A validated cloud-based genomic platform for co-ordinated, expedient global analysis of SARS-CoV-2 genomic epidemiology

Authors

Daniel Gyamfi Amoako,[1,2] Nguyen To Anh, [3] Jasmine Bastable,[4] Marc Brouard,[5] Constanza Campano Romero,[6] Andres Castillo Ramirez,[6] Bede Constantinides,[5] Derrick W. Crook,[4,5] Phan Manh Cuong,[7] Moussa Moise Diagne,[8] Amadou Diallo,[8] Nguyen Thanh Dung,[9] Laura Dunn,[4] Le Van Duyet,[7] Josie Everatt,[1] Katherine Fletcher,[5] Philip W. Fowler,[5] Mailie Gail,[10] Jessica Gentry,[4] Saheer Gharbia,[11] Hospital for Tropical Diseases SARS-CoV-2 testing team,[9] Nguyen Thi Thu Hong,[3] Martin Hunt,[5,13] Zamin Iqbal, [13,14] Katie Jeffery,[4] Dikeledi Kekana,[1] Thomas Kesteman,[3] Jeff Knaggs,[5,13] Marcela Lopes Alves,[5] Dinh Nguyen Huy Man,[9] Amy J. Mathers,[12] Nghiem My Ngoc,[9] Sarah Oakley,[4] Hardik Parikh,[12] Tim E.A. Peto, [4,5] T. Phuong Quan,[5] Marcelo Rojas Herrera,[6] Nicholas Sanderson,[5] Vitali Sintchenko,[10] Jeremy Swann,[5] Junko Takata, [4,5] Nguyen Thi Tam,[3] Le Van Tan,[3] Pham Ngoc Thach,[7] Ndeye Marieme Top,[8] Nguyen Thu Trang,[3] Van Dinh Trang,[7] Robert Turner,[5] H. Rogier van Doorn,[3] Anne von Gottberg,[1,15] Jeremy Westhead,[5] Nicole Wolter,[1,15] Bernadette C. Young[4,5,16]

Author Affiliations

1. The National Institute for Communicable Diseases (NICD), a division of the National Health Laboratory Service, Johannesburg, Republic of South Africa
2. College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa
3. Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam.
4. Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom
5. Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

6. El Instituto de Salud Pública de Chile, Santiago, Chile
7. National Hospital for Tropical Diseases, Hanoi, Vietnam
8. The Institut Pasteur de Dakar, Dakar, Senegal
9. Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam
10. Microbial Genomics Reference Laboratory, Institute of Clinical Pathology and Medical Research, New South Wales Health Pathology, Sydney, Australia
11. National Infection Service, UK Health Security Agency, London, United Kingdom
12. Division of Infectious Diseases, School of Medicine, University of Virginia, Charlottesville, United States of America
13. European Molecular Biology Laboratory – European Bioinformatics Institute, Hinxton, United Kingdom
14. Milner Centre for Evolution, University of Bath, United Kingdom
15. Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
16. NIHR Oxford Biomedical Research Centre, Oxford, United Kingdom

Corresponding author: Junko Takata (junko.takata@ndm.ox.ac.uk), Nuffield Department of Medicine, University of Oxford, Oxford, OX3 9DU, 01865 220856. All authors listed alphabetically.

Keywords: COVID-19, SARS-CoV-2, public health genomics, bioinformatics

Abstract

Background

Viral sequencing has made critical contributions to our understanding of and response to the COVID-19 pandemic, but sequencing capacity and bioinformatic expertise remain limited in many settings. This proof-of-principle study aimed to demonstrate the utility of a cloud-based sequencing analysis pipeline, the Tiled Amplicon Pipeline (TAP), for rapid and collaborative SARS-CoV-2 sequencing across seven globally distributed sites.

Methods

In this cross-sectional study from July to August 2022, seven international sites submitted all SARS-CoV-2 sequence data generated over a two-week period to our cloud platform. No patient identifying information was uploaded, and human reads were removed prior to upload to the cloud. Users could opt in to share sample information with collaborators via a tagging system. The pipeline performed sequence assembly, lineage identification and relatedness analysis.

Results

Seven sites contributed 5,432 sequences, of which 5,342 (98.3%) were from clinical samples and 90 (1.7%) were controls. 4,470/5,342 (83.7%) clinical samples had sufficient coverage for lineage assignment. Omicron lineages dominated, with BA.5, BA.4 and BA.2 comprising the vast majority, consistent with contemporary epidemiological observations at the time. Phylogenetic analysis demonstrated low diversity within lineages, and genotypically identical or highly similar sequences were recovered from globally disparate sites.

72 **Conclusions**

73 A cloud-based analysis platform like TAP addresses bioinformatic bottlenecks and
74 facilitates international capacity building and collaboration in pathogen surveillance,
75 enhancing global epidemic and pandemic preparedness.

76 **Importance**

77 The COVID-19 pandemic showed the potential of whole-genome sequencing in informing
78 public health responses, but also highlighted the significant global disparity of
79 sequencing capacity and bioinformatic expertise. In this study, we describe the
80 development and validation of a cloud-based, automated sequencing analysis pipeline for
81 SARS-CoV-2 – the Tiled Amplicon Pipeline (TAP). 5,432 sequences from clinical samples
82 across six continents were collected over a two-week period and processed through TAP,
83 demonstrating its power to co-ordinate secure data sharing and aggregated global
84 analysis. TAP is designed to have the capability to assemble other viruses beyond SARS-
85 CoV-2 that use tiled amplicon sequencing, highlighting its potential to be adapted quickly
86 for use in future outbreaks. Tools such as TAP can simplify access to and interpretation
87 of genomic data, which in turn contributes to a growing global landscape of genomic data
88 sharing and surveillance that underpin pandemic preparedness.

Introduction

Viral genome sequencing has proven pivotal to understanding the evolution of the SARS-CoV-2 virus during the COVID-19 pandemic and in shaping the public health response. International genomic surveillance and data sharing initiatives have together made it possible to track the emergence of variants globally [1], to demonstrate the impact of travel restrictions on viral dynamics across continents [2] or within countries [3, 4], and to identify transmission routes within hospitals [5]. Successive waves of infection driven by new variants showed that rapidly detecting new lineages is critical for understanding disease epidemiology and guiding subsequent public health responses [6], as well as informing the development of vaccines and therapeutics such as neutralising antibodies [7-9].

However, the pandemic also highlighted marked global variability in sequencing capacity and cost [10], with much of the SARS-CoV-2 sequencing undertaken at centralised reference laboratories [11]. Several important challenges remain in expanding genomic surveillance, including (1) access to sequencing technology, especially in lower- and middle-income countries, such as access to consumables, complex import requirements, or failure of supply chains; (2) availability of bioinformatic expertise; and (3) interpretability of results generated using a plethora of different wet lab and bioinformatics protocols [12-14]. There is a need for accessible genomic surveillance infrastructure that can be used by researchers, microbiologists, and clinicians from any location, to deliver an up-to-date, global perspective of genomic evolution of pathogens with epidemic or pandemic potential.

113 One solution to both the shortage of bioinformatic expertise and the lack of global
114 interpretability of results is web-accessible analysis infrastructure [15]. The Global
115 Pathogen Analysis Service (GPAS) was rapidly set up in 2021 in response to the COVID-
116 19 pandemic by the University of Oxford as a cloud-based, globally accessible web
117 platform. GPAS provided fast and secure access to a comprehensive SARS-CoV-2 genomic
118 analysis pipeline, delivering genome assembly, variant calling and lineage classification
119 from raw sequence data. Users retained control of their data but could opt in to share
120 their data to facilitate comparisons in a wider context, empowering laboratories to
121 control their own analysis. GPAS was validated to UKAS ISO 15189:2012 standards in a
122 UK tertiary hospital clinical microbiology laboratory, and was commissioned by the UK
123 Health Security Agency (UKHSA) for over a year to support Pillar 1 national surveillance
124 (testing in UKHSA laboratories and NHS hospitals, for healthcare workers and those with
125 a clinical need) and various clinical trials. GPAS provided proof-of-principle that standard
126 microbiology laboratories without bioinformatics expertise can generate timely outputs
127 for local surveillance and automatically submit sequences to public repositories.
128
129 In 2024, GPAS was upgraded as the Tiled Amplicon Pipeline (TAP) with major
130 enhancements to several software components. Since January 2025 TAP has been
131 deployed as the SARS-CoV-2 pipeline on EIT Pathogena (version 1.2.0), a multi-pipeline
132 genomic pathogen analysis platform that is free of charge for users in low- and middle-
133 income countries and accessible in all settings with internet access [16]. This service
134 demonstrates the potential of cloud-based platforms to overcome barriers to
135 democratising effective genomic surveillance, making advanced genomic analysis
136 available to resource-limited laboratories.
137

138 In this paper, we report the results of a collaborative, cross-sectional SARS-CoV-2
139 sequencing study across seven globally distributed sites, which demonstrate the utility
140 of a cloud-based sequencing platform in providing a quality assured, rapid, and
141 integrated global snapshot of viral diversity. We further describe the components of the
142 pipeline and discuss its potential for rapid adaptation in response to future viral
143 pandemics.
144

Materials and Methods

Pipeline development

TAP was deployed to a cloud platform and controlled via a Command Line Interface (CLI) tool. Full methodology is described in the Appendix (Supplementary Appendix A). At upload, a universally unique identifier (UUID) is generated and assigned to each sample. The mapping between the UUID and the user's sample identifier is downloaded and only held by the user, and FASTQ header lines are also truncated, ensuring that no potentially personally identifiable information is transmitted to the cloud platform. Tight access control to data applies within TAP, and by default data is not shared with other users unless explicitly authorised by the data owners.

Following FASTQ file upload to the cloud, processing commences automatically with no further user input required. Any reads mapping to the human genome are first removed [17], reducing the risk of any human reads being retained. Genome assembly is performed by an amplicon-aware genome assembly tool Viridian v1.3.1 [18], which scaffolds per-amplicon *de novo* assemblies into a single whole genome consensus assembly in FASTA format. Viridian was configured to use the SARS-CoV-2 reference genome Wuhan-Hu-1 (MN908947.3) for assembly and variant calling [19], and a library of seven amplicon primer schemes (AmpliSeq v1; ARTIC versions 3, 4.1, 5.3.2 (400), 5.2.0 (1200); Midnight 1200; VarSkip v1a-2b). When a primer scheme is not specified by the user, Viridian can automatically infer the most likely primer scheme from within its library.

After assembly, amino acid mutations are identified using Nextclade [20], and Pango lineages are assigned with Pangolin version 4.3.1 [21]. Aligned sequences are compared

170 using a novel algorithm, FindNeighbour5, which identifies single nucleotide variant
171 (SNV) distances between sequences without conflicting variant calls (0 SNVs) as well as
172 those differing by one, two and three SNVs [22, 23]. The main outputs from the pipeline
173 are presented in an access-restricted user interface portal and downloadable as a
174 summary file containing the lineage assignment, a list of related samples (according to
175 permissions), and metrics of the genome assembly (e.g. coverage, mean depth, amplicon
176 dropouts, etc). Detailed intermediate files such as VCF or FASTA files are also available
177 for download.
178
179 **Pipeline validation**
180 GPAS was validated for both Illumina and Oxford Nanopore Technologies (ONT)
181 sequencing platforms in 2021, using various datasets including a 'truth set' of cultured
182 SARS-CoV-2 samples sequenced with multiple platforms and library preparations,
183 annotated with manually curated variant calls (described in further detail in [18] and
184 [24]), in addition to community samples collected in Northumbria, UK. Overall, GPAS
185 showed negligible adjusted false call rates (less than 1/100,000 nucleotides) with respect
186 to the 'truth set', and had high concordance (less than 1.9/1,000,000 discordant events)
187 with the ARTIC assembly pipeline in use at the time.
188
189 TAP was subsequently validated against the same 'truth set' of cultured SARS-CoV-2
190 samples, which achieved 100% concordance with expected Pango lineages using the
191 same Pangolin version. Concordance of TAP and GPAS was further compared using data
192 from the current study, which showed 99.4% concordance in lineage calls (using the
193 same Pangolin and Viridian versions), and nucleotide call discordance of 0.59/1,000,000

194 sites. Further details of the validation and comparison processes are described in
195 Supplementary Appendix B.
196
197 **Sample frame, sequencing, and upload**
198 Between July and August 2022, seven sequencing centres participated in a two-week
199 sequencing pilot across seven countries: Senegal; Chile; South Africa; New South Wales
200 (NSW), Australia; Vietnam; United Kingdom (UK); and Virginia, United States (USA).
201 Centres were either accredited clinical microbiology and/or public health laboratories.
202 All clinical samples in which SARS-CoV-2 was detected by PCR in respiratory samples
203 were eligible for inclusion and underwent in-house genomic sequencing at participating
204 sites, with no more than one submission from an individual patient. Sequencing platform
205 (either Illumina or ONT) and primer schemes were chosen by participating sites, who
206 followed their established sequencing protocols.
207
208 Raw sequences were uploaded in FASTQ format to GPAS, the working name of the
209 pipeline at the time, along with limited associated metadata (sample identifier,
210 instrument platform, sample type [clinical or control], collection date, and country).
211 Submitting sites verified the run quality reports of each sequencing batch including a
212 review of positive and negative controls, and batches were passed or failed accordingly.
213 All passed samples were tagged to release them to the shared data pool for subsequent
214 aggregate analysis; explicit permission was given from all collaborators to configure data
215 access controls such that all submitters could access and view each other's sequences,
216 metadata, and analytical outputs. For this study, automatic detection of primer scheme
217 was enabled. All sequences were uploaded to the European Nucleotide Archive (project
218 accession: PRJEB70597) at the time of the study.

Data analysis

FASTQ files from the study were redownloaded from the ENA and run through EIT Pathogena (version 1.2.0 pre-release) on 6th November 2024. A maximum likelihood phylogeny of aligned sequences was constructed for the three largest Pango lineages in the sample set (BA.5, BA.4 and BA.2) using IQTree version 2.3.6, assuming a general time reversible nucleotide substitution model with gamma rate heterogeneity [25]. All other analyses were conducted in RStudio version 2023.06.0+421.

227 # Results

228 **Primer scheme detection**

229 5,432 sequences were shared across seven sites (Supplementary Table S1), with date of
230 collection ranging from April to July 2022. Of submitted samples, 90 (1.7%) were controls
231 and 5,342 (98.3%) were sequences from clinical samples. Primer schemes were
232 successfully auto-detected by Viridian for six out of seven sites; for the remaining site
233 (NSW, Australia), the primer scheme was not auto-detected as it used a modified
234 Midnight-1200 scheme to account for amplicon dropouts, and was therefore explicitly
235 specified.

236

237 **Aggregate analysis of global genetic epidemiology**

238 Among the clinical samples included in the analysis, 4,782/5,342 (90.0%) were
239 assembled with at least 70% genome coverage (Table 1). 4,470/5,342 (83.7%) clinical
240 samples were assembled with sufficient coverage and post-assembly quality to be
241 assigned a lineage (Table 2). 4,442/4,470 (99.4%) were Omicron variants, with BA.5,
242 BA.4 and BA.2 being the most common. A small number of Delta variant sequences were
243 identified (0.3%, 15/4,470), all of which were collected prior to July 2022.

Centre	Submitted samples	Controls	<50% coverage*	50-70% coverage	>70% coverage	Earliest sample collection date	Latest sample collection date
Senegal	197	0	9	10	178	13th July 2022	20th July 2022
Chile	1205	20	4	6	1175	16th June 2022	22nd July 2022
South Africa	202	0	7	6	189	17th May 2022	24th June 2022
NSW, Australia	1194	31	132	0	1031	21st June 2022	18th July 2022
Vietnam	316	3	14	7	292	1st April 2022	14th July 2022
United Kingdom	1818	36	280	85	1417	12th July 2022	26th July 2022
Virginia, USA	500	0	0	0	500	7th June 2022	28th June 2022
Total	**5432**	**90**	**446**	**114**	**4782**		

Table 1: Samples submitted by study centre, with date of collection and genome coverage. Earliest and latest sample dates exclude controls. NSW = New South Wales.

* includes samples that could not be assembled

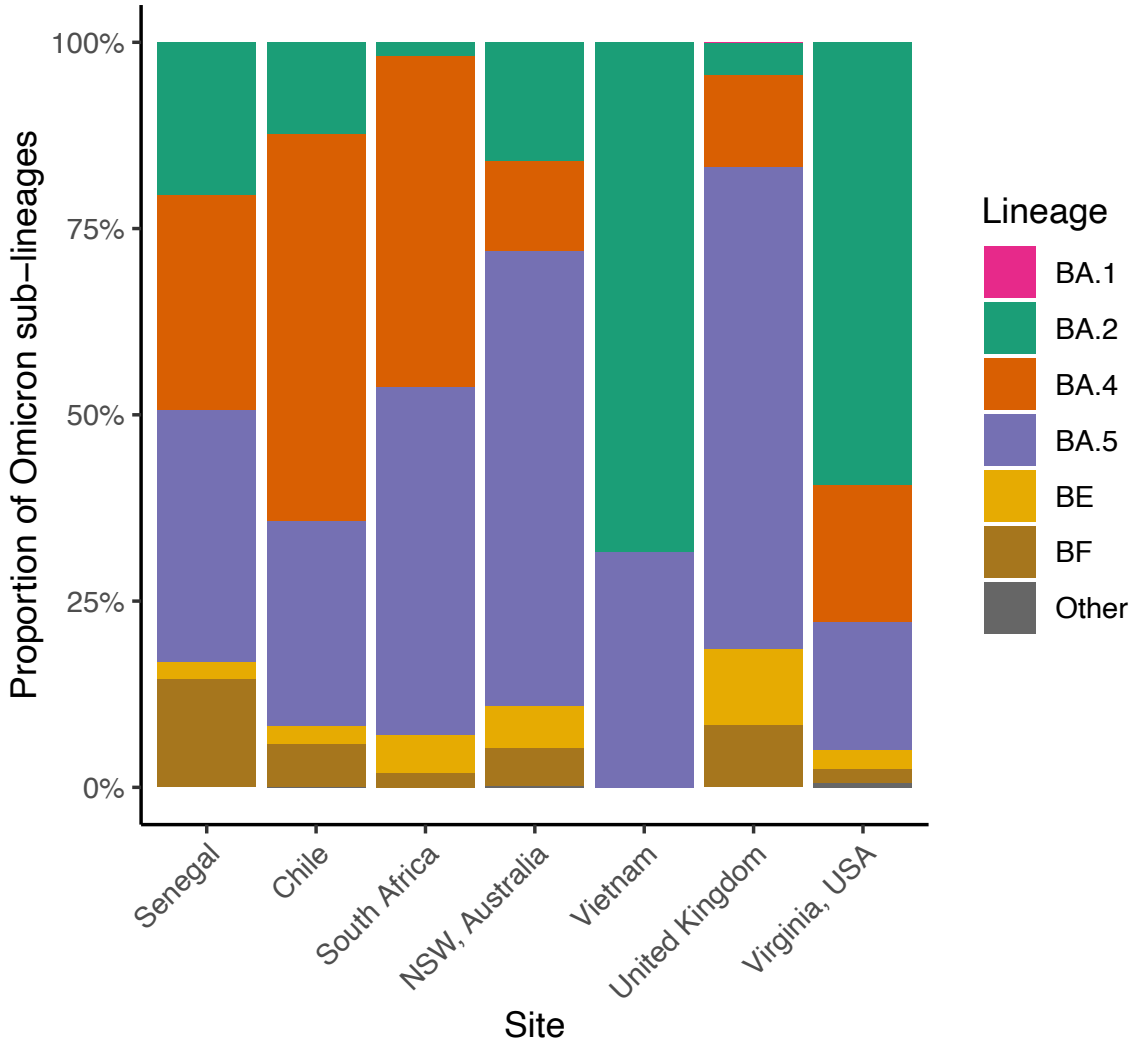
Centre	Omicron							Delta	Alpha	Other*	Total
	BA.1	BA.2	BA.4	BA.5 and related				AY.5, AY.57, B.1.617.2	B.1.1.7		
				BA.5	BE	BF	Other (BG, BK)				
Senegal	0	17	24	28	2	12	0	0	0	0	83
Chile	0	136	570	302	26	63	1	0	0	5	1103
South Africa	0	3	82	77	8	3	0	0	0	3	176
NSW, Australia	0	165	124	629	58	52	2	0	0	1	1031
Vietnam	2	228	0	6	0	0	0	15	0	0	251
United Kingdom	2	57	165	864	137	111	0	0	1	2	1339
Virginia, USA	0	289	89	84	12	9	3	0	0	1	487
Total	4	895	1054	1990	243	250	6	15	1	12	4,470

Table 2: SARS-CoV-2 lineage of samples by study centre (where a lineage was assigned by Pangolin). NSW = New South Wales.

* Other includes: P.1, P.1.2, A, A.2.2, B.1, B.1.429, B.28, C.37, DE.1, XAM, XAN, XAS, XAZ, XBA, P.2

253 **Global mixing across multiple Omicron lineages**

254 For clinical samples collected between 1st June 2022 and 31st July 2022, the proportion
255 of samples assigned to the different Omicron sub-lineages varied substantially by study
256 site (Figure 1). Each of BA.2, BA.4 or BA.5 dominated (constituted >50% of samples from)
257 at least one site, while in Senegal no single lineage dominated, and BA.4 and BA.5 were
258 equally common in South Africa (44.3% and 46.9% respectively).



259

260

261 **Figure 1: Omicron sub-lineages by study site.** Proportion of samples assigned to
262 Omicron sub-lineages within each study site, where collection date was in June or July
263 2022.

Maximum likelihood phylogenies for each of the three most common lineages revealed global mixing of Omicron lineages (Figure 2). There were no examples of sub-lineages being completely geographically restricted to one site; one sub-lineage of BA.4 was found predominantly in Chile, but an example of this sub-lineage was also identified in the United Kingdom.

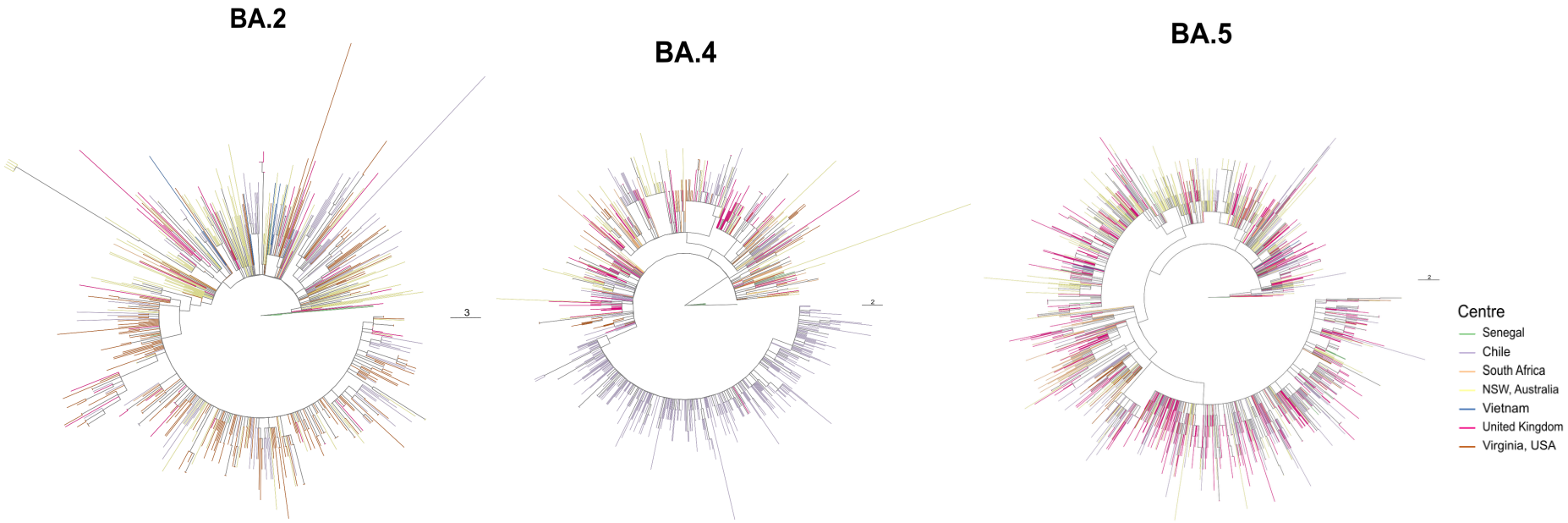
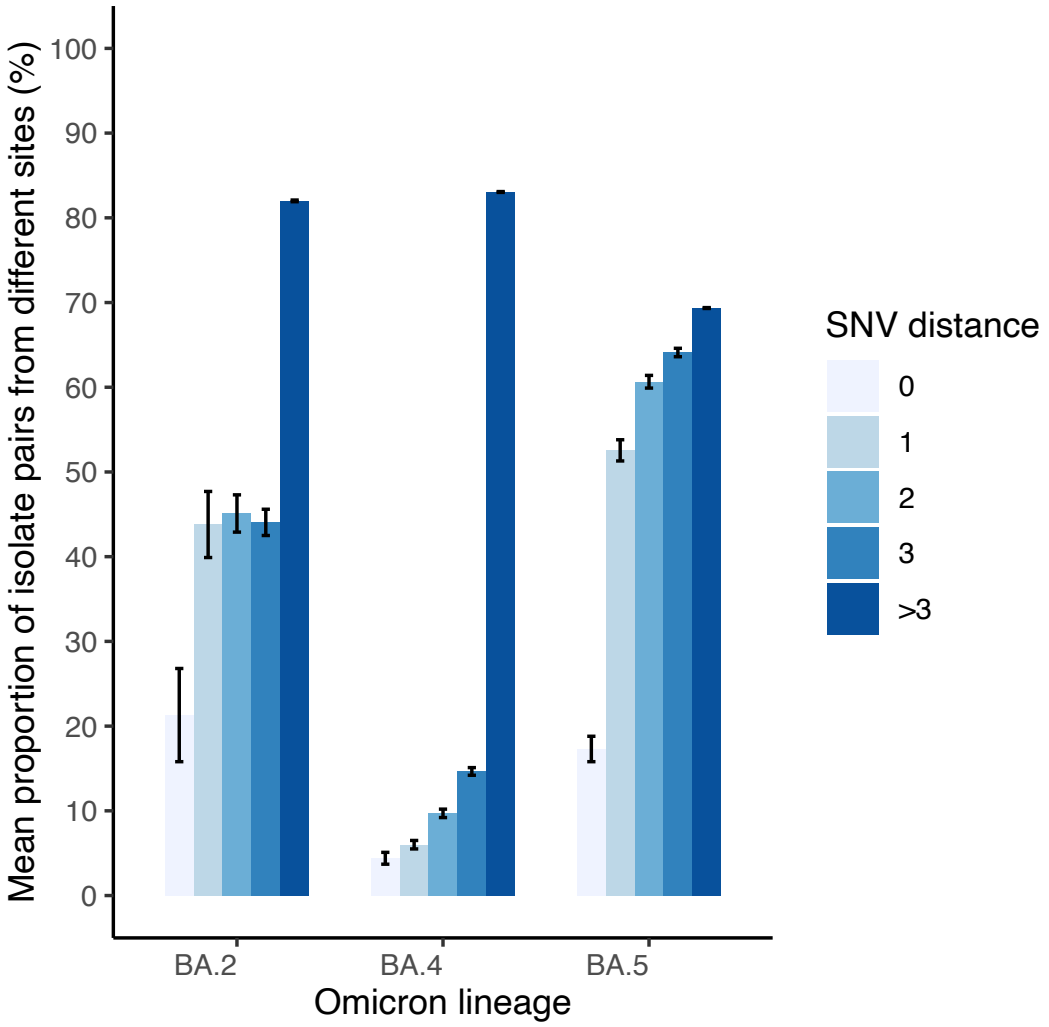


Figure 2: Maximum likelihood phylogeny of isolates from major SARS-CoV-2 lineages (BA.2, BA.4, BA.5), with branches coloured by sequencing site. Branch lengths are scaled to represent substitutions across the full SARS-CoV-2 genome with corresponding scale bars.

274 Similarly, a proportion of genotypically identical (SNV=0) isolate pairs were found in
275 different sites for all 3 major lineages, particularly for the BA.2 lineage, where 21.3% of
276 SNV=0 pairs were found in different sites (Figure 3, Supplementary Table 2). This
277 proportion increased with greater SNV distances. BA.4 showed the least amount of global
278 dispersal below SNV <3, corresponding to the majority of SNV <3 isolate pairs in this
279 lineage being from Chile where there was evidence of a dominant sub-lineage as
280 illustrated in Figure 2.



281

282 **Figure 3:** Proportion of unique isolate pairs that are from different sites at each SNV
283 distance for each major Omicron lineage. Bars indicate 95% confidence intervals.

284

285 **Cloud processing time and performance**

286 At the time of the study, cloud processing performance on GPAS was measured by time
287 elapsed between upload and completion of analysis. Median cloud processing time per
288 sample was 30.6 minutes (IQR: 13.1 – 69.9); variability was observed, with several
289 batches from Australia, UK and USA taking longer than 1000 minutes (Supplementary
290 Figure 1). These were caused by either a data centre outage during an extreme heat event
291 in the UK, or as a result of a software update prematurely applied to the upload portal,
292 which was promptly resolved by improvements in the client side CLI software. Other data
293 upload issues were identified and addressed case-by-case, including metadata file
294 formatting errors, the availability of upload clients compatible with all required
295 operating systems, and user interface display errors preventing batch release.

296

297 In TAP, median cloud processing time per sample for a typical batch of 100 was 14.5
298 minutes (IQR: 13.8 – 15.0) for Illumina and 15.6 minutes (IQR: 15.2 – 16.7) for ONT; due
299 to parallelisation, processing a single batch of 100 samples was achievable in 19.7
300 minutes (Illumina) and 20.4 minutes (ONT) (Supplementary Figure 2). Processing all
301 5,432 samples took 766 minutes (12.7 hours). CPU time per sample is approximately 3
302 minutes (noting that this occurs in parallel, i.e. 2 CPUs working for 4 minutes would be
303 reported as 8 minutes CPU time by the workflow manager, despite only taking 4
304 minutes); majority of analysis time is associated with other pipeline activities including
305 network file transfers or reading data from disk into memory. Peak RAM usage for typical
306 use of the pipeline was approximately 5GB.

307 ## Discussion

308 This pilot study demonstrates the potential of globally synchronous data processing and
309 analysis in informing public health, through a unified protocol of genome assembly,
310 variant calling and relatedness analysis. The aggregated analysis from this study
311 conforms to contemporary observations of SARS-CoV-2 genetic epidemiology at the time
312 [26-28], and demonstrates cosmopolitan global mixing of Omicron lineages between
313 study sites. Even a BA.4 sub-lineage found almost exclusively in Chile was not completely
314 restricted to that site and could be identified in a geographically distant site (UK).
315 Similarly, up to 21% of genotypically identical (SNV=0) sequence pairs were identified in
316 different countries over the same two-month period. Such mixing of lineages likely
317 reflects the relaxation of travel restrictions in the participating sites at the time of the
318 study, and captures the rapid global dispersal of Omicron with successive selective
319 sweeps of new lineages over short time scales [29]. These observations highlight the
320 potential of a cloud-based sequencing pipeline in facilitating data sharing and generating
321 co-ordinated insights that could inform real-time decision-making.

322

323 While automatic primer scheme detection was used for this study, our results show the
324 limitation in this approach, where a modified scheme was not detected to be amongst the
325 current Viridian primer scheme library. Although this illustrates the sensitivity of
326 Viridian's auto-detection, it is inevitable that with viral evolution, primer schemes will
327 continue to be modified or newly developed to maintain sequencing coverage of an
328 evolving target genome. Viridian has the capability to accommodate custom primer
329 schemes when provided with a suitable scheme definition in browser extensible data
330 (BED) format. This capability will therefore form the basis of future pipeline iterations
331 and will allow the pipeline to pivot in future to assemble other viruses beyond SARS-CoV-

332 2 that use tiled amplicon sequencing. In turn, this ability to rapidly adapt existing
333 infrastructure for emerging threats, rather than build bespoke solutions from scratch
334 each time, supports a proactive and 'Always On' approach to pandemic preparedness
335 [30].
336
337 Our study is limited by the opportunistic sampling frame used and is likely to be
338 geographically incomplete, with a preponderance for countries that have historically
339 contributed more to global sequencing efforts. Similarly, our study has the limitation of
340 lacking longitudinal data on how these findings changed over time. Despite these
341 shortcomings, we have demonstrated how even a simply structured sampling across time
342 and space is still well placed to rapidly identify replacement by new lineages. Such
343 limitations in sampling frame diversity and global disparity in sequencing capacity were
344 reflected during the pandemic itself: sequencing data produced in high-income countries
345 was up to ten-fold higher than that achieved in low- and middle-income countries during
346 the first two years [11], and data was not collected evenly over the course of pandemic.
347 Well-designed, suitably powered, and representative sampling is a key part of national
348 and international pandemic preparedness planning, as exemplified by a cohort design in
349 the UK [27].
350
351 The world is now better prepared to urgently initiate genomic pathogen surveillance. The
352 available sequencing platform capacity is much improved, and the achievements of global
353 data aggregation (as illustrated by INSDC, GISAID and Pathoplexus databases) have
354 alleviated a major obstacle to effective genomic surveillance and data sharing. In turn, the
355 integration of genomic data with disease manifestation and severity data would enable
356 the ready investigation of associations between genomic variation and clinical outcome

357 as suggested by others [31], and facilitate predictive modelling for anticipating the course
358 of future epidemics [32].
359
360 A remaining obstacle, especially in low-resource settings, is establishing local turnkey
361 bioinformatics to ensure standardised, quality-assured outputs that do not depend on in-
362 house bioinformatics expertise. In TAP, we have developed an exemplar of such a cloud-
363 based service, with simple ingestion of local, raw sequence data in FASTQ format, and
364 flexible privacy protections for data sharing to facilitate local or international
365 comparative analyses. Outputs from TAP such as the consensus FASTA sequence can in
366 turn be submitted to international repositories such as GISAID which do not take raw
367 FASTQ files directly. While factors such as internet bandwidth may limit real-world
368 performance, TAP is capable of rapid parallel data processing, with a run time of 15
369 minutes per sample and 20 minutes in total for a batch of 100 typical SARS-CoV-2
370 genomes. The EIT Pathogena platform ensures sustainability of technical and
371 infrastructural support to the pipeline and is currently freely available to low- and
372 middle-income countries, supporting long-term use across research, clinical, or
373 surveillance settings. Such a service, in turn, contributes to the growing global landscape
374 of genomic data sharing that will underpin future pandemic preparedness.
375

376 ## Conclusions

377 The COVID-19 pandemic has demonstrated the multiple ways in which viral sequencing
378 can inform pandemic responses, but also the inequitable distribution of resources and
379 capabilities. A service such as TAP contributes to the democratisation of genomics,
380 enabling researchers and laboratories, regardless of their location or bioinformatics
381 expertise, to participate actively in global surveillance efforts. By providing accessible

382 and user-friendly tools for sequence assembly, lineage assignment, and data sharing, it
383 promotes inclusive collaboration, harmonises data, and ultimately enhances pandemic
384 preparedness.

Author contributions

Study conceptualisation: BCY, DWC

Study design: BCY, DWC

Software building and computational infrastructure: MB, BC, MH, ZI, PWF, JK, MLA, NS, JS, RT, JW

Data collection: DA, NTA, JB, CCR, ACR, PMC, MMD, AD, NTD, LD, LVD, JE, KF, MG, JG, SG, Hospital for Tropical Diseases SARS-CoV-2 testing team, NTTH, KJ, DK, TK, DNHM, AJM, NMN, SO, HP, MRH, VS, NTTa, LVT, PNT, NMT, NTTr, VDT, HRVD, AvG, NW

Data analysis: JT, BCY, BC, TEAP, TPQ

Manuscript writing: JT, BCY, BC

Manuscript editing and review: all authors

Funding statement

GPAS is a non-profit organisation, and GPAS cloud infrastructure was supported by a donation from Oracle Corporation. For this study, all GPAS services were offered free of charge to all sites, and staff costs for the analysis team were met by the University of Oxford, with support from the National institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and the NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance (NIHR200915), a partnership between the UK Health Security Agency (UKHSA) and the University of Oxford. The views expressed are those of the author(s) and not necessarily those of the NIHR, UKHSA or the Department of Health and Social Care. TAP is deployed on the EIT Pathogena platform, funded by the Ellison Institute of Technology Oxford. EIT Pathogena is free at the point of use for users in low- and middle-income countries.

Sequencing activities for NICD are supported by a conditional grant from the South African National Department of Health as part of the emergency COVID-19 response; a cooperative agreement between the National Institute for Communicable Diseases of the National Health Laboratory Service and the United States Centers for Disease Control and Prevention (NU51IP000930); the South African Medical Research Council (SAMRC) with funds received from the South African Department of Science and Innovation; the African Society of Laboratory Medicine (ASLM) and Africa Centers for Disease Control and Prevention through a sub-award from the Bill and Melinda Gates Foundation grant number INV-018978; Africa PGI, the UK Foreign, Commonwealth and Development Office and Wellcome (Grant no 221003/Z/20/Z); and the Department of Health and Social Care's Fleming Fund using UK aid. NICD sequencing was also supported by The Coronavirus Aid, Relief, and Economic Security Act (CARES ACT) through the Centers for Disease Control and Prevention (CDC) and the COVID International Task Force (ITF) funds through the CDC under the terms of a subcontract with the African Field Epidemiology Network (AFENET) AF-NICD-001/2021.

Genomic surveillance conducted in Vietnam was supported by the Wellcome Trust (222574/Z/21/Z). L.V.T. is supported by the Wellcome Trust of Great Britain (204904/Z/16/Z and 226120/Z/22/Z).

Ethical statement

SARS-CoV-2 sequencing was performed for public health surveillance and ethical approval for secondary analysis was not required. This determination was reviewed by the University of Oxford Joint Research Office. No patient identifying information was

shared as part of this study. Sample identifiers remain with the submitter, and are never kept on GPAS/TAP, or shared with other users. GPAS/TAP generates identifiers for each sequence and writes a file linking anonymised ID to the submitted identifiers. Only the user submitting sequence data has access to these records. The GPAS/TAP upload client selects only SARS-CoV-2 sequence reads and discards all human sequences.

Data availability

Submitted sequencing reads (free from human reads) for all included samples are available from European Nucleotide Archive study accession PRJEB70597.

Conflicts of interest

AvG and NW have received grant funding from Sanofi and The Bill and Melinda Gates Foundation. PWF and DWC receive consultancy fees from the Ellison Institute of Technology, Oxford. JT and TEAP receive funding from the Ellison Institute of Technology, Oxford.

Acknowledgments

- Microbial Genomics Reference Laboratory, New South Wales Health Pathology, Sydney, Australia
- El Instituto de Salud Pública de Chile, Chile
- Centre for Respiratory Diseases and Meningitis (CRDM), National Institute for Communicable Diseases (NICD), a division of the National Health Laboratory Service, South Africa
- The Institut Pasteur de Dakar, Senegal

- Oxford University Hospitals NHS Trust, Oxford, United Kingdom
- University of Virginia, Charlottesville, United States of America
- Oxford University Clinical Research Unit, Vietnam
- Hospital for Tropical Diseases SARS-CoV-2 testing team, HTD Vietnam: Le Manh Hung, Nguyen Le Nhu Tung, Nguyen Thanh Phong, Vo Minh Quang, Pham Thi Ngoc Thoa, Nguyen Thanh Truong, Tran Nguyen Phuong Thao, Dao Phuong Linh, Ngo Tan Tai, Ho The Bao, Vo Trong Vuong, Huynh Thi Kim Nhung
- Oracle Global Health Business Unit

467 **References**

468 1. Li J, Lai S, Gao GF, Shi W. The emergence, genomic diversity and global spread of
469 SARS-CoV-2. Nature. 2021;600(7889):408-18.

470 2. Wilkinson E, Giovanetti M, Tegally H, San JE, Lessells R, Cuadros D, et al. A year of
471 genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. Science.
472 2021;374(6566):423-31.

473 3. McCrone JT, Hill V, Bajaj S, Pena RE, Lambert BC, Inward R, et al. Context-specific
474 emergence and growth of the SARS-CoV-2 Delta variant. Nature. 2022;610(7930):154-
475 60.

476 4. Raghwani J, du Plessis L, McCrone JT, Hill SC, Parag KV, Theze J, et al. Genomic
477 Epidemiology of Early SARS-CoV-2 Transmission Dynamics, Gujarat, India. Emerg Infect
478 Dis. 2022;28(4):751-8.

479 5. Snell LB, Fisher CL, Taj U, Stirrup O, Merrick B, Alcolea-Medina A, et al. Combined
480 epidemiological and genomic analysis of nosocomial SARS-CoV-2 infection early in the
481 pandemic and the role of unidentified cases in transmission. Clin Microbiol Infect.
482 2022;28(1):93-100.

483 6. Updated working definitions and primary actions for SARS-Cov-2 variants. WHO
484 Technical Advisory Group on Virus Evolution; 2023.

485 7. Tao K, Tzou PL, Nouhin J, Gupta RK, de Oliveira T, Kosakovsky Pond SL, et al. The
486 biological and clinical significance of emerging SARS-CoV-2 variants. Nat Rev Genet.
487 2021;22(12):757-73.

488 8. Tuekprakhon A, Nutalai R, Dijokaite-Guraliuc A, Zhou D, Ginn HM, Selvaraj M, et
489 al. Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1
490 serum. Cell. 2022;185(14):2422-33 e13.

491 9. Aggarwal A, Akerman A, Milogiannakis V, Silva MR, Walker G, Stella AO, et al.
492 SARS-CoV-2 Omicron BA.5: Evolving tropism and evasion of potent humoral responses
493 and resistance to clinical immunotherapeutics relative to viral variants of concern.
494 EBioMedicine. 2022;84:104270.

495 10. Merhi G, Koweyes J, Salloum T, Khoury CA, Haidar S, Tokajian S. SARS-CoV-2
496 genomic epidemiology: data and sequencing infrastructure. Future Microbiol.
497 2022;17:1001-7.

498 11. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, et al.
499 Global disparities in SARS-CoV-2 genomic surveillance. Nat Commun. 2022;13(1):7003.

500 12. Carter LL, Yu MA, Sacks JA, Barnadas C, Pereyaslov D, Cognat S, et al. Global
501 genomic surveillance strategy for pathogens with pandemic and epidemic potential
502 2022-2032. Bull World Health Organ. 2022;100(4):239-A.

503 13. Inzaule SC, Tessema SK, Kebede Y, Ogwell Ouma AE, Nkengasong JN. Genomic-
504 informed pathogen surveillance in Africa: opportunities and challenges. Lancet Infect
505 Dis. 2021;21(9):e281-e9.

506 14. Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, et al. Global landscape of SARS-
507 CoV-2 genomic surveillance and data sharing. Nat Genet. 2022;54(4):499-507.

508 15. Ohlsen EC, Hawksworth AW, Wong K, Guagliardo SAJ, Fuller JA, Sloan ML, et al.
509 Determining Gaps in Publicly Shared SARS-CoV-2 Genomic Surveillance Data by
510 Analysis of Global Submissions. Emerg Infect Dis. 2022;28(13):S85-S92.

511 16. EIT Pathogena [Available from: https://eit-pathogena.com.

512 17. Constantinides B, Hunt M, Crook DW. Hostile: accurate decontamination of
513 microbial host sequences. Bioinformatics. 2023;39(12).

514 18. Hunt M, Hinrichs AS, Anderson D, Karim L, Dearlove BL, Knaggs J, et al.
515 Addressing pandemic-wide systematic errors in the SARS-CoV-2 phylogeny. bioRxiv.
516 2024:2024.04.29.591666.

517 19. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus
518 associated with human respiratory disease in China. Nature. 2020;579(7798):265-9.

519 20. Aksamentov I, Roemer C, Hodcroft EB, Neher RB. Nextclade: clade assignment,
520 mutation calling and quality control for viral genomes. Journal of Open Source Software.
521 2021;6(67):3773.

522 21. O'Toole A, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment
523 of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol.
524 2021;7(2):veab064.

525 22. Mazariegos-Canellas O, Do T, Peto T, Eyre DW, Underwood A, Crook D, et al.
526 BugMat and FindNeighbour: command line and server applications for investigating
527 bacterial relatedness. BMC Bioinformatics. 2017;18(1):477.

528 23. FindNeighbour5 [Available from: https://github.com/oxfordmmm/FN5.

529 24. Constantinides B, Webster H, Rodger G, Hunt M, Supasa P, Dejnirattisai W, et al. A
530 diverse reference set of cultured SARS-CoV-2 genomes sequenced using various
531 amplification methods and instrument platforms 2024 [Available from:
532 https://www.ebi.ac.uk/biostudies/studies/S-BSST1334.

533 25. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler
534 A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the
535 Genomic Era. Mol Biol Evol. 2020;37(5):1530-4.

536 26. Tosta S, Moreno K, Schuab G, Fonseca V, Segovia FMC, Kashima S, et al. Global
537 SARS-CoV-2 genomic surveillance: What we have learned (so far). Infect Genet Evol.
538 2023;108:105405.

539 27. Foulkes S, Monk EJM, Sparkes D, Hettiarachchi N, Milligan ID, Munro K, et al.
540 Early Warning Surveillance for SARS-CoV-2 Omicron Variants, United Kingdom,
541 November 2021-September 2022. Emerg Infect Dis. 2023;29(1):184-8.

542 28. Xu Y, Liu T, Li Y, Wei X, Wang Z, Fang M, et al. Transmission of SARS-CoV-2
543 Omicron Variant under a Dynamic Clearance Strategy in Shandong, China. Microbiol
544 Spectr. 2023;11(2):e0463222.

545 29. Markov PV, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, et al. The
546 evolution of SARS-CoV-2. Nat Rev Microbiol. 2023;21(6):361-79.

547 30. van der Westhuizen HM, Soundararajan S, Berry T, Agus D, Carmona S, Ma P, et
548 al. A consensus statement on dual purpose pathogen surveillance systems: The always
549 on approach. PLOS Glob Public Health. 2024;4(11):e0003762.

550 31. Oude Munnink BB, Worp N, Nieuwenhuijse DF, Sikkema RS, Haagmans B,
551 Fouchier RAM, et al. The next phase of SARS-CoV-2 surveillance: real-time molecular
552 epidemiology. Nat Med. 2021;27(9):1518-24.

553 32. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, et al. Assessing
554 transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature. 2021;593(7858):266-
555 9.
556