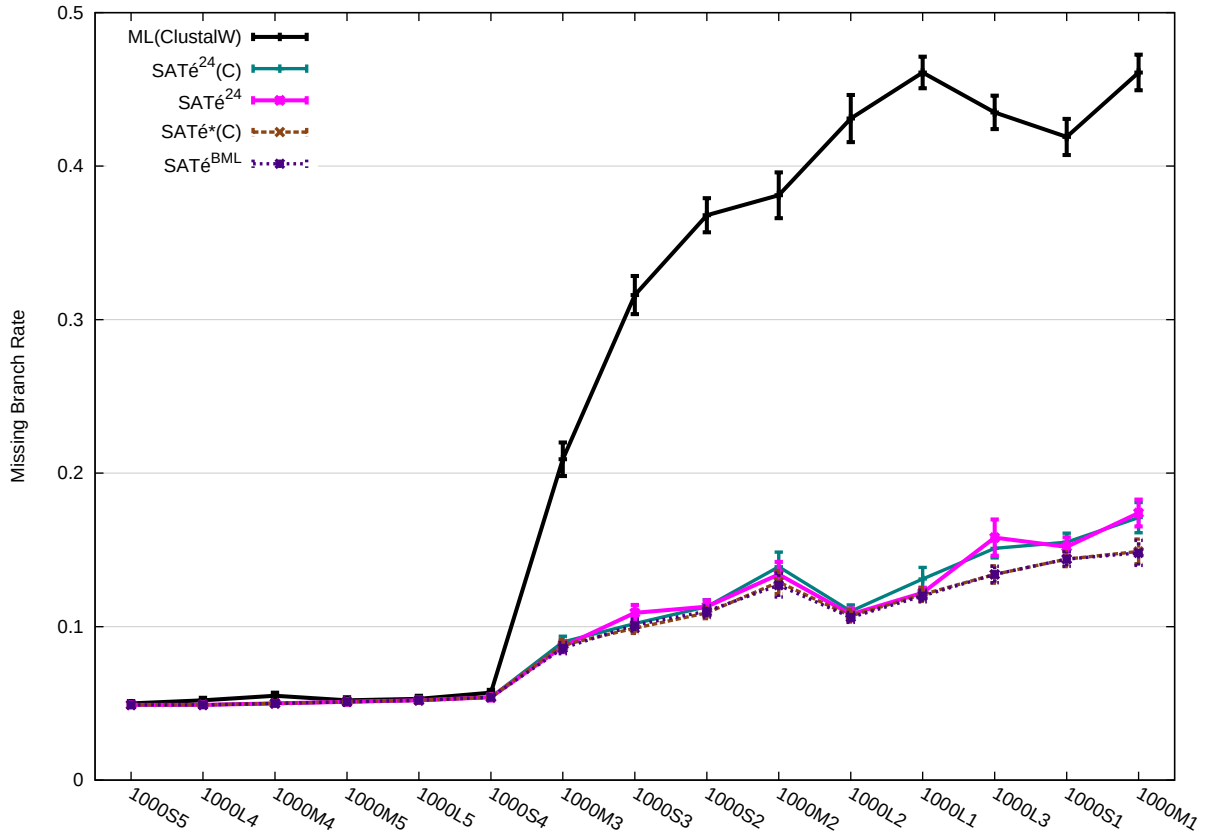



Figure S7: **Comparison of SATé missing branch rates (on 1000-taxon models) among variants of SATé using either the ClustalW starting tree/alignment pair or the tree/alignment pair with the best ML score.** The x-axis has the fifteen 1000-taxon models from easy to hard, based on missing branch rates. All data points include standard error bars. There are $n = 20$ datapoints for each reported average and standard error. SATé$^{24}$(C) performs CT-5 proposals for 24 hours using an ML(ClustalW) starting tree/alignment pair. SATé*(C) is run until no more improvements with CT-5 proposals can be found in 24 hours using an ML(ClustalW) starting tree/alignment pair. SATé$^{BML}$ is the best likelihood method out of the set of SATé$^{24}$, SATé$^{24}$(C), and SATé*(C).

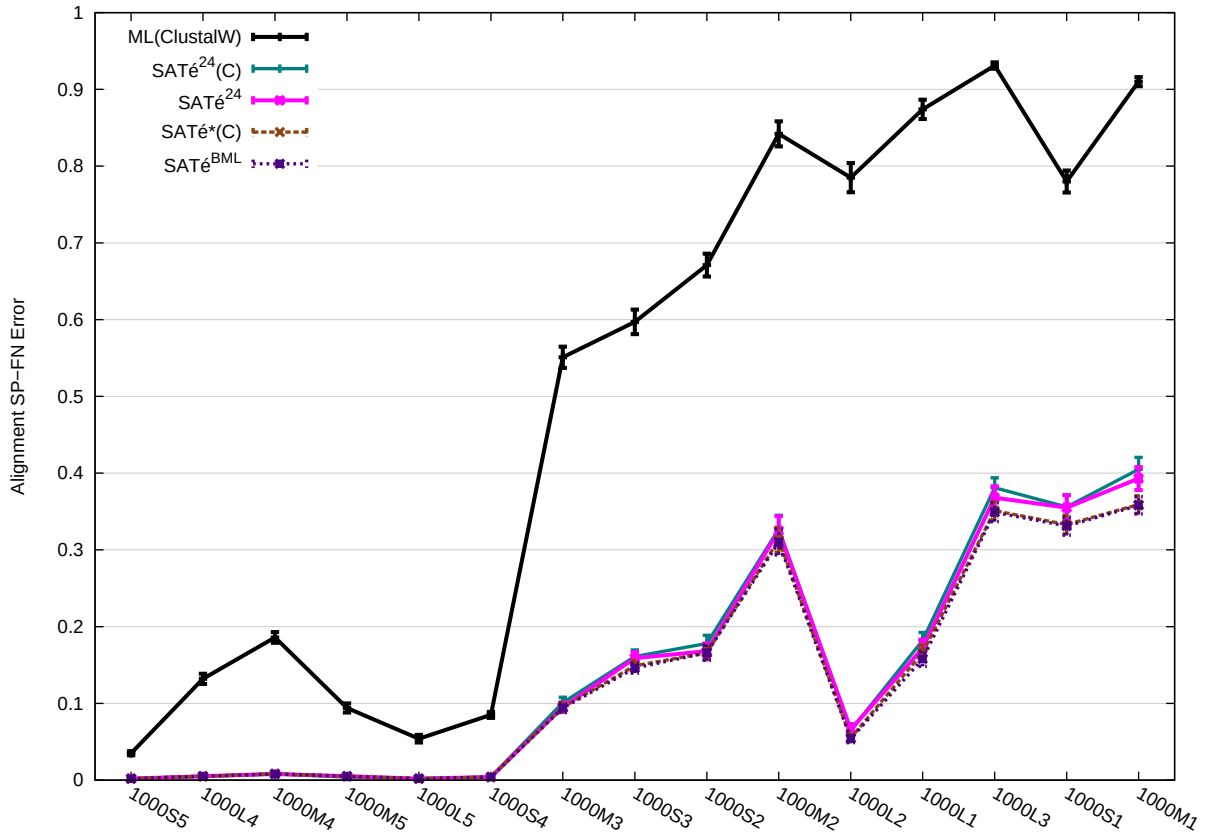



Figure S8: **Comparison of SATé SP-FN error (on 1000-taxon models) among variants of SATé using either the ClustalW starting tree/alignment pair or the tree/alignment pair with the best ML score.** The x-axis has the fifteen 1000-taxon models from easy to hard, based on missing branch rates. All data points include standard error bars. There are $n = 20$ datapoints for each reported average and standard error. SATé$^{24}$(C) performs CT-5 proposals for 24 hours using an ML(ClustalW) starting tree/alignment pair. SATé*(C) is run until no more improvements with CT-5 proposals can be found in 24 hours using an ML(ClustalW) starting tree/alignment pair. SATé$^{BML}$ is the best likelihood method out of the set of SATé$^{24}$, SATé$^{24}$(C), and SATé*(C).

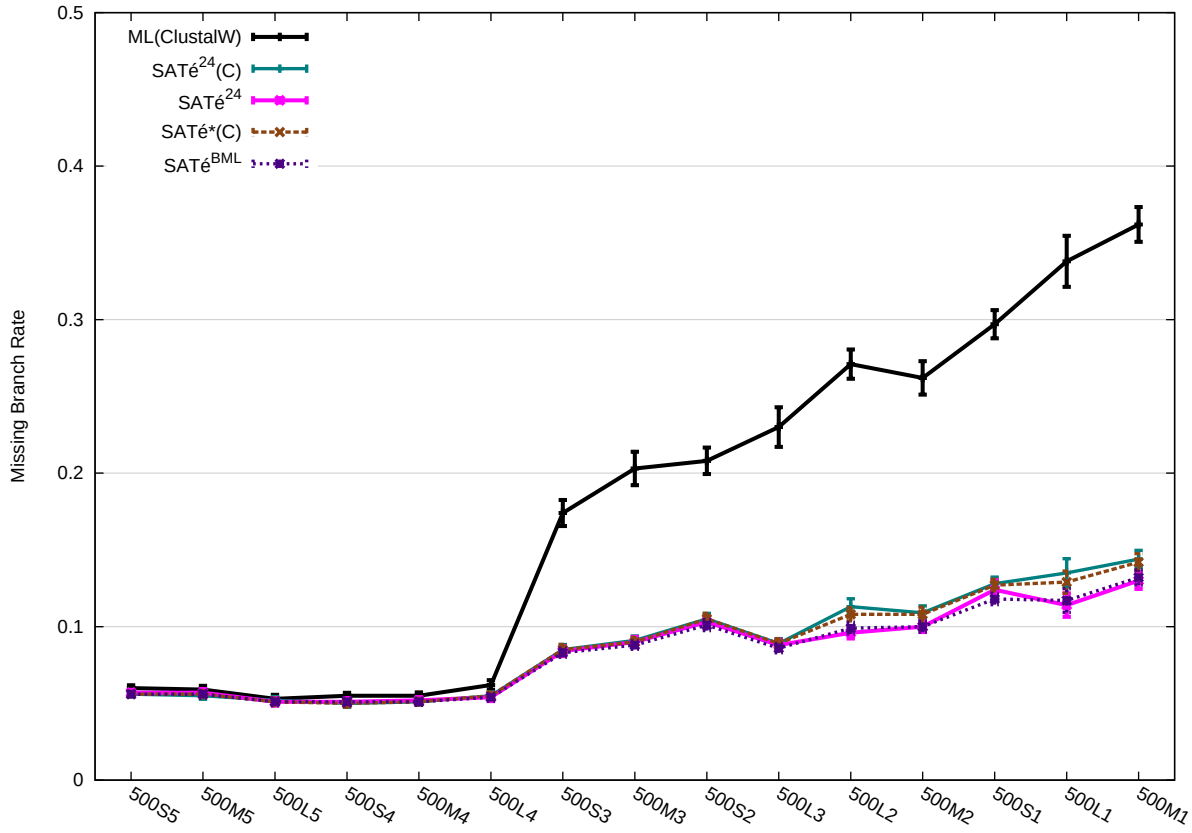



Figure S9: **Comparison of SATé missing branch rates (on 500-taxon models) among variants of SATé using either the ClustalW starting tree/alignment pair or the tree/alignment pair with the best ML score.** The x-axis has the fifteen 500-taxon models from easy to hard, based on missing branch rates. All data points include standard error bars. There are $n = 20$ datapoints for each reported average and standard error. SATé$^{24}$(C) performs CT-5 proposals for 24 hours using an ML(ClustalW) starting tree/alignment pair. SATé*(C) is run until no more improvements with CT-5 proposals can be found in 24 hours using an ML(ClustalW) starting tree/alignment pair. SATé$^{BML}$ is the best likelihood method out of the set of SATé$^{24}$, SATé$^{24}$(C), and SATé*(C).

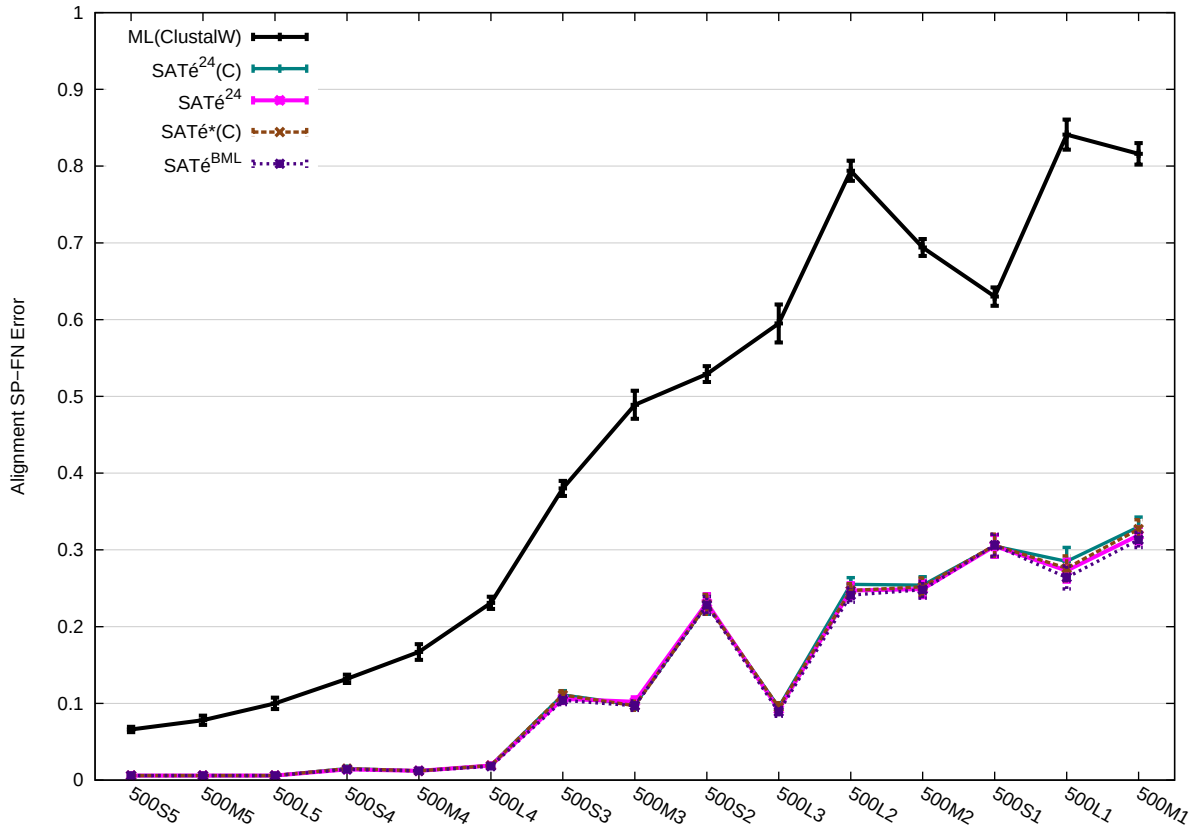



Figure S10: **Comparison of SATé SP-FN error (on 500-taxon models) among variants of SATé using either the ClustalW starting tree/alignment pair or the tree/alignment pair with the best ML score** The x-axis has the fifteen 500-taxon models from easy to hard, based on missing branch rates. All data points include standard error bars. There are $n = 20$ datapoints for each reported average and standard error. SATé$^{24}$(C) performs CT-5 proposals for 24 hours using an ML(ClustalW) starting tree/alignment pair. SATé*(C) is run until no more improvements with CT-5 proposals can be found in 24 hours using an ML(ClustalW) starting tree/alignment pair. SATé$^{BML}$ is the best likelihood method out of the set of SATé$^{24}$, SATé$^{24}$(C), and SATé*(C).

