

UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Scalabe Phylo-Pangenomics

Permalink

<https://escholarship.org/uc/item/408649w0>

ISBN

9798190914948

Author

Kramer, Alexander

Publication Date

2026-08-25

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

SCALABLE PHYLO-PANGENOMICS

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY
in
BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by
Alexander M. Kramer

August 2026

The Dissertation of Alexander M. Kramer is
approved:

Professor Russell Corbett-Detig, chair

Professor David Haussler

Professor Benedict Paten

Donald Smith
Interim Vice Provost and Dean of Graduate Studies

Copyright © by
Alexander M. Kramer
2026

TABLE OF CONTENTS

Chapter 1. Introduction	1
1.1 The scale of genome sequencing	1
1.2 Phylogenetics in genomic epidemiology	1
1.3 The limits of <i>de novo</i> inference and the case for online phylogenetics	2
1.4 Mutation-annotated trees, the UShER ecosystem, and parsimony at scale . . .	3
1.5 Utilizing phylogenies of millions of genomes	3
1.6 Reference bias and the pangenome paradigm	4
1.7 Mutation-annotated networks: lossless, phylogeny-aware pangenomes	5
1.8 Aims and contributions	6
Chapter 2. Pandemic-Scale Tree Inference	7
2.1 Introduction	7
2.2 Background	9
2.3 Methods	14
2.3.1 Ground truth phylogeny construction	14
2.3.2 Simulated Data Preparation	16
2.3.3 Experimental Setup	17
2.3.4 Analyses Using Real SARS-CoV-2 Data Under Pandemic Time Constraints	19
2.3.5 Analyses Without Pandemic Time Constraints	19

2.3.6 Correlation of Parsimony and Likelihood on Large SARS-CoV-2 Phylogenies	20
2.4 Results and Discussion	21
2.4.1 Online Phylogenetics is an Alternative to <i>de novo</i> Phylogenetics for Ongoing Studies	21
2.4.2 Analyses Using Simulated Data Suggest that Online Phylogenetics is More Accurate for SARS-CoV-2	23
2.4.3 Analyses Using Real Data Suggest that Online Phylogenetics is More Efficient than <i>de novo</i> and Produces Similarly Optimal Phylogenies	26
2.4.4 Under Pandemic Time Constraints, matOptimize has Equivalent or More Favorable Metrics Compared to ML and Pseudo-ML Methods for SARS- CoV-2 Phylogenies	29
2.4.5 Parsimony Optimization with matOptimize Produces Comparable SARS-CoV-2 Trees to the Most Thorough ML Methods	30
2.4.6 Parsimony and Likelihood are Strongly Correlated When Optimizing Large SARS-CoV-2 Phylogenies	32
2.5 Conclusions	34
Chapter 3. Large-Scale Tree Analysis and Visualization	38
3.1 ShUSHER: privacy-preserving placement in the browser	38
3.1.1 Motivation: placement without data disclosure	38
3.1.2 Design and workflow	40
3.1.3 Privacy guarantee and outlook	41
3.2 Treenome Browser: co-visualization of phylogenies and genomes	42

3.2.1 Motivation	42
3.2.2 Description	42
3.2.3 Implementation	43
3.3 Applications	44
3.4 Discussion	45
Chapter 4. Phylo-pangenomic Analysis with Panmap	46
4.1 Introduction to Panmap	46
4.2 Results	48
4.2.1 Panmap overview	48
4.2.2 Phylogenetic compression yields indexes up to 600× smaller than alternatives	49
4.2.3 Panmap enables phylogenetic placement directly from sequencing reads	51
4.2.4 Panmap achieves high placement accuracy across coverage levels . . .	54
4.2.5 Phylogenetic placement enables scalable pangenome-guided read alignment	58
4.2.6 Panmap assembles higher quality genomes than single-reference methods	59
4.2.7 Panmap resolves haplotype mixtures across diverse viral species	62
4.2.8 Panmap outperforms existing tools on real wastewater surveillance data	63
4.2.9 Application to ancient DNA phylogenetic analysis	66
4.3 Discussion	69

4.4 Methods	73
4.4.1 Overview of Methods	73
4.4.2 Reference PanMAN structure and genome decoding	73
4.4.3 Seed sketches of genome sequences	74
4.4.4 Indexing k-mer seeds	75
4.4.5 Indexing optimizations	75
4.4.6 Alignment-free phylogenetic placement	76
4.4.7 Mutation spectrum prior for genotype likelihoods	76
4.4.8 Genotype likelihoods and consensus assembly	77
4.4.9 Phylogenetic placement in Single-sample mode	77
4.4.10 Read alignment in Single-sample mode	78
4.4.11 Read scoring in Metagenomic mode	79
4.4.12 Abundance estimation in Metagenomic mode	80
4.4.13 Experimental setup and data preparation	81
4.4.14 Panmap index evaluation	82
4.4.15 Phylogenetic placement evaluation, Single-sample mode	84
4.4.16 Haplotype abundance evaluation with simulated data	85
4.4.17 Genome assembly evaluation with real data	85
4.4.18 SARS-CoV-2 lineage abundance estimation from wastewater data ..	87
4.4.19 Ancient environmental DNA placement	88
4.4.20 Panmap-assigned reads validation from Wang et al. dataset	89
4.4.21 Per-sample placement on <i>Mammuthus</i> PanMAN	90
Chapter 5. Conclusion	91

Appendix A. Supplementary Figures and Table for Chapter 2	93
Appendix B. Supplementary Figures for Chapter 4	98
Appendix C. Supplementary Tables for Chapter 4	109
Bibliography	110

LIST OF FIGURES

Figure 2.1. De novo versus online phylogenetics	22
Figure 2.2. Accuracy of online optimization on simulated data	25
Figure 2.3. Parsimony versus ML and pseudo-ML optimization on real data	27
Figure 2.4. De novo matOptimize compared with the most thorough ML inference	32
Figure 2.5. Linear relationship between likelihood and parsimony improvement .	34
Figure 3.1. The ShUSHER web interface	41
Figure 3.2. Treenome Browser for the global SARS-CoV-2 tree	43
Figure 4.1. Overview of Panmap’s indexing and analysis modes	51
Figure 4.2. Phylogenetic placement accuracy and performance of Panmap	57
Figure 4.3. Consensus assembly accuracy versus single-reference workflows	61
Figure 4.4. Abundance estimation in simulated and wastewater mixtures	65
Figure 4.5. Reads assigned to nodes in two ancient DNA datasets	69
Figure S2.1. Temporal patterns in simulated SARS-CoV-2 data may affect Robinson-Foulds (RF) distance normalization	95
Figure S2.2. Most branches in the ground truth phylogeny are extremely short ...	96
Figure S2.3. Resolution of the ground truth phylogeny	97
Figure S4.1. Multithreading scalability of Panmap	99
Figure S4.2. Placement accuracy versus read length	100

Figure S4.3. Computational resources for placing samples onto a PanMAN with 8 million genomes	100
Figure S4.4. Evaluation of haplotype abundance estimation in simulated HIV (A) and RSV (B) mixture samples	101
Figure S4.5. Comparison of haplotype abundance estimation accuracy between Panmap and WEPP	102
Figure S4.6. Comparison of SARS-CoV-2 lineage abundance estimation between Panmap, WEPP, and Freyja	103
Figure S4.7. Wastewater estimated and clinically derived SARS-CoV-2 lineage proportions	104
Figure S4.7 (continued). Wastewater lineage proportions (continued)	105
Figure S4.8. Damage plots of Panmap-assigned ancient DNA reads	106
Figure S4.9. <i>Mammuthus</i> mitochondrial phylogeny	107
Figure S4.10. Panmap single-sample scoring metric evaluation	108

LIST OF TABLES

Table 4.1. Pangenome index size, build time, and memory across indexing tools .	53
Table S2.1. Replicate comparison of parsimony and likelihood scores	94

Abstract

Scalable phylo-pangenomics

Alexander M. Kramer

The COVID-19 pandemic generated genomic data at unprecedented scale, with tens of millions of SARS-CoV-2 genomes deposited in public repositories and thousands of new sequences added each day. This dissertation develops methods for analyzing genomic datasets at this scale, unified by the idea that encoding genomes according to their evolutionary relationships can make otherwise intractable computations practical. First, I evaluate online phylogenetic inference, in which new genomes are continuously added to an existing tree, and compare parsimony-based methods with maximum-likelihood approaches under pandemic time constraints. For densely sampled SARS-CoV-2 genomes, online inference with UShER and matOptimize produces trees comparable to established maximum-likelihood methods while requiring orders of magnitude less time and memory. I then develop tools that make phylogenies containing millions of genomes useful for downstream analysis and visualization. ShUShER enables privacy-preserving phylogenetic placement within a web browser, allowing laboratories to analyze sensitive sequences without transmitting them to an external server. Treenome Browser co-visualizes the genomic variation of millions of samples alongside their phylogenetic relationships by operating directly on a compressed mutation-annotated tree. Finally, I describe Panmap, which uses Pangenome Mutation-Annotated Networks (PanMANs) to place, align, and genotype sequencing reads and to estimate haplotype abundances against reference collections containing up to millions of genomes. Panmap produces indexes hundreds of times smaller

than graph-based alternatives, improves genome reconstruction over single-reference workflows at low coverage, and supports applications ranging from pathogen genome assembly to ancient environmental DNA analysis. Together, these results show that evolutionary history can serve not only as an object of inference but as a scalable computational infrastructure for phylogenetic and pangenomic analyses as genomic datasets continue to grow.

ACKNOWLEDGMENTS

I'm very grateful for the numerous mentors, colleagues, family, and friends that have supported me throughout my PhD. Thank you to my advisor, Russ, and my committee members for their guidance and wisdom. Thank you to all the researchers and public health workers who collected, sequenced, and shared the many millions of genomes that make methods like these possible.

The text of this dissertation includes reprints of the following previously published material:

- Kramer, A. M., Thornlow, B., Ye, C., De Maio, N., McBroome, J., Hinrichs, A. S., Lanfear, R., Turakhia, Y., & Corbett-Detig, R. (2023). Online phylogenetics with matOptimize produces equivalent trees and is dramatically more efficient for large SARS-CoV-2 phylogenies than *de novo* and maximum-likelihood implementations. *Systematic Biology*, 72(5), 1039-1051.
- Kramer, A. M., Sanderson, T., & Corbett-Detig, R. (2023). Treenome Browser: co-visualization of enormous phylogenies and millions of genomes. *Bioinformatics*, 39(1), btac772.
- Kramer, A. M., Turakhia, Y., & Corbett-Detig, R. (2021). ShUSHER: private browser-based placement of sensitive genome samples on phylogenetic trees. *Journal of Open Source Software*, 6(66), 3677.

- Kramer, A. M., Zhang, A., Ayala, N., de Sanctis, B., Karim, L. M., Hinrichs, A. S., Walia, S., Turakhia, Y., & Corbett-Detig, R. (2026). Panmap: Scalable phylogeny-guided alignment, genotyping, and placement on pangenomes. *bioRxiv*.

Professor Russell Corbett-Detig directed and supervised the research which forms the basis for the dissertation. Co-authors have given permission to include the above material.

My contributions to each reprinted publication were as follows. For the online phylogenetics and matOptimize study reproduced in Chapter 2, I share first authorship with Bryan Thornlow; I implemented and carried out benchmarking experiments, analyzed the results, and co-wrote the manuscript. For the Treenome Browser and ShUSHER tools reproduced in Chapter 3, I designed and implemented both tools, designed and performed the analyses, and wrote both manuscripts. For the Panmap study reproduced in Chapter 4, I share first authorship with Alan Zhang; I contributed to the design of the method, implemented single-sample mode of the software, performed experiments and analyses, and co-wrote the manuscript.

CHAPTER 1. INTRODUCTION

1.1 The scale of genome sequencing

Whole genome sequencing is now a routine tool in infectious disease surveillance, and the COVID-19 pandemic dramatically changed its scale. Over the course of the pandemic, laboratories worldwide deposited tens of millions of SARS-CoV-2 genomes in public repositories, adding thousands of new sequences each day (McBroome et al., 2021). SARS-CoV-2 became the most densely sequenced organism on record, and the number of genomes available quickly outgrew existing computational phylogenetic and genomic methods. This dissertation develops methods for analyzing genomic data at such scales.

Pandemic-scale data create demands that ordinary comparative genomics does not. Sequences arrive continuously rather than in one batch, analyses must be updated within days to stay useful for public health, and the sheer number of sequences rules out many existing methods due to excessive time or memory cost. These constraints force a reconsideration of representations and algorithms for phylogenomics at scale.

1.2 Phylogenetics in genomic epidemiology

Phylogenetic inference reconstructs the evolutionary relationships among sampled genomes, and it underlies much of genomic epidemiology. Comprehensive pathogen phylogenies support genomic surveillance, the designation and tracking of

variants and lineages (Rambaut et al., 2020), the reconstruction of transmission links, the separation of community transmission from independent introductions, and the detection of recombination, convergent evolution, and rate variation. Subsampling can hide the relationships an analysis is meant to find, so phylogenies should ideally contain every available sequence. If a complete and current tree of all sampled genomes can be maintained, a large class of public-health questions becomes answerable.

1.3 The limits of *de novo* inference and the case for online phylogenetics

Standard phylogenetic inference is *de novo*: all sequence data are gathered first, and a tree is inferred once from scratch. This fits the historical norm in which a dataset is collected for one study and rarely grows afterward, and most phylogenetics software is built accordingly (Price et al., 2010; Minh et al., 2020; Kozlov et al., 2019). In settings where sequences accumulate every day, such as during a pandemic, and a current tree of millions of genomes is needed daily, inferring such a tree from scratch each day is too costly.

Online phylogenetics offers an alternative (Gill et al., 2020). New sequences are continually placed onto an existing tree, and the augmented tree may be optimized periodically rather than rebuilt. Placing query sequences onto a fixed reference tree has a long methodological history (Matsen et al., 2010; Berger et al., 2011; Barbera et al., 2019), but the pandemic posed a new problem: maintaining a single, continuously growing phylogeny of all sampled genomes rather than placing a few queries onto a stable backbone. Online phylogenetics particularly suits densely sampled pathogen data, where a new genome is usually a close relative of one already in the tree.

1.4 Mutation-annotated trees, the UShER ecosystem, and parsimony at scale

Online phylogenetics of SARS-CoV-2 is enabled by phylogenetic compression via mutation-annotated trees (MATs). A MAT is a phylogeny whose branches carry the mutations inferred to have occurred along them, so that any sampled genome's variants can be recovered by accumulating mutations from the root. Closely related genomes differ at few sites, so a MAT is much smaller than the equivalent alignment or variant-call file (VCF). The MAT underlies UShER, which adds samples to a tree by maximum parsimony (Turakhia et al., 2021); the matUtils toolkit and the daily-updated global SARS-CoV-2 MAT database (McBroome et al., 2021); the Taxonium tree-exploration tool (Sanderson, 2022); and matOptimize, which optimizes a MAT under maximum parsimony (Ye et al., 2022).

Sampling density also informs the choice of optimality criterion. Maximum likelihood is usually preferred over maximum parsimony because it can account for multiple substitutions at one site along one branch and offers a more biologically relevant model of the data. That advantage shrinks as branches get shorter, and in a densely sampled phylogeny (e.g., the global SARS-CoV-2 tree) branches are typically extremely short. This suggests that parsimony may be accurate enough for pandemic-scale inference while running far faster and using less memory than likelihood and pseudo-likelihood implementations (Minh et al., 2020; Kozlov et al., 2019; Price et al., 2010). Validating this hypothesis, and defining when parsimony-based online inference can stand in for likelihood-based *de novo* inference, is the subject of Chapter 2.

1.5 Utilizing phylogenies of millions of genomes

A phylogenetic tree is only useful to the extent researchers can analyze it, and trees of millions of genomes strain previously available tools for placement and visualization. While Chapter 2 describes the inference and validation of huge trees, Chapter 3 addresses the practical usage of these trees for downstream analysis. First, public health laboratories often hold sequences they cannot send to an outside server for privacy or governance reasons, which rules out server-based placement. Second, existing tools could display a large phylogeny (Sanderson, 2022) or show a genome alignment next to a reference annotation (Lee et al., 2022), but none could display the complete genetic variation in million-genome datasets alongside the tree relating them. Chapter 3 describes two web browser tools that address these gaps: ShUSHER for privacy-preserving sample placement, and Treenome Browser for co-visualization of genomes and phylogeny at scale.

1.6 Reference bias and the pangenome paradigm

The methods discussed in Chapters 2 and 3, like most genomic analyses, describe each genome relative to a single linear reference. That convention introduces reference bias, where variation missing from or poorly represented by the reference is under-ascertained. Pangenome methods reduce this bias by utilizing a collection of genomes, whether a species, population, or clade, instead of a single reference (Computational Pan-Genomics Consortium, 2018). The common approach represents a pangenome as a sequence graph, where nodes are sequence segments, edges join segments adjacent in one or more genomes, and each genome is a path through the graph. Tools such as vg (Garrison et al., 2018), Minigraph (Li et al., 2020), and Progressive Cactus (Armstrong et al., 2020) build and index such graphs, and Mini-

graph-Cactus (Hickey et al., 2024) was used to construct the draft human pangenome reference (Liao et al., 2023).

A defining feature of practical pangenome representations is losslessness, meaning every input genome can be reconstructed exactly. Graph formats provide lossless encoding, but graph construction is expensive, and these methods have so far been applied to pangenomes with tens or hundreds of genomes, not the millions available for densely sequenced pathogens (Liao et al., 2023; Hickey et al., 2024). Existing pangenome formats also focus on representing variation and largely ignore the evolutionary relationships and mutational histories among the genomes they contain (Walia et al., 2026).

An UShER-style mutation-annotated tree, though built from pangenome-scale data, is not a pangenome *per se*. It is anchored to a single linear reference and is lossy because it only utilizes single nucleotide variants (SNVs). For a low-diversity, low-indel pathogen such as SARS-CoV-2 this loss is negligible for many analyses, but it discards the indels and structural variation that a lossless pangenome retains.

1.7 Mutation-annotated networks: lossless, phylogeny-aware pangenomes

The Pangenome Mutation-Annotated Network (PanMAN) (Walia et al., 2026) is an alternative representation that, in contrast to sequence graph-based approaches, organizes and compresses a pangenome according to its evolutionary relationships. Its constituent tree(s) annotate substitutions together with inferred indels and structural mutations, and multiple such trees connect into a network through edges that can capture recombination and horizontal gene transfer. The result is lossless, since every input genome can be recovered exactly, and it stays compact through the same

evolutionary compression that makes the UShER MAT efficient. A SARS-CoV-2 pangenome of about eight million sequences fits in a few hundred megabytes (Walia et al., 2026). Chapter 4 presents Panmap, a tool that operates on PanMANs to map reads and assemble genomes using their evolutionary structure.

1.8 Aims and contributions

Chapter 2 concerns pandemic-scale phylogenetic inference. Using simulated and real SARS-CoV-2 data, it compares online with *de novo* inference and parsimony with likelihood-based optimization under pandemic time constraints. Online, parsimony-based inference with UShER and matOptimize produces equivalent or more optimal trees than established likelihood implementations while using orders of magnitude less time and memory.

Chapter 3 concerns the analysis and visualization of large phylogenies and presents two open-source tools. ShUShER runs UShER inside a web browser, so users can place sensitive sequences without transmitting data. Treenome Browser shows the genomes of millions of samples alongside the tree relating them, using the compression of the mutation-annotated tree and hardware-accelerated rendering to stay responsive on trees of more than twelve million sequences.

Chapter 4 concerns scalable phylo-pangenomics. It generalizes from reference-based, lossy mutation-annotated trees to lossless mutation-annotated pangenomes, and presents Panmap, which uses the evolutionary structure in a PanMAN to place, align, and genotype raw sequencing reads against pangenomes of up to millions of genomes, and to estimate metagenomic abundances.

CHAPTER 2. PANDEMIC-SCALE TREE INFERENCE

2.1 Introduction

USHER and matOptimize work together to infer trees at pandemic scale. USHER places new samples on a tree by parsimony (Turakhia et al., 2021), and matOptimize optimizes a tree by parsimony (Ye et al., 2022). This chapter asks whether trees built this way are accurate enough to use in place of maximum likelihood, specifically for large SARS-CoV-2 phylogenies. It compares online parsimony inference against maximum-likelihood and *de novo* methods on SARS-CoV-2, using simulated data and real genomes. Parsimony matches or exceeds the likelihood tools on tree accuracy while running thousands of times faster and using far less memory. A primary reason for this are that parsimony scores and likelihoods are highly correlated when branches are short, which is the case under dense sampling. The rest of this chapter is a reproduced form of the following publication:

Kramer, A. M., Thornlow, B., Ye, C., et al. (2023). Online phylogenetics with matOptimize produces equivalent trees and is dramatically more efficient for large SARS-CoV-2 phylogenies than *de novo* and maximum-likelihood implementations. *Systematic Biology* 72(5):1039–1051.

Phylogenetics has been foundational to SARS-CoV-2 research and public health policy, assisting in genomic surveillance, contact tracing, and assessing emergence and spread of new variants. However, phylogenetic analyses of SARS-CoV-2

have often relied on tools designed for *de novo* phylogenetic inference, in which all data are collected before any analysis is performed and the phylogeny is inferred once from scratch. SARS-CoV-2 data sets do not fit this mold. There are currently over 14 million sequenced SARS-CoV-2 genomes in online databases, with tens of thousands of new genomes added every day. Continuous data collection, combined with the public health relevance of SARS-CoV-2, invites an “online” approach to phylogenetics, in which new samples are added to existing phylogenetic trees every day. The extremely dense sampling of SARS-CoV-2 genomes also invites a comparison between likelihood and parsimony approaches to phylogenetic inference. Maximum likelihood (ML) and pseudo-ML methods may be more accurate when there are multiple changes at a single site on a single branch, but this accuracy comes at a large computational cost, and the dense sampling of SARS-CoV-2 genomes means that these instances will be extremely rare because each internal branch is expected to be extremely short. Therefore, it may be that approaches based on maximum parsimony (MP) are sufficiently accurate for reconstructing phylogenies of SARS-CoV-2, and their simplicity means that they can be applied to much larger data sets. Here, we evaluate the performance of *de novo* and online phylogenetic approaches, as well as ML, pseudo-ML, and MP frameworks for inferring large and dense SARS-CoV-2 phylogenies. Overall, we find that online phylogenetics produces similar phylogenetic trees to *de novo* analyses for SARS-CoV-2, and that MP optimization with UShER and matOptimize produces equivalent SARS-CoV-2 phylogenies to some of the most popular ML and pseudo-ML inference tools. MP optimization with UShER and matOptimize is thousands of times faster than presently available implementations of

ML and online phylogenetics is faster than *de novo* inference. Our results therefore suggest that parsimony-based methods like UShER and matOptimize represent an accurate and more practical alternative to established ML implementations for large SARS-CoV-2 phylogenies and could be successfully applied to other similar data sets with particularly dense sampling and short branch lengths.

2.2 Background

The widespread availability and extreme abundance of pathogen genome sequencing has made phylogenetics central to combatting the COVID-19 pandemic. Communities worldwide have implemented genomic surveillance by systematically sequencing the genomes of a percentage of local cases (Deng et al. 2020; Lu et al. 2020a; Meredith et al. 2020; Park et al. 2021). This has been important in tracing local transmission chains (Bluhm et al. 2020; Lam 2020), understanding the genetic makeup of viral populations within local communities (Gonzalez-Reiche et al. 2020; Franceschi et al. 2021; Thornlow et al. 2021a), uncovering the means by which viral lineages have been introduced to new areas (Castillo et al. 2020), and measuring the relative spread of specific variants (Skidmore et al. 2021; Umair et al. 2021). Phylogenetic approaches for better understanding the proximate evolutionary origins of the virus (Li et al. 2020), as well as to identify recombination events (Jackson et al. 2021; Turakhia et al. 2021b) and instances of convergent evolution (Kalantar et al. 2020; Peng et al. 2021) have greatly informed our understanding of the virus. Phylogenetic visualization software including Auspice (Hadfield et al. 2018) and Taxonium (Sanderson 2022) have also become widely used for public health purposes. All of these applications require a phylogeny.

A comprehensive, up-to-date phylogenetic tree of SARS-CoV-2 is important for public health officials and researchers. A tree containing all available sequences can sometimes facilitate identification of epidemiological links between samples that might otherwise be lost in subsampled phylogenies. Conversely, these approaches can often rule out otherwise plausible transmission histories. Such information can also help to identify the likely sources of new viral strains in a given area (Moreno et al. 2020; Tang et al. 2021). Additionally, using up-to-date information enables us to find and track novel variants of concern and clades which are growing quickly (Annava-jhala et al. 2021; Tegally et al. 2021), and to measure the spread of known variants at both local and global scales. Furthermore, comprehensive phylogenies can improve the ability of computational methods to find recombinant viral genomes (Turakhia et al. 2021b), natural selection at homoplasious positions (van Dorp et al. 2020), variation in mutation rates (De Maio et al. 2021), and systematic recurrent errors (Turakhia et al. 2020). This also facilitates naming lineages of interest, which has been particularly important in tracking variants of concern throughout the pandemic (e.g., B.1.1.7 or “Alpha” and B.1.617.2 or “Delta”) (Rambaut et al. 2020).

SARS-CoV-2 presents a unique set of challenges for phylogenetic analyses. The unprecedented pace and scale of whole-genome sequence data has forced the phylogenetics community to place runtime and scalability at the center of every inference strategy. More than 14 million SARS-CoV-2 genome sequences are currently available, with tens of thousands being added each day. Prior to the pandemic, *de novo* phylogenetics, that is, inferring a phylogeny from scratch, has been the standard, as there has rarely been a need to re-infer or improve pre-existing phylogenies on a

daily basis. Daily inference of a tree of millions of samples from scratch, however, is extremely costly, and has brought a renewed focus on methods for adding new samples to existing phylogenetic trees (Matsen et al. 2010; Berger et al. 2011; Izquierdo-Carrasco et al. 2014; Fourment et al. 2018; Barbera et al. 2019). This approach has been called “online phylogenetics” (Gill et al. 2020), and has important advantages in the context of the pandemic and beyond. Online phylogenetics is appealing for the genomic surveillance of pathogens, because iterative optimization should decrease computational expense, allowing good estimates of phylogenies to be made readily available. It may be particularly effective in a pandemic setting, where new samples are closely related to existing samples in the tree, and existing samples are often identical to inferred ancestral states. That is, the ancestor of a newly sequenced genome has often effectively already been observed. The methods described here may be less useful for reconstructing deeper evolutionary histories of other large genomic data sets, for example organisms like insects or certain pathogens like HIV-1.

Second, SARS-CoV-2 genomes are much more closely related than sequences in most other phylogenetic analyses. Because the relative advantages of maximum likelihood (ML) methods decrease for closely related samples and long branches are relatively rare in the densely sampled SARS-CoV-2 phylogeny (Felsenstein 1978; Hendy and Penny 1989; Philippe et al. 2005), this suggests that phylogenetic inferences based on maximum parsimony, a much faster and simpler phylogenetic inference method, could be better suited for online phylogenetic analyses of SARS-CoV-2 genomes (Wertheim et al. 2022). The principle of maximum parsimony is that the tree with the fewest mutations should be favored, and it is sometimes

described as a non-parametric phylogenetic inference method (Sullivan and Swofford 2001; Kolaczkowski and Thornton 2004). Additionally, because parsimony-based tree optimization does not require estimation of ancestral character state uncertainty at all positions in the phylogeny like ML optimization does, parsimony uses much less memory.

Here, we evaluate approaches that would enable one to maintain a fully up-to-date and comprehensive global phylogeny of SARS-CoV-2 genome sequences (McBroome et al. 2021). Specifically, we investigate tradeoffs between online and *de novo* phylogenetics and between maximum parsimony, ML, and pseudo-ML approaches when the aim is for an analysis to scale to millions of sequences, with tens of thousands of new sequences being added daily. We chose to compare maximum parsimony, ML, and pseudo-ML (and omit other approaches like neighbor-joining) because they were the most effective methods at inferring large SARS-CoV-2 phylogenies based on previous analyses (Lanfear and Mansfield 2020), and because most distance-based methods are quadratic in memory usage so cannot scale to estimating trees from data sets of more than a few hundred thousand sequences (Wang et al. 2022). We mimic the time-course of the pandemic by introducing increasingly large numbers of SARS-CoV-2 genome sequences proportionately to their reported sampling dates.

