Published online 11 August 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 3 **1**
https://doi.org/10.1093/nargab/lqab075

Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows

Frédéric Lemoine [1,2,*] **and Olivier Gascuel** [1,3]

[1]Unité de Bioinformatique Évolutive, Département de Biologie Computationnelle, Institut Pasteur, Paris, France, [2]Hub de Bioinformatique et Biostatistique, Département de Biologie Computationnelle, Institut Pasteur, Paris, France and [3]Institut de Systématique, Evolution, Biodiversité, ISYEB, UMR 7205, Muséum National d'Histoire Naturelle, CNRS, SU, EPHE, UA, Paris, France

Received May 14, 2021; Revised July 09, 2021; Editorial Decision August 02, 2021; Accepted August 03, 2021

ABSTRACT

Phylogenetics is nowadays at the center of numerous studies in many fields, ranging from comparative genomics to molecular epidemiology. However, phylogenetic analysis workflows are usually complex and difficult to implement, as they are often composed of many small, reccuring, but important data manipulations steps. Among these, we can find file reformatting, sequence renaming, tree re-rooting, tree comparison, bootstrap support computation, etc. These are often performed by custom scripts or by several heterogeneous tools, which may be error prone, uneasy to maintain and produce results that are challenging to reproduce. For all these reasons, the development and reuse of phylogenetic workflows is often a complex task. We identified many operations that are part of most phylogenetic analyses, and implemented them in a toolkit called Gotree/Goalign. The Gotree/Goalign toolkit implements more than 120 user-friendly commands and an API dedicated to multiple sequence alignment and phylogenetic tree manipulations. It is developed in Go, which makes executables easily installable, integrable in workflow environments, and parallelizable when possible. Moreover, Go is a compiled language, which accelerates computations compared to interpreted languages. This toolkit is freely available on most platforms (Linux, MacOS and Windows) and most architectures (amd64, i386) on GitHub at https://github.com/evolbioinfo/gotree, Bioconda and DockerHub.

INTRODUCTION

Increase in computer power and development of bioinformatics methods that handle very large datasets make it possible to perform phylogenetic analyses at an unprecedented scale. For example, it is common to perform phylogenetic studies involving several thousand trees or several thousand taxa (e.g. (1) or (2)). Such studies often present and run complex pipelines involving many steps and several tools and scripts, which makes them hard to describe and share. For example, the current COVID-19 pandemics and the diverse SARS-CoV-2 data analysis workflows that flourish demonstrate the need to manipulate sequences, alignments, and phylogenetic trees in an easy and automated way (e.g. https://github.com/roblanf/sarscov2phylo/).

Figure 1 displays an example of a phylogenetic workflow inspired from (3) and described in details in the Results section. The main objective of this workflow is to analyse a phylogenomic dataset to infer a species tree and compare it to a reference tree. It contains 19 steps among which the majority is not constituted by usual computer intensive tasks (e.g. tree inference and multiple sequence alignment), but rather by alignment and tree manipulations such as downloading, renaming, reformating, comparing, rerooting, annotating, etc. These tasks are (i) repetitive (found in many similar workflows), (ii) tedious to implement (many ways to do it) and (iii) error prone, which makes this workflow difficult to implement, describe and reproduce.

That is why we think a common framework was needed to help implement such analyses in a concise and reproducible way.

Several tools, either in the form of command line or APIs, already exist to manipulate phylogenetic trees and/or alignments. For example, Newick-utilities (10) is a command line tool dedicated to phylogenetic tree manipulation in newick format. ETE toolkit (11) is a Python API for the analysis and visualization of trees. Ape (12) is a package dedicated to the analyses of phylogenetics and evolution in the

*To whom correspondance should be addressed. Tel: +33 145 688 778; Email: frederic.lemoine@pasteur.fr
Present address: Olivier Gascuel, Institut de Systématique, Evolution, Biodiversité, ISYEB, UMR 7205, Muséum National d'Histoire Naturelle, CNRS, SU, EPHE, UA, Paris, France.