## Supporting Tables

Table S1: **Parameters used to simulate datasets**.The parameters used to evolve sequences on trees for the 37 different models. See the Materials and Methods section for details on how the gap-length categories were defined.

| Model | Taxa | Gap Length | Tree Height | Gap Prob |
|---|---|---|---|---|
| 1000L1 | 1000 | long | 35 | 4.3E-06 |
| 1000L2 | 1000 | long | 35 | 2E-06 |
| 1000L3 | 1000 | long | 30 | 1E-05 |
| 1000L4 | 1000 | long | 5 | 1.5E-05 |
| 1000L5 | 1000 | long | 5 | 7.5E-06 |
| 1000M1 | 1000 | medium | 35 | 8.2E-06 |
| 1000M2 | 1000 | medium | 30 | 1E-05 |
| 1000M3 | 1000 | medium | 20 | 7.5E-06 |
| 1000M4 | 1000 | medium | 5 | 3E-05 |
| 1000M5 | 1000 | medium | 5 | 1.5E-05 |
| 1000S1 | 1000 | short | 35 | 8.2E-06 |
| 1000S2 | 1000 | short | 35 | 4.3E-06 |
| 1000S3 | 1000 | short | 30 | 5E-06 |
| 1000S4 | 1000 | short | 5 | 1.5E-05 |
| 1000S5 | 1000 | short | 5 | 7.5E-06 |
| 500L1 | 500 | long | 25 | 1.6E-05 |
| 500L2 | 500 | long | 20 | 2E-05 |
| 500L3 | 500 | long | 20 | 1E-05 |
| 500L4 | 500 | long | 5 | 4E-05 |
| 500L5 | 500 | long | 5 | 2E-05 |
| 500M1 | 500 | medium | 25 | 1.6E-05 |
| 500M2 | 500 | medium | 20 | 2E-05 |
| 500M3 | 500 | medium | 20 | 1E-05 |
| 500M4 | 500 | medium | 5 | 4E-05 |
| 500M5 | 500 | medium | 5 | 2E-05 |
| 500S1 | 500 | short | 25 | 1.6E-05 |
| 500S2 | 500 | short | 20 | 2E-05 |
| 500S3 | 500 | short | 20 | 1E-05 |
| 500S4 | 500 | short | 5 | 4E-05 |
| 500S5 | 500 | short | 5 | 2E-05 |
| 100L1 | 100 | long | 15 | 3E-05 |
| 100L2 | 100 | long | 2 | 0.0002 |
| 100M1 | 100 | medium | 15 | 5E-05 |
| 100M2 | 100 | medium | 7 | 0.0001 |
| 100M3 | 100 | medium | 4 | 0.0001 |
| 100S1 | 100 | short | 10 | 0.0001 |
| 100S2 | 100 | short | 2 | 0.001 |

Table S2: **Empirical statistics for the 500 and 1000-taxon moderate-to-difficult datasets** Empirical statistics for the true alignments and trees arising from the moderate-to-difficult models. The percent indels of an alignment is the percentage of cells in the alignment matrix that lacked nucleotides. The p-distance of a pair of aligned sequences was calculated by dividing the number of sites where the two sequences had different nucleotides by the number of sites in which both sequences had nucleotides. The average p-distance of an alignment is the average p-distance of all pairs of aligned sequences in that alignment. A gap in an alignment is any maximal contiguous sequence of cells lacking nucleotides within the alignment. The average or maximum gap length of an alignment is the average or maximum length, respectively, of all gaps within the alignment. The degree of resolution in the PIMT is the number of internal branches in the tree divided by $t - 3$, where $t$ is the number of taxa in the tree (a fully resolved unrooted tree on $t$ taxa has $t - 3$ internal branches). Each statistic is the average of the $n = 20$ replicates for a model. Maximum standard deviations (max std dev) are reported for each statistic.

| | True Alignment Statistics (Set-wise) | | | | | | PIMT Statistics (Branch-wise) | | PIMT Statistics |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Avg. p-dist (%)** | **Max. p-dist (%)** | **Indels (%)** | **No. of Cols** | **Avg. Gap Len** | **Median Gap Len** | **Avg. p-dist (%)** | **Indels (%)** | **Resolution (%)** |
| 1000L1 | 69.5 | 76.9 | 73.2 | 3817.5 | 13.6 | 10.7 | 25.3 | 0.16 | 99.7 |
| 1000L2 | 69.6 | 76.9 | 57.7 | 2406.9 | 11.6 | 9.3 | 25.1 | 0.08 | 99.6 |
| 1000L3 | 68.7 | 76.3 | 85.2 | 7042.8 | 20.0 | 16.1 | 22.8 | 0.35 | 99.7 |
| 1000M1 | 69.5 | 76.9 | 74.4 | 3965.0 | 10.1 | 8.0 | 24.8 | 0.19 | 99.6 |
| 1000M2 | 68.4 | 76.2 | 74.2 | 3972.3 | 10.3 | 7.9 | 22.1 | 0.19 | 99.5 |
| 1000M3 | 66.0 | 74.1 | 62.8 | 2722.6 | 7.6 | 5.6 | 18.0 | 0.11 | 99.4 |
| 1000S1 | 69.4 | 76.8 | 53.0 | 2141.2 | 4.0 | 3.4 | 24.5 | 0.09 | 99.7 |
| 1000S2 | 69.3 | 76.8 | 35.0 | 1546.0 | 2.9 | 2.4 | 23.8 | 0.04 | 99.7 |
| 1000S3 | 68.6 | 76.3 | 37.0 | 1595.2 | 2.9 | 2.4 | 22.3 | 0.05 | 99.6 |
| **max std dev** | 1.2 | 1.0 | 4.5 | 605.33 | 1.35 | 1.46 | 1.8 | 0.03 | 0.7 |
| 500L1 | 67.0 | 74.9 | 80.5 | 5419.3 | 17.0 | 13.0 | 21.2 | 0.50 | 99.5 |
| 500L2 | 65.7 | 73.9 | 80.9 | 5475.9 | 16.9 | 12.7 | 19.3 | 0.52 | 99.4 |
| 500L3 | 65.8 | 74.1 | 68.6 | 3306.9 | 12.9 | 10.3 | 19.4 | 0.26 | 99.4 |
| 500M1 | 67.4 | 74.8 | 70.9 | 3522.3 | 9.1 | 6.8 | 22.7 | 0.32 | 99.5 |
| 500M2 | 65.8 | 74.0 | 69.7 | 3394.5 | 9.0 | 6.5 | 19.4 | 0.30 | 99.5 |
| 500M3 | 65.7 | 73.7 | 53.5 | 2185.2 | 6.8 | 4.9 | 19.4 | 0.15 | 99.4 |
| 500S1 | 67.3 | 75.2 | 48.7 | 1962.2 | 3.6 | 3.0 | 22.2 | 0.15 | 99.5 |
| 500S2 | 65.5 | 73.7 | 48.0 | 1935.8 | 3.6 | 2.9 | 19.1 | 0.15 | 99.5 |
| 500S3 | 65.6 | 73.6 | 31.7 | 1468.2 | 2.7 | 2.0 | 19.4 | 0.08 | 99.5 |
| **max std dev** | 1.6 | 1.4 | 5.0 | 643.47 | 1.37 | 1.45 | 1.6 | 0.07 | 0.8 |

Table S3: **Empirical statistics for the 500 and 1000-taxon easy models.** Empirical statistics, as described in Table S2, for the true alignments and trees arising from the easy models. $n = 20$ for each entry.

| | True Alignment Statistics (Set-wise) | | | | | | PIMT Statistics (Branch-wise) | | PIMT Statistics |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Avg. p-dist (%)** | **Max. p-dist (%)** | **Indels (%)** | **No. of Cols** | **Avg. Gap Len** | **Median Gap Len** | **Avg. p-dist (%)** | **Indels (%)** | **Resolution (%)** |
| 1000L4 | 50.0 | 60.8 | 58.6 | 2446.2 | 11.4 | 9.2 | 6.5 | 0.08 | 97.6 |
| 1000L5 | 49.6 | 60.6 | 42.6 | 1764.8 | 10.4 | 8.0 | 6.5 | 0.04 | 97.4 |
| 1000M4 | 49.5 | 60.6 | 60.5 | 2570.6 | 7.6 | 5.8 | 6.6 | 0.10 | 97.6 |
| 1000M5 | 49.9 | 60.2 | 44.2 | 1810.0 | 6.2 | 4.4 | 6.8 | 0.05 | 97.8 |
| 1000S4 | 50.1 | 60.8 | 24.6 | 1328.1 | 2.5 | 2.0 | 6.7 | 0.03 | 97.9 |
| 1000S5 | 49.8 | 61.1 | 14.1 | 1165.2 | 2.3 | 2.0 | 6.5 | 0.01 | 97.8 |
| **max std dev** | 1.2 | 1.0 | 4.5 | 605.33 | 1.35 | 1.46 | 1.8 | 0.03 | 0.7 |
| 500L4 | 49.9 | 60.7 | 69.6 | 3390.2 | 13.3 | 10.0 | 7.5 | 0.27 | 98.3 |
| 500L5 | 49.7 | 60.6 | 51.0 | 2075.3 | 10.9 | 8.6 | 7.1 | 0.12 | 97.8 |
| 500M4 | 49.1 | 60.5 | 52.5 | 2154.6 | 6.7 | 4.9 | 7.0 | 0.15 | 98.0 |
| 500M5 | 49.5 | 60.5 | 35.6 | 1568.2 | 5.7 | 4.2 | 7.3 | 0.07 | 98.1 |
| 500S4 | 49.2 | 60.6 | 31.3 | 1459.8 | 2.7 | 2.0 | 7.2 | 0.07 | 98.2 |
| 500S5 | 49.8 | 60.4 | 18.7 | 1231.0 | 2.3 | 1.9 | 7.3 | 0.04 | 98.0 |
| **max std dev** | 1.6 | 1.4 | 5.0 | 643.47 | 1.37 | 1.45 | 1.6 | 0.07 | 0.8 |

Table S4: **Empirical statistics for the 100-taxon model conditions**. We report the empirical statistics for the 100-taxon model conditions used in the first ALIFRITZ and BAli-Phy experiment. $n = 20$ for each entry.

| | True Alignment Statistics (Set-wise) | | | | | | PIMT Statistics (Branch-wise) | | PIMT Statistics |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Avg. p-dist (%)** | **Max. p-dist (%)** | **Indels (%)** | **No. of Cols** | **Avg. Gap Len** | **Median Gap Len** | **Avg. p-dist (%)** | **Indels (%)** | **Resolution (%)** |
| 100L1 | 62.3 | 71.0 | 56.4 | 2459.7 | 11.2 | 8.1 | 20.5 | 0.8 | 99.7 |
| 100L2 | 32.2 | 44.4 | 53.6 | 2281.9 | 11.4 | 8.5 | 4.3 | 0.7 | 95.9 |
| 100M1 | 62.6 | 71.0 | 55.1 | 2316.8 | 7.1 | 4.8 | 20.4 | 0.8 | 99.7 |
| 100M2 | 53.1 | 63.3 | 53.6 | 2262.9 | 7.0 | 4.8 | 12.7 | 0.8 | 99.3 |
| 100M3 | 45.0 | 56.4 | 38.9 | 1681.9 | 5.9 | 4.2 | 8.1 | 0.4 | 98.4 |
| 100S1 | 58.0 | 67.5 | 40.4 | 1698.2 | 3.1 | 2.6 | 16.1 | 0.6 | 99.0 |
| 100S2 | 32.0 | 43.6 | 57.2 | 2418.3 | 4.6 | 3.6 | 4.6 | 1.1 | 96.8 |
| **max std dev** | 2.0 | 1.5 | 5.9 | 359.6 | 1.3 | 1.3 | 1.9 | 0.2 | 2.0 |

Table S5: **Empirical statistics for the second ALIFRITZ and BAli-Phy experiments.** Each dataset has 100 taxa and consists of a single replicate. We also report the resolution of the majority consensus tree of the trees sampled by BAli-Phy. $n = 1$ for each value in the table.

| | True Alignment Statistics (Set-wise) | | | | | | PIMT Statistics (Branch-wise) | | PIMT Statistics | BAli-Phy Statistics |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Avg. p-dist (%) | Max. p-dist (%) | Indels (%) | No. of Cols | Avg. Gap Len | Median Gap Len | Avg. p-dist (%) | Indels (%) | Resolution (%) | Resolution (%) |
| 100L1 | 62.8 | 70.8 | 54.4 | 2287 | 10.6 | 7.0 | 21.2 | 0.8 | 100.0 | 93.8 |
| 100L2 | 34.3 | 44.9 | 60.6 | 2720 | 13.1 | 9.0 | 5.2 | 1.0 | 96.9 | 99.0 |
| 100M1 | 61.9 | 69.2 | 58.5 | 2505 | 7.5 | 6.0 | 20.5 | 0.9 | 100.0 | 93.8 |
| 100M2 | 53.4 | 63.2 | 52.2 | 2184 | 7.6 | 5.0 | 12.0 | 0.8 | 100.0 | 99.0 |
| 100M3 | 44.3 | 55.6 | 31.8 | 1496 | 4.6 | 4.0 | 7.5 | 0.3 | 100.0 | 99.0 |
| 100S1 | 59.0 | 68.2 | 38.7 | 1650 | 2.8 | 2.0 | 16.7 | 0.5 | 99.0 | 94.8 |
| 100S2 | 32.8 | 42.9 | 62.7 | 2738 | 5.4 | 4.0 | 5.6 | 1.4 | 95.9 | 97.9 |

Table S6: **Informative gap information for the 500 and 1000-taxon model conditions.** The number of informative gaps per sequence, mean informative gap length, and median informative gap length statistics for the true alignments from each 500 and 1000-taxon simulated model condition. Averages and standard deviations are shown ($n = 20$ for each value). All taxon sequences in all of our simulations had informative gaps.

| | Number of informative gaps per sequence | | Mean informative gap length | | Median informative gap length | |
|---|---|---|---|---|---|---|
| **Model condition** | **Average** | **Std dev** | **Average** | **Std dev** | **Average** | **Std dev** |
| 1000L1 | 88.9 | 11.4 | 15.2 | 1.5 | 12.1 | 1.5 |
| 1000L2 | 41.7 | 8.2 | 17.0 | 1.9 | 13.1 | 2.3 |
| 1000L3 | 174.8 | 17.0 | 17.2 | 1.2 | 12.9 | 1.3 |
| 1000L4 | 44.7 | 6.9 | 16.4 | 2.2 | 12.9 | 2.3 |
| 1000L5 | 22.5 | 2.8 | 17.5 | 3.1 | 12.7 | 3.7 |
| 1000M1 | 163.6 | 15.7 | 9.1 | 0.6 | 6.5 | 0.5 |
| 1000M2 | 160.4 | 16.8 | 9.0 | 0.7 | 6.5 | 0.7 |
| 1000M3 | 99.5 | 15.0 | 8.6 | 0.8 | 6.5 | 0.9 |
| 1000M4 | 88.5 | 12.2 | 8.6 | 0.8 | 6.3 | 0.8 |
| 1000M5 | 47.7 | 7.1 | 8.3 | 1.1 | 5.7 | 1.1 |
| 1000S1 | 156.8 | 17.1 | 3.6 | 0.2 | 2.9 | 0.3 |
| 1000S2 | 75.1 | 12.4 | 3.7 | 0.3 | 2.9 | 0.4 |
| 1000S3 | 84.9 | 13.4 | 3.5 | 0.3 | 2.8 | 0.4 |
| 1000S4 | 43.6 | 6.5 | 3.7 | 0.4 | 3.0 | 0.5 |
| 1000S5 | 21.8 | 3.8 | 3.8 | 0.6 | 3.0 | 0.6 |
| 500L1 | 129.7 | 16.9 | 16.4 | 1.3 | 12.2 | 1.5 |
| 500L2 | 137.1 | 15.0 | 16.2 | 1.3 | 12.3 | 1.4 |
| 500L3 | 70.2 | 10.0 | 16.1 | 1.3 | 12.2 | 1.4 |
| 500L4 | 71.4 | 7.8 | 16.4 | 1.4 | 12.2 | 1.9 |
| 500L5 | 31.6 | 4.0 | 16.2 | 2.1 | 12.1 | 2.7 |
| 500M1 | 139.5 | 14.7 | 8.9 | 0.6 | 6.2 | 0.4 |
| 500M2 | 131.1 | 16.7 | 8.6 | 0.6 | 6.0 | 0.6 |
| 500M3 | 67.4 | 7.9 | 8.6 | 0.9 | 6.2 | 1.0 |
| 500M4 | 63.2 | 11.2 | 8.5 | 1.2 | 6.2 | 1.2 |
| 500M5 | 33.1 | 6.5 | 8.2 | 1.3 | 6.2 | 1.1 |
| 500S1 | 129.5 | 15.0 | 3.7 | 0.3 | 3.0 | 0.4 |
| 500S2 | 123.8 | 14.0 | 3.7 | 0.3 | 3.0 | 0.0 |
| 500S3 | 64.8 | 7.2 | 3.6 | 0.4 | 2.9 | 0.4 |
| 500S4 | 63.2 | 10.1 | 3.5 | 0.4 | 2.7 | 0.5 |
| 500S5 | 30.7 | 4.4 | 3.7 | 0.5 | 2.9 | 0.5 |

