Syst. Biol. 0(0):1–0, 2021

DOI:10.1093/sysbio/syab069

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; and* [3]*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA*
[*]*Correspondence to be sent to: Department of Medicine, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA;*
E-mail: jwertheim@health.ucsd.edu.

Received 05 May 2021; reviews returned 05 August 2021; accepted 11 August 2021
Associate Editor: Vincent Savolainen

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.[Ebolavirus; epidemic; HIV; homoplasy; mumps virus; perfect phylogeny; SARS-CoV-2; virus; West Nile virus; Yule–Harding model; Zika virus.]

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). Homoplasy (Wake et al. 2011) is absent. Real-world data sets rarely allow reconstruction of perfect phylogenies, but algorithms can be modified to search efficiently for "near-perfect" trees when a small amount of homoplasy is present (Fernandez-Baca and Lagergren 2003; Awasthi et al. 2012). In this article, we address how best to measure accuracy in such "near-perfect" trees, what factors guarantee accuracy is high, and whether real data sets 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 (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 $\xi \leq 1$, 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 nonzero 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 data sets (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 data sets 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 objective motivates placing a premium on minimizing false negatives (i.e., on deciphering all such transmission events) and thereby maximizing resolution. Increased phylogenetic resolution is achievable by analyzing longer genomic fragments from viruses with faster evolutionary rates (Dudas and Bedford 2019). However, understanding the false positive rate remains a key issue in characterizing phylogenetic accuracy (Felsenstein and Kishino 1993), particularly in the special case of a poorly resolved tree with few—but well-supported—clades.

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 article is organized as follows: "Materials and Methods" are divided into two parts: first, mathematical theory (with proofs in the Supplementary material available on Dryad at http://dx.doi.org/10.6076/D12S3M), 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 $\mathrm{Prob}(\hat{T}=T)$ (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, $N_{\mathrm{FP}}=N_{\mathrm{FN}}$ (Berry and Gascuel 1996; Smirnov and Warnow 2021); otherwise they can contribute differentially to error. The Robinson–Foulds (RF) distance (Robinson and Foulds 1981), $d_{\mathrm{RF}}=N_{\mathrm{FP}}+N_{\mathrm{FN}}$, 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):

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

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 multistate characters, the notion of a perfectly evolved characters is stronger than that of 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 are 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/nonbinary) tree on leaf set $[n]$.

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

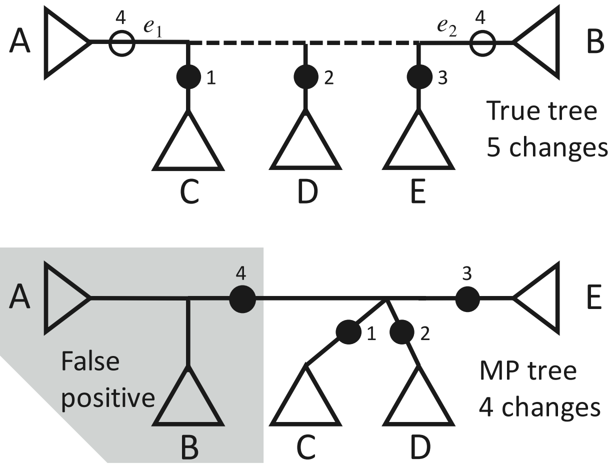

FIGURE 1. How a false positive split is inferred by maximum parsimony (MP). On true tree (top) sites 1–3 are binary and "perfect"; that is, they have only a single change (locations marked by filled circles), but site 4 is binary and homoplastic, changing twice (open circles), on edges e_1 and e_2. The dotted line is the path between the two homoplastic changes in site 4. If no perfect sites change along the dotted line path on the true tree, a false positive split is inferred on the MP tree (bottom).

(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 material available on Dryad).

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).

In other words, we are interested in "false splits" (i.e., splits in the reconstructed MP tree that are not present in the—underlying and unknown—true tree T). The split corresponding to f (by condition C-i) should *not* be in T (condition C-ii); however, condition C-iii would lead MP to add this false split into the reconstructed tree based on the other "true splits" since the false split is compatible with all of the latter.

