Addressing pandemic-wide systematic errors in the SARS-CoV-2 phylogeny

Hunt, M.; Hinrichs, A.S.; Anderson, D.; Karim, L.; Dearlove, B.L.; Knaggs, J.; ... ; Iqbal, Z.

Citation
Hunt, M., Hinrichs, A. S., Anderson, D., Karim, L., Dearlove, B. L., Knaggs, J., … Iqbal, Z. (2026). Addressing pandemic-wide systematic errors in the SARS-CoV-2 phylogeny. *Nature Methods*, *23*(3), 653-662. doi:10.1038/s41592-025-02947-1

Version:	Publisher's Version
License:	Creative Commons CC BY 4.0 license
Downloaded from:	https://hdl.handle.net/1887/4303055

Note: To cite this publication please use the final published version (if applicable).

nature methods

Resource

https://doi.org/10.1038/s41592-025-02947-1

Addressing pandemic-wide systematic errors in the SARS-CoV-2 phylogeny

Received: 18 October 2024

Accepted: 27 October 2025

Published online: 9 February 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

The majority of SARS-CoV-2 genomes obtained during the pandemic were derived by amplifying overlapping windows of the genome ('tiled amplicons'), reconstructing their sequences and fitting them together. This leads to systematic errors in genomes unless the software is both aware of the amplicon scheme and of the error modes of amplicon sequencing. Additionally, over time, amplicon schemes need to be updated as new mutations in the virus interfere with the primer binding sites at the end of amplicons. Thus, waves of variants swept the world during the pandemic and were followed by waves of systematic errors in the genomes, which had significant impacts on the inferred phylogenetic tree.

Here we reconstruct the genomes from all public data as of June 2024 using an assembly tool called Viridian (https://github.com/iqbal-lab-org/viridian), developed to rigorously process amplicon sequence data. With these high-quality consensus sequences we provide a global phylogenetic tree of 4,471,579 samples, viewable at https://viridian.taxonium.org. We provide simulation and empirical validation of the methodology, and quantify the improvement in the phylogeny.

On the eve of the SARS-CoV-2 pandemic, had one commissioned a poll of phylogeneticists on whether their methods were adequate for current public health needs, the overall response would have been in the affirmative. At that point, most people were analyzing relatively small datasets ($n < 5{,}000$), usually carefully curated and generally studied by people working closely with those obtaining and processing the clinical samples, or indirectly, via national public health organizations. Data were usually small and clean, and there was limited urgency. One year later, all of these statements would no longer be true. The SARS-CoV-2 pandemic placed unprecedented strains on the genomics and bioinformatics communities in terms of scale, turnaround time and coordination. In every dimension, tools and systems were pushed far beyond expectations. Despite significant efforts and innovations, numerous steps in the process (from patient to global phylogenies and dashboards) required prioritizing speed and practicality over absolute accuracy. This was the right thing to do at the time as it enabled real-time management decisions to be taken; however, as there was no unified genome assembly or quality control (QC) process, the end result has been that the set of SARS-CoV-2 genomes, on which future evolutionary and vaccine analyses will be based, contain a large number of systematic errors[1,2]. The goal of this study is to re-assemble all publicly available SARS-CoV-2 raw sequence data with a single analysis workflow to remove the majority of these errors, thereby building a higher quality phylogenetic tree for all our benefit.

Unlike the sequencing of bacterial genomes after culture (where the details of sequencing and assembly can stay the same over reasonably long periods) the specifics of viral sequencing and assembly during the pandemic had to keep changing, as we describe below. This resulted in a myriad of inconsistencies across the globe and errors in consensus sequences. A fundamental constraint on sequencing of SARS-CoV-2 was the fact that viral load in patient samples was generally very low and highly variable, as a result of which the most common way to sequence was via tiled amplicons (as had been carried out previously for other viruses[3]). Here, the genome is divided into overlapping 'tiles', each of which is independently PCR-amplified, guided by PCR primers at either end of the tile. That this was possible at all was thanks to two things: the early release of the genome sequence[4,5] and Quick's[6] rapid production of a set of primers, the first 'ARTIC' (acronym referring to a consortium) primer scheme. A feature of any tiled amplicon scheme is that, as the virus evolves, eventually mutations within primer binding

e-mail: zi245@bath.ac.uk

https://doi.org/10.1038/s41592-025-02947-1

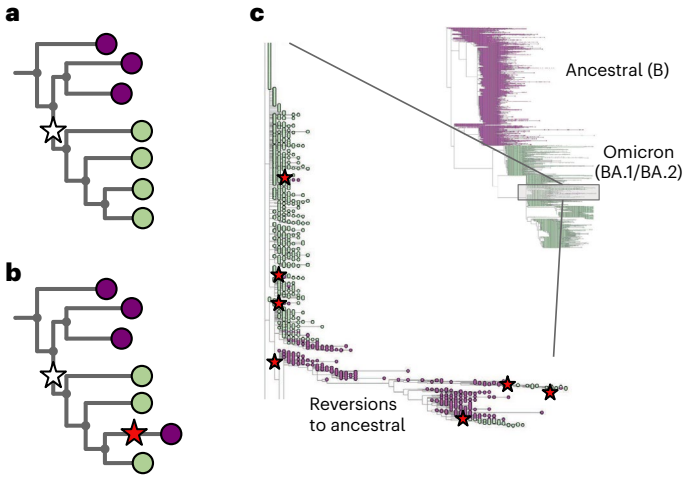

Fig. 1 | Assemblers which wrongly default to the reference base in the absence of data cause reversions in the phylogeny. a, Cartoon phylogeny built from perfect genomes, with leaves colored by genotype at a specific position X (purple, ancestral base; green, derived base). Just one mutation at this site, shown as a white star, is needed to explain the data. **b**, Cartoon showing the effect of assembly software assuming that a genome is identical to the reference genome when there are no data; here the amplicon containing position X is dropped in the lowest-but-one genome on the tree, creating one lone purple leaf. The tool which infers the phylogeny looks for a parsimonious explanation for this color distribution, and concludes it was caused by a mutation (white star) followed by a 'reversion' back to the ancestral base (red star). Errors in assembly caused by reference-bias tend to create enrichments of reversions. **c**, Part of the current UShER SARS-CoV-2 phylogeny, colored by genotype at genome position 22813 (spike codon 417). Blow-up shows multiple reversions back to the ancestral purple. A nonexhaustive set of artifactual mutations (such as reversions, unreversions and re-reversions) are shown with red stars, where there is a flip back and forth from green to/from purple.

sites will lead to failed amplification of the associated tile, creating gaps in the genome sequence data ('dropouts'). This is to be expected and necessitates the development of an updated scheme with new primers; however, as shown in Supplementary Fig. 1, many genome assembly software pipelines implicitly made the false assumption that in the absence of data (no reads from an amplicon) one should infer the sequence as being that of the reference genome, which in the case of SARS-CoV-2 is also the ancestral sequence. Thus, at various points during the pandemic, researchers analyzing the phylogeny would find a sudden crop of genomes 'reverting to the ancestor'.

In Fig. 1a we show part of a tree with the leaves colored to show what base that genome has at a specific position (purple for the ancestral base and green for the derived (new) base caused by a mutation shown as a white star). One single mutation explains that data. In Fig. 1b, we show the impact of wrongly assigning the ancestral base at the lowest-but-one leaf (fourth purple down). Here, the most parsimonious way to explain this is with a second mutation (red star) 'reverting' back to the ancestral purple. In Fig. 1c we show part of the global SARS-CoV-2 phylogeny hosted at taxonium.org (accessed 9 April 2024), zoomed in to show where Omicron branches from the ancestor. Leaves are colored by the genotype of genome position 22813 (codon 417) in the spike gene (again purple is ancestral). In the blow-up we see within the green (Omicron) clade, a striking spray of purple that does not sit cleanly in any subclade. Patterns like this, caused by systematic assembly errors, have been previously shown to occur in the SARS-CoV-2 phylogeny[1]. Such errors can have a considerable effect on our inferences about the underlying biology—in this case, K417N is a mutation that affects antibody escape[7] and systematic errors like this can lead to misinterpretation. However, although one can use a reversion count as a metric of whether we suspect there are assembly problems, reversions are not always errors. For example, SARS-CoV-2 has a C-to-T mutation bias[8,9] (strictly a C-to-U, as it is an RNA virus, but we convert to DNA space for phylogenetics), so if you have a T to C mutation on a phylogenetic branch leading to a large clade, you may expect to see multiple reversions back to T in that clade.

There are several other possible technical artifacts that can arise (for example, primer dimers[10], interactions between amplicons[11] or primers binding in noncanonical sites[12]), which should be expected and handled, otherwise additional errors will result. Unfortunately, these errors often correlated with individual sequencing centers, which themselves correlated with local prevalence of particular lineages at particular times. In addition, where amplicon dropout was incomplete, the likelihood of wrongly imputing the reference genome at a particular position becomes a function of decreasing amounts of sample RNA, creating a false relationship between genotype and viral load[13].

Because of amplicon dropouts, as the pandemic progressed and sequential waves of variants of concern (VOCs) arose, the ARTIC primer scheme was updated multiple times to restore amplification, as well as a slew of alternative options (for example, Midnight[14], AmpliSeq (Thermo Fisher Scientific) and VarSkip; https://github.com/nebiolabs/VarSkip). Each VOC wave brought mutations in primer bindings sites leading to amplicon dropouts, and a subsequent wave of artifacts in genomes as these were mishandled (Fig. 2). New amplicon schemes were then introduced, and gradually taken up, solving previous dropout problems, but also followed by smaller waves of new artifacts in the genomes, sometimes caused by primers not being correctly trimmed and being incorporated into assemblies. It is no exaggeration to say that since this issue was first raised[2], thousands of person-hours of time have been spent manually looking through trees and genomes trying to decide whether strange phenomena are artifacts or not. Some of us (R.C.D. and A.H.) have been maintaining the global phylogenetic tree of SARS-CoV-2 since 2021 (ref. 15), and the only way we have been able to maintain the integrity of the tree has been to (1) completely mask 150 nucleotide positions in the genome, as they are systematically too often wrong to ever be trusted, and (2) systematically mask (ignore) certain mutations on specific branches of the tree. As artifacts ebbed and flowed, and were discovered by analysts, the masking had to be updated (Fig. 2 and Supplementary Fig. 2). After the mammoth global efforts to sequence and collate these SARS-CoV-2 genomes, the richest dataset of any pathogen to date, it is critical to now reprocess and clean these data, providing a firm foundation for future discoveries.

As of June 2024, there were approximately 6 million SARS-CoV-2 raw sequence datasets deposited in the European Nucleotide Archive (ENA)/Sequence Read Archive (SRA), very few of which had metadata recording the primer scheme and the assembly pipeline used (data from COG-UK being a notable but geographically localized exception). In this paper we will describe our amplicon-aware assembly and QC processes, with which we reprocessed these genomes and measured the improvements in the genomes and phylogeny, and provide these data as a resource for the whole community.

Results

We set out to reprocess all available SARS-CoV-2 sequence read data, generating new consensus genomes through an assembly workflow designed for tiled amplicon schemes with a rigorous QC process, and thereby build a global phylogeny that minimizes the need for masking unreliable parts of the genome and tree.

To this end, we created Viridian, an efficient amplicon-aware assembler to consistently handle Illumina, Oxford Nanopore and Ion Torrent reads. As publicly shared sequence data do not generally have metadata logging the primer scheme used, Viridian first identifies the amplicon scheme from the input reads. In light of this, with knowledge of where primers bind, it then makes consensus sequences for each

https://doi.org/10.1038/s41592-025-02947-1

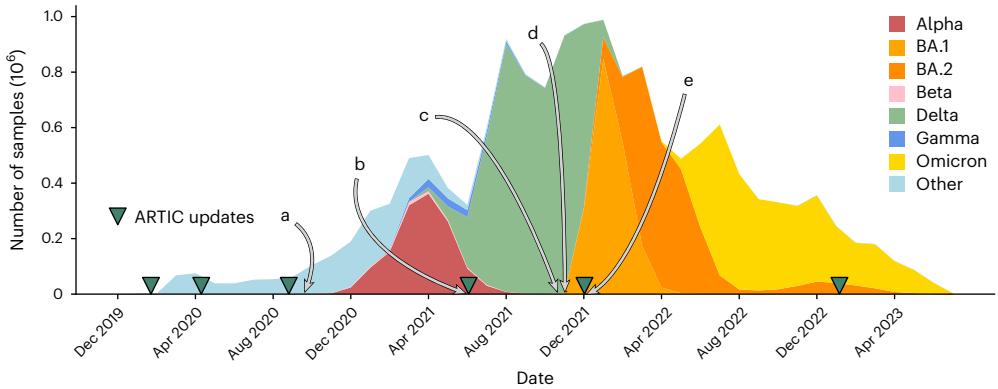

Fig. 2 | Timeline of the SARS-CoV-2 pandemic from December 2019 to July 2023. Selected events relating to problems with sequencing and consensus calling, labeled **a**–**e**. Releases of ARTIC primers schemes (v.1, 2, 3, 4, 4.1 and 5.3.2) are marked with green triangles. **a**, Primer dimers cause amplicon dropouts[10] and 28% of GISAID[30] sequences deposited in September 2020 have at least one gap of length at least 200 bp (ref. 31). **b**, A 9-bp deletion in the primer binding region of ARTIC v.3 amplicon 73 causes missing data[32]. **c**, Dropouts causing artifacts at spike 95 and 142 (ref. 13). **d**, ARTIC v.4 roll out triggers artifactual mutations in some pipelines[12]. **e**, Omicron samples cause ARTIC v.4 amplicon dropout, triggering the update to ARTIC v.4.1 (ref. 33).

amplicon by building a partial-order alignment graph of the reads using Racon[16], an approach that will detect indels more robustly than one based on pileups. Viridian then merges the per-amplicon consensuses into a single consensus and calls variants. To evaluate the confidence of each position in this consensus, it remaps the reads to the consensus, identifies unsupported positions, and using this, finally outputs a high-quality sequence that has low-quality bases masked. The emphasis throughout is on minimizing errors, in particular where amplicon primers bind, producing a consensus sequence where all unmasked positions should be correct.

