Title:

Online Phylogenetics using Parsimony Produces Slightly Better Trees and is Dramatically More Efficient for Large SARS-CoV-2 Phylogenies than *de novo* and Maximum-Likelihood Approaches

Authors:

Bryan Thornlow[1,2,*], Cheng Ye[3], Nicola De Maio[4], Jakob McBroome[1,2], Angie S. Hinrichs[2], Robert Lanfear[5], Yatish Turakhia[3], Russell Corbett-Detig[1,2,*]

Affiliations:

[1]Department of Biomolecular Engineering, University of California, Santa Cruz; Santa Cruz, CA 95064, USA
[2]Genomics Institute, University of California, Santa Cruz; Santa Cruz, CA 95064, USA
[3]Department of Electrical and Computer Engineering, University of California, San Diego; San Diego, CA 92093, USA
[4]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus; Cambridge CB10 1SD, UK
[5]Department of Ecology and Evolution, Research School of Biology, Australian National University; Canberra, ACT 2601, Australia
*Correspondence to: bthornlo@ucsc.edu and rucorbet@ucsc.edu

Abstract:
Phylogenetics has been foundational to SARS-CoV-2 research and public health policy, assisting in genomic surveillance, contact tracing, and assessing emergence and spread of new variants. However, phylogenetic analyses of SARS-CoV-2 have often relied on tools designed for *de novo* phylogenetic inference, in which all data are collected before any analysis is performed and the phylogeny is inferred once from scratch. SARS-CoV-2 datasets do not fit this mould. There are currently over 5 million sequenced SARS-CoV-2 genomes in public databases, with tens of thousands of new genomes added every day. Continuous data collection, combined with the public health relevance of SARS-CoV-2, invites an "online" approach to phylogenetics, in which new samples are added to existing phylogenetic trees every day. The extremely dense sampling of SARS-CoV-2 genomes also invites a comparison between Likelihood and Parsimony approaches to phylogenetic inference. Maximum Likelihood (ML) methods are more accurate when there are multiple changes at a single site on a single branch, but this accuracy comes at a large computational cost, and the dense sampling of SARS-CoV-2 genomes means that these instances will be extremely rare. Therefore, it may be that approaches based on Maximum Parsimony (MP) are sufficiently accurate for reconstructing phylogenies of SARS-CoV-2, and their simplicity means that they can be applied to much larger datasets. Here, we evaluate the performance of *de novo* and online phylogenetic approaches, and ML and MP frameworks, for inferring large and dense SARS-CoV-2 phylogenies. Overall, we find that online phylogenetics produces similar phylogenetic trees to *de novo* analyses for SARS-CoV-2, and that MP optimizations produce more accurate SARS-CoV-2 phylogenies than do ML optimizations. Since MP is thousands of times faster than presently available implementations of ML and online phylogenetics is faster than *de novo*, we therefore propose that, in the context of comprehensive genomic epidemiology of SARS-CoV-2, MP online phylogenetics approaches should be favored.

Key words:
SARS-CoV-2, phylogenetics, parsimony, maximum likelihood, optimization

Main Text:

The widespread availability and extreme abundance of pathogen genome sequencing has made phylogenetics central to combatting the pandemic. Communities worldwide have begun implementing genomic surveillance, the systematic genetic sequencing of a percentage of local cases (Deng et al. 2020; Lu et al. 2020a; Meredith et al. 2020; Park et al. 2021). This has been invaluable in tracing local transmission chains (Bluhm et al. 2020; Lam 2020), understanding the genetic makeup of viral populations within local communities (Gonzalez-Reiche et al. 2020; Franceschi et al. 2021; Thornlow et al. 2021a), uncovering the means by which viral lineages have been introduced to new areas (Castillo et al. 2020), and measuring the relative spread of specific variants (Skidmore et al. 2021; Umair et al. 2021). Phylogenetic approaches for better understanding the proximate evolutionary origins of the virus (Li et al.), as well as to identify recombination events (Jackson et al. 2021; Turakhia et al. 2021b) and instances of convergent evolution (Kalantar et al. 2020; Peng et al. 2021) have greatly informed our understanding of the virus. Phylogenetic visualization software including Auspice (Hadfield et al. 2018) and Taxonium (Sanderson) have also become widely used for public health purposes.

A comprehensive, up-to-date phylogenetic tree of SARS-CoV-2 is important for public health officials and researchers. A tree containing all available sequences can facilitate identification of epidemiological links between samples that might otherwise be obscured in subsampled phylogenies. Such information can also help to identify the likely sources of new viral strains in a given area (Moreno et al. 2020; Tang et al. 2021). Additionally, using up-to-date information enables us to find and track quickly growing clades and novel variants of concern (Annavajhala et al. 2021; Tegally et al. 2021), as well as to measure the spread of known variants at both global and community scales. This also facilitates naming lineages of interest, which has been especially important in tracking variants of concern during the pandemic (*e.g.* B.1.1.7 or "Alpha" and B.1.617.2 or "Delta") (Rambaut et al. 2020). Having all available mutational, geographic, and temporal information about SARS-CoV-2 also enables more comprehensive research to be done regarding the prevalence of recombination (Turakhia et al. 2021b) and the effects of specific mutations on viral phenotype (Khan et al. 2021; Tian et al. 2021), furthering our collective understanding of the virus.

SARS-CoV-2 presents a unique set of phylogenetic challenges. First, the unprecedented pace and scale of whole-genome sequence data has forced the phylogenetics community to place runtime and scalability at the center of every inference strategy. More than 5 million SARS-CoV-2 genome sequences are currently available, with tens of thousands being added each day. Prior to the pandemic, *de novo* phylogenetics, or approaches that infer phylogenies from scratch, have been the standard, as there has rarely been a need to re-infer or improve pre-existing phylogenies on a daily basis. Re-inferring a tree of more than 5 million samples daily, however, is extremely costly, and has brought a renewed focus on methods for adding new samples to existing phylogenetic trees (Matsen et al. 2010; Berger et al. 2011; Fourment et al. 2018; Barbera et al. 2019). This approach has been called "online phylogenetics" (Gill et al. 2020), and has important advantages in the context of the pandemic and beyond. Online phylogenetics is appealing for the genomic surveillance of any pathogen, because iterative optimization should decrease computational expense, allowing good estimates of phylogenies to be made readily available. .

Second, SARS-CoV-2 genomes are much more closely related than sequences in most other phylogenetic analyses. Because the advantages of maximum likelihood methods decrease for closely related samples and long branches are relatively rare in the densely sampled SARS-CoV-2 phylogeny (Felsenstein 1978; Hendy and Penny 1989; Philippe et al. 2005), this suggests that phylogenetic inferences based on maximum parsimony, a much faster and simpler phylogenetic inference method, could be better suited for online phylogenetic analyses of SARS-CoV-2 genomes (Wertheim et al. 2021). The principle of maximum parsimony is that the tree with the fewest mutations should be favored, and it is sometimes described as a non-parametric phylogenetic inference method (Sullivan and Swofford 2001; Kolaczkowski and Thornton 2004). Additionally, because parsimony-based tree optimization does not require estimation of uncertainty at all positions in the phylogeny like ML optimization does, parsimony uses much less memory.

Here, we evaluate approaches that would enable one to maintain a fully up-to-date and comprehensive global phylogeny of SARS-CoV-2 genome sequences (McBroome et al. 2021). Specifically, we investigate tradeoffs between online and *de novo* phylogenetics and between maximum parsimony and maximum likelihood approaches. We mimic the time-course of the pandemic by introducing increasingly large numbers of SARS-CoV-2 genome sequences proportionately to their reported sampling dates. Results from our comparisons demonstrate that for the purposes of SARS-CoV-2 phylogenetics, in which samples are numerous and closely related and inference speed is of high significance, parsimony-based online phylogenetics applications are clearly most favorable and are also the only immediately available methods

capable of producing daily phylogenetic estimates of all available SARS-CoV-2 genomes (Turakhia et al. 2021a). As similarly vast datasets containing millions of increasingly closely related samples will soon be available for many species and pathogens, we expect that the approaches we evaluate here will become increasingly central to phylogenetic inference.