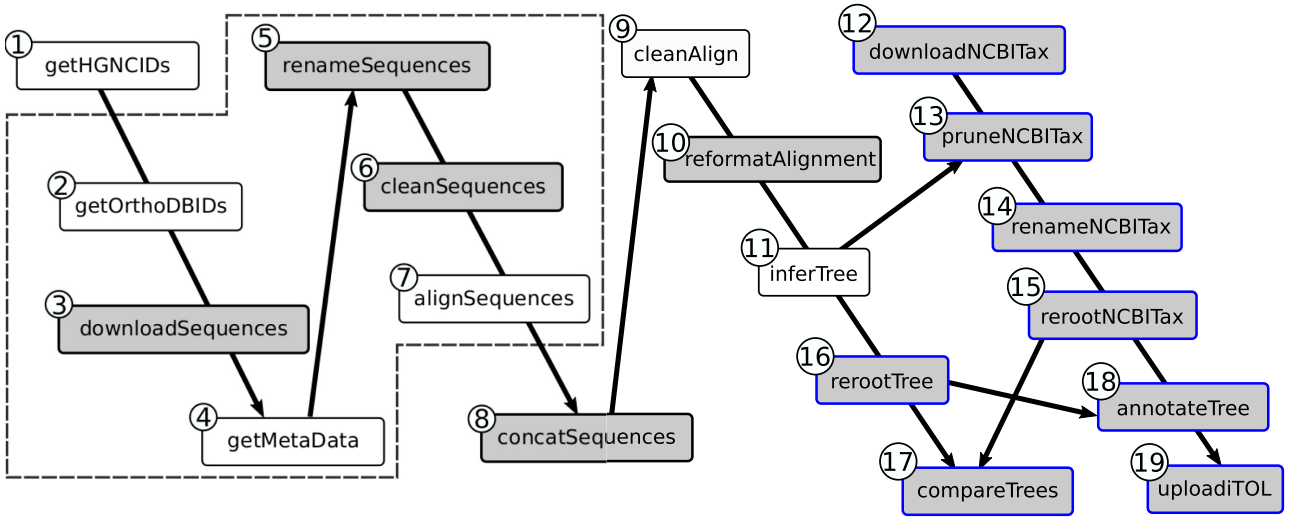

Figure 1. Structure of our use case phylogenomic workflow. It is made of 19 steps (processes), represented as boxes in the figure. Gray boxes represent processes performed using Goalign, and Gray boxes with blue contour represent processes performed using Gotree. Processes are linked by lines if results of the upstream process are needed for the execution of the downstream process. The steps represented by boxes surrounded by the dashed line are executed for each gene in parallel. The steps are the following: **1**) match RefSeq, NCBI and HGNC (4) gene identifiers; **2**) get OrthoDB (5) identifiers of orthologous groups corresponding to HGNC identifiers; **3**) download sequences of each group; **4**) get species name of each sequence (from OrthoDB id); **5**) rename the sequences using the species names; **6**) clean the sequences (e.g. removing special characters); **7**) align the sequences (MAFFT (6)); **8**) concatenate all alignments in a single large alignment; **9**) clean the alignment (BMGE (7)); **10**) reformat the alignment into Phylip format; **11**) infer the phylogenetic tree (IQ-TREE (8)); **12**) download the NCBI taxonomy in Newick format; **13**) keep only the 25 species of interest from NCBI taxonomy; **14**) change species names that differ between OrthoDB and NCBI taxonomy; **15–16**) Reroot NCBI taxonomy and inferred tree; **17**) compare both trees in terms of common bi-partitions; **18**) annotate inferred tree with NCBI taxonomy clades; **19**) upload the annotated tree to iTOL (9).