We performed three evaluations of Viridian against two existing ARTIC workflow implementations: ARTIC-ILM (for Illumina) and ARTIC-ONT (for Nanopore) (Methods). The data used were (1) simulated data; (2) a 'truth set' of 67 runs from 27 isolates with known results; and (3) a larger dataset (n = 12,287, 'Early Omicron') from multiple countries in Africa from November 2021 to March 2022 that includes the emergence of the Omicron variant.

Primer scheme identification

We first evaluated our method for identifying primer schemes (Methods) using two datasets where we knew the correct primer scheme; these consisted of 8,000 simulated genomes and 67 curated truth genomes. There were zero errors. We then used 2,341,118 Illumina and 122,410 Oxford Nanopore samples where the ENA/SRA metadata had an ARTIC primer scheme version entry of 3 or 4, and compared to the call from Viridian (Supplementary Table 1). There was 99.7% agreement for Illumina and 98.2% for Oxford Nanopore samples. A manual investigation of a subset (n = 20) of the discordances concluded that the remaining errors were likely metadata errors in the ENA/SRA; in 19 of 20 cases, the pileups were categorical that Viridian was correct, and in the remaining one, the data were inconclusive (Supplementary Text 3 and Figs. 3–7). Note that both the truth set and the ENA/SRA data contain samples where tagmentation during the library preparation caused fragmented reads, confirming that the method worked there too.

Simulations

We simulated a SARS-CoV-2 tree of 8,000 genomes, including single-nucleotide polymorphism (SNP) errors in primers and amplicon dropouts. Illumina and Nanopore reads were simulated from each genome, from simulated amplicons using the ARTIC v.4 scheme. To evaluate the accuracy of resulting consensus sequences from ARTIC-ILM, ARTIC-ONT and Viridian, a novel pipeline was developed called CTE (COVID truth evaluation; Methods), which evaluates each consensus sequence using the truth to classify each position in the genome as correct or as an error. Results were highly consistent across all tools and amplicon schemes (Supplementary Table 2a–d). For Illumina data, ARTIC-ILM called all 395,799 SNPs and Viridian called 395,795 SNPs. With Nanopore data, ARTIC-ONT called 394,152 SNPs and Viridian 395,748. The 'missed' SNPs were called as Ns, not as reference bases, except for one Nanopore SNP called as heterozygous by Viridian. Although there were overall very few errors, ARTIC-ONT had notably more indel errors than Viridian (54 ARTIC compared to zero Viridian errors; Supplementary Table 2c,d).

Empirical truth dataset

The tools were compared on a truth dataset of 67 high-quality sequencing runs from 28 samples, comprising a mix of Illumina and Nanopore reads and ARTIC (v.3, v.4 and v.4.1) and Midnight amplicon schemes. The 'truth', including all expected SNPs in all runs, was determined by manual inspection of reads mapped to the reference genome. Similarly to the simulations, all tools performed well, with few errors (Supplementary Tables 3 and 4), and Viridian performing better with respect to indels on Nanopore data (43 ARTIC errors compared to 1 Viridian error; Supplementary Table 4e,f). Across the whole truth set there was a total of 1,696 SNPs, of which Viridian called 1,688 and ARTIC-ILM/ONT called 1,689. ARTIC-ILM/ONT had 1,989,650 correct reference calls, and Viridian 1,988,410. Missed SNPs and differences in reference calls were due to masking with Ns. We measured the peak RAM and total CPU time of each truth set run. Viridian had mean peak RAM usage of 444 MB and mean CPU time of 154 s, whereas ARTIC-ILM and ARTIC-ONT used 1.45 GB of RAM and took 366 s, and 1.80 GB of RAM, and took 561 s, respectively (Supplementary Table 5 and Supplementary Fig. 8).

African 'Early Omicron' dataset

Next, we evaluated our own empirical dataset, sequenced and assembled at the Centre for Epidemic Response and Innovation in South Africa, with samples from November 2021 to March 2022, including VOCs Alpha, Beta and Delta, and also encompassing the emergence of the Omicron variant. The 12,287 samples were from South Africa (n = 8,645), Angola (n = 957), Mozambique (n = 619), Mauritius (n = 488), Malawi (n = 480), Cameroon (n = 344), Zimbabwe (n = 333), Ethiopia (n = 232), Uganda (n = 102) and Namibia (n =83) (and four with unknown country), and include Illumina (n = 9,935) and Nanopore (n = 2,352) runs, using either ARTIC (n = 11,070 including v.3.4 and 4.1) or Midnight (n = 1,217) amplicon schemes (Supplementary Table 6). Each sample was processed with Viridian and ARTIC-ILM/ARTIC-ONT as appropriate, and the results compared to our original assemblies[17] which have previously been shared to the UShER[18,19] SARS-CoV-2 phylogeny via

https://doi.org/10.1038/s41592-025-02947-1

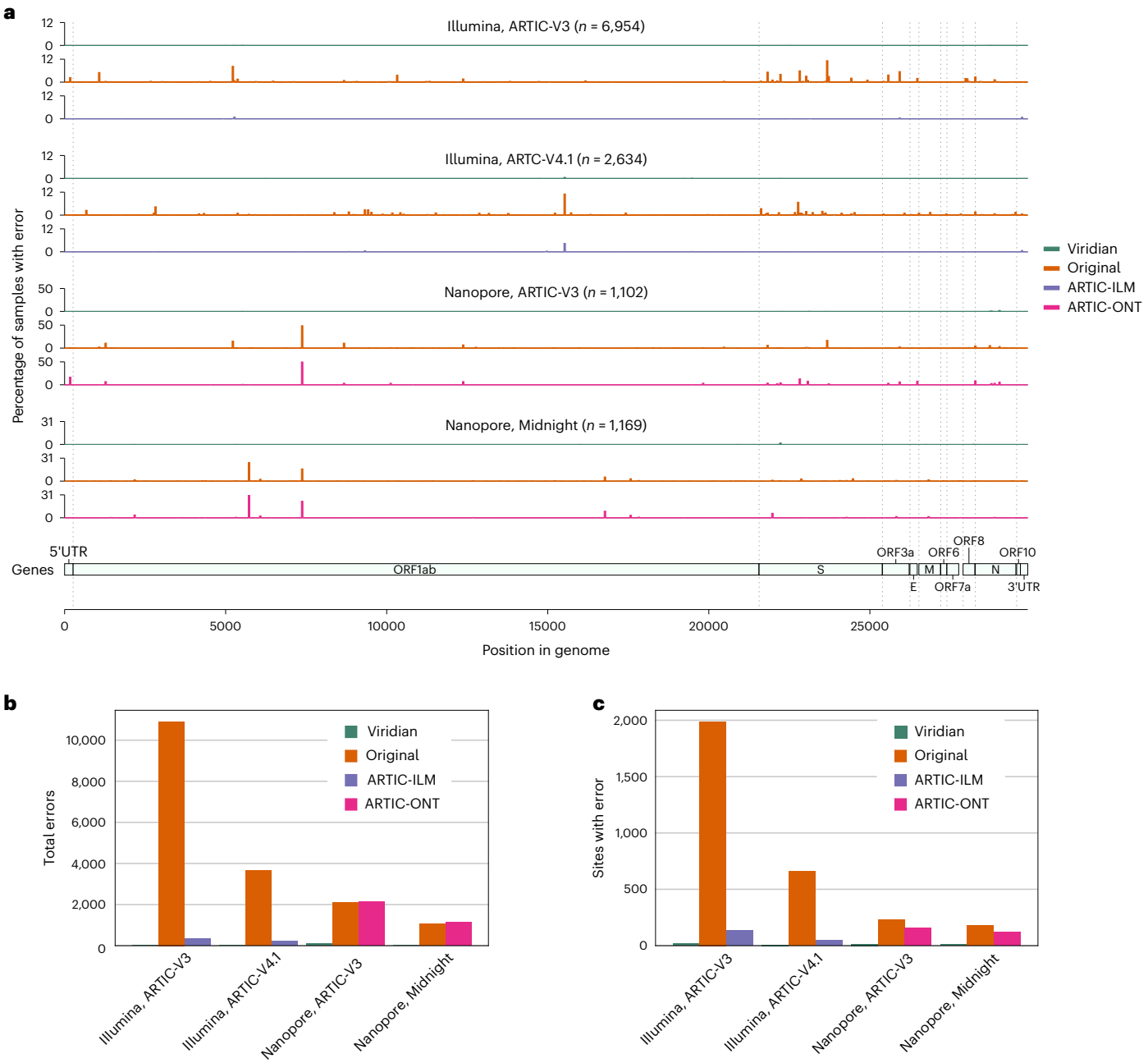

Fig. 3 | Errors in consensus sequences from the 'Early Omicron' African dataset, split by sequencing technology and amplicon scheme. a, Plots show the percentage of consensus sequences with an error, taking the maximum value in windows of length 50 bp. Error here is defined as where the consensus sequence has an A/C/G/T call, the read depth passes Viridian's default filters (Methods) and the reads support a different A/C/G/T call. Results are shown for Viridian, the original assemblies and for the ARTIC-ILM and ARTIC-ONT assembly workflows. **b**, Total errors called by each method, summarizing the data shown in **a** for each dataset. **c**, Total number of sites with at least one error.

GISAID. We scanned all positions in all consensus assemblies for 'hard errors', where the majority of the reads disagreed with the consensus (for example, the consensus called an A but most reads say G; Methods). We found systematic positional errors (which were specific to primer scheme and sequencing technology) in the original consensuses and the ARTIC-ONT assemblies. The errors were substantially reduced in the ARTIC-ILM workflow although some did remain. By contrast the errors were almost completely removed by Viridian. This is summarized in Fig. 3a, showing errors across the genome and total error counts and sites with errors. Depending on the dataset, total Viridian errors ranged from 31 to 86, whereas ARTIC had 219–2,148 errors, and the original assemblies 1,069–10,909 (Fig. 3b and Supplementary Table 7). The total number of positions in the genome where at least one sample had one error followed a similar pattern (Fig. 3c and Supplementary Table 7).

Assembly and evaluation of the global data

We processed all Illumina, Nanopore and Ion Torrent SARS-CoV-2 sequencing runs from the ENA/SRA as of 2 March 2023, keeping all 3,960,704 that passed QC (Methods) and produced a consensus sequence using Viridian. We also obtained all matching entries from GenBank, giving an 'intersection set' of 3,311,456 samples with both a Viridian and GenBank consensus sequence. We then built a tree of each of these three datasets (all 3,960,704 Viridian sequences, Intersection/Viridian (the Viridian assemblies of the intersection set), and

https://doi.org/10.1038/s41592-025-02947-1

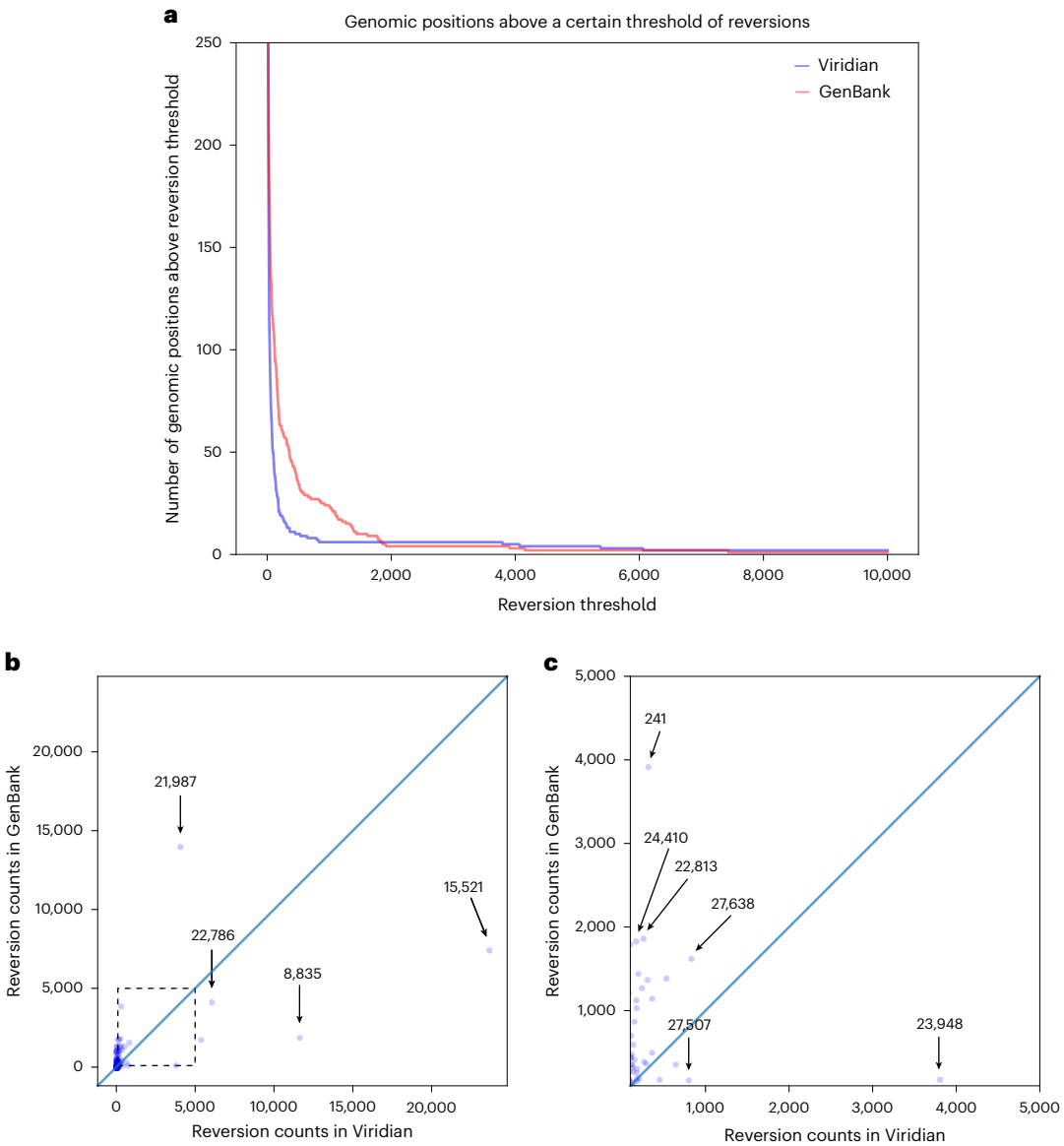

Fig. 4 | Most variable sites cause fewer reversions in the Viridian tree than the GenBank tree. a, Plot showing how many positions in the genome have at least *N* reversions in each tree (Viridian in blue, GenBank in red). Viridian curve drops faster, having fewer positions that create many reversions. **b**, Scatter-plot comparing count of reversion mutations found in the GenBank dataset and Viridian dataset. Note that (0, 0) is slightly indented from the origin of the plot. Each point represents a position of the SARS-CoV-2 genome. Three points below the line $y = x$ are highlighted (labeled by genomic coordinates 22786, 8835 and 15521) where Viridian has particularly high numbers of reversions, and one (labeled 21987) for GenBank. **c**, Blow-up of dotted square from **b** showing vast majority of variable sites in the genome lie above the line $y = x$.

