


# A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees

Jakob McBroome,<sup>†,1,2</sup> Bryan Thornlow,<sup>†,1,2</sup> Angie S. Hinrichs,<sup>2</sup> Alexander Kramer,<sup>1,2</sup> Nicola De Maio,<sup>3</sup> Nick Goldman <sup>3</sup> David Haussler,<sup>1,2</sup> Russell Corbett-Detig,<sup>\*,1,2</sup> and Yatish Turakhia<sup>\*,1,2</sup>

<sup>1</sup>Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, USA

<sup>2</sup>Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA

<sup>3</sup>European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, United Kingdom

<sup>†</sup>These authors contributed equally to this work.

\* **Corresponding authors:** E-mails: yturakhi@ucsc.edu; rucorbet@ucsc.edu.

**Associate editor:** Jian Lu

## Abstract

The vast scale of SARS-CoV-2 sequencing data has made it increasingly challenging to comprehensively analyze all available data using existing tools and file formats. To address this, we present a database of SARS-CoV-2 phylogenetic trees inferred with unrestricted public sequences, which we update daily to incorporate new sequences. Our database uses the recently proposed mutation-annotated tree (MAT) format to efficiently encode the tree with branches labeled with parsimony-inferred mutations, as well as Nextstrain clade and Pango lineage labels at clade roots. As of June 9, 2021, our SARS-CoV-2 MAT consists of 834,521 sequences and provides a comprehensive view of the virus' evolutionary history using public data. We also present matUtils—a command-line utility for rapidly querying, interpreting, and manipulating the MATs. Our daily-updated SARS-CoV-2 MAT database and matUtils software are available at [http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/USHER\\_SARS-CoV-2/](http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/USHER_SARS-CoV-2/) and <https://github.com/yatisht/usher>, respectively.

**Key words:** COVID-19, SARS-CoV-2 phylogenetics, genomic surveillance.

The COVID-19 pandemic has inspired unprecedented levels of genome sequencing for a single pathogen (Hodcroft et al. 2021). Over a million SARS-CoV-2 genomes have been sequenced worldwide so far, and tens of thousands of new genomes are getting uploaded daily (Maxmen 2021). These data have enabled scientists to closely track the evolution and transmission dynamics of the virus at global and local scales (Deng et al. 2020; Chaillon and Smith 2021; da Silva Filipe et al. 2021). However, the scale of these data is posing serious computational challenges for comprehensive phylogenetic analyses (Hodcroft et al. 2021). Platforms like Nextstrain (Hadfield et al. 2018) have been invaluable in studying viral transmission networks and genomic surveillance efforts, but they only provide subsampled SARS-CoV-2 trees consisting of a tiny fraction of available data, omitting phylogenetic relationships with most available sequences. A single, comprehensive SARS-CoV-2 reference tree of all available data could not only facilitate detailed and unambiguous phylogenetic analyses at global, country, and local levels but may also help promote consistency of results across different research groups (Turakhia et al. 2020).

The massive volume of SARS-CoV-2 data also poses numerous data sharing challenges with existing file formats, such as Fasta or Variant Call Format (VCF), which are bulky and

necessitate network speeds and computational capabilities that are beyond the reach of many research and scientific groups.

## New Approaches

In this work, we simultaneously address the issue of maintaining a comprehensive SARS-CoV-2 reference tree and its associated data processing, sharing, and analysis challenges. Specifically, we are maintaining and openly sharing a daily-updated database of mutation-annotated trees (MATs) containing global SARS-CoV-2 sequences from public databases, without any downsampling (other than for quality control, see [supplementary methods, Supplementary Material online](#)), including annotations for Nextstrain clades (Hadfield et al. 2018) and Pango lineages (Rambaut et al. 2020; [supplementary fig. 1, Supplementary Material online](#)). The MAT is an extremely efficient data format proposed recently (Turakhia, Thornlow, Hinrichs, De Maio, et al. 2021) which uses a form of phylogenetic compression (Ané and Sanderson 2005) to facilitate sharing of extremely large genome sequence data sets. An uncompressed MAT of 834,521 SARS-CoV-2 public sequences requires only 65 MB to store and encodes more information than in a 43 GB VCF containing a single-nucleotide variation of all sequences (the MAT format does not

handle insertions and deletions [Turakhia, Thornlow, Hinrichs, De Maio, et al. 2021]) and a 38 MB Newick file containing the phylogenetic tree topology.