R language. Buddy Suite (13) is another python framework (command line) to manipulate phylogenetic trees and alignments. BIO++ (14) is a C++ library for the analysis of sequences and trees. Lastly, Phyx (15) is a command line tool written in C++ that performs phylogenetic analyses on trees and sequences. Some of these tools (e.g. ape, ETE toolkit, BIO++) are mainly developer oriented and may be difficult to use for non-programmer phylogenetic analysts who are not familiar with R, python or C++. Others are command line only (e.g. Newick-utilities), and may be of limited use for developers who want to use part of them in their own programs. They are all implemented either in C, C++, Python or R. To our knowledge, no phylogenetic tree toolkit currently exists, which mix alignment and tree manipulation, in a command line environment and via an API, and proposes a large diversity of commands. In particular, none exists for the Go programming language for which several bioinformatics libraries and tools already exist such as biogo (16) and Vcfanno (17).

In this context, we developed Gotree/Goalign, a toolkit dedicated to burdensome and repetitive phylogenetic tasks, which (i) consists of two user-friendly executables, `gotree` and `goalign`, integrating state of the art phylogenetic commands and requiring no programming skills to use, (ii) provides a set of chainable commands that are integrable in workflows (e.g. Nextflow (18) or Snakemake (19)), (iii) is straightforward to install via static binaries available for most platforms and architectures, and (iv) provides a public API accessible to developers wanting to manipulate phylogenetic trees and multiple sequence alignments in Go.

The Gotree/Goalign toolkit is already used in several public workflows such as Grapevine (the phylogenetic pipeline for the COG-UK project https://github.com/COG-UK/grapevine), ARTIC-EBOV (The ARTIC Ebola virus phylogenetic analysis protocol https://github.com/artic-network/artic-ebov), KOVID-TREES-NF (a Nextflow (18) workflow for SARS-CoV-2 phylogenetics https://github.com/MDU-PHL/kovid-trees-nf), bioconvert (a generic bioinformatic file format converter https://github.com/bioconvert/bioconvert), and in phylogenetic studies (20–26).

MATERIALS AND METHODS

Alignment manipulation

Goalign implements more than 60 commands to manipulate sequences and multiple sequence alignments. A subset of these commands is given in Table 1.

The most obvious (and maybe the most widely used) commands reformat an input sequence file into any supported format, which is essential to integrate different tools in the same analysis workflow. The supported input and output formats are: Fasta, Phylip, Clustal, Nexus, and tool specific formats (e.g. TNT and PaML input formats). It is worth noting that Goalign supports compressed input files (.gz, .bz and .xz) and remote files through http(s).

A second kind of commands is designed to extract summary statistics about input alignments, such as the length, the number of sequences, the frequency of each character, the list of sequence names, etc.

The third kind of commands aims at modifying input alignments (e.g. extract sequences or sites, clean, translate, rename, shuffle, mask, concatenate, append, etc.) or produc-

Table 1. Subset of commands implemented in Gotree/Goalign

Goalign Command	Description
reformat	Converts between Nexus, Clustal, Fasta and Phylip
rename	Modifies or cleans sequence names
clean	Removes sites/sequences
codonalign	Aligns by codons using a protein alignment
concat	Concatenates several alignments
mask	Masks parts of the alignment
translate	Translates input alignment in amino-acids
extract	Extracts subsequences from input alignments
build seqboot	Builds bootstrap alignments
compute distance	Computes distance matrix
stats	Computes statistics about input alignments
Gotree Command	**Description**
reformat	Converts between Newick, Nexus and PhyloXML
prune	Removes user-defined tips
collapse	Collapses branches
reroot	Reroots trees
compare trees	Compares tree topologies (bipartitions)
compute support	Computes branch supports (30) and (31)
stats	Computes statistics about input trees
matrix	Computes patristic distance matrix
upload itol	Uploads a tree to iTOL (9)

ing new alignments (build bootstrap alignments, align by codon, align pairwise, etc.).

Finally, some commands are designed to compute distance matrices for nucleotidic and proteic alignments. To this aim, Goalign implements main nucleotidic evolutionary models (JC, K2P, F81, F84 and TN93) and amino-acid matrices (DAYHOFF, JTT, MtRev, LG, WAG). Distances between amino-acid sequences are computed using maximum likelihood as in PhyML (27) and FastME (28).

Tree manipulation

Gotree implements more than 60 commands dedicated to the manipulation of phylogenetic trees (see Table 1 for a list of a few of them).