We evaluate potential online phylogenetics approaches by iteratively adding samples to existing trees and optimizing the augmented phylogeny with different tools that have been proposed for this purpose during the pandemic. In particular, we compare online matOptimize, IQ-TREE 2, RAxML-NG, and FastTree 2. Between each optimization step, we use USHER (Turakhia et al. 2021a) to add samples to trees

by maximum parsimony. matOptimize is a parsimony optimization approach that uses subtree pruning and regrafting (SPR) moves to minimize the total mutations in the final tree topology (Ye et al. 2022). IQ-TREE 2 uses nearest neighbor interchange (NNI) and optionally stochastic tree moves to find the tree with the highest likelihood given an input multiple sequence alignment (Minh et al. 2020). RAxML-NG is a ML approach that uses SPR moves to search tree-space for higher likelihood phylogenies (Kozlov et al. 2019). FastTree 2 uses a pseudo-likelihood approach that employs minimum-evolution SPR and/or NNI moves and ML NNI moves while using several heuristics to reduce the search space (Price et al. 2010). The likelihood-based approaches evaluated here report branch lengths in substitutions per site. Parsimony-based matOptimize reports branch lengths in total substitutions, which can be converted to the latter by dividing by the alignment length. Reporting unambiguous branch lengths in this way is possible because UShER and matOptimize resolve potential ambiguities in ancestral sequences by parsimony. When there are ambiguous ancestral states, both matOptimize and UShER will prefer to assign new mutations at the node farthest from the root. This is similar to DELTRAN (Agnarsson and Miller 2008; Swofford and Maddison 1987). Additionally, when UShER places new samples with ambiguous nucleotides in the consensus sequence, it resolves each to the most parsimonious state given the guide tree. If there is more than one equally parsimonious nucleotide, UShER will select the reference nucleotide if it is one of the possibilities. This decision is motivated by the observation that infrequent ambiguous states arise for a variety of technical reasons and typically the reference nucleotide will be correct (De Maio et al. 2020; Turakhia et al. 2020). The branch lengths reported by UShER and matOptimize may be interpreted

as is or used as an initial estimate for other distance measures, for example in the construction of time trees (Sanderson 2021).

Results from our comparisons demonstrate that for the purposes of SARS-CoV-2 phylogenetics, in which samples are numerous and closely related and inference speed is of high significance, parsimony-based online phylogenetics applications are clearly most favorable and are also the only immediately available methods capable of producing daily phylogenetic estimates of all available SARS-CoV-2 genomes (Turakhia et al. 2021a). We note that matOptimize is used to maintain such a phylogeny comprising over 14 million genomes as of February 2023 (McBroome et al. 2021). As global genomic data collection accelerates further, we expect parsimony-based online approaches to become increasingly central to phylogenetic inference for data sets with similar properties to SARS-CoV-2, for example Mpox, RSV, and *Mycobacterium tuberculosis*.

USHER and matOptimize are available through Anaconda at <https://anaconda.org/bioconda/usher> and on Github at <https://github.com/yatisht/usher>. Our analyses use Github commit 66ca5ff which corresponds to the nearest conda version of 0.5.0.

2.3 Methods

2.3.1 Ground truth phylogeny construction

We first developed a global phylogeny, hereafter the “starting tree,” which we used in subsequent analyses. We began by downloading VCF and FASTA files corresponding to 18 March 2021 from our own daily updated database (McBroome et al. 2021). The VCF file contains pairwise alignments of each of the 434,063

samples to the SARS-CoV-2 reference genome. We then implemented filters, retaining only sequences containing at least 28,000 non-N nucleotides, and fewer than 2 non-[ACGTN-] characters. We used UShER to create a phylogeny from scratch using only the remaining 366,492 samples. To remove potentially erroneous sequences, we iteratively pruned this tree of highly divergent internal branches with branch parsimony scores greater than 30, then terminal branches with branch parsimony scores greater than 6, until convergence, resulting in a final global phylogeny containing 364,427 samples. The branch parsimony score indicates the total number of substitutions along a branch. Similar filters based on sequence divergence are used by existing SARS-CoV-2 phylogenetic inference methods. For full reproducibility, files used for creating the global phylogeny can be found in subrepository 1 on the project GitHub page (Thornlow et al. 2021b).

Following this, we created a “ground truth” tree by optimizing the starting tree to compare against the results of inference methods on simulated data. We used matOptimize, FastTree 2, and maximum parsimony (MP) IQ-TREE 2. MP IQ-TREE 2 uses parsimony as the optimality criterion in contrast to the ML mode used in other experiments, which was infeasible on a data set of this size. In these optimization experiments, we used experimental versions of MP IQ-TREE 2 that allow finer control of parsimony parameters (specific versions are listed in the supplemental Github repository). In one of our Ground Truth Optimization Experiments, we used the starting tree and its corresponding alignment and ran 5 iterations of MP IQ-TREE 2, varying the SPR radius from 20 to 100 in increments of 20. Experiments on a small data set indicated that there is little or no improvement in parsimony

score beyond a radius of 100. Separately, we tested another strategy that applied 2 iterations of MP IQ-TREE 2 to the starting tree, the 1st iteration using an SPR radius of 20 and the 2nd using a radius of 100. Finally, we tested a strategy of 6 iterations of pseudo-likelihood optimization with FastTree 2 followed by 2 iterations of parsimony optimization with matOptimize. The tree produced by this strategy, hereafter the “ground truth” tree, had the highest likelihood of all the strategies we tested. This tree (after_usher_optimized_fasttree_iter6.tree) and files for these optimization experiments can be found in subrepository 2.

In the multifurcating ground truth tree of 364,427 samples, there are 265,289 unique (in FASTA sequence) samples. There are 447,643 nodes in the tree. For reference, a full binary tree with the same number of leaves has 728,853 nodes. 23,437 of the 29,903 sites in the alignment are polymorphic (they display at least 2 non-ambiguous nucleotides). Out of 83,216 inferred ancestral nodes, 68,261 (82%) have as a child a sampled node with identical genotype (branch length zero). Homoplasies are common in these data. In the starting tree, 19,090 sites display a mutation occurring on at least 2 different branches, and 4976 sites display a mutation occurring more than 10 times in the tree. Approximately half of the internal nodes are resolved, having exactly 2 nodes as immediate descendants (Supplementary Fig. S2.3).

2.3.2 Simulated Data Preparation

To generate our simulated data, we used the SARS-CoV-2 reference genome (GISAID ID: EPI_ISL_402125; GenBank ID: MN908947.3) (Shu and McCauley 2017; Sayers et al. 2021) as the root sequence and used phastSim (De Maio et al. 2022) to simulate according to the ground truth phylogeny described above. Intergenic

regions were evolved using phastSim using the default neutral mutation rates estimated in ref. (De Maio et al. 2021), with position-specific mean mutation rates sampled from a gamma distribution with $\alpha = \beta = 4$, and with 1% of the genome having a 10-fold increase mutation rate for one specific mutation type (SARS-CoV-2 hypermutability model described in ref. [De Maio et al. 2022]). Evolution of coding regions was simulated with the same neutral mutational distribution, with a mean nonsynonymous/synonymous rate ratio of $\omega = 0.48$ as estimated in (Turakhia et al. 2021a), with codon-specific ω values sampled from a gamma distribution with $\alpha = 0.96$ and $\beta = 2$. Rates for each intergenic and coding region were not normalized in order to have the same baseline neutral mutation rate distribution across the genome.

2.3.3 Experimental Setup

We tested *de novo* and online matOptimize, ML IQ-TREE 2, and FastTree2 on this simulated alignment. The Simulated Online Experiments began by using UShER to infer a small tree *de novo* from the 1st batch of samples, followed by alternating steps of optimization using 1 of the 3 evaluated methods and placement of additional samples with UShER. For our Simulated De Novo Experiments, we supplied each software package with an alignment corresponding to all samples in that batch and its predecessors (or VCF for matOptimize) without a guide tree. For both cases, each tree is larger than its predecessor by ~ 5000 samples, and each tree necessarily contains all samples in the immediately preceding tree. We terminated experiments that took longer than 24 h to complete to mimic the time constraints of continual inference during a pandemic. For FastTree 2, we used 2 rounds of minimum-evolution subtree-prune-regraft (SPR) moves (-spr 2), maximum SPR length of 1000 (-sprlength 1000), zero

rounds of minimum-evolution NNIs (-nni 0), default settings for ML nearest neighbor interchanges, and the Generalised Time Reversible+Gamma (GTR+G) substitution model (-gtr -gamma). Previous analyses with FastTree 2 on large SARS-CoV-2 data sets show that an SPR radius of 1000 has negligible effect on execution time but can produce more optimal phylogenies than the default setting of 10 (Lanfear and Mansfield 2020). For IQ-TREE 2, we used a branch length minimum of 0.000000001 (-blmin 1e-9), zero rounds of stochastic tree search (-n 1), and the GTR+G substitution model (-m GTR+G). With these parameters, IQ-TREE 2 constructs a starting parsimony tree and then performs hill-climbing NNI steps to optimize likelihood, avoiding the significant time overhead of stochastic search. We ran matOptimize and UShER with default parameters. For each iteration of our Simulated De Novo Experiment with UShER+matOptimize, we started with an empty tree, added all samples with UShER, then optimized with matOptimize. We ran all matOptimize analyses using an instance with 15 CPUs and 117.2 GB of RAM, and we ran all IQ-TREE 2 and FastTree 2 analyses on an instance with 31 CPUs and 244.1 GB of RAM, but we limited each command to 15 threads for equivalence with matOptimize.

To assess the effectiveness of each method, we computed the Robinson–Foulds (RF) distance (Robinson and Foulds 1981) and quartet similarity of each optimization to the ground truth tree, pruned to contain only the samples belonging to that batch. To calculate each RF distance, we used the -O (collapse tree) argument in matUtils extract (McBroome et al. 2021) and then used the dist.topo command in the ape package in R (Paradis and Schliep 2019), comparing the collapsed optimized tree and the pruned, collapsed ground truth tree at each iteration. We computed normalized RF distances

by dividing each distance by the expression $I(T_i) + N_i - 3$. The value $I(T_i)$ is the number of internal edges in the pruned ground truth tree for iteration i , and N_i is the number of taxa, with $N_i - 3$ representing the maximum possible number of internal edges in the inferred tree (Steel and Penny 1993). We computed normalized quartet similarities to the ground truth tree using the metric described in (Asher and Smith 2022) which does not penalize excess resolution in the inferred tree. Files for all simulated data experiments can be found in subrepository 4.

2.3.4 Analyses Using Real SARS-CoV-2 Data Under Pandemic Time Constraints

To mimic pandemic-style phylogenetics, we separated a total of 233,326 real SARS-CoV-2 samples from the starting tree of 364,427 samples into 50 batches of ~5000 by sorting according to the date of sample collection. We then set up 2 experiments for each of the 3 software packages (matOptimize [conda version 0.5.0], ML IQ-TREE 2 [multicore version 2.1.3 COVID-edition], and FastTree 2 [Double Precision version 2.1.10]). We repeated our iterative experiments using online and *de novo* matOptimize, IQ-TREE 2 and FastTree 2 on real data, using the same strategies as with simulated data, terminating experiments that took more than 24 h. In these Real Online and De Novo Experiments, we measured the Jukes-Cantor (JC) likelihood of each tree, as well as the runtime and peak memory usage of each program. Files for all real data experiments can be found in subrepository 3.

2.3.5 Analyses Without Pandemic Time Constraints

Eliminating the 24-h runtime restriction, we also performed 3 *de novo* iterative experiments on both real and simulated data. In these Real and Simulated Unrestricted

Experiments, we tested UShER + matOptimize, ML IQ-TREE 2 with stochastic search, and RAxML-NG on iterations of ~4.5k, ~8.9k, and ~13.2k samples, allowing each to run for up to 14 days. For runs that did not terminate within this time (the 2nd and 3rd iterations of RAxML-NG), we used the best tree inferred during the run for comparisons. We ran IQ-TREE 2 without the “-n 1” parameter as in previous experiments, enabling stochastic search. We ran IQ-TREE 2 and RAxML-NG under the GTR+G model with the smallest minimum branch length parameter that did not cause numerical errors. To compare the trees in the Real Unrestricted Experiments, we computed log-likelihoods under the GTR+G model for all trees, fixing the model parameters to those estimated by IQ-TREE 2 during tree inference. We also compared the log-likelihoods of the trees under the parameters estimated by RAxML-NG for the 1st iteration, but could not do so for the 2nd and 3rd iterations which did not terminate in under 2 weeks. We allowed optimization of branch lengths during likelihood calculation. For the UShER+matOptimize trees, before computing likelihoods, we converted the branch lengths into units of substitutions per site by dividing each branch length by the alignment length (29,903). To compare the trees inferred in the Simulated Unrestricted Experiments, we computed the RF distance and quartet similarity of each tree to the corresponding ground truth tree described above.

2.3.6 Correlation of Parsimony and Likelihood on Large SARS-CoV-2 Phylogenies

We also performed an experiment on the starting tree of 364,427 samples to measure the correlation between the 2 optimality metrics used in our main experiments, parsimony and JC likelihood. In our Correlation Optimization Experiment, we performed 6 rounds of FastTree 2, measured the parsimony score and JC likelihood of the

resulting trees, and computed the Pearson correlation coefficient between parsimony and likelihood.

2.4 Results and Discussion

2.4.1 Online Phylogenetics is an Alternative to de novo Phylogenetics for Ongoing Studies

The vast majority of phylogenetics during the pandemic has consisted of *de novo* phylogenetics approaches (Hadfield et al. 2018; Li et al. 2020; Lu et al. 2020a, 2020b; Meredith et al. 2020), in which each phylogeny is inferred using only genetic variation data, and without a guide tree (Fig. 2.1). This strategy for phylogenetic inference has long been the default, as in most instances in the past, data are collected just once for a project, and more relevant data are rarely going to be made available in the near future. This process is well characterized and has been foundational for many phylogenetics studies (Hug et al. 2016; Parks et al. 2018; Lu et al. 2020b), and most phylogenetics software is developed with *de novo* phylogenetics as the primary intended usage.

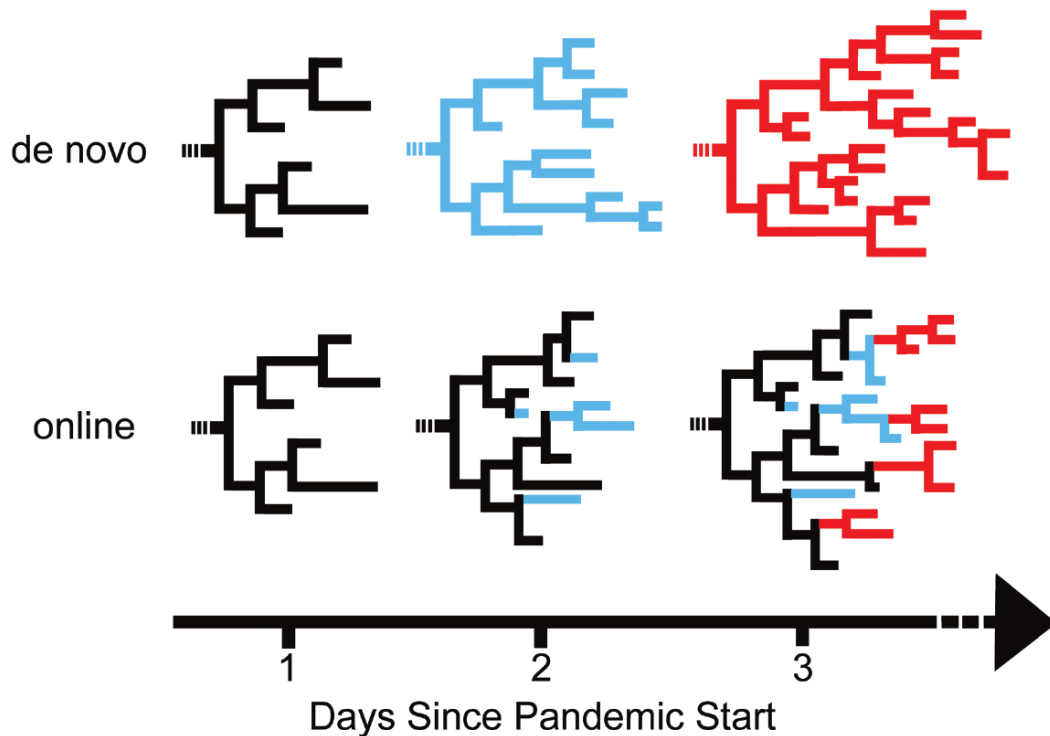


Figure 2.1. Phylogenies may be optimized from scratch using *de novo* phylogenetics or iteratively using online phylogenetics. In *de novo* phylogenetics (top), trees are repeatedly re-inferred from scratch. Conversely, online phylogenetics (bottom) involves placement of new samples as they are collected. Online methods may be particularly well suited to pathogen datasets like SARS-CoV-2 where close relatives of ancestral samples already exist in the phylogeny. Intermittent optimization steps (not depicted) after new samples are placed can help overcome errors from previous iterations. Online phylogenetics is expected to be much faster and require less memory than *de novo* phylogenetics.

A challenging aspect of pandemic phylogenetics is the need to keep up with the pace of data generation as genome sequences continuously become available. To evaluate phylogenetics applications in the pandemic (Fig. 2.1), we split 233,326 samples dated from 23 December 2019 through 11 January 2021 into 50 batches according to their date of collection. Each batch contains roughly 5000 samples. Samples in each batch were collected within a few days of each other, except in the 1st months of the

pandemic when sample collection was more sparse. We also constructed a data set of otherwise similar data simulated from a known phylogeny (see Methods). The intent of this scheme is to roughly approximate the data generation and deposition that occurred during the pandemic. All data sets are available from the repository associated with this project (Thornlow et al. 2021b), for reproducibility and so that future methods developers can directly compare their outputs to our results. We performed online and *de novo* phylogenetics using a range of inference and optimization approaches. Since thousands of new sequences are added to public sequence repositories each day, we terminated any phylogenetic inference approaches that took more than 24 h, because such phylogenies would be obsolete for some public health applications by the time they were inferred.

2.4.2 Analyses Using Simulated Data Suggest that Online Phylogenetics is More Accurate for SARS-CoV-2

We first compared matOptimize (conda version 0.5.0) (Ye et al. 2022), ML IQ-TREE 2 (Minh et al. 2020), and FastTree 2 (Price et al. 2010) using both online and *de novo* phylogenetics strategies using simulated data that we designed to closely mimic real SARS-CoV-2 data sets. All Simulated Online Experiments used UShER (Turakhia et al. 2021a) to add new sequences to the previous tree (see Methods) as to the best of our knowledge it is the only software package that is fast enough to perform under real-time constraints. We chose these 3 tools to cover several different approaches and based on their widespread usage among SARS-CoV-2 phylogenetics applications. For example, matOptimize is part of the UShER suite (Turakhia et al. 2021a), IQ-TREE

2 is used by (COVID-19 Genomics UK (COG-UK) Consortium 2020, Lanfear and Mansfield 2020) and FastTree 2 is used by (Hadfield et al. 2018).

Simulating an alignment based on a known tree ensures that there is a ground truth for comparison to definitively assess each optimization method. We used an inferred global phylogeny as a template to simulate a complete multiple sequence alignment using phastSim (De Maio et al. 2022). We subsampled this simulated alignment into 50 progressively larger sets of samples, ranging in number of samples from 4676 to 233,326, to examine each of the 3 optimization methods in both online and *de novo* phylogenetics. For each iteration, we condensed identical samples, collapsed very short branches, and computed the RF distance to the condensed, collapsed ground truth global tree on which the simulation was based (see Methods). We also computed a quartet similarity metric (Asher and Smith 2022) to the ground truth tree. The ground truth tree was pruned to contain only the relevant samples for each iteration.

All Simulated Online Experiments noticeably outperformed their *de novo* counterparts in the majority of iterations. Overall, online matOptimize produced phylogenies with the lowest normalized RF distance to the ground truth for the majority of iterations (Fig. 2.2a). Online IQ-TREE 2 performed similarly in RF distance but was able to complete only 24 of the 50 iterations due to its extreme computational resource requirements. When measured by quartet similarity, FastTree 2 produced trees closest to ground truth for the majority of the 14 iterations it completed (Fig. 2.2b). *De novo* IQ-TREE 2 performs relatively worse when measured by quartet similarity, perhaps due to errors early in tree inference that significantly affect the quartet similarity. For example, for the 9th phylogeny of 39,621 sequences, which was the last phylogeny

produced using under 200 GB of RAM in under 24 h by all 6 methods, we found normalized RF distances of 0.0232, 0.0376, and 0.0297 for *de novo* USHER+matOptimize, FastTree 2, and IQ-TREE 2, respectively, and distances of 0.0220, 0.0296, and 0.0225 for online matOptimize, FastTree 2, and IQ-TREE 2, respectively. The distance of all methods to the ground truth tree are exceptionally small, possibly due to the “near perfect” phylogenetic properties of SARS-CoV-2 data (Wertheim et al. 2022).

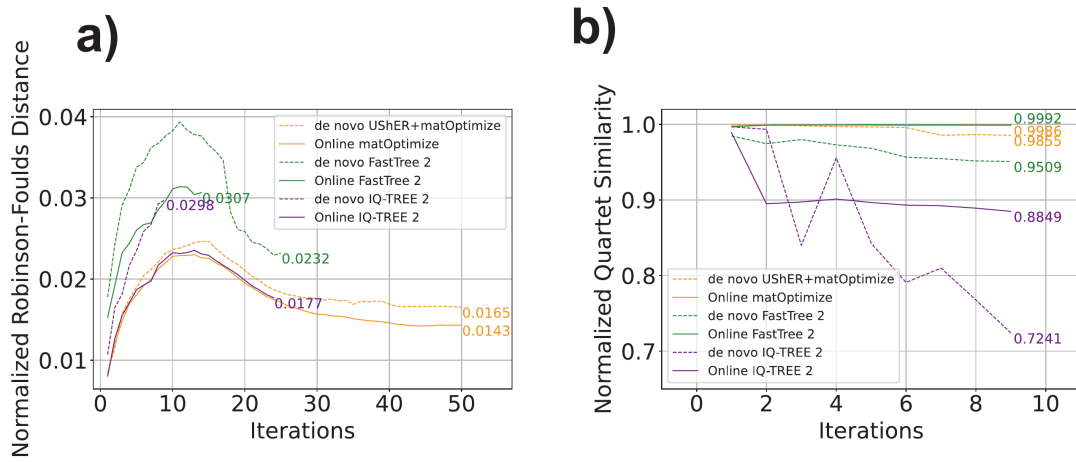


Figure 2.2. Online matOptimize produces highly similar phylogenies to ground truth on simulated data. For each batch of samples in our Simulated Online and De Novo Experiments, we calculated the normalized Robinson–Foulds (RF) distance between the tree produced by a given optimization software and the ground truth tree pruned to contain only the relevant samples, first collapsing near-zero branch lengths into multifurcations in both trees. To normalize RF distances, we divided each distance by $I(T_i) + N_i - 3$, where $I(T_i)$ is the number of internal edges in the pruned ground truth tree for iteration i , and N_i is the number of taxa (see Supplementary Fig. S2.1). We computed a quartet similarity metric to the ground truth tree for ten iterations of each method, normalizing by the maximum possible score of the metric as described in (Asher and Smith 2022). We terminated FastTree and IQ-TREE 2 after the first phylogeny that took more than 24 h to optimize.

There are many possible explanations for the improved performance of online phylogenetics on simulated data relative to *de novo* approaches. One possibility is that smaller trees are easier to optimize, so online phylogenetics will tend to estimate

accurate initial trees to which later samples are added (this occurs because the radius for SPR moves when optimizing a large tree is too small to find improvements that are more readily applied when the tree contains fewer samples). In online phylogenetics, these improvements carry over to subsequent trees, while in *de novo* phylogenetics, they do not. The radius is defined as the phylogenetic distance of the search space when moving a node to a more optimal position. As the phylogeny increases in size, the distance from a node to its optimal position is likely to also increase, necessitating a larger SPR move radius to make equivalent improvements in larger trees. Additionally, the temporal nature of sample placement likely aids the performance of online approaches, with newer samples added to the tree after older samples.

2.4.3 Analyses Using Real Data Suggest that Online Phylogenetics is More Efficient than de novo and Produces Similarly Optimal Phylogenies

While analyses using simulated data offer the ability to compare to a known ground truth, assessing the performance of each method on real SARS-CoV-2 data may more accurately reflect practical use of each method. Therefore, we also tested each optimization strategy on 50 progressively larger sets of real SARS-CoV-2 samples in our Real Online and De Novo Experiments and calculated the parsimony score and likelihood of each optimized tree, as well as the run-time and peak RAM usage of each software package used (Fig. 2.3). To accomplish this, we subsampled our global phylogeny, which was produced using stringent quality control steps (see Methods), as before, to mimic the continuous accumulation of samples over the course of the pandemic.

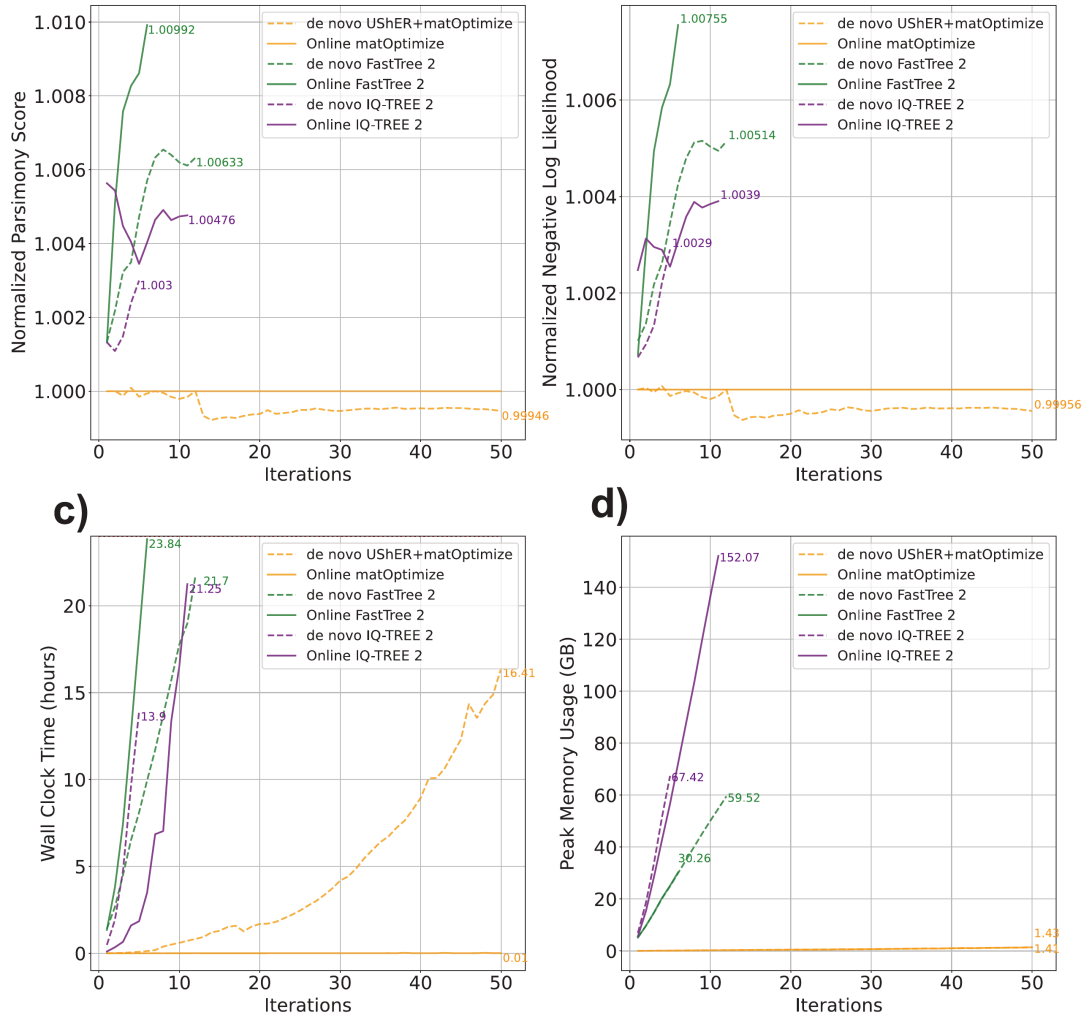


Figure 2.3. In practice, optimization by parsimony with matOptimize is more effective for SARS-CoV-2 data than optimization by established ML or pseudo-ML methods. For each Real Online and De Novo Experiment, we calculated (A) the parsimony score of each tree using matUtils, (B) the log-likelihood of each tree using IQ-TREE 2, (C) runtime and (D) peak memory usage of each optimization. (A) and (B) are normalized by the value obtained for the matOptimize online approach such that all other methods are expressed as a ratio. Strategies that surpassed 24 h (C) or the allowable RAM usage (D) were terminated prior. In most cases, with the notable exception of FastTree 2, online phylogenetics (solid lines) perform better than *de novo* phylogenetics (dashed lines). We ran all matOptimize analyses using an instance with 15 CPUs and 117.2 GB of RAM, and we ran all IQ-TREE 2 and FastTree 2 analyses on an instance with 31 CPUs and 244.1 GB of RAM, but limited each command to 15 threads for equivalence with matOptimize.