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

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

The probability of a false split is then given by the following theorem (see Supplementary material available on Dryad for proof).

Theorem 1. *For each $T \in B(n)$, and $k \geq 1$ 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 Supplementary material available on Dryad).

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:

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

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

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

for all i between 1 and $n-3$ (see Supplementary material available on Dryad for proof).

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

(i)

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

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 $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$, satisfies

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

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

Notice also from Fig. 3 that as ξ increases from 0 the estimate of $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ 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:

$$\mathrm{FP}_{T_n}\sim 4/(3m),$$

which is independent of ξ (and n). Details are provided in the Supplementary material available on Dryad.

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 Seq-Gen v. 1.3.4 (Rambaut and Grassly 1997), with either JC69, HKY, or GTR models (and base frequencies and rate matrix parameters set or estimated from data), and with one of four across-site-rate (ASR) variation models: no variation, invariant sites model, gamma model, or free-rate model ranging from 2 to 10 bins (Kalyaanamoorthy et al. 2017)—the free-rate model was implemented in Seq-Gen by using 2–10 site partitions; iv) reconstruction of estimated tree $\hat{T}$ [using PAUP 4.0a, build 166: Swofford (2003)) for MP with options "hsearch add=simple swap=no nreps=r;contree "all/strict"; and using IQ-TREE 2 (v. 2.0.6) (Minh et al. 2020) for ML with options "-m JC+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 in the Dryad repository).

A typical data set size of $n=513$ (chosen to allow perfectly symmetrical trees plus one outgroup, when such were needed), and $m=1000$ was used to model trees large enough to potentially satisfy the near-perfect assumptions, and to have a sufficient number of sites to infer a range of accuracy when combined with $\bar{\lambda}$ values ranging from 10^{-5} to 0.316 substitutions per site. Gamma shape parameters were set at 0.1, 1.0, and 10.0, which encompasses distributions ranging from highly variable to nearly constant. For edge length variation this range encompasses what we observed in the empirical virus data sets. For ASR variation, it captures much of the range of inferred values we have seen in the literature. Finally, R was generally 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 (Swofford 2003) was used for MP bootstrapping (same heuristic search as above but with 100 replicates × 10 subreplicates); IQ-TREE 2 (Minh et al. 2020) 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, nonhomoplastic 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 Felsenstein's bootstrapping (1000 replicates), ultrafast bootstrapping (10,000 replicates), transfer bootstrap exchange (TBE; 1000 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 Felsenstein's bootstrap, ultrafast bootstrapping, TBE, SH-aLRT, and aBayes.

All Python scripts related to perfect tree simulations are available in the Dryad repository.

Virus Data Sets

Viral phylogenies were obtained from the NextStrain (Hadfield et al. 2018) website (accessed May 5, 2020) (Table 1). Phylograms were downloaded for dengue virus, dengue 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 et al. (2021) (see Data S1 of the Supplementary material available on Dryad at https://dx.doi.org/10.6076/D12S3M 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 of the Supplementary material available on Dryad).

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, Table S1 of the Supplementary material available on Dryad). For each virus, we used IQ-TREE 2 to infer the six rate parameters of a GTR substitution model with empirical base frequencies. The optimal site-to-site rate variation model, including free-rate models, was determined using the Bayesian information criterion (BIC) in IQ-TREE 2 (Kalyaanamoorthy et al. 2017). These models were used to parameterize sequence simulation in Seq-Gen, as described above.

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

$$1+\frac{m\overline{\lambda}}{\alpha_e} \qquad (5)$$

which was shown by Bedford and Hartl (2008) for an equivalent parameterization. Virus trees were preprocessed, setting any edge lengths $<1.1\times10^{-6}$ 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

Table 1. Parameters of 11 empirical virus phylogenies.

Abbreviations	Virus	Leaves	Sites	Resolution
DENV	Dengue virus	1197	10,264	0.8795
DENV-1	Dengue virus serotype 1	1067	10,264	0.8160
EBOV	Ebolavirus	1610	18,164	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	15,782	0.7009
MuV	Mumps virus	458	15,154	0.2961
RSV	Respiratory syncytial virus	997	14,986	0.6121
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2	583	29,668	0.2324
WNV	West Nile virus	2512	10,395	0.5960
ZIKV	Zika virus	543	10,320	0.5453

the true and estimated trees are nearly binary; therefore $N_{\mathrm{FP}} \cong N_{\mathrm{FN}}$.