Table S7: **Biological dataset cleaning statistics.** We report the percentage of sites in each original curated alignment that were comprised of only indel letters and thus had no nucleotides. We also report the percentage of taxa in each original curated alignment that were more than 50% unsequenced. These sites and taxa were removed during our biological dataset cleaning procedure to produce the cleaned curated alignments used throughout our experiments.

| Dataset | Sites without nucleotides (%) | Taxa with more than 50% unsequenced sites (%) |
|---|---|---|
| 23S.M.aa_ag | 4.4 | 30.7 |
| 23S.M | 2.4 | 2.8 |
| 16S.M | 4.2 | 0.0 |
| 16S.M.aa_ag | 5.9 | 14.8 |
| 23S.E | 3.8 | 0.0 |
| 23S.E.aa_ag | 6.9 | 53.2 |

Table S8: **Biological dataset curation statistics.** Bootstraps were computed over 500 replicates. The median gap length was always 1. Top 2 average refers to the average of the 23SM.aa_ag and 23S.M datasets. Top 4 average refers to the average of the 23SM.aa_ag, 23S.M, 16S.M, and 16S.M.aa_ag datasets. All 6 average refers to the average of all of the biological datasets. $n = 1$ for all values in the table.

| Dataset | Taxa | Sites | 90% bootstrap resolution | 75% bootstrap resolution | 50% bootstrap resolution | Average p-distance (%) | Indels (%) | Average gap len |
|---|---|---|---|---|---|---|---|---|
| 23S.M.aa_ag | 263 | 10305 | 50.0 | 60.0 | 78.5 | 37.7 | 83.5 | 34.2 |
| 23S.M | 278 | 10738 | 48.7 | 61.1 | 73.1 | 37.7 | 83.7 | 31.9 |
| 16S.M | 901 | 4722 | 35.6 | 46.9 | 65.1 | 35.9 | 78.1 | 17.2 |
| 16S.M.aa_ag | 1028 | 4907 | 32.4 | 42.2 | 61.2 | 34.2 | 82.6 | 22.0 |
| 23S.E | 117 | 9079 | 57.9 | 65.8 | 77.2 | 29.6 | 59.7 | 12.6 |
| 23S.E.aa_ag | 144 | 8619 | 58.2 | 64.5 | 77.3 | 30.3 | 61.1 | 13.5 |
| **Top 2 average** | 270.5 | 10521.5 | 49.4 | 60.5 | 75.8 | 37.7 | 83.6 | 33.1 |
| **Top 4 average** | 617.5 | 7668.0 | 41.7 | 52.6 | 69.5 | 36.4 | 82.0 | 26.3 |
| **All 6 average** | 455.2 | 8061.7 | 47.1 | 56.8 | 72.1 | 34.2 | 74.8 | 21.9 |

Table S9: **Averages and standard deviations of SATé$^{24}$ runtimes and normalized log likelihoods by stage.** Stage 1 likelihoods are normalized to the likelihood of the RAxML(MAFFT) analysis. Stage 2 likelihoods are given as the amount of relative improvement over the Stage 1 likelihoods. $n = 20$ for each value in the table.

| | **Runtime (hours)** | | | | | | | | | **Normalized Log Likelihood (%)** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **SATé$^{24}$ Stage 1** | | | **SATé$^{24}$ Stage 2** | | | **SATé$^{24}$ Total** | | | **SATé$^{24}$ Stage 1** | | **SATé$^{24}$ Stage 2 Improvement** | |
| **Model** | **Avg** | **Min** | **Max** | **Avg** | **Min** | **Max** | **Avg** | **Min** | **Max** | **Avg** | **Std dev** | **Avg** | **Std dev** |
| 1000L1 | 60.1 | 42.1 | 83.7 | 29.2 | 25.4 | 32.9 | 89.3 | 70.9 | 113.7 | 99.429 | 0.603 | 1.652 | 0.680 |
| 1000L2 | 44.3 | 32.4 | 64.3 | 28.1 | 25.5 | 31.3 | 72.5 | 58.4 | 95.6 | 99.126 | 0.774 | 0.784 | 0.448 |
| 1000L3 | 77.4 | 56.5 | 108.4 | 29.5 | 25.7 | 32.4 | 106.9 | 87.2 | 140.4 | 99.694 | 0.363 | 1.572 | 0.451 |
| 1000L4 | 20.2 | 18.4 | 22.9 | 26.4 | 25.1 | 27.9 | 46.6 | 44.7 | 49.9 | 99.955 | 0.035 | 0.002 | 0.009 |
| 1000L5 | 19.3 | 17.2 | 22.6 | 26.0 | 24.9 | 27.0 | 45.3 | 42.6 | 48.3 | 99.989 | 0.010 | 0.007 | 0.012 |
| 1000M1 | 61.1 | 43.7 | 88.9 | 28.2 | 25.4 | 32.1 | 89.2 | 69.9 | 119.3 | 98.338 | 1.268 | 1.149 | 0.574 |
| 1000M2 | 48.1 | 32.6 | 73.2 | 28.8 | 26.1 | 33.6 | 76.9 | 61.3 | 100.1 | 99.173 | 0.777 | 0.956 | 0.425 |
| 1000M3 | 31.0 | 25.5 | 39.2 | 27.0 | 24.9 | 29.5 | 58.0 | 50.5 | 67.9 | 99.520 | 0.391 | 0.201 | 0.207 |
| 1000M4 | 23.3 | 19.7 | 27.0 | 26.5 | 25.5 | 28.0 | 49.8 | 45.2 | 55.0 | 99.850 | 0.073 | 0 | 0 |
| 1000M5 | 19.7 | 17.3 | 21.2 | 25.9 | 24.8 | 27.8 | 45.5 | 43.4 | 47.0 | 99.951 | 0.033 | 0.013 | 0.050 |
| 1000S1 | 46.2 | 34.3 | 64.5 | 27.1 | 25.1 | 29.7 | 73.3 | 60.3 | 92.4 | 97.972 | 0.927 | 0.656 | 0.532 |
| 1000S2 | 36.9 | 25.5 | 50.0 | 27.1 | 25.5 | 29.6 | 64.0 | 51.0 | 77.4 | 98.836 | 0.867 | 0.640 | 0.488 |
| 1000S3 | 33.6 | 25.3 | 42.0 | 26.3 | 25.2 | 28.1 | 59.9 | 51.2 | 68.7 | 98.699 | 1.005 | 0.400 | 0.348 |
| 1000S4 | 18.8 | 16.3 | 23.5 | 26.0 | 24.8 | 27.5 | 44.8 | 41.8 | 49.9 | 99.948 | 0.042 | 0.001 | 0.003 |
| 1000S5 | 18.4 | 16.7 | 21.1 | 26.2 | 24.7 | 30.1 | 44.6 | 42.3 | 50.8 | 99.986 | 0.014 | 0.005 | 0.007 |
| 500L1 | 18.6 | 12.5 | 24.8 | 25.8 | 24.7 | 27.1 | 44.4 | 37.5 | 50.5 | 99.478 | 0.953 | 1.593 | 0.689 |
| 500L2 | 14.4 | 11.8 | 20.1 | 25.6 | 24.8 | 26.6 | 40.0 | 37.0 | 45.7 | 99.910 | 0.281 | 1.193 | 0.592 |
| 500L3 | 11.4 | 8.1 | 18.4 | 25.2 | 24.6 | 26.2 | 36.6 | 33.5 | 43.2 | 99.930 | 0.224 | 0.654 | 0.227 |
| 500L4 | 9.1 | 7.6 | 11.1 | 25.0 | 24.5 | 25.8 | 34.2 | 32.4 | 35.9 | 99.875 | 0.065 | 0.020 | 0.030 |
| 500L5 | 7.1 | 6.1 | 9.1 | 24.9 | 24.4 | 25.4 | 32.0 | 30.5 | 33.9 | 99.967 | 0.029 | 0.017 | 0.022 |
| 500M1 | 13.2 | 9.6 | 18.0 | 25.6 | 24.8 | 26.9 | 38.8 | 34.7 | 43.3 | 98.753 | 1.559 | 1.459 | 0.557 |
| 500M2 | 11.2 | 9.7 | 13.7 | 25.2 | 24.5 | 26.3 | 36.4 | 34.5 | 39.7 | 99.539 | 0.564 | 0.922 | 0.342 |
| 500M3 | 10.5 | 7.7 | 13.4 | 25.0 | 24.4 | 25.7 | 35.5 | 32.3 | 38.6 | 99.549 | 0.487 | 0.609 | 0.349 |
| 500M4 | 8.8 | 7.1 | 11.9 | 24.8 | 24.4 | 25.5 | 33.7 | 31.7 | 37.1 | 99.848 | 0.082 | 0.004 | 0.017 |
| 500M5 | 7.0 | 5.9 | 8.9 | 24.7 | 24.3 | 25.4 | 31.7 | 30.4 | 33.7 | 99.957 | 0.036 | 0.013 | 0.013 |
| 500S1 | 11.8 | 8.3 | 17.7 | 25.1 | 24.4 | 26.1 | 37.0 | 33.7 | 42.6 | 98.312 | 1.097 | 0.778 | 0.573 |
| 500S2 | 9.2 | 7.8 | 12.6 | 25.0 | 24.4 | 25.6 | 34.1 | 32.5 | 37.0 | 98.324 | 1.549 | 0.425 | 0.397 |
| 500S3 | 10.2 | 8.0 | 12.5 | 24.8 | 24.4 | 25.4 | 35.0 | 32.7 | 37.5 | 99.251 | 0.558 | 0.496 | 0.267 |
| 500S4 | 7.6 | 6.1 | 9.4 | 25.0 | 24.4 | 25.4 | 32.5 | 30.9 | 34.1 | 99.806 | 0.096 | 0.002 | 0.007 |
| 500S5 | 6.7 | 5.8 | 8.5 | 24.7 | 24.3 | 25.2 | 31.4 | 30.3 | 33.2 | 99.943 | 0.039 | 0.003 | 0.008 |

Table S10: **Number of proposals made and accepted by SATé[24].** The number of proposals made and the number of proposals accepted by SATé[24] for model conditions grouped by number of taxa and difficulty type. $n = 180$ for values in the moderate-to-difficult rows of the table. $n = 120$ for values in the easy rows of the table.

| | **Number of proposals made** | | **Number of proposals accepted** | |
|---|---|---|---|---|
| **Model condition group** | **Average** | **Std dev** | **Average** | **Std dev** |
| 1000 moderate-to-difficult | 5.4 | 1.6 | 2.3 | 1.2 |
| 500 moderate-to-difficult | 18.6 | 4.5 | 3.4 | 1.8 |
| 1000 easy | 13.0 | 2.1 | 0.7 | 1.3 |
| 500 easy | 31.6 | 5.5 | 1.2 | 2.0 |

Table S11: **Results for alignment/tree proposals accepted by SATé[24] for the moderate-to-difficult models.** The missing branch rate and alignment SP-FN error for three different alignment/tree pairs encountered during a SATé[24] run: the initial tree/alignment pair, the first accepted pair, and the final pair. Averages and standard errors are shown; $n = 20$ for all values.

| | Missing branch rate (%) | | | | | | Alignment SP-FN error (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Starting tree | | First accepted tree | | Final tree | | Starting alignment | | First accepted alignment | | Final alignment | |
| **Model** | **Average** | **Std dev** | **Average** | **Std dev** | **Average** | **Std dev** | **Average** | **Std dev** | **Average** | **Std dev** | **Average** | **Std dev** |
| 1000L1 | 24.9 | 6.4 | 15.3 | 4.1 | 12.2 | 1.5 | 31.8 | 10.2 | 22.8 | 6.4 | 17.2 | 5.0 |
| 1000L2 | 17.3 | 6.8 | 11.6 | 3.2 | 10.8 | 1.7 | 14.1 | 6.5 | 9.9 | 5.1 | 6.6 | 3.2 |
| 1000L3 | 26.4 | 5.5 | 19.6 | 5.2 | 15.8 | 5.3 | 48.7 | 9.9 | 43.9 | 7.8 | 36.8 | 6.6 |
| 1000M1 | 30.6 | 6.1 | 20.1 | 4.2 | 17.4 | 3.9 | 46.6 | 9.3 | 44.4 | 5.6 | 39.3 | 6.7 |
| 1000M2 | 21.0 | 5.8 | 14.8 | 3.8 | 13.4 | 3.7 | 37.6 | 8.9 | 37.1 | 7.4 | 32.5 | 8.4 |
| 1000M3 | 9.9 | 1.8 | 8.4 | 1.6 | 8.7 | 1.6 | 9.6 | 3.1 | 12.8 | 2.2 | 9.5 | 2.8 |
| 1000S1 | 24.6 | 6.1 | 17.4 | 5.0 | 15.2 | 2.8 | 36.3 | 9.0 | 38.7 | 4.7 | 35.5 | 7.3 |
| 1000S2 | 16.6 | 4.0 | 12.0 | 1.9 | 11.3 | 1.9 | 20.5 | 6.5 | 21.0 | 3.5 | 16.8 | 4.5 |
| 1000S3 | 14.1 | 3.2 | 10.5 | 1.8 | 10.9 | 2.3 | 16.7 | 3.7 | 17.1 | 3.6 | 15.9 | 3.5 |
| 500L1 | 17.3 | 6.1 | 12.8 | 4.1 | 11.4 | 3.4 | 37.7 | 11.6 | 30.5 | 8.1 | 27.3 | 6.5 |
| 500L2 | 13.2 | 3.5 | 10.1 | 1.7 | 9.6 | 1.7 | 30.6 | 7.9 | 27.1 | 4.7 | 24.7 | 4.3 |
| 500L3 | 10.9 | 3.5 | 9.1 | 1.8 | 8.8 | 1.5 | 13.0 | 4.0 | 11.0 | 3.1 | 9.2 | 2.6 |
| 500M1 | 19.7 | 4.4 | 14.7 | 3.3 | 13.0 | 2.4 | 41.2 | 8.7 | 35.9 | 5.9 | 31.9 | 4.8 |
| 500M2 | 14.8 | 3.6 | 10.6 | 1.9 | 10.0 | 1.6 | 28.1 | 6.7 | 27.7 | 4.2 | 24.9 | 4.8 |
| 500M3 | 11.1 | 2.5 | 9.2 | 1.6 | 9.0 | 1.5 | 12.6 | 5.0 | 12.2 | 2.4 | 10.2 | 2.7 |
| 500S1 | 16.3 | 4.1 | 13.1 | 2.1 | 12.4 | 2.8 | 33.5 | 7.4 | 34.2 | 4.9 | 30.5 | 6.4 |
| 500S2 | 12.9 | 3.2 | 10.2 | 1.9 | 10.3 | 1.6 | 22.3 | 5.9 | 26.5 | 4.9 | 23.1 | 5.2 |
| 500S3 | 9.9 | 1.7 | 8.7 | 1.5 | 8.4 | 1.4 | 12.8 | 3.8 | 13.0 | 2.5 | 10.6 | 2.6 |