Results and Discussion:

Online phylogenetics is an alternative to de novo phylogenetics for ongoing studies.

The vast majority of phylogenetics during the pandemic has consisted of *de novo* phylogenetics approaches, in which each phylogeny is inferred using only genetic variation data, and without a guide tree (Fig. 1). This strategy for phylogenetic inference has long been the default, as in most instances in the past, data is collected just once for a project, and more relevant data is rarely going to be made available in the near future. This process is well characterized and has been foundational for many phylogenetics studies (Hug et al. 2016; Parks et al. 2018; Lu et al. 2020b), and most phylogenetics software is developed with *de novo* phylogenetics as the primary intended usage.

A challenging aspect of pandemic phylogenetics is the need to keep up with the pace of data generation as genome sequences continuously become available. To evaluate phylogenetics applications in the pandemic (Fig. 1), we split 233,326 samples dated from December 23, 2019 through January 11, 2021 into 50 batches according to their date of collection. Each batch contains roughly 5,000 samples. Samples in each batch were collected within a few days of each other, except in the first months of the pandemic when sample collection was more sparse. We also constructed a dataset of otherwise similar data simulated from a known phylogeny (see Methods). The intent of this scheme is to roughly approximate the data generation and deposition that occurred during the pandemic. All datasets are available from the repository associated with this project (Thornlow et al. 2021b), for reproducibility and so that future methods developers can directly compare their outputs to our results. We performed online and *de novo* phylogenetics using a range of inference and optimization approaches. Since thousands of new sequences are added to public sequence repositories each day, we terminated any phylogenetic inference approaches that took more than 24 hours, because such phylogenies would be somewhat obsolete by the time they were inferred.

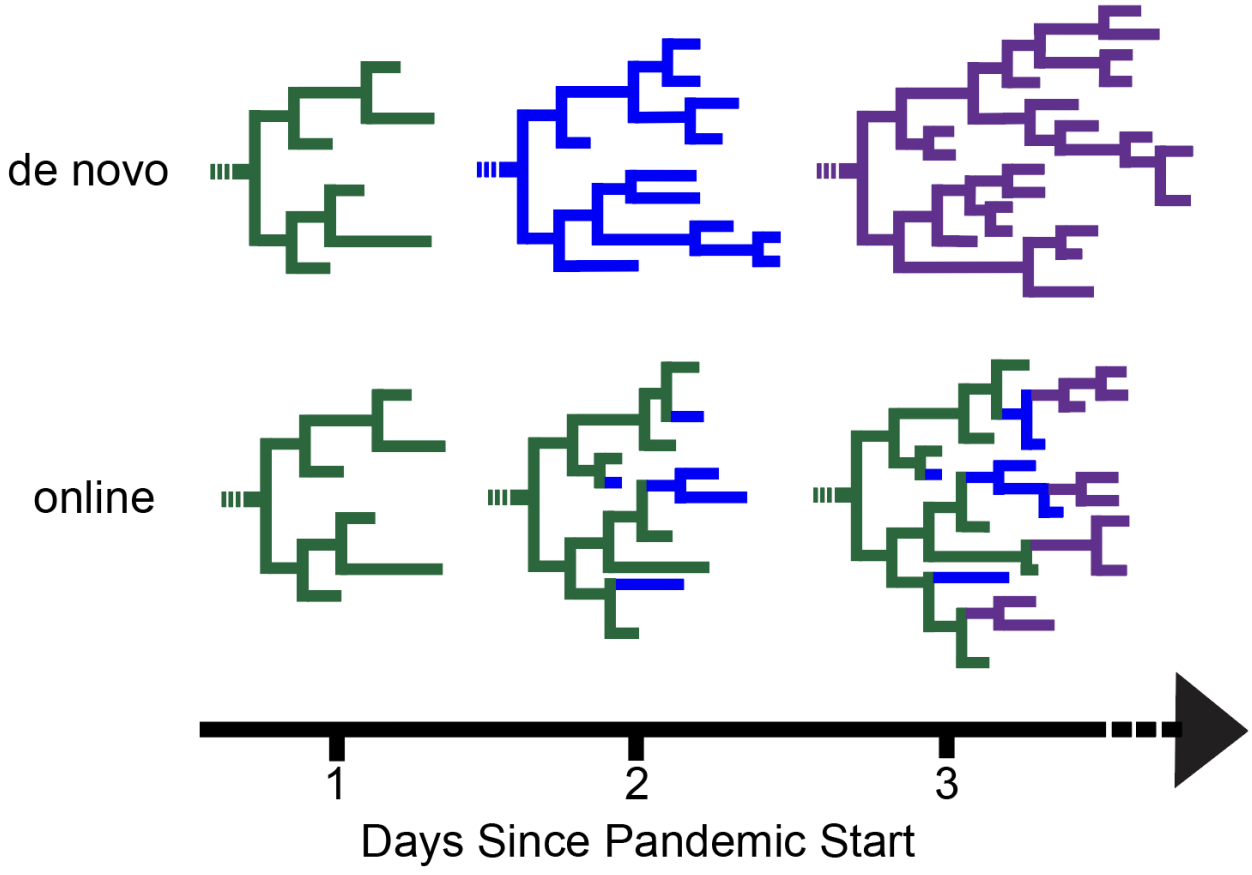

Figure 1: Phylogenies may be optimized from scratch using de novo phylogenetics or iteratively using online phylogenetics. *In de novo phylogenetics (top), trees are repeatedly re-inferred from scratch. Conversely, online phylogenetics (bottom) involves placement of new samples as they are collected along with repeated re-optimization of the tree. Online phylogenetics is also much faster and requires less memory than de novo phylogenetics due to its starting from an already partially optimized tree.*

Analyses using simulated data suggest that online phylogenetics is more accurate.

We first compared matOptimize (commit 66ca5ff) (Turakhia et al. 2021a), IQ-TREE 2 (Minh et al. 2020), and FastTree 2 (Price et al. 2010) using both online and *de novo* phylogenetics strategies using simulated data that we designed to closely mimic real SARS-CoV-2 datasets. All online phylogenomics workflows used UShER (Turakhia et al. 2021a) to add new sequences to the previous tree (see Methods) as to our knowledge it is the only software package that is fast enough to perform under real time constraints. We chose these three tools based on their widespread usage among SARS-CoV-2 phylogenetics applications (*e.g.* matOptimize is part of the UShER suite (Turakhia et al. 2021a), IQ-TREE 2 is used by (COVID-19 Genomics UK (COG-UK) Consortium 2020; Lanfear and Mansfield 2020) and FastTree 2 is used by (Hadfield et al. 2018)) as well as to cover several different methodologies. matOptimize uses subtree pruning and regrafting (SPR) moves to find the tree with the fewest total mutations. IQ-TREE 2 uses nearest neighbor interchange (NNI) and stochastic moves to find the tree with the greatest likelihood given an alignment and substitution model. FastTree 2 uses a pseudo-likelihood approach involving minimum-evolution SPR moves and maximum likelihood NNIs.

Simulating an alignment based on a known tree ensures that there is a ground truth for comparison to definitively assess each optimization method. We used an inferred global phylogeny as a template to simulate a complete multiple sequence alignment using phastSim (De Maio et al. 2021b). We subsampled this simulated alignment into 50 progressively larger sets of samples, ranging in number of samples from 4,676 to 233,326 (see Methods), to examine each of the three optimization methods in both online and *de novo* phylogenetics. We then computed the Robinson-Foulds distance for unrooted trees of each iteration, after condensing identical samples and collapsing very short branches, to the global mutation-annotated tree on which the simulation was based, pruned to contain only the relevant samples, and normalized by the maximum possible Robinson-Foulds distance between the trees (Fig. 2) (Steel and Penny 1993).