To accompany this database, we present *matUtils*—a toolkit for rapidly querying, interpreting, and manipulating the MATs included in our database or constructed with USHER (Turakhia, Thornlow, Hinrichs, De Maio, et al. 2021). Using *matUtils*, common operations in genomic surveillance and contact tracing efforts, including annotating an MAT with new clades, extracting specific subtrees, or converting the MAT to standard Newick or VCF format, can be performed in a matter of seconds to minutes even on a laptop. We also provide a web interface for *matUtils* through the UCSC SARS-CoV-2 Genome Browser (Fernandes et al. 2020). Together, our SARS-CoV-2 database and *matUtils* toolkit can simultaneously democratize and accelerate pandemic-related research.

## Results and Discussion

### A Daily-Updated MAT Database of Global SARS-CoV-2 Sequences

To aid the scientific community studying the mutational and transmission dynamics of the SARS-CoV-2 virus and its different variants, we are maintaining a daily-updated database of SARS-CoV-2 MATs composed of public data. Starting with the final Newick tree release dated November 13, 2020, of Rob Lanfear's *sarscov2phylo* (<https://github.com/roblanf/sarscov2phylo>, last accessed September 6, 2021) that is rerooted to Wuhan/Hu-1 (GenBank MN908947.3, RefSeq NC\_045512.2), we have set up an automated pipeline to aggregate public sequences available through GenBank (Clark et al. 2007), COG-UK (Nicholls et al. 2021), and the China National Center for Bioinformation on a daily basis and incorporate them into our MAT using USHER (supplementary methods, Supplementary Material online). GISAID data (Shu and McCauley 2017) are not included in our MATs because its usage terms do not allow redistribution. Similar to GISAID, our database is subject to the sampling bias resulting from the vast disparity in the sequencing efforts of various countries (Cyranoski 2021, supplementary fig. 1B, Supplementary Material online). We also use the *matUtils annotate* command (supplementary methods, Supplementary Material online) to add Nextstrain clade and Pango lineage annotations to individual branches of our MAT. As of June 9, 2021, our MAT consists of 834,521 sequences, includes 14 Nextstrain clade and 895 Pango lineage annotations for all samples, and is only 65 MB, or 14 MB when gzip-compressed (supplementary fig. 1 and table S1, Supplementary Material online). To our knowledge, this is the most comprehensive representation of the SARS-CoV-2 evolutionary history using publicly available sequences. It can be freely used to study evolutionary and transmission dynamics of the virus at global, country, and local levels, and can be visualized using the *Cov2Tree* tool (<https://cov2-tree.org/>, last accessed September 6, 2021) developed by Theo Sanderson.

### *matUtils* Provides a Wide Range of Functions to Analyze and Manipulate MATs

We have created a high-performance command-line utility called *matUtils* for performing a wide range of operations on MATs for rapid interpretation and analysis in genomic surveillance and contact tracing efforts. *matUtils* is distributed with the USHER package (Turakhia, Thornlow, Hinrichs, De Maio, et al. 2021) and uses the same MAT format as USHER. *matUtils* is organized into five different subcommands: *annotate*, *summary*, *extract*, *uncertainty*, and *introduce* (fig. 1), described briefly below. We provide detailed instructions for the usage of each module on our wiki (<https://usher-wiki.readthedocs.io/en/latest/matUtils.html>, last accessed September 6, 2021).