However, when $\xi << \xi^*$, then $\mathrm{FP}_T << \mathrm{FN}_T$, 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 $\xi \leq 1$, 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,b). 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 JC69 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 $\xi \leq 1$, the expected number of substitutions per site is ≤ 1; the number of edges on which a site changes state is approximately Poisson 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 $\xi \leq 1$, 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,

$$\mathrm{FP}_T \cong \xi/3,$$

which implies FP_T is small when ξ is small enough (e.g., $\mathrm{FP}_T < 0.05$ 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 $\xi \leq 1$, 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, $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$, generated from parameters based on the inferred tree. In the remainder of this article, the "expected false positive rate" will generally refer to $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$. 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:

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

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 I, is monotonically decreasing in

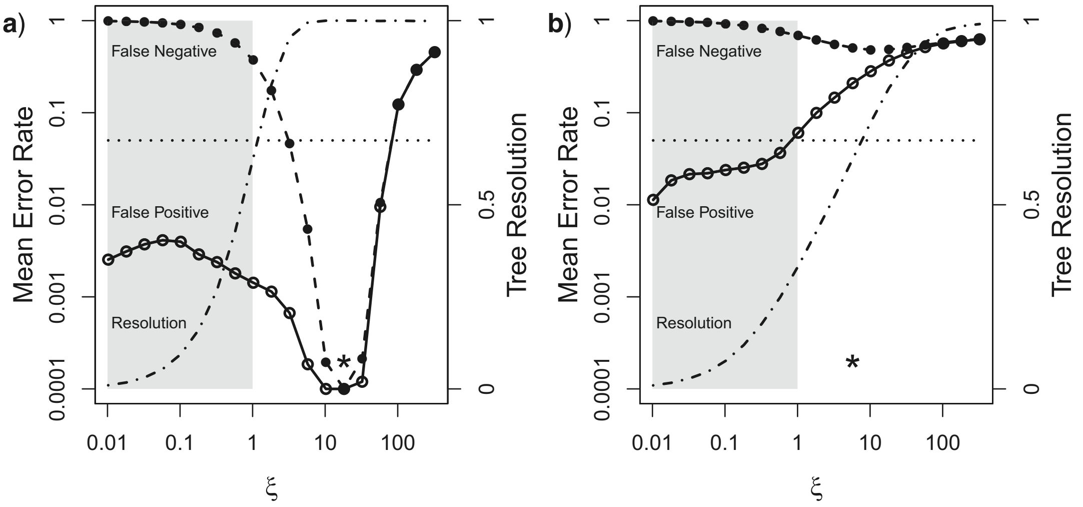

FIGURE 2. Accuracy of maximum parsimony phylogeny reconstruction in simulations over a wide range of per site tree length, ξ, and other parameters. Solid and dashed curves are mean false positive and negative error rates, respectively (log scale left); dashed sigmoidal curve is fractional resolution 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 ($\xi \leq 1.0$) 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 $\alpha_e = \alpha_{\mathrm{ASR}} = 0.25$).

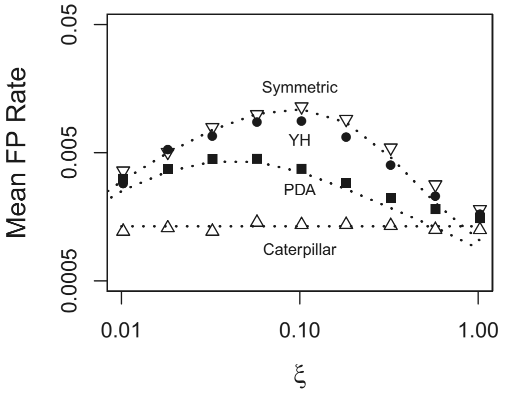

FIGURE 3. Mean false positive (FP) rate in four tree models. Fit to theoretical predictions from Equation 6 (or the limit expression of $4/3m$ for caterpillar trees: see Methods) are shown by dashed lines. Each point is mean of 1000 replicates × 100 sub-replicates. Simulation conditions were $n=513,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