All online phylogenetics methods noticeably outperformed their *de novo* counterparts. Overall, online matOptimize produced phylogenies with the lowest Robinson-Foulds distance to the ground truth for the majority of iterations (Fig. 2). Online IQ-TREE 2 performed similarly, but was able to complete only 25 of the 50 iterations due to its extreme computational resource requirements. For example, for the 14th phylogeny, which was the last phylogeny produced using under 200 GB of RAM in under 24 hours by all six methods, we found Robinson-Foulds distances of 1696, 2590, and 2130 for *de novo* matOptimize, FastTree 2, and IQ-TREE 2 respectively, and distances of 1557, 2111, and 1618 for online matOptimize, FastTree 2, and IQ-TREE 2, respectively.

There are several possible explanations for the improved performance of online phylogenetics relative to *de novo* approaches. First, the radius for SPR moves when optimizing a large tree is insufficiently large to find improvements that are more readily applied when the tree contains fewer samples as in early rounds of online phylogenetics. In online phylogenetics, these improvements carry over to subsequent trees, while in *de novo*, they do not. The radius is defined as the phylogenetic distance of the search space when moving a node to a more optimal position. As the phylogeny increases in size, the distance from a node to its optimal position is likely to also increase, necessitating a larger SPR move radius to make equivalent improvements in larger trees. Second, large clades consisting primarily of samples with branch length zero might further reduce the ability of optimization methods to find improvements by indirectly limiting search space due to the increased number of edges when represented internally as a bifurcating tree. It may sometimes be possible to explore moves across such tree regions during online phylogenetics in early iterations when the polytomy is relatively small. Third, online phylogenetics facilitates tree optimization by providing an exceptionally good starting tree that has already been heavily optimized in previous iterations. We expect that this approach will typically outperform parsimony and neighbor-joining starting trees that are used in most *de novo* phylogenetic inference approaches. Finally, it is also possible that UShER produces more optimal starting trees than the other phylogenetics inference packages we evaluated.

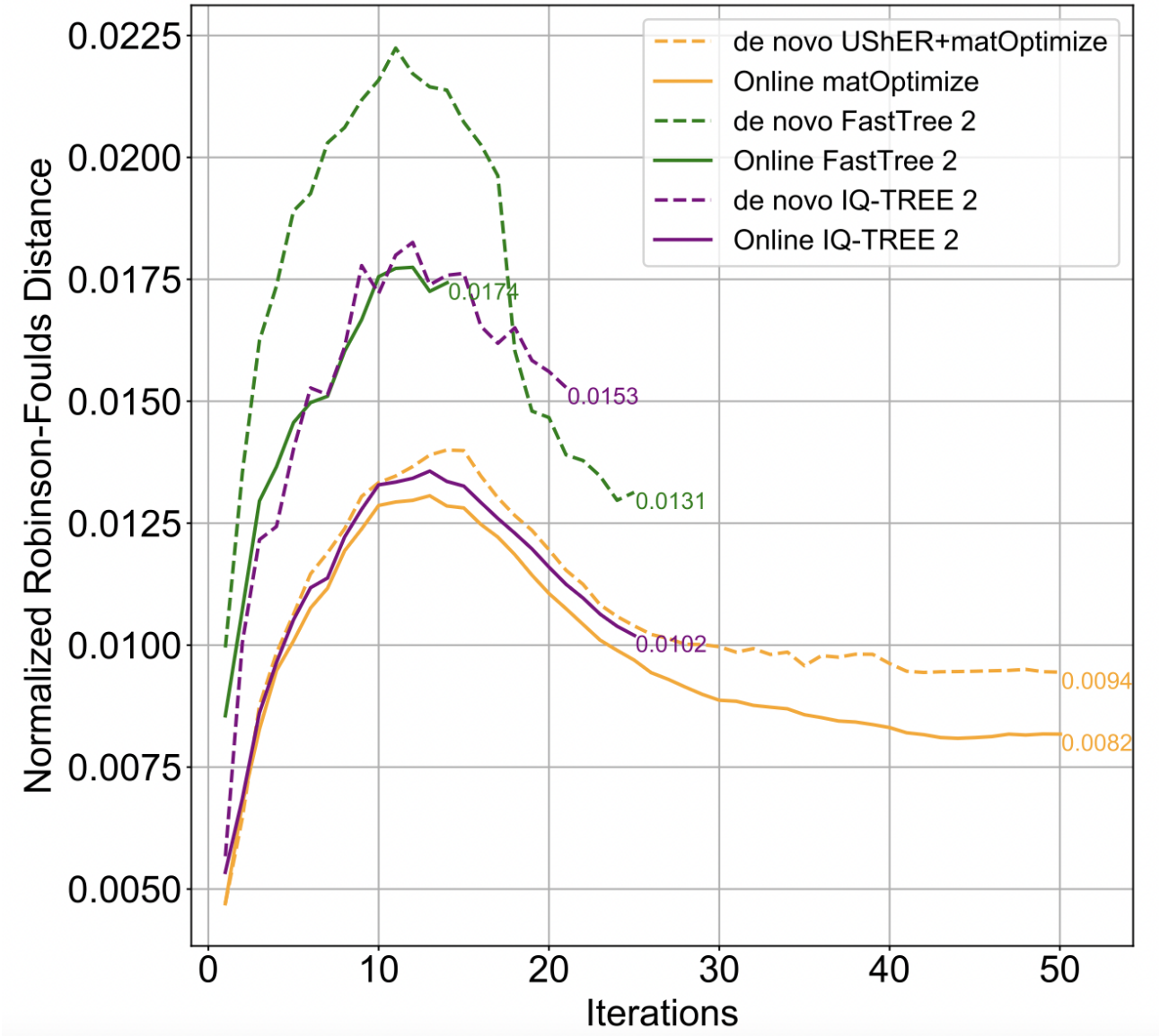

Figure 2: Online matOptimize produces phylogenies most similar to ground truth on simulated data. *For each batch of samples, we calculated the Robinson-Foulds distance between the tree produced by a given optimization software and the ground truth tree pruned to contain only the relevant samples. We then normalized these values by the maximum possible Robinson-Foulds distance between the two trees, which is equal to 2n-6 where n equals the number of samples in each tree (Steel and Penny 1993). We terminated FastTree and IQ-TREE after the first phylogeny that took more than 24 hours to optimize.*

Analyses using real data suggest that online phylogenetics is more efficient than *de novo* and produces similarly optimal phylogenies.

While analyses using simulated data offer the ability to compare to a known ground truth, assessing the performance of each method on real SARS-CoV-2 data may more accurately reflect practical use of each method. Therefore, we also tested each optimization strategy on 50 progressively larger sets of real SARS-CoV-2 samples and calculated the parsimony score and likelihood of each optimized tree, as well as the run-time and peak RAM usage of each software package used (Fig. 3). To accomplish this, we subsampled our global phylogeny, which was produced using stringent quality control steps (see Methods), as before, to mimic the continuous accumulation of samples over the course of the pandemic.

Online optimizations are generally much faster than *de novo* phylogenetic inference. For example, IQ-TREE 2 achieves a roughly four-fold faster run-time for online optimizations compared to inferring the tree *de novo* (Fig. 3c). The 11th iteration, which has 47,819 sequences and was the last to be completed by both

online and *de novo* IQ-TREE 2, took 22 hours 50 minutes for *de novo* IQ-TREE 2 but only 5 hours 26 minutes for online IQ-TREE 2. *De novo* UShER+matOptimize was the only *de novo* method to finish all trees in fewer than 24 hours, but its speed for each daily update pales in comparison to online matOptimize. Online matOptimize is several orders of magnitude faster than its *de novo* counterpart, and its optimizations for the largest phylogenies take roughly 30 seconds, while *de novo* tree inference with UShER can take several hours for trees consisting of more than 100,000 samples (Fig. 3c). However, whether a software package is used for online or *de novo* phylogenetics does not strongly affect its peak memory usage.