As for Goalign, Gotree reformatting commands are the first to be used in phylogenetic workflows and constitute the glue between all phylogenetic tools, which often support heterogeneous, specific formats. Gotree supports the following input and output formats: Newick, Nexus and PhyloXML. Moreover, Gotree supports compressed input files (.gz), and remote files from any urls or from dedicated servers (iTOL (9) and Treebase (29), See Supp. Text 1 for examples of such commands).

The second kind of commands implemented in Gotree produces summary statistics about input phylogenetic trees, such as the number of tips, number of branches, average branch length and branch support, sum of branch lengths, number of cherries, tree balance indices (Colless, Sackin), and patristic distance matrix.

A third kind of commands modifies the input tree, e.g. modifying branch lengths and supports, collapsing branches by length or support, removing lists of tips, rerooting the tree in different ways, etc.

Other commands generate new trees or new data on the trees, such as bootstrap support computations (Felsenstein's Bootstrap Proportions (30) and Transfer Bootstrap Expectation (31)) or random tree generation using different models.

The last kind of commands compares trees, for example comparing topologies (bipartition distance), or comparing tip names.

Put together, these commands provide the user with a large panel of possibilities adapted to many different situations, a few of which are described in the Results section.

Implementation of Gotree/Goalign

The Gotree/Goalign toolkit is implemented in the Go (https://golang.org/) programming language, and thus takes advantage of (i) being available on major operating systems (Windows, Linux, MacOS), (ii) the richness of the Go standard library and the massive amount of available packages (http, hashmaps, compression, regex, channels, concurrency, gonum, etc.), (iii) the distributed nature of Go packages, providing phylogenetic packages easily accessible to any developers.

It consists of two executables `gotree` and `goalign` providing all the commands in the same spirit as `git` or `docker`.

All the commands are implemented in the Unix mindset, having one atomic command per task, reading data from the standard input and writing results to the standard output when possible.

Integration in workflows

The Gotree/Goalign toolkit has been developed to be easily integrated in phylogenetic workflows, using main workflow managers such as Nextflow (18) or Snakemake (19). Two main characteristics facilitate this behavior.

First of all, Gotree/Goalign consists of a large set of command line tools that can be easily chained, and that avoid manual interventions as much as possible. This feature makes them easily usable in workflows as all commands are atomic, easily citable, describable and reproducible.

Moreover, they are made available through three channels that are compatible with the diversity of use in workflow managers, i.e., multi-plateform executables, Docker and Singularity containers, and Bioconda packages.

RESULTS: USE CASE IN PHYLOGENOMICS

In this use case, we analyze a phylogenomic dataset inspired from (3), in which the authors analyze a set of 1730 genes from primates. They infer the species tree either from individual gene trees using ASTRAL III (32) or from gene concatenation using maximum likelihood. Our use case is inspired from the concatenation study, inferring a species tree from available groups of primate orthologous proteins in OrthoDB (5) after having mapped the RefSeq identifiers of the 1730 analyzed genes to their OrthoDB counterpart.

The workflow is displayed in more detail in Figure 1 (code available at https://github.com/evolbioinfo/gotree_usecase and in Supp. Data 1), and consists of the following main

