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

Martin Hunt[1-4], Angie S. Hinrichs[5], Daniel Anderson[1], Lily Karim[5,6], Bethany L Dearlove[7], Jeff Knaggs[1-4], 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], Theo Sanderson[9], Christopher Ruis[10,11], Nicola de Maio[1], Lucas N. Amenga-Etego[12], Dominic S. Y. Amuzu[12], Martin Avaro[13], Gordon A. Awandare[12], Reuben Ayivor-Djanie[14,15], Matthew Bashton[16], Elizabeth M Batty[17,18], Yaw Bediako[12], Denise De Belder[19], Estefania Benedetti[13], Andreas Bergthaler[7], Stefan A. Boers[20], Josefina Campos[19], Rosina Afua Ampomah Carr[15,21], Facundo Cuba[19], Maria Elena Dattero[13], Wanwisa Dejnirattisai[22], Alexander Dilthey[23], Kwabena Obeng Duedu[15,24], Lukas Endler[7], Ilka Engelmann[25], Ngiambudulu M. Francisco[26], Jonas Fuchs[27], Etienne Z. Gnimpieba[28], Soraya Groc[29], Jones Gyamfi[15,30], Dennis Heemskerk[20], Torsten Houwaart[23], Nei-yuan Hsiao[31], Matthew Huska[32], Martin Hölzer[32], Arash Iranzadeh[33], Hanna Jarva[34], Chandima Jeewandara[35], Bani Jolly[36,37], Rageema Joseph[33], Ravi Kant[38,39,40], Karrie Ko Kwan Ki[41], Satu Kurkela[34], Maija Lappalainen[34], Marie Lataretu[32], Chang Liu[42,43], Gathsaurie Neelika Malavige[35], Tapfumanei Mashe[44], Juthathip Mongkolsapaya[18,42,43], Brigitte Montes[29], Jose Arturo Molina Mora[45], Collins M. Morang'a[12], Bernard Mvula[46], Niranjan Nagarajan[47,48], Andrew Nelson[49], Joyce M. Ngoi[12], Joana Paula da Paixão[26], Marcus Panning[27], Tomas Poklepovich[19], Peter K. Quashie[12], Diyanath Ranasinghe[35], Mara Russo[13], James Emmanuel San[50,51], Nicholas D. Sanderson[2,3], Vinod Scaria[37,52], Gavin Screaton[2], Tarja Sironen[38,39], Abay Sisay[53], Darren Smith[16], Teemu Smura[38,39], Piyada Supasa[42,43], Chayaporn Suphavilai[47], Jeremy Swann[2], Houriiyah Tegally[54], Bryan Tegomoh[55,56,57], Olli Vapalahti[38,39], Andreas Walker[58], Robert J Wilkinson[9,59,60], Carolyn Williamson[33], IMSSC2 Laboratory Network Consortium[61], Tulio de Oliveira[54,62], Timothy EA Peto[2], Derrick Crook[2], Russell Corbett-Detig[5,6], and Zamin Iqbal[1,63]

[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
[6]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA
[7]Institute for Hygiene and Applied Immunology, Center for Pathophysiology, Infectiology and Immunology, Medical University of Vienna, Vienna 1090, 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]West African Centre for Cell Biology of Infectious Pathogens (WACCBIP), University of Ghana, Accra, Ghana
[13]Servicio de Virus Respiratorios, Instituto Nacional Enfermedades Infecciosas, ANLIS "Dr. Carlos G. Malbrán", Buenos Aires, Argentina
[14]Laboratory for Medical Biotechnology and Biomanufacturing, International Centre for Genetic

Engineering and Biotechnology, Tristie, Italy
[15]Department of Biomedical Sciences, University of Health and Allied Sciences, Ho, Ghana
[16]The Hub for Biotechnology in the Built Environment, Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, NE1 8ST, UK
[17]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK
[18]Mahidol-Oxford Tropical Medicine Research Unit, Bangkok, Thailand
[19]Unidad Operativa Centro Nacional de Genómica y Bioinformática, ANLIS “Dr. Carlos G. Malbrán”, Buenos Aires, Argentina
[20]Dept. Medical Microbiology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA, Leiden, The Netherlands
[21]Department of Computational Medicine and Bioinformatics, University of Michigan, Michigan, Ann Arbor, MI, USA
[22]Division of Emerging Infectious Disease, Research Department, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkoknoi, Bangkok 10700, Thailand
[23]Institute of Medical Microbiology and Hospital Hygiene, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[24]College of Life Sciences, Birmingham City University, Birmingham, UK
[25]Pathogenesis and Control of Chronic and Emerging Infections, Univ Montpellier, INSERM, Etablissement Français du Sang, Virology Laboratory, CHU Montpellier, Montpellier, France
[26]Grupo de Investigação Microbiana e Imunológica, Instituto Nacional de Investigação em Saúde (National Institute for Health Research), Luanda, Angola
[27]Institute of Virology, Freiburg University Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany
[28]Biomedical Engineering Department, University of South Dakota, Sioux Falls, SD 57107
[29]Virology Laboratory, CHU Montpellier, Montpellier, France
[30]School of Health and Life Sciences, Teesside University, Middlesbrough, UK
[31]Divison of Medical Virology, University of Cape Town and National Health Laboratory Service
[32]Genome Competence Center (MF1), Robert Koch Institute, Nordufer 20, 13353 Berlin, Germany
[33]Computational Biology Division, University of Cape Town
[34]HUS Diagnostic Center, Clinical Microbiology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
[35]Allergy Immunology and Cell Biology Unit, Department of Immunology and Molecular Medicine, University of Sri Jayewardenepura, Nugegoda, Sri Lanka
[36]Karkinos Healthcare Private Limited (KHPL), Aurbis Business Parks, Bellandur, Bengaluru, Karnataka, 560103, India
[37]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, Uttar Pradesh, India
[38]Department of Veterinary Biosciences, University of Helsinki, 00014 Helsinki, Finland
[39]Department of Virology, University of Helsinki, 00014 Helsinki, Finland
[40]Department of Tropical Parasitology, Institute of Maritime and Tropical Medicine, Medical University of Gdansk, 81-519 Gdynia, Poland
[41]Department of Microbiology, Singapore General Hospital, Singapore
[42]Chinese Academy of Medical Science (CAMS) Oxford Institute (COI), University of Oxford, Oxford, UK
[43]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK
[44]Health System Strengthening Unit, World Health Organisation, Harare, Zimbabwe
[45]Centro de investigación en Enfermedades Tropicales & Facultad de Microbiología, Universidad de Costa Rica, Costa Rica
[46]Public Health Institute of Malawi, Ministry of Health, Malawi
[47]Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore
[48]Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[49]Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, NE1 8ST, UK
[50]Duke Human Vaccine Institute, Duke University, Durham, NC 27710
[51]University of KwaZulu Natal, Durban, South Africa, 4001

[52]Vishwanath Cancer Care Foundation (VCCF), Neelkanth Business Park Kirol Village, West Mumbai, Maharashtra, 400086, India
[53]Department of Medical Laboratory Sciences, College of Health Sciences, Addis Ababa University, P.O.Box 1176, Addis Ababa, Ethiopia
[54]Centre for Epidemic Response and Innovation (CERI), Stellenbosch University, South Africa
[55]Centre de Coordination des Opérations d'Urgences de Santé Publique, Ministere de Sante Publique, Cameroun
[56]University of California, Berkeley, Berkeley, California, USA
[57]Nebraska Department of Health and Human Services, Lincoln, Nebraska, USA
[58]Institute of Virology, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[59]Centre for Infectious Diseases Research in Africa, University of Cape Town
[60]Imperial College London, UK
[61]Consortium - please see supplementary for details
[62]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), University of KwaZulu-Natal, South Africa
[63]Milner Centre for Evolution, University of Bath, UK

Abstract

The SARS-CoV-2 genome occupies a unique place in infection biology – it is the most highly sequenced genome on earth (making up over 20% of public sequencing datasets) with fine scale information on sampling date and geography, and has been subject to unprecedented intense analysis. As a result, these phylogenetic data are an incredibly valuable resource for science and public health. However, the vast majority of the data was sequenced by tiling amplicons across the full genome, with amplicon schemes that changed over the pandemic as mutations in the viral genome interacted with primer binding sites. In combination with the disparate set of genome assembly workflows and lack of consistent quality control (QC) processes, the current genomes have many systematic errors that have evolved with the virus and amplicon schemes. These errors have significant impacts on the phylogeny, and therefore over the last few years, many thousands of hours of researchers time has been spent in "eyeballing" trees, looking for artefacts, and then patching the tree.

Given the huge value of this dataset, we therefore set out to reprocess the complete set of public raw sequence data in a rigorous amplicon-aware manner, and build a cleaner phylogeny. Here we provide a global tree of 3,960,704 samples, built from a consistently assembled set of high quality consensus sequences from all available public data as of March 2023, viewable at `https://viridian.taxonium.org`. Each genome was constructed using a novel assembly tool called Viridian (`https://github.com/iqbal-lab-org/viridian`), developed specifically to process amplicon sequence data, eliminating artefactual errors and mask the genome at low quality positions. We provide simulation and empirical validation of the methodology, and quantify the improvement in the phylogeny.

Phase 2 of our project will address the fact that the data in the public archives is heavily geographically biased towards the Global North. We therefore have contributed new raw data to ENA/SRA from many countries including Ghana, Thailand, Laos, Sri Lanka, India, Argentina and Singapore. We will incorporate these, along with all public raw data submitted between March 2023 and the current day, into an updated set of assemblies, and phylogeny. We hope the tree, consensus sequences and Viridian will be a valuable resource for researchers.