We also found that online phylogenetics strategies produce trees very similar in both parsimony score and likelihood to their *de novo* counterparts, with differences of less than 1% in all cases (Fig. 3a-b). For example, in the 11th iteration, online IQ-TREE 2 produces a tree with a parsimony score of 32,005, whereas *de novo* IQ-TREE 2 produces a tree with parsimony score 32,149. Our results suggest that in addition to the computational savings that allow online phylogenetics approaches to continuously stay up-to-date, online phylogenetics approaches also produce trees with similar parsimony scores and likelihoods to their *de novo* counterparts.

Parsimony-based optimization methods have favorable metrics compared to ML methods for SARS-CoV-2 phylogenies.

In the case of both *de novo* and online phylogenetics, the parsimony-based matOptimize outperforms both FastTree 2 and IQ-TREE 2 in runtime and peak memory usage. For the sixth iteration (26,486 samples), which was the largest phylogeny inferred by all online methods in under 24 hours and using under 200 GB of RAM, online FastTree 2 required nearly 24 hours and 30.3 GB of RAM, and online IQ-TREE 2 required 1 hour 45 minutes and 72 GB of RAM. By contrast, matOptimize used only 6 seconds and 0.15 GB of RAM. This iteration contained roughly 10% as many samples as the 50th and final iteration (233,326 total samples), which online matOptimize completed in 32 seconds using 1.41 GB of RAM at peak usage. Even this largest tree represents only a very small fraction of the more than 5 million currently available SARS-CoV-2 genomes, indicating that, among the approaches we evaluated, matOptimize is the only viable option for maintaining a comprehensive SARS-CoV-2 phylogeny via online phylogenetics.

In addition to its scalability, matOptimize outperforms ML optimization methods in both the parsimony and likelihood scores of the trees that it infers. For the sixth iteration (26,486 samples), we found parsimony scores of 16,130, 16,179, and 16,290 for online matOptimize, IQ-TREE 2, and FastTree 2 respectively. While all methods produce phylogenies with parsimony scores within 1% of each other, matOptimize is consistently the lowest. However, matOptimize was developed to optimize by parsimony, while the other methods were developed for ML optimizations. Unexpectedly, we found log-likelihood scores of -233,414.277, -233,945.528, and -235,177.396 for matOptimize, IQ-TREE 2, and FastTree 2 respectively, indicating that matOptimize produces preferable phylogenies based on likelihood as well. We used a Jukes-Cantor (JC) model to calculate likelihoods due to time constraints in calculation for more complex substitution models, but a Generalised Time Reversible (GTR) model with specified rate parameters produced strongly correlated likelihoods (Fig. S1). Specifically, we fit a generalized linear model using a Gamma family (inverse link function) to predict the likelihood of the tree under the JC model using the iteration of tree construction and the GTR likelihood as predictors. We examined the six trees from the first and second iteration (12 in total). We found that the GTR likelihood was significantly correlated with the JC likelihood ($p < 2.27 \times 10^{-5}$).

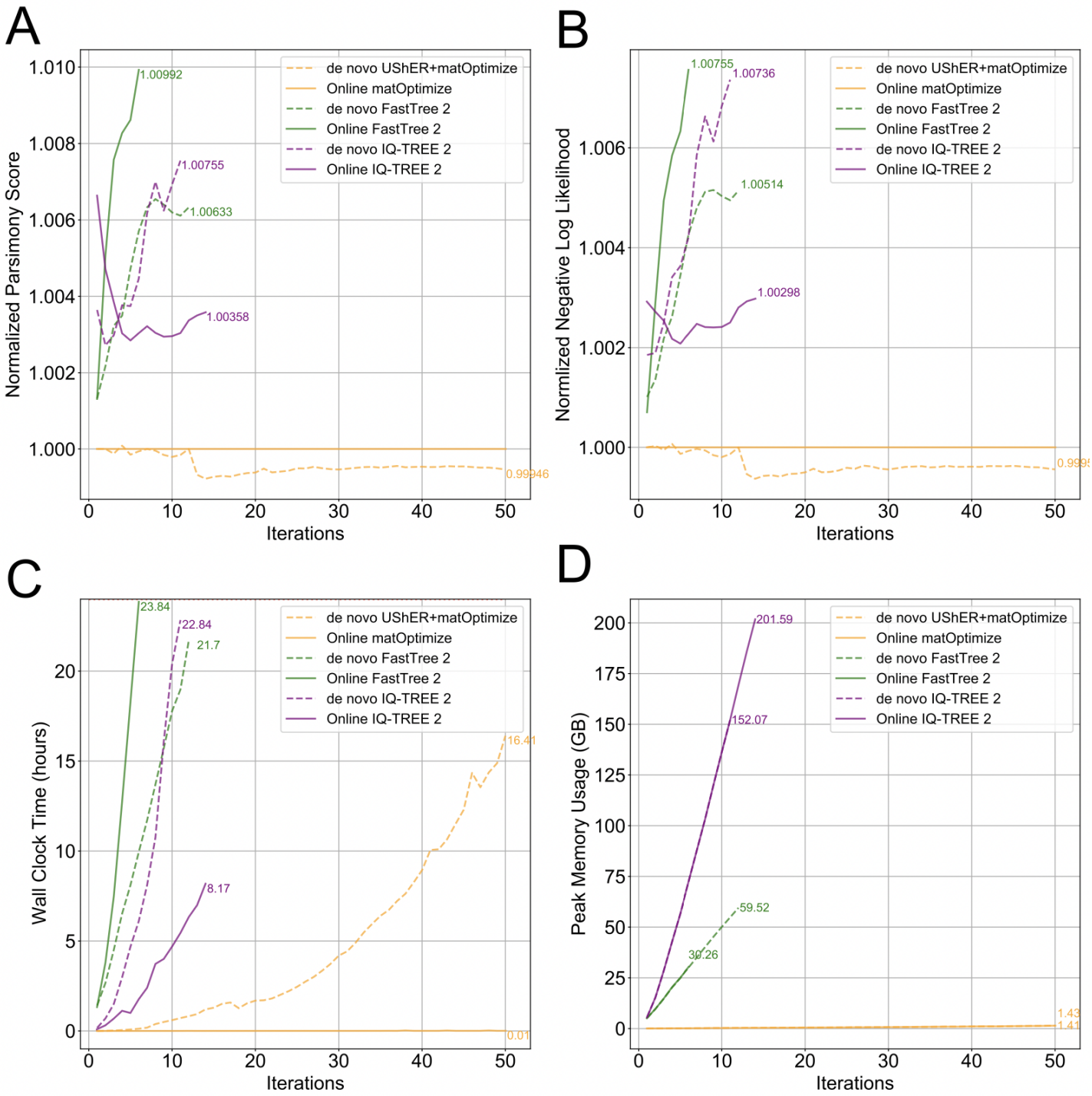

Figure 3: In practice, optimization by parsimony is more effective for SARS-CoV-2 data than optimization by ML. *We calculated (A) the parsimony score for each tree using matUtils, (B) the log-likelihood of each tree using IQ-TREE 2, (C) runtime and (D) peak memory usage of each optimization. (A) and (B) are normalized by the value obtained for the UShER/matOptimize online approach such that all other methods are expressed as a ratio. Strategies that surpassed 24 hours (C) or the allowable RAM usage (D) were terminated prior. In most cases, with the notable exception of FastTree 2, online phylogenetics (solid lines) perform better than de novo phylogenetics (dashed lines). We ran all matOptimize analyses using an instance with 15 CPUs and 117.2 GB of RAM, and we ran all IQ-TREE 2 and FastTree 2 analyses on an instance with 31 CPUs and 244.1 GB of RAM, but limited each command to 15 threads for equivalence with matOptimize.*

Parsimony and likelihood are strongly correlated when optimizing large SARS-CoV-2 phylogenies.

