nature methods

Article

https://doi.org/10.1038/s41592-025-02932-8

Rate variation and recurrent sequence errors in pandemic-scale phylogenetics

Received: 21 September 2024

Accepted: 24 October 2025

Published online: 9 February 2026

Check for updates

Nicola De Maio[1] ✉, Myrthe Willemsen[1,7], Samuel Martin[1], Zihao Guo[1,8], Abhratanu Saha[1,9], Martin Hunt[1,2,3,4], Nhan Ly-Trong[5], Bui Quang Minh[5], Zamin Iqbal[1,6] & Nick Goldman[1]

Phylogenetic analyses of genome sequences from infectious pathogens reveal essential information regarding their evolution and transmission, as seen during the coronavirus disease 2019 pandemic. Recently developed pandemic-scale phylogenetic inference methods reduce the computational demand of phylogenetic reconstruction from genomic epidemiological datasets, allowing the analysis of millions of closely related genomes. However, widespread homoplasies, due to recurrent mutations and sequence errors, cause phylogenetic uncertainty and biases. We present algorithms and models to substantially improve the computational performance and accuracy of pandemic-scale phylogenetics. In particular, we account for, and identify, mutation rate variation and recurrent sequence errors. We reconstruct a reliable and public sequence alignment and phylogenetic tree of >2 million severe acute respiratory syndrome coronavirus 2 genomes encapsulating the evolutionary history and global spread of the virus up to February 2023.

Genomic epidemiology has become a vital tool in regional, national and global health, as exemplified during the coronavirus disease 2019 (COVID-19) pandemic[1–4]. It is a priority for national and international research, and its role in the interpretation and control of transmission of diverse human pathogens is expected to increase in future. Genomic epidemiology can reveal details about infectious pathogen biology[5], evolution[6], transmission[7] and effectiveness of containment measures[3]. These inform downstream research, but also have immediate effects for policymakers, in vaccine manufacturing, and in other areas.

Analysis of genomic epidemiological data relies heavily on phylogenetics. However, well-established phylogenetic methods have mostly been developed for inter-species evolutionary biology. Consequently, phylogenetic analysis of genomic epidemiological data with these methods is particularly challenging[8,9]. Part of this challenge is due to the unprecedented size of such datasets, in particular those of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes, with currently more than 20 million SARS-CoV-2 genomes shared globally. With ongoing improvement and widespread adoption of genome sequencing technologies, the sizes of genomic epidemiology datasets are expected to further increase in the future. Recently developed pandemic-scale phylogenetic methods, such as UShER[10,11] and MAPLE[12], address the problem of computational demand in large genomic epidemiology datasets by using algorithms and statistical tools specifically developed for this type of data. However, other difficulties affecting the analysis of such data remain. Of these, homoplasies (apparently recurring nucleotide substitutions along the phylogenetic tree) substantially affect the accuracy and reliability of phylogenetic inference[8]. We identified two predominant factors contributing to widespread homoplasies in SARS-CoV-2: (a) highly mutable nucleotides and genome positions[13], and (b) recurrent sequence errors[14,15].

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridgeshire, UK. [2]Nuffield Department of Medicine, University of Oxford, Oxford, UK. [3]National Institute of Health Research Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK. [4]Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford, Oxford, UK. [5]School of Computing, College of Engineering, Computing and Cybernetics, Australian National University, Canberra, Australian Capital Territory, Australia. [6]Milner Centre for Evolution, University of Bath, Bath, UK. [7]Present address: University Medical Center Utrecht, Utrecht, the Netherlands. [8]Present address: Télécom SudParis, Institut Polytechnique de Paris, Palaiseau, France. [9]Present address: Institut Curie, PSL Research University, CNRS, Paris, France. ✉e-mail: demaio@ebi.ac.uk

Addressing these two phenomena is expected to lead to more reliable phylogenetic estimation.

(a) Accounting for heterogeneity in evolutionary rates is common and consequential in phylogenetics[16]. In popular phylogenetic models of rate variation, each genome position is allowed to evolve at any of a given number of rates[17], which usually leads to more accurate inference[18]. These approaches, however, substantially increase computational demand[19] and might not be able to account for the extreme level of mutation rate variation observed in genomic epidemiological datasets[13]. To overcome these problems, here we propose a computationally efficient model in which each genome position is assigned its own estimated mutation rate, similar to the SLR model[20] of heterogeneous selection pressure. This approach leverages the large size of available datasets to model in detail highly variable mutation rates.

(b) Recurrent sequence errors are ubiquitous in SARS-CoV-2 genomic data[14,15], and are highly detrimental to phylogenetic estimation due to the low levels of sequence divergence in such data[15]. Here, we extend previously proposed phylogenetic sequence error models (which typically assume homogeneity in errors along the genome[21,22]) to explicitly model and infer position-specific sequence error probabilities. This allows us to not only effectively identify positions with recurrent errors, which can then be masked from the sequence alignment[14,15,23], but also to account for these errors during phylogenetic inference itself.

We implemented these models within MAPLE, together with algorithmic improvements, to reduce the computational demand of the software. We use these tools, millions of publicly shared SARS-CoV-2 genomes, and recent advancements in consensus calling methods to infer the recurrent mutation and error landscape of the virus and to reconstruct a publicly available, reliable and global SARS-CoV-2 phylogenetic tree.

Models and algorithms

Models and algorithms presented here are discussed in detail in Methods and have been implemented within our approximate maximum likelihood phylogenetic software MAPLE v0.6.8 (https://github.com/NicolaDM/MAPLE/).

Models of rate variation and recurrent sequence errors

In addition to the GTR[24] substitution model, we consider three extensions:

1. The more general nonstationary, nonreversible UNREST nucleotide model[25], which can account for the nonstationary genome evolution of SARS-CoV-2 in human hosts[13]. Because MAPLE avoids matrix exponentiation[12], our estimation is not affected by numerical instability. Given the low divergence in the considered datasets, we approximate the root nucleotide frequencies as identical to the observed nucleotide frequencies in the alignment.
2. We extend the UNREST model to account for rate variation. Differently from classical random variable models in phylogenetics[17], our approach allows one rate per genome position, assigning one free parameter to each site to model its position-specific substitution rate. This approach is highly parameterized, and, therefore, only suitable for large genomic datasets, and allows us to account for elevated variation in substitution rates, and in particular highly mutable sites, which are a cause of extensive homoplasies in SARS-CoV-2 (refs. 13,15). This model is described in detail in Methods.
3. We extend the rate variation UNREST model above to account for recurrent sequence errors. In this model each position of the genome is assigned a site-specific error probability free parameter. Again, this site-specific sequence error model is highly parameterized and is only suitable for large datasets. More details about this model are given in Methods.

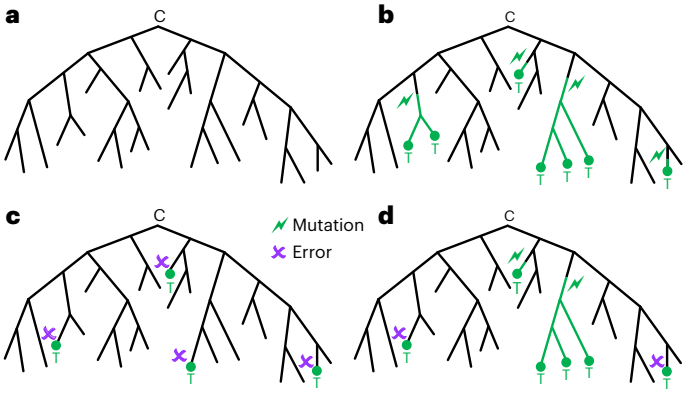


Fig. 1 | Graphical examples of the effects of substitution rate and error probability variation. In these examples, nucleotide C is the root state, while nucleotide T (shown as green in the terminal nodes and on branches) is the derived nucleotide. For simplicity, we show a very small phylogeny. In practice, in SARS-COV-2 phylogenetic trees substitutions are more numerous and much sparser, and so there is less uncertainty in substitution history and parameter inference. **a**, When mutation rate and error probability at a given genome position are both 0, the corresponding alignment column will have only one nucleotide and there will be no substitutions. **b**, The presence of mutations (highlighted with a lightning bolt and connected to their descendant observed genomes by green lines), but not sequence errors, is expected to result in a balanced proportion of observed substitutions in internal and terminal branches (here as an example we have two mutations on internal branches and two on terminal ones). **c**, A site with no mutations (0 mutation rate) but with recurrent errors (shown as crosses) results in an enrichment of substitutions at terminal branches (here we observe four substitutions on terminal branches). **d**, A site with positive mutation rate and error probability is expected to result in substitutions at internal branches, but also an enrichment of substitutions at terminal branches (here one observed substitution at an internal branch and three at terminal ones). When investigating real data, we do not know which terminal branch substitutions are the results of mutations and which are the result of errors. However, we use the enrichment of substitutions at terminal branches relative to substitutions on internal branches to infer the prevalence of errors and mutations at the considered genome position.

We only consider the rate variation model as an extension of the UNREST substitution model. Our rationale is that using a more restricted substitution model such as GTR in conjunction with rate variation would mean that differences in substitution rates across nucleotides that are not accounted for by the GTR model (for example, elevated C-to-T and G-to-T mutation rates increasing genome T content over time[13]) might lead to wrongly inferred rate variation (for example, higher rates at C and G genome positions). Similarly, we only consider the error model as an extension of the rate variation model, that is, we never estimate error probabilities without also estimating rate variation. The reason is that, without accounting for rate variation, the error model might wrongly interpret excess mutations at a site as the result of high error probability.

We efficiently infer the values of the thousands of free parameters in our more complex models using expectation maximization (EM; Methods). We give a graphical example in Fig. 1 to explain how we can estimate mutation rate and error probability jointly for every position of the genome. Without sequence errors, we expect substitutions to happen along the phylogenetic tree at the same rate for every branch, and so on terminal and internal branches indiscriminately (for example, Fig. 1b) and proportionally to their length. A recurrent sequence error instead is expected to cause an enrichment only in the number of substitutions observed at terminal branches (for example, Fig. 1c,d). We estimate site-specific substitution rates and error probabilities by distinguishing substitutions at internal versus terminal branches, while at the same time accounting for different substitution rates for

different nucleotides, for branch lengths, and for uncertainty in the nucleotide substitution history.

Local references

MAPLE performs phylogenetic analysis of vast genomic epidemiological data by representing closely related observed genome sequences, reconstructed ancestral genomes and their uncertainty, in terms of differences with respect to a reference genome[12]. However, as an outbreak progresses, divergence with respect to the reference genome increases, and as such the computational demand of storing and processing genetic data under this paradigm also increases. In principle, information could be stored more efficiently: a mutation-annotated tree[26] can represent observed genomic sequences efficiently by recording estimated mutation events on the tree. This means that an observed genome can be reconstructed by tracing, along a phylogenetic tree, the inferred mutational history separating this genome from the root, ancestral genome. This way, each mutation event only needs to be represented once in the phylogenetic tree, and does not need to be explicitly represented in the descendants of the branch where the mutation occurred.

The problem with using a mutation-annotated tree in phylogenetics is, however, that information for two distant nodes of the tree cannot be directly compared. For any comparison, one needs to traverse all of the branches in the tree separating the two considered nodes and update their information ('translating' it so that both are represented based on the same background reference) accordingly. We found that while this approach leads to a reduction in memory consumption, it did not substantially reduce runtime. Instead, our algorithm selects a small number of tree nodes and uses them as local references (Extended Data Fig. 1). A useful reference is not too similar to another reference, and is used to represent sufficiently many genomes. For this reason we select nodes based on the number of mutations separating them from their most recent reference ancestor, and based on the number of genomes descending from them. Genetic sequence information and uncertainty at nodes in the tree is then represented in terms of their closest reference ancestor. The advantage of this approach is that, when comparing information across different nodes in the tree, only a small number of 'translations' needs to be carried out (one for each traversed local reference), while substantial computational demand is saved thanks to the more concise representation of genetic sequence information.

Other added features

In Methods we present additional improvements we made to MAPLE, such as allowing sample placement onto an initial tree for online phylogenetic inference, improved topological search, efficient tree root search, fast and accurate branch length estimation, and parallelization of the subtree prune and regraft (SPR) search.

Computational demand

Our improvements lead to approximately halving the runtime and memory demand of MAPLE (Extended Data Table 1). Runtime can further be reduced by parallelizing the topological search, but since this is currently done in a distributed memory framework, it also leads to an increase in memory demand (Extended Data Table 1 and Extended Data Fig. 2). Also, expectedly, attempting to parallelize with too many cores leads to a reduction in performance.

Our rate variation and error models lead to limited additional computational cost compared to the more basic models (Extended Data Table 1 and Fig. 2a,b). Due to the low divergence of the considered datasets, our EM approach infers thousands of model parameters (such as site-specific rates and error probabilities) at limited additional time and memory demand. However, more complex models in MAPLE can still lead to some increase in runtime, for example, due to the increased complexity of the tree search associated with the lowering of the likelihood cost of the most recurrent mutation events.

Simulation-based benchmark

To assess the accuracy of our methods, we aimed at simulating realistic SARS-CoV-2 genome data including both mutation rate variation and position-specific error probabilities (Methods). Knowing the exact phylogenetic tree used in simulations and the position of sequence errors generated in the genome data, we can assess MAPLE's accuracy in estimating both. We find that our more complex models, and in particular the error model, substantially improve phylogenetic tree inference accuracy (Fig. 2c). MAPLE needs sufficiently large datasets to reliably identify position-specific error parameters and errors in the input genome sequences (Fig. 2d,e; regarding real data see Methods and Extended Data Fig. 3), with accuracy increasing as more sequences are considered. With 200,000 genomes, we have approximately 92% precision and 83% sensitivity in estimating individual sequence errors in our simulations.

Toward reliable, global and public SARS-CoV-2 alignment and phylogeny

Currently, millions of SARS-CoV-2 genome sequencing raw read datasets are shared publicly[27]. This represents a unique opportunity to reliably and consistently call millions of consensus genomes, and to corroborate putative recurrent consensus sequence errors.

To validate our methods, we collected millions of SARS-CoV-2 genome sequencing read datasets (Methods). First, we investigated highly recurrent sequence errors. Then, we filtered out putative recurrent sequence errors and possibly contaminated samples to obtain a reliable, global and public SARS-CoV-2 sequence alignment and phylogeny (Methods).

