Biological causes and impacts of rugged tree landscapes in phylodynamic inference

Jiansi Gao[a,1], **Marius Brusselmans**[b], **Luiz M. Carvalho**[c,d,e], **Marc A. Suchard**[f,g,h], **Guy Baele**[b], **and Frederick A. Matsen IV**[a,i,j,k,1]

Phylodynamic analysis has been instrumental in elucidating the epidemiological and evolutionary dynamics of pathogens. The Bayesian approach to phylodynamics integrates out phylogenetic uncertainty, which is typically substantial in phylodynamic datasets due to low genetic diversity. Bayesian phylodynamic analysis does not, however, scale with modern datasets, partly due to difficulties in traversing tree space. Here, we set out to characterize tree space of phylodynamic inference and assess its impacts on analysis difficulty and key biological estimates. By running extensive Bayesian analyses of 15 classic large phylodynamic datasets and carefully analyzing the posterior samples, we find that the posterior landscape in tree space ("tree landscape") is diffuse yet rugged, leading to widespread tree sampling problems that usually stem from the sequences in a small part of the tree. We develop clade-specific diagnostics to show that a few sequences—including putative recombinants and recurrent mutants—frequently drive tree space ruggedness and sampling problems, although existing data-quality tests show limited power to detect such sequences. The sampling problems can significantly impact phylodynamic inferences or even distort major biological conclusions; the impact is usually stronger on "local" estimates (*e.g.*, introduction history) that are associated with particular clades than on "global" parameters (*e.g.*, demographic trajectory) that are governed by the general tree shape. We evaluate existing and newly-developed MCMC diagnostics, and offer strategies for optimizing MCMC settings and mitigating impacts of the sampling problems. Our findings highlight the need for and directions to develop efficient traversal over the rugged tree landscape, ultimately advancing scalable and reliable phylodynamics.

phylodynamics | tree space | viral evolution | epidemiology | Markov chain Monte Carlo

Significance Statement

Bayesian phylodynamics is central to epidemiological studies but computationally challenging. One major challenge is posed by the need to explore a vast and complex tree space. Phylodynamic datasets usually comprise many sequences with limited genetic diversity, distinct from traditional phylogenetic datasets, warranting characterization of phylodynamic tree space and inference performance. Here, we demonstrate that: 1) phylodynamic tree landscape is highly rugged, leading to widespread tree sampling problems that are frequently driven by a few sequences, and; 2) these problems can distort biological conclusions. We develop new diagnostics to identify the problematic sequences and highlight potential solutions to mitigate their impacts. We offer strategies to optimize phylodynamic analysis workflows and to develop algorithms for navigating tree space ruggedness, thereby facilitating infectious disease investigation.

Phylodynamic inference combines genetic data with epidemiological information via phylogenetic modeling, and is now a key component of infectious disease studies for elucidating pathogen biology and transmission dynamics. During the COVID-19 pandemic, unprecedented volume of SARS-CoV-2 genomic data proved invaluable for identifying variants of concern (1, 2) and assessing intervention measures (3, 4). Phylodynamic datasets comprising densely sampled heterochronous genomic sequences from measurably evolving populations, such as pathogens and somatic cells (5), are now increasingly available. However, despite the sheer volume of data, limited genetic diversity among the sequences leads to significant uncertainty in phylogenetic inference (6, 7).

Bayesian phylodynamic inference incorporates various sources of epidemiological and biological information (8), and accommodates phylogenetic uncertainty naturally by integrating over the space of possible phylogenetic trees ("tree space"). This integration is typically achieved via Markov chain Monte Carlo (MCMC), an algorithm that approximates the joint posterior by stochastic simulation. Many key evolutionary and epidemiological inferences, such as the demographic trajectory and geographic dispersal history of the pathogen, depend on the underlying phylogeny, making efficient exploration of tree space crucial for scalable phylodynamics. However, due to the challenges in navigating and traversing the vast and complex tree space and the computational costs of evaluating each step, Bayesian phylodynamic analysis of modern genomic datasets is often computationally prohibitive.

The tree landscape can also be rugged—comprising complex topographical structures including local peaks (9), terraces (10), and ridges (11)—leading to further challenges in efficiently traversing the tree space and to tree sampling problems. Indeed, tree searches may get trapped at local peaks or wander on terraces and ridges, resulting in a lack of convergence between replicate runs and/or inadequate mixing within each run (Fig. 1). Accordingly, MCMC diagnosis is typically treated as an integral step in Bayesian phylogenetic analysis. Traditional diagnostics—*e.g.*, the effective sample size (ESS), which uses the number of effectively independent samples to quantify within-chain mixing—mostly focus on continuous model parameters (12), while recent studies (13–17) advocate for diagnosing the tree topology as well given its central role in phylogenetic studies. The ruggedness in tree space and how well tree search explores it is also a long-standing phylogenetic research focus (9, 18–20), some (*e.g.*, 21) even predating Bayesian phylogenetics. However, all these studies focus on characterizing the tree space of phylogenomic inference with divergent species and assessing the impact on estimating the tree topology itself.

Author affiliations: [a]Computational Biology Program, Fred Hutchinson Cancer Center, Seattle, WA 98109; [b]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium; [c]Institute of Ecology and Evolution, University of Edinburgh, Edinburgh, United Kingdom; [d]School of Applied Mathematics, Getulio Vargas Foundation, Rio de Janeiro, Brazil; [e]Department of Biostatistics, University of North Carolina Chapel Hill, Chapel Hill, NC 27599; [f]Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095; [g]Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA 90095; [h]Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA 90095; [i]Howard Hughes Medical Institute, Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109; [j]Department of Genome Sciences, University of Washington, Seattle, WA 98109; [k]Department of Statistics, University of Washington, Seattle, WA 98109

The authors declare no competing interests.

[1]To whom correspondence may be addressed. E-mail: jsigao@gmail.com (JG) or matsen@fredhutch.org (FAM).

DRAFT

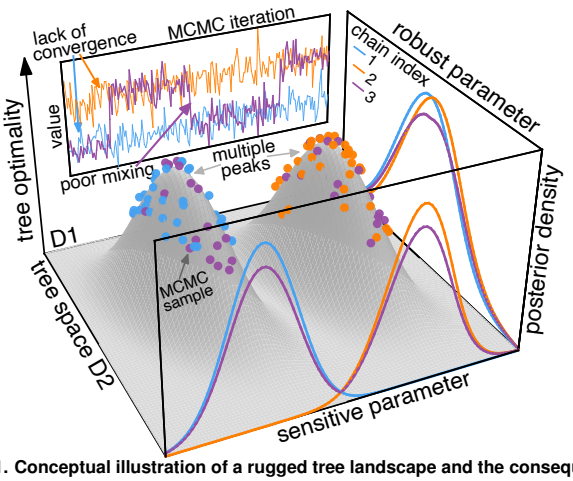

Fig. 1. Conceptual illustration of a rugged tree landscape and the consequent sampling problems. The *x*- and *y*-axes correspond to a two-dimensional representation of the tree space, while the *z*-axis indicates the tree optimality score (*e.g.*, Bayesian posterior density). This tree landscape is multimodal with two local peaks depicted. Three hypothetical MCMC replicate chains are run to sample the posterior: chains 1 and 2 randomly start from the left and right peaks, respectively, and then are both stuck on the corresponding peak for the entire chain, while chain 3 is able to cross the valley (*i.e.*, commute between the peaks) but only occasionally. Therefore, if only one of the chains is run, then chain 1 or 2 would suggest satisfactory sampling by itself, whereas chain 3 would uncover the two peaks and show inadequate mixing due to the infrequent commute. When multiple chains are run, comparing them will reveal the lack of convergence in sampling the tree. The tree space ruggedness and tree sampling problems may lead to poor sampling of the model parameters that depend on the tree, which may be manifested in the MCMC trace plot (back panel), resulting in discrepant posterior distributions inferred among the chains (front panel). Conversely, some other parameters may be relatively robust to tree sampling problems such that their estimates are still congruent among the chains (right panel) despite the ruggedness and poor tree sampling.

Phylodynamic analyses distinguish from traditional phylogenetic analyses by focusing on estimating epidemiological and evolutionary parameters while inferring a dated phylogeny of heterochronously sampled viral sequences with limited genetic diversity under a population-process tree prior (*e.g.*, the coalescent model). A comprehensive survey on tree space of phylodynamic inference is thus much needed. Does the limited genetic diversity lead to a less rugged phylodynamic tree landscape compared to its phylogenomic counterpart? How efficiently does MCMC explore phylodynamic tree space, and how much of the computational cost can be attributed to the ruggedness? Do tree sampling problems occur frequently in phylodynamic analysis and how robust are focal phylodynamic estimates to these problems? Which diagnostics excel in assessing the sampling performance and identifying the underlying causes of sampling problems? Are there any particular biological factors that may contribute to the ruggedness, and can we identify them for a given dataset before running a costly phylodynamic analysis?

Here, to answer these critical questions, we comprehensively characterize phylodynamic tree space and assess its impacts on phylodynamic analysis difficulty and key biological inferences. By running exhaustively long Bayesian phylodynamic analyses of large viral genomic datasets and carefully analyzing the posterior outputs, we demonstrate that the tree space typically has a diffuse yet highly rugged landscape, resulting in widespread sampling problems. We develop new MCMC diagnostics to dissect these tree sampling problems and identify the sequences that are the main drivers of the problems ("problematic sequences"); we show that removing these problematic sequences leads to dramatically improved MCMC performance. We also evaluate various data-quality tests as alternative quick ways to detect the problematic sequences prior to a phylodynamic analysis and reveal their limited efficacy. We go on to present empirical evidence illustrating that key phylodynamic estimates can be strongly impacted by tree sampling problems. Finally, we offer strategies to help researchers identify problematic sequences and set up efficient phylodynamic analyses, enhancing the potential of phylodynamic methods for addressing important biological questions with modern datasets.

