SARS-CoV-2 sequencing with cloud-based analysis illustrates expedient co-ordinated surveillance of viral genomic epidemiology across six continents

Authors

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

Affiliations

1. The National Institute for Communicable Diseases (NICD), a division of the National Health Laboratory Service, Johannesburg,, Republic of South Africa
2. Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam.
3. Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
4. El Instituto de Salud Pública de Chile, Chile
5. Oxford University Hospitals NHS Trust, Oxford, United Kingdom
6. National Hospital for Tropical Diseases, Hanoi, Vietnam
7. The Institut Pasteur de Dakar, Senegal
8. Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam
9. Microbial Genomics Reference Laboratory, Institute of Clinical Pathology and Medical Research, New South Wales Health Pathology, Sydney, Australia

10. Division of Infectious Diseases, School of Medicine, University of Virginia, Charlottesville, United States of America
11. European Bioinformatics Institute, Hinxton, UK
12. Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

Corresponding author: Bernadette Young (Bernadette.young@ndm.ox.ac.uk)

Abstract

Viral sequencing has been critical in the COVID-19 pandemic response, but sequencing and bioinformatics capacity remain inconsistent. To examine the utility of a cloud-based sequencing analysis platform for SARS-CoV-2 sequencing, we conducted a cross-sectional study incorporating seven countries in July 2022. Sites submitted sequential SARS-CoV-2 sequences over two weeks to the Global Pathogen Analysis Service (GPAS). The GPAS bioinformatics cloud platform performs sequence assembly plus lineage and related sample identification. Users can share information with collaborators while retaining data ownership. Seven sites contributed sequencing reads from 5,346 clinical samples, of which 4,799/5,346 (89.8%) had a lineage identified. Omicron lineages dominated, with the vast majority being BA.5, BA.4 and BA.2, commensurate with contemporary genomic epidemiological observations. Phylogenetic analysis demonstrated low within-lineage diversity, and highly similar sequences present in globally disparate sites. A cloud-based analysis platform like GPAS addresses bioinformatics bottlenecks and facilitates collaboration in pathogen surveillance, enhancing epidemic and pandemic preparedness.

Introduction

Viral genomic sequencing has been pivotal in understanding and responding to SARS-CoV-2 viral evolution during the COVID-19 pandemic. Global sharing of data and comparison of samples has facilitated public health surveillance by tracking the emergence of variants globally,[1] demonstrating the impact of travel restrictions on viral dynamics across continents,[2] and of travel and quarantine on spread within countries,[3,4] as well as demonstrating transmission routes within hospitals.[5] As successive waves driven by new viral variants emerged, it became clear that rapidly detecting new lineages is critical both for understanding and predicting disease epidemiology as well as for guiding both public health measures and treatment of infection, including the use of neutralising antibodies.[6,7] However there remains marked global variability in sequencing costs and capacity.[8] Much SARS-CoV-2 sequencing has been undertaken at centralised reference laboratories, but the speed and power of this tool is greatly enhanced when researchers and clinicians at any location can contribute to the global view of viral evolution and surveillance.[9]

Several important challenges remain in expanding pathogen genomic surveillance, including access to sequencing technology, the availability of bioinformatic resources and the global interpretability of results generated using different bioinformatics approaches.[10,11,12] One solution to the relative shortage of bioinformatic expertise in many settings is web-accessible analysis infrastructure.[13] The Global Pathogen Analysis Service (GPAS) is a cloud-based globally web-accessible platform that provides secure and rapid access to comprehensive SARS-CoV-2 sequencing genome assembly and analysis software. GPAS supports multiple sequencing platforms and tiling PCR primer schemes, and has been validated against widely used methods. Users can opt to share their data on this platform, allowing those who opt in to view their own results in a wider context, facilitating comparisons for rapid and comprehensive assessments of viral data. The platform is designed for use by clinical and public health laboratory scientists and has been established and used for over a year in a tertiary hospital based clinical microbiology laboratory to ISO15189:2012 standards. Scaling

up this simple-to-use assembly, variant calling and analysis tool for SARS-CoV-2 empowers laboratories to control the analysis of their sequence data. It provides proof of principle that standard microbiology laboratories without bioinformatics expertise from across the world can generate outputs for local surveillance and automated submission to public databases. Such a service is an exemplar of how cloud-based services, accessible in all settings with internet access, offer great promise for overcoming an important barrier to effective genomic surveillance.

The primary aim of this project was to demonstrate that a globally broad snapshot of SARS-CoV-2 viral diversity can be simultaneously gathered, processed and shared through a cloud-based software, enabling aggregation for combined descriptive analysis.

Methods

Sample frame and sequencing: Sequencing centres participated in a two-week pilot in July to August 2022. Centres were from seven countries: Senegal; Chile; South Africa; New South Wales (NSW), Australia; Vietnam; United Kingdom; Virginia, United States (USA). Centres were either accredited clinical microbiology or public health laboratories. All clinical samples in which SARS-CoV-2 was detected were eligible for inclusion and underwent in-house genomic sequencing at participating sites during the pilot were eligible for inclusion, with no more than one submission from an individual patient.

Sequencing was performed using either Illumina or Oxford Nanopore Technology (ONT) platforms. Multiple primer schemes were supported (ARCTIC V3, V4 or 4.1, and Midnight). Sequencing platform and primer schemes were chosen by participating sites, who followed their established sequencing protocols.

Bioinformatic methods: Unselected raw sequences were submitted to GPAS in FASTQ format with sequencing information (instrument platform, primer scheme) and limited

113 associated metadata (sample name, whether sequence data was a control or study sample,
114 collection date, country of origin). At upload, a Universally Unique IDentifier (UUID) was
115 randomly generated and assigned to each sample. The mapping between the centre's sample
116 identifier and the GPAS UUID was held only by the submitting organisation, ensuring that no
117 potentially personally identifiable information was transmitted to the GPAS platform. When
118 results were returned to the submitter after analysis, they could hence be re-linked to the
119 subject's record. At upload, only reads mapping to SARS-CoV-2 were kept,[14] guaranteeing
120 that no human reads were retained and pushed to the cloud. Tight access control to data
121 applied within GPAS. With the permission of all collaborators, data access controls were
122 configured such that all submitters could access each other's sequences, metadata and
123 analytical output data for the duration of the investigation.
124
125 Sequence data analysis was performed by the GPAS SARS-CoV-2 bioinformatics pipeline.
126 Genome assembly was performed by an amplicon-aware *de novo* genome assembly tool
127 Viridian v0.3.7.[15] Viridian applies quality control (QC) at both a read and amplicon level,
128 to produce a consensus assembly. Each genomic position is called A,C,G or T only where 10
129 or more reads passing amplicon-aware QC provide at least 70% support, otherwise it is called
130 ambiguously as N. After assembly, clade assignment and amino acid mutation calls were
131 made using Nextclade,[16] and Pango lineages assigned with Pangolin.[17] Samples were
132 further labelled according to the United Kingdom Health Security Agency (UKHSA)
133 nomenclature to define and track variants of concern using aln2type.[18] The GPAS QC suite
134 provided read-level QC output for each sample, reporting the number of mapped reads per
135 sample, median read depth, reference coverage at read depths of 10 and 20, and the number
136 of amplicons failing quality control. Consensus genome assemblies were aligned by Viridian
137 to the Wuhan-Hu-1 reference genome (MN908947.3).[19] Aligned sequences were compared
138 using findNeighbour4, which identifies sequences without conflicting variant calls (0 single
139 nucleotide variants (SNVs)), as well as those differing by one, two and three SNVs.[20,21]
140

Submitting sites verified the run quality reports of each sequencing batch, including review of positive and negative controls, and outputs were passed or failed accordingly. All passed samples were tagged for aggregate analysis, and released to the GPAS shared data pool, which identifies other samples within three SNVs in the shared data pool. The GPAS platform recorded the total number of sequencing runs and total number of samples. Analysis duration was also monitored, along with descriptive reports of any unexpected issues with the platform.

Phylogenetic analysis: In Oxford, genome assemblies in the shared data pool were downloaded from the GPAS portal. 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) in using RAxML [22] (assuming a general time reversible (GTR) nucleotide substitution model).

156 **Results**

157 All sites tagged their sequencing data and results granting access for aggregated analysis.
158 5,437 sequences were shared (Table 1, Table S1). As different sequencing centres worked
159 with samples collected at different times, the date of collection of submitted sequences varied
160 by centre, ranging from April to July 2022. Of submitted samples 91 (1.7%) were controls
161 and 5,346 (98.3%) were clinical samples Among these, 4,797/5,346 (89.7%) were assembled
162 with at least 70% coverage (at a depth of 10 reads), and 4,799/5,346 (89.8%) were assembled
163 with sufficient coverage to be assigned a lineage (Table 2). 4,775/4,799 (99.5%) were
164 Omicron variants, with BA.5, BA.4 and BA.2 being the most common. A small number of
165 Delta variant sequences were identified (23/4,799 0.5%), all of which were collected prior to
166 July 2022.

167

168 Table 1: Samples submitted by study centre, with date of collection and genome coverage
169 (with minimum depth of 10 reads)

Centre	Submitted	Controls	<50% coverage	50-70% coverage	>70% coverage	Earliest sequence	Latest sequence
Senegal	197	0	14	10	173	13th July 2022	20th July 2022
Chile	1205	20	4	7	1174	16th June 2022	21st July 2202
South Africa	202	0	9	4	189	17th May 2022	24th June 2022
NSW, Australia	1194	31	34	23	1106	21st June 2022	18th July 2022
Vietnam	316	3	20	5	288	1st April 2022	14th July 2022
United Kingdom	1823	37	319	100	1367	12th July 2022	26th July 2022
Virginia, USA	500	0	0	0	500	7th June 2022	29th June 2022
Total	5437	91	400	149	4797	1st April 2022	26th July 2022

170

171

172

173

174

175

176 Table 2: SARS-CoV-2 lineage by study centre (where a lineage was assigned by Pangolin)

	Omicron							Delta	Alpha	Other	
Centre	BA.5	BA.4	BA.2	BE.1	BF.1	BA.1	BG.2	Delta	B.1.1.7	B.28	Total
Senegal	69	47	34	5	18	0	0	0	0	0	173
Chile	397	607	145	20	4	0	0	0	0	0	1,173
South Africa	89	86	3	6	0	5	0	0	0	0	189
NSW, Australia	736	130	184	54	3	0	0	0	0	0	1,107
Vietnam	8	0	256	0	0	4	0	22	0	0	290
United Kingdom	974	169	57	140	23	2	0	0	1	1	1,367
Virginia, USA	95	94	296	11	1	1	1	1	0	0	500
Total	2,368	1,133	975	236	49	12	2	23	1	1	4,799

177

178 Comparing only clinical samples collected between 1st June 2022 and 31st July 2022, the
179 proportion of samples assigned to the different Omicron sub-lineages varied substantially by
180 study site (Figure 1). Each of BA.2, BA.4 or BA.5 dominated (constituted >50% of samples
181 from) at least one site, while Senegal had no single dominant lineage. For each of the three
182 most common lineages, a maximum likelihood phylogeny was constructed to compare
183 diversity within and between sites (Figure 2).

184

185 Phylogenetic analysis revealed global mixing of Omicron lineages. There were no examples
186 of sub-lineages being completely geographically restricted within this study. One sub-lineage
187 in BA.4 was predominantly found in Chile (Figure 2), however examples of this sub-lineage
188 were also identified in the United Kingdom. Similarly, a sub-lineage of BA.2 was mainly
189 found in Vietnam (Figure 2) but again an example of this genotype was found in the United
190 Kingdom.

191

192 Highly genetically similar samples were identified within lineages from globally distinct
193 sites, with branch lengths corresponding to ≤ 0.01% of the genome (approximately three
194 nucleotides) separating many samples. This limited diversity was further explored by
195 examining the relative likelihood of samples being from the same site with increasing
196 genomic variation. By this measure, the greatest dispersal was seen within the BA.4 lineage

197 (Figure 3). In total 357/1,133 of BA.4 samples were genotypically identical to another
198 sequenced sample. Among these, 244/357 (68.4%) were only identical to other samples from
199 the same study centre. Allowing for just one single nucleotide variant (SNV) reduced the
200 proportion of closely related samples to the same study centre (and country) to 302/634
201 (47.6%). The proportions reduced to 279/819 (34.1%) and to 201/923 (21.8%) when allowing
202 for two and three SNVs respectively. Only 210/1,133 (18.5%) did not have a related sample
203 allowing up to three SNVs. The other major lineages BA.2 and BA.5 likewise showed
204 evidence of genotypically identical sequences identified at different study sites, with
205 increasing evidence of dispersal as SNV threshold increased (Figure 3).
206
207 Cloud processing performance was measured by time elapsed from upload until completion
208 of processing. The median sample processing time was 30.6 minutes (IQR 13.1-69.9
209 minutes). The distribution of processing times was much wider than expected, and processing
210 delays occurred for several batches from Australia, United Kingdom and United States
211 (Figure S1). Some of these coincided with data centre outages during an extreme heat event
212 in the United Kingdom, while other delays occurred due to errors in a software update for the
213 upload portal, which were resolved by improvements implemented in the user side command
214 line interface software. Other data upload issues were identified and addressed in real time,
215 including metadata file formatting errors, availability of upload clients for all required
216 operating systems, and user interface display errors preventing batch release.
217
218 **Discussion**
219 This pilot study reveals the potential for cloud-based synchronous global data processing and
220 analysis for public health. The GPAS platform provides an example of unified assembly,
221 variant calling and analysis of data from participating centres, which can facilitate both local
222 and globally aggregated analysis. In this cross-sectional observational study, the aggregate
223 analysis conforms to contemporary observations of SARS-CoV-2 genetic

224 epidemiology[23,24,25,26,27], though these observations remain limited by global disparities
225 in sequencing volume.
226
227 The opportunistic sampling frame used means this study is likely to be geographically
228 incomplete, but even with this limitation we observed cosmopolitan global mixing of
229 Omicron lineages between study sites. Even a BA.4 sub-lineage which demonstrated
230 extensive replacement in one geographical location (Chile) was not completely restricted to
231 that site and could be identified in a geographically distant site (United Kingdom). During the
232 period of the study, the United Kingdom site performed genomic sequencing on all clinical
233 isolates which had tested positive for SARS-CoV-2 on PCR, and despite being dominated by
234 BA.5, was also able to identify examples of a BA.2 sub-lineage that was otherwise restricted
235 to Vietnam. Such mixing is likely the result of movement for travel and migration, as SARS-
236 CoV-2 related travel restrictions had been relaxed in the countries with participating sites.
237
238 Given evidence of global mixing and limited diversity between sites, viral genome
239 sequencing is limited in its ability to inform studies of local viral transmission in this
240 Omicron dominated era, where both rapid global mixing and transmission drives spread.[26]
241 Allowing for up to three different single nucleotide variants, the majority of isolates from the
242 dominant lineages BA.2, BA.4 and BA.5 had at least one closely related sample within each
243 countries' contemporaneously sequenced samples. Depending on lineage, between 10.0%
244 and 31.5% of genotypically identical sequences in this study were identified in different
245 countries and continents. The use of a shared bioinformatics analysis pipeline, as described
246 here, highlights how lineages can be simultaneously and confidently mapped across distant
247 sites. This captures the evidence of well described rapid global dispersal of Omicron lineages
248 with successive selective sweeps of these new Omicron variants over short time scales.[27]
249
250 These observations highlight the potential for deriving insights from viral sequencing at a few
251 globally distributed sites. While this study pertains to a specific point in the pandemic, it

raises the question of how to better design genomic surveillance balancing the public health need and available resources. It is recognised that a preponderance of sequence data came from high-income countries, who sequenced a proportion of cases ten-fold higher than that achieved in low- and middle-income counties during the first two years of the pandemic.[9] Nor was the data collected evenly over the course of pandemic. In retrospect this was not optimal. Better designed representative sampling would be immensely empowering and is likely possible through well-designed and suitably powered sampling frames. Ideally, this would be addressed through pandemic preparedness planning, enabling sequencing to be focussed at selected sentinel sites globally cognisant of national jurisdictions and the benefit of longitudinal data, as exemplified by a cohort design in the UK.[25]

Pathogen sequencing played a highly valuable role in the COVID-19 pandemic and remains a critical global and national capability. This is highlighted by studies of viral transmission that have yielded important insights into the modes of pathogen spread.[1-5] In addition, identification of emerging variants-of-concern guided public health and policy responses,[28] as well as informing assessments about the impact of emerging variants on the impact of therapeutics (such as monoclonal antibodies) and vaccines.[7,29] The present cross-sectional study has the limitation of lacking longitudinal data on how these findings changed over time. Nevertheless, this study demonstrates how simply structured sampling across the world is well placed to rapidly identify replacement by new lineages. The integration of genomic data with subject disease manifestation and severity data would enable the ready investigation of associations between genomic variation and severity of disease, as suggested by others [30] and predictive modelling for anticipating the course of future epidemics as described so effectively early in the SARS-CoV-2 pandemic [31].

The world is now better prepared to urgently initiate genomic pathogen surveillance. The available sequencing platform capacity is much improved, and the achievements of global data aggregation (illustrated by INSDC and GISAID databases) has alleviated a major

obstacle to effective genomic surveillance. A remaining obstacle is local turnkey bioinformatics to perform sequence assembly, variant calling and analysis. These are needed to deliver standardised and quality assured outputs – specifically the SARS-CoV-2 viral sequence FASTA file – for phylogenetic analysis either for global analysis by an international entity such as GISAID, or for local and national analyses. GPAS has developed an exemplar cloud-based, web-enabled service design that is optimised for operation by a laboratory scientist and removes dependence on in-laboratory bioinformatics expertise. Thus, this work reveals the benefit of establishing software for simply and automatically ingesting sequence data from a sequencer in a local lab and yielding data ready for local or international analysis. A software infrastructure such as the GPAS platform fills a gap in delivering global genomic surveillance. Flexible privacy protections allows users to control sharing of genomic data. GPAS is freely available to low- and middle-income countries, reducing some of the barriers to analysis services in low resource settings. Remaining challenges include high speed computer network access and local epidemiological tools, and expertise for detailed epidemiological analysis.

Conclusions

The COVID-19 pandemic has demonstrated the multiple ways in which viral sequencing can inform pandemic responses, but resources and capabilities are not equitably distributed. A service like GPAS democratises SARS-CoV-2 genomics, enabling researchers and laboratories, regardless of their location or level of bioinformatics expertise, to participate actively in global surveillance efforts. By providing accessible and user-friendly tools for sequence assembly, lineage assignment, and data sharing, it fosters collaboration, harmonises data, and ultimately enhances pandemic preparedness.

Funding statement

GPAS is free at the point of use for users in low- and middle-income countries. GPAS was developed specifically to minimise the need for bioinformatics staff (a scarce and expensive resource), to help bring genomic insights within reach for laboratories with limited resources.

GPAS operates as a non-profit organisation. All GPAS SARS-CoV-2 services were offered free of charge to all sites for the purposes of this study. Staff costs for the 5C analysis team were met by the University of Oxford.

The cloud infrastructure used in this study was donated by Oracle Corporation.

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 or shared with other users. GPAS 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 upload client selects only SARS-CoV-2 sequence reads and discards non SARS-CoV-2 reads (including 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.

Authors

All authors have seen and approved the manuscript

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

References

1. Li J, Lai S, Gao GF, Shi W. The emergence, genomic diversity and global spread of SARS-CoV-2. Nature. 2021 Dec;600(7889):408-418. doi: 10.1038/s41586-021-04188-6.
2. Wilkinson E, Giovanetti M, Tegally H, San JE, Lessells R, Cuadros D, et al. A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. Science. 2021 Oct 22;374(6566):423-431. doi: 10.1126/science.abj4336.
3. McCrone JT, Hill V, Bajaj S, Pena RE, Lambert BC, Inward R, et al. Context-specific emergence and growth of the SARS-CoV-2 Delta variant. Nature. 2022 Oct;610(7930):154-160. doi: 10.1038/s41586-022-05200-3.
4. Raghwani J, du Plessis L, McCrone JT, Hill SC, Parag KV, Thézé J, et al. Genomic Epidemiology of Early SARS-CoV-2 Transmission Dynamics, Gujarat, India. Emerg Infect Dis. 2022 Apr;28(4):751-758. doi: 10.3201/eid2804.212053.

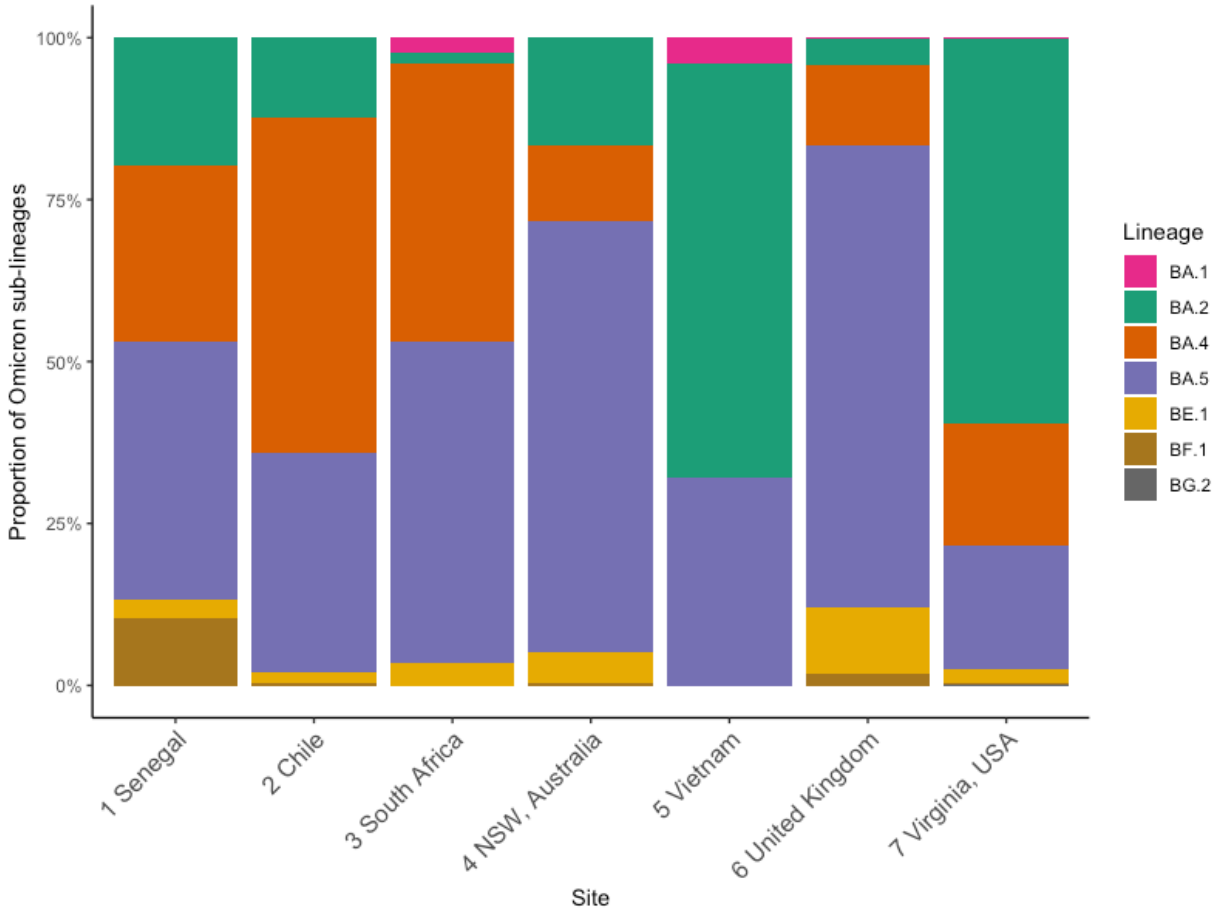
387 5. Snell LB, Fisher CL, Taj U, Stirrup O, Merrick B, Alcolea-Medina A, et al.
388 Combined epidemiological and genomic analysis of nosocomial SARS-CoV-2
389 infection early in the pandemic and the role of unidentified cases in transmission. Clin
390 Microbiol Infect. 2022 Jan;28(1):93-100. doi: 10.1016/j.cmi.2021.07.040
391 6. Tao K, Tzou PL, Nouhin J, Gupta RK, de Oliveira T, Kosakovsky Pond SL, et al. The
392 biological and clinical significance of emerging SARS-CoV-2 variants. Nat Rev
393 Genet. 2021 Dec;22(12):757-773. doi: 10.1038/s41576-021-00408-x.
394 7. Tuekprakhon A, Nutalai R, Dijokaite-Guraliuc A, Zhou D, Ginn HM, Selvaraj M, et
395 al. Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and
396 BA.1 serum. Cell. 2022 Jul 7;185(14):2422-2433.e13. doi:
397 10.1016/j.cell.2022.06.005. Epub 2022 Jun 9. PMID: 35772405; PMCID:
398 PMC9181312.
399 8. Merhi G, Koweyes J, Salloum T, Khoury CA, Haidar S, Tokajian S. SARS-CoV-2
400 genomic epidemiology: data and sequencing infrastructure. Future Microbiol. 2022
401 Sep;17:1001-1007. doi: 10.2217/fmb-2021-0207.
402 9. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, et al.
403 Global disparities in SARS-CoV-2 genomic surveillance. Nat Commun. 2022 Nov
404 16;13(1):7003. doi: 10.1038/s41467-022-33713-y
405 10. Global genomic surveillance strategy for pathogens with pandemic and epidemic
406 potential, 2022–2032. Geneva: World Health Organization; 2022. Licence: CC BY-
407 NC-SA 3.0 IGO
408 11. Inzaule SC, Tessema SK, Kebede Y, Ogwell Ouma AE, Nkengasong JN. Genomic-
409 informed pathogen surveillance in Africa: opportunities and challenges. Lancet Infect
410 Dis. 2021 Sep;21(9):e281-e289. doi: 10.1016/S1473-3099(20)30939-7
411 12. Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, et al. Global landscape of SARS-
412 CoV-2 genomic surveillance and data sharing. Nat Genet. 2022 Apr;54(4):499-507.
413 doi: 10.1038/s41588-022-01033-y.

414 13. Ohlsen EC, Hawksworth AW, Wong K, Guagliardo SAJ, Fuller JA, Sloan ML, et al.
415 Determining Gaps in Publicly Shared SARS-CoV-2 Genomic Surveillance Data by
416 Analysis of Global Submissions. Emerg Infect Dis. 2022 Dec;28(13):S85-S92. doi:
417 10.3201/eid2813.220780.
418 14. Hunt M, Swann J, Constantinides B, Fowler PW, Iqbal Z. ReadItAndKeep: rapid
419 decontamination of SARS-CoV-2 sequencing reads. Bioinformatics. 2022 Jun
420 13;38(12):3291-3293. doi: 10.1093/bioinformatics/btac311.
421 15. https://github.com/iqbal-lab-org/viridian_workflow/wiki
422 16. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment,
423 mutation calling and quality control for viral genomes. Journal of Open Source
424 Software, 2021 6(67): 3773, https://doi.org/10.21105/joss.03773
425 17. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al.
426 Assignment of epidemiological lineages in an emerging pandemic using the pangolin
427 tool. Virus Evol. 2021 Jul 30;7(2):veab064. doi: 10.1093/ve/veab064.
428 18. https://github.com/connor-lab/aln2type
429 19. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus
430 associated with human respiratory disease in China. Nature. 2020
431 Mar;579(7798):265-269. doi: 10.1038/s41586-020-2008-3
432 20. Mazariegos-Canellas O, Do T, Peto T, Eyre DW, Underwood A, Crook D, et al
433 BugMat and FindNeighbour: command line and server applications for investigating
434 bacterial relatedness. BMC Bioinformatics. 2017 Nov 13;18(1):477. doi:
435 10.1186/s12859-017-1907-2.
436 21. https://github.com/davidhwyllie/findNeighbour4
437 22. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis
438 of large phylogenies Bioinformatics. 2014; 30(9),1312-3 doi:
439 10.1093/bioinformatics/btu033.
440 23. World Health Organisation Coronavirus (COVID-19) Dashboard,
441 https://covid19.who.int [Accessed 19th April 2023]

442 24. Tosta S, Moreno K, Schuab G, Fonseca V, Segovia FMC, Kashima S, et al. Global
443 SARS-CoV-2 genomic surveillance: What we have learned (so far). Infect Genet
444 Evol. 2023 Mar;108:105405. doi: 10.1016/j.meegid.2023.105405.
445 25. Foulkes S, Monk EJM, Sparkes D, Hettiarachchi N, Milligan ID, Munro K, et al.
446 Early Warning Surveillance for SARS-CoV-2 Omicron Variants, United Kingdom,
447 November 2021-September 2022. Emerg Infect Dis. 2023 Jan;29(1):184-188. doi:
448 10.3201/eid2901.221293.
449 26. Xu Y, Liu T, Li Y, Wei X, Wang Z, Fang M, et al. Transmission of SARS-CoV-2
450 Omicron Variant under a Dynamic Clearance Strategy in Shandong, China. Microbiol
451 Spectr. 2023 Mar 14;11(2):e0463222. doi: 10.1128/spectrum.04632-22.
452 27. Markov PV, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, et al. The
453 evolution of SARS-CoV-2. Nat Rev Microbiol. 2023 Apr 5. doi: 10.1038/s41579-
454 023-00878-2.
455 28. WHO Technical Advisory Group on SARS-CoV-2 Virus Evolution, Updated working
456 definitions and primary actions for SARS-Cov-2 variants, World Health Organisation,
457 March 2023. https://www.who.int/publications/m/item/updated-working-definitions-
458 and-primary-actions-for--sars-cov-2-variants [Accessed 19th April, 2023]
459 29. Aggarwal A, Akerman A, Milogiannakis V, Silva MR, Walker G, Stella AO, et al.
460 SARS-CoV-2 Omicron BA.5: Evolving tropism and evasion of potent humoral
461 responses and resistance to clinical immunotherapeutics relative to viral variants of
462 concern. EBioMedicine. 2022 Oct;84:104270. doi: 10.1016/j.ebiom.2022.104270.
463 30. Oude Munnink BB, Worp N, Nieuwenhuijse DF, Sikkema RS, Haagmans B,
464 Fouchier RAM, et al The next phase of SARS-CoV-2 surveillance: real-time
465 molecular epidemiology. Nat Med. 2021 Sep;27(9):1518-1524. doi: 10.1038/s41591-
466 021-01472-w.
467 31. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, et al. Assessing
468 transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature. 2021
469 May;593(7858):266-269. doi: 10.1038/s41586-021-03470-x.

470 **Figures**

471



472

473

474 **Figure 1:** Omicron sub-lineages by study site. Proportion of samples assigned to Omicron

475 sub-lineage within each study site, where collection date was in June or July 2022.

476

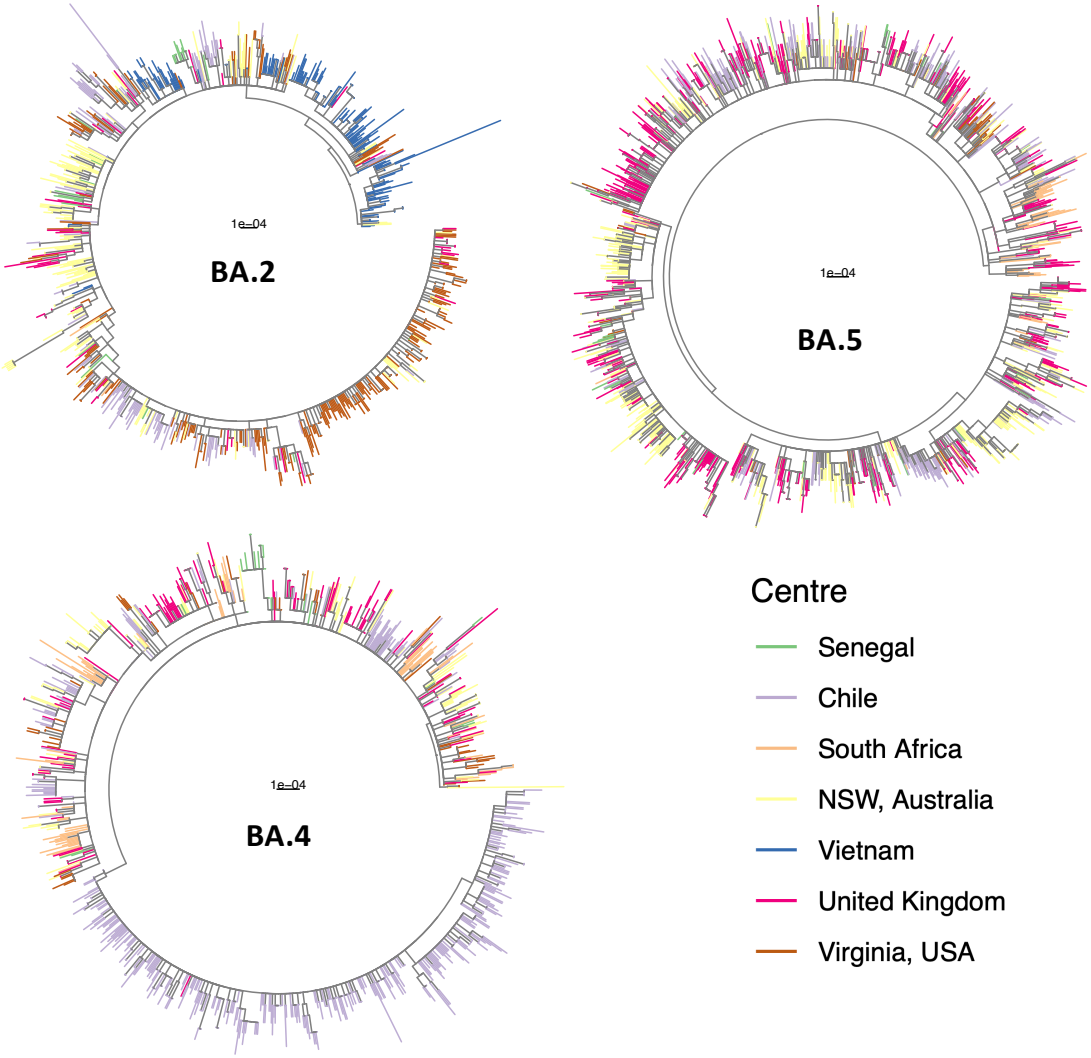


Figure 2: Maximum likelihood phylogeny of isolates from each major lineage (BA.2, BA.4, BA.5) found in seven centres. Final branches are coloured according the centre in which they were sequenced.

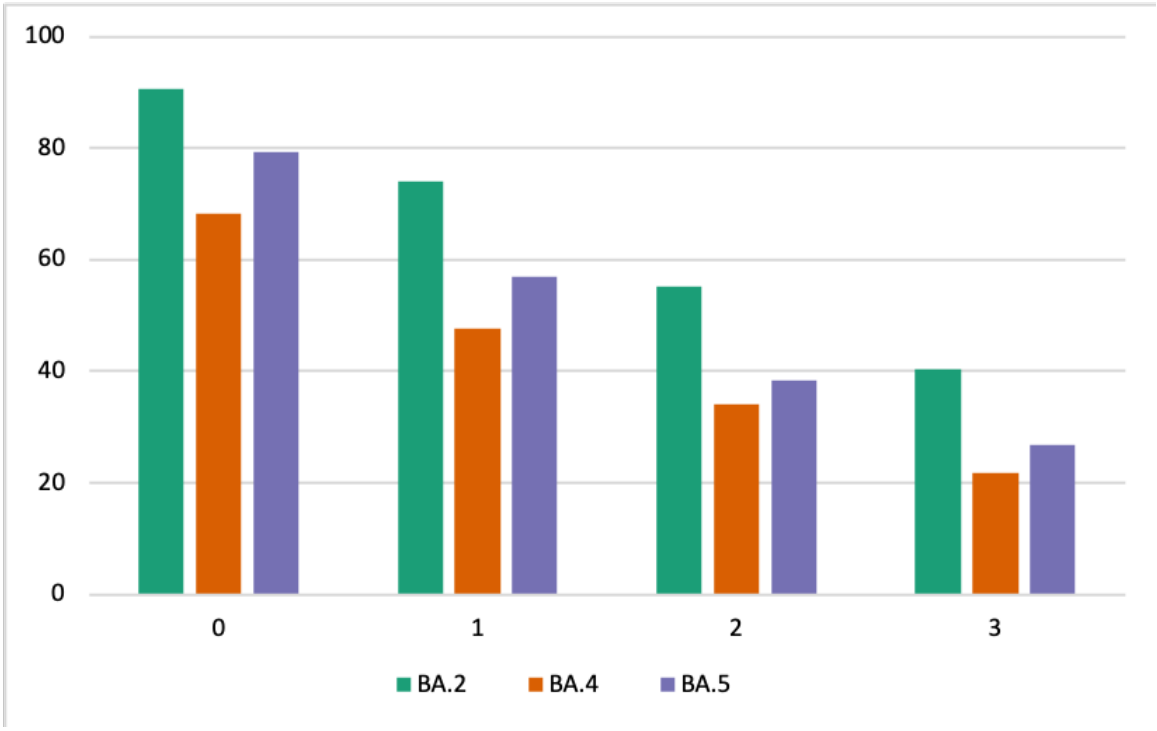
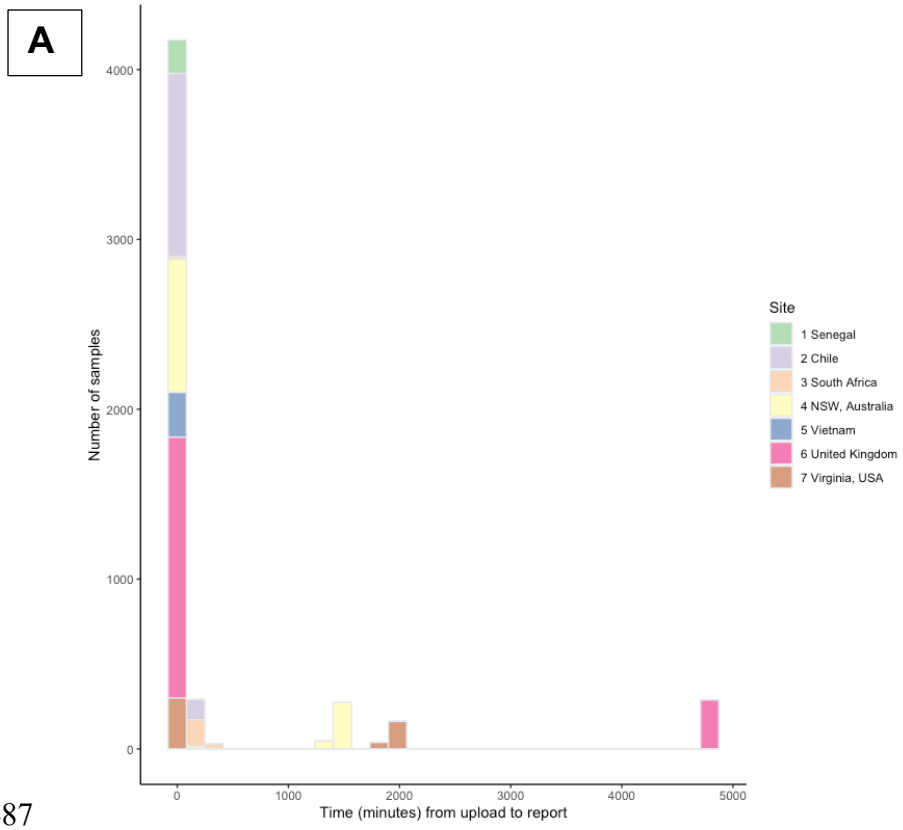


Figure 3: Percentage of related samples which are restricted to the same study site, allowing for 0, 1, 2 and 3 SNVS as the threshold for relatedness, for 3 major lineages (BA.2, BA.4 and BA.5).



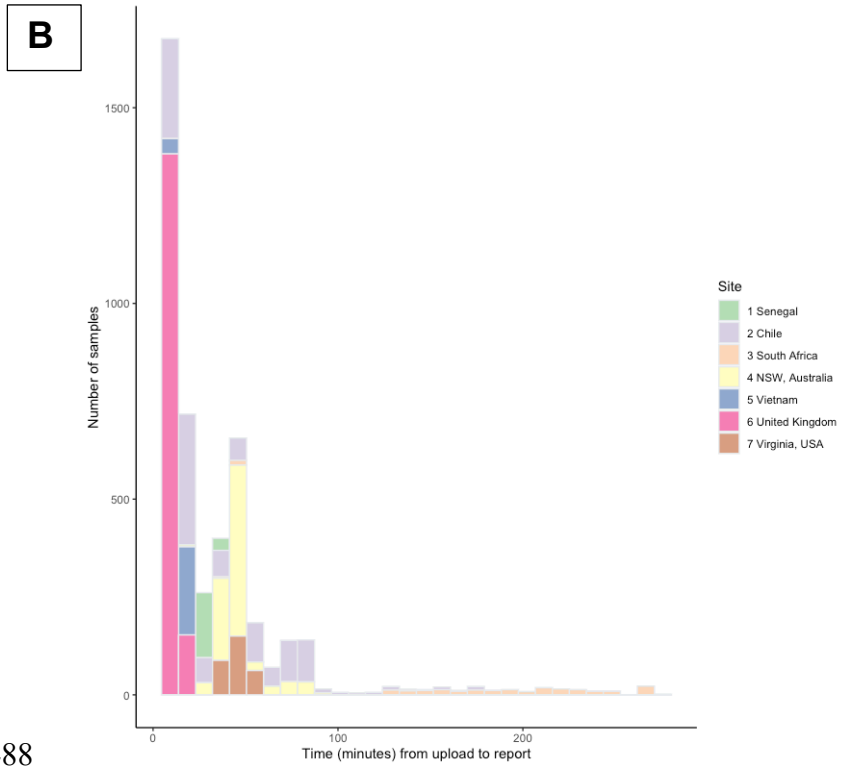


Figure S1: Distribution of processing time (in minutes) for each sample in GPAS pipeline according to submitting site (a) all samples and (b) samples processed in under 1000 minutes