#### *Annotate*

This function annotates clades in an MAT. One of the central uses of phylogenetics during the pandemic is to trace the emergence and spread of new viral lineages. Nextstrain (Hadfield et al. 2018), Pango (Rambaut et al. 2020), and GISAID (Shu and McCauley 2017) provide different nomenclatures for SARS-CoV-2 variants that have been used widely in genomic surveillance. Our MAT format provides the ability to annotate internal branches of the tree with an array of clade names, one for each clade nomenclature. *matUtils annotate* provides two methods for annotation: 1) directly providing the mappings of each clade name to its corresponding node or 2) providing a set of representative sample names for each clade from which the clade roots can be automatically inferred (supplementary methods, Supplementary Material online). Both methods ensure that the clades remain monophyletic, but we use the second approach to label Nextstrain clades and Pango lineages in our SARS-CoV-2 MAT database since it can be automated using available data (supplementary methods, Supplementary Material online). *matUtils annotate* has high congruence with Nextstrain clades and Pango lineage annotations (supplementary table S1, Supplementary Material online).

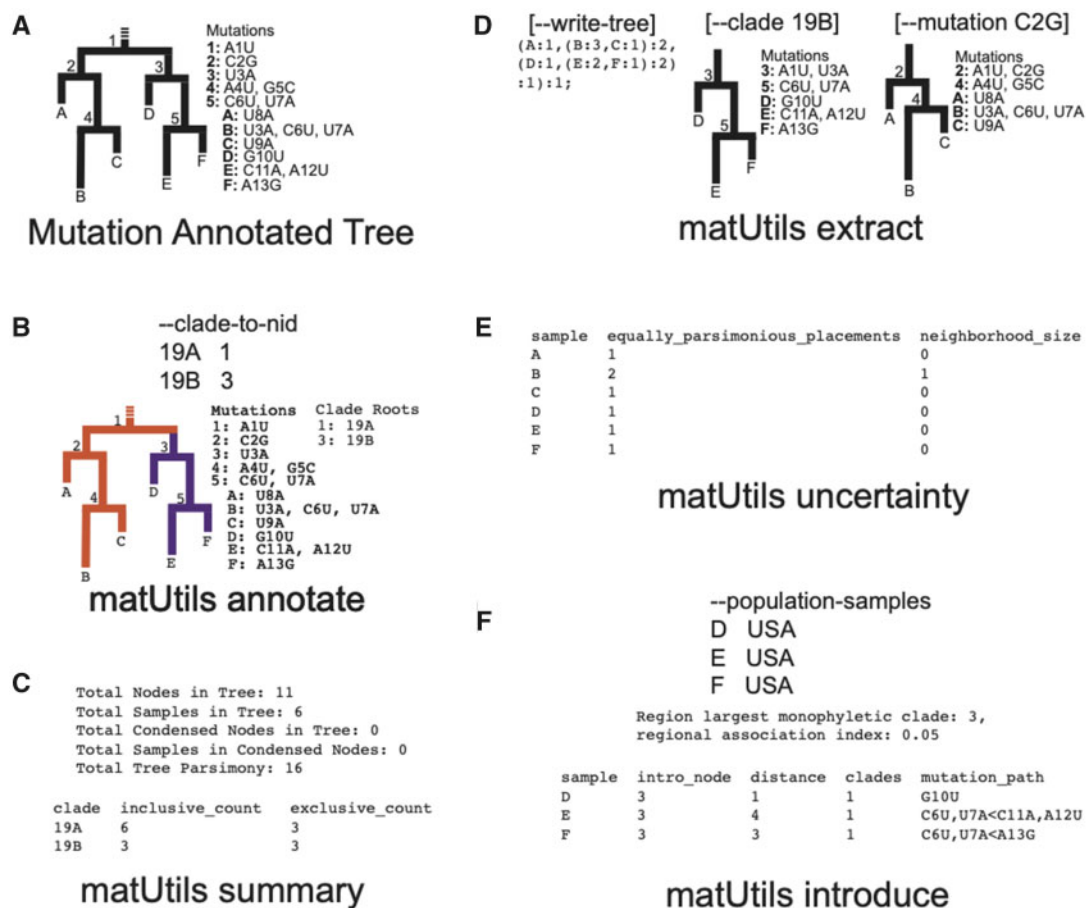
Once clades are annotated on an MAT, the USHER placement tool (Turakhia, Thornlow, Hinrichs, De Maio, et al. 2021) can assign each newly placed sequence to its corresponding Pango lineage. This is being used as a feature in Pangolin 3.0 (<https://github.com/cov-lineages/pangolin/releases/tag/v3.0>, last accessed September 6, 2021) to perform clade assignments in a fully phylogenetic framework.