Table S12: **Results for alignment/tree proposals accepted by SATé[24] for the easy models.** The missing branch rate and alignment SP-FN error for three different alignment/tree pairs encountered during a SATé[24] run as in Table S11. Averages and standard errors are shown; $n = 20$ for each value. For one model, 1000M4, SATé[24] did not accept any proposals, which is marked in the table as "N/A".

| | Missing branch rate (%) | | | | | | Alignment SP-FN error (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Starting tree | | First accepted tree | | Final tree | | Starting alignment | | First accepted alignment | | Final alignment | |
| Model | Average | Std dev | Average | Std dev | Average | Std dev | Average | Std dev | Average | Std dev | Average | Std dev |
| 1000L4 | 4.9 | 0.7 | 5.4 | 0.2 | 4.9 | 0.7 | 0.5 | 0.2 | 0.7 | 0.2 | 0.5 | 0.2 |
| 1000L5 | 5.2 | 0.8 | 5.1 | 0.9 | 5.2 | 0.8 | 0.3 | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 |
| 1000M4 | 5.0 | 0.6 | N/A | N/A | 5.0 | 0.6 | 0.8 | 0.2 | N/A | N/A | 0.8 | 0.2 |
| 1000M5 | 5.1 | 0.8 | 4.9 | 0.8 | 5.1 | 0.8 | 0.4 | 0.1 | 0.8 | 0.2 | 0.5 | 0.3 |
| 1000S4 | 5.4 | 0.9 | 5.4 | 0.6 | 5.4 | 0.9 | 0.4 | 0.1 | 0.5 | 0.2 | 0.4 | 0.1 |
| 1000S5 | 4.9 | 0.7 | 4.8 | 0.5 | 4.9 | 0.7 | 0.2 | 0.1 | 0.3 | 0.2 | 0.2 | 0.1 |
| 500L4 | 5.5 | 1.1 | 5.7 | 0.8 | 5.4 | 1.1 | 1.7 | 0.5 | 2.8 | 0.9 | 1.9 | 0.7 |
| 500L5 | 5.1 | 1.1 | 5.1 | 0.9 | 5.1 | 1.1 | 0.6 | 0.2 | 0.7 | 0.3 | 0.6 | 0.2 |
| 500M4 | 5.2 | 1.0 | 3.5 | 0.0 | 5.2 | 1.0 | 1.2 | 0.4 | 3.1 | 0.0 | 1.2 | 0.4 |
| 500M5 | 5.7 | 1.0 | 5.5 | 1.0 | 5.7 | 1.1 | 0.6 | 0.2 | 0.7 | 0.3 | 0.6 | 0.2 |
| 500S4 | 5.1 | 1.1 | 4.9 | 0.0 | 5.1 | 1.1 | 1.4 | 0.4 | 2.9 | 0.0 | 1.4 | 0.4 |
| 500S5 | 5.6 | 0.7 | 5.9 | 1.0 | 5.7 | 0.8 | 0.6 | 0.2 | 0.9 | 0.2 | 0.6 | 0.2 |

Table S13: **SATé[24]'s performance relative to the next-best method on the moderate-to-difficult models.** Dunn-Ŝidák-corrected p-values are reported – to control the setwise error rate – for one-tailed, paired t-tests on SATé[24] and the two-phase method that came closest, on average, to matching SATé[24]'s performance. Results are reported for both missing branch rate and alignment (SP-FN) error ($n = 40$ for each test) for all of the moderate-to-difficult models. Models for which SATé[24] was significantly better than the next best tree estimator (setwise $\alpha = 0.05$) are shown in blue. Models in which SATé[24] was significantly better (setwise $\alpha = 0.05$) than the next best alignment estimator are marked with an asterisk (*).

| | Improvement in missing branch rate | | Improvement in alignment (SP-FN) error | |
|---|---|---|---|---|
| **Model** | **Next best tree** | **Dunn-Ŝidák p-value** | **Next best alignment** | **Dunn-Ŝidák p-value** |
| 1000L1* | RAxML(Muscle) | <0.0001 | Muscle | 0.0108 |
| 1000L2 | RAxML(Muscle) | 0.0014 | Prank+GT | 0.8958 |
| 1000L3 | RAxML(MAFFT) | <0.0001 | Muscle | 0.6427 |
| 1000M1 | RAxML(Prank+GT) | <0.0001 | Prank+GT | 0.3948 |
| 1000M2* | RAxML(Prank+GT) | <0.0001 | Prank+GT | 0.0007 |
| 1000M3 | RAxML(MAFFT) | 0.1982 | Prank+GT | 0.2715 |
| 1000S1 | RAxML(Prank+GT) | <0.0001 | Prank+GT | 0.6952 |
| 1000S2* | RAxML(Prank+GT) | <0.0001 | Prank+GT | 0.0112 |
| 1000S3 | RAxML(Prank+GT) | <0.0001 | Prank+GT | 0.0848 |
| 500L1* | RAxML(MAFFT) | <0.0001 | MAFFT | <0.0001 |
| 500L2* | RAxML(MAFFT) | <0.0001 | MAFFT | <0.0001 |
| 500L3* | RAxML(MAFFT) | 0.0004 | MAFFT | 0.0001 |
| 500M1* | RAxML(MAFFT) | <0.0001 | MAFFT | <0.0001 |
| 500M2* | RAxML(MAFFT) | <0.0001 | Prank+GT | 0.0016 |
| 500M3* | RAxML(MAFFT) | 0.0037 | Prank+GT | 0.0056 |
| 500S1* | RAxML(MAFFT) | 0.0011 | Prank+GT | 0.0117 |
| 500S2 | RAxML(MAFFT) | <0.0001 | Prank+GT | 0.9503 |
| 500S3* | RAxML(MAFFT) | 0.0070 | Prank+GT | 0.0098 |

Table S14: **p-values for Spearman rank-correlation analyses for the 1000-taxon moderate-to-difficult models.** The $p$-values for the Spearman rank correlation coefficients for the normalized maximum likelihood scores and the missing branch rates and for the normalized maximum likelihood scores and the alignment SP-FN errors are shown for the moderate-to-difficult models. We normalized the log likelihood scores of trees estimated for each alignment by dividing them by the log likelihood score of the RAxML(MAFFT) tree. We computed the missing branch rates of maximum likelihood trees computed for the true alignment, MAFFT, Prank+GT, Muscle, and ClustalW, as well as for all trees generated during SATé[24] analyses. We also computed the alignment SP-FN error rates for MAFFT, Prank+GT, Muscle, ClustalW, and SATé[24] alignments. Thus, for every dataset, we computed statistics for a varying number of alignments and trees. For these 1000-taxon datasets, the number of alignments and trees for which we computed statistics ranged from 9 to 15 with an average of 11.4 for each dataset. We calculated these statistics for each dataset, and we reported aggregate statistics for the datasets within each model. The average ($n = 20$ for each average), minimum, and maximum p-values for the Spearman rank-correlation coefficients relating normalized likelihood scores with missing branch rate and alignment error for the moderate-to-difficult models are shown. Maxima and minima are from the 20 replicates per model. Models in which the correlation between the maximum likelihood score and the missing branch rate was significant (p-value under $\alpha = 0.05$) are marked in blue, and models in which the correlation between the ML score and the alignment (SP-FN) error rate is significant are marked with an asterisk (*).

| Model | p-value for norm. likelihood vs. missing branch rate | | | p-value for norm. likelihood vs. alignment SP-FN error | | |
|---|---|---|---|---|---|---|
| | Average | Minimum | Maximum | Average | Minimum | Maximum |
| 1000L1* | 0.0030 | <0.0001 | 0.0255 | 0.0010 | <0.0001 | 0.0113 |
| 1000L2* | 0.0025 | <0.0001 | 0.0138 | 0.0002 | <0.0001 | 0.0022 |
| 1000L3* | 0.0161 | <0.0001 | 0.1872 | 0.0102 | <0.0001 | 0.0985 |
| 1000M1* | 0.0087 | <0.0001 | 0.1206 | 0.0128 | <0.0001 | 0.0878 |
| 1000M2* | 0.0032 | <0.0001 | 0.0275 | 0.0056 | <0.0001 | 0.0558 |
| 1000M3* | 0.0131 | <0.0001 | 0.1065 | 0.0040 | <0.0001 | 0.0293 |
| 1000S1* | 0.0082 | <0.0001 | 0.0725 | 0.0206 | <0.0001 | 0.1205 |
| 1000S2* | 0.0030 | <0.0001 | 0.0225 | 0.0026 | <0.0001 | 0.0350 |
| 1000S3* | 0.0021 | <0.0001 | 0.0092 | 0.0040 | <0.0001 | 0.0483 |

Table S15: **p-values for Spearman rank-correlation analyses for the 500-taxon moderate-to-difficult models.** The $p$-values for the Spearman rank correlation coefficients for the normalized maximum likelihood scores and the missing branch rates and for the normalized maximum likelihood scores and the alignment SP-FN errors are shown for the 500-taxon moderate-to-difficult models, as described inTable S14. For these 500-taxon datasets, the number of alignments and trees for which we computed statistics ranged from 15 to 35 with an average of 24.6 for each dataset. The values in the table are reported for $n = 20$ datasets per model. Models in which the correlation between the maximum likelihood score and the missing branch rate was significant (p-value under $\alpha = 0.05$) are marked in blue, and models in which the correlation between the ML score and the alignment (SP-FN) error rate is significant are marked with an asterisk (*).

| Model | p-value for norm. likelihood vs. missing branch rate | | | p-value for norm. likelihood vs. alignment SP-FN error | | |
|---|---|---|---|---|---|---|
| | Average | Minimum | Maximum | Average | Minimum | Maximum |
| 500L1* | 0.0044 | <0.0001 | 0.0582 | 0.0002 | <0.0001 | 0.0012 |
| 500L2* | 0.0044 | <0.0001 | 0.0232 | 0.0010 | <0.0001 | 0.0114 |
| 500L3* | 0.0375 | <0.0001 | 0.5064 | <0.0001 | <0.0001 | 0.0001 |
| 500M1* | 0.0017 | <0.0001 | 0.0282 | 0.0027 | <0.0001 | 0.0434 |
| 500M2* | 0.0045 | <0.0001 | 0.0193 | 0.0025 | <0.0001 | 0.0310 |
| 500M3* | 0.0127 | <0.0001 | 0.0783 | 0.0006 | <0.0001 | 0.0064 |
| 500S1* | 0.0184 | <0.0001 | 0.2520 | 0.0288 | <0.0001 | 0.3302 |
| 500S2* | 0.0124 | <0.0001 | 0.1675 | 0.0045 | <0.0001 | 0.0306 |
| 500S3* | 0.0898 | <0.0001 | 0.6456 | 0.0007 | <0.0001 | 0.0093 |

Table S16: **p-values for Spearman rank-correlation analyses for the easy models.** The $p$-values for the Spearman rank correlation coefficients for the normalized maximum likelihood scores and the missing branch rates and for the normalized maximum likelihood scores and alignment SP-FN errors are shown for the easy models, as described in Table S14. For these datasets, the number $n$ of alignments and trees for which we computed statistics ranged from 12 to 52 with an average of 28.3 for each dataset. The values in the table are reported for $n = 20$ datasets per model. Models in which the correlation between the ML score and the alignment (SP-FN) error rate is significant are marked with an asterisk (*).

| Model | p-value for norm. likelihood vs. missing branch rate | | | p-value for norm. likelihood vs. alignment SP-FN error | | |
|---|---|---|---|---|---|---|
| | Average | Minimum | Maximum | Average | Minimum | Maximum |
| 1000L4* | 0.2867 | 0.0029 | 0.9494 | 0.0015 | <0.0001 | 0.0102 |
| 1000L5* | 0.4090 | 0.0253 | 0.9797 | 0.0008 | <0.0001 | 0.0115 |
| 1000M4* | 0.1948 | 0.0027 | 0.8992 | 0.0015 | <0.0001 | 0.0289 |
| 1000M5* | 0.3808 | 0.0055 | 0.9444 | 0.0071 | <0.0001 | 0.0741 |
| 1000S4* | 0.2514 | 0.0079 | 0.8344 | 0.0001 | <0.0001 | 0.0008 |
| 1000S5 | 0.4951 | 0.0265 | 0.9839 | 0.0660 | <0.0001 | 0.7758 |
| 500L4 | 0.3607 | <0.0001 | 0.9951 | 0.0523 | <0.0001 | 0.5805 |
| 500L5* | 0.3977 | 0.0027 | 0.8525 | 0.0018 | <0.0001 | 0.0228 |
| 500M4* | 0.2905 | 0.0165 | 0.9984 | 0.0046 | <0.0001 | 0.0878 |
| 500M5* | 0.4340 | 0.0272 | 0.9907 | 0.0018 | <0.0001 | 0.0310 |
| 500S4* | 0.3627 | 0.0071 | 0.8314 | 0.0024 | <0.0001 | 0.0202 |
| 500S5* | 0.3662 | 0.0151 | 0.9814 | 0.0026 | <0.0001 | 0.0413 |

Table S17: **Runtime in hours for variants of SATé.** SATé$^{BML}$ is the best likelihood method out of the set of SATé$^{24}$, SATé$^{24}$(C), and SATé*(C). $n = 20$ for each value in the table.

