bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

COVFlow: virus phylodynamics analyses from selected SARS-CoV-2 sequences

Gonché Danesh[1,*], Corentin Boennec[1], Laura Verdurme[2], Mathilde Roussel[2], Sabine Trombert-Paolantoni[2], Benoit Visseaux[2], Stéphanie Haim-Boukobza[2], Samuel Alizon[1,3]

[1] MIVEGEC, CNRS, IRD, Université de Montpellier

[2] Laboratoire CERBA, Saint Ouen L'Aumône, France

[3] Center for Interdisciplinary Research in Biology (CIRB), College de France, CNRS, INSERM, Université PSL, Paris, France

[*] corresponding author: gonche.danesh@ird.fr

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Phylodynamic analyses generate important and timely data to optimise public health response to SARS-CoV-2 outbreaks and epidemics. However, their implementation is hampered by the massive amount of sequence data and the difficulty to parameterise dedicated software packages. We introduce the COVFlow pipeline, accessible at https://gitlab.in2p3.fr/ete/CoV-flow, which allows a user to select sequences from the Global Initiative on Sharing Avian Influenza Data (GISAID) database according to user-specified criteria, to perform basic phylogenetic analyses, and to produce an XML file to be run in the Beast2 software package. We illustrate the potential of this tool by studying two sets of sequences from the Delta variant in two French regions. This pipeline can facilitate the use of virus sequence data at the local level, for instance, to track the dynamics of a particular lineage or variant in a region of interest.

Keywords: COVID-19, molecular epidemiology, sequence database, phylogenetics, public health

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 Introduction

SARS-CoV-2 full genome sequences were made available since 2020 through the database created by the Global Initiative on Sharing Avian Influenza Data (GISAID) [6, 15]. This allowed the timely monitoring of variants of concerns (VoC) with platforms such as CoVariants (`CoVariants`), outbreak.info [16], or CoV-Spectrum [4], and the realisation of phylogenetic analyses, e.g. via NextStrain [10].

Phylogenies represent a powerful means to analyse epidemics because there is an intuitive parallel between a transmission chain and a time-scaled phylogeny of infections, which is the essence of the field known as 'phylodynamics' [7]. As illustrated in the case of the COVID-19 pandemic, state-of-the-art analyses allow one to investigate the spatio-temporal spread of an epidemic [5], superspreading events [2], and even detect differences in transmission rates between variants [27].

Phylodynamic analyses involve several technical steps to go from dates virus sequence data to epidemiological parameter estimates, which can make them difficult to access to a large audience. Furthermore, the amount of data shared greatly overcomes the capacities of most software packages and imposes additional selection steps that further decrease the accessibility of these approaches. To address these limitations, we introduce the COVflow pipeline which covers all the steps from filtering the sequence data according to criteria of interest (e.g. sampling data, sampling location, virus lineage, or sequence quality) to generating a time-scaled phylogeny and an XML configuration file for a BDSKY model [25] to be run in the Beast2 software package [3].

Some pipelines already exist to assess sequence quality, infer an alignment, and infer a time-scaled phylogeny such as NextClade [1] and Augur [12]. However, these do not include a data filtration step based on metadata characteristics. Furthermore, performing a phylodynamic analysis from the output files they generate requires dedicated skills. The COVFlow pipeline addresses these two limitations and integrates all the steps present in separate software packages to go from the raw sequence data and metadata to the XML to be run in Beast2.

In this manuscript, we present the architecture of the pipeline and apply it to data from the French epidemic. Focusing on sequences belonging to the Delta variant collected in France in two regions, Ile-de-France, and Provence-Alpes-Cote-d'Azur, by a specific French laboratory (CERBA), we illustrate the accessibility, flexibility, and public health relevance of the COVFlow pipeline.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