Investigation of recurrent sequence errors

We collected SARS-CoV-2 genomes with publicly available sequencing raw read data, a consensus sequence in GenBank and a Viridian[28] consensus genome. We consider Viridian genomes here since they have been consistently assembled and called using an approach that addresses reference biases affecting consensus genomes from public databases[28]. After filtering out low-coverage genomes and samples that might have been affected by contamination or mixed infections, our collection contained 2,993,121 genomes (Methods). For these, we created two alignments, one containing the GenBank consensus genomes, and one containing the Viridian ones. Given that we want to investigate the presence and abundance of recurrent sequence errors in these alignments, for now we do not perform any prior masking of alignment columns (for example, refs. 14,23). From each alignment we inferred a phylogeny and recurrent sequence errors with MAPLE (Fig. 3a,b). We then investigated read data and the phylogeny to corroborate candidate recurrent sequencing or bioinformatics artifacts in the consensus genomes. In particular, we focused on positions inferred to contain errors with a frequency around or above 0.01% of genomes. We considered the following red flags of possible recurrent artifacts:

1. The presence, at the considered sample and genome position, of multiple nucleotides supported at high frequency (>5%, which is higher than that of simple sequencing errors[29]) by reads. This pattern can be, for example, a result of primers binding at unintended sites of homology[14].
2. Elevated numbers of IUPAC ambiguity characters in the consensus sequences. In the presence of high-frequency alternative nucleotides, consensus calling pipelines can include ambiguity characters instead of calling a nucleotide. Recurrent ambiguity characters are, therefore, a symptom of recurrent alternative nucleotides in the reads at the considered genome position.
3. Frequent different consensus sequence calls by Viridian and GenBank at the considered position. This suggests that different protocols for read mapping, filtering and trimming, or consensus

https://doi.org/10.1038/s41592-025-02932-8

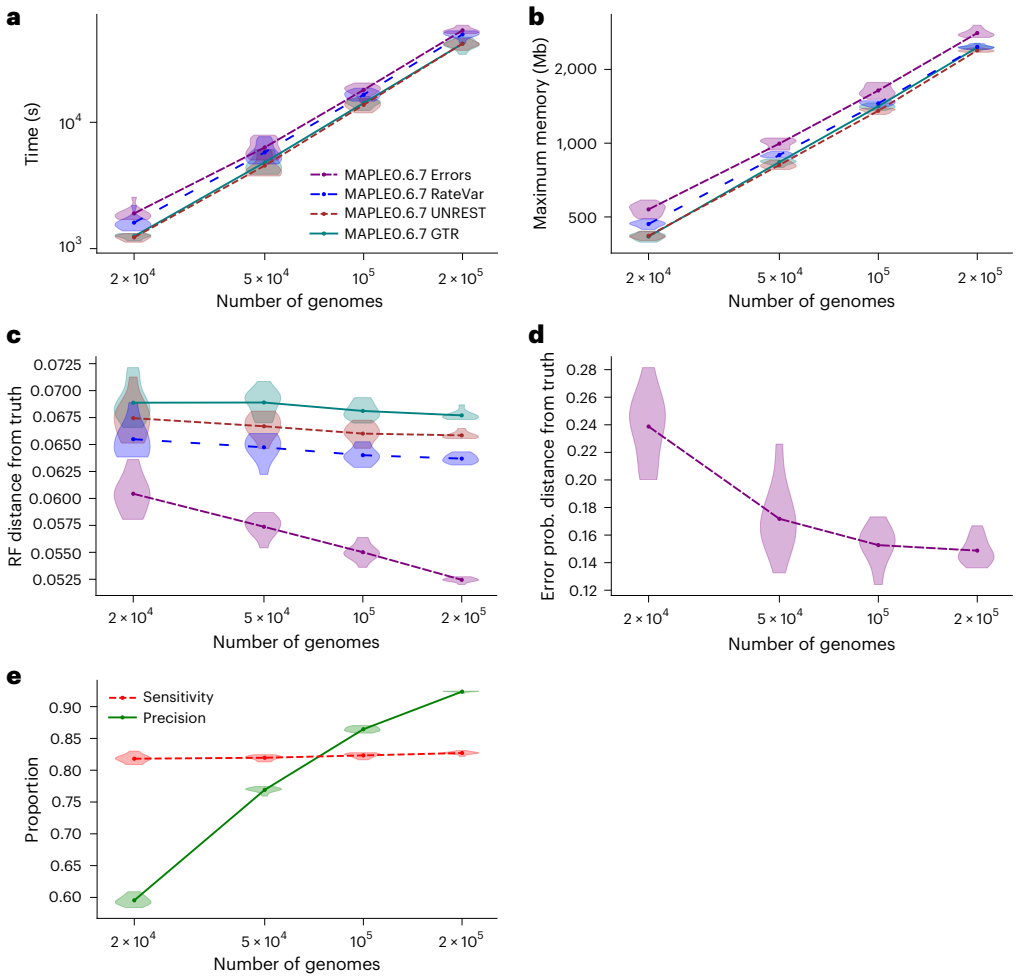


Fig. 2 | Simulation-based benchmark. Assessment of different models using SARS-CoV-2-like genome data simulated with rate variation and recurrent sequence errors (Methods). On the *x* axis are genome subset sizes. For each size and model considered we ran ten replicates. Dots show the mean across replicates, while violin plots show variation between replicates. The four models considered are GTR, UNREST, UNREST with rate variation ('RateVar'), and UNREST with rate variation and sequence errors ('Errors'). **a**,**b**, Time demand (**a**) and maximum memory demand (**b**) for non-parallelized tree inference from scratch. **c**, Normalized Robinson–Foulds (RF) distances (accounting for multifurcations; Methods) between simulated ('true') phylogenetic trees and estimated ones; lower values represent more accurate estimations of tree topology. **d**, Euclidean distance between the vectors of estimated and simulated position-specific sequence error rates; lower values represent more accurate estimates of position-specific error parameters. prob., probability. **e**, Precision of inference of individual sequence errors (true positives divided by total inferred errors) and sensitivity (true positives divided by total simulated errors). We used a probability threshold of 50% to define inferred sequence errors within input genome sequences. The total number of simulated errors was ≈0.22 per genome. All analyses were run on one core of an Intel Xeon Gold 6252 Processor at 2.10 GHz.

calling thresholds might systematically affect the consensus sequences at the considered position.

4. Phylogenetic clustering of apparent substitutions caused by putative errors. This can happen if errors are genotype dependent (for example, a real mutation might affect primer binding, or an indel might affect quality of read alignment to the reference genome), or if errors are caused by sequencing protocols whose adoption might vary by time and location. We used Taxonium[30] to visually investigate mutation-annotated phylogenies inferred by MAPLE.
5. High mutation rate estimated by MAPLE at the considered position. MAPLE's error model assumptions can be substantially violated if the frequency of errors at a given position is too high (for example, >1%), if they are correlated with other errors, or if they are not uniformly distributed along the phylogenetic tree. In these cases, portions of the recurrent errors can be interpreted by MAPLE as genuine mutations, leading to high estimated mutation rates at positions affected by recurrent errors.

While analyzing the GenBank consensus genome alignment, we noticed that many sites of high-frequency mutations (those with hundreds of thousands of descendants) had large numbers of artificial reversions to the reference genome, resulting from genomes with low coverage at these positions. Some of such sites with the most putative erroneous reversions in GenBank consensus genomes are positions 210, 7124, 9053, 21618, 22813, 23854, 26107, 26577, 27373, 28881, 28916, 29402 and 29742 of the SARS-CoV-2 Wuhan-Hu-1 reference genome MN908947.3 (ref. 31). This appears to be the result of common reference biases in consensus calling pipelines, which are addressed by Viridian[28]. Masking these and many other positions from the GenBank genomes would remove too much essential information, so we focused the rest of our analyses on the Viridian consensus sequences, and used these as the basis to create a reliable global SARS-CoV-2 alignment and phylogeny.

A few positions of the Viridian genomes were inferred to contain recurrent errors, particularly positions 8835 and 15521, which were previously also highlighted as problematic, as they are frequently affected by artifacts caused by ARTIC V4 primers binding to unintended genome locations[14,23]. In Extended Data Table 2, we list positions inferred to be

https://doi.org/10.1038/s41592-025-02932-8

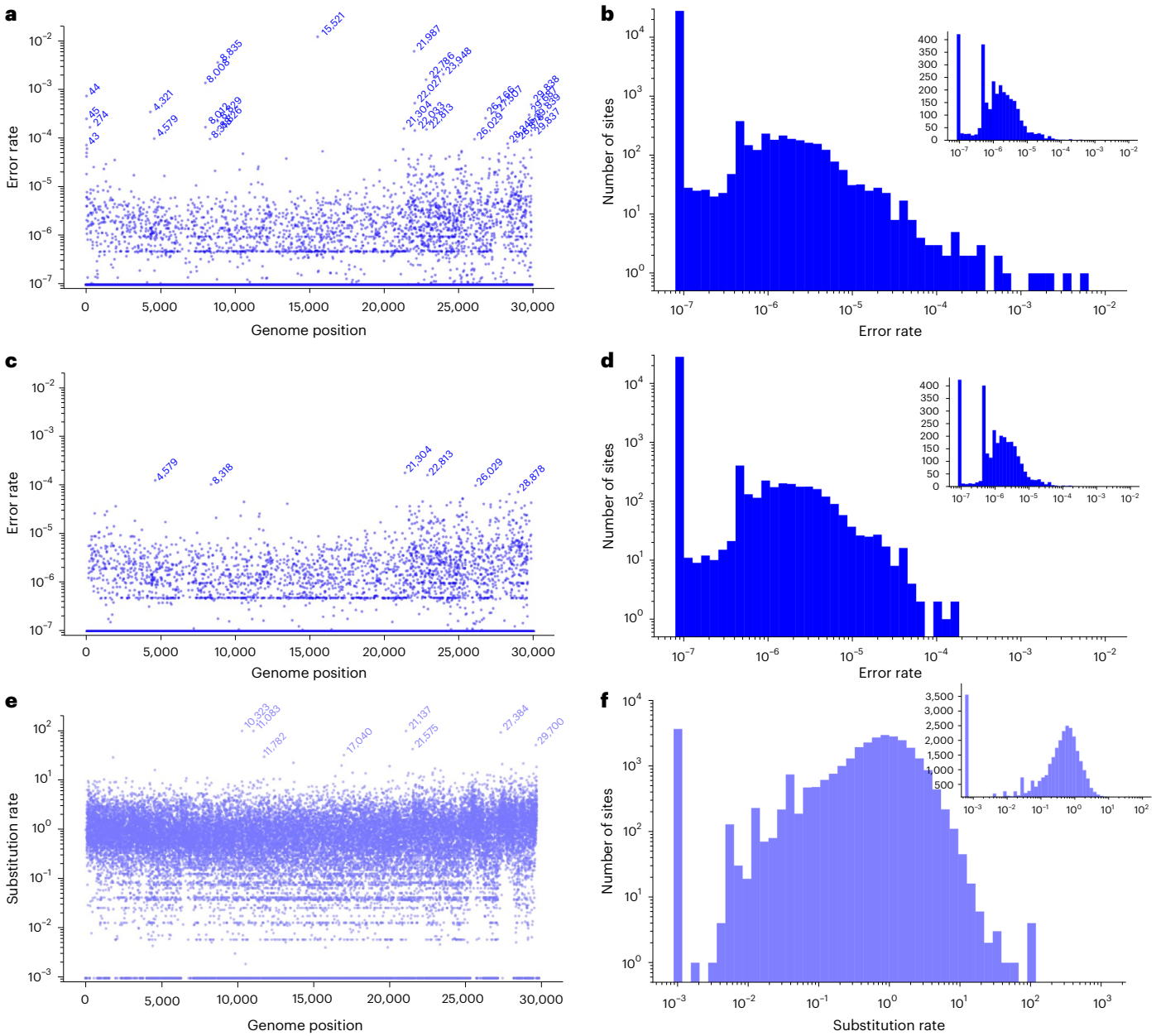


Fig. 3 | Estimated position-specific SARS-CoV-2 error and substitution rates. Site-specific values of the sequence error probabilities (minimum 10^{-7} allowed) and relative substitution rates (maximum 100 and minimum 0.001 allowed) estimated under the rate variation model in MAPLE v0.6.8 from the alignment of >2 million Viridian consensus SARS-CoV-2 genomes. **a**, Distribution of inferred error probabilities (*y* axis, logarithmic scale) along the genome (*x* axis). High values of error probabilities (>0.00007) are annotated with the corresponding genome position. Here estimates are obtained from the initial alignment without any column masking. **b**, Histogram of position-specific error probabilities from **a**; the plot uses a logarithmic scale along the *y* axis (and *x* axis), while the inset shows the same histogram with a linear *y* axis. **c**, As in **a**, but here error probabilities are estimated from the final alignment with columns masked as in Extended Data Table 2. **d**, As in **b**, but showing the error probabilities from **c**. **e**, Distribution of the inferred substitution rates (*y* axis, logarithmic scale) along the genome (*x* axis). Values > 30 are annotated with their genome position; estimates derived from the same run as **c**. **f**, Histogram of the position-specific substitution rates from **e**. Note logarithmic axes, and linear *y* axis in the inset. See also Extended Data Fig. 5 for comparisons of site-wise estimates of error and substitution rates.

error-prone in SARS-CoV-2 Viridian genomes and that we mask from the alignment.

Other positions inferred by MAPLE to contain moderate numbers of errors (in the order of hundreds; Fig. 3c,d) showed no red flags corroborating possible recurrent consensus sequence errors. We suspect that these might be caused by other phenomena resulting in an enrichment in the number of substitutions on terminal branches: for example, deleterious recurrent mutations. Positions that seem most affected by this pattern include 4579, 8318, 11522, 13424, 21255, 21550, 22487, 23416, 25163, 25166 and 26029.

Although we now focus only on Viridian consensus genomes—due to the previously mentioned highly recurrent reference-biased errors in GenBank genomes—in cases where the analysis of GenBank genomes is of specific interest, in addition to masking relevant positions highlighted in Extended Data Table 2, we also recommend masking the following positions that we identified with a similar approach to the one just described: 241 (likely thousands of errors), 12422 (hundreds of errors), 27904 (thousands of errors), and the cluster 19209, 19210, 19212, 19214 and 19217 (hundreds of highly correlated errors).