Online optimizations are generally much faster than *de novo* phylogenetic inference. For example, IQ-TREE 2 achieves a roughly 7-fold faster run-time for online optimizations compared to inferring the tree *de novo* (Fig. 2.3c). The 5th iteration, which has 22,012 sequences and was the last to be completed by both online and *de novo* IQ-TREE 2, took 13 h 53 min for *de novo* IQ-TREE 2 but only 1 h 50 min for online IQ-TREE 2. *De novo* UShER+matOptimize was the only *de novo* method to finish all trees in fewer than 24 h, but its speed for each daily update pales in comparison to online matOptimize. Online matOptimize is several orders of magnitude faster than its *de novo* counterpart, and its optimizations for the largest phylogenies take roughly 30 s, while *de novo* tree inference with UShER can take several hours for trees consisting of more than 100,000 samples (Fig. 2.3c). However, whether a software package is used for online or *de novo* phylogenetics does not strongly affect its peak memory usage.

We also found that online phylogenetics strategies produce trees very similar in both parsimony score and likelihood to their *de novo* counterparts, with differences of less than 1% in all cases (Fig. 2.3a and b). For example, in the 5th iteration containing 22,012 sequences, online IQ-TREE 2 produces a tree with a parsimony score of 13,393, whereas *de novo* IQ-TREE 2 produces a tree with parsimony score of 13,387. Our results suggest that in addition to the computational savings that allow online phylogenetics approaches to continuously stay up-to-date, online phylogenetics approaches also produce trees with similar parsimony scores and likelihoods to their *de novo* counterparts.

2.4.4 Under Pandemic Time Constraints, matOptimize has Equivalent or More Favorable Metrics Compared to ML and Pseudo-ML Methods for SARS-CoV-2 Phylogenies

In the case of both *de novo* and online phylogenetics, the parsimony-based matOptimize outperforms both FastTree 2 and IQ-TREE 2 in runtime and peak memory usage. For the 6th iteration (26,486 samples) of our Real Online and De Novo Experiments, which was the largest phylogeny inferred by all online methods in under 24 h and using under 200 GB of RAM, online FastTree 2 required nearly 24 h and 30.3 GB of RAM, and online IQ-TREE 2 required 3 h 29 min and 72 GB of RAM. By contrast, matOptimize required only 6 s and 0.15 GB of RAM. This iteration contained roughly 10% as many samples as the 50th and final iteration (233,326 total samples), which online matOptimize completed in 32 s using 1.41 GB of RAM at peak usage. Even this largest tree represents only a very small fraction of the more than 14 million currently available SARS-CoV-2 genomes, indicating that, among the approaches we evaluated, matOptimize is the only viable option for maintaining a comprehensive SARS-CoV-2 phylogeny via online phylogenetics.

In addition to its scalability, matOptimize outperformed ML and pseudo-ML optimization methods under 24-h time constraints in both the parsimony and likelihood scores of the trees that it inferred. For the 6th iteration (26,486 samples) of our Real Online Experiments, we found parsimony scores of 16,130, 16,195, and 16,290 for matOptimize, IQ-TREE 2, and FastTree 2, respectively. While all methods produced phylogenies with parsimony scores within 1% of each other, matOptimize was consistently the lowest. However, matOptimize was developed to optimize by parsimony, while the other methods optimize for likelihood. Unexpectedly, we found that the

trees from matOptimize had better log-likelihood than those from likelihood-based software implementations. For the 6th iteration (26,486 samples) of our Real Online Experiments, we found log-likelihood scores of $-233,414.277$, $-234,130.859$, and $-235,177.396$ for matOptimize, IQ-TREE 2, and FastTree 2, respectively. IQ-TREE 2 and FastTree 2 inferred trees using a Generalized Time Reversible (GTR) model, but due to time constraints in calculation we used a JC model to compare likelihoods across methods. However, a GTR model with specified rate parameters produced strongly correlated likelihoods (Supplementary Table S2.1, Dryad: <https://doi.org/10.7291/D13Q2J>) (Thornlow et al. 2022). Indeed, we were able to calculate likelihoods under the GTR model for the first 5 iterations of the Real De Novo Experiments (from 4676 to 22,012 samples) for both the matOptimize and IQ-TREE 2 trees. In every case, the tree from matOptimize had a slightly better likelihood score than the tree from IQ-TREE 2.

2.4.5 Parsimony Optimization with matOptimize Produces Comparable SARS-CoV-2 Trees to the Most Thorough ML Methods

We also compared the performance of *de novo* inference with UShER+matOptimize to state-of-the-art methods without a 24-h limit on runtime in our Real and Simulated Unrestricted Experiments. In 3 iterations of increasing size ($\sim 4.5k$, $\sim 8.9k$, and $\sim 13.2k$ samples), we inferred trees from real and simulated data using UShER+matOptimize, IQ-TREE 2 with stochastic search enabled, and RAxML-NG. We allowed each experiment to run for up to 2 weeks. All programs completed successfully on the 1st iteration. RAxML-NG did not terminate within 2 weeks for the 2nd and 3rd iterations. On real data, we found that UShER+matOptimize produced trees

with higher log-likelihoods than IQ-TREE 2 and RAxML-NG across all 3 iterations (Fig. 2.4a). Under the substitution model parameters estimated by IQ-TREE 2, the log-likelihoods for the 1st iteration were $-73,780.756$, $-73,828.271$, and $-73,782.289$ for UShER+matOptimize, IQ-TREE 2, and RAxML-NG, respectively. Under the parameters estimated by RAxML-NG, the log-likelihoods for the 1st iteration were $-73,754.894$, $-73,801.935$, and $-73,756.246$ for UShER+matOptimize, IQ-TREE 2, and RAxML-NG, respectively. On simulated data, UShER+matOptimize produced trees closer to the ground truth than the other methods when measured by quartet similarity across all 3 iterations (Fig. 2.4b). By RF distance, the UShER+matOptimize trees were closest to the ground truth for the 2nd and 3rd iterations, but the RAxML-NG tree was closest to ground truth in the 1st iteration (Fig. 2.4c). On the 1st iteration of simulated data, we found normalized RF distances of 0.0084, 0.0117, and 0.0071 for UShER+matOptimize, IQ-TREE 2, and RAxML-NG, respectively. We found normalized quartet similarities of 0.999, 0.997, and 0.969 for UShER+matOptimize, IQ-TREE 2, and RAxML-NG, respectively. We therefore conclude that parsimony-based tree inference with matOptimize can perform equivalently to state of the art ML approaches and can do this in a tiny fraction of the time, making it by far the most suitable approach for pandemic-scale phylogenetics of SARS-CoV-2.

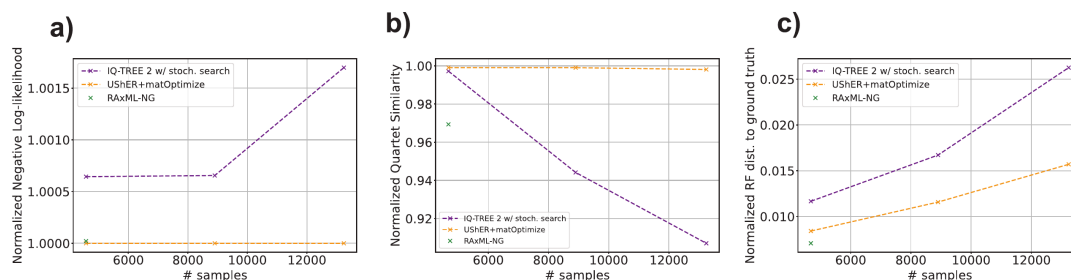


Figure 2.4. De novo matOptimize produces comparable trees to the most thorough ML inference programs. We ran these methods for up to 2 weeks each to infer trees *de novo* from the 3 smallest iterations of real and simulated data. For real data (A), log-likelihoods were computed under the model parameters estimated by IQ-TREE 2 at each iteration. Values are normalized by the value of the matOptimize approach such that all other methods are expressed as a ratio. For simulated data (B, C), the reported quartet similarities are normalized by the maximum value of the metric described in (Asher and Smith 2022). RF distances (C) are normalized by dividing by $I(T_i) + N_i - 3$, where $I(T_i)$ is the number of internal edges in the pruned ground truth tree for iteration i , and N_i is the number of taxa. For all panels, the 2nd and 3rd iterations of RAxML-NG (which did not terminate within 2 weeks) are omitted.

2.4.6 Parsimony and Likelihood are Strongly Correlated When Optimizing Large SARS-CoV-2 Phylogenies

In our Correlation Optimization Experiment on the starting tree of 364,427 SARS-CoV-2 genomes, we found that after each of 6 iterations of FastTree 2 optimization, the JC likelihood and parsimony improvements are strongly linearly correlated (Fig. 2.5). This suggests that changes achieved by maximizing parsimony will also optimize likelihood for SARS-CoV-2 data. That is, for extremely densely sampled phylogenies in which long branches are rare, parsimony and likelihood are highly correlated, as are the effects of tree moves to optimize either. However, despite the strength of this correlation, we find an extreme disparity in practical usage when optimizing by either metric. Parsimony-based methods are far more time-

and data-efficient, and presently available ML and pseudo-ML approaches quickly become prohibitively computationally expensive. For example, while the 6 iterations of FastTree did result in large improvements in both likelihood and parsimony score, the resulting tree would be out of date long before the 10.5-day optimization had completed. Moreover, we applied matOptimize to the tree output by the 6th iteration of FastTree, achieving a parsimony score of 293,866 (improvement of 288) and a JC log-likelihood of -3483329.485 (improvement of 2318.535) in just 16 min, indicating that even after 10.5 days, additional optimization was still possible. This suggests that, for the purposes of optimizing even moderately large SARS-CoV-2 trees, parsimony-based methods should be heavily favored due to their increased efficiency.

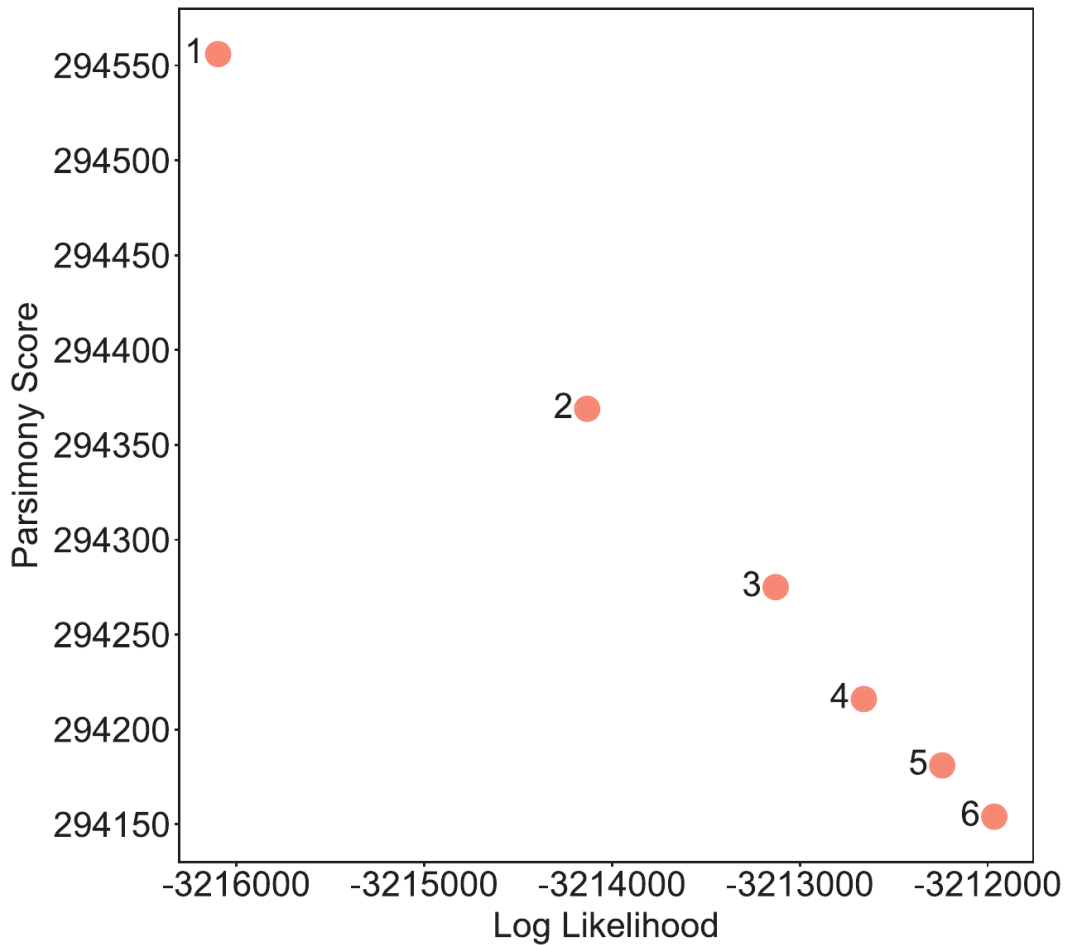


Figure 2.5. Improvement in likelihood and parsimony have a linear relationship for our optimized global tree. We optimized our initial global tree using 6 iterations of FastTree and measured the total parsimony and the likelihood after each, finding a linear relationship (Pearson correlation, $\rho = -1.0$, $P < 2.9 \times 10^{-7}$).

2.5 Conclusions

The SARS-CoV-2 pandemic has made phylogenetics central to efforts to combat the spread of the virus, but has posed challenges for many commonly used phylogenetics frameworks. A major component of this effort relies on a comprehensive, up-to-date, global phylogeny of SARS-CoV-2 genomes. However, the scale

and continuous growth of the data have caused difficulties for standard *de novo* phylogenetic methods. Here, we find that online phylogenetics methods are practical, pragmatic, and accurate for inferring daily phylogenetic trees from a large and densely sampled virus outbreak.

One counterintuitive result is that parsimony-based optimization outperformed sufficiently efficient ML and pseudo-ML approaches in the majority of our experiments regardless of whether phylogenies were evaluated using parsimony or likelihood. This might be a consequence of the fact that parsimony scores and likelihoods are strongly correlated across phylogenies inferred via a range of phylogenetic approaches. The extremely short branches (Supplementary Fig. S2.2) on SARS-CoV-2 phylogenies mean that the probability of multiple mutations occurring at the same site on a single branch is negligible. Stated another way, SARS-CoV-2 is approaching a “limit” where parsimony and likelihood are nearly equivalent. Similar observations have been made in other fields that, in certain cases, nonparametric methods based on parsimonious solutions perform similarly to statistical methods that assume an underlying model (Huelsenbeck et al. 2000; Ree et al. 2005; Nylander et al. 2008). In turn, because of its relative efficiency, matOptimize is able to search more of the possible tree space in the same amount of time, thereby resulting in trees with better likelihoods and lower parsimony scores than trees optimized using currently available ML software packages. We emphasize that this does not bear on the relative merits of the underlying principles of ML and MP, but instead reflects the utility of methods that have been applied during the pandemic. Nevertheless, this observation does suggest that in some cases, MP optimization may provide a fast and accurate starting point for

ML optimization methods. Indeed, many popular phylogenetic software packages such as RAxML (Stamatakis 2014) and IQ-TREE (Minh et al. 2020) already use stepwise-addition parsimony trees as initial trees for their optimization or in conjunction with likelihood during inference as in PhyML 3.0 (Guindon et al. 2010). Our results suggest that further optimization of initial trees using MP may provide benefits in speed and accuracy for some data sets, even when the target is an estimate of the ML tree.

As sequencing technologies progress, sample sizes for phylogenetic analyses of major pathogens and highly studied organisms will necessarily continue to increase. Today, SARS-CoV-2 represents an extreme with respect to the total number of samples relative to the very short branch lengths on the phylogeny. However, the global sequencing effort during the pandemic suggests that the public health sphere has a strong interest in the increased application of whole-genome sequencing to study the genomic contents, evolution, and transmission history of major and emerging human pathogens. In addition to our global tree of SARS-CoV-2 with ~14 million samples, we use online phylogenetics with UShER and matOptimize to maintain 3 other publicly available viral phylogenies, with a web server for sample placement hosted at <https://genome.ucsc.edu/cgi-bin/hgPhyloPlace>. As of February 2023, our Mpox tree contains 3751 genomes, our RSV-A tree contains 2592 genomes, and our RSV-B tree contains 1915 genomes. In addition, we have applied UShER and matOptimize to the bacterial pathogen *Mycobacterium tuberculosis*, constructing a *de novo* tree with 10,248 genomes (Ye et al. 2022). All of these pathogen trees share relatively dense sampling and short average branch lengths, so the results of this paper suggest they are well-suited to inference using online parsimony methods. We expect that million-sample

data sets will become commonplace in the near future, and that parsimony-based methods like `matOptimize` show promise for many pathogens beyond SARS-CoV-2. Recently developed parsimony-based likelihood approximations may ultimately be similarly scalable and accurate (De Maio et al. 2023). Online phylogenetics using both of these methods will be a fruitful avenue for future development and application to accommodate these data sets.

CHAPTER 3. LARGE-SCALE TREE ANALYSIS AND VISUALIZATION

This chapter presents two browser tools for working with trees of millions of genomes. ShUSHER runs UShER in the browser, so researchers can place their own sequences on the global tree without sending them to a server. Treenome Browser shows the genome of every sample next to its place in the tree, drawing each mutation at its position in the genome and across the samples that carry it. Both tools depend on the UShER mutation-annotated tree (MAT) used in Chapter 2.

This chapter is a reproduced form of the following publications:

Kramer, A. M., Sanderson, T., & Corbett-Detig, R. (2023). Treenome Browser: co-visualization of enormous phylogenies and millions of genomes. *Bioinformatics* 39(1):btac772.

Kramer, A. M., Turakhia, Y., & Corbett-Detig, R. (2021). ShUSHER: private browser-based placement of sensitive genome samples on phylogenetic trees. *Journal of Open Source Software* 6(66):3677.

3.1 ShUSHER: privacy-preserving placement in the browser

3.1.1 Motivation: placement without data disclosure

The methods of Chapter 2 are valuable for not only for phylogenetic inference (e.g., for maintaining the global SARS-CoV-2 tree) but also for analysis of new samples (by, e.g., public health researchers) through placement onto large reference

phylogenies. Some typical questions a genomic surveillance lab asks are: given a genome that has been sequenced, where does it sit on the global tree, which known samples are its closest relatives, and does it fall inside a notable lineage or cluster? UShER answers these questions through phylogenetic placement on huge trees, placing new samples onto existing mutation-annotated trees (MATs) without re-inferring the surrounding phylogeny (Turakhia et al., 2021). UCSC exposes this capability to the public through a web tool hosted alongside the UCSC Genome Browser at usher.bio, which lets users upload sequences and receive their placement on a global tree of millions of SARS-CoV-2 genomes.

That tool, like most bioinformatics web services, works by uploading the user's data to a server, performing the computation there, and returning the result. For most sequences this is acceptable, but in some jurisdictions a viral consensus genome is treated as Protected Health Information (PHI), because in principle it can be linked to the infected individual from whom it was sequenced. Where that is the policy, or where data governance rules or other regulations forbid transmitting sequence data outside a controlled environment, sending a genome to an external server is not permitted.

Previously, the alternative to using usher.bio was to download the UShER tool locally and run it from the command line. This works, but it excludes those researchers who need to use UShER on sensitive data but who do not have extensive computational expertise. For example, a hospital epidemiologist or a public health analyst is often not a command-line bioinformatician, and requiring them to build and install a C++ program, obtain and update a multi-gigabyte reference tree, and manage numerous file formats is a substantial barrier.

ShUSHER (Shh: private Ultrafast Sample placement on Existing tRees) resolves this by removing the server from the computation. It is a browser based application that runs the UShER placement algorithm entirely on the user's own machine, inside the web page. The user visits a URL and is presented with a familiar web interface, but none of their data is transmitted anywhere. Genomic data is processed, aligned, placed, and visualized locally using a custom WebAssembly and JavaScript port of UShER.

3.1.2 Design and workflow

UShER is a mature C++ program built for a server or workstation, whereas a browser executes JavaScript in a sandbox with no access to native binaries. Rather than reimplement placement in JavaScript, ShUSHER compiles the existing UShER C++ source to WebAssembly (using Emscripten), a portable bytecode that can be executed at near native speed within the browser sandbox. The code running in the browser is the UShER algorithm and a React interface that orchestrates the placement. Placement also requires preprocessing: users supply consensus sequences, typically in raw FASTA format, which must first be aligned to the SARS-CoV-2 reference and reduced to single-nucleotide variants. ShUSHER performs this step in-browser as well, reusing components from Nextclade (CITE), so that the entire pipeline runs without transmitting the sequence.

The user navigates to <https://shusher.gi.ucsc.edu/> (Fig. 3.1), selects a reference tree, and provides their samples. ShUSHER downloads the global SARS-CoV-2 tree from the daily-updated database of Chapter 2 (McBroome et al., 2021). Then, the tool aligns the samples, places each on the tree, and extracts the subtrees surrounding the

placed samples. These subtrees are then displayed with Auspice, the tree-visualization component of Nextstrain (Hadfield et al., 2018), via the auspice.us application, also locally in the browser.

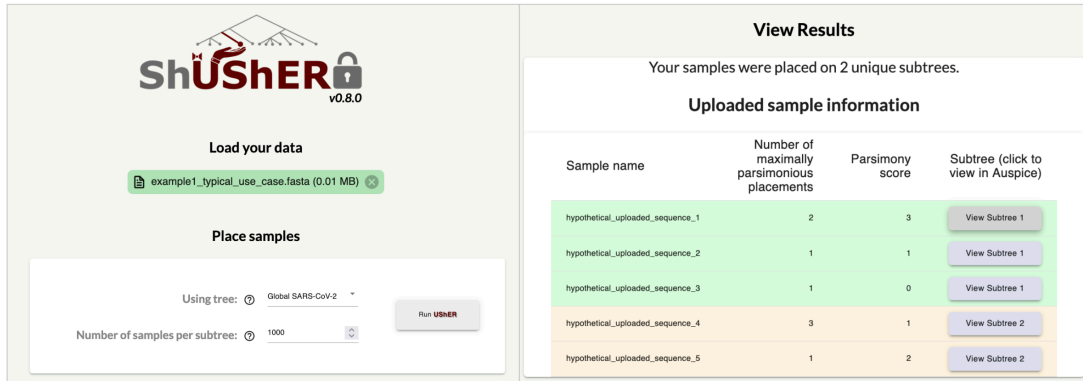


Figure 3.1. The ShUSHER web interface. Sequences are loaded from a local FASTA file (left panel), which never leaves the browser; the user selects a reference tree and the number of samples to include in each returned subtree, then runs UShER. The results panel (right) lists each uploaded sample together with the number of maximally parsimonious placements found for it, its parsimony score, and a button that opens the surrounding subtree in Auspice. Samples placed on the same subtree are grouped, so that related uploads can be inspected together.

3.1.3 Privacy guarantee and outlook

The advantage provided by ShUSHER is that the user’s sequence data are never transmitted over the Internet. This comes at a cost, because the computation is bounded by what is achievable in the web browser sandbox. These constraints make ShUSHER a tool for interactive, laboratory-scale placement rather than bulk processing of large query sets, for which local command-line UShER remains more appropriate. ShUSHER therefore makes placement on the global tree available to laboratories that cannot use a server-based service. The web application and source code is freely available at <https://shusher.gi.ucsc.edu/> and <https://github.com/amkram/shusher>.

3.2 Treenome Browser: co-visualization of phylogenies and genomes

3.2.1 *Motivation*

A phylogeny of millions of genomes and the genetic variation underlying it are two views of the same data, but existing tools could show only one at a time. UShER and matOptimize maintain a continuously growing SARS-CoV-2 tree of millions of samples (Chapter 2), and Taxonium renders trees of that size interactively (Sanderson, 2022), but a tree alone does not show where in the genome its defining mutations fall. Conversely, tools such as Gingr (Treangen et al., 2014) and the UCSC Genome Browser can place a multiple sequence alignment beside a phylogeny (Lee et al., 2022), but not at the scale of millions of sequences. No existing tool could display the complete genetic variation of a million-sample dataset alongside the tree relating it. Treenome Browser fills this gap: it displays the genetic variants of every sample (relative to a reference genome) aligned with its position in the tree, and remains responsive on trees of more than twelve million sequences. It is available at cov2tree.org for the global SARS-CoV-2 tree and at taxonium.org for user-provided trees.

3.2.2 *Description*

Treenome Browser draws each mutation as a vertical line, positioned horizontally at the mutation's coordinate in a reference genome and spanning vertically the set of samples in the phylogeny that carry it (Fig. 3.2). Both nucleotide and amino acid mutations are supported. When the whole tree is in view, the genetic signatures of major clades emerge as bands of shared mutations; zooming or searching to a clade or an individual sample reveals finer structure, and the genome axis can be navigated to

examine particular sites or regions. This co-visualization supports applications such as variation-informed primer design, uncovering technical artifacts in sequence assembly or tree construction, and investigating molecular evolution. The SARS-CoV-2 build at cov2tree.org is preloaded with reference annotations (RefSeq NC_045512v2) curated by the UCSC Genome Browser (Lee et al., 2022); user-provided mutation-annotated trees built with UShER can be prepared with TaxoniumTools and loaded at taxonium.org.

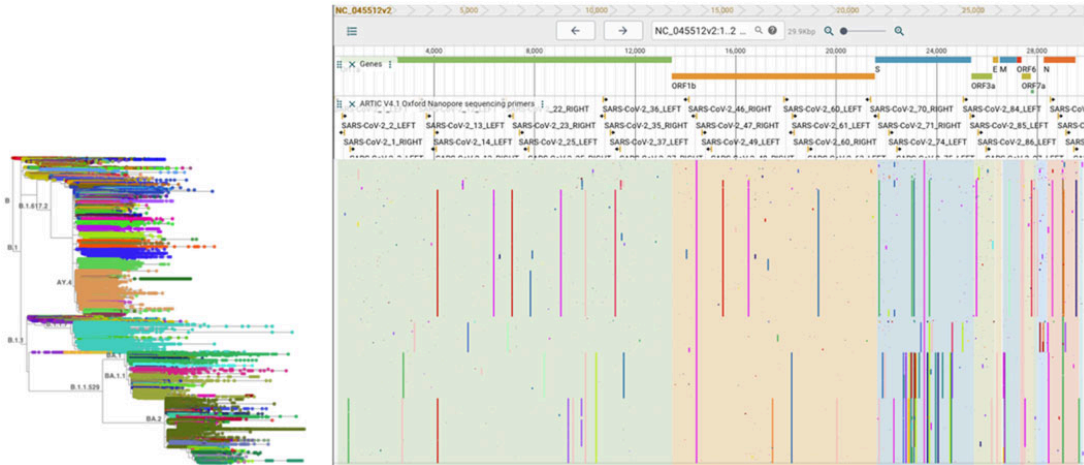


Figure 3.2. Treenome Browser for the global SARS-CoV-2 tree with 5.9 million genomes, shown with the ARTIC Primers v4.1 annotation track. The phylogeny is shown at left; colored lines represent amino acid mutations relative to the reference genome, drawn at their position along the genome (top axis).

3.2.3 Implementation

Treenome Browser scales to huge trees by exploiting the mutation-annotated tree itself. Because a mutation-annotated tree encodes each mutation once, on the branch where it arose, rather than repeating it in every genome that inherits it, it is often many times smaller than the equivalent VCF, and Treenome Browser uses this phylogenetic compression directly for display. Its core algorithm decodes the tree to

compute the on-screen position of each mutation: the vertical span of every subclade is precomputed by dynamic programming, and these spans are combined with horizontal coordinates during a single root-to-leaf traversal to draw every mutation without ever enumerating the full genotype of each leaf. Rendering uses the `deck.gl` framework, which drives WebGL to draw these large mutation sets far faster than conventional web visualization. Performance is further improved by subsampling: Treenome Browser reuses the sparsified trees Taxonium already displays when zoomed out (Sanderson, 2022) and discards mutations spanning clades too small to perceive at the current zoom level, excluding only imperceptible detail while keeping every individual node and mutation reachable. The reference genome and annotation tracks are shown in an interactive JBrowse 2 panel (Buels et al., 2016), with SARS-CoV-2 annotations fetched as bigBed and bigWig files through the UCSC Genome Browser API. Treenome Browser is implemented in React as a feature addition to Taxonium; source code is available at <https://github.com/theosanderson/taxonium>.