Results

We collected 15 large phylodynamic datasets from representative high-impact empirical studies and performed comprehensive Bayesian phylodynamic analyses. Specifically, for each dataset we ran 10 independent MCMC replicate chains using `BEAST` (35), including five chains with 400 million iterations each ("400M run") and the other five with one billion iterations each ("1B run"). We specified the models and priors, as well as the MCMC proposals, following the original studies; the proposals, including those on the tree, are mostly identical to the defaults of `BEAUti`, reflecting common empirical practices. We then analyzed the posterior outputs to visualize tree space and quantify sampling performance via various existing and newly developed MCMC diagnostics. We detail these analyses in Materials and Methods, and in Suppl. Section S2.

Biological Properties of Phylodynamic Datasets Underlie Tree Inference Difficulties. To quantify the genetic diversity in the sequence alignments, we performed mutation mapping over the posterior trees. For the majority of the datasets, most branches in the tree experienced zero or one mutation across the alignment, and most sites are invariant or parsimony-uninformative (Suppl. Fig. S1). Such limited genetic diversity leads to high uncertainty in tree topology inference, manifested in both the large fraction of weakly supported clades that are subtended by short branches in the summary tree (Fig. 2, inset pie charts, and Suppl. Fig. S3) and a diffuse posterior tree space (Fig. 3 and Suppl. Fig. S4). Notable exceptions include the avian influenza virus (AIV) and Lassa virus (LASV) datasets, where most branches experienced mutations at tens or hundreds of sites, leading to little uncertainty in the summary tree topology. In addition, due to the fluctuating prevalence of viruses and temporal biases in the sampling process, the number of active phylogenetic lineages varies heavily over time (Suppl. Fig. S2). Complex evolutionary and epidemiological processes, such as selection and clonal interference, likely further contribute to the highly imbalanced tree shape (Fig. 2). These biological properties distinguish tree inference in viral phylodynamic analysis from that in phylogenomic analysis of divergent species.

Phylodynamic Tree Landscape is Highly Rugged and Tree Sampling Problems are Widespread. We visualize the posterior tree space as a two-dimensional plot via multidimensional scaling (MDS), a dimensionality reduction technique that converts the high-dimensional tree space into a low-dimensional representation (36). MDS requires a pairwise tree distance matrix as its input; as we focus on the posterior space defined by the discrete tree topology, we computed both the Robinson-Foulds (RF; 37) and rooted subtree prune-and-regraft (rSPR; 38) distances between all pairs of sampled posterior trees. For most datasets, the tree space is multimodal, and tree sampling problems are widespread (RF: Fig. 3; rSPR: Suppl. Fig. S5). In seven of the 15 datasets, independent chains fail to converge; *i.e.*, each chain is stuck in a region of tree space (composed of one or several

DRAFT

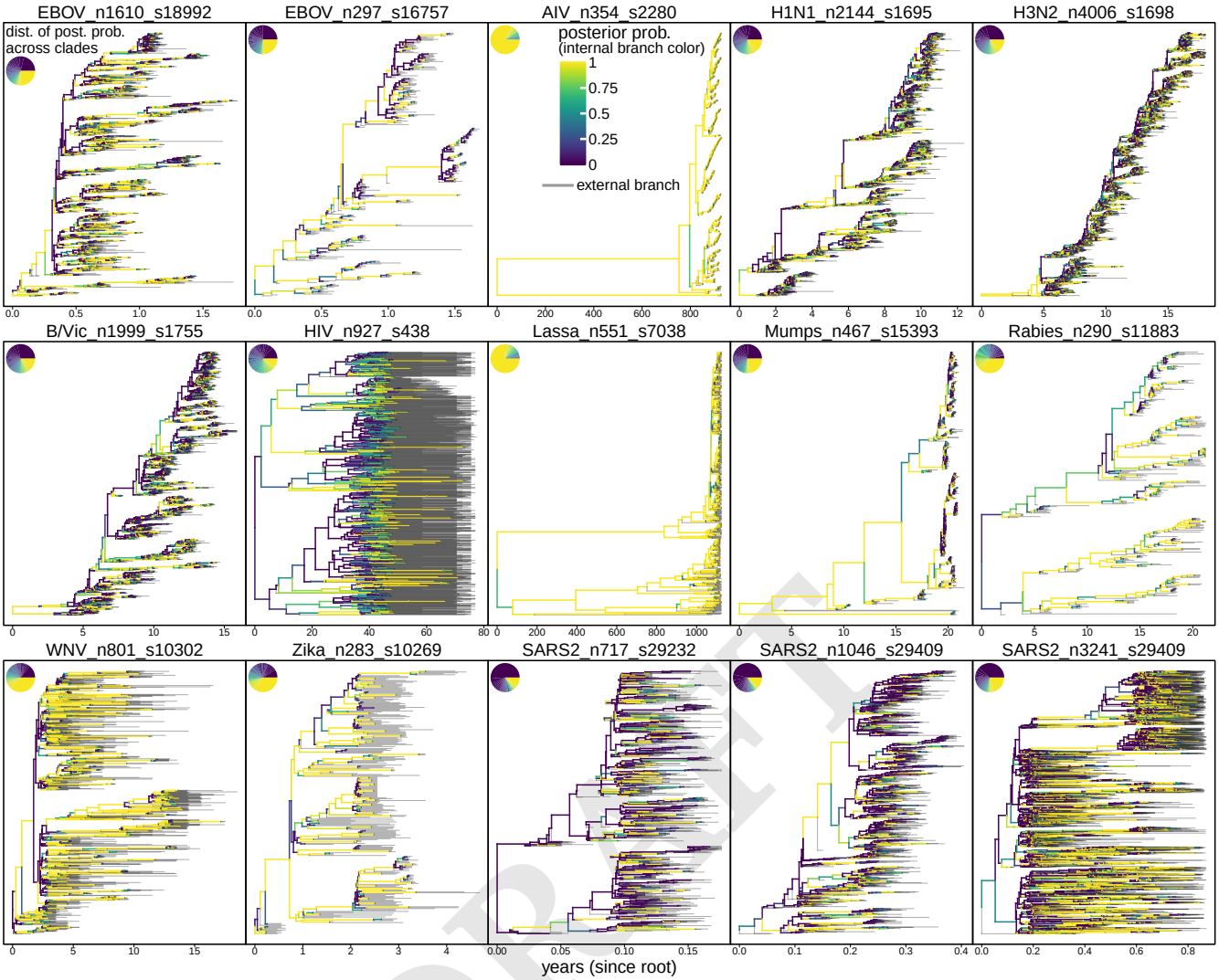

Fig. 2. Dated viral phylogenies are typically highly imbalanced and their inference is highly uncertain. Each panel presents the summary (maximum-clade credibility; MCC) time tree of an empirical dataset. Branch colors indicate the posterior probability of the subtending nodes (yellow: high; purple: low; gray: external branches). The inset pie chart (in the top-left corner of each panel) summarizes the distribution of the inferred posterior probabilities across all nodes in the corresponding MCC tree. Across nearly all datasets, most clades in the MCC tree are weakly supported and subtended by short branches. The tree shape is highly imbalanced with the heterochronous tips unevenly distributed through time. The dataset name labeled on top of each panel is constructed to indicate the virus, number of sequences (n), and number of sites (s) in the alignment of the dataset. Dataset sources (numbered by the panel order): 1 (22); 2 (23); 3 (24); 4–6 (25); 7 (26); 8 (27); 9 (28); 10 (29); 11 (30); 12 (31); 13 (32); 14 (33), and; 15 (34). See Suppl. Table S1 and Section S2 for detailed description of the datasets.

peaks) and are unable to move between the regions even after one billion iterations. Five other datasets exhibit slow mixing; *i.e.*, the chain is able to commute between the peaks but tens of millions of iterations are usually required for each commute. The joint log-posterior density ("posterior density") values are effectively indistinguishable between the peaks (Suppl. Fig. S4) so that no global mode can be established, revealing the diffuse yet rugged tree landscape. The MCC tree summarized from each chain also frequently resides on different peaks (Fig. 3, triangles), while the MCC tree summarized from all the chains combined mostly resides on one of the peaks towards the corner of the MDS plot (Fig. 3, open circle).

Tree topology diagnostics. In addition to inspecting the tree space visually, we quantify it and the tree sampling performance via MCMC diagnostics focusing on tree topologies. Specifically, we evaluate within-chain mixing with tree ESS (13, 16) and between-chain convergence with tree potential scale reduction factor (tree PSRF; 9, 39) as well as the average standard deviation of clade frequencies (ASDCF; 40). These topology diagnostics corroborate the MDS plots, although overall they appear to be more conservative in revealing the sampling problems (Suppl. Fig. S6). Following the conventional cutoffs of the diagnostics (ESS below 200; PSRF above 1.1; ASDCF above 0.01), four of the 15 datasets exhibit poor mixing in sampling the tree, while 12 datasets (primarily identified by the ASDCF) fail to converge.