Intersection/GenBank (the GenBank assemblies of the intersection set)) using MAFFT[20] and UShER (reverting deletions to the ancestral sequence and excluding insertions; Methods). Supplementary Fig. 9 provides an overview of how the samples were processed to make the trees. Note that these trees:

(1) are built from unmasked consensus genomes, unlike the current UShER global SARS-CoV-2 phylogeny, which pre-masks a list of 'problematic sites' in the genome where the community has determined assemblies may be unreliable, and
(2) do not have any forcible masking of particular mutations on the branches of specific VOCs, unlike the current public SARS-CoV-2 tree.

To assess the improvement in accuracy of a tree built from Viridian sequences, we next compared the Viridian and GenBank intersection set trees.

Ns and Pango assignment. A scatter-plot comparing the number of Ns in the Viridian versus GenBank assemblies (Supplementary HTML file) showed very little correlation, and a strong enrichment of points where there were many more Ns in the Viridian assembly—$n = 1{,}604{,}389$ (53.4%) of GenBank assemblies had no Ns, compared to $n = 1{,}197{,}638$ (39.8%) of Viridian assemblies. There were more Ns in the GenBank assembly for 9% of samples versus 49% samples with more Ns in the Viridian assembly; of those samples with more Ns in the Viridian assembly, 29% had zero Ns in the GenBank assembly. This is consistent with the known issue that for some software pipelines, portions of the reference sequence had been used to fill in dropouts for a large number of sequences, and this effect alone will have been a significant cause of reversions in the tree. Nevertheless, analysis at the lineage level using Pangolin showed very strong agreement, with only 0.98% ($n = 29{,}475$) of samples having discordant assignments. Of the mismatches, the majority (77%) were parent–child, with Viridian assembly the child

https://doi.org/10.1038/s41592-025-02947-1

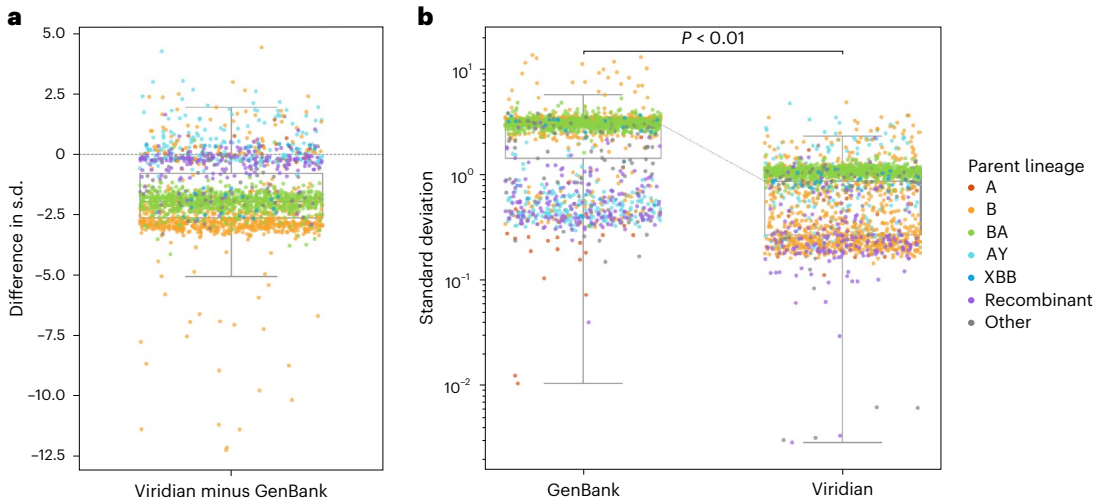

Fig. 5 | Comparison of uncertainty in growth estimates for different lineages when based on either the Viridian or GenBank tree. a,**b**, The same data are represented in two ways; each point represents one lineage. **a**, Plot shows the difference in s.d. of posterior density of relative growth rate estimate Δ log *R* (s.d. using the Viridian tree minus s.d. using the GenBank tree). Negative values here show that on average, the Viridian tree yields lower uncertainty than the GenBank tree. **b**, Plot shows the s.d. of the posterior density of relative growth rate estimate Δ log *R* based on the GenBank tree (left) and Viridian tree (right). The median s.d. of strain growth rate using the GenBank tree is 2.967, while the median s.d. using the Viridian tree is 0.859. This difference is statistically significant ($P = 2.85 \times 10^{-411}$, two-sided paired *t*-test; test statistic is 55, degrees of freedom is 2,120). Box-plots show first and third quartiles (lower and upper boundaries of box) and whiskers are set to the farthest point that is within 1.5 × the interquartile range from the box. Legend labels denote parent lineage.

(more specific) in 60% of those. Only 0.01% ($n = 287$) mismatched at the variant level. No Viridian assembly was 'unassigned', compared to 87 of the GenBank assemblies. Analysis of the results by collection date, country, technology and primer scheme revealed no category enriched for disagreements.

Indel calls. In samples where Viridian and GenBank assemblies result in the same Pangolin variant, indel calls are generally concordant and either very dominant or very rare; the mean discordance across indels for each variant was less than 1% for all variants except Zeta (1.1%), Lamba (1.4%), Omicron BA.3 (5.7%) and Theta (33%), which all had low number of sequences in this dataset (n ranging between 6 and 107). The characterizing insertion of TAC after position 21990 (S:YY144-145TSN) in Mu is an exception, found in 90% of Viridian assemblies but only 60% of GenBank assemblies. In samples where Viridian/GenBank have mismatched WHO variant calls, we see fewer indels per sample in GenBank versus Viridian (Supplementary HTML File). Notable differences at variant-defining indel sites; in particular, for samples assigned Delta for the Viridian assembly and Omicron for the GenBank assembly, we see two Delta-defining indels that are present in the Viridian assemblies, but absent in the GenBank assemblies. We show in Supplementary Fig. 10 those positions where there is discordance between Viridian and GenBank.

Reversions. One of the key signals of artifactual problems used during the pandemic, was finding positions in the genome (or branches of the tree) with very large numbers of reversions. We therefore used Matutils[15] and custom scripts to count the number of reversions in both trees, and plot this in two ways. In Fig. 4a, we show one minus the cumulative density function of reversions in the two trees, showing that the Viridian tree has far fewer positions with many reversions. To understand which positions are problematic, in Fig. 4b we show a scatter-plot comparing number of reversions at each position of the genome, in the Viridian and GenBank trees, with a blow-up of the central region in Fig. 4c. The main issue for phylogenetic analysis is positions with large numbers of reversions, so we care more about the graph away from the origin. We see that apart from a handful of positions far to the right and below the line $y = x$, all positions have fewer reversions in the Viridian tree. In other words, a smaller set of positions can be masked in the Viridian tree than in the GenBank tree to greatly reduce the number of reversions. For example, the GenBank tree has 63 positions with 200 or more reversions, while the Viridian tree has only 20. Supplementary Fig. 11 shows the specific example of genome position 22813 (introduced earlier in Fig. 1), comparing the current UShER global phylogeny with the Viridian tree.

Improved accuracy of lineage growth rate estimates. We ran PyR_0, a hierarchical Bayesian regression model that measures growth rates of SARS-CoV-2 lineages using genetic, temporal and geographical data[21]. When we ran this model on the Viridian tree, precision improved more than threefold on average compared to running the model on a GenBank tree. B- and BA-descended lineages had the largest decrease in the uncertainty of their growth rate measurements (Fig. 5). Improvements in precision occurred while maintaining accuracy. Supplementary Figs. 12–14 provide more detail.

Final global tree and masking

We updated our global sample list to include data from the ENA/SRA as of 28 June 2024, making a final global tree of the Viridian consensus sequences containing 4,471,579 samples. Tree construction was carried out, as is normal with UShER, by batching the samples, and then alternating adding a batch to the tree and optimizing the tree. In the process of doing this, we noted how the order in which samples were passed to UShER had a very significant effect on the deep structure of the tree. Passing them in in random order resulted in the initial tree being constructed with recombinant genomes, resulting in considerable misplacement of the VOCs. We determined that the best approach was first to construct a tree with samples with no missing data, passed in in temporal order, then to add lower quality samples later (Methods). After constructing the tree, we masked positions in the problematic -sites set, which includes highly homoplasic sites in addition to sites previously observed to be reversion-prone in SARS-CoV-2, and masked 31 reversions that occurred 200 or more times in the tree (this choice of 200 allowed us to exclude position 11083, which is highly homoplasic and one of the first problematic sites), but did not include 23040 where there have been true reversions multiple times in Omicron. After masking to remove artifactual reversions and highly homoplasic sites, we ran matOptimize[22], which

rapidly searches the tree for opportunities to reduce the total number of mutations in the tree by performing branch moves that group similar sequences together, thus maximizing parsimony.

Given the size of the tree, it would not be possible to use classical Felsenstein bootstrapping to measure phylogenetic uncertainty, so instead we use SPRTA[23], a method that shifts from assessing confidence of clades/groupings of taxa to instead looking at confidence of evolutionary histories (whether a lineage evolved from a specific other lineage or not). Supplementary Fig. 15 shows a histogram of confidences of nodes in the tree (raw data in Supplementary Table 8). We provide a second version of the phylogeny in a supplementary file, storing the SPRTA uncertainty information as metadata within file such that it can be detected by the taxonium viewer and explored interactively (screenshots in Supplementary Fig. 16). The user can ask taxonium to show low-support nodes, or for a specific node, to see what alternative places in the phylogeny they might equally well be placed (Supplementary Fig. 16b).

Effect on evolutionary and epidemiological analysis

The primary aim of this study is to provide a high-quality resource (assemblies and phylogeny), with less 'ad hoc masking', with the intention that it reduces systematic error and noise in downstream work of others. We give two example applications.

First, to estimate the effect of the reduced number of sequence/assembly artifacts in the Viridian assemblies on epidemiological analysis, we used geographic metadata for each sample and a pandemic-scale cluster estimation algorithm (matUtils, Cluster-Tracker[24]), to compare the number of inferred unique SARS-CoV-2 viral introductions in each country using the GenBank and Viridian data (Supplementary Table 9). The expectation would be that removing artifactual errors would reduce the number of small clusters, caused by errors pushing genomes out of the larger clusters they truly belong in, creating artificial 'introductions'. We found, for every country except Slovakia, there were more inferred introductions with the GenBank assemblies. The effect is more pronounced in highly sampled geographic regions, especially the USA (15,026 versus 13,626 introductions and 7,281 versus 6,676 singleton clusters for GenBank versus Viridian; Supplementary Fig. 17). As predicted, we see fewer small introductions with Viridian, and at the far right (note log scales) the very largest clusters are slightly larger.

Second, we quantified the extent to which the higher quality assemblies would affect estimates of differing mutational spectra of different VOCs[25]. In all cases the spectra were very similar (that is the effect was limited), but interestingly in Alpha there had been an odd T > A context (labeled with an arrow in Supplementary Fig. 18a) that was elevated above all others with the August 2022 UShER tree, which was gone in the Viridian data (Supplementary Fig. 18b). The difference in G > T mutations that had been observed previously between Omicron and non-Omicron is still very much present (Supplementary Fig. 19), confidence intervals (shown as error bars) do not always overlap the $x = y$ line, so there are minor differences in the exact values, but the overall trend and conclusions are unchanged.

Discussion

The pandemic was met with an unprecedented globally distributed sequencing effort that imposed substantial challenges for comparing and jointly analyzing data produced by thousands of labs with heterogeneous sampling, molecular, bioinformatic and analysis protocols. In particular, the downstream effect of using multiple variable-quality genome assembly workflows, inconsistent QC criteria and the inevitable coevolution of virus and amplicon schemas, led to systematic errors in genomes, and therefore the phylogeny.

Here we present Viridian, a fast, low-resource viral assembly tool specifically designed for tiled amplicon data and use it to produce a high-quality sequence dataset of all publicly deposited SARS-CoV-2 data from January 2020 through to June 2024. With this we were able to build a much higher quality phylogenetic tree, needing less masking, than the current phylogeny.