3.3 Applications

Treenome Browser has several potential applications. Because every mutation is drawn at its coordinate in the genome and across the samples that carry it, a candidate primer region can be checked against the sites that vary among circulating lineages, since variation under a primer compromises binding. The same view can reveal technical artifacts from sequence assembly or tree construction, which appear as implausible mutation patterns: a mutation recurring on many unrelated branches, for instance, or mutations clustering at sites already known to be problematic. Recurrent

and reversional mutations are visible together, which is useful for studying homoplasy and convergent evolution.

ShUSHER gives a laboratory the local phylogenetic context of a new sample, meaning its nearest relatives in the global tree and its lineage assignment, without transmitting the sequence. Jurisdictions that treat viral consensus genomes as protected health information cannot use server-based placement services, and ShUSHER is intended for them. It can be applied to any pathogen for which a mutation-annotated reference tree exists.

3.4 Discussion

ShUSHER and Treenome Browser address two obstacles to using a phylogeny of millions of genomes: placing samples when the data cannot be shared, and reading the genetic variation alongside the tree. Both remain interactive at this scale by using the same mutation-annotated representation that makes the inference in Chapter 2 tractable, in which each mutation is stored once on the branch where it arose rather than repeated in every descendant genome.

Both tools, however, describe variation relative to a single linear reference genome. This is adequate for a low-diversity, low-indel pathogen such as SARS-CoV-2, but it under-represents the insertions, deletions, and structural variation that a single reference cannot encode, and it inherits the reference bias common to all reference-based analysis. Chapter 4 removes this constraint, replacing the single reference with a lossless, phylogeny-aware pangenome.

CHAPTER 4. PHYLO-PANGENOMIC ANALYSIS WITH PAN-MAP

This chapter moves from single reference sequences to pangenomes. It presents Panmap, which maps, aligns, and genotypes sequencing reads against mutation-annotated pangenomes (PanMANs) of up to millions of genomes, and estimates metagenomic abundances. A PanMAN, unlike an UShER mutation-annotated tree, is lossless. It encodes the full sequences of every input genome, including the indels and structural variation that a single reference leaves out.

This chapter is a reproduced form of the following publication:

Kramer, A. M., Zhang, A., Ayala, N., de Sanctis, B., Karim, L. M., Hinrichs, A. S., Walia, S., Turakhia, Y., & Corbett-Detig, R. (2026). Panmap: Scalable phylogeny-guided alignment, genotyping, and placement on pangenomes. *bioRxiv*. <https://doi.org/10.64898/2026.03.29.711974>

4.1 Introduction to Panmap

Pangenome approaches have emerged as a powerful alternative to single-reference methods, jointly leveraging many reference genomes to capture greater genetic diversity and improve variant identification. Because they encode known variants within populations, species, or clades (Marschall et al. 2018), reference pangenomes enable more sensitive and accurate analyses than single-reference approaches (Ebler et al. 2020) and have become essential in fields such as human genomics, infectious

disease surveillance, and agriculture (Sirén et al. 2021; Zhou et al. 2022; Rice et al. 2023).

Most pangenome methods represent variation as sequence graphs, encoding genomes as paths through networks of shared and divergent sequences (Liao et al. 2023). Tools like VG (Garrison et al. 2018), Progressive Cactus (Armstrong et al. 2020), and Minigraph (Li et al. 2020) exemplify this framework. However, graph-based approaches face significant scalability challenges. They often require extensive memory and time for read alignment and variant identification, especially for large microbial pangenomes that may contain thousands or millions of whole genomes. In addition, sequence graphs typically represent variation structurally but do not explicitly encode the evolutionary history relating sequences, limiting opportunities to leverage ancestral sequence information during analysis.

An alternative paradigm, used by tools like UShER and tskit, represents genomic variation as genealogical structures, encoding variants as mutations along evolutionary branches (Kelleher et al. 2018; McBroome et al. 2021; Turakhia et al. 2021). The Pangenome Mutation-Annotated Network (PanMAN) format uses this approach for lossless pangenome representation, capturing point mutations, indels, and structural variants as mutation-annotated genealogies. These genealogies may be single trees for minimally recombining species, or phylogenetic networks for more complex evolutionary histories (Walia et al. 2026). By encoding ancestral sequences at internal nodes, PanMAN achieves superior compression while providing a richer reference structure that includes both sampled diversity and inferred ancestral sequences at the internal nodes in the network. However, fully leveraging such genealogically struc-

tured pangenomes for common tasks like read alignment, phylogenetic placement, and metagenomic inference requires new methods to efficiently index and utilize mutation-annotated phylogenetic networks as a reference object.

Here we present Panmap, a tool that indexes k-mers in PanMANs to perform phylogenetic placement, read alignment, genotyping, and mixed sample haplotype deconvolution from sequencing reads. We apply it to pathogen genome assembly, wastewater metagenomic surveillance, and phylogenetic placement of ancient environmental DNA, demonstrating equivalent or better performance while scaling to reference pangenomes intractable for existing tools. Panmap is available through Bioconda and Docker, with source code and documentation at <https://github.com/amkram/panmap>.

4.2 Results

4.2.1 Panmap overview

Panmap builds a compact k-mer index from the sequence content and phylogenetic structure of a PanMAN, supporting placement, alignment, genotyping, consensus assembly, and metagenomic deconvolution from sequencing reads. It provides two processing modes (Figure 4.1). In single-sample mode, k-mer seeds (see Methods) from all reads are scored jointly across the phylogeny to find a close reference haplotype, which may be a sampled genome or an inferred ancestor. Panmap can then optionally align reads to that sequence and produce a consensus assembly with genotype calls. In metagenomic mode, reads are scored independently, yielding per-read node assignments and haplotype abundance estimates (see Methods).

4.2.2 Phylogenetic compression yields indexes up to 600× smaller than alternatives

Panmap’s efficiency results from its phylogenetically compressed k-mer index. Because closely related genomes share most k-mer seeds, representing genomes in a phylogeny allows Panmap to store only the seed differences along each branch rather than storing seeds independently for every genome. Panmap traverses the entire reference PanMAN phylogeny once and encodes only the seed differences between parent and child nodes. These delta encodings capture changes in k-mer seed presence along each branch, exploiting the similarity among closely related genomes for dramatic compression and extremely efficient computation of the index.

Panmap’s index outperforms alternatives in construction time, index size, and memory required. We compared the Panmap index to IPK/EPIK (Romashchenko et al. 2023), which builds phylo-k-mer databases for phylogenetic placement, and VG Giraffe (Garrison et al. 2018; Sirén et al. 2021), which indexes pangenome graphs for read alignment (Table 4.1, Table S4.2). For 4,000 respiratory syncytial virus (RSV) genomes, Panmap’s 5.7 MB index built in 4 seconds, versus VG Giraffe’s 3.5 GB (614× larger) in 6 hours (5,580× slower) and IPK/EPIK’s 2.1 GB (368× larger) in 3 hours (1,005× slower). For 20,000 SARS-CoV-2 genomes, Panmap’s 7.3 MB index was 614× smaller than VG Giraffe’s 4.5 GB, and built in 3 seconds versus 7.5 hours. For *Mycobacterium tuberculosis*, Panmap indexed the 400-genome pangenome about 520× faster than VG, but the compression advantage was less pronounced (240 MB vs 500 MB, 2.1× smaller), likely because the *M. tuberculosis* genomes included in the PanMAN are not as closely related as for the other pathogens considered. IPK/EPIK failed to index both the SARS-CoV-2 and *M. tuberculosis* pangenomes. We performed

all of the above experiments with one thread each, though Panmap also supports multithreaded indexing and placement (Figure S4.1). These results imply that Panmap will scale favorably as data sets continue to grow in size.

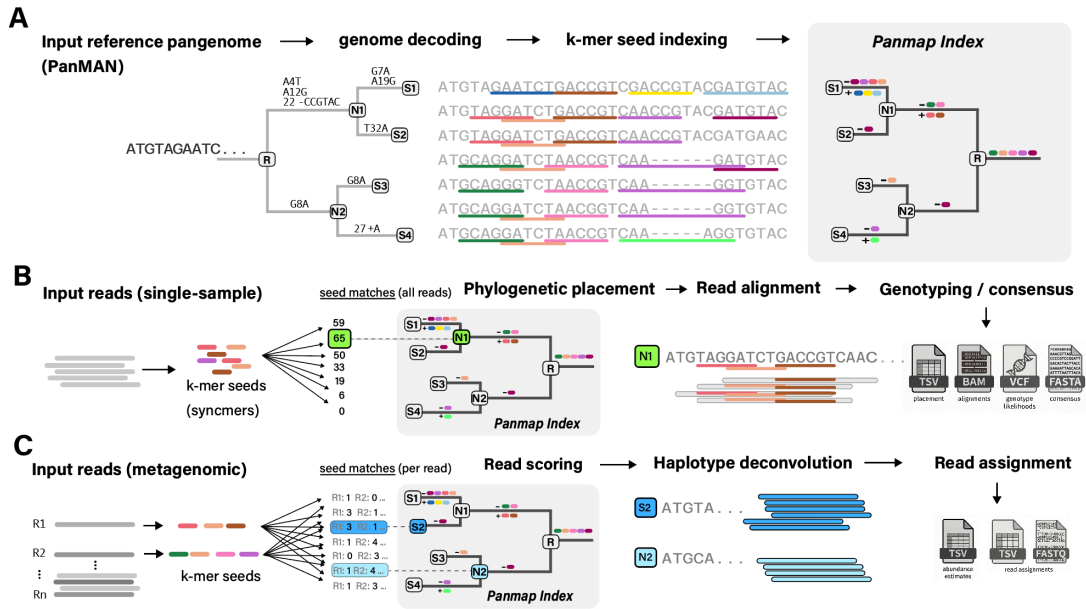


Figure 4.1. Overview of Panmap’s indexing and analysis modes. (A) Pangenome indexing. Panmap takes a reference pangenome in PanMAN format (a mutation-annotated phylogeny encoding SNPs, insertions, and deletions along each branch), decodes the full genome sequence at each node, extracts k-mer seeds, and stores only the seed differences between parent and child nodes in a phylogenetically compressed index. (B) Single-sample mode. Input reads are decomposed into k-mer seeds, and Panmap scores all seeds collectively against every node in the index during a single tree traversal to identify the best-matching reference (here, node N1 with 65 seed matches). Raw counts are shown, though Panmap uses a containment metric in practice; see Methods). Reads are then aligned to the placed node’s genome sequence (decoded from the PanMAN), and Panmap produces a placement report (TSV), read alignments (BAM), genotype likelihoods (VCF), and a consensus assembly (FASTA). Each stage is optional: for example, Panmap can report only the phylogenetic placement without proceeding to alignment or assembly. (C) Metagenomic mode. Each read is scored independently against the index, and per-read scores are used to identify the reference haplotypes present in a mixed sample (here, S2 and N2). Reads are deconvolved by competitive assignment to their best-matching haplotypes, and Panmap estimates relative abundances and outputs abundance estimates (TSV) and per-read assignments (FASTQ).

4.2.3 Panmap enables phylogenetic placement directly from sequencing reads

Phylogenetic placement positions query sequences within existing reference phylogenies, enabling sample identification and phylogenetic profiling without *de novo* tree reconstruction (Czech et al. 2022). Most existing approaches require some form of preprocessing before placement. Alignment-based methods, such as EPA-ng (Barbera et al. 2019) and pplacer (Matsen et al. 2010), require queries to be pre-aligned to a reference multiple sequence alignment. EPIK (Romashchenko et al. 2023) enables alignment-free placement using phylo-k-mers but still requires assembled query sequences. USHER (Turakhia et al. 2021) achieves rapid maximum-parsimony placement on mutation-annotated trees but requires VCF input derived from prior variant calling. For ancient DNA, pathPhynder (Martiniano et al. 2022) places samples onto phylogenies from pre-aligned BAM files, while euka (Vogel et al. 2023) and Soibean (Vogel et al. 2024) use VG-format pangenome graphs for taxonomic detection and species-level identification from ancient environmental DNA, but are restricted to curated mitochondrial references for select taxa.

Table 4.1. For three pathogen species, we indexed pangenomes with Panmap, VG Giraffe, and IPK/EPIK. We measured the construction time, memory usage, and size after GZIP of the indexes. The Panmap index records k-mer seed differences across all genomes and is used for phylogenetic placement and abundance estimation. For read alignment to a pangenome graph, VG Giraffe indexes k-mers (minimizers), a graph Burrows-Wheeler transform (GBWT), and node distances. IPK builds the phylo-k-mer index used by EPIK for phylogenetic placement. All programs were run with one CPU core on a dedicated computing cluster node and allowed up to 1 TB of runtime memory. We built all Panmap indexes in single-sample mode with default parameters (k=19, s=8, l=3) for all species except *M. tuberculosis*, for which we used k=51, s=8, l=1. Metagenomic mode's index is nearly identical but also includes seed orientations, which incurs a small size overhead (Table S4.1).

Species (# genomes)	Format	Input size	Tool	Index size	vs. Panmap	Index time (1 thread)	Peak memory
RSV (4,000)	PanMAN	709 KB	Panmap	5.7 MB	1×	4 s	207.6 MB
SARS-CoV-2 (20,000)	PanMAN	2.3 MB	Panmap	9.4 MB	1×	8 s	416.6 MB
SARS-CoV-2 (8 million)	PanMAN	366 MB	Panmap	2.5 GB	1×	1 hr 20 m	135.2 GB
<i>M. tuberculosis</i> (400)	PanMAN	5.1 MB	Panmap	240 MB	1×	2 m 51 s	17.7 GB
RSV (4,000)	GFA	3.0 MB	VG Giraffe	3.5 GB	614×	6 h 12 m	359.0 GB
SARS-CoV-2 (20,000)	GFA	11 MB	VG Giraffe	4.5 GB	479×	7 h 39 m	374.7 GB
<i>M. tuberculosis</i> (400)	GFA	93 MB	VG Giraffe	500 MB	2.1×	24 h 43 m	178.4 GB
RSV (4,000)	FASTA MSA	9.6 MB	IPK / EPIK	2.1 GB	368×	1 h 7 m	115.4 GB
SARS-CoV-2 (20,000)	FASTA MSA	120 MB	IPK / EPIK	(failed during branch optimization)			
<i>M. tuberculosis</i> (400)	FASTA MSA	511 MB	IPK / EPIK	(exceeded 1 TB RAM limit)			

Because all of these methods depend on a prior alignment, assembly, or variant calling step using one or multiple references, reads that map poorly to those references may

be discarded before any phylogenetic signal can be extracted. Panmap instead matches k-mer seeds directly against the entire pangenome, so divergent, short, or damaged reads can still contribute to placement without requiring a complete alignment (see Methods).

4.2.4 Panmap achieves high placement accuracy across coverage levels

Phylogenetic placement using Panmap's single-sample mode is fast and sensitive for simulated and real sequencing data, remaining accurate even for samples with very low coverage depth (Figure 4.2). For simulation experiments, we selected 70 leaf nodes at random in each PanMAN, applied random SNPs and indels at two mutation rates to generate new descendant genomes, simulated short reads based on each novel genome, then placed the collection of reads for that sample with Panmap. For real data experiments, we chose up to 100 PanMAN leaf nodes per species, downloaded corresponding FASTQ data from the Sequence Read Archive, then removed each chosen node from the PanMAN before placing the reads with Panmap.

For both experiments we measured accuracy relative to the sample's known position in the reference tree. The parsimony cost of a placement is the number of mutations (substitutions and indels) separating the sample from the node it is placed on. Our reported accuracy score is the parsimony cost of Panmap's placement minus the cost at the original position: a positive score means Panmap's placement requires that many more mutations, zero means a tie, and a negative score means a closer (more parsimonious) fit than the original. For ease of interpretation we clamp the score to $\max(0, \text{score})$, collapsing every equal-or-better placement to zero. A negative score before clamping may therefore reflect minor errors in the reference tree's topology.

For SARS-CoV-2, RSV, and *M. tuberculosis*, Panmap achieved accurate placement across all coverage levels. On simulated SARS-CoV-2 data (n=140, two mutation rates), placement accuracy was stable across coverages, with a median score of 3–5 mutations from the true phylogenetic position. For RSV, simulated placement (n=140) achieved a median score of 2–3 across all coverages, and real RSV data (n=20) had a median score of 4–5. Phylogenetic placement of individual SARS-CoV-2 and RSV samples completed in 0.3–0.7 seconds across all coverage levels, using 148–183 MB of RAM. For the 400-genome *M. tuberculosis* PanMAN, placement on real (n=15) and simulated data (n=140) achieved a median score of 0–2 mutations.

Panmap achieves accurate placement even with very short reads. On simulated 100× coverage data, RSV and SARS-CoV-2 placement accuracy was stable at default parameters (k=19, s=8, l=3) across all read lengths down to 40 bp (Figure S4.2). For RSV, varying the linked syncmer parameter (l=1, 2, 3, 4) had minimal effect, but for SARS-CoV-2, where the larger pangenome presents greater ambiguity, linked syncmers (l > 1) substantially improved placement at all read lengths (Figure S4.2, right). This demonstrates that chaining syncmers into linked seeds (Panmap defaults to l=3) can provide additional specificity for accurate placement on larger, more diverse pangenomes, even when individual reads are shorter than typical Illumina output.

We additionally benchmarked placement and assembly with Panmap against a PanMAN containing 8 million SARS-CoV-2 genomes. We first indexed the PanMAN with Panmap using 16 threads, which took 11 min 34 s with peak memory usage of 115 GB. We then placed a 100× coverage sample on the PanMAN, which took 64 seconds using 1 thread (peak memory 17.6 GB) and 23 seconds using 16 threads (peak memory

17.6 GB). We then assembled the same sample by running the full Panmap algorithm rather than stopping after placement. This mode requires loading the full PanMAN into memory rather than just the index and took 223 seconds (peak memory 65.6 GB) using 1 thread and 173 seconds (peak memory 65.7 GB) using 16 threads. Because index loading is parallelized, additional threads speed up single-sample placement substantially; for full consensus generation the runtime is dominated by loading the full PanMAN into memory, so additional threads provide more modest benefits.. For large PanMANs, we therefore recommend processing samples in batch mode (`--batch`) to amortize this overhead. Using 16 threads, Panmap placed a batch of 70 SARS-CoV-2 samples at 100× coverage in 201 seconds (~20.9 samples per minute), with a peak RAM usage of 21.6 GB.

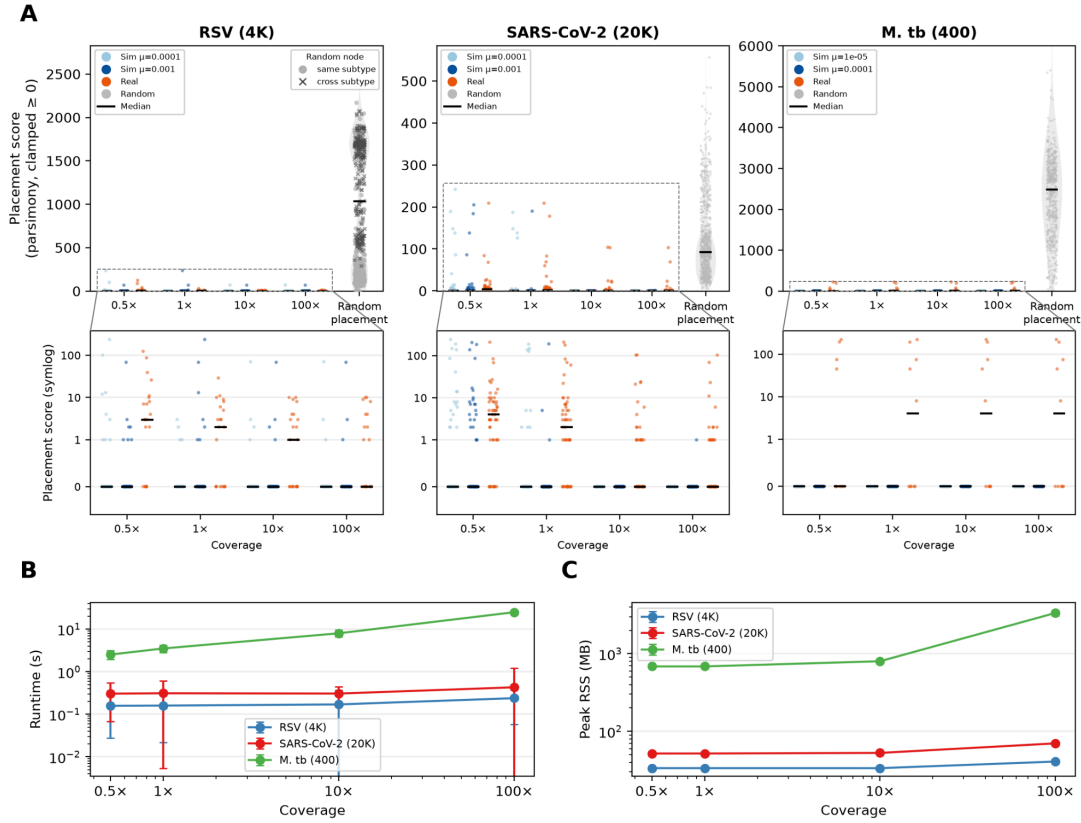


Figure 4.2. Phylogenetic placement accuracy and performance of Panmap on simulated and real sequencing data. (A) Placement accuracy. For each of RSV (4,000-genome PanMAN), SARS-CoV-2 (20,000-genome PanMAN), and *M. tuberculosis* (400-genome PanMAN), we selected 70 leaf nodes, applied random mutations at two species-specific mutation rates to generate descendant genomes, simulated 150 bp paired-end reads at four estimated coverage levels (0.5 $\times$, 1 $\times$, 10 $\times$, 100 $\times$), removed each node from the tree, and placed the reads with Panmap (70 unique samples $\times$ 2 mutation rates $\times$ 4 coverages = 560 placements per species). We performed the same experiments with real data. Violin plots show the distribution of parent-relative placement scores; a score of zero indicates placement at least as parsimonious as the original position. The gray violins show the distributions of random placements for comparison (the multimodal distribution of RSV random placements arises because the reference PanMAN contains two divergent RSV subtypes, RSV-A and RSV-B). Under Poisson sampling, estimated coverage levels of 0.5 $\times$, 1 $\times$, 10 $\times$, and 100 $\times$ correspond to approximately 39%, 63%, $\sim$ 100%, and $\sim$ 100% of the genome covered by at least one read, respectively. (B) and (C) show runtime (wall-clock seconds) and peak memory usage as a function of coverage, colored by species.

4.2.5 Phylogenetic placement enables scalable pangenome-guided read alignment

Read alignment to reference sequences is a foundational task in genomic analysis (Reinert et al. 2015; Kim et al. 2020; Alser et al. 2021). In single-sample mode, after k-mer based phylogenetic placement, Panmap aligns all input reads to the placement taxon's genome with the Minimap2 library (Li et al. 2018). The default alignment mode is fast and achieves high mapping rates in most cases. We also provide an optional more sensitive alignment mode that uses classic ancient DNA parameters for bwa-aln (Li and Durbin 2009), which performs well when the input sequences are extremely short (<70bp) or highly damaged, as is typical of ancient DNA (Oliva et al. 2021). For applications where a full graph representation is unnecessary, aligning to the single placement sequence substantially accelerates pangenome-based read alignment. While pangenome graphs can accommodate tens or hundreds of large genomes (e.g., human), they struggle with thousands or millions of samples, even for species with small genomes (Walia et al. 2026).

We compared Panmap to VG Giraffe for read mapping on large pathogen pangenomes and found graph-based mapping computationally challenging at this scale (Table S4.2). For example, after a 7h 39m indexing step, VG Giraffe took 26 minutes and 21 GB of memory to map 25 150bp reads onto a graph of 20,000 SARS-CoV-2 genomes using 25 threads. We also evaluated VG Giraffe's haplotype sampling mode (Sirén et al. 2024), which improves mapping speed by restricting alignment to a set of 4 synthetic haplotypes (default setting) but still requires the 7h 39m indexing step. Using 1 thread, VG Giraffe completed haplotype sampling and alignment of 19,934 simulated SARS-CoV-2 reads (~100× coverage) in 1m 27s, with peak memory usage

of 470.9 MB. Using the 20,000-genome PanMAN, Panmap completed the same task in 1.42 seconds and 135 MB of memory, after a 3 second indexing step.

4.2.6 Panmap assembles higher quality genomes than single-reference methods

Genome assembly with Panmap is more accurate than single-reference-guided methods for RSV, SARS-CoV-2, and *M. tuberculosis*, especially at very low depths of coverage (Figure 4.3). We evaluated Panmap’s accuracy using real sequencing reads from public SRA datasets with reference pangenomes containing 4,000 RSV, 20,000 SARS-CoV-2, and 400 *M. tuberculosis* genomes. For each sample, we performed leave-one-out evaluation: the sample’s known genome was removed from the pangenome before placement and assembly. Ground truth sequences were obtained directly from the PanMAN leaf nodes (the original GenBank submissions). We sub-sampled reads to coverages ranging from 0.5× to 100× and assembled each sample with Panmap and separately with a field standard single-reference consensus assembly workflow for each species: BWA-MEM + iVar consensus (Li 2013; Grubaugh et al. 2019) for RSV and SARS-CoV-2, and Clockwork (Hunt et al. 2022) for *M. tuberculosis*. We measured accuracy as the fraction of the ground truth genome (excluding 150 bp at each end) correctly reconstructed by each method. At 0.5× coverage, Panmap correctly genotyped a median of 91% of the RSV genome compared to 15% for BWA + iVar, which can only call bases at positions with read support. At 100× coverage, Panmap reaches 97% genotyping accuracy while BWA + iVar plateaus at 82%, because the standard RSV-A (NC_038235.1) and RSV-B (NC_001781.1) references are too divergent from contemporary circulating strains to achieve complete read mapping. For SARS-CoV-2, both methods converge near 100% at 100× coverage, but Panmap

maintains >99% accuracy even at 0.5× depth where BWA + iVar covers only ~30% of the genome. For *M. tuberculosis*, both Panmap and Clockwork achieve >99.5% median accuracy at all coverage levels, reflecting the species' lower genetic diversity (more samples are similar to the standard reference genome). However, Panmap achieves slightly better accuracy (median >99.8%) in a fraction of the runtime.

Panmap's advantage for genome assembly stems from pangenome-guided reference selection: by using phylogenetic placement to choose from hundreds or thousands of sequences, reads map more completely, and uncovered positions inherit bases from a near-identical reference rather than being left as ambiguous. Across the full pangenome reference sets for RSV and SARS-CoV-2, Panmap selects a reference that yields more accurate genotyping than the MinHash-based method Mash Screen, in part because it can choose an inferred ancestral node rather than the nearest single leaf; for *M. tuberculosis*, whose genome is much larger than the viral species, both reference selection methods performed similarly (Figure S4.3).