| | **SATé$^{24}$(C)** | | | **SATé*(C)** | | | **SATé$^{24}$** | | | **SATé$^{BML}$** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Avg** | **Max** | **Min** | **Avg** | **Max** | **Min** | **Avg** | **Max** | **Min** | **Avg** | **Max** | **Min** |
| 1000L1 | 42.4 | 53.1 | 36.5 | 87.8 | 148.1 | 65.2 | 89.3 | 113.7 | 70.9 | 162.3 | 228.3 | 125.7 |
| 1000L2 | 38.1 | 43.8 | 33.5 | 85.7 | 141.7 | 59.0 | 72.5 | 95.6 | 58.4 | 146.5 | 199.7 | 114.2 |
| 1000L3 | 42.1 | 49.2 | 36.4 | 112.6 | 167.0 | 66.0 | 106.9 | 140.4 | 87.2 | 205.3 | 287.2 | 141.1 |
| 1000L4 | 33.7 | 34.7 | 31.9 | 88.9 | 207.8 | 56.3 | 46.6 | 49.9 | 44.7 | 127.1 | 246.2 | 93.9 |
| 1000L5 | 33.8 | 36.0 | 32.4 | 87.4 | 135.1 | 57.3 | 45.3 | 48.3 | 42.6 | 124.0 | 172.5 | 93.3 |
| 1000M1 | 40.5 | 44.5 | 36.1 | 119.6 | 169.7 | 67.9 | 89.2 | 119.3 | 69.9 | 195.7 | 260.7 | 133.3 |
| 1000M2 | 38.5 | 45.3 | 30.9 | 85.6 | 149.5 | 56.6 | 76.9 | 100.1 | 61.3 | 151.4 | 226.1 | 112.5 |
| 1000M3 | 32.8 | 35.5 | 30.9 | 80.0 | 116.0 | 57.7 | 58.0 | 67.9 | 50.5 | 131.2 | 170.3 | 105.8 |
| 1000M4 | 34.1 | 37.1 | 32.1 | 79.7 | 158.5 | 58.0 | 49.8 | 55.0 | 45.2 | 120.6 | 200.5 | 96.1 |
| 1000M5 | 33.8 | 35.5 | 32.6 | 81.2 | 158.0 | 57.1 | 45.5 | 47.0 | 43.4 | 118.1 | 194.5 | 93.3 |
| 1000S1 | 37.4 | 42.5 | 33.1 | 96.8 | 172.1 | 62.9 | 73.3 | 92.4 | 60.3 | 158.8 | 228.3 | 114.7 |
| 1000S2 | 35.8 | 39.9 | 31.5 | 89.7 | 190.9 | 58.7 | 64.0 | 77.4 | 51.0 | 143.7 | 250.2 | 103.4 |
| 1000S3 | 34.5 | 38.9 | 31.6 | 83.5 | 140.5 | 57.4 | 59.9 | 68.7 | 51.2 | 135.0 | 197.1 | 106.2 |
| 1000S4 | 33.4 | 35.5 | 31.1 | 79.5 | 184.1 | 55.4 | 44.8 | 49.9 | 41.8 | 116.0 | 219.0 | 90.5 |
| 1000S5 | 33.8 | 35.5 | 32.2 | 89.7 | 229.6 | 57.5 | 44.6 | 50.8 | 42.3 | 125.6 | 265.5 | 92.4 |
| 500L1 | 28.9 | 31.0 | 26.5 | 61.6 | 82.4 | 51.1 | 44.4 | 50.5 | 37.5 | 102.2 | 126.2 | 86.4 |
| 500L2 | 27.8 | 29.8 | 26.2 | 64.4 | 129.4 | 50.9 | 40.0 | 45.7 | 37.0 | 101.7 | 165.9 | 86.2 |
| 500L3 | 26.9 | 27.8 | 26.2 | 51.8 | 53.1 | 50.2 | 36.6 | 43.2 | 33.5 | 86.2 | 93.4 | 82.3 |
| 500L4 | 27.0 | 27.6 | 26.1 | 59.2 | 100.4 | 50.9 | 34.2 | 35.9 | 32.4 | 90.9 | 132.2 | 81.4 |
| 500L5 | 26.9 | 27.7 | 26.4 | 67.4 | 102.2 | 50.6 | 32.0 | 33.9 | 30.5 | 97.0 | 131.0 | 79.1 |
| 500M1 | 27.7 | 29.3 | 26.3 | 56.3 | 78.7 | 51.0 | 38.8 | 43.3 | 34.7 | 92.4 | 117.6 | 83.7 |
| 500M2 | 26.9 | 28.9 | 25.9 | 56.7 | 77.5 | 50.2 | 36.4 | 39.7 | 34.5 | 90.9 | 112.5 | 83.8 |
| 500M3 | 26.8 | 27.9 | 25.9 | 57.7 | 101.4 | 50.4 | 35.5 | 38.6 | 32.3 | 91.1 | 134.7 | 81.9 |
| 500M4 | 27.5 | 28.6 | 26.6 | 71.5 | 124.1 | 51.7 | 33.7 | 37.1 | 31.7 | 102.3 | 153.4 | 82.1 |
| 500M5 | 27.0 | 28.2 | 26.3 | 73.5 | 173.3 | 50.6 | 31.7 | 33.7 | 30.4 | 102.7 | 201.3 | 79.6 |
| 500S1 | 27.1 | 28.6 | 25.9 | 54.3 | 79.6 | 50.8 | 37.0 | 42.6 | 33.7 | 89.0 | 113.1 | 84.1 |
| 500S2 | 26.6 | 28.2 | 25.6 | 51.2 | 52.8 | 50.2 | 34.1 | 37.0 | 32.5 | 83.4 | 86.3 | 81.2 |
| 500S3 | 27.2 | 29.4 | 26.2 | 57.0 | 100.8 | 50.7 | 35.0 | 37.5 | 32.7 | 89.7 | 134.9 | 81.6 |
| 500S4 | 27.2 | 28.3 | 26.5 | 63.9 | 100.8 | 51.0 | 32.5 | 34.1 | 30.9 | 93.8 | 131.3 | 79.7 |
| 500S5 | 26.9 | 27.6 | 26.3 | 72.0 | 149.4 | 50.7 | 31.4 | 33.2 | 30.3 | 101.0 | 179.8 | 79.0 |

Table S18: **Alignment SP-FN errors for biological datasets with respect to the curated alignment.** SATé$^{BML}$ is the best likelihood method for SATé run until no improvements can be found with CT-5 proposals and with either all two-phase starting/tree alignment pairs or just RAxML(ClustalW). "Top 2 average" refers to the average of the 23SM.aa_ag and 23S.M datasets. "Top 4 average" refers to the average of the 23SM.aa_ag, 23S.M, 16S.M, and 16S.M.aa_ag datasets. "All 6 average" refers to the average of all of the biological datasets. $n = 1$ for all values.

| Dataset | SATé$^{BML}$ | MAFFT | Prank+GT | Muscle | Clustal | Two-phase average |
|---|---|---|---|---|---|---|
| 23S.M.aa_ag | 28.4 | 28.3 | 44.1 | 35.6 | 47.6 | 38.9 |
| 23S.M | 29.3 | 28.6 | 44.9 | 34.5 | 46.2 | 38.6 |
| 16S.M | 22.0 | 21.8 | 42.4 | 32.0 | 42.6 | 34.7 |
| 16S.M.aa_ag | 22.7 | 22.6 | 40.7 | 31.1 | 38.2 | 33.2 |
| 23S.E | 21.2 | 18.5 | 35.5 | 21.5 | 38.5 | 28.5 |
| 23S.E.aa_ag | 22.2 | 19.5 | 37.3 | 22.8 | 30.0 | 27.4 |
| **Top 2 average** | 28.9 | 28.4 | 44.5 | 35.0 | 46.9 | 38.7 |
| **Top 4 average** | 25.6 | 25.3 | 43.0 | 33.3 | 43.7 | 36.3 |
| **All 6 average** | 24.3 | 23.2 | 40.8 | 29.6 | 40.5 | 33.5 |

Table S19: **Missing branch rates for estimated trees on biological datasets with respect to three reference trees.** The three reference trees are the maximum likelihood trees estimated by RAxML on the curated alignment, keeping all edges with bootstrap support at least 90%, 75%, or 50%. SATé$^{BML}$ is the best likelihood method for SATé run until no improvements can be found with CT-5 proposals and with either all two-phase starting/tree alignment pairs or just RAxML(ClustalW). "Top 2 average" refers to the average of the 23SM.aa_ag and 23S.M datasets. "Top 4 average" refers to the average of the 23SM.aa_ag, 23S.M, 16S.M, and 16S.M.aa_ag datasets. "All 6 average" refers to the average of all of the biological datasets. $n = 1$ for all values in the table.

| | **Reference tree: 90% RAxML(CuratedAln) tree** **Missing branch rate (%)** | | | | | |
|---|---|---|---|---|---|---|
| **Dataset** | **SATé$^{BML}$** | **MAFFT** | **Prank+GT** | **Muscle** | **Clustal** | **Two-phase average** |
| 23S.M.aa_ag | 3.8 | 6.2 | 14.6 | 13.1 | 7.7 | 10.4 |
| 23S.M | 4.5 | 5.2 | 8.2 | 9.7 | 9.0 | 8.0 |
| 16S.M | 3.1 | 2.2 | 8.4 | 14.4 | 6.6 | 7.9 |
| 16S.M.aa_ag | 2.7 | 2.7 | 5.7 | 15.4 | 7.5 | 7.8 |
| 23S.E | 1.5 | 3.0 | 4.5 | 4.5 | 16.7 | 7.2 |
| 23S.E.aa_ag | 4.9 | 6.1 | 25.6 | 9.8 | 19.5 | 15.2 |
| **Top 2 average** | 4.2 | 5.7 | 11.4 | 11.4 | 8.3 | 9.2 |
| **Top 4 average** | 3.5 | 4.1 | 9.2 | 13.1 | 7.7 | 8.5 |
| **All 6 average** | 3.4 | 4.2 | 11.2 | 11.1 | 11.2 | 9.4 |
| | **Reference tree: 75% RAxML(CuratedAln) tree** **Missing branch rate (%)** | | | | | |
| **Dataset** | **SATé$^{BML}$** | **MAFFT** | **Prank+GT** | **Muscle** | **Clustal** | **Two-phase average** |
| 23S.M.aa_ag | 8.3 | 10.9 | 19.2 | 17.3 | 16.0 | 15.9 |
| 23S.M | 11.3 | 11.9 | 14.3 | 14.9 | 15.5 | 14.1 |
| 16S.M | 5.9 | 5.7 | 11.4 | 20.2 | 10.9 | 12.1 |
| 16S.M.aa_ag | 5.1 | 4.2 | 7.6 | 21.5 | 12.5 | 11.4 |
| 23S.E | 4.0 | 6.7 | 6.7 | 6.7 | 18.7 | 9.7 |
| 23S.E.aa_ag | 6.6 | 7.7 | 27.5 | 11.0 | 23.1 | 17.3 |
| **Top 2 average** | 9.8 | 11.4 | 16.8 | 16.1 | 15.8 | 15.0 |
| **Top 4 average** | 7.7 | 8.2 | 13.1 | 18.5 | 13.7 | 13.4 |
| **All 6 average** | 6.9 | 7.8 | 14.4 | 15.3 | 16.1 | 13.4 |
| | **Reference tree: 50% RAxML(CuratedAln) tree** **Missing branch rate (%)** | | | | | |
| **Dataset** | **SATé$^{BML}$** | **MAFFT** | **Prank** | **Muscle** | **Clustal** | **Two-phase average** |
| 23S.M.aa_ag | 15.2 | 16.7 | 24.5 | 25.5 | 22.1 | 22.2 |
| 23S.M | 17.4 | 17.9 | 19.9 | 17.9 | 20.9 | 19.2 |
| 16S.M | 13.0 | 12.0 | 17.4 | 28.4 | 17.8 | 18.9 |
| 16S.M.aa_ag | 10.8 | 11.0 | 13.9 | 28.1 | 18.8 | 17.9 |
| 23S.E | 5.7 | 10.2 | 11.4 | 10.2 | 23.9 | 13.9 |
| 23S.E.aa_ag | 12.8 | 16.5 | 33.9 | 20.2 | 32.1 | 25.7 |
| **Top 2 average** | 16.3 | 17.3 | 22.2 | 21.7 | 21.5 | 20.7 |
| **Top 4 average** | 14.1 | 14.4 | 18.9 | 25.0 | 19.9 | 19.5 |
| **All 6 average** | 12.5 | 14.0 | 20.2 | 21.7 | 22.6 | 19.6 |

Table S20: **Comparison of SATé$^{24}$ with CT-5 proposals versus SATé$^{24}$ with CT-1 proposals on moderate-to-difficult simulated datasets.** Normalized likelihoods (normalized by RAxML(MAFFT) likelihood), missing branch rates, alignment SP-FN errors, and runtimes for SATé$^{24}$ with CT-5 proposals and SATé$^{24}$ with CT-1 proposals. Standard deviations are shown for normalized likelihoods and runtimes, and standard errors are shown for missing branch rates and alignment SP-FN errors ($n = 20$ for all values).

| | **Normalized log likelihood (%)** | | | | **Missing branch rate (%)** | | | | **Alignment SP-FN error (%)** | | | | **Runtime (hours)** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **SATé$^{24}$ with CT-5** | | **SATé$^{24}$ with CT-1** | | **SATé$^{24}$ with CT-5** | | **SATé$^{24}$ with CT-1** | | **SATé$^{24}$ with CT-5** | | **SATé$^{24}$ with CT-1** | | **SATé$^{24}$ with CT-5** | | **SATé$^{24}$ with CT-1** | |
| **Model** | **Avg** | **Std dev** | **Avg** | **Std dev** | **Avg** | **Std err** | **Avg** | **Std err** | **Avg** | **Std err** | **Avg** | **Std err** | **Avg** | **Std dev** | **Avg** | **Std dev** |
| 1000L1 | 97.77673 | 0.81122 | 98.98487 | 0.43190 | 12.2 | 0.3 | 19.3 | 1.7 | 17.2 | 1.1 | 26.4 | 2.0 | 89.3 | 12.1 | 90.1 | 13.3 |
| 1000L2 | 98.34172 | 1.02367 | 98.93070 | 0.62447 | 10.8 | 0.4 | 14.4 | 1.4 | 6.6 | 0.7 | 11.7 | 1.6 | 72.5 | 10.8 | 74.1 | 11.6 |
| 1000L3 | 98.12181 | 0.53100 | 99.30126 | 0.38647 | 15.8 | 1.2 | 22.3 | 1.3 | 36.8 | 1.5 | 45.1 | 2.0 | 106.9 | 14.0 | 108.9 | 13.6 |
| 1000M1 | 97.18962 | 1.13773 | 98.17254 | 1.26749 | 17.4 | 0.9 | 27.2 | 1.7 | 39.3 | 1.5 | 43.5 | 1.8 | 89.2 | 13.0 | 92.0 | 13.0 |
| 1000M2 | 98.21738 | 0.54477 | 99.03781 | 0.68216 | 13.4 | 0.8 | 19.7 | 1.5 | 32.5 | 1.9 | 36.7 | 2.1 | 76.9 | 11.5 | 77.7 | 10.5 |
| 1000M3 | 99.31944 | 0.28622 | 99.47553 | 0.36749 | 8.7 | 0.4 | 9.7 | 0.5 | 9.5 | 0.6 | 9.4 | 0.6 | 58.0 | 4.3 | 59.2 | 4.1 |
| 1000S1 | 97.31542 | 0.93642 | 97.86084 | 0.93162 | 15.2 | 0.6 | 22.0 | 1.3 | 35.5 | 1.6 | 34.7 | 2.1 | 73.3 | 8.7 | 75.2 | 9.6 |
| 1000S2 | 98.19578 | 0.67528 | 98.72529 | 0.76377 | 11.3 | 0.4 | 15.2 | 1.0 | 16.8 | 1.0 | 19.1 | 1.4 | 64.0 | 7.4 | 64.9 | 7.6 |
| 1000S3 | 98.29939 | 0.86298 | 98.61158 | 0.90487 | 10.9 | 0.5 | 13.7 | 0.8 | 15.9 | 0.8 | 15.6 | 0.9 | 59.9 | 4.4 | 62.1 | 4.8 |
| 500L1 | 97.88506 | 1.09651 | 99.02303 | 0.88606 | 11.4 | 0.8 | 13.8 | 1.3 | 27.3 | 1.5 | 31.6 | 2.1 | 44.4 | 3.5 | 44.5 | 3.8 |
| 500L2 | 98.71665 | 0.64857 | 99.44499 | 0.50816 | 9.6 | 0.4 | 10.6 | 0.4 | 24.7 | 1.0 | 26.2 | 0.9 | 40.0 | 2.1 | 40.7 | 1.7 |
| 500L3 | 99.27586 | 0.32729 | 99.66086 | 0.22633 | 8.8 | 0.3 | 9.6 | 0.6 | 9.2 | 0.6 | 10.2 | 0.8 | 36.6 | 2.3 | 37.1 | 2.1 |
| 500M1 | 97.29438 | 1.31315 | 98.35116 | 1.39257 | 13.0 | 0.5 | 15.1 | 0.8 | 31.9 | 1.1 | 33.0 | 1.6 | 38.8 | 2.0 | 39.0 | 2.3 |
| 500M2 | 98.61660 | 0.58778 | 99.29137 | 0.50391 | 10.0 | 0.4 | 12.4 | 0.7 | 24.9 | 1.1 | 26.2 | 1.2 | 36.4 | 1.4 | 36.7 | 1.8 |
| 500M3 | 98.94030 | 0.49606 | 99.28160 | 0.43894 | 9.0 | 0.3 | 9.8 | 0.4 | 10.2 | 0.6 | 10.1 | 0.8 | 35.5 | 1.5 | 35.8 | 1.5 |
| 500S1 | 97.53346 | 0.73006 | 98.16044 | 0.96684 | 12.4 | 0.6 | 14.9 | 0.9 | 30.5 | 1.4 | 30.4 | 1.6 | 37.0 | 2.3 | 37.4 | 2.7 |
| 500S2 | 97.89945 | 1.30475 | 98.27918 | 1.50646 | 10.3 | 0.4 | 12.1 | 0.6 | 23.1 | 1.2 | 21.9 | 1.1 | 34.1 | 1.1 | 34.4 | 1.3 |
| 500S3 | 98.75497 | 0.48868 | 99.08902 | 0.45624 | 8.4 | 0.3 | 9.1 | 0.4 | 10.6 | 0.6 | 10.7 | 0.7 | 35.0 | 1.3 | 35.5 | 1.1 |