While our comparisons of online and *de novo* as well as parsimony-based and ML optimizations of cumulative pandemic-style data demonstrated practical performance, the largest trees completed by all methods in these experiments represent only a small fraction of available SARS-CoV-2 data. It is also crucial that we identify the optimal ways to produce a large phylogeny from already aggregated data. We therefore evaluated phylogenetic inference methods for optimizing a tree of 364,427 SARS-CoV-2 genome sequences, without constraining methods according to time or memory requirements. We optimized this global phylogeny using matOptimize (Turakhia et al. 2021a), IQ-TREE 2 (Minh et al. 2020), and FastTree 2 (Price et al. 2010). Overall, we found that matOptimize produced the tree with the lowest parsimony score across all methods in roughly one hour (Table 1).

We found that after each of the six iterations of FastTree 2 optimization, the likelihood and parsimony improvements are strongly linearly correlated (Fig. 4). This suggests that changes achieved by maximizing parsimony will also optimize likelihood for SARS-CoV-2 data. That is, for extremely densely sampled phylogenies wherein long branches are especially rare, parsimony and likelihood of phylogenies, and tree moves to optimize either are highly correlated. However, despite the strength of this correlation, we find an extreme disparity in practical usage when optimizing by either metric. Parsimony-based methods are far more time- and data-efficient, and presently-available ML approaches quickly become prohibitively expensive. For example, while the 6 iterations of FastTree did result in large improvements in both likelihood and parsimony score, the resulting tree would be out of date long before the 10.5-day optimization had completed. Moreover, we applied matOptimize to the tree output by the sixth iteration of FastTree, achieving a parsimony score of 293,866 (improvement of 288) in just 16 minutes, indicating that even after 10.5 days, additional optimization was still possible. This suggests that, for the purposes of optimizing even moderately large SARS-CoV-2 trees, parsimony-based methods should be heavily favored due to their increased efficiency.

Method	Iterations	Runtime (H:M:S)	Final Parsimony Score (Percent Change from Starting Tree)
IQ-TREE 2	2	24:30:52	294,258 (0.67)
FastTree 2	6	252:02:49	294,154 (0.71)
matOptimize	1	1:12:03	294,022 (0.75)

Table 1: We applied each of the three optimization methods to a starting tree of 364,427 SARS-CoV-2 samples, which had an initial parsimony score of 296,247. We first ran 2 iterations of IQ-TREE 2 optimization, using an SPR radius of 20 on the first and 100 on the second. We also used an SPR radius of 10 on one iteration of matOptimize, and six iterations of pseudo-likelihood optimization using FastTree 2, which we terminated after roughly 10.5 days.

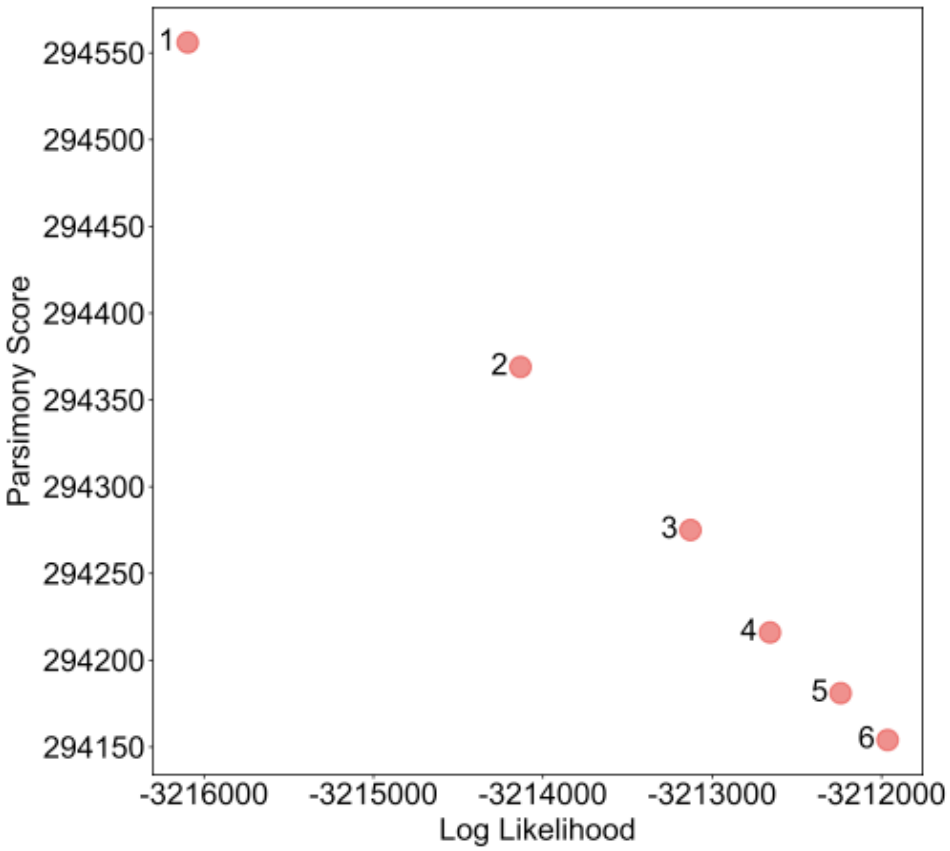

Figure 4. Improvement in likelihood and parsimony have a linear relationship for our optimized global tree. *We optimized our initial global tree using 6 iterations of FastTree and measured the total parsimony and the likelihood after each, finding a linear relationship (Pearson correlation, rho = -1.0, $p < 2.9\times10^{-7}$).*

Conclusions

The SARS-CoV-2 pandemic has made phylogenetics central to efforts to combat the spread of the virus, but has posed challenges for many commonly used phylogenetics frameworks. A major component of this effort relies on a comprehensive, up-to-date, global phylogeny of SARS-CoV-2 genomes. However, the scale and continuous growth of the data have caused difficulties for standard *de novo* phylogenetic methods. Here, we find that online phylogenetics methods are practical, pragmatic, and accurate for inferring daily phylogenetic trees from a large and densely-sample virus outbreak.

One counterintuitive result is that parsimony-based optimizations outperform sufficiently efficient ML approaches regardless of whether phylogenies are evaluated using parsimony or likelihood. This might be a consequence of the fact that parsimony scores and likelihoods are strongly correlated across phylogenies inferred via a range of phylogenetic approaches. The extremely short branches on SARS-CoV-2 phylogenies mean that the probability of multiple mutations occurring at the same site on a single branch is negligible. Stated another way, SARS-CoV-2 is approaching a “limit” where parsimony and likelihood are nearly equivalent. In turn, because of their relative efficiency, parsimony-based methods are able to search more of the possible tree space in the same amount of time, thereby resulting in trees with better likelihoods and lower parsimony scores than trees optimized using currently-available ML software packages. We emphasize that this does not bear on the relative merits of the underlying principles of ML and MP, but instead reflects the utility of methods that have been applied during the pandemic. Nevertheless, this observation does suggest that in some cases, MP optimization may provide a fast and accurate starting point for ML optimization methods. Indeed, many popular phylogenetics software, such as RAxML (Stamatakis 2014) and IQ-TREE (Minh et al. 2020) already use stepwise-addition parsimony trees as starting trees for their optimization. Our results suggest that further optimization of these starting trees using MP may provide benefits in speed *and* accuracy for some datasets, even when the target is an estimate of the ML tree.

As sequencing technologies progress and become more readily available, sample sizes for phylogenetic analyses of major pathogens and highly-studied organisms will necessarily continue to increase. Today, SARS-CoV-2 represents an extreme with respect to the total number of samples relative to the very

short branch lengths on the phylogeny. However, the global sequencing effort during the pandemic suggests that the public health sphere has a strong interest in the increased application of whole-genome sequencing to study the genomic contents, evolution, and transmission history of major and emerging human pathogens. We expect that million-sample datasets will become commonplace in the near future. Online phylogenetics, using parsimony-based tree inference and optimization, or greatly accelerated likelihood approximations, will be a fruitful avenue for future development and application to accommodate these datasets.