Introduction

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 analysing relatively small datasets (N<5000), usually carefully curated and generally studied by people working closely with those obtaining and processing the clinical samples, or indirectly, via national public health organisations. Data were usually small, 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 (i.e. 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, since there was no unified genome assembly or 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 vast majority of these errors, thereby building a higher quality

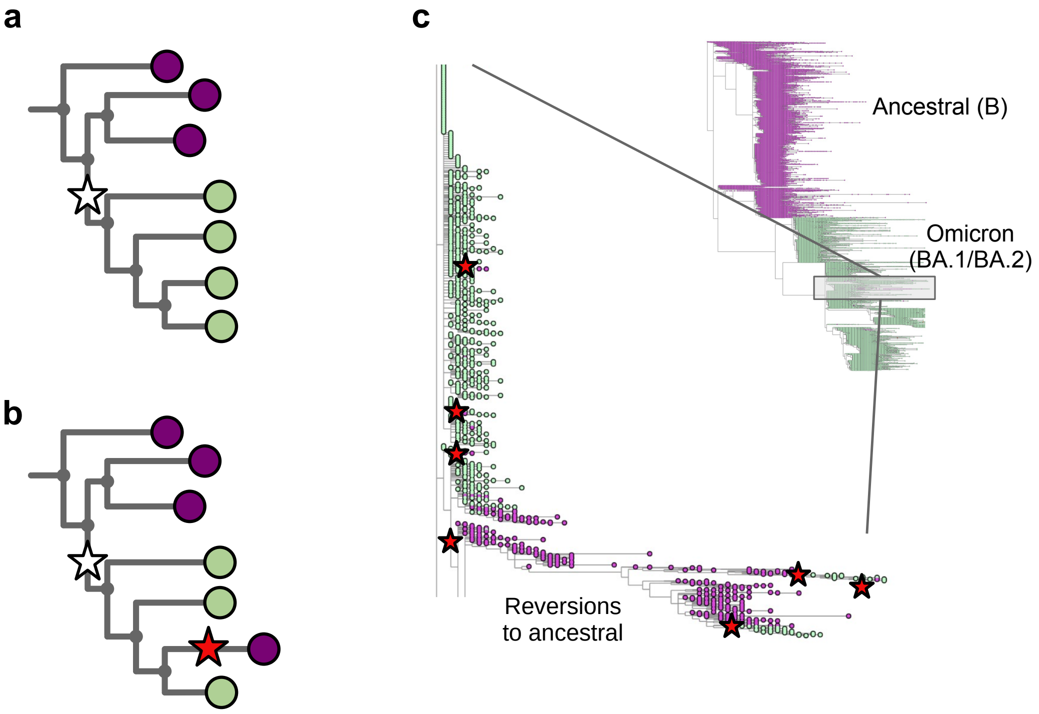

Figure 1: Assemblers which wrongly default to the reference base in the absence of data cause reversions in the phylogeny. a) Cartoon phylogeny built from perfect genomes, with leaves coloured by genotype at a specific position X (purple - ancestral base, green - derived base). Just one mutation at this 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 is 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 colour 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, coloured by genotype at genome position 22813 (spike codon 417). Blow-up shows multiple reversions back to the ancestral purple. A non-exhaustive set of artefactual mutations (reversions, unreversions, re-reversions etc) are shown with red stars, where there is a flip back and forth from green to/from purple.

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 done 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 et al's rapid production of a set of primers, the first "ARTIC" (acronym referring to a consortium) primer scheme [6]. A feature of any tiled amplicon scheme is that, as the virus evolves, eventually mutations within primer-binding 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. Additionally, 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 analysing the phylogeny would find a sudden crop of genomes "reverting to the ancestor".

In Figure 1a we show part of a tree with the leaves coloured 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 Figure 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 Figure 1c we show part of the global SARS-CoV-2 phylogeny hosted at taxonium.org (accessed 9th April 2024), zoomed in to show where Omicron branches from the ancestor. Leaves are coloured 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 are in general more likely to be due to assembly artefact than multiple independent reversions. Such errors can have considerable impact 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 a number of other possible technical artefacts that can arise (e.g. primer dimers [10], interactions between amplicons [17], or primers binding in non-canonical sites [15]) which should be expected and handled, otherwise additional errors will result. Unfortunately, these errors often correlated with individual sequencing centres, 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 [14].

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 (Midnight [18], AmpliSeq (Thermo Fisher Scientific), VarSkip (`https://github.com/nebiolabs/VarSkip`, etc). Each VOC wave brought mutations in primer bindings sites leading to amplicon dropouts, and a subsequent wave of artefacts in genomes as these were mishandled (see Figure 2). New amplicon schemes were then introduced, and gradually taken up, solving previous dropout problems, but also followed by smaller waves of new artefacts 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 if strange phenomena are artefacts or not. Some of us (RCD, AH) have been maintaining the global phylogenetic tree of SARS-CoV-2 since 2021 [19], and the only way we have been able to maintain the integrity of the tree has been to a) completely mask 150 nucleotide positions in the genome, as they are systematically too often wrong to ever be trusted, and b) systematically mask (ie ignore) certain mutations on specific branches of the tree. As artefacts ebbed and flowed, and were discovered by analysts, the masking had to be updated (see Figure 2 and Supplementary Figure S1). After the mammoth global efforts

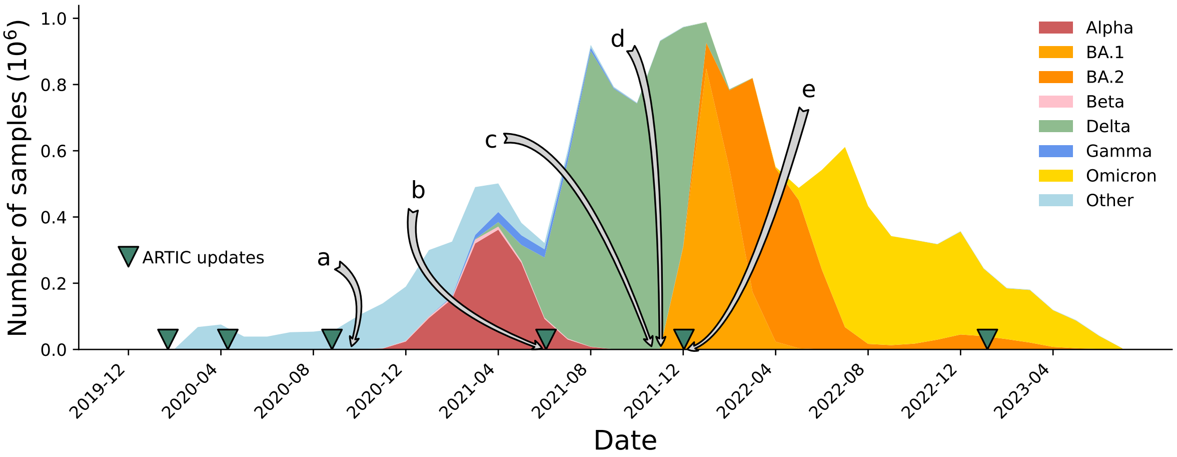

Figure 2: Timeline of the SARS-CoV-2 pandemic from December 2019 to July 2023, with selected events relating to problems with sequencing and consensus calling labelled a-e. Releases of ARTIC primers schemes (versions 1, 2, 3, 4, 4.1, 5.3.2) are marked with green triangles. a) Primer dimers cause amplicon dropouts [10] and 28% of GISAID [11] sequences deposited in September 2020 have at least one gap of length at least 200bp [12]. b) A 9bp deletion in the primer binding region of ARTIC V3 amplicon 73 causes missing data [13]. c) Dropouts causing artefacts at Spike 95 and 142 [14]. d) ARTIC v4 roll out triggers artifactual mutations in some pipelines [15]. e) Omicron samples cause ARTIC v4 amplicon dropout, triggering the update to ARTIC v4.1 [16].

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 this data, providing a firm foundation for future discoveries.

As of March 2023, there were approximately 5.3 million SARS-CoV-2 raw sequence datasets deposited in the SRA/ENA, very few of which had metadata recording the primer scheme and the assembly pipeline used (data from COG-UK being a notable but geographically localised exception). In this paper we will describe our amplicon-aware assembly and QC processes, 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 quality-control 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. Since publicly shared sequence data does 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 amplicon by building a partial-order alignment graph of the reads using Racon [20], an approach which 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 minimising 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) (see 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=12287, "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 (see 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 with the call from Viridian (Supplementary tables 1,2). 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 19/20 cases, the pileups were categorical that Viridian was correct, and in the remaining one, the data were inconclusive (supplementary text and Figures S2-6). Note that both the truth set and the ENA/SRA data contain samples where tagmentation during the library preparation caused fragmented reads, confirming the method worked there too.

Simulations

We simulated a SARS-CoV-2 tree of 8,000 genomes, including SNP errors in primers and amplicon dropouts. Illumina and Nanopore reads were simulated from each genome, from simulated amplicons using the ARTIC v4 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, see 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 Tables S3a-d). Although there were overall very few errors, ARTIC-ONT had notably more indel errors than Viridian (Supplementary Tables S3c,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 (v3, v4, v4.1) 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 S4,5), and Viridian performing better with respect to indels on Nanopore data (Supplementary Table S5e,f). We measured the peak RAM and total CPU time of each truth set run. Viridian had mean peak RAM usage of 444MB and mean CPU time 154s, whereas ARTIC-ILM and ARTIC-ONT used 1.45GB of RAM and took 366s, and 1.80GB of RAM, and took 561s respectively (Supplementary Table S6, Supplementary Figure S7).

African "Early Omicron" dataset

Next we evaluated our own empirical dataset, sequenced and assembled at CERI in South Africa, with samples from November 2021 to March 2022, including Variants of Concern 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), Namibia (N=83) (and 4 with unknown country), and include Illumina (N=9,935) and Nanopore (N=2,352) runs, using either ARTIC (N=11,070 including versions 3,4, and 4.1) or Midnight (N=1,217) amplicon schemes (Supplementary Table S7). Each sample was processed with Viridian and ARTIC-ILM/ARTIC-ONT as appropriate, and the results compared with our original assemblies [21] which have previously been shared to the UShER [22, 23] SARS-CoV-2 phylogeny via 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, see 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 significantly reduced in the ARTIC-ILM workflow although some did remain. By contrast the errors were completely removed by Viridian. This is summarised in Figure 3.