#### *Summary*

This function provides a brief summary of the available data in the input MAT file and is meant to serve as a typical first step in any MAT-based analysis. It provides a count of the total number of samples in the MAT, the size of each annotated clade, the total parsimony score (i.e., the sum of mutation events on all branches of the MAT), the number of distinct mutations, phylogenetically informed translation of mutations, and other similar statistics.

#### *Extract*

Many SARS-CoV-2 phylodynamic studies involve restricting the analysis to a smaller tree of interest. Although it can be



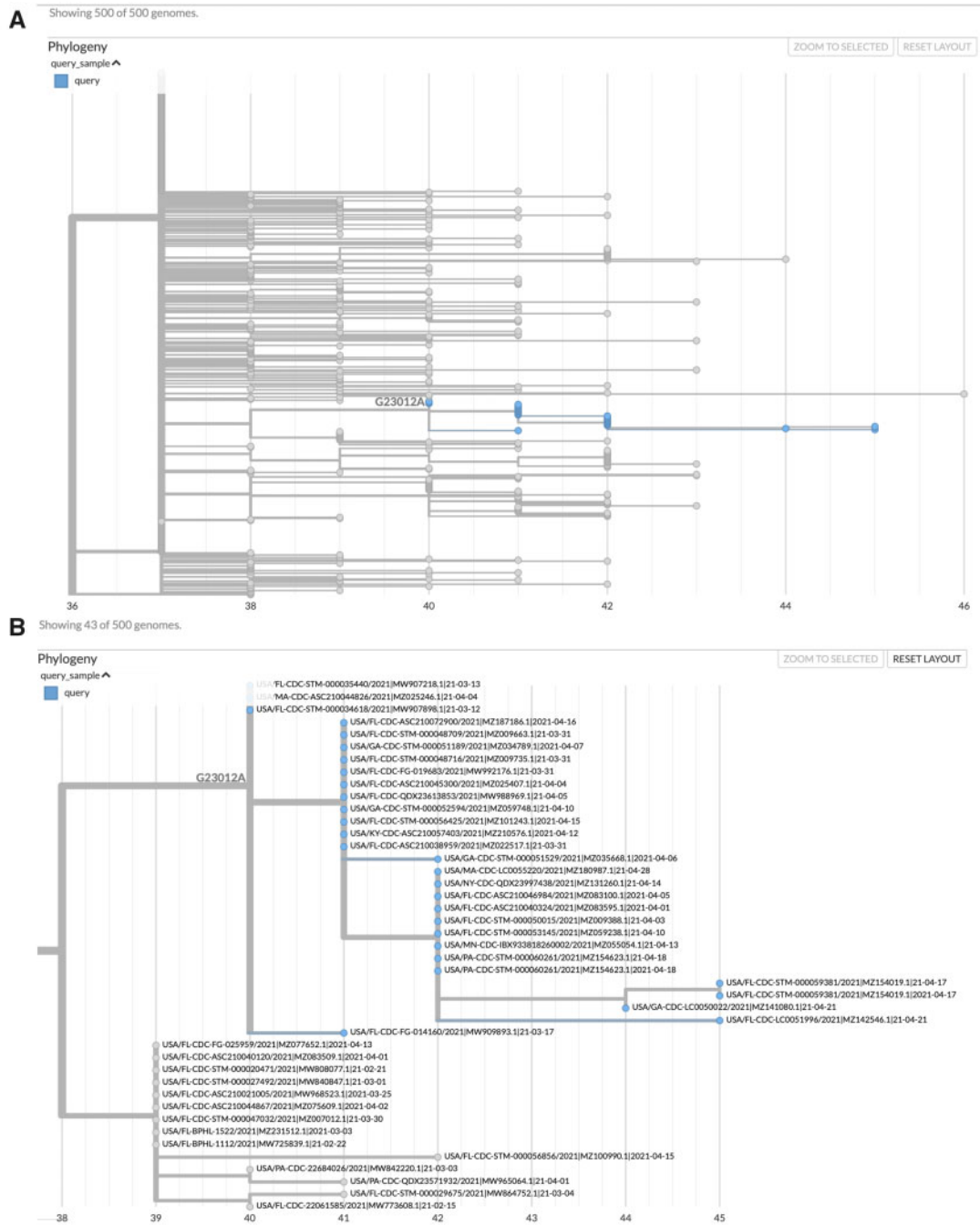
**FIG. 1.** *matUtils* functions enable fast, user-friendly analysis of MATs. (A) An example MAT with tree topology corresponding to the MAT on the left and the mutation annotations on each node shown on the right. (B) *matUtils annotate* allows the user to annotate internal nodes with clade names. In this example, nodes 1 and 3 are annotated with clade names 19A and 19B, respectively. This MAT serves as an input to commands shown in panels C–F. (C) *matUtils summary* outputs sample-, clade-, and tree-level statistics for the input MAT. (D) *matUtils extract* allows users to convert an MAT to Newick format (left), subset the MAT for a specified clade (center) or mutation (right), among other functions. (E) *matUtils uncertainty* outputs parsimony scores, equally parsimonious placements and neighborhood sizes for each sample of an input MAT. Sample B has two equally parsimonious placements, as it could also be placed as a descendant of node 5 with terminal mutations C2G, A4U, and G5C. (F) *matUtils introduce* can take a list of samples of interest as input and output the largest monophyletic clade and regional association index associated with the input population, along with their predicted introduction nodes and paths. In all panels, user input commands are shown in large fonts (e.g., “*matUtils annotate*”) and output text from these commands is shown in monospaced fonts.