ξ and m, and depends on the shape of T. Simulations indicate that the approximation is close for $\xi \leq 1$ (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 of the Supplementary material available on Dryad). $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ 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 $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ (Figs. 2, 4 and Fig. S4 of the Supplementary material available on Dryad). The difference between predicted $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ 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 $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ around $\xi = 0.1$; instead, $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ increases monotonically with ξ and eventually exceeds 5% for the simulated data set sizes. The impact of ASR variation is deleterious at all values of ξ, but even when ASR variation is large (gamma shape parameter $\alpha_{\mathrm{ASR}} = 0.1$), the false positive rate remains slightly below 5% for simulated data set sizes in the absence of EL variation (Fig. S4 of the Supplementary material available on Dryad).

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 $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ substantially, though it still remains well below 5% under our typical simulation conditions when $\xi \leq 1$ (Fig. S5 of the Supplementary material available on Dryad).

Thus, the near-perfect tree length of $\xi \leq 1$ is a region in which rate variation appears to have less of an impact on false positive rates than when tree lengths are longer.

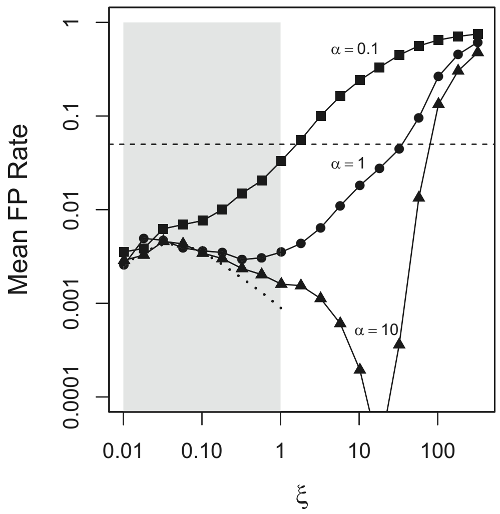

FIGURE 4. Effect of edge length variation on expected false positive (FP) rate for different values of the shape parameter of the edge length gamma distribution, α_e. Smaller values of α_e correspond to higher rate variation. ASR variation is assumed absent. The dashed curve is the prediction from Eqn. (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 $\xi \leq 1$.

This suggests that the definition of near-perfect zone in practice can include substantial rate variation.

Expected False Positive Rates in Virus Phylogenies

We estimated key parameters from the trees and underlying data for 11 empirical virus phylogenies (Table 1, Table S1 of the Supplementary material available on Dryad) and used simulation to estimate expected false positive rates (Fig. 5). 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 $\xi \leq 1.0$, and six of those had $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}] \leq 0.05$ irrespective of random tree model. $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ was generally lower for PDA versus YH models. As expected, $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ increased roughly with ξ, despite the large differences in these data sets.

Epidemics with young crown group ages on the order of years or decades (e.g., Zika virus, West Nile virus, and mumps virus) had expected false positive rates below 5%, even though West Nile virus had a ξ slightly above 1. Viruses encompassing single epidemics (e.g., SARS-CoV-2 in China, EBOV in West Africa, and HIV-1 in Cambodia) also had expected false positive rates below 5%. Remarkably, HIV-1 had a low expected false positive rate even though the tree was constructed using the fewest number of sites in our sample (from only a single partial gene). Number of site affects accuracy through the $\Omega(T_n, \xi, m)$ term in Eqn. 6.

Trees with lowest levels of resolution (Table 1) had the highest expected false positive rates. For example, dengue virus serotype 1, which does not represent a single epidemic, had low phylogenetic resolution, a $\xi > 1$, and a correspondingly high expected false positive rate. The phylogenetically more diverse dengue virus tree representing all four DENV serotypes had an even higher tree length and expected false positive rate.

The measles virus tree was an outlier with $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ above 5%, even though its tree length was below one. Notable, MeV had the fewest taxa of any virus analyzed (Table 1) and subsequently lower phylogenetic resolution. This combination of factors implies sensitivity to the assumption of large n in our results.