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$^{\text{nd}}$ March 2023, keeping all 3,960,704 that passed QC (see 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 data sets – all 3,960,704 Viridian sequences, Intersection/Viridian (i.e. the Viridian assemblies of the intersection set), and Intersection/GenBank (i.e. the GenBank assemblies of the intersection set) – using MAFFT [24] and UShER (reverting deletions to the ancestral sequence and excluding insertions, see Methods). Note that these trees

a) 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

b) do not have any forcible masking of particular mutations on the branches of specific Variants of Concern, 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 scatterplot 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

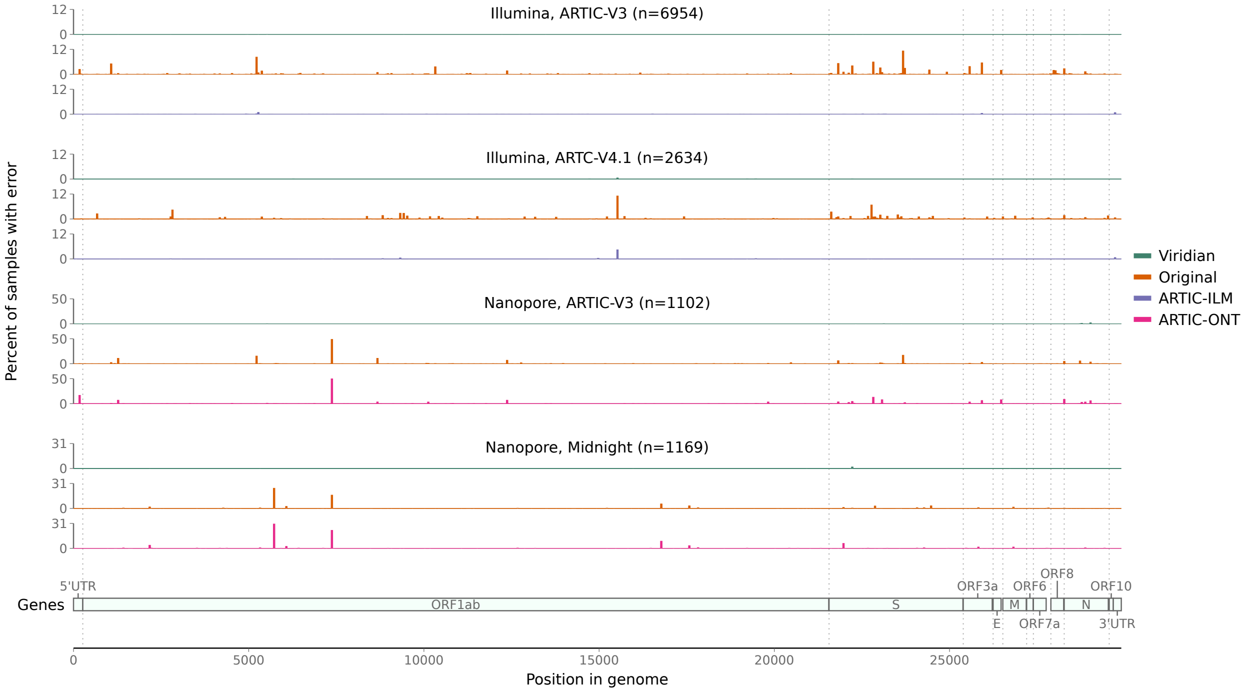

Figure 3: Errors across the genome in consensus sequences from the "Early Omicron" African dataset, split by sequencing technology and amplicon scheme. Plots show the percent of consensus sequences with an error (y-axis), taking the maximum value in windows of length 50bp (x-axis). Error here is defined as where the consensus sequence has an A/C/G/T call, the read depth passes Viridian's default filters (see 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.

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 (i.e. more specific) in 60% of those. Only 0.01% (N=287) mismatched at the variant level. No Viridian assembly was "Unassigned", compared with 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 characterising 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

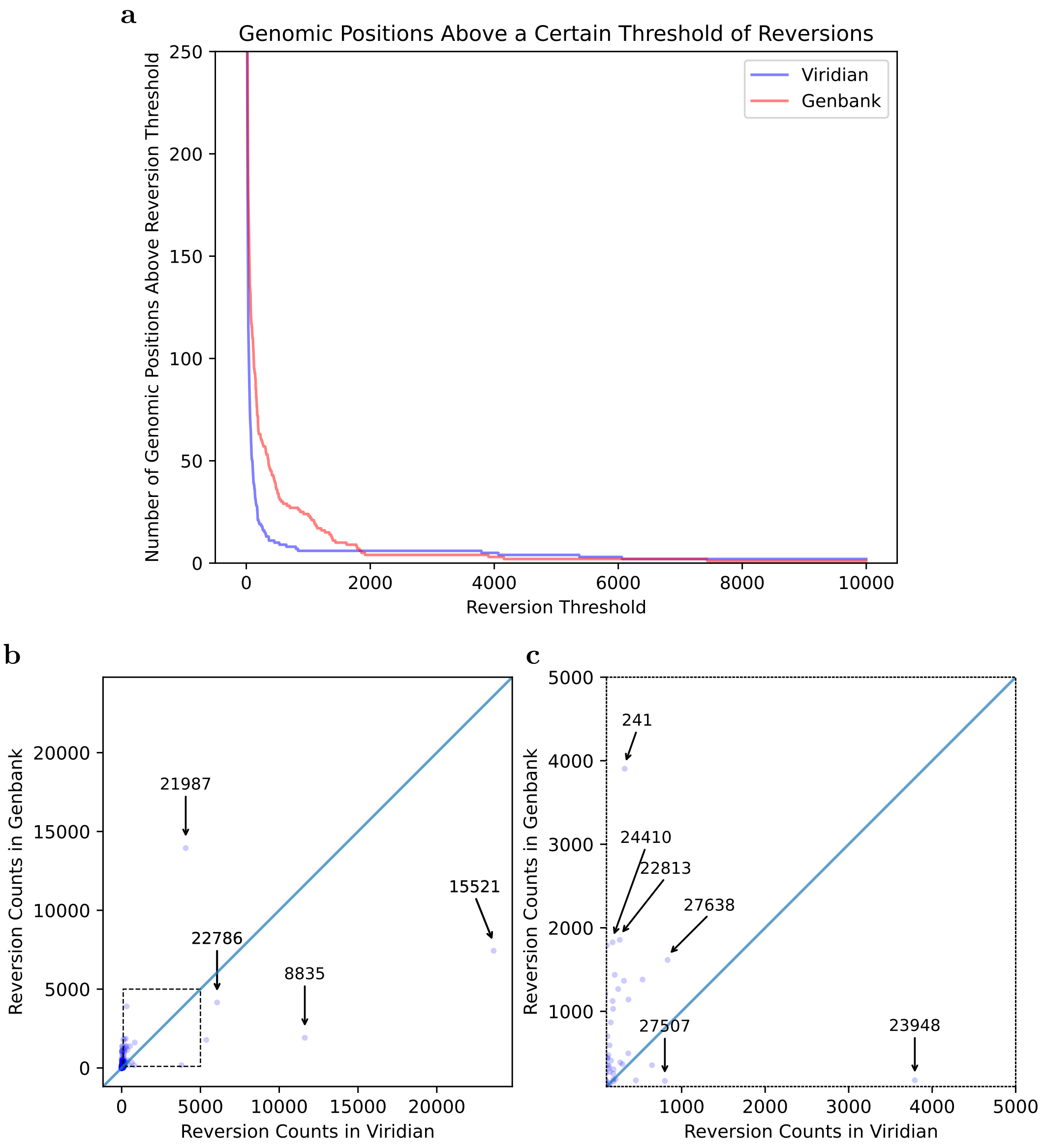

Figure 4: Most variable sites cause fewer reversions in the Viridian tree than the GenBank tree. a) Plot showing how many positions in the genome (y axis) have at least N reversions (x axis) in each tree (Viridian in blue, GenBank in red). Viridian curve drops faster, having fewer positions that create many reversions. b) Scatterplot comparing count of reversion mutations found in GenBank Dataset (y axis) and Viridian dataset (x axis). Note (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 (labelled by genomic coordinates: 22786, 8835, 15521) where Viridian has particularly high numbers of reversions, and one (labelled 21987) for GenBank. c) Blow up of dotted square from panel b) showing vast majority of variable sites in the genome lie above the line y=x.

the GenBank assemblies. We show in Supplementary Figure S8 those positions where there is discordance between Viridian and GenBank.

Reversions

One of the key signals of artefactual 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 [19] and custom scripts to count the number of reversions in both trees, and plot this in two ways. In Figure 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 Figure 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 Figure 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 in order 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. See also Supplementary Figure S9 for the specific example of genome position 22813 (introduced earlier in Figure 1), comparing the current UShER global phylogeny with the Viridian tree.

Final global tree and masking

Our final global tree of the Viridian consensus sequences contains 3,960,704 samples. Tree construction was done, as is normal with UShER, by batching the samples, and then alternating adding a batch to the tree and optimising 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 (see Methods). After constructing the tree, we masked positions in the Problematic Sites set, which includes highly homoplasic sites

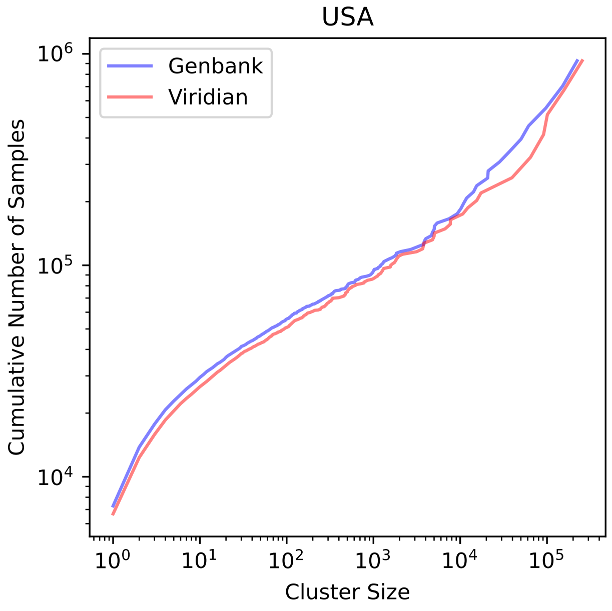

Figure 5: Cumulative Distribution of the number of samples in USA stratified by cluster size.

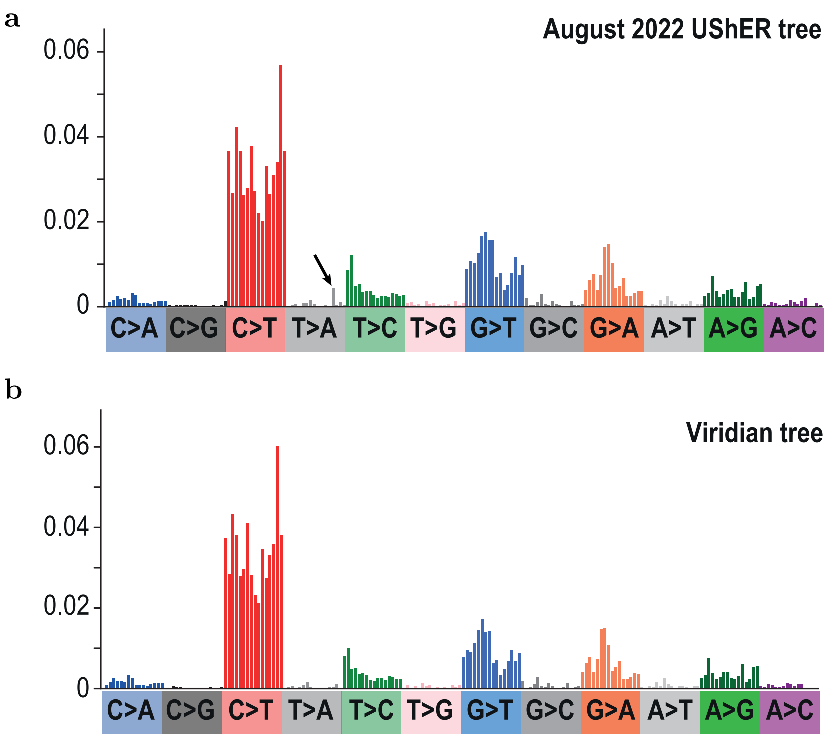

Figure 6: Comparison of Alpha variant mutational spectra calculated using (a) the August 2022 UShER tree [26] and (b) the Viridian tree. Colours show different mutation types (for example C mutating to T, labelled as C>T) and bars show individual surrounding contexts (for example an upstream A and a downstream A). Spectra are rescaled by the availability of the starting nucleotide triplet. The arrow shows a contextual mutation that is unexpectedly elevated in the August 2022 UShER tree; this elevation is not present in the Viridian tree.

