XX (2021), **0**, 0, *pp.* 1–35
doi:XX/XX/output

Accuracy in near-perfect virus phylogenies

Joel O. Wertheim[1,*], Mike Steel[2], and Michael J. Sanderson,[3]

[1] *Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA*
[2] *Biomathematics Research Center, School of Mathematics and Statistics, University of Canterbury, Christchurch, 8041, New Zealand*
[3] *Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721 USA*

**To whom correspondence should be addressed: jwertheim@health.ucsd.edu; sanderm@email.arizona.edu*

Abstract

Phylogenetic trees from real-world data often include short edges with very few substitutions per site, which can lead to partially resolved trees and poor accuracy. Theory indicates that the number of sites needed to accurately reconstruct a fully resolved tree grows at a rate proportional to the inverse square of the length of the shortest edge. However, when inferred trees are partially resolved due to short edges, "accuracy" should be defined as the rate of discovering false splits (clades on a rooted tree) relative to the actual number found. Thus, accuracy can be high even if short edges are common. Specifically, in a "near-perfect" parameter space in which trees are large, the tree length ξ (the sum of all edge lengths), is small, and rate variation is minimal, the expected false positive rate is less than $\xi/3$; the exact value depends on tree shape and sequence length. This expected false positive rate is far below the false negative rate for small ξ and often well below 5% even when some assumptions are relaxed. We show this result analytically for maximum parsimony and explore its extension to maximum likelihood using theory and simulations. For hypothesis testing, we show that measures of split "support" that rely on bootstrap resampling consistently imply weaker support than that implied by the false positive rates in near-perfect trees. The near-perfect parameter space closely fits several empirical studies of human virus diversification during outbreaks and epidemics, including Ebolavirus, Zika virus, and SARS-CoV-2, reflecting low substitution rates relative to high transmission/sampling rates in these viruses.

Key words: Perfect phylogeny, Homoplasy, Yule-Harding model, virus, SARS-CoV-2, Ebola virus, Zika virus

Introduction

A "perfect phylogeny" is an evolutionary tree constructed from discrete character data in which no character state evolves more than once (Gusfield, 1997; Fernandez-Baca and Lagergren, 2003)—that is, homoplasy (Wake et al., 2011) is absent. Perfect phylogenies rarely exist for real-world datasets, but algorithms can be modified to search efficiently for "near-perfect" trees when a small amount of homoplasy is allowed (Fernandez-Baca and Lagergren, 2003; Awasthi et al., 2012). In this paper, we address how best to measure accuracy in such "near-perfect" trees, what factors guarantee accuracy is high, and whether real datasets with such minimal levels of homoplasy even exist.

The concept of perfect and near-perfect phylogenies played a key role in early attempts to understand the connections among phylogenetic tree reconstruction methods, such as maximum likelihood (ML), maximum parsimony (MP), and maximum compatibility. In a landmark paper, Felsenstein (Felsenstein, 1973) showed that a sufficient condition for ML and MP to infer the same tree was for the expected number of substitutions on edges of the tree to be very small. Then, "[i]f our assumption were true that evolutionary change is improbable during the relevant period of time, most characters should be uniform over the group. A few would show a single change of state during the evolution of the group. But only very rarely would we find more than one change of state, so that few or no characters would show convergence." This last statement may have been the first hint of a probabilistic description of "near-perfect phylogeny". This condition can be stated more formally as $\xi \leqslant 1$, where ξ is the expected number of substitutions per site summed over the entire tree (i.e., the tree length per site). Homoplasy is rare but has a non-zero probability of occurring.

Felsenstein's concluding comment on near-perfect phylogenies was skeptical: "Real data is certainly not like this..." (Felsenstein, 1973). Homoplasy has since been viewed as a commonplace feature of phylogenetic datasets (Wake et al., 2011) and, reasonably enough, most phylogenetic theory has been developed with this sentiment as an implicit assumption. However, extensive surveys of genetic diversity in RNA viruses have revealed that some viral phylogenies, particularly those associated with outbreaks and epidemics, do exhibit small per site total tree lengths consistent with near-perfect phylogenies (Dudas and Bedford, 2019). These datasets often comprise full-length viral genomes from RNA viruses, which are typically 10–30 kb in length and have a substitution rate of around 10^{-3} substitutions/site/year.

The potential of these data to yield fully resolved phylogenies has been of particular interest in epidemiology, because internal nodes in viral trees represent transmission events (Campbell et al., 2018; Grubaugh et al., 2019; Dudas and Bedford, 2019). This motivates placing a premium on minimizing false negatives (i.e., on deciphering all such transmission events). However, understanding the false positive rate remains a key issue in characterizing phylogenetic accuracy overall (Felsenstein and Kishino, 1993).

Here we explore what assumptions comprise "near-perfect" phylogenies and decouple the false-positive and false-negative components of accuracy in such trees. In particular, by focusing on a mathematically tractable case in which tree size is large yet tree length is small, we will show that the false positive rate can be very good, even when the false negative rate is not: most of the clades inferred are probably correct, even though the tree may be only partly resolved. We also survey a set of viral phylogenies that have many properties of this near-perfect space and estimate their accuracy. Finally, we briefly consider phylogenetic "support" measures in relation to accuracy in near-perfect data. Whereas accuracy relates to the overall performance of a tree estimator relative to the true tree, support relates to the probability of making a mistake in deciding about some aspect of that tree—typically the presence of a particular split—using a statistically based decision rule such as the bootstrap support value or a posterior probability (Felsenstein, 1985; Felsenstein and Kishino, 1993; Hillis and Bull, 1993; Efron et al., 1996; Susko, 2008, 2009; Alfaro and Holder, 2006; Simmons and Norton, 2014).

This paper is organized as follows. "Materials and Methods" are divided into two parts: first, mathematical theory (with proofs in the Supplement), and second, simulation protocols, data, and data analysis. "Results" begin with a more expository description of the theory, illustrated with simulation results, and then describes results from analyses of robustness and support, and data analyses. Following these is the Discussion.

Materials and Methods I. Theory

Definitions of Accuracy

Given a true unrooted binary tree, T, and an estimated tree, $\hat{T}$, a strict measure of accuracy is just $\text{Prob}(\hat{T} = T)$ (Huelsenbeck and Hillis, 1993; Erdös et al., 1999). In large trees it is useful to measure partial agreement, such as the proportion of nontrivial splits on $\hat{T}$ that are also on T, out of a possible $n - 3$ (Yang, 1998).

A still more nuanced definition of accuracy is useful when either T or $\hat{T}$ is only partially resolved (not binary), that is, when the number of nontrivial splits, $C(T)$, is less than $n - 3$ (Warnow, 2013). Let N_{FP} be the number of splits on $\hat{T}$ but not T (false positives), and let N_{FN} be the number of splits on T but not $\hat{T}$ (false negatives). When both trees are binary, $N_{FP} = N_{FN}$ (Berry and Gascuel, 1996; Smirnov and Warnow, 2021); otherwise they can contribute differentially to error. The Robinson–Foulds (RF) distance (Robinson and Foulds, 1981), $d_{RF} = N_{FP} + N_{FN}$, combines both errors in one measure of overall accuracy. Here we distinguish between these errors explicitly by defining false positive and negative rates (Smirnov and Warnow, 2021):

$$\begin{aligned} FP_T &= \mathbb{E}[N_{FP}/C(\hat{T})], \\ FN_T &= \mathbb{E}[N_{FN}/C(T)]. \end{aligned} \tag{1}$$

Both error rates are expectations over some generating model for the data, described next.

Evolutionary Model

Let $B(n)$ denote the set of unrooted binary phylogenetic trees with leaf set $[n] = \{1, 2, \ldots, n\}$. Note that a tree $T \in B(n)$ has $2n - 3$ edges. Consider a Jukes-Cantor model (JC69; Felsenstein, 2004), with rate parameter λ, in which the probability of a state change between the endpoints of an edge e, denoted p_e, is given by $p_e = p$, where $p = \frac{3}{4}(1 - \exp(-4\lambda/3))$. Assume further that all edges have the same value of λ. Let ξ denote the expected number of state changes per character in T. Thus $\xi = \lambda \cdot (2n - 3)$.

A *character* refers to the assignment of states to the taxa at a given site of an alignment. We will say that a character evolves 'perfectly' on T if there is a single change of state across one interior edge (say e) and no change of state on any other edge of T. Thus, a character that evolves perfectly on T is homoplasy-free, and the two notions are equivalent for binary characters. However, for multi-state characters, the notion of a perfectly evolved characters is stronger than that of being being merely homoplasy-free. We deal here with this stronger notion for two reasons: firstly, it simplifies the mathematical analysis, and second, the expected proportion of homoplasy-free characters that not perfectly evolved under the models we consider tends to zero as the number of taxa becomes large.

We will say that a character f evolves on T with c *edge changes on* $e_1, \ldots, e_c$ if state changes occur on edges $e_1, \ldots, e_c$ and on no other edge of T. More briefly, we say that f evolves on T with c edge changes if f evolves with c edge changes for some set of c distinct edges of T (mostly we will deal with the case $c = 2$).

Recall that a *split* refers to a bipartition of the leaf set $[n]$ into two nonempty subsets (and splits are induced by binary characters). A character that has evolved perfectly on T produces a split, and these splits (across a set of perfectly evolved characters) are compatible and so form a (generally unresolved/non-binary) tree on leaf set $[n]$.

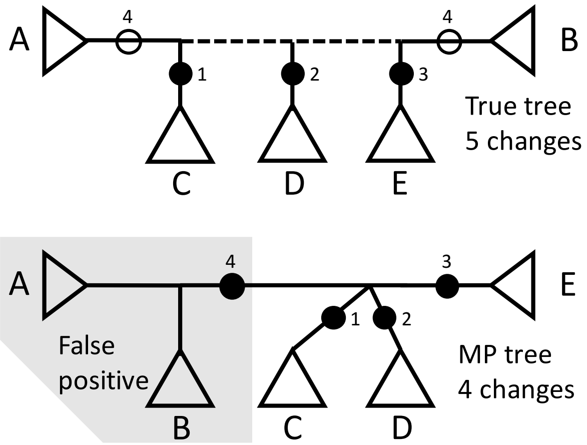

Fig. 1. How a false positive split is inferred by maximum parsimony (MP). On true tree (top) sites 1–3 are binary and "perfect"; that is, they have only a single change (locations marked by black circles), but site 4 is binary and homoplastic, changing twice (open circles). The dotted line is the path between the two homoplastic changes in site 4. As long as perfect sites do not change along the dotted line path, a false positive split is inferred on the MP tree (bottom).

Probability of False Splits

Suppose that m characters evolve on T and that, of these m characters, k of them are perfectly evolved on T (note that more than one of these characters may correspond to the same split of T). Next, consider a single additional character f which has evolved on T with 2 edge changes, on e_1, e_2 (there is no restriction that these must be interior edges). Under certain conditions, the MP tree for these characters will include a false split (false positive)—a split not on T (Fig. 1). In particular, a false split occurs if no perfect character changes state along the path between e_1 and e_2 (see Lemma 1 in the Supplementary Information).

Let $\Phi_T^{(k)}$ be the probability that a character f that has evolved on T with 2 edge changes generates a false split under MP, which means:

(C-i) it is a binary character,

(C-ii) the corresponding split is not a split of T, and

(C-iii) the split described by f is compatible with k characters that are perfectly evolved on T (by the Markovian process described above).

Given a tree $T \in B(n)$, let $d_T(e_1, e_2)$ denote the number of edges of T that lie strictly within the path between e_1 and e_2 (i.e., excluding e_1 and e_2). Thus, e_1 and e_2 are adjacent if and only if $d_T(e_1, e_2) = 0$. In addition, let $\varphi_T = (\varphi_T(0), \varphi_T(1), \ldots, \varphi_T(n-3))$, where $\varphi_T(i)$ is the number of (unordered) pairs of edges $\{e, e'\}$ of T for which $d_T(e, e') = i$. Finally, for i between 1 and $n-3$, let

$$\tilde{\varphi}_T(i) = \frac{\varphi_T(i)}{\binom{2n-3}{2}}. \tag{2}$$

The probability of a false split is then given by the following theorem (see SI for proof).

Theorem 1 For each $T \in B(n)$, and $k \geqslant 1$ we have:

$$\Phi_T^{(k)} = \frac{1}{3} \cdot \sum_{i=1}^{n-3} \tilde{\varphi}_T(i) \left(1 - \frac{i}{(n-3)}\right)^k .$$

Theorem 1 shows that for fixed k and n, the shape of T plays a significant role in determining $\Phi_T^{(k)}$; in particular, unbalanced trees (such as caterpillars) will have a smaller value of $\Phi_T^{(k)}$ than more balanced trees. Indeed, it is possible to calculate the value of $\Phi_T^{(k)}$ exactly for the two extreme cases of caterpillar trees and fully-balanced trees to determine the extent of this dependence (see SI).

Estimating the Expected False Positive Rate

Given a binary phylogenetic tree T, and m characters evolved randomly on T by the model described earlier, the *false positive rate* (FP_T) is the expected value of the ratio of false splits to all splits in the estimated tree (Eqn. 1; here we assume that if the reconstructed tree is a star, this proportion [which is technically 0/0] is zero). Recall that ξ is the expected number of state changes in the tree T per character, under the model described earlier. FP_T is a function of the three parameters T (specifically, its shape and number of leaves), m, and λ (equivalently, FP_T is a function of T, m, and ξ).

In general, it is mathematically complicated to describe FP_T in terms of these parameters. However, when the number of leaves in a tree grows faster than the number of perfectly compatible characters, it is possible to state a limit result to provide an approximation to FP_T for large trees.

In the following theorem, we consider the following setting:

(I) $m\xi = \Theta(n^\beta)$ for some $0 < \beta < \frac{1}{2}$, and

(II) $m\xi^2 = O(1)$,

where $O(1)$ refers to dependence on n (thus $m\xi^2$ is not growing with n). Note that Condition (I) implies that the number of perfectly evolved characters grows with the number of leaves, but at a rate that is slower than linearly. Conditions (I) and (II) imply that ξ decreases as n increases.

In this setting, we show that the false positive rate is (asymptotically) of the form $\frac{\xi}{3}$ times a function Ω that involves T (via its shape), m, and ξ. If we now treat ξ as a variable, then for $\xi = 0$, the function Ω is close to 1 (for large n) and so FP_T initially grows like $\xi/3$. However, as ξ increases, Ω begins to decline at an increasing rate, resulting in the false positive rate reaching a maximum value before starting to decrease.

To describe this result, we need to define this function Ω. Let

$$\Omega(T_n, \xi, m) = \sum_{i=1}^{n-4} \tilde{\varphi}_{T_n}(i) \cdot \frac{e^{-i\mu/(n-3)} - e^{-\mu}}{1 - i/(n-3)},$$

where:

$$\mu = \frac{1}{2}m\xi$$

and where $\tilde{\varphi}_{T_n}(i)$ is given in Eqn. (2). For example, for any caterpillar tree, we have $\tilde{\varphi}_{T_n}(i) = 4(n-2-i)/\binom{2n-3}{2})$.

Notice that $\Omega(T_n, \xi, m)$ depends on T_n only via the coefficients $\tilde{\varphi}_{T_n}(i)$, and this dependence is linear. Thus, if $\mathcal{D}$ is a distribution on trees (e.g. the PDA or YH), then the expected value of $\Omega(T_n, \xi, m)$ is given by:

$$\mathbb{E}_{\mathcal{D}}[\Omega(T_n, \xi, m)] = \sum_{i=1}^{n-4} \mathbb{E}_{\mathcal{D}}[\tilde{\varphi}_{T_n}(i)] \cdot \frac{e^{-i\mu/(n-3)} - e^{-\mu}}{1 - i/(n-3)}. \tag{3}$$

For the PDA distribution, the term $\mathbb{E}_{PDA}[\tilde{\varphi}_{T_n}(i)]$ has an explicit exact value, namely,

$$\mathbb{E}_{PDA}[\tilde{\varphi}_{T_n}(i)] = \frac{(i+3)2^i(2n-i-4)!(n-2)!}{(2n-4)!(n-i-3)!\binom{2n-3}{2}}, \tag{4}$$

for all i between 1 and $n-3$ (see SI for proof).

Theorem 2 For each $n \geqslant 1$, let T_n be a binary phylogenetic tree with n leaves, and suppose that Conditions (I) and (II) hold.

(i)

$$FP_{T_n} = \frac{\xi}{3} \cdot \Omega(T_n, \xi, m) \cdot (1 + o(1)),$$

where $o(1)$ is a term that tends to 0 as n grows.

(ii) If T_n is sampled from a distribution $\mathcal{D}$ (e.g. PDA, YH), then the expected value of FP_{T_n}, denoted $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$, satisfies

$$\mathbb{E}_{\mathcal{D}}[FP_{T_n}] = \frac{\xi}{3} \cdot \mathbb{E}_{\mathcal{D}}\left[\Omega(T_n, \xi, m)\right] \cdot (1 + o(1)).$$

Remarks: Note that FP_{T_n} depends only on the shape of the tree T_n (and not on how its leaves are labelled), thus for a tree distribution $\mathcal{D}$ on either the class of caterpillar trees, or symmetric trees, we have $FP_{T_n} = \mathbb{E}_{\mathcal{D}}[FP_{T_n}]$.

Notice also from Fig. 3 that as ξ increases from 0 the estimate of $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ given by $\frac{\xi}{3} \cdot \Omega(T_n, \xi, m)$ for the YH, PDA distributions and for symmetric trees initially increases (approximately linearly) with ξ but then begins to decrease with increasing ξ. By contrast, when T_n has the caterpillar tree shape, the estimate of FP_{T_n} appears to be constant as ξ increases from 0 (see Fig. 3). Indeed, when T_n is a caterpillar tree, the expression for FP_{T_n} in Theorem 2(i) reduces to the following remarkably simple expression as n becomes large:

$$FP_{T_n} \sim 4/(3m),$$

which is independent of ξ (and n). Details are provided in the SI.