Methods

We first developed a "global phylogeny", from which all analyses in this study were performed. We began by downloading .vcf and .fasta files corresponding to March 18, 2021, from our own daily-updated database (McBroome et al. 2021). These files contain 434,063 samples. We then implemented filters, retaining only sequences containing at least 28,000 non-N nucleotides, and fewer than two non-[ACGTN-] characters. We used UShER to create a phylogeny from scratch using only the remaining 366,492 samples. We iteratively pruned this tree of internal branches with parsimony scores greater than 30, then terminal branches with parsimony scores greater than 6, until convergence, resulting in a final global phylogeny containing 364,428 samples. For full reproducibility, files used for all analyses can be found in subrepository 1 on the project GitHub page (Thornlow et al. 2021b).

Following this, we further optimized this global phylogeny, hereafter the "starting tree", using matOptimize, FastTree 2, and IQ-TREE 2. We used the starting tree and its corresponding alignment and ran five iterations of IQ-TREE 2, varying the SPR radius from 20 to 100 in increments of 20. Separately, we also optimized the starting tree using two iterations of IQ-TREE 2, the first iteration using an SPR radius of 20 and the second using a radius of 100. We also ran one iteration of matOptimize on the starting tree using an SPR radius of 10, and six iterations of FastTree 2 on the starting tree. We then ran matOptimize again on the tree output by the sixth FastTree 2 iteration. Files for all analyses can be found in subrepository 2.

To mimic pandemic-style phylogenetics, we separated the samples in the starting tree into batches of ~5,000 by sorting according to the date of sample collection. We then set up two frameworks for each of the three software packages (matOptimize (commit 66ca5ff), IQ-TREE 2 (multicore version 2.1.3 COVID-edition), and FastTree 2 (Double Precision version 2.1.10)). In online phylogenetics, we began with a tree created from scratch using UShER, optimized using the software package of choice, and added samples with UShER and re-optimizing after each batch. In *de novo* phylogenetics, we supplied each software package with an alignment corresponding to all samples in that batch and its predecessors (or .vcf for matOptimize) without a guide tree. For both cases, each tree is larger than its predecessor by ~5,000 samples, and each tree necessarily contains all samples in the immediately preceding tree. For FastTree 2, we used 2 rounds of subtree-prune-regraft (SPR) moves, maximum SPR length of 1000, zero rounds of minimum evolution nearest neighbor interchanges (NNI), and the Generalised Time Reversible + Gamma (GTR+G) substitution model. For IQ-TREE 2, we used a branch length minimum of 0.000000001, zero rounds of stochastic tree search, and the GTR+G substitution model. We ran all matOptimize analyses using an instance with 15 CPUs and 117.2 GB of RAM, and we ran all IQ-TREE 2 and FastTree 2 analyses on an instance with 31 CPUs and 244.1 GB of RAM, but we limited each command to 15 threads for equivalence with matOptimize. Files for all analyses can be found in subrepository 3.

To generate our simulated data, we used the SARS-CoV-2 reference genome (GISAID ID: EPI_ISL_402125; GenBank ID: MN908947.3) (Shu and McCauley 2017; Sayers et al. 2021) as the root sequence and used phastSim (De Maio et al. 2021b) to simulate according to the global tree optimized by six iterations of FastTree 2 and one iteration of matOptimize (*after_usher_optimized_fasttree_iter6.tree* available in Subrepository 2 (Thornlow et al. 2021b)). We chose this tree as the ground truth for our simulation experiments because it had the greatest log-likelihood and lowest parsimony of any of our optimizations of the starting tree. Intergenic regions were evolved using phastSim (De Maio et al. 2021b) using the default neutral mutation rates estimated in ref. (De Maio et al. 2021a), with position-specific mean mutation rates sampled from a gamma distribution with alpha=beta=4, and with 1% of the genome having a 10-fold increase mutation rate for one specific mutation type (SARS-CoV-2 hypermutability model described in ref. (De Maio et al. 2021b)). Evolution of coding regions was simulated with the same neutral mutational distribution, with a mean nonsynonymous/synonymous rate ratio of omega=0.48 as estimated in (Turakhia et al. 2021a), with codon-specific omega values sampled from a gamma distribution with alpha=0.96 and beta=2. Rates for each intergenic and coding region were not normalized in order to have the same baseline neutral mutation rate distribution across the genome.

We repeated our iterative experiments using *de novo* and online matOptimize, IQ-TREE 2 and FastTree 2 on this simulated alignment, using the same strategies as before. However, instead of computing parsimony and likelihood scores, we computed the Robinson-Foulds (RF) distance (Robinson and Foulds 1981) of each optimization to the ground truth tree, pruned to contain only the samples belonging to that batch. To calculate each RF distance, we used the -R (resolve polytomies) and -O (collapse tree) arguments in matUtils extract (McBroome et al. 2021) and then used the dist.topo command in the *ape* package in R (Paradis and Schliep 2019), comparing the collapsed optimized tree and the pruned, collapsed ground truth tree at each iteration. In Figure 2, we expressed the RF distances as a proportion of the total possible RF distance, which is equivalent to two times the number of samples in the trees minus six (Steel and Penny 1993).

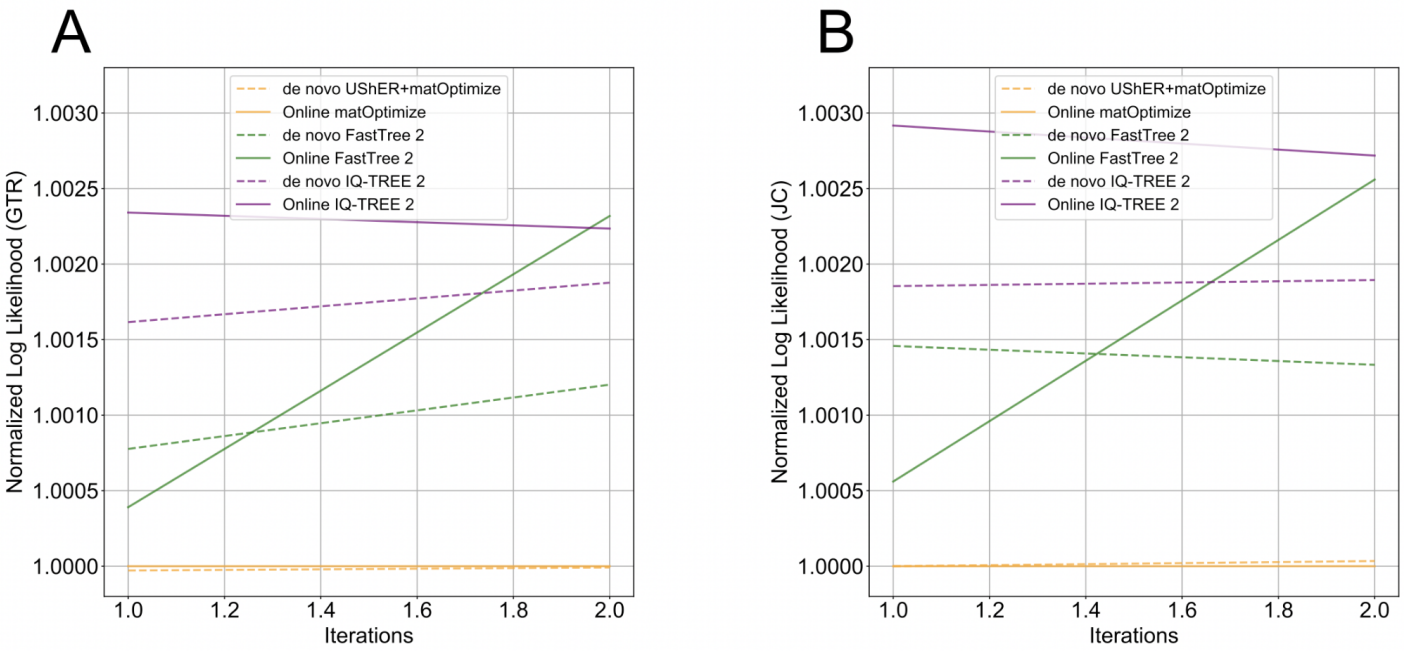

Figure S1: Log likelihoods calculated using Generalised Time Reversible (GTR) and Jukes-Cantor (JC) models are correlated. *We calculated log likelihoods for each de novo and online method as in Figure 2B using (A) GTR+G and (B) JC models, which suggest that relative performance of each method is consistent across models, and significantly correlated with each other. All values are normalized by the value obtained for the UShER/matOptimize online approach, such that other methods are expressed as a ratio.*

Acknowledgments: We gratefully acknowledge the authors from the originating laboratories responsible for obtaining each sample, as well as the submitting laboratories where the genome data were generated and shared, on which this research is based.