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 (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, we ran matOptimize [25] to improve the structure of the tree in the absence of artefactual reversions and highly homoplasic sites.

Impact 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, in order to estimate the effect of the reduced number of sequence/assembly artefacts in the Viridian assemblies on epidemiological analysis, we used geographic metadata for each sample and a pandemic-scale cluster estimation algorithm (matUtils, Cluster-Tracker [27]), to compare the number of inferred unique SARS-CoV-2 viral introductions in each country using the GenBank and Viridian data (Supplementary Table S8). The expectation would be that removing artefactual 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 United States (15,026 versus 13,626 introductions and 7,281 versus 6,676 singleton clusters for GenBank vs Viridian); see Figure 5. As predicted, we see fewer small introductions with Viridian, and at the far right (note log scales) the very largest clusters are slightly larger.

Secondly, we quantified the extent to which the higher quality assemblies would affect estimates of differing mutational spectra of different Variants of Concern [26]. In all cases the spectra were very similar (i.e. the effect was limited), but interestingly in Alpha there had been an odd T>A context (labelled with an arrow in Figure 6a) that was elevated above all others with the August 2022 UShER tree, which was gone in the Viridian data (Figure 6b).

The difference in G>T mutations that had been observed previously between Omicron and non-Omicron is still very much present; see Supplementary Figure S10 – 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 co-evolution 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 March 2023. 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 evolutionary and epidemiological analyses. Second, that Viridian itself will prove useful, providing a significant improvement for Nanopore (and marginal for Illumina) compared with the ARTIC workflow, and a standardised 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 artefact from biology. Viridian will work for tiled amplicon sequencing of non-segmented viruses where a consensus is the desired output (i.e. 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 (eg SARS-CoV-2).

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. Quality control within a single lab 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 standardisation (and adoption) of metadata around sampling,

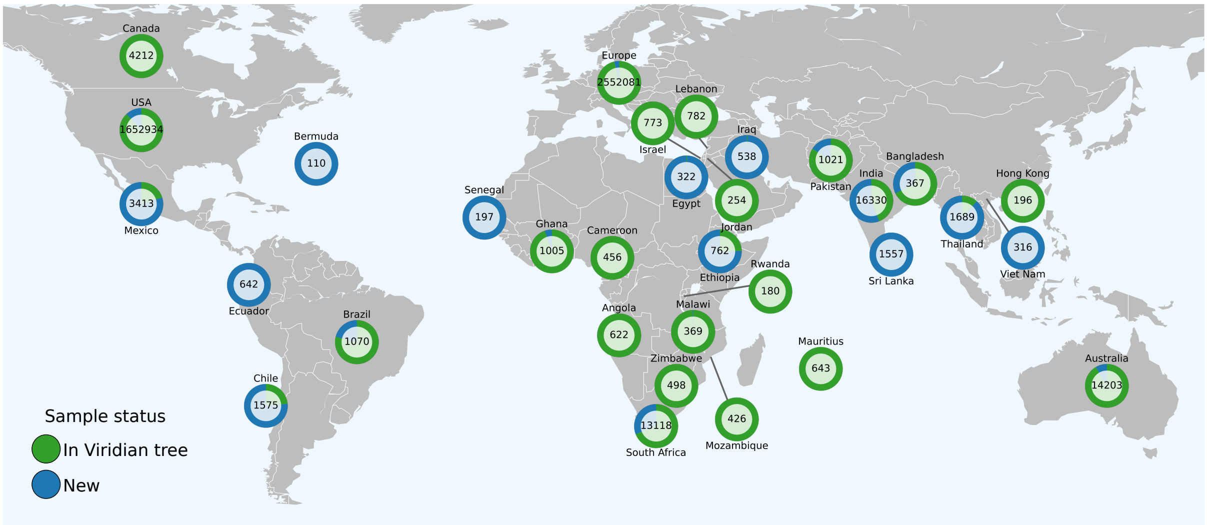

Figure 7: Geographical distribution of samples in this and the next release. Numbers show the total number of samples for each country, excluding QC failures. The proportion coloured green represents samples already included in the global Viridian tree (from our data freeze in March 2023). Blue shows new samples retrieved from the ENA/SRA on 19th March 2024 that are not in the tree, but will be included in the next update to this study. See Supplementary Figure S11 for the per-country counts of Europe. Only countries with a total of at least 100 samples are shown.

assembly and QC, and also multinational "simulations" of pandemics to better prepare for integrating data from different pipelines.

Returning to our project, since the data in the ENA/SRA is heavily biased towards a few high income countries (especially USA and UK), we realised it was important to increase the geographical breadth of our dataset, and reached out to scientists around the globe inviting them to join our collaboration. This preprint represents Phase 1 of this project. Our team has now submitted pre-existing raw sequence data to the ENA/SRA from Austria, Germany, Ghana, India, Netherlands, South Africa and Sri Lanka. Once the final submissions (from Argentina, Laos, and Singapore) have arrived, we will process all data in the ENA/SRA (taking us up to mid 2024) and make a new release (and update this paper). The worldwide distribution of samples is shown in Figure 7, and a breakdown of Europe is in Supplementary Figure S11 (raw data is in Supplementary Table S9).

The pandemic was a global catastrophe with a huge cost in life, and the immense efforts of health professionals on the front lines, public health officials and the (sometimes ad hoc) networks of scientists and bioinformaticians has left many exhausted. However it is also a story of tremendous achievement and solidarity. In doing this work and building this collaboration, it has been striking how everyone has been determined to make the most of this vast resource of SARS-CoV-2 genomic data and build the cleanest and most correct assemblies and phylogeny as possible, to the benefit of us all. It has been a privilege to work together to produce these high quality resources, which was only possible because raw sequence data was deposited in the ENA/SRA.

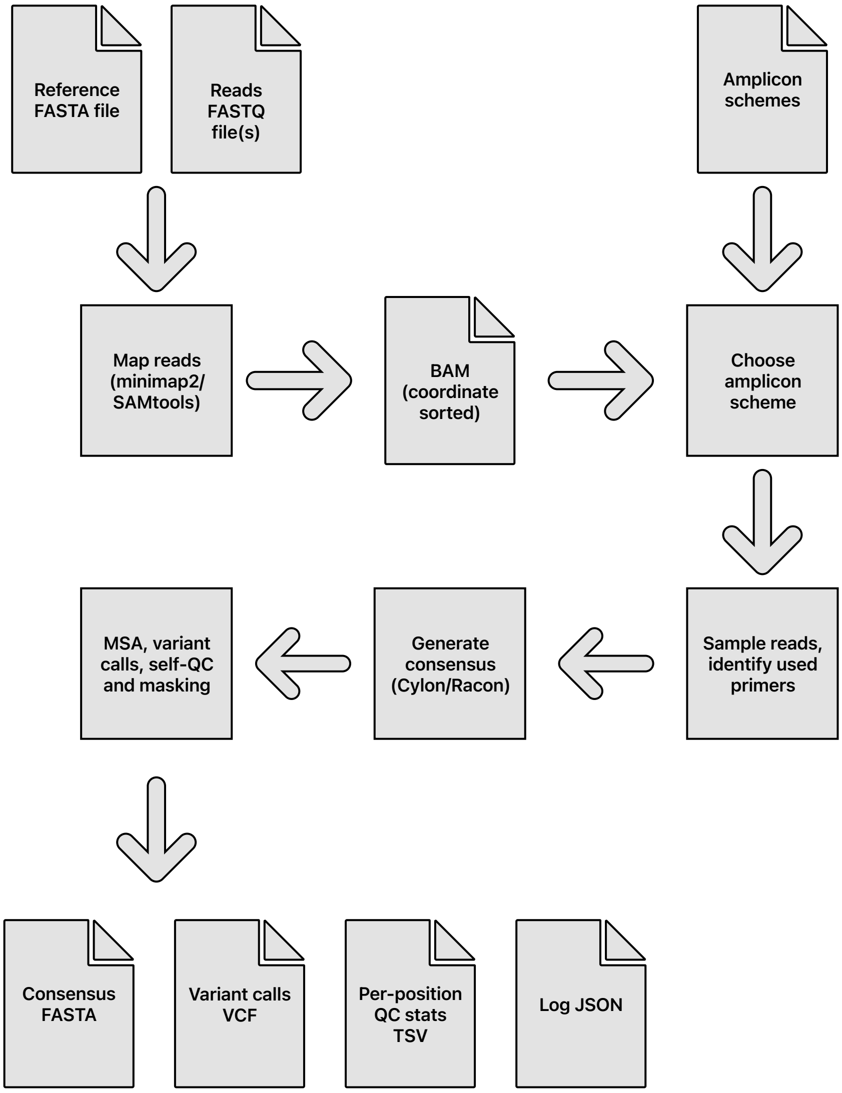
Reference FASTA file
Reads FASTQ file(s)
Amplicon schemes
Map reads (minimap2/ SAMtools)
BAM (coordinate sorted)
Choose amplicon scheme
MSA, variant calls, self-QC and masking
Generate consensus (Cylon/Racon)
Sample reads, identify used primers
Consensus FASTA
Variant calls VCF
Per-position QC stats TSV
Log JSON

Figure 8: Overview of the Viridian pipeline, from input sequencing reads to output files.

Methods

Viridian pipeline

The main stages of the assembly process are: 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, to output a final masked consensus sequence. An overview of the pipeline is shown in Figure 8.

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 v1; ARTIC versions 3, 4.1, 5.3.2_400, 5.2.0_1200 [30]; Midnight 1200 [18]; and VarSkip v1a-2b (`https://github.com/nebiolabs/VarSkip`).