Real-valued variable diagnostics. The topology diagnostics show promise for identifying tree sampling problems but are computationally expensive to obtain. Alternatively, MCMC diagnosis of `BEAST` analysis is typically conducted by loading the posterior log files into `Tracer` (12) and inspecting the computed ESS values of the logged continuous variables. In light of this standard empirical practice, we sought to investigate the potential of MCMC diagnostics computed with the variables that are logged by default, such as the posterior density and evolutionary rate, in assessing tree sampling performance. In addition, we also develop new tree-based variable diagnostics computed with the parsimony score of (*i.e.*, the minimum number of mutations required over) each sampled tree; this score is model-independent, determined solely by tree topology given the alignment, and insensitive noise in parameter sampling. While the diagnostics of some continuous

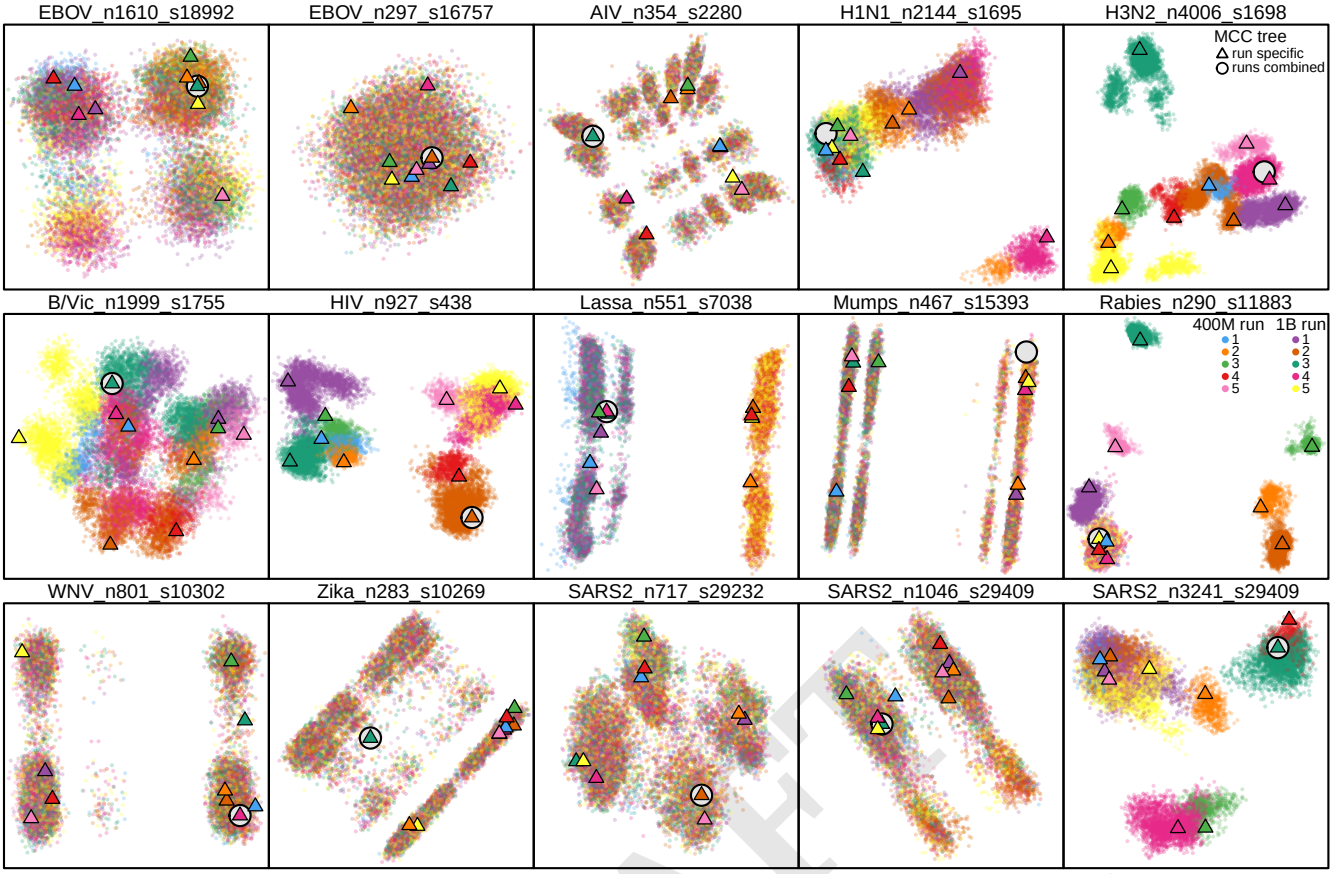

Fig. 3. Phylodynamic tree space is diffuse yet highly rugged, leading to widespread tree sampling problems. Each panel shows the MDS visualization of the posterior tree space of a dataset, generated based on the pairwise RF distances computed over the posterior trees. Each dot represents a sampled tree with its color indicating the corresponding MCMC chain index. A triangle represents the MCC tree of each chain, while an open circle indicates the MCC tree summarized across all chains. For most datasets the tree space is multimodal. In seven datasets (*e.g.*, HIV), the chains fail to converge even after one billion iterations, revealed by the clear separation of colors. Five other datasets (*e.g.*, EBOV_n1610_s18992) exhibit a pattern of slow mixing within each chain, indicated by the unevenly mixed colors among the peaks.

variables (including the substitution model parameters and root age) almost never indicate any sampling problem, the variable diagnostics overall exhibit a greater rate of inadequate mixing and a lower rate of convergence failure compared to the topology diagnostics (Suppl. Fig. S6). In contrast, the parsimony-score ESS and PSRF corroborate the MDS plots more closely, indicating inadequate mixing and convergence failure for 12 and seven datasets, respectively.

Tree Sampling Problems Can Be Dissected by Clade-Specific MCMC Diagnostics. To quantify the difference between trees in distinct peaks and identify the clades constituting tree sampling problems, we explore various clade-specific diagnostics. We first examined the within-chain mixing and between-chain convergence of sampling each clade using clade ESS and standard deviation of clade frequency (SDCF), respectively. The clade ESS is computed by treating the occurrence of each clade as a binary variable, whose value is one when that clade is present in the sampled tree and zero otherwise (15). For most datasets, less than 10% of the clades exhibit poor sampling (Suppl. Fig. S6), demonstrating that the problems with sampling the tree likely come from a small part of it. We also assessed the MCMC performance of clade-specific continuous variables—including the node age, subtending branch length, and branch-specific evolutionary rate—by computing their ESSs and PSRFs. Similar to the clade occurrence, in most datasets less than 20% of the clades exhibit sampling problems for each clade-specific variable (Suppl. Fig. S6). Finally, we evaluated the association between sites in the sequence alignment and tree sampling problems by treating the log-likelihood of each site ("site likelihood") as a continuous variable to compute its ESS and PSRF; similar to the clade-specific patterns, for most datasets a small fraction (below 10%) of the sites are associated with the sampling problems (Suppl. Fig. S6). An outlier to this pattern is the HIV dataset, where most of the clades and sites exhibit poor sampling, likely caused by problems with sampling the substitution model parameters that occur exclusively to this dataset.

A Few Problematic Sequences Can Strongly Impact Phylodynamic Analysis; MCMC Diagnostics Can Identify Them. Among the 12 datasets with tree sampling problems, the problems appear to stem from a small fraction of the sequences in all but the HIV dataset. For the five datasets where fewer than 10 problematic sequences ("tips") are identified by the clade-specific diagnostics, we investigate whether these problematic tips are the direct causes of tree sampling problems by pruning them from each sampled tree and then re-assessing MCMC performance via the diagnostics re-computed with the pruned trees. For all five datasets, the sampling problems are largely attenuated after removing the identified problematic tips, and in four of these datasets merely 1–3 such tips need to be removed (Fig. 4 and Suppl. Figs. S9–S12).

To look at this effect in detail, we present the LASV dataset (27) here as an example, where pruning off just two of the 551 sequences completely reshapes the tree space and effectively resolves the sampling problems (Fig. 4). At least two peaks existed in the posterior tree space of this dataset (Fig. 4E, Left); four of the 10 chains were stuck on the left peak, another

DRAFT

four chains were stuck on the right peak, while the remaining two chains were able to cross the valley but only once in the entire chain. Despite having seemingly adequate within-chain mixing according to the ESSs, the exceptionally high branch-rate PSRFs associated with two tips ("088" and "091") indicate strong discrepancies among the chains (Fig. 4C, Left). The peaks in the tree space (Fig. 4B) correspond exactly to the alternative topological positions of these two tips (Fig. 4A); 091 is sister to 088 in all the trees in the right peak, whereas those in the left peak have 091 being sister to the cherry (036,046). Thus, the two chains crossed the valley by effectively moving 091 from being 088's sister to being sister to (036,046), and once the move happened, 088 wobbled between being the base of clades X and Y (Fig. 4A), dividing the left peak into two commutable sub-peaks (Fig. 4B, brown and yellow).

Tree sampling problems can strongly impact key biological estimates. Both the evolutionary rate over and time span of the branches subtending the two problematic tips are largely determined by the subtree topology; the mean estimates vary by 2–4 times and 95% credible intervals (CIs) barely overlap between the two topologies (Fig. 4C and Suppl. Fig. S7C). The evolutionary rate over branch 088 roughly halves when 088 stops being sister to 091 (Fig. 4C, Right top, green to brown), whose mean estimate decreases from the 99.5 to 0.5 percentile of the distribution of the mean rates across branches (Fig. 4C, Right, light gray). In addition, age estimates of the nodes (*i.e.*, tMRCAs) close to those two tips in the tree are also affected by the sampling problems, with the tMRCA of clade X (Fig. 4D, Right) inferred to be significantly younger when 088 and 091 are sisters (green) than otherwise (yellow).