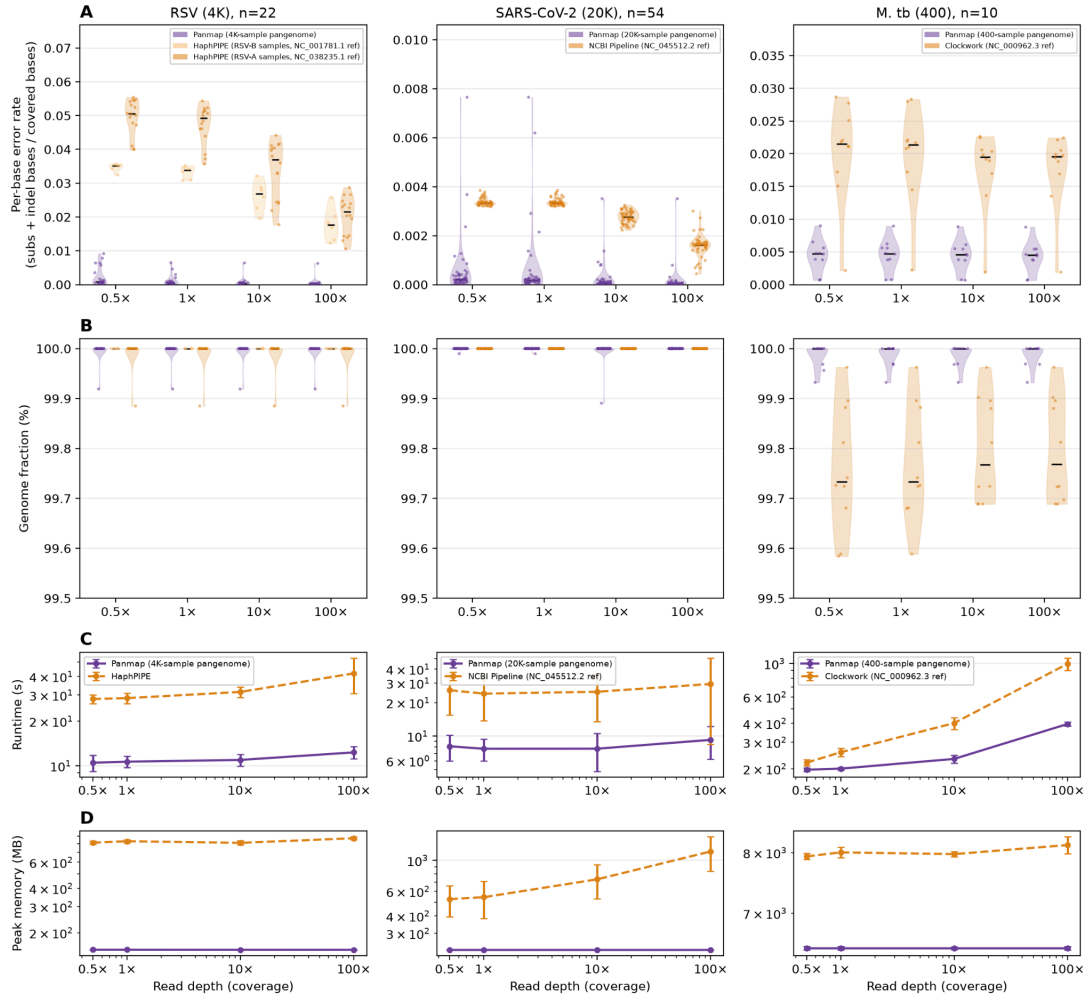


Figure 4.3. Consensus assembly accuracy and performance of Panmap versus field-standard single-reference workflows on real sequencing data. We subsampled real Illumina reads to four depths (0.5–100×) and assembled each with Panmap against the full PanMAN and with the species’ field-standard workflow: RSV (4,000-genome PanMAN vs. HaphPIPE with subtype-matched NC_038235.1/RSV-A or LR699737.1/RSV-B), SARS-CoV-2 (20,000-genome PanMAN vs. the NCBI SARS-CoV-2 Variant Calling pipeline with NC_045512.2), and *M. tuberculosis* (400-genome PanMAN vs. Clockwork with NC_000962.3). For each sample, we removed its own genome and any identical copies from the PanMAN, then scored the consensus against that sample’s GenBank genome, ignoring 250 bp at each end; n is the number of samples per species. (A) Per-base error rate = (substitutions + inserted + deleted bases) / covered bases. (B) Fraction of ground-truth genome reconstructed. (C) Single-threaded runtime and (D) peak memory versus coverage (median ± SD).

4.2.7 Panmap resolves haplotype mixtures across diverse viral species

In metagenomic mode, Panmap accurately estimates haplotype abundances in simulated and real SARS-CoV-2 mixture samples, even when haplotypes carry variants absent from the reference tree. We evaluated metagenomic Panmap using the published SARS-CoV-2 PanMAN containing 20,000 SARS-CoV-2 sequences spanning 1,000 Pango lineages, as described in Walia et al. 2026. We simulated read mixtures with 1,000 $\times$ coverage from up to 10 sampled haplotypes, each haplotype containing up to 20 introduced SNPs. We evaluated Panmap on the simulated samples and compared estimated haplotype abundances to the truth with a 3 bp difference tolerance (see methods) (Figure 4.4A). Abundance estimates closely matched their simulated proportions across all mutation levels, with root mean square error (RMSE) values of 0.009, 0.011, 0.011, 0.020, and 0.036 for samples containing 0, 1, 5, 10, and 20 SNPs per haplotype, respectively. As expected, increasing mutation levels led to more haplotypes not in the truth set to be estimated to be present (“others” category), with cumulative abundances of 0.27%, 0.60%, 0.95%, 2.03%, and 3.68% across the 40 replicates at each SNP level, reflecting the added noise introduced by unaccounted mutations. Despite this trend, false positives remained low in most scenarios, and true haplotypes were accurately estimated even in samples containing 200 total unaccounted mutations, implying that our method will be applicable for populations that are less densely sampled than major human pathogens. Simulations conducted on an RSV PanMAN with 4,000 samples and an HIV PanMAN with 20,000 samples yielded concordant results (Figure S4.4). For HIV, we demonstrated that our method can be applied to more divergent species that present challenges for single-reference based

approaches, such as Freyja (Karthikeyan et al. 2022), WEPP (Gangwar et al. 2025), and the method described in Pipes et al. 2022. We compared Panmap specifically to WEPP as it has been shown to be the most accurate and generalizable among single-reference based tools (Gangwar et al. 2025). As WEPP requires an UShER-format mutation-annotated tree (MAT), we constructed an HIV MAT following the viral_usher pipeline (Hinrichs) using the same 20,000 samples as the published HIV PanMAN and using K03455.1 as the root. Our method achieved substantially lower weighted haplotype and peak distances than WEPP (Figure S4.5). While WEPP relies on a single root reference and is susceptible to information loss specifically through alignment failure of highly divergent reads, our approach utilizes the entire pangenomic reference.

4.2.8 Panmap outperforms existing tools on real wastewater surveillance data

We next evaluated metagenomic Panmap’s performance on 22 real SARS-CoV-2 wastewater samples collected at Point Loma between November 27, 2021, and February 7, 2022, sequenced using SARS-CoV-2 targeted tiled amplicons (Karthikeyan et al. 2022). We evaluated our estimated lineage proportions by comparing them to those derived from clinical sequences in San Diego with a 7-day temporal shift as a proxy for ground truth (following Karthikeyan et al. and Gangwar et al.). This window was selected for its high epidemic surveillance efforts and high sequence submission rates. We used the same published SARS-CoV-2 PanMAN described above for Panmap, and a subsampled SARS-CoV-2 UShER-format MAT containing the same 20,000 sequences as the PanMAN for WEPP. Compared with WEPP and Freyja, Panmap achieved higher accuracy than Freyja and comparable accuracy to

WEPP (Figures 4.4B, S4.6, S4.7), while demonstrating a >10X speedup in runtime (Figure 4.4C).

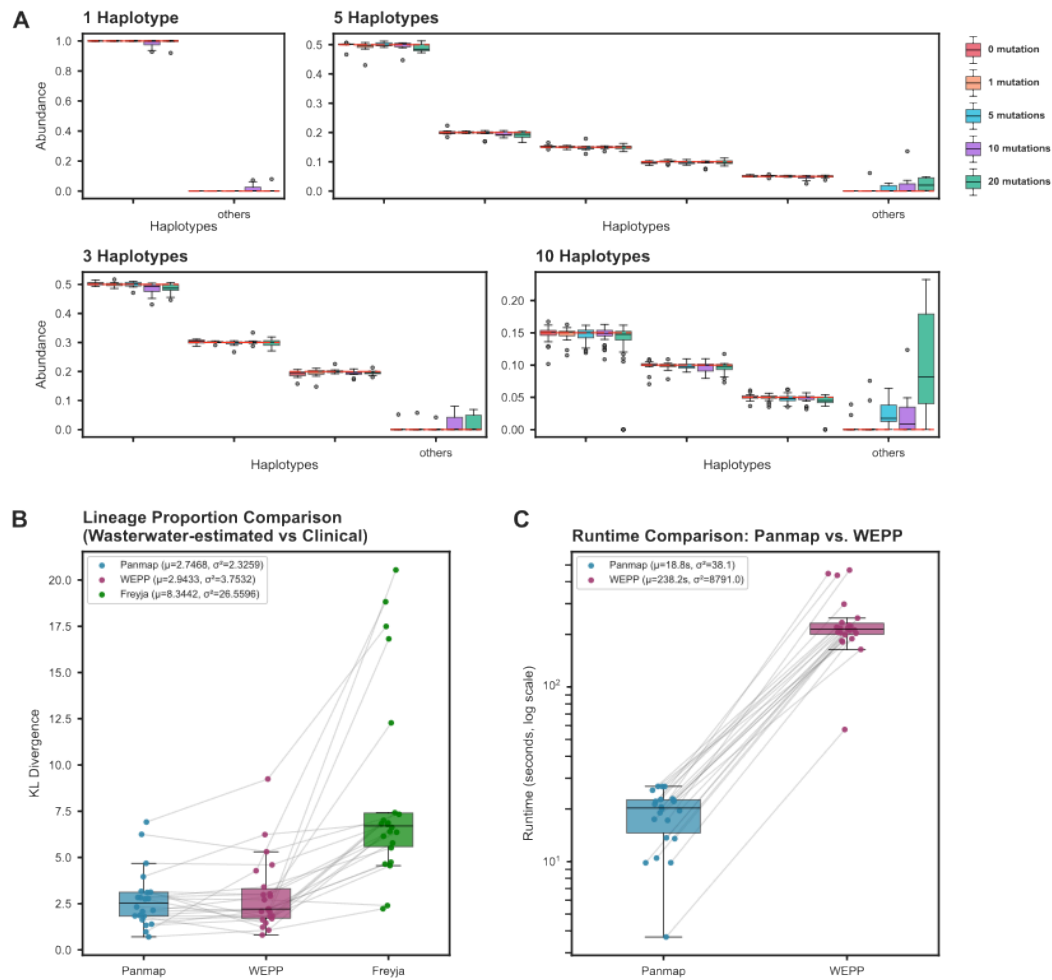


Figure 4.4. Evaluation of abundance estimation in simulated and wastewater SARS-CoV-2 mixture samples. (A) Estimated haplotype abundance in simulated SARS-CoV-2 mixture samples with 1, 3, 5, or 10 haplotypes, introducing 0, 1, 5, 10, or 20 single-nucleotide polymorphisms (SNPs) per haplotype. Ten replicates were simulated per composition and mutation level. Red dashed lines denote true simulated abundances; the ‘others’ category represents accumulated false-positive haplotype abundance. (*10-haplotype samples: 4 haplotypes at 15%, 2 at 10%, 4 at 5% abundance) (B) Kullback-Leibler (KL) divergences between lineage proportions estimated by Panmap, WEPP, and Freyja from 22 wastewater samples, compared to one-week-delayed clinically observed lineage proportions. (C) Runtime comparison between Panmap and WEPP in seconds. Gray lines connect results from the same wastewater sample.

4.2.9 Application to ancient DNA phylogenetic analysis

Panmap’s metagenomic mode demonstrates utility for ancient environmental DNA (aeDNA) analysis. We constructed a vertebrate mitochondrial genome pangenome containing 15,655 sequences representing 7,893 unique species and indexed it with Panmap (see Methods). To assign eDNA reads to specific taxa, we implemented a competitive mapping approach in Panmap that filters low-confidence mappings, uneven coverage, and phylogenetically ambiguous reads matching multiple distantly related taxa (see Methods). This method is similar to the LCA-based workflows commonly used in aeDNA studies (Wang et al. 2021; Kjær et al. 2022; Wang et al. 2022; De Sanctis et al. 2025). Unlike those methods, which map against multi-terabyte reference databases, Panmap maps against a compressed mitochondrial pangenome and therefore assigns only mitochondrial reads. This yields a much smaller index, and we demonstrate this enables rapid taxonomic and hotspot identification. The resulting reduced set of assigned reads can be additionally validated downstream through more computationally demanding methods such as full-read alignment and BLAST (Altschul et al. 1990).

With these implementations, Panmap filters and assigns reads and read clusters; as described below, these assignments recover phylogenetic signals concordant with the original studies. We applied this approach to two recently published ancient sedimentary DNA datasets: circumpolar ancient permafrost DNA (Wang et al. 2021; Miller and Simpson 2022) and 2-million-year-old ancient sediment DNA (Kjær et al. 2022). Although our phylogenetic placement evaluated all vertebrates, we focused on the Elephantidae family for both studies, as genera within this family were found in

both previous studies, and are of high ecological interest (Wang et al. 2021; Miller and Simpson 2022).

For the Wang et al. (2021) data, we obtained adapter-trimmed reads and placement results directly from the authors and performed quality control on the FASTQ files before applying our method (see methods). In accordance with the original paper's findings, we identified a clear read assignment hotspot in the Elephantidae clade, specifically at the *Mammuthus* nodes (shown across all Wang et al. 2021 samples in Figure 4.5A). Panmap processed ~9 billion reads in the original 651 samples in 8 minutes using 16 threads, and assigned a total of 4,098 reads to nodes within the Elephantidae family, more reads than we assigned to any other families. Of these, 3,803 (~93%) were alignable to the *Mammuthus Primigenius* mitochondrial reference genome using bwa-aln with ancient-DNA parameters, about 4.5 times as many as the 846 mammoth mitochondrial reads that Wang et al. (2021) identified using the traditional competitive mapping approach. There are several reasons for this. First, Panmap compares each read against many closely related references simultaneously, including inferred ancestral sequences at internal nodes rather than only observed genomes. Second, only nodes with maximal placement score to a read are considered for the taxa LCA filter. As expected, because Panmap was applied using mitochondrial reference sequences, fewer overall reads were assigned than in whole-genome reference approaches. In the same dataset, the original study assigned ~332,000 reads to the elephantid whole genomes, compared with the 3,803 elephantid reads Panmap recovered.

To be conservative, we excluded the remaining ~7% of the reads that were not alignable from all downstream analyses. The retained reads were further validated to be Elephantidae (see Methods) . Approximately 93% had exclusive BLAST hits to Elephantidae mitochondria, with sequence divergence and damage patterns similar to those identified by Wang et al. (2021) (Figure S4.8, Table S4.6). Finally, for each Wang et al. 2021 sample separately, we placed the 3,803 reads that are alignable to the *M. primigenius* mitochondrial genome to a *Mammuthus*-specific PanMAN containing 78 *Mammuthus* mitochondrial genomes. This allowed direct comparison to the original per-sample pathPhynder (Martiniano et al. 2022) placement results. Panmap achieved on average more than 2 branches deeper placement than pathPhynder, yielding more informative placements overall (Figure S4.9, Table S4.5).

For the 2-million-year-old sample from Kjaer et al., we downloaded the published raw data, performed quality control, and applied our method as above by placing reads from all samples combined into our vertebrate PanMAN. Our placement shows a concordant cluster of reads assigned to the mastodon node within the Elephantidae and Mammutidae families (Figure 4.5B). Of the 755 reads that Panmap assigned to this clade, 233 were alignable to the *Mammut americanum* mitochondrial reference, and, as expected, most of these were assigned only to *M. americanum* nodes. Few reads were alignable to the other Elephantidae mitogenomes, which is consistent with the weak, background-level signals. The original Kjaer et al. study noted that the mastodon placement was supported by only two single-nucleotide polymorphisms (SNPs), and this was later confirmed in further work (Vogel et al. 2024). Our analysis provides further evidence supporting this placement.

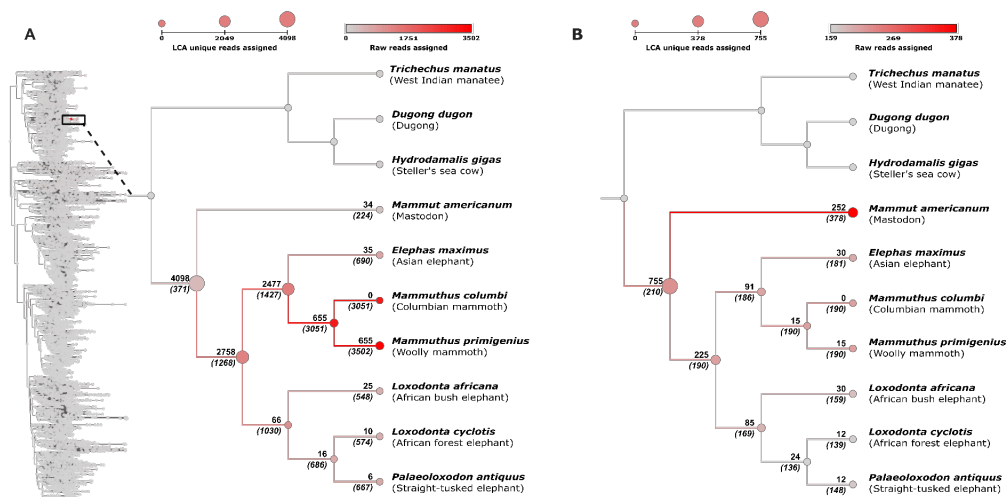


Figure 4.5. Number of reads assigned to nodes for two datasets processed by Panmap. Note that reads may be assigned to more than one node within a family. Numbers above branches indicate LCA counts, or the total number of reads assigned only to that node or any node below it, i.e. not including any reads which are also assigned elsewhere. Numbers in parentheses below branches indicate the count of raw reads assigned directly to the node (which may also be assigned elsewhere). Node color scales with raw read count, and node size scales with LCA read count. (A) Read assignments for the Wang et al. dataset. The full vertebrate mitochondrial PanMAN is shown on the left, with nodes colored by raw read assignment count; the highlighted Elephantidae clade is expanded on the right. (B) Read assignments for the Kjær et al. dataset, restricted to the Elephantidae clade.

4.3 Discussion

Panmap addresses a central limitation of existing pangenome tools by enabling phylogenetic placement, genotyping, and assembly against reference pangenomes containing tens of thousands of genomes, a scale at which graph-based and alignment-based approaches become impractical. By incorporating evolutionary information, Panmap improves the resolution and efficiency of studying new samples, achieving better accuracy than single reference-guided genotyping methods for *M. tuberculosis*,

RSV, and SARS-CoV-2. The accuracy gap widens at low coverage, where a poorly chosen reference introduces systematic errors that sparse read data cannot correct. We expect this advantage to grow as pangenomes continue to become larger and more densely sampled. Panmap can also place and align reads to inferred ancestral sequences at internal nodes, providing suitable references even for samples that diverge from all observed genomes in the pangenome.

Panmap consistently produced smaller indexes than similar methods that index pangenomes. Panmap's indexes for the SARS-CoV-2 and RSV pangenomes were hundreds or thousands of times smaller than VG Giraffe's graph-based indexes. However, Panmap's index was only about 30% smaller than VG Giraffe's for *Mycobacterium tuberculosis*. One reason for this is that the pangenome has fewer, though larger, genomes, making Panmap's compression advantage less apparent than in the viral pangenomes comprising thousands of samples. Another reason is that the pangenome has a more complex (Table S4.4) or simply longer evolutionary history, reducing its phylogenetic compressibility. As Panmap builds on PanMAN and its concept of "evolutionary compression", the compression advantage of Panmap's k-mer index is greatest when the overall sequence diversity is minimized and the phylogenetic relatedness is maximized in the reference pangenome, and more evident as the number of individual genomes increases. However, Panmap is likewise currently restricted to encode within-species microbial genomes and low-recombination eukaryotic genomic systems, such as mitochondria.

Panmap's metagenomic mode broadens its utility to mixed-sample haplotype deconvolution from bulk sequencing data, addressing a long-standing bottleneck in

population and microbial genomics.. In simulated SARS-CoV-2 mixtures, abundance estimates remained accurate even with constituent haplotypes carrying up to 20 unobserved single-nucleotide polymorphisms (SNPs), demonstrating robustness against incomplete reference pangenome representation. This robustness is crucial for surveillance of under-sampled pathogen populations, where reference phylogenies often lag behind circulating diversity. Concordant results across RSV and HIV pangenomes further support this generalizability, particularly for HIV where Panmap handles the elevated sequence divergence that causes single reference-based methods, such as WEPP, to lose accuracy. On real wastewater data, Panmap achieved KL divergence comparable to WEPP in estimating clinically derived lineage proportions (Karthikeyan et al. 2022; Gangwar et al. 2025), while running 10X faster than WEPP. This favorable accuracy-speed trade-off supports its utility for high-throughput wastewater-based epidemiology. Although off-target reads are negligible in the amplicon-targeted samples analyzed here because of the additional preprocessing, shotgun metagenomic samples can contain a large fraction of off-target reads, which Panmap's `--discard` option removes before abundance estimation (see Methods). This has minor effects on the resulting estimates, though not strictly zero, as highly divergent but informative reads may occasionally be discarded while some off-target reads may exceed the threshold by chance and be retained.

The ancient eDNA application demonstrates that Panmap's metagenomic framework generalizes beyond within-species haplotype deconvolution to cross-species taxonomic assignment against a mitochondrial pangenome tree. By competitively mapping reads against a vertebrate mitochondrial pangenome of nearly 16,000

sequences, Panmap recovered substantially more high-confidence Elephantidae reads than the commonly-used two-step ancient environmental DNA competitive mapping and phylogenetic placement approach, while producing placements concordant with published findings (Wang et al. 2021; Kjær et al. 2022). Panmap scores and assigns reads without a separate alignment step and uses a compact mitochondrial pangenome rather than a multi-terabyte reference database, offering a practical, rapid alternative to LCA-based workflows for the taxonomic screening of degraded and low-complexity ancient DNA. One current limitation is that the PanMAN and Panmap index used here are mitochondrial genomes only, so Panmap assigns only mitochondrial reads; for analyses that additionally require nuclear data, however, it can serve as a fast pre-filtering step that identifies taxonomic hotspots to target subsequent, more in-depth analyses. Extending Panmap to directly support whole-genome reference databases represents an important direction for future development.

Panmap's phylogenetic approach to pangenomic analysis offers advantages over both single-reference and graph-based methods: it scales to millions of genomes, produces highly compressed indexes, and leverages evolutionary relationships to improve both placement and genotyping accuracy. For metagenomic applications, the same phylogenetic framework enables haplotype deconvolution and abundance estimation with reduced reference size and computational cost. These capabilities make Panmap particularly well-suited for real-time pathogen lineage surveillance, where rapid turnaround and scalability are essential, as well as for analyses of ancient and environmental DNA through conserved organelle- or marker-gene-based taxonomic

assignment, such as the mitochondrial pangenome demonstrated here or other conserved loci including 16s rRNA and chloroplast genomes.

4.4 Methods

4.4.1 Overview of Methods

Panmap accepts two inputs: a reference pangenome in PanMAN format (Walia et al. 2026) and sequencing reads in FASTQ format. The PanMAN encodes both a phylogeny and genomic variation, including point mutations, insertions, deletions, and structural variants. PanMANs may be constructed from multiple sequence alignments using `panmanUtils` (<https://turakhia.ucsd.edu/panman>). Panmap supports single or paired-end reads, and is optimized for short reads.

Panmap has two modes. Single-sample mode assumes all reads originate from one haplotype; it jointly scores reads to identify an optimal reference, aligns reads to that reference, and produces a genotyped consensus assembly. Metagenomic mode handles mixed samples containing multiple haplotypes; it scores each read independently against every node in the phylogeny and then either infers the constituent haplotypes and their relative abundances from the score distribution, or applies competitive mapping filters for eDNA read assignment. The following sections describe components shared by both modes, followed by mode-specific methods and experiment details.

4.4.2 Reference PanMAN structure and genome decoding

A PanMAN losslessly encodes the full sequence of every sample and every inferred ancestor as a phylogeny annotated with mutations along its branches. Any of these sequences can be decoded by applying the encoded nucleotide-level mutations (point mutations and short indels) and block-level mutations (larger structural variants, where homologous blocks may be gained, lost, or inverted) along the branches from the root to that sequence's node. . Panmap decodes each node's genome during indexing by performing a single depth-first traversal of the phylogeny, maintaining one genome in memory, along with its block presence and orientation state, and incrementally updating it at each node. This traversal-based approach enables rapid indexing of pangenomes containing many thousands of sequences.

4.4.3 Seed sketches of genome sequences

For sample placement and read alignment, we consider a subset of k-mers, called seeds, from each genome or haplotype in a PanMAN. Retained seeds for each sequence are chosen under a syncmer scheme (Edgar 2021) so that the similarity between haplotypes can be approximated by the similarity of their seed sketches. Syncmer seeds are parameterized by two values, k and s , selecting as a seed any k -mer whose lexicographically smallest s -mer begins at either the start or end of the k -mer. In practice, we order k -mers and s -mers by a hashed integer value rather than their lexicographic ordering. User-supplied values of k and s , or default values $k=19$ and $s=8$, determine the syncmer parameterization and may be tuned to affect the density and length of sampled seeds. Following the minimizer-based approach of (Ekim et al. 2023) we also define l -linked syncmers: ordered tuples of l consecutive syncmers treated as a single matching unit. Linked syncmers represent longer genomic regions

than individual syncmers while remaining robust to local variation. Panmap uses 3-linked syncmers ($l=3$) by default.

4.4.4 Indexing k -mer seeds

We compactly index the presence and hashed sequences of all seeds (linked syncmers) in the reference pangenome by compressing them according to phylogenetic structure. Because similar sequences are expected to share many seeds, we use the phylogenetic tree to annotate seeds only at the node where they first appear. During the depth-first traversal described above, we maintain and update a list of seed starting positions in memory for each node's decoded genome. The index is stored on disk as a depth-first list of nodes and a list of "mutations" for each node that encodes insertions or deletions of hashed seeds relative to the parent node's seed list. We do not store seed positions in the on-disk index, and in single-sample mode we store only the changes in seed frequencies at each node. For metagenomic mode, the index additionally stores seed orientations and runs of gap characters; metagenomic mode indexes are approximately $2\times$ the size of single-sample mode indexes on disk (Table S4.1). At runtime, Panmap loads the index into memory and uses it to reconstruct the seed state at each node.

4.4.5 Indexing optimizations

We employ several additional strategies to make seed indexing faster and compact. During traversal, we expect each taxon to have nearly the same list of seeds as its parent, except for a small number of changes near mutations to the nucleotide sequence. Thus, we only recompute seeds within k bases (the length of a syncmer)

upstream of each nucleotide mutation at a given taxon, retaining the rest of the seeds from the parent. The indexing algorithm is multithreaded by partitioning the depth-first traversal into equal segments that are processed independently and then merged. To write compressed index data to disk and decode it during placement, we use Cap'n Proto (<https://github.com/capnproto/capnproto>) to serialize the index followed by a Zstandard (<https://github.com/facebook/zstd>) compression layer (compression level 7 by default).

4.4.6 Alignment-free phylogenetic placement

In both single-sample and metagenomic modes, Panmap performs phylogenetic placement of query sequences by counting k-mer seeds matches between input reads and a reference pangenome. Because a full read aligner often enforces base-level, end-to-end or edit distance-bounded scoring, it may discard reads whose match to a reference is partial or sub-threshold even when shared k-mers are present (Günther & Nettelblad 2019), an effect that grows for short, divergent reads such as ancient DNA (Dolenz et al. 2024). Panmap's k-mer seed matching relaxes this scoring requirement, so reads that contribute shared k-mers still inform placement.

4.4.7 Mutation spectrum prior for genotype likelihoods

We compute a mutation spectrum prior for a reference pangenome by performing a depth-first traversal of its phylogeny, improving genotyping accuracy by incorporating evolutionary context. During traversal, we count substitutions between each parent and child node, recording their mutation types (e.g., A→C, G→T) in a 4×4 substitution matrix. Substitutions are counted only in aligned regions present in both

parent and child. After traversal, we normalize each count by the total observations of each reference base to obtain substitution probabilities. The mutation spectrum is scaled to an expected branch length (user-specified, or 0.0011 by default, matching minimap2's mutation rate).

4.4.8 Genotype likelihoods and consensus assembly

Panmap computes initial genotype likelihoods internally using bcftools mpileup and bcftools call (`--ploidy 1 -cA`), then applies the mutation spectrum prior for nucleotide substitutions. The posterior probability of genotype g given data D is:

$$P(g \mid D) \propto L(g \mid D) \times P(g \mid r)$$

where $L(g \mid D)$ is the genotype likelihood from bcftools and $P(g \mid r)$ is the prior probability that reference base r mutates to variant base g , derived from the mutation spectrum. For indels, we use unmodified posterior probabilities from bcftools. The consensus sequence is called from the final posterior genotype probabilities.