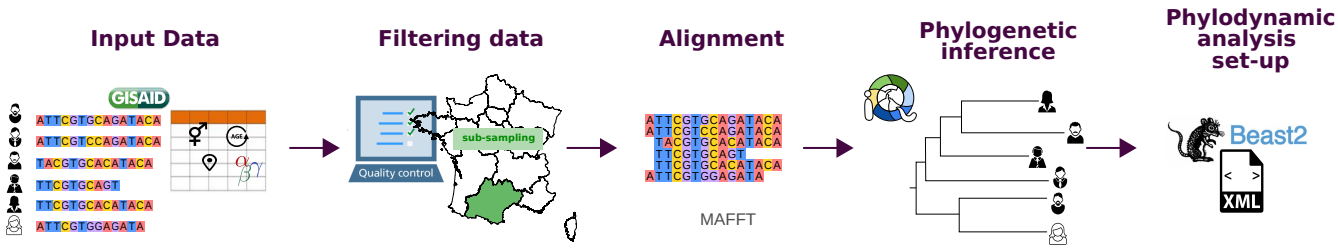


Figure 1: **Structure of the COVFlow pipeline.** The input data correspond to FASTA sequences and metadata provided by the GISAID. The data filtering is done using a YAML configuration file. The sequence alignment is performed with MAFFT and the phylogenetic inference with IQ-TREE. The pipeline generates an XML file that can be directly used with Beast2.

2 Methods

COVFlow is a bioinformatics pipeline for phylogenetic and phylodynamic analysis of SARS-CoV-2 genome sequences. It is based on the Snakemake workflow management system [20] and its dependencies are easily installed via a conda virtual environment. Snakemake ensures reproducibility, while Conda (https://docs.conda.io/en/latest/) and Bioconda [8] allow for version control of the programs used in the pipeline. Overall, the pipeline is easy to install and avoids dependency conflicts.

Pipeline configuration

The pipeline workflow is configured using a configuration file, in a YAML format. The configuration file must contain the path to the sequence data file, the path to the metadata file, and the prefix chosen for the output files. Each parameter of the pipeline following steps has a default value, which can be modified by the user in the YAML configuration file.

2.1 Input data

The input data analysed by COVFlow are sequence data and metadata, corresponding to patient properties, that can be downloaded from the Global Initiative on Sharing All Influenza Data (GISAID, https://www.gisaid.org/). The sequence data are in a FASTA format file. The metadata downloaded contains details regarding the patient's sequence ID (column named 'strain'), the sampling dates (column 'date'), the region, country and division where the sampling has been made (columns respectively named 'region', 'country', and 'division'). It also lists the virus lineage assigned by the Pangolin tool [21], and the age and sex of the patient (columns respectively named 'pango_lineage', 'age', and

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

'sex').

Data filtering

The first step implemented in the pipeline performs quality filtering. By default, genomic sequences that are shorter than $27,000$ bp, or that have more than $3,000$ missing data (i.e. N bases) and more than 15 non-ATGCN bases are excluded. These parameter values can be modified by the user. Sequences belonging to non-human or unknown hosts are also excluded. Sequences for which the sampling date is more recent than the submission date, or for which the sampling date is unclear (e.g. missing day) are also excluded. Finally, duplicated sequences and sequences that are flagged by the Nextclade tool [1] with an overall bad quality (Nextclade QC overall status 'bad' or 'mediocre') are also removed.

The sequence data is then further filtered following the user's criteria. These include Pangolin lineages, sampling locations (regions, countries, or divisions), and sampling dates. In addition to specifying the maximum and/or minimum sampling dates, the user can specify a sub-sampling scheme of the data with a number or percentage of the data per location and/or per month. For example, the user can decide to keep $x\%$ of the data per country per division per month or to keep y sequence data per division. Finally, more specific constraints can be given using a JSON format file with three possible actions: i) keep only rows (i.e. sequences) that match or contain a certain value, ii) remove rows that match or contain a certain value, and iii) replace the value of a column by another value for specific rows with a column that matches or contains a certain value. The last action can be used to correct the metadata, for instance, if the division field is not filled in but can be inferred from the names of the submitting laboratory. The JSON file can be composed of multiple key-value pairs, each belonging to one of the three actions. For example, the user can specify to keep only male patients and to remove data from one particular division while setting the division of all the samples submitted by a public hospital from the Paris area (i.e. the APHP) to the value 'Ile-de-France'.