https://doi.org/10.1038/s41592-025-02932-8

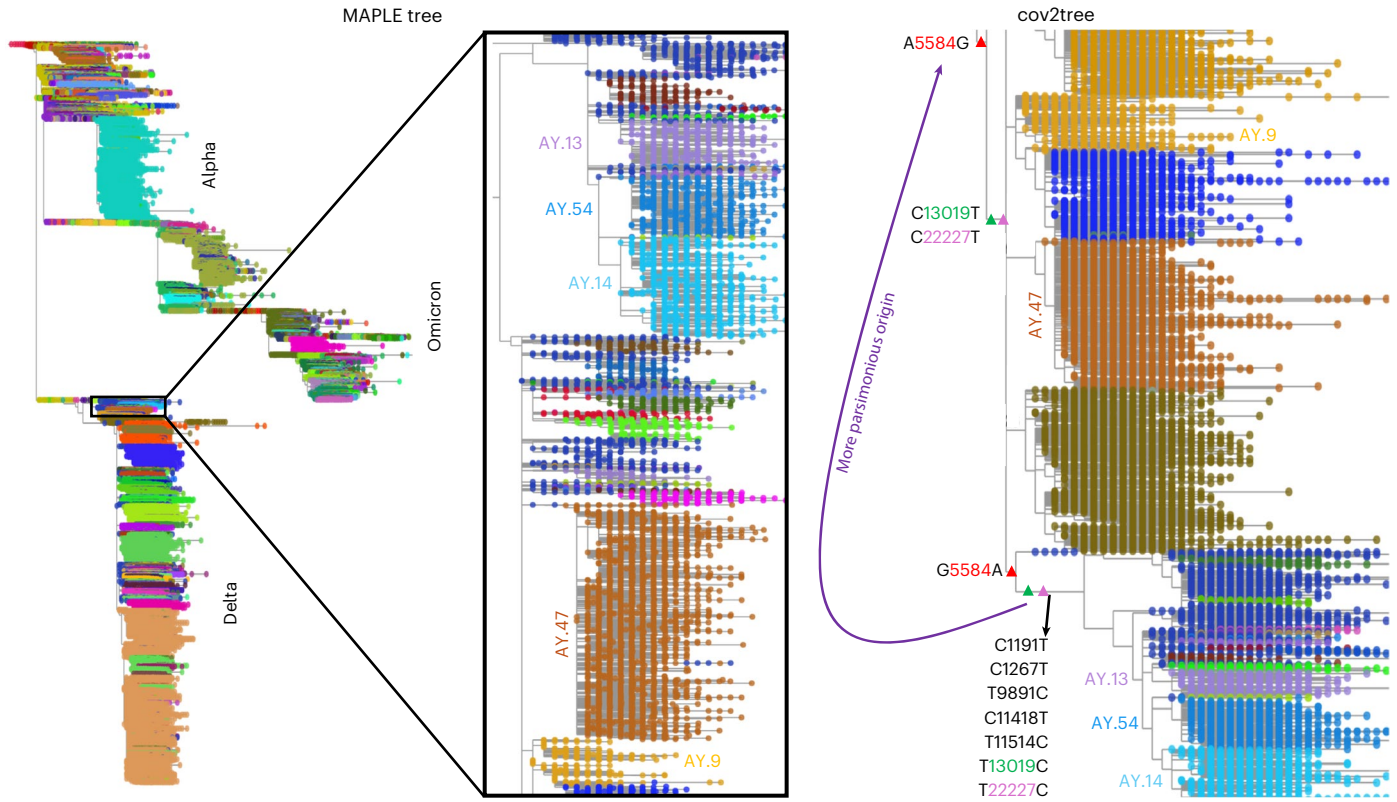


Fig. 4 | Comparison of the evolutionary histories of the AY SARS-COV-2 lineage inferred by MAPLE and cov2tree. On the left is the global SARS-CoV-2 phylogenetic tree inferred by MAPLE from Viridian genomes. In the center is a zoom-in of part of the tree showing the early evolution of the AY lineage. Phylogenetic tips are colored according to the Pango[6] lineage assigned by Pangolin[33] v4.3 (with pangolin-data v1.21) to the corresponding genomes. We also show the names of some of these lineages. On the right is the corresponding part of the phylogenetic tree for cov2tree[26,30]. We use colored triangles to annotate inferred mutation events on the branches of the phylogenetic tree, with mutations at the same position shown in the same color. We represent these mutations with their genome position and nucleotides (for example 'A5584G' means 'nucleotide A mutating into G at position 5584'). Three mutations in upstream branches are reverted downstream leading to the clade containing lineages AY.13, AY.14 and AY.54. The purple arrow shows a possible regrafting of this clade, leading to a more parsimonious evolutionary history avoiding two of these reversions.

Alignment masking and tree inference

With the aim to create a global and as reliable as possible SARS-CoV-2 alignment and phylogeny, we aligned all available Viridian SARS-CoV-2 genomes, masked putative error-prone positions, and filtered out low-coverage and highly heterozygous samples that could be affected by contamination or mixed infection. See Methods for a detailed description of our data preparation. The final alignment contains 2,072,111 Viridian consensus genomes.

From this alignment, we inferred a phylogenetic tree using MAPLE under an UNREST substitution model with rate variation. Inference of the initial tree with one core took about 4 days, 17 h and 18.5 GB of maximum memory; topological improvement search parallelized over 14 cores took about 5 days, 13 h and 440 GB of maximum memory. All analyses were run on an Intel Xeon Gold 6252 Processor at 2.10 GHz. The alignment, metadata, inferred tree and inferred substitution rates have been uploaded on Zenodo[32].

Consistent with previous analyses[13], genome positions 10323, 11083, 21137 and 27384 are inferred to be those with the highest substitution rate (Fig. 3e). In particular, we infer >11,000 separate mutation events at position 11083 in our phylogenetic tree.

We compared our phylogenetic tree with the global, public cov2tree[26,30]. It is not simple to do this systematically due to differences in the genomes included, consensus calling method and alignment masking approach, so we focused on global patterns of major SARS-CoV-2 variants' evolution. While we see strong overlaps, we also notice some differences. For example, we infer one of the major Delta branches to be composed of two sister clades, the first one containing lineages AY.13, AY.14 and AY.54 among others, and the second, larger one, containing lineages AY.9 and AY.47 among many others (Fig. 4). Cov2tree instead infers the first clade to be a sub-clade of the second one. Investigating the cov2tree tree and mutational history, it appears that the latter represents a local tree optimum, and that removing and reattaching the first clade in a similar way to the MAPLE tree leads to a more parsimonious evolutionary history, with two fewer substitution events (Fig. 4). This is exactly the kind of tree improvement that we have designed MAPLE to perform with its SPR search[12], and this example shows the positive impact it has for inferring phylogenetic trees.

It is also useful to investigate the consistency of Pango[6] lineage assignments (made using Pangolin[33]) in our tree: inconsistent assignments, for example showing multiple independent origins of the same lineage, can reveal phylogenetically uncertain lineage histories where our models might help identify more likely scenarios. An example is given in Fig. 5, showing that MAPLE infers two independent origins for genomes in the lineage AY.43.3 which is defined by a mutation C to T at position 19955 within lineage AY.43. MAPLE infers that C19955T occurred independently two times within AY.43, and in total 380 times across the whole tree. Clustering all AY.43.3 genomes in one clade would require replacing one of the occurrences of C19955T with other rarer mutations (Fig. 5).

This approach, however, also highlights possible issues in our tree and alignment. For example, we infer multiple clades of Pango

https://doi.org/10.1038/s41592-025-02932-8

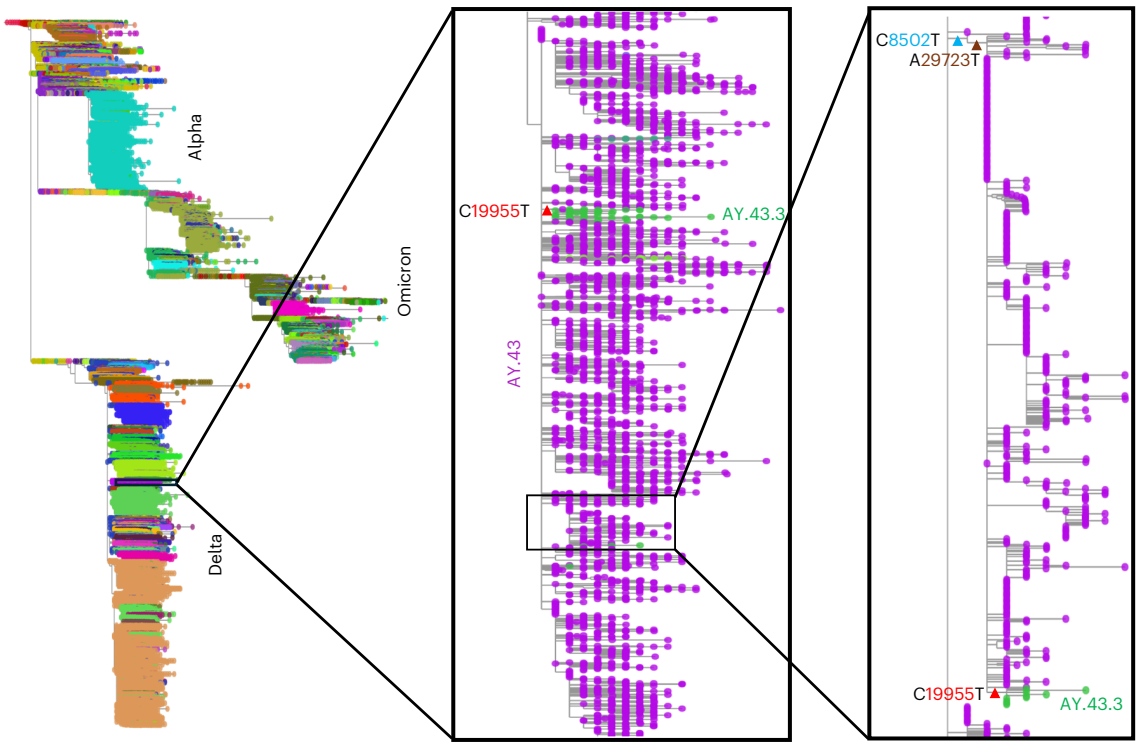


Fig. 5 | Origin of genomes assigned to lineage AY.43.3. Tree and notation here are as in Fig. 4. In the center, we show that the majority of genomes in lineage AY.43.3 are inferred to belong to a clade separated from the rest of the AY.43 genomes by mutation C19955T. However, in the right panel we show that MAPLE infers a smaller AY.43.3 clade to have originated from a separate C19955T mutation. MAPLE infers C19955T to be a higher frequency mutation (380 separate occurrences in the tree) than the other two mutations (C8502T, 83 occurrences, and A29723T, 33 occurrences) one would need to duplicate to cluster all the AY.43.3 genomes into the same clade.

sublineages BA.1.1 and BA.1.15 within BA.1 (Fig. 6). This seems to be due to some of these genomes having the consensus sequences of rare recombinants. Investigating read data, we see low coverage and heterozygosity at these putative recombination sites; altogether, these are the hallmarks of consensus sequence artifacts caused by contamination or mixed infections.

Discussion

We introduced a pandemic-scale phylogenetic framework to account for, and estimate, rate variation and recurrent sequence errors. This approach, implemented within the software MAPLE, can lead to improved phylogenetic inference at limited additional computational demand, and allowed us to efficiently investigate recurrent consensus sequence errors. We find that recurrent errors inferred by MAPLE are typically corroborated by sequencing read data. We also implemented algorithmic improvements to MAPLE and parallelized its most computationally demanding tasks. We use these methods to reconstruct a reliable, global, public alignment and phylogenetic tree from >2 million curated and filtered SARS-CoV-2 genomes. These advances not only are useful in studying SARS-CoV-2 evolution and spread, but also can be more generally applied within genomic epidemiology, and therefore enhance our preparedness for future pandemics.

There are of course some limitations to our methods. One is the risk of over-parameterization; to prevent this it is important to analyze sufficiently many sequences (our simulations suggest approximately >50,000 genomes: Fig. 2d; but analysis of convergence on real data suggests that substantially more data might be needed: Extended Data Fig. 3) when considering these models of rate variation and recurrent sequence errors. Furthermore, while MAPLE's approximations are essential in reducing computational demand and allowing pandemic-scale likelihood-based phylogenetics, they can also lead to biases at higher evolutionary distances[12]. In particular, very high sequence error probability at a given position can lead MAPLE to phylogenetically cluster sequences sharing the same error, interpreting these errors as the result of real mutation events, and therefore underestimating the number of sequence errors at the considered position. We observed this phenomenon particularly at positions 8835 and 15521, which have putative consensus sequence errors in hundreds of thousands of genomes. Similar issues can arise when other assumptions of our model are violated, for example, if errors at different positions are correlated, or if errors are phylogenetically clustered.

While we showed that these advances can lead to more realistic phylogenetic inference, we also have to consider other issues in our alignment and tree. The most prominent of these seems to be due to a small number of genomes whose consensus sequence is seemingly affected by contamination. Even a small number of such genomes can noticeably affect the inferred evolutionary history of major clades. While we tried to filter out samples potentially affected by contamination, our preprocessing filtered out the majority of available genomes (around 3 million of around 5 million), so using more strict filters might be counterproductive. Further improvements here are clearly warranted, and one option might be to consider model-based inference of contamination in read data. Alternatively, stricter consensus calling thresholds could be used, for example more strictly masking regions with relatively low depth or high heterozygosity. However, a probably more appealing avenue could be using a phylogenetic approach, not requiring the use of read data, to identify putative recombinant and contaminated genomes (consensus genomes affected by contamination would not be likely distinguishable from true recombinants with small numbers of descendants), as in for example ref. 34. Additionally, including time data into pandemic-scale phylogenetic inference

https://doi.org/10.1038/s41592-025-02932-8

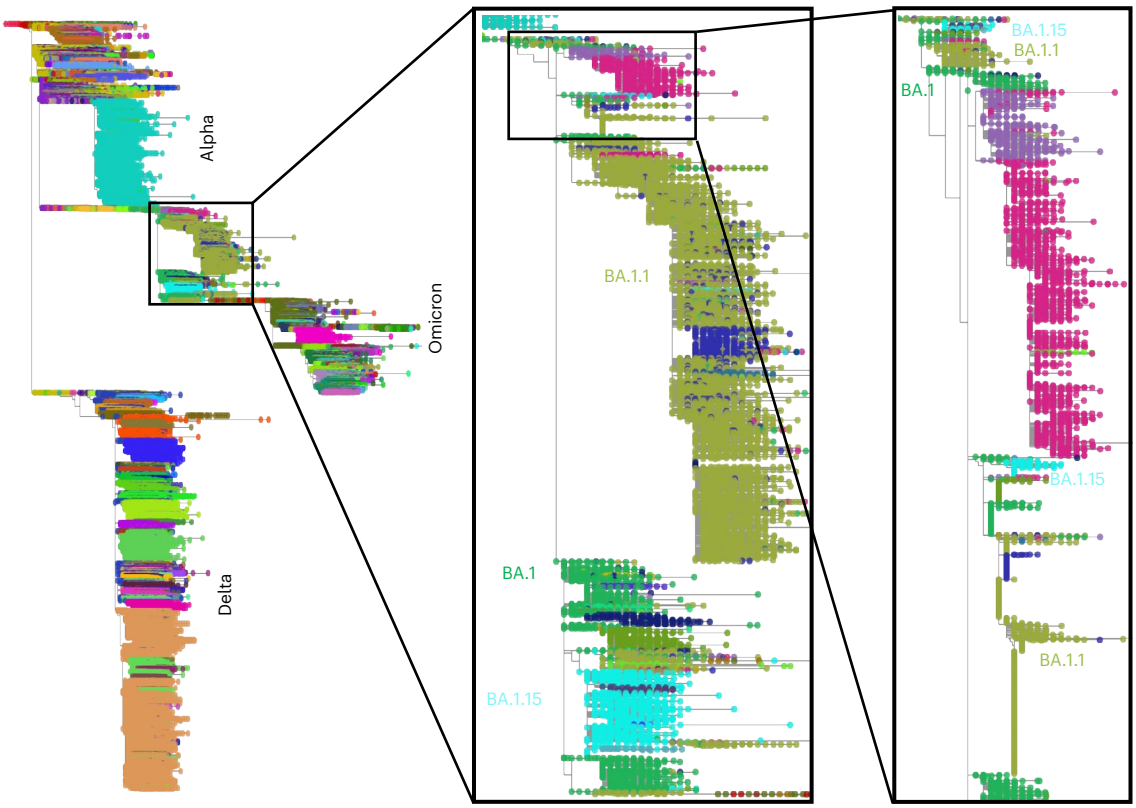


Fig. 6 | Issues with inferring the history of lineages BA.1.1 and BA.1.15. Tree and notation here are as in Fig. 4. As shown in the zoom-ins, lineages BA.1.1 and BA.1.15 appear to have emerged multiple times within parent lineage BA.1, in what we might call 'duplicate' clades. Upon investigation, this seems to be caused by possible artifacts in a few consensus sequences, such as SRR18978380 and ERR8227846, due to contamination. Most of the samples in these smaller clades have very low coverage and ambiguous consensus sequence at the genomic loci of the mutations distinguishing the duplicate clades, meaning that they could equivalently be placed in more than one of these clades.