4.4.9 Phylogenetic placement in Single-sample mode

In single-sample mode, Panmap identifies an optimal reference haplotype by collectively comparing input reads to every sequence in the pangenome. Rather than scoring each read individually (as in metagenomic mode), single-sample Panmap collects k-mer seeds from all reads and compares them to the indexed seeds of each node. A single depth-first traversal of the phylogeny assigns each node a similarity score approximating its sequence identity to the input sample. During traversal, we maintain a running list of seeds at the current node by applying indexed changes to the parent's seed list, avoiding recomputation. The highest-scoring node, whether a

sampled genome at a leaf or an inferred ancestral haplotype at an internal node, is selected as the reference for downstream alignment and genotyping.

Our default similarity metric is a log-scaled containment index. For each node i , we compute:

$$S_i = \frac{\sum_{s \in R \cap G_i} \log(1 + r_s)}{\sum_{s \in R} \log(1 + r_s)}$$

where R is the set of unique seeds extracted from all reads, G_i is the set of seeds present in genome i , and r_s is the frequency of seed s across all reads. This metric measures the fraction of read seed content contained in each genome, with log-scaling so that high-coverage seeds do not dominate the score.

We compute this metric dynamically across all nodes by incrementally updating the numerator using only the seed changes between parent and child nodes stored in the index. Panmap can alternately use several other scoring metrics, including log-scaled cosine similarity and raw seed matches, though containment performed best in our experiments (Figure S4.10).

4.4.10 Read alignment in Single-sample mode

After a suitable reference is chosen, in single-sample mode all input reads are aligned to this sequence using a seed-chain-align approach. Alignment seeds are by default recomputed with smaller k-mer size, or optionally, seeds identified during placement may be passed directly to the chaining and alignment functions of the Minimap2 library (Li et al. 2018). Panmap also supports a bwa-aln (Li and Durbin 2009) mode that improves mapping sensitivity for very short or damaged reads, such as those from ancient DNA samples (Figure 4.3). Because bwa-aln outperforms other

alignment algorithms for ancient DNA (Oliva et al. 2021), we recommend this mode for such samples.

4.4.11 Read scoring in Metagenomic mode

For mixed samples containing multiple haplotypes, Panmap’s metagenomic mode scores each read independently against every node in the phylogeny, rather than jointly scoring all reads. Read scores can be either the raw count of seed hits or a pseudo-chaining score (raw count by default). From the query reads, Panmap generates a seed-to-read hash map, where each key is a seed hash and each value is the list of read indices containing that seed. This enables rapid retrieval and updating of reads affected by a seed change during tree traversal. A child node effectively inherits all read scores unaffected by seed changes from its parent.

We implemented several strategies to further optimize computational efficiency during read scoring:

Read score annotations: To avoid storing read scores for every read against every node, Panmap only records read scores when they are updated at a node. The complete read scores for any node can be reconstructed by traversing from the root to that node and sequentially applying the recorded score changes.

Query de-duplication: Sequencing datasets often contain many identical reads, particularly for amplicon protocols. Panmap collapses exact duplicate reads into unique sequences with associated counts before further processing. Reads are then sketched into seeds, and reads sharing identical seed sketches are further collapsed. This two-stage deduplication substantially reduces the number of reads to be placed and scored.

Seed pruning and tree condensation: Only seeds that are present in both the query and the reference tree contribute to read scoring; therefore, query-only or reference-only seeds can be discarded with no impact. Panmap discards query-only seeds from the query and prunes seed changes of reference-only seeds from the tree. After reference-only seeds are pruned, Panmap further condenses the tree by merging nodes that have no seed mutations post-pruning. This greatly reduces wasteful iterations and lookups, as well as memory required for seed storage.

4.4.12 Abundance estimation in Metagenomic mode

We implemented an expectation–maximization (EM) algorithm accelerated by SQUAREM (Du and Varadhan 2020) internally to estimate the haplotypes present and their proportions (following Pipes et al. 2022). To reduce the search space for the EM algorithm, we select a subset of the N most probable haplotypes (default N=1000) by identifying the nodes with the highest seed overlap coefficients (OC) with the query, where the overlap coefficient between a reference R and a query Q is:

$$\text{OC}(R, Q) = \frac{|S(R) \cap S(Q)|}{\min(|S(R)|, |S(Q)|)} = \frac{|S(R) \cap S(Q)|}{|S(R)|},$$

where $S(R)$ is the set of unique seeds sketched from the reference, and $S(Q)$ is the set of unique seeds sketched from all query reads in the sample. A probability matrix of size (Q, N) , where p_{qn} is the probability of query q being generated from the genome at node n , is estimated by

$$p(q, n) = \varepsilon^{s(q,n)} \times (1 - \varepsilon)^{m(q) - s(q,n)},$$

where ε is the error rate (default: 0.005), $s(q, n)$ is the score of query q against node n , and $m(q)$ is the total number of seeds sketched from query q (i.e. the maximum possible score query q can have).

As a large fraction of reads in a shotgun metagenomic sample come from non-target organisms, Panmap provides a “--discard” option that removes, prior to abundance estimation, any read q whose maximum score across all nodes, normalized by its maximum attainable score $m(q)$, falls below a user-specified ratio (we recommend “--discard 0.9”). This applies the same normalization as the eDNA low-mappability filter ($p(r)$, defined under Ancient environmental DNA placement) to the abundance-estimation mode.

An abundance vector of size N is initialized with uniform proportions for probable haplotypes. After a round of EM converges, haplotypes with an estimated proportion less than a threshold (default to 0.5%) are discarded. An abundance vector for the remaining haplotype is initialized with uniform proportions, and another round of EM is performed. This repeats until all haplotypes are above the threshold proportion after an EM convergence.

4.4.13 Experimental setup and data preparation

To evaluate the performance and accuracy of Panmap, we conducted a series of experiments using both real and simulated genomic datasets. We used reference PanMANs of *M. tuberculosis*, SARS-CoV-2, HIV, and RSV genomes, as well as a vertebrate and *Mammuthus* mitochondrial genomes. We evaluated Panmap on the pathogen pangenomes with real and simulated sequencing data, and used real ancient DNA reads for the Elephantidae pangenome.

Virus and bacteria genomes. For *M. tuberculosis*, SARS-CoV-2, HIV, and RSV, we downloaded reference pangenomes in PanMAN format from <https://zenodo.org/records/17781629>, comprising 400, 20K, 20K, and 4K genomes, respectively. We also downloaded the 8M-genome SARS-CoV-2 PanMAN from the same archive.

Vertebrate mitochondrial genomes. We obtained all vertebrate mitochondrial genomes from RefSeq. We generated the MSA and phylogenetic tree of these genomes using TWILIGHT and DIPPER, respectively, which were then passed on to panmanUtils to generate the PanMAN.

Mammuthus mitochondrial genomes. We obtained 78 *Mammuthus* spp. mitochondrial genomes and their inferred phylogeny from ref. (Martiniano et al. 2022). We then aligned the 78 sequences using TWILIGHT, and used the resulting multiple sequence alignment, together with the Martiniano et al. Newick tree as a guide, to construct a PanMAN.

4.4.14 Panmap index evaluation

We evaluated Panmap’s index in comparison to VG Giraffe and EPIK, two other tools that index k-mers in pangenomes. EPIK requires a database of phylo-k-mers (Romashchenko et al. 2023) built with the related tool IPK from a phylogenetic tree and a multiple sequence alignment. For read alignment to a pangenome graph in GFA format, VG Giraffe indexes k-mer minimizers, a graph Burrows-Wheeler transform, and node distances (Sirén et al. 2021; Sirén and Paten 2022). We provided each tool one CPU thread and measured the runtime, memory usage, and total size after GZIP of all index files needed to run the respective tool. All experiments were performed on a reserved node in a computing cluster, and we determined runtime and memory

usage from the “elapsed (wall clock) time” and “maximum resident set size” reported by GNU time with the argument “-v”.

To build indexes with VG Giraffe, we first extracted Graphical Fragment Assembly (GFA) files from each reference PanMAN pangenome, then indexed the graph with `vg autoindex --workflow giraffe`, which produces a GBZ file (indexed graph Burrows-Wheeler transform), a .dist file (node distances in the graph), and a .min file (indexed k-mer minimizers).

We used IPK to build indexes for EPIK, beginning with a FASTA format multiple genome alignment and a phylogenetic tree, both extracted from the reference PanMAN pangenomes. We first optimized branch lengths of the starting phylogeny by running IQ-TREE with the “-te” argument, which performs maximum-likelihood optimization on a fixed tree topology. We used the GTR+G model for branch optimization, then supplied the best model parameters estimated by IQ-TREE to IPK with the “--ar-config” option for each species. We ran IPK with “-m GTR”, “-k 8”, and “--use-unrooted”, choosing a k-mer size of 8 based on the IPK documentation’s recommendation for large trees. IPK can use either RAxML-ng or PhyML for ancestral state reconstruction, and we first tried the default RAxML-ng mode for all species. For 20,000 SARS-CoV-2 genomes, IPK in RAxML-ng mode failed due to an error in branch length optimization, so we tried again with a different tree containing no branch lengths, which was unsuccessful for the same reason. For 400 *M. tuberculosis* genomes, IPK failed in RaxML-ng mode due to exceeding 1 TB RAM usage. We tried again with the PhyML mode, which supports datasets up to 400 sequences, though this run did not complete within 3 days.

4.4.15 Phylogenetic placement evaluation, Single-sample mode

We evaluated Panmap's phylogenetic placement on both simulated and real sequencing data. For simulations, we selected 70 leaf nodes at random from each reference PanMAN, applied random SNPs and indels with wgsim (indel fraction 0.1) at two per-site mutation rates ($\mu = 10^{-4}$ and 10^{-3} for RSV and SARS-CoV-2; $\mu = 10^{-5}$ and 10^{-4} for *M. tuberculosis*) to generate novel descendant genomes, then simulated 150 bp paired-end reads at four estimated coverage levels (0.5×, 1×, 10×, and 100×), yielding 560 placements per species. For real data, we retrieved assembled genomes from GenBank and corresponding raw reads from the Sequence Read Archive, filtering for an estimated coverage depth of at least 500× and no more than five N characters in the assembly for the viral species; for *M. tuberculosis*, we selected complete genome assemblies with paired-end Illumina data available in SRA. We aligned each sample's reads back to its assembly and discarded samples with more than 5 variants identified by bcftools call. The final real dataset comprised 51 SARS-CoV-2, 25 RSV, and 10 *M. tuberculosis* samples (Table S4.3).

For simulated samples, we placed reads directly onto the unmodified PanMAN and compared Panmap's placement to the true parent of the simulated descendant. For real samples, we used a leave-one-out approach: we temporarily removed the sample's leaf from the PanMAN, subsampled its reads to the target coverage depth, and placed them with Panmap. In both cases, we quantified placement accuracy as the difference in parsimony score between Panmap's placed node and the sample's expected position (placed minus original), adjusted to $\max(0, \text{score})$ so that zero indicates an equivalent

or better placement. Parsimony costs were computed by aligning the sample's reference assembly to each candidate node with minimap2 and counting SNPs and indels.

We quantify placement accuracy by genomic distance rather than topological distance because a PanMAN's branch structure is construction-dependent and the number of branches between nodes does not always reflect their sequence divergence (near-identical genomes can be separated by many branches, especially in large, low-divergence pangenomes). Genomic distance (SNPs and indels) is independent of the tree topology and directly measures divergence from the true position.

4.4.16 Haplotype abundance evaluation with simulated data

To evaluate Panmap's ability to estimate haplotype proportions in simulated mixture samples, we randomly selected sets of 1, 3, 5, and 10 haplotypes from a reference PanMAN and simulated read mixtures for each haplotype level. To assess the effect of variants unrecorded in the tree, we also introduced 0, 1, 5, 10, and 20 single-nucleotide polymorphisms (SNPs) to each haplotype in a sample, generating 10 replicates for each combination of haplotype number and mutation level (200 total replicates). 150-bp paired-end reads at 1000× average depth were simulated using InSilicoSeq and its pre-built NovaSeq error model. Given our objective is estimating major cluster abundance rather than exact sequences, we defined "acceptable resolution" for lineage tracking as haplotypes within ≤ 3 bp of the true sequence.

This evaluation was performed for SARS-CoV-2, RSV, and HIV, using pan-MANs published by Walia et al. (<https://zenodo.org/records/17781629>).

4.4.17 Genome assembly evaluation with real data

We evaluated consensus genome assembly using Panmap and commonly used reference-based pipelines on real sequencing data. For SARS-CoV-2 and RSV, we compared against BWA-MEM + iVar consensus, using the NC_045512.2 reference for SARS-CoV-2, the RSV-A reference (NC_038235.1) for RSV-A samples, or the RSV-B reference (NC_001781.1) for RSV-B samples (determined by competitive mapping of the sample's reads to each reference before evaluation). For *M. tuberculosis*, we compared against Clockwork variant_call_one_sample(Trimmomatic → minimap2 → Cortex + bcftools → Minos adjudication; NC_000962.3 reference). Each method produced a consensus FASTA, which we compared to the ground truth consensus FASTA. We performed quality control filtering and trimming on all reads with fastp prior to evaluation, and trimmed primers for SARS-CoV-2 amplicon protocols with iVar.

We constructed ground truth genomes for 25 RSV genomes, 51 SARS-CoV-2 genomes, and 10 *M. tuberculosis* genomes. Because assembly quality of viral data is variable, for SARS-CoV-2 and RSV we assembled ground truths from the full QC'd read sets (high coverage, >500×) using the Broad Institute assemble_denovo pipeline (SPAdes *de novo* assembly → scaffold against reference → read-back refinement). For *M. tuberculosis*, we used sequences stored in the PanMAN directly (losslessly identical to a published Genbank assembly). We subsampled each dataset to ~0.5×, 1×, 10×, and 100× estimated coverage based on genome length. For each assembled consensus by Panmap, BWA+iVar, and Clockwork, we compared each assembly to its ground truth using minimap2 asm20 alignment (- -cs), quantifying SNP and indel errors and

genome coverage. The final real dataset comprised 74 SARS-CoV-2, 25 RSV, and 10 M. tuberculosis samples.

4.4.18 SARS-CoV-2 lineage abundance estimation from wastewater data

We evaluated Panmap’s capability to estimate lineage abundance using 22 SARS-CoV-2 wastewater samples, collected by Karthikeyan et al. at Point Loma (San Diego, California) between November 27, 2021 and February 7, 2022. We chose this period due to high local epidemic surveillance and sequencing submission activity. For Panmap, we used a SARS-CoV-2 PanMAN comprising 20,000 samples and 1,000 Pango lineages (<https://zenodo.org/records/17781629>), with the most recent sample submitted on January 5, 2024. For comparison with WEPP, we used the next available UShER-format Mutation-Annotated Tree (MAT) on the UCSC Genome Browser (Casper et al. 2026) after this date (public-2024-02-01.all.masked.pb.gz).

We preprocessed the wastewater reads by aligning to the Wuhan-1 SARS-CoV-2 reference (NC_045512.2) using minimap2 (minimap -a - -sam-hit-only -MD -2 -x sr) (Li et al. 2018; Li 2021) and trimming the primer sequences using ivar (ivar trim -e -q 1 -m 80 -x 3) (Grubaugh et al. 2019).

For Panmap, we converted the trimmed bam files to FASTQ format and removed trimmed regions using samtools. As the wastewater data were generated using tiled amplicons, metagenomic Panmap removes reads whose seed relative frequencies are less than 0.01 compared to the total number of reads associated with the same amplicon to mitigate spurious PCR and sequencing errors.

As these are amplicon-targeted samples, and reads that failed to align to the Wuhan-1 reference were discarded during this preprocessing, reads supplied to

Panmap can be assumed to be almost entirely of SARS-CoV-2 origin, the fraction of off-target sequences is therefore negligible and the “--discard” filter described above was not applied here.

We ran WEPP and Freyja with the default settings and commands as described in their documentation, following the same mapping and trimming procedures described above.

4.4.19 Ancient environmental DNA placement

We obtained the adaptor-trimmed circumpolar ancient permafrost DNA dataset from ref. (Wang et al. 2021) and preprocessed the reads using fastp (Chen 2023) and sga (Simpson and Durbin 2012) to deduplicate exact substrings, trim poly Gs, and remove reads with low quality, low complexity, and short length. We used the following specific commands and parameters:

```
fastp -V -D -A --dup_calc_accuracy 5 -g -x --poly_g_min_len 5 --
poly_x_min_len 5 -q 30 -e 25 -l 30 -y -p -w 8
sga preprocess --dust-threshold=30 -m
sga index --algorithm=ropebwt; sga filter --no-kmer-check
```

We implemented several competitive mapping-based filters for eDNA taxonomic assignment. Let r denote a sequencing read and $S(r)$ the set of unique seeds sketched from r , with $|S(r)|$ denoting the total seed count.

Low seed-hit filter. For each read r , let $\sigma(r)$ denote its maximum alignment score against the reference tree. The seed-hit ratio is defined as:

$$\rho(r) = \frac{\sigma(r)}{|S(r)|}$$

Reads for which $\rho(r) < \tau_m$ (default $\tau_m = 0.7$) are discarded as low mappability.

Phylogenetic ambiguity filter. Let $T(r,l)$ denote the set of taxa to which read r is assigned at taxonomic rank l . A read is discarded as phylogenetically ambiguous if:

$$|T(r, l)| > N$$

where N is a user-defined threshold (default $N=1$) and l defaults to the family rank.

Evenness of coverage filter. For a reference node v , let $R(v)$ denote the set of reads assigned to v , and let $S(v)$ be the set of seeds sketched from the reference sequence of v , with $|S(v)|$ the total reference seed count. Define $h(s)$ as the number of hits to seed $s \in S(v)$ from reads in $R(v)$, and let $D(v)$ denote the mean number of hits per seed:

$$D(v) = \frac{1}{|S(v)|} \sum_{s \in S(v)} h(s)$$

The observed seed coverage is:

$$C_{\text{obs}(v)} = \frac{|\{s \in S(v) : h(s) \geq 1\}|}{|S(v)|}$$

Under a Poisson model with rate $D(v)$, the expected seed coverage is:

$$C_{\text{exp}(v)} = 1 - e^{-D(v)}$$

Node v is discarded if the coverage ratio $C_{\text{obs}(v)}/C_{\text{exp}(v)} < \tau_c$, where τ_c defaults to 0.2.

Reads assigned to any discarded node are subsequently removed as well.

For ancient DNA placement, we recommend and use seed parameters $k=15$, $s=8$, $l=1$. Shorter seeds are used here to account for ancient DNA's divergence and damage.

4.4.20 Panmap-assigned reads validation from Wang et al. dataset

We validated Panmap-assigned reads by first aligning them to the *Mammuthus primigenius* mitochondrial reference genome (GenBank: DQ188829.2) on the PanMAN using bwa-aln (Li and Durbin 2009) with parameters tuned for ancient DNA (bwa-aln -t 8 -l 1024 -q 15 -n 0.02) (Martiniano et al. 2022). We then performed nucleotide BLAST (Altschul et al. 1990) on all the alignable reads using the core_nt database (blastn -evalue 0.05 -max_target_seqs 1000). For a sequence, BLAST hits were retained only if their bit scores fell within the top 10% of the maximum score observed for the corresponding sequence, following Slon et al. (Slon et al. 2017). Reads were considered to have exclusively Elephantidae mitochondrial hits if their retained matches were only hits to mitochondrial genomes of Elephantidae species.

4.4.21 Per-sample placement on *Mammuthus* PanMAN

We placed each sample of Panmap-assigned reads from the Wang et al. data onto a *Mammuthus* species PanMAN. Samples were processed using haplotype deconvolution as described above. Although most samples returned a single ambiguous haplotype group (i.e. haplotypes indistinguishable from the given data), for samples estimated to contain more than one haplotype group, the most abundant group was selected. Each sample is then placed at the last common ancestor of all haplotypes within the selected group.

CHAPTER 5. CONCLUSION

In Chapter 2, I worked with Bryan Thornlow and co-authors to ask whether parsimony-based online phylogenetics is accurate enough to replace maximum likelihood for SARS-CoV-2. We compared online with *de novo* inference, and parsimony with likelihood, under the time constraints that a pandemic imposes, using both real genomes and simulated data. Online optimization with UShER and matOptimize produced trees equivalent to or more optimal than those from IQ-TREE 2, FastTree 2, and RAxML-NG. Parsimony was sufficient because dense sampling produces very short branches, and we found that parsimony scores and likelihoods improved together across successive rounds of optimization.

In Chapter 3, I built two browser-based tools for analyzing trees of this size. Laboratories that cannot transmit patient-derived sequence were unable to use server-based placement services, so I compiled UShER to WebAssembly and developed ShUShER, which aligns, places, and displays a user's samples entirely within the browser. Existing tools could show a multiple sequence alignment beside a phylogeny, but not for millions of sequences, so I built Treenome Browser, which decodes the mutation-annotated tree directly for display and remains responsive on trees of more than twelve million sequences.

In Chapter 4, I worked with Alan Zhang and co-authors to move from the single linear reference-based structure used in the earlier chapters to a lossless phylogenetic pangenome method for analyzing new samples. Panmap indexes a PanMAN by storing

only the k-mer seeds that change along each branch, and uses that index to place reads, align them, call genotypes, and estimate haplotype abundances.

Mutation-annotated trees and networks store mutations once, on the branch where they arise, so that their underlying genomic sequences are recovered by accumulating differences from the root rather than storing each full sequence independently. Because these structures record changes between nodes and their parents, placement, tree optimization, rendering, and seed indexing can all be computed by updating a running state along branches instead of examining every genome in the dataset. As this thesis demonstrates, phylogenetic compression is not only a useful method for reducing storage space, but also a valuable principle for making computational pangenomic analyses faster and more scalable.

APPENDIX A. SUPPLEMENTARY FIGURES AND TABLE FOR CHAPTER 2

The figures and table in this appendix are the supplementary material of the matOptimize study reproduced in Chapter 2, referred to in that chapter as Supplementary Figures S2.1–S2.3 and Supplementary Table S2.1. Scripts and data for that study are available on GitHub (<https://github.com/bpt26/parsimony>) and in the Dryad Digital Repository (<https://doi.org/10.7291/D13Q2J>). Captions are reproduced verbatim, except that references to figures of the original publication have been renumbered to match this dissertation.

Table S2.1. The best-performing methods on real data yield equivalent results across replicates under Generalised Time Reversible (GTR) and Jukes-Cantor (JC) likelihoods over five iterations. De novo IQ-TREE 2 and *de novo* UShER+matOptimize performed the best in terms of parsimony and JC likelihood on real SARS-CoV-2 data. We report our original experiment (Replicate 1) and a repeated experiment (Replicate 2) for the first five iterations of both of these methods. We specified a different random seed in each matOptimize run, and IQ-TREE 2 uses the system clock as a random seed. We computed JC and GTR likelihoods for all runs, using the GTR model parameters selected by IQ-TREE 2 in each experiment.

IQ-TREE 2 Replicate 1 (left) and IQ-TREE 2 Replicate 2 (right):

# samples	Parsimony score	JC likelihood	GTR likelihood	Parsimony score	JC likelihood	GTR likelihood
4,676	3,022	-76,693.571	-73,301.5662	3,022	-76,688.407	-73,291.8025
8,902	5,524	-106,704.839	-100,977.2868	5,526	-106,717.690	-100,968.6700
13,241	8,062	-137,083.559	-128,610.3219	8,065	-137,108.520	-128,616.3814
17,941	10,905	-171,090.515	-159,483.9097	10,903	-171,068.926	-159,502.9303
22,012	13,387	-200,737.415	-186,372.7342	13,382	-200,652.723	-186,340.0237

matOptimize Replicate 1 (left) and matOptimize Replicate 2 (right):

# samples	Parsimony score	JC likelihood	GTR likelihood	Parsimony score	JC likelihood	GTR likelihood
4,676	3,018	-76,642.464	-73,251.466	3,018	-76,642.482	-73,250.200
8,902	5,518	-106,607.769	-100,858.276	5,517	-106,597.636	-100,849.985
13,241	8,050	-136,894.070	-128,403.412	8,049	-136,894.066	-128,397.885
17,941	10,879	-170,725.142	-159,111.944	10,880	-170,723.755	-159,112.517
22,012	13,347	-200,140.552	-185,784.193	13,347	-200,148.601	-185,788.145

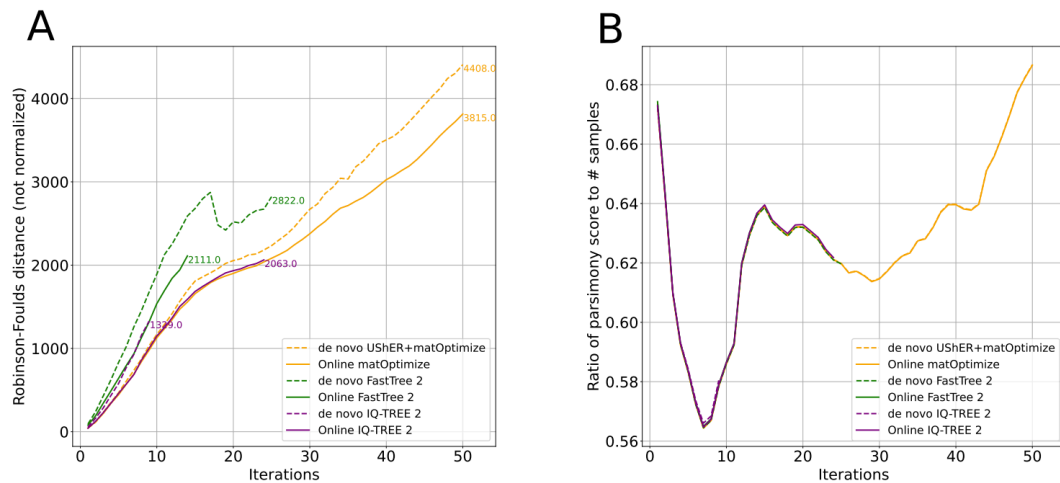


Figure S2.1. Temporal patterns in simulated SARS-CoV-2 data may affect Robinson-Foulds (RF) distance normalization. The RF distances for each tree in Figure 2.2 are normalized based on the number of internal edges in the ground truth tree. While the raw RF distances are approximately continuously increasing (A), they do not increase linearly with the normalization factor, leading to the pattern observed in Figure 2.2. A potential explanation for this is the variation in sequence diversity over simulated time. The ratio of the number of total mutations in the tree (parsimony score) to the number of samples in the inferred trees at each iteration (B) approximates the average divergence between samples in each tree. The initial drop in divergence per sample may contribute to the more rapid increase in RF distance because there is less phylogenetic signal to facilitate the resolution of correct topologies. As the divergence subsequently increases, tree inference improves before the RF distances stabilize and begin to increase approximately linearly.

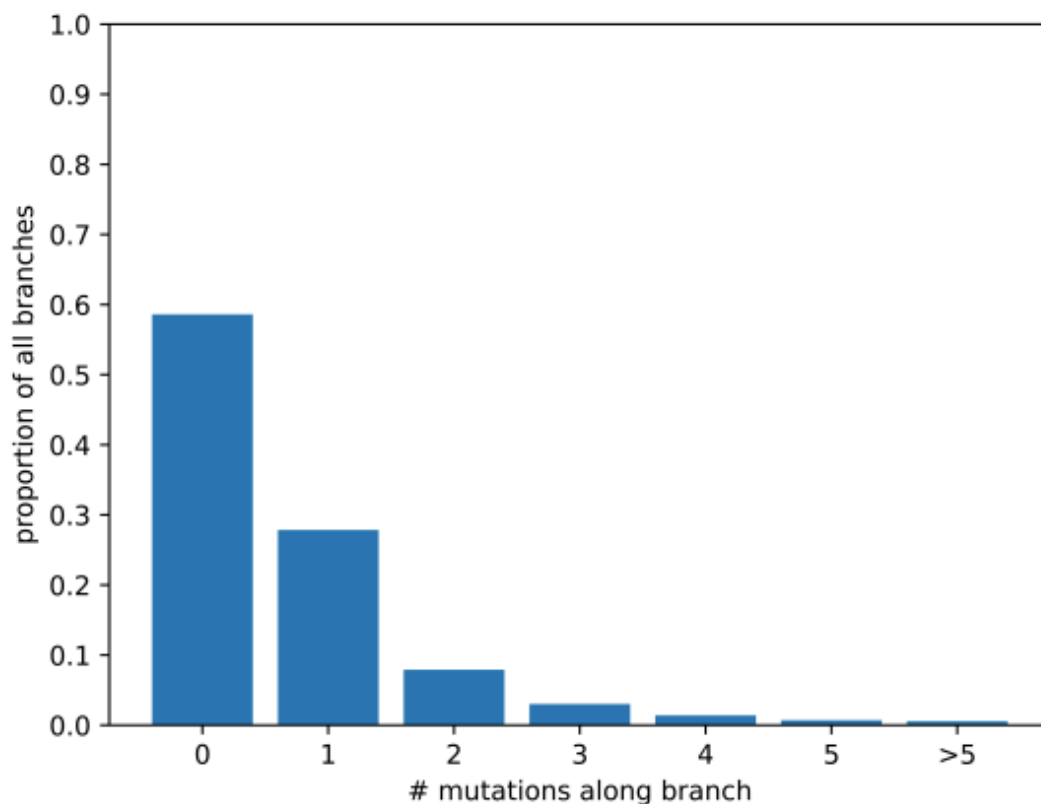


Figure S2.2. Most branches in the ground truth phylogeny are extremely short. In our optimized global SARS-CoV-2 phylogeny, the majority of branch lengths are zero. This low amount of divergence yields many identical nodes in the tree and demonstrates that the probability of observing multiple mutations at a single site along the same branch is negligible. These characteristics may help explain the ability of parsimony-based inference methods to outperform likelihood optimization on SARS-CoV-2 data.