We hope for three outcomes. First, that this resource will provide a valuable substrate for detailed methodological, evolutionary and epidemiological analyses. This has already happened, with de Maio et al. developing new methods for handling mutation rate variation and sequencing errors in large phylogenies[26]. Second, that Viridian itself will prove useful, providing a significant improvement for Nanopore (and marginal for Illumina) compared to the ARTIC workflow, and a standardized single workflow and output format for Illumina, Nanopore and Ion Torrent. Third, that in future epidemics or pandemics, the tools and ideas from this paper will serve to reduce the amount of time spent poring over trees and trying to distinguish artifact from biology. Viridian will work for tiled amplicon sequencing of nonsegmented viruses where a consensus is the desired output (not in circumstances where multiple strains should be identified) and a single reference can be used. In other words, situations where there is limited structural variation or hypervariability, such as a particular outbreak or a recent zoonosis (such as SARS-CoV-2). Successful Viridian testing on 181 mpox samples using the data from Chen et al.[27] (Supplementary Table 10) is described in Supplementary Text 11 and Supplementary Figs. 20–23.

We note that a similar approach (amplicon-by-amplicon assembly followed by remapping for QC) has been previously used for HIV (https://github.com/neherlab/hivwholeseq?tab=readme-ov-file#1-mappingfiltering-sample-by-sample). An alternative approach, more robust to handling hypervariable regions, is to do amplicon assembly followed by de novo scaffolding of amplicons without use of a reference. This method was implemented in the tool Lilo, used for African Swine Fever Virus[28].

Despite all this, bioinformatic methods can only go so far. QC within a single laboratory is relatively easy, especially if one can use molecular protocols, such as negative controls and using synthetic spike-ins[29]; however, maintaining quality levels from distributed sequencing and assembly on a national and global scale is much harder. Our approach (uniform reprocessing) is actually the simplest, providing the raw data remains available; however, it is not a viable approach mid-pandemic when there is barely enough time to keep up with incoming data. We therefore advocate for improved standardization (and adoption) of metadata around sampling, assembly and QC, and also multinational 'simulations' of pandemics to better prepare for integrating data from different pipelines.

As the data in the ENA/SRA is heavily biased toward a few high-income countries (especially the USA and UK), we realized that it was important to increase the geographical breadth of our dataset. Our team submitted pre-existing raw sequence data to the ENA/SRA from Argentina, Austria, Germany, Ghana, India, the Netherlands, South Africa, Singapore and Sri Lanka. The worldwide distribution of samples is shown in Supplementary Figs. 24 and 25 (raw data are in Supplementary Table 11). It has been a privilege to work together to produce these high-quality resources for the benefit of all, which was only possible because raw sequence data were deposited in the ENA/SRA.

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/s41592-025-02947-1.

References

1. Turakhia, Y. et al. Stability of SARS-CoV-2 phylogenies. *PLoS Genet.* **16**, e1009175 (2020).
2. De Maio, N. et al. Issues with sars-cov-2 sequencing data. *Virlogical.org* https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473 (2020).

3. Henn, M. R. et al. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.* **8**, e1002529 (2012).
4. Holmes, E. Novel 2019 coronavirus genome. *Virological.org* https://virological.org/t/novel-2019-coronavirus-genome/319/1 (2020).
5. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
6. Quick, J. ncov-2019 sequencing protocol v.1. *protocols.io* https://doi.org/10.17504/protocols.io.bbmuik6w (2020).
7. Alenquer, M. et al. Signatures in SARS-CoV-2 spike protein conferring escape to neutralizing antibodies. *PLoS Pathog.* **17**, e1009772 (2021).
8. Di Giorgio, S., Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.* **6**, eabb5813 (2020).
9. De Maio, N. et al. Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biol. Evol.* **13**, evab087 (2021).
10. Itokawa, K., Sekizuka, T., Hashino, M., Tanaka, R. & Kuroda, M. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PLoS ONE* **15**, e0239403 (2020).
11. Cerutti, L. Missing g21987a mutation in SARS-CoV-2 delta variants due to non-specific amplification by artic v3 primers. *Virological.org* https://virological.org/t/missing-g21987a-mutation-in-sars-cov-2-delta-variants-due-to-non-specific-amplification-by-artic-v3-primers/764 (2021).
12. Sanderson, T. et al. Systematic errors associated with some implementations of artic v4 and a fast workflow to prescreen samples for new problematic sites. *Virological.org* https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/16 (2021).
13. 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).
14. Freed, N. E., Vlková, M., Faisal, M. B. & Silander, O. K. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biol. Meth. Protoc.* **5**, bpaa014 (2020).
15. 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).
16. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
17. Viana, R. et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
18. Turakhia, Y. et al. Ultrafast sample placement on Existing tRees (USHER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).
19. Hinrichs, A., Ye, C., Turakhia, Y. & Corbett-Detig, R. The ongoing evolution of USHER during the SARS-CoV-2 pandemic. *Nat. Genet.* **56**, 4–7 (2024).
20. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
21. Obermeyer, F. et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).
22. Ye, C. et al. matOptimize: a parallel tree optimization method enables online phylogenetics for SARS-CoV-2. *Bioinformatics* **38**, 3734–3740 (2022).
23. De Maio, N., Ly-Trong, N., Minh, B. Q. & Goldman, N. Assessing phylogenetic confidence at pandemic scales. *Nature* **647**, 472–478 (2025).
24. McBroome, J., Martin, J., de Bernardi Schneider, A., Turakhia, Y. & Corbett-Detig, R. Identifying SARS-CoV-2 regional introductions and transmission clusters in real time. *Virus Evol.* **8**, veac048 (2022).
25. Ruis, C. et al. A lung-specific mutational signature enables inference of viral and bacterial respiratory niche. *Microb. Genom.* **9**, mgen001018 (2023).
26. De Maio, N. et al. Rate variation and recurrent sequence errors in pandemic-scale phylogenetics. *Nat. Methods* https://doi.org/10.1038/s41592-025-02932-8 (2025).
27. Chen, N. F. G. et al. Development of an amplicon-based sequencing approach in response to the global emergence of mpox. *PLoS Biol.* **21**, e3002151 (2023).
28. Warr, A. et al. No part gets left behind: tiled Nanopore sequencing of whole ASFV genomes stitched together using Lilo. Preprint at *bioRxiv* https://doi.org/10.1101/2021.12.01.470769v1 (2021).
29. Lagerborg, K. A. et al. Synthetic DNA spike-ins (SDSIs) enable sample tracking and detection of inter-sample contamination in SARS-CoV-2 sequencing workflows. *Nat. Microbiol.* **7**, 108–119 (2021).
30. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Chall.* **1**, 33–46 (2017).
31. Cotten, M., Lule Bugembe, D., Kaleebu, P. & V.T. Phan, M. Alternate primers for whole-genome SARS-CoV-2 sequencing. *Virus Evol.* **7**, veab006 (2021).
32. Murall, C. L. et al. Recent evolution and international transmission of SARS-CoV-2 clade 19B (Pango A lineages). *Virological.org* https://virological.org/t/recent-evolution-and-international-transmission-of-sars-cov-2-clade-19b-pango-a-lineages/711 (2021).
33. Quick, J. Sars-cov-2 v4.1 update for omicron variant, https://community.artic.network/t/sars-cov-2-v4-1-update-for-omicron-variant/342 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025

Martin Hunt[1,2,3,4], Angie S. Hinrichs[5], Daniel Anderson[1], Lily Karim[5,6], Bethany L. Dearlove[7], Jeff Knaggs[1,2,3,4], Bede Constantinides[2,4], Philip W. Fowler[2,3,4], Gillian Rodger[2,4], Teresa Street[2,3], Sheila Lumley[2,8], Hermione Webster[2,4], Theo Sanderson[9], Christopher Ruis[10,11], Benjamin Kotzen[12], Nicola de Maio[1], Lucas N. Amenga-Etego[13], Dominic S. Y. Amuzu[13], Martin Avaro[14], Gordon A. Awandare[13], Reuben Ayivor-Djanie[15,16], Timothy Barkham[17], Matthew Bashton[18], Elizabeth M. Batty[19,20], Yaw Bediako[13], Denise De Belder[21], Estefania Benedetti[14],

Andreas Bergthaler[7], Stefan A. Boers[22], Josefina Campos[21], Rosina Afua Ampomah Carr[16,23], Yuan Yi Constance Chen[17], Facundo Cuba[21], Maria Elena Dattero[14], Wanwisa Dejnirattisai[24], Alexander Dilthey[25], Kwabena Obeng Duedu[16,26], Lukas Endler[7], Ilka Engelmann[27], Ngiambudulu M. Francisco[28], Jonas Fuchs[29], Etienne Z. Gnimpieba[30], Soraya Groc[31], Jones Gyamfi[16,32], Dennis Heemskerk[22], Torsten Houwaart[25], Nei-yuan Hsiao[33], Matthew Huska[34], Martin Hölzer[34], Arash Iranzadeh[35], Hanna Jarva[36], Chandima Jeewandara[37], Bani Jolly[38,39], Rageema Joseph[33], Ravi Kant[40,41,42], Karrie Ko Kwan Ki[43], Satu Kurkela[36], Maija Lappalainen[36], Marie Lataretu[34], Jacob Lemieux[12], Chang Liu[44,45], Gathsaurie Neelika Malavige[37], Tapfumanei Mashe[46], Juthathip Mongkolsapaya[20,44,45], Brigitte Montes[31], Jose Arturo Molina Mora[47], Collins M. Morang'a[13], Bernard Mvula[48], Niranjan Nagarajan[49,50], Andrew Nelson[51], Joyce M. Ngoi[13], Joana Paula da Paixão[28], Marcus Panning[29], Tomas Poklepovich[21], Peter K. Quashie[13], Diyanath Ranasinghe[37], Mara Russo[14], James Emmanuel San[52,53], Nicholas D. Sanderson[2,3], Vinod Scaria[39,54], Gavin Screaton[2], October Michael Sessions[55], Tarja Sironen[40,41], Abay Sisay[56], Darren Smith[18], Teemu Smura[40,41], Piyada Supasa[44,45], Chayaporn Suphavilai[49], Jeremy Swann[2], Houriiyah Tegally[57], Bryan Tegomoh[58,59,60], Olli Vapalahti[40,41], Andreas Walker[61], Robert J. Wilkinson[9,62,63], Carolyn Williamson[33], Xavier Zair[55], IMSSC Laboratory Network Consortium*, Tulio de Oliveira[57,64], Timothy EA Peto[2], Derrick Crook[2], Russell Corbett-Detig[5,6] & Zamin Iqbal[1,65] ✉

[1]European Molecular Biology Laboratory - European Bioinformatics Institute, Hinxton, UK. [2]Nuffield Department of Medicine, University of Oxford, Oxford, UK. [3]National Institute of Health Research Oxford Biomedical Research Centre, John Radcliffe Hospital, Headley Way, Oxford, UK. [4]Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford, Oxford, UK. [5]Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. [6]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA. [7]Institute for Hygiene and Applied Immunology, Center for Pathophysiology, Infectiology and Immunology, Medical University of Vienna, Vienna, Austria. [8]Department of Infectious Diseases and Microbiology, John Radcliffe Hospital, Oxford, UK. [9]Francis Crick Institute, London, UK. [10]Victor Phillip Dahdaleh Heart & Lung Research Institute, University of Cambridge, Cambridge, UK. [11]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK. [12]Department of Infectious Diseases, Massachusetts General Hospital, Boston, MA, USA. [13]West African Centre for Cell Biology of Infectious Pathogens (WACCBIP), University of Ghana, Accra, Ghana. [14]Servicio de Virus Respiratorios, Instituto Nacional Enfermedades Infecciosas, ANLIS "Dr. Carlos G. Malbrán", Buenos Aires, Argentina. [15]Biomanufacturing Group, International Centre for Genetic Engineering and Biotechnology, Trieste, Italy. [16]Department of Biomedical Sciences, University of Health and Allied Sciences, Ho, Ghana. [17]Tan Tock Seng Hospital, Singapore, Singapore. [18]The Hub for Biotechnology in the Built Environment, Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, UK. [19]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [20]Mahidol-Oxford Tropical Medicine Research Unit, Bangkok, Thailand. [21]Unidad Operativa Centro Nacional de Genómica y Bioinformática, ANLIS "Dr. Carlos G. Malbrán", Buenos Aires, Argentina. [22]Center of Infectious Diseases, Medical Microbiology and Infection Control, Leiden University Medical Centre, Albinusdreef 2, Leiden, The Netherlands. [23]Department of Computational Medicine and Bioinformatics, University of Michigan, Michigan, Ann Arbor, MI, USA. [24]Division of Emerging Infectious Disease, Research Department, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkoknoi, Thailand. [25]Institute of Medical Microbiology and Hospital Hygiene, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [26]College of Life Sciences, Birmingham City University, Birmingham, UK. [27]Pathogenesis and Control of Chronic and Emerging Infections, Univ Montpellier, INSERM, Virology Laboratory, CHU Montpellier, Montpellier, France. [28]Grupo de Investigação Microbiana e Imunológica, Instituto Nacional de Investigação em Saúde (National Institute for Health Research), Luanda, Angola. [29]Institute of Virology, Freiburg University Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany. [30]Biomedical Engineering Department, University of South Dakota, Sioux Falls, SD, USA. [31]Virology Laboratory, CHU Montpellier, Montpellier, France. [32]School of Health and Life Sciences, Teesside University, Middlesbrough, UK. [33]Divison of Medical Virology, University of Cape Town and National Health Laboratory Service, Cape Town, South Africa. [34]Genome Competence Center (MF1), Robert Koch Institute, Nordufer 20, Berlin, Germany. [35]Computational Biology Division, University of Cape Town, Cape Town, South Africa. [36]HUS Diagnostic Center, Clinical Microbiology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland. [37]Institute of Allergology and Immunology, University of Sri Jayewardenepura, Gangodawila, Nugegoda, Sri Lanka. [38]Karkinos Healthcare Private Limited (KHPL), Aurbis Business Parks, Bellandur, Bengaluru, India. [39]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India. [40]Department of Veterinary Biosciences, University of Helsinki, Helsinki, Finland. [41]Department of Virology, University of Helsinki, Helsinki, Finland. [42]Department of Tropical Parasitology, Institute of Maritime and Tropical Medicine, Medical University of Gdansk, Gdynia, Poland. [43]Department of Microbiology, Singapore General Hospital, Singapore, Singapore. [44]Chinese Academy of Medical Science (CAMS) Oxford Institute (COI), University of Oxford, Oxford, UK. [45]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [46]Health System Strengthening Unit, World Health Organisation, Harare, Zimbabwe. [47]Centro de investigación en Enfermedades Tropicales & Facultad de Microbiologia, Universidad de Costa Rica, San José, Costa Rica. [48]Public Health Institute of Malawi, Ministry of Health, Malawi, Malawi. [49]Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [50]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [51]Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, UK. [52]Duke Human Vaccine Institute, Duke University, Durham, NC, England. [53]University of KwaZulu Natal, Durban, South Africa. [54]Vishwanath Cancer Care Foundation (VCCF), West Mumbai, India. [55]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore. [56]Department of Medical Laboratory Sciences, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia. [57]Centre for Epidemic Response and Innovation (CERI), Stellenbosch University, Stellenbosch, South Africa. [58]Centre de Coordination des Opérations d'Urgences de Santé Publique, Ministere de Sante Publique, Yaoundé, Cameroun. [59]University of California, Berkeley, Berkeley, CA, USA. [60]Nebraska Department of Health and Human Services, Lincoln, NE, USA. [61]Institute of Virology, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [62]Centre for Infectious Diseases Research in Africa, University of Cape Town, Cape Town, South Africa. [63]Imperial College London, London, UK. [64]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), University of KwaZulu-Natal, Durban, South Africa. [65]Milner Centre for Evolution, University of Bath, Bath, UK. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: zi245@bath.ac.uk