(see for example ref. 35) could improve inference, for example, by helping us place genomes with substantial sequence uncertainty into more likely regions of the phylogenetic tree.

Here we focused on the inference of individual phylogenetic trees using a maximum likelihood approach. This approach ignores uncertainty in the inference of the tree, however, for example due to multiple plausible mutational histories, or multiple possible placements of incomplete genomes. While some of this uncertainty can be captured using multiple inference runs with different starting trees (Extended Data Fig. 4), to address these issues in a more efficient and interpretable manner, we are developing an approach to assess and represent phylogenetic uncertainty, particularly in genomic epidemiology[36]. An annotated tree using this representation of uncertainty for the data and maximum likelihood inference considered here is available from ref. 32.

Implementation of the presented algorithms and models in C++ is ongoing[37], and will substantially improve their efficiency. We plan to use OpenMP[38] to implement a shared-memory parallel tree search and substantially reduce memory usage. We additionally plan to automatically infer the optimal number of cores for parallelizing the tree search. The use of multiple local references is also the key to generalizing our approach to arbitrary phylogenetic scales, for example, switching between parsimony-like approximations and standard phylogenetic likelihood techniques across different branches of the tree, depending on their length.

In conclusion, we have presented advances that allow phylogenetic analysis of genomic epidemiological data at massive scale and high accuracy, which will be key in tracking transmission and pathogen evolution in a world where genome sequencing plays an ever-more relevant role in combating infectious disease.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-025-02932-8.

References

1. Gonzalez-Reiche, A. S. et al. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* **369**, 297–301 (2020).
2. Lu, J. et al. Genomic epidemiology of SARS-CoV-2 in Guangdong province, China. *Cell* **181**, 997–1003 (2020).
3. Hodcroft, E. B. et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712 (2021).
4. Vöhringer, H. S. et al. Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature* **600**, 506–511 (2021).
5. Volz, E. et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell* **184**, 64–75 (2021).
6. Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
7. Lemieux, J. E. et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**, eabe3261 (2021).

8. Morel, B. et al. Phylogenetic analysis of SARS-CoV-2 data is difficult. *Mol. Biol. Evol.* **38**, 1777–1791 (2021).
9. Hodcroft, E. B. et al. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* **591**, 30–33 (2021).
10. Turakhia, Y. et al. Ultrafast Sample placement on existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genetics* **53**, 809–816 (2021).
11. Ye, C. et al. matOptimize: a parallel tree optimization method enables online phylogenetics for SARS-CoV-2. *Bioinformatics* **38**, 3734–3740 (2022).
12. De Maio, N. et al. Maximum likelihood pandemic-scale phylogenetics. *Nat. Genet.* **55**, 746–752 (2023).
13. De Maio, N. et al. Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biol. Evol.* **13**, evab087 (2021).
14. De Maio, N. et al. Issues with SARS-CoV-2 sequencing data. *Discourse* https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/ (2020).
15. Turakhia, Y. et al. Stability of SARS-CoV-2 phylogenies. *PLoS Genet.* **16**, e1009175 (2020).
16. Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* **21**, 428–444 (2020).
17. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).
18. Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**, 367–372 (1996).
19. Stamatakis, A. Phylogenetic models of rate heterogeneity: a high performance computing perspective. In *Proc. 20th IEEE/ACM International Parallel and Distributed Processing Symposium* (IEEE Computer Society, 2006).
20. Massingham, T. & Goldman, N. Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**, 1753–1762 (2005).
21. Felsenstein, J. *Inferring Phylogenies* Vol. 2 (Sinauer Associates, 2004).
22. Kuhner, M. K. & McGill, J. Correcting for sequencing error in maximum likelihood phylogeny inference. *G3* **4**, 2545–2552 (2014).
23. De Maio, N. et al. Masking strategies for SARS-CoV-2 alignments. *Discourse* https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480/ (2020).
24. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**, 57–86 (1986).
25. Yang, Z. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105–111 (1994).
26. McBroome, J. et al. A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. *Mol. Biol. Evol.* **38**, 5819–5824 (2021).
27. Harrison, P. W. et al. The COVID-19 data portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic Acids Res.* **49**, W619–W623 (2021).
28. Hunt, M. et al. Addressing pandemic-wide systematic errors in the SARS-CoV-2 phylogeny. *Nat. Methods* https://doi.org/10.1038/s41592-025-02947-1 (2025).
29. Bull, R. A. et al. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat. Commun.* **11**, 6272 (2020).
30. Sanderson, T. Taxonium, a web-based tool for exploring large phylogenetic trees. *eLife* **11**, e82392 (2022).
31. Wu, F. et al. A new coronavirus associated with human respiratory disease in china. *Nature* **579**, 265–269 (2020).
32. De Maio, N. Global (2M) SARS-CoV-2 genomes dataset, from Viridian, processed with MAPLE. *Zenodo* https://doi.org/10.5281/zenodo.12733488 (2024).
33. O'Toole, Á. et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **7**, veab064 (2021).
34. Turakhia, Y. et al. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature* **609**, 994–997 (2022).
35. Sanderson, T. Chronumental: time tree estimation from very large phylogenies. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.27.465994 (2021).
36. De Maio, N., Ly-Trong, N., Martin, S., Minh, B. Q. & Goldman, N. Assessing phylogenetic confidence at pandemic scales. *Nature* **647**, 472–478 (2025).
37. Ly-Trong, N., Bielow, C., De Maio, N. & Minh, B. Q. CMAPLE: efficient phylogenetic inference in the pandemic era. *Mol. Biol. Evol.* **41**, msae134 (2024).
38. Dagum, L. & Menon, R. OpenMP: an industry standard API for shared-memory programming. *IEEE Comput. Sci. Eng* **5**, 46–55 (1998).

Methods

Genome lists

Here we provide a summary of how genome sequences and ancestral likelihoods are stored in MAPLE, which will be relevant to describe many of the method developments presented here. For a full description of the MAPLE approach see ref. 12.

We represent genome sequences and ancestral sequence likelihood in terms of differences with respect to a reference—either local or global. Instead of likelihood vectors[21], we use genome lists. An entry of a genome list represents the state of a portion of the genome, either a single site, or a collection of contiguous sites. An entry of a genome list is defined as a tuple $(\tau, s, l, \mathbf{v})$ containing:

- An entry type τ; the allowed types are 'R', to represent a stretch of contiguous sites identical to the reference; type 'N' to represent a stretch of contiguous sites containing no descendant sequence information, that is with the four nucleotides having the same partial likelihoods; entries of type R and N are essential for concisely representing observed and ancestral genomes. Types 'A', 'C', 'G' and 'T' represent a single site where the represented nucleotide is the state of the considered (possibly ancestral) genome with negligeable uncertainty. Type 'O' is used for a single site where multiple nucleotides have non-negligible partial likelihoods, that is, type O entries are used to represent uncertainty at the few genome positions and tree nodes where multiple states are plausible.
- An entry position s representing the site(s) of the genome the entry refers to. If the entry represents a stretch of sites, this element is the last one (from 5′ to 3′) of these sites. In this case, the first site of the stretch can be inferred by the s element of the previous entry in the genome list, since entries in the genome list are sorted according to s and since they form a partition of the considered genome.
- The branch length l representing the evolutionary distance between the considered node (which stores the genome list) and the location in the tree that the partial likelihoods contained or represented by the genome list entry refer to. For example, an entry of type R might represent a stretch of nucleotides observed at the child c of the considered node n, in case the other child node of n has type N at these same positions. In this example l represents the branch length separating n and c.
- A vector of relative partial likelihoods $\mathbf{v}$, only present in genome list entries of type O. The difference between this vector and a standard likelihood vector in phylogenetics is that $\mathbf{v}$ is normalized (its entries sum up to 1). This is because, while performing the tree search, MAPLE only utilizes and stores relative likelihoods, not absolute ones.

Modeling rate variation

Variation in evolutionary rate along the genome can heavily impact phylogenetic inference[16]. As such, models of evolutionary rate variation[18] have been an essential part of phylogenetic analyses in the last decades. Here we first review the most popular phylogenetic models of rate variation and then present the details of our approach.

Gamma random variable model of rate variation. In likelihood-based phylogenetics, variation in evolutionary rates is typically modeled using gamma-distributed random variable models, where each genome position is allowed to evolve under any of a certain number (typically four) of rates. Under this model, the likelihood is integrated over every possible rate assignment for every site, and a discretized gamma distribution is used to model the distribution of rates across categories[17].

In more detail, for any site s of the alignment, for any rate category c and for any node n of the tree, one calculates a set of likelihoods $l_i^{n,s,c}$, one for each nucleotide i; to do this, the substitution rate from nucleotide i to nucleotide j for category c is defined as equation (1):

$$q_{i,j}^{c} = r_c q_{i,j} \tag{1}$$

where r_c is the rate of category c, and $q_{i,j}$ is the background substitution rate from i to j, which is shared between categories. Likelihoods $l_i^{n,s,c}$ are initialized as usual for each terminal node, site and category. Then using the Felsenstein pruning algorithm[39] and traversing the child nodes before the parent nodes, one calculates likelihoods at internal nodes, where the likelihoods of class c, at each node and site, are calculated using the substitution rates $q_{i,j}^{c}$ of category c and the likelihoods of category c from the children nodes. This means that normally the time and memory demands of this approach are C times higher than the pruning algorithm without rate variation, with C the number of categories considered.

A major problem when trying to use this model with pandemic-scale datasets and algorithms is that, under this model, mutation events in one area of the tree can impact the relative likelihoods of different rate categories in other distant parts of the tree. This is not compatible with the algorithms and data structures in MAPLE, which are based on assumptions of 'locality': changes in the tree typically only affect relative likelihoods in a small phylogenetic neighborhood of the affected nodes. For these reasons, here we used a different model of rate variation, detailed in the next section.

Site-specific model of rate variation in MAPLE. Instead of the traditional gamma model of rate variation, we propose a model similar to CAT[19]. In CAT, alignment sites are grouped into a number of classes, and each class of sites has one (and only one) rate assigned to it; this speeds up likelihood calculations since for each site only one set of likelihoods needs to be computed. In CAT, sites are grouped into a small number of classes to reduce the number of rate parameters to avoid over-parameterization[19]. In the case of pandemic-scale phylogenetics, since we have millions of observations for each site, we argue that grouping sites into classes is more likely to lead to under-parameterization, and we assign to each site s its own specific substitution rate r_s. Site rates r_s are estimated in MAPLE using EM (see below).

The most important advantage of this site-specific model compared to the gamma model is that it can be implemented in MAPLE almost without affecting its computational demand. To calculate the likelihood cost of no mutation happening on a stretch of the genome identical to the reference (see ref. 12), we precompute the sums of site-specific substitution rates of the reference nucleotides for prefix segments of the genome (portions of the genomes made of all nucleotides from the first one up to a given position).

Another advantage of our approach is that it allows us to estimate biologically informative site-specific rates (Fig. 3). Rates inferred from our Viridian alignment of 2 million genomes have been uploaded to Zenodo[32].

Sequence error model

Recurrent sequence errors adversely impact phylogenetic inference, especially with low divergence datasets such as collections of SARS-CoV-2 genomes[14,15]. We previously developed methods to identify and mask recurrent errors[14,15,23] to help prevent biases in phylogenetic and other downstream analyses[23]. These approaches have, however, proved untenable in the long term due to excessive computational demand and the elevated number of genome positions affected by recurrent errors.

Here, we use a different approach. Firstly, we focus on genomes with publicly shared raw read data, which allows us to consistently run a tailored, state-of-the-art pipeline for calling consensus genomes; this prevents reference-biased calls and, therefore, reduces the number of recurrent errors in our alignment. The availability of read data also allows us to corroborate putative consensus sequence errors. Secondly, we perform phylogenetic and sequence error inference

jointly, using a recurrent sequence error model within MAPLE, which is detailed in this section. Compared to previous methods, this approach has several advantages:

- It allows the estimation of site-specific error probabilities across the genome and the inference of individual sequence errors.
- It can be paired with a rate variation model (see above) to disentangle the effects of substitution rate variation and sequence error probability variation.
- Model parameter and sequence error estimation can be performed at the same time as phylogenetic inference.
- It can be used on large datasets of millions of genomes, as it adds limited burden to MAPLE's computational demand.

While traditional phylogenetic models of sequence errors often assume a homogeneous error probability across the genome[21,22], we use site-specific error probabilities (see also ref. 40). Like previous phylogenetic error models, our approach modifies the likelihoods of terminal nodes. For example, without errors, the partial likelihood vector corresponding to an observed nucleotide 'A' is defined as (1, 0, 0, 0), where we use the alphabetical order of nucleotides. When accounting for sequence errors with a homogeneous error probability ε, the same likelihood vector becomes $(1-\varepsilon, \varepsilon/3, \varepsilon/3, \varepsilon/3)$[21,22]. With site-specific error probabilities, each site s of the genome has its own error probability ε_s, so that, for site s, the partial likelihood vector becomes $(1-\varepsilon_s, \varepsilon_s/3, \varepsilon_s/3, \varepsilon_s/3)$.

Often, modeling site-specific error probabilities with one free parameter per site would result in over-parameterization. However, just as for the substitution rates (see above), the large numbers of sequences we consider compensate for the large number of parameters and allow us to avoid under-parameterization, because error probabilities in SARS-CoV-2 are likely highly site specific[14,15]. We estimate error probabilities at the same time as rate variation using EM (see below).

To efficiently implement the site-specific error model, we modify some of the algorithms and data structures in MAPLE. First, we add a flag to each entry of genome lists (which are equivalent to likelihood vectors); a positive flag means that the data/likelihoods represented by the entry refer to a terminal node and, therefore, need to be corrected by the probability of a sequence error. Otherwise, a negative flag means that no correction for possible sequence errors is required. This flag does not simply distinguish terminal node entries from internal node entries; this is because the observation of a nucleotide in MAPLE can be carried over from a terminal node likelihood to its parent likelihood, in case the sibling node contains no information at the considered site. In this case, the parent node genome list entry will also have a positive flag. We use flags instead of calculating likelihoods explicitly under the error model because we want to avoid calculation of likelihoods for long stretches of the genome.

When creating new genome lists (combining two genome lists into one), or when comparing two genome lists to calculate likelihood costs (for example, the cost of an SPR move or the addition of a sample to the initial tree), we need to account for possible sequence errors only when at least one of the two flags of the two genome list entries considered is positive. To do this, we proceed as typical in phylogenetic sequence error models[21,22], except that, using the assumption of short branches and low error probabilities, we use a first-order approximation and ignore terms that are quadratic (or higher order) in either the branch length or the error probability. For example, given a branch length t, we ignore terms that are in the order of magnitude of t^2, ε_s^2 or $t \times \varepsilon_s$. A substantial benefit of this approximation is that we can compute likelihood costs and relative likelihoods for long stretches of the genome in constant time. For example, when two considered genome list entries are of type R (they represent a stretch of the genome identical to the reference), then their parent node will also be of type R at the same positions, since any alternative would require two mutations, two errors or one error and one mutation at the same position; all these alternative scenarios have negligible probability compared to the history with no mutations and no errors. Conversely, when comparing two genome list entries of type R to calculate a likelihood score (for example, for calculating the likelihood cost of a placement), we can integrate the cost of no mutations and no errors happening on the considered stretch of the genome in constant time by using precomputed total sums of error probabilities for genome prefixes, in a very similar way to accounting for rate variation in MAPLE (see above and ref. 12).