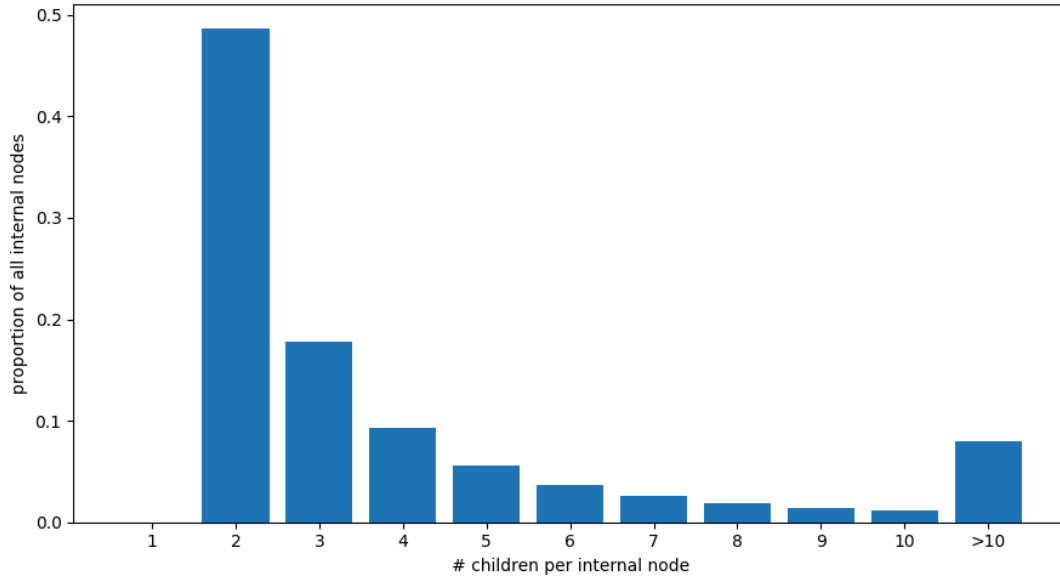


Figure S2.3. Resolution of the ground truth phylogeny. This histogram shows the proportion of all internal nodes of the simulated, optimized ground truth phylogeny that have a given number of direct descendants (children). For example, nearly half of the internal nodes are resolved, with exactly two children.

APPENDIX B. SUPPLEMENTARY FIGURES FOR CHAPTER 4

The figures in this appendix are the supplementary figures of the Panmap manuscript reproduced in Chapter 4, referred to in that chapter as Figure S4.1 through Figure S4.10. They are reproduced from the supplementary material of the preprint (<https://doi.org/10.64898/2026.03.29.711974>).

Fig. S8. Multithreading scalability (≤ 16 threads) — wall-clock time and peak memory vs thread count.

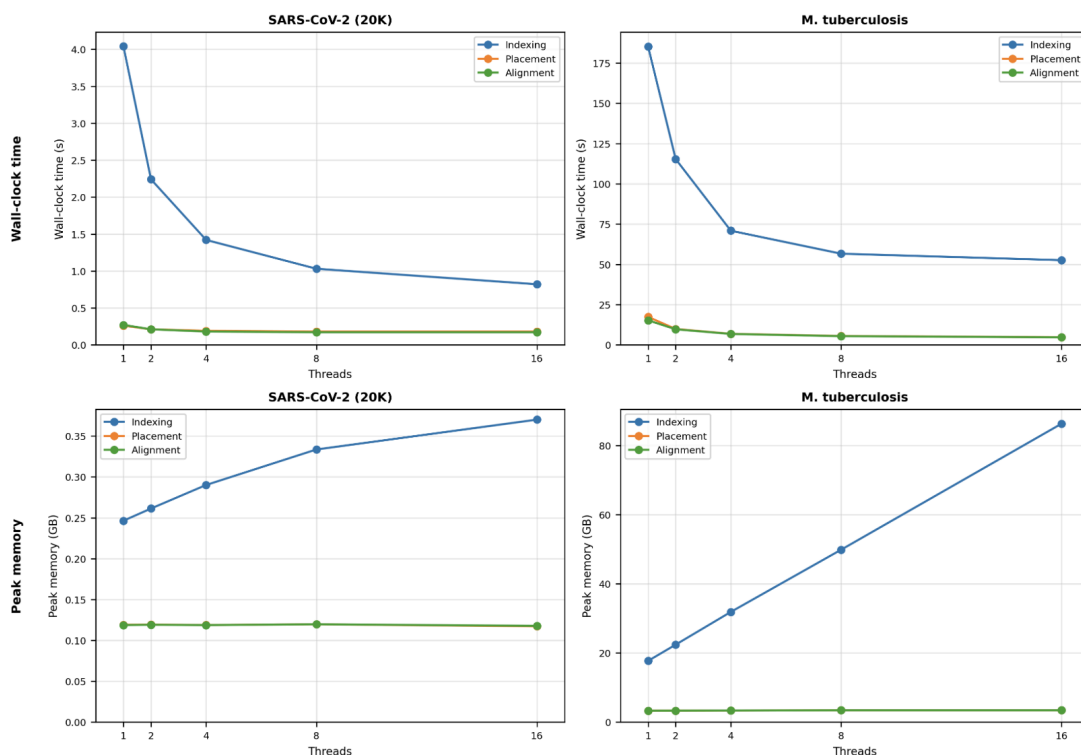


Figure S4.1. Multithreading scalability of Panmap. Wall-clock time and peak memory usage vs. thread count for Panmap’s three stages (indexing, placement, and alignment) on SARS-CoV-2 (20,000-genome PanMAN, left) and *M. tuberculosis* (400-genome PanMAN, right). Index construction benefits from multithreading: for SARS-CoV-2, wall time drops from ~4s (1 thread) to ~2.3s (2 threads) to ~1.5s (4 threads) to <1s (8 threads); for *M. tuberculosis*, from ~175s to ~110s to ~75s to ~55s over the same progression. Returns diminish beyond 8 threads, with 8-to-16 yielding only marginal improvement. Placement and alignment runtimes are effectively constant across thread counts. Peak indexing memory grows with thread count, particularly for *M. tuberculosis* (~18 GB at 1 thread to ~80 GB at 16), while placement and alignment memory remains flat. Based on these profiles, 4–8 threads offers the best trade-off between speed and memory for index construction.

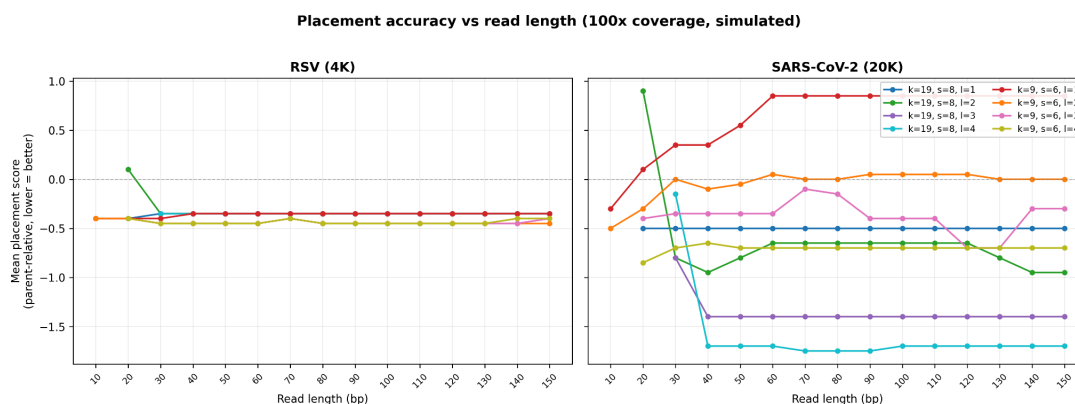


Figure S4.2. Placement accuracy vs. read length for simulated 100× coverage data. Mean parent-relative placement score (lower is better; zero indicates placement equivalent to the true position) is shown for RSV (4,000-genome PanMAN, left) and SARS-CoV-2 (20,000-genome PanMAN, right). Each line represents a different combination of k-mer size (k) and linked syncmer chain length (l). For RSV, all configurations achieve stable, near-optimal placement across all read lengths down to ~20 bp. For SARS-CoV-2, linked syncmers ($l \geq 2$) maintain substantially better accuracy at read lengths >40 bp. This demonstrates that chaining syncmers into linked seeds provides additional specificity for placement on larger, more diverse pangenomes, and that Panmap can accurately place reads far shorter than standard Illumina output.

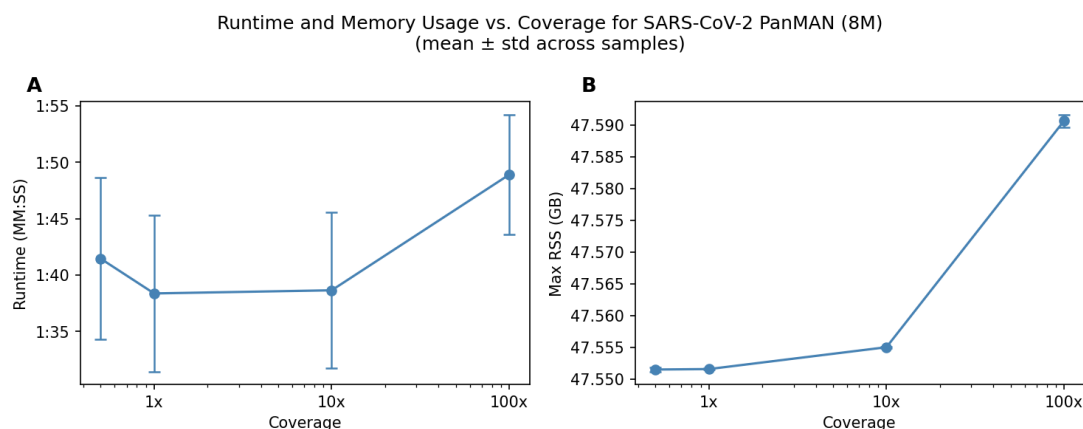


Figure S4.3. Computational resources for placing samples onto a PanMAN with 8 million genomes. (A) Average runtime for placing a single SARS-CoV-2 sample with 0.5X, 1X, 10X, and 100X coverage. (B) Average max RSS usage for placing a single SARS-CoV-2 sample.

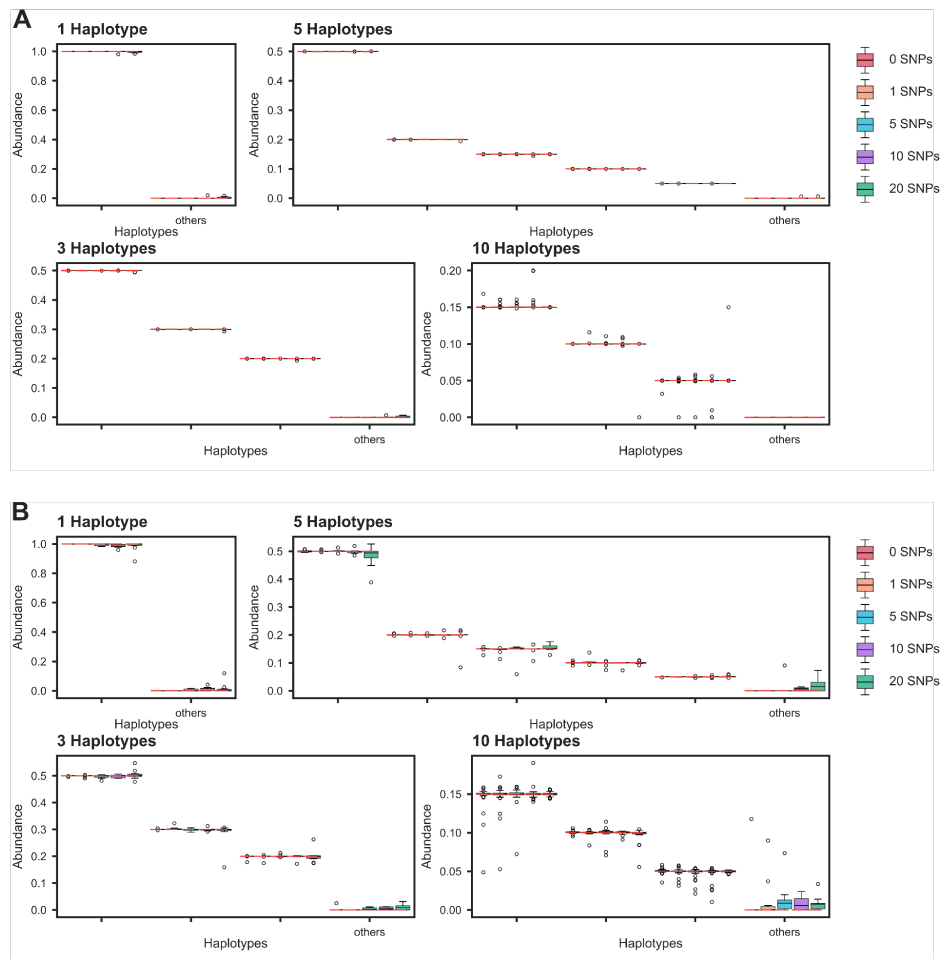


Figure S4.4. Evaluation of haplotype abundance estimation in simulated HIV (A) and RSV (B) mixture samples. Estimated haplotype abundance in simulated HIV and RSV mixture samples with 1, 3, 5, or 10 haplotypes, introducing 0, 1, 5, 10, or 20 single-nucleotide polymorphisms (SNPs) per haplotype. Ten replicates were simulated per composition and mutation level. Red dashed lines denote true simulated abundances; the ‘others’ category represents accumulated false-positive haplotype abundance. (*10-haplotype samples: 4 haplotypes at 15%, 2 at 10%, 4 at 5% abundance).

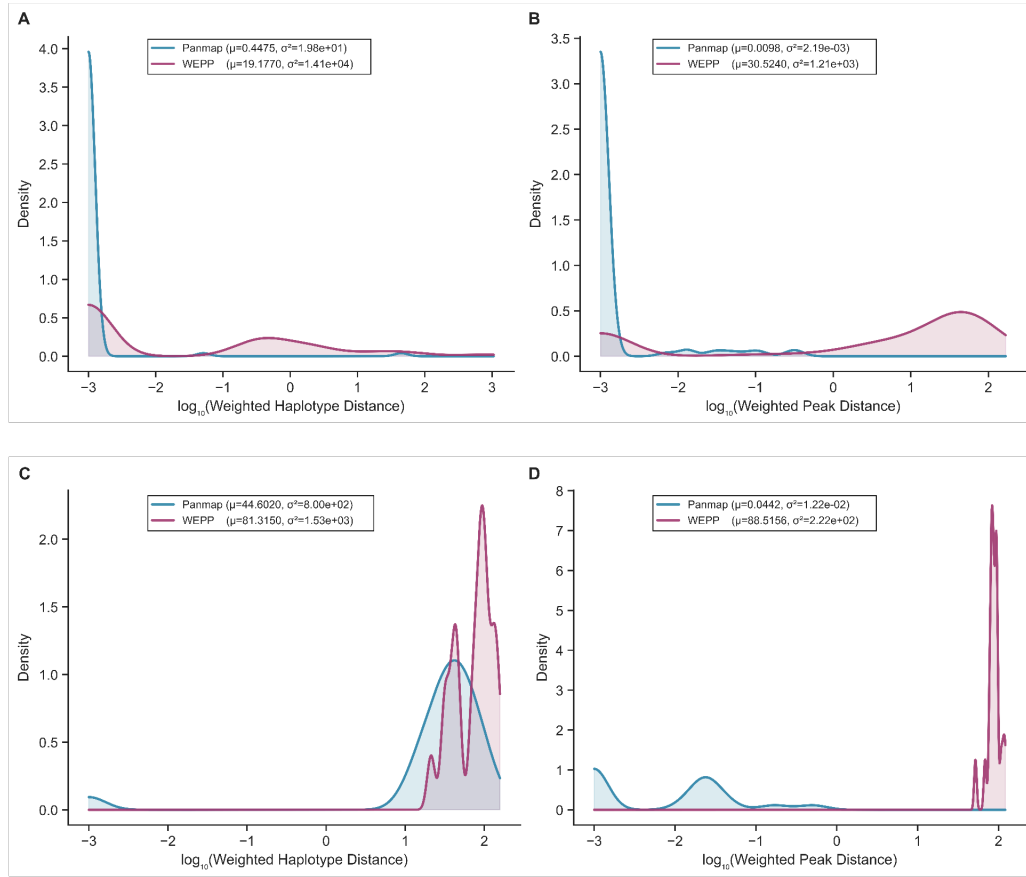


Figure S4.5. Comparison of haplotype abundance estimation accuracy between Panmap and WEPP. Density plots show Weighted Haplotype Distance (WHD) and Weighted Peak Distance (WPD) for both methods. WHD and WPD quantify sensitivity and precision of haplotype inference, respectively, with lower values indicating greater agreement with true haplotype compositions (Gangwar et al.). (A, B) Samples containing 1, 2, 3, 5, or 10 haplotypes (20 replicates per mixture level; plots aggregate all replicates). (C, D) Samples containing 50 haplotypes under limit-of-detection testing (20 replicates total).

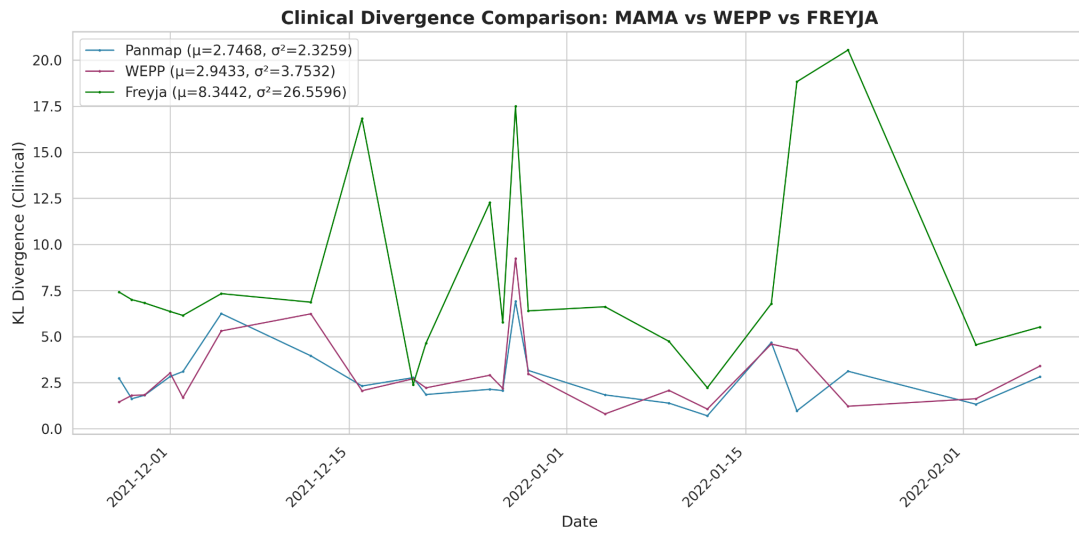


Figure S4.6. Comparison of SARS-CoV-2 lineage abundance estimation between Panmap, WEPP, and Freyja. Time series of KL divergences between wastewater estimated lineage proportions and one-week-delayed clinically observed lineage proportions.

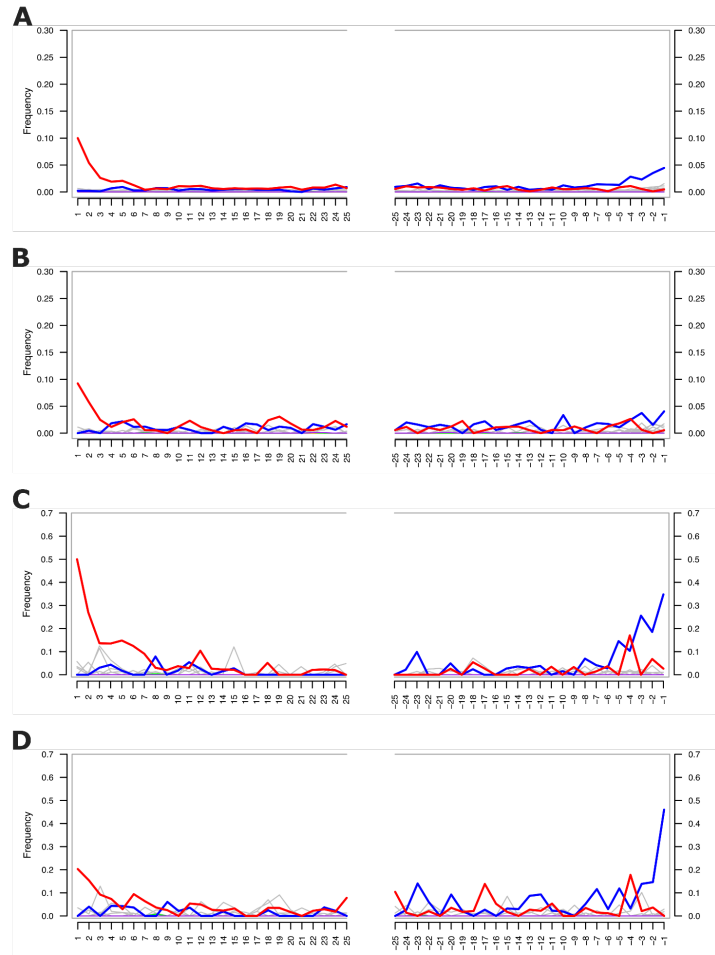


Figure S4.8. Damage plots of assigned reads generated using mapDamage (Supp. Reference 1). Red lines indicate C-to-T substitution frequencies, blue lines G-to-A substitution frequencies, and gray lines all other substitutions. (A) Panmap-assigned reads from Wang et al. samples mapped to a *Mammuthus primigenius* reference genome (GenBank: DQ188829.2). (B) Competitive mapping-assigned reads from Wang et al. samples mapped to the same reference. (C) Panmap-assigned reads from Kjær et al. samples mapped to a *Mammot americanum* reference genome (GenBank: KY364233.1). (D) Competitive mapping-assigned reads from Kjær et al. samples mapped to the same reference.

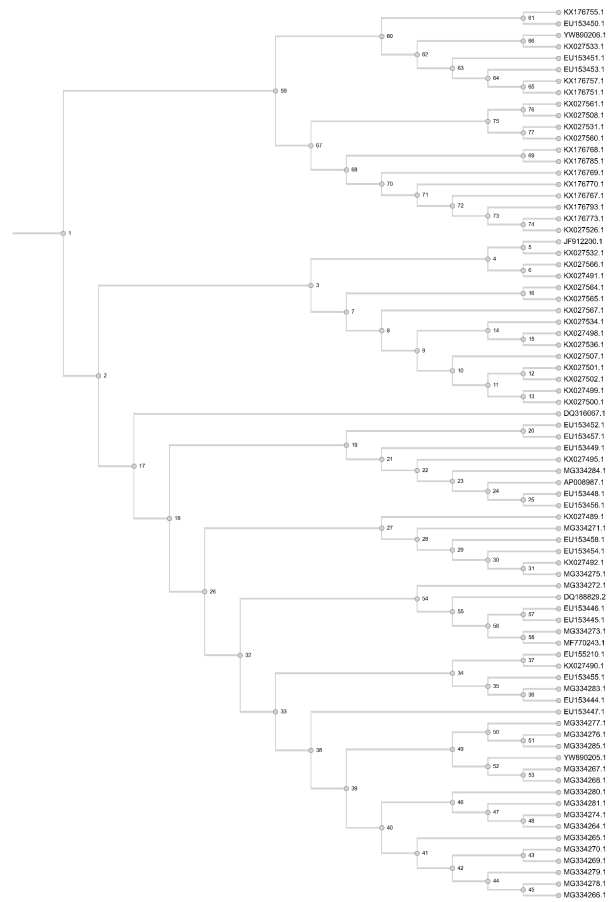


Figure S4.9. *Mammuthus* mitochondrial phylogeny. Tree topology of the mammoth mitochondrial genomes. Referenced by Table S4.4.

Fig. S1. Placement score distributions across log scoring metrics and coverages (simulated reads). Score = $\max(0, \text{placed_dist} - \text{parent_dist})$; 0 is exact. Numbers below x-axis: % exact placements.

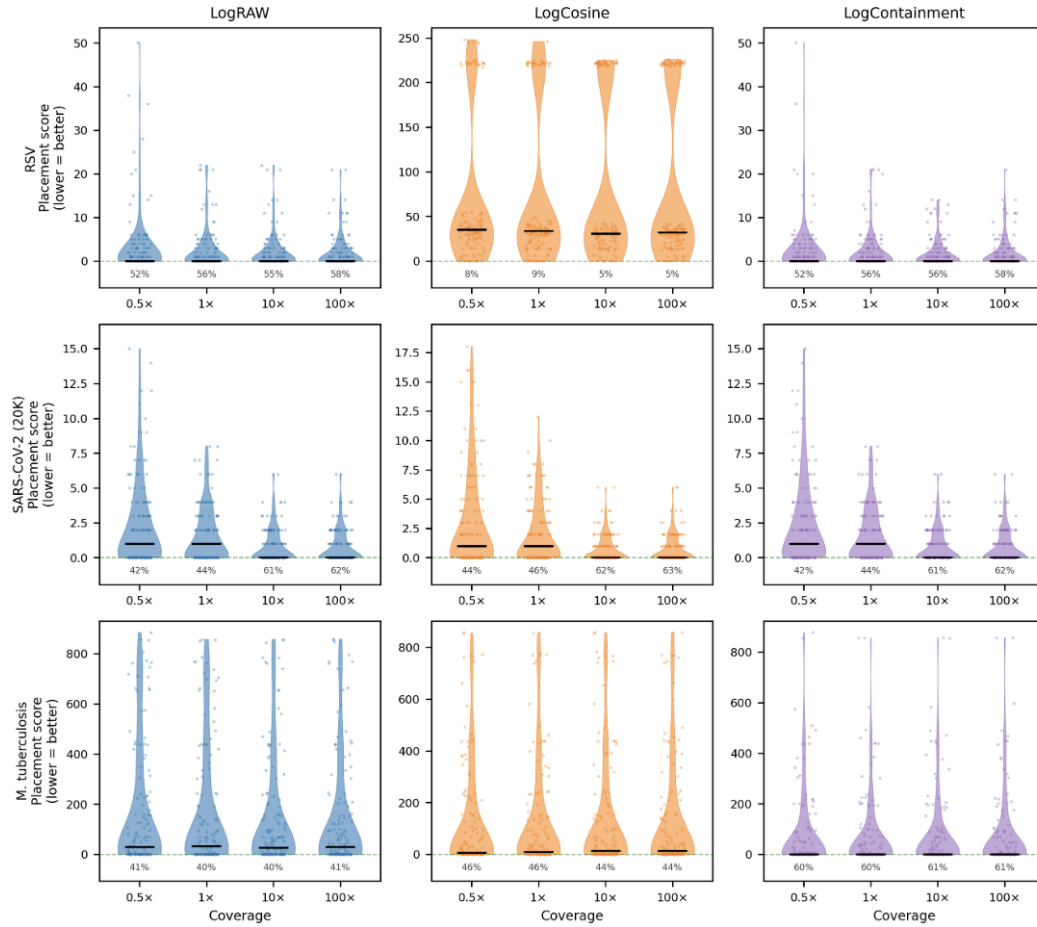


Figure S4.10. Panmap single-sample scoring metric evaluation. Comparison of three placement scoring metrics on simulated data across RSV (4,000 genomes), SARS-CoV-2 (20,000 genomes), and *M. tuberculosis* (400 genomes) at four coverage levels. Violin plots show the distribution of parent-relative placement scores (lower is better; zero indicates placement at least as parsimonious as the true position); black bars indicate medians; percentages show exact-match rates (score = 0). Log containment consistently outperforms the other two metrics. These results motivated our choice of Log containment as Panmap's default placement metric.