https://doi.org/10.1038/s41592-025-02947-1

IMSSC Laboratory Network Consortium

Barbara Biere[66], Ralf Dürrwald[66], Christin Mache[66], Djin-Ye Oh[66], Jessica Schulze[66], Marianne Wedde[66], Thorsten Wolff[66], Stephan Fuchs[67], Torsten Semmler[67], Sofia Paraskevopoulou[67], Romy Kerber[68], Stefan Kröger[68], Walter Haas[68], Konrad Bode[69], Victor Corman[70], Michael Erren[71], Patrick Finzer[72], Roger Grosser[73], Manuel Haffner[74], Beate Hermann[75], Christina Kiel[76], Andi Krumbholz[77], Thomas Lorentz[77], Kristian Meinck[78], Andreas Nitsche[79], Markus Petzold[80], Thomas Schwanz[81], Florian Szabados[82], Friedemann Tewald[83] & Carsten Tiemann[84]

[66]Influenza and other Respiratory Viruses Unit, Robert Koch Institut, Berlin, Germany. [67]Genome Competence Centre Unit, Robert Koch Institut, Berlin, Germany. [68]Respiratory Infections Unit, Robert Koch Institut, Berlin, Germany. [69]MVZ Labor Dr Limbach, Heidelberg, Germany. [70]Institute of Virology, Charité-University Medicine, Berlin, Germany. [71]MVZ Laborzentrum Weser-Ems, Osnabrück, Germany. [72]MVZ Düsseldorf-Centrum, Düsseldorf, Germany. [73]Labor Dr Wisplinghoff, Köln, Germany. [74]MVZ Labor Dr Kirkamm, Mainz, Germany. [75]MVZ Dianovis, Greiz, Germany. [76]MVZ Labor Dessau, Dessau-Roßlau, Germany. [77]Labor Dr Krause, Kiel, Germany. [78]IMD-Laborverbund, Greifswald, Germany. [79]Highly Pathogenic Viruses Unit, Robert Koch Institut, Berlin, Germany. [80]Institut für Medizinische Mikrobiologie und Hygiene, Institut für Virologie, TU Dresden, Germany. [81]Institut für Medizinische Mikrobiologie und Hygiene, Universitätsmedizin Mainz, Mainz, Germany. [82]Laborarztpraxis Osnabrück, Georgsmarienhütte, Germany. [83]Labor Enders, Stuttgart, Germany. [84]Labor Krone, Bad Salzuflen, Germany.

Methods

Viridian pipeline

The main stages of the assembly process are to identify the amplicon scheme; sample the reads per amplicon; generate a consensus sequence by overlapping a consensus built for each amplicon; determine variants by aligning the consensus to the reference sequence; mask low-quality bases using read mapping to the consensus; and output a final masked consensus sequence. An overview of the pipeline is shown in Supplementary Fig. 26.

Amplicon scheme identification. The amplicon scheme is automatically identified from the reads, from the built-in set of schemes (users can optionally add their own): AmpliSeq (v.1); ARTIC (v.3, 4.1, 5.3.2_400, 5.2.0_1200)[34]; Midnight 1200[14]; and VarSkip (v.1a-2b; (https://github.com/nebiolabs/VarSkip).

The reads are mapped to the reference genome (default SARS-CoV-2 MN908947.3) using minimap2 (ref. 35) with options `-x map-ont` (Nanopore) or `-x sr` (Illumina/Ion Torrent). SAMtools[36,37] is used to make a sorted by coordinate and indexed BAM file, which by default is deleted at the end of the run but can be kept using the option `–keep_bam`. This BAM file is parsed using pysam (https://github.com/pysam-developers/pysam) to determine read depth across the genome and which amplicon scheme is the best match to the reads. Mappings flagged as secondary or supplementary are ignored. If reads are paired, then only proper read pairs are used. The pipeline is stopped at this stage if (by default) less than half of the genome has more than 20× read depth.

For each amplicon scheme under consideration, a normalized score is calculated based on the positions of mapped fragment ends. Throughout, 'fragment' means the mapped portion of an unpaired read, or the leftmost to rightmost mapping coordinates of a proper read pair. The idea is that fragment end mapping positions are expected to stack up at the left end of left primers and the right end of right primers, as the reads are from amplicon sequencing. The score is an overall measure of how close the fragment ends are to the primer ends.

At each position in the genome, the number of fragments with leftmost mapped end at that position is counted. These counts are used to score each amplicon scheme separately in turn (Supplementary Fig. 27). For each position in the genome, the distance to the nearest left end of a left primer in the scheme is found, moving to the left of that position. For example, if there is a left primer at position 100–130, then (assuming no other primers in this region), position 103 would have a distance of 3 (Supplementary Fig. 27a). Then at that position, we find how many fragments had their left end mapped at that position, and add that number to a counter of nearest distances. For example, if there were 20 fragments with left end at position 103, then 20 would be added to the counter for distance 3. The process is repeated similarly for right primers: for each position in the genome, the distance to the nearest right end of a right primer is found, moving to the right of that position (Supplementary Fig. 27a,b). The end result is a count of mapped fragment ends at each distance from a primer (Supplementary Fig. 27b,c).

The distance is normalized by taking the distance as a percent of the mean amplicon length for the scheme, and the count of fragment ends is normalized by taking the percent of total fragment ends. The results are binned, so that for each integer i in the range 0–100, we know the percent of fragments $f(i)$ ending normalized distance in the interval $[i, i+1)$ from a primer. The score is defined as

$$\sum_{i=0}^{100}(f(i)-i).$$

This is similar to calculating the area between the observed fragment counts and the line $y=x$ (Supplementary Fig. 27d), but negative values are allowed. The maximum possible score for perfect reads is 5,050, because $f(i)=100$ for all i and the score is then

$$\sum_{i=0}^{100}(100-i)=5,050.$$

Intuitively, a scheme that matches the reads will have fragment ends close to the primer ends, resulting in an initial steep curve. Conversely, a scheme that is not related to the reads should approximately follow the line $y=x$. Therefore, measuring the divergence from the $y=x$ line provides a reliable measure of how well the scheme and reads agree. Supplementary Fig. 27d shows cartoons of a matching and nonmatching scheme, and Supplementary Fig. 28 for a real example output by Viridian. Viridian chooses the scheme with the highest score; however, if the best score is less than 250, or less than double the second-best score, then the run is stopped and the sample is considered to be failed. For context, ERR8959196, shown in Supplementary Fig. 28, had best score of 4,290 and second-best score of 464. The default cutoffs, scores seen in simulated reads and in the empirical truth dataset are discussed in more depth in the supplementary material (Supplementary Tables 12 and 13 and Supplementary Figs. 29 and 30).

Read sampling. Once the amplicon scheme is known, reads are sampled to a target depth of (by default) 1,000× for each amplicon, or using all reads for an amplicon if the mean depth is less than 1,000×. If a fragment matches to more than one amplicon, then it is assigned randomly to one of the amplicons (the random number generator is seeded so that results are deterministic).

Within an amplicon, where there is more than one left primer (and similarly in the following description for right primers), the number of fragments supporting that primer is counted. Here, support is counted as the left fragment end being within 5 bp of the start of the primer. A primer is excluded from the remainder of the pipeline if it is supported by fewer than 20 fragments. The exception is that if no left primers for the amplicon have support, then all left primers are kept. The result is an inferred amplicon scheme, consisting of a subset of the original primers from the chosen scheme.

Each fragment is assigned to a left and right primer pair within its designated amplicon. These are chosen by taking the rightmost left primer and leftmost right primer that contain the fragment. In summary, at this point in the pipeline we have a set of reads for each amplicon with mean coverage 1,000× (or lower if there were not enough reads sequenced for an amplicon). Where an amplicon has more than one left and/or right primer, the set of reads is further split into sets for each primer pair.

Assembly. A consensus sequence is generated using a separate module called cylon (https://github.com/iqbal-lab-org/cylon). The overall method is to generate a consensus for each amplicon, overlap these consensus sequences into contigs, then scaffold against the reference sequence to output a final consensus sequence for the genome (Supplementary Fig. 31). It takes the inferred amplicon scheme (as described in the previous section) and a set of sampled reads for each amplicon. Reads are further sub-sampled for each amplicon from the 1,000× reads, with a target depth of (by default) 150× for Illumina and 250× for Nanopore or Ion Torrent.

A consensus sequence is generated for each amplicon by iteratively running Racon[16] until no more corrections are made, up to a maximum of ten runs. As Racon uses a partial-order alignment graph around the reference, this is a more reliable method of assembling sequence that contains indels than using mapping/pileup. If the input reads are paired, then each read pair is merged where possible using NGMerge[38] before running Racon. During testing, merging read pairs was found to improve the accuracy of Racon. In each Racon iteration, reads are mapped using minimap2 with options `-x map-ont` (Nanopore) or `-x sr` (Illumina/Ion Torrent). Racon options `–no-trimming –window-length W` are used, where `W` is the length of the amplicon plus 100 to avoid any erroneous indels at window ends. If no sequence

is returned from Racon, then the amplicon is classed as failed. The sampled reads are mapped back to the consensus sequence and all positions with less than 5× depth are masked with Ns. If the resulting sequence is shorter than 30 bp or has more than 50% Ns then the amplicon is failed.

Once there is a consensus sequence for each amplicon, adjacent amplicons are merged. First, amplicons are mapped to the reference genome using minimap2, and those with no mapping in the correct orientation are classified as failed and removed. If there is a perfect sequence match of at least 10 bp between adjacent amplicons, it is used to join them. Otherwise, if the minimap2 match coordinates imply that adjacent amplicons overlap (the reference positions overlap), then those matches are used. Finally, if the minimap2 matches do not have overlapping reference positions—for example, if one or both of the amplicons have a truncated consensus sequence—then a contig break is placed between the two amplicons.

Note that the start and end of the consensus sequence from each amplicon is excluded by this overlapping method, meaning that unreliable regions of consensus sequences that were inferred from reads starting or ending with primers are excluded. The only exception to this is where an amplicon is dropped, the next amplicon will include primer sequence; however, this is masked later in the QC stage. The amplicon overlapping is repeated for each adjacent pair of amplicons, stitching together a consensus sequence.

Once all possible adjacent amplicons have been merged, the result is one or more contig(s). When there is more than one contig, the position in the reference of each contig is determined using `nucmer` from the MUMmer software package[39]. The contigs are scaffolded, putting an estimated number of Ns between them based on the mapping coordinates. As there could be insertions or deletions in the sample, this number of Ns is not reliable, but it is corrected during the next stage.

Variant calling. Variants are called with respect to the reference genome using the function `make_truth_vcf` from the tool `varifier`[40]. This globally aligns the cylon consensus sequence to the reference genome to identify variants. As the amplicon schemes do not cover the complete reference genome, false-positive deletions are excluded from the start and end of the genome using the options `–global_align_min_coord`, `–global_align_max_coord` to restrict to coordinates within the amplicon scheme. Gaps in the consensus (that is, strings of Ns) are corrected to be the same length as the corresponding portion of the reference sequence using the option `–sanitise_truth_gaps`. These incorrect lengths can arise from failed amplicons, where the amplicon overlapping algorithm cannot always determine the exact gap length. For Nanopore and Ion Torrent reads, indels of length 1 or 2 are removed from the consensus sequence using the option `–indel_max_fix_length 2`. This removes false-positive indels caused by the error model of those technologies, at the cost of excluding real calls; however, in most cases any true-positive call that is removed will be masked later in the QC and masking stage of the pipeline.

The end result of this stage is a VCF file of variants, a consensus sequence with consistent gap lengths and the alignment of the reference and consensus sequences.

QC and masking. During read sampling to 1,000× read depth per amplicon, each fragment (read pair or single unpaired read) is allocated to a left and right primer, by taking the smallest primer range that spans the entire fragment. For each amplicon and each primer pair within that amplicon, all reads for that primer pair are mapped to the consensus sequence using minimap2 (with the same options as the original run of minimap2) and then pileup is run to gather coverage statistics. Keeping the reads partitioned in this way means that at each genome position, the results from one pileup run can be counted as either inside a primer ('bad' coverage) or not inside a primer ('good' coverage). This is outlined in Supplementary Fig. 32. Pileup is calculated using the `pileup` function from pysam with the `stepper` option set to `samtools`, and `ignore_overlaps` and `compute_baq` set to `False`.