Aligning and masking

The set of sequences resulting from the data filtering is then divided into temporary FASTA files with a maximum number of 200 sequences per file. For each subset, sequences are aligned to the

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reference genome MN908947.3 using MAFFT v7.305 [14] with the 'keeplength' and 'addfragments' options. All the aligned sequences are then aggregated into a single file. Following earlier studies, the first 55 and last 100 sites of the alignment are then masked to improve phylogenetic inference (http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473).

Inferring and time-scaling a phylogeny

A maximum-likelihood phylogenetic tree is estimated using IQ-TREE v2.1.2 [19] under a GTR substitution model from the alignment. The resulting phylogeny is time-scaled using TreeTime v0.8.1 [23]. By default, the tree is rooted using two ancestral sequences (Genbank accession numbers MN908947.3 and MT019529.1) as an outgroup, which is then removed, with a fixed clock rate of $8 \cdot 10^{-4}$ substitutions per position per year [22] and a standard deviation of the given clock rate of 0.0004. These parameters can be modified by the user. The output phylogeny is in a Newick format file.

BDSKY XML file generating

The Bayesian birth-death skyline plot (or BDSKY) method allows the inference of the effective reproductive number from genetic data or directly from a phylogenetic tree, by estimating transmission, recovery, and sampling rates [25]. This method allows these parameters to vary through time and is implemented within the BEAST2 software framework [3].

Performing a BDSKY analysis requires setting an XML file specifying the parameters for the priors. As in any Bayesian analysis, this step is extremely important. The default settings in BEAST2 have been chosen to minimise the risk of errors. COVFlow builds on most of these with some modifications to fit the needs of large SARS-CoV-2 phylogenies.

The most important change has to do with the inference of the phylogeny. This can be done by BEAST2 but to minimise computation speed and allow for the analysis of large phylogenies, the pipeline sets the time-scaled phylogeny from the previous step in the XML file.

The default XML file assumes that there are two varying effective reproductive numbers to estimate, with a lognormal prior distribution, $\text{LogNorm}(M = 0, S = 1)$, and a starting value of 1. This prior is adapted to such virus epidemics and, as we will see below, can be edited if needed. The default prior

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for the rate of end of the infectious period is a uniform distribution, Uniform$(10, 300)$, with a starting value of 70 years^{-1}, and is assumed to be constant over time. This yields infectious periods varying from 1.2 to 36.5 days, which is relevant with the biology of SARS-CoV-2 infections [29]. Usually, little or no sampling effort is made before the first sample was collected. Therefore, by default, we assume two sampling proportions: before the first sampling date it is set to zero, and after the default prior is a beta distribution, Beta$(\alpha = 1, \beta = 1)$, with a starting value of 0.01. The non-zero sampling proportion is assumed to remain constant during the time the samples were collected. The method can also estimate the date of origin of the index case, which corresponds to the total duration of the epidemic. The default prior for this parameter prior is a uniform distribution Uniform$(0, 2)$ years.

In the COVFlow configuration file, the user can modify the distribution shapes, the starting values, the upper and lower values, and the dimensions for each of these parameters to estimate. The length of the MCMC chain and the sampling frequency, which are by default set to $10,000,000$ and $100,000$ respectively, can also be modified.