Estimation of model parameters with EM

Early versions of MAPLE used a simple, heuristic approach to substitution rate parameter estimation. We now instead use EM to jointly estimate substitution rates, including site-specific ones, and error probabilities, if required. EM parameter optimization is performed after the estimation of the initial tree, and after each round of SPR tree topology improvement. After initial tree estimation, parameter optimization is started from an empirical global substitution rate matrix, uniform site-specific rates and a uniform site-specific error probability equal to the reverse of the genome length (equivalent to one expected error per genome).

Our EM approach is inspired by Klosterman et al.[41] (whose notation we follow here), with the difference that we exploit the shortness of the considered tree branches to reduce computational demand. We first detail our EM approach for the basic scenario that sequence errors and rate variation are not modeled, then describe the case with sequence errors, and finally we describe the modifications needed to model site-specific rates and error probabilities.

Given a branch of length T, and given two nucleotide assignments, a for the top of the branch and b (possibly with $b = a$) for the bottom, we define the substitution probability of such an assignment as $M_{ab}(T)$. Assuming $T \ll 1$, we approximate $M_{ab}(T) \approx Tq_{ab}$ if $a \neq b$ and $M_{ab}(T) \approx 1 + Tq_{aa}$ otherwise. The expected number of substitutions from nucleotide i to $j \neq i$ on the considered branch is defined as equation (2):

$$c_{ij} = \frac{1}{M_{ab}(T)} \int_0^T M_{ai}(t) q_{ij} M_{jb}(T-t) dt \tag{2}$$

while the expected waiting time in nucleotide i on the same branch is defined as equation (3):

$$w_i = \frac{1}{M_{ab}(T)} \int_0^T M_{ai}(t) M_{ib}(T-t) dt \tag{3}$$

(see ref. 41 for more details). Again, following the assumption of short branches, we use first-order approximations:

- In the case $a = b$, we have $w_a \approx T$ and $w_i \approx 0$ for $i \neq a$. We also have $c_{ij} \approx 0$ for any $i \neq j$.
- In the case $a \neq b$ we have $w_i \approx T/2$ for $i = a$ or $i = b$, and $w_i \approx 0$ otherwise. We also have $c_{ij} \approx 1$ if $i = a$ and $j = b$, and $c_{ij} \approx 0$ otherwise.

Parameters c_{ij} and w_i can be calculated efficiently across genome lists, similarly to how likelihood costs are calculated in MAPLE. Then, the EM estimates of the substitution rates are calculated by traversing the tree and adding up, for each branch encountered, for each genome position, and for each pair of nucleotide assignments a, b for the two ends of the considered branch, the value $p_{ab}c_{ij}$ to a total count (that is, over all branches) C_{ij} for each nucleotide pair i,j, and the value $p_{ab}w_i$ to a total count W_i for each nucleotide i, where p_{ab} is the posterior probability of the assignment a, b for the considered branch. These updates are done in constant time for each genome list entry, no matter the number of positions represented by the entry. In the case that the site-specific model of rate variation (see above) is used, we estimate one rate for each genome position by keeping track of site-specific substitution counts and waiting times. Once we complete the traversal of the tree, the new estimate of the substitution rate q_{ij} will be $q_{ij} = C_{ij}/W_i$.

EM with a sequence error model. When we model recurrent sequence errors (see above), the estimation of model parameters with EM becomes more complex. In addition to branch-specific expectations c_{ij} and w_i, and global counts C_{ij} and W_i, we also track branch-specific expected errors e_{ij} and global error counts E_{ij}. Even though we do not consider nucleotide-specific error probabilities, we keep the i,j subscript for consistency and for possible future extensions. Here we assume, for simplicity, constant substitution rates and error probability along the genome. The further modifications needed to model site-wise variation are discussed in the following section.

EM counts are calculated as:

- In the case $a = b$ we have $w_i \approx T$ for $i = a$ and $w_i \approx 0$ otherwise. We also have $c_{ij} \approx 0$ and $e_{ij} \approx 0$ for any $i \neq j$.
- In the case $a \neq b$ and b does not derive from a terminal node observation (that is, it could not be the result of a sequence error), we have $w_i \approx T/2$ for $i = a$ and $i = b$, and $w_i \approx 0$ otherwise. We also have $c_{ij} \approx 1$ if $i = a$ and $j = b$, and $c_{ij} \approx 0$ otherwise. Parameter $e_{ij} \approx 0$ for any $i \neq j$.
- In the case $a \neq b$ and b could be the result of a sequence error, we have that $\frac{\varepsilon/3}{q_{ab}T+\varepsilon/3}$ is approximately the probability that the substitution was caused by an error, while $\frac{q_{ab}T}{q_{ab}T+\varepsilon/3}$ is the approximate probability that a mutation caused it. From this, we have that $w_a \approx T\frac{q_{ab}T/2+\varepsilon/3}{q_{ab}T+\varepsilon/3}$ and $w_b \approx T\frac{q_{ab}T/2}{q_{ab}T+\varepsilon/3}$, while $w_i \approx 0$ if $i \neq a$ and $i \neq b$. We also have $c_{ij} \approx \frac{q_{ab}T}{q_{ab}T+\varepsilon/3}$ if $i = a$ and $j = b$, and $c_{ij} \approx 0$ otherwise. Finally, $e_{ij} \approx \frac{\varepsilon/3}{q_{ab}T+\varepsilon/3}$ for $i = a$ and $j = b$, and $e_{ij} \approx 0$ otherwise.

We traverse the tree and add up, for each branch encountered and each pair of nucleotide assignments a, b for the two ends of the considered branch, the value $p_{ab}c_{ij}$ to the total count C_{ij} for each i,j; the value $p_{ab}w_i$ to the total count W_i for each i; and the value $p_{ab}e_{ij}$ to the total count E_{ij} for each i,j. Again, these updates are done in constant time for each genome list entry (possibly covering thousands of genome positions) considered. Once we complete the traversal of the tree, the new estimate of the substitution rate q_{ij} is $q_{ij} = C_{ij}/W_i$. The new estimate of the error probability is $\varepsilon = \sum_{i,j}E_{ij}/K$, where K is the number of times informative characters (non-'N' characters) are observed at terminal nodes (we do not count w_i on branches/positions above type 'N' characters).

EM with site variation. When modeling site-specific rates and error probabilities, we need to keep separate counts for each genome position, that is, instead of W_i we define total waiting time for character i and site s as W_i^s; we add to this count only the waiting times w_i calculated at position s of the alignment. We similarly define and update position-specific counts $C_{i,j}^s$, $E_{i,j}^s$ and K^s.

Convergence of EM parameter inference. While there are some theoretical guarantees regarding the convergence of EM[42], in some practical applications convergence might be too slow (see for example ref. 43). Figure 2d shows that error rate parameters inferred from simulated data are quite close to the correct ones, in particular with datasets containing at least 50,000 SARS-CoV-2 genomes. Simulated data are, however, less complex than real data. When we compare rate parameters inferred from subsets of the real data to those inferred from the full dataset, we see that considerably more genomes are needed for convergence of the error rate parameters, while mutation rate parameter inference converges much more quickly (Extended Data Fig. 3).

An interesting question is, given an alignment column and a tree, how does the EM algorithm infer both a mutation rate and an error rate? (Extended Data Fig. 5). We give here a simple example with an analytical description of this convergence, as well as some practical results.

For simplicity, we consider the scenario where mutations between any pair of nucleotides have the same rate (a Jukes–Cantor model[44]), and where all branches of the tree have the same length $0 < b \ll 1$ (that is, short divergence, as in our SARS-CoV-2 scenario). Such a binary rooted tree relating G genomes will necessarily have G terminal branches and $G-1$ internal branches. With the assumption of short divergence, we can assume that the number and type of substitutions inferred on a given tree from a given alignment do not depend on the assumed mutation rate r or error rate ε at the considered site. So, we can assume that at the considered site we have $S_I \ll G$ substitutions on internal branches and $S_T \ll G$ substitutions on terminal branches. The EM procedure will, therefore, only depend on parameters G, S_I and S_T, as well as on the mutation rate $r_0 > 0$ and error rate $\varepsilon_0 > 0$ used to initialize the EM inference.

In these circumstances, the number of expected errors will be approximately equal to equation (4):

$$S_T\frac{\varepsilon_0}{\varepsilon_0 + r_0 b} \tag{4}$$

since, if we assume an error rate ε_0 and a mutation rate r_0, a terminal substitution on a branch of length b has an approximate probability of $\varepsilon_0/(\varepsilon_0 + r_0 b)$ of being an error. From this, we get the next EM estimate of the error rate as defined in equation (5):

$$\varepsilon_1 = \frac{S_T}{G}\frac{\varepsilon_0}{\varepsilon_0 + r_0 b} \tag{5}$$

by dividing by the number of sequences in the alignment. Similarly, the expected number of mutations will be as defined in equation (6):

$$S_I + S_T\frac{r_0 b}{\varepsilon_0 + r_0 b}, \tag{6}$$

leading to the next mutation rate inference as shown in equation (7):

$$r_1 = \frac{S_I + S_T\frac{r_0 b}{\varepsilon_0 + r_0 b}}{b(2G-1)}. \tag{7}$$

Note that from this and from equation (4) it follows that, as shown in equation (8):

$$r_1 b = \frac{S_I + S_T\left(1 - \frac{\varepsilon_0}{\varepsilon_0 + r_0 b}\right)}{2G-1} = \frac{S_I + S_T - G\varepsilon_1}{2G-1} \tag{8}$$

and similarly for $i > 1$, we have equation (9):

$$\varepsilon_i = \frac{S_T}{G}\frac{\varepsilon_{i-1}}{\varepsilon_{i-1} + r_{i-1} b} \tag{9}$$

and equation (10):

$$r_i b = \frac{S_I + S_T - G\varepsilon_i}{2G-1} \tag{10}$$

Let us consider first the case $\frac{S_T}{G} < \frac{S_I}{G-1}$. In this case the proportion of substitutions on internal branches is higher than on terminal branches, meaning that the mutation rate alone should be enough to explain all substitutions; therefore, we would expect $\lim_{i\to\infty}\varepsilon_i = 0$. Indeed, we have that in this case, for any $i > 1$, as shown in equation (11):

$$\begin{aligned}\varepsilon_i &= \frac{S_T}{G}\frac{\varepsilon_{i-1}}{\varepsilon_{i-1} + r_{i-1} b} = \varepsilon_{i-1}\frac{S_T}{G\varepsilon_{i-1} + G\frac{S_I + S_T - G\varepsilon_{i-1}}{2G-1}} \\ &= \varepsilon_{i-1}\frac{(2G-1)S_T}{G(2G-1)\varepsilon_{i-1} + G(S_I + S_T - G\varepsilon_{i-1})} \\ &= \varepsilon_{i-1}\frac{GS_T + (G-1)S_T}{G(G-1)\varepsilon_{i-1} + GS_I + GS_T}.\end{aligned} \tag{11}$$

Since we assumed that $\frac{S_T}{G} < \frac{S_I}{G-1}$ and, therefore, $(G-1)S_T < GS_I$, and since $G(G-1)\varepsilon_{i-1}$ is positive, we have that the denominator $G(G-1)\varepsilon_{i-1} + GS_I + GS_T$ is larger than the numerator $GS_T + (G-1)S_T$, and so $\varepsilon_i < \varepsilon_{i-1}$, and more specifically that $\varepsilon_i < C\varepsilon_{i-1}$ with a constant $C < 1$, meaning that ε_i converges exponentially to 0.

Let us now consider the second scenario in which $\frac{S_T}{G} > \frac{S_I}{G-1}$, or equivalently $S_T(G-1) > S_I G$. In this case the mutations on internal branches are not enough to justify the number of substitutions observed on terminal branches, so we would expect a positive error rate to be inferred, in particular equal to $\frac{S_T}{G} - \frac{S_I}{G-1}$. Let us define as $D_i = \varepsilon_i - (\frac{S_T}{G} - \frac{S_I}{G-1})$ the difference between the current estimate of ε and its expected limit. We then have that for any $i > 1$, as shown in equation (12):

$$\begin{aligned}
\varepsilon_i &= \varepsilon_{i-1}\frac{(2G-1)}{G}\frac{S_T}{(G-1)\varepsilon_{i-1} + S_I + S_T} \\
&= \varepsilon_{i-1}\frac{(2G-1)}{G}\frac{S_T}{(G-1)\left(\frac{S_T}{G} - \frac{S_I}{G-1} + D_{i-1}\right) + S_I + S_T} \\
&= \varepsilon_{i-1}\frac{(2G-1)S_T}{(2G-1)S_T + G(G-1)D_{i-1}} \\
&= \varepsilon_{i-1}\left(1 - \frac{G(G-1)D_{i-1}}{(2G-1)S_T + G(G-1)D_{i-1}}\right) \\
&= \left(\frac{S_T}{G} - \frac{S_I}{G-1} + D_{i-1}\right)\left(1 - \frac{G(G-1)D_{i-1}}{(2G-1)S_T + G(G-1)D_{i-1}}\right) \\
&= \frac{S_T}{G} - \frac{S_I}{G-1} + D_{i-1}\left(1 - \frac{G(G-1)\left(\frac{S_T}{G} - \frac{S_I}{G-1} + D_{i-1}\right)}{(2G-1)S_T + G(G-1)D_{i-1}}\right) \\
&= \frac{S_T}{G} - \frac{S_I}{G-1} + D_{i-1}\left(1 - \frac{(G-1)S_T - GS_I + G(G-1)D_{i-1}}{(2G-1)S_T + G(G-1)D_{i-1}}\right) \\
&= \frac{S_T}{G} - \frac{S_I}{G-1} + D_{i-1}\frac{GS_T + GS_I}{(2G-1)S_T + G(G-1)D_{i-1}}.
\end{aligned} \tag{12}$$

Since this time we assumed that $S_T(G-1) > GS_I$, we have that the coefficient $\frac{GS_T+GS_I}{(2G-1)S_T+G(G-1)D_{i-1}}$ is <1 and, therefore, that $D_i < D_{i-1}C$ with a constant $C < 1$, meaning that D_i converges exponentially to 0 and, therefore, ε_i converges exponentially to $\frac{S_T}{G} - \frac{S_I}{G-1}$, concluding our argument.

