

## Phylogenetics

# Treenome Browser: co-visualization of enormous phylogenies and millions of genomes

Alexander M. Kramer <sup>1,2,\*</sup>, Theo Sanderson<sup>3</sup> and Russell Corbett-Detig<sup>1,2,\*</sup>

<sup>1</sup>Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064, USA, <sup>2</sup>Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA and <sup>3</sup>The Francis Crick Institute, London NW1 1AT, UK

\*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on September 27, 2022; revised on November 22, 2022; editorial decision on November 26, 2022; accepted on November 29, 2022

## Abstract

**Summary:** Treenome Browser is a web browser tool to interactively visualize millions of genomes alongside huge phylogenetic trees.

**Availability and implementation:** Treenome Browser for SARS-CoV-2 can be accessed at [cov2tree.org](https://cov2tree.org), or at [taxoniuum.org](https://taxoniuum.org) for user-provided trees. Source code and documentation are available at [github.com/theosanderson/taxoniuum](https://github.com/theosanderson/taxoniuum) and [docs.taxoniuum.org/en/latest/treenome.html](https://docs.taxoniuum.org/en/latest/treenome.html).

**Contact:** alex.kramer@ucsc.edu or rucorbet@ucsc.edu

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Huge genomic datasets are increasingly common and widely available. The ongoing coronavirus disease 2019 pandemic has particularly stressed existing bioinformatic infrastructure for the analysis of genomic data. We developed UShER (Turakhia *et al.*, 2021) to enable the inference of a continuously growing phylogeny containing millions of SARS-CoV-2 genomes (McBroome *et al.*, 2021) and Taxoniuum to enable interactive exploration of phylogenies of this size (Sanderson, 2022). Gingr (Treangen *et al.*, 2014) and the UCSC Genome Browser can display a multiple sequence alignment alongside a phylogeny (Lee *et al.*, 2022). However, no existing tool can simultaneously visualize million sample phylogenies with their underlying genome sequences.

We present Treenome Browser, a visualization tool for exploring genetic variation in millions of genomes alongside a huge phylogeny. Treenome Browser uses an innovative phylogenetic compression technique to interactively display the genome of each sample aligned with its phylogenetic position, remaining performant on trees with over 12 million sequences. The web application is available at [cov2tree.org](https://cov2tree.org) for the global SARS-CoV-2 tree and at [taxoniuum.org](https://taxoniuum.org) for user-provided trees.

## 2. Description

Treenome Browser displays mutations as vertical lines spanning the mutation's presence among samples in the phylogeny, drawn at their horizontal position in an associated reference genome (Fig. 1). Both amino acid and nucleotide mutations are supported. When the whole tree is visible, the genetic signatures of major clades in the tree become apparent. Users can search or manually zoom to

samples in the tree to inspect smaller clades and individual genomes. The genome alignment can also be navigated to analyze sites or regions of interest. Potential applications of Treenome Browser include variation-informed primer design, uncovering technical artifacts in tree construction or primary sequence assembly and investigating molecular evolutionary processes. We provide a special build of Treenome Browser for the global SARS-CoV-2 tree at [cov2tree.org](https://cov2tree.org). This site is preloaded with annotations for the reference genome (RefSeq NC\_045512v2) curated by the UCSC Genome Browser (Lee *et al.*, 2022). We also support user-provided mutation-annotated trees created with UShER, which can be produced using the TaxoniuumTools package (Sanderson 2022) and loaded at [taxoniuum.org](https://taxoniuum.org).

## 3. Implementation

The mutation-annotated tree data structure enables Treenome Browser to scale to huge trees. This data structure is used by UShER and Taxoniuum to encode genomic data in the form of a phylogeny with the inferred emergence of mutations annotated on each node. The genotype data encoded in a mutation-annotated tree is often many times smaller than the same data in VCF format (Supplementary Table S1). We use this phylogenetic compression to efficiently display mutations. Compared to an existing tool with similar functionality, Treenome Browser is much faster and scales to much larger phylogenomic datasets (Supplementary Table S2).

The core algorithm used by Treenome Browser decodes a mutation-annotated tree to compute the on-screen position of each mutation in the tree. To compute vertical positions, the vertical span of each subclade of the tree is first stored using dynamic programming. These pre-computed positions are then used along with