Extension to 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 Supplementary material available on Dryad 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 $\xi \leq 1$, a JC69 model, and no edge length or ASR variation, with trees inferred by IQ-TREE 2 (Minh et al. 2020) under the same model, are close to the equation's predictions (Fig. S6 of the Supplementary material available on Dryad). Nonetheless, some differences were observed, which tended to imply better accuracy for MP. These differences could largely be attributed to technical or implementation issues in ML software. 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. S6 of the Supplementary material available on Dryad). Second, ML programs set hard numerical lower bounds strictly greater than zero on edge lengths, often (by default) on the same order as $\bar{\lambda}$ for the virus data sets, so these must be reset downward to obtain correct tree likelihoods (Morel et al. 2021). 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 (IQ-TREE 2 issues a warning about this). We saw this in ML searches roughly when $\xi \geq 0.1$, 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 $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$. By collapsing short edge lengths inferred by ML to be less than $1/m$, this behavior can be mitigated (Fig. S6 of the Supplementary material available on Dryad).

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 $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$

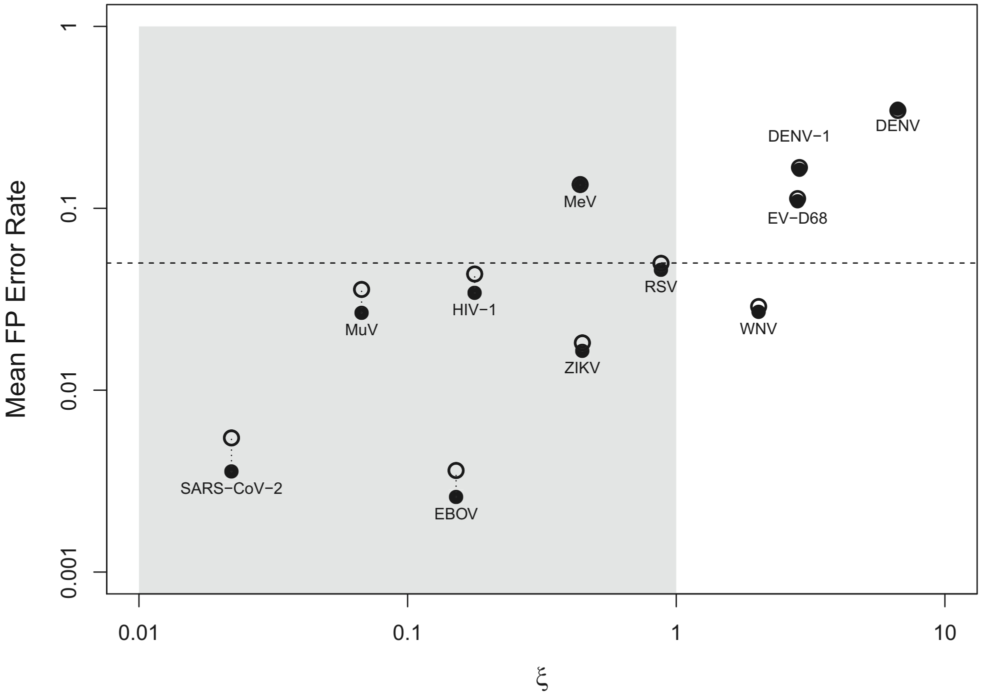

FIGURE 5. Expected false positive (FP) rates, $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$, for 11 empirical virus phylogenetic data sets (Table 1) for maximum parsimony (MP) inference, estimated by simulation using parameters estimated from the data (Table S1). Abbreviations given in Table 1. Simulation experiments used either a Yule–Harding random tree distribution (open circles) or PDA distribution (closed circles: some data points have indistinguishable differences between random tree models). Each point is mean of 500 replicates × 100 sub-replicates. The near-perfect zone of $\xi \le 1.0$ is shaded. Horizontal dashed line indicates a 0.05 expected false positive rate.

comparable with ML but with much faster running times.

