1   **Title:**

2

3   Online Phylogenetics using Parsimony Produces Slightly Better Trees and is Dramatically More Efficient

4   for Large SARS-CoV-2 Phylogenies than *de novo* and Maximum-Likelihood Approaches

5

6   **Authors:**

7

8   Bryan Thornlow[1,2,*,#], Alexander Kramer[1,2,#], Cheng Ye[3], Nicola De Maio[4], Jakob McBroome[1,2], Angie S.

9   Hinrichs[2], Robert Lanfear[5], Yatish Turakhia[3], Russell Corbett-Detig[1,2,*]

10

11   **Affiliations:**

12

13   [1]Department of Biomolecular Engineering, University of California, Santa Cruz; Santa Cruz, CA

14   95064, USA

15   [2]Genomics Institute, University of California, Santa Cruz; Santa Cruz, CA 95064, USA

16   [3]Department of Electrical and Computer Engineering, University of California, San Diego; San

17   Diego, CA 92093, USA

18   [4]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI),

19   Wellcome Genome Campus; Cambridge CB10 1SD, UK

20   [5]Department of Ecology and Evolution, Research School of Biology, Australian National

21   University; Canberra, ACT 2601, Australia

22   *Correspondence to bthornlo@ucsc.edu and rucorbet@ucsc.edu

23   [#]These authors contributed equally to this work.

24

25

26

27 **Abstract:**

28 Phylogenetics has been foundational to SARS-CoV-2 research and public health policy, assisting in

29 genomic surveillance, contact tracing, and assessing emergence and spread of new variants. However,

30 phylogenetic analyses of SARS-CoV-2 have often relied on tools designed for *de novo* phylogenetic

31 inference, in which all data are collected before any analysis is performed and the phylogeny is inferred

32 once from scratch. SARS-CoV-2 datasets do not fit this mould. There are currently over 10 million

33 sequenced SARS-CoV-2 genomes in online databases, with tens of thousands of new genomes added

34 every day. Continuous data collection, combined with the public health relevance of SARS-CoV-2, invites

35 an "online" approach to phylogenetics, in which new samples are added to existing phylogenetic trees

36 every day. The extremely dense sampling of SARS-CoV-2 genomes also invites a comparison between

37 likelihood and parsimony approaches to phylogenetic inference. Maximum likelihood (ML) methods are

38 more accurate when there are multiple changes at a single site on a single branch, but this accuracy

39 comes at a large computational cost, and the dense sampling of SARS-CoV-2 genomes means that

40 these instances will be extremely rare because each internal branch is expected to be extremely short.

41 Therefore, it may be that approaches based on maximum parsimony (MP) are sufficiently accurate for

42 reconstructing phylogenies of SARS-CoV-2, and their simplicity means that they can be applied to much

43 larger datasets. Here, we evaluate the performance of *de novo* and online phylogenetic approaches, and

44 ML and MP frameworks, for inferring large and dense SARS-CoV-2 phylogenies. Overall, we find that

45 online phylogenetics produces similar phylogenetic trees to *de novo* analyses for SARS-CoV-2, and that

46 MP optimizations produce more accurate SARS-CoV-2 phylogenies than do ML optimizations. Since MP

47 is thousands of times faster than presently available implementations of ML and online phylogenetics is

48 faster than *de novo*, we therefore propose that, in the context of comprehensive genomic epidemiology

49 of SARS-CoV-2, MP online phylogenetics approaches should be favored.

50

51 **Key words:**

52 SARS-CoV-2, phylogenetics, parsimony, maximum likelihood, optimization

53

54   **Introduction:**

55

56   The widespread availability and extreme abundance of pathogen genome sequencing has made

57   phylogenetics central to combatting the pandemic. Communities worldwide have begun implementing

58   genomic surveillance by systematically sequencing the genomes of a percentage of local cases (Deng et

59   al. 2020; Lu et al. 2020a; Meredith et al. 2020; Park et al. 2021). This has been invaluable in tracing local

60   transmission chains (Bluhm et al. 2020; Lam 2020), understanding the genetic makeup of viral

61   populations within local communities (Gonzalez-Reiche et al. 2020; Franceschi et al. 2021; Thornlow et

62   al. 2021a), uncovering the means by which viral lineages have been introduced to new areas (Castillo et

63   al. 2020), and measuring the relative spread of specific variants (Skidmore et al. 2021; Umair et al.

64   2021). Phylogenetic approaches for better understanding the proximate evolutionary origins of the virus

65   (Li et al. 2020), as well as to identify recombination events (Jackson et al. 2021; Turakhia et al. 2021b)

66   and instances of convergent evolution (Kalantar et al. 2020; Peng et al. 2021) have greatly informed our

67   understanding of the virus. Phylogenetic visualization software including Auspice (Hadfield et al. 2018)

68   and Taxonium (Sanderson 2021a) have also become widely used for public health purposes.

69   A comprehensive, up-to-date phylogenetic tree of SARS-CoV-2 is important for public health

70   officials and researchers. A tree containing all available sequences can sometimes facilitate identification

71   of epidemiological links between samples that might otherwise be obscured in subsampled phylogenies.

72   Conversely, these approaches can often rule out otherwise plausible transmission histories. Such

73   information can also help to identify the likely sources of new viral strains in a given area (Moreno et al.

74   2020; Tang et al. 2021). Additionally, using up-to-date information enables us to find and track quickly

75   growing clades and novel variants of concern (Annavajhala et al. 2021; Tegally et al. 2021), as well as to

76   measure the spread of known variants at both global and community scales. Furthermore,

77   comprehensive phylogenies can facilitate identification of recombinant viral genomes (Turakhia et al.

78   2021b), natural selection at homoplasious positions (van Dorp et al. 2020), variation in mutation rates

79   (De Maio et al. 2021a), and systematic recurrent errors (Turakhia et al. 2020). This also facilitates

80   naming lineages of interest, which has been especially important in tracking variants of concern during

81   the pandemic (*e.g.* B.1.1.7 or "Alpha" and B.1.617.2 or "Delta") (Rambaut et al. 2020).

82   SARS-CoV-2 presents a unique set of phylogenetic challenges. First, the unprecedented pace

83   and scale of whole-genome sequence data has forced the phylogenetics community to place runtime

84   and scalability at the center of every inference strategy. More than 10 million SARS-CoV-2 genome

85   sequences are currently available, with tens of thousands being added each day. Prior to the pandemic,

86   *de novo* phylogenetics, or approaches that infer phylogenies from scratch, have been the standard, as

87   there has rarely been a need to re-infer or improve pre-existing phylogenies on a daily basis. Re-inferring

88   a tree of more than 10 million samples daily, however, is extremely costly, and has brought a renewed

89   focus on methods for adding new samples to existing phylogenetic trees (Matsen et al. 2010; Berger et

90   al. 2011; Izquierdo-Carrasco et al. 2014; Fourment et al. 2018; Barbera et al. 2019). This approach has

91   been called "online phylogenetics" (Gill et al. 2020), and has important advantages in the context of the

92   pandemic and beyond. Online phylogenetics is appealing for the genomic surveillance of any pathogen,

93   because iterative optimization should decrease computational expense, allowing good estimates of

94   phylogenies to be made readily available.

95   Second, SARS-CoV-2 genomes are much more closely related than sequences in most other

96   phylogenetic analyses. Because the advantages of maximum likelihood methods decrease for closely

97   related samples and long branches are relatively rare in the densely sampled SARS-CoV-2 phylogeny

98   (Felsenstein 1978; Hendy and Penny 1989; Philippe et al. 2005), this suggests that phylogenetic

99   inferences based on maximum parsimony, a much faster and simpler phylogenetic inference method,

100   could be better suited for online phylogenetic analyses of SARS-CoV-2 genomes (Wertheim et al. 2021).

101   The principle of maximum parsimony is that the tree with the fewest mutations should be favored, and it

102   is sometimes described as a non-parametric phylogenetic inference method (Sullivan and Swofford

103   2001; Kolaczkowski and Thornton 2004). Additionally, because parsimony-based tree optimization does

104   not require estimation of ancestral character state uncertainty at all positions in the phylogeny like ML

105   optimization does, parsimony uses much less memory.

106   Here, we evaluate approaches that would enable one to maintain a fully up-to-date and

107   comprehensive global phylogeny of SARS-CoV-2 genome sequences (McBroome et al. 2021).

108   Specifically, we investigate tradeoffs between online and *de novo* phylogenetics and between maximum

109   parsimony and maximum likelihood approaches when the aim is for an analysis to scale to millions of

110   sequences, with tens of thousands of new sequences being added daily. We chose to compare

111   maximum parsimony and maximum likelihood (and omit other approaches like neighbor-joining) because

112   they were the most effective methods at inferring large SARS-CoV-2 phylogenies based on previous

113   analyses (Lanfear and Mansfield 2020), and because the most efficient distance-based methods are

114   quadratic in memory usage so cannot scale to estimating trees from datasets of more than a few

115   hundred thousand sequences (Wang et al. 2022). We mimic the time-course of the pandemic by

116   introducing increasingly large numbers of SARS-CoV-2 genome sequences proportionately to their

117   reported sampling dates.

118   We evaluate potential online phylogenetics approaches by iteratively adding samples to existing

119   trees and optimizing the augmented phylogeny with different tools that have been proposed for this

120   purpose during the pandemic. In particular, we evaluate matOptimize, IQ-TREE 2, and FastTree 2.

121   Between each optimization step, we use UShER (Turakhia et al. 2021a) to add samples to trees by

122   maximum parsimony. matOptimize is a parsimony optimization approach that uses subtree pruning and

123   regrafting (SPR) moves to minimize the total mutations in the final tree topology (Ye et al. 2022) . IQ-

124   TREE 2 uses nearest neighbor interchange (NNI) to find the tree with the highest likelihood given an

125   input multiple sequence alignment (Minh et al. 2020). FastTree 2 uses a pseudo-likelihood approach that

126   employs minimum-evolution SPR and/or NNI moves and maximum-likelihood NNI moves while using

127   several heuristics to reduce the search space (Price et al. 2010). The likelihood-based approaches

128   evaluated here report branch lengths in substitutions per site. Parsimony-based matOptimize reports

129   branch lengths in total substitutions, which can be converted to the latter by dividing by the genome

130   length. These branch lengths may be interpreted as is or used as an initial estimate for other distance

131   measures, for example in the construction of time trees (Sanderson 2021b).

132   Results from our comparisons demonstrate that for the purposes of SARS-CoV-2 phylogenetics,

133   in which samples are numerous and closely related and inference speed is of high significance,

134   parsimony-based online phylogenetics applications are clearly most favorable and are also the only

135   immediately available methods capable of producing daily phylogenetic estimates of all available SARS-

136   CoV-2 genomes (Turakhia et al. 2021a). We note that matOptimize is used to maintain such a phylogeny

137   comprising over 9 million genomes as of April 2022 (McBroome et al. 2021). As similarly vast datasets

138   will soon be available for many species and pathogens, we expect that online approaches using

139   parsimony or pseudo-likelihood optimization will become increasingly central to phylogenetic inference.

140

141   **Results and Discussion:**

142

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

144

145   The vast majority of phylogenetics during the pandemic has consisted of *de novo* phylogenetics

146   approaches (Hadfield et al. 2018; Li et al. 2020; Lu et al. 2020a, 2020b; Meredith et al. 2020), in which

147   each phylogeny is inferred using only genetic variation data, and without a guide tree (Fig. 1). This

148   strategy for phylogenetic inference has long been the default, as in most instances in the past, data are

149   collected just once for a project, and more relevant data are rarely going to be made available in the near

150   future. This process is well characterized and has been foundational for many phylogenetics studies

151   (Hug et al. 2016; Parks et al. 2018; Lu et al. 2020b), and most phylogenetics software is developed with

152   *de novo* phylogenetics as the primary intended usage.

153   A challenging aspect of pandemic phylogenetics is the need to keep up with the pace of data

154   generation as genome sequences continuously become available. To evaluate phylogenetics

155   applications in the pandemic (Fig. 1), we split 233,326 samples dated from December 23, 2019 through

156   January 11, 2021 into 50 batches according to their date of collection. Each batch contains roughly 5,000

157   samples. Samples in each batch were collected within a few days of each other, except in the first

158   months of the pandemic when sample collection was more sparse. We also constructed a dataset of

159 otherwise similar data simulated from a known phylogeny (see Methods). The intent of this scheme is to

160 roughly approximate the data generation and deposition that occurred during the pandemic. All datasets

161 are available from the repository associated with this project (Thornlow et al. 2021b), for reproducibility

162 and so that future methods developers can directly compare their outputs to our results. We performed

163 online and *de novo* phylogenetics using a range of inference and optimization approaches. Since

164 thousands of new sequences are added to public sequence repositories each day, we terminated any

165 phylogenetic inference approaches that took more than 24 hours, because such phylogenies would be

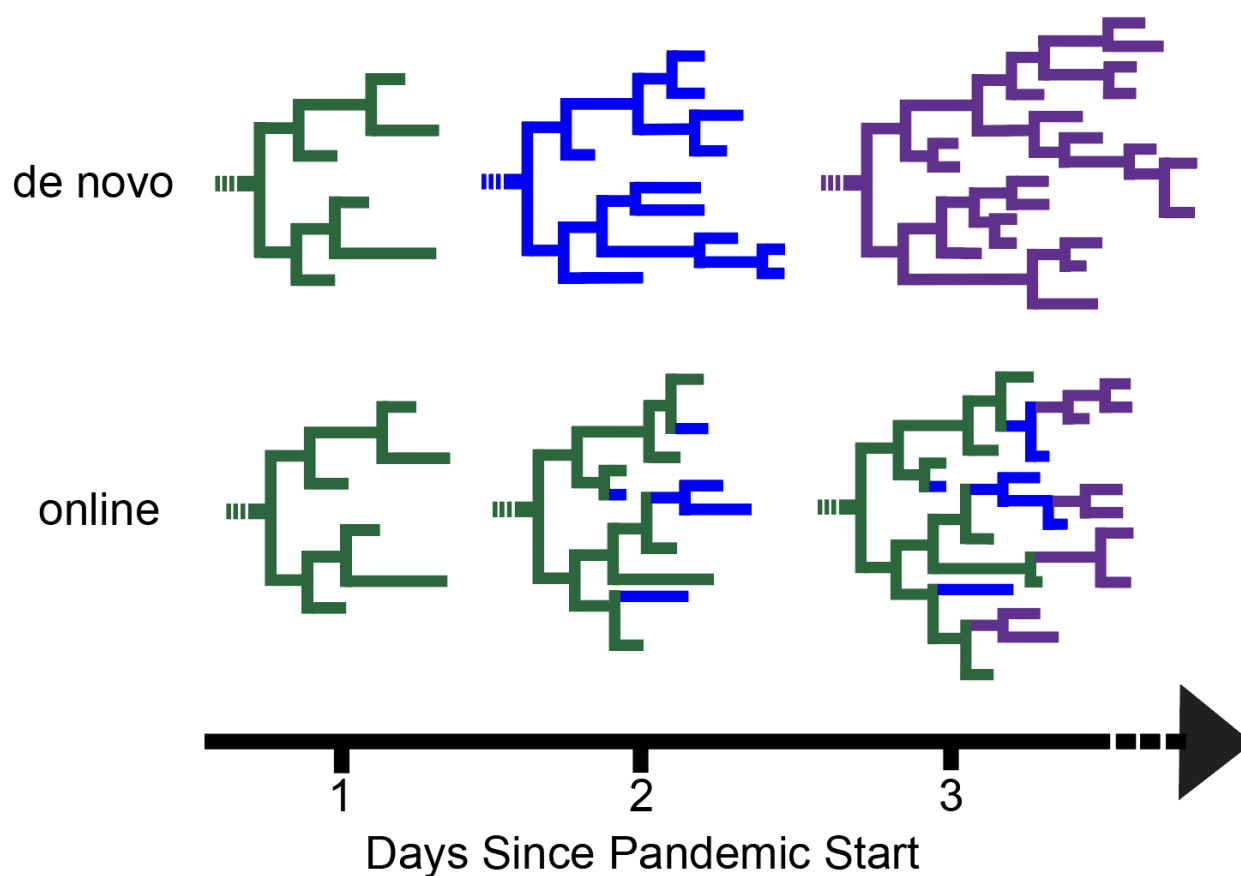166 obsolete for some public health applications by the time they were inferred.

167

168

169 **Figure 1: Phylogenies may be optimized from scratch using de novo phylogenetics or iteratively**

170 **using online phylogenetics.** *In de novo phylogenetics (top), trees are repeatedly re-inferred from*

171 *scratch. Conversely, online phylogenetics (bottom) involves placement of new samples as they are*

172 *collected. Intermittent optimization steps (not depicted) after new samples are placed can help overcome*

173 *errors from previous iterations. Online phylogenetics is expected to be much faster and require less*

174 *memory than de novo phylogenetics.*

175

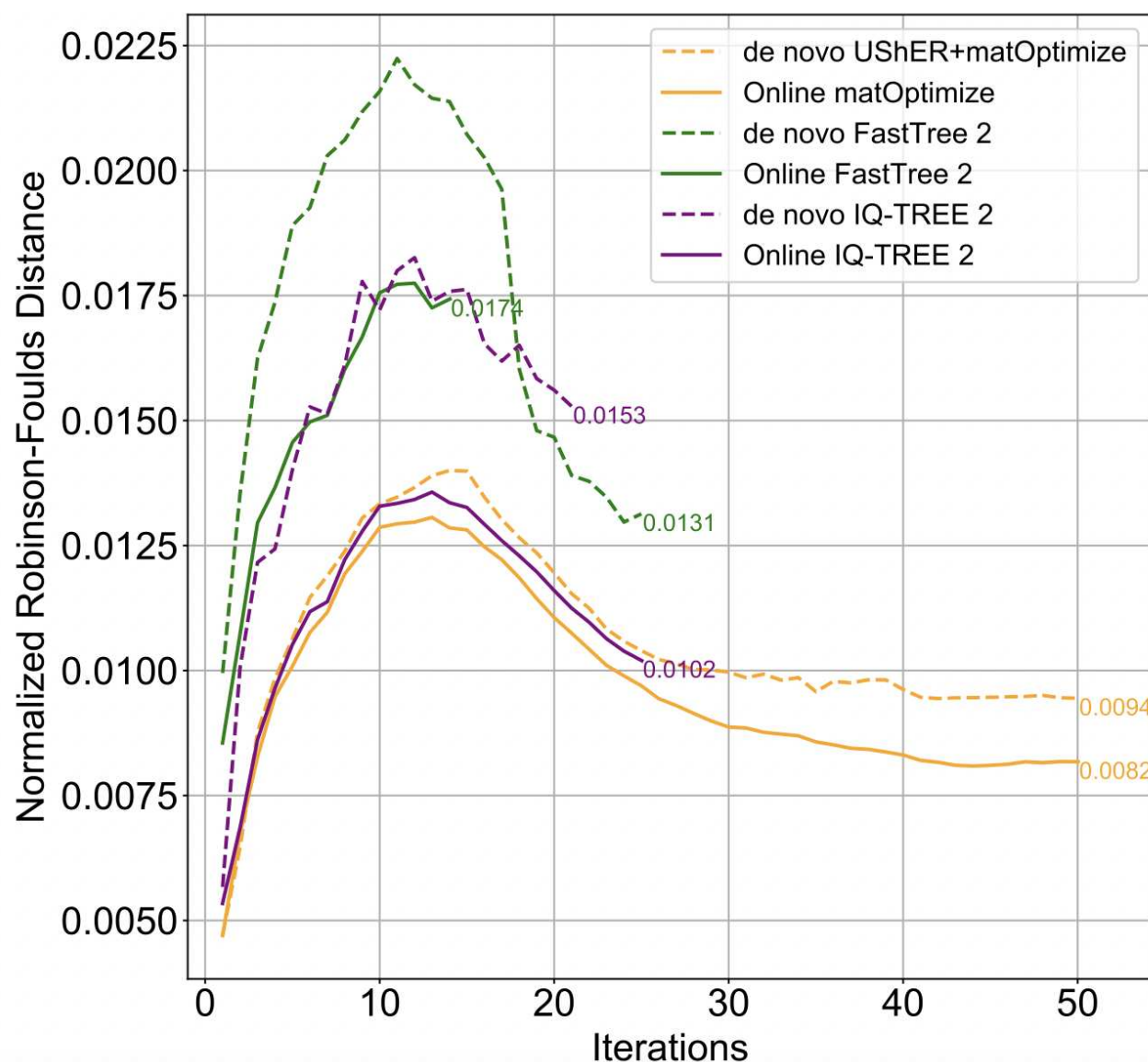176 **Analyses using simulated data suggest that online phylogenetics is more accurate for SARS-**

177 **CoV-2.**

178

179   We first compared matOptimize (commit 66ca5ff, conda version 0.4.8) (Ye et al. 2022), IQ-TREE

180 2 (Minh et al. 2020), and FastTree 2 (Price et al. 2010) using both online and *de novo* phylogenetics

181 strategies using simulated data that we designed to closely mimic real SARS-CoV-2 datasets. All online

182 phylogenomics workflows used UShER (Turakhia et al. 2021a) to add new sequences to the previous

183 tree (see Methods) as to our knowledge it is the only software package that is fast enough to perform

184 under real time constraints. We chose these three tools based on their widespread usage among SARS-

185 CoV-2 phylogenetics applications (*e.g.* matOptimize is part of the UShER suite (Turakhia et al. 2021a),

186 IQ-TREE 2 is used by (COVID-19 Genomics UK (COG-UK) Consortium 2020; Lanfear and Mansfield

187 2020) and FastTree 2 is used by (Hadfield et al. 2018)) as well as to cover several different

188 methodologies.

189   Simulating an alignment based on a known tree ensures that there is a ground truth for

190 comparison to definitively assess each optimization method. We used an inferred global phylogeny as a

191 template to simulate a complete multiple sequence alignment using phastSim (De Maio et al. 2021b). We

192 subsampled this simulated alignment into 50 progressively larger sets of samples, ranging in number of

193 samples from 4,676 to 233,326 (see Methods), to examine each of the three optimization methods in

194 both online and *de novo* phylogenetics. We then computed the Robinson-Foulds distance for unrooted

195 trees of each iteration, after condensing identical samples and collapsing very short branches, to the

196 global mutation-annotated tree on which the simulation was based, pruned to contain only the relevant

197 samples, and normalized by the maximum possible Robinson-Foulds distance between the trees (Fig. 2,

198 Fig. S3) (Steel and Penny 1993).

199   All online phylogenetics methods noticeably outperformed their *de novo* counterparts. Overall,

200   online matOptimize produced phylogenies with the lowest Robinson-Foulds distance to the ground truth

201   for the majority of iterations (Fig. 2). Online IQ-TREE 2 performed similarly, but was able to complete

202   only 25 of the 50 iterations due to its extreme computational resource requirements. For example, for the

203   14th phylogeny of 60,571 sequences, which was the last phylogeny produced using under 200 GB of

204   RAM in under 24 hours by all six methods, we found Robinson-Foulds distances of 1696, 2590, and

205   2130 for *de novo UShER*+matOptimize, FastTree 2, and IQ-TREE 2 respectively, and distances of 1557,

206   2111, and 1618 for online matOptimize, FastTree 2, and IQ-TREE 2, respectively.

207   There are several possible explanations for the improved performance of online phylogenetics

208   relative to *de novo* approaches. First, the radius for SPR moves when optimizing a large tree is

209   insufficiently large to find improvements that are more readily applied when the tree contains fewer

210   samples as in early rounds of online phylogenetics. In online phylogenetics, these improvements carry

211   over to subsequent trees, while in *de novo*, they do not. The radius is defined as the phylogenetic

212   distance of the search space when moving a node to a more optimal position. As the phylogeny

213   increases in size, the distance from a node to its optimal position is likely to also increase, necessitating

214   a larger SPR move radius to make equivalent improvements in larger trees. Second, large clades

215   consisting primarily of samples with branch length zero might further reduce the ability of optimization

216   methods to find improvements by indirectly limiting search space due to the increased number of edges

217   when represented internally as a bifurcating tree. It may sometimes be possible to explore moves across

218   such tree regions during online phylogenetics in early iterations when the polytomy is relatively small.

219   Third, online phylogenetics facilitates tree optimization by providing an exceptionally good initial tree that

220   has already been heavily optimized in previous iterations. We expect that this approach will typically

221   outperform parsimony and neighbor-joining initial trees that are used in most *de novo* phylogenetic

222   inference approaches. Finally, because each online experiment began with a small tree inferred *de novo*

223   by stepwise sample addition with UShER, it is possible that these initial trees are more optimal than initial

224   trees produced during *de novo* inference by the other software we evaluated, perhaps because UShER

225   prefers the reference nucleotide in cases of ambiguous internal character states.

226

227

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 (see Figure S3), 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.

235

236

237 **Analyses using real data suggest that online phylogenetics is more efficient than *de novo* and**

238 **produces similarly optimal phylogenies.**

239

240 While analyses using simulated data offer the ability to compare to a known ground truth,

241 assessing the performance of each method on real SARS-CoV-2 data may more accurately reflect

242 practical use of each method. Therefore, we also tested each optimization strategy on 50 progressively

243 larger sets of real SARS-CoV-2 samples and calculated the parsimony score and likelihood of each

244 optimized tree, as well as the run-time and peak RAM usage of each software package used (Fig. 3). To

245 accomplish this, we subsampled our global phylogeny, which was produced using stringent quality

246 control steps (see Methods), as before, to mimic the continuous accumulation of samples over the

247 course of the pandemic.

248 Online optimizations are generally much faster than *de novo* phylogenetic inference. For

249 example, IQ-TREE 2 achieves a roughly four-fold faster run-time for online optimizations compared to

250 inferring the tree *de novo* (Fig. 3c). The 11th iteration, which has 47,819 sequences and was the last to

251 be completed by both online and *de novo* IQ-TREE 2, took 22 hours 50 minutes for *de novo* IQ-TREE 2

252 but only 5 hours 26 minutes for online IQ-TREE 2. *De novo* UShER+matOptimize was the only *de novo*

253 method to finish all trees in fewer than 24 hours, but its speed for each daily update pales in comparison

254 to online matOptimize. Online matOptimize is several orders of magnitude faster than its *de novo*

255 counterpart, and its optimizations for the largest phylogenies take roughly 30 seconds, while *de novo* tree

256 inference with UShER can take several hours for trees consisting of more than 100,000 samples (Fig.

257 3c). However, whether a software package is used for online or *de novo* phylogenetics does not strongly

258 affect its peak memory usage.

259 We also found that online phylogenetics strategies produce trees very similar in both parsimony

260 score and likelihood to their *de novo* counterparts, with differences of less than 1% in all cases (Fig. 3a-

261 b). For example, in the 11th iteration containing 47,819 sequences, online IQ-TREE 2 produces a tree

262 with a parsimony score of 32,005, whereas *de novo* IQ-TREE 2 produces a tree with parsimony score

263 32,149. Our results suggest that in addition to the computational savings that allow online phylogenetics

264 approaches to continuously stay up-to-date, online phylogenetics approaches also produce trees with

265 similar parsimony scores and likelihoods to their *de novo* counterparts.

266

267 **Under pandemic time constraints, parsimony-based optimization methods have favorable metrics**

268 **compared to ML methods for SARS-CoV-2 phylogenies.**

269

270 In the case of both *de novo* and online phylogenetics, the parsimony-based matOptimize

271 outperforms both FastTree 2 and IQ-TREE 2 in runtime and peak memory usage. For the sixth iteration

272 (26,486 samples), which was the largest phylogeny inferred by all online methods in under 24 hours and

273 using under 200 GB of RAM, online FastTree 2 required nearly 24 hours and 30.3 GB of RAM, and

274 online IQ-TREE 2 required 1 hour 45 minutes and 72 GB of RAM. By contrast, matOptimize used only 6

275 seconds and 0.15 GB of RAM. This iteration contained roughly 10% as many samples as the 50th and

276 final iteration (233,326 total samples), which online matOptimize completed in 32 seconds using 1.41 GB

277 of RAM at peak usage. Even this largest tree represents only a very small fraction of the more than 10

278 million currently available SARS-CoV-2 genomes, indicating that, among the approaches we evaluated,

279 matOptimize is the only viable option for maintaining a comprehensive SARS-CoV-2 phylogeny via

280 online phylogenetics.

281 In addition to its scalability, matOptimize outperforms ML optimization methods under 24-hour

282 time constraints in both the parsimony and likelihood scores of the trees that it infers. For the sixth

283 iteration (26,486 samples), we found parsimony scores of 16,130, 16,179, and 16,290 for online

284 matOptimize, IQ-TREE 2, and FastTree 2 respectively. While all methods produce phylogenies with

285 parsimony scores within 1% of each other, matOptimize is consistently the lowest. However,

286 matOptimize was developed to optimize by parsimony, while the other methods were developed for ML

287 optimizations. Unexpectedly, we found log-likelihood scores of -233,414.277, -233,945.528, and -

288 235,177.396 for matOptimize, IQ-TREE 2, and FastTree 2 respectively, indicating that matOptimize

289 produces preferable phylogenies based on likelihood as well. We used a Jukes-Cantor (JC) model to

290 calculate likelihoods due to time constraints in calculation for more complex substitution models, but a

291   Generalised Time Reversible (GTR) model with specified rate parameters produced strongly correlated

292   likelihoods (Fig. S1). Specifically, we fit a generalized linear model using a Gamma family (inverse link

293   function) to predict the likelihood of the tree under the JC model using the iteration of tree construction

294   and the GTR likelihood as predictors. We examined the six trees from the first and second iteration (12 in

295   total). We found that the GTR likelihood was significantly correlated with the JC likelihood ($p < 2.27 \times 10^{-5}$).

296

297

298   **Parsimony optimization produces comparable or more favorable SARS-CoV-2 trees than the most**

299   **thorough maximum likelihood methods.**

300

301   We also compared the performance of *de novo* inference with UShER+matOptimize to state-of-

302   the-art methods without a 24-hour limit on runtime. In three iterations of increasing size (~4.5k, ~8.9k,

303   and ~13.2k samples), we inferred trees from real and simulated data using UShER+matOptimize, IQ-

304   TREE 2 with stochastic search enabled, and RAxML-NG. With the parameters used here, IQ-TREE 2

305   performs stochastic NNI moves in addition to hill-climbing NNI. RAxML-NG is a maximum likelihood

306   approach that uses SPR moves to search tree-space for higher likelihood phylogenies (Kozlov et al.

307   2019). We allowed each experiment to run for up to two weeks. All programs completed successfully on

308   the first iteration. RAxML-NG did not terminate within two weeks for the second and third iterations. On

309   real data, we found that UShER+matOptimize produced trees with higher log-likelihoods than IQ-TREE 2

310   and RAxML-NG across all three iterations (Fig. 4A). Under the substitution model parameters estimated

311   by IQ-TREE 2, the log-likelihoods for the first iteration were -73780.756, -73828.271, and -73782.289 for

312   UShER+matOptimize, IQ-TREE 2, and RAxML-NG respectively. Under the parameters estimated by

313   RAxML-NG, the log-likelihoods for the first iteration were -73754.894, -73801.935, and -73756.246 for

314   UShER+matOptimize, IQ-TREE 2, and RAxML-NG respectively. On simulated data,

315   UShER+matOptimize produced trees closer to the ground truth than the other methods when measured

316   by quartet distance across all three iterations (Fig. 4B). By RF distance, the UShER+matOptimize trees

317   were closest to the ground truth for the second and third iterations, but the RAxML-NG tree was closest

318   to ground truth in the first iteration (Fig. 4C). We therefore conclude that parsimony-based tree inference

319   can perform equivalently or better than state of the art maximum likelihood approaches but do this in a

320   tiny fraction of the time, making it by far the most suitable approach for pandemic-scale phylogenetics of
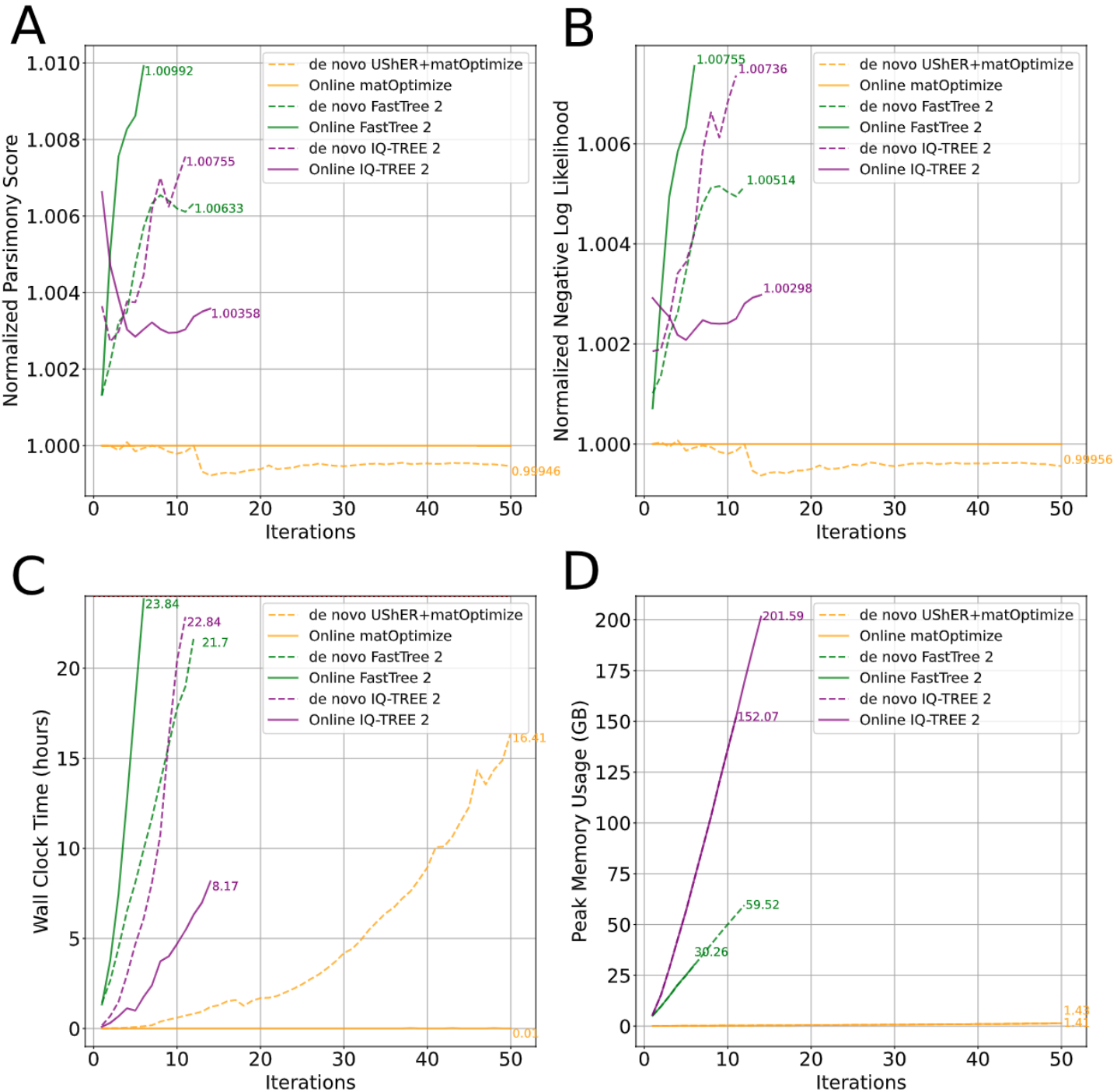
321   SARS-CoV-2.

322

323

324 ***Figure 3: In practice, optimization by parsimony is more effective for SARS-CoV-2 data than***

325 ***optimization by ML.*** *We calculated (A) the parsimony score for each tree using matUtils, (B) the log-*

326 *likelihood of each tree using IQ-TREE 2, (C) runtime and (D) peak memory usage of each optimization.*

327 *(A) and (B) are normalized by the value obtained for the matOptimize online approach such that all other*

328 *methods are expressed as a ratio. Strategies that surpassed 24 hours (C) or the allowable RAM usage*

329 *(D) were terminated prior. In most cases, with the notable exception of FastTree 2, online phylogenetics*

330 *(solid lines) perform better than de novo phylogenetics (dashed lines). We ran all matOptimize analyses*

331 *using an instance with 15 CPUs and 117.2 GB of RAM, and we ran all IQ-TREE 2 and FastTree 2*

332 *analyses on an instance with 31 CPUs and 244.1 GB of RAM, but limited each command to 15 threads*
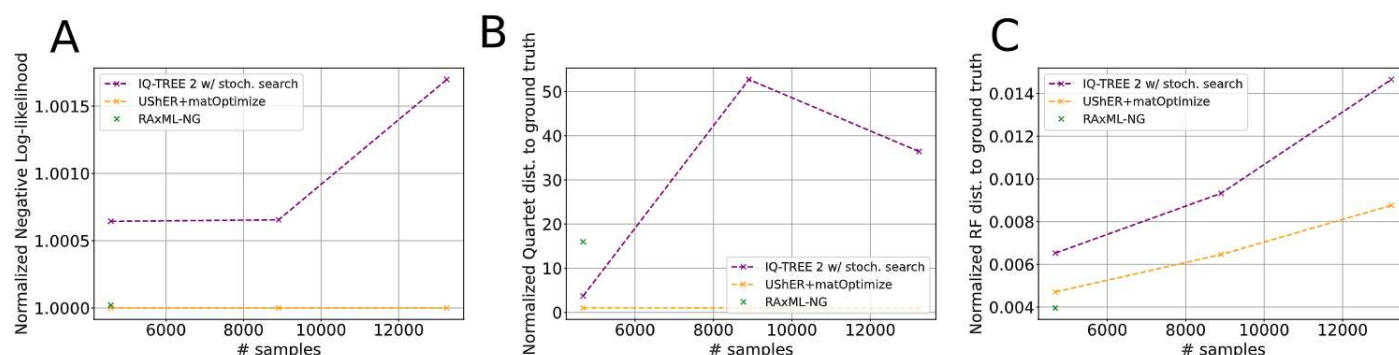
333 *for equivalence with matOptimize.*

334

335

336 ***Figure 4: de novo matOptimize produces similar or more favorable trees compared to the most***

337 ***thorough maximum-likelihood inference program****s. We ran these methods for up to two weeks each*

338 *to infer trees de novo from the three smallest iterations of real and simulated data. For real data **(A),** log-*

339 *likelihoods were computed under the model parameters estimated by IQ-TREE 2 at each iteration.*

340 *Values are normalized by the value of the matOptimize approach such that all other methods are*

341 *expressed as a ratio. For simulated data **(B, C)**, the reported quartet distances **(B)** are similarly*

342 *normalized by the value of the matOptimize approach such that other methods are expressed as a ratio.*

343 *RF distances **(C)** are normalized by the maximum possible RF distance of 2n-6, where n is the number*

344 *of leaves in each tree. For all panels, the second and third iterations of RAxML-NG (which did not*

345 *terminate within two weeks) are omitted.*

346

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

349

350   While our comparisons of online and *de novo* as well as parsimony-based and ML optimizations

351   of cumulative pandemic-style data demonstrated practical performance, the largest trees completed by

352   all methods in these experiments represent only a small fraction of available SARS-CoV-2 data. It is also

353   crucial that we identify the optimal ways to produce a large phylogeny from already aggregated data. We

354   therefore evaluated phylogenetic inference methods for optimizing a tree of 364,427 SARS-CoV-2

355   genome sequences, without constraining methods according to time or memory requirements. We

356   optimized this global phylogeny using matOptimize (Ye et al. 2022), IQ-TREE 2 (Minh et al. 2020), and

357   FastTree 2 (Price et al. 2010). Overall, we found that matOptimize produced the tree with the lowest

358   parsimony score across all methods in roughly one hour (Table 1).

359   We found that after each of the six iterations of FastTree 2 optimization, the likelihood and

360   parsimony improvements are strongly linearly correlated (Fig. 5). This suggests that changes achieved

361   by maximizing parsimony will also optimize likelihood for SARS-CoV-2 data. That is, for extremely

362   densely sampled phylogenies wherein long branches are especially rare, parsimony and likelihood of

363   phylogenies, and tree moves to optimize either are highly correlated. However, despite the strength of

364   this correlation, we find an extreme disparity in practical usage when optimizing by either metric.

365   Parsimony-based methods are far more time- and data-efficient, and presently available ML approaches

366   quickly become prohibitively expensive. For example, while the 6 iterations of FastTree did result in large

367   improvements in both likelihood and parsimony score, the resulting tree would be out of date long before

368   the 10.5-day optimization had completed. Moreover, we applied matOptimize to the tree output by the

369   sixth iteration of FastTree, achieving a parsimony score of 293,866 (improvement of 288) in just 16

370   minutes, indicating that even after 10.5 days, additional optimization was still possible. This suggests

371   that, for the purposes of optimizing even moderately large SARS-CoV-2 trees, parsimony-based

372   methods should be heavily favored due to their increased efficiency.

373

374

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)

375

376 **Table 1:** We applied each of the three optimization methods to a starting tree of 364,427 SARS-CoV-2

377 samples, which had an initial parsimony score of 296,247. We first ran 2 iterations of IQ-TREE 2

378 optimization, using an SPR radius of 20 on the first and 100 on the second. We also used an SPR radius

379 of 10 on one iteration of matOptimize, and six iterations of pseudo-likelihood optimization using FastTree

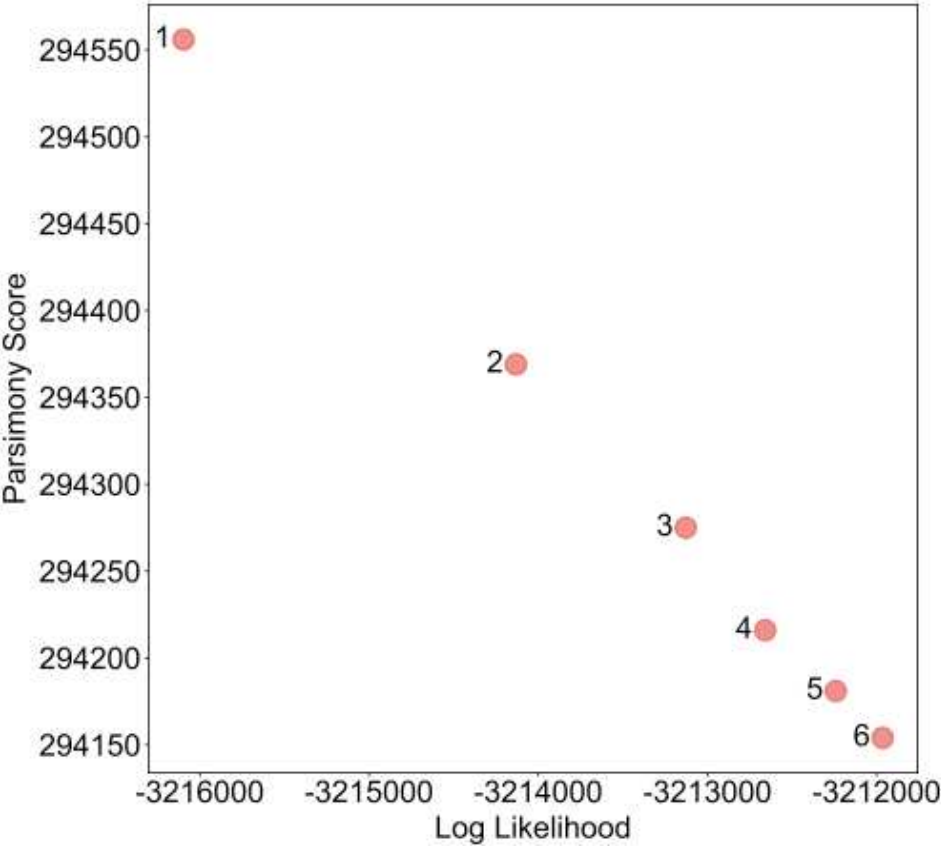380 2, which we terminated after roughly 10.5 days.

381

382 *Figure 5. Improvement in likelihood and parsimony have a linear relationship for our optimized*

383 *global tree. We optimized our initial global tree using 6 iterations of FastTree and measured the total*

384 *parsimony and the likelihood after each, finding a linear relationship (Pearson correlation, rho = -1.0, p <*

385 *2.9×10^{-7}).*

386

387 **Conclusions**

388

389 The SARS-CoV-2 pandemic has made phylogenetics central to efforts to combat the spread of

390 the virus, but has posed challenges for many commonly used phylogenetics frameworks. A major

391 component of this effort relies on a comprehensive, up-to-date, global phylogeny of SARS-CoV-2

392 genomes. However, the scale and continuous growth of the data have caused difficulties for standard *de*

393 *novo* phylogenetic methods. Here, we find that online phylogenetics methods are practical, pragmatic,

394 and accurate for inferring daily phylogenetic trees from a large and densely-sample virus outbreak.

395 One counterintuitive result is that parsimony-based optimizations outperform sufficiently efficient

396 ML approaches regardless of whether phylogenies are evaluated using parsimony or likelihood. This

397 might be a consequence of the fact that parsimony scores and likelihoods are strongly correlated across

398 phylogenies inferred via a range of phylogenetic approaches. The extremely short branches (Fig. S2) on

399 SARS-CoV-2 phylogenies mean that the probability of multiple mutations occurring at the same site on a

400 single branch is negligible. Stated another way, SARS-CoV-2 is approaching a "limit" where parsimony

401 and likelihood are nearly equivalent. In turn, because of their relative efficiency, parsimony-based

402 methods are able to search more of the possible tree space in the same amount of time, thereby

403 resulting in trees with better likelihoods and lower parsimony scores than trees optimized using currently-

404 available ML software packages. We emphasize that this does not bear on the relative merits of the

405 underlying principles of ML and MP, but instead reflects the utility of methods that have been applied

406 during the pandemic. Nevertheless, this observation does suggest that in some cases, MP optimization

407 may provide a fast and accurate starting point for ML optimization methods. Indeed, many popular

408 phylogenetics software, such as RAxML (Stamatakis 2014) and IQ-TREE (Minh et al. 2020) already use

409   stepwise-addition parsimony trees as initial trees for their optimization. Our results suggest that further

410   optimization of these initial trees using MP may provide benefits in speed *and* accuracy for some

411   datasets, even when the target is an estimate of the ML tree.

412   As sequencing technologies progress and become more readily available, sample sizes for

413   phylogenetic analyses of major pathogens and highly-studied organisms will necessarily continue to

414   increase. Today, SARS-CoV-2 represents an extreme with respect to the total number of samples

415   relative to the very short branch lengths on the phylogeny. However, the global sequencing effort during

416   the pandemic suggests that the public health sphere has a strong interest in the increased application of

417   whole-genome sequencing to study the genomic contents, evolution, and transmission history of major

418   and emerging human pathogens. We expect that million-sample datasets will become commonplace in

419   the near future. Parsimony-based methods like matOptimize show promise for huge datasets with short

420   branch lengths. Similarly, recently developed parsimony-based likelihood approximations may ultimately

421   be similarly scalable and accurate (De Maio et al. 2022). Online phylogenetics using both of these

422   methods will be a fruitful avenue for future development and application to accommodate these datasets.

423

424   **Methods**

425

426   We first developed a "global phylogeny", from which all analyses in this study were performed.

427   We began by downloading VCF and FASTA files corresponding to March 18, 2021 from our own daily-

428   updated database (McBroome et al. 2021). The VCF file contains pairwise alignments of each of the

429   434,063 samples to the SARS-CoV-2 reference genome. We then implemented filters, retaining only

430   sequences containing at least 28,000 non-N nucleotides, and fewer than two non-[ACGTN-] characters.

431   We used UShER to create a phylogeny from scratch using only the remaining 366,492 samples. To

432   remove potentially erroneous sequences, we iteratively pruned this tree of highly divergent internal

433   branches with branch parsimony scores greater than 30, then terminal branches with branch parsimony

434   scores greater than 6, until convergence, resulting in a final global phylogeny containing 364,427

435   samples. The branch parsimony score indicates the total number of substitutions along a branch. Similar

436   filters based on sequence divergence are used by existing SARS-CoV-2 phylogenetic inference

437   methods. For full reproducibility, files used for creating the global phylogeny can be found in

438   subrepository 1 on the project GitHub page (Thornlow et al. 2021b).

439   Following this, we tested several optimization strategies on this global phylogeny, hereafter the

440   "starting tree". We used matOptimize, FastTree 2, and maximum parsimony (MP) IQ-TREE 2. MP IQ-

441   TREE 2 uses parsimony as the optimality criterion in contrast to the maximum likelihood mode used in all

442   other experiments, which was infeasible on a dataset of this size. In these optimization experiments, we

443   used experimental versions of MP IQ-TREE 2 that allow finer control of parsimony parameters (specific

444   versions are listed in the supplemental Github repository). In one experiment, we used the starting tree

445   and its corresponding alignment and ran five iterations of MP IQ-TREE 2, varying the SPR radius from

446   20 to 100 in increments of 20. Experiments on a small dataset indicated that there is little or no

447   improvement in parsimony score beyond a radius of 100. Separately, we tested another strategy that

448   applied two iterations of MP IQ-TREE 2 to the starting tree, the first iteration using an SPR radius of 20

449   and the second using a radius of 100. Finally, we tested a strategy of six iterations of pseudo-likelihood

450   optimization with FastTree 2 followed by two iterations of parsimony optimization with matOptimize. The

451   tree produced by this strategy, hereafter the "ground truth" tree, had the highest likelihood of all the

452   strategies we tested. This tree (after_usher_optimized_fasttree_iter6.tree) and files for these optimization

453   experiments can be found in subrepository 2.

454   In the multifurcating ground truth tree of 364,427 samples, there are 265,289 unique (in FASTA

455   sequence) samples. There are 447,643 nodes in the tree. For reference, a full binary tree with the same

456   number of leaves has 728,853 nodes. 23,437 of the 29,903 sites in the alignment are polymorphic (they

457   display at least two non-ambiguous nucleotides). Homoplasies are common in these data. In the starting

458   tree, 19,090 sites display a mutation occurring on at least two different branches, and 4,976 sites display

459   a mutation occurring more than ten times in the tree.

460   To mimic pandemic-style phylogenetics, we separated a total of 233,326 samples from the

461   starting tree of 364,427 samples into 50 batches of ~5,000 by sorting according to the date of sample

462   collection. We then set up two frameworks for each of the three software packages (matOptimize

463   (commit 66ca5ff, conda version 0.4.8), maximum-likelihood IQ-TREE 2 (multicore version 2.1.3 COVID-

464   edition), and FastTree 2 (Double Precision version 2.1.10)). The online phylogenetics frameworks began

465   by using UShER to infer a small tree *de novo* from the first batch of samples, followed by alternating

466   steps of optimization using one of the three evaluated methods and placement of additional samples with

467   UShER. In *de novo* phylogenetics, we supplied each software package with an alignment corresponding

468   to all samples in that batch and its predecessors (or VCF for matOptimize) without a guide tree. For both

469   cases, each tree is larger than its predecessor by ~5,000 samples, and each tree necessarily contains all

470   samples in the immediately preceding tree. For FastTree 2, we used 2 rounds of subtree-prune-regraft

471   (SPR) moves (-spr 2), maximum SPR length of 1000 (-sprlength 1000), zero rounds of minimum

472   evolution nearest neighbor interchanges (-nni 0), and the Generalised Time Reversible + Gamma

473   (GTR+G) substitution model (-gtr -gamma). For IQ-TREE 2, we used a branch length minimum of

474   0.000000001 (-blmin 1e-9), zero rounds of stochastic tree search (-n 0), and the GTR+G substitution

475   model (-m GTR+G). With these parameters, IQ-TREE 2 constructs a starting parsimony tree and then

476   performs hill-climbing NNI steps to optimize likelihood, avoiding the significant time overhead of

477   stochastic search. We ran all matOptimize analyses using an instance with 15 CPUs and 117.2 GB of

478   RAM, and we ran all IQ-TREE 2 and FastTree 2 analyses on an instance with 31 CPUs and 244.1 GB of

479   RAM, but we limited each command to 15 threads for equivalence with matOptimize. Files for all

480   simulated data experiments can be found in subrepository 3.
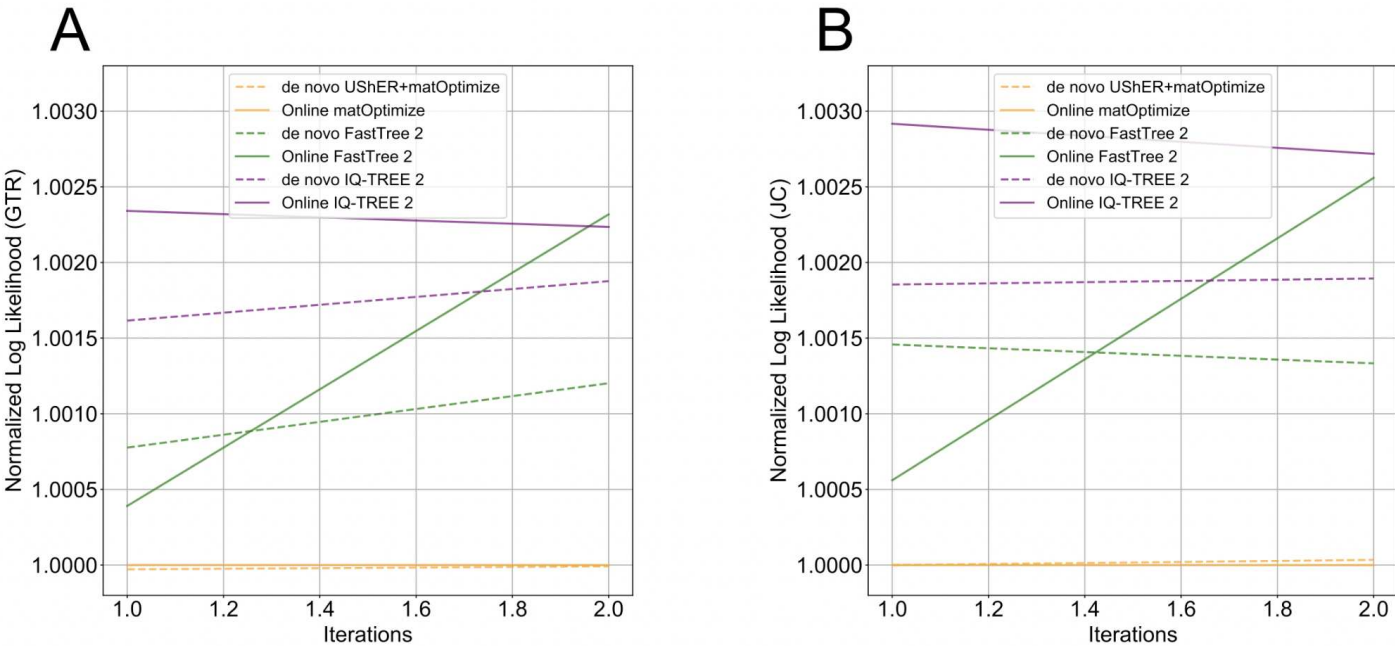
481   To generate our simulated data, we used the SARS-CoV-2 reference genome (GISAID ID:

482   EPI_ISL_402125; GenBank ID: MN908947.3) (Shu and McCauley 2017; Sayers et al. 2021) as the root

483   sequence and used phastSim (De Maio et al. 2021b) to simulate according to the ground truth phylogeny

484   described above. Intergenic regions were evolved using phastSim using the default neutral mutation

485   rates estimated in ref. (De Maio et al. 2021a), with position-specific mean mutation rates sampled from a

486   gamma distribution with alpha=beta=4, and with 1% of the genome having a 10-fold increase mutation

487   rate for one specific mutation type (SARS-CoV-2 hypermutability model described in ref. (De Maio et al.

488   2021b)). Evolution of coding regions was simulated with the same neutral mutational distribution, with a

489   mean nonsynonymous/synonymous rate ratio of omega=0.48 as estimated in (Turakhia et al. 2021a),

490   with codon-specific omega values sampled from a gamma distribution with alpha=0.96 and beta=2.

491   Rates for each intergenic and coding region were not normalized in order to have the same baseline

492   neutral mutation rate distribution across the genome.

493   We repeated our iterative experiments using *de novo* and online matOptimize, IQ-TREE 2 and

494   FastTree 2 on this simulated alignment, using the same strategies as before. However, instead of

495   computing parsimony and likelihood scores, we computed the Robinson-Foulds (RF) distance (Robinson

496   and Foulds 1981) of each optimization to the ground truth tree, pruned to contain only the samples

497   belonging to that batch. To calculate each RF distance, we used the -O (collapse tree) argument in

498   matUtils extract (McBroome et al. 2021) and then used the dist.topo command in the *ape* package in R

499   (Paradis and Schliep 2019), comparing the collapsed optimized tree and the pruned, collapsed ground

500   truth tree at each iteration. We computed normalized RF distances as a proportion of the total possible

501   RF distance, which is equivalent to two times the number of samples in the trees minus six (Steel and

502   Penny 1993).

503   Eliminating the 24-hour runtime restriction, we also repeated the first three *de novo* iterative

504   experiments on both real and simulated data to compare UShER+matOptimize, IQ-TREE 2 with

505   stochastic search, and RAxML-NG. These iterations of ~4.5k, ~8.9k, and ~13.2k samples were allowed

506   to run for up to 14 days. For runs that did not terminate within this time (the second and third iterations of

507   RAxML-NG), we used the best tree inferred during the run for comparisons. We ran IQ-TREE 2 and

508   RAxML-NG under the GTR+G model with the smallest minimum branch length parameter that did not

509   cause numerical errors. To compare the trees inferred from real data, we computed log-likelihoods under

510   the GTR+G model for all trees, fixing the model parameters to those estimated by IQ-TREE 2 during tree

511   inference. We also compared the log-likelihoods of the trees under the parameters estimated by RAxML-

512   NG for the first iteration, but could not do so for the second and third iterations which did not terminate in

513   under two weeks. We allowed optimization of branch lengths during likelihood calculation. For the

514   UShER+matOptimize trees, before computing likelihoods, we converted the branch lengths into units of

515   substitutions per site by dividing each branch length by the alignment length (29,903). To compare the

516   trees inferred from simulated data, we computed the RF and quartet distances of each tree to the

517   corresponding ground truth tree described above.

518

519

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

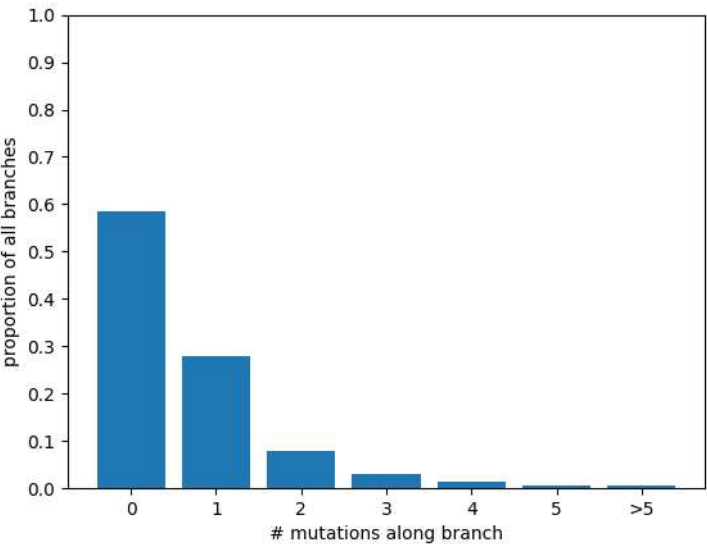*value obtained for the matOptimize online approach, such that other methods are expressed as a ratio.*

525

526 **Figure S2: Most branches in the ground truth phylogeny are extremely short.** *In our optimized*

527 *global SARS-CoV-2 phylogeny, the majority of branch lengths are zero. This low amount of divergence*

528 *yields many identical nodes in the tree and demonstrates that the probability of observing multiple*

529 *mutations at a single site along the same branch is negligible. These characteristics may help explain the*

530 *ability of parsimony-based inference methods to outperform likelihood optimization on SARS-CoV-2 data.*
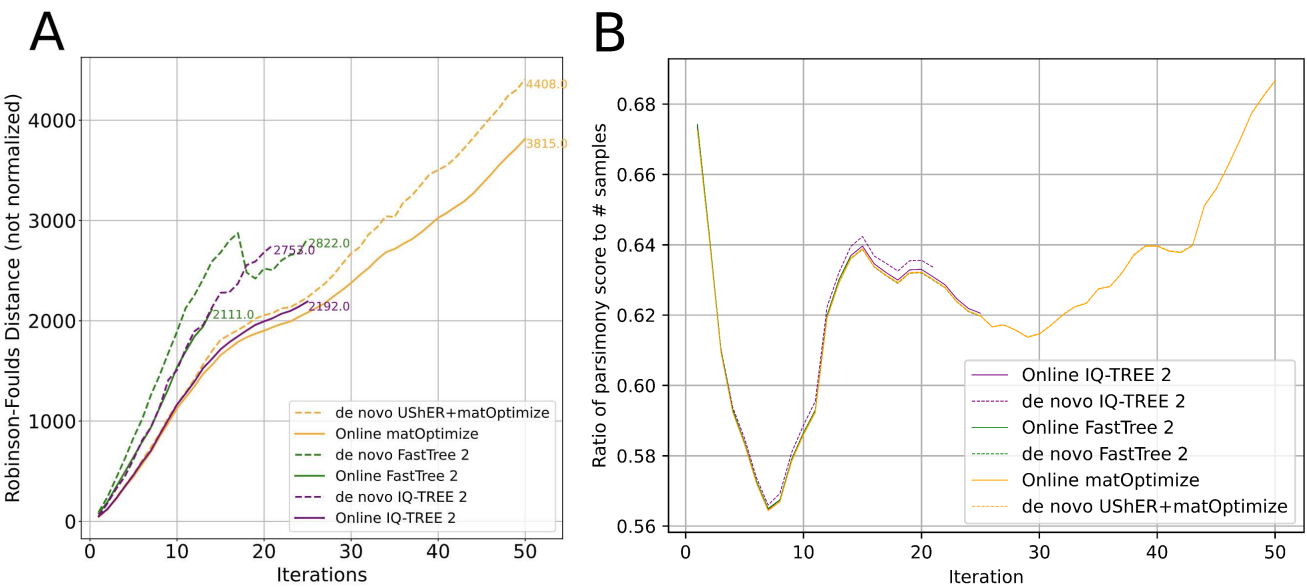

531

532 **Figure S3: Temporal patterns in simulated SARS-CoV-2 data may affect Robinson-Foulds (RF)**

533 **distance normalization**. *The RF distances for each tree in Figure 2 are normalized against the*

534 *maximum possible RF distance for that tree. While the raw RF distances are approximately continuously*

535 *increasing (A), they do not increase linearly with the maximum RF distance, leading to the pattern*

536 *observed in Figure 2. A potential explanation for this is the variation in sequence diversity over simulated*

537 *time. The ratio of the number of total mutations in the tree (parsimony score) to the number of samples in*

538 *the inferred trees at each iteration (B) approximates the average divergence between samples in each*

539 *tree. The initial drop in divergence per sample may contribute to the more rapid increase in RF distance*

540 *because there is less phylogenetic signal to facilitate the resolution of correct topologies. As the*

541 *divergence subsequently increases, tree inference improves before the RF distances stabilize and begin*

542 *to increase approximately linearly.*

543

555

556

557

558 **References:**

559   Annavajhala M.K., Mohri H., Wang P., Nair M., Zucker J.E., Sheng Z., Gomez-Simmonds A., Kelley A.L.,

560   Tagliavia M., Huang Y., Bedford T., Ho D.D., Uhlemann A.-C. 2021. A Novel and Expanding SARS-

561   CoV-2 Variant, B.1.526, Identified in New York. medRxiv.

562   Barbera P., Kozlov A.M., Czech L., Morel B., Darriba D., Flouri T., Stamatakis A. 2019. EPA-ng:

563   Massively Parallel Evolutionary Placement of Genetic Sequences. Syst. Biol. 68:365–369.

564   Berger S.A., Krompass D., Stamatakis A. 2011. Performance, accuracy, and Web server for evolutionary

565   placement of short sequence reads under maximum likelihood. Syst. Biol. 60:291–302.

566   Bluhm A., Christandl M., Gesmundo F., Klausen F.R., Mančinska L., Steffan V., França D.S., Werner

567   A.H. 2020. SARS-CoV-2 transmission routes from genetic data: A Danish case study. PLOS ONE.

568   15:e0241405.

569   Castillo A.E., Parra B., Tapia P., Acevedo A., Lagos J., Andrade W., Arata L., Leal G., Barra G., Tambley

570   C., Tognarelli J., Bustos P., Ulloa S., Fasce R., Fernández J. 2020. Phylogenetic analysis of the first

571   four SARS-CoV-2 cases in Chile. J. Med. Virol. 92:1562–1566.

572   COVID-19 Genomics UK (COG-UK) Consortium. 2020. An integrated national scale SARS-CoV-2

573   genomic surveillance network. Lancet Microbe. 1:e99–e100.

574   De Maio N., Kalaghatgi P., Turakhia Y., Corbett-Detig R., Minh B.Q., Goldman N. 2022. Maximum

575   likelihood pandemic-scale phylogenetics. bioRxiv.:2022.03.22.485312.

576   De Maio N., Walker C.R., Turakhia Y., Lanfear R., Corbett-Detig R., Goldman N. 2021a. Mutation Rates

577   and Selection on Synonymous Mutations in SARS-CoV-2. Genome Biol. Evol. 13.

578   De Maio N., Weilguny L., Walker C.R., Turakhia Y., Corbett-Detig R., Goldman N. 2021b. phastSim:

579   efficient simulation of sequence evolution for pandemic-scale datasets. bioRxiv.

580   Deng X., Gu W., Federman S., du Plessis L., Pybus O.G., Faria N.R., Wang C., Yu G., Bushnell B., Pan

581   C.-Y., Guevara H., Sotomayor-Gonzalez A., Zorn K., Gopez A., Servellita V., Hsu E., Miller S.,

582   Bedford T., Greninger A.L., Roychoudhury P., Starita L.M., Famulare M., Chu H.Y., Shendure J.,

583   Jerome K.R., Anderson C., Gangavarapu K., Zeller M., Spencer E., Andersen K.G., MacCannell D.,

584   Paden C.R., Li Y., Zhang J., Tong S., Armstrong G., Morrow S., Willis M., Matyas B.T., Mase S.,

585   Kasirye O., Park M., Masinde G., Chan C., Yu A.T., Chai S.J., Villarino E., Bonin B., Wadford D.A.,

586   Chiu C.Y. 2020. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern

587   California. Science. 369:582–587.

588   Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading.

589   Syst. Biol. 27:401–410.

590   Fourment M., Claywell B.C., Dinh V., McCoy C., Matsen F.A. Iv, Darling A.E. 2018. Effective Online

591   Bayesian Phylogenetics via Sequential Monte Carlo with Guided Proposals. Syst. Biol. 67:490–502.

592   Franceschi V.B., Caldana G.D., de Menezes Mayer A., Cybis G.B., Neves C.A.M., Ferrareze P.A.G.,

593   Demoliner M., de Almeida P.R., Gularte J.S., Hansen A.W., Weber M.N., Fleck J.D., Zimerman R.A.,

594   Kmetzsch L., Spilki F.R., Thompson C.E. 2021. Genomic epidemiology of SARS-CoV-2 in Esteio,

595   Rio Grande do Sul, Brazil. BMC Genomics. 22:371.

596   Gill M.S., Lemey P., Suchard M.A., Rambaut A., Baele G. 2020. Online Bayesian Phylodynamic

597   Inference in BEAST with Application to Epidemic Reconstruction. Mol. Biol. Evol. 37:1832–1842.

598   Gonzalez-Reiche A.S., Hernandez M.M., Sullivan M.J., Ciferri B., Alshammary H., Obla A., Fabre S.,

599   Kleiner G., Polanco J., Khan Z., Alburquerque B., van de Guchte A., Dutta J., Francoeur N., Melo

600   B.S., Oussenko I., Deikus G., Soto J., Sridhar S.H., Wang Y.-C., Twyman K., Kasarskis A., Altman

601   D.R., Smith M., Sebra R., Aberg J., Krammer F., García-Sastre A., Luksza M., Patel G., Paniz-

602   Mondolfi A., Gitman M., Sordillo E.M., Simon V., van Bakel H. 2020. Introductions and early spread

603   of SARS-CoV-2 in the New York City area. Science. 369:297–301.

604   Hadfield J., Megill C., Bell S.M., Huddleston J., Potter B., Callender C., Sagulenko P., Bedford T., Neher

605   R.A. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 34:4121–4123.

606   Hendy M.D., Penny D. 1989. A Framework for the Quantitative Study of Evolutionary Trees. Syst. Biol.

607   38:297–309.

608   Hug L.A., Baker B.J., Anantharaman K., Brown C.T., Probst A.J., Castelle C.J., Butterfield C.N.,

609   Hernsdorf A.W., Amano Y., Ise K., Suzuki Y., Dudek N., Relman D.A., Finstad K.M., Amundson R.,

610   Thomas B.C., Banfield J.F. 2016. A new view of the tree of life. Nat Microbiol. 1:16048.

611   Izquierdo-Carrasco F., Cazes J., Smith S.A., Stamatakis A. 2014. PUmPER: phylogenies updated

612   perpetually. Bioinformatics. 30:1476–1477.

613   Jackson B., Boni M.F., Bull M.J., Colleran A., Colquhoun R.M., Darby A.C., Haldenby S., Hill V., Lucaci

614   A., McCrone J.T., Nicholls S.M., O'Toole Á., Pacchiarini N., Poplawski R., Scher E., Todd F.,

615   Webster H.J., Whitehead M., Wierzbicki C., COVID-19 Genomics UK (COG-UK) Consortium, Loman

616   N.J., Connor T.R., Robertson D.L., Pybus O.G., Rambaut A. 2021. Generation and transmission of

617   interlineage recombinants in the SARS-CoV-2 pandemic. Cell. 184:5179–5188.e8.

618   Kalantar K.L., Carvalho T., de Bourcy C.F.A., Dimitrov B., Dingle G., Egger R., Han J., Holmes O.B.,

619   Juan Y.-F., King R., Kislyuk A., Lin M.F., Mariano M., Morse T., Reynoso L.V., Cruz D.R., Sheu J.,

620   Tang J., Wang J., Zhang M.A., Zhong E., Ahyong V., Lay S., Chea S., Bohl J.A., Manning J.E., Tato

621   C.M., DeRisi J.L. 2020. IDseq—An open source cloud-based pipeline and analysis service for

622   metagenomic pathogen detection and monitoring. Gigascience. 9.

623   Khan A., Zia T., Suleman M., Khan T., Ali S.S., Abbasi A.A., Mohammad A., Wei D.-Q. 2021. Higher

624   infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y

625   mutants: An insight from structural data. J. Cell. Physiol. 236:7045–7057.

626   Kolaczkowski B., Thornton J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics

627   when evolution is heterogeneous. Nature. 431:980–984.

628   Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-

629   friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35:4453–4455.

630   Lam T.T.-Y. 2020. Tracking the Genomic Footprints of SARS-CoV-2 Transmission. Trends Genet.

631   36:544–546.

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

633   Li X., Giorgi E.E., Marichannegowda M.H., Foley B., Xiao C., Kong X.-P., Chen Y., Gnanakaran S.,

634   Korber B., Gao F. 2020. Emergence of SARS-CoV-2 through recombination and strong purifying

635   selection. Sci Adv. 6.

636   Lu J., du Plessis L., Liu Z., Hill V., Kang M., Lin H., Sun J., François S., Kraemer M.U.G., Faria N.R.,

637   McCrone J.T., Peng J., Xiong Q., Yuan R., Zeng L., Zhou P., Liang C., Yi L., Liu J., Xiao J., Hu J.,

638   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

639   M., Quick J., Song T., Rambaut A., Loman N., Raghwani J., Pybus O.G., Ke C. 2020a. Genomic

640   Epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell. 181:997–1003.e9.

641   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

642   F., Wang L., Hu T., Zhou H., Hu Z., Zhou W., Zhao L., Chen J., Meng Y., Wang J., Lin Y., Yuan J.,

643   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.

644   2020b. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus

645   origins and receptor binding. Lancet. 395:565–574.

646   Matsen F.A., Kodner R.B., Armbrust E.V. 2010. pplacer: linear time maximum-likelihood and Bayesian

647   phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics. 11:538.

648   McBroome J., Thornlow B., Hinrichs A.S., Kramer A., De Maio N., Goldman N., Haussler D., Corbett-

649   Detig R., Turakhia Y. 2021. A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2

650   Mutation-Annotated Trees. Mol. Biol. Evol. 38:5819–5824.

651   Meredith L.W., Hamilton W.L., Warne B., Houldcroft C.J., Hosmillo M., Jahun A.S., Curran M.D., Parmar

652   S., Caller L.G., Caddy S.L., Khokhar F.A., Yakovleva A., Hall G., Feltwell T., Forrest S., Sridhar S.,

653   Weekes M.P., Baker S., Brown N., Moore E., Popay A., Roddick I., Reacher M., Gouliouris T.,

654   Peacock S.J., Dougan G., Török M.E., Goodfellow I. 2020. Rapid implementation of SARS-CoV-2

655   sequencing to investigate cases of health-care associated COVID-19: a prospective genomic

656   surveillance study. Lancet Infect. Dis. 20:1263–1271.

657 Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R.

658   2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic

659   Era. Mol. Biol. Evol. 37:1530–1534.

660 Moreno G.K., Braun K.M., Riemersma K.K., Martin M.A., Halfmann P.J., Crooks C.M., Prall T., Baker D.,

661   Baczenas J.J., Heffron A.S., Ramuta M., Khubbar M., Weiler A.M., Accola M.A., Rehrauer W.M.,

662   O'Connor S.L., Safdar N., Pepperell C.S., Dasu T., Bhattacharyya S., Kawaoka Y., Koelle K.,

663   O'Connor D.H., Friedrich T.C. 2020. Revealing fine-scale spatiotemporal differences in SARS-CoV-2

664   introduction and spread. Nat. Commun. 11:5558.

665 Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary

666   analyses in R. Bioinformatics.

667 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.,

668   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.

669   Genomic Surveillance of SARS-CoV-2: Distribution of Clades in the Republic of Korea in 2020.

670   Osong Public Health Res Perspect. 12:37–43.

671 Parks D.H., Chuvochina M., Waite D.W., Rinke C., Skarshewski A., Chaumeil P.-A., Hugenholtz P. 2018.

672   A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.

673   Nat. Biotechnol. 36:996–1004.

674 Peng J., Liu J., Mann S.A., Mitchell A.M., Laurie M.T., Sunshine S., Pilarowski G., Ayscue P., Kistler A.,

675   Vanaerschot M., Li L.M., McGeever A., Chow E.D., Marquez C., Nakamura R., Rubio L., Chamie G.,

676   Jones D., Jacobo J., Rojas S., Rojas S., Tulier-Laiwa V., Black D., Martinez J., Naso J., Schwab J.,

677   Petersen M., Havlir D., DeRisi J., IDseq Team. 2021. Estimation of secondary household attack

678   rates for emergent spike L452R SARS-CoV-2 variants detected by genomic surveillance at a

679   community-based testing site in San Francisco. Clin. Infect. Dis.

680 Philippe H., Zhou Y., Brinkmann H., Rodrigue N., Delsuc F. 2005. Heterotachy and long-branch

681   attraction in phylogenetics. BMC Evol. Biol. 5:50.

682 Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2--approximately maximum-likelihood trees for large

683   alignments. PLoS One. 5:e9490.

684 Rambaut A., Holmes E.C., O'Toole Á., Hill V., McCrone J.T., Ruis C., du Plessis L., Pybus O.G. 2020. A

685   dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature

686   Microbiology. 5:1403–1407.

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

688 Sanderson T. 2021a. taxonium: Explore very large trees in the browser. Github.

689   https://github.com/theosanderson/taxonium

690 Sanderson T. 2021b. Chronumental: time tree estimation from very large phylogenies.

691   bioRxiv.:2021.10.27.465994.

692 Sayers E.W., Cavanaugh M., Clark K., Pruitt K.D., Schoch C.L., Sherry S.T., Karsch-Mizrachi I. 2021.

693   GenBank. Nucleic Acids Res. 49:D92–D96.

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

695   Eurosurveillance. 22.

696 Skidmore P.T., Kaelin E.A., Holland L.R.A., Maqsood R. 2021. Emergence of a SARS-CoV-2 E484K

697   variant of interest in Arizona. medRxiv.

698 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large

699   phylogenies. Bioinformatics. 30:1312–1313.

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

701 Sullivan J., Swofford D.L. 2001. Should We Use Model-Based Methods for Phylogenetic Inference When

702   We Know That Assumptions About Among-Site Rate Variation and Nucleotide Substitution Pattern

703   Are Violated? Systematic Biology. 50:723–729.

704 Tang J.W., Toovey O.T.R., Harvey K.N., Hui D.D.S. 2021. Introduction of the South African SARS-CoV-2

705   variant 501Y.V2 into the UK. J. Infect. 82:e8–e10.

706 Tegally H., Wilkinson E., Giovanetti M., Iranzadeh A., Fonseca V., Giandhari J., Doolabh D., Pillay S.,

707   San E.J., Msomi N., Mlisana K., von Gottberg A., Walaza S., Allam M., Ismail A., Mohale T., Glass

708   A.J., Engelbrecht S., Van Zyl G., Preiser W., Petruccione F., Sigal A., Hardie D., Marais G., Hsiao

709   N.-Y., Korsman S., Davies M.-A., Tyers L., Mudau I., York D., Maslo C., Goedhals D., Abrahams S.,

710   Laguda-Akingba O., Alisoltani-Dehkordi A., Godzik A., Wibmer C.K., Sewell B.T., Lourenço J.,

711   Alcantara L.C.J., Kosakovsky Pond S.L., Weaver S., Martin D., Lessells R.J., Bhiman J.N.,

712   Williamson C., de Oliveira T. 2021. Detection of a SARS-CoV-2 variant of concern in South Africa.

713   Nature. 592:438–443.

714 Thornlow B., Hinrichs A.S., Jain M., Dhillon N., La S., Kapp J.D., Anigbogu I., Cassatt-Johnstone M.,

715   McBroome J., Haeussler M., Turakhia Y., Chang T., Olsen H.E., Sanford J., Stone M., Vaske O.,

716   Bjork I., Akeson M., Shapiro B., Haussler D., Kilpatrick A.M., Corbett-Detig R. 2021a. A new SARS-

717   CoV-2 lineage that shares mutations with known Variants of Concern is rejected by automated

718   sequence repository quality control. bioRxiv.

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

720 Tian F., Tong B., Sun L., Shi S., Zheng B., Wang Z., Dong X., Zheng P. 2021. Mutation N501Y in RBD of

721   Spike Protein Strengthens the Interaction between COVID-19 and its Receptor ACE2.

722   bioRxiv.:2021.02.14.431117.

723 Turakhia Y., Thornlow B., Hinrichs A.S., De Maio N., Gozashti L., Lanfear R., Haussler D., Corbett-Detig

724   R. 2021a. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics

725   for the SARS-CoV-2 pandemic. Nat. Genet. 53:809–816.

726 Turakhia Y., Thornlow B., Hinrichs A.S., Mcbroome J. 2021b. Pandemic-Scale phylogenomics reveals

727   elevated recombination rates in the SARS-CoV-2 spike region. bioRxiv.

728 Umair M., Ikram A., Salman M., Khurshid A., Alam M., Badar N., Suleman R., Tahir F., Sharif S.,

729   Montgomery J., Whitmer S., Klena J. 2021. Whole-genome sequencing of SARS-CoV-2 reveals the

730   detection of G614 variant in Pakistan. PLoS One. 16:e0248371.

731 Wang W., Barbetti J., Wong T., Thornlow B., Corbett-Detig R., Turakhia Y., Lanfear R., Minh B.Q. 2022.

732   DecentTree: Scalable Neighbour-Joining for the Genomic Era. bioRxiv.:2022.04.10.487712.

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

734 Ye C., Thornlow B., Hinrichs A., Torvi D., Lanfear R., Corbett-Detig R., Turakhia Y. 2022. matOptimize: A

735   parallel tree optimization method enables online phylogenetics for SARS-CoV-2.

736   bioRxiv.:2022.01.12.475688.

737

738

739

740

741