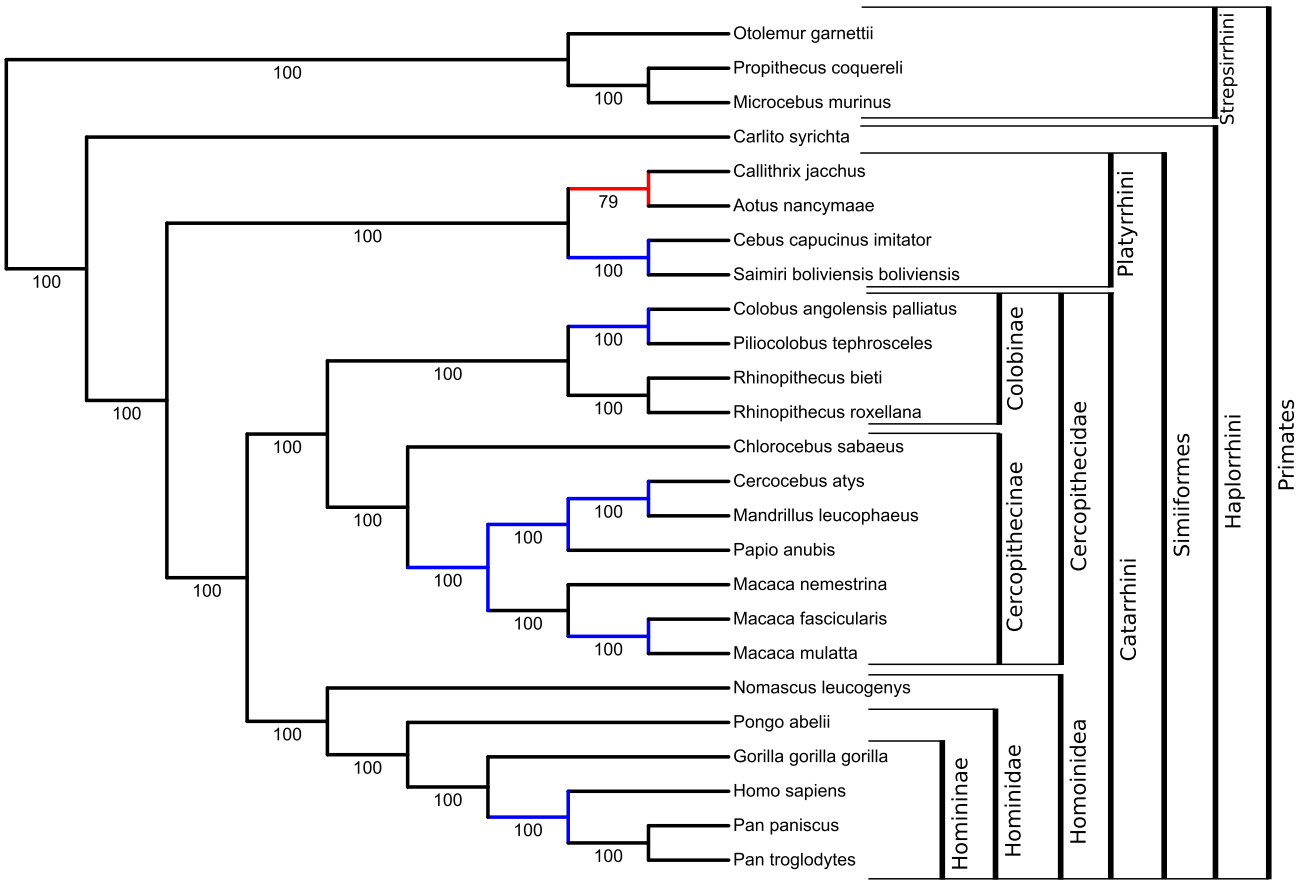

Figure 2. Tree inferred from the concatenation of 1315 orthologous proteins in 25 primates using maximum likelihood, with Felsenstein's bootstrap supports of the internal branches. Visualisation of tree topology with branch supports comes from iTOL (9) after upload by Gotree. Branch colors and clade annotations have been added independently. The red branch is the only one that contradicts the NCBI taxonomy, and blue branches are resolved in the inferred tree, but unresolved in the NCBI taxonomy. The topology is identical to the tree inferred with maximum likelihood from gene concatenation in (3).

steps (details of the tools and options are given in Supplementary Figure S1): (i) mapping RefSeq, HGNC (4) and OrthoDB identifiers ; (ii) retrieving each group of orthologous proteins in OrthoDB if they are present in at least 90% of the 25 primate species (Supp. Data 2) and in a single copy (this results in 1315 orthologous groups, listed in Supp. Data 3); (iii) for each orthologous group, downloading the sequences of all proteins and their associated metadata; (iv) renaming, cleaning and aligning sequences using MAFFT (6); (v) concatenating the alignments of the 1315 orthologous groups (this results in a single alignment of 674 089 amino-acids); (vi) cleaning the alignment using BMGE (7) (this keeps 516 999 sites); (vii) inferring the species tree using IQ-TREE (8); (viii) downloading the NCBI taxonomy; (ix) comparing the inferred tree with the NCBI taxonomy and (x) uploading both trees to iTOL (9). Each step and their parameters are described in detail in Supplementary Figure S1.