computationally challenging to identify samples most closely related to a given sample or cluster from over a million other sequences, it is straightforward to retrieve subtrees from a comprehensive phylogeny. *matUtils extract* provides an efficient and robust suite of options for subtree selection from an MAT. A user can use *matUtils extract* to subsample an MAT to find samples that contain a mutation of interest, are members of a specific clade, have a name matching a specific regular expression pattern (such as the expression “[IND\*|India\*]” to select samples from India), among other criteria (supplementary methods, Supplementary Material online). *matUtils extract* also includes options to identify from an MAT sequences which have descended from long internal branches in the tree, which can sometimes arise from recombination (Jackson et al. 2021; Turakhia, Thornlow, Hinrichs, McBroome, et al. 2021), or those with an unusually high parsimony score, which are indicative of low-quality sequences (Mai and Mirarab 2018). Notably, *matUtils extract*

can produce an output Auspice v2 JSON that is compatible with the Auspice tree visualization tool (Hadfield et al. 2018; fig. 2, supplementary methods, Supplementary Material online). *matUtils extract* can also convert an MAT into other file formats, such as a Newick for its corresponding phylogenetic tree and a VCF for its corresponding genome variation data. *matUtils extract* also provides an option to resolve all polytomies in an MAT arbitrarily, similar to the *muti2di* functionality in *ape* (Paradis and Schliep 2019), for compatibility with phylogenetic tools that do not allow polytomies.

#### Uncertainty

A fundamental concern in SARS-CoV-2 phylogenetics is topological uncertainty (Hodcroft et al. 2021), which may result from contaminated sequences or sample mixtures (Turakhia, Thornlow, Hinrichs, De Maio, et al. 2021). The impact of this concern depends on the biological context of the analysis. *matUtils uncertainty* provides a topological uncertainty statistic