Materials and Methods II. Simulations, Data, and Data Analyses

Main Simulation Pipeline

Simulations were run to assess goodness of fit and robustness of mathematical predictions under various regimes of model parameters and tree inference criteria (MP or ML), as well as to estimate expected accuracy in empirical data sets. Each of R simulation replicates (with r sub-replicate tree searches in each) consisted of the following sequence of steps: (i) generation of a random binary tree T with n leaves according to either a "proportional-to-distinguishable-arrangement" (PDA) or Yule-Harding (YH) model (Aldous, 2001) (as well as the two extreme cases of completely unbalanced caterpillar trees, and completely balanced symmetric trees); (ii) assignment of edge lengths of T according to a gamma distribution with shape parameter α_e and mean $\bar{\lambda}$; (iii) generation of a sequence alignment of m sites using either JC69, HKY or GTR models (using Seq-Gen v. 1.3.4, with base frequencies, rate matrix parameters, invariant site parameter or gamma shape parameter set or estimated from empirical data); (iv) reconstruction of estimated tree $\hat{T}$ [using PAUP 4.0a (build 166) for MP with options 'hsearch add=simple swap=no nreps=r;contree all/strict'; and using IQ-Tree2 (v. 2.0.6) (Minh et al., 2020) for ML with options '-m HKY+FQ -nt 1 -redo -mredo –polytomy -blmin 1e-9', replicated r times, followed by strict consensus]; (v) tallying N_{FP} and N_{FN} from T and $\hat{T}$ and computing error rates. Mean rates across replicates were then tallied. All steps except (iii) and (iv) used custom PERL scripts (available at https://github.com/sanderm53/pperfect). Generally, R was set to 1000 and r to 100.

Support Simulations

Phylogenetic support measures were estimated in trees simulated via the main pipeline described above with $n = 513$, $m = 1000$, a JC69 model with no rate variation, and PDA random trees. Ten values of λ in the interval $[10^{-5}, 0.31622]$ were analyzed. PAUP was used for MP bootstrapping (same heuristic search as above but with 100 replicates × 10 subreplicates); IQTree2 was used (50 random tree replicates) for SH-aLRT ('-alrt 1000'), aBayes, and ultrafast bootstrapping ('-B 1000'), with additional options enforcing minimum branch lengths of 10^{-9} and collapsed polytomies. Mean support across replicates was computed.

Perfect four-taxon alignments were generated in which each of the five branches had a single, non-homoplastic nucleotide substitution in the alignment and all other sites were constant. Alignment lengths ranged between 40 nt and 30,000 nt. ML trees were inferred in IQTree2 with a JC69 model, minimum branch lengths of 10^{-9}, and collapsed polytomies. Clade support was determined using ultrafast bootstrapping (10,000 replicates), SH-aLRT (10,000 replicates), and aBayes. Full Bayesian inference was also performed in MrBayes v3.2.7 (Ronquist and Huelsenbeck, 2003) with a single run per replicate of 2.5 million generations, with the first 10% of generations discarded as burnin.

Alignments for larger perfect symmetrical and asymmetrical (caterpillar) trees were generated with 8, 16, 32, 64, and 128 taxa. Each branch, including terminal branches, had a single nonhomoplastic nucleotide substitution in the alignment with all other sites constant. Alignment lengths ranged from 236 to 32,768 nt. ML trees were inferred as described above for the four-taxon alignments, and support was assessed by ultrafast bootstrapping, SH-aLRT, and aBayes.

Virus Datasets

Viral phylogenies were obtained from the NextStrain (Hadfield et al., 2018) website (accessed 05 May 2020) (Table 1). Phylograms were downloaded for dengue virus, sengue virus serotype 1, Ebolavirus (Dudas et al., 2017), Enterovirus 68 (Dyrdak et al., 2019), measles morbillivirus, mumps virus, respiratory syncytial virus, West Nile virus (Hadfield et al., 2019), and Zika virus. In addition, we also analyzed an iatrogenic HIV-1 outbreak in

Cambodia (Rouet et al., 2018) and the first wave of the SARS-CoV-2 epidemic in China (Pekar et al., 2021). The SARS-CoV-2 phylogeny is the ML tree used in Pekar (Pekar et al., 2021) (see Data S1 for list of GISAID Accession IDs). Publicly available genomic sequences (or genetic sequences for HIV-1) were downloaded from GenBank and aligned with mafft v7.407 (Katoh and Standley, 2013) (accession numbers can be found in Data S2).

False positive rates for the virus phylogenies were estimated with our simulation pipeline, setting parameters to values estimated from published trees and publicly available sequences used to construct them (Table 1, S1). IQTree2 was used to infer the six rate parameters of a GTR substitution model with empirical base frequencies, and either ASR variation following a gamma distribution with shape parameter $\hat{\alpha}_{\text{ASR}}$, or an invariant sites model with parameter p_{invar} ('GTR+F+G4' or 'GTR+F+I'). Model fit was assessed using the Bayesian Information Criterion (BIC) in IQTree2 (Kalyaanamoorthy et al., 2017).

Edge length (per site) variation was assumed to follow a gamma distribution: $\lambda \sim \Gamma(\alpha_e, \alpha_e/\overline{\lambda})$ having mean $\overline{\lambda}$ and variance $\overline{\lambda}^2/\alpha_e$. The distribution of substitutions is a mixture of Poisson and gamma distributions, which is a negative binomial with a variance to mean ratio of

$$1 + \frac{m\overline{\lambda}}{\alpha_e} \tag{5}$$

which was shown by Bedford and Hartl for an equivalent parameterization (Bedford and Hartl, 2008). Virus trees were preprocessed, setting any edge lengths $< 1.1 \times 10^{-6}$ to zero, assuming these reflected ML numeric artifacts. Then, $\overline{\lambda}$ was estimated from the observed sum of per site edge lengths divided by $2n - 3$, and Eqn. 5 was then used to estimate α_e.

Ideally, we would fit the data to the random tree model, but standard methods either assume binary trees or model polytomies with an a priori assumption about the tree model itself (e.g., Bortolussi et al., 2006). Therefore, we repeated simulations using both PDA and YH models.

Results

Overview of Results on Accuracy

Simulations of tree inference with MP, over a large range of tree lengths, ξ, and other parameters illustrate several known results (Fig. 2) and perhaps a few less well known ones. First, resolution of the inferred tree increases with tree length. Second, "overall" accuracy, as measured by the RF distance, is optimal at an intermediate tree length, ξ^* (Yang, 1998; Bininda-Emonds et al., 2001; Steel and Leuenberger, 2017). Moreover, when $\xi >> \xi^*$, the false positive error rate, FP_T, is similar to the false negative rate, FN_T, as might be expected because the true and estimated trees are nearly binary; therefore $N_{FP} \cong N_{FN}$.

However, when $\xi << \xi^*$, $FP_T << FN_T$, and the false positive error rate can remain quite good (< 0.05) over a large range of ξ even when the false negative error rate is very high. However, the range of tree lengths for which this result holds depends critically on rate variation across edges and sites. When $\xi \leqslant 1$, the false positive rate is low and insensitive to the presence of rate variation; but, when $\xi > 1$, the false positive rate is much more sensitive to rate variation—high when variation is present and low when absent (contrast Fig. 2A and Fig. 2B). In real-world data, as ξ increases, we expect that evidence of rate variation will become more apparent. Key elements of these findings can be shown analytically in a "near-perfect" zone described by a simple evolutionary model.

Overview of the Mathematical Theory

First we define "near-perfect" more formally. Assume the data consist of an alignment of m independent and identically distributed nucleotide sites that have evolved according to a Jukes-Cantor model (Felsenstein, 2004) on an unrooted binary tree T, with n leaves. Each of the $2n-3$ edges of T have length λ, and thus the total tree length is $\xi = \lambda(2n-3)$. When n is large and $\xi \leqslant 1$, the expected number of substitutions per site is $\leqslant 1$; the number of edges on which a site changes state is approximately Poisson

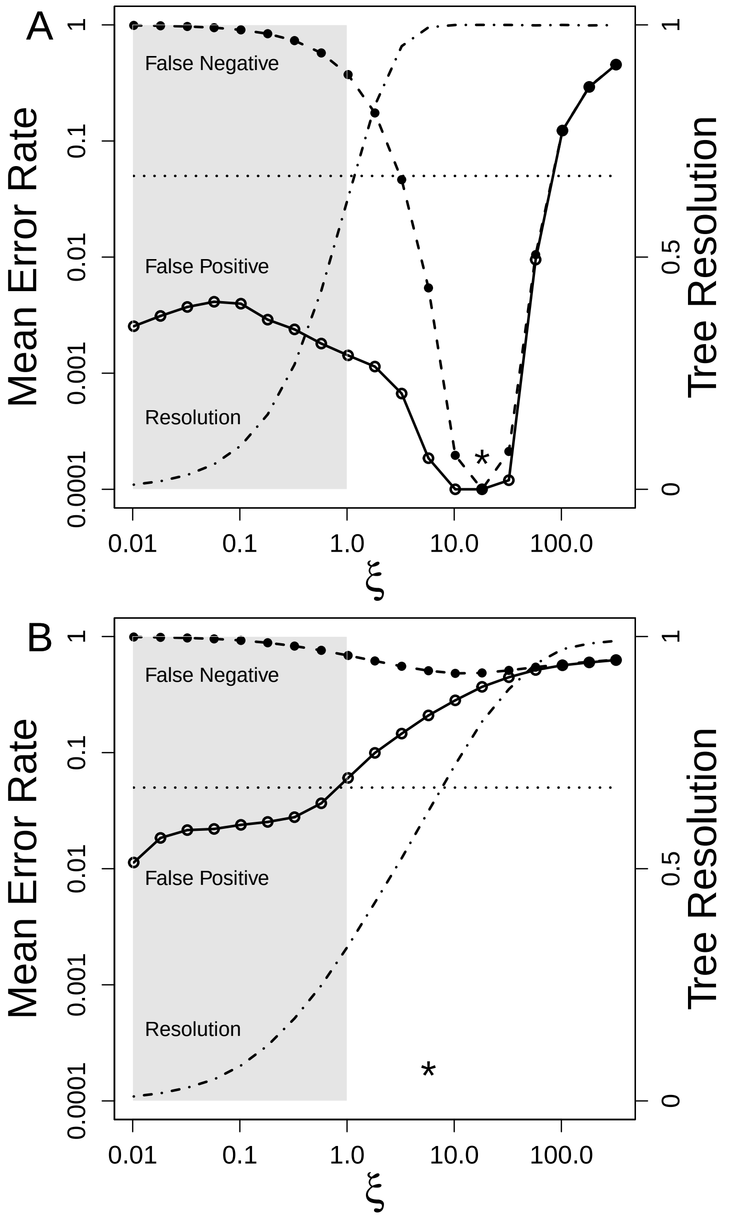

Fig. 2. Accuracy of maximum parsimony phylogeny reconstruction in simulations over a wide range of tree length, ξ, and other parameters. Open circles are mean false positive error rates; closed circles are mean false negative error rates (log scale left); dashed sigmoidal curve is fractional resolution of estimated tree (linear scale right). Trees are generated by a random proportional-to-distinguishable-arrangement (PDA) model for 513 taxa, from which a sequence alignment length of 1000 sites is generated. The dotted horizontal line is placed at an error rate of 0.05. Asterisk marks the location of the optimal tree length with best overall Robinson–Foulds accuracy, ξ^*. Each point is mean of 1000 replicates × 100 sub-replicates (see Methods). "Near-perfect" values of $\xi \leqslant 1.0$ are shaded. A) JC69 model with no edge length or across-site-rate variation. Because of y-axis log scaling, two y values of zero were set to 0.0001. B) JC69 model with substantial edge length and across-site-rate variation, both modeled as a gamma distribution with shape parameters $\alpha_e = \alpha_{ASR} = 0.25$).

distributed with mean ξ; and the probability of more than one change on an edge is low, meaning multiple changes at a site occur on distinct edges. Though these conditions will generate alignments dominated by "perfect" sites exhibiting no homoplasy, a few sites may exhibit homoplasy even with $\xi \leqslant 1$, which motivates the term "near- perfect". Under these conditions, tree reconstruction methods will tend to infer relatively unresolved trees unless the number of sites is very large.

Rare sites that exhibit homoplasy can introduce false positive splits on the inferred tree (Fig. 1). A naïve argument using Equation 1 might suggest that FP_T would depend on ξ roughly as $O(\xi^2)/O(\xi) = O(\xi)$, namely the ratio of the expected numbers of sites having changes on two edges (i.e., those that are potentially homoplastic and misleading) to those sites having only a single change (those that are reliable), for sufficiently small ξ. But because only one-third of those two-edge sites are actually homoplastic in a JC69 model,

$$FP_T \cong \xi/3,$$

which implies FP_T is small when ξ is small enough (e.g., $FP_T < 0.05$ whenever $\xi < 0.15$).

This approximation can be improved further by recognizing that not all two-edge homoplastic sites induce false positives, depending on their position in the true tree (Fig. 1). Given the evolutionary model, the probability that k perfect sites, and another site f that has evolved with two edge changes will produce a "false positive" under MP is denoted $\Phi_T^{(k)}$ (Theorem 1 above). Because this probability is often less than one, FP_T can remain below 0.05 at higher values of ξ than the naïve argument suggests.

If the true tree were known with some precision, the first part of Theorem 2 could be used directly to calculate false positive rates. However, in the "near-perfect" parameter space of large n and $\xi \leqslant 1$, estimates of the true tree are likely to be only partially resolved (Fig. 2). We therefore derive the expected false positive rate for a distribution, $\mathcal{D}$, of randomly generated trees of size n, $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$, generated from parameters based on the inferred tree. In the remainder of this paper, the "expected false positive rate" will generally refer to $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$. We assume that $\mathcal{D}$ is usually either a "proportional-to-distinguishable-arrangement" (PDA) or Yule-Harding (YH) distribution (Aldous, 2001), but also consider the two extreme cases of completely unbalanced (caterpillar) trees, and completely balanced (symmetric) trees. Unlike PDA and YH trees, these last two have a constant tree shape (with random leaf labels). From the second part of Theorem 2, we see that, for a JC69 model and trees inferred with MP, the following approximation holds increasingly well as n increases:

$$\mathbb{E}_{\mathcal{D}}[FP_{T_n}] \cong \frac{\xi}{3} \cdot \mathbb{E}_{\mathcal{D}}[\Omega(T_n, \xi, m)] \tag{6}$$

given the assumption that ξ is sufficiently small and the number of sites does not grow too quickly with the size of the tree. The function $\Omega(T_n, \xi, m)$, defined in Materials and Methods, is monotonically decreasing in ξ and m, and depends on the shape of T. Simulations indicate that the approximation is close for $\xi \leqslant 1$ (Fig. 3), but if many equally parsimonious trees are present, the search algorithm should take a strict consensus of a broad sample of those solutions (Fig. S3). $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ is better on average for PDA than YH trees, and both are bounded between a theoretical worst case error rate for symmetric and best case error rate for caterpillar trees. In fact, the expected false positive rate for the latter is just $4/(3m)$ in the limit of large n, which is independent of ξ.

Robustness to Violation of Assumptions

Violations of assumptions tend to increase the expected false positive rate above the predictions of Equation 6. For example, adding edge length (EL) variation or across-site-rate (ASR) variation increases $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ (Figs. 2, 4 and Fig. S4). The difference between predicted $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ based on Eqn. (6), with no edge length variation, and simulation-based estimates with edge length variation included is small when $\xi << 1$ but increases substantially as ξ increases. When edge length variation is large (gamma shape parameter $\alpha_e = 0.1$), there is no longer a local maximum value of $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ around $\xi = 0.1$; instead, $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ increases monotonically with ξ and eventually exceeds 5% for the simulated dataset sizes. The impact of ASR variation is deleterious at all values of ξ,

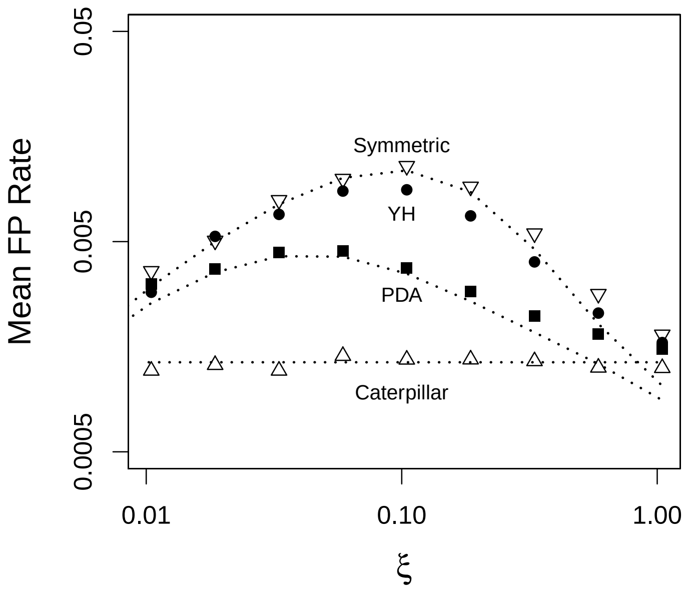

Fig. 3. Mean false positive rate in four tree models. Fit to theoretical predictions from Equation 6 (or the limit expression of $4/3m$ for caterpillar trees: see Methods) are shown by dashed lines. Each point is mean of 1000 replicates × 100 sub-replicates. Simulation conditions were $n = 513, m = 1000$, with a JC69 model. Predicted values are not known for YH model.

but even when ASR variation is large (gamma shape parameter $\alpha_{\text{ASR}} = 0.1$), the false positive rate remains slightly below 5% for simulated dataset sizes in the absence of EL variation (Fig. S4).

Departure of the substitution model from the JC69 model assumed in the "near-perfect" zone can also increase the expected false positive rate. For example, a strong transition–transversion bias increases $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ substantially, though it still remains well below 5% under our typical simulation conditions when $\xi \leqslant 1$ (Fig. S5).

Thus, the near-perfect tree length of $\xi \leqslant 1$ is a region in which rate variation appears to have less of an impact on false positive rates than when tree lengths are longer. This suggests that the definition of near-perfect zone in practice can include substantial rate variation as long as $\xi \leqslant 1$.

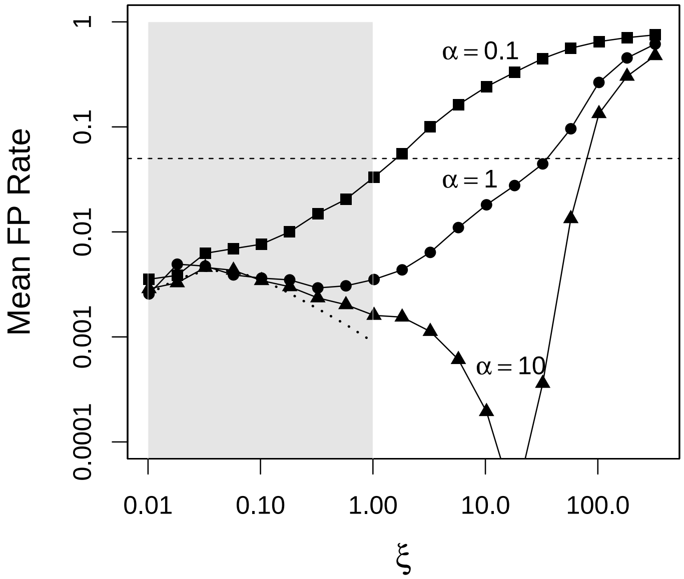

Fig. 4. Effect of edge length variation on expected false positive rate for different values of the shape parameter of the edge length gamma distribution, α_e. Smaller values of α_e correspond to higher rate variation. ASR variation is assumed absent. The dashed curve is the prediction from Eqn. (6), in which both sources of variation are absent. Simulation conditions assumed PDA trees with $n = 513$, $m = 1000$, 1000 replicates, 100 subreplicates. Gray rectangle shows "near-perfect" values of $\xi \leqslant 1$.

Expected False Positive Rates in Virus Phylogenies

We estimated key parameters from the trees and underlying data for 11 empirical virus phylogenies (Table 1, S1) and used simulation to estimate expected false positive rates (Figs. 5, S6). The studies span a wide range of tree size and resolution and alignment length, and their tree lengths span three orders of magnitude. Seven of these viruses fell within the "near-perfect" tree length zone of $\xi \leqslant 1.0$, and five of those had $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ below 5% for two different models of ASR variation (see Methods) and both random tree models. $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ was uniformly lower for PDA vs. YH models, and for invariant sites vs. gamma models (Fig. S6). As expected, lower $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ were generally observed for lower values of ξ.

Epidemics with young crown group ages on the order of years or decades (e.g., Zika

Table 1. Parameters of 11 empirical virus phylogenies

Abbreviation	Study	Leaves	Sites	Resolution
DENV	Dengue virus	1197	10264	0.8795
DENV-1	Dengue virus serotype 1	1067	10264	0.8160
EBOV	Ebolavirus	1610	18164	0.3632
EV-D68	Enterovirus 68	824	7293	0.8029
HIV-1	Human immunodeficiency virus type 1	189	1038	0.2193
MeV	Measles morbillivirus	109	15782	0.7009
MuV	Mumps virus	458	15154	0.2961
RSV	Respiratory syncytial virus	997	14986	0.6121
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2	583	29668	0.2324
WNV	West Nile virus	2512	10395	0.5960
ZIKV	Zika virus	543	10320	0.5453

virus, West Nile virus, and mumps virus) have an expected false positive rate below 5% expected in near-perfect trees, even though West Nile virus had a ξ slightly above 1. At the other extreme, dengue virus serotype 1, which does not represent a single epidemic, had a $\xi > 1$ and a correspondingly high expected false positive rate, and the phylogenetically more diverse dengue virus data representing all four serotypes had an even higher tree length and expected false positive rate.

The HIV-1 and measles virus trees both were outliers in having $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ above 5%, even though their tree lengths were below one. These two trees had the fewest taxa (Table 1), possibly indicating sensitivity to the assumption of large n in our results. Moreover, the HIV-1 tree was constructed with the fewest sites (representing only a single partial gene), which affects accuracy through the $\Omega(T_n, \xi, m)$ term in Eqn. 6. It may also lead to a poor estimate of the ASR gamma shape parameter, though $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ was about 5% when using the invariant sites model.

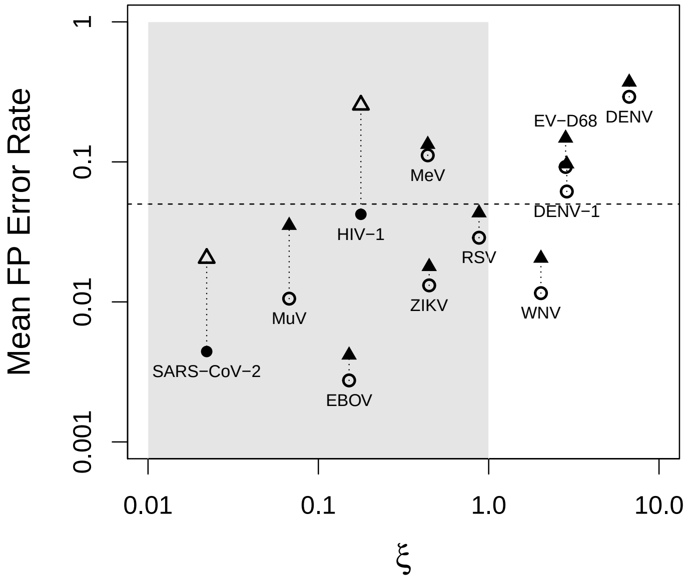

Fig. 5. Expected false positive rates, $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$, for 11 empirical virus phylogenetic datasets (Table 1) for maximum parsimony (MP) inference, estimated by simulation using parameters estimated from the data (Table S1). Abbreviations given in Table 1. Simulation experiments assumed ASR variation according to either an invariant sites model (circles) or a gamma distributed model (triangles) and a Yule–Harding random tree distribution (each point is mean of 500 replicates × 100 sub-replicates). Model point with higher likelihood is shaded. False positive rates assuming PDA random trees are uniformly slightly better (Fig. S6). The near-perfect zone of $\xi \leqslant 1.0$ is shaded. Horizontal dashed line indicates a 0.05 expected false positive rate.

Extension to Maximum Likelihood (ML) Inference

Theoretical results hint that ML and MP should reconstruct the same tree under "near-perfect" assumptions. For example, ML provably converges to MP when there are enough constant characters in an alignment, a condition similar to $\xi << 1$ (Tuffley and Steel, 1997, Thm. 3). Further arguments presented in the SI support this conjecture.

We used simulation to check how well Equation 6, derived for MP, predicted the expected false positive rate under ML inference in the near-perfect zone. Simulations with $\xi \leqslant 1$, a JC69 model, and no edge length or ASR variation, with trees inferred by IQTree2 (Minh et al., 2020) under the same model, are close to the equation's predictions (Fig. S7). Nonetheless, some differences were observed, which tended to imply better accuracy for

MP. These differences could largely be attributed to technical or implementation issues. First, the computational expense of ML searches makes it tempting to undertake fewer replicate searches for local optima, but this was as critical to improve the fit to Equation 6 for ML as it was for MP (Fig. S7). Second, ML programs set hard numerical lower bounds strictly greater than zero on edge lengths, often (by default) on the same order as $\bar{\lambda}$, so these must be reset downward to obtain correct tree likelihoods (Morel et al., 2020). Finally, inferred edge lengths that are larger than these programs' lower bounds but still smaller than about $1/m$ tend to be included in the ML tree despite weak evidence (IQTree2 issues a warning about this). We saw this in ML searches roughly when $\xi \geqslant 0.1$, when three-state sites become more common in alignments than they were at lower values of ξ. Even without homoplasy, ML tends to over-resolve trees in a way that elevates $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$. By collapsing short edge lengths inferred by ML to be less than $1/m$, this behavior can be mitigated (Fig. S7).

In general, ML is expected to be more accurate than MP under more realistic model conditions and higher rates, something we observed commonly in simulations in which $\xi > \xi^*$. However, simulations also suggest that in the near-perfect zone, MP can achieve an $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ comparable with ML but with much faster running times.

Accuracy and Support in Near-perfect Trees

False positive "accuracy", defined as $1 - \mathbb{E}_{\mathcal{D}}[FP_{T_n}]$, is very high in the near-perfect zone of small tree lengths, whereas conventional support values are quite variable in this zone under the same simulation conditions (Fig. 6). At very low ξ, the average MP bootstrap support is about the theoretically expected 64% for a single nonhomoplastic substitution supporting an edge (Felsenstein, 1985). Model-based support measures had higher values, with aBayes (Anisimova et al., 2011) being greater than ultrafast bootstrap (Hoang et al., 2018), which, in turn, was greater than SH-aLRT (Guindon et al., 2010), but only aBayes was close to $1 - \mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ across the range of tree lengths in the near-perfect zone. Notably, aBayes is the only one of the four metrics that is not based on resampling.

We explored other factors impacting support in the boundary case of perfect trees. For sequence length, we computed standard support metrics in an ML framework in perfect four-taxon datasets, in which each branch was defined by a single change, and alignments range between 40 nt and 30,000 nt (Fig. S8). As observed for MP, non-parametric ML bootstrap support is approximately 63%, regardless of sequence length, in accordance with theoretical predictions (Felsenstein, 1973). Of the ML model-based support metrics, aBayes provided higher values than ultrafast bootstrap and SH-aLRT, both of which rely on bootstrap resampling. The aBayes support reached $\geqslant$95% for alignments as short as 100 nt, which tracked the full Bayesian posterior support estimates that had support $\geqslant$95% in alignments as short as 60 nt. The discrepancy between the Bayesian estimates and those that use bootstrap resampling, in light of our other results, suggests that resampling methods used in the presence of splits defined by only a single informative site may fail to integrate relevant information about low tree lengths.

On the other hand, in perfect trees from 8–128 taxa, in which the mean edge length remained the same (but therefore ξ grew with n), mean SH-aLRT and aBayes support was unchanged, but mean ultrafast bootstrap support increased (Fig. S9).

Discussion

In this paper, we study a “near-perfect” parameter space for phylogenetic inference on large trees with small tree lengths and no rate variation within or between sites or edges. The “near-perfect” tree length of $\xi \leqslant 1$ means that few sites exhibit homoplasy and, for MP inference, the false positive rate can be much better than the false negative rate and well under 5% for typical datasets with thousands of sites. The near-perfect conditions defined here to allow mathematical derivations appear to be sufficient but not necessary. For example, with no rate variation, the false positive rate can be very good even when $\xi > 1$ (Fig. 2A, S5), and, if $\xi < 1$, a substantial level of rate variation can be present

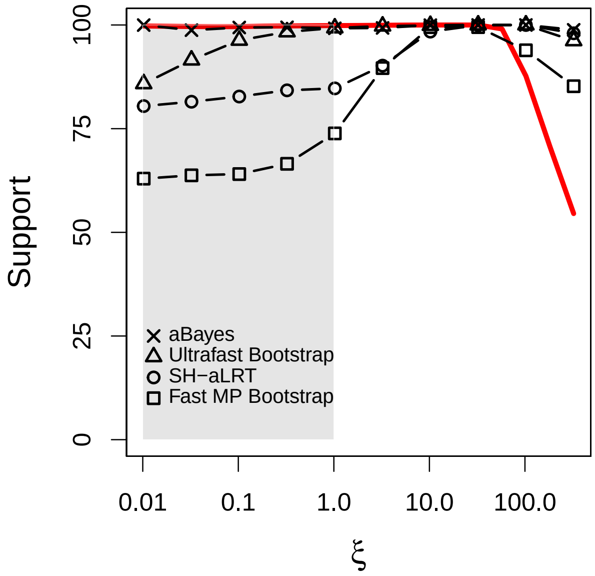

Fig. 6. Statistical support measures compared to expected false positive accuracy. The red curve is the mean value of $(1 - \mathbb{E}_{\mathcal{D}}[FP_{T_n}]) \times 100$ in simulations. The near-perfect parameter space is shaded.

without elevating the false positive rate by nearly as much as when $\xi > 1$ (Fig. 2,4, S4).

The second case is clearly more relevant in real-world data. The 11 empirical virus datasets all had substantial rate variation and showed a general increase in false positive rate with ξ, with almost all rates below 5% occurring when $\xi \leqslant 1$, much like the predicted patterns seen in Fig. 2B and Fig. 4. This accords with our simulation results suggesting that the good "near-perfect" false positive rates may emerge even when relaxing the strict near-perfect assumption of no rate variation—as long as $\xi \leqslant 1$.

These and many other empirical findings about RNA virus phylogenies sampled intensively in epidemics postdate much of the extensive body of other work on accuracy and support in phylogenetics. Not surprisingly, little note has been made about the stark contrast between false positive and false negative rates in phylogenies in which tree length is well below the optimal tree length for "overall accuracy", since published examples have been relatively rare. The goal of much of the field of phylogenetics is, after all, to maximize tree resolution, even if this effort requires adding (or switching to) sequence data with more variation and thus longer tree lengths.

Because "near-perfect" datasets reflect a combination of the number of taxa and sites, evolutionary rate and time parameters, and assumptions about the substitution model, they also implicitly reflect sampling of the true tree, which is particularly relevant in epidemic trees in which sampling is far below disease incidence. Sampling can continue over time, increasing n, and the viruses continue to evolve over time, increasing the depth of the tree. Both of these increase ξ but in different ways; therefore, it is possible for the same RNA virus to have near-perfect and not near-perfect datasets depending on the study. For example, the SARS-Cov2 dataset we included had $n = 583$ and $\xi = 0.02$, well within the "near-perfect" zone, but a much more intensively sampled tree over a longer period of time (Lanfear, 2020) with $n = 147156$ has a tree length of $\xi = 3.89$ (after collapsing any edges with $\lambda \leqslant 1.1 \times 10^{-6}$), which is remarkably small for such a large tree but lies just outside our definition of near-perfect.

Other mathematical results on phylogenetic accuracy have largely focused on either the limiting case of infinite sequence length ("consistency"), or the number of sites needed for accurate inference (the "sequence length requirement"). For MP, for example, the shortest edge length is critical and $\lim_{m\to\infty} \text{Prob}(\hat{T}_{MP} = T) = 1$ as long as $\lambda_{\text{min}} > \xi^2/(1-\xi)$ (Steel, 2000, Thm. 1(A)). More generally, let m' be the number of sites needed for $\text{Prob}(\hat{T}_{MP} = T)$ to exceed some fixed required accuracy. For the neighbor–joining method m' grows exponentially with n (Lacey and Chang, 2006); for ML, m' is polynomial or better in n, depending on edge lengths (Roch and Sly, 2017). Moreover, m' also grows as $O(1/\lambda^2_{\text{min}})$ for ML and some more ad hoc estimators (Erdös et al., 1999; Roch, 2019), implying again that short edges tend to degrade accuracy when accuracy is defined in terms of total agreement between T and $\hat{T}$, unlike here.

A cryptic factor affecting the false positive rate is tree shape. Highly asymmetric trees have better expected false positive rates than highly symmetric trees, because expected path lengths are longer and it is harder to induce false positive splits by chance (Fig. 1). Thus, a random sample of PDA trees will have a better $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$ than more symmetrical YH trees. Differences in tree shape among RNA virus phylogenies have long been noted (Grenfell et al., 2004), such as the typically more asymmetric influenza trees.

Perfect and near-perfect phylogenies have been studied as discrete optimization problems (Gusfield, 1997; Fernandez-Baca and Lagergren, 2003) in which the goal is to find an optimal tree when, at most, some small number of sites exhibit homoplasy. Little of this work has considered accuracy per se, but Gronau et al. (Gronau et al., 2012) highlighted the connection between short edge lengths and false positives, and developed a "fast converging" algorithm (i.e., having an $O(\text{poly}(n))$ sequence length requirement) that returns a tree with short edges collapsed when they do not meet a threshold probability of being correct, thus minimizing false positives. The connection between this tree and those built by more conventional methods is unclear, but it may be a promising approach for trees in the near-perfect zone.

Model-based phylogenetic inference methods such as ML and Bayesian inference are generally regarded as theoretically superior to MP, especially for datasets that fit substitution models much more complex than our "near-perfect" JC69 model with no rate variation. Though our mathematical results for expected false positive rates were derived for MP, there is both relevant theory and considerable simulation evidence to suggest that in the near-perfect zone, the ML expected false positive rate is approximated by the MP theory, both in terms of its absolute value and its shape as a function of tree length. As ξ increases, especially above ξ^*, ML consistently has better accuracy than MP, but we conjecture that the false positive rates of MP and ML differ much less as ξ gets very small. Further work is needed to test this conjecture.

The connection between the false positive rate as a measure of accuracy and conventional measures of phylogenetic support appears to be sensitive to the choice of support method when $\xi << 1$ (Fig. 6). The aBayes method corresponds well to what is implied by $1 - \mathbb{E}_{\mathcal{D}}[FP_{T_n}]$, but resampling methods using either likelihood or parsimony correspond less well. The connection between phylogenetic accuracy and support in frequentist and Bayesian settings has been studied in detail (Felsenstein, 1985; Hillis and Bull, 1993; Felsenstein and Kishino, 1993; Efron et al., 1996; Susko, 2008, 2009; Alfaro and Holder, 2006; Simmons and Norton, 2014), but remains somewhat fraught. We hesitate to draw firm conclusions without a formal analysis of support in the "near-perfect" parameter space but note the variability in support estimates we found (Fig. 6).

The low false positive rate in near-perfect trees suggests that phylogenies describing viral epidemics in this zone can be interpreted directly without defaulting to identifying clades with strong support values. Frequent convergent evolution, and recombination in positive-strand RNA viruses, can complicate phylogenetic inference and may increase the false positive rate in real-world trees (Morel et al., 2020). Nonetheless, if individual clade support needs to be invoked, we recommend Bayesian approaches that do not rely on bootstrap resampling of sparse substitutions.

The benefit of real-time viral genomic sequencing for public health action became apparent during the 2014–2015 West African Ebola epidemic (Gire et al., 2014), and is a critical component of tracking the COVID-19 pandemic (Oude Munnink et al., 2020; Grubaugh et al., 2021). Consequently, the viruses responsible for these diseases, Ebolavirus and SARS-CoV-2, epitomize near-perfect phylogenetic trees in our analysis. We can expect a greater intensity of genomic sequencing accompanying future viral outbreaks, increasing the importance and relevance of near-perfect phylogenies.

Acknowledgements

We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared via GISAID the SARS-CoV-2 genomic sequence data on which this research is based. A complete list acknowledging the authors who submitted data analyzed in this study can be found in Data S1. MJS thanks the

University of Arizona's HPC facility, Bio5 Institute, and Rod Wing's lab for computing support. JOW was supported by an NIH-NIAID R01 AI135992.

Disclosure Statement

The authors have no conflicts of interest related to this work.

References

Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statistical Science 16:23–34.

Alfaro, M. E. and M. T. Holder. 2006. The posterior and the prior in Bayesian phylogenetics. Annual Review of Ecology Evolution and Systematics 37:19–42.

Anisimova, M., M. Gil, J. F. Dufayard, C. Dessimoz, and O. Gascuel. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Systematic Biology 60:685–99.

Awasthi, P., A. Blum, J. Morgenstern, and O. Sheffet. 2012. Additive approximation for near-perfect phylogeny construction. Pages 25–36 *in* Approximation, randomization, and combinatorial optimization. Algorithms and techniques (M. Goemans, K. Jansen, J. Rolim, and L. Trevisan, eds.). Springer, Berlin.

Bedford, T. and D. L. Hartl. 2008. Overdispersion of the molecular clock: Temporal variation of gene-specific substitution rates in Drosophila. Molecular Biology and Evolution 25:1631–1638.

Berry, V. and O. Gascuel. 1996. On the interpretation of bootstrap trees: Appropriate threshold of clade selection and induced gain. Molecular Biology and Evolution 13:999–1011.

Bininda-Emonds, O. R. P., S. G. Brady, J. Kim, and M. J. Sanderson. 2001. Scaling of accuracy in extremely large phylogenetic trees. Pacific Symposium on Biocomputing 6:547–558.

Bortolussi, N., E. Durand, M. Blum, and O. François. 2006. apTreeshape: Statistical analysis of phylogenetic tree shape. Bioinformatics 22:363–364.

Campbell, F., C. Strang, N. Ferguson, A. Cori, and T. Jombart. 2018. When are pathogen genome sequences informative of transmission events? PLoS Pathog 14:e1006885.

Dudas, G. and T. Bedford. 2019. The ability of single genes vs full genomes to resolve time and space in outbreak analysis. BMC Evol Biol 19:232.

Dudas, G., L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, F. Bielejec, S. L. Caddy, M. Cotten, J. D'Ambrozio, S. Dellicour, A. Di Caro, J. W. Diclaro, S. Duraffour, M. J. Elmore, L. S. Fakoli, O. Faye, M. L. Gilbert, S. M. Gevao, S. Gire, A. Gladden-Young, A. Gnirke, A. Goba, D. S. Grant, B. L. Haagmans, J. A. Hiscox, U. Jah, J. R. Kugelman, D. Liu, J. Lu, C. M. Malboeuf, S. Mate, D. A. Matthews, C. B. Matranga, L. W. Meredith, J. Qu, J. Quick, S. D. Pas, M. V. T. Phan, G. Pollakis, C. B. Reusken, M. Sanchez-Lockhart, S. F. Schaffner, J. S. Schieffelin, R. S. Sealfon, E. Simon-Loriere, S. L. Smits, K. Stoecker, L. Thorne, E. A. Tobin, M. A. Vandi, S. J. Watson, K. West, S. Whitmer, M. R. Wiley, S. M. Winnicki, S. Wohl, R. Wolfel, N. L. Yozwiak, K. G. Andersen, S. O. Blyden, F. Bolay, M. W. Carroll, B. Dahn, B. Diallo, P. Formenty, C. Fraser, G. F. Gao, R. F. Garry, I. Goodfellow, S. Gunther, C. T. Happi, E. C. Holmes, B. Kargbo, S. Keita, P. Kellam, M. P. G. Koopmans, J. H. Kuhn, N. J. Loman, N. Magassouba, D. Naidoo, S. T. Nichol, T. Nyenswah, G. Palacios, O. G. Pybus, P. C. Sabeti, A. Sall, U. Stroher, I. Wurie, M. A. Suchard, P. Lemey, and A. Rambaut. 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544:309–315.

Dyrdak, R., M. Mastafa, E. B. Hodcroft, R. A. Neher, and J. Albert. 2019. Intra- and interpatient evolution of enterovirus D68 analyzed by whole-genome deep sequencing. Virus Evolution 5:vez007.

Efron, B., E. Halloran, and S. Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Sciences of the United States of America 93:13429–13434.

Erdös, P. L., M. A. Steel, L. A. Szekely, and T. J. Warnow. 1999. A few logs suffice to build (almost) all trees (i). Random Structures and Algorithms 14:153–184 mAR.

Felsenstein, J. 1973. Maximum likelihood and minimum steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22:240–249.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Press, Sunderland, MA.

Felsenstein, J. and H. Kishino. 1993. Is there something wrong with the bootstrap on phylogenies? a reply to Hillis and Bull. Systematic Biology 42:182–192.

Fernandez-Baca, D. and J. Lagergren. 2003. A polynomial-time algorithm for near-perfect phylogeny. Siam Journal on Computing 32:1115–1127.

Gire, S. K., A. Goba, K. G. Andersen, R. S. Sealfon, D. J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas, S. Wohl, L. M. Moses, N. L. Yozwiak, S. Winnicki, C. B. Matranga, C. M. Malboeuf, J. Qu, A. D. Gladden, S. F. Schaffner, X. Yang, P. P. Jiang, M. Nekoui, A. Colubri, M. R. Coomber, M. Fonnie, A. Moigboi, M. Gbakie, F. K. Kamara, V. Tucker, E. Konuwa, S. Saffa, J. Sellu, A. A. Jalloh, A. Kovoma, J. Koninga, I. Mustapha, K. Kargbo, M. Foday, M. Yillah, F. Kanneh, W. Robert, J. L. Massally, S. B. Chapman, J. Bochicchio, C. Murphy, C. Nusbaum, S. Young, B. W. Birren, D. S. Grant, J. S. Scheiffelin, E. S. Lander, C. Happi, S. M. Gevao, A. Gnirke, A. Rambaut, R. F. Garry, S. H. Khan, and P. C. Sabeti. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science 345:1369–72.

Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303:327–32.

Gronau, I., S. Moran, and S. Snir. 2012. Fast and reliable reconstruction of phylogenetic trees with indistinguishable edges. Random Structures and Algorithms 40:350–384.

Grubaugh, N. D., E. B. Hodcroft, J. R. Fauver, A. L. Phelan, and M. Cevik. 2021. Public health actions to control new SARS-CoV-2 variants. Cell 184:1127–1132.

Grubaugh, N. D., J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, and K. G. Andersen. 2019. Tracking virus outbreaks in the twenty-first century. Nat Microbiol 4:10–19.

Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–21.

Gusfield, D. 1997. Algorithms on strings, trees and sequences. Cambridge University Press, New York.

Hadfield, J., A. F. Brito, D. M. Swetnam, C. B. F. Vogels, R. E. Tokarz, K. G. Andersen, R. C. Smith, T. Bedford, and N. D. Grubaugh. 2019. Twenty years of West Nile virus spread and evolution in the Americas visualized by Nextstrain. PLoS Pathogens 15:e1008042.

Hadfield, J., C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. 2018. Nextstrain: Real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123.

Hillis, D. M. and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology 42:182–192.

Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh. 2018. UFBoot2: Improving the ultrafast bootstrap approximation. Molecular Biology and Evolution 35:518–522.

Huelsenbeck, J. P. and D. M. Hillis. 1993. Success of phylogenetic methods in the 4-taxon case. Systematic Biology 42:247–264.

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

Katoh, K. and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–80.

Lacey, M. R. and J. T. Chang. 2006. A signal-to-noise analysis of phylogeny estimation by neighbor-joining: insufficiency of polynomial length sequences. Mathematical Biosciences 199:188–215.

Lanfear, R. 2020. A global phylogeny of SARS-CoV-2 sequences from GISAID.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution 37:1530–1534.

Morel, B., P. Barbera, L. Czech, B. Bettisworth, L. Hubner, S. Lutteropp, D. Serdari, E. G. Kostaki, I. Mamais, A. M. Kozlov, P. Pavlidis, D. Paraskevis, and A. Stamatakis. 2020. Phylogenetic analysis of SARS-CoV-2 data is difficult. Molecular Biology and Evolution .

Oude Munnink, B. B., D. F. Nieuwenhuijse, M. Stein, A. O'Toole, M. Haverkate, M. Mollers, S. K. Kamga, C. Schapendonk, M. Pronk, P. Lexmond, A. van der Linden, T. Bestebroer, I. Chestakova, R. J. Overmars, S. van Nieuwkoop, R. Molenkamp, A. A. van der Eijk, C. GeurtsvanKessel, H. Vennema, A. Meijer, A. Rambaut, J. van Dissel, R. S. Sikkema, A. Timen, M. Koopmans, and t. Dutch-Covid-19 response. 2020. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nature Medicine 26:1405–1410.

Pekar, J., M. Worobey, N. Moshiri, K. Scheffler, and J. O. Wertheim. 2021. Timing the SARS-CoV-2 index case in Hubei province. Science .

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53:131–147.

Roch, S. 2019. Hands-on introduction to sequence-length requirements in phylogenetics. Pages 47–86 *in* Bioinformatics and Phylogenetics: Seminal Contributions of Bernard Moret (T. Warnow, ed.). Springer International Publishing, Cham.

Roch, S. and A. Sly. 2017. Phase transition in the sample complexity of likelihood-based phylogeny inference. Probability Theory and Related Fields 169:3–62.

Ronquist, F. and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–4.

Rouet, F., J. Nouhin, D. P. Zheng, B. Roche, A. Black, S. Prak, M. Leoz, C. Gaudy-Graffin, L. Ferradini, C. Mom, S. Mam, C. Gautier, G. Lesage, S. Ken, K. Phon, A. Kerleguer, C. Yang, W. Killam, M. Fujita, C. Mean, D. Fontenille, F. Barin, J. C. Plantier, T. Bedford, A. Ramos, and V. Saphonn. 2018. Massive iatrogenic outbreak of Human Immunodeficiency Virus Type 1 in rural Cambodia, 2014–2015. Clin Infect Dis 66:1733–1741.

Simmons, M. P. and A. P. Norton. 2014. Divergent maximum-likelihood-branch-support values for polytomies. Molecular Phylogenetics and Evolution 73:87–96.

Smirnov, D. and T. Warnow. 2021. Phylogeny estimation given sequence length heterogeneity. Syst. Biol. 70:268–282.

Steel, M. 2000. Sufficient conditions for two tree reconstruction techniques to succeed on sufficiently long sequences. SIAM Journal on Discrete Mathematics 14:36–48.

Steel, M. and C. Leuenberger. 2017. The optimal rate for resolving a near-polytomy in a phylogeny. Journal of theoretical biology 420:174–179.

Susko, E. 2008. On the distributions of bootstrap support and posterior distributions for a star tree. Systematic Biology 57:602–612.

Susko, E. 2009. Bootstrap support is not first-order correct. Systematic Biology 58:211–223.

Tuffley, C. and M. Steel. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bulletin of Mathematical Biology 59:581–607.

Wake, D. B., M. H. Wake, and C. D. Specht. 2011. Homoplasy: From detecting pattern to determining process and mechanism of evolution. Science 331:1032–1035.

Warnow, T. 2013. Large-scale multiple sequence alignment and phylogeny estimation. Pages 85–146 *in* Models and algorithms for genome evolution (C. Chauve, N. El-Mabrouk, and E. Tannier, eds.). Springer, London.

Yang, Z. 1998. On the best evolutionary rate for phylogenetic analysis. Systematic Biology 47:125–133.

XX (2021), **0**, 0, *pp.* 1–34
doi:XX/XX/output

Supplementary Information: Accuracy in near-perfect virus phylogenies

Joel O. Wertheim, Mike Steel, and Michael J. Sanderson

Mathematical Results and Proofs

Definitions and preliminary observations

Let $B(n)$ denote the set of unrooted binary phylogenetic trees with leaf set $[n] = \{1, 2, \ldots, n\}$. Thus, $B(n) = (2n-5)!!$. Note that a tree $T \in B(n)$ has $2n-3$ edges. Consider a Jukes-Cantor model in which all edges have the same value of λ. Thus, the probability of a state change between the endpoints of an edge e, denoted p_e, is given by $p_e = p$ where $p = \frac{3}{4}(1 - \exp(-4\lambda/3))$.

A *character* refers to the assignment of states to the taxa at a given site of an alignment. Let us say that a character evolves 'perfectly' on T if there is a single change of state across one interior edge (say e) and no change of state any other edge of T. The probability that this occurs for any given interior edge e is: $p_e \prod_{e' \neq e}(1 - p_{e'}) = p(1-p)^{2n-4}$. Note that this probability does not depend on the particular choice of e or on the shape of T. Note also that a perfectly evolved character has at least two species present in each state (otherwise the change would have been on a pendant edge of T).

We will say that a character f evolves on T with c *edge changes on* $e_1, \ldots, e_c$ if state changes occur on edges $e_1, \ldots, e_c$ and on no other edge of T. More briefly, we say that f evolves on T with c edge changes if f evolves with c edge changes for some set of c distinct edges of T (mostly we will deal with the case $c = 2$). The probability that a character evolves on T with 2 edge changes on e_1, e_2 is $p^2(1-p)^{2n-5}$, while the probability that a character evolves on T with 2 state changes is $\binom{2n-3}{2}p^2(1-p)^{2n-5}$.

We pause here to make an observation: If a data set consists of characters each of which have evolved on T with either 1 or 2 edge changes, then the maximum parsimony tree(s) and the maximum compatibility tree(s) for this data set will be exactly the same. Moreover, if there are sufficiently many constant characters, then any maximum likelihood tree will also be one of these trees.

Recall that a *split* refers to a bipartition of the leaf set $[n]$ into two non-empty subsets (and splits are induced by binary characters). A character that has evolved perfectly on T produces a split, and these splits (across a set of perfectly evolved characters) are compatible and so form a (generally unresolved/non-binary) tree on leaf set $[n]$.

Now suppose that m characters evolve on T and suppose that, of these m characters, k of them are perfectly evolved on T (note that more than one of these characters may correspond to the same split of T).

Next, consider a single additional character f which has evolved on T with two edge changes on e_1, e_2 (there is no restriction that these are interior edges). The probability that this character f is a binary character is $\frac{1}{3}$ (as noted earlier). Moreover, in that case, the bipartition induced by f corresponds to a split of T if and only if e_1 and e_2 are adjacent.

The question now arises as to whether we can discover that f gives a false split of T (without knowing T), just on the basis of the other k perfectly-evolved characters.

We next introduce some further notation. We will let ξ denote the expected number of state changes in the tree T per character, under the model described. Thus ξ equals the per-edge rate λ times the number of edges of T, and so

$$\xi = \lambda \cdot (2n-3).$$

We will also use the following standard notation throughout: we write $f(n) \sim g(n)$ if the ratio $f(n)/g(n)$ tends to 1 as n becomes large.

Given a tree $T \in B(n)$, let $d_T(e_1, e_2)$ denote the number of edges of T that lie strictly within the path between e_1 and e_2 (i.e. excluding e_1 and e_2). Thus, e_1 and e_2 are adjacent if and only if $d_T(e_1, e_2) = 0$

Next, given $T \in B(n)$, let $\varphi_T = (\varphi_T(0), \varphi_T(1), \dots, \varphi_T(n-3))$ where $\varphi_T(i)$ is the number of (unordered) pairs of edges $\{e, e'\}$ of T for which $d_T(e, e') = i$. Thus $\varphi_T(0) = 3(n-2)$, however the other values comprising φ_T depend on the shape of T. In particular, for a complete balanced binary tree with $n = 2^h$ leaves we have $\varphi_T(i) = 0$ for all $i \geqslant 2\log_2(n) - 2$, while for a caterpillar tree with the same number of leaves we have $\varphi_T(i) > 0$ for all $i \leqslant n-3$.

Note that the sequence φ_T is a topological invariant (i.e. it depends only on the unlabelled shape of the tree) and does not depend on any other parameters mentioned above. Clearly, for any $T \in B(n)$ we have:

$$\sum_{i=0}^{n-3} \varphi_T(i) = \binom{2n-3}{2}, \tag{1}$$

since both sides of this equation count the number of pairs of edges of T. For i between 1 and $n-3$, we will let

$$\tilde{\varphi}_T(i) = \frac{\varphi_T(i)}{\binom{2n-3}{2}}. \tag{2}$$

Thus, Eqn. (1) translates to the identity $\sum_{i=1}^{n-3} \tilde{\varphi}_T(i) = 1$. Thus $\tilde{\varphi}_T(i)$ is the probability that a pair on edges (selected uniformly at random from all pairs) has exactly i edges lying on the path strictly between the two selected edges.

Next we state a simple combinatorial result that will be useful in the proof of the first theorem.

Lemma 1 Let f be a binary character that has evolved on T by 2 edge changes on e_1 and e_2, and let f' be a character that has perfectly evolved on T by a change on a single interior edge e. Then f and f' are compatible (i.e. induce compatible splits) if and only if e does not lie on the path in T that is strictly between e_1 and e_2 (i.e. the red edges in Fig. 1).

Proof: First suppose that e lies on one of the red edges shown in Fig. 1, say between

t_i and t_{i+1} for $i \in \{1, \ldots, r-1\}$. Let x_j be a leaf of t_j for $j \in \{0, i, i+1, r+1\}$. Then if f is a binary character arising from changes on (just) e_1 and e_2 then f splits the set $\{x_0, x_i, x_{i+1}, x_{r+1}\}$ as $x_0x_{r+1}|x_ix_{i+1}$ while f' splits this same set as $x_0x_i|x_{i+1}x_{r+1}$. Since these two partial splits are incompatible, so too are f and f'.

Next suppose that e lies in one of the green subtrees of Fig. 1 (including the stem edge of that subtree), say subtree t_i for $i \in \{0, 1, \ldots, r+1\}$. We consider the possible cases: (1) $e \in \{e_1, e_2\}$, or (2) e is an edge of t_0 or t_{r+1} or (3) e is an edge of t_i or its stem edge for $i \in \{1, \ldots, r\}$. Let $X_i \subset [n]$ denote the subset of leaves of T that are leaves of t_i.

In case (1), it suffices by symmetry to consider the case $e = e_1$. Then f' induces the full split $X_0|([n] - X_0)$ while f induces the full split $(X_0 \cup X_{r+1})|([n] - (X_0 \cup X_{r+1}))$ and these two splits are compatible, since the set on the left of the first split is contained in the set on the left of the second split.

In case (2), it suffices by symmetry to consider the case where e is an edge of t_0. In that case f' induces a full split of the for $Y|([n] - Y)$ where $Y \subseteq X_0$, and this split is again compatible with the full split $(X_0 \cup X_{r+1})|([n] - (X_0 \cup X_{r+1}))$ induced by f since Y is a subset of the left part of this split.

In case (3), suppose that e is an edge of t_i or its stem edge. In that case, f' induces the split $W|([n] - W)$ where $W \subseteq X_i$, and since $X_0 \cup X_{r+1} \subseteq [n] - W$ it follows that f is compatible with f'. This completes the proof of Lemma 1. □

Next, let $\Phi_T^{(k)}$ be the probability that a character f that has evolved on T with 2 edge changes has the following three additional properties:

(C-i) it is a binary character,

(C-ii) the corresponding split is not a split of T, and

(C-iii) the split described by f is compatible with k characters that are perfectly evolved on T (by the Markovian process described above).

The point of these conditions is that we might add the split corresponding to f to the other k compatible spits which still be compatible, but create a 'false positive' split in the resulting tree (note that we do not know a-priori which splits are the perfectly evolved ones, as we are assuming that T is not known).

We now present an exact expression for $\Phi_T^{(k)}$. Firstly, let $\mathcal{C}_T$ be the collection of all (unordered) pairs of non-adjacent edges in T. Thus $|\mathcal{C}_T| = \binom{2n-3}{2} - 3(n-2)$. For $\{e_1, e_2\} \in \mathcal{C}_T$, let

$$R_T(e_1, e_2) = \frac{d_T(e_1, e_2)}{n-3}, \tag{3}$$

Thus $R_T(e_1, e_2)$ is the proportion of interior edges of T that lie between two non-adjacent edges e_1 and e_2.

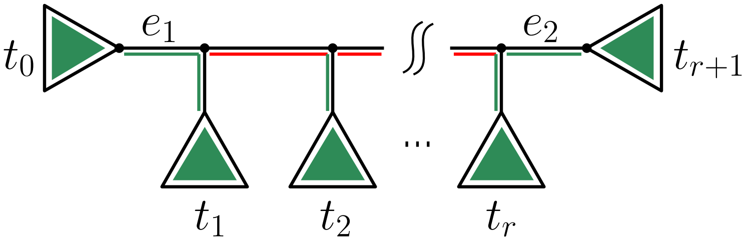

Fig. 1. A representation of T determined by the pair $\{e_1, e_2\}$ on which changes of state occur for character f. When f is a binary character, a perfectly evolved character f' will be compatible with f provided f' corresponds to a change of state on an edge in the green portions of the tree; otherwise, if the change is on one of the $r-1$ edges marked in red, the two characters will be incompatible.

Next, let μ_T denote the average value of $d_T(e, e')$ over all pairs of edges $\{e, e'\}$ sampled from T. That is:

$$\mu_T = \frac{1}{\binom{2n-3}{2}} \sum_{\{e,e'\}} d_T(e, e'). \tag{4}$$

THEOREM 2 For each $T \in B(n)$, and $k \geqslant 1$ we have:

(i)

$$\Phi_T^{(k)} = \frac{1}{3} \cdot \frac{1}{\binom{2n-3}{2}} \cdot \sum_{\{e_1,e_2\} \in \mathcal{C}_T} (1 - R_T(e_1,e_2))^k,$$

$$= \frac{1}{3} \cdot \sum_{i=1}^{n-3} \tilde{\varphi}_T(i) \left(1 - \frac{i}{(n-3)}\right)^k.$$

(ii)

$$\Phi_T^{(k)} \geqslant \frac{1}{3} \cdot \left(1 - \frac{\mu_T}{n-3}\right)^k - \frac{1}{(2n-3)}.$$

Proof: Part (i): Under the Jukes-Cantor model, the (conditional) probability that a randomly evolved character f is binary, given that is has evolved on T with 2 edge changes is $\frac{1}{3}$. To see this, let e_1, e_2 denote the two edges on which there are state changes. Then if the state at the left-hand end vertex of e_1 (as shown in Fig. 1) is α and the state on the right end vertex of e_1 is β then for f to be a binary character we require e_2 to involve the transition from β back to α (which is one of the three possible other states that β can change to on e_2, and each change has equal probability under the Jukes-Cantor model).

Now consider a single random character f' that has perfectly evolved on T. Since (a) each of the $n-3$ interior edges of T has the same probability of being the 'change edge' associated with f', and (b) f' is compatible with f if and only if the associated change edge for that character does not lie on the path that is (strictly) between e_1 and e_2 (by Lemma 1) and (c) the proportion of interior edges of T that do not lie strictly between the edges e_1 and e_2 is $1 - R_T(e_1, e_2)$, it follows that the probability that f' is compatible with f is $1 - R_T(e_1, e_2)$.

Thus, for k characters that have independently perfectly evolved on T the probability that each of them is compatible with f is $(1 - R_T(e_1, e_2))^k$.

Finally, each choice for f across all pairs of edges of T (nor just the non-adjacent pairs) has equal probability, and if we let $\eta(e_1, e_2) = 1$ if $\{e_1, e_2\}$ are not adjacent, and 0 otherwise, then the probability that f satisfies properties (C-i), (C-ii) and (C-iii) above is the average value of $\frac{1}{3} \cdot \eta(e_1, e_2) \cdot (1 - R_T(e_1, e_2))^k$ across all $\binom{2n-3}{2}$ pairs of edges $\{e_1, e_2\}$ of T, which gives the first expression in Theorem 2(i), and the second expression follows directly.

Part (ii). Let $Q_T^{(k)} = \cdot \sum_{i=0}^{n-3} \tilde{\varphi}_T(i) \left(1 - \frac{i}{(n-3)}\right)^k$. Then $Q_T^{(k)}$ is the expected value of $(1 - R_T(e_1, e_2))^k$ when a pair of edges of T is selected uniformly at random from the set of all such pairs. Since $f(x) = (1-x)^k$ is a convex function, Jensen's inequality then gives:

$$Q_T^{(k)} = \mathbb{E}[(1 - R_T(e_1, e_2))^k] \geqslant (1 - \mathbb{E}[R_T(e_1, e_2)])^k = (1 - \mu_T)^k. \tag{5}$$

Now $\Phi_T(k) = \frac{1}{3}\left(Q_T^{(k)} - \frac{3(n-2)}{\binom{2n-3}{2}}(1-0)^k\right)$, since there are $3(n-2)$ pairs of adjacent edges in T, and $\frac{1}{3} \cdot \frac{3(n-2)}{\binom{2n-3}{2}} = \frac{1}{(2n-3)}$, and so applying this to the Eqn. 5 gives the result claimed.

□

Remarks: Theorem 2(i) shows that for fixed k and n, the shape of T plays a significant role in determining $\Phi_T^{(k)}$; in particular, pectinate trees (such as caterpillars) will have a smaller value of $\Phi_T^{(k)}$ than more balanced trees. Indeed, it is possible to exactly calculate the value of $\Phi_T^{(k)}$ for the two extreme families: caterpillars and fully-balanced trees to determine the extent of this dependence, as we describe in the next section. Notice also that the exact computation of $\Phi_T^{(k)}$ for a given tree T with n leaves involves a summation of just $O(n^2)$ terms, so could be calculated fairly easily.

Notice also that in the special case where $k = 1$, Theorem 2(i) simplifies (via Eqn. (1)) as follows.

Corollary 1

$$\Phi_T^{(1)} = \frac{1}{3}\left(1 - \frac{\mu_T}{n-3}\right).$$

Exact values of $\Phi_T^{(k)}$ for classes of trees

Proposition 1

(i) Let T_n denote a caterpillar tree with n leaves. Then $\Phi_{T_4} = 0$ and for $n \geqslant 5$:

$$\Phi_{T_n}^{(k)} = \frac{4}{3(2n-3)(n-2)(n-3)^k} \sum_{j=1}^{n-4} \left(j^{k+1} + j^k\right)$$

Thus, for fixed k we have:

$$\Phi_{T_n}^{(k)} = \frac{2}{3(k+2)} + o(1),$$

where $o(1)$ is a term that tends to 0 as n grows.

(ii) Let T^h denote any tree in $B(2^h+1)$ obtained by attaching a leaf to the root of a complete balanced binary tree of height h. Then

$$\varphi_{T^{(h)}}(i) = P(i,h) + Q(i,h), \tag{6}$$

where:

$$P(i,h) = \begin{cases} 2^{i+2}(2^{h-i}-1), & \text{if } i < h; \\ 0, & \text{if } i \geqslant h \end{cases}$$

and

$$Q(i,h) = 2^i \cdot \sum_{j=1}^{h-1} 2^{h-j-1} S(i,j),$$

where

$$S(i,j) = \begin{cases} i-1, & \text{for } 1 \leqslant i \leqslant j+1; \\ 2j-(i-1), & \text{for } j+1 < i \leqslant 2j; \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

(iii) Let $\mathcal{D}$ be a probability distribution on $B(n)$. Then

$$\mathbb{E}_{\mathcal{D}}[\Phi_T^{(k)}] = \frac{1}{3} \cdot \sum_{i=1}^{n-3} \mathbb{E}_{\mathcal{D}}[\tilde{\varphi}_T(i)] \left(1 - \frac{i}{(n-3)}\right)^k. \tag{8}$$

In particular, when $\mathcal{D}$ is the PDA distribution on $B(n)$

$$\mathbb{E}_{PDA}[\tilde{\varphi}_T(i)] = \frac{(i+3)2^i(2n-i-4)!(n-2)!}{(2n-4)!(n-i-3)!\binom{2n-3}{2}}. \tag{9}$$

Proof: Part (i): For $n \geqslant 4$ any path of length ℓ in T_n that lies strictly between two edges $\{e, e'\}$ is a path of interior edges, and there are precisely four such pairs of edges that correspond to the same path. Moreover, the number of interior edge paths of T_n of length $n-3-j$ is $j+1$ for each j between 0 and $n-4$. Thus,

$$\Phi^{(k)}_{T_n} = \frac{1}{3\binom{2n-3}{2}} \sum_{j=0}^{n-4} 4(j+1)\left(\frac{j}{n-3}\right)^k,$$

and straightforward algebraic manipulation leads to the stated expression. The last part of (i) follows from the asymptotic identity: $\sum_{j=1}^{n} j^k \sim \frac{n^{k+1}}{k}$.

Part (ii)

For $h \geqslant 1$, let U_i^h be the set of edges e of T^h, excluding the stem edge, for which there are exactly $i \geqslant 0$ edges strictly between e and the stem edge. If we also let $u^{(h)}(i) = |U_i^h|$ then it is easily seen that:

$$u_i^{(h)} = \begin{cases} 2^{i+1}, & \text{for } 0 \leqslant i \leqslant h-1; \\ 0, & \text{for } i \geqslant h. \end{cases} \tag{10}$$

For $h \geqslant 1$ and $i \geqslant 1$, let N_i^h denote collection of pairs of edges in T^h that are separated by exactly i edges, and let $n_i^{(h)} = |N_i^h|$. Thus, $n_i^{(h)} = \varphi_{T^{(h)}}(i)$. Observe that:

$$n_i^{(h)} = 0 \Leftrightarrow i > 2h-2.$$

Notice also that deleting the stem edge of T^{h+1} (and its incident vertices) produces two copies of T^h - we will call these L and R ('left' and 'right').

The set N_i^{h+1} can be partitioned into three classes:

Class 1: A pair of edges consisting of an edge of T^{h+1} at distance i from the stem edge, together with the stem edge. This set has size $u_i^{(h+1)}$, where $u_i^{(h)}$ is as given above.

Class 2: A pair of edges that lie entirely within L or entirely within R. This set has size $2 \cdot n_i^{(h)}$.

Class 3: A pair of edges $\{e_1, e_2\}$ with e_1 in L and e_2 in R with the distance between e_1 and e_2 being i.

A little care is required to enumerate Class 3. One subcase is that e_1 is not the stem edge of L and e_2 is not the stem edge of R. In that case, the number of choices is

$$\sum_{(l,r)\geqslant(0,0):r+l=i-2,r,l\leqslant h-1} u_r^{(h)}u_l^{(h)},$$

which equals the expression $Q(i,h)$ given in Part (ii). The complementary subcase where e_1 or e_2 is the stem edge of L or R contributes

$$u_i^{(h+1)}.$$

Combining the above gives the recursion:

$$n_i^{(h+1)} = u_i^{(h+1)} + 2n_i^{(h)} + [Q(i,h) + u_i^{(h+1)}]$$

which simplifies to:

$$n_i^{(h+1)} = 2n_i^{(h)} + 2u_i^{(h+1)} + Q(i,h).$$

or equivalently,

$$n_i^{(h)} = 2n_i^{(h-1)} + 2u_i^{(h)} + Q(i,h-1).$$

Solving this recursion (noting that $u_i^{(h)} = 0$ for $i \geqslant h$) leads to the expression for $\varphi_{T^{(h)}}$ given in Part (ii).

Part (iii): Eqn. (8) follows by linearity of expectation (and interchanging the order of expectation operators).

The expression for $\mathbb{E}[\tilde{\varphi}_T(i)]$ involves an argument in enumerative combinatorics. Let $N(n,k)$ denote the number of (unordered) forests consisting of k rooted binary trees whose leaf sets are disjoint and contain a total of n leaves (we allow a single labeled leaf to be a rooted tree of size 1, otherwise the root of each tree has degree 2). It is easily seen that $N(n,2)$ is precisely the number of rooted binary trees $(2n-3)!!$, since deleting the root of such a tree, produces a forest of two trees with disjoint leaf sets. It turns out there is an exact formula for $N(n,r)$ (from ∼1990):

$$N(n,r) = \frac{(2n-r-1)!}{(n-r)!(r-1)!2^{n-r}},$$

for $r = 1, \ldots, n$ (see e.g. (3), p. 105).

An *ordered forest* is a forest with a linear ordering on its component trees. If $O(n, r)$ denotes the number of ordered forests consisting of r trees then

$$O(n,r) = r!N(n,r) = r\frac{(2n-r-1)!}{(n-r)!2^{n-r}}.$$

Notice that:

$$\frac{1}{2}\sum_{r=4}^{n} O(n,r) = (2n-5)!! \times \left[\binom{2n-3}{2} - 3(n-2)\right].$$

To see this observe that twice the right-hand side counts the number of pairs $(T, (e_1, e_2))$ where $T \in B(n)$ and (e_1, e_2) is an ordered pair of non-adjacent edges. Notice that if we delete the path connecting e_1 and e_2 we obtain an ordered forest of rooted trees on the same leaf set as T (*cf.* Fig. 1) . This provides a bijection between these two sets in which the number of strictly edges between e_1 and e_2 in T is equal to $i - 3$ where i is the number of forests.

Thus,

$$\varphi_T(i-3) = \frac{1}{2} \times \frac{O(n,i)}{(2n-5)!!} = \frac{i2^{i-3}(2n-i-1)!(n-2)!}{(2n-4)!(n-i)!},$$

and rearranging, gives the result.

□

To illustrate Part (i), with $n = 5$ we have $\Phi_{T_5} = \frac{8}{63}(\frac{1}{2})^k$. Notice that this converges to zero exponentially fast with k (and in general for fixed n this will be the case).

Also, observe that for general values of n and with $k = 1$ (say) we have:

$$\Phi_T^{(k)} = \frac{2(n-4)}{3(2n-3)(n-2)}\left(1 + \frac{2n-7}{3}\right).$$

Thus, for the simulation involving trees with $n = 500$ leaves, if these were caterpillars (instead of YH trees), for $k = 1$ we would expect $\Phi_T^{(k)}$ to be very close to $\frac{2}{9}$ (in agreement with the asymptotic claim in the Part (i) of Proposition 1) which is lower than the simulated values on YH trees, as expected.

To illustrate Part (ii), $\varphi_{T^3}(2) = P(2,3) + Q(2,3) = 2^4(2^1-1) + 2^2(2+1) = 28$. Notice also that for general h we obtain (as expected)

$$\varphi_{T^h}(2h-2) = 2^{2h-2},$$

and $\varphi_{T^h}(i) = 0$ for $i > 2h-2$.

Notice that for h large, $P(i,h)$ is negligible relative to $Q(i,h)$.

A lower bound on the expected value of $\Phi_T^{(k)}$ for the PDA and YH distributions

Proposition 2 Let T be a tree in $B(n)$ sampled from the PDA or YH distribution. Then, for all $n \geqslant 4$ we have:

(i)

$$\mathbb{E}_{PDA}[\mu_T] = \frac{2}{(n-2)} \cdot \left[\frac{2^{n-2}n!}{(2n-3)!!} - 2n + 2\right].$$

Moreover,

$$\mathbb{E}_{PDA}[\mu_T] = \sqrt{\pi} \cdot \sqrt{n} + o(1).$$

(ii)

$$\mathbb{E}_{PDA}[\Phi_T^{(k)}] \geqslant \frac{1}{3}e^{-k\sqrt{\pi}/\sqrt{n}} - o(1),$$

where $o(1)$ is a term that tends to 0 as n grows.

(iii)

$$\mathbb{E}_{YH}[\Phi_T^{(k)}] \geqslant \frac{1}{3}e^{-ck\log(n)/n} - o(1),$$

where $o(1)$ is a term that tends to 0 as n grows, and c is a constant (independent of n).

Proof: Part (i): For $T \in B(n)$ let $NA(T)$ be the collection of (unordered) pairs $\{e_1, e_2\}$ of non-adjacent edges of T.

The first step is to apply a classic technique in enumerative combinatorics; namely, we count a certain set Ω in two different ways, to obtain an equation. Here, we take the set

Ω to be the collection of triples $(T, \sigma, \{e_1, e_2\})$ where $T \in B(n)$, σ is a split of $[n]$ that corresponds to an interior edge of T, and $\{e_1, e_2\} \in NA(T)$ with the edge of T corresponding to split σ being strictly within the path connecting e_1 and e_2.

Summing first over all choices of T we have:

$$|\Omega| = \sum_{T \in B(n)} \sum_{\{e_1,e_2\} \in NA(T)} d_T(e_1, e_2). \tag{11}$$

and so,

$$\mathbb{E}_{PDA}[\mathbb{E}_{\mathcal{C}_T}[R_T]] = \frac{|\Omega|}{|B(n)| \cdot |\mathcal{C}_T|}, \tag{12}$$

where $|\mathcal{C}_T| = \binom{2n-3}{2} - 3(n-2)$ is the number of pairs of non-adjacent edges in T.

On the other hand, we can count $|\Omega|$ by first summing over all choices of the split σ (stratified by the size k of the smaller half of the split) and then counting for each such σ the number of T and $\{e_1, e_2\}$. This gives:

$$|\Omega| = \frac{1}{2} \sum_{k=1}^{n} \binom{n}{k} [rb(k) \cdot rb(n-k)] \cdot [(2k-2) \cdot (2(n-k)-2)]. \tag{13}$$

where $rb(m) = (2m-3)!!$ is the number of rooted trees on a leaf set of size m.

To see this, note that $\binom{n}{k}$ is the number of ways to partition $[n]$ into a split σ of two sets of size k and $n-k$, and $\frac{1}{2}rb(k)rb(n-k)$ is the number of trees in $B(n)$ that contain this split (the factor $\frac{1}{2}$ is to remedy the double counting that occurs). Finally, given T and σ there are now $(2k-2)(2(n-k)-2)$ choices of $\{e_1, e_2\}$ where e_1 is an edge of the rooted subtree of T on the leaf set of size k and e_2 is an edge of the rooted subtree of T of size $n-k$.

We will apply generating function techniques to calculate an exact expression for term on the right hand side of Eqn. (13). Notice can rewrite Eqn. (13) as follows:

$$|\Omega| = 2n![x^n] \sum_{k=1}^{n} ((k-1)\frac{rb(k)}{k!} x^k \cdot (n-k-1)\frac{rb(n-k)}{(n-k)!} x^{n-k}$$

where $[x^n]f(x)$ denote the coefficient of x^n in $f(x)$, and hence, even more compactly,

$$|\Omega| = 2n![x^n]F(x)^2 \tag{14}$$

where

$$F(x) := \sum_{i \geqslant 1} (i-1)\frac{rb(i)}{i!}x^i. \tag{15}$$

Let $rb(n) = (2n-3)!!$ (the number of rooted binary trees on a leaf set of size n) and let $\nu(x) = \sum_{n \geqslant 1} \frac{rb(n)}{n!}x^n$ denote the associated exponential generating function. It is well known (see e.g. (3)) that

$$\nu(x) = \frac{1}{2}\nu^2(x) + x \tag{16}$$

which has the unique solution

$$\nu(x) = 1 - \sqrt{1-2x}. \tag{17}$$

We will use the following relationships which follow easily from Eqn. (16):

$$\nu^2(x) = 2\nu(x) - 2x, \tag{18}$$

$$\nu(x)\nu'(x) = \nu'(x) - 1, \tag{19}$$

where $\nu'(x) = \frac{d}{dx}\nu(x)$.

Recall, $F(x)$ from Eqn. (15). We have:

$$F(x) = x\nu'(x) - \nu(x).$$

Thus,

$$F(x)^2 = x^2\nu'(x)^2 - 2x\nu(x)\nu'(x) + \nu^2(x),$$

and this simplifies by Eqns. (18) and (19)) to:

$$F(x)^2 = x^2\nu'(x)^2 - 2x\nu'(x) + 2\nu(x). \tag{20}$$

In order to determine the expression in Eqn. (14) we consider the coefficient of x^n in $F(x)^2$ (i.e. $[x^n]F(x)^2$) by considering the three terms on the right of Eqn. (20). For the first term of Eqn. (20), Eqn. (17)) gives: $\nu'(x)^2 = \frac{1}{1-2x}$ and so

$$[x^n]x^2\nu'(x)^2 = [x^{n-2}]\nu'(x)^2 = [x^{n-2}](1-2x)^{-1} = 2^{n-2}. \tag{21}$$

For the second term of Eqn. (20), we have:

$$[x^n]2x\nu'(x) = 2[x^{n-1}]\nu'(x) = 2\frac{n \cdot rb(n)}{n!}. \tag{22}$$

For the third term of Eqn. (20), we have:

$$[x^n]2\nu(x) = 2[x^n]\nu(x) = 2\frac{rb(n)}{n!}. \tag{23}$$

Substituting the expressions on the right hand side of Eqns. (18), (19) and (20) into the corresponding terms in Eqn (20) allows us now to determine the expression for $|\Omega|$ given by Eqn. (14), as follows:

$$|\Omega| = 2n! \left[2^{n-2} - 2\frac{n \cdot rb(n)}{n!} + 2\frac{rb(n)}{n!}\right]$$

which simplifies slightly to:

$$|\Omega| = 2\left[2^{n-2}n! - 2(n-1)rb(n)\right]. \tag{24}$$

By substituting the expression for $|\Omega|$ given by Eqn. (24) into Eqn. (12) and rearranging terms, gives the claimed expression for $\mathbb{E}_{PDA}[\mathbb{E}_{\mathcal{C}_T}[R_T]]$.

The second claim in Part (i) of Theorem 2 follows from the asymptotic equivalence:

$$\frac{rb(n)}{n!} \sim \frac{1}{2\sqrt{\pi}} 2^n n^{-3/2}$$

together with some standard algebraic manipulation.

Part (ii): We apply Theorem 2(Part (ii)) to Part (i) of the current theorem. WIth $k \leqslant \gamma\sqrt{n}$ we obtain:

$$\mathbb{E}_{PDA}[\Phi_T^{(k)}] \geqslant \frac{1}{3}\left(1 - \frac{\sqrt{\pi}}{\sqrt{n}}\right)^{\gamma\sqrt{n}} \sim \frac{1}{3}\exp(-\gamma\sqrt{\pi}).$$

Part (iii): We apply Proposition 4 of (1) (where $\beta = 0$ corresponds to the YH distribution). This result shows that in a tree $T \in B(n)$ sampled according to the YH distribution the maximum distance D from any leaf to the root is less than

$(4.31 + \epsilon)\log(n)$ with a probability converging to 1 as n grows. Now μ_T is always less than twice the distance from any leaf in T to any other leaf in T, and this inter-leaf distances is bounded by $2D$. The result now follows from Theorem 2(Part (ii)). □

Remarks:

Note that Parts (ii) and (iii) imply that for each fixed k, $\mathbb{E}_{PDA}[\Phi_T^{(k)}]$ and $\mathbb{E}_{YH}[\Phi_T^{(k)}]$ both converge to $\frac{1}{3}$ as n grows, in contrast to the caterpillar case of Part (i) of Proposition 1, where the limit is smaller (e.g. for $k = 1$, the limit is $\frac{2}{9}$).

Notice also that the lower bounds on $\mathbb{E}_{\mathcal{D}}[\Phi_T^{(k)}]$ allow k to grow faster with n for YH trees than PDA (essentially because in a Yule tree, there are on average fewer edges on paths between leaves, so fewer 'red' edges on the path between e_1, e_2 (see Fig. 1). In this lower bound we are undercounting by ignoring the red portions to the left of e_1 and right of e_2 by considering only cases where e_1 and e_2 are both pendant edges (this undercounting is valid since we are stating a lower bound on $\mathbb{E}_{\mathcal{D}}[\Phi_T^{(k)}]$).

The expected number of such false splits

Note that Theorem 2 describes the probability that a single character satisfies the three conditions (C-i)–(C-iii) above, conditional on this character having evolved on T with 2 edge changes (thus $\Phi_T^{(k)}$ should be viewed as a conditional probability). This raises two further natural question: for a given character, what is the probability p_T that such a character will evolve so as to satisfy conditions (Ci)–(Ciii)? And how many such characters should we expect to see? For these questions, the number k of perfectly evolved characters should now be treated as a random variable, K. Thus, let K be the random variable corresponding to the number of distinct (and nontrivial) splits generated from those characters (from within the m characters in total) that have perfectly evolved on T. The variable K lies between 0 and m. Let μ_K be the expected value of K.

Conditional on $K = k$, the value p_T is simply given (ignoring terms involving λ of higher than quadratic power) by:

$$p_T = \left[\binom{2n-3}{2} - 3(n-2)\right]\lambda^2\Phi_T^{(k)}. \tag{25}$$

In particular,

$$p_T \sim 2n^2\lambda^2\Phi_T^{(k)}. \tag{26}$$

We have

$$\mu_K = (n-3)\left(1-\left(1-(\lambda+O(\lambda^2))\right)^m\right).$$

Provided that $\lambda m << 1$ we may approximate this last equation by:

$$\mu_K \approx (n-3)\lambda m.$$

Thus, provided that m large and $\lambda m < 1$ and using using μ_K to estimate K (which is reasonable since m is large, and we are also assuming above that $\lambda n << 1$) we obtain:

$$K \approx (n-3)\lambda m$$

and solving for λ in this equation, and substituting the estimate of λ into Eqn. (26) gives:

$$p_T \sim \frac{2K^2}{m^2}\Phi_T^{(K)}.$$

Let us now multiply through by m (the number of characters) to get.

$$mp_T \sim \frac{2K^2}{m}\Phi_T^{(K)}. \tag{27}$$

The term on the left (mp_T) has a natural interpretation - it is simply the expected number of characters that have evolved on T by 2 edge changes and that also satisfy the three conditions (C-i)–(C-iii) above. By Equation (27), this can be approximated by $\frac{2K^2}{m}\Phi_T^{(K)}$. Note that this quantity is independent of λ (provided this is sufficiently small that the above approximations are reasonable), and n also is not involved in the first term in the product.

Note that one cannot interpret mp_T as the expected number of false splits in a maximum compatibility tree involving the K perfectly evolved characters and the additional characters with 2 changes. This is because some of these latter characters may be incompatible with each other (even though they are compatible with the K perfectly evolved characters).

The case where $\ell > 1$ binary characters evolve with 2 edge changes on T

So far we have considered the impact of a single binary character that has evolved on T with 2 edge changes. What happens if there is more than one such character? In particular, when will two such characters (which induce false splits) be compatible with each other? The following result provides a concise characterisation. This result is also relevant to case for more than two such characters, since a collection of binary characters is compatible if and only if every pair of characters is compatible.

Proposition 3 Suppose that f and f' are two binary characters that have evolved on T, each with with 2 edge changes (say e_1, e_2 for f and e'_1 and e'_2 for f'). Let P (respectively, P') denote the set of edges in the path in T consisting of e_1, e_2 (respectively e'_1 and e'_2) and the edges on the path between them. Then f and f' are compatible if and only if either one of the following two conditions holds:

(c_1) $P \subseteq P'$ or $P' \subseteq P$.

(c_2) e does not lie in P' and e' does not lie in P.

Proof: The proof involves a case analysis. In particular, we show that:

(1) if (c_1) holds then f and f' are compatible;

(2) if (c_1) fails but (c_2) holds, then f and f' are compatible; and

(3) if (c_1) and (c_2) both fail to hold, then f and f' are incompatible.

To simplify notation we first introduce some conventions. We will let σ (respectively σ') denote the split of $[n]$ induced by (reversing state) changes on e_1 and e_2 (respectively on e_1' and e_2'). For subsets $V_1, \ldots, V_r$ of $[n]$ we write $V_1 \cdots V_r|-$ as shorthand for the split $(V_1 \cup \cdots \cup V_r)|([n] - (V_1 \cup \cdots V_r))$, and if a subtree within T has leaf set $A \subset [n]$ we will denote this subtree by writing $t(A)$. We will also assume that e_1, e_2, e_1', e_2' are four distinct edges (we deal with the case where this fails at the end).

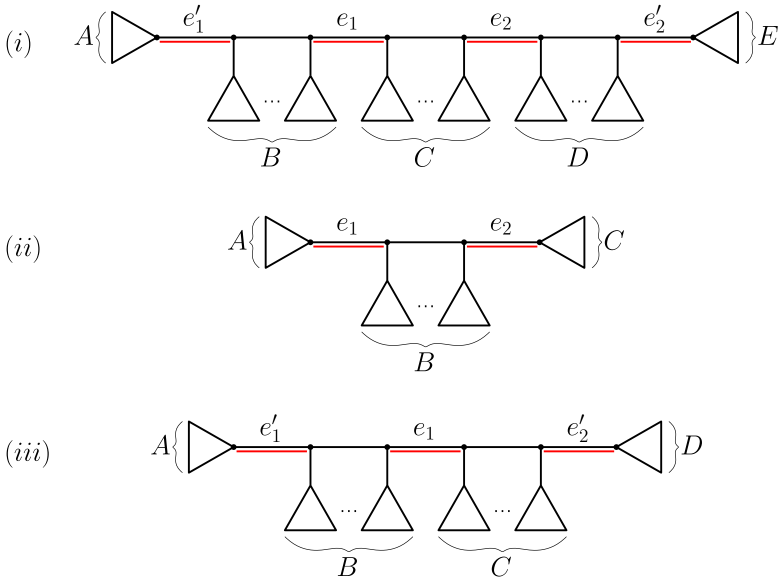

Fig. 2. The cases for the proof of Proposition 3

Case (1): Suppose that (c_1) holds. Without loss of generality we may assume that $P \subseteq P'$ and that the order of the path from e_1' to e_2' passes through e_1 and e_2 in this order. Thus we can represent T as in Fig. 2(i) where A, B, C, D, E denote the (unions of the) leaf sets of the corresponding subtrees determined by this arrangement of the four edges.

Thus σ is the split $ABDE|C$, and σ' is the split $AE|BCD$. Since the second half of he first split (namely C) is a subset of the second half of the second split (BCD) these two splits are compatible. Note that this argument also holds if one or both of the following identities applies: $e_1 = e_1'$ or $e_2 = e_2'$.

Case (2): Now suppose that (c_1) fails but (c_2) holds. In this case, by considering the path between e_1 and e_2 we can represent T as in Fig. 2(ii) and e_1' and e_2' either both lie the subtree $t(A)$ or $t(C)$ or else e_1' lies in $t(B_i)$ and e_2' lies in $t(B_j)$ for some i, j (we allow $i = j$)

First suppose that e_1', e_2' both lie within $t(A)$. Then $\sigma' = A'|-$ for a subset of A' of A and since $\sigma = AC|B$ these two splits are compatible. A similar argument holds if e_1', e_2' both lie within $t(C)$. Alternatively, suppose that e_1' is an edge in $t(B_i)$ and e_2' is an edge of $t(B_j)$ (we allow $i = j$). Then $\sigma' = B'|-$ for a subset of B, and since $\sigma = AC|-$ these two splits are also compatible (since $B' \cap (A \cup C) = \emptyset$).

Case (3): Finally suppose that Case (c_1) and Case (c_2) both fail. Without loss of generality we may assume that e_1 lies within P'. In this case we can represent tree T as shown in Fig. 2(iii). Thus $\sigma' = (A \cup D)|-$.

By the assumption that Case (c_1) and Case (c_2) both fail, it follows that e_2 lies in one of the following trees $t(A), t(D), t(B_i)$, or $t(C_j)$, for some i, j. By symmetry, there are just two sub-cases to consider: (i) e_2 lies in $t(A)$ or (ii) e_2 lies in $t(B_i)$. In subcase (i) $\sigma = CDA'|-$ where A' is a proper subset of A, and C is the union of the leaf sets in $C_1, \ldots C_s$. Thus $\sigma = CDA'|-$ is incompatible with $\sigma' = AD|-$ since neither of the sets on the left of the split contains the other.

In subcase (ii), $\sigma = B'CD|-$ for a proper subset B' of B (the strict containment is because e_1 and e_2 are not adjacent) and so again σ and σ' are incompatible.

Finally, we assumed that e_1, e_2, e_1', e_2' are four distinct edges. Otherwise (since $e_1 \neq e_2$ and $e_1' \neq e_2'$) we either have: $\{e_1, e_2\} = \{e_1', e_2'\}$, in which case condition (c_1) holds, and the two characters f and f' induce the same split and so are compatible, or we may assume without loss of generality that $e_1 = e_1'$ and $e_2 \neq e_2'$. A similar (though simpler) case analysis to the above leads to the same conclusions as before.

□

Remark: A consequence of Proposition 3 and our earlier results is that if n is large in comparison to k then a small number (ℓ) of 2-state characters evolved randomly on T are likely to be (i) compatible with each other, (ii) compatible with the k perfectly evolved characters, and (iii) not be splits of T, and thus show up as false splits on the reconstructed (max compatabiilty or max parsimony) tree. To see this, observe first that the probability that (c_2) occurs tends to 1 as $n \to \infty$ for fixed k (this holds for any choice of T, or a randomly sampled tree from the PDA or YH distribution†. Second, a collection of binary characters is compatible if and only if each pair of them is compatible, and so if ℓ and k are fixed (or grow sufficiently slowly with n, depending on the tree shape) the probability that all ℓ false splits of T are present in the MP or MC tree (along with the splits corresponding to the k perfectly compatible characters of T) tends to 1 as $n \to \infty$.

Estimating the false positive rate

Given a binary phylogenetic tree T, and m characters evolved randomly on T by the model described earlier, the *false positive rate* (FP_T) is the expected value of the proportion of non-trivial splits in the reconstructed tree (using MC, say) that are not in T (here we assume that if the reconstructed tree is a star, this proportion (which is technically 0/0) is zero). Recall that ξ is the expected number of state changes in the tree T per character, under the model described earlier. FP_T is a function of the three parameters T (specifically, its shape and number of leaves), m and λ (equivalently, FP_T is a function of T, m and ξ).

In general, it is mathematically complicated to describe FP_T in terms of these parameters. However, when the number of leaves in a tree grows faster than the number of perfectly compatible characters, it is possible to state a limit result, in order to provide an approximation to FP_T for large trees.

In the following theorem we consider the following setting:

I. $m\xi = \Theta(n^\beta)$ for some $0 < \beta < \frac{1}{2}$, and

II. $m\xi^2 = O(1)$,

where $O(1)$ refers to dependence on n (thus $m\xi^2$ is not growing with n). Note that Condition (I) implies that the number of perfectly-evolved characters grows with the number of leaves, but at a rate that is slower than linearly. When $\beta \geqslant \frac{1}{2}$, Condition (I) also provides a positive probability that more than one perfectly-evolved character will give rise to the same non-trivial split. Conditions (I) and (II) imply that ξ decreases as n increases.

We will show that in this setting, the false positive rate is (asymptotically) of the form $\frac{\xi}{3}$ times a function Ω that involves T (via its shape), m and ξ. To describe this result, we need to define this function Ω. Let

$$\Omega(T_n, \xi, m) = \sum_{i=1}^{n-4} \tilde{\varphi}_{T_n}(i) \cdot \frac{e^{-i\mu/(n-3)} - e^{-\mu}}{1 - i/(n-3)},$$

where

$$\mu = \frac{1}{2} m\xi$$

and where $\tilde{\varphi}_{T_n}(i)$ is given in Eqn. (2). For example, for any caterpillar tree, we have $\tilde{\varphi}_{T_n}(i) = 4(n-2-i)/\binom{2n-3}{2})$.

Notice that $\Omega(T_n, \xi, m)$ depends on T_n only via the coefficients $\tilde{\varphi}_{T_n}(i)$, and this dependence in linear. Thus, if $\mathcal{D}$ is a distribution on trees (e.g. the PDA or YH) then the expected value of $\Omega(T_n, \xi, m)$ is given by:

$$\mathbb{E}_{\mathcal{D}}[\Omega(T_n, \xi, m)] = \sum_{i=1}^{n-4} \mathbb{E}_{\mathcal{D}}[\tilde{\varphi}_{T_n}(i)] \cdot \frac{e^{-i\mu/(n-3)} - e^{-\mu}}{1 - i/(n-3)}. \qquad (28)$$

For the PDA distribution, the term $\mathbb{E}_{PDA}[\tilde{\varphi}_T(i)]$ has an explicit exact value, given by Eqn. (9).

Theorem 3 For each $n \geqslant 1$, let T_n be a binary phylogenetic tree with n leaves, and suppose that Conditions (I) and (II) hold.

(i)

$$FP_{T_n} = \frac{\xi}{3} \cdot \Omega(T_n, \xi, m)(1 + o(1))$$

where $o(1)$ is a term that tends to 0 as n grows.

(ii) If T_n is sampled from a distribution $\mathcal{D}$ (e.g. PDA, YH) then the expected value of FP_{T_n}, denoted $\mathbb{E}_{\mathcal{D}}[FP_{T_n}]$, satisfies

$$\mathbb{E}_{\mathcal{D}}[FP_{T_n}] = \frac{\xi}{3} \cdot \mathbb{E}_{\mathcal{D}}\left[\Omega(T_n, \xi, m)\right](1 + o(1)).$$

(iii) If T_n is a caterpillar tree, then $FP_{T_n} = 4(1+o(1))/(3m)$.

Proof: Part (i): For convenience, we will write T in place of T_n The number X_1 of perfectly evolved characters on T has a binomial distribution (m trials, with the probability of success of $\frac{\xi}{2}(1+o(1))$ (since the proportion of interior edge in T is $(n-3)/(2n-3) \sim \frac{1}{2}$) and so has expected value $\frac{1}{2}m\xi(1+o(1))$). As before, let K denote the number of distinct non-trivial splits of T_n that correspond to (one or more) of these perfectly evolved characters. By Condition (I) it follows that each perfectly evolved character is (with probability $\rightarrow 1$) generated at most once, and so K is approximated by X_1, which in turn is approximated (for large n) by a Poisson distribution with mean $\mu = \frac{1}{2}m\xi$.

Consider now the splits that arise from 2-edge-change characters on T. Denote the number of false splits by X_2, and the number of true splits by X_2'. The expected value of X_2' is of order $m\xi^2/n$ and so it converges to zero as n grows, by Condition (II). By Proposition 3 and Condition (II), the probability that every pair of the X_2 false splits is compatible (with each other) tends to 1 as n grows.

Next, consider splits (true or false) that arise from characters involving 3 or more changes on different edges. If X_3 denotes the number of such characters, then the expected value of X_3 is bounded above by a constant (independent of n) times $m\xi^3$, and by Conditions (I) and (II), the ratio of this to K is of order ξ^2 (independent of n).

Thus,

$$FP_T = \frac{1}{3}\cdot\frac{\xi^2 m}{2}(1+o(1))\cdot\sum_{i=1}^{n-4}\tilde{\varphi}_T(i)\sum_{k=0}^{n-3}\frac{\left(1-\frac{i}{n-3}\right)^k}{k+O(1)}\mathbb{P}(K=k), \tag{29}$$

where K has a Poisson distribution with expected value $\mu = \frac{m\xi}{2}$, and $O(1)$ refers to a term that is bounded in n (by Condition (II)) and accounts for any non-trivial splits induced by characters that are not perfectly evolved and in the reconstructed tree (as well as for splits from perfectly evolved characters that are 'lost' in a strict consensus by being the only such split in a path between the two edges of a 2-change character), while $o(1)$ refers to a term the tends to zero as n grows due to characters that involve 3 or more edge changes).

Thus, $\mathbb{P}(K=k) = e^{-\mu}\mu^k/k!$, and so, under Conditions (I) and (II) we have:

$$\sum_{k=0}^{n-3} \frac{\left(1-\frac{i}{n-3}\right)^k}{k+O(1)}\mathbb{P}(K=k) \sim e^{-\mu}\sum_{k=0}^{\infty}\frac{(\rho\mu)^k}{k!(k+1)},$$

where $\rho := 1 - \frac{i}{n-3}$. We now apply the identity: $\sum_{k=0}^{\infty}\frac{x^k}{k!(k+1)} = \frac{e^x-1}{x}$, with $x = \rho\mu$ to obtain:

$$e^{-\mu}\sum_{k=0}^{\infty}\frac{(\rho\mu)^k}{k!(k+1)} = e^{-\mu}\left(\frac{e^{\rho\mu}-1}{\rho\mu}\right) = \frac{e^{\mu(\rho-1)}-e^{-\mu}}{\rho\mu} = \frac{e^{-i\mu/(n-3)}-e^{-\mu}}{(1-i/(n-3))\mu}. \tag{30}$$

Since $\mu = m\xi/2$, notice that we can write the term $\frac{\xi^2 m}{2}$ in Eqn. (29) as $\xi\mu$ and so, from Eqns. (29) and (30), we have:

$$FP_T \sim \frac{1}{3}\xi\mu \cdot \sum_{i=1}^{n-4}\tilde{\varphi}_T(i)\frac{e^{-i\mu/(n-3)}-e^{-\mu}}{(1-i/(n-3))\mu}. \tag{31}$$

Finally, canceling the term μ in the numerator and denominator on the right-hand-side of Eqn. (31) gives the expression in Part (i).

Part (ii): This follows directly from Part (i) by linearity of expectation.

Part (iii): When T_n is a caterpillar tree we have:

$$\tilde{\varphi}_{T_n}(i) = \frac{4(n-2-i)}{\binom{2n-3}{2}},$$

for $1 \leqslant i \leqslant n-3$. We can rewrite this as: $\tilde{\varphi}_{T_n}(i) = \frac{4}{2n-3}\cdot\left(1-\frac{i}{n-2}\right)$, and substituting this into Part (i) we obtain:

$$FP_{T_n} = \frac{\xi}{3}\cdot\frac{4}{(2n-3)}\sum_{i=1}^{n-4}\left(1-\frac{i}{n-2}\right)\cdot\frac{e^{-i\mu/(n-3)}-e^{-\mu}}{(1-i/(n-3))}\cdot(1+o(1)).$$

As $n\to\infty$ we have the asymptotic identity:

$$FP_{T_n} \sim \frac{\xi}{3}\cdot\frac{4}{(2n-3)}\sum_{i=1}^{n-4}(e^{-i\mu/(n-3)}-e^{-\mu}) \sim \frac{2\xi}{3n}\cdot\left(\left(\sum_{i=1}^{\infty}x_n^i\right)-(n-4)e^{-\mu}\right), \tag{32}$$

where $x := e^{-\mu/(n-3)}$.

Now, x converges to 1 from below as n grows (under Conditions (I) and (II)), and so applying the identity $\sum_{i=1}^{\infty}x^i = x/(1-x)$ (for $0<x<1$), together with the identity

$1 - x \sim \mu/(n-3) \sim \mu/n = m\xi/2n$ (since $\mu = \frac{1}{2}m\xi$) gives:

$$FP_{T_n} \sim \frac{2\xi}{3n} \cdot \frac{2nx_n}{m\xi} - (2\xi/3)e^{-\mu} \sim 4/(3m), \tag{33}$$

noting that the term $(2\xi/3)e^{-\mu}$ is asymptotically negligible relative to the term it follows in Eqn. (33) under Conditions (I) and (II). This completes the proof.

□

Appendix 1: Links between ML and MP

Given $T \in B(n)$ and a data set D consisting of a sequence of m characters. Let n_i be the number of these characters that have parsimony score i on T, for $i \geqslant 0$ (thus n_0 is the number of constant-state characters present in the data, n_1 is the number of characters present that could have perfectly evolved on T etc). We will assume that (i) each character has a unique most-parsimonious representation on T and (ii) no edge is involved in a state change for more than one character (under a most-parsimonious representation on T).

These conditions are reasonable under conditions (1) and (2) described earlier for large n, where $n_0 >> n_1 >> n_2 \ldots$ and where m grows more slowly than n so changes on edges are likely to be 'well-spaced'. We consider a likelihood setting where an edge counted by n_i has probability p_i and an edge not used in any most parsimonious reconstruction has probability ν.

Note that

$$\sum_{i=0}^{k} n_i = m \text{ and } \sum_{i=1}^{k} in_i = ps(T, D),$$

where $ps(T, D)$ is the parsimony score of the tree T for the data. We will assume that $ps(T, D) << m$.

Now, we can write the likelihood function for T as follows:

$$L(T|D) = Q_0^{m-n_0} \cdot Q_1^m \cdot q_1^{n_1} q_2^{2n_2} \cdots q_k^{kn_k}, \tag{34}$$

where $Q_0 = \nu^{2n-3-ps(T,D)}$, $Q_1 = \prod_{i=1}^{k}(1-p_i)^{in_i}$, $q_i = \frac{p_i}{1-p_i}$. Clearly, $L(T|D)$ is maximised by setting $\nu = 1$ (regardless of the other parameters). Moreover, the log likelihood critical values for p_i obtained by solving the equations:

$$\frac{\partial ln(L)}{\partial p_i} = 0,$$

gives $p_i = \frac{1}{m-n_0}$ for all $i \geqslant 1$; in particular, the optimal p_i values are all equal, and applying Eqn. (34) shows that $L(T|D)$ is then a monotone decreasing function of $ps(T, D)$. In this case, as noted in (2) (p. 353), the MP tree(s) maximize this optimal likelihood score.

Appendix 2: The probability a false split does **not** appear

Now consider any given binary tree T, a character f that has evolved on T by 2 edge changes on e_1, e_2, together with k perfectly evolved characters on T (with all the characters independently evolved under the Jukes-Cantor model with p_e constant across the edges of T).

Let $\Psi_T^{(k)}(e_1, e_2)$ be the probability that f satisfies conditions (C-i) and (C-ii) above (i.e. it is binary, and corresponds to a split that is not in T), and this split does **not** occur in either the MP (or MC) tree for the data that consists of f together with k perfectly evolved characters. In the case of ties (in constructing the MP or MC tree) these are broken uniformly.

Proposition 4 Then:

$$\Psi_T^{(k)}(e_1, e_2) = \frac{1}{3}\left(1 - (1 - R_T(e_1, e_2))^k - \frac{1}{2}kR_T(e_1, e_2)(1 - R_T(e_1, e_2))^{k-1}\right).$$

Proof: Let $\mathcal{B}$ be a binomial random variable consisting of k trials, with the probability of success on each trial of $R_T(e_1, e_2) = \frac{d_T(e_1,e_2)}{n-3}$ (as defined earlier). Then $\Psi_T^{(k)}(e_1, e_2) = \frac{1}{3}\left(1 - \mathbb{P}(\mathcal{B} = 0) - \frac{1}{2}\mathbb{P}(\mathcal{B} = 1)\right)$. The factor of 1/2 is to allow for the breaking of a tie when $\mathcal{B} = 1$ between the two MC (or MP) trees. □

Observe that if e_1 and e_2 are adjacent, then $R_T(e_1, e_2) = 0$ and so $\mathbb{P}(\mathcal{B} = 0) = 1$ and thus $\Psi_T^{(k)}(e_1, e_2) = 0$.

Next, let $\Psi_T^{(k)}$ be the expected value of $\Psi_T^{(k)}(e_1, e_2)$ across all pairs of edges $\{e_1, e_2\}$. Thus, $\Psi_T^{(k)}$ is a natural quantity to compare with the earlier $\Phi_T^{(k)}$. We then have:

Corollary 2

$$\Psi_T^{(k)} = \frac{1}{3}\cdot\frac{1}{\binom{2n-3}{2}}\cdot\sum_{\{e_1,e_2\}\in\mathcal{C}_T}\left[1-(1-R_T(e_1,e_2))^k-\frac{1}{2}kR_T(e_1,e_2)(1-R_T(e_1,e_2))^{k-1}\right].$$

Remark: Notice that $\Phi_T^{(k)}$ is a monotone decreasing function of k (and thus $\mathbb{E}_{\mathcal{D}}[\Phi_T^{(k)}]$ is also). Also, since we can write $\Psi_T^{(k)} = \frac{1}{2}(\mathbb{P}(\mathcal{B}\geqslant 1)+\mathbb{P}(\mathcal{B}\geqslant 2))$ it follows that $\Psi_T^{(k)}$ is a monotone increasing function of k (and so $\mathbb{E}_{\mathcal{D}}[\Psi_T^{(k)}]$ is also).

Notice also that

$$\Phi_T^{(k)} + \Psi_T^{(k)} \leqslant \pi_n \tag{35}$$

and

$$\lim_{k\to\infty}(\Phi_T^{(k)} + \Psi_T^{(k)}) = \pi_n, \tag{36}$$

where $\pi_n = \frac{1}{3}\cdot\left(1-\frac{3(n-2)}{\binom{2n-3}{2}}\right)$ is the conditional probability that a character f is binary and that the split induced by this binary character is incompatible with T given that f has evolved on T with 2 edge changes. To see this, observe that there are $\binom{2n-3}{2}-3(n-2)$ pairs of non-adjacent edges in T, all 2 edge change characters have the same probability. Eqns. (35) and (36) also apply if Φ and Ψ are replaced by their expected value over a tree distribution $\mathcal{D}$.

References

Aldous, D. (1995). Probability distributions on cladograms. In: Random Discrete Structures, eds. D. Aldous and R. Pemantle, 1-18. Springer: IMA Volumes in Mathematics and its Applications 76.

Goldman, N. (1990). Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analysis. Syst. Zool. 39(4):345–361.

Steel, M. (2016). Phylogeny: Discrete and random processes in Evolution. SIAM.

Table 1. Full Set of Parameters of 11 Virus Phylogenies Sampled[1]

Abbrev	n	m	ξ	α_e	Res	r_{AC}	r_{AG}	r_{AT}	r_{CG}	r_{CT}	r_{GT}	f_A	f_C	f_G	f_T	α_{ASR}	p_{INV}
DENV	1197	10264	6.68	0.08	0.88	1.15	7.83	1.55	0.80	23.74	1	0.32	0.21	0.26	0.21	0.44	0.29
DENV-1	1067	10264	2.88	0.31	0.82	1.03	7.02	1.31	0.74	18.86	1	0.32	0.21	0.26	0.21	0.38	0.45
EBOV	1610	18164	0.15	0.51	0.36	0.99	7.16	0.62	0.33	8.48	1	0.32	0.21	0.20	0.27	0.29	0.66
EV-D68	824	7293	2.83	0.29	0.80	1.44	31.73	1.88	1.71	46.62	1	0.32	0.20	0.21	0.27	0.26	0.53
HIV	189	1038	0.18	0.45	0.22	1.51	4.24	0.66	0.20	3.39	1	0.39	0.16	0.21	0.23	0.02	0.79
MeV	109	15782	0.44	0.23	0.70	1.39	9.47	0.62	0.40	13.81	1	0.29	0.24	0.23	0.23	0.44	0.55
MuV	458	15154	0.07	0.23	0.30	0.63	3.04	0.16	0.14	3.62	1	0.31	0.22	0.20	0.27	0.07	0.75
RSV	997	14986	0.88	0.29	0.61	1.40	8.61	1.35	0.25	15.78	1	0.39	0.18	0.16	0.28	0.42	0.52
SARS-CoV-2	583	29668	0.02	0.66	0.23	0.44	1.26	0.18	0.39	3.81	1	0.30	0.18	0.20	0.32	0.02	0.90
WNV	2512	10395	2.03	0.57	0.60	1.08	5.22	1.14	0.54	16.05	1	0.27	0.22	0.29	0.22	0.37	0.52
ZIKV	543	10320	0.45	0.33	0.55	0.99	4.02	1.19	0.74	10.76	1	0.27	0.22	0.29	0.21	0.66	0.45

[1] n=number of leaves; m=number of sites; ξ=tree length; Res=Fractional resolution on tree; r_{XY}=rate parameter of the GTR matrix from base X to Y;f_X=frequency of base X; α_e=edge rate variation gamma shape parameter; α_{ASR}=across-site rate variation gamma shape parameter; p_{INV}=fraction of invariant sites

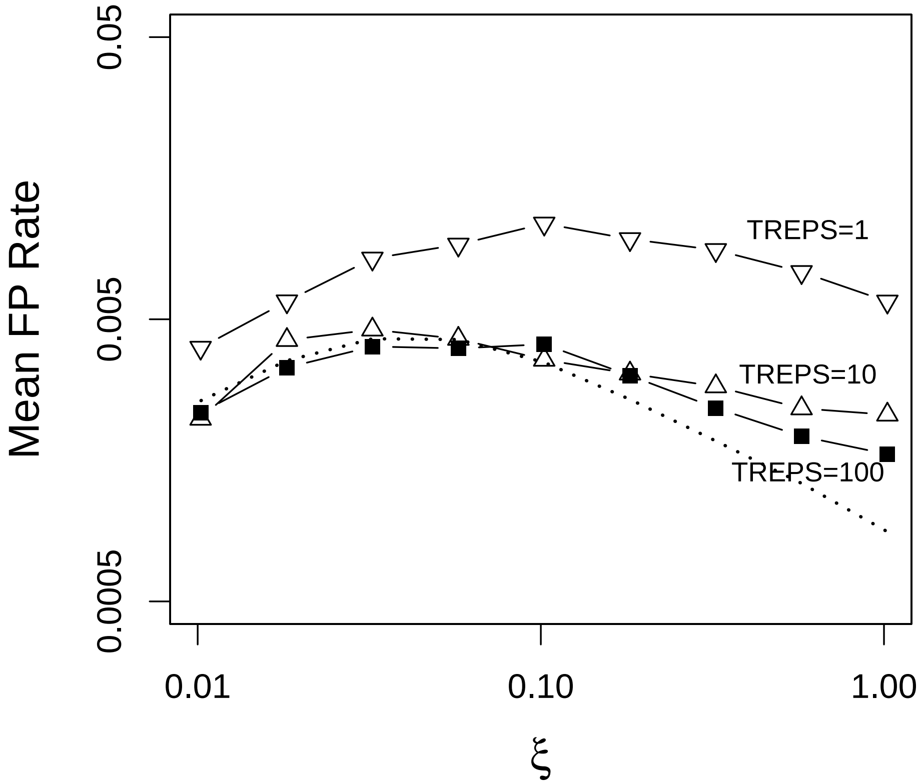

Fig. 3. Mean expected false positive rate in the near-perfect zone of $\xi \leqslant 1$ for three different breadths of MP tree search algorithm for a PDA distribution of trees (1,10, and 100 random addition sub-replicates per simulation replicate; 1000 simulation replicates). Predicted values from Eqn. 2 (main text) is given by dashed line.

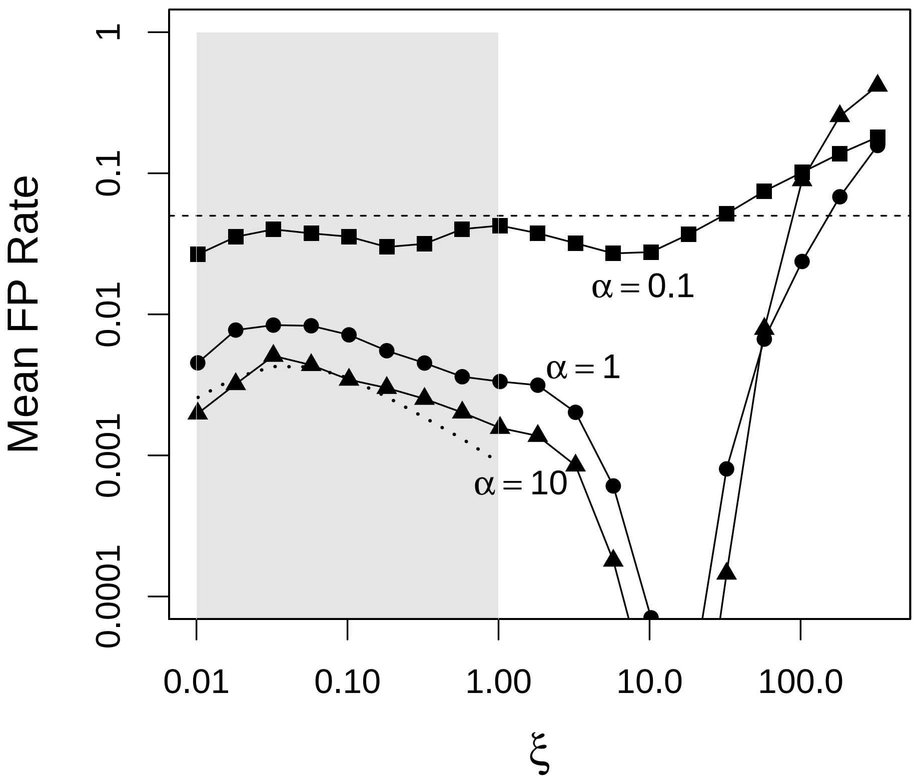

Fig. 4. Effect of across-site-rate variation on expected false positive rate for MP inference for different values of the α_{ASR} shape parameter of the ASR gamma distribution. Smaller α values have higher rate variation. Edge length variation is assumed absent. Dashed curve is the prediction from Eqn. 2 (main text), in which both sources of variation are absent. Tree search algorithm has 100 subreplicates per simulation replicate; 1000 simulation replicates.

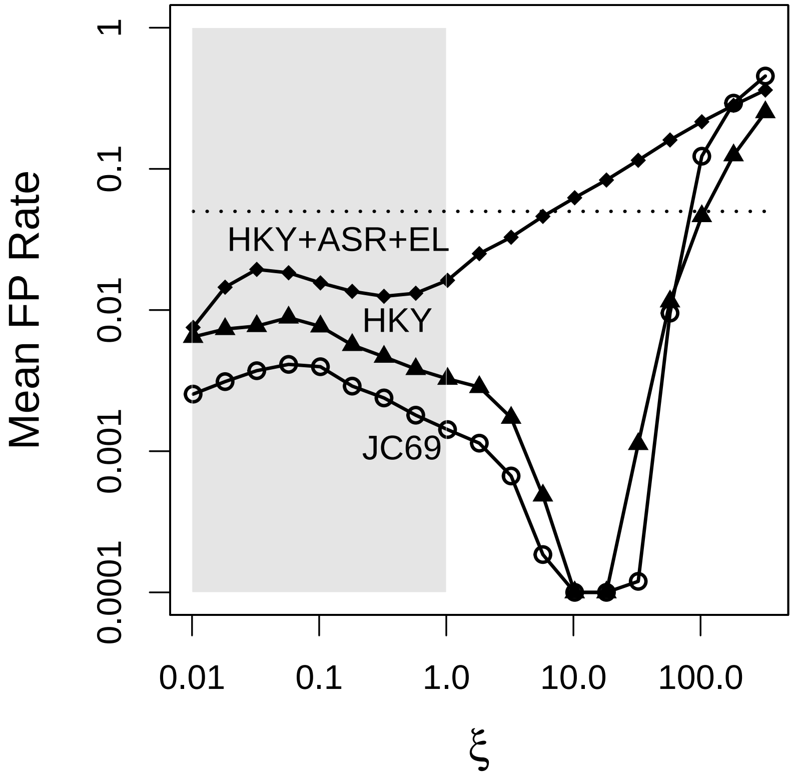

Fig. 5. Effect of model complexity on expected false positive rates for MP. JC69 model (open circles) corresponds to near-perfect assumptions in shaded box ($\xi \leqslant 1$). HKY model (closed triangles) has a transition:transversion ratio of 5 and equal base frequencies. HKY+ASR+EL model (closed diamonds) has a transition:transversion ratio of 5 and also rate variation across sites ($\alpha_{ASR} = 1$) and edges (($\alpha_e = 1$)). Points are means of 1000 replicates × 100 sub-replicates.

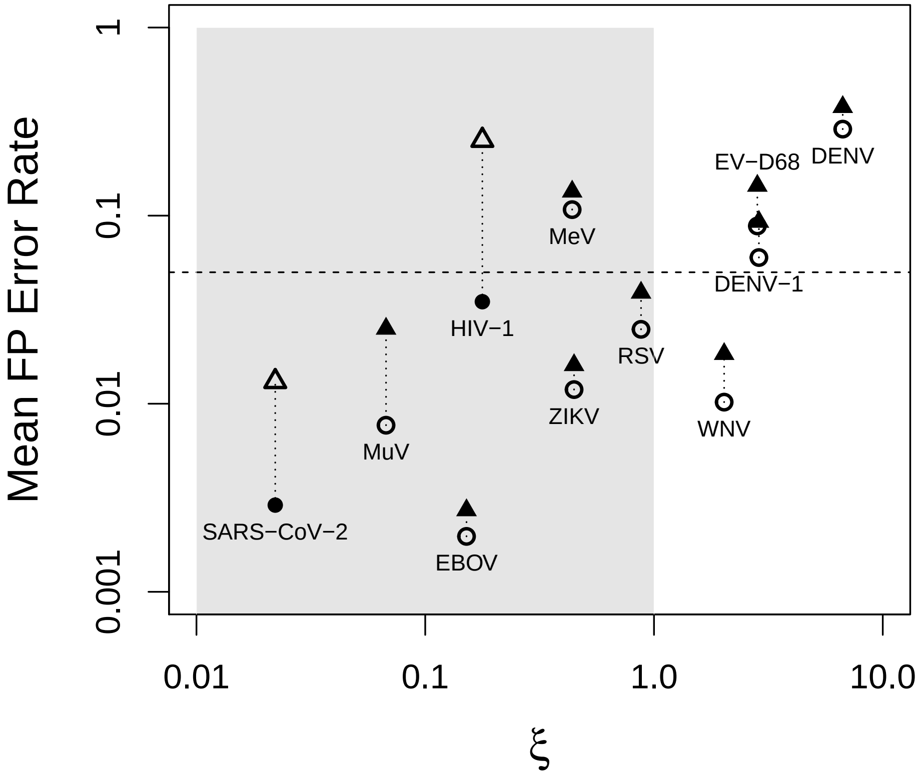

Fig. 6. Mean false positive rates estimated for 11 viral phylogenetic data sets, as a function of ξ and other parameters estimated from the data. Simulations used parameters given in Table 1, and are for MP assuming ASR variation follows either an invariant sites model (circles) or a gamma distributed model (triangles), with higher likelihood model point shaded, all assuming a PDA distribution of trees (100 subreplicates per simulation replicate; 500 simulation replicates). Near-perfect zone of $\xi \leqslant 1$ is shaded. Horizontal dashed line indicates a 0.05 FP rate.

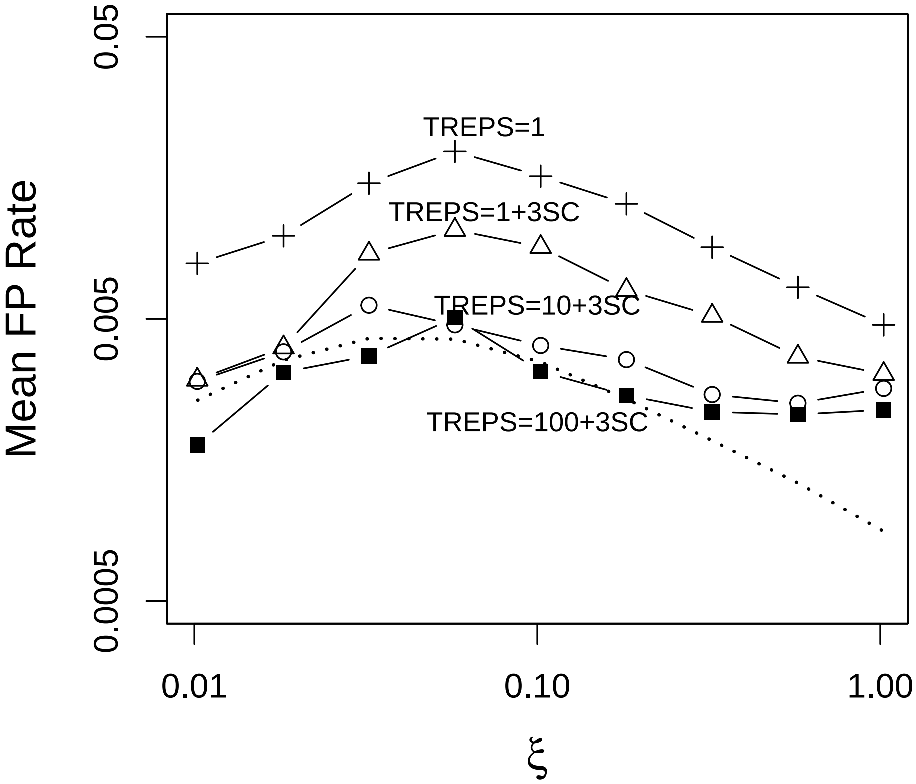

Fig. 7. Mean false positive rate in the near-perfect zone of $\xi \leqslant 1$ for maximum likelihood searches in IQTree2 using different search breadths and corrections. Each point is the mean in 500 replicate simulations with a PDA distribution of trees. Open plusses: one subreplicate search per simulation replicate; open triangles: one subreplicate plus 3-state correction; open circles: 10 subreplicate searches plus 3-state correction and trees combined via strict consensus. closed squares:100 subreplicate searches plus 3-state correction and trees combined via strict consensus. The “3-state correction” collapses edges less than $1/m$ substitutions per site, prior to consensus ($m = 1000$ sites). For comparison to theory for MP, predicted values from Eqn. 2 are given by dashed line.

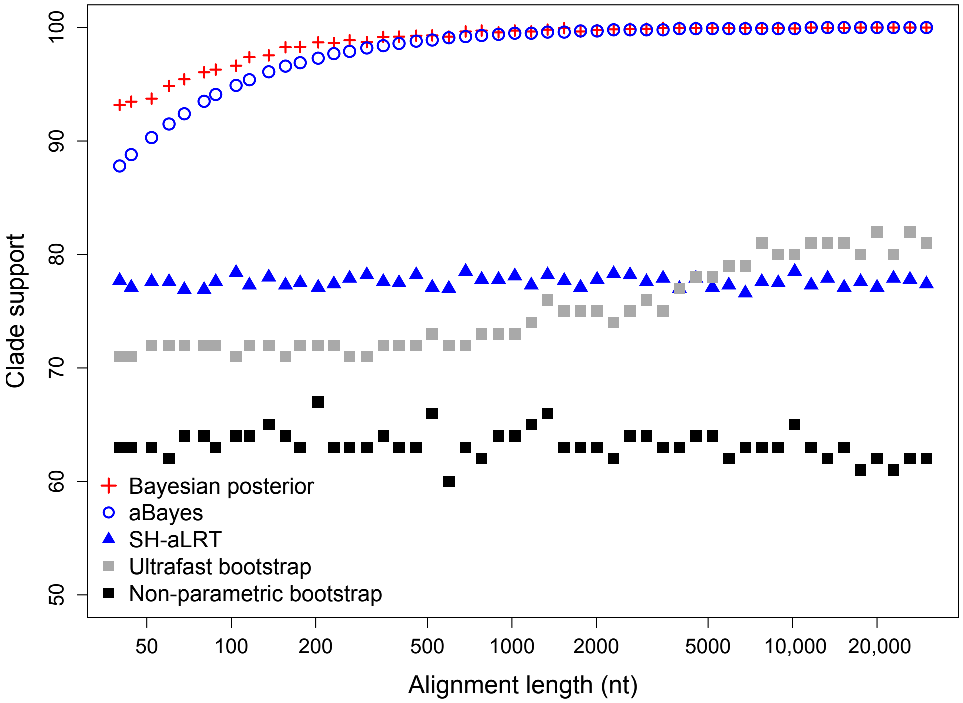

Fig. 8. Five measures of clade support estimated on perfect four-taxon trees as alignment length varies. Each edge of the tree has exactly one site with a substitution on that edge; all other sites are constant. Support metrics using bootstrap resampling are denoted with solid shapes.

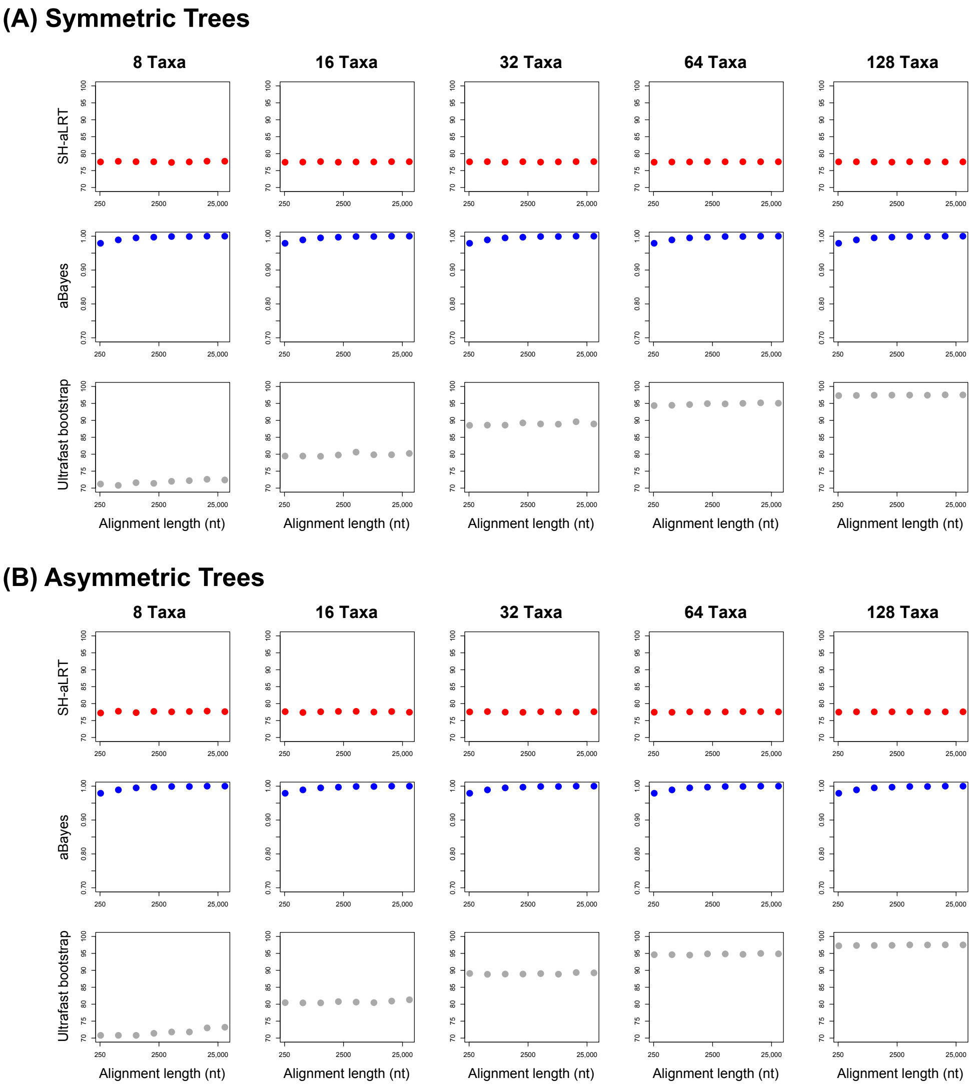

Fig. 9. Three measures of clade support estimated on perfect trees of different sizes and alignment lengths. Each edge of the tree has exactly one site with a substitution on that edge; all other sites are constant. (A) Simulations on perfectly symmetric trees. (B) Simulations on perfectly asymmetric (caterpillar) trees.