Since it involves many very different steps, this example is challenging to implement and to run fully automatically, without manual intervention. Apart from tree inference and multiple sequence alignment, most of the operations required to develop this workflow are available in the Gotree/Goalign toolkit and are straightforward to execute (Supplementary Figure S1). The implementation of the workflow, made of 19 steps and ∼350 lines of code, is fully homogeneous in terms of file formats and input/outputs. Furthermore, the ability to run the full analysis, from the input data to the output tree visualization, without manual intervention facilitates its description and reproducibility.

The resulting primate phylogeny, given in Figure 2, is in almost complete agreement with the NCBI taxonomy, with 15 out of 16 branches of the NCBI taxonomy found in our tree. It is worth noting that the only conflicting branch (in red), grouping Callitrix and Aotus, has a lower bootstrap support than other branches and is also present in the tree inferred using maximum likelihood from gene concatenation in (3). In fact, our topology and theirs (3) are identical.

DISCUSSION

We developed the Gotree/Goalign toolkit to simplify the manipulation of phylogenetic trees and alignments, and to facilitate the development of reproducible phylogenetic workflows. Importantly, it is not a wrapper around major tree inference and multiple sequence alignment software, i.e. it does not take care of aligning sequences and inferring trees, but rather takes care of all the other complex, tedious and numerous tree and sequence operations that are

necessary in phylogenetic workflows. The Gotree/Goalign toolkit is developed in Go and is easily installable on major operating systems and provides a public API usable in any Go project.

It is already used in several projects, and we are confident that Gotree/Goalign will be able to build a community interested in adding new functionnalities.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

First, we thank the Gotree/Goalign users for their feedbacks, which allow us to add new functionnalities and improve existing ones. We also thank all the *Evolutionary Bioinformatics* lab: Anna Zhukova, Marie Morel, Luc Blassel and Jakub Voznica for testing and using Gotree/Goalign extensively.

FUNDING

PRAIRIE [ANR-19-P3IA-0001 to O.G.]. Funding for open access charge: Institute Pasteur.
Conflict of interest statement. None declared.