Pileup results are aggregated at each position in the consensus sequence. This is used with the reference genome/consensus sequence alignment to output a tab-delimited report with read depth details at each position (split into separate counts for good and bad coverage). The good coverage is used to generate a masked consensus sequence, where untrustworthy positions are replaced with Ns. If the majority of reads disagree with the consensus position, or fewer than 20 reads in total agree with the consensus, then it is masked. At positions where there is evidence of more than one allele (by default an allele is counted as present if is supported by at least 20% of reads) then the consensus base is replaced with an ambiguous IUPAC code (for example, 'R' to mean 'A' or 'G').

Output files. The final masked consensus sequence is written in FASTA format, plus other files with additional information. Plots of read depth across the genome and scheme identification scoring are made. All QC results are written to a tab-delimited file with one position per row, including detailed read depth information. A log file in JSON format is written, with a high-level results summary section that includes all command line parameters, run time, version information and consensus sequence statistics. It also contains detailed information such as the multiple sequence alignment (MSA) between the reference and consensus, amplicon details (such as chosen primers and number of matching reads) and genome-wide read depth statistics.

Simulated data

We developed a Snakemake[41] pipeline to simulate tiled amplicon sequencing with PCR artifacts, to compare the assembly accuracy of Viridian to the Connor Laboratory (https://github.com/connor-lab/ncov2019-artic-nf) and Epi2me laboratories (https://github.com/epi2me-labs/wf-artic) ARTIC Nextflow workflows. First, to get a realistic tree 'shape' truth assemblies are simulated from a reference genome and reference phylogeny[15] using PhastSim[42] and obtained truth variant calls using varifier[40]. The primer sequences of the ARTIC v.4 amplicon scheme are then mapped to the truth assembly of each sample using the `aln` command of `bwa`[43] to get the start and end positions of each amplicon and check for sequence mismatches in primer binding regions. If one or more mismatches are identified, one of two possible PCR artifacts are simulated with equal probability: either the primer sequence containing the mismatch is replaced with the reference sequence, or the amplicon is assigned a read depth of 0. Random amplicon dropout is simulated with probability 0.001 and the sequencing depth of all other amplicons is drawn from a normal distribution ($\mu = 500$, s.d. 20). Reads are then simulated from each amplicon at the selected sequencing depths using ART[44] for Illumina and Badread[45] with `–identity 94,98.5,3` for Nanopore. The reads of each amplicon are aggregated such that there is one FASTQ of Illumina and one of Nanopore reads per sample and the reads are assembled using the Connor lab pipeline and Viridian workflow for Illumina and Epi2me labs pipeline and Viridian workflow for Nanopore. Finally, a new tool called COVID truth evaluation (CTE; https://github.com/iqbal-lab-org/covid-truth-eval), which is described in detail later, was used to generate TSV files that summarize the assembly accuracy for each tool.

Empirical truth set

Combined nasal and oropharyngeal specimens were identified during routine sequencing at Oxford University Hospitals NHS Foundation Trust as part of Pillar 1 national surveillance in the UK. Specimens were selected representing the Pango lineages B, B.1, B.1.1.7, B.1.1.7 (E484K), B.1.214.2, B.1.351, B.1.525, B.1.617.2, B.28, BA.1, P.1 and P.2. These were retrieved and cultured at the University of Oxford, generating abundant virus stocks. RNA from these virus stocks was sequenced

using Illumina and Oxford Nanopore instruments with both ARTIC and Oxford Nanopore Technologies (ONT) Midnight protocols, in addition to sequence-independent single primer amplification, forming the dataset deposited in ENA projects PRJEB50520 and PRJEB51850 (ref. 46). Sequencing was performed at the University of Oxford except where otherwise stated below.

Viral culture. Vero cells were maintained in DMEM high-glucose medium supplemented with 1% fetal bovine serum, 2 mM GlutaMAX, 100 IU ml^{-1} penicillin–streptomycin and 2.5 µg ml^{-1} amphotericin B at 37 °C, 5% CO_2 in a humidified atmosphere before inoculation with 200 µl of throat swab fluid. Cells were then incubated at 37 °C, with daily monitoring for cytopathic effects. When cytopathic effects reached 80%, virus-containing supernatants were collected through centrifugation at 3,000 rpm at 4 °C and stored at −80 °C in single-use aliquots. Virus titers were quantified by a focus-forming assay on Vero cells. Spike genes were sequenced to verify protein sequence integrity. Ref. 47 provides more details.

Extraction. Viral RNA was extracted from 200 µl and 400 µl volumes of Coplan viral transport medium on the KingFisher Flex system (Thermo Fisher) using the MagMAX Viral/Pathogen II Nucleic Acid Isolation kit (IVD). Two wash steps were incorporated and extracts were eluted in 50 µl.

PCR. PCR tests were performed by Oxford University Hospitals NHS Foundation Trust using two PCR assays: Altona RealStar (targeting E and S genes; Altona Diagnostics) and Thermo Fisher TaqPath assay (targeting S and N genes, and ORF1ab; Thermo Fisher).

Sequence-independent single primer amplification. Viral RNA was extracted as described above then complementary DNA was prepared using a SISPA approach[48]. In brief, first RNA was reverse-transcribed with SuperScript III Reverse Transcriptase (Life Technologies) using Sol-Primer A (5′-GTTTCCCACTGGAGGATA-N9-3′)[49]. Then 5 µl of cDNA and 1 µl (100 pmol $µl^{-1}$) primer B (5′-GTTTCCCACTGGAGGATA-3′) were added to a 50-µl reaction using AccuTaq LA (Sigma), according to the manufacturer's instructions. PCR conditions were 98 °C for 30 s, followed by 30 cycles of 94 °C for 15 s, 50 °C for 20 s, and 68 °C for 5 min, and a final step of 68 °C for 10 min. Amplified cDNA was purified using a 1:1 ratio of AMPure XP beads (Beckman Coulter) and quantified using the Qubit High Sensitivity dsDNA kit (Thermo Fisher Scientific).

SISPA Oxford Nanopore sequencing. SISPA products were sequenced following a previously described protocol[50] using ONT native barcoding (EXP-NBD104) and ligation sequencing (SQK-LSK109) kits with R9.4.1 flow cells.

ARTIC v.3 Illumina sequencing. Libraries were prepared using the NEBNext ARTIC SARS-CoV-2 Library Prep kit, following standard protocol with cDNA Amplicon and Ligation Bead Clean-ups (v.3.0 7/21). Manual library normalization was performed to ensure even sample coverage, based on the library's DNA concentration and average size, as measured by the Qubit (Thermo Fisher Scientific) and 2200 TapeStation (Agilent Technologies). Paired-end sequencing was performed using the MiSeq reagent kit v.2, with 2 × 250 bp, and one water control on each run. NEBNext Multiplex Oligos for Illumina (96 Unique Dual Index Primer Pairs) were used.

ARTIC v.4.1 Illumina sequencing. Libraries were sequenced at the University of Northumbria following the ARTIC V4.1 CoronaHiT-Illumina protocol[51], using an Illumina NextSeq 550.

ARTIC v.3 Oxford Nanopore sequencing. Sequencing was performed using the ARTIC LoCost protocol and v.3 primers using R9.4.1 flow cells. Final library concentration was quantified by the High Sensitivity dsDNA kit Qubit (Thermo Fisher Scientific).

ONT Midnight Oxford Nanopore sequencing. Libraries were prepared using ONT Midnight RT-PCR Expansion kits (EXP-MRT001) and rapid barcoding (SQK-RBK110.96), following manufacturer protocols. R9.4.1 flow cells were used.

Manual curation. All reads were mapped to the reference genome MN908947.3 using minimap2 with the `-x` preset `map-ont` for Nanopore reads and `sr` for Illumina. A sorted BAM file was made using `samtools sort`. This was used to make an unfiltered set of variant calls by piping the output of `samtools mpileup` into `bcftools call -vm`. Each sample was curated manually, using Artemis[52] to view the mapped reads and infer a truth set of variant calls. Although the unfiltered calls from bcftools were used as a guide, the whole genome for every sample was inspected for variant calls. In rare cases where the Nanopore and Illumina reads disagreed at a position, it was flagged as 'unknown'. The VCF files and metadata are available at https://github.com/iqbal-lab-org/covid-truth-datasets.

Consensus accuracy evaluation

The accuracy of results of the simulated data and truth set were evaluated using a new tool CTE. It can evaluate either a VCF file of variant calls, or a consensus sequence, by comparing it with a 'truth' consensus sequence. If the input is a VCF file, the consensus sequence to be evaluated is made by applying the variants to the reference sequence. It makes a MSA of the consensus, truth, and reference sequences using MAFFT[20]. Each position in the genome is classified by comparing the base calls of the MSA, to verify the accuracy of the consensus sequence. The most common case is that the truth nucleotide is equal to the reference nucleotide, and the consensus also called the reference nucleotide. The possibilities for the truth are a reference call, 'homozygous' SNP (that is, A, C, G or T, which is different from the reference), 'heterozygous' SNP (that is, a mix of A, C, G, T), indel, dropped amplicon or an N. Although rare, an N is used when the truth is unknown, as described above in the manual curation section. The possibilities for the consensus call are the same, except each nucleotide call could be correct or incorrect (the same as or different from the truth nucleotide). CTE reports the total count of each combination seen in the input sample.

Dropped amplicons are known in the truth data; however, they must be estimated from the consensus sequence that is under evaluation. As tools can use different methods to mask a nucleotide or an entire amplicon, defining a position with an N as part of a dropped amplicon, or simply masked, is ambiguous. CTE uses the minimum possible range of coordinates we would expect to be Ns if an amplicon is dropped, ranging from one past the end of the previous amplicon to the position before the start of the next amplicon. If a run of Ns contains this range of coordinates for a given amplicon, then it is considered as dropped in the sequence under evaluation. Hence there is some ambiguity between 'called as N' and 'dropped' when interpreting the output of CTE.

Africa dataset

The Africa dataset comprises a total of 12,287 samples, each of which has a 'GISAID' assembly, and either Illumina (n = 9,935) or ONT (n = 2,352) sequencing reads, with primer schemes ARTIC v.3 or 4, or MIDNIGHT-1200 (Supplementary table 6). All samples were processed with Viridian and ARTIC-ILM/ONT, producing a consensus sequence. Systematic positional errors were then identified using Viridian, which was run on each consensus sequence from Viridian and ARTIC-ILM/ONT using the option `–force_consensus`. This skips the de novo consensus building stage, instead using the provided assembly. The final QC stage is run as normal, which provides a method to evaluate the

input assembly. In particular, positions where the consensus sequence is not supported by reads can be identified. Figure 3a was generated using the branch of the Viridian code https://github.com/martinghunt/viridian/tree/qc_plots.

Global dataset

Metadata for all sequencing runs with taxon ID 2697049 were downloaded using the ENA portal query https://www.ebi.ac.uk/ena/portal/api/search?result=read_run&query=tax_id=2697049&fields=all&limit=10000000 on 2 March 2023. These runs were filtered to only keep those with `library_strategy` equal to `AMPLICON`, `library_source` equal to `VIRAL RNA`, `host` empty or equal to `homo sapiens`, and `instrument_platform` one of `ILLUMINA`, `OXFORD_NANOPORE` or `ION_TORRENT`. The resulting 5,288,952 sequencing runs were downloaded using either `prefetch/fasterq-dump` from the SRA-toolkit (https://github.com/ncbi/sra-tools) or `enaDataGet` (https://github.com/enasequence/enaBrowserTools). They were processed with Viridian, with 4,395,655 passing its QC requirements and producing a consensus sequence. These were further filtered for quality, requiring no more than three 'heterozygous' base calls (none of A, C, G, T, N) and no more than 5,000 Ns. The N count was taken from the consensus sequence after aligning to the reference using MAFFT, as described in the Trees section later. A further round of filtering was applied based on requiring a reliable date for each sequencing run, using where available the collection date from the ENA/SRA, COVID-19 Genomics UK Consortium (COG-UK) and GISAID. Runs with no collection date from any source were removed. Where dates conflicted for a given sample, the order of preference used was the date with highest resolution, then COG-UK, GISAID and finally ENA/SRA. At this stage, there were 3,960,704 runs, which is the set of runs used to compare with GenBank sequences (see next paragraph). Finally, the data were updated on 28 June 2024, adding all new runs that passed the same QC requirements, making a total of 4,484,157 consensus sequences.

All GenBank genomes were downloaded on 23 May 2023 using the Datasets tool (https://github.com/ncbi/datasets) with parameters `download virus genome taxon SARS-CoV-2`. The genome and metadata files (`genomic.fna.gz`, `data_report.jsonl.gz`) were extracted from the downloaded zip file. Genomes with host taxon ID ('`host`' → '`taxId`') 9606 (human), were kept. The genomes were matched to sequencing runs from the ENA/SRA using the run accession. Only GenBank genomes that matched to a single run that also belonged to the set of 3,960,704 Viridian consensus sequences (from the initial data obtained on 2 March 2023) were kept. This resulted in an 'intersection set' of 3,006,407 runs with both a Viridian consensus sequence and GenBank genome.

Primer scheme validation

As the COG-UK metadata includes the ARTIC primer scheme version, we used their project PRJEB37886 (included in the global dataset) to validate the scheme calls from Viridian. The ARTIC primer scheme version used was obtained from the SRA metadata using `efetch` (https://www.ncbi.nlm.nih.gov/books/NBK179288/) to download metadata for experiments in batches using the options `-format xml -db sra -input ids.txt`, where `ids.txt` is the name of the file containing a list of experiment accessions. The primer scheme version was extracted for each experiment from the value of the `artic_primer_version` tag in the `EXPERIMENT_ATTRIBUTES` section of the XML data. Each `efetch` command was attempted twice (failures were common), resulting in a total of 2,485,169 primer scheme calls from ENA/SRA metadata. We then restricted to Illumina and Nanopore samples that passed Viridian (the 4,395,655 samples described earlier), and only included ENA/SRA primer scheme values of `3/ARTIC v3` for ARTIC v.3 and `4/4.1alt/ARTIC v4` for ARTIC v.4. This was a total of 2,341,118 samples.