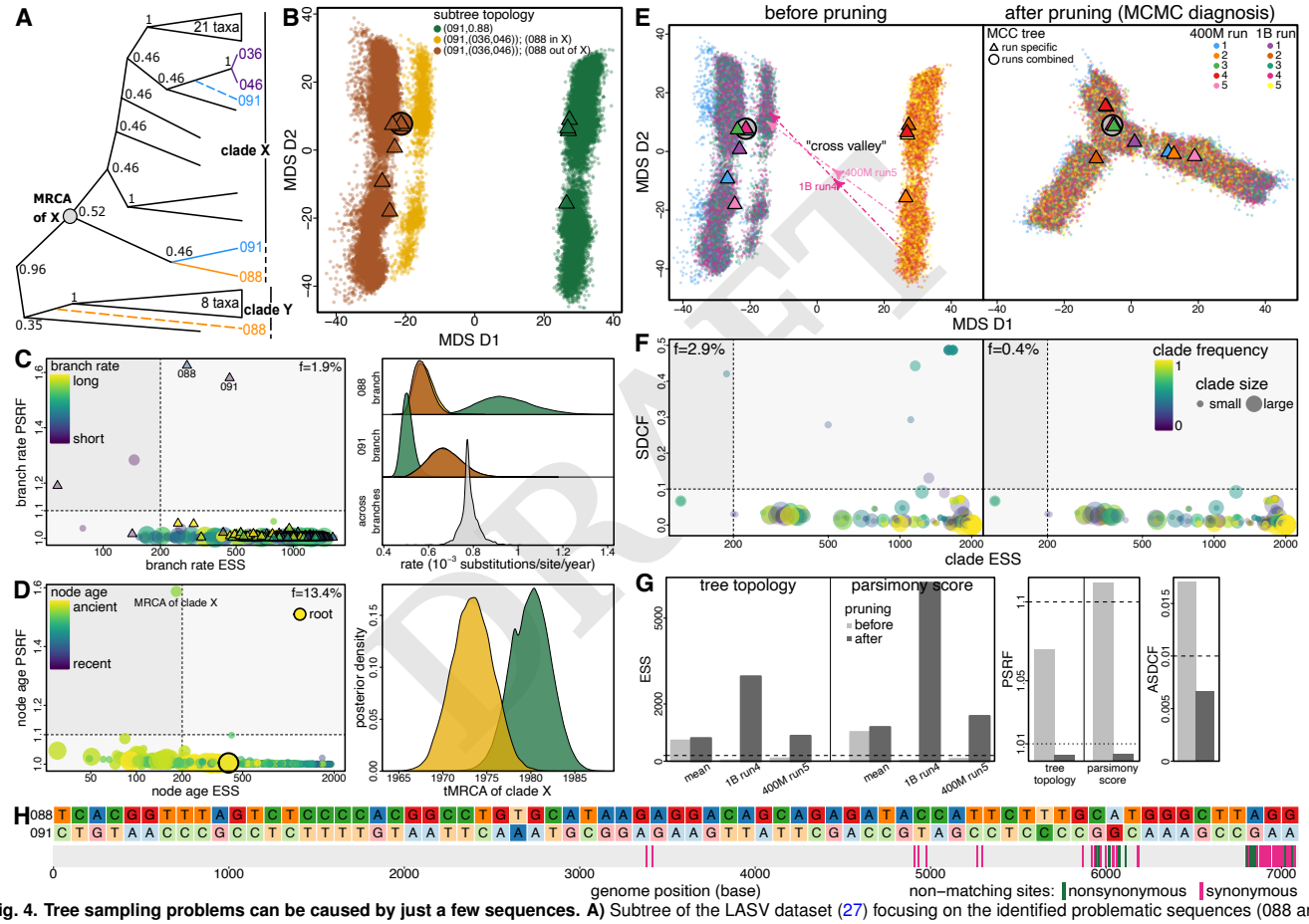

Fig. 4. Tree sampling problems can be caused by just a few sequences. A) Subtree of the LASV dataset (27) focusing on the identified problematic sequences (088 and 091). Dashed branches indicate the alternative positions of 088 and 091 when they are not sisters. The sequence names (*e.g.*, "088") used here are shorthands of their full names, following how they were referred to in the original study (41); see Suppl. Table S2 for the full names and GenBank IDs. **B)** MDS representation of the posterior tree space (based on RF distance) with each tree colored by the topology of its focal subtree. When 088 and 091 are sisters (green), they are always basal to clade X. When 091 is in turn sister to (036,046) (purple in **A**), 088 alternates between the base of clades X and Y (yellow and brown). **C)** Evolutionary rate estimates over the branches subtending these two sequences are largely determined by the subtree topology. (Left) MCMC diagnostics computed for the rate over each branch. PSRF (*y*-axis) and ESS (*x*-axis) quantify between-chain convergence and within-chain mixing, respectively. Dashed lines indicate the conventional cutoffs; PSRF above 1.1 (horizontal dashed line) indicates lack of convergence and ESS below 200 (vertical dashed line) indicates inadequate mixing. Circles and triangles represent internal and external branches, respectively, with color indicating the inferred branch rate value (yellow: high; purple: low) and size proportional to the number of descendant tips. A small fraction ($f = 1.9\%$) of the dots reside outside the bottom-right quadrant surrounded by the dashed lines, indicating that those branch rates fail to be estimated reliably. (Right) Posterior distribution of the branch rate (top: 088; middle: 091) conditioned on the subtree topology (colored same as in **B**). **D)** tMRCA estimates can also be affected by tree sampling problems. (Left) PSRF of the tMRCA of clade X reveals convergence failure. (Right) The tMRCA of clade X is inferred to be significantly younger when 088 and 091 are sisters than when they are not. **E)** Posterior tree space before (Left) and after (Right) pruning the two problematic sequences, with the dot colored by the chain index (*i.e.*, the dots in Left of **E** are identical to those in **B** other than their colors). (Left) Before pruning, at least two peaks were present, and the sampler only crossed the valley once in two (1B run4 and 400M run5; dashed arrows) of the 10 chains, with the MCC trees of six chains in the left peak and the other four in the right peak. (Right) The valley segregating the peaks vanishes after pruning, and the MCC tree summarized from all chains combined moves close to the center. **F)** Clade occurrence diagnostics before (Left) and after (Right) the pruning. SDCF (*y*-axis) and clade ESS (*x*-axis) quantify between-chain convergence and within-chain mixing, respectively. Each dot represents a clade, whose color indicates the clade frequency averaged over all chains. **G)** Tree-based diagnostics quantify MCMC performance before (light gray) and after (dark gray) the pruning, corroborating the visual pattern in (**E**). For the tree and parsimony-score ESSs (Left), we show both the average over all chains ("mean") and the values under the two chains where the valley was crossed once before pruning. **H)** Sites that are incompatible between 088 and 091 and their positions in LASV's L segment. (Top) Nucleotide states at the incompatible sites. Darker cells indicate the mutations inferred to occur on the subtending branch of 088 or that of 091 when they are sisters. (Bottom) Positions of the incompatible sites marked by colored vertical bars (green: nonsynonymous; pink:synonymous).

After removing the two problematic tips, the valley segregating the peaks and the separation among chains in the tree space both vanish (Fig. 4E, Right). For the two chains that previously crossed the valley once, the tree and parsimony-score ESSs both increase by at least an order of magnitude after pruning (Fig. 4G, Left), while the tree and parsimony-score PSRFs, as well as the ASDCF, drop to well below the corresponding cutoffs (Fig. 4G, Middle and Right). The dramatic changes in these diagnostics strongly corroborate the visual contrast seen in the MDS plots (Fig. 4E). Moreover, after pruning, non-converging clades (SDCF above 0.1) no longer exist and the fraction of clades with poor MCMC performance decreases from 2.9% to 0.4% (Fig. 4F), demonstrating the attenuation of sampling problems. The fraction of sites that are associated with the sampling problems also decreases from 2.4% to 0.3%, with all the site-specific PSRFs that were above the cutoff dropping below it (Suppl. Fig. S7D).

To explore the biological causes of the local peaks in this dataset, we examined the incompatible sites between the two problematic sequences and located them in the L segment of LASV. The 71 incompatible sites are distributed extremely unevenly along the segment, where only two of them are in the starting two-thirds of the segment ("$S\frac{2}{3}$") while the remaining 69 occur sporadically in the ending one-third ("$E\frac{1}{3}$")—especially the last 5%—of the segment (Fig. 4H). Furthermore, the per-site Hamming distances between 088 and other taxa in clade X (excluding 091) in $E\frac{1}{3}$ are on average 3 times those in $S\frac{2}{3}$ (Suppl. Fig. S8A). Recombination analyses performed with `GARD` (42) and `3SEQ` (43) both detect recombination in this dataset; `GARD` infers one breakpoint at genomic position 2925, while 342 of the 346 recombinant triplets inferred by `3SEQ` involve 088. In addition, over 10% of sites in either 088's and 091's genome are missing (Suppl. Fig. S8B), and surprisingly, the missing sites predominantly occur at the third codon position. Notably, a sequence from clade Y was used as the reference when computing a consensus sequence for 088 in the original study (41). Together, these lines of evidence suggest that these two problematic sequences might each have been produced by taking a consensus sequence from mixed viral samples due to superinfection or contamination.