REFERENCES

1. Boussau,B., Szöllősi,G.J., Duret,L., Gouy,M., Tannier,E. and Daubin,V. (2013) Genome-scale coestimation of species and gene trees. *Genome Res.*, **23**, 323–330.
2. Jetz,W., Thomas,G., Joy,J., Hartmann,K. and Mooers,A. (2012) The global diversity of birds in space and time. *Nature*, **491**, 444–448.
3. Vanderpool,D., Minh,B.Q., Lanfear,R., Hughes,D., Murali,S., Harris,R.A., Raveendran,M., Muzny,D.M., Hibbins,M.S., Williamson,R.J. and et,al. (2020) Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLoS Biol.*, **18**, e3000954.
4. Povey,S., Lovering,R., Bruford,E., Wright,M., Lush,M. and Wain,H. (2001) The HUGO gene nomenclature committee (HGNC). *Hum. Genet.*, **109**, 678–680.
5. Kriventseva,E.V., Kuznetsov,D., Tegenfeldt,F., Manni,M., Dias,R., Simão,F.A. and Zdobnov,E.M. (2018) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.*, **47**, D807–D811.
6. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
7. Criscuolo,A. and Gribaldo,S. (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.*, **10**, 210.
8. Minh,B.Q., Schmidt,H.A., Chernomor,O., Schrempf,D., Woodhams,M.D., Von Haeseler,A. and Lanfear,R. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, **37**, 1530–1534.
9. Letunic,I. and Bork,P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296.
10. Junier,T. and Zdobnov,E.M. (2010) The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics*, **26**, 1669–1670.
11. Huerta-Cepas,J., Dopazo,J. and Gabaldón,T. (2010) ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*, **11**, 24.
12. Paradis,E., Claude,J. and Strimmer,K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.
13. Bond,S.R., Keat,K.E., Barreira,S.N. and Baxevanis,A.D. (2017) BuddySuite: command-line toolkits for manipulating sequences, alignments, and phylogenetic trees. *Mol. Biol. Evol.*, **34**, 1543–1546.
14. Dutheil,J., Gaillard,S., Bazin,E., Glémin,S., Ranwez,V., Galtier,N. and Belkhir,K. (2006) Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, **7**, 188.
15. Brown,J.W., Walker,J.F. and Smith,S.A. (2017) Phyx: phylogenetic tools for unix. *Bioinformatics*, **33**, 1886–1888.
16. Kortschak,R.D. and Adelson,D.L. (2015) bíogo: a simple high-performance bioinformatics toolkit for the Go language. bioRxiv doi: https://doi.org/10.1101/005033, 27 March 2015, preprint: not peer reviewed.
17. Pedersen,B.S., Layer,R.M. and Quinlan,A.R. (2016) Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol.*, **17**, 118.
18. Tommaso,P.D., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
19. Köster,J. and Rahmann,S. (2012) Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
20. Bigot,T., Guglielmini,J. and Criscuolo,A. (2019) Simulation data for the estimation of numerical constants for approximating pairwise evolutionary distances between amino acid sequences. *Data Brief.*, **25**, 104212.
21. Theys,K., Lemey,P., Vandamme,A.-M. and Baele,G. (2019) Advances in visualization tools for phylogenomic and phylodynamic studies of viral diseases. *Front. Public Health*, **7**, 208.
22. Guglielmini,J., Bourhy,P., Schiettekatte,O., Zinini,F., Brisse,S. and Picardeau,M. (2019) Genus-wide Leptospira core genome multilocus sequence typing for strain taxonomy and global surveillance. *PLoS Neglect. Trop. Dis.*, **13**, e0007374.
23. Modi,V. and Dunbrack,R.L. (2019) A structurally-validated multiple sequence alignment of 497 human protein kinase domains. *Sci. Rep. UK*, **9**, 1–16.
24. Baidaliuk,A., Lequime,S., Moltini-Conclois,I., Dabo,S., Dickson,L.B., Prot,M., Duong,V., Dussart,P., Boyer,S., Shi,C. *et al.* (2020) Novel genome sequences of cell-fusing agent virus allow comparison of virus phylogeny with the genetic structure of *Aedes aegypti* populations. *Virus Evolution*, **6**, veaa018.
25. Dalai,S.C., Junqueira,D.M., Wilkinson,E., Mehra,R., Kosakovsky Pond,S.L., Levy,V., Israelski,D., de Oliveira,T. and Katzenstein,D. (2018) Combining phylogenetic and network approaches to identify HIV-1 transmission links in San Mateo county, California. *Front. Microbiol.*, **9**, 2799.
26. Turakhia,Y., Thornlow,B., Hinrichs,A.S., Maio,N.D., Gozashti,L., Lanfear,R., Haussler,D. and Corbett-Detig,R. (2021) Ultrafast Sample placement on Existing tRees (USHER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.*, **53**, 809–816.
27. Guindon,S., Dufayard,J.-F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *System. Biol.*, **59**, 307–321.
28. Lefort,V., Desper,R. and Gascuel,O. (2015) FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.*, **32**, 2798–2800.
29. Piel,W.H., Donoghue,M., Sanderson,M. and Netherlands,L. (2000) TreeBASE: a database of phylogenetic information. In *Proceedings of the 2nd International Workshop of Species*. Vol. **2000**.
30. Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.
31. Lemoine,F., Entfellner,J.-B.D., Wilkinson,E., Correia,D., Felipe,M.D., De Oliveira,T. and Gascuel,O. (2018) Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, **556**, 452–456.
32. Zhang,C., Rabiee,M., Sayyari,E. and Mirarab,S. (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, **19**, 15–30.