Discordant samples for manual inspection were chosen by taking all Illumina samples with ENA/SRA scheme v.3 and Viridian scheme v.4, sorting by run accession, and taking five equally spaced runs from the list. The same method was used for Illumina with ENA/SRA v.4 and Viridian v.3, and then similarly for Oxford Nanopore samples, totaling 20 samples for manual inspection. Reads were mapped using minimap2 with the option `-a` to make SAM output, and the preset `-x` of `sr` (Illumina) or `map-ont` (Nanopore). A sorted BAM file was made using SAMtools, and then manually inspected with Artemis.

Trees

Trees were built using MAFFT and UShER[18] and visualized with taxonium[53]. Each sequence was aligned to the reference using MAFFT with the option `–keeplength` to force the alignment to be the same length as the reference genome, by only allowing gaps in the query sequence. The alignment was modified by forcing any gaps in the query sequence to be the same as the reference sequence. The resulting sequences were batched into sets of size 100,000. A VCF file was made for each batch with `faToVcf`, with the option `-includeNoAltN`. A tree was built by adding each batch in turn using `usher-sampled` and the option `–sort-before-placement-3`. The final tree was optimized with the UShER command `matOptimize` and the options `-m 0.000000001 -r 8 -T 20`. Finally, the taxonium input file was generated using the script `usher_to_taxonium` from `taxoniumtools`[53]. The processing of input sequences to obtain taxonium input was implemented in a pipeline called Ushonium (https://github.com/martinghunt/ushonium).

To maintain an accurate tree structure, we ordered the samples by first using the samples with zero N or heterozygous calls, sorted by collection date. Then the remaining samples were used, again sorted by collection date. An exception to the date ordering was the 12,953 samples (3,876 of these were in the intersection set of 3,006,407 samples) where the GISAID date was given priority over other sources, which were added at the end instead of using the date. Using the highest quality consensus sequences first meant that UShER did not have to impute any ambiguous positions in any sequences. Sorting in date order meant that recombinant genomes (which emerged later in the pandemic) were not added to the tree too early, as they could be placed in an incorrect clade and then cause structural errors.

The global Viridian tree was built in two stages. A first version of the tree was built from the runs up to the 2 March 2023, using the order described above (highest quality and earliest collection date first). Positions in the problematic-sites set (https://github.com/W-L/ProblematicSites_SARS-CoV2) were masked globally in the tree, and 31 reversions found to occur at least 200 times in the tree were also masked globally (all masked positions are listed in Supplementary Table 14). matOptimize was run following the masking to join branches that had been split by the masked substitutions or reversions. This tree was used as a starting point to update using the second batch of data from 28 June 2024, with the same ordering method. The problematic-sites positions were masked in new sequences before they were added to the tree. After the new sequences were added, in addition to masking the 31 reversions that occurred at least 200 times in the first version of the tree before masking, we added branch-specific masking for regions in BA.1 and BA.2.86 in which mafft misinterprets a deletion and insertion in close proximity as a series of substitutions. Positions 6513, 6515, 22195, 22197-8, 22202 and 22204 were masked in the BA.1 branch. Positions 21610, 21612-3, 21615-7, 21619-21, 21624-7, 21629, 21632, 21637 and 21639-41 were masked in BA.2.86. matOptimize was run after masking. 12,578 duplicate runs were removed from the tree that came from shared samples, to make a final tree with 4,471,579 unique samples/runs. We note that there are only 14 duplicate runs in the intersection tree, which were not removed.

Measuring uncertainty in the global tree

We ran SPRTA on our tree with the Jukes–Cantor model, obtaining measurements of uncertainty and alternative placements of nodes which correspond almost exactly to alternative equally

parsimonious trees. To do this, MAPLE v.0.7.2 was run with the options `--doNotOptimiseBLengths --doNotImproveTopology --numTopologyImprovements doNotReroot`, which prevent any tree improvement in MAPLE, so that the output tree is the same as the input tree. The option `–normalizeInputBLen 0.000033` was used, which rescales the branch lengths to match the length unit used in MAPLE (expected substitution per site versus the unit of number of substitutions used by UShER). A JC69 model[54] was used with the option `--model JC`. SPRTA was run with the option `--SPRTA`, while representing alternative placements in the output tree and metadata (using the option `–networkOutput`). The other options used were `--largeUpdate --estimateMAT --numCores 10 --reference NC_045512.2.fa`.

PyR_0 analysis

PyR_0 was run using Python v.3.10. Code is available via GitHub at https://github.com/broadinstitute/pyro-cov?tab=readme-ov-file.

Analysis was conducted using the matched Viridian tree and GenBank tree of the intersection dataset. PyR_0 estimates growth rate of lineages using a hierarchical regression model (see ref. 21 for details); based on this, the standard deviation of strain growth rate was aggregated across regions (countries or first-level country divisions (for example, state or province) if the first-level division has at least 50 samples) by summing region-specific standard deviations. A paired *t*-test was conducted on the standard deviation in growth rate estimates using the Viridian tree versus GenBank tree. Supplementary Manhattan plots (spike protein and whole genome) only show mutations that appeared in both Viridian and GenBank trees, and a paired *t*-test was conducted on the growth rate estimates for each mutation. An unpaired *t*-test was also conducted on the full set of mutations, including those that only appear in the Viridian or GenBank trees, though no statistically significant results were found. Accompanying each Manhattan plot (Supplementary Figs. 12 and 13) is a plot of the ratio of growth-related mutations to all mutations, where growth-related mutations are defined as those which are at least one s.d. from zero. Fisher's exact test was performed to analyze the difference in proportions of growth-related mutations in each annotated subdomain/reading frame of the spike protein/whole genome (respectively). To produce Supplementary Fig. 14, rank was assigned according to the mean of the posterior density of the relative growth rate of a strain compared to the ancestral strain (denoted by *R*/*RA*) divided by the standard deviation of said posterior. Δlog*R*is the common log of the *R*/*RA* growth rate estimate. Mutation relative growth rate describes the relative growth rate conferred by a mutation compared to no mutation.

Calculation of mutational spectra and proportions of G > T mutations

Mutational spectra were calculated as reported previously[25]. In brief, all mutations downstream of the corresponding lineage root node are identified. The contexts of these mutations are calculated in the genomic sequence at the start of the corresponding phylogenetic branch, taking into account mutations that have arisen on ancestral branches in the phylogenetic tree. Mutational spectra were rescaled by the genomic composition in the lineage root ancestor as described previously[25]. Confidence intervals on the proportion of G > T mutations were calculated using a Wilson score interval incorporating the calculated proportion and the number of sampled mutations.

Software versions

Package versions used for the simulations were: Snakemake (v.7.8.5)[41], PhastSim (v.0.0.4)[42], ART (v.2016.06.05)[44], Badread git commit (c2bdcbe)[45], ARTIC Illumina workflow git commit (8af5152) from https://github.com/connor-lab/ncov2019-artic-nf, Epi2me wf-artic git commit (218aa1d) from https://github.com/epi2me-labs/wf-artic, CTE git commit (9cd94b8) from https://github.com/iqbal-lab-org/covid-truth-eval, Nextflow (v.21.04.3)[55], bwa git commit (c77ace7)[43], (htslib v1.14)[56], SAMtools (v.1.14)[36], BEDTools (v.2.30.0)[57], joblib (v1.1.0) from https://github.com/joblib/joblib, numpy (v.1.22.1)[58], pandas (v1.4.0)[59], pysam (v.0.18.0) at https://github.com/pysam-developers/pysam and tqdm (v.4.62.3) from https://github.com/tqdm/tqdm.

The ARTIC-ILM pipeline used was git commit (8af5152) from https://github.com/connor-lab/ncov2019-artic-nf. The ARTIC-ONT pipeline used was git commit (218aa1d) from https://github.com/epi2me-labs/wf-artic. Version 4.3 of Pangolin and v.1.21 of Pangolin-data were used for the intersection dataset. Version 1.29 of Pangolin-data was used on the final Viridian global tree. MAPLE (v.0.7.2) was used to measure uncertainty in the global tree.

Viridian (v.1.0.0 or v.1.1.0) was used to process all runs. The only difference between these versions is v.1.1.0 added support for unpaired Illumina reads. The versions of tools used by Viridian were: Cylon git commit (57d559a), minimap2 git commit (b0b199f), MUMmer (v.4.0.0rc1), NGMerge git commit (224fc6a), Racon git commit (a2cfcac), Varifier git commit (8bc8726). Ushonium git commit (b024320) was used, with dependencies MAFFT (v.7.520), UShER git commit (2df81ee) and taxoniumtools (v.2.0.111).

Reporting summary

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

Data availability

The global Viridian tree is hosted at https://viridian.taxonium.org. All other additional files are available from figshare and some are provided as supplementary tables and files. Supplementary data file on figshare at https://doi.org/10.6084/m9.figshare.30453716.v1 (ref. 60), which is a TSV file containing metadata of all 5,959,032 sequencing runs considered in this study. Supplementary Tables 1–14 in one xlsx file on figshare at https://doi.org/10.6084/m9.figshare.28987784.v2 (ref. 61). Supplementary Table 1. Summary of counts of amplicon schemes in INSDC metadata and the scheme called by Viridian. Supplementary Table 2. Accuracy of Viridian, ARTIC-ILM and ARTIC-ONT on simulated data. Supplementary Table 3. Accuracy of Viridian, ARTIC-ILM and ARTIC-ONT on Illumina truth dataset. Supplementary Table 4. Accuracy of Viridian, ARTIC-ILM and ARTIC-ONT on Nanopore truth dataset. Supplementary Table 5. Run times and RAM usage on the truth dataset. Supplementary Table 6. Metadata for the African dataset. Supplementary Table 7. Counts of sites with errors in the African dataset. Supplementary Table 8. Confidence of nodes in the global Viridian tree. Supplementary Table 9. Numbers of inferred viral introductions. Supplementary Table 10. mpox data. Supplementary Table 11. Country counts in the Viridian global tree, and number of new samples since the tree was built. Supplementary Table 12. Viridian amplicon scheme scores using simulated data. Supplementary Table 13. Viridian amplicon scheme scores on the truth dataset. Supplementary Table 14. Positions masked when building the global Viridian tree. Supplementary HTML file on figshare at https://doi.org/10.6084/m9.figshare.25713198 (ref. 62) comparison of Viridian and GenBank assemblies. All Viridian consensus sequences that are in the global tree, split over two tar archive files on figshare (https://doi.org/10.6084/m9.figshare.25713225 (ref. 63) and https://doi.org/10.6084/m9.figshare.27194637 (ref. 64)), which contain the sequences split over multiple xzipped FASTA files. These are the same batched FASTA files used when building the trees. The Viridian global tree of 4,471,579 sequences, in JSONL and `.pb` format on figshare at https://doi.org/10.6084/m9.figshare.27194547 (ref. 65). The GenBank and Viridian intersection trees in JSONL and `.pb` format on figshare at https://doi.org/10.6084/m9.figshare.25713285 (ref. 66). All other Viridian consensus sequences that are not in the global tree, split over two xzipped FASTA files on figshare at https://doi.org/10.6084/m9.figshare.25713342 and https://doi.org/10.6084/m9.figshare.27194652 (refs. 67,68). The output TSV file from Maple/

SPRTA run on the global Viridian tree on figshare at https://doi.org/10.6084/m9.figshare.28985573.v1 (ref. 69). The Viridian global tree with Maple/SPRTA data added in JSONL format on figshare at https://doi.org/10.6084/m9.figshare.29097608 (ref. 70).

Code availability

Viridian is freely available under the MIT license at https://github.com/iqbal-lab-org/viridian, and code v.1.5.1 is archived at https://doi.org/10.5281/zenodo.17257005 (ref. 71). Code used for analysis and to generate figures is available at https://github.com/martinghunt/viridian-paperand is archived at https://doi.org/10.5281/zenodo.17279235 (ref. 72).