Table S21: **Comparison of SATé$^{24}$ with CT-5 proposals versus SATé$^{24}$ with CT-1 proposals on easy simulated datasets.** Normalized likelihoods (normalized by RAxML(MAFFT) likelihood), missing branch rates, alignment SP-FN errors, and runtimes for SATé$^{24}$ with CT-5 proposals and SATé$^{24}$ with CT-1 proposals. Standard deviations are shown for normalized likelihoods and runtimes, and standard errors are shown for missing branch rates and alignment SP-FN errors ($n = 20$ for all values).

| | **Normalized log likelihood (%)** | | | | **Missing branch rate (%)** | | | | **Alignment SP-FN error (%)** | | | | **Runtime (hours)** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **SATé$^{24}$ with CT-5** | | **SATé$^{24}$ with CT-1** | | **SATé$^{24}$ with CT-5** | | **SATé$^{24}$ with CT-1** | | **SATé$^{24}$ with CT-5** | | **SATé$^{24}$ with CT-1** | | **SATé$^{24}$ with CT-5** | | **SATé$^{24}$ with CT-1** | |
| **Model** | **Avg** | **Std dev** | **Avg** | **Std dev** | **Avg** | **Std err** | **Avg** | **Std err** | **Avg** | **Std err** | **Avg** | **Std err** | **Avg** | **Std dev** | **Avg** | **Std dev** |
| 1000L4 | 99.95265 | 0.04142 | 99.95320 | 0.04023 | 4.9 | 0.2 | 4.9 | 0.2 | 0.5 | 0.0 | 0.5 | 0.0 | 46.6 | 1.3 | 47.4 | 1.8 |
| 1000L5 | 99.98229 | 0.01194 | 99.98571 | 0.00994 | 5.2 | 0.2 | 5.2 | 0.2 | 0.2 | 0.0 | 0.2 | 0.0 | 45.3 | 1.3 | 46.2 | 2.0 |
| 1000M4 | 99.84962 | 0.07343 | 99.84962 | 0.07343 | 5.0 | 0.1 | 5.0 | 0.1 | 0.8 | 0.0 | 0.8 | 0.0 | 49.8 | 1.9 | 50.6 | 2.4 |
| 1000M5 | 99.93819 | 0.06837 | 99.95049 | 0.03163 | 5.1 | 0.2 | 5.1 | 0.2 | 0.5 | 0.1 | 0.4 | 0.0 | 45.5 | 0.9 | 46.3 | 1.3 |
| 1000S4 | 99.94727 | 0.04083 | 99.94773 | 0.04122 | 5.4 | 0.2 | 5.4 | 0.2 | 0.4 | 0.0 | 0.4 | 0.0 | 44.8 | 1.8 | 45.4 | 1.9 |
| 1000S5 | 99.98051 | 0.01676 | 99.98390 | 0.01325 | 4.9 | 0.2 | 4.8 | 0.2 | 0.2 | 0.0 | 0.2 | 0.0 | 44.6 | 1.7 | 44.7 | 1.5 |
| 500L4 | 99.85457 | 0.05501 | 99.87023 | 0.06075 | 5.4 | 0.3 | 5.5 | 0.2 | 1.9 | 0.2 | 1.7 | 0.1 | 34.2 | 0.9 | 34.6 | 1.1 |
| 500L5 | 99.94990 | 0.03571 | 99.95813 | 0.02962 | 5.1 | 0.2 | 5.2 | 0.2 | 0.6 | 0.0 | 0.6 | 0.0 | 32.0 | 0.8 | 32.1 | 0.8 |
| 500M4 | 99.84460 | 0.07621 | 99.84529 | 0.07696 | 5.2 | 0.2 | 5.1 | 0.2 | 1.2 | 0.1 | 1.2 | 0.1 | 33.7 | 1.4 | 34.2 | 1.5 |
| 500M5 | 99.94475 | 0.03156 | 99.95090 | 0.03163 | 5.7 | 0.2 | 5.5 | 0.2 | 0.6 | 0.0 | 0.6 | 0.0 | 31.7 | 1.0 | 32.1 | 1.1 |
| 500S4 | 99.80381 | 0.09301 | 99.80302 | 0.09158 | 5.1 | 0.3 | 5.0 | 0.2 | 1.4 | 0.1 | 1.4 | 0.1 | 32.5 | 1.0 | 32.4 | 0.9 |
| 500S5 | 99.94011 | 0.03661 | 99.94185 | 0.03778 | 5.7 | 0.2 | 5.6 | 0.2 | 0.6 | 0.0 | 0.6 | 0.0 | 31.4 | 0.8 | 31.7 | 0.7 |

Table S22: **Log likelihoods and missing branch rates for SATé[24] run using different CT-$i$ proposals and different starting tree/alignment pairs.** SATé$^{BML}$ is the best likelihood method for SATé run until no improvements can be found with CT-5 proposals and with either all two-phase starting/tree alignment pairs or just RAxML(ClustalW). Average refers to the average for a statistic across all biological datasets. $n = 1$ for all values in the table.

| | **Log likelihood** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **SATé$^{BML}$** | **SATé* CT-5** | **SATé*(C) CT-5** | **SATé* CT-4** | **SATé*(C) CT-4** | **SATé* CT-3** | **SATé*(C) CT-3** | **SATé* CT-2** | **SATé*(C) CT-2** | **SATé* CT-1** | **SATé*(C) CT-1** |
| 23S.M.aa_ag | -215770.7 | -215830.6 | -215770.7 | -217796.7 | -216504.4 | -217959.5 | -216520.3 | -218270.6 | -218327.3 | -217638.2 | -218292.5 |
| 23S.M | -229244.1 | -229244.5 | -229244.1 | -230412.7 | -230413.7 | -230359.7 | -231154.1 | -231548.2 | -231551.6 | -232427.2 | -232441.4 |
| 16S.M | -265172.9 | -265703.6 | -265172.9 | -265664.1 | -265855.0 | -265365.1 | -265647.0 | -266217.5 | -267191.3 | -266112.9 | -266109.5 |
| 16S.M.aa_ag | -262755.4 | -262755.4 | -263035.7 | -263413.2 | -263040.3 | -264351.0 | -263595.3 | -264686.5 | -264574.5 | -264833.6 | -264389.7 |
| 23S.E | -179066.0 | -179493.1 | -179066.0 | -180254.3 | -180171.0 | -180634.9 | -180744.2 | -180697.4 | -180914.2 | -181030.3 | -181233.4 |
| 23S.E.aa_ag | -178412.0 | -178412.0 | -178610.3 | -179140.2 | -179186.9 | -180176.3 | -180106.4 | -180590.2 | -180625.6 | -180892.5 | -180756.7 |
| | **Reference tree: 90% RAxML(CuratedAln) tree Missing branch rate (%)** | | | | | | | | | | |
| **Model** | **SATé$^{BML}$** | **SATé* CT-5** | **SATé*(C) CT-5** | **SATé* CT-4** | **SATé*(C) CT-4** | **SATé* CT-3** | **SATé*(C) CT-3** | **SATé* CT-2** | **SATé*(C) CT-2** | **SATé* CT-1** | **SATé*(C) CT-1** |
| 23S.M.aa_ag | 3.8 | 4.6 | 3.8 | 7.7 | 3.8 | 8.5 | 4.6 | 8.5 | 5.4 | 5.4 | 4.6 |
| 23S.M | 4.5 | 4.5 | 4.5 | 5.2 | 5.2 | 5.2 | 3.7 | 3.7 | 3.7 | 3.0 | 3.7 |
| 16S.M | 3.1 | 2.8 | 3.1 | 2.8 | 3.1 | 2.8 | 2.5 | 3.1 | 3.1 | 3.1 | 2.8 |
| 16S.M.aa_ag | 2.7 | 2.7 | 3.3 | 3.6 | 3.3 | 2.7 | 3.6 | 2.4 | 3.3 | 2.7 | 3.0 |
| 23S.E | 1.5 | 3.0 | 1.5 | 3.0 | 1.5 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 1.5 |
| 23S.E.aa_ag | 4.9 | 4.9 | 6.1 | 4.9 | 6.1 | 6.1 | 4.9 | 4.9 | 6.1 | 8.5 | 7.3 |
| **Average** | 3.4 | 3.8 | 3.7 | 4.5 | 3.9 | 4.7 | 3.7 | 4.3 | 4.1 | 4.3 | 3.8 |
| | **Reference tree: 75% RAxML(CuratedAln) tree Missing branch rate (%)** | | | | | | | | | | |
| **Model** | **SATé$^{BML}$** | **SATé* CT-5** | **SATé*(C) CT-5** | **SATé* CT-4** | **SATé*(C) CT-4** | **SATé* CT-3** | **SATé*(C) CT-3** | **SATé* CT-2** | **SATé*(C) CT-2** | **SATé* CT-1** | **SATé*(C) CT-1** |
| 23S.M.aa_ag | 8.3 | 10.9 | 8.3 | 10.9 | 9.6 | 14.1 | 10.3 | 14.1 | 10.3 | 10.3 | 9.0 |
| 23S.M | 11.3 | 10.1 | 11.3 | 10.1 | 10.1 | 10.7 | 9.5 | 9.5 | 9.5 | 9.5 | 10.1 |
| 16S.M | 5.9 | 5.7 | 5.9 | 5.2 | 5.7 | 6.2 | 5.7 | 6.2 | 6.2 | 5.9 | 6.7 |
| 16S.M.aa_ag | 5.1 | 5.1 | 5.3 | 5.5 | 6.5 | 4.2 | 6.0 | 4.8 | 5.8 | 4.4 | 4.4 |
| 23S.E | 4.0 | 6.7 | 4.0 | 5.3 | 4.0 | 4.0 | 5.3 | 6.7 | 6.7 | 6.7 | 5.3 |
| 23S.E.aa_ag | 6.6 | 6.6 | 7.7 | 7.7 | 7.7 | 7.7 | 7.7 | 7.7 | 7.7 | 12.1 | 8.8 |
| **Average** | 6.9 | 7.5 | 7.1 | 7.5 | 7.3 | 7.8 | 7.4 | 8.2 | 7.7 | 8.1 | 7.4 |
| | **Reference tree: 50% RAxML(CuratedAln) tree Missing branch rate (%)** | | | | | | | | | | |
| **Model** | **SATé$^{BML}$** | **SATé* CT-5** | **SATé*(C) CT-5** | **SATé* CT-4** | **SATé*(C) CT-4** | **SATé* CT-3** | **SATé*(C) CT-3** | **SATé* CT-2** | **SATé*(C) CT-2** | **SATé* CT-1** | **SATé*(C) CT-1** |
| 23S.M.aa_ag | 15.2 | 17.2 | 15.2 | 17.2 | 15.2 | 20.1 | 15.2 | 19.6 | 15.2 | 17.2 | 14.7 |
| 23S.M | 17.4 | 15.9 | 17.4 | 15.4 | 15.4 | 15.4 | 14.4 | 14.9 | 15.9 | 13.9 | 15.4 |
| 16S.M | 13.0 | 13.0 | 13.0 | 12.1 | 12.1 | 12.8 | 11.8 | 13.0 | 12.6 | 13.3 | 12.8 |
| 16S.M.aa_ag | 10.8 | 10.8 | 11.0 | 10.7 | 12.1 | 9.9 | 12.0 | 10.7 | 11.3 | 9.6 | 10.8 |
| 23S.E | 5.7 | 11.4 | 5.7 | 9.1 | 6.8 | 8.0 | 8.0 | 11.4 | 11.4 | 10.2 | 8.0 |
| 23S.E.aa_ag | 12.8 | 12.8 | 16.5 | 12.8 | 12.8 | 15.6 | 12.8 | 12.8 | 14.7 | 16.5 | 18.3 |
| **Average** | 12.5 | 13.5 | 13.1 | 12.9 | 12.4 | 13.6 | 12.4 | 13.7 | 13.5 | 13.5 | 13.3 |

Table S23: **Alignment SP-FN errors and runtimes in hours for SATé run using different CT-$i$ proposals and different starting tree/alignment pairs.** SATé$^{BML}$ is the best likelihood method for SATé run until no improvements can be found with CT-5 proposals and with either all two-phase starting/tree alignment pairs or just RAxML(ClustalW). Average refers to the average for a statistic across all biological datasets. $n = 1$ for all values in the table.

| | **Alignment SP-FN error** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **SATé$^{BML}$** | **SATé* CT-5** | **SATé*(C) CT-5** | **SATé* CT-4** | **SATé*(C) CT-4** | **SATé* CT-3** | **SATé*(C) CT-3** | **SATé* CT-2** | **SATé*(C) CT-2** | **SATé* CT-1** | **SATé*(C) CT-1** |
| 23S.M.aa_ag | 28.4 | 29.9 | 28.4 | 26.7 | 28.6 | 28.1 | 26.5 | 27.7 | 27.8 | 27.0 | 28.6 |
| 23S.M | 29.3 | 29.3 | 29.3 | 27.0 | 27.0 | 26.9 | 27.5 | 26.6 | 26.6 | 28.4 | 28.4 |
| 16S.M | 22.0 | 21.3 | 22.0 | 21.3 | 21.1 | 20.5 | 20.7 | 21.4 | 22.0 | 21.3 | 22.1 |
| 16S.M.aa_ag | 22.7 | 22.7 | 23.2 | 22.4 | 22.5 | 21.9 | 22.6 | 22.4 | 22.7 | 22.5 | 22.9 |
| 23S.E | 21.2 | 22.5 | 21.2 | 20.7 | 20.7 | 20.1 | 20.1 | 19.7 | 19.4 | 18.5 | 19.8 |
| 23S.E.aa_ag | 22.2 | 22.2 | 21.5 | 22.4 | 20.9 | 19.9 | 22.7 | 21.4 | 19.9 | 19.8 | 19.9 |
| **Average** | 24.3 | 24.6 | 24.3 | 23.4 | 23.5 | 22.9 | 23.3 | 23.2 | 23.1 | 22.9 | 23.6 |
| | **Runtime in hours** | | | | | | | | | | |
| **Model** | **SATé$^{BML}$** | **SATé* CT-5** | **SATé*(C) CT-5** | **SATé* CT-4** | **SATé*(C) CT-4** | **SATé* CT-3** | **SATé*(C) CT-3** | **SATé* CT-2** | **SATé*(C) CT-2** | **SATé* CT-1** | **SATé*(C) CT-1** |
| 23S.M.aa_ag | 117.1 | 67.2 | 52.0 | 66.4 | 53.1 | 64.2 | 52.9 | 91.7 | 55.1 | 70.4 | 52.5 |
| 23S.M | 118.7 | 68.9 | 52.0 | 69.5 | 51.8 | 69.0 | 53.6 | 70.9 | 52.5 | 70.7 | 53.7 |
| 16S.M | 193.7 | 106.0 | 96.5 | 104.3 | 63.0 | 102.9 | 65.0 | 106.9 | 62.8 | 136.9 | 67.6 |
| 16S.M.aa_ag | 196.0 | 114.5 | 91.7 | 117.4 | 67.3 | 114.0 | 68.3 | 112.6 | 69.5 | 118.7 | 80.3 |
| 23S.E | 133.6 | 60.0 | 75.6 | 60.2 | 51.4 | 59.4 | 101.1 | 60.3 | 76.2 | 36.4 | 51.1 |
| 23S.E.aa_ag | 134.9 | 61.8 | 75.6 | 62.7 | 101.0 | 61.9 | 51.8 | 63.0 | 52.2 | 62.2 | 76.0 |
| **Average** | 149.0 | 79.7 | 73.9 | 80.1 | 64.6 | 78.6 | 65.4 | 84.3 | 61.4 | 82.6 | 63.5 |

Table S24: **Runtimes in hours for all two-phase methods on biological datasets.** Average refers to the average for a statistic across all biological datasets. $n = 1$ for all values in the table.