Accuracy and Support in Near-Perfect and Perfect Trees

False positive "accuracy," defined as $1-\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$, 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 bootstrap support for MP 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 our $1-\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ false positive accuracy across the range of tree lengths in the near-perfect zone. Notably, aBayes is the only one of the 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 data sets, in which each branch was defined by a single change, and alignments range between 40 nt and 30,000 nt (Fig. S7 of the Supplementary material available on Dryad). As observed for MP, Felsenstein's ML bootstrap support is approximately 63%, regardless of sequence length, in accordance with theoretical predictions (Felsenstein 1985). Transfer bootstrap exchange (TBE) (Lemoine et al. 2018; Lutteropp et al. 2020) values were indistinguishable from Felsenstein's bootstrap. Of the other 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 ≥95% for alignments as short as 100 nt, which tracked the full Bayesian posterior support estimates that had support ≥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 to 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. S8 of the Supplementary material available on Dryad). The TBE method was developed to correct for a downward bias of bootstrap values often seen in large trees. As expected, TBE exceeds conventional bootstrap support as taxon number increases. However, this increase is modest in perfectly symmetrical trees compared with perfectly asymmetric trees and only surpasses 95% in the largest asymmetric trees (Fig. S8 of the Supplementary material available on Dryad).

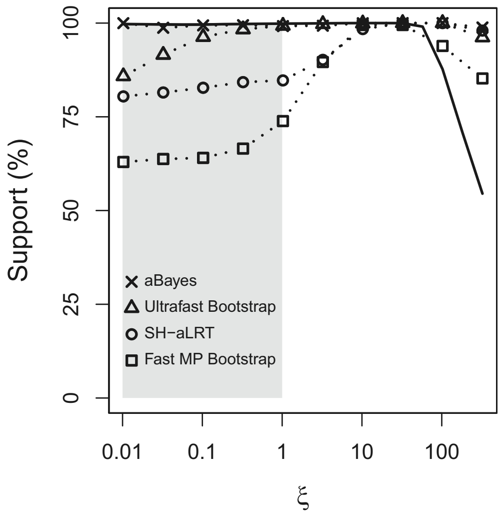

FIGURE 6. Statistical support measures compared to expected false positive accuracy, as a function of tree length. The solid curve is the mean value of $(1-\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}])\times 100$ in simulations. The near-perfect parameter space is shaded.

DISCUSSION

In this article, 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 $\xi \leq 1$ 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 data sets 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, Fig. S5 of the Supplementary material available on Dryad), and, if $\xi < 1$, a substantial level of rate variation can be present without elevating the false positive rate by nearly as much as when $\xi > 1$ (Figs. 2, 4, Fig. S4 of the Supplementary material available on Dryad).

The second case is clearly more relevant in real-world data. The 11 empirical virus data sets all had substantial rate variation and showed a general increase in false positive rate with ξ, with almost all rates below 5% occurring when $\xi \leq 1$, much like the predicted patterns seen in Figures 2B and 4. This observation 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 $\xi \leq 1$.

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" data sets 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 data sets depending on the study. For example, the SARS-Cov2 data set 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 $n=147,156$ has a tree length of $\xi=3.89$ (after collapsing any edges with $\lambda \leq 1.1\times 10^{-6}$), which is remarkably small for such a large tree but lies just outside our definition of near-perfect. This finding suggests that large-scale phylogenetic approaches for SARS-CoV-2 surveillance are appropriate (Ferreira et al. 2021; Turakhia et al. 2021) and that such approaches are unlikely to falsely suggest close relatedness (i.e., transmission clusters) where none exists.

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 $\lim_{m\to\infty}\mathrm{Prob}(\hat{T}_{\mathrm{MP}}=T)=1$ as long as $\lambda_{\min} > \xi^2/(1-\xi)$ (Steel 2000, Thm. 1(A)). More generally, let m' be the number of sites needed for $\mathrm{Prob}(\hat{T}_{\mathrm{MP}}=T)$ 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 $O(1/\lambda^2_{\min})$ 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 $\hat{T}$, in contrast to our findings 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 $\mathbb{E}_{\mathcal{D}}[\mathrm{FP}_{T_n}]$ 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. (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 building 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 data sets 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 $1-\mathbb{E}_{\mathcal{D}}[\text{FP}_{T_n}]$, 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 we do note the variability in support estimates we found and suspect that Bayesian measures may be better reflections of false positive accuracy in practice (Fig. 6). If individual clade support needs to be invoked in near-perfect viral phylogenies, we recommend Bayesian approaches that do not rely on bootstrap resampling of sparse substitutions. In near-perfect trees, Bayesian approaches can make use of the limited amount of genetic diversity to draw strongly supported inference, as opposed to bootstrapping approaches which require multiple sites supporting a clade before inferring similarly strong support. When phylogenetically informative data are limited, as in near-perfect trees, the consistency of the data supporting a clade appears more relevant than their prevalence.