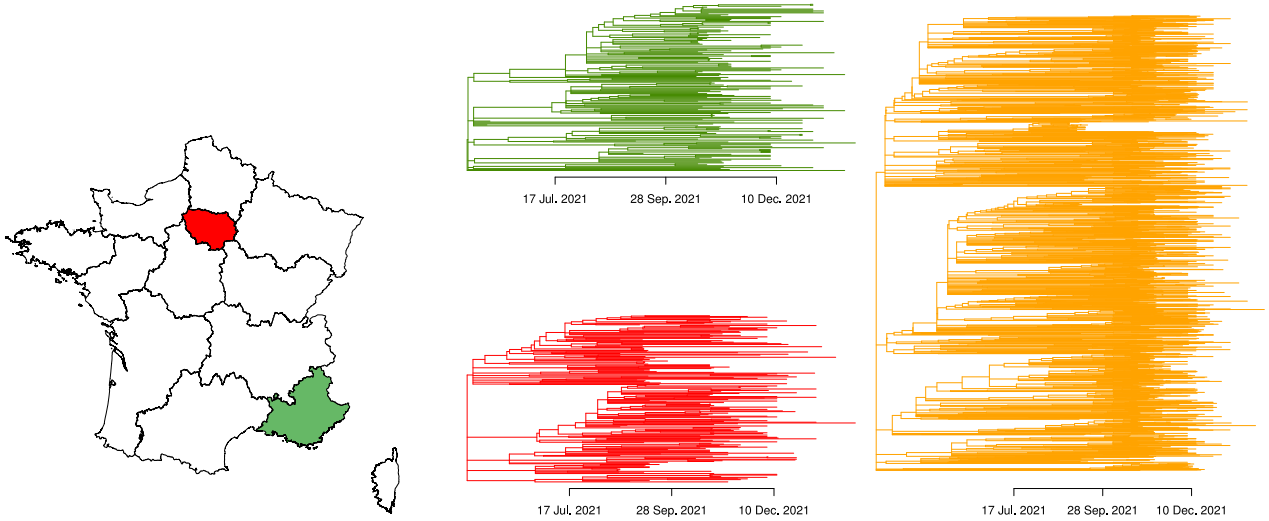
The BEAST2 inference itself is not included in the pipeline. The reason for this is that a preliminary step (i.e. installing the BDSKY package) needs to be performed by the user. Similarly, the analysis of the BEAST2 output log files needs to be performed by the user via Tracer or a dedicated R script.

3 Results

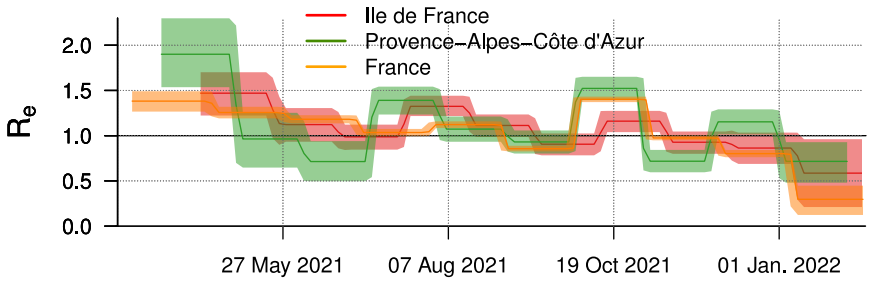
We illustrate the potential of the COVFlow pipeline by performing a phylodynamic analysis of a specific COVID-19 lineage, here the Delta variant (Pango lineage B.1.617.2), in two regions of a country, here Ile-de-France and Provence-Alpes-Côte d'Azur in France (Figure 2(a)).

We downloaded sequence data and metadata from the GISAID platform for the GK clade corresponding to the lineage B.1.617.2 available on the April 22, 2022, which amounted to $4,212,049$ sequences. Using the pipeline and the editing of the its JSON file, we cleaned the sequence data, selected the data collected by a specific large French laboratory (CERBA), selected the data from two regions of interest (the Ile-de-France region for the first analysis and the Provence-Alpes-Côte d'Azur region for the second analysis), and sub-sampled the data to keep up to 50 sequences per month. For the third analysis, which included the whole country, we sub-sampled the data to keep up to 50

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



(a) (b)



(c)

Figure 2: **Analysing the SARS-CoV-2 Delta variant epidemics in French regions using the COVFlow pipeline.** a) Geographical sub-sampling using at most 50 sequences per month for the Delta variant in Ile-de-France (IdF, in red), Provence-Alpes-Côte d'Azur (PACA, in green), and in all of France collected by CERBA laboratory. b) Time-scaled phylogenies generated using sub-sampled data from IdF (in red), PACA (green), and all of France (in orange). c) Temporal variations of the effective reproductive number (R_e) of the Delta variant in IdF (red), in PACA (green), and France (orange) estimated using Beast2. The last panel was generated using Beast2. In panel c, the lined show the median values and the shaded area the 95% highest posterior density.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequences per month per French region. This resulted in the selection of 176 SARS-CoV-2 genomes for Ile-de-France (IdF), 221 genomes for Provence-Alpes-Côte d'Azur (PACA), and $1,575$ genomes for France. In this example, the two regions were chosen because they had some of the highest coverage in the dataset, while being in different parts of France. The other parameters of the pipeline were default except that for the number of windows for the effective reproductive numbers in the BDSKY analysis which was set to 10.