| Dataset | RAxML(MAFFT) | RAxML(Prank+GT) | RAxML(Muscle) | RAxML(ClustalW) | Total two-phase |
|---|---|---|---|---|---|
| 23S.M.aa_ag | 2.8 | 5.0 | 4.1 | 2.1 | 14.0 |
| 23S.M | 5.1 | 5.9 | 4.1 | 2.2 | 17.3 |
| 16S.M | 9.4 | 12.8 | 17.9 | 8.8 | 48.8 |
| 16S.M.aa_ag | 16.7 | 15.7 | 15.5 | 10.1 | 58.0 |
| 23S.E | 1.1 | 7.0 | 0.8 | 2.0 | 10.9 |
| 23S.E.aa_ag | 1.3 | 7.8 | 1.1 | 2.4 | 12.6 |
| **Average** | 6.1 | 9.0 | 7.3 | 4.6 | 26.9 |

Table S25: **Percent of datasets on which BAli-Phy crashed on 4GB machines due to memory limitations.** $n = 20$ for all values in the table.

| Model | Percent of datasets that BAli-Phy crashed on (%) |
|---|---|
| 100L1 | 15 |
| 100L2 | 35 |
| 100M1 | 30 |
| 100M2 | 15 |
| 100M3 | 30 |
| 100S1 | 20 |
| 100S2 | 50 |

Table S26: **Comparison of alignment errors for two-phase and coestimation methods.** SATé$^{BML}$ is the best likelihood method for SATé run until no improvements can be found with CT-5 proposals and with either all two-phase starting/tree alignment pairs or just RAxML(ClustalW). The four model conditions for which ALIFRITZ had not yet reported any ML trees are marked "N/A". We report results for both the posterior decoding alignment and the MAP alignment from BAli-Phy's MCMC walk. $n = 1$ for all values in the table.

| Model | SATé$^{BML}$ | MAFFT | Prank+GT | Muscle | ClustalW | BAli-Phy posterior-decoding | BAli-Phy MAP | ALIFRITZ |
|---|---|---|---|---|---|---|---|---|
| 100L1 | 21.3 | 20.1 | 41.7 | 30.6 | 54.1 | 12.4 | 15.3 | N/A |
| 100L2 | 1.7 | 1.9 | 1.7 | 2.4 | 12.9 | 1.1 | 1.8 | 11.0 |
| 100M1 | 31.8 | 29.2 | 63.2 | 39.3 | 56.9 | 29.0 | 29.0 | N/A |
| 100M2 | 12.1 | 17.5 | 13.1 | 15.9 | 39.1 | 6.1 | 7.9 | N/A |
| 100M3 | 3.3 | 4.0 | 3.1 | 3.3 | 8.5 | 2.3 | 3.0 | 16.4 |
| 100S1 | 27.8 | 29.4 | 35.5 | 39.4 | 40.9 | 12.0 | 13.6 | N/A |
| 100S2 | 13.4 | 19.5 | 13.4 | 18.0 | 27.3 | 6.9 | 9.2 | 47.5 |
| Average | 15.9 | 17.4 | 24.5 | 21.3 | 34.2 | 10.0 | 11.4 | N/A |

Table S27: **Comparison of tree errors for two-phase and coestimation methods.** SATé$^{BML}$ is the best likelihood method for SATé run until no improvements can be found with CT-5 proposals and with either all two-phase starting/tree alignment pairs or just RAxML(ClustalW). The four model conditions for which ALIFRITZ had not yet reported any ML trees are marked "N/A". We report results for both the majority consensus tree and MAP tree computed from from BAli-Phy's partially completed MCMC walk. Since the consensus tree is not usually binary, we also give the "FP" rate for BAli-Phy. $n = 1$ for all values in the table.

| | **Missing branch rate (%)** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **RAxML(TrueAln)** | **SATé$^{BML}$** | **RAxML(MAFFT)** | **RAxML(Prank+GT)** | **RAxML(Muscle)** | **RAxML(ClustalW)** | **BAli-Phy majority consensus** | **BAli-Phy MAP** | **ALIFRITZ** |
| 100L1 | 12.4 | 15.5 | 12.4 | 29.9 | 12.4 | 26.8 | 16.5 | 15.5 | N/A |
| 100L2 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 4.3 | 2.1 | 2.1 | 4.3 |
| 100M1 | 3.1 | 13.4 | 17.5 | 33.0 | 14.4 | 25.8 | 42.3 | 41.2 | N/A |
| 100M2 | 6.2 | 5.2 | 6.2 | 6.2 | 5.2 | 6.2 | 5.2 | 5.2 | N/A |
| 100M3 | 5.2 | 5.2 | 6.2 | 4.1 | 4.1 | 6.2 | 5.2 | 4.1 | 6.2 |
| 100S1 | 11.5 | 11.5 | 13.5 | 24.0 | 13.5 | 18.8 | 17.7 | 17.7 | N/A |
| 100S2 | 3.2 | 2.2 | 3.2 | 2.2 | 4.3 | 7.5 | 2.2 | 2.2 | 7.5 |
| Average | 6.2 | 7.8 | 8.7 | 14.5 | 8.0 | 13.6 | 13.0 | 12.6 | N/A |

| | **BAli-Phy majority consensus tree "FP" rate (%)** |
|---|---|
| 100L1 | 11.0 |
| 100L2 | 4.2 |
| 100M1 | 38.5 |
| 100M2 | 4.2 |
| 100M3 | 4.2 |
| 100S1 | 14.1 |
| 100S2 | 4.2 |
| Average | 11.5 |

Table S28: **Log likelihoods, missing branch rates, alignment SP-FN errors, and runtime in hours for SATé* run using different CT-$i$ proposals and different starting tree/alignment pairs for the second ALIFRITZ and BAli-Phy experiment.** SATé$^{BML}$ is the best likelihood method for SATé run until no improvements can be found with CT-5 proposals and with either all two-phase starting/tree alignment pairs or just RAxML(ClustalW). $n = 1$ for all values in the table.

| | **Log likelihood** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **SATé$^{BML}$** | **SATé* CT-5** | **SATé*(C) CT-5** | **SATé* CT-4** | **SATé*(C) CT-4** | **SATé* CT-3** | **SATé*(C) CT-3** | **SATé* CT-2** | **SATé*(C) CT-2** | **SATé* CT-1** | **SATé*(C) CT-1** |
| 100L1 | -94913.7 | -94913.7 | -94921.7 | -94934.7 | -94932.3 | -94932.7 | -94953.0 | -95070.8 | -95062.3 | -95630.3 | -95558.1 |
| 100L2 | -44938.8 | -44938.8 | -45106.6 | -44938.8 | -45129.3 | -44938.8 | -45120.4 | -44938.8 | -45161.8 | -44938.8 | -45176.8 |
| 100M1 | -91137.5 | -91137.5 | -91461.6 | -91185.3 | -91521.7 | -91397.3 | -91502.7 | -91718.9 | -91909.8 | -92224.8 | -92435.5 |
| 100M2 | -71253.7 | -71253.7 | -71257.7 | -71286.1 | -71286.1 | -71390.3 | -71387.1 | -71500.8 | -71508.6 | -71553.0 | -71562.6 |
| 100M3 | -51273.3 | -51273.3 | -51301.5 | -51283.4 | -51337.0 | -51283.4 | -51374.0 | -51283.4 | -51386.1 | -51283.4 | -51410.9 |
| 100S1 | -83525.2 | -83647.4 | -83525.2 | -83803.4 | -83556.6 | -84021.9 | -84077.9 | -84614.2 | -84397.8 | -85074.8 | -85077.7 |
| 100S2 | -46844.7 | -46844.7 | -48797.1 | -46844.7 | -48896.0 | -46844.7 | -49467.0 | -46844.7 | -49567.4 | -46844.7 | -49629.3 |
| | **Missing branch rate (%)** | | | | | | | | | | |
| **Model** | **SATé$^{BML}$** | **SATé* CT-5** | **SATé*(C) CT-5** | **SATé* CT-4** | **SATé*(C) CT-4** | **SATé* CT-3** | **SATé*(C) CT-3** | **SATé* CT-2** | **SATé*(C) CT-2** | **SATé* CT-1** | **SATé*(C) CT-1** |
| 100L1 | 15.5 | 15.5 | 14.4 | 13.4 | 12.4 | 12.4 | 13.4 | 12.4 | 11.3 | 11.3 | 9.3 |
| 100L2 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 |
| 100M1 | 13.4 | 13.4 | 19.6 | 11.3 | 18.6 | 13.4 | 16.5 | 9.3 | 21.6 | 15.5 | 17.5 |
| 100M2 | 5.2 | 5.2 | 6.2 | 5.2 | 5.2 | 6.2 | 5.2 | 6.2 | 5.2 | 6.2 | 5.2 |
| 100M3 | 5.2 | 5.2 | 5.2 | 4.1 | 5.2 | 4.1 | 4.1 | 4.1 | 5.2 | 4.1 | 6.2 |
| 100S1 | 11.5 | 17.7 | 11.5 | 15.6 | 17.7 | 17.7 | 13.5 | 14.6 | 14.6 | 24.0 | 18.8 |
| 100S2 | 2.2 | 2.2 | 5.4 | 2.2 | 5.4 | 2.2 | 4.3 | 2.2 | 3.2 | 2.2 | 4.3 |
| | **Alignment SP-FN error (%)** | | | | | | | | | | |
| **Model** | **SATé$^{BML}$** | **SATé* CT-5** | **SATé*(C) CT-5** | **SATé* CT-4** | **SATé*(C) CT-4** | **SATé* CT-3** | **SATé*(C) CT-3** | **SATé* CT-2** | **SATé*(C) CT-2** | **SATé* CT-1** | **SATé*(C) CT-1** |
| 100L1 | 21.3 | 21.3 | 19.7 | 16.4 | 15.8 | 13.7 | 14.7 | 15.4 | 14.6 | 16.6 | 16.4 |
| 100L2 | 1.7 | 1.7 | 2.4 | 1.7 | 2.1 | 1.7 | 1.9 | 1.7 | 2.1 | 1.7 | 2.3 |
| 100M1 | 31.8 | 31.8 | 37.0 | 29.8 | 35.7 | 28.5 | 27.9 | 28.1 | 30.8 | 25.9 | 26.2 |
| 100M2 | 12.1 | 12.1 | 12.1 | 10.6 | 10.6 | 12.4 | 12.3 | 11.0 | 11.2 | 13.1 | 13.4 |
| 100M3 | 3.3 | 3.3 | 3.8 | 3.1 | 3.8 | 3.1 | 3.9 | 3.1 | 3.9 | 3.1 | 3.9 |
| 100S1 | 27.8 | 26.3 | 27.8 | 26.8 | 28.1 | 26.7 | 29.1 | 26.0 | 26.0 | 35.5 | 40.9 |
| 100S2 | 13.4 | 13.4 | 19.0 | 13.4 | 20.6 | 13.4 | 17.1 | 13.4 | 18.5 | 13.4 | 19.3 |
| | **Runtime in hours** | | | | | | | | | | |
| **Model** | **SATé$^{BML}$** | **SATé* CT-5** | **SATé*(C) CT-5** | **SATé* CT-4** | **SATé*(C) CT-4** | **SATé* CT-3** | **SATé*(C) CT-3** | **SATé* CT-2** | **SATé*(C) CT-2** | **SATé* CT-1** | **SATé*(C) CT-1** |
| 100L1 | 99.6 | 51.4 | 49.1 | 75.4 | 49.0 | 51.4 | 49.2 | 51.5 | 49.2 | 51.4 | 49.2 |
| 100L2 | 97.1 | 25.1 | 72.2 | 25.1 | 48.3 | 25.1 | 48.3 | 25.1 | 48.4 | 25.2 | 48.3 |
| 100M1 | 97.3 | 49.0 | 48.4 | 49.0 | 48.2 | 49.0 | 48.3 | 49.1 | 48.2 | 49.1 | 48.4 |
| 100M2 | 98.7 | 50.6 | 48.7 | 50.6 | 48.7 | 50.6 | 48.7 | 50.5 | 48.7 | 26.5 | 48.8 |
| 100M3 | 96.8 | 48.7 | 48.2 | 24.7 | 48.2 | 24.6 | 48.2 | 24.6 | 48.2 | 24.6 | 48.2 |
| 100S1 | 99.1 | 51.0 | 48.9 | 51.0 | 48.9 | 51.1 | 49.0 | 51.1 | 49.0 | 26.9 | 24.9 |
| 100S2 | 73.0 | 24.8 | 48.3 | 24.8 | 48.3 | 24.8 | 48.3 | 24.9 | 48.4 | 24.9 | 48.3 |

Table S29: **Runtimes for two-phase methods and SATé$^{BML}$in experiments 5 and 6.** SATé$^{BML}$ is the best likelihood method for SATé run until no improvements can be found with CT-5 proposals and with either all two-phase starting/tree alignment pairs or just RAxML(ClustalW). $n = 1$ for all values in the table. The runtime for BAli-Phy and ALIFRITZ was 337.1 hours per dataset. This was insufficient for stationarity or convergence on any dataset.

| | Runtime (hours) | | | | | |
|---|---|---|---|---|---|---|
| **Model** | **RAxML(TrueAln)** | **SATé$^{BML}$** | **MAFFT** | **Prank+GT** | **Muscle** | **ClustalW** |
| 100L1 | 0.8 | 99.6 | 1.0 | 0.4 | 0.9 | 0.9 |
| 100L2 | 0.0 | 97.1 | 0.1 | 0.6 | 0.1 | 0.2 |
| 100M1 | 0.1 | 97.3 | 0.2 | 0.4 | 0.2 | 0.1 |
| 100M2 | 0.6 | 98.7 | 0.7 | 0.4 | 0.7 | 0.7 |
| 100M3 | 0.0 | 96.8 | 0.1 | 0.3 | 0.0 | 0.1 |
| 100S1 | 0.7 | 99.1 | 0.8 | 0.4 | 0.9 | 0.8 |
| 100S2 | 0.1 | 73.0 | 0.1 | 0.4 | 0.1 | 0.2 |
| Average | 0.3 | 94.5 | 0.4 | 0.4 | 0.4 | 0.4 |

Table S30: **Number of BAli-Phy iterations in two-week long analyses of first replicate dataset from each 100 taxon model.** $n = 1$ for all values in the table.

| Model | Iterations |
|---|---|
| 100L1 | 2531 |
| 100L2 | 2430 |
| 100M1 | 2329 |
| 100M2 | 2818 |
| 100M3 | 3397 |
| 100S1 | 2979 |
| 100S2 | 2852 |

Table S31: **Results for the relaxed Gblocks experiment.** Columns in the datasets were eliminated (masked) with relaxed Gblocks. Some datasets had all sites masked, making tree estimation and error calculation impossible. The first column shows the percentage ($n = 180$ for each moderate-to-difficult model condition and $n = 120$ for each easy model condition) of datasets with 100% sites masked. The second, third, and fourth data columns show the total average, minimum model condition average and maximum model condition average increase in missing branch rate for trees estimated on the remaining nonempty masked datasets relative to the missing branch rate for the datasets that included all columns in the alignment. We also report the average and standard deviation of percentage of sites that were masked in the alignments with some remaining sites.

| | | **Datasets with empty masked alignments (%)** | **Increase in missing branch rate (%)** | | | **Original sites masked (%)** | |
|---|---|---|---|---|---|---|---|
| | **Methods** | | **Average** | **Minimum** | **Maximum** | **Average** | **Std dev** |
| **1000-taxon moderate-to-difficult** | RAxML(TrueAln) | 0 | 29.6 | 14.2 | 47.5 | 95.5 | 3.4 |
| | SATé[24] | 3.3 | 32.6 | 15.4 | 48.8 | 94.8 | 4.1 |
| | Prank+GT | 2.8 | 33.7 | 17.1 | 46.1 | 95.9 | 3.5 |
| | RAxML(MAFFT) | 0.6 | 30.3 | 12.3 | 41.6 | 95.1 | 4.1 |
| | RAxML(MUSCLE) | 0.6 | 15.6 | 8.0 | 18.9 | 85.9 | 4.7 |
| | RAxML(ClustalW) | 22.2 | 22.0 | 6.8 | 35.8 | 89.5 | 7.4 |
| **500-taxon moderate-to-difficult** | RAxML(TrueAln) | 0 | 16.5 | 8.1 | 26.1 | 92.8 | 5.6 |
| | SATé[24] | 0 | 16.7 | 7.4 | 25.9 | 90.7 | 6.6 |
| | Prank+GT | 0 | 25.1 | 11.6 | 41.6 | 93.6 | 5.6 |
| | RAxML(MAFFT) | 0 | 11.5 | 5.4 | 17.3 | 86.5 | 9.3 |
| | RAxML(MUSCLE) | 0 | 9.6 | 5.3 | 14.6 | 81.1 | 7.3 |
| | RAxML(ClustalW) | 3.3 | 14.1 | 2.3 | 29.2 | 81.1 | 12.0 |
| **1000-taxon easy** | RAxML(TrueAln) | 0 | 1.9 | 0.7 | 26.1 | 57.7 | 17.3 |
| | SATé[24] | 0 | 1.8 | 0.8 | 25.9 | 57.1 | 17.6 |
| | Prank+GT | 0 | 1.8 | 0.6 | 41.6 | 57.6 | 17.5 |
| | RAxML(MAFFT) | 0 | 1.7 | 0.7 | 17.3 | 54.4 | 17.1 |
| | RAxML(MUSCLE) | 0 | 1.5 | 0.7 | 14.6 | 54.5 | 17.2 |
| | RAxML(ClustalW) | 0 | 1.0 | 0.5 | 29.2 | 42.0 | 13.2 |
| **500-taxon easy** | RAxML(TrueAln) | 0 | 2.0 | 0.9 | 2.9 | 61.4 | 15.3 |
| | SATé[24] | 0 | 2.0 | 0.9 | 2.9 | 60.8 | 15.5 |
| | Prank+GT | 0 | 1.9 | 1.1 | 2.8 | 61.4 | 15.4 |
| | RAxML(MAFFT) | 0 | 1.7 | 0.9 | 2.5 | 56.8 | 15.6 |
| | RAxML(MUSCLE) | 0 | 1.5 | 0.6 | 2.4 | 57.2 | 15.9 |
| | RAxML(ClustalW) | 0 | 1.0 | 0.5 | 1.8 | 42.5 | 12.2 |

Table S32: **Results for the stringent Gblocks experiment.** Columns in the datasets were eliminated (masked) with stringent Gblocks. Some datasets had all sites masked, making tree estimation and error calculation impossible. In particular, only one ClustalW alignment out of 180 had any sites at all after masking on 1000-taxon moderate-to-difficult model conditions. The first column shows the percentage ($n = 180$ for each moderate-to-difficult model condition and $n = 120$ for each easy model condition) of datasets with 100% sites masked. The second, third, and fourth data columns show the total average, minimum model condition average and maximum model condition average increase in missing branch rate for trees estimated on the remaining non-empty masked datasets relative to the missing branch rate for the datasets that included all columns in the alignment. We also report the average and standard deviation of percentage of sites that were masked in the alignments with some remaining sites.

| | | **Percent of datasets with empty masked alignments (%)** | **Increase in missing branch rate (%)** | | | **Original sites masked (%)** | |
|---|---|---|---|---|---|---|---|
| | **Methods** | | **Average** | **Minimum** | **Maximum** | **Average** | **Std dev** |
| **1000-taxon moderate-to-difficult** | RAxML(TrueAln) | 78.9 | 74.0 | 69.6 | 77.3 | 99.3 | 0.4 |
| | SATé[24] | 85.0 | 73.0 | 69.9 | 77.0 | 99.3 | 0.4 |
| | Prank+GT | 88.9 | 70.3 | 67.2 | 74.5 | 99.4 | 0.3 |
| | RAxML(MAFFT) | 91.7 | 62.6 | 54.4 | 70.0 | 99.1 | 0.5 |
| | RAxML(MUSCLE) | 75.0 | 64.3 | 52.7 | 73.7 | 99.1 | 0.4 |
| | RAxML(ClustalW) | 99.4 | 65.2 | 65.2 | 65.2 | 99.2 | 0 |
| **500-taxon moderate-to-difficult** | RAxML(TrueAln) | 2.9 | 68.8 | 62.4 | 73.5 | 99.1 | 0.7 |
| | SATé[24] | 10.7 | 68.4 | 59.5 | 73.7 | 99.0 | 0.7 |
| | Prank+GT | 13.6 | 64.4 | 58.6 | 66.2 | 99.0 | 0.5 |
| | RAxML(MAFFT) | 0.7 | 62.8 | 56.3 | 68.3 | 98.1 | 1.6 |
| | RAxML(MUSCLE) | 3.6 | 61.5 | 50.9 | 68.9 | 98.9 | 0.6 |
| | RAxML(ClustalW) | 7.1 | 56.0 | 53.0 | 60.6 | 98.5 | 0.7 |
| **1000-taxon easy** | RAxML(TrueAln) | 0 | 15.9 | 4.0 | 73.5 | 84.3 | 13.1 |
| | SATé[24] | 0 | 17.5 | 4.0 | 73.7 | 84.6 | 13.2 |
| | Prank+GT | 0 | 17.6 | 4.1 | 66.2 | 85.0 | 12.9 |
| | RAxML(MAFFT) | 0 | 15.4 | 3.6 | 68.3 | 83.5 | 13.4 |
| | RAxML(MUSCLE) | 0 | 17.1 | 4.1 | 67.2 | 84.4 | 13.0 |
| | RAxML(ClustalW) | 0.8 | 18.5 | 4.5 | 53.3 | 83.5 | 12.9 |
| **500-taxon easy** | RAxML(TrueAln) | 0 | 17.3 | 6.8 | 34.4 | 87.7 | 9.0 |
| | SATé[24] | 0 | 18.7 | 6.8 | 34.8 | 88.2 | 8.8 |
| | Prank+GT | 0 | 18.8 | 7.1 | 36.5 | 88.4 | 8.7 |
| | RAxML(MAFFT) | 0 | 16.9 | 6.5 | 32.0 | 86.5 | 9.7 |
| | RAxML(MUSCLE) | 0 | 18.4 | 6.4 | 34.5 | 87.3 | 9.5 |
| | RAxML(ClustalW) | 0 | 20.1 | 7.0 | 41.9 | 87.0 | 8.7 |

Table S33: **Results for the 75% masked alignment experiment.** We report the changes in missing branch rates for trees estimated from "masked" alignments. Masked alignments were produced by eliminating columns from estimated alignments where the proportion of taxa having gaps in the column was greater than 75%. We report the ($n = 180$ for each moderate-to-difficult model condition and $n = 120$ for each easy model condition) total average, minimum model condition average and maximum model condition average change in missing branch rate for trees estimated using the masked datasets relative to the missing branch rate for the datasets that included all columns in the alignment. Negative values indicate an improvement in the missing branch rate, and positive values indicate that the missing branch rate increases. We also report the average and standard deviation of percentage of sites in the alignment that were masked.

| | | **Change in missing branch rate (%)** | | | **Original sites masked (%)** | |
|---|---|---|---|---|---|---|
| | **Methods** | **Average** | **Minimum** | **Maximum** | **Average** | **Std dev** |
| **1000-taxon moderate-to-difficult** | RAxML(TrueAln) | -0.06 | -0.23 | 0.10 | 61.49 | 16.73 |
| | SATé[24] | 1.03 | -0.08 | 2.76 | 52.53 | 16.39 |
| | Prank+GT | 0.03 | -0.21 | 0.20 | 60.02 | 16.54 |
| | RAxML(MAFFT) | -0.06 | -0.36 | 0.11 | 52.21 | 15.83 |
| | RAxML(MUSCLE) | 0.03 | -0.17 | 0.39 | 35.79 | 9.18 |
| | RAxML(ClustalW) | -0.046 | -0.21 | 0.12 | 7.20 | 2.63 |
| **500-taxon moderate-to-difficult** | RAxML(TrueAln) | -0.005 | -0.22 | 0.26 | 61.51 | 16.14 |
| | SATé[24] | 0.09 | -0.08 | 0.30 | 50.14 | 17.05 |
| | Prank+GT | 0.05 | -0.51 | 0.51 | 58.79 | 15.72 |
| | RAxML(MAFFT) | -0.11 | -0.29 | 0.20 | 45.31 | 18.76 |
| | RAxML(MUSCLE) | -0.02 | -0.28 | 0.44 | 38.20 | 9.57 |
| | RAxML(ClustalW) | -0.01 | -0.12 | 0.08 | 8.58 | 4.14 |
| **1000-taxon easy** | RAxML(TrueAln) | 0.04 | -0.05 | 0.15 | 40.81 | 17.20 |
| | SATé[24] | 0.03 | 0.00 | 0.04 | 40.31 | 17.27 |
| | Prank+GT | 0.05 | -0.01 | 0.10 | 40.80 | 17.16 |
| | RAxML(MAFFT) | 0.06 | -0.04 | 0.17 | 37.69 | 16.23 |
| | RAxML(MUSCLE) | 0.00 | -0.09 | 0.04 | 37.53 | 15.88 |
| | RAxML(ClustalW) | 0.07 | -0.05 | 0.21 | 23.45 | 9.52 |
| **500-taxon easy** | RAxML(TrueAln) | 0.02 | -0.16 | 0.08 | 43.18 | 17.07 |
| | SATé[24] | 0.00 | -0.18 | 0.12 | 42.44 | 16.73 |
| | Prank+GT | 0.02 | -0.06 | 0.09 | 43.08 | 16.96 |
| | RAxML(MAFFT) | 0.05 | -0.06 | 0.19 | 39.08 | 16.56 |
| | RAxML(MUSCLE) | 0.01 | -0.11 | 0.13 | 38.99 | 16.41 |
| | RAxML(ClustalW) | -0.01 | -0.09 | 0.11 | 22.79 | 9.34 |

Table S34: **Results for the 50% masked alignment experiment.** We report the changes in missing branch rates for trees estimated from "masked" alignments, as described in Table S33. Masked alignments were produced by eliminating columns from estimated alignments where the proportion of taxa having gaps in the column was greater than 50%. $n = 180$ for each moderate-to-difficult model condition and $n = 120$ for each easy model condition.

| | | Change in missing branch rate (%) | | | Original sites masked (%) | |
|---|---|---|---|---|---|---|
| | **Methods** | **Average** | **Minimum** | **Maximum** | **Average** | **Std dev** |
| **1000-taxon moderate-to-difficult** | RAxML(TrueAln) | -0.08 | -0.25 | 0.22 | 61.62 | 16.74 |
| | SATé[24] | 1.12 | 0.04 | 2.94 | 54.76 | 16.55 |
| | Prank+GT | 0.07 | -0.18 | 0.46 | 60.26 | 16.62 |
| | RAxML(MAFFT) | 0.00 | -0.31 | 0.28 | 53.83 | 15.89 |
| | RAxML(MUSCLE) | 0.08 | -0.32 | 0.73 | 36.55 | 9.10 |
| | RAxML(ClustalW) | 0.068 | -0.19 | 0.51 | 9.80 | 4.07 |
| **500-taxon moderate-to-difficult** | RAxML(TrueAln) | -0.015 | -0.23 | 0.25 | 61.79 | 16.21 |
| | SATé[24] | 0.11 | -0.06 | 0.32 | 52.80 | 17.33 |
| | Prank+GT | 0.16 | -0.36 | 0.56 | 59.20 | 15.84 |
| | RAxML(MAFFT) | -0.06 | -0.39 | 0.13 | 46.85 | 18.95 |
| | RAxML(MUSCLE) | 0.01 | -0.46 | 0.31 | 39.75 | 9.76 |
| | RAxML(ClustalW) | 0.02 | -0.15 | 0.20 | 13.68 | 6.34 |
| **1000-taxon easy** | RAxML(TrueAln) | 0.08 | -0.02 | 0.13 | 40.96 | 17.25 |
| | SATé[24] | 0.06 | 0.04 | 0.10 | 40.48 | 17.30 |
| | Prank+GT | 0.06 | 0.00 | 0.12 | 40.95 | 17.20 |
| | RAxML(MAFFT) | 0.07 | -0.03 | 0.19 | 37.97 | 16.36 |
| | RAxML(MUSCLE) | 0.01 | -0.05 | 0.08 | 37.86 | 16.05 |
| | RAxML(ClustalW) | 0.14 | 0.00 | 0.30 | 27.51 | 11.27 |
| **500-taxon easy** | RAxML(TrueAln) | 0.01 | -0.03 | 0.07 | 43.39 | 17.15 |
| | SATé[24] | 0.05 | -0.10 | 0.20 | 42.69 | 16.85 |
| | Prank+GT | 0.06 | -0.02 | 0.23 | 43.29 | 17.05 |
| | RAxML(MAFFT) | 0.03 | -0.07 | 0.11 | 39.43 | 16.65 |
| | RAxML(MUSCLE) | 0.06 | -0.10 | 0.18 | 39.39 | 16.55 |
| | RAxML(ClustalW) | 0.09 | -0.07 | 0.27 | 27.62 | 11.16 |

# Supplemental References


(S1) M. J. Sanderson, *Bioinformatics* **19**, 301 (2003).

(S2) J. Stoye, D. Evers, F. Meyer, *Bioinformatics* **14**, 157 (1998).

(S3) L. Nakhleh, *et al.*, *Proceedings of the 7th Pacific Symposium on BioComputing (PSB02)* (World Scientific Pub, 2002), pp. 211–222.

(S4) F. Rodriguez, J. Oliver, A. Marin, J. Medina, *Journal of Theoretical Biology* **142**, 485 (1990).

(S5) The Nematode Branch of the Assembling the Tree of Life Project: NemATOL (2008). Website at `nematol.unh.edu/tree/tree1/vIch20ct682_c1w1.aln`.

(S6) D. L. Swofford, PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10 (Sinauer Associates, Sunderland, MA, 2003).

(S7) J. Cannone, *et al.*, *BMC Bioinformatics* **3** (2002). `http://www.rna.ccbb.utexas.edu`.

(S8) J. Thompson, D. Higgins, T. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).

(S9) K. Katoh, K. Kuma, H. Toh, T. Miyata, *Nucleic Acids Res.* **33**, 511 (2005).

(S10) R. Edgar, *BMC Bioinformatics* **5**, 113 (2004).

(S11) A. Loytynoja, N. Goldman, *Proc. Natl Acad. Sci. USA* **102**, 10557 (2005).

(S12) S. Nelesen, K. Liu, D. Zhao, C. R. Linder, T. Warnow, *Pacific Symposium on Biocomputing* (2008), vol. 13, pp. 15–24.

(S13) A. Stamatakis, *Bioinformatics* **22**, 2688 (2006).

(S14) G. Talavera, J. Castresana, *Systematic Biology* **56**, 564 (2007).

(S15) R. Fleissner, D. Metzler, A. von Haeseler, *Syst. Biol.* **54**, 548 (2005).

(S16) B. Redelings, M. Suchard, *Syst. Biol.* **54**, 401 (2005).

(S17) G. Lunter, I. Miklós, A. Drummond, J. L. Jensen, J. Hein, *BMC Bioinformatics* **6**, 83 (2005).

(S18) A. Drummond, A. Rambaut, *BMC Evolutionary Biology* **7**, 214 (2007).

(S19) R. R. Sokal, F. J. Rohlf, *Biometry* (W.H. Freeman, San Francisco, CA, 1995), third edn.

(S20) M. Litzkow, *Usenix Summer Conference* (1987), pp. 381–384.

(S21) U. Roshan, B. Moret, T. Williams, T. Warnow, *Proc. 3rd Computational Systems Biology Conf. (CSB'05)* (Proceedings of the IEEE, 2004), pp. 98–109.

(S22) R. Durbin, S. Eddy, A. Krogh, G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids* (Cambridge University Press, 1998).

(S23) I. Holmes, R. Durbin, *J. Comput. Biol.* **5**, 493 (1998).

(S24) Á. Novák, I. Miklós, R. Lyngsø, J. Hein, *Bioinformatics* **24**, 2403 (2008).

(S25) A. Schwartz, L. Pachter, *Bioinformatics* **23**, e24 (2007).