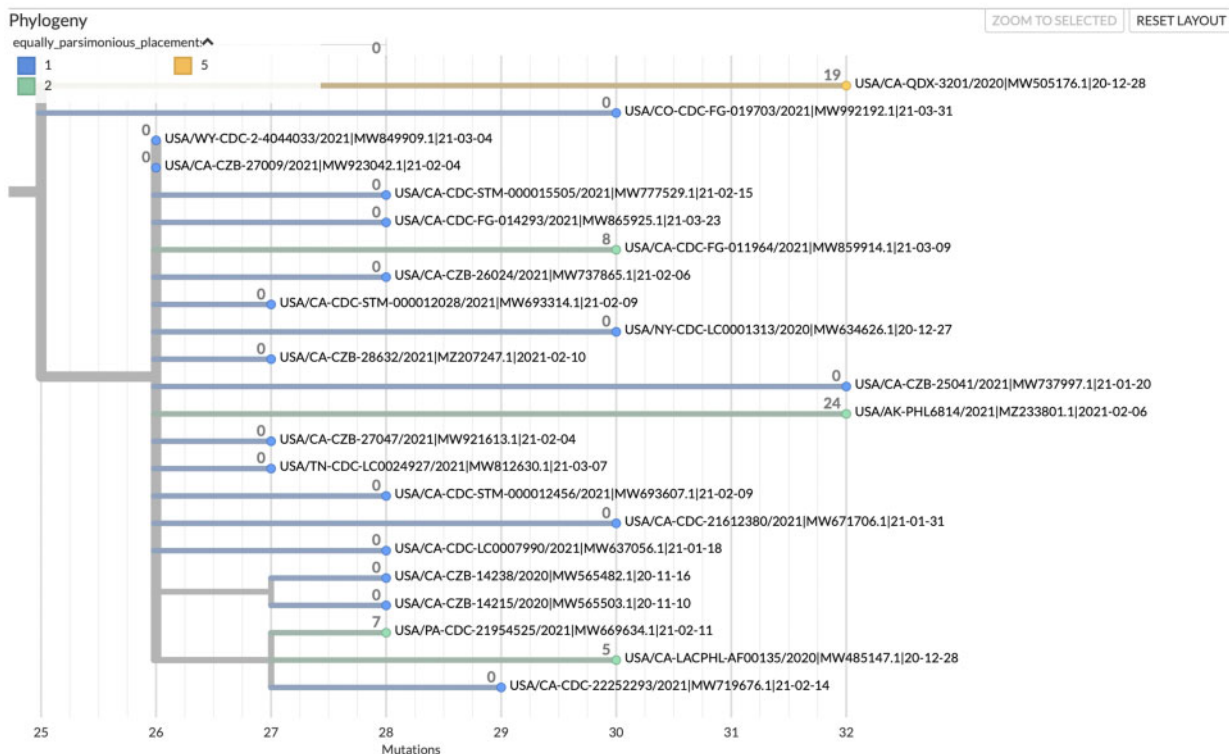
**FIG. 2.** *matUtils* can generate informative visuals with Auspice. The above trees represent a clade of related B.1.1.7 samples from the United States which secondarily acquired the potentially important spike protein mutation E484K, which is caused by the nucleotide mutation G23012A. These trees were obtained by running the command “*matUtils* extract -i public-2021-06-09.all.masked.nextclade.pangolin.pb.gz -c B.1.1.7 -m G23012A -H '(USA.\*)' -N 500 -j clade\_trees -d clade\_out,” which selects all samples from clade B.1.1.7 which acquired this mutation and are from the United States, then identifies the minimum set of 500 sample subtrees which contain all of these samples, creating an Auspice v2 format JSON for each subtree (Hadfield et al 2018). This results in 35 distinct subtree JSON files of 500 samples each in the output directory. Panel A represents the entirety of subtree six as viewed with Auspice (Hadfield et al 2018), including blue highlights and a branch label where our mutation of interest occurred. Panel B is zoomed in on this subtree and its sister clade; at this scale, we can read individual sample names and observe that this specific strain has been actively spreading in the United States during April 2021.

that computes the number of equally parsimonious placements that exist for each specified sample in the input MAT. Importantly, *matUtils* also allows the user to calculate equally parsimonious positions for already placed samples. This is accomplished by pruning the sample from the tree and placing the sample back to the tree using the placement module of

USHER (Turakhia, Thornlow, Hinrichs, De Maio, et al. 2021; supplementary methods, Supplementary Material online). *matUtils uncertainty* additionally records the number of mutations separating the two most distant equally parsimonious placements, reflecting the distribution of placements across the tree (supplementary methods, Supplementary Material

## mutation\_annotated\_tree

Showing 23 of 50 genomes.