The first output of the pipeline is the time-scaled phylogeny inferred from the sequences. In Figure 2(b), we show the one for each of the two French regions considered and the one for France. This already allow us to visualise the origin of the epidemic associated with the sequences sampled. Furthermore, the shape of the phylogeny can reflect the number of introductions in the locality studied.

The second output of the pipeline is the XML file for a BDSKY model that can be run into Beast2. In Figure 2(c), we show the temporal variations in the effective reproduction number (R_e), that is the average number of secondary infections caused by an infected individual at a given date. If $R_e < 1$, the epidemic is decreasing and if $R_e > 1$ it is growing. The results show that the importation of the Delta variant epidemic seems to occurred earlier and more frequently in PACA than in IdF in early 2021. Furthermore, in early July 2021, we see that the epidemic wave started in PACA before IdF. This is consistent with the beginning of the school holidays and PACA being a densely populated region in the summer. During the summer 2021, the epidemic growth in these two regions, IdF and PACA, was more important than the French average. In the fall 2021, the epidemic growth was again stronger and earlier in PACA than in IdF. Furthermore, given the superposition of the French and PACA R_e curves, it is possible that PACA drove the national epidemic at the time. Finally, we see that contrarily to IdF or France, PACA experience a period of Delta variant growth at the end of 2021. This final result is consistent with large French surveillance data obtained through variant-specific PCR tests that show a high proportion of infections with the S:L452R mutation in PACA compared to the other French regions during the Omicron wave [24].

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4 Discussion

The COVID-19 pandemic constitutes a qualitative shift in terms of generation, sharing, and analysis of virus genomic sequence data. The GISAID initiative allowed the rapid sharing of SARS-CoV-2 sequence data, which is instrumental for local, national, and international public health structures that need to provide timely reports on the sanitary situation. At a more fundamental level, this genomic data is also key to furthering our understanding of the spread and evolution of the COVID-19 pandemic [18], especially in low-resource countries [28].

We elaborate the COV-flow pipeline, which allows users to perform all the steps from raw sequence data to phylodynamics analyses. In particular, it can select sequences from the GISAID datased based on metadata, perform a quality check, align the sequences, infer a phylogeny, root this phylogeny into time, and generate an XML file for Beast2 analysis (we also provide scripts to analyse the outputs). Furthermore, COV-flow can also readily allow the implementation of subsampling schemes per location and per date. This can help balance the dataset and also be extremely useful to perform sensitivity analyses and explore the robustness of the phylodynamic results.

A future extension will consist in including other Beast2 population dynamics models, for instance, the Bayesian Skyline model, which is not informative about R_0 but is less sensitive to variations in sampling intensity as it assumes sampling is negligible. Another extension will be to use other databases to import SARS-CoV-2 genome data, e.g. that published by NCBI, via LAPIS (Lightweight API for Sequences).

Beast2 can simultaneously infer population dynamics parameters and phylogenies, which is an accurate way to factor in phylogenetic uncertainty [3]. However, this global inference is particularly computationally heavy and is out of reach for large data sets. To circumvent this problem, we perform the phylogenetic inference first using less accurate software packages and then impose the resulting phylogeny into the Beast2 XML file. An extension of the pipeline could offer the user to also perform the phylogenetic inference, for instance by using the so-called 'Thorney Beast' (`https://beast.community/thorney_beast`) implemented in Beast 1.10 [26].

Finally, it is important to stress that phylogenetic analyses are always dependent on the sampling scheme [9, 11, 13, 17]. If most of the sequences come from contact tracing in dense clusters, the analysis

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

will tend to overestimate epidemic spread. This potential bias can be amplified by the sequence selection feature introduced in the pipeline. An advantage of COVFlow is that it can perform spatio-temporal subsampling but additional studies are needed to identify which are the most appropriate subsampling schemes to implement.

Acknowledgement

The authors acknowledge further support from the CNRS, the IRD and the itrop HPC (South Green Platform) at IRD montpellier, which provided HPC resources that contributed to the results reported here (https://bioinfo.ird.fr/).