Several key features are shared among these datasets, for which removing a few problematic sequences significantly improves the sampling performance. First, a problematic tip usually wobbles between alternative topological positions that are similarly parsimonious and 5–10 nodes apart on the tree, creating peaks in tree space with comparable posterior density. Multiple such tips may lead to distinct peaks combinatorially; *e.g.*, the three problematic sequences identified in the SARS-CoV-2 Origin dataset (32) give rise to at least eight peaks (Suppl. Fig. S9). This SARS-CoV-2 dataset also demonstrates that, due to its extremely limited genetic diversity, even a single recurrent mutation (either bona fide or introduced by sequencing error) can make the sequence problematic for tree inference, behaving like a recombinant. Second, the branch-specific variables associated with the problematic tip, such as the time span or evolutionary rate, tend to vary significantly between the alternative topological positions. The clades as well as the variables associated with the branches in the vicinity of the problematic tip can also be strongly distorted by its wobble. While these behaviors hinder efficient exploration of tree space, they provide a way of detecting the problematic sequences via the clade-specific diagnostics, in particular the branch-rate (and -length) diagnostics, which are more interpretable than the clade ESS or SDCF as one problematic sequence can distort the sampling of many clades while simultaneously rendering them to be poorly supported. Finally, when billions or more iterations are on average required to commute between the peaks, the chains will appear to be stuck on different peaks and thus exhibit lack of convergence; when the commute is easier and occurs more frequently (but not frequently enough), the outstanding sampling problem will instead be inadequate mixing within each chain.

Limited Success of Root-to-Tip Regression and Recombination Tests in Identifying Problematic Sequences. We have demonstrated that clade-specific diagnostics can be used to identify problematic sequences in the five datasets where fewer than 10 such sequences exist. However, for the other seven datasets where more clades exhibit poor sampling (despite still being a small fraction, ≈10%, of all the clades), in which case a stable backbone topology cannot be established from the sampled trees, we are unable to pinpoint the key problematic sequences via MCMC diagnosis. In addition, although shorter MCMC chains might be sufficient for the purpose of detecting problematic sequences, hours or even days of computation would be required to perform this preliminary check.

Accordingly, we investigate whether problematic sequences can be identified without running the phylodynamic analyses but with simpler and faster data-quality tests. To that end, we conducted the root-to-tip (RTT) regression (44) and recombination tests with each dataset and compared the detected outlier and recombinant sequences with those identified by the MCMC diagnostics. The RTT regression—a test for assessing the temporal signal in pathogen genetic data and identifying outlier sequences whose sampling dates or genetic composition are incongruent with the rest—reveals strong temporal signal in 11 of the 15 datasets (Suppl. Fig. S19), among which nine have been identified with tree sampling problems. The RTT regression detects outliers in all these nine datasets, seven of which with more than 10.

However, despite the large number of RTT outliers, pruning them from the sampled trees has *little* impact on the sampling performance of tree or other parameters (Suppl. Figs. S10–S18), considerably less than that under the MCMC-diagnosis approach for the datasets where both approaches are applicable (Suppl. Figs. S10–S12); the RTT outliers barely overlap with the problematic sequences identified by MCMC diagnosis. The only exception is the rabies virus (RABV) dataset, where five of the seven RTT outliers have also been identified to be problematic by MCMC diagnosis, and both approaches lead to significantly improved sampling performance, although the improvement under the MCMC-diagnosis approach is more substantial (Suppl. Fig. S12).

To explore the potential for recombination tests to identify the problematic sequences, we analyzed each dataset with `3SEQ`, a quick method that examines all sequence triplets for potential recombinants without performing any phylogenetic inference. `3SEQ` detects recombination in six datasets, five of which have been identified with tree sampling problems. In three datasets pervasive recombinations were detected by `3SEQ`; *e.g.*, at an extreme, 346 and 271 putative recombinants are found in the LASV and RABV datasets, respectively, accounting for 62.8% and 93.4% of all sequences in each dataset. Additional tests with `GARD`, a likelihood-based phylogenetic method, also detect recombination in these two datasets, confirming the strong recombination signal. However, the high percentages of

DRAFT

putative recombinants in these two datasets pose challenges on how to properly purge the putative recombinants to avoid their impacts on downstream phylodynamic analysis. Interestingly, usually a small set of sequences are inferred to be the putative parents of most recombinant triplets detected by `3SEQ`. For example, the problematic sequence 088, which led to the sampling problems of the LASV dataset, is involved in 98.8% of the detected recombinant triplets. Similarly, one of the problematic sequences in the RABV dataset that has been identified by MCMC diagnosis is involved in 72.7% of the triplets. These results suggest that such quick recombination tests may detect key recombinants, making them a useful preliminary check in phylodynamic analysis.

Impact of Tree Space Ruggedness on Phylodynamic Inferences. To assess the biological impacts of tree sampling problems, we focus on three focal phylodynamic inferences, including the 1) effective population size over time (*i.e.*, demographic trajectory), 2) geographic dispersal dynamics, and 3) tMRCAs and evolutionary rates. Many `BEAST` phylodynamic analyses use the skygrid non-parametric coalescent model (45), which divides the history into time intervals and estimates the effective population size of each interval, constituting the temporal demographic trajectory. The demographic trajectory is a key epidemiological parameter, upon which conclusions about the scale and timing of outbreaks can be drawn.

We find that the inferred demographic trajectories are *barely distinguishable* among the chains, except for a few datasets where the inferred curve is shifted to be slightly older or younger (Suppl. Fig. S20). Moreover, for each dataset, the demographic trajectory—including both its accuracy and precision—summarized from the first 10 million post-burnin iterations of each chain (2% of the full analysis) is virtually identical to that summarized from the full analysis (Suppl. Fig. S21). Consequently, if the demographic trajectory is the primary interest, then MCMC chains that are orders of magnitude shorter than those needed for adequately sampling the tree would be sufficient.

By contrast, the geographic phylodynamic estimates can be sensitive to tree sampling problems. For eight datasets where the sampling locations are available, we assessed the impact on the inferred number of dispersal events across all areas as well as between each pair of areas by reconstructing a parsimonious dispersal history conditioning on each posterior tree. Although most of the event numbers are robust to tree sampling problems (Suppl. Fig. S22), the inferred numbers over several biologically important dispersal routes differ significantly among the chains (Fig. 5 and Suppl. Figs. S23–S28). For example, our analyses of the HIV dataset reveals that the number of dispersal events inferred between two focal geographic areas (Bwamanda and Kinshasa) in the early spread history of HIV (26) exhibits large discrepancies among the chains (Fig. 5): for some chains (*e.g.*, 400M run4) the median estimate is 3 while the 95% CI includes no more than 5 events, while for some others (*e.g.*, 1B run3) the median estimate is 8 and the posterior probability for there being fewer than 5 events is smaller than 5%. Therefore, in this case tree sampling problems directly lead to convergence failure in the inferred dispersal history.

For most datasets, the root age (Suppl. Fig. S29) and the tMRCA of most clades appear to be sampled satisfactorily (Suppl. Fig. S6, Bottom), while the mean evolutionary rate estimates differ significantly among the chains in five of the 15 datasets (Suppl. Fig. S31). Both the root age and mean evolutionary rate require at least 100 million post-burnin iterations from each chain to reach the accuracy and precision comparable with those under the full analysis (Suppl. Figs. S30 and S32), much longer than the chain length required for the demographic trajectory. Similar to the geographic inference, the tMRCA estimates of certain clades can also be sensitive to tree sampling. For example, in the HIV dataset, the posterior median and 95% CI of the root age and origin time of HIV subtype B differ significantly among chains (Suppl. Fig. S33). Consistent with our discovery that tree sampling problems usually stem from a small part of the tree, the biological results here suggest that the ruggedness and consequent sampling problems tend to impact the clade-specific estimates (*e.g.*, clade tMRCA and geographic source) more strongly than those global parameters that are largely determined by the general tree shape and lineage fluctuation pattern (*e.g.*, the demographic trajectory).

Discussion

Bayesian phylodynamics integrates complex biological models and heterogeneous information with various degrees of uncertainty, but is often computationally challenging for modern datasets. Recent computational tools (46, 47) developed for outbreak phylogenetics have primarily focused on accelerating

DRAFT

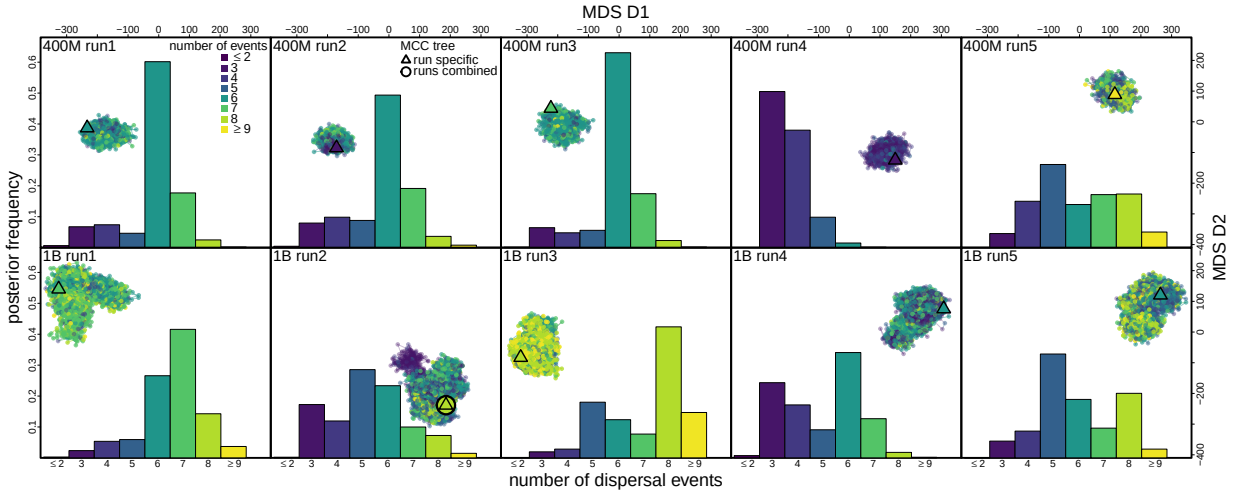

Fig. 5. Distinct peaks in tree space can lead to different biological conclusions. The number of dispersal events inferred between two focal geographic areas (Bwamanda and Kinshasa) in a study investigating the early spread history of HIV (26). Each panel corresponds to an MCMC chain, with the top row shows the posterior of each 400M run and the bottom row shows that of each 1B run. The dots in each panel visualize the MDS tree space; each dot represents a posterior tree, colored according to the parsimony number of dispersal events (η) over this particular route on that tree. A triangle represents the MCC tree of each chain, while an open circle indicates the MCC tree summarized across all chains. The histogram in each panel summarizes the posterior distribution of η in that chain.