Funding: This work was supported by National Institutes of Health (R35GM128932 to R.C.D., T32HG008345 (B.T. and J.M.), F31HG010584 to B.T.), Alfred P. Sloan Foundation fellowship, University of California Office of the President Emergency COVID-19 Research Seed Funding (R00RG2456 to R.C.-D.), European Molecular Biology Laboratory (to N.D.M.), Australian Research Council (DP200103151 to R.L.), Chan-Zuckerberg Initiative grant (to R.L.), and by Eric and Wendy Schmidt by recommendation of the Schmidt Futures program.

Competing interests: R.L. works as an advisor to GISAID. The remaining authors declare no competing interests.

References:

Annavajhala M.K., Mohri H., Wang P., Nair M., Zucker J.E., Sheng Z., Gomez-Simmonds A., Kelley A.L., Tagliavia M., Huang Y., Bedford T., Ho D.D., Uhlemann A.-C. 2021. A Novel and Expanding SARS-CoV-2 Variant, B.1.526, Identified in New York. medRxiv.

Barbera P., Kozlov A.M., Czech L., Morel B., Darriba D., Flouri T., Stamatakis A. 2019. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. Syst. Biol. 68:365–369.

Berger S.A., Krompass D., Stamatakis A. 2011. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. Syst. Biol. 60:291–302.

Bluhm A., Christandl M., Gesmundo F., Klausen F.R., Mančinska L., Steffan V., França D.S., Werner A.H. 2020. SARS-CoV-2 transmission routes from genetic data: A Danish case study. PLOS ONE. 15:e0241405.

Castillo A.E., Parra B., Tapia P., Acevedo A., Lagos J., Andrade W., Arata L., Leal G., Barra G., Tambley C., Tognarelli J., Bustos P., Ulloa S., Fasce R., Fernández J. 2020. Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. J. Med. Virol. 92:1562–1566.

COVID-19 Genomics UK (COG-UK) Consortium. 2020. An integrated national scale SARS-CoV-2 genomic surveillance network. Lancet Microbe. 1:e99–e100.

De Maio N., Walker C.R., Turakhia Y., Lanfear R., Corbett-Detig R., Goldman N. 2021a. Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. Genome Biol. Evol. 13.

De Maio N., Weilguny L., Walker C.R., Turakhia Y., Corbett-Detig R., Goldman N. 2021b. phastSim: efficient simulation of sequence evolution for pandemic-scale datasets. bioRxiv.

Deng X., Gu W., Federman S., du Plessis L., Pybus O.G., Faria N.R., Wang C., Yu G., Bushnell B., Pan C.-Y., Guevara H., Sotomayor-Gonzalez A., Zorn K., Gopez A., Servellita V., Hsu E., Miller S., Bedford T., Greninger A.L., Roychoudhury P., Starita L.M., Famulare M., Chu H.Y., Shendure J., Jerome K.R., Anderson C., Gangavarapu K., Zeller M., Spencer E., Andersen K.G., MacCannell D., Paden C.R., Li Y., Zhang J., Tong S., Armstrong G., Morrow S., Willis M., Matyas B.T., Mase S., Kasirye O., Park M., Masinde G., Chan C., Yu A.T., Chai S.J., Villarino E., Bonin B., Wadford D.A., Chiu C.Y. 2020. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science. 369:582–587.

Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Syst. Biol. 27:401–410.

Fourment M., Claywell B.C., Dinh V., McCoy C., Matsen F.A. Iv, Darling A.E. 2018. Effective Online Bayesian Phylogenetics via Sequential Monte Carlo with Guided Proposals. Syst. Biol. 67:490–502.

Franceschi V.B., Caldana G.D., de Menezes Mayer A., Cybis G.B., Neves C.A.M., Ferrareze P.A.G., Demoliner M., de Almeida P.R., Gularte J.S., Hansen A.W., Weber M.N., Fleck J.D., Zimerman R.A., Kmetzsch L., Spilki F.R., Thompson C.E. 2021. Genomic epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil. BMC Genomics. 22:371.

Gill M.S., Lemey P., Suchard M.A., Rambaut A., Baele G. 2020. Online Bayesian Phylodynamic Inference in BEAST with Application to Epidemic Reconstruction. Mol. Biol. Evol. 37:1832–1842.

Gonzalez-Reiche A.S., Hernandez M.M., Sullivan M.J., Ciferri B., Alshammary H., Obla A., Fabre S., Kleiner G., Polanco J., Khan Z., Alburquerque B., van de Guchte A., Dutta J., Francoeur N., Melo B.S., Oussenko I., Deikus G., Soto J., Sridhar S.H., Wang Y.-C., Twyman K., Kasarskis A., Altman D.R., Smith M., Sebra R., Aberg J., Krammer F., García-Sastre A., Luksza M., Patel G., Paniz-Mondolfi A., Gitman M., Sordillo E.M., Simon V., van Bakel H. 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. Science. 369:297–301.

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.

Hendy M.D., Penny D. 1989. A Framework for the Quantitative Study of Evolutionary Trees. Syst. Biol. 38:297–309.

Hug L.A., Baker B.J., Anantharaman K., Brown C.T., Probst A.J., Castelle C.J., Butterfield C.N., Hernsdorf A.W., Amano Y., Ise K., Suzuki Y., Dudek N., Relman D.A., Finstad K.M., Amundson R., Thomas B.C., Banfield J.F. 2016. A new view of the tree of life. Nat Microbiol. 1:16048.

Jackson B., Boni M.F., Bull M.J., Colleran A., Colquhoun R.M., Darby A.C., Haldenby S., Hill V., Lucaci A., McCrone J.T., Nicholls S.M., O'Toole Á., Pacchiarini N., Poplawski R., Scher E., Todd F., Webster H.J., Whitehead M., Wierzbicki C., COVID-19 Genomics UK (COG-UK) Consortium, Loman N.J., Connor T.R., Robertson D.L., Pybus O.G., Rambaut A. 2021. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell.

Kalantar K.L., Carvalho T., de Bourcy C.F.A., Dimitrov B., Dingle G., Egger R., Han J., Holmes O.B., Juan Y.-F., King R., Kislyuk A., Lin M.F., Mariano M., Morse T., Reynoso L.V., Cruz D.R., Sheu J., Tang J., Wang J., Zhang M.A., Zhong E., Ahyong V., Lay S., Chea S., Bohl J.A., Manning J.E., Tato C.M., DeRisi J.L. 2020. IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. Gigascience. 9.

Khan A., Zia T., Suleman M., Khan T., Ali S.S., Abbasi A.A., Mohammad A., Wei D.-Q. 2021. Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: An insight from structural data. J. Cell. Physiol. 236:7045–7057.

Kolaczkowski B., Thornton J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 431:980–984.

Lam T.T.-Y. 2020. Tracking the Genomic Footprints of SARS-CoV-2 Transmission. Trends Genet. 36:544–546.

Lanfear R., Mansfield R. 2020. roblanf/sarscov2phylo: 13-11-20. .