The reads are mapped to the reference genome (default SARS-CoV-2 MN908947.3) using minimap2 [31] with options `-x map-ont` (Nanopore) or `-x sr` (Illumina/Ion Torrent). SAMtools

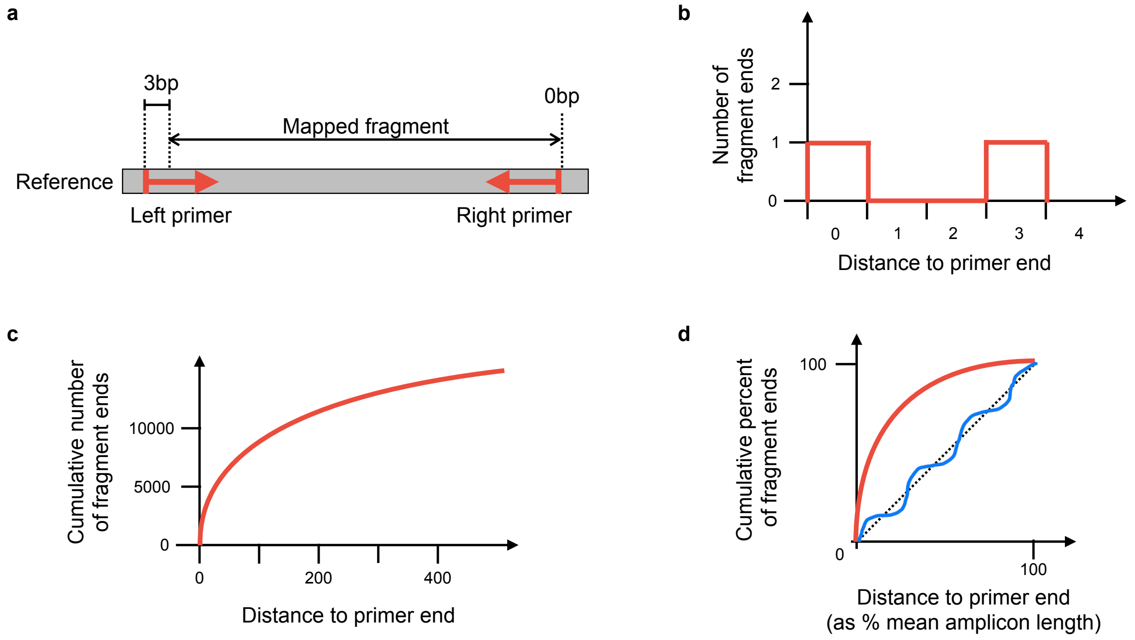

Figure 9: Method to score an amplicon scheme, using mapped fragments. **a)** Example of one mapped fragment, where its left end is 3bp from the start of the primer, and its right end is 0bp from the end of the right primer. **b)** The plot generated from the fragment in a). The right end of the fragment increments the counter for zero distance from a primer, and the left end of the fragment increments the counter for 3bp distance from a primer. The information from all fragments in the sample is added in this way, to make the distribution of distances from nearest primer ends. **c)** The cumulative plot from b) after adding all fragments. **d)** Plot c) is normalised by taking distance to primer end as a percentage of the mean amplicon length (x axis), and fragment counts as percent of total fragments (y axis). The red line indicates a typical curve where the reads match the scheme, whereas the blue line shows a scheme that does not match. The scheme's score is the sum of differences between the calculated line and the $y = x$ line (shown as a dashed line).

[32, 33] 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 20X read depth.

For each amplicon scheme under consideration, a normalised 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, since 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 (Figure 9). 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 (see Figure 9(a)). 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 for right primers, resulting in a count of mapped fragment ends at each distance from a primer (see Figure 9(b,c)).

The distance is normalised by taking the distance as a percent of the mean amplicon length for the scheme, and the count of fragment ends is normalised 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 normalised 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 $y = x$ (Figure 9(d)), but negative values are allowed. The maximum possible score for perfect reads is 5050, because f(i) = 100 for all i and the score is then

$$\sum_{i=0}^{100} (100 - i) = 5050.$$

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. See Figure 9(d) for cartoons of a matching and non-matching scheme, and 10 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 Figure 10 had best score 4290 and second best score 464.

Read sampling. Once the amplicon scheme is known, reads are sampled to a target depth of (by default) 1000X for each amplicon, or using all reads for an amplicon if the mean depth is less than 1000X. 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 5bp 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 1000X (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.

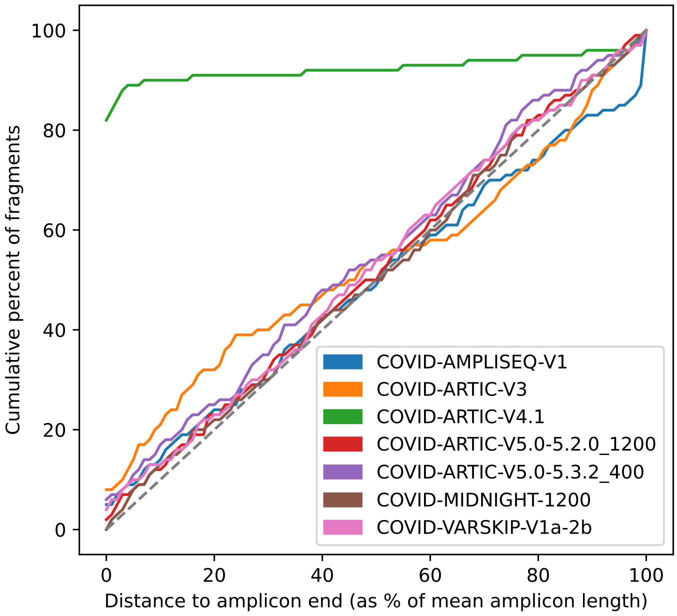

Figure 10: Example scheme identification score plot from Viridian. Made from run accession ERR8959196, which is Nanopore reads sequenced using ARTIC-V4.1 primers.

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 (Figure 11). 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 1000X reads, with a target depth of (by default) 150X for Illumina and 250X for Nanopore or Ion Torrent.

A consensus sequence is generated for each amplicon by iteratively running Racon [20] until no more corrections are made, up to a maximum of 10 runs. If the input reads are paired, then each read pair is merged where possible using NGMerge [34] 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 5X depth are masked with Ns. If the resulting sequence is shorter than 30bp 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 10bp 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

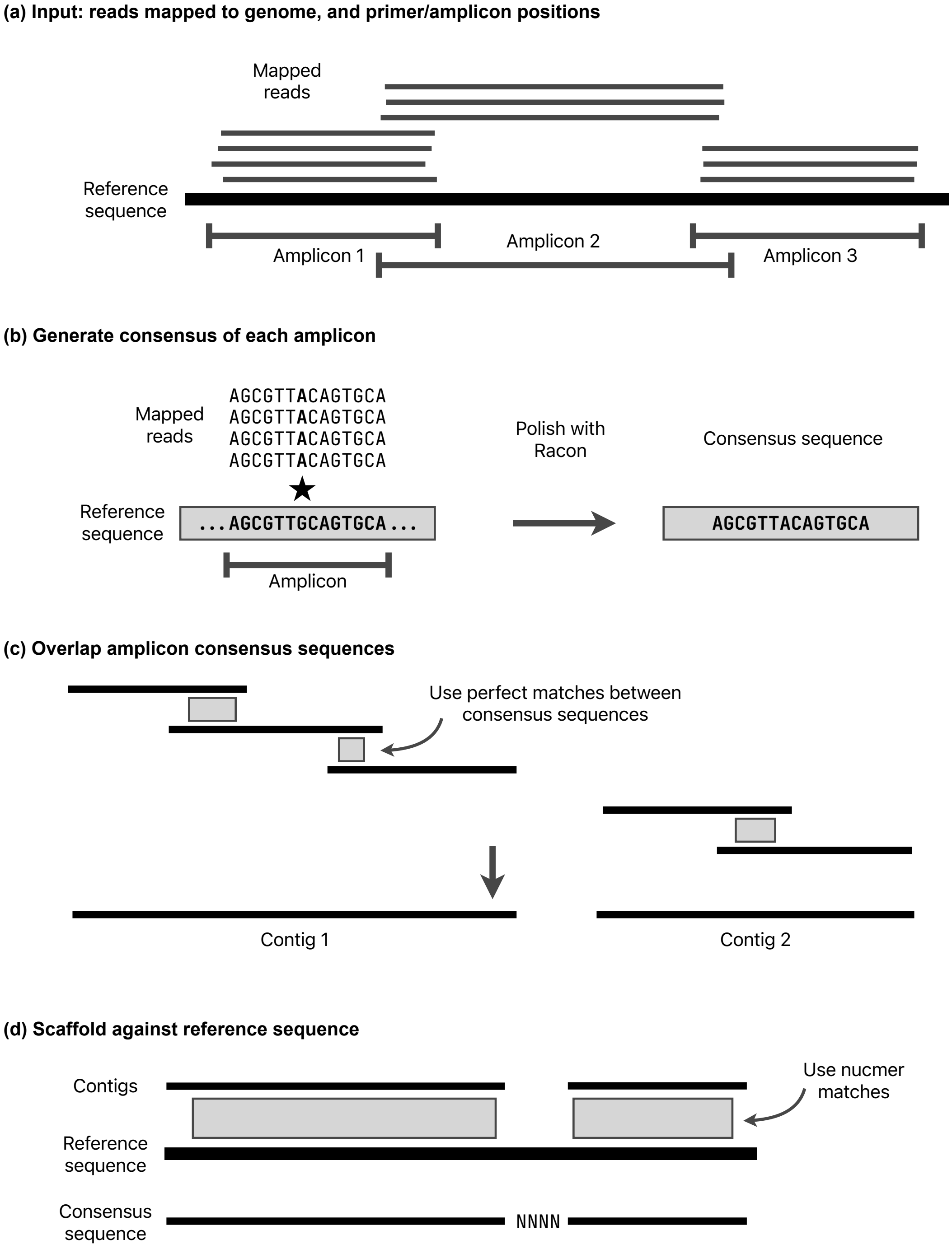

Figure 11: Consensus sequence construction methods. See main text for details. a) The starting point is primer and amplicon positions, and reads mapped to the consensus sequence. b) The consensus sequence of each amplicon is generated independently, using Racon. c) The amplicon sequences are overlapped using perfect matches (if they exist), making contigs. d) The contigs are scaffolded against the reference genome, adding gaps where needed.