**Fig. 3.** matUtils uncertainty statistics reveal low-quality sample placements. This Auspice view of an example subtree is annotated with both equally parsimonious placements (in color) and neighborhood size (branch label integers). Eighteen of our 23 samples in the subtree have a single placement and a neighborhood size of 0, indicating high placement certainty for those samples. Of the five samples with multiple equally parsimonious placements, one sample has five equally parsimonious placements with an NSS value of 19, indicating a high level of placement uncertainty for this sample spanning a relatively large neighborhood.

online). The output file is compatible as “drag-and-drop” meta-data with the Auspice platform, which allows for a rapid visualization of potentially problematic placements (fig. 3).

### Introduce

Public health officials are often concerned about the number of new introductions of the virus genome in a given country or local area. To aid this analysis, *matUtils* introduce can calculate the association index (Wang et al. 2001) or the maximum monophyletic clade size statistic (Salemi et al. 2005; Parker et al. 2008) for arbitrary sets of samples, along with simple heuristics for approximating points of introduction into a region (supplementary methods, Supplementary Material online).

### matUtils Enables Rapid Analysis of a Comprehensive SARS-CoV-2 Global Tree and Its Web Interface

The *matUtils* toolkit is designed to scale efficiently to SARS-CoV-2 phylogenies containing millions of samples. Using *matUtils*, common pandemic-relevant operations described in the earlier section can be performed in the order of seconds to minutes with the current scale of SARS-CoV-2 data (supplementary tables S2–S9, Supplementary Material online). For example, it takes only 5 s to summarize the information

contained in our June 9, 2021 SARS-CoV-2 MAT of 834,521 samples and only 15 s to extract the mutation paths from the root to every sample in the MAT (supplementary table S2, Supplementary Material online). Since *matUtils* is primarily designed to work with the newly proposed and information-rich MAT format, it does not have direct counterparts in other bioinformatic software packages currently, but its efficiency is similar or better than state-of-the-art tools that offer comparable functionality (supplementary tables S2–S9, Supplementary Material online). For example, *matUtils* is able to resolve polytomies in an 834,521 sample tree in 9 s, a task which takes over 37 min using *ape* (Paradis and Schliep 2019; supplementary table S3, Supplementary Material online). *matUtils* is also very memory-efficient, requiring less than 1.4 GB of main memory for most tasks, making it possible to run even on laptop devices.

Certain functions of *matUtils* (such as extracting subtrees of provided sample names or identifiers) have also been ported to UCSC SARS-CoV-2 Genome Browser (Fernandes et al. 2020) and are available from <https://genome.ucsc.edu/cgi-bin/hgPhyloPlace> (last accessed September 6, 2021). Our database and utility fill a critical need for open, public, rapid analysis of the global SARS-CoV-2 phylogeny by health departments and research groups across the world, with highly efficient file formats that do not require high-speed

internet connectivity or large storage devices, and tools capable of rapidly performing large-scale analyses on laptops.

## Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

## Acknowledgments

The authors thank Rob Lanfear for reviewing this manuscript and his valuable feedback. They thank Cheng Ye for his help in parallelizing VCF extraction. They also thank all the laboratories that submit data to public databases. J.M., B.T., and R.C.-D. were supported by R35GM128932 and by an Alfred P. Sloan Foundation fellowship to R.C.-D., J.M., and B.T. were funded by T32HG008345 and F31HG010584. The UCSC Genome Browser is funded by NHGRI, currently with grant 5U41HG002371. The SARS-CoV-2 database is funded by generous individual donors including Eric and Wendy Schmidt by recommendation of the Schmidt Futures program. A.K. is supported by CDC award BAA 200-2021-11554. N.D.M. and N.G. are funded by the European Molecular Biology Laboratory (EMBL). Y.T. is funded through Schmidt Futures Foundation SF 857 and NIH grant 5R01HG010485.

## Data Availability

Our daily-updated SARS-CoV-2 MAT database and matUtils software are available at [http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/USHER\\_SARS-CoV-2/](http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/USHER_SARS-CoV-2/) and <https://github.com/yatisht/usher>, respectively. For benchmarking, the exact commands used for each comparison can be found in Supplementary Tables S2–S9, and the input data used for each comparison can be found at [https://github.com/bpt26/matutils\\_benchmarking/](https://github.com/bpt26/matutils_benchmarking/).