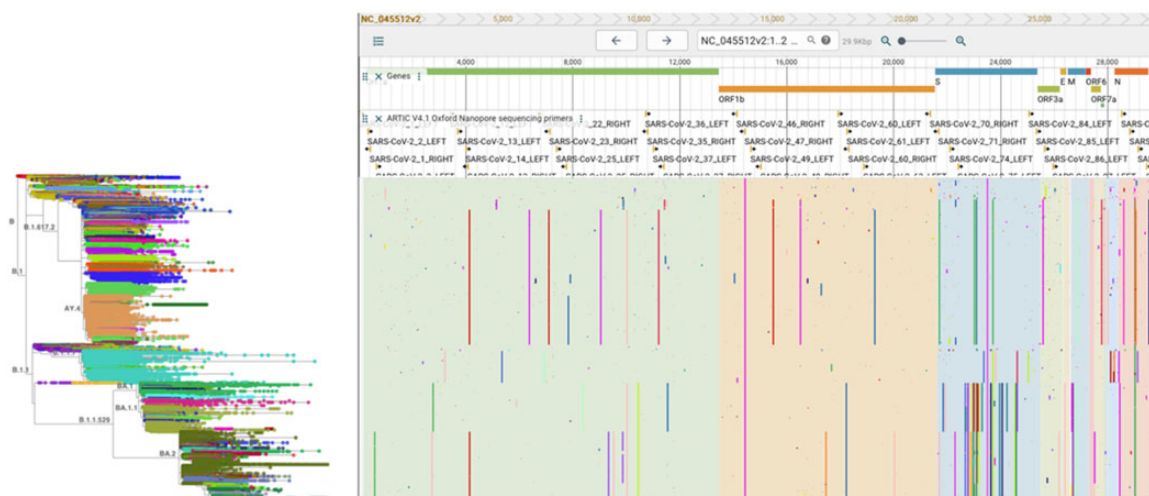


Fig. 1. Treenome Browser for the global SARS-CoV-2 tree with 5.9 million genomes, shown with the ARTIC Primers v4.1 annotation track (<https://community.artic.network/t/sars-cov-2-v4-1-update-for-omicron-variant/342>). Colored lines represent amino acid mutations relative to the reference genome

horizontal positions computed on-the-fly to draw each annotated mutation on screen as it is encountered in a traversal of the tree from root to leaves. We use the deck.gl framework (<https://github.com/visgl/deck.gl>) to render the mutations along the genome, which exploits WebGL to render large datasets much more quickly than traditional approaches for web visualization. This phylogenetically informed process displays on screen the complete set of amino acid and/or nucleotide substitutions in every leaf genome without explicitly enumerating the mutations in each genome, increasing rendering speed and reducing memory usage (Supplementary Table S3).

Treenome Browser further increases performance by subsampling nodes and mutations. Because many nodes occupy essentially the same spatial position when the tree is zoomed out, Taxonium dynamically displays sparsified trees from the full tree (Sanderson, 2022). Treenome Browser also uses these subsampled trees when displaying mutations to avoid unneeded computation. This tool further filters mutations by discarding those spanning very small clades at a given zoom level. These subsampling steps exclude imperceptible data based on zoom level while allowing users to view any individual node or mutation.

The reference genome and annotation tracks are displayed in an interactive JBrowse 2 (Buels et al., 2016) panel. Annotations are loaded into Treenome Browser automatically for the global SARS-CoV-2 tree at [cov2tree.org](https://cov2tree.org). We use the UCSC Genome Browser API to fetch each bigBed and bigWig annotation file curated by UCSC for SARS-CoV-2 (Fernandes et al., 2020). These annotations and user-provided annotation files are displayed by JBrowse 2. Treenome Browser is implemented in React as a feature addition to the Taxonium project. Source code is available at [https://github.com/theosanderson/taxonum/tree/master/taxonum\\_web\\_client](https://github.com/theosanderson/taxonum/tree/master/taxonum_web_client).

## Acknowledgements

The authors gratefully acknowledge the many developers of the open-source projects enabling the development of Treenome Browser and those who have generated and submitted the sequence data visualized by our tool. We also

thank Jakob McBroome, Angie Hinrichs and Yatish Turakhia for their feedback and helpful discussions.

## Funding

This work was supported by the Centers for Disease Control [BAA 200-2021-11554 to A.K. and R.C.D.]. T.S. is supported by the Wellcome Trust [210918/Z/18/Z] and the Francis Crick Institute which receives its core funding from Cancer Research UK [FC001043], the UK Medical Research [FC001043] Council and the Wellcome Trust [FC001043].

*Conflict of Interest:* none declared.

## Data availability

Files used in our supplemental experiments are available at [http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/USHER\\_SARS-CoV-2/2022/09/21/](http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/USHER_SARS-CoV-2/2022/09/21/). The latest Taxonium mutation-annotated tree is available at [https://cov2tree.nyc3.cdn.digitaloceanspaces.com/latest\\_public.json.gz](https://cov2tree.nyc3.cdn.digitaloceanspaces.com/latest_public.json.gz).

## References

- Lee, B.T. et al. (2022) The UCSC genome browser database: 2022 update. *Nucleic Acids Res.*, 50, D1115–D1122.
- McBroome, J. et al. (2021) A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. *Mol. Biol. Evol.*, 38, 5819–5824.
- Sanderson, T. (2022) Taxonium: a web-based tool for exploring large phylogenetic trees. *eLife*, 11, e82392.
- Treangen, T.J. et al. (2014) The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.*, 15, 524.
- Turakhia, Y. et al. (2021) Ultrafast sample placement on existing tRees (USHER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.*, 53, 809–816.