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. This finding supports the current practice in SARS-CoV-2 nomenclature, whereby clades (e.g., denoting variants or migration events) are defined with reference to specific synapomorphies (Rambaut et al. 2020; Worobey et al. 2020; O'Toole et al. 2021). We acknowledge that 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. 2021).

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.

In conclusion, we have shown that many RNA virus datasets satisfy assumptions used to derive results on near-perfect phylogenetic accuracy. These criteria include sufficiently low substitution rates across a large enough tree and no recombination. Any set of genomes sampled in a clade on a short enough time scale, or highly conserved regions of genomes sampled across a deeper clade, can also satisfy the first assumption, but recombination would remain problematic in many taxa. Springer et al. (2020) illustrate a potential path forward in their study of "low-homoplasy" retroelement characters in mammal genomes. They pursue a species tree inference approach to such data, which would likely be "near-perfect" were it not for recombination. It may be possible to derive additional results on accuracy when local near-perfect trees (or sub-alignments) are combined under the multispecies coalescent (Liu et al. 2019).

Supplementary Material

Data available from the Dryad Digital repository: http://dx.doi.org/10.6076/D12S3M.

Acknowledgments

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 of the Supplementary material available on Dryad. M.J.S. thank the University of Arizona's HPC facility, Bio5 Institute, and Rod Wing's lab for computing support.

Funding

This work was supported by the NIH-NIAID [R01 AI135992 to J.O.W.].

References

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

Alfaro M.E., Holder M.T. 2006. The posterior and the prior in Bayesian phylogenetics. Annu. Rev. Ecol. Evol. Syst. 37:19–42.

Anisimova M., Gil M., Dufayard J.F., Dessimoz C., Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst. Biol. 60:685–699.

Awasthi P., Blum A., Morgenstern J., Sheffet O. 2012. Additive approximation for near-perfect phylogeny construction. In: Goemans M., Jansen K., Rolim J., Trevisan L., editors. Approximation, randomization, and combinatorial optimization. Algorithms and techniques. Berlin: Springer. p. 25–36.

Bedford T., Hartl D.L. 2008. Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in *Drosophila*. Mol. Biol. Evol. 25:1631–1638.

Berry V., Gascuel O. 1996. On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. Mol. Bio. Evol. 13:999–1011.

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

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

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

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

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

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

Efron B., Halloran E., Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93:13429–13434.

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

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. Sunderland, MA: Sinauer Press.

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

Fernandez-Baca D., Lagergren J. 2003. A polynomial-time algorithm for near-perfect phylogeny. SIAM J. Comput. 32:1115–1127.

Ferreira R.-C., Wong E., Gugan G., Wade K., Liu M., Baena L.M., Chato C., Lu B., Olabode A.S., Poon A.F.Y. 2021. CoVizu: rapid analysis and visualization of the global diversity of SARS-CoV-2 genomes. bioRxiv. https://doi.org/10.1101/2021.07.20.453079.

Gire S.K., Goba A., Andersen K.G., Sealfon R.S., Park D.J., Kanneh L., Jalloh S., Momoh M., Fullah M., Dudas G., Wohl S., Moses L. M., Yozwiak N.L., Winnicki S., Matranga C.B., Malboeuf C. M., Qu J., Gladden A. D., Schaffner S.F., Yang X., Jiang P.P., Nekoui M., Colubri A., Coomber M.R., Fonnie M., Moigboi A., Gbakie M., Kamara F. K., Tucker V., Konuwa E., Saffa S., Sellu J., Jalloh A.A., Kovoma A., Koninga J., Mustapha I., Kargbo K., Foday M., Yillah M., Kanneh F., Robert W., Massally J.L., Chapman S.B., Bochicchio J., Murphy C., Nusbaum C., Young S., Birren B.W., Grant D. S., Scheiffelin J.S., Lander E.S., Happi C., Gevao S.M., Gnirke A., Rambaut A., Garry R.F., Khan S.H., Sabeti P.C. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science 345:1369–1372.

Grenfell B.T., Pybus O.G., Gog J.R., Wood J.L., Daly J.M., Mumford J.A., Holmes E.C. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303:327–332.

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

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

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

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

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

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

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

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

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Bio. Evol. 35:518–522.

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

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

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

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

Lanfear R. 2020. A global phylogeny of SARS-CoV-2 sequences from GISAID. Zenodo. doi: 10.5281/zenodo.3958883.

Lemoine F., Domelevo Entfellner J.B., Wilkinson E., Correia D., Davila Felipe M., De Oliveira T., Gascuel O. 2018. Renewing

Felsenstein's phylogenetic bootstrap in the era of big data. Nature 556:452–456.

Liu L., Anderson C., Pearl D., Edwards S. 2019. Modern phylogenomics: building phylogenetic trees using the multispecies coalescent model. Methods Mol. Biol. 1910:211–239.

Lutteropp S., Kozlov A.M., Stamatakis A. 2020. A fast and memory-efficient implementation of the transfer bootstrap. Bioinformatics 36:2280–2281.

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

Morel B., Barbera P., Czech L., Bettisworth B., Hubner L., Lutteropp S., Serdari D., Kostaki E.G., Mamais I., Kozlov A.M., Pavlidis P., Paraskevis D., Stamatakis A. 2021. Phylogenetic analysis of SARS-CoV-2 data is difficult. Mol. Biol. Evol. 38:1777–1791.

O'Toole Á., Scher E., Underwood A., Jackson B., Hill V., McCrone J.T., Colquhoun R., Ruis C., Abu-Dahab K., Taylor B., Yeats C., Du Plessis L., Maloney D., Medd N., Attwood S.W., Aanensen D.M., Holmes E.C., Pybus O.G., Rambaut A. 2021. Assignment of epidemiological lineages in an emerging pandemic using the Pangolin tool. Virus Evol. veab064. doi: 10.1093/ve/veab064.

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

Pekar J., Worobey M., Moshiri N., Scheffler K., Wertheim J.O. 2021. Timing the SARS-CoV-2 index case in Hubei province. Science 372:412–417.

Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Rambaut A., Holmes E.C., O'Toole Á., Hill V., McCrone J.T., Ruis C., du Plessis L., Pybus O.G. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5:1403–1407.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Roch, S. 2019. Hands-on introduction to sequence-length requirements in phylogenetics. In: Warnow T., editor. Bioinformatics and phylogenetics: seminal contributions of Bernard Moret. Springer International Publishing. p. 47–86.

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

Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

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

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

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

Springer M. S., Molloy E. K., Sloan D. B., Simmons M. P., Gatesy J. 2020. ILS-aware analysis of low-homoplasy retroelement insertions: inference of species trees and introgression using quartets. J. Hered. 111:147–168.

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

Steel M., Leuenberger C. 2017. The optimal rate for resolving a near-polytomy in a phylogeny. J. Theor. Biol. 420:174–179.

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

Susko E. 2009. Bootstrap support is not first-order correct. Syst. Biol. 58:211–223.

Swofford D.L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland: Sinauer Associates.

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

Turakhia Y., Thornlow B., Hinrichs A.S., De Maio N., Gozashti L., Lanfear R., Haussler D., Corbett-Detig R. 2021. Ultrafast sample placement on existing trees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat. Genet. 53:809–816.

Wake D.B., Wake M.H., Specht C.D. 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. In: Chauve C., El-Mabrouk N., Tannier E., editors. Models and algorithms for genome evolution. London: Springer. p. 85–146.

Worobey M., Pekar J., Larsen B.B., Nelson M.I., Hill V., Joy J.B., Rambaut A., Suchard M.A., Wertheim J.O., Lemey P. 2020. The emergence of SARS-CoV-2 in Europe and North America. Science 370:564–570.

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