Li X., Giorgi E.E., Marichann M.H., Foley B., Xiao C., Kong X.-P., Chen Y., Korber B., Gao F. Emergence of SARS-CoV-2 through Recombination and Strong Purifying Selection. .

Lu J., du Plessis L., Liu Z., Hill V., Kang M., Lin H., Sun J., François S., Kraemer M.U.G., Faria N.R., McCrone J.T., Peng J., Xiong Q., Yuan R., Zeng L., Zhou P., Liang C., Yi L., Liu J., Xiao J., Hu J., Liu T., Ma W., Li W., Su J., Zheng H., Peng B., Fang S., Su W., Li K., Sun R., Bai R., Tang X., Liang M., Quick J., Song T., Rambaut A., Loman N., Raghwani J., Pybus O.G., Ke C. 2020a. Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell. 181:997–1003.e9.

Lu R., Zhao X., Li J., Niu P., Yang B., Wu H., Wang W., Song H., Huang B., Zhu N., Bi Y., Ma X., Zhan F., Wang L., Hu T., Zhou H., Hu Z., Zhou W., Zhao L., Chen J., Meng Y., Wang J., Lin Y., Yuan J., Xie Z., Ma J., Liu W.J., Wang D., Xu W., Holmes E.C., Gao G.F., Wu G., Chen W., Shi W., Tan W. 2020b. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 395:565–574.

Matsen F.A., Kodner R.B., Armbrust E.V. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics. 11:538.

McBroome J., Thornlow B., Hinrichs A.S., De Maio N., Goldman N., Haussler D., Corbett-Detig R., Turakhia Y. 2021. A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. bioRxiv.:2021–2004.

Meredith L.W., Hamilton W.L., Warne B., Houldcroft C.J., Hosmillo M., Jahun A.S., Curran M.D., Parmar S., Caller L.G., Caddy S.L., Khokhar F.A., Yakovleva A., Hall G., Feltwell T., Forrest S., Sridhar S., Weekes M.P., Baker S., Brown N., Moore E., Popay A., Roddick I., Reacher M., Gouliouris T., Peacock S.J.,

Dougan G., Török M.E., Goodfellow I. 2020. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect. Dis. 20:1263–1271.

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.

Moreno G.K., Braun K.M., Riemersma K.K., Martin M.A., Halfmann P.J., Crooks C.M., Prall T., Baker D., Baczenas J.J., Heffron A.S., Ramuta M., Khubbar M., Weiler A.M., Accola M.A., Rehrauer W.M., O'Connor S.L., Safdar N., Pepperell C.S., Dasu T., Bhattacharyya S., Kawaoka Y., Koelle K., O'Connor D.H., Friedrich T.C. 2020. Revealing fine-scale spatiotemporal differences in SARS-CoV-2 introduction and spread. Nat. Commun. 11:5558.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics.

Park A.K., Kim I.-H., Kim J., Kim J.-M., Kim H.M., Lee C.Y., Han M.-G., Rhie G.-E., Kwon D., Nam J.-G., Park Y.-J., Gwack J., Lee N.-J., Woo S., No J.S., Lee J., Ha J., Rhee J., Yoo C.-K., Kim E.-J. 2021. Genomic Surveillance of SARS-CoV-2: Distribution of Clades in the Republic of Korea in 2020. Osong Public Health Res Perspect. 12:37–43.

Parks D.H., Chuvochina M., Waite D.W., Rinke C., Skarshewski A., Chaumeil P.-A., Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36:996–1004.

Peng J., Liu J., Mann S.A., Mitchell A.M., Laurie M.T., Sunshine S., Pilarowski G., Ayscue P., Kistler A., Vanaerschot M., Li L.M., McGeever A., Chow E.D., Marquez C., Nakamura R., Rubio L., Chamie G., Jones D., Jacobo J., Rojas S., Rojas S., Tulier-Laiwa V., Black D., Martinez J., Naso J., Schwab J., Petersen M., Havlir D., DeRisi J., IDseq Team. 2021. Estimation of secondary household attack rates for emergent spike L452R SARS-CoV-2 variants detected by genomic surveillance at a community-based testing site in San Francisco. Clin. Infect. Dis.

Philippe H., Zhou Y., Brinkmann H., Rodrigue N., Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5:50.

Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 5:e9490.

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. Nature Microbiology. 5:1403–1407.

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

Sanderson T. taxodium: Explore very large trees in the browser. Github.

Sayers E.W., Cavanaugh M., Clark K., Pruitt K.D., Schoch C.L., Sherry S.T., Karsch-Mizrachi I. 2021. GenBank. Nucleic Acids Res. 49:D92–D96.

Shu Y., McCauley J. 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. 22.

Skidmore P.T., Kaelin E.A., Holland L.R.A., Maqsood R. 2021. Emergence of a SARS-CoV-2 E484K variant of interest in Arizona. medRxiv.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Steel M.A., Penny D. 1993. Distributions of tree comparison metrics—some new results. Syst. Biol.

Sullivan J., Swofford D.L. 2001. Should We Use Model-Based Methods for Phylogenetic Inference When We Know That Assumptions About Among-Site Rate Variation and Nucleotide Substitution Pattern Are Violated? Systematic Biology. 50:723–729.

Tang J.W., Toovey O.T.R., Harvey K.N., Hui D.D.S. 2021. Introduction of the South African SARS-CoV-2 variant 501Y.V2 into the UK. J. Infect. 82:e8–e10.

Tegally H., Wilkinson E., Giovanetti M., Iranzadeh A., Fonseca V., Giandhari J., Doolabh D., Pillay S., San E.J., Msomi N., Mlisana K., von Gottberg A., Walaza S., Allam M., Ismail A., Mohale T., Glass A.J., Engelbrecht S., Van Zyl G., Preiser W., Petruccione F., Sigal A., Hardie D., Marais G., Hsiao N.-Y., Korsman S., Davies M.-A., Tyers L., Mudau I., York D., Maslo C., Goedhals D., Abrahams S., Laguda-Akingba O., Alisoltani-Dehkordi A., Godzik A., Wibmer C.K., Sewell B.T., Lourenço J., Alcantara L.C.J., Kosakovsky Pond S.L., Weaver S., Martin D., Lessells R.J., Bhiman J.N., Williamson C., de Oliveira T. 2021. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature. 592:438–443.

Thornlow B., Hinrichs A.S., Jain M., Dhillon N., La S., Kapp J.D., Anigbogu I., Cassatt-Johnstone M., McBroome J., Haeussler M., Turakhia Y., Chang T., Olsen H.E., Sanford J., Stone M., Vaske O., Bjork I., Akeson M., Shapiro B., Haussler D., Kilpatrick A.M., Corbett-Detig R. 2021a. A new SARS-CoV-2 lineage that shares mutations with known Variants of Concern is rejected by automated sequence repository quality control. bioRxiv.

Thornlow B., roblanf, Corbett-Detig R., Turakhia Y., Cheng Y. 2021b. bpt26/parsimony: .

Tian F., Tong B., Sun L., Shi S., Zheng B., Wang Z., Dong X., Zheng P. 2021. Mutation N501Y in RBD of Spike Protein Strengthens the Interaction between COVID-19 and its Receptor ACE2. bioRxiv.:2021.02.14.431117.

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

Turakhia Y., Thornlow B., Hinrichs A.S., Mcbroome J. 2021b. Pandemic-Scale phylogenomics reveals elevated recombination rates in the SARS-CoV-2 spike region. bioRxiv.

Umair M., Ikram A., Salman M., Khurshid A., Alam M., Badar N., Suleman R., Tahir F., Sharif S., Montgomery J., Whitmer S., Klena J. 2021. Whole-genome sequencing of SARS-CoV-2 reveals the detection of G614 variant in Pakistan. PLoS One. 16:e0248371.

Wertheim J.O., Steel M., Sanderson M.J. 2021. Accuracy in near-perfect virus phylogenies. Syst. Biol.