Bioinformatics, 38(3), 2022, 856–860
https://doi.org/10.1093/bioinformatics/btab725
Advance Access Publication Date: 21 October 2021
Applications Note

Phylogenetics

Optimizing viral genome subsampling by genetic diversity and temporal distribution (TARDiS) for phylogenetics

Simone Marini[1,2], Carla Mavian[2,3], Alberto Riva [4], Mattia Prosperi[1], Marco Salemi[2,3,*] and Brittany Rife Magalis [2,3,*]

[1]Department of Epidemiology, University of Florida, Gainesville, FL 32611, USA, [2]Emerging Pathogens Institute, University of Florida, Gainesville, FL 32611, USA, [3]Department of Pathology, University of Florida, Gainesville, FL 32611, USA and [4]Bioinformatics Core, Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL 32611, USA

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on January 20, 2021; revised on September 10, 2021; editorial decision on September 28, 2021; accepted on October 18, 2021

Abstract

Summary: TARDiS is a novel phylogenetic tool for optimal genetic subsampling. It optimizes both genetic diversity and temporal distribution through a genetic algorithm.

Availability and implementation: TARDiS, along with example datasets and a user manual, is available at https://github.com/smarini/tardis-phylogenetics

Contact: salemi@pathology.ufl.edu or brittany.rife@epi.ufl.edu

1 Introduction

Viral genetic sequence data can be used to trace viral evolutionary patterns, as well as spatiotemporal events and dynamics for viral and bacterial pathogens (Grenfell *et al.*, 2004). Tools such as NextStrain (Hadfield *et al.*, 2018) are now routinely utilized to monitor the phylodynamics of epidemics based on real-time deposition of pathogen sequences in databases (e.g. GenBank, HIVdatabases, GISAID) (Lednicky *et al.*, 2021; Mavian *et al.*, 2020a,b,c; Shu and McCauley, 2017; Wilkinson *et al.*, 2019). Not unlike traditional epidemiological analysis, however, these methods can be affected significantly by sampling bias (Hall *et al.*, 2016), and sampling during outbreaks are rarely performed randomly from a representative, stratified population (Rife *et al.*, 2017). Not only do the quality and quantity of sequences vary per country, but even regional sample collection policies tend to be inconsistent over time, as exemplified by the inherent sampling bias of SARS-CoV-2 strains, collected through convenience sampling and sequenced during the early pandemic phase (Mavian *et al.*, 2020a). Moreover, continuous generation of new sequences can very quickly approach information overload. For example, at the time of writing, more than a million sequences have been deposited in GISAID (SARS-CoV-2) database, with a number of countries either over or under represented compared of their actual infection prevalence. In such cases, full dataset analyses cannot be accomplished, as computational tools are not designed to handle hundreds of thousands of sequences. In order to reduce computational complexity, subsampling must often be performed (Hong *et al.*, 2020), typically using an approach that maximizes genetic diversity among subpopulations (Chernomor *et al.*, 2015), e.g. countries or regions (Hong *et al.*, 2020), which increases phylogenetic signal in the dataset, thus improving phylodynamic inference over convenience sampling. Besides enhancing signal for statistical phylogenetic inference, reliable estimates of significant events in the context of space and time also require sufficient temporal signal in the dataset (Hall *et al.*, 2016), or distribution of sampling over time, to calibrate reliable molecular clocks (Rambaut *et al.*, 2016). Despite the fact that sampling strategies pose a significant threat to conclusions drawn from phylodynamic inference, this problem has received so far insufficient attention (Frost *et al.*, 2015). Hall *et al.* (2016) were able to demonstrate that sampling sequences uniformly with respect to both space and time can allow for accurate reconstruction of the changing demographics of a dynamic, structured population. To date, however, there currently exists no tool to aid researchers to optimize subsampling with respect to space, time and genetic diversity. In what follows, we introduce TARDiS (Temporal And diveRsity Distribution Sampler), a machine learning approach designed to optimize phylogenetic subsampling according to both genetic diversity and temporal distribution for user-defined subpopulations.

2 Materials and methods

TARDiS implements a genetic algorithm (GA) (Falcón-Cardona and Coello, 2020; Kramer, 2017) optimizing genetic diversity and time sampling distributions criteria for any set of viral or bacterial genomes.

The output consists of user-defined number n of optimally subsampled genomes from a complete dataset of N genomes. Briefly, the algorithm is initialized as a population of random individuals. Each individual is a solution to the problem, i.e. a subsample of size n genomes. Each individual is characterized by a fitness score, reflecting how well that particular individual (solution) performs on the given problem. In our case, fitness is measured as a combination of genetic diversity (i.e. how diverse are the genomes represented by the individual), and time distribution (i.e., how evenly distributed are the genomes represented by the individual along the epidemic timeline).

In the event that whole-genome sequences are not available, we recommend a test for signal in the dataset (e.g., using likelihood mapping; Strimmer and von Haeseler, 1997) to ensure optimal TARDiS performance. In the absence of sufficient signal, or sufficient number of informative sites in the alignment, less variation will be observed for the calculation of genetic diversity across subsamples, rendering TARDiS more reliant on the temporal optimization function. While this scenario will ultimately provide a more uniform distribution of sampling times, downstream phylodynamic analysis may prove unreliable, as the phylogenetic tree reconstructed from the sequence data will not be able to provide significant support for individual relationships among sampled sequences. It is important to note that testing for phylogenetic signal is often recommended prior to analysis regardless of genomic fragmentation (Strimmer and von Haeseler, 1997).

2.1 Genetic algorithm principles

2.1.1 Genetic diversity maximization

We aim to recover a subsample of genomes as genetically diverse as possible. To do so, we first need to calculate the genetic distance between all possible genome pairs, represented by a square distance matrix D, with N rows and columns. Users can provide their own distance matrix or let TARDiS compute it using the Jukes–Cantor nucleotide substitution model. We calculate the genetic diversity fitness F_{gd} of a subset s with a total of n genomes as

$$F_{gd}(s) = \frac{\sum_{(i,j,i\neq j)}^{D_s} dist(i,j)}{dist_{max}(D,n)},$$

where s is the genome subset representing a single individual (solution); D_s is the diversity matrix for the n genomes included in s; i and j are a genome pair $\in s$, with $i \neq j$, and a total of $(f = (n^2 - n))/(2)$ pairs; $dist(i,\ j)$ is the genetic distance of the $(i,\ j)$ genome pair; and $dist_{max}(D,n)$ is the sum of the genetic distances of the top f elements of the distance matrix D, i.e. the sum of the maximum genetic diversities of the whole distance matrix D. It represents a theoretical upper bound to force genetic diversity fitness in the $[0,1]$ interval, with a higher value representing a better F_{gd}.

2.1.2 Time distribution optimization

Our objective is to recover a subsample of n genomes that are distributed as evenly along the considered time interval as possible. Intuitively, if $n = 100$ for an infected population that is increasing at an exponential rate over the course of 10 days (and every case was represented), we would consider one genome per day beginning with patient zero at day one. We can thus calculate the ideal time distribution I_{td} as a date vector of n elements, starting with the first available date d_f, ending with the last available date d_l, and having the remaining $n-2$ elements distanced with a $(d_f - d_l)/(n-1)$ interval. The worst possible time distribution W_{td}, on the other hand, is a time distribution concentrated into a single specific date (i.e., all samples collected on the same day). We measure the time distribution fitness F_{td} for a single individual (solution) as

$$F_{td}(s) = 1 - \frac{\sum_i^n |time(g_i) - t_i|}{\sum_i^n |t_w - t_i|},$$

where $time(g_i)$ is the collection date of the ith genome, with i being a genome included in s; t_i is ith date in I_{td}; and t_w is ith date in W_{td}. In other words, F_{td} is bound in the $[0,1]$ interval, with a higher value representing a better F_{td}.

The final fitness F of a specific individual s is calculated as

$$F(s) = F_{gd}(s) \times w_{gd}(s) + F_{td}(s) \times w_{td},$$

where w_{gd} and w_{td} are user-defined weights to set the importance of genetic diversity and time distribution, respectively.

While all values between 0 and 1 are available to the user for the w_{gd} and w_{gd} parameters, we have only considered scenarios wherein genetic diversity is considered in the presence ($w_{gd} = 0.5$, $w_{td} = 0.5$) or absence ($w_{gd} = 1$, $w_{td} = 0$) of time, with 0.5 being the default value in TARDiS, owing to the improved performance of the former over the latter in the described simulation. There are various circumstances where the alternative ($w_{gd} = 1$, $w_{td} = 1$) or time-weighted only ($w_{gd} = 0$, $w_{td} = 1$) parameter settings might be more relevant. For example, if more emphasis is placed on the evaluation of the changing level of diversity over the course of an epidemic, a more representative sampling for each month of the epidemic might be more appropriate. Alternatively, if sample collection is limited to the early phase of an epidemic, with very limited temporal spread, weighting according to sampling time would not necessarily improve significantly the dataset over purely diversity-based weighting. Deviation of parameter values from 0, 0.5 and 1 have not been assessed because the impact of this deviation is likely to depend on the dataset used. Hence, we would be unable to prescribe values outside of those reported herein that would be generalizable to all viral sequence data.

2.1.3 GA operators

Once a population is generated, fitness is calculated for each individual. Individuals are then chosen and combined to produce a new population in an iterative fashion. To generate a novel individual, TARDiS uses three operators: selection, mutation, and crossover. The selection operator is based on deterministic tournament selection with $k = 5$ (Falcón-Cardona and Coello, 2020). Briefly, two sets of k individuals are randomly chosen, and the individual with the highest fitness is selected from each set. The crossover operator combines two tournament winners a and b into a new individual c by keeping all the g genomes $\in (a \cup b)$- i.e., the shared genomes- and randomly selecting n–g genomes $\in (a \cap b) - (a \cup b)$- i.e., the genomes pertaining to a or b, but not both. To help avoid local maxima, each newly generated individual c has a 0.08 probability of mutating (Falcón-Cardona and Coello, 2020; Kramer, 2017). A mutation is defined as swapping a genome of individual c with one randomly chosen from the remaining genome pool- i.e., genomes $\notin (a, b)$. Note also that users define both a fraction of the population that is randomly created (and thus not evolved) for each generation, and a fraction of best genomes (ranked by fitness) to be copied without modifications in the next generation (elitism).

2.1.4 Group/spatial subsampling

Whereas subsampling over several groups of genomes can be performed independently, TARDiS is designed to take grouping factors into consideration (i.e., traits, such as geographical origin assigned to taxa in a.csv input file), resulting in optimal subsamples for each prespecified group, which is compiled into a single, ready-to-use alignment. In the following case study, we show how a geographical constraint can be added in the subsampling process.

2.2 Case study: subsampling a rising epidemic

We simulated a growing epidemic using a stochastic, agent-based model (Lequime *et al.*, 2020) with limited migration between 10 subpopulations, or regions (a, . . ., j).

2.2.1 Population dynamics

Each subpopulation was allowed to emerge from the initially infected population (a) with a mean probability of [initial] infection of 0.02 (standard deviation [sd] of 0.005). Each infected individual

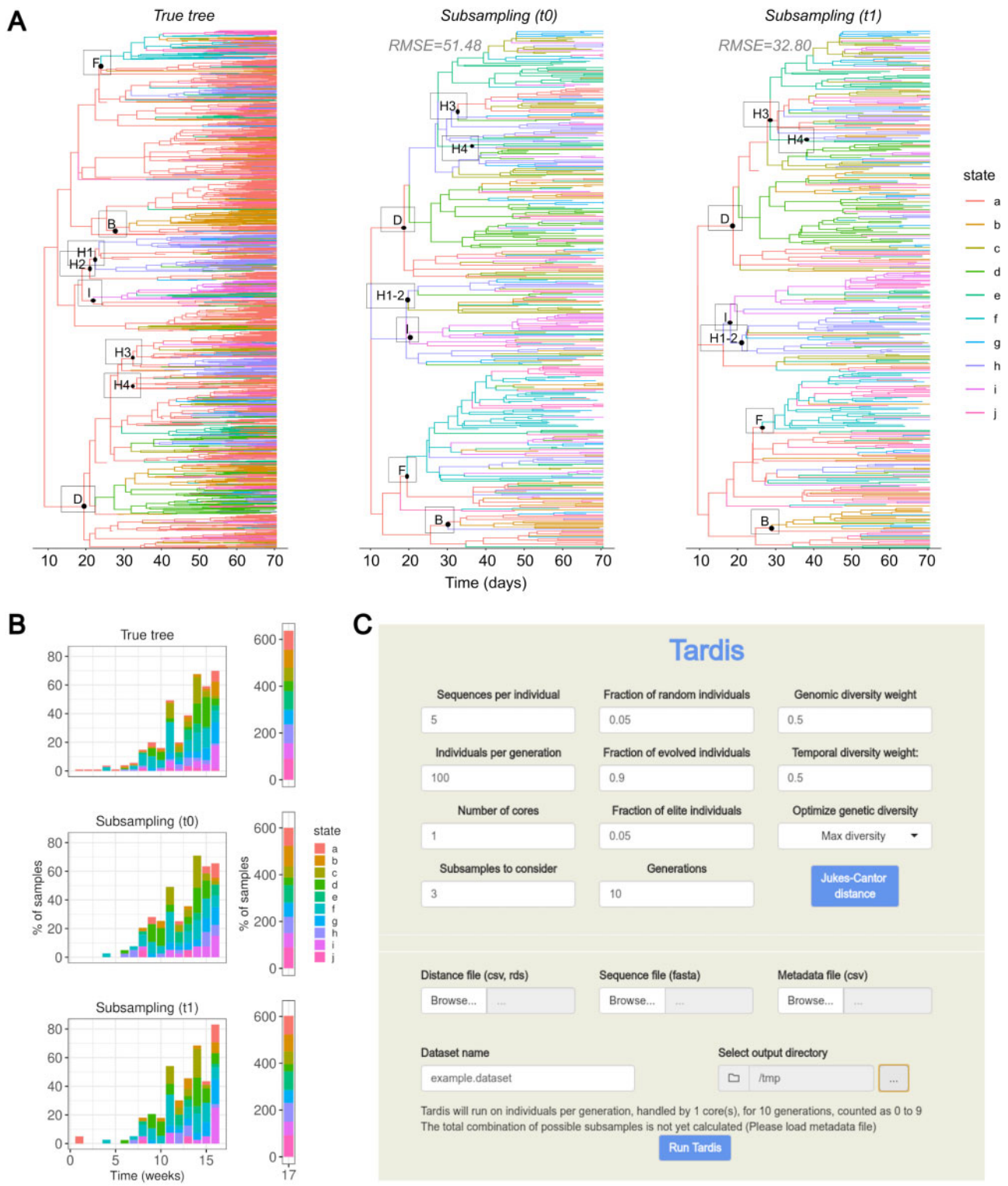

Fig. 1. (**A**) True and subsampled trees with representative clades. Eight representative clades were chosen for which the majority of taxa consisted of a single subpopulation, or state, and were consistent across true and subsampled trees. RMSE was calculated for the true TMRCAs and estimated TMRCAs across the five representative clades for the subsampled tree with (*t*1) and without consideration of time (*t*0). (**B**) Temporal distribution of samples per subpopulation (summing to 100%) for true and subsampled trees with (*t*1) and (*t*0) without consideration of time. (**C**) Screenshot of the TARDiS graphical user interface. Note, among the options, the number of individuals per each generation of the GA; the number of (best) solutions to output ('subsamples to consider'); the fractions of random, evolved, or elite individuals for each generation; and the options for genetic diversity optimization—besides the default maximum diversity, users can select to optimize toward the mean or median of the dataset

within a subpopulation was then allowed to migrate to another subpopulation with a mean probability of 0.01 (sd = 0.005). The number of contacts for each individual was picked from a normal distribution with a mean of 4 (sd = 2). The probability of transmission (when a contact occurs) was provided in the form of a threshold function: prior to 5 days (sd = 3), the host was not able to transmit, but after that time, the individual was able transmit with a mean probability of 0.05 (sd = 0.005), representing an incubation

period for the simulated virus. Each infected individual was removed from the simulation (representing death, recovery, etc.) after 14 days. The described parameters resulted in a basic reproductive number (R_0) of ~1.6 for the epidemic. The simulated epidemic was run for 365 days or until a total of 10,000 hosts were infected. For each of the 10 subpopulations, individuals belonging to that subpopulation were binned according to week of removal (i.e. 7-day intervals) and subsampled according to an exponential distribution (rate = 5), representing idealistic sampling of a population proportional to the size of the epidemic and resulting in a range [37, 844] of sampled individuals for each subpopulation (state). The original transmission tree was pruned, leaving only the remaining sampled individuals.

A molecular clock, or constant evolutionary rate across all branches of the tree, was assumed, allowing branches separating nodes within the tree to be scaled in both time and genetic distance. Nucleotide (A, C, G, T) sequences were thus simulated along the tree using a general time reversible evolutionary model, allowing for unequal nucleotide frequencies and differing rates of substitution for each pair of nucleotides, such as is commonly observed for RNA viral genomes edited by APOBEC-mediated deamination (represented as C–>T) (Mourier *et al.*, 2021). The rate matrix was defined as (0.32512, 1.07402, 0.26711, 0.25277, 2.89976, 1.00000) and nucleotide frequencies (0.299, 0.183, 0.196, 0.322). A gamma distribution of rate variation across sites (alpha = 2.35) was also used, with a proportion (0.60) of sites considered to be invariable to accommodate for evolutionary variation, such as across codon positions in a coding sequence. Branch lengths were scaled by a factor of 8E−04 (approximate evolutionary rate in substitutions/site/year), representing that of RNA viruses such as beta coronaviruses (Nakagawa and Miyazawa, 2020).

Migration rates for each of the 10 subpopulations were calculated as a function of the number of transitions between subpopulation states (non-reversible) along each branch within the tree and the frequency (F) of the initial subpopulation among tree tips- i.e., for w branches with transitions between subpopulations- and x branches with specifically transitions from i (node at earlier time point) to j (node at more recent time point)

$$R_{ij} = \frac{x \times F_i}{w}.$$

We ran TARDiS on a single simulated dataset, subsampling 40 genomes per region (with the exception of region f, with 37 genomes available) for 50 generations, with a population of 1000 individuals per generation, of which 85% were evolved, 10% were newly generated, and 5% were elite. The phytools package (Revell, 2012) in R (R Core Team, 2020) was used for joint likelihood reconstruction of discrete ancestral states (Pupko *et al.*, 2000) according to subpopulation for each internal node of the subsampled trees. Transition rates among discrete states along tree nodes were considered to be equal *a priori*. Migration rates between states were then re-estimated, as previously described. We compared the results obtained both with ($w_{gd} = 1$, $w_{td} = 1$) and without considering time distribution ($w_{gd} = 1$, $w_{td} = 0$). Our simulation indicated that better results are obtained by considering time distribution: the overall migration rate root mean squared error (RMSE) decreased by 17% (0.035–0.029). Eight representative clades were then chosen for which the majority of taxa consisted of a single subpopulation and were consistent across true and subsampled trees (Fig. 1A). The RMSE for the estimated time of the most recent common ancestor (TMRCA), representing the upper limit of the timing of introduction into that particular region, decreased by 43.4% [from 25.37 considering only genetic diversity ($t0$) to 14.36 if we include time distribution ($t1$)]. The addition of a temporal weighting component for an exponentially growing population can act to both increase and decrease representation of earlier time points (e.g. weeks 15 and 12, respectively; Fig. 1B). However, representation of week 1 of the epidemic was increased from 0% to 5%. As the early stages of an epidemic, and time nearing the root of the tree, represent periods of high epidemiological and phylogenetic uncertainty, sample representation during this time is critical for reliable phylodynamic inference and thus contributed to the loss of error in our estimates.

3 Implementation

TARDiS is implemented as a command-line tool based on NextFlow, suitable for analyzing large datasets in a high-performance computing environment, and as a graphical user interface based on R/Shiny for ease-of-use and experimentation (Fig. 1). Along with example datasets and a user manual, TARDiS is available at https://github.com/smarini/tardis-phylogenetics

Funding

This work was supported in part by the National Institutes of Health (NIH) R01 AI145552 (M.S. and M.P.), NIH R21 AI138815 (M.P. and M.S.), National Science Foundation (NSF) RAPID DEB2028221 (M.P. and M.S.), NSF RAPID DMS2028728 (M.S.), and the Stephany W. Holloway University Chair in AIDS Research.

Conflict of Interest: none declared.

References

Chernomor,O. *et al.* (2015) Split diversity in constrained conservation prioritization using integer linear programming. *Methods Ecol. Evol.*, **6**, 83–91.

Falcón-Cardona,J.G. and Coello,C.A.C. (2020) Indicator-based multi-objective evolutionary algorithms: a comprehensive survey. *ACM Comput. Surv.*, **53**, 1–35.

Frost,S.D. *et al.* (2015) Eight challenges in phylodynamic inference. *Epidemics*, **10**, 88–92.

Grenfell,B.T. *et al.* (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, **303**, 327–332.

Hadfield,J. *et al.* (2018) Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, **34**, 4121–4123.

Hall,M.D. *et al.* (2016) The effects of sampling strategy on the quality of reconstruction of viral population dynamics using bayesian skyline family coalescent methods: a simulation study. *Virus Evol.*, **2**, vew003.

Hong,S.L. *et al.* (2020) In search of covariates of HIV-1 subtype b spread in the united states—a cautionary tale of large-scale bayesian phylogeography. *Viruses*, **12**, 182.

Kramer,O. (2017) *Genetic Algorithm Essentials*, Vol. **679**. Springer.

Lednicky,J. *et al.* (2021) Earliest detection to date of SARS-CoV-2 in Florida: identification together with influenza virus on the main entry door of a university building, february 2020. *PLoS One*, **16**, e0245352.

Lequime,S. *et al.* (2020) nosoi: a stochastic agent-based transmission chain simulation framework in R. *Methods Ecol. Evol.*, **11**, 1002–1007.

Mavian,C. *et al.* (2020a) Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-CoV-2 infections unreliable. *Proc. Natl. Acad. Sci. USA*, **117**, 12522–12523.

Mavian,C. *et al.* (2020b) A snapshot of SARS-CoV-2 genome availability up to April 2020 and its implications: data analysis. *JMIR Public Health Surveill.*, **6**, e19170.

Mavian,C. *et al.* (2020c) Toxigenic vibrio cholerae evolution and establishment of reservoirs in aquatic ecosystems. *Proc. Natl. Acad. Sci. USA*, **117**, 7897–7904.

Mourier,T. *et al.* (2021) Host-directed editing of the SARS-CoV-2 genome. *Biochem. Biophys. Res. Commun.*, **538**, 35–39.

Nakagawa,S. and Miyazawa,T. (2020) Genome evolution of SARS-CoV-2 and its virological characteristics. *Inflamm. Regen.*, **40**, 17.

Pupko,T. *et al.* (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.*, **17**, 890–896.

Rambaut,A. *et al.* (2016) Exploring the temporal structure of heterochronous sequences using tempest (formerly path-o-gen). *Virus Evol.*, **2**, vew007.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Revell,L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, **3**, 217–223.
Rife,B.D. *et al.* (2017) Phylodynamic applications in 21 st century global infectious disease research. *Global Health Res. Policy*, **2**, 13.
Shu,Y. and McCauley,J. (2017) Gisaid: global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance*, **22**, 30494.
Strimmer,K. and von Haeseler,A. (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl. Acad. Sci. USA*, **94**, 6815–6819.
Wilkinson,E. *et al.* (2019) The effect of interventions on the transmission and spread of HIV in South Africa: a phylodynamic analysis. *Sci. Rep.*, **9**, 1–12.