The scenario we just considered is very simplified, as in our actual EM procedure we use an UNREST model instead of a Jukes–Cantor one, we account for uncertainty in the reconstruction of the substitution history, and, most importantly, we do not assume that all branches have the same length. This last assumption is particularly important. While during EM we do account for different branch lengths, we however assume a constant tree, and we do not optimize branch lengths at the same time as model parameters but rather alternate between optimization of the two. This means that we are not guaranteed convergence to an overall optimal pair of tree and model parameters, because optimal branch lengths depend on model parameters and vice versa. For example, typically the higher the probability that a substitution on a terminal branch is the result of a sequencing error, the shorter the maximum likelihood length of that branch will be. Conversely, the shorter a terminal branch is, the higher the probability that a substitution on that branch is the result of a sequencing error rather than a mutation. Analysis of real data confirms that convergence to a global optimum for tree and model parameters jointly is not guaranteed, although overall the impact of initial parameters is limited. For example, starting from initial values of the site-wise error rates spanning five orders of magnitude, we find that the mean of the estimated site-wise error rates is always between around 2×10^{-7} and 3×10^{-7} (Extended Data Table 3).

Other added features

Sample placement on partial initial trees for online phylogenetic inference. We have implemented in MAPLE the option of using an initial tree, either complete (relating all the sequences in the input alignment) or incomplete. If the initial tree is incomplete, MAPLE will add the missing samples to the tree using maximum likelihood stepwise addition, and, since stepwise addition is typically not sufficient on its own to achieve accurate tree inference[45], it will then perform SPR searches to improve the tree topology. We have also implemented the option to perform a faster, partial SPR search targeting only nodes whose genome lists have been affected during stepwise addition. If only a very few samples have been added to the initial tree, this partial SPR search can be very fast.

These features not only allow users to specify custom initial trees, but also allow online phylogenetic inference[46] (where new samples are added to a tree as they become available, avoiding costly repeated reestimation of a large phylogeny from scratch) and placement of individual samples onto a target tree[10].

Improvements to the placement and SPR search algorithms. In earlier versions of MAPLE, placement and SPR searches were divided into two stages. In the first stage, we traversed the tree to find a preliminary optimal placement. Then, in the second stage, we optimized the exact location of the placement and the new branch lengths near the initial optimal placement[12].

We have now implemented more thorough searches, similarly divided in two stages, but in which the first stage returns a set of preliminary near-optimal placements (instead of just one). The second stage then optimizes branch lengths for all these candidates, and selects the best placement after this optimization. This is similar to the 'baseball' heuristic of the metagenomic query mapper pplacer[47] and other phylogenetic placement tools[48], and improves tree inference accuracy at very limited additional computational demand.

Tree rooting. We have implemented an efficient dynamic programming search of the maximum likelihood tree root location. This algorithm traverses the tree from parents (starting from the current tree root) to children nodes. As we traverse nodes, we keep track of how phylogenetic likelihoods are affected by moving the tree root from a parent node to its child, with the consequent inversion of mutation events on the branch separating the two. With this approach, we can not only search for the maximum likelihood rooting of the tree, but also assess the probability of this rooting against others[36].

Branch length optimization. We have implemented in MAPLE a more efficient and accurate procedure for branch length optimization, based on approximating the derivative of the phylogenetic relative log-likelihood as a function of branch length.

Given a branch of the phylogenetic tree, assume that G_2 is the genome list of the bottom node of the branch (representing the relative likelihoods of the subtree below the considered branch), and G_1 is the genome list for the top node of the branch (representing the relative likelihoods of the subtree complementary to the subtree below the considered branch). Our aim is to efficiently calculate the branch length value t that minimizes the likelihood cost of merging G_1 and G_2. This likelihood cost is the product of the likelihood cost of each individual intersection entry of G_1 and G_2 (ref. 12), so this is the same as minimizing the sum of the log-likelihoods of each intersection entry. To do this, we look for the values of t for which the derivative of the likelihood cost is 0.

Assuming that E_1 is any considered entry of G_1, and E_2 is the corresponding considered entry of G_2, we can approximate the derivative with respect to t of the log-likelihood costs as:

- If E_1 and E_2 represent the same nucleotide, or the same stretch of nucleotides, the log-likelihood cost is $\approx \sum_i - r_i(t + t_0)$ where r_i is the substitution rate of the considered nucleotide at position i, and t_0 is the sum of the branch length elements of E_1 and E_2 (these are used in the presence of 'N' characters in the alignment). The derivative is then $-\sum_i r_i$.

- If E_1 and E_2 represent different nucleotides i and j, the log-likelihood cost is $\approx \log q_{ij} + \log(t + t_0)$, where q_{ij} is the substitution rate from i to j. Its derivative is $\approx 1/(t + t_0)$.
- If either E_1 or E_2 is of type N (that is, it provides no sequence information), the log-likelihood cost is zero, and so is its derivative.
- If either E_1 or E_2 is of type O (meaning that multiple nucleotides are possible at this site and node), then we have to consider the different relative likelihoods of different nucleotides. Here we consider the most complex case of both E_1 and E_2 being of type O, the other scenarios being simplifications of the following. We call p_1 the relative likelihood vector of E_1 and p_2 the relative likelihood vector of E_2. The log-likelihood cost is $\approx \log(\sum_i p_1(i)p_2(i) + (t + t_0)\sum_{i,j} q_{ij}p_1(i)p_2(j))$. We do not know a priori if $p_i(i)$ and $p_2(j)$ are large or small compared to t and t_0. However, in the case that the coefficient of the t, $\sum_{i,j} q_{ij}p_1(i)p_2(j)$, is negative, then we know that $\sum_i p_1(i)p_2(i)$ is the dominant term inside the logarithm. In fact, $(t + t_0)\sum_{i,j} q_{ij}p_1(i)p_2(j)$ is equal to $(t + t_0)\sum_{i\neq j} q_{ij}p_1(i)p_2(j) + (t + t_0)\sum_i q_{ii}p_1(i)p_2(i)$ of which the first term is positive and the second term is negative (and, therefore, larger in absolute value as we assumed that the sum of the two terms is negative). Because we assume that none of the substitution rates q_{ii} is very large (that is, it is not of the order of magnitude of the inverse of a typical branch length, which would break the assumptions of MAPLE), then $\sum_i q_{ii}p_1(i)p_2(i)$ and, therefore, $\sum_{i,j} q_{ij}p_1(i)p_2(j)$ is of the same or smaller order of magnitude than $\sum_i p_1(i)p_2(i)$, so we obtain $|(t + t_0)\sum_{i,j} q_{ij}p_1(i)p_2(j)| \ll \sum_i p_1(i)p_2(i)$. By expressing the log-likelihood cost as $\approx \log(1 + (t + t_0)\frac{\sum_{i,j} q_{ij}p_1(i)p_2(j)}{\sum_i p_1(i)p_2(i)})$, we can then approximate it as $(t + t_0)\frac{\sum_{i,j} q_{ij}p_1(i)p_2(j)}{\sum_i p_1(i)p_2(i)}$, which has derivative with respect to t of $\frac{\sum_{i,j} q_{ij}p_1(i)p_2(j)}{\sum_i p_1(i)p_2(i)}$. We cannot necessarily use this approximation when $\sum_{i,j} q_{ij}p_1(i)p_2(j) > 0$; instead, in this case we can express the log-likelihood cost as $c + \log(t + a)$ for some positive values a and c, whose derivative is $1/(t + a)$.

In all cases, the derivative of the approximate log-likelihood cost of an entry is, therefore, in one of the two following forms:

- $-r$ for some positive cumulative substitution rate r, or
- $\frac{1}{t+a}$ for some positive value a.

Summing across all entry intersections, the total approximate log-likelihood derivative becomes $-R + \sum_{k=1}^{K} \frac{1}{t+a_k}$, for some positive R, K and a_k values. This expression is well defined for $t > 0$ and is strictly decreasing as t increases, with limit $-R$ at $t \to +\infty$ and limit $+\infty$ at either $t = 0$ or some value below 0. This means that either there is only one $t > 0$ for which this derivative is 0, and therefore $\sum_{k=1}^{K} \frac{1}{t+a_k} = R$, which then corresponds to the optimal branch length, or the optimal branch length is $t = 0$.

To find the value of t for which $\sum_{k=1}^{K} \frac{1}{t+a_k} = R$, in practice we use a bisection algorithm. Initial upper and lower bounds for the bisection algorithm are set to $t = K/R - \max_k(a_k)$ and $t = K/R - \min_k(a_k)$, with exceptions in case any of these values is negative.

This approach for inferring a branch length only requires one genome list comparison, but can require several calculations of the term $\sum_{k=1}^{K} \frac{1}{t+a_k}$. This is typically more efficient and more precise than our previous branch length inference approach, which used a bisection algorithm applied to the likelihood function itself and required a genome list comparison for each branch length evaluated during the bisection search[12]. To benchmark this approach for branch length inference, we compare it with our previous one (MAPLE v0.1.3) and with the deeper but slower optimization procedure in IQ-TREE (v3.0.1)[49]. We consider three sub-alignments of our global Viridian SARS-CoV-2 alignment (see below) of size suitable for IQ-TREE, containing the genomes from lineage B.1.429 (8,284 genomes), BA.1.17.2 (13,014 genomes) and AY.4.2.2 (11,003 genomes) respectively, as assigned by Pangolin[33] v4.3 with pangolin-data v1.21. These three sub-alignments are available from GitHub via https://github.com/NicolaDM/MAPLE/tree/main/example_files/. On each of these three alignments, we ran MAPLE and IQ-TREE with UNREST model (MAPLE option '--model = UNREST' and IQ-TREE options '--m UNREST -keep-ident -nt 16 -blmin 1e-13') to:

- Infer a fixed background tree topology using MAPLE v0.6.8 with rate variation (option '--rateVariation').
- Use this topology as fixed initial tree topology in MAPLE (options '--inputTree --noFastTopologyInitialSearch --numTopologyImprovements = 0 --doNotReroot') and IQ-TREE (option '--te') and infer only model parameters and branch lengths with MAPLE v0.1.3, MAPLE v0.6.8 with rate variation, MAPLE v0.6.8 without rate variation, IQ-TREE with rate variation (options '--mset UNREST --mrate E,I,G,I+G') and IQ-TREE without rate variation.
- Evaluate the likelihood of the estimated set of branch lengths using IQ-TREE with fixed input tree and branch lengths (options '--te --blfix') and either with or without rate variation, depending on the option used for branch length inference.

We find that our more recent procedure in MAPLE v0.6.8 estimates branch lengths with considerably higher likelihood than our old approach in MAPLE v0.1.3, and closer to IQ-TREE estimates (Extended Data Table 4).

Parallelization. The most time-intensive and memory-intensive part of the MAPLE algorithm is the topological improvement search through SPR moves[12]. We have now parallelized this step using the Python package 'multiprocessing' (https://docs.python.org/3/library/multiprocessing.html) to speed up the analysis of large collections of genomes. MAPLE can now partition the nodes of an initial phylogenetic tree into nonoverlapping sets, and assign each set to a different core. Each core then performs SPR searches for replacement of its assigned nodes in parallel. When all cores have finished their search, the possible SPR moves found for each node replacement are ranked based on the likelihood improvement they are expected to bring to the tree. Starting from the most promising (that is, the greatest likelihood improvement) one, a new serial SPR search is then performed, with the best-found SPR moves being this time actually implemented by modifying the tree before moving to the next most promising node to be replaced. Because the number of advantageous SPR moves found is typically small compared to the total number of nodes in the tree, this final serial step is typically very fast compared to the initial parallelized SPR search.

Parallelization is currently implemented in a distributed memory framework, which limits the efficiency of parallelization for large numbers of cores due to the time and memory cost of duplicating the data (Extended Data Table 1 and Extended Data Fig. 2). We are currently planning a shared-memory implementation of this same algorithm in CMAPLE[37].

Robinson–Foulds distances. We have implemented an efficient Robinson–Foulds distance[50] calculation in MAPLE for large phylogenetic trees. We use Day's 1985 linear time and memory algorithm[51]. Note that other approaches might be more efficient (see for example fastRF[52] and Hash-RF[53]), but our implementation is particularly useful when comparing trees with low divergence. This is because, unlike other implementations, we distinguish between branches with positive length and branches with 0 length. We consider a branch of length 0 as nonexistent or noninformative, that is, the clade defined by the branch is not considered part of the tree topology. Our implementation allows the specification of a minimum branch length threshold under which branches are considered absent from the tree (effectively of length 0); this helps preventing biases in tree distance calculations when considering trees that are effectively multifurcating, even though they might

be represented as binary, as usually is the case for the output trees of popular phylogenetic methods[49,54,55].

SARS-CoV-2 genome datasets

Viridian and GenBank SARS-CoV-2 genome dataset. To investigate possible recurrent errors in different consensus sequence calling pipelines, we created a dataset of genomes for which we could find raw read data and consensus sequences both from Viridian and GenBank. The approach to data collection is described in ref. 28. We separated the Viridian and GenBank consensus sequences to create two separate alignments. The genomes were aligned to the reference MN908947.3 using MAFFT v7.505 with options '--auto --keeplength --addfragment', which remove nucleotides from the alignments that are inserted with respect to the reference, thus creating a multiple sequence alignment of the same length as the reference genome. We then filtered out genomes with any of:

- 3,000 or more 'N' or '-' characters in either consensus sequence,
- 50 or more blocks of one or more consecutive 'N' or '-' characters in either consensus sequence, or
- 5 or more IUPAC ambiguity characters in either consensus sequence,

because these genomes might have been affected by low coverage or contamination. This resulted in two alignments each with 2,993,121 consensus genomes—one containing Viridian consensus sequences, and the other GenBank consensus sequences for the same genomes. We then ran MAPLE v0.6.7 with our error model multiple times on each alignment, iteratively masking positions that appeared most strongly affected by recurrent sequence errors (see the main text).

After this, we created a second alignment containing all the high-quality Viridian genomes available to us, and masking initial (1–78) and final (29769–29903) positions and alignment columns affected by putative recurrent errors (Extended Data Table 2). As before, we removed genomes that might have been affected by contamination, mixed infections and low coverage, this time with more stringent thresholds:

- 2,500 or more 'N' or '-' characters,
- 30 or more blocks of one or more 'N' or '-' characters, or
- 4 or more IUPAC ambiguity characters.

We also filtered out genomes based on coverage and heterozygosity patterns in the read data, that is, we removed genomes with any of:

- >2,500 positions with coverage ≤100,
- >1,500 positions with coverage ≤20,
- >30 positions with >5% minor allele frequency,
- >7 positions with >10% minor allele frequency, or
- >2 positions with >20% minor allele frequency.

From the resulting alignment, we masked deletions of length ≤30 bp, that is, we converted them to the reference sequence. While deletions do not usually cause problems to phylogenetic inference, and are instead usually interpreted as missing data, we observed that, in our dataset, errors (either in sequence or alignment) in a few genomes at positions of common deletions caused wrong substitutions to be inferred on ancestral branches, in addition to causing substantial phylogenetic uncertainty and instability (with consequently longer runtime for phylogenetic inference). We did not mask longer deletions in the consensus genomes as these often represent parts of the genomes with low coverage, and as such do not cluster together phylogenetically and so do not seem to cause these problems.

The resulting global Viridian alignment contains 2,072,111 genomes. From this alignment, we estimated a phylogenetic tree in MAPLE v0.6.8 under an UNREST model with rate variation in three steps:

1. We estimated an initial tree with one core, with options '--noFastTopologyInitialSearch --numTopologyImprovements 0'.
2. Then, starting from the tree inferred in the first step ('--inputTree' option), we used 14 cores in parallel to improve the tree topology with the option '--largeUpdate'.
3. Finally, starting from the tree inferred in the second step, we performed a deeper run of topology improvement with 14 cores with options '--largeUpdate --noFastTopologyInitialSearch --thresholdLogLKtopology 28.0 --allowedFailsTopology 8'.