References

34. Tyson, J. R. et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using Nanopore. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.04.283077v1 (2020).
35. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
36. Li, H. et al. The Sequence Alignment Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
37. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
38. Gaspar, J. M. NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinform.* **19**, 536 (2018).
39. Marçais, G. et al. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
40. Hunt, M. et al. Minos: variant adjudication and joint genotyping of cohorts of bacterial genomes. *Genome Biol.* **23**, 147 (2022).
41. Köster, J. & Rahmann, S. Snakemake: a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
42. De Maio, N. et al. phastSim: efficient simulation of sequence evolution for pandemic-scale datasets. *PLoS Comput. Biol.* **18**, e1010056 (2022).
43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
44. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
45. Wick, R. Badread: simulation of error-prone long reads. *J. Open Source Softw.* **4**, 1316 (2019).
46. Constantinides, B. et al. A diverse reference set of cultured sars-cov-2 genomes sequenced using various amplification methods and instrument platforms. *BioStudies* https://doi.org/10.6019/S-BSST1334 (2024).
47. Nutalai, R. et al. Potent cross-reactive antibodies following Omicron breakthrough in vaccinees. *Cell* **185**, 2116–2131.e18 (2022).
48. Greninger, A. L. et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time Nanopore sequencing analysis. *Genome Med.* **7**, 99 (2015).
49. Kafetzopoulou, L. E. et al. Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Euro. Surveill.* **23**, 1800228 (2018).
50. Lewandowski, K. et al. Metagenomic Nanopore sequencing of influenza virus direct from clinical respiratory samples. *J. Clin. Microbiol.* **58**, e00963–19 (2019).
51. Baker, D. J. et al. CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes. *Genome Med.* **13**, 21 (2021).
52. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**, 464–469 (2012).
53. Sanderson, T. Taxonium, a web-based tool for exploring large phylogenetic trees. *eLife* **11**, e82392 (2022).
54. Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* (eds. Munro, H. N.) 21–132 (Elsevier, 1969).
55. Tommaso, P. D. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
56. Bonfield, J. K. et al. HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience* **10**, giab007 (2021).
57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
58. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
59. McKinney, W. et al. Data structures for statistical computing in Python. In *Proc. 9th Python in Science Conference* Vol. 445, 51–56 (SciPy, 2010).
60. Hunt, M., Hinrichs, A. S. & Iqbal, Z. Viridian sequencing run metadata. *Figshare* https://doi.org/10.6084/m9.figshare.30453716 (2025).
61. Hunt, M. et al. Supplementary tables 1–14. *Figshare* https://doi.org/10.6084/m9.figshare.28987784.v2 (2025).
62. Dearlove, B., Endler, L. & Bergthaler, A. Supplementary html file. *Figshare* https://doi.org/10.6084/m9.figshare.25713198 (2025).
63. Hunt, M. Viridian consensus sequences in the global tree, batch 1 of 2. *Figshare* https://doi.org/10.6084/m9.figshare.25713225 (2025).
64. Hunt, M. Viridian consensus sequences in the global tree, batch 2 of 2. *Figshare* https://doi.org/10.6084/m9.figshare.27194637 (2025).
65. Hunt, M., Hinrichs, A. S., Corbett-Detig, R. & Iqbal, Z. Global tree of Viridian consensus sequences. *Figshare* https://doi.org/10.6084/m9.figshare.27194547 (2025).
66. Hunt, M., Hinrichs, A. S., Corbett-Detig, R. & Iqbal, Z. GenBank and Viridian intersection dataset trees. *Figshare* https://doi.org/10.6084/m9.figshare.25713285 (2025).
67. Hunt, M. Viridian consensus sequences not in the global tree, batch 1 of 2. *Figshare* https://doi.org/10.6084/m9.figshare.25713342 (2025).
68. Hunt, M. Viridian consensus sequences not in the global tree, batch 2 of 2. *Figshare* https://doi.org/10.6084/m9.figshare.27194652 (2025).
69. De Maio, N. Output tsv file from maple/sprta run on global Viridian tree. *Figshare* https://doi.org/10.6084/m9.figshare.28985573.v1 (2025).
70. De Maio, N. Viridian global tree with maple/sprta data added. *Figshare* https://doi.org/10.6084/m9.figshare.29097608 (2025).
71. Hunt, M., Knaggs, J., Constantinides, B., Westhead, J. & Iqbal, Z. iqbal-lab-org/viridian (v.1.5.1). *Zenodo* https://doi.org/10.5281/zenodo.17257005 (2025).
72. Hunt, M., Karim, L. & Dearlove, B. iqbal-lab-org/viridian-paper (2025-11-06). *Zenodo* https://doi.org/10.5281/zenodo.17279235 (2025).

Acknowledgements

We thank E. Hodcroft, R. Neher, D. MacCannell, N. Goldman, T. Bedford, R. Colquhoun, A. Rambaut, R. Davies and R. Lanfear for discussions and advice. We also thank Z. Waheed, N. Rahman and K. Gueye for help with ENA submissions. We thank the Global Pathogen Analysis Service team in Oxford, who used an early version (v.0.3.7) of Viridian in production for over a year, providing valuable feedback. We thank the microbiology laboratory staff of the John Radcliffe Hospital, Oxford University Hospitals NHS Trust, for providing assistance with sample processing. We thank the IMSSC2 Laboratory Network Consortium members at the Robert Koch Institute for providing raw data sequences, the Sequencing Core

Facility of the Genome Competence Center (MF1) at the Robert Koch Institute for providing excellent sequencing services, and we thank all laboratories contributing to the German SARS-CoV-2 surveillance. We acknowledge high-performance computing services provided by the Robert Koch Institute. We also thank PHA4GE (Public Health Alliance for Genomic Epidemiology; https://pha4ge.org) for bringing some of us together. We gratefully acknowledge all data contributors (the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative). Finally, we thank N. Goldman, who gave us the idea for the start of the introduction. We acknowledge the following funding. 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; M.H. and J.K.) 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 NIHR, the Department of Health or the UKHSA. We acknowledge grants R35GM128932 (R.C.D.), T32HG012344 (L.K.) and U24HG002371 and BAA 200-2021-11554 (both A.H.). We also acknowledge Academy of Finland grant 336490, VEO—European Union's Horizon 2020 (grant number 874735), Finnish Institute for Health and Welfare (THL), the Jane and Aatos Erkko Foundation (all to R.K.). We acknowledge Wellcome (222574 and 203135), Francis Crick Institute (UKRI, C2112; Wellcome CC2112; Cancer Research UK CC2112) and support in part from the Biomedical Research Centre of Imperial College NHS Trust (all to R.W.). We acknowledge the National Institutes of Health [5P20GM103443-20] to E.Z.G. We acknowledge the Bundesministerium für Bildung und Forschung (German Ministry for Science and Education) and the following grants: IMS-RKI, IMS-NRZ/KL and EU4Health (IMS-HERA1, agreement numbe: ECDC/HERA/2021/008 ECD.12222 and IMS-HERA2, project number 101113012), which funded M.L. and M. Huska and the Integrated Molecular Surveillance for SARS-CoV-2 (IMSSC2) Laboratory Network. We acknowledge the Fondation Botnar Research Award (program grant no. 6063) and the UK Cystic Fibrosis Trust (Innovation Hub Award 001) to C.R. Vicerrectoría de Investigación, Universidad de Costa Rica, grant no. C0196 Project: Protocolo bioinformático y de inteligencia artificial para el apoyo de la vigilancia epidemiológica basada en laboratorio del virus SARS-CoV-2 mediante la identificación de patrones genómicos y clínico-demográficos en Costa Rica a JAMM. We acknowledge the Wellcome Trust Sir Henry Wellcome Postdoctoral Fellowship (210918/Z/18/Z to T.S.). We acknowledge the Chinese Academy of Medical Sciences Innovation Fund for Medical Science, China (grant number 2018-I2M-2-002) to G.S. and J.M. We also acknowledge Schmidt Futures, the Red Avenue Foundation and the Oak Foundation. We acknowledge the charitable donation of L. Ellison to Oxford University toward pathogen surveillance, which supported the work of D. Crook and P. Fowler.

Author contributions

M. Hunt wrote the final implementation of Viridian, developed the primer scheme identification system, assembled the genomes, developed the pipeline for tree-building and performed all analyses not listed below. A.H. analyzed and quality controlled the phylogenies and wrangled metadata. D.A. developed the simulation framework and performed analyses. L.K. performed the reversion analyses and geographical/introduction analysis. B.D. analyzed the assemblies, Ns, indels, and Pango assignments. J.K. contributed to the first implementation of Viridian and initial exploratory work. B.K. performed the PyR_0 analysis, with supervision from J.L., P.S., W.D., C.L., J.M. and G.R.S. isolated and cultured virus stock used to construct the empirical truth set. H.W., G.R., T.S. and S.L. sequenced the empirical truth set, and B.C. analyzed and quality controlled the data. P.W.F. carried out independent testing on different simulations and empirical data. M. Hunt, J.K., B.C., T.P. and D.C analyzed the empirical truth set results. T.S. integrated the phylogeny into taxonium. N.d.M carried out extensive QC analyses of the genomes. C.R. performed the mutation spectrum analysis. H.T., S.E.J. and T.d.O. collated the 'early Omicron' dataset. All other authors collected samples, sequenced genomes and shared data with the archives. Z.I. conceived of the project. Z.I. and R.C-D. supervised the project. M. Hunt, A.H., L.K., D.A., T.S., R.C.-D. and Z.I. wrote the paper. All authors reviewed the manuscript.

Funding

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

Competing interests

G. Screaton is on the GSK Vaccines Scientific Advisory Board, consults for AstraZeneca, and is a founding member of RQ Biotechnology. P. Fowler, D. Crook and Z. Iqbal have consulted for the Ellison Institute of Technology. B. Jolly is employed by Karkinos Healthcare Private Limited. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-025-02947-1.

Correspondence and requests for materials should be addressed to Zamin Iqbal.

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

Reprints and permissions information is available at www.nature.com/reprints.

nature portfolio

Corresponding author(s): Zamin Iqbal

Last updated by author(s): 10/21/2025

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	We downloaded all tiled amplicon SARS-CoV-2 sequence data from the ENA/SRA as of June 2024.
Data analysis	Package versions used for the simulations were: Snakemake v7.8.5, PhastSim v0.0.4, ART v2016.06.05, Badread git commit c2bdcbe, ARTIC Illumina workflow git commit 8af5152 from https://github.com/connor-lab/ncov2019-artic-nf, Epi2me wf-artic git commit 218aa1d from https://github.com/epi2me-labs/wf-artic, CTE git commit 9cd94b8 from https://github.com/iqbal-lab-org/covid-truth-eval, Nextflow v21.04.3, bwa git commit c77ace7, htslib v1.14, samtools v1.14, BEDTools v2.30.0, joblib v1.1.0 from https://github.com/joblib/joblib, numpy v1.22.1, pandas v1.4.0, pysam v0.18.0 at https://github.com/pysam-developers/pysam tq,dm v4.62.3 from https://github.com/tqdm/tqdm. Version 4.3 of Pangolin, and version 1.21 of pangolin-data were used for the intersection dataset. Version 1.29 of pangolin-data was used on the final Viridian global tree. MAPLE version 0.7.2 was used to measure uncertainty in the global tree. Viridian v1.0.0 or v1.1.0 was used to process all runs. The only difference between these versions is v1.1.0 added support for unpaired Illumina reads. The versions of tools used by Viridian were: Cylon git commit 57d559a, minimap2 git commit b0b199f, MUMmer v4.0.0rc1, NGMerge git commit 224fc6a, Racon git commit a2cfcac, Varifier git commit 8bc8726. Ushonium git commit b024320 was used, with dependencies MAFFT v7.520, UShER git commit 2df81ee, and taxoniumtools v2.0.111. We also ran version 3.10 of pyR0 from https: //github.com/broadinstitute/pyro-cov?tab=readme-ov-file.

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

Supplementary text and figures S1-9 are in the supplementary PDF file.

The global Viridian tree is hosted at https://viridian.taxonium.org.

All other additional files are available from Figshare:
• Supplementary table S1[60], https://doi.org/10.6084/m9.figshare.27195261 - this is a TSV file containing metadata of all 5,959,032 sequencing runs considered in this study
• Supplementary tables S2-15 in one xlsx file[61], https://doi.org/10.6084/m9.figshare.28987784, details below:

S2 - Summary of counts of amplicon schemes in INSDC metadata and the scheme called by Viridian
S3 - Accuracy of Viridian, ARTIC-ILM and ARTIC-ONT on simulated data
S4 - Accuracy of Viridian, ARTIC-ILM and ARTIC-ONT on Illumina truth data set
S5 - Accuracy of Viridian, ARTIC-ILM and ARTIC-ONT on Nanopore truth data set
S6 - Run times and RAM usage on the truth data set
S7 - Metadata for the African data set
S8 - Counts of sites with errors in the African data set
S9 - Confidence of nodes in the global viridian tree
S10 - Numbers of inferred viral introductions
S11 - mpox data
S12 - Country counts in the Viridian global tree, and number of new samples since the tree was built
S13 - Viridian amplicon scheme scores using simulated data
S14 - Viridian amplicon scheme scores on the truth data set
S15 - Positions masked when building the global Viridian tree

• Supplementary HTML file, https://doi.org/10.6084/m9.figshare.25713198-comparison of Viridian and GenBank assemblies
• All Viridian consensus sequences that are in the global tree, split over two tar archive files (https://doi.org/10.6084/m9.figshare.25713225, https://doi.org/10.6084/m9.figshare.27194637), which contain the sequences split over multiple xzipped FASTA files. These are the same batched FASTA files used when building the trees.
• The Viridian global tree of 4,471,579 sequences, in JSONL and .pb format[65], https://doi.org/10.6084/m9.figshare.27194547
• The GenBank and Viridian intersection trees in JSONL and .pb format[66], https://doi.org/10.6084/m9.figshare.25713285
• All other Viridian consensus sequences that are not in the global tree, split over two xzipped FASTA files - https://doi.org/10.6084/m9.figshare.25713342, https://doi.org/10.6084/m9.figshare.27194652.
• The output TSV file from Maple/SPRTA run on the global Viridian tree[69], https://doi.org/10.6084/m9.figshare.28985573.v1
• The Viridian global tree with Maple/SPRTA data added, in JSONL format[70], https://doi.org/10.6084/m9.figshare.29097608

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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 analyse all public tiled-amplicon genome data for SARS-CoV-2, about 6 million genomes.
Data exclusions	None (except for our criteria that we only analyse tiled amplicon data)
Replication	We measure the error rates of different software processes (amplicon assembly by Viridian and state-of-the-art alternatives), first on simulated data (n=8000), then a small manually curated truth set of 67 genomes, then a larger independent dataset (Early Omicron data from Africa, n=12,287), and then via various indirect methods (number of reversions in the tree, comparison of amplicon scheme identification with metadata) on further independent datasets (all other sequenced SARS-CoV-2). In doing so we take care to consider key covariates (sequencing technology, primer scheme version, and also Pango/lineage). This is in a sense replication of those error estimates. But beyond that, this is not a paper making claims that A causes B, or is predictive of B, so there is no replication in the normal sense of for example GWAS.
Randomization	N/A. There is no randomization because we are not selecting from a population, or assigning to different groups, or comparing groups, or doing experiments.
Blinding	There was no blinding because there was no group allocation. This is not a study where we compare groups.

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

Methods

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