the evaluation of each tree. The orthogonal computational challenge in phylodynamics posed by the need to explore a vast and complex tree space has yet to be assessed. Another critical question is how well the underlying tree needs to be sampled for reliable inference of key phylodynamic parameters. Here, to address these questions that are crucial for scalable phylodynamics, we perform a comprehensive empirical survey to characterize phylodynamic tree space and tree sampling performance as well as their biological impacts.

First, we show that phylodynamic trees are different from the phylogenetic trees for which tree sampling behavior has been explored. Phylodynamic trees are usually markedly imbalanced, composed of many weakly supported nodes subtended by short branches, making them distinct from the typical highly supported ultrametric species trees inferred from phylogenomic datasets. Despite high phylogenetic uncertainty, the largely diffuse phylodynamic tree space comprises many rugged areas that are rich in local peaks, resulting in widespread tree sampling problems including inadequate mixing within each MCMC chain and lack of convergence among the chains (Fig. 3). The greater prevalence of tree sampling problems observed here compared to that reported in a previous study (14) quantifying phylogenetic MCMC performance across alignments of divergent species is consistent with a statement therein that the convergence is more difficult with more taxa and shorter branches. The ruggedness in tree space has frequently been reported in phylogenetic literature (9, 19, 20), but it may seem counterintuitive given the limited genetic diversity in phylodynamic datasets.

Indeed, we show that the limited genetic diversity may well be a main driver of the ruggedness and widespread sampling problems, as the substantial ambiguity greatly inflates the tree space (9), inducing terraces (10) everywhere and wide and nearly flat basins separating the local peaks. Moreover, when many closely related sequences differ by at most one nucleotide (*e.g.*, in the SARS-CoV-2 datasets), we demonstrate that a single recurrent mutation can create extra peaks in the tree space that are small but enough to cause sampling problems.

Dissecting tree sampling problems via existing and newly-developed clade-specific MCMC diagnostics, we demonstrate that usually a small fraction of the clades are poorly sampled (Suppl. Fig. S6) and thus the sampling problems arise from a small part of the tree. We then show that a few sequences causing the poorly sampled clades disproportionately contribute to the ruggedness and sampling problems. Specifically, in one third of the datasets with sampling problems, simply removing 1–3 problematic sequences identified by the clade-specific diagnostics drastically reshapes the tree space—where the valleys segregating the peaks vanish—and largely resolves the sampling problems (Fig. 4 and Suppl. Fig. S9–S11). Although previous studies (48, 49) have reported that a small number of genes can strongly influence phylogenetic inference, the undue influence of a few sequences on tree space complexity and tree search difficulty has not been revealed.

We also demonstrate that tree sampling problems can conspicuously impact key biological inferences. Although the demographic trajectory, root age, and tMRCA of most clades are usually robust to tree sampling problems, estimates of the tMRCA and branch-specific evolutionary rate associated with certain clades vary significantly between the trees in different peaks in the tree space (Fig. 4), leading to discrepant estimates of the mean evolutionary rate in at least five datasets (Suppl. Fig. S31). Similarly, despite the overall satisfactory sampling of the number of geographic events over most dispersal routes, the inferred geographic history over biologically important dispersal routes can be strongly influenced by tree sampling problems (Fig. 5). These biological results indicate that tree space ruggedness tends to impact local (*i.e.*, clade-specific) estimates more strongly than the global parameters that depend more on the general tree shape, consistent with our discovery that a small part of the tree leads to the sampling problems.

Importantly, the sampling problems may threaten to distort major biological conclusions, especially when the clades of primary interest are poorly sampled. For example, in our analyses of the HIV dataset, in five of the 10 chains the origin time of HIV-1 subtype B is inferred to be around 1960—consistent with the time estimates reported in previous studies (26) and the hypothesis that the subtype B was seeded by the Haitian professionals who worked in the DRC after its independence in 1960—whereas that time estimate in the other five chains is significantly earlier than 1960 (Suppl. Fig. S33), challenging that hypothesis. Moreover, the numbers of geographic events over at least six major dispersal routes involving Kinshasa, the putative origin location of HIV-1 group M (26), differ dramatically among the chains (Fig. 5 and Suppl. Figs. S24–S28), indicating distinct early spread history of HIV-1. Poor tree sampling may also mislead the testing of key epidemiological hypotheses that depend upon certain phylogenetic relationships. A focal aim of the original study of the LASV dataset (41) is to identify potential human-to-human transmission pairs—as distinguished from zoonotic spillovers—among the inferred phylogenetic sisters based on the number of substitutions between them; the non-converging sister relationship between 088 and 091 revealed here and the extremely uneven substitution distribution along their genomes (Fig. 4) raise the concern of spurious transmission pairs produced by tree sampling problems.

We confirm that traditional and tree-based MCMC diagnostics are a useful tool for quantifying tree sampling performance in phylodynamic inference; however, care is needed in applying cutoffs. The MCMC diagnostics tend to be more conservative than the MDS plot in revealing tree sampling problems; this may be attributed to the (mis)use of the rule-of-thumb cutoffs, which have been suggested to be overly lenient for assessing tree sampling (15, 16). Moreover, we suspect that the power of tree diagnostics can be undermined by the diffuseness of phylodynamic tree space, where tree distance saturates quickly across MCMC iterations, resulting in misleadingly small autocorrelation (*i.e.*, high ESS) and indistinguishable within- and between-chain variations (*i.e.*, low PSRF). Conversely, the ASDCF suggests more pervasive tree convergence failure compared to the MDS plot or tree PSRF, consistent with a previous report (14). However, the traditional cutoff (40) may be less applicable when most clades are weakly supported (see also 15). Thus, although the diagnostics allow an automated analysis pipeline, we advocate complementing them with visual inspection of the posterior tree space via the MDS plot.

The diagnostics computed with the continuous variables reveal strong discrepancies among the sampling performance of various model parameters. While the diagnostics of root age, population size, and substitution model parameters almost never exhibit any poor sampling, the least satisfactory diagnostics across all continuous variables (frequently the branch-rate model parameters, tree length, or posterior density) usually indicate worse mixing compared to the tree,

DRAFT

suggesting additional problems with sampling the rates and times. This also shows the limited power of existing variable diagnostics in assessing tree sampling, consistent with previous studies (16, 17).

Here we propose new diagnostics based on a classical statistic—the parsimony score—and find them effective in identifying tree sampling problems. These new diagnostics are model-independent and avoid expensive tree comparisons. We recommend adding them to the standard diagnostic set especially when the topology is a focal parameter.

The distinct sampling efficiencies across various parameters and the topology also suggests room for adapting MCMC settings to answer questions of interest for a given phylodynamic study. Indeed, as demonstrated in our analyses, although the billion-iteration-long analyses are still frequently far from adequate in sampling the tree, merely 10 million (post-burnin) iterations are usually sufficient to obtain reliable estimates of the demographic trajectory. In principle, although all parameters depend on the tree, many distinct trees may correspond to virtually identical values for a given parameter such that an inadequate or non-converging marginalization over trees may well appear inconsequential in estimating that parameter (Fig. 1, right panel). In practice, instead of simply checking the ESSs computed with the variables logged by default, or going to the other extreme, exhaustively examining every possible diagnostic, empirical studies may focus on diagnosing the parameters underlying their main biological conclusions. When the biological interests concern certain topological relationship or clades, we advocate for applying the corresponding clade-specific diagnostics.

When the focal estimates are affected by tree sampling problems, our discovery that removing a few problematic sequences can significantly improve tree sampling performance highlights a potential solution. We have demonstrated that the clade-specific diagnostics developed here are useful in identifying the problematic sequences, but they can only be applied after the costly phylodynamic analysis and are only effective when the number of such sequences is relatively small. An efficient and versatile method for detecting the problematic sequences would be very useful. Unfortunately, our exploration of using quick and simple data-quality tests for that purpose showed limited success.

The weak association between the outliers detected by these tests and the problematic sequences identified under the MCMC-diagnosis approach may reflect heterogeneous origins of the problematic sequences. Some of these sequences may be true recombinants, making the recombination tests (*e.g.*, `3SEQ`) useful, while some others may result from mismatches between sequences and their metadata, which could potentially be caught by the RTT regression. Errors introduced by sequencing or bioinformatic pipelines may give rise to some other problematic sequences, which are hard to detect without the raw data. These problematic factors are ever more concerning for phylodynamic inference as empirical studies perform increasingly comprehensive analyses leveraging all available sequences that have been sporadically collected from various outbreaks and generated by different sequencing and bioinformatic procedures with varying error rates (6, 50).