The alignment, metadata, inferred tree and inferred substitution rates have been uploaded to Zenodo[32]. We also uploaded an annotated tree that can be visualized in Taxonium (https://taxonium.org/), and more efficiently the Taxonium Desktop App[30].

Simulated SARS-CoV-2 sequence data. We simulated SARS-CoV-2 genomes evolved according to a known ('true') background phylogeny and substitution model. We used as the background tree the publicly available (26 October 2021) global SARS-CoV-2 phylogenetic tree from http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/UShER_SARS-CoV-2/ (ref. 26) representing the evolutionary relationship of 2,250,054 SARS-CoV-2 genomes as obtained using UShER[10].

We used phastSim (v0.0.3)[56] to simulate sequence evolution along this tree according to SARS-CoV-2 nonstationary neutral mutation rates[13] and using the SARS-CoV-2 MN908947.3 reference as the root sequence. We also simulated variation in substitution rates, using in phastSim a gamma distribution of rates with $\alpha = 0.2$, leading to approximately the same substitution rate variance as in the substitution rates estimated with MAPLE from the GISAID real SARS-CoV-2 genome dataset considered in ref. 12.

Ambiguity characters were introduced in the alignment to mimic sequence incompleteness observed in real SARS-CoV-2 genome data. For each simulated sequence we sampled one random sequence from the GISAID real SARS-CoV-2 genome dataset considered in ref. 12 and copy-pasted from it the stretches of 'N' and gap '-' characters into the simulated sequence. Additionally, we counted the number of isolated ambiguous characters in the real sequence, and we masked an equal number of randomly selected single-nucleotide polymorphisms (differences with respect to the reference genome) in the simulated sequence. See ref. 12 for more detail.

Finally, we added sequence errors to the simulated alignment with a custom script; we used the same site-specific sequence error probabilities estimated from the GISAID SARS-CoV-2 genome dataset in ref. 12 using MAPLE with rate variation and sequence error model.

All these steps were meant to recreate a dataset similar to the real SARS-CoV-2 one, in particular in terms of phylogenetic complexity. However, phylogenetic inference might be considerably more uncertain in real data[36] due to unaccounted-for sources of complexity[57] such as recombination[34] and contamination.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The SARS-CoV-2 whole-genome alignment, metadata, inferred tree and inferred substitution rates have been uploaded on Zenodo via https://doi.org/10.5281/zenodo.12733488 (ref. 32).

Code availability

Models and algorithms have been implemented within the free and open-source phylogenetic software MAPLE (v0.6.8)[58] available from GitHub via https://github.com/NicolaDM/MAPLE/. The script MapleDataProcessing.py used to process the data is also available there.

References

39. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
40. Kozlov, O. *Models, Optimizations, and Tools for Large-Scale Phylogenetic Inference, Handling Sequence Uncertainty, and Taxonomic Validation*. Ph.D. thesis, Karlsruhe Institute of Technology (2018).
41. Klosterman, P. S. et al. XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics* **7**, 428 (2006).
42. Wu, C. F. J. On the convergence properties of the EM algorithm. *Ann. Stat.* **11**, 95–103 (1983).
43. De Maio, N., Holmes, I., Schlötterer, C. & Kosiol, C. Estimating empirical codon hidden Markov models. *Mol. Biol. Evol.* **30**, 725–736 (2013).
44. Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* Vol. 3 (ed. Munro, H. N.) 22–132 (Academic, 1969).
45. Collienne, L., Barker, M., Suchard, M. A. & Matsen IV, F. A. Phylogenetic tree instability after taxon addition: empirical frequency, predictability, and consequences for online inference. *Syst. Biol.* **74**, 101–111 (2025).
46. Kramer, A. M. et al. 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. *Syst. Biol.* **72**, 1039–1051 (2023).
47. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538 (2010).
48. Zhang, J., Kapli, P., Pavlidis, P. & Stamatakis, A. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* **29**, 2869–2876 (2013).
49. Wong, T. K. et al. IQ-TREE 3: phylogenomic inference software using complex evolutionary models. Preprint at *EcoEvoRxiv* https://doi.org/10.32942/X2P62N (2025).
50. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
51. Day, W. H. Optimal algorithms for comparing trees with labeled leaves. *J. Classif.* **2**, 7–28 (1985).
52. Pattengale, N. D., Gottlieb, E. J. & Moret, B. M. Efficiently computing the robinson-foulds metric. *J. Comput. Biol.* **14**, 724–735 (2007).
53. Sul, S.-J. & Williams, T. L. A randomized algorithm for comparing sets of phylogenetic trees. In *Proc. 5th Asia-Pacific Bioinformatics Conference* (eds Sankoff, D. et al.) 121–130 (World Scientific, 2007).
54. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree 2—approximately maximum-likelihood trees for large alignments. *PloS ONE* **5**, e9490 (2010).
55. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
56. De Maio, N. et al. phastsim: efficient simulation of sequence evolution for pandemic-scale datasets. *PLoS Comput. Biol.* **18**, e1010056 (2022).
57. Trost, J. et al. Simulations of sequence evolution: how (un)realistic they are and why. *Mol. Biol. Evol.* **41**, msad277 (2024).
58. De Maio, N. & Ly-Trong, N. NicolaDM/MAPLE: Maple 0.2.1. *Zenodo* https://doi.org/10.5281/zenodo.7584634 (2025).
59. Sanderson, T. & Barrett, J. C. Variation at spike position 142 in SARS-CoV-2 delta genomes is a technical artifact caused by dropout of a sequencing amplicon. *Wellcome Open Res.* **6**, 305 (2021).

Acknowledgements

We thank R. Corbett-Detig, Y. Turakhia, C. O'Cathail and C. Bielow for their help in the development of MAPLE and for useful discussions on these topics. N.D.M., S.M. and N.G. were supported by the European Molecular Biology Laboratory (EMBL) and by MRC grant MR/Z503769/1. M.W. and Z.I. were supported by EMBL. Z.G. and A.S. were funded by the Higher Education, Research and Innovation Department of the French Embassy to the United Kingdom. N.L.-T. and B.Q.M. were supported by a Chan Zuckerberg Initiative grant (no. EOSS4-0000000312) for Essential Open Source Software for Science. M.H. was supported by EMBL, the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with the UK Health Security Agency (UKHSA; NIHR200915) and the NIHR Biomedical Research Centre, Oxford. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health or the UKHSA.

Author contributions

N.D.M. conceived and implemented the methods and analyses, performed simulation tests, benchmarks and real-data analyses, and wrote the paper. M.W. implemented an early version of the error model in MAPLE. S.M. investigated the effect of random starting trees on the phylogenetic inference of MAPLE and implemented the rate variation model in CMAPLE. Z.G. performed preliminary analysis of recurrent errors in different SARS-CoV-2 datasets. A.S. worked on early implementations of the rate variation model in MAPLE. M.H. supported the construction and investigation of the real SARS-CoV-2 datasets. N.L.-T. contributed to the implementation of the rate variation model in CMAPLE. B.Q.M. supervised the implementation of methods in CMAPLE. Z.I. supervised the creation and analysis of real SARS-CoV-2 datasets. N.G. supervised the work and wrote the paper. All authors provided support during the drafting of the paper.

Funding

Open access funding provided by European Molecular Biology Laboratory (EMBL).

Competing interests

All authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-025-02932-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-025-02932-8.

Correspondence and requests for materials should be addressed to Nicola De Maio.

Peer review information *Nature Methods* thanks Karthik Gangavarapu and Alexandros Stamatakis for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team.

https://doi.org/10.1038/s41592-025-02932-8

Extended Data Table 1 | Performance of different algorithms in analysing a 100,000 SARS-CoV-2 genomes subsample of the full Viridian dataset (see Methods)

Method, task	Cores	Model	Runtime	Max memory	Log-likelihood
MAPLE v0.2.1, initial tree	1	GTR	1h:49m	4.11G	-2,398,633.2
> MAPLE v0.6.7, initial tree	1	GTR	2h:05m	1.74G	-2,400,287.4
> MAPLE v0.6.7, initial tree	1	UNREST	1h:56m	1.78G	-2,354,447.9
MAPLE v0.2.1, tree improvement	1	GTR	27h:25m	5.03G	-2,395,799.3
> MAPLE v0.6.7, tree improvement	1	GTR	10h:53m	2.69G	-2,390,275.5
> MAPLE v0.6.7, tree improvement	1	UNREST	16h:50m	2.63G	-2,346,497.9
> MAPLE v0.6.7, tree improvement	1	+Rate variation	19h:39m	2.66G	-2,173,639.2
> MAPLE v0.6.7, tree improvement	1	+Error model	18h:06m	2.68G	-2,170,842.1
> > MAPLE v0.6.7, tree improvement	3	+Error model	6h:33m	27.91G	-2,170,791.3
> > MAPLE v0.6.7, tree improvement	6	+Error model	4h:37m	42.30G	-2,170,791.3
> > MAPLE v0.6.7, tree improvement	12	+Error model	3h:12m	67.76G	-2,170,791.3
> > MAPLE v0.6.7, tree improvement	24	+Error model	3h:26m	118.49G	-2,170,791.3

Analyses possible prior to the work in this paper use MAPLE v0.2.1, while analyses using features presented here use MAPLE v0.6.7 (> represents single-core runs, and > > parallelized runs). The first three rows refer to initial tree construction: this is performed by stepwise addition (adding one sample at a time to the initial tree), after sorting samples based on both their divergence from the reference and the completeness of their sequence data (see[12]). Initial tree inference is listed here separately from tree improvement since the former cannot be parallelized in MAPLE v0.6.7. Tree improvement in MAPLE v0.2.1 was started from the MAPLE v0.2.1 initial tree. Tree improvement in MAPLE v0.6.7 with GTR model was started from the initial tree from MAPLE v0.6.7 with GTR model. All other tree improvement runs were started from the initial tree from MAPLE v0.6.7 with UNREST model. A '+' in the Model column indicates increasingly complex models, that is +Rate variation is combined with the UNREST model, +Error model is added to UNREST+Rate variation.

https://doi.org/10.1038/s41592-025-02932-8

Extended Data Table 2 | Masked Viridian genome alignment positions

Position	Minor variants	Ambiguities	Consensus differences	Phylogenetic clustering	Substitution rate	Likely errors	Notes
76		**154**	**30 A vs. 93**		0.5	< 100	**Low coverage**
78		18	**10 A, 15 C, 187 G vs. 80 A, 30 C, 51 G**		1.5	< 200	**Low coverage**
274	✓	**391**	**355 A vs. 9**	✓	1.4	≈ 300	**Low coverage of CA mutations**
4321	✓	**628**	707716 T vs. 709324	✓	4.6	Hundreds	**Heterozygous reversions in BA.2**
8008		239	**20135 A vs. 2**	✓	**26**	≈ 20000	**Low coverage, correlated with 8012**
8012		10	**4678 C vs. 1**	✓	**13**	≈ 4600	**Low coverage, correlated with 8008**
8826		77	**131 C, 237 G vs. 31 C, 1 G**	✓	**7.2**	< 400	**Low coverage, correlated with 8829**
8829		**301**	**801 A vs. 33**	✓	1.6	Hundreds	**Low coverage, correlated with 8826**
8835	✓	**52213**	**357989 C vs. 68802**	✓	**100**	≈ 360000	**ARTIC4 alternative primer binding[14], correlated with 15521**
15510		**713**	**1916 C vs. 733**	✓	**34**	≈ 1900	**Assembly difficulties**
15521	✓	**31294**	**756265 A vs. 188634**	✓	**100**	≈ 750000	**ARTIC4 alternative primer binding[14], correlated with 8835**
15854	✓	55	**424 A vs. 8**	✓	1.4	≈ 400	
17259	✓	141	**7898 G vs. 6677**	✓	**15**	≈ 7000	
19413	✓	**1332**	**1532 G vs. 74**	✓	**32**	≈ 1500	
19672	✓	**199**	**365 A vs. 182**	✓	0.09	< 400	**Possible primer trimming issue, correlated with heterozygosity at 19668**
21650	✓	**1110**	**95 C vs. 4**	✓	0.7	< 100	
21987	✓	**34374**	421580 G vs. 451142	✓	**100**	Thousands	**ARTIC3 amplicon dropout[59]**
22027		11	**1548 G vs. 11**	✓	**13**	≈ 1500	**Mapping issues near deletion 22029–34**
22033		82	**2364 T vs. 712**	✓	1.3	≈ 1500	**Mapping issues near deletion 22029–34**
22786	✓	**18127**	643395 C vs. 671746	✓	**100**	Thousands	**Issues downstream of AC mutation**
22882	✓	**1881**	1069445 G vs. 1109262	✓	**100**	Thousands	**Issues downstream of TG mutation**
23118	✓	**39690**	**528 T vs. 68**	✓	1.6	≈ 500	
23948	✓	**1257**	1361584 T vs. 1374120	✓	**25**	Thousands	**Issues downstream of GT mutation**
25296	✓	**888**	**177 C vs. 9**	✓	2.2	< 200	
25324	✓	22	**418 A vs. 0**	✓	0.7	≈ 400	**Correlated with 25336**
25336	✓	**1212**	**413 A vs. 572**	✓	1.9	≈ 400	**Correlated with 25324**
26530	✓	**4098**	653150 G vs. 659866	✓	**22**	Thousands	**Heterozygous reversions in BA.1**
26766	✓	**4190**	**1688 C vs. 2141**	✓	1.9	≈ 1700	
27507	✓	**842**	41636 C vs. 44029	✓	**40**	Hundreds	**Issues downstream of AC mutation**
29687		10	**251 C vs. 69**	✓	3.6	< 300	**Low coverage**

We also masked (not included in the table) position 25202 for computational convenience (it contains no alternative nucleotide calls, but 534,413 ambiguous calls), positions 22195, 22197, 22198, 22202, and 22204 due to recurrent alignment errors causing five linked artificial substitutions at these positions in about 50,000 genomes, and positions 28245, 28247, 28249, 28351, 28253 and 28254 for similar reasons. For the other positions we highlight in bold font what we consider red flags in the consensus alignment, read data, and the phylogenetic inference. 'Minor variants': > 5% frequency minor variants at these positions were systematically observed in the mapped reads of genomes inferred by MAPLE to contain a consensus sequence error. 'Ambiguities': number of ambiguous nucleotide characters counted in Viridian consensus sequences — higher numbers highlighted in red. 'Consensus differences': differences in numbers of alternative nucleotides in Viridian vs. GenBank consensus sequences (for example '30 A vs. 93' means non-reference A occurs 30 times in Viridian sequences vs. 93 times in Genbank data); large differences are highlighted in bold font. 'Phylogenetic clustering': phylogenetic clustering of errors and substitutions at the considered positions, assessed by visually inspecting the estimated phylogenetic tree annotated with inferred errors and substitutions. 'Substitution rate': substitution rate estimated by MAPLE: the genome-wide average rate is 1.0 and the maximum allowed is 100; high rates are highlighted in bold font. 'Likely errors': our estimate of the number of Viridian consensus sequence errors at the considered position, after taking all evidence into account. 'Notes': additional information, such as apparent or known causes of the errors at these positions. For these preliminary analyses, positions 1–72 and 29769–29903 were preventively masked due to being not called in a large fraction of the genomes.

https://doi.org/10.1038/s41592-025-02932-8

Extended Data Table 3 | Error rate inference from different initial error rate values

Initial error rate	Estimated mean error rate / 10^{-7}	log-likelihood
10^{-8}	1.95	-18,702,132
2×10^{-8}	2.00	-18,702,081
5×10^{-8}	2.06	-18,702,018
10^{-7}	2.11	-18,701,985
2×10^{-7}	2.17	-18,701,941
5×10^{-7}	2.23	-18,701,904
10^{-6}	2.29	-18,701,890
2×10^{-6}	2.43	-18,701,863
5×10^{-6}	2.65	-18,701,883
10^{-5}	2.83	-18,701,951
2×10^{-5}	2.86	-18,701,952
5×10^{-5}	2.88	-18,701,924
10^{-4}	2.95	-18,701,942
2×10^{-4}	3.00	-18,702,019
5×10^{-4}	3.04	-18,702,099
10^{-3}	3.04	-18,702,112

Inferred mean error rates and total likelihoods were estimated with MAPLE v0.6.8. All runs correspond to the full real SARS-CoV-2 dataset, on which we ran MAPLE with rate variation and site-specific error rates, and using the same pre-estimated fixed initial tree topology (but we re-estimated the branch lengths).