The authors thank the Experimental and Theoretical Evolution team from Maladies Infectieuses et Vecteurs: Écologie, Génétique, Évolution et Contrôle, University of Montpellier, for discussion, as well as the EMERGEN consortium (complete member list in Supplementary Materials).

This project was supported by the Agence Nationale de la Recherche Maladies Infectieuses Émergentes to the MODVAR project (grant no. ANRS0151).

Authors contributions

GD and SA conceived the study, GD built the pipeline and performed the analyses, CB contributed to the implementation of the pipeline, LV, MR, STP, BV, and SHB contributed genetic sequence data, SA and GD wrote a first version of the manuscript.

Data and scripts

The sequences analysed were generated by CERBA and uploaded to GISAID.

The R scripts, along with all the files generated by the pipeline and used for the analyses (XML files, FASTA alignments, time-scaled phylogenies) are provided in Supplementary Materials.

The pipeline itself can be accessed on the Git public repository https://gitlab.in2p3.fr/ete/CoV-flow

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] Ivan Aksamentov, Cornelius Roemer, Emma B. Hodcroft, and Richard A. Neher. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software*, 6(67):3773, 2021.

[2] Samuel Alizon. Superspreading genomes. *Science*, 371(6529):574–575, February 2021. Publisher: American Association for the Advancement of Science Section: Perspective.

[3] Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, et al. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014.

[4] Chaoran Chen, Sarah Nadeau, Michael Yared, Philippe Voinov, Ning Xie, Cornelius Roemer, and Tanja Stadler. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics*, 38(6):1735–1737, 2022.

[5] Louis du Plessis, John T. McCrone, Alexander E. Zarebski, Verity Hill, Christopher Ruis, Bernardo Gutierrez, Jayna Raghwani, Jordan Ashworth, Rachel Colquhoun, Thomas R. Connor, Nuno R. Faria, Ben Jackson, Nicholas J. Loman, Áine O'Toole, Samuel M. Nicholls, Kris V. Parag, Emily Scher, Tetyana I. Vasylyeva, Erik M. Volz, Alexander Watts, Isaac I. Bogoch, Kamran Khan, the COVID-19 Genomics UK (COG-UK) Consortium, David M. Aanensen, Moritz U. G. Kraemer, Andrew Rambaut, and Oliver G. Pybus. Establishment & lineage dynamics of the SARS-CoV-2 epidemic in the UK. preprint, Epidemiology, October 2020.

[6] Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(1):33–46, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gch2.1018.

[7] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–32, 2004.

[8] Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A Chapman, Jillian Rowe, Christopher H

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tomkins-Tinch, Renan Valieris, and Johannes Köster. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, 15(7):475–476, 2018.

[9] Stéphane Guindon and Nicola De Maio. Accounting for spatial sampling patterns in Bayesian phylogeography. *Proceedings of the National Academy of Sciences*, 118(52):e2105273118, 2021.

[10] James Hadfield, Colin Megill, Sidney M. Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A. Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, December 2018. Publisher: Oxford Academic.

[11] Matthew D. Hall, Mark E. J. Woolhouse, and Andrew Rambaut. The effects of sampling strategy on the quality of reconstruction of viral population dynamics using Bayesian skyline family coalescent methods: A simulation study. *Virus Evolution*, 2(1):vew003, 2016.

[12] John Huddleston, James Hadfield, Thomas R. Sibley, Jover Lee, Kairsten Fay, Misja Ilcisin, Elias Harkins, Trevor Bedford, Richard A. Neher, and Emma B. Hodcroft. Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. *Journal of Open Source Software*, 6(57):2906, 2021.

[13] Michael D. Karcher, Luiz Max Carvalho, Marc A. Suchard, Gytis Dudas, and Vladimir N. Minin. Estimating effective population size changes from preferentially sampled genetic sequences. *PLOS Computational Biology*, 16(10):e1007774, 2020.

[14] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 01 2013.