## References

Ané C, Sanderson MJ. 2005. Missing the forest for the trees: phylogenetic compression and its implications for inferring complex evolutionary histories. *Syst Biol*. 54(1):146–157.

Chaillon A, Smith DM. 2021. Phylogenetic analyses of SARS-CoV-2 B.1.1.7 lineage suggest a single origin followed by multiple exportation events versus convergent evolution. *Clin Infect Dis*. Advance Access published March 26, 2021, doi:10.1093/cid/ciab265

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al.; Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450(7167):203–218.

Cyranoski D. 2021. Alarming COVID variants show vital role of genomic surveillance. *Nature* 589(7842):337–338.

da Silva Filipe A, Shepherd JG, Williams T, Hughes J, Aranday-Cortes E, Asamaphan P, Ashraf S, Balcazar C, Brunker K, Campbell A, et al.; COVID-19 Genomics UK (COG-UK) Consortium. 2021. Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. *Nat Microbiol*. 6(1):112–122.

Deng X, Gu W, Federman S, Plessis L D, Pybus OG, Faria NR, Wang C, Yu G, Bushnell B, Pan C-Y, et al. 2020. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* 369(6503):582–587.

Fernandes JD, Hinrichs AS, Clawson H, Gonzalez JN, Lee BT, Nassar LR, Raney BJ, Rosenbloom KR, Nerli S, Rao AA, et al. 2020. The UCSC SARS-CoV-2 genome browser. *Nat Genet*. 52(10):991–998.

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34(23):4121–4123.

Hodcroft EB, Maio ND, Lanfear R, MacCannell DR, Minh BQ, Schmidt HA, Stamatakis A, Goldman N, Dessimoz C. 2021. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* 591(7848):30–33.

Jackson B, Boni MF, Bull MJ, Collieran A, Colquhoun RM, Darby A, Haldenby S, Hill V, Lucaci A, McCrone JT, et al. 2021. Generation and transmission of inter-lineage recombinants in the SARS-CoV-2 pandemic. *Cell*. Advance Access published August 17, 2021, doi:10.1101/2021.06.18.21258689.

Mai U, Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* 19(Suppl 5):272.

Maxmen A. 2021. One million coronavirus sequences: popular genome site hits mega milestone. *Nature* 593(7857):21.

Nicholls SM, Poplawski R, Bull MJ, Underwood A, Chapman M, Abu-Dahab K, Taylor B, Jackson B, Rey S, Amato R, et al. 2021. CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *Gen Biol*. 22(1):196.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35(3):526–528.

Parker J, Rambaut A, Pybus OG. 2008. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect Genet Evol*. 8(3):239–246.

Rambaut A, Holmes EC, O’Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. 5(11):1403–1407.

Salemi M, Lamers SL, Yu S, de Oliveira T, Fitch WM, McGrath MS. 2005. Phylodynamic analysis of human immunodeficiency virus type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. *J Virol*. 79(17):11343–11352.

Shu Y, McCauley J. 2017. GISAID: global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance* 22:30494.

Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, Hinrichs AS, Fernandes JD, Borges R, Slodkowitz G, et al. 2020. Stability of SARS-CoV-2 phylogenies. *PLoS Genet*. 16:e1009175.

Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, Haussler D, Corbett-Detig R. 2021. Ultrafast Sample placement on Existing tRees (USHER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat Genet*. 53(6):809–816.

Turakhia Y, Thornlow B, Hinrichs A, McBroome J, Ayala N, Ye C, Maio ND, Haussler D, Lanfear R, Corbett-Detig R. 2021. Pandemic-scale phylogenomics reveals elevated recombination rates in the SARS-CoV-2 spike region. Available from: <https://www.biorxiv.org/content/10.1101/2021.08.04.455157v1>

Wang TH, Donaldson YK, Brettell RP, Bell JE, Simmonds P. 2001. Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. *J Virol*. 75(23):11686–11699.