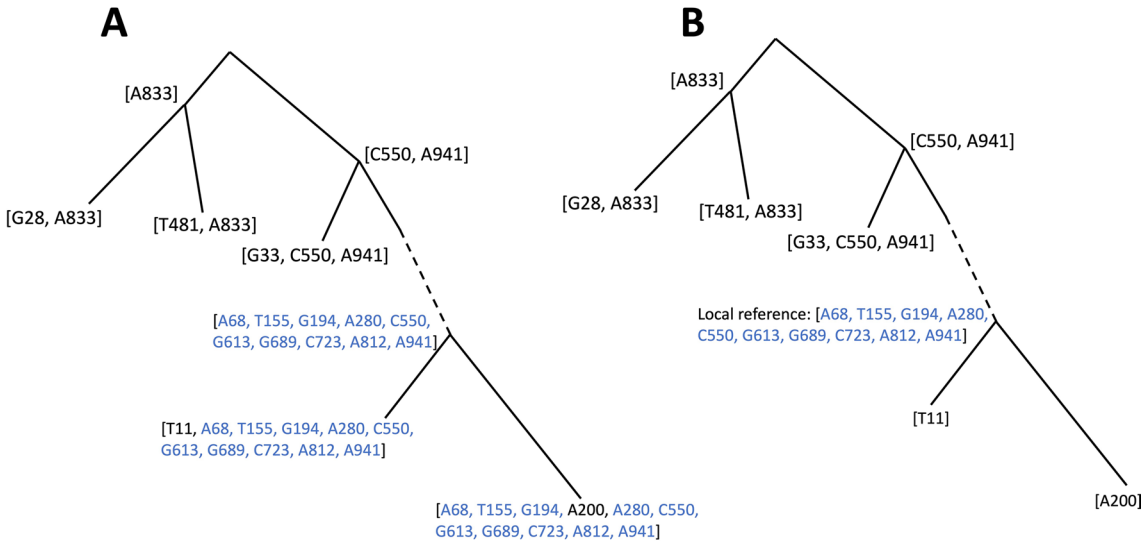
https://doi.org/10.1038/s41592-025-02932-8

Extended Data Table 4 | Likelihood of branch lengths estimated with different approaches in MAPLE

Dataset	MAPLE v0.1.3	MAPLE v0.6.8	MAPLE v0.6.8 rate variation
AY.4.2.2	− 81.26	− 2.739	+ 0.5454
B.1.429	− 444.3	− 19.93	+ 0.8764
BA.1.17.2	− 622.3	− 4.278	− 0.7378

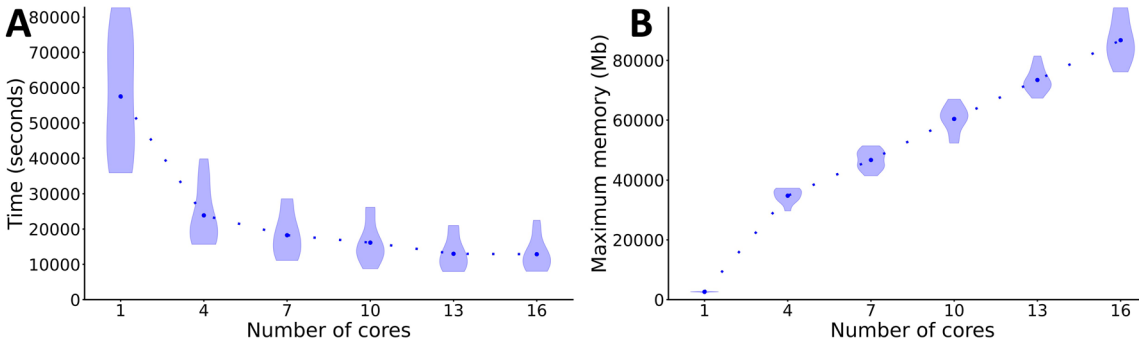
Values shown are the differences between the log-likelihood of the branch lengths estimated by the considered method (but with likelihood evaluated in IQ-TREE) and the log-likelihood of the branch lengths estimated by IQ-TREE. Higher values correspond to higher likelihoods, which are interpreted as better branch length estimates. For the column corresponding to MAPLE with rate variation, the log-likelihood is evaluated by, and compared to, IQ-TREE with rate variation.

https://doi.org/10.1038/s41592-025-02932-8



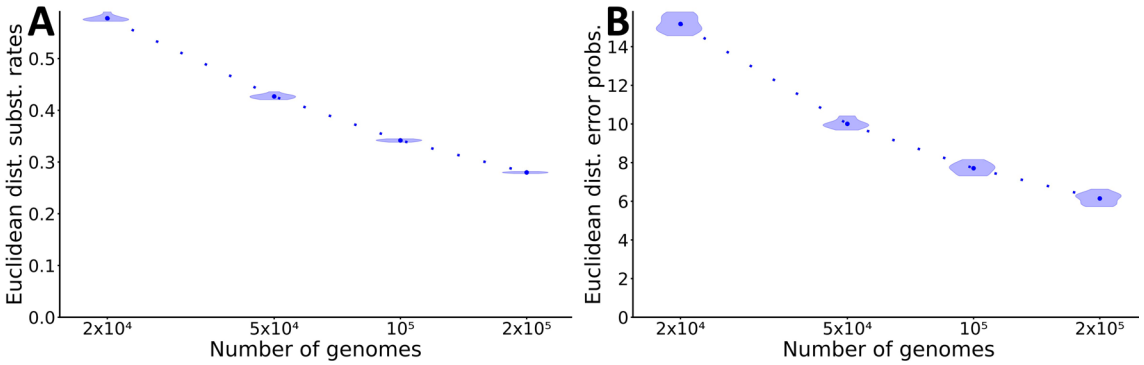
Extended Data Fig. 1 | Graphical representation of local references. Previously, we represented observed and ancestral genomes in terms of differences with respect to a fixed reference genome[12], as exemplified in **A**. Here, for graphical simplicity, the ancestral genome of the root of the tree is assumed to be the reference genome. We highlight in blue the mutations accumulated along the dashed branch which in this case are observed and represented in all the descendants of this branch (along with subsequent mutations T11 and A200, represented in black). Our more concise representation of the same tree and genome evolution history is showcased in **B**. Here a new local reference is defined at the bottom of the dashed branch; the genome of a local reference is represented in terms of its differences with respect to its parent reference (in this case the root genome). All genomes downstream of a local reference (except those that are eventually downstream of a further local reference) are represented in terms of their differences with respect to this local reference. This makes the representation of these genomes more concise, and their comparison faster.

https://doi.org/10.1038/s41592-025-02932-8



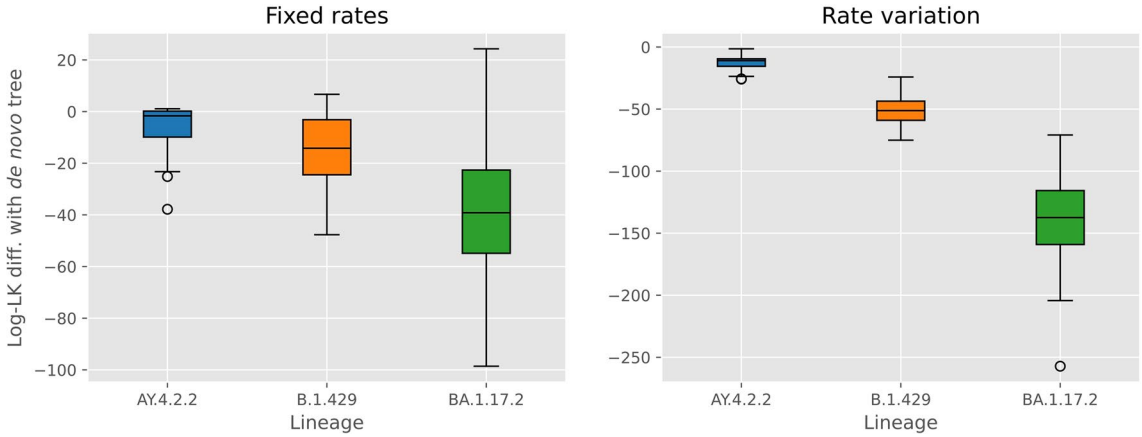
Extended Data Fig. 2 | Computational demand of parallelized SPR search in MAPLE. A Time and **B** maximum total RAM usage of tree optimization in MAPLE v0.6.7 on 100,000 real SARS-CoV-2 genomes. On the X-axis we show the number of cores used by MAPLE. For these analyses we used an UNREST substitution model with rate variation and site-specific error rates. For each number of cores considered we ran 10 replicates. Dots show the mean across replicates, while violin plots show variation between replicates.

https://doi.org/10.1038/s41592-025-02932-8



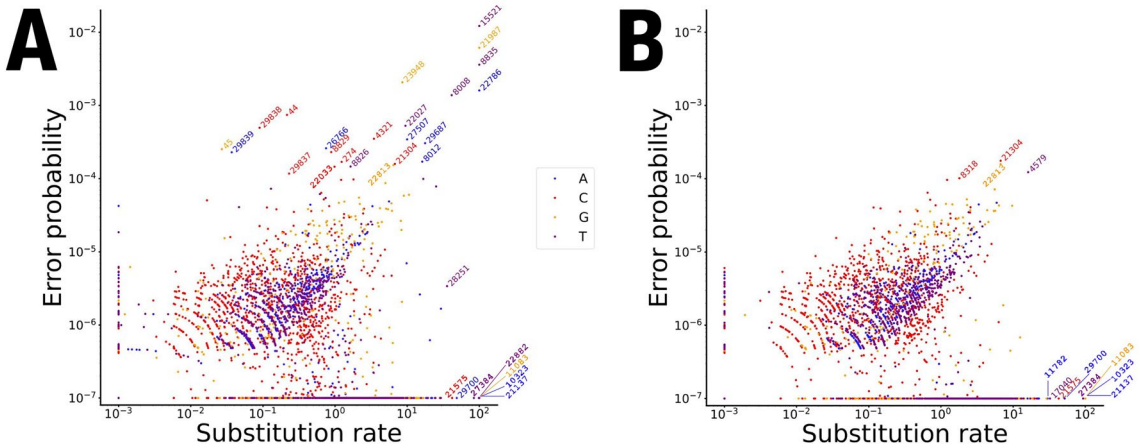
Extended Data Fig. 3 | Convergence of site-specific parameters. Normalised Euclidean distance between the vectors of inferred parameters from smaller datasets and the same vectors of inferred parameter values from the entire datasets. Smaller distances represent better convergence. Here we consider real SARS-CoV-2 genomes, with numbers of genomes on the X-axis. **A** Substitution rate parameters. **B** Error probability parameters. Both vectors have the same length as the reference SARS-CoV-2 genome MN908947.3 (29,903 parameters). We inferred all parameters with MAPLE v0.6.7 with the UNREST substitution model with rate variation and site-specific error rates. For each dataset size we ran 10 replicates. Dots show the mean across replicates, while violin plots show variation between replicates.

https://doi.org/10.1038/s41592-025-02932-8



Extended Data Fig. 4 | Convergence of phylogenetic inference in MAPLE starting from random trees. Log-likelihood difference between trees inferred in MAPLE v0.7.4.7 from random starting trees, and the tree inferred from scratch (the *de novo* tree). Trees were inferred with either no rate variation along the genome (left) or rate variation along genome (right). Negative numbers represent cases in which the tree inferred from scratch has a higher log-likelihood than the tree inferred from a random starting tree. On the X-axis we represent the three SARS-CoV-2 lineages whose genomes are considered (these are the same datasets considered in the Methods). Each box plot represents results from running 100 MAPLE runs from 100 distinct random starting trees. Random starting trees were obtained by permuting the leaves of the tree inferred from scratch. The three lines in each boxplot represent the first, second and third quartiles of the data points. The whiskers extend from the box to the farthest data point lying within 1.5 times the inter-quartile range from the box. Outlier data points are shown as circles.

https://doi.org/10.1038/s41592-025-02932-8



Extended Data Fig. 5 | Comparison of estimated substitution rates and error rates. A Initial sitewise substitution rate and error rate estimates from the initial unmasked alignment of > 2M real SARS-CoV-2 genomes. Note the logarithmic scales on both axes. Colours indicate the reference nucleotide at each position. We highlight positions with the highest estimated rates. **B** As **A** but using the estimates from the final masked alignment.

nature portfolio

Corresponding author(s): Nicola De Maio

Last updated by author(s): Oct 16, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data.
Data analysis	MAFFT v7.505 https://mafft.cbrc.jp/alignment/server/index.html . MAPLE v0.1.3, v0.2.1, v0.6.7, v0.6.8, and v0.7.4.7 https://github.com/NicolaDM/MAPLE . IQ-TREE v3.0.1 https://github.com/iqtree/iqtree2 , phastSim v0.0.3 https://github.com/NicolaDM/phastSim , Pangolin v4.3 (with pangolin-data v1.21) https://cov-lineages.org/resources/pangolin.html , Taxonium v2.0.115 https://github.com/theosanderson/taxonium .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The SARS-CoV-2 whole genome alignment, metadata, inferred tree, and inferred substitution rates have been uploaded on Zenodo, see https://doi.org/10.5281/zenodo.12733487

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not collected.
Reporting on race, ethnicity, or other socially relevant groupings	Not collected.
Population characteristics	Not relevant.
Recruitment	Not relevant.
Ethics oversight	Not relevant.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We included in the study all Viridian-called SARS-CoV-2 genomes available to us.
Data exclusions	We filtered out potentially contaminated sequences and masked alignment columns affected by recurrent sequence errors, as described in the manuscript.
Replication	To aid reproducibility we share data and software used for the analyses.
Randomization	Not relevant since no experimental groups were defined.
Blinding	Not relevant since no group allocation was performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	Not relevant.
Novel plant genotypes	Not relevant.
Authentication	Not relevant.