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 [35]. The contigs are scaffolded, putting an estimated number of Ns between them based on the mapping coordinates. Since 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` [36]. This globally aligns the cylon consensus sequence to the reference genome to identify variants. Since 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 (ie 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 1000X 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 Figure 12. 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

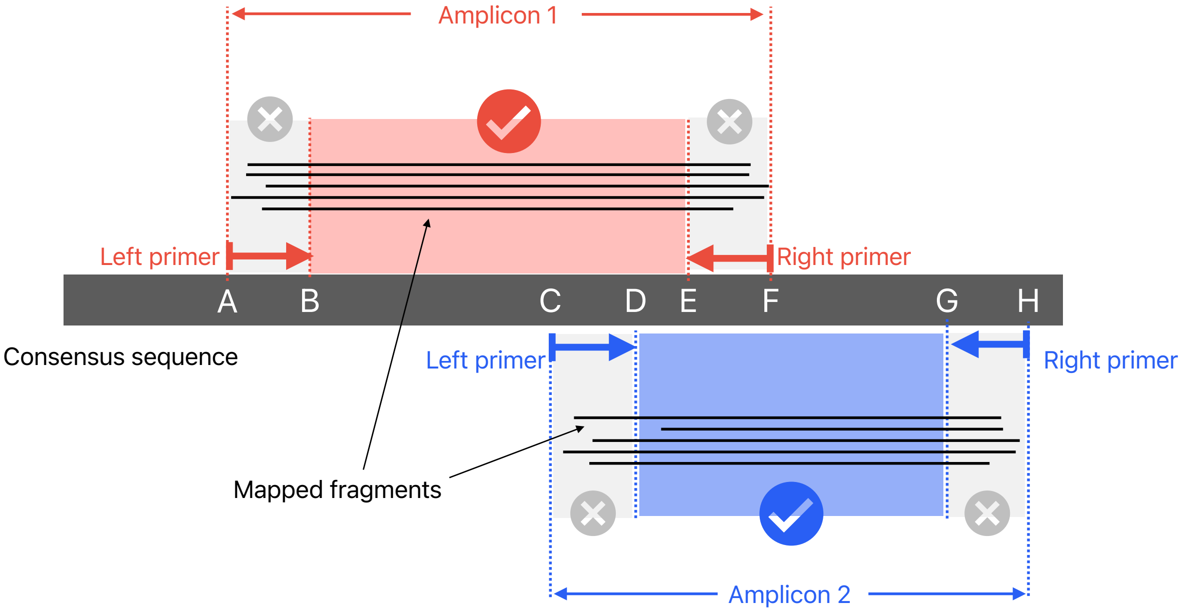

Figure 12: Consensus sequence pileup/masking methods. Two amplicons are shown with fragments (either illumina read pairs, or unpaired nanopore reads) mapped to the consensus. The fragments from amplicon 1 contribute to pileup at B-E, and do not count towards the primer regions A-B or E-F. Similarly, the fragments from amplicon 2 contribute to coverage at D-G (but not to C-D or G-H).

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 MSA between the reference and consensus, amplicon details (chosen primers, number of matching reads etc), and genome-wide read depth statistics.

Simulated data

We developed a Snakemake [37] pipeline to simulate PCR artefacts for 8000 SARS-CoV-2 samples, to compare the assembly accuracy of Viridian to the Connor Lab (`https://github.com/connor-lab/ncov2019-artic-nf`) and Epi2me labs (`https://github.com/epi2me-labs/wf-artic`) ARTIC Nextflow workflows. Firstly, truth assemblies are simulated from a reference genome and reference phylogeny using PhastSim [38] and truth variant calls obtained using varifier [36]. The primer sequences of the ARTIC v4 amplicon scheme are then mapped to the truth assembly of each sample using the `aln` command of `bwa` [39] 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 artefacts 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 ($\mu = 500, SD = 20$). Reads are then simulated from each amplicon at the selected sequencing depths using ART [40] for Illumina and Badread [41] 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 Eval (`https://github.com/iqbal-lab-org/covid-truth-eval`), which is described in detail later, was used to generate TSV files that summarise 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 (OUH) as part of Pillar 1 national surveillance in the United Kingdom. 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 ONT Midnight protocols, in addition to sequence-independent single-primer amplification, forming the dataset deposited in ENA projects PRJEB50520 and PRJEB51850 [42]. Sequencing was performed at the University of Oxford except where otherwise stated below.

Viral culture. Vero cells were maintained in Dulbecco's Modified Eagle Medium (DMEM) high glucose supplemented with 1% fetal bovine serum, 2mM Glutamax, 100 IU/ml penicillin-streptomycin, and 2.5μg/ml amphotericin B at 37°C, 5% CO2 in a humidified atmosphere before inoculation with 200μl of throat swab fluid. Cells were then incubated at 37°C, with daily monitoring for cytopathic effects (CPE). When CPE reached 80%, virus-containing supernatants were harvested 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 in order to verify protein sequence integrity. Refer to [43] for more details.

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

PCR. PCR tests were performed by OUH using two PCR assays: Altona RealStar (targeting E and S genes; Altona Diagnostics, Liverpool, UK) and Thermo Fisher TaqPath assay (targeting S and N genes, and ORF1ab; Thermo Fisher, Abingdon, UK).

Sequence Independent Single Primer Amplification (SISPA). Viral RNA was extracted as described above then Complementary DNA (cDNA) was prepared using a SISPA approach [44]. In brief, firstly RNA was reverse-transcribed with SuperScript III Reverse Transcriptase (Life Technologies, UK) using Sol-Primer A (5'-GTTTCCCACTGGAGGATA-N9-3') [45]. Then 5μL of cDNA and 1μL (100pmol/μL) Primer B (5'-GTTTCCCACTGGAGGATA-3') were added to a 50μL reaction using AccuTaq LA (Sigma, Poole, United Kingdom), according to manufacturer's instructions. PCR conditions were 98°C for 30s, followed by 30 cycles of 94°C for 15s, 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, Brea, California, US) and quantified using the Qubit High Sensitivity dsDNA kit (Thermo Fisher Scientific, UK).

SISPA Oxford Nanopore sequencing. SISPA products were sequenced following a previously described protocol [46] using Oxford Nanopore Technologies (ONT) native barcoding (EXP-NBD104) and ligation sequencing (SQK-LSK109) kits with R9.4.1 flow cells.

ARTIC V3 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 (Version 3.0 7/21). Manual library normalisation 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, UK) and 2200 TapeStation (Agilent Technologies, USA). Paired-end sequencing was performed using the MiSeq reagent kit v2, with 2×250bp, and one water control on each run. NEBNext® Multiplex Oligos for Illumina® (96 Unique Dual Index Primer Pairs) were used.

ARTIC V4.1 Illumina sequencing. Libraries were sequenced at the University of Northumbria following the ARTIC V4.1 CoronaHiT-Illumina protocol [47], using an Illumina NextSeq 550.

ARTIC V3 Oxford Nanopore sequencing. Sequencing was performed using the ARTIC LoCost protocol and v3 primers using R9.4.1 flow cells. Final library concentration was quantified by the High Sensitivity dsDNA kit Qubit (Thermo Fisher Scientific, UK).

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 [48] 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 (Covid Truth Eval). 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 multiple sequence alignment (MSA) of the consensus, truth, and reference sequences using MAFFT [24]. 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 (ie A, C, G, T that is different from the reference), "heterozygous" SNP (ie 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 (ie 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. Since 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=9935) or ONT (N=2352) sequencing reads, with primer schemes ARTIC Version 3 or 4, or MIDNIGHT-1200 (Supplementary table S7). All samples were processed with Viridian and ARTIC-ILM/ONT, producing a consensus sequence.

Global dataset

Metadata for all sequencing runs with taxon ID 2697049 was 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 2nd 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 3 "heterozygous" base calls (ie 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. The final number of samples remaining was 3,960,704.

All GenBank genomes were downloaded on 23rd 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, ie 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 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

Since 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 version 3 and `4/4.1alt/ARTIC v4` for ARTIC version 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 version 3 and Viridian scheme version 4, sorting by run accession, and taking 5 equally spaced runs from the list. The same method was used for Illumina with ENA/SRA version 4 and Viridian version 3, and then similarly for Oxford Nanopore samples, totalling 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 [22] and visualised with taxonium [49]. 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` [49]. The processing of input sequences to obtain taxonium input was implemented in a pipeline called Ushonium (`https://github.com/martinghunt/ushonium`).

In order 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, since they could be placed in an incorrect clade and then cause structural errors.

Calculation of mutational spectra and proportions of G>T mutations

Mutational spectra were calculated as reported previously [26]. Briefly, 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, i.e. 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 [26]. Confidence intervals on the proportion of G>T mutations were calculated using Wilson score interval incorporating the calculated proportion and the number of sampled mutations.

Software versions

Package versions used for the simulations were: Snakemake v7.8.5 [37], PhastSim v0.0.4 [38], ART v2016.06.05 [40], Badread git commit c2bdcbe [41], 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 [50], bwa git commit c77ace7 [39], htslib v1.14 [51], samtools v1.14 [32], BEDTools v2.30.0 [52], joblib v1.1.0 from `https://github.com/joblib/joblib`, numpy v1.22.1 [53], pandas v1.4.0 [54], pysam v0.18.0 at `https://github.com/pysam-developers/pysam` and tqdm v4.62.3 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 version 1.21 of pangolin-data were used.

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. Usho-nium git commit b024320 was used, with dependencies MAFFT v7.520, UShER git commit 2df81ee, and taxoniumtools v2.0.111.

Data availability

Viridian is freely available under the MIT license at `https://github.com/iqbal-lab-org/viridian`. Supplementary text and figures S1-9 are in the supplementary PDF file. Code used for analysis and to generate figures is availble at `https://github.com/martinghunt/viridian-paper`. The global Viridian tree is hosted at `https://viridian.taxonium.org`. All other additional files are available from Figshare:

- Supplementary table S1, `https://doi.org/10.6084/m9.figshare.25712982` - this is a TSV file containing metadata of all 5,288,952 sequencing runs considered in this study

- Supplementary tables S2-9 in one xlsx file, `https://doi.org/10.6084/m9.figshare.25713045`:

  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 - Numbers of inferred viral introductions

S9 - Country counts in the Viridian global tree, and number of new samples since the tree was built

- Supplementary HTML file, `https://doi.org/10.6084/m9.figshare.25713198` - comparison of Viridian and GenBank assemblies
- All 3,960,705 Viridian consensus sequences that are in the global tree, in a single 196MB tar archive file (`https://doi.org/10.6084/m9.figshare.25713225`), which contains the sequences split over multiple xzipped FASTA files
- The Viridian global tree of 3,960,705 sequences, in JSONL and `.pb` format, `https://doi.org/10.6084/m9.figshare.25713261`
- The GenBank and Viridian intersection trees in JSONL and `.pb` format, `https://doi.org/10.6084/m9.figshare.25713285`
- All other Viridian consensus sequences that are not in the global tree in a single zxipped FASTA file, `https://doi.org/10.6084/m9.figshare.25713342`.

Author contributions

Martin 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. Angie Hinrichs analysed and quality controlled the phylogenies and wrangled metadata. Daniel Anderson developed the simulation framework and performed analyses. Lily Karim performed the reversion analyses and geographical/introduction analysis. Bethany Dearlove analysed the assemblies, Ns, indels, and Pango assignments. Jeff Knaggs contributed to the first implementation of Viridian and initial exploratory work. Piyada Supasa, Wanwisa Dejnirattisai, Chang Liu, Juthathip Mongkolsapaya and Gavin R. Screaton isolated and cultured virus stock used to construct the empirical truth set. Hermione Webster, Gillian Rodger, Teresa Street and Sheila Lumley sequenced the empirical truth set, and Bede Constantinides analysed and quality controlled the data. Philip Fowler did independent testing on different simulations and empirical data. Martin Hunt, Jeff Knaggs, Bede Constantinides, Tim Peto, Derrick Crook analysed the empirical truth set results. Theo Sanderson integrated the phylogeny into taxonium. Nicola de Maio did extensive quality control analyses of the genomes. Christopher Ruis performed the mutation spectrum analysis. Houriiyah Tegally, San Emmanuel James, Tulio de Oliveira collated the “early Omicron” dataset. All other authors collected samples, sequenced genomes and shared data with the archives. Zamin Iqbal conceived of the project. Zamin Iqbal and Russell Corbett-Detig supervised the project. Martin Hunt, Angie Hinrichs, Lily Karim, Daniel Anderson, Theo Sanderson, Russell Corbett-Detig and Zamin Iqbal wrote the paper. All authors reviewed the manuscript.

Acknowledgements

We would like to thank Emma Hodcroft, Richard Neher, Duncan MacCannell, Nick Goldman, Trevor Bedford, Rachel Colquhoun, Andrew Rambaut, Robert Davies, and Rob Lanfear for discussions and advice. We also thank Zahra Waheed, Nadim Rahman and Khadim Gueye for help with ENA submissions. We thank the Global Pathogen Analysis Service team in

Oxford, who used an early version (v0.3.7) 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 Robert Koch Institute for providing excellent sequencing services, and we thank all labs contributing to the German SARS-CoV-2 surveillance. We acknowledge high-performance computing services provided by the Robert Koch Institute. We gratefully acknowledge all data contributors, i.e., 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 would like to thank Nick Goldman, who gave us the idea for the start of the introduction.

Funding

Martin Hunt and Jeff Knaggs were supported by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with the UK Health Security Agency (UKHSA) (NIHR200915), and supported by the NIHR Biomedical Research Centre, Oxford. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health or the UKHSA. Russell Corbett-Detig was supported in part by R35GM128932. Lily Karim acknowledges support from T32HG012344. Angie Hinrichs was supported by U24HG002371 and BAA 200-2021-11554. Theo Sanderson receives funding from the Wellcome Trust through a Sir Henry Wellcome Postdoctoral Fellowship (210918/Z/18/Z). We acknowledge the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Science (CIFMS), China (grant number: 2018-I2M-2-002) to Gavin Screaton and Juthathip Mongkolsapaya, Schmidt Futures, the Red Avenue Foundation, and the Oak Foundation. Ravi Kant was supported by the Academy of Finland (grant number 336490), VEO—European Union's Horizon 2020 (grant number 874735), Finnish Institute for Health and Welfare (THL), the Jane and Aatos Erkko Foundation. We acknowledge the charitable donation of Larry Elliott to Oxford University towards pathogen surveillance, which supported the work of Derrick Crook and Philip Fowler. Robert Wilkinson was supported by Wellcome (222574, 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. Etienne Z. Gnimpieba was supported by the National Institutes of Health [5P20GM103443-20]. Marie Lataretu, Matthew Huska, and the Integrated Molecular Surveillance for SARS-CoV-2 (IMSSC2) Laboratory Network were supported by 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 number: ECDC/HERA/2021/008 ECD.12222 and IMS-HERA2, project number: 101113012).

Jose Arturo Molina Mora: Vicerrectoría de Investigación, Universidad de Costa Rica, grant #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.

Conflict of Interest

Gavin Screaton sits on the GSK Vaccines Scientific Advisory Board, consults for AstraZeneca, and is a founding member of RQ Biotechnology.

References

[1] Yatish Turakhia, Nicola De Maio, Bryan Thornlow, Landen Gozashti, Robert Lanfear, Conor R. Walker, Angie S. Hinrichs, Jason D. Fernandes, Rui Borges, Greg Slodkowicz, Lukas Weilguny, David Haussler, Nick Goldman, and Russell Corbett-Detig. Stability of SARS-CoV-2 phylogenies. *PLOS Genetics*, 16(11):e1009175, November 2020.

[2] Nicola De Maio, Conor Walker, Rui Borges, Lukas Weilguny, Greg Slodkowicz, and Nick Goldman. Issues with sars-cov-2 sequencing data, `https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473`. May 2020.

[3] Matthew R. Henn, Christian L. Boutwell, Patrick Charlebois, Niall J. Lennon, Karen A. Power, Alexander R. Macalalad, Aaron M. Berlin, Christine M. Malboeuf, Elizabeth M. Ryan, Sante Gnerre, Michael C. Zody, Rachel L. Erlich, Lisa M. Green, Andrew Berical, Yaoyu Wang, Monica Casali, Hendrik Streeck, Allyson K. Bloom, Tim Dudek, Damien Tully, Ruchi Newman, Karen L. Axten, Adrianne D. Gladden, Laura Battis, Michael Kemper, Qiandong Zeng, Terrance P. Shea, Sharvari Gujja, Carmen Zedlack, Olivier Gasser, Christian Brander, Christoph Hess, Huldrych F. Günthard, Zabrina L. Brumme, Chanson J. Brumme, Suzane Bazner, Jenna Rychert, Jake P. Tinsley, Ken H. Mayer, Eric Rosenberg, Florencia Pereyra, Joshua Z. Levin, Sarah K. Young, Heiko Jessen, Marcus Altfeld, Bruce W. Birren, Bruce D. Walker, and Todd M. Allen. Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. *PLOS Pathogens*, 8(3):e1002529, March 2012.

[4] Edward Holmes. Novel 2019 coronavirus genome, `https://virological.org/t/novel-2019-coronavirus-genome/319/1`. January 2020.

[5] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, Ming-Li Yuan, Yu-Ling Zhang, Fa-Hui Dai, Yi Liu, Qi-Min Wang, Jiao-Jiao Zheng, Lin Xu, Edward C. Holmes, and Yong-Zhen Zhang. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269, March 2020.

[6] Josh Quick. ncov-2019 sequencing protocol v.1, www.protocols.io, `https://dx.doi.org/10.17504/protocols.io.bbmuik6w`. January 2020.

[7] Marta Alenquer, Filipe Ferreira, Diana Lousa, Mariana Valério, Mónica Medina-Lopes, Marie-Louise Bergman, Juliana Gonçalves, Jocelyne Demengeot, Ricardo B. Leite, Jingtao Lilue, Zemin Ning, Carlos Penha-Gonçalves, Helena Soares, Cláudio M. Soares, and Maria João Amorim. Signatures in SARS-CoV-2 spike protein conferring escape to neutralizing antibodies. *PLOS Pathogens*, 17(8):e1009772, August 2021.

[8] Salvatore Di Giorgio, Filippo Martignano, Maria Gabriella Torcia, Giorgio Mattiuz, and Silvestro G. Conticello. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Science Advances*, 6(25):eabb5813, June 2020.

[9] Nicola De Maio, Conor R Walker, Yatish Turakhia, Robert Lanfear, Russell Corbett-Detig, and Nick Goldman. Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. *Genome Biology and Evolution*, 13(5):evab087, May 2021.

[10] Kentaro Itokawa, Tsuyoshi Sekizuka, Masanori Hashino, Rina Tanaka, and Makoto Kuroda. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PLOS ONE*, 15(9):e0239403, September 2020.

[11] Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(1):33–46, January 2017.

[12] Matthew Cotten, Dan Lule Bugembe, Pontiano Kaleebu, and My V.T. Phan. Alternate primers for whole-genome SARS-CoV-2 sequencing. *Virus Evolution*, 7(1):veab006, January 2021.

[13] Carmen Lia Murall, Fatima Mostefai, Jean-Christophe Grenier, Raphaël Poujol, Julie Hussin Hussin, Sandrine Moreira, B. Jesse Shapiro Shapiro, and the CoVSeQ consortium. Recent evolution and international transmission of SARS-CoV-2 clade 19B (Pango A lineages), `https://virological.org/t/recent-evolution-and-international-transmission-of-sars-cov-2-clade-19b-pango-a-lineages/711`. June 2021.

[14] Theo Sanderson and Jeffrey C. Barrett. Variation at Spike position 142 in SARS-CoV-2 Delta genomes is a technical artifact caused by dropout of a sequencing amplicon. *Wellcome Open Research*, 6:305, November 2021.

[15] Theo Sanderson, Nicola De Maio, Angie S. Hinrichs, Adriano de Bernardi Schneider, Conor Walker, Nick Goldman, Yatish Turakhia, Robert Lanfear, and Russell Corbett-Detig. Systematic errors associated with some implementations of artic v4 and a fast workflow to prescreen samples for new problematic sites, `https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/16`. November 2021.

[16] Josh Quick. Sars-cov-2 v4.1 update for omicron variant. December 2021.

[17] Lorenzo Cerutti. Missing g21987a mutation in sars-cov-2 delta variants due to non-specific amplification by artic v3 primers, `https://virological.org/t/missing-g21987a-mutation-in-sars-cov-2-delta-variants-due-to-non-specific-amplification-by-artic-v3-primers/764`. October 2021.

[18] Nikki E Freed, Markéta Vlková, Muhammad B Faisal, and Olin K Silander. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biology Methods and Protocols*, 5(1):bpaa014, January 2020.

[19] Jakob McBroome, Bryan Thornlow, Angie S Hinrichs, Alexander Kramer, Nicola De Maio, Nick Goldman, David Haussler, Russell Corbett-Detig, and Yatish Turakhia. A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees. *Molecular Biology and Evolution*, 38(12):5819–5824, December 2021.

[20] Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 27(5):737–746, May 2017.

[21] Raquel Viana, Sikhulile Moyo, Daniel G. Amoako, Houriiyah Tegally, Cathrine Scheepers, Christian L. Althaus, Ugochukwu J. Anyaneji, Phillip A. Bester, Maciej F. Boni, Mohammed Chand, Wonderful T. Choga, Rachel Colquhoun, Michaela Davids, Koen Deforche, Deelan Doolabh, Louis Du Plessis, Susan Engelbrecht, Josie Everatt, Jennifer Giandhari, Marta Giovanetti, Diana Hardie, Verity Hill, Nei-Yuan Hsiao, Arash Iranzadeh, Arshad Ismail, Charity Joseph, Rageema Joseph, Legodile Koopile, Sergei L. Kosakovsky Pond, Moritz U. G. Kraemer, Lesego Kuate-Lere, Oluwakemi Laguda-Akingba, Onalethatha Lesetedi-Mafoko, Richard J. Lessells, Shahin Lockman, Alexander G. Lucaci, Arisha Maharaj, Boitshoko Mahlangu, Tongai Maponga, Kamela Mahlakwane, Zinhle Makatini, Gert Marais, Dorcas Maruapula, Kereng Masupu, Mogomotsi Matshaba, Simnikiwe Mayaphi,

Nokuzola Mbhele, Mpaphi B. Mbulawa, Adriano Mendes, Koleka Mlisana, Anele Mnguni, Thabo Mohale, Monika Moir, Kgomotso Moruisi, Mosepele Mosepele, Gerald Motsatsi, Modisa S. Motswaledi, Thongbotho Mphoyakgosi, Nokukhanya Msomi, Peter N. Mwangi, Yeshnee Naidoo, Noxolo Ntuli, Martin Nyaga, Lucier Olubayo, Sureshnee Pillay, Botshelo Radibe, Yajna Ramphal, Upasana Ramphal, James E. San, Lesley Scott, Roger Shapiro, Lavanya Singh, Pamela Smith-Lawrence, Wendy Stevens, Amy Strydom, Kathleen Subramoney, Naume Tebeila, Derek Tshiabuila, Joseph Tsui, Stephanie Van Wyk, Steven Weaver, Constantinos K. Wibmer, Eduan Wilkinson, Nicole Wolter, Alexander E. Zarebski, Boitumelo Zuze, Dominique Goedhals, Wolfgang Preiser, Florette Treurnicht, Marietje Venter, Carolyn Williamson, Oliver G. Pybus, Jinal Bhiman, Allison Glass, Darren P. Martin, Andrew Rambaut, Simani Gaseitsiwe, Anne Von Gottberg, and Tulio De Oliveira. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature*, 603(7902):679–686, March 2022.

[22] Yatish Turakhia, Bryan Thornlow, Angie S. Hinrichs, Nicola De Maio, Landen Gozashti, Robert Lanfear, David Haussler, and Russell Corbett-Detig. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics*, 53(6):809–816, June 2021.

[23] Angie Hinrichs, Cheng Ye, Yatish Turakhia, and Russell Corbett-Detig. The ongoing evolution of UShER during the SARS-CoV-2 pandemic. *Nature Genetics*, 56(1):4–7, January 2024.

[24] Kazutaka Katoh and Daron M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, April 2013.

[25] Cheng Ye, Bryan Thornlow, Angie Hinrichs, Alexander Kramer, Cade Mirchandani, Devika Torvi, Robert Lanfear, Russell Corbett-Detig, and Yatish Turakhia. matOptimize: a parallel tree optimization method enables online phylogenetics for SARS-CoV-2. *Bioinformatics*, 38(15):3734–3740, 06 2022.

[26] Christopher Ruis, Thomas P. Peacock, Luis M. Polo, Diego Masone, Maria Soledad Alvarez, Angie S. Hinrichs, Yatish Turakhia, Ye Cheng, Jakob McBroome, Russell Corbett-Detig, Julian Parkhill, and R. Andres Floto. A lung-specific mutational signature enables inference of viral and bacterial respiratory niche. *Microbial Genomics*, 9(5), May 2023.

[27] Jakob McBroome, Jennifer Martin, Adriano de Bernardi Schneider, Yatish Turakhia, and Russell Corbett-Detig. Identifying SARS-CoV-2 regional introductions and transmission clusters in real time. *Virus Evolution*, 8(1):veac048, 06 2022.

[28] Amanda Warr, Caitlin Newman, Nicky Craig, Ingrida Vendelė, Rizalee Pilare, Lilet Cariazo Cruz, Twinkle Galase Barangan, Reildrin G. Morales, Tanja Opriessnig, Virginia Mauro Venturina, Milagros R. Mananggit, Samantha Lycett, Clarissa YJ Domingo, and Christine Tait-Burkard. No part gets left behind: Tiled nanopore sequencing of whole ASFV genomes stitched together using Lilo. *bioRxiv*, December 2021.

[29] Kim A. Lagerborg, Erica Normandin, Matthew R. Bauer, Gordon Adams, Katherine Figueroa, Christine Loreth, Adrianne Gladden-Young, Bennett M. Shaw, Leah R. Pearlman, Daniel Berenzy, Hannah B. Dewey, Susan Kales, Sabrina T. Dobbins, Erica S. Shenoy, David Hooper, Virginia M. Pierce, Kimon C. Zachary, Daniel J. Park, Bronwyn L. MacInnis, Ryan Tewhey, Jacob E. Lemieux, Pardis C. Sabeti, Steven K. Reilly, and Katherine J. Siddle. Synthetic DNA spike-ins (SDSIs) enable sample tracking and detection of inter-sample

contamination in SARS-CoV-2 sequencing workflows. *Nature Microbiology*, 7(1):108–119, December 2021.

[30] John R. Tyson, Phillip James, David Stoddart, Natalie Sparks, Arthur Wickenhagen, Grant Hall, Ji Hyun Choi, Hope Lapointe, Kimia Kamelian, Andrew D. Smith, Natalie Prystajecky, Ian Goodfellow, Sam J. Wilson, Richard Harrigan, Terrance P. Snutch, Nicholas J. Loman, and Joshua Quick. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore, September 2020.

[31] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018.

[32] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[33] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, January 2021.

[34] John M. Gaspar. NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics*, 19(1):536, December 2018.

[35] Guillaume Marçais, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, 14(1):e1005944, January 2018.

[36] Martin Hunt, Brice Letcher, Kerri M. Malone, Giang Nguyen, Michael B. Hall, Rachel M. Colquhoun, Leandro Lima, Michael C. Schatz, Srividya Ramakrishnan, Zamin Iqbal, and CRyPTIC consortium. Minos: variant adjudication and joint genotyping of cohorts of bacterial genomes. *Genome Biology*, 23(1):147, July 2022.

[37] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 08 2012.

[38] Nicola De Maio, William Boulton, Lukas Weilguny, Conor R. Walker, Yatish Turakhia, Russell Corbett-Detig, and Nick Goldman. phastSim: Efficient simulation of sequence evolution for pandemic-scale datasets. *PLOS Computational Biology*, 18(4):e1010056, April 2022.

[39] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 05 2009.

[40] Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 2012.

[41] Ryan Wick. Badread: simulation of error-prone long reads. *Journal of Open Source Software*, 4(36):1316, 2019.

[42] Bede Constantinides, Hermione Webster, Gillian Rodger, Martin Hunt, Piyada Supasa, Wanwisa Dejnirattisai, Chang Liu, Juthathip Mongkolsapaya, Gavin R. Screaton, and Derrick Crook. A diverse reference set of cultured sars-cov-2 genomes sequenced using various

amplification methods and instrument platforms, `https://doi.org/10.6019/S-BSST1334`. *BioStudies*, February 2024.

[43] Rungtiwa Nutalai, Daming Zhou, Aekkachai Tuekprakhon, Helen M. Ginn, Piyada Supasa, Chang Liu, Jiandong Huo, Alexander J. Mentzer, Helen M.E. Duyvesteyn, Aiste Dijokaite-Guraliuc, Donal Skelly, Thomas G. Ritter, Ali Amini, Sagida Bibi, Sandra Adele, Sile Ann Johnson, Bede Constantinides, Hermione Webster, Nigel Temperton, Paul Klenerman, Eleanor Barnes, Susanna J. Dunachie, Derrick Crook, Andrew J. Pollard, Teresa Lambe, Philip Goulder, Neil G. Paterson, Mark A. Williams, David R. Hall, Juthathip Mongkolsapaya, Elizabeth E. Fry, Wanwisa Dejnirattisai, Jingshan Ren, David I. Stuart, and Gavin R. Screaton. Potent cross-reactive antibodies following Omicron breakthrough in vaccinees. *Cell*, 185(12):2116–2131.e18, June 2022.

[44] Alexander L. Greninger, Samia N. Naccache, Scot Federman, Guixia Yu, Placide Mbala, Vanessa Bres, Doug Stryke, Jerome Bouquet, Sneha Somasekar, Jeffrey M. Linnen, Roger Dodd, Prime Mulembakani, Bradley S. Schneider, Jean-Jacques Muyembe-Tamfum, Susan L. Stramer, and Charles Y. Chiu. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, 7(1):99, December 2015.

[45] Liana E. Kafetzopoulou, Kyriakos Efthymiadis, Kuiama Lewandowski, Ant Crook, Dan Carter, Jane Osborne, Emma Aarons, Roger Hewson, Julian A. Hiscox, Miles W. Carroll, Richard Vipond, and Steven T. Pullan. Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Eurosurveillance*, 23(50), December 2018.

[46] Kuiama Lewandowski, Yifei Xu, Steven T. Pullan, Sheila F. Lumley, Dona Foster, Nicholas Sanderson, Alison Vaughan, Marcus Morgan, Nicole Bright, James Kavanagh, Richard Vipond, Miles Carroll, Anthony C. Marriott, Karen E. Gooch, Monique Andersson, Katie Jeffery, Timothy E. A. Peto, Derrick W. Crook, A. Sarah Walker, and Philippa C. Matthews. Metagenomic Nanopore Sequencing of Influenza Virus Direct from Clinical Respiratory Samples. *Journal of Clinical Microbiology*, 58(1):e00963–19, December 2019.

[47] Dave J. Baker, Alp Aydin, Thanh Le-Viet, Gemma L. Kay, Steven Rudder, Leonardo De Oliveira Martins, Ana P. Tedim, Anastasia Kolyva, Maria Diaz, Nabil-Fareed Alikhan, Lizzie Meadows, Andrew Bell, Ana Victoria Gutierrez, Alexander J. Trotter, Nicholas M. Thomson, Rachel Gilroy, Luke Griffith, Evelien M. Adriaenssens, Rachael Stanley, Ian G. Charles, Ngozi Elumogo, John Wain, Reenesh Prakash, Emma Meader, Alison E. Mather, Mark A. Webber, Samir Dervisevic, Andrew J. Page, and Justin O'Grady. CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes. *Genome Medicine*, 13(1):21, December 2021.

[48] Tim Carver, Simon R. Harris, Matthew Berriman, Julian Parkhill, and Jacqueline A. McQuillan. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, 28(4):464–469, February 2012.

[49] Theo Sanderson. Taxonium, a web-based tool for exploring large phylogenetic trees. *eLife*, 11:e82392, November 2022.

[50] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017.

[51] James K Bonfield, John Marshall, Petr Danecek, Heng Li, Valeriu Ohan, Andrew Whitwham, Thomas Keane, and Robert M Davies. HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience*, 10(2):giab007, 02 2021.

[52] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 01 2010.

[53] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[54] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.