APPENDIX C. SUPPLEMENTARY TABLES FOR CHAPTER 4

Chapter 4 cites supplementary Tables S4.1 through S4.6. These are published in the supplementary material of the preprint reproduced in that chapter (<https://doi.org/10.64898/2026.03.29.711974>). Their contents are:

- **Table S4.1.** Metagenomic mode performance: index size, construction time, and peak memory for the metagenomic-mode index, which additionally stores seed orientations.
- **Table S4.2.** VG Giraffe read-mapping performance on the pathogen pangenome graphs.
- **Table S4.3.** Sequence Read Archive accessions used for the RSV, SARS-CoV-2, and *M. tuberculosis* evaluations.
- **Table S4.4.** Sources and construction details of the reference PanMANs.
- **Table S4.5.** Per-sample mammoth placement results for the Wang et al. (2021) dataset.
- **Table S4.6.** Validation of Panmap-assigned reads by alignment and BLAST.

BIBLIOGRAPHY

- Agnarsson, I., & Miller, J. A. (2008). Is ACCTRAN better than DELTRAN?. *Cladistics*, 24, 1032–1038.
- Alser, M., Rotman, J., Deshpande, D., Taraszka, K., Shi, H., Baykal, P. I., Yang, H. T., Xue, V., Knyazev, S., Singer, B. D., Balliu, B., Koslicki, D., Skums, P., Zelikovsky, A., Alkan, C., Mutlu, O., & Mangul, S. (2021). Technology dictates algorithms: recent developments in read alignment. *Genome Biology*, 22(1), 249.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Annavajhala, M. K., Mohri, H., Wang, P., Nair, M., Zucker, J. E., Sheng, Z., Gomez-Simmonds, A., Kelley, A. L., Tagliavia, M., Huang, Y., Bedford, T., Ho, D. D., & Uhlemann, A.-C. (2021). Emergence and expansion of SARS-CoV-2 B.1.526 after identification in New York. *Nature*, 597, 703–708. <https://doi.org/10.1038/s41586-021-03908-2>
- Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E. D., ... Paten, B. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833), 246–251.

- Asher, R. J., & Smith, M. R. (2022). Phylogenetic signal and bias in paleontology. *Systematic Biology*, *71*, 986–1008.
- Barbera, P., Kozlov, A. M., Czech, L., Morel, B., Darriba, D., Flouri, T., & Stamatakis, A. (2019). EPA-ng: massively parallel evolutionary placement of genetic sequences. *Systematic Biology*, *68*(2), 365–369.
- Berger, S. A., Krompass, D., & Stamatakis, A. (2011). Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology*, *60*, 291–302.
- Bluhm, A., Christandl, M., Gesmundo, F., Klausen, F. R., Mančinská, L., Steffan, V., França, D. S., & Werner, A. H. (2020). SARS-CoV-2 transmission routes from genetic data: a Danish case study. *Plos One*, *15*, e241405.
- Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elisk, C. G., Lewis, S. E., Stein, L., & Holmes, I. H. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology*, *17*, 66. <https://doi.org/10.1186/s13059-016-0924-1>
- Casper, J., Speir, M. L., Raney, B. J., Perez, G., Nassar, L. R., Lee, C. M., Hinrichs, A. S., Gonzalez, J. N., Fischer, C., Diekhans, M., Clawson, H., Benet-Pages, A., Barber, G. P., Vaske, C. J., Baren, M. J. van, Wang, K., Rodriguez, Y. J. P., Jenkins-Kiefer, J. A., Chalamala, M., ... Haeussler, M. (2026). The UCSC Genome Browser database: 2026 update. *Nucleic Acids Research*, *54*, D1331–D1335.
- Castillo, A. E., Parra, B., Tapia, P., Acevedo, A., Lagos, J., Andrade, W., Arata, L., Leal, G., Barra, G., Tampley, C., Tognarelli, J., Bustos, P., Ulloa, S., Fasce, R., &

- Fernández, J. (2020). Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. *Journal of Medical Virology*, *92*, 1562–1566.
- Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta*, *2*, e107.
- COVID-19 Genomics UK (COG-UK) Consortium. (2020). An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe*, *1*, e99–e100.
- Czech, L., Stamatakis, A., Dunthorn, M., & Barbera, P. (2022). Metagenomic analysis using phylogenetic placement—a review of the first decade. *Frontiers in Bioinformatics*, *2*, 871393. <https://doi.org/10.3389/fbinf.2022.871393>
- De Maio, N., Boulton, W., Weilguny, L., Walker, C. R., Turakhia, Y., Corbett-Detig, R., & Goldman, N. (2022). phastSim: efficient simulation of sequence evolution for pandemic-scale datasets. *PLOS Computational Biology*, *18*(4), e1010056. <https://doi.org/10.1371/journal.pcbi.1010056>
- De Maio, N., Kalaghatgi, P., Turakhia, Y., Corbett-Detig, R., Minh, B. Q., & Goldman, N. (2023). Maximum likelihood pandemic-scale phylogenetics. *Nature Genetics*, *55*, 746–752. <https://doi.org/10.1038/s41588-023-01368-0>
- De Maio, N., Walker, C. R., Turakhia, Y., Lanfear, R., Corbett-Detig, R., & Goldman, N. (2021). Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biology and Evolution*, *13*, evab87.
- De Maio, N., Walker, C., Borges, R., Weilguny, L., Slodkiewicz, G., & Goldman, N. (2020). Issues with SARS-CoV-2 sequencing data. *Virological.org*. <https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/1>

- De Sanctis, B., Mirchandani, C., Dong, H., Macleod, R., Corbett-Detig, R., & Wang, Y. (2025). Bamdam: a post-mapping authentication toolkit for ancient metagenomics. *Genome Biology*, 26, 413.
- Deng, X., Gu, W., Federman, S., Plessis, L. du, Pybus, O. G., Faria, N. R., Wang, C., Yu, G., Bushnell, B., Pan, C.-Y., Guevara, H., Sotomayor-Gonzalez, A., Zorn, K., Gopez, A., Servellita, V., Hsu, E., Miller, S., Bedford, T., Greninger, A. L., ... Chiu, C. Y. (2020). Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science*, 369, 582–587.
- Dolenz, S., Valk, T. van der, Jin, C., Oppenheimer, J., Sharif, M. B., Orlando, L., Shapiro, B., Dalén, L., & Heintzman, P. D. (2024). Unravelling reference bias in ancient DNA datasets. *Bioinformatics*, 40(7), btae436. <https://doi.org/10.1093/bioinformatics/btae436>
- Dorp, L. van, Richard, D., Tan, C. C. S., Shaw, L. P., Acman, M., & Balloux, F. (2020). No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nature Communications*, 11, 5986.
- Du, Y., & Varadhan, R. (2020). SQUAREM: an R package for off-the-shelf acceleration of EM, MM and other EM-like monotone algorithms. *Journal of Statistical Software*, 92, 1–41.
- Ebler, J., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Korbel, J., Eichler, E. E., Zody, M. C., Dilthey, A. T., & Marschall, T. (2020). Pangenome-based genome inference. *Biorxiv*. <https://doi.org/10.1101/2020.11.11.378133>
- Edgar, R. (2021). Synckmers are more sensitive than minimizers for selecting conserved k-mers in biological sequences. *Peerj*, 9, e10805.

- Ekim, B., Sahlin, K., Medvedev, P., Berger, B., & Chikhi, R. (2023). Efficient mapping of accurate long reads in minimizer space with mapquik. *Genome Research*, 33, 1188–1197.
- Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27, 401–410.
- Fourment, M., Claywell, B. C., Dinh, V., McCoy, C., Matsen IV, F. A., & Darling, A. E. (2018). Effective online Bayesian phylogenetics via sequential Monte Carlo with guided proposals. *Systematic Biology*, 67, 490–502.
- Franceschi, V. B., Caldana, G. D., Menezes Mayer, A. de, Cybis, G. B., Neves, C. A. M., Ferrareze, P. A. G., Demoliner, M., Almeida, P. R. de, Gulate, J. S., Hansen, A. W., Weber, M. N., Fleck, J. D., Zimmerman, R. A., Kmetzsch, L., Spilki, F. R., & Thompson, C. E. (2021). Genomic epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil. *BMC Genomics*, 22, 371.
- Gangwar, P., Katte, P., Bhatt, M., & Turakhia, Y. (2025). WEPP: phylogenetic placement achieves near-haplotype resolution in wastewater-based epidemiology. *Medrxiv*, 2025.06.09.25329287.
- Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9), 875–879.
- Gill, M. S., Lemey, P., Suchard, M. A., Rambaut, A., & Baele, G. (2020). Online Bayesian phylodynamic inference in BEAST with application to epidemic reconstruction. *Molecular Biology and Evolution*, 37, 1832–1842.

- Gonzalez-Reiche, A. S., Hernandez, M. M., Sullivan, M. J., Ciferri, B., Alshammary, H., Obla, A., Fabre, S., Kleiner, G., Polanco, J., Khan, Z., Albuquerque, B., Guchte, A. van de, Dutta, J., Francoeur, N., Melo, B. S., Oussenko, I., Deikus, G., Soto, J., Sridhar, S. H., ... Bakel, H. van. (2020). Introductions and early spread of SARS-CoV-2 in the New York City area. *Science*, 369, 297–301.
- Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., De Jesus, J. G., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S., Gurfield, N., Van Rompay, K. K. A., Isern, S., Michael, S. F., Coffey, L. L., Loman, N. J., & Andersen, K. G. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biology*, 20, 8.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59, 307–321.
- Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *Plos Genetics*, 15(7), e1008302. <https://doi.org/10.1371/journal.pgen.1008302>
- Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34, 4121–4123. <https://doi.org/10.1093/bioinformatics/bty407>
- Hendy, M. D., & Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Systematic Biology*, 38, 297–309.

- Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., Marschall, T., Li, H., & Paten, B. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, 42, 663–673. <https://doi.org/10.1038/s41587-023-01793-w>
- Hinrichs, A. S. (n.d.). *viral_usher: Easily build an UShER tree of genomes for any virus in RefSeq/GenBank*. Retrieved https://github.com/AngieHinrichs/viral_usher
- Huelsenbeck, J. P., Rannala, B., & Larget, B. (2000). A Bayesian framework for the analysis of cospeciation. *Evolution*, 54, 352–364.
- Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, 1, 16048.
- Hunt, M., Letcher, B., Malone, K. M., Nguyen, G., Hall, M. B., Colquhoun, R. M., Lima, L., Schatz, M. C., Ramakrishnan, S., CRyPTIC consortium, & Iqbal, Z. (2022). Minos: variant adjudication and joint genotyping of cohorts of bacterial genomes. *Genome Biology*, 23, 147.
- Izquierdo-Carrasco, F., Cazes, J., Smith, S. A., & Stamatakis, A. (2014). PUMPER: phylogenies updated perpetually. *Bioinformatics*, 30, 1476–1477.
- Jackson, B., Boni, M. F., Bull, M. J., Colleran, A., Colquhoun, R. M., Darby, A. C., Haldenby, S., Hill, V., Lucaci, A., McCrone, J. T., Nicholls, S. M., O'Toole, A., Pacchiarini, N., Poplawski, R., Scher, E., Todd, F., Webster, H. J., Whitehead, M., Wierzbicki, C., ... COVID-19 Genomics UK (COG-UK) Consortium. (2021).

- Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell*, 184, 5179–5188.e8.
- Kalantar, K. L., Carvalho, T., Bourcy, C. F. A. de, Dimitrov, B., Dingle, G., Egger, R., Han, J., Holmes, O. B., Juan, Y.-F., King, R., Kislyuk, A., Lin, M. F., Mariano, M., Morse, T., Reynoso, L. V., Cruz, D. R., Sheu, J., Tang, J., Wang, J., ... DeRisi, J. L. (2020). IDseq—an open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *Gigascience*, 9, giaa111.
- Karthikeyan, S., Levy, J. I., De Hoff, P., Humphrey, G., Birmingham, A., Jepsen, K., Farmer, S., Tubb, H. M., Valles, T., Tribelhorn, C. E., Tsai, R., Aigner, S., Sathe, S., Moshiri, N., Henson, B., Mark, A. M., Hakim, A., Baer, N. A., Barber, T., ... Knight, R. (2022). Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature*, 609, 101–108.
- Kelleher, J., Thornton, K. R., Ashander, J., & Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *Plos Computational Biology*, 14(11), e1006581.
- Kim, J., Ji, M., & Yi, G. (2020). A review on sequence alignment algorithms for short reads based on next-generation sequencing. *IEEE Access*, 8, 189811–189822.
- Kjær, K. H., Winther Pedersen, M., De Sanctis, B., De Cahsan, B., Korneliussen, T. S., Michelsen, C. S., Sand, K. K., Jelavić, S., Ruter, A. H., Schmidt, A. M. A., Kjeldsen, K. K., Tesakov, A. S., Snowball, I., Gosse, J. C., Alsos, I. G., Wang, Y., Dockter, C., Rasmussen, M., Jørgensen, M. E., ... Willerslev, E. (2022). A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature*, 612, 283–291.

- Kolaczowski, B., & Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, *431*, 980–984.
- Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*, 4453–4455.
- Lam, T. T.-Y. (2020). Tracking the genomic footprints of SARS-CoV-2 transmission. *Trends in Genetics*, *36*, 544–546.
- Lanfear, R., & Mansfield, R. (2020). A global phylogeny of SARS-CoV-2 sequences from GISAID, version 3.0. *Github*, *Roblanf/sarscov2phylo*.
- Lee, B. T., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, C. M., Muthuraman, P., Nassar, L. R., Nguy, B., Pereira, T., Perez, G., Raney, B. J., Rosenbloom, K. R., Schmelter, D., Speir, M. L., ... Kent, W. J. (2022). The UCSC Genome Browser database: 2022 update. *Nucleic Acids Research*, *50*, D1115–D1122.
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Arxiv*.
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100.
- Li, H. (2021). New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, *37*, 4572–4574.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

- Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21(1), 265.
- Li, X., Giorgi, E. E., Marichannegowda, M. H., Foley, B., Xiao, C., Kong, X.-P., Chen, Y., Gnanakaran, S., Korber, B., & Gao, F. (2020). Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Science Advances*, 6, eabb9153.
- Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., ... Paten, B. (2023). A draft human pangenome reference. *Nature*, 617(7960), 312–324.
- Lu, J., Plessis, L. du, Liu, Z., Hill, V., Kang, M., Lin, H., Sun, J., François, S., Kraemer, M. U. G., Faria, N. R., McCrone, J. T., Peng, J., Xiong, Q., Yuan, R., Zeng, L., Zhou, P., Liang, C., Yi, L., Liu, J., ... Ke, C. (2020). Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell*, 181, 997–1003.e9.
- Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., ... Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, 395, 565–574.
- Marschall, T., Marz, M., Abeel, T., Dijkstra, L., Dutilh, B. E., Ghaffaari, A., Kersey, P. J., Kloosterman, W. P., Mäkinen, V., Novak, A. M., Paten, B., Porubsky, D., Rivals, E., Alkan, C., Baaijens, J. A., Bakker, P. I. W. de, Boeva, V., Bonnal, R. J. P.,

- Chiaromonte, F., ... Schönhuth, A. (2018). Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19(1), 118–135.
- Martiniano, R., De Sanctis, B., Hallast, P., & Durbin, R. (2022). Placing ancient DNA sequences into reference phylogenies. *Molecular Biology and Evolution*, 39(2), msac17. <https://doi.org/10.1093/molbev/msac017>
- Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11, 538.
- McBroome, J., Thornlow, B., Hinrichs, A. S., Kramer, A., De Maio, N., Goldman, N., Haussler, D., Corbett-Detig, R., & Turakhia, Y. (2021). A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. *Molecular Biology and Evolution*, 38(12), 5819–5824.
- Meredith, L. W., Hamilton, W. L., Warne, B., Houldcroft, C. J., Hosmillo, M., Jahun, A. S., Curran, M. D., Parmar, S., Caller, L. G., Caddy, S. L., Khokhar, F. A., Yakovleva, A., Hall, G., Feltwell, T., Forrest, S., Sridhar, S., Weekes, M. P., Baker, S., Brown, N., ... Goodfellow, I. (2020). Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infectious Diseases*, 20, 1263–1272.
- Miller, J. H., & Simpson, C. (2022). When did mammoths go extinct?. *Nature*, 612, E1–E3.
- Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Haeseler, A. von, & Lanfear, R. (2020). IQ-TREE 2: new models and efficient

- methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*, 1530–1534.
- Moreno, G. K., Braun, K. M., Riemersma, K. K., Martin, M. A., Halfmann, P. J., Crooks, C. M., Prall, T., Baker, D., Baczenas, J. J., Heffron, A. S., Ramuta, M., Khubbar, M., Weiler, A. M., Accola, M. A., Rehrauer, W. M., O'Connor, S. L., Safdar, N., Pepperell, C. S., Dasu, T., ... Friedrich, T. C. (2020). Revealing fine-scale spatiotemporal differences in SARS-CoV-2 introduction and spread. *Nature Communications*, *11*, 5558.
- Nylander, J. A. A., Olsson, U., Alström, P., & Sanmartín, I. (2008). Accounting for phylogenetic uncertainty in biogeography: a Bayesian approach to dispersal-vicariance analysis of the thrushes (Aves: Turdus). *Systematic Biology*, *57*, 257–268.
- Oliva, A., Tobler, R., Llamas, B., & Souilmi, Y. (2021). Additional evaluations show that specific BWA-aln settings still outperform BWA-mem for ancient DNA data alignment. *Ecology and Evolution*, *11*(24), 18743–18748.
- Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526–528.
- Park, A. K., Kim, I.-H., Kim, J., Kim, J.-M., Kim, H. M., Lee, C. Y., Han, M.-G., Rhie, G.-E., Kwon, D., Nam, J.-G., Park, Y.-J., Gwack, J., Lee, N.-J., Woo, S., No, J. S., Lee, J., Ha, J., Rhee, J., Yoo, C.-K., & Kim, E.-J. (2021). Genomic surveillance of SARS-CoV-2: distribution of clades in the Republic of Korea in 2020. *Osong Public Health and Research Perspectives*, *12*, 37–43.
- Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarszewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome

- phylogeny substantially revises the tree of life. *Nature Biotechnology*, 36, 996–1004.
- Peng, J., Liu, J., Mann, S. A., Mitchell, A. M., Laurie, M. T., Sunshine, S., Pilarowski, G., Ayscue, P., Kistler, A., Vanaerschot, M., Li, L. M., McGeever, A., Chow, E. D., Marquez, C., Nakamura, R., Rubio, L., Chamie, G., Jones, D., Jacobo, J., ... IDseq Team. (2021). Estimation of secondary household attack rates for emergent spike L452R SARS-CoV-2 variants detected by genomic surveillance at a community-based testing site in San Francisco. *Clinical Infectious Diseases*, 74, 32–39.
- Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., & Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology*, 5, 50.
- Pipes, L., Chen, Z., Afanaseva, S., & Nielsen, R. (2022). Estimating the relative proportions of SARS-CoV-2 haplotypes from wastewater samples. *Cell Reports Methods*, 2, 100313.
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. *Plos One*, 5, e9490.
- Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., Plessis, L. du, & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, 5, 1403–1407.
- Ree, R. H., Moore, B. R., Webb, C. O., & Donoghue, M. J. (2005). A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution*, 59, 2299–2311.

- Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of next-generation sequencing reads. *Annual Review of Genomics and Human Genetics*, 16, 133–151.
- Rice, E. S., Alberdi, A., Alfieri, J., Athrey, G., Balacco, J. R., Bardou, P., Blackmon, H., Charles, M., Cheng, H. H., Fedrigo, O., Fiddaman, S. R., Formenti, G., Frantz, L. A. F., Gilbert, M. T. P., Hearn, C. J., Jarvis, E. D., Klopp, C., Marcos, S., Mason, A. S., ... Warren, W. C. (2023). A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants. *BMC Biology*, 21(1), 267.
- Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53, 131–147.
- Romashchenko, N., Linard, B., Pardi, F., & Rivals, É. (2023). EPIK: precise and scalable evolutionary placement with informative k-mers. *Bioinformatics*, 39(12), btad692. <https://doi.org/10.1093/bioinformatics/btad692>
- Sanderson, T. (2021). Chronumental: time tree estimation from very large phylogenies. *Biorxiv*, 2021.10.27.465994.
- Sanderson, T. (2022). Taxonium, a web-based tool for exploring large phylogenetic trees. *Elife*, 11, e82392. <https://doi.org/10.7554/eLife.82392>
- Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Schoch, C. L., Sherry, S. T., & Karsch-Mizrachi, I. (2021). GenBank. *Nucleic Acids Research*, 49, D92–D96.
- Shu, Y., & McCauley, J. (2017). GISAID: global initiative on sharing all influenza data—from vision to reality. *Euro Surveillance*, 22, 30494.

- Simpson, J. T., & Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, 22, 549–556.
- Sirén, J., & Paten, B. (2022). GBZ file format for pangenome graphs. *Bioinformatics*, 38, 5012–5018.
- Sirén, J., Eskandar, P., Ungaro, M. T., Hickey, G., Eizenga, J. M., Novak, A. M., Chang, X., Chang, P.-C., Kolmogorov, M., Carroll, A., Monlong, J., & Paten, B. (2024). Personalized pangenome references. *Nature Methods*. <https://doi.org/10.1038/s41592-024-02407-2>
- Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J. A., Hickey, G., Chang, P.-C., Carroll, A., Gupta, N., Gabriel, S., Blackwell, T. W., Ratan, A., Taylor, K. D., Rich, S. S., Rotter, J. I., Haussler, D., Garrison, E., & Paten, B. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, 374(6574), abg8871. <https://doi.org/10.1126/science.abg8871>
- Skidmore, P. T., Kaelin, E. A., Holland, L. R. A., Maqsood, R., Wu, L. I., Mellor, N. J., Blain, J. M., Harris, V., LaBaer, J., Murugan, V., & Lim, E. S. (2021). Genomic sequencing of SARS-CoV-2 variant B.1.243.1, Arizona, USA. *Emerging Infectious Diseases*, 27, 2718–2720.
- Slon, V., Hopfe, C., Weiß, C. L., Mafessoni, F., Rasilla, M. de la, Lalueza-Fox, C., Rosas, A., Soressi, M., Knul, M. V., Miller, R., Stewart, J. R., Derevianko, A. P., Jacobs, Z., Li, B., Roberts, R. G., Shunkov, M. V., Lumley, H. de, Perrenoud, C., Gušić, I., ... Meyer, M. (2017). Neandertal and Denisovan DNA from Pleistocene sediments. *Science*, 356, 605–608.

- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30, 1312–1313.
- Steel, M. A., & Penny, D. (1993). Distributions of tree comparison metrics—some new results. *Systematic Biology*, 42, 126–141.
- Sullivan, J., & Swofford, D. L. (2001). Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated?. *Systematic Biology*, 50, 723–729.
- Swofford, D., & Maddison, P. (1987). Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences*, 87, 199–229.
- Tang, J. W., Toovey, O. T. R., Harvey, K. N., & Hui, D. D. S. (2021). Introduction of the South African SARS-CoV-2 variant 501Y.V2 into the UK. *Journal of Infection*, 82, e8–e10.
- Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N., Mlisana, K., Gottberg, A. von, Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A. J., Engelbrecht, S., Van Zyl, G., ... Oliveira, T. de. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592, 438–443.
- Thornlow, B., Cheng, Y., De Maio, N., McBroome, J., Hinrichs, A., Lanfear, R., Turakhia, Y., Corbett-Detig, R., & Kramer, A. (2022). Online phylogenetics with matOptimize for SARS-CoV-2 [Dataset]. *Dryad Digital Repository*. <https://doi.org/10.7291/D13Q2J>

- Thornlow, B., Hinrichs, A. S., Jain, M., Dhillon, N., La, S., Kapp, J. D., Anigbogu, I., Cassatt-Johnstone, M., McBroome, J., Haeussler, M., Turakhia, Y., Chang, T., Olsen, H. E., Sanford, J., Stone, M., Vaske, O., Bjork, I., Akeson, M., Shapiro, B., ... Corbett-Detig, R. (2021). A new SARS-CoV-2 lineage that shares mutations with known variants of concern is rejected by automated sequence repository quality control. *Biorxiv*.
- Thornlow, B., Lanfear, R., Kramer, A., Corbett-Detig, R., Turakhia, Y., & Ye, C. (2021). Online phylogenetics using parsimony supplemental repository. *Github*, *Bpt26/parsimony*.
- Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, *15*, 524.
- Turakhia, Y., De Maio, N., Thornlow, B., Gozashti, L., Lanfear, R., Walker, C. R., Hinrichs, A. S., Fernandes, J. D., Borges, R., Slodkiewicz, G., Weilguny, L., Haussler, D., Goldman, N., & Corbett-Detig, R. (2020). Stability of SARS-CoV-2 phylogenies. *Plos Genetics*, *16*, e1009175.
- Turakhia, Y., Thornlow, B., Hinrichs, A. S., & McBroome, J. (2021). Pandemic-scale phylogenomics reveals elevated recombination rates in the SARS-CoV-2 spike region. *Biorxiv*.
- Turakhia, Y., Thornlow, B., Hinrichs, A. S., 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. *Nature Genetics*, *53*(6), 809–816. <https://doi.org/10.1038/s41588-021-00862-7>

- Umair, M., Ikram, A., Salman, M., Khurshid, A., Alam, M., Badar, N., Suleman, R., Tahir, F., Sharif, S., Montgomery, J., Whitmer, S., & Klena, J. (2021). Whole-genome sequencing of SARS-CoV-2 reveals the detection of G614 variant in Pakistan. *Plos One*, *16*, e248371.
- Vogel, N. A., Rubin, J. D., Pedersen, A. G., Sackett, P. W., Pedersen, M. W., & Renaud, G. (2024). Soibean: high-resolution taxonomic identification of ancient environmental DNA using mitochondrial pangenome graphs. *Molecular Biology and Evolution*, *41*, msae203.
- Vogel, N. A., Rubin, J. D., Swartz, M., Vlieghe, J., Sackett, P. W., Pedersen, A. G., Pedersen, M. W., & Renaud, G. (2023). euka: robust tetrapodic and arthropodic taxa detection from modern and ancient environmental DNA using pangenomic reference graphs. *Methods in Ecology and Evolution*, *14*, 2717–2727.
- Walia, S., Motwani, H., Tseng, Y.-H., Smith, K., Corbett-Detig, R., & Turakhia, Y. (2026). Compressive pangenomics using mutation-annotated networks. *Nature Genetics*, *58*, 445–453.
- Wang, W., Barbetti, J., Wong, T., Thornlow, B., Corbett-Detig, R., Turakhia, Y., Lanfear, R., & Minh, B. Q. (2022). DecentTree: scalable neighbour-joining for the genomic era. *Biorxiv*, 2022.04.10.487712.
- Wang, Y., Korneliussen, T. S., Holman, L. E., Manica, A., & Pedersen, M. W. (2022). ngsLCA—a toolkit for fast and flexible lowest common ancestor inference and taxonomic profiling of metagenomic data. *Methods in Ecology and Evolution*, *13*, 2699–2708.

- Wang, Y., Pedersen, M. W., Alsos, I. G., De Sanctis, B., Racimo, F., Prohaska, A., Coissac, E., Owens, H. L., Merkel, M. K. F., Fernandez-Guerra, A., Rouillard, A., Lammers, Y., Alberti, A., Denoeud, F., Money, D., Ruter, A. H., McColl, H., Larsen, N. K., Cherezova, A. A., ... Willerslev, E. (2021). Late Quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature*, *600*, 86–92.
- Wertheim, J. O., Steel, M., & Sanderson, M. J. (2022). Accuracy in near-perfect virus phylogenies. *Systematic Biology*, *71*, 426–438.
- Ye, C., Thornlow, B., Hinrichs, A., Kramer, A., Mirchandani, C., Torvi, D., Lanfear, R., Corbett-Detig, R., & Turakhia, Y. (2022). matOptimize: a parallel tree optimization method enables online phylogenetics for SARS-CoV-2. *Bioinformatics*, *38*, 3734–3740.
- Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., Zhang, J., Lyu, H., Lin, T., Gao, Q., Saha, S., Mueller, L., Fei, Z., Städler, T., Xu, S., ... Huang, S. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, *606*(7914), 527–534.