Accordingly, instead of simply removing the detected outliers, we advocate for re-examining the raw data—such as the raw reads and the bioinformatic pipeline generated the sequences—to check for potential errors. If no clear error can be found, or if these outlier sequences are of particular biological interest, we suggest running the phylodynamic analysis with and without them as a sensitivity test to ensure the robustness of key estimates. Alternatively, if the problematic sequences are only identified by the clade-specific diagnostics after a phylodynamic analysis where certain focal parameters are sampled poorly, studies may follow our approach here to confirm the contribution of the identified sequences to the sampling problems by pruning them from the posterior trees and re-assessing the sampling performance. Once confirmed, the analysis can be rerun with those sequences corrected (or removed) by using the (pruned) last sampled tree as a starting tree. To facilitate this process, if resources allow, we suggest running many (*e.g.*, 10) checkpointed (51) replicate chains concurrently and performing MCMC diagnosis before they run too long (after passing the burnin). If the identified sequences only cause sampling problems but are not erroneous or recombinants, instead of rerunning the full analysis, studies may pursue a sequential approach, running a second analysis to re-infer the phylodynamic parameters marginalizing over the pruned trees (26, 52, 53).

Indeed, sequences may happen to be equivocally supported at various positions in the tree and only appear problematic due to the difficulty of moving between those positions. These sequences may be perceived as "rogue" taxa (54)—taxa with varying positions across trees that are similarly good—and many methods (55–57) have been developed for detecting them from the inferred tree set. However, all these studies focus on pruning the rogues to obtain a better consensus tree, while their impact on tree search has yet to be explored. Investigating the paths and conditions that allow commutes between the peaks will bring to light how to make these commutes more frequently, which is not only crucial for Bayesian inference, but may also benefit maximum-parsimony and maximum-likelihood estimators to avoid getting trapped by local peaks and find the tallest peak more efficiently.

Some problematic sequences only require a single SPR move to commute between the alternative positions, but that specific move is proposed infrequently as both the subtree to prune and the branch to regraft on are selected using simple random proposals; that is, when a dataset contains 10 times more sequences, a specific tree move needs to wait 100 times more iterations for it to be proposed. Rather than simple random proposals authors have proposed extending (40), guided (58, 59), or adaptive (60) tree proposals. However, even when the needed move is by chance proposed, the acceptance rate tends to be very low, contradicting the equivalent posterior densities among the peaks. This contradiction may be explained by our discovery that the branch duration and rate associated with a problematic sequence usually vary significantly between the alternative positions; the default tree proposals only attempt to alter the topology (while maintaining the node ages) but not jointly modify the rates associated with the affected branches, resulting in rejections due to the mismatch among branch-specific variables instead of bad topology. A tree proposal that simultaneously adjusts the rates on the branches affected by the topological change—aiming to preserve the product of rate and time on those branches—would likely increase the acceptance rate; such joint proposals have been recently proposed (61, 62) and deserve more attention. In addition, as revealed by our LASV results, the trees between certain peaks are sometimes separated by multiple SPR moves, in which case the needed commute would be more easily achieved if a set of tree moves could be applied

DRAFT

jointly. Similar to the classic 'DS1–DS11' datasets (40) that have been widely used as benchmarks in phylogenetic studies, we believe that the datasets compiled here and the large number of posterior samples will facilitate the benchmarking and training of new phylodynamic inference algorithms, including some implementing the solutions suggested above.

Our study highlights the need to develop and adopt best practices in phylodynamic analysis, and sheds light on methodological advances towards scalable phylodynamics. We have comprehensively characterized the ruggedness in phylodynamic tree space and dissected tree sampling problems, culminating in a clear path towards efficient tree space exploration. We are optimistic that Bayesian phylodynamic inference—with continuing advances in rigorous analysis pipeline and robust computational tools—will be increasingly instrumental in pathogen biology studies and infectious disease control.

Materials and Methods

We briefly summarize the methods and analyses here, and provide more details, as well as supplementary results, in SI Appendix.

Data Curation and Phylodynamic Analyses. To compile the empirical datasets (except the SARS-CoV-2 ones), we obtained the sequence alignment and sampling times from the `BEAST` XML files provided in the original studies. For the SARS-CoV-2 datasets, we acquired the sequences from GISAID (63) according to their IDs. We performed Bayesian phylodynamic analyses of each dataset by running 10 independent MCMC replicate chains using `BEAST` v1.10.5 (35) with `BEAGLE` v3.2.0 (64) enabled, including five 400-million-iteration chains intended to imitate the best practice in most empirical studies and five 1-billion-iteration chains aimed to establish the "ground-truth" posterior distribution. In these analyses, we specified models and priors, as well as the MCMC proposals and their weights, as in the original studies; the proposals and weights are mostly identical to the defaults in `BEAUti` v1.10.5 under the specified models. We discarded a conservatively long burnin (first 200 million iterations) from each chain to focus on the post-burnin behavior, and sampled every 100 thousand iterations.

Tree Space Characterization and MCMC Diagnosis. We characterize the MCMC performance and explore the biological impacts by examining three types of variables, including: 1) continuous model parameters (*e.g.*, evolutionary rate), statistics (*e.g.*, tree length), and probability densities (*e.g.*, joint posterior); 2) clade-specific (*e.g.*, node age) and site-specific (*e.g.*, site likelihood) variables, and; 3) the tree. For each continuous variable, we assessed its within-chain mixing by effective sample size (ESS), computed for each chain (using the post-burnin samples) with an `R` (65) implementation of `Tracer`'s (12) ESS calculation routine provided by the `essTracer` function in the `convenience` package (15). To report a mean ESS across all chains for each variable, we truncated the five 1B runs so that all the chains have identical length. We assessed the between-chain convergence by potential scale reduction factor (PSRF; 39), computed with the `Rhat` function (66) in the `rstan` (67) `R` package.

We used `TreeSummary` (a `BEAST` tool) to post-process the posterior trees to generate the clade-specific variable samples. To compute the alignment-wise and site-specific parsimony scores given each sampled tree, we used the `parsimony` function in `R` package `phangorn` (68). We computed the branch-specific parsimony scores by first calling `phangorn`'s `ancestral.pars` function to reconstruct the parsimonious mutation history for each site given the tree, and then for each branch counting the parsimony number of mutations across all sites. Similar to the alignment parsimony score, we obtained the total number of parsimony dispersal events over each tree using the `parsimony` function. The number of pairwise parsimony dispersal events is computed by first reconstructing the parsimonious geographic history given each tree using the `ancestral.pars` function, and then counting the number of events between each pair of areas.

We characterize the posterior tree space visually and quantitatively with three types of summaries, including: 1) the MCC summary tree, 2) a two-dimensional MDS plot, and 3) tree MCMC diagnostics. We summarized the MCC tree of each dataset from the sampled trees combined across all chains using `TreeAnnotator` v1.10.5 (69); we also summarized the MCC tree from each chain as well. We calculated the pairwise RF distances between rooted trees by counting the number of incompatible clades (instead of splits) obtained via `ape`'s (70) `prop.part` function, and the pairwise rSPR distances using `rspr` (38, 71). We computed the ASDCF by first obtaining the SDCF of each clade, calculated as the standard deviation of the estimated frequencies of that clade among all chains, and then taking the mean over all clades (40).

Assessing Impact of the Problematic Sequences. We used `TreePruner`, a `BEAST` tool that takes the posterior trees as its input, to prune the identified problematic sequences from each tree. To perform the RTT regression (44), we first conducted maximum-likelihood (ML) phylogenetic inference of each dataset using `IQ-TREE` v2.3.2 (72). We then rooted the inferred ML tree by minimizing the root mean squared residuals (RMS), using `ape`'s `rtt` function. Finally, we conducted a linear regression between the estimated RTT genetic distances and the corresponding sampling times using `R`'s `lm` function. We analyzed each dataset with two recombination detection methods, `GARD` (42) and `3SEQ` (43). The `GARD` analyses were run with HyPhy v2.5 (73) under the default settings. We manually parallelized the `3SEQ` analyses so that the sequences could be evaluated concurrently, demonstrating the practicality of such test for serving as a quick check.

Data, Materials, and Software Availability. All data and code necessary to reproduce our results are available in the GitHub repository (https://github.com/jsigao/ssstree_supparchive) and the Zenodo repository (https://doi.org/10.5281/zenodo.15574519) .

ACKNOWLEDGMENTS. We thank Andrew Magee for valuable discussions. This project was partially supported by US National Institutes of Health grants R01 AI162611 and R01 AI153044. Scientific Computing Infrastructure at Fred Hutch was funded by ORIP grant S10OD028685. Dr. Matsen is an Investigator of the Howard Hughes Medical Institute. G.B. acknowledges support from the Research Foundation - Flanders ("Fonds voor Wetenschappelijk Onderzoek - Vlaanderen," G0E1420N, G098321N), from the European Union Horizon 2023 RIA project LEAPS (grant agreement no. 101094685), and from the DURABLE EU4Health project 02/2023-01/2027 which is co-funded by the European Union (call EU4H-2021-PJ4) under Grant Agreement No. 101102733.

1. E Volz, et al., Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell* **184**, 64–75 (2021).
2. JT McCrone, et al., Context-specific emergence and growth of the sars-cov-2 delta variant. *Nature* **610**, 154–160 (2022).
3. L du Plessis, et al., Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
4. J Gao, MR May, B Rannala, BR Moore, New phylogenetic models incorporating interval-specific dispersal dynamics improve inference of disease spread. *Mol. Biol. Evol.* **39**, msac159 (2022).
5. T Stadler, OG Pybus, MP Stumpf, Phylodynamics for cell biologists. *Science* **371**, eaah6266 (2021).
6. Y Turakhia, et al., Stability of SARS-CoV-2 phylogenies. *PLoS Genet.* **16**, e1009175 (2020).
7. B Morel, et al., Phylogenetic analysis of SARS-CoV-2 data is difficult. *Mol. Biol. Evol.* **38**, 1777–1791 (2021).
8. GW Hassler, et al., Data integration in Bayesian phylogenetics. *Annu. review statistics its application* **10**, 353–377 (2023).
9. C Whidden, FA Matsen IV, Quantifying MCMC exploration of phylogenetic tree space. *Syst. Biol.* **64**, 472–491 (2015).
10. MJ Sanderson, MM McMahon, M Steel, Terraces in phylogenetic tree space. *Science* **333**, 448–450 (2011).
11. FF Nascimento, Md Reis, Z Yang, A biologist's guide to Bayesian phylogenetic analysis. *Nat. Ecol. & Evol.* **1**, 1446–1454 (2017).
12. A Rambaut, AJ Drummond, D Xie, G Baele, MA Suchard, Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901 (2018).
13. R Lanfear, X Hua, DL Warren, Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. *Genome Biol. Evol.* **8**, 2319–2332 (2016).
14. SM Harrington, V Wishingrad, RC Thomson, Properties of Markov chain Monte Carlo performance across many empirical alignments. *Mol. Biol. Evol.* **38**, 1627–1640 (2021).
15. LG Fabreti, S Höhna, Convergence assessment for Bayesian phylogenetic analysis using MCMC simulation. *Methods Ecol. Evol.* **13**, 77–90 (2022).
16. A Magee, M Karcher, FA Matsen IV, VM Minin, How trustworthy is your tree? Bayesian phylogenetic effective sample size through the lens of Monte Carlo error. *Bayesian Anal.* **19**, 565 (2024).
17. M Brusselmans, et al., On the importance of assessing topological convergence in Bayesian phylogenetic inference. *Virus Evol.* **10**, veae081 (2024).
18. E Mossel, E Vigoda, Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* **309**, 2207–2209 (2005).
19. RG Beiko, JM Keith, TJ Harlow, MA Ragan, Searching for convergence in phylogenetic Markov chain Monte Carlo. *Syst. Biol.* **55**, 553–565 (2006).

DRAFT

20. JM Brown, RC Thomson, The behavior of Metropolis-coupled Markov chains when sampling rugged phylogenetic distributions. *Syst. Biol.* **67**, 729–734 (2018).
21. DR Maddison, The discovery and importance of multiple islands of most-parsimonious trees. *Syst. Biol.* **40**, 315–328 (1991).
22. G Dudas, et al., Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315 (2017).
23. P Mbala-Kingebeni, et al., Ebola virus transmission initiated by relapse of systemic Ebola virus disease. *New Engl. J. Med.* **384**, 1240–1247 (2021).
24. M Worobey, GZ Han, A Rambaut, A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature* **508**, 254–257 (2014).
25. T Bedford, et al., Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* **523**, 217 (2015).
26. NR Faria, et al., The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346**, 56–61 (2014).
27. R Klitting, et al., Predicting the evolution of the Lassa virus endemic area and population at risk over the next decades. *Nat. Commun.* **13**, 5596 (2022).
28. LH Moncla, et al., Repeated introductions and intensive community transmission fueled a mumps virus outbreak in Washington State. *Elife* **10**, e66448 (2021).
29. M Viana, et al., Effects of culling vampire bats on the spatial spread and spillover of rabies virus. *Sci. Adv.* **9**, eadd7437 (2023).
30. S Dellicour, et al., Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nat. Commun.* **11**, 5620 (2020).
31. ND Grubaugh, et al., Travel surveillance and genomics uncover a hidden Zika outbreak during the waning epidemic. *Cell* **178**, 1057–1071 (2019).
32. JE Pekar, et al., The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. *Science* **377**, 960–966 (2022).
33. DS Candido, et al., Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**, 1255–1260 (2020).
34. P Lemey, et al., Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature* **595**, 713–717 (2021).
35. MA Suchard, et al., Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
36. DM Hillis, TA Heath, KS John, Analysis and visualization of tree space. *Syst. Biol.* **54**, 471–482 (2005).
37. DF Robinson, LR Foulds, Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
38. C Whidden, N Zeh, RG Beiko, Supertrees based on the subtree prune-and-regraft distance. *Syst. Biol.* **63**, 566–581 (2014).
39. A Gelman, DB Rubin, Inferences from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–511 (1992).
40. C Lakner, P Van Der Mark, JP Huelsenbeck, B Larget, F Ronquist, Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Syst. biology* **57**, 86–103 (2008).
41. L Kafetzopoulou, et al., Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science* **363**, 74–77 (2019).
42. SL Kosakovsky Pond, D Posada, MB Gravenor, CH Woelk, SD Frost, GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096–3098 (2006).
43. HM Lam, O Ratmann, MF Boni, Improved algorithmic complexity for the 3SEQ recombination detection algorithm. *Mol. Biol. Evol.* **35**, 247–251 (2018).
44. A Rambaut, TT Lam, L Max Carvalho, OG Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
45. MS Gill, et al., Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013).
46. N De Maio, et al., Maximum likelihood pandemic-scale phylogenetics. *Nat. Genet.* **55**, 746–752 (2023).
47. P Varilly, et al., Delphy: scalable, near-real-time Bayesian phylogenetics for outbreaks. *bioRxiv* pp. 2025–03 (2025).
48. JM Brown, RC Thomson, Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Syst. Biol.* **66**, 517–530 (2017).
49. XX Shen, CT Hittinger, A Rokas, Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat. Ecol. & Evol.* **1**, 0126 (2017).
50. M Hunt, et al., Addressing pandemic-wide systematic errors in the SARS-CoV-2 phylogeny. *bioRxiv* (2024).
51. MS Gill, P Lemey, MA Suchard, A Rambaut, G Baele, Online Bayesian phylodynamic inference in BEAST with application to epidemic reconstruction. *Mol. Biol. Evol.* **37**, 1832–1842 (2020).
52. DD Pollock, DJ Zwickl, JA McGuire, DM Hillis, Increased taxon sampling is advantageous for phylogenetic inference. *Syst. Biol.* **51**, 664 (2002).
53. J Gao, MR May, B Rannala, BR Moore, Model misspecification misleads inference of the spatial dynamics of disease outbreaks. *Proc. National Acad. Sci.* **120**, e2213913120 (2023).
54. M Wilkinson, Majority-rule reduced consensus trees and their use in bootstrapping. *Mol. Biol. evolution* **13**, 437–444 (1996).
55. AJ Aberer, D Krompass, A Stamatakis, Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst. Biol.* **62**, 162–166 (2013).
56. MR Smith, Using information theory to detect rogue taxa and improve consensus trees. *Syst. Biol.* **71**, 1088–1094 (2022).
57. J Klawitter, RR Bouckaert, AJ Drummond, Skeletons in the forest: Using entropy-based rogue detection on Bayesian phylogenetic tree distributions. *bioRxiv* pp. 2024–09 (2024).
58. S Höhna, AJ Drummond, Guided tree topology proposals for Bayesian phylogenetic inference. *Syst. Biol.* **61**, 1–11 (2012).
59. C Zhang, JP Huelsenbeck, F Ronquist, Using parsimony-guided tree proposals to accelerate convergence in Bayesian phylogenetic inference. *Syst. Biol.* **69**, 1016–1032 (2020).
60. X Meyer, Adaptive tree proposals for Bayesian phylogenetic inference. *Syst. Biol.* **70**, 1015–1032 (2021).
61. R Zhang, A Drummond, Improving the performance of Bayesian phylogenetic inference under relaxed clock models. *BMC evolutionary biology* **20**, 1–28 (2020).
62. J Douglas, R Zhang, R Bouckaert, Adaptive dating and fast proposals: Revisiting the phylogenetic relaxed clock model. *PLoS Comput. Biol.* **17**, e1008322 (2021).
63. Y Shu, J McCauley, GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
64. DL Ayres, et al., BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst. Biol.* **68**, 1052–1061 (2019).
65. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), (2024).
66. A Vehtari, A Gelman, D Simpson, B Carpenter, PC Bürkner, Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian analysis* **16**, 667–718 (2021).
67. Stan Development Team, RStan: the R interface to Stan (2024) R package version 2.32.6.
68. KP Schliep, phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
69. G Baele, et al., HIPSTR: highest independent posterior subtree reconstruction in TreeAnnotator X. *bioRxiv* pp. 2024–12 (2024).
70. E Paradis, K Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
71. C Whidden, RG Beiko, N Zeh, Fixed-parameter algorithms for maximum agreement forests. *SIAM J. on Comput.* **42**, 1431–1466 (2013).
72. BQ Minh, et al., IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
73. SL Kosakovsky Pond, et al., HyPhy 2.5——a customizable platform for evolutionary hypothesis testing using phylogenies. *Mol. biology evolution* **37**, 295–299 (2020).

DRAFT