[15] Shruti Khare, Céline Gurry, Lucas Freitas, Mark B. Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, Joses Ho, Raphael TC Lee, Winston Yeo, GISAID Core Curation Team, and Sebastian Maurer-Stroh. GISAID's Role in Pandemic Response. *China CDC Weekly*, 3(49):1049–1051, 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[16] Alaa Abdel Latif, Julia L. Mullen, Manar Alkuzweny, Ginger Tsueng, Marco Cano, Emily Haag, Jerry Zhou, Mark Zeller, Nate Matteson, Chunlei Wu, Kristian G. Andersen, Andrew I. Su, Karthik Gangavarapu, Laura D. Hughes, and the Center for Viral Systems Biology. outbreak.info: Lineage comparison, 2021.

[17] Stilianos Louca, Angela McLaughlin, Ailene MacPherson, Jeffrey B Joy, and Matthew W Pennell. Fundamental Identifiability Limits in Molecular Epidemiology. *Molecular Biology and Evolution*, 38(9):4010–4024, 2021.

[18] Michael A. Martin, David VanInsberghe, and Katia Koelle. Insights from SARS-CoV-2 sequences. *Science*, 371(6528):466–467, 2021.

[19] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, 02 2020.

[20] F Mölder, KP Jablonski, B Letcher, MB Hall, CH Tomkins-Tinch, V Sochat, J Forster, S Lee, SO Twardziok, A Kanitz, A Wilm, M Holtgrewe, S Rahmann, S Nahnsen, and J Köster. Sustainable data analysis with snakemake [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Research*, 10(33), 2021.

[21] Áine O'Toole, Emily Scher, Anthony Underwood, Ben Jackson, Verity Hill, John T McCrone, Rachel Colquhoun, Chris Ruis, Khalil Abu-Dahab, Ben Taylor, Corin Yeats, Louis du Plessis, Daniel Maloney, Nathan Medd, Stephen W Attwood, David M Aanensen, Edward C Holmes, Oliver G Pybus, and Andrew Rambaut. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution*, 7(2), 07 2021. veab064.

[22] Andrew Rambaut. Phylodynamic Analysis | 176 genomes | 6 Mar 2020, January 2020. Library Catalog: virological.org.

[23] Pavel Sagulenko, Vadim Puller, and Richard A Neher. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, 4(1), 01 2018. vex042.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.17.496544; this version posted February 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[24] Mircea T. Sofonea, Bénédicte Roquebert, Vincent Foulongne, David Morquin, Laura Verdurme, Sabine Trombert-Paolantoni, Mathilde Roussel, Jean-Christophe Bonetti, Judith Zerah, Stéphanie Haim-Boukobza, and Samuel Alizon. Analyzing and Modeling the Spread of SARS-CoV-2 Omicron Lineages BA.1 and BA.2, France, September 2021–February 2022. *Emerging Infectious Diseases*, 28(7), 2022.

[25] Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci USA*, 110(1):228–33, 2013.

[26] Marc A. Suchard, Philippe Lemey, Guy Baele, Daniel L. Ayres, Alexei J. Drummond, and Andrew Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1), 2018.

[27] Erik Volz, Verity Hill, John T. McCrone, Anna Price, David Jorgensen, Áine O'Toole, Joel Southgate, Robert Johnson, Ben Jackson, Fabricia F. Nascimento, Sara M. Rey, Samuel M. Nicholls, Rachel M. Colquhoun, Ana da Silva Filipe, James Shepherd, David J. Pascall, Rajiv Shah, Natasha Jesudason, Kathy Li, Ruth Jarrett, Nicole Pacchiarini, Matthew Bull, Lily Geidelberg, Igor Siveroni, Ian Goodfellow, Nicholas J. Loman, Oliver G. Pybus, David L. Robertson, Emma C. Thomson, Andrew Rambaut, and Thomas R. Connor. Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. *Cell*, 0(0), November 2020. Publisher: Elsevier.

[28] Eduan Wilkinson, Marta Giovanetti, Houriiyah Tegally, James E. San, Richard Lessells, and *et al.* A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science*, 374(6566):423–431, 2021.

[29] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, Lulu Guan, Yuan Wei, Hui Li, Xudong Wu, Jiuyang Xu, Shengjin Tu, Yi Zhang, Hua Chen, and Bin Cao. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, March 2020.