Syst. Biol. 66(5):769–785, 2017

DOI:10.1093/sysbio/syx051
Advance Access publication June 8, 2017

More on the Best Evolutionary Rate for Phylogenetic Analysis

Seraina Klopfstein[1,2,*], Tim Massingham[3], and Nick Goldman[3]

[1]*Naturhistorisches Museum der Burgergemeinde Bern, Bernastr. 15, CH-3005 Bern, Switzerland;* [2]*University of Bern, Institute of Ecology and Evolution, Baltzerstr. 6, CH-3012 Bern, Switzerland; and* [3]*European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton CB10 1SD, UK*

Correspondence to be sent to: Naturhistorisches Museum der Burgergemeinde Bern, Bernastr. 15, CH-3005 Bern, Switzerland; E-mail: klopfstein@nmbe.ch.

Received 19 August 2016; reviews returned 19 May 2017; accepted 24 May 2017
Associate Editor: Jeffrey Townsend

Abstract.—The accumulation of genome-scale molecular data sets for nonmodel taxa brings us ever closer to resolving the tree of life of all living organisms. However, despite the depth of data available, a number of studies that each used thousands of genes have reported conflicting results. The focus of phylogenomic projects must thus shift to more careful experimental design. Even though we still have a limited understanding of what are the best predictors of the phylogenetic informativeness of a gene, there is wide agreement that one key factor is its evolutionary rate; but there is no consensus as to whether the rates derived as optimal in various analytical, empirical, and simulation approaches have any general applicability. We here use simulations to infer optimal rates in a set of realistic phylogenetic scenarios with varying tree sizes, numbers of terminals, and tree shapes. Furthermore, we study the relationship between the optimal rate and rate variation among sites and among lineages. Finally, we examine how well the predictions made by a range of experimental design methods correlate with the observed performance in our simulations.

We find that the optimal level of divergence is surprisingly robust to differences in taxon sampling and even to among-site and among-lineage rate variation as often encountered in empirical data sets. This finding encourages the use of methods that rely on a single optimal rate to predict a gene's utility. Focusing on correct recovery either of the most basal node in the phylogeny or of the entire topology, the optimal rate is about 0.45 substitutions from root to tip in average Yule trees and about 0.2 in difficult trees with short basal and long-apical branches, but all rates leading to divergence levels between about 0.1 and 0.5 perform reasonably well.

Testing the performance of six methods that can be used to predict a gene's utility against our simulation results, we find that the probability of resolution, signal-noise analysis, and Fisher information are good predictors of phylogenetic informativeness, but they require specification of at least part of a model tree. Likelihood quartet mapping also shows very good performance but only requires sequence alignments and is thus applicable without making assumptions about the phylogeny. Despite them being the most commonly used methods for experimental design, geometric quartet mapping and the integration of phylogenetic informativeness curves perform rather poorly in our comparison. Instead of derived predictors of phylogenetic informativeness, we suggest that the number of sites in a gene that evolve at near-optimal rates (as inferred here) could be used directly to prioritize genes for phylogenetic inference. In combination with measures of model fit, especially with respect to compositional biases and among-site and among-lineage rate variation, such an approach has the potential to greatly improve marker choice and should be tested on empirical data. [Experimental design; phylogenomics.]

Experimental design in phylogenetics had a rather slow start, mostly because marker choice was for a long time limited by practical considerations such as sequencing costs and the availability of primers for polymerase chain reaction and Sanger sequencing (Simon et al. 1994; Hillis et al. 1996; Goldman 1998; Prum et al. 2015). In the beginning of the phylogenomic era, choosing among loci was deemed unnecessary given the fast decrease in sequencing costs and corresponding increase in the amount of sequence data that could be produced within a project's budget. However, the original belief that the tree of life could be resolved by sheer volume of data was brought into doubt by contradicting phylogenies resulting from different data sources or analysis approaches, even when thousands of genes were included (Rokas et al. 2003; Phillips et al. 2004; Philippe et al. 2009, 2011)—"Big Data" alone is not sufficient to make up for inadequate experimental design or unrealistic analysis. Most empirical studies continue to rely on a handful of molecular markers, and even though laboratory costs for producing large numbers of loci are no longer prohibitive, there are still significant computational restrictions if we are to use realistic models of evolution and profit from the most recent developments in analysis methodology (Song et al. 2012; Baele and Lemey 2013; Dell'Ampio et al. 2014). Choosing the right markers for phylogenetic analysis is thus just as important as ever; furthermore, genome-scale data and thus thousands of candidate loci are currently established even for nonmodel taxa through transcriptome and genome projects (Genome 10K Community of Scientists 2009; Misof et al. 2014; Haberer et al. 2016), which offers the necessary basis for doing so efficiently (Regier et al. 2008, 2013; Betancur-R. et al. 2014; Doyle et al. 2015).

The best criteria for choosing loci to address a particular phylogenetic question are still under debate, but the most commonly cited attribute is the evolutionary rate of a gene, or, in other words, the level of divergence it shows for a specific pair or set of taxa (Yang 1998; Betancur-R. et al. 2014; Doyle et al. 2015). If a gene (or more to the point, most of its sites) evolves too slowly with respect to a particular split in the tree, it will show too few differences to contain enough information to correctly infer the respective relationships. If it evolves too fast, the substitutions will become saturated: different mutations at the same site cannot be distinguished and the gene loses its discriminatory power. Exactly how slow or fast is optimal for resolving a specific phylogenetic problem remains unclear, and the answer might depend strongly on how that problem is framed. Analytical attempts to identify optimal rates are scarce. Goldman (1998) developed an experimental design framework based on the Fisher information matrix and, among other questions, used it to infer the optimal rate for estimating the length of a particular branch in a model phylogeny, as well as splitting times in clock trees; Felsenstein (2004) used standard large-sample theory to identify the length of a single branch which can be estimated with the lowest coefficient of variation. Both these approaches focused on the rate that achieves maximum precision for branch length estimates, but the relationship between the inference of branch lengths and of the tree topology remains unclear. In terms of topology inference, different nodes in a tree might be best resolved by genes with different rates due to their varying ages, the lengths of the surrounding branches, and probably also the shape of the subtree they are subtending. Townsend (2007) derived the optimal rate for correctly reconstructing the topology of a symmetric four-taxon tree by examining the probability of observing unreversed synapomorphies when the length of the internal branch in the tree approaches zero. The same scenario, but factoring in the length of the branch in question, was used by Susko and Roger (2012).

It remains unclear how relevant the results from such asymptotic studies are with respect to the correct recovery of empirical tree topologies, which is the main question of interest in experimental design in phylogenomics. Empirical studies that compared loci from phylogenomic data sets by their rate and phylogenetic performance found contradicting results, but this might partially be explained by the fact that they only examined relative rates among their loci and usually did not to report absolute rates in relation to the divergence times they were looking at. Arranging 62 nuclear protein-coding genes by their substitution rates at second codon positions, Regier et al. (2008) found that relatively slower genes showed higher congruence among their respective gene trees than faster genes and that excluding increasing numbers of the faster genes improved node support. Similar results were obtained in other studies which excluded the relatively faster genes or positions (Philippe et al. 2000; Nozaki et al. 2007). Although, in contrast, Salichos and Rokas (2013) found higher congruence among relatively faster loci, this result was later reported as an artefact caused by shorter average sequence lengths in the slower genes in their data set (Betancur-R. et al. 2014). Examining the performance of the three criteria evolutionary rate, clock-likeness, and fit to the evolutionary model in improving phylogenetic signal in two phylogenomic data sets, Doyle et al. (2015) found that filtering by rate was largely outperformed by the other two criteria and that it did not improve concordance among loci nor fit to a reference topology. Given that each study used a different set of genes and examined different taxonomic groups, it is possible that they simply looked at different parts of the rate spectrum, from genes too fast to resolve relatively deeps splits to genes too slow to have accumulated enough information about relatively recent divergences.

Establishing a generally optimal rate for topology inference only makes sense if this rate is sufficiently robust across a reasonable range of phylogenetic settings. One important determinant of the optimal rate is probably the shape of the tree or, more specifically, the relative length of the more basal versus the more apical branches (Rohlf et al. 1990). "Bushy" trees with short basal and long-apical branches are more difficult to resolve because the signal about the sequence of basal splits tends to be erased on the longer apical branches; they probably require lower rates than "stemmy" trees with long branches near the root, but the extent of this effect is unclear. In addition, the number of terminals might have a profound impact on the optimal rate. One might argue that the increase in total tree length that comes with the addition of terminals leads to a lowered probability of unreversed synapomorphies providing unequivocal evidence for more basal splits in the tree, and the optimal rate should thus decrease with increasing numbers of terminals (Klopfstein et al. 2010). On the other hand, adding taxa to a phylogenetic problem can expose multiple substitutions along long branches and might thus facilitate the use of faster evolving sites; larger trees should thus have an increased optimal rate compared with trees of the same depth but with a more sparse taxon sampling (Townsend and Leuenberger 2011).

Yang (1998) used simulations to establish optimal rates across a range of tree shapes and tree sizes, examining several four-taxon trees and a set of Yule trees with higher numbers of taxa. He found a strong dependency of the optimal rate on tree shape and tree size and an asymmetric bell-curve depicting performance over different evolutionary rates. The performance of both parsimony and maximum likelihood (ML) methods was found to increase rather steeply at first as increasing rates led to more sites starting to vary and become informative. It then peaked at a range of rates dependent on the tree examined, and dropped rather slowly at higher rates: the worst effects only arose with highly saturated sequences. Yang focused on unrooted trees and measured the rate of evolution by the total tree length, but did not specifically control for tree height; his results concerning different tree shapes and numbers of terminals are thus not easily comparable among each other and to other phylogenetic settings, as tree length increases both with the evolutionary rate of a gene, with the number of taxa in the tree, and with a tree shape with long-terminal branches. Despite this gap in our understanding of optimal evolutionary rates, several methods have been suggested which aim to predict phylogenetic informativeness (PI). Goldman's (1998) approach using Fisher information can also be applied to predict the information content for genes with different rates, at least when assuming that topology recovery and branch-length estimation are sufficiently related (San Mauro et al. 2012) and when focusing on a specific phylogenetic context. Susko and Roger (Susko 2011; Susko and Roger 2012) developed an asymptotic framework that can be used to obtain approximations of the probabilities of correct resolution of a split of interest ("probability of resolution" hereafter) and applied it to questions of taxon sampling and sequence lengths, but it can also be used to examine resolution probabilities under different evolutionary rates. It requires specification of the split in question and the length of the branch and is thus also rather problem specific. Townsend (2007) used an asymptotic quartet case to derive informativeness profiles of different markers over time based on their site-specific rates. His method only requires input of a vector of site rates; it makes the implicit assumption that the optimal rate derived for the specific quartet case is universal enough to provide a good measure across phylogenetic settings. A later extension of the method (Townsend et al. 2012) estimates both the probability of signal for the true relationships and noise supporting the false relationships and requires similar information to Susko and Roger's method. An entirely different approach is taken by quartet mapping (Strimmer and von Haeseler 1997) that directly assesses the information about randomly selected quartets of taxa present in a sequence alignment, no matter whether the trees supported by these quartets are compatible with one another or not. Quartet mapping comes in two varieties: geometric quartet mapping, which counts the number of site patterns in favor of one of the three possible topologies of an unrooted tree of four taxa (Nieselt-Struwe and von Haeseler 2001), and likelihood mapping, which calculates the likelihood score under each possible topology (Strimmer and von Haeseler 1997).

In a simulation approach similar to the one employed by Yang (1998), we here aim to first establish the best rate for phylogenetic inference under a variety of realistic phylogenetic scenarios, teasing apart the effects of tree shape and number of terminals on one hand and evolutionary rate on the other. We also address two ways in which rates can vary, first among the nucleotide positions of a gene (among-site rate variation, ASRV) and second among the branches of the tree (among-lineage rate variation, ALRV). We find that the best rate is surprisingly robust to different settings, which implies that methods which aim to make general predictions about phylogenetic performance based on the evolutionary rate of a gene have a high potential. Finally, we evaluate six such experimental design methods by contrasting their predictions with the observed performance of different rates in our simulations and discuss the implications for marker choice for phylogenetic inference.

Materials and Methods

We used a combination of bash scripts, R scripts (R Core Team 2014), and the program PhyML (Guindon and Gascuel 2003) to first simulate trees and sequences, then analyze the data under ML, and finally summarize the success of tree inference. We then contrasted our simulation results with predictions obtained by six methods in experimental design. All scripts are available from the Dryad repository (http://dx.doi.org/10.5061/dryad.s342d). Calculations were conducted on the Unix clusters of the EMBL-European Bioinformatics Institute (EMBL-EBI) and on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

Simulation Trees

Instead of using highly specific phylogenetic settings like the quartet trees used in previous studies (Yang 1998; Townsend 2007; Fischer and Steel 2009; Susko and Roger 2012), we here focus on trees obtained under a model of cladogenesis which might more closely resemble trees from typical phylogenetic studies, including a varying number of terminals. To that end, we simulated pure-birth trees (Yule trees) with the birth rate set so that it maximizes the probability of observing the target number of taxa (i.e., the birth rate was set to the logarithm of the number of taxa divided by two), and the sampling fraction set to one. We made use of the sim.bd.taxa.age function in the "TreeSim" package in R (Stadler 2011), which produces rooted and ultrametric trees with a fixed time because the most recent common ancestor (MRCA). Trees were simulated with 4–200 taxa (Table 1). In contrast to Yang (1998), who focused on total tree length, by keeping the time because the MRCA (tree depth) fixed, we are able to directly compare optimal rates across trees of different shapes and with different numbers of terminals. Evolutionary rates are given as the divergence that a gene shows from the root to any of the tips (assuming that the tree is ultrametric); denoting the age of the root of the tree by T, a rate is thus referred to as the number of substitutions with respect to the time unit T. A rate of $1.0/T$ would thus translate into a pairwise distance of 2.0 substitutions between two taxa which go back to the MRCA of the group (Fig. 1a). For most analyses, we focused on trees with 10 or 100 taxa. One thousand trees were simulated for each combination of settings unless stated otherwise.

Yule trees are comparatively easy to resolve because of their relatively even distribution of branch lengths. To produce more difficult trees and investigate the impact of tree shape, we used Ornstein–Uhlenbeck (OU) transformations (Blomberg et al. 2003) with the parameter d of the process, often interpreted as the strength of selection, set to $10^{-0.5}$, $10^{-0.25}$, 1 (which preserves the original branch lengths), $10^{0.25}$ and $10^{0.5}$. These values produce a wide range of trees from highly "bushy" with short internal and long-external branches to very "stemmy" trees with relatively short external

TABLE 1. Summary of simulation settings

Parameter	Details	Values	Sample size
Tree size	Number of taxa	4, 5, 6, 7, 10, 20, 50, 100, 200	1000 trees; data sets per tree: 100 (for 4, 5, 6, 10 taxa) and 10 (for >10 taxa)
Tree shape	Relative length of more basal branches versus branches closer to the tips, simulated by Ornstein–Uhlenbeck transformation with parameter *d*	$d = 10^{-0.5}, 10^{-0.25}, 1, 10^{0.25}, 10^{0.5}$	1000 trees, 100 (10-taxon trees) and 10 data sets (100-taxon trees) per tree
#bp	Number of sites	20, 100, 200, 500, 1000, 10,000	1000 trees, 1 data set per tree
α (ASRV)	Parameter of the gamma distribution of among-site rate variation, discretized in four categories	0.1, 0.3, 0.5, 0.7, 1, 10, 10,000	1000 trees of 10 taxa and 100 trees of 100 taxa, 1 data set per tree
Clock variance (ALRV)	Variance of a lognormal distribution with mean 1.0 from which rate multipliers were drawn for each branch of the tree	0.0, 0.001, 0.01, 0.05, 0.1, 1.0, 10	1000 trees of 10 taxa and 100 trees of 100 taxa, 1 data set per tree

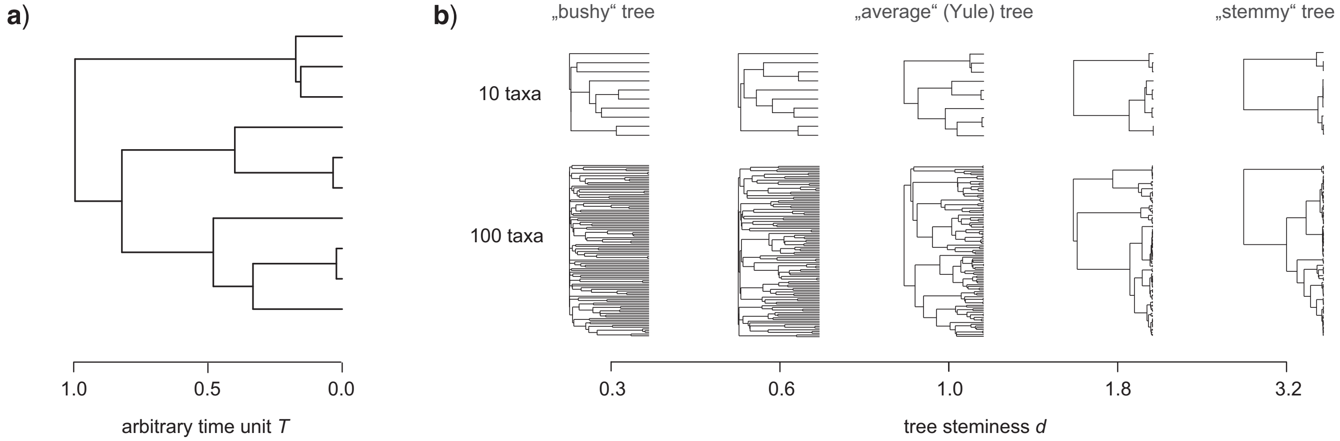


FIGURE 1. Simulation trees. (a) An example tree with 10 taxa as derived under the Yule process with fixed tree depth (here, 1.0) and target number of terminals. We report evolutionary rates with respect to the arbitrary time unit given as the distance between root and tips in rooted, ultrametric trees, with a rate of $0.5/T$ representing 0.5 expected substitutions per site between root and tip, or 1.0 expected substitutions per site between two taxa whose common ancestor is at the root. (b) Tree shapes in trees with (top) 10 or (bottom) 100 taxa as varied via the Ornstein–Uhlenbeck transformation of branch length with parameter *d* (steminess).

branches (Fig. 1b). The γ statistic of Pybus and Harvey (2000) was calculated as a measure of tree shape using the "ape" package in R (Paradis et al. 2004), and we found that our transformations resulted in trees with average γ values of −2.7, −1.8, −0.1, 1.8, and 3.1 for the different OU-transformations for 10-taxon trees, and −9.8, −6.8, −0.1, 11.8, and 16.1 for trees of 100 taxa.

Most empirical data sets deviate from the assumption of a strict molecular clock, showing varying levels of rate variation among the branches of the tree. Such ALRV can lead to systematic biases in phylogenetic inference due to effects such as long-branch attraction (LBA; Felsenstein 1978). In order to simulate ALRV, we randomly drew rate multipliers for each branch from a lognormal distribution with a mean of 1.0 and a range of variances (Table 1). A reasonable range of ALRV was taken from a comparison of more than 700 loci which were examined for their clocklikeness in a recent transcriptome study in ichneumonid hymenopterans (Seraina Klopfstein, unpublished data).

Sequence Simulation

We simulated nucleotide sequences on the simulated trees using the simSeq function in the "phangorn" package in R (Schliep 2011). A range of evolutionary rates was examined from 0.025 to 10 expected substitutions from root to any tip. In the most basic setup, we simulated sequences under the Jukes–Cantor model with a single substitution rate for all sites. Sequence length was chosen in most cases to be 500 bp and also varied from 20 bp to 10,000 bp in an attempt to evaluate its impact (Table 1). To model ASRV, we used a discretized gamma distribution with four rate categories, for values of α of 0.1, mimicking strong ASRV, up to 10,000 which means virtually no ASRV (Yang 1994).

To obtain a more precise estimate of the optimal rate for trees with different numbers of terminals and different shapes, we simulated 100 (for trees with 4, 5, 6, or 10 taxa) or 10 (for trees of 20 taxa and more) sequence alignments for each tree-rate combination (Table 1). As we only observed small differences in the variance of the resulting best-rate estimates, we only produced a single alignment from each tree for the remaining analyses.

Measuring the Success of Phylogenetic Inference

Each simulated sequence alignment was analyzed under ML in PhyML (Guindon and Gascuel 2003), using the correct substitution model, no ASRV and the nearest-neighbor-interchange tree-rearrangement algorithm. For the data sets simulated under ASRV, the α parameter of the gamma distribution was estimated using four categories, as in the simulations.

The success of the ML method was measured in two ways. First, the correct recovery of the most basal split in each tree was examined; this measure is of particular interest in cases where the deepest splits in a phylogeny are of most importance. As PhyML infers unrooted trees from the sequence alignments, we restricted this analysis to those trees in which the first split produces two groups each of minimum size 2, as a split between a single taxon and the rest is always recovered. Especially in trees with small numbers of taxa, this excluded a rather large proportion of simulations. We thus also performed the analysis for the next split after the root in those cases, but found no differences to the simpler rejection-sampling approach besides an increase in variance in the optimal rates (results not shown). As using the second split as well can strongly influence the age of the split in question, we then focused on simulation trees with a basal split that can be analyzed for recovered monophyly.

Second, as an overall measure of phylogenetic accuracy, we examined the proportion of correctly recovered splits. This proportion was calculated from the Penny and Hendy (1985) topological distance (which is based on the well-known Robinson–Foulds distance) as implemented in ape (Paradis et al. 2004).

We used 15 discrete rates (between $0.025/T$ and $10.0/T$) in our simulations, so our estimate for the optimal rate is only an approximation; we thus refer to it as the "best" instead of the "optimal" rate. Based on the above-mentioned criteria of topological accuracy, we choose either the one rate outperforming the others, or the mean of multiple rates in cases where several rates achieved the same success score. Note that in the cases of very easy or very difficult trees, there were often several rates that performed equally well or badly, and we would in these cases expect the result to tend towards the mean of the examined rates (which is $1.5/T$). However, in all cases, the best rate was clearly below this value. Violin plots (see Fig. 2) were created using the "vioplot" package in R (Hintze and Nelson 1998).

Performance of Methods for Experimental Design

We contrasted the performance of ML on the simulated data sets of different evolutionary rates with the predictions by six different methods for experimental design in phylogenetics: Fisher information (Goldman 1998), probability of resolution (Susko and Roger 2012), integration of PI profiles (Townsend 2007), signal-noise analysis (Townsend et al. 2012), and two variants of quartet mapping (Strimmer and von Haeseler 1997; Nieselt-Struwe and von Haeseler 2001). For this analysis, we considered three disparate tree shapes (d values of $10^{-0.5}$, 1, and $10^{0.5}$, i.e., a very bushy, an average, and a very stemmy tree). As some methods are computationally rather costly, especially the calculation of Fisher information, we focused on trees with 10 taxa.

Fisher information (Goldman 1998) was calculated using the software "EDIBLE" (Massingham and Goldman 2000) available on github (https://github.com/timmassingham/EDIBLE). We applied the D-criterion (Geuten et al. 2007), as it best reflects overall informativeness on branch lengths of the phylogeny. The calculation of the Fisher information matrix requires specification of the phylogenetic scenario. The trees complete with relative branch lengths used for simulating the data sets were provided to the method; we thus only addressed cases in which a good guess about the tree topology and relative branch lengths can be made beforehand. We compared this measure of informativeness to both success measures specified before, that is, recovery of the most basal split and proportion of correctly recovered splits per tree.

The probability of resolution (PR) (Susko 2011; Susko and Roger 2012) was calculated using the program "pr4design" provided by the authors. This method is specific for one split of interest in the tree, and we specified as such the most basal split and provided the program with the true length of the corresponding branch under the respective evolutionary rate, the true substitution model, and the sequence length. We could not come up with a straightforward way of using the PR measure in the context of the recovery of all splits in the tree; we thus also used the most basal split in this latter context, assuming that the recovery of most splits crucially depends on getting that first divergence right.

Calculation of PI profiles (Townsend 2007) only requires a vector of evolutionary rates at each site, with respect to a time unit of choice. The true rates used in the simulations were provided and the PI scores were calculated using a custom R script. PI profiles allow measuring information over different time periods. When comparing to the success of resolving the most basal split of the phylogeny, we used the PI score estimated for the root of the tree; when comparing to the overall phylogenetic accuracy, we took the integral under

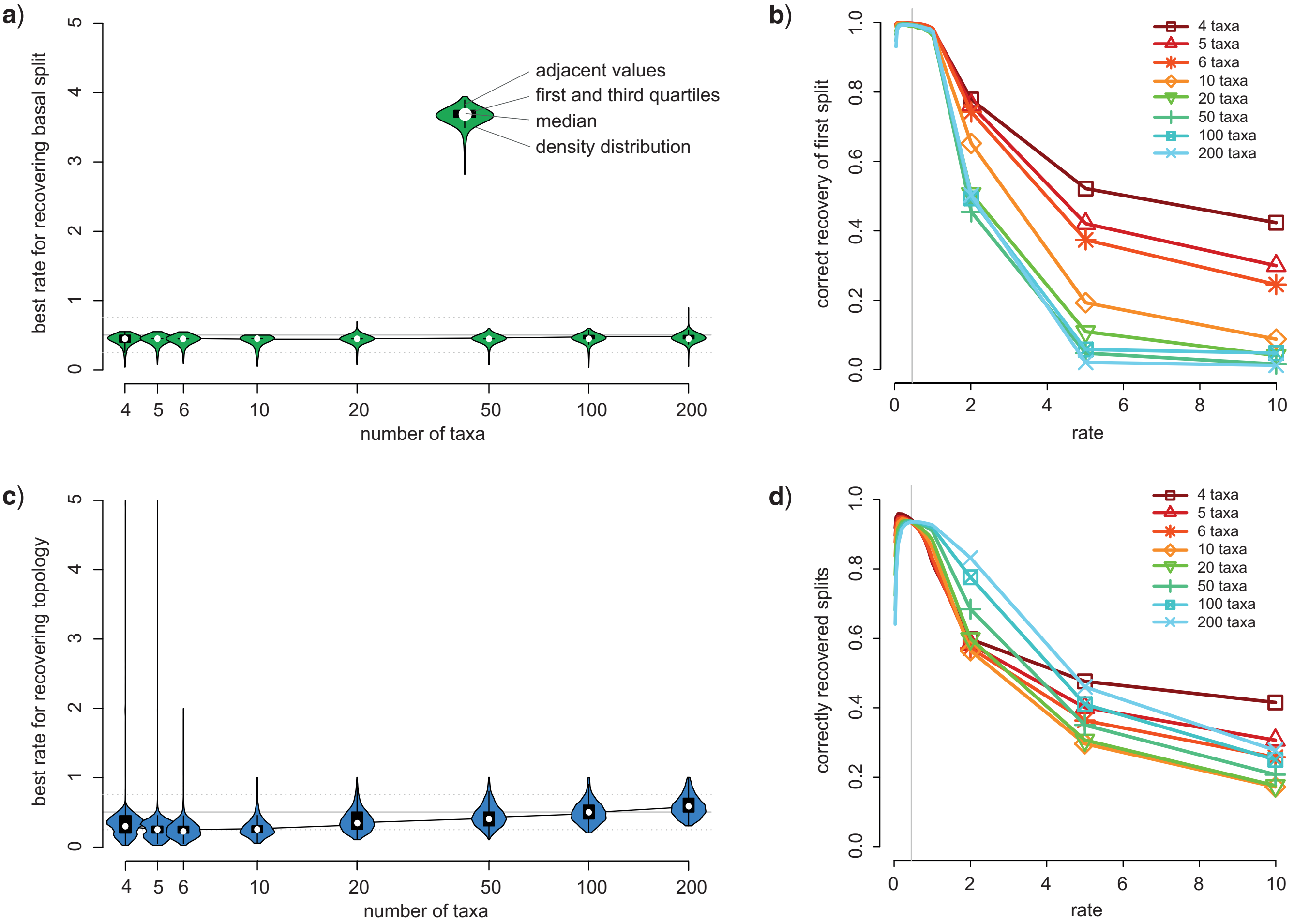


FIGURE 2. Taxon sampling and the best rates for inferring the most basal split in a tree (a, b) or for recovering most of the splits in the topology (c, d). Sequences of length 500 bp were simulated on Yule trees of depth 1.0, under the Jukes–Cantor model and without among-site rate variation. Violin plots (a and c) depict the median, first and third quartiles, whiskers (which as in a box plot denote the highest and lowest values within 1.5 times the distance between first and third quartile), and density distributions of the best rates. The solid and dotted grey horizontal lines are intended as guides for the eye; they are at a rate of 0.5 (solid line) and 0.25 and 0.75 (dotted lines). Performance curves (b and d) are shown for 15 different rates (i.e., 0.025, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 5.0, and 10.0 expected substitutions between root and tip), with symbols omitted at low rates to improve readability.

the PI curve between the root and the mid-depth of the tree; given that PI profiles do not account for the negative effects of homoplasy (Klopfstein et al. 2010), we decided to use this more conservative approach which puts the emphasis on the deeper nodes instead of calculating the integral over the entire time period.

To calculate Townsend et al.'s (2012) signal to noise measure, we obtained the length of the basal branch as for PR. It was provided to the method along with the rates used in the simulations, the state space (four in our case as we simulated nucleotide sequences under Jukes–Cantor—highly uneven base frequencies would mean a reduction in effective state space) and the sequence length. A combination of bash scripts and a Perl script provided by the authors (http://phydesign.townsend.yale.edu/instructions.html) was employed to obtain an estimate for the potential excess support for the true relationship (for definitions of the terms, see Townsend et al. 2012). This approach has also been chosen in previous applications of the method (e.g., Prum et al. 2015).

Quartet mapping requires no *a priori* information about evolutionary rates or tree shapes but instead is based on the sequence alignments. We wrote R scripts to calculate geometric and likelihood quartet mapping from the simulated alignments, defining site patterns for the former according to Nieselt-Struwe and von Haeseler (2001) and making use of the "pml" and "optim.pml" function of phangorn for the latter. We followed Misof et al. (2013) in considering a quartet as being resolved if the star topology was rejected (i.e., if at least one of the three values of the quartet-simplex was below 1/6). There are 210 different four-taxon combinations in a tree of size 10. To estimate how many quartets are needed in order to obtain a good approximation of the tree-likeness measure, we sampled

10 random trees for each tree shape. For each of these 30 trees, we calculated 1,000 quartets from each of the 15 different rate alignments, and obtained the number of quartets necessary to get within a 2.5% range of the final tree-likeness (Supplementary File S1 available on Dryad at http://dx.doi.org/10.5061/dryad.s342d). We ended up sampling 100 quartets for the 10-taxon trees (a preliminary analysis of trees of 100 taxa gave values of 200 quartets necessary to get within the same range).

Results

Best Rates and Taxon Sampling

The best rate for phylogenetic inference is very stable across a wide range of number of terminals, especially if we measure success by the recovery of the most basal split in the tree (Fig. 2a, b). For Yule trees with a sampling fraction of 1.0 and an average shape ($d=1$), genes that show a level of divergence of around $0.45/T$ perform best, with a very small increase in the mean from $0.43/T$ for four taxa to $0.46/T$for 200 taxa. Examining the success at each rate (Fig. 2b) shows that a range of rates performs almost equally well, with every rate between $0.025/T$ and $1.0/T$ expected substitutions per site having a chance higher than 95% to correctly resolve the most basal split, regardless of the number of taxa. A strong drop in performance is however observed between a rate of $1.0/T$ and $2.0/T$. This drop is steeper for larger trees, with fewer than half of the most basal splits being recovered correctly at rates of $2.0/T$ and higher in trees with 50 or more taxa. When considering the proportion of correctly recovered splits in the tree (Fig. 2c, d), we find an increase in the best rate from about $0.25/T$ to $0.58/T$ (for trees of 5–200 taxa). The drop in performance at rates of $2.0/T$ and higher is now a bit steeper for smaller trees (Fig. 2d). In contrast to the recovery of the most basal split, we here also observe a lower performance of very low rates. Recoveries of at least 90% of all splits are observed for rates between $0.1/T$ and $0.9/T$.

Best Rates and Tree Shape

The relative lengths of the early branches in comparison to the later ones in the tree have a large impact on the performance of phylogenetic inference, even though the best rate remains within the range observed for average Yule trees (Fig. 3). Trees with very short basal and long-apical branches ("bushy" trees) are more difficult to resolve and require lower evolutionary rates for correct recovery of the most basal split (with a mean rate of $0.28/T$ and $0.19/T$for trees with 10 and 100 taxa, respectively; Fig. 3a, b, e, f). For the bushiest tree with 100 terminals, successful recovery of the most basal split was extremely low due to very short basal branches, reaching a maximum of 6.7% for the best rate of $0.2/T$. The best rate increases for highly stemmy trees with long basal and short apical branches (to $0.50/T$ and $0.49/T$, respectively). The drop between rate $1.0/T$and $2.0/T$ is present both for trees with 10 and with 100 taxa, and is steeper for the more difficult trees. When considering the proportion of correctly recovered splits, the picture remains the same for trees with 10 taxa and for bushy to average trees of 100 taxa; however, in the case of very stemmy trees with 100 taxa, the very short apical branches require high rates to be resolved (Fig. 3g, h).

Best Rates and the Number of Sites

The correct recovery of difficult phylogenetic problems depends on the sequence length as expected, with a steady increase in the probability of resolving the most basal and the other splits from 20–10, 000 bp at rates below $5.0/T$ (Fig. 4). However, although the range of rates that perform well increases somewhat when more sites are sampled, the best rate remains constant with increasing amounts of information (Supplementary File S2 available on Dryad). Even more strikingly, there is virtually no improvement when sites are added if they evolve at very high rates (from $5.0/T$) in more difficult settings, such as the bushy tree of 10 taxa and in the average Yule tree of 100 taxa (Fig. 4b, c; even when the number of basepairs was increased to 10,000).

Best Rates with Among-Site and Among-Lineage Rate Variation

The variation in the evolutionary rate among sites in an alignment has a large impact on general performance and best rates (Fig. 5). Values of α of 10 or 10, 000 (which reflects virtually no ASRV; also compare Fig. 3) result in the now familiar picture of good performance of rates of about $0.1/T$ to $0.9/T$ and a steep drop between a rate of $1.0/T$ and $2.0/T$; however, a value of α around 1.0 represents a turning point below which higher rates perform almost equally well, even though the best (average) rate only increases from 0.41/T at $\alpha=10, 000$ to $0.62/T$ for $\alpha=0.1$ in the case of the difficult tree of 10 taxa. At the most extreme among-site rate variation, we furthermore observe irregular behavior with an initial drop in performance at very low rate, then an increase again for rates $2.0/T$–$5.0/T$, and a final slow decrease. The shape of these curves to a large extent matches the number of sites evolving at favorable rates ($0.1/T$ to $0.5/T$), as shown in Figure 5g.

Variation in evolutionary rates among the branches of the tree also has only a very minor effect on the best rate (Fig. 6). This effect was most pronounced for the large tree of 100 taxa where extensive ALRV (variance of 10.0) led to a slight decrease in the best rate. Both for average and bushy trees of 10 taxa, the impact of ALRV on the best rate was negligible, even though the overall success at correctly reconstructing the trees decreased with increasing ALRV.

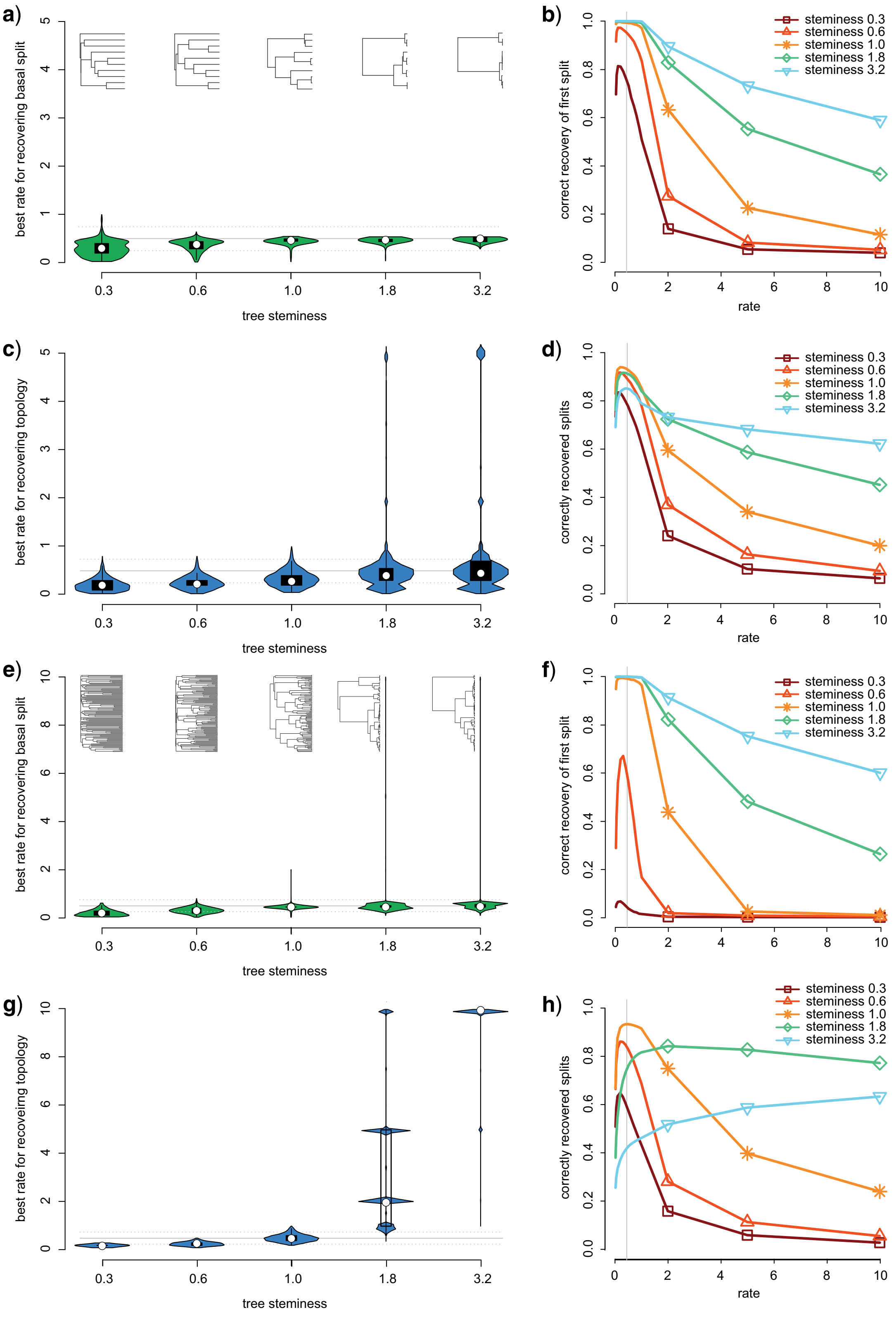


FIGURE 3. Tree shape and the best rates for inferring the most basal split in a tree (a, b, e, and f) or for recovering most of the splits in the topology (c, d, g, and h). Two tree sizes were included, with 10 (a–d) or 100 taxa (e–h). Relative branch lengths were manipulated using Ornstein–Uhlenbeck transformations on Yule trees using values for parameter d of $10^{-0.5}$, $10^{-0.25}$, 1, $10^{0.25}$, and $10^{0.5}$ to create very bushy, average, and very stemmy trees; typical examples of the resulting tree shapes are shown in (a) and (c). Note that only 15 different rates were tested; some of the violin plots thus show discontinuous density distributions (stemmy trees in c and g) where different replicates had high but different optimal rates. Symbols were omitted at low rates in plots b, d, f, and h to improve readability.

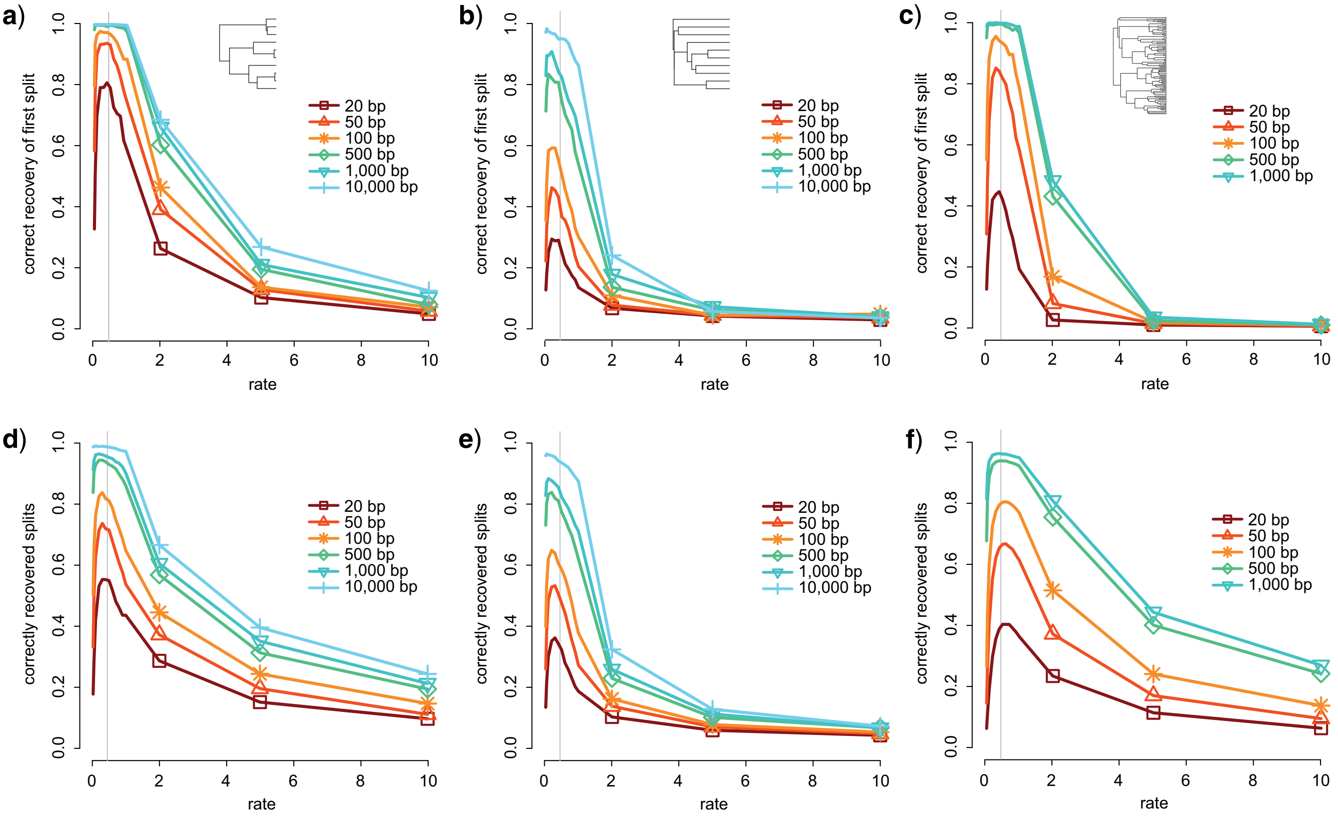


FIGURE 4. Success curves for different evolutionary rates under a range of data set sizes. The top row (a–c) shows the probability of correctly recovering the most basal split in an average tree of 10 taxa (a), a bushy tree of 10 taxa (b), and an average tree with 100 taxa (c); the bottom row (d–f) repeats the same sequence for the proportion of correctly recovered splits. Symbols were omitted at low rates to improve readability.

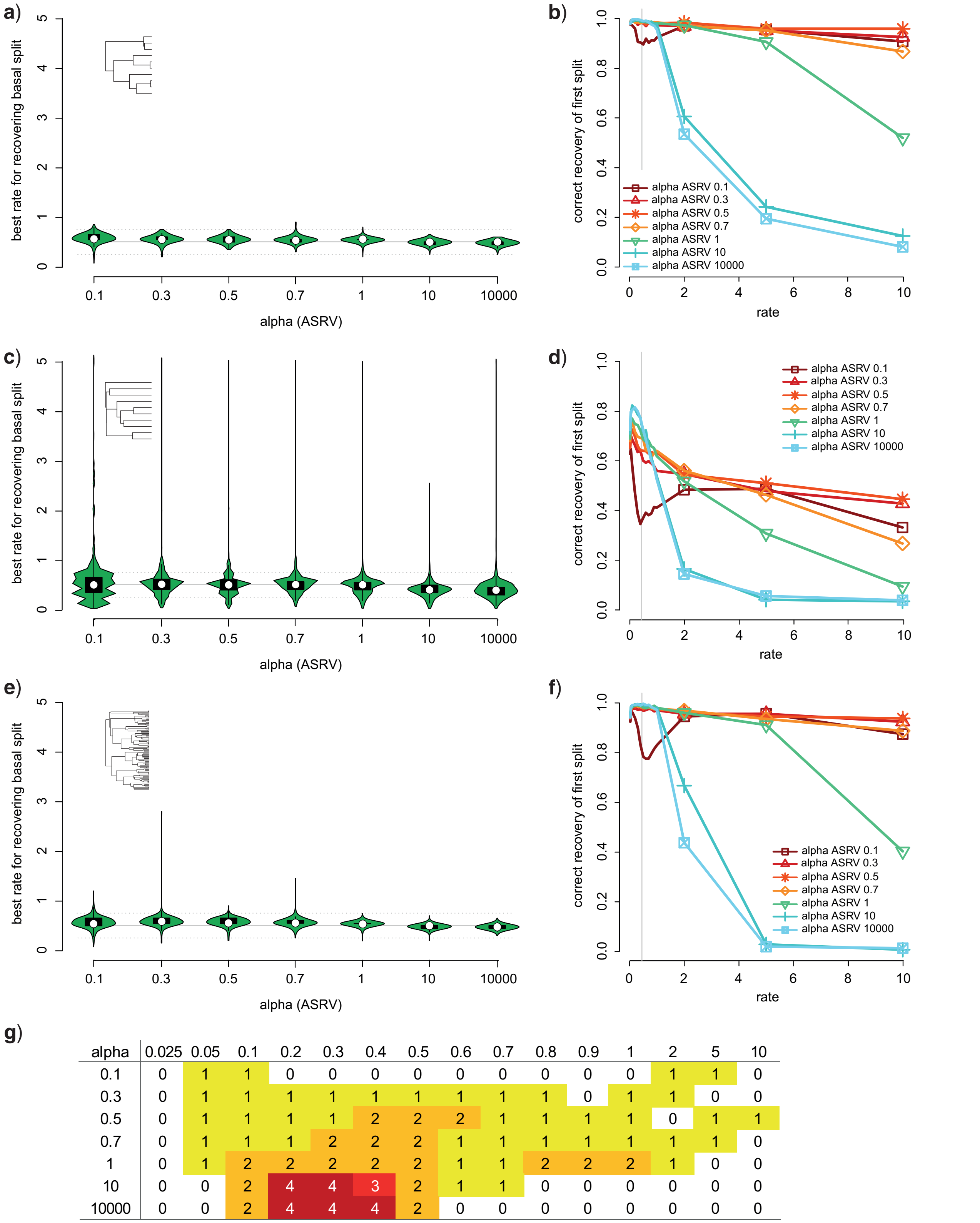
Comparing Experimental Design Methods

We contrasted the observed performance of simulated data sets which evolved at different evolutionary rates with predictions made by six methods for experimental design in phylogenetics (Fig. 7). We consider a method as more successful the more closely its predictions follow the curve of actual performance, measured in terms of the correct recovery of either the most basal split (Fig. 7a–c) or of all splits in the tree (Fig. 7d–f). The maxima and minima of the score obtained by each method was adjusted to the maximum and minimum performance for comparability, and the sum of the differences to the actual performance was used to create a ranking of the methods (Table 2; see also Discussion, below). Considering the recovery of the most basal split (Fig. 7a–c), PR performed best, followed by likelihood quartet mapping, signal-noise analysis, and Fisher information. The first two worked especially well for more difficult trees, whereas the latter had especially good predictive power in the case of the stemmy tree. At the other end of the scale, geometric quartet mapping overrates the contribution of slower versus faster sites, whereas an opposite bias is observed for PI profiles. When comparing the different methods to the overall recovery rate of the correct topology (Fig. 7d–f, Table 2), PR and likelihood quartet mapping outperform all the others, whereas Fisher information now obtains better scores than the signal-noise analysis. Once more, PI profiles and geometric quartet mapping perform least well.

DISCUSSION

The Best Rate for Phylogenetic Analysis

We find that the best rate for phylogenetic inference is surprisingly stable across a range of tree sizes and shapes. The optimal rate for recovering the most basal split in Yule trees with 4–200 taxa is around 0.45 substitutions per site between root and tip (tree age T), with the range of $0.1/T$ to $0.9/T$ showing good performance. Focusing on more difficult tree shapes with short basal and long-apical branches, the optimal rate drops to about $0.2/T$, and the range of good performance is between about $0.05/T$ and $0.4/T$. These values are similar to previous results from analytical or simulation approaches to determine optimal rates for phylogenetic inference. Felsenstein (2004), equation 13.34) found that a branch of length 0.719 substitutions per site can be estimated with the lowest coefficient of



alpha	0.025	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	2	5	10
0.1	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0
0.3	0	1	1	1	1	1	1	1	1	1	0	1	1	0	0
0.5	0	1	1	1	1	2	2	2	1	1	1	1	0	1	1
0.7	0	1	1	1	2	2	2	1	1	1	1	1	1	1	0
1	0	1	2	2	2	2	2	1	1	2	2	2	1	0	0
10	0	0	2	4	4	3	2	1	1	0	0	0	0	0	0
10000	0	0	2	4	4	4	2	0	0	0	0	0	0	0	0

FIGURE 5. ASRV and the best rates for inferring the most basal split in a tree. Average 10-taxon trees (a and b), average 100-taxon trees (c and d), and bushy 10-taxon trees (e and f) were analyzed (g and h); examples of such trees are given as inlaid figures. Note that only 15 different rates were tested, which is why some of the violin plots show discontinuous density distributions (e). The heat map (g) shows how many of the four categories with which the gamma distribution was approximated contained sites that evolve at a near-optimal rate (0.1–0.5 expected substitutions between root and a tip of the tree), for values of α from 0.1–10,000 and for root-tip distances 0.025–10. Symbols were omitted at low rates in panes b, d, and f to improve readability.

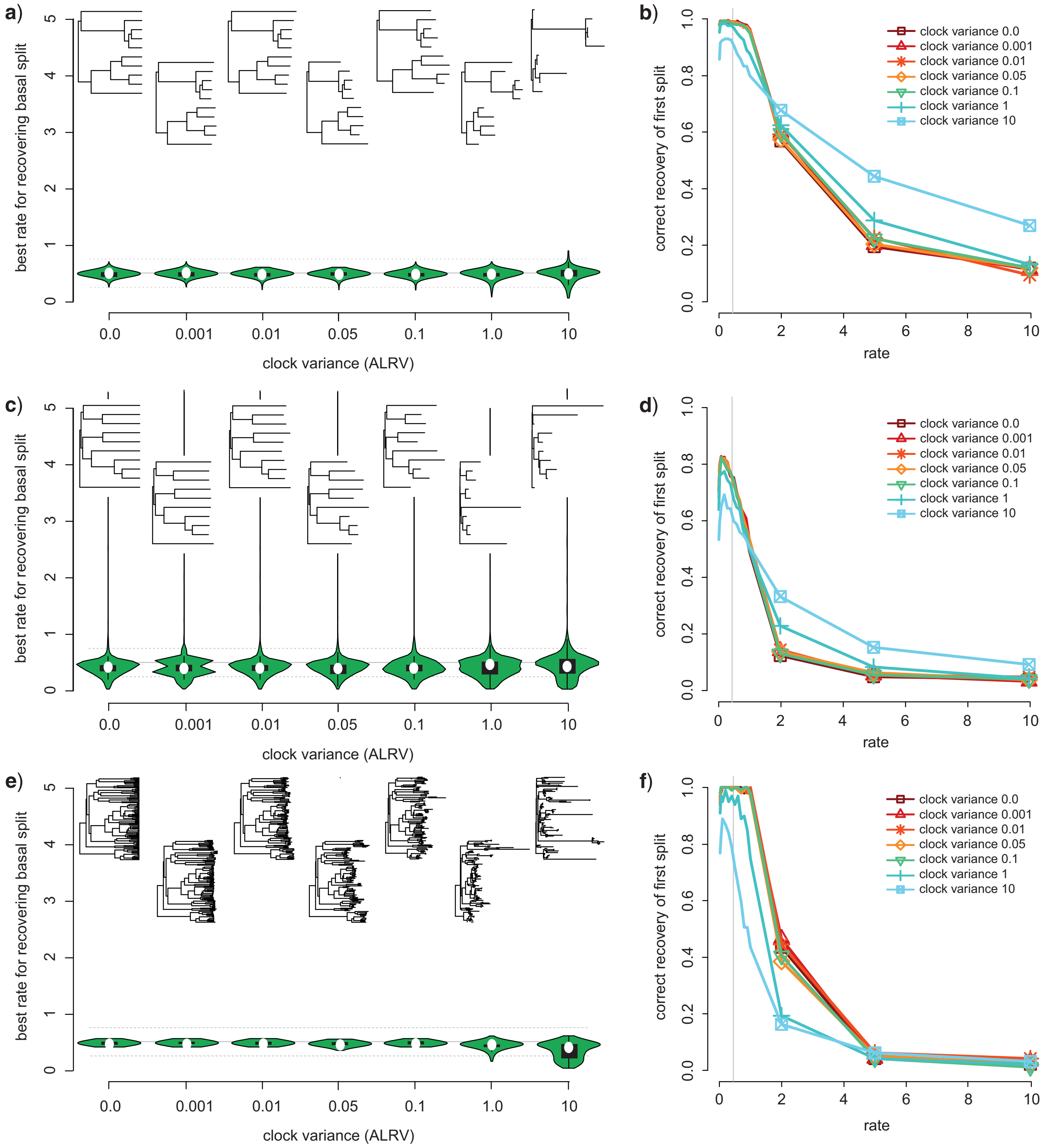


FIGURE 6. ALRV and the best rates for inferring the most basal split in a tree. Average 10-taxon trees (a, b), average 100-taxon trees (c, d), and bushy 10-taxon trees (e, f) were analyzed (g, h); examples of such trees are given as inlaid figures. ALRV is given as the variance of a lognormal distribution from which rate multipliers were drawn for each branch ("clock variance"). Symbols were omitted at low rates in panes (b, d, and f) to improve readability.

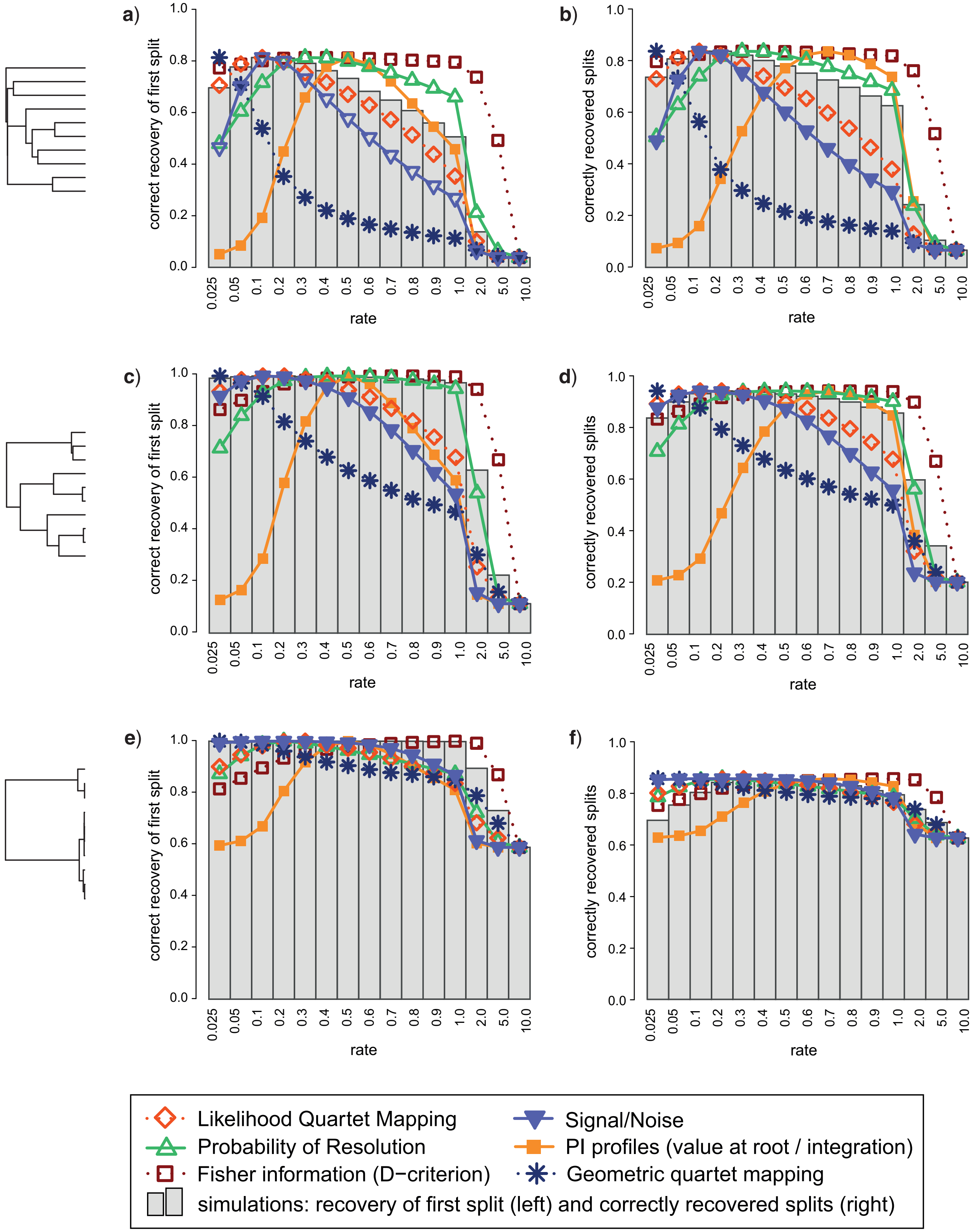


FIGURE 7. Comparison between the performance of data sets of 500 bp evolving at different evolutionary rates in simulations (bar plots) and the predicted informativeness based on six methods for experimental design. The left column (a, c, and e) shows the probability of correctly recovering the most basal split in a bushy (a), average (c), and stemmy tree (e) of 10 taxa; the right column (b, d, and f) repeats the same sequence for the proportion of correctly recovered splits.

TABLE 2. Ranking of experimental design methods

Criterion/method	Bushy tree		Medium tree		Stemy tree		Total
	Σdiffs[a]	Rank	Σ diffs	Rank	Σ diffs	Rank	Rank
Recovery of most basal split							
Likelihood quartet mapping	0.726	1	1.474	3	1.027	4	3
Probability of resolution	1.335	2	0.734	1	0.993	3	1
Signal-noise analysis	1.840	3	2.210	4	0.753	1	3
Fisher information	2.303	4	1.147	2	0.860	2	3
PI profile quantification	2.898	5	4.593	6	2.332	6	5
Geometric quartet mapping	4.922	6	3.890	5	1.094	5	6
Proportion of correctly recovered splits							
Likelihood quartet mapping	1.198	2	1.096	3	0.476	2	2
Probability of resolution	0.904	1	0.603	1	0.421	1	1
Signal-noise analysis	2.339	4	1.676	4	0.509	3	4
Fisher information	1.788	3	0.997	2	0.514	4	3
PI profiles	3.573	5	3.363	6	0.844	6	5
Geometric quartet mapping	5.374	6	3.168	5	0.649	5	6

[a] Σ diffs: sum of (absolute) differences between prediction and observed performance in simulation.

variation under a Jukes–Cantor model; this corresponds to a rate of about $0.36/T$ in the rooted case. Using Fisher information, Goldman (1998) estimated the optimal rate for inferring the branch lengths of a model phylogeny including the great apes and two species of monkeys. The optimal rate was inferred in relation to a particular gene, the ψη-pseudogene; under midpoint rooting and using the branch lengths provided for that gene, his result would correspond roughly to a rate of $0.6/T$. Townsend (2007) showed that a rate of $0.25/T$ maximizes the joint probabilities of a single substitution along the internal branch of a symmetric four-taxon tree and no subsequent substitution if the length of the internal branch approaches zero. Finally, Yang (1998) found an optimal tree length of about 0.5 to 1.0 in a variety of four-taxon trees, which corresponds to a rate of about $0.15/T$ to $0.3/T$ under midpoint rooting, depending on the tree shape. He observed a steady increase in the optimal tree length when he added taxa, but as he worked with unrooted trees, he did not specifically control for tree height. Our results indicate that this increase was probably only due to the added taxa and not because of an increase in the optimal rate. The only context in which higher rates should be preferred are large, very stemmy trees and thus numerous short branches towards the tips, and only if the measure of success is the overall phylogenetic accuracy and not the recovery of the most basal splits.

The observed robustness of optimal rates towards varying numbers of taxa is reassuring for experimental design methods focusing on evolutionary rates, but note that our analysis assumed taxon addition in a random fashion with respect to average Yule trees. Typical approaches to taxon sampling in phylogenetic studies, however, often aim at maximizing taxonomic or phylogenetic diversity (Höhna 2011). Our results on tree shapes obtained for 10 and 100 taxa (Fig. 3) suggest that such a sampling scheme, compared with the random addition of taxa, will lead to a shift of the best rate towards lower values, especially for larger numbers of taxa (compare Fig. 3a, e at a tree steminess below 1.0). As a consequence, tree shape and taxon sampling need to be considered together when comparing optimal rates under different taxon sampling regimes.

As expected, sampling more sites increases the overall performance of ML in inferring the correct tree at lower rates, but leaves the optimal rate virtually unchanged. Interestingly, for inferring the most basal split in more difficult trees (e.g., a bushy 10-taxon tree or an average tree with 100 taxa, Fig. 4b, c), adding sites does not improve the performance of high-rate data sets at all, with no difference in the success for 100 bp versus 10,000 bp at rates of $5.0/T$ or more. Sampling even a very large number of completely saturated sites is thus not going to improve phylogenetic inference, which might help explain the observed failure of certain genomic data sets to resolve difficult phylogenetic questions.

Rate variation among sites has only a minor effect on the optimal rates, but changes the performance curves rather drastically, at least at high-average rates. Performance does not drop as steeply as for data sets without ASRV, which is most probably due to the presence of a sufficient number of sites in the data set which evolve at near-optimal rates and thus contain enough information about the deeper splits (as evidenced in Fig. 6g). In general, genes with a high amount of ASRV should thus be preferred, especially at higher levels of divergence. This finding is in agreement with observations in numerous empirical studies that found a good performance of genes with sites that evolve at a wide range of different rates; typically, these are genes that show relaxed selection at the amino-acid level (e.g., matK in Müller et al. 2006; Hilu et al. 2014) (CAD in Klopfstein et al. 2013). Such genes also have the potential to resolve nodes over a broader range of divergences. However, this will only hold true if the approach used to model ASRV is appropriate, such as in a simulation setting where the true model is known. If an inferior model is used, it is likely that the method will struggle to correctly identify the more slowly-evolving sites and extract their information (Yang 1996).

Rate variation among lineages has long been reported to cause serious challenges to phylogenetic inference methods, mostly through LBA and similar effects (Felsenstein 1978; Swofford et al. 2001). Although we in general observe a decrease in performance with increasing ALRV, the best rate remains virtually unchanged. This might come as a surprise, given that LBA is expected to be more severe at high rates where the differences in branch lengths are more pronounced. However, another effect might counteract LBA: exceedingly short branches as they occur under strong ALRV can only be resolved if sufficient evolution takes place on them, which favors higher rates. We also note that the causes of effects attributed to LBA are still not clear (Parks and Goldman 2014).

From Simulations to Empirical Data

Simulation studies inevitably make many simplifying assumptions, and the extent to which their results can be transferred to empirical studies requires careful consideration. We aimed to cover a wide range of settings inspired by empirical examples, but specific cases might fall outside of that range. Furthermore, our focus was on evolutionary rates as a determinant of performance, whereas other properties of a gene might be equally or even more important. In order for readers to see the limitations of what we have done, and to indicate avenues for future improvements, we here highlight some aspects in which our simulation study may not match empirical analyses.

The tree shapes we used in the simulations reflect a wide range of phylogenetic settings (Mooers and Heard 1997; Nee 2006). Besides mimicking a scenario where selection leads to faster evolution in the early phase and a slowdown once a trait is close to an adaptive peak (Lande 1976), bushy trees can also result from adaptive radiations with an initially increased diversification rate until most available niches are filled (Purvis et al. 2011), or from the sampling scheme typical to most phylogenetic studies which aim to include representatives from higher taxonomic levels (Höhna 2011). Stemmy trees can for instance indicate periods of increased extinction or result from sampling schemes that mix intra- with interspecific sampling, as in alpha-taxonomic studies or in metagenomics (Pons et al. 2006). It is self-evident that for a specific phylogenetic setting, more accurate optimal rates could be obtained using simulations similar to ours; but given the wide range of scenarios covered here and the high constancy of the optimal rate across that range, such an approach is unlikely to yield very different results.

Our simulation settings for ASRV were more limited, as we simply adopted a discretized gamma distribution with four rate categories. This approach is also the most commonly used one when modelling ASRV for phylogenetic inference (Yang 1996), but it is frequently combined with data partitioning (Huelsenbeck et al. 1996; Brown and Lemmon 2007). As an example, protein-coding genes are often partitioned into codon positions, as these evolve under strongly differing selective regimes; their ASRV profiles often show multiple peaks, a situation which cannot be approximated by a gamma distribution. Our simulations nevertheless provide useful insights: we demonstrate that performance correlates less with the average rate of a data set with strong ASRV, but rather with the number of sites that evolve at near-optimal rates. This result is likely applicable to all different kinds of ASRV patterns.

When measuring the performance of different data sets, we focused on two measures, first the recovery of the most basal split and second the number of nodes in the tree that were correctly resolved. These are aspects that many phylogenetic studies are interested in, but in cases where most nodes are comparatively easy to resolve (e.g., due to long subtending branches) one might want to focus on a specific node when designing the experiment. In such cases, we recommend transforming evolutionary rates with respect to the depth of the node in question. Other studies might be more interested in the accuracy of branch-length estimations, for instance in the context of molecular dating; this aspect is not covered here, and the relationship between informativeness about the topology on one hand and about branch lengths on the other remains unclear (Geuten et al. 2007; San Mauro et al. 2012).

In this study, we inferred the phylogenies under the correct evolutionary models, thus not examining model misspecification, which is potentially the most severe limitation of our simulation study. Model misspecification can pertain to any aspect of the model, such as the substitution model (e.g., through nonstationarity and nucleotide composition biases, Jermiin et al. 2004; Jayaswal et al. 2014; Klopfstein et al. 2015), the ASRV model including the partitioning scheme (Yang 1996; Brown and Lemmon 2007), or even the assumption that all included markers share a single evolutionary history (Edwards 2009). The optimal rates estimated here might thus only apply if an appropriate evolutionary model is chosen. More work needs to be done on how different kinds of model-misspecification interact with evolutionary rates, but we predict that in many cases, the negative effects of inadequate modelling will be more severe at higher rates.

Predicting PI

We have provided a comparison between six methods that aim to predict PI on one hand and the observed performance in our simulations on the other. To measure the degree of fit between the two, we used distances to the actual performance after normalization; one could imagine different criteria, for example, whether the ranking of the different rates is close to the true performance. However, a method which shows a nearly linear correspondence between predicted and actual performance will arguably be the most powerful one. The tested methods differ strongly in their assumptions, implementation, requirements in terms of input, and performance. The two methods that outperformed the others under most phylogenetic scenarios and especially in the case of difficult trees are Susko and Roger's PR (2012) and likelihood quartet mapping (Strimmer and von Haeseler 1997). Neither has been used in phylogenomics before. The two methods rely on different input: the former on the length of the branch in question, its position in the tree and the evolutionary rates, and the latter on the sequence alignment. Likelihood quartet mapping is thus especially suited in cases where insufficient *a priori* knowledge about the tree is available, but sequences have already been collected. Preliminary analyses showed that 100 quartets for small (10 taxa) and about 200 for large (100 taxa) should be sufficient to obtain a good estimate of the tree-likeness of the data set (Supplementary File S1available on Dryad), which makes this method computationally feasible even for large numbers of genes.

The two next-best methods are the signal-and-noise analysis (Townsend et al. 2012) which is very similar in its requirements to PR, and Fisher information (Goldman 1998). The signal-noise analysis has been used recently for experimental design in phylogenomic studies at shallow and deep levels of divergence (Mendoza et al. 2015; Prum et al. 2015) and given our results certainly is a promising approach. Although statistically very sound, Fisher information is computationally prohibitive for trees of more than about 10 taxa and requires specification of a full model tree. It remains to be examined how close this model tree needs to be to the true tree in order for the method to be effective. This is also the case for the estimates of the length of the branch in question for the resolution probability and the signal-noise analysis.

Ironically, the two last-ranked methods in our comparison are the ones that have been used most often in to predict PI, that is, geometric quartet mapping and PI profiles. The former has been implemented in the matrix reduction software MARE (Misof et al. 2013) and has been applied in several projects (Meusemann et al. 2010; Dell'Ampio et al. 2014; Misof et al. 2014). In our analyses, it in all cases attributed the highest informativeness to the slowest sites. One might argue that such a bias is conservative as faster sites will tend to be more strongly influenced by systematic errors like model misspecification, but it is likely that a method with such a strong bias eliminates too much phylogenetic signal; we showed here that using likelihoods instead of site patterns in quartet mapping is a much better alternative. The integration of PI curves offers a very convenient approach due to the ease of computation and availability of a web application (López-Giráldez and Townsend 2011) and has been used in numerous contexts (e.g., Townsend et al. 2008). However, the bias towards underrating slow and overrating fast sites which has been reported previously (Klopfstein et al. 2010) was confirmed here. The two quantitative ways in which PI profiles were interpreted in this study, that is, integration of the curves or comparison of absolute scores at the root, thus cannot be recommended. An alternative interpretation has been suggested for PI profiles which examines the shape of the PI curves in a qualitative way (Townsend and Leuenberger 2011; Prum et al. 2015); we have not examined this approach here.

To estimate the predictive power of the six methods, we examined the fit of the prediction curves obtained from each method after scaling them to the range of our success measures; using the ranking of the different rates instead gave a very similar picture (results not shown). But note that our analyses tested the performance of the six methods under the assumption that these data sets were analyzed with the appropriate model. This best-case scenario ignores the impact of model misfit and thus only provides an assessment of the upper limit of performance that these methods can achieve. The actual performance will certainly be lower and might be misled by systematic errors in a similar way as the phylogenetic inference algorithms themselves, and our predictions on informativeness by any method should be interpreted accordingly.

Implications for Experimental Design in Phylogenetics and Phylogenomics

Our dual approach of establishing optimal rates under a variety of settings and testing methods to predict PI allows us to make recommendations for experimental design in phylogenetics, regardless if it is undertaken before conducting the laboratory experiment, for instance for Sanger sequencing or for approaches using target enrichment, or afterwards in the bioinformatics pipeline. Choosing genes (or positions within genes) based on their evolutionary rates is a commonly used method (Philippe et al. 2000; Nozaki et al. 2007; Regier et al. 2008; Chen et al. 2015), even though there is some discussion about its effectiveness (Salichos and Rokas 2013; Doyle et al. 2015). Our results have important implications for such data-choice approaches in phylogenomics. First, the asymmetric bell-shaped success curve already detected by Yang (1998) necessitates the filtering to be two-sided; both genes that are too slow evolving and those with too high-evolutionary rates show reduced performance. Furthermore, as a range of rates have very similar levels of performance, absolute rates in expected substitutions between root and tip should be calculated, instead of sorting genes by rate, and genes preferred that have many sides within this near-optimal range. The results presented in this study can provide guidelines. Given that ASRV is often high especially in protein-coding genes and a number of sites might be under purifying selection and thus do not vary at all, it seems reasonable to estimate site-specific rates and filtering the genes according to the number of sites evolving at favorable rates, instead of relying on average rate estimates across all sites of a gene.

Using the evolutionary rate as a guide in phylogenomic experimental design might already improve efficiency, but there are noteworthy alternatives, especially likelihood quartet mapping (Strimmer and von Haeseler 1997) and PR (Susko 2011; Susko and Roger 2012). The latter is especially suited in cases where a rather precise idea already exists about the phylogeny, whereas the former only requires a sequence alignment. These methods have to our knowledge not been applied for experimental design in phylogenomics; the future will show how well they perform on real data sets instead of simulated data. In any case, other aspects that might influence a marker's PI should also be included, most of all aspects of model-fit (Doyle et al. 2015). The relative merits of rate-based filtering remain to be established in empirical data and likely differ between data sets.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.s342d.

Funding

This work was supported by the Swiss National Science Foundation [PZ00P3_154791 to S.K.] and by EMBL (to S.K., T.M., and N.G.). T.M.'s contribution to this work was concluded prior to joining Oxford Nanopore Technologies.

Acknowledgments

We would like to thank associate editor Jeffery Townsend and three anonymous reviewers for constructive criticism which helped to greatly improve the manuscript.

References

Baele G., Lemey P. 2013. Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency. Bioinformatics 29:1970–1979.

Betancur-R. R., Naylor G.J.P., Orti G. 2014. Conserved genes, sampling error, and phylogenomic inference. Syst. Biol., 63:257–262.

Blomberg S.P., Garland T.J., Ives A.R. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745.

Brown J.M., Lemmon A.R. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst. Biol. 56:643–655.

Chen M.-Y., Liang D., Zhang P. 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. Syst. Biol. 64:1104–1120.

Dell'Ampio, E., Meusemann K., Szucsich N.U., Peters R.S., Meyer B., Borner J., Petersen M., Aberer A.J., Stamatakis A., Walzl M.G., Minh B.Q., von Haeseler A., Ebersberger I., Pass G., Misof B. 2014. Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. Mol. Biol. Evol. 31:239–249.

Doyle V.P., Young R.E., Naylor G.J.P., Brown J.M. 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64:824–837.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA: Sinauer Associates.

Fischer M., Steel M.A. 2009. Sequence length bounds for resolving a deep phylogenetic divergence. J. Theor. Biol. 256:247–252.

Genome 10K Community of Scientists 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. J. Hered. 100:659–674.

Geuten K., Massingham T., Darius P., Smets E., Goldman N. 2007. Experimental design criteria in phylogenetics: where to add taxa. Syst. Biol. 56:609–622.

Goldman N. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. R. Soc. Lond. B Biol. Sci. 265:1779–1786.

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

Haberer G., Mayer K.F.X., Spannagl M. 2016. The big five of the monocot genomes. Curr. Opin. Plant. Biol. 30:33–40.

Hillis D.M., Mable B.K., Larson A., Davis S.K., Zimmer E.A. 1996. Nucleic acids IV: sequencing and cloning. In: Hillis D.M., Moritz C., Mable B.K., editors. Molecular systematics. Sunderland, MA: Sinauer. p. 321–381.

Hilu K.W., Black C.M., Oza D. 2014. Impact of gene molecular evolution on phylogenetic reconstruction: a case study in the rosids (superorder Rosanae, Angiosperms). PLoS One 9:e99725.

Hintze J.L., Nelson R.D. 1998. Violin plots: a box plot-density trace synergism. Am. Stat. 52:181–184.

Höhna S. 2011. Inferring speciation and extinction rates under different sampling schemes. Mol. Biol. Evol. 28:2577–2589.

Huelsenbeck J.P., Bull J.J., Cunningham C.W. 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11:152–158.

Jayaswal V., Wong T.K.F., Ronbinson J., Poladian L., Jermiin L.S. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol. 63:726–742.

Jermiin L., Ho S.Y.W., Ababneh F., Robinson J., Larkum A.W.D. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53: 638–643.

Klopfstein S., Kropf C., Quicke D.L.J. 2010. An evaluation of phylogenetic informativeness profiles and the molecular phylogeny of Diplazontinae (Hymenoptera, Ichneumonidae). Syst. Biol. 59:226–241.

Klopfstein S., Vilhelmsen L., Heraty J.M., Sharkey M.J., Ronquist F. 2013. The hymenopteran tree of life: evidence from protein-coding genes and objectively aligned ribosomal data. PLoS One 8:e69344.

Klopfstein S., Vilhelmsen L., Ronquist F. 2015. A nonstationary Markov model detects directional evolution in hymenopteran morphology. Syst. Biol. 64:1089–1103.

Lande R. 1976. Natural selection an random genetic drift in phenotypic evolution. Evolution 30:314–334.

López-Giráldez F., Townsend J.P. 2011. PhyDesign: an online application for profiling phylogenetic informativeness. BMC Evol. Biol. 11:152.

Massingham T., Goldman N. 2000. EDIBLE: experimental design and information calculations in phylogenetics. Bioinformatics 16:2000.

Mendoza C.G., Naumann J., Samain M.-S., Goetghebeur P., De Smet Y., Wanke S. 2015. A genome-scale mining strategy for recovering novel rapidly-evolving nuclear single-copy genes for addressing shallow-scale phylogenetics in Hydrangea. BMC Evol. Biol. 15:132.

Meusemann, K., von Reumont B.M., Simon S., Roeding F., Strauss S., Kück P., Ebersberger I., Walzl M., Pass G., Breuers S., Achter V., von Haeseler A., Burmester T., Hadrys H., Waegele J.W., Misof B. 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol. Biol. Evol. 27:2451–2464.

Misof B., Liu S.L., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer A.J., Aspock

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

Misof B., Meyer B., Von Reumont B.M., Kück P., Misof K., Meusemann K. 2013. Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. BMC Bioinformatics, 14:348.

Mooers A.O., Heard S.B. 1997. Inferring evolutionary process from phylogenetic tree shape. Q. Rev. Biol. 72:31–54.

Müller K.F., Borsch T., Hilu K.W. 2006. Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: Contrasting *matK*, *trnT-F*, and *rbcL* in basal angiosperms. Mol. Phylogenet. Evol. 41:99–117.

Nee S. 2006. Birth-death models in macroevolution. Annu. Rev. Ecol. Evol. Syst. 37:1–17.

Nieselt-Struwe K., von Haeseler A. 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. Mol. Biol. Evol. 18:1204–1219.

Nozaki H., Iseki M., Hasegawa M., Misawa K., Nakada T., Sasaki N., Watanabe M. 2007. Phylogeny of primary photosynthetic eukaryotes as deduced from slowly evolving nuclear genes. Mol. Biol. Evol. 24:1592–1595.

Paradis E., Claude J., Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Parks S.L., Goldman N. 2014. Maximum likelihood inference of small trees in the presence of long branches. Syst. Biol. 63:798–811.

Penny D., Hendy M.D. 1985. The use of tree comparision metrics. Syst. Zool. 34:75–82.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 9:e1000602.

Philippe H., Derelle R., Lopez P., Pick K., Borchiellini C., Boury-Esnault N., Vacelet J., Renard E., Houliston E., Quéinnec E., Da Silva C., Wincker P., Le Guyader H., Leys S., Jackson D.J., Schreiber F., Erpenbeck D., Morgenstern B., Wörheide G., Manuel M. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19:706-712

Philippe H., Lopez P., Brinkmann H., Budin K., Germot A., Laurent J., Moreira D., Müller M., Le Guyader H. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc. R. Soc. Lond. B Biol. Sci. 267:1213–1221.

Phillips M.J., Delsuc F., Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.

Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S., Sumlin W.D., Vogler A.P. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–609.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Moriarty Lemmon E., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573.

Purvis A., Fritz S.A., Rodriguez J.J., Harvey P.H., Grenyer R. 2011. The shape of mammalian phylogeny: patterns, processes and scales. Philos. Trans. R. .Soc. Lond. B Biol. Sci. 366:2462–2477.

Pybus O.G., Harvey P.H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. Lond. B Biol. Sci. 267:2267–2272.

R Core Team. 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Regier J.C., Shultz J.W., Ganley A.R.D., Hussey A., Shi D., Ball B., Zwick A., Stajich J.E., Cummings M.P., Martin J.W., Cunningham C.W. 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst. Biol. 57:920–938.

Rohlf J., Chang W.S., Sokal R.R., Kim J. 1990. Accuracy of estimated phylogenies: effects of tree topology and evolutionary model. Evolution 44:1671–1684.

Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

San Mauro D., Gower D.J., Cotton J.A., Zardoya R., Wilkinson M., Massingham T. 2012. Experimental design in phylogenetics: testing predictions from expected information. Syst. Biol. 61: 661–674.

Schliep K.P. 2011. Phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.

Simon C., Frati F., Beckenbach A.T., Crespi B.J., Liu H., Flook P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Ann. Entomol. Soc. Am. 87:651–701.

Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in the eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. U.S.A. 109:14942–14947.

Stadler T. 2011. Simulating trees on a fixed number of extant species. Syst. Biol. 60:676–684.

Strimmer K., von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. U.S.A. 94:6815–6819.

Susko E. 2011. Large sample approximations of probabilities of correct evolutionary tree estimation and biases of maximum likelihood estimation. Stat. Appl. Genet. Mol. Biol. 10:Article 10.

Susko E., Roger A.J. 2012. The probability of correctly resolving a split as an experimental design criterion in phylogenetics. Syst. Biol. 61:811–821.

Swofford D.L., Waddell P.J., Huelsenbeck J.P., Foster P.G., Lewis P.O., Rogers J.S. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.

Townsend J.P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Townsend J.P., Leuenberger C. 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Syst. Biol. 60: 358–365.

Townsend J.P., López-Giráldez F., Friedman R. 2008. The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree. J. Mol. Evol. 67:437–447.

Townsend J.P., Su Z., Tekle Y.I. 2012. Phylogenetic signal and noise: predicting the power of a dataset to resolve phylogeny. Syst. Biol